Quantitative genetic variation: a post-modern view

Martin Farrall*

Department of Cardiovascular Medicine, University of Oxford, The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK

Received January 21, 2004; Revised and Accepted February 3, 2004

It has become commonplace to map individual quantitative trait loci (QTL) in experimental organisms; the means (line-crosses and dense maps of markers) and motivation (the close relationship between continuous physiological traits and common, complex diseases) are self-evident. Progress in mapping human QTL has been more gradual, an inevitable consequence of genetic mapping in a natural population setting. The common objective of these studies has been to understand the molecular mechanisms underlying individual QTL. Recent theoretical and practical advances shift this focus to a more comprehensive or genomic perspective on quantitative variation. Fisher’s infinitesimal model of adaptive evolution, which satisfied quantitative geneticists for over 50 years, has been modified in the light of data from QTL mapping experiments in plants and animals. The resulting exponential model provides a pleasing empirical fit to the distribution of QTL effect sizes, predicts that a large amount of quantitative variation will be explained by a limited number of genes and suggests a new mathematical framework for linkage mapping. Molecular analysis of QTL suggests that coding variants (e.g. allozymes) underlie a fraction of quantitative variation and that variants that affect gene expression (expression QTL, eQTL) have a substantial role. This is supported by genomic experiments that combine expression profiling with classical genetic mapping approaches to reveal a remarkable wealth of quantitative heritable variation in the transcriptome and that cis- and trans-acting regulatory factors are organized in networks reflecting pleiotropy. It is hoped that these advances will enhance our understanding of the genetic basis of complex inherited diseases.

INTRODUCTION

Epidemiological studies have successfully identified biochemical, physiological and anthropometric risk factors for common, complex diseases such as coronary artery disease. Quantitative trait loci (QTL) substantially influence many of these continuous traits. For instance, roughly half the inter-individual variation in LDL-cholesterol levels is heritable and elevated levels are a renowned risk factor for atherosclerosis. Such ‘intermediate phenotypes’ provide a critical link between genetic variation and clinical outcome. The technical challenge that understandably preoccupies researchers is how to reliably identify individual genes, one at a time, to explain chunks of this heritability. To this end, there has been great interest in devising experimental designs and statistical genetic methodologies that can be used to efficiently map QTLs in humans and experimental line-crosses. Progress in mapping QTLs in plants and animals has been rapid, and it is hoped that the sequencing of various genomes will accelerate the slower positional cloning phase of moving from linkage to locus (1).

In humans, progress is more sedate, despite a range of study designs, statistical methodologies and a cornucopia of genetic markers. This is unsurprising as human studies are carried out on natural out-bred populations which, combined with the range of non-inherited factors (e.g. diet, medication) enjoyed by our species, has a major impact on power.

The primary objective of mapping studies in both humans and experimental organisms has been to identify the molecular basis of individual QTLs. In the reductionist tradition, higher levels of organization of quantitative variation are assumed to be determined by the lower levels and this is reflected in most of the statistical approaches used to analyse QTL mapping data. In this article, I will review recent theoretical and practical advances that provide a genomic perspective on quantitative genetic variation and allow studies at a higher level, closer to the genotype–phenotype interface. The first advance involves theoretical models of quantitative variation and evolutionary adaptation to provide a framework to understand the totality of

*To whom correspondence should be addressed. Tel: +44 1865287601; Fax: +44 1865287501; Email: mfarrall@well.ox.ac.uk
QTL effects, the second entails recent insights into the molecular basis of quantitative variation following the union of genomics technologies with classical genetic mapping.

THEORETICAL MODELS OF QUANTITATIVE GENETIC VARIATION

The study of quantitative variation enthralled researchers as it allows them to peek at the mechanics of evolution. The central process of adaptation, the concept that an organism's genome evolves to become tuned to its environment, was studied mathematically by Fisher (2) in a classic thesis that shaped opinion for over half a century. Fisher used a random mutation model, random with respect to their impact on phenotypes, with the reasonable premise that mutations of large effect would tend to be deleterious and so be rapidly eliminated from the population by natural selection. This expectation was central to his adaptation model, which predicted that heritable traits would be specified by an innumerable number of minute effects. This infinitesimal theory has had a resounding impact on the field and supported the gradualism hypothesis that evolutionary change is an imperceptible, fluid process. However, over the last 20 years or so evidence has been steadily accumulating from QTL mapping experiments in plants and animals that doesn't fit comfortably with this theory (3). Many of these experiments involve line crosses constructed from inbred lines with divergent phenotypes; the unnatural selection process that leads to their creation is expected to have fixed spontaneous mutations associated with quantitative variation. The common findings from these experiments are the localization of individual genes (QTLs) with quantifiable, measurable effects—some loci might even explain 10 or 20% of the inherited variability. These results might have surprised Sir Ronald Fisher (1890–1962), how could these variants with such a substantial influence on continuous traits have arisen under the infinitesimal model? Yet they are consistent with Alan Robertson’s (1920–1989) insight that the distribution of allelic effects might be exponential (4). Two modifications to Fisher’s adaptation model have been proposed (5,6) which suggest a solution to this inconsistency. Firstly, Kimura (5) proposed that random mutations of large effect would have a high probability of fixation if they were favourable. This balances Fisher’s expectation that most favourable mutations that arise will have small effects with correspondingly small fixation probabilities and Kimura predicts that mutations of an intermediate size will drive adaptation. Secondly, Orr (6) proposed that the early steps in the random adaptive walk to the optimum would tend to be longer than those that follow. When the adaptation process begins, there is plenty of adaptive space in which mutations with large effects can be tested (by natural selection), later on in the process when the organism is close to the optimum state the space is cramped so successful mutations must have smaller effects (Fig. 1). Orr’s model combines both modifications and leads to the robust prediction that the distribution of effect sizes is exponential; quantitative variation is determined by a few QTL of (relatively) large effect and an increasing number of genes of progressively smaller effects. It also follows that we expect that the oldest variants in a species would tend to be associated with the largest effects. It is encouraging that the exponential model of QTL effect size has been supported by theoretical studies from an entirely different field, models of metabolite fluxes through metabolic pathways (7).

The exponential model has been applied in various guises to the analysis of human complex traits. Complex segregation analysis, a statistical modeling technique to investigate the

Figure 1. (A) A representation of a two-dimensional Fisher–Orr adaptation model. Each dimension represents the fitness conferred by a quantitative trait; mutations are pleiotropic as they simultaneously influence both traits. The present phenotypic position for the population is shown by the outer black circle, the red central point represents the optimal phenotype and the grey dashed lines show intermediate states. Green arrows represent random mutations of different sizes and illustrate Fisher’s view that large mutations are usually disadvantageous; the largest mutation would be disadvantageous whatever the angle of displacement. Blue arrows show an idealized adaptation process under Orr’s model with four sequential mutations and show how early mutations (1 and 2) tend to have larger effects than later mutations (3 and 4). (B) An exponential distribution of QTL effect sizes was modeled for 250 QTLs with a gamma distribution with scale parameter = 0.03 and shape parameter 0.3 and assuming that the gene effects are additive and jointly sum to 50% heritability. The size of each QTL effect (heritability) is shown on the abscissa, the rank of the QTL (largest, next largest etc.) is shown on the ordinate. So, for example, the largest QTL (rank = 1) is associated with a heritability of 12%, the fifth largest QTL (rank = 5) confers a heritability of 3.55%.
genetic architecture of human quantitative variation, is often performed using a mixed model (8), in which the segregation of a major gene is modeled on a polygenic background. This model represents an extreme example of the exponential model and analyses have often suggested major genes of substantial effect (>10% variance), suggesting that the granularity of quantitative traits is not as fine as predicted by infinitesimal theory. Morton (9) later outlined an explicit exponential model, the beta model, by building on the implications of Sewell Wright (10). This model has been used to give rough estimates of the number of number of oligogenes underlying complex diseases (11).

One urgent application of the exponential model is in QTL linkage mapping. The classic approach to detecting QTL in line crosses involves interval mapping to assess the statistical support to an individual QTL at a specific location (12). This univariate approach is obviously limited in its scope and multivariate methods based on regression models have been subsequently developed (13). A promising, alternative Bayesian approach starts with the premise that there are scores or even hundreds of QTL scattered throughout the genome, and that they have a wide range of effect sizes (14). The analysis involves the simultaneous estimation of effect sizes across all the QTL that conform to an exponential (L-shaped) distribution; the final model might therefore be constrained to be consistent with the total heritable variation in the cross providing a neat and tidy solution. This approach is attractive as it uses information from the total genome-screen rather than focusing on the handful of largest LOD scores. It would also seem to have an obvious application in the study of complex human diseases, in the analysis of both linkage and gene-association data.

To illustrate the distribution of effect sizes under the exponential model, I have calculated the expected distribution of gene effect sizes for a hypothetical quantitative trait assuming that the total heritability conferred by 250 QTL is 50% (Fig. 2). We see that the largest QTL effect is predicted to confer a heritability of 12%, the second largest 8%, the third largest 6% and so on. The 10 genes of largest effect are predicted to jointly explain 43% of the total variation, i.e. 85% of the total heritability. Similar calculations assuming that 1000 QTL jointly confer 50% heritability predict that the gene with the largest effect confers a heritability of 4.5% and 37 QTL jointly explain 85% of the total heritability. This model is certainly speculative but I hope not ridiculous. For instance, the exponential model has been used in a regression analysis of human blood pressure data in an isolated Adriatic population and suggested that roughly a dozen QTL could explain 25% of the heritability with the rest attributable to 400 genes of small effect (15).

We now consider the implications of these observations in the context of human mapping studies. If we adhere to the infinitesimal theory, then all such mapping experiments are hopeless as we would never have enough power to detect QTL conferring tiny effects, even though the genome would be saturated in them! Under the exponential model, it seems plausible that QTL explaining 5% or more of the total heritability might exist. Genes of this magnitude would need enormous numbers of families to be reliably detected in genome-wide linkage screens (of the order of 30,000 unselected sib-pairs). This formidable hurdle may be serendipitously straddled as anthropometric traits (e.g. stature) are often measured as part of complex disease studies and this information can be combined with genome-wide mapping information (16–19). It is recognized that a fraction of these cohorts, the portion with the most extreme phenotypes, contains the majority of the linkage information (20,21); indeed this selective genotyping strategy has been used to good effect to identify human QTLs (e.g. 22). Yet the medium-term future of gene-mapping studies is widely anticipated to depend on gene-association studies with dense SNP maps (23). We might reasonably hope that, in time, genome-wide association screens will sift out sets of provisional loci that include the largest gene effects which can feed confirmatory studies so that eventually individual genes associated with effects of say >1% heritability will be reliably identified. It is then conceivable that one day we will be in a position to compile lists of genes that explain majority of the heritability of quantitative traits (e.g. 60 and 88% for the 1000 and 250 QTL models discussed above). For the residual tiny genes, it seems unlikely that many of them will ever be systematically mapped. This will place a hard limit on the scope of genetic prediction for complex diseases.

THE MOLECULAR COMPLEXITY OF QUANTITATIVE VARIATION

In a seminal study of a superficially straightforward quantitative trait in Drosophila, alcohol dehydrogenase (Adh) enzyme activity, Stam and Laurie (24) discovered a complicated molecular basis for the allelic variation in both coding (fast/slow—threonine/lysine 192) and non-coding (intronic and downstream) regions of the gene. They also noted how several allelic effects may be coupled to operate as ‘super-alleles’ that effectively segregate as a single major gene. Non-coding quantitative trait variants have been implicated in other Drosophila traits (3) as well as in maize and tomatoes (25). These findings resonate to some extent with the experiences of complex human disease mappers, for instance in the study of calpain-10 and type 2 diabetes, in which multiple intronic variants are associated with susceptibility (26) or the association between the INS VNTR and type 1 diabetes believed to involve transcriptionally active variation (27,28). Other examples include regulatory variants in the RANTES gene that modify HIV-1 infection and progression (29) and RUNX1 binding site variation associated with systemic lupus erythematosis (30). Some doubt on the genomic importance of regulatory variants followed an examination of a list of 27,027 mutations recorded in simple mendelian genes which included a modest proportion (0.8%) labeled ‘regulatory variants’ (31). However, much of the data (Human Gene Mutation Database, http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html) is based on studies in which exons have been systematically searched for mutations in cohorts of patients with rare monogenic diseases; the resulting database provides an excellent resource for clinical and molecular geneticists in this respect. However, the ascertainment bias for the mutations has consequences, for instance the database does not include the INS VNTR or calpain-10 SNPs that have
been associated with diabetes, and extrapolation of these findings to QTL or susceptibility genes for complex diseases is of questionable value.

The potential of gene expression variation in driving evolution was anticipated by Mary-Claire King and Alan Wilson (32), who considered the striking similarity of human and chimpanzee protein sequences (99% identity), which cannot explain the profound differences in anatomy and behavior that have evolved over the last 5 million years or so. They proposed that variants affecting gene regulation rather than structural protein sequence could account for these differences. This hypothesis has been supported experimentally by micro-array studies of fish (33), primates and mice (34) examining transcriptome variation across species. However, the King–Wilson hypothesis is unlikely to explain all the differences and the imminent completion of the chimpanzee genome sequencing project is hoped to inform this fascinating debate (35).

Figure 2. (A) A schematic representation of the exon (shown by vertical bars)—intron structure of the human ACE gene. The location of two ancestral recombination breakpoints in white Europeans are shown. (B) Measured haplotype analysis of white European ACE haplotypes grouped into four clades (A–D); means and 95% confidence intervals for standardized ACE activities associated with each clade are shown. Clades A (shown in red) and B (blue) differ at dozens of polymorphic sites and are associated with low and high ACE activities. Clade C was derived from clades A and B following the 5’ ancestral recombination breakpoint; its effect on ACE activity is intermediate. Clade D was derived from clades A and B following the 3’ ancestral recombination breakpoint; it is phenotypically indistinguishable from clade B. (C) Measured haplotype analysis of four ACE SNPs in a black African population. The population frequency and mean (and standard error) standardized ACE activity for each haplotype is shown. The effect of substituting a SNP can be assessed by comparing similar haplotypes. For instance, A-A-ins-A and A-A-ins-G differ at the 31958 SNP; this change is associated with a significant increase in ACE activity. An additive allele-substitution model fits the data well and points out the biggest effects at the 31958 and 31839 sites.
In our laboratory, we have studied the quantitative genetics of the human angiotensin-1 converting enzyme (ACE) as it provides a tractable system that we modestly hope can trail in the footsteps of Adh. ACE is a component of the renin–angiotensin system (RAS), which has an important role in salt and water homeostasis, the maintenance of vascular smooth muscle tone and blood pressure. The enzyme is expressed in the endothelium of many tissues and is membrane-bound, cleavage results in circulating activity which is readily assayed in plasma or serum samples thus facilitating it’s genetic epidemiology. Complex segregation analysis of circulating ACE activity demonstrated marked heritability attributable to a major gene (36). An intronic I/D polymorphism (insertion of an alu sequence) is strongly associated with ACE activity (37); linkage analysis has confirmed that the association is tightly linked to the ACE gene (38,39). Quantitative immunosassays suggest that variable gene expression and not structural differences that modify enzyme kinetics underlies the ACE quantitative variation (40). Studies of rat strains have shown that levels of gene transcription varies with ACE genotype, suggesting that variation of plasma ACE levels is due to differing levels of transcription in endothelial cells (41). Re-sequencing studies have investigated the extensive polymorphism in ACE but no significant coding variants have been found in populations of European or African ancestry (42–44). Haplotyping studies have shown that in white Europeans, ACE haplotypes have a ‘block’ structure that can be decomposed into a limited number of clades on the basis of two ancestral recombination events (45–47) (Fig. 2B). Haplotyping studies show that majority of the ACE-linked quantitative variation maps between the two breakpoints (∼30% of the total trait variability), with a minor portion mapping upstream from the 5′ breakpoint (42,45,47) (Fig. 2B). However, the ‘block’ structure means that there is restricted diversity within each clade, which provides a natural limit to the resolution of mapping within this population. Studies of ACE quantitative variation in populations of African ancestry (43,48,49) show that there is greater sequence and haplotype diversity. Two intronic SNPs (31839insC and A31958G) that map to the central ACE region flanked by ancestral recombination breakpoints show the strongest association (∼20% of total variability); a minor portion (∼5% of total variability) maps to the ACE promoter (42) (Fig. 2C). The important conclusion from these studies is that multiple, common, non-coding polymorphisms underlie a major human quantitative trait locus

It seems that the scope of quantitative genetics is expanding in the post-modern era with the renaissance of genetical genomics (50) (Fig. 3). Expression profiling experiments have produced exciting results in yeast (51) by revealing extensive polygenic variation in gene expression in this simple eukaryote studied in a controlled environment. The abundance of information that can be gleaned from the transcriptome is striking, 38% of genes that showed differential expression in the two yeast strains showed quantitative genetic variation; the QTLs formed two groups, cis-modulation of single genes and trans-modulation of multiple genes (51). The latter group has a direct bearing on the widespread phenomenon of pleiotropy. The scope of these observations is implied by expression array studies of quantitative inherited variation in maize, a murine F2 intercross and lymphoblastoid cell-lines from human CEPH families (52). The union of transcriptomics and genomics has been used to identify individual QTLs associated with complex disease, namely complement factor 5, as a susceptibility gene for experimental allergic asthma in mice (53).

A recent study suggested that most of the quantitative variation in expression differences in a yeast cross-mapped to trans-acting loci (56). Surprisingly, only a minor portion of these loci appeared to co-localize with transcription factors, in fact the trans-acting variation appeared to be dispersed across a range of molecular functions. This finding jars with the recent identification of PHF11, a QTL that influences IgE levels and asthma (57) and contains two PHD (plant homeodomain) zinc fingers and is presumed to regulate transcription. Complementary studies (58,59) detail technical approaches for detailed investigation of cis-acting polymorphism. Such advances in technology are essential in order to deal with the flood of SNPs that affect gene expression that might be detected in genomic screens.
CONCLUSIONS

Since the 1980s, QTL mapping studies in experimental organisms has led to a steady stream of comprehensively mapped traits (3,60). Human QTL mapping has been quietly tracking this well-worn path despite the problems of working in natural populations. We might hope that the technical challenges of moving from linkage to locus will ease with advances in genome databases and experimental methods (1). The big question is the extent to which this ‘bottom-up’ approach will illuminate the broader perspective in which the evolutionary context of pleiotropic characteristics is studied. The emergence of a complementary, pragmatic anti-reductionist approach in which the genotype–phenotype interface is modeled at a genomic level is hoped to provide such insights.

ACKNOWLEDGEMENTS

Many thanks to Richard Copley and Mark Lathrop for their helpful comments and pointers. I am grateful to the British Heart Foundation, European Union Framework V, National Kidney Research Fund and Medical Research Council for their support.

REFERENCES


10. Wright, S. (1968)


