Bioinformatic Tools for Next Generation DNA Sequencing

Development and Analysis of Model Systems

Sønderkær, Mads

Publication date:
2012

Document Version
Early version, also known as pre-print

Bioinformatic Tools for Next Generation DNA Sequencing

Development and Analysis of Model Systems

PhD thesis

By

Mads Sønderkær

August 2012

AALBORG UNIVERSITY
DENMARK

Preface

This thesis is submitted to The Doctoral School of Engineering and Science at Aalborg University as documentation of PhD studies that was carried out at the Department of Biotechnology, Chemistry and Environmental Engineering under the supervision of Associate Professor Kåre Lehmann Nielsen and was funded by The Danish Council for Strategic Research, Programme Committee on Strategic Growth Technologies under the project "*SEQNET-software tools for next generation sequencing*"; grant number 2106-07-0021 and Aalborg University.

In the current thesis, an introduction to DNA sequencing technologies and analysis of next generation sequencing data is given in chapter 1. Following this, a description of the development and application of custom tools for DeepSAGE gene expression analysis is given in chapter 2. In the following three chapters, a description of development and use of the two plant model systems *Lotus japonicus* and *Solanum tuberosum* in transcriptome data analysis and an analysis of variance in gene expression data sets is presented. Finishing, a general discussion of the obtained results is presented in chapter 6. Lists of published and planned publications related to the current thesis can be found in appendices A, C and D. A complete list with descriptions of all programs developed as a part of this thesis can be found in appendix B. Moreover, all programs can be found on the enclosed CD. Due to the size of next generation sequencing data sets, none used in the current thesis have been enclosed. These are of course available by request to the author (mson@bio.aau.dk).

First and foremost, I would like to thank my supervisor Associate Professor Kåre Lehmann Nielsen, for his excellent and insightful guidance throughout this PhD study, and for finding time for our sometimes long but fruitful discussions whenever needed. I also thank Assistant Professor Stig Uggerhøj Andersen, PhD and Associate Professor Elena Simona Radutoiu, PhD from Århus University for the collaboration with the *Lotus japonicus* project – I truly hope all our efforts will be fruitful. Thanks also go out to members of the potato genome sequencing consortia, especially members of the transcriptome group – I learned a lot working with you guys. I would also like to thank the members of the Functional Genomics Group for all your support, especially to Anne, for doing all the work in the lab, I didn't have time for, and to Mette for the careful and highly appreciated reading and commenting of the thesis.

Last but not least, I would like to thank my lovely wife Jannie. The topic of this thesis has developed like a tidal wave – you are the main reason I was able to able to surf.

Mads Sønderkær, August 2012

Summary

Next generation DNA sequencing has within the past years led to an increase of biological data with several orders of magnitude, particularly within genomics and transcriptomics. Therefore, extensive bioinformatic frameworks have to be carefully developed to ensure correct biological interpretation of the data. This thesis concerns the development of such a bioinformatic framework to store and analyze DNA sequence based transcriptome data for two plant model systems, namely *Lotus japonicus* (*L. japonicus*) and *Solanum tuberosum* (potato).

First, a bioinformatic framework for data preprocessing and storage was developed. Following this, algorithms for sequence filtering and sequence error correction was developed for tag based sequencing data, such as serial analysis of gene expression (SAGE) data and these where subsequently evaluated. It was shown that the frequently applied sequence quality filtering methods performs poorly because they lead to a high degree of data loss and a low degree of reduction of the data complexity. The optimal pre-processing method was concluded to be a combination of SAGEscreen sequence error correction and singleton removal. To maximize the fraction of data useable for biological interpretations, methods for tag annotation (the process of matching a short 21 base pair (bp) sequence tag to its corresponding mRNA transcript) was developed and evaluated. Here it was found that allowing a single nucleotide mismatch, no decrease in fidelity could be detected when analyzing real data sets. However, allowing a single mismatch significantly increased the amount of tags that could be annotated, probably reflecting the real genetic variance of strains and individuals versus model sequences in the databases.

The developed methods were implemented in the data analysis of a tag based transcriptome study of the interaction between *L. japonicus* and nitrogen fixating bacteria. The purpose of the study was to elucidate the complex signaling pathway involved in the symbiotic development of *L. japonicus* nodules. *L. japonicus* wild type and mutant plants with altered nodule developments at different stages were sampled and analyzed, hereby enabling detection of differences between the mutants and wild type plants at different stages in nodulation facilitating an elucidation of the complex signaling pathway. First, a transcriptome model of *L. japonicus* based on the genome sequence was constructed greatly improving the annotation of sequence tags. Secondly, the tag based transcriptome data set was compared with a similar gene expression data set generated using cDNA microarray technology. A fairly good correlation between the results of the two data sets was found. Following this, a time series analysis of the wild type plant revealed an early specific induction of several genes involved in defense or cell wall metabolism. Furthermore, a transcript encoding an Asparagine synthetase was found to be ~ 50 fold up-regulated in nodules leading to further investigation of the expression levels of genes known to be involved in asparagine assimilation. Results indicate that Asparagine synthetase is a key enzyme in asparagine assimilation, possibly by direct incorporation of ammonia, and thus important for nitrogen uptake in symbiotic legumes.

Transcriptome data sets are by nature "noisy", i.e. variation in the observed gene expression levels exist between samples. Due to the cost of transcriptomic experiments, studies investigating this have been performed using a relative low number of replicates. Here, a high replicate tag based transcriptome experiment (2 x 47) of two potato cultivars is presented. Data

was analyzed to investigate the variation of tag based gene expression data and its impact on determination of differential gene expression. Moreover, the technical variation of the tag based method DeepSAGE, and mRNAseq was compared. Analysis of the high replicate data set revealed that a substantial amount of variation was present in the tag based transcriptome data set causing a large decrease in the power to detect differential expression when lowering the number of replicates. However, the specificity of the detection was maintained. The comparison between data sets generated using DeepSAGE and mRNAseq revealed that the observed variation in DeepSAGE data sets to a large extent was technical, and not caused by differences between biological samples. Importantly, almost no technical variation was found in the mRNAseq data set.

Finally, a large part of the current PhD-project has been in connection with the potato genome sequencing project, which was a joint effort between 26 international research groups. The genome sequence was published in July 2011 in Nature. As a part of AAU contribution to the potato genome sequencing project, and of the current thesis, an algorithm for prediction of un-translated gene regions (UTRs) in the genome sequence was developed. Furthermore, experimental validation of gene models, investigation of the quality of the potato genome annotation, manual gene validation and curation of a small subset of genes, an overview transcriptome analysis of potato gene expression, and a detailed transcriptome analysis of the starch metabolism genes, comparing the two genotypes sequenced was performed. It was found that similar results of experimental UTR prediction were obtained when using the developed algorithm and when using Cufflinks as an mRNA prediction method. The potato genome was the first larger eukaryotic genome that was annotated using an mRNAseq assisted method. Analyses presented here, showed that this improved the quality of gene annotation, since a substantial amount of noise (mis-called gene transcripts) that could be filtered out was present in the *de novo* based annotation. The quality of the annotation was assessed by manual curation of 167 gene models. This analysis showed that 74 % of the gene models were correctly predicted and that an additional 18 % could be manually curated based on evidence provided by the mRNAseq data. Finally, the expression analysis of genes involved in starch metabolism revealed several candidate gene loci, which could explain the phenotypic differences between the tubers of the two genotypes. Moreover, a wide dynamic range and a large degree of tissue specific expression between different gene loci were observed. These results could facilitate the development of higher yielding potato cultivars, e.g. in regards to selection or in the form of gene modification.

Resume

Næste generation DNA sekventering har indenfor de seneste år ledt til en forøgelse af biologiske data med adskillige størrelsesordner især indenfor *genomics* og *transcriptomics*. Omfattende og omhyggelig udvikling af bioinformatiske værktøjer til understøttelse af korrekt biologisk fortolkning af data er derfor nødvendig. Denne afhandling drejer sig om udvikling af sådanne bioinformatiske værktøjer til lagring og analyse af DNA-sekvensbaseret transkriptomdata indenfor to plantemodelsystemer, *Lotus japonicus* (*L. japonicus*) og *Solanum tuberosum* (kartoffel).

Først blev bioinformatiske værktøjer til preprocessering og lagring af data udviklet. Efterfølgende blev algoritmer til sekvensfiltrering og sekvensfejlretning af tagbaseret sekvensdata såsom *serial analysis of gene expression* (SAGE) udviklet og evalueret. Det blev påvist, at den oftest anvendte filteringsmetode baseret på sekvenskvalitet præsterer dårligt, da de i høj grad medfører tab af data og kun reducerer datakompleksiteten i lav grad. Det blev konkluderet, at den optimale preprocesseringsmetode var en kombination af SAGEscreen fejlretning og fjernelse af *singletons*. For at maksimere mængden af data, som er brugbart til biologisk fortolkning, blev metoder til tagannotering (det at matche et kort 21 basepar (bp) sekvenstag til dets tilsvarende mRNA-transkript) udviklet og evalueret. Her blev det fundet at tilladelse af en enkelt forkert baseparsing ikke medførte til et fald i nøjagtigheden af tagannoteringen, når rigtige datasæt blev analyseret. Derimod medførte tilladelse af en enkelt forkert baseparsing en signifikant forøgelse af antallet af sekvenstags som kunne annotereres. Dette reflekterer formodentligt den genetiske variation af stammer og individer versus modelsekvenserne i databaserne.

De udviklede metoder blev implementeret i dataanalysen af et sekvenstagbaseret transkriptomstudie af samspillet mellem *L. japonicus* og nitrogenfikserende bakterier. Formålet med studiet var at belyse den komplekse kemiske signalvej involveret i den symbiotiske udvikling af *L. japonicus nodules*. Prøver af *L. japonicus* vildtype- og mutantplanter med ændret udvikling af *nodules* på forskellige stadier blev indsamlet og analyseret. Hermed blev detektion af forskelle mellem mutantplanterne og vildtypeplanterne på forskellige stadier af nodulationen muliggjort, hvilket kan hjælpe til belysningen af den komplekse kemiske signalvej. Først blev en transkriptommodel af *L. japonicus* baseret på genomsekvensen lavet. Denne forbedrede i høj grad annotering af sekvenstags. Hernæst blev det tagbaserede transkriptomdatasæt sammenlignet med lignende et genekspressionsdatasæt genereret med cDNA microarray teknologien. En ganske god korrelation mellem resultaterne fra de to datasæt blev fundet. Efterfølgende afslørede en genekspressionsanalyserne af en tidsserie af vildtypeplanterne en specifik tidlig induktion af adskillige gener involveret i forsvar og cellevægsmetabolisme. Ydermere blev et transkript, som koder for en asparaginsyntase, fundet til at være ~ 50 gange opreguleret i *nodules*. Dette førte til en ydereligere undersøgelse af genekspressionsniveauerne af gener som er involveret i assimilation af asparagin. Resultaterne indikerer at asparaginsyntase er et centralt enzym for assimilation af asparagin, muligvis ved direkte inkorporation af ammoniak, og hermed vigtig for optagelsen af nitrogen i symbiotiske bælgplanter.

Transkriptomdatasæt er i naturen "*noisy*". Det vil sige, at der er variation i de observerede genekspressionsværdier mellem forskellige prøver. Grundet de høje omkostninger er genek-

spressionsstudier, hvor variationen er blevet undersøgt, blevet udført med et relativt lavt antal af replikater. Her præsenteres et højt replikeret (2 x 47) sekvenstagbaseret transkriptomekperiment af to kartoffelsorter. Data blev analyseret for at undersøge variationen af sekvenstagbaserede genekspressionsdata og dens indflydelse på bestemmelsen af differentielt regulerede gener. Ydermere blev den tekniske variation i datasæt generet med den sekvenstagbaserede metode DeepSAGE og mRNAseq sammenlignet. Analysen af det højt replikerede datasæt viste, at der var en væsentlig mængde variation i sekvenstagbaseret transkriptomdata. Dette medfører en stor reduktion i evnen til at detektere differentiel ekspression, når antallet af replikater formindskes. Dog opretholdels dektektionsspecificiteten. Sammenligning af datasæt genereret med DeepSAGE eller mRNAseq metoderne viste at den observerede variation i DeepSAGE data i høj grad var af teknisk karakter og ikke skyldtes forskelle mellem biologiske prøver. Nok så vigtigt blev der næsten ikke fundet nogen teknisk variation i mRNAseq datasættet.

En stor del af denne ph.d. er lavet i forbindelse med kartoffelgenomsekventeringsprojeketet, som var en fælles indsats mellem 26 internationale forskningsgrupper. I juli 2011 blev genomsekvensen publiceret i Nature. Som en del af AAUs bidrag til kartoffelgenomsekventeringsprojeketet og som en del af denne afhandling, blev en algoritme til forudsigelse af utranslaterede regioner (UTRer) i genomsekvensen udviklet. Ydermere blev eksperimentel validering af genmodeller, undersøgelse af kvaliteten af annoteringen af genomsekvensen, manuel validering og kurering af en mindre delmængde gener, en overbliksanalyse af kartoffelens genekspression og en detaljeret transkiptomanalyse af stivelsessyntesen, hvor de to sekventerede genotyper blev sammenlignet. Lignende resultater blev opnået for forudsigelse af UTR regioner ved brug af den udviklede algoritme til sammenligning med Cufflinks, som er en metode til forudsigelse af mRNA transkripter. Kartoffelgenomet var det første større eukaryote genom, som blev annoteret ved brug af en mRNAseq understøttet metode. Analyser som præsenteres her viste at dette gav en annotering af gener i høj kvalitet, men at en anseelig del af støj (forkert annoterede gentranskripter), som er nødvendigt at filtrere for var til stede i annoteringen. Kvaliteten af annoteringen blev estimeret ved manuel kurering af 167 genmodeller. Denne analyse viste at 74 % af genmodellerne var korrekt forudsagte, og at ydereligere 18 % kunne kureres manuelt baseret på mRNAseq datasættet. Endeligt afdækkede ekspressionensanalysen af gener involveret i stivelsessyntesen adskillige kandidatgener, som kunne forklare den fænotypiske forskel af knolde fra de to forskellige genotyper. Desuden blev der observeret en bred dynamisk rækkevide af ekspressionsniveauer samt en stor grad af vævsspecifik ekspression mellem forskellige genloci. Disse resultater kan bidrage til udviklingen af højere ydende kartoffelsorter, fx i relation til forædling eller genmodifikation.

List of abbreviations

Abbreviations used in the current thesis are listed below in alphabetical order.

Abbreviation	Description
AAU	Aalborg University
AMI	after *Mesorhizobium loti* inoculation
ATP	Adenosine-5'-triphosphate
AU	Århus University
BAC	Bacterial Artificial Chromosome
BGI	Beijing Genomics Institute
BLAST	Basic Local Alignment Search Tool
bp	Base Pair
BWT	Burrow-Wheeler Transformation
cDNA-AFLP	Complementary DNA - Amplified fragment length polymorphism, an AFLP-based transcript profiling method (Bachem *et al.*, 1996).
CDS	Coding Sequence
Chip-seq	chromatin immunoprecipitation with next generation sequencing
CPM	Count per million
CV	Coefficient of Variance. Normalized measure of dispersion (STDV/EXP) (Hendricks & Robey, 1936).
D	Dispersion
DDBJ	DNA Data Bank of Japan
ddNTPs	dideoxynucleotide triphosphates
DE	Differential gene Expression
DM	DM1-3 516R44. A double monoploid potato variety
DNA	Deoxyribonucleic acid
EC number	Enzyme Commission number
EST	Expressed Sequence Tag
EXP	Mean Expression Level
FKPM	Expected Fragments per Kilobase of transcript per Million fragments sequenced
GA	Illumina Genome Analyzer
GA	Genome Analyzer
Gb	Giga base
GO	Gene Ontology
HT-SuperSAGE	high throughput SuperSAGE
IT	Infection Threads
kb	Kilobase
KEGG	The Kyoto Encyclopedia of Genes and Genomes
Mb	Mega base
miRNA	Micro RNA
MPSS	Massively Parallel Signature Sequencing
mRNA	Messenger Ribonucleic acid
mRNAseq	Sequencing of messenger RNA using Next Generation Sequencing
N50	The size *N* such that 50% of the genome is contained in contigs of size *N* or greater (Zimin *et al.*, 2009).
NcRNA	Non coding Ribonucleic acid
NF	Nodulation factor
NGS	Next Generation Sequencing
NSE	Normalized standard error
nt	Nucleotide
PC	Principal Component
PCA	Principal Component Analysis
PCR	polymerase chain reaction
PE	Parried-end
PET	Paired-end tag
PGSC	The Potato Genome Sequencing Consortium
PLS-da	Partial Least Squares Discriminant Analysis partial least squares regression (PLS)
PLS-R	Partial Least Squares Regression
PUT	PlantGDB-assembled Unique Transcripts
PUT	PlantGDB-assembled unique transcripts
R^2	Goodness of fit
RH	RH89-039-16. A diploid heterozygous potato variety
RKPM	Reads per Kilobase of exon model Per Million mapped read

SAGE	Serial Analysis of Gene Expression
SFF	Standard Flowgram Format
SNP	Single Nucleotide Polymorphism
STDV	Standard Deviation
TAIR	The Arabidopsis Information Resource
TC	Tentative consensus sequence
UTR	Un-translated Region
WGS	Whole Genome Shotgun
ρ_p	Pearson's coefficient of correlation
ρ_s	Spearman's coefficient of correlation

Contents

Motivation

The “omics” research field represents a shift of paradigm in molecular biology. Previously molecular biology data was scarce because sample handling and manipulation was expensive and time consuming. Consequently, data analysis and validation was relatively easy because of the limited number of samples and measurements. In contrast, “omics” technologies provide huge amounts of data, often noisy, which cannot be organized, stored and analyzed without extensive bioinformatic frameworks that have to be carefully developed to ensure correct biological interpretation of the data. Next generation DNA sequencing has since its introduction in 2005 profoundly accelerated this trend by increasing the cost/efficiency ratio of data production with several orders of magnitude, particularly within genomics and transcriptomics.

This thesis concerns the development of such a bioinformatic framework to store and analyze DNA sequence based transcriptome data within four main projects: 1) Custom tools development for DeepSAGE gene expression analysis of multiple elite potato cultivars during development, drought and disease stress. 2) Data analysis of *Lotus japonicus* interaction with nitrogen fixating bacteria. 3) A high replicate experiment of two potato cultivars to analyze the variation of tag-based gene expression data and its impact on determination of differential gene expression; and 4) Gene expression analysis and experimental gene prediction and validation of potato cultivars DM and RH as part of the potato genome sequencing.

The development of bioinformatic frameworks for the analysis of transcriptome data has been a "hot topic" in research during the course of this thesis. Therefore, a number of bioinformatic tools and algorithms have been developed by other groups and these have been integrated into this project where useful.

Chapter 1

Background

1.1 Development of DNA Sequencing Technologies

1.1.1 First generation Technologies – the Past

DNA sequencing was pioneered by Sanger, Nicklen, and Coulson who developed the chain termination method in 1977 (Sanger, Nicklen & Coulson, 1977) using dideoxynucleotide triphosphates (ddNTPs), and by Gilbert and Maxam, who developed a method based on chemical modification of the DNA and subsequent cleavage at specific bases (Maxam & Gilbert, 1977). Sanger sequencing was used to determine the DNA sequence the 5375 bp genome of the bacteriophage phi X 174 in 1977 - the first genome ever sequenced (Sanger *et al.*, 1977). Further improvements to the method including semi-automation were developed by Smith *et al.* in the 1980's introducing four-color Sanger sequencing, using four fluorescently labeled ddNTPs for each DNA base, enabling optical detection (Smith *et al.*, 1986). This method in combination with capillary electrophoresis, was the key technology behind the first fully automated DNA sequencing system, the ABI 370, marketed by Applied Biosystems Inc. (Now Life Technologies) in 1986 (Marziali & Akeson, 2001). Through the 1990s the Sanger technology was refined and improved driven by the Human genome sequencing project. In 1998, the first two sequencing systems utilizing 96 capillary array electrophoresis (Huang, Quesada & Mathies, 1992) was marketed, namely the ABI 3700 by Applied Biosystems Inc. and the MegaBace by Amersham Pharmacia Biotech (now GE Healthcare Life Sciences). Additional improvements of the Sanger sequencing method continued though the 00's, enabling parallelized sequencing of up to 384 DNA fragments up to ~ 1,000 bp in length with accuracy higher than 99.99 % (Shendure & Ji, 2008).

In 1988, the first "sequencing by synthesis" method was reported by Eward D. Hyman, namely pyrosequencing (Hyman, 1988). The method utilizes real-time detection of pyrophosphate release upon nucleotide (nt) incorporation, a concept invented by Pål Nyrén (Nyren, 1987; Nyren & Lundin, 1985). It took his and the teams of Ronaghi and Uhlen more than ten years of development, before they could introduce a working sequencing method in 1996 (Nyren, 2007; Ronaghi, Uhlen & Nyren, 1998; Ronaghi *et al.*, 1996). In 1999, the first automated pyrosequencing system was made commercially available by Pyrosequencing AB (later Biotage AB, now a part of Qiagen) (Nyren, 2007).

A major hallmark for DNA sequencing was the initiation of the Human Genome Project, which formally began in 1990 (U.S. Department of Energy Genome Programs, 2011). A draft sequence was completed in 2001 (Lander *et al.*, 2001; Venter *et al.*, 2001) and the finished sequence was completed in (Collins *et al.*, 2004) 2003. The project ended up costing nearly 3 billion US dollars (the National Human Genome Research Institute, 2010), but also triggered research efforts in fundament bioinformatic algorithms and in development of cheaper and increasingly higher-throughput sequencing techniques (Ansorge, 2009). One of the results of these efforts became what we today refer to as next generation sequencing (NGS) technologies.

1.1.2 Next (Second) Generation Technologies – the Present

In the beginning of this millennium, several NGS technologies were marketed. This has caused an explosion of generated DNA sequence data; well-illustrated by the number total of nucleotides deposited in GenBank® (Benson *et al.*, 2011) in the past two decades, cf. Figure 1-1.

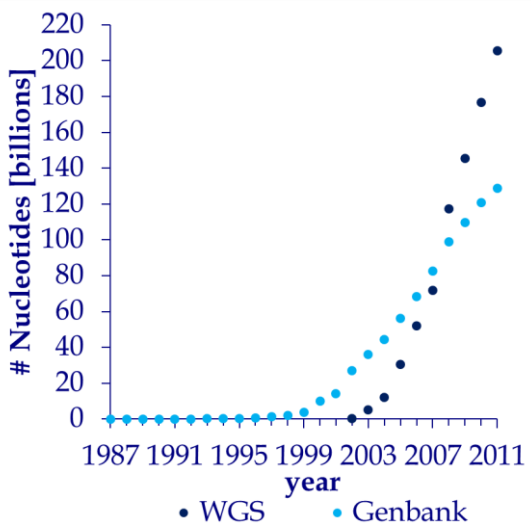


Figure 1-1 Number of nucleotides deposited in GenBank® (Benson *et al.*, 2011). Statistics retrieved from the DNA Data Bank of Japan (DDBJ) (DNA Data Bank of Japan, 2011). WGS = whole genome shotgun sequences.

For many years, the cost reduction of DNA sequencing halved ~ every two years, which is similar to the increase of computer power, as stated by Moore in 1965 (Moore, 1965). However, the introduction of NGS technologies have caused a further dramatic cost reduction of DNA sequencing, illustrated by costs associated with DNA sequencing performed at the National Human Genome Research Institute (Wetterstrand, 2011), cf. Figure 1-2.

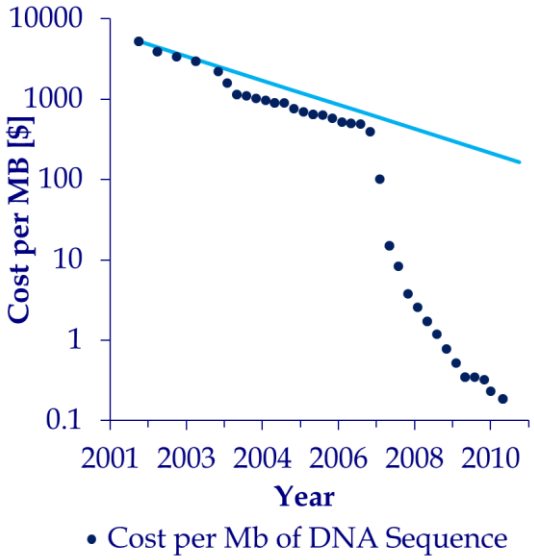


Figure 1-2 Costs associated with DNA sequencing performed at the National Human Genome Research Institute (Wetterstrand, 2011). The light blue line represents cost of sequencing following the same pattern as Moore's law (Moore, 1965). Notice the logarithmic scale of the Y-axis.

The fact that NGS technologies have a broad range of applications, cf. section 1.2, and already have inspired novel uses beyond the original purpose were some of the reasons why Nature Methods in 2007 selected NGS as method of the year (Nature Methods, 2008). While a complete review of all current and future sequencing technologies is beyond the scope of this

thesis the three NGS technologies dominating today's market; 454 Sequencing (19%), Illumina Sequencing (60 %), and SOLiD™ Sequencing (19 %) (Herper, 2010), will be described in detail and compared, while others only will be described shortly.

1.1.2.1 454 Sequencing

In 2005, the first NGS system, 454 Sequencing, was marketed by 454 Life Sciences (now Roche) (Margulies *et al.*, 2005). The system is based on the same principles as pyrosequencing, cf. section 1.1.1, hence pyrophosphate detection (Nyren & Lundin, 1985).

An overview of the workflow of 454 sequencing is given in Figure 1-3. Library construction is accomplished by DNA fractionation using nebulization, and subsequently enzymatically blunt-ending and adaptor ligation, or by *in vitro* transposition using the Nextera technology (Osborne & Slatter, 2011) . One of the adaptors contains a 5' biotin tag, enabling binding onto streptavidin coated beads. Hereafter, clonal amplification of the DNA fragments is carried out by so-called emulsion polymerase chain reaction (PCR), where the beads are separated by a w/o emulsion and the amplification occurs in oil droplets containing a PCR reaction mixture (Dressman *et al.*, 2003). The emulsion is then broken, and the beads are subsequently treated with denaturant for removal of untethered DNA strands, and finally hybridization-based enrichment of template-carrying beads is performed, cf. Figure 1-3 (Margulies *et al.*, 2005). The enriched beads are loaded in a picotiter plate containing 28 µm diameter wells only allowing one bead per well. This enables a fixed bead position at which each sequencing reaction can be monitored. Smaller beads containing immobilized sequencing enzymes (ATP sulfurylase, luciferase, and apyrase) are also added. At each sequencing cycle, single species nucleotides are flowed across the plate. At strands where the DNA polymerase-catalyzed addition of one or more nucleotides is possible, pyrophosphates are released. This enables oxidization of luceferin by the action of ATP sulfurylase and luciferase, hereby omitting light, which is detected by a CCD sensor (Ronaghi, Uhlen & Nyren, 1998). The sequencing cycle is finalized by degradation of unincorporated nucleotides by the action of apyrase (Margulies *et al.*, 2005). Image and signal processing occurs as part of a sequencing run, and the end output is Standard Flowgram Format (SFF) files containing the flowgrams for individual reads, the base called read sequences, and per-base quality scores (The DNA Sequencing Facility, 2010).

A major limitation of the 454 technology is base determination in homopolymeric regions (Huse *et al.*, 2007). This is caused by the fact that multiple nucleotide incorporations can occur in the same sequencing cycle, and the number of bases incorporation therefore must be determined by the signal intensity. As a consequence the dominant errors of 454 sequencing are insertion/deletion errors (Shendure & Ji, 2008). Margulies *et al.* initially reported linearity in signal to number of bases was preserved up to 8 nucleotide homopolymers (Margulies *et al.*, 2005). However, a recent study by Gilles *et al.* done using the newest 454 sequencing platform, the GS FLX+ system, showed that insertion/deletion errors contributed to 94% of an overall error rate of 0.53% on the first 100 bp (Gilles *et al.*, 2011).

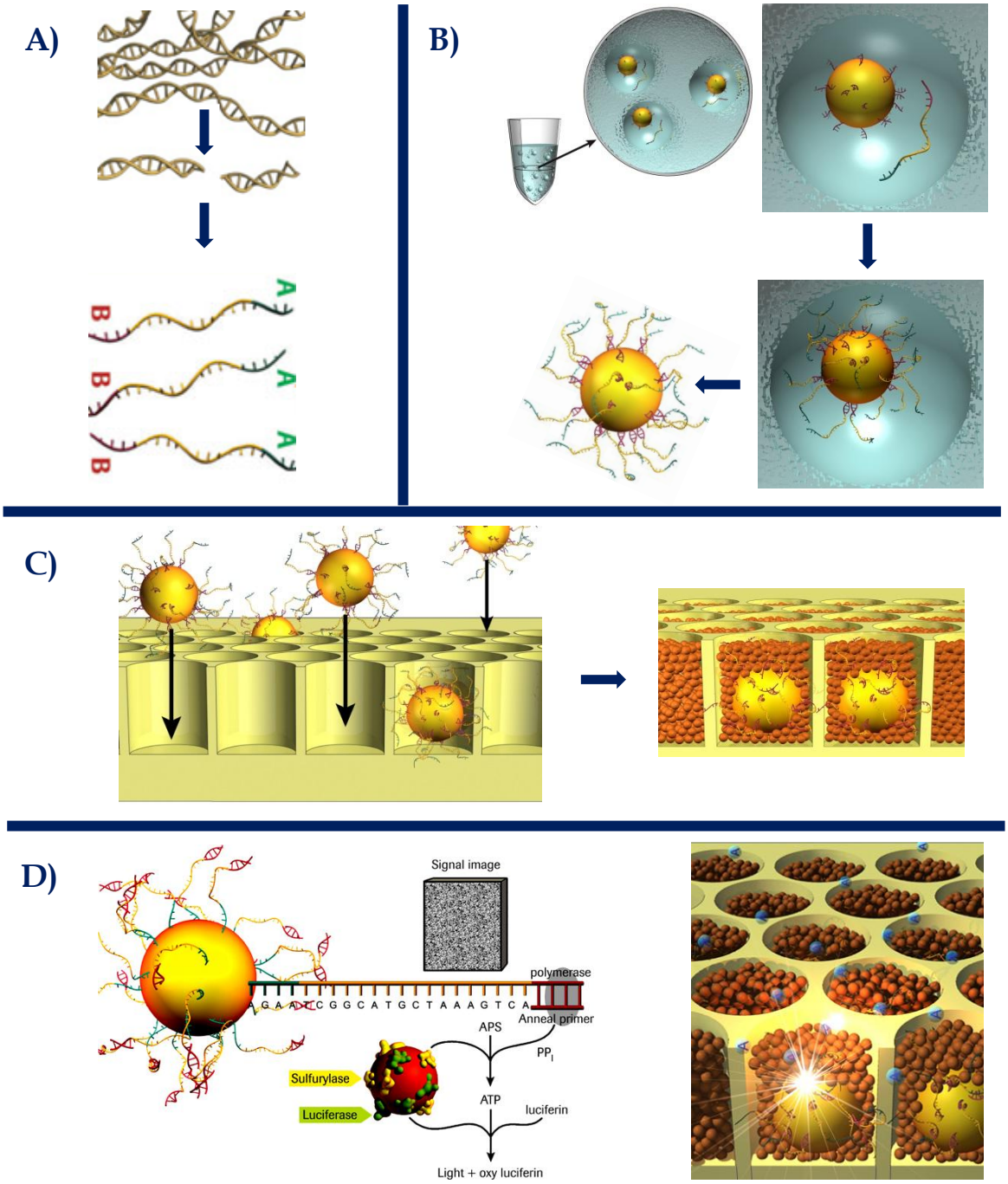


Figure 1-3 Workflow of 454 sequencing. **A)** DNA Library construction by fractionation using nebulization, and subsequently blunt-ending and adaptor ligation. **B)** One of the adaptors contains a 5' biotin tag, enabling binding onto streptavidin coated beads where clonal amplification of the DNA fragments is carried out by emulsion PCR. Beads are separated by a w/o emulsion enabling amplification in oil droplets containing a PCR reaction mixture. The emulsion is subsequently broken followed by enrichment of template-carrying beads. **C)** Enriched beads are loaded in a picotiter plate containing 28 µm diameter wells only allowing one bead per well. Smaller beads containing immobilized sequencing enzymes (ATP sulfurylase, luciferase, and apyrase) are subsequently added. **D)** At each sequencing cycle, single species nucleotides are flown across the plate. Pyrophosphates are released in wells where nucleotide addition occurs, enabling oxidization of luceferin by the action of ATP sulfurylase and luciferase, hereby omitting light, which is detected by a CCD sensor. The sequencing cycle is finalized by degradation of unincorporated nucleotides by the action of apyrase. See text for further details. Images reprinted with permission from 454 Sequencing© Roche Diagnostics.

These results are similar to those of a study by Huse *et al.* performed using the first version of the 454 sequencing platform, the GS20 (Huse *et al.*, 2007). One advantages of the 454 sequencing system is the read length, which according to Roche is up to 1 kb on the newest platform GS FLX Titanium XL+ with the newest chemicals (Roche Diagnostics, 2010), and the study by Gilles *et al.* only showed an increase in the overall error rate to 1.07 % of ~ 550 bp reads (Gilles *et al.*, 2011).

1.1.2.2 Illumina Sequencing

The Illumina sequencing technology is based on the ideas of Shankar Balasubramanian, and David Klenerman from the 1990s (Illumina, 2011a), and work done by Turcatti *et al.* (Turcatti *et al.*, 2008; Fedurco *et al.*, 2006). Illumina sequencing is also a "sequencing by synthesis" method, but opposite to 454 sequencing it utilizes reversible termination chemistry of nucleotide analogues (Bentley *et al.*, 2008). The company Solexa was founded in 1998. After 8 years of research and development and merger with the molecular clustering technology company Manteia in 2004 and the instrumentation company Lynx Therapeutics in 2005, Solexa could launch their first DNA-sequencer, the Genome Analyzer, to the market in 2006. A year later Solexa was acquired by Illumina.

An overview of the workflow of Illumina sequencing is given in Figure 1-4. Like 454 sequencing, library construction is accomplished by DNA fractionation using nebulization, and subsequently enzymatically blunt-ending and adaptor ligation. However, DNA amplification is performed on the glass surface of a flowcell using solid-phase bridge PCR (Fedurco *et al.*, 2006; Adessi *et al.*, 2000). Here the adaptor flanked DNA fragments are bound to an oligonucleotide covered surface. Altering cycles of *bst* polymerase extension and denaturation using formamide creates copies of the template DNA fragment, and the immobilization ensures that all amplicons originating from the single molecule template are clustered together on the surface. Each cluster consists of ~ clonal 1,000 copies of the template. Using the latest version of Illumina sequencing systems, the Hiseq 2000, it is possible to amplify ~ 25 million amplicons at distinguishable locations in each of the 8 lanes of the flowcell, which in the end results in a throughput up to 55 gigabases (Gb) per day. The 8 lanes enable parallel sequencing of eight independent libraries, cf. Figure 1-4. After cluster generation, the amplicons are single stranded, and a sequencing primer is hybridized to one of the adaptors flanking the DNA fragment of interest. At each cycle a single base is incorporated with chemically modified nucleotides by a modified DNA polymerase (Bentley *et al.*, 2008). A 3'-*O*-azidomethyl blocking group ensures that only one base is incorporated, and one of four fluorescent labels enables detection of the different DNA bases (Bentley *et al.*, 2008; Turcatti *et al.*, 2008). After acquisition of four images at different channels, the sequencing cycle ends with chemical cleavage of the fluorophore and the blocking group, enabling base incorporation at the next sequencing cycle. After subsequent image analysis and base calling, and filtering of poor quality reads, the end output is sequence files in Illumina's FASTQ format. Due to the use of modified polymerase and nucleotides, the most frequent sequence error for Illumina sequencing has been reported to be substitution (Dohm *et al.*, 2008; Hutchison, 2007). The Read-length of Illumina is limited by factors, such as incomplete cleavage of the fluorophore or blocking group causing signal decay and dephasing (Whiteford *et al.*, 2009; Shendure & Ji, 2008).

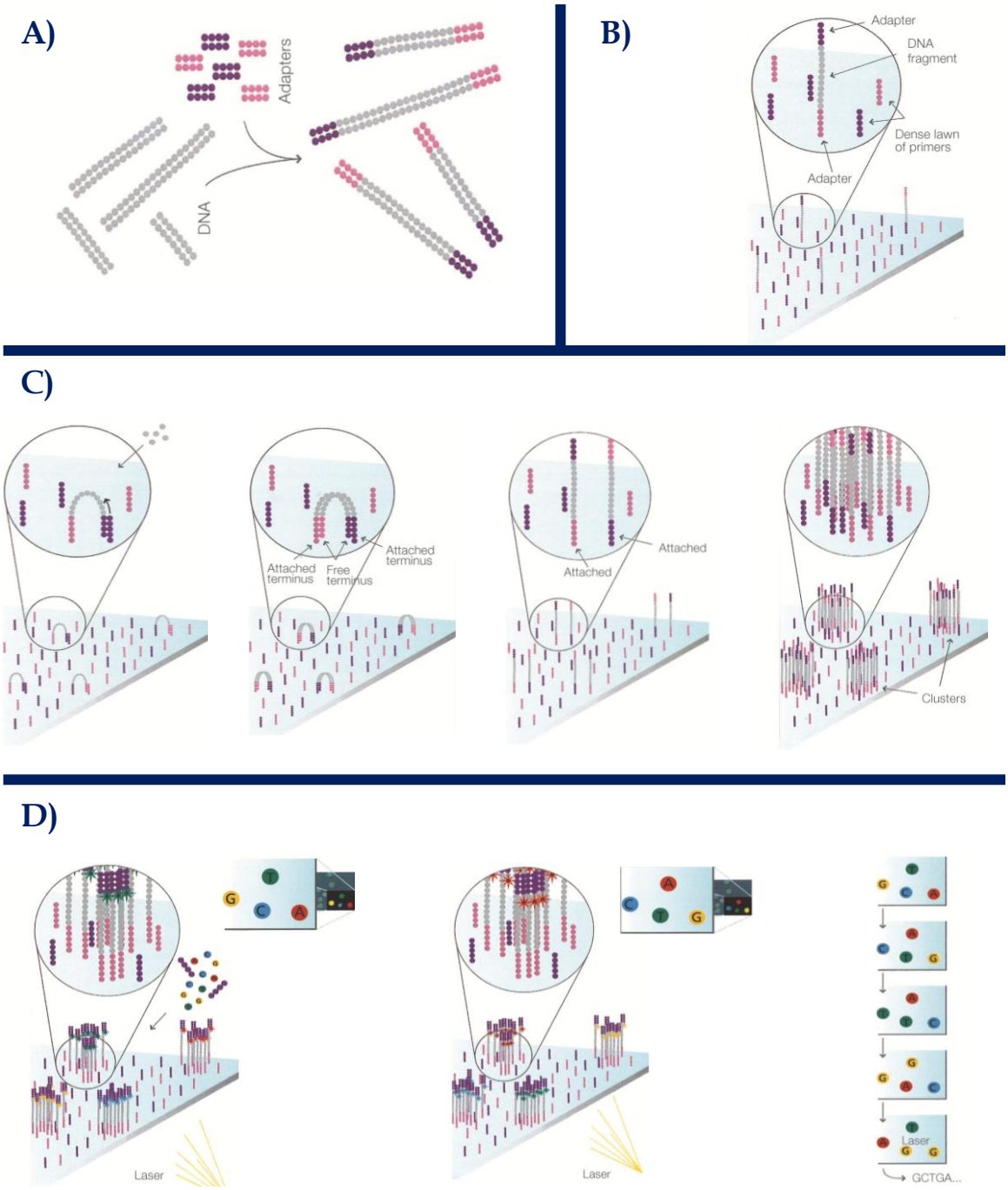


Figure 1-4 Workflow of Illumina sequencing. Like 454 sequencing. **A)** DNA Library construction by DNA fractionation using nebulization, and subsequently blunt-ending and adaptor ligation. **B)** Adaptor flanked DNA fragments are bound to an oligo covered surface. **C)** Altering cycles of *bst* polymerase extension and denaturation with formamide creates clusters consisting of ~ clonal 1,000 copies of the template. **D)** Sequencing is initiated by hybridization of a sequencing primer. At each cycle a single base is incorporated with chemically modified nucleotides by a modified DNA polymerase, and one of four fluorescent labels enables detection of the different DNA. After acquisition of four images at different channels, the sequencing cycle ends with chemical cleavage of the fluorophore and the blocking group, enabling base incorporation at the next sequencing cycle. Subsequently, images are converted into sequence files by image analysis, base calling, and sequence filtering. See text for further details. Images reprinted with permission from Illumina, Inc.

Illumina's first sequencer, the Genome Analyzer Classic, could only sequence up to 36 base pairs. However, today Illumina sequencing is able to produce up to 150 bp reads (Illumina, 2011a).

1.1.2.3 SOLiD™ Sequencing

The technology behind the SOLiD™ (Sequencing by Oligo Ligation and Detection) platform was first described in 2005 by Shendure *et al.* (Shendure *et al.*, 2005). The first commercially available sequencer was released in October 2007 (Mardis, 2008).

An overview of the workflow of SOLiD sequencing is given in Figure 1-5. Library construction is similar to that of both 454 and Illumina sequencing, and may be constructed in several ways to produce, adaptor-flanked fragments (Shendure & Ji, 2008). The SOLiD technology also requires DNA amplification, and like 454 sequencing this is performed using emulsion PCR (Dressman *et al.*, 2003) where DNA fragments are bound to paramagnetic beads. Prior to sequencing, the emulsion is broken, and amplicons bearing beads enriched and immobilized to the surface of a specially treated glass slide (Mardis, 2008), generating a dense, disordered array. The SOLiD technology is also a "sequencing by synthesis" method, but is unique due to several reasons. Firstly, after annealing of a sequencing primer, the synthesis of DNA is not driven by a DNA-polymerase, but a ligase (Shendure *et al.*, 2005; Housby & Southern, 1998). At each step, matching fluorescently labeled octamer originating from a degenerate set is ligated to the DNA fragment. Fluorophores on the different octamers are correlated to a specific position within the octamer. After image acquisition in four different channels, chemical cleavage of the octamer between the fifth and sixth base is performed removing the fluorophore. Multiple ligation steps enable sequencing of every fifth base of the DNA fragment. Following several rounds of ligation, image acquisition and cleavage, the DNA is denatured, enabling annealing of a new sequencing primer at a different position on the adaptor sequence, and a new set of ligation steps (Shendure *et al.*, 2005). A second unique feature is that the fluorophores are correlated to dinucleotides, and not just a single base. This combined with an alternate use of sequencing primers and octamer sets, where the fluorophores correspond to different positions on the octamer, enables that each base is sequenced twice, and miscalled bases can be corrected (Heinz, 2010; McKernan *et al.*, 2008). According to the Life Technologies, this enables sequencing with an accuracy of 99.9999% or above for the majority of base calls on their latest sequencer, the 5500 Series SOLiD™ System (Life Technologies, 2011). After subsequent analysis, the end output is sequence files in SOLiDs color space format csfasta (Life Technologies, 2008). However, these can be converted into regular sequence files in FASTA format using dynamic programming (Li & Durbin, 2009).

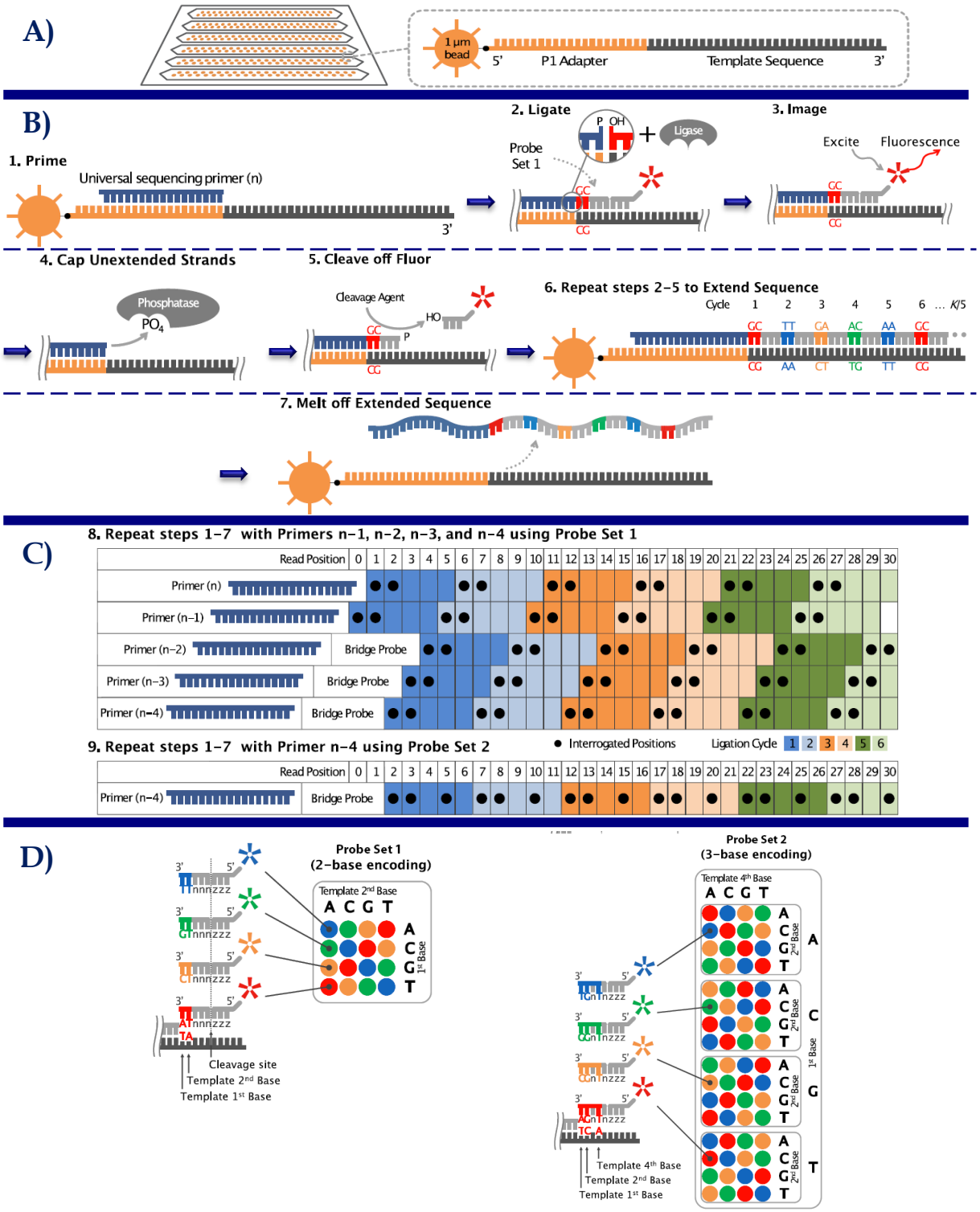


Figure 1-5 Workflow of SOLiD sequencing. Library construction and DNA amplification is similar to 454 sequencing, cf. section 1.1.2.1. **A)** Amplicons bearing beads are enriched and immobilized to the surface of a specially treated glass slide. **B)** Ligase drive DNA synthesis is performed after annealing of a sequencing primer (1). At each step, matching fluorescently labeled octamer originating from a degenerate set is ligated to the DNA fragment (2). After image acquisition in four different channels (3), un-extended strands are capped (4) the fluorophore is removed by chemical cleavage (5). Multiple ligation steps enable sequencing of every fifth base of the DNA fragment (6). Finally, extended strands are melted off. **C)** Multiple sequence rounds using different sequence primers and probe set enables double determination of every base. **D)** The Exact Call Chemistry completing the 2-base encoding ensures high raw sequence accuracy. See text for more details. Images reprinted with permission from Life Technologies Corp.

1.1.2.4 Paired-end Sequencing

All three platforms described above are more or less limited by short read lengths, cf. sections 1.1.2.1 and 1.1.2.2. However, this limitation has been partly overcome by the development of paired-end sequencing, which can be performed using all three sequencing systems. Paired-end tags (PETs) are shorter sequences originating from the two ends of a target DNA (Fullwood *et al.*, 2009). Paired-end sequencing was already described in 1981 by Hong (Hong, 1981), and the first use of paired-end sequencing was described by Ewards and Caskey in 1990 (Edwards & Caskey, 1991). There are multiple ways of constructing a paired-end library. One is the clone based method, where the target sequence is ligated with adaptors containing MmeI restriction sites immediately next to the target sequence. Following amplification in *E. coli,* purification and MmeI digestion, the tag containing vector is recircularized, hereby joining the two sequence tags. After subsequent amplification in *E. coli*, the PET constructs can be purified using restriction digestion (Ng, Wei & Ruan, 2007). A second method was introduced by Shendure *et al.* concurrently with the introduction of the technology behind SOLiD sequencing (Shendure *et al.*, 2005). Here the target DNA fragments are directly circularized with linker oligonucleotides hereby joining the two ends of the target DNA. The linker sequence contains two restriction sites (e.g. MmeI) flanking the two ends of the target DNA, enabling restriction digestion to release the tag-linker-tag construct for sequencing (Shendure *et al.*, 2005). These two methods can create libraries with long inserts (up to 20 kb) between the two sequence tags (454 sequencing, 2011), which are often referred to as mate pair libraries (Fullwood *et al.*, 2009). Additional to these methods, short insert libraries (200-500 bp) can also be paired-end sequenced using Illumina sequencing. Here paired-end libraries are made using adaptors with two different sequencing primers. Paired-end is performed by first sequencing the target DNA utilizing the first sequencing primer. After subsequent product denaturation, bridging, and second strand synthesis, the opposite strand is cleaved providing a template for a round second sequencing utilizing the second sequencing primer (Illumina, 2011b; Bentley *et al.*, 2008).

1.1.2.5 Comparison of Next Generation Sequencing Platforms

The three described sequencing technologies have advantages and limitation in regards to terms such as cost, throughput, read length, and practical aspects. These are all compared in Table 1-1.

Table 1-1 Comparison of next generation sequencing platforms. PE = Paired-end, MP = mate pair. *Using the Nextera library preparation, the time consumption is shortened by 1 day. [1]Information as provided by company. [2]Information based on review by Pareek *et al.* (Pareek, Smoczynski & Tretyn, 2011).

Company	Roche Diagnostics	Illumina, Inc.	Life Technologies
Platform			
Sequencing system	GS FLX Titanium+	HiSeq 2000	SOLiD 5500xl
Estimated system cost[2]	$ 500,000	$400,000	525,000
Cost per Mb[2]	$ 84.39	$5.97	$5.81
Advantages	Longest read lengths among NGS platforms	Very high throughput	high throughput and accuracy
Limitations	Challenging sample prep. Problematic base determination in homopolymeric regions. Sequential reagent washing causes error accumulation.	Signal decay and dephasing limits read length and causes lower accuracy at the end of reads	Challenging sample prep.
Library & template preparation			
Sample requirements	1 µg for shotgun libraries, 5 µg for PE libraries	1 µg for single or paired-end libraries	<2 µg for shotgun libraries, 5–20 µg for PE libraries
Amplification method	Emulsion PCR	Bridge amplification	Emulsion PCR
Time of library prep. & amplification	3-4 days*	2 days*	2-4.5 days
PE insert size	3,8, and 20 kb	200-500 bp (PE) 2-5 kb (MP)	600bp-6kb
Sequencing			
Sequencing technology	Pyrosequencing	Reversible Dye Terminators	Oligonucleotide Probe Ligation
Detection Method	Light emission from secondary reactions initiated by pyrophosphate release	Fluorescent emission from incorporated dye-labeled nucleotides	Fluorescent emission from ligated dye-labeled oligonucleotides
Run time	10 hours	2-11 days	1-7 days
Maximum libraries without multiplexing	16 gaskets	16 (2 flowcells)	12 (2 flowchips)
Multiplexing barcode number	132	12	96
Maximum samples with multiplexing	2112	192	1152
Sequencing statistics[1]			
Read length	700-1000 bp single (100 x 100) bp MP	100 bp single (100 x 100) bp PE (36 x 36) bp MP	75 bp single (75 x 35) bp PE (60 x 60) bp MP
Raw accuracy	99.997%	99.5% at 100bp	Up to 99.99%
Throughput per day	~730 MB	Up to 55 Gb (for a 2 x 100 bp run)	~10-15 Gb

1.1.2.6 Other Next Generation Sequencing Platforms

Although the commercial market of DNA sequencing is dominated by the three platforms described above (Herper, 2010), other NGS technologies have been and are being developed, and some have already been commercialized. In the following, two of these, the Ion Torrent System by Life Technologies and the Heliscope by Helicos BioScience will be described. For a more complete list of sequencing technologies see review by Zhang *et al.* (Zhang *et al.*, 2011).

Ion Torrent (Life Technologies)

In February 2010, Ion Torrent introduced a sequencer with a novel detection system not based not on light emission and optics but ion detection (Rusk, 2011). The technology builds on the work by Rothberg *et al.* and utilizes the fact that nucleotide incorporation by DNA-polymerase results in hydrolysis of the nucleotide triphosphate, causing release of a single proton. This produces a shift in pH that scales with the number of nucleotides incorporated. (Rothberg *et al.*, 2011)

Library preparation and the sequencing scheme are similar to that of 454 sequencing. Library preparation is performed using clonal amplification on beads. Sequencing is performed in a dense microwell array containing ion sensitive field-effect transistors enabling real time measurement of the change in pH, which is converted into a voltage. The four nucleotides are sequentially flowed over the array, producing a voltage change where incorporation of one or more nucleotides occurs (Rothberg *et al.*, 2011). Rothberg *et al.* reported that the technology can produce 100 bp reads with a raw accuracy of 98.90 % (Rothberg *et al.*, 2011), which is similar to other NGS platforms (Bentley *et al.*, 2008; Margulies *et al.*, 2005; Shendure *et al.*, 2005). However, as for 454 sequencing base calling in homopolymeric regions is problematic (Rothberg *et al.*, 2011).

In the six months the technology has been commercially available, it has improved dramatically. In July 2011 Ion Torrent announced the release of the Ion 316™ chip, which can produce 100 Mega bases (Mb) in a 2 hour run. This is a tenfold increase in throughput than the original Ion 314™ chip, and Ion Torrent stats that the Ion 318™ chip coming out later in 2011 will be able to produce 1 Giga base (Gb), another tenfold increase. These promising improvements are partially achieved by the development of better field-effect transistors; a development following Moore's law (Moore, 1965), the author of which, Gordon Moore, had his genome sequenced. For a further technology overview see work by Rothberg *et al.* (Rothberg *et al.*, 2011) and (Ion Torrent Systems, 2011).

Heliscope™ Single Molecule Sequencer

The major difference between Heliscope™ Single Molecule Sequencer and other NGS platforms is that it utilizes sequencing of single DNA molecules. The technology was introduced by Braslavsky *et al.* in 2003 (Braslavsky *et al.*, 2003), and later licensed by Helicos BioScience, which could introduce the first sequencing platform based on this technology in 2007. Library construction is done by DNA shearing and subsequent polyadenylation (Ozsolak *et al.*, 2010). Sequencing is performed by hybridizing the DNA fragments to covalently bound PolyT-oligonucleotides on a flow cell. Sequencing is performed in a similar fashion as Illumina sequencing, cf. section 1.1.2.2, with single nucleotide extension followed by detection and cleavage of fluorophores (Harris *et al.*, 2008). According to Helicos BioScience, Heliscope™ Single Molecule Sequencer is comparable with other current NGS sequencing systems with regards to accuracy. However, the system is currently limited by shorter read length (25- 55 bp), and lower throughput compared to the Illumina and SOLiD sequencing systems (Helicos BioSciences, 2008). A unique feature of this system is direct RNA sequencing (Ozsolak & Milos, 2011a; Ozsolak & Milos, 2011b). For a further technology overview see work by Harris *et al.* (Harris *et al.*, 2008), Pushkarev *et al.* (Pushkarev, Neff & Quake, 2009), Ozsolak *et al.* (Ozsolak & Milos, 2011a; Ozsolak & Milos, 2011b) and (Helicos BioSciences Corporation, 2008).

1.1.3 Third Generation Technologies – the Future

The next (or second) generation sequencing technologies have had and still have a huge impact on life Sciences, cf. section 1.2. Inspired by this; in 2004 the National Human Genome Research Institute launched research programs to further accelerate the development of sequencing technologies, hereby lowering the cost for sequencing a genome to less than $100,000, a goal already reached in 2009 (National Human Genome Research Institute, 2011; Wetterstrand, 2011). The ultimate goal is the "$1,000 genome" (National Human Genome Research Institute, 2011), a goal not that far away, since the price in April 2011 was less than $17,000 (Wetterstrand, 2011). This search for further innovation to enable routine sequencing of genomes has triggered research of novel DNA sequencing technologies. There are several definitions of "third generation" or "next-next generation" sequencing technologies (Niedringhaus *et al.*, 2011; Pareek, Smoczynski & Tretyn, 2011; Schadt, Turner & Kasarskis, 2010), but the one used by Schadt *et al.* states that third generation sequencing technologies are able to perform single molecule sequencing without pausing between read steps (Schadt, Turner & Kasarskis, 2010) clearly separates second and third generation sequencing technologies. Especially, the need for signal enhancement to enable reliable base detection i.e. DNA amplification is a limiting factor for NGS technologies. Amplification of DNA can introduce sequence errors and can also change the relative abundance of different DNA fragments (Pareek, Smoczynski & Tretyn, 2011).

PacBio RS System (Pacific Biosciences)

In April 2011, the first third generation sequencing system, the PacBio RS System from Pacific Biosciences became available (Pacific Biosciences, 2011). The technology is based on direct observation of a single molecule of DNA polymerase using zero-mode waveguide technology (Eid *et al.*, 2009; Levene *et al.*, 2003). The PacBio RS System was recently used by Chin *et al.* in the investigation of the source of a cholera outbreak on Haiti. Here average read lengths between 700 and 1,000 bp with raw accuracies ranging from 81 % to 83 % was reported (Chin *et al.*, 2011). While certainly an attractive read length for a DNA sequencing system, the raw error rate is inferior to existing sequencing technologies and limits its usefulness.

Nanopore Sequencing Technologies

Research is also conducted in an entirely different sequencing technology based on nanopore structures. The concepts and potentials of nanopore sequencing were reviewed by Branton *et al.* in 2008 (Branton *et al.*, 2008). The idea behind nanopore sequencing is that base detection should be performed by conductivity measurements either across or through a nanoscale pore. In theory, the chemical differences of each base would result in an altered current flow through the pore, which could be detected and used for base determination (Niedringhaus *et al.*, 2011). Although still in development, the nanopore approach could potentially become the "fourth-generation" sequencing technology.

1.2 Applications and Challenges for Analysis of Next Generation Sequencing Data

1.2.1 "–Omics" Based Research Fields

In cellular and molecular biology nouns ending with –ome have the sense "all of the specified constituents of a cell, considered collectively or in total" (Dictionary Oxford English, 2010), and hence the -omics suffix signifies the measurement of the entire collection of biological molecules or information. "-Omics" spans over a wide range of research fields, which are constantly increasing. However, the four major fields are studies of collection of the three molecule types of the central dogma of molecular biology (Crick, 1970) and the product of the action of these namely (Schneider & Orchard, 2011) :

- Genomics - the quantitative study of DNA and genomes, with elements such as protein coding genes and regulatory elements.
- Transcriptomics - the quantitative study of transcribed RNA, such as messenger RNA (mRNA) and micro RNA (miRNA).
- Proteomics - the quantitative study of protein abundance.
- Metabolomics - the quantitative study of metabolites.

These four fields have given rise to numerous related research fields. A few examples are physiomics - the study of the functional behavior of the physiological state of an individual or species (Bassingthwaighte, 2007), pharmacogenomics - the study of how variations in the human genome affect the response to medicines (Martin & Martin, 2008), and nutrigenomics - the study of how different foods can interact with genes to increase the risk of common chronic diseases (MedicineNet.com, 2003).

Since "omics" studies requires simultaneous measurement of thousands or even millions of variables (genes, mRNAs, proteins etc. depending on the field), the measurement techniques for "omics" studies need to be high-throughput. The developments of NGS technologies have therefore been a major factor in advances in genomics and transcriptomics, which e.g. have enhanced our understanding of the pathogenesis of many human diseases (Cappola & Margulies, 2011; Daly, 2010; Novelli *et al.*, 2010; Hardy & Singleton, 2009), and have made personalized medicine to become a possibility in the future (Lunshof *et al.*, 2010). However, with these advances and new possibilities that have emerged, bioinformatic challenges have followed, and the benefits of the NGS technologies cannot be utilizes before a bioinformatic framework for data handling and analysis exists (Zhang *et al.*, 2011; Pop & Salzberg, 2008). In the following sections central aspects of genomics and transcriptomics and the associated bioinformatic challenges will be described.

1.2.2 Analysis of Next Generation Sequencing Genomics Data

A key step in data analysis of a variety of applications utilizing NGS data is often initial alignment or mapping of reads to a reference or assembly of the short reads into larger continuous sequences (e.g. a genome sequence or a collection of mRNA transcripts) (Flicek & Birney, 2009). Examples of applications such applications are: genome re-sequencing and subsequent identification of variations (Levy *et al.*, 2007), identification of protein binding sites on the DNA

combining chromatin immunoprecipitation with NGS (Chip-seq) (Johnson *et al.*, 2007), gene expression profiling (RNAseq) (Mortazavi *et al.*, 2008), identification of the genome-wide methylation pattern (epigenomics) (Cokus *et al.*, 2008; Callinan & Feinberg, 2006). In the following sections, some of the bioinformatic challenges, developments, and current solutions for alignment and assembly of NGS reads will be described.

1.2.2.1 Genome Alignment of NGS Data

The development of NGS technologies caused a need for new alignment algorithms to be developed for several reasons. NGS data characteristics differ from Sanger sequencing data in several aspects; the most obvious being a shorter read length (especially for Illumina and SOLiD data), but also elements such as error rates and types (e.g. insertion/deletion errors of 454 data) is different, cf. section 1.1.2. Earlier generation and widely used alignment programs such as Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990) were primarily designed for protein or DNA sequence alignment and search through a database in order to find homologue sequences. Genome alignment programs differ, since they should be optimized for determination of the most likely source for a sequencing read within a genome sequence. This difference causes an assumption change regarding the number of expected mismatches (Flicek & Birney, 2009). Moreover, process speed is a crucial factor due to the massive and ever growing volume of NGS data.

Most of today's short read alignment programs utilize a multistep procedure. The first step identifies a small subset of likely matching places on the reference genome for the read utilizing heuristic methods. In a second step a more accurate alignment algorithm such as the Smith-Waterman algorithm (Smith & Waterman, 1981) is used on the subset to identify the most likely place for the read on the genome (Flicek & Birney, 2009). To facilitate fast searching in the first step of the algorithm, most first short read alignment programs designed for NGS data create so-called indices for the reads, the genome sequence, or both (Horner *et al.*, 2010; Li & Homer, 2010). Some of the first short read alignment programs such as MAQ (Li, Ruan & Durbin, 2008) and short oligonucleotide alignment program (SOAP) (Li *et al.*, 2008) utilize hash-based methods for the first step (Li & Homer, 2010; Flicek & Birney, 2009). A hash table is a data structure enabling indexing and facilitating fast searching of non-sequential data such as DNA sequences (Flicek & Birney, 2009). Further development of short read alignment programs has resulted in the implementation of algorithms based on a data structure called the FM-index (Ferragina & Manzini, 2000) and the utilization of Burrow-Wheeler transformation (BWT) (Burrows & Wheeler, 1994). These have the advantage that alignment to identical copies of a substring in the reference is only needed to be done once, whereas with a hash table index, an alignment must be performed for each copy (Li & Homer, 2010). This is reflected in the speed of popular BWT based programs such as Bowtie (Langmead *et al.*, 2009), Burrows-Wheeler Alignment tool (BWA) (Li & Durbin, 2009), SOAP2 (Li *et al.*, 2009), and the mapping tool of the CLC Genomics Workbench (CLC Bio, 2010) that all are between 10 and 30 times faster than hash based programs (Flicek & Birney, 2009).

To improve the quality of alignment, most of today's alignment software take additional information into account during sequence alignment. Paired-end sequence information enables the possibility of alignment of an ambiguous read, if the mate aligns unambiguously. This effect causes paired-end alignment to outperform single read alignment, both in terms of sensitivity and specificity (Li & Homer, 2010). Base quality information has also been shown to

improve alignment accuracy (Flicek & Birney, 2009), due to the possibility of lower penalization of low quality bases (Smith, Xuan & Zhang, 2008). Moreover, the color space information of SOLiD, which most alignment algorithms are able to handle can also improve alignment accuracy because each position is interrogated twice. Sequence alignment can be performed entirely in color space, but according to results by Li and Homer a better solution is to perform color-aware Smith-Waterman alignment in the second part of the alignment process (Li & Homer, 2010).

1.2.2.2 *De novo* Genome Assembly of NGS Data

When a reference sequence is not available, it is necessary to reconstruct the entire genome from the sequenced reads. All NGS technologies produce reads that most often are several magnitudes shorter than the target (genome) sequence. The challenge of assembling these short fragments into longer contigs is overcome by oversampling of random fragments of the target sequence, and subsequent *de novo* assembly of the reads (Flicek & Birney, 2009). The concept and challenges of *de novo* assembly of genomes was nicely described by Pevzner *et al.* in 2001 stating that: "*Children like puzzles, and they usually assemble them by trying all possible pairs of pieces and putting together pieces that match. Biologists assemble genomes in a surprisingly similar way; the major difference being that the number of pieces is larger*" (Pevzner, Tang & Waterman, 2001). This metaphor to a jigsaw puzzle highlights a major challenge of *de novo* assembly, namely the number of pieces (reads) to assemble. A second complication to solving the puzzle is the presence of near identical repeat regions in the genome (Pop, 2009), which could give rise to the addition of "*and all the pieces are blue sky*" to Pevsner's statement. Assembly of perfect repeat regions is only possible if reads spanning the regions exist, while assembly of inexact repeat regions is possible by high-stringency alignments. However, this is highly complicated by the presence of sequence errors. (Miller, Koren & Sutton, 2010)

Assembly programs designed for Sanger data such as Arachne (Batzoglou *et al.*, 2002), and Celera Assembler (Myers *et al.*, 2000) utilizes an overlap / layout /consensus approach for assembly. Here, heuristic all-against-all pairwise comparisons are performed followed by construction of an overlap graph and finally multiple sequence alignment to determine the precise layout and consensus sequence (Miller, Koren & Sutton, 2010). This approach has been adapted in the Newbler (Margulies *et al.*, 2005) software distributed with the 454 sequencing platform, and by the short read assembler Edena (Hernandez *et al.*, 2008). A different approach is used by some of the first assembly programs for short read NGS data such as SSAKE (Warren *et al.*, 2007), VCAKE (Jeck *et al.*, 2007), and SHARCGS (Dohm *et al.*, 2007). They all utilize a variant of a greedy algorithm, where reads are chosen to form seeds for contig formation. These seeds are continuously extended by identification of overlapping reads while the extension is unambiguous. These programs all produces relatively short contigs (a few kilobases) (Dohm *et al.*, 2007; Jeck *et al.*, 2007; Warren *et al.*, 2007) terminated at ambiguous regions caused by ubiquitous repeat regions (Flicek & Birney, 2009). This challenge was largely overcome by the employment of de Bruijn graph-based approaches, which is today's most widely used approach and is also named an Eulerian approach (Miller, Koren & Sutton, 2010; Flicek & Birney, 2009). It utilizes a de Bruijn graph, cf. Figure 1-6.

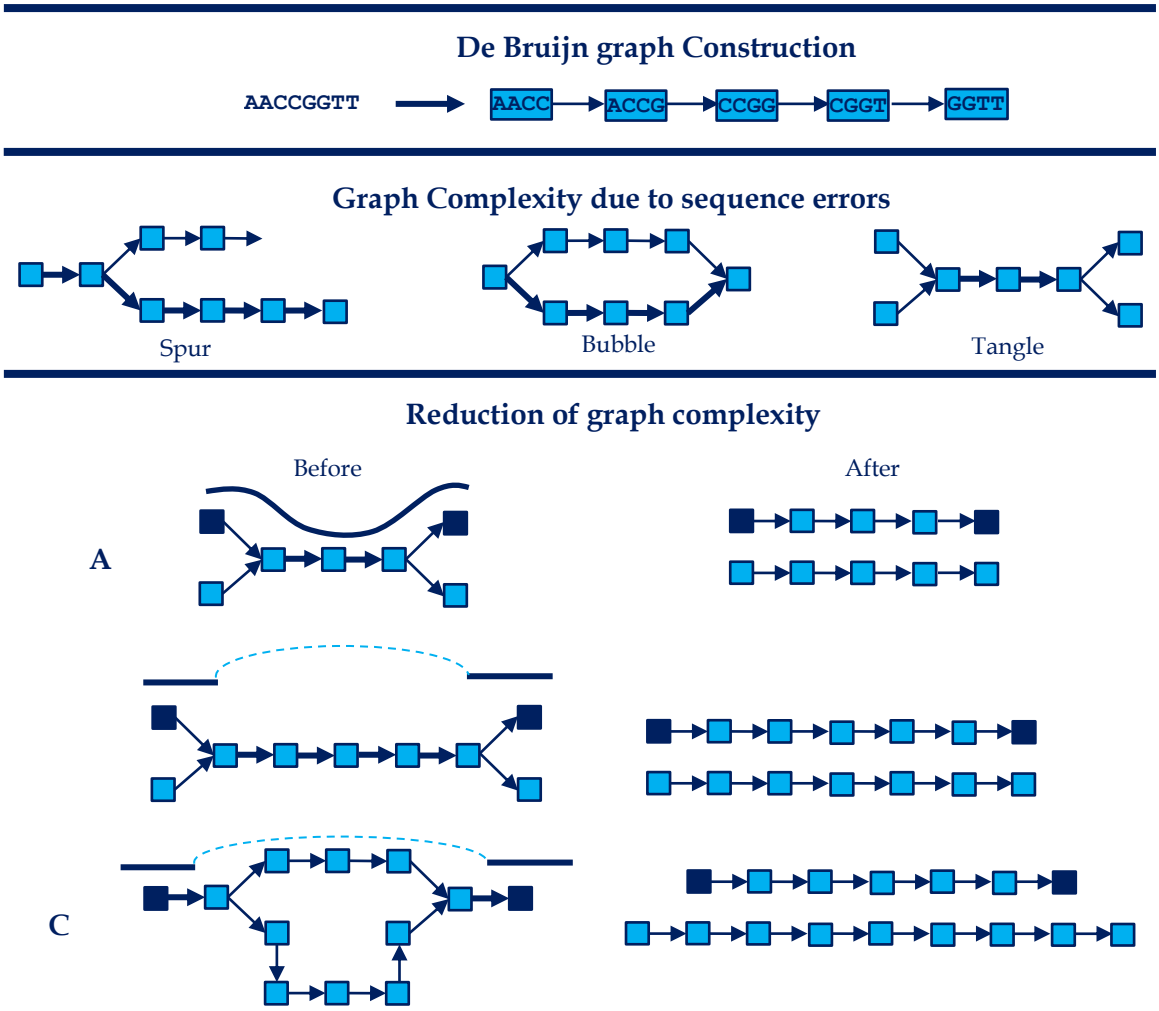


Figure 1-6 Sequence assembly using de Bruijn graphs. **Top:** A Read represented by a K-mer graphs (K=4), where the graph has a node for every K-mer in the read and an edge for every pair of K-mers that overlap by K-1 bases in the read. The path is simple to construct because the K-mers are larger than the repeats (2 bp) in the read. Real sequence data uses longer K-mers. **Middle:** Sequence errors and repetitive elements cause complexity in the de Bruijn graph. Spurs are caused by sequence errors toward the end of a read, Bubbles are caused by sequence errors in the middle and repeat sequences can lead to tangles. **Bottom:** Graph complexity can be solved e.g. when (A) a read threading joins paths across collapsed repeats that are shorter than the read lengths, (b) when mate threading joins paths across collapsed repeats that are shorter than the paired-end distance of a read, or (c) when path following chooses one path if its length fits the paired-end constraint. Non-branching path shown in the figure could be simplified to single edges or nodes (graph simplification). Edges represented in more reads are drawn with thicker arrows. Reads are shown as lines and the paired-end distance is shown in dashed lines. Figure revised from (Miller, Koren & Sutton, 2010).

A de Bruijn graph is compact representation based on short K-mers (a sequence of K base calls) ideally for high coverage, short read data sets, one of the reasons being that the time to construct the graph scales linearly with the number of reads compared to quadratic time increase as in the case for all-against-all pairwise comparisons (Zerbino & Birney, 2008). The nodes of the graph represent all fixed-length sub-sequences from a larger sequence, and the edges represent all fixed-length overlaps between consecutive sub-sequences in the larger sequence, cf. Figure 1-6. Given error-free data with K-mers covering the entire genome and spanning all repeat regions, the graph contains a path that traverses each edge exactly once, representing the assembled genome. This perfect read path would be easy to find, however the graphs from real sequencing data are more complex due to the presence of sequence errors and repetitive elements, cf. Figure 1-6. EULER (Chaisson, Brinza & Pevzner, 2009; Chaisson & Pevzner, 2008;

Chaisson, Pevzner & Tang, 2004) Velvet (Zerbino & Birney, 2008), Allpaths (Butler *et al.*, 2008), AbySS (Simpson *et al.*, 2009), and SOAPdenovo (Li *et al.*, 2010c) are the most widely used freely available de Bruijn graph software programs today. Although there are differences in the algorithms, they all share a common set of features:

- Error detection and correction based on sequence composition of the reads, e.g. K-mer frequency (Chaisson, Brinza & Pevzner, 2009; Zerbino & Birney, 2008).
- Graph construction to represent reads and their shared sequence, either representation of K-mers as graph nodes (Li *et al.*, 2010c; Chaisson, Brinza & Pevzner, 2009; Simpson *et al.*, 2009; Zerbino & Birney, 2008), or representation of simple paths as graph nodes (Butler *et al.*, 2008).
- Graph simplification by reduction of simple paths into single nodes (Li *et al.*, 2010c; Chaisson, Brinza & Pevzner, 2009; Zerbino & Birney, 2008).
- Recognition of spurs, cf. Figure 1-6, caused by sequencing error toward one end of a read, and bubbles, cf. Figure 1-6, caused by sequencing error toward the middle of a read, and by polymorphisms in the genome and subsequent removal of error induced paths (Chaisson, Brinza & Pevzner, 2009; Simpson *et al.*, 2009; Butler *et al.*, 2008; Zerbino & Birney, 2008).
- Simplification of tangles, cf. Figure 1-6, caused by repeats in the genome sequence e.g. by using individual sequence read information or paired-end distances as constraints on path distance (Li *et al.*, 2010c; Chaisson, Brinza & Pevzner, 2009).

The initial output is of all assemblers is contig assemblies. However, EULER Velvet, Allpaths, and SOAPdenovo also use paired-end and mate-pair information to order the contigs and create scaffolds if possible (Miller, Koren & Sutton, 2010). Although the de Bruijn data structure is compressed, the memory overhead can still be a challenge when assembling large eukaryotic genomes. Some of the assemblers are designed to overcome this problem. AbySS distributes the graph and computations across a computer grid (Simpson *et al.*, 2009), Allpaths uses database implementation (Butler *et al.*, 2008), and SOAPdenovo uses a more space-efficient graph structure (Li *et al.*, 2010c).

Lin *et al.* (Lin *et al.*, 2011) have recently compared the performance of SSAKE, VCAKE, Edena, Velvet, AbySS, and SOAPdenovo for different genomes (ranging from 100 kb to 100 Mb) under different conditions. Read lengths, read error rate, sequence coverage, use of paired-end information, and GC content of the genome all affected the quality of the assemblies. No assembler outperformed all others under all conditions with respect to all performance measures. However, AbySS in general had the lowest assembly error rates and SOAPdenovo generated the longest N50 lengths[1]. Moreover, these two assemblers also were the most efficient in regards to runtime and memory usage.

Today high quality *de novo* assemblies of most bacterial sized genomes can be performed using existing algorithms. Furthermore, a few larger and more complex eukaryotic genomes such as the human (Li *et al.*, 2010b), the giant panda (Li *et al.*, 2010a) and the potato genome (The Potato Genome Sequencing Consortium *et al.*, 2011) have been assembled *de novo*. However, assembly of the latter also showed the limitations of current algorithms in regards to haplotype reconstruction of highly heterozygous genomes, i.e. ploidity induced complexity. *De novo* assembly of potato, which is a highly heterozygous genome, was only possible for a doubled monoploid (DM), and not for the diploid genotype (RH) sequenced (The Potato Genome Sequencing Consortium *et al.*, 2011).

1 N50 is defined as the size N such that 50% of the genome is contained in contigs of size N or greater (Zimin *et al.*, 2009).

1.2.3 Analysis of Next Generation Sequencing Transcriptomics Data

The differences of cell type and state in an eukaryotic organism is not encoded directly by the genome sequence, but rather the diverse patterns of gene expression (Pepke, Wold & Mortazavi, 2009), why the study of the transcriptome is essential for understanding of cell development and disease (Wang, Gerstein & Snyder, 2009). In the past, transcriptome analysis has mostly been quantitative (gene expression profiling), but with the development of NGS technologies transcriptomics have gotten multiple functions such as: transcript discovery, determination of the transcriptional structure of genes (start sites, un-translated regions (UTRs), splicing patterns and other post-transcriptional modifications (Pepke, Wold & Mortazavi, 2009; Wang, Gerstein & Snyder, 2009). In the following the developments, today's applications and bioinformatic challenges in transcriptomics will be described.

1.2.3.1 Transcriptomics before Next Generation Sequencing

The field of transcriptomics saw a series of breakthroughs in the 1990's with the development of new analytical techniques such as differential display (Liang & Pardee, 1992), Serial Analysis of Gene Expression (SAGE) (Velculescu *et al.*, 1995), DNA microarray technologies (Lockhart *et al.*, 1996; Schena *et al.*, 1995), cDNA-AFLP (Bachem *et al.*, 1996), random Expressed Sequence Tag (EST) sequencing (Adams *et al.*, 1991), and massively parallel signature sequencing (MPSS) (Brenner *et al.*, 2000). While it is out of the scope of this thesis to describe all these techniques, only two technologies (Ishii *et al.*, 2000) for transcriptome analyses in the past decade namely the most dominating technology microarray and SAGE will be described.

The technology of DNA microarrays, cf. Figure 1-7 panel A, is based on hybridization between a fluorescently labeled target DNA and DNA probes with known sequences that are fixed on spots on a solid base, e.g. a glass slide (Schena *et al.*, 1995). Isolated mRNA transcripts are converted to complementary DNA (cDNA), during which they are fluorescently labeled. Following removal of unbound material, the fluorescent signal intensity from each spot on the array is measured. After pre-processing steps such as background correction, the signal is dependent on the original amount of target sequence (Müller, Neumaier & Hoffmann, 2008). Two-color microarrays (Shalon, Smith & Brown, 1996) are hybridized with cDNA prepared from two samples using two different fluorophores, enabling relative intensity comparison between the samples. Microarrays can contain thousands to millions spots, enabling global relative detection of transcripts.

The SAGE technology produces so-called sequence tags from mRNA transcripts enabling digital measurements of the mRNA transcript abundance (Velculescu *et al.*, 1995), cf. Figure 1-7 panel B. Following cDNA conversion and immobilization to streptavidin beads, sequences are digested with the restriction enzyme NlaIII (called the anchoring enzyme), which recognizes 5′-CATG-3′. Following ligation of a linker containing a recognition site of the Type IIS restriction endonuclease BsmFI, the fragment is cleaved 15 bp away in the 3′ direction from the recognition site releasing the sequence the tag. After removal of the linker fragment, tags are concatenated, cloned into a plasmid vector, and sequenced using Sanger sequencing. SAGE libraries usually contained between 10 and 100 thousand tags (Matsumura *et al.*, 2005). SAGE data analysis includes tag annotation, i.e. mapping the tags back to the mRNA transcripts they originated from, which is complicated by the short nature of the SAGE tags (Müller, Neumaier &

Hoffmann, 2008). The original method described by Velculescu *et al.* in 1995 (Velculescu *et al.*, 1995), producing 14-15 base pair tags, were later improved by Saha *et al.* to 18 base pair tags (long-SAGE) (Saha *et al.*, 2002) and Matsumura *et al.* to 26 base pairs (Matsumura *et al.*, 2005), easing the process of tag annotation (Müller, Neumaier & Hoffmann, 2008).

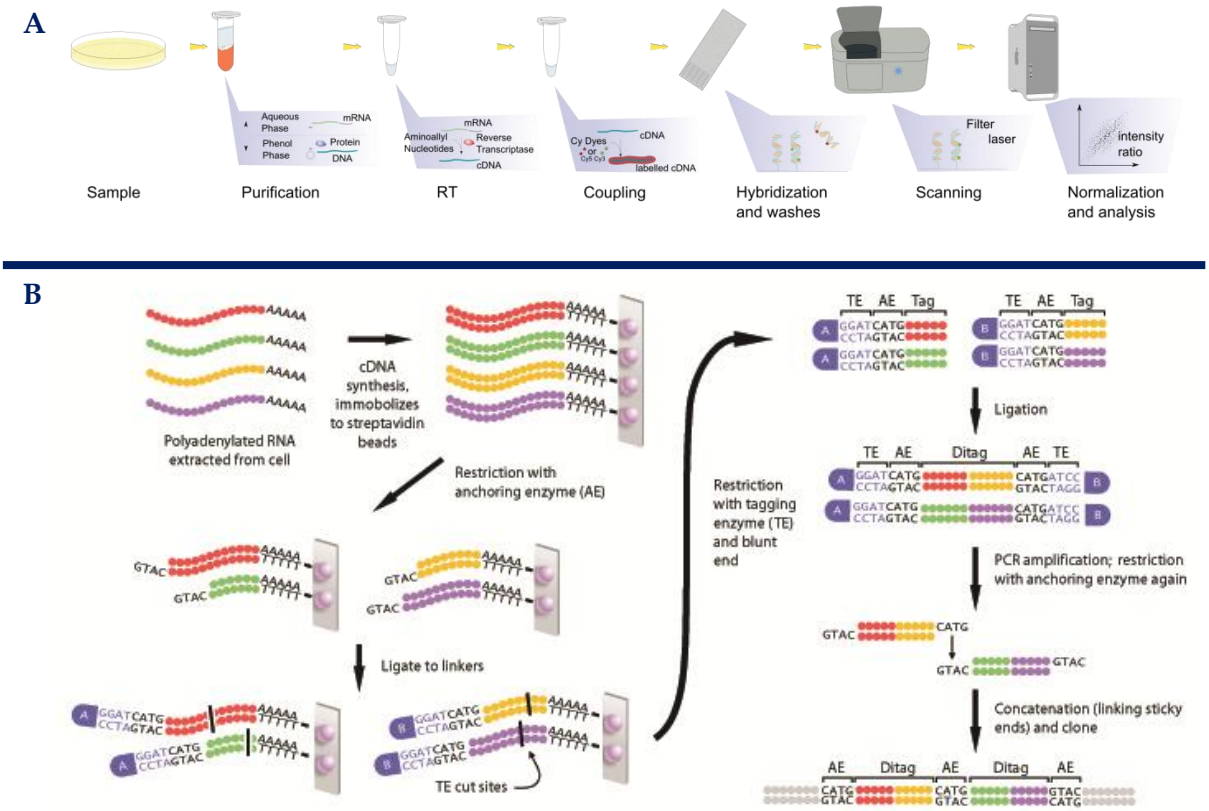


Figure 1-7 The procedure of the DNA microarrays and Serial Analysis of Gene Expression (SAGE). **A)** In the DNA microarray technology, mRNA transcripts are fluorescently labeled during cDNA synthesis. The transcripts are subsequently hybridizes to the array and following removal of unbound material; the fluorescent signal intensity from each spot on the array is measured. **B)** In the SAGE technology, following cDNA synthesis and immobilization to streptavidin beads, sequences are digested with a restriction enzyme. Hereafter, ligation of a linker containing a recognition site of the Type IIS restriction endonuclease BsmFI is performed and the fragment is cleaved 15 bp away in the 3′ direction from the recognition site releasing the sequence the tag. After removal of the linker fragment, tags are concatenated, cloned into a plasmid vector. DNA microarray illustration adapted from: http://en.wikipedia.org/wiki/File:Microarray.svg. SAGE illustration by Jiang Long from "The Science Creative Quarterly", available at: http://www.scq.ubc.ca/painless-gene-expression-profiling-sage-serial-analysis-of-gene-expression/.

The SAGE and DNA microarray technologies both have weaknesses and strengths. DNA microarray data is produced from an analog signal and high background levels and saturation of the signal limits the dynamic range of detection (Wang, Gerstein & Snyder, 2009; Okoniewski & Miller, 2006). SAGE data, on the other hand, is digital counting data and in theory has unlimited dynamic range and is more suitable for a comparison of different data sets (Matsumura *et al.*, 2005). In contrast to microarrays that only measure the transcript abundance of transcripts matching oligos on the chip (a closed system), SAGE can measure transcript abundance without prior knowledge of the investigated transcriptome, and is therefore an open system (Müller, Neumaier & Hoffmann, 2008). However, a small fraction of mRNA transcripts do not contain a restriction site for the anchoring enzyme hereby omitting these from being measured by SAGE, cf. Figure 1-8. This can in part be overcome by construction additional libraries using a different anchoring enzyme (Saha *et al.*, 2002). An alternative could also be to use a combination of two or more

anchoring enzymes. Based on an *in silico* analysis of *Homo sapiens* (*H. sapiens*), *Arabidopsis thaliana* (*A. thaliana*) and *Solanum tuberosum* (*S. tuberosum*) mRNA sequences presented here, this could in theory ensure that between 96.6 % and 99.9 % of these transcripts could be detected by SAGE. Although SAGE experiments do not require costly equipment, they are more expensive and time consuming than microarrays (a full protocol requires 10-14 days) (Matsumura *et al.*, 2005). This limits the ability of SAGE to be high throughput, and is one of the major reasons why the microarray technology became dominating in the 1990's (Marioni *et al.*, 2008).

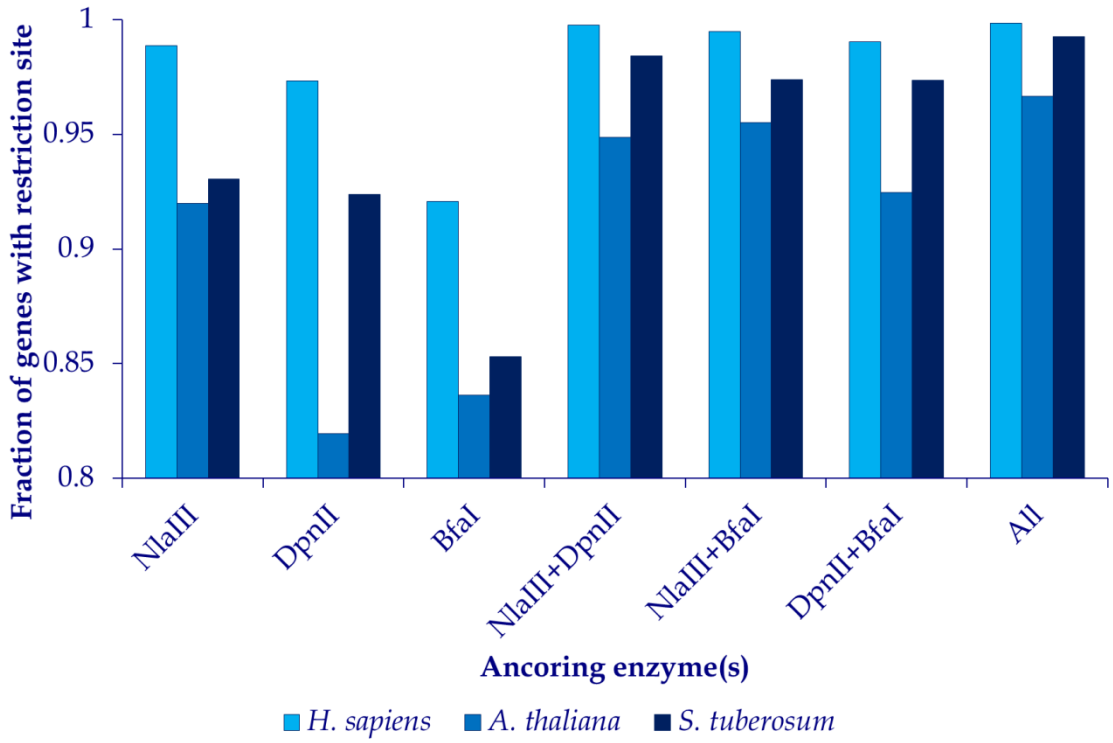


Figure 1-8 Fraction of genes with recognition site for different anchoring enzymes and combinations of these. The fraction of mRNA transcripts from *H. sapiens, A. thaliana*, or *S. tuberosum* having a restriction site for one of the three frequently anchoring enzymes NlaIII (CATG), DpnII (GTAC), BfaI (CTAG). Combinations of two or all three enzymes require the presence of at least one of the enzymes recognition site. Notice the limited Y-axis. Extraction of sequence tags was performed using *GlobalSagemap.pl* (cf. appendix B). The fraction of sequences having a restriction site was subsequently calculated. Sequences used:
A thaliana: Representative cDNA gene models from the TAIR10 (Dec 2010) release.
Found at ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR10_blastsets/.
H. Sapiens: RefSeq mRNA sequences (version 12/092011).
Found at: ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/.
S. tuberosum: Representative transcript models from the Genome Annotation v3.4
Found at: http://potatogenomics.plantbiology.msu.edu/index.html

1.2.3.2 Next Generation Tag Based Transcriptomics

The arrival of NGS technologies opened possibilities for further development of tag based transcriptome methods. Already in 2006, the group of Kåre L. Nielsen introduced the DeepSAGE method, the first utilization of NGS to SAGE by adaption of the LongSAGE method (Saha *et al.*, 2002) to the 454 sequencing platform (Nielsen, Hogh & Emmersen, 2006). Here, di-tag concatenation, clone picking, and sequence template preparation could be omitted from the library preparation protocol hereby shortening the time consumption. Moreover, the larger throughput of 454 sequencing at the time increased the SAGE library sizes to ~ 315,000 tags per sequencing run equivalent to 3 microarray experiments (Lu *et al.*, 2004). In 2008, Nielsen *et al.* introduced sample multiplexing to the protocol (Nielsen, 2008). The same year, adaptation to the Illumina platform and further simplification of the DeepSAGE method, by omitting the creation of di-tags, was presented by the same group in the work by Annabeth H. Petersen (Petersen, 2008). Due to the fact that large parts of the work presented in this thesis are based on this method, it will be shortly described and compared to other similar methods. The current DeepSAGE protocol for library preparation, cf. Figure 1-9, is initialized by mRNA capture to oligo(dT) beads and subsequent cDNA synthesis. Hereafter, the sequences are NlaIII digested, and a biotinylated adaptor sequence containing a MmeI recognition site is ligated to the bead bound sequence. Next, the ligation product is digested with MmeI, releasing a sequence containing a 17-19 bp sequence tag with a two nucleotide over hang originating from the original mRNA transcript. Following purification by capture on streptavidin beads, ligation of a second adaptor containing a 3 nucleotide barcode sequence results in the final amplicon, which is compatible for sequencing on the Illumina sequencing platform. The 3 nucleotide barcode sequence is non-redundant, i.e. one barcode sequence cannot be made from another by one substitution, hereby minimizing sample cross contamination due to sequencing errors in the barcode sequence (Petersen, 2008). Today, this method is implemented and widely used at Aalborg University. Using the Genome AnalyzerII_x sequencing platform, ~0.1 billion sequence tags are routinely generated per sequence run (information based on all sequencing runs performed at Aalborg University, data not shown). This is roughly equal to 96 SAGE libraries with an average size of ~1 million tags per run.

Similar methods to that of DeepSAGE have later been published. One is Tag-seq, by Morrissy *et al.* published in 2009 (Morrissy *et al.*, 2009). This is basically an exact copy of the DeepSAGE method, however without sequence barcoding enabling multiplexing of samples. Another similar method is high throughput SuperSAGE (HT-SuperSAGE), developed by Matsumura *et al.* and published in 2010 (Matsumura *et al.*, 2010). It is a Further development of the SuperSAGE method (Matsumura *et al.*, 2005) adapting this to the Illumina platform. The only obvious difference between DeepSAGE and HT-SuperSAGE having an impact on the type of data generated by the two methods is the use of different tagging enzymes, where MmeI is utilized in DeepSAGE resulting in 21 bp tags and EcoP15I is utilized in HT-SuperSAGE resulting in longer 26 bp tags. The impact of tag length on tag-to-gene annotation will be discussed in section 1.2.3.3.

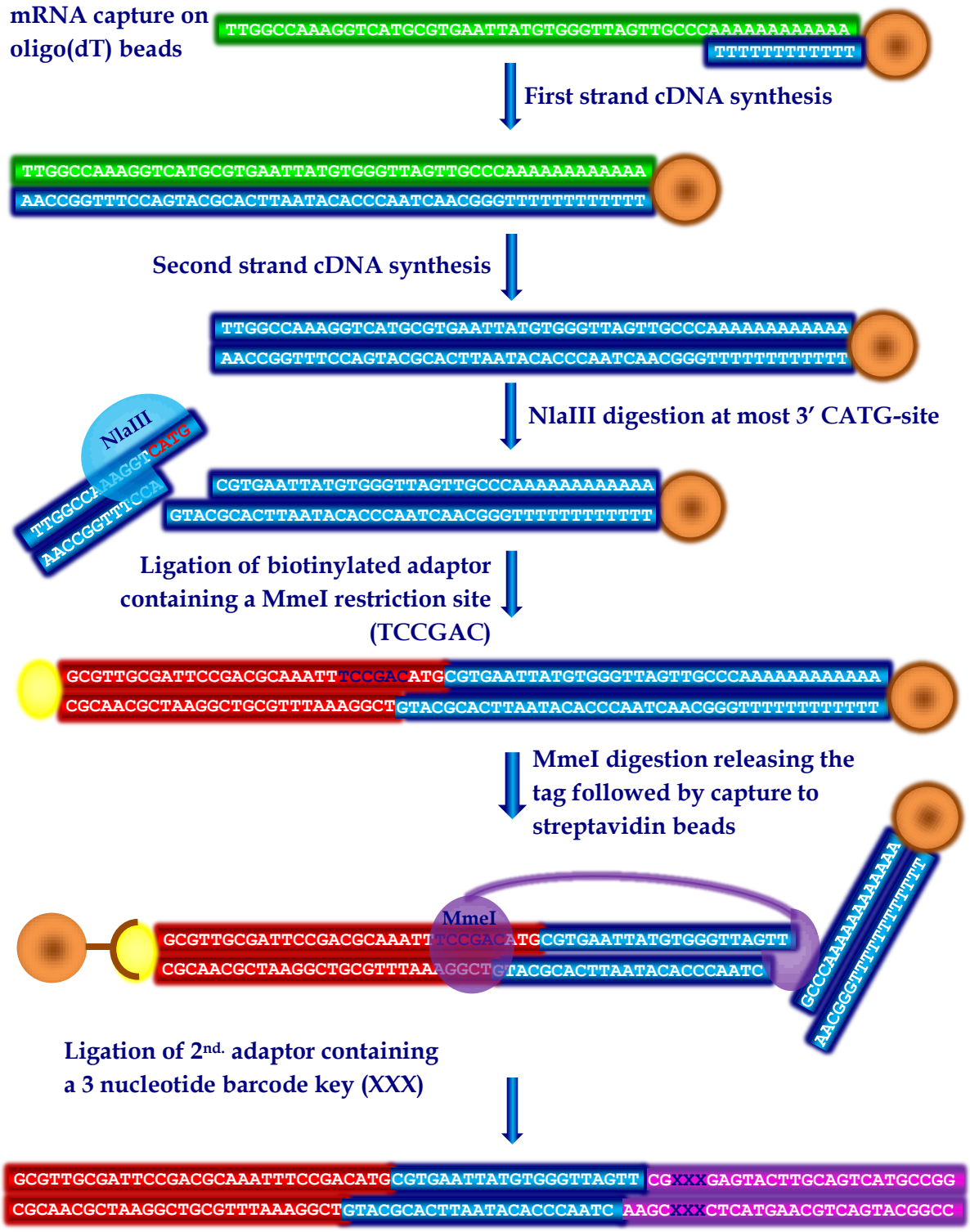


Figure 1-9 Overview of the DeepSAGE protocol. First, mRNA capture to oligo(dT) beads and subsequent cDNA synthesis are performed. Hereafter, sequences are NlaIII digested and a biotinylated adaptor sequence containing a MmeI recognition site is ligated to the bead bound sequence. Next, the ligation product is digested with MmeI, releasing a sequence containing a 17-19 bp sequence tag with a two nucleotide overhang, which is ligated to a second adaptor containing a non-redundant 3 nucleotide barcode containing sequence that enables multiplexing of samples. The final amplicon contains a sequence primer site enabling sequencing on the Illumina platform. Figure made with inspiration from (Petersen, 2008).

1.2.3.3 Annotation of Sequence Tags - Comparison of DeepSAGE and SuperSAGE

A crucial step in SAGE data analysis is reliable tag-to-gene-annotation, in order to extract biological knowledge (Saha *et al.*, 2002). The specificity of the tag-to-gene annotation is dependent on the length of the tag, and has been reported to improve when increased from 14 bp to 21 bp (Saha *et al.*, 2002), and further improve when increased from 21 bp to 26 bp (Matsumura *et al.*, 2005). Based on an *in silico* analysis of 50 tags (Matsumura *et al.*, 2005) the authors behind the SuperSAGE method, which produces 26 bp tags compared to the shorter 21 bp tags produced by Deep-SAGE, claim that quote: "*the 26 bp DNA tag sequence greatly improves the efficiency of gene annotation of the tags*" (Matsumura *et al.*, 2005), and "*These 26-bp tags allow a much better and unambiguous tag-to-gene identification, which is just not possible with shorter tags*" *(Matsumura et al., 2010)*. However, when performing *in silico* analysis of non-redundant transcript sets (1 transcript per gene) containing sequences originating from *H. sapiens*[2], *A. thaliana*[3], or *S. tuberosum*[4] only minor improvements in the uniqueness of the tag-to-gene annotation using longer tag sequences is detected, cf. Figure 1-10.

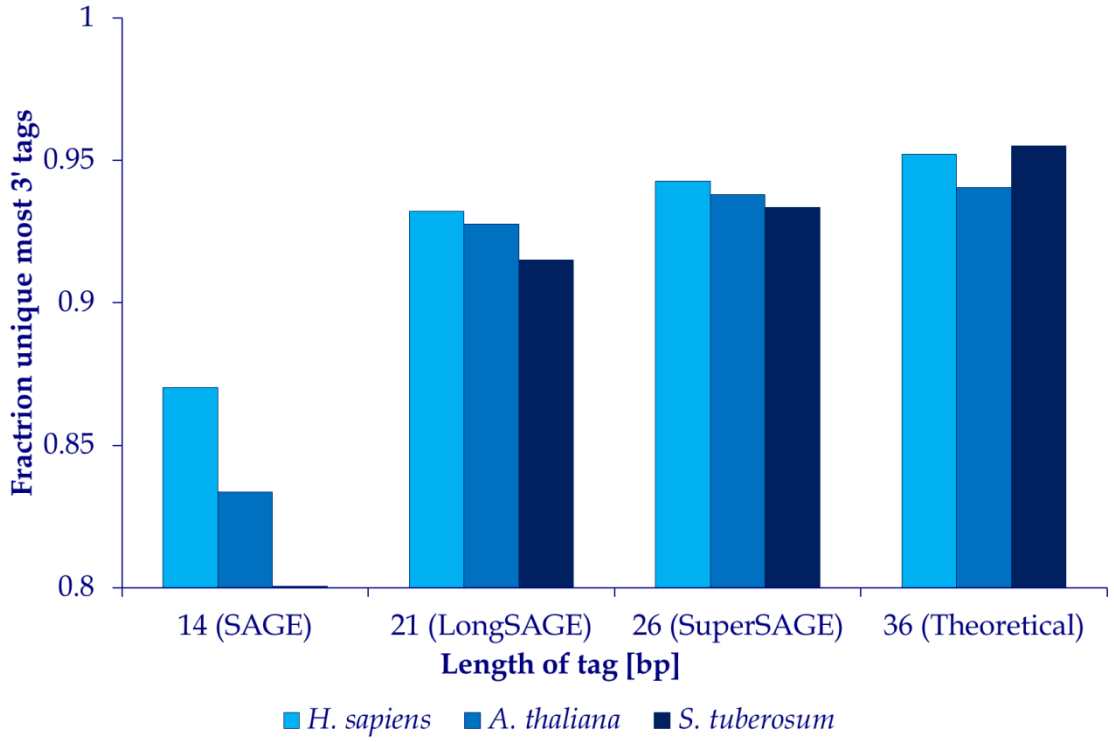


Figure 1-10 Fraction of unique most 3' sequences tags from *H. sapiens*, *A. thaliana* or *S. tuberosum* mRNA transcripts. Using *GlobalSagemap.pl.* The most 3' tag following an NlaIII restriction site (CATG) was extracted from each sequence, while this in theory is the tag extracted by SAGE methods using this tagging enzyme. The number of genes matching each tag was counted and divided by the total number of different tags. For the *H. sapiens* data set one transcript from each protein coding gene was subsequently extracted using *NonRedundantRefSeq.pl*. Notice the limited Y-axis.

On average 83.5 %, 92.5 %, 93.8 %, and 95.0 % of mRNA transcripts have a unique most 3' tag (known as the canonical tag) using 14 bp (SAGE), 21 bp (DeepSAGE), 26 bp (HT-

[2] RefSeq mRNA sequences (version 12/092011). Found at: ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/.
[3] Representative cDNA gene models from the TAIR10 (Dec 2010) release. Found at: ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR10_blastsets/.
[4] Potato: Representative transcript models from the Genome Annotation v3.4. Found at: http://potatogenomics.plantbiology.msu.edu/index.html

SuperSAGE), and 36 bp (theoretical) tags, respectively. This indicates that non-uniqueness of tags mostly is caused by the presence of duplicated genes and large gene families that are highly similar, and that the uniqueness of tags is only slightly improved when increasing the tag length to more than 21 nucleotides.

Another challenge for correct tag-to-gene annotation is the presence of sequence errors in the data. For example, with a 17 bp DeepSAGE tag (omitting the CATG cut site of NlaIII), and with an average read accuracy of 99 %, (roughly that of the Genome AnalyzerII by Illumina (Illumina, 2009)) there is an 84 % change of producing an error free tag. Hence 16 % of the data potentially contains 1 or more sequence errors. To overcome this, error correcting algorithms have been developed. One such is SAGEscreen by Akmaev and Wang (Akmaev, 2008; Akmaev & Wang, 2004). SAGEscreen is a is a multi-step procedure, which empirically estimates error rates based from highly abundant tags, and performs statistical testing to detect possible error containing tags (Akmaev & Wang, 2004).

1.2.3.4 Next Generation Transcriptomics – mRNAseq

By replacing DNA with RNA purification and addition of a cDNA synthesis step, protocols for the next generation sequencing version of Expressed Sequence Tag for transcriptome analysis, RNAseq, was developed soon after the release of all three major NGS technologies. One of the first studies using NGS for transcriptome analysis was the work of Emrich *et al.* published late 2006 (Emrich *et al.*, 2007). Here, they combined laser capture micro dissection with a single run of 454 sequencing to produce more than 261,000 ESTs from maize. The data was used for gene discovery and annotation, and at the time, it increased the number of maize ESTs deposited in Genbank by more than 40 % (Emrich *et al.*, 2007), showing the potential of NGS in transcriptomics. Shortly after, a study by the same group was published where 454 transcriptome sequencing was used for discovery of single nucleotide polymorphisms (SNPs) (Barbazuk *et al.*, 2007). The following year in 2008, several studies utilizing RNAseq were published. Among these were gene expression profiling studies in mouse (Mortazavi *et al.*, 2008), yeast (Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008), and human (Sultan *et al.*, 2008). The latter study by Sultan *et al.* further demonstrated the ability of RNAseq to investigate alternative splicing events. Moreover, the potential of RNAseq for investigation of cytosine methylation (an epigenetic regulation) and small RNAs was shown in an *A. thaliana* study by Lister *et al.* (Lister *et al.*, 2008). Even though the RNAseq technology was at an early stage, it was found in a study by Marioni *et al.* to produce highly replicable results with low technical variation at least comparable and in some ways superior when compared to the microarray technology (Marioni *et al.*, 2008). These breakthrough studies are some of the reasons why RNAseq was called: *"A revolutionary tool for transcriptomics"* and was *"expected to revolutionize the manner in which eukaryotic transcriptomes are analyzed"* in a review by Wang, Gerstein and Snyder in 2009 (Wang, Gerstein & Snyder, 2009). Several later studies (Bradford *et al.*, 2010; Tang *et al.*, 2009; Cloonan *et al.*, 2008; Marioni *et al.*, 2008; Mortazavi *et al.*, 2008) comparing gene expression micro arrays and quantitative mRNAseq also reported good correspondence between the two methods in regards to gene expression and fold changes, but with mRNAseq outperforming the microarray technology in regards to gene detection rates. Following this new approach, there has been a fast development of algorithms and methods for specific analysis of RNAseq derived data nicely reviewed by Garber *et al.* in 2011 (Garber *et al.*, 2011). Core parts of RNAseq analysis involve read alignment, transcript assembly and transcript quantification. In the following sections, methods for analysis of these core parts will be described. Follow-

ing sequencing, the mRNAseq reads are most often either aligned to an existing genome sequence or assembled *de novo* (Wang, Gerstein & Snyder, 2009). This result in the challenges described in 1.2.2.1 and 1.2.2.2 for alignment and *de novo* assembly, respectively, but additional challenges arise when dealing with mRNAseq data.

1.2.3.5 Genome Alignment of mRNAseq Data

If an annotated genome sequence for the organism being investigated is available, the first step of mRNAseq data analysis often is often read alignment (Garber *et al.*, 2011; Wang, Gerstein & Snyder, 2009). In this regard, two additional bioinformatic challenges (or opportunities) arise, namely mapping of RNAseq data reads spanning exon-exon boundaries and mapping of reads containing parts of the PolyA tail of the mRNA transcript.

By identifying reads that contain multiple A's or T's at the end and a matching remaining part, enable detection of the 3' gene boundary at the nucleotide level (Wang, Gerstein & Snyder, 2009). Two early studies in yeast showed the potential of RNAseq for analysis of UTRs, leading to the discovery of several not previous analyzed regions, and showing extensive heterogeneity of the 3' end, both locally within a few base pair window, but also showing distinct regions of polyA addition in several genes (Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008).

In early RNAseq studies, alignment of exon-exon spanning reads where performed by compiling an extra "junction library" containing predicted junction sequences (Wang, Gerstein & Snyder, 2009; Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008). By expanding this library to represent all possible exon combinations within each gene, Sultan *et al.* investigated alternative splicing in the human transcriptome hereby confirming 90,145 junctions, and detecting 4,096 novel junctions (Sultan *et al.*, 2008). However, this method, using un-spliced alignment tools, relies on existing gene and transcript annotation, and fails to detect splicing events involving new exons (Garber *et al.*, 2011). This has led to development of so-called "spliced aligners", enabling mapping of intron spanning reads requiring large gaps in the alignment (Garber *et al.*, 2011). Two general classes of spliced aligners exist today, the "exon first" and the "seed and extend" aligners (Garber *et al.*, 2011). Aligners such as MapSplice (Wang *et al.*, 2010), SpliceMap (Au *et al.*, 2010) and Tophat (Trapnell, Pachter & Salzberg, 2009), the latter being one of the first "seed and extend" spliced aligners, use a two-step alignment process. First, reads are mapped using an un-spliced alignment algorithm. Following this, unmapped reads are split into shorter segments, and regions of the genome near parts with read coverage are then searched for possible spliced connections. Spliced aligners such as QPALMA (De Bona *et al.*, 2008) and GSNAP (Wu & Nacu, 2010) utilizes the "seed and extend" method, where reads are split into shorter parts, which give rise to candidate regions for alignment on the genome. Secondly a more sensitive alignment method, such as Smith-Waterman alignment in the case of QPALMA (De Bona *et al.*, 2008), is used to determine the optimal alignment on the genome. Today, the "first exon" methods are faster, due to the initial use of fast un-spliced alignment, but these are biased towards less optimal un-spliced alignment (e.g. less optimal alignment to an un-spliced pseudogene, instead of correct optimal spliced alignment to the expressed gene) (Garber *et al.*, 2011).

1.2.3.6 Transcriptome Reconstruction Using mRNAseq Data

Reconstruction of the transcriptome (i.e. all transcripts including different splice variants) of an organism has become possible, since the arrival of mRNAseq. Reconstruction can either

be genome sequence assisted or be performed completely independent (similar to *de novo* genome assembly) (Garber *et al.*, 2011). Assembling the transcriptome has additional challenges compared to genome assembly, cf. section 1.2.2.2. Firstly, the transcriptome is represented by unequal coverage of several orders of magnitude due to difference in expression between genes. Secondly, several transcript variants can exist for each gene. Moreover, mRNAseq data also contain sequences originating from incompletely or mis-spliced mRNA that contain intronic sequences complicating the assembly problem even further (Garber *et al.*, 2011).

First generation algorithms for genome assisted transcriptome reconstruction such as G-Mo.R-se (Denoeud *et al.*, 2008) relied on primary identification of exons as regions with read coverage, and secondly establishment of connections between exons by the use of spliced reads across the regions with coverage (Yassoura *et al.*, 2009; Denoeud *et al.*, 2008). However, this method has proven to be insufficient to reconstruct lowly expressed genes and genes with multiple transcript variants (Garber *et al.*, 2011). Newer genome assisted transcriptome reconstruction methods use information from longer spliced reads directly to assemble the transcriptome. These algorithms include Cufflinks (Trapnell *et al.*, 2010) and Scripture (Guttman *et al.*, 2010). Both are graph based methods and both uses Tophat (Trapnell, Pachter & Salzberg, 2009) for the initial spliced read alignment to the genome. A difference between the two, is that Scripture provides maximum sensitivity by reporting all isoforms compatible with the read (Guttman *et al.*, 2010), whereas Cufflinks provides maximum precision by only reporting the minimal number of compatible isoforms (Trapnell *et al.*, 2010).

Direct *de novo* assembly of the mRNAseq reads is independent of a genome sequence and hereby enables transcriptome analyses in organism without an existing genome model (Garber *et al.*, 2011). TransAbySS (Robertson *et al.*, 2010) is an algorithm and analysis pipeline based on the *de novo* assembler AbySS (Simpson *et al.*, 2009), which also has been used to assemble mRNAseq data (Birol *et al.*, 2009). The major difference between TransAbySS and AbySS is that TransAbySS use a variable k-mer strategy in order to deal with the difference in gene expression levels and multiple transcript isoforms (Robertson *et al.*, 2010).

1.2.4 Analysis of Sequence Based Transcriptome data

Quantitative transcriptomics or gene expression profiling has long been the most widely used transcriptomic application (Garber *et al.*, 2011). As mentioned, DNA microarrays have long been the choice of method, when preforming global gene expression analyses, and DNA microarray data analysis methods are well-established (Garber *et al.*, 2011). However, sequence based transcriptome data is digital count data, why the developed methods for microarray analyses cannot be directly transferred. The primary goal of a both mRNAseq and tag based transcriptome experiments is to produce a list of tags, transcripts or genes with a corresponding expression value for each sample in the experiment. Though both tag based methods and RNAseq data by nature is digital count data, several systematic biases need to be taken into account for proper assessment of the transcript distribution in a sample. These will be discussed in section 1.2.4.1. The objective for the downstream analysis is to extract biological knowledge in order to make conclusions or new hypotheses based on the experiment. While transcriptome data contains thousands or millions of variables, the perhaps few biological interesting variables are well hidden. Methods for extraction of biological knowledge from transcriptome data will be discussed sections 1.2.4.2, 1.2.4.3, and 1.2.4.4. In the following sec-

tions the terms "gene" and "transcript" will be used, but the methods described for estimation of gene expression and extraction of biological knowledge are equally valid for data based on sequence tags if otherwise is not mentioned.

1.2.4.1 Estimation of Gene Expression Values

When estimating the gene expression using RNAseq data, two major biases need to be taken into account. Firstly, the number of sequences originating from each transcript in each sequence library will be a function of the relative abundance of the transcript (the factor that needs to be estimated), but also of the transcript length (longer transcripts produces more fragments), and the total number of sequences in each library, the latter is also valid for tag based transcriptome data. These two major factors were recognized and taken into account by Mortazavi *et al.* (Mortazavi *et al.*, 2008), when they defined the RPKM expression value (reads per kilobase of exon model per million mapped reads), cf. equation (1-1)

$$\mathrm{RKPM} = \frac{10^9 \cdot \mathrm{C}}{\mathrm{N} \cdot \mathrm{L}} \qquad \textbf{(1-1)}$$

Where:

RPKM = Expression value of gene [reads per kb of exon model per million mapped reads]
C = number of mappable reads that fall onto the gene's exons [#]
N = Total number of mappable reads in the experiment [#]
L = The sum of the exons [bp] in base pairs

Later, Trapnell *et al.* defined the analogous FPKM value (expected fragments per kilobase of transcript per million fragments sequenced) accounting for paired-end sequence data (Trapnell *et al.*, 2010).

Biases in the read distribution caused by library preparation, which could affect the estimation of the expression level have also been investigated. Dohm *et al.* studied general biases in Illumina DNA sequencing, and found that GC rich regions where overrepresented (Dohm *et al.*, 2008) ; i.e. GC rich transcripts could be overrepresented. Mortazavi *et al.* noted that fragmentation of the RNA prior to reverse transcription resulted in a more uniform coverage within each gene (Mortazavi *et al.*, 2008). The groups of Li, Jiang, and Wong and Hansen, Brenner, and Wong have also studied non-uniformity of the read coverage (Hansen, Brenner & Dudoit, 2010; Li, Jiang & Wong, 2010). Li *et al.* suggested using a Poisson model with variable rates to take the non-uniformity into account, and proposed two models for estimation of these rates (Li, Jiang & Wong, 2010). Hansen and colleagues on the other hand, showed that the non-uniformity was caused by the use of random hexamer primers during cDNA synthesis, and provided a weighting scheme to account for this phenomenon (Hansen, Brenner & Dudoit, 2010).

Several developments have been made in regards to quantification accuracy of transcripts originating from large gene families or genes with multiple isoforms. Parts of these transcripts are nearly or completely identical, causing transcript assignment to be challenging at the least. One strategy, named ALEXA-seq, proposed by Griffith and colleagues relies on uniquely matching reads to estimate the isoform-level expression (Griffith *et al.*, 2010). However, this method fails if no unique exon for an isoform exists (Garber *et al.*, 2011). To overcome this, statistical models for estimation of transcript isoforms best explaining the observed read distribution have been developed. Examples of these developments are the work by Jiang and Wong (Jiang & Wong, 2009), the work of Li *et al.* (Li *et al.*, 2009), the implementation in the program MISO

by Katz *et al.* (Katz *et al.*, 2010), and the implementation in Cufflinks (Trapnell *et al.*, 2010). Often, the quantitative transcriptomics is simplified to the gene level, where gene expression is defined as the sum of the expression of all the isoforms of a gene (Garber *et al.*, 2011). While this abundance is difficult to estimate, two common simplified schemes for calculation of the gene expression exist (Garber *et al.*, 2011). One is the ”exon intersection method” where the expression is based on reads mapping to constitutive exons of a gene (Bullard *et al.*, 2010). This method has analogies to expression calculation using DNA microarrays (Garber *et al.*, 2011). The other approach is the “exon union” method, where the expression is based on reads mapping to any exon of a gene. This method is implemented in e.g. ALEXA-seq (Griffith *et al.*, 2010) and Cufflinks (Trapnell *et al.*, 2010).

1.2.4.2 Determination of Differential Gene Expression

A classic setup for transcriptome experiments is a comparative study of e.g. “disease vs. healthy” or “treated vs. non-treated” biological samples. With this setup the motive is often to detect genes that have differential expression (DE) between the two states, and from this set make biological interpretations. Statistical models and methods for DE identification using DNA microarray are well established (Nature Genetics Editors, 2005) and implemented (e.g. in the R package Limma by Gordon Smyth (Smyth, 2004)), due to the extensive use of these in the past decade. The development of NGS technologies has motivated development of statistical methods for DE identification based on sequence data. This development has primarily been inspired by statistical methods developed for SAGE data, due to the digital character of the data, but also in part by methods developed for microarray data (Robinson, McCarthy & Smyth, 2010).

Several methods for detection of differential expressed genes in SAGE data were developed in the late 1990s up until the arrival of NGS based transcriptomics, outdating SAGE. However, today’s methods are further developments of the SAGE analysis methods, making a description of these worthwhile mentioning. Following the assumption that reads are independently sampled from a population with fixed gene abundances, the distribution of the read counts could be approximated by the Poisson distribution, which e.g. was used in the work of Madden *et al.* (Madden *et al.*, 1997). Zhang and colleagues modeled the data using a binomial distribution, (Zhang *et al.*, 1997) and others used an approximation of the normal distribution (Man, Wang & Wang, 2000; Kal *et al.*, 1999; Michiels *et al.*, 1999; Madden *et al.*, 1997). Furthermore bayesian: approaches were also used (Lal *et al.*, 1999; Chen *et al.*, 1998; Audic & Claverie, 1997). In a comparative study by Man *et al.* of several of the above mentioned methods, these performed equally well for higher (>20) tag counts (Man, Wang & Wang, 2000). Later, it was realized that SAGE data was overdispersed, i.e. more variation could be observed than that explained by sampling (Blackshaw *et al.*, 2003). The overdispersion needs to be taken into account, to ensure high specificity and sensitivity of DE identification. Baggerly *et al.* were among the first to account for what they termed “between library variation”, with their t_w-test, which was based on the beta-binomial distribution (Baggerly *et al.*, 2003). They later used logistic regression with overdispersion to accommodate more than a two group comparison (Baggerly *et al.*, 2004), which is closely related to the log-linear model approach developed by Lu *et al.* (Lu, Tomfohr & Kepler, 2005).

Some early mRNAseq studies used for DE identification, like the one by Marioni *et al.* (Marioni *et al.*, 2008) and the one by Bloom *et al.* (Bloom *et al.*, 2009) failed to take overdispersion into account, and based their approaches on the Poisson distribution. These methods were later implemented into the R package DEGseq by Wang et al (Wang *et al.*, 2009). However, other studies showed that

mRNAseq data also were overdispersed (Nagalakshmi *et al.*, 2008). In 2007, Robinson and Smyth introduced a model based on the negative binomial distribution, which is applicable for both SAGE and mRNAseq data (Robinson & Smyth, 2008; Robinson & Smyth, 2007). Initially inspired by the work of Smyth on microarray data (Smyth, 2004), they used a single estimation of the dispersion (Robinson & Smyth, 2008) for the entire gene set giving the mean-variance relationship described by equation (1-2).

$$\sigma^2 = \mu + \alpha \cdot \mu^2 \qquad (1\text{-}2)$$

Where:
σ^2 = Variance of gene expression
μ = Mean of gene expression
α = Dispersion parameter

Later, they expanded their model, by calculating a tag-wise estimation of the dispersion, which is squeezed towards the common dispersion (Robinson & Smyth, 2007). This methodology was implemented in the R package (Team, 2012) EdgeR in 2009 (Robinson, McCarthy & Smyth, 2010), which since has been widely for DE identification in mRNAseq studies (cited 44 times in Web of Science at the time of writing). In 2010, Robinson and Oshlack introduced a gene expression normalization method, which accounts for the bias caused by genes uniformly highly expressed in one condition, hereby occupying a large part of "sequence space" in that library, and hereby lowering the measured expression level of other genes (Robinson & Oshlack, 2010). This normalization method, which is implemented in EdgeR (Robinson, McCarthy & Smyth, 2010) ensures (under the assumption that most genes are not DE genes) that equally expressed genes are assigned the same level of expression (Robinson & Oshlack, 2010). A similar approach, also based on the negative binomial distribution was introduced by Simon and Huber and implemented in the R package DESeq (Anders & Huber, 2010). Here, they use local regression for determination of the mean-variance relationship. The latest method developed (at the time of writing), is that of Hardcastle and Kelly (Hardcastle & Kelly, 2010). Their method is also based on the negative binomial distribution but use an empirical bayesian approach to establish posterior probabilities of multiple models of differential expression. In their study, they provide a comparison of Bayes (Hardcastle & Kelly, 2010) EdgeR (Robinson, McCarthy & Smyth, 2010), the overdispersed log-linear model of Lu Tomfohr, and Kepler (Lu, Tomfohr & Kepler, 2005), the overdispersed logistic model of Baggerly *et al.* (Baggerly *et al.*, 2004), DEGseq (Wang *et al.*, 2009), and DESeq (Anders & Huber, 2010). They showed that DEGseq in general perform poorly compared to all other methods, especially if the data have a high proportion of unidirectional differential expression (either up- or down- regulated genes) (Hardcastle & Kelly, 2010). They also show that their method, BaySeq, and EdgeR outperform the other methods and that BaySeq and EdgeR perform almost identically (Hardcastle & Kelly, 2010).

1.2.4.3 Visualization of Gene Expression Patterns

To elucidate biological relevant gene expression patterns, such as tissue specific expression or differential expression of genes belonging to the same biological pathway, in large transcriptome data sets, other analysis methods are needed. Of these, hierarchical clustering and principal component analysis (PCA) are two widely used methods, why these will be shortly described.

The general purpose of clustering is visualization of similar gene expression profiles (clustering of samples) and genes with similar gene expression patterns across samples (clustering of

genes) (Eisen *et al.*, 1998). Although other commonly used clustering methods like k-means clustering (MacQueen, 1967), self-organizing maps (Kohonen, 1982) and support vector machines (Boser, Guyon & Vapnik, 1992) exist, hierarchical clustering is the most widely used clustering method for gene expression data (de Hoon *et al.*, 2004). This is illustrated by the fact that as of time of writing, the bioinformatic software documented by Eisen in 1998 (Eisen *et al.*, 1998) is the third most cited article in PNAS (National Academy of Sciences, 2011) and it has been used in more than 4,000 biological or biomedical publications. An enhancement of the method has later been developed by de Hooen *et al.* and implemented in the open source clustering software Cluster 3.0 (de Hoon *et al.*, 2004). The output of a cluster analysis is displayed graphically as a heatmap (Eisen *et al.*, 1998). The heatmap is appended cluster trees to its margin facilitating visual inspection of gene and sample relationships (Weinstein, 2008), which is intuitive for biologists (Eisen *et al.*, 1998). A small example of a heatmap is shown in Figure 1-11, where leaf and tuber samples co-cluster based on Fructose-bisphosphate aldolase gene expression of 8 loci.

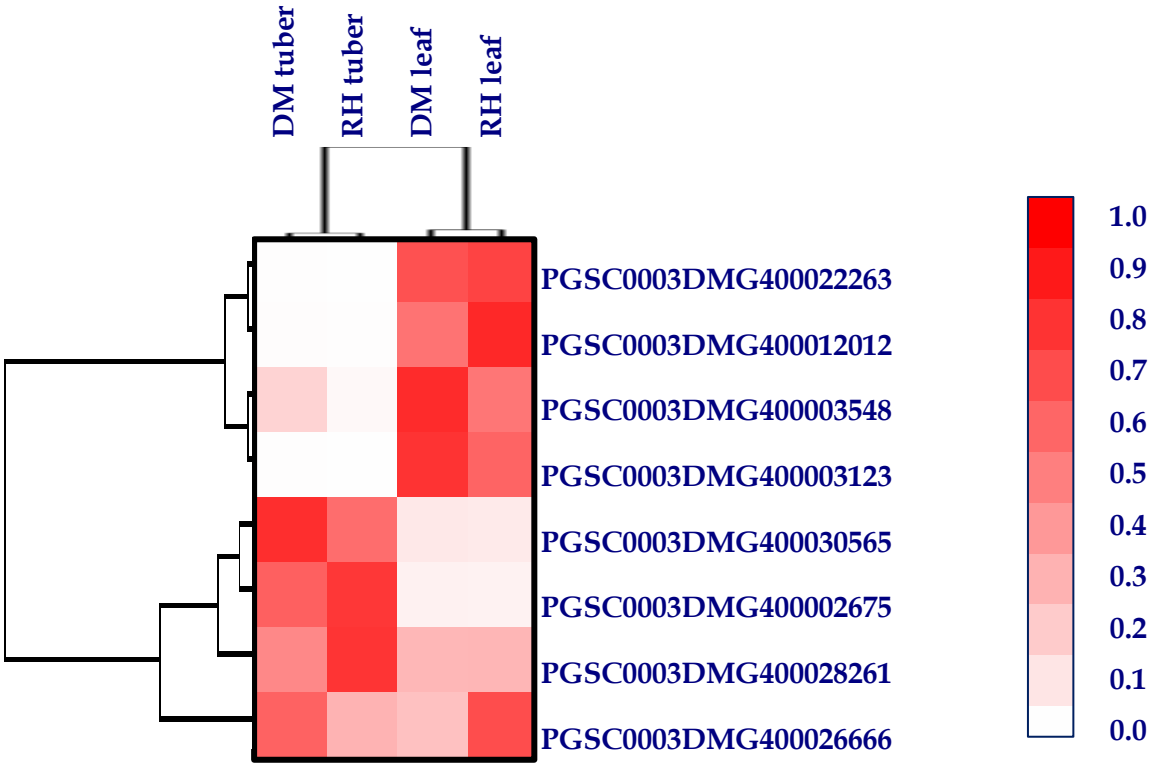


Figure 1-11 Hierarchical clustering of the gene expression measured as FKPM of the 8 loci in the potato genome that are annotated as Fructose-bisphosphate aldolase in four gene expression profiles. The gene expression profiles are from leaf and tuber tissue from the two varieties (DM and RH) used for the genome sequence of potato (The Potato Genome Sequencing Consortium *et al.*, 2011). Subsequent clustering, expression values are normalized between 0 and 1 for visualization purposes. The expression data was retrieved from: http://potatogenomics.plantbiology.msu.edu/index.html.

Principal component analysis (PCA) is a different multivariate analysis method, which can be used to explore transcriptome data. It is a mathematical algorithm that reduces dimensionality of a data set by identifying directions in the data, so-called principal components (PCs), where the variation is maximal (Jolliffe, 2002). The data can be represented using only a few PCs, instead of thousands of variables, while most of the variation in the data is retained (Jolliffe, 2002). The output of a PCA is score plots, which is any bi-plot of PCs plotted against each other, also named the "map of samples" and loadings plots, showing how much each variable (gene) contributes to each PC. By visual inspection of these plots, it is possible to detect

similarities (and differences) between samples, possible groupings of the samples and possible sample outliers (Ringnér, 2008). A small example of a PCA using the gene expression from the same 8 Fructose-bisphosphate aldolase loci is shown in Figure 1-12.

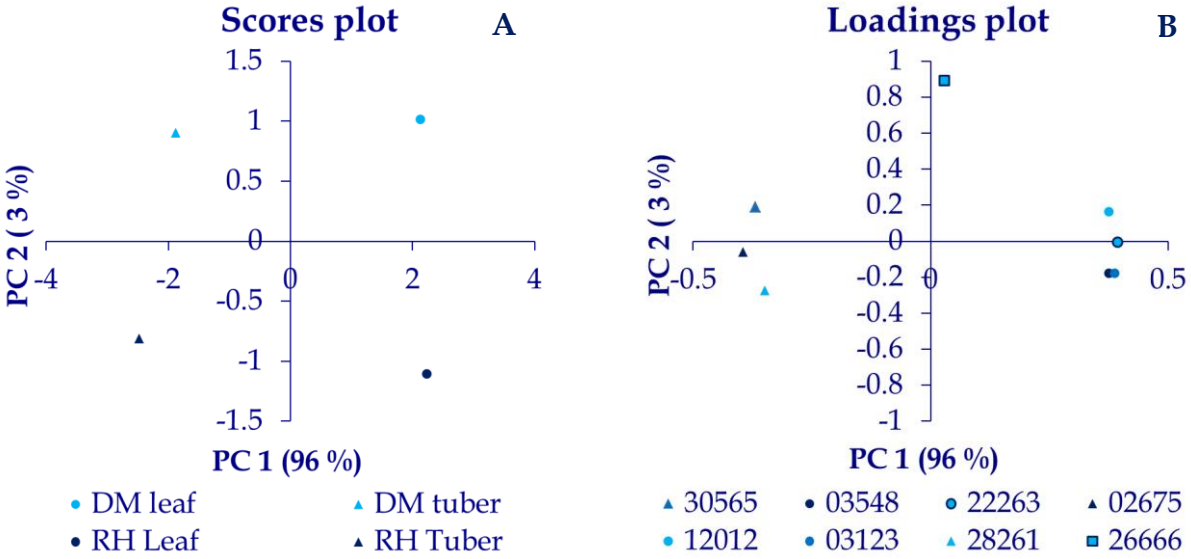


Figure 1-12 PCA analysis of the same 8 loci of Fructose-bisphosphate aldolase clustered in Figure 1-11. **A)** Scores plot of the 1st and 2nd principal components. PC1 splits the samples according to tissue type, accounting for 96 % of the explained variance. PC2 splits the samples according to genotype accounting for 3 % of the variance. **B)** Loadings plot of the 1st and 2nd principal components. Genes marked with triangles are positively correlated with tuber samples, i.e. are highly expressed in tuber samples, while genes marked with circles are highly expressed in leaves. The gene marked with a square has similar expression in both tubers and leaves, but have higher expression in the DM genotype compared to the RH genotype. PC = principal components. The explained variance is given in percentage. The expression data was retrieved from: http://potatogenomics.plantbiology.msu.edu/index.html.

One challenge of PCA is to decide which type of data normalization to use. Mean centering (subtraction of the mean of a variable from each data point) of the data, is normally performed hereby centering the data at origo. Moreover, standardization can be performed by to unit variance scaling, also named auto scaling (Esbensen, 2000), where each variable is weighted with the inverse standard deviation, hereby normalizing the variance of each gene to 1, cf. equation (1-3) (Esbensen, 2000). This transformation equals the weight of all genes to the PCA model regardless of their original variance and expression level (Esbensen, 2000). The danger of this normalization is that "noise variables" are over emphasized. However, by using selective weighting, or choosing an offset in the standardization, this can be overcome (Esbensen, 2000). Other scaling methods have been used (van den Berg *et al.*, 2006), and "Pareto scaling" (Eriksson, 1999), cf. equation (1-4), have been found to be practically useful (Eriksson *et al.*, 2004; Atif *et al.*, 2003), as it reduces the relative importance of large expression values, but keep the structure of the data partially intact (van den Berg *et al.*, 2006).

$$\text{Auto scaling:} \quad \mathrm{Xnorm_{ij}} = \frac{\mathrm{X_{ij} - Xmean_i}}{\mathrm{s_i}} \qquad \textbf{(1-3)}$$

$$\text{Pareto scaling:} \quad \mathrm{Xnorm_{ij}} = \frac{\mathrm{X_{ij} - Xmean_i}}{\sqrt{\mathrm{s_i}}} \qquad \textbf{(1-4)}$$

Where:

$\mathbf{X_{ij}}$ = Raw gene expression for gene i in sample j

$\mathbf{Xnorm_{ij}}$ = Normalized gene expression for gene i in sample j

$\mathbf{Xmean_i}$ = Mean gene expression for gene i $= \frac{1}{J}\sum_{j=1}^{J} X_{ij}$

$\mathbf{s_i}$ = Standard deviation for gene i $= \sqrt{\frac{1}{J}\sum_{j=1}^{J} \frac{(X_{ij} - Xmean)^2}{J-1}}$

PCA identifies directions in the data with large variation, and not directions relevant for separating classes (Ringnér, 2008). However, this is possible using partial least squares discriminant analysis (PLS-da). PLS-da is a classification method based on partial least squares regression (PLS-R), which relates variations in one or several variables (e.g. control vs. disease) to the variations of predictor variables (genes) (Esbensen, 2000). This method have for example been used for variable selection for further analysis (Lê Cao, Boitard & Besse, 2011) and for cancer classification using gene expression profiling (Tang *et al.*, 2010).

1.2.4.4 Ontological Assisted Data Analysis

Classification of genes into groups (e.g. by enzymatic function or pathway) can simplify the data into higher order information and facilitate generation of biological hypotheses based on transcriptomic studies (Tian *et al.*, 2005). This classification requires a framework linking genes and groups. The two most important frameworks are the Gene Ontology (GO) (Ashburner *et al.*, 2000) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa *et al.*, 2009; Kanehisa & Goto, 2000). Gene Ontologies are controlled vocabularies for gene and protein roles in cells. Three overall ontologies exist: name biological process, molecular function, and cellular component (Ashburner *et al.*, 2000). GO terms are organized in a hierarchical tree-like structure called a directed acyclic graph, where a node can have several parents (Khatri & Draghici, 2005). The KEGG pathway database is a representation of higher order gene functions in terms of a network of interacting molecules (such as a biological pathway) (Kanehisa & Goto, 2000). These frameworks enable statistical testing of groups of genes, which is less vulnerable to false positives. Several bioinformatic tools have been developed for performing such tests either based on KEGG or GO terms (see for example review of 14 such tools by Khatri and Draghici (Khatri & Draghici, 2005)). A typical approach of many of these tools is enrichment analysis. Here a list of DE genes is compared to a background list (e.g. all genes of an organism, all genes on a microarray chip, or all transcripted genes observed in an experiment). Overrepresented annotations GO terms or pathways (e.g. determined by a Fisher's exact test, a hyper geometric test or a X^2 test) in the list of DE genes are then highlighted (Huang, Sherman & Lempicki, 2009). Examples of such tools are the BiNGO plugin (Maere, Heymans & Kuiper, 2005) for the molecular interaction networks visualization tool Cytoscape (Shannon *et al.*, 2003), and a database for annotation, visualization, and integrated discovery (DAVID) that enables both GO term and KEGG pathway analysis (Huang, Sherman &

Lempicki, 2009; Dennis Jr. *et al.*, 2003). Analogues method has been developed for pathway analysis of genome-wide association studies (O'Dushlaine *et al.*, 2011; Jia *et al.*, 2010; O'Dushlaine *et al.*, 2009). Here, all SNPs in a pathway instead of single SNPs are tested for association to a condition (e.g. disease). Although ontological approaches is a promising strategy for identification of biological processes responsible for a studied phenotype (Huang, Sherman & Lempicki, 2009), these approaches suffer from some important limitations (Khatri & Draghici, 2005). Firstly, all existing functional annotation databases are incomplete, i.e. only a subset of the known genes are functionally annotated (King *et al.*, 2003). Secondly, a large part of the functional annotation have been electronically inferred without human involvement (Khatri & Draghici, 2005), illustrated by two of the most well annotated genomes *A. thaliana* and *H. sapiens*, where at the time of writing, 25 % and 57 % of GO terms, respectively were solely electronically inferred (The Gene Ontology Consortium, 2011).

1.2.5 Software for NGS Data Analysis

In the following, bioinformatic tools for NGS data are listed, cf. Table 1-2. The development of NGS software is rapid, why the list quickly becomes outdated. Therefore, links to current lists of NGS software are given at the bottom of the table.

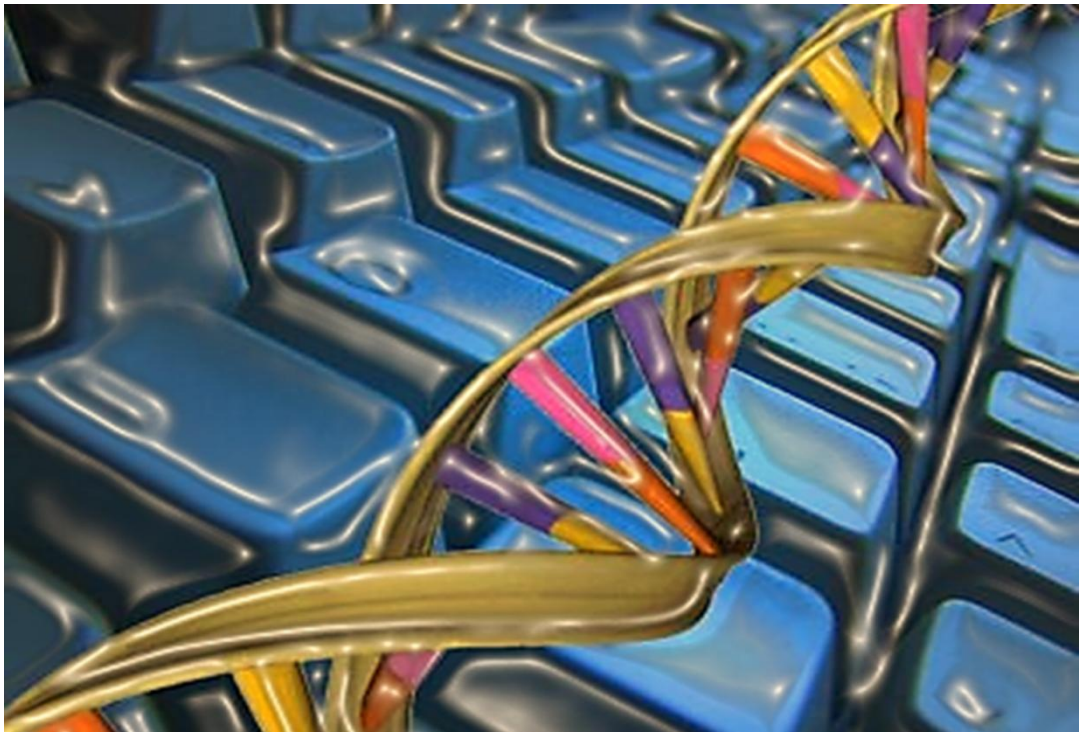
Table 1-2 List of bioinformatic software used to analyze next-generation sequencing data. Links to current lists are given at the bottom.

Name	Algorithm Principal / Statistical model	Data type	Citation
Read Aligners			
MAQ	Hash-based	Illumina, SOLiD	(Li, Ruan & Durbin, 2008)
SOAP	Hash-based	Illumina	(Li *et al.*, 2008)
CLC Genomics Workbench	Hash-based	All	(CLC Bio, 2010)
Bowtie	FM-index and BWT	Illumina	(Langmead *et al.*, 2009)
BWA	FM-index and BWT	Illumina, 454	(Li & Durbin, 2009)
SOAP2	FM-index and BWT	Illumina	(Li *et al.*, 2009)
De novo Genome Assembly			
Arachne	Overlap/layout/consensus	Sanger	(Batzoglou *et al.*, 2002)
Celera Assembler	Overlap/layout/consensus	Sanger, 454, Illumina	(Myers *et al.*, 2000)
Newbler	Overlap/layout/consensus	Sanger, 454	(Margulies *et al.*, 2005)
Edena	Overlap/layout/consensus	Illumina	(Hernandez *et al.*, 2008)
SSAKE	Iterative extension	Illumina	(Warren *et al.*, 2007)
VCAKE	Iterative extension	Illumina	(Jeck *et al.*, 2007)
SHARCGS	Iterative extension	Illumina	(Dohm *et al.*, 2007)
EULER	de Bruijn Graph	Sanger, 454	
Velvet	de Bruijn Graph	All	(Zerbino & Birney, 2008)
Allpaths	de Bruijn Graph	Illumina, SOLiD	(Butler *et al.*, 2008)
AbySS	de Bruijn Graph	Illumina, SOLiD	(Simpson *et al.*, 2009)
SOAPdenovo	de Bruijn Graph	Illumina	(Li *et al.*, 2010c)
Spliced Genome Aligners			
MapSplice	Exon first	Paired-end	(Wang *et al.*, 2010)
SpliceMap	Exon first	Paired-end	(Au *et al.*, 2010)
Tophat	Seed and extend	Paired-end	(Trapnell, Pachter & Salzberg, 2009)
QPALMA	Seed and extend	Paired-end	(De Bona *et al.*, 2008)
GSNAP	Seed and extend	Paired-end	(Wu & Nacu, 2010)

Transcriptome Reconstruction			
G-Mo.R-se	Genome assisted	Paired-end	(Denoeud *et al.*, 2008)
Cufflinks	Genome assisted	Paired-end	(Trapnell *et al.*, 2010)
Scripture	Genome assisted	Paired-end	(Guttman *et al.*, 2010)
TransAbySS	de novo assembly	Paired-end	(Robertson *et al.*, 2010)
Estimation of isoform-level expression			
ALEXA-seq	relies on uniquely matching reads		(Griffith *et al.*, 2010)
MISO	statistical models for the observed read distribution		(Katz *et al.*, 2010)
Cufflinks	statistical models for the observed read distribution		(Trapnell *et al.*, 2010)
Determination of Differential Gene Expression			
DEGseq	Poisson distribution		(Wang *et al.*, 2009)
EdgeR	Negative binomial distribution / tag-wise dispersion		(Robinson, McCarthy & Smyth, 2010)
DESeq	Negative binomial distribution		(Anders & Huber, 2010)
BaySeq	Negative binomial distribution / Bayesian approach		(Hardcastle & Kelly, 2010)
Identification of Gene Expression Patterns			
Cluster 3.0	Hierarchical clustering		(de Hoon *et al.*, 2004)
The Unscrambler	Principal component analysis		(Wass, 2005)
Current lists of NGS software			
Wikipedia.org: http://en.wikipedia.org/wiki/List_of_sequence_alignment_software Seqanswers.com: http://seqanswers.com/wiki/Software/list Gene ontology tools: http://www.geneontology.org/GO.tools.shtml			

Chapter 2

Bioinformatic Framework for Tag Based Transcrip-tomics – Initial Work



2.1 Primary Data Processing

The Illumina Genome Analyzer (GA) sequencing system produces images files as primary output from a sequencing run, cf. Figure 1-4. The conversion of these image files into sequence files requires bioinformatic hands on time every time a sequencing run has completed. To facilitate and automate the primary data processing an analysis pipeline, which will be described following sections, was made. Subsequently, tag based sequencing data requires additional pre-processing steps, namely sequence tag extraction and annotation. These will be described in the following sections. Finally, the developed methods for sequence filtering, error correction, and tag annotation will be evaluated.

Following a sequencing run on an Illumina Genome Analyzer (GA), initial data analysis is required for generation of sequence data. The initial data analysis illustrated in Figure 2-1 consists of three steps:

- Image analysis; where clusters are located on the image and cluster intensities, positions, and noise estimates are calculated.
- Base calling; where the sequence of bases is read from each cluster, a confidence level for each base is calculated, and read filtering is performed.
- Sequence analysis; where sequence files in FASTQ format are generated. Here, reference alignment and data visualization is also possible.

Although Illumina provides a pipeline for these three steps, it is made up by a collection of command line modules (either Perl or Python scripts or C++ executables), and hence requires a high amount of human input to execute. To enable an automation of the entire data preprocessing and data storage requiring minimal user input, an analysis pipeline was made. The pipeline also includes data preprocessing steps that are unique for tag based transcriptome data. Initially, recorded images were transferred to the analysis pipeline server, where the entire preprocessing was performed. Later, Illumina introduced Real Time Analysis (RTA), enabling automated image analysis and base calling during a sequence run. However, RTA does not facilitate custom settings for sequence filtering, why base calling allowing custom settings is still a part of the pipeline, cf. Figure 2-1. The pipeline can be started immediately after a sequencing run has been initiated, and will automatically be executed when the sequence run is completed. Minimal user input is required to start the pipeline; only three options need to be se set, cf. Figure 2-2. First, the type of run is selected; either single read, paired-end read or single reads containing tag based sequencing data. Choosing the latter enables the additional preprocessing steps for tag based sequencing data. Second, the type of sequence filtering is chosen, either no filtering, passing all reads, or default sequence filtering, hereby enabling chastity filtering in the sequence analysis step. The Chastity per base is defined as the ratio of the brightest intensity over the sum of the brightest and second brightest intensity, where the default setting is 0.6 (Illumina, 2009).

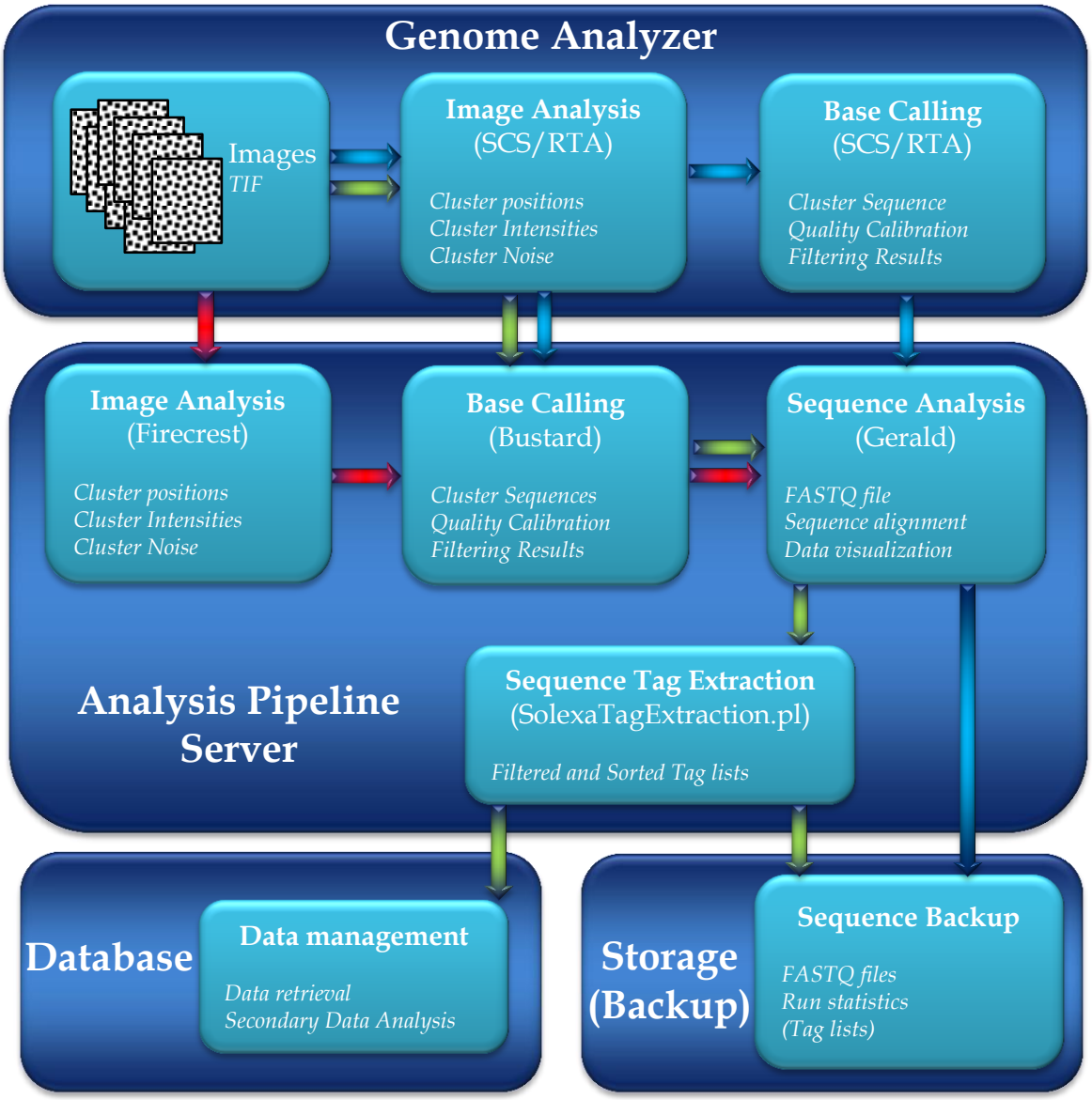


Figure 2-1 Overview of the primary data processing of Illumina sequence data. Before introduction of Real time analysis (RTA), image files were transferred to an analysis server where image analysis, base calling and sequence analysis were performed (red arrows). After introduction of RTA, Image analysis and base calling is performed on the Genome Analyzer analysis computer, only requiring sequence analysis to be performed on the analysis server (blue lines). For tag based sequence data non-default settings for the base calling is used, requiring this to be performed on the analysis server. Following sequence generation tag lists are created and uploaded to a database (green lines). The last step for all types is data backup.

Finally, settings for extraction of tags from the sequence files including an additional optional filtering step are set, where default settings correspond to tags generated with the DeepSAGE method and no additional filtering is performed. The pipeline runs the data preprocessing steps automatically, where after sequence backup and upload of eventual tag libraries to a database is performed.

1)

```
###################### Choose type of run ######################
     (1) Tag based Single read
     (2) Single read (but no tag based sequencing)
     (3) Paired-End

Select 1, 2 or 3
```

2)

```
#################### Set quality filter settings ################
     (1) No filtering (e.g. for tag based sequencing runs)
     (2) Normal chastity filtering
     (Q) Quit program

Select 1, 2 or Q
```

3)

```
################### Settings for tag extraction #################

Default settings = No filtering!
Current settings for tag extraction...

Cut off value (Phred Score): 0
Maximum number of bases allowed below cut off: 17
Length of tags extracted: 17

Change settings:
[Y/N]:
```

4)

```
###################
Genome Analyzer is running...
Total runtime: 5 hours and 34 min
```

Figure 2-2 Screen shots from the analysis pipeline program requiring minimum user input. **1)** Type of run is selected. **2)** Type of sequence filtering is selected. **3)** Settings for sequence tag extraction are selected. **4)** The program has started, now waiting for the sequence run to complete. The analysis automatically starts when the sequence run is complete and all files have been transferred to the analysis server.

Although automation of the primary data processing might seem a trivial thing, it has proven to facilitate productivity in the research group, where it has been employed. Firstly, it has enabled that the initial data processing can be started by the technician starting the sequence run. Secondly, the time demanding human intervention for the bioinformatic analysis has been reduced to merely insuring that the analysis completed successfully, and that the data quality is o.k. Thirdly, it minimizes risk of errors and simplifies troubleshooting, since input parameters are only states once. The described analysis was designed for the Genome AnalyzerII$_x$. Later, a similar analysis pipeline has been developed for the Illumina HiScan and HiSeq sequencing systems. These are currently employed at Klinisk Genetisk Afdeling, Vejle Sygehus[5], and at the Department of Biotechnology, Chemistry and Environmental Engineering at Aalborg University.

Following the initial preprocessing steps, sequence files containing tag based transcriptome data are subjected to automatically extraction of sequence tags, which are counted tabulated and sorted using the sequence barcode key information. The DeepSAGE method uses MmeI as the tagging enzyme and sequences containing a 17-19 bp tag following by adaptor sequence are therefore produced. These are recognized using pattern matching, and the first 17

[5] Personal communication: Annabeth Høegh Petersen: Annabeth.Hogh.Petersen@slb.regionsyddanmark.dk

bp are extracted even though the tags are up to 19 bp to simplify downstream data analysis, cf. Figure 2-3.

A)

```
1)
@ILLUMINA_0000:5:1:28:282#0/1
ACTAGTAATCAGAAACACGGTTGAGATCGTATGCCGTCTTCTGCTTGAAAAAACAACAACACCGAACAGAAC
+ILLUMINA_0000:5:1:28:282#0/1
`a]_ZGa\Y_^[]TPG_aYUGRV_S_U_RTWXWZQFG[XV_QNUX]BBBBBBBBBBBBBBBBBBBBBBBBB
2)
@ILLUMINA_0000:5:1:28:1664#0/1
CATGAGACTTAGACTTCAACGATGGAGATCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAAAAAAAAACAA
+ILLUMINA_0000:5:1:28:1664#0/1
`bbbbbbIb`bb`]bbbbbbTbbb^bSY\_bVa^YXG`]abS]^Y\V_abbb_\BBBBBBBBBBBBBBBBB
3)
@ILLUMINA_0000:5:1:28:1652#0/1
TTAAATACAACATTTTTCGAGCGAGATCGTATGCCGGCTTCTGCCTCAAAAAAAAAAAAAAAAACATAACATA
+ILLUMINA_0000:5:1:28:1652#0/1
``a[`^aaabb``Y_N]a^`^``a^`^^^Q^Z_a\GDURS\SFWHLGTb_R]G]bbbbZBBBBBBBBBBBBB
```

B)

AAT TTA TCC TAG TGT CTC CAA CGG ATG ACA AGC GTT CTT GCG GAC GGA

Figure 2-3 **A)** Sequence tag extraction from an Illumina FASTQ file. Sequence tags (marked in turquoise) are extracted from sequence containing a tag, a valid 3 bp barcode sequence, and the first part of the adaptor sequence (bps used are marked in yellow). Notice that only 17 bps are extracted from the second sequence, even though the tag is 19 bp. If a Phred score filter is used, cf. section 2.2.1, only passing tags with all bases having a score >= 20, the third sequence will be filtered out due to the base marked in red with the quality "N", which equals a Phred score of 14. **B)** Barcode sequences used in the DeepSAGE Solexa protocol. Notice that one sequence error in a barcode cannot create the sequence of another barcode sequence.

Following extraction, sequence tags are counted and tabulated hereby creating a tag list for each sample based on the barcode sequence.

As mentioned by Saha *et al.* reliable tag-to-gene-annotation is a crucial step in order to extract biological knowledge (Saha *et al.*, 2002). To facilitate fast annotation, complying with restraints caused by the SAGE methods (sequence tags must be preceded by a recognition site for the anchoring enzyme) an annotation algorithm was made, and implemented in Perl (*Global-Sagemap.pl*). The recognition site for the anchoring enzyme can be set, facilitating annotation of tag lists created using different SAGE methods. In the case of DeepSAGE, the recognition site is set to "CATG". Tags are extracted from a sequence collection (e.g. a set of mRNA transcripts, ESTs or even a whole genome sequence) and saved to three "virtual tag lists" along with enclosed functional annotation, cf. Figure 2-4. Tags matching multiple genes will retrieve multiple annotations. The three virtual tag lists contain the most 3' 17 bp tag downstream of a CATG site, the most 5' 17 bp reversed tag upstream of a CATG site, and all 17 bp tags both up- and downstream of all CATG sites, respectively. Tag lists are then matched against the three virtual tag lists hierarchically. First, tags are matched against the most 3' 17 bp tags, since these in theory are the ones generated by SAGE methods. Secondly, tags are matched against the most 5' 17 bp reversed tag upstream of a CATG site because the orientation of e.g. ESTs in a sequence database can be unknown possibly leading to an inversion. Therefore, these tags have the potential of in fact being the most 3' tag. Finally, tags are matched against all internal tags; here taking several factors into account. The longest transcript variant of a gene is often the one represented in sequence databases (e.g. RefSeq (Maglott *et al.*, 2000)), but is not necessarily the most abundant. Here, an internal tag would be the representative tag of a gene. Furthermore, internal tags can also be created by incomplete diges-

tion of the anchoring enzyme, or transcripts having multiple PolyA sites, which have been reported earlier (Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008). This will also be illustrated later in the current thesis, cf. Figure 5-7 in section 5.2.2. Tags are first annotated using perfect matching against the virtual tag lists, and secondly, if chosen, annotated allowing a single nucleotide difference. This feature facilitates e.g. annotation of tags originating from organisms with a high SNP frequency, which is not represented in the model sequence collection.

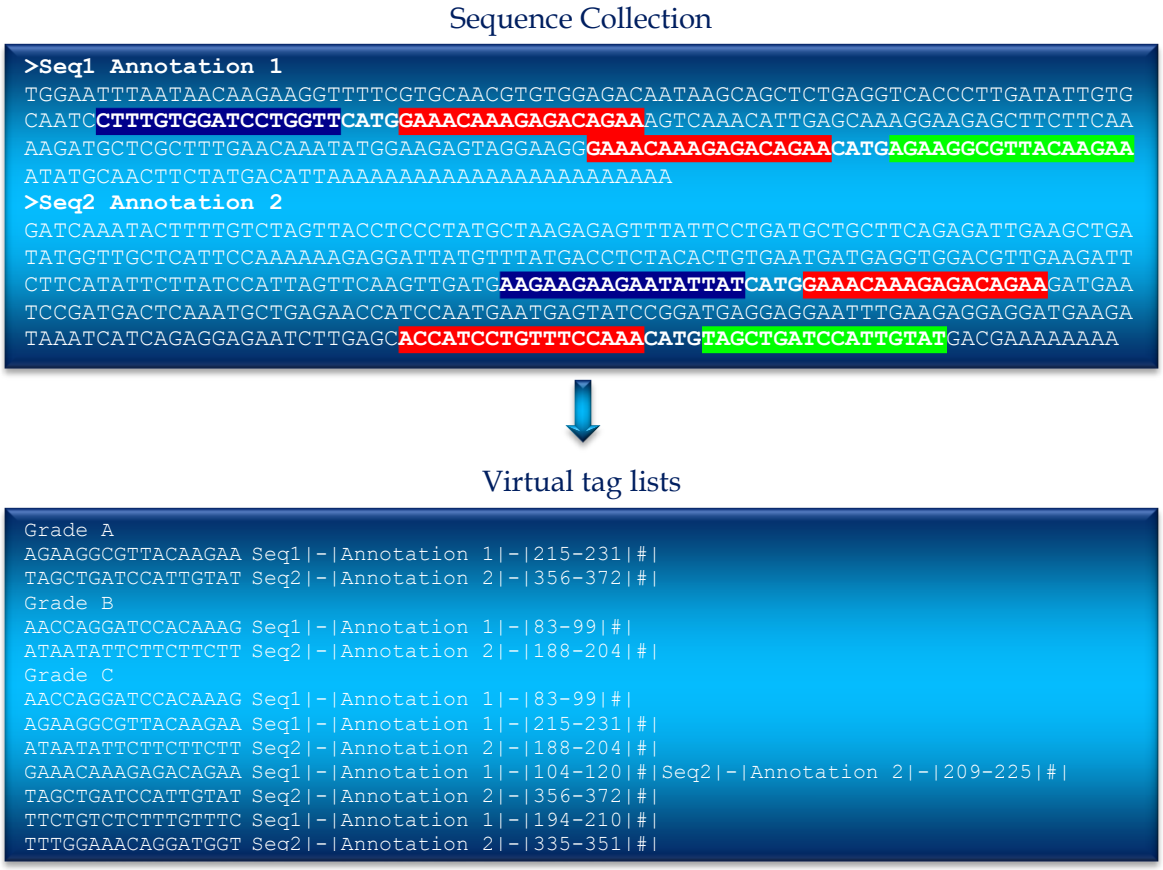


Figure 2-4 Extraction of tags into virtual tag lists. Tags are extracted from a sequence collection (e.g. a set of mRNA transcripts, ESTs or even a whole genome sequence) and saved to three "virtual tag lists" along with enclosed functional annotation, cf. Figure 2-4. Tags matching multiple genes will have multiple annotations. The three virtual tag lists contain the most 3' 17 bp tag downstream of a CATG site (marked in green), the most 5' 17 bp tag upstream of a CATG site (marked in blue), and all 17 bp tags both up- and downstream of all CATG sites (marked in red).

2.2 Evaluation of Primary Data Processing

Methods for sequence filtering, sequence error, and tag annotation are steps in the data pre-processing, which affects the processed data differently depending on chosen settings. An evaluation of how the different steps affect the data will be described in the following sections.

2.2.1 Evaluation of Sequence Filtering and Sequence Error Correction

As mentioned, an average read accuracy of 99 % for each base only leads to an 84 % change of producing a 17 bp error free tag. A large amount of low abundance error tags are therefore produced. This complicates data analysis, because another variable is added to the data set each time a new error tag is created. Therefore, filtering out or correcting for these errors can facilitate easier downstream data analysis, by reducing the data complexity. There are several strategies for filtering out error tags. The filtering can be performed using sequence quality information, based on the hypothesis that error tags arise from low quality sequences. Here, the standard chastity filter from the Illumina pipeline can be employed. Moreover, a simple Phred Score Quality (PSQ) filter was developed as a part of the tag extraction. Here, a minimum base quality score (converted into the Phred score format (Ewing *et al.*, 1998)) and a maximum allowed number of bases below the cutoff base quality score are used as filtering criteria. Although 16 % of all generated tags are generated due to sequencing errors (with an average read accuracy of 99 %), the change of generating the same tag by sequencing errors multiple times is small. Therefore, filtering can also simply be based on tag counts, filtering out all low abundance tag at a user defined cutoff value. Moreover, error correction using the SAGEscreen algorithm (Akmaev & Wang, 2004) can be employed. This has the advantage that error tags are detected and their counts are subsequently added to the true tag from which they most likely originate from, and hereby maintaining the information of these tags. An evaluation of how the different filtering strategies and sequence error correction affect the data structure is described in the following.

2.2.1.1 Methods

The data analysis was performed on 12 tag libraries all originating from *S. tuberosum* leaf samples. The 12 libraries were sequenced in a single lane on a Genome AnalyzerII$_x$. Image analysis and base calling were performed using the GAPipeline version 1.5.1 software with default filtering settings (Chastity threshold of 0.6, on the first 25 cycles) and with no filtering, respectively. Sequence tag extraction was performed using *SolexaTagExtraction.pl* with and without PSQ filtering, hereby creating four data sets in total. All 12 libraries in each set were combined into a single library using *CombineLibraryCounts.pl*. The resulting four libraries were subjected to SAGEscreen error correction using the algorithm implemented in the CLC Genomics Workbench version 4.8. Tags were annotated using a sequence collection consisting of predicted mRNA transcripts in the genome sequence of version 3.4 of *S. tuberosum* Group Phureja DM1-3 516R44[6] (PGSC mRNAs), Tentative consensus sequences from

[6] available at: http://potatogenomics.plantbiology.msu.edu/index.html

TIGR *S. tuberosum* Gene Indices version 13 (Quackenbush *et al.*, 2000)[7] (TCs), and PlantGDB-assembled unique transcripts from PlantGDB version 157a (Dong, Schlueter & Brendel, 2004)[8] (PUTs). To reduce redundancy, TC and PUT sequences were compared to PGSC mRNAs using blastN (Camacho *et al.*, 2009; Altschul *et al.*, 1990). TCs and PUTs Sequences with a significant hit (E-value < $1*10^{-30}$) adopted the ID of the matched PGSC mRNA sequence, hereby creating a sequence collecting with multiple sequences with the same ID.

2.2.1.2 Results

PSQ and chastity filtering, SAGEscreen error correction, singleton removal, and combinations hereof were applied on either raw sequence data or tag lists. The effect of these was investigated in regards to data complexity (the total number of unique tags) and data preservation (remaining tag counts). The results are summarized in Table 2-1, Table 2-2, and Table 2-3. Under the assumption that a "true" error-free tag originating from a mRNA transcript is far more likely to have a match in a sequence data base than a tag originating from a sequence error, the quality of filtering and error correction can be evaluated by investigating the effect of these on tags that can be annotated contra tags that cannot.

Table 2-1 Total Number of unique tags and total tag count for 12 tag libraries originating from a single sequence file. Statistics are shown for all tags, and tags with an annotation (see section 2.2.1.1 for details). Percentages are given in relation to the number of unique tags and the total tag count of the original unfiltered sample, respectively. SAGEscreen enables annotation of additional tags, why percentages can exceed 100 %. PSQ = Phred Score Quality filter, SS = SAGEscreen error correction, and SR = Singleton removal.

Sequence filter	Error correction	Unique tags		Tag Count	
All tags					
None	None	1,746,171	(100 %)	14,193,217	(100.0 %)
None	SS	631,238	(36.1 %)	14,193,217	(100.0 %)
PSQ	none	660,611	(37.8 %)	10,987,789	(77.4 %)
PSQ	SS	195,894	(11.2 %)	10,987,789	(77.4 %)
Chastity	None	677,941	(38.8 %)	11,771,085	(82.9 %)
Chastity	SS	199,990	(11.5 %)	11,771,085	(82.9 %)
Chastity + PSQ	none	576,178	(33.0 %)	10,839,945	(76.4 %)
Chastity + PSQ	SS	186,933	(10.7 %)	10,839,945	(76.4 %)
None	SR	409,275	(23.4 %)	12,856,321	(90.6 %)
None	SS + SR	221,145	(12.7 %)	13,783,125	(97.1 %)
Annotated tags					
None	None	76,165	(100.0 %)	8,616,730	(100 %)
None	SS	73,455	(96.4 %)	9,625,819	(112 %)
PSQ	none	74,722	(98.1 %)	7,487,777	(86.9 %)
PSQ	SS	72,736	(95.5 %)	8,019,347	(93.1 %)
Chastity	None	75,033	(98.5 %)	8,053,943	(93.5 %)
Chastity	SS	73,036	(95.9 %)	8,604,954	(99.9 %)
Chastity + PSQ	none	74,395	(97.7 %)	7,487,767	(86.9 %)
Chastity + PSQ	SS	72,620	(95.3 %)	7,933,961	(92.1 %)
None	SR	71,259	(93.6 %)	8,611,824	(99.9 %)
None	SS+SR	70,601	(92.7 %)	9,622,965	(112 %)

Filtering methods, SAGEscreen error correction, and singleton removal all considerably reduce the data complexity (between 61.2 % and 89.3 % of all unique tags are removed, cf. Table 1-1). Using sequence filtering, this reduction in data complexity is at the expense of some data loss, since 17.1 % and 22.6 % of the data is discarded using chastity and PSQ filtering, respectively. When comparing the sequence filtering methods with simple singleton remov-

[7] available at: http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=potato
[8] available at: http://www.plantgdb.org/download/download.php?dir=/Sequence/ESTcontig//Solanum_tuberosum/previous_version/157a

al, it is worth noticing that singleton removal on its own has the highest level of data complexity reduction with the lowest amount of data loss (only 9.4 % of the data is discarded). Under the described assumption, high quality sequence filtering should preferably filter out sequences with tags that cannot be annotated, while these are less likely to originate from "true" tags. However, both sequence filtering methods discard a great deal of annotated tags (23.2 % and 35.2 % of filtered sequences contain annotated tags using chastity and PSQ filtering, respectively; cf. Table 2-2). This leads to a 6.5 % and 13.1 % reduction in the total tag count of annotated tags using chastity and PSQ filtering, respectively. Simple singleton removal outperform sequence filtering; only 0.4 % of the total amount of filtered tags can be annotated leading to nearly no reduction in the total count of annotated tags (99.9% remain, cf. Table 2-1).

Table 2-2 Effect of filtering methods. Number of total filtered tags, annotated tags and not annotated tags using PSQ, chastity filtering or both. Percentages are given in relation to the total number of filtered tags. PSQ = Phred Score Quality filter.

Sequence filter	Filtered tags	Annotated tags		Not annotated tags	
Chastity	2,422,132	562,787	(23.2 %)	1,859,345	(76.8 %)
PSQ	3,205,428	1,128,953	(35.2 %)	2,076,475	(64.8 %)
Chastity + PSQ	3,353,272	1,128,963	(33.7 %)	2,224,309	(66.3 %)

SAGEscreen error correction has the advantage, that it causes no data loss. It has a slightly better ability to reduce the data complexity compared to the sequence filtering methods, cf. Table 2-1. However, a major advantage is that SAGEscreen is an error correction method, and not a filtering. This leads to a higher fraction of the total tag count that can be annotated (a 12 % increase in the total annotated tag count, cf. Table 2-1, which correspond to 7.1 % of the entire data set, cf. Table 2-3). There is a very low amount of error corrected tags that were annotated prior to error correction (0.3 % of the total tag count of error corrected tags, cf. Table 2-3). These tags are potentially true tags that should not have been error corrected. However, among these 75.7 % and 88.5 % of all different tags were only observed once or less than three times, respectively. This indicates that these tags are likely to be error tags of more abundant true tags. In total, 11.2 % of the entire data set was error corrected, indicating that the per base sequence error frequency of the data is 0.7 %. This is similar to the raw accuracy for the Genome AnalyzerIIx reported by Illumina (Cappelletti, 2009).

Table 2-3 Effect of SAGEscreen error correction. Tag count of annotated tags, not annotated tags, and all tags corrected using SAGEscreen error correction. The increase in tags that can be annotated after SAGEscreen error correction is given in the bottom.

SAGEscreen corrected tags	Total tag count	Fraction of Corrected tags [%]	Fraction of total tag count [%]
Tags with annotation	4,871	0.30 %	0.03 %
Tag with no annotation	1,595,313	99.7 %	11.2 %
Total	1,600,184	100 %	11.3 %
Additional annotated tags	1,009,089	63.3 %	7.1 %

2.2.1.3 Conclusions

Based on the above analysis of different pre-processing methods including sequence filtering, error correction, and singleton removal, it is concluded that performing sequence filtering (either PSQ or chastity) leads to the highest degree of data loss and the lowest degree of data complexity reduction. Furthermore, the filtering is less specific towards tags that cannot

be annotated (and hence have a higher chance of being an error tag or at least a non-informative tag), why this should be avoided. This can be explained by the fact that sequence filtering relies on the sequence quality. Although a base has a low quality score, the chance that is determined correctly is still high. In the above analysis a Phred score cutoff value of 20 was used, only allowing a 1 % chance of a sequence error to occur. This leads to data loss, since many "true" tags are filtered out. Furthermore, the sequence error can arise during amplification of the sequence library and hence not be reflected in the sequence quality. On the other hand, SAGEscreen error correction results in a higher degree of data complexity reduction without data loss. Moreover, the method is specific towards error tags correcting non-annotated tags with nearly 100 % specificity, leading to an increase in the amount of interpretable data, where the sequence filtering methods on the other hand led to data loss. The original SAGEscreen software is freely available for academia (Akmaev & Wang, 2004). However, this was developed for smaller SAGE libraries, and is not well-suited for large tag based transcriptome data sets based on NGS. As a consequence of this analysis, the SAGEscreen method was reinvented and implemented in the CLC Genomics workbench in a much more efficient implementation. CLC Genomics workbench is a commercial software package and hence not freely available. If this is unavailable, simple singleton removal results in at least the same degree of data complexity reduction, however without the benefit of increase in interpretable data. The optimal pre-processing method is concluded to be a combination of SAGEscreen error correction and singleton removal.

2.2.2 Evaluation of Sequence Tag Annotation Specificity

As mentioned, tags can be annotated allowing a single mismatch to facilitate e.g. annotation of tags originating from species with a high SNP frequency among different genotypes. However, when allowing a single mismatch there is a potential risk that annotation will become less specific, i.e. more tags will be unambiguously matched to the sequence collection. If tag matching was totally random, there would be a 55 % and ~0 % chance that 100,000 17 nt tags all were unique using perfect matching and allowing a single nucleotide mismatch. This improves to 65 % and 11 % using 21 nt tags, which is the sequence length produced using the SuperSAGE method (Matsumura *et al.*, 2010). However, as previously shown, cf. section 1.2.3.3, tags are not randomly selected from "tag space" (due to the forces of evolution), and hence are far less likely all to be unique. An evaluation of allowing a single nucleotide mismatch is described in the following.

2.2.2.1 Methods

Non-redundant collections of mRNA transcripts from *H. sapiens*[9], *A. thaliana*[10], *and S. tuberosum*[11] were downloaded. For the *H. sapiens* data set, one transcript from each protein coding gene was subsequently extracted using *NonRedundantRefSeq.pl.* Using *GlobalSagemap-V30.pl,* virtual tag lists for each sequence collection were made using standard settings for analysis of DeepSAGE tags.

[9] RefSeq mRNA sequences (version 12/092011). Found at: ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/
[10] Representative cDNA gene models (TAIR10 (Dec 2010) release). Found at: ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR10_blastsets/.
[11] Potato: Representative transcript models from the Genome Annotation v3.4 Found at: http://potatogenomics.plantbiology.msu.edu/index.html

To investigate the theoretically maximum loss of specificity in tag matching, all theoretical 17 nt DeepSAGE tags from non-redundant sequence collections of mRNA transcripts from *A. thaliana* (33,602 sequences), *H. sapiens* (19,754), and *S. tuberosum* (39,031 sequences) were extracted. The distribution of number of genes matched per tags was subsequently calculated for tag annotation for all theoretical tags using perfect matching and allowing a single nucleotide mismatch, respectively.

To investigate the actual loss of specificity in tag matching a combined library originating from 12 tag libraries all originating from *S. tuberosum* leaf samples was created not using sequence filtering, but using SAGEscreen error correction and singleton removal, cf. section 2.2.1.1 for details. The library was annotated against the *S. tuberosum* mRNA transcript sequence collection, using the hierarchical scheme described in section 2.1 allowing a single mismatch. Finally, the distributions of number of genes matched using perfect matching and allowing a single mismatch was calculated.

2.2.2.2 Results

The theoretically maximum loss of specificity in tag matching was investigated for non-redundant sequence collections of mRNA transcripts from *A. thaliana*, *H. sapiens*, and *S. tuberosum*. The fraction of uniquely matching tags was calculated for all theoretical DeepSAGE tags using perfect matching and matching with one nt mismatch, cf. Figure 2-5.

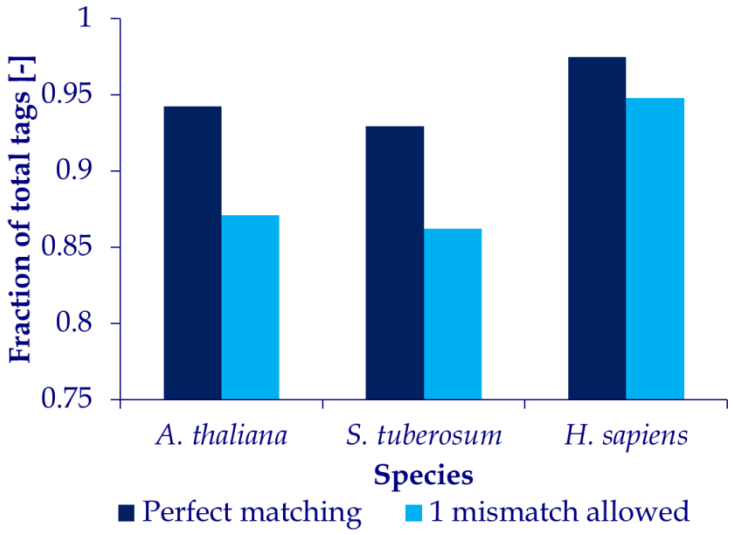


Figure 2-5 Theoretical tag matching specificity indicated as the fraction of uniquely matching sequence tags using perfect matching allowing a single nucleotide mismatch, respectively. All theoretical 17 nt DeepSAGE tags from non-redundant sequence collections of mRNA transcripts from *A. thaliana* (33,602 sequences), *H. sapiens* (19,754), and *S. tuberosum* (39,031 sequences) were extracted and the distribution of the number of genes matched per tags was subsequently calculated for tag annotation of all tags using perfect matching and allowing a single nucleotide mismatch (see section 2.2.2.1 for details). Notice the limited Y-axis.

There is a 7-8 % decrease in the fraction of uniquely matching tags in the two largest sequence collections (*A. thaliana* with 33,602 sequences, and *S. tuberosum* with 39,031 sequences), cf. Figure 2-5, but only a 3 % decrease of the smaller collection of *H. sapiens* mRNA transcripts that contains 19,754 sequences. This theoretically maximum loss in matching specificity should be taken into account, when deciding whether or not to allow a mismatch. The loss in specificity should be related to the actual gain of informative tags, i.e. gain of tags that can be annotated. This was investigated by annotating a combined library from 12 DeepSAGE

libraries using perfect matching and subsequently allowing 1 nt mismatch, cf. Table 2-4 and Figure 2-6.

Table 2-4 Percentages of tags that can be annotated using perfect matching and subsequently matching allowing 1 nt mismatch. Numbers are shown for the most 3' tag, the most 5' reverse complemented (RV) tag, and internal tags.

	% of unique tags	% of total tag count
Perfect Matching		
Most 3'	5.4 %	30 %
Most 5' RV	1.5 %	1.5 %
Internal	16 %	18 %
Total annotated	23 %	50 %
Additional matching with 1 Mismatch Allowed		
Most 3'	9.0 %	7.2 %
Most 5' RV	0.8 %	1.5 %
Internal	9.9 %	5.0 %
Total annotated	42 %	64 %

As seen in Table 2-4 there is a 28 % increase (from 50 % to 64 % of the total tag count) of tags that can be annotated. Moreover, when comparing the specificity of tag matching, there is hardly any difference between the specificity of tags matched perfectly and tags matched allowing one nt. difference, cf. Figure 2-6, although the theoretical specificity should be lower, cf. Figure 2-5. This can be caused by the hierarchical annotation scheme, where annotation using perfect matching is attempted first, only followed by annotation allowing a mismatch if annotation using perfect matching failed. The scheme ensures that uniquely matching tags using perfect matching, which become ambiguous allowing a single mismatch, will be uniquely matched. This can be particularly useful in cases where several homologues sequences are found in the sequence collection.

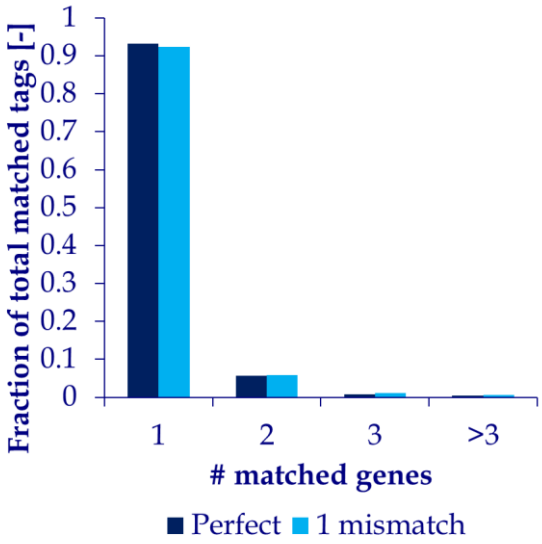


Figure 2-6 Actual tag matching specificity indicated as the fraction of sequence tags matching 1, 2, 3, or >3 gens using perfect matching and subsequently allowing a single nucleotide mismatch if annotation is not possible using perfect matching. A combined library originating from 12 tag libraries all originating from *S. tuberosum* leaf samples was annotated against the *S. tuberosum* mRNA transcript sequence collection, using the hierarchical scheme described in section 2.1 (see text for more details).

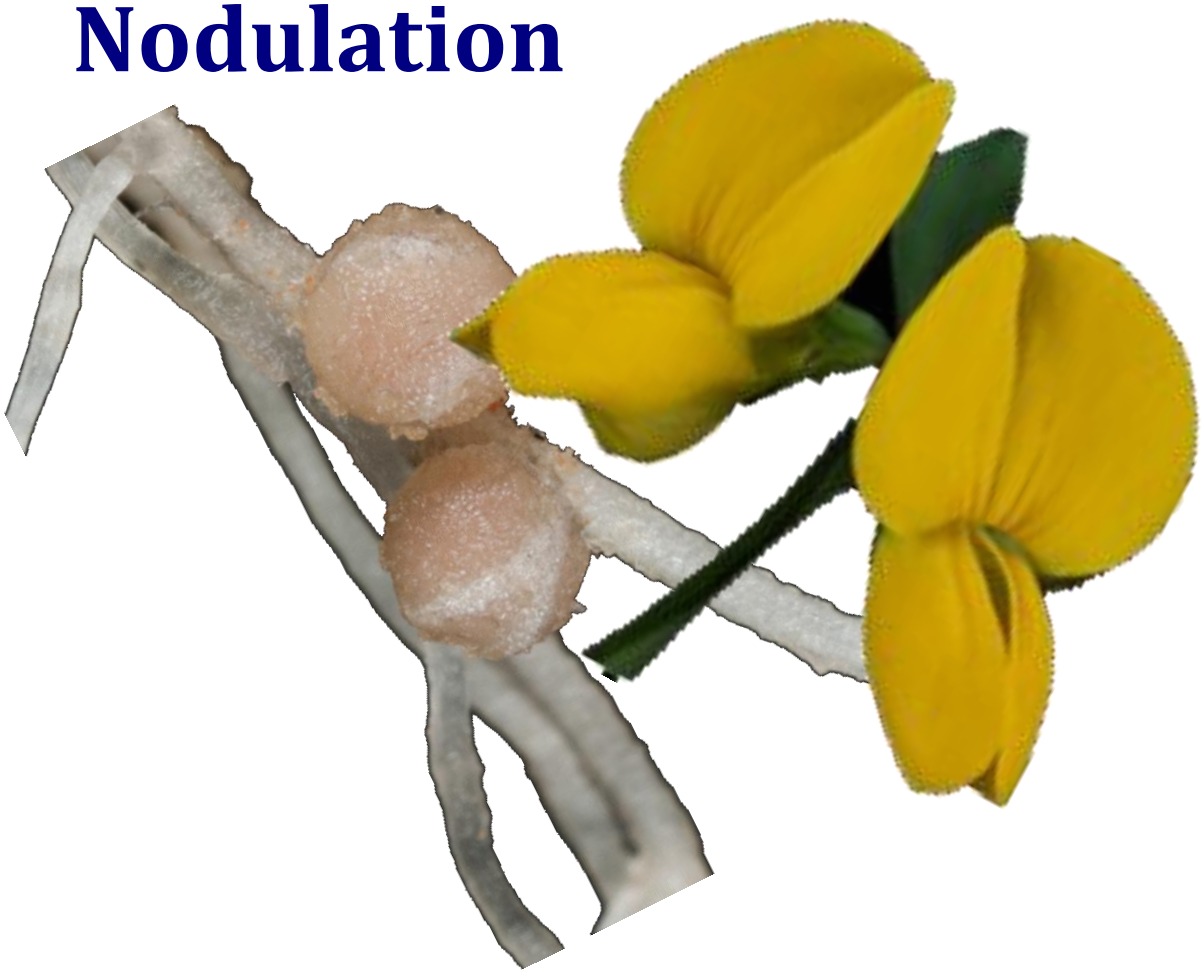
2.2.2.3 Discussion and Conclusions

Although there is a theoretical possibility for a decrease in the specificity of tag annotation allowing a single nt. mismatch, no decrease can be detected when performing the annotation on real data sets. Moreover, allowing a single nt. mismatch in the annotation significantly increases the amount of tags that can be annotated. This observed increase is most likely not an effect of additional annotation of tags generated by sequence errors, since the data was SAGEscreen error corrected prior to annotation. The increase is more likely to be an effect of additional annotation of tags containing a SNP compared to the sequence found in the database. The fact that *S. tuberosum* is highly heterozygous between genotypes (The Potato Genome Sequencing Consortium *et al.*, 2011) supports this. Therefore annotation allowing a single nt. mismatch is recommended. Although the most 3' tag in theory should be the representative tag of a transcript, the use of internal tags is recommend. As mentioned, this is not always the case due to factors such as incomplete digestion of the anchoring enzyme, or multiple PolyA sites and alternative splicing of transcripts. The use of internal tags will to some extent account for these, and it has e.g. been shown by Robinson *et al.*, that the amount of unique tags with an annotation increases from 55 % using only the most 3' tag to 71 % of all unique tags including internal tags in an *A. thaliana* data set (Robinson *et al.*, 2004).

The reliability of tag to gene mapping has been investigated by several others (see review by Wang for details (Wang, 2007)). Although the annotation method can affect the efficiency of tag to gene annotation, the presence of a high quality collection of transcripts is vital. Construction of such a sequence collection is greatly facilitated by the presence of a genome sequence, which will be illustrated in the following chapters.

Chapter 3

Transcriptome Analysis of *Lotus japonicus* During Nodulation



3.1 Introduction to the Transcriptome Analysis of *Lotus japonicus* During Nodulation

In 2007, the first larger transcriptome study, which was to be analyzed using the DeepSAGE technology, was initiated at Aalborg University (AAU). The study was an investigation of the transcriptome from relevant *Lotus japonicus* (*L. japonicus*) tissues in response to the symbiotic interaction with the soil bacteria *Mezorhizobium loti* (*M. loti*). During this interaction, named nodulation, *L. japonicus* develops root derived organs known as nodules. These organs provide a suitable environment for the bacterial enzyme nitrogenase (Oldroyd, 2007), which catalyzes the agricultural important process nitrogen fixation. The project is a joint effort between Århus University (AU) and AAU. *L. Japonicus* plants were grown in a greenhouse of Århus University. Sampling and DeepSAGE library preparation were performed by Annabeth H. Pedersen at Aalborg University and described in (Petersen, 2008). As a part of the current thesis, the transcriptome data analysis was performed. Parts of the data analysis and data interpretation have been performed in close collaboration with Stig Uggerhøj Andersen, PhD and Elena Simona Radutoiu, PhD. Because the data set was the first larger data set produced at AAU, it was meant to serve as a pilot study for future large scale tag based transcriptome studies, for example the study described later in section 4.1.2. Several different data analysis methods were attempted, some of which later implemented in standard data analysis pipelines. However, following the fast development of bioinformatic software within the last three years, some methods developed for the current project were later substituted. Since it would be out of the scope of the current thesis to describe all analyses performed, only analyses used to obtain the status of the study of the time of writing or analyses that highlight central aspects of DeepSAGE transcriptome data will be described.

3.1.1 Legumes and their Agricultural Importance

The *Fabaceae* family commonly known as legumes is second only to the *Gramineae* family (grasses) in importance in regards to human of food, livestock fodder, and raw materials for industry (Graham & Vance, 2003). Contributing to 65 % of the world's total production of grain legumes soybean (*Glycine max*) is by far the most agricultural important legume with most of the production used for either oil extraction or livestock fodder, cf. Table 3-2 (Wang *et al.*, 2003). As source for human foods bean (*Phaseolus vulgaris*), pea (*Pisum sativum*), chickpea (*Cicer arietinum*), broad bean (*Vicia faba*), pigeon pea (*Cajanus cajan*), cowpea (*Vigna unguiculata*), and lentils are the most important legumes, cf. Table 3-1.

Table 3-1 World's production of legumes in 2010. Source: (FAOSTAT, 2011a)

Legume	Production [Ton]
Soybeans	2,621,577,298
Alfalfa for forage and silage	254,254,327
Peanuts, with shell	37,665,245
Beans, green	17,662,028
Peas, green	15,073,796
Chick peas	10,943,281
Cow peas, dry	5,568,383
Lentils	4,641,139
Broad beans, horse beans, dry	4,316,371
Pigeon peas	3,680,314

Besides from oil production where legumes (mostly soybean and peanuts) contribute to more than 30 % of the world's total production, cf. Table 3-2, legumes have a wide range of uses; e.g. for bread and snacks as flour, in milks, yogurt, and infant formula in liquid form, and in the preparation of biodegradable plastics (Graham & Vance, 2003).

Table 3-2 World production of plant oils in 2010. Legume types are marked in bold. Source: (FAOSTAT, 2011b)

Source	Production [ton]	
Palm	45,097,422	(31 %)
Soybean	39,762,356	(27 %)
Rapeseed	22,596,247	(15 %)
Sunflower	12,629,071	(8.6 %)
Palm kernel	5,647,422	(3.9 %)
Peanut	5,129,196	(3.5 %)
Cottonseed	4,621,393	(3.2 %)
Coconut (copra)	3,497,564	(2.4 %)
Olive, virgin	3,269,249	(2.2 %)
Maize	2,312,771	(1.6 %)
Sesame	977,215	(0.7 %)
Linseed	613,619	(0.4 %)
Safflower	131,621	(0.1 %)

A highly significant factor in agricultural production of legumes is their ability to fixate nitrogen (N) in symbiosis with soil bacteria (Oldroyd *et al.*, 2011). N is widely spread in nature but the biologically active forms of N are often the limiting factor in plant growth (Oldroyd *et al.*, 2011). Industrial N fixation to produce artificial fertilizer accounts for ~ 50 % of fossil fuel usage in agriculture (Oldroyd, 2007). It is estimated that 40-60 million tons of N are fixated by cultivated legumes annually; representing a value of ~ 10 billion US$ in fertilizer (Udvardi *et al.*, 2005). Non-legume plant model species such as *Arabidopsis thaliana* lack the ability to provide insights to symbiotic nitrogen fixation, and therefore cannot be used easily e.g. in legume breeding programs (Udvardi *et al.*, 2005). However, agricultural legumes are also relatively poor model systems for genetics and genomics research, due to the often large genome sizes and tetraploidity of cultivated legume genomes, and difficulties in regards to plant transformation and regeneration (Udvardi *et al.*, 2005). Due to this, two species, *L. japonicus* (Handberg & Stougaard, 1992) and *Medicago truncatula* (*M. truncatula*) (Barker *et al.*, 1990), have been chosen as model plants for legume research.

In regards to N fixation, the main difference between the two model plants is that *L. japonicus* develops determinate nodules arising from the central cortex with a transient meristem, whereas *M. truncatula* develops indeterminate nodules arising from inner cortical cells adjacent to the endodermis with a tip-growing meristem (Oldroyd *et al.*, 2011).

3.1.2 The Model Legume Plant *Lotus Japonicus*

L. japonicus, which is depicted in cf. Figure 3-1, was first recognized at Kyoto in Japan centuries ago. Its natural habitat is in East- and Central-Asia (Marquez, 2005). Many ecotypes of *L. japonicus* can be found living in its natural habitat. Among these are the ecotypes Gifu (accession B-129) and Miyakojima (MG-20). Gifu was originally collected in the 1950s by Professor Isao Hirayoshi on a riverbank in the town of Gifu and later used by Kurt Handberg and Jens Stougaard to establish *L. japonicus* as a model plant for legume research (Marquez, 2005; Handberg & Stougaard, 1992).

Figure 3-1 Anatomy of *L. japonicus* **(A).** Flowers **(B)** and root nodules, housing rhizobial bacteria, on *L. japonicus* roots **(C)** are depicted. Picture Sources: A: (CBM S.c.r.l., 2007), B: (LMU, 2012), C: (John Innes Centre, 2011).

As a legume model plant for legume research, *L. japonicus* has several advantages such as a small plant size, large seed set, short life cycle (~ three months), high transformability self-pollinating proliferation (Marquez, 2005). Moreover, it has a small (~ 470 MB) and simple genome (diploid 2n=12) (Sato *et al.*, 2008; Marquez, 2005). In regards to bacterial symbiosis, both *M. loti* and *Bradyrhizobium* sp. can nodulate *L. japonicus*. However, *Bradyrhizobium* sp. only induces ineffective nodules (Marquez, 2005). The ecotype Miyakojima found on the island of the same name has been crossed with Gifu to establish populations for map-based cloning, due to a high degree of polymorphisms (Marquez, 2005; Kawaguchi, 2000). Moreover, MG-20 was the ecotype used for the sequencing of the *L. japonicus* genome, which was published in 2008 (Sato *et al.*, 2008).

3.1.3 Nodule development and Function in *Lotus japonicus*

Legume N-fixation requires both bacterial infection and nodule organogenesis and these needs to be both spatial and temporal coordinated (Oldroyd *et al.*, 2011), cf. Figure 3-2 panel A. Both processes require initial plant recognition of bacterial produced signaling molecules, named nodulation factors (NF), cf. Figure 3-2 panel B. Nod factors, which in many ways act like plant hormones, consist of a β 1–4-linked N-acetylglucosamine residues backbone with an N-linked fatty acid moiety attached to the non-reducing terminal sugar (Oldroyd & Downie, 2008). Several additional modifications to this basic structure can occur, and these differ between species of *Rhizobia* (Oldroyd & Downie, 2008). The plant signaling receptors are likely to be membrane bound receptor-like kinases with N-acetylglucosamine–binding lysin motifs (LysM) (Limpens *et al.*, 2003; Madsen *et al.*, 2003; Radutoiu *et al.*, 2003). In *L. japonicus*, the "Nod-factor receptor 1" (NFR1) and NFR5, which have extracellular LysM resembling domains have been shown to be required for the early recognition of Nod factors, and hence are likely to be the signaling receptors (Madsen *et al.*, 2003; Radutoiu *et al.*, 2003).

The signaling pathway for nodule organogenesis is continued via calcium oscillations in the nuclear region that are activated by Nod factor recognition (Ehrhardt, Wais & Long, 1996). Several other proteins are also required for activation of this signaling pathway. Among these are "symbiosis receptor-like kinase" (SYMRK) (Stracke *et al.*, 2002), components of the nuclear pore, such as NENA (Groth *et al.*, 2010), Nucleoporin133 (NUP133) (Kanamori *et al.*, 2006), NUP85 (Saito *et al.*, 2007) , and Castor and Pollux, which are two cation channels located on the nuclear envelope (Charpentier *et al.*, 2008). Downstream signaling of the calcium oscillations involves a calcium and calmodulin dependent protein kinase (CCaMK) (Tirichine *et al.*, 2006; Lévy *et al.*, 2004; Mitra *et al.*, 2004). CCaMK phosphorylates a protein of unknown function (CYCLOPS) (Yano *et al.*, 2008). Several transcription factors that regulate the gene expression downstream of CCaMK and CYCLOPS activation have been identified In *L. Japonicus*. These include "nodulation signaling pathway 1" (NSP1), NSP2 (Heckmann *et al.*, 2006; Kaló *et al.*, 2005), and "nodule inception" (NIN) (Schauser *et al.*, 1999). Furthermore, "ERF required for nodulation 1" (ERN1) have been found to be essential in *M.* truncatula (Middleton *et al.*, 2007), and two homologues genes also belonging to the AP2-EREBP transcription factor family, (ERN2 and ERN3) have been identified as trans-acting factors that regulate the expression of an early nodulin gene, *ENOD11* (Andriankaja *et al.*, 2007; Middleton *et al.*, 2007). In *L. japonicus*, the gene *LjERF1* also encoding an AP2-EREBP transcription factor has been found to be a positive regulator of nodulation (Asamizu *et al.*, 2008). Downstream of calcium oscillations, it has been shown that cytokinin and auxin signaling in the inner/mid cortex is necessary for further development of nodule organogenesis. Here, the histidine kinase gene LHK1 encoding a cytokinin receptor is essential (Murray *et al.*, 2007; Tirichine *et al.*, 2007). Oldroyd *et al.* concluded that localized cytokinin signaling in the root cortex and pericycle, leads to a localized suppression of polar auxin transport and down regulation of a auxin inducible promoter GH3 indicating that low auxin levels are associated with initiation of the nodule tissue development (Oldroyd *et al.*, 2011; Takanashi, Sugiyama & Yazaki, 2011; Grunewald *et al.*, 2009; Pacios-Bras *et al.*, 2003).

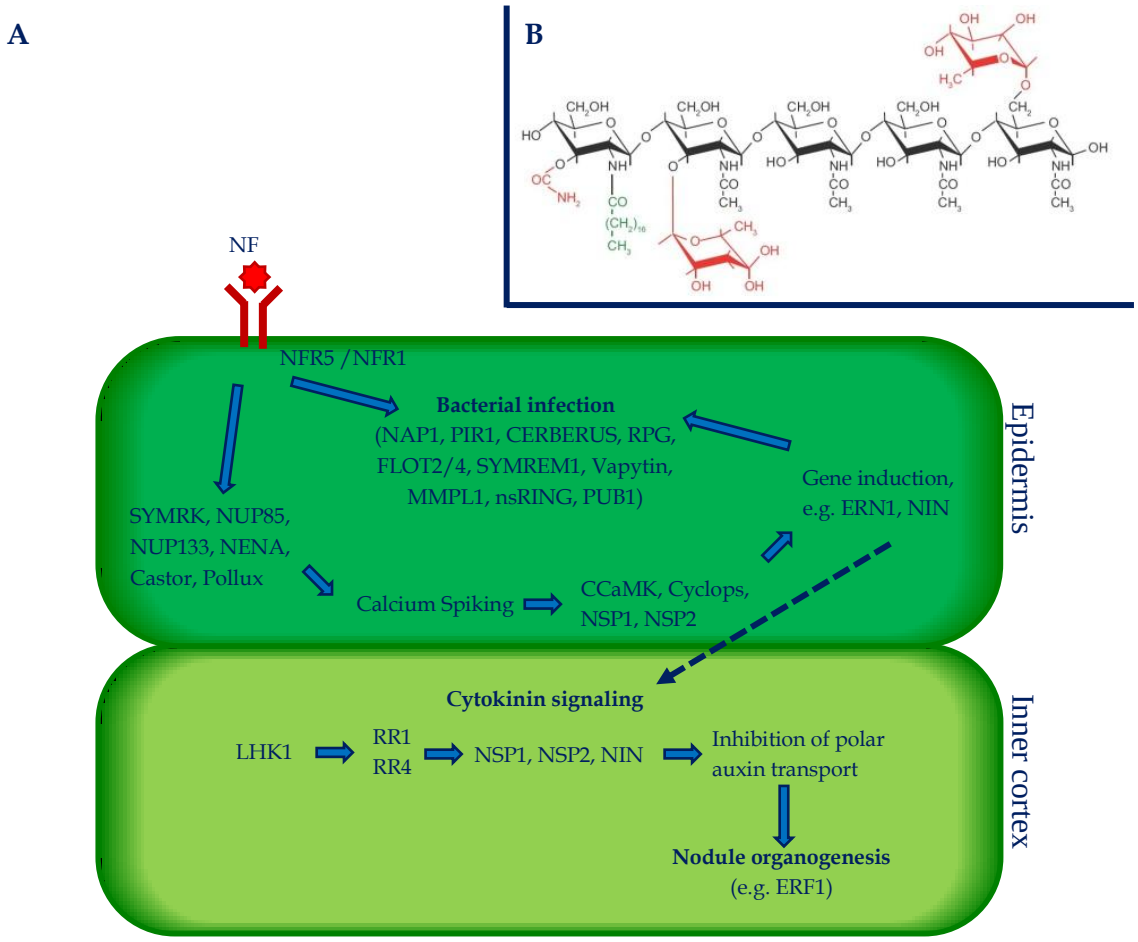


Figure 3-2 **A)** Gene signaling pathways for nodule organogenesis and bacterial root infection. Epidermal cells perceive bacterial Nod factors (NF) through receptor-like kinases causing calcium spiking via a suite of proteins. The calcium oscillation signal is perceived by several genes (e.g. CCaMK and NIN) inducing gene expression, which facilitates the downstream signaling pathway. This initiates bacterial infection at the epidermis and the promotion of cell division in the cortex via an unknown signal (dotted line). Here, Cytokinin signaling induces suppression of polar auxin transport promoting nodule organogenesis. **B)** A representative nodulation factor produced by *M. loti* is shown with a backbone of β 1–4-linked N-acetylglucosamine residues (black), an N-linked acyl group (green), and other host-specific decorations (red). NF = Nodulation factor. Figure revised from (Oldroyd *et al.*, 2011).

During nodule organogenesis, there are regions of high auxin levels at the root tip versus high cytokinin levels further up the root at the transition between the proximal meristem (a region of cell division) and the elongation zone, (a region of cell expansion) (Oldroyd *et al.*, 2011). This auxin/cytokinin ratio is maintained by the auxin induced response regulators RR7 and RR15, that suppress cytokinin signaling and the cytokinin induced transcription factor SHY2 that suppress the expression of auxin transporters (Dello Ioio *et al.*, 2008; Müller & Sheen, 2008). Several transcription factors are induced downstream of *LHK1*. Among these are the response regulator RR1 and RR4 that function as a positive and negative regulator of cytokinin signaling (Plet *et al.*, 2011; Gonzalez-Rizzo, Crespi & Frugier, 2006). Moreover, NIN and NSP2 are highly induced by cytokinin (Gonzalez-Rizzo, Crespi & Frugier, 2006).

The most common mode for rhizobial infection is the formation of infection threads but the infection can also occur via root hairs, via cracks in the epidermis, or via interstitial infections between epidermal cells (Gage, 2004). Three E3 ubiquitin ligases, Plant U-box protein 1 (PUB1), CERBERUS, and nodule specific RING finger protein (nsRING), have been found to be involved with bacterial infection (Mbengue *et al.*, 2010; Yano *et al.*, 2009; Shimomura *et al.*, 2006). In *M. truncatula*, Mbengue *et al.* found that PUB1 is phosphorylated by the Nod-factor receptor LYK3, and

that PUB1 is a negative regulator of bacterial infection (Mbengue *et al.*, 2010), whereas both CERBERUS and nsRING were found to be required for infection (Yano *et al.*, 2009; Shimomura *et al.*, 2006). After NF recognition, several molecular processes occur to ensure bacterial infection. Firstly, the polar root growth is interrupted, and sometimes accompanied by swelling at the root-hair tip. Here after, growth is resumed to form a branch (Esseling, Lhuissier & Emons, 2003). This results in the root hair bending hereby entrapping the bacteria in a so-called infection pocket (Geurts, Fedorova & Bisseling, 2005). Here, the bacteria divide and form colonies (Oldroyd *et al.*, 2011). Secondly, less than 10 minutes after NF recognition cytoskeletal changes and changes in microtubule organization are induced (Weerasinghe *et al.*, 2003). This sets up the framework for directional cell expansion (Petrášek & Schwarzerová, 2009). Actin rearrangements occurs via the ARP2/3 complex (Smith & Oppenheimer, 2005), which needs the SCAR/WAVE complex to be activated. Components of the SCAR/WAVE complex are encoded by the *NAP1* and *PIR1* genes, and mutations in these genes have been correlated with loss of normal actin rearrangements (Yokota *et al.*, 2009). Localized plant cell wall degradation is necessary for the infection without causing cell rupture (Ridge & Rolfe, 1985), correlating well with the inducement of the cell wall degrading enzymes pectinmethylesterase and polygalacturonase found in other legumes (Lievens *et al.*, 2002; Muñoz *et al.*, 1998). Changes to the plant plasma membrane also occur during bacterial infection. Specialized compartments possibly associated with receptor functions, named lipid rafts, are formed (Oldroyd *et al.*, 2011). The flotillins FLOT2 and FLOT4, which are markers for lipid rafts are required for infection thread initiation (Haney & Long, 2010; Langhorst *et al.*, 2008; Kioka, Ueda & Amachi, 2002), and Symbiotic Remorin 1 (SYMREM1) also associated with formation of lipid rafts is important during bacterial colonization (Lefebvre *et al.*, 2010).

Later in the symbiotic development of nodules, the *Rhizobia* are induced to differentiate into bacteroids and express nitrogenase, hereby enabling N fixation (Suganuma *et al.*, 2003). In *L. japonicus*, symbiosomes usually contain two or more bacteroids (Oldroyd *et al.*, 2011). Several genes, in which mutations have been shown to affect bacteroid development or N fixation, have been identified both in *L. japonicus* and in the *Rhizobia* (Kawaguchi *et al.*, 2002; Fischer, 1994). N fixation requires specific induction of bacterial *nif* genes that are directly involved in the N fixation process and *fix* genes that affect plant symbiosis in regards to N-fixation (Long, 2001). Several of these have been identified and reviewed by Fischer (Fischer, 1994). Moreover, N fixation is accompanied by the down-regulation of bacterial NH_4^+ assimilation into amino acids (Patriarca, Tatè & Iaccarino, 2002). Of plant supplied components that are important for bacteroid development is the integral membrane protein "Stationary Endosymbiont Nodule 1" (SEN1) (Hakoyama *et al.*, 2012; Suganuma *et al.*, 2003). Prior to N fixation, a very high nodule specific expression (~10 % of all produced mRNAs) of symbiotic plant leghemoglobins occurs (Ott *et al.*, 2009; Ott *et al.*, 2005). Leghemoglobins are crucial for N fixation and have been shown to buffer free O_2, hereby maintaining high flux to sites of respiration and avoiding inactivation of the oxygen-labile enzyme nitrogenase (Ott *et al.*, 2005). Mutations in other plant genes have also been shown to cause non N fixating (fix$^-$) phenotypes. Among these is the *FEN1* gene that encodes homocitrate synthase (Hakoyama *et al.*, 2009). This can be explained by the fact that plant provided homocitrate is incorporated into the FeMo cofactor of dinitrogen reductase in bacteroids, which has been shown to be important for N fixation (Hakoyama *et al.*, 2009). Another gene found to be crucial is the *SST1* gene encoding a sulfate transporter located on the symbiosome membrane in nodules. The encoded protein ensures the transport of essentially important sulfate from the plant cell cytoplasm to the intracellular Rhizobia (Krusell *et al.*, 2005).

3.2 Transcriptome Data Analysis of *Lotus japonicus* During Nodulation

3.2.1 Introduction to the Transcriptome Analysis of *Lotus japonicus* During Nodulation

The purpose of the current study was to elucidate the complex signaling pathway involved in the symbiotic development of *L. japonicus* nodules and subsequent nitrogen fixation. To facilitate this, *L. japonicus* wild type plant (Gifu B-129) and mutant plants with altered nodule developments at different stages were chosen for transcriptome analysis. The selected mutant plants were:

- *nfr5*, which has a mutation in the *NFR5* gene causing a lack of the activation domain in NFR5. This inactivates the Nod factor signaling receptor and hereby blocks nodulation at a very early stage (Madsen *et al.*, 2003; Radutoiu *et al.*, 2003).
- *nin*, which has a mutation in the *NIN* gene is also arrested at the stage of bacterial recognition, but subsequent to calcium oscillations (Schauser *et al.*, 1999).
- *snf1*, which has a single amino acid mutation in the gene encoding CCaMK. This mutant spontaneously develops nodules without rhizobial infection (Tirichine *et al.*, 2006).
- *sym11*, which has a mutation in the gene encoding SEN1, which is essential for nitrogen fixation activity and bacteroid differentiation (Hakoyama *et al.*, 2012; Suganuma *et al.*, 2003). The *sym11* mutant is arrested late in nodulation and produce nodules that in the beginning are pinkish, indicating that leghemoglobin synthesis has initiated before further development is terminated (Sandal *et al.*, 2006; Schauser *et al.*, 1998).

The experimental setup enabled a detailed transcriptome analysis, where differences between the mutants and wild type plants at different stages in nodulation could be identified, potentially leading to the elucidation of the complex signaling pathway involved in the symbiotic development of *L. japonicus* nodules and subsequent nitrogen fixation. Wild type and mutant plants were inoculated with *M. loti* and tissue samples were collected at different time points after infection, cf. Table 3-3.

Table 3-3 Overview of *L. japonicus* tissue samples used in the transcriptome analysis. MAFF = Plants are inoculated with *M. loti* strain MAFF 303099. Root window is the root sector containing elongating root hairs, ca. cm 1 to 3 on the root starting from the tip. This zone is the most responsive to the presence of Nod factor produced by symbiotic bacteria (Høgslund *et al.*, 2009; Oldroyd & Downie, 2008). The Root infection is the zone of infection thread initiation located ca. cm 1 to 5 on the root from the tip (Høgslund *et al.*, 2009).

ID	Genotype	Treatment	Tissue	Time after treatment
01	Wild Type	None	Root	0 hours
02	Wild Type	None	Root Window	0 hours
03	Wild Type	MAFF	Root Window	8 hours
04	Wild Type	MAFF	Root Window	16 hours
05	Wild Type	MAFF	Root Window	24 hours
06	Wild Type	MAFF	Root Window	48 hours
07	Wild Type	MAFF	Root	3 days
08	Wild Type	MAFF	Root Infection Zone	3 days
09	Wild Type	MAFF	Root and Nodules	7 days
10	Wild Type	MAFF	Root and Nodules	14 days
11	Wild Type	MAFF	Nodules	14 days
12	Wild Type	MAFF	Nodules	21 days
13	nfr5	None	Root Window	0 hours
14	nfr5	MAFF	Root Window	8 hours
15	nfr5	MAFF	Root Window	16 hours
16	nfr5	MAFF	Root Window	24 hours
17	nfr5	MAFF	Root Window	48 hours
18	nin	None	Root Window	0 hours
19	nin	MAFF	Root Window	8 hours
20	nin	MAFF	Root Window	16 hours
21	nin	MAFF	Root Window	24 hours
22	nin	MAFF	Root Window	48 hours
25	snf1	None	Nodules	21 days
26	sym11	MAFF	Nodules	21 days

3.2.2 Bioinformatic Resources Used for Data Interpretation

As mentioned, a good transcriptome sequence model is crucial if biological interpretations are to be made from transcriptome data. As described earlier, especially tag based transcriptome analyses require high quality transcriptome models in order to make biological interpretations, due to complications regarding annotation of tags. The primary bioinformatic resources used in the current project has been Tentative Consensus sequences (TCs) from the TIGR Gene Index Project (Quackenbush *et al.*, 2000)[12], PlantGDB-assembled unique transcripts (PUTs) from PlantGDB (Dong, Schlueter & Brendel, 2004)[13], both are collections of EST clusters, and the predicted transcriptome sequences based on the gene prediction of the *L. japonicus* genome sequence released by the Kazusa DNA Research Institute (Sato *et al.*, 2008)[14]. All three resources have improved throughout the course of the current project facilitating data interpretation, cf. Figure 3-3. However, due to the time scope of the current thesis, not all analyses have been performed using the latest versions of these resources.

[12] Available at : http://compbio.dfci.harvard.edu/cgi-bin/tgi/T_release.pl?gudb=l_japonicus
[13] Available at : http://www.plantgdb.org/download/download.php?dir=/Sequence/ESTcontig/Lotus_japonicus
[14] Available at: http://www.kazusa.or.jp/lotus/index.html

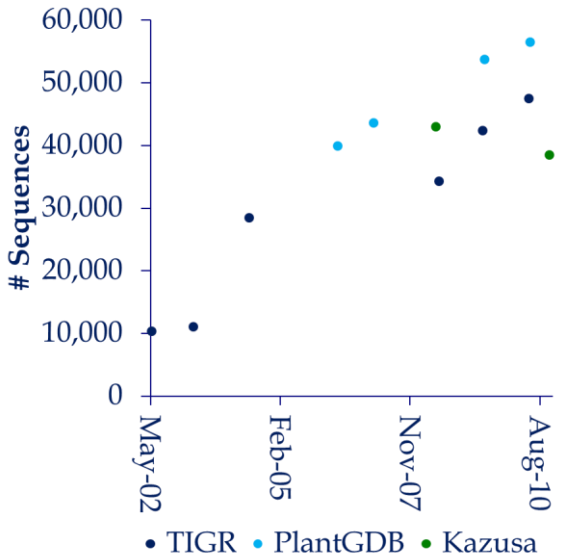


Figure 3-3 Development of sequence collections used for annotation of sequence tags. The number of sequences found in the different versions of Tentative Consensus sequences from the TIGR Gene Index Project (Quackenbush *et al.*, 2000), PlantGDB-assembled unique transcripts from PlantGDB (Dong, Schlueter & Brendel, 2004), and CDS sequences based on the gene prediction of the *L. japonicus* genome sequence released by the Kazusa DNA Research Institute (Sato *et al.*, 2008) are shown, respectively.

Later, in section 3.2.4, the quality of the difference sequence collections in regards to tag annotation will be discussed.

3.2.3 Tissue Sampling and Primary Data Processing

L. japonicus plants were prepared and grown in the greenhouse of Århus University in Ølsted. A total of 71 samples of roots, root window root infection zone and nodules were harvested and submitted to the DeepSAGE library preparation protocol. The root window is defined at the root sector containing elongating root hairs, ca. cm 1 to 3 on the root starting from the tip. This zone is the most responsive to the presence of Nod factor produced by symbiotic bacteria (Høgslund *et al.*, 2009; Oldroyd & Downie, 2008). The root infection is the zone of infection thread initiation located ca. cm 1 to 5 on the root from the tip (Høgslund *et al.*, 2009). For further details regarding tissue sampling and DeepSAGE library preparation readers are referred to (Petersen, 2008) (Petersen, 2008). With a few exceptions, cf. Table 3-4, three libraries representing biological replicates were produced for each genotype at each time point. 8-12 libraries were pooled and sequenced in a single lane on an Illumina GA sequencer. All samples were sequenced twice, initially producing 6 libraries for each genotype and time point. Raw image data was preprocessed automatically using *SolexaTagExtractionPipeline.pl* by which image analysis and base calling were performed using the Genome Analyzer Pipeline version 0.3.0 omitting chastity filtering – otherwise default settings, followed by sequence tag extraction from FASTQ files, and subsequent counting and tabulating of sequence tags, hereby generating raw tag list for each sample based on the barcode sequence. As mentioned, the *L. japonicus* data set was the first larger DeepSAGE data set produced at AAU. Therefore, several different filtering and error correction methods (evaluated in 2.2.1) were applied. However, if not stated otherwise, the data was filtered and error corrected as follows: The two technical replicates from each sequence run was combined using *CombineLibraryCounts.pl*, and singleton tags was subsequently removed using *CutoffLibs.pl*. Hereafter, each sample was subjected to SAGEscreen error correction using the implemented algorithm in the CLC Genomics

Workbench V 4.9 (requiring conversion of tabular tag lists into FASTQ files, which can be imported into the CLC Genomics workbench using *TagLists2FASTQ.pl*, export of SAGEscreen error corrected lists from the CLC Genomics Workbench in .csv format, and conversion of tag lists in .csv format into tabular tag lists using *CLCcsv2taglist.pl*). On average 97.1 % of the tag sequences were retained, while lowering the data complexity to 46.1 % of the raw number of unique tag sequences, cf. Table 3-4.

Table 3-4 DeepSAGE Library sizes after filtering and SAGEscreen sequence error correction. Sequences retained after filtering and SAGEscreen sequence error correction are given in percentages. On average 97.1 % of the tag sequences were retained, while lowering the data complexity to 46.1 % of the raw number of unique tag sequences. Failed = Sample failed during library preparation. WS = Replicate was identified to originate form a different sample and therefore excluded. Sample type for each ID is given in Table 3-3.

ID	Replicate 1	Replicate 2	Replicate 3
01	556,376 (97.1 %)	729,387 (96.3 %)	214,703 (93.5 %)
02	321,618 (92.4 %)	409,123 (96.1 %)	172,297 (90.4 %)
03	452,817 (94.2 %)	308,029 (92.3 %)	21,303 (79.0 %)
04	175,156 (92.4 %)	Failed	28,550 (79.0 %)
05	259,795 (92.8 %)	225,338 (92.7 %)	99,346 (87.8 %)
06	72,203 (82.9 %)	3,026,799 (99.2 %)	61,389 (81.0 %)
07	276,060 (92.8 %)	80,196 (80.3 %)	59,905 (86.6 %)
08	WS	799,596 (97.2 %)	357,849 (93.9 %)
09	83,931 (86.7 %)	184,340 (91.1 %)	55,761 (89.8 %)
10	839,545 (95.1 %)	10,987 (65.2 %)	109,934 (87.1 %)
11	454,428 (95.8 %)	319,912 (94.2 %)	204,324 (90.0 %)
12	5,278,105 (99.5 %)	1,069,629 (96.9 %)	Failed
13	4,793,445 (99.4 %)	WS	111,426 (85.8 %)
14	1,874,942 (98.5 %)	2,606,682 (98.8 %)	124,010 (91.6 %)
15	28,802 (72.3 %)	42,402 (78.1 %)	1,050,631 (97.0 %)
16	50,173 (81.6 %)	118,136 (84.5 %)	402,989 (94.7 %)
17	925,930 (96.7 %)	54,926 (81.2 %)	287,082 (92.1 %)
18	15,421 (67.4 %)	55,881 (82.5 %)	149,341 (85.1 %)
19	47,583 (76.2 %)	Failed	205,255 (91.2 %)
20	28,514 (76.0 %)	1,523,266 (98.1 %)	476,357 (94.3 %)
21	12,390 (63.9 %)	3,769,347 (99.4 %)	106,360 (88.9 %)
22	56,099 (77.7 %)	Failed	50,233 (80.2 %)
25	70,312 (84.6 %)	49,754 (83.7 %)	28,772 (79.7 %)
26	77,946 (79.8 %)	15,436 (65.4 %)	65,385 (82.2 %)
27	WS	WS	WS

It was discovered that during the library preparation, when samples were pooled, inadequate concentration measurements had been performed. This gave rise to very large size differences between the libraries resulting in nearly a 600 fold difference between the smallest and largest library, cf. Figure 3-4.

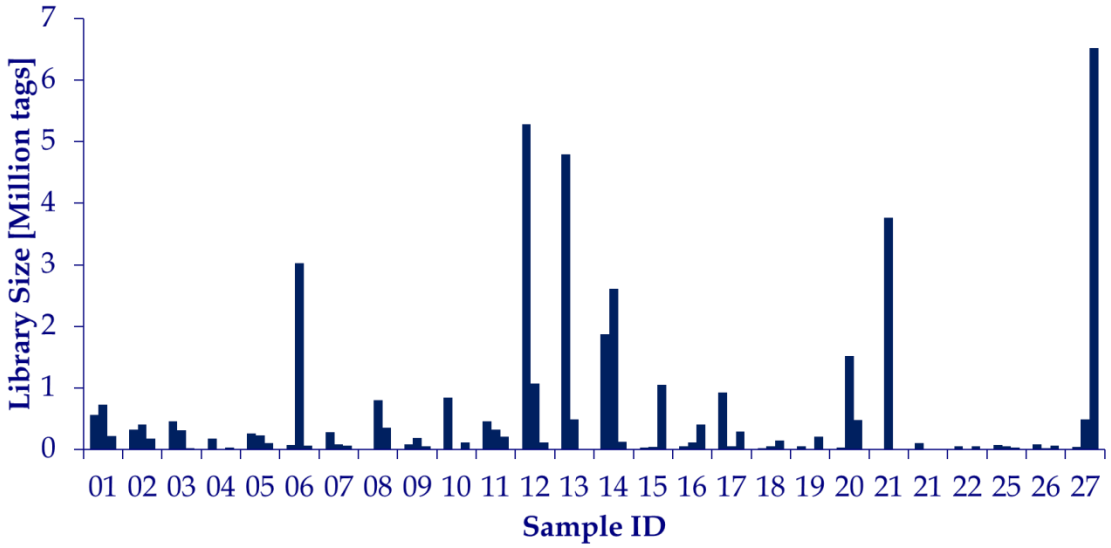


Figure 3-4 DeepSAGE Library sizes after filtering illustrating the large difference between individual libraries. The difference between the smallest and largest library is nearly 600 fold. Sample type for each ID is given in Table 3-3.

The difference in size between the libraries affects the data in several ways. Firstly, more genes are detected in libraries that have been sequenced more deeply. These genes will easily be detected as differentially expressed when the large libraries are compared to smaller, where the chance of detecting these genes is smaller. Moreover, the large difference can cause problems during normalization of the expression values. Using simple normalization of tag counts, a gene observed twice in a library of 1 million tags and once in a library of 100,000 tags will be normalized to 2 and 10 CPM, respectively. The resulting observed 5-fold difference is more likely to be an artifact of the normalization, than a true differential expression of the gene.

3.2.4 Generation of a *Lotus japonicus* Transcriptome Model for Tag Annotation

As mentioned, this project was a joint effort between Århus University and Aalborg University. Therefore, it was chosen to compare the expression profiles from this study with similar expression profiles (same genotype, tissue type, developmental stage, and treatment) found in a microarray study by the same research group at AU (Høgslund *et al.*, 2009). A direct comparison required the use of the transcriptome model only containing CDSes from the gene prediction of the *L. japonicus* genome sequence (Kazusa-CDS V1) released by the Kazusa DNA Research Institute (Sato *et al.*, 2008) for the annotation of sequence tags, since Kazusa-CDS V1 was used to generate the *L. japonicus* geneChip® (Lotus1a520343) [15]. Therefore, the quality of the available *L. japonicus* transcript models in regards to tag annotation was investigated. All samples were pooled and tags were initially annotated against *L. japonicus* TCs version 5 (LJGI-5) from the TIGR Gene Index Project (Quackenbush *et al.*, 2000) and Kazusa-CDS V1, cf. Figure 3-5. Later in the project, an update to the TCs (LJGI-6) and PUTs from PlantGDB (Dong, Schlueter & Brendel, 2004) version 177a (May 26th 2010) were implemented, why annotation statistics for these are also shown in Figure 3-5.

[15] Affymatrix Custom Array, accession A-AFFY-90. For further details see: http://www.ebi.ac.uk/arrayexpress/arrays/A-AFFY-90

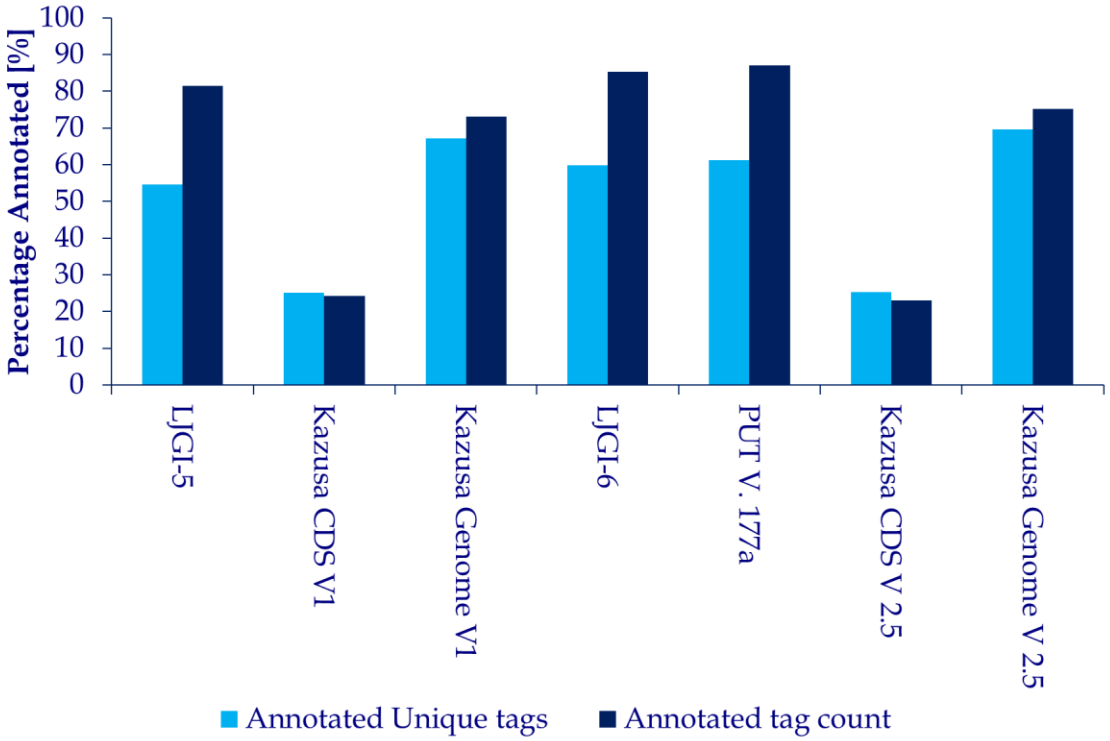


Figure 3-5 Annotation of all DeepSAGE libraries against different *L. japonicus* transcriptome models. The fraction of unique tags, and total tag counts that could be annotated are given. Tags were matched against tentative consensus sequences from the TIGR Gene Index Project version 5 and 6, CDS and genome models from the *L. japonicus* genome sequence released by the Kazusa DNA Research Institute version 1 and 5, and PUTs from PlantGDB version 177a allowing a single mismatch.

From Figure 3-5 it is clear that Kazusa-CDS V1 only containing 25 % of the different tags observed in data set is inferior to LJ-GI5 that contains 55 % of all unique tags in the data set. The difference is even larger, when comparing the total number of tags that can be annotated. Here, 23 % compared to 81 % could be annotated using Kazusa-CDS V1 and LJ-GI5, respectively.

The difference observed in EST collections (both TCs and PUTs) for unique tags, and total tag count is easily explained by the fact that highly expressed genes are more likely to be represented in an EST collection. However, the low number of tags that match Kazusa-CDS V1 and the difference observed between Kazusa-CDS V1 and the genome model itself could arise from either missing or mis-annotated gene models or the fact that the CDS model does not contain information of the un-translated regions (UTRs). The latter seems likely, since SAGE methods produce tags that are located in the 3′ end of the transcript. A previous study by Pesole *et al.*, showed that the average length of plant 3′ UTRs is ~ 200 bp (Pesole *et al.*, 2001). Therefore, the primary SAGE tag produced is likely to be located in the 3′ UTR. To investigated this, an *in silico* analysis was performed using *ExtentCDSmodels.pl* and *GlobalSAGEmap.pl* (cf. appendix B) adding different amounts of adjacent nucleotides to the CDS models in both the 5′and 3′ ends, cf. Figure 3-6.

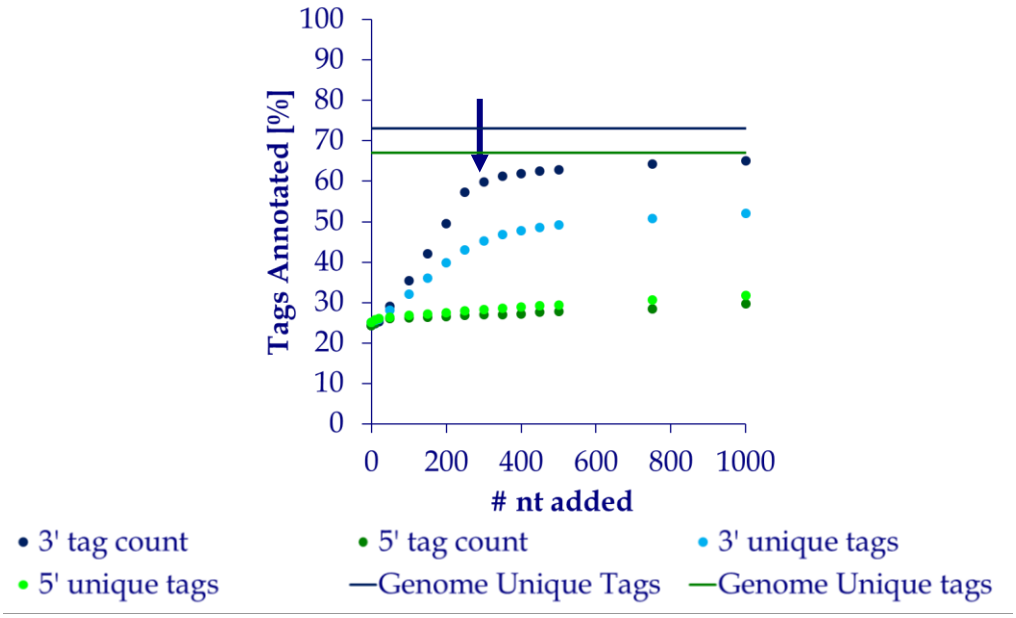


Figure 3-6 Improvement of the Kazusa transcriptome model by nucleotide addition in the 3' end of the transcript. Percentage of unique tags and the total tag count matching the Kazusa transcriptome model after addition of different amounts of adjacent nucleotides to the CDS models in both the 5'and 3' end. The arrow indicates the point (at 250 nt) where the rate of annotated tags / nt added decreases. Based on this 250 nt was chosen as the length to extent the CDSes.

As expected, adding nucleotides to the CDS models in the 3' end clearly improves the number of tags that can be annotated; while addition in the 5' end (which was included to estimate if observed improvements were random) does not, cf. Figure 3-6. At ~250 nt, the rate of annotated tags / nt added decreases. This implies that only few CDS have longer 3'UTR regions. This result is in agreement with those found by Pesole *et al.*, since 200 bp could seem as a good estimate of the average length of the 3' UTR. Although the annotation of tags improves, it does not reach the same level as the entire genome sequence, cf. Figure 3-6. This difference is likely to be explained by genes that are not annotated in the genome sequence, and to a lesser extend random DNA contamination in the DeepSAGE data set. However, this was not investigated further. Based on these results it was chosen to incorporate a sequence collection comprising of the CDS models with 250 bp added in the 3' end (from here on named "Kazusa-CDS+250nt") for the annotation of sequence tags.

Initial biological interpretation of DeepSAGE data is dependent on functional annotations of the sequences that the sequence tags are matched against (CDS, EST mRNA etc.). Therefore, since no functional annotation existed for the sequences in the Kazusa CDS model, such had to be made. Shortly, taxonomy names of all 94,180 species belonging to the *Viridiplantae* kingdom was retrieved at The NCBI Taxonomy Homepage[16] and Uniprot Reference Clusters (Uniref100) based on Release 57.0 of 24-Mar-09 of UniProtKB/Swiss-Prot was retrieved from the European Bioinformatics Institute (EMBL-EBI)[17]. Using *GetTaxIDs.pl* all *Viridiplantae* Uniref100 clusters was extracted. These were subsequently used subjects for BLASTX Similarity searches (Camacho *et al.*, 2009; Altschul *et al.*, 1990) of all sequences in "Kazusa-CDS+250nt". Using *CreateID2nameTable.pl*, *GetBestUniRefBLASTResult.pl*, and *AddUniRefAnnotation.pl*, functional annotations were added to sequences with a significant match (E-value =< $1*10^{-5}$). Matches

[16] Search string used at http://www.ncbi.nlm.nih.gov/taxonomy: "Viridiplantae[SubTree] AND species[Rank] NOT uncultured[prop] AND ("above species level"[prop] OR specified[prop])". Resulting file with taxon names from the search was subsequently downloaded.
[17] Current version available at: ftp://ftp.ebi.ac.uk/pub/databases/uniprot/uniref/uniref100/. For details regarding releases see: http://www.uniprot.org/news/2009/03/24/release.

with functional annotations containing phrases such as "uncharacterized", "hypothetical", "predicted protein", and "whole genome shotgun sequence of line" were omitted if possible. The latter phrase was included to avoid *Vitis vinifera* sequences having non-informative functional annotations. Later, the same procedure was performed on LJGI-6, since this had outdated functional annotation, making searches using the Uniref100 IDs impossible in some cases.

3.2.5 Comparison of DeepSAGE and Microarray Data Sets

To facilitate biological interpretation based on this study combined with results earlier obtained by the collaborating research group at Århus University in a microarray study, the two data sets were compared (Høgslund *et al.*, 2009). In the scope of the current thesis, it was chosen to investigate the ability of the two methods to determine gene expression levels, and more importantly the ability to detect differential expression. The ability to determine gene expression levels with high confidence is reflected in difference between the correlation observed between sample replicates and different biological samples, respectively. Therefore this was analyzed. Furthermore, the correlation between the two data sets was investigated based on an analysis of the correlation between the observed expression values, and an analysis of the correlation between changes observed in the gene expression in the two data sets.

3.2.5.1 Methods

Processed data from the study by Høgslund *et al.* (Høgslund *et al.*, 2009) was downloaded at the ArrayExpress archive hosted by EMBL-EBI[18], and samples corresponding to a sample in the DeepSAGE data set in regards to same genotype, tissue type, developmental stage, and treatment were extracted, cf. Table 3-5.

Table 3-5 Overview of *L. japonicus* DeepSAGE and Affymatrix libraries used in the comparative study. WT = wild type. MAFF = Plants are inoculated with *M.loti* strain MAFF 303099. Affymatrix sample ID covers a triplicate. Each sample begins with F, M or W. See sample and data Relationship found at http://www.ebi.ac.uk/arrayexpress/experiments/E-TABM-715/sdrf.html for further details.

DeepSAGE	Affymatrix*	Genotype	Treatment	Tissue	Time after treatment
02	11	WT	None	Root Window	0 hours
05	05	WT	MAFF	Root Window	24 hours
07	15	WT	MAFF	Root	3 days
09	34	WT	MAFF	Root and Nodules	7 days
11	33	WT	MAFF	Nodules	14 days
12	31	WT	MAFF	Nodules	21 days
16	02	nfr5	MAFF	Root Window	24 hours
18	09	nin	None	Root Window	0 hours
21	03	nin	MAFF	Root Window	24 hours
26	29	sym11	MAFF	Nodules	21 days

DeepSAGE libraries were annotated against the "Kazusa-CDS+250nt" sequence collection allowing 1 mismatch and discarding non-uniquely matching tags. Hereafter, the mean expression level (EXP) and standard deviation (STDV) for each gene were calculated by summing matching tag counts for each gene using and subsequently applying equations (3-1)

[18] Available at: http://www.ebi.ac.uk/arrayexpress/experiments/E-TABM-715

and (3-2) implemented in *Tag2GeneCounts.pl* and *MeanAndSTD.pl*. Note that standard deviations are multiplied by the ratio between the library size and the sample group size.

$$\mathrm{EXP} = \frac{\sum_{i=1}^{N}(\text{tag count}_i)}{\sum_{i=1}^{N}(\text{lib size}_i)} \quad (3\text{-}1)$$

$$\mathrm{STDV} = \sqrt{\sum_{i=1}^{N}\left(\left(\frac{\text{tag count}_i}{\text{lib size}_i} - \mathrm{EXP}\right)^2 \cdot \frac{\text{lib size}_i}{\sum_{i=1}^{N}(\text{lib size}_i)}\right)} \quad (3\text{-}2)$$

Where:
EXP = Mean expression level
N = # replicates
STDV = standard deviation

Firstly, only probes representing a gene in the *L. japonicus* genome were kept. Secondly, probes matching multiple transcripts (marked with "_s_at") and probes that are identical, or highly similar, to un-related sequences (marked with "_x_at") were discarded (similar to discarding non-uniquely matching tags in the DeepSAGE data set). Thirdly, due to the fact that gene and clone IDs frequently have been updated or collapsed into new names and the identifiers therefore differ between the Affymatrix geneChip® and the genome annotation, only probes, which could be unambiguously linked between the two were kept. Therefore, out of 17,385 genes observed in the DeepSAGE samples 6,988 genes could be compared with the Affymatrix study.

The difference between the variance observed between sample replicates and different biological samples was calculated by firstly calculating the mean of Pearson's coefficients of variance (Pearson, 1895) (ρ_p) for each library and libraries being replicates of the same sample (reflecting internal correlation between replicates, named internal ρ_p, and secondly for each library and libraries not being replicates of the same sample (reflecting correlation between biological samples, named external ρ_p. Hereafter, the mean internal and external ρ_p for each biological sample were calculated. Affymatrix expression values were inverse log transformed prior to calculations.

Due to the nature of Affymatrix and DeepSAGE data, a linear relationship between the expression values is not expected. Therefore, the Spearman correlation (ρ_s) coefficients between corresponding Affymatrix and DeepSAGE samples were calculated for observed expression and for the observed fold change difference in the gene expression between sample 02 (Wild type day 0 after infection), and samples 05,07,09,11, and 12 (Wild type day 1,3,7,14, and 21 after infection), respectively. In cases where gene expression was not observed in a sample, the expression level was set to the lowest observed expression value in the sample. Moreover, gene expression values were categorized as noisy or not. Expression values were categorized as noisy either if expression was not observed or the coefficient of variance (CV) was > 1 in one of the samples compared, cf. equation (3-3) (Hendricks & Robey, 1936).

$$CV = \frac{\sqrt{\sigma^2}}{\mu} \qquad (3\text{-}3)$$

Where:
CV = Coefficient of variance
σ^2 = Variance of the mean expression
μ = Mean expression

3.2.5.2 Results

The ability of DeepSAGE and Affymatrix to determine gene expression levels was compared by analyzing the difference between the correlation observed between sample replicates and different biological samples, respectively, cf. Figure 3-7. For DeepSAGE, the average internal ρ_p was found to be 0.84 ± 0.01 and the external ρ_p was found to be 0.54 ± 0.05. These are significantly different (Student's T-test P-value = $1.7 \cdot 10^{-7}$). However, less variation was observed for the Affymatrix data set reflected by a higher average internal ρ_p (0.95 ± 0.00), and a more significant difference between internal and external ρ_p (Student's T-test P-value = $7.5 \cdot 10^{-13}$). This could indicate that the DeepSAGE data set in general is more "noisy", which should be accounted for during the data analysis when making biological interpretations. Another interesting observation can be made from Figure 3-7. In both data sets, the external ρ_p is lowest for samples originating from nodule tissue (samples 11, 12 and 27). This reflects the larger difference between this tissue and other tissues sampled in the experiment, indicating that nodules are a highly differentiated tissue type.

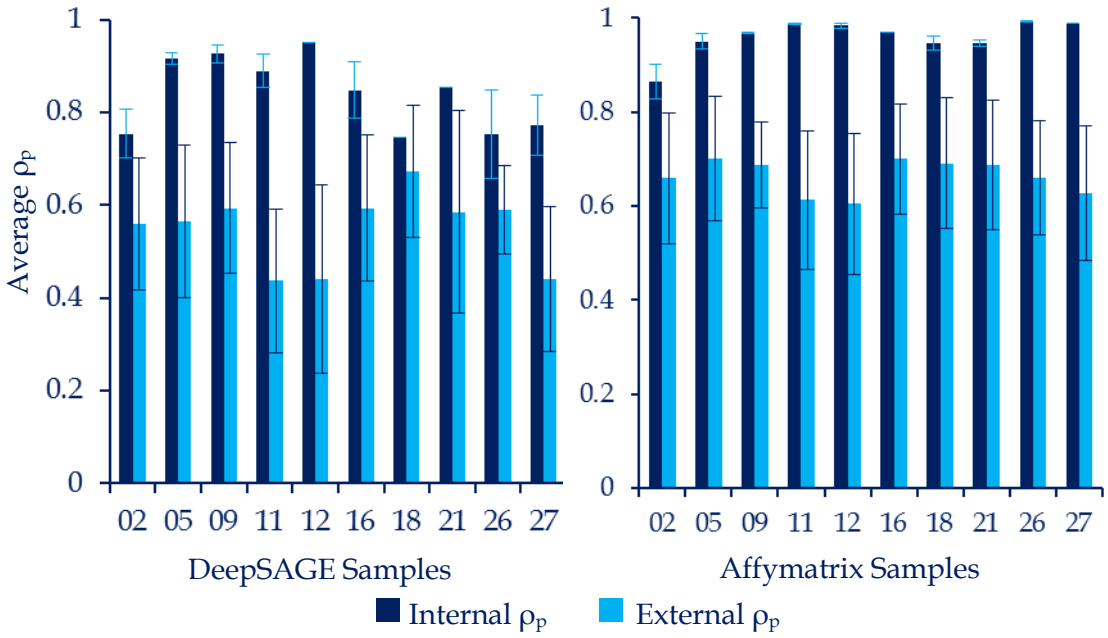


Figure 3-7 Average internal Pearson's correlation between replicates (internal ρ_p), and between biological samples (External ρ_p) for corresponding samples (same genotype, developmental stage, and treatment) from the DeepSAGE study (left) and the Affymatrix study (right) (Høgslund *et al.*, 2009). Both data sets have significantly lower external ρ_p. Moreover, the lowest external ρ_p is observed for nodule samples (11 and 12), reflecting a highly differentiated tissue type.

When comparing the DeepSAGE and Affymatrix expression data directly, cf. Figure 3-8 panel A, a fairly good correlation between expression values can be observed. Other studies have reported similar correlations between gene expression data obtained by digital methods (SAGE methods or mRNAseq) and microarrays, even when comparing libraries originating from the same tissue sample (Wang, Gerstein & Snyder, 2009; Ishii *et al.*, 2000), and not as in this case where

only corresponding samples in regards to genotype, developmental stage, and treatment are compared. Because of this, biological interpretations based on a combination of the DeepSAGE and the Affymatrix data seems valid even though the libraries originate from two different studies.

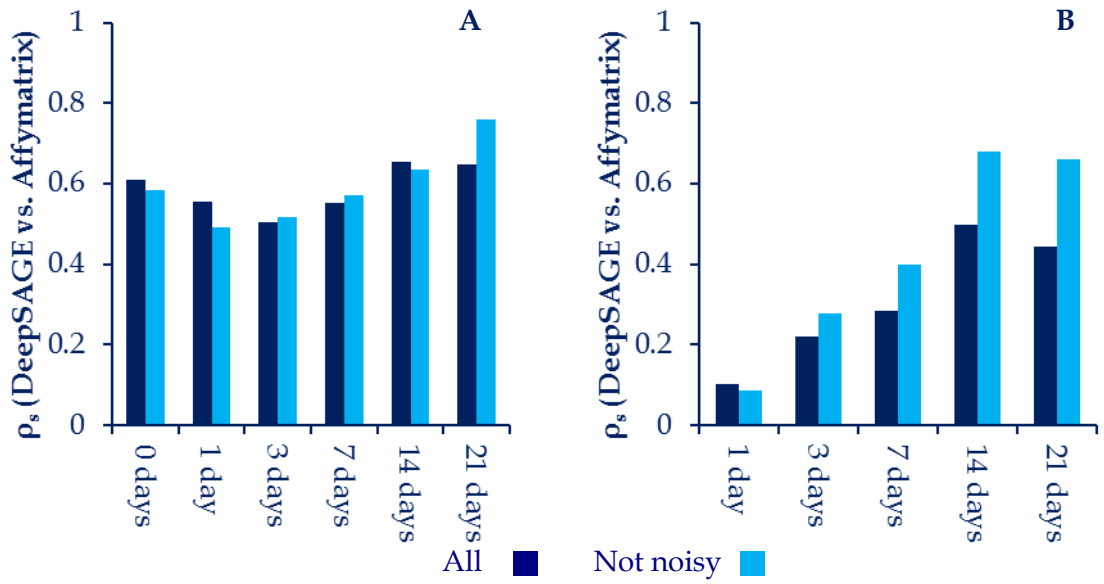


Figure 3-8 Correlation between the DeepSAGE and Affymatrix data sets. **A:** Spearman's correlation coefficients (ρ_s) between corresponding Affymatrix and DeepSAGE. **B:** ρ_s of fold changes (in relation to the "0 day" sample) between corresponding Affymatrix and DeepSAGE. Genes are categorized as not noisy if expression is observed and CV < 1 in all samples compared.

However, when comparing which genes display differential expression in the two data sets, the correlation is much lower, cf. Figure 3-9 and Figure 3-8 panel B. The observed fold change measured by DeepSAGE vs. Affymatrix from a time series of the wild type plant at between expression values day 0 and day 1, 3, 7, 14, and 21, respectively is shown in Figure 3-9. Good correlation between DeepSAGE and Affymatrix is indicated by a majority of fold changes located in the 1st and 3rd quadrant of the plots. However, at day 1, only 51 % of all and 53 % of the well determined genes show changes in the same direction in the DeepSAGE and Affymatrix data set. This would be expected when comparing two entirely unrelated data sets, and is also indicated in the very low correlation (ρ_s = 0.1), cf. Figure 3-8 panel B. The correlation improves throughout the time series reaching a maximum after 14 days, which is the first time point where the nodules are fully developed. Throughout the time series, more genes are down- than up-regulated. This reflects the differentiation and specialization of the nodule tissue. This fact is also reflected in the high number of genes that are up-regulated in the later time points compared to day 0, but poorly determined, cf. Figure 3-9. The majority of these are caused by either very low or no expression at day 0, which is poorly determined showing that these genes are highly nodule specific.

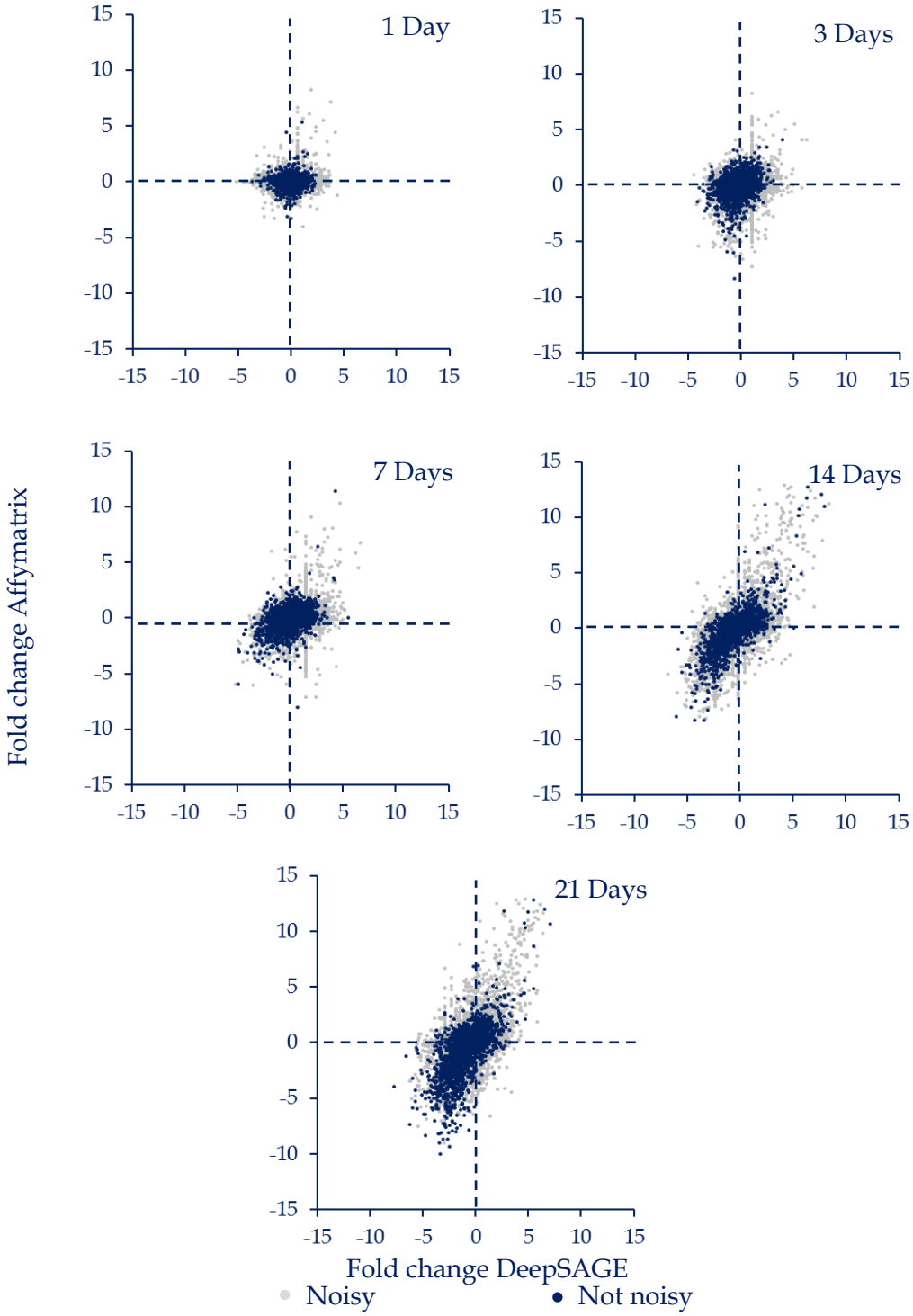


Figure 3-9 Log2 Fold change differences in gene expression in the DeepSAGE (X-axis) and Affymatrix (Y-axis) data sets. Wild type root window (sample 02) is compared with samples reflecting the nodule development (samples 05,07,09,11 and 12). More genes are down- than up-regulated reflecting the differentiation and specialization of the nodule tissue. Genes are categorized as noisy if expression is not observed or if CV > 1 in one of the samples compared.

The large size differences between the DeepSAGE libraries were the cause to some concern, because these potentially could influence the integrity of the data, and hereby have a negative influence the ability to draw biological interpretations, cf. section 3.2.3. However, the results from the comparison of the DeepSAGE and the similar Affymatrix data were compatible with similar studies (Wang, Gerstein & Snyder, 2009; Ishii *et al.*, 2000). Taking into account the complications regarding direct gene to gene comparison due to different annotations of the Affyma-

trix and DeepSAGE data, and the fact that corresponding libraries did not originate from the same biological sample, but only similar samples in regards to genotype, developmental stage, and treatment, the magnitude of the correlation between seems to be very reasonable.

3.2.6 Methods for Analysis of *Lotus japonicus* Time Series

In the following section, methods used for the data analysis of both the wild type and mutant time series, which are described in sections 3.2.7 and 3.2.8, are described. First, a data matrix of tag counts from all samples was made using *CompareSage.pl*. To minimize data loss due to lack of annotation, tags were annotated in hierarchical manner against LJ-GI6, "Kazusa-CDS+250nt", and PUT V177a. Tags with no annotation were subsequently removed and not used in further analyses. Hereafter, tag counts were summed to each gene using *Tag2GeneCounts.pl*, and treated as follows:

For detection of differential expression the mean expression level (EXP) and standard deviation (STDV) for each gene was calculated using and *MeanAndSTD.pl*. Samples were compared to the uninoculated state for the same tissue type; either root or root window (samples 01 and 02). Differential expression was determined using the Z-test (P-value =< 0.01, Bonferroni corrected) and a minimum fold change difference of 2. Genes found to be differentially expressed at any state were subsequently subjected to hierarchical pairwise complete-linkage clustering using Euclidean distance as distance measure using the clustering software Cluster 3.0 (de Hoon *et al.*, 2004). Clustering was performed on both genes and samples. For visualization purposes, EXP values were subsequently normalized to the maximum expression of each gene.

For principal component analysis (PCA), data was normalized to counts per million using *NormaliseTagTable.pl* accounting for differences in library sizes, and subsequently mean centered. Hereafter, PCA was performed on un-scaled and Pareto scaled data, cf. equation (1-4) using the software program The Unscrambler v 9.8 (Wass, 2005).

3.2.7 Data Analysis of Wild Type Time Series Samples

The wild type time series consisted of samples from 0 hours to 21 days after *M. loti* inoculation (AMI). Data analysis was firstly performed on the entire data set. Hereafter, more focused analyses were perform on samples originating from earlier time points (0-48 hours AMI, samples 01-06) because these could be compared with samples from the mutant genotypes that are arrested early in nodulation (*nfr5* and *nin*, samples 13-17, and 18-22, respectively) and on samples from later time points AMI (3-21 days after infection, samples 07-12) because these could potentially give new insight in the transcriptome of the fully developed nodule. Moreover, samples from the fully developed wild type nodules could be compared with corresponding samples from the two mutant genotypes that are arrested later in nodulation, or producing non-functional nodules (*snf1* and *sym11*).

Firstly, the gene expression of genes known to be involved in bacterial infection and/or nodule organogenesis was investigated. Several known gene expression patterns could be confirmed by the DeepSAGE transcriptome analysis, cf. Figure 3-10. The NOD factor receptor genes *NFR1* and *NFR5* are up-regulated at 8 and 16 hours AMI in the root window, respectively and are not expressed in functional nodules. This up-regulation indicates the involvement of these genes in the early parts of bacterial infection and nodule organogenesis. Early up-regulation of the genes *SYMRK, NUP133, NENA, NUP85, Castor*, and *ENOD11,*

which are known to be a part of early downstream signaling in nodule organogenesis could also be confirmed, cf. Figure 3-10. Especially, a co-regulation of the *SYMRK* and *ENOD11* genes is clear. The gene *CCaMK*, which is a part of the signaling downstream of calcium oscillations has maximum expression after 16 hours but is already induced 8 hours AMI. This, together with up-regulation of the *ENOD11, NIN* and *LHK1* genes indicate that calcium oscillations have occurred within the first 8 hours of the time series. The co-regulation of the *CCaMK*, *NIN* and *LHK1* genes was subsequently confirmed by quantitative PCR (data not shown, analysis performed at AU). The nucleoporins *NUP133, NENA, NUP85* and the ion channel *Castor* have maximum expression between 16 and 48 hours AMI. This could indicate that these are involved in or regulated by calcium oscillations. Another interesting observation is that the transcription factor NSP1, is specifically up-regulated 48 hours AMI, and subsequently down-regulated. This is well in line with the fact that NSP1 is a transcription factor that regulates gene expression down-stream of CCaMK activation. Furthermore, a similar expression pattern can be observed for the gene *LjERF1*. Asamizu *et al.* found that *LjERF1* expression was slightly up-regulated compared to the un-inoculated in early time points (1.8, 1.7 and 1.2 fold state after 3, 24 and 48 hours AMI, respectively), and the gene was subsequently down-regulated at a later state (10, 2.5 and 2 fold down-regulated at 4, 7 and 12 days AMI, respectively) (Asamizu *et al.*, 2008). The results of the current study are not in line with the findings by Asamizu *et al.*, but show a specific up-regulation at 2 and 3 days AMI. The down-regulation of *LjERF1* expression, later in nodule organogenesis, is however in line with the results of Asamizu *et al*. Some known expression patterns could not be confirmed by the current DeepSAGE transcriptome study. Several of these can be explained by low expression of the mRNA transcript, close to the noise limit. An example is expression of *NSP2* which is only detected 21 days AMI at a very low level (0.6 CPM). This is believed to be an artifact of the sequencing depth of the libraries. Moreover, for several of the genes where the known expression could be verified, the regulation could not be significantly determined due to relative low expression of the genes. CCaMK has been shown to phosphorylate CYCLOPS. However, the *CYCLOPS* gene is up-regulated at a later stage in the nodule organogenesis and has maximum expression 14 days AMI.

Several of gene expression patterns of genes known to be involved in bacterial infection could be confirmed, cf. Figure 3-10. Expression of two out of the three E3 ubiquitin ligases known to be involved could be detected. Firstly, PUB1 expression is induced end subsequently repressed at an early stage (8-16 hours AMI). This is well in line with the fact that PUB1 is a negative regulator of bacterial infection. Secondly, the observed gene expression pattern of LJnsRING, which is up-regulated at a later stage of bacterial infection, is well in line with the results found by Shimomura *et al.* (Shimomura *et al.*, 2006). Another interesting observation is that the expression of genes known to be involved in actin rearrangements and changes in the plant plasma membrane are co-expressed, cf. Figure 3-10. These include *NAP1*, *FLOT1*, and *SYMREM1* that all are induced a 3 days AMI. Finally, genes known to be essential for the final symbiotic nodule development such as *FEN1*, *SEN1* and *SST1* are induced late. Of these the expression of *SST1* is the most prominent. This gene is highly expressed at 14 and 12 days AMI and is up-regulated ~60 fold compared to the un-inoculated state.

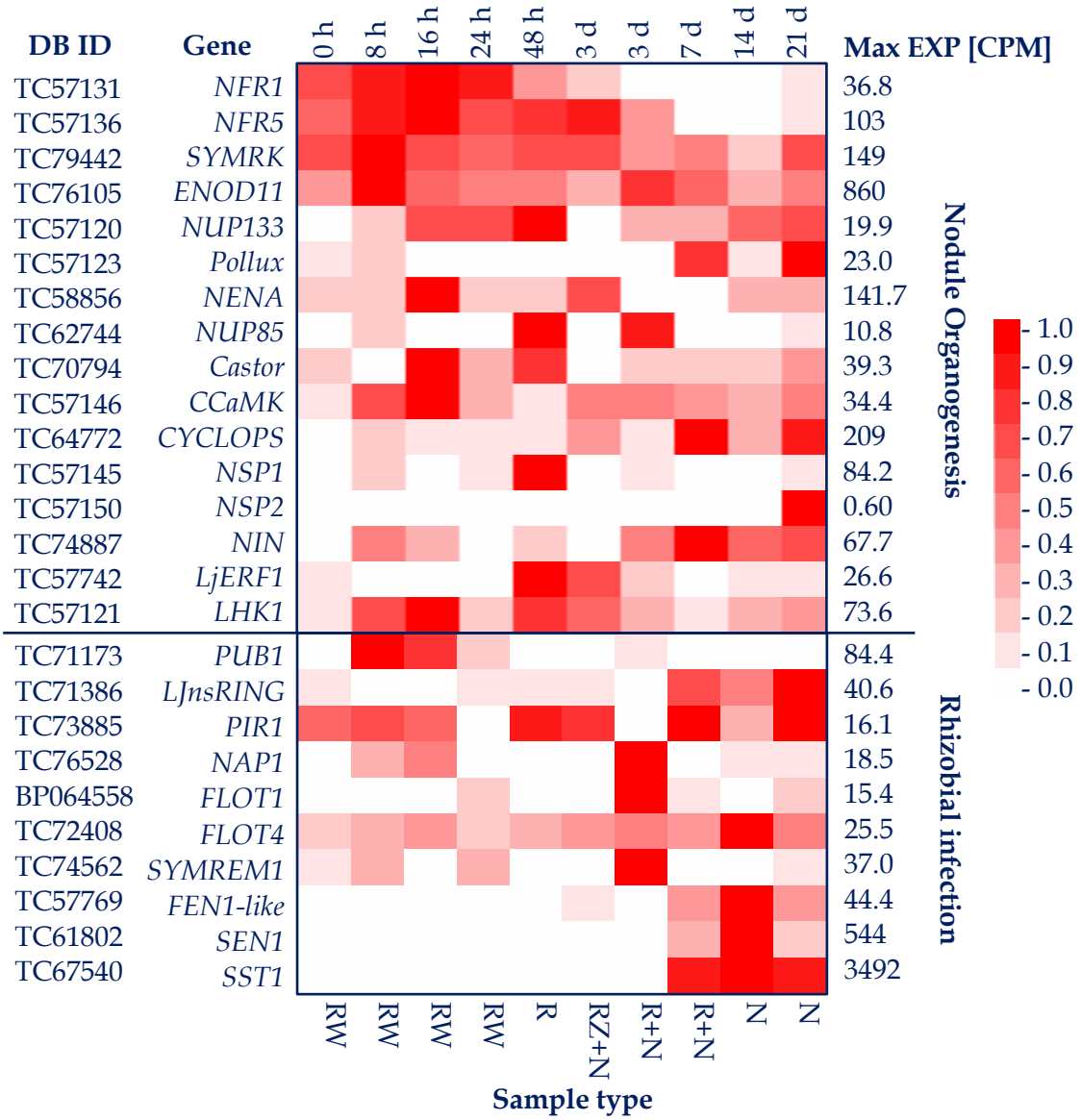


Figure 3-10 Relative mRNA expression levels of transcripts encoded by genes known to be a part of *L. japonicus* nodule organogenesis (top) and rhizobial infection (bottom). Time after inoculation with *M. loti* is given at the top, sequence ID and gene name is given on the left, and maximum expression level is given on the right. Colors represent the expression level relative to the maximum observed value for each transcript. h = hour, d = day, CPM = counts per million, RW = root window, R = root, and RZ = root infection zone.

To further elucidate the *L. japonicus* transcriptome during nodulation a PCA was performed on the wild type time series, cf. Figure 3-11 panel A. The first principal component nicely split the late samples (samples 09-12, 7-21 days AMI) from the early samples (samples 01-08, 0-24 hours AMI). Among the 100 transcripts with the highest PC1 values, cf. Figure 3-11 panel B, several transcripts encoded by genes known to be involved in nitrogen fixation could be found confirming that the PCA model reflected nodule organogenesis. Among these were 14 genes encoding Leghemoglobins (all included in the top 30 list of genes with the highest PC1 values) and several genes encoding known nodulins (ENOD11, ENOD18, ENOD-36A, and ENOD-16).

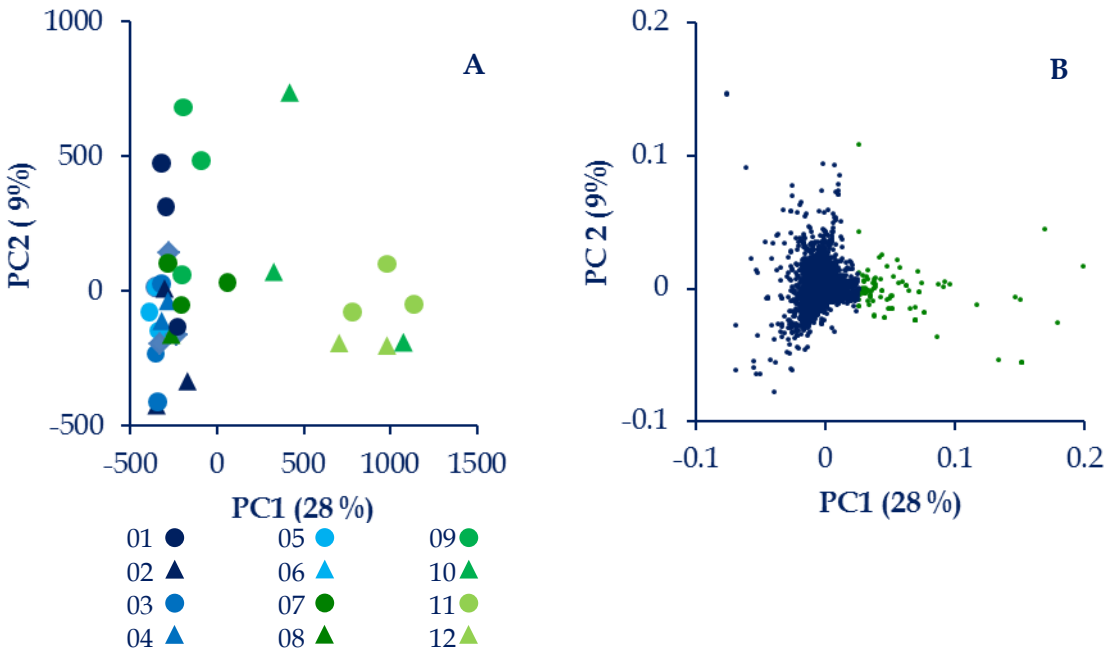
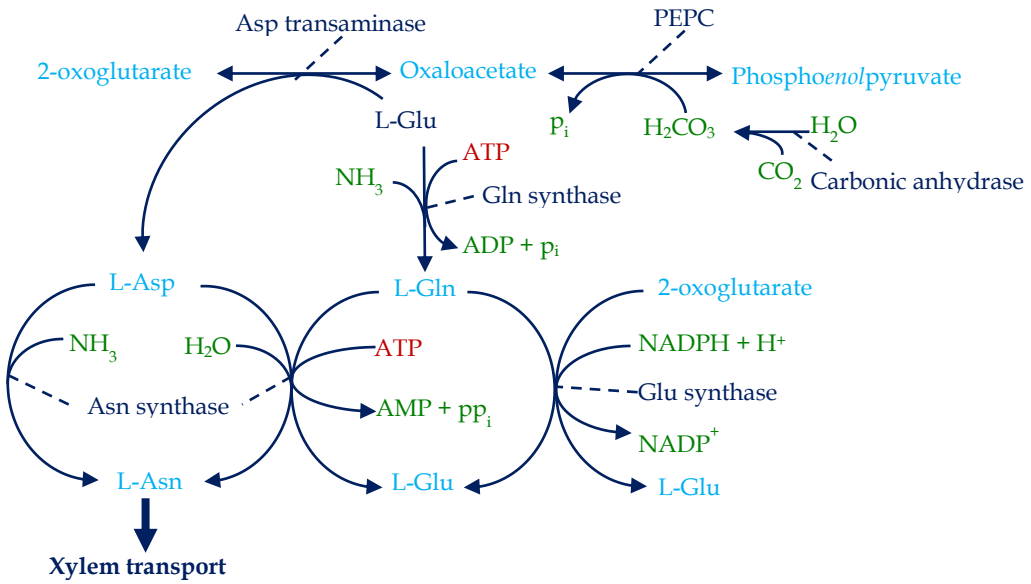


Figure 3-11 PCA plot of all 32 samples from the wild type time series ranging from 0 hours to 21 days after inoculation with *M. loti* strain MAFF 303099. **A:** Scores plot of PC1 vs. PC2 showing the split of early samples (samples 01-06 corresponding to 0-48 hours after infection marked in blue) and late samples (samples 07-12 corresponding to 3-21 days after infection marked in green). **B:** Loadings plot showing the splitting of genes. The top 100 genes ranked by the PC1 value, which are highly expressed in nodules, are marked in green. Data was pareto scaled prior to PCA. PC = Principal component. Numbers in percentages indicate variance explained by the PCA model.

Interestingly, TC57163 had the highest PC1 value indicating that this transcript is the most indicative transcript for nitrogen fixation. The transcript encodes an Asparagine (Asn) synthetase (EC 6.3.5.4), which is 50-55 fold up-regulated in nodules at 14 and 21 days after inoculation compared to day 0, respectively. The transcript accounts for ~ 5 % of the transcriptome in both cases. This finding lead to a further investigation of the expression levels of genes known to be involved in Asn metabolism. It is well-known that temperate legumes in contrast to many other plant species, utilizes asparagine, rather than glutamine (Gln) to transport reduced nitrogen within the plant (Shi *et al.*, 1997). In the case in of *L. japonicus*, asparagine can account for 86% of the nitrogen transported from root to shoot (Lea & Miflin, 1980). One possible reason for this difference is that Asn provides a more economic means for nitrogen transport, due to its lower N:C ratio (2:4) compared to that of Gln (2:5) (Waterhouse *et al.*, 1996). It is thought that the NH_4^+ produced by N fixation in general is assimilated by the action of Gln synthetase (EC 6.3.1.2) in conjunction with NADH-Glu synthase (EC 1.4.1.14), hereby producing Gln and glutamate (Glu), which then is subsequently converted into Asn by the action of aspartic acid (Asp) aminotransferase (EC 2.6.1.1) and Asn synthetase (Carvalho *et al.*, 2003) using 2 ATP and 1 oxaloacetate molecules, cf. Figure 3-12.



EC Number	Enzyme	Reaction
4.2.1.1	Carbonic anhydrase	$CO_2 + H_2O \rightleftharpoons H_2CO_3$
4.1.1.31	PEPC	Phosphoenolpyruvate + $HCO_3^- \rightleftharpoons p_i$ + oxaloacetate
6.3.1.2	Glutamine synthase	ATP + L-Glu + $NH_3 \rightleftharpoons$ ADP + p_i + L-Gln
1.4.1.13	Glutamate synthase (NADPH)	L-Gln + 2-oxoglutarate + NADPH + $H^+ \rightleftharpoons$ 2 L-Glu + $NADP^+$
2.6.1.1	Aspartate transaminase	L-Asp + 2-oxoglutarate $\rightleftharpoons$ oxaloacetate + L-Glu
6.3.5.4	Asparagine synthase	ATP + L-Asp + L-Gln + $H_2O \rightleftharpoons$ AMP + pp_i + L-Asn + L-Glu
		A) L-Gln + $H_2O \rightleftharpoons$ L-Glu + NH_3
		B) ATP + L-Asp + $NH_3 \rightleftharpoons$ AMP + pp_i + L-Asn

Net reaction using Gln as intermediate:
2 ATP + L-Glu + oxaloacetate + NH_3 + $H_2O \rightleftharpoons$ ADP + p_i + 2-oxoglutarate + AMP + pp_i + L-Asn
Net reaction using NH_3 directly:
ATP + NH_3 + oxaloacetate + L-Glu = 2-oxoglutarate + AMP + pp_i + L-Asn

Figure 3-12 Enzymatic reactions of *L. japonicus* asparagine assimilation. The Enzyme Commission number (EC number), name of the catalyzing enzyme and the chemical reactions are given in the bottom. Net reactions using Gln and NH_3, respectively, are given at the bottom. *L. japonicus* Asparagine synthases are multi-domain proteins that catalyze two reactions (A and B). Major products are given in light blue, enzymes in light blue, and consumption of ATP in red. Both the normal reaction (6.3.5.4) of Asparagine synthase and the alternative reaction where NH_3 is used directly as N donor are shown. PEPC = Phospho*enol*pyruvate carboxylase, p_i = phosphate, and pp_i = diphosphate.

Interestingly, it is also known that Asn synthase also can use NH_4^+ directly as an N donor although the Km value is 40-fold higher (Hirel & Lea, 2001). This provides a second enzymatic step in which NH_4^+ is assimilated. Following this, either an increase in Gln or Asn synthetase activity in nodules could be expected in order to remove the toxic levels of NH_4^+produced in the symbiomes. However, such induction is not observed for any of the 6 transcripts encoding Gln synthases that were detected in the data set. Moreover, the most abundant transcript (TC61595) is even ~7 times down-regulated in the nodule tissues compared to day 0. Furthermore Hirel *et al.* found that overexpression of Gln synthase led to a severe decrease in biomass production and Asn assimilation, implying that Gln synthase activity actually is a negative regulator of nitrogen fixation (Hirel *et al.*, 1997). This is also supported by a study by Harrison *et al.* where it was shown that a reduction in Gln synthetase in *L. japonicus nodules* leads to an increase of the amino acid content in the nodules, primarily due to an increase of Asn (Harrison *et al.*, 2003) and a study by Carvalho *et al.*, who found a negative correlation between Gln- and Asn synthase expression in *M. truncatula* (Carvalho *et al.*, 2003). All this implies that Asn assimilation in the nodules is (at least partly) performed by a different mechanism that does not

involve an increase in Gln synthetase. The observed high induction of Asn synthase along with the high NH_4^+ concentration in the nodule due to the action of the bacterial nitrogenase in the bacteroid could imply that Asn assimilation in the nodules is performed by asparagine synthase by direct use of NH_4^+. This hypothesis is supported by several observations in the data set. Firstly, a large induction of the enzymes involved in the Asn assimilation pathway would be expected. This is observed for the most abundant transcripts encoding Carbonic anhydrase (~ 150 fold up-regulated), Phospho*enol*pyruvate carboxylase (~ 10 fold up-regulated), Aspartate transaminase (~ 9 fold up-regulated), and as mentioned (Asn) synthetase (~50 fold up-regulated). Secondly, also supporting this hypothesis is the fact that the direct incorporation of NH_4^+ by Asn Synthase saves 1 ATP compared to incorporation of NH_4^+ using Gln as intermediate. However, some Gln synthase activity is still needed, due to the need for Asp, which is converted by aspartate transaminase from Glu. Glu is the product of the GS-GOGAT cycle, hereby introducing the need for Gln synthase activity. This is supported by the results of a study by Carvalho *et al.*, who found that an increase in Asn Synthase expression is not maintained when Gln synthesis is completely inhibited (Carvalho *et al.*, 2003). Interestingly, expression of Glutamate synthase the second enzyme in the GS-GOGAT cycle that catalyzes the conversion of Gln to Glu is also induced in nodules. The most abundant transcript, LjSGA_057226.2, is between 3.2 and 20 fold up-regulated compared to the uninoculated state. Carvalho *et al.* estimated that a combined NH_4^+ assimilation both using Gln as intermediate and direct uptake of NH_4^+ by Asn Synthase could save ~ 17 % in energy requirements (Carvalho *et al.*, 2003). A second hypothesis that can be made based on the results of this and other studies is that high concentrations of Gln is a negative regulator of NH_4^+ assimilation. Both Asn Synthase and Glu Synthase were found to be highly up-regulated in the current study. Here the extreme induction of Asn Synthase could be explained by the fact that the activity of the enzyme provides NH_4^+ assimilation under the consumption of Gln in contrast to Gln synthase, which produces Gln. As mentioned this is also well in line with the results that overexpression of Gln synthase led to a severe decrease in Asn assimilation (Hirel *et al.*, 1997). Finally, as seen in Figure 3-12 it is clear that since Asn is the major transport compound and if the conversion of Asp to Asn is the rate limiting step of Asn assimilation (and hereby NH_4^+ assimilation), induction of Asn Synthase would automatically lead to a higher rate of NH_4^+ assimilation.

Besides genes known to be involved in nitrogen fixation the PCA also revealed induction of genes so far unknown to be involved in nodule organogenesis. The 3 transcripts PUT-177a-Lotus_japonicus-52140, TC59466, and AW720139 (sorted by the descending values of PC1 in the PCA, cf. Figure 3-11) were also all highly induced and expressed in the nodule tissues. They were between 23 and 137 fold up-regulated compared to day 0 and accounted for between 1 and 3 % of all transcripts in the nodule tissues. Initial functional annotation of these all resulted in non-informative annotations such as “Putative uncharacterized protein”. However, a second BLASTX similarity search (Camacho *et al.*, 2009; Altschul *et al.*, 1990) indicated that all transcripts were homologue to extensin-like proteins. Extensins are hydroxyproline-rich glycoproteins located in the cell wall, which e.g. have found to be important for the development of root hairs in tomato (Bucher *et al.*, 1997), and play a role in cell wall development and resistance to bacterial infection (Lamport *et al.*, 2011). This could indicate that special nodule cell wall metabolism processes are maintained in mature nodules to maintain the structure of this tissue.

Secondly, a more focused analysis of the earlier time points in the wild type time course (0-48 hours AMI) was performed using PCA to elucidate changes in the *L. japonicus* transcriptome early in nodulation, cf. Figure 3-13.

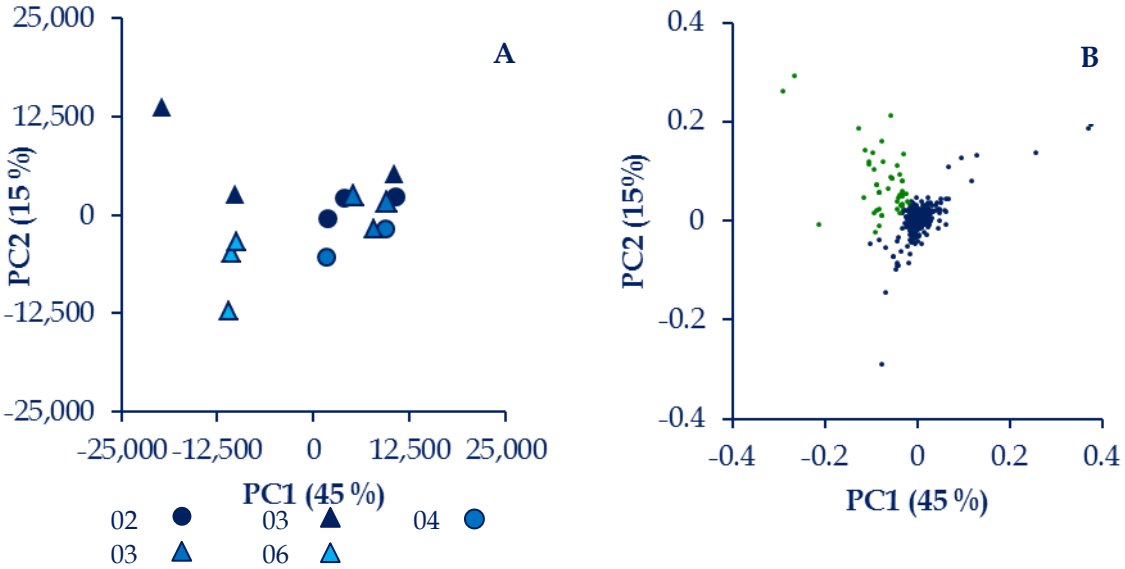


Figure 3-13 Transcriptome of *L. japonicus* early in nodulation. PCA plot of samples from the wild type time series ranging from 0-48 hours after inoculation with *M. loti* strain MAFF 303099 (AMI). **A:** Scores plot of PC1 vs. PC2 showing the grouping of samples originating from 8 hours (top left) and 48 (bottom left) AMI. **B:** Loadings plot showing the splitting of genes. The top 50 genes nearest to the top left corner, which are highly expressed 8 AMI, are marked in green. PC = Principal component. Numbers in percentages indicate variance explained by the PCA model.

Once again, the PCA was able to highlight interesting expression patterns. The samples 8 hours AMI split out in the top left corner in the scores plot, cf. Figure 3-13 panel A. Genes regulated very early in the nodulation process, are potential candidates for being involved in the nodulation pathway. Therefore, the cause of this split was investigated further. Genes identified by the PCA to be specifically up-regulated 8 hours AMI and subsequently down-regulated are shown in Figure 3-14.

Figure 3-14 Genes identified by PCA to be specifically up-regulated in the root window at 8 hours AMI. The 50 genes most correlated with samples 8 hours AMI are listed. Genes are ranked by the distance to the top left corner in the loading plot, which indicate high correlation cf. Figure 3-13B. The heat map shows the relative expression early in the time series (0-72 hours) for each gene compared to the maximum expression level in the entire time series. Genes marked in red are related to defense response, genes marked in green are related to cell wall development, and genes marked in blue can related to both defense response and to cell wall development. * Root infection zone. ** Root and nodules.

Based on the functional annotation, it is clear that two processes are dominating the *L. japonicus* transcriptome at 8 hours, namely a defense response and cell wall synthesis/development. Since extensins are known to be a part of both defense and root hair morphology, their up-regulation is interesting. It could indicate that the plant recognizes the bacterial infection and starts a defense response, which is subsequently shut down, possibly due to rhizobial signaling. On the other hand the extensin reduction could be a part of the root hair development indicating that changes in the cell wall, which is needed for bacterial infection and nodule organogenesis are already occurring.

3.2.8 Data Analysis of Mutant Time Series Samples

The investigation of the time series of the wild type transcriptome revealed several interesting changes. In order to further elucidate the nodulation process, the gene expression of genes found to be regulated in the wild type time series were investigated in the time series of the two mutants: *nfr5* and *nin*. Furthermore the gene expression of genes found to be involved in the nitrogen fixation in mature wild type nodule tissue were compared to the gene expression in nodule tissue in the two mutants: *snf1* and *sym11*. Firstly, the transcriptome profiles in *nfr5* and *nin* of genes known to be involved in nodule organogenesis and/or bacterial infection were investigated and compared to the transcriptome profile of the wild type plant, cf. Figure 3-15. NFR5 was found to be induced at 8 and 16 hours AMI in the wild type plant. The same expression pattern is seen for both the *nfr5* and *nin* mutants. However, the expression level in the *nfr5* mutant is significantly lower than the expression levels in the wild type and *nin* genotypes throughout the entire time series (student's t-test p-value = $1.7 \cdot 10^{-3}$ and $1.5 \cdot 10^{-3}$, respectively). This could indicate that the activation of NFR5 positively regulates the expression of the *NFR5* gene since the NFR5 protein in the *nfr5* mutant lacks the activation domain. The Nucleoporin NUP133, and the cation channel Castor were found to be induced between 16 and 24 hours AMI in the wild type plant. Interestingly, the same gene expression pattern can be observed in both the *nfr5* and the *nin* genotype. This could indicate the induction of these genes is independent of nod factor signaling, since they are also induced in the *nfr5* mutant. The expression pattern of *NIN* was similar in all three genotypes, cf. Figure 3-15. However, the expression was observed to be 2.7±0.5 fold less in the *nfr5* mutant at 8 and 16 hours AMI compared to the wild type and *nin* mutant. This could indicate that nod factor signaling is a positive regulator of *NIN* expression, which then further drives the nodulation process. An interesting gene expression pattern was found for the *LHK1* gene in the mutant genotypes, cf. Figure 3-15. An induction co-regulated with *NIN* expression was observed at 8 and 16 hours in the wild type plant. Surprisingly, this is also observed in the *nfr5* mutant, although *LHK1* expression is a part of the cytokinin and auxin signaling occurring downstream of calcium oscillations, which does not occur in the *nfr5* mutant. The expression of *LHK1* in the *nin* mutant is not differentially regulated between 0 and 48 hours, but is continuously highly expressed compared to the wild type and *nfr5* mutant. This could indicate that expression of *NIN* is a part of the regulation of *LHK1*, and that timing of *LHK1* expression is important for successful nodule organogenesis. Another interesting observation is the early induction of Castor, and especially Pollux in the wild type plant, which is lacking in both mutant plants, cf. Figure 3-15. Charpentier *et al.*, concluded that quantitative gene expression of the two ion channels Castor and Pollux plays a critical role in modulating the nuclear envelope membrane potential (Charpentier *et al.*, 2008). These are therefore likely to play a part in the signaling pathway involving calcium spiking, and the lack of early induction of these genes in the mutant plants could therefore be a key factor to their non-nodulating phenotype.

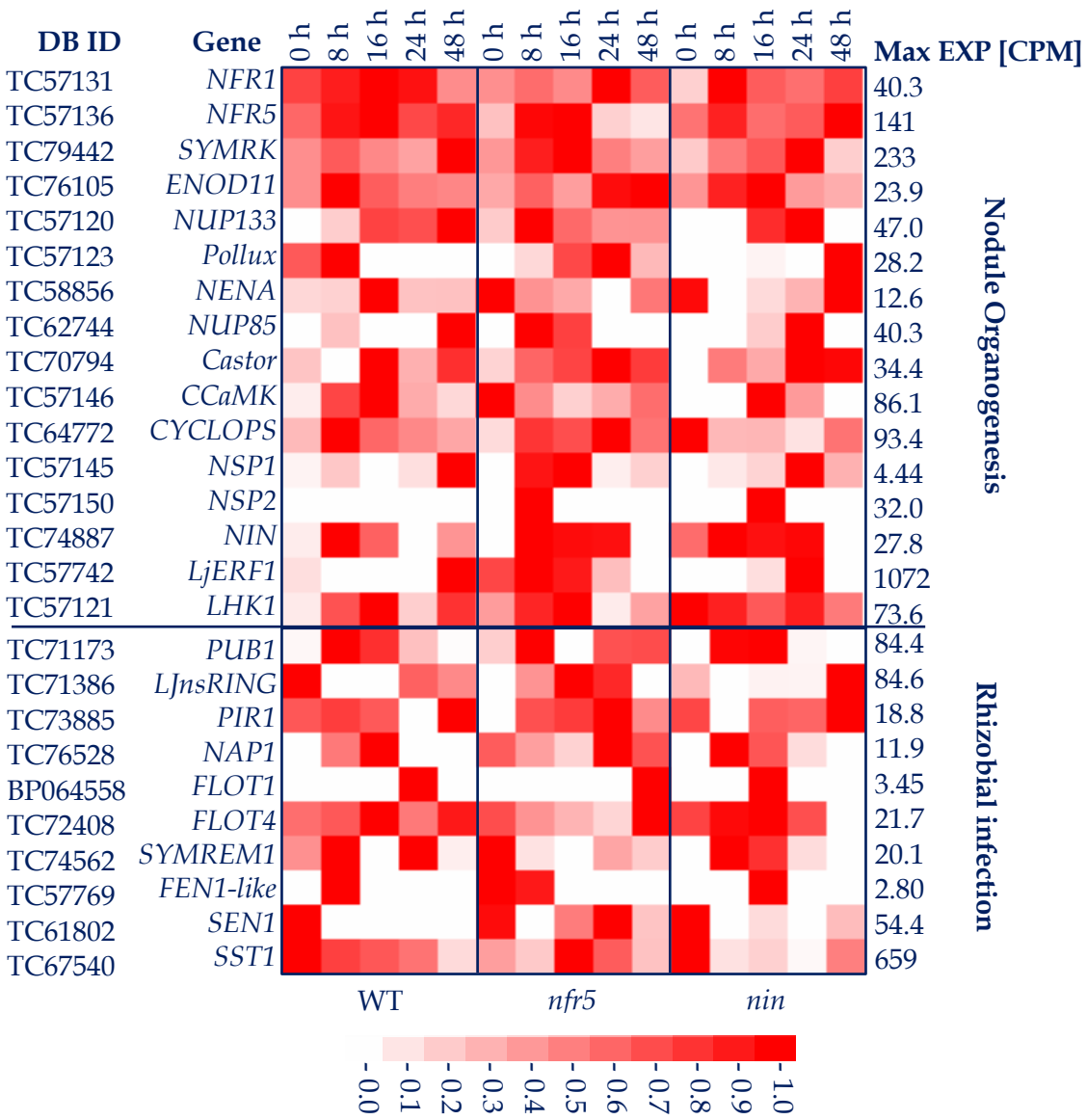


Figure 3-15 Relative mRNA expression levels of transcripts encoded by genes known to be a part of *L. japonicus* nodule organogenesis. Transcriptome profiles from 0 to 48 hours AMI from the wild type (WT), *nfr5*, and *nin* genotypes are shown. Sequence ID and gene name is given on the left. The maximum expression level measured as counts per million (CPM) is given on the right. The expression level is relative to the maximum observed value for each transcript indicated by the color bar in the bottom.

When investigating the genes known to be involved in bacterial infection, only *PUB1* was found to be induced within the first 48 hours AMI in the wild type plant, cf. Figure 3-15. The expression of this gene in the mutant genotypes are well in line with the fact that *PUB1* has found to phosphorylated by the NFR5 homologue in *M. truncatula* (Mbengue *et al.*, 2010) since *PUB1* is induced in the *nin* mutant, but does not seem to be induced in the *nfr5* mutant. This could indicate that *PUB1* expression is regulated by nod factor recognition, and although that the expression of the gene has been found to be a negative regulator of bacterial infection; it is induced early after nod factor recognition. Mbengue *et al.* speculated that *PUB1* could be a key regulator of downstream signaling of *NFR5* (Mbengue *et al.*, 2010). This hypothesis is well in line with the gene expression pattern found in the current study.

The gene expression of genes identified by the PCA to be specifically up-regulated 8 hours AMI and subsequently down-regulated in the wild type plant was further investigated in the

mutant plants. Here, the same clear induction of genes involved in defense and cell wall metabolism at 8 hours AMI could not be observed, cf. Figure 3-16. On average, the 50 genes most correlated with the wild type sample 8 hours AMI were 3.4 fold up-regulated at 8 hours AMI and subsequently 5.0 fold down-regulated at 16 hours AMI. By visual inspection of the heatmap in Figure 3-16, it is clear that the expression pattern found in the *nin* mutant is more similar to that of the wild type plant compared to the expression pattern found in the *nfr5* mutant. Although the same sharp up-regulation at 8 hours and subsequent down-regulation at 16 hours seen in the wild type plant cannot be observed in the *nin* mutant, many genes are induced at 16 and 24 hours and subsequently repressed at 48 hours, although not to the same extent as seen in the wild type plant. The expression pattern seen in the *nfr5* mutant is more random since a lot of the genes are slightly up- or down-regulated; not following the specific expression pattern seen in the wild type plant. These results indicate that Nod-factor signaling is important for controlling the observed defense and cell wall metabolism response seen in the wild type, since the gene expression pattern is dependent on nod factor recognition by *NFR5*. The results also indicate that *NIN* expression could be a part of a signaling pathway controlling the timing of the observed response. Furthermore, the results indicate that the induction of extensins is not a part of a less specific defense response towards the invading bacteria, but is more likely to reflect coordinated cell wall development.

Lastly, the gene expression in nodule tissue 21 days AMI was compared between the wild type, and the *snf1*, and *sym11* mutants. Due to the fact that no global expression profiles were available for the mutant genotypes taken at the time of *M. loti* inoculation, it was not possible to state conclusions regarding induction or repression of genes in the mutant genotypes. However, direct comparisons of absolute gene expression levels between the wild type and the mutant genotypes were possible. Firstly, when comparing the gene expression profiles globally, the *sym11* profile is more similar to the wild type compared to the *snf1* mutant (ρ_s = 0.59 and 0.69, respectively). In fact, the gene expression profile of the *snf1* nodules is more similar to the wild type root window sample at the time of *M. loti* inoculation (ρ_s = 0.64). Furthermore, the expression levels of the genes found by PCA to be indicative of N_2 fixation in the wild type plant, cf. Figure 3-11, and genes known to be involved in asparagine synthesis, cf. Table 3-6, nearly all had expression levels similar to that of the wild type root window sample at the time of *M. loti* inoculation. This indicates that even though the *snf1* mutant spontaneously develops nodules without rhizobial infection, the transcriptome of effective nodules infected with *Rhizobia* is highly influenced by the presence of the bacteria. However, a few genes seemed to be induced in the *snf1* mutant nodules, based on the fact that their relative high expression was in the same range as that of the wild type nodules. Interestingly, several transcripts annotated as nodulins (e.g. TC62028, TC68196, TC61802, and TC71278) were among these. This could imply that the expression of these genes is involved in processes controlling and maintaining the nodule structure and that their expression is induced downstream of *CCaMK* expression, not affected by the lack of bacterial infection. Supporting this theory is the fact that the three transcripts encoding extensin-like proteins found to be induced in wild type nodules (PUT-177a-Lotus_japonicus-52140, TC59466, and AW720139) were expressed at similar levels in the *snf1* mutant.

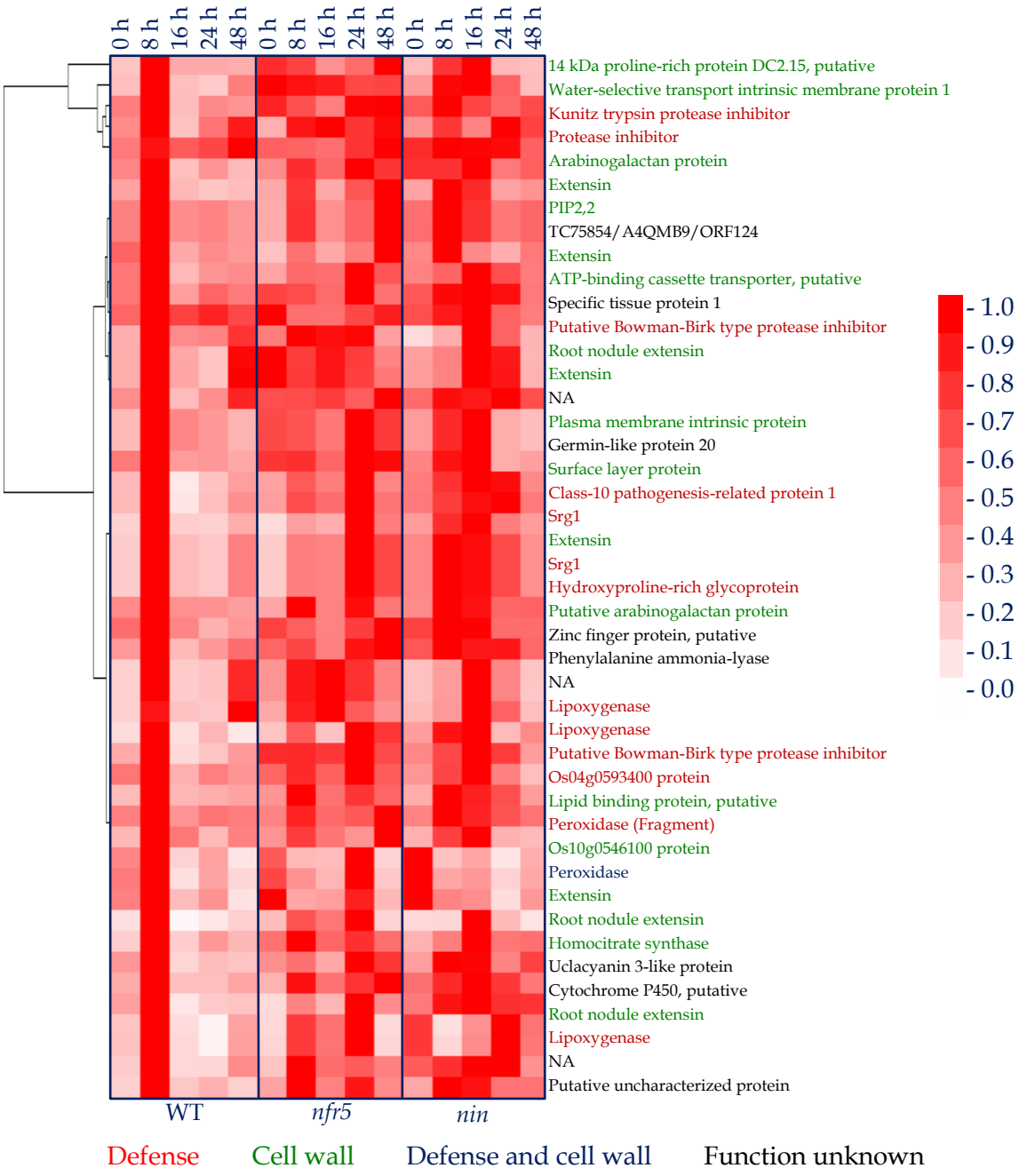


Figure 3-16 Genes identified by PCA to be specifically up-regulated 8 hours AMI in the wild type time series. The 50 genes most correlated with samples 8 hours AMI are listed. Genes are ranked by the distance to the top left corner in the loading plot, which indicate high correlation cf. Figure 3-13 panel B. The heat map shows the relative expression early in the time series (0-48 hours) for each gene compared to the maximum expression level in the wild type (WT), *nfr5*, and *nin* genotypes. Genes marked in red are related to defense response, genes marked in green are related to cell wall development, genes marked in blue can related to both defense response and to cell wall development.

SEN1 gene expression is required for bacterial differentiation into nitrogen-fixing bacteroids in *L. japonicus* nodules (Suganuma *et al.*, 2003). Interestingly, *SEN1* gene expression is induced 7 days AMI in the wild type plant and is constitutively expressed hereafter, cf. Figure 3-10. No *SEN1* gene expression could be detected in nodules 21 days AMI in the *sym11* mutant harboring a mutation in the *SEN1* gene. This implies that the mutation in *sym11* most likely has caused the gene to be silenced (e.g. by a mutation in the gene promoter region). Alternative, the mutation can have caused alterations in the gene structure (e.g. by a mutation in an in-

tron/exon boundary) causing the normally indicative Deep SAGE tag of the *SEN1* transcript to be omitted during mRNA splicing.

Table 3-6 Absolute expression levels of genes known to be a part of asparagine synthesis. Expression levels are given as CPM of the two most abundant transcripts of each enzymatic reaction in wild type (WT) nodules are given. Expression values indicating a 2-fold or higher induction compared to the WT root window sampled are marked in bold and red. PEPC = Phospho*enol*pyruvate carboxylase.

EC	ID	Name	WT (02)	WT (12)	snf1 (25)	sym11 (26)
2.4.1.13	TC69747	Sucrose synthase	774	**1,833**	242	539
	TC80030		776	**1,828**	242	539
4.2.1.1	TC70132	Carbonic anhydrase	82	**1,843**	709	**1,395**
	TC61311		81	**1,820**	669	**1,342**
4.1.1.31	TC73893	PEPC	321	**642**	13	88
	TC57636		143	116	168	145
6.3.1.2	TC61595	Glutamine synthetase	746	104	309	271
	TC74374		8	**71**	**71**	**129**
1.4.1.14	LjSGA_057226.2	glutamate synthase	12	**41**	0	13
	TC76734		8	**22**	0	13
2.6.1.1	AW719338	Aspartate aminotransferase	50	**349**	77	**258**
	TC63353		18	**156**	0	0
6.3.5.4	TC57163	Asparagine synthetase	885	**44,668**	551	1,398
	BP077614		7	**385**	0	13

The *sym11* mutant has a pinkish nodule phenotype, which is believed to be caused by leghemoglobin synthesis (Sandal *et al.*, 2006; Schauser *et al.*, 1998). The data of this study supports this. On average, Leghemoglobin transcripts were found to be between 10 and 27 times more abundant in wild type nodules compared to the wild type root sample at the time of *M. loti* inoculation (data not shown). In the *sym11* mutant, Leghemoglobin transcripts were found to be between 2.5 and 6.4 times higher compared to the wild type root sample at the time of *M. loti* inoculation (data not shown). This implies that an induction of leghemoglobin expression has occurred in the *sym11* mutant, however not to the same degree as in wild type nodules. Like the *snf1* mutant, expression of transcripts likely to be involved in cell wall metabolism was found at similar levels compared to the wild type plant. The function of *SEN1* is unknown, but it is predicted to have a transport function, and it has been shown that *SEN1* gene expression is involved in symbiosome development (Hakoyama *et al.*, 2012; Suganuma *et al.*, 2003). Therefore, it was not surprising that several genes involved in asparagine synthesis, and hereby nitrogen fixation, were not induced in *sym11* nodules, cf. Table 3-6, but were found to be expressed at similar levels as the wild type root window sample at the time of *M. loti* inoculation. These results imply that *SEN1* expression is necessary for regulating the transcriptome associated with Asn synthesis at a late stage in nodule organogenesis. Moreover, several transcripts matching nodulins (e.g. TC68196, TC62028, TC74943, and TC64787), and genes involved cell wall metabolism (e.g. PUT-177a-Lotus_japonicus-52140, TC59466, and AW720139) found to be induced in the wild type plant were also induced in the *sym11* mutant. However, the expression of other nodulins transcripts (e.g. TC63124 and TC61802) were found at similar levels as the wild type root window sample at the time of *M. loti* inoculation, hereby indicating that these nodulins genes are part pathways affected by *SEN1* expression. The results indicate that the expression of genes involved in cell wall metabolism are not

affected by the lack of *SEN1* expression, but the development of the N_2 fixation machinery is. This makes the *SEN1* gene and the genes encoding nodulins, which were found to be affected by the lack of *SEN1* expression, to be candidate genes for members of the signaling pathway in the development of the N_2 fixation machinery.

3.3 Summary and Conclusions

3.3.1 Challenges of the Data Analysis

The gene expression study of *L. japonicus* wild type and mutant plants during nodulation was the first large-scale study performed at AAU using the DeepSAGE technology. Several experiences were gained and methods developed during the time course of the data analysis. Firstly, substantial efforts were made in order to lower the data complexity and hereby facilitate data analysis. As described in section 2.1, data preprocessing step such as singleton removal and SAGEscreen error correction significantly reduces the data complexity while retaining most of the data. In the case of the *L. japonicus* transcriptome data set, which comprised of 71 samples in total, 221,404 different tags were observed. After data preprocessing 115,858 different tags were retained, hereby obtaining a 50 % reduction in data complexity. Moreover, tag counts were summed to each gene resulting in only 39,902 different genes, hereby reducing the data complexity to 18 % of the original value. An additional advantage to this approach was that the large variation found for lower abundant secondary tags (e.g. caused by incomplete NlaIII digestion) was avoided. In total, 91 % of all tag counts were retained after all data preprocessing steps, showing that the developed methods for data preprocessing facilitate lowering of the data complexity while retaining most of the data.

A major challenge in the data analysis was the insufficient annotation of the *L. japonicus* genome. The genome sequence of an organism can potentially provide useful information for transcriptome studies. However, to transfer the information hidden in a genome sequence to transcriptomics studies, a high quality gene annotation of the genome sequence is required. Currently, this is not the case for the *L. japonicus* genome sequence. The annotation is built entirely on *ab initio* predictions of coding sequences, hereby failing to provide information of the un-translated regions. The current study exemplifies that annotation of the un-translated regions is crucial if the genome sequence is to be useful for tag based transcriptome studies. In the current study an *in silico* approach was used to predict the average length of the 3' UTR, which then was subsequently used to improve the gene annotation. In section 5.2.2, more refined methods to predict the un-translated regions using mRNAseq will be presented. A second challenge in the data analysis was the lack of functional annotation of the sequence collections used. High quality functional annotation of the transcriptome of an organism is the second step needed to facilitate more advanced data analyses such ontological assisted data analysis. Hence, efforts were made to improve (or in the case of the sequence collection based on the genome sequence – provide) the gene ontology (GO) annotation of the sequence collection. GO terms are available for the tentative consensus sequences from TIGR. However, only 19 % of the sequences have annotations for a biological process, and GO terms were not available for any of the other sequence collections used. Therefore, gene ontologies were constructed for the sequence collections used by transferring the GO term assigned to the transcripts functional annotation[19], providing functional categories to the Kazusa-CDS and PlantGDB PUTs, and improving the annotation of the TIGR TCs (54 % of the sequences acquired a GO term). Despite these efforts, subsequent ontological assisted data analyses did not provide further insight to the nodulation process, possibly due to the lack of sufficient functional annotation of the genes involved. Therefore, the details of this

[19] The Uniref100 GO annotation is available at: ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/gp_association.goa_uniprot.gz.

analysis are not presented in the current thesis. However, using the Cytoscape (Shannon *et al.*, 2003) plugin BiNGO (Maere, Heymans & Kuiper, 2005), it was possible to assign a significance criteria for the overrepresentation of cell wall metabolism genes found to be specifically induced 8 hours AMI in the wild type plant, cf. Figure 3-14. It was found that "cell wall organization or biogenesis" (GO term 71455) was overrepresented (P-value = $2.3 \cdot 10^{-5}$, FDR corrected). This shows the potential of ontological assisted data analysis, since this pattern was not detected using a method relying on pair wise testing.

During the course of the data analysis, the integrity of the data set was questioned due to the large differences in library size and the discovery of 5 mis-labeled samples. However, the results of the comparison between the DeepSAGE and Affymatrix data showing correlation at expectable levels between the data sets provided evidence for the data integrity of the remaining parts of the data. The discovery of mis-labeled samples was originally based on the results of a PCA analysis. This showed grouping of replicates originating from widely different samples, while most replicates grouped nicely according to the sample origin (data not shown). Following this, correlation analyses were performed to identify 5 mis-labeled samples, which then could be excluded from the data set. PCA analysis was found to be an excellent method to provide first glance insight into the data, and at the same time facilitate detection of outlier samples.

Initially, an analysis strategy based on detection of differentially expressed genes between the un-inoculated state and the different states AMI. The statistical model behind the Z-test was initially chosen for determination of differential expression. However, it quickly became evident that the Z-test fails to take into account that the data set is overdispersed. This resulted in a very large number of genes to be called as differentially expressed. In the wild type time series, more than 30 % of the observed genes were determined to be differentially expressed between the un-inoculated state and a state AMI. The majority of these were found to be lowly expressed and often poorly determined, why this strategy was abandoned. The poor determination was found to be partly caused by the large differences in library size, since detection limit varied between the libraries. Following development in statistical models for detection of differentially expressed genes, the pairwise test implemented in EdgeR (Robinson, McCarthy & Smyth, 2010) was attempted (data not shown). By taking gene wise dispersion into account, the detection of differentially genes was found to be more robust, which was concluded based on the fact that most lowly expressed and poorly determined genes called to be differentially expressed using the Z-test were not called as differentially expressed using EdgeR. However, a large fraction of genes found to be a part of expression patterns detected using PCA (e.g. the induction of defense and cell wall metabolism genes at 8 hours AMI in the wild type time series), was not detected as significantly differentially expressed. Although a total of 1,401 genes were found to be differentially expressed in the wild type time series, most of these was regulated at the later stages of nodulation, reflecting the large phenotypical differences between the samples compared. In fact, no genes were found to be differentially expressed at 8 hours AMI in the wild type time series using EdgeR. This highlights why a traditional pairwise comparison strategy can fail to elucidate biological meaningful changes in the transcriptome, since it relies on gene wise testing. The P-value threshold for significant differential expression is selected arbitrarily and is also affected by e.g. choice of correction method for multiple testing (which often becomes more stringent, when more genes are observed). The choice of significance threshold is often chosen so a managea-

ble number of genes are called as differentially expressed. However, whether these changes are biological relevant is not certain. As a consequence, expression patterns involving insignificant changes (when compared using a pairwise test) of a large set of genes that combined are indicative of a difference between two samples may fail to be elucidated. In the current study, PCA was successfully used several times, elucidating gene expression patterns that gave biological meaning, potentially highlighting key components of *L. japonicus* nodulation.

3.3.2 New Insight to the Nodulation of *Lotus Japonicus*

Despite the challenges described, the current study of the *L. japonicus* transcriptome did elucidate novel insight into nodulation and identified new candidates for key genes of the nodulation signaling pathway and the regulation of N_2 fixation in mature nodules. Perhaps most interestingly, is the discovery of the massive induction of an Asparagine synthase in mature *L. japonicus* nodules. The gene if this is a major candidate for being a key regulator of the symbiotic N_2 fixation in *L. japonicus* nodules, why further studies (such as knockout or overexpression studies) should be performed in the future. The experimental setup for the current transcriptome study was designed to elucidate all stages of the nodulation process. However, this was not achieved. Initially, the analysis strategy for elucidating the transcriptome of *L. japonicus* during nodulation was to compare the wild type transcriptome with that of the *nfr5* mutant, which is arrested very early in the nodulation and that of the *nin* mutant, which is arrested at a later state subsequent to calcium oscillations. Hereby, it seemed possible to detect genes involved in the early parts of the signaling pathway (by detecting genes that were regulated in wild type and the *nin* mutant), and genes involved later in the signaling pathway downstream of *NIN* expression (by detecting genes that were regulated only in wild type). However, even though the first sample was collected only 8 hours AMI, the data indicates that induction of *NIN* already has occurred, cf. Figure 3-10. As a consequence, this study to some extent fails to elucidate the nodulation signaling pathway prior to NIN expression. As an example, it is clear that the observed induction of defense and cell wall metabolism genes in the wild type plant is not found the *nfr5* mutant, making this a potential early response prior to *NIN* expression, cf. Figure 3-16. However, although the induction seems lower and at a later state (8-16 hours AMI in the *nin* mutant instead of 8 hours in the wild type), it can to some extent be found in the *nin* mutant, cf. Figure 3-16. Therefore, it is inconclusive whether the observed response is prior to *NIN* expression or to some extent is affected by the expression of the *NIN* gene. To elucidate this, a more focused experimental setup is needed, e.g. with sampling every hour for the first 16 hours AMI. In the case of the experimental setup for the comparison of the wild type plant and the *snf1* and *sym11* mutants, only a simple pairwise analysis was possible, hereby only reflecting the resulting transcriptome caused by the mutated genes. In the case of the *snf1* mutant, the comparison enabled detection of genes that were induced solely by *CCaMK* expression, not dependent on rhizobial presence. Several nodulins and genes likely to be involved in controlling and maintaining the nodule structure were detected, hereby implying that *CCaMK* expression is a key component in the signaling pathway of these processes. The analysis of the transcriptome of *sym11* nodules indicated that *SEN1* expression is necessary for the development of the N_2 fixation machinery but not for controlling and maintaining the nodule structure. Furthermore, some nodulins that are potential candidates for member of the signaling pathway controlling the development of the N_2 fixation machinery were identified. The current study has only provided a snapshot of the *L. japonicus* transcriptome at the late stages of nodulation,

why further investigations are needed. A detailed time series experiment of the *sym11* mutant (or other mutants arrested late in nodulation) is a possibility. Here, the developmental stage at which mutants are arrested in nodule development could be pinpointed more precisely, and more members of the nodulation signaling pathway could possibly be identified.

Chapter 4

Analysis of Variance in Tag Based Transcriptome Data

4.1 Introduction to the Analysis of Variance in Tag Based Transcriptome Data

4.1.1 Noise in Gene Expression Data

Transcriptome data sets are by nature "noisy", i.e. variation in the observed gene expression levels exist between samples. Overall, the observed variance can be divided into two major sources, namely technical and biological variation (Chen *et al.*, 2004). Both of these broad categories can be further divided into several sources of variation.

In the following, the definition of biological variance by Bartlett is used, i.e.: "*A component of the variance in biochemical measurements determined by the physiology of the subjects observed*" (Bartlett, 1999). Several levels of biological variance exist. Firstly, variation between different biological groups is found. Depending on how these biological groups are defined, the elucidation of differences in the transcriptome reflected as phenotypical differences between these groups is most often the goal of a transcriptome study. Comparison of biological groups can range from species vs. species, within the same species (e.g. different tissues or developmental stages) down to comparison of gene expression profiles at the single cell level. Even at the single cell level, several sources of biological variance can be considered (Elowitz *et al.*, 2002; Kuznetsov, Knott & Bonner, 2002; Swain, Elowitz & Siggia, 2002). In this regard, Raser and O'shea identified four potential sources of variation, namely 1) the inherent stochasticity of biochemical processes 2) variation caused by differences in the internal states of a population of cells (e.g. due to cell cycle progression) 3) subtle micro environmental differences (e.g. morphogen gradients), and 4) ongoing genetic mutations (Raser & O'Shea, 2005). Moreover, they defined "intrinsic noise" as local within a cell, and "extrinsic noise" that causes differences between two cells (Raser & O'Shea, 2005). However, the current thesis focuses on biological variation at a more global scale; namely between biological replicates of samples originating from the same species and tissue type. Therefore, "intrinsic noise" will refer to differences observed between samples originating from the same biological replicate (e.g. one leaf from a single plant), and "extrinsic noise" will refer to differences observed between samples originating from differences between biological replicates (different plants).

Both microarray and sequence based transcriptome studies are multi-step produces where each step (which in both cases broadly can be divided into sampling, RNA purification, library preparation and measurement) is a potential source of noise. Several studies have investigated sources of technical variation of the microarray technology (Chen *et al.*, 2004; Novak, Sladek & Hudson, 2001), and other studies have described sequence based technologies in regards to reproducibility and accuracy (e.g. studies by Matsumura *et al.* and Nielsen *et al.* based on tag based sequencing (Nielsen, Hogh & Emmersen, 2006; Matsumura *et al.*, 2005) and the work by Mortazavi *et al.* based on RNAseq (Mortazavi *et al.*, 2008)). Moreover, several studies comparing the different methods have been conducted, such as the comparison between RNAseq and microarray (Bradford *et al.*, 2010; Marioni *et al.*, 2008) and comparisons between tag based sequencing methods and microarray (Ishii *et al.*, 2000). Due to the cost of transcriptomic experiments, these studies have been performed using a minimum number of replicates (either technical or biological depending on the experiment). The reduction of replication adversely affects the estimation of the gene expression, and hereby also affects the ability to determine differential expression (Kendziorski *et al.*, 2003).

Following the price of transcriptomic experiments, biological samples have often been pooled to reduce the effect of biological variance. Kendziorski *et al.* investigated how many biological samples that should be pooled and how many technical replicates of this were needed to gain the same quality in the estimation of the gene expression level as in a non-pooling experiment. They concluded that pooling of biological samples is an advantage in regards to estimation of gene expression, especially when the biological variance between samples is large compared to the technical (Kendziorski *et al.*, 2003).

4.1.2 The Current Study – an Analysis of Variance

In the following, an analysis of variance in tag based transcriptome data will be presented. Here, three data sets will be used. The first dataset contains two biological groups, each consisting of 47 libraries. The libraries originate from leaf tissue of two different field grown potato cultivars (cv. Kuras, and cv. Kardal). Each library represents a true biological replicate (they originate from different plants). This experiment was designed to investigate the overall variance (both technical and biological) of DeepSAGE data sets.

The second data set consists of 3x3 DeepSAGE libraries and 3x3 mRNAseq libraries. Here, tissue from three leaves from three different plants was homogenized and subsequently divided into three samples. From these nine samples, RNA was purified, where after one half of the sample was used for DeepSAGE library preparation, and the other half was used for RNAseq library preparation, hereby creating 9 DeepSAGE and 9 mRNAseq libraries. This experiment was designed to investigate the contributions of intrinsic noise (variance observed between samples originating from the same leaf) and extrinsic noise (variance observed between samples originating from different leaves) to the overall variation in DeepSAGE and mRNAseq data sets, respectively.

Finally, the variation was investigated in a collection of several large scale DeepSAGE data sets consisting of more than 2,000 samples and more than 1.2 billion tags in total. The total collecting has been made as a part of a research program named "*Developing potato into a high-efficient, low-maintenance and multipurpose crop*" (from her on referred to as the large scale DeepSAGE project (LSDS-project)[20]. The collection consists of samples originating from leaf and tuber tissue of either field grown or greenhouse grown *S. tuberosum* plants from 14 different modern cultivars, cf. Figure 4-1. For each time point, cultivar, and treatment three biological replicates originating from different plants were sampled. This analysis was aimed to elucidate effects of technical and biological variance on the ability to detect differentially expressed genes between different biological groups in a larger scale.

[20] The research program is funded by The Danish Council for Strategic Research, Programme Commission on Health, Food and Welfare, Grant 2101-07-0116

Yield

Drought resistance

Late blight resistance

Series	Purpose of study	Year	Growth condition	# samples	Total size of libraries [million tags]
2	Yield	2008	Field	280	203
3	Drought resistance	2008	Controlled field conditions*	672	341
6	Drought resistance	2009	Controlled field conditions*	185	271
4	Late blight resistance	2008	Green house	485	228
5	Late blight resistance	2008	Green house	413	117
Total				2,013	1,221

Figure 4-1 DeepSAGE Data sets a part of the research program named "*Developing potato into a high-efficient, low-maintenance and multipurpose cop*" used for data analysis of variance of tag based transcriptome data. * Plants for the drought resistance data set were field grown with controlled watering during the measurement period.

4.2 Methods

4.2.1 Plant Material and Library Preparation

Leaf and tuber tissue were collected from 3 non-neighboring plants. For each biological replicate, leaf tissue was represented by the 3rd and 4th fully developed leaf, and either 5 whole tubers (if the tuber < 20 mm) or 5 10x10x10 mm^3 pieces (if the tuber > 20 mm) were collected for tuber tissues. Samples were immediately stored in liquid N_2. RNA purifications were performed using the Ambion® RNAqueous® Kit (Life Technologies) according to the manufacturer's instructions.

4.2.2 Tag based sequencing library preparation and data pre-processing

On average, 2 µg of total RNA per sample were used for the DeepSAGE library preparation (Petersen, 2008). Samples were diluted to a final concentration of 10 nM and sets of 12 samples with different identification keys were pooled and subsequently sequenced for 36 cycles on either a Genome Analyzer or a Genome AnalyzerIIx according to the manufacturer's instructions. Image analysis and base calling was performed using the GAPipeline version 0.3 or 1.5.1 software omitting chastity filtering, otherwise default settings. Tag lists were generated using the automatic pipeline described in section 2.1. No sequence error correction was performed. To facilitate data storage and retrieval, tag libraries for the LSDS-project were subsequently renamed and uploaded to a database using *sampleNameConversion.pl* and *DatabaseUpload.pl*. Library nomenclature was made according to the scheme outlined in Table 4-1.

Table 4-1 Tag libraries in the LSDS-project were named using the nomenclature: "*Sort.Harvesttime.Replica.Tissue.Location.Series.Type*". Possible nomenclatures for each part are listed. An example is 10.4.2.2.K.2.1.

ID part	ID Nomenclature	ID explanation	
Sort	# [1-14]	1 = Bintje 2 = Desiree 3 = Dianella 4 = Ditta 5 = Jutlandia 6 = Kardal 7 = Karnico	8 = Kuras 9 = Matador 10 = Toluca 11 = Sarpo Mira 12 = Signum 13 = Spunta 14 = 97-HGP-01
Harvest time	#	Time of harvest in [Hours] or [Days]	
Replica	# [1-5]	Replicate number	
Tissue	# [1-3]	1 = leaf 2 = tuber 3 = root	
Location	A, K or V	K = KMC, Brande A = AKV, Langholt V = LKF, Vandel	
Series	# [2-6]	2 = Yeild-KMC-2008 3 = Drought-Vandel-2008 4 = Late_blight-Vandel-2008_1 5 = Late_blight-Vandel-2008_1	
Type	# [1-3]	1 = control 2 = drought 3 = Late blight	

Tags were annotated using a sequence collection consisting of predicted mRNA transcripts in the genome sequence version 3.4 of *S. tuberosum* Group Phureja DM1-3 516R44[21] (PGSC mRNAs), tentative consensus sequences from TIGR *S. tuberosum* Gene Indices version 13 (Quackenbush *et al.*, 2000)[22] (TCs), and PlantGDB-assembled unique transcripts from PlantGDB version 157a (Dong, Schlueter & Brendel, 2004)[23] (PUTs). To reduce redundancy, TC and PUT sequences were compared to PGSC mRNAs using blastN (Camacho *et al.*, 2009; Altschul *et al.*, 1990). TCs and PUTs Sequences with a significant hit (E-value < $1 \cdot 10^{-30}$) adopted the ID of the matching PGSC mRNA sequence, hereby creating a sequence collecting with multiple sequences with the same ID. Tags with no annotation were subsequently removed and not used in further analyses. Hereafter, expression values were calculated for each gene by summing matching tag counts using *Tag2GeneCounts.pl*.

4.2.3 mRNAseq Library Preparation and Data pre-processing

1-2 µg of total RNA per sample were used for mRNAseq library preparation according to the manufacturer's instructions (Cat # RS-930-1001 Rev. D, Illumina Inc.). Each library was sequenced for 72 cycles in one lane on a Genome AnalyzerII$_x$ according to the manufacturer's instructions. Image analysis and base calling was performed using the GAPipeline version 1.5.1 software with chastity filtering (default settings). Reads were mapped to the potato genome (*S. tuberosum Group Phureja* DM1-3 516R44 (CIP801092) Version 3 DM, Version 2.1.9), with Minimum 80 % similarity in 90 % of the read length. To facilitate data analysis by the R package EdgeR (Robinson, McCarthy & Smyth, 2010), expression values were calculated as number of reads mapped to each gene. Read mapping and calculation of expression values were performed using the CLC Genomics Workbench v 4.8.

4.2.4 Comparison of High Replicate Biological Groups

Two biological groups, each consisting of 47 DeepSAGE libraries originating from leaf tissue from different plants of two potato cultivars (cv. Kuras, and cv. Kardal) were compared. Firstly, a PCA of the gene expression of all 96 samples was performed using The Unscrambler X.1. Data was mean centered and Pareto scaled prior to analysis. Differential gene expression (DE) between the two groups was determined in three ways; using the Z-test, or using the exact test (Robinson & Smyth, 2008) implemented in the R package EdgeR (Robinson, McCarthy & Smyth, 2010) either using a common or a tagwise estimate of the dispersion. For DE using the Z-test, the mean expression and standard deviation for each gene was calculated using equation (3-1) and (3-2) implemented in *MeanAndSTD.pl*. DE was defined as P-value cutoff < 0.05, Bonferroni corrected (Bonferroni, 1935). Determination of DE using the exact test was performed in R version 2.1.15 (Team, 2012) using the EdgeR package version 2.6.2, and defined as P-value < 0.05, FDR corrected (Benjamini & Hochberg, 1995). For detailed description of settings used in EdgeR, readers are referred to the explanation of programs enclosed in appendix B. The normalized standard error (NSE), defined in equation (4-1) was subsequently calculated for each gene.

[21] available at: http://potatogenomics.plantbiology.msu.edu/index.html
[22] available at: http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=potato
[23] available at: http://www.plantgdb.org/download/download.php?dir=/Sequence/ESTcontig//Solanum_tuberosum/previous_version/157a

$$\frac{\sigma}{\mu \cdot \sqrt{n}} \qquad (4\text{-}1)$$

Where:
σ = Standard deviation of the mean expression
μ = Mean expression
n = Number of libraries in a biological group

To investigate the effect of replicate number on DE determination, libraries were randomly selected from each biological group using *SubsamplingOfReplicates.pl*, creating groups with 3, 6, 12, 24 or 36 replicates. DE was subsequently determined between biological groups of equal size using the exact test with tagwise dispersion estimates (P-value < 0.05, FDR corrected). The subsampling of libraries was performed 3 times for each replicate size. True differential expression (TDE) was defined as DE found using all 47 libraries. For each replicate size, the average sensitivity (observed TDE / total TDE) and specificity (observed TDE / observed DE) were calculated.

To investigate the effect of library size on DE determination, 15 triplicates were made from each biological group, combining the libraries according to size, and subsequently compared. Furthermore, a triplicate was made for each group by summing the expression values from 15 libraries (hereby creating 3 libraries of similar size). Subsequently, DE, sensitivity and specificity was calculated as described above.

4.2.5 Comparison of DeepSAGE and mRNAseq Libraries

Tissue from three leaves from different plants was homogenized and subsequently divided into three samples. From these, RNA was purified, where after one half of the sample was used for DeepSAGE library preparation, and the other half was used for RNAseq library preparation, hereby creating 9 DeepSAGE and 9 mRNAseq libraries. Sampling, library preparation, sequencing, and data preprocessing were performed as described in section 4.2.1-4.2.3. Only genes observed to be expressed by both methods were included in the analysis. The mRNAseq libraries were sequenced by a greater depth, and were therefore subsampled to the same size as the corresponding DeepSAGE library using *TaglistSubsampling.pl* to facilitate a fair comparison of the two methods. For both DeepSAGE and mRNAseq libraries 30 triplicates were constructed; 3 triplicates containing samples originating from the same leaf and 27 triplicates (all possible combinations) containing samples originating from different leaves. The averages of average normalized expression level, the measured variance, and the estimated tagwise dispersion calculated for each triplicate using the EdgeR package version 2.6.2 were hereafter calculated for triplicates containing samples originating from either the same leaf or different leaves, respectively. A non-linar regression analysis was performed for the gene expression values and the corresponding observed variance, using the non-linar models listed in Table 4-2. Data was fitted using the GRG-nonlinear solver function in Microsoft Excel 2010, and the goodness of fit (R^2), cf. equation (4-4), was subsequently calculated for each fit.

$$SS_{Total} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad (4\text{-}2)$$

$$SS_{Reg} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (4\text{-}3)$$

$$R^2 = 1 - \frac{SS_{Reg}}{SS_{Total}} \quad (4\text{-}4)$$

Where:

y_i = Observed variance for gene i
$\bar{y}$ = Average variance for all genes
$\hat{y}_i$ = Estimated variance by the regression model for gene i
R^2 = Goodness of fit for the regression model

Table 4-2 Regression models used to fit the observed variation in the DeepSAGE and mRNAseq data sets. σ^2 = variance, μ = average gene expression, CV = coefficient of variance (σ/μ). D = dispersion.

Model	Variance	Coefficient of Variance
Poisson	$\sigma^2 = \mu$	$CV = \frac{\sqrt{\mu}}{\mu}$
Negative Binomial	$\sigma^2 = \mu + D \cdot \mu^2$	$CV = \frac{\sqrt{\mu}}{\mu} + \sqrt{D}; D \geq 0$
Taylor Polynomial	$\sigma^2 = \mu + D_1 \cdot \mu^2 + D_2 \cdot \mu$	$CV = \frac{\sqrt{\mu}(1+\sqrt{D_2})}{\mu} + \sqrt{D_1}; D_1 \geq 0, D_2 \geq 0$

4.2.6 Analysis of the Large Scale DeepSAGE Library Collection

Following data preprocessing described in section 4.2.2 expression values was combined, ordered, and tabulated for each biological group using *CompareSage.pl*, hereby creating a tab separated file with expression levels from each replicate for each biological group. Hereafter the data set was filtered omitting biological groups that were represented by less than 3 samples, contained one or more libraries with an original tag count of less than 100,000, or where the size difference between the smallest and the largest library was larger than 5-fold. Using *CalculateTagWiseDispersions.pl*, which utilizes the R package EdgeR (Robinson, McCarthy & Smyth, 2010), the mean expression, observed variance, a common dispersion estimate for the library, and tagwise dispersion estimates for each gene were calculated. The normalized standard error (NSE) was calculated for each gene in each library and the number of genes above different NSE thresholds was calculated for each library using *NoGenesAboveCVcutoff.pl*. The difference in observed variation between the different data sets in the DeepSAGE library collection was investigated by testing for a difference in the average tagwise dispersion of each library between the different data sets, using student's T-test. The results were subsequently visualized using a box and whisker plot.

The level of control of gene expression was investigated by comparing the average observed CV with an estimated CV of each gene in all biological groups of the LSDS-project. Firstly, all biological groups were filtered for lowly expressed genes, only including genes with an average expression > 30 CPM. Hereafter, the average variance of each gene, measured as CV was calculated, and the observed variance was estimated using the Taylor series polynomial model, cf. Table 4-2. Finally the difference between the estimated and the observed variance was calculated and the genes deviating most from the model (either being more or less variable) were extracted.

4.3 Results

4.3.1 Comparison of High Replicate Biological Groups

To investigate if relevant variance was present in the data set, i.e. a difference between the two biological groups could be detected, a PCA of the gene expression values was performed. The two cultivars nicely split out in the PCA scores plot by the first two principal components (PCs), cf. Figure 4-2 panel A. This shows that the primary non-random variance in the data set is caused by differences in the gene expression between the two cultivars. However, the fact that the two cultivars do not split out based on a single PC indicates, that the difference in the global gene expression between the two cultivars is caused by subtle changes in many genes. This is also high-lighted by the fact that there are only very few genes that clearly split out in the loadings plot indicating an up-regulated in Kuras or Kardal, respectively (seen by the light blue and dark blue loadings in Figure 4-2 panel B, which represent genes found to be differentially expressed). However, their location in the loadings plot is clearly correlated with the splitting of the biological groups in the scores plot.

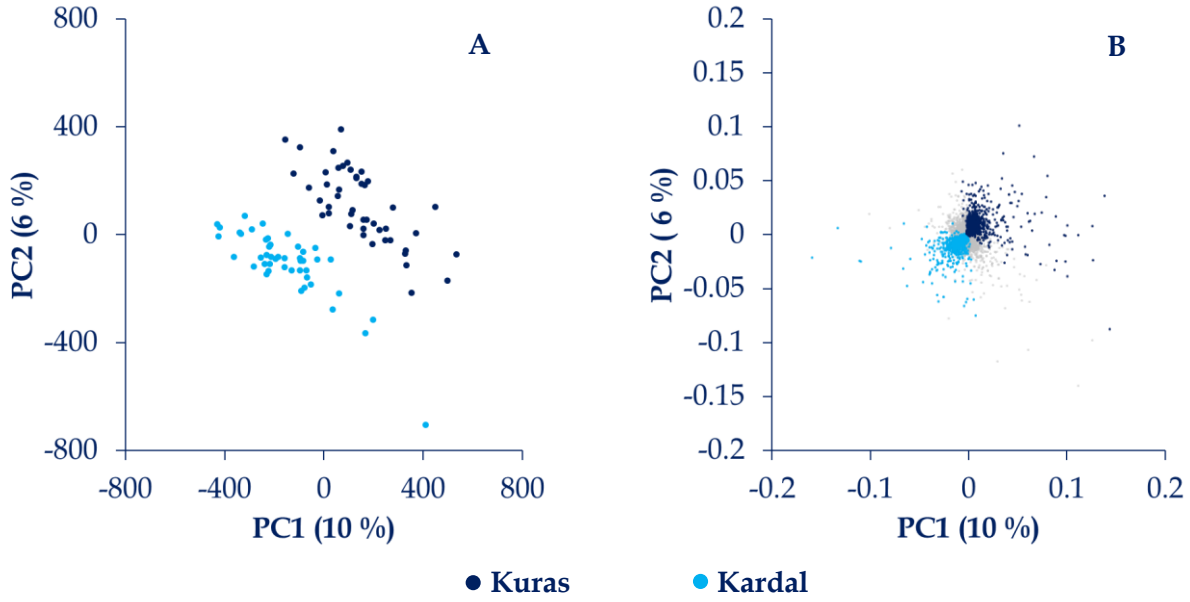


Figure 4-2 PCA of the high-replicate data set (2 x 47 replicates) containing two different biological groups (cv. Kuras and cv Kardal). **(A)** PCA scores plot showing the splitting of the samples from the two different biological groups by the two first PCs. **(B)** PCA loadings plot showing genes found to be up-regulated in Kuras (dark blue) or in Kardal (light blue), and genes not differentially expressed (light grey). There is no clear splitting of the genes in the loadings plot correlated with the splitting of the samples in the scores plot. This indicates, that the difference between the two biological groups is made up of small changes in the gene expression of many genes. Differential expression was determined using EgdeR with tagwise dispersion (cf. Figure 4-3 panel D). Data was mean centered and pareto scaled prior to PCA. PC = Principal component. Numbers in percentages indicate variance explained by the PCA model.

Differential gene expression (DE) between the two high-replicate groups was determined using the Z-test and the exact test implemented in EdgeR using common or tag was wise dispersion estimates, respectively, cf. Figure 4-3. Here, a typical pattern for gene expression data can be observed where the largest fold change differences are found for lowly expressed genes. The noise level in the data set was elucidated by an investigation of the fraction of "noisy" genes (genes with expression values having a CV > 1), cf. Figure 4-3 panel A. Only 5,876 genes out of the 40,438 observed (14 %) had a CV below 1.

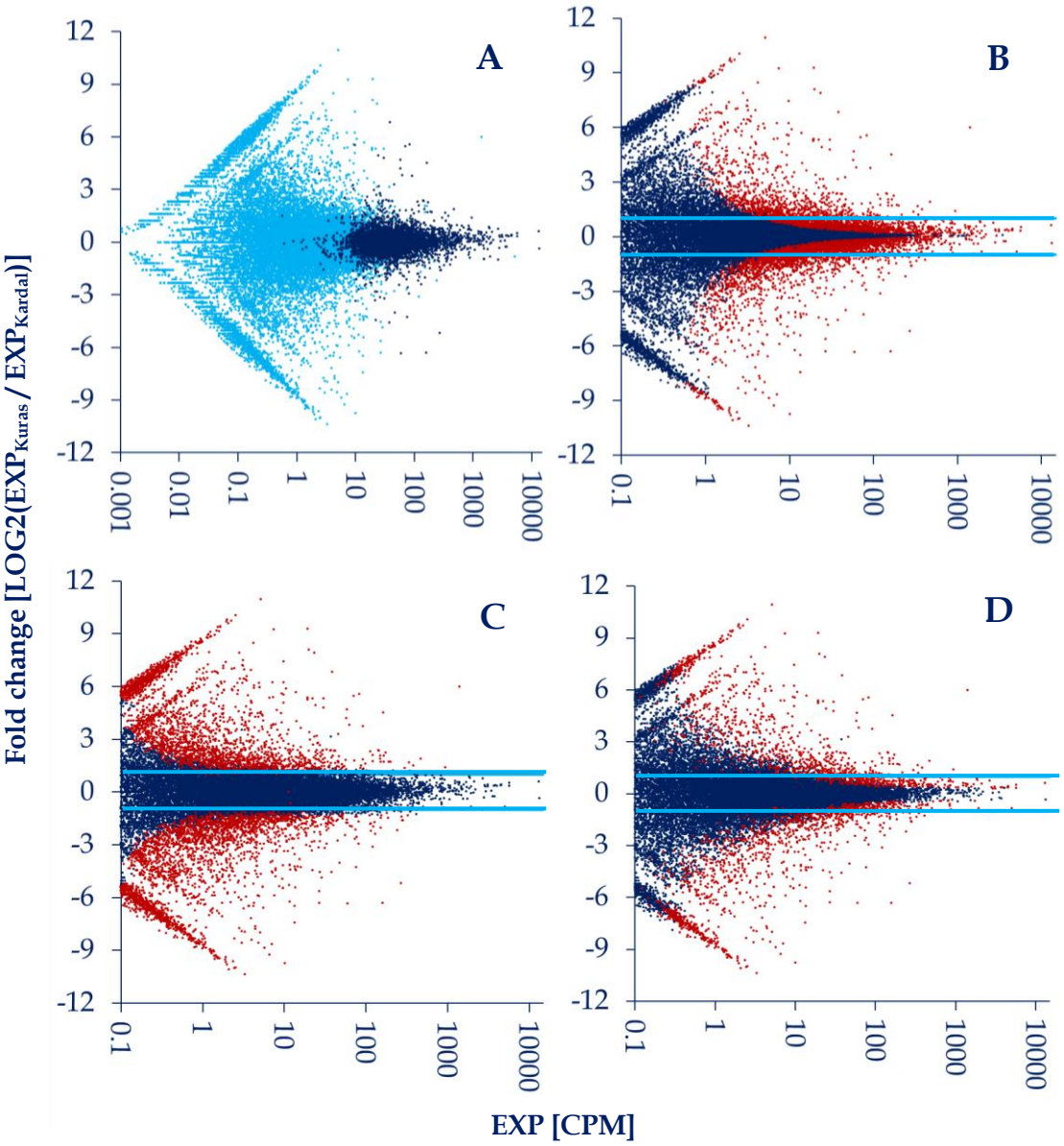


Figure 4-3 Detection of differential expression between the two high-replicate groups. Plots of the log-fold change against the log-expression for each gene (similar to the MA-plot used in data analysis of cDNA microarray experiments (Dudoit *et al.*, 2002)). A typical pattern for gene expression data can be observed where the largest fold change differences are found for lowly expressed genes. **(A)** Estimation of the observed gene expression levels. Genes with CV > 1 are marked in light blue are considered to have a noisy gene expression. Genes with CV ≤ 1 are marked in dark blue. 86 % of all the genes observed have a "noisy" gene expression. Genes found to be differentially expressed (marked in red) using the Z-test **(B),** EdgeR with a common estimate of the dispersion **(C),** or EdgeR with a tagwise estimate of the dispersion **(D).** Genes with a P-value < 0.05 after correction for multiple testing were defined as differentially expressed. The horizontal blue lines indicate a fold 2 change. EXP = mean Expression. CPM = counts per million.

For each gene, the CV is well determined due to the high number of replicates (n = 47). The high number of replicates ensures a good estimation of the mean expression, because NSE → 0 for $\mu \rightarrow \infty$. Due to this, 77 % of all genes detected were well determined (defined as NSE < 1). Therefore it can be concluded that the between library variation in the DeepSAGE data set is relatively large. Knowing this, it is also clear that the Z-test, which does not take the dispersion into account is not a good choice of statistical test for differential expression, and that a large fraction of the 7,866 genes detected as differentially expressed using the Z-test most likely are false positives. Even more, the more stringent Bonferroni correction was used to correct for multiple testing contrary to the FDR correction method, which was used for cor-

rection of multiple testing using the exact test. Even so, in the case of 4,449 genes the p-value was found to be ≈ 0, and these would hence be determined as DE no matter the choice of multiple testing correction method. It can also be observed that a large fraction of highly expressed genes, with little fold change difference in expression between the two groups have been called as DE using the Z-test, cf. Figure 4-3 panel B. Due to the low difference in expression, a large fraction of these are unlikely to be biological relevant, when phenotypical differences between the cultivars are to be elucidated. Using the exact test 5,393 and 2511 genes where determined as DE using a common dispersion estimate or a tagwise dispersion estimate, respectively, cf. Figure 4-3 panels C and D. When comparing the observed and estimated variance, the observed variance is generally underestimated for low expression levels and overestimated for high expression levels using a common dispersion estimate, cf. Figure 4-4.

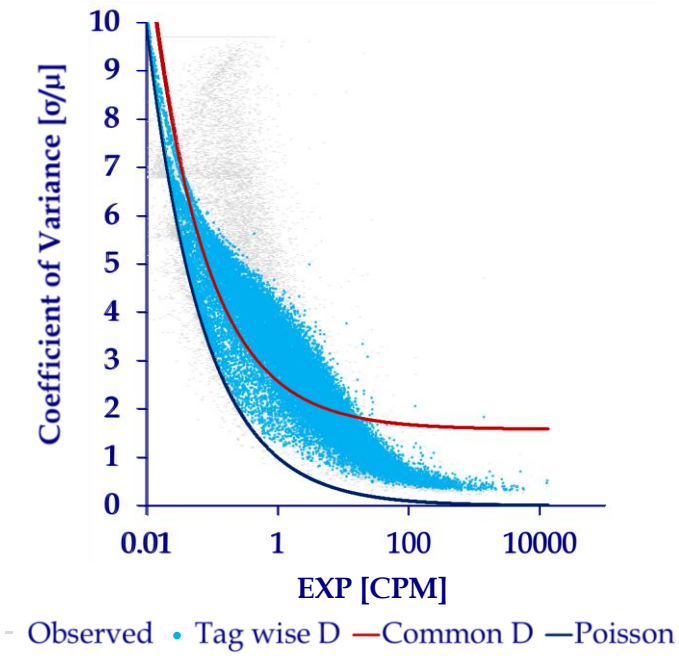


Figure 4-4 Noise measured as CV (σ/μ) as a function of gene expression level. There is a clear variance-expression dependency. In general using the Poisson model (dark blue line) underestimates the dispersion (D). Tagwise estimation of the dispersion (light blue dots) out-performers a common dispersion estimate (red line) only underestimating the dispersion for very lowly expressed genes (EXP < 1 CPM) whereas a common dispersion estimate underestimates the dispersion for lowly expressed and overestimates it for highly expressed genes. EXP = mean gene expression CPM = counts per million.

This result in a much larger fraction of lowly expressed genes and a smaller fraction of highly expressed genes being determined as DE using a common dispersion estimate compared to using a tagwise estimate, cf. Figure 4-3 panels C and D. In general, a tagwise dispersion estimate out-performs common dispersion estimation only underestimating the dispersion for very lowly expressed genes (EXP < 1 CPM). This is also seen in the goodness of fit for the two models, where R^2 = 0.67 for tagwise dispersion estimation and R^2 = 0.46 for common dispersion estimation.

The effect of replicate number on DE determination was investigated by random selection of libraries from each biological group, hereby creating groups with 3 to 36 replicates followed by calculation of the average specificity and sensitivity of the DE determination, cf. Figure 4-5. The sensitivity drops dramatically for genes at all expression levels when lowering the

number of replicates. In fact, only an average of 1 % of the genes found to be differentially expressed between the biological groups using all 47 replicates were also determined as DE only using triplicates to represent the biological groups, i.e. the number of type II errors is high, cf. Figure 4-5 panel A. However, the specificity is maintained when lowering the number of replicates, cf. Figure 4-5 panel B. Except for extremely lowly expressed genes (< 1 CPM), nearly all the DE genes found only using triplicates were true positives, i.e. the number of type I errors is low. Not surprisingly, the specificity of detection of genes found to be differentially expressed with the highest significance, i.e. with the lowest P-values, is retained to a higher degree when lowering the number of replicates, cf. Figure 4-5 panel C.

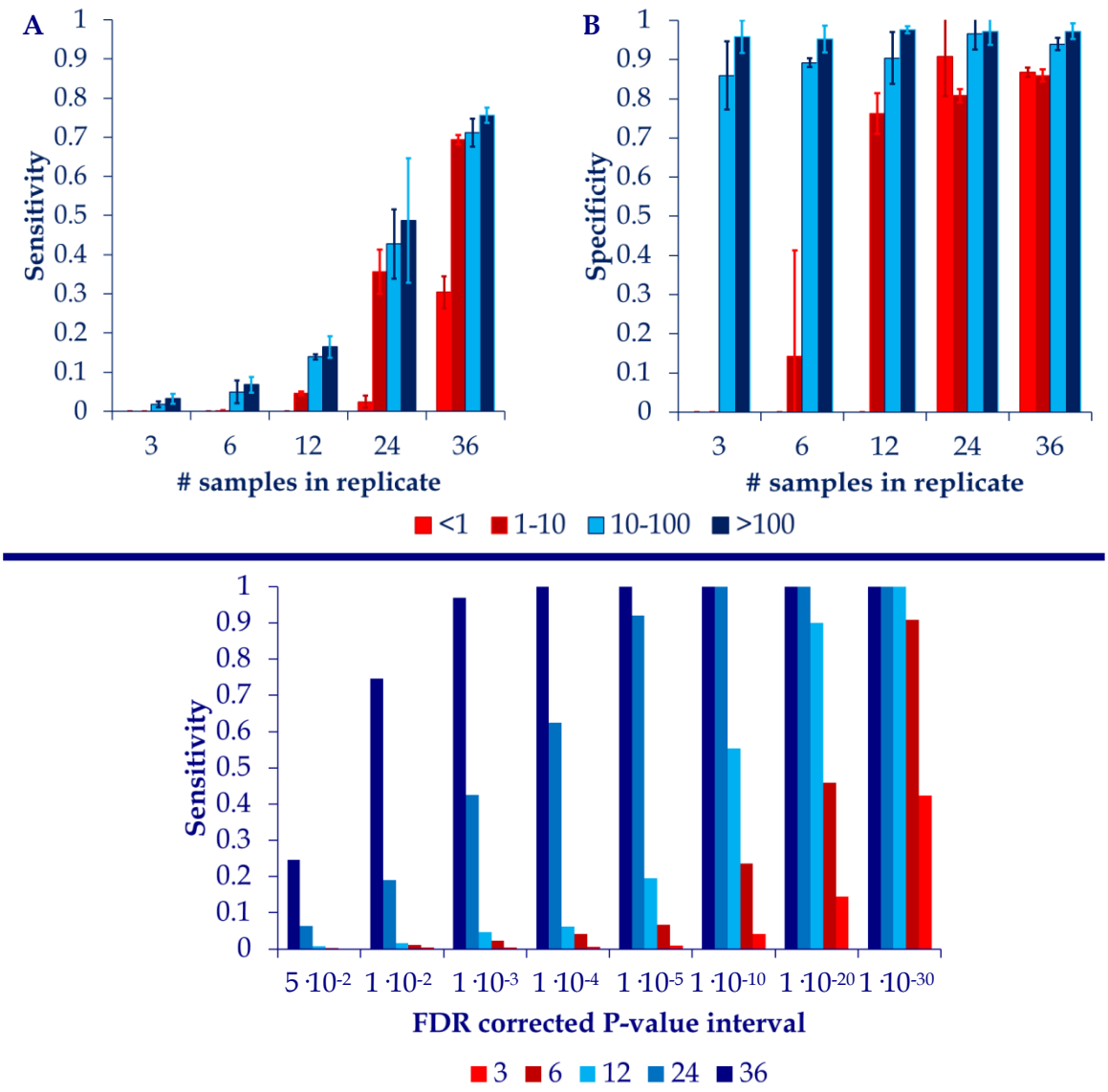


Figure 4-5 Effect of replicate number on the determination of differential expression for genes grouped at different expression levels. **(A)** The sensitivity drops dramatically for genes in all expression levels when lowering the number of replicates. **(B)** However, the specificity is maintained at all expression levels except for extremely lowly expressed genes (light red bars). **(C)** Specificity group by the original P-value found using all 47 replicates for different replicate sizes. In general, the sensitivity of DE genes determined with highest significance (lowest P-values) is retained to a higher degree compared to genes determined as DE close to the P-value cutoff 0.05. The exact test with tagwise dispersion estimation was used for DE determination.

The effect of the library size of a triplicate on DE determination was investigated by ordering the 47 libraries of the high-replicate groups according to size, and creating triplicates with an

average library size between 500,000 and 1.7 million tags, cf. Figure 4-6 panel A, where after specificity and sensitivity of the DE determination was calculated, cf. Figure 4-6 panel B.

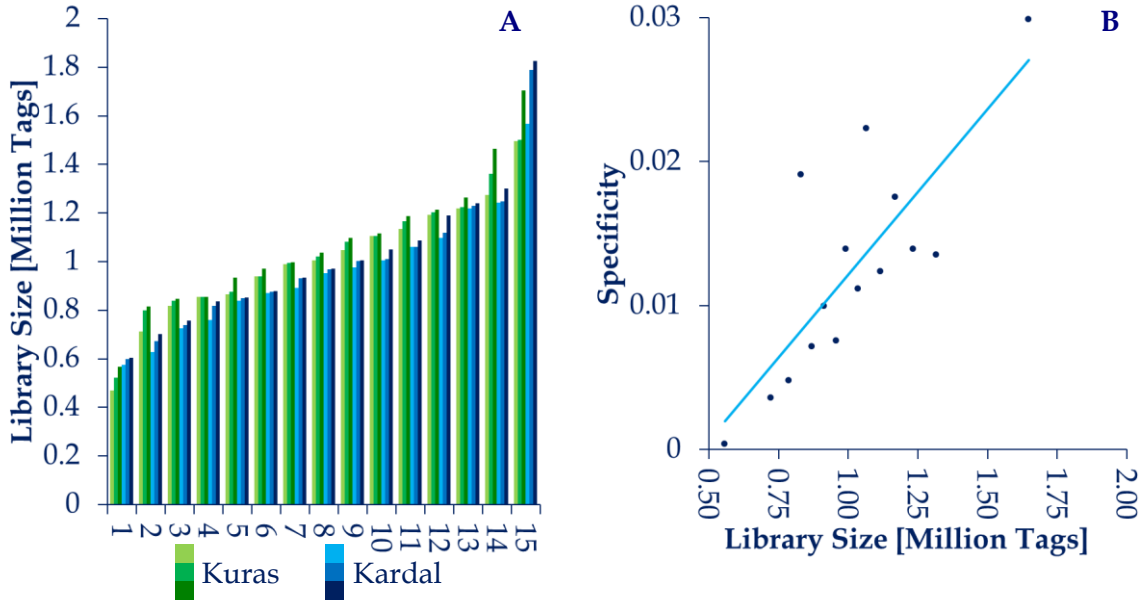


Figure 4-6 Effect of the library size on the determination of differential gene expression. **(A)** Grouping of libraries from the high-replicate groups into triplicates according to library size. **(B).** The specificity of the DE determination in triplicates is correlated with library size. Between 1 and 89 genes were found to be differentially expressed in the triplicates out of the 2511 genes found using all 47 replicates, The light blue line indicates the positive linar correlation between the number of genes detected and the library size (ρ_p = 0.64).

Although there is a slight improvement of the sensitivity with increasing library size of a triplicate, only 3.5 % of the genes detected as DE using all 47 replicates are detected in the largest triplicate. This indicates that the low specificity found in the triplicates only to a low extent is caused by sequencing depth, and more likely is caused by between library variance. The specificity was once again found to be high in the triplicates with an average specificity of 0.88 ± 0.08.

To investigate whether the number of replicates or the between library variance were the cause of the low sensitivity observed for DE detection using triplicates, the 47 replicates from each biological group were pooled into 3 replicates of equal size (hereby including almost the entire data set), and DE determination between the biological groups was performed following calculation of the sensitivity and the specificity of the DE determination, cf. Figure 4-7. In total, 1,449 genes were found to be differentially regulated both when the biological groups consisted of large triplicate and when they consisted of 47 smaller libraries, generally showing good agreement in the DE determination between the two different types of division of the data set (either many small or few large libraries) indicated by good correlation between the significance of the DE detection, cf. Figure 4-7 panel A. Most of the genes not detected as DE, using a large pooled triplicate, were lowly expressed, cf. Figure 4-7 panel B. In general, the sensitivity was much higher for the large pooled replicate compared to the smaller random replicates. Therefore, the lack of sensitivity in a random triplicate from the high-replicate group is meant to be caused by a large between library variation, which cannot be accounted for in a triplicate.

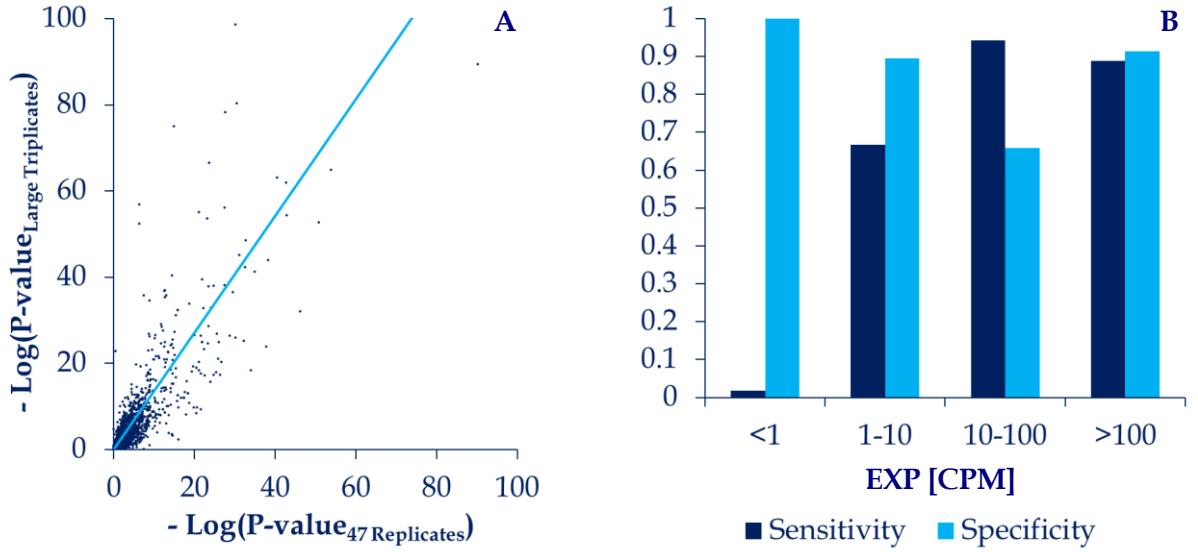


Figure 4-7 The effect of the number of replicates vs. the between library variance on DE determination. **(A)** Correlation between the significance of the difference between the two biological groups calculated using either biological groups consisting of 47 small replicates or a large triplicate of pooled libraries. There is a good linear correlation (ρ_p = 0.71). The P-values found using a large triplicate is on average 1.4 times lower than those found using 47 small replicates. **(B)** Sensitivity and specificity of the DE determination using biological groups consisting of large triplicate of pooled libraries compared to a biological group consisting of 47 smaller replicates.

The source of this variation can both be caused biological differences between the samples and differences caused by the sample preparation. This will be elucidated in the following section.

4.3.2 Comparison of DeepSAGE and mRNAseq Libraries

The experiment described in the following was designed to investigate the contributions of intrinsic noise (variance observed between samples originating from the same leaf) and extrinsic noise (variance observed between samples originating from different leaves to the overall variation in DeepSAGE and mRNAseq data sets, respectively. The data set consists of 3x3 DeepSAGE libraries and 3x3 mRNAseq libraries.

Firstly, the correlation between corresponding mRNAseq and DeepSAGE libraries was investigated, cf. Figure 4-8. Since both methods are count based, a linear correlation between the observed expression values measured by DeepSAGE and mRNAseq was expected. However, the coefficient of determination (ρ_p) is only 0.530, cf. Figure 4-8 panel A. Since there is a good linear correlation internally between either DeepSAGE or mRNAseq libraries, cf. Figure 4-8 panels C and D, the poorer correlation between the two methods must primarily be caused by the fundamental differences in the library preparation and data analysis that exist between DeepSAGE and mRNAseq. It was attempted to minimize differences in the data analysis between the two data sets. For instance, the mRNAseq libraries were subsampled to the same size as the DeepSAGE libraries. The subsampling has nearly no effect on the correlation between the DeepSAGE and mRNAseq data, since there are nearly perfect correlation between the full size and the subsampled mRNAseq libraries, cf. Figure 4-8 B. However, differences in the data analysis that cannot be avoided do exist. Here, mapping of mRNAseq reads to a genome sequence vs. annotation of DeepSAGE tags should be mentioned. E.g. how randomly matching mRNAseq reads are treated when calculating expression values plays a role. It was chosen to include randomly matching mRNAseq reads, while it is similar

to the distribution of non-unique DeepSAGE tags to multiple genes. The effect of this was however not investigated further.

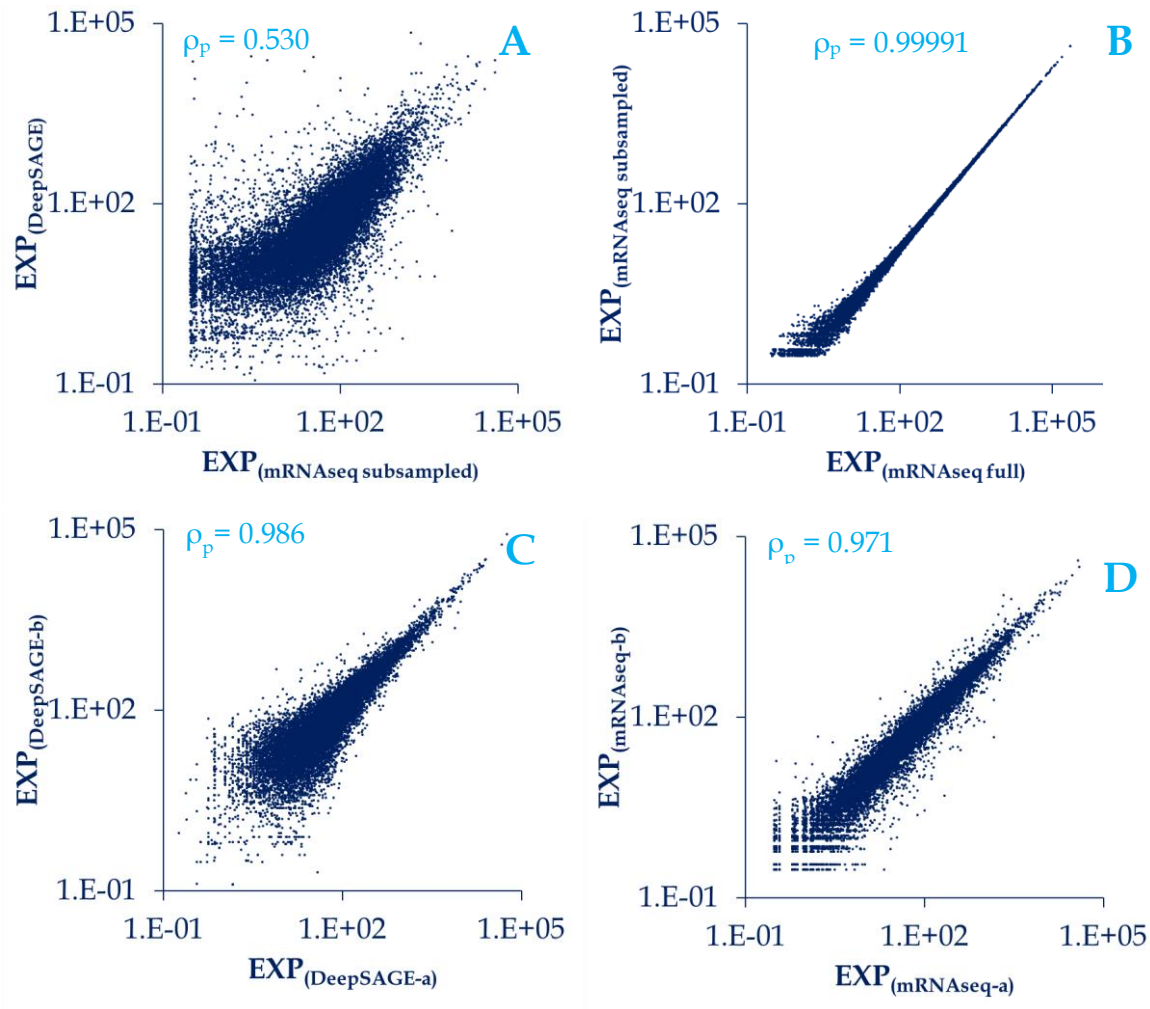


Figure 4-8 Comparison between gene expression measured by mRNAseq and DeepSAGE. **(A)** There is only medium linear correlation between average expression of subsampled mRNAseq libraries, and DeepSAGE libraries. **(B)** There is nearly perfect linear correlation between average expression of full-size mRNAseq libraries, and the mRNAseq libraries that have been subsampled to the same size as the DeepSAGE libraries. There is a good linear correlation internally between both DeepSAGE **(C)** and mRNAseq libraries **(D)**. The observed differences between mRNAseq and DeepSAGE must therefore be caused by fundamental differences in the library preparation and data analysis between the two methods. EXP = mean expression.

The intrinsic and extrinsic noise in both mRNAseq and DeepSAGE libraries were investigated by an analysis of the observed variance in triplicates originating from the same or different leaves, respectively; cf. Figure 4-9. Here, the observed CV and estimated CVs are plotted against the gene expression level from the mRNAseq or DeepSAGE data set, respectively.

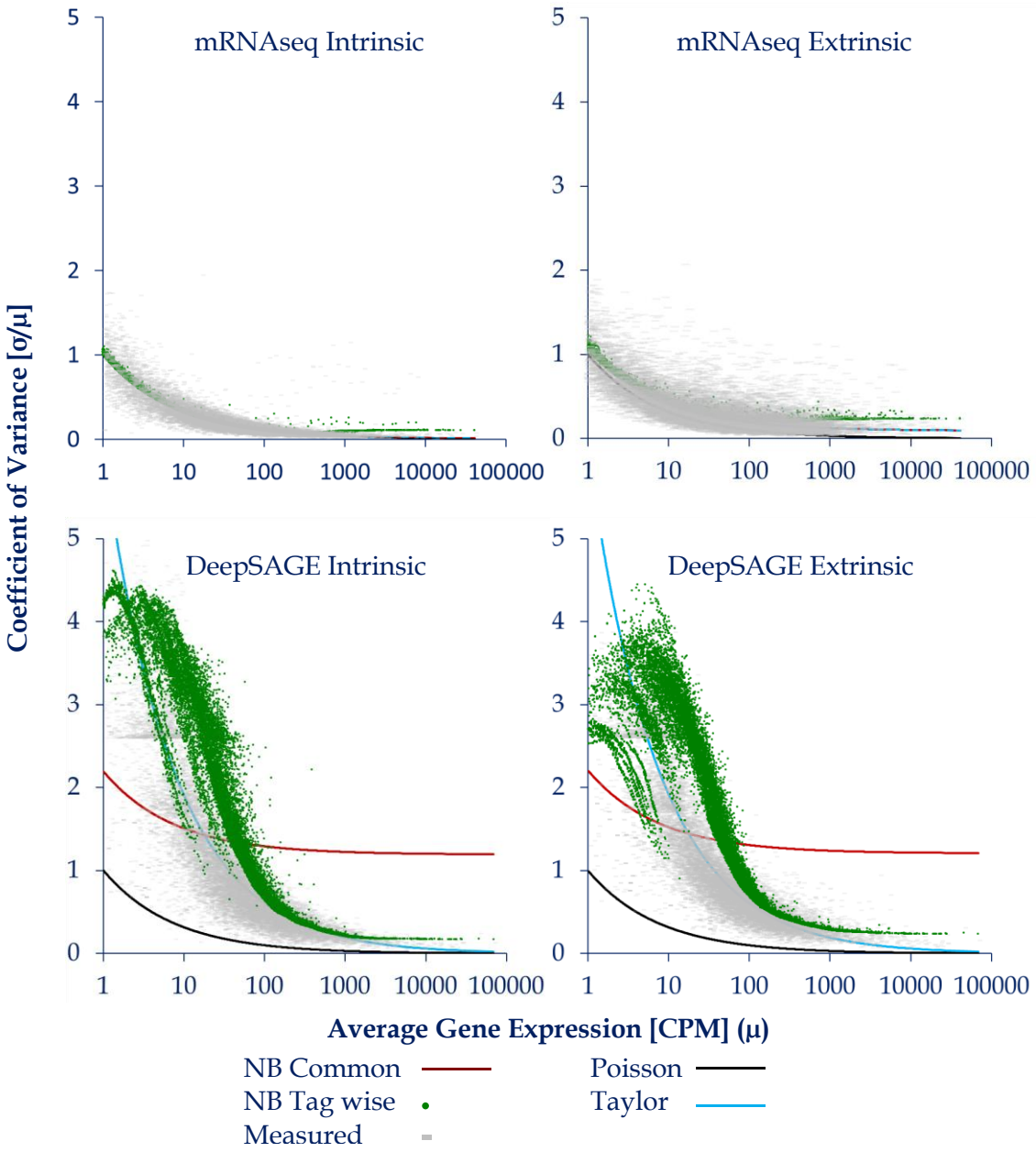


Figure 4-9 Intrinsic and Extrinsic noise in mRNAseq and DeepSAGE gene expression data sets. The average measured CV (marked in light grey) from triplicates originating from the same leaf (representing intrinsic noise) and from different leaves (representing intrinsic + extrinsic noise) plotted against average gene expression level in mRNAseq and DeepSAGE data sets, respectively. The average CV is fitted to different statistical models using non-linear regression. The regression models and the results are listed in Table 4-3. NB = negative binomial. Taylor = Taylor polynomial.

It is clear that the intrinsic noise in the mRNAseq data sets is very low compared to the DeepSAGE data set since the observed variance measured as CV is lower, cf. Figure 4-9. In fact, the variance of the data is well modeled using the Poisson distribution that only accounts for the sampling process (R^2 = 0.902, cf. Table 4-3). This is also shown by the results and the goodness of fit (R^2) of the non-linear regression of the other models. Both the negative binomial and Taylor polynomial models have parameters very close to 0 for the mRNAseq intrinsic data sets, and are hereby both reduced to the Poisson model. Furthermore, the addition of additional parameters does not improve the goodness of fit, cf. Table 4-3. The tagwise estimation of the dispersion (D) only models the variation in the data slightly better than the poison model (R^2 = 0.910, cf. Table 4-3). Together, this indicates that there is very low intrinsic noise, ad i.e. low technical variance in the mRNAseq data set.

Table 4-3 Regression models used to fit the observed variation in the DeepSAGE and mRNAseq data sets. Either the entire data set (top) or only genes with an expression level ≥ 100 (bottom) was used for the regression analysis. σ^2 = variance, μ = average gene expression, CV = coefficient of variance (σ/μ). R^2 = Goodness of fit. * Parameters are tagwise estimated. ** Tagwise estimates are calculated using the entire data set.

Data set/Model	mRNAseq Intrinsic		mRNAseq Extrinsic		DeepSAGE Intrinsic		DeepSAGE Extrinsic	
	Parameters	R^2	Parameters	R^2	Parameters	R^2	Parameters	R^2
Poisson $\sigma^2 = \mu$	None	0.902	None	0.466	None	-0.443	None	-0.492
NB Common $\sigma^2 = \mu + D \cdot \mu^2$	$D = 3.8*10^{-6}$	0.902	$D_1 = 8.44*10^{-3}$	0.618	$D_1 = 1.423$	0.229	$D_1 = 1.458$	0.232
Taylor Polynomial $\sigma^2 = \mu + D1 \cdot \mu^2 + D2 \cdot \mu$	$D_1 = 3.8*10^{-6}$ $D_2 = 0$	0.902	$D_1 = 8.44*10^{-3}$ $D_2 = 0$	0.618	$D_1 = 0$ $D_2 = 5.031$	0.773	$D_1 = 0$ $D_2 = 5.069$	0.777
NB tagwise $\sigma^2 = \mu + D_{(gene)} \cdot \mu^2$	NA*	0.910	NA*	0.673	NA*	0.406	NA*	0.178
Poisson For $\mu \geq 100$	None	0.014	None	-0.92	None	-3.21	None	-4.098
NB Common For $\mu \geq 100$	$D = 3.9*10^{-6}$	0.026	$D_1 = 1.17*10^{-2}$	-0.03	$D_1 = 0.087$	0.167	$D_1 = 0.117$	0.142
Taylor Polynomial For $\mu \geq 100$	$D_1 = 1.1*10^{-4}$ $D_2 = 0$	0.049	$D_1 = 1.17*10^{-2}$ $D_2 = 0$	-0.03	$D_1 = 2.7*10^{-3}$ $D_2 = 3.770$	0.446	$D_1 = 0{,}015$ $D_2 = 3.419$	0.355
NB tagwise For $\mu \geq 100$	NA**	0.219	NA**	0.101	NA**	0.046	NA**	0.243

There is a much larger increase in the observed variance towards lower expression levels in the DeepSAGE data set compared to the mRNAseq data set, cf. Figure 4-9. In fact CV is > 1 for most genes with an expression below 100 CPM, and hence the standard deviation is larger than the estimated expression level. This has a large impact on the variance estimate using the different models. Since the Poisson distribution, only takes sampling variation into account, it is a very poor model of the observed variation in the DeepSAGE data. This is seen by the extreme low goodness of fit measure ($R^2 < 0$, cf. Table 4-3), indicating that the regression error of the model is larger than setting the variance of all genes to the average variance in the data set (a horizontal line in the plot). For the same reason, the negative binomial distribution with a common dispersion estimate is also a very poor model of the observed variance. A common dispersion estimate adds a fixed value $(\sqrt{D})$ to the CV estimate of the Poisson distribution (a vertical displacement of the Poisson distribution graphs in Figure 4-9). This fails to account for the more rapid large increase in CV towards lower expression levels observed in the DeepSAGE data sets, resulting in poor regression results ($R^2 \approx 0.23$ for the DeepSAGE data sets, cf. Table 4-3). It is clear that there is an additional expression dependent factor in the relationship between the observed variance and the expression level in the DeepSAGE data set, other than the linear relationship accounted by the Poisson distribution ($\sigma^2 = \mu$). This is somewhat taken into account by tagwise estimation of the dispersion implemented in EdgeR improving the goodness of fit to 0.4 for the DeepSAGE intrinsic noise data set, cf. Table 4-3. However, the tagwise estimation method implemented in EdgeR attempts to "squeeze" the tagwise dispersion towards the estimated common dispersion, which obviously is a poor dispersion estimate. This results in over estimation of the dispersion, and hence increases the change of false negatives when testing for differential expression in the DeepSAGE data sets. A more suited estimate of the observed dispersion seems to be a simple Taylor series with an extra term ($D_2 \cdot \mu$) compared to the negative binomial model with a common estimate of the dispersion. The regression results using this model is far better than

the other models ($R^2 \approx 0.77$ for the DeepSAGE data sets, cf. Table 4-3). However, the model underestimates the dispersion for larger expression values, especially for the DeepSAGE data set cf. Figure 4-10 and Table 4-3.

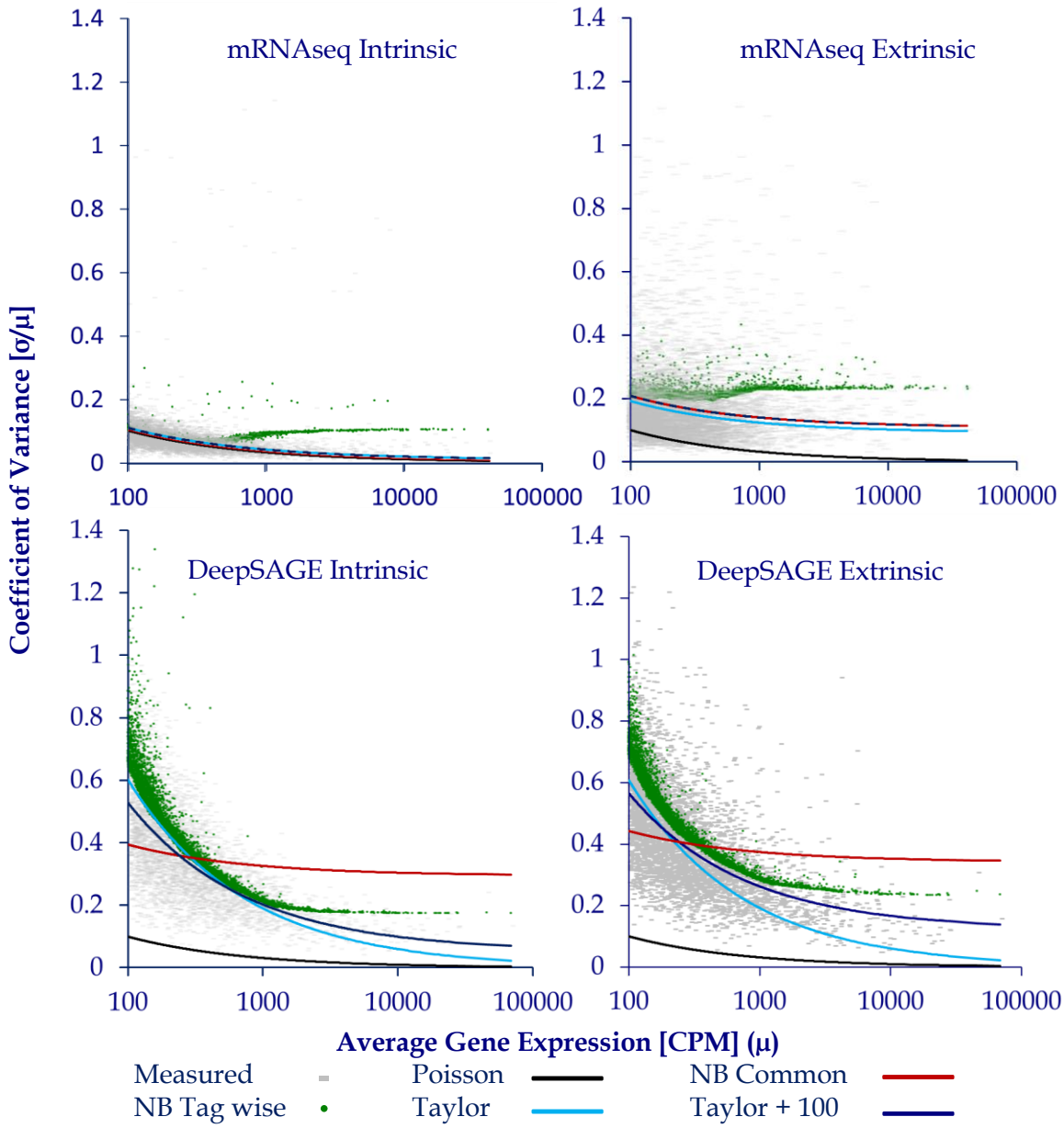


Figure 4-10 Intrinsic and Extrinsic noise in mRNAseq and DeepSAGE gene expression data sets for highly expressed genes (EXP > 100). The average measured CV (marked in light grey) from triplicates originating from the same leaf (representing intrinsic noise) and from different leaves (representing intrinsic + extrinsic noise) plotted against average gene expression level in mRNAseq and DeepSAGE data sets, respectively. The average CV is fitted to different statistical models using non-linear regression only including genes with a gene expression level >= 100. The regression results are listed in Table 4-3. Expression independent variance can be observed in both the DeepSAGE and mRNAseq extrinsic datasets, reflecting biological differences between different plants. NB = negative binomial. Taylor = Taylor polynomial using the entire data set. Taylor + 100 = Taylor polynomial using only genes with an expression level > 100 for the regression.

This is caused by the fact that the majority of genes have low expression levels, why the regression model favors a good fit at lower expression levels. In general, all regression models have a poor goodness of fit, indicating that the average variance in the data set fits the data equally well. This implies that there is a base level of variance for all genes which is expression independent. When comparing the variance of the mRNAseq data sets from the same

and different leaves, respectively, cf. Figure 4-9, it is clear that there is an increase in the observed variance for a large fraction of the genes. This additional variation accounts for differences in the gene expression between different plants. It was hypothesized that the genes with the largest increase in variance of their expression was of a certain kind, or belonged to a certain group (e.g. genes involved in photosynthesis or defense, which are known to vary greatly between individuals). An initial investigation of this was made by manual inspection of the annotation of the genes with the largest difference in CV between the intrinsic and extrinsic mRNAseq data sets, but no conclusions could be made, and the hypothesis was not investigated further using this data set.

The average biological variance in the data sets (defined as the additional variation observed in the data sets with triplicates originating from different plants compared to the observed variation in the data sets with triplicates originating from the same leaf) was quantified at gene different expression levels in both the DeepSAGE and mRNAseq data sets, cf. Figure 4-11.

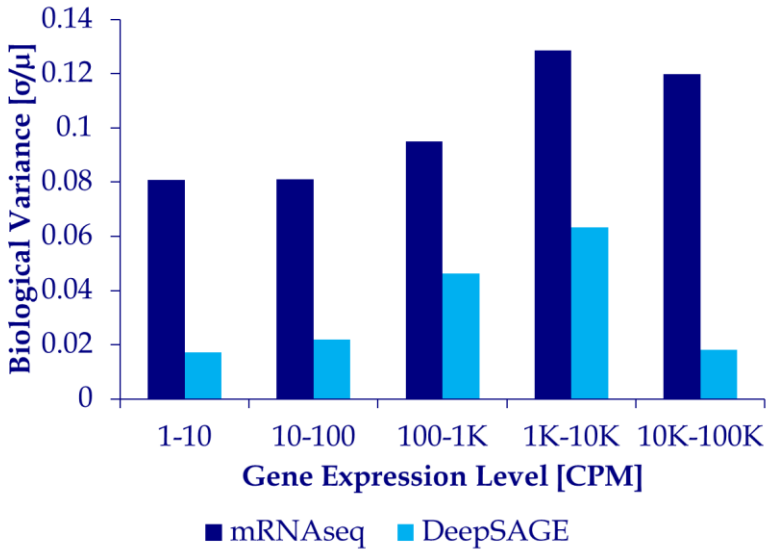


Figure 4-11 Biological variance between field grown *S. tuberosum* plants measured by mRNAseq and DeepSAGE. The biological variance is estimated as the average difference between the observed variance in data sets with triplicates originating from the same leaf, representing intrinsic noise and data sets with triplicates originating from different leaves. The overall average observed biological variance in the mRNAseq data set was found to be 8 % but only 3 % in the DeepSAGE data set. The coefficient of variance (CV) is used as variance measure.

It seems like there is a tendency of higher observed biological variance in genes with high expression values, except for genes with the most abundant expression (> 10,000 CPM), which can be explained by poor determination due to the low number of genes in this category. The increase in the estimation of the biological variance for genes with higher expression could be explained by the relatively lower contribution of sampling variation for higher expression values ($\sqrt{\mu}/\mu \rightarrow 0$ for $\mu \rightarrow \infty$). The overall average observed biological noise (measured as CV) in the mRNAseq data set was found to be 8 % but only 3 % in the DeepSAGE data set. For the mRNAseq data set, the biological noise can also be estimated using the estimates of the common dispersion. Here, the biological noise was found to be 9 %, which is well in line with the result found using the observed variation.

4.3.3 Analysis of a large scale DeepSAGE Library Collection

Although the variance might be difficult to evaluate in a low replicate (such as a triplicate) biological group, an analysis of a data set consisting of many triplicates might contribute to the elucidation of variance sources in DeepSAGE data. To estimate the overall quality of the determination of the gene expression level, the amount of "noisy" gene expression (genes with CV > 1) was investigated in data sets of the LSDS-project containing 499 biological groups, cf. Figure 4-12. On average, CV of the average gene expression was below 1 of ~ 8,000 genes in each biological group, which on average corresponded to approximately half of the genes observed. This large between library variation in the biological groups naturally reduces the ability to detect genes that are differentially expressed between the biological groups.

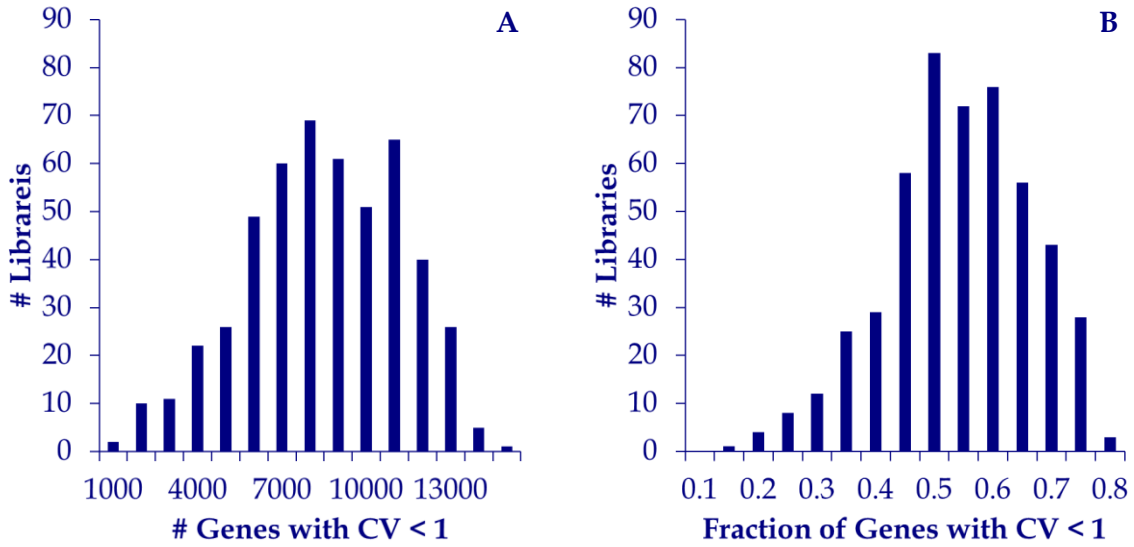


Figure 4-12 Amount of genes with relatively low in between library variation (CV < 1) in the large-scale collection of DeepSAGE libraries. **(A)** Distribution of libraries based on the total amount of genes with CV <1. **(B)** Distribution of libraries based on the fraction of genes with CV < 1. The average amount and the average fraction of genes with low in between library variation in each library was found to be 7972 ± 2767 and 0.51 ± 0.12, respectively.

The quality of the estimation of the mean expression level in each group is dependent on the number of replicates in the biological groups, cf. equation (4-1). The average fraction of well determined genes with a low NSE is low in the biological groups of the LSDS-project, cf. Figure 4-13. This of course impairs the ability to detect differences between the biological groups, why only relatively larger differences in the gene expression between the biological groups can be expected to be detected. To improve the average fraction of well determined gens, the immediate solution would be to increase the number of replicates in the biological groups (since NSE $\rightarrow 0$ for $n \rightarrow \infty$).

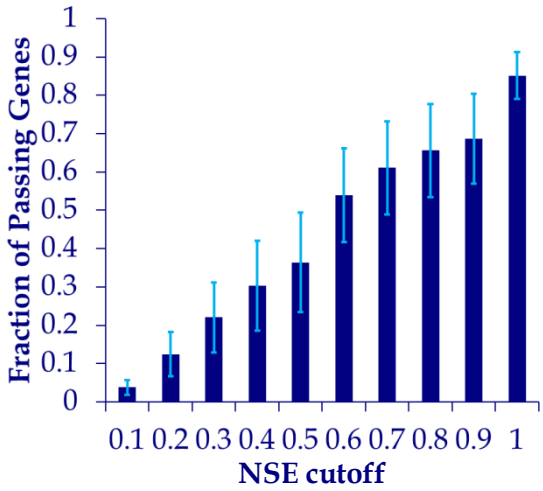


Figure 4-13 Quality of mean expression level determination in LSDS-project biological groups. The average fraction of genes left after applying different cutoffs of the normalized standard error (NSE), defined as $\frac{\sigma}{\mu \cdot \sqrt{n}}$, cf. equation (4-1). The error bars indicate the standard deviation. The number of libraries (n) in each biological group is 3.

Different growth conditions obviously have an effect on the variation in the gene expression data, e.g. since larger differences in growth conditions exist for field grown plants compared to greenhouse grown plants. To elucidate the effect of the different growth conditions used in the LSDS-project, the variance in data sets with different growth conditions was investigated, cf. Figure 4-14 and Table 4-4.

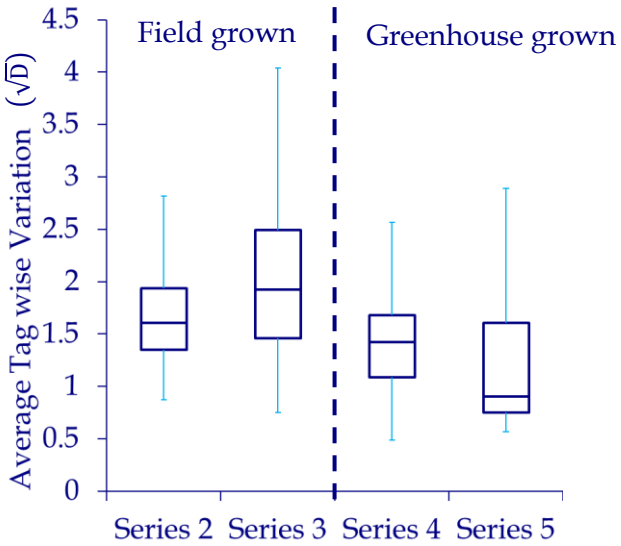


Figure 4-14 Effect of different growth conditions on the between library variation of biological groups in the LSDS-project. The average tagwise variance estimate $(\sqrt{D})$, which was used as an overall measure for the variance in each biological group, was calculated using EdgeR. The estimated variance in series 4 and 5, which contains greenhouse grown samples are significantly lower than the estimated variance in series 2 and 3, which contains field grown samples, cf. Table 4-4.

Table 4-4 Difference in the average tagwise variance between different data sets with libraries originating from field grown potato plants (series 2 and 3) or green house grown potato plants (series 4 and 5). Difference in the average tagwise variance $(\sqrt{D})$ between the data sets was tested using a student's T-test. The resulting P-values after Bonferroni correction for multiple testing are listed.

Data set	Series 3 (Field)	Series 4 (Greenhouse)	Series 5 (Greenhouse)
Series 2 (Field)	$2.87 \cdot 10^{-5}$	$3.55 \cdot 10^{-3}$	$2.96 \cdot 10^{-10}$
Series 3 (Field)		$8.66 \cdot 10^{-13}$	$9.72 \cdot 10^{-23}$
Series 4 (Greenhouse)			$4.44 \cdot 10^{-3}$

As expected, the largest between library variation was observed in data sets with field grown potato plants, cf. Figure 4-14. In general the variation between libraries is large, only in data series 5, which consist of greenhouse grown plants, is the average tagwise variation estimate below 1. A significant difference between field grown and greenhouse grown data sets was observed, cf. Table 4-4. The average tagwise variation was between 1.1 and 2.1 larger in the data sets with field grown potato plants. Again, this reduces the ability to detect genes that are differentially expressed between the biological groups.

Finally, the level of control of gene expression was investigated by comparing the average observed CV with an estimated CV of gene with the highest average expression in all 499 biological groups of the LSDS-project. Because the variance is expression dependent, a direct comparison of the variance measured as CV is not possible. Therefore, the observed variance was compared the estimated variance modeled by the simple Taylor series, cf. Figure 4-15.

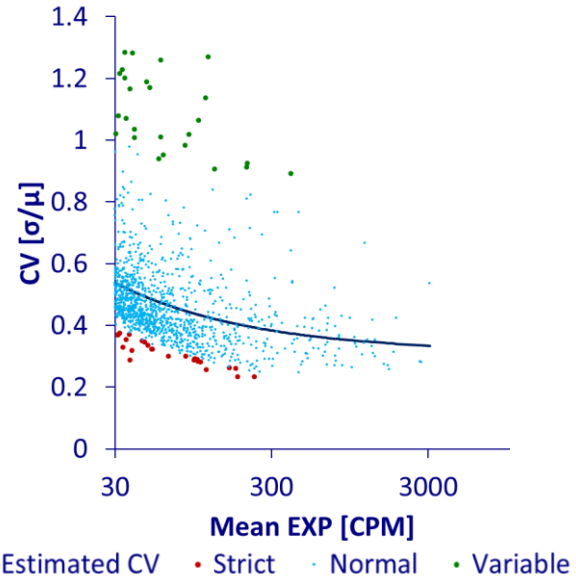


Figure 4-15 Gene expression control of highly expressed genes observed in biological groups in the LSDS-project. The average observed CV of all 499 biological groups is plotted against the average gene expression level for highly expressed genes (EXP > 30 CPM). The expected variance is estimated using the simple Taylor series model (dark blue line), cf. Table 4-3. Genes with the lowest observed variance compared to the expected (marked in red) are assumed to be under strict control, and the genes with the highest observed variance compared to the estimated (marked in green) are assumed to have a more variable gene expression, and therefore be under less control. EXP = average gene expression. CPM = counts per million.

The expression of genes with a lower observed variance compared to the estimated variance differ less between the biological replicates, and the expression is therefore assumed to be more strictly controlled, and vice versa for genes with a higher observed variance compared to the estimated. The 25 genes assumed to be under the strictest control, and the 25 genes having the most variable gene expression are listed in Table 4-5.

Table 4-5 Genes found to be under strict control (left), and genes found to have a more variable gene expression compared to the expected (right). [1]PGSC0003DMG is omitted from all PGSC IDs.

PGSC ID[1]	Annotation	CV Diff.	PGSC ID[1]	Annotation	CV Diff.
400027199	Gene of unknown function	-0.226	400010146	Kunitz-type tuber invertase inhibitor	0.844
400009893	Gene of unknown function	-0.198	400014104	Patatin-2-Kuras 4	0.786
400020228	Protein phosphatase 2a, regulatory subunit	-0.194	400010283	Class I chitinase	0.771
401021841	Gene of unknown function	-0.172	400010136	Stigma expressed protein	0.761
400007579	Coiled-coil domain-containing protein	-0.171	400010143	Cysteine protease inhibitor 1	0.708
402025662	Gene of unknown function	-0.167	400031877	Metallocarboxypeptidase inhibitor	0.701
402001978	Apoptosis-inducing factor	-0.165	400002261	Conserved gene of unknown function	0.698
400001390	Gene of unknown function	-0.165	400010137	Cysteine protease inhibitor 1	0.685
400010923	Gene of unknown function	-0.162	400032250	Non-specific lipid-transfer protein	0.683
400024391	Tyrosine-specific transport protein	-0.162	400010145	Cysteine protease inhibitor 9	0.679
400024389	ATPase	-0.162	400022430	Polyphenoloxidase	0.651
400025165	ISPH protein	-0.161	400019517	Chitin-binding lectin 1	0.630
401009242	I-box binding factor	-0.156	400011751	2-oxoglutarate-dependent dioxygenase	0.575
400020570	26S protease regulatory subunit 6B homolog	-0.156	400009513	Aspartic protease inhibitor 5	0.550
400004594	Serine/threonine-protein kinase PBS1	-0.152	400028048	Pectin methylesterase inhibitor isoform	0.546
400027765	Peroxisomal biogenesis factor	-0.151	400004547	Proteinase inhibitor type-2 P303.51	0.537
402000056	Bypass1	-0.151	400004548	Proteinase inhibitor type-2 CE-VI57	0.537
400024037	Ribosomal protein L19	-0.150	400019110	Chalcone synthase 2	0.527
400046276	Conserved gene of unknown function	-0.148	400002880	Proline-rich protein	0.527
400025777	Poly(A)-binding protein	-0.148	400009511	Aspartic protease inhibitor 8	0.517
400025368	Gene of unknown function	-0.147	400013010	24K germin	0.514
400031295	4F5 protein family protein	-0.146	401001384	Protein GAST1	0.500
400007830	Eukaryotic translation initiation factor 5A-1/2	-0.146	400003044	Osmotin	0.485
400012630	Cell division cycle protein cdt2	-0.145	400027944	RNA binding protein	0.484
400030476	26S protease regulatory subunit 6B homolog	-0.145	400012019	DC1.2 homologue	0.481

Since apoptosis is a process that needs to be very strictly controlled, it is interesting that PGSC0003DMG402001978, which encodes an apoptosis-inducing factor, seems to be under strict control. Moreover, genes involved the cell division cycle (such as PGSC0003DMG400031295 and PGSC0003DMG400012630) are also found to be strictly controlled. Several protease inhibitors, which are involved in plant defense, were found to have a highly variable gene expression. This could imply that these genes are only expressed if an individual plant is subjected to a specific attack (e.g. a specific fungus). However, since this class of proteins is also utilized as storage proteins in the tuber, the extreme expression needed in tuber tissue to support storage protein synthesis might eliminate the possibility of strict expression control in other tissues as well.

4.4 Discussion

Elucidation of differences in the transcriptome reflected as phenotypical differences between biological groups is most often the goal of a transcriptome study. The analyses presented here, clearly show that a substantial amount of noise is present in DeepSAGE gene expression data, which complicates this goal. This noise was also present internally in the biological groups represented by 47 replicates; cf. Figure 4-4, hereby showing that the source of the noise is not poor estimation of the mean expression due to a low replicate number, but that it must originate from something else.

The comparison between the mRNAseq and DeepSAGE data sets revealed that a substantial fraction of the observed variance between libraries must originate from biases caused by the DeepSAGE library preparation procedure, since this noise was not present in the mRNAseq data set, cf. Figure 4-9. In fact, the intrinsic noise in the mRNAseq data set was well modeled by the Poisson distribution, and hence is likely to originate from sampling, cf. Table 4-3. Therefore it can also be concluded that the technical variation of mRNAseq data is extremely low. The additional noise in the DeepSAGE data compared to the mRNAseq data must originate from differences between the library preparation procedures of the two methods. DeepSAGE library preparation consists of several additional steps compared to that of mRNAseq, including digestion with the anchoring and tagging enzymes, cf. Figure 1-9. These additional steps must inevitably lead to additional loss of sample material prior to PCR amplification. Furthermore, 28 PCR cycles are used in the DeepSAGE procedure, whereas only 15 cycles are used in the mRNAseq procedure, which could be suspected to cause an amplification bias. Others have shown that biases such as a GC content bias exist in the SAGE procedure (Margulies, Kardia & Innis, 2001), but since these presumably affect all technical samples equally, it cannot be used to explain the intrinsic noise observed. PCR amplification could be investigated as a potential source of noise by replication of the PCR amplification step on the same sample that has been subject to the prior steps in the DeepSAGE procedure. Technical replicates are hereby created, and a subsequent analysis of the variation between the resulting libraries could be performed. If the observed variation would be in the same order of that found within triplicates originating from the same leaf in the current study, cf. Figure 4-9, it could be concluded that the PCR amplification was the source of the noise. This could in part be explained by a "PCR founder effect". If the chance of low abundant transcripts to be amplified in the first cycles of the PCR amplification is low, some low abundant transcripts might be amplified in some samples, but not in others. Following 28 cycles of PCR amplification, these low abundant transcripts would either be represented by an artificially high tag count (e.g. 20) or not at all. This pattern has in fact been observed by the author in the current study and in other DeepSAGE data sets for many low abundant transcripts. If the bias is caused by overamplification, lowering the cycle number in the PCR amplification would be the obvious choice to decrease the noise. Matsumura *et al.* have shown that as low as 10 cycles is enough to produce a sufficient amount of tags (Matsumura *et al.*, 2010), although this must be sample dependent. However they used 2-3 times the amount of starting material, resulting in the requirement of two additional PCR cycles to gain the same amount of product when using the same amount of starting material as in the DeepSAGE procedure. If a significant lower amount of noise would be the result of the investigation, PCR amplification could be ruled out as the noise source. The variation would then have to be caused by

steps in the procedure prior to PCR amplification. As described above, random loss of material in the steps prior to PCR amplification, followed by PCR amplification can lead to a founder effect. This would especially affect low abundance transcripts, which would not be represented in all samples, and would lead to the same tag count pattern as described above.

The comparison between the mRNAseq data and the DeepSAGE data also showed that the negative binomial distribution with a common dispersion estimate is sufficient to account for the biological variance found in the mRNAseq data set, cf. Table 4-3. The extrinsic noise seems to be independent of the gene expression level and can therefore be accounted for by a common dispersion estimate, cf. Figure 4-9. However, the analysis also showed that this model provides a poor dispersion estimate in the DeepSAGE data, since there is a significant amount of expression level dependent intrinsic noise present in the data. Here, tagwise estimation provides a far better dispersion estimate, but due to the fact that the tagwise dispersion is "squeezed" towards a common dispersion estimate, which obviously is wrong, this model also performs relatively poorly, cf. Table 4-3. In fact an extension of the negative binomial model using a simple Taylor series with only one extra expression dependent term is a better model of the variance observed in the DeepSAGE data sets. A possible implementation of this could be to replace the tagwise dispersion estimates found using the algorithm implemented in EdgeR with the expression dependent dispersion estimate found using the Taylor series model prior to DE determination. Another possible solution would be to apply the algorithm of Anders and Huber implemented in the DESeq package (Anders & Huber, 2010). Here, the dispersion is estimated using mean-dependent local regression to avoid biases in the DE detection over the dynamic range of gene expression. This of course provides more flexibility than the Taylor Series model and is likely to fit the mean-variance in the data better, since the Taylor Series fits the variance in the entire data set only using 2 parameters. However, the data presented here, suggests that these two parameters are enough to fit the mean-variance dependency. Furthermore, the Taylor Series have the advantage of simplicity, and immediately provides interpretable indicators of the variance structure in the data set. The CV formula in Table 4-2 can be divided into the sum of three terms, namely $\sqrt{D_1} + \frac{\sqrt{\mu}}{\mu} + \frac{\sqrt{\pi \cdot D_2}}{\mu}$. Here, $\sqrt{D_1}$ accounts for the biological variance, $\frac{\sqrt{\mu}}{\mu}$ accounts for the sampling variance and $\frac{\sqrt{\pi \cdot D_2}}{\mu}$ accounts for the short noise, which the data presented here suggest is technical variance. Following this, the CV at $\mu = 1$ is $\left(\sqrt{D_1} + 1 + \sqrt{D_2}\right)$, and $CV \rightarrow \sqrt{D_1}$ for $\mu \rightarrow \infty$. Therefore $\sqrt{D_1}$ can be interpreted as a measure of biological variance, and $\sqrt{D_2}$ can be interpreted as a measure of the shot noise originating from technical variance. The model underestimates the variance for highly expressed genes, and only using genes with an expression level > 100 CPM shown an improvement for the highly expressed genes, especially in the case of the DeepSAGE data sets, cf. Figure 4-10. This result favors the method of Anders and Huber, where the variance is estimated using local regression.

The analysis of the high-replicate groups showed that the sensitivity dropped dramatically, when lowering the replicate number to a triplicate, cf. Figure 4-5. This is naturally a cause of concern, since nearly all biological groups in the LSDS-project are represented by triplicates only. On the other hand, it is encouraging that the specificity of the genes determined as DE was found to be high, and that the sensitivity of the most significantly DE genes was retained to a higher degree. This means, that genes found to be DE using a low replicate number are

likely to be true positives, and that these are truly the genes with the highest differential expression between the biological groups. Therefore these are most likely to be the cause of the phenotypical differences between the biological groups. Moreover, the gene expression analysis was performed on two very similar biological groups (both healthy leaf tissue), and large differences in the gene expression can therefore not be expected. When larger phenotypical differences are present, DE detection between biological groups represented by triplicates only is more fruitful. For instance, a comparison of a control group and a late blight infected group (both leaf tissue from cv. Kuras) 11 days after infection yields ~ 500 DE genes (data no shown). Therefore, the results of the current study challenge the fundamental assumption that most genes are not differentially expressed, and that the phenotypical differences observed are explained by a small subset of the transcriptome. This assumption is used in various parts of the data analysis; for example during calculation of gene expression levels when accounting for biases introduced by the RNA composition (highly expressed genes in one biological group consume a substantial proportion of the sequencing power and the remaining genes in the library are therefore under sampled) (Robinson & Oshlack, 2010). Undoubtedly, a higher number of replicates would lead to a higher number of DE genes detected, and it could be speculated that using an even higher number of replicates, would lead to detection of almost every gene as DE, even though the differences in the gene expression levels would be very small between the biological groups. This is supported by the PCA analysis, where the two biological groups clearly split out, but where the loadings plot indicate that this splitting is caused by many genes, cf. Figure 4-2. Now, the question of how many DE genes are needed to explain the phenotypical differences between the biological groups arises, i.e. how many DE genes are biological relevant?

Although there is correlation between library size and the power to detect differentially expressed genes, raising the sequencing depth of each library, does not improve the power of detection considerably, cf. Figure 4-6. It seems, that in order to capture the variance found internally in the biological groups either a high number of replicates or pooling of many samples is needed, cf. Figure 4-7. However, if the internal variance in a biological group to a large extent is caused by technical variance, which the data suggests, pooling samples prior to library preparation will not lower the variance. Here, the only solution would be to pool multiple libraries prior to sequencing, which would be extremely costly and time consuming, and therefore not an option in reality.

At first glance the large collection of DeepSAGE libraries in the LSDS-project provides the possibility to answer several interesting biological questions regarding variance in gene expression data. One interesting fundamental question of transcriptomics is whether all genes are under the same amount of regulatory control. The investigation of the biological variance of the expression could elucidate this. However, the analysis of the data quality showed that the expression of a significant amount of the genes observed, was “noisy”, and an in-depth analysis of the variance therefore only seemed possible for relatively abundant transcripts, because the analysis requires a reliable estimate of the biological variance of each gene. Among the genes found to be under stricter regulation were genes known to be a part of apoptosis or the cell division cycle, and several protease inhibitors were found to have a more variable than other genes with the same expression level. These results make biological sense and support the hypothesis that certain genes are under stricter regulatory control than others. However, the hypothesis cannot be confirmed due to the level of noise found in data

set. Due to this, only the genes with the highest expression were included in the analysis, and intuitively these are less variable than lowly expressed genes. Several low abundant transcripts are by nature more variable. For example a response, which requires a rapid transcriptional change, is easier to employ by altering the expression of e.g. a lowly expressed transcription factor, than altering the expression of an abundant protein.

Finally, an overall analysis of the variance observed in the different data set was performed, cf. Figure 4-14. Here it was shown that additional variance is present in data sets consisting of field grown potato plants compared to data sets consisting of greenhouse grown plants. This is of course to no surprise, but must be taken into account when analysing the data. As a consequence, it should not be expected that small differences between cultivars are elucidated easily. Therefore comparisons should be performed between biological groups with larger phenotypical differences exist. From these hypotheses can be formulated and tested between biological groups where smaller phenotypical differences exist. Here, the size of the data set is an obvious advantage, since it provides the possibility to pool many samples; e.g. many cultivars with similar phenotypes for a trait and test this biological group against another pool of other cultivars with different phenotypes for that trait.

Chapter 5

Genome Sequence and Analysis of the Tuber Crop Potato

5.1 Introduction to the Potato Genome Sequencing Project

In 2005, the establishment of an international consortium capable of sharing the tasks needed to determine the genome sequence of *S. tuberosum* was initiated. The Potato Genome Sequencing Consortium (PGSC) was originally a collaboration between 13 research groups from China, India, Poland, Russia, the Netherlands, Ireland, Argentina, Brazil, Chile, Peru, USA, New Zealand, and the UK (Visser *et al.*, 2009). Today, the PGSC is a global community with members from 26 research groups worldwide, including the Functional Genomics group at Aalborg University (The Potato Genome Sequencing Consortium *et al.*, 2011). The PGSC originally started out by sequencing a diploid heterozygous potato variety named RH89-039-16 (from here on referred to as RH). The sequencing of RH was based on a bacterial artificial chromosome (BAC) clone library consisting of 78,000 clones. The strategy for sequencing RH was a BAC-by-BAC strategy consisting of shotgun sequencing of individual adjacent BACs (Visser *et al.*, 2009). However, the heteozygocity of RH limited the overall progress of the assembly of the RH genome (The Potato Genome Sequencing Consortium *et al.*, 2011). This fact and the advent of NGS technologies made the PGSC change sequencing strategy. In 2008, sequencing of DM1-3 516R44 (from here on referred to as DM), a doubled monoploid was initiated. The strategy for sequencing DM was chosen to be whole genome shotgun (WGS), and already in September 2009, the first draft genome assembly for DM was made public available (PGSC, 2009).

In 2009, the Functional Genomics group at AAU (Kåre Lehmann Nielsen and I) joined the PGSC. Initial involvement included sequencing of 17 mRNAseq libraries from various tissues of RH. The initial primary aim of our involvement was to perform gene expression analysis of the RH mRNAseq libraries. However, following sequencing of both DM and RH mRNAseq libraries, multiple opportunities arose. In the following, analyses performed as a part of the current PhD-project is described. Firstly, the gene annotation of the first draft genome sequence was built on homology-based, EST-based and *ab initio* gene prediction, and lacked information about un-translated regions (UTRs). The mRNAseq data made a prediction of UTRs possible and an experimental validation of gene models. These analyses are described in the sections 5.2.2 and 5.2.3, respectively. At a later state in the genome sequencing project, genome assisted transcriptome reconstruction was performed to improve the gene annotation. To investigate the quality of this annotation, manual gene validation and curation of a small subset was performed. These analyses are described in the sections 5.2.4 and 5.2.5, respectively. In section 5.2.6, the overall gene expression analysis of DM and RH mRNAseq libraries from various tissues is described. Furthermore, a detailed transcriptome analysis of the starch metabolism genes, comparing the DM and RH genotypes is described in section 5.2.7.

Throughout the improvement of the genome sequence assembly and annotation, I have been involved and performed a great deal of data quality control. This sometimes underestimated bioinformatic task is an absolute necessity, but is sometimes forgotten. While out of the scope of the current thesis, this will not be described in details. However, a few comments and suggestions regarding this will be given at the end of this chapter. Throughout the chapter, specific references to tables and figures will be given to the published article e.g. as (cf. Fig-

ure, 4C p. 5 (The Potato Genome Sequencing Consortium *et al.*, 2011)) for a citation in the main article and (cf. Supplementary Table 4., pp. 35-40 (The Potato Genome Sequencing Consortium *et al.*, 2011)) for a citation in the supplementary text. The full article can be found in appendix A and the supplementary text can be found on the enclosed CD in the file "*Genome Sequence and Analysis of the Tuber Crop Potato Supplementary text.pdf*". Many parts of the analyses for the published article have of course been a collaborative effort involving many members of the PGSC. However, the analyses presented in the current thesis are solely performed by the author if nothing else is stated.

5.1.1 The Potato – a Food Crop, it's Genetics, Physiology and Genome

Potato (*Solanum tuberosum* L.) is a member of the *Solanaceae* family, which also includes other economically important species such as tomato (*Solanum lycopersicum*), pepper (*Capsicum annuum*), eggplant (*Solanum melongena*), and tobacco (*Nicotiana tabacum*). *S. tuberosum* is a eudicot belonging to the asterid clade, which represents 25% of flowering plants. The publication of the potato genome in July 2010 was the first in the asterid clade (The Potato Genome Sequencing Consortium *et al.*, 2011). The taxonomic classification of *Solanum tuberosum* L is somewhat discussed (Gopal & Khurana, 2006). However, Huamán and Spooner have proposed that all landrace populations of cultivated potato plants are a single species, *S. tuberosum*, which can be divided into 8 cultivar groups, and that all modern cultivars of potato belong to the tuberosum group (Huamán & Spooner, 2002). *S. tuberosum* is tetraploid, and the genome is highly heterozygous. The haploid potato genome consists of 12 chromosomes with a total length estimated to be 844 million base pairs by flow cytometry (Arumuganathan & Earle, 1991). Modern Potato plants suffer from acute inbreeding depression and are susceptible to many devastating pests and pathogens, exemplified by late blight disease caused by the oomycete *Phytophthora infestans* (The Potato Genome Sequencing Consortium *et al.*, 2011). The potato plant can grow ~ 1 m tall and produces underground tubers, cf. Figure 5-1. The tubers are highly specialized storage organs (Chapman, 1958), and are formed by elongation of stolons followed by swelling of the stolon tip (Li, 1985).

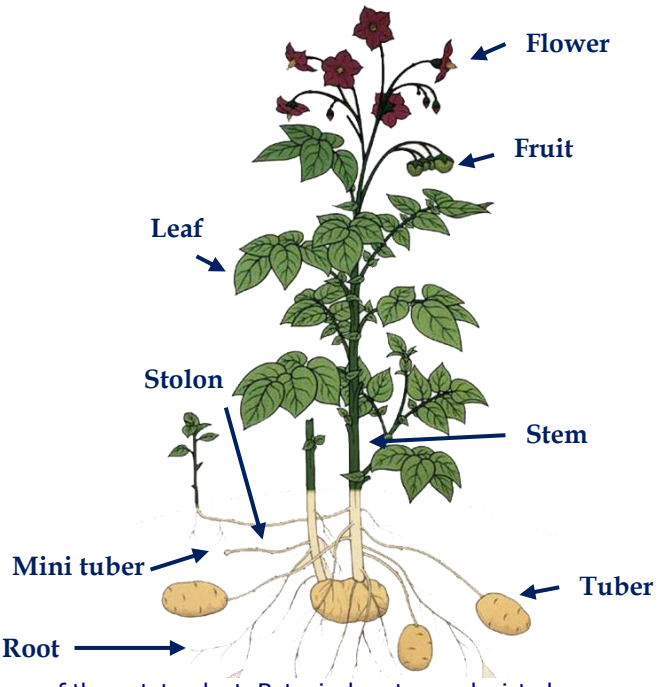


Figure 5-1 Anatomy of the potato plant. Botanical parts are depicted

Potato is the world's third most important food crop after rice and wheat, cf. Figure 5-3 panel A (FAOSTAT, 2010). It is central to global food security, for instance due to its global distribution, cf. Figure 5-2.

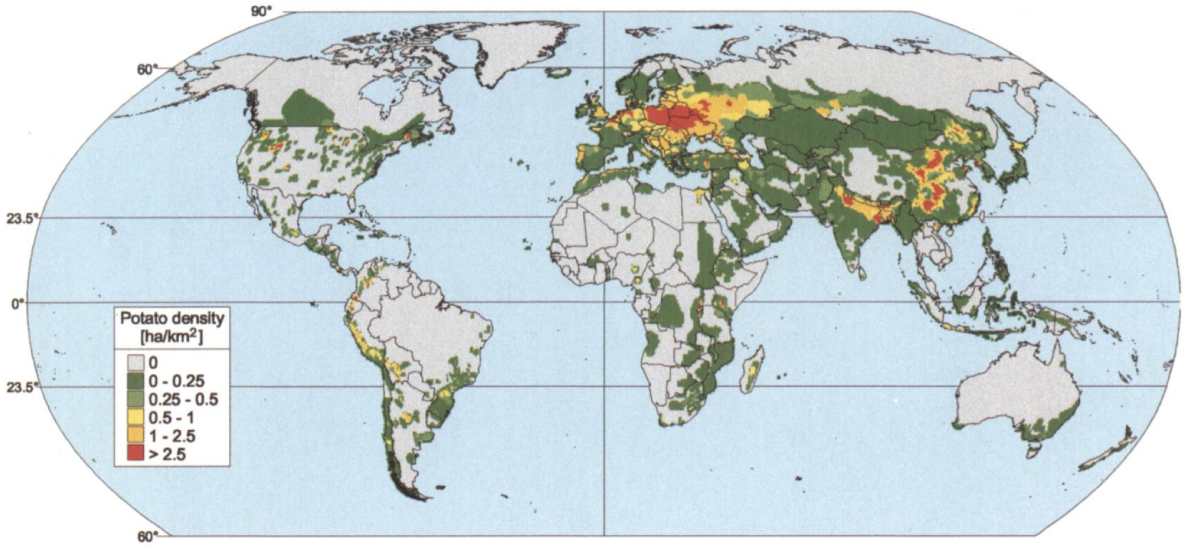


Figure 5-2 Global potato distribution. The relative potato area [ha/km²] is shown. Source: (Hijmans, 2001).

Potato outperforms rice, wheat, and other cereals in terms of total yield with a production 4 and 6 times greater per hectare than rice and wheat, respectively (FAOSTAT, 2011a), cf. Figure 5-3 panel B. More interestingly in regards to food production, potato fields produce more energy, namely 11.7 million kcal/Ha compared to rice and wheat fields, which only produce 10.3 and 8.3 million kcal/Ha, respectively (FAOSTAT, 2011a; FAOSTAT, 2010), cf. Figure 5-3 panel C. Furthermore, these numbers include lower yielding gourmet potato varieties. This is not the case for wheat and rice, why the potential energy difference could be even greater. Most potatoes are consumed in Asia, hence potatoes are an important food crop in this high populated region, cf. Table 5-1. In the western world, potatoes are a big part of the diet. Europeans

for example, on average eat 94 kg potatoes per year. Within the last 40 years the potato has become a more and more important food crop in low income food deficit countries, whose share of the worldwide potato food supply have risen from ~ 11 % in the early 1960's to more than 40 % % in 2007, cf. Figure 5-3 panel D (FAOSTAT, 2010).

Table 5-1 Potato consumption in 2007 by region. The total consumption and the consumption per capita are shown. Source (FAOSTAT, 2010).

Region	Total [million ton]	Per capita [kg/capita/year]
Africa	13.5	14.2
Northern America	19.5	57.0
Central America	2.3	15.6
South America	11.0	28.9
Asia	93.9	23.7
Europe	66.8	91.4
Oceania	1.5	53.3
World	208.7	31.7

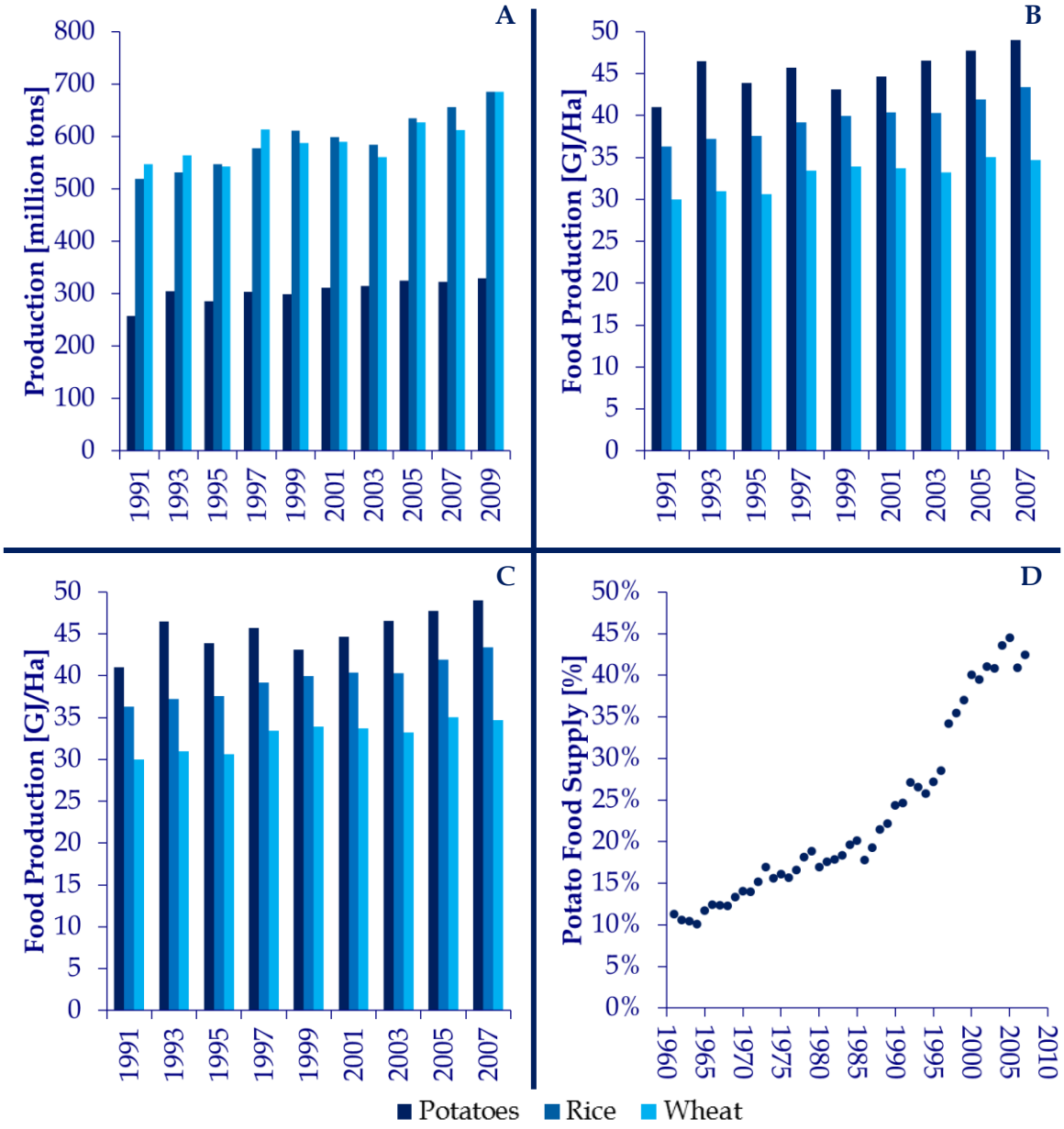


Figure 5-3 Production and yield for the world's three most important food crops; wheat rice, and potato in different years. **A)** Production in million tons. **B)** Production yield measured as tons per hectare. **C)** Production yield measured as energy per hectare. **D)** Food deficit countries' share of the worldwide potato food supply. GJ = Giga Joule. Ha = hectare. Source: (FAOSTAT, 2010).

5.2 Data Analysis of mRNAseq Data for the Potato Genome Sequence Project

5.2.1 Versions and Nomenclature of the Potato Genome Sequence

Throughout the course of the potato genome project, multiple versions of the genome sequence, genome annotation, scaffold to super scaffold mapping, and super scaffold to pseudomolecules mappings were released to consortium members. Within little more than a year 6 versions of the genome sequence and accompanying annotation files were released internally in the consortium, cf. Table 5-3. This was caused by continuous improvements and error corrections made. However, this causes some of the analyses presented in the current thesis to have been conducted on outdated and not public available versions. Some of the analyses presented here, have even given rise to version updates (e.g. the update from version 3.1 to 3.2). In Table 5-3, versions of the genome sequence and the gene annotation are listed along with comments on the differences. To secure identifier uniqueness and to some extent facilitate conversion between data sets and versions, a naming strategy of the different data types was made. All identifiers must have the following structure: PGSCxxxxYYzVIIIIIIII. Here, "xxxx" is a unique, four character identifier for the dataset (version of the genome assembly). "YY" is a two letter code for the strain or cultivar (DM or RH). "z" is one of the object type identifiers listed in Table 5-2.

Table 5-2 Identifiers for different object types used in the potato genome sequencing consortium (PGSC).

Identifier	Object type
S	Scaffold
N	Gap
I	Intron
E	Exon
G	Gene
T	Transcript
P	Peptide
C	CDS (complete coding sequence)
J	CDS segment (coding sequence from one exon)
B	Super scaffold
L	Linkage map
M	Pseudomolecule (golden path)
O	Singleton contig
H	3' UTR region
F	5' UTR region

Each identifier ends with a nine character and zero padded identifier for that entity. This should be unique in dataset xxxx. Version numbers (V) should be included in this nine digit number at the left hand side. The number of digits should remain constant. Moreover, in version 3.4 some genes are split into several genes due to non-overlapped of transcripts, and the name PGSC0003DMG#########.# is used. For example, PGSC0003DMG200000007 is split into PGSC0003DMG400000007.1 and PGSC0003DMG400000007.2

Table 5-3 Versions of the potato genome sequence and gene annotation. The data set version number X.Y indicates the version of the genome assembly (X) and the version of the gene annotation (Y), i.e. DM 3.4 contains version 3 of the genome sequence, and version 4 of the gene annotation. NcRNAs = non-protein coding mRNA transcripts.

Data set	Release Date	Comments
DM 2.0	2009/09/18	First version released to the consortium. Genome assembly only contains scaffolds. Gene annotation is a combination of homology-based, EST-based, and *ab initio* gene prediction methods. Each gene contains one transcript variant, and only information about the coding sequence is given (no UTRs are predicted). Annotation contains 40,322 gene models.
DM 3.0	2009/12/22	Updated and improved genome assembly. The gene annotation is performed by the same procedure as version 2.0. Genome assembly only contains scaffolds. Annotation contains 40,842 gene models of which 27,099 are supported by mRNAseq.
DM 3.1	2010/03/23	Updated version of the gene annotation; Cufflinks predicted transcripts are now incorporated. Genes predicted by Cufflinks can contain more than one transcript variant and information about the UTR regions. Super scaffold information is added for the version 3 genome assembly. Annotation contains 47,352 gene models encoding 70,885 mRNA protein coding transcripts of which 47,190 have been predicted using mRNAseq data. Moreover, 26,896 ncRNAs are also predicted using mRNAseq data.
DM 3.2	2010/04/08	Correction of errors in the version 3.1 annotation, found by the author of the current thesis. Coding sequences, which are not on the same strand as their corresponding transcripts are removed. Annotation contains 47,352 gene models encoding 69,456 mRNA protein coding transcripts of which 45,761 have been predicted using mRNAseq data. Moreover 28,325 ncRNAs are also predicted using mRNAseq data.
DM 3.3	2010/08/30	Updated version of the 3.1 gene annotation after quality control. Version 3.3 differs from 3.1 in 6 aspects: 1. Genes crossing gaps are only kept if transcripts are supported by EST or protein support. 2. The genes of which length is smaller than 300bp are excluded. 3. The transcripts related with transposon and related gene loci are excluded. 4. Only one copy of genes sharing the same coordinate but on different strand with CDS overlapping coding sequences is included. 5. Identical protein sequences originating from multiple genes are flagged. 6. Only one sequence is kept in the protein fasta file, when identical proteins exist from the same locus due to multiple transcript variants that encode the same peptide. Annotation contains 41,197 gene models containing 62,491 mRNA protein coding transcripts of which 41,915 have been predicted using mRNAseq data and 23,586 ncRNAs also predicted using mRNAseq data.
DM 3.3b-j	2010/09/02 - 2010/11/11	9 updates of the 3.3 gene annotation. These contain iterative quality filterings, which were incorporated in version 3.4.
DM 3.4	2010/11/15	Updated version of the 3.3 gene annotation after additional quality control. Version 3.4 differs from 3.3 in 3 aspects: 1. Stricter homology criteria for genes, which length is greater than 10k. (EST identity: >=95%, EST coverage: >=90%, identity >=80%, and coverage >=80% with protein database) and intron length must be smaller than median size + 2*(Standard Deviation). 2. Some genes are split into several genes due to non-overlapped of transcripts. 3. All the noncoding transcripts are excluded in this version. Annotation contains 39,031 gene models containing 56,218 mRNA protein coding transcripts of which 37,581 have been predicted using mRNAseq data. Moreover this version contains a collection of representative transcripts. The transcript giving rise to the longest protein sequence is chosen for each gene model. This version was the version used for the publication of the potato genome sequence in Nature (The Potato Genome Sequencing Consortium *et al.*, 2011).

5.2.2 Experimental Prediction of Un-translated Regions using mRNAseq data in the Version 3.0 Data Set

The role of mRNA un-translated regions (UTRs) is crucial in many post-transcriptional regulatory pathways (Pesole *et al.*, 2001). These pathways that control mRNA localization, stability and translation efficiency are often mediated by cis-acting RNA elements that are generally located in 5′ UTRs (Schwartz *et al.*, 2006; Pesole *et al.*, 2000; Sonenberg, 1994). Hence, annotation of mRNA transcripts including the UTRs in a genome sequence can facilitate more complex analyses of transcriptome regulation. Moreover, annotation of UTRs can greatly improve the tag to gene mapping in tag based transcriptomics, since the primary SAGE tag often is located in the 3′ UTR. As mentioned, the first gene annotations of version 2.0 and 3.0 of the potato genome sequence were based on a combination of homology-based, EST-based, and *ab initio* gene prediction methods, cf. Table 5-3, why the gene models lacked information of the UTRs. The experimental prediction presented here, is based on a reference assembly of mRNAseq libraries to the DM genome sequence. 5' and 3′ UTR regions are predicted by up/downstream extension of the first/last exon, respectively. The extension is stopped when the mRNAseq support (the coverage of mapped reads) drops below a threshold value, cf. Figure 5-4 panel B.

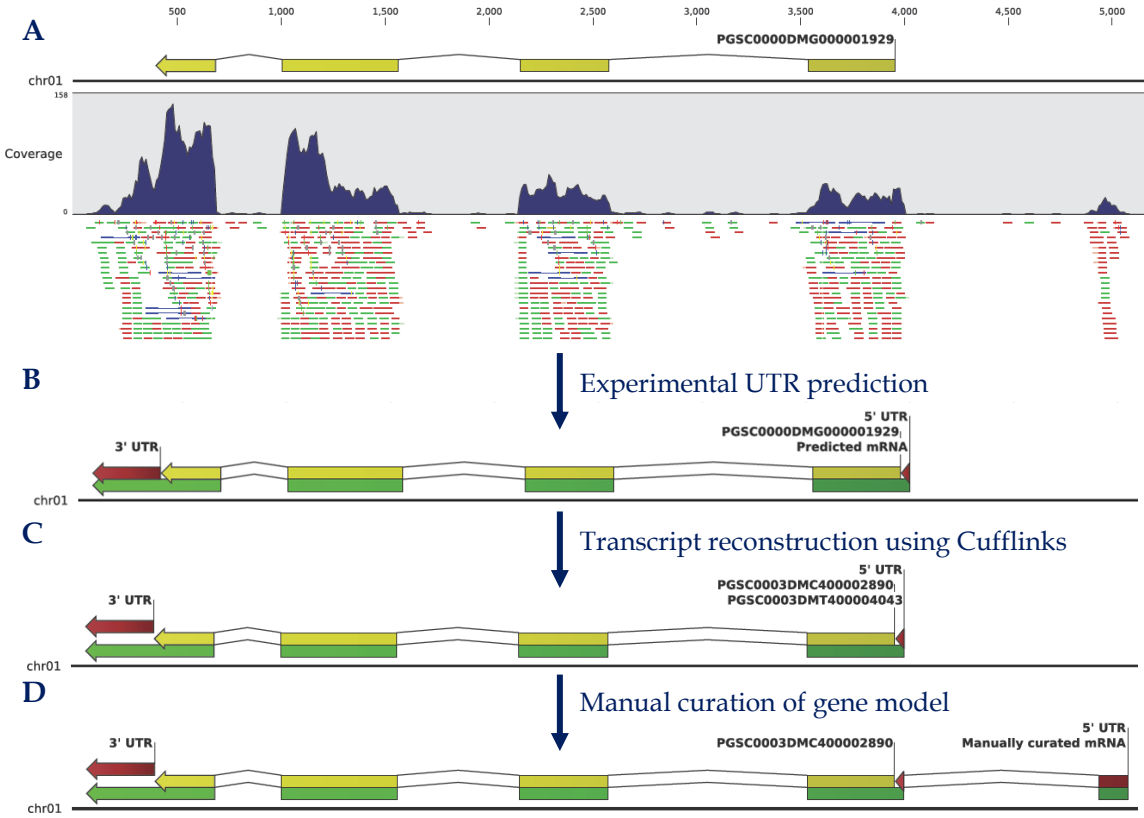


Figure 5-4 Gene prediction of a gene region encoding an invertase in the V3 potato genome. **A)** Annotation in DM V3.0. This gene annotation only contains information on the coding part of the mRNA transcript (marked in yellow). Coverage is shown below the gene model. Reads are indicated as follows: Green = Forward matching, red: reverse complement matching, yellow = randomly matching. **B)** Result of the experimental UTR prediction described in the current section. 5' and 3′ UTR regions (marked in red) are predicted by up/downstream extension of the first/last exon, respectively. The extension is stopped when the coverage of mapped reads drops to 0. **C)** Result of transcript reconstruction using Cufflinks. This is the annotation for this gene model in the final version (DM V3.4), which matches the annotation resulting from the experimental UTR prediction. **D)** Result of annotation after manual curation of a gene model containing introns in the 5′UTR. The exon has been missed both in version 3.0 after UTR prediction and in the final version (V3.4).

5.2.2.1 Methods

All 32 DM and 16 RH RNAseq libraries (NCBI Sequence Read Archive (SRA030516; study SRP005965[24]) and the European Nucleotide Database ArrayExpress Database (E-MTAB-552; study ERP000527[25]), respectively) were mapped to the DM V3.0 genome sequence using the CLC Genomics workbench 3.7.1. Reference assembly was perform using un-gapped alignment (end trimming allowed), and random match mode, otherwise default settings. Coverage values were reported and tabulated for each position. Using *PredictUTRs.pl* 3′ and 5′ UTRs were predicted and annotations of these were added to a GFF annotation file. To optimize UTR predictions, these were performed using varying values of minimum coverage of the first base in the UTR and minimum coverage of the UTR. The length distribution of the UTRs was compared to UTRs of *Arabidopsis thaliana (A. thaliana)*[26]. Using *GetUTRsAndExons.pl* and *CreatemRNAseqs.pl* a fasta file containing the predicted mRNA transcripts was created. Detailed description on script usage and resulting files can be found in appendix B.

5.2.2.2 Results

Using varying values of minimum coverage of the first base in the UTR and minimum coverage of the UTR between 47% and 61 % of the 40,842 gene models had 5′ UTRs and/or 3′ UTRs predicted, cf. Table 5-4. Analysis made by Beijing Genomics Institute (BGI) showed that 27,099 of the gene models had mRNAseq support. 93 % of these had mRNAseq support in the 3′ and/or 5′ end of the predicted CDS (minimum coverage = 1, cf. Table 5-4), making UTR prediction possible.

Table 5-4 Experimental gene prediction using different parameters for the minimum start coverage (S) and the minimum coverage (C). If the minimum coverage gets below the threshold value, the extension in the 3′ or 5′ end of the predicted transcript is stopped. If the coverage at the end of the predicted CDS is below the start coverage, the extension is not started. The number of 5′ and 3′ UTRs, and the mean length of these are shown. See text for more details.

Start coverage	Minimum coverage	# 5'UTR	# 3'UTR	Mean 5'UTR length [bp]	Mean 3'UTR length [bp]
S = 1	C = 1	24,957	25,208	223	354
S = 5	C = 1	21,145	21,550	253	398
S = 10	C = 1	19,374	19,942	263	413
S = 10	C = 5	19,374	19,942	168	296
S = 10	C = 10	19,374	19,942	134	251

To optimize the experimental UTR prediction, the length distribution of the UTRs was compared to predicted UTRs of *A. thaliana*, cf. Figure 5-5. Not surprisingly, more stringent parameters result in shorter predicted UTRs. Especially, a higher threshold value for the minimum UTR coverage will produce shorter UTRs, whereas a higher minimum start coverage does not seem to affect the length to the same extent, cf. Figure 5-5. In general, the UTRs are predicted to be longer in *S. tuberosum* compared *A. thaliana*. This may not be surprising, since *A. thaliana* is a faster growing plant. There is therefore a higher evolutionary pressure on *A. thaliana* to shorten unneeded translated sequence compared to *S. tuberosum*. This was also confirmed, when comparing gene features such as mRNA-, CDS-, exon-, and UTR-length, and number of exons per gene of the version 3.2 potato gene annotation against *V. vinifera*

[24] Available at: http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP005965
[25] Available at: http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-552
[26] Available at UTRdb: http://utrdb.ba.itb.cnr.it/advdownload

(grape) and *A. thaliana*, cf. section 5.2.4.2. Here, *V. vinifera* and *S. tuberosum* both had longer UTRs than *A. thaliana*, while the overall statistics for the other features were similar.

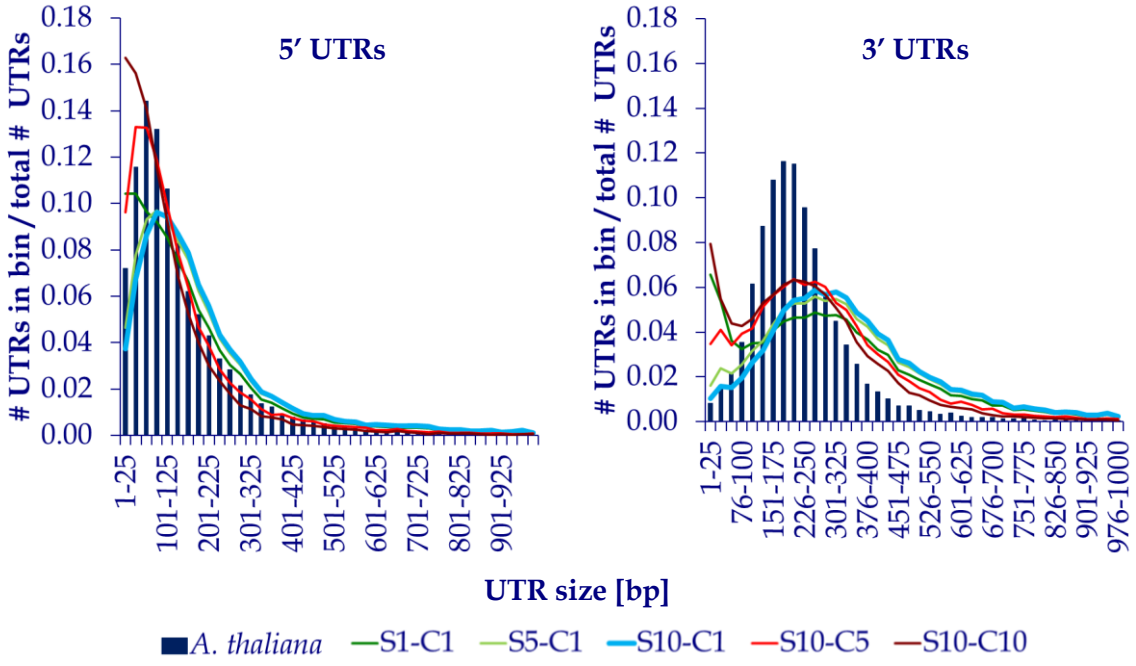


Figure 5-5 Distributions of the predicted 5'UTR and 3'UTR length using different parameters. Data is shown of combinations of the start coverage (S) and the minimum UTR coverage (C). Discrete distributions are shown as solid lines for visualization purposes.

When comparing the length distribution of the predicted UTRs in the V3.0 data set, with the set of representative transcript models in the final version (V3.4), these are also similar, cf. Figure 5-6. The predicted UTRs of the V 3.0 data set are longer than the UTRs in the data set of representative transcript models in the V3.4 data set. Reasons for this are the coverage threshold values used for prediction, both in the presented method and for Cufflinks (Trapnell *et al.*, 2011), which was used to predict a part of the transcripts in the V3.4 data set. Another plausible reason is that transcript prediction using Cufflinks is more specific, i.e. contains fewer errors. This can lead to a more correct prediction of the coding sequences, which then becomes longer hereby shortening the UTR of the mRNA transcript.

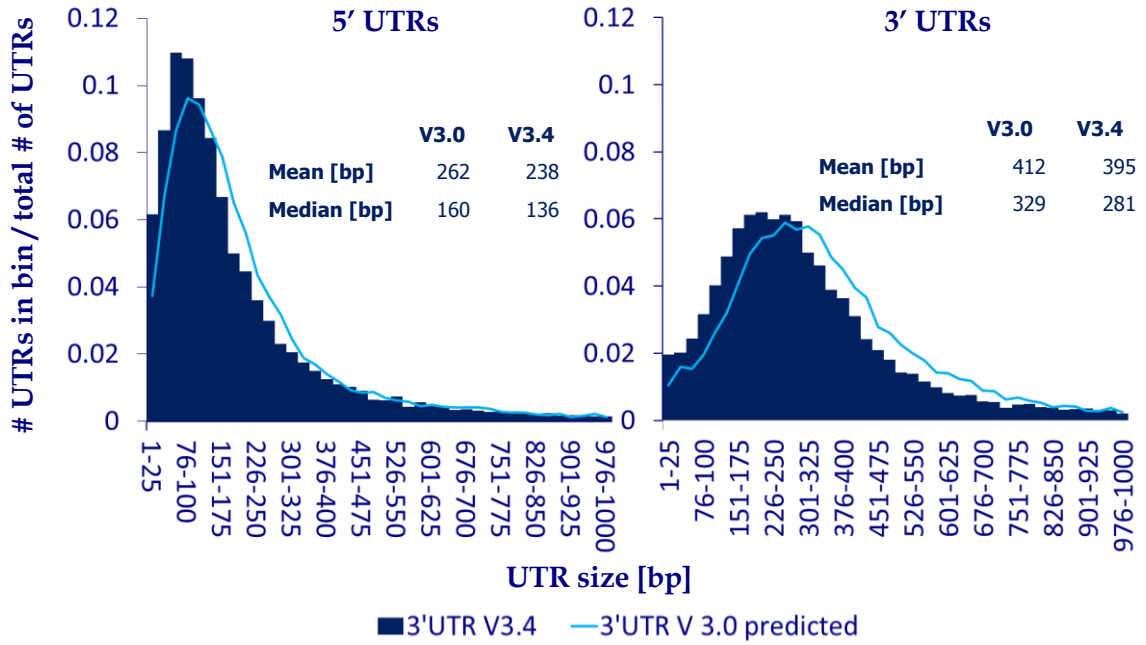


Figure 5-6 Distributions of the predicted 5′UTR and 3′UTR length for the V3.0 data set and the final data set of representative transcript models in the V3.4 data set. A start coverage of 10 and a minimum coverage of 1 were used for the prediction in the V3.0 data set. The mean and median UTR sizes are given. Gene models with no UTRs are omitted.

Manual inspection of several gene model mappings revealed multiple polyA sites in regions spanning several hundred kilo bases, cf. Figure 5-7. This supports the fact that there is extensive heterogeneity in the 3′ end of eukaryotic mRNA transcripts, which also have been reported in other studies (Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008).

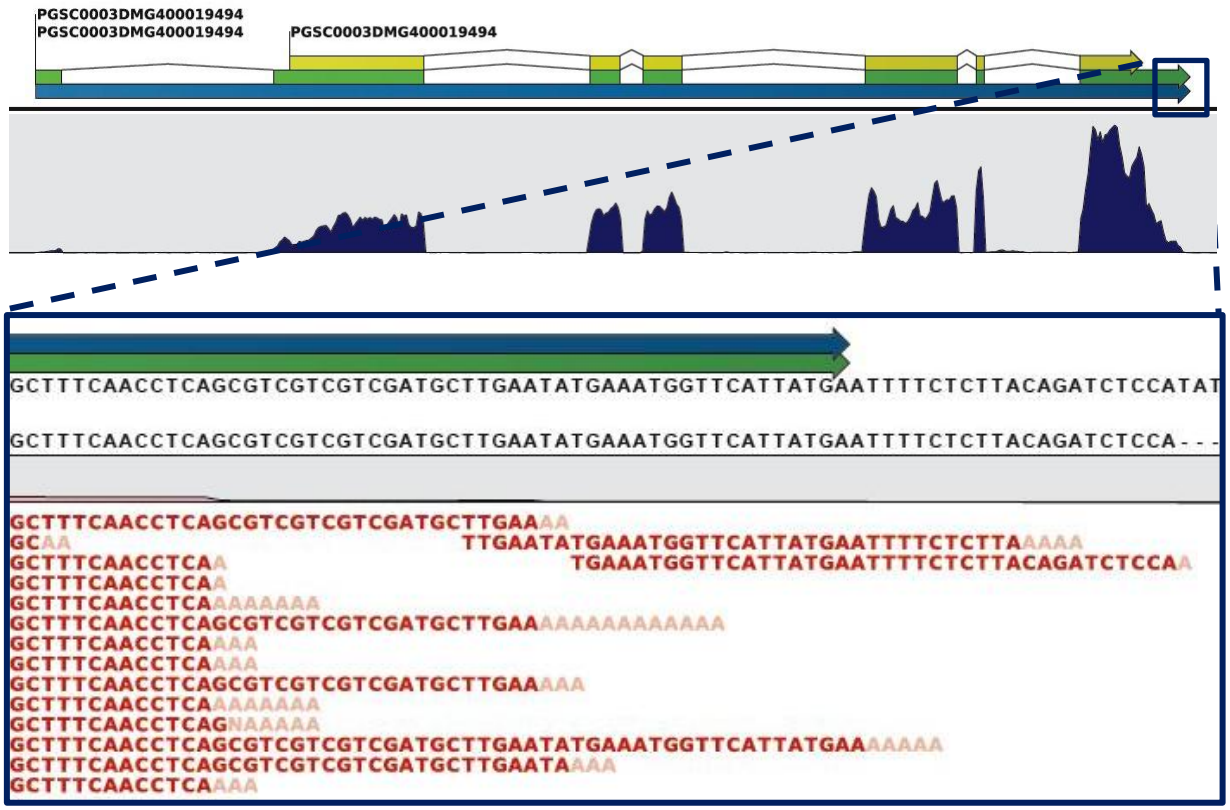


Figure 5-7 Example of multiple polyA sites in a potato gene model. **Top:** Gene model showing the predicted gene region in blue, the predicted mRNA transcript in green, and the coding sequence in yellow. The mRNAseq coverage is shown below the gene model. Notice the sharp coverage drop at exon/intron boundaries, compared to a more steady coverage decrease in the 3' end of the transcript. **Bottom:** Close up of reads matching at the 3' end of the transcript in a region spanning more than 70 bp showing multiple polyA sites of the transcript. Example is from the V3.4 gene annotation of a gene model encoding an Invertase.

5.2.2.3 Discussion and Conclusion

The results presented here indicates that the experimental developed UTR prediction method give similar results to those obtained using Cufflinks for mRNA transcript (and hereby UTR) prediction. However, the method has some limitations. Firstly, the method relies on *a prio* gene prediction using other methods. Errors introduced by these method, will not be corrected and can affect the UTR prediction. Moreover, the method fails to call intron containing UTRs correctly. An example of this is seen in Figure 5-4 panel D (however, Cufflinks also failed to predict the transcript model correctly). Pesole *et al.* have shown that 15 % of 5' UTRs but only 4 % of 3' UTRs in green plants contain introns (Pesole *et al.*, 2001), why this limitation will affect a higher number of 5'UTR predictions.

Initial work was started by the author to discover genes that had been missed in the V3.0 annotation using mRNAseq. However, due to the mentioned limitations of the UTR prediction method, and the release of Cufflinks (which was 3 months prior to the release of the V 3.0 data set (Trapnell *et al.*, 2011)), this work was not incorporated in later versions of the annotation. The UTR prediction method could in theory be used to predict UTRs for genes not based on Cufflinks predictions in later versions. However, these genes will most likely have little or no mRNAseq support, since no gene has been called using Cufflinks, why the quality of a prediction would be questionable.

Another approach proposed by the author to improve or in some cases predict 3' UTRs, was to use information from observed SAGE tags. The SAGE tag is indicative of the mRNA transcript, and the 3'UTR could be extended to the location of the SAGE tag in cases of low expressed gene models missing mRNAseq support. This analysis could be facilitated by the large collection of DeepSAGE libraries from LSDS-project, cf. section 4.1.2. This method would however, suffer from the same limitations as the mRNAseq supported UTR prediction.

5.2.3 Experimental Gene Model Validation of the Version 3.0 Data Set

The mRNAseq data set from DM and RH libraries was also used to validate how many genes were supported by mRNAseq, and how many gene structures could be fully validated by mRNAseq. The experimental gene validation was also based on a reference assembly of mRNAseq libraries to the DM genome sequence. Un-mapped reads for this initial reference assembly were subsequently mapped to the predicted mRNA transcripts, because these reads are potential exon-exon spanning reads, cf. Figure 5-8. Full validation of a gene was defined as: "*The gene must have 100 % coverage and all predicted exon-exon boundaries must have reads spanning the boundary when mapping to the predicted mRNA transcript*".

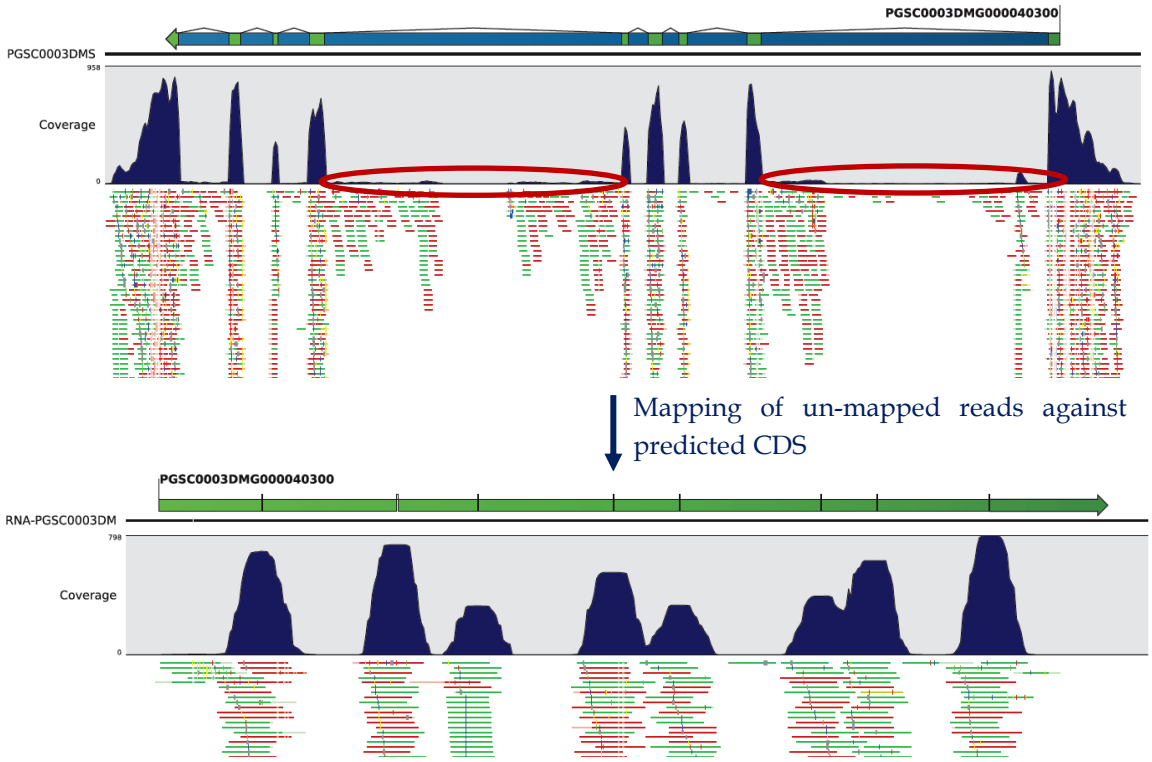


Figure 5-8 Example of gene model validation in the 3.0 potato genome annotation. The gene model (indicated in blue) PGSC0003DMG000040300 encoding an ELMO domain-containing protein is fully validated. **Top:** Initial reference assembly against the V 3.0 genome sequence. Reads map with high coverage at all exon regions (indicated in green), verifying the gene structure. **Bottom:** Subsequent mapping of un-mapped reads to the predicted mRNA transcript sequence. Reads map at the exon/exon boundaries verifying the gene structure. Coverage is shown below the gene model. Reads are indicated as follows: Green = Forward matching, red: reverse complement matching, yellow = randomly matching. Red circles indicate intron regions with coverage.

5.2.3.1 Methods

Initial reference assembly was performed as described in section 5.2.2.1, and coverage values were reported and tabulated for each position. Un-mapped reads were used as input for read mapping against the predicted coding sequences of the V3.0 genome annotation, and coverage values were reported and tabulated for each position. Using *ReadCoverage.pl* and *ExonExonBoundaryCoverage.pl t*he following statistics were calculated for each gene and individual CDS region: Mean and max coverage, fraction of CDS with read coverage and coverage of start and end position (based on reads mapped to the predicted mRNA transcripts). Detailed description on script usage and resulting files can be found in appendix B. Similarity searches of predicted protein and CDS sequences were performed. Protein sequences were matched against the UniRef100 database (release-2010_01)[27] using BlastP (Camacho *et al.*, 2009; Altschul *et al.*, 1990) (E-value <= 10). CDS sequences were matched against the PlantGDB-assembled Unique Transcript-fragment derived from *S. tuberosum* mRNAs (release 157a - based on GenBank release 157)[28] using BlastN (Camacho *et al.*, 2009; Altschul *et al.*, 1990) (E-value <= 10). In both cases, the best hit (if any) was tabulated and reported.

5.2.3.2 Results

Of the 40,842 gene models in the DM v.3.0 genome 9,076 (22 %) were not observed with a single read, leaving 31,766 (78 %) supported by at least one mRNAseq read, cf. Figure 5-9. 26,787 and 23,459 had mRNAseq support in 50 % and 90 % of the model, respectively. In total, 16,697 (52 % of gene models with mRNAseq support, 41 % of all gene models) could be fully validated having mRNAseq support in the entire predicted mRNA transcript region of the gene model and having all exon/exon boundaries confirmed.

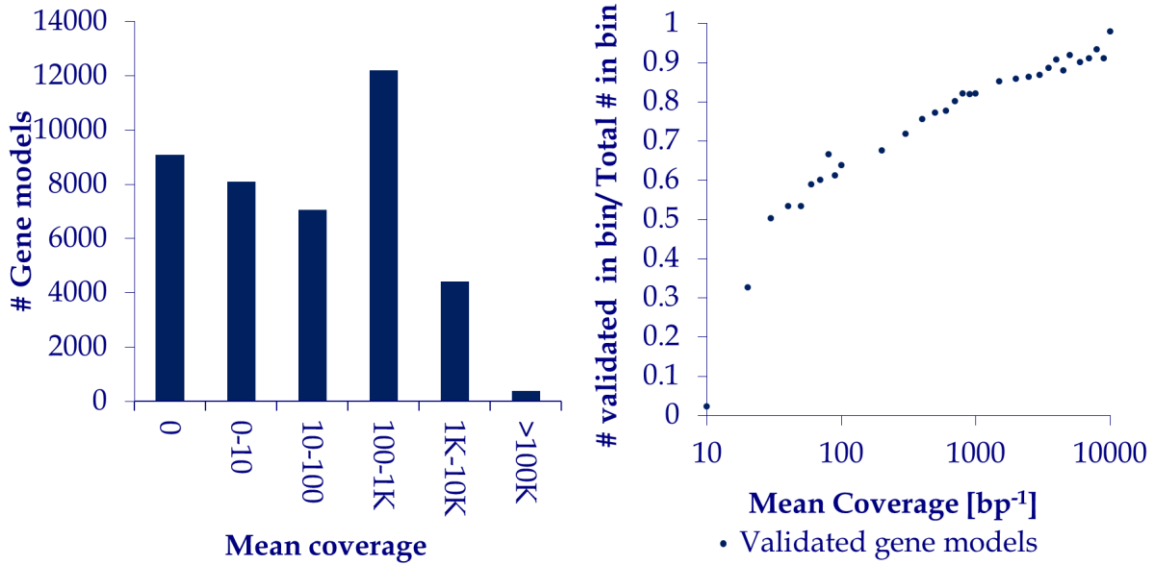


Figure 5-9 Gene model validation of the gene models in V3.0. **Left:** Distribution of gene models according to observed mean coverage. **Right:** Fraction of gene models fully validated. Gene models are binned according to mean coverage, and the fraction of validated models within each bin is shown.

[27] Available at: ftp://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2010_01/uniref/uniref2010_01.tar.gz
[28] Available at: http://www.plantgdb.org/download/Download/Sequence/ESTcontig/Solanum_tuberosum/previous_version/157a/Solanum_tuberosum.mRNA.PUT.fasta.bz2

There is a clear positive correlation between the mean coverage and the validation ratio of the gene models, cf. Figure 5-9. This indicates that the validation rate of gene models with low expression is under estimated.

To assess the quality of the 9,076 gene models with no mRNAseq support, these were compared to genes with mRNAseq support in more than 50 % of the predicted CDS (26,787 gene models), cf. Figure 5-10. Gene models with no mRNAseq support are less similar to known mRNA transcripts and protein sequences, cf. Figure 5-10 panels A and B. Moreover, the length of the protein coding sequence in the gene model is shorter in gene models with mRNAseq support.

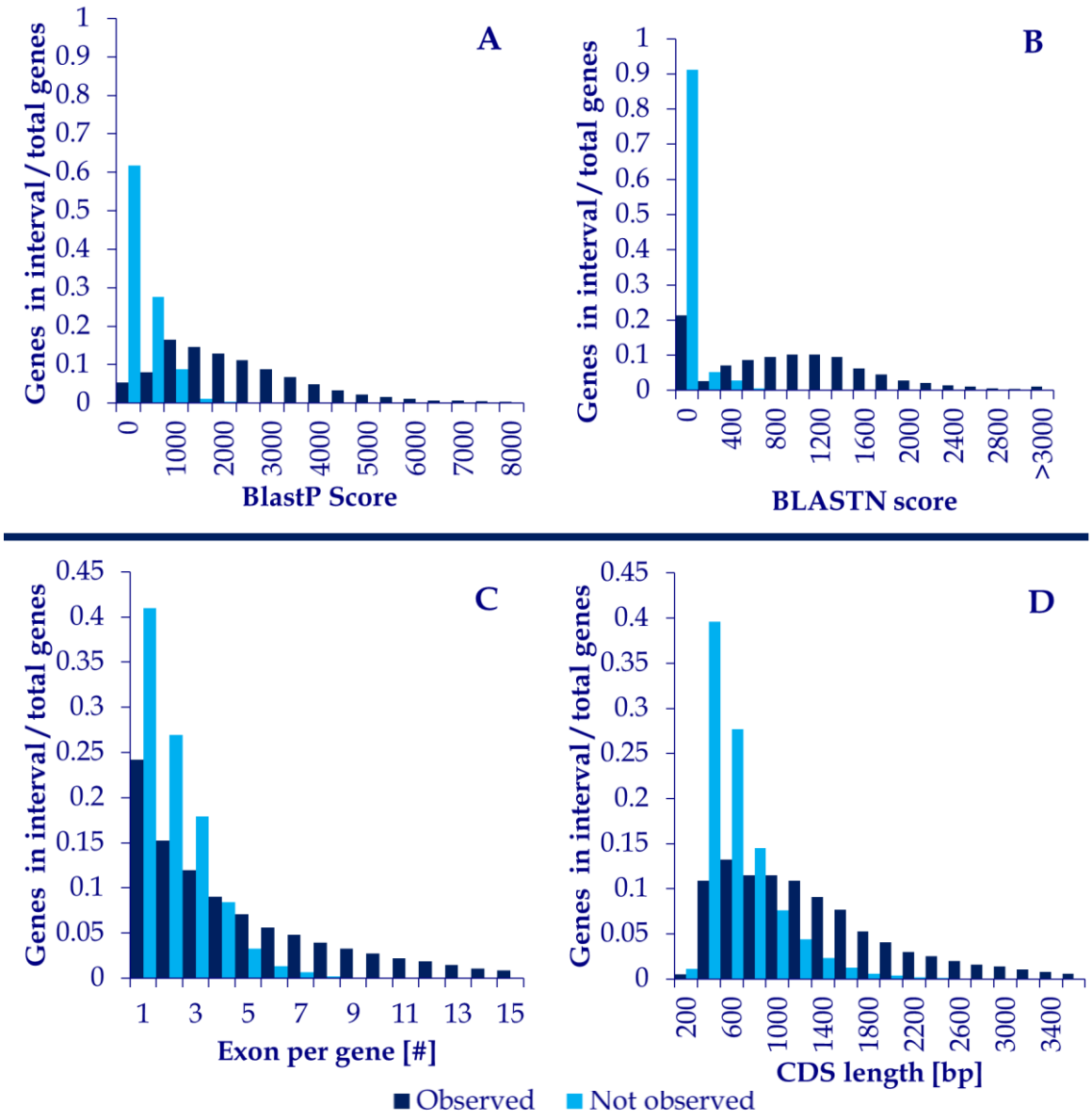


Figure 5-10 Comparison between version 3.0 gene models with (dark blue) and without (light blue) mRNAseq support. **A)** Predicted protein sequences similarity to proteins in the UniRef100 database. **B)** Predicted CDS sequences similarity to PlantGDB-assembled unique transcript-fragments. **C)** Number of exons per gene model. **D)** Total length of the coding sequence (CDS).

5.2.3.3 Discussion and Conclusions

For a large fraction of the gene models (78 %), it could be concluded that these were supported by mRNAseq reads and hence located in transcripted parts of the genome. However, based on the results, it was not possible to give an overall estimate of how many gene models were predicted with the correct exon/intron structure. This is caused by the correlation between read coverage and validation. The validation is based on a data set consisting 53 samples originating from different tissues and conditions (cf. Supplementary Table 4, pp. 35-40 and Supplementary text section 6, pp. 7-9 (The Potato Genome Sequencing Consortium *et al.*, 2011)). It contains more than 743 million mRNAseq reads (representing 39 billion bases), of which 92 % map to the genome. Although this data set covers a wide range of developmental stages and tissue types, it still does not capture the entire transcriptome of *S. tuberosum*, to a sufficient amount for complete gene validation. One reason is the dynamic range of the transcriptome, making gene validation of lowly expressed genes difficult. For a significant amount of the gene models (22 %), no mRNAseq support could be detected. This can in part be explained by genes, which are not expressed in the validation data set. However, the features of the gene models with no mRNAseq support differ from those with mRNAseq support, cf. Figure 5-10. If the gene models with no mRNAseq support in fact were correctly predicted functional genes, there would be no difference between these and gene models with mRNAseq support. However, the observed difference could indicate that a fraction of the gene models with no mRNAseq support, in fact are pseudogenes or mis-predicted genes. Few other genome-wide studies have characterized pseudogenes in plants (Zou *et al.*, 2009; Zou *et al.*, 2009; Benovoy & Drouin, 2006). The study by Zou *et al.*, in *Oryza sativa* (rice) and *A. thaliana* indicated that plant genomes can contain a substantial amount of pseudogenes (Zou *et al.*, 2009). Furthermore, they showed that a small fraction (2-5%) of the pseudogenes was expressed, however at lower levels, and that these pseudogenes were likely to be the product of a recent pseudogenization event (Zou *et al.*, 2009). These findings from *O. sativa* support the possibility that the gene models with no mRNAseq support could be pseudogenes, but it also shows that pseudogenization is a continuous process, hereby complicating the annotation process of these.

The experimental gene model validation of the V3.0 data set indicated that the majority of the gene models predicted using a combination of *ab initio* gene prediction, and protein and EST alignments were correct. However, this analysis also indicated that some gene models most likely were mis-predicted. Hence, improvements of the annotation was needed. This led to the incorporation of mRNAseq assisted gene prediction by Cufflinks in the V3.1 data set

5.2.4 Statistics of Gene Models in the Version 3.2 Annotation

With the incorporation of mRNAseq assisted gene prediction by Cufflinks (Trapnell *et al.*, 2011) in the V3.1 data set new features in the annotation was introduced. Firstly, gene models predicted by Cufflinks had UTR annotations; secondly, some gene models predicted by Cufflinks had multiple transcripts, and thirdly some mRNA transcripts predicted using Cufflinks were annotated as non-coding, due to non-existing or extremely short CDS annotations. To investigate the overall quality of this annotation update, overall statistics, which

will be presented in the current section, were calculated. During work presented here, multiple errors in the annotation were discovered, why an update of the annotation was made, resulting in version 3.2. Most of the analyses presented in published article were based on this version, and later updated to match the final version.

5.2.4.1 Methods

Gene annotation files in GFF format of *A. thaliana* (TAIR9)[29] and *V. vinifera* (IGGP_12x)[30] were downloaded and compared to the potato genome annotation DM version 3.2. Descriptive statistics such as mRNA, CDS, exon and UTR lengths, and number exons per transcripts were calculated for each mRNA transcripts and subsequently reported and tabulated using *StatsFromGFF.pl*.

5.2.4.2 Results

The DM V3.2 gene annotation contains slightly more gene models than the DM V3.0 gene annotation, cf. Table 5-5. However, of the 28,792 gene models predicted using Cufflinks 6,510 were novel. Of these 2,320 encoded a protein, while 4,190 only encoded non-protein mRNA transcripts (ncRNAs). 17,762 gene models in V3.0 could be verified by Cufflinks prediction and were replaced, cf. Table 5-5. 10,043 gene models were predicted both by GLEAN and Cufflinks, but with alternate mRNA transcripts. 13,037 GLEAN predicted gene models did not overlap with a Cufflinks predicted gene model in the V3.2 gene annotation. The amount of 5' and 3' UTRs containing intron regions (15 % and 7 %, respectively), is in agreement with the findings of Pesole *et al.*, who found intron regions in 15 % of 5' UTR regions and 4 % of 3'UTR regions in the taxonomic collection "other *viridiplantae*" (Pesole *et al.*, 2001). 42 % of all gene models had more than 1 mRNA transcript predicted, cf. Figure 5-11 panel A. When comparing number of exons per gene, and the lengths of mRNA transcripts, CDS sequences, 5'UTRs and 3'UTRs between *A. thaliana*, *V. vinifera* and *S. tuberosum*, these were generally similar, cf. Figure 5-11. However, there were differences. Relative to the total number of genes, *S. tuberosum* and *A. thaliana* have nearly twice the amount of single exon genes than *V. vinifera*; cf. Figure 5-11 panel B.

[29] Available at : ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR9_genome_release/TAIR9_gff3/TAIR9_GFF3_genes.gff
[30] Available at: http://www.genoscope.cns.fr/externe/Download/Projets/Projet_ML/data/12X/annotation/Vitis_vinifera_annotation.gff.gz

Table 5-5 Descriptive statistics for the DM V3.2 gene annotation. *Statistics only calculated for Cufflinks predicted protein coding transcripts. **Statistics for UTRs only calculated for Cufflinks predicted transcripts. NcRNA = Non-protein coding transcript.

Annotation	Type	Number
Gene		
	Total predicted gene models	47,352 (100 %)
	Genes only predicted by Cufflinks	24,272 (51 %)
	Genes only predicted by GLEAN	13,037 (28 %)
	Genes predicted by both GLEAN & Cufflinks	10,043 (21 %)
	Genes containing protein coding mRNA transcripts	43,162 (92 %)
	Genes predicted by Cufflinks only containing ncRNA transcripts	4,190 (8 %)
	Mean number of predicted mRNA transcripts per gene model	2.1
mRNA Transcripts		
	Total predicted mRNA transcripts	97,781 (100 %)
	Protein coding mRNA transcripts	69,456 (71 %)
	NcRNA transcripts	28,325 (29 %)
	Single exon mRNA transcripts	37,980 (39 %)
	Mean length [bp]	1259
	Median length [bp]	957
	Mean mapping depth [reads/bp]*	235
	Median mapping depth [reads/bp]*	26.6
Exons		
	Mean exon length [bp]	369.5
	Median exon length [bp]	189
	Mean # exons per transcript	3.9
UTRs*		
	Mean 5’ UTR length [bp]	234
	Mean 3’ UTR length [bp]	403
	5'UTRs with introns	3,905 (15%)
	3'UTRs with introns	1,901 (7 %)

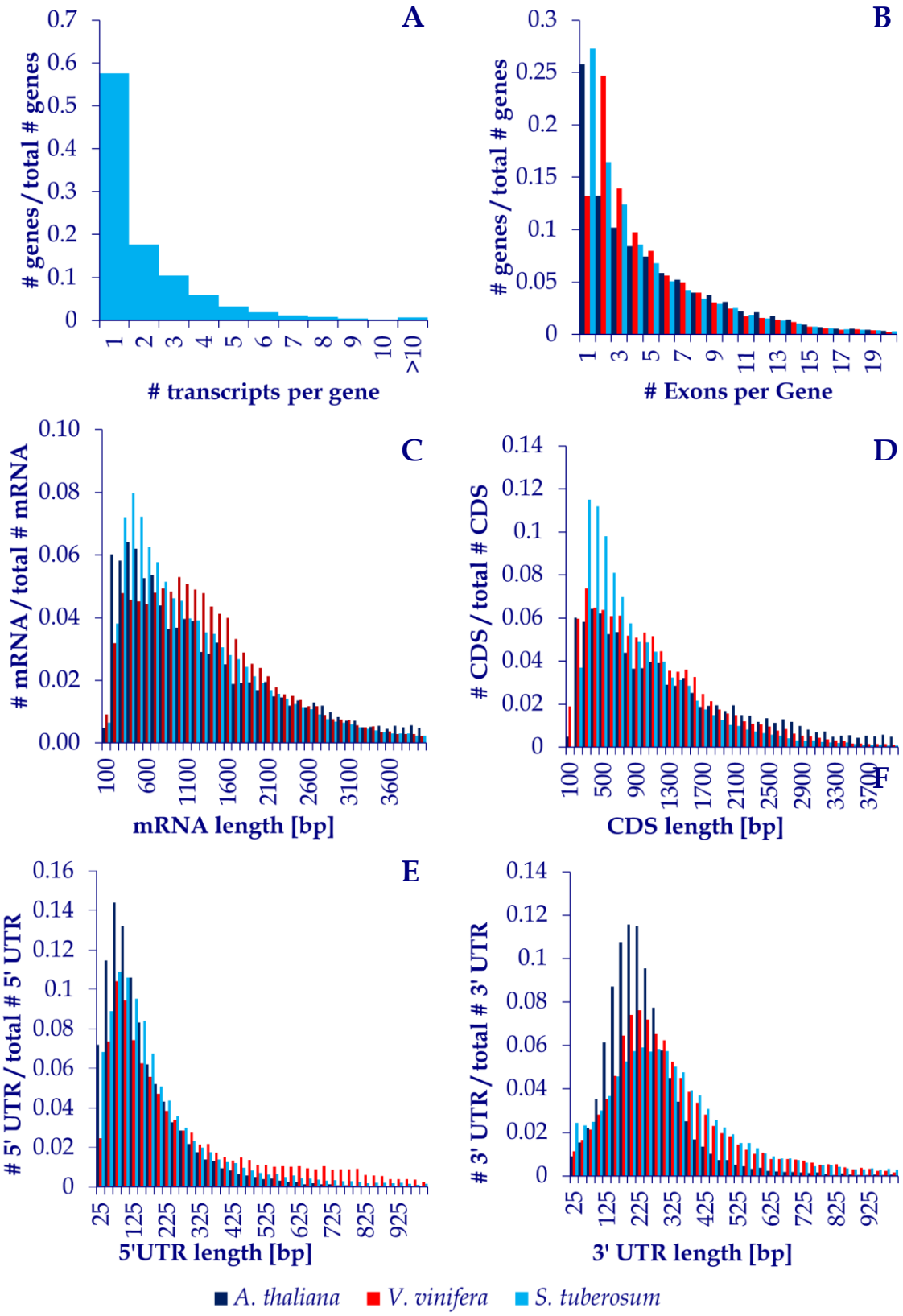


Figure 5-11 Descriptive statistics for the DM V3.2 gene annotation compared to gene annotations from *A. thaliana* and *V. Vinifera*. **A)** Number of transcripts per gene for gene models predicted using Cufflinks. **B)** Number of exons per gene. **C)** Length of predicted mRNA transcripts. **D)** Length of predicted CDS sequences. **E)** Length of predicted 5' UTRs. **F)** Length of predicted 3' UTRs. Only transcripts with predicted UTRs are shown in panels E and F.

V. vinifera clearly has a higher number of longer mRNA transcripts, cf. Figure 5-11 panel C, and also a higher number of longer CDS sequences although not to the same extent, cf. Figure 5-11 panel D. The distributions of UTR lengths are very similar for *V. vinifera* and *S. tuberosum*, both having longer UTRs compared to *A. thaliana*.

Some mRNA transcripts predicted using Cufflinks were annotated as non-coding, due to non-existing or extremely short CDS annotations. Descriptive statistics for these were compared to protein coding mRNA transcripts in the DM V3.2 annotation, cf. Figure 5-12.

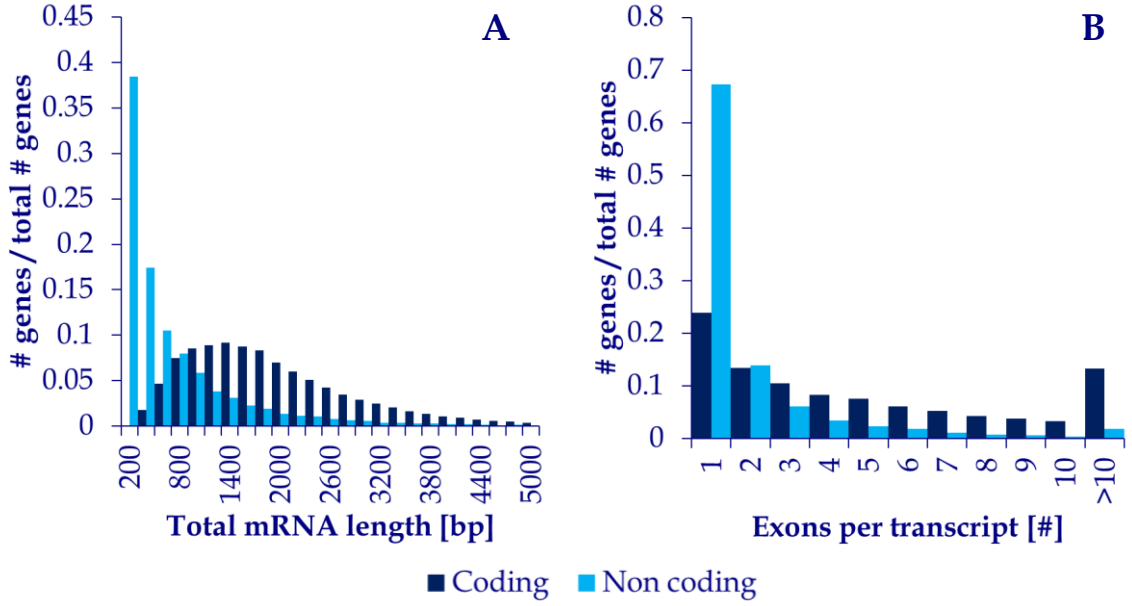


Figure 5-12 Descriptive statistics for protein coding and non-protein coding mRNA transcripts predicted by Cufflinks in the DM V3.2 gene annotation. **A)** Total length of mRNA transcript. **B)** Number of exons per mRNA transcript.

Non-protein coding mRNA transcripts were generally found to be shorter and having fewer exons than protein coding mRNA transcripts. Moreover, there were 2.8 times more single exon ncRNAs than protein coding transcripts. Of the 28,325 predicted ncRNAs most (83 %) were located in regions also encoding a protein coding transcript. This indicates that these are predictions of mis-spliced mRNAs or short alternative splicing variants not giving rise to a functional protein sequence. Of the last 4,782 ncRNAs originating from genes only encoding ncRNAs 3,420 (71 %) had a significant hit to a Uniref100 protein (BLASTX (Camacho *et al.*, 2009; Altschul *et al.*, 1990) similarity search (E-value cutoff $1 \cdot 10^{-5}$) data not shown). This could indicate that these in fact are protein coding, but errors (e.g. in the exon structure) in the annotation have caused the gene to be called as non-coding.

5.2.4.3 Discussion

The overall descriptive statistics of the V3.2 annotation were similar to that of other gene annotation from plant genomes. This provided evidence of a high quality annotation. However, several issues were discovered. The incorporation of Cufflinks predicted transcripts into the V3.2 gene annotation gave both advantages and disadvantages. Compared the combined *ab initio* sequence alignment based method used in the V3.0 annotation, the Cufflinks prediction gave more direct evidence of transcription (and hence indication of the existence of a gene). Additionally, the Cufflinks prediction enabled prediction of 6,510 novel gene

models compared to the V3.0 gene annotation. However, along with these advantages came some disadvantages. Although Cufflinks provides maximum precision only reporting the minimal number of compatible isoforms (Trapnell *et al.*, 2010), there is no filtering of transcripts most likely to be the product of mis-splicing or premature mRNAs. This result in a substantial amount of reads matching the intron regions of highly expressed genes, cf. Figure 5-8. In some cases, mis-splicing can be recognized if some intron regions have coverage in the entire region, indicating that the region was not spliced, and others have no or very few reads matching, indicating a near perfect precision of splicing these regions out. 29 % of all Cufflinks predicted transcripts were ncRNAs, and most of these were located in gene regions also encoding protein coding mRNAs. Several studies have shown the importance of ncRNAs e.g. in respect to regulation of gene expression, and that most of these ncRNAs are located in near proximity or even in the intron region of the target gene (Au *et al.*, 2011; Carra *et al.*, 2011; van Bakel *et al.*, 2010). Hence, some of the ncRNAs predicted in the potato genome are most likely correctly annotated. However, it is extremely difficult to distinguish these from the mentioned artifacts of mis-spliced or premature mRNAs, one reason being that some ncRNAs are polyA negative (Wang *et al.*, 2011; Wu *et al.*, 2008), why evidence of polyadenylation cannot be used to distinguish mis-spliced or premature mRNAs from true ncRNA transcripts. Such distinction would properly require manual annotation on a gene by gene basis. Due to this, it was chosen to eliminate ncRNAs in later versions of the annotation. Moreover, it was chosen to create a representative transcript set where the mRNA encoding the longest protein sequence was chosen for each gene. The ncRNAs could also be the product of mis-predicted genes where the ncRNA in fact is a part of a larger mRNA transcript, but the evidence of linkage between exons has been insufficient. An extreme case of this is given later in Figure 5-13 (see section 5.2.5). Although correction of these cases also would have to be done manually, the process could be facilitated by flagging gene models containing ncRNAs not being a part of the representative mRNA transcript. This would facilitate both correction of gene models (by including extra exon regions into the mRNA transcript), and discovery of true ncRNA transcripts. Furthermore, some neighboring genes only encoding ncRNAs, had significant matches to the same or similar Uniref protein sequence. This can indicate that these potentially are parts of the same protein coding, which have been mis-predicted. This could be used to flag genes for manual curation.

The V3.2 data set became the set, which most of the analyses presented in the published article were based on. Later versions of the annotation were updates after different filterings were performed. For example, the V3.2 annotation contained a substantial amount of transposon related genes, which were eliminated in later versions. Moreover, genes of which length was smaller than 300bp were excluded, cf. Table 5-3. These filterings were performed on a global scale, and it is in the author's conviction, that further improvements of the gene annotation would require manual interference on a gene by gene basis. The overall quality of the gene annotation was evaluated based on descriptive statistics. To acquire a more detailed evaluation of the quality of the gene annotation, it was chosen to perform a manual curation of a small subset of genes. The subset chosen for this was genes involved in starch metabolism, cf. section 5.2.5. Therefore, the analysis could be combined with a detailed analysis of the gene expression of these genes described in section 5.2.7.

5.2.5 Manual Gene Model Curation of Starch Metabolism Genes

As mentioned, a small subset of genes was chosen to acquire a more detailed evaluation of the quality of the gene annotation. It was chosen to perform manual curation of genes involved in starch metabolism. The curation was based on a read mapping of all DM and RH RNAseq libraries and was performed by visual inspection of the gene models of interest.

5.2.5.1 Methods

Genes encoding proteins involved in starch metabolism was initially identified by a BLASTP (Camacho *et al.*, 2009; Altschul *et al.*, 1990) similarity search of protein sequences known to be involved in starch metabolism[31] against DM V3.2 peptide sequences (E-value =< $1 \cdot 10^{-20}$). In total, 54 sequences encoding proteins from 23 different enzymatic reactions were used for the searched, cf. Table 5-6.

Table 5-6 Protein sequences known to be involved in starch metabolism used for identification of starch metabolism genes in the DM V3.2 gene annotation.

E.C.	description	GenBank Accessions
2.4.1.1	Phosphorylase	CAA43490, CAA36612, AAA33809
2.4.1.13	Sucrose synthase	AAA63451, AAA63452, AAA33841
2.4.1.14	Sucrose-phosphate synthase	BAA00570
2.4.1.18	1,4-alpha-glucan branching enzyme	CAA78283, CAA52036
2.4.1.21	Starch synthase	CAA49463, CAA80358, CAA52917, CAA53741, CAA54265, AAA50305
2.4.1.25	4-alpha-glucanotransferase	AAA91883
2.7.1.1	Hexokinase	AAA91884, CAA71442, CAA64173
2.7.1.4	Fructokinase	CAA63966
2.7.1.90	Diphosphate--fructose-6-phosphate 1-phosphotransferase	AAC26113, CAB40746
2.7.7.27	Glucose-1-phosphate adenylyltransferase	AAD25541, AAF14186, CAB76673, CAB76674
2.7.7.9	UTP--glucose-1-phosphate uridylyltransferase	CAB93680, CAB93681
2.7.9.4	Alpha-glucan, water dikinase	AAK84008, AAL55635
3.1.3.11	Fructose-bisphosphatase	AAN15317, AAN15318, AAN15319
3.1.3.24	Sucrose-phosphate phosphatase	AAO34668, AAO67719
3.2.1.1	Alpha-amylase	AAQ17074, AAR99599
3.2.1.2	Beta-amylase	CAA61241
3.2.1.26	Beta-fructofuranosidase	AAU00726, ABA40442, ABB29926, ABB99399, ABC01905
3.2.1.68	Isoamylase	ABS52706, ABS52707, ABY58016
4.1.2.13	Fructose-bisphosphate aldolase	ACC93586, ABY89288, ACD13788
5.3.1.9	Glucose-6-phosphate isomerase	ACD50895
5.4.2.2	Phosphoglucomutase	ACT09058, ACZ66259
-	ADP/ATP translocator	NP_001234018,NP_001105434
-	Glucose-6-phosphate translocator	AAC08526,AAO19451

Moreover, an internal PGSC functional annotation database[32] was used for a keyword search to identify genes with functional annotations involved in starch metabolism (e.g. "Starch

[31] Initial list of protein sequences was compiled by Bjorn Klosterman at Wageningen University, the Netherlands.
[32] Available at: http://www.compbio.dundee.ac.uk/geneweb/search. Registration needed

synthase" and "Alpha-amylase"). Hereafter, all DM V3.2 proteins belonging to the same putative orthologous group[33] as a protein already identified were included. Finally, identified starch metabolism protein sequences in the DM V3.2 gene annotation were used in a second BLASTP similarity search against all DM V3.2 protein sequences. Proteins with a significant match (E-value =< $1 \cdot 10^{-20}$) were included.

All 32 DM and 16 RH RNAseq libraries (NCBI Sequence Read Archive (SRA030516; study SRP005965[34]) and the European Nucleotide Database ArrayExpress Database (E-MTAB-552; study ERP000527[35]), respectively) were mapped to the DM V3.0 genome sequence using the CLC Genomics workbench 3.7.1. Reference assembly was perform using un-gapped alignment (end trimming allowed), and random match mode, otherwise default settings. Identified starch metabolism gene regions were extracted. Manual curation of the gene models was performed by visual inspection, confirming or correcting all start and stop sites and all intron/exon boundaries. Each gene model was considered to be correct, if the start and stop sites and the intron/exon structure could be validated for 1 transcript model.

[33] Putative orthologous groups of 12 plant species were identified using OrthoMCL with default parameters (Li, Stoeckert Jr. & Roos, 2003). Species included were: *Arabidopsis thaliana, Brachypodium distachyon, Carica papaya, Chlamydomonas reinhardtii, Glycine max, Oryza sativa, Physcomitrella patens, Populus trichocarpa, Solanum tuberosum, Sorghum bicolor, Vitis vinifera* and *Zea mays*. Identification was performed by Brett Whitty at Michigan State University. A prediction based on the V3.4 version is available at:
http://potatogenomics.plantbiology.msu.edu/data/12_plants_all_orthomcl_parsed.txt.zip.
[34] Available at: http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP005965
[35] Available at: http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-552

5.2.5.2 Results

In total, 167 gene models were identified as starch metabolism genes, cf. Table 5-7. On average, 6 genes were identified to encode proteins for each enzymatic reaction, ranging between 1 and 23 proteins, cf. Table 5-7. Results from the manual gene curation are summarized in Table 5-8.

Table 5-7 Identified starch metabolism gene models. The number of proteins for each enzymatic reaction is given in parenthesis after the Enzyme Commission (EC) classification

EC Number / Gene ID	# Transcripts	Comments
2.4.1.1 Phosphorylase (6)		
PGSC0003DMG200002479	2	Correct
PGSC0003DMG200003495	1	Gene split, joined with PGSC0003DMG200007782
PGSC0003DMG200007782	1	Gene split, joined with PGSC0003DMG200003495
PGSC0003DMG200009711	2	Correct
PGSC0003DMG200028382	4	Correct
PGSC0003DMG200031765	3	Correct
2.4.1.13 Sucrose synthase (6)		
PGSC0003DMG200002895	2	Correct
PGSC0003DMG200006672	5	Correct
PGSC0003DMG200013546	4	Correct
PGSC0003DMG200013547	1	Correct
PGSC0003DMG200016730	2	Correct
PGSC0003DMG200031046	1	Exon/Intron structure manually curated
2.4.1.14 Sucrose-phosphate synthase (5)		
PGSC0003DMG200019060	5	Exon/Intron structure manually curated
PGSC0003DMG200026428	1	Correct
PGSC0003DMG200027936	1	Correct
PGSC0003DMG200029891	1	Gene split, joined with PGSC0003DMG200029892
PGSC0003DMG200029892	1	Gene split, joined withPGSC0003DMG200029891
2.4.1.18 1,4-alpha-glucan branching enzyme (4)		
PGSC0003DMG200002510	3	Correct
PGSC0003DMG200002712	1	Correct
PGSC0003DMG200009981	4	Correct
PGSC0003DMG200022307	3	Genome sequence error, manually curated
2.4.1.21 Starch synthase (7)		
PGSC0003DMG200001328	2	Correct
PGSC0003DMG200008322	1	Exon/Intron structure manually curated
PGSC0003DMG200012111	2	Correct
PGSC0003DMG200013540	5	Exon/Intron structure manually curated
PGSC0003DMG200016481	2	Correct
PGSC0003DMG200018552	3	Correct
PGSC0003DMG200030619	2	Exon/Intron structure manually curated
2.4.1.25 4-alpha-glucanotransferase (2)		
PGSC0003DMG200002195	4	Correct
PGSC0003DMG200016589	3	Correct
2.7.1.1 Hexokinase (6)		
PGSC0003DMG200000295	2	Exon/Intron structure manually curated

PGSC0003DMG200002525	1	Correct
PGSC0003DMG200009861	3	Correct
PGSC0003DMG200013187	1	Correct
PGSC0003DMG200016521	4	Correct
PGSC0003DMG200030624	3	Correct
2.7.1.116-phosphofructokinase (8)		
PGSC0003DMG200010749	2	Correct
PGSC0003DMG200016208	2	Correct
PGSC0003DMG200017413	1	Correct
PGSC0003DMG200019734	1	Correct
PGSC0003DMG200023631	2	Gene fusion, manually curated
PGSC0003DMG200025455	5	Exon/Intron structure manually curated
PGSC0003DMG200027554	1	Not curated due to low coverage
PGSC0003DMG200029304	1	Correct
2.7.1.4 Fructokinase (7)		
PGSC0003DMG200010277	2	Correct
PGSC0003DMG200020361	1	Correct
PGSC0003DMG200024246	4	Correct
PGSC0003DMG200026916	1	Correct
PGSC0003DMG200027017	2	Correct
PGSC0003DMG200028311	2	Correct
PGSC0003DMG200030653	2	Correct
2.7.1.90 Diphosphate--fructose-6-phosphate 1-phosphotransferase (4)		
PGSC0003DMG200000669	2	Exon/Intron structure manually curated
PGSC0003DMG200010007	3	Correct
PGSC0003DMG200016726	1	Correct
PGSC0003DMG200029309	1	Correct
2.7.7.27 Glucose-1-phosphate adenylyltransferase (6)		
PGSC0003DMG200000735	2	Exon/Intron structure manually curated
PGSC0003DMG200009026	2	Correct
PGSC0003DMG200015952	3	Correct
PGSC0003DMG200025218	1	Not curated due to low coverage
PGSC0003DMG200031084	1	Correct
PGSC0003DMG200046891	1	Correct
2.7.7.9 UTP--glucose-1-phosphate uridylyltransferase (4)		
PGSC0003DMG200008445	2	Correct
PGSC0003DMG200013333	5	Correct
PGSC0003DMG200030031	4	Correct
PGSC0003DMG200031123	5	Correct
2.7.9.4 Alpha-glucan, water dikinase (3)		
PGSC0003DMG200007677	2	Correct
PGSC0003DMG200008503	2	Correct
PGSC0003DMG200016613	3	Correct
3.1.3.11 Fructose-bisphosphatase (6)		
PGSC0003DMG200010788	2	Correct
PGSC0003DMG200019188	1	Correct

PGSC0003DMG200019189	1	Correct
PGSC0003DMG200020363	1	Correct
PGSC0003DMG200024109	5	Correct
PGSC0003DMG200030370	3	Correct
3.1.3.24 Sucrose-phosphate phosphatase (3)		
PGSC0003DMG200033584	1	Correct
PGSC0003DMG200021341	3	Correct
PGSC0003DMG200028134	3	Correct
3.2.1.1 Alpha-amylase (5)		
PGSC0003DMG200007974	2	Correct
PGSC0003DMG200009891	4	Correct
PGSC0003DMG200017626	6	Gene fusion, manually curated
PGSC0003DMG200020603	5	Correct
PGSC0003DMG200025153	3	Correct
3.2.1.2 Beta-amylase (8)		
PGSC0003DMG200000169	2	Correct
PGSC0003DMG200001549	1	Correct
PGSC0003DMG200001855	1	Correct
PGSC0003DMG200010664	2	Correct
PGSC0003DMG200012129	2	Correct
PGSC0003DMG200020509	2	Gene fusion, manually curated
PGSC0003DMG200024145	1	Exon/Intron structure manually curated
PGSC0003DMG200026199	4	Correct
3.2.1.26 Beta-fructofuranosidase (23)		
PGSC0003DMG200001596	1	Exon/Intron structure manually curated
PGSC0003DMG200002583	1	Correct
PGSC0003DMG200002756	2	Correct
PGSC0003DMG200004463	2	Exon/Intron structure manually curated
PGSC0003DMG200004790	1	Correct
PGSC0003DMG200008388	1	Correct
PGSC0003DMG200008942	1	Gene fusion, manually curated
PGSC0003DMG200008943	2	Gene fusion, manually curated
PGSC0003DMG200008974	1	Not curated due to low coverage
PGSC0003DMG200009257	3	Correct
PGSC0003DMG200009936	1	Correct
PGSC0003DMG200011037	1	Correct
PGSC0003DMG200013088	4	Correct
PGSC0003DMG200013856	1	Correct
PGSC0003DMG200019494	4	Correct
PGSC0003DMG200022270	2	Correct
PGSC0003DMG200026107	2	Correct
PGSC0003DMG200026530	3	Correct
PGSC0003DMG200027925	1	Correct
PGSC0003DMG200028252	3	Gene fusion, manually curated
PGSC0003DMG200033142	3	Exon/Intron structure manually curated
PGSC0003DMG200042880	1	Correct
PGSC0003DMG200046915	1	Not curated due to low coverage

3.2.1.41 Pullulanase (1)		
PGSC0003DMG200031073	2	Correct
3.2.1.68 Isoamylase (5)		
PGSC0003DMG200000954	1	Correct
PGSC0003DMG200007274	3	Exon/Intron structure manually curated
PGSC0003DMG200017932	1	Not curated, partial model
PGSC0003DMG200020699	6	Correct
PGSC0003DMG200030253	2	High degree of random matching, not curated
3.6.1.1 Inorganic diphosphatase (16)		
PGSC0003DMG200002126	1	Not curated due to low coverage
PGSC0003DMG200002775	1	Correct
PGSC0003DMG200003103	2	Correct
PGSC0003DMG200003514	2	Exon/Intron structure manually curated
PGSC0003DMG200004999	1	Correct
PGSC0003DMG200005858	1	Exon/Intron structure manually curated
PGSC0003DMG200007913	1	Correct
PGSC0003DMG200008932	3	Correct
PGSC0003DMG200012223	3	Correct
PGSC0003DMG200014208	1	Correct
PGSC0003DMG200017725	1	Not curated due to low coverage
PGSC0003DMG200025085	2	Correct
PGSC0003DMG200026784	1	Correct
PGSC0003DMG200028529	2	Correct
PGSC0003DMG200030682	1	Correct
4.1.2.13 Fructose-bisphosphate aldolase (9)		
PGSC0003DMG200034489	1	Not curated due to low coverage
PGSC0003DMG200002675	2	Correct
PGSC0003DMG200003123	2	Correct
PGSC0003DMG200003548	2	Correct
PGSC0003DMG200012012	1	Correct
PGSC0003DMG200022263	3	Gene fusion, manually curated
PGSC0003DMG200026665	6	Correct
PGSC0003DMG200026666	5	Correct
PGSC0003DMG200028261	2	Correct
PGSC0003DMG200030565	2	Correct
5.3.1.9 Glucose-6-phosphate isomerase (4)		
PGSC0003DMG200009848	1	High degree of random matching, not curated
PGSC0003DMG200012910	2	Correct
PGSC0003DMG200015341	2	Exon/Intron structure manually curated
PGSC0003DMG200030128	1	High degree of random matching, not curated
5.4.2.2 Phosphoglucomutase (5)		
PGSC0003DMG200001912	3	Correct
PGSC0003DMG200009842	1	Correct
PGSC0003DMG200015902	1	Correct
PGSC0003DMG200017367	3	High degree of random matching, not curated
PGSC0003DMG200024250	2	Correct

ADP/ATP translocator (8)		
PGSC0003DMG200004065	2	Correct
PGSC0003DMG200006806	1	Correct
PGSC0003DMG200008225	2	Exon/Intron structure manually curated
PGSC0003DMG200011790	1	Correct
PGSC0003DMG200013596	1	Correct
PGSC0003DMG200021860	1	Not curated due to low coverage
PGSC0003DMG200031867	1	Correct
PGSC0003DMG200032824	1	Correct
Glucose-6-phosphate translocator (6)		
PGSC0003DMG200001041	4	Correct
PGSC0003DMG200005269	1	Correct
PGSC0003DMG200005602	1	Correct
PGSC0003DMG200012710	2	Correct
PGSC0003DMG200025495	4	Correct
PGSC0003DMG200044320	1	Exon/Intron structure manually curated

Each gene model was considered to be correct, if the start and stop sites and the intron/exon structure could be validated for at least 1 transcript model. This was the case for 123 gene models (74 %). In all cases except for 5, the transcript model encoding the longest protein sequence was correct. 31 gene models (17 %) could be manually corrected based on the visual inspection of the read mappings. An example of such correction is given in Figure 5-13.

Table 5-8 Result summery of manual curation of 167 identified starch metabolism gene models in the V3.2 gene annotation.

Result of manual validation	# Gene models
Gene models with correct transcript model	123 (74 %)
Gene models where the transcript having the longest CDS is correct	118 (70 %)
Manually corrected Gene models	31 (18 %)
Annotated gene model contain more than 1 gene (gene fusion)	7 (7.0 %)
Correct gene model is annotated as multiple genes (split gene)	4 (2.3 %)
Gene models not validated	13 (8 %)
Total identified starch metabolism gene models	167 (100 %)

Corrections could often be made to the intron/exon structure, either by combining predicted transcripts or by changing exon annotations a few nucleotides. In nearly every case, this led to a longer predicted CDS sequence. BLASTP similarity searches showed that these longer CDS sequences were more similar to known *S. tuberosum* protein sequences or other plant orthologs (data not shown). This verified that the longer CDS sequences were not modular proteins, but likely to be correctly curated. Other frequent errors found were gene fusions (2 genes annotated as 1) found in 7 cases and gene splits (1 gene annotated as 2) found in 4 cases. 13 gene models (8 %) could not be validated, either due to low coverage or a high degree of random matching.

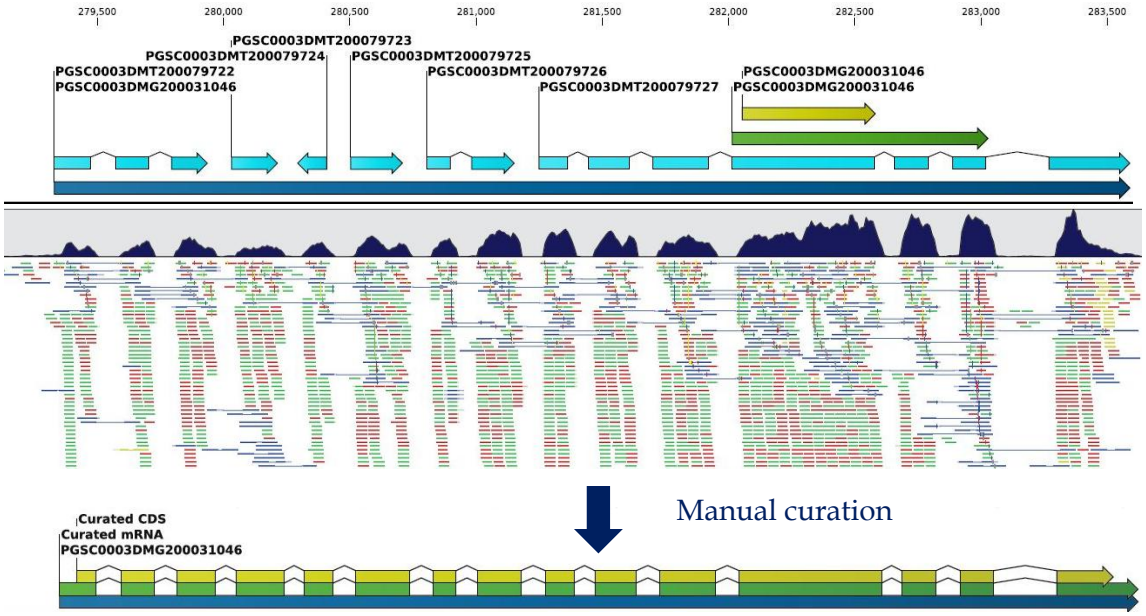


Figure 5-13 Manual curation of the gene model PGSC0003DMG200013046. The original gene model (marked in dark blue) contains 6 ncRNAs (marked in light blue) and a single coding mRNA transcript (marked in green) with a protein coding sequence (marked in yellow), which only covers a small part of the transcribed part of the gene. Manual curation enabled prediction of a single protein coding mRNA transcript covering the entire gene model with a coding sequence covering most of it.

5.2.5.3 Discussion and Conclusions

The manual curation of the starch metabolism gene models confirmed that most gene models containing multiple mRNA transcripts most likely were the product of mis-splicing or premature mRNAs. In most cases, a single transcript variant could account for the mapping of mRNAseq reads to the gene model with coverage far greater than other transcript variants predicted. As expected due to polyA positive mRNA enrichment, no evidence was found for the presence of ncRNAs in close proximity or in the intron regions of any of the gene models investigated. However, this does not conclude that any of the starch metabolism genes are regulated by ncRNAs such as miRNAs and further analyses of miRNA could elucidate this. In nearly every case (96 %), the mRNA transcript variant encoding the longest protein sequence had the most mRNAseq support, and hence was most likely to be the functional transcript of the gene. This later provided an argument when the non-redundant set of representative mRNA transcripts was made for the V3.4 gene annotation, because this only contained the mRNA transcript encoding the longest protein sequence for each gene. Several gene fusions and gene splits were discovered in the starch metabolism gene set. This discovery led to further analyses by the author to improve the gene annotation in regards to these errors. Firstly, annotated genes containing annotations of non-overlapping CDS sequences were identified, and these gene models were split in the V3.4 gene annotation. No further efforts were made to correct gene splits. However, a method to detect possible gene splits would be to investigate protein similarity searches of neighboring gene models. Cases where the CDS sequence of two neighboring matches e.g. the first and last part of the same or similar proteins could be flagged for further investigation. This method would however still require manual curation. In 11 % of the investigated gene models corrections were made to the mRNA transcript annotation, resulting in a longer CDS sequence and similarity searches indicated that the corrections made did not result in modular protein sequences. The short-

ened CDS sequences were often caused by minor errors in the intron/exon structure of the predicted mRNA transcript, but these have major effect on the predicted CDS e.g. due to errors causing frame shifts. The correction of these would require manual curation, but a semi-automated method could be used to detect genes potentially having errors and at the same time flag gene models that are likely to have a correctly annotated mRNA transcript. Using a BLASTP (Camacho *et al.*, 2009; Altschul *et al.*, 1990) output, ratios between the alignment length, the length of the query protein (the predicted CDS), and the length of subject protein (the best hit in the database) could be calculated. Ratios ≈ 1 would indicate a correctly annotated CDS, whereas ratios far from 1 would indicate gene model potentially having errors in the predicted mRNA transcript, cf. Figure 5-15.

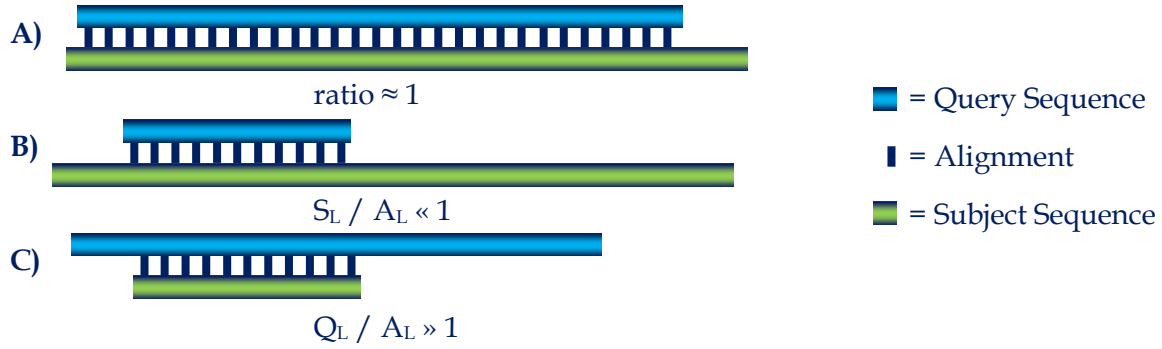


Figure 5-14 Outline of scheme to detect gene models potentially containing errors in the predicted mRNA transcript affecting the length of the CDS. **A)** Both the ratio between the alignment length and length of the predicted CDS (the query sequence) and the ratio between the alignment length and the length of the matched protein in the database (the subject sequence) are ≈ 1, indicating that the predicted mRNA transcript is likely to be correct. **B)** The ratio between the alignment length and the length of the matched protein in the database is « 1, indicating that the predicted mRNA transcript contains exon/intron boundary errors, which shortens the length of the CDS. **C)** The ratio between the alignment length and length of the predicted CDS is » 1 indicating that the CDS is a modular protein. S_L = length of subject sequence. A_L = alignment length. QL = length of query sequence.

Although the subset only accounts for 0.3 % of the entire gene set of the V3.2 annotation, and it therefore is impossible to give statistically significant conclusions regarding the quality of the entire annotations, the annotation quality of the starch metabolism genes provides some evidence of the quality of the entire gene set. As described later in section 5.2.7, the starch metabolism genes have expression levels that differs several orders of magnitude. Therefore, they are a good representation of the entire gene set in regards to expression level. Since the gene prediction using Cufflinks is dependent on sequence coverage (and hence expression level) it is likely that the annotation quality of the starch metabolism gene set does reflect the entire V3.2 gene annotation. 74 % of the starch metabolism gene models had a correctly annotated mRNA transcript, and it is the author's belief that this is a good estimate for the number of correct gene models in the entire gene set. 18 % of the starch metabolism gene models could be corrected by manual curation. However this would be very time costly even with the above suggestions for flagging potentially error containing gene models. It is the author's belief that the final version of the gene annotation where additional filtering steps were incorporated (some of which were applied do to the results of this analysis) is of high quality, and that additional improvements would require manual curations.

5.2.6 An Overview of Potato Gene Expression

The data analysis of the mRNAseq gene expression data for RH and DM, which will be described in the following section was the basis for several detailed studies described both in the main and supplementary text of the published article. Among these is the section regarding tuber biology in the main text, cf. (The Potato Genome Sequencing Consortium *et al.*, 2011). Here, the gene expression analysis of starch metabolism genes was a significant contribution. This analysis will be described in section 5.2.7. In the current section, an overview analysis of the DM and RH transcriptomes focused on gene expression differences between genotypes and tissues will be described.

5.2.6.1 Methods

The primary data analysis was performed in collaboration with Brett Whitty at Michigan State University and is also described in the methods section in the published article. Shortly; All 32 DM and 16 RH RNAseq libraries (NCBI Sequence Read Archive (SRA030516; study SRP005965[36]) and the European Nucleotide Database ArrayExpress Database (E-MTAB-552; study ERP000527[37]), respectively) were mapped to the DM V3.0 genome sequence using Tophat (Trapnell, Pachter & Salzberg, 2009). The DM V3.4 representative mRNA transcripts gene annotation was used as input for calculation of expression values given as fragments per kilobase per million mapped reads (FPKM) using Cufflinks (Trapnell *et al.*, 2011). Cufflinks was run with default settings, with a maximum intron length of 15,000. FPKM values were reported and tabulated for each transcript.

A subset of the gene expression data set was selected, omitting samples from whole plant or stressed tissues, and genes with a FPKM value less than 5 leaving 26,219 transcripts. This set was subjected to complete linkage hierarchical clustering using the clustering software Cluster 3.0 (de Hoon *et al.*, 2004). Clustering was performed on both transcripts and samples using uncentered Pearson's correlation as distance measure. For visualization purposes, FPKM values were subsequently normalized to the maximum expression of each transcript. The same subset was subjected to principal component analysis (PCA). This was performed using the nonlinear iterative partial least squares algorithm using the software program The Unscrambler v 9.8 (Wass, 2005). PCA was performed on an auto-scaled and a centered only data set, respectively.

5.2.6.2 Results

The expression values of all genes with a maximal FPKM value ≥ 5 from healthy tissues only (26.219 genes and 34 samples) were subjected to PCA to initially analyze the gene expression changes and elucidate whether biologically relevant variation was present in the data. As seen in Figure 5-15 panel A, when performing PCA on a centered only data set (essentially assuming that the higher expression of a gene the more important the gene is) the greatest directions of variation (PC1 and PC2) separates the samples into four groups of tissues: tuber derived, green, dark, and flower tissues. Hence, the data contains variance that is relevant in order to study tissue function and biology in terms of gene expression. Furthermore, the older the tissue, the more diverged the tissues become from each other in the PCA plot. This is

[36] Available at: http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP005965
[37] Available at: http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-552

in agreement with the hypothesis that the more differentiated (and specialized) a tissue becomes, the more extreme its gene expression profile becomes. The centered only PCA also shows that tissues in the dark group have fewer genes with extreme expression values than the three other groups, since this group is located in the center of the PCA scores plot. Even when investigating the next PCs, this group does not split out. PC3 for example, splits out RH and DM tuber tissues. Noticeably the callus sample, which is un-differentiated tissue belong to the dark group. When performing the PCA analysis on an auto-scaled data set (essentially assuming that all changes in gene expression is equally important irrespectively of absolute transcript level) the greatest directions of variation (PC1 and PC2) separates the RH samples from the DM samples, cf. Figure 5-15 panel B. Moreover, this PCA splits out the dark group, which can indicate that this tissue group is characterized by changes in expression of many genes, although not the genes with the most extreme expression values.

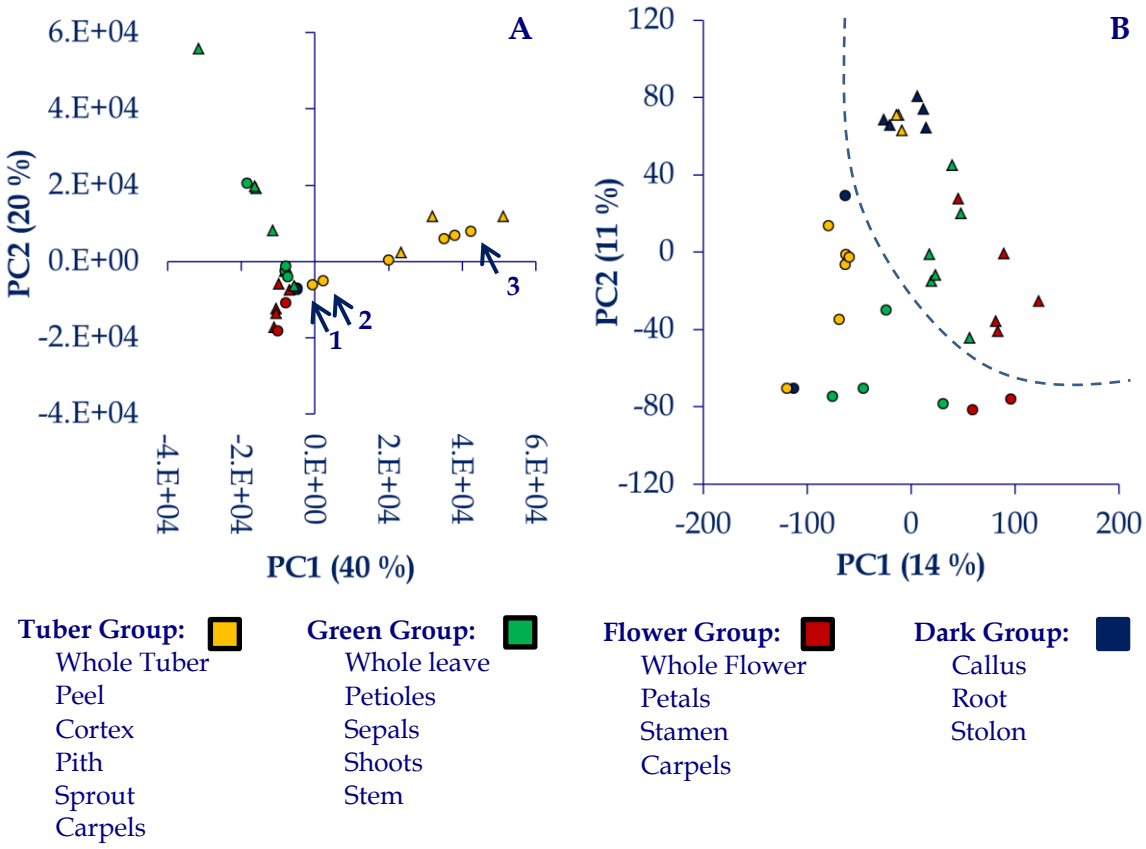


Figure 5-15 Scores plots of the principal component analysis (PCA) performed on a centered only and an auto-scaled data set consisting of FKPM values ≥ 5 from 34 DM or RH mRNAseq libraries originating from various tissues and developmental stages. **(A)** The centered only PCA shows separation of samples into four of tissue groups (Flower, green tissue, dark tissue and tuber tissue). **(B)** The auto-scaled PCA shows separation of RH and DM samples (indicated by a dashed line). Triangles indicate DM samples. Circles indicate RH samples. The tuber group marked in yellow contains whole tuber, peel, cortex pith, and sprout samples. The dark group marked in blue contains callus root and stolon samples. The flower group marked in red contains whole flower, petal, stamen, and carpel samples. The green group marked in green contains whole leave, petiole, sepal and shoot samples. Percentages indicate amount of explained variance. The plot shows that older and more diverged tissues have more extreme PC values. This is exemplified by RH tuber sprout (1), RH young tuber (2), and RH mature tuber (3). PC = principal component.

Subsequently, the data set was subjected to hierarchically clustering. Importantly, the samples of healthy tissues cluster very nicely together in the four functional groups determined by the PCA analysis and with obvious biological sense, cf. Figure 5-16. The clustering was

inspected for structure and several interesting sub-clusters of genes that showed tissue specific expression were identified. To confirm that these sub-clusters contain the genes that caused the separation in tissue groups in the PCA analysis, the ten most extreme genes of tuber, green and flower group (based on the loadings plot of the centered only PCA) were confirmed to be part of the relevant sub-clusters. Interestingly, while tissue specific sub-clusters are common to DM and RH for the flower, stolon, and green tissues, sub-clusters specific for tuber are separated into DM and RH specific sub-clusters, respectively. An example of that the clustering reflects the tissue types are seen in the bottom part of the dendrogram, where genes that are highly expressed in flower associated tissues are clustered. These genes subsequently split up in a group highly expressed in all flower associated tissues (red square), and a group with specific expression in the stamen and flowers (green square), cf. Figure 5-16. Interestingly, the RH stolon and young tuber samples are found in a sub-cluster, while the RH mature tuber, peel, pith and cortex samples are found in a different sub-cluster. It is therefore possible to quickly identify genes with differential expression between these two sub-clusters. These genes are potentially important for tuber development.

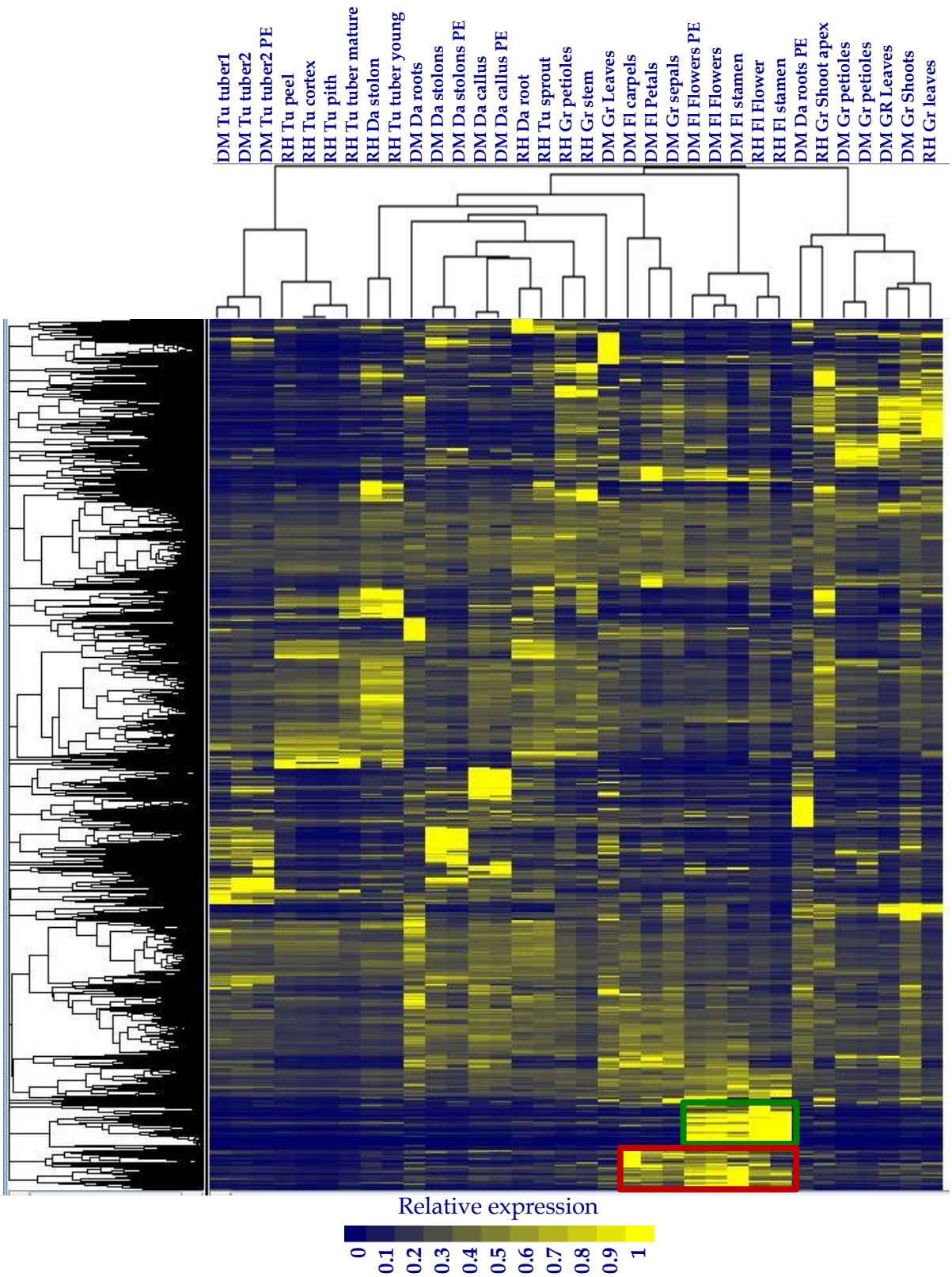


Figure 5-16 Dendrogram of hierarchically clustering of FKPM values from 26.219 genes from healthy tissues of RH and DM. Subsequent to clustering, expression values were normalized to the maximum expression level of each gene for visualization purposes, hereby ranging between 0 (dark blue) and 1 (yellow). Generally samples cluster according to tissue type (Tu = tuber, Da = dark, Gr = green, and Fl = flower tissues, respectively). The green square indicates genes with specific expression in flowers and stamen. The red square indicates genes with specific expression in all flower associated tissues.

5.2.6.3 Discussion and Conclusions

It is not surprising that the gene expression that characterizes tissues is dominated by a relatively small subset of genes that are relatively highly expressed (e.g. RubisCo in green tissues and storage proteins, like patatins in tubers). However, the variation structure of the data additionally suggests, that the difference between genotypes is made up of by a very large number of subtle changes in gene expression affecting a much larger number of genes; and that the effect of these changes combined are in fact greater than those changes causing tissue difference. This is an important issue when performing global gene expression analyses, where the importance of absolute versus relative gene expression changes is an ongoing discussion. It should be emphasized that this does not mean that individual low level expressed genes cannot convey important biological change in relation to tissue development and function, but rather that the dominating differences in gene expression between tissues are caused by subsets of highly expressed genes and between species are dominated by smaller changes in a much wider set of genes. This was also observed in the analysis of the high-replicate groups, where two different potato cultivars were compared, cf. section 4.3.1. The observed division of RH and DM tuber tissues in the hierarchical clustering could possibly reflect the high degree of selection on tuber characteristics that has been imposed on the RH genotype, which is closer related to modern European potato cultivars than the DM genotype.

The result of the gene expression overview was used as the starting point for many more detailed analyses of small subset of genes performed by different research group in the PGSC; some of which are described in the published article (The Potato Genome Sequencing Consortium *et al.*, 2011). Both the clustering and the PCA proved to facilitate these analyses by enabling a quick overview of subset of genes, and by highlighting genes with interesting expression patterns in regards to the analyses. Since the data set contained no biological replicates, it was chosen not to perform detection of differentially expressed genes, although methods for this are available (e.g. EdgeR (Robinson, McCarthy & Smyth, 2010)). Regardless, the lack of biological replicates, it was possible to answer several biological questions, and propose new biological hypotheses based on the gene expression data. An example of this is the gene expression analysis of *S. tuberosum* starch metabolism, which will be described in the following section.

5.2.7 Gene Expression Analysis of Starch Metabolism

Current commercial potato cultivars generally have a high starch content, and hence high yield. This seems a logical consequence of the high degree of selection on tuber characteristics that has been imposed on the modern European potato, cf. Figure 5-17. However, how this selection has affected the transcriptome is unknown. Therefore, a gene expression analysis comparing the starch metabolism transcriptome of RH, which is closer to modern potato cultivars with that of DM, could provide important clues to which genes are important for accumulation of starch in the tubers, and hence could be potential candidate genes for manipulation of the starch metabolism in *S. tuberosum*.

Figure 5-17 Potato tubers from the DM (left) and RH (right) genotypes. The differences between the genotypes reflect the selection of tuber characteristics that has been imposed on the modern European potato.

Starch accumulation in potato tubers is a result of metabolic processes that are highly redundant in regards to gene isoforms and biological pathways, cf. Figure 5-18 and Table 5-7. Starch synthesis in the tuber starts in the cytosol by conversion of sucrose into glucose-phosphates. These, glucose-1-phosphate or glucose-6-phosphate, are transported into the amyloplast (Fettke *et al.*, 2010; Tauberger *et al.*, 2000) and by direct incorporation of glucose-1-phosphate or via ADP-Glucose converted into starch, cf. Figure 5-18. Starch breakdown occurs via phosphorylytic or hydrolytic (by the action of α-amylase and β-amylase) reactions. In plants, starch synthesis takes place not only in storage organs but also in leaves, where transient starch produced during the day is consumed during the night. The expression of many genes of starch metabolism has been analyzed in *S. tuberosum* but completion of the genome sequence has enabled the creation of a more complete overview of gene isoform usage in starch metabolism. Here, a description of the starch metabolism transcriptome in leaves, tubers, and stolons will be described.

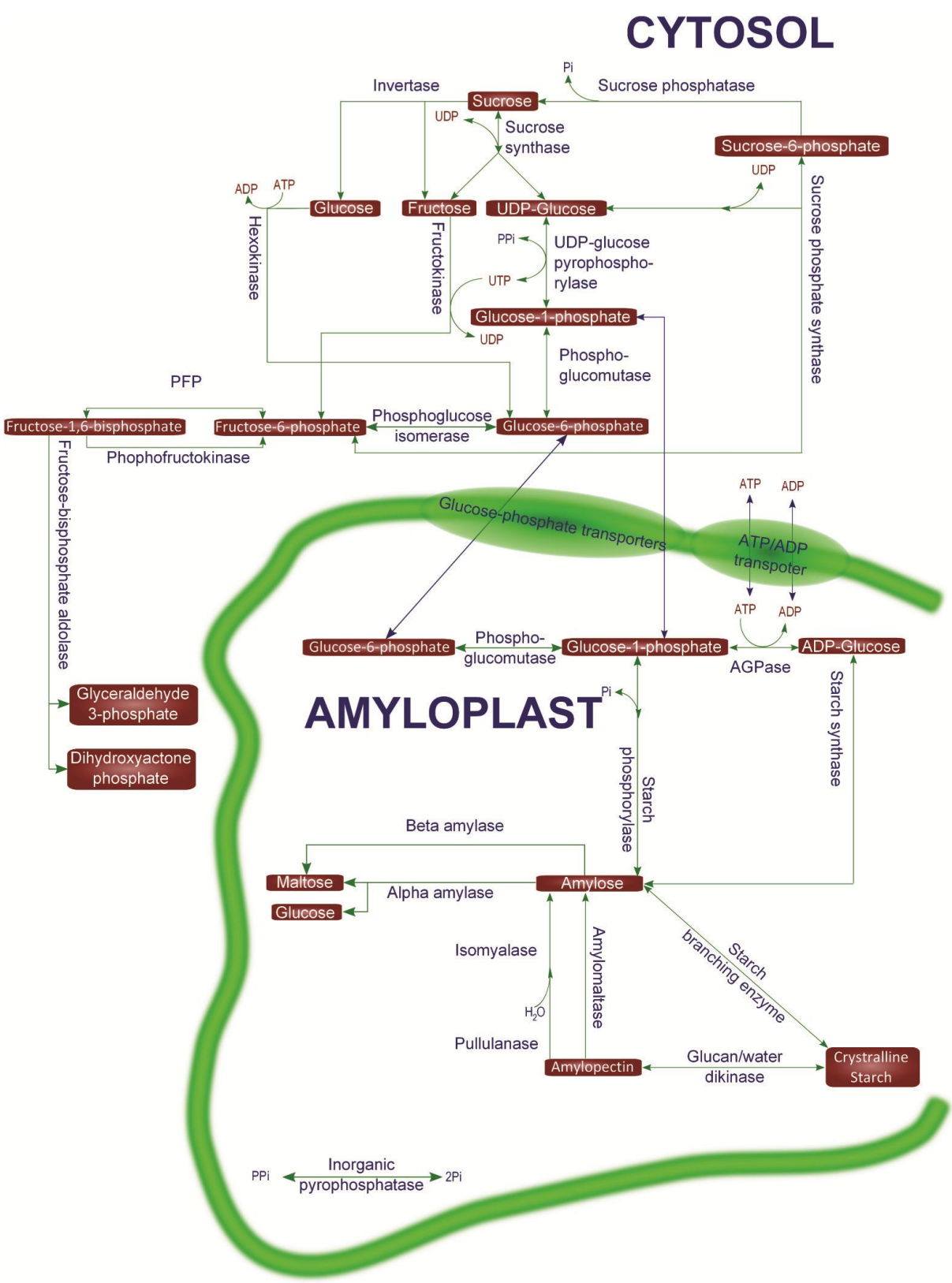


Figure 5-18 Overview of the *S. tuberosum* starch metabolism. Starch synthesis starts in the cytosol by conversion of sucrose into glucose-phosphates. These are transported into the amyloplast and by direct incorporation of glucose-1-phosphate or via ADP-Glucose converted into starch. Starch breakdown occurs via phosphorylytic or hydrolytic (by the action of α-amylase and β-amylase) reactions. Green arrows indicate conversion. Blue arrows indicate transport. PFP = Pyrophosphate-fructose-6-phosphate-1-phosphotransferase.

5.2.7.1 Methods

DM and RH RNAseq libraries from leaf, tuber and stolon (NCBI Sequence Read Archive (SRA030516; study SRP005965[38]) and the European Nucleotide Database ArrayExpress Database (E-MTAB-552; study ERP000527[39]), respectively) were subjected to RNAseq analysis using the CLC Genomics workbench version 3.7.1. Reads were mapped to the DM V3.0 genome sequence. Representative mRNA transcripts of the DM V3.2 gene annotation was used as annotation, incorporating the manual curations of starch metabolism genes described in section 5.2.5. Expression values were calculated and reported as reads per kilobase per million mapped reads (RPKM) and tabulated for each transcript. The expression values were subjected to complete linkage hierarchical clustering using the clustering software Cluster 3.0 (de Hoon *et al.*, 2004). Clustering was performed on both transcripts and samples using un-centered Pearson's correlation as distance measure. For visualization purposes, RPKM values were normalized to the maximum expression of each transcript. The expression values were also subjected to principal component analysis (PCA). This was performed using the nonlinear iterative partial least squares algorithm using the software program The Unscrambler v 9.8 (Wass, 2005). PCA was performed on an auto-scaled and a centered only data set, respectively. For 6 samples (1 from each tissue and genotype) the absolute expression levels and relative expression values (to the maximum of each transcript) was visualized as heat maps using the online visualization tool Prometra[40]

5.2.7.2 Results

An overview of the entire transcriptome of the starch metabolism genes from 6 samples (DM tuber, RH mature tuber, DM and RH stolons, and DM and RH leaves) is depicted in Figure 5-19. Interesting observations can be made from both the relative and absolute expression values. Firstly, by looking at the absolute expression levels a wide dynamic range between different gene loci can be observed. This can be exemplified by many of the enzymes where only one or a few of the isoforms are significantly expressed. A good example is starch synthase where the expression level of PGSC0003DMG200012111 is between 17 and 49 times higher than any of the other 6 gene isoforms. Secondly, by looking at the relative expression levels a large degree of tissue specific expression can be observed. A good example is Fructose-bisphosphate Aldolase, which is a part of the glycolysis. Here, there are 4 loci mostly expressed in leaves, two loci being expressed both in stolon and tubers, and 1 loci mostly being expressed in stolons. Another is starch phosphorylase, where the gene loci all show tissue specific expression, the two loci with the highest expression being tuber specific

When investigating all 15 available samples, both a hierarchical clustering, cf. Figure 5-20 and a PCA, cf. Figure 5-21, clearly split out the samples in the three tissue groups; leaves, stolons, and tubers. The clustering is also able to split the samples according to genotypes, splitting RH and DM tuber samples in two different sub-clusters. It is clear from the resulting dendrogram of the clustering that most genes are either highly expressed in stolons and developmental closely related tissues (such as young tubers), or show specific expression in tubers or leaves, cf. Figure 5-20.

[38] Available at: http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP005965
[39] Available at: http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-552
[40] Aavailable online at: https://prometra.cebitec.uni-bielefeld.de/cgi-bin/login.cgi

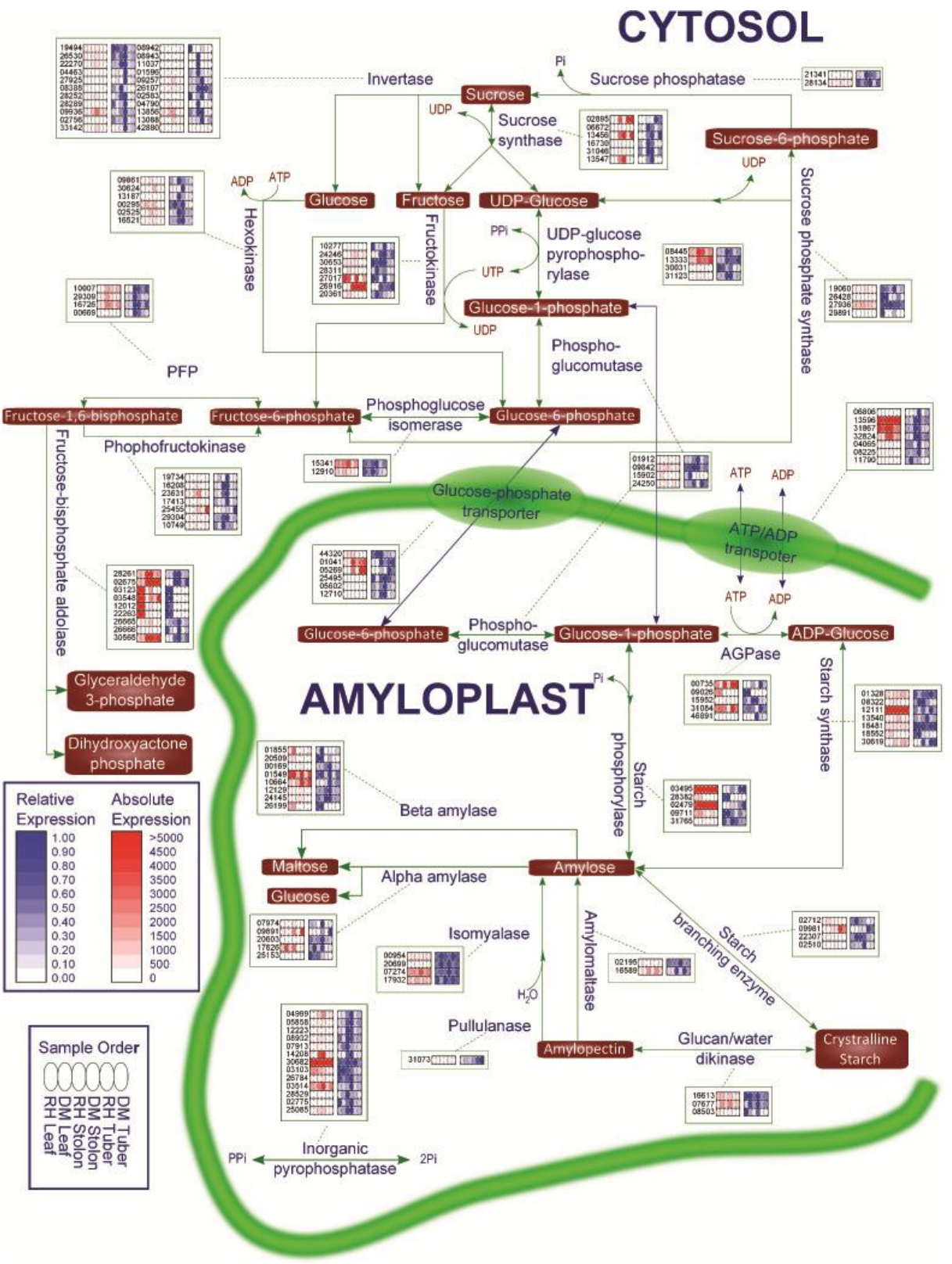


Figure 5-19 An overview of the entire transcriptome of the starch metabolism genes from 6 samples (DM tuber, RH mature tuber, DM and RH stolons, and DM and RH leaves). For each enzymatic reaction, the relative (marked in blue) and the absolute expression in RKPM (marked in red is shown). Gene loci are denoted with the last 5 digits of the PGSC identifier (PGSC0003DMG2000xxxxx). Green arrows indicate conversion. Blue arrows indicate transport. PFP = Pyrophosphate-fructose-6-phosphate-1-phosphotransferase.

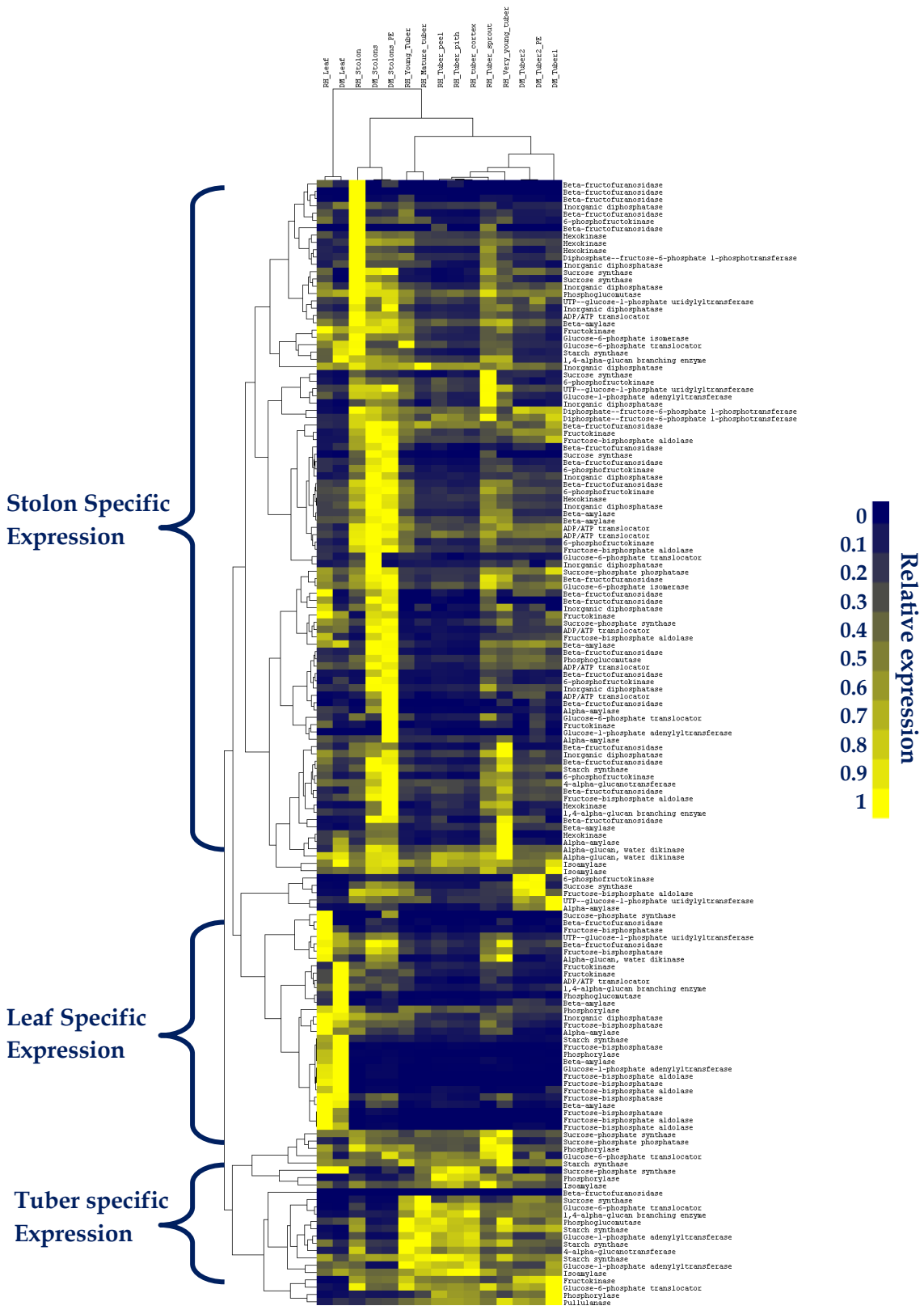


Figure 5-20 Resulting dendrogram of a hierarchical clustering of *S. tuberosum* starch metabolism gene expression in 15 RH or DM samples from tuber, leaf, and stolon tissues. Loci in the top part of the dendrogram are highly expressed in stolons and developmentally closely related tissues such as young tubers. Loci with leaf or tuber specific gene expression is marked in the bottom part of the dendrogram Clustering was performed on absolute RKPM values, and subsequently normalized to the highest expression level of each gene for visualization purposes. Notice that the samples cluster according to tissue type. PE = sampled sequenced with paired end sequencing.

The majority of the genes involved in starch metabolism have the highest expression in the stolon tissue. However, these are often less expressed than the gene loci showing leaf or tuber specific expression. This together with the smaller amount of loci showing leaf or tuber specific expression shows the specialization of the transcriptome caused by tissue development.

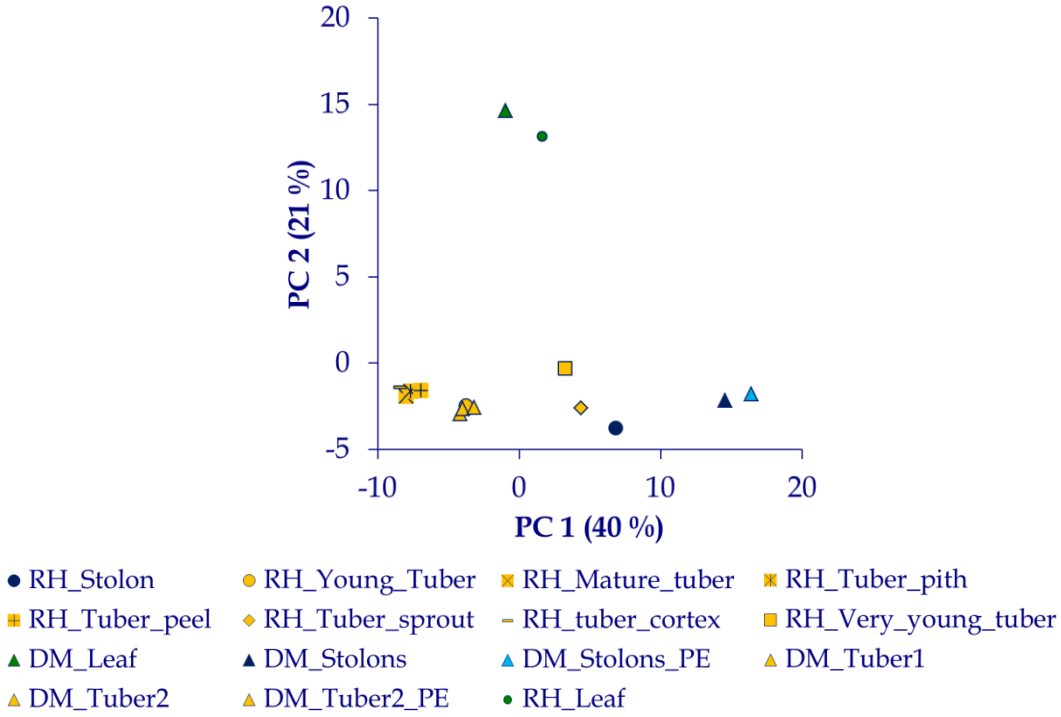


Figure 5-21 PCA of all 15 stolon (blue), tuber (yellow) or leaf (green) samples from RH and DM. The PCA was performed on an auto-scaled data set of RKPM expression values. Explained variance is given a percentage after the principal component (PC). All DM samples are marked by triangles. PE = paired end.

An interesting observation can be made from the PCA scores plot, cf. Figure 5-21. The first principal component (PC1), which explains most of the variance in the data set (40 %), seems to capture variation in the data caused by tissue development and differentiation over time. Looking at RH tuber samples, these are organized from right to left as stolon -> sprout -> very young -> young -> mature tuber, reflecting the developmental stage of this tissue. This variation is larger than the variation between leaf and tuber samples captured by PC2, cf. Figure 5-21. This indicates that there is a larger overall difference in starch metabolism gene expression comparing tuber tissue at different stages than tuber and leaf tissue taken at the same developmental stage. In the published article, we state that "*considerably greater levels of α-amylase (10–25-fold) and β-amylase (5–10-fold) mRNAs were found in DM tubers compared to RH*", (cf. p. 4 (The Potato Genome Sequencing Consortium *et al.*, 2011)). This conclusion was based a comparison between normalized expression levels of the most abundant loci of α- and β- amylase from DM tubers and the RH mature tuber sample, and differ slightly from the expression levels given as FKPM values listed in Table 5-9. Here, expression levels from RH tuber samples from different developmental stages are given along with biological and technical replicates from DM tuber samples. A great dynamic range can be observed for the expression levels of the 2 highest expressed loci of α- and β-amylase (PGSC0003DMG200009891, PGSC0003DMG200017626, PGSC0003DMG200001549, and PGSC0003DMG200010664, respectively.

Table 5-9 FKPM values for loci encoding α-amylase or β-amylase. Expression values for the two most expressed loci of each enzyme are marked in bold. * "PGSC0003DMG2000" is omitted from all PGSC IDs. ** Fold difference in expression level between the average expression level of the two DM samples and the mature RH tuber sample. "-" indicates higher expression in DM, "+" indicates higher expression in RH.

PGSC ID*	Amylase	DM tuber Tissues		RH Tuber Tissues				Fold Diff.**
		Sample 1	Sample 2	Sprout	Very young	Young	Mature	
09891	Alpha	5148	9586	1733	909	328	153	-48
17626	Alpha	459	187	1607	5317	403	128	-2.5
07974	Alpha	53	28	29	35	8	5	-8.1
20603	Alpha	20	19	51	117	39	41	+2.1
25153	Alpha	5	2	56	63	48	14	+4.0
10664	Beta	3625	4671	6956	18294	1229	1975	-2.1
01549	Beta	3730	3279	3393	3012	1285	657	-5.3
01855	Beta	725	311	141	379	14	27	-19
12129	Beta	120	133	335	288	240	108	-1.2
24145	Beta	196	128	378	418	318	176	+1.0
00169	Beta	66	67	247	250	122	47	-1.4
20509	Beta	17	31	0	152	5	19	-1.3
26199	Beta	3	8	72	27	15	4	-1.4

5.2.7.3 Discussion and Conclusions

Candidate gene loci, which could explain the phenotypical differences between DM and RH tubers, were detected in the starch metabolism transcriptome. Several gene loci related to starch synthesis had higher expression in RH mature tuber compared to DM tubers (e.g. AGPase, starch synthase, and starch branching enzyme), and gene loci of α- and β-amylase with relevant expression levels in tubers were more expressed in DM compared to RH, which indicates a higher hydrolytic starch degradation in DM tuber compared to RH. One of the conclusions of starch metabolism in the published article is that these gene expression differences between RH and DM are consistent with the concept that increasing tuber yield may be partially attained by selection for decreased activity of the hydrolytic starch degradation pathway (The Potato Genome Sequencing Consortium *et al.*, 2011). This is truly a plausible explanation for the phenotypical differences, and gives rise for further investigations of the gene expression of these candidate genes. In this regard, Scheidig *et al.* identified a *S. tuberosum* β-amylase gene (PCT-BMYI), and showed that silencing of this gene gave a starch-excess phenotype in leaves (Scheidig *et al.*, 2002). The corresponding homologue in the DM genome is PGSC0003DMG200001855. This gene shows leaf specific expression, cf. Figure 5-19. This is well in line with the results obtained be Scheidig *et al.* (Scheidig *et al.*, 2002), and could indicate that low expression of genes with tuber specific expression encoding β-amylase would lead to an increase in the starch content of the tubers. In regards to the importance of α-amylase, Cochrane *et al.* showed a positive correlation between the α-amylase activity and the amount of reducing sugars (degraded starch) during storage (Cochrane *et al.*, 1991). This again indicates that that low expression of genes with tuber specific expression encoding α-amylase would lead to an increase in the starch content of the tubers. When comparing the phosphorylytic and hydrolytic degradation pathways, it seems reasonable that low expression of α- and β-amylase involved in the hydrolytic pathway have been unknowingly selected for rather than low expression of starch phosphorylase. The equilibrium state of the hydrolytic reactions of α- and β-amylase are shifted far towards maltose and glucose and is in practice irreversible,

whereas Fettke *et al.* showed that the reaction catalyzed by starch phosphorylase also can be a part of starch synthesis (Fettke *et al.*, 2010).

The gene expression of some of the candidate genes showing differential expression between DM and RH tuber are most likely relevant for the starch content of potato tubers (and hence the yield). However, conclusions from a comparison between mature RH tubers and the DM tubers should be made with caution. Although a direct comparison between the DM and RH is relevant, it is difficult to achieve developmental facing, i.e. get samples from both genotypes at the same developmental stage. Although DM is fairly vigorous, it grows slower than the RH genotype (The Potato Genome Sequencing Consortium *et al.*, 2011). The PCA analysis of all tuber samples from both DM and RH indicates that the transcriptome of the DM tubers are more similar to the RH young tuber, and not the mature tuber, which the comparison has been made on. The analysis also showed that this developmental facing of samples is important, when comparing gene expression profiles between different genotypes, why the largest amount of variation in the starch metabolism gene expression data set could be explained by the developmental stage of the tuber samples, cf. Figure 5-21.

The analysis of *S. tuberosum* starch metabolism genes can facilitate the development of higher yielding potato cultivars, e.g. in regards to selection or in the form of gene modification. At first it might seem a daunting task to manipulate the carbon flux in such a redundant pathway containing many gene isoforms in nearly every enzymatic step. However, most gene loci are likely to have little influence of the net reaction because of low expression and consequently low abundance of the enzymes they encode. An investigation of the gene expression at several developmental stages is of course highly relevant, why several of the starch metabolism genes show larges variation in the expression levels over time. Moreover, several gene loci show tissue specific gene expression. Therefore, it seems possible to affect the starch metabolism in the sink organs without affecting the starch metabolism in the leaves, which would properly have large consequences for the plant vigor. This can be exemplified by the work study of Scheidig *et al.* who identified a β-amylase gene, which gave a starch-excess leaf phenotype when knocked out (Scheidig *et al.*, 2002). The same gene was in this study shown to have leaf specific expression. Therefore, by selecting the properly expressed gene loci (in this case often tuber specific), it could be possible to manipulate the carbon flux in the starch metabolism by altering the expression of these loci and hence the concentration of the enzymes they encode. However, this analysis in line with the results by Scheidig *et al.* (Scheidig *et al.*, 2002) also showed that the genes encoding enzymes of the starch metabolism have a great dynamic range in their expression at different developmental stages. This complicates the task of manipulation of the starch metabolism.

5.3 Conclusions and Perspectives

As stated in the published article, the potato genome sequence provides a new resource for use in breeding since many traits of interest to plant breeders are quantitative in nature and the genome sequence will simplify both their characterization and deployment in cultivars (The Potato Genome Sequencing Consortium *et al.*, 2011). It is the author's opinion that although the release of the sequence genome sequence is a major milestone for potato genomic research, it should not be the last. The quality of the genome sequence and accompanying gene annotation is of high quality taken into account that the annotation was made automatically. Substantial efforts were made to improve both the linking of scaffolds to super scaffolds and finally construction of pseudomolecules. During a ~2 year period, improvements were made to the gene annotation incorporating a new method for gene prediction based on NGS (mRNAseq) data. After several rounds of filterings, the final annotation is to the author's opinion as good as it can get using global and automatic methods. A further improvement would require far more human intervention, and would therefore be both costly and time consuming. However, as the potato research community starts to use the genome sequence, an opportunity arises to improve the gene annotation. When researchers perform detailed analyses on a low number of genes, errors can easily be discovered and manually corrected. However, to incorporate these corrections into future updates, a frame work for manual editing and error reporting needs to exist. Such frame work does not exist today, because no funding exists within the PGSC for maintenance and updating of the genome sequence and annotation. Hopefully, such framework will be made ensuring that the first version of the genome sequence and gene annotation is not engraved in stone, but corrections and improvements can be made using the combined efforts of the potato research community. Currently funding for such projects is hard to achieve, exemplified by the lack of funding to properly the most used and valued plant genetics database, namely The Arabidopsis Information Resource (TAIR) (Abbott, 2009)[41]. However to ensure the usefulness of a genome sequence, such databases are needed. In regards to the potato genome sequence, a possible solution could be a joint effort from the entire Solanaceae research community. Here, The Sol Genomics Network (SGN) (Bombarely *et al.*, 2011), which already exists could be an excellent bioinformatic frame work to facilitate not just potato breeding, but breeding of species in the entire Solanaceae family.

[41] Available at: http://www.arabidopsis.org/

Chapter 6

Discussion & Conclusions

Detailed discussion of the results can be found in the individual chapters. Here, an overall discussion and future perspectives of the accomplishments of the current PhD project will be presented. The fundamental basis of the discussion will be the overall theme of the current thesis; namely development of a bioinformatic framework for DNA sequence based transcriptomics to facilitate biological interpretations.

The overall goal of almost any molecular biology study is elucidation of the sources causing phenotypical differences between individuals. This is the case in all research fields from human disease to plant research. Using the central dogma as a framework, the phenotypical differences can be caused by differences at different levels, i.e. starting from the DNA → RNA → protein → metabolite and ending at the phenotype level. With the development of "omics" research fields, it has become possible to perform a study at all the levels – so why choose transcriptomics? Obviously, it can be argued that biological relevance of the observed differences increases the closer to the phenotypical level the study is performed. However, with today's technologies, the ability to measure the different "omes" is inversely proportional to the biological relevance, so the simple answer to the above question is "because it's possible"! Due to the amplification cascade of the central dogma, the dynamic range is in most cases far too large to capture for today's measurement techniques already at the proteome level (e.g. in the case of mass spectroscopy). However, with the development of next generation sequencing technologies, it has become possible to nearly capture the entire "omes" of both of the lower levels, i.e. the genome and the transcriptome. When interpreting the results of a transcriptome study, the correlation between the transcriptome and the proteome, and the level of complexity that comes here of should of course always be considered. The mRNA levels are the combined result of mRNA transcription and degradation. Following this, protein levels are the combined result of protein translation and degradation, and adding to the complexity their biological function is further the result of several other mechanisms such as post-translatoric modifications and protein translocations. Therefore, the relevance of the results of transcriptome studies could be questioned. However, Schwanhäusser *et al.* recently showed in a study performed on mammalian tissue culture cell, that a good correlation between mRNA and protein levels exists, and that mRNA levels are the most important factor when predicting the protein levels (Schwanhüusser *et al.*, 2011). Therefore, it is obvious that findings from transcriptome analyses always need to be confirmed and further investigated, either with other higher level "omics" methods (proteonomics or metabolomics), or more detailed analyses to confirm, that the differences found in the transcriptome, in fact are causal for the phenotype in question.

There are different kinds of transcriptome studies. These depend on the experimental setup each having different expected outcomes and a different set of pitfalls, some of which have been highlighted in the current thesis. The first kind can be described as a pioneer study, where absolutely nothing is known about the phenotype investigated. With today's standards in DNA sequencing, if the transcriptome of an organism is completely unknown, it is possible to quickly generate a model of it, either though genome or transcriptome sequencing, followed by *de novo* assembly and annotation. In the case of more complex eukaryotic organisms transcriptome sequencing would in most cases be advantageous. Although the challenges of the assembly are greater, cf. section 1.2.3.6, the size of the transcriptome is in these cases far smaller than that of the genome and less sequencing efforts are therefore

needed. Secondly, the annotation process is simpler due to the lack of the intron-exon gene structure found in the genome facilitating a more precise transcript prediction. Furthermore, the same data set used for transcriptome reconstruction can also be used for the gene expression study itself. When performing a pioneer study, the focus of the transcriptome analysis should be to generate hypotheses that can explain the phenotype in question. To facilitate this hypothesis generation, careful design of the experiment is needed, e.g. by choosing conditions or individuals that clearly split out the phenotype in question to facilitate the elucidation of its reflection in the transcriptome. A second kind of transcriptome study is to increase the knowledge of a phenotype, which has been partially explained. An example could be a complex signaling or metabolic pathways, where some genes have been found to be involved, but the exact order and interaction between the members of the pathway has not been elucidated. The analysis presented in the current thesis of *L. japonicus* during nodulation of is a fine example of such study. Nodulation is extremely complex and involves several concurrent processes, where members interact and proteins can have multiple functions in more than one pathway. This of course complicates the interpretation of a gene expression analysis, and although some is known about the nodulation signaling pathway, much is still left to be unraveled. Nodulation leads to a large phenotypical difference (roots vs. nodules), and differences between the start and end state are therefore easily detected. However, a comparison between these states will only lead to few clues to the elucidation of the transcriptome of the nodulation process itself. The current study was designed to enable an investigation of the entire nodulation process. However, the conclusion must be that the time resolution of the sampling was insufficient to acquire this goal. Since the design of any study is limited by time and cost, researchers have to choose whether to know little about a lot or choose to design a more detailed experiment focusing on a part of the process. It is the author's belief that the latter is more fruitful, i.e. *divide et impera.* Two focused transcriptome studies on different parts of a process would elucidate more than two more global studies. In the case of the *L. japonicus* study, this could have been brought into effect by performing a more detailed analysis with higher sampling frequency of either the earlier time points focusing more on the initiation of the nodulation process or the later time points focusing more on nitrogen fixation. It could seem that the excitement caused by the development of next generation sequencing technologies, have caused researchers to choose the more global approach when designing experiments - a choice which should be performed with great thought. In addition, the experimental conditions should be considered. As the analysis of the LSDS-project data sets showed, cf. section 4.2.6, reality comes with a price! Not surprisingly, the comparison between the data sets with the field grown or the greenhouse grown plants showed that additional biological variance between libraries representing the same biological group could be observed when growth conditions were under less stringent control. This unavoidable extra biological variation must be taken into consideration, in regards to number of biological replicates and pooling of samples, when designing the experiment. Again, time and costs restrict the design. Due to the noise found in gene expression data, it is the author's belief that a less global analysis containing less biological groups represented by more replicates is more likely to produce targets for further research than a more global analysis with more biological groups represented by less biological replicates. The final type of transcriptome analysis, which is represented in the current thesis, is a very detailed and less global analysis only involving a small subset of the transcriptome. When outlining the

manuscript for the potato genome sequencing paper, the extensive mRNAseq data set provided the opportunity to perform both a global and several potentially interesting more focused analyses of the *S. tuberosum* transcriptome. However, it was chosen to describe an already well-studied process, namely starch metabolism. This choice was of course made to highlight what phenotypical trait is unique for *S. tuberosum* compared to other plants with published genome sequences, namely the tuber, and here starch metabolism is a key metabolic pathway. This relatively simple process (only consisting of 23 enzymatic reactions, which were discovered to be encoded by 167 genes) made an in-depth analysis possible. Although the data set did not contain any replicates (besides biological groups representing different tissue types consisting of a sample from each of the two genotypes sequenced), a simple comparison between three tissues, only using six samples, provided new knowledge of an already well-studied process and new target genes for further research. This highlights the power of a transcriptome analysis, but in the same time also the need and usefulness of a well-annotated transcriptome.

Both the gene expression analysis of *L. japonicus* during nodulation, and the gene expression analysis of *S. tuberosum* clearly showed that transcript annotation is crucial for a successful transcriptome study; especially in regards to tag annotation when using tag based methods such as DeepSAGE, which has been the method of choice to generate most of the data sets analyzed in the current project. Results of a substantial part of the work presented here have facilitated improvements to the genome annotation for the model systems *L. japonicus* and *S. tuberosum*. Since the representative SAGE tag often is located in the 3'UTR region, improvement of the annotation of UTR regions have been a focus area in the presented work, and several algorithms dependent on prior annotation of the CDS regions were developed to facilitate this, cf. sections 3.2.4 and 5.2.2. However, the rapid development within algorithms for spliced alignment of mRNAseq reads and the lowering of sequencing costs have facilitated in higher quality genome assisted transcriptome reconstruction including annotation of the UTR regions. The potato genome sequence was the first published major eukaryotic genome that was annotated using an mRNAseq sequence data assisted method. Results presented here, clearly show the need for improvements of algorithms for genome assisted transcriptome reconstruction to reduce the number of mis-annotated transcripts caused by the noise found in mRNAseq data, cf. section 5.2.4. Whether this is possible or manual curation will be needed to a great extent to ensure a high quality genome annotation is unclear. However, semi-automated methods to flag potentially mis-annotated transcript models have been presented here, cf. section 5.2.5. These could facilitate less time consumption for manual curation.

The majority of the data sets analyzed in the current thesis were generated using the DeepSAGE technology. Comparisons of this were made to both the cDNA microarray and mRNAseq technologies. The comparison between Affymatrix and DeepSAGE data was not made between libraries originating from the same biological sample, but only similar samples in regards to genotype, developmental stage, and treatment. Here a fairly good correlation between the results of the two methods was shown; cf. section 3.2.5. A more direct comparison was performed against the other digital count method, mRNAseq, which is today's most widely used method for transcriptome analyses. Results showed that mRNAseq outperforms DeepSAGE in regards to reproducibility using the same sequencing power. Nearly

no technical variance was found in the mRNAseq data, whereas a substantial amount was found in the DeepSAGE data. In 2008, at the beginning of the current project the choice of technology for the transcriptome analyses presented here was obvious. The DeepSAGE technology outperformed mRNAseq in several aspects, cf. Table 6-1. Perhaps most importantly, a DeepSAGE library was 10 times cheaper to generate compared to an mRNAseq library. This was primarily caused by the lack of the possibility to multiplex samples during sequencing using the mRNAseq technology (although it could be customly designed). Moreover, the recommended amount of total RNA was higher for mRNAseq than for DeepSAGE, and there was not a large difference in read length between the two methods.

Table 6-1 Comparison of the DeepSAGE and mRNAseq technologies using Illumina sequencing. Statistics from 2008 and 2012 are based on running costs of the Genome Analyzer and HiSeq2000 platforms, respectively. * 2 x 150 bp sequencing runs yielding high quality data are routinely performed at AAU. **Estimated costs include chemical costs for mRNA purification and library preparation and sequencing costs. It does not include salary for technical personnel. ***The 2012 price for DeepSAGE is ~ 1,000 DKK.

	DeepSAGE 2008	mRNAseq 2008	mRNAseq 2012
Input requirements [µg total mRNA]	2	1-10	0.1 - 4
Read length [bp]	21	36	PE 2 x 100*
Estimated cost per library [DKK]**	1,300***	15,000	1,250
Max # samples per run	128	8	96

Following the development of the Illumina sequencing technology within the last four years including a 6 times reduction in the price for library preparation, lowering the RNA input requirements and the acquired ability to multiplex samples during sequencing, the mRNAseq technology has surpassed tag based methods such as the DeepSAGE technology in nearly all aspects. There are still minute advantages in regards to price and multiplexing for the DeepSAGE technology. However, the mRNAseq technology has several advantages over the DeepSAGE technology, such as: Simpler library preparation, simpler data analysis due to more accurate assigning of reads to transcripts facilitated by longer read length, additional possibilities such as genome annotation and transcriptome reconstruction, and as shown in the current thesis – significantly lower amount of technical variation. Therefore, the conclusion must be that mRNAseq is the best choice of today for transcriptome analyses.

Chapter 7

References

7.1 List of References

1 454 sequencing 2011, , *Whole Genome Sequencing* [Homepage of Roche Diagnostics Corporation], [Online]. Available: http://454.com/applications/whole-genome-sequencing/index.asp [2011, 07/26].

2 Abbott, A. 2009, "Plant genetics database at risk as funds run dry", *Nature,* vol. 462, no. 7271, pp. 258-259.

3 Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., Kerlavage, A.R., McCombie, W.R. & Venter, J.C. 1991, "Complementary DNA sequencing: Expressed sequence tags and human genome project", *Science,* vol. 252, no. 5013, pp. 1651-1656.

4 Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J.J., Mayer, P. & Kawashima, E. 2000, "Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms.", *Nucleic acids research,* vol. 28, no. 20.

5 Akmaev, V.R. 2008, "Correction of technology-related artifacts in serial analysis of gene expression.", *Methods in molecular biology (Clifton, N.J.),* vol. 387, pp. 133-142.

6 Akmaev, V.R. & Wang, C.J. 2004, "Correction of sequence-based artifacts in serial analysis of gene expression", *Bioinformatics,* vol. 20, no. 8, pp. 1254-1263.

7 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. 1990, "Basic local alignment search tool", *Journal of Molecular Biology,* vol. 215, no. 3, pp. 403-410.

8 Anders, S. & Huber, W. 2010, "Differential expression analysis for sequence count data", *Genome biology,* vol. 11, no. 10.

9 Andriankaja, A., Boisson-Dernier, A., Frances, L., Sauviac, L., Jauneau, A., Barker, D.G. & De Carvalho-Niebel, F. 2007, "AP2-ERF transcription factors mediate nod factor-dependent Mt ENOD11 activation in root hairs via a novel cis-regulatory motif", *Plant Cell,* vol. 19, no. 9, pp. 2866-2885.

10 Ansorge, W.J. 2009, "Next-generation DNA sequencing techniques", *New Biotechnology,* vol. 25, no. 4, pp. 195-203.

11 Arumuganathan, K. & Earle, E.D. 1991, "Nuclear DNA content of some important plant species", *Plant Molecular Biology Reporter,* vol. 9, no. 3, pp. 208-218.

12 Asamizu, E., Shimoda, Y., Kouchi, H., Tabata, S. & Sato, S. 2008, "A positive regulatory role for LjERF1 in the nodulation process is revealed by systematic analysis of nodule-associated transcription factors of Lotus japonicus", *Plant Physiology,* vol. 147, no. 4, pp. 2030-2040.

13 Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G. 2000, "Gene ontology: Tool for the unification of biology", *Nature genetics,* vol. 25, no. 1, pp. 25-29.

14 Atif, U., Earll, Eriksson, L., Johansson, E., Lord, P. & Margrett, S. 2003, "Analysis of gene expression datasets using partial least-squares discriminant analysis and principal-component analysis" in *Euroqsar 2002 Designing Drugs and Crop Protectants: Processes, Problems and Solutions*, eds. M.G. Ford, D. Livingstone, J. Dearden & H.V.d. Waterbeemd, Blackwell Science Inc, , pp. 369-370-373.

15 Au, K.F., Jiang, H., Lin, L., Xing, Y. & Wong, W.H. 2010, "Detection of splice junctions from paired-end RNA-seq data by SpliceMap", *Nucleic acids research,* vol. 38, no. 14, pp. 4570-4578.

16 Au, P.C.K., Zhu, Q.-., Dennis, E.S. & Wang, M.-. 2011, "Long non-coding RNA-mediated mechanisms independent of the RNAi pathway in animals and plants", *RNA Biology,* vol. 8, no. 3, pp. 404-414.

17 Audic, S. & Claverie, J.-. 1997, "The significance of digital gene expression profiles", *Genome research,* vol. 7, no. 10, pp. 986-995.

18 Bachem, C.W.B., Van Der Hoeven, R.S., De Bruijn, S.M., Vreugdenhil, D., Zabeau, M. & Visser, R.G.F. 1996, "Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: Analysis of gene expression during potato tuber development", *Plant Journal,* vol. 9, no. 5, pp. 745-753.

19 Baggerly, K.A., Deng, L., Morris, J.S. & Aldaz, C.M. 2004, "Overdispersed logistic regression for SAGE: Modelling multiple groups and covariates", *BMC Bioinformatics,* vol. 5.

20 Baggerly, K.A., Deng, L., Morris, J.S. & Aldaz, C.M. 2003, "Differential expression in SAGE: Accounting for normal between-library variation", *Bioinformatics,* vol. 19, no. 12, pp. 1477-1483.

21 Barbazuk, W.B., Emrich, S.J., Chen, H.D., Li, L. & Schnable, P.S. 2007, "SNP discovery via 454 transcriptome sequencing", *Plant Journal,* vol. 51, no. 5, pp. 910-918.

22 Barker, D.G., Bianchi, S., Blondon, F., Dattée, Y., Duc, G., Essad, S., Flament, P., Gallusci, P., Génier, G., Guy, P., Muel, X., Tourneur, J., Dénarié, J. & Huguet, T. 1990, "Medicago truncatula, a model plant for studying the molecular genetics of the Rhizobium-legume symbiosis", *Plant Molecular Biology Reporter,* vol. 8, no. 1, pp. 40-49.

23 Bartlett, W.A. 1999, 01/01/1999-last update, *Biological Variation: Information Site for Laboratory Medicine* [Homepage of Ninewells Hospital & Medical School], [Online]. Available: http://biologicalvariation.com/ESW/Files/Biological_Variation_update_ed.pps [2012, 05/01].

24 Bassingthwaighte, J. 2007, 05/11-2007-last update, *Defining the Physiome* [Homepage of the NSR Physiome Project], [Online]. Available: http://www.physiome.org/About/index.html#physiome [2011, 07/15].

25 Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P. & Lander, E.S. 2002, "ARACHNE: A whole-genome shotgun assembler", *Genome research,* vol. 12, no. 1, pp. 177-189.

26 Benjamini, Y. & Hochberg, Y. 1995, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", *Journal of the Royal Statistical Society.Series B (Methodological),* vol. 57, no. 1, pp. 289-300.

27 Benovoy, D. & Drouin, G. 2006, "Processed pseudogenes, processed genes, and spontaneous mutations in the Arabidopsis genome", *Journal of Molecular Evolution,* vol. 62, no. 5, pp. 511-522.

28 Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Sayers, E.W. 2011, "GenBank", *Nucleic acids research,* vol. 39, no. Database issue, pp. D32-7.

29 Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M.J., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M.D., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Catenazzi, M.C.E., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K.V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Huw Jones, T.A., Kang, G.-., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ling Ng, B., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Pike, A.C., Chris Pinkard, D., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Chiva Rodriguez, A., Roe, P.M., Rogers, J., Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Ernest Sohna Sohna, J., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R. & Smith, A.J. 2008, "Accurate whole human genome sequencing using reversible terminator chemistry", *Nature,* vol. 456, no. 7218, pp. 53-59.

30 Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin, R.D., Zhao, Y., Hirst, M., Schein, J.E., Horsman, D.E., Connors, J.M., Gascoyne, R.D., Marra, M.A. & Jones, S.J.M. 2009, "De novo transcriptome assembly with ABySS", *Bioinformatics,* vol. 25, no. 21, pp. 2872-2877.

31 Blackshaw, S., Kuo, W.P., Park, P.J., Tsujikawa, M., Gunnersen, J.M., Scott, H.S., Boon, W.M., Tan, S.S. & Cepko, C.L. 2003, "MicroSAGE is highly representative and reproducible but reveals major differences in gene expression among samples obtained from similar tissues.", *Genome biology,* vol. 4, no. 3.

32 Bloom, J.S., Khan, Z., Kruglyak, L., Singh, M. & Caudy, A.A. 2009, "Measuring differential gene expression by short read sequencing: Quantitative comparison to 2-channel gene expression microarrays", *BMC Genomics,* vol. 10.

33 Bombarely, A., Menda, N., Tecle, I.Y., Buels, R.M., Strickler, S., Fischer-York, T., Pujar, A., Leto, J., Gosselin, J. & Mueller, L.A. 2011, "The sol genomics network (solgenomics.net): Growing tomatoes using Perl", *Nucleic acids research,* vol. 39, no. SUPPL. 1, pp. D1149-D1155.

34 Bonferroni, C.E. 1935, "Il calcolo delle assicurazioni su gruppi di teste" in *Studi in Onore del Professore Salvatore Ortu Carboni*, pp. 13-60.

35 Boser, B.E., Guyon, I.M. & Vapnik, V.N. 1992, *A training algorithm for optimal margin classifiers*, ACM, Pittsburgh, Pennsylvania, United States.

36 Bradford, J.R., Hey, Y., Yates, T., Li, Y., Pepper, S.D. & Miller, C.J. 2010, "A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling", *BMC Genomics,* vol. 11, no. 1.

37 Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S.B., Krstic, P.S., Lindsay, S., Ling, X.S., Mastrangelo, C.H., Meller, A., Oliver, J.S., Pershin, Y.V., Ramsey, J.M., Riehn, R., Soni, G.V., Tabard-Cossa, V., Wanunu, M., Wiggin, M. & Schloss, J.A. 2008, "The potential and challenges of nanopore sequencing", *Nature biotechnology,* vol. 26, no. 10, pp. 1146-1153.

38 Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S.R. 2003, "Sequence information can be obtained from single DNA molecules", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 100, no. 7, pp. 3960-3964.

39 Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J.-. & Corcoran, K. 2000, "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays", *Nature biotechnology,* vol. 18, no. 6, pp. 630-634.

40 Bucher, M., Schroeer, B., Willmitzer, L. & Riesmeier, J.W. 1997, "Two genes encoding extensin-like proteins are predominantly expressed in tomato root hair cells", *Plant Molecular Biology,* vol. 35, no. 4, pp. 497-508.

41 Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. 2010, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments", *BMC Bioinformatics,* vol. 11.

42 Burrows, M. & Wheeler, D.J. 1994, "A block-sorting lossless data compression algorithm", *SRC Research Report,* vol. 124.

43 Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C. & Jaffe, D.B. 2008, "ALLPATHS: De novo assembly of whole-genome shotgun microreads", *Genome research,* vol. 18, no. 5, pp. 810-820.

44 Callinan, P.A. & Feinberg, A.P. 2006, "The emerging science of epigenomics.", *Human molecular genetics,* vol. 15 Spec No 1, pp. R95-101.

45 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. 2009, "BLAST+: Architecture and applications", *BMC Bioinformatics,* vol. 10.

46 Cappelletti, M. 2009, 17-09-2009-last update, *Innovations in Genomic Analysis: Downstream analysis of Illumina Sequencing Data* [Homepage of Illumina Inc.], [Online]. Available: http://mi.caspur.it/workshop_NGS09/docs/Cappelletti_NGS09.pdf [2011, 08-12].

47 Cappola, T.P. & Margulies, K.B. 2011, "Functional genomics applied to cardiovascular medicine", *Circulation,* vol. 124, no. 1, pp. 87-94.

48 Carra, A., Gambino, G., Urso, S. & Nervo, G. 2011, *Non Coding RNAs and Gene Silencing in Grape*, Springer Berlin Heidelberg.

49 Carvalho, H.G., Lopes-Cardoso, I.A., Lima, L.M., Melo, P.M. & Cullimore, J.V. 2003, "Nodule-specific modulation of glutamine synthetase in transgenic Medicago truncatula leads to inverse alterations in asparagine synthetase expression", *Plant Physiology,* vol. 133, no. 1, pp. 243-252.

50 CBM S.c.r.l. 2007, , *Plant Resistence Genes Project* [Homepage of CBM S.c.r.l.], [Online]. Available: http://prgdb.cbm.fvg.it/plants.php [2012, 02/12/2012].

51 Chaisson, M., Pevzner, P. & Tang, H. 2004, "Fragment assembly with short reads", *Bioinformatics,* vol. 20, no. 13, pp. 2067-2074.

52 Chaisson, M.J., Brinza, D. & Pevzner, P.A. 2009, "De novo fragment assembly with short mate-paired reads: Does the read length matter?", *Genome research,* vol. 19, no. 2, pp. 336-346.

53 Chaisson, M.J. & Pevzner, P.A. 2008, "Short read fragment assembly of bacterial genomes", *Genome research,* vol. 18, no. 2, pp. 324-330.

54 Chapman, H.W. 1958, "Tuberization in the Potato Plant", *Physiologia Plantarum,* vol. 11, no. 2, pp. 215-224.

55 Charpentier, M., Bredemeier, R., Wanner, G., Takeda, N., Schleiff, E. & Parniske, M. 2008, "Lotus japonicus Castor and Pollux are ion channels essential for perinuclear calcium spiking in legume root endosymbiosis", *Plant Cell,* vol. 20, no. 12, pp. 3467-3479.

56 Chen, H., Centola, M., Altschul, S.F. & Metzger, H. 1998, "Characterization of gene expression in resting and activated mast cells", *Journal of Experimental Medicine,* vol. 188, no. 9, pp. 1657-1668.

57 Chen, J.J., Delongchamp, R.R., Tsai, C.-., Hsueh, H.-., Sistare, F., Thompson, K.L., Desai, V.G. & Fuscoe, J.C. 2004, "Analysis of variance components in gene expression data", *Bioinformatics,* vol. 20, no. 9, pp. 1436-1446.

58 Chin, C.-., Sorenson, J., Harris, J.B., Robins, W.P., Charles, R.C., Jean-Charles, R.R., Bullard, J., Webster, D.R., Kasarskis, A., Peluso, P., Paxinos, E.E., Yamaichi, Y., Calderwood, S.B., Mekalanos, J.J., Schadt, E.E. & Waldor, M.K. 2011, "The origin of the Haitian cholera outbreak strain", *New England Journal of Medicine,* vol. 364, no. 1, pp. 33-42.

59 CLC Bio 2010, 11/05 2010-last update, *CLC Bio* [Homepage of CLC BIO], [Online]. Available: http://www.clcbio.com/index.php [2010, 27/05 2010].

60 Cloonan, N., Forrest, A.R.R., Kolle, G., Gardiner, B.B.A., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., Robertson, A.J., Perkins, A.C., Bruce, S.J., Lee, C.C., Ranade, S.S., Peckham, H.E., Manning, J.M., McKernan, K.J. & Grimmond, S.M. 2008, "Stem cell transcriptome profiling via massive-scale mRNA sequencing", *Nature Methods,* vol. 5, no. 7, pp. 613-619.

61 Cochrane, M.P., Duffus, C.M., Allison, M.J. & Mackay, G.R. 1991, "Amylolytic activity in stored potato tubers. 2. The effect of low-temperature storage on the activities of α-and β-amylase and α-glucosidase in potato tubers", *Potato Research,* vol. 34, no. 3, pp. 333-341.

62 Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M. & Jacobsen, S.E. 2008, "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning", *Nature,* vol. 452, no. 7184, pp. 215-219.

63 Collins, F.S., Lander, E.S., Rogers, J. & Waterson, R.H. 2004, "Finishing the euchromatic sequence of the human genome", *Nature,* vol. 431, no. 7011, pp. 931-945.

64 Crick, F. 1970, "Central dogma of molecular biology", *Nature,* vol. 227, no. 5258, pp. 561-563.

65 Daly, A.K. 2010, "Genome-wide association studies in pharmacogenomics", *Nature Reviews Genetics,* vol. 11, no. 4, pp. 241-246.

66 De Bona, F., Ossowski, S., Schneeberger, K. & Rätsch, G. 2008, "Optimal spliced alignments of short sequence reads", *Bioinformatics,* vol. 24, no. 16, pp. i174-i180.

67 de Hoon, M.J.L., Imoto, S., Nolan, J. & Miyano, S. 2004, "Open source clustering software", *Bioinformatics,* vol. 20, no. 9, pp. 1453-1454.

68 Dello Ioio, R., Nakamura, K., Moubayidin, L., Perilli, S., Taniguchi, M., Morita, M.T., Aoyama, T., Costantino, P. & Sabatini, S. 2008, "A genetic framework for the control of cell division and differentiation in the root meristem", *Science,* vol. 322, no. 5906, pp. 1380-1384.

69 Dennis Jr., G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. & Lempicki, R.A. 2003, "DAVID: Database for Annotation, Visualization, and Integrated Discovery.", *Genome biology,* vol. 4, no. 5.

70 Denoeud, F., Aury, J.-., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., Morgante, M., Valle, G., Wincker, P., Scarpelli, C., Jaillon, O. & Artiguenave, F. 2008, "Annotating genomes with massive-scale RNA sequencing", *Genome biology,* vol. 9, no. 12.

71 Dictionary Oxford English 2010, 07/01-last update, "*-ome, comb. form*". [Homepage of Oxford University Press], [Online]. Available: http://www.oed.com/view/Entry/131183.

72 DNA Data Bank of Japan 2011, 06/20-last update, *DDBJ Database Release Histry* [Homepage of DNA Data Bank of Japan], [Online]. Available: http://www.ddbj.nig.ac.jp/breakdown_stats/dbgrowth-e.html [2011, 07/18].

73 Dohm, J.C., Lottaz, C., Borodina, T. & Himmelbauer, H. 2008, "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing", *Nucleic acids research,* vol. 36, no. 16, pp. e105.

74 Dohm, J.C., Lottaz, C., Borodina, T. & Himmelbauer, H. 2007, "SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing", *Genome research,* vol. 17, no. 11, pp. 1697-1706.

75 Dong, Q., Schlueter, S.D. & Brendel, V. 2004, "PlantGDB, plant genome database and analysis tools", *Nucleic acids research,* vol. 32, no. DATABASE ISS., pp. D354-359.

76 Dressman, D., Yan, H., Traverso, G., Kinzler, K.W. & Vogelstein, B. 2003, "Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 100, no. 15, pp. 8817-8822.

77 Dudoit, S., Yang, Y.H., Callow, M.J. & Speed, T.P. 2002, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments", *Statistica Sinica,* vol. 12, no. 1, pp. 111-139.

78 Edwards, A. & Caskey, C.T. 1991, "Closure strategies for random DNA sequencing", *Methods,* vol. 3, no. 1, pp. 41-47.

79 Ehrhardt, D.W., Wais, R. & Long, S.R. 1996, "Calcium spiking in plant root hairs responding to rhizobium modulation signals", *Cell,* vol. 85, no. 5, pp. 673-681.

80 Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., DeWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. & Turner, S. 2009, "Real-time DNA sequencing from single polymerase molecules", *Science,* vol. 323, no. 5910, pp. 133-138.

81 Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. 1998, "Cluster analysis and display of genome-wide expression patterns", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 95, no. 25, pp. 14863-14868.

82 Elowitz, M.B., Levine, A.J., Siggia, E.D. & Swain, P.S. 2002, "Stochastic gene expression in a single cell", *Science,* vol. 297, no. 5584, pp. 1183-1186.

83 Emrich, S.J., Barbazuk, W.B., Li, L. & Schnable, P.S. 2007, "Gene discovery and annotation using LCM-454 transcriptome sequencing", *Genome research,* vol. 17, no. 1, pp. 69-73.

84 Eriksson, L., Antti, H., Gottfries, J., Holmes, E., Johansson, E., Lindgren, F., Long, I., Lundstedt, T., Trygg, J. & Wold, S. 2004, "Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm)", *Analytical and Bioanalytical Chemistry,* vol. 380, no. 3 SPEC.ISS., pp. 419-429.

85 Eriksson, L. 1999, *Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS),* Umeå : Umetrics, Umeå.

86 Esbensen, K. 2000, *Multivariate data analysis : in practice ; an introduction to multivariate data analysis and experimental design,* 5th edn, Oslo : CAMO ASA, Oslo.

87 Esseling, J.J., Lhuissier, F.G.P. & Emons, A.M.C. 2003, "Nod factor-induced root hair curling: Continuous polar growth towards the point of nod factor application", *Plant Physiology,* vol. 132, no. 4, pp. 1982-1988.

88 Ewing, B., Hillier, L., Wendl, M.C. & Green, P. 1998, "Base-calling of automated sequencer traces using phred. I. Accuracy assessment", *Genome research,* vol. 8, no. 3, pp. 175-185.

89 FAOSTAT 2011a, 05/17/2011-last update, *Crops* [Homepage of Food and agriculture organization of the united nations], [Online]. Available: http://faostat.fao.org/site/567/default.aspx#ancor [2011, 12/11].

90 FAOSTAT 2011b, 12/21/2011-last update, *Crops processed* [Homepage of Food and Agriculture Organization of the United Nations], [Online]. Available: http://faostat.fao.org/site/636/default.aspx#ancor [2012, 02/08/2012].

91 FAOSTAT 2010, 06/02/2011-last update, *Crops Primary Equivalent* [Homepage of Food and agriculture organization of the united nations], [Online]. Available: http://faostat.fao.org/site/609/default.aspx#ancor [2011, 12/11].

92 Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. 2006, "BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies", *Nucleic acids research,* vol. 34, no. 3.

93 Ferragina, P. & Manzini, G. 2000, "Opportunistic data structures with applications", *Annual Symposium on Foundations of Computer Science - Proceedings*, pp. 390.

94 Fettke, J., Albrecht, T., Hejazi, M., Mahlow, S., Nakamura, Y. & Steup, M. 2010, "Glucose 1-phosphate is efficiently taken up by potato (Solanum tuberosum) tuber parenchyma cells and converted to reserve starch granules", *The New phytologist,* vol. 185, no. 3, pp. 663-675.

95 Fischer, H.M. 1994, "Genetic regulation of nitrogen fixation in rhizobia.", *Microbiological reviews,* vol. 58, no. 3, pp. 352-386.

96 Flicek, P. & Birney, E. 2009, "Sense from sequence reads: methods for alignment and assembly.", *Nature methods,* vol. 6, no. 11 Suppl, pp. S6-S12.

97 Fullwood, M.J., Wei, C.-., Liu, E.T. & Ruan, Y. 2009, "Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses", *Genome research,* vol. 19, no. 4, pp. 521-532.

98 Gage, D.J. 2004, "Infection and invasion of roots by symbiotic, nitrogen-fixing rhizobia during nodulation of temperate legumes", *Microbiology and Molecular Biology Reviews,* vol. 68, no. 2, pp. 280-300.

99 Garber, M., Grabherr, M.G., Guttman, M. & Trapnell, C. 2011, "Computational methods for transcriptome annotation and quantification using RNA-seq", *Nature Methods,* vol. 8, no. 6, pp. 469-477.

100 Geurts, R., Fedorova, E. & Bisseling, T. 2005, "Nod factor signaling genes and their function in the early stages of Rhizobium infection", *Current opinion in plant biology,* vol. 8, no. 4, pp. 346-352.

101 Gilles, A., Meglecz, E., Pech, N., Ferreira, S., Malausa, T. & Martin, J. 2011, "Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing", *BMC Genomics,* vol. 12, no. 1, pp. 245.

102 Gonzalez-Rizzo, S., Crespi, M. & Frugier, F. 2006, "The Medicago truncatula CRE1 cytokinin receptor regulates lateral root development and early symbiotic interaction with Sinorhizobium meliloti", *Plant Cell,* vol. 18, no. 10, pp. 2680-2693.

103 Gopal, J. & Khurana, P.S.M. (eds) 2006, *Handbook of potato production, improvement, and postharvest management*, 1st. edn, The Hawirth Press Inc., Binghanton.

104 Graham, P.H. & Vance, C.P. 2003, "Legumes: Importance and constraints to greater use", *Plant Physiology,* vol. 131, no. 3, pp. 872-877.

105 Griffith, M., Griffith, O.L., Mwenifumbo, J., Goya, R., Morrissy, A.S., Morin, R.D., Corbett, R., Tang, M.J., Hou, Y.-., Pugh, T.J., Robertson, G., Chittaranjan, S., Ally, A., Asano, J.K., Chan, S.Y., Li, H.I., McDonald,

H., Teague, K., Zhao, Y., Zeng, T., Delaney, A., Hirst, M., Morin, G.B., Jones, S.J.M., Tai, I.T. & Marra, M.A. 2010, "Alternative expression analysis by RNA sequencing", *Nature Methods,* vol. 7, no. 10, pp. 843-847.

106 Groth, M., Takeda, N., Perry, J., Uchid, H., Dräxl, S., Brachmann, A., Sato, S., Tabata, S., Kawaguchi, M., Wang, T.L. & Parniskea, M. 2010, "NENA, a Lotus japonicus homolog of Sec13, is required for rhizodermal infection by arbuscular mycorrhiza fungi and rhizobia but dispensable for cortical endosymbiotic development", *Plant Cell,* vol. 22, no. 7, pp. 2509-2526.

107 Grunewald, W., Van Noorden, G., Van Lsterdael, G., Beeckman, T., Gheysen, G. & Mathesius, U. 2009, "Manipulation of auxin transport in plant roots during rhizobium symbiosis and nematode parasitism", *Plant Cell,* vol. 21, no. 9, pp. 2553-2562.

108 Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., Rinn, J.L., Lander, E.S. & Regev, A. 2010, "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs", *Nature biotechnology,* vol. 28, no. 5, pp. 503-510.

109 Hakoyama, T., Niimi, K., Yamamoto, T., Isobe, S., Sato, S., Nakamura, Y., Tabata, S., Kumagai, H., Umehara, Y., Brossuleit, K., Petersen, T.R., Sandal, N., Stougaard, J., Udvardi, M.K., Tamaoki, M., Kawaguchi, M., Kouchi, H. & Suganuma, N. 2012, "The integral membrane protein SEN1 is required for symbiotic nitrogen fixation in lotus japonicus nodules", *Plant and Cell Physiology,* vol. 53, no. 1, pp. 225-236.

110 Hakoyama, T., Niimi, K., Watanabe, H., Tabata, R., Matsubara, J., Sato, S., Nakamura, Y., Tabata, S., Jichun, L., Matsumoto, T., Tatsumi, K., Nomura, M., Tajima, S., Ishizaka, M., Yano, K., Imaizumi-Anraku, H., Kawaguchi, M., Kouchi, H. & Suganuma, N. 2009, "Host plant genome overcomes the lack of a bacterial gene for symbiotic nitrogen fixation", *Nature,* vol. 462, no. 7272, pp. 514-517.

111 Handberg, K. & Stougaard, J. 1992, "Lotus japonicus, an autogamous, diploid legume species for classical and molecular genetics", *Plant Journal,* vol. 2, no. 4, pp. 487-496.

112 Haney, C.H. & Long, S.R. 2010, "Plant flotillins are required for infection by nitrogen-fixing bacteria", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 107, no. 1, pp. 478-483.

113 Hansen, K.D., Brenner, S.E. & Dudoit, S. 2010, "Biases in Illumina transcriptome sequencing caused by random hexamer priming.", *Nucleic acids research,* vol. 38, no. 12.

114 Hardcastle, T.J. & Kelly, K.A. 2010, "BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data", *BMC Bioinformatics,* vol. 11.

115 Hardy, J. & Singleton, A. 2009, "Genomewide association studies and human disease", *New England Journal of Medicine,* vol. 360, no. 17, pp. 1759-1768.

116 Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., DiMeo, J., Efcavitch, J.W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S.R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H. & Xie, Z. 2008, "Single-molecule DNA sequencing of a viral genome", *Science,* vol. 320, no. 5872, pp. 106-109.

117 Harrison, J., Pou De Crescenzo, M.-, Sené, O. & Hirel, B. 2003, "Does lowering glutamine synthetase activity in nodules modify nitrogen metabolism and growth of Lotus japonicus?", *Plant Physiology,* vol. 133, no. 1, pp. 253-262.

118 Heckmann, A.B., Lombardo, F., Miwa, H., Perry, J.A., Bunnewell, S., Parniske, M., Wang, T.L. & Downie, J.A. 2006, "Lotus japonicus nodulation requires two GRAS domain regulators, one of which is functionally conserved in a non-legume", *Plant Physiology,* vol. 142, no. 4, pp. 1739-1750.

119 Heinz, B. 2010, 07/01-last update, *A Theoretical Understanding of 2 Base Color Codes and Its Application to Annotation, Error Detection, and Error Correction* [Homepage of Applied Biosystems], [Online]. Available: http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_058265.pdf [2011, 07/26].

120 Helicos BioSciences 2008, , *tSMS™ Performance* [Homepage of Helicos BioSciences Corporation], [Online]. Available: http://www.helicosbio.com/Technology/TrueSingleMoleculeSequencing/tSMStradePerformance/tabid/151/Default.aspx [2011, 07/29].

121 Helicos BioSciences Corporation 2008, , *True Direct DNA Measurement* [Homepage of Helicos BioSciences Corporation], [Online]. Available: http://www.helicosbio.com/Portals/0/Documents/Helicos%20tSMS%20Technology%20Primer.pdf [2011, 07/28].

122 Hendricks, W.A. & Robey, K.W. 1936, "The Sampling Distribution of the Coefficient of Variation", *The Annals of Mathematical Statistics,* vol. 7, no. 3, pp. pp. 129-132.

123 Hernandez, D., Frantois, P., Farinelli, L., ǂsteros, M. & Schrenzel, J. 2008, "De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer", *Genome research,* vol. 18, no. 5, pp. 802-809.

124 Herper, M. 2010, 23/02-last update, *A decade's perspective on DNA sequencing technology* [Homepage of Forbes.com LLC], [Online]. Available: http://www.forbes.com/2010/02/22/health-genome-illumina-gene-diagnosis-by-dna.html [2011, 26/07].

125 Hijmans, R.J. 2001, "Global distribution of the potato crop", *American Journal of Potato Research,* vol. 78, no. 6, pp. 403-412.

126 Hirel, B. & Lea, P.J. 2001, "Ammonium assimilation" in *Plant Nitrogen*, eds. P.J. Lea & J.F.M. Morof Gaudry, Springer-Verlag, Berlin, pp. 79-80-99.

127 Hirel, B., Phillipson, B., Murchie, E., Suzuki, A., Kunz, C., Ferrario, S., Limami, A., Chaillou, S., Deleens, E., Brugière, N., Chaumont-Bonnet, M., Foyer, C. & Morot-Gaudry, J. 1997, "Manipulating the pathway of ammonia assimilation in transgenic non-legumes and legumes", *Zeitschrift für Pflanzenernährung und Bodenkunde,* vol. 160, no. 2, pp. 283-290.

128 Høgslund, N., Radutoiu, S., Krusell, L., Voroshilova, V., Hannah, M.A., Goffard, N., Sanchez, D.H., Lippold, F., Ott, T., Sato, S., Tabata, S., Liboriussen, P., Lohmann, G.V., Schauser, L., Weiller, G.F., Udvardi, M.K. & Stougaard, J. 2009, "Dissection of symbiosis and organ development by integrated transcriptome analysis of Lotus japonicus mutant and wild-type plants", *PLoS ONE,* vol. 4, no. 8.

129 Hong, G.F. 1981, "A method for sequencing single-stranded cloned DNA in both directions", *Bioscience reports,* vol. 1, no. 3, pp. 243-252.

130 Horner, D.S., Pavesi, G., Castrignano, T., De Meo, P.D., Liuni, S., Sammeth, M., Picardi, E. & Pesole, G. 2010, "Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing", *Briefings in bioinformatics,* vol. 11, no. 2, pp. 181-197.

131 Housby, J.N. & Southern, E.M. 1998, "Fidelity of DNA ligation: A novel experimental approach based on the polymerisation of libraries of oligonucleotides", *Nucleic acids research,* vol. 26, no. 18, pp. 4259-4266.

132 Huamán, Z. & Spooner, D.M. 2002, "Reclassification of landrace populations of cultivated potatoes (Solanum sect. Petota)", *American Journal of Botany,* vol. 89, no. 6, pp. 947-965.

133 Huang, D.W., Sherman, B.T. & Lempicki, R.A. 2009, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources", *Nature Protocols,* vol. 4, no. 1, pp. 44-57.

134 Huang, X.C., Quesada, M.A. & Mathies, R.A. 1992, "DNA sequencing using capillary array electrophoresis", *Analytical Chemistry,* vol. 64, no. 18, pp. 2149-2154.

135 Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L. & Welch, D.M. 2007, "Accuracy and quality of massively parallel DNA pyrosequencing", *Genome biology,* vol. 8, no. 7.

136 Hutchison, C.A.,3rd 2007, "DNA sequencing: bench to bedside and beyond", *Nucleic acids research,* vol. 35, no. 18, pp. 6227-6237.

137 Hyman, E.D. 1988, "A new method of sequencing DNA", *Analytical Biochemistry,* vol. 174, no. 2, pp. 423-436.

138 Illumina 2009, 10/13/2099-last update, *De novo Assembly Using Illumina Reads* [Homepage of Illumina Inc.], [Online]. Available: http://www.illumina.com/Documents/products/technotes/technote_denovo_assembly_ecoli.pdf [2012, 09/01].

139 Illumina, I. 2011a, , *History of Solexa Sequencing* [Homepage of Illumina, Inc.], [Online]. Available: http://www.illumina.com/technology/solexa_technology.ilmn [2011, 07/21].

140 Illumina, I. 2011b, , *Paired-end Sequencing Assay* [Homepage of Illumina, Inc.], [Online]. Available: http://www.illumina.com/technology/paired_end_sequencing_assay.ilmn [2011, 07/26].

141 Illumina, I. 2009, , *product documentation* [Homepage of Illumina, Inc.], [Online]. Available: http://www.illumina.com/pagesnrn.ilmn?ID=275 [2009, 10/12].

142 Ion Torrent Systems 2011, , *Technology: How does it work?* [Homepage of Ion Torrent Systems, Inc.], [Online]. Available: http://www.iontorrent.com/technology-how-does-it-work-more/ [2011, 07/31].

143 Ishii, M., Hashimoto, S.-., Tsutsumi, S., Wada, Y., Matsushima, K., Kodama, T. & Aburatani, H. 2000, "Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis", *Genomics,* vol. 68, no. 2, pp. 136-143.

144 Jeck, W.R., Reinhardt, J.A., Baltrus, D.A., Hickenbotham, M.T., Magrini, V., Mardis, E.R., Dangl, J.L. & Jones, C.D. 2007, "Extending assembly of short DNA sequences to handle error", *Bioinformatics,* vol. 23, no. 21, pp. 2942-2944.

145 Jia, P., Wang, L., Meltzer, H.Y. & Zhao, Z. 2010, "Common variants conferring risk of schizophrenia: A pathway analysis of GWAS data", *Schizophrenia research,* vol. 122, no. 1-3, pp. 38-42.

146 Jiang, H. & Wong, W.H. 2009, "Statistical inferences for isoform expression in RNA-Seq", *Bioinformatics,* vol. 25, no. 8, pp. 1026-1032.

147 John Innes Centre 2011, 19/12/2012-last update, *Legumes give nitrogen-supplying bacteria special access pass* [Homepage of John Innes Centre], [Online]. Available: http://news.jic.ac.uk/2011/12/legumes-give-bacteria-access/ [2012, 02/12/2012].

148 Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. 2007, "Genome-wide mapping of in vivo protein-DNA interactions", *Science,* vol. 316, no. 5830, pp. 1497-1502.

149 Jolliffe, I.T. 2002, *Principal Component Analysis,* 2nd edn, Springer, Secaucus, NJ, USA.

150 Kal, A.J., Van Zonneveld, A.J., Benes, V., Van Den Berg, M., Koerkamp, M.G., Albermann, K., Strack, N., Ruijter, J.M., Richter, A., Dujon, B., Ansorge, W. & Tabak, H.F. 1999, "Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources", *Molecular biology of the cell,* vol. 10, no. 6, pp. 1859-1872.

151 Kaló, P., Gleason, C., Edwards, A., Marsh, J., Mitra, R.M., Hirsch, S., Jakab, J., Sims, S., Long, S.R., Rogers, J., Kiss, G.B., Downie, J.A. & Oldroyd, G.E.D. 2005, "Nodulation signaling in legumes requires NSP2, a member of the GRAS family of transcriptional regulators", *Science,* vol. 308, no. 5729, pp. 1786-1789.

152 Kanamori, N., Madsen, L.H., Radutoiu, S., Frantescu, M., Quistgaard, E.M.H., Miwa, H., Downie, J.A., James, E.K., Felle, H.H., Haaning, L.L., Jensen, T.H., Sato, S., Nakamura, Y., Tabata, S., Sandal, N. & Stougaard, J. 2006, "A nucleoporin is required for induction of Ca2+ spiking in legume nodule development and essential for rhizobial and fungal symbiosis", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 103, no. 2, pp. 359-364.

153 Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. 2009, "KEGG for representation and analysis of molecular networks involving diseases and drugs", *Nucleic acids research,* vol. 38, no. SUPPL.1, pp. D355-D360.

154 Kanehisa, M. & Goto, S. 2000, "KEGG: Kyoto Encyclopedia of Genes and Genomes", *Nucleic acids research,* vol. 28, no. 1, pp. 27-30.

155 Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. 2010, "Analysis and design of RNA sequencing experiments for identifying isoform regulation", *Nature Methods,* vol. 7, no. 12, pp. 1009-1015.

156 Kawaguchi, M., Imaizumi-Anraku, H., Koiwa, H., Niwa, S., Ikuta, A., Syono, K. & Akao, S. 2002, "Root, root hair, and symbiotic mutants of the model legume Lotus japonicus", *Molecular Plant-Microbe Interactions,* vol. 15, no. 1, pp. 17-26.

157 Kawaguchi, M. 2000, "Lotus japonicus 'Miyakojima' MG-20: An early-flowering accession suitable for indoor handling", *Journal of Plant Research,* vol. 113, no. 1112, pp. 507-509.

158 Kendziorski, C.M., Zhang, Y., Lan, H. & Attie, A.D. 2003, "The efficiency of pooling mRNA in microarray experiments.", *Biostatistics (Oxford, England),* vol. 4, no. 3, pp. 465-477.

159 Khatri, P. & Draghici, S. 2005, "Ontological analysis of gene expression data: Current tools, limitations, and open problems", *Bioinformatics,* vol. 21, no. 18, pp. 3587-3595.

160 King, O.D., Foulger, R.E., Dwight, S.S., White, J.V. & Roth, F.P. 2003, "Predicting gene function from patterns of annotation", *Genome research,* vol. 13, no. 5, pp. 896-904.

161 Kioka, N., Ueda, K. & Amachi, T. 2002, "Vinexin, CAP/ponsin, ArgBP2: A novel adaptor protein family regulating cytoskeletal organization and signal transduction", *Cell structure and function,* vol. 27, no. 1, pp. 1-7.

162 Kohonen, T. 1982, "Self-organized formation of topologically correct feature maps", *Biological cybernetics,* vol. 43, no. 1, pp. 59-69.

163 Krusell, L., Krause, K., Ott, T., Desbrosses, G., Krämer, U., Sato, S., Nakamura, Y., Tabata, S., James, E.K., Sandal, N., Stougaard, J., Kawaguchi, M., Miyamoto, A., Suganuma, N. & Udvardi, M.K. 2005, "The sulfate transporter SST1 is crucial for symbiotic nitrogen fixation in Lotus japonicus root nodules", *Plant Cell,* vol. 17, no. 5, pp. 1625-1636.

164 Kuznetsov, V.A., Knott, G.D. & Bonner, R.F. 2002, "General statistics of stochastic process of gene expression in eukaryotic cells", *Genetics,* vol. 161, no. 3, pp. 1321-1332.

165 Lal, A., Lash, A.E., Altschul, S.F., Velculescu, V., Zhang, L., McLendon, R.E., Marra, M.A., Prange, C., Morin, P.J., Polyak, K., Papadopoulos, N., Vogelstein, B., Kinzler, K.W., Strausberg, R.L. & Riggins, G.J. 1999, "A public database for gene expression in human cancers", *Cancer research,* vol. 59, no. 21, pp. 5403-5407.

166 Lamport, D.T.A., Kieliszewski, M.J., Chen, Y. & Cannon, M.C. 2011, "Role of the Extensin Superfamily in Primary Cell Wall Architecture", *Plant Physiology,* vol. 156, no. 1, pp. 11-19.

167 Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coul-

son, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la, B.M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de, J.P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S. & Chen, Y.J. 2001, "Initial sequencing and analysis of the human genome", *Nature,* vol. 409, no. 6822, pp. 860-921.

168 Langhorst, M.F., Jaeger, F.A., Mueller, S., Sven Hartmann, L., Luxenhofer, G. & Stuermer, C.A.O. 2008, "Reggies/flotillins regulate cytoskeletal remodeling during neuronal differentiation via CAP/ponsin and Rho GTPases", *European journal of cell biology,* vol. 87, no. 12, pp. 921-931.

169 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. 2009, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome", *Genome biology,* vol. 10, no. 3.

170 Lê Cao, K.-., Boitard, S. & Besse, P. 2011, "Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems", *BMC Bioinformatics,* vol. 12.

171 Lea, P.J. & Miflin, B.J. 1980, "Transport and metabolism of asparagine and other nitrogen compounds within the plant.
" in *The biochemistry of plants*, eds. P.K. Stumpf & E.E. Conn, Academic Press, New York, pp. 169-570-604.

172 Lefebvre, B., Timmers, T., Mbengue, M., Moreau, S., Hervé, C., Tóth, K., Bittencourt-Silvestre, J., Klaus, D., Deslandes, L., Godiard, L., Murray, J.D., Udvardi, M.K., Raffaele, S., Mongrand, S., Cullimore, J., Gamas, P., Niebel, A. & Ott, T. 2010, "A remorin protein interacts with symbiotic receptors and regulates bacterial infection", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 107, no. 5, pp. 2343-2348.

173 Levene, H.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G. & Webb, W.W. 2003, "Zero-mode waveguides for single-molecule analysis at high concentrations", *Science,* vol. 299, no. 5607, pp. 682-686.

174 Lévy, J., Bres, C., Geurts, R., Chalhoub, B., Kulikova, O., Duc, G., Journet, E.-., Ané, J.-., Lauber, E., Bisseling, T., Dénarié, J., Rosenberg, C. & Debellé, F. 2004, "A Putative Ca2+ and Calmodulin-Dependent Protein Kinase Required for Bacterial and Fungal Symbioses", *Science,* vol. 303, no. 5662, pp. 1361-1364.

175 Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., Lin, Y., MacDonald, J.R., Pang, A.W.C., Shago, M., Stockwell, T.B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S.A., Busam, D.A., Beeson, K.Y., McIntosh, T.C., Remington, K.A., Abril, J.F., Gill, J., Borman, J., Rogers, Y.-., Frazier, M.E., Scherer, S.W., Strausberg, R.L. & Venter, J.C. 2007, "The diploid genome sequence of an individual human", *PLoS Biology,* vol. 5, no. 10, pp. 2113-2144.

176 Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. & Dewey, C.N. 2009, "RNA-Seq gene expression estimation with read mapping uncertainty", *Bioinformatics,* vol. 26, no. 4, pp. 493-500.

177 Li, H. & Homer, N. 2010, "A survey of sequence alignment algorithms for next-generation sequencing", *Briefings in Bioinformatics,* vol. 11, no. 5, pp. 473-483.

178 Li, H. & Durbin, R. 2009, "Fast and accurate short read alignment with Burrows-Wheeler transform", *Bioinformatics,* vol. 25, no. 14, pp. 1754-1760.

179 Li, H., Ruan, J. & Durbin, R. 2008, "Mapping short DNA sequencing reads and calling variants using mapping quality scores", *Genome research,* vol. 18, no. 11, pp. 1851-1858.

180 Li, J., Jiang, H. & Wong, W.H. 2010, "Modeling non-uniformity in short-read rates in RNA-Seq data", *Genome biology,* vol. 11, no. 5.

181 Li, L., Stoeckert Jr., C.J. & Roos, D.S. 2003, "OrthoMCL: Identification of ortholog groups for eukaryotic genomes", *Genome research,* vol. 13, no. 9, pp. 2178-2189.

182 Li, P.H. 1985, *Potato physiology,* Academic Press.

183 Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O.A., Leung, F.C.-., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C.C., Lam, T.T.-., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M.W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T.-., Yiu, S.-., Liu, S., Zhang, H., Li, D., Huang, Y., Wang, X., Yang, G., Jiang, Z., Wang, J., Qin, N., Li, L., Li, J., Bolund, L., Kristiansen, K., Wong, G.K.-., Olson, M., Zhang, X., Li, S., Yang, H., Wang, J. & Wang, J. 2010a, "The sequence and de novo assembly of the giant panda genome", *Nature,* vol. 463, no. 7279, pp. 311-317.

184 Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., Qian, W., Ren, Y., Tian, G., Li, J., Zhou, G., Zhu, X., Wu, H., Qin, J., Jin, X., Li, D., Cao, H., Hu, X., Blanche, H., Cann, H., Zhang, X., Li, S., Bolund, L., Kristiansen, K., Yang, H., Wang, J. & Wang, J. 2010b, "Building the sequence map of the human pan-genome", *Nature biotechnology,* vol. 28, no. 1, pp. 57-62.

185 Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J. & Wang, J. 2010c, "De novo assembly of human genomes with massively parallel short read sequencing", *Genome research,* vol. 20, no. 2, pp. 265-272.

186 Li, R., Yu, C., Li, Y., Lam, T.-., Yiu, S.-., Kristiansen, K. & Wang, J. 2009, "SOAP2: An improved ultrafast tool for short read alignment", *Bioinformatics,* vol. 25, no. 15, pp. 1966-1967.

187 Li, R., Li, Y., Kristiansen, K. & Wang, J. 2008, "SOAP: Short oligonucleotide alignment program", *Bioinformatics,* vol. 24, no. 5, pp. 713-714.

188 Liang, P. & Pardee, A.B. 1992, "Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction", *Science,* vol. 257, no. 5072, pp. 967-971.

189 Lievens, S., Goormachtig, S., Herman, S. & Holsters, M. 2002, "Patterns of pectin methylesterase transcripts in developing stem nodules of Sesbania rostrata", *Molecular Plant-Microbe Interactions,* vol. 15, no. 2, pp. 164-168.

190 Life Technologies 2011, , *SOLiD™ System accuracy with the Exact Call Chemistry module* [Homepage of Life Technologies Corporation], [Online]. Available: http://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_091372.pdf [07/28, 2011].

191 Life Technologies 2008, 03/01-last update, *SOLiD™ Data Format and File definitions Guide* [Homepage of Life Technologies], [Online]. Available: http://marketing.appliedbiosystems.com/mk/submit/SOLID_KNOWLEDGE_RD?_JS=T&rd=dm [2011, 07/26].

192 Limpens, E., Franken, C., Smit, P., Willemse, J., Bisseling, T. & Geurts, R. 2003, "LysM Domain Receptor Kinases Regulating Rhizobial Nod Factor-Induced Infection", *Science,* vol. 302, no. 5645, pp. 630-633.

193 Lin, Y., Li, J., Shen, H., Zhang, L., Papasian, C.J. & Deng, H.-. 2011, "Comparative studies of de novo assembly tools for next-generation sequencing technologies", *Bioinformatics,* vol. 27, no. 15, pp. 2031-2037.

194 Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. & Ecker, J.R. 2008, "Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis", *Cell,* vol. 133, no. 3, pp. 523-536.

195 LMU 2012, 02/12/2012-last update, *Lotus japonicus* [Homepage of Institute of Genetics of the University of Munich], [Online]. Available: http://www.genetik.bio.lmu.de/research/parniske/models/lotusjaponicus/index.html [2012, 02/12/2012].

196 Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. & Brown, E.L. 1996, "Expression monitoring by hybridization to high-density oligonucleotide arrays", *Nature biotechnology,* vol. 14, no. 13, pp. 1675-1680.

197 Long, S.R. 2001, "Genes and signals in the rhizobium-legumes symbiosis", *Plant Physiology,* vol. 125, no. 1, pp. 69-72.

198 Lu, J., Tomfohr, J.K. & Kepler, T.B. 2005, "Identifying differential expression in multiple SAGE libraries: An overdispersed log-linear model approach", *BMC Bioinformatics,* vol. 6.

199 Lu, J., Lal, A., Merriman, B., Nelson, S. & Riggins, G. 2004, "A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips", *Genomics,* vol. 84, no. 4, pp. 631-636.

200 Lunshof, J.E., Bobe, J., Aach, J., Angrist, M., Thakuria, J.V., Vorhaus, D.B., Hoehe, M.R. & Church, G.M. 2010, "Personal genomes in progress: from the human genome project to the personal genome project.", *Dialogues in clinical neuroscience,* vol. 12, no. 1, pp. 47-60.

201 MacQueen, J.B. 1967, "Some Methods for Classification and Analysis of Multivariate Observations", *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, eds. L.M.L. Cam & J. Neyman, University of California Press, , pp. 281.

202 Madden, S.L., Galella, E.A., Zhu, J., Bertelsen, A.H. & Beaudry, G.A. 1997, "SAGE transcript profiles for p53-dependent growth regulation", *Oncogene,* vol. 15, no. 9, pp. 1079-1085.

203 Madsen, E.B., Madsen, L.H., Radutoiu, S., Olbryt, M., Rakwalska, M., Szczyglowski, K., Sato, S., Kaneko, T., Tabata, S., Sandal, N. & Stougaard, J. 2003, "A receptor kinase gene of the LysM type is involved in legume perception of rhizobial signals", *Nature,* vol. 425, no. 6958, pp. 637-640.

204 Maere, S., Heymans, K. & Kuiper, M. 2005, "BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks", *Bioinformatics,* vol. 21, no. 16, pp. 3448-3449.

205 Maglott, D.R., Katz, K.S., Sicotte, H. & Pruitt, K.D. 2000, "NCBI's LocusLink and RefSeq", *Nucleic acids research,* vol. 28, no. 1, pp. 126-128.

206 Man, M.Z., Wang, X. & Wang, Y. 2000, "POWER_SAGE: Comparing statistical tests for SAGE experiments", *Bioinformatics,* vol. 16, no. 11, pp. 953-959.

207 Mardis, E.R. 2008, "The impact of next-generation sequencing technology on genetics", *Trends in genetics : TIG,* vol. 24, no. 3, pp. 133-141.

208 Margulies, E.H., Kardia, S.L. & Innis, J.W. 2001, "Identification and prevention of a GC content bias in SAGE libraries.", *Nucleic acids research,* vol. 29, no. 12, pp. E60-60.

209 Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.-., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. & Rothberg, J.M. 2005, "Genome sequencing in microfabricated high-density picolitre reactors", *Nature,* vol. 437, no. 7057, pp. 376-380.

210 Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. 2008, "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays", *Genome research,* vol. 18, no. 9, pp. 1509-1517.

211 Marquez, A.J. 2005, *Lotus japonicus Handbook,* Springer, Dordrecht.

212 Martin, E. & Martin, E. 2008, *A dictionary of biology,* .

213 Marziali, A. & Akeson, M. 2001, "New DNA sequencing methods", *Annual Review of Biomedical Engineering,* vol. 3, pp. 195-223.

214 Matsumura, H., Yoshida, K., Luo, S., Kimura, E., Fujibe, T., Albertyn, Z., Barrero, R.A., Krüger, D.H., Kahl, G., Schroth, G.P. & Terauchi, R. 2010, "High-throughput superSAGE for digital gene expression analysis of multiple samples using next generation sequencing", *PLoS ONE,* vol. 5, no. 8.

215 Matsumura, H., Ito, A., Saitoh, H., Winter, P., Kahl, G., Reuter, M., Krüger, D.H. & Terauchi, R. 2005, "SuperSAGE", *Cellular microbiology,* vol. 7, no. 1, pp. 11-18.

216 Maxam, A.M. & Gilbert, W. 1977, "A new method for sequencing DNA", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 74, no. 2, pp. 560-564.

217 Mbengue, M., Camut, S., de Carvalho-Niebel, F., Deslandes, L., Solène, F., Klaus-Heisen, D., Moreau, S., Rivas, S., Timmers, T., Hervé, C., Cullimore, J. & Lefebvre, B. 2010, "The medicago truncatula E3 ubiquitin ligase PUB1 interacts with the LYK3 symbiotic receptor and negatively regulates infection and nodulation", *Plant Cell,* vol. 22, no. 10, pp. 3474-3488.

218 McKernan, K., Blanchard, A., Kotler, L. & Costa, G. 2008, *Reagents, methods, and libraries for bead-based sequencing*, 435/6.12 edn, C12Q 1/68 20060101 C12Q001/68, USA.

219 MedicineNet.com 2003, 05/04-last update, *Definition of Nutrigenomics* [Homepage of MedicineNet, Inc], [Online]. Available: http://www.medterms.com/script/main/art.asp?articlekey=23241 [2011, 07/14].

220 Michiels, E.M.C., Oussoren, E., Van Groenigen, M., Pauws, E., Bossuyt, P.M.M., Voûte, P.A. & Baas, F. 1999, "Genes differentially expressed in medulloblastoma and fetal brain", *Physiological Genomics,* vol. 1999, no. 1, pp. 83-91.

221 Middleton, P.H., Jakab, J., Penmetsa, R.V., Starker, C.G., Doll, J., Kalo, P., Prabhu, R., Marsh, J.F., Mitra, R.M., Kereszt, A., Dudas, B., VandenBosch, K., Long, S.R., Cook, D.R., Kiss, G.B. & Oldroyd, G.E. 2007,

"An ERF transcription factor in Medicago truncatula that is essential for Nod factor signal transduction", *The Plant Cell,* vol. 19, no. 4, pp. 1221-1234.

222 Miller, J.R., Koren, S. & Sutton, G. 2010, "Assembly algorithms for next-generation sequencing data", *Genomics,* vol. 95, no. 6, pp. 315-327.

223 Mitra, R.M., Gleason, C.A., Edwards, A., Hadfield, J., Downie, J.A., Oldroyd, G.E.D. & Long, S.R. 2004, "A Ca2+/calmodulin-dependent protein kinase required for symbiotic nodule development: Gene identification by transcript-based cloning", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 101, no. 13, pp. 4701-4705.

224 Moore, G.E. 1965, "Cramming more components onto integrated circuits", vol. 38, no. 8, pp. 114-115,116,117.

225 Morrissy, A.S., Morin, R.D., Delaney, A., Zeng, T., McDonald, H., Jones, S., Zhao, Y., Hirst, M. & Marra, M.A. 2009, "Next-generation tag sequencing for cancer gene expression profiling", *Genome research,* vol. 19, no. 10, pp. 1825-1835.

226 Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. 2008, "Mapping and quantifying mammalian transcriptomes by RNA-Seq", *Nature Methods,* vol. 5, no. 7, pp. 621-628.

227 Müller, B. & Sheen, J. 2008, "Cytokinin and auxin interaction in root stem-cell specification during early embryogenesis", *Nature,* vol. 453, no. 7198, pp. 1094-1097.

228 Müller, H., Neumaier, M. & Hoffmann, G. 2008, "Gene expression studies with microarrays and SAGE: Biological and analytical principles", *LaboratoriumsMedizin,* vol. 32, no. 5, pp. 308-316.

229 Muñoz, J.A., Coronado, C., Pérez-Hormaeche, J., Kondorosi, A., Ratet, P. & Palomares, A.J. 1998, "MsPG3, a Medicago sativa polygalacturonase gene expressed during the alfalfa-Rhizobium meliloti interaction", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 95, no. 16, pp. 9687-9692.

230 Murray, J.D., Karas, B.J., Sato, S., Tabata, S., Amyot, L. & Szczyglowski, K. 2007, "A cytokinin perception mutant colonized by Rhizobium in the absence of nodule organogenesis", *Science,* vol. 315, no. 5808, pp. 101-104.

231 Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H.J., Remington, K.A., Anson, E.L., Bolanos, R.A., Chou, H.-., Jordan, C.M., Halpern, A.L., Lonardi, S., Beasley, E.M., Brandon, R.C., Chen, L., Dunn, P.J., Lai, Z., Liang, Y., Nusskern, D.R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G.M., Adams, M.D. & Venter, J.C. 2000, "A whole-genome assembly of Drosophila", *Science,* vol. 287, no. 5461, pp. 2196-2204.

232 Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. & Snyder, M. 2008, "The transcriptional landscape of the yeast genome defined by RNA sequencing", *Science,* vol. 320, no. 5881, pp. 1344-1349.

233 National Academy of Sciences 2011, 09/08-2011-last update, *Most-Cited Articles as of September 1, 2011 -- updated monthly* [Homepage of National Academy of Sciences], [Online]. Available: http://www.pnas.org/reports/most-cited [2011, 10/10].

234 National Human Genome Research Institute 2011, 02/22-last update, Genome Technology Program [Homepage of National Human Genome Research Institute], [Online]. Available: http://www.genome.gov/10000368 [2011, 07/28].

235 Nature Genetics Editors (ed) 2005, *The Chipping Forecast III*, Nature Publishing Group.

236 Nature Methods 2008, "Method of the year", *Nature Methods,* vol. 5, no. 1, pp. 1.

237 Ng, P., Wei, C.L. & Ruan, Y. 2007, "Paired-end diTagging for transcriptome and genome analysis", *Current protocols in molecular biology / edited by Frederick M.Ausubel ...[et al.],* vol. Chapter 21, pp. Unit 21.12.

238 Niedringhaus, T.P., Milanova, D., Kerby, M.B., Snyder, M.P. & Barron, A.E. 2011, "Landscape of next-generation sequencing technologies", *Analytical Chemistry,* vol. 83, no. 12, pp. 4327-4341.

239 Nielsen, K.L. 2008, "DeepSAGE: higher sensitivity and multiplexing of samples using a simpler experimental protocol.", *Methods in molecular biology (Clifton, N.J.),* vol. 387, pp. 81-94.

240 Nielsen, K.L., Hogh, A.L. & Emmersen, J. 2006, "DeepSAGE--digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples", *Nucleic Acids Res.,* vol. 34, no. 19, pp. e133.

241 Novak, J.P., Sladek, R. & Hudson, T.J. 2001, "Characterization of variability in large-scale gene expression data: Implications for study design", *Genomics,* vol. 79, no. 1, pp. 104-113.

242 Novelli, G., Predazzi, I.M., Mango, R., Romeo, F. & Mehta, J.L. 2010, "Role of genomics in cardiovascular medicine", *World journal of cardiology,* vol. 2, no. 12, pp. 428-436.

243 Nyren, P. 2007, "The history of pyrosequencing", *Methods in molecular biology (Clifton, N.J.),* vol. 373, pp. 1-14.

244 Nyren, P. 1987, "Enzymatic method for continuous monitoring of DNA polymerase activity", *Analytical Biochemistry,* vol. 167, no. 2, pp. 235-238.

245 Nyren, P. & Lundin, A. 1985, "Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis", *Analytical Biochemistry,* vol. 151, no. 2, pp. 504-509.

246 O'Dushlaine, C., Kenny, E., Heron, E., Donohoe, G., Gill, M., Morris, D. & Corvin, A. 2011, "Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility", *Molecular psychiatry,* vol. 16, no. 3, pp. 286-292.

247 O'Dushlaine, C., Kenny, E., Heron, E.A., Segurado, R., Gill, M., Morris, D.W. & Corvin, A. 2009, "The SNP ratio test: Pathway analysis of genome-wide association datasets", *Bioinformatics,* vol. 25, no. 20, pp. 2762-2763.

248 Okoniewski, M.J. & Miller, C.J. 2006, "Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations", *BMC Bioinformatics,* vol. 7.

249 Oldroyd, G.E.D., Murray, J.D., Poole, P.S. & Downie, J.A. 2011, *The rules of engagement in the legume-rhizobial symbiosis*.

250 Oldroyd, G.E.D. & Downie, J.A. 2008, *Coordinating nodule morphogenesis with rhizobial infection in legumes*.

251 Oldroyd, G.E.D. 2007, "Nodules and hormones", *Science,* vol. 315, no. 5808, pp. 52-53.

252 Osborne, R. & Slatter, A. 2011, *METHODS AND COMPOSITIONS FOR POLYNUCLEOTIDE LIBRARY PRODUCTION, IMMORTALIZATION AND REGION OF INTEREST EXTRACTIO* , 506/26 edn, C40B 50/06 20060101 C40B050/06, US.

253 Ott, T., Sullivan, J., James, E.K., Flemetakis, E., Günther, C., Gibon, Y., Ronson, C. & Udvardi, M. 2009, "Absence of symbiotic leghemoglobins alters bacteroid and plant cell differentiation during development of lotus japonicus root nodules", *Molecular Plant-Microbe Interactions,* vol. 22, no. 7, pp. 800-808.

254 Ott, T., Van Dongen, J.T., Günther, C., Krusell, L., Desbrosses, G., Vigeolas, H., Bock, V., Czechowski, T., Geigenberger, P. & Udvardi, M.K. 2005, "Symbiotic leghemoglobins are crucial for nitrogen fixation in legume root nodules but not for general plant growth and development", *Current Biology,* vol. 15, no. 6, pp. 531-535.

255 Ozsolak, F. & Milos, P.M. 2011a, "RNA sequencing: Advances, challenges and opportunities", *Nature Reviews Genetics,* vol. 12, no. 2, pp. 87-98.

256 Ozsolak, F. & Milos, P.M. 2011b, "Transcriptome profiling using single-molecule direct RNA sequencing.", *Methods in molecular biology (Clifton, N.J.),* vol. 733, pp. 51-61.

257 Ozsolak, F., Kapranov, P., Foissac, S., Kim, S.W., Fishilevich, E., Monaghan, A.P., John, B. & Milos, P.M. 2010, "Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation", *Cell,* vol. 143, no. 6, pp. 1018-1029.

258 Pacific Biosciences 2011, 04/27-last update, *Pacific Biosciences Begins Shipments of Commercial PacBio RS Systems* [Homepage of Thomson Reuters], [Online]. Available: http://www.reuters.com/article/2011/04/27/idUS255979+27-Apr-2011+BW20110427 [2011, 07/31].

259 Pacios-Bras, C., Schlaman, H.R.M., Boot, K., Admiraal, P., Langerak, J.M., Stougaard, J. & Spaink, H.P. 2003, "Auxin distribution in Lotus japonicus during root nodule development", *Plant Molecular Biology,* vol. 52, no. 6, pp. 1169-1180.

260 Pareek, C.S., Smoczynski, R. & Tretyn, A. 2011, "Sequencing technologies and genome sequencing", *Journal of Applied Genetics,* .

261 Patriarca, E.J., Tatè, R. & Iaccarino, M. 2002, "Key role of bacterial NH4 + metabolism in rhizobium-plant symbiosis", *Microbiology and Molecular Biology Reviews,* vol. 66, no. 2, pp. 203-222.

262 Pearson, K. 1895, "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material", *Philosophical Transactions of the Royal Society of London.(A.),* vol. 186, pp. 343-414.

263 Pepke, S., Wold, B. & Mortazavi, A. 2009, "Computation for ChIP-seq and RNA-seq studies.", *Nature methods,* vol. 6, no. 11 Suppl, pp. S22-32.

264 Pesole, G., Mignone, F., Gissi, C., Grillo, G., Licciulli, F. & Liuni, S. 2001, "Structural and functional features of eukaryotic mRNA untranslated regions", *Gene,* vol. 276, no. 1-2, pp. 73-81.

265 Pesole, G., Grillo, G., Larizza, A. & Liuni, S. 2000, "The untranslated regions of eukaryotic mRNAs: structure, function, evolution and bioinformatic tools for their analysis.", *Briefings in bioinformatics,* vol. 1, no. 3, pp. 236-249.

266 Petersen, A.H. 2008, *Transcriptome analysis of potato tuber lifecycle*, Department of Biotechnology, Chemistry and Environmental Engineering, Aalborg University.

267 Petrášek, J. & Schwarzerová, K. 2009, "Actin and microtubule cytoskeleton interactions", *Current opinion in plant biology,* vol. 12, no. 6, pp. 728-734.

268 Pevzner, P.A., Tang, H. & Waterman, M.S. 2001, "An Eulerian path approach to DNA fragment assembly", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 98, no. 17, pp. 9748-9753.

269 PGSC 2009, 09/23/2009-last update, *Potato Genome Sequence Released by International Group of Scientists* [Homepage of The Potato Genome Sequencing Consortium], [Online]. Available: http://www.potatogenome.net/images/2/2e/PGSC_Press_Release_0909.pdf [2011, 12/11].

270 Plet, J., Wasson, A., Ariel, F., Le Signor, C., Baker, D., Mathesius, U., Crespi, M. & Frugier, F. 2011, "MtCRE1-dependent cytokinin signaling integrates bacterial and plant cues to coordinate symbiotic nodule organogenesis in Medicago truncatula", *Plant Journal,* vol. 65, no. 4, pp. 622-633.

271 Pop, M. 2009, "Genome assembly reborn: Recent computational challenges", *Briefings in Bioinformatics,* vol. 10, no. 4, pp. 354-366.

272 Pop, M. & Salzberg, S.L. 2008, "Bioinformatics challenges of new sequencing technology", *Trends in Genetics,* vol. 24, no. 3, pp. 142-149.

273 Pushkarev, D., Neff, N.F. & Quake, S.R. 2009, "Single-molecule sequencing of an individual human genome", *Nature biotechnology,* vol. 27, no. 9, pp. 847-850.

274 Quackenbush, J., Liang, F., Holt, I., Pertea, G. & Upton, J. 2000, "The TIGR Gene Indices: Reconstruction and representation of expressed gene sequences", *Nucleic acids research,* vol. 28, no. 1, pp. 141-145.

275 Radutoiu, S., Madsen, L.H., Madsen, E.B., Felle, H.H., Umehara, Y., Grønlund, M., Sato, S., Nakamura, Y., Tabata, S., Sandal, N. & Stougaard, J. 2003, "Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases", *Nature,* vol. 425, no. 6958, pp. 585-592.

276 Raser, J.M. & O'Shea, E.K. 2005, "Molecular biology - Noise in gene expression: Origins, consequences, and control", *Science,* vol. 309, no. 5743, pp. 2010-2013.

277 Ridge, R.W. & Rolfe, B.G. 1985, "Rhizobium sp. degradation of legume root hair cell wall at the site of infection thread origin", *Applied and Environmental Microbiology,* vol. 50, no. 3, pp. 717-720.

278 Ringnér, M. 2008, "What is principal component analysis?", *Nature biotechnology,* vol. 26, no. 3, pp. 303-304.

279 Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y.S., Newsome, R., Chan, S.K., She, R., Varhol, R., Kamoh, B., Prabhu, A.-., Tam, A., Zhao, Y., Moore, R.A., Hirst, M., Marra, M.A., Jones, S.J.M., Hoodless, P.A. & Birol, I. 2010, "De novo assembly and analysis of RNA-seq data", *Nature Methods,* vol. 7, no. 11, pp. 909-912.

280 Robinson, M.D., McCarthy, D.J. & Smyth, G.K. 2010, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.", *Bioinformatics (Oxford, England),* vol. 26, no. 1, pp. 139-140.

281 Robinson, M.D. & Oshlack, A. 2010, "A scaling normalization method for differential expression analysis of RNA-seq data", *Genome biology,* vol. 11, no. 3.

282 Robinson, M.D. & Smyth, G.K. 2008, "Small-sample estimation of negative binomial dispersion, with applications to SAGE data", *Biostatistics,* vol. 9, no. 2, pp. 321-332.

283 Robinson, M.D. & Smyth, G.K. 2007, "Moderated statistical tests for assessing differences in tag abundance", *Bioinformatics,* vol. 23, no. 21, pp. 2881-2887.

284 Robinson, S.J., Cram, D.J., Lewis, C.T. & Parkin, I.A. 2004, "Maximizing the efficacy of SAGE analysis identifies novel transcripts in Arabidopsis.", *Plant Physiology,* vol. 136, no. 2, pp. 3223-3233.

285 Roche Diagnostics 2010, , *GS FLX+ System* [Homepage of Roche Diagnostics GmbH], [Online]. Available: http://454.com/downloads/GSFLX+_InstumentSpecSheet.pdf [2011, 02/21].

286 Ronaghi, M., Uhlen, M. & Nyren, P. 1998, "A sequencing method based on real-time pyrophosphate", *Science (New York, N.Y.),* vol. 281, no. 5375, pp. 363, 365.

287 Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. & Nyren, P. 1996, "Real-time DNA sequencing using detection of pyrophosphate release", *Analytical Biochemistry,* vol. 242, no. 1, pp. 84-89.

288 Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., Hoon, J., Simons, J.F., Marran, D., Myers, J.W., Davidson, J.F., Branting, A., Nobile, J.R., Puc, B.P., Light, D., Clark, T.A., Huber, M., Branciforte, J.T., Stoner, I.B., Cawley, S.E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J.A., Namsaraev, E., McKernan, K.J., Williams, A., Roth, G.T. & Bustillo, J. 2011, "An integrated semiconductor device enabling non-optical genome sequencing", *Nature,* vol. 475, no. 7356, pp. 348-352.

289 Rusk, N. 2011, "Torrents of sequence", *Nat Meth,* vol. 8, no. 1, pp. 44-44.

290 Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, K.W. & Velculescu, V.E. 2002, "Using the transcriptome to annotate the genome", *Nat.Biotechnol.,* vol. 20, no. 5, pp. 508-512.

291 Saito, K., Yoshikawa, M., Yano, K., Miwa, H., Uchida, H., Asamizu, E., Sato, S., Tabata, S., Imaizumi-Anraku, H., Umehara, Y., Kouchi, H., Murooka, Y., Szczyglowski, K., Downie, J.A., Parniske, M., Hayashi, M. & Kawaguchi, M. 2007, "Nucleoporin85 is required for calcium spiking, fungal and bacterial symbioses, and seed production in Lotus japonicus", *Plant Cell,* vol. 19, no. 2, pp. 610-624.

292 Sandal, N., Petersen, T.R., Murray, J., Umehara, Y., Karas, B., Yano, K., Kumagai, H., Yoshikawa, M., Saito, K., Hayashi, M., Murakami, Y., Wang, X., Hakoyama, T., Imaizumi-Anraku, H., Sato, S., Kato, T., Chen, W., Hossain, M.S., Shibata, S., Wang, T.L., Yokota, K., Larsen, K., Kanamori, N., Madsen, E., Radutoiu, S., Madsen, L.H., Radu, T.G., Krusell, L., Ooki, Y., Banba, M., Betti, M., Rispail, N., Skøt, L., Tuck, E., Perry, J., Yoshida, S., Vickers, K., Pike, J., Mulder, L., Charpentier, M., Müller, J., Ohtomo, R., Kojima, T., Ando, S., Marquez, A.J., Gresshoff, P.M., Harada, K., Webb, J., Hata, S., Suganuma, N., Kouchi, H., Kawasaki, S., Tabata, S., Hayashi, M., Parniske, M., Szczyglowski, K., Kawaguchi, M. & Stougaard, J. 2006, "Genetics of symbiosis in Lotus japonicus: Recombinant inbred lines, comparative genetic maps, and map position of 35 symbiotic loci", *Molecular Plant-Microbe Interactions,* vol. 19, no. 1, pp. 80-91.

293 Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M. & Smith, M. 1977, "Nucleotide sequence of bacteriophage phi X174 DNA", *Nature,* vol. 265, no. 5596, pp. 687-695.

294 Sanger, F., Nicklen, S. & Coulson, A.R. 1977, "DNA sequencing with chain-terminating inhibitors", *Proc.Natl.Acad.Sci.U.S.A,* vol. 74, no. 12, pp. 5463-5467.

295 Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., Sasamoto, S., Watanabe, A., Ono, A., Kawashima, K., Fujishiro, T., Katoh, M., Kohara, M., Kishida, Y., Minami, C., Nakayama, S., Nakazaki, N., Shimizu, Y., Shinpo, S., Takahashi, C., Wada, T., Yamada, M., Ohmido, N., Hayashi, M., Fukui, K., Baba, T., Nakamichi, T., Mori, H. & Tabata, S. 2008, "Genome structure of the legume, Lotus japonicus", *DNA Research,* vol. 15, no. 4, pp. 227-239.

296 Schadt, E.E., Turner, S. & Kasarskis, A. 2010, "A window into third-generation sequencing.", *Human molecular genetics,* vol. 19, no. R2, pp. R227-240.

297 Schauser, L., Roussis, A., Stiller, J. & Stougaard, J. 1999, "A plant regulator controlling development of symbiotic root nodules", *Nature,* vol. 402, no. 6758, pp. 191-195.

298 Schauser, L., Handberg, K., Sandal, N., Stiller, J., Thykjær, T., Pajuelo, E., Nielsen, A. & Stougaard, J. 1998, "Symbiotic mutants deficient in nodule establishment identified after T-DNA transformation of Lotus japonicus", *Molecular and General Genetics,* vol. 259, no. 4, pp. 414-423.

299 Scheidig, A., Fröhlich, A., Schulze, S., Lloyd, J.R. & Kossmann, J. 2002, "Downregulation of a chloroplast-targeted β-amylase leads to a starch-excess phenotype in leaves", *Plant Journal,* vol. 30, no. 5, pp. 581-591.

300 Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. 1995, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray", *Science,* vol. 270, no. 5235, pp. 467-470.

301 Schneider, M.V. & Orchard, S. 2011, "Omics technologies, data and bioinformatics principles", *Methods in molecular biology,* vol. 719, pp. 3.

302 Schwanhüusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. & Selbach, M. 2011, "Global quantification of mammalian gene expression control", *Nature,* vol. 473, no. 7347, pp. 337-342.

303 Schwartz, A.M., Komarova, T.V., Skulachev, M.V., Zvereva, A.S., Dorokhov, Y.L. & Atabekov, J.G. 2006, "Stability of plant mRNAs depends on the length of the 3′-untranslated region", *Biochemistry (Moscow),* vol. 71, no. 12, pp. 1377-1384.

304 Shalon, D., Smith, S.J. & Brown, P.O. 1996, "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization", *Genome research,* vol. 6, no. 7, pp. 639-645.

305 Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. 2003, "Cytoscape: A software Environment for integrated models of biomolecular interaction networks", *Genome research,* vol. 13, no. 11, pp. 2498-2504.

306 Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. & Church, G.M. 2005, "Accurate multiplex polony sequencing of an evolved bacterial genome", *Science (New York, N.Y.),* vol. 309, no. 5741, pp. 1728-1732.

307 Shendure, J. & Ji, H. 2008, "Next-generation DNA sequencing", *Nat Biotech,* vol. 26, no. 10, pp. 1135-1145.

308 Shi, L., Twary, S.N., Yoshioka, H., Gregerson, R.G., Miller, S.S., Samac, D.A., Gantt, J.S., Unkefer, P.J. & Vance, C.P. 1997, "Nitrogen assimilation in alfalfa: Isolation and characterization of an asparagine synthetase gene showing enhanced expression in root nodules and dark-adapted leaves", *Plant Cell,* vol. 9, no. 8, pp. 1339-1356.

309 Shimomura, K., Nomura, M., Tajima, S. & Kouchi, H. 2006, "LjnsRING, a novel RING finger protein, is required for symbiotic interactions between Mesorhizobium loti and Lotus japonicus", *Plant and Cell Physiology,* vol. 47, no. 11, pp. 1572-1581.

310 Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M. & Birol, I. 2009, "ABySS: A parallel assembler for short read sequence data", *Genome research,* vol. 19, no. 6, pp. 1117-1123.

311 Smith, A.D., Xuan, Z. & Zhang, M.Q. 2008, "Using quality scores and longer reads improves accuracy of Solexa read mapping", *BMC Bioinformatics,* vol. 9.

312 Smith, L.G. & Oppenheimer, D.G. 2005, *Spatial control of cell expansion by the plant cytoskeleton*.

313 Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B. & Hood, L.E. 1986, "Fluorescence detection in automated DNA sequence analysis", *Nature,* vol. 321, no. 6071, pp. 674-679.

314 Smith, T.F. & Waterman, M.S. 1981, "Identification of common molecular subsequences", *Journal of Molecular Biology,* vol. 147, no. 1, pp. 195-197.

315 Smyth, G.K. 2004, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments", *Statistical Applications in Genetics and Molecular Biology,* vol. 3, no. 1.

316 Sonenberg, N. 1994, "mRNA translation: Influence of the 5' and 3' untranslated regions", *Current Opinion in Genetics and Development,* vol. 4, no. 2, pp. 310-315.

317 Stracke, S., Kistner, C., Yoshida, S., Mulder, L., Sato, S., Kaneko, T., Tabata, S., Sandal, N., Stougaard, J., Szczyglowski, K. & Parniske, M. 2002, "A plant receptor-like kinase required for both bacterial and fungal symbiosis", *Nature,* vol. 417, no. 6892, pp. 959-962.

318 Suganuma, N., Nakamura, Y., Yamamoto, M., Ohta, T., Koiwa, H., Akao, S. & Kawaguchi, M. 2003, "The Lotus japonicus Sen1 gene controls rhizobial differentiation into nitrogen-fixing bacteroids in nodules", *Molecular Genetics and Genomics,* vol. 269, no. 3, pp. 312-320.

319 Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H. & Yaspo, M.-. 2008, "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome", *Science,* vol. 321, no. 5891, pp. 956-960.

320 Swain, P.S., Elowitz, M.B. & Siggia, E.D. 2002, "Intrinsic and extrinsic contributions to stochasticity in gene expression", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 99, no. 20, pp. 12795-12800.

321 Takanashi, K., Sugiyama, A. & Yazaki, K. 2011, "Involvement of auxin distribution in root nodule development of Lotus japonicus", *Planta,* vol. 234, no. 1, pp. 73-81.

322 Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K. & Surani, M.A. 2009, "mRNA-Seq whole-transcriptome analysis of a single cell", *Nature Methods,* vol. 6, no. 5, pp. 377-382.

323 Tang, K.-., Yao, W.-., Li, T.-., Li, Y.-. & Cao, Z.-. 2010, "Cancer classification from the gene expression profiles by discriminant kernel-pls", *Journal of Bioinformatics and Computational Biology,* vol. 8, no. SUPPL. 1, pp. 147-160.

324 Tauberger, E., Fernie, A.R., Emmermann, M., Renz, A., Kossmann, J., Willmitzer, L. & Trethewey, R.N. 2000, "Antisense inhibition of plastidial phosphoglucomutase provides compelling evidence that potato tuber amyloplasts import carbon from the cytosol in the form of glucose-6-phosphate", *The Plant Journal : for cell and molecular biology,* vol. 23, no. 1, pp. 43-53.

325 Team, R.D.C. 2012, "R: A Language and Environment for Statistical Computing", .

326 The DNA Sequencing Facility 2010, 07/10-last update, *454 GS FLX Seqeuncer* [Homepage of The DNA Sequencing Facility, Department of Genetics, University of Pennsylvania], [Online]. Available: http://www.med.upenn.edu/genetics/dnaseq/454sequencer.shtml [2011, 07/20].

327 The Gene Ontology Consortium 2011, 10/15-2011-last update, *Current Annotations* [Homepage of The Gene Ontology Consortium], [Online]. Available: http://www.geneontology.org/GO.current.annotations.shtml [2011, 10/16].

328 the National Human Genome Research Institute 2010, 30/10-last update, *The Human Genome Project Completion: Frequently Asked Questions* [Homepage of the National Human Genome Research Institute], [Online]. Available: http://www.genome.gov/11006943 [2011, 07/19].

329 The Potato Genome Sequencing Consortium, Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang (Principal, S., Li (Principal, R., Wang (Principal, J., Orjeda (Principal, G., Guzman, F., Torres, M., Lozano, R., Ponce, O., Martinez, D., De la Cruz, G., Chakrabarti (Principal, S.K., Patil, V.U., Skryabin (Principal, K.G., Kuznetsov, B.B., Ravin, N.V., Kolganova, T.V., Beletsky, A.V., Mardanov, A.V., Di Genova, A., Bolser, D.M., Martin (Principal, D.M.A., Li, G., Yang, Y., Kuang, H., Hu, Q., Xiong, X., Bishop, G.J., Sagredo (Principal, B., Mejía, N., Zagorski (Principal, W., Gromadka, R., Gawor, J., Szczesny, P., Huang (Principal, S., Zhang, Z., Liang, C., He, J., Li, Y., He, Y., Xu, J., Zhang, Y., Xie, B., Du, Y., Qu (Principal, D., Bonierbale, M., Ghislain, M., del Rosario Herrera, M., Giuliano (Principal, G., Pietrella, M., Perrotta, G., Facella, P., O'Brien, K., Feingold (Principal, S.E., Barreiro, L.E., Massa, G.A., Diambra, L., Whitty, B.R., Vaillancourt, B., Lin, H., Massa, A.N., Geoffroy, M., Lundback, S., DellaPenna, D., Robin Buell (Principal, C., Sharma, S.K., Marshall, D.F., Waugh, R., Bryan (Principal, G.J., Destefanis, M., Nagy, I., Milbourne (Principal, D., Thomson, S.J., Fiers, M., Jacobs (Principal, J.M.E., Nielsen (Principal, K.L., Sønderkær, M., Iovene, M., Torres, G.A., Jiang (Principal, J., Veilleux, R.E., Bachem (Principal, C.W.B., de Boer, J., Borm, T., Kloosterman, B., van Eck, H., Datema, E., te Lintel Hekkert, B., Goverse, A., van Ham, R.C.H.J. & Visser, R.G.F. 2011, "Genome sequence and analysis of the tuber crop potato", *Nature,* vol. 475, pp. 189-189–195.

330 Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S. & Park, P.J. 2005, "Discovering statistically significant pathways in expression profiling studies", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, no. 38, pp. 13544-13549.

331 Tirichine, L., Sandal, N., Madsen, L.H., Radutoiu, S., Albrektsen, A.S., Sato, S., Asamizu, E., Tabata, S. & Stougaard, J. 2007, "A gain-of-function mutation in a cytokinin receptor triggers spontaneous root nodule organogenesis", *Science,* vol. 315, no. 5808, pp. 104-107.

332 Tirichine, L., Imaizumi-Anraku, H., Yoshida, S., Murakami, Y., Madsen, L.H., Miwa, H., Nakagawa, T., Sandal, N., Albrektsen, A.S., Kawaguchi, M., Downie, A., Sato, S., Tabata, S., Kouchi, H., Parniske, M., Kawasaki, S. & Stougaard, J. 2006, "Deregulation of a Ca2+/calmodulin-dependent kinase leads to spontaneous nodule development", *Nature,* vol. 441, no. 7097, pp. 1153-1156.

333 Trapnell, C., Roberts, A., Pertea, G., Williams, B., Mortazavi, A., Kwan, G., Van Baren, J., Salzberg, S., Wold, B. & Pachter, L. 2011, 11/30/2011-last update, *Cufflinks Transcript assembly, differential expression, and differential regulation for RNA-Seq* [Homepage of The Laboratory for Mathematical and Computational Biology at UC Berkeley, the computational genomics group at the Institute of Genetic Medicine at Johns Hopkins University, and the Wold Lab at Caltec], [Online]. Available: http://cufflinks.cbcb.umd.edu/index.html [2011, 12/18].

334 Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., Salzberg, S.L., Wold, B.J. & Pachter, L. 2010, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation", *Nature biotechnology,* vol. 28, no. 5, pp. 511-515.

335 Trapnell, C., Pachter, L. & Salzberg, S.L. 2009, "TopHat: Discovering splice junctions with RNA-Seq", *Bioinformatics,* vol. 25, no. 9, pp. 1105-1111.

336 Turcatti, G., Romieu, A., Fedurco, M. & Tairi, A.-. 2008, "A new class of cleavable fluorescent nucleotides: Synthesis and optimization as reversible terminators for DNA sequencing by synthesis", *Nucleic acids research,* vol. 36, no. 4.

337 U.S. Department of Energy Genome Programs 2011, 31/05-last update, *Human Genome Project Information* [Homepage of U.S. Department of Energy Genome Programs], [Online]. Available: http://genomics.energy.gov [2011, 19/07].

338 Udvardi, M.K., Tabata, S., Parniske, M. & Stougaard, J. 2005, "Lotus japonicus: Legume research in the fast lane", *Trends in plant science,* vol. 10, no. 5, pp. 222-228.

339 van Bakel, H., Nislow, C., Blencowe, B.J. & Hughes, T.R. 2010, "Most "dark matter" transcripts are associated with known genes", *PLoS Biology,* vol. 8, no. 5.

340 van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K. & van der Werf, M.J. 2006, "Centering, scaling, and transformations: Improving the biological information content of metabolomics data", *BMC Genomics,* vol. 7.

341 Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. 1995, "Serial analysis of gene expression", *Science,* vol. 270, no. 5235, pp. 484-487.

342 Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., bu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di, F.,V, Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J.,

Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N. & Nodell, M. 2001, "The sequence of the human genome", *Science,* vol. 291, no. 5507, pp. 1304-1351.

343 Visser, R.G.F., Bachem, C.W.B., de Boer, J.M., Bryan, G.J., Chakrabati, S.K., Feingold, S., Gromadka, R., van Ham, R.C.H.J., Huang, S., Jacobs, J.M.E., Kuznetsov, B., de Melo, P.E., Milbourne, D., Orjeda, G., Sagredo, B. & Tang, X. 2009, "Sequencing the Potato genome: Outline and first results to come from the Elucidation of the sequence of the world's third most important food crop", *American Journal of Potato Research,* vol. 86, no. 6, pp. 417-429.

344 Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., MacLeod, J.N., Chiang, D.Y., Prins, J.F. & Liu, J. 2010, "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.", *Nucleic acids research,* vol. 38, no. 18.

345 Wang, L., Feng, Z., Wang, X., Wang, X. & Zhang, X. 2009, "DEGseq: An R package for identifying differentially expressed genes from RNA-seq data", *Bioinformatics,* vol. 26, no. 1, pp. 136-138.

346 Wang, S.M. 2007, "Understanding SAGE data", *Trends in Genetics,* vol. 23, no. 1, pp. 42-50.

347 Wang, T.L., Domoney, C., Hedley, C.L., Casey, R. & Grusak, M.A. 2003, "Can we improve the nutritional quality of legume seeds?", *Plant Physiology,* vol. 131, no. 3, pp. 886-891.

348 Wang, X., Song, X., Glass, C.K. & Rosenfeld, M.G. 2011, "The long arm of long noncoding RNAs: roles as sensors regulating gene transcriptional programs.", *Cold Spring Harbor perspectives in biology,* vol. 3, no. 1.

349 Wang, Z., Gerstein, M. & Snyder, M. 2009, "RNA-Seq: a revolutionary tool for transcriptomics", *Nature reviews.Genetics,* vol. 10, no. 1, pp. 57-63.

350 Warren, R.L., Sutton, G.G., Jones, S.J.M. & Holt, R.A. 2007, "Assembling millions of short DNA sequences using SSAKE", *Bioinformatics,* vol. 23, no. 4, pp. 500-501.

351 Wass, J.A. 2005, "The unscrambler - A different approach to multivariate data analysis and experimental design", *Scientific Computing and Instrumentation,* vol. 22, no. 7, pp. 37-39.

352 Waterhouse, R.N., Smyth, A., Massonneau, A., Presser, L.M. & Clarkson, D.T. 1996, "Molecular cloning and characterisation of asparagine synthetase from Lotus japonicus: Dynamics of asparagine synthesis in N-sufficient conditions", *Plant Molecular Biology,* vol. 30, no. 5, pp. 883-897.

353 Weerasinghe, R.R., Collings, D.A., Johannes, E. & Allen, N.S. 2003, "The distributional changes and role of microtubules in Nod factor-challenged Medicago sativa root hairs", *Planta,* vol. 218, no. 2, pp. 276-287.

354 Weinstein, J.N. 2008, "A postgenomic visual icon", *Science,* vol. 319, no. 5871, pp. 1772-1773.

355 Wetterstrand, K.A. 2011, 02/04-last update, *DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program* [Homepage of National Human Genome Research Institute, NIH], [Online]. Available: http://www.genome.gov/sequencingcosts [2011, 07/26].

356 Whiteford, N., Skelly, T., Curtis, C., Ritchie, M.E., Löhr, A., Zaranek, A.W., Abnizova, I. & Brown, C. 2009, "Swift: Primary data analysis for the Illumina Solexa sequencing platform", *Bioinformatics,* vol. 25, no. 17, pp. 2194-2199.

357 Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J. & Bähler, J. 2008, "Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution", *Nature,* vol. 453, no. 7199, pp. 1239-1243.

358 Wu, Q., Kim, Y.C., Lu, J., Xuan, Z., Chen, J., Zheng, Y., Zhou, T., Zhang, M.Q., Wu, C.-. & Wang, S.M. 2008, "Poly A- transcripts expressed in Hela cells", *PLoS ONE,* vol. 3, no. 7.

359 Wu, T.D. & Nacu, S. 2010, "Fast and SNP-tolerant detection of complex variants and splicing in short reads", *Bioinformatics,* vol. 26, no. 7, pp. 873-881.

360 Yano, K., Shibata, S., Chen, W.-., Sato, S., Kaneko, T., Jurkiewicz, A., Sandal, N., Banba, M., Imaizumi-Anraku, H., Kojima, T., Ohtomo, R., Szczyglowski, K., Stougaard, J., Tabata, S., Hayashi, M., Kouchi, H. & Umehara, Y. 2009, "CERBERUS, a novel U-box protein containing WD-40 repeats, is required for formation of the infection thread and nodule development in the legume-Rhizobium symbiosis", *Plant Journal,* vol. 60, no. 1, pp. 168-180.

361 Yano, K., Yoshida, S., Müller, J., Singh, S., Banba, M., Vickers, K., Markmann, K., White, C., Schuller, B., Sato, S., Asamizu, E., Tabata, S., Murooka, Y., Perry, J., Wang, T.L., Kawaguchi, M., Imaizumi-Anraku, H., Hayashi, M. & Parniske, M. 2008, "CYCLOPS, a mediator of symbiotic intracellular accommodation", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 105, no. 51, pp. 20540-20545.

362 Yassoura, M., Kaplana, T., Fraser, H.B., Levin, J.Z., Pfiffner, J., Adiconis, X., Schroth, G., Luo, S., Khrebtukova, I., Gnirke, A., Nusbaum, C., Thompson, D.-., Friedman, N. & Regev, A. 2009, "Ab initio construction of

a eukaryotic transcriptome by massively parallel mRNA sequencing", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 106, no. 9, pp. 3264-3269.

363 Yokota, K., Fukai, E., Madsen, L.H., Jurkiewicz, A., Rueda, P., Radutoiu, S., Held, M., Hossain, M.S., Szczyglowski, K., Morieri, G., Oldroyd, G.E.D., Downie, J.A., Nielsen, M.W., Rusek, A.M., Sato, S., Tabata, S., James, E.K., Oyaizu, H., Sandal, N. & Stougaard, J. 2009, "Rearrangement of actin cytoskeleton mediates invasion of lotus japonicus roots by Mesorhizobium loti", *Plant Cell,* vol. 21, no. 1, pp. 267-284.

364 Zerbino, D.R. & Birney, E. 2008, "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs", *Genome research,* vol. 18, no. 5, pp. 821-829.

365 Zhang, J., Chiodini, R., Badr, A. & Zhang, G. 2011, "The impact of next-generation sequencing on genomics", *Journal of Genetics and Genomics,* vol. 38, no. 3, pp. 95-109.

366 Zhang, L., Zhou, W., Velculescu, V.E., Kern, S.E., Hruban, R.H., Hamilton, S.R., Vogelstein, B. & Kinzler, K.W. 1997, "Gene expression profiles in normal and cancer cells", *Science,* vol. 276, no. 5316, pp. 1268-1272.

367 Zimin, A.V., Delcher, A.L., Florea, L., Kelley, D.R., Schatz, M.C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C.P., Sonstegard, T.S., Marçais, G., Roberts, M., Subramanian, P., Yorke, J.A. & Salzberg, S.L. 2009, "A whole-genome assembly of the domestic cow, Bos taurus", *Genome biology,* vol. 10, no. 4.

368 Zou, C., Lehti-Shiu, M.D., Thibaud-Nissen, F., Prakash, T., Buell, C.R. & Shiu, S.-. 2009, "Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice", *Plant Physiology,* vol. 151, no. 1, pp. 3-15.

APENDICES

A) Published Articles

Genome Sequence and Analysis of the Tuber Crop Potato

The current appendix contains the resulting published article from the potato genome sequencing project (The Potato Genome Sequencing Consortium *et al.*, 2011). The supplementary text (500 pages in total) with supporting tables and figures, and detailed description of methods can be found on the enclosed CD in the file "*Genome Sequence and Analysis of the Tuber Crop Potato Supplementary text.pdf*".

ARTICLE

doi:10.1038/nature10158

Genome sequence and analysis of the tuber crop potato

The Potato Genome Sequencing Consortium*

Potato (*Solanum tuberosum* L.) is the world's most important non-grain food crop and is central to global food security. It is clonally propagated, highly heterozygous, autotetraploid, and suffers acute inbreeding depression. Here we use a homozygous doubled-monoploid potato clone to sequence and assemble 86% of the 844-megabase genome. We predict 39,031 protein-coding genes and present evidence for at least two genome duplication events indicative of a palaeopolyploid origin. As the first genome sequence of an asterid, the potato genome reveals 2,642 genes specific to this large angiosperm clade. We also sequenced a heterozygous diploid clone and show that gene presence/absence variants and other potentially deleterious mutations occur frequently and are a likely cause of inbreeding depression. Gene family expansion, tissue-specific expression and recruitment of genes to new pathways contributed to the evolution of tuber development. The potato genome sequence provides a platform for genetic improvement of this vital crop.

Potato (*Solanum tuberosum* L.) is a member of the Solanaceae, an economically important family that includes tomato, pepper, aubergine (eggplant), petunia and tobacco. Potato belongs to the asterid clade of eudicot plants that represents ~25% of flowering plant species and from which a complete genome sequence has not yet, to our knowledge, been published. Potato occupies a wide eco-geographical range[1] and is unique among the major world food crops in producing stolons (underground stems) that under suitable environmental conditions swell to form tubers. Its worldwide importance, especially within the developing world, is growing rapidly, with production in 2009 reaching 330 million tons (http://www.fao.org). The tubers are a globally important dietary source of starch, protein, antioxidants and vitamins[2], serving the plant as both a storage organ and a vegetative propagation system. Despite the importance of tubers, the evolutionary and developmental mechanisms of their initiation and growth remain elusive.

Outside of its natural range in South America, the cultivated potato is considered to have a narrow genetic base resulting originally from limited germplasm introductions to Europe. Most potato cultivars are autotetraploid ($2n = 4x = 48$), highly heterozygous, suffer acute inbreeding depression, and are susceptible to many devastating pests and pathogens, as exemplified by the Irish potato famine in the mid-nineteenth century. Together, these attributes present a significant barrier to potato improvement using classical breeding approaches. A challenge to the scientific community is to obtain a genome sequence that will ultimately facilitate advances in breeding.

To overcome the key issue of heterozygosity and allow us to generate a high-quality draft potato genome sequence, we used a unique homozygous form of potato called a doubled monoploid, derived using classical tissue culture techniques[3]. The draft genome sequence from this genotype, *S. tuberosum* group Phureja DM1-3 516 R44 (hereafter referred to as DM), was used to integrate sequence data from a heterozygous diploid breeding line, *S. tuberosum* group Tuberosum RH89-039-16 (hereafter referred to as RH). These two genotypes represent a sample of potato genomic diversity; DM with its fingerling (elongated) tubers was derived from a primitive South American cultivar whereas RH more closely resembles commercially cultivated tetraploid potato. The combined data resources, allied to deep transcriptome sequence from both genotypes, allowed us to explore potato genome structure and organization, as well as key aspects of the biology and evolution of this important crop.

Genome assembly and annotation

We sequenced the nuclear and organellar genomes of DM using a whole-genome shotgun sequencing (WGS) approach. We generated 96.6 Gb of raw sequence from two next-generation sequencing (NGS) platforms, Illumina Genome Analyser and Roche Pyrosequencing, as well as conventional Sanger sequencing technologies. The genome was assembled using SOAPdenovo[4], resulting in a final assembly of 727 Mb, of which 93.9% is non-gapped sequence. Ninety per cent of the assembly falls into 443 superscaffolds larger than 349 kb. The 17-nucleotide depth distribution (Supplementary Fig. 1) suggests a genome size of 844 Mb, consistent with estimates from flow cytometry[5]. Our assembly of 727 Mb is 117 Mb less than the estimated genome size. Analysis of the DM scaffolds indicates 62.2% repetitive content in the assembled section of the DM genome, less than the 74.8% estimated from bacterial artificial chromosome (BAC) and fosmid end sequences (Supplementary Table 1), indicating that much of the unassembled genome is composed of repetitive sequences.

We assessed the quality of the WGS assembly through alignment to Sanger-derived phase 2 BAC sequences. In an alignment length of ~1 Mb (99.4% coverage), no gross assembly errors were detected (Supplementary Table 2 and Supplementary Fig. 2). Alignment of fosmid and BAC paired-end sequences to the WGS scaffolds revealed limited (≤0.12%) potential misassemblies (Supplementary Table 3). Extensive coverage of the potato genome in this assembly was confirmed using available expressed sequence tag (EST) data; 97.1% of 181,558 available Sanger-sequenced *S. tuberosum* ESTs (>200 bp) were detected. Repetitive sequences account for at least 62.2% of the assembled genome (452.5 Mb) (Supplementary Table 1) with long terminal repeat retrotransposons comprising the majority of the transposable element classes, representing 29.4% of the genome. In addition, subtelomeric repeats were identified at or near chromosomal ends (Fig. 1). Using a newly constructed genetic map based on 2,603 polymorphic markers in conjunction with other available

*Lists of authors and their affiliations appear at the end of the paper.

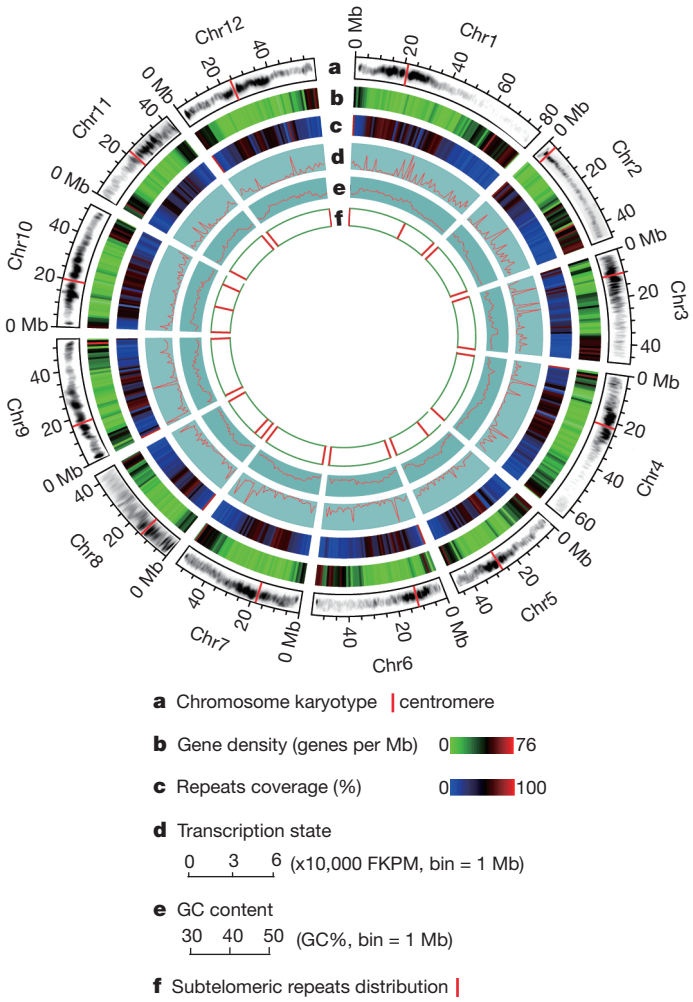


Figure 1 | The potato genome. **a**, Ideograms of the 12 pseudochromosomes of potato (in Mb scales). Each of the 12 pachytene chromosomes from DM was digitally aligned with the ideogram (the amount of DNA in each unit of the pachytene chromosomes is not in proportion to the scales of the pseudochromosomes). **b**, Gene density represented as number of genes per Mb (non-overlapping, window size = 1 Mb). **c**, Percentage of coverage of repetitive sequences (non-overlapping windows, window size = 1 Mb). **d**, Transcription state. The transcription level for each gene was estimated by averaging the fragments per kb exon model per million mapped reads (FPKM) from different tissues in non-overlapping 1-Mb windows. **e**, GC content was estimated by the per cent G+C in 1-Mb non-overlapping windows. **f**, Distribution of the subtelomeric repeat sequence CL14_cons.

genetic and physical maps, we genetically anchored 623 Mb (86%) of the assembled genome (Supplementary Fig. 3), and constructed pseudomolecules for each of the 12 chromosomes (Fig. 1), which harbour 90.3% of the predicted genes.

To aid annotation and address a series of biological questions, we generated 31.5 Gb of RNA-Seq data from 32 DM and 16 RH libraries representing all major tissue types, developmental stages and responses to abiotic and biotic stresses (Supplementary Table 4). For annotation, reads were mapped against the DM genome sequence (90.2% of 824,621,408 DM reads and 88.6% of 140,375,647 RH reads) and in combination with *ab initio* gene prediction, protein and EST alignments, we annotated 39,031 protein-coding genes. RNA-Seq data revealed alternative splicing; 9,875 genes (25.3%) encoded two or more isoforms, indicative of more functional variation than represented by the gene set alone. Overall, 87.9% of the gene models were supported by transcript and/or protein similarity with only 12.1% derived solely from *ab initio* gene predictions (Supplementary Table 5).

Karyotypes of RH and DM suggested similar heterochromatin content[6] (Supplementary Table 6 and Supplementary Fig. 4) with large blocks of heterochromatin located at the pericentromeric regions (Fig. 1). As observed in other plant genomes, there was an inverse relationship between gene density and repetitive sequences (Fig. 1). However, many predicted genes in heterochromatic regions are expressed, consistent with observations in tomato[7] that genic 'islands' are present in the heterochromatic 'ocean'.

Genome evolution

Potato is the first sequenced genome of an asterid, a clade within eudicots that encompasses nearly 70,000 species characterized by unique morphological, developmental and compositional features[8]. Orthologous clustering of the predicted potato proteome with 11 other green plant genomes revealed 4,479 potato genes in 3,181 families in common (Fig. 2a); 24,051 potato genes clustered with at least one of the 11 genomes. Filtering against transposable elements and 153 nonasterid and 57 asterid publicly available transcript-sequence data sets yielded 2,642 high-confidence asterid-specific and 3,372 potato-lineage-specific genes (Supplementary Fig. 5); both sets were enriched for genes of unknown function that had less expression support than the core Viridiplantae genes. Genes encoding transcription factors, self-incompatibility, and defence-related proteins were evident in the asterid-specific gene set (Supplementary Table 7) and presumably contribute to the unique characteristics of asterids.

Structurally, we identified 1,811 syntenic gene blocks involving 10,046 genes in the potato genome (Supplementary Table 8). On the basis of these pairwise paralogous segments, we calculated an age distribution based on the number of transversions at fourfold degenerate sites (4DTv) for all duplicate pairs. In general, two significant groups of blocks are seen in the potato genome (4DTv ~0.36 and ~1.0; Fig. 2b), suggesting two whole-genome duplication (WGD) events. We also identified collinear blocks between potato and three rosid genomes (*Vitis vinifera*, *Arabidopsis thaliana* and *Populus trichocarpa*) that also suggest both events (Fig. 2c and Supplementary Fig. 6). The ancient WGD corresponds to the ancestral hexaploidization (γ) event in grape (Fig. 2b), consistent with a previous report based on EST analysis that the two main branches of eudicots, the asterids and rosids, may share the same palaeo-hexaploid duplication event[9]. The γ event probably occurred after the divergence between dicots and monocots about 185 ± 55 million years ago[10]. The recent duplication can therefore be placed at ~67 million years ago, consistent with the WGD that occurred near the Cretaceous–Tertiary boundary (~65 million years ago)[11]. The divergence of potato and grape occurred at ~89 million years ago (4DTv ~0.48), which is likely to represent the split between the rosids and asterids.

Haplotype diversity

High heterozygosity and inbreeding depression are inherent to potato, a species that predominantly outcrosses and propagates by means of vegetative organs. Indeed, the phenotypes of DM and RH differ, with RH more vigorous than DM (Fig. 3a). To explore the extent of haplotype diversity and possible causes of inbreeding depression, we sequenced and assembled 1,644 RH BAC clones generating 178 Mb of non-redundant sequence from both haplotypes (~10% of the RH genome with uneven coverage) (Supplementary Tables 9–11). After filtering to remove repetitive sequences, we aligned 99 Mb of RH sequence (55%) to the DM genome. These regions were largely collinear with an overall sequence identity of 97.5%, corresponding to one single-nucleotide polymorphism (SNP) every 40 bp and one insertion/deletion (indel) every 394 bp (average length 12.8 bp). Between the two RH haplotypes, 6.6 Mb of sequence could be aligned with 96.5% identity, corresponding to 1 SNP per 29 bp and 1 indel per 253 bp (average length 10.4 bp).

Current algorithms are of limited use in *de novo* whole-genome assembly or haplotype reconstruction of highly heterozygous genomes

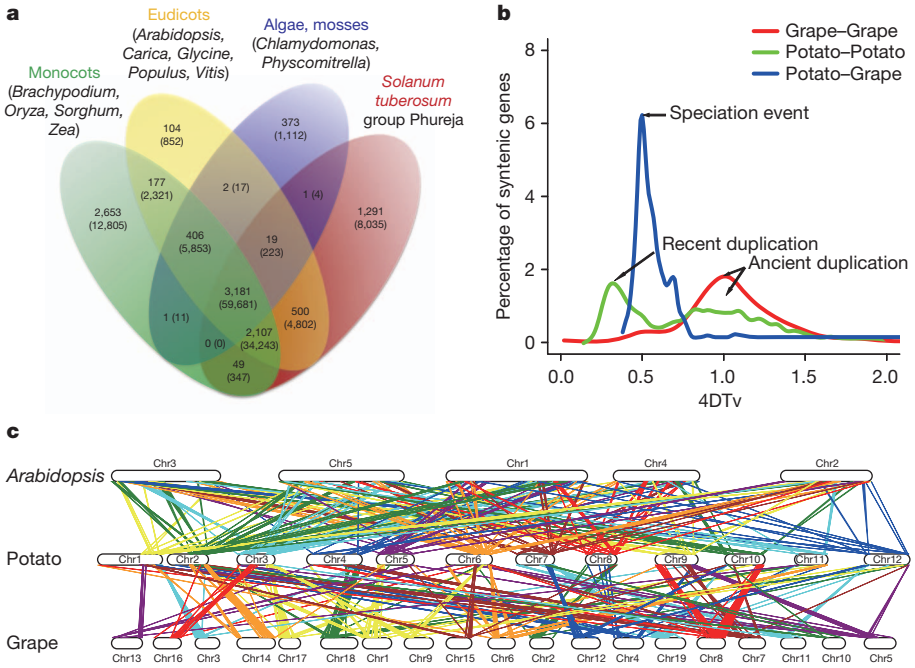


Figure 2 | Comparative analyses and evolution of the potato genome. **a**, Clusters of orthologous and paralogous gene families in 12 plant species as identified by OrthoMCL[33]. Gene family number is listed in each of the components; the number of genes within the families for all of the species within the component is noted within parentheses. **b**, Genome duplication in dicot genomes as revealed through 4DTv analyses. **c**, Syntenic blocks between *A. thaliana*, potato, and *V. vinifera* (grape) demonstrating a high degree of conserved gene order between these taxa.

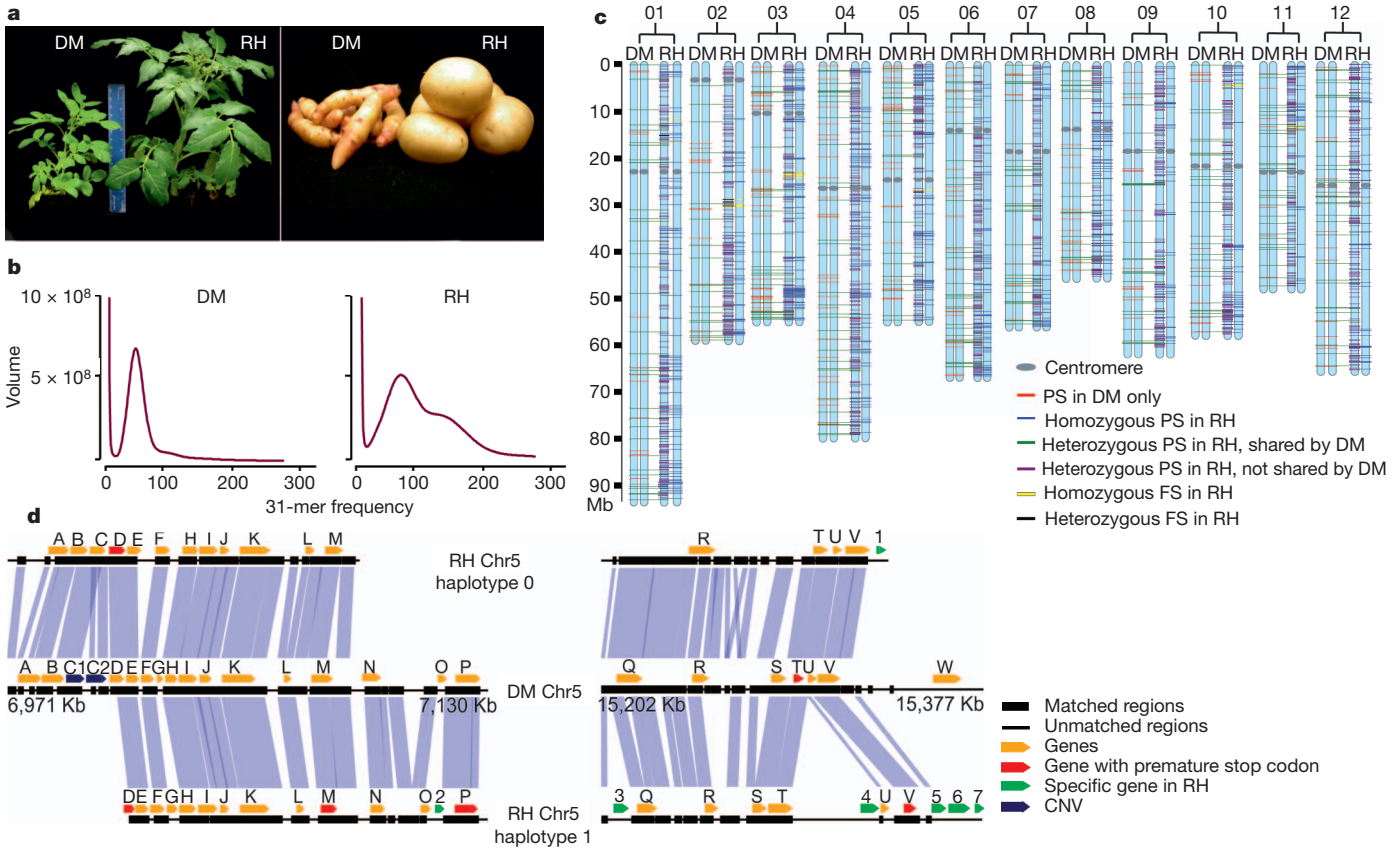


Figure 3 | Haplotype diversity and inbreeding depression. **a**, Plants and tubers of DM and RH showing that RH has greater vigour. **b**, Illumina K-mer volume histograms of DM and RH. The volume of K-mers (*y*-axis) is plotted against the frequency at which they occur (*x*-axis). The leftmost truncated peaks at low frequency and high volume represent K-mers containing essentially random sequencing errors, whereas the distribution to the right represents proper (putatively error-free) data. In contrast to the single modality of DM, RH exhibits clear bi-modality caused by heterozygosity. **c**, Genomic distribution of premature stop, frameshift and presence/absence variation mutations contributing to inbreeding depression. The hypothetical RH pseudomolecules were solely inferred from the corresponding DM ones. Owing to the inability to assign heterozygous PS and FS of RH to a definite haplotype, all heterozygous PS and FS were arbitrarily mapped to the left haplotype of RH. **d**, A zoom-in comparative view of the DM and RH genomes. The left and right alignments are derived from the euchromatic and heterochromatic regions of chromosome 5, respectively. Most of the gene annotations, including PS and RH-specific genes, are supported by transcript data.

such as RH, as shown by K-mer frequency count histograms (Fig. 3b and Supplementary Table 12). To complement the BAC-level comparative analysis and provide a genome-wide perspective of heterozygosity in RH, we mapped 1,118 million whole-genome NGS reads from RH (84× coverage) onto the DM assembly. A total of 457.3 million reads uniquely aligned providing 90.6% (659.1 Mb) coverage. We identified 3.67 million SNPs between DM and one or both haplotypes of RH, with an error rate of 0.91% based on evaluation of RH BAC sequences. We used this data set to explore the possible causes of inbreeding depression by quantifying the occurrence of premature stop, frameshift and presence/absence variants[12], as these disable gene function and contribute to genetic load (Supplementary Tables 13–16). We identified 3,018 SNPs predicted to induce premature stop codons in RH, with 606 homozygous (in both haplotypes) and 2,412 heterozygous. In DM, 940 premature stop codons were identified. In the 2,412 heterozygous RH premature stop codons, 652 were shared with DM and the remaining 1,760 were found in RH only (Fig. 3c and Supplementary Table 13). Frameshift mutations were identified in 80 loci within RH, 49 homozygous and 31 heterozygous, concentrated in seven genomic regions (Fig. 3c and Supplementary Table 14). Finally, we identified presence/absence variations for 275 genes; 246 were RH specific (absent in DM) and 29 were DM specific, with 125 and 9 supported by RNA-Seq and/or Gene Ontology[13] annotation for RH and DM, respectively (Supplementary Tables 15 and 16). Collectively, these data indicate that the complement of homozygous deleterious alleles in DM may be responsible for its reduced level of vigour (Fig. 3a).

The divergence between potato haplotypes is similar to that reported between out-crossing maize accessions[14] and, coupled with our inability to successfully align 45% of the BAC sequences, intra- and inter-genome diversity seem to be a significant feature of the potato genome. A detailed comparison of the three haplotypes (DM and the two haplotypes of RH) at two genomic regions (334 kb in length) using the RH BAC sequence (Fig. 3d and Supplementary Tables 17 and 18) revealed considerable sequence and structural variation. In one region ('euchromatic'; Fig. 3d) we observed one instance of copy number variation, five genes with premature stop codons, and seven RH-specific genes. These observations indicate that the plasticity of the potato genome is greater than revealed from the unassembled RH NGS. Improved assembly algorithms, increased read lengths, and *de novo* sequences of additional haplotypes will reveal the full catalogue of genes critical to inbreeding depression.

Tuber biology

In developing DM and RH tubers, 15,235 genes were expressed in the transition from stolons to tubers, with 1,217 transcripts exhibiting >5-fold expression in stolons versus five RH tuber tissues (young tuber, mature tuber, tuber peel, cortex and pith; Supplementary Table 19). Of these, 333 transcripts were upregulated during the transition from stolon to tuber, with the most highly upregulated transcripts encoding storage proteins. Foremost among these were the genes encoding proteinase inhibitors and patatin (15 genes), in which the phospholipase A function has been largely replaced by a protein storage function in the tuber[15]. In particular, a large family of 28 Kunitz protease inhibitor genes (KTIs) was identified with twice the number of genes in potato compared to tomato. The KTI genes are distributed across the genome with individual members exhibiting specific expression patterns (Fig. 4a, b). KTIs are frequently induced after pest and pathogen attack and act primarily as inhibitors of exogenous proteinases[16]; therefore the expansion of the KTI family may provide resistance to biotic stress for the newly evolved vulnerable underground organ.

The stolon to tuber transition also coincides with strong upregulation of genes associated with starch biosynthesis (Fig. 4c). We observed several starch biosynthetic genes that were 3–8-fold more highly expressed in tuber tissues of RH compared to DM (Fig. 4c). Together this suggests a stronger shift from the relatively low sink strength of the ATP-generating general carbon metabolism reactions towards the plastidic starch synthesis pathway in tubers of RH, thereby causing a flux of carbon into the amyloplast. This contrasts with the cereal endosperm where carbon is transported into the amyloplast in the form of ADP-glucose via a specific transporter (brittle 1 protein[17]). Carbon transport into the amyloplasts of potato tubers is primarily in the form of glucose-6-phosphate[18], although recent evidence indicates that glucose-1-phosphate is quantitatively important under certain conditions[19]. The transport mechanism for glucose-1-phosphate is unknown and the genome sequence contains six genes for hexose-phosphate transporters with two highly and specifically expressed in stolons and tubers. Furthermore, an additional 23 genes encode proteins homologous to other carbohydrate derivative transporters, such as triose phosphate, phosphoenolpyruvate, or UDP-glucuronic acid transporters and two loci with homologues for the brittle 1 protein. By contrast, in leaves, carbon-fixation-specific genes such as plastidic aldolase, fructose-1,6-biphosphatase and distinct leaf isoforms of starch synthase, starch branching enzyme, starch phosphorylase and ADP-glucose pyrophosphorylase were upregulated. Of particular interest is the difference in tuber expression of enzymes involved in the hydrolytic and phosphorolytic starch degradation pathways. Considerably greater levels of α-amylase (10–25-fold) and β-amylase (5–10-fold) mRNAs were found in DM tubers compared to RH, whereas α-1,4 glucan phosphorylase mRNA was equivalent in DM and RH tubers. These gene expression differences between the breeding line RH and the more primitive DM are consistent with the concept that increasing tuber yield may be partially attained by selection for decreased activity of the hydrolytic starch degradation pathway.

Recent studies using a potato genotype strictly dependent on short days for tuber induction (*S. tuberosum* group Andigena) identified a potato homologue (*SP6A*) of *A. thaliana FLOWERING LOCUS T* (*FT*) as the long-distance tuberization inductive signal. SP6A is produced in the leaves, consistent with its role as the mobile signal (S. Prat, personal communication). *SP/FT* is a multi-gene family (Supplementary Text and Supplementary Fig. 7) and expression of a second *FT* homologue, *SP5G*, in mature tubers suggests a possible function in the control of tuber sprouting, a photoperiod-dependent phenomenon[20]. Likewise, expression of a homologue of the *A. thaliana* flowering time MADS box gene *SOC1*, acting downstream of *FT*[21], is restricted to tuber sprouts (Supplementary Fig. 8). Expression of a third *FT* homologue, *SP3D*, does not correlate with tuberization induction but instead with transition to flowering, which is regulated independently of day length (S. Prat, personal communication). These data indicate that neofunctionalization of the day-length-dependent flowering control pathway has occurred in potato to control formation and possibly sprouting of a novel storage organ, the tuber (Supplementary Fig. 9).

Disease resistance

Potato is susceptible to a wide range of pests and pathogens and the identification of genes conferring disease resistance has been a major focus of the research community. Most cloned disease resistance genes in the Solanaceae encode nucleotide-binding site (NBS) and leucine-rich-repeat (LRR) domains. The DM assembly contains 408 NBS-LRR-encoding genes, 57 Toll/interleukin-1 receptor/plant R gene homology (TIR) domains and 351 non-TIR types (Supplementary Table 20), similar to the 402 resistance (*R*) gene candidates in *Populus*[22]. Highly related homologues of the cloned potato late blight resistance genes *R1*, *RB*, *R2*, *R3a*, *Rpi-blb2* and *Rpi-vnt1.1* were present in the assembly. In RH, the chromosome 5 R1 cluster contains two distinct haplotypes; one is collinear with the R1 region in DM (Supplementary Fig. 10), yet neither the DM nor the RH R1 regions are collinear with other potato R1 regions[23,24]. Comparison of the DM potato *R* gene sequences with well-established gene models (functional *R* genes) indicates that many NBS-LRR genes (39.4%) are pseudogenes owing to indels, frameshift mutations, or premature stop

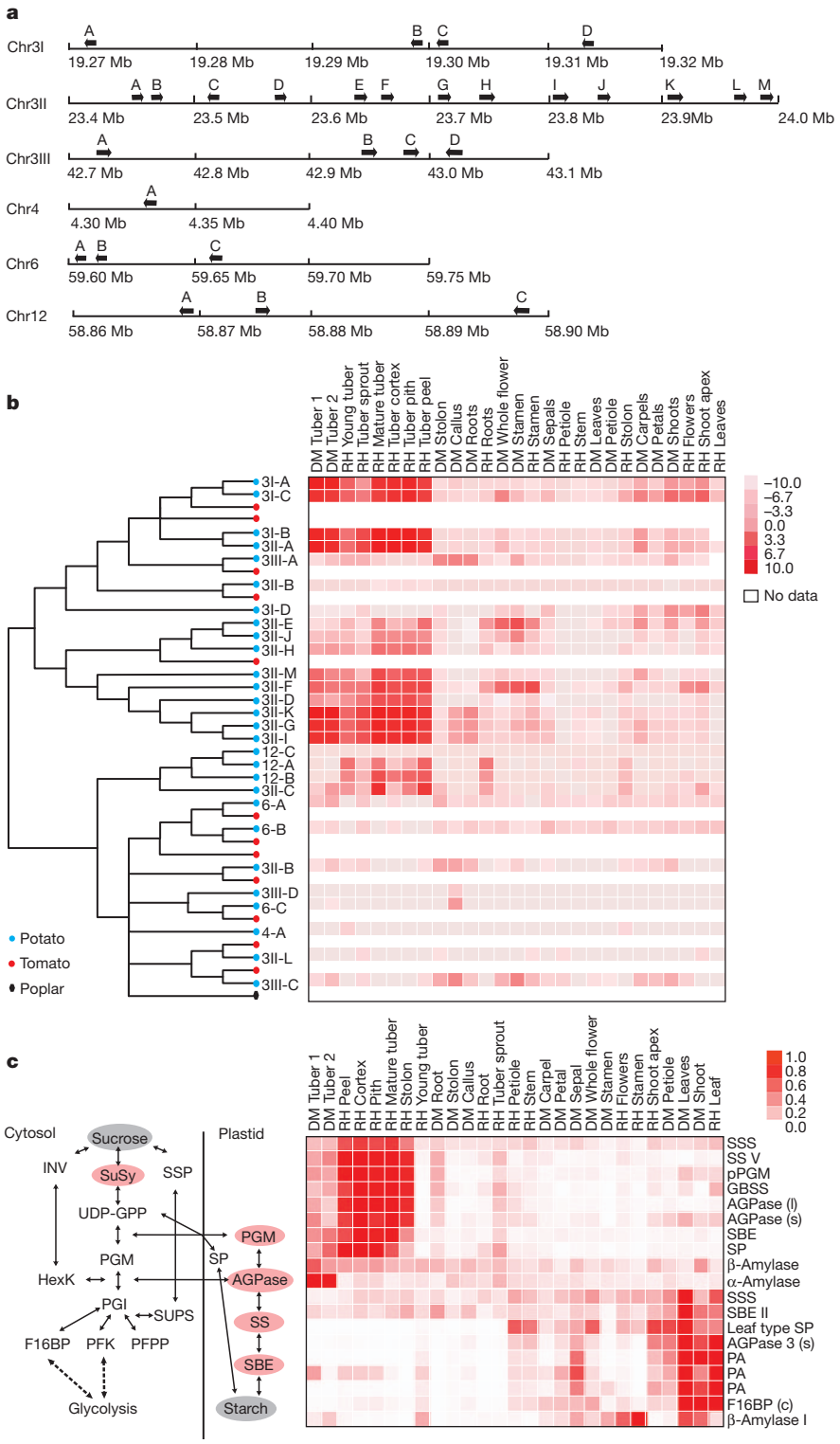


Figure 4 | Gene expression of selected tissues and genes. **a**, KTI gene organization across the potato genome. Black arrows indicate the location of individual genes on six scaffolds located on four chromosomes. **b**, Phylogenetic tree and KTI gene expression heat map. The KTI genes were clustered using all potato and tomato genes available with the *Populus* KTI gene as an out-group. The tissue specificity of individual members of the highly expanded potato gene family is shown in the heat map. Expression levels are indicated by shades of red, where white indicates no expression or lack of data for tomato and poplar. **c**, A model of starch synthesis showing enzyme activities is shown on the left. AGPase, ADP-glucose pyrophosphorylase; F16BP, fructose-1,6-biphosphatase; HexK, hexokinase; INV, invertase; PFK, phosphofructokinase; PFPP, pyrophosphate-fructose-6-phosphate-1-phosphotransferase; PGI, phosphoglucose isomerase; PGM, phosphoglucomutase; SBE, starch branching enzyme; SP, starch phosphorylase; SPP, sucrose phosphate phosphatase; SS, starch synthase; SuSy, sucrose synthase; SUPS, sucrose phosphate synthase; UDP-GPP, UDP-glucose pyrophosphorylase. The grey background denotes substrate (sucrose) and product (starch) and the red background indicates genes that are specifically upregulated in RH versus DM. On the right, a heat map of the genes involved in carbohydrate metabolism is shown. ADP-glucose pyrophosphorylase large subunit, AGPase (l); ADP-glucose pyrophosphorylase small subunit, AGPase (s); ADP-glucose pyrophosphorylase small subunit 3, AGPase 3 (s); cytosolic fructose-1,6-biphosphatase, F16BP (c); granule bound starch synthase, GBSS; leaf type L starch phosphorylase, Leaf type SP; plastidic phosphoglucomutase, pPGM; starch branching enzyme II, SBE II; soluble starch synthase, SSS; starch synthase V, SSV; three variants of plastidic aldolase, PA.

codons including the *R1*, *R3a* and *Rpi-vnt1.1* clusters that contain extensive chimaeras and exhibit evolutionary patterns of type I *R* genes[25]. This high rate of pseudogenization parallels the rapid evolution of effector genes observed in the potato late blight pathogen, *Phytophthora infestans*[26]. Coupled with abundant haplotype diversity, tetraploid potato may therefore contain thousands of *R*-gene analogues.

Conclusions and future directions

We sequenced a unique doubled-monoploid potato clone to overcome the problems associated with genome assembly due to high levels of heterozygosity and were able to generate a high-quality draft potato genome sequence that provides new insights into eudicot genome evolution. Using a combination of data from the vigorous, heterozygous diploid RH and relatively weak, doubled-monoploid DM, we could directly address the form and extent of heterozygosity in potato and provide the first view into the complexities that underlie inbreeding depression. Combined with other recent studies, the potato genome sequence may elucidate the evolution of tuberization. This evolutionary innovation evolved exclusively in the *Solanum* section *Petota* that encompasses ~200 species distributed from the southwestern United States to central Argentina and Chile. Neighbouring *Solanum* species,

including the *Lycopersicon* section, which comprises wild and cultivated tomatoes, did not acquire this trait. Both gene family expansion and recruitment of existing genes for new pathways contributed to the evolution of tuber development in potato.

Given the pivotal role of potato in world food production and security, the potato genome provides a new resource for use in breeding. Many traits of interest to plant breeders are quantitative in nature and the genome sequence will simplify both their characterization and deployment in cultivars. Whereas much genetic research is conducted at the diploid level in potato, almost all potato cultivars are tetraploid and most breeding is conducted in tetraploid material. Hence, the development of experimental and computational methods for routine and informative high-resolution genetic characterization of polyploids remains an important goal for the realization of many of the potential benefits of the potato genome sequence.

METHODS SUMMARY

DM1-3 516 R44 (DM) resulted from chromosome doubling of a monoploid ($1n = 1x = 12$) derived by anther culture of a heterozygous diploid ($2n = 2x = 24$) *S. tuberosum* group Phureja clone (PI 225669)[27]. RH89-039-16 (RH) is a diploid clone derived from a cross between a *S. tuberosum* 'dihaploid' (SUH2293) and a diploid clone (BC1034) generated from a cross between two *S. tuberosum* × *S. tuberosum* group Phureja hybrids[28] (Supplementary Fig. 11). Sequence data from three platforms, Sanger, Roche 454 Pyrosequencing, and Illumina Sequencing-by-Synthesis, were used to assemble the DM genome using the SOAPdenovo assembly algorithm[4]. The RH genotype was sequenced using shotgun sequencing of BACs and WGS in which reads were mapped to the DM reference assembly. Superscaffolds were anchored to the 12 linkage groups using a combination of *in silico* and genetic mapping data. Repeat sequences were identified through sequence similarity at the nucleotide and protein level[29]. Genes were annotated using a combined approach[30] on the repeat masked genome with *ab initio* gene predictions, protein similarity and transcripts to build optimal gene models. Illumina RNA-Seq reads were mapped to the DM draft sequence using Tophat[31] and expression levels from the representative transcript were determined using Cufflinks[32].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 11 January; accepted 3 May 2011.
Published online 10 July 2011.

1. Hijmans, R. J. Global distribution of the potato crop. *Am. J. Potato Res.* **78,** 403–412 (2001).
2. Burlingame, B., Mouillé, B. & Charrondiére, R. Nutrients, bioactive non-nutrients and anti-nutrients in potatoes. *J. Food Compost. Anal.* **22,** 494–502 (2009).
3. Paz, M. M. & Veilleux, R. E. Influence of culture medium and *in vitro* conditions on shoot regeneration in *Solanum phureja* monoploids and fertility of regenerated doubled monoploids. *Plant Breed.* **118,** 53–57 (1999).
4. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–272 (2010).
5. Arumuganathan, K. & Earle, E. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9,** 208–218 (1991).
6. Tang, X. *et al.* Assignment of genetic linkage maps to diploid *Solanum tuberosum* pachytene chromosomes by BAC-FISH technology. *Chromosome Res.* **17,** 899–915 (2009).
7. Peters, S. A. *et al. Solanum lycopersicum* cv. Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements. *Plant J.* **58,** 857–869 (2009).
8. Albach, D. C., Soltis, P. S. & Soltis, D. E. Patterns of embryological and biochemical evolution in the Asterids. *Syst. Bot.* **26,** 242–262 (2001).
9. Tang, H. *et al.* Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18,** 1944–1954 (2008).
10. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449,** 463–467 (2007).
11. Fawcett, J. A., Maere, S. & Van de Peer, Y. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proc. Natl Acad. Sci. USA* **106,** 5737–5742 (2009).
12. Lai, J. *et al.* Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature Genet.* **42,** 1027–1030 (2010).
13. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
14. Gore, M. A. *et al.* A first-generation haplotype map of maize. *Science* **326,** 1115–1117 (2009).
15. Prat, S. *et al.* Gene expression during tuber development in potato plants. *FEBS Lett.* **268,** 334–338 (1990).
16. Glaczinski, H., Heibges, A., Salamini, R. & Gebhardt, C. Members of the Kunitz-type protease inhibitor gene family of potato inhibit soluble tuber invertase *in vitro*. *Potato Res.* **45,** 163–176 (2002).
17. Shannon, J. C., Pien, F. M. & Liu, K. C. Nucleotides and nucleotide sugars in developing maize endosperms: synthesis of ADP-glucose in *brittle-1*. *Plant Physiol.* **110,** 835–843 (1996).
18. Tauberger, E. *et al.* Antisense inhibition of plastidial phosphoglucomutase provides compelling evidence that potato tuber amyloplasts import carbon from the cytosol in the form of glucose-6-phosphate. *Plant J.* **23,** 43–53 (2000).
19. Fettke, J. *et al.* Glucose 1-phosphate is efficiently taken up by potato (*Solanum tuberosum*) tuber parenchyma cells and converted to reserve starch granules. *New Phytol.* **185,** 663–675 (2010).
20. Sonnewald, U. Control of potato tuber sprouting. *Trends Plant Sci.* **6,** 333–335 (2001).
21. Yoo, S. K. *et al. CONSTANS* activates *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* through *FLOWERING LOCUS T* to promote flowering in *Arabidopsis*. *Plant Physiol.* **139,** 770–778 (2005).
22. Kohler, A. *et al.* Genome-wide identi?cation of *NBS* resistance genes in *Populus trichocarpa*. *Plant Mol. Biol.* **66,** 619–636 (2008).
23. Ballvora, A. *et al.* Comparative sequence analysis of *Solanum* and *Arabidopsis* in a hot spot for pathogen resistance on potato chromosome V reveals a patchwork of conserved and rapidly evolving genome segments. *BMC Genomics* **8,** 112 (2007).
24. Kuang, H. *et al.* The *R1* resistance gene cluster contains three groups of independently evolving, type I *R1* homologues and shows substantial structural variation among haplotypes of *Solanum demissum*. *Plant J.* **44,** 37–51 (2005).
25. Kuang, H., Woo, S. S., Meyers, B. C., Nevo, E. & Michelmore, R. W. Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. *Plant Cell* **16,** 2870–2894 (2004).
26. Haas, B. J. *et al.* Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* **461,** 393–398 (2009).
27. Haynes, F. L. in *Prospects for the Potato in the Developing World: an International Symposium on Key Problems and Potentials for Greater Use of the Potato in the Developing World* (ed. French, E. R.) 100–110 (International Potato Center (CIP), 1972).
28. van Os, H. *et al.* Construction of a 10,000-marker ultradense genetic recombination map of potato: providing a framework for accelerated gene isolation and a genomewide physical map. *Genetics* **173,** 1075–1087 (2006).
29. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **25,** 4.10.1–4.10.14 (2004).
30. Elsik, C. G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8,** R13 (2007).
31. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
32. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28,** 511–515 (2010).
33. Li, L., Stoeckert, C. J. Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge the assistance of W. Amoros, B. Babinska, R. V. Baslerov, B. K. Bumazhkin, M. F. Carboni, T. Conner, J. Coombs, L. Daddiego, J. M. D'Ambrosio, G. Diretto, S. B. Divito, D. Douches, M. Filipiak, G. Gianese, R. Hutten, E. Jacobsen, E. Kalinska, S. Kamoun, D. Kells, H. Kossowska, L. Lopez, M. Magallanes-Lundback, T. Miranda, P. S. Naik, A. N. Panteleeva, D. Pattanayak, E. O. Patutina, M. Portantier, S. Rawat, R. Simon, B. P. Singh, B. Singh, W. Stiekema, M. V. Sukhacheva and C. Town in providing plant material, generating data, annotation, analyses, and discussions. We are indebted to additional faculty and staff of the BGI-Shenzhen, J Craig Venter Institute, and MSU Research Technology Support Facility who contributed to this project. Background and preliminary data were provided by the Centre for BioSystems Genomics (CBSG), EU-project (APOPHYS EU-QLRT-2001-01849) and US Department of Agriculture National Institute of Food and Agriculture SolCAP project (2008-55300-04757 and 2009-85606-05673). We acknowledge the funding made available by the "863" National High Tech Research Development Program in China (2006AA100107), "973" National Key Basic Research Program in China (2006CB101904, 2007CB815703, 2007CB815705, 2009CB119000), Board of Wageningen University and Research Centre, CAPES - Brazilian Ministry of Education, Chinese Academy of Agricultural Sciences (seed grant to S.H.), Chinese Ministry of Agriculture (The "948" Program), Chinese Ministry of Finance (1251610601001), Chinese Ministry of Science and Technology (2007DFB30080), China Postdoctoral Science Foundation (20070420446 to Z.Z.), CONICET (Argentina), DAFF Research Stimulus Fund (07-567), CONICYT-Chile (PBCT-PSD-03), Danish Council for Strategic Research Programme Commission on Health, Food and Welfare (2101-07-0116), Danish Council for Strategic Research Programme Commission on Strategic Growth Technologies (Grant 2106-07-0021), FINCyT ((099-FINCyT-EQUIP-2009)/(076-FINCyT-PIN-2008), Préstamo BID no. 1663/OC-PE, FONDAP and BASAL-CMM), Fund for Economic Structural Support (FES), HarvestPlus Challenge Program, Indian Council of Agricultural Research, INIA-Ministry of Agriculture of Chile, Instituto Nacional de Innovacion Agraria-Ministry of Agriculture of Peru, Instituto Nacional de Tecnología Agropecuaria (INTA), Italian Ministry of Research (Special Fund for Basic Research), International Potato Center (CIP-CGIAR core funds), LBMG of Center for Genome Regulation and Center for Mathematical Modeling, Universidad de Chile (UMI 2807 CNRS), Ministry of Education and Science of Russia (contract 02.552.11.7073), National Nature Science Foundation of China (30671319, 30725008, 30890032, 30971995), Natural Science Foundation of Shandong Province in China (Y2006D21), Netherlands Technology Foundation (STW), Netherlands Genomics Initiative (NGI), Netherlands Ministries of Economic Affairs (EZ) and Agriculture (LNV), New Zealand Institute for Crop & Food Research Ltd

Strategic Science Initiative, Perez Guerrero Fund, Peruvian Ministry of Agriculture-Technical Secretariat of coordination with the CGIAR, Peruvian National Council of Science and Technology (CONCYTEC), Polish Ministry of Science and Higher Education (47/PGS/2006/01), Programa Cooperativo para el Desarrollo Tecnológico Agroalimentario y Agroindustrial del Cono Sur (PROCISUR), Project Programa Bicentenario de Ciencia y Tecnología - Conicyt, PBCT - Conicyt PSD-03, Russian Foundation for Basic Research (09-04-12275), Secretaría de Ciencia y Tecnología (SECyT) actual Ministerio de Ciencia y Tecnología (MINCyT), Argentina, Shenzhen Municipal Government of China (CXB200903110066A, ZYC200903240077A, ZYC200903240076A), Solexa project (272-07-0196), Special Multilateral Fund of the Inter-American Council for Integral Development (FEMCIDI), Teagasc, Teagasc Walsh Fellowship Scheme, The New Zealand Institute for Plant & Food Research Ltd Capability Fund, UK Potato Genome Sequencing grant (Scottish Government Rural and Environment Research and Analysis Directorate (RERAD), Department for Environment, Food and Rural Affairs (DEFRA), Agriculture and Horticulture Development Board - Potato Council), UK Biotechnology and Biological Sciences Research Council (Grant BB/F012640), US National Science Foundation Plant Genome Research Program (DBI-0604907 /DBI-0834044), Virginia Agricultural Experiment Station USDA Hatch Funds (135853), and Wellcome Trust Strategic award (WT 083481).

Author Contributions A.D.G., A.G., A.N.M., A.V.B., A.V.M., B.B.K., B.K., B.R.W., B.S., B.t.L.H., B.V., B.X., B.Z., C.L., C.R.B., C.W.B.B., D.F.M., D. Martinez, D. Milbourne, D.M.A.M., D.M.B., D.D., D.M., E.D., F.G., G.A.M., G.A.T., G.D.I.C., G.G., G.J. Bishop, G.J. Bryan, G.L., G.O., G.P., G.Z., H.K., H.L., H.v.E., I.N., J.d.B., J.G., J.H., J.J., J.M.E.J., J.W., J.X., K.L.N., K.O'B., L.D., L.E.B., M.B., M.D., M.d.R.H., M.F., M. Geoffroy, M.Ghislain, M.I., M.P., M.S., M.T., N.M., N.V.R., O.P., P.F., P.N., P.S., Q.H., R.C.H.J.v.H., R.E.V., R.G., R.G.F.V., R. Lozano, R. Li, S.C., S.E.F., S.H., S.J.T., S.K.C., S.K.S., S.L., S.P., S.Y., T.B., T.V.K., V.U.P., X. Xiong, X. Xu, Y.D., Y.H., Y.L., Y.Y., Y.Z. and Z.Z. were involved in experimental design, data generation and/or data analysis. A.N.M., B.K., C.R.B., C.W.B.B., D.D., D. Milbourne, D.M.A.M., D.M.B., E.D., G.G., G.J. Bishop, G.J. Bryan, G.O., H.L., I.N., J.d.B., J.J., J.M.E.J., K.L.N., M.B., M.F., M.D., M.S., O.P., R.C.H.J.v.H., R.E.V., R.G.F.V., R. Lozano, R.W., S.E.F., S.H., S.J.T., S.K.S., T.B. and X. Xu wrote the manuscript. B.S., C.R.B., C.W.B.B., D.F.M., D. Milbourne, D.M.A.M., D.Q., G.G., G.J. Bishop, G.J. Bryan, G.O., G.P., J.M.E.J., J.W., K.G.S., R.G.F.V., R. Li, R.W., S.E.F., S.H., S.K.C., S.Y., W.Z. and Y.D. supervised data generation/analysis and managed the project. C.R.B., C.W.B.B., G.J. Bryan, G.O., J.M.E.J. and S.H. are members of The Potato Genome Sequencing Consortium Steering Committee.

Author Information BAC and fosmid end sequences have been deposited in the GSS division of GenBank (BAC: GS025503–GS026177, GS262924–GS365942, GS504213–GS557003; fosmid: FI900795–FI901529, FI907952–FI927051, GS557234–GS594339, GS635316–GS765761). DM Illumina GA2 WGS and Roche 454 sequences have been deposited in the NCBI Sequence Read Archive (SRA029323) and EBI Short Read Archive (ERP000411) respectively. RH NGS sequences have been deposited in the EBI Short Read Archive (ERP000627). DM and RH RNA-Seq reads have been deposited in the NCBI Sequence Read Archive (SRA030516; study SRP005965) and the European Nucleotide Database ArrayExpress Database (E-MTAB-552; study ERP000527), respectively. The DM Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AEWC00000000. The version described in this paper is the First Version, AEWC01000000. Genome sequence and annotation can be obtained and viewed at http://potatogenome.net. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.H. (huangsanwen@caas.net.cn), C.R.B. (buell@msu.edu) or R.G.F.V. (Richard.Visser@wur.nl).

The Potato Genome Consortium (Participants are listed alphabetically by institution.)

BGI-Shenzhen Xun Xu[1], Shengkai Pan[1], Shifeng Cheng[1], Bo Zhang[1], Desheng Mu[1], Peixiang Ni[1], Gengyun Zhang[1], Shuang Yang (Principal Investigator)[1], Ruiqiang Li (Principal Investigator)[1], Jun Wang (Principal Investigator)[1]; **Cayetano Heredia University** Gisella Orjeda (Principal Investigator)[2], Frank Guzman[2], Michael Torres[2], Roberto Lozano[2], Olga Ponce[2], Diana Martinez[2], Germán De la Cruz[2]; **Central Potato Research Institute** S. K. Chakrabarti (Principal Investigator)[3], Virupaksh U. Patil[3]; **Centre Bioengineering RAS** Konstantin G. Skryabin (Principal Investigator)[4], Boris B. Kuznetsov[4], Nikolai V. Ravin[4], Tatjana V. Kolganova[4], Alexey V. Beletsky[4], Andrei V. Mardanov[4]; **CGR-CMM, Universidad de Chile** Alex Di Genova[5]; **College of Life Sciences, University of Dundee** Daniel M. Bolser[6], David M. A. Martin (Principal Investigator)[6]; **High Technology Research Center, Shandong Academy of Agricultural Sciences** Guangcun Li[7], Yu Yang[7]; **Huazhong Agriculture University** Hanhui Kuang[8], Qun Hu[8]; **Hunan Agricultural University** Xingyao Xiong[9]; **Imperial College London** Gerard J. Bishop[10]; **Instituto de Investigaciones Agropecuarias** Boris Sagredo (Principal Investigator)[11], Nilo Mejía[11]; **Institute of Biochemistry & Biophysics** Wlodzimierz Zagorski (Principal Investigator)[12], Robert Gromadka[12], Jan Gawor[12], Pawel Szczesny[12]; **Institute of Vegetables & Flowers, Chinese Academy of Agricultural Sciences** Sanwen Huang (Principal Investigator)[13], Zhonghua Zhang[13], Chunbo Liang[13], Jun He[13], Ying Li[13], Ying He[13], Jianfei Xu[13], Youjun Zhang[13], Binyan Xie[13], Yongchen Du[13], Dongyu Qu (Principal Investigator)[13]; **International Potato Center** Merideth Bonierbale[14], Marc Ghislain[14], Maria del Rosario Herrera[14]; **Italian National Agency for New Technologies, Energy & Sustainable Development** Giovanni Giuliano (Principal Investigator)[15], Marco Pietrella[15], Gaetano Perrotta[15], Paolo Facella[15]; **J Craig Venter Institute** Kimberly O'Brien[16]; **Laboratorio de Agrobiotecnología, Instituto Nacional de Tecnología Agropecuaria** Sergio E. Feingold (Principal Investigator)[17], Leandro E. Barreiro[17], Gabriela A. Massa[17]; **Laboratorio de Biología de Sistemas, Universidad Nacional de LaPlata** Luis Diambra[18]; **Michigan State University** Brett R. Whitty[19], Brieanne Vaillancourt[19], Haining Lin[19], Alicia N. Massa[19], Michael Geoffroy[19], Steven Lundback[19], Dean DellaPenna[19], C. Robin Buell (Principal Investigator)[19]; **Scottish Crop Research Institute** Sanjeev Kumar Sharma[20]†, David F. Marshall[20]†, Robbie Waugh[20]†, Glenn J. Bryan (Principal Investigator)[20]†; **Teagasc Crops Research Centre** Marialaura Destefanis[21], Istvan Nagy[21], Dan Milbourne (Principal Investigator)[21]; **The New Zealand Institute for Plant & Food Research Ltd** Susan J. Thomson[22], Mark Fiers[22], Jeanne M. E. Jacobs (Principal Investigator)[22]; **University of Aalborg** Kåre L. Nielsen (Principal Investigator)[23], Mads Sønderkær[23]; **University of Wisconsin** Marina Iovene[24], Giovana A. Torres[24], Jiming Jiang (Principal Investigator)[24]; **Virginia Polytechnic Institute & State University** Richard E. Veilleux[25]; **Wageningen University & Research Centre** Christian W. B. Bachem (Principal Investigator)[26], Jan de Boer[26], Theo Borm[26], Bjorn Kloosterman[26], Herman van Eck[26], Erwin Datema[27], Bas te Lintel Hekkert[27], Aska Goverse[28,29], Roeland C. H. J. van Ham[27,28] & Richard G. F. Visser[26,28]

[1]BGI-Shenzhen, Chinese Ministry of Agricultural, Key Lab of Genomics, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China. [2]Cayetano Heredia University, Genomics Research Unit, Av Honorio Delgado 430, Lima 31, Peru and San Cristobal of Huamanga University, Biotechnology and Plant Genetics Laboratory, Ayacucho, Peru. [3]Central Potato Research Institute, Shimla 171001, Himachal Pradesh, India. [4]Centre Bioengineering RAS, Prospekt 60-letya Oktyabrya, 7-1, Moscow 117312, Russia. [5]Center for Genome Regulation and Center for Mathematical Modeling, Universidad de Chile (UMI 2807 CNRS), Chile. [6]College of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, UK. [7]High Technology Research Center, Shandong Academy of Agricultural Sciences, 11 Sangyuan Road, Jinan 250100, P. R. China. [8]Huazhong Agriculture University, Ministry of Education, College of Horticulture and Forestry, Department of Vegetable Crops, Key Laboratory of Horticulture Biology, Wuhan 430070, P. R. China. [9]Hunan Agricultural University, College of Horticulture and Landscape, Changsha, Hunan 410128, China. [10]Imperial College London, Division of Biology, South Kensington Campus, London SW7 1AZ, UK. [11]Instituto de Investigaciones Agropecuarias, Avda. Salamanca s/n, Km 105 ruta 5 sur, sector Los Choapinos. Rengo, Región del Libertador Bernardo O'Higgins, Código Postal 2940000, Chile. [12]Institute of Biochemistry and Biophysics, DNA Sequencing and Oligonucleotides Synthesis Laboratory, PAS ul. Pawinskiego 5a, 02-106 Warsaw, Poland. [13]Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Key Laboratory of Horticultural Crops Genetic Improvement of Ministry of Agriculture, Sino-Dutch Joint Lab of Horticultural Genomics Technology, Beijing 100081, China. [14]International Potato Center, P.O. Box 1558, Lima 12, Peru. [15]Italian National Agency for New Technologies, Energy and Sustainable Development (ENEA), Casaccia Research Center, Via Anguillarese 301, 00123 Roma, Italy and Trisaia Research Center, S.S. 106 Ionica - Km 419,50 75026 Rotondella (Matera), Italy. [16]J Craig Venter Institute, 9712 Medical Center Dr, Rockville, Maryland 20850, USA. [17]Laboratorio de Agrobiotecnología, Estación Experimental Agropecuaria Balcarce, Instituto Nacional de Tecnología Agropecuaria (INTA) cc276 (7620) Balcarce, Argentina. [18]Laboratorio de Biología de Sistemas, CREG, Universidad Nacional de LaPlata, 1888, Argentina. [19]Michigan State University, East Lansing, Michigan 48824, USA. [20]Scottish Crop Research Institute, Genetics Programme, Invergowrie, Dundee DD2 5DA, UK. [21]Teagasc Crops Research Centre, Oak Park, Carlow, Ireland. [22]The New Zealand Institute for Plant & Food Research Ltd., Private Bag 4704, Christchurch 8140, New Zealand. [23]University of Aalborg (AAU), Department of Biotechnology, Chemistry and Environmental Engineering, Sohngaardsholmsvej 49, 9000 Aalborg, Denmark. [24]University of Wisconsin-Madison, Department of Horticulture, 1575 Linden Drive, Madison, Wisconsin 53706, USA. [25]Virginia Polytechnic Institute and State University, Department of Horticulture, 544 Latham Hall, Blacksburg, Virginia 24061, USA. [26]Wageningen University and Research Centre, Dept. of Plant Sciences, Laboratory of Plant Breeding, Droevendaalsesteeg 1, 6708PB Wageningen, Netherlands. [27]Wageningen University and Research Centre, Applied Bioinformatics, Plant Research International, Droevendaalsesteeg 1, 6708PB Wageningen, Netherlands. [28]Centre for BioSystems Genomics, Droevendaalsesteeg 1, 6708PB Wageningen, Netherlands. [29]Wageningen University and Research Centre, Dept. of Plant Sciences, Laboratory of Nematology, Droevendaalsesteeg 1, 6708PB Wageningen, Netherlands. †Present address: The James Hutton Institute, Invergowrie, Dundee, DD2 5DA, UK (S.K.S., D.F.M., R.W., G.J. Bryan).

METHODS

DM whole-genome shotgun sequencing and assembly. Libraries were constructed from DM genomic DNA and sequenced on the Sanger, Illumina Genome Analyser 2 (GA2) and Roche 454 platforms using standard protocols (see Supplementary Text). A BAC library and three fosmid libraries were end sequenced using the Sanger platform. For the Illumina GA2 platform, we generated 70.6 Gb of 37–73 bp paired-end reads from 16 libraries with insert lengths of 200–811 bp (Supplementary Tables 21 and 22). We also generated 18.7 Gb of Illumina mate-pair libraries (2, 5 and 10 kb insert size). In total, 7.2 Gb of 454 single-end data were generated and applied for gap filling to improve the assembly, of which 4.7 Gb (12,594,513 reads) were incorporated into the final assembly. For the 8 and 20 kb 454 paired-end reads, representing 0.7 and 1.0 Gb of raw data respectively, 90.7 Mb (511,254 reads) and 211 Mb (1,525,992 reads), respectively, were incorporated into the final assembly.

We generated a high-quality potato genome using the short read assembly software SOAPdenovo[4] (Version 1014). We first assembled 69.4 Gb of GA2 paired-end short reads into contigs, which are sequence assemblies without gaps composed of overlapping reads. To increase the assembly accuracy, only 78.3% of the reads with high quality were considered. Then contigs were further linked into scaffolds by paired-end relationships (~300 to ~550 bp insert size), mate-pair reads (2 to approximately 10 kb), fosmid ends (~40 kb, 90,407 pairs of end sequences) and BAC ends (~100 kb, 71,375 pairs of end sequences). We then filled gaps with the entire short-read data generated using Illumina GA2 reads. The primary contig N_{50} size (the contig length such that using equal or longer contigs produces half of the bases of the assembled genome) was 697 bp and increased to 1,318 kb after gap-filling (Supplementary Tables 23 and 24). When only the paired-end relationships were used in the assembly process, the N_{50} scaffold size was 22.4 kb. Adding mate-pair reads with 2, 5 and 10 kb insert sizes, the N_{50} scaffold size increased to 67, 173 and 389 kb, respectively. When integrated with additional libraries of larger insert size, such as fosmid and BAC end sequences, the N_{50} reached 1,318 kb. The final assembly size was 727 Mb, 93.87% of which is non-gapped sequence. We further filled the gaps with 6.74 fold coverage of 454 data, which increased the N_{50} contig size to 31,429 bp with 15.4% of the gaps filled.

The single-base accuracy of the assembly was estimated by the depth and proportion of disconcordant reads. For the DM v3.0 assembly, 95.45% of 880 million usable reads could be mapped back to the assembled genome by SOAP 2.20 (ref. 34) using optimal parameters. The read depth was calculated for each genomic location and peak depth for whole genome and the CDS regions are 100 and 105, respectively. Approximately 96% of the assembled sequences had more than 20-fold coverage (Supplementary Fig. 1). The overall GC content of the potato genome is about 34.8% with a positive correlation between GC content and sequencing depth (data not shown). The DM potato should have few heterozygous sites and 93.04% of the sites can be supported by at least 90% reads, suggesting high base quality and accuracy.

RH genome sequencing. Whole-genome sequencing of genotype RH was performed on the Illumina GA2 platform using a variety of fragment sizes and reads lengths resulting in a total of 144 Gb of raw data (Supplementary Table 25). These data were filtered using a custom C program and assembled using SOAPdenovo 1.03 (ref. 4). Additionally, four 20-kb mate-pair libraries were sequenced on a Roche 454 Titanium sequencer, amounting to 581 Mb of raw data (Supplementary Table 26). The resulting sequences were filtered for duplicates using custom Python scripts.

The RH BACs were sequenced using a combination of Sanger and 454 sequencing at various levels of coverage (Supplementary Tables 9–11). Consensus base calling errors in the BAC sequences were corrected using custom Python and C scripts using a similar approach to that described previously[35] (Supplementary Text). Sequence overlaps between BACs within the same physical tiling path were identified using megablast from BLAST 2.2.21 (ref. 36) and merged with megamerger from the EMBOSS 6.1.0 package[37]. Using the same pipeline, several kilobase-sized gaps were closed through alignment of a preliminary RH whole-genome assembly. The resulting non-redundant contigs were scaffolded by mapping the RH whole-genome Illumina and 454 mated sequences against these contigs using SOAPalign 2.20 (ref. 34) and subsequently processing these mapping results with a custom Python script. The scaffolds were then ordered into superscaffolds based on the BAC order in the tiling paths of the FPC map. This procedure removed 25 Mb of redundant sequence, reduced the number of sequence fragments from 17,228 to 3,768, and increased the N50 sequence length from 24 to 144 kb (Supplementary Tables 9 and 10).

Construction of the DM genetic map and anchoring of the genome. To anchor and fully orientate physical contigs along the chromosome, a genetic map was developed *de novo* using sequence-tagged-site (STS) markers comprising simple sequence repeats (SSR), SNPs, and diversity array technology (DArT). SSR and SNP markers were designed directly from assembled sequence scaffolds, whereas polymorphic DArT marker sequences were searched against the scaffolds for high-quality unique matches. A total of 4,836 STS markers including 2,174 DArTs, 2,304 SNPs and 358 SSRs were analysed on 180 progeny clones from a backcross population ((DM × DI) × DI) developed at CIP between DM and DI (CIP no. 703825), a heterozygous diploid *S. tuberosum* group Stenotomum (formerly *S. stenotomum* ssp. *goniocalyx*) landrace clone. The data from 2,603 polymorphic STS markers comprising 1,881 DArTs, 393 SNPs and 329 SSR alleles were analysed using JoinMap 4 (ref. 38) and yielded the expected 12 potato linkage groups. Supplementary Fig. 3 represents the mapping and anchoring of the potato genome, using chromosome 7 as an example.

Anchoring the DM genome was accomplished using direct and indirect approaches. The direct approach employed the ((DM × DI) × DI) linkage map whereby 2,037 of the 2,603 STS markers comprised of 1,402 DArTs, 376 SNPs and 259 SSRs could be uniquely anchored on the DM superscaffolds. This approach anchored ~52% (394 Mb) of the assembly arranged into 334 superscaffolds (Supplementary Table 27 and Supplementary Fig. 3).

RH is the male parent of the mapping population of the ultra-high-density (UHD) linkage map[28] used for construction and genetic anchoring of the physical map using the RHPOTKEY BAC library[39]. The indirect mapping approach exploited *in silico* anchoring using the RH genetic and physical map[28,40], as well as tomato genetic map data from SGN (http://solgenomics.net/). Amplified fragment length polymorphism markers from the RH genetic map were linked to DM sequence scaffolds via BLAST alignment[36] of whole-genome-profiling sequence tags[41] obtained from anchored seed BACs in the RH physical map, or by direct alignment of fully sequenced RH seed BACs to the DM sequence. The combined marker alignments were processed into robust anchor points. The tomato sequence markers from the genetic maps were aligned to the DM assembly using SSAHA2 (ref. 42). Positions of ambiguously anchored superscaffolds were manually checked and corrected. This approach anchored an additional ~32% of the assembly (229 Mb). In 294 cases, the two independent approaches provided direct support for each other, anchoring the same scaffold to the same position on the two maps.

Overall, the two strategies anchored 649 superscaffolds to approximate positions on the genetic map of potato covering a length of 623 Mb. The 623 Mb (~86%) anchored genome includes ~90% of the 39,031 predicted genes. Of the unanchored superscaffolds, 84 were found in the N90 (622 scaffolds greater than 0.25 Mb), constituting 17 Mb of the overall assembly or 2% of the assembled genome. The longest anchored superscaffold is 7 Mb (from chromosome 1) and the longest unanchored superscaffold is 2.5 Mb.

Identification of repetitive sequences. Transposable elements (TEs) in the potato genome assembly were identified at the DNA and protein level. RepeatMasker[29] was applied using Repbase[43] for TE identification at the DNA level. At the protein level, RepeatProteinMask[29,44] was used in a WuBlastX[36] search against the TE protein database to further identify TEs. Overlapping TEs belonging to the same repeat class were collated, and sequences were removed if they overlapped >80% and belonged to different repeat classes.

Gene prediction. To predict genes, we performed *ab initio* predictions on the repeat-masked genome and then integrated the results with spliced alignments of proteins and transcripts to genome sequences using GLEAN[30]. The potato genome was masked by identified repeat sequences longer than 500 bp, except for miniature inverted repeat transposable elements which are usually found near genes or inside introns[45]. The software Augustus[46] and Genscan[47] was used for *ab initio* predictions with parameters trained for *A. thaliana*. For similarity-based gene prediction, we aligned the protein sequences of four sequenced plants (*A. thaliana*, *Carica papaya*, *V. vinifera* and *Oryza sativa*) onto the potato genome using TBLASTN with an *E*-value cut-off of 1×10^{-5}, and then similar genome sequences were aligned against the matching proteins using Genewise[48] for accurately spliced alignments. In EST-based predictions, EST sequences of 11 *Solanum* species were aligned against the potato genome using BLAT (identity ≥0.95, coverage ≥0.90) to generate spliced alignments. All these resources and prediction approaches were combined by GLEAN[30] to build the consensus gene set. To finalize the gene set, we aligned the RNA-Seq from 32 libraries, of which eight were sequenced with both single- and paired-end reads, to the genome using Tophat[31] and the alignments were then used as input for Cufflinks[32] using the default parameters. Gene, transcript and peptide sets were filtered to remove small genes, genes modelled across sequencing gaps, TE-encoding genes, and other incorrect annotations. The final gene set contains 39,031 genes with 56,218 protein-coding transcripts, of which 52,925 nonidentical proteins were retained for analysis.

Transcriptome sequencing. RNA was isolated from many tissues of DM and RH that represent developmental, abiotic stress and biotic stress conditions (Supplementary Table 4 and Supplementary Text). cDNA libraries were constructed (Illumina) and sequenced on an Illumina GA2 in the single- and/or paired-end

mode. To represent the expression of each gene, we selected a representative transcript from each gene model by selecting the longest CDS from each gene. The aligned read data were generated by Tophat[31] and the selected transcripts used as input into Cufflinks[32], a short-read transcript assembler that calculates the fragments per kb per million mapped reads (FPKM) as expression values for each transcript. Cufflinks was run with default settings, with a maximum intron length of 15,000. FPKM values were reported and tabulated for each transcript (Supplementary Table 19).

Comparative genome analyses. Paralogous and orthologous clusters were identified using OrthoMCL[49] using the predicted proteomes of 11 plant species (Supplementary Table 28). After removing 1,602 TE-related genes that were not filtered in earlier annotation steps, asterid-specific and potato-lineage-specific genes were identified using the initial OrthoMCL clustering followed by BLAST searches (*E*-value cut-off of 1×10^{-5}) against assemblies of ESTs available from the PlantGDB project (http://plantgdb.org; 153 nonasterid species and 57 asterid species; Supplementary Fig. 5 and Supplementary Table 29). Analysis of protein domains was performed using the Pfam hmm models identified by InterProScan searches against InterPro (http://www.ebi.ac.uk/interpro). We compared the Pfam domains of the asterid-specific and potato-lineage-specific sets with those that are shared with at least one other nonasterid genome or transcriptome. A Fisher's exact test was used to detect significant differences in Pfam representation between protein sets.

After removing the self and multiple matches, the syntenic blocks ($\geq$5 genes per block) were identified using MCscan[9] and i-adhore 3.0 (ref. 50) based on the aligned protein gene pairs (Supplementary Table 8). For the self-aligned results, each aligned block represents the paralogous segments pair that arose from the genome duplication whereas, for the inter-species alignment results, each aligned block represents the orthologous pair derived from the shared ancestor. We calculated the 4DTv (fourfold degenerate synonymous sites of the third codons) for each gene pair from the aligned block and give a distribution for the 4DTv value to estimate the speciation or WGD event that occurred in evolutionary history.

Identification of disease resistance genes. Predicted open reading frames (ORFs) from the annotation of *S. tuberosum* group Phureja assembly V3 were screened using HMMER V.3 (http://hmmer.janelia.org/software) against the raw hidden Markov model (HMM) corresponding to the Pfam NBS (NB-ARC) family (PF00931). The HMM was downloaded from the Pfam home page (http://pfam.sanger.ac.uk/). The analysis using the raw HMM of the NBS domain resulted in 351 candidates. From these, a high quality protein set ($<1 \times 10^{-60}$) was aligned and used to construct a potato-specific NBS HMM using the module 'hmmbuild'. Using this new potato-specific model, we identified 500 NBS-candidate proteins that were individually analysed. To detect TIR and LRR domains, Pfam HMM searches were used. The raw TIR HMM (PF01582) and LRR 1 HMM (PF00560) were downloaded and compared against the two sets of NBS-encoding amino acid sequences using HMMER V3. Both TIR and LRR domains were validated using NCBI conserved domains and multiple expectation maximization for motif elicitation (MEME)[51]. In the case of LRRs, MEME was also useful to detect the number of repeats of this particular domain in the protein. As previously reported[52], Pfam analysis could not identify the CC motif in the N-terminal region. CC domains were thus analysed using the MARCOIL[53] program with a threshold probability of 90 (ref. 52) and double-checked using paircoil2 (ref. 54) with a *P*-score cut-off of 0.025 (ref. 55). Selected genes ($\pm$1.5 kb) were searched using BLASTX against a reference *R*-gene set[56] to find a well-characterized homologue. The reference set was used to select and annotate as pseudogenes those peptides that had large deletions, insertions, frameshift mutations, or premature stop codons. DNA and protein comparisons were used.

Haplotype diversity analysis. RH reads generated by the Illumina GA2 were mapped onto the DM genome assembly using SOAP2.20 (ref. 34) allowing at most four mismatches and SNPs were called using SOAPsnp. Q20 was used to filter the SNPs owing to sequencing errors. To exclude SNP calling errors caused by incorrect alignments, we excluded adjacent SNPs separated by $<$5 bp. SOAPindel was used to detect the indels between DM and RH. Only indels supported by more than three uniquely mapped reads were retained. Owing to the heterozygosity of RH, the SNPs and indels were classified into heterozygous and homozygous SNPs or indels.

On the basis of the annotated genes in the DM genome assembly, we extracted the SNPs located at coding regions and stop codons. If a homozygous SNP in RH within a coding region induced a premature stop codon, we defined the gene harbouring this SNP as a homozygous premature stop gene in RH. If the SNP inducing a premature stop codon was heterozygous, the gene harbouring this SNP was considered a heterozygous premature stop codon gene in RH. In addition, both categories can be further divided into premature stop codons shared with DM or not shared with DM. As a result, the numbers of premature stop codons are 606 homozygous PS genes in RH, 1,760 heterozygous PS genes in RH but not shared with DM, 288 PS in DM only, and 652 heterozygous premature stop codons in RH and shared by DM.

To identify genes with frameshift mutations in RH, we identified all the genes containing indels of which the length could not be divided by 3. We found 80 genes with frameshift mutations, of which 31 were heterozygous and 49 were homozygous.

To identify DM-specific genes, we mapped all the RH Illumina GA2 reads to the DM genome assembly. If the gene was not mapped to any RH read, it was considered a DM-specific gene. We identified 35 DM-specific genes, 11 of which are supported by similarity to entries in the KEGG database[57]. To identify RH-specific genes, we assembled the RH Illumina GA2 reads that did not map to the DM genome into RH-specific scaffolds. Then, these scaffolds were annotated using the same strategy as for DM. To exclude contamination, we aligned the CDS sequences against the protein set of bacteria with the *E*-value cut-off of 1×10^{-5} using Blastx. CDS sequences with $>$90% identity and $>$90% coverage were considered contaminants and were excluded. In addition, all DM RNA-seq reads were mapped onto the CDS sequences, and CDS sequences with homologous reads were excluded because these genes may be due to incorrect assembly. In total, we predicted 246 RH specific genes, 34 of which are supported by Gene Ontology annotation[17].

34. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25,** 1966–1967 (2009).
35. Chaisson, M., Pevzner, P. & Tang, H. Fragment assembly with short reads. *Bioinformatics* **20,** 2067–2074 (2004).
36. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
37. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277 (2000).
38. Van Ooijen, J. W. in *JoinMap 4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations* (ed. Kyazma, B. V.) (Wageningen, 2006).
39. Borm, T. J. *Construction and Use of a Physical Map of Potato.* PhD thesis, Wageningen Univ. (2008).
40. Visser, R. G. F. *et al.* Sequencing the potato genome: outline and first results to come from the elucidation of the sequence of the world's third most important crop. *Am. J. Potato Res.* **86,** 417–429 (2009).
41. Van der Vossen, E. *et al.* in *Whole Genome Profiling of the Diploid Potato Clone RH89-039-16* (Plant & Animal Genomes XVIII Conference, 2010).
42. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11,** 1725–1729 (2001).
43. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462–467 (2005).
44. Jiang, Z., Hubley, R., Smit, A. & Eichler, E. E. *DupMasker*: a tool for annotating primate segmental duplications. *Genome Res.* **18,** 1362–1368 (2008).
45. Kuang, H. *et al.* Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: new functional implications for MITEs. *Genome Res.* **19,** 42–56 (2009).
46. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32,** W309–W312 (2004).
47. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
48. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14,** 988–995 (2004).
49. Chen, F., Mackey, A. J., Vermunt, J. K. & Roos, D. S. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE* **2,** e383 (2007).
50. Simillion, C., Janssens, K., Sterck, L. & Van de Peer, Y. i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics* **24,** 127–128 (2008).
51. Bailey, T. L. & Elkan, C. The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3,** 21–29 (1995).
52. Mun, J. H., Yu, H. J., Park, S. & Park, B. S. Genome-wide identification of NBS-encoding resistance genes in *Brassica rapa. Mol. Genet. Genomics* **282,** 617–631 (2009).
53. Delorenzi, M. & Speed, T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics* **18,** 617–625 (2002).
54. McDonnell, A. V., Jiang, T., Keating, A. E. & Berger, B. Paircoil2: improved predictions of coiled coils from sequence. *Bioinformatics* **22,** 356–358 (2006).
55. Porter, B. W. *et al.* Genome-wide analysis of *Carica papaya* reveals a small *NBS* resistance gene family. *Mol. Genet. Genomics* **281,** 609–626 (2009).
56. Sanseverino, W. *et al.* PRGdb: a bioinformatics platform for plant resistance gene analysis. *Nucleic Acids Res.* **38,** D814–D821 (2010).
57. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32,** D277–D280 (2004).

B) Perl Programs

Throughout the course of the current PhD project, numerous algorithms were developed for analysis of next generation sequencing data. In the following, all programs mentioned are alphabetically listed with a short description. All programs are Perl command line applications. A small help menu is printed, when only entering the name of the program (shown in italics in the following table). All programs have been tested under Perl v5.14.2. Moreover, details of for settings of central commands in the R-package EgdeR (Robinson, McCarthy & Smyth, 2010) are given[42]. Programs implementing the EdgeR package have been tested under R version 2.15.0 (Team, 2012) using EgdeR version 2.6.7.

Program	Description
AddUniRefAnnotation.pl	Program to add functional annotations to headers in a FASTA sequence file. Usage: -l input file gene ID -> Uniref name (made using GetBestUniRefBLASTResult.pl) -f input protein FASTA file
CalculateTagWiseDispersions.pl	Program to calculate: mean expression, observed variance, a common dispersion estimate for the library, and tagwise dispersion estimates for each gene. The program implements the R-package EgdeR. Additional EgdeR settings, which are automatically set: prior.n = 25 / number of libraries trend = "movingave" Usage: -i input file with list of files Settings for calculation of tagwise dispersion -l number of libraries (default 3) -p prop.used (default 0.05) -n Minimum counts per million -m minimum libraries
CLCcsv2taglist.pl	Program to convert multiple CLC csv tag count in csv format into a tag list. Usage: -c cut off value (Default: 1) -i Input file containing list of input file names
CombineLibraryCounts.pl	Program to create multiple files with combined tag count from multiple tag files. Usage: -i Input file containing list of input file names in the format: Outputname1 -> input1 -> input2 -o Output filename (optional) -p print file counts [Y/N] If -i is not given Inputfile1 Inputfile2 ... InputfileN are added as subsequent arguments.
CompareSage.pl	Program to generate tag matrix for comparisons of DeepSAGE experiments Usage: -o Output filename (optional) -c Cut off value (No cutoff = 1)

[42] The EgdeR manual can be found at: http://www.bioconductor.org/packages/release/bioc/manuals/edgeR/man/edgeR.pdf

	-i Input file containing list of input file names If -i is not given Inputfile1 Inputfile2 ... InputfileN are added as subsequent arguments. OBS: If cut off value is set > 1 all count below cut off will be printed as "cut off value -1" to avoid false positives.
CreateID2nameTable.pl	Program to create a table with ID -> annotation from a FASTA file Usage: -i input file
CreatemRNAseqs.pl	Program to create a FASTA file with mRNA sequences based on annotations in a GFF file Exon sequences are extracted from a BLAST database using "GetUTRsAndExons.pl". A BLAST database with exon sequences is created using is subsequently created using the "makeblastdb" command from the BLAST+ command line application[43]. Usage: -i input GFF file -b Input nucleotide BLAST database with exon sequences
CutoffLibs.pl	Program to reduce tag lists by a cut off value Input is a tab delimited tag file with absolute tag counts, and output is a tab delimited tag file with absolute tag counts above cut off value. The original library size is printed in button of file Usage: -c cut off value (Default: 1) -i Input file containing list of input file names If -i is not given Inputfile1 Inputfile2 ... InputfileN are added as subsequent arguments.
DatabaseUpload.pl	Program to upload tag files to database for the LSDS project. No options need to be set to run the program.
ExonExonBoundaryCoverage.pl	Program to read the Exon/exon boundary coverage based on a GFF file and a coverage file. The coverage file is exported in CSV format from the CLC genomics Workbench and transformed into a tabular file using "ReduceCoverage.pl". Usage: -i input coverage file -g input GFF file with predicted CDS annotations -l length on each side of Boundary to calculate the coverage (default 10 bp) -c minimum coverage
ExtentCDSmodels.pl	Program to extract the first exon in the 5'end or the last exon in the 3'end of CDS models. Input files are a FASTA file with the genome sequence, a FASTA file with the CDS sequences and an annotation file with CDS models in GFF format. The program is made customly for the genome sequence of *Lotus Japonicus* (miyakogusa.jp - http://www.kazusa.or.jp/lotus/) Usage: -c CDS Sequence file -g Genome sequence file -a Annotation file in GFF format -t Number of nucleotides to add in 3' end -f Number of nucleotides to add in 5' end
GetBestUniRefBLASTResult.pl	Program to Extract the best hit for each query from a BLAST output file. The BLAST output must be in the format using the option outfmt 7.

[43] BLAST+ command line applications can be found at : ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/

	The "Input ID -> description file" is created using "CreateID2nameTable.pl". The hit with the lowest E-value is selected. Hits with the phrases : uncharacterized, Hypthetical, Hypothetical, or Predicted protein are avoided if possible. Usage: -i Input BLAST file -i Input ID -> description file -o Output file name (optional)
GetTaxIDs.pl	Program to limit a Uniref100 FASTA file to sequences from a specific taxonomic division. List with file names is downloaded from NCBI Taxonomy. (http://www.ncbi.nlm.nih.gov/taxonomy) in the format "Taxon names". For example to get all sequences belonging to the kingdom "Viridiplantae" search: "Viridiplantae[SubTree] AND species[Rank] NOT uncultured[prop] AND ("above species level"[prop] OR specified[prop])" Usage: -i Input file with taxon names to be extracted -f Input FASTA file to extract sequences from -o Output file name (optional)
GetUTRsAndExons.pl	Program to extract exon sequences based on the information provided in an annotation file in GFF3 format. A BLAST database with scaffold sequences is created using the "makeblastdb" command from the BLAST+ command line application. The BLAST database must be located in the same directory Usage: -i input GFF file with exon annotated regions -b BLAST database containing scaffold sequences
GlobalSagemap.pl	Program to extract SAGE tags from transcript sequences and match these to tags in a tag file. All tags, most 3-prime tags, and most 5-prime tags are extracted separately. All tags, most 3-prime tags, and most 5-prime tags are extracted separately. Input is a FASTA file with transcript sequences to match against (-s transcript sequences file) and a tag list file (-t tag file). OBS: Tags most be in first column and not contain the enzyme recognition motif (e.g. CATG). In the output file, it is listed whether a 3-prime (Grade A), a 5-prime match (Grade B), or an inside match (Grade C) was found. Usage: Creation if Virtual tag lists: -s Sequence contig file -l Tag size (default: 17 not including the CATG ID key) -e Enzyme recognition motif (default: CATG) Annotation of tag list: -t tag file (to be annotated) -n Number of libraries in tag file (Default: 2) -c Cut off value (default: 1 = Not found are printed as 0) -m 1 mismatch allowed [Y/N] (default: N) -r Tags have been extracted before [Y/N] (Default: N) -a Max number of annotations printed for each tag (Default: 3) -u Print Unknown Y/N] (Default: Y) -o Output file name (Optional)
MeanAndSTD.pl	Program to calculate mean expression values and standard deviations. Input is a data matrix with absolute tag counts that can contain multiple sample groups, each with multiple libraries. Output is a tab delimited data matrix with mean expression values and standard deviations (per 1 10 100... million). The mean expression and the standard deviation are calculated using the formula found in equations (3-1) and (3-2), respectively. Usage:

	-i input tag file -l list with sample -> sample group -r Relative tag count denominator (Tag count / X million) (default: 1 million) -n Number of libraries in the data matrix file (must be given)
NoGenesAboveCVcutoff.pl	Program to calculate the number of genes in a library, where the standard error of the mean expression value is below a cutoff value Usage: -i Input file containing list of input file names -c standard cutoff -s sample number (set to one for CV and not standard error) -o Output filename (optional) If -i is not given Inputfile1 Inputfile2 ... InputfileN are added as subsequent arguments.
NonRedundantRefSeq.pl	Program to make a non-redundant RefSeq FASTA file. If multiple sequences exit for a gene, the longest one is printed. Only mature messenger RNA (mRNA) transcripts (NM_) and model mRNA provided by a genome annotation processes (XM_) are included. Usage: -i input FASTA file -n Number of letters to print in each line in the output FASTA file
NormaliseTagTable.pl	Program to normalize absolute tag counts in a data matrix that can contain multiple libraries. Tag counts are normalized according to library size (e.g. to counts per million). Input is a tab delimited data matrix with absolute tag counts and output is a tab delimited data matrix with relative tag counts (pr. 1 10 100... million). Usage: -i input tag file -m Minimum tag count (default: 1) -r Relative tag count denominator (Tag count / X million) (default: 1 million) -d Number of decimals in relative count (default 0) -f Normalization factor: Actual or Original [A/O]: (Default: O) -n Number of libraries in file
PredictUTRs.pl	program to Predict 3' and 5' UTR regions based on a GFF file and a coverage file Coverage file is exported in CSV format from the CLC genomics Workbench and transformed into a tabular file using "ReduceCoverage.pl". The annotation of the CDS sequence is extended in the 5' and 3' end if a minimum coverage (-c option) is found in the coverage file. A small gap with no coverage can be allowed (-l option). Usage: -i input coverage file -g input GFF file with predicted CDS annotations -l maximum length of non-covered gap (default 0 bp) -c minimum coverage -s minimum coverage of 1. nucleotide in UTR region
ReadCoverage.pl	Program to calculate the mean and max coverage, fraction of CDS with read coverage, and coverage of CDS start and end positions. Each CDS can is validated it has coverage of mRNAseq reads over threshold limits (-t,-s and -m options). The coverage file is exported in CSV format from the CLC genomics Workbench and transformed into a tabular file using "ReduceCoverage.pl". The exon/exon boundary coverage file is created using "ExonExonBoundaryCoverage.pl". Usage: -i input coverage file -l input GFF list of genes -t Threshold value (Default: 1) -s 2. Threshold value (Default: 1)

	-m Minimum % of CDS covered to pass -g Add boundary coverage from CDS-CDS boundary file "Y/N" -b Exon/exon boundary coverage file
sampleNameConversion.pl	Program to convert sample IDs names into database IDs for the LSDS project Usage: -i Input file -s Sample ID format AKV (2008): 1 KMC (2008): 2 Vandel drought (2008): 3 Vandel Late Blight l 1 (2008): 4 Vandel Late Blight 2 (2008): 5 Vandel drought (2009): 6 -o Output file name (Optional)
ReduceCoverage.pl	Program to convert a coverage file exported in CSV format from the CLC genomics Workbench into a tabular file. All positions with below the threshold value are not printed. Usage: -i input CSV formatted file -c Threshold value
SolexaTagExtraction.pl	Program to Extract sequence tags from FASTQ sequence files and create tabular lists with counts. Multiple output files are created, one for each barcode. Cutoff values are based on Phred scores which is translated into the Illumina quality format output files are located in subfolders C1-C8 (from the 8 lanes on a flow cell). Output files are named based on the Barcode sequence. Barcode sequences must be placed in the same directory as the FASTQ files. Usage: -c Cutoff value (Default: Phred Score 20) -m Max bases below cut off value (Default: 1 bases) -b Barcode file (Default: solexa-keys.txt) -t Tag length (Default: 17) -p print progress on screen [Y/N] (Default: \"N\") -q Quality score scheme: 1 = Solexa 2 = Illumina 1.3+ 3 = Illumina 1.5+ 4 = Illumina 1.8+ (default)
SolexaTagExtractionPipeline.pl	Program to automate the solexa pipeline, and tag extraction Program is designed for minimum user input. Steps in script: 1) Validation of presence of necessary files and directories 2) Optional change of settings for base calling and tag extraction 3) Base calling using goat_pipeline.py script 4) Creation of s_[1-8]_sequence.txt files using gerald.pl script 5) Tag extraction using SolexaTagExtractionAuto.pl script 6) Optional cleanup of created analysis directory (Only the "Analysis directory, Not the "Data" directory") Additional optional settings: -c Configurations library (Default: "Configurations") -s files have been automatically copied [Y/N] Tag extraction settings (Default settings = no filtering) -v Cutoff value (Default: Phred Score 0) -m Max bases below cut off value (Default: 17 bases) -b Barcode file (Default: solexa-keys.txt) -t Tag length (Default: 17)
StatsFromGFF.pl	Program to create a statistics table from a GFF file. Input is a GFF file with annotations of genes on a genome and output is a tab delimited file with information on each mRNA transcript.

	Usage: -i input GFF file -o Output filename (optional). Default: "STATS-input file name"
SubsamplingOfReplicates.pl	Program to randomly extract a subset of replicates from a data matrix file. Usage: -i input tag file -l list with sample -> sample group -n Number of libraries to be subsampled from each replicate group
Tag2GeneCounts.pl	Program to sum tag counts from multiple tags matching the same gene. If a tag matches multiple transcripts, the count is divided to all transcripts Usage: -i input tag file -d Number if decimals to use (Default 3) -o Output filename (optional)
TagLists2FASTQ.pl	Program to create a Sequence file in FASTQ format from a Tag list. Input is a list with names of tag files. Usage: -k sample key to use (default AAA) -i List with tag tables -o Output file name (Optional). Default: FASTQ-"Tag list"
TaglistSubsampling.pl	Program to generate a random sub sample of a tag list. Input is tag list and output is a sub sample of the tag list, based on the tag count of each sample. Tags are drawn from the tag list without replacement (A tag with tag count 2, can be subsampled to times) Usage: -i Input file -n total tag count of subsample -o Output file name (Optional). Default: Subsample_"Sub sample size"-"Tag table"
EgdeR (Robinson, McCarthy & Smyth, 2010) Descriptions are from the EgdeR manual[44]	**Command:** estimateCommonDisp **Description:** Maximizes the negative binomial conditional common likelihood to give the estimate of the common dispersion across all tags for the unadjusted counts provided. **Command:** estimateTagwiseDisp Description: Estimates tagwise dispersion values by an empirical Bayes method based on weighted conditional maximum likelihood. Settings: • *prior.n:* numeric scalar, smoothing parameter that indicates the weight to give to the common likelihood compared to the individual tag's likelihood; default 'getPriorN(object)' gives a value for 'prior.n' that is equivalent to giving the common likelihood 20 prior degrees of freedom in the estimation of the tag/genewise dispersion. • trend: method for allowing the prior distribution for the dispersion to be abundance-dependent. Possible values are "none", "movingave" and "tricube". "none" means no trend. "movingave" applies a moving average smoother to the local likelihood values. "tricube" applies tricube weighting to locally smooth the common likelihood. • prop.used: optional scalar giving the proportion of all

[44] The EgdeR manual can be found at: http://www.bioconductor.org/packages/release/bioc/manuals/edgeR/man/edgeR.pdf

	tags/genes to be used for the locally weighted estimation of the tagwise dispersion, allowing the dispersion estimates to vary with abundance expression level). For each tag/gene the estimate of its dispersion is based on the closest 'prop.used' of all of the genes to that gene, where 'closeness' is based on similarity in expression level. (Robinson, McCarthy & Smyth, 2010)

C) Complete Publication List

The work presented in the current thesis, especially regarding development of custom tools for data analysis of tag based transcriptomics have resulted in several peer review publications. Although these studies are related to the theme of the current thesis, it was chosen not to describe them in detail in order to keep the scope of the thesis as narrow as possible.

- Gyetvai, G., **Sønderkær, M.**, Göbel, U., Basekow, R., Ballvora, A., Imhoff, M., Kersten, B., Nielsen, K.-. & Gebhardt, C. 2012, "The transcriptome of compatible and incompatible interactions of potato (Solanum tuberosum) with Phytophthora infestans revealed by DeepSAGE analysis", *PLoS ONE*, vol. 7, no. 2.
- Nordlund, J., Kiialainen, A., Karlberg, O., Berglund, E.C., Göransson-Kultima, H., **Sønderkær, M.**, Nielsen, K.L., Gustafsson, M.G., Behrendtz, M., Forestier, E., Perkkiö, M., Söderhäll, S., Lönnerholm, G. & Syvänen, A. 2012, "Digital gene expression profiling of primary acute lymphoblastic leukemia cells", *Leukemia*, vol. 26, no. 6, pp. 1218-1227.
- Mortensen, S.A., **Sønderkær, M.**, Lynggaard, C., Grasser, M., Nielsen, K.L. & Grasser, K.D. 2011, "Reduced expression of the DOG1 gene in Arabidopsis mutant seeds lacking the transcript elongation factor TFIIS", *FEBS letters*, vol. 585, no. 12, pp. 1929-1933.
- Dueholm, M.S., Petersen, S.V., **Sønderkær, M.**, Larsen, P., Christiansen, G., Hein, K.L., Enghild, J.J., Nielsen, J.L., Nielsen, K.L., Nielsen, P.H. & Otzen, D.E. 2010, "Functional amyloid in pseudomonas", *Molecular microbiology*, vol. 77, no. 4, pp. 1009-1020.
- Kaminski, K.P., Petersen, A.H., **Sønderkær, M.**; Pedersen, L.H., Pedersen, H., Feder, C., Nielsen, K.L. "Transcriptome analysis suggests that starch synthesis may proceed via multiple metabolic routes in high yielding potato cultivars", *PLoS ONE*, accepted with major revisions.
- Kloster, B.M., Bilgrau, A.E., Rodrigo-Domingo, M., Bergkvist, K.S., Schmitz, A., **Sønderkær, M.**, Bødker, J.S., Falgreen, S., Nyegaard, M., Johnsen H.E., Nielsen, K.L., Dybkaer, K., Bøgsted, M. "A model system for assessing and comparing the ability of exon microarray and tag sequencing to detect genes specific for malignant B-cells". *BMC Genomics*, submitted.

D) Planned Publications

The work presented in chapters 3 and 4 in the current thesis, is planned to be submitted to peer review journals. Below, tentative titles and small description of these are given.

Results from the Transcriptome Analysis of *Lotus japonicus* During Nodulation

Sønderkær, M., Petersen, A.H., Andersen, S.U., Nielsen, K.L. Stougaard, J. and Radutoiu, E.S

The two major findings of this study, namely the early specific cell wall metabolism and defense response only found in the wild type *L. japonicus* genotype and the dramatic up-regulation of an Asparagine synthase gene will be described. The data set and the analysis hereof might also become a part of a larger study also incorporating proteonomics data. However, a tentative title for a manuscript of the transcriptome study could be: "*Gene expression profiling of Lotus Japonicus during nodulation reveals early plant response to bacterial symbiosis and major induction of an Asparagine synthase in nitrogen fixating plant nodules*".

Results from the Analysis of Variance in Tag Based Transcriptome Data

Sønderkær, M. and Nielsen, K.L.

Firstly, the analysis of the high replicate biological groups represented by an unprecedented number of replicates will be presented. The major findings, namely a major drop in specificity but retainment of the sensitivity when lowering the replicate number will be presented. Moreover, the fact that a difference between two closely related biological groups (in this case two different potato cultivars) can be made of by subtle changes in the expression of many genes. This challenges the general assumption in data analysis of gene expression data sets, that most genes are not differentially expressed. Moreover, the comparison between mRNAseq and DeepSAGE will be presented, highlighting the low technical variation found in the mRNAseq data set. A tentative title of the manuscript could be: "*Biological and technical variation in gene expression using sequence tag based methods*".