Quantitative traits in plants: beyond the QTL

Ilan Paran[1] and Dani Zamir[2]

[1]Department of Plant Genetics and Breeding, Agricultural Research Organization, The Volcani Center, PO Box 6, Bet Dagan 50250, Israel
[2]Department of Field and Vegetable Crops, Faculty of Agriculture, The Hebrew University of Jerusalem, PO Box 12, Rehovot 76100, Israel

Phenotypic variation for quantitative traits results from segregation at multiple quantitative trait loci (QTL), the effects of which are modified by the internal and external environments. Because of their favorable genetic attributes (e.g. short generation time, large families and tolerance to inbreeding), plants are often used to test new concepts in quantitative trait analysis. Thus far, the molecular basis underlying allelic variation at QTL is similar to the identified variation for simple mendelian loci; namely, alterations in gene expression or protein function. Further comprehensive dissection of complex phenotypes will depend on our ability to link genetic components of the QTL variation to genomic databases.

Genetic variation in Nature often takes the form of a quantitative phenotypic range, with an approximately normal distribution, rather than of qualitative phenotypes that fall into discrete categories. The genetic variation underlying quantitative phenotypes, such as human intelligence, body weight, plant yield, etc., results from the segregation of numerous quantitative trait loci (QTL), each explaining a portion of the total variation, and whose expression is modified by interactions with other genes and by the environment [1].

With the advent of co-dominant DNA markers, it became possible to construct saturated genetic maps and to locate QTL for numerous phenotypes in plants, animals and humans [2]. Plants are used as model organisms for the study of quantitative traits because they are particularly amenable to high-resolution mapping and positional cloning. QTL effects and DNA marker polymorphisms are maximized by crossing diverse phenotypes, which often belong to different species or sub-species, and by constructing large experimental populations. Accurate estimates of the mean phenotypic values are facilitated by replicated tests in different environments, which can be achieved through clonal propagation of the genotypes, as well as by evaluating permanent mapping resource populations, such as RECOMBINANT INBREDS or INTROGRESSION LINES (Fig. 1; see Glossary). Such populations, composed of fixed genotypes, can be evaluated by different laboratories for a wide range of traits, thereby creating a comprehensive phenotypic database [3]. The effect of a single QTL, as well as of interactions between QTL, can be efficiently studied by constructing nearly isogenic lines (NILs) that differ only at a single QTL region. Segregating populations, on the order of thousands of individuals, derived from crossing such NILs, can be used to narrow down the position of the QTL to a small genomic region in which candidate genes can be found. Finally, the identity of a QTL is validated by complementation tests by genetic transformation. This paper asserts that QTL have the same molecular basis as regular mendelian genes and that to explore the genetic basis of multiple complex phenotypes, we have to find ways to integrate quantitative genetic information into genomic databases.

The nature of QTL variation

The attributes of QTL mapping in plants have facilitated investigations into the molecular bases of several QTL in *Arabidopsis*, rice, maize and tomato, and revealed that in diverse species, orthologous genetic networks can control related complex phenotypes (Table 1).

In *Arabidopsis*, a LONG-DAY FLOWERING plant, *EDI* (*early day-length insensitive*) is a major flowering-time QTL. NILs that differ for *EDI* alleles of two ecotypes were used for high-resolution mapping and positional cloning of the gene [4]. *EDI* was found to be a novel allele of a blue-light photoreceptor, cryptochrome-2 (*CRY2*). This allele increased protein stability as a result of a single amino-acid substitution that led to early flowering in short days. Similarly, an association study of *Arabidopsis* ecotypes linked quantitative variation in hypocotyl elongation to a

Glossary

Anthesis: The time of expansion of a flower.
Apical dominance: The ability of the apical meristem to prevent side shoots or buds from developing while it is growing.
Transposon tagging: A gene marked by an inserted transposable element where the gene can be cloned directly by isolating the sequences flanking the site of insertion.
Cultivar: A variety of plant produced through selective breeding by humans and maintained by cultivation.
Donor/recurrent parent: In backcross breeding (repeated crossing of an F1 hybrid to one of its parents), the parent that contributes the desired genes is the donor, and the parent to which the genes are transferred is the recurrent one.
Introgression lines: A set of nearly isogenic lines (NILs) developed through a succession of backcrosses, where each line carries a single defined chromosome segment from a divergent genome.
Long-day/short-day flowering: Flowering that requires exposure to light for a period shorter or longer than a critical length.
Mitotic index: The fraction of the total number of cells in a tissue which are actively engaged in mitosis.
Recombinant inbreds: A population of homozygous individuals that is obtained by repeated self crossing from an F1 hybrid, and that contains ~50% of each of the parental genomes in different combinations.

Corresponding author: Dani Zamir (zamir@agri.huji.ac.il).

Table 1. Cloned quantitative trait loci (QTL) in plants

Plant	QTL	Phenotype	Underlying genes	Nature of allelic variation	Ref.
Arabidopsis	*EDI*	Flowering time	Cryptochrome photoreceptor (*CRY2*)	Altered protein function	[4]
Arabidopsis	*PHYA*	Hypocotyl elongation	Phytochrome-A	Altered protein function	[5]
Rice	*Hd1*	Flowering time	Transcription factor (*CONSTANS*)	Loss of function	[7]
Rice	*Hd6*	Flowering time	Protein kinase (*CK2a*)	Loss of function	[8]
Rice	*Hd3a*	Flowering time	*FLOWERING LOCUS T* (*FT*)	Unknown	[9]
Maize	*tb1*	Plant architecture	Transcription factor	Expression level	[12]
Maize	*Dwarf8*	Flowering time	Transcription factor (*GIBBERELLIN INSENSITIVE*)	Unknown	[14]
Tomato	*Brix9–2-5*	Sugar content	Apoplastic invertase (*LIN5*)	Altered protein function	[15]
Tomato	*fw2.2*	Fruit weight	Regulatory gene	Expression level	[16]
Tomato	*Ovate*	Fruit shape	Regulatory gene	Loss of function	[19]

single amino-acid substitution within the photoreceptor phytochrome-A [5].

Rice, which diverged from *Arabidopsis* 150 Myr ago [6] is a SHORT-DAY FLOWERING plant. The two species provide an example of some conservation of developmental pathways that control flowering time. In rice, at least 14 QTL regulate the time of flowering. These QTL were detected in several mapping populations derived from a single cross of rice *Oryza sativa japonica* and *O. sativa indica* CULTIVARS, which differ in flowering time (heading date). Three heading date QTL, *Hd1*, *Hd3a* and *Hd6*, were mapped to a high resolution using NILs and isolated by a map-based cloning approach [7–9]. All three were homologous to *Arabidopsis* genes involved in the control of flowering time. *Hd1* is an ortholog of the *Arabidopsis* flowering-time transcription factor *CONSTANS*. However, in contrast to *CONSTANS*, which promotes flowering under long-day conditions in *Arabidopsis*, *Hd1* delays flowering under long days in rice [10]. Rice *Hd6* encodes the α subunit of protein kinase CK2; this protein kinase phosphorylates a transcription factor involved in the circadian clock, and its reduced expression effects flowering time in *Arabidopsis* [11]. For both *Hd1* and *Hd6*, the cause of the natural variation is loss of function owing to deletion (*Hd1*) or a premature stop codon (*Hd6*). *Hd3a* is a rice homolog of the *FLOWERING TIME* (*FT*) gene that promotes flowering under long-day conditions in *Arabidopsis*. Similar to the regulation of *FT* by *CONSTANS* in *Arabidopsis*, *Hd3a* is regulated by *Hd1* in rice. QTL analysis revealed that although the rice and *Arabidopsis* genes operate in an opposite manner, some functions affecting photoperiodic response are conserved.

Wild species have been instrumental in revealing the nature of QTL variation as exemplified by teosinte, the wild progenitor of modern maize. Maize exhibits much stronger APICAL DOMINANCE than teosinte, and this variation is controlled by *teosinte branched1* (*tb1*). *tb1* was initially detected by QTL analysis in an F2 population, but it was subsequently cloned by TRANSPOSON TAGGING [12]. *tb1* is a transcription factor that, in cultivated maize, suppresses growth of lateral branches. Allelic variation at *tb1* is confined to the regulatory region of the QTL and

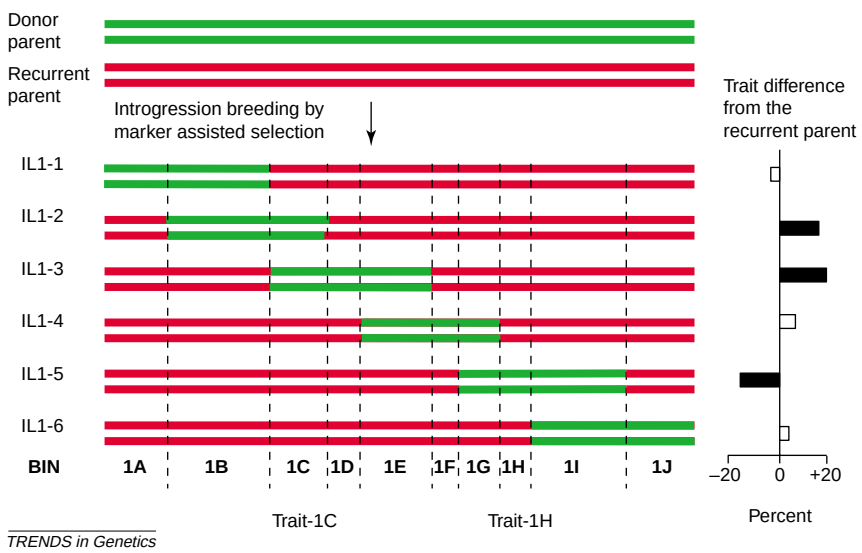

Fig. 1. Introgression-line (IL) bin mapping of a quantitative trait locus (QTL). An exotic accession represented by a pair of homologous chromosomes of the DONOR PARENT (green chromosomes), the RECURRENT PARENT (red chromosome) and six introgression lines (ILs) for chromosome 1. The ILs, which are homozygous for the red chromosomes in the rest of the genome, constitute a set of nearly isogenic lines for chromosome 1. ILs are produced through successive introgression backcrossing and marker-assisted selection to produce a set of recurrent parent lines with single introgressed segments (see Ref. [3] for details). Six ILs covering chromosome 1 create ten mapping bins (1A–1J), each with a unique donor parent composition. A QTL is mapped by phenotyping all lines in randomized replicated trials and presenting the results as percent difference from the recurrent parent (0 value; black bars indicate a significant difference from the common control; empty bars indicate non-significant differences). Combined analysis of the data for all lines defines a QTL that increases the phenotypic value to bin 1C and a reducing QTL to bin 1H. The mapping scheme presented here was adopted from actual tomato data [21]. Such mapping can be obtained to a very high resolution, depending on the number of recombinants available, and the QTL interval can then be superimposed on the physical sequence map. A current bioinformatics challenge is to develop frameworks to explore, on line, the genetic components of the QTL variation for a wide range of phenotypes obtained in different laboratories that study a common genetic resource.

results in the accumulation of a greater amount of transcript of the maize allele than the teosinte allele. Furthermore, sequence comparison of *tb1* in maize and teosinte accessions revealed that selection during domestication at this QTL was confined to the regulatory region [13]. Quantitative variation for flowering time in 92 inbred maize lines was investigated by an association study at *Dwarf8*, which is the ortholog of the wheat gene responsible for the 'Green Revolution' and was determined as a candidate in previous QTL mapping studies. Sequence polymorphism at this locus revealed significant association of several deletions and insertions in the coding and noncoding regions with flowering time. However, the exact cause of variation has not yet been established [14].

In tomato, many QTL mapping studies have been performed for fruit traits such as sugar content, weight and shape. *Brix9–2-5*, a QTL that increases sugar content of the tomato, was delimited through fine-mapping to 484 bp of the apoplastic invertase (*LIN5*), which operates in sugar transport to the developing fruit [15]. Altered enzyme activity as a result of amino-acid substitutions in the gene was determined as the cause for the variation between the cultivated and wild-species alleles (E. Fridman, unpublished).

A major fruit-size QTL in tomato is *fw2.2*, which exemplifies regulation at the transcription level. The cultivated tomato allele of *fw2.2* contributes to the large increase in fruit size that occurred with the domestication of tomato [16]. A series of subsequent studies conducted on NILs that differ for *fw2.2* provided clues as to the function and molecular basis of its action. *fw2.2* is a negative regulator of cell division and is transcribed at low levels at pre- and post-ANTHESIS in all floral organs of both the cultivated and wild plants. Comparative sequencing of the *fw2.2* locus in the genus *Lycopersicon* indicated that the fruit-weight phenotype is associated with variation in a few nucleotides in the promoter region [17]. Although the fruit of the *fw2.2* NILs did not differ in their cell size, their patterns of cell division during fruit development did [18]. In the small-fruited parent, there was an increase in the MITOTIC INDEX immediately after anthesis that was significantly higher than in the large-fruited line. However, cell division in the small-fruited line declined rapidly to levels that were significantly lower than its large-fruited counterpart. Similarly, transcript comparison between the NILs indicated a different pattern of gene expression for the two alleles. Transcript level in the large-fruited line increased to a high level immediately after anthesis but declined rapidly. By contrast, the increase in transcript level in the small-fruited line was slower but sustained for a longer period of fruit development and resulted in twice as many transcripts as the large-fruited line. These results provide evidence that subtle changes in transcript quantity as well as in the timing of gene expression (heterochronic allelic variation) are correlated with natural variation at *fw2.2*. Both *fw2.2* and the recently isolated QTL affecting fruit shape, *OVATE*, are encoded by previously uncharacterized plant genes. *OVATE* controls the transition from round- to pear-shaped fruit, and the cause for the variation is loss of function of the protein because of a premature stop codon [19].

Beyond the QTL

Quantitative variation in experimental and natural populations has been a subject of study for more than a century. However, knowledge of the molecular basis of traits showing continuous distribution was lacking because the factors that regulate the variation had not been identified. The use of NILs, which isolate a single QTL region, transformed the task of QTL cloning into one similar to that performed for simple mendelian traits, with the exception that phenotyping requires more-detailed replicated measurements (Fig. 1). The sequencing of plant genomes, the availability of thousands of markers and improvements in genotyping and phenotyping technologies will enhance map-based cloning of QTL in the future. Other methods for QTL isolation, which do not necessarily require detailed linkage information, are transposon tagging and association studies with candidate genes.

Information about map positions of QTL is already included in genomic databases in the form of a map position and confidence intervals. However, map positions of QTL are only a small fraction of the information that is generated by replicated multi-trait measurements that combine quantitative genetics with marker analysis (in the past ten years more than 500 papers have been published on QTL mapping). The challenge we are facing now is how to develop a framework for presenting, *in silico*, the range of statistical outputs that result from QTL studies; for example, homozygous, heterozygous, pleiotropic, epistatic and environmental effects. This framework, which can be based on the genetic or physical sequence map, will form a basis for further integration of QTL databases with genome information that includes gene content, expression and function.

Variation in QTL alleles in plants has been identified in both coding and regulatory regions of single genes – similar to the variation identified in numerous genes that control qualitative traits. A QTL, therefore, can be regarded as an intermediate stage in the genetic analyses, between a statistically defined locus and a mendelian gene. Powerful genetic tools developed in plants have demonstrated that this transition from a QTL to a mendelian gene is feasible not only for major QTL, but also for minor ones, such as *Hd6* [20]. The ability to associate QTL with sequences has led to the revelation of new functions for known interacting genes and to the discovery of phenotypes for 'genes of unknown function'. The large number of plant QTL that have been mapped to a high resolution will evolve in the coming years to associations of complex phenotypes with their underlying factors. This will create a framework in which to examine how the elusive biological networks interact to create a phenotype.

References

1 Mackay, T.F.C. (2001) The genetic architecture of quantitative traits. *Annu. Rev. Genet.* 35, 303–339
2 Glazier, A.M. *et al.* (2002) Finding genes that underline complex traits. *Science* 298, 2345–2349
3 Zamir, D. (2001) Improving plant breeding with exotic genetic libraries. *Nat. Rev. Genet.* 2, 983–989
4 El-Din-El-Assal, S. *et al.* (2001) A QTL for flowering time in *Arabidopsis* reveals a novel allele of CRY2. *Nat. Genet.* 29, 435–440

5 Maloof, J.N. *et al.* (2001) Natural variation in light sensitivity of *Arabidopsis*. *Nat. Genet.* 29, 441–446
6 Liu, H. *et al.* (2001) Comparative genomics between rice and *Arabidopsis* shows scant collinearity in gene order. *Genome Res.* 11, 2020–2026
7 Yano, M. *et al.* (2000) *Hd1*, a major photoperiod sensitivity quantitative trait locus in rice is closely related to the *Arabidopsis* flowering time gene CONSTANS. *Plant Cell* 12, 2473–2483
8 Takahashi, Y. *et al.* (2001) *Hd6*, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the alpha subunit of protein kinase *CK2*. *Proc. Natl. Acad. Sci. U. S. A.* 98, 7922–7927
9 Kojima, S. *et al.* (2002) *Hd3a*, a rice ortholog of the *Arabidopsis* FT gene, promotes transition to flowering downstream of *Hd1* under short-day conditions. *Plant Cell Physiol.* 43, 1096–1105
10 Samach, A. and Gover, A. (2001) Photoperiodism: the consistent use of CONSTANS. *Curr. Biol.* 11, R651–R654
11 Sugano, S. *et al.* (1998) Protein kinase *CK2* interacts with and phosphorylates the *Arabidopsis* circadian clock-associated 1 protein. *Proc. Natl. Acad. Sci. U. S. A.* 95, 11020–11025
12 Doebley, J. *et al.* (1997) The evolution of apical dominance in maize. *Nature* 386, 485–488
13 Wang, R.L. *et al.* (1999) The limits of selection during maize domestication. *Nature* 398, 236–239
14 Thornsberry, J.M. *et al.* (2001) *Dwarf8* polymorphisms associate with variation in flowering time. *Nat. Genet.* 28, 286–289
15 Fridman, E. *et al.* (2000) A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. *Proc. Natl. Acad. Sci. U. S. A.* 97, 4718–4723
16 Frary, A. *et al.* (2000) *fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. *Science* 289, 85–88
17 Nesbitt, T.C. and Tanksley, S.D. (2002) Comparative sequencing in the genus *Lycopersicon*: implications for the evolution of fruit size in the domestication of cultivated tomatoes. *Genetics* 162, 365–379
18 Cong, B. *et al.* (2002) Natural alleles at a tomato fruit size quantitative trait locus differ by heterochronic regulatory mutations. *Proc. Natl. Acad. Sci. U. S. A.* 99, 13606–13611
19 Liu, J. *et al.* (2002) A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proc. Natl. Acad. Sci. U. S. A.* 99, 13302–13306
20 Maloof, J.N. (2003) QTL for plant growth and morphology. *Curr. Opin. Plant Biol.* 6, 85–90
21 Fridman, E. *et al.* (2002) Two tightly linked QTL modify tomato sugar content via different physiological pathways. *Mol. Genet. Genomics* 266, 821–826

0168-9525/03/$ - see front matter
doi:10.1016/S0168-9525(03)00117-3

Genome Analysis

An evolutionary approach reveals a high protein-coding capacity of the human genome

Anton Nekrutenko[1], Wen-Yu Chung[2] and Wen-Hsiung Li[1,3]

[1]Department of Ecology and Evolution, The University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA
[2]Institute of Information Science, Academia Sinica, Taipei, 115 Taiwan
[3]Computational and Evolutionary Genomics, Genomics Research Center, Academia Sinica, Taipei, 115 Taiwan

We developed a new evolutionary method for identifying exons from genomic sequences and found 19 000 potential coding exons that are absent from all existing annotations of the human genome. Of these, 13 700 satisfied very stringent criteria and can with confidence be considered as novel exons. Evidently, a large number of new human genes can be identified using evolutionary approaches.

Although considerable progress has been made in developing tools for *ab initio* prediction of protein-coding genes, current methods have high false-positive and false-negative rates [1]. It is also unclear how to classify computationally predicted exons that do not match other data, such as expressed sequence tag (EST) or protein sequences. For example, the Mouse Genome Sequencing Consortium used several gene-finding tools that predicted between 14 006 and 48 462 genes, whereas the 'consensus' dataset contained 22 011 genes [2]. Although new gene-finders such as SGP [3], TwinScan [4], DoubleScan [5] and SLAM [6] use comparative information, they were designed to infer gene structures rather than to detect the evolutionary signals of coding regions.

Thus, we aimed to develop an algorithm that complements existing tools by testing the protein-coding potential of a conserved genomic region using a new evolutionary approach. Indeed, we developed the K_A/K_S ratio test [7] (K_A is the rate of substitution per nonsynonymous site, where a base change will lead to a change in amino acid, and K_S is the rate of substitution per synonymous site). The false-positive rate of this test was estimated to be 3% by using computer simulation, and the false-negative rate was estimated to be 8% by using a set of known exons of orthologous human and mouse genes [7]. The method is based on two assumptions: (1) mammalian species, such as human and mouse, share a vast majority of their genes [2,8], and (2) most genes are subject to much stronger selective constraints on nonsynonymous changes than on synonymous ones [9,10]. In this study, we implemented our method (for details see http://nekrut.uchicago.edu/kaks).

Before using our method to find new exons, we estimated the rate at which it recovers known exons that are conserved between human and mouse. We used the dataset of Korf *et al.* [4], which is the most comprehensive comparative dataset currently available. First, we aligned human and mouse sequences using megablast [11] and identified 1860 known exons that were conserved between the two species. Next, we applied our procedure to the alignments from the previous

Corresponding author: Wen-Hsiung Li (whli@uchicago.edu).