Gene regulation and the emergence of phenotypes – a network approach

Genregulering og emergens av fenotyper – en nettverksstudie

Philosophiae Doctor (PhD) Thesis

Niklas Mähler

Dept. of Chemistry, Biotechnology and Food Science
Norwegian University of Life Sciences

Ås 2016

Norwegian University of Life Sciences

Thesis number 2016:52
ISSN 1894-6402
ISBN 978-82-575-1372-6

Summary

The emergence of complex traits in living organisms has been of interest to biologists since the early days of biology. Domestication and breeding has resulted in the most remarkable transformations, such as the grass teosinte turning into the cornstalks of today, or the variation that can be seen among different breeds of dogs. In the mid 19th century, Gregor Mendel uncovered the basics of genetic inheritance and was able to explain the passing down of traits in peas. Not all traits are this simple to dissect, however, since they are controlled by several genes; these are collectively called complex traits. Due to the large number of genes in a species, it is simply not possible to explore the space of all gene combinations exhaustively. New sequencing technologies however makes it possible obtain so-called omics data on multiple aspects of a biological system and these data can be integrated in order to narrow down the search space and focus on the functional gene combinations. Moreover, several of these data types are much closer to the phenotype than data on variation in genome sequence. The changes in genome sequence are manifested as changes in gene expression levels, or changes in protein sequences in turn leading to changes in protein function, protein interactions, metabolite levels and gene regulation. In order to obtain a complete picture of how phenotypes change based on changes in genome sequence, these intermediate layers must be included as well.

In this thesis, we aim to shine some light on gene regulation and the emergence of complex traits. In paper I, gene regulation in the cyanobacterium *Synechocystis* is explored by integrating regulatory motifs with co-expression networks, and a web tool is developed to make the results interactively available to the research community. Paper II investigates the sexual dimorphism in *Populus tremula* using data on phenotype, gene expression, and genotype. In paper III, the focus is directed towards the genetic component of gene expression variation and how this can be understood in the context of a co-expression network. Finally, paper IV expands on paper III by adding genotype–phenotype associations, in addition to eQTLs and gene expression, in order to dissect leaf shape in *Populus tremula*.

Sammanfattning

Uppkomsten av komplexa egenskaper i levande organismer har intresserat biologer under lång tid. Domesticering och avel har resulterat i dramatiska förändringar av arter, till exempel förvandlingen av gräset teosinte till dagens majsstänglar, eller variationen vi kan se mellan olika hundraser. I mitten av 1800-talet upptäckte Gregor Mendel de grundläggande principerna bakom genetisk nedärvning. Dock är inte all egenskaper lika enkla att förklara då de kontrolleras av mer än en gen. Denna typ av egenskaper kallas gemensamt för komplexa egenskaper. På grund av det stora antalet gener i en organism är det helt enkelt inte möjligt att utforska alla genkombinationer för att försöka förklara dessa egenskaper. Nya sekvenseringsteknologier gör det dock möjligt att samla så kallade omics-data som kan fånga olika aspekter av biologiska system, och detta data kan kombineras för att reducera antalet genkombinationer till de som mest troligt bidrar till själva egenskapen. Utöver detta så är många av dessa datatyper "närmre" den slutgiltiga egenskapen jämfört med genomsekvensen. Förändringar i genomsekvensen uttrycker sig som förändringar i genuttrycksnivåer, eller förändringar i proteinsekvenser som i sin tur leder till förändringar i funktion hos proteinerna, interaktioner mellan proteinerna, metabolitnivåer och reglering av gener. För att kunna få en komplett bild av hur komplexa egenskaper förändras baserat på förändringar i genomesekvensen måste dessa mellanliggande lager av reglering inkluderas.

I denna avhandling undersöker vi genreglering of komplexa egenskaper för att försöka få en klarare bild av hur detta fungerar. I artikel I undersöker vi genreglering i cyanobakterien *Synechocystis* genom att integrera regulatoriska motiv med co-uttrycksnätverk. Även en webbapplikation utvecklades för att tillgängliggöra resultaten. Artikel II ser på könsdimorfism i asp (*Populus tremula*) genom att använda data på fenotyper, genuttryck, samt genotyp. I artikel III riktas fokus mot den genetiska komponenten av variation i genuttryck och hur en klarare bild kan erhållas genom att se på detta ur perspektivet av ett co-uttrycksnätverk. Slutligen expanderar artikel IV på resultaten från artikel III genom att lägga till genotyp–fenotyp–associationer för att försöka förklara skillnad i bladform hos asp.

What I cannot create,
I do not understand.

RICHARD FEYNMAN

Acknowledgements

First and foremost I would like to thank my supervisor, Torgeir. When I started to get genuinely interested in bioinformatics back in 2010, he was the one who gave me my first bioinformatics project, and see where that led me. Always pushing me just enough, and having a knack for asking me the right questions.

I am forever grateful to my parents, always supporting me in whatever endeavour I have embarked on, and for always making sure I have both feet on the ground.

Thank you, Rickard and Peter, for the late-night Battlefield sessions in the past year that have worked wonders for me in order to blow off some steam.

Without all the people at Umeå Plant Science Centre I don't think I would be where I am today. It was here that I got the inspiration to pursue a PhD in the first place. Thanks to David for nice discussions and for lending out his couch whenever I visited, and a big thanks to Jing for contributing to the more hard-core genetics side of things. A special thank you goes out to Nat and Tiggy for providing me with interesting data and for guiding me through the more hairy biology. Working together with you has been a huge source of inspiration to me, and for that you have my gratitude.

Last but not least, the Biostatistics group. What an amazing collection of people, always having their doors and minds open. It has been a privilege sharing the workplace with you, and I will miss the coffee breaks in the appendix, the fishing stories, the marathon/Birken/factor discussions, the food festivals, the Bollywood dancing, as well as all the electric car propaganda. There were seldom boring moments, much thanks to the party committee led by Navreet and Guro. You are all awesome!

June 6, 2016
Ås, Norway
Niklas Mähler

List of papers

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

I. Mähler N, Cheregi O, Funk C, Netotea S, Hvidsten TR (2014) *Syn*ergy: A Web Resource for Exploring Gene Regulation in *Synechocystis* sp. PCC6803. PLoS One 9: e113496

II. Robinson KM[*], Delhomme N[*], Mähler N, Schiffthaler B, Önskog J, Albrectsen BR, Ingvarsson PK, Hvidsten TR, Jansson S, Street NR (2014) *Populus tremula* (European aspen) shows no evidence of sexual dimorphism. BMC Plant Biol 14: 276

III. Mähler N, Terebieniec BK, Wang J, Ingvarsson PK, Street NR, Hvidsten TR (2016) The genetic architecture of gene expression natural variation in a forest tree suggests buffering of central genes. Manuscript

IV. Robinson KM[*], Mähler N[*], Terebieniec BK, Hvidsten TR, Street NR (2016) A systems genetics approach to understanding the control of natural variation of leaf morphology in European aspen. Manuscript

[*]Equal contribution

Paper contributions

These are my contributions to the papers included in this thesis.

I. Performed all analysis, implemented the web application, and drafted the paper.

II. Ran and designed the support vector machine analysis and wrote the corresponding parts of the paper.

III. Performed the eQTL mapping, constructed the co-expression network, performed all analyses related to these, and drafted the manuscript.

IV. Performed GWA and all gene expression related analysis, and drafted the manuscript.

Table of Contents

1

Introduction

The aim of this thesis is to shed light on the complex matter that is complex traits, and how these traits emerge from networks of interacting genes.

1.1 Background

Some phenotypes are simple to explain from a genetic point of view. Perhaps the most famous example are the experiments conducted by Gregor Mendel between 1856 and 1863. He crossed pea plants (*Pisum sativum*) having different properties; some were green, some were yellow, some had wrinkly seeds while others had smooth seeds. He then observed how these traits were passed down to the next generation of pea plants and thus laid the foundation for what we today refer to as the laws of Mendelian inheritance. This is something that is so fundamental in biology today that parts of it are taught already in primary school. Perhaps Mendel was just lucky, because most traits are much more complex than the ones he described [1–3]. One example of a phenotype that

has been notoriously difficult to explain is human height. Human height has a very high heritability, i.e. a large portion of the variation in human height can be explained by genetic differences [4], but studies that have tried to identify the genetic factors that contribute to the variation in human height have not succeeded to explain things as well and as simple as Mendel explained the colour of his peas. In a recent study, a large team of researchers identified 697 genetic variants that were associated with human height, and together these variants only explained 20 percent of the variation [3]. This is the definition of a complex trait; a trait that can only be explained by the combination of a large number of small effects that individually can be very hard to detect. Another way of putting it is that the complex traits are emergent properties arising from a combination of smaller factors that in themselves are not as complex. These types of traits are often also referred to as non-Mendelian or polygenic, indicating that they are determined by multiple genes, and possibly the interaction between these genes. It is important to note, however, that the individual effects that contribute to the trait do follow the laws of Mendel—it is just that the combination of these effects manifests themselves in a way that does not allow the trait to be dissected in the same way as pea colour.

In the case of Mendel's pea plants, the traits were very visible and the difference between plants was easy to assess. There is however more subtle variation in all natural populations. In humans, variation might be manifested as susceptibility to disease, and in trees, this could be something as minuscule as the texture of the bark or the width of the leaves. This variability can be a result of genotype alone, environmental factors alone, or a combination of these. A long lived debate in this area of research is the nature versus nurture debate, i.e. whether a particular trait arises from genetics or from the environment. Today the general consensus is that most traits are a result of both genotype and the environment that the genotype is subjected to. To add additional complexity, there is also interaction between the genotype and the environment (or G$\times$E), a phenomenon where different genotypes respond differently to changes in environment [5]. Twin studies have often been used to study these types of interactions. For example, if identical twins separated at birth remain very similar, albeit not absolutely identical, for a particular phenotypic trait, this

trait is inferred to be under tight genetic control [6]. Consequently, the trait is then also highly heritable and is environmentally invariant.

Figure 1 shows an overview of how information travels from the DNA via RNA to proteins and metabolites (the central dogma of molecular biology) and how these interact to give rise to complex phenotypes. This introduction will go through this figure one concept at a time and explain the underlying biology, the data we retrieve from this biology, and finally, methods used to analyse the data.

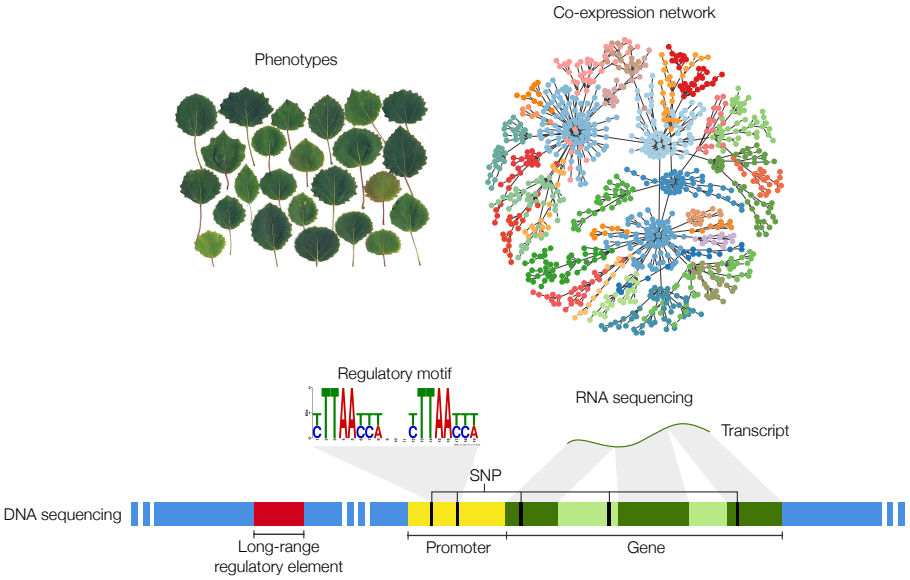

Figure 1: An overview of the different types of data that has been used in this thesis. As the foundation we have the genome sequence with its genes and variation. This gives rise to complex traits by expressing genes that in turn interact with each other in biochemical pathways that in the end can be observed as, for example, leaf shape.

1.2 Genetic variation

1.2.1 The molecule of life

All living organisms have at least one thing in common: they have a genome. It will not look the same in different species, or even individuals of the same species, but the fundamentals are the same; there are four nucleotides, adenine (A), cytosine (C), guanine (G) and thymine (T), that form the molecule deoxyribonucleic acid, more commonly known as DNA. The well known double-helix structure of DNA was discovered in the 1950s, and in the paper by Watson and Crick [7] the second sentence reads: "This structure has novel features which are of considerable biological interest." This might be one of the biggest understatements in modern science. The DNA is organised into larger units called chromosomes, and the number of chromosomes vary from species to species. Humans, for example, have 23 chromosomes and is a diploid organism—it has two copies of each of the chromosomes. European aspen (*Populus tremula*) has 19 chromosomes, and is also a diploid organism. Being diploid means that every gene (and most other pieces of DNA for that matter) exist in two copies—two alleles. Whenever a cell divides, the genetic information has to be copied so that each of the daughter cells gets their own copy of the genome. With this process, perhaps *the* most fundamental property of biology manifests itself—erroneous copying of DNA. Without errors in this process, life as we know it would not evolve. These errors introduce variation into the genetic material, and this variation can take different shapes. Errors in the DNA are known as mutations, and one type of mutation are single nucleotide polymorphisms (SNPs). As the name implies, this type of mutation changes a single base in the genome into another, and these are the type of mutations this thesis will mostly focus on. However, we will also look at more elaborate mutations such as the duplication of genes or the entire genome.

The central dogma of molecular biology states that information flows from DNA to protein via messenger ribonucleic acid (mRNA), and information cannot flow from protein to DNA [8]. When we talk about genes in this context, we mean the parts of DNA that are transcribed into mRNA, and eventually translated

into protein. Since the DNA alphabet only contains four letters, and the protein alphabet contains twenty letters, there is not a one-to-one relationship between mRNA and protein, but units of three nucleotides (codons) define one amino acid, which constitute the building blocks of proteins. Proteins then act as the workers and the building blocks of the cell. The parts of the DNA that are translated into proteins are referred to as coding DNA, while other parts of the DNA are referred to as non-coding. Non-coding regions of the genome can also be transcribed and mostly have regulatory functions, but also act as structural elements, for example 16S ribosomal RNA [9].

In the mid 19th century, traits were believed to be blended when inherited, but Mendel's experiments showed that this was not always the case. From experiments he concluded that there must be different variants of some hidden factor that give rise to the differences in traits in the offspring generation. These factors are what we today refer to as genes, and the variants of these genes are alleles.

If mutations are introduced into coding regions of the genome, one of three things might happen: no effect at all (silent mutation), an amino acid substitution (mis-sense mutation), or the introduction of a stop codon that will prematurely halt the translation process (non-sense mutation). Fifteen years ago, these types of mutations were the focus of biological studies as everything outside of genes was largely discarded as non-functional "junk DNA". Since then, with the arrival of cheap and high-throughput sequencing technologies, the focus and understanding has changed. Although the majority of "junk DNA" is not expressed and translated explicitly, it does facilitate or influence the expression of genes and can contribute to the control of when and at what levels genes are expressed. These parts of the genome consist of, amongst other components, promoters, enhancers, and non-coding RNAs (microRNA, long non-coding RNA [lncRNA or lincRNA], transport RNA [tRNA], ribosomal RNA [rRNA], etc.) which all have different roles in regulating gene expression. tRNA and rRNA are integral components in translation of mRNA to protein, while other types of non-coding RNA have been shown to have regulatory properties [10,11].

It might sound as though regulatory DNA is something that has been discovered during the past fifteen years, but this is not the case at all. Regulatory elements in non-coding regions of the genome have been known and, to some extent, elucidated since at least the 1960's with the description of the regulation of the *lac* operon by François Jacob and Jacques Monod [12]. Even though these types of regulatory mechanisms have been known for a long time, it is only the developments in the past 10 years or so that have made large scale analysis of these types of regulatory mechanisms possible. This component of the genome is today commonly referred to as the regulatory genome, and a plethora of studies have emerged that identify and elucidate the biological function of this in more detail [13], such as the ENCODE project that has the goal of identifying all functional elements in the human genome [14][1]. Gene expression and some more details of the regulatory genome will be presented in more detail in section 1.3.

Given the diversity of the genome in terms of function, it is very hard to predict what effect different mutations will have on individual phenotypes. While it is easy to predict the effect that mutations in coding regions will have on the amino-acid composition of a protein, predicting the effects that this change will have on protein function is less simple. To then understand how that altered function will later influence phenotype is substantially harder again. Understanding, from sequence alone, the effect of mutations that modify gene regulation are much harder still, and they usually require extensive experimental validation [15]. New efforts, such as ENCODE, will enable researchers to more easily determine what effect mutations will have.

Humans and chimpanzees share as much as 99% of the coding regions of the genome, and a lot of research has focused on discovering the genomic differences that give rise to the phenotypic differences between humans and chimpanzees. Several studies have found that most of these differences are located in non-coding regions, i.e. potential regulatory regions [16,17]. So far, most of this variation has only been quantified; developing an understanding of how these differences determine functional effects is a challenge at least an order of magnitude more complicated. Increasingly, efforts are being poured into the

[1] One could also argue the projects like ENCODE help drive the technological development.

problem of predicting the effect of mutations in non-coding regions. In the past few years we have seen the development of tools that try to predict the effect that SNPs will have on transcription factor binding affinity [18,19], as well as tools aiming to predict general regulatory effects [20] based on information in existing databases.

Another important source of genetic variation are gene and genome duplications. Returning to the comparison between humans and chimpanzees, studies have shown that gene duplication plays an important role in explaining phenotypic differences [21,22]. In addition, it has been shown that single gene and whole genome duplications play an important role in speciation in plants, i.e. the formation of new species [23,24], and that they likely explain Darwin's "abominable mystery"—the explosive radiation of species in the angiosperm lineage [25,26]. Approximately 15% of angiosperm speciation events are accompanied by a genome duplication event [27], and all flowering plants share at least one genome duplication event in their evolutionary history [28,29]. In *Populus* species (poplars, aspens, and cottonwoods), a whole genome duplication event occurred about 65 million years ago [30].

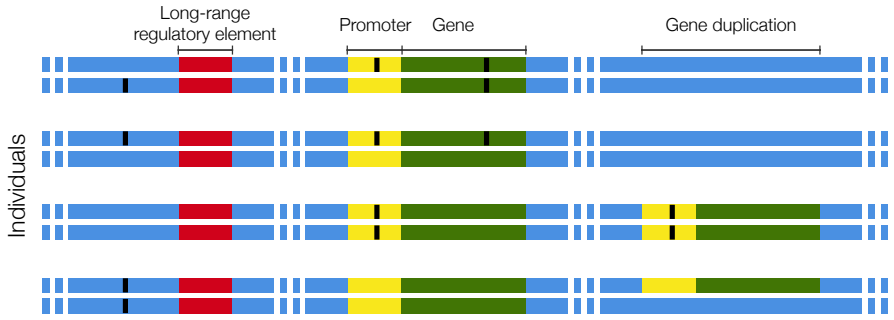

Figure 2: Schematic example of genetic variation. Each of the four diploid individuals has two alleles for each locus representing intergenic sequences (blue), long-range regulatory elements (red), promoters (yellow), and genes (green). Polymorphisms where one of the alleles does not match the reference is indicated by black lines. A gene duplication is illustrated as well where individual three has a duplication of both alleles while individual four only has a duplication of one of the alleles.

1.2.2 Quantifying genetic variation

Technological advancements in the past two decades have led to a revolution in biology. Genome sequencing, i.e. the process of determining the order of nucleotides in the genome, has become very affordable. The $1000 human genome has been a long-time vision, and during my PhD period, became a reality [31][2]. It has never been this cheap or easy to obtain the complete genome sequence of an organism, and this clearly has huge potential for characterising the genetic variation among individuals in a population.

The process of sequencing an individual involves extracting the DNA, randomly fragmenting the DNA, and then determining the sequence of nucleotides for each DNA fragment. The sequencing is then performed until the mean number of sequenced fragments, or reads, for each position in the genome reaches the required depth. There are a number of ways that genomic variation can be quantified from high-throughput sequencing data, but the most common approach today is to align the sequencing reads against a reference genome, that is, a genome sequence that has already been determined. With this approach it is possible to quantify genetic variation by comparing the read sequences with the reference sequence. In the case of diploid organisms we expect to see two alleles for each locus. If the locus is homozygous, i.e. the two alleles are identical, then the reads originating from that locus should be identical. Conversely, if the locus is heterozygous, i.e. the two alleles are different, then the reads should ideally divide into two groups of equal size. Depending on the number of reads that support the variant and the quality of the reads, the variant will be detected, or called.

Different types of prior knowledge can be incorporated in the variant calling in order to increase precision, such as known variants from databases such as dbSNP [32]. Working with non-model, or even non-human organisms, often mean that these types of resources are not available, at least not to the same extent.

[2] Depending somewhat on how you count.

1.3 Gene expression

Genetic variation does not really have any significance if it does not manifest itself in a way that alters phenotype, and in an evolutionary perspective, affects survival or reproductive fitness. One way that it can manifest itself is through gene expression. Genes are pieces of DNA that are transcribed into messenger ribonucleic acid (mRNA), and subsequently translated into proteins. This is a very simplified view of the biological reality involved in these processes. At every step of the process there are different forks in the road that can be taken, and each of these forks will change the fate of that gene. Such forks can lead to a gene eventually being translated into protein, or it might result in splicing out part of the gene before translation into a protein, therefore effectively producing an alternative protein from the same gene, or it could lead to the degradation of the mRNA, among numerous other examples. These processes are also highly dynamic, responding to different kinds of stimuli, such as environmental changes.

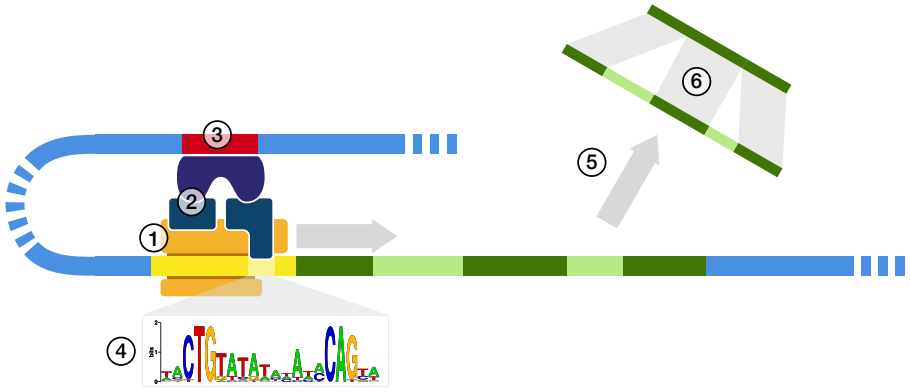

Figure 3: Schematic overview of gene expression. The transcriptional machinery including the RNA polymerase (1) is recruited to the promoter of the gene to be expressed by transcription factors (2) that bind to the promoter and possibly enhancers (3) through specific motifs in the DNA (4). The DNA is then translated to RNA by the RNA polymerase (5) and introns are spliced out (6) before the mature mRNA is translated into protein. This is a simplified view of how coding regions are transcribed.

1.3.1 Regulation of gene expression

As stated in section 1.2, the regulatory genome has received ever more attention throughout the last 15 years. New studies increasingly identify examples where protein sequence is identical between vastly different phenotypes, but where changes in gene regulation is instead responsible for the phenotypic variation. Examples of this are beak length [33] and beak shape [34] in Darwin's finches, and the previously mentioned differences between humans and chimpanzees [16,17]. In the study of Darwin's finches by Lamichhaney et al. [34], one of the genes that was associated with the differences in beak shape was a transcription factor (TF). Transcription factors are proteins that bind to promoter regions upstream of genes and that consequently recruit the transcriptional machinery involving the RNA polymerase (figure 3). The regions of a genome that transcription factors recognise are commonly referred to as motifs. They are short DNA sequences with a specific composition that is meant to match the binding residues in the active site of a transcription factor protein. These motifs are often degenerate, i.e. some positions in the motif can have a number of different DNA bases without affecting function, and can thus be difficult to detect [35]. It is also very difficult to predict the effect of a single mutation in one of these binding sites. One recent study characterised the effect on gene expression by somatic mutations in cancer tumours and found that many of the genes displayed altered regulation as a result of mutations in transcription factor binding sites [36]. This study emphasises the important role that the regulatory genome plays in complex disease.

There are several computational approaches for identifying regulatory motifs, and perhaps the most common method is to compare regulatory sequences (e.g. promoters) thought to be used by the same transcription factor, and identify common regions corresponding to binding sites (motifs) among these sequences. Sequences to consider could be the promoters of genes with similar expression profiles (co-expressed genes) or of genes involved in the same biochemical pathway. One way of increasing detection power is by including regulatory sequences from multiple related species in addition to the species studied, so called phylogenetic footprinting [37,38]. The promoter regions of orthologous

genes—the same gene in different species—are compared, and assuming that transcription factor binding sites accumulate mutations slower than surrounding, non-functional, regions, sites are identified.

Transcription factors play one role in the regulation of gene expression. They can either activate or repress gene expression, and in many cases several transcription factors, both activators and repressors, are involved in determining the final regulatory output for a gene, i.e. how highly expressed it will be. Combinatorial relationships make it very difficult to test, or even computationally explore, all regulatory mechanisms in order to explain the expression patterns of a gene. In a study that attempted to dissect the combinatorial nature of gene expression regulation [39], the authors were able to explain gene expression inside transcriptional modules computationally based on the expression of the regulators. In this case, the dimensionality of the search space was reduced by limiting the number of studied genes to those that were expressed in a particular tissue—aspen leaves. Furthermore, the regulators considered for each module were determined in an iterative fashion, where a new regulator was added only if it increased the predictive power of the model.

Most genes are transcribed at some point in the lifetime of an organism, but this could possibly be at a single time point in a specific tissue. Thus, it is easy to see that there must be a very complex regulatory system orchestrating transcription. Transcription factors have to be expressed, and they in turn regulate the expression of some other gene(s) that in turn might act in a feedback loop to regulate the expression of itself. This quickly scales to form a complex network that is not easy to disentangle. To further complicate things, even if all factors needed to transcribe a particular gene are available, the gene might still not be expressed if the 3D structure of the DNA is not arranged in a configuration that allows access to the transcription factor binding sites, for example. In order for e.g. enhancers to act properly, they need to be physically close to the gene it acts upon (figure 3). The 3D structure of DNA is part of what is known as epigenetics; the heritable changes in gene expression that are not caused by changes in DNA sequence [40]. Another type of epigenetic modification that influences gene expression is the methylation of promoter regions which can block transcription factors from binding [41,42].

Due to this combined complexity, most studies of regulatory networks have so far been limited to smaller sets of genes [43].

1.3.2 Quantifying gene expression

Similar to genome sequencing, the estimation of gene expression received a big boost from the development of high-throughput sequencing. In the case of gene expression, instead of extracting and sequencing the DNA, the mRNA is extracted, reverse transcribed into complementary DNA (cDNA) and this is then sequenced—a process referred to as RNA-Sequencing. This effectively creates a snapshot of the abundance of all transcribed RNAs—the transcriptome—in a tissue of interest at the time of extraction. The last part of this sentence is something that is very important to consider, and we will come back to this in section 1.6. With the previously very popular microarray technology, relative quantification of transcript abundance was also possible, but limited to the genes that were included on the array, among other limitations. With RNA-Sequencing, *all* mRNA in the cell can, theoretically, be sequenced, regardless of whether the gene expressing it has previously been identified and annotated, or not.

The data from RNA-Sequencing is similar to that from DNA-Sequencing in that it consists of sequence reads based on a set of template sequences. In addition to being able to measure the expression of all genes and not only known genes, the dynamic range of RNA-Sequencing is significantly wider compared to microarrays since the signal does not get saturated, and the noise levels are lower [44]. In order to quantify gene expression, reads are aligned to either a reference transcriptome (all known RNAs in the organisms) or a reference genome (all DNA). Both approaches have their advantages and disadvantages, but perhaps the most obvious disadvantage of using a reference transcriptome is that only known gene products will be detected. Aligning to a reference genome means that no prior information about known genes is used; if reads map to an unannotated region of the genome, then something is expressed in that region. One problem of aligning reads to the genome is that splice junctions have to be handled. A splice junction is the border between an

exon and an intron, and these are not present in the sequenced mature mRNA (figure 3).

A problem that exists for both alignment methods is multi-mapping reads—reads that map to multiple locations in the genome/transcriptome. Duplicated genes, for example, might result in multi-mapping reads. Even if the genes have diverged when it comes to their regulation, their coding sequences can be more or less identical, and given a read produced from either of the duplicated copies one cannot confidently say from which gene the read originated.

Another problem slightly related to multi-mapping reads is that of alternative splicing variants. Splice variants are mRNAs that are produced by the same gene, but they have different composition of exons. Some variants might be missing an exon that other transcripts have, for example, and these transcripts are even harder to separate than duplicated genes. One possible way around the problem is to look at the expression of exons, and not genes or transcripts as a whole [45].

A more recent approach to read alignment is a collection of methods referred to as "alignment free" that are utilised in software such as sailfish [46] and kallisto [47]. The principle of these methods is to not care about the exact location of every read, instead to focus on which transcript the read is compatible with. These types of methods are quite new and have yet to be thoroughly tested, but they are in any case very interesting simply due to their speed; kallisto is 150–350 times faster than software that traditionally has been used for quantifying gene expression in RNA-Seq data [47].

We have developed our own pipeline for read processing and mapping, that was utilised throughout this thesis [48].

1.3.3 Co-expression networks

When two genes have similar expression profiles they are said to be co-expressed. By a gene's expression profile we mean its quantified expression across a number of tissues, time points, conditions, or treatments. Co-expression can be seen as a manifestation of the underlying regulatory network—if two genes are regulated

by the same factors, it is expected that these genes also are co-expressed. In contrast to the regulatory network, the co-expression network is simple to construct, with the simplest approach being to calculate the correlation between all pairs of genes. A co-expression network can be represented as a graph structure where the vertices are genes and the edges represent the degree of co-expression (figure 4). The consensus from a large number of studies is that co-expression networks (and biological networks in general) often are scale-free. What this means is that there are few genes with many connections to other genes (high degree centrality) in the network and more genes with few connections (low degree centrality) [49]. If the network is disturbed by random perturbations, such as mutations, genes with a high degree centrality are less likely to be targeted, due to their low frequency [50]. Consequently, a scale-free network will be robust against random perturbations. Another measure of centrality is betweenness centrality, which is a measure of how often a node is part of the shortest paths between all pairs of nodes in the network (figure 4). Co-expression networks have also been found to be modular, that is, there are sub-networks in the global network that are more tightly connected to the inside of the module than to the outside. More often than not it is the case that these modules are enriched in functional categories such as Gene Ontology [51] terms or Kyoto Encyclopedia of Genes and Genomes [52] pathways [53,54].

Studies have shown that co-expression networks can be useful vehicles in capturing and describing biologically relevant gene expression signatures. One example is a study performed in lake whitefish (*Coregonus clupeaformis*) where researchers found network modules that were correlated with dwarfism [55]. Another study identified gene expression signatures common across cancer tumour types using a co-expression network approach [56].

As previously stated, the co-expression network is a manifestation of the underlying regulatory network, but it is also important to remember that the co-expression network is only a very simple representation of the correlation in gene expression levels. It only captures the state of mRNA in the cell at the time of sampling, and it is not necessarily a representation of the corresponding protein abundance, or of protein–protein interactions. Two proteins that *can* interact *will* not necessarily interact just because their corresponding

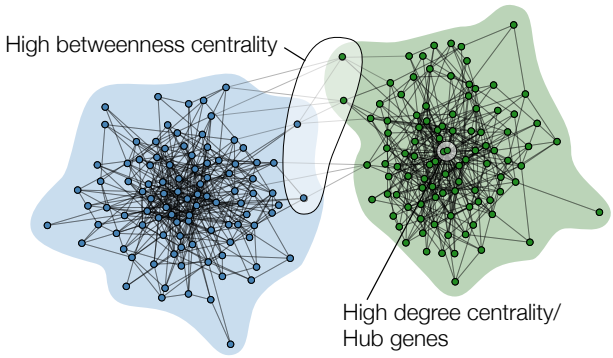

Figure 4: If the expression of two genes is correlated they are said to be co-expressed. This can be represented as a graph, or network, structure where each node represents a gene, and the edges between nodes represent significant co-expression. Modules in the network are defined as sub-networks that have a stronger connections to genes inside the module compared to genes outside the module. The modules are here represented by background colour. Nodes are said to have a high degree centrality if they have many connections to other genes, and these nodes can also be referred to as hub nodes. Nodes with a high betweenness centrality are genes that act as a connection between many other pairs of nodes in the network. These are typcially nodes that connect modules with each other.

transcripts are expressed at the same time [57]. However, if two genes are expressed simultaneously in a sufficiently high number of different conditions, it is likely that they share at least some of their regulatory mechanisms. One study exploring this in *Arabidopsis* used network cliques—sub-networks where all nodes are connected to all other nodes—to identify potential transcription factor binding sites [58]. They found that regulatory motifs identified in cliques that contained many genes targeted by the transcription factor E2Fa in many cases corresponded to the previously verified binding site of that transcription factor.

1.4 Association studies

Association studies, in this context, refers to the association of genetic variants with a phenotype of interest. This phenotype can range from very clear ones, such as human height, to more abstract phenotypes, such as gene expression. In the following sub-sections I will present the concept of association studies

and what we have learned from those so far.

1.4.1 Genome-wide association

I dare to bet that most people have come into contact with genome wide association studies (GWAS) at one time or another. Whenever you see headlines in the news such as "the obesity gene has been found", it is likely that the underlying study is a GWA study. It is also likely that the sensational headline is not quite true. What researchers have done in cases like this is to collect populations of individuals; those that have the phenotype of interest, such as a disease, and another population of healthy people. The genomes of these individuals are then sequenced or otherwise assayed for genetic variants and the researchers then ask themselves: can we identify variants that can be used to predict if an individual will be healthy or diseased?

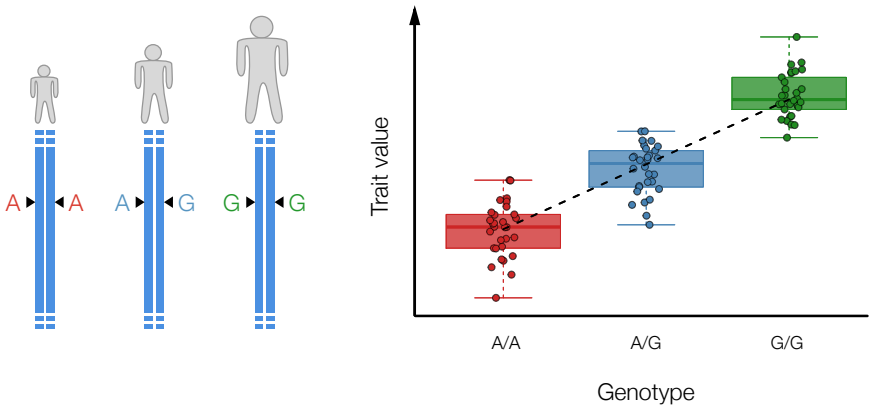

Figure 5: A schematic example of an association study with three different individuals with three different genotypes at a particular locus (left): A/A, A/G, and G/G. These genotypes explain the height of these individuals where the G allele is associated with higher individuals. Associating this locus with the height of individuals in a population might yield the plot to the right. The dashed line is then fitted to the data to minimise the distances between all data points and this line. If the slope of this line is significantly different from zero we say that the association is significant. The effect size is the slope of the fitted line and the variance explained by the SNP is related to the amount of variation of the data points around the line. The closer the points are to the line, the more of the phenotypic variation is explained by the SNP. Finally, the significance of the association is the probability of the slope of the line being different from zero.

Prediction is usually performed by way of relatively simple linear regression models where the disease status or the quantitative phenotype acts as the response variable and the genetic variant as the explanatory variable. A straight line is then fitted to the data and represents the predicted trait value for a given genotype (figure 5). These types of linear models assume that there is an additive effect, i.e. the contribution of an allele adds up to explain the phenotype. In figure 5, this is shown by the G allele adding to the trait value in a way that two copies of the G allele has twice the effect compared to having one copy of the G allele, in relation to having no G allele. It is not hard to imagine that GWA studies often require a huge number of tests. In humans, for example, we expect to find one SNP every 1.9 kilobases in the genome [59], and this would result in more than 1.5 million tests if one were to test the association to the phenotype of interest for each and every genetic variant. This has a few consequences, but mainly it requires computational power that has only become widely available quite recently. More importantly, the multiple testing burden of GWA studies often becomes quite heavy.

Multiple hypothesis testing is a statistical problem stemming from the fact that we expect to see random associations when performing a large number of tests. In the case of GWA in humans, suppose that we are associating 1.5 million variants with a particular phenotype. Applying the commonly used p-value threshold of 0.05 means that we would expect 75,000 false positive associations, i.e. associations that are due to purely random chance as a direct result of the number of tests. To control for this the p-values can be adjusted according to the number of tests performed using approaches such as the Bonferroni correction, where the obtained p-values are multiplied by the number of tests. Depending on the context, other less stringent methods are often preferred, like methods that control the false discovery rate (FDR) [60]. In studies involving genomic and gene expression data, there is extensive correlation structure in the data. Using relatively simple approaches such as Bonferroni or FDR correction will not take the correlation structure into account and can thus be overly conservative, and this can be overcome by using permutation tests.

In order to detect associations with a very low effect, i.e. a slope close to zero (see figure 5), a large number of samples are needed. With fewer samples,

only the most obvious associations will be detected, i.e. those with a high effect. For example, in the study of human height mentioned previously, more than 250,000 individuals were included in the study [3], and a meta-analysis of almost the same magnitude was performed to find a genetic explanation to body mass index [2]. In both cases the phenotypic variance explained by individual variants was very low (below 1%). These small effects would not be detectable in a smaller study.

1.4.2 eQTL mapping

Expression quantitative trait locus (eQTL) mapping is related to the traditional GWAS, as just described, but the phenotypes here are of the more abstract kind, namely gene expression. The problems of GWAS get even bigger for this type of association study since not only do we have a large number of genetic variants, we also have a large number of phenotypes. The phenotypes in this case are measures of gene expression for every transcribed gene. If we are to consider the expression of every gene in the human genome together with all the genetic variants in the genome, we have to perform approximately 30 billion tests. Not only does this results in a multiple testing problem, but it also causes purely computational problems. Not too long ago, this many tests would have been practically impossible to perform due to the computational resources needed, but with the increase in computational power, coupled with clever methods [61], this is now relatively easy to do.

QTL mapping is typically divided into two categories: linkage mapping and association mapping. Linkage mapping is usually used when family information is available, such as in a controlled cross. It relies on known markers and operates by performing a cross and observing how genetic markers associate with changes in the trait of interest. For the work described in this thesis we have instead used natural populations of plants, for which we do not have family information and where a naturally breeding collection of individuals are considered, rather than a controlled cross between two individuals. This approach is referred to as association mapping, or linkage disequilibrium mapping. This method is related to GWA in that a large number of genetic

markers (typically SNPs) are statistically tested to determine whether they are significantly associated with variation in a phenotype; the phenotype in this case being gene expression levels. Linkage disequilibrium (LD) is the non-random association between different loci. The idea is that the SNPs used for the association are in LD with the factor that is actually responsible for the phenotype. This way, the causal variant itself does not necessarily have to be included in the association, as long as a variant that is in LD with it is included.

eQTLs can be classified as either local or distant. A local eQTL is close to the gene that it is associated with while a distant one is far away, either on the same chromosome or on a different chromosome than the associated gene. The distance threshold where local becomes distant is however somewhat arbitrary. In our eQTL analysis in paper III, we classify SNPs within 100 kilobases from the transcription start site to be local, based on the distance distribution of eQTL on the same chromosome as the associated gene. The division into local and distant is a purely structural one as opposed to a functional definition. A more functional definition also exists, where eQTLs are classified depending on *how* they act on the associated gene. eQTLs are said to act either in *cis* or in *trans*, with *cis*-eQTLs acting directly on gene expression while *trans*-eQTL act indirectly on the associated gene. An example of a *cis* mechanism could be a variant that modifies a transcription factor binding motif in the promoter of a gene, while a *trans* effect could be something so subtle as affecting the abundance of a certain co-factor that is required for expression of the associated gene. Consequently, a *cis*-eQTL should act in an allele specific manner. If a transcription factor binding site gets disrupted in only one allele, only the transcription of that allele will be affected. Conversely, *trans*-eQTLs will have the same effect on both alleles. Due to the indirect mechanism of *trans*-eQTL, these are generally of lower effect (remember the slope from figure 5), and this is something that has been reported by numerous studies (references in [62,63]), although there are exceptions [64]. Normally *cis* acting variants are local to the associated gene while *trans* effects are more distant. Some studies opt to only consider local eQTL, like [65], and this is to some extent a tactical decision in that it makes the computational problem a bit easier since fewer

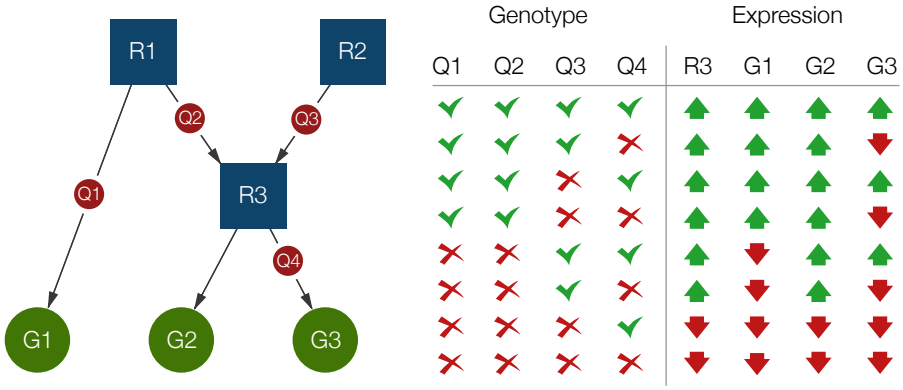

Figure 6: Simplified example of when eQTL effects and gene regulation is masked. A green checkmark means the regulatory link is enabled, while a red cross means it is disabled. Green arrows indicate up-regulation of the gene while a red arrow indicates down-regulation of the gene. In the regulatory network, the regulators R1 and R2 are always on, while regulator R3 is on as long as at least one of the eQTLs Q2 or Q3 enables the signal. The expression of G1 only depends on Q1, and this eQTL is thus easily detected by standard eQTL mapping methods since there is a perfect relationship between the genotype and the expression. Due to the dual regulators and eQTLs for R3, there is no perfect relationship between the eQTLs Q2 and Q3 and either R3 or G2. The regulation of G3 is even more complicated where R3 needs to be expressed, and at the same time Q4 must enable the signal. No perfect relationship between G3 and any of the eQTLs exist even though Q4 is *cis*-acting and Q2 and Q3 are both *trans*-acting.

tests have to be performed, and consequently, the multiple testing problem becomes slightly less of a problem since the number of markers considered for each gene is much smaller than the total number of markers.

The first study of the genetics underlying gene expression variation was performed in yeast in 2002 [66] and included 3,312 genetic markers and 6,215 genes. At the time this was a big feat, but today we are able to run association tests for all genes in the genome and all markers as demonstrated by the human Genotype Tissue Expression project (GTEx; [65]) with a total of about 6.8 million SNPs and using both coding and non-coding genes (53,934 genes in total).

1.4.2.1 Biology gets complicated quickly

Complex traits are the result of the interactions between many different factors. When it comes to eQTLs, the most common approach is to consider pairs of genes and genetic variants one by one. A better approach would be to analyse combinations of genetic variants and how they affect gene expression in concert. However, it is not possible to do this in an exhaustive manner due to computational complexity and multiple testing. In figure 6 a simple example of how the regulation could be hidden from traditional analysis methods is shown. The gene G1 is perfectly correlated with the genotype of the eQTL Q1, and thus the traditional approach is perfectly capable of detecting this relationship. It does not take much before this becomes too complicated though. R3 is dependent on two eQTLs, Q2 and Q3. The expression of R3 is not perfectly correlated with neither Q2 nor Q3, but in combination these eQTLs fully explain the expression of R3. In other words, a model that takes all pairs of SNPs into account would be needed to detect this relationship. Since G2 is directly regulated by R3, the dissection of G2 would need the same model as R3. Finally, G3 could only be dissected if all triplets of SNPs were taken into account. This is a very simplified example, but it highlights the inherent difficulties of systems genetics. In paper III we work with about 3.2 million SNPs and about 20,000 genes resulting in about 64 billion models. This would be able to capture the expression of G1. In order to dissect the expression of R3 and G2 we would need to create models using all pairs of SNPs against all genes and this would result in 1.02×10^{17} models. The expression of G3 is explained by three eQTLs, and in order to test all SNP triplets, we would have to investigate 1.09×10^{23} models. Assuming that we are able to calculate 10 million models per second—which is about the same speed as we achieved in paper III—computing all models for pairs of SNPs would take more than 300 *years*, and all models of SNP triplets would take more than 340 *million* years. Moreover, this is not even the worst part since the ridiculous number of tests would need a correspondingly strict correction for multiple testing. In order for any effect, no matter how large, to be significant, an enormous amount of sequenced and phenotyped individuals is needed. This can be viewed as the Catch 22 of genomics, where we have biological complexity on one side and

limited data availability and computational power on the other.

Machine learning is class of methods that can be used in order to identify patterns in large data sets. In paper II we use a support vector machine (SVM) approach to classify samples as male or female based on gene expression. Omics data have a dimensionality problem with a large number of variables (e.g. genes) compared to the number of observations. An SVM will very likely perform very well on this kind of data, but it will not generalise, i.e. new observations will not be classified with a very high accuracy. This can be alleviated somewhat by limiting the model using cross-validation, but instead the model will likely have a bad performance for all data instead. In order to use methods like this, the data must be limited to smaller data sets with a higher signal to noise ratio.

As seen in figure 6, the complexity of regulation often results in redundancy in the regulatory network, redundancy that can act as a buffer for random mutations [67]. Here gene duplications play a role as well since with two copies of the same gene, any detrimental mutations to one of them will most likely not affect the organism in a drastically negative way. Not only does this protect the organism, but it can also hide the regulatory mechanism from traditional analysis methods. One way to think of this is that simplicity would be bad for biology in general. If something is easy to disentangle, then a very small perturbation, like a mutation, could possibly disrupt the whole system. This is part of why we, in paper III, hypothesise that genes that are central in the co-expression network have evolved more redundancy in their regulation. By having more redundancy, these genes will not be affected as easily by random mutations, and this is the same idea underlying the hypothesis of scale-free biological networks (section 1.3.3).

1.4.3 Genetic variants in an evolutionary context

The genetic variants that are used for association mapping are not static in evolutionary time and their current state in a population reflects the evolutionary history and outcome within extant individuals of the popoulation being studied. Mutations (i.e. markers) could slowly be removed from the population

if they have a detrimental effect on reproductive fitness (negative or purifying selection), they could become fixed if they have a beneficial effect (positive selection), there could be selection to maintain a mix of alleles (balancing selection), or they could be under no selection pressure (selectively neutral) and just drift through the population (genetic drift). These concepts can be used to put the genetic variation into an evolutionary context that can help understand the process that have acted on mutations since their point of origin in the history of a species. In the context of eQTLs, a recent study showed signs of eQTLs being under negative selection and that the effect size of the eQTLs were negatively correlated with their frequency [68], a finding in common to those reported here in paper III.

1.5 Integration of different types of data

From the sections on gene expression and association studies above, we see that it is possible to explain some of the variability in complex traits using omics resources by themselves. The natural follow-up question is to ask whether we can gain even more from combining omics data. Between the genome and the phenotype of interest there are many regulatory steps: genes will be expressed (or not expressed), proteins might be degraded prematurely (or accumulate), and all these effects act on each other in a complex network (figure 7). A single analysis method, e.g. GWA, will simply not be able to capture the whole truth. It will generate a genomic variant that is associated with your trait of interest, but everything in between will essentially be a black box. By integrating different types of data, the black box can be illuminated. Furthermore, combining data can constrain our search space an thus alleviate some of the problems with computational power and multiple testing discussed previously. The approach of combining different levels of omics data is known as systems genetics [69].

Most GWAS variants found so far are located in non-coding regions of the genome, and it is thus hard to assign function to these variants. One approach to annotate these non-coding variants could be to combine GWA with eQTL mapping. This way genes can be associated with gene expression if genetic

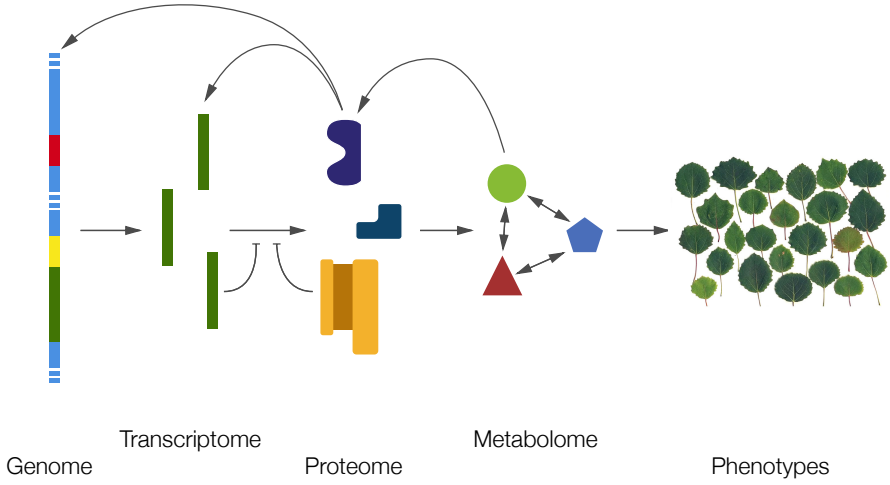

Figure 7: The different types of regulatory layers and how they can interact in order to give rise to complex traits. Genetic information is transferred to downstream layers through transcription into RNA. This in turn is translated into protein. Proteins then act together in order to produce and modify metabolites, as well as interacting with RNA and the DNA to regulate transcription. All this, together with environmental factors, give rise to phenotypes; some more complex than others.

variants are shared between the two studies, and phenotypes can consequently be associated with genes through guilt by association [70–72]. So far, most studies have focused only on protein coding genes leaving non-coding GWA variants without functional annotation. With RNA-Sequencing as the dominating technology for estimating gene expression together with the encouragement from the community to make data publicly available, it will be possible to revisit these studies as the annotations of the regulatory genome improve.

To gain even more understanding as to how complex traits emerge, information from even more regulatory layers must be included. This can be done with two main approaches: multi-staged analysis or a meta-dimensional analysis [73]. The multi-staged analysis is based on using data in a hierarchical manner, e.g. identify SNPs that are significantly associated with the phenotype of interest, and associate that subset of SNPs with gene expression levels, i.e. eQTL mapping. In this way, the number of SNPs to consider is significantly decreased

compared to a genome-wide eQTL mapping approach. The expression of the genes associated with genetic variants can then be used to investigate protein expression and perhaps to examine a subset of a protein interaction network in order to obtain a more complete picture of the emergence of phenotypes. In a meta-dimensional analysis, data from different layers are combined into a simultaneous analysis in order to consider multiple relationships at the same time, as opposed to the multi-staged analysis approach. For an in-depth review of data integration see [73].

In section 1.4.2.1, we mention that it is not feasible to do an exhaustive search for combinatorial effects. This problem could get even worse when including more data, but at the same time, more data can also help mitigate these issues. By layering the different types of data, things that most likely do not have an effect on the phenotype of interest can be excluded. By systematically integrating different types of data, the search space can be limited drastically, and ultimately it could actually be feasible to perform an exhaustive search of this reduced space.

The dimensionality of individual data sets can also be reduced in many cases. For example, if there are SNPs in the genome that are inherited together (i.e. in LD), including all SNPs in such a region will not provide any additional information, but rather take away statistical power due to an increased multiple testing burden. If only one SNP is used as a representative for a linkage block, it is important to remember that this might very well not be the causal SNP. Again, integration of other data types such as transcription factor binding information might be used to select the likely causal SNP.

1.6 Limitations

Although the strategies for elucidating the emergence of complex traits discussed in this thesis are promising, there are practical limitations that cannot be circumvented, at least not easily. In cases where natural populations are studied, we are dependent on capturing the right moment when the biochemical process of interest is active. If we do not sample at the critical time point, we risk

missing the processes that actually define the trait of interest. Not only is the sampling time point critical for capturing e.g. a developmental stage, it is also important to consider daily fluctuations that can have an effect on gene expression. This is clearly demonstrated in a recent study in rice where gene expression was quantified in the field, and it was shown that even short term variations in temperature and sunlight levels modified gene expression in a reproducible manner [74].

When it comes to geographical range, we are limited to the range of sampling, and cannot do inferences beyond that. For example, in papers III and IV, we were limited to the range of Sweden, while aspens are spread out across more or less the whole northern hemisphere. In that context, Sweden is a very small part of the total distribution range, and it might not be very surprising that we did not see any pronounced population structure in the data.

In order to have enough statistical power in association studies, large sample sizes are needed. A problem with using forest trees is that it is very expensive and time consuming to maintain a large population of trees. Ideally, we want to grow them in a controlled environment in order to minimise environmental effects, but this is clearly not a practical approach. In the study of human height that has been mentioned several times in this thesis, they used more than 250,000 individuals, but this was a meta-study [3], i.e. a study collecting data from previous studies, avoiding the hassle of collecting the data themselves. Even so, a meta-study of the same magnitude in *P. tremula* is not possible today, since the amount of data generated is not even close to that of human studies. The importance of good annotations such as previous efforts to elucidate regulatory mechanisms also play a big role [36]. As of yet, most of these efforts have been directed towards human studies (e.g. ENCODE [14]), quite understandably. While the results from these annotation efforts are not directly transferable to other species, information regarding general characteristics of genome structure and function can most likely be transferred to other species. This has to some extend been done already, but in the other direction. For example, the fruit fly *Drosophila melanogaster* has been used as a model organism for human genetics and disease for over a century [75].

The results of associations studies are just that—associations. A genetic variant that is associated with a particular phenotype is not necessarily the causative variant. It might be that it in turn is associated with the causal variant through linkage disequilibrium (LD). In the case of plants, a variant in LD with the causal variant might be good enough in many cases where marker assisted selection can be employed in breeding. However, if more control of the phenotype is needed, a variant in LD is not of much help. If the variant is not causal, it is likely that mutating this position will not result in a corresponding change in phenotype. Strategies to filter out the causal variants from association studies include integrating different types of data in order to single out the most likely candidate genes or loci, but there are several challenges associated with this kind of data integration. The individual data sets themselves have their own issues to begin with. There are systematic biases, normalisation issues, and correlation structures that are not trivial to deal with, and that can eat up a considerable portion of resources available to a project. Something a bit more abstract that could help with finding causal variants is transparency when it comes to publication. This could potentially help minimise confirmation bias and consequently the number of false positives in circulation [76].

2

Paper summaries

This chapter will give a short summary of each paper included in the thesis. Paper I deals with gene regulation in a cyanobacterium, while papers II–IV considers aspects of gene expression, genotype, and phenotype in the deciduous tree *Populus tremula* (European aspen).

2.1 Paper I — Gene regulation in a cyanobacterium

Synechocystis is a fresh water cyanobacterium and it is one of the most studied cyanobacteria to date, being used as a model system for nitrogen fixation and photosynthesis amongst other things. Even though the genome of *Synechocystis* was sequenced already in 1996, most of the genes in its genome are still annotated as having unknown function. In this paper, we created the web application *Syn*ergy to enable researchers working with *Synechocystis* to explore the gene expression and the gene regulation of this organism *in silico* in order to find potential candidate genes for e.g. knock-out experiments. We collected 371

microarray experiments from public sources and constructed a co-expression network. As mentioned in section 1.3.3, a co-expression network is simply a manifestation of the underlying regulatory network, so in order to form a link between co-expression and co-regulation, potential regulatory motifs were identified using phylogenetic footprinting. This method is based on the alignment of regulatory regions of orthologous genes from related organisms. In this case, 22 genomes from the Chroococcales taxon were used for the phylogenetic footprinting, and this resulted in a set of 4,977 potential regulatory motifs. In the paper we show that co-expression network neighbourhoods of regulatory proteins were enriched for regulatory motifs, thus providing a possible regulatory link between these regulators and the co-expressed genes. The user of the web application can then investigate whether their gene set of interest is co-expressed, and whether this to some extent can be explained by shared regulatory motifs. In order to make the application as useful as possible, the gene identifiers used were the well established identifiers from CyanoBase (http://genome.microbedb.jp/cyanobase/; [77]).

As part of a sanity check of the integrated data, a couple of case studies were conducted where both previously published results were confirmed, and also potentially novel regulatory relationships were presented.

*Syn*ergy is publicly available at http://synergy.plantgenie.org.

2.2 Paper II — Two-class phenotype prediction

The majority of angiosperms are monoecious or hermaphroditic, i.e. each individual has both male and female flowers, or the flowers have both male and female organs, respectively. This is not the case for about 4% of flowering plants, including the genus *Populus*, which is dioecious (with a few exceptions). It is thought that the formation of a dioecious species evolved from hermaphrodites and thus have not yet evolved distinct sex chromosomes, as is the case in mammals. When the sexes are separated, constraints on the phenotype are released as the individual sexes adapt to a new fitness optimum, and this would then give rise to sexual dimorphism, i.e. phenotypic differences between the

sexes. Previous studies in *Populus* species have shown that there is likely a sex determining region on chromosome 19, but so far no study has looked at global phenotype and gene expression patterns to look for sexual dimorphism.

In paper II we show that there are no significant phenotypic differences in a range of phenotypes in *Populus tremula*. The phenotypes included different biomass traits such as height and diameter, and also a range of secondary metabolites involved in herbivore defence. In addition to the more classical phenotypes, differences in gene expression were investigated by identifying differentially expressed genes. In addition, we also attempted to classify samples as male or female with a support vector machine (SVM) trained on genes inside sliding windows across the genome. The rationale behind the sliding window approach comes from the fact that there is a consensus that there is a sex determining locus. This locus might contain more than one gene, and individual genes might not be able to fully explain the sex division. However, in this case the SVM analysis did not result in any gene combinations that could predict sex any better than single genes. Only two individual genes were found to be significantly differentially expressed between the sexes, and one of those was located in a region previously linked to sex determination. As reported by Pakull et al. [78], and independently discovered by us, part of this gene is deleted in females and thus gives rise to the difference in expression between sexes.

2.3 Paper III — Genetic basis of gene expression variation

Natural variation is perhaps the most important aspect of biology; without variation there would be no evolution. Some of the observed variability can be explained by environmental factors while some can be explained by genetic factors. In paper III we take a closer look at the natural variation in gene expression in *Populus tremula* by performing eQTL mapping and constructing a co-expression network from gene expression data from a natural population of *P. tremula* spanning the distribution range of this species in Sweden. In

total, RNA-Sequencing and DNA-Sequencing data from 81 distinct genotypes were used for the study.

One of the main goals of this study was to see whether eQTLs could explain the structure of the co-expression network. Since the data originates from natural populations of unrelated individuals, the pairwise gene expression correlations were low, but we were still able to identify distinct co-expression modules. Genes whose expression was associated with genetic variants (eGenes) were less central in the co-expression network than what would be expected by chance, and there was also a negative relationship between the centrality of eGenes and the eQTL effect size. A general hypothesis when it comes to biological networks is that central genes are critical for the organism to function correctly, and that disruptions to these genes could have a negative impact on fitness. We hypothesise that these central genes have more regulatory redundancy than genes that are peripheral in the network. The regulation is governed by many small-effect eQTLs that in concert offers genetic buffering of the regulation of these genes. Due to a relatively small population size, we are not able to detect these small effect size eQTLs, and a given, but practically difficult, follow-up would be to collect more data in order to test this hypothesis.

2.4 Paper IV — Leaf shape and systems genetics

Leaves in plants are the main organs for photosynthesis and carbon fixation, and the morphology of leaves affects photosynthetic efficiency. Furthermore, leaves are often one of the most recognisable traits of a plant. In paper IV we applied a systems genetics approach where data on genotype, gene expression, and phenotype were integrated in order to understand the control of natural variation of leaf shape in *Populus tremula*. In contrast to paper II and paper III where a one-gene-at-a-time approach was followed, paper IV focused on explaining the complexity of leaf traits as they emerge from the interaction of many genes.

Three different leaf traits were considered: circularity, indent width, and leaf area. Only a handful of SNPs were significant in GWA for indent width and leaf

area, while none were significant for circularity. Two of the traits, circularity and indent width, were highly heritable. Of the SNPs with the highest significance, most were located in untranslated regions of genes, indicating that they might be exerting their effects through gene expression. However, very few of the SNPs were also eQTL SNPs in paper III. Furthermore, correlating gene expression values with the leaf traits did not result in any significant correlations, indicating that neither single-SNP nor single-gene approaches for dissecting leaf shape are viable.

We took these results as support for the infinitesimal model, i.e. that these traits are controlled by numerous variants of small effect size. Consequently, we employed a gene set enrichment approach where sets of genes associated with the top GWAS results were tested, as well as gene sets based on gene ontology terms. Here, several gene sets with a common functional role had a significant association to each of the three traits, emphasising the need to go beyond single-SNP and single-gene approaches in order to understand complex traits.

3

Discussion

In order to decipher the emergence of complex traits, the most common approaches that are used today, such as GWAS and eQTL mapping, are not enough by themselves. To be able to find the factors that contribute to complex traits, all layers of regulation must be taken into account. However, this is not a trivial task. Firstly, the limitations for the individual data types must be accounted for. Secondly, the data must be integrated in a way that maximises the information we are able to get out, while at the same time minimising the number of false associations.

This thesis has explored some of the strategies that can be used in order to unravel the emergence of complex traits, utilising gene expression data, genotype data, and phenotype data. In doing this we have seen the potential in integrating data from different sources in order to get a more complete picture of gene regulation and the emergence of phenotypes, but we have also seen that there is a long way ahead of us. In no way have we exhausted the possibilities with the data we have worked with.

3.1 Future perspectives

The future holds much in store when it comes to the analysis of complex traits. With sequencing costs already being low, they will probably get even lower. Furthermore, new technologies, such as nanopore sequencing [79] that enables sequencing of longer reads, will allow even more accurate quantification of gene expression and identification of genetic variation. With the short-read technologies that dominate the market today, there are a lot of ambiguities when it comes to e.g. the expression of splice variants and allele-specific expression; problems that have yet to be solved. Disruption of splicing has been associated with several human diseases [80], and may play an important role in environmental adaptation in plants [81]. If the whole mRNA molecule can be sequenced in one go, the expression of each and every splice variant could be determined with much better accuracy than any of the techniques employed today. However, one should not underestimate the computational challenges that usually follow with new technologies. It might be easy in theory, but just as for short-read sequencing, there will surely be some hurdles to pass on the way.

When it comes to sequencing, a high quality reference genome is a vital component in order to map genetic variation or quantify gene expression. Up until today, reference genomes are simply a long string of characters effectively representing a single haplotype in a single individual. Projects such as the 1000 genomes project [82] make it possible to deviate from this path and construct reference genomes that not only represent a consensus genome sequence, but that also represent the variation present in populations of individuals. There have been several studies to date reporting reference allele bias in RNA-Sequencing data, i.e. reads originating from the reference allele will map more confidently to the reference genome compared to a read originating from an allele that contains polymorphisms relative to the reference [83–85]. This could be alleviated by having a reference genome format that represents known variation, together with compatible software. The latest release of the human reference genome (GRCh38) is a step in this direction with alternative loci available for selected parts of the genome that are too complex

to be represented by a single sequence. In order to include all known genetic variation from e.g. the 1000 Genomes Project a number of associated challenges must be overcome, and this is something that likely will move forward slowly.

The association studies that are used today have one very obvious limitation: these are simply statistical association between a genetic variant and a trait of interest. Due to the lack of independence among genetic variance stemming from linkage disequilibrium, the variant that is associated to the trait of interest might not be the causative variant. Testing this could be accomplished using the relatively new and much hyped CRISPR/Cas9 technology [86]. Briefly, this can be seen as molecular scissors and glue that can cut and paste in the genome in order to insert, change, or delete parts of the DNA. With this, it would be possible to test the phenotypic effect of variants on a large scale in order to find the causative variants in genome wide association studies. Old cloning techniques are able to do the same thing, but they are very laborious, and in cases when traits are polygenic, it is often not feasible to generate anything more complex than a double mutant. With CRISPR/Cas9 it is possible to test several variants at the same time using a multiplex strategy, i.e. targeting multiple loci in a single experiment [87], which would be a direct requirement in order to verify multiple causal variants underlying complex traits. Furthermore, it is possible to perform allele specific modifications, where a point mutation is introduced in one allele while leaving the rest of the genetic background the same [88]. This technology is still very young, but owing to the great impact it has had on the scientific community, a wide range of publicly available tools have been developed in order to aid the community in designing CRISPR/Cas9 experiments [89–94], making this a far more accessible alternative compared to similar technologies of a more proprietary nature [95].

One aspect that has not been taken into account at all in this thesis is the effect of epigenetics. In section 1.3.1, it is mentioned that there are a number of factors that are required in order for RNA to be transcribed from DNA. One additional factor is epigenetics, i.e. modifications “on top of” the DNA that does not change the actual DNA sequence but still affect regulation. One example of an epigenetic modification is methylation which is the addition of a methyl group to the DNA backbone. If this methylation occurs in a transcription

factor binding site, it can block the binding of the transcription factor and consequently repress the expression of the gene [96]. Since most of the significant variants that are identified in GWAS are located in intergenic regions, not much information is provided initially by the GWA alone. Integrating these kinds of results with epigenetic data can contribute to a better understanding of regulatory mechanisms that connect genomic variation and higher order phenotypes [97].

In the end, these methods and technologies should be used in order to improve the situation for people and the environment. With climate change being a very real and imminent threat to the future of our species, we will need to develop improved crop varieties that are able to grow in environments that would normally be too harsh for the crop varieties of today. Most stress related traits are complex, and breeding strategies used are basically just trial-and-error in order to randomly identify something that will be ever so slightly more tolerant to e.g. drought. If researchers instead are able to, with these new technologies, dissect the genetic background of these complex traits, new crop varieties could be generated much more rapidly with e.g. marker assisted selection or by simply modifying the genome using something like CRISPR/Cas9. The ultimate goal would be to some day be able to pinpoint causal variants and be able to say that "if we change this from a C to a T we will get 5% higher yield in arid conditions". This scenario is probably quite far away at the moment, and given the complexity of biology, it is not even certain that we will get there—but at least we will not be bored.

References

1. Barsh GS. What controls variation in human skin color? PLoS biology. Public Library of Science; 2003;1: E27. doi:10.1371/journal.pbio.0000027

2. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010;42: 937–948. doi:ng.686 [pii]\r10.1038/ng.686

3. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nature Genetics. 2014;46: 1173–1186. doi:10.1038/ng.3097

4. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era — concepts and misconceptions. Nature Reviews Genetics. 2008;9: 255–266. doi:10.1038/nrg2322

5. Falconer DS, Mackay TFC. Introduction to Quantitative Genetics. 4th ed. Pearson; 1996.

6. Bouchard T, Lykken D, McGue M, Segal N, Tellegen A. Sources of human psychological differences: the Minnesota Study of Twins Reared Apart. Science. American Association for the Advancement of Science; 1990;250: 223–228. doi:10.1126/science.2218526

7. Watson JD, Crick FHC. Molecular structure of nucleic acids [Internet]. 1953.

pp. 737–738. doi:10.1097/BLO.0b013e3181468780

8. Crick F. On protein synthesis. Symposia of the Society for Experimental Biology. 1958;12: 138–63. Available: http://www.ncbi.nlm.nih.gov/pubmed/13580867

9. Brimacombe R, Stiege W. Structure and function of ribosomal RNA. Biochemical Journal. 1985;229: 1–17. Available: http://www.ncbi.nlm.nih.gov/pubmed/8722015

10. Gurtan AM, Sharp PA. The role of miRNAs in regulating gene expression networks. Journal of Molecular Biology. Elsevier Ltd; 2013;425: 3582–3600. doi:10.1016/j.jmb.2013.03.007

11. Yoon J-H, Abdelmohsen K, Gorospe M. Posttranscriptional Gene Regulation by Long Noncoding RNA. Journal of Molecular Biology. Elsevier B.V. 2013;425: 3723–3730. doi:10.1016/j.jmb.2012.11.024

12. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology. 1961;3: 318–356. doi:10.1016/S0022-2836(61)80072-7

13. Istrail S, De-Leon SBT, Davidson EH. The regulatory genome and the computer. Developmental Biology. 2007;310: 187–195. doi:10.1016/j.ydbio.2007.08.009

14. Feingold E, Good P, Guyer M, Kamholz S, Liefer L, Wetterstrand K, et al. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306: 636–640. doi:10.1126/science.1105136

15. Stern DL. Perspective: Evolutionary Developmental Biology and the Problem of Variation. Evolution. 2000;54: 1079. doi:10.1554/0014-3820(2000)054[1079:PEDBAT]2.0.CO;2

16. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, et al. Forces shaping the fastest evolving regions in the human genome. PLoS Genetics. 2006;2: 1599–1611. doi:10.1371/journal.pgen.0020168

17. Polavarapu N, Arora G, Mittal VK, McDonald JF. Characterization and potential functional significance of human-chimpanzee large INDEL variation.

Mobile DNA. 2011;2: 13. doi:10.1186/1759-8753-2-13

18. Macintyre G, Bailey J, Haviv I, Kowalczyk A. Is-rSNP: A novel technique for in silico regulatory SNP detection. Bioinformatics. 2011;27: i524–i530. doi:10.1093/bioinformatics/btq378

19. Zuo C, Shin S, Keleş S. atSNP: transcription factor binding affinity testing for regulatory SNP detection. Bioinformatics. 2015;31: 3353–3355. doi:10.1093/bioinformatics/btv328

20. Makarov V, O'Grady T, Cai G, Lihm J, Buxbaum JD, Yoon S. Antools: A comprehensive and versatile annotation toolkit for genomic variants. Bioinformatics. 2012;28: 724–725. doi:10.1093/bioinformatics/bts032

21. Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, et al. A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature. 2005;437: 88–93. doi:10.1038/nature04000

22. Blekhman R, Oshlack A, Gilad Y. Segmental duplications contribute to gene expression differences between humans and chimpanzees. Genetics. 2009;182: 627–630. doi:10.1534/genetics.108.099960

23. Lynch M. The Evolutionary Fate and Consequences of Duplicate Genes. Science. 2000;290: 1151–1155. doi:10.1126/science.290.5494.1151

24. Vallejo-Marín M, Buggs RJA, Cooley AM, Puzey JR. Speciation by genome duplication: Repeated origins and genomic composition of the recently formed allopolyploid species Mimulus peregrinus. Evolution. 2015;69: 1487–1500. doi:10.1111/evo.12678

25. Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, et al. Widespread genome duplications throughout the history of flowering plants. Genome Research. 2006;16: 738–749. doi:10.1101/gr.4825606

26. Friedman WE. The meaning of Darwin's "abominable mystery". American Journal of Botany. 2009;96: 5–21. doi:10.3732/ajb.0800150

27. Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. The frequency of polyploid speciation in vascular plants. Proceedings

of the National Academy of Sciences of the United States of America. 2009;106: 13875–13879. doi:10.1073/pnas.0811575106

28. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, et al. Ancestral polyploidy in seed plants and angiosperms. Nature. 2011;473: 97–100. doi:10.1038/nature09916

29. Mühlhausen S, Kollmar M. Whole genome duplication events in plant evolution reconstructed and predicted using myosin motor proteins. BMC evolutionary biology. 2013;13: 202. doi:10.1186/1471-2148-13-202

30. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). Science. 2006;313: 1596–1604. doi:10.1126/science.1128691

31. Check Hayden E. Is the $1,000 genome for real? Nature. 2014; doi:10.1038/nature.2014.14530

32. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001;29: 308–11. doi:10.1093/nar/29.1.308

33. Abzhanov A, Kuo WP, Hartmann C, Grant BR, Grant PR, Tabin CJ. The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. Nature. 2006;442: 563–567. doi:10.1038/nature04843

34. Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A, et al. Evolution of Darwin's finches and their beaks revealed by genome sequencing. Nature. 2015;518: 371–375. doi:10.1038/nature14181

35. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics (Oxford, England). 2000;16: 16–23. doi:10.1093/bioinformatics/16.1.16

36. Mathelier A, Lefebvre C, Zhang AW, Arenillas DJ, Ding J, Wasserman WW, et al. Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. Genome biology. Genome Biology; 2015;16: 84. doi:10.1186/s13059-015-0648-7

37. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT. Embryonic ϵ and γ globin genes of a prosimian primate (Galago crassicau-

datus). Journal of Molecular Biology. 1988;203: 439–455. doi:10.1016/0022-2836(88)90011-3

38. Blanchette M, Tompa M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Research. 2002;12: 739–748. doi:10.1101/gr.6902

39. Street NR, Jansson S, Hvidsten TR. A systems biology model of the regulatory network in Populus leaves reveals interacting regulators and conserved regulation. BMC plant biology. 2011;11: 13. doi:10.1186/1471-2229-11-13

40. Richards EJ. Inherited epigenetic variation–revisiting soft inheritance. Nature reviews Genetics. 2006;7: 395–401. doi:10.1038/nrg1834

41. Siegfried Z, Eden S, Mendelsohn M, Feng X, Tsuberi BZ, Cedar H. DNA methylation represses transcription in vivo. Nature genetics. 1999;22: 203–206. doi:10.1038/9727

42. Medvedeva YA, Khamis AM, Kulakovskiy IV, Ba-Alawi W, Bhuyan MSI, Kawaji H, et al. Effects of cytosine methylation on transcription factor binding sites. BMC genomics. BMC Genomics; 2014;15: 119. doi:10.1186/1471-2164-15-119

43. Chai LE, Loh SK, Low ST, Mohamad MS, Deris S, Zakaria Z. A review on the computational approaches for gene regulatory network construction. Computers in Biology and Medicine. Elsevier; 2014;48: 55–65. doi:10.1016/j.compbiomed.2014.02.011

44. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics. 2009;10: 57–63. doi:10.1038/nrg2484

45. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome Research. 2012;22: 2008–2017. doi:10.1101/gr.133744.111

46. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature Biotechnology. Nature Publishing Group; 2014;32: 462–464.

doi:10.1038/nbt.2862

47. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016; doi:10.1038/nbt.3519

48. Delhomme N, Mähler N, Schiffthaler B, Sundell D, Mannepperuma C, Hvidsten TR, et al. Guidelines for RNA-Seq data analysis. Epigenesys. 2014; Available: http://www.epigenesys.eu/en/protocols/bio-informatics/1283-guidelines-for-rna-seq-data-analysis

49. Barabási A-L, Albert R. Emergence of Scaling in Random Networks. Science. 1999;286: 509–512. doi:10.1126/science.286.5439.509

50. Whitacre JM. Biological robustness: Paradigms, mechanisms, systems principles. Frontiers in Genetics. 2012;3: 1–15. doi:10.3389/fgene.2012.00067

51. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25: 25–29. doi:10.1038/75556

52. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000;28: 27–30. doi:10.1093/nar/28.1.27

53. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008;9: 559. doi:10.1186/1471-2105-9-559

54. Langfelder P, Luo R, Oldham MC, Horvath S. Is my network module preserved and reproducible? PLoS Computational Biology. 2011;7. doi:10.1371/journal.pcbi.1001057

55. Filteau M, Pavey S a, St-Cyr J, Bernatchez L. Gene coexpression networks reveal key drivers of phenotypic divergence in lake whitefish. Molecular biology and evolution. 2013;30: 1384–96. doi:10.1093/molbev/mst053

56. Doig TN, Hume D a, Theocharidis T, Goodlad JR, Gregory CD, Freeman TC. Coexpression analysis of large cancer datasets provides insight into the cellular phenotypes of the tumour microenvironment. BMC genomics. BMC Genomics; 2013;14: 469. doi:10.1186/1471-2164-14-469

57. Vogel C, Marcotte EM. Insights into the regulation of protein abundance

from proteomic and transcriptomic analyses. Nature Reviews Genetics. Nature Publishing Group; 2012;13: 227–232. doi:10.1038/nrg3185

58. Zheng X, Liu T, Yang Z, Wang J. Large cliques in Arabidopsis gene coexpression network and motif discovery. Journal of plant physiology. 2011;168: 611–618. doi:10.1016/j.jplph.2010.09.010

59. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001;409: 928–933. doi:10.1038/35057149

60. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995; Available: http://www.jstor.org/stable/10.2307/2346101

61. Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28: 1353–1358. doi:10.1093/bioinformatics/bts163

62. Breitling R, Li Y, Tesson BM, Fu J, Wu C, Wiltshire T, et al. Genetical Genomics: Spotlight on QTL Hotspots. PLoS Genetics. 2008;4: e1000232. doi:10.1371/journal.pgen.1000232

63. Kliebenstein D. Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. Annual review of plant biology. 2009;60: 93–114. doi:10.1146/annurev.arplant.043008.092114

64. Clement-Ziza M, Marsellach FX, Codlin S, Papadakis MA, Reinhardt S, Rodriguez-Lopez M, et al. Natural genetic variation impacts expression levels of coding, non-coding, and antisense transcripts in fission yeast. Molecular Systems Biology. 2014;10: 764–764. doi:10.15252/msb.20145123

65. Ardlie KG, Deluca DS, Segre AV, Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015;348: 648–660. doi:10.1126/science.1262110

66. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science. 2002;296: 752–5.

doi:10.1126/science.1069516

67. Rutherford SL. From genotype to phenotype: buffering mechanisms and the storage of genetic information. BioEssays. 2000;22: 1095–1105. doi:10.1002/1521-1878(200012)22:12<1095::AID-BIES7>3.0.CO;2-A

68. Josephs EB, Lee YW, Stinchcombe JR, Wright SI. Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. Proceedings of the National Academy of Sciences of the United States of America. 2015;112: 15390–15395. doi:10.1073/pnas.1503027112

69. Civelek M, Lusis AJ. Systems genetics approaches to understand complex traits. Nature reviews Genetics. Nature Publishing Group; 2013;15: 34–48. doi:10.1038/nrg3575

70. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 2010;466: 714–9. doi:10.1038/nature09266

71. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen P a C, Monlong J, Rivas M a, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501: 506–11. doi:10.1038/nature12531

72. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518: 197–206. doi:10.1038/nature14177

73. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nature Reviews Genetics. Nature Publishing Group; 2015;16: 85–97. doi:10.1038/nrg3868

74. Plessis A, Hafemeister C, Wilkins O, Gonzaga ZJ, Meyer RS, Pires I, et al. Multiple abiotic stimuli are integrated in the regulation of rice gene expression under field conditions. eLife. 2015;4: e08411. doi:10.7554/eLife.08411

75. Stephenson R, Metcalfe NH. Drosophila melanogaster: a fly through its history and current use. The journal of the Royal College of Physicians of Edinburgh. 2013;43: 70–5. doi:10.4997/JRCPE.2013.116

76. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J,

Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. Nature Publishing Group; 2014;508: 469–476. doi:10.1038/nature13127

77. Nakao M, Okamoto S, Kohara M, Fujishiro T, Fujisawa T, Sato S, et al. CyanoBase: the cyanobacteria genome database update 2010. Nucleic Acids Res. 2010;38: D379——81. doi:10.1093/nar/gkp915

78. Pakull B, Kersten B, Lüneburg J, Fladung M. A simple PCR-based marker to determine sex in aspen. Mendel R, editor. Plant Biology. 2015;17: 256–261. doi:10.1111/plb.12217

79. Schneider GF, Dekker C. DNA sequencing with nanopores. Nature Biotechnology. Nature Publishing Group; 2012;30: 326–328. doi:10.1038/nbt.2181

80. Tazi J, Bakkour N, Stamm S. Alternative splicing and disease. Biochimica et Biophysica Acta - Molecular Basis of Disease. Elsevier B.V. 2009;1792: 14–26. doi:10.1016/j.bbadis.2008.09.017

81. Staiger D, Brown JWS. Alternative Splicing at the Intersection of Biological Timing, Development, and Stress Responses. The Plant Cell. 2013;25: 3640–3656. doi:10.1105/tpc.113.113803

82. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526: 68–74. doi:10.1038/nature15393

83. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics. 2009;25: 3207–3212. doi:10.1093/bioinformatics/btp579

84. Stevenson KR, Coolon JD, Wittkopp PJ. Sources of bias in measures of allele-specific expression derived from RNA-sequence data aligned to a single reference genome. BMC genomics. 2013;14: 536. doi:10.1186/1471-2164-14-536

85. Panousis NI, Gutierrez-Arcelus M, Dermitzakis ET, Lappalainen T. Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies.

Genome Biology. 2014;15: 467. doi:10.1186/s13059-014-0467-2

86. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science. 2012;337: 816–821. doi:10.1126/science.1225829

87. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science. 2013;339: 819–823. doi:10.1126/science.1231143

88. Smith C, Abalde-Atristain L, He C, Brodsky BR, Braunstein EM, Chaudhari P, et al. Efficient and allele-specific genome editing of disease loci in human iPSCs. Molecular Therapy. 2014;23: 570–577. doi:10.1038/mt.2014.226

89. Heigwer F, Zhan T, Breinig M, Winter J, Brügemann D, Leible S, et al. CRISPR library designer (CLD): software for multispecies design of single guide RNA libraries. Genome Biology. Genome Biology; 2016;17: 55. doi:10.1186/s13059-016-0915-2

90. Chari R, Mali P, Moosburner M, Church GM. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. Nature methods. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. 2015;12: 823–826. doi:10.1038/nmeth.3473

91. Heigwer F, Kerr G, Boutros M. E-CRISP: fast CRISPR target site identification. Nature methods. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. 2014;11: 122–3. doi:10.1038/nmeth.2812

92. MacPherson CR, Scherf A. Flexible guide-RNA design for CRISPR applications using Protospacer Workbench. Nature biotechnology. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. 2015;33: 805–6. doi:10.1038/nbt.3291

93. Montague TG, Cruz JM, Gagnon JA, Church GM, Valen E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. Nucleic acids research. 2014;42: W401–7. doi:10.1093/nar/gku410

94. Wong N, Liu W, Wang X. WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. Genome Biology. BioMed Central;

2015;16: 218. doi:10.1186/s13059-015-0784-0

95. Bortesi L, Fischer R. The CRISPR/Cas9 system for plant genome editing and beyond. Biotechnology Advances. Elsevier B.V. 2015;33: 41–52. doi:10.1016/j.biotechadv.2014.12.006

96. Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Research. 2013;23: 555–567. doi:10.1101/gr.147942.112

97. Farh KK-h, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. Nature Publishing Group; 2015;518: 337–343. doi:10.1038/nature13835

Paper I

PLOS ONE

RESEARCH ARTICLE

*Syn*ergy: A Web Resource for Exploring Gene Regulation in *Synechocystis* sp. PCC6803

Niklas Mähler[1], Otilia Cheregi[2], Christiane Funk[2,3], Sergiu Netotea[2,3,4]*, Torgeir R. Hvidsten[1,3]

1. Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway, **2.** Department of Chemistry, Umeå University, Umeå, Sweden, **3.** Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, Umeå, Sweden, **4.** Computational Life Science Cluster, Umeå University, Umeå, Sweden

*sergiu.netotea@umu.se

OPEN ACCESS

Citation: Mähler N, Cheregi O, Funk C, Netotea S, Hvidsten TR (2014) *Syn*ergy: A Web Resource for Exploring Gene Regulation in *Synechocystis* sp. PCC6803. PLoS ONE 9(11): e113496. doi:10.1371/journal.pone.0113496

Editor: Leonardo Mariño-Ramírez, National Institutes of Health, United States of America

Received: July 25, 2014

Accepted: October 24, 2014

Published: November 24, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. Microarray data is available from KEGG (http://www.genome.jp/kegg/expression/). Full genomes available at CyanoBase (http://genome.microbedb.jp/cyanobase/). All source code for *Syn*ergy, including the database, is available at GitHub (http://github.com/maehler/Synergy).

Funding: SN and TRH were funded by the Swedish Research Council (VR, http://www.vr.se/) grant number 2011-5811, and The Swedish Governmental Agency for Innovation Systems (VINNOVA, http://www.vinnova.se/) in parts through the UPSC Berzelii Centre for Forest Biotechnology. CF and OC are grateful for funding from the Swedish Energy Agency and Umeå University (Solar Fuels, http://solarfuels.eu/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Despite being a highly studied model organism, most genes of the cyanobacterium *Synechocystis* sp. PCC 6803 encode proteins with completely unknown function. To facilitate studies of gene regulation in *Synechocystis*, we have developed *Syn*ergy (http://synergy.plantgenie.org), a web application integrating co-expression networks and regulatory motif analysis. Co-expression networks were inferred from publicly available microarray experiments, while regulatory motifs were identified using a phylogenetic footprinting approach. Automatically discovered motifs were shown to be enriched in the network neighborhoods of regulatory proteins much more often than in the neighborhoods of non-regulatory genes, showing that the data provide a sound starting point for studying gene regulation in *Synechocystis*. Concordantly, we provide several case studies demonstrating that *Syn*ergy can be used to find biologically relevant regulatory mechanisms in *Synechocystis*. *Syn*ergy can be used to interactively perform analyses such as gene/motif search, network visualization and motif/function enrichment. Considering the importance of *Synechocystis* for photosynthesis and biofuel research, we believe that *Syn*ergy will become a valuable resource to the research community.

Introduction

Cyanobacteria are the only prokaryotic organisms that produce oxygen in the process of photosynthesis, and are the ancestors of higher plant chloroplasts. Not

only did cyanobacteria establish the aerobic Earth's atmosphere, they also play a crucial role in the global biochemical cycle today by fixing CO_2 and producing half of the global biomass. Being prokaryotes, cyanobacteria can be genetically modified easily and due to their fast photoautotrophic growth, they have a great potential for large scale production of renewable biofuels [1, 2] and other valuable products [1, 3, 4]. The popularity of the cyanobacteria phylum in photosynthesis and biotechnology research is reflected in the high number of sequenced cyanobacterial genomes available in Cyanobase (http://genome.microbedb.jp/cyanobase/) [5] and other public databases [6]. After the genome of the unicellular fresh water cyanobacterium *Synechocystis* sp. PCC 6803 (hereafter *Synechocystis*) was sequenced in 1996 [7], large amounts of gene expression data have been generated from cells exposed to diverse experimental conditions. Identifying groups of genes with similar expression patterns (i.e. co-expressed genes) in such data sets allows inference of functional and regulatory similarities among genes. For example, light response in *Synechocystis* has been studied using gene co-expression networks [8–10]. While these studies give insight into how cells react to single modifications, only the integration of multiple transcriptome data sets will allow a holistic understanding of the cellular response. The first meta-analysis of transcriptomics data in *Synechocystis* used a co-expression network inferred from 163 different environmental and genetic perturbations to identify a large number of genes (referred to as the Core Transcriptional Response) that are commonly regulated under most perturbations [9]. The growing interest in integrated transcriptome analysis has also led to the development of a web database, CyanoEXpress [11]. Although this tool comprises a vast set of experimental data, and integrates microarray data obtained with different experimental platforms, its use is restricted to the visualization and analysis of gene expression clusters. However, genes regulated by the same transcription factor (i.e. co-regulated genes) should not only be co-expressed, but also contain similar *cis*-regulatory elements in their promoter region. In *Synechocystis*, co-expression has not yet been linked with motif discovery in order to obtain a more mechanistic understanding of gene regulation.

We have developed *Syn*ergy, a web resource for exploring *Synechocystis* gene regulation, which integrates co-expression network analysis with motif analysis. *Syn*ergy is available at http://synergy.plantgenie.org. Considering the importance of *Synechocystis* as a model organism in biofuel production [2] and photosynthetic research [12, 13], we believe *Synergy* will become a valuable resource to many researchers.

Results and Discussion

In this article we provide an integrated analysis of co-expression networks, promoter motifs and existing gene function annotations in *Synechocystis*. See Figure 1 for an overview.

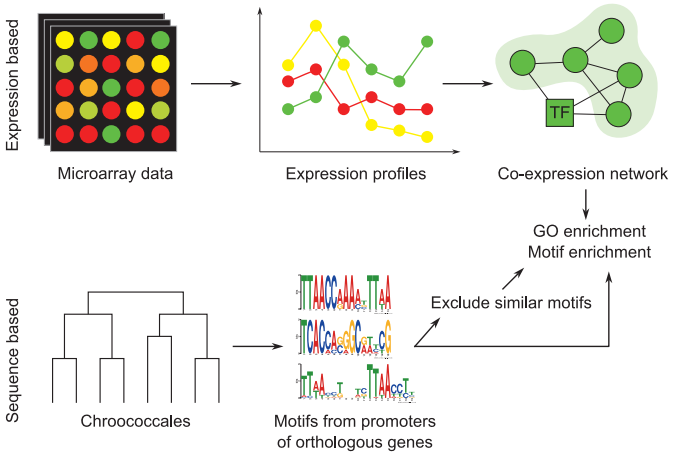

Figure 1. Overview of the data and methods used in the study. A co-expression network was inferred from gene expression, and promoter motifs were identified *de novo* from the genome sequences of orthologous species. The motif information was used to investigate if transcription factor neighborhoods were enriched for motifs compared to random network neighborhoods.

doi:10.1371/journal.pone.0113496.g001

Co-expression network inference

Co-expression networks were inferred from 371 individual microarray experiments obtained from KEGG Expression (Table 1; http://www.genome.jp/kegg/expression/; [14]). We used locally corrected mutual information scores (CLR scores, see Materials and Methods) to measure co-expression between pairs of genes, and constructed co-expression networks by linking genes with a CLR score above a preset threshold. Thus, a co-expression network is a set of nodes representing genes, which are connected by links representing co-expression above a threshold. Since some of the expression values were missing in the published data, we decided to investigate their impact by inferring two different networks; one based on a subset of samples that contained expression values for all the genes across all microarrays (subset co-expression), and another one based on all microarrays (complete co-expression). The subset co-expression network contained 3,077 genes (i.e. nodes) and 59,595 links with a CLR score above 4.0, while the corresponding complete co-expression network contained 3,067 nodes and 52,081 links.

Figure 2 shows a simplified version of the complete co-expression network where highly connected sub-networks are collapsed into single nodes (clusters) that thus represent several co-expressed genes (see Materials and Methods). Some of these clusters are associated with Gene Ontology (GO) [15] terms that are assigned more often to genes in that cluster than what one would expect by chance

Table 1. References to the microarrays used in this study.

Reference	Arrays	Conditions
[38]	18	3
[39]	20	4
[40]	4	1
[41]	22	2
[42]	11	3
[43]	46	11
[44]	144	12
[45]	38	10
[46]	4	1
[47]	4	1
[48]	14	4
[49]	28	14
[50]	18	9
Total	371	

All data can be found at http://www.genome.jp/kegg/expression/.

doi:10.1371/journal.pone.0113496.t001

(false discovery rate (FDR) [16] corrected p-value <0.05 or, equivalently, q-value <0.05). We will refer to such statistically significant overrepresentation as *enrichment*. The dominating clusters in the network display genes encoding proteins related to energy metabolism, photosynthesis, translation and protein folding. These clusters stand out not only because they contain genes with stringent regulation under the majority of stress conditions tested, but also because these genes encode proteins with inter-functional dependency. As also previously noticed [9], the expression of ribosomal genes is correlated with the expression of energy producing pathways (photosynthesis and energy metabolism); shutting down the major energy producing pathways will result in temporary translational stop. Protection from reactive oxygen species (ROS) is of tremendous importance for an oxygen-producing organism like *Synechocystis*, which is reflected by the central location of the cluster representing genes coding for enzymes involved in protein folding.

Co-expression networks can be used to quantify the importance of a gene by reporting several different measures of *network centrality* calculated for the node representing that gene. The *degree centrality* of a node is defined as the fraction of all nodes in the network that are directly connected to it (i.e. neighbors). The *betweenness centrality* of a node is the fraction of times that node is in the shortest path between two other nodes in the network (the shortest path between two nodes in a network is the fewest number of links needed to travel from one node to the other). The 40 genes with the highest *degree*- and *betweenness- centrality* (average centrality of 0.179 and 0.008, respectively) in the complete co-expression network were both enriched for genes encoding proteins involved in the photosynthetic processes (GO:0015979: *photosynthesis*, $q<0.001$ and $q<0.05$, respectively). The

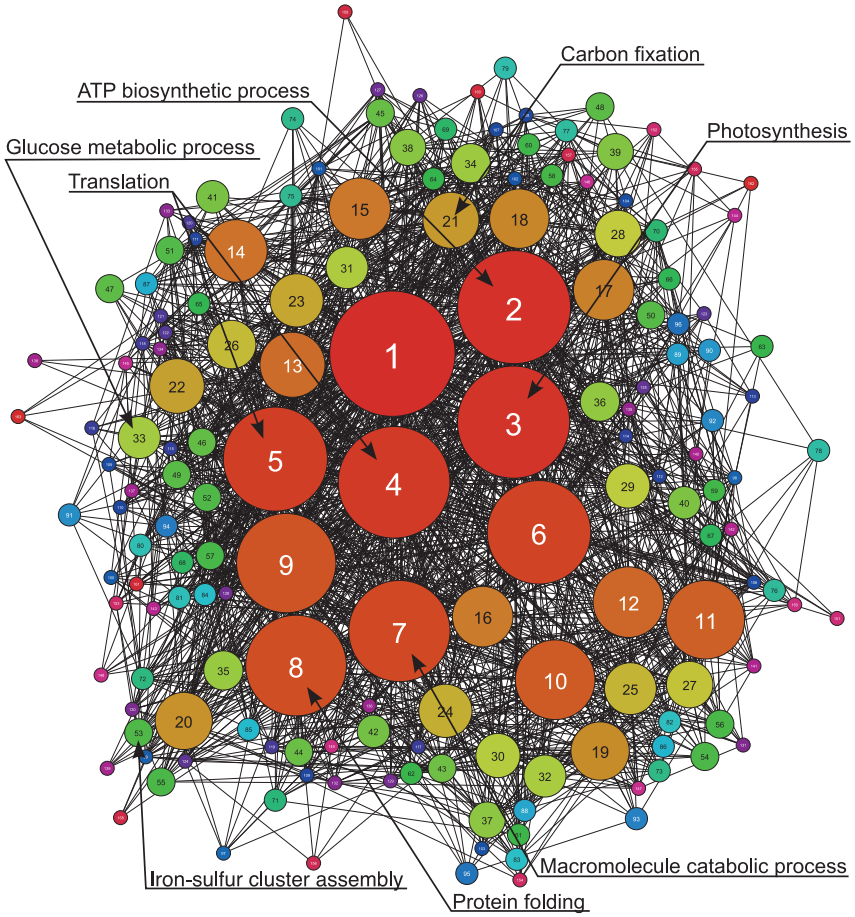

Figure 2. Clustered co-expression network. A clustered co-expression network derived from the complete co-expression network at a CLR threshold of 4.0. Each node corresponds to a set of clustered genes. The size of the nodes is proportional to the number of genes in the cluster. Two clusters are linked if they share at least one co-expressed gene pair. The annotations correspond to the most significantly enriched GO terms in the clusters ($q<0.05$).

doi:10.1371/journal.pone.0113496.g002

complete results are available in file S1. The central role of these *photosynthesis* related genes within the gene regulation of *Synechocystis* is also supported by the relatively central location of its gene cluster (Cluster 3) in Figure 2. Functional enrichment of co-expression in the model plant *Arabidopsis thaliana* has also found a cluster of genes encoding proteins involved in *photosynthesis* in a central position

[17]. This confirms the high conservation of *photosynthesis* related genes; in particular the regulation of these genes is highly conserved.

Phylogenetic footprinting

Transcription factors (TFs) bind to regulatory elements in the promoter region of genes or operons to enhance or repress their transcription. Phylogenetic footprinting was used to identify conserved DNA motifs within promoters of orthologous genes, which would indicate functional regulatory elements. We identified 8,961 groups of orthologous genes in 22 Chroococcales genomes (see file S2 for a list of organisms) and searched for conserved DNA promoter motifs using *de novo* motif finding (see Materials and Methods). Since motifs were discovered from each group of orthologous genes independently, the resulting motif set contained as many as 15,306 motifs that could be mapped to *Synechocystis* promoters, of which many were very similar or even identical. To obtain a more representative motif set, we inferred a *motif similarity network*, identified clusters in this network and compiled a final library of 4,977 *central motifs*; one motif from each cluster (see Materials and Methods). This extensive motif set displays good coverage of the *Synechocystis* promoters; already at a *q*-value threshold of 0.10 (i.e. less than 10% of the motif mappings are expected to be false positives), virtually every gene had at least one motif mapped and almost every motif in the library was mapped to at least one promoter (Figure 3).

Motif enrichment in co-expression network neighborhoods of regulatory genes

A major aim of our study was to integrate co-expression networks and regulatory motifs in order to describe gene regulation in *Synechocystis*. To this end, we rely on the assumption that genes encoding TFs are co-expressed with their target genes and that the target genes contain a specific binding site, which is used by the TF to initiate transcription. Consequently, we tested this assumption for each gene annotated with a regulatory function or DNA binding by first identifying all genes directly connected to that putative TF (i.e. the *TF neighborhood*) and then by calculating to what degree motifs occurred more often in this neighborhood than what one would expect by chance (i.e. enriched motifs). This analysis was performed for different network CLR thresholds and motif *q*-values in the complete co-expression network and in the subset network (where experiments with missing values were removed) using all discovered motifs and the non-redundant set of central motifs. Figure 4 shows that the library of central motifs resulted in more TF neighborhoods with enriched motifs ($q<0.05$) than the set of all motifs, which on one hand can be explained by the multiple hypothesis correction procedure, but on the other hand also indicates that the reduced set of central motifs covers all motif variants. Also, TF neighborhoods in the complete co-expression network contained enriched motifs more often than in the subset network, indicating that our network inference procedure copes well with data sets having missing values. Based on these

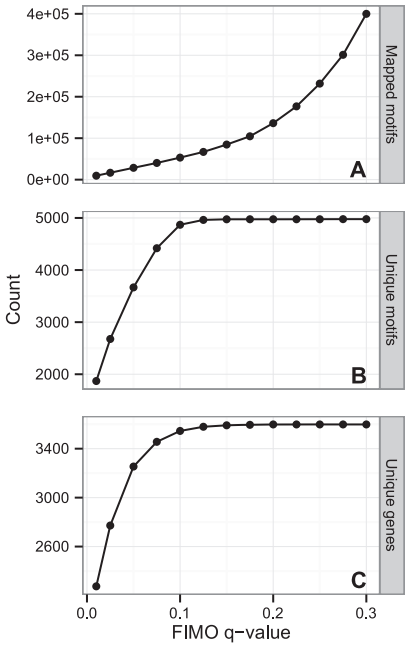

Figure 3. Central motifs mapped to *Synechocystis* promoters. The plots show the total number of times the central motifs were mapped to promoters (A), the number of unique motifs that were mapped (B) and the number of unique genes the motifs were mapped to (C) for different FIMO *q*-value thresholds.

doi:10.1371/journal.pone.0113496.g003

results, all analyses are henceforth based on the complete network and the central motifs. Interestingly, there is a relationship between the network CLR threshold and the motif *q*-value threshold, where stricter CLR thresholds require more generous *q*-value thresholds in order to maximize the number of motif-enriched TF neighborhoods. The highest number of enriched TF neighborhoods with the lowest *p*-values was observed in the complete network with a CLR threshold of four and a motif *q*-value of 0.15. Here, 105 of the 136 investigated genes with a regulatory function (77%), and 87 of the 118 investigated DNA binding genes (74%), had at least one enriched motif in its neighborhood. In total, 387 and 445 motifs were enriched in these analyses, respectively. These results are statistically highly significant, both, compared to neighborhoods of ordinary genes in the network ($p=0.001$) and compared to TF neighborhoods in randomized networks ($p<0.001$). Thus, we can conclude that co-expression and motif information to a large degree concur in *Synechocystis*. The fact that these two completely independent data sets agree so well also strengthens any biological insight inferred from our data.

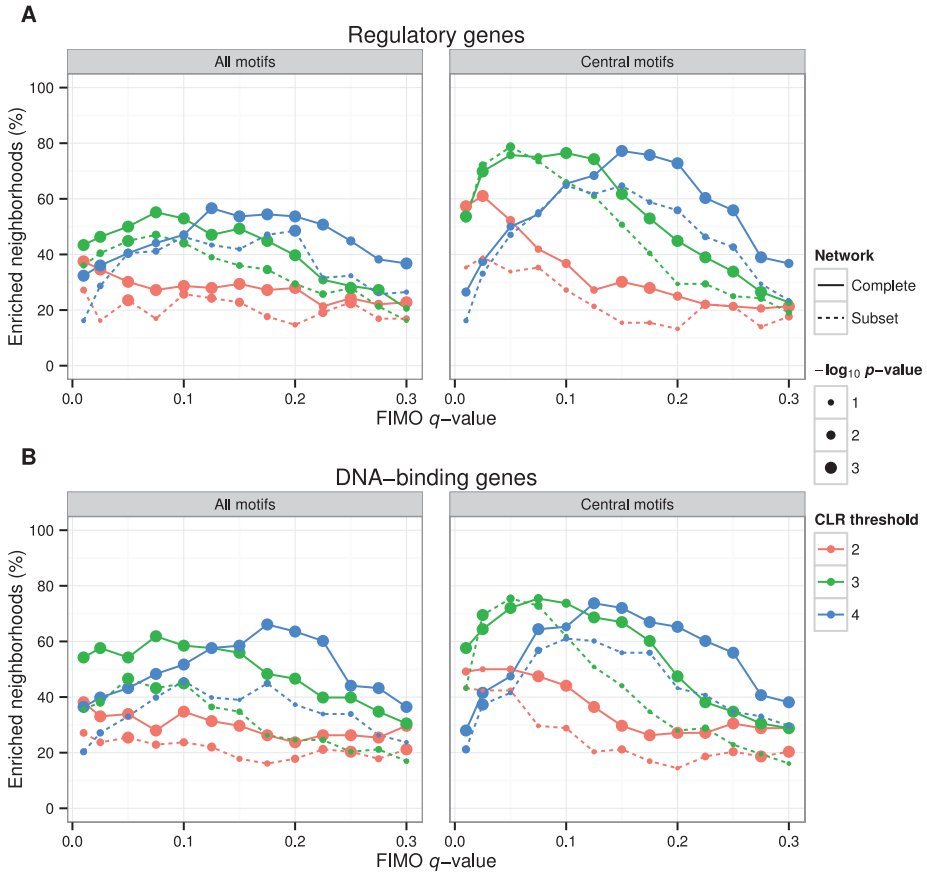

Figure 4. Gene co-expression neighborhoods with significant motif enrichment. The figure plots the fraction of neighborhoods for regulatory genes (A) and DNA-binding genes (B) with at least one significantly enriched motif ($q<0.05$) against the q-value threshold for mapping motifs to the genome. The fractions are calculated from the total number of genes in the respective groups that have gene expression data (118 DNA-binding genes and 136 regulatory genes). Plots are shown for all motifs and the subset of central motifs as well as for the complete and subset co-expression networks with different CLR thresholds. *P*-values are given for each combination of parameters and indicate the probability of observing the reported fraction of enriched neighborhoods in randomized networks.

doi:10.1371/journal.pone.0113496.g004

Conservation of co-expression in photosynthesis genes

Cyanobacteria are the evolutionary origin of the plant chloroplast. *Synechocystis* therefore is an important model system for studying photosynthesis. We investigated to what extent the co-expression of *Synechocystis* genes coding for

photosynthetic proteins is conserved in plants. 64 *Synechocystis* genes were annotated with the GO term *photosynthesis* (GO:0015979), of which 62 genes formed a connected co-expression subnetwork (CLR threshold of three, Figure 5A). 35 of these *Synechocystis* genes had at least one ortholog in *A. thaliana* ($E<$1e-5), resulting in 30 unique *A. thaliana* gene models (file S3). We analyzed these genes in the comparative network tool ComPlEx [18], and indeed confirmed that all these genes formed a co-expression cluster with the same CLR threshold of three. Moreover, this co-expression network was highly conserved also in *Oryza sativa* and *Populus trichocarpa* (Figure 5B and 5C).

Web application

We have created a web tool for integrated analysis of co-expression networks and regulatory motifs called *Syn*ergy (http://synergy.plantgenie.org). Available tools include an interactive co-expression network viewer, Gene Ontology and motif enrichment tools, precompiled gene lists and the ability to export annotated gene lists.

The natural starting point on the web site is the *gene search tool*. From here, the user can search for genes of interest or upload a list of genes (Figure 6A). There is also the possibility of using precompiled gene lists; genes annotated to a GO category, genes associated with a motif, genes in a co-expression cluster (Figure 2)

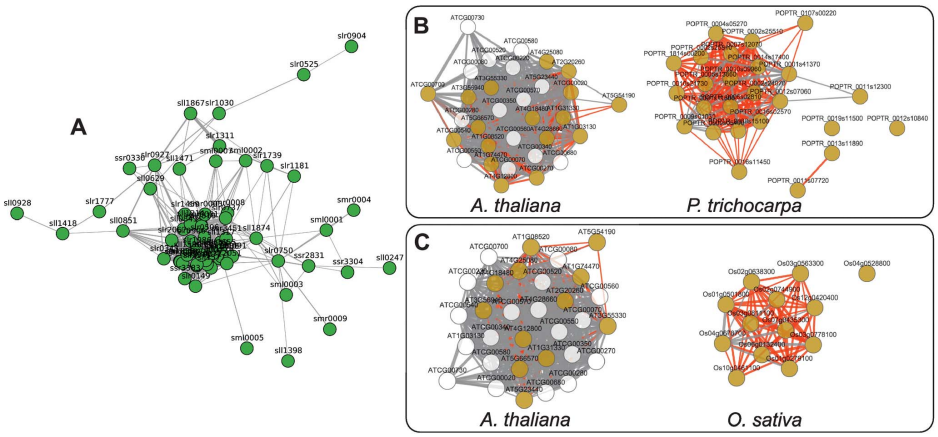

Figure 5. Conservation of photosynthesis genes. Co-expressed genes related to photosynthesis in *Synechocystis* (A) were BLASTed against *A. thaliana*. The orthologs (BLAST E-value <1e-5) were compared against *P. trichocarpa* (B) and *O. sativa* (C) using the network comparison tool ComPlEx. This revealed conservation of co-expression across all four species. Note that the *A. thaliana* genes given in white color were not measurably expressed in the other species.

doi:10.1371/journal.pone.0113496.g005

Figure 6. Web application screenshots. Gene search interface (A), network viewer (B), gene details (C) and motif details (D).

doi:10.1371/journal.pone.0113496.g006

and genes in the immediate co-expression neighborhood of a regulatory gene. For each of these gene lists, GO and motif enrichment have been pre-calculated.

Genes of interest can be added to the gene basket and these genes will be available throughout the application. The gene basket page allows the user to manage the gene basket and to calculate GO and motif enrichment for the genes currently in the basket.

The network viewer features the possibility to view and explore co-expression among sets of genes (Figure 6B). Genes that are co-expressed with the gene(s) in the current co-expression network can be found by expanding the network at any selected CLR threshold. It is also possible to export the networks in the Graph Modelling Language (GML) file format, or as publication quality PDFs.

Gene expression profiles of a chosen set of genes can be plotted across the 371 experiments and later downloaded as publication quality PDFs.

For each gene name there is a dedicated page detailing annotations, the expression profile and a list of motifs in the promoter (Figure 6C). Correspondingly, there is a dedicated page for each motif containing the motif logo, the set of genes that contain the motif in their promoters, the possibility of searching for this motif in existing motif databases and the position specific probability matrix for use in other software (Figure 6D).

To make sure that feedback from users reaches the developers by the shortest path possible, a public issue tracker is available at Github (https://github.com/maehler/Synergy/issues). Here, users can file tickets for bugs and enhancements. Documentation for the tools can be found at http://synergy.plantgenie.org/documentation.

Below we describe a number of case studies that illustrate different uses of *Syn*ergy:

Case study 1: identification of genes regulated by a known transcription factor

*Syn*ergy can be used to analyze motif occurrences in order to find candidate genes regulated by a known transcription factor. Previously, a spaced motif in the upstream region of genes involved in phosphate limitation had been identified in *Synechocystis* as well as the transcription factor recognizing this motif [19]. The consensus motif contained the direct repeat sequence [CT]TTAA[CT][CT][TA]NNN[CT]TTAA[CT][CT][TA] (Figure 7). Comparing the central region of the motif (TTAA[CT][CT][TA]NNN[CT]TTAA) with existing motifs in *Syn*ergy identified the motif NP_442272.1_1 (*E*-value 1.61e-5). A total of 56 genes contained this motif in their promoter sequence, including

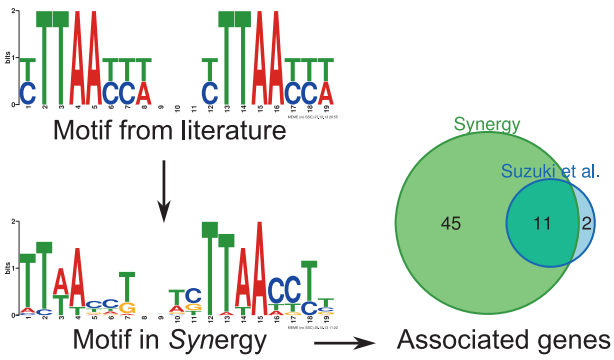

Figure 7. *Syn*ergy case study 1. A regulatory motif and its transcription factor were extracted from the literature [19]. Searching for the motif in *Syn*ergy identified a number of genes that were experimentally determined to be regulated by this transcription factor.

doi:10.1371/journal.pone.0113496.g007

slr0447 (*urtA*), *slr1247* (*pstS2*) and *sll0679* (*sphX*) that have been reported to be up- or down-regulated under phosphate limiting conditions [19]. However, *slr1247* and *sll0679* are leading genes in two operons according to information in Cyanobase. Assuming that the downstream genes in these operons are also regulated by the motif, we identified 11 of the 13 genes reported by [19].

Case study 2: motif analysis to reveal protein function

*Syn*ergy further can be used to investigate the relationship between a set of genes by integrated analysis of both motifs and co-expression. A search for genes coding for proteins related to the two photosystems in the *Syn*ergy gene search tool resulted in 51 genes that subsequently were tested for regulatory motif enrichment. The motif NP_441569.1_8 was ranked as the second most enriched motif (q-value <0.001), and its best match in the Prodoric database was MX000068 in *Bacillus subtilis*. A sigma factor is known to bind to this motif, and using protein BLAST revealed a number of sigma factors with highly significant E-values ($<$1e-10) in *Synechocystis*.

With this information in hand, a new gene search was performed, in which all genes coding for proteins annotated as sigma factors were added to the existing selection of genes. Looking at the co-expression network for these genes revealed that genes coding for photosystems together with those coding for sigma factors formed a connected subnetwork (CLR threshold of three). Our analysis thus supports previous data showing that sigma factors play a vital role in controlling the stoichiometry of the photosystems within the thylakoid membrane [20, 21].

Case study 3: functional role of hypothetical proteins

*Syn*ergy can be used to assign functions to unknown or hypothetical proteins based on co-expressed genes with known function. The CP12 protein encoded by *ssl3364* is highly conserved in all photosynthetic organisms, but is annotated as a hypothetical protein in Cyanobase. In higher plants and algal species (reviewed by [22]) it was found to be involved in the thioredoxin-mediated regulation of the Calvin-Benson cycle [22]. Moreover, additional functions are hypothesized for this protein in plants [22] and a comparative analysis of 126 cyanobacterial genomes reveals functional diversity among its orthologues [23]. A co-expression neighborhood analysis of *ssl3364* (CLR threshold of four with an expansion threshold of five) generated a densely connected cluster of 54 genes and 798 links. The neighborhood is dominated by genes encoding proteins of the oxidative stress response like chaperones and proteases, and is enriched in genes coding for enzymes involved in protein folding (GO:0006457, q-value <0.01). We hypothesize a new biological function for the CP12 protein in *Synechocystis*, *i.e.* protection from oxidative stress, similar to the function of its orthologues in *A. thaliana* and *Chlamydomonas reinhardtii*, which have been shown to protect Calvin-Benson enzymes from oxidative stress [24].

Case study 4: TF neighborhoods contain biologically relevant motifs

We have shown that the neighborhoods of TFs in our co-expression networks contain common motifs more often than by chance (enriched motifs). To see whether experimental data support that these automatically discovered promoter motifs in fact bind TFs, external motif databases were explored. The gene *sll0998*, for example, encodes a LysR family transcription regulator. In the co-expression network (complete network, CLR threshold of 4) this TF is connected to eight neighboring genes with three enriched motifs in their promoters ($q<0.05$). One of the motifs was NP_440076.1_5. Searching for motifs similar to NP_440076.1_5 in Prodoric resulted in the motif MX000155 known to be regulated by OxyR in *E. coli*. Using protein BLAST to search for homologs of OxyR in *Synechocystis* gave a highly significant hit ($E=1e\text{-}26$) to the protein product of *sll0998*.

Conclusions

We have developed a web tool, *Syn*ergy, allowing interactive analysis of the *Synechocystis* genome by integrating co-expression networks, regulatory elements and existing knowledge such as functional annotations and known regulatory genes and elements. Furthermore, we have demonstrated the usefulness of this tool in finding both previously published and new biologically relevant regulatory links in *Synechocystis*.

Materials and Methods

Microarray data

A total of 371 individual microarray experiments were downloaded from Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg/expression/). All of the data were based on the Takara microarray chips that covers 83% (3,079/3,726) of the genes in *Synechocystis* [25]. The data were combined into a single data set and normalized with the limma package [26] in R; a software environment for statistical computing and graphics.

Annotations

Gene annotations were retrieved from Cyanobase. In total, 146 genes were annotated as coding for enzymes with a regulatory function. In this study, these genes were treated as coding for known transcription factors. In Cyanobase, there were also functional annotations translated into GO terms. In total, 2,040 *Synechocystis* genes were annotated to 2,076 GO terms.

Co-expression inference

Mutual Information (MI) and Context Likelihood of Relatedness (CLR) were used to infer co-expression networks from the microarray data. MI is a metric that does not assume linearity or continuity when measuring the dependence between

two variables. This makes it possible to detect relationships that would be undetected by other methods, such as the Pearson correlation coefficient. CLR then finds the most statistically significant co-expression neighbors of each gene based on the local background distribution of MI scores to all other genes [27]. From the z-scores produced by the CLR algorithm, a co-expression network was constructed. A co-expression network can be defined as a collection of nodes (genes) and links (co-expression relationships) where the links are weighted according to the strength of the co-expression.

To account for the large number of missing values in the complete dataset, two different co-expression networks were constructed: the complete co-expression network using all samples (i.e. all 371 microarray experiments) and the subset co-expression network using only the samples with no missing values (67 samples).

Phylogenetic footprinting

MEME [28] was used to find potential regulatory motifs in groups of orthologs (so-called phylogenetic footprinting). The proteomes of 22 organisms in the Chroococcales taxon (file S2) were downloaded from NCBI and clustered with OrthoMCL [29]. MEME was then used to find conserved motifs in the promoter regions of the corresponding genes in each group. A promoter was defined as the 400 bp sequence upstream of the transcription start site, and the promoters were retrieved using Regulatory Sequence Analysis Tools (RSAT) [30]. MEME was instructed to find motifs between 8 and 20 bp in length with an *E*-value threshold of 100. The MEME motifs were then mapped back to the *Synechocystis* promoters using FIMO [31] and motifs with a *q*-value below 0.3 were kept.

The phylogenetic footprinting approach resulted in many motifs that were similar to each other. To eliminate duplicates, a motif similarity network was constructed. The similarities were calculated by CompariMotif [32] using the consensus motifs derived from the position specific scoring matrices (PSSMs) as input. The motif network was then clustered using MCL [33]. The motif with the highest *betweenness centrality* was chosen as a representative motif from each cluster (central motif).

Motif and GO enrichment

To calculate enrichment of motifs or GO terms in a set of genes, Fisher's exact test was used. The test was implemented using the Python library scipy (v0.13.3) (http://www.scipy.org). To correct for multiple testing, false discovery rate (FDR) adjustment was used and *q*-values were reported.

Motif enrichment in network neighborhoods

For genes of interest, the immediate co-expression neighborhood was extracted and motif overrepresentation was calculated for these neighbors. The analysis was performed on genes annotated with *regulatory function* and genes annotated with *DNA-binding*. As a negative control, 1,000 random gene lists with 100 genes in

each were used. In all gene sets, genes without expression values were excluded since they will not be present in the co-expression networks. Both, the complete and the subset co-expression networks were used with CLR thresholds of 3, 4 and 5. We also tested different sets of motifs mapped to the genome as defined by different FIMO *q*-value thresholds. For each neighborhood and parameter combination, motif enrichment was calculated using Fisher's exact test and FDR correction as described above, excluding the gene from which the neighborhood was created. If a neighborhood had at least one overrepresented motif with $q<0.05$, the neighborhood was considered to be enriched. To test for significance of the enrichment in the context of networks, motif enrichment was also performed in networks where node labels had been randomly shuffled.

Web application implementation

The *Syn*ergy web application was developed with the PHP framework CodeIgniter (http://ellislab.com/codeigniter). The network viewer was implemented with the JavaScript library Cytoscape.js (http://cytoscape.github.io/cytoscape.js/), the successor of the Flash interface Cytoscape Web [34].

TOMTOM [35] was used for comparing motifs to known regulatory elements in other organisms. The PRODORIC [36] and RegTransBase [37] prokaryotic motif databases were downloaded from the MEME website.

Supporting Information

File S1. GO enrichment of the genes with the highest centrality.
doi:10.1371/journal.pone.0113496.s001 (XLS)

File S2. Number of coding regions vs. genome size for the organisms used during the phylogenetic footprinting.
doi:10.1371/journal.pone.0113496.s002 (XLS)

File S3. Best sequence alignments with *Arabidopsis* genes.
doi:10.1371/journal.pone.0113496.s003 (XLS)

Author Contributions

Conceived and designed the experiments: SN NM TRH. Analyzed the data: CF OC TRH SN NM. Wrote the paper: CF OC TRH NM SN.

References

1. **Lee H-S, Vermaas WFJ, Rittmann BE** (2010) Biological hydrogen production: prospects and challenges. Trends Biotechnol 28: 262–271. Available: http://www.ncbi.nlm.nih.gov/pubmed/20189666. Accessed 2014 Mar 21.
2. **Machado IMP, Atsumi S** (2012) Cyanobacterial biofuel production. J Biotechnol 162: 50–56. Available: http://www.ncbi.nlm.nih.gov/pubmed/22446641. Accessed 2013 May 28.

3. **Lindberg P, Park S, Melis A** (2010) Engineering a platform for photosynthetic isoprene production in cyanobacteria, using Synechocystis as the model organism. Metab Eng 12: 70–79. Available: http://www.ncbi.nlm.nih.gov/pubmed/19833224. Accessed 2013 Jun 2.

4. **Englund E, Pattanaik B, Ubhayasekera SJK, Stensjö K, Bergquist J, et al.** (2014) Production of Squalene in Synechocystis sp. PCC 6803. PLoS One 9: e90270. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3953072&tool=pmcentrez&rendertype=abstract. Accessed 2014 Apr 7.

5. **Nakao M, Okamoto S, Kohara M, Fujishiro T, Fujisawa T, et al.** (2010) CyanoBase: the cyanobacteria genome database update 2010. Nucleic Acids Res 38: D379–81. Available: http://www.ncbi.nlm.nih.gov/pubmed/19880388.

6. **Fujisawa T, Okamoto S, Katayama T, Nakao M, Yoshimura H, et al.** (2014) CyanoBase and RhizoBase: databases of manually curated annotations for cyanobacterial and rhizobial genomes. Nucleic Acids Res 42: D666–70. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965071&tool=pmcentrez&rendertype=abstract. Accessed 2014 Apr 11.

7. **Kaneko T, Sato S, Kotani H, Tanaka a, Asamizu E, et al.** (1996) Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions (supplement). DNA Res 3: 185–209. Available: http://www.ncbi.nlm.nih.gov/pubmed/8905238.

8. **Aurora R, Hihara Y, Singh AK, Pakrasi HB** (2007) A network of genes regulated by light in cyanobacteria. OMICS 11: 166–185. Available: http://www.ncbi.nlm.nih.gov/pubmed/17594236. Accessed 2014 Apr 7.

9. **Singh AK, Elvitigala T, Cameron JC, Ghosh BK, Bhattacharyya-Pakrasi M, et al.** (2010) Integrative analysis of large scale expression profiles reveals core transcriptional response and coordination between multiple cellular processes in a cyanobacterium. BMC Syst Biol 4: 105. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2924297&tool=pmcentrez&rendertype=abstract.

10. **Miranda H, Cheregi O, Netotea S, Hvidsten TR, Moritz T, et al.** (2013) Co-expression analysis, proteomic and metabolomic study on the impact of a Deg/HtrA protease triple mutant in Synechocystis sp. PCC 6803 exposed to temperature and high light stress. J Proteomics 78: 294–311. Available: http://www.ncbi.nlm.nih.gov/pubmed/23063787. Accessed 2014 Apr 11.

11. **Hernandez-prieto MA, Futschik ME** (2012) CyanoEXpress: A web database for exploration and visualisation of the integrated transcriptome of. Bioinformation 8.

12. **Knoop H, Zilliges Y, Lockau W, Steuer R** (2010) The metabolic network of Synechocystis sp. PCC 6803: systemic properties of autotrophic growth. Plant Physiol 154: 410–422. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2938163&tool=pmcentrez&rendertype=abstract. Accessed 2014 Jan 28.

13. **Knoop H, Gründel M, Zilliges Y, Lehmann R, Hoffmann S, et al.** (2013) Flux balance analysis of cyanobacterial metabolism: the metabolic network of Synechocystis sp. PCC 6803. PLoS Comput Biol 9: e1003081. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3699288&tool=pmcentrez&rendertype=abstract. Accessed 2014 Jan 24.

14. **Kanehisa M** (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28: 27–30. Available: http://nar.oxfordjournals.org/cgi/content/long/28/1/27. Accessed 2013 May 27.

15. **Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al.** (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29. Available: http://www.ncbi.nlm.nih.gov/pubmed/10802651.

16. **Benjamini Y, Hochberg Y** (1995) Controlling the False Discovery Rate: A practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. Available: http://www.jstor.org/stable/10.2307/2346101. Accessed 2013 Jun 24.

17. **Mentzen WI, Wurtele ES** (2008) Regulon organization of Arabidopsis. BMC Plant Biol 8: 99. Available: http://www.biomedcentral.com/1471-2229/8/99. Accessed 2014 Jul 11.

18. **Netotea S, Sundell D, Street NR, Hvidsten TR** (2014) ComPlEx: conservation and divergence of co-expression networks in A. thaliana, Populus and O. sativa. BMC Genomics 15: 106. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3925997&tool=pmcentrez&rendertype=abstract. Accessed 2014 Feb 26.

19. **Suzuki S, Ferjani A, Suzuki I, Murata N** (2004) The SphS-SphR two component system is the exclusive sensor for the induction of gene expression in response to phosphate limitation in

synechocystis. J Biol Chem 279: 13234–13240. Available: http://www.ncbi.nlm.nih.gov/pubmed/14707128. Accessed 2013 Dec 19.

20. **Tozawa Y, Teraishi M, Sasaki T, Sonoike K, Nishiyama Y, et al.** (2007) The plastid sigma factor SIG1 maintains photosystem I activity via regulated expression of the psaA operon in rice chloroplasts. Plant J 52: 124–132. Available: http://www.ncbi.nlm.nih.gov/pubmed/17651366. Accessed 2014 Feb 26.

21. **Shimizu M, Kato H, Ogawa T, Kurachi A, Nakagawa Y, et al.** (2010) Sigma factor phosphorylation in the photosynthetic control of photosystem stoichiometry. Proc Natl Acad Sci U S A 107: 10760–10764. Available: http://www.pnas.org/content/107/23/10760.short. Accessed 2014 Feb 26.

22. **López-Calcagno PE, Howard TP, Raines C a** (2014) The CP12 protein family: a thioredoxin-mediated metabolic switch? Front Plant Sci 5: 9. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3906501&tool=pmcentrez&rendertype=abstract. Accessed 2014 Mar 26.

23. **Stanley D, Raines C, Kerfeld C** (2013) Comparative analysis of 126 cyanobacterial genomes reveals evidence of functional diversity among homologs of the redox-regulated CP12 protein. Plant Physiol 161: 824–835. Available: http://www.plantphysiol.org/content/161/2/824.short. Accessed 2014 Apr 7.

24. **Marri L, Thieulin-Pardo G, Lebrun R, Puppo R, Zaffagnini M, et al.** (2014) CP12-mediated protection of Calvin-Benson cycle enzymes from oxidative stress. Biochimie 97: 228–237. Available: http://www.ncbi.nlm.nih.gov/pubmed/24211189. Accessed 2014 Mar 20.

25. **Los DA, Zorina A, Sinetova M, Kryazhov S, Mironov K, et al.** (2010) Stress Sensors and Signal Transducers in Cyanobacteria. Sensors 10: 2386–2415.

26. **Smyth GK** (2005) Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, editors. Bioinformatics and Compuational Biology Solutions Using R and Bioconductor. Springer. pp. 397–420.

27. **Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, et al.** (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 5: e8. Available: http://www.ncbi.nlm.nih.gov/pubmed/17214507.

28. **Bailey TL, Williams N, Misleh C, Li WW** (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34: W369–W373. Available: http://dx.doi.org/10.1093/nar/gkl198.

29. **Li L, Stoeckert CJJ, Roos DS** (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=403725&tool=pmcentrez&rendertype=abstract. Accessed 2014 Mar 19.

30. **Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, et al.** (2011) RSAT 2011: regulatory sequence analysis tools. Nucleic Acids Res 39: W86–W91. Available: http://dx.doi.org/10.1093/nar/gkr377.

31. **Grant CE, Bailey TL, Noble WS** (2011) FIMO: scanning for occurrences of a given motif. Bioinformatics 27: 1017–1018. Available: http://www.ncbi.nlm.nih.gov/pubmed/21330290. Accessed 2013 Mar 20.

32. **Edwards RJ, Davey NE, Shields DC** (2008) CompariMotif: quick and easy comparisons of sequence motifs. Bioinformatics 24: 1307–1309. Available: http://www.ncbi.nlm.nih.gov/pubmed/18375965. Accessed 2013 Jul 4.

33. **Enright AJ, Van Dongen S, Ouzounis C a** (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575–1584. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=101833&tool=pmcentrez&rendertype=abstract.

34. **Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, et al.** (2010) Cytoscape Web: an interactive web-based network browser. Bioinformatics 26: 2347–2348. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2935447&tool=pmcentrez&rendertype=abstract. Accessed 2014 Jan 31.

35. **Gupta S, Stamatoyannopoulos J a, Bailey TL, Noble WS** (2007) Quantifying similarity between motifs. Genome Biol 8: R24. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1852410&tool=pmcentrez&rendertype=abstract. Accessed 2014 Jan 20.

36. **Munch R** (2003) PRODORIC: prokaryotic database of gene regulation. Nucleic Acids Res 31: 266–269. Available: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkg037. Accessed 2014 Feb 13.

37. **Cipriano MJ, Novichkov PN, Kazakov AE, Rodionov DA, Arkin AP, et al.** (2013) RegTransBase – a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes. BMC Genomics 14: 213. Available: http://www.biomedcentral.com/1471-2164/14/213.

38. **Suzuki I, Kanesaki Y, Mikami K, Kanehisa M, Murata N** (2001) Cold-regulated genes under control of the cold sensor Hik33 in Synechocystis. Mol Microbiol 40: 235–244. Available: http://www.ncbi.nlm.nih.gov/pubmed/11298290.

39. **Hihara Y, Kamei A, Kanehisa M, Kaplan A, Ikeuchi M** (2001) DNA Microarray Analysis of Cyanobacterial Gene Expression during Acclimation to High Light. Plant Cell 13: 793–806. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=135531&tool=pmcentrez&rendertype=abstract.

40. **Yoshimura H, Yanagisawa S, Kanehisa M, Ohmori M** (2002) Screening for the target gene of cyanobacterial cAMP receptor protein SYCRP1. Mol Microbiol 43: 843–853.

41. **Hihara Y, Sonoike K, Kanehisa M, Ikeuchi M** (2003) DNA microarray analysis of redox-responsive genes in the genome of the cyanobacterium Synechocystis sp. strain PCC 6803. J Bacteriol 185: 1719–1725. Available: http://jb.asm.org/cgi/content/abstract/185/5/1719.

42. **Kobayashi M, Ishizuka T, Katayama M, Kanehisa M, Bhattacharyya-Pakrasi M, et al.** (2004) Response to oxidative stress involves a novel peroxiredoxin gene in the unicellular cyanobacterium Synechocystis sp. PCC 6803. Plant cell Physiol 45: 290–299. Available: http://www.ncbi.nlm.nih.gov/pubmed/15047877.

43. **Paithoonrangsarid K, Shoumskaya MA, Kanesaki Y, Satoh S, Tabata S, et al.** (2004) Five histidine kinases perceive osmotic stress and regulate distinct sets of genes in Synechocystis. J Biol Chem 279: 53078–53086. Available: http://www.ncbi.nlm.nih.gov/pubmed/15471853.

44. **Kucho K, Okamoto K, Tsuchiya Y, Nomura S, Nango M, et al.** (2005) Global Analysis of Circadian Expression in the Cyanobacterium Synechocystis sp. Strain PCC 6803. J Bacteriol 187: 2190–2199. Available: http://jb.asm.org/cgi/content/abstract/187/6/2190.

45. **Shoumskaya MA, Paithoonrangsarid K, Kanesaki Y, Los DA, Zinchenko V V, et al.** (2005) Identical Hik-Rre systems are involved in perception and transduction of salt signals and hyperosmotic signals but regulate the expression of individual genes to different extents in synechocystis. J Biol Chem 280: 21531–21538. Available: http://www.ncbi.nlm.nih.gov/pubmed/15805106.

46. **Panichkin VB, Arakawa-Kobayashi S, Kanaseki T, Suzuki I, Los DA, et al.** (2006) Serine/threonine protein kinase SpkA in Synechocystis sp. strain PCC 6803 is a regulator of expression of three putative pilA operons, formation of thick pili, and cell motility. J Bacteriol 188: 7696–7699. Available: http://jb.asm.org/cgi/content/long/188/21/7696. Accessed 2013 Jun 19.

47. **Kanesaki Y, Los DA, Suzuki I, Murata N** (2010) Sensors and Signal Transducers of Environmental Stress in Cyanobacteria. In: Pareek A, Sopory SK, Bohnert HJ, editors. Abiotic Stress Adaptation in Plants SE - 2. Springer Netherlands. pp. 15–31. Available: http://dx.doi.org/10.1007/978-90-481-3112-9_2.

48. **Prakash JSS, Sinetova M, Zorina A, Kupriyanova E, Suzuki I, et al.** (2009) DNA supercoiling regulates the stress-inducible expression of genes in the cyanobacterium Synechocystis. Mol Biosyst 5: 1904–1912.

49. **Panichkin** (2008) Ser/Thr protein kinases are involved in cold-signal transduction in a cyanobacterium. Available: http://www.genome.jp/kegg-bin/get_htext?htext=Exp_DB&hier=1. Accessed 2014 Jan 21.

50. **Prakash JSS, Krishna PS, Sirisha K, Kanesaki Y, Suzuki I, et al.** (2010) An RNA helicase, CrhR, regulates the low-temperature-inducible expression of heat-shock genes groES, groEL1 and groEL2 in Synechocystis sp. PCC 6803. Microbiology 156: 442–451. Available: http://www.ncbi.nlm.nih.gov/pubmed/19926653.

Paper II

Robinson et al. BMC Plant Biology 2014, 14:276
http://www.biomedcentral.com/1471-2229/14/276

BMC Plant Biology

RESEARCH ARTICLE Open Access

Populus tremula (European aspen) shows no evidence of sexual dimorphism

Kathryn M Robinson[1†], Nicolas Delhomme[1†], Niklas Mähler[2], Bastian Schiffthaler[1], Jenny Önskog[1], Benedicte R Albrectsen[1,3], Pär K Ingvarsson[4], Torgeir R Hvidsten[1,2], Stefan Jansson[1] and Nathaniel R Street[1*]

Abstract

Background: Evolutionary theory suggests that males and females may evolve sexually dimorphic phenotypic and biochemical traits concordant with each sex having different optimal strategies of resource investment to maximise reproductive success and fitness. Such sexual dimorphism would result in sex biased gene expression patterns in non-floral organs for autosomal genes associated with the control and development of such phenotypic traits.

Results: We examined morphological, biochemical and herbivory traits to test for sexually dimorphic resource allocation strategies within collections of sexually mature and immature *Populus tremula* (European aspen) trees. In addition we profiled gene expression in mature leaves of sexually mature wild trees using whole-genome oligonucleotide microarrays and RNA-Sequencing.

Conclusions: We found no evidence of sexual dimorphism or differential resource investment strategies between males and females in either sexually immature or mature trees. Similarly, single-gene differential expression and machine learning approaches revealed no evidence of large-scale sex biased gene expression. However, two significantly differentially expressed genes were identified from the RNA-Seq data, one of which is a robust diagnostic marker of sex in *P. tremula*.

Keywords: Sexual dimorphism, RNA-Sequencing, transcriptomics, *Populus tremula*, dioecious

Background

Sexual dimorphism, the differentiation of both primary (*i.e.* gonads) and secondary (other morphological, behavioural and physiological) sex characteristics is the norm in animal systems [1]. In angiosperms the majority of extant species are co-sexual, being either monoecious or hermaphroditic (*i.e.* they bear separate male and female flowers or have either flowers containing both sexual organs, respectively). However, ~4% of plant species are dioecious [2,3], with different individuals producing only male or female flowers, and it is thought that dioecy evolved from ancestral hermaphrodites, which inherently lack sex chromosomes [4]. In several animal systems including nematodes, insects and mammals, sex determination is well characterised [5], whereas the molecular mechanisms underlying dioecious sex determination in plants remain largely unresolved [4,6]. The emergence of dioecy appears to have occurred relatively recently in many plant species, with sex determining loci being located in small regions of reduced recombination where there may not yet have been adequate time for heteromorphic sex chromosomes to have evolved [4].

Evolutionary theory suggests that sexual dimorphism arises after release from a co-sexual state as each sex adapts to a new fitness optimum following the removal of constraints previously imparted by the other sex – *i.e.* that trade-offs necessarily exist between the male and female functions in a monoecious state [4,7,8]. With the exception of sex-determining loci (or chromosomes), males and females share the same genome. Thus sexually dimorphic phenotypes that are not controlled by genes within the sex determining loci/chromosome must result from differential expression regulation of autosomal genes involved in the development and control of those traits [1]. Examples of expected sexual trade-offs include differential optimal strategies of resource allocation to growth and secondary metabolites (such as phenolic compounds) given production of either pollen or seeds; for example, females may allocate more carbon to secondary metabolites at the

* Correspondence: nathaniel.street@umu.se
†Equal contributors
[1]Department of Plant Physiology, Umeå Plant Science Centre, Umeå University, 901 87 Umeå, Sweden
Full list of author information is available at the end of the article

expense of stem growth in order to protect seeds from predators and pathogens [9,10], resulting in males and females experiencing contrasting selective pressures [8,11].

The genus *Populus* includes poplars, aspens, and cottonwoods and is a well-established model system [12] with a high quality genome sequence available for *P. trichocarpa* [13,14]. *Populus* species and hybrids have numerous industrial and silvicultural uses [15,16] and are often keystone species [17,18]. In *Populus,* dioecy is the common condition with the only exception being the monoecious, hermaphroditic *P. lasiocarpa* (see citations in [19]). There are also rare cases of gender reversion, perfect (bisexual) flower formation and even mature seed catkin formation on male trees [19-21 and citations in 22]. *Populus* species do not have heteromorphic sex specific chromosomes [22], and the molecular mechanism of sex determination remains undetermined, although sex is genetically determined [23]. In *P. trichocarpa* there is substantial evidence that the sex-determining locus is located in the peritelomeric region of chromosome 19 [22,23]. For all *Populus* genetic maps where sex has been included as a marker during map construction, there is always a single sex-linked locus that is located on chromosome 19. However, its location on that chromosome varies in different sections of the genus. There are also contrasting reports as to which sex is heterogametic [22,24-26]. In the aspens it is now well established that the sex determination locus is located in the pericentromeric region of chromosome 19 [24-28]. Pakull *et al*. [28] recently identified that Potri.019G047300, a gene that the same group had previously identified as a candidate in the sex determination locus [24], is either completely or partially deleted specifically in females, a finding that we independently discovered and detail below.

There is a current lack of knowledge of whether global or specific patterns of sex biased gene expression exist in non-reproductive tissues of dioecious plant species [4]. To date, this has been investigated in a single study of *Silene latifolia* [29], which considered only 22 ESTs. Here we addressed this question using *P. tremula,* which produces high amounts of phenolic-based secondary metabolites that have been implicated in defence against herbivores and pathogens [30,31] making it a suitable model system to test for sexually dimorphic differences in resource allocation to growth and defence. We explored global gene expression patterns in combination with a set of diagnostic phenotypes in non-reproductive tissues (leaves) of sexually mature *P. tremula*. The same phenotypes were additionally assayed in sexually immature trees. Gene expression was profiled using both whole genome oligonucleotide microarrays and RNA-Sequencing (RNA-Seq). The expression data were used for both individual gene differential expression tests as well as a machine learning approach to test for genomic regions containing combinations of genes exhibiting sex-related expression differences.

Results

Phenotypic analysis reveals no evidence of sexual dimorphism in P. tremula

We found no evidence of sexual dimorphism in tree height or diameter (Additional file 1) in either the Umeå Aspen collection (UmAsp; [32]) or the Swedish Aspen (SwAsp; [33]) samples or for height increment, a measure of vigour, in the juvenile SwAsp samples (Figure 1a, Additional file 1). Similarly, we found no statistical evidence of sexual dimorphism for leaf area (Figure 1b), leaf nutritional quality (nitrogen and carbon content and their ratio, Figure 1c) or specific secondary metabolites (total phenolics and condensed tannins, Figure 2a-b) in either the UmAsp or SwAsp samples (Additional file 1). All SwAsp phenotypic data except carbon and nitrogen concentration were generated by Robinson *et al*. [34], who showed that these, and other, traits had a substantial degree of heritability (clonal repeatability), a result that could only be obtained from high quality phenotypic data, negating the possibility that

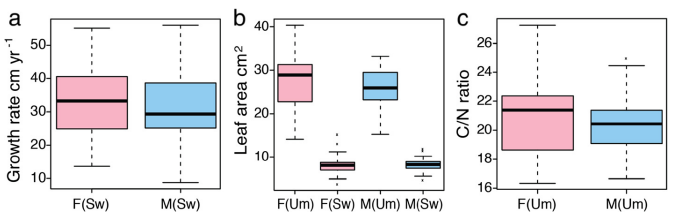

Figure 1 Growth and resource allocation in male and female *Populus temula* trees. Boxes representing females (F) are coloured pink and males (M) coloured blue, in the Umeå Aspen (Um) and Swedish Aspen (Sw) collections. **(a)** Growth rate calculated as height increment over five years in SwAsp. Analysis of Variance (ANOVA) results showed no significant sex differences ($F_{1,45}$ = 0.448, P =0.507) **(b)** Individual leaf area in UmAsp and SwAsp. ANOVA results showed no significant sex differences for samples from either the Um ($F_{1,38}$ = 0.958, P =0.334) or Sw ($F_{1,44}$ = 0.012, P =0.914) collections. **(c)** Foliar carbon/nitrogen ratio in Um. ANOVA results showed no significant sex differences ($F_{1,38}$ = 0.631, P =0.432).

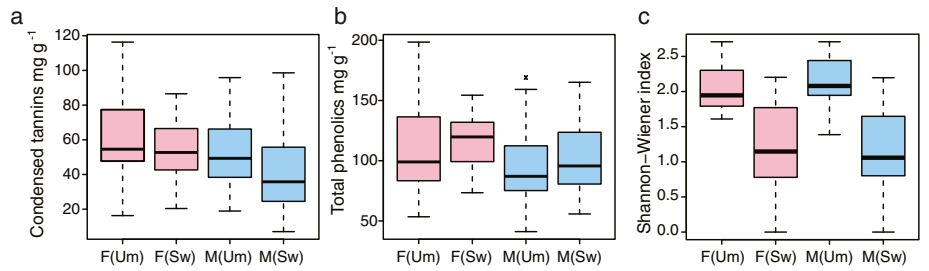

Figure 2 Secondary metabolite and herbivory phenotypes in male and female *Populus tremula* in the Umeå Aspen (Um) and Swedish Aspen (Sw) collections. Boxes representing females (F) are coloured pink and males (M) coloured blue. **(a)** Foliar condensed tannins. Analysis of Variance (ANOVA) results showed no significant sex differences for samples from either the Um ($F_{1,38}$ = 1.667, P =0.203) or Sw ($F_{1,45}$ = 2.764, P =0.103) collections **(b)** Foliar total phenolic concentrations. ANOVA results showed no significant sex differences for samples from either the Um ($F_{1,38}$ = 01941, P =0.172) or Sw ($F_{1,45}$ = 2.561, P =0.117) collections. **(c)** Shannon-Wiener index of arthropod herbivore diversity. ANOVA results showed no significant sex differences for samples from either the Um ($F_{1,38}$ = 0.659 P =0.422) or Sw ($F_{1,45}$ = 0.074, P =0.787) collections.

the observed lack of significant sexual dimorphism resulted from low data quality.

Herbivorous insects display no sexual preference

Arthropods are common folivores on *P. tremula* and numerous aspen-associated morphospecies have been recorded [34]. We found no statistically significant sex-related differences for arthropod abundance, species richness, feeding guild abundances, or the Shannon-Wiener diversity index in either the UmAsp or SwAsp samples (Figure 2c, Additional file 1). We also found no statistically significant sex-related differences in the arthropod community of UmAsp and SwAsp analysed by non-parametric Multivarite Analysis of Variance (MANOVA; UmAsp: $F_{1,38}$ = 0.325, P =0.808; SwAsp: $F_{1,45}$ = 0.825, P =0.5, Additional file 1).

Transcript profiling reveals no global patterns of sex-biased expression

We profiled gene expression in mature leaves of male and female *P. tremula* from the UmAsp collection using whole genome oligonucleotide microarrays (Figure 3) and RNA-Sequencing (RNA-Seq; (Figure 4). The samples used for RNA-Seq profiling were collected in two years and a Principle Component Analysis (PCA) analysis revealed

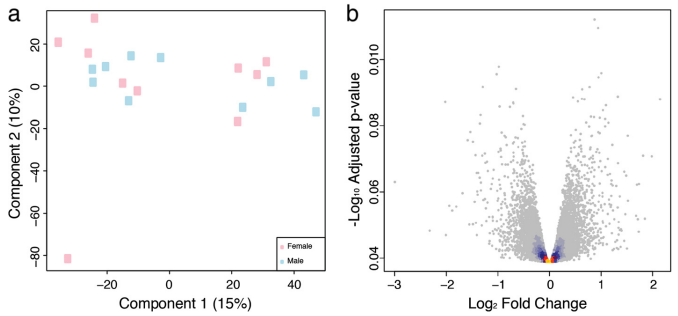

Figure 3 Overview of microarray gene expression patterns in male and female *Populus tremula* trees from the Umeå Aspen collection. **(a)** Principal Component Analysis plot of the microarray data with samples classified by sex (male in blue, female in pink). The percentage variance explained by each component is shown in parenthesis for each axis. The female sample shown at the bottom left of the plot was classified as an outlier and excluded from statistical analyses. **(b)** Volcano plot of the negative $\log_{10}$ p-value (y-axis) plotted against $\log_2$-fold change (x-axis) showing the results of differential expression analysis comparing male to female trees assayed using whole-genome Agilent oligonucleotide microarrays. Technical noise was accounted for in the statistical model by including factors for slide and sub-array within slide and the effect of sex was tested after removal of variance due to those technical effects. Non-significant genes are coloured to indicate density, which is shaded from yellow (high) to blue (low). No genes were significant (note that 0.01 on the y-axis corresponds to a p-value of 0.977).

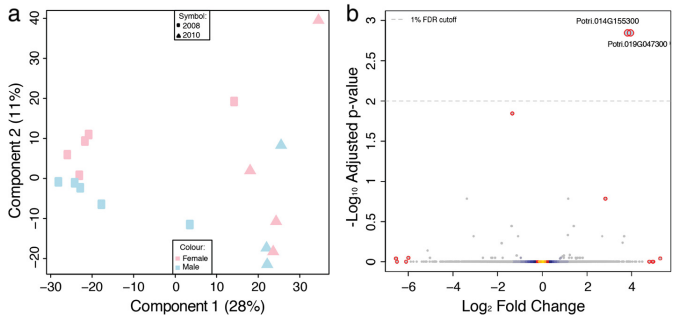

Figure 4 Overview of RNA-Sequencing gene expression patterns in male and female *Populus tremula* trees from the Umeå Aspen collection. (a) Principal Component Analysis plot of the RNA-Sequencing (RNA-Seq) expression data with samples classified by sex (male in blue, female in pink) and by year of sampling (2008 as squares and 2010 as triangles). The percentage variance explained by each component is shown in parenthesis for each axis. **(b)** Volcano plot of the negative $\log_{10}$ p-value (y-axis) (*i.e.* the log odds ratio) plotted against $\log_2$-fold change (x-axis) showing the results of differential expression analysis assayed using RNA-Seq comparing male to female trees. The statistical model included factors for year of sampling and sex and the effect of sex was tested after removal of the year effect. Significant genes are shown in blue where expression was higher in males. For the two significant genes at a 1% False Discover Rate (FDR) cut-off, the obtained p-value was $<1e^{-10}$ and was therefore set to 0. As a result the log odds value is infinite and was therefore replaced with the next largest log odds +1. Non-significant genes are coloured to indicate density, which is shaded from yellow (high) to blue (low). The dashed horizontal line represents a 1% FDR. The four genes with the smallest p-values (regardless of significance) and the four genes with the highest and lowest non-significant fold change values are circled in red. These genes are represented in Additional file 2. The gene identifiers for the two statistically significant genes are shown (identifiers refer to V3 of the *P. trichocarpa* genome).

clear differences between samples from the two years (Figure 4a). A total of 1,138 genes were identified as significantly differentially expressed between years (Figure 5).

Despite many genes having relatively high mean fold-changes between sexes in the RNA-Seq data (Figure 4b), the within-sex variation for those genes was high resulting in non-significant statistical test results. To further explore this, we examined the variance among samples for the four genes with the lowest and highest fold change values and for the four genes with the smallest p values regardless of fold change (of which only two were statistically significant) in the RNA-Seq data. Variance for genes with high between-sex fold-change values was high (Additional file 2) and only two genes (see below) were statistically significantly differentially expressed between males and females.

We applied a machine learning approach, support vector machines (SVMs), to sliding windows of contiguous genes in the *P trichocarpa* genome to identify any regions where the combination of expression patterns for all genes within the window were predictive of sex. No statistically significant gene combinations that were predictive of sex were identified.

Potri.019G047300 is not present in females and is located in the sex determination locus

In contrast to the clear influence resulting from year of sampling, differential expression analysis identified only

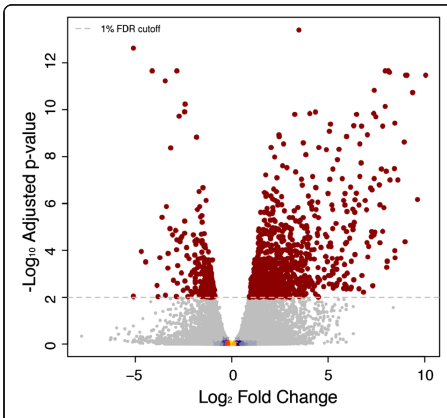

Figure 5 Differential expression of year of sampling effect. Volcano plot of the negative $\log_{10}$ p-value (y-axis) - *i.e.* log odds ratio - plotted against $\log_2$-fold change (x-axis) showing the results of differential expression analysis assayed using RNA-Sequencing comparing samples collected from trees in 2008 and 2010. Gene expression was assayed using samples collected from trees in 2008 and 2010 by RNA-Sequencing. Significant genes are shown as larger brown points. Non-significant genes are coloured to indicate density, which is shaded from yellow (high) to blue (low). The dashed horizontal line represents a 1% False Discovery Rate.

two statistically significant sexually dimorphic differences in the RNA-Seq dataset (Figure 4b; Potri.014G155300, FDR adjusted p-value 0.00: Potri.019G047300, FDR adjusted p-value 0.00) and none in the microarray dataset (Figure 3b). These two genes were not represented in the v1.1 genome annotation that was used for the array design, therefore excluding the possibility to cross-validate the result in the microarray dataset. However, Pakull *et al.* [28] provide an excellent and completely independent confirmation of this finding for Potri.019G047300.

Potri.014G155300 has no functional annotation but contains Pfam (Protein family) domains associated with cellulose synthase activity. This gene has highest sequence similarity to the *Arabidopsis thaliana* homolog AT2G32540, which is annotated as "Cellulose synthase-like B4". More interestingly, the second gene (Potri.019G047300) is one of seven candidate genes identified within the sex determination locus of *P. tremuloides* by Kersten *et al.* [24] and was recently shown by the same authors to have a partial or complete deletion in female aspens [28] resulting in expression only being observed in males. The gene has no current functional description in poplar but contains WD40 domains and shows highest sequence similarity based homology to the *A. thaliana* gene AT5G16750 (TORMOZEMBRYO DEFECTIVE, TOZ). In *A. thaliana* this gene is required for regulated division planes and embryo development [35] and is thought to be involved in 18S rRNA biogenesis and RNA methylation. We examined the expression of the seven candidates highlighted by Kersten *et al.* [24] within our data, revealing that this was the only gene displaying any evidence of differential expression between sexes (Figure 6). The gene was expressed more highly in male than female trees. Examination of Affymetrix gene expression microarray data represented at the poplar eFP resource (http://bar.utoronto.ca/efppop/cgi-bin/efpWeb.cgi; [36]) shows that this gene has high expression in male catkins and low expression in female catkins for the three array probes representing this gene (PtpAffx.113801.1.S1_s_at, PtpAffx.212175.1.S1_at; probe-to-gene links were obtained from PopArray [37], http://aspendb.uga.edu/). However, as these data represent expression in *P. balsamifera* and as this gene is not deleted in female *P. trichocarpa* trees (as suggested by the presence of the complete gene structure in the assembled genome sequence) these results require caution for extrapolation to the aspens.

We used genomic re-sequencing data (collected for another study, but available on request) from two of the assayed trees, one male and one female, to further explore this locus. Genomic DNA sequencing reads (2x100 bp paired-end reads generated from a 300 bp insert library and sequenced using standard procedures on the Illumina HiSeq 2000 platform) were aligned to the reference *P. trichocarpa* genome sequence and only uniquely

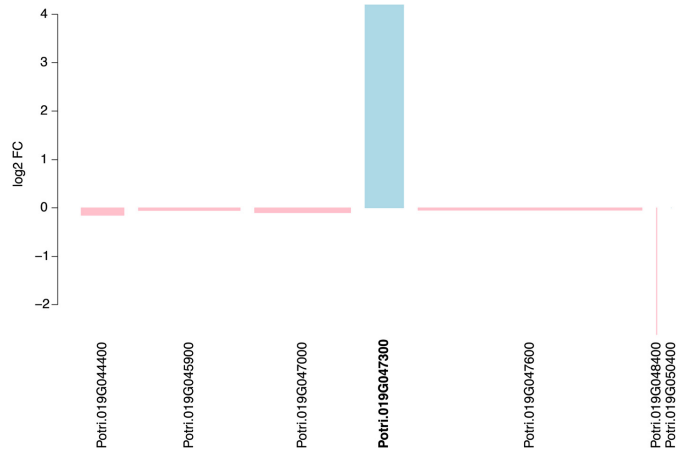

Figure 6 Bar graph showing differential expression of seven candidate genes from the sex determination locus of *Populus tremuloides* [24]. The width of each bar indicates the mean expression level of each gene relative to the gene with the highest expression (Potri.019G047600), which had an expression value of 240, such that narrow bars represent low expression and wide bars high expression. The y-axis indicates the $\log_2$ fold-change between male and female trees. Expression values represent variance stabilising transformation normalised read counts. Genes displaying higher expression in females are shown in pink and those with higher expression in males in blue. The gene model identifier for the only one of these genes that was identified as significantly differentially expressed in the RNA-Seq data is marked in bold text.

mapping reads were considered. This revealed that there is a deletion of this region in the female individual (Figure 7), which is in agreement with the results recently reported by Pakull *et al.* [28] and that explains the lack of any RNA-Seq reads being produced from female individuals in this region. As such females appear to be homozygous for absence of this locus. Corresponding plots based on RNA-Seq reads from all individuals assayed are available in Additional file 3. A single female individual (226.1) showed expression of the TOZ gene. We have been unable to confirm the sex of this tree as it has not flowered again since sex was originally determined. Repeating the above analyses with or without this individual did not affect the results obtained (see the R analysis HTML report on the PopGenIE FTP site [38]).

No evidence of biased sex ratio in P. tremula

We observed no sex bias in the *P. tremula* collections studied. The sex ratio of the SwAsp samples was 1:1 (female:male, where 52 trees of a total 116 in the collection are of known sex, Additional file 4). In the UmAsp samples the sex ratio was 1:1.1 (where 42 trees of 350 are of known sex, Additional file 4).

Discussion

In dioecious species, evolutionary theory suggests that males and females may have contrasting optimal strategies of resource investment to maximise reproductive success. As a result, natural selection would result in the emergence of sexual dimorphism in phenotypic, biochemical and ecological traits associated with contrasting resource allocation and utilisation as each sex evolves towards fitness optima. If phenotypic sexual dimorphism does arise, there will be concomitant dimorphism in gene expression patterns in the corresponding tissue(s) associated with those phenotypic traits. Such dimorphic gene expression patterns will be independent of any differential gene expression associated with sex determination and the control of reproductive tissue development. As such, although those genes may in some cases be located within the sex determination region or chromosome, it is likely that many such regulated genes will be autosomal.

In the current study our primary interest was to test the hypothesis that male and female *P. tremula* individuals invest resources differentially, resulting in sexual dimorphism. To this end a number of morphological and biochemical traits (Additional file 1) were selected to be

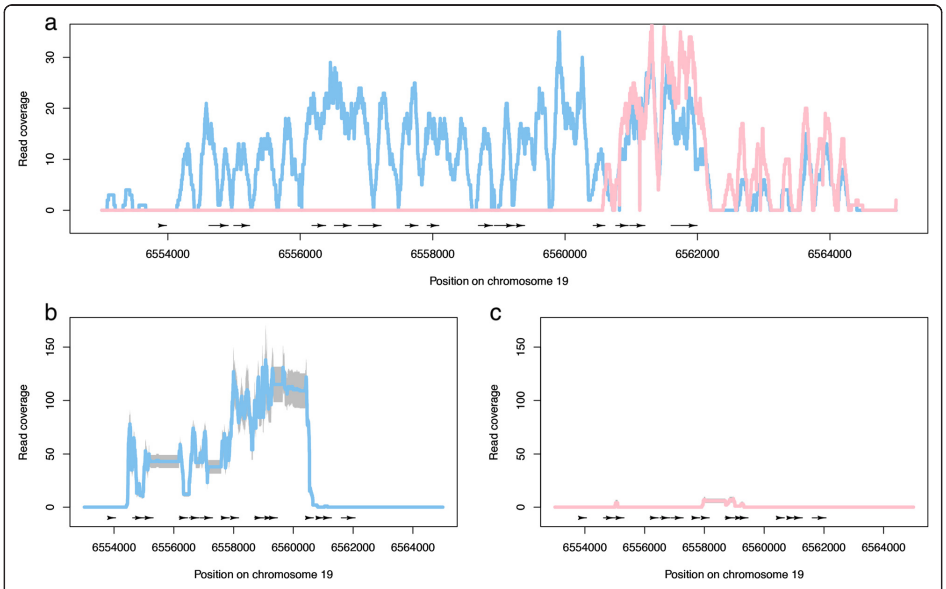

Figure 7 Genomic DNA and RNA-Sequencing read coverage (y axis) for the region of chromosome 19 (x axis) including Potri.019G047300. **(a)** Read coverage of uniquely mapping genomic DNA reads from a male (blue, 229.1) and female (pink, 349.2) individual. Black arrows represent exons with arrow direction indicating strand. **(b,c)** Read coverage of uniquely mapping RNA-Sequencing reads for male (b, n = 8) and female (c, n = 9) individuals. The coloured line represents the average per base pair read coverage across all individuals with grey indicating ± two standard deviations.

diagnostic of such dimorphism in leaves sampled from a set of wild-growing, sexually mature *P. tremula* individuals (the UmAsp collection) and a set of common-garden, sexually immature and clonally replicated individuals (the SwAsp collection, see materials and methods). We focused on leaves as these are the primary point of interaction between aspens and the majority of their associated herbivores as well as representing the site of energy assimilation and therefore carbohydrate production for utilisation in primary (growth-associated) and secondary metabolism.

P. tremula ***shows no phenotypic evidence of sexual dimorphism***

Height and diameter are often used as proxies for fitness based on the assumption that faster growing and larger individuals are better equipped to out-compete their neighbours, allowing greater resource acquisition that can be invested in sexual reproduction [39]. We found no statistical evidence supporting phenotypic differences between males and females for any of the phenotypic traits that we assayed in either the sexually mature UmAsp or sexually immature SwAsp samples. These results contrast with observations in Pauley [40] who reported a strong male biased sex ratio within a collection of superior-growth individuals of five North American *Populus* species. This was interpreted as potential evidence that males may display more vigorous growth. In *P. euphratica* growth traits showed variable differences between sexes among sample plots with no consistent statistically significant difference between sexes for assayed growth traits [41]. In the cross-species meta-analysis presented in Cornelissen & Stillin [10], males in general exhibited larger leaves, lower concentrations of secondary metabolites and higher growth rates. However, and in agreement with our results, there was no sexual dimorphism for height or nutrient concentrations. In *P. deltoides,* Farmer [42] observed that males were taller than females but did not have greater stem diameter. Citations within Farmer detail observations that the height of *P. tremula* x *P. tremuloides* seedling cohorts was correlated to the proportion of males, but also that no differences in vigour between sexes had been identified in *P. tremuloides*. Our results are also in agreement with those reported for *P. tremuloides* by Mitton & Grant [43] and Stevens & Esser [44]. Based on the current limited number of publications examining sexual dimorphism we would conclude that it is not yet possible to ascertain whether any generalisations can be formed regarding the presence or absence of sexual dimorphism for growth or defence related traits in *Populus*.

Several studies have additionally reported higher herbivore loads associated with increased growth in males [45-48], however we found no such reports in *Populus*. Although the meta-analysis presented in Cornelissen & Stiling [10] found that, in general, males suffered higher arthropod abundances, showed evidence of reduced levels of secondary metabolites and increased growth rates, it is not possible to extrapolate such generalised findings as being relevant to a specific species. Our own data identified no statistical evidence of sexual dimorphism in arthropod abundance, diversity or folivore herbivory damage in *P. tremula* in concordance with a lack of dimorphism in assayed growth and defence related phenotypes.

The majority of current evidence for sexual dimorphism in *Populus* has been identified in response to stressful environmental conditions, for example under drought, salinity [49-51], UV-B radiation [52], chilling stress [53], or differential nutrient availability [54-57] where females were found to be more sensitive. However, these studies typically used small sample sizes, in some cases being restricted to only a single individual of either sex. They also profile response to short term, acute stress exposure in most cases. This is in contrast to the approach taken here where we sample a collection of wild-growing trees. In these conditions individuals would have been exposed to various short to long-term stress events. We were interested to know whether evidence of dimorphism is present under such conditions in addition to knowing if there is evidence of sexual dimorphism for resource allocation to growth in sexually immature trees. In *Salix* it has been reported that evidence for sexual dimorphism varies through the growing season [11]. Such reports can lead to the general impression that sexual dimorphism is common or expected. However, bias against the publication of negative results potentially means that many such examples of a lack of dimorphism have remained unreported. The variable presence of evidence for sexual dimorphism also cautions against over-extrapolation of such results until multiple conditions and seasonal sampling points have been considered for each species and each geographic area of interest.

At both the national (SwAsp) and local (UmAsp) scales we believe that our sampling represents an unbiased representation of wild-growing mature trees, with sampling taking place with no knowledge of, or consideration for, sex or the presence of flowering. It is, of course, possible that studies testing more specific hypotheses, for example along an elevational cline (as reported for *Salix* [11]), may uncover evidence for shifting sex ratios or for sexual dimorphism. Indeed we see weak evidence for this within the SwAsp collection (Additional file 4) suggesting that further studies are needed in *P. tremula* before general conclusions can be drawn. We would caution against extrapolation of these findings beyond *P. tremula* growing in natural conditions within the geographic range covered by our sampling. To allow more general conclusions to be drawn for other *Populus* species, members of the Salicaeae and, more widely, other dioecious herbaceous species, will require equivalently detailed investigation and publication.

Environment affected gene expression more than sex

We profiled gene expression in leaves of sexually mature *P. tremula* individuals from the UmAsp collection to test the hypothesis that sexually dimorphic phenotypic traits would also be revealed by concomitant differential gene expression between males and females in non-reproductive tissues for genes associated with those phenotypes. In agreement with the above morphological and biochemical phenotypic results, we found no reliable evidence of large-scale sexually dimorphic (sex-biased) differential expression (Figures 3 and 4). In contrast, clear evidence of an effect of sampling collection was found (Figures 4a and 5). As samples from the two years were collected on different dates and from different heights within the canopy we cannot determine whether environmental/climatic variation between years or height in the canopy accounted for this difference. Significantly differentially expressed genes between the sample collections were over-represented for Gene Ontology (GO) biological process categories primarily involved in cellulose biosynthesis and glucan and lipid metabolism, most likely reflecting the slightly different sampling dates, with year-to-year variance in climatic conditions affecting the rate of leaf development and maturity. This exemplifies that in *P. tremula* leaves, changes in environmental conditions influence expression to a greater extent than the sex of an individual and that our expression data was of sufficient quality to identify biological effects influencing gene expression patterns.

The primary aim of this study was to identify patterns of sexually dimorphic gene expression associated with the morphological and biochemical traits profiled. As such, we would have expected relatively large numbers of genes to be involved should dimorphism have been present. For example, if females invest more resources into chemical defences produced via secondary metabolism, there would be corresponding sexually dimorphic differences in the expression of genes involved in secondary metabolism. Here we present gene expression results generated using *P. tremula* RNA-Seq read alignments to the *P. trichocarpa* reference genome. On the basis of a number of considered factors we do not believe that this biased our results: firstly, the vast majority - over 90% - of RNA-Seq reads aligned to the *P. trichocarpa* genome, suggesting that the two species have an almost entirely overlapping gene space and that sequence divergence within coding regions is not high enough to impact read alignment; secondly, we have also used a draft assembly of the *P. tremula* genome (available at the PopGenIE FTP resource [38]; ftp://popgenie.org/popgenie/UPSC_genomes/UPSC_Draft_Assemblies/Current/Genome/) to confirm that the vast majority of annotated CDS regions in *P. trichocarpa* can be aligned to the draft assembly and that analysis of the RNA-Seq data aligned to this draft genome does not produce different results; lastly, alignment of *P. tremuloides* and *P. tremula* x *P. tremuloides* genetic maps to the *P. trichocarpa* chromosomes suggests that there have been no major genome rearrangements between aspens and *P. trichocarpa* [24,27], although microsynteny has not been examined to date. As such, although there may be a small number of genes unique to, or highly variable between, each species, differences between the two species are not sufficient to affect the results of global-scale expression pattern analyses. We would caution that studies aiming specifically to identify the gene(s) underlying sex determination, where genetic mapping suggests a single locus is involved and for which a single or small number of genes are likely involved, could substantially benefit from use of species-specific genome sequences.

Potri.019G047300 *is absent in females and is located in the sex determination locus*

The proposed peritelomeric sex determination locus on chromosome 19 of *P. trichocarpa* represents a region of reduced recombination [23]. Kersten *et al*. [24] recently provided evidence of a similar region of reduced recombination in the pericentromeric sex-linked locus of chromosome 19 in *P. tremuloides*. One of the two genes that we identified as being highly, and exclusively, significantly differentially expressed between sexes in the RNA-Seq data (Potri.019G047300) is located in that identified sex determination locus of *P. tremuloides*. It is one of seven candidate genes identified by Kersten *et al.* [24] on the basis of Gene Ontology and other annotation evidence as having the potential to be involved in sex determination, primarily due to annotated involvement in floral organ development. This was the only one of those seven genes with evidence of differential expression between sexes in our data (Figure 6). Pakull *et al*. [28] recently refined this finding, reporting a complete or partial deletion of this gene in female *P. tremuloides* and *P. tremula* individuals. Here we present independent confirmation of this finding, supported by both genomic DNA and RNA-Seq results (Figure 7). It is unclear what the biological influence of differential expression of the gene in leaves might be. Our results clearly show that this single gene did not result in any larger-scale downstream patterns of sex-biased expression and examination of expression evidence at the PopGenIE [58] org and poplar eFP resources showed that expression of this gene varies between tissues and through the growth cycle, suggesting that expression is not merely constitutively fixed in males. This is certainly a finding that deserves future attention.

Due to reduced recombination rates in sex-determination loci, all genes within a locus will, on average, be co-inherited [22]. Such a case could be identifiable as a region of the genome where a contiguous set of genes would have

consistently sex-biased expression, resulting from either presence/absence differences for genes present only in the W-linked (or Y-linked) haplotype, or expression level differences for genes present in both haplotypes, but with fixed *cis*-acting differences between the Z and W (or X and Y) haplotypes. As the degree of expression bias may be small on a gene-by-gene basis, single gene analysis methods may lack the sensitivity to detect such differences but methods considering combinations of genes may succeed. For example, such a situation could have been possible for all seven of the candidate genes in the *P. tremuloides* sex determination locus discussed above. We therefore applied a machine learning approach to identify any sets of collinear genes (within sliding windows) that were predictive of sex. However, no statistically significant combinations of weakly predictive genes or synergistically predictive genes were identified.

Conclusions

We present an assessment of sex ratio and the lack of sexual dimorphism based on two independent samplings of Swedish *P. tremula*. Our sample of 87 was more comprehensive than almost all previous such assessments in *Populus* and, as such, we feel that the results obtained are an accurate representation for *P. tremula*. We identified no evidence that sex has served as a significant selective pressure affecting gross-scale morphological, biochemical or herbivorous insect interaction traits expected to be diagnostic of differential resource investment and allocation strategies. Correspondingly, there was no evidence for sex-biased patterns of gene expression associated with those, or any other, traits.

Although no evidence of large-scale patterns of sexually dimorphic gene expression patterns were identified, a previously identified candidate gene for sex determination in *P. tremuloides* [24] showed exclusive expression in males due to the homozygous absence of the locus in female individuals, an observation warranting future attention.

Methods

Phenotypic, morphological and biochemical traits

We examined the incidence of flowering in a collection of (sexually mature) wild, mature aspen (*Populus tremula* L.) trees in Sweden, the Umeå Aspen collection (UmAsp; [32]). In addition we used sexually immature clonal copies of trees of known sex from the Swedish Aspen (SwAsp) collection growing in a common garden experiment near Sävar, Umeå in Sweden, that were propagated and planted as described previously [33].

Umeå aspen collection (UmAsp)

Twenty-two trees bore male flowers and 20 trees bore female flowers in the spring of 2007. Tree sex was determined by visual examination of catkins and was confirmed by returning to each tree to record whether female trees retained catkins post pollination when male catkins had died. A description of trees and their geographic coordinates, together with sampling dates, is provided in Additional file 5. Tree height was measured in 2007 (when the collection was established) using a vertex dendrometer and trunk circumference was measured at breast height (1.3 m). Sampling took place on 22-25 June 2008. Six branches, each bearing approximately 60 leaves, were cut 4-5 m above ground level, in a transect from east to west across the canopy, or the nearest feasible positions, for morphological and herbivore community analyses. Sampled branches were sealed into plastic bags and kept at 4°C prior to morphological and arthropod analyses. A second sample of ten undamaged leaves from the west of the canopy was frozen in liquid nitrogen and stored at -80°C prior to RNA extraction. Following RNA extraction, samples were freeze-dried and used in assays of total phenolics and condensed tannins as described in [34] and leaf carbon and nitrogen against aspartame, wheat and atropine standards (Flash EA 1112 NC Soil Analyser, Thermo Fisher Scientific, Milan). Ten undamaged leaves were taken from each branch sample and scanned for image analysis conducted with LAMINA [59] to obtain leaf area. The same ten leaves were dried and weighed to calculate specific leaf area. From each sample bag, forty leaves were removed at random and examined for arthropod herbivore specimens and leaf modifications caused by known arthropods on aspen [34], from which arthropod species richness was calculated as the total number of morphospecies and arthropod abundance as the total number of individuals. Herbivores were also classified and summed by feeding guilds based on utilisation of the plant tissue: leaf-chewers, leaf-miners, gall-makers and leaf-rollers. Arthropod herbivore diversity was calculated for each genotype with the Shannon-Wiener index [60] using the diversity function from the package vegan [61] in R [62].

Swedish aspen collection (SwAsp)

In the Sävar common garden, clonal copies of 23 genotypes within the SwAsp collection originated from female and 24 from male trees. Sex was determined based on available flowering observations of the original trees from which the common garden trees were cloned. To determine sex, catkins were removed and examined using a binocular microscope. The cloned trees in this common garden experiment have not yet reached sexual maturity. Trees were measured annually in autumn for height with a measuring pole and for diameter at 30 cm from ground level using digital calipers. Growth rate over a five year period was calculated as (Height (2011) – Height (2006))/5, when the trees were between two and seven years old. Ten undamaged, mature

leaves were harvested for measurement of leaf area and specific leaf area (leaf area/dry mass), and a further ten leaves were harvested, dried and assayed for condensed tannins and total phenolics as described in [34]. Leaf nitrogen and carbon content were analysed on an available subset of six male and six female genotypes harvested on 29 June 2010. Counts of all arthropod herbivores on each tree in the SwAsp common garden were conducted on 27 – 29 June 2008 as described in [34]. Morphospecies of folivorous arthropods were summed from the replicates of each genotype. Arthropod species richness was calculated as the sum of arthropod morphospecies on each SwAsp genotype. Arthropod herbivore diversity (Shannon-Wiener index), abundance, species richness and feeding guild abundances were calculated on each genotype using the same methods as the UmAsp samples. Details of clone geographic origins, sex, replication in the common garden and phenotypic data collected are provided in Additional file 5.

Statistical analysis

Statistical analyses were conducted and figures generated in R conducted [R Core Development Team reference]. Statistical significance for all tests was determined at $\alpha \leq 0.05$. Dependent variables (tree phenotypes) were tested for normality and homogeneity of variance using Anderson-Darling and equal variance (Bartlett) tests to meet the assumptions of analysis of variance (ANOVA). Where transformation using Box-Cox powers or log-transformation did not result in improvement of the distribution of a dependent variable, a two-tailed Mann–Whitney *U*-test was applied. In SwAsp, the latitude of origin for each genotype was initially applied as a covariate, to account for phenotypic variation associated with latitude, however no significant effect of sex was identified for any response variable ($P > 0.1$), therefore final analyses were conducted without a covariate. ANOVA or Mann–Whitney U-tests tested the effect of sex (independent variable) on each phenotypic trait (response variable). To test for potential environmental influences partitioned by sex (independent variable) in UmAsp trees, the response variables latitude, longitude, and elevation were used in separate one-way ANOVAs but sex had no significant effect on the responses ($P > 0.5$ in all cases), therefore environmental factors were not considered necessary in analyses of phenotypic traits. In each of UmAsp and SwAsp, arthropod community composition was compared between male and female trees using non-parametric multivariate analysis of variance (npMANOVA; [63]). A Bray-Curtis dissimilarity matrix constructed from counts of arthropod herbivores on aspen genotypes (response variable) and tested for effects of tree sex (independent variable) using npMANOVA in the adonis function implemented in the R package vegan [61]. The p-value for significance was determined from 999 permutations of the data matrix.

Gene expression analysis

Sample collection for microarray and RNA-Seq analysis

Sample collection from the UmAsp trees is described above and sample details are given in Additional file 5. Briefly, ten mature leaves produced from pre-formed, overwintered buds were collected per tree, from ten male and ten female trees on June 29 2009 and used to perform whole genome oligonucleotide microarray hybridisations. For RNA-Seq analysis we used a combination of a set of samples that had been collected in 2008 (five male and five female individuals collected 22-25 June) and additional samples collected in 2010 (three male and four female individuals collected 11 August). All samples consist of pools of ten leaves collected from ten buds (one leaf per bud avoiding the first and last emergent leaf) collected by removing a length of branch from either the base of the tree canopy (2009 and 2010 samples) or from a branch at a height of 4-5 m (2008 samples).

RNA extraction

Total RNA was extracted from 0.5 g tissue using a modified version of the CTAB method [64] as described in [65]. Briefly, the ten sampled leaves were ground under liquid nitrogen using a pestle and mortar and 0.5 g of ground material was then used for RNA extraction. Precipitated RNA was further purified using an RNeasy Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. RNA concentration and purity were measured using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and integrity was analysed on an Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbronn, Germany). For each set of samples (*i.e.* all samples used for microarray or RNA-Seq analysis) all RNA extractions were performed together on the same day with the order of male and female samples randomised.

Microarray hybridisation and analysis

We used the Agilent v1.0 4x44k *Populus* gene expression oligonucleotide microarray (Agilent Technologies, Waldbronn, Germany), as detailed in the Gene Expression Omnibus platform ID GPL16040. We used the cDNA synthesis, amplification, microarray hybridisation and washing protocols supplied by Agilent (Agilent Technologies, Waldbronn, Germany) with no modifications. All hybridisations were performed using only one sample and using Cy3. Ten male and ten female individuals were profiled and the respective samples were randomised on arrays with two male and female samples run on each slide and with the position of males and females randomised between the four array sections per array slide. Arrays were scanned at 5 μm resolution, using a Scanarray 4000 microarray analysis system scanner (Perkin-Elmer, Boston, MA, USA). Spot data were extracted using

GenePix (v5, Axon Instruments Inc, Union City, CA, USA). Microarray normalisation and analyses were performed using the Bioconductor [66] limma package [67] in R [62]. Microarray annotations were obtained from the PopArray resource [37] and were based on V2 of the genome annotation. The microarrays were first background corrected using the normexp method implemented in the backgroundCorrect function. Then, a between microarray quantile normalisation was performed using the normalizeBetweenArrays function. A Principle Component Analysis (PCA) plot was used for quality control and this identified one sub array assaying a female individual as a clear outlier and this sample was therefore eliminated and not used for the statistical analyses. These were conducted by fitting a linear model taking into account batch effects for slide and position of sub-array within slide to the data in order to identify genes with a high probability of differential expression between sexes. FDR-adjusted P values were used to assess the significance of differential expression.

RNA sequencing and analysis

Total RNA preparations were sent to the Science for Life Laboratory (SciLifeLab, Stockholm, Sweden) for sequencing. Paired-end (2 × 100 bp) RNA-Seq data were generated using standard Illumina protocols and kits (TruSeq SBS KIT-HS v3, FC-401-3001; TruSeq PE Cluster Kit v3, PE-401-3001) and all sequencing was performed using the Illumina HiSeq 2000 platform. We generated data from 8 male individuals (five sampled in 2008 and three in 2010) and 9 female individuals (five sampled in 2008 and four in 2010). For sequencing, samples were recoded (from 1-17) with males and females randomised to avoid bias due to sample handling order. Samples were multiplexed by the addition of a unique barcode sequence and all samples were profiled on two lanes of the same flowcell with male and female samples and samples from 2008 and 2010 randomised between the two lanes. Briefly, the sequencing protocol involved DNase 1 digestion of total RNA, mRNA isolation by use of oligo(dT) beads, mRNA fragmentation, first and second strand cDNA synthesis, end-repair, A-tailing, bar-coded adapter ligation and PCR amplification. Sequencing libraries were quality checked using an Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbronn, Germany) before sequencing. The quality of the raw sequence data was assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Data were then filtered to remove adapters and trimmed for quality using Trimmomatic (v0.32; [68]; settings TruSeq3-PE-2.fa:2:30:10 LEADING:3 SLIDINGWINDOW:5:20 MINLEN:50). Residual ribosomal RNA (rRNA) contamination was assessed and filtered using SortMeRNA (v1.9; [69]; settings -n 6 -a 8 -v) using the rRNA sequences provided with SortMeRNA (rfam-5 s-database-id98.fasta, rfam-5.8 s-database-id98.fasta, silva-bac-16 s-database-id85.fasta, silva-euk-18 s-database-id95.fasta, silva-bac-23 s-database-id98.fasta and silva-euk-28 s-database-id98.fasta). After both filtering steps, FastQC was run again to ensure that no technical artefacts were introduced. Filtered reads were aligned to v3.0 of the *P. trichocarpa* genome (retrieved from the Phytozome [70] resource) using STAR (v2.3.1e [71]; non default settings: –OutQSconversion -31 –outReadsUnmapped Fastx –alignIntronMax 11000). The annotations obtained from the *P. trichocarpa* v3.0 GFF file were modified to generate 'synthetic' gene models; *i.e.* for each gene a non-redundant set of all exons from all transcripts was defined, with overlapping exons merged where necessary. This gene-model GFF file and the OSA read alignments were used as input to the HTSeq (http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html) htseq-count python utility to calculate exon-based read count values. The htseq-count utility takes only uniquely mapping reads into account. Statistical analysis of single-gene differential expression between sexes was performed in R (v3.1.0 [62]) using the Bioconductor (v2.14 [66]) DESeq and DESeq2 packages (v1.16.0 [72] and v1.4.5 [73]). For the DESeq/DESeq2 analyses, a two-factor linear model was fitted with the factors Sex and Year where Year was included as a blocking factor and the effect of Sex was tested after removal of the Year effect. FDR adjusted p-values were used to assess significance. The normalised read counts obtained from DESeq2 were used for all subsequent expression analyses, *e.g.* PCA, which were performed in R, with the exception of the differential gene expression analyses, which were performed using DESeq as it has been shown to be the most conservative of the currently available methods with the lowest false discovery rate [74]. An overview of the data, including raw and post-QC read counts and alignment rates is given in Additional file 6.

We analysed the RNA-Seq dataset using read alignments to both v2.0 and v3.0 of the *P. trichocarpa* genome assembly and annotation, yielding similar results in both cases. Similarly we analysed the microarray dataset using probe annotations based on v1.0 and v2.0 of the genome and assembly with similar gene-level results in both cases. We have also analysed the microarray data at the probe level, again yielding similar results.

Support vector machine identification of sex-predictive gene combinations

We used both the microarray data and normalised RNA-Seq expression values to test for the presence of contiguous gene combinations (*i.e.* windows of genes located next to each other within the genome) that were predictive of sex. We applied a sliding window across the genome with a window size of 10 genes (other window sizes were also tested with similar results). In total our expression data included 30,709 and 20,557 genes in

the RNA-Seq and microarray datasets, respectively. The criterion for accepting a gene inside a window was that it had at least 5 samples with non-zero expression values. Furthermore, only windows with at least 4 accepted genes were included. The Python module scikit-learn [75] was used to train SVMs with a radial basis function (RBF) kernel parameterised by C and γ. This approach has previously been shown effective on gene expression data [76]. Since the optimal values of these parameters are not known prior to training, a grid search was performed in a parameter space consisting of $\gamma = \{10^{-4}, 5 \cdot 10^{-4}, 10^{-3}, 5 \cdot 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ *and* $C = \{1, 10, 10^3, 5 \cdot 10^3, 10^4, 5 \cdot 10^4, 10^5\}$. For each genomic window, a double cross validation (CV) was performed where the outer CV was a leave-one-out and the inner was a 2-fold CV. The inner CV was used to train the SVM (*i.e.* estimate the parameters), and parameters with the smallest prediction error were used to predict the test data from the outer CV. The error rate was measured as the fraction of incorrect sex predictions. To validate the error rates, a permutation test was performed where 10,000 random genomic windows from all scaffolds were used in the same machine learning approach, but where the sex assignments were shuffled.

Availability of supporting information

Microarray data has been deposited to the Gene Expression Omnibus (GEO) under the accession ID GSE46219. Raw RNA-Seq data has been deposited to the European Nucleotide Archive (ENA) under the accession ID ERP002471.

Raw RNA-Seq fastq, the synthetic exon GFF3 file used for read alignment and HTSeq analysis, read alignment BAM files and other associated outputs from the gene expression analysis can be downloaded from the PopGenIE (*Populus* Genome Integrative Explorer; [58]) FTP resource [38]). The FTP site includes RData files for both gene expression datasets as well as an HTML transcript of the analyses performed, which we highly encourage readers to examine as all analysis details are included in addition to a number of summary plots exploring the dataset. To facilitate future meta-analyses, all phenotype data used in this study is also available at the FTP site. The data pre-processing source code is available through our public git repository accessible at https://bioinformatics.upsc.se. The RNA-Seq expression data presented here has been integrated in the exImage and exPlot expression visualisation tools at PopGenIE.org [58], where they are called the "Expression diversity (RNASeq)" dataset.

Additional files

Additional file 1: Statistical analyses of phenotypic and biochemical traits in the UmAsp and SwAsp samples. Phenotypic trait means and standard deviations for female and male individuals from the UmAsp collection (sheet1) and the SwAsp collection (sheet2), with results of one-way ANalyses Of VAriance (ANOVA). Where data could not be transformed to meet the assumptions of variance structure for ANOVA, a Mann–Whitney *U* test was conducted. For the extended phenotype of the arthropod community, non-parametric MANOVA (npMANOVA) results are shown for each of the UmAsp collection (sheet 1) and SwAsp collection (sheet 2).

Additional file 2: PDF image containing dot plot representations of per-sample expression values of the four genes with the smallest p values (regardless of significance) when testing for the effect of sex (top row), the four genes with the highest fold-change between males and females (middle row) and the lowest fold change (bottom row). Bold text gene identifiers in the top row of plots indicate the two statistically significant genes. The genes represented in these figures are those circled in red in Figure 4b. Expression values represent variance stabilising transformation normalised read counts derived using HTSeq and DESeq2. Black lines represent the median expression value per sex. For each gene male and female samples are plotted separately with males represented by blue dots and females by pink dots. The position along the x-axis of the plot has no meaning and merely separates male from female samples. Note that the y-axis is a log scale, for which a pseudo count was added to every value to avoid infinite values from the log transformation.

Additional file 3: PDF file containing individual plots of per base pair read coverage for reads aligning uniquely to the Potri.019G047300 locus.

Additional file 4: PDF file containing plots further exploring sex ratio of individuals and populations in the Swedish Aspen collection in relation to elevation, latitude and marker based population structure.

Additional file 5: Sex, longitude and latitude of clone origin for UmAsp and SwAsp, and elevation for UmAsp, samples used in the current study. The year of sampling for phenotype, microarray and RNA-Seq analysis is indicated. For UmAsp trees the longitude, latitude and elevation values represent the location of the actual tree sampled. For SwAsp samples they represent the origin of the original clone that was used to establish the clonal common garden experiment at the Skogforsk research station, Sävar, near Umeå, (63.896054°N, 20.549321°E). All UmAsp clones flowered in 2007. For the SwAsp samples, the number of clonal replicates present in the common garden is shown.

Additional file 6: Overview of RNA-Seq data including quality control metrics and correspondence between sample and ENA submission IDs.

Abbreviations

cDNA: Complementary DNA; CDS: Coding DNA sequence; DNA: Deoxyribonucleic acid; ENA: European nucleotide archive; FTP: File transfer protocol; GEO: Gene expression omnibus; GFF: General feature format; GO: Gene ontology; PCA: Principal component analysis; QA: Quality assessment; rRNA: Ribosomal RNA; RNA: Seq – RNA-sequencing; RNA: Ribonucleic acid; SVM: Support vector machines; SwAsp: Swedish aspen; UmAsp: Umeå aspen; VST: Variance stabilising transformation.

Competing interests

All authors declare that they have no competing interests.

Authors' contributions

KMR and NRS collected all leaf samples. KMR performed all morphological, biochemical and herbivore analyses. NRS performed all RNA extractions and microarray hybridisations. ND and NM performed the RNASeq and microarray expression analyses. NM, JÖ and TRH performed the machine learning analyses. BS performed the analysis of read alignments for the TOZ gene. SJ, PI and BA supervised and designed the project, which was originally conceived by SJ. NRS, KMR and ND prepared the manuscript with assistance from all authors. All authors approved the final manuscript.

Acknowledgements

We thank Yvan Fracheboud for use of data on flowering in the UmAsp collection, Agneta Olsson for assistance in collecting samples used for the

arthropod analysis. This work was supported by funds from the Swedish Research Council (VR), the Swedish Governmental Agency for Innovation Systems (VINNOVA), The Swedish Research Council (FORMAS) and, in parts, through the UPSC Berzelii Centre for Forest Biotechnology. NRS is supported by the Trees and Crops for the Future (TC4F) project.

Author details
[1]Department of Plant Physiology, Umeå Plant Science Centre, Umeå University, 901 87 Umeå, Sweden. [2]Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, 1432 Ås, Norway. [3]Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, DK 1871 Frederiksberg C, Denmark. [4]Department of Ecology and Environmental Science, Umeå Plant Science Centre, Umeå University, 901 87 Umeå, Sweden.

Received: 28 June 2014 Accepted: 6 October 2014
Published online: 16 October 2014

References

1. Parsch J, Ellegren H: **The evolutionary causes and consequences of sex-biased gene expression.** *Nat Rev Genet* 2013, **14:**83–87.
2. Ainsworth C: **Boys and girls come out to play: the molecular biology of dioecious plants.** *Ann Bot* 2000, **86:**211–221.
3. Heslop-Harrison JSP, Schwarzacher T: **Organisation of the plant genome in chromosomes.** *Plant J* 2011, **66:**18–33.
4. Charlesworth D: **Plant sex chromosome evolution.** *J Exp Bot* 2013, **64:**405–420.
5. Williams T, Carroll S: **Genetic and molecular insights into the development and evolution of sexual dimorphism.** *Nat Rev Genet* 2009, **10:**797–804.
6. Diggle PK, Di Stilio VS, Gschwend AR, Golenberg EM, Moore RC, Russell JRW, Sinclair JP: **Multiple developmental processes underlie sex differentiation in angiosperms.** *Trends Genet* 2011, **27:**368–376.
7. Obeso J: **The costs of reproduction in plants.** *New Phytol* 2002, **155:**321–348.
8. Shine R: **Ecological causes for the evolution of sexual dimorphism: a review of the evidence.** *Q Rev Biol* 1989, **64:**419–461.
9. Lloyd D, Webb CJ: **Secondary sex characters in plants.** *Bot Rev* 1977, **43:**177–216.
10. Cornelissen T, Stiling P: **Sex-biased herbivory: a meta-analysis of the effects of gender on plant-herbivore interactions.** *Oikos* 2005, **111:**488–500.
11. Dudley LS: **Ecological correlates of secondary sexual dimorphism in Salix glauca (Salicaceae).** *Am J Bot* 2006, **93:**1775–1783.
12. Jansson S, Douglas CJ: **Populus: a model system for plant biology.** *Annu Rev Plant Biol* 2007, **58:**435–458.
13. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen G-L, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, *et al*: **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray).** *Science* 2006, **313:**1596–1604.
14. Wullschleger SD, Weston DJ, DiFazio SP, Tuskan GA: **Revisiting the sequencing of the first tree genome: populus trichocarpa.** *Tree Physiol* 2013, **33:**357–364.
15. Bradshaw HD, Ceulemans R, Davis J, Stettler R: **Emerging model systems in plant biology: poplar (Populus) as a model forest tree.** *J Plant Growth Regul* 2000, **19:**306–313.
16. Pulford ID, Watson C: **Phytoremediation of heavy metal-contaminated land by trees–a review.** *Environ Int* 2003, **29:**529–540.
17. Whitham TG, DiFazio SP, Schweitzer JA, Shuster SM, Allan GJ, Bailey JK, Woolbright SA: **Extending genomics to natural communities and ecosystems.** *Science (80-)* 2008, **320:**492–495.
18. Latva-Karjanmaa T, Suvanto L, Leinonen K, Rita H: **Sexual reproduction of european aspen (Populus tremula L.) at prescribed burned site: the effects of moisture conditions.** *New For* 2006, **31:**14.
19. Boes TK, Strauss SH: **Floral phenology and morphology of black cottonwood, Populus trichocarpa (Salicaceae).** *Am J Bot* 1994, **81:**562–567.
20. Lester DT: **Variation in sex expression in Populus tremuloides Michx.** *Silvae Genet* 1963, **12:**141–151.
21. Rowland DL, Garner ER, Jespersen M: **A rare occurrence of seed formation on male branches of the dioecious tree, populus deltoides.** *Am Midl Nat* 2002, **147:**185–187.
22. Tuskan GA, DiFazio S, Faivre-Rampant P, Gaudet M, Harfouche A, Jorge V, Labbé JL, Ranjan P, Sabatti M, Slavov G, Street N, Tschaplinski TJ, Yin T: **The obscure events contributing to the evolution of an incipient sex chromosome in Populus: a retrospective working hypothesis.** *Tree Genet Genomes* 2012, **8:**559–571.
23. Yin T, DiFazio SP, Gunter LE, Zhang X, Sewell MM, Woolbright SA, Allan GJ, Kelleher CT, Douglas CJ, Wang M, Tuskan GA: **Genome structure and emerging evidence of an incipient sex chromosome in Populus.** *Genome Res* 2008, **18:**422–430.
24. Kersten B, Pakull B, Groppe K, Lueneburg J, Fladung M: **The sex-linked region in Populus tremuloides Turesson 141 corresponds to a pericentromeric region of about two million base pairs on P. trichocarpa chromosome 19.** *Plant Biol (Stuttg)* 2014, **16:**411–418.
25. Paolucci I, Gaudet M, Jorge V, Beritognolo I, Terzoli S, Kuzminsky E, Muleo R, Scarascia Mugnozza G, Sabatti M: **Genetic linkage maps of Populus alba L. and comparative mapping analysis of sex determination across Populus species.** *Tree Genet Genomes* 2010, **6:**863–875.
26. Pakull B, Groppe K, Meyer M, Markussen T, Fladung M: **Genetic linkage mapping in aspen (Populus tremula L. and Populus tremuloides Michx.).** *Tree Genet Genomes* 2009, **5:**505–515.
27. Pakull B, Groppe K, Mecucci F, Gaudet M, Sabatti M, Fladung M: **Genetic mapping of linkage group XIX and identification of sex-linked SSR markers in a Populus tremula × Populus tremuloides cross.** *Can J For Res* 2011, **41:**245–253.
28. Pakull B, Kersten B, Lüneburg J, Fladung M: **A simple PCR-based marker to determine sex in aspen.** *Plant Biol (Stuttg)* 2014, doi:10.1111/plb.12217.
29. Zluvova J, Zak J, Janousek B, Vyskot B: **Dioecious Silene latifolia plants show sexual dimorphism in the vegetative stage.** *BMC Plant Biol* 2010, **10:**208.
30. Osier T, Lindroth R: **Effects of genotype, nutrient availability, and defoliation on aspen phytochemistry and insect performance.** *J Chem Ecol* 2001, **27:**1289–1313.
31. Boeckler A, Gershenzon J, Unsicker S: **Phenolic glycosides of the Salicaceae and their role as anti-herbivore defenses.** *Phytochemistry* 2011, **72:**1497–1509.
32. Fracheboud Y, Luquez V, Bjorken L, Sjodin A, Tuominen H, Jansson S: **The control of autumn senescence in European aspen.** *Plant Physiol* 2009, **149:**1982–1991.
33. Luquez V, Hall D, Albrectsen BR, Karlsson J, Ingvarsson P, Jansson S: **Natural phenological variation in aspen (Populus tremula): the SwAsp collection.** *Tree Genet Genomes* 2007, **4:**279–292.
34. Robinson K, Ingvarsson P, Jansson S, Albrectsen B: **Genetic variation in functional traits influences arthropod community xomposition in aspen (Populus tremula L.).** *PLoS One* 2012, **7:**e37679.
35. Griffith ME, Mayer U, Capron A, Ngo QA, Surendrarao A, McClinton R, Jürgens G, Sundaresan V: **The TORMOZ gene encodes a nucleolar protein required for regulated division planes and embryo development in Arabidopsis.** *Plant Cell* 2007, **19:**2246–2263.
36. Wilkins O, Nahal H, Foong J, Provart NJ, Campbell MM: **Expansion and diversification of the Populus R2R3-MYB family of transcription factors.** *Plant Physiol* 2009, **149:**981–993.
37. Tsai CJ, Ranjan P, DiFazio SP, Tuskan GA, Johnson VE, Joshi CP: **Poplar genome microarrays**. In *Genet Genomics Breed Poplar*. Edited by Joshi CP. CRC Press Boca Raton: Science Publishers, Inc; 2011:112–127.
38. *PopGenIE (Populus Genome Integrative Explorer) File Transfer Protocol site.* (ftp://popgenie.org/popgenie).
39. Price P: **The plant vigor hypothesis and herbivore attack.** *Oikos* 1991, **62:**244.
40. Pauley SS: **Sex and vigor in Populus.** *Science (80-)* 1948, **108:**302–303.
41. Petzold A, Pfeiffer T, Jansen F, Eusemann P, Schnittler M: **Sex ratios and clonal growth in dioecious Populus euphratica Oliv., Xinjiang Prov., Western China.** *Trees* 2012, **27:**729–744.
42. Farmer RE: **Sex ratio and sex-related characteristics in eastern cottonwood.** *Silvae Genet* 1964, **13:**116–118.
43. Mitton JB, Grant MC: **Observations on the ecology and evolution of quaking aspen, populus tremuloides, in the Colorado Front Range.** *Am J Bot* 1980, **67:**202.
44. Stevens M, Esser S: **Growth–defense tradeoffs differ by gender in dioecious trembling aspen (Populus tremuloides).** *Biochem Syst Ecol* 2009, **37:**567–573.
45. Jing S, Coley P: **Dioecy and herbivory: the effect of growth rate on plant defense in Acer Negundo.** *Oikos* 1990, **58:**369.

46. Boecklen WJ, Price PW, Mopper S: **Sex and drugs and herbivores: sex-biased herbivory in Arroyo Willow (Salix Lasiolepis).** *Ecology* 1990, **71:**581–588.
47. Hjältén J: **Plant sex and hare feeding preferences.** *Oecologia* 1992, **89:**253–256.
48. Boecklen W, Hoffman T: **Sex-biased herbivory in *Ephedra trifurca*: the importance of sex-by-environment interactions.** *Oecologia* 1993, **96:**49–55.
49. Jiang H, Peng S, Zhang S, Li X, Korpelainen H, Li C: **Transcriptional profiling analysis in Populus yunnanensis provides insights into molecular mechanisms of sexual differences in salinity tolerance.** *J Exp Bot* 2012, **63:**3709–3726.
50. Xu X, Yang F, Xiao X, Zhang S, Korpelainen H, Li C: **Sex-specific responses of Populus cathayana to drought and elevated temperatures.** *Plant Cell Environ* 2008, **31:**850–860.
51. Chen F, Chen L, Zhao H, Korpelainen H, Li C: **Sex-specific responses and tolerances of Populus cathayana to salinity.** *Physiol Plant* 2010, **140:**163–173.
52. Xu X, Zhao H, Zhang X, Hänninen H, Korpelainen H, Li C: **Different growth sensitivity to enhanced UV-B radiation between male and female Populus cathayana.** *Tree Physiol* 2010, **30:**1489–1498.
53. Zhang S, Jiang H, Peng S, Korpelainen H, Li C: **Sex-related differences in morphological, physiological, and ultrastructural responses of Populus cathayana to chilling.** *J Exp Bot* 2011, **62:**675–686.
54. Randriamanana TR, Nybakken L, Lavola A, Aphalo PJ, Nissinen K, Julkunen-Tiitto R: **Sex-related differences in growth and carbon allocation to defence in Populus tremula as explained by current plant defence theories.** *Tree Physiol* 2014, **34:**471–487.
55. Zhao H, Li Y, Zhang X, Korpelainen H, Li C: **Sex-related and stage-dependent source-to-sink transition in Populus cathayana grown at elevated CO(2) and elevated temperature.** *Tree Physiol* 2012, **32:**1325–1338.
56. Wang X, Curtis P: **Gender-specific responses of Populus tremuloides to atmospheric CO2 enrichment.** *New Phytol* 2001, **150:**675–684.
57. Li L, Zhang Y, Luo J, Korpelainen H, Li C: **Sex-specific responses of Populus yunnanensis exposed to elevated CO2 and salinity.** *Physiol Plant* 2013, **147:**477–488.
58. Sjödin A, Street NR, Sandberg G, Gustafsson P, Jansson S: **The populus genome integrative explorer (PopGenIE): a new resource for exploring the Populus genome.** *New Phytol* 2009, **182:**1013–1025.
59. Bylesjö M, Segura V, Soolanayakanahally RY, Rae AM, Trygg J, Gustafsson P, Jansson S, Street NR: **LAMINA: a tool for rapid quantification of leaf size and shape parameters.** *BMC Plant Biol* 2008, **8:**82.
60. Whittaker RH: **Evolution and measurement of species diversity.** *Taxon* 1972, **21:**213–251.
61. Dixon P: **VEGAN, a package of R functions for community ecology.** *J Veg Sci* 2003, **14:**927–930.
62. *R: A Language and Environment for Statistical Computing.* (http://www.r-project.org).
63. Anderson MJ: **A new method for non-parametric multivariate analysis of variance.** *Austral Ecol* 2001, **26:**32–46.
64. Chang S, Puryear J, Cairney J: **A simple and efficient method for isolating RNA from pine trees.** *Plant Mol Biol Report* 1993, **11:**113–116.
65. Street NR, Skogström O, Sjödin A, Tucker J, Rodríguez-Acosta M, Nilsson P, Jansson S, Taylor G: **The genetics and genomics of the drought response in Populus.** *Plant J* 2006, **48:**321–341.
66. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5:***R80.*
67. Smyth G: **limma: Linear Models for Microarray Data**. In *Bioinforma Comput Biol Solut Using R Bioconductor.* Edited by Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S. New York: Springer; 2005:397–420. Statistics for Biology and Health.
68. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics* 2014, **30:**btu170.
69. Kopylova E, Noé L, Touzet H: **SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data.** *Bioinformatics* 2012, **28:**3211–3217.
70. Goodstein D, Shu S, Howson R, Neupane R, Hayes R, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar D: **Phytozome: a comparative platform for green plant genomics.** *Nucleic Acids Res* 2012, **40**(Database issue):D1178–D1186.
71. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29:**15–21.
72. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11:**R106.
73. Love MI, Huber W, Anders S: *Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.* Cold Spring Harbor Labs Journals; 2014. bioRxiv 2014 doi:10.1101/002832.
74. Soneson C, Delorenzi M: **A comparison of methods for differential expression analysis of RNA-seq data.** *BMC Bioinformatics* 2013, **14:**91.
75. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E: **Scikit-learn: machine learning in Python.** *J Mach Learn Res* 2011, **12:**2825–2830.
76. Önskog J, Freyhult E, Landfors M, Rydén P, Hvidsten TR: **Classification of microarrays; synergistic effects between normalization, gene selection and machine learning.** *BMC Bioinformatics* 2011, **12:**390.

doi:10.1186/s12870-014-0276-5
Cite this article as: Robinson *et al.*: *Populus tremula* (European aspen) shows no evidence of sexual dimorphism. *BMC Plant Biology* 2014 **14**:276.

Paper III

The genetic architecture of gene expression natural variation in a forest tree suggests buffering of central genes

Niklas Mähler[1], Barbara K Terebieniec[2], Jing Wang[3], Pär K Ingvarsson[3], Nathaniel R Street[2,*] and Torgeir R Hvidsten[1,2,*]

[*]Equal contribution

[1]Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, 1430 Ås, Norway
[2]Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, 901 87 Umeå, Sweden
[3]Umeå Plant Science Centre, Department of Ecology and Environmental Science, Umeå University, 901 87 Umeå, Sweden

Abstract

Several eQTL studies in plant species, including forest tree, have investigated general properties of the genetic architecture of gene expression variation. Most of these studies used controlled crosses and it is unclear whether their findings extend to natural populations of unrelated individuals. Here we utilize RNA-Sequencing to assay gene expression in winter buds undergoing bud flush in a natural population of *Populus tremula*. Expression Quantitative Trait Locus (eQTL) mapping identified 164,290 significant eQTLs paring 6,241 unique genes (eGenes) with 147,419 unique SNPs (eSNPs). We found approximately four times as many local as distant eQTLs with local eQTLs having significantly higher effect size. eQTLs were primarily located in regulatory regions of genes (UTRs or flanking regions) regardless of whether they were local or distant, and whether the local eQTL was closest to the associated eGene or some other gene. We used the gene expression data to infer a co-expression network and utilized the eQTLs to explain the structure of the network. Although pairwise gene expression similarity from natural variation was expectedly lower than from tissue atlases or developmental gradients, the network displayed modularity and scale-freeness. Although we found eGenes in the core of 28 of 38 network modules, eGenes were generally underrepresented in cores, and overrepresented in the periphery of the network, with a negative correlation between effect size and network connectivity. We hypothesize that network modules are explained by a few central regulators under the control of eSNPs of low effect size, matching expectations that biological networks are buffered against single mutations inducing highly negative effects. As such, most eSNPs fine-tune the regulation of individual genes while a few eSNPs combinatorially control regulators that determine the modular structure of networks, thus explaining why eGenes generally have low connectivity.

Introduction

A central aim of biology is to understand how genomes encode emergent phenotypes and how genetic variation results in natural variation in phenotypes within populations. While the biological function of many genes can be, and has been, ascertained through use of mutant screens, such approaches tend to identify genes essential to a biological process. There is no clear consensus as to whether variation

in these genes underlies natural variation in phenotypic traits. Additionally, many traits are complex, being controlled by the action and interaction of numerous genes, often rendering mutational approaches inadequate or infeasible. In such cases turning to nature and employing population genetics and systems biology approaches to uncover genetic and genomic co-variation between assayed traits (genomic markers, gene expression, protein abundance, metabolites etc.) and phenotype represents an appealing alternative approach.

The availability of massively parallel sequencing technologies affords new possibilities for addressing biological questions, for example enabling the generation of *de novo* genome assemblies and of population-wide resequencing data that can be used to perform genome-wide association studies (GWAS), even in species with large genomes that harbour high levels of polymorphism or that display rapid linkage disequilibrium (LD) decay. The use of genome-wide resequencing data allows the discovery of, effectively, all genetic polymorphisms within an individual, which can then be used as markers for association mapping. These genetic markers, of which single nucleotide polymorphisms (SNPs) are currently the most commonly considered, can then be used to perform association or linkage mapping with the aim of identifying the subset of polymorphisms that contribute to the control of phenotypic variation among individuals. The majority of SNPs are expected to have no, or insignificant (*i.e.* are selectively neutral), consequence. However, an implication of the infinitesimal model [1] is the expectation that the control of variation for any given phenotype will be contributed to by a large number of loci, each of small effect size, although it is not clear how this holds across the scale of molecular to complex, integrative phenotypes (*e.g.* from transcription to morphology) or whether this pattern contrasts between adaptive and selectively neutral phenotypic trait variation.

Advances in sequencing technologies have concordantly revolutionised transcriptomics studies, especially so in non-model organisms. Following the seminal work of [2] and [3], numerous early studies in a range of species established that there is a significant heritable component underlying natural variation of gene expression levels among individuals within populations [4–16] and that natural variation in expression underlies a number of phenotypes [17–24]. Given these findings, it became apparent that gene expression values could be considered in the same way as any other quantitative phenotype and be subjected to linkage or association mapping to identify polymorphisms contributing to expression level variation among individuals [25], as first reported in [19], with the identified loci termed expression Quantitative Trait Locus (eQTL; [6]) or, less commonly, expression level polymorphisms (ELPs; [26]). eQTLs are classified as either local or distant acting depending on the physical location of the associated polymorphism in relation to the gene that the eQTL is mapped for: local eQTLs are usually defined as being located within a specified physical distance of the gene location (typically up to 2 Mbp, although this varies depending on species) while distant eQTLs represent polymorphisms that are located beyond that threshold distance or on another chromosome. eQTLs can further be classified as acting in *cis* or *trans*: *cis* eQTLs act in an allele specific manner and are usually considered to be local, although long-range *cis* interactions can occur, for example when a polymorphism is located in an enhancer that is physically distant from the gene of interest; *trans* acting eQTLs affect both alleles of a gene and are most commonly located distant to that gene. There continues to be strong interest in eQTLs as they offer potential functional links between phenotypes and underlying molecular mechanisms. Importantly, the majority of polymorphisms that have been associated to phenotypes using GWAS in a wide range of species are located outside of protein coding or transcribed regions [27–30], suggesting that they influence expression rather than altering protein or transcript function. eQTLs therefore have potential for explaining how such polymorphisms ultimately influence phenotype, their evolutionary signatures, as well as offering new insights into the nature of expression regulation [31,32].

There have been a number of previous eQTL studies conducted using plant species including *Arabidopsis thaliana* [33–38], maize [39–41] and rice [42,43], and in forest tree species [7,44–46]. A number of general observations have been made concerning the genetic architecture of gene expression variation using the expanding body of evidence from these and other eukaryotic systems, including that a greater number of local eQTLs are typically identified and that these individually explain a larger proportion of gene expression variance than do distant eQTLs [34,47–50]. However, with the exception of human studies, the majority of previous work was conducted using controlled, often interspecific, crosses and it is not clear how generally applicable the conclusions from these studies are for natural populations of unrelated individuals. Few studies have considered whether observed, heritable variation is adaptive [51,52] and there is a lack of consensus as to whether or not natural variation in gene expression is selectively neutral [52].

Species in the Populus family have been established as a powerful model system for forest tree genomics due to their relatively small genome, rapid growth, propensity for clonal propagation and ease of genetic transformation [53]. The genome of *P. trichocarpa* (black cottonwood) was the first tree to be sequenced [54] and, to date, has been used as a reference genome for studies of all members of the family. *P. tremula* (European aspen) has many features that render it a particularly useful model for population genetics and speciation studies [55,56]. To facilitate exploitation of *P. tremula* as a model system, we have produced a draft *de novo* assembly of the *P. tremula* genome (available at http://popgenie.org; [57]) and resequencing data (Wang *et al.* In prep.) for all individuals comprising the Swedish Aspen (SwAsp) collection [58]. Here, we utilise this resource in combination with population-wide RNA-Seq data assaying gene expression in winter buds undergoing bud flush. We used these expression data to perform eQTL mapping and to construct a co-expression network, with the results integrated to provide insight into the genetic architecture of natural variation in gene expression levels.

Results

Population level gene expression similarity

We utilised the northern common garden (located at 63.9° N, near Umeå, Sweden) of the Swedish Aspen (SwAsp) collection [58] which comprises 116 *Populus tremula* genotypes sampled across the species distribution range in Sweden (56.2° to 66.4° N, Figure 1A). We have previously shown that the SwAsp collection represents abundant genetic variation, that linkage disequilibrium (LD) is low [56,59] and that there is minimal population structure (Wang *et al.* In prep.) and have recently performed whole genome re-sequencing of the collection (Wang *et al.* In prep.). These resequencing data were aligned to a *de novo* assembly of the *P. tremula* genome (available at PopGenIE.org; [57]) to perform SNP calling and genotyping, from which 4,509,654 SNPs were identified after stringent filtering for use in downstream studies such as GWAS (Wang *et al.* In prep.).

To enable examination of the genetic architecture of natural variation in gene expression within the SwAsp collection, we generated RNA-Seq expression data from winter buds at the point of spring bud flush for 219 individuals (clonal replicates), representing 86 genotypes, and first examined the distribution of broad-sense heritability (H^2) and Q_{ST} (Figure 1B,C respectively) for all expressed

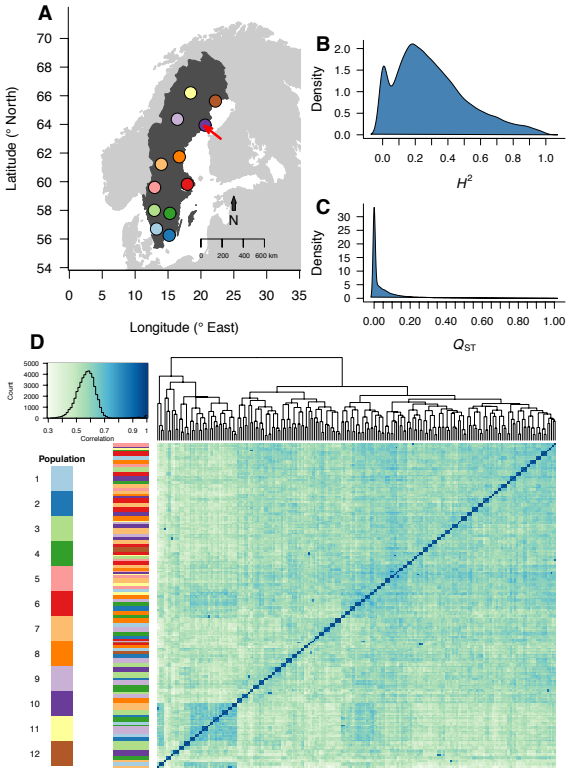

Figure 1. (A) Map of the original locations of the SwAsp populations. The red arrow points to the location of the common garden used in this study. (B) Distribution of gene expression heritability. (C) Distribution of gene expression Q_{ST}. (D) Sample clustering based on all samples, including biological replicates. The heatmap represents the sample correlation matrix based on the 500 genes with the highest expression variance. Darker colour indicates higher correlation. The coloured bar represent the populations the samples belong to. The small clusters on the diagonal correspond to biological replicates of each genotype.

annotated genes. H^2 ranged from 0.0 to 1.0 with a mean (± s.d) of 0.30 (0.22) and with 5,924 genes (17%) having $H^2 > 0.5$. There was a weak positive correlation with median expression (Pearson r = 0.09, $p < 2.2\times10^{-16}$) and a positive correlation to expression variance (Pearson r = 0.43, $p < 2.2\times10^{-16}$). Q_{ST} ranged from 0.0 to 1.0 with a mean (± s.d) of 0.06 (0.12) andhad a weak negative correlation with expression variance (Pearson r = -0.02, $p < 4.5\times10^{-4}$) and a positive correlation with median expression (Pearson r = 0.18, $p < 2.2\times10^{-16}$). To further examine whether population structure was apparent on the basis of expression variation among genotypes, we performed hierarchical clustering of all individuals (Figure 1A) or genotypes (Figure S1).

We selected the 500 genes with the highest H^2 (0.88-1.0, 0.93 ± 0.03) and Q_{ST} (0.54-1.0, 0.71 ± 0.13) and subjected these to Gene Ontology (GO) enrichment analysis to determine their biological relevance. Genes with high H^2 were overrepresented for categories including protein phosphorylation (GO:0006468), while high Q_{ST} genes were enriched in terms including translation (GO:0006412) and gene expression (GO:0010467). Likewise, we considered the 500 genes with the lowest values, which

identified no significant overrepresentation for low H^2 genes and overrepresentation of terms including amino acid activation (GO:0043038)) among the 11,895 genes with a Q_{ST} of zero.

We performed a regression analysis to ascertain whether a set of geographic (latitude, longitude, elevation), climatic (temperature, precipitation) or other factors (time since sample collection) significantly explained the global patterns of gene expression similarity among genotypes (Figure S2), as identified by performing a PCA of the expression data. None of the gene expression principal components (PCs) were significantly explained by environmental factors, with the only significant results found between PCs 2, 5 and 7 and the number of hours from collecting branches from the field until bud sampling for RNA extraction (see Materials and methods), which explained 6.6%, 3.2%, and 2.1% of expression variance, respectively.

We subsequently filtered expression values to remove unexpressed genes and uninformative expression profiles with low variance. Of 35,154 annotated genes, 20,835 were expressed in all samples, including biological replicates, with 23,183 genes expressed in all genotypes when considering genotype means. Filtering to remove uninformative expression retained 22,306 genes while 12,848 were removed, representing both genes that were not expressed in our bud samples (6,736 genes with median expression of zero of which 2,385 had no detectable expression at all), or that were weakly expressed (1,762 genes with variance < 0.05 and median expression < 2), together with genes that had stable expression among genotypes (4,350 genes with expression variance < 0.05 and median expression >= 2). The latter potentially represent genes with canalised gene expression. Analysis of this set of stably expressed genes identified enrichment for GO categories including protein transport (GO:0015031, p = 6.8e-11) and protein localisation (GO:0008104, $p = 2.2\times10^{-10}$). In contrast, the 500 genes with the highest variance were enriched for GO categories related to protein phosphorylation (GO:0006468, $p < 10^{-6}$), chitin metabolic process (GO:0006030, $p < 10^{-4}$), and cell wall macromolecule catabolic process (GO:0016998, $p < 10^{-4}$). Comparing the variance of these 500 genes with mean F_{ST} calculated using SNPs within those genes revealed no apparent relationship.

eQTL mapping

We performed eQTL mapping (after accounting for hidden factors and populations structure) using linear modelling as implemented in the R-package Matrix eQTL [60]. We define an eQTL as a significant association between a SNP (termed an eSNP) and the expression of a gene (termed an eGene). Furthermore, we classified an eQTL as *local* if the eSNP was located on the same chromosome and not more than 100 Kbp from the associated eGene, and as *distant* otherwise. We did not consider whether eQTLs acted in *cis* or *trans*.

In total 164,290 eQTLs were identified at 5% empirical FDR (Materials and methods): 131,103 local and 33,187 distant. These eQTLs represented pairwise relationships between 6,241 unique genes (eGenes) and 147,419 unique SNPs (eSNPs). The genomic context of eSNPs was determined by overlapping the eSNP positions with gene annotations. After normalizing for feature length, the majority of local eSNPs were located within untranslated regions (UTRs) and up- or down-stream (regulatory) regions of genes, with distinctly lower representation within exons than introns (Figure 2B). The genomic contexts of local and distant eSNPs were largely similar, although there were distinctly more eSNPs located within intergenic regions for distant eQTLs. Dividing the local eQTLs on the basis of whether they were located within or near the associated eGene or within or near any other gene inside of our local distance threshold revealed that approximately half were located near the eGene itself. There was a clear tendency for a local eSNP to be located proximal to the transcription

start site (TSS) or the stop codon (Figure S3). eGenes had significantly higher heritability (difference in median of 0.24, Mann-Whitney $p < 2.2\times10^{-16}$) than non eGenes (Figure 2D), with this trend being slightly higher for local than distant eQTLs (Figure S5). There were no significant differences in median expression between eGenes and non eGenes or for local or distant eGenes. eGenes with at least one local eQTL were enriched for GO categories related to tRNA metabolic process (GO:0006399, $p = 1.5\times10^{-5}$), ncRNA metabolic process (GO:0034660, $p = 2.6\times10^{-5}$) and organonitrogen compound biosynthetic process (GO:1901566, $p = 2.2\times10^{-5}$) while eGenes with at least one distant eQTL were enriched for protein phosphorylation (GO:0006468, $p = 0.0064$).

The vast majority of eSNPs were associated with a single eGene (132,258 eSNPs) with a maximum of six eGenes associated with a single eSNP (Figure S6A). In contrast only 1,248 of the 6,241 eGenes

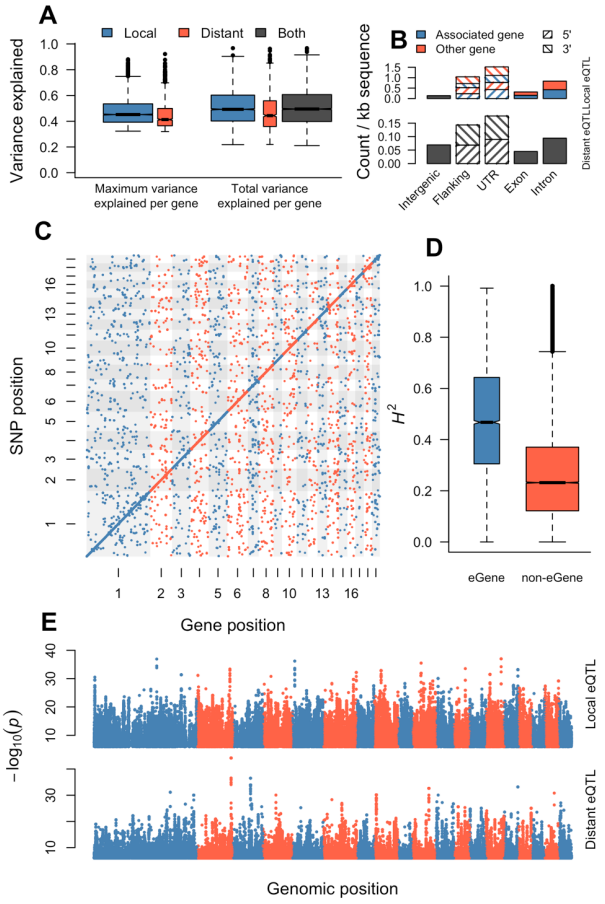

Figure 2. (A) Expression variance explained (R^2) for local and distant eQTLs. Box plots show the maximum variance explained by a single eQTL for each gene and the total variance explained by all eQTLs for each gene. The widths of the boxes are proportional to the number of genes represented. (B) Genomic context of local and distant eSNPs. When an eSNP overlapped with features on both strands, both of them were counted. For both local and distant eQTLs the features are based on the gene that is closest to the eSNP, and furthermore, for local eQTLs, the features are divided into whether the eSNP is located in or near the same gene that it is associated to or not. Flanking regions represent 2 kb upstream and downstream from the gene. (C) Scatter plot showing the positions of all significant eQTLs in this study. No evidence of eQTL hotspots can be observed. Numbers indicate chromosome. (D) Broad sense heritability distributions for eGenes and non-eGenes. (E) Manhattan plots for local eQTLs (upper) and distant eQTLs (lower).

were associated with a single eSNP, while there were eGenes associated with up to 1,547 eSNPs (Figure S6B). To partially account for linkage, we fitted linear models between the expression of each eGene and all the significant eSNPs for that gene, both local and distant. The use of a linear model masks eSNPs that contain identical/redundant information and thus effectively identifies haplotype blocks present in all individuals (which we refer to as 'unique eSNPs'), while also producing a measure of how well the combination of eSNPs explains the expression of the corresponding eGene (in terms of percentage variance explained, %VE). Of the 4,993 eGenes associated with more than one eSNP, 4,703 were also associated to more than one unique eSNP. The adjusted %VE for the combination of eQTLs was, in general, higher than for single eSNPs. Local eSNPs explained significantly more of the variance than did distant eSNPs (local mean adjusted %VE = 51, distant mean adjusted %VE = 47, Mann-Whitney $p < 2.2\times10^{-16}$, Figure 2A) and also had higher statistical significance (Figure 2E). There was a positive correlation between the maximum %VE of the eSNPs associated with an eGene and gene expression H^2 (Pearson r = 0.47, $p < 2.2\times10^{-16}$).

To detect possible hotspots, we plotted eSNP positions against the genomic positions of the associated genes (Figure 2C), as well as the number of eSNPs and eGenes for 100 kb genomic windows along the chromosome (Figure S7). We did not identify any clear hotspots in our data.

Co-expression network

We used our genotype mean gene expression values to calculate a co-expression network and subsequently considered the network characteristics for genes with and without a mapped eQTL (eGenes). The network was constructed using the R-package WGCNA (see Materials and methods), and, similar to other biological networks, it had a good scale-free fit ($R^2 = 0.97$). One notable feature of the network compared to the type of networks considered by systems biology analyses, often inferred from different tissues or perturbations, was that the distribution of pairwise gene expression correlations was relatively narrow. We compared our network with the *P. tremula* expression atlas (exAtlas; [57]), which represents different tissues collected from a single genotype, and observed that the correlation distribution for the exAtlas samples was much wider than that of our population expression data (Kolmogorov-Smirnov D = 0.14, $p < 2.2\times10^{-16}$; Figure 3A).

Clustering analysis of the co-expression network identified 38 co-expression modules (two examples are shown in Figure 3C). These were enriched for a number of different Gene Ontology (GO) categories including translation (modules 9, 10, and 14), photosynthesis (module 22) and oxidation-reduction process (module 29; for all results see supplementary file 1). Despite the narrow distribution of correlation values, the modules were reasonably well defined, as shown by examination of the normalized connectivity difference (K_{diff}), *i.e.* the difference between intra- and inter-modular connectivity (see Materials and methods). All modules exhibited a positive mean K_{diff}, with only 157 genes (0.7%) having a negative K_{diff}. This was in stark contrast to genes assigned to the 'junk' module (*i.e.* all genes not assigned to any well-defined module), where there were 480 genes with negative K_{diff} (29%).

eGenes are under-represented in network module cores

To test whether eQTLs explained the structure of the co-expression network we examined the relationship of eGenes to network connectivity. In general, eGenes had lower connectivity and betweenness centrality than non-eGenes (Mann-Whitney $p < 2.2\times10^{-16}$ for both, Figure 3B). Moreover, genes with a positive K_{diff} were significantly under-represented for eGenes (hypergeometric test p =

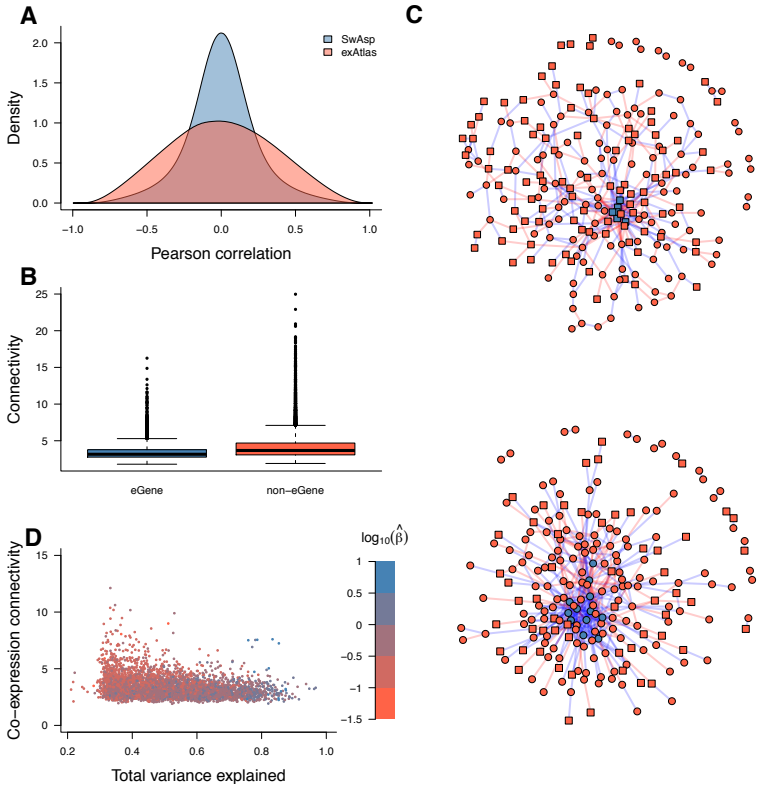

Figure 3. (A) Co-expression correlation distribution of all pairs of genes for the SwAsp data and the exAtlas data. (B) Distribution of co-expression connectivity for eGenes and non-eGenes. (C) Network module 23 and module 22 with core genes coloured blue and eGenes are indicated by square nodes. In module 23 (upper network), all core genes are eGenes, while for module 22, there are very few eGenes in the core. Red edges indicate negative correlation in expression while blue edges indicate positive correlation. (D) eGene connectivity plotted against the total variance explained of all eQTL associated with each eGene. Colors indicate the effect size where more blue represents a more positive effect and red is a more negative effect.

1.7e-27). We defined the core of each module to be the 10% of genes in the module with the highest normalized K_{diff} while also having an intra-modular connectivity >1. Using this definition, all 38 modules contained at least one core gene, with the percentage of core genes ranging from 2-10% (Supplementary file 1). Among the module cores, 28 contained at least one eGene, with 29 module cores being significantly under-represented for eGenes (hypergeometric test, 5% FDR). This further emphasises that eGenes were not central in the network. Subsequently, we tested whether the periphery of the network (see Materials and methods) was overrepresented for eGenes. Sixty-four of 142 genes defined as being peripheral were eGenes, representing a significant enrichment (hypergeometric test $p = 2.6\times10^{-5}$).

In addition to eGenes having generally lower connectivity, there was also a negative relationship between the effect size of eQTLs and co-expression connectivity, where eGenes with the highest connectivities were associated with the smallest effect sizes (Pearson r = -0.15, $p < 10^{-10}$, Figure 3D). eGenes also had higher expression heritability (H^2) than non-eGenes (difference in median heritability 0.24, Mann-Whitney $p < 2.2\times10^{-16}$). In addition, H^2 correlated positively with eQTL effect size (Pearson r = 0.59, $p < 2.2\times10^{-16}$) and negatively with connectivity (Pearson r = -0.30, $p < 2.2\times10^{-16}$).

eGenes had higher expression variance than non-eGenes (median variance of 0.075 versus 0.048, $p < 2.2\times10^{-16}$). In particular, 60 of the 75 genes with variance > 2 (hypergeometric test $p < 2.2\times10^{-16}$) and 178 of the 250 genes with variance > 1 ($p < 2.2\times10^{-16}$), were eGenes (Figure S8). There was a weak negative relationship between network connectivity and gene expression variance (Pearson r = -0.08, $p < 2.2\times10^{-16}$). Transcription factors had higher connectivity than non-transcription factors (Mann-Whitney p-value $< 2.2\times10^{-16}$).

Paralogs with diverged expression are more likely to be eGenes

In *P. tremula* 3,910 paralog pairs were detected (Delhomme *et al.* In prep.), with 2,140 of these (4,185 unique genes) passing our gene expression and variance filtering criteria. These paralogs were significantly under-represented for eGenes, with 1,078 of the 4,185 genes having at least one associated eSNP (hypergeometric test p = 0.0004). Comparing the expression correlation of paralog pairs to that of random gene pairs showed that paralogs exhibited conserved regulation (permutation p-value < 0.001). We compared the expression correlation distributions of paralog pairs containing 0, 1, and 2 eGenes (Figure 4) and found that a higher number of eGenes in a pair was associated with lower expression correlation ($p < 10^{-10}$). Excluding the paralog genes did not alter the fact that eGenes had significantly lower connectivity than non-eGenes.

Population genetics signatures of eQTLs

We examined allele frequencies and compared these for all SNPs as well as considering specifically local and distant eSNPs, revealing that both distant and local were more represented at higher frequencies compared to all non eSNPs (Figure 5A).

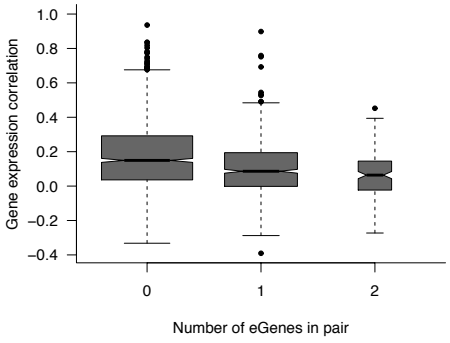

Figure 4. Correlation within paralog pairs as a function of the number of eGenes in the paralog pair. The widths of the boxes are proportional to the number of genes in each set. The mean correlations for paralog pairs with 0, 1, or 2 eGenes were 0.17, 0.10, and 0.06, respectively. The correlation difference for paralog pairs with 0 and 1 eGenes was significant, as well as for paralog pairs with 1 or 2 eGenes (Mann-Whitney $p = 4.8\cdot 10^{-16}$ and p = 0.005, respectively).

We detected a significant negative correlation between eQTL effect size and allele frequency (Spearman's rho=-0.290, P < 0.001) (Figure 5B). We also found significantly positive correlation between effect size and the standardized integrated haplotype score (|iHS|) [61] of eSNPs. Positive selection signals, revealed by |iHS|>2.0, were observed for 6.9% (5,885 eSNPs) of all tested eSNPs (84,956) (Figure 5C). We used one example eSNP (Potra001809:6322) with the highest |iHS| value as a proxy to explore the extent of positive selection at this eSNP (Figure S9). We calculated the extended haplotype homozygosity (EHH) [62] for both the ancestral and derived allele at Potra001809:6332, finding that haplotype homozygosity decayed substantially more rapidly for the derived allele compared to the ancestral allele (Figure S9b).

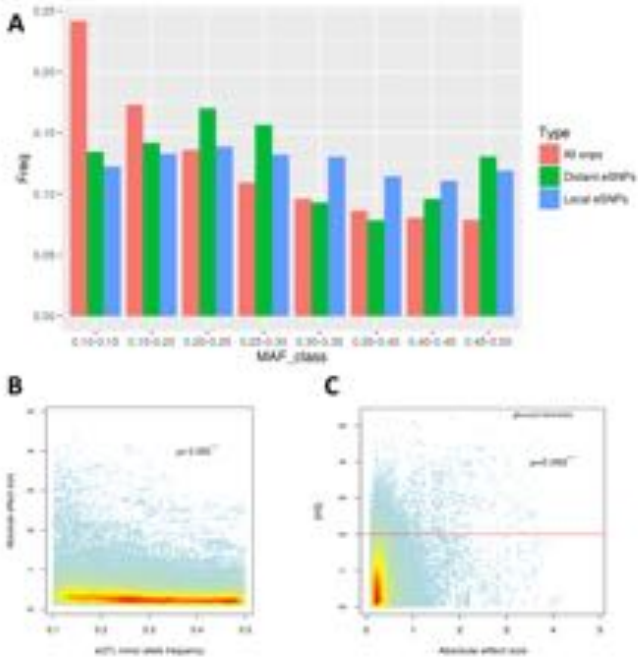

Figure 5. (A) Comparison of minor allele frequency between all SNPs, distant eSNPs and local eSNPs. (B) The relationship between minor allele frequency and effect size (absolute value of beta) of eQTLs. (C) The relationship between effect size (absolute value of beta) and the absolute integrated haplotype score (|iHS|) of eQTLs. The red horizontal line indicates the threshold of positive selection signal (|iHS|>2.0). The black dot indicates the eSNP (Potra001809:6322) with the highest |iHS| value.

Discussion

Expression heritability and population differentiation

Our data identified prevalent heritability in gene expression levels, in line with observations in a number of species [4–16,52]. The majority of these studies reported narrow sense heritability (h^2) estimates, however [20] also reported significant H^2 for the majority of expressed genes. We observed 5,924 genes (17%) having H^2>0.5, representing ~10% more of expressed genes than reported in [52], although direct comparisons are hard given that we calculate H^2 rather than h^2. Although a relatively large proportion of genes had a large fraction of their variance explained by genetic factors (accepting that our estimates may suffer from over-inflation due to the relatively small number of genotypes available), for most genes the largest fraction of variance remained unexplained despite removal of hidden confounders, suggesting that there remained a large influence of environment on expression levels (plasticity).

Q_{ST} values were universally low, showing that there is no apparent differentiation between sub-populations for expression levels. This is not surprising given the lack of population structure at the genome level (Wang *et al.* In prep) and that low values of Q_{ST} have been reported for all phenotypes except the date of bud set (and other traits that are linked or confounded by this, such as height; [63]). However, as just noted, a substantial proportion of expression variance did likely result from environmental effects. As we attempted to remove systematic confounding effects, the expression variation observed will have resulted from environmental effects that hidden confounder removal was not able to account for, combined with inherent transcriptional noise. Although there was no apparent structuring of expression variation among genotypes on the basis of sub-population of origin (*i.e.* population structure), some evidence of clustering was apparent (Figure 1D), but this was not the case for the expression data after hidden confounder removal (Figure S10). Reassuringly, biological replicates of a genotype were consistently clustered together for both the original and the adjusted data, indicating that our data were reliable and that reproducible biological differences (in the current case genetic differences) in gene expression among genotypes could be detected. We were not able to identify the underlying factors responsible for the clustering, but were able to exclude a set of climatic, environmental and experimental factors that represented the most obvious candidates. However, it is of course possible that these environmental factors, especially latitude, may explain the expression of some genes. Despite not being able to identify the causes of the observed clustering, we did identify over-representation of GO categories in the set of the 500 most variable genes, suggesting that this represents a signature with biological relevance.

[64] recently reported a reanalysis of two existing datasets assaying gene expression among natural accessions of *A. thaliana* [65,66] observing that thousands of genes displayed clear present/absent expression among accessions. In contrast, we did not find any genes displaying such a pattern of expression variation (Figure S11), an observation that we also confirmed in an independent *P. tremula* dataset [67], albeit containing substantially fewer genotypes.

eQTLs

In line with other published work, we found more local than distant eQTLs, with local eQTLs explaining significantly more of the variance than distant eQTLs [34,47–50] (Figure 2A). Although each eSNP typically was associated with only one gene, many genes were controlled by more than one

unique eSNP and the combined variance explained by these eSNPs was higher than for the individual eSNPs, both for local and distant eQTLs. This suggests that the observed expression variation resulted from the combined effects of several eSNPs, consistent with the infinitesimal model. However, combining local and distant eQTLs did not improve on only combining local eQTLs, which is surprising since one would expect local and distant eQTLs to not be in linkage and to contribute independent information. On the other hand, there were far fewer distant eQTLs affecting fewer genes and being of lower effect. Looking specifically at the 762 eGenes with both local and distant eQTLs, we did indeed see an average increase in the proportion of variance explained by all eSNPs compared to only local eSNPs (an increase of 0.04 %VE after adjustment).

Local eQTLs were typically located in regulatory regions (Figure 2B), with UTRs having the highest density of local eSNPs (~1.5 eSNP per kb sequence) followed by flanking regions (2kb up- and downstream, ~1 per kb sequence) and introns (~0.75 per kb sequence). About half of the local eQTLs were located in or near the associated gene, suggesting a direct role in affecting the expression of that gene. Most of the remaining local eSNPs were located near another gene, with only a few being intergenic. eSNPs located near the associated eGene and those located near another gene were distributed fairly equally across different genomic contexts, with highest representation of regulatory regions. This indicates that eSNPs located near to a gene other than the one for which the eQTL was mapped (but still classified as local) may act indirectly (potentially through additional regulators) on the eGene and, if this is the case, that the indirect action of the eSNP is most often accomplished by affecting expression rather than by changing coding sequence. This also extends to many distant eQTLs, which had a remarkably similar distribution among genomic contexts to that of local eQTLs. However, distant eQTLs were somewhat more common in exon regions compared to local ones, as would be expected by indirectly acting eQTLs, and much more often located in intergenic regions, potentially representing distantly acting regulatory elements such as enhancers. [68] reported introns as the most frequent SNP location for distant eQTLs in humans, followed by intergenic regions. Our results only partially agree with this, with distant eQTLs being far more commonly associated with UTRs and flanking regions, followed by introns and intergenic regions.

Previous eQTL studies in both plants and a range of other organisms have frequently identified distinct eQTL hotspots [19,34,36,47,50], which represent loci where numerous *trans*-acting effects are co-located, with a stringent definition requiring that more such eQTLs are co-located than would be expected by chance. In our data distant eQTLs were distributed fairly equally across the genome (Figure 2C), with a maximum of six eGenes associated to a single eSNP and with the vast majority of eSNPs associated to a single gene. Analysing genomic windows of 100 kbp did reveal clear peaks of eSNPs on chromosomes 6, 11 and 18, however, these were not associated with corresponding eGene peaks (Figure S7). Thus, our data do not support the presence of eQTL hotspots. It has also been reported that hotspots are often associated with polymorphisms that alter developmental timing, which was not the case in our study.

The relatively small size of our population limits the power available for GWAS mapping and overinflates effect sizes due to the 'Winner's curse'. Thus, our ability to detect eQTLs associated with rare alleles were limited. Before performing the association mapping we therefore filtered SNPs on minor allele and major genotype frequencies (Materials and methods). This filtering accomplished two goals: Firstly, we removed alleles observed in a single (or very few) individuals that often result in highly significant p-values (see Figure S12 for an example); Secondly, we increased our power to detect other, more frequent eSNPs, by reducing the number of variables going into the association mapping. We do not claim that the removed SNPs are not important but, rather, that we lack the detection

power to distinguish *bona fide* low-frequency eSNPs (false negatives) from spurious ones (false positives). Having discussed what SNPs we *did not* detect, it is interesting to analyse the properties of SNPs we *did* detect. There was a relatively high correlation between both the effect size of, and the variance explained by, an eSNP and expression heritability (H^2) of the associated eGenes (Rs ranging from 0.40 to 0.60). Furthermore, we observed a substantially higher H^2 for eGenes than for non-eGenes (Figure 2D). Since expression heritability measure the fraction of gene expression variance that can be be explained by genetic variation, it is maybe not surprising that this measure is a good proxy for whether a gene will be detected as an eGene or not, and how strong the association will be. However, it is reassuring that low expression variance among genotype replicates coupled with high variation among genotypes (i.e. the hallmark of a high H^2) is consistently associated with detection of eSNPs in our data.

Co-expression

It remains an open question whether there are distinct characteristics associated with genes displaying natural variation in expression and whether coexpression networks representing natural variation in gene expression resemble the coexpression networks more typically considered by systems biology studies. Here we analyzed such characteristics using a gene co-expression network together with eQTL mapping within a natural population of unrelated individuals of the outbreeding forest tree species *P. tremula*. After removing hidden confounders from the expression data, the assumption is that a significant portion of the variance in expression of single genes, and of significant co-expression between genes, has a genetic basis.

The pairwise expression correlations underlying our co-expression network were low (mean 0.00 +/- s.d. 0.12), at least compared to those observed in most systems biology studies, for example in the *P. tremula* exAtlas network [57] where all samples originated from different tissues in a single clone (mean correlation 0.01 +/- s.d. 0.36) (Figure 3A). The fact that correlation values were generally low in our network is not surprising given that we are assaying natural variation among genotypes of the same species sampled under controlled conditions with the specific intention to limit environmental noise. It would be somewhat surprising if gene expression between individuals varied to the same extent as when exposing a single individual to often high effect perturbations or between developmental stages or tissues. Despite the low correlations, our natural variation co-expression network displayed typical properties characteristic to biological networks [69]; the network was scale-free with hubs and distinct modules. Specifically, we identified 38 co-expression modules that were enriched for a number of functional categories, suggesting that they represent biologically meaningful units. A central aim of this study was to ascertain to what degree eQTLs could explain these network modules.

Genes with the highest expression variance, and the highest heritability, in our data were enriched for eGenes. However, perhaps surprisingly, genes co-expressed with many other genes (high connectivity) were under-represented for eGenes, while genes with few co-expression partners (located in the network periphery) were enriched for eGenes (Figure 3B). Furthermore, eGene connectivity was negatively correlated with variance explained by the associated eSNPs (Figure 3D). Thus there seems to be a need to consolidate the expectation that eSNPs should explain the co-expression network with our observation that network hubs were underrepresented by eGenes and that eGenes of higher connectivity were controlled by eSNPs of lower effect size. While this may seem paradoxical, a moment's pause reveals many plausible explanations for these observations.

The simplest explanation for a cluster of co-expressed genes (*i.e.* a network module) is that it results from variation (either expression or altered protein structure) in a single regulator, with that variation being caused by a single eSNP. If *a single* regulator induces the observed co-expression among module members, then it is almost inevitable that a module will be underrepresented by eGenes. This is in agreement with our observation that eGenes were under-represented in module cores. In fact, it would be sufficient that at least one eGene be present within each module core in order for the eSNPs to explain the co-expression data: in our data we observed at least one eGene in 28 of the 38 module cores. In this scenario, if co-expression between gene pairs were high, one would expect the eSNP(s) controlling the regulator to also be identified as a distant eQTL for the other coexpressed genes in the module. However, pairwise correlation in our data was low and eSNPs were generally associated with a single gene; most likely the direct target of the eSNP. Unlike network modules, the expression variation of genes with few expression partners (with a rare expression profile) is most likely explained by the presence of a directly-acting local eSNP, thus explaining the observed enrichment of eGenes within the network periphery.

The above assumption that regulators are controlled by a single eSNP would not be favored by selection as it offers no opportunity for buffering. Regulatory networks are known to be redundant, scale-free and to have an inherent ability to buffer against single mutations of large negative effect [70–72]. Genes with the ability to cause such large negative effects would, therefore, be observed as hubs in a co-expression network, and we would expect them to be controlled by complex regulatory mechanisms involving additive effects. This, in turn, would imply that many SNPs, each of small effect size, would contribute to the control of expression variation of those hub genes. Under this assumption we would therefore lack power to detect such eSNPs given our current sample size, and these would represent false negatives (*i.e.* would be unobserved). This lack of detection power could explain why the core of 10 of the 38 modules within our network did not contain eGenes.

Just as nature abhors a vacuum, biology appears to abhore simplicity. As such, it should be expected that neither of the above explanations would account for all variation in gene expression. In fact, although we mention additive effects above, several other types of complex regulatory mechanisms can preclude our ability to detect eSNPs. Epistatic interactions between eSNPs, such as AND-logics, would result in very little correlation between the individual AND-interacting eSNPs and the expression of the controlled gene. Likewise, the expression of a gene will frequently not only be a function of the associated eSNPs but also of the expression of the regulators acting through those SNPs, which are themselves affected by other eSNPs. Furthermore, the regulator(s) that initiated an observed co-expression module may no longer be expressed at the sampling point used for RNA isolation, and thus the eSNPs that determined the expression variation of that regulator will no longer be visible to us (or a 'ghost signature' may remain, but be too weak to reach significance).

In conclusion, we propose that a modular co-expression network from a natural population should be determined by a relatively small set of highly connected regulators displaying expression variation due to eSNPs. Since these regulators will be hubs, with potentially huge negative effects, they will be buffered by complex regulatory mechanisms making them difficult to detect by small association studies testing the effect of individual SNPs. Supporting this hypothesis, we find that eGenes are underrepresented in module cores and that transcription factors have higher connectivity than non-transcription factors. Nonetheless, we do find eGenes in the core of 28 of 38 modules, in principle being adequate to explain these modules. In addition, we find a large number of eGenes with low connectivity, fine tuning the expression of individual genes. [73] similarly found that conserved genes and hubs in human protein-protein interaction networks were less likely to be associated with a

detectable eQTL and that the effect size of eQTLs were negatively correlated with connectivity in the protein-protein interaction network. In this study, we confirmed similar observations in the context of a natural variation and co-expression networks.

Our results support the hypothesis that natural variation in gene expression follows the infinitesimal model, with there being remarkably few examples of a single eSNP controlling a large proportion of the expression variance of a gene. As we argue, this may arise from selective processes that maintain biological robustness. However, we cannot extrapolate these findings to suggest whether this is the rule for all cases. For example, we observed a snapshot of expression during the process of bud flush, a trait that does not appear to be under strong directional selection. To ascertain the general relevance of our findings, similar studies should be performed assaying expression for phenotypic traits that do show evidence of selection, such as the timing of autumn bud set within the SwAsp collection [63], although the specifics of how to sample comparative material while controlling for confounding effects to enable expression profiling presents significant challenges in such a case.

Salicaceae species underwent a recent whole-genome duplication that remains represented by a large number of paralogous gene pairs. If many of these duplicated genes are functionally redundant or in the process of diverging, one would expect them to be overrepresented for eGenes as sub- or neo-functionalization requires derived SNPs to drive expression differences. However, we saw an under-representation of eGenes in paralog pairs, suggesting full divergence and fixation of most of the paralogs associated with the whole-genome duplication - that is that each of the genes in a paralog pair has diverged expression and that the SNPs that initially induced that expression divergence have reached fixation. Interestingly, we found progressively lower expression correlation between paralog pairs containing zero, one or two eGenes (Figure 4) indicating that eSNPs drive paralogs away from their present state of higher co-expression than randomly selected gene pairs (comparing Figure 3A to Figure 4 clearly shows higher co-expression for paralogous pairs). The fixation of such eSNPs will, over evolutionary time, seal the fate of these paralogs through the process of sub-, neo- or, most commonly, non-functionalization. It could be argued that paralogous eGenes would be loners in the co-expression network, because they are evolving away from their ancestral function in "search of" a new functional role, and that such drifting paralogs could be responsible for the observation that eGenes were enriched at the network periphery. However, removing paralogs from the analysis did not change the fact that eGenes have lower connectivity than non-eGenes, thus refuting this hypothesis.

Signatures of selection

In contrast to the pattern reported in [31], our results suggest that eSNPs, both local and distant, have been maintained at higher frequencies compared to the global set of SNPs not associated with gene expression in our data (Figure 5A). Rather than this suggesting that eSNPs were under weaker purifying selection, these results more likely reflect the lower power of our study to detect the true positive eSNPs of low allele frequency. Thus, future work will need to consider the influence of the variation in ascertainment power across allele frequencies during the detection of eSNPs and/ or eQTLs.

We detected a significant negative correlation between eQTL effect size and eSNP allele frequency (Figure 5B), suggesting that prevalent purifying selection may have been acting on expression variation of the eQTLs detected in our study. Nonetheless, we also observed a significantly positive correlation between effect size and |iHS| (Figure 5C), suggesting the action of positive selection on these eSNPs. These results indicate that positive selection has also been involved in shaping the regulation of gene

expression variation among individuals. In combination, our results further suggest that variation in gene expression is controlled in a similar manner to most quantitative traits, by the combined action of a large number of components (eSNPs in the case of gene expression and genes in the case of phenotypes) and that any given locus, in general, explains only a very small fraction of the genetic variance. Although further work addressing this point explicitly is required, our results also likely indicate that expression variation is an important contributor to standing phenotypic variation.

Materials and methods

Samples

We collected branches form the SwAsp collection common garden in north of Sweden on 27th May 2012, before natural bud break but as close to the point of natural spring bud break as possible. Branches were placed in the greenhouse facility at the Umeå Plant Science Centre under conditions selected to induce rapid bud break (24 h light, temperature of 20 °C and humidity 50-70 %). At a defined point of emergence (Figure S13), buds were harvested, flash frozen in liquid nitrogen and stored at -80℃ until used for RNA isolation. Only terminal buds were sampled (i.e. no lateral buds were included). The time from the day branches were placed in the greenhouse until bud flush sampling ranged from one day to eight days (Figure S14A) and there was a high, positive correlation to bud flush date recorded in the field for the same year (Figure S14B; $r=0.776$, $p < 2.2\times10^{-16}$). As has previously been reported [63], there was no apparent Q_{ST} for bud flush, either in the field or for the greenhouse material (Q_{ST} 0.13 and 0.07 respectively), however H^2 was high (H^2 = 0.82 and 0.71 respectively).

RNA isolation

One to two buds per clonal replicate were ground using one 3 mm stainless steel bead (Qiagen, Redwood city, USA) in Corning^{R} 96 well PP 1.2 ml cluster tubes (Sigma-Aldrich, St. Louis, USA) using a Mixer Mill MM400 (Retsch, Haan, Germany) at 20 Hz for 2 x 15 sec . Total RNA was extracted from all samples according to [74] with the omission of the L spermidine. Buffer volumes were adjusted according to starting material (70 - 130 mg). RNA isolation was performed using one extraction with CTAB buffer followed by one chloroform : isoamyl alcohol IAA (24:1) extraction. All other steps were performed as in [74]. DNA contamination was removed using DNA-free^{TM} DNA removal Kit (Life Technologies, Carlsbad, USA). RNA purity was measured using a NanoDrop 2000 (Thermo Scientific, Wilmington, USA) and RNA integrity was assessed using the Plant RNA Nano Kit for the Bioanalyzer (Agilent Technologies, Santa Clara, USA).

RNA-Sequencing and analysis

RNA-Sequencing was performed as in [67]. Briefly, paired-end (2 × 100 bp) RNA-Seq data were generated using standard Illumina protocols and kits (TruSeq SBS KIT-HS v3, FC-401-3001; TruSeq PE Cluster Kit v3, PE-401-3001) and all sequencing was performed using the Illumina HiSeq 2000 platform at the Science for Life Laboratory, Stockholm, Sweden. Raw data is available at the European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena) with accession number ERP014886.

RNA-Seq FASTQ-files were pre-processed and aligned to v1.0 of the *P. tremula* reference genome (available at http://popgenie.org) as in [75]. In short, reads were quality and adapter trimmed using Trimmomatic v0.32 [76], rRNA matching reads were filtered using SortMeRNA v1.9 [77], reads were

aligned to the v1.0 *P. tremula* reference genome using STAR 2.4.0f1 [78] and read counts were obtained using htseq-count from HTSeq [79]. FastQC [80] was used to track read quality throughout the process. Normalised gene expression values were obtained by applying a variance stabilising transformation (vst) to the raw counts from HTSeq, as implemented in the DESeq2 R-package [81].

Gene expression H^2 and Q_{ST}

We calculated repeatability as an assumed upper bound estimate of broad sense heritability of gene expression (see [82] for discussion) from the variance estimates in our data according to the equation

$$H^2 = \frac{V_G}{V_P}$$

where V_G is the genetic component of the variance calculated as the expression variance between genotypes for a particular gene (*i.e.* variance among genotype means) and V_P is the total phenotypic variance calculated as the sum of V_G and V_E, where V_E is the environment component of the variance calculated as the expression variance within genotypes for a particular gene (i.e. the mean variance among clonal replicates). Point estimates of H^2 were obtained using the repeatability function from the heritability R package [83].

Population differentiation (Q_{ST}; [84]) was calculated as

$$Q_{ST} = \frac{V_{between}}{V_{between} + 2V_{genetic}}$$

where $V_{between}$ is the variance among populations and $V_{genetic}$ is the genetic variance among genotypes as computed using the lmer function from the lme4 R package [85] using the formula

```
expression ~ 1 + (1|population) + (1|clone)
```

where *expression* is the expression of a gene, *population* is a factor representing the population of each sample, and *clone* is a factor representing genotype replicates. As we use repeatability as an upper bound estimate of H^2, our Q_{ST} estimates are conservative [86].

Hidden confounder removal

Gene expression data was adjusted for hidden confounders before mapping eQTLs and constructing the co-expression network. Hidden confounders in the gene expression data was accounted for by regressing out the 9 first principal components (PCs) of the gene expression data [87–89]. The number of components to remove was determined by running the eQTL mapping with 0 to 20 PCs removed and selecting the number of components that yielded the largest number of significant eQTLs (Benjamini-Hochberg $p < 0.05$) (Figure S15). This approach is based on the assumption that the number of identified eQTLs will increase if the removed PCs are removing unwanted, systematic variation (i.e. noise) rather than informative biological variation [87–89].

eQTL mapping

eQTL mapping was performed by associating gene expression with biallelic SNPs using the R package Matrix eQTL v2.1.1 [60]. Before doing the association, genes were filtered on variance so that only genes with a gene expression variance above 0.05 were included. SNPs were also filtered on minor allele frequency (MAF) and major genotype frequency (MGF); any SNPs with MAF < 0.1 or MGF > 0.9 were excluded to avoid spurious associations. The first genotype principal component based on independent SNPs (see Wang *et al.* In prep.) was used as a covariate in the linear model used by Matrix eQTL to account for the weak signature of population structure. Permutation testing was used to determine eQTL significance whereby genotype sample labels were permuted 1000 times and the maximum absolute t-statistic from Matrix eQTL was recorded for each gene across all SNPs for each permutation. Empirical p-values were calculated with the empPvals function in the qvalue R-package [90], and q-values (empirical FDR) were calculated with the qvalue function in the same package.

When determining the genomic context of eSNPs, there were some cases where introns overlapped exons as a result of overlapping gene model being present on the same strand. These 41 eSNPs were discarded from the counting. Another type of overlap that was discarded were cases where an eSNP overlapped a gene feature, but no sub-feature inside that gene (e.g. UTR, exon or intron). These 1961 eSNPs were excluded from the counting. Since many of the features overlap (e.g. exon and untranslated regions), the priority for counting was untranslated region, exon/intron, upstream/downstream and intergenic.

Co-expression

The R-package WGCNA [91] was used for constructing a co-expression network. The input gene expression values were per-gene genotype means. We chose to use the unsigned network type for this study with the motivation that we did not want to discard negative relationships. By looking at this from an eQTL perspective, an eSNP can be positively associated with one gene while negatively associated with another. There is a relationship between these genes that would be missed if we used the signed or the signed-hybrid approaches. Using the unsigned approach, we assure that genes with strong negative correlation end up in the same network modules. A soft thresholding power of 5 was used to calculate adjacencies. The topological overlap matrix (TOM) was generated using the TOMsimilarity function with the signed approach in order to take negative edges into account (see [91] for details). In order to identify network modules, hierarchical clustering was applied to the TOM dissimilarity matrix (1 – TOM) and the resulting dendrogram was divided into modules using the cutreedynamic function. The connectivity of the network was then defined as the adjacency sum for each node, i.e. the weights of the edges that are connected to this node. This concept was applied to modules as well to obtain measures of intra- and inter-modular connectivity, i.e. the connectivity based on edges connecting the gene with other genes inside the same module, and connectivity based on edges connecting the gene with genes outside of the module.

To define the periphery of the network we applied a hard edge-threshold to the network where only gene-pairs with an absolute Pearson correlation > 0.22 were linked, which corresponded to the top 1% most correlated gene pairs. Genes were then classed as peripheral if they linked to only one other gene.

References

1. Fisher RA. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans. R. Soc. Edinburgh. 1919;52:399–433.

2. Sandberg R, Yasuda R, Pankratz DG, Carter TA, Del Rio JA, Wodicka L, et al. Regional and strain-specific gene expression mapping in the adult mouse brain. Proc. Natl. Acad. Sci. 2000;97:11038–43.

3. Primig M, Williams RM, Winzeler EA, Tevzadze GG, Conway AR, Hwang SY, et al. The core meiotic transcriptome in budding yeasts. Nat. Genet. 2000;26:415–23.

4. Jin W, Riley RM, Wolfinger RD, White KP, Gurgel GP, Gibson G. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. 2001;29:389–95.

5. Oleksiak MF, Churchill GA, Crawford DL. Variation in gene expression within and among natural populations. Nat. Genet. 2002;32:261–6.

6. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature. 2003;422:297–302.

7. Kirst M, Myburg AA, Leon JPG De, Kirst ME, Scott J, Sederoff R. Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. Plant Physiol. 2004;135:2368–78.

8. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, et al. Genetic analysis of genome-wide variation in human gene expression. Nature. 2004;430:743–7.

9. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. Nature. 2005;437:1365–9.

10. Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, et al. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat Genet. 2005;37:243–53.

11. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, et al. Genome-wide associations of gene expression variation in humans. PLoS Genet. 2005;1:0695–704.

12. DeCook R, Lall S, Nettleton D, Howell SH. Genetic Regulation of Gene Expression During Shoot Development in Arabidopsis. Genetics. 2006;172:1155–64.

13. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KCC, et al. A genome-wide association study of global gene expression. Nat Genet. Nature Publishing Group; 2007;39:1202–7.

14. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. 2008;24:408–15.

15. Kim J, Gibson G. Insights from GWAS into the quantitative genetics of transcription in humans. Genet. Res. (Camb). 2010;92:361–9.

16. Powell JE, Henders AK, McRae AF, Kim J, Hemani G, Martin NG, et al. Congruence of Additive and Non-Additive Effects on Gene Expression Estimated from Pedigree and SNP Data. PLoS Genet. 2013;9.

17. Wang RL, Stec A, Hey J, Lukens L, Doebley J. The limits of selection during maize domestication. Nature. 1999;398:236–9.

18. Carroll SB. Endless forms: the evolution of gene regulation and morphological diversity. Cell. 2000;101:577–80.

19. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science (80-.). 2002;296:752–5.

20. Ayroles JF, Carbone MA, Stone E a, Jordan KW, Lyman RF, Magwire MM, et al. Systems genetics of complex traits in Drosophila melanogaster. Nat. Genet. 2009;41:299–307.

21. Mackay TFC, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. Nat. Rev. Genet. Nature Publishing Group; 2009;10:565–77.

22. Liao B-Y, Weng M-P, Zhang J. Contrasting genetic paths to morphological and physiological evolution. Proc. Natl. Acad. Sci. 2010;107:7353–8.

23. Hines HM, Papa R, Ruiz M, Papanicolaou A, Wang C, Nijhout HF, et al. Transcriptome analysis reveals novel patterning and pigmentation genes underlying *Heliconius* butterfly wing pattern variation. BMC Genomics. 2012;13:288.

24. Richards CL, Rosas U, Banta J, Bhambhra N, Purugganan MD. Genome-Wide Patterns of Arabidopsis Gene Expression in Nature. Gibson G, editor. PLoS Genet. Public Library of Science; 2012;8:e1002662.

25. Jansen RCC, Nap JPJP. Genetical genomics: the added value from segregation. Trends Genet. 2001;17:388–91.

26. Doerge RW. Mapping and analysis of quantitative trait loci in experimental populations. Nat. Rev. Genet. 2002;3:43–52.

27. Flint J, and Mackay TFC. Genetic architecture of quantitative traits in flies, mice and humans. Genome Res. 2009;19:723–33.

28. Hindorff L a, Sethupathy P, Junkins H a, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U. S. A. 2009;106:9362–7.

29. Ku CS, Loy EY, Pawitan Y, Chia KS. The pursuit of genome-wide association studies: where are we now? J. Hum. Genet. Nature Publishing Group; 2010;55:195–206.

30. Wray GA. The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. Nature Publishing Group; 2007;8:206–16.

31. Josephs EB, Lee YW, Stinchcombe JR, Wright SI. Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. Proc. Natl. Acad. Sci. U. S. A. 2015;112:15390–5.

32. Lappalainen T. Functional genomics bridges the gap between quantitative genetics and molecular biology. Genome Res. 2015;25:1427–31.

33. Kliebenstein DJ, West MAL, van Leeuwen H, Kim K, Doerge RW, Michelmore RW, et al. Genomic Survey of Gene Expression Diversity in Arabidopsis thaliana. Genetics. 2006;172:1179–89.

34. Keurentjes JJB, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G, Snoek LB, et al. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proc. Natl. Acad. Sci. National Academy of Sciences; 2007;104:1708–13.

35. van Leeuwen H, Kliebenstein DJJ, West MALA, Kim K, van Poecke R, Katagiri F, et al. Natural variation among Arabidopsis thaliana accessions for transcriptome response to exogenous salicylic acid. Plant Cell Online. Department of Plant Sciences, University of California, Davis, California 95616.: American Society of Plant Biologists; 2007;19:2099–110.

36. West M a L, Kim K, Kliebenstein DJ, Van Leeuwen H, Michelmore RW, Doerge RW, et al. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. Genetics. 2007;175:1441–50.

37. Zhang X, Cal AJ, Borevitz JO. Genetic architecture of regulatory variation in *Arabidopsis thaliana*. Genome Res. 2011;21:725–33.

38. Lowry DB, Logan TL, Santuari L, Hardtke CS, Richards JH, DeRose-Wilson LJ, et al. Expression quantitative trait locus mapping across water availability environments reveals contrasting associations with genomic features in Arabidopsis. Plant Cell. 2013;25:3266–79.

39. Swanson-Wagner R a, DeCook R, Jia Y, Bancroft T, Ji T, Zhao X, et al. Paternal dominance of trans-eQTL influences gene expression patterns in maize hybrids. Science. 2009;326:1118–20.

40. Holloway B, Luck S, Beatty M, Rafalski J-A, Li B. Genome-wide expression quantitative trait loci (eQTL) analysis in maize. BMC Genomics. 2011;12:336.

41. Fu J, Cheng Y, Linghu J, Yang X, Kang L, Zhang Z, et al. RNA sequencing reveals the complex regulatory network in the maize kernel. Nat. Commun. Nature Publishing Group; 2013;4:2832.

42. Wang J, Yu H, Xie W, Xing Y, Yu S, Xu C, et al. A global analysis of QTLs for expression variations in rice shoots at the early seedling stage. Plant J. 2010;63:1063–74.

43. Wang J, Yu H, Weng X, Xie W, Xu C, Li X, et al. An expression quantitative trait loci-guided co-expression analysis for constructing regulatory network using a rice recombinant inbred line population. J. Exp. Bot. 2014;65:1069–79.

44. Kirst M, Basten CJ, Myburg AA, Zeng ZB, Sederoff RR. Genetic Architecture of Transcript-Level Variation in Differentiating Xylem of a Eucalyptus Hybrid. Genetics. 2005;169:2295–303.

45. Drost DR, Benedict CI, Berg A, Novaes E, Novaes CRDB, Yu Q, et al. Diversification in the genetic architecture of gene expression and transcriptional networks in organ differentiation of Populus. Proc. Natl. Acad. Sci. U. S. A. 2010;107:8492–7.

46. Kullan AR, van Dyk MM, Hefer C a, Jones N, Kanzler A, Myburg A a. Genetic dissection of growth, wood basic density and gene expression in interspecific backcrosses of Eucalyptus grandis and E. urophylla. BMC Genet. 2012;13:60.

47. Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc. Natl. Acad. Sci. U. S. A. 2005;102:1572–7.

48. Hughes KA, Ayroles JF, Reedy MM, Drnevich JM, Rowe KC, Ruedi EA, et al. Segregating Variation in the Transcriptome: Cis Regulation and Additivity of Effects. Genetics. 2006;173:1347–55.

49. Meiklejohn CD, Parsch J, Ranz JM, Hartl DL. Rapid evolution of male-biased gene expression in Drosophila. Proc. Natl. Acad. Sci. U. S. A. 2003;100:9894–9.

50. Potokina E, Druka A, Luo Z, Wise R, Waugh R, Kearsey M. Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. Plant J. 2008;53:90–101.

51. Whitehead A, Crawford DL. Variation within and among species in gene expression: raw material for evolution. Mol. Ecol. 2006;15:1197–211.

52. Leder EH, McCairns RJS, Leinonen T, Cano JM, Viitaniemi HM, Nikinmaa M, et al. The evolution and adaptive potential of transcriptional variation in sticklebacks--signatures of selection and widespread heritability. Mol. Biol. Evol. 2015;32:674–89.

53. Jansson S, Douglas CJ. Populus: A Model System for Plant Biology. Annu. Rev. Plant Biol. 2007;58:435–58.

54. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). Science (80-.). 2006;313:1596–604.

55. Wang J, Street NR, Scofield DG, Ingvarsson PK. Variation in linked selection and recombination drive genomic divergence during allopatric speciation of European and American aspens. Mol. Biol. Evol. 2016;msw051.

56. Wang J, Street NR, Scofield DG, Ingvarsson PK. Natural Selection and Recombination Rate Variation Shape Nucleotide Polymorphism Across the Genomes of Three Related Populus Species. Genetics. 2016;202:1185–200.

57. Sundell D, Mannapperuma C, Netotea S, Delhomme N, Lin Y, Sjödin A, et al. The Plant Genome Integrative Explorer Resource: PlantGenIE.org. New Phytol. 2015;208:1149–56.

58. Luquez V, Hall D, Albrectsen BR, Karlsson J, Ingvarsson P, Jansson S. Natural phenological variation in aspen (Populus tremula): the SwAsp collection. Tree Genet. Genomes. 2008;4:279–92.

59. Ingvarsson PK. Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., *Salicaceae*). Genetics. 2005;169:945–53.

60. Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28:1353–8.

61. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4:0446–58.

62. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419:832–7.

63. Robinson KM, Ingvarsson PK, Jansson S, Albrectsen BR. Genetic variation in functional traits influences arthropod community composition in aspen (Populus tremula L.). PLoS One. 2012;7:e37679.

64. Zan Y, Shen X, Forsberg SKG, Carlborg Ö. Genetic regulation of transcriptional variation in wild-collected Arabidopsis thaliana accessions. bioRxiv. 2016;

65. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, et al. Patterns of population epigenomic diversity. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013;495:193–8.

66. Dubin MJ, Zhang P, Meng D, Remigereau M-S, Osborne EJ, Paolo Casale F, et al. DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. Elife. eLife Sciences Publications Limited; 2015;4:e05255.

67. Robinson KM, Delhomme N, Mähler N, Schiffthaler B, Onskog J, Albrectsen BR, et al. Populus tremula (European aspen) shows no evidence of sexual dimorphism. BMC Plant Biol. 2014;14:276.

68. Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, et al. Heritability and genomics of gene expression in peripheral blood. Nat. Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;46:430–7.

69. Barabasi A-L, Oltvai ZNZN, Barabási A-L. Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 2004;5:101–13.

70. Macneil LT, Walhout AJM. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. Genome Res. 2011;21:645–57.

71. Whitacre JM, Bender A. Networked buffering: a basic mechanism for distributed robustness in complex adaptive systems. Theor. Biol. Med. Model. 2010;7:20.

72. Whitacre JM. Biological robustness: Paradigms, mechanisms, systems principles. Front. Genet. 2012;3:1–15.

73. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2013;

74. Chang S, Puryear J, Cairney J. A simple and efficient method for isolating RNA from pine trees. Plant Mol. Biol. Report. 1993;11:113–6.

75. Delhomme N, Mähler N, Schiffthaler B, Sundell D, Mannepperuma C, Hvidsten TR, et al. Guidelines for RNA-Seq data analysis. Epigenesys. 2014;

76. Bolger A, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;1–7.

77. Kopylova E, Noé L, Touzet H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28:3211–7.

78. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2012;29:15–21.

79. Anders S, Pyl PT, Huber W. HTSeq - A Python framework to work with high-throughput sequencing data. Bioinformatics. 2014;31:166–9.

80. Andrews S. FastQC [Internet]. 2016 [cited 2016 Apr 30]. Available from: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

81. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

82. Dohm MR. Repeatability estimates do not always set an upper limit to heritability. Funct. Ecol. 2002;16:273–80.

83. Kruijer W, Boer MP, Malosetti M, Flood PJ, Engel B, Kooke R, et al. Marker-based estimation of heritability in immortal populations. Genetics. 2015;199:379–98.

84. Spitze K. Population structure in Daphnia obtusa: quantitative genetic and allozymic variation. Genetics. 1993;135:367–74.

85. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Softw. 2015;67:1–48.

86. Manier MK, Palumbi SR. Intraspecific divergence in sperm morphology of the green sea urchin, Strongylocentrotus droebachiensis: implications for selection in broadcast spawners. BMC Evol. Biol. 2008;8:283.

87. Hyun MK, Ye C, Eskin E. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. Genetics. 2008;180:1909–25.

88. Pickrell JK, Marioni JC, Pai A a, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. Nature Publishing Group; 2010;464:768–72.

89. Mostafavi S, Battle A, Zhu X, Urban AE, Levinson D, Montgomery SB, et al. Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. PLoS One. 2013;8:e68141.

90. Storey J. qvalue: Q-value estimation for false discovery rate control. 2015.

91. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.

Supplementary files

Supplementary file 1: https://figshare.com/s/e73d155d398b0e170fa0

Supplementary figures

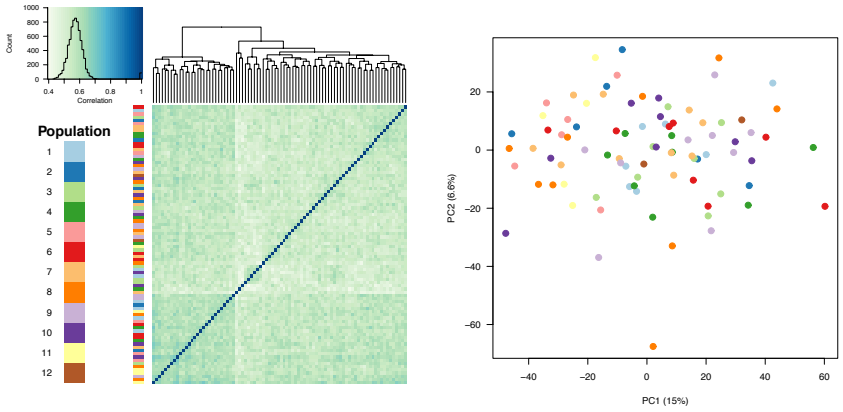

Figure S1. Clustering of genotypes. (Left) Heatmap of the sample correlation matrix based on the 500 most variably expressed genes. Darker colour indicates higher correlation. The coloured bar represent the populations the genotypes belong to. (Right) The two first principal components from a principal component analysis (PCA) based on all genes. Again, colours represent the genotype population. The percentages in the axis labels indicate the amount of variance explained by each component.

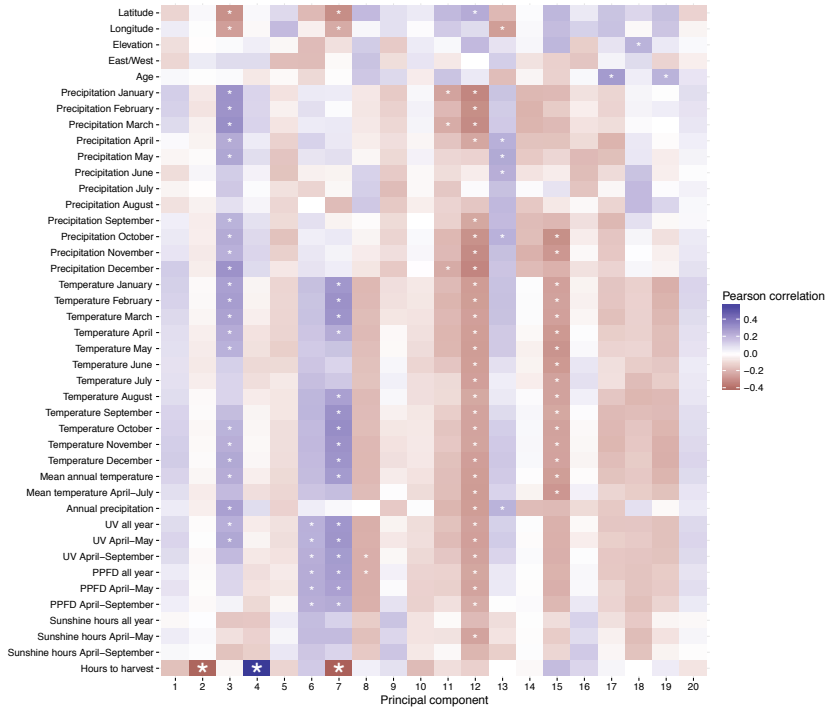

Figure S2. Correlations between gene expression principal components (PCs) and environmental variables. The values in each tile represent the Pearson correlation between the gene expression PC (x-axis) and environmental variable (y-axis). Small asterisks represent a nominal p-value < 0.05 while large asterisks represent Benjamini-Hochberg (BH) adjusted p-values < 0.05. The only factor with significant correlations to expression PCs was "Hours to harvest", which is the number of hours into the sampling period that the buds were harvested. It was significantly associated with PC4 (BH-adjusted p = 4.6e-6), PC7 (BH-adjusted p = 0.030) and PC2 (BH-adjusted p = 0.033).

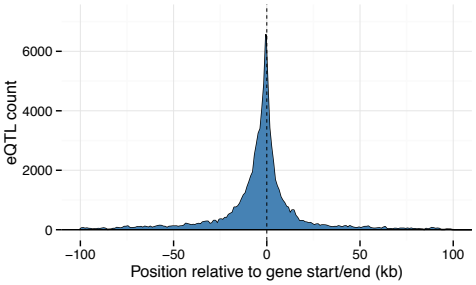

Figure S3. Distribution of local eSNPs relative to the gene that they are associated to. The zero-position represents both the start and the end of the gene feature, i.e. intragenic features are not shown.

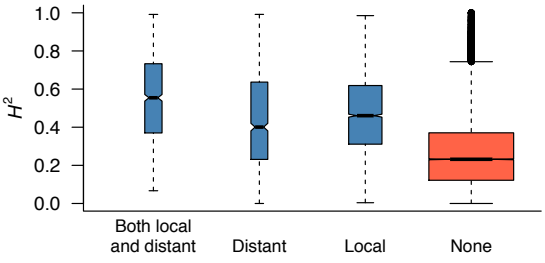

Figure S4. Heritability distributions for four non-overlapping sets of genes: genes with both local and distant eQTLs, genes with only distant eQTLs, genes with only local eQTLs, and genes with no significant eQTLs.

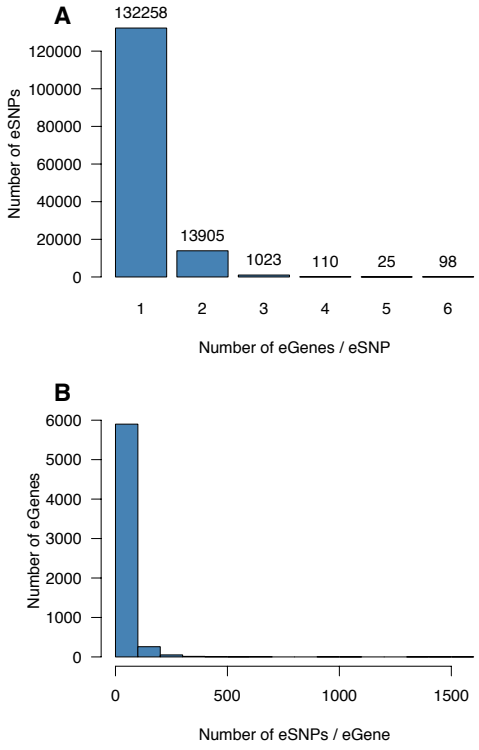

Figure S6. (A) Bar plot of the number of associated eGenes per eSNP. (B) Histogram of the number of associated eSNPs per eGene.

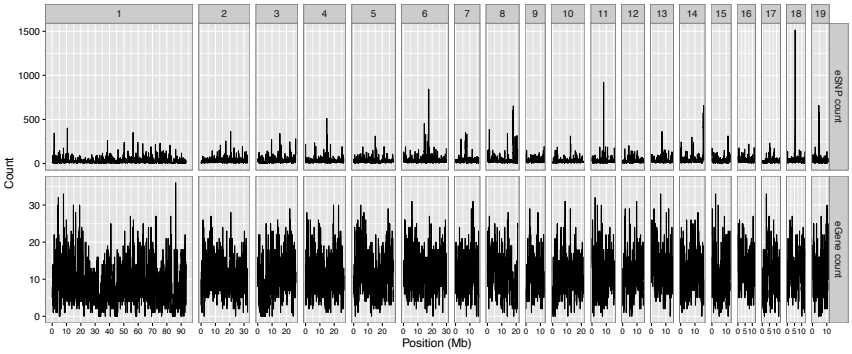

Figure S7. Sliding window approach to detect eQTL hotspots. The upper panel shows the number of eSNPs in genomic windows of 100 kb for each of the 19 chromosomes. The lower panel shows the number of unique genes that are associated to each genomic window (nominal eQTL p-value < 1e-6).

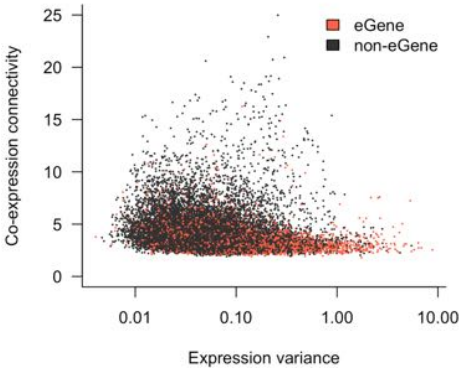

Figure S8. Gene expression variance plotted against co-expression network connectivity. eGenes are indicated by red points.

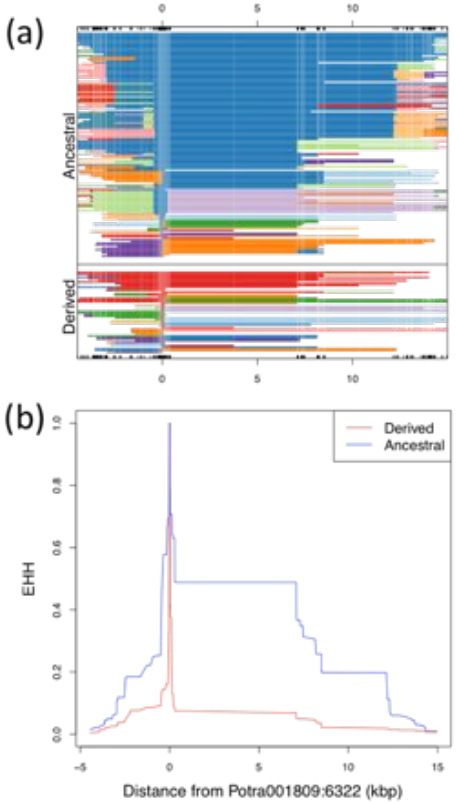

Figure S9. The decay of haplotype homozygosity in the eSNP (Potra001809:6322) with the highest |iHS| value. (a) shows the decay of haplotypes in single regions near the eSNP (Potra001809:6322) for both ancestral and derived alleles. Adjacent haplotypes with the same color carry identical genotypes everywhere between the eSNP and the candidate site. Haplotypes are no longer plotted beyond the points at which they become unique. (b) Extended haplotype homozygosity (EHH) plot for the eSNP. EHH of the ancestral allele (blue curve) is much higher than the EHH of the derived allele (red curve), suggesting that the haplotypes with ancestral allele were the targets of selective sweeps.

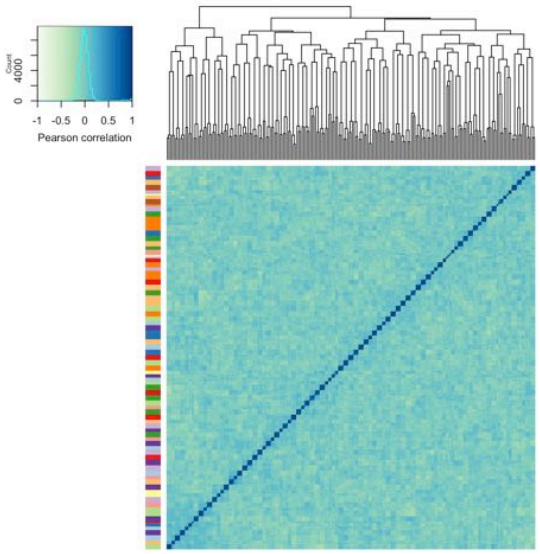

Figure S10. Heatmap representing the sample clustering based on gene expression values after hidden confounder removal. The 500 most variable genes in the original expression data were used for calculating the sample correlations (i.e. the same genes as in Figure 1D). The colour bar represents the population of the samples with the same colour scheme as in Figure 1D. The small clusters on the diagonal represents genotype replicates.

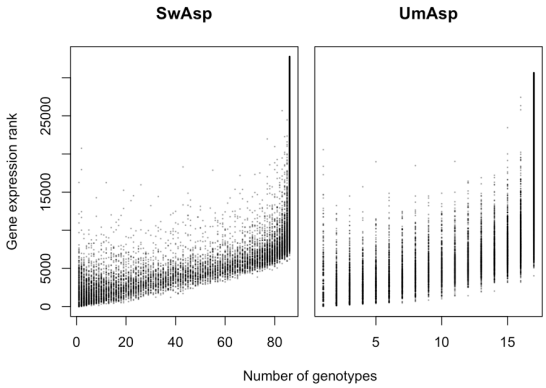

Figure S11. Relationship between gene expression ranks and the number of genotypes with detectable expression of the genes. The number of genotypes that a gene was expressed in was determined by counting the number of genotypes with non-zero expression for each gene. The gene expression ranks were calculated by ranking the mean gene expression values where the mean was calculated only considering samples with non-zero expression.

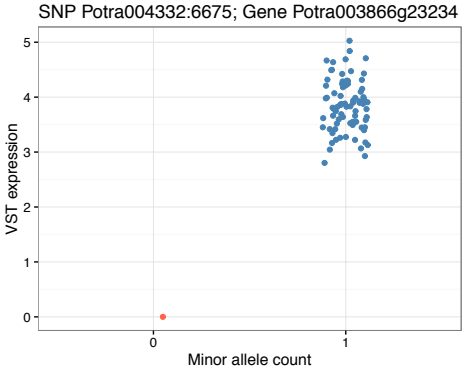

Figure S12. Motivating example for major genotype filtering. In this case the SNP is heterozygous in all samples but one, resulting in a minor allele frequency close to 0.5, but a major genotype frequency close to one. The resulting association turns out very significant, but is only supported by a single sample.

Figure S13. Representative photo of the sampled bud flush stage.

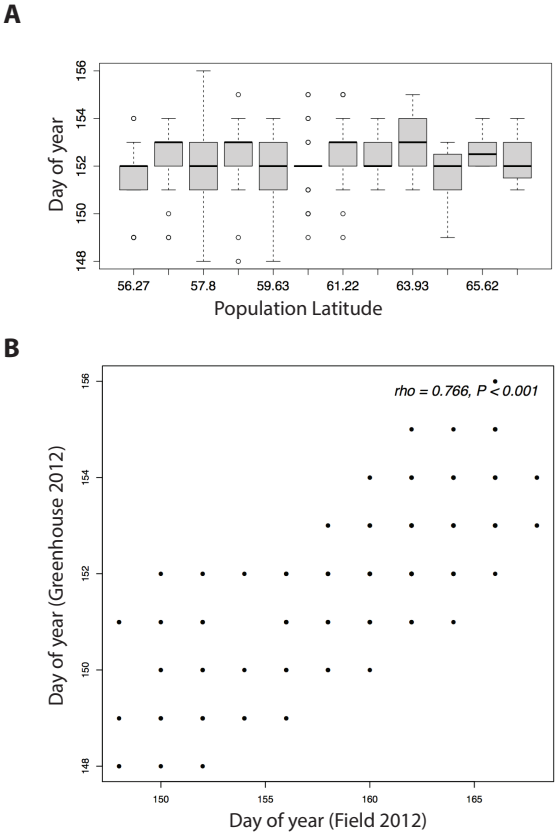

Figure S14. (A) Box plot distributions of the Julian day of sampling for the SwAsp sub-populations. (B) The relationship between Julian day of bud flush for the greenhouse sampled buds and Julian day of bud flush in the field for the same year (2012).

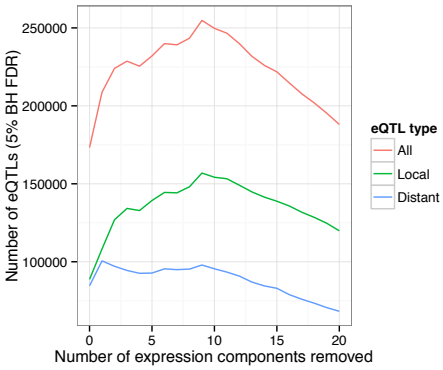

Figure S15. The number of eQTLs detected as a function of the number of principal components regressed out of the gene expression data.

Paper IV

A systems genetics approach to understanding the control of natural variation of leaf morphology in European aspen

Kathryn M Robinson[+,1], Niklas Mähler[+,2], Barbara K Terebieniec[1], Torgeir R Hvidsten[1,2], Nathaniel R Street[*1]

[*] Corresponding author: Nathaniel R Street: nathaniel.street@umu.se
[+] Equal contribution

[1] Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, 901 87 Umeå, Sweden
[2] Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, 1430 Ås, Norway

Abstract

Background: A central aim of biological investigation is to understand how genomes encode information controlling emergent, complex phenotypes and the genetic architecture underlying natural variation of such traits among individuals. There are still relatively few studies exploring the genetic architecture of complex traits in natural populations with no consensus understanding of how genetic architecture is linked to the evolutionary history of a trait. Here, we focused on natural variation in leaf size and shape in a natural population of European aspen (*Populus tremula*) sampled across the distribution range of Sweden. Leaf morphology traits show no evidence of clinal variation of population differentiation, appearing to be selectively neutral. As such they serve as a useful contrast to studies performed on highly adaptive traits, such as the timing of autumn bud set.

Results: We assayed leaf size and shape variation in replicated common garden experiments of the Swedish Aspen (SwAsp) collection, finding no correlation between leaf shape and a range of climatic, geographic and biological traits and no evidence of population differentiation. We utilised a collection of genome-wide SNPs to perform a Genome-Wide Association Study (GWAS), identifying overlapping sets of SNPs from the replicated gardens/years in order to reduce false positives. These results were integrated with RNA-Sequencing data assaying gene expression in the SwAsp collection that has previously been used to map expression QTL (eQTL). Using these data, we identified SNPs associated to leaf physiognomy phenotypes, to gene expression (eQTL) and correlations between phenotype and gene expression levels.

Conclusions: The genetic architecture underlying natural variation in leaf morphology traits within the study population was in agreement with the infinitesimal model, with individual SNPs explaining little of the heritable trait variation. There was no evidence of correlation between gene expression and the phenotypic traits considered, although overlapping sets of morphological and expression associated SNPs were identified. We conclude that variation among genotypes in leaf morphology is controlled by the effect of many SNPs, each of small effect, each of which results in a small-scale modulation of expression patterns contributing to the control of the leaf developmental program.

Keywords: Leaf shape; RNA-Sequencing; transcriptomics; Populus tremula; association mapping, natural variation, candidate gene

Introduction

For many years the accepted dogma was that variation between species, and genetically controlled aspects of variation among individuals of the same species, resulted largely from non-synonymous changes in the protein coding sequence of genes: Resultant changes in protein structure/function then perpetuated a change in phenotype. Our understanding of the causal factors underlying natural variation and speciation has been revolutionised during the past decade and, most recently, is being further refined by discoveries arising from the application of high throughput sequencing approaches. For example, genomics has provided extensive evidence that divergence in regulation and expression network structure are key components of both within species variation and divergence between species [1–8]. This has required a change in emphasis from the identification of causal polymorphisms within the protein coding regions of genes to that of identifying sequence variations that modify gene regulation and expression. The ability to determine whether the source of such regulatory diversity derives from polymorphisms lying in *cis* (local effects) or in *trans* (distant effects) additionally enhances our understanding of the genetic architecture of gene expression diversity [4, 9].

While reductionist molecular biology approaches have taught us much about the function and role of numerous individual genes, we still know relatively little about the mechanisms underlying natural variation and how interacting networks of genes result in the emergent properties of phenotypes [9]. Complex polymorphic traits are not the result of genes acting independently but rather are emergent properties of a polygenic, dynamic system of interactions among genes and between genes and the environment [10]. While genetic approaches have provided insight into the genetic architecture of complex traits, knowledge of the causal genetic polymorphisms has remained limited. For example it is not known if the majority of causative polymorphisms lie in protein-coding, promoter, intron or inter-genetic regions of the genome, how often the control of polymorphic traits is determined by *cis* (*i.e.* proximal) or *trans* (*i.e.* distal) effects and how these patterns relate to selection and adaptive trait variation.

Partly driven by the above questions and by the availability of new technologies, the previous decade saw an explosion of interest in genome-wide association studies (GWAS), particularly in the field of human medical research. Such studies aim to identify causative genetic polymorphisms contributing to the control of complex quantitative phenotypes (for example height or obesity). Results from the numerous studies performed have, perhaps, been less insightful than hoped [11]. For the majority of traits with moderate to high heritability considered to date, potentially causative SNPs explain little of the phenotypic diversity that exists and often lack biological interpretation [12–14], leading many to ask where the 'missing heritability' lies [15]. As a result, attention is shifting towards approaches that integrate multiple forms that can both identify functionally important SNPs and provide biological context or insight as to their mode of action. These approaches also represent a means to minimise wasted effort on false-positives, which can be particularly important considering the substantial effort invested in functional validation of identified candidates. Integrative approaches include various combination of the relationship between SNPs and eQTL, differentially expressed genes and constructed expression network structure, phenotypic trait correlations and association mapping or QTL results (for example [16–20]).

Systems genetics formalises such integrative approaches as the study of systems biology within a population genetics context [9]. As an analogy, GWAS provides a two-dimensional view of a system: Adding gene expression data from the same individuals transforms this view into a three-dimensional one, allowing previously hidden properties to be seen. In contrast to many previous systems biology approaches where, for example, gene expression networks are constructed from

a diverse set of experimental conditions, systems genetics profiles expression among individuals of a population. An expression network is then constructed based on the correlation structure of expression variation across the population. In this way, hubs in the network structure represent key components determining variation within the population. Two key aspects of the systems genetics approach are that it places emphasis on the fact that complex traits are not the result of genes acting independently but rather are emergent properties of a dynamic system of interactions between genes and the environment and that the link between phenotype, expression and genomic variants can provide functional insight when identified variants lie outside of protein coding regions [10].

As model systems for genetic studies, plants have a number of distinct advantages over animals: Clonal propagation results in the ability to precisely calculate heritability, phenotypic plasticity, genotype and environment effects and their interaction (GxE) within a single generation [15]. To date the plant field has focused largely on candidate gene based association studies. A major limitation of such an approach is that only genes for which there is pre-existing knowledge are considered. An assumption is also made that those genes are involved in the determination of natural variation. Additionally, a functional candidate gene approach is typically combined with a screen for SNPs within coding regions. This is counterintuitive as the functional evidence used to identify candidate genes is often the presence of phenotypic variation associated with differences in gene expression – and expression differences are rarely caused by SNPs within coding regions. A substantial risk of such an approach is falsely concluding that a gene does not play a role in controlling trait variation simply because no SNPs within the coding region associate with the phenotype of interest. This risk is especially high in species with rapid decay of linkage.

European aspens have been shown to contain extremely high levels of genetic variation [21–23], to have no significant population structure ([24]; Wang *et al.* 2016 in prep) and to be suitable for high-resolution association mapping [25]. *P. tremula* can also be grown in tissue culture and is amenable to genetic transformation, meaning that functional confirmation can be generated in the genetic background studied. In this context, variation in leaf shape of European aspen represents an excellent model system as we have previously shown that many leaf physiognomy traits have high levels of heritable variation [26]. Leaves are the direct energy source sustaining the majority of complex life on earth. Humans interact with and recognise plant species largely through the shape of their leaves and, as such, leaf shape forms an important component of the relationship we share with the living world. Indeed, leaf shape was historically one of the key features used by Linnaeus and others to classify and identify species and long-term historical changes in leaf shape recorded in the fossil record provide insight into historical climatic conditions, subsequently allowing extrapolation of past trends to future changes in response to climate change. Leaf shape varies distinctly both between and, often, among species with some identifiable global trends such as the narrowing and more defined serration of leaves toward latitudinal extremes [27].

Here, we have taken a systems genetics approach using the Swedish Aspen collection [28] to investigate variation in a number of traits associated with leaf shape and size. These traits were selected to serve as a contrast to other highly adaptive traits that we have also been investigating (Wang *et al.* 2016 in prep) as leaf shape shows no evidence of clinal variation and has likely not been a target of positive selection. We have integrated population-wide genomic resequencing and RNA-Sequencing data together with morphological phenotypic traits to explore the genetic architecture of leaf shape variation.

Results

Phenotypes

We sampled mature, pre-formed leaves in the Swedish Aspen (SwAsp) collection at two common garden experiments in the north (Sävar, Västerbotten) and south (Ekebo, Skåne) of Sweden in two years (2008 and 2011). There was considerable variation in leaf shape represented among the SwAsp genotypes (Figure 1A) and we used a digital image analysis method [26] to measure a number of traits indicative of leaf physiognomy (Figure 1B). We calculated clonal repeatability as an upper-bound estimate of broad sense heritability (H^2), which was relatively high for all traits and for which shape traits (including circularity and measures of indent size and number) had distinctly higher values than those related to size (measures of length, width and area), accounting for the bimodal distribution observed in Figure 1C (see supplementary file S1 for details of all traits). We calculated Q_{ST}, which revealed no evidence of population differentiation for any of the traits considered (Figure 1D, supplementary file S1). As many of the dimension traits have a high degree of redundancy, we performed a dimension reduction using principle component analysis (PCA), for which PC1 was related primarily to size variation with PC2 and PC3 relating to components of shape (Figure 1E). ANOVA tests comparing traits and PCs in the two gardens/years revealed a variable degree of environmental variance (supplementary table S1), with size traits and PC1 having a greater variation than shape traits and PCs (consider the F values in sheet 1 of supplementary file S1). Although significant GxE was observed for a number of both size and shape traits, the percentage of variance explained by GxE (as indicated by F values) was considerably smaller for the shape related traits. As such, leaf shape is under tight genetic control and is a relatively invariant feature of a genotype. This is confirmed by the higher correlation between shape trait values for the two gardens/years than for size related traits (Figure 1F).

Genome wide association mapping

We performed a genome wide association (GWA) analysis to identify links between single nucleotide polymorphism (SNP) variants and indent width, leaf area, and leaf circularity for all of the four datasets. A total of 4.5 million SNPs (detailed in Wang et al. 2016, In prep) were associated to the phenotypes separately for each garden and year. After multiple testing correction and combining all significant associations from all four datasets, 65 SNPs were significant for indent width, 39 for leaf area and no SNPs were significant for circularity at a 5% false discovery rate (FDR; Figure 2). The significant SNPs for indent width had a wide range of contexts from coding SNPs to non-coding and intergenic SNPs, but judging from their location, many of them are likely to be linked. 44 of the 65 indent width SNPs were located within 2 kbp of a total of 15 genes, and all but one of the SNPs with a significant association to leaf area were located within 2 kbp of 5 genes. Neither of these gene sets were enriched for any GO terms.

To identify a set of the most consistent SNPs in order to minimise false positives, we selected the top 1000 SNPs (ranked by adjusted p-value but disregarding significance) for each trait and garden/year (figure 2). We refer to these SNPs as phenotypic trait associated SNPs or simply pSNPs. The genomic context of pSNPs identified in at least two gardens/years are shown in Figure 3 We also identified the set of genes associated with the pSNPs (genes located within 2kb of the SNP) present in at least two of the gardens/years (Materials and methods): 684 SNPs were associated with 252 genes for circularity (enriched in inorganic anion transport, GO:0015698, $p = 0.005$), 715 SNPs with 236 genes for indent width (enriched in hexose metabolic process, GO:0019318, $p = 0.007$), and 66 SNPs with 43 genes for leaf area (enriched in sucrose metabolic process, GO:0031324, $p = 0.0003$).

Gene expression

After filtering gene expression values on variance and adjusting for hidden confounders (Materials and methods), correlations were computed between the expression of 22,306 genes and the three morphological traits for each garden and year. After correcting for multiple testing, no genes were significantly correlated to any of the traits at 5% FDR (Figure 4). We therefore investigated the top 1000 genes from each garden/year (ranked by adjusted p-value but disregarding significance) and refer to these as expression-phenotype correlated genes or epGenes. Overlaps between the epGenes from each garden/year were relatively high (Figure 4). A GO enrichment analysis showed that the intersection of genes correlating with circularity was enriched for "ATP biosynthetic process" (GO:0006754; 31 genes, $p = 0.008$), intersecting genes correlating with indent width were enriched for "cellular protein localization" (GO:0034613; 7 genes, $p = 0.0018$), while no significant GO terms were found for leaf area.

Gene set enrichment analysis (GSEA)

Since no individual genes displayed expression that correlated significantly with any of the phenotypes, we employed gene set enrichment analysis (GSEA, [29]) to test whether any *sets of genes* were significantly enriched at the extreme ends of the gene list, *i.e.* the list of all 22,306 genes sorted by correlation to a phenotype (Materials and methods).

For each of the three phenotypes, we tested two types of gene sets (Tables 1-3 for circularity, indent width and leaf area, respectively). Firstly, we tested genes associated with the pSNPs that were found in all, or at least two, of the gardens/years (see Venn diagrams in Figure 2). The only significant set from this analysis was the genes associated with the 98 pSNPs discovered for circularity in all gardens/years ($p = 0.027$). Secondly, we tested genes with a common functional role (Gene Ontology annotation). Here we found several significant associations including GO terms related to amino acid biosynthesis and transport, as well as DNA repair and carbohydrate metabolism

Data integration

We combined data on pSNPs, epGenes, and eQTLs (referred to as pairs of eSNPs and eGenes, see Paper III) to identify genes and genetic variants that were associated with leaf phenotypes. For each phenotype, we first intersected genes associated with pSNPs identified in at least two gardens/years, epGenes identified in at least two gardens/years and eGenes associated with significant eSNPs in the eQTL mapping. This resulted in three genes each for circularity and indent width, and none for leaf area. The three genes for circularity were Potra000998g08306, Potra001379g11776 and Potra009203g26307, annotated as disease resistance protein, synthase mitochondrial F1 complex assembly factor, and CASP-like protein, respectively. The three genes associated with indent width were Potra000351g01289, Potra163617g27107 and Potra000727g05700, all of which are unannotated. Potra163617g27107 does however have an ortholog in *Arabidopsis thaliana* that is annotated as being part of the mitochondrial outer membrane translocase complex (GO:0005742).

Next, we intersected the GSEA-significant GO categories in Tables 1-3 with the pSNPs and eQTLs. This revealed that several functional categories significantly associated to phenotypes through expression correlation also contained many eGenes. However, they contained very few genes with pSNPs and there was almost no overlap between these pSNPs and eSNPs.

Discussion

For more than a century there has been much unresolved and speculative discussion as to the evolutionary significance, if any, of leaf shape *per se* and of natural variation of leaf shape within populations [30, 31]. While global and historical trends in leaf shape have been identified and attributed to climatic conditions [27], there remains a rather spectacular paucity of evidence linking standing leaf shape variation to aspects of plant fitness [30]. In the present study we identified considerable genetic control of leaf physiognomy traits (Figure 1), in particular traits associated with leaf shape – indeed these are the highest values we have observed for any of the numerous phenotypes we have considered to date, regardless of whether or not those traits have evidence of being under positive selection ([32], Wang *et al,* 2016 in prep). There was considerable observable variation in leaf shape both within local populations and within the distribution range of aspen across Sweden, with no identified clinal trends or evidence of population differentiation (Figure 1, supplementary file S1). Such patterns of variation may indicate that variation in the trait is selectively neutral or that balancing selection is acting to maintain within population variation. Here, we identified very few significant associations between SNPs and phenotypic traits, with no significant associations being identified in more than one garden/year for any of the traits considered. Even when identifying overlap without considering significance but taking the top 1000 SNP associations per garden/year per trait (pSNPs, Figure 2), there was still relatively little overlap, although there were greater numbers of SNPs in common among two or more gardens/years for the traits with higher H^2. Although these overlaps were low, GO over-representation tests identified significant enrichment of categories for genes located within 2kb of the pSNPs, indicating that there was biologically functional meaning to the sets of genes, albeit without readily explainable functional interpretation. Taking the total variance explained for the sets of overlapping SNPs (data not shown) reveals that very little of the total genetically controlled variance was explained by identified SNP associations. If one assumes that SNPs of high effect size would have been detected as significant, this result suggests that there remain a very large number of SNPs of small effect size that, in combination, control variation in leaf size and shape – a finding in support of the infinitesimal model. This is an area where there is little consensus in the literature, with various studies reporting genetic architectures for complex traits spanning a range from few SNPs of large effect to many SNPs of small effect size. One likely explanation for these contrasting results is that genetic architecture contrasts depending on the extent to which a trait has been the target of positive selection, with strongly adaptive traits (such as bud flush for aspen within Sweden) more likely to be explained by a small number of SNPs of large effect (Wang *et al*. 2016, In prep). The case of leaf physiognomy in aspen appears to be far more similar to that of human height, for which there appears to be a vast number of small effect SNPs underlying the height variation among individuals within a population.

For the three traits considered, the greatest proportion of associated SNPs identified in at least two gardens/years were located in UTR regions and the fewest in intergenic regions (Figure 3). The distributions of the proportion of SNPs within exons, introns and flanking regions were less consistent between the three traits, although with so few SNPs considered it is hard to interpret whether these differences are meaningful. The presence of the largest proportion of SNPs within UTRs is likely indicative of these acting to modify expression, suggesting that gene expression variation should be associated with variation in these traits. We utilised a resource profiling gene expression from winter buds undergoing induced spring bud flush in controlled conditions in an attempt to identify gene expression variation associated with trait variance. As for pSNPs, there were relatively few consistent correlations present between gene expression levels and phenotypic traits (Figure 4), none of which were significant. Similar to pSNPs, a greater degree of overlap among gardens/years was observed for the more heritable traits (circularity and indent width) compared to leaf area; when considering the top 1000 expression-phenotype correlated genes

(epGenes). In the case of circularity and indent width, significant over-representations for GO categories was present, suggesting biologically relevant functional links for these genes. In contrast, no such signal was detected for leaf area, which is likely due to the lower heritability and greater degree of environmental variability for leaf shape, possibly suggesting that different mechanisms, and therefore sets of genes, influenced leaf area at the two gardens and in the two years.

The power of systems genetics lies in the integration of expression, eQTL and phenotypic GWAS results. In our case such integration proved to be of minimal value due to the low number of trait-SNP and trait-expression associations detected. Of the three genes linked for all data types (overlap between genes close to pSNPs, epGenes and eGenes) for circularity and indent width, none of these are known to function in the control of leaf shape or to have any role in leaf development. However, these do represent our best leads for further downstream investigation.

The infinitesimal model presents a number of analysis challenges. The first is that it can be extremely challenging to identify statistically significant associations between causal SNPs (true positives) and phenotypes, especially in cases such as ours where the millions of SNPs considered creates something of a multiple testing nightmare. Also, applying methods that employ a one gene or one SNP at a time strategy to explain traits will likely prove futile. A more realistic approach is to utilise gene set enrichment analysis (GSEA, [29]), where one can identify sets of genes that together correlate significantly with a trait, despite there being no significant individual genes. Here, we utilised the list of genes sorted by expression correlation to traits to test whether gene sets (1) located close to discovered pSNPs (i.e. top 1000 SNPs identified in at least two gardens/years) or (2) with a common functional role (*i.e.* genes annotated with the same Gene Ontology category) displayed such properties (Tables 1-3). The former (1) identified only one significant association between circularity and the 98 pSNPs discovered for circularity in all gardens/years (Figure 2). The latter (2) revealed several functional categories with significant associations to traits. Moreover, several of these categories included high numbers of genes with associated eQTLs (eGenes). However, the same categories contained very few genes located close to phenotypic trait associated pSNPs, and very few of these pSNPs were in turn associated with eQTLs (i.e. few pSNPs were also eSNPs). Thus, although we identified some gene sets for which expression were significantly associated with traits and that many of these genes also had mapped eQTL associations, there was a very low correspondence between gene expression and genome sequence driven discoveries. In part, such low correspondence will be due to the fact that each SNP affecting gene expression is of small effect size (Paper III) and that expression variation of an individual gene contributing to the control of the phenotypic traits also explains only a small fraction of the total phenotypic variation. Another important factor relates to the fact that we have only a single snapshot measure of gene expression and that an eQTL can only be identified for this snapshot: another snapshot measure would be expected to identify additional SNPs associated with expression variance.

It may seem somewhat paradoxical that pSNP association results suggest that gene expression is likely to be the primary mechanism driving leaf shape variation (inferred from genomics context, Figure 3) but that we identified so few correlation links between expression and the traits, and furthermore that so few of the pSNPs were eSNPs. However, there are a number of possible explanations for this lack of correspondence, some of which are already alluded to above. Although the molecular mechanisms that give rise to leaf shape variation remain entirely unknown [30], leaf development is a temporal process with many temporally separated components, variation in any of which will contribute to variation in final leaf form. As such, a single snapshot of gene expression during this developmental program will be insufficient to capture all relevant links between gene expression and final leaf form. To overcome this limitation would require a time series sampling strategy, with samples being collected at multiple developmentally equivalent stages for all individuals in the population and with eQTL mapping being performed for the

snapshot measures of expression of all genes at each time point as well as for expression trajectories – a strategy that is currently intractable (or at least not fundable). In the current case of profiling expression from leaves that are produced and developmentally arrested in winter buds there are added complications. It is very likely that the period before bud set comprises an important developmental period where much of the pattern formation leading to final leaf shape is occurring. However, as genotypes set bud at different times and as climatic and environmental conditions vary considerably during the period in which bud set occurs, collecting developmentally equivalent samples of all genotypes would be extremely challenging. Before bud set the problem is also compounded by the fact that within a single bud the multiple leaves produced will be at different developmental stages. As such there would be a very high chance that the averaged expression snapshot that would be obtained from extracting RNA from a whole bud would negate any meaningful developmental expression signature (the Simpson's paradox). Even after bud set, when all leaves are arrested at an equivalent developmental stage, similar problems exist at the leaf level as leaf development also varies spatially, for example with cell production being more prolific and continuing for longer at the leaf base than at the tip. Here again, a single sample from an entire leaf will potentially mask or negate the ability to associate expression variation to phenotypic traits. In both cases, spatially resolved expression profiling, for example using laser capture microdissection, would offer a solution although performing such an experiment for replicated samples all individuals in a population would be a daunting undertaking and would likely suffer a multitude of confounding technical factors. New or improved sampling techniques and expression profiling approaches combined with falling costs of generating expression data will help to overcome these limitations or to make more comprehensive experimental designs feasible within the future, however they will not overcome the fact that the signature of SNP to expression to phenotype will be weak and hard to detect for any trait following the infinitesimal model.

Our results have shown that variation in the shape of pre-formed leaves in European aspen is controlled by numerous SNPs each of very small effect. These SNPs were primarily located in UTR regions of genes, suggesting that they induce variation in leaf form through modulation of gene expression patterns. However, very few of these coincided with eQTL and there was no significant correlation detected between a snapshot of gene expression during spring bud flush and the phenotypic traits considered.

Our findings highlight the challenges faced when employing a systems genetics strategy, especially for traits controlled according to the infinitesimal model. As such, although the approach holds great potential for providing functional insight to the link between SNPs and phenotypic trait variation, the biological characteristics of the study system may present severe limitations to this potential.

Materials and Methods

Leaf shape phenotyping

Leaf size and shape parameters were measured in a natural population of *Populus tremula*, the Swedish Aspen (SwAsp) collection, growing in common gardens at Sävar, northern Sweden (63.9°N, 20.5°E) and Ekebo, southern Sweden (55.9°N, 13.1°E). The common garden trials comprised of natural (wild-growing) aspen genotypes collected in 2003 across ten latitudinal degrees, which were cloned and planted in 2004 in a randomised block design in each garden [28]. Leaf samples were harvested in Sävar on 14 July 2008 and 28 June 2011 and in Ekebo on 18 July 2008 and 4 August 2011, when leaves were fully expanded and mature, but prior to bud set and before the occurrence of substantial damage due to herbivory or the presence of fungal rust infection. Ten undamaged leaves per replicate tree were sampled randomly across the canopy, avoiding leaves from the first or last leaf in a leaf cohort originating from a single bud. In total, in 430, 444, 326 and 393 trees were sampled in Ekebo 2008, Ekebo 2011, Sävar 2008 and Sävar 2011 respectively, comprising between 1 and 8 (median = 3) clonal replicates. One hundred and thirteen genotypes were sampled in both years in Ekebo and in 2011 in Sävar, and 111 genotypes were sampled in 2008 in Sävar. Leaves were stored at 4° - 8°C immediately after harvest. Petioles were removed at the leaf base and the sample of ten leaves per tree was scanned in colour at 300 dpi using with a CanoScan 4400F. A 5x4cm Post-it note was scanned as a scale image. The resulting images were analysed using LAMINA [26] to obtain leaf size and shape metrics (see supplementary file S2 for a list and descriptions). Median values of the ten leaves per tree were calculated for each leaf size and shape metric. Principal components analysis (PCA) was employed using the prcomp function within the R programming environment [33] for the size and the shape trait sets separately and in combination (see supplementary file S2 for the classification of obtained traits as either size or shape related). The first three principal components for leaf shape (PC1, PC2 and PC3) were used as unique leaf shape phenotypes and the first principal component for leaf size metrics (PC1) was used as a summary phenotype of leaf size. The first three components of the entire data set (Size and shape PC1, 2 and 3) were used to summarise all leaf phenotypes into three reduced descriptors. The PCA loadings are provided in supplementary file S2. For genome-wide association mapping, median values of each leaf shape and size phenotype were calculated for each aspen genotype for which there were three or more clonal replicates.

Statistical analyses

All statistical analyses were conducted in R. Phenotypic data were examined for homogeneity of variance. No data transformations were required to meet the assumptions of a normal distribution. Pearson correlations were used for all phenotypic correlations calculated.

We calculate clonal repeatability (R) and used this to provide an upper-bound estimate or broad sense heritability (H^2 – see materials and methods). We refer to this trait as H^2 rather than R as this probably allows a more intuitive interpretation for readers, however we not that the two are not the same (see (Dohm, 2002 for dicussion). Estimates of broad-sense heritability (H^2) and their 95% confidence intervals, including all clonal replicates, was calculated as

$$H^2 = V_G / (V_G + V_E)$$

where V_G and V_E are genetic and environmental variance components, using the heritability function in the R package 'Heritability'. To estimate population differentiation, QST , the following formula was used:

$$Q_{ST}=V_{pop}/(V_{pop}+(2*V_{geno}))$$

where V_{pop} is the population and V_{geno} is the genotype genetic variance components.

Genetic correlations between phenotypes were calculated as

$$r_{G(AB)} = V_{G(AB)}/ \sqrt{(V_{G(A)} \times V_{G(B)})}$$

where $r_{G(AB)}$, the genetic correlation of phenotype A and phenotype B, was calculated from the $V_{G(AB)}$, the genetic covariance in phenotype A and phenotype B, $V_{G(A)}$ and $V_{G(B)}$ were the genetic variances of phenotypes A and B respectively.

Genetic (clonal) variation for each phenotype between years and common gardens were investigated using separate analyses of variance (ANOVA) models where phenotype was the dependent variable. Analyses were conducted only on genotypes with three or more replicate trees per genotype. To examine common garden effects in the same year, garden, genotype, and their interaction were considered independent variables in the following models:

$$Phenotype_{2008} \sim Garden + Genotype + Garden \times Genotype$$

$$Phenotype_{2011} \sim Garden + Genotype + Garden \times Genotype$$

where $Phenotype_{Year}$ indicates that the analysis was conducted on the phenotypic data from one year to compare the two gardens. To compare the phenotypic data from the two gardens in one year, the phenotypic response for a given garden (Ekebo or Sävar) was partitioned into variance of the independent variables year and genotype and their interaction:

$$Phenotype_{Ekebo} \sim Year + Genotype + Year \times Genotype$$

$$Phenotype_{Sävar} \sim Year + Genotype + Year \times Genotype$$

where $Phenotype_{Garden}$ indicates phenotypic data were taken from only one garden in each model. ANOVA models were implemented in the aov function in R. All effects were considered significant at $P<0.05$.

RNA-Sequencing data

The RNA-Seq data used in this study has been described previously (Paper III). It consists of 219 samples distributed among 86 distinct genotypes. The same type of gene expression filtering and adjustment were used in this paper as in Paper III. Genes were required to have an expression variance above 0.05, and the first nine gene expression principal components were regressed out from the data. This left 22,306 genes for further analysis. The data has been uploaded to the European Nucleotide Archive (ENA) with accession number ERP014886.

Genome wide association mapping

A total of 4.5 million SNPs were considered for the GWA, previously described in Wang *et al.* (2016, in prep). A univariate linear mixed model was applied to the data using GEMMA [35] and included the first principal component based on independent SNPs as a random effect in order to account for population structure (Wang *et al.* 2016 in prep) as well as the built-in estimation of a relatedness matrix. GEMMA produces different statistics for significance, and in this study we used p-values based on a likelihood ratio test. These p-values were consequently Benjamini-Hochberg adjusted for multiple testing for each garden and year separately using the p.adjust

function in R. To associate genes with SNPs, the v1.0 *Populus tremula* annotations from the PopGenIE.org web resource were used [36], and any gene within 2 kbp of a SNP were said to be associated with that SNP.

Gene set enrichment analysis

Gene set enrichment analysis was implemented in R according to [29]. In short, a gene set was tested for significant association to a phenotype based on gene expression correlation. The expression profile of each of the 22,306 genes were correlated to the phenotype and sorted by correlation (from positive to negative values). A running sum was produced from the top of the list by adding the correlation to the sum if the gene is part of the gene set and subtracting if it is not. The test statistic is then the maximum absolute value of this running sum. This value is also used to represent the leading edge of the gene set, i.e. the genes that contribute the most to the enrichment. "Geneset %" and "Total %" in Tables 1-3 represents the portion of the data that contribute the most to the enrichment (the leading edge subset). A larger value of "Geneset %" indicates that a large portion of the genes in the gene set contribute to the enrichment, and a small value of "Total %" indicates that the leading-edge subset is tightly clustered at one extreme of the correlation distribution. Significance was determined by a permutation strategy where the phenotype sample labels were permuted. This process was repeated 1000 times, and the fraction of permuted tests that had a higher score than the score from the original data (one-sided test) was used as the p-value for the enrichment. The p-values for GSEA based on GO gene sets were not adjusted for multiple tests due to the high level of dependence in the data. Therefore these p-values should be interpreted with care, but can be used as a relative ranking metric of the different gene sets.

Gene Ontology enrichment

The R-package topGO (http://bioconductor.org/packages/release/bioc/html/topGO.html) was used to perform GO enrichment analysis. In all cases the background used for the enrichments were the set of expressed genes (22,306 genes), and the classic test was used with the Fisher test statistic. In order for a GO term to be considered enriched, the gene set tested had to contain at least two genes annotated to that particular term, regardless of p-value.

Acknowledgements

NRS is supported by the Trees and Crops for the Future (TC4F) project. The authors also would like to acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure (NGI), and Uppmax for providing assistance in massive parallel sequencing and computational infrastructure.

References

1. Carroll SB: **Endless forms: the evolution of gene regulation and morphological diversity.** *Cell* 2000, **101**:577–580.

2. Brem RB, Yvert G, Clinton R, Kruglyak L: **Genetic dissection of transcriptional regulation in budding yeast**. *Science (80-)* 2002, **296**:752–755.

3. Ayroles JF, Carbone MAA, Stone EA, Jordan KW, Lyman RF, Magwire MM, Rollmann SM, Duncan LH, Lawrence F, Anholt RRH, Mackay TFC: **Systems genetics of complex traits in**

Drosophila melanogaster. *Nat Genet* 2009, **41**:299–307.

4. Mackay TFC, Stone EA, Ayroles JF: **The genetics of quantitative traits: challenges and prospects**. *Nat Rev Genet* 2009, **10**:565–577.

5. Liao B-Y, Weng M-P, Zhang J: **Contrasting genetic paths to morphological and physiological evolution**. *Proc Natl Acad Sci* 2010, **107**:7353–7358.

6. Hines HM, Papa R, Ruiz M, Papanicolaou A, Wang C, Nijhout HF, McMillan WO, Reed RD: **Transcriptome analysis reveals novel patterning and pigmentation genes underlying Heliconius butterfly wing pattern variation**. *BMC Genomics* 2012, **13**:288.

7. Richards CLC, Rosas U, Banta J, Bhambhra N, Purugganan MMD, Gibson G, Pigliucci M, Schlichting C, Jones C, Schwenk K, Leakey A, Ainsworth E, Bernard S, Markelz C, Ort D, Kammenga J, Herman M, Ouborg N, Johnson L, Breitling R, Chapman M, Leebens-Mack J, Burke J, Richards CLC, Hanzawa Y, Ehrenreich I, Purugganan MMD, Anderson JJ, Mitchell-Olds T, Boer T De, et al.: **Genome-Wide Patterns of Arabidopsis Gene Expression in Nature**. *PLoS Genet* 2012, **8**:e1002662.

8. Wang RL, Stec A, Hey J, Lukens L, Doebley J: **The limits of selection during maize domestication.** *Nature* 1999, **398**(March):236–239.

9. Civelek M, Lusis AJ: **Systems genetics approaches to understand complex traits.** *Nat Rev Genet* 2013, **15**:34–48.

10. Swami M: **Systems genetics: Networking complex traits**. *Nat Rev Genet* 2009, **10**:219–219.

11. Hardy J, Singleton A: **Genomewide Association Studies and Human Disease**. *N Engl J Med* 2009, **360**:1759–1768.

12. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461**:747–753.

13. Goldstein DB: **Common Genetic Variation and Human Traits**. *N Engl J Med* 2009, **360**:1696–1698.

14. Cirulli ET, Goldstein DB: **Uncovering the roles of rare variants in common disease through whole-genome sequencing.** *Nat Rev Genet* 2010, **11**:415–425.

15. Ingvarsson PK, Street NR: **Association genetics of complex traits in plants.** *New Phytol* 2011, **189**:909–22.

16. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KCC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WOC: **A genome-wide association study of global gene expression**. *Nat Genet* 2007, **39**:1202–1207.

17. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavaré S, Deloukas P, Dermitzakis ET: **Population genomics of human gene expression.** *Nat Genet* 2007, **39**:1217–1224.

18. Veyrieras J-BB, Kudaravalli S, Kim SYY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK: **High-resolution mapping of expression-QTLs yields insight into human gene regulation.** *PLoS Genet* 2008, **4**.

19. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, Dermitzakis ET: **Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations**. *PLoS Genet* 2010, **6**.

20. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ: **Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS**. *PLoS Genet* 2010, **6**.

21. Ingvarsson PK: **Nucleotide Polymorphism and Linkage Disequilibrium Within and Among Natural Populations of European Aspen (Populus tremula L., Salicaceae)**. *Genetics* 2005, **169**:945–953.

22. Wang J, Street NR, Scofield DG, Ingvarsson PK: **Natural Selection and Recombination Rate Variation Shape Nucleotide Polymorphism Across the Genomes of Three Related Populus Species.** *Genetics* 2015, **202**:1185–1200.

23. Wang J, Street NR, Scofield DG, Ingvarsson PK: **Variation in linked selection and recombination drive genomic divergence during allopatric speciation of European and American aspens**. *Mol Biol Evol* 2016:msw051.

24. Hall D, Luquez V, Garcia VMM, St Onge KRR, Jansson S, Ingvarsson PKK: **ADAPTIVE POPULATION DIFFERENTIATION IN PHENOLOGY ACROSS A LATITUDINAL GRADIENT IN EUROPEAN ASPEN (POPULUS TREMULA, L.): A COMPARISON OF NEUTRAL MARKERS, CANDIDATE GENES AND PHENOTYPIC TRAITS.** *Evol Int J Org Evol* 2007.

25. Ingvarsson PKK, Garcia V, Luquez V, Hall D, Jansson S: **Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 locus in European aspen (Populus tremula, Salicaceae).** *Genetics* 2008, **178**:2217–2226.

26. Bylesjö M, Segura V, Soolanayakanahally RY, Rae AM, Trygg J, Gustafsson P, Jansson S, Street NR: **LAMINA: a tool for rapid quantification of leaf size and shape parameters.** *BMC Plant Biol* 2008, **8**:82.

27. Peppe DJ, Royer DL, Cariglino B, Oliver SY, Newman S, Leight E, Enikolopov G, Fernandez-Burgos M, Herrera F, Adams JM, Correa E, Currano ED, Erickson JM, Hinojosa LF, Hoganson JW, Iglesias A, Jaramillo CA, Johnson KR, Jordan GJ, Kraft NJB, Lovelock EC, Lusk CH, Niinemets U, Peñuelas J, Rapson G, Wing SL, Wright IJ, Niinemets Ü, Peñuelas J, Rapson G, et al.: **Sensitivity of leaf size and shape to climate: global patterns and paleoclimatic applications**. *New Phytol* 2011, **190**:724–739.

28. Luquez V, Hall D, Albrectsen BR, Karlsson J, Ingvarsson P, Jansson S: **Natural phenological**

variation in aspen (Populus tremula): the SwAsp collection. *Tree Genet Genomes* 2008, **4**:279–292.

29. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102**:15545–15550.

30. Chitwood DH, Sinha NR: **Evolutionary and Environmental Forces Sculpting Leaf Development**. *Curr Biol* 2016, **26**:R297–R306.

31. Kidner CA, Umbreen S: **Why is Leaf Shape so Variable?** *Int J Plant Dev Biol* 2010, **4**:64–75.

32. Robinson K, Ingvarsson P, Jansson S, Albrectsen B: **Genetic Variation in Functional Traits Influences Arthropod Community Composition in Aspen (Populus tremula L.)**. *PLoS One* 2012, **7**:e37679.

33. **R: A language and environment for statistical computing** [http://www.r-project.org]

34. Dohm RM: **Repeatability estimates do not always set an upper limit to heritability**. *Funct Ecol* 2002, **16**:273–280.

35. Zhou X, Stephens M: **Efficient multivariate linear mixed model algorithms for genome-wide association studies.** *Nat Methods* 2014, **advance on**.

36. Sundell D, Mannapperuma C, Netotea S, Delhomme N, Lin Y-C, Sjödin A, Van de Peer Y, Jansson S, Hvidsten TR, Street NR: **The Plant Genome Integrative Explorer Resource: PlantGenIE.org**. *New Phytol* 2015, **208**:1149–1156.

Figures and Tables

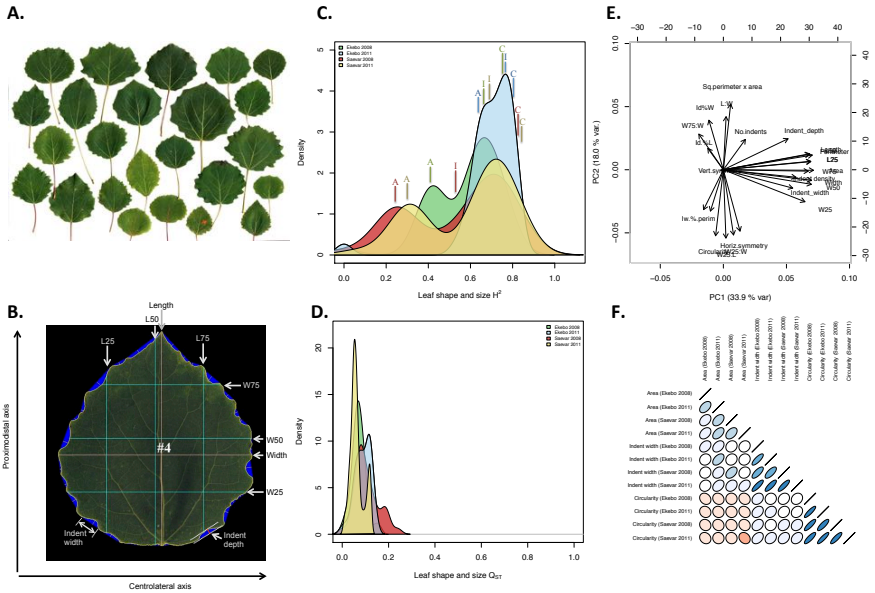

Figure 1 Leaf physiognomy variation in the Swedish Aspen collection. (A.) Natural genetic variation of physiognomic traits is apparent in a selection of the SwAsp genotypes. (B.) Leaf size metrics as measured by LAMINA (Bylesjö et al, 2008). Density distribution of (C.) heritability (H2) and (D.) QST values, with independent distributions for Ekebo 2008 (green), Ekebo 2011 (blue), Sävar 2008 (red) and Sävar 2011 (gold). Arrows on the H2 distributions indicate the H2 values for the phenotypic traits. A = leaf area, I = indent width, and C = circularity. Biplot (E.) of the first two components of principal component analysis of all size and shape phenotypes in all years and gardens. Correlation plot (E) showing the positive (blue) or negative (red) correlations between selected phenotypes in all years and gardens. Narrower ellipses represented higher correlation r values.

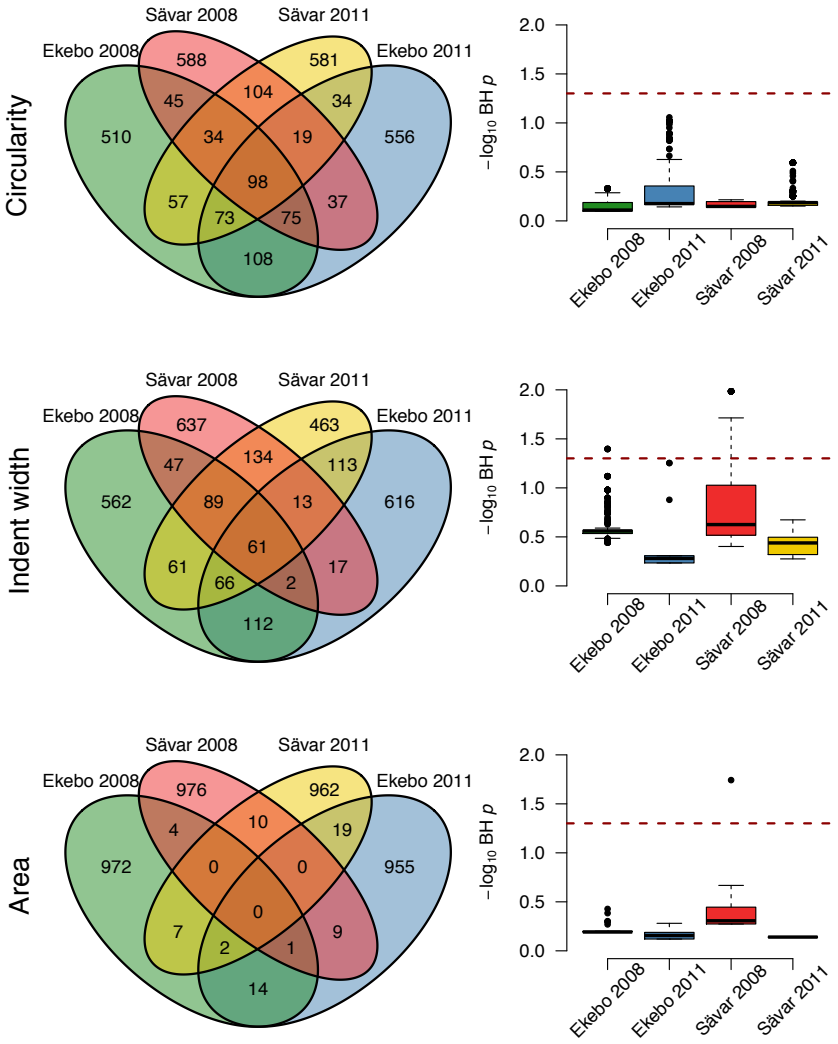

Figure 2. GWAS results for the three traits in the different gardens and years. The Venn diagrams show the top 1000 pSNPs based on p-value for each garden and year and how they overlap. The boxplots show the distribution of adjusted p-values for pSNPs that are found among the top 1000 SNPs in at least two of the gardens/years. The dashed line represents 5% FDR.

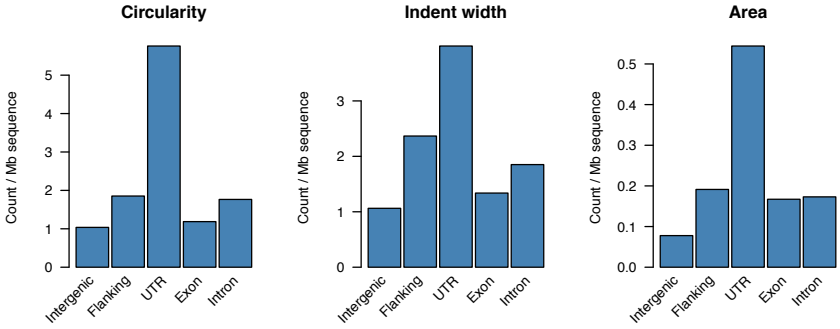

Figure 3. Genomic context of pSNPs found in at least two gardens/years (Figure 2) normalised by total feature length. In cases where SNPs overlapped several features, they were prioritised in the following way: UTR, exon, intron, flanking, intergenic. Exon counts should thus represent coding SNPs.

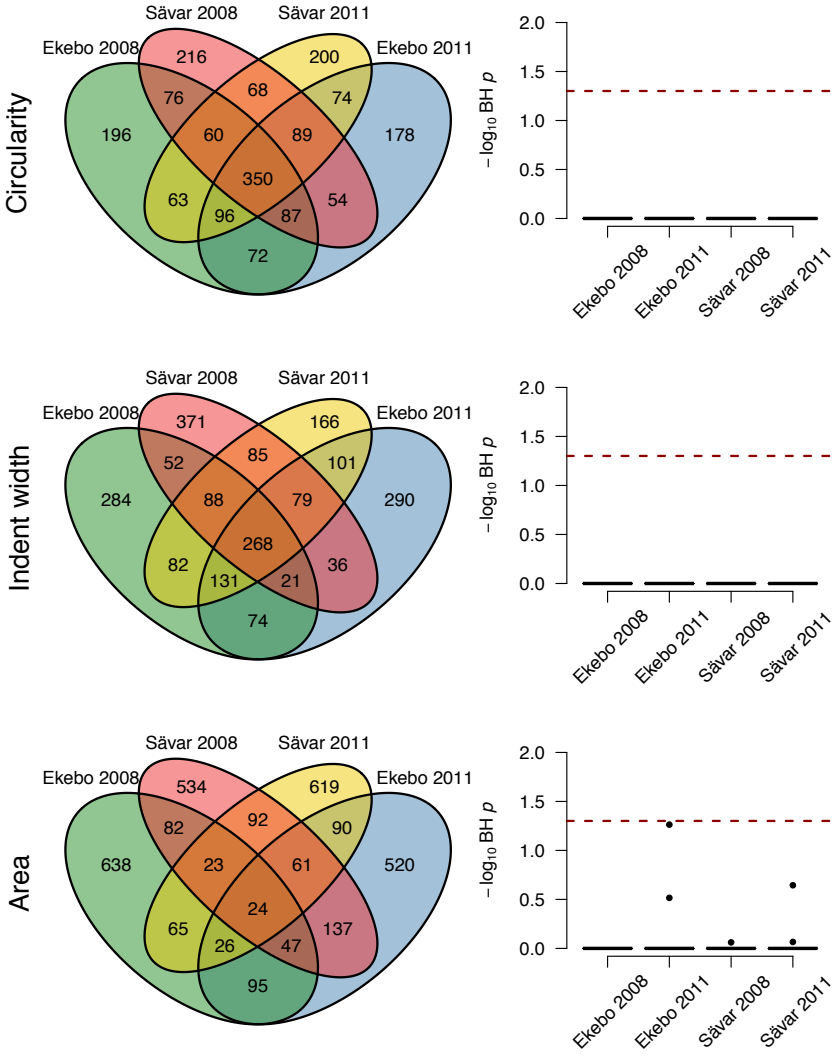

Figure 4. Correlation between gene expression profiles and the three traits for the different gardens and years. The Venn diagrams show the top 1000 epGenes based on Pearson correlation for each garden and year and how they overlap. The boxplots show the distribution of Benjamini-Hochberg adjusted p-values for correlations of the genes that are found among the top 1000 genes in at least two of the gardens/years. The dashed line represents 5% FDR.

Table 1. GSEA results for circularity. Included are GO terms with a p-value < 0.05 in at least two of the populations. The statistics presented here is the enrichment with the lowest p-value. Included are also the gene sets based on genes located within 2kb of the pSNPs discovered in at least two of the gardens/years ("pairs") and all four gardens/years ("intersection") (Figure 4). The columns are Name: name of the gene set; Description: description of the gene set; Size: the number of genes in the geneset that were expressed (Materials and methods); # eGenes: number of eGenes in the gene set; # pSNP genes: the number of genes in the gene set that were located within 2kb of a pSNP; # pSNPs: the number of pSNPs associated with the genes in the gene set; # eSNPs: the number of pSNPs that were also eSNPs; p-value: p-value of the enrichment; Geneset %: percent of the gene set that is part of the leading edge subset; Total %: the percentage of all genes that are included in the leading edge subset.

Name	Description	Size	# eGenes	# pSNP genes	# pSNPs	# eSNPs	p-value	Geneset %	Total %
GO:0008654	phospholipid biosynthetic process	11	4	0	0	0	0.001	45.45	13.43
GO:0003333	amino acid transmembrane transport	20	7	0	0	0	0.002	55.00	20.13
GO:0004722	protein serine/threonine phosphatase activity	35	8	1	2	0	0.002	54.29	20.86
GO:0006470	protein dephosphorylation	53	11	1	2	0	0.003	50.94	27.71
GO:0016829	lyase activity	25	9	0	0	0	0.003	36.00	13.48
GO:0010333	terpene synthase activity	19	7	0	0	0	0.004	36.84	9.26
GO:0006281	DNA repair	66	21	0	0	0	0.004	31.82	15.09
GO:0016310	phosphorylation	15	2	0	0	0	0.006	40.00	11.45
GO:0007165	signal transduction	197	52	2	11	0	0.007	35.53	25.02
GO:0008138	protein tyrosine/serine/threonine phosphatase activity	16	2	0	0	0	0.009	31.25	11.58
GO:0015171	amino acid transmembrane transporter activity	15	3	0	0	0	0.01	66.67	20.13
GO:0003779	actin binding	28	9	0	0	0	0.012	42.86	23.07
GO:0030246	carbohydrate binding	106	28	0	0	0	0.012	39.62	24.33
GO:0005515	protein binding	2774	724	31	113	0	0.015	26.06	22.51
GO:0016301	kinase activity	32	12	0	0	0	0.021	31.25	10.49
GO:0016627	oxidoreductase activity, acting on the CH-CH group of donors	32	14	0	0	0	0.022	53.13	30.50
GWAS SNP-gene associations	intersection	33	12	33	135	0	0.027	39.39	22.35
GO:0008408	3'-5' exonuclease activity	15	5	0	0	0	0.03	46.67	16.70
GO:0036459	ubiquitinyl hydrolase activity	21	5	0	0	0	0.03	42.86	24.01
GO:0016746	transferase activity, transferring acyl groups	36	12	0	0	0	0.043	50.00	23.56
GO:0050661	NADP binding	53	14	0	0	0	0.043	39.62	21.99
GWAS SNP-gene associations	pairs	195	54	195	484	2	0.098	26.15	20.50

Table 2. GSEA results for indent width. See Table 1 for column name descriptions.

Name	Description	Size	# eGenes	# pSNP genes	# pSNPs	# eSNPs	p-value	Geneset %	Total %
GO:0015171	amino acid transmembrane transporter activity	15	3	0	0	0	0.001	80.00	29.99
GO:0006812	cation transport	49	17	0	8	0	0.001	38.78	22.69
GO:0003779	actin binding	28	9	0	0	0	0.002	53.57	17.40
GO:0016762	xyloglucan:xyloglucosyl transferase activity	23	4	1	1	0	0.003	47.83	10.98
GO:0003333	amino acid transmembrane transport	20	7	0	0	0	0.004	70.00	30.85
GO:0004842	ubiquitin-protein transferase activity	84	12	2	2	0	0.004	34.52	17.90
GO:0006855	drug transmembrane transport	47	12	0	0	0	0.005	36.17	14.58
GO:0008233	peptidase activity	18	4	0	0	0	0.006	44.44	22.89
GO:0016746	transferase activity, transferring acyl groups	36	12	0	0	0	0.007	41.67	19.65
GO:0004222	metalloendopeptidase activity	32	11	0	0	0	0.008	46.88	23.21
GO:0006073	cellular glucan metabolic process	23	4	1	1	0	0.01	47.83	10.98
GO:0048544	recognition of pollen	74	26	0	0	0	0.011	36.49	22.01
GO:0001522	pseudouridine synthesis	15	6	0	0	0	0.021	73.33	28.53
GO:0005874	microtubule	17	5	0	0	0	0.024	52.94	12.92
GO:0005488	binding	311	82	1	1	0	0.025	33.76	23.38
GO:0005200	structural constituent of cytoskeleton	17	5	0	0	0	0.026	47.06	7.46
GO:0048046	apoplast	49	6	1	1	0	0.028	40.82	23.14
GO:0008272	sulfate transport	15	1	1	5	0	0.029	26.67	11.76
GO:0007165	signal transduction	197	52	1	5	0	0.042	27.92	20.51
GWAS SNP-gene associations	pairs	160	40	160	510	31	0.086	16.25	10.23
GWAS SNP-gene associations	intersection	13	1	13	112	4	0.135	23.08	9.55

Table 3. GSEA results for area. See Table 1 for column name descriptions.

Name	Description	Size	# eGenes	# pSNP genes	# pSNPs	# eSNPs	p-value	Geneset %	Total %
GO:0016762	xyloglucan:xyloglucosyl transferase activity	23	4	0	0	0	0.001	60.87	21.65
GO:0006139	nucleobase-containing compound metabolic process	34	16	0	0	0	0.002	50.00	22.98
GO:0030145	manganese ion binding	24	7	0	0	0	0.002	50.00	21.51
GO:0000413	protein peptidyl-prolyl isomerization	33	12	1	1	0	0.002	48.48	25.96
GO:0005488	binding	311	82	0	0	0	0.002	35.37	23.24
GO:0006073	cellular glucan metabolic process	23	4	0	0	0	0.003	47.83	13.11
GO:0006259	DNA metabolic process	14	6	0	0	0	0.004	64.29	28.89
GO:0003755	peptidyl-prolyl cis-trans isomerase activity	33	12	1	1	0	0.004	45.45	20.40
GO:0006629	lipid metabolic process	130	34	0	0	0	0.009	43.08	27.80
GO:0008408	3'-5' exonuclease activity	15	5	0	0	0	0.01	60.00	18.24
GO:0000287	magnesium ion binding	90	31	1	3	0	0.012	45.56	24.73
GO:0006006	glucose metabolic process	16	6	0	0	0	0.012	43.75	20.17
GO:0006281	DNA repair	66	21	0	0	0	0.012	28.79	17.57
GO:0016310	phosphorylation	15	2	0	0	0	0.02	46.67	24.51
GO:0048046	apoplast	49	6	0	0	0	0.021	26.53	9.36
GO:0022891	substrate-specific transmembrane transporter activity	49	13	0	0	0	0.034	34.69	16.96
GO:0008233	peptidase activity	18	4	0	0	0	0.041	50.00	22.08
GWAS SNP-gene associations	pairs	28	5	28	51	0	0.052	28.57	15.61

Supplementary file S1: Microsoft Excel file with six sheets. Available at https://figshare.com/s/c40f29949c545f788e22.

Sheet 1. **H2.** Broad-sense heritability (H^2) values, and upper and lower 95% confidence intervals for all leaf shape and size phenotypes and their PCA summaries.

Sheet 2. **QST.** Q_{ST} values, and upper and lower 95% confidence intervals for all leaf shape and size phenotypes and their PCA summaries.

Sheet 3. **Ekebo 2008 x 2011.** Analyses of variance tables for all leaf shape and size phenotypes, from the model: $Phenotype_{Ekebo} \sim Year + Genotype + Year \times Genotype$. Analyses were conducted only on genotypes with three or more replicate trees per genotype.

Sheet 4. **Saevar 2008 x 2011.** Analyses of variance tables for all leaf shape and size phenotypes, from the model: $Phenotype_{Saevar} \sim Year + Genotype + Year \times Genotype$. Analyses were conducted only on genotypes with three or more replicate trees per genotype.

Sheet 5. **2008 Saevar x Ekebo.** Analyses of variance tables for all leaf shape and size phenotypes, from the model: $Phenotype_{2008} \sim Garden + Genotype + Garden \times Genotype$. Analyses were conducted only on genotypes with three or more replicate trees per genotype.

Sheet 6. **2011 Saevar x Ekebo.** Analyses of variance tables for all leaf shape and size phenotypes, from the model: $Phenotype_{2011} \sim Garden + Genotype + Garden \times Genotype$. Analyses were conducted only on genotypes with three or more replicate trees per genotype.

Supplementary file S2: Microsoft Excel file with three sheets. Available at https://figshare.com/s/d952ccec6a3a26f60795.

Sheet 1. **Leaf shape and size metrics.** Leaf metrics for leaf shape and size phenotypes are listed separately, with their definitions. For further details, please see [26].

Sheet 2. **Sampling information.** A brief summary of the numbers of trees and the numbers of genotypes represented in each common garden and sampling year.

Sheet 3. **PCA loadings.** Principal component analyses loadings for the first components of **(a)** shape, **(b)** size and **(c)** shape and size phenotypes.