Plant Biotechnology Journal

Plant Biotechnology Journal (2010) **8**, pp. 10–27 doi: 10.1111/j.1467-7652.2009.00460.x

Review article

Expression quantitative trait loci analysis in plants

Arnis Druka[1], Elena Potokina[2], Zewei Luo[3], Ning Jiang[3], Xinwei Chen[1], Mike Kearsey[3] and Robbie Waugh[1,*]

[1]*Genetics, Scottish Crop Research Institute, Invergowrie, Dundee, UK*
[2]*Vavilov All-Russian Institute of Plant Industry, St Petersburg, Russia*
[3]*School of Biosciences, The University of Birmingham, Birmingham, UK*

Received 27 August 2009;
revised 24 September 2009;
accepted 29 September 2009.
**Correspondence* (fax 44 1382 568587;
e-mail robbie.waugh@scri.ac.uk)

Keywords: expression quantitative trait loci, transcript-level variation, genetical genomics, transcript-derived markers.

Summary

An expression Quantitative Trait Locus or eQTL is a chromosomal region that accounts for a proportion of the variation in abundance of a mRNA transcript observed between individuals in a genetic mapping population. A single gene can have one or multiple eQTLs. Large scale mRNA profiling technologies advanced genome-wide eQTL mapping in a diverse range of organisms allowing thousands of eQTLs to be detected in a single experiment. When combined with classical or trait QTLs, correlation analyses can directly suggest candidates for genes underlying these traits. Furthermore, eQTL mapping data enables genetic regulatory networks to be modelled and potentially provide a better understanding of the underlying phenotypic variation. The mRNA profiling data sets can also be used to infer the chromosomal positions of thousands of genes, an outcome that is particularly valuable for species with unsequenced genomes where the chromosomal location of the majority of genes remains unknown. In this review we focus on eQTL studies in plants, addressing conceptual and technical aspects that include experimental design, genetic polymorphism prediction and candidate gene identification.

Introduction

Phenotypic differences among individuals are partly the result of sequence polymorphisms that produce altered (or absent) proteins and partly the result of qualitative and quantitative differences in gene expression that generate varying amounts of protein in a cell or tissue. While variation within coding sequences is largely immune to environmental stimuli, gene expression at the transcriptional level is frequently considered the opposite, providing variation in the location, timing and/or abundance of individual mRNA. However, such variation in mRNA abundance is not solely determined by environment—a large number of studies have shown that genotypic variation in regulatory sequences can have a profound effect on comparative levels of gene expression among alleles. In the majority of cases, levels of gene expression are equated with the steady-state abundance of individual mRNA transcripts that have been determined in a specific sample at a given point in time. Abundance information can be captured using a variety of techniques and at a range of scales ranging from quantitative reverse transcription polymerase chain reaction (RT-PCR) (Czechowski *et al.*, 2004), through DNA microarrays (Schena *et al.*, 1995) to massively parallel signature sequencing (MPSS) (Brenner *et al.*, 2000) currently facilitated by next generation sequencing (NGS) (Wall *et al.*, 2009). If transcript levels are measured across a population of plants, the recorded variation in mRNA transcript abundance for each gene may be treated as a heritable trait that can be subjected to statistical genetic analyses. This can, in turn, locate and identify the underlying genetic factors that control the observed variation. The terms genetical genomics (GG) or expression quantitative trait loci (eQTLs) (Jansen and Nap, 2001) have been coined to describe this type of analysis.

The first published large scale eQTL mapping experiments (in yeast and mouse) involved small experimental populations and mostly described the results of genetic mapping (Brem *et al.*, 2002; Schadt *et al.*, 2003). These were soon followed by more focused studies on specific

complex traits and aimed at a better understanding of the molecular networks underlying the trait QTLs and experimental validation of inferred candidate genes (Mehrabian *et al.*, 2005; Yang *et al.*, 2009). Recently the eQTL approach has been extended to genome-wide association studies in humans, mostly addressing complex disease-related traits (e.g. Emilsson *et al.*, 2008, for the recent review see Cookson *et al.*, 2009), and to traits in *Drosophila* such as aggressive behaviour (Edwards *et al.*, 2009).

In this review we attempt to summarize the work that has been published surrounding eQTL analysis, with a focus on plant species. Published studies have been conducted in *Arabidopsis* (DeCook *et al.*, 2006; Keurentjes *et al.*, 2007; West *et al.*, 2007), eucalyptus (Kirst *et al.*, 2004, 2005), maize (Schadt *et al.*, 2003; Shi *et al.*, 2007), wheat (Jordan *et al.*, 2007) and poplar (Street *et al.*, 2006), but because of familiarity we have chosen to focus our discussion around barley, a large genome (5300 Mbp) (Bennett and Smith, 1976) true diploid ($2n = 2x = 7$) crop plant. For more than 100 years barley has been a model plant for genetics research. It has benefited from worldwide collaborations of barley researchers leading to the early development of extensive Expressed Sequence Tag (EST) collections (Zhang *et al.*, 2004), community Bacterial Artificial Chromosome libraries (Yu *et al.*, 2000), high-throughput single nucleotide polymorphism (SNP) mapping platforms and the first publicly available Affymetrix GeneChip (Barley1: Close *et al.*, 2004) for a plant species. The fact that barley also has well established and widely distributed reference populations (Kleinhofs *et al.*, 1993; Wenzl *et al.*, 2006) that have been subject to extensive phenotypic and genotypic characterization (and subsequent genetic analysis) makes it particularly valuable for studies aimed at exploring the potential of eQTL analyses.

The eQTL studies are gaining popularity in plant genetics because they represent a potential mechanism to short-cut the tedious process of positional cloning, especially for genes underlying quantitative characters (Hansen *et al.*, 2008). The data generated from eQTL experiments can be partitioned into two essential components of genomic analysis. Firstly, they can be used to generate the information required to construct a robust and comprehensive sequence-based genetic framework map (West *et al.*, 2006; Luo *et al.*, 2007; Potokina *et al.*, 2008b) and, secondly, they provide data for eQTL analysis itself which is directly coupled to candidate gene identification (Shi *et al.*, 2007; Druka *et al.*, 2008a). Furthermore, the potential to exploit highly complex eQTL datasets using 'systems analyses' is significant and beginning to generate genome-wide appraisals of specific biological phenomena (Jansen *et al.*, 2009). As our investigations are beginning to facilitate the identification of genes underlying biological traits, we have been actively addressing the strengths and the weaknesses of adopting an eQTL approach. Consequently, we attempt here to provide a balanced perspective on where and when eQTL analysis may be an appropriate strategy and guidelines for effective experimental design and execution.

What exactly are eQTLs?

Expression QTLs are those genetic regions identified by applying QTL analysis methods to data on the abundance of transcripts of particular genes in samples taken from different individuals (genotypes) in a segregating population, or populations with other genetic structures. Transcript abundance is used as a measure of the level of that gene's expression in each individual and can be analysed as a trait (an 'eTrait') just like other phenotypes (pTraits) such as plant height or yield. Expression is normally measured for many thousands of genes simultaneously and hence there are thousands of eTraits recorded.

As with pQTL analyses, eQTL analysis requires genetic markers which can be genotyped in all individuals in the population and used to form a framework genetic map of the whole genome. These markers and their map locations may have been developed in the population before the eQTL study or may be developed *de novo* entirely or in part from the expression data itself. A high quality genetic linkage map is a critical component of such experiments, because the map resolution, marker density and polymorphism will condition the quality of pQTL and eQTL analyses, and how we interpret the impact of allelic variation on physiological processes through transcriptional and other molecular networks (Sieberts and Schadt, 2007). The outcome of this analysis is a statistical association between genetic markers located at specific regions of the genome and the 'transcript abundance' of the assayed gene. The significance of the association can be recorded as the Logarithm Of Odds (LOD) score or Likelihood Ratio Statistic (LRS), and plotted relative to each test position covered by the genetic markers across the genome. The values of the test statistic (LOD or LRS) needed to achieve significance depend on the population size, population type and the proportion of non-genetic variation for the trait. The resulting eQTL plot indicates the likely genetic location(s) of DNA sequence variation (i.e. eQTL) that causes the observed variation in transcript abundance across this population.

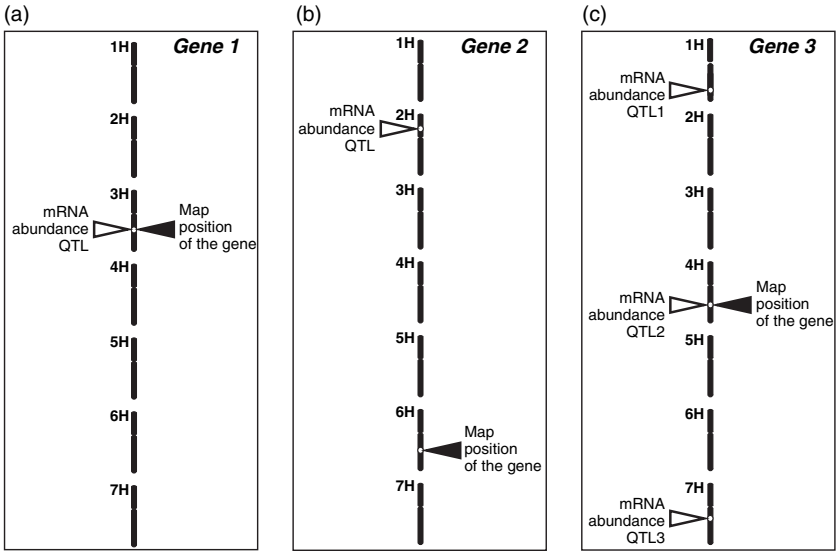

Figure 1 The mode of regulation of gene expression inferred by expression quantitative trait loci (eQTL) analysis. Panel (a): *cis*-regulation is considered the likely cause of observed genotype-dependent mRNA accumulation if the positions of the eQTL and the gene coincide. Panel (b): If they do not, a *trans*-factor encoded by an eQTL locus (chromosome 2H) is the most likely cause of the observed differences. Panel (c): if multiple eQTL are observed, with one coinciding with the location of the gene, combined *cis*- and *trans*-regulation can be inferred.

Expression QTLs are empirically divided into two classes *cis* and *trans* (Figure 1). In the former, the sequence variation controlling transcript levels is assumed to be determined by the sequence variation that lies within or in the close proximity of the gene. In terms of classical molecular genetics such DNA sequence variants are called *cis*-elements; hence a *cis*-eQTL coincides with the location of the underlying gene. In case of *trans*-eQTL, the observed location of the eQTL does not coincide with the location of the gene. This implies that the observed eQTL represents the position of a locus that controls the expression of the target gene. Genes underlying *trans*-eQTLs are assumed to encode *trans*-acting factors—typically proteins that by binding to *cis*-elements of other genes, that control their mRNA expression. Thus a *trans*-eQTL could, for example, represent the location of a transcription factor that controls the expression of the target either alone or, potentially, the correlated expression of several functionally related genes. In reality, target gene expression can be controlled by a combination of both *cis*- and *trans*-acting elements (Figure 1).

It must be emphasized that all QTL studies rely on natural genetical variation in the population under study. Thus, different populations may reveal different QTLs while the absence of QTL for a trait in one population is not evidence that QTL variation for that trait does not exist in other populations.

Technologies for eQTL mapping

Instead of looking at a single gene, a typical eQTL experiment involves large scale mRNA profiling of thousands of genes (Jansen and Nap, 2001). Several platforms are suitable for eQTL analysis. In barley, the Affymetrix Barley1 GeneChip (Close *et al.*, 2004) and custom Agilent microarrays have been used, but other options are summarized in the Table 1. The technologies differ in the way the microarrays are fabricated and this can offer advantages for specific applications. For example, the basic building blocks of Affymetrix' GeneChip microarrays are 25-base long oligodeoxynucleotide probes that are synthesized at specific locations on a coated quartz surface by photolithography (see http://www.affymetrix.com/about_affymetrix/outreach/educator/microarray_curricula.affx#1_1). Each 25-mer is called a feature. Over a million features per microarray are usually available for synthesis, allowing multiple (typically 22) probes per gene (the probe-set). The Barley1 GeneChip has c. 23 000 probe sets corresponding to assembled EST unigenes.

In contrast, the basic building blocks of Agilent microarrays are 60-base long oligodeoxyribonuclotides that are printed on glass slides using Agilent's proprietary SurePrint™ technology. Currently, the slides generally contain either $1 \times 244K$, $2 \times 105K$, $4 \times 44K$ or $8 \times 15K$ probes. The Agilent gene expression platform is fully customizable; ready-to-go probes can be ordered to be synthesized on the slide, or alternatively custom sequences can be used to design probes by 'eArray', an online probe design tool (http://www.chem.agilent.com/en-US/products/instruments/dnamicroarrays/pages/gp50660.aspx). An extensive pool of pre-designed probes is also available from the eArray depository. Experimental design using Agilent microarrays can incorporate either a two-dye labelling protocol or single-dye labelling. Currently, two barley custom Agilent

Table 1 Commercially available parallel, high-throughput gene expression analysis platforms

Company	Detection method	No. probes per gene	Probe length (mer)	No. channels	No. genes detected simultaneously
Affymetrix	Fluorescence; hybridization to array	Multiple (10–20 pairs)	25	1	10 000–100 000
Agilent	Fluorescence; hybridization to array	Single	60	1 or 2	15 000–244 000
Nimblegen	Fluorescence; hybridization to array	Multiple (up to 20)	45–60	1 or 2	10 000–100 000
Illumina	Fluorescence; tags on beads	Single	50	1	10 000+
ABI	Fluorescence; gel electrophoresis	Single	60	1	10 000+
Sequenom	Mass spectrometry	Single	60–90	2	Up to 1000

arrays, 15K and 44K have been designed by ourselves, and a 44K array has been designed by Agilent. Based on various assumptions we estimate that the Agilent 44K microarray contains approximately 38 000 barley genes, the Barley1 GeneChip around 18 000, and the 15K microarray 15 000 barley genes. Because all platforms are gene-based, data integration from different experiments is relatively straight forward.

When choosing an expression platform, genome representation, microarray performance and costs all have to be considered. By design, Affymetrix arrays provide data with more ideal statistical properties than the Agilent platform because individual mRNA signal detection is based on multiple probes and in many cases multiple probe sets per gene. However, relatively low gene content and high price tag can outweigh its design advantage. Agilent's flexible customization could also, of course, be used to design a cheap multi-probe array. Microarray platforms are currently available for most major plant species or can be easily developed for any plant with available EST collections using a specific vendor's array design approach.

Next generation sequencing methodologies, such as that provided by Illumina's Solexa platform, offer an increasingly attractive alternative for tag-based transcript abundance studies (Wall *et al.*, 2009). Despite the relative expense (at the moment) these platforms offer the opportunity to analyse any species—with or without a genome sequence—using an open platform that is capable of recording the abundance of all mRNAs in a given sample. We are unaware of any example of eQTL analysis published to date using this approach but suggest it is likely in the near future that NGS will be utilized for this purpose.

Experimental design

Factors influencing the design of eQTL experiments are essentially those that are important in all QTL studies. They fall into three main categories; type of population, population size and the organization of replication and randomization. All QTL studies require some level of appropriate replication and this implies that individual genotypes can be replicated. Plant geneticists typically use populations of homozygous lines derived from the F_1 generation of bi-parental crosses which are either inbred directly from the F_1s or via backcrossing to one (or both) of the parents. An individual from a natural or artificial outbreeding population can be used as a surrogate for the biparental F_1. Homozygosity may be achieved by inbreeding, typically by single seed descent (SSD) to produce recombinant inbred lines (RIL) or by production of doubled haploid lines (DHL). Such populations can only provide information about the additive effects of QTL because every genotype is homozygous. However, many inbred lines from a RIL or DH population can be crossed *inter se* to create a pseudo F_2 population in which every genotype is replicated as opposed to being unique as in a true F_2. Alternatively they can be crossed to testers. Such crosses produce reproducible and hence replicable F_2's from which the dominance effects of QTL can also be studied (Dupuis and Siegmund, 1999; Liu and Zeng, 2000).

The accuracy of locating eQTL in such populations depends directly on the amount of recombination that has occurred in the production of the inbred lines. At present there is no environmental treatment that can increase the rate of recombination easily. At meiosis each bivalent chromosomes typically has around two chiasmata and hence there is a high probability that a chromosome derived from any individual meiosis has one or no crossovers with most genes failing to recombine at all. Using microspore culture to produce DHL from an F_1 involves just a single round of meioses in the F_1 and hence recombination is minimized. Inbreeding to produce RIL from an F_2 by SSD effectively doubles the amount of recombination because crossing over can occur in several generations, although the effects of crossing over in generating recombinants are considerably reduced as homozygosity increases. If it is possible to randomly mate the F_2 for one or two generations before inbreeding, then the amount of

recombination is considerably increased but such a procedure may be technically difficult and time consuming particularly with some inbreeding species.

Population sizes should be as large as possible because size affects the number of recombinants that can be sampled for QTL location as well as increasing the statistical power of the analysis. Increasing population size has corresponding increases in costs which are particularly significant for gene expression analysis. It has been suggested that no QTL study should involve less than 200 lines and several recent publications have used many more (Schön *et al.*, 2004; West *et al.*, 2007). However, there is no convincing statistical justification why much smaller populations cannot be used and many studies have successfully used 100 lines or less. Gene expression traits typically have high heritabilities which ensure a high level of power for detecting their underlying genetic variants. Providing significance levels are set such that the genome wide false-positive detection rates are low (typically ≤5%) for that population, then one simply fails to detect more eQTLs in small populations. However, 95% of those actually detected are real; but these should always be individually confirmed by other means such as by use of different samples of RIL/DHL or near isogenic lines from the same population.

In all experiments, and those involving eQTL are no exception, it is essential to have an appropriate measure of replicate error variation to prove that genetic variation exists. If considering a DH or RIL populations for example, one would expect that, for a given segregating marker, half the population of lines ($n/2$) would consist of the one homozygote and the other half the second homozygote. Given the availability of expression data for a gene on all lines one could test for any marker whether the two homozygotes differ in expression for a particular gene by a *t*-test with $n - 2$ df. However, the error mean squared (the denominator) in such a *t*-test will include variation from environmental and genetical sources from all other marker loci, so it provides a very conservative test. In case of a single *cis*- or *trans*-acting eQTL this may not be an issue because one marker will extract all or most of the genetic variation, but for multiple eQTLs the residual genetical variation of other QTL will result in a weaker test. Ideally, therefore it is important to have replicates of individual genotypes, i.e. the DH lines or RIL because they will allow the estimation of the true within line variation. Within a given budget, replication reduces the total numbers of lines that can be studied, with a concomitant reduction in power and precision. Thus, although it is important to have an appropriate measure of environmental error, it is not necessary to replicate all lines; providing that the error is determined for a randomly chosen sample of lines it is possible to perform a test of significance. Obviously the power of this test increases with the number of lines replicated but beyond *c*. 30 replicated lines (error df = 30) power increases very little.

It is also essential that care is taken to ensure that the replicates of all genotypes are unbiased i.e. they incorporate all of the non-genetic factors that could cause lines to differ. Such differences may include effects such as: technical variation in sampling, preparing, assaying and recording material on a slide; true biological environmental variation such as is found among genetically identical full-sibs from a parent within a line; environmental or maternal (or epigenetic) effects attributable to different parents within a line (whether it is sensible to consider generations beyond the parents is debatable but even under rigorously controlled conditions, maternal effects can be very considerable). Unfortunately, these factors are often ignored in the experimental design resulting in serious underestimation of the non-genetic variation and inflation of the genetic components. It is generally agreed that the focus of the design should be on biological rather than technical variation (Kerr and Churchill, 2001; Churchill, 2002; Kendziorski *et al.*, 2003).

To avoid unnecessary replication of slides it is possible to pool samples from different parents and individuals on a single slide as long as the replicate slide involves a pool from different parents and individuals (Churchill, 2002; Kendziorski *et al.*, 2003; Simon and Dobbin, 2003). Alternatively, a distant-pair design has been proposed for two-colour microarrays by Fu and Jansen (2006). The design uses genetic marker information to identify pairs of individuals with maximum dissimilarity across the mapping population and improves the efficiency of eQTL studies. This approach has recently been adopted in studying the interaction between barley and leaf rust (X. Chen, pers. Commun.).

Tissue sampling

As individual mRNA abundances are dynamic traits that are subject to developmental, environmental and physical cues, to minimize the contribution of non-genetic factors, tissue sampling for RNA extraction should, ideally, attempt to avoid these sources of variation. For example, the presence of highly pleiotropic major developmental genes (e.g. a major dwarfing gene) segregating in a population may

be expected to partition a significant portion of observed eQTL variation according to the dwarfing allele present in individuals in the population. While it is possible to sample developmentally 'equivalent' material across a period of time (DeCook *et al.*, 2006; Jordan *et al.*, 2007) and obtain what appears to be meaningful eQTL data, our recommendation would be to attempt to minimize this variation as much as is possible. For example, Druka *et al.* (2008b) harvested seed from the reference barley Steptoe × Morex (*St/Mx)* population grown in the same environment, sieved seed to obtain equivalent size fractions, pre-conditioned all seed in parallel, germinated three replicates of *c*. 30 seed and chose only three germinating seed from each replication that they considered by eye to be developmentally and physiologically equivalent for pooling and RNA isolation. They subsequently combined equal quantities of RNA from each of three biological replicates prior to analysis by Affymetrix chip hybridization. As a result they minimized the potential sources of non-genetic variation and obtained a robust and reproducible dataset that was considered fit for eQTL analysis.

A framework genetic map

A high quality genetic map is required for interrogation of the transcript abundance data. It can have three origins as follows: (i) a legacy map constructed prior to the experiment, (ii) a *de novo* map inferred from the expression data itself (see below), (iii) an independently generated map using genetic markers derived from a subset of the genes used for mRNA profiling (e.g. SNPs). Because of the way they were derived (e.g. using restriction fragment length polymorphisms, amplified fragment length polymorphism or single sequence repeat), legacy maps (while temptingly convenient) frequently suffer from the anonymity of the majority of the markers that were used in their construction, from high levels of inherent type 1 error and from missing data. These compromise the accurate association of eQTL and their underlying genes. In contrast, bi-allelic markers derived from the expression dataset itself [i.e. single feature polymorphism (SFP), gene expression markers (GEM) or transcript-derived marker (TDM)—see below] can represent a highly efficient gene-based marker system (Rostoks *et al.*, 2005; West *et al.*, 2006; Luo *et al.*, 2007; Potokina *et al.*, 2008b). Maps based on these marker types optimize the use of the transcript abundance data and minimize type 1 error, as the same dataset is used for both map construction and eQTL analysis. Finally, current advances in high-throughput SNP-based DNA genotyping technologies enable quick and efficient generation of high quality, robust, transferable and saturated genetic linkage maps (T. Close, unpublished results). Our experience indicates a preference for either of the latter two approaches.

Methods for eQTL analysis

The major challenges in modelling and analysis of experimental data for mapping regulators of genome-wide gene expression lie in the high-dimensionality and complex correlation structure of the source microarray data. Otherwise eQTL analysis shares virtually the same statistical principles and approaches as conventional pQTL (Lan *et al.*, 2003). As noted, microarray experiments provide two sources of information essential for eQTL analysis: genome-wide genetic markers and the transcript abundance phenotype of every gene that corresponds to a probe or probe-set on the array (Figure 2). Data modelling and analysis therefore play fundamental roles in extracting statistically manageable information from the raw hybridization signals and in integrating the information into eQTL.

Prediction of genetic polymorphisms from microarray data

The ability to identify sequence polymorphisms from gene expression microarray data has useful implications in at least two aspects. Firstly, it improves both accuracy and precision in calculating gene expression indices by excluding probes containing genetic polymorphisms, while in turn, improving the statistical power of eQTL analysis (Alberts *et al.*, 2007). Secondly, it enables the concomitant generation of an abundant collection of reliable genetic markers that can be used as the framework for subsequent eQTL genetic analysis (Luo *et al.*, 2007). The markers derived from microarray experiments fall into three classes as follows: (i) single feature polymorphism (Borevitz *et al.*, 2003), (ii) GEM (West *et al.*, 2006), (iii) TDM (Potokina *et al.*, 2008b).

Single feature polymorphism SFP (Borevitz *et al.*, 2003): to distinguish hybridization signals associated with any molecular alteration from background, the signals generally need to be collected from the sequences where mutation occurs. The probe-set design of Affymetrix microarrays perfectly meets this need. Statistical methods have been developed to detect polymorphisms between target sequences and probes by testing for non-uniformity of hybridization intensity among every feature in a probe-set for a given gene. The principle for RNA-based tem-

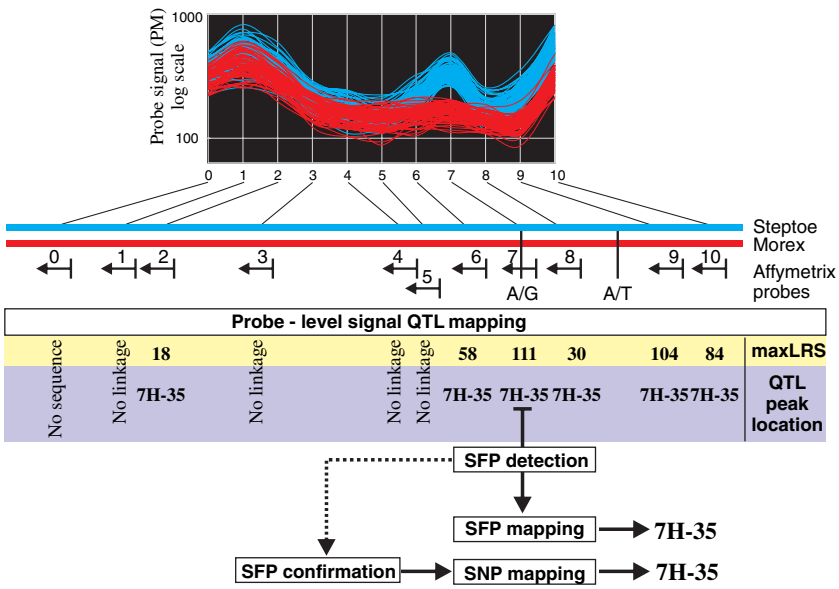

Figure 2 Framework for probe-level analysis: single feature polymorphism (SFP)'s GEM and TDM. mRNA abundance signals from a single GeneChip probeset can be used directly for eQTL mapping (Druka *et al.*, 2008a,b) or SFP inference (Luo *et al.*, 2007). The line plot (top) shows the individual probe signal values from one probset extracted from 150 Affymetrix CEL files from the *St*/*Mx* doubled haploid recombinant line population. Line colours (red/blue) were assigned based on the clear separation of the signals recorded from probe #7 (numbered short black arrows). The position of two single nucleotide polymorphisms (SNP) between *St* and *Mx* are shown relative to the location of the probes on the consensus sequence. In this example, six of the 11 probes identified eQTL (yellow panel, maxLRS) at the same locus (7H-35, i.e. chromosome 7H position 35 cM, violet panel). Polymorphism between parental alleles showed that the strongest eQTL (probe #7) contained a SNP in the target sequence and therefore represented an SFP. Both the SFP and SNP mapped to the same genetic location. The other five probes detected variation in transcript abundance but were not sufficiently discrete to be classified as GEM. The strong probe #7 eQTL is therefore because of the combined effect of the SNP and mRNA abundance differences and would also have been classified as a TDM.

plates is an extension of that described by Winzeler *et al.* (1998) who pioneered the development of a high-throughput genotyping platform by hybridization of labelled total genomic DNA to oligonucleotide arrays. In yeast, this approach has proved useful in linkage analysis, the dissection of QTL, and in assessing species population structure (Hazen and Kay, 2003). Recently, it has been applied to organisms with more complex genomes, such as *Arabidopsis thaliana* (Borevitz *et al.*, 2003), to assess the molecular basis of natural phenotypic variation. This type of sequence variation detected by a single probe in an oligonucleotide array was termed as SFP (Borevitz *et al.*, 2003) (Figure 2).

By attempting to integrate genetic polymorphism screening and gene expression analysis, Ronald *et al.* (2005) proposed the concept of simultaneous genotyping and gene expression analysis with microarray data. They hybridized cRNA from two parental yeast strains and their segregants onto yeast Affymetrix GeneChip arrays, and developed a method for identifying SFP and consequently genotyping the yeast strains (essentially by combining a *k*-means clustering and a mixture model analysis). The approach was based on the proposition that the presence of polymorphism in a perfect-match (PM) probe sequence on Affymetrix arrays between parental strains, one of which was assumed to have the same sequence as the probe, would lead to a detectable difference between the observed PM value of the probe and its predicted value. The predicted value comes from the 'positional-dependent-nearest-neighbour' model (Zhang *et al.*, 2003). This idea has been modified and implemented to predict SFP using gene expression data profiled by Affymetrix GeneChip arrays in more complicated species such as *Arabidopsis* (West *et al.*, 2006) and barley (Cui *et al.*, 2005; Rostoks *et al.*, 2005). However, it is much more challenging statistically to predict SFP from gene expression data than from genomic DNA microarray data. This is because the effect of genetic polymorphism within a transcript molecule on the hybridization signal is compounded by the abundance of the transcripts from the gene represented by the probe. Luo *et al.* (2007) tested the performance of various methods by implementing them to predict SFP from expression data of 22 801 open reading frame, which were profiled using Affymetrix microarrays on two barley cultivars and their double haploid offspring. They found that a large proportion of the predicted SFP

from these methods were more a reflection of variation in gene expression than genuine sequence polymorphisms. The use of these GEM may however result in a serious bias in eQTL analysis because autocorrelation between markers (i.e. GEM) and trait phenotypes (i.e. gene expression level) may lead to the false prediction of *cis*-regulators. Thus, a statistical method that enables robust prediction of probes bearing genuine sequence polymorphisms from gene expression data is clearly desirable.

Wang *et al.* (2009) proposed a Bayesian statistical approach for detecting SFP's in transcript sequences and for predicting SFP genotypes when tested in a segregating population derived from genetically divergent parental lines. This was achieved by modelling a PM value from Affymetrix cRNA hybridization experiments as a product of the binding affinity between the transcript and probe sequences and the abundance of the transcript. They analysed two independent microarray datasets (RNA hybridizations from barley and yeast) and demonstrated that their method provided significantly improved robustness and accuracy for predicting SFP reflecting genuine sequence polymorphism, when compared with five other statistical methods. Their method was appropriate for predicting SFP from expression microarray data and from genomic DNA microarray data. By comparing predicted SFP with those where sequence information was available, they showed that all the methods applied stringent selection criteria to SFP and thus only a small fraction of probes on arrays recorded SFP. The approach effectively maintained both false-positive and false-negative rates at a low level.

Gene expression markers (West *et al.*, 2006): these are based on the clear-cut difference in the level of transcript abundance between two genotypes. Using *Arabidopsis* RIL, West *et al.* (2006) showed that a subset of 324 (from 1431) genes exhibiting a >2-fold difference in a level of gene expression observed between the two parental genotypes had non-overlapping expression levels among the progenies of the mapping population: i.e. for each of the offspring the inherited parental allele can be unequivocally predicted by the mRNA transcript level. Accordingly, they proposed that a GEM can be used as a genetic marker that identifies the location of the DNA sequences that regulate the expression of its corresponding gene. Luo *et al.* (2007) used a weighted average for all features per probeset on Affymetrix arrays to determine comparative eQTL 'levels'. The same principle operates for all other array types. Clearly, an important question is whether the mapped GEM (i.e. regulating locus) coincides with the position of the gene itself, because only GEM that reflect polymorphism in local regulatory sequences (perhaps, a promoter) will place the gene at the 'correct' place on the genetic map (i.e. regulation in *cis*).

Potokina *et al.* (2008b, 2009) observed that 95% of GEM in barley were the result of *cis*-regulatory polymorphisms. Thus, in this case, GEM reflect polymorphic sites in or near a gene that trigger dramatic changes in the levels of gene expression and are detected indirectly as extreme differences in transcript abundance. GEM separate both the parents and progeny of a bi-parental cross into two distinct groups each containing one of the parental alleles. This has an important consequence: to map a gene by transcript profiling of a segregating population requires no prior information on whether the expression marker represents a nucleotide polymorphism in the probe itself (i.e. SFP) or simply indicates a regulatory polymorphism. In case of Agilent arrays that typically comprise one probe per gene, regulatory polymorphisms will be the dominant source of the observed variation because the long oligonucleotide probes used are relatively insensitive to single nucleotide changes (Hughes *et al.*, 2001).

Transcript-derived markers (Potokina *et al.*, 2008b): these are a catch-all class where the hybridization signal reported by a probe is used alone as a gene-specific marker i.e. not distinguishing between SFP and GEM (Figure 2). These authors identified 1596 barley TDM that were successfully employed to construct a robust genetic map of the St/Mx population that was subsequently used for eQTL analysis. The barley results support previous reports that eQTL with the highest LOD scores are generally attributable to *cis*-eQTL (Figure 3.) (Gibson and Weir, 2005; Hubner *et al.*, 2005; Yamashita *et al.*, 2005; West *et al.*, 2007; Druka *et al.*, 2008a).

The importance of gene-based linkage maps

In species with large and/or unsequenced genomes, gene-base linkage maps have considerable added value over traditional marker-based maps. This is because conservation of synteny between these species and fully sequenced model genomes can help validate predicted gene locations based on eQTL analysis. For example, using the latest barley EST assembly (Harvest 35; http://www.harvest-web.org/hweb/bin/wc.dll?hwebProcess~hmain~&versid=5assembly) the total number of reciprocal barley EST-rice genome-barley EST hits is just under 18 000. This translates into *c.* 43% of rice genes having good barley homologues, or 39% of the predicted total number of barley genes having

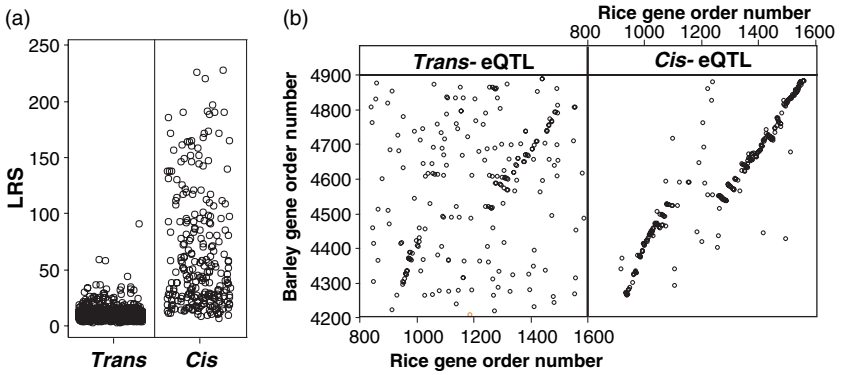

Figure 3 Inference and validation of *cis*-expression quantitative trait loci (*cis*-eQTLs) and their use as genetic markers. Panel (a): about 1000 genes that have both single nucleotide polymorphism (SNP)-based mapping data and reveal significant eQTLs can be separated in two groups: a '*trans*' group that have the position of a gene on a different chromosome from its eQTL, and a '*cis*' group where a gene and eQTL co-locate. Individual value plots show clear differences of the distribution of Likelihood Ratio Statistics (LRS) in these two groups—most of the high LRS values are associated with the '*cis*' group and can be confirmed as *cis*-eQTLs. Panel B: The observed chromosomal order of eQTLs in barley from each group (*cis* vs. *trans*) was plotted against the order of their rice homologs on the rice genome pseudomolecules. One chromosome is shown, with two major blocks of conserved synteny observed for the *cis*-regulated genes (high LRS) and a more random distribution for the *trans*-regulated genes (modified from Druka *et al.*, 2008a,b). Note, that *cis*-eQTLs with low LRS values are not included in this analysis.

good rice homologues. Aligning the location and order of genes on the rice genome sequence to the genetic order of the corresponding *cis*-eQTL on the barley map reveals the same synteny blocks as observed in barley SNP-based maps (Figure 3) (Druka *et al.*, 2008a; T. Close, unpublished results), supporting the validity of predicting gene positions based on *cis*-eQTL.

Assessing gene expression from microarray data

An important question surrounds how consistently gene expression can be evaluated by different technical platforms. It has been documented that commercial microarrays are more technically consistent than non-commercial microarrays (Coughlan *et al.*, 2004; Chen *et al.*, 2007) and that one-dye platforms are typically more consistent than two-dye platforms (Kuo *et al.*, 2006). We compared the expression of 15 208 barley 'unigenes' profiled using Agilent one-dye and two-dye microarrays with that of the same group of genes from Affymetrix microarrays (N. Jiang, unpublished results). We observed a significant discrepancy between the expression levels evaluated from the two types of microarray. Correlation coefficients in estimated expression indices were as low as 65% between the platforms with the one-dye system generating a higher level of concordance than the two-dye system. Despite this observation, using a different algorithm, X. Chen (unpublished results) very effectively used normalized ratios derived from 72 distant-pair hybridizations (Fu and Jansen 2006) to map eQTL related to the interaction between barley and *Puccinia hordei* Otth., suggesting that the analytical approach used for interpretation of the data may be key to the degree of correspondence observed.

Is this surprising? Not especially. Jiang *et al.* (2008) explored seven main stream methods developed for extracting gene expression levels from Affymetrix microarray datasets in both yeast and barley and found that these methods can be divided into two clusters. The methods within each cluster reveal correlation coefficients of ≥95%, but the correlation is reduced to *c*. 70% between the two clusters. In addition, the number of genes called to be 'significantly differentially expressed' between the same set of genotypes varies substantially among the different data extraction methods when subject to the same rate of false-positives (Storey and Tibshirani, 2003).

Genome-wide eQTL analysis

Regulatory polymorphisms revealed as eQTLs in GG studies are major determinants of quantitative phenotypic traits (Stamatoyannopoulos, 2004). But how complex should we expect these traits to be?

Analysis of the barley St/Mx eQTL data from 'germinating embryo' and 'first true leaf' provides several useful numerical insights. Of the 22 840 probe sets on the Affymetrix GeneChip, 70% (15 967) showed significant expression in embryo tissue based on the 'Present' call by the Affymetrix' MAS 5.0 algorithm. Of these genes, 81%

had one or more eQTL that could be detected by composite interval mapping at a genome wide False Discovery Rate of 5%, resulting in a total of 23 738 eQTL. Among the 36% (5764) of genes with just one significant eQTL, 2291 (40%) had expression differences that fell into two clear, non-overlapping expression classes among the parents and DHL. SNP data indicated that in the majority (95%) of cases, these two classes matched the SNP genotype of the target gene used as probe. These were assumed to represent *cis*-acting expression regulators and this was supported by their high LOD scores, a recognized indicator of *cis*-regulation (Druka *et al.*, 2008a). The remaining 5% were candidates for *trans*-regulators. However, in two cases the putative eQTL mapped to locations occupied by known duplicates of the target gene so it is possible that others may also represent the location of duplicated genes.

When comparing the location of the eQTL to that of the gene itself in barley, slightly <30% of the observed eQTL were in *cis* and the rest in *trans*. A global transcript profiling study in Arabidopsis using the Bay × Sha RIL population correspondingly reported that approximately one-third of the eTraits were *cis*-eQTL (West *et al.*, 2007) whereas in another Arabidopsis study employing the Ler × Cvi RIL population, 50% *trans*-eQTL and 50% *cis*-eQTL were detected (Keurentjes *et al.*, 2007). However, the number of eQTL detected in each study also varied dramatically (over 36 000 in one study and 4000 in the other). In a more focused study in eucalyptus, 1067 of 2608 genes revealed 1655 eQTL in a backcross progeny of 91 individuals (Kirst *et al.*, 2004) of which 821 were single eQTL (49%). A quarter of the detected eQTL were single eQTL in a study focused on cell wall digestibility in maize (Shi *et al.*, 2007) (Table 2). Overall, the proportion and number of *cis*- and *trans*-eQTL identified in an experiment depends on many factors, such as the number of lines used, the degrees of replication and the inherent genetic architecture of the population under study (Hansen *et al.*, 2008). These factors influence the statistical power (higher power allows the robust detection of more *trans*-eQTL) and the number and ratio of *cis*- vs. *trans*-eQTL detected in a population. Indeed, *trans*-eQTL may themselves be the elaboration of divergent *cis*-regulation at a polymorphic regulatory locus (Kliebenstein, 2009). Differential regulation may also be tissue and/or pathway-specific.

Mapping pQTL often requires the selection of parental lines that contrast for the phenotype of interest. While this is also true for eQTL mapping, if parental transcript variation is used exclusively to select differentially expressed genes for use in subsequent studies, many informative genes would end up being be overlooked. This is because transcript-level variation between two homozygous inbred parents does not accurately predict transcript variation in their progeny and is usually significantly underestimated (West *et al.*, 2007). Keurentjes *et al.* (2007) calculated the heritability values from parental data and a RIL population to be 28.6% and 74.7% respectively. Similar findings were observed in barley by X. Chen (unpublished results) who identified 1037 significant differentially expressed genes between St and Mx, but were able to genetically map 9557 genes using the St/Mx population. These observations illustrate quite clearly how genome-wide eQTL studies can provide significantly more information on the biological activity of genes than expression studies on the parental lines alone.

Table 2 Summary of eQTL studies performed on different plant species

	Population				No. genes/eQTL				
Plant species	Parental lines	No. progeny	Type	Tissue used for array	Genes analysed	Genes mapped	Single eQTL	*cis*-eQTL	*cis*-eQTL (%)
Barley	St × Mx	139	DH	Germinating embryos	15967	12933	5764		29–39
Maize	Stiff stalk × lancaster	76	F3	Ear leaf tissue	18805	6481	NA	NA	NA
	Flint × Flint (AS18 × AS07)	40	RIL	5-week old stems	439	89	23	NA	NA
Eucalptus	(tree G50 × E. globulus)x tree 678.2.1	91	Pseudo-backcross	20-month old xylem	2608	1067	821	NA	NA
Arabidopsis	Bay-0 × Sha	211	RIL	6-week old plants	22746	15664		5127	32%
	Ler × Cvi	160	RIL	Aerial parts of seedlings	24065	4066		1875	46%

Source: barley: Potokina *et al.* (2008a,b); maize: Schadt *et al.* (2003) and Shi *et al.* (2007); eucalyptus: Kirst *et al.* (2006), Arabidopsis: West *et al.* (2007) and Keurentjes *et al.*, (2007).

eQTL, expression quantitative trait loci; RIL, recombinant inbred lines; DHL, doubled haploid lines. NA, not available.

The eQTL analyses summarized above (in barley) have used inbred lines. The genetic structure of these lines greatly eases the challenges in theoretical modelling and statistical analysis of the experimental data and, at the same time, limits predictability of the analysis. For example, the significance of dominance effects were neither considered nor explored. By using more genetically complex populations such as F_2's, pedigrees or even sporadic samples from natural populations, these questions may be tackled. However, microarray data collected from such population types can also raise new issues. For example, statistical power for detecting linkage between markers and expression data may be lowered because of a reduced contrast in the phenotypic effect (expression level) between marker genotype classes. Moreover, more complicated statistical approaches need to be developed to cope with new patterns of variation (e.g. from multiple alleles) embedded in the corresponding microarray datasets.

Hotspots

Almost all studies conducted to date reveal that eQTLs are not evenly distributed across genetic maps. When eQTL cluster in a specific region more than expected by chance, the region is frequently declared 'an eQTL hotspot'. eQTL hotspots may be a reflection of regions that are either gene-rich or recombine infrequently (such as genetic centromeres). This type of hotspot is generally of little functional interest. A biologically meaningful eQTL hotspot would represent, for example, the location of a master transcriptional regulator that controls the expression of a suite of genes that act in the same biological process or pathway. To differentiate between these two possibilities, gene ontology and enrichment analysis can facilitate to identify any underlying biological links. When significant enrichment of a 'functional category' is observed, the potential biological pathways relevant to the functional category can be inferred and relevant experiments designed for further investigation.

When eQTL studies are performed in different populations using similar tissues or treatments, the expectation is that consistent eQTL hotspots would represent the same biological pathways. eQTL studies performed in the same population but using different tissues or treatments should however yield complementary results, that reflect the dynamic nature of the transcriptome. Using mRNA from germinating embryos, Potokina *et al.* (2008b) observed several regions on chromosomes 2H, 5H and 7H which had many more eQTL than expected by chance alone based on a uniform distribution of genes per cM. Interestingly, in the same population using *Puccinia hordei* infected seedling leaves 18 h post-infection, eQTL hotspots on two different linkage groups, 1H and 3H, were observed (X. Chen, pers. commun.). While the eQTL hotspots in the barley embryo experiment appeared to be biased towards genomic regions that exhibit little recombination and hence have more genes per cM, in the pathogen challenged tissues this did not appear to be the case (X. Chen, pers. commun.). In the former, some eQTL hotspots did however correlate with the known location of 'malting quality' QTL (a trait expressed and measured in this tissue), while in pathogen challenged tissues at least two of the three hotspots were enriched for mRNA related to general 'pathogen responsive genes'. eQTL hotspots have also been recorded in *Arabidopsis* and while the biological pathways represented at all hotspots have not been identified (Keurentjes *et al.*, 2007; West *et al.*, 2007), one co-located with the well known ERECTA locus which is responsible for pleiotropic effects on many traits including morphological differences (Koornneef *et al.*, 2004). While such observations provide a potential opportunity to unravel the genetic control of important phenotypic traits, in general, these types of study are currently at a very early stage.

Limited pleiotropy

The number and distribution of detected eQTLs, as illustrated in the previous comparison, may vary between specific tissues or stages of development. Potokina *et al.* (2008a) compared eQTL data derived from two different tissues using the same segregating population: germinating embryos and seedling leaves. After conducting a highly selective comparison between the two datasets the experiments yielded 1498 and 1134 robust eQTLs, in embryos and leaves respectively. Five hundred and fifty-one of the eQTLs were common to both tissues (Potokina *et al.*, 2008a). They suggested that the cause of the observed tissue-specific *cis*-eQTLs lay in the phenomenon known as limited pleiotropy of *cis*-regulatory mutations. Limited pleiotropy describes a situation where the effect of *cis*-regulatory variation is spatially or temporally limited. For example, in the barley dataset, 34 genes were detected that revealed a reciprocal change in the parent that contributes the allele with the most abundant transcript in the two tissues sampled. Limited pleiotropy deserves more attention because it may be relevant to

some age-related disorders and/or tissue-specific syndromes. In plants, limited pleiotropy may be important in adaptive traits such as floral organ pigmentation or quantitative age-related plant resistance to various foliar pathogens. Limited pleiotropy may reflect the occurrence of transcription factors whose influence on gene expression is either tissue or stage specific and modulated by polymorphism in the transcription factor binding sites of their target genes.

Candidate gene identification using eQTL data sets

An eQTL study can provide useful data for the identification of candidate genes for pQTLs, particularly, but not necessarily, when eQTL data is generated from the same population under similar conditions as the pQTL data. Of most interest are *cis*-eQTLs underlying pQTLs because they fulfil two important criteria that are required for them to be considered candidates: they co-locate with the pQTL and they are differentially expressed. They are therefore candidates with respect to both position and transcriptional regulation.

A straightforward approach to identifying candidate genes for a trait of interest is to correlate phenotypic trait measurements with mRNA abundance values of all of the genes present on the microarray assay platform. The minimal experimental constraint is that the same genetically fixed population has been used to obtain both the trait and mRNA abundance values. Correlation analysis returns a list of correlates (probes or probe sets) and their respective correlation coefficients. Correlates with the highest absolute correlation coefficient can be considered potential candidate genes for the trait. Logically, most highly correlated eQTL should fall into the region containing the pQTL. However, only one, if any, would be causal, with the correlation observed from other genes almost certainly the result of their physical linkage and regulation in *cis*.

The correlation approach was used to simulate identification of *Rpg1*, a gene that confers resistance in barley to the wheat stem rust pathogen *Puccinia graminis* f. sp. *Tritici* (Druka *et al.*, 2008b). In that study, the abundance of the recently cloned *Rpg1* mRNA was represented by a specific probe-set on the array and was one of the top correlates with stem rust resistance. Both *Rpg1* mRNA abundance and stem rust resistance had single, strong and coinciding QTL. While this was a good and testable example, the correlation approach also appeared to be informative for traits that had multiple QTL. In the same study Druka *et al.* (2008b) partitioned the original, quantitative resistance data into four principal components (PC1–PC4) then used the individual PC data for correlation analysis. They identified *Hsp25* as a candidate for the PC3 stem rust response trait. Both PC3 and *Hsp25* had three significant correlated QTL mapping to chromosomes 3H, 5H and 7H. Recently, in an independent yeast 2-hybrid screen, another small heat-shock protein, HSP17 was identified as an interacting partner with RPG1 (Kleinhofs *et al.*, 2009), supporting the original mRNA-trait correlation-based hypothesis that suggested the involvement of small heat-shock proteins in the barley–*Puccinia graminis* f. sp. *Tritici* interaction.

To infer candidate genes based on correlation, directionality has to be taken into account with transcript abundance data considered in the context of the underlying trait biology. The reason for selecting positively correlated mRNA as candidates for *Rpg1* was based on the assumption that increased resistance should positively correlate with the amount of mRNA from the resistance gene. In this particular case, prior information was available—*Rpg1* is a dominant race-specific resistance gene. In reality, directionality is not always intuitive, especially when dealing with quantitative traits. Despite this, attempts should be made to consider correlations in the context of the target biological process. Directionality plays a central role in the interpretation of observed associations and subsequently for the systems-based assembly of gene regulatory networks. In the *Rpg1* example, differential transcript abundance *per se* is unlikely to be the root cause of the observed phenotypic variation because a stop codon in the susceptibility allele results in the production of a non-functional protein. The stop codon may, however, also result in a reduction in steady-state mRNA abundance through the non-sense mediated mRNA decay mechanism (Brogna and Wen, 2009). The message is that care should be taken when interpreting mRNA abundance as the cause of phenotypic variation.

Correlation analysis provides an overview of potential genes associated with a trait. Further analysis involves the putative function of the correlated genes and whether there are multiple coinciding eQTLs, or hotspots, that may indicate that the causal gene is a *trans*-acting 'master regulator' which may not be represented on the array. If such eQTL hotspots predominantly consists of *trans*-eQTLs that have annotations from previous studies suggesting some form of functional relatedness (e.g. mainly genes involved in pathogen response), then a master regulatory locus may be inferred. Such loci are exemplified by the *sub1* locus in rice which controls the activity of an ethylene

response factor that has significant *trans* effects (Fukao *et al.*, 2006; Xu *et al.*, 2006) and the *ERECTA* locus, mentioned previously, that exerts secondary effects on many developmental processes (Keurentjes *et al.*, 2007). Similarly, a locus on barley chromosome 2H containing a putative regulatory 'master locus' affecting the expression of other genes associated with programmed cell death has been proposed (Druka *et al.*, 2008b). If making such inferences, particularly in small populations, it is important to exclude the possibility that chance co-segregation is responsible for the correlation.

When characterizing a 'master regulatory' locus to identify the underlying gene, it is important to realize that this gene may not be represented on the array. However, if the expression level of a reliably identified *trans*-eQTL is essentially bi-allelic (i.e. it is a single strong eQTL) then it is a powerful surrogate that can be used to characterize and refine the location of the target locus. It may be exploited to effectively reduce the number of positional candidate genes that can feasibly be used for functional validation. As in traditional positional cloning projects, increasing the size of the experimental population and integrating other sources of information is essential.

Allelic imbalance

Individual transcript abundance measurements cannot distinguish the relative contribution of each allele in a heterozygous individual (e.g. when assessing heterosis in maize hybrids, Swanson-Wagner *et al.*, 2006). Assignment is generally based on genetic segregation of transcript abundance by quantitative genetics analysis and requires extensive genotypic and phenotypic (transcript abundance) information across the population. However, an independent test of *cis*- or *trans*-regulation that directly assays the level of allele expression in F1 hybrids is the allele imbalance assay (Cowles *et al.*, 2002; Yan *et al.*, 2002; Wittkopp *et al.*, 2004). The rationale is that in an F_1 hybrid, alleles at a given locus will be in an otherwise identical genetic, transcriptional and environmental background which should in principle remove almost all non-genetic effects (Figure 4). Allele imbalance assays exploit SNP between the coding sequence of the target alleles in heterozygous individuals combined with quantitative RT-PCR, single base extension and capiliary electrophoresis (or pyrosequencing) to obtain an accurate measure of the levels of expression of each allele (by comparison to a titration curve constructed from mixed parental DNA). This approach has been used to identify novel alleles and quantify transcriptional variation for alleles of genes involved in biotic and abiotic stress tolerance in barley (von Korff *et al.*, 2009). There, nineteen of the 30 genes assayed revealed allelic expression level differences of up to 19-fold because of *cis*-regulatory variation. Imbalance assays have also been used in polyploids to quantify relative expression levels of homoeologous transcripts from the A, B and D genomes of wheat (Stamati *et al.*, 2009). Of course, NGS of whole transcriptomes can also quantify allele-specific expression and will no doubt have great potential in gaining understanding in areas of biology that include epigenetics and the molecular basis of heterosis (Guo *et al.*, 2006, 2008).

Prospects

Increasing interest in eQTL studies is largely because of the prospect of reducing the time and effort required to identify genes underlying quantitative traits. However, simple questions such as 'what is a *cis*- (or a *trans*-) eQTL?' remain unanswered. We can assay how they are elaborated—but not the cause. We would argue that understanding the basis of allelic imbalances in gene expression is academically interesting, but not core to the practical application of the approach. So what issues must be considered carefully prior to undertaking an eQTL experiment and what outcomes should be expected?

First, most published studies have used existing bi-parental populations, existing molecular marker-based linkage maps and potentially valuable phenotypic datasets. Given a likely 'error-rate' of *c.* 5%–15% in legacy molecular marker data (see Luo *et al.*, 2007) we consider this a dangerous strategy that may lead to erroneous interpretation of what is after all an expensive dataset to collect. Given the established principles of using the same transcript abundance dataset to generate a more than sufficient quantity of reliable bi-allelic marker information we would recommend using *de novo* TDM for map construction. In most species, residual error can frequently be spotted as single marker double recombinants and subsequently corrected in the dataset. This clean and consistent map can then be used as a framework for eQTL and phenotypic analysis. Of course, rigour must be applied when establishing the thresholds for assigning an eQTL as a TDM, but the statistical framework for extracting robust data are freely available.

Second, greater equivalence of the sampled tissues will reduce noise in the system. Non-equivalence, for example the result of physiological, developmental or tissue com-

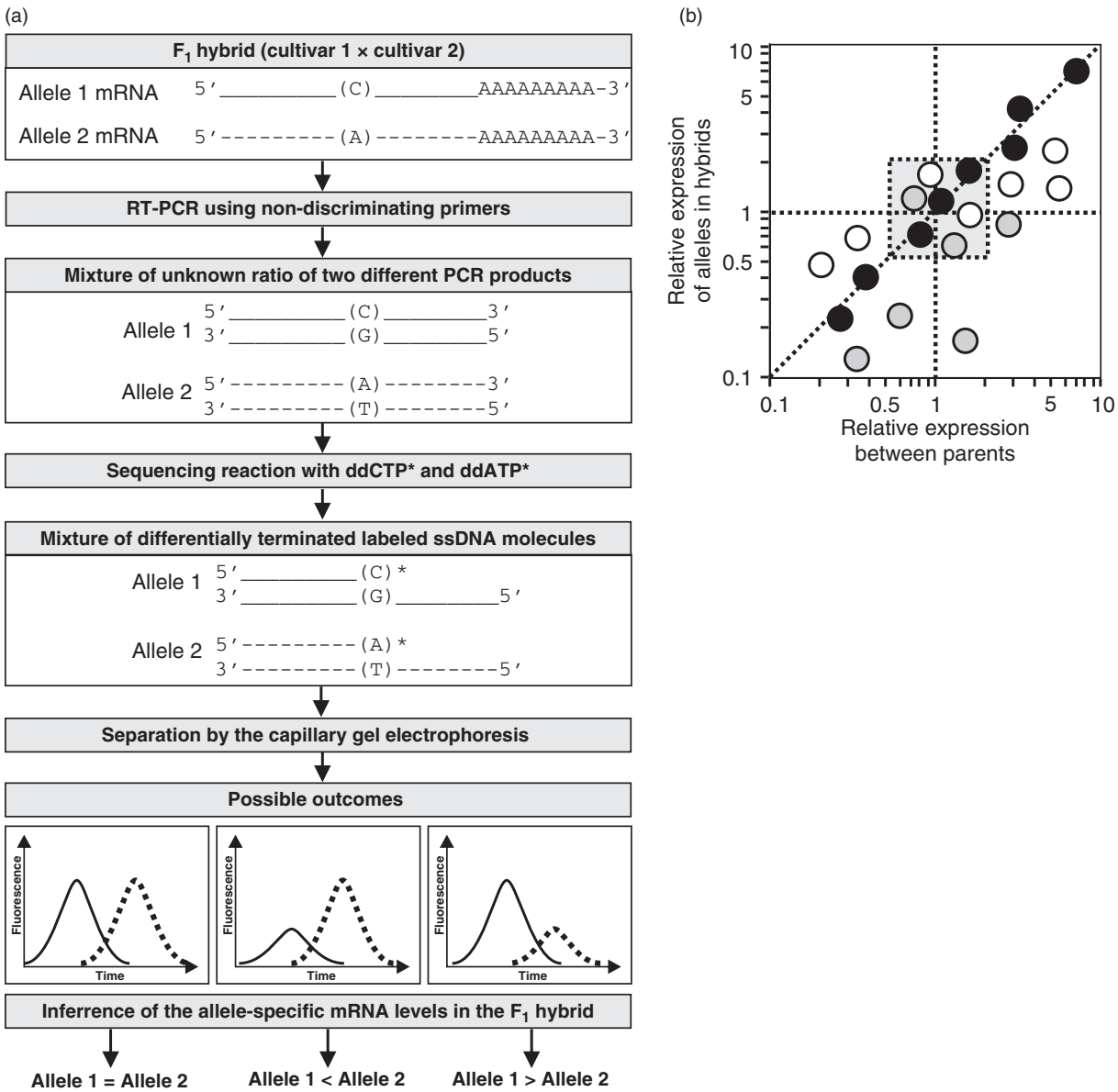

Figure 4 Principle of the allele imbalance assay (adapted from Yan *et al.*, 2002) Panel (a): quantification of *cis*-regulation in F1 hybrids. As both alleles in an F1 hybrid are in an identical physiological, developmental and environmental background, differential allele expression is detected by comparing the peak heights of differently labelled single nucleotide polymorphism alleles on electropherograms. After correction for peak height differences obtained from equivalent quantities of genomic DNA (i.e. alleles), the observed differences are regarded as a true reflection of *cis*-regulation. Panel (b): an extension of the analysis of panel (a), further differentiating *cis*- from *trans*-regulation (modified from Wittkopp *et al.*, 2004) Genes that lie along diagonal are likely to be regulated 100% in *cis*- (black circles), on the dotted horizontal line 100% in *trans* (white circles) and those off the diagonal (grey circles) are partially *cis*- and *trans*-regulated.

plexity differences segregating in a population will influence the general usefulness of the dataset. For example, segregation of a major pleiotropic developmental gene—-such as a dwarfing gene—may influence the value of a dataset for its intended application. Our recommendation would be to take all possible steps towards maintaining 'equivalence' of the sampled tissue. In practice, this may, for example, start the generation before the eQTL experiment is conducted to minimize the impact of 'maternal effects'.

Third, eQTL analysis in crop plants with unsequenced genomes has relied on the use of microarray platforms developed from EST-based gene discovery programs. Because EST collections do not contain all genes in a gen-

ome, any information derived from these platforms will be partial. Furthermore, EST discovery will bias gene discovery towards mRNA that are 'relatively abundant' within the tissues sampled. We do not consider this too much of an impediment. Many genes that are 'causal' to a given phenotype are likely to be of a class that initiate a series of subsequent effects (e.g. *trans*-eQTL or transcription factors) elaborated through the expression of a cascade of downstream events. Being able to monitor the expression of genes in the cascade provides information both on the likely genetic location of the causal gene (that does not need to be on the assay platform) and information on the network of genes that respond to it. Thus, while the information will never be complete, for many applications the richness of the derived datasets is such that positional information will provide a platform for causal gene identification. In addition, given the information is gene based, it is particularly powerful for species—e.g. the large genome small grain cereals—where genomic models exist and can be exploited directly through conservation of synteny. Establishing the predicted gene content of any region thus becomes both easy and powerful. This issue may soon disappear through the use of NGS for transcript profiling, which in addition to being an open system, has brought the potential to conduct comprehensive eQTL studies in any species.

Once a robust eQTL dataset is obtained on a given population, its use can extend to traits not generally assayed in the tissue that has been sampled. A clear example of this is the data referred to above relating to *Rpg1* (Druka *et al.*, 2008b). There, the eQTL dataset used for analysis was from uninfected germinating embryo. The resistance phenotype would not normally have been assayed in that tissue but despite this the most highly correlated sequences made clear biological sense. Consequently, we use this dataset widely in our laboratory as a quick point of reference for any new phenotypic trait subject to investigation. This is important because eQTL experiments can be expensive and only a small fraction of the information is generally utilized by the originating laboratory. To provide broader community access, the barley GG data set has been integrated into the GeneNetwork (http://www.genenetwork.org). This enables straightforward testing of multiple genetic hypotheses using pre-compiled higher-order phenotypic trait information, mRNA abundance, genotype or custom data sets (Druka *et al.*, 2008a). Currently eQTL data and a limited number of barley tissue types are represented in the GeneNetwork. Our intention is to extend the data set by generating and integrating novel phenotypic, mRNA abundance and genotype data using different recombinant line populations and tissue types.

Despite the progress reported above, eQTL analysis in plants remains in its infancy. The studies conducted to date have nevertheless been sufficiently encouraging to stimulate interest and activity across the plant genetics community. Robust statistical algorithms now exist and are accessible to the research community for marker discovery and general eQTL analyses. However, the real promise of these highly dimensional datasets most likely lies in systems genetic interpretation of biological processes or phenomena (Jansen *et al.*, 2009). As the ultimate outcome of such analyses will be 'candidate genes' that are supported by various lines of evidence, the need for efficient strategies for validating their role, often in polygenic processes, remains pressing.

Acknowledgements

Much of the work referred to above was supported by grants from the BBSRC/RERAD (Ref: SCR/910/04) to R. Waugh and M.J. Kearsey and RERAD Program 1, WP1.1. to R. Waugh and A. Druka.

References

Alberts, R., Terpstra, P., Li, Y., Breitling, R., Nap, J.-P. and Jansen, R.C. (2007) Sequence polymorphisms cause many false *cis* eQTL. *PLoS ONE*, **2**(7), e622.

Bennett, M.D. and Smith, J.B. (1976) Nuclear DNA amounts in angiosperms. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **274**, 227–274.

Borevitz, J.O., Liang, D., Plouffe, D., Chang, H.S., Zhu, T., Weigel, D., Berry, C.C., Winzeler, E. and Chory, J. (2003) Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* **13**, 513–523.

Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J. and Corcoran, K. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634.

Brogna, S. and Wen, J. (2009) Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat. Struct. Mol. Biol.* **16**, 107–113.

Chen, J., Agrawal, V., Rattray, M., West, M.A., St Clair, D.A., Michelmore, R.W., Coughlan, S.J. and Meyers, B. (2007) A comparison of microarray and MPSS technology platforms for expression analysis of Arabidopsis. *BMC Genomics*, **8**, 414.

Churchill, G.A. (2002) Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* **32**, 490–495.

Close, T.J., Wanamaker, S.I., Caldo, R.A., Turner, S.M., Ashlock, D.A., Dickerson, J.A., Wing, R.A., Muehlbauer, G.J., Kleinhofs, A. and Wise, R.P. (2004) A new resource for cereal genomics: 22K barley GeneChip comes of age. *Plant Physiol.* **134**, 960–968.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M. and Lathrop, M. (2009) Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10**, 184–194.

Coughlan, S., Agrawal, V. and Meyers, B. (2004) A comparison of global gene expression measurement technologies in *Arabidopsis thaliana*. *Comp. Funct. Genom.* **5**, 245–252.

Cowles, C.R., Hirschhorn, J.N., Altshuler, D. and Lander, E.S. (2002) Detection of regulatory variation in mouse genes. *Nat. Genet.* **32**, 432–437.

Cui, X.P., Xu, J., Asghar, R., Condamine, P., Svensson, J.T., Wanamaker, S., Stein, N., Roose, M. and Close, T.J. (2005) Detecting single-feature polymorphisms using oligonucleotide arrays and robustified projection pursuit. *Bioinformatics*, **21**, 3852–3858.

Czechowski, T., Bari, R.P., Stitt, M., Scheible, W.R. and Udvardi, M.K. (2004) Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant J.* **38**, 366–379.

DeCook, R., Lall, S., Nettleton, D. and Howell, S.H. (2006) Genetic regulation of gene expression during shoot development in Arabidopsis. *Genetics*, **172**, 1155–1164.

Druka, A., Druka, I., Centeno, A.G., Li, H., Sun, Z., Thomas, W.T., Bonar, N., Steffenson, B.J., Ullrich, S.E., Kleinhofs, A., Wise, R.P., Close, T.J., Potokina, E., Luo, Z., Wagner, C., Schweizer, G.F., Marshall, D.F., Kearsey, M.J., Williams, R.W. and Waugh, R. (2008a) Towards systems genetic analyses in barley: integration of phenotypic, expression and genotype data into GeneNetwork. *BMC Genet.* **18**, 9–73.

Druka, A., Potokina, E., Luo, Z.W., Bonar, N., Druka, I., Zhang, L., Marshall, D.F., Steffenson, B.J., Close, T.J., Wise, R.P., Kleinhofs, A., Williams, R.W., Kearsey, M.J. and Waugh, R. (2008b) Exploiting regulatory variation to identify genes and loci underlying quantitative traits in barley. *Theor. Appl. Genet.* **117**, 261–272.

Dupuis, J. and Siegmund, D. (1999) Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics*, **151**, 373–386.

Edwards, A.C., Ayroles, J.F., Stone, E.A., Carbone, M.A., Lyman, R.F. and Mackay, T.F. (2009) A transcriptional network associated with natural variation in *Drosophila* aggressive behavior. *Genome Biol.* **10**, R76.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., Mouy, M., Steinthorsdottir, V., Eiriksdottir, G.H., Bjornsdottir, G., Reynisdottir, I., Gudbjartsson, D., Helgadottir, A., Jonasdottir, A., Jonasdottir, A., Styrkarsdottir, U., Gretarsdottir, S., Magnusson, K.P., Stefansson, H., Fossdal, R., Kristjansson, K., Gislason, H.G., Stefansson, T., Leifsson, B.G., Thorsteinsdottir, U., Lamb, J.R., Gulcher, J.R., Reitman, M.L., Kong, A., Schadt, E.E. and Stefansson, K. (2008) Genetics of gene expression and its effect on disease. *Nature*, **452**, 423–428.

Fu, J. and Jansen, R.C. (2006) Optimal design and analysis of genetic studies on gene expression. *Genetics*, **172**, 1993–1999.

Fukao, T., Xu, K.N., Ronald, P.C. and Bailey-Serres, J. (2006) A variable cluster of ethylene response factor-like genes regulates metabolic and developmental acclimation responses to submergence in rice. *Plant Cell*, **18**, 2021–2034.

Gibson, G. and Weir, B. (2005) The quantitative genetics of transcription. *Trends Genet.* **21**, 616–623.

Guo, M., Rupe, M.A., Yang, X., Crasta, O., Zinselmeier, C., Smith, O.S. and Bowen, B. (2006) Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis. *Theor. Appl. Genet.* **113**, 831–845.

Guo, M., Yang, S., Rupe, M., Hu, B., Bickel, D.R., Arthur, L. and Smith, O. (2008) Genome-wide allele-specific expression analysis using Massively Parallel Signature Sequencing (MPSS™) reveals *cis*- and *trans*-effects on gene expression in maize hybrid meristem tissue. *Mol. Plant Biol.* **66**, 551–563.

Hansen, B.G., Halkier, B.A. and Kliebenstein, D.J. (2008) Identifying the molecular basis of QTLs: eQTL add a new dimension. *Trends Plant Sci.* **13**, 72–77.

Hazen, S.P. and Kay, S.A. (2003) Gene arrays are not just for measuring gene expression. *Trends Plant Sci.* **8**, 413–416.

Hubner, N., Wallace, C.A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., Musilova, A., Kren, V., Causton, H., Game, L., Born, G., Schmidt, S., Müller, A., Cook, S.A., Kurtz, T.W., Whittaker, J., Pravenec, M. and Aitman, T.J. (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.* **37**, 243–253.

Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., Lefkowitz, S.M., Ziman, M., Schelter, J.M., Meyer, M.R., Kobayashi, S., Davis, C., Dai, H., He, Y.D., Stephaniants, S.B., Cavet, G., Walker, W.L., West, A., Coffey, E., Shoemaker, D.D., Stoughton, R., Blanchard, A.P., Friend, S.H. and Linsley, P.S. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotech.* **19**, 342–347.

Jansen, R.C. and Nap, J.P. (2001) Genetical genomics: the added value from segregation. *Trends Genet.* **17**, 388–391.

Jansen, R.C., Tesson, B.M., Fu, J., Yang, Y. and McIntyre, L.M. (2009) Defining gene and QTL networks. *Curr. Opin. Plant Biol.* **12**, 241–246.

Jiang, N., Leach, L.J., Hu, X., Potokina, E., Jia, T., Druka, A., Waugh, R., Kearsey, M.J. and Luo, Z. (2008) Methods for evaluating gene expression from Affymetrix microarray datasets. *BMC Bioinformatics*, **9**, 284.

Jordan, M.C., Somers, D.J. and Banks, T.W. (2007) Identifying regions of the wheat genome controlling seed development by mapping expression quantitative trait loci. *Plant Biotechnol. J.* **5**, 442–453.

Kendziorski, C., Zhang, Y., Lan, H. and Attie, A.D. (2003) The efficiency of mRNA pooling in microarray experiments. *Biostatistics*, **4**, 465–477.

Kerr, K. and Churchill, G.A. (2001) Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–201.

Keurentjes, J.J.B., Fu, J., Terpstra, I.R., Garcia, J.M., van den Ackerveken, G., Basten Snoek, L., Peeters, A.J.M., Vreugdenhil, D., Koornneef, M. and Jansen, R.C. (2007) Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proc. Natl Acad. Sci. USA*, **104**, 1708–1713.

Kirst, M., Myburg, A.A., De Leon, J.P., Kirst, M.E., Scott, J. and Sederoff, R. (2004) 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.* **135**, 2368–2378.

Kirst, M., Basten, C.J., Myburg, A.A., Zeng, Z.B. and Sederoff, R.R. (2005) Genetic architecture of transcript-level variation in differentiating xylem of a eucalyptus hybrid. *Genetics*, **169**, 2295–2303.

Kleinhofs, A., Kilian, A., Saghai Maroof, M.A., Biyashev, R.M., Hayes, P., Chen, F.Q., Lapitan, N., Fenwick, A., Blake, T.K., Kanazin, V., Ananiev, E., Dahleen, L., Kudrna, D., Bollinger, D., Knapp, S.J., Liu, B., Sorrells, M., Heun, M., Franckowiak, J.D., Hoffman, D., Skadsen, R. and Steffenson, B.J. (1993) A molecular, isozyme and morphological map of the barley (*Hordeum vulgare*) genome. *Theor. Appl. Genet.* **86**, 705–712.

Kleinhofs, A., Brueggeman, R., Nirmala, J., Zhang, L., Mirlohi, A., Druka, A., Rostoks, N. and Steffenson, B.J. (2009) Barley stem rust resistance genes: structure and function. *Plant Genome*, **2**, 109–120.

Kliebenstein, D.J. (2009) Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTL. *Annu. Rev. Plant Biol.* **60**, 93–114.

Koornneef, M., Alonso-Blanco, C. and Vreugdenhil, D. (2004) Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annu. Rev. Plant Biol.* **55**, 141–172.

von Korff, M., Radovic, S., Choumane, W., Stamati, K., Udupa, S.M., Grando, S., Ceccarelli, S., Mackay, I., Powell, W., Baum, M. and Morgante, M. (2009) Asymmetric allele-specific expression in relation to developmental variation and drought stress in barley hybrids. *Plant J.* **59**, 14–26.

Kuo, W.P., Liu, F., Trimarchi, J., Punzo, C., Lombardi, M., Sarang, J., Whipple, M.E., Maysuria, M., Serikawa, K., Lee, S.Y., McCrann, D., Kang, J., Shearstone, J.R., Burke, J., Park, D.J., Wang, X., Rector, T.L., Ricciardi-Castagnoli, P., Perrin, S., Choi, S., Bumgarner, R., Kim, J.H., Short, G.F., Freeman, M.W., III, Seed, B., Jensen, R., Church, G.M., Hovig, E., Cepko, C.L., Park, P., Ohno-Machado, L. and Jenssen, T.K. (2006) A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat. Biotechnol.* **24**, 832–840.

Lan, H., Stoehr, J.P., Nadler, S.T., Schueler, K.L., Yandell, B.S. and Attie, A.D. (2003) Dimension reduction for mapping mRNA abundance as quantitative traits. *Genetics*, **164**, 1607–1614.

Liu, Y. and Zeng, Z.B. (2000) A general mixture model approach for mapping quantitative trait loci from diverse cross designs involving multiple inbred lines. *Genet. Res.* **75**, 345–355.

Luo, Z.W., Potokina, E., Druka, A., Wise, R., Waugh, R. and Kearsey, M.J. (2007) SFP genotyping from detects *cis*-acting expression regulators. *Genetics*, **176**, 789–800.

Mehrabian, M., Allayee, H., Stockton, J., Lum, P.Y., Drake, T.A., Castellani, L.W., Suh, M., Armour, C., Edwards, S., Lamb, J., Lusis, A.J. and Schadt, E.E. (2005) Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nat. Genet.* **37**, 1224–1233.

Potokina, E., Druka, A., Luo, Z., Moscou, M., Wise, R., Waugh, R. and Kearsey, M. (2008a) Tissue-dependent limited pleiotropy affects gene expression in barley. *Plant J.* **56**, 287–296.

Potokina, E., Druka, A., Luo, Z.W., Wise, R., Waugh, R. and Kearsey, M.J. (2008b) eQTL analysis of 16,000 barley genes reveals a complex pattern of genome wide transcriptional regulation. *Plant J.* **53**, 90–101.

Potokina, E., Druka, A., Luo, Z., Waugh, R. and Kearsey, M.J. (2009) Transcriptome analysis of barley (*Hordeum Vulgare* L.) using the Affymetrix Barley1 Genechip *Russ. J. Genet.* **45**(11), 81–92.

Ronald, J., Akey, J.M., Whittle, J., Smith, E.N., Yvert, G. and Kruglyak, L. (2005) Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res.* **15**, 284–291.

Rostoks, N., Borevitz, J.O., Hedley, P.E., Russell, J., Mudie, S., Morris, J., Cardle, L., Marshall, D.F. and Waugh, R. (2005) Single-feature polymorphism discovery in the barley transcriptome. *Genome Biol.* **6**, R54.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., Linsley, P.S., Mao, M., Stoughton, R.B. and Friend, S.H. (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297–302.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schön, C.C., Utz, H.F., Groh, S., Truberg, B., Openshaw, S. and Melchinger, A.E. (2004) Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. *Genetics*, **167**, 485–498.

Shi, C., Uzarowska, A., Ouzunova, M., Landbeck, M., Wenzel, G. and Lübberstedt, T. (2007) Identification of candidate genes associated with cell wall digestibility and eQTL (expression quantitative trait loci) analysis in a Flint x Flint maize recombinant inbred line population. *BMC Genomics*, **8**, 22.

Sieberts, S.K. and Schadt, E.E. (2007) Moving toward a system genetics view of disease. *Mamm. Genome*, **18**, 389–401.

Simon, R.M. and Dobbin, K. (2003) Experimental design of DNA microarray experiments. *BioTechniques (Suppl.)*, **34**, 16–21.

Stamati, K., Mackay, I. and Powell, W. (2009) A quantitative genomic imbalance gene expression assay in hexaploid species: wheat (Triticum aestivum). *Genome*, **52**, 89.

Stamatoyannopoulos, J.A. (2004) The genomics of gene expression. *Genomics*, **84**, 449–457.

Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide study. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Street, N.R., Skogstrom, O., Sjodin, A., Tucker, J., Rodriguez-Acosta, M., Nilsson, P., Jansson, S. and Taylor, G. (2006) The genetics and genomics of the drought response in Populus. *Plant J.* **48**, 321–341.

Swanson-Wagner, R.A., Jia, Y., DeCook, R., Borsuk, L.A., Nettleton, D. and Schnable, P.S. (2006) All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. *Proc. Natl Acad. Sci. USA*, **103**, 6805–6810.

Wall, P.K., Leebens-Mack, J., Chanderbali, A.S., Barakat, A., Wolcott, E., Liang, H., Landherr, L., Tomsho, L.P., Hu, Y., Carlson, J.E., Ma, H., Schuster, S.C., Soltis, D.E., Soltis, P.S., Altman, N. and Depamphilis, C.W. (2009) Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics*, **10**, 347.

Wang, M., Hu, X., Li, G., Leach, L.J., Potokina, E., Druka, A., Waugh, R., Kearsey, M.J. and Luo, Z.W. (2009) Robust detection and genotyping single feature polymorphisms from gene expression data. *PLoS Comput. Biol.* **5**, e1000317.

Wenzl, P., Li, H., Carling, J., Zhou, M., Raman, H., Paul, E., Hearnden, P., Maier, C., Xia, L., Caig, V., Ovesná, J., Cakir, M., Poulsen, D., Wang, J., Raman, R., Smith, K.P., Muehlbauer, G.J., Chalmers, K.J., Kleinhofs, A., Huttner, E. and Kilian, A. (2006) A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and agricultural traits. *BMC Genomics*, **7**, 206–228.

West, M.A.L., van Leeuwen, H., Kozik, A., Kliebenstein, D.J., Doerge, R.W., St Clair, D.A. and Michelmore, R.W. (2006) High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis. *Genome Res.* **16**, 787–795.

West, M.A.L., Kim, K., Kliebenstein, D.J., van Leeuwen, H., Michelmore, R.W., Doerge, R.W. and St Clair, D.A. (2007) Global eQTL mapping reveals the complex genetic architecture of transcript level variation in Arabidopsis. *Genetics*, **175**, 1441–1450.

Winzeler, E.A., Richards, D.R., Conway, A.R., Goldstein, A.L., Kalman, S., McCullough, M.J., McCusker, J.H., Stevens, D.A., Wodicka, L., Lockhart, D.J. and Davis, R.W. (1998) Direct allelic variation scanning of the yeast genome. *Science*, **281**, 1194–1197.

Wittkopp, P.J., Haerum, B.K. and Clark, A.G. (2004) Evolutionary changes in *cis* and *trans* gene regulation. *Nature*, **430**, 85–88.

Xu, K., Xu, X., Fukao, T., Canlas, P., Maghirang-Rodriguez, R., Heuer, S., Ismail, A.M., Bailey-Serres, J., Ronald, P.C. and Mackill, D.J. (2006) *Sub1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature*, **442**, 705–708.

Yamashita, S., Wakazono, K., Nomoto, T., Tsujino, Y., Kuramoto, T. and Ushijima, T. (2005) Expression quantitative trait loci analysis of 13 genes in the rat prostate. *Genetics*, **171**, 1231–1238.

Yan, H.Y.W., Velculescu, V.E., Vogelstein, B. and Kinzler, K.W. (2002) Allelic variation in human gene expression. *Science*, **297**, 1143.

Yang, X., Deignan, J.L., Qi, H., Zhu, J., Qian, S., Zhong, J., Torosyan, G., Majid, S., Falkard, B., Kleinhanz, R.R., Karlsson, J., Castellani, L.W., Mumick, S., Wang, K., Xie, T., Coon, M., Zhang, C., Estrada-Smith, D., Farber, C.R., Wang, S.S., van Nas, A., Ghazalpour, A., Zhang, B., Macneil, D.J., Lamb, J.R., Dipple, K.M., Reitman, M.L., Mehrabian, M., Lum, P.Y., Schadt, E.E., Lusis, A.J. and Drake, T.A. (2009) Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nat. Genet.* **41**, 415–423.

Yu, Y., Tompkins, J.P., Waugh, R., Frisch, D.A., Kudrna, D., Kleinhofs, A., Brueggeman, R.S., Muehlbauer, G.J., Wise, R.P. and Wing, R.A. (2000) A bacterial artificial chromosome library for barley (*Hordeum vulgare* L.) and the identification of clones containing putative resistance genes. *Theor. Appl. Genet.* **101**, 1093–1099.

Zhang, L., Miles, M.F. and Aldape, K.D. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.* **21**, 818–821.

Zhang, H.., Sreenivasulu, N., Weschke, W., Stein, N., Rudd, S., Radchuk, V., Potokina, E., Scholz, U., Schweizer, P., Zierold, U., Langridge, P., Varshney, R.K., Wobus, U. and Graner, A. (2004) Large-scale analysis of the barley transcriptome based on expressed sequence tags. *Plant J.* **40**, 276–290.

