Title
Dissecting complex traits using the Drosophila Synthetic Population Resource

Journal
Trends in Genetics, 30(11)

ISSN
0168-9525

Authors
Long, Anthony D
Macdonald, Stuart J
King, Elizabeth G

Publication Date
2014-11-01

DOI
10.1016/j.tig.2014.07.009

Peer reviewed

Published in final edited form as:
Trends Genet. 2014 November ; 30(11): 488–495. doi:10.1016/j.tig.2014.07.009.

Dissecting Complex Traits Using the *Drosophila* Synthetic Population Resource

Anthony D. Long,
Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, CA 92697

Stuart J. Macdonald, and
Department of Molecular Biosciences, University of Kansas, Lawrence, KS 66045

Elizabeth G. King
Division of Biological Sciences, University of Missouri, Columbia, MO 65211

Abstract

For most complex traits we have a poor understanding of the positions, phenotypic effects, and population frequencies of the underlying genetic variants contributing to their variation. Recently, several groups have developed multi-parent advanced intercross mapping panels in different model organisms in an attempt to improve our ability to characterize causative genetic variants. These panels are powerful and are particularly well suited to the dissection of phenotypic variation generated by rare alleles and loci segregating multiple functional alleles. We describe studies using one such panel, the *Drosophila* Synthetic Population Resource, and the implications for our understanding of the genetic basis of complex traits. In particular, we note that many loci of large effect appear to be multiallelic. If multiallelism is a general rule, analytical approaches designed to identify multiallelic variants should be a priority for both genome wide association studies and multi-parental panels.

Keywords

Multi-parental Population; multiallelism; *Drosophila melanogaster*; complex traits; missing heritability

A Central Question in Biology

A striking feature of biology is that within populations there is a great deal of inter-individual phenotypic variation. This is true for characters of economic importance in domesticated animals and plants (e.g., yield, growth rate, taste), evolutionary interest (e.g., ability to evade predators, traits influencing reproductive success), and medical relevance in humans (e.g., blood pressure, risk of heart disease, predisposition to schizophrenia).

Corresponding Author: A.D. Long (tdlong@uci.edu).

Although variation in these complex traits is in part due to environmental differences experienced during development and growth, estimates of narrow sense heritability tell us that segregating genetic differences often contribute upwards of 50% to the total phenotypic variation [1,2]. Describing the molecular genetic architecture of complex trait variation is of central importance to biology. Do the most prominent genes contributing to variation in a complex trait explain a large or small fraction of the variation? Are causative genes generally biallelic or do they tend to harbor allelic series? Are causative polymorphic nucleotides generally rare or intermediate in frequency? How important are genotype-by-environment or genotype-by-genotype (epistasis) interactions? Are causative sites generally amino acid polymorphisms, *cis*-regulatory variants, or some other molecular class of variation?

Despite the myriad of well-defined questions central to understanding complex trait variation, we have made only modest progress in empirically addressing them. Throughout the 1990s, the community attempted to dissect complex traits using QTL (Quantitative Trait Locus) mapping in the F_1 offspring of a pair of inbred lines. By and large, the approach taken mimicked an experimental program laid out in two landmark papers, one describing a QTL mapping experiment in tomato [3] and a second describing the statistical machinery allowing for interval mapping of any complex trait given some dense set of genetic markers [4]. Once inexpensive PCR-based markers became widely available, QTL were mapped for hundreds of cross/character combinations. Experiments typically mapped a handful of QTL capable of explaining the bulk of genetic variation segregating in a cross, but failed to localize causative sites to intervals smaller than ∼10cM, typically too large to consider positional cloning. After the turn of the millennium, QTL mapping for the genetic dissection of trait variation was supplanted by genome-wide association studies (GWAS), initially driven by the commercial availability of arrays capable of genotyping hundreds of thousands of SNPs (Single Nucleotide Polymorphisms) relatively cost-effectively in humans. Association studies were not a new idea, but a key paper by the Welcome Trust Case Control Consortium that genotyped thousands of cases and controls for several important genetic diseases significantly altered the intellectual landscape [5]. Over the past decade hundreds of GWAS studies have been carried out in humans, and over 1000 replicable SNP-phenotype associations have been detected [6]. A striking observation is that most SNPs significantly associated with complex trait variation in humans explain just a tiny fraction of the known heritable variation in that trait [7], and even associations identified in massive meta-analyses [8] only account for a small fraction of the heritability. The factors responsible for this missing heritability in outbred animal species remain elusive, but experiments with model systems represent a promising avenue for obtaining insight into these phenomena.

Multi-Parental Populations in Model Systems

A strategy that has been recently utilized for the dissection of complex traits in several model systems is the use of Multi-Parental Populations (MPPs). MPPs start with *k* highly-inbred founder strains and through an *n* generation cross create individuals who are genetic mosaics of those founders (Figure 1). Initiating the MPPs from several founder strains greatly expands the number of natural haplotypes segregating at any given gene, providing a

more representative picture of standing variation within the species than mapping experiments initiated from only two founders. After several generations of intercrossing, individuals from a highly-recombinant, synthetic F_n population are then inbred via full-sibling mating to yield Recombinant Inbred Lines (RILs). Multiple generations of intercrossing prior to RIL extraction reduces the expected size of founder chromosomal fragments, resulting in increased QTL mapping resolution. Advanced intercross resources are now available for several organisms, varying in the number of inbred founders, the crossing design, and the number of lines available for mapping (Table 1).

The populations from which MPPs are derived are synthetic in the sense that they are made by intercrossing a modest number of highly inbred and highly characterized founder strains. As such, the genetic architecture of these populations is not completely representative of variation segregating in the natural population from which the founder strains were obtained. The largest such distinction between synthetic and natural populations is the contribution of rare variants to standing variation. In natural populations rare variants as a class may make significant contributions to standing variation, but they are difficult to detect via any sort of association study scheme [9]. In contrast, in a synthetic population rare variants either fail to be sampled or are at a frequency of *1/k*. Thus, they can be studied when sampled, but their contribution to standing variation is difficult to estimate since their frequency in the source population from which the founders are derived is unknown. Despite this limitation, by simultaneously examining several natural alleles at every gene, synthetic populations can elucidate general features of the genetic architecture of complex traits, and these features can guide future work in natural populations.

Strengths of MPPs for Complex Trait Dissection

Users of MPP RILs measure some complex trait of interest in the panel and statistically identify the most likely genomic position of genetic factors affecting the trait. There are key differences between GWAS and MPP frameworks that affect mapping resolution, the power to detect rare alleles, and the ability to identify multiallelic loci.

GWAS seek to detect associations between single biallelic SNPs and a phenotype measured in a panel (e.g., [5,10]). Because linkage disequilibrium (LD) in a population is likely to be modest, the hope is to genotype SNPs at a sufficiently high density that the causative SNP itself is among the genotyped set [11], or barring that is in strong LD with a genotyped site [9]. By contrast, in an MPP, at any given genomic location the statistical test carried out is analogous to a one-way ANOVA with levels corresponding to the *k* founder haplotypes. This testing approach is effective because if the advanced intercross lines are properly constructed, all but regions of the genome tightly linked to the locus being tested are randomized with respect to founder types. Unlike a GWAS that typically tests the effect of a SNP on a phenotype, an MPP tests the effect of a local haplotype a few cM in size on a phenotype, effectively integrating over all the causative variants in that gene region. The implication is that the estimated effects associated with the *k* founder types at any given location are the additive effects of those gene-size or larger alleles with respect to the remainder of the genome. The differences in statistical testing philosophies result in qualitatively different sets of possible inferences in GWAS and MPP frameworks.

Molecular Nature of Functional Allelic Variation

One clear benefit of the MPP design is that it is agnostic as to the molecular nature of functional alleles underlying trait variation. GWAS rely on dense marker sets to ensure causative sites are either genotyped or indirectly targeted via strong LD with a genotyped marker. Currently this can only be achieved cost-effectively with SNP arrays [5] or Illumina resequencing [10,12], resulting in markers that tend to be biallelic SNPs or small INDELs. If causative changes represent other classes of molecular variation, such as repeats, large structural variants, transposable elements, and so on, these are only queried in a GWAS if they happen to be in strong LD with a genotyped SNP. Because the statistical test employed in an MPP is a haplotype-based test of an entire genomic region, the precise molecular nature of the causative allele has no effect on the ability of the experiment to identify it.

The downside is that an MPP is unable to localize an association to the precise molecular variant, thus association studies potentially offer higher resolution than those based on MPPs. Following mapping in an MPP, a 95% confident interval on the location of causative sites will typically encompass 0.5-5 Mb [13-17]. Thus localizing QTL to windows containing several (or several dozen) genes. By contrast, when a significant association is detected in a collection of natural chromosomes, the causative site is likely very close to the marker. Although resolution in an MPP is much higher than that achieved by traditional two-parent QTL mapping, resolution is a weakness of MPP resources compared to GWAS.

Rare Causative Loci

In a collection of natural chromosomes, SNPs can be at any frequency, with rare alleles dominating [18-20]. GWAS have modest power to identify intermediate-frequency causative variants provided the sample size is large. By contrast, if causative alleles happen to be rare (i.e., < 1-5%), as is likely to be the case under a mutation selection balance model [21,22], GWAS have very poor power. This is largely due to the poor power of statistical tests when groups being compared are of vastly unequal sample sizes. By contrast, provided a rare allele is sampled in the founders of an MPP, its frequency will often be fairly common in the mapping panel (with an expectation of $1/k$). Thus, power to detect rare alleles in an MPP is fairly high, provided they are sampled [15]. Unfortunately, while MPPs sample a larger proportion of standing variation than two-parent QTL mapping, the actual fraction of segregating genetic variation captured is unknown.

Allelic Heterogeneity at Causative Loci

If a single gene harbors multiple rare causative variants in weak LD with one another, a GWAS is poorly powered to detect such variants, and such genes may be invisible to the GWAS approach. This scenario, of multiple causative biallelic variants at a single gene, is the architecture one expects to exist if the variation in a complex trait is maintained by recurrent deleterious mutations [21,22], and is exactly the pattern observed for single-gene, Mendelian diseases such as phenylketonuria. In a GWAS framework an individual association test must reach a strict statistical threshold to be declared significant. For a gene harboring multiple causative variants having low LD with one another, each individual variant will account for only a fraction of the variation accounted for by the entire gene, so GWAS power is reduced. A corollary is that under allelic heterogeneity GWAS will

underestimate the effect size of a gene, as it will tend to highlight only the largest effect SNP in a region. “Burden tests” that integrate over potentially causative SNPs may offer a solution, but although their development is an active area of research, it is a difficult statistical problem [22-24]. By contrast, in an MPP, the size of founder blocks is generally larger than a gene, so the *k*-way ANOVA carried out at each genomic position automatically considers the possibility of multiallelism. One possible explanation for both the missing heritability in humans [7], and the observation that QTL fine-mapped in MPPs generally explain much larger fractions of the genetic variation [17] than GWAS loci [25] is the failure of routine GWAS methods to account for the possibility of allelic heterogeneity at causative genes.

MPP Case Study: The *Drosophila* Synthetic Population Resource

We developed and distribute the *Drosophila* Synthetic Population Resource (DSPR; www.flyrils.org; [14,15]; Box 1), the largest MPP resource available for an animal model system. The initial mapping experiment employing the DSPR was a proof-of-principal study to empirically demonstrate its power and accuracy [14]. Pioneering work from Cathy Laurie and colleagues demonstrated that sites segregating at the *Alcohol dehydrogenase* (*Adh*) locus, in particular the *Fast/Slow* allozyme polymorphism, have a major effect on the observed activity of the ADH enzyme *in vivo* [26-28]. We measured ADH enzyme activity in all DSPR RILs, and mapped a large-effect QTL that implicated just 22 genes, including *Adh*. It is quite clear that ADH activity in the RILs is strongly correlated with the allele present at the nonsynonymous site underlying the *Fast/Slow* polymorphism. Because the founder haplotype means at the *Adh*-associated QTL do not clearly fall into just two groups, other sites in the region are additionally responsible for conferring some activity variation. One of these sites is likely to be the intronic ∇1 insertion/deletion event that influences the amount of ADH protein produced [27], along with at least one other functional variant predicted to reside at *Adh* [28].

In our ADH activity study, we were able to map a number of loci influencing the phenotype in *trans*. The existence of such loci was predicted [29] but they were not previously localized with any precision. Collectively, our single experiment was able to map eight QTL influencing ADH activity, explaining approximately 60% of the genetic variation for the trait. We have subsequently used the DSPR to map QTL explaining large fractions of the heritable variation for traits such as nicotine resistance [30], and the response to widely-used chemotheraphy drugs [31,32]. Several other groups are also using the DSPR to study additional metabolic, morphological, and behavioral traits (www.FlyRILs.org/Projects). These experiments have enjoyed success, routinely identifying QTL contributing >5% of the genetic variance, where the additive contribution of all mapped QTL explains a large fraction of the heritable variation. This stands is in stark contrast to the picture typically observed in human GWAS studies [7]. The haplotype-based tests used in an MPP such as the DSPR allow the sampled collection of causative sites at a gene to be tested as a group. Thus, while the effects of the individual causative sites might not be distinguishable, causative genes are powerfully-mapped using MPP. In a GWAS, small-effect causative sites at a gene are unlikely to reach genome-wide significance given the large number of SNPs tested and therefore go undetected [25].

Empirical Estimates of Allele Number at eQTL

Transcript abundances of the vast majority of genes are complex traits with heritabilities comparable to visible phenotypes. We characterized the transcriptome in female head tissue using the DSPR, examining more than 11,000 transcripts in ∼600 genotypes [33]. Expression QTL (eQTL) mapping allowed us to describe the genetic architecture of thousands of complex phenotypes, potentially leading to general conclusions regarding the frequencies, numbers of alleles, and effect sizes of loci contributing to complex traits. In addition, a comprehensive map of the genetic variants underlying the expression of all transcripts in a specific tissue brings us closer to the ultimate goal of linking together variation at multiple organizational levels. Overall, we identified local, *cis*-eQTL for ∼70% of all transcripts. As observed in other eQTL studies, *cis*-eQTL have large effect sizes (median = 24% of phenotypic variation), but by no means explain all of the variance in transcript abundances. A substantial portion of the genetic variance remains undiscovered, presumably residing in multiple, small-effect *trans* variants.

A principal advantage of MPP designs is the ability to use a haplotype-based analysis to assess the extent of multiallelism. A major result of our study is the implication that the majority of eQTL are multiallelic, with founder genotype means falling into more than two allelic classes (Figure 2). We estimate three or more alleles are segregating for 95% of *cis*-eQTLs, suggesting allelic heterogeneity may be the norm for complex traits. Other studies have also found evidence for multiallelism at the level of eQTLs in *Drosophila* [34], *Arabidopsis* [35], and humans [36,37], although not in mice [17]. Simply distinguishing between a biallelic model and any multiallelic model is relatively straightforward. However, estimating the number of alleles at a QTL and categorizing founder genotype means into allelic groups remains a significant analytical challenge. Our simulations suggest that when the actual number of alleles is greater than three, we are currently unable to accurately estimate the true value (Figure 2; [33]). This is due both to uncertainty in founder genotype mean estimation and possible shortcomings of model comparison methods.

Mapping Designs

When using a panel of MPP RILs, the experimental design to be employed is an important consideration. The simplest format is to directly phenotype the RILs (e.g., [14,30]), but this may be problematic for fitness-related traits, such as many behavioral and reproductive characters, that are expected to suffer from inbreeding depression. A second strategy is to phenotype the F_1 of crosses between pairs of A and B RILs (e.g., [31,33]). This approach ensures an outbred, heterozygous set of experimental individuals. However, under this design there are 64 possible diploid genotypes at any given position, or 16 additive effects to estimate, so even when several hundred genotypes are examined parameter estimation can be a challenge. A third approach is to cross each RIL to a handful of inbred reference strains and phenotype the F_1 [38]. This approach allows for more convenient parameter estimation as well as the potential to estimate background specific epistasis, so it holds considerable promise, although a more thorough empirical assessment is still required, and certain designs may be more or less powerful for particular traits and/or questions.

From QTL to Causative Variants

Making the transition from statistically mapped QTL to the molecular genetic architecture of complex traits is the ultimate goal, and several avenues are available to accomplish this.

QTL Phasing

For QTL contributing >5% to standing variation, provided several hundred RILs are examined, both empirically-motivated simulations [15] and experiments suggest QTL can be resolved to intervals of 1-2 cM. This is considerably greater resolution than traditional QTL mapping, and in flies such intervals will typically implicate <50 genes, and often suggest clear candidates [14,30-32,39]. An appealing theoretical aspect of the DSPR is the idea of "phasing" mapped QTL, thereby reducing the candidate causative variant list to just a few dozen candidates. Phasing is based on the idea that at a QTL peak it is possible that the average phenotype associated with different founder genotypes falls into two distinct categories, each plausibly representing a single allelic state (as depicted in Figure 3a). When two clear classes present themselves, it is then possible to identify a handful of variants segregating among the founders that are "in-phase" with those means (Figure 3a). For large-effect, apparently bi-allelic QTL, phasing has been empirically demonstrated to work. We used this approach to identify the *Fast/Slow* variant at *Adh* [14], and it has been successfully used in the collaborative cross to identify candidate causative SNPs [39]. If allelic heterogeneity exists then phasing has much less utility (Figure 3b). If, for example, the four odd-numbered founder alleles each associated with a low phenotypic value in Figure 3a are each due to an independent *cis*-regulatory mutation that partially abolishes expression, then phasing cannot be successfully used to identify candidate causative variants.

Moving forward, an important empirical question is what fraction of mapped QTL in the DSPR have a genetic architecture consistent with allelic heterogeneity as opposed to a simple biallelic inheritance. For the traditional traits studied to date, mapped QTL often do not appear to be bi-allelic. These observations are in agreement with the results of *cis*-eQTL mapping experiments where mapped QTL are also rarely biallelic. In concert, these observations suggest that the fraction of QTL segregating for several, as opposed to two, functional alleles may be large for many of the traditional morphological, physiological, and behavioral traits of interest to biologists. An implication is that statistical tests designed to detect genes harboring multiple causative SNPs in GWAS datasets may be more suited to uncovering the genetic basis of complex trait variation than the current, almost universally-employed SNP-centric tests.

Leveraging Multiple Reference Populations

A second community resource for dissecting complex traits in *Drosophila* is the *Drosophila* Genetic Reference Panel (DGRP; [10,12]), a collection of roughly 200 unrelated, completely resequenced inbred lines. The approach to dissection with the DGRP is a GWAS, with some extra power obtained via the replicate phenotyping of inbred lines. As mentioned above, GWAS approaches are most successful in the case of an intermediate frequency biallelic SNP. For example, if there is a causative SNP at intermediate frequency contributing at least 15% (7%) to the variation among line means the DGRP has a ∼75% (12%) chance of

detecting and localizing it (at $\alpha=10^{-6}$ to control the genome-wide false positive rate; see Box 2). A potentially powerful approach would be to use the DGRP to validate QTL identified in the DSPR. If the DSPR can be used to narrow a candidate interval to 1-2 cM, a lower statistical threshold can be used in the DGRP, muting the problem of strict, genomewide correction for multiple testing. However, if QTL are generally multiallelic, localization using the DGRP will be challenging. Imagine a hypothetical gene that contributes 5% to standing variation, where the effect comes from four SNPs in linkage equilibrium, each contributing 1.25% to variation (see Figure 3b). The DSPR has excellent power to map and localize such a QTL (we estimate the power to map a QTL contributing 5% to the variance in line means given 600 RILs phenotyped is ∼90% [15]). However, because each SNP would have only a small effect, the DGRP is unlikely to validate them. We anxiously await the results of experiments that measure the same complex traits in both the DSPR and the DGRP and empirically determine the level of correspondence.

Functional Genomics

Historically the *Drosophila* complex trait community has used the binary GAL4-UAS::RNAi system [40] to suggest a candidate gene is indeed causal. The premise is that if a phenotype changes between the control and the RNAi knockdown strains, one can be reasonably confident they have the correct gene. Such experiments have been fairly convincing when the knockdown is confined to a very precise tissue and/or temporal window, but perhaps less so when the knockdown is constitutive (e.g., an actin GAL4 driver), or the UAS::RNAi construct may be reasonably expected to exhibit pleiotropy (e.g., when knocking down a transcription factor), or the trait of interest might be influenced by a range of genetic pathways (e.g., a behavioral or life history trait).

The ultimate proof of causation will require replacing a candidate gene region in an isogenic background - ideally multiple such backgrounds - with several different natural alleles. There is hope that the CRISPR-Cas9 system will make such precise genome editing routine [41], but such experiments will likely remain sufficiently laborious that it would seem prudent to identify a very strong candidate gene before embarking on such an experimental program. We believe two types of community data will greatly aid researchers in this regard. First, access to eQTLs for several tissues and developmental time points will be critical to the resolution of candidate genes under complex trait QTL. There is much speculation that a large fraction of variation in complex traits is ultimately regulatory in origin [42-44], and a *cis*-regulatory eQTL that overlaps a QTL for a trait, and where there is a strong correlation among the founder effects at the two mapped loci, would represent strong evidence for the target gene making a regulatory contribution to trait variation. Second, given the important role of chromatin accessibility in gene regulation, and the strong association between eQTL and DNase I-hypersensitive sites [45,46], we suggest that maps of DNase I sensitivity QTL (dsQTL) for a range of tissues or timepoints would represent a valuable resource for investigators dissecting complex traits to the level of variants in transcription factor binding sites.

Lessons from Model Systems

The DSPR, and similar frameworks available in other model systems, are critical for research into the genetic basis of complex, biomedically relevant traits. In both model systems and human populations, identifying the genes responsible for complex trait variation implicates new genes and new pathways in the control of trait variation. By leveraging the battery of tools available, model systems have the potential to go beyond what is possible in human populations, dissecting the fundamental properties of trait architecture at a mechanistic, functional level. Current and forthcoming projects using the DSPR, and work on mammalian MPPs, consistently demonstrate high mapping power, and often map QTL contributing much greater fractions of the genetic variation than do GWAS studies in human populations. An interesting observation that has already arisen from animal MPPs is that mapped QTL can fail to exhibit segregation patterns indicative of simple biallelic inheritance [17,33]. Instead, patterns are suggestive of the multiallelism typically found at genes underlying Mendelian disorders, merely with a significant reduction in penetrance of the individual causative variants. If multiallelism is also common at complex trait-associated genes in humans, as has been suggested (cf. [47]), one plausible contributor to the missing heritability problem is the failure of typical human GWAS analysis to explicitly consider this allelic heterogeneity [22]. If multiple, small-effect causative mutations are present at a trait-associated gene, especially if they are relatively rare and on different haplotypes, SNP-by-SNP testing will fail to tag these variants in a single test, and fail to identify all but the largest associations. One major lesson for human GWAS from studies in model systems is to prioritize developing analytical approaches designed to detect multiallelic causative genes. Over the next decade, multiparent advanced intercross resources such as the DSPR are likely to uncover other general lessons that can aid in the dissection of complex traits in humans.

Acknowledgments

NIH NRSA F32 GM099382 (EGK), NIH R01 OD010974 (SJM and ADL), NIH R01 GM085260 (SJM), NIH R01 GM085251 & Borchard Foundation (ADL).

Citations

(2 citations that are in press are in “red”)

1. Falconer, DS.; Mackay, TFC. Introduction to Quantitative Genetics. 4. Longman; 1996.
2. Roff, DA. Evolutionary Quantitative Genetic. Kluwer Academic Pub; 1997.
3. Paterson AH, et al. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature. 1988; 335:721–726. [PubMed: 2902517]
4. Lander ES, Botstein D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics. 1989; 121:185–199. [PubMed: 2563713]
5. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]
6. Welter D, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research. 2014; 42:D1001–6. [PubMed: 24316577]
7. Manolio TA, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–753. [PubMed: 19812666]

8. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet. 2014; 46:234–244. [PubMed: 24509480]
9. Spencer CCA, et al. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 2009; 5:e1000477. [PubMed: 19492015]
10. Mackay TFC, et al. The Drosophila melanogaster Genetic Reference Panel. Nature. 2012; 482:173–178. [PubMed: 22318601]
11. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996; 273:1516–1517. [PubMed: 8801636]
12. Huang W, et al. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Res. 2014; 24:1193–1208. [PubMed: 24714809]
13. Kover PX, et al. A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in Arabidopsis thaliana. PLoS Genet. 2009; 5:e1000551. [PubMed: 19593375]
14. King EG, et al. Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. Genome Res. 2012; 22:1558–1566. [PubMed: 22496517]
15. King EG, et al. Properties and power of the Drosophila Synthetic Population Resource for the routine dissection of complex traits. Genetics. 2012; 191:935–949. [PubMed: 22505626]
16. Valdar W, et al. Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat Genet. 2006; 38:879–887. [PubMed: 16832355]
17. Rat Genome Sequencing and Mapping Consortium. et al. Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. Nat Genet. 2013; 45:767–775. [PubMed: 23708188]
18. Fu YX. Statistical properties of segregating sites. Theor Popul Biol. 1995; 48:172–197. [PubMed: 7482370]
19. Nelson MR, et al. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. Science. 2012; 337:100–104. [PubMed: 22604722]
20. Tennessen JA, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012; 337:64–69. [PubMed: 22604720]
21. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? The American Journal of Human Genetics. 2001; 69:124–137.
22. Thornton KR, et al. Properties and modeling of GWAS when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. PLoS Genet. 2013; 9:e1003258. [PubMed: 23437004]
23. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5:e1000384. [PubMed: 19214210]
24. Wu MC, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89:82–93. [PubMed: 21737059]
25. Park JH, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet. 2010; 42:570–575. [PubMed: 20562874]
26. Laurie CC, et al. Associations between DNA sequence variation and variation in expression of the Adh gene in natural populations of Drosophila melanogaster. Genetics. 1991; 129:489–499. [PubMed: 1683848]
27. Laurie CC, Stam LF. The effect of an intronic polymorphism on alcohol dehydrogenase expression in Drosophila melanogaster. Genetics. 1994; 138:379–385. [PubMed: 7828821]
28. Stam LF, Laurie CC. Molecular dissection of a major gene effect on a quantitative trait: the level of alcohol dehydrogenase expression in Drosophila melanogaster. Genetics. 1996; 144:1559–1564. [PubMed: 8978044]
29. King JJ, McDonald JF. Post-translational control of alcohol dehydrogenase levels in Drosophila melanogaster. Genetics. 1987; 115:693–699. [PubMed: 3108072]
30. Marriage T, et al. Fine-mapping Nicotine Resistance Loci in *Drosophila* Using a Multiparent Advanced Generation Intercross Population. Genetics. 2014; xxx:yyyy–yyyy.
31. Kislukhin G, et al. The Genetic Architecture of Methotrexate Toxicity Is Similar in Drosophila melanogaster and Humans. G3: Genes/Genomes/Genetics. 2013; 3:1301–1310.

32. King EG, et al. Using Drosophila melanogaster to identify chemotherapy toxicity genes. G3: Genes/Genomes/Genetics. 2014; xxx:yyyy–yyyy.
33. King EG, et al. Genetic Dissection of the Drosophila melanogaster Female Head Transcriptome Reveals Widespread Allelic Heterogeneity. PLoS Genet. 2014; 10:e1004322. [PubMed: 24810915]
34. Gruber JD, Long AD. Cis-regulatory variation is typically polyallelic in Drosophila. Genetics. 2008; 181:661–670. [PubMed: 19064705]
35. Zhang X, et al. Genetic architecture of regulatory variation in *Arabidopsis thaliana*. Genome Res. 2011; 21:725–733. [PubMed: 21467266]
36. Lappalainen T, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501:506–511. [PubMed: 24037378]
37. Powell JE, et al. Congruence of additive and non-additive effects on gene expression estimated from pedigree and SNP data. PLoS Genet. 2013; 9:e1003502. [PubMed: 23696747]
38. Macdonald SJ, Long AD. Joint estimates of quantitative trait locus effect and frequency using synthetic recombinant populations of *Drosophila melanogaster*. Genetics. 2006; 176:1261–1281. [PubMed: 17435224]
39. Aylor DL, et al. Genetic analysis of complex traits in the emerging Collaborative Cross. Genome Res. 2011; 21:1213–1222. [PubMed: 21406540]
40. Duffy JB. GAL4 system in Drosophila: A fly geneticist's swiss army knife. genesis. 2002; 34:1–15. [PubMed: 12324939]
41. Gratz SJ, et al. Highly Specific and Efficient CRISPR/Cas9-Catalyzed Homology-Directed Repair in Drosophila. Genetics. 2014; 196:961–971. [PubMed: 24478335]
42. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science. 1975; 188:107–116. [PubMed: 1090005]
43. Gibson G, Weir B. The quantitative genetics of transcription. Trends Genet. 2005; 21:616–623. [PubMed: 16154229]
44. Gilad Y, et al. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. 2008; 24:408–415. [PubMed: 18597885]
45. Degner JF, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012; 482:390–394. [PubMed: 22307276]
46. Gaffney DJ, et al. Dissecting the regulatory architecture of gene expression QTLs. Genome Biol. 2012; 13:R7. [PubMed: 22293038]
47. McClellan J, King MC. Genetic Heterogeneity in Human Disease. Cell. 2010; 141:210–217. [PubMed: 20403315]
48. Cridland JM, et al. Abundance and Distribution of Transposable Elements in Two Drosophila QTL Mapping Resources. Mol Biol Evol. 2013; 30:2311–2327. [PubMed: 23883524]
49. Baird NA, et al. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE. 2008; 3:e3376. [PubMed: 18852878]
50. Collaborative Cross Consortium. The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics. 2012; 190:389–401. [PubMed: 22345608]
51. Huang X, et al. Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. Proc Natl Acad Sci USA. 2011; 108:4488–4493. [PubMed: 21368205]
52. Threadgill DW, Churchill GA. Ten years of the Collaborative Cross. Genetics. 2012; 190:291–294. [PubMed: 22345604]
53. Cubillos FA, et al. High-resolution mapping of complex traits with a four-parent advanced intercross yeast population. Genetics. 2013; 195:1141–1155. [PubMed: 24037264]
54. McMullen MD, et al. Genetic properties of the maize nested association mapping population. Science. 2009; 325:737–740. [PubMed: 19661427]

Box 1

Characteristics of the *Drosophila* Synthetic Population Resource

The DSPR is derived from two synthetic populations of *D. melanogaster*, populations A and B, each derived from a set of eight founders (seven of the founders are unique to a population, while one founder is shared by both A and B). The synthetic populations were created by mass-mating *trans*-heterozygotes obtained from a series of round-robin crosses within a set of eight founders (Figure 1). The populations were immediately split into two pairs of subpopulations (A1, A2, B1, and B2) and the subpopulations maintained for an additional 50 generations of mass mating at an estimated census population size of 500-2000 individuals. At generation 50 we established ∼2500 brother-sister lines that were subsequently inbred for 25 generations via full-sib mating to obtain ∼1700 RILs divided roughly equally between the four synthetic subpopulations. The DSPR is highly powered to detect QTL by virtue of the large number of RILs available [15]. All 15 founder lines were Illumina resequenced using paired end reads to roughly 50×, allowing the identification of virtually every SNP segregating between the RILs in non-repetitive parts of the genome. We have since also identified 7,104 transposable elements segregating in the founders [48], and this data is part of the current release (Release 3; www.FlyRILs.org). Each of the RILs were genotyped using RADseq [14,49] to obtain roughly 10K SNP markers. As the SNP density achieved using RADseq was denser than the average size of a recombinant fragment, we were able to construct a Hidden Markov Model (HMM). For any given RIL at any given location in the genome, the HMM confidently estimates the probability it harbors material derived from each of the 8 possible founders [14].

The number of generations of recombination in an MPP before initiating RIL creation results in a trade-off between founder representation and mapping resolution. The DSPR underwent 50 generations of mass mating in large population cages. This method of creating mosaic genomes via free recombination is contrasted with the three generation fixed funnel mating scheme utilized by the Collaborative Cross (Figure 1). There are advantages and disadvantages to both approaches. The mass mating strategy results in shared recombination events between lines, with early recombination events being represented in a larger number of lines. In addition, during the mass-mating phase, both selection and drift can act on the populations, leading to uneven founder representation in the panel [14]. The Mouse Collaborative Cross panel does not share these features, by virtue of their use of the fixed funnel crossing design [50]. Despite the disadvantages of 50 generations of free recombination there is a major advantage – increased mapping resolution. In the DSPR, recombinant segment sizes average just 3cM and QTL are typically mapped to ∼1-2cM.

Box 2

Power to Detect a Causative Variant Contributing to a Continuous Trait in a Collection of Inbred Lines

We set up a simple simulation model to determine the expected power to detect a single causative variant in a panel of 200 inbred lines. We simulated a single SNP at a frequency of 33% that contributed either 7% or 10% to the variance between line means. We then generated the remaining variance from a normal distribution to simulate a phenotype for each line. We obtained a p-value for each simulated SNP using a t-test. To control the false positive rate given the roughly 3 million expected tests for a genome scan, we used a significance threshold of p = 1e-6. This simulation was performed 1000 times and the power is the resulting proportion of significant tests. The code to reproduce this simulation is below. The code can be easily modified to consider different allele frequencies for the causative site, difference thresholds for significance, and different percent variance between line means explained by the causative site.

```
#### R code ####
# Assume:
# a) 200 Inbred Lines are available
# b) Causative SNP at a frequency of 33%
# c) Causative site contributes 7% or 10% of the
# variance between line means. This translates to 3.5% or
# 5% of Vt if the trait's h^2 is 50%.
# d) alpha is 1e-6 to control FPR @~3M tests. Human GWAS
# use 1e-7.
##################
I <- c(rep(1,66),rep(0,134))
vI <- var(I)
PerVar <- 10
alpha <- 1e-6
vE <- ((100-PerVar)/PerVar)*vI
hit <- 0
for (i in 1:1000){
Y <- I + rnorm(200,0,sqrt(vE))
df <- data.frame(Y=Y,I=I)
out <- t.test(Y~I,data=df)
hit <- hit + (out[[3]] < alpha)
}
cat("%Var Line Means=",PerVar, "\t","Power=",hit/10, "%\n")
##################
```

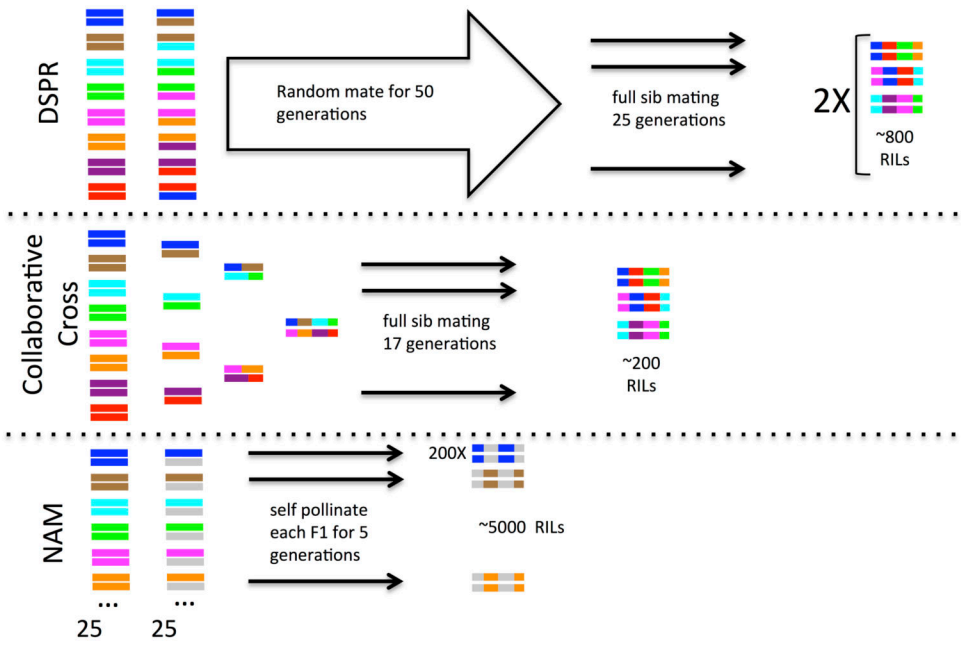

Figure 1.

Some common advanced intercross resources highlighting different crossing schemes. The *Drosophila* Synthetic Population Resource (DSPR) generates an F_1 using a “round-robin” crossing scheme between the 8 highly inbred founder strains. The F_1 are then randomly mated for 50 generations to maximize recombination, and RILs generated by 25 generations of sibling mating. The DSPR involved two parallel crossing schemes each initiated from a different set of 8 founders. The Mouse Collaborative Cross (CC) also starts with 8 highly inbred founders, but mates them using a 3 generation “funnel” cross, followed by roughly 17 generations of brother-sister mating of sub-lines. The CC achieves more even founder representation than the DSPR, at the expense of the size of recombinant fragments. The Nested Association Mapping (NAM) resource of maize starts with 25 inbred founders, but crosses each founder to a common parent (B73) to create 25 highly recombined populations. Unlike the CC and DSPR where 8 chromosomes are segregating in a panel, in NAM only 2 chromosomes segregate in each of the 25 panels.

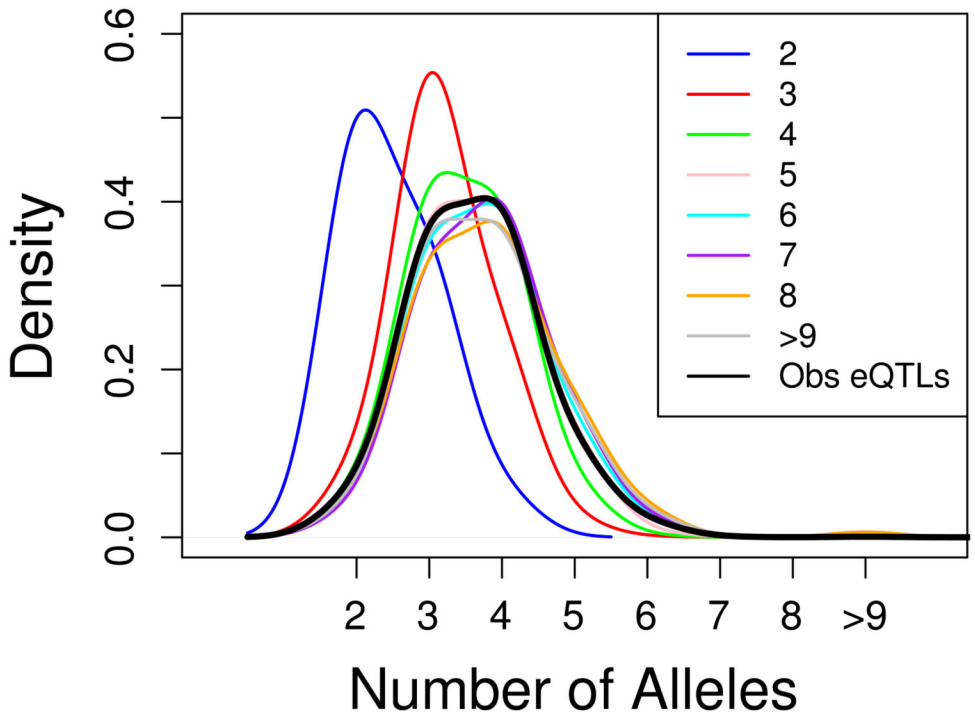

Figure 2.

Density plot of the estimated number of alleles over replicate simulations under models where the true number of alleles (different colored lines) take on different values. Also plotted (thick black line) is the observed distribution of the number of estimated alleles over all detected cis-eQTL. The observed data most closely matches the distribution of the simulated data when QTLs segregate 5 alleles (pink line).

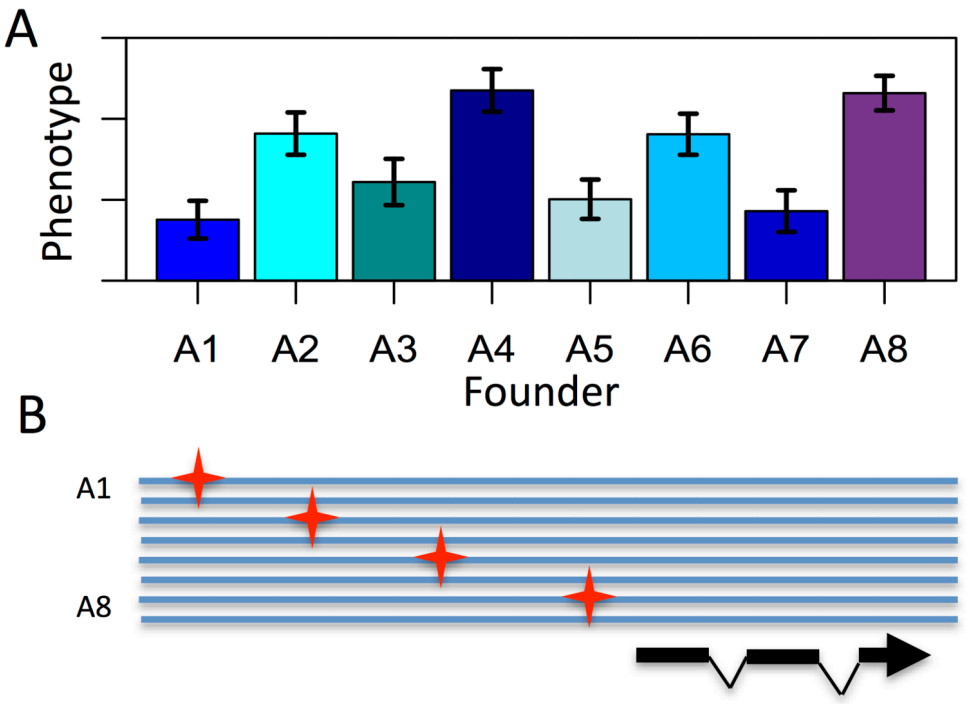

Figure 3.

A) A simulated set of founder means for a QTL explaining 5% of the variance in a set of RILs. Underlying the QTL is a single biallelic SNP, with the low allele present in odd-numbered founders and the high allele present in even-numbered founders. Error bars are SEMs based on 80 RILs per founder allele at the location of the mapped QTL. The plot suggests an underlying biallelic architecture. In the case of a single intermediate frequency causative site the number of "in phase" SNPs is quite modest. For example, for a biallelic causative site with the high allele present in 4 out of 8 founders localized to 1.5cM the expected number of candidate causative SNPs is <10. B) The scenario in which the four low alleles of panel A are not due to a single minor allele, but four independent rare cis-regulatory alleles all in the same gene. Such allelic heterogeneity is expected under a mutation-selection balance model. Under this model each causative SNP only explains ∼1% of the variation between line means. Furthermore, causative SNPs that are private in the DSPR are likely rare in the population as a whole or in a resource like the DGRP. Such causative SNPs are extremely difficult to detect using a GWAS approach as they are both rare and of subtle effect.

Table 1

Characteristics of Available Multi-parental Populations

Panel	Species	K^a	k^b	RILs[c]	n^d	Ref.
AMPRIL http://arabidopsis.info/CollectionInfo?id=138	*Arabidopsis thaliana*	8	8	~500	2	[51]
MAGIC http://mus.well.ox.ac.uk/19genomes/magic.html	*Arabidopsis thaliana*	19	~10[e]	~700	4	[13]
DSPR http://flyrils.org	*Drosophila melanogaster*	15	8	~1700	25	[14,15]
CC http://compgen.unc.edu/wp/?page_id=99	*Mus musculus*	8	8	50-150	3	[39,52]
SGRP-4X	*Saccharomyces cerevisiae*	4	4	0[f]	12	[53]
NAM http://www.panzea.org/http://maizecoop.cropsci.uiuc.edu/nam-rils.php	*Zea mays*	26	2[g]	5000	2	[54]

[a] Total number of founders contributing to resource.

[b] Number of founders per cross.

[c] Number of RILs distributed via website.

[d] Number of generations of recombination prior to initiating RIL construction.

[e] The crosses to create the MAGIC panel were partly random. On average ~10 founders contribute to each RIL.

[f] The *S. cerevisiae* resource does not distribute RILs, but instead the outbred population.

[g] NAM has 25 founders all crossed to B73 in 2 way crosses with 200 RILs per cross.