

Genetical Genomics: A Few Critical Points

S. Abdolhamid Angaji

Department of Biology, Tarbiat Moallem University, Tehran, Iran

Abstract: The regions identified through quantitative trait loci (QTL) methods are quite large and a considerable amount of effort is required to clone functional genes. Fortunately, during the past two decades, numerous advances in both technology and methodology have greatly improved our ability to collect, measure, and analyze the data necessary to clone gene. The genomic analysis of transcripts as individual phenotypes has led to the emerging field of expression QTL (eQTL) analysis. If gene expression levels are heritable, one can consider each gene's expression level to be a quantitative trait that can be combined with marker data for a linkage study to identify the loci influence variation in the gene's expression. A locus affecting the expression level of a gene has been named eQTL, and the genetic variation causing differences in gene expression are called expression-level polymorphisms.

Key words: Forward genetics, QTL mapping, Expression QTL

INTRODUCTION

There are two general approaches to understanding the function of a gene: forward genetics and reverse genetics. Forward genetics has the goal of trying to find the genetic basis of a phenotype or trait, reverse genetics is aimed at finding the possible phenotypes that may be derived from a specific genetic sequence that is detailed in a DNA sequencing (Pekosz *et al.*, 1999).

Forward genetics refers to a process where studies are initiated to determine the genetic underpinnings of observable phenotypic variation. It begins with a well-characterized phenotype and then works toward identifying the gene(s) responsible for the phenotype. In many cases the observable variation has been induced using a DNA damaging agent (mutagen) such as T-DNA tagging, transposon tagging and gene or enhancer traps which require inserting foreign DNA into a host genome. Genetic mapping approaches such as quantitative trait loci (QTL) mapping and association mapping are also forward genetic approaches and are often used because gene transfer is not required (Tierney and Lamour, 2005; White *et al.*, 2007).

The preliminary aim of QTL mapping is to produce a comprehensive 'framework' (also 'skeletal' or 'scaffold' linkage map) that covers all chromosomes evenly in order to identify markers flanking those QTLs that control traits of interest. There are several more steps required, because even the closest markers flanking a QTL may not be tightly linked to a gene of interest (Michelmore, 1995). This means that recombination can occur between a marker and QTL, thus reducing the reliability and usefulness of the marker. By using larger population sizes and a greater number of markers, more tightly-linked markers can be identified; this process is termed 'high-resolution mapping' (also 'fine mapping'). Therefore, high-resolution mapping of QTLs may be used to develop reliable markers for marker-assisted selection (at least <5 cM but ideally <1 cM away from the gene) and also to discriminate between a single gene or several linked genes (Michelmore, 1995; Mohan *et al.*, 1997).

Three main approaches lead to the cloning of genes of interest. Classical methods such as positional cloning (Rommens *et al.*, 1989, Xu *et al.*, 2006) and insertional mutagenesis (Bechtold *et al.*, 1993) have been used with success to identify major genes, i.e. genes having a major effect on the phenotype. However, these methods are limited by genome size and/or by the lack of transposons in the species being studied (Byrne and McMullen, 1996).

Our ability to identify the molecular basis of quantitative traits is being enhanced by genomic methodologies such as transcriptomics, metabolomics, and proteomics (Keurentjes *et al.*, 2006). The most mature of these approaches is the use of microarrays to measure global transcript levels in mapping populations and map expression QTLs (eQTLs) (Kliebenstein, 2009).

Corresponding Author: S. Abdolhamid Angaji, Department of Biology, Tarbiat Moallem University, Tehran, Iran
Email: Angaji@tmu.ac.ir

Genetical Genomics:

Jansen and Nap (2001) outlined the use of gene expression data in QTL analysis and the approach was termed “genetical genomics”. The basic idea of eQTL is to monitor the expression levels for all genes in breeding experiments, and then find association between the quantitative phenotype and defined patterns of gene expression. In the other word, By developing microarray chips on mapping populations it is now possible to map QTL involved in regulation of gene expression, it is to say, QTL mapping is combined with expression profiling of individual genes in a segregating (mapping) population (Jansen and Nap, 2001). In this approach, total mRNA or cDNA of the organ/tissue from each individual of a mapping population is hybridized onto a microarray carrying a high number of cDNA fragments representing the species/tissue of interest and quantitative data are recorded reflecting the level of expression of each gene on the filter (Gupta and Varshney, 2004). In plants, eQTLs were first identified in maize (Schadt *et al.*, 2003). The dedicated software tool “Expressionview” has also been developed to combine visualization of gene expression data with QTL mapping (Fischer *et al.*, 2003; Cardiff, 2007).

Because eQTL analysis uses segregation population, it is possible to determine whether expression of a target gene is regulated in *cis* (mapping of the differentially regulated candidate gene within the eQTL) or *trans* (the candidate gene is located outside the corresponding eQTL, or in a transcription factor of the gene). The latter gene product is of interest because more than one QTL can be connected to such a *trans*-acting factor (genes acting on the transcription of other genes) (Schadt *et al.*, 2003).

For plant species, such as Arabidopsis and rice, where genome sequence is available, it will be possible to map the individual genes comprising these profiles to specific chromosome regions. A comparison of the coding and regulatory regions of genes comprising the eQTL between the parental strains and progeny could suggest candidate genes. Shortly after genetical genomics was shown to be a powerful tool for identifying candidate genes for traits of economic value in plants (Schadt *et al.*, 2003), a strategy was proposed which relied in identifying *cis*-regulated eQTL that co-localize with a trait QTL. The rationale is that for any trait transcriptionally regulated by a given genomic region (defined by its QTL) there should be a corresponding *cis*-acting eQTL for gene that controls it. Instead of relying on the detection of anonymous markers correlated with a trait, such as in traditional QTL analysis, this approach identifies actual genes. This approach was first applied by Schadt *et al.* (2003) who identified four candidate genes whose eQTL co-localized with several obesity related QTLs in mice. The homologous in human genome had been previously linked to obesity (Lembertas *et al.*, 1997). Following this study, Schadt and his colleagues proposed and demonstrated novel strategies for identifying the genes that control complex, quantitative phenotypes, based on gene expression analysis of segregating population, using specific models (Schadt *et al.*, 2005). The motivation was that an association between transcription levels and trait phenotypic value (demonstrated by the co-localization of QTL and eQTL) may have several origins, some of which are not of interest, and some that may indicate specific target genes. The two most easily confoundable models are the 1) causative and 2) reactive models. In the first, a genetic polymorphism (QTL) causes changes in transcription levels at a given gene, which results in phenotypic variation in a trait of interest. In the second, reactive model, the genetic polymorphism causes a change in the phenotype, which has consequence in the transcript level of one or more genes. In both models, genetic polymorphism, eQTL and QTL may be co-localized. In a third, independent model, the genetic polymorphism causes variation in gene expression and trait, but both are unrelated to each other (Jordan *et al.*, 2007; Varshney and Tuberosa, 2007; Winter and Kahl, 1995).

In most studies, it was found that correlation between candidate gene expression and trait variation is frequently higher than that derived from the QTL analysis. For instance, Kirst and his colleagues (2004) have identified individual QTLs that explained ~20% of the phenotypic variance in tree height. However, certain transcript levels associated with the QTL explained more than 30% of the phenotypic variance. In other study carried out by Schadt *et al.* (2005) in an F₂ mice population, four QTLs that explained together ~39% of phenotypic variance in adiposity related traits were detected. However, transcript levels detected in certain genes explained over 60% of the trait phenotypic variance. Several factors may explain this discrepancy; 1) traditional QTL analysis accounts mostly for additive sources of variance that contribute to the phenotype. So, the QTL analysis of the traits has underestimated the effect of genetic locus. In contrast, analysis of transcript abundance in segregating populations may account not only for additive but also non-additive genetic effects that are not captured in traditional studies (Gibson *et al.*, 2004; Varshney and Tuberosa, 2007), or 2) correlation coefficients between trait and expression may be overestimated (for unknown reason) (Varshney and Tuberosa, 2007).

Some Challenges:

Cost:

Due to the high cost for profiling RNA samples of an entire mapping population, transcriptome profiling based on microarrays is better suited for studies involving a limited number of samples extracted from congenic strains differing at key genomic regions (e.g. NILs) and/or bulked RNA samples obtained from the tails of mapping populations. More recently, cDNA-AFLPs have been used as an alternative to microarrays to identify eQTLs in *Arabidopsis* (Vuylsteke *et al.*, 2006). As compared to microarrays, the cDNA-AFLP approach (i) has a relatively low start-up cost and requires no prior sequence information (Breyne *et al.*, 2003), (ii) avoids bias for abundant transcripts, and (iii) can distinguish the expression of highly homologous genes (Breyne and Zabeau, 2001; Breyne *et al.*, 2003). However, distinct drawbacks of a cDNA-AFLP platform are the limited coverage of the transcriptome and the identification of differential genes, a procedure which requires purification and sequencing of individual AFLP fragments.

Threshold:

The main challenge of any genetical genomics study is to define an appropriate threshold to declare presence of an eQTL. In traditional QTL analysis, entire genome is scanned for significant marker associations and several hundred statistical tests are carried out (Lander and Botstein, 1989; Varshney and Tuberosa, 2007). Strategies to address the problems generated by the multiple number of tests have been proposed and the most popular one is permutation tests. In a study, permutation tests was carried out for 40 genes randomly selected from the set in the microarray, and used the most conservative threshold as defined by null distribution, among those genes (Kirst *et al.*, 2005). When detecting eQTLs from microarray data, since thousands of expression traits are simultaneously test on hundreds or thousands of loci on the whole genome, this problem is magnified by 2-3 orders of magnitude, with the analysis of hundreds or thousands of genes (Deng *et al.*, 2007; Varshney and Tuberosa, 2007).

False Discovery Rate (FDR) may be preferred in this context to control false-positive results. The FDR is the proportion of false positives among all genes that we consider significant. FDR can be viewed as an equivalent of a P-value in experiments with multiple hypotheses testing. In microarray experiments we test simultaneously null-hypotheses for all genes. If there are 20000 genes on a chip, then by using P-value=0.05 we will consider 5% genes significant even if null-hypotheses are true for all genes (i.e., no differential expression). It means that we will get 1000 false positives. However, many issues remain to be addressed more carefully and thoroughly, for example, whether it is safe to consider the distribution of genomewide significance threshold the same for different traits (Deng *et al.*, 2007).

REFERENCE

- Bechtold, N., J. Ellis and G. Pelletier, 1993. In planta *Agrobacterium* mediated gene transfer by infiltration of adult *Arabidopsis thaliana* plants. *Acad. Sci. Paris*, 3: 1194-1199.
- Breyne, P. and M. Zabeau, 2001. Genome-wide expression analysis of plant cell cycle modulated genes, *Curr. Opin. Plant Biol.*, 4: 136-142.
- Breyne, P., R. Dreesen, B. Cannoot, D. Rombaut, K. Vandepoele, S. Rombauts, R. Vanderhaeghen, D. Inze and M. Zabeau, 2003. Quantitative cDNA-AFLP analysis for genome-wide expression studies, *Mol. Genet. Genomics*, 269: 173-179.
- Byrne, P.F. and M.D. McMulle, 1996. Defining genes for agricultural traits: QTL analysis and the candidate gene approach. *Probe.*, 7: 24-27.
- Cardiff, R.D., 2007. Comparative pathobiology of Breast Cancer. IOS Press.
- Deng, H.W., H. Shen and Y. Liu, 2007. Current topics in human genetics: studies in complex diseases. World Scientific Publishing Co. Ptc. Ltd.
- Fischer, G., S.M. Ibrahim, G.A. Brockmann, J. Pahnke, E. Bartocci, H.J. Thiesen, P. Serrano-Fernandez and S. Moller, 2003. Expressionview: visualization of quantitative trait loci and gene-expression data in Ensembl. *Genome Biol.*, 4: R477.
- Gibson, G., R. Riley-Berger, L. Harshman, A. Kopp, S.T. Vacha, S. Nuzhdin and M. Wayne, 2004. Extensive Sex-Specific Nonadditivity of Gene Expression in *Drosophila melanogaster*. *Genetics*, 167: 1791-1799.
- Gupta, P.K. and R.K. Varshney, 2004. Cereal Genomics. Kluwer Academic Publisher.
- Jansen, R.C. and J.P. Nap, 2001. Genetical genomics: The added value from segregation. *Trends in Genetics*, 17: 388-391.

- Jordan, M.C., D.J. Somers and T.W. Banks, 2007. Identifying regions of the wheat genome controlling seed development by mapping expression quantitative trait loci. *Plant Biotechnol. J.*, 5: 442-453.
- Keurentjes, J.B., J. Fu, C.H. Ric De Vos, A. Lommen, R.D. Hall, R.J. Bino, L.H.W van der Plas, R.C. Jansen, D. Vreugdenhil and M. Koornneef, 2006. The genetics of plant metabolism. *Nat. Genet.*, 38: 842-49.
- Kirst, M., A.A. Myburg, J.P. De León, M.E. Kirst, J. Scott and R. Sederoff, 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(4): 2368-2378.
- Kirst, M., J.C.J. Basten, A.A. Myburg, Zhao-Bang Zeng and R.R. Sederoff, 2005. Genetic Architecture of Transcript-Level Variation in Differentiating Xylem of a Eucalyptus Hybrid. *Genetics*, 169: 2295-2303.
- Kliebenstein, D., 2009. Quantitative Genomics: Analyzing Intraspecific Variation Using Global Gene Expression Polymorphisms or eQTLs. *Annu. Rev. Plant Biol.*, 60: 93-114.
- Lembertas, A.V., L. Perusse, Y.C. Changnon, J.S. Fisler, C.H. Warden, D.A. Purcell-Huynh, F.T. Dionne, J. Gagnon, A. Nadeau, A.J. Lusis and C. Bouchard, 1997. Identification of an obesity quantitative trait locus on mouse chromosome 2 and evidence of linkage to body fat and insulin on the human homologous region 20q. *Clin Invest.*, 100(5): 1240-1247.
- Michelmore, R., 1995. Molecular approaches to manipulation of disease resistance genes. *Annu Rev Phytopathol*, 33: 393-427.
- Mohan, M., S. Nair, A. Bhagwat, T.G. Krishna, M. Yano, C.R. Bhatia and T. Sasaki, 1997. Genome mapping, molecular markers and marker-assisted selection in crop plants. *Mol Breed.*, 3: 87-103.
- Pekosz, A., B. He and R.A. Lamb, 1999. Reverse genetics of negative-strand RNA viruses: Closing the circle. *PNAS*, 96(16): 8804-8806.
- Rommens, J.M., M.S. Jannuzzi, B.S. Kerem, M.L. Drumm, G. Melmer, M. Dean, R. Rozmahel, J.L. Cal, D. Kenedy, N. Hideka, M. Zsiga, M. Buchwald, J.R. Riordn, L.C. Tsui and F.S. Cellins, 1989. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science*, 245: 1059-1065.
- Schadt, E.E., S.A. Monks, T.A. Drake, A.J. Lusis, N. Che, V. Colinayo, T.G. Ruff, S.B. Milligan, J.R. Lamb, G. Cavet, P.S. Linsley, M. Mao, R.B. Stoughton and S.H. Friend, 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422: 297-301.
- Schadt, E.E., J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S.K. Sieberts, S. Monks, M. Reitman, C. Zhang, P.Y. Lum, A. Leonardson, R. Thieringer, J.M. Metzger, L. Yang, J. Castle, H. Zhu, S.F. Kash, T.A. Drake, A. Sachs and A.J. Lusis, 2005. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37: 710-717.
- Tierney, M.B. and K.H. Lamour, 2005. An Introduction to Reverse Genetic Tools for Investigating Gene Function. *The Plant Health Instructor*. DOI: 10.1094/PHI-A-2005-1025-01.
- Varshney, R.K. and R. Tuberosa, 2007. *Genomics-assisted Crop Improvement: Genomics approaches and platforms*, Volume 1. Springer.
- Vuylsteke, M., H. Van den Daele, A. Vercauteren, M. Zabeau and M. Kuiper, 2006. Genetic dissection of transcriptional regulation by cDNA-AFLP. *Plant J.*, 45: 439-446.
- White, T.L., W.T. Adams and D.B. Neale, 2007. *Forest Genetics*, CABI, pp: 682.
- Winter, P. and G. Kahl, 1995. Molecular marker technologies for plant improvement. *World J. Microbiol. Biotechnol.*, 11: 438-448.
- Xu, J.R., X. Zhao and R.A. Dean, 2006. From genes to genomes; a new paradigm for studying fungal pathogenesis in *Magnaporthe oryzae* L. *Advances in genetics*, 57: 175-218.