Populus transcriptomics – from noise to biology

ANDREAS SJÖDIN

Akademisk avhandling

som med vederbörligt tillstånd av rektorsämbetet vid Umeå universitet för avläggande av Teknologidektorsexamen i Växter cell- och molekyläriobiologi, framläggs till offentligt försvar i Arbetslivsinstitutets hörsal, Umeå Universitet, fredagen den 30 november 2007 klockan 10.00
Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent:
Dr. Jan Lohmann
Max Planck Institute for Developmental Biology
Tübingen, Germany

Umeå Plant Science Centre
Department of Plant Physiology
Umeå university
Sweden 2007
Abstract
DNA microarray analysis today is not just generation of high-throughput data, much more attention is paid to the subsequent efficient handling of the generated information. In this thesis, a pipeline to generate, store and analyse Populus transcriptional data is presented. A public Populus microarray database - UPSC–BASE - was developed to gather and store transcriptomic data. In addition, several tools were provided to facilitate microarray analysis without requirements for expert-level knowledge. The aim has been to streamline the workflow from raw data through to biological interpretation.

Differentiating noise from valuable biological information is one of the challenges in DNA microarray analysis. Studying gene regulation in free-growing aspen trees represents a complex analysis scenario as the trees are exposed to, and interacting with, the environment to a much higher extent than under highly controlled conditions in the greenhouse. This work shows that, by using multivariate statistics and experimental planning, it is possible to follow and compare gene expression in leaves from multiple growing seasons, and draw valuable conclusions about gene expression from field-grown samples.

The biological information in UPSC-BASE is intended to be a valuable transcriptomic resource also for the wider plant community. The database provides information from almost a hundred different experiments, spanning different developmental stages, tissue types, abiotic and biotic stresses and mutants. The information can potentially be used for both cross-experiment analysis and for comparisons against other plants, such as Arabidopsis or rice. As a demonstration of this, microarray experiments performed on Populus leaves were merged and genes preferentially expressed in leaves were organised into regulons of co-regulated genes. Those regulons were used to define genes of importance in leaf development in Populus. Taken together, the work presented in this thesis provides tools and knowledge for large-scale transcriptional studies and the stored gene expression information has been proven to be a valuable information resource for in-depth studies about gene regulation.

Keywords: Populus, microarray, transcriptomics, UPSC-BASE, leaf development
Populus transcriptomics – from noise to biology

ANDREAS SJÖDIN

Umeå Plant Science Centre
Department of Plant Physiology
Umeå university
Sweden 2007
Abstract

DNA microarray analysis today is not just generation of high-throughput data, much more attention is paid to the subsequent efficient handling of the generated information. In this thesis, a pipeline to generate, store and analyse *Populus* transcriptional data is presented. A public *Populus* microarray database - UPSC–BASE - was developed to gather and store transcriptomic data. In addition, several tools were provided to facilitate microarray analysis without requirements for expert-level knowledge. The aim has been to streamline the workflow from raw data through to biological interpretation.

Differentiating noise from valuable biological information is one of the challenges in DNA microarray analysis. Studying gene regulation in free-growing aspen trees represents a complex analysis scenario as the trees are exposed to, and interacting with, the environment to a much higher extent than under highly controlled conditions in the greenhouse. This work shows that, by using multivariate statistics and experimental planning, it is possible to follow and compare gene expression in leaves from multiple growing seasons, and draw valuable conclusions about gene expression from field-grown samples.

The biological information in UPSC-BASE is intended to be a valuable transcriptomic resource also for the wider plant community. The database provides information from almost a hundred different experiments, spanning different developmental stages, tissue types, abiotic and biotic stresses and mutants. The information can potentially be used for both cross-experiment analysis and for comparisons against other plants, such as *Arabidopsis* or rice. As a demonstration of this, microarray experiments performed on *Populus* leaves were merged and genes preferentially expressed in leaves were organised into regulons of co-regulated genes. Those regulons were used to define genes of importance in leaf development in *Populus*. Taken together, the work presented in this thesis provides tools and knowledge for large-scale transcriptional studies and the stored gene expression information has been proven to be a valuable information resource for in-depth studies about gene regulation.
Sammanfattning


En av de stora utmaningarna i analys av mikromatriser är att kunna särskilja bruset från värdefull biologisk information. Att studera träd som växer utomhus är komplext eftersom de interagerar med omgivningen i mycket större utsträckning än vad som är fallet i växthusets kontrollerade miljö. Det här arbetet visar att det är möjligt, med hjälp av avancerad statistik och god försoksplanering, att följa och jämföra genuttrycket i blad från aspar utomhus under flera år för att dra värdefulla slutsatser om geners reglering.

List of papers

The thesis is based on the following papers, which will be referred to in the text by their Roman numerals.


V. Sjödin A, Wissel K, Bylesjö M, Trygg J and Jansson S: Global expression profiling in leaves of free-growing aspen (manuscript)

VI. Sjödin A, Street NR, Bylesjö M, Trygg J, Gustafsson P and Jansson S: A cross-species transcriptomics approach to identify genes involved in leaf development (manuscript)
Also published by the author but not printed in the thesis


Abbreviations

ANOVA Analysis of variance
BASE BioArray Software Environment
DNA Deoxyribonucleic acid
EST Expressed Sequence Tag
Fv/Fm Optimum quantum yield
GEO Gene Expression Omnibus
GMOD Generic Model Organism Database
GO Gene Ontology
GSEA Gene Set Enrichment Analysis
JGI Joint Genome Institute
KEGG Kyoto encyclopaedia of Genes and Genomes
LIMS Laboratory information management system
LOWESS Locally weighted scatterplot smoothing
MASQOT MicroArray Spot Quality cOnTrol
MeSH Medical Subject Heading
MIAME Minimum Information About a Microarray Experiment
MIPS Munich Information Center for Protein Sequences
miRNA microRNA
MPSS Massive Parallel Signature Sequencing
mRNA messenger RNA
NCBI National Center for Biotechnology Information
OPLS Orthogonal projections to latent structures
PCR Polymerase Chain Reaction
PI Principal Investigator
PICME Platform for Integrated Clone Management
POP1 UPSC Populus microarray generation 1
POP2 UPSC Populus microarray generation 2
PLS-DA Partial Least Squares Discriminant Analysis
QTL Quantitative trait locus
RLS Restricted Linear Scaling
RNA Ribonucleic Acid
RT-PCR Reverse Transcription Polymerase Chain Reaction
SAGE Serial Analysis of Gene Expression
UBC University of British Colombia
UPSC Umeå Plant Science Centre
VSN Variance Stabilisation Normalisation
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of papers</td>
<td>vii</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>ix</td>
</tr>
<tr>
<td>Contents</td>
<td>xi</td>
</tr>
</tbody>
</table>

### 1 Background

1.1 Gene regulation                                                   | 3    |
1.2 Microarray technology                                             | 4    |
1.2.1 Alternative transcriptional technologies                       | 7    |
1.2.2 Data analysis                                                  | 7    |
1.2.3 Storing and sharing the information                            | 9    |
1.3 *Populus* as model organism                                       | 9    |
1.3.1 Why study mature free-growing *Populus* trees                   | 11   |

### 2 Results achieved

2.1 Building the infrastructure for transcriptomics                   | 13   |
2.1.1 Comparison of different *Populus* microarray platforms          | 14   |
2.1.2 UPSC-BASE                                                      | 15   |
2.1.3 Improving the hybridisation quality of microarray slides        | 17   |
2.1.4 Image analysis                                                 | 17   |
2.1.5 Normalisation                                                  | 19   |
2.1.6 *In silico* reference designs                                   | 21   |
2.2 Turning UPSC-BASE into a valuable resource                        | 22   |
2.2.1 Integrating transcriptomics with physiology                    | 22   |
2.2.2 Trees in future climate                                        | 27   |
2.2.3 Applying stress treatments                                     | 28   |
2.2.4 How to use stored information                                  | 29   |
2.3 Reconstruction of gene regulatory networks in *Populus*          | 30   |

### 3 Conclusions

3.1 The next generation Populus genome browser                        | 34   |
3.2 Future perspectives                                              | 35   |

### 4 Acknowledgement

Bibliography                                                           | 39   |
A dramatic shift has been seen in biology during the last decade. Modern biology has entered a genomic era with new genomes being sequenced at an ever-increasing speed. High-throughput technologies allow the simultaneous measurement of thousands of characteristics in biological systems such as proteins (Perez and Nolan, 2002), metabolites (Lu et al., 2006) and RNA (Schena et al., 1995; Lu et al., 2005). The amount of public data available has open possibilities to study thousands of genes at the same time. This drift from a univariate view of biology to systems biology demands totally different competences from researchers. It is no longer enough to be skilled at lab bench work only; basic knowledge in bioinformatics and data handling are needed. Modern biology can’t continue to evolve without the support of solid bioinformatic infrastructures.

This thesis aims to increase knowledge of how to handle and analyse microarrays in a concise, streamlined procedure. There are no golden rules for microarrays analysis and no method fits all circumstances. The research presented in the following chapters should be seen as recommendations rather than golden rules. It covers a journey of setting up an infrastructure for *Populus* transcriptomics. It is very important for researchers to trust public repositories and to achieve data with high quality. Focus in this work has therefore been to standardise and automate the work flow from hybridisation through to candidate gene list generation. The final part of the thesis shows how to use generated data to draw biological conclusions and to define regulons to understand gene regulation machinery in *Populus* leaves.
Background

What is happening inside a *Populus* leaf, from bud burst in the spring until it finally changes colour from green to yellow or red in the autumn? *Populus* trees and other multi-cellular organisms have evolved complex gene regulatory mechanisms in order to coordinate signals between cells. Plants cannot move as rapidly as animals. The fastest way is to spread seeds but this is still far from rapid, and plants are therefore more dependent on regulation to cope with changes in environmental conditions. Until recently it was only possible to study a handful of genes or proteins at the same time. This can be compared to sampling random balls from an urn and drawing conclusion about a small, spotlighted area without knowing the relation to the rest of the system. The drift that has occurred over the last decade that has seen a shift from such gene-wise approaches to large-scale studies is an effect of technology and computer development and improvement. Measuring gene regulation is not the problem any longer and new unknown connection between previously separated pathways has been discovered. The main problem remains to understand the biology behind the building blocks and the connections within the signalling system. The expression of many genes is regulated after transcription (i.e. by microRNAs) and increased mRNA concentration doesn’t directly imply increased protein levels. Nevertheless, mRNA levels can be quantitatively measured by transcriptional methods, such as Northern blots or microarrays and even if mRNA levels are not always informative beyond the level of the transcriptome, in many cases they will be and mRNA profiling indicates which proteins have the potential to be translated within cells.

1.1 Gene regulation

The central dogma, which describes the route from DNA to RNA to protein, forms the basis of molecular biology. See Figure 1.1 for a simplified overview of involved steps. The genetic information of a cell is stored in the DNA consisting of thousands of genes. Each gene serves as a blueprint of how to build a protein. The nucleus contains the DNA, which is organised into chromosomes. This information is identical for all cells of an organism and the DNA must be precisely and accurately replicated before a cell divides. For protein-coding genes, the information is
transcribed into RNA. The non-coding parts are removed and transported out to the cytosol where the RNA finally is translated into proteins.

Gene regulation is a research field progressing fast to understand how the regulation systems are connected. The discovery of a group of regulatory elements, transcription factors (TF), can explain some parts but a lot of questions still remains unanswered. It has been shown that expression of mRNAs not always reflects the protein levels (Anderson and Seilhamer, 1997). The recent discovery of microRNA (miRNA) has given some evidence that it might be a major gene regulator (Bartel, 2004; Zhang et al., 2006, 2007; Kim and Nam, 2006). It might, in addition to turn over rates, be one of the missing links explaining the inconsistency between mRNA and protein expression levels. It has been shown in other organisms that they contain about 1-5% miRNA genes (Lai et al., 2003; Lim et al., 2003) compared to the total protein-coding genes. miRNA might also be responsible to control 30% of the protein-coding genes (Lewis et al., 2005).

1.2 Microarray technology

Quantification of large numbers of messenger RNA (mRNA) transcripts using microarray technology provides detailed insight into cellular processes involved in the regulation of gene expression. Large-scale generation of gene regulation information was not possible until recently. Array technology (larger dot blots) was already in use in mid 1980s (Augenlicht et al., 1984, 1987) but didn’t become important until the mid 1990s, when cDNA microarrays emerged (DeRisi et al., 1996; Lockhart et al., 1996). Pioneering work on plant microarrays was presented a decade ago.
using *Arabidopsis thaliana* (Schena *et al.*, 1995) followed some years later by work in *Populus spp.* (hereafter *Populus*) (Hertzberg *et al.*, 2001). For information and discussion about the laboratory part of microarray technology see Figure 2 in Paper IV (Sjodin *et al.*, 2006). Although the laboratory part stages of microarray technology are important, the major work-effort lies in the subsequent data processing and analysis.

Microarray experiments are extremely powerful and provide information on a genome-wide scale. However, they are also very complex and time-consuming compared to other biological techniques. The outcome is large and complicated data sets requiring large endeavours to analyse and validate. To cope with those problems the microarray community agreed to comply with standard rules to simplify data sharing and comparison of experiments. Minimum Information About a Microarray Experiment (MIAME) (Brazma *et al.*, 2001) stated what information and data should always be accessible for published microarray experiments. A schematic overview of the different steps is shown in Figure 1.2 with indication of the main coverage for each paper. For a more extensive description see Figure 2 and 3 in Paper IV.

The following data needs to be stored and provided for all microarray publications:

1. Raw data for each hybridisation
2. Normalised data
3. Sample annotation
4. Experimental design
5. Annotation of the array elements
6. Laboratory and data processing protocols

The raw data for each hybridisation includes both raw images from the scans and extracted raw intensity data. The raw intensity data is filtered and normalised to remove systematic noise. Normalised data is what is interesting to distribute and store as it is what is used for later analysis. Information describing sample annotation and design of experiments are very important to allow others to understand and interpret the biological results. The remaining points are more infrastructure-orientated to describe how the experiment was performed. However, without proper annotation information it is impossible to draw valid biological conclusions.

The first version of the MIAME standard had mainly the medical field in mind and a special updated version for plants was later published to better cope with
Figure 1.2. Schematic overview of methods used in analysis of microarrays. It is indicated what part is covered in the different papers.
1.2. MICROARRAY TECHNOLOGY

the special situation for plants (Zimmermann et al., 2006). When are microarrays an appropriate method to use? Microarray experiments shouldn’t be performed without spending time understanding the technique and to design appropriate experiments. Beginners to the use of microarray technology are usually recommended to consult an experienced expert before beginning to prepare samples. There are many examples of new users wasting both time and money trying to do microarrays with inappropriate biological samples. Microarrays are nice screening tools for hypothesis generation for later gene-wise experimental approaches.

1.2.1 Alternative transcriptional technologies

Several alternatives to cDNA microarray have emerged for analysing gene expression. The most closely related is spotted oligonucleotides, which have the advantage of allowing more gene-specific measurements in comparison to EST based platforms. Affymetrix has patented a special method allowing probes to be synthesised in situ. Those slides are supposed to be more reproducible between different laboratories, although they are much more expensive.

Different sequencing approaches exist as alternatives to closed microarray platforms. Quantitative real-time reverse-transcription PCR (qRT-PCR) (Heid et al., 1996) is usually used for smaller gene expression comparisons. For larger scale studies, Expressed Sequence Tag (EST) library construction has been widely used for gene expression profiling (Bhalerao et al., 2003; Sterky et al., 2004; Kohler et al., 2003; Brosche et al., 2005; Ranjan et al., 2004; Ralph et al., 2006). The high cost of sequencing has led to other alternatives, such as Serial Analysis of Gene Expression (SAGE) (Velculescu et al., 1995) and Massive Parallel Signature Sequencing (MPSS)(Brenner et al., 2000; Reinartz et al., 2002), which both aim to maximise the amount of information that can be extracted per sequence run. Recently, techniques based on de novo sequencing (454 and Alexa™) have allowed increased high-throughput profiling at a feasible cost. From here on the term microarray will refer only to spotted cDNA microarrays unless stated otherwise.

1.2.2 Data analysis

Analysis of microarrays has been covered extensively in the literature (for recent review see Allison et al. (2006)). The process can be summarized in five key components: design, pre-processing, inference, classification and validation. The experimental design is the basis and affects efficiency downstream in experiments. Biological replication and sample size need to be considered as well as if the samples should be pooled or not. The pre-processing stages including image analysis,
normalisation and transformation is an important part of this thesis. These steps have a huge impact on the final conclusion from microarray experiments.

Classification of samples can be divided in supervised where *a priori* information about the samples are used and unsupervised where no known information is provided. Unsupervised clustering using hierarchical clustering (Eisen *et al.*, 1998) has been widely used in microarray experiments. This approach provides a good overview of the data structure, but supervised methods are, in general, more powerful for classification. Using known information about systems improves result generation and interpretation. Most analysis method initially generate a, often long, list of genes and such lists are hard to summarise in an objective way. Applying statistical tests with tools looking for over-represented gene categories in classification systems (Hosack *et al.*, 2003; Banerjee and Zhang, 2002), such as Gene Ontology (GO) (Ashburner *et al.*, 2000), Munich Information Center for Protein Sequences (MIPS) (Mewes *et al.*, 1999; Schoof *et al.*, 2002) and Kyoto encyclopaedia of genes and genomes (KEGG) (Kanehisa and Goto, 2000) will outperform most individuals in interpreting gene functions. Checking only gene presence in results doesn’t take expression values into consideration. Lately Gene Set Enrichment Analysis (GSEA) has emerged as a complement using both classification system and expression information to find consistent, but small changes in categories.

Validation has been regarded as an important part of microarray results. However, it still remains unclear how validation should be achieved. To what level should the results obtained in validation agree with the initial results? Both sampling variability and measurement errors negatively influence validation. Most validation studies apply Northern blots or real-time PCR to confirm a handful of positive results with validation is rarely performed on genes not showing interesting patterns (negative controls) as well as those showing interesting patterns (positive control).

Commercial software for microarray analysis, such as Genespring, Genepix, has played an important role in many microarray labs to provide simple and extensive tools. The drawback of commercial software, in addition to expensive license fees, is the time lag it takes for companies to include new analysis methods presented in the literature and the black box nature of implemented analysis methods. The bioinformatic fields have a tradition of home-made software. In most cases the free software can provide the same or more extensive set of analysis tools in user-friendly graphical interfaces (Dudoit *et al.*, 2003). The TIGR multi expression viewer (Saeed *et al.*, 2003) provides many clustering and classification algorithms in an user-friendly software. The Bioconductor suite (Gentleman *et al.*, 2004), mainly written in the statistical language R (Ihaka and Gentleman, 1996), is the option for more advanced users and it offers an extensive collection of methods and algorithms. The suite was originally presented in 2002 consisting of 20 packages for microarray analysis. The project has now grown to >100 packages for technologies, such as microarray, proteomics, mass spectrometry, SAGE, cell-based assays and
1.3. **POPULUS AS MODEL ORGANISM**

1.2.3 **Storing and sharing the information**

Computer-aided tools are required for handling and analysing the tremendous amounts of data produced by microarray experiments. Mainly two types of database are needed, an annotation database containing information about biological material on the slide, and a storage and analysis database of expression data. For the later it can be divided into databases storing raw data, databases containing data analysis tools and finally databases containing curated, analysed data. Public repositories such as Arrayexpress (Brazma et al., 2003, 2006), Gene expression omnibus (GEO) (Barrett et al., 2007; Edgar et al., 2002), CIBEX (Ikeo et al., 2003) are designed to store MIAME-compliant microarray raw and processed data. Submission of microarray data to one of the two databases is now demanded by many scientific journals. Both databases are mainly created to provide information about experiments and how to reproduce the results. However no tools are provided for performing the actual analysis within the database system. BioArray Software Environment (BASE) (Saal et al., 2002; Troein et al., 2006) and similar systems provide both storage and analysis possibilities. Curated databases with processed data have gained increased attention lately mainly due to their user-friendly approach. Connecting to public databases, such as Genevestigator (Zimmermann et al., 2004; Grennan, 2006; Zimmermann et al., 2005), AthaMap (Galuschka et al., 2007; Bulow et al., 2006; Steffens et al., 2005, 2004), PathoPlant (Bulow et al., 2007), The Botany Array Resource (Winter et al., 2007; Toufighi et al., 2005) allows extraction of information for genes of interest. In this way, information about gene expression can be accessed without needing to perform additional laboratory work.

1.3 **Populus as model organism**

Around 35–40 plants have been partly sequenced until now. The two first completed were thale cress (*Arabidopsis thaliana*) (Arabidopsis Genome Initiative, 2000) and rice (*Oryza sativa*, japonica cultivar-group) (International Rice Genome Sequencing Project, 2005). Another variety of rice, (Yu et al., 2005, 2002), grape (*Vitis vinifera*) (The French Italian Public Consortium for Grapevine Genome Characterization, 2007) and *Populus trichocarpa* (Tuskan et al., 2006) have also been sequenced and assembled but are currently awaiting final genome presentation. The draft genome sequence of *Populus trichocarpa* was presented in 2006. *Populus* plays an important role as a forest tree, producing timber, pulp, paper, and energy (Tuskan and
Walsh, 2001), as well as being a keystone species in many ecosystems. The genome sequence, together with complementary genomic resources (Segerman et al., 2007; Andersson et al., 2004; Sterky et al., 2004), high growth rate, clonal propagation and the possibility to transform (Bradshaw et al., 2000; Taylor, 2002; Wullschleger et al., 2002; Brunner et al., 2004; Jansson and Douglas, 2007) have resulted in poplar emerging as the model system for tree research.

![Graph showing the number of Populus publications in Pubmed from 1976 to 2005.](image)

**Figure 1.3.** Number of *Populus* publications in Pubmed 1976-2005.

The number of publications in Pubmed classified as *Populus* in Medical Subject Heading (MeSH) field was extracted using Hubmed (Eaton, 2006). The figure 1.3 shows the increasing number of *Populus* publications in Pubmed. The records have grown from less than ten to hundreds papers per year. The UPSC *Populus* genome programme was started around the first surge of publications and a second major increase in publication came around the presentation of the *Populus* genome draft. In summary, *Populus* has emerged as a plant model system where *Arabidopsis* is not well suited.
1.3.1 Why study mature free-growing Populus trees

Studies to depict how environmental and developmental factors influence gene expression can addressed in several ways. The following list summarises three main groups:

1. Monitor effects of mutations.
2. Expose plants to highly controlled conditions and monitor effects.
3. Monitor effects of exposing plants to natural, uncontrolled conditions.

The first two strategies are used frequently while the third has been much less frequently employed. The main reason for this are the considerable challenges faced in performing this kind of approach. Weather and biotic interactions may have a major effect and bias gene regulation of interest. The result is highly dependent on selection of the measured parameters, but a proper experimental design can separate developmental factors from environmental factors influencing gene expression. Multivariate statistics has been applied on gene expression to study plants grown under uncontrolled, highly variable, conditions with good outcome (Wissel et al., 2003).

Controlled conditions in the greenhouse represent a highly over-simplified representation of the complex conditions that exist in the field. Plants in the field are exposed to different biotic (fungi, herbivore etc.) and abiotic (rain, wind, nutrient deficiency etc.) conditions. Kulheim et al. (2002) shows nicely the power of field experiments to reveal phenotypic differences that are undetectable in controlled conditions. The natural variation in environmental conditions offers unique experimental possibilities that are impossible to mimic in greenhouses. In the field, trees are in their natural environment and not in artificial binary light-dark and temperature condition.

Another parameter to consider is the difference between juvenile and mature plants. For instance, in medicine, research is performed on individuals matching required criteria. It is usually not a good idea to study cancer treatments in children for a cancer where most cases occur in more elderly persons. In tree biology most research is performed on very young plants. As a result, studying mature plants in their natural conditions will provide an important complement to the more usual research carried out in greenhouses on young, small plants.
Chapter 2

Results achieved

This thesis covers both generation and biological interpretation of microarray data. Robust laboratory protocols are essential for the generation of good-quality data for the final interpretation of results. The initial laboratory stages involved in the use of microarrays are not the focus here, with the main achievements of this thesis regarding the analysis stages, from information extraction in properly designed microarray experiments through to drawing final biological conclusions. This chapter is divided in three sections; method development, biological studies and large-scale data mining. The first section describes the different Populus microarray platforms available, data extraction from microarray slides (image analysis), removal of systematic noise (normalisation) and storage of data. The biological study starts with the integration of genomic tools and physiology on free-growing Populus tremula (aspen) trees, which is the most novel part of the biological studies in this thesis. The other microarray experiments presented are more generally described to give a short introduction of Populus leaf data generated in this project. The data is used in the third section to extract information from multiple experiments that would not have been visible in individual experiments. For the third section, UPSC-BASE was screened for leaf microarray experiments and groups of co-regulated genes ( regulons) in Populus leaves were defined to find candidate genes important for leaf development. Second, public data from several different microarray platforms were merged for large-scale stress comparisons.

2.1 Building the infrastructure for transcriptomics

Huge amounts of data are generated when laboratory and analysis of transcriptomic techniques have been set up and are running. For a detailed overview of the laboratory and analysis pipeline, see Paper IV (Sjodin et al., 2006). Each single microarray slide is digitalised and stored in large raw TIFF-image files. Those files need to be quantified to extract raw data results. In total, the final product will, depending on the number of scans for each slide and the resolution of the captured image, need large amounts of storage space. It is hard to keep track of all files without a proper infrastructure and storing the information as files on individual personal computers should be avoided. For larger groups of scientists working on
the same microarray platform, a well designed database system is the only possible solution. In addition to storing data, the database will also encourage information sharing and provide a safe backup system.

### 2.1.1 Comparison of different *Populus* microarray platforms

The different microarray platforms available for use in *Populus* were compared before setting up the large-scale, cross-platform comparison (meta-analysis). All public EST sequences in GenBank (Benson *et al.*, 2007) were downloaded and nucleotide BLAST searches were performed against the poplar genome. Out of 45555 predicted genes (gene models) in *Populus* (Tuskan *et al.*, 2006) the calculation gave 24687 genes that were supported by at least 1 EST. Not surprisingly, the most represented gene is Ribulose bisphosphate carboxylase, small chain (eugene3.01230087), which was represented by 2287 ESTs.

<table>
<thead>
<tr>
<th>Number of spots</th>
<th>Gene models</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPSC POP1 13872</td>
<td>8916</td>
</tr>
<tr>
<td>UPSC POP2 25648</td>
<td>13882</td>
</tr>
<tr>
<td>Helsinki 9K 4</td>
<td>5536</td>
</tr>
<tr>
<td>Helsinki 10K 4</td>
<td>5536</td>
</tr>
<tr>
<td>PICME3 4</td>
<td>8004</td>
</tr>
<tr>
<td>PICME4 4</td>
<td>8547</td>
</tr>
<tr>
<td>UBC 4</td>
<td>9409</td>
</tr>
</tbody>
</table>

**Table 2.1.** Comparison of different *Populus* microarray platforms. The columns describes number of clones spotted and gene models represented on different *Populus* cDNA microarrays.

<table>
<thead>
<tr>
<th></th>
<th>Helsinki 10K</th>
<th>PICME4</th>
<th>UBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPSC POP2</td>
<td>3834</td>
<td>5830</td>
<td>6471</td>
</tr>
<tr>
<td>Helsinki 10K</td>
<td>5160</td>
<td>3071</td>
<td></td>
</tr>
<tr>
<td>PICME4</td>
<td>4705</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2.2.** Number of gene models common to different *Populus* cDNA microarrays

Table 2.1 shows coverage of gene models by the different *Populus* microarray platforms and Table 2.2 show gene models in common. The UPSC POP2 microarray contains all clones from the POP1 microarray. The two Helsinki microarrays contains exactly the same of ESTs as the 10K version of the microarray, with the 10K version only containing more control clones. The majority of the EST clones from the Helsinki microarray are also present on the PICME microarrays. In a broader perspective, 1817 gene models were represented on the latest versions of all of the platforms. POP2 is the microarray platform with largest gene model support
2.1. BUILDING THE INFRASTRUCTURE FOR TRANSCRIPTOMICS

(30%). In conclusion, cDNA microarrays only cover a small subset of the predicted gene models in *Populus*. As a complement to the cheaper cDNA microarrays, NimbleGen and Affymetrix has developed a full genome array based on *Populus* genome release version 1.0. (Ramirez-Carvajal *et al.*, 2007)

2.1.2 UPSC-BASE

BioArray Software Environment (BASE) (Saal *et al.*, 2002; Troein *et al.*, 2006) was selected as the UPSC microarray database because it has large community support in the academic field and is being developed at a national university. The system is build upon a comprehensive open-source MIAME-compliant web-based database solution. The database uses multiple-layer architectures with a database storing the data in the background. Scripts are responsible of sending information between the server and the viewable layer in the web browser.

**Modifications and improvements of the database**

UPSC-BASE is built upon the BASE version 1.2.10 release. BASE was the best available option for large-scale storage of microarray data but several modifications focusing on improving the interface to make it more user-friendly and fit the special needs for the microarray community at UPSC were implemented. A short summary of the most important changes is presented here.

The access level has been modified to better fit a large research centre with several visitors. The database categorises each user into one of four different levels.

1. Administrative users
2. Principal Investigators (PIs)
3. Scientists (Postdoc and PhD students)
4. Visiting scientists

Storing information in databases can seem frightening where limiting access to data is important and a common misunderstanding is that data is accessible for others persons. Restrictions are used in databases to limit access to functions and data for certain types of users. Specific user levels allow the user interface to be as simple as possible.
The second major modification apparent for the user is that experiments are requested and not created as in the original BASE version. This allows the administrator together with a steering group to have total control over experiments added to the database. Only PI users have the right to request experiments and are responsible for adding the data in the database once the data has been generated. Information about produced microarray slides is imported directly by the administrator and individual slides are identified by unique bar codes. This modification provides a tight connection between the production, ordering and storing of microarray slides. The result is that end-users don’t need to define slide layouts and slide generation as long correct bar codes are provided to the database.

The third user-friendly modification is the batch-wise import function. In the original installation, all data, such as TIFF images and original raw data files, had to be imported one by one. This was a very time-consuming procedure for large experiments and import to UPSC-BASE can instead be achieved by first uploading all data files to the ftp-server and then linking the data using a spreadsheet file. The batch-import is flexible and can be extended to any number of slides and handle multiple scans for each slide. If the data files are named consistently, R-code for generating ‘skeleton’ spreadsheet files can be used, which minimises the amount of work needed to input all MIAME-required fields of the experiment.

In addition to those major improvements several minor modifications have been made to streamline the daily work. The Laboratory Information Management System (LIMS) is extended to allow more advanced external analysis tools. Frequent users can set up analysis tools to run in sequence which saves a lot of time for routine analysis. The alternative would be waiting for each analysis step to finish before starting the next. The automatic analysis procedure is often sufficient for non-experienced microarray users to study expression patterns of favourite genes. All experiments in UPSC-BASE are automatically transformed using Restricted Linear Scaling (RLS) (Ryden et al., 2006), and subsequently normalised using the step-wise normalisation method (Wilson et al., 2003). This data is denoted ‘auto-finalized’ in the database and can be compared between experiments. This also simplifies extended searches performed between experiments.

Finally, extended flexibility for data sharing has been implemented to promote more open communities to improve the results. The permission structure was modified to allow permission levels handling multiple groups throughout the database system. In the installation, the user can be added as a second (additional) group to the experiment with the correct access privileges, thus encourage collaboration between different research groups.

The many modifications have unfortunately made it impossible to upgrade the database to newer none-modified versions of BASE. This is a very common problem for scientist. Provided software are often not adequate in the standard version and
modifications are performed to fit special needs. Unfortunately, these modifications make it impossible to use subsequent updates from the community (Swertz and Jansen, 2007). A lot of these problems could be avoided if more software used module-based extension possibilities.

2.1.3 Improving the hybridisation quality of microarray slides

Sample collection, extraction of RNA and cDNA synthesis all contribute sources of error. Hybridisation is the first step in the microarray pipeline introducing bias purely from non-biological sources. Hybridisation was originally carried out on microscope slides with spotted cDNA using cover slips. The hybridisation solution was not mobile and spatial patterns in the hybridisation pattern were normal. Using a framed cover slip (which raises the cover slip from the array surface slightly) significantly increased the quality of microarray slide hybridisations. In addition to quality problems in hybridisations, the time consumption for manual hybridisation and washing was a bottle neck. Only a handful of microarrays could be performed per day, even for a trained person. Automatic systems, such as the Automated Slide Processor (Amersham, Sweden) advanced the microarray technology from labour-intensive to high-throughput. Reproducible results can be achieved with minor human involvement resulting in better quality data being generated, which allows for easier biological interpretations.

2.1.4 Image analysis

The focus in the microarray field has been more on statistical analysis rather than improving the quality of microarray slides. Image analysis is a very crucial step in the work flow (Yang et al., 2001, 2002a) and problems will have a large impact on later biological interpretations. For instance the glass slide is frequently contaminated during the preparation steps, see Figure 2.1. This generates spots that do not reflect the true expression level of the gene.

The obvious question is ‘What makes a good spot good?’: Typically the quality assessment of spots is performed manually by visual inspection. Considering that a normal sized cDNA microarray contains 25000 spots and a decent experiment has 20 slides results in more than half a million spots to inspect. The main problem is to decide if a spot looks strange due to technical issues or rather due to biological effects. It is often also almost impossible to differentiate between the two possible sources of noise and all incorrect measurements will affect the final analysis result in the normalisation and hypothesis tests. Spots affected by technical issues will thus affect other spots during analysis and the biological interpretation might be
misleading or even incorrect. Existing image analysis software for microarrays, such as GenePix, Imagen, QuantArray and ScanAlyze, are all highly dependent on user interaction. Alternative, more advanced, semi-automatic software such as UCSF Spot (Jain et al., 2002) and WaveRead (Bidaut et al., 2006) can handle batch wise image analysis but do not include an advanced routine for spot quality classification. Bluefuse (BlueGene) offers a commercial software that offers user-friendly, high-throughput image analysis although with the limitation it isn’t free.

**MASQOT**

MicroArray Spot Quality cOnTrol (MASQOT), described more extensively in Paper I (Bylesjo et al., 2005) was developed to streamline image analysis. Briefly, the image analysis is divided into spot characterisation, spot quality assessment, and spot classification. The main aim was to mimic human, visual inspection and provide significant time-savings for the biologist. A training dataset was based on 5 slides from the *Populus* second generation microarray (POP2). Roughly 80000 spots (split in half into training and test set) were manually inspected by 3 microarray users and the quality of each spot was assigned as bad (flagged) or not bad (not flagged). Multivariate methods were applied to find spot characterisations important for classification of bad spots and finally class discrimination of the classes were found using PLS-DA. For the training set containing 38627 spots 98.6% of the spots are correctly predicted and in the independent test set containing 39421 spots 98.1% of the spots are correctly predicted. This proves that the methods robustness in finding and differentiating good spots from bad spots.
A robust methodology for quality control is nice but is useless if nobody knows how to use the method. Just as most people playing a computer game do not care how the game is working to produce the graphics, microarray users don’t want to study complex algorithms in order to classify spots as good or bad. The solution for MASQOT was to implement the algorithm in an application called MASQOT-GUI, presented in Paper II (Bylesjo et al., 2006). The interface is designed to make it easy for the beginner to get started using the software. The focus has been on providing a tool with batch-mode possibilities using as few options as possible in default settings. The experienced user still has other more detailed options to handle specific problematic cases.

**Restricted linear scaling**

Microarray slides are digitalised and converted to 16-bit TIFF images in high-resolution scanners. This limits the intensity range to $2^{16} = 65536$ discrete steps, which is a problem for comparing lowly and highly expressed genes. Physical microarray slides are routinely scanned multiple times to extend the linear range before importing into UPSC-BASE and the multiple scans are then merged using a regression method, Restricted Linear Scaling (RLS) (Ryden et al., 2006). The implemented RLS is a slightly modified version compared to the algorithm presented by Dudley et al. (2002) handling the problems associated with missing and saturated signals, which occur in most types of microarray experiments. The implemented method has a limitation of not using all data values from the slides. Only data points within the linear range are used. The constrained model (CM) (Bengtsson et al., 2004) uses all information from separate scans and might be a candidate for implementation in BASE in the future as a complement to RLS.

**2.1.5 Normalisation**

Microarrays are an expensive technology and replication was not always routinely used. When microarrays were first introduced, several single-array methods for inferring differential expression were developed and they were sometimes used without applying normalisation. The simplest method only considered ratios below or above a pre-determined cut-off, often referred as the fold-change method. An obvious shortcoming of this approach is the use of an arbitrary cut-off without considering the variability in the data. Ratio-based expression values from microarrays don’t have constant variation within the intensity range or between different slides. Lowly expressed spots tend to have higher variation due to random error. Three single-slide experiment methods, usually referred as Chen’s (Chen et al., 1997), Churchill’s and Newton’s, aimed to handle those problems. Quite rapidly those
single slide methods become obsolete as the experiment sizes grew. The two later, not even published. At the same time, need of proper normalisation methods started to receive more attention in the literature and the underlying technical problems of microarray technology became apparent. A first approach was to apply median normalisation to centre the expression ratios on the overall median expression. However this method had the severe limitation of assuming linearity between logged expression ratios and total intensity levels. Different levels of dye incorporation leads to unbalanced channel intensities and a non-linear relationship appears for low intensity spots. The reason is mainly due to the fact that they are very close to the detection limit. The most commonly used approach to handle those problem is to use LOWESS normalisation (Yang et al., 2002b) that was originally used on global slide area, but later applied on the print-tip groups to handle spatial problem in a discrete approach (Smyth, 2004). Continuous methods using 2D sliding windows (Dudoit and Yang, 2002), optimised local intensity-dependent normalisation (Futschik and Crompton, 2004, 2005) and neural nets normalisation (Tarca et al., 2005) emerged to take care the spatial effects that were often seen for cDNA microarray hybridisations. All the described methods have several parameters affecting the outcome of normalisation. It is hard to choose settings to gain optimal results. Step-wise normalisation performs multiple normalisation methods in sequence and only keeps steps improving the normalisation results. In this way parameter modification is no longer an issue and focus can be on interpreting the data. For a more general overview of normalisation methods see Smyth and Speed (2003).

**OPLS microarray normalisation**

Normalisation using Orthogonal Projections to Latent Structures (OPLS) is in contrast to other normalisation methods mentioned so far as it is not ratio-based or performed within single microarray slides. Approaches to incorporate between-slide normalisation for microarrays, such as spline normalisation (Smyth and Speed, 2003), have been presented in the literature but not really implemented in standard analysis pipelines. Using a two-step approach of ratio-based normalisation followed by between-slide normalisation might be problematic. The initial within-slide normalisation might introduce slide dependency that was not visible before the normalisation step. The Variance Stabilisation Normalisation (VSN) method (Huber et al., 2002, 2003) takes a different approach in performing channel-wise linear and non-linear transformation to reduce dependence between mean value and variance. None of those methods takes the sample and experimental design into consideration. The idea presented in Paper III is to use information on the experimental design in the normalisation process, similar to normalisation based on ANalysis Of Variance (ANOVA) (Kerr et al., 2000). The integration of experimental design in normalisation makes non-correlated variation separable from variation.
of interest (Bylesjo et al., 2007). Unwanted correlation might result from dye, slide and other technical effects. Using the OPLS normalisation method will reduce the noise and make data interpretation clearer.

2.1.6 *In silico* reference designs

Pair-wise relationship in cDNA microarrays makes experimental design more complex than for many other fields. Most early studies hybridised a common reference in one channel of the microarrays to circumvent the problem. The advantage with that strategy is mainly easier data interpretation. Figure 2.2 and Table 2.3 shows the number of biological samples and microarray slides needed for loop and common reference designs (Churchill, 2002). In this particular small example, each biological sample is measured twice as many times for the same number of slides and cost. The bottle necks in microarray experiments are often either economy or amount of biological material. In both cases it is rational to skip the common reference design and instead perform loop or more advanced experimental set-ups. (Vinciotti et al., 2005)

![Graphical representation of two commonly used experimental designs for microarray experiments. A. Pooled reference design containing three different samples compared to a mixture of them. B. A loop design containing three samples compared against each other.](image)

<table>
<thead>
<tr>
<th></th>
<th>Common reference</th>
<th>Loop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Units</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Slides</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Measurements</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

*Table 2.3.* Summary of number of slides and biological material needed for two types of microarray experimental designs, common reference and loop design.
The problem of interpretation in loop designs might be circumvented by using transformation to \textit{in silico} references (Diaz \textit{et al.}, 2003). Paper IV (Sjodin \textit{et al.}, 2006) shows how two experiments using different designs in the end will give the same results. The published \textit{in silico} algorithm used all samples used in microarray experiments. An extended algorithm able to cope with subsets of the samples collection was developed to handle special cases. The method was initially used in Druart \textit{et al.} (2007) to improve the interpretability of the results. The experiment was performed using a pooled reference design. Unfortunately, the pooled reference was not an equal distribution of the biological samples used in the study. The solution was to remove the reference sample from the analysis and focus on biological relevant samples by constructing an \textit{in silico} pool as a reference point. The second problem with the experimental set-up was the huge difference in expression levels through the time-series. Small changes within shorter time periods were masked by dramatic changes in gene expression over the complete series. The time-series was divided into smaller biological relevant time spans by applying the \textit{in silico} reference method. Important smaller changes within each time period could then be visualised.

### 2.2 Turning UPSC-BASE into a valuable resource

A chain is only as strong as the weakest link: A well designed database might be very impressive in its architecture, but totally worthless without interesting content. The following section will describe some of the work to transform the empty, useless database into something valuable for plant scientists. UPSC-BASE was, in the beginning, made to gather microarray results produced from the UPSC \textit{Populus} microarray. The main contribution is still in-house results, but its general architecture has allowed the content to be extended to other \textit{Populus} microarray platforms and other plant species, such as rice and \textit{Arabidopsis}. UPSC-BASE contains at present several published experiments with full open-access to raw data and additional non-published experiments searchable for an external users. This allows extraction of microarray data about interesting genes valuable for planning other experimental approaches.

#### 2.2.1 Integrating transcriptomics with physiology

Transcriptional studies of free-growing \textit{Populus} trees are the laboratory contribution in this thesis. A key difference between \textit{Populus} and \textit{Arabidopsis} is the lifespan and the possibilities to study the plant longer in natural conditions. A perennial plant, such as \textit{Populus}, has a high probability of being exposed to all kinds of biotic and abiotic stresses and conditions. Instead of growing plants in the greenhouse under
2.2. TURNING UPSC-BASE INTO A VALUABLE RESOURCE

different, artificial, constant conditions, the trees were studied in the field with genomic tools. This approach is much less common and has several limitations and complications. However it is a good complement to studies performed in the greenhouse and it will give novel information about the responses caused by different natural conditions.

![Figure 2.3](image-url).

**Figure 2.3.** The *Populus* tree T2 over the growing season. A. Leaf in early June (Phase Ia) B. Fully expanded leaf in mid June (Phase Ic) C. Mid season leaf (Phase II) D. Senescent leaf (Phase III) E-H Overview of the tree corresponding to the same leaf ages as A-D.

The studies of free-growing *Populus tremula* trees started in autumn 1999, as a pre-study of the *Populus* senescence project (Bhalerao *et al.*, 2003; Andersson *et al.*, 2004; Keskitalo *et al.*, 2005). Two *Populus* trees, T1 and T2 were included in various periods of the study. During the first year, samples were collected twice per week. From 2003 to 2006 samples were collected on a daily basis during the growing season, which in Umeå is from late May to early October. See Figure 2.3 for an overview of the growing season in northern Sweden. The growing season of the leaf can roughly be divided in three main phases. Phase I for leaf development during the first month, Phase II when mature leaves are mainly producing energy and Phase III for leaf senescence. In addition to leaf sample collection, several measurements have been performed for leaf length, chlorophyll and florescence. Finally an overview picture was taken of the tree for weather and leaf-state documentation.
Influence of weather on gene regulation in *Populus* leaves

Temperature has a major impact on processes in living organisms. The temperature has been tracked at a weather station at Umeå University campus. Temperature sum was calculated on daily mean temperatures higher than +1°C and this cut-off was based on experiments done on willow trees (Hannerz, 1999). Between 1999 and 2006 the temperature sum showed quite a consistent pattern with some exceptions, see Figure 2.4. The largest difference was seen in 2002, where the spring was unusually warm in late April and early May. Smaller difference in temperature were also observed in 1999 and 2004. In 2004 the temperature showed a peak in early May before it become colder again. The opposite pattern was true for 1999 when it took a long time for the temperature sum to reach normal levels.

![Figure 2.4. Temperature sum recorded at Umeå University during 1999-2006. Most of the years are overlayed with exception of 2002 and smaller time periods of 1999 and 2004.](image)

Weather parameters were used for both seasonal (Wissel *et al.*, 2003; Sjodin *et al.*, 2007b)(Paper V) and autumn senescence projects (Andersson *et al.*, 2004; Bhalerao *et al.*, 2003; Keskitalo *et al.*, 2005)(Paper VII) to keep track of environmental conditions. An extensive weather experiment containing ten samples from two different years is presented in Paper V. The samples were chosen to have large variation in weather parameters and to be within the Phase II. The modelling results showed temperature to be the major impact on gene expression. This could resulted for several reasons, such as too small a dataset to allow separation of noise from biological information. To refine results, the time should been spanning a longer period to capture more difference in weather conditions and ultimately from even more years. Wissel *et al.* (2003) shows the validity of the method for gene-wise modelling and this approach should work at a larger scale, such as using microarrays.
Physiological measurements

Bringing genomic tools out in the field highlights several issues to be solved. For example, electric power needs to be provided during the hours it takes to measure photosynthesis parameters for multiple leaves. Also the weather has an influence on how efficiently the measurements can be made. It is tricky handling electronic equipment during windy or rainy conditions. All those kind of problems had to be solved before starting field studies. An overview of the experimental procedure and equipment can be seen in Figure 2.5.

Figure 2.5. A. Physiological measurements and sample collection performed at lower branches of the tree. B. Dark adaptation on small leaf area for photosynthesis measurements. C. Experimental set up of photosynthesis measurements including laptop and camera. D. Dark adaptation of complete leaves. E. Digitalised images of photosynthesis parameters overlaid on the leaf area.

Briefly, leaf length was measured using a ruler and area from a digital photograph every day through the growing season. At the same time chlorophyll content was estimated using a non-destructive method (van den Berg and Perkins, 2004). The results are partly presented in Figure 2 in Paper V (Sjodin et al., 2007b). In addition, during one complete growing season three leaves were dark-adapted for
30 minutes and analysed using a FluorCam fluorometer. The fluorometer uses a range of flashing lights to measure photosynthesis. One of the most common parameters used in fluorescence is the Fv/Fm ratio. A leaf is kept in the dark to process all residual energy, indicated by small amounts of fluorescence. Then the leaf is flashed with bright light and the fluorescence signal will increase to maximum level (Fm) as photosystem II cannot use all of this energy. The difference between the maximum and minimum fluorescence is the variable fluorescence, Fv. The ratio between Fm and Fv indicates the proportion of the maximum possible fluorescence, which was used for photosynthesis. In normal cases the efficiency is about 80%. As presented in Figure 2.6, Optimum quantum yield (Fv/Fm) increased steadily during the two first weeks and was then stable until the start of senescence. Chlorophyll content on the other hand took almost one month to reach maximum level and breakdown started earlier.

![Figure 2.6](image)

**Figure 2.6.** Optimum quantum yield (Fv/Fm) and chlorophyll content in *Populus* leaves over the growing season of 2004.

An interesting observation was made in 2004 when bud burst was almost two weeks earlier than normal due to warm weather (described above), see Figure 2.4. Unusually cold weather then stopped leaf development and leaves were finally mature around the same time as normal.

**Connecting physiology and genomics**

Measurements in the field give information about the physiological state of the tree during the season. A genomic approach using microarrays was applied and initially presented in Keskitalo et al. (2005); Bhalerao et al. (2003); Andersson et al. (2004). Paper V is based on same procedure but it is extended to the complete growing
season and not only the senescence phase. Weather data were first analysed with multivariate methods to find eleven dates with similar weather to minimise the environmental effect. Before the data were analysed a second generation of the *Populus* microarray became available. Another microarray experiment was performed to cover more genes and to achieve higher time resolution from multiple years. The results showed a developmental trend over the season with high correlation to temperature sum for the first month. To confirm the effect of weather on gene regulation, ten days from two different years were chosen. The results showed that approximately four times more genes are involved in leaf development compared to responses to weather differences, which gives an indication of how *Populus* gene regulation is distributed between developmental processes and stress responses.

2.2.2 Trees in future climate

Several publications during the last years have reported an increase in global atmospheric CO$_2$ concentration. Knowing the effect of CO$_2$ concentration is of high relevance because it might have a major impacts on the response to the environment. It has been shown that higher [CO$_2$] leads to faster leaf development and higher biomass (Taylor *et al.*, 2003). The underlying reasons for those effects were not known and two parallel projects, POPFACE (Calfapietra *et al.*, 2003) and Biosphere (Barron-Gafford *et al.*, 2005; Murthy *et al.*, 2005; Zabel *et al.*, 1999) aimed at finding out how gene expression is regulated in response to altered atmospheric [CO$_2$]. The POPFACE experiment used two levels of CO$_2$ concentration; ambient (350-380ppm) and elevated (550ppm) with the elevated level being selected as it represented the predicted concentration for 2050 (Paper VIII) (Taylor *et al.*, 2005) while Biosphere had three different levels; 400, 800 and 1200ppm (Paper IX) (Druart *et al.*, 2006). Both studies report minor responses to elevated [CO$_2$] at the transcriptome level. Leaves are probably only growing faster and longer and therefore expand more but no major difference can been seen in mature leaves. Reported difference in gene expression might even be an artefact of different physiological leaf ages instead of direct treatment effect (Taylor 2005). It seems strange to continue studying mature leaves in elevated CO$_2$ when the main effect is increased biomass. Leaf developmental series would be of more interest if more leaf transcriptional studies should be performed. Druart *et al.* (2006) found larger differences in wood samples and further experiments in elevated CO$_2$ should focus on increased biomass in wood tissues.
2.2.3 Applying stress treatments

Microarrays are especially suitable for studying large regulation differences in stress treatments. Early transcription profiling studies on *Arabidopsis* were performed for several common stresses, such as dehydration, salinity and cold (Oono et al., 2003; Seki et al., 2002; Kreps et al., 2002). *Populus* shows, in comparison to *Arabidopsis*, much greater genome sequence difference between different genotypes (Marron et al., 2002; Tschaplinski et al., 1998; Harvey and Driessche, 1997), which makes it a good model for studying the influence of different genetic architectures on induced stress responses. Street et al. (2006) (Paper XIII) contribute with valuable transcriptional information about contrasting drought responses in two *Populus* genotypes. Changes in gene expression are often usually heavily induced by stress treatments and only some are directly involved in adjustment while the rest may be secondary consequences. Many microarray studies don’t take variability within or between species into consideration. However, the approaches, such as in Street et al. (2006), of combining genetic mapping and expression tools will ultimately be needed to identify the underlying genetic control of the primary response.

Trees growing in natural conditions are exposed for various other living organisms. The EST library of senescent leaves and bark contains large amount of sequences not originating from *Populus* (Segerman et al., 2007). The presence of non *Populus* ESTs is easy to explain as a result of infection in senescent leaves of the free-growing tree used for construction of the EST library. However, it is harder to know why bark from the greenhouse contains high amounts of foreign ESTs. It is important to understand the similarity and difference between responses derived from different conditions. Controlled studies have been performed examining virus infection (Smith et al., 2004), fungi (Rinaldi et al., 2007) and herbivore attack (Ralph et al., 2006; Babst et al., 2005) to characterise different stress responses. Those individual datasets create a solid base to study stress.

All stress experiments in UPSC-BASE and external published datasets were merged in order to provide a better understanding of the relation between different stress treatments (Figure 2.7). Briefly, individual clones on the microarrays were aggregated based on gene model information and the intercept between all used microarray platforms were analysed. The included *Populus* microarray platforms are the above decribed cDNA microarrays and the Nimblegen full *Populus* genome microarray (Tuskan et al., 2006). The CO$_2$ microarray experiment (Taylor et al., 2005) is not a stress treatment in the traditional meaning but was included in the comparison to relate it to abiotic and biotic stresses. As expected, CO$_2$ treatment did not show any major gene expression changes compared to other stresses. One fungi treatment and ozone are influencing gene expression in similar ways. Another interesting observation is the location of senescence between cold treatment and mosaic virus infection. The reason might be that senescence is affected by multiple
abiotic and biotic factors or that certain stresses induce a response more akin to senescence. Acute drought samples showed a similar behaviour to the senescence and mosaic virus samples, which is far away from the other, less severe drought samples of the chronic drought treatment. This stress comparison is mainly included to show the power of merging data from multiple experiments and platforms to provide an overview that is not seen in the separate experiments alone. Individual genes responsible for the separation between stress treatments may have a huge economical potential for industry. Understanding how trees respond to different stresses is necessary to inform improvement of trees to produce new varieties that are stress-resistant and biomass productive.

![Figure 2.7](image)

**Figure 2.7.** Overview of relationship between stress treatments in *Populus*. Ozon and UV seem to group together in the same way as cold groups together with senescence and acute drought with mosaic virus infection.

### 2.2.4 How to use stored information

The information in PopulusDB (Segerman *et al.*, 2007; Sterky *et al.*, 2004) is mainly used in other projects to produce digital Northernns (Ewing *et al.*, 1999). EST libraries constructed with non-subtractive techniques allow comparisons between abundance of ESTs in different tissues and stresses. Interesting results from microarray experiments can be evaluated against the EST distribution as a means of
confirmation. This method has been used to classify and validate results in wood cell death (Moreau et al., 2005), drought stress (Paper XIII) (Street et al., 2006), protease distribution (Paper XII) (Garcia-Lorenzo et al., 2006), seasonal gene regulation (Paper V) (Sjodin et al., 2007b) and for leaf specific genes (Paper VI) (Sjodin et al., 2007a). It provides an external validation without needing to undertake any extra laboratory work. It provides some weight to the results, even if it provides only circumstantial validation support.

Another way to maximize the produced data is to re-analyse with different views and questions. Results from the first microarray experiment from UPSC, a wood formation series (Hertzberg et al., 2001) have been used extensively to direct other in-depth results and as a basis for selecting genes to targets by gene silencing (Schrader et al., 2004b,a; Moyle et al., 2002; Israelsson et al., 2003, 2005). Several research projects are spin-offs from this initial wood transcriptional study, but so far no large-scale database data mining has been performed for Populus microarray data.

2.3 Reconstruction of gene regulatory networks in Populus

The aim of the outcome from this thesis was, from the beginning, to find regulons in Populus leaves. A regulon is defined as a group of co-regulated genes (Dong and Horvath, 2007; Zhang and Horvath, 2005; Yip and Horvath, 2007). Genes with highly correlated expression levels are biologically interesting, since they are assumed to have common regulatory mechanisms. Paper VI shows how generated microarray data was used to find co-regulated genes in Populus leaves (Sjodin et al., 2007a). Understanding complete systems instead of reporting a list of isolated genes adds another dimension to biological research. Here the focus on regulons (modules in network terminology) as opposed to individual genes. Tissue specificity of gene regulation was predicted in addition to complete leaf regulon determination. Briefly, an experiment containing several different tissues was used, and modelled using OPLS to find genes with significantly higher expression in specific groups of tissues. 1116 gene models were significantly more highly expressed in leaves samples. Not surprisingly, leaves showed high amounts of photosynthesis related genes together with cell differentiation, structural molecule activity and lipid metabolism, as seen as the Plant Gene Ontology overview in Figure 3 in paper VI.

Those genes are, according to the Populus dataset, very characteristic for leaves. However drawing conclusions from only one dataset might be very dangerous. Results were confirmed using two additional dataset, from Populus (Tuskan et al., 2006) and Arabidopsis (Schmid et al., 2005). The two Populus datasets had around 451 gene models in common, corresponding to 40% of the gene models in the
Figure 2.8. Overview of the over-represented Gene Ontology groups in leaf specific genes. The colour indicates how significant the over-representation is within each Gene Ontology category and the size is proportional to number of included genes in each category.
initial dataset from POP2 microarrays. The leaf specific genes contained high over-
representation of Gene Ontology categories connected to photosynthesis. Figure
2.8 shows the three different categories for the full Gene Ontology. It is in principle
the same figure as shown in Paper VI but the complexity in better visualised in
the full Gene Ontology. Especially cellular component pinpoints the connection to
photosynthesis components.

All available leaf experiments in UPSC-BASE (Sjödin et al., 2006) were downloaded
and step-wise normalised. To not bias the results based on different EST abundance
on the microarray slide, the results were aggregated based on gene model and
networks of gene expression data were constructed by calculating correlation for all
gene pairs. Gene expression network analysis aims to identify modules and hub genes
(i.e. highly connected genes) present in the dataset. Microarray data is often noisie
and correlation values were weighted to emphasise strong correlations (Zhang and
Horvath, 2005). Groups of genes with similar patterns of expression were identified
inside the network. Those groups of genes will be defined as modules and represents
the regulons of gene expression in Populus leaves. Each of them can be seen as a
new network with tight internal connections. Hub genes have been shown to be
important key regulatory genes (Lehner et al., 2006; Carlson et al., 2006).

The focus in the paper was on gene regulation of leaf development and transcription
factors controlling gene expression. Developing tissues share many common
processes and gene models particularly helpful for discriminating leaf tissues from
other tissues were selected to pinpoint regulation that might be specific to leaves.
The constructed networks based on this selection showed clear separation between
two main types of genes; photosynthesis and cell differentiation. Results were con-
firmed using external experiments from Arabidopsis (Schmid et al., 2005), multiple
Populus microarray platforms (Matsubara et al., 2006; Tuskan et al., 2006; Street
et al., 2006) and digital Northern (Segerman et al., 2007; Sterky et al., 2004; Ewing
et al., 1999) to not be biased by microarray results from a single source. A more
novel approach in Paper VI is the combination of microarray results and QTL for
leaf development (Street et al., 2006) to screen for candidate genes that may be
responsible for the control of leaf development. This approach has huge potential,
but at the moment it is very limited due to lack of publicly available QTL data
from published articles. The information provided in QTL publications is nearly
always not sufficient to allow re-analysis or re-plotting of the results. Other fields in
biology should learn from the microarray community and develop their own stan-
dard rules for publishing corresponding to the MIAME standard (Brazma et al.,
2001). The outcome of this collocation analysis was the identification of two good
candidates transcription factors (YABBY and zinc finger homeodomain) that have
been shown to be important in leaf development.
Conclusions

In the first part of this thesis I show how to build a pipeline for streamlined microarray analysis from actual lab work to interpretation of regulated genes. Setting up routines for hybridisation, scanning, image analysis and normalisation significantly increases the quality of the generated biological data. Securing the data quality is a crucial task to allow other scientists to trust and use the produced data. As a final part of the work flow, a database was developed for the storing and analysis of data. The database provides large value for experimental biologist by letting them access advanced methods from a web interface.

In the second part of the thesis I show how the created infrastructure tools can be used to convert the database to a useful resource of transcriptomic information. First, as a comprehensive study bringing genomic tools out in the field and then a general overview of other project spanning different stresses and developmental stages. The goal with all of the experiments has always been to span as many different developmental stages and environmental conditions as possible to be able to reconstruct the gene regulatory network to predict regulons in *Populus* leaves. The last paper connects all earlier *Populus* leaf microarray experiments in a meta-analysis to achieve this goal. The regulon information is a valuable resource for constructing gene markers for fast fingerprinting using cheaper low-throughput technologies, such as RT-PCR. This will make it possible to perform large-scale fingerprinting in populations of trees, such as SwAsp (Luquez et al., 2007; Hall et al., 2007).

The draft genome has made it possible to work on putative gene models instead of contigs based on EST sequences. It has helped the *Populus* field to take a step forward, but the immature state of the *Populus* genome is still something of a bottleneck. Gathering the *Populus* community to work in an open-access environment is necessary if *Populus* should gain any importance for plant science outside the existing small *Populus* field. More focus should put on learning from the mistakes and solutions made by the *Arabidopsis* community. The main issue to be solved is the data sharing problem. (Last, 2003; Cech et al., 2003) *Populus* will never be a widely-used model organism if the data sharing issue isn’t resolved. An important decision to take is a uniform naming of gene models. At present the gene models in *Populus* are named according to best gene model prediction from four different softwares. Names containing up to 50 characters are not very convenient to handle.
This systems should be replaced by an implementation of the relative position name system used by the the Arabidopsis Genome Initiative (Mayer et al., 1999). Solving the name issues in Populus would simplify work significantly and make the genome far more accessible to the wider community.

It was clear from the microarray platform comparison that the overlaps between different microarrays were minor compared to the total amount of predicted genes in the Populus genome. According to the calculation only 24687 of the 45555 official gene models have EST support. Constructing a universal, affordable full genome Populus microarray is urgently required to allow more advanced large-scale meta-analysis and to confirm the validity of the gene models.

Although this work has developed tools and generated information about gene regulation in Populus leaves, it is still not known in detail how gene regulatory systems works during leaf development or autumn senescence. The processes are too complex to be solved by a single approach. Further development and integration of methods studying transcriptomic, metabolomics and proteomics as well and physiological data is needed. At the same time ‘omics’ technologies can’t only be used alone on plants in artificial environments. To understand the complexity in biological systems the modern ‘omics’ field needs to merge with genetics and ecology. The will allow furthering of understanding of how variation within and between different biological systems is achieved and controlled.

3.1 The next generation Populus genome browser

UPSC-BASE is a valuable resource for transcriptomics in Populus. However, end-users need to have knowledge about the microarray technology to handle and sort stored information. There is now enough microarray data publicly available in Populus to allow creation of a Populus resource similar to Genevestigator. This resource would allow extraction of gene expression data for genes of interest without the need to understand advanced analysis tools.

During the work with microarray and sequence data for Populus it has become clear that the present official JGI (Joint Genome Institute) genome browser doesn’t fulfil the demands of advanced or the average user of the genome resource. A major problem with the browser is a lack of data sharing and description of how the data has been collected and treated. The Arabidopsis Resource (TAIR) have recently decided to exchange their genome browser to an open-source solution used by many model systems. Efforts need to be taken if Populus to ensure that poplar remains as the model systems for trees now that additional genomes such as grape and eucalyptus are on the way to being sequenced. A Populus genome browser based on the GMOD (Generic Model Organism Database) project (Stein et al., 2002)
would solve those problems. The GMOD project includes WormBase (Stein et al., 2002; Harris et al., 2003, 2004; Chen et al., 2004, 2005), FlyBase (Crosby et al., 2007), Mouse Genome Informatics (Eppig et al., 2005, 2007), Gramene (Ware et al., 2002a,b), the Rat Genome Database (Twigger et al., 2007), TAIR (Huala et al., 2001), EcoCyc (Karp et al., 2007; Keseler et al., 2005), DictyBase (Chisholm et al., 2006; Kreppel et al., 2004), wFleaBase (Colbourne et al., 2005) and Saccharomyces Genome Database (Dwight et al., 2004; Cherry et al., 1997). Collecting data from various public resources in a public database makes it possible to spend effort on biological interpretation instead of bioinformatic methods to merge data. This resource is valuable both for bioinformatics performing global-scale analyses as well for 'favourite gene' studies. BLAST searches against subsets of gene models, such as the Jamboree set, are easily performed in contrast to the JGI genome browser, where all BLAST searches are performed against unfiltered genome sequence. A search at JGI will always result in multiple gene models at the same physical position on and results need to be filtered manually. A genome browser database will allow integration of additional data sources, such as QTL or microarray data. It would then be possible to easily extract expression and QTL information about interesting genome regions. As an integrated resource, end-users would achieve the most out of generated biological data.

3.2 Future perspectives

In the future, tools and information presented in this thesis could be used in several ways. An option for further work would be to combine transcriptomics with genome sequence information so that it would be possible to find and understand how key regulatory motifs control gene expression. Finally, it would be interesting to study in greater detail key genes in leaf development, both in natural populations and in transgenic plants. More work is needed before we can understand these complex processes and ultimately produce trees optimised for growth in specific conditions.
Acknowledgement

Stefan Janson  first I would like to thank you as my supervisor believed in me from the beginning and gave me the opportunity study whatever I wanted. . . (mainly outside my PhD project)

Collaborators and co-authors  My PhD time wouldn’t be the same would all national and international collaboration I have been involved in.

Jansson group  Former and present members of the Jansson group during my PhD: Kirsten, Jenny A, Ulrika, Rupali, Carsten, Johanna, Frank, Abdul, Oskar, Virginia, Yvan, Jakob, Hanna, Lars, Roger . . . Martin is not to be forgotten. Thanks also for the biochemistry part, Maribel, Adriana and Agnieszka for fun moments.

Jan Karlsson  my former master thesis supervisor that together with Marisa and Jenny J (tried to) introduced me to the plant biology. I am not really sure that you really succeeded . . .

Umeå Plant Science Centre  I will miss all nice colleagues in the enormous fika room and your biological knowledge. You are welcome to contact Sjödin Consulting for technical support and questions . . .

My project students  during the years several students helped me to escape from the lab bench and to be outside in the rain together with the Aspen tree. Tack så mycket, Johan, Daniel, Sara, Sylvia, Therese, Ali and Julia.

TBi plant team : The first engineering students going for plant science, Linus, Jens and Lars. Now we are just waiting for you Linus . . . I am also grateful to see how many that followed us to USCP and the plant field.

The chemometric nerds  Max Bylesjö and Mattias Rantalainen, thanks for the scientific and social discussions at different bars around the world . . .

The fish tank  My office has moved several times during my PhD time but the last 4 years have been in the Fish tank with smaller modification of the crew. Version 1.01 (Charleen, Nathalie, Oskar), 2.0 (Charleen, Micke) and 2.5 (Nat, Micke). You made the work so much funnier during the boring periods. ’Julen varar ända fram till påskas!!!’ or in our case for several years.
The French mafia for introducing me to all strange French food and culture. I promise to be a better French language student in the future. Shout ‘putain d’ordinateur de merde’ and you will have free computer support...

Midsummer Eve heroes Matthieu and Gaia (2004) and Nadia (2003), your leaf sampling on the Midsummer Eve was invaluable! I still owe you a big favour. Just tell me when you need it!!

Nadia Goué for always being my hardest critics. The cup quote ’It is never too late to be what you might have been’ motive me through the dark months of writing. I hope you found your place in the country of sun rise to let people around you enjoying your happy face. I miss you and the ‘gympa’...

Oskar Skogström it has been nice to have you here during all years in Umeå. Even if you have been a little bit ’invisible’ the last years...

Natalie Druart and Charleen Moreau you made a brave attempt to teach me French. However I am sorry to tell I almost only remember ’Ta Gueule!!!’...maybe because it was the only time you both got silent in the office.

Sofia Frankki for convincing me to stay in Umeå and all nice moments together during the years. You will always remind me about Star Wars.

Mélanie & Yohann thanks for nice company and dinners. You gave me a new perspective on French people ...

Rozenn Le Hir you finally realized that nothing is impossible as long you want and dare to do it...Be brave girl!

Nathaniel Street deserves special thank for nice company around the world (Umeå, Southampton, Gatlinburg, Pretoria and Venice). The last months of work wouldn’t have been possible or done without your help and motivation. It is sad that we couldn’t work together for longer time. Let’s enjoy Sweden now!

Joséli Schwambach Hej, hej, hemskt mycket hej! The Lofoten trip was a great vacation (at the right moment to forget work). I miss the all nice tea breaks with you. Next time in Brasil!

Dephine Gendre for all cool moments and travels. I will miss you and your honest opinions in our discussions! Nobody can climb like you...ps. Don’t forget to call me when you want to open a café!!!

Sunday movie It has been a really nice to finish the weekend with nice company in front of a good movie. Keep the traditions going!

Family here is your opportunity to understand what I have been doing the last years in Umeå...ps. Det finns en svensk sammanfattning...å tack lillasyster för det fina omlaget! Troligen är det fler som lägger märke till ditt arbete än mitt...


