CellPress

Review

Detecting Polygenic Evolution: Problems, Pitfalls, and Promises

Maren Wellenreuther[1,2,*] and Bengt Hansson[1,*]

Unraveling the genetic basis of organismal form and function remains one of the major goals of evolutionary biology. Theory has long supported a model of polygenic evolution in which quantitative traits are underpinned by many genes of small effect, but empirical methods have lacked the power to detect causative loci when effect sizes are small or moderate. We (i) review traditional approaches used for identifying the molecular basis of phenotypic traits, to highlight the inherent problems and pitfalls that bias them towards the detection of large-effect loci. We then (ii) outline the promises of recent statistical frameworks to detect polygenic signatures of trait evolution, and discuss some of the first studies in evolutionary biology employing these approaches. Lastly, we (iii) outline future directions and point to areas that still need development.

The Search for the Loci that Matter in Evolution

A fundamental goal in evolutionary biology is to identify the genes shaping phenotypes [1]. Achieving this goal has been anything from straightforward, however. Theoreticians have long described phenotypic evolution as a slow process that is driven by weak selection that typically extends long time-periods. The mathematical interpretation of this process is the infinitesimal model, which was introduced in 1918 by Fisher when he demonstrated that the inheritance and evolution of **quantitative traits** (see Glossary) proceeds via selection on an infinite number of unlinked and non-epistatic polygenes of small effect [2,3]. An abundance of theoretical treatments have since emerged corroborating that the majority of quantitative traits are caused by many genes of small and equal effect, suggesting that evolutionary change can be represented as a flux in allele frequency changes of these polygenes (e.g., [4–6]).

While theoretical models overwhelmingly support a model of **polygenic** evolution, the empirical demonstration of polygenes has proven difficult [7]. In the early days, the demonstration of polygenes was hampered by a lack of molecular knowledge and technologies, and it was only after 1980 that it was possible to use polymorphic marker systems [e.g., allozymes, amplified fragment length polymorphisms (AFLPs), microsatellites] to initiate the search for the genes responsible for quantitative phenotypic variation within a formalized framework [8]. Mapped loci via this framework were redubbed **quantitative trait loci** (QTLs), and thereupon became a popular research pursuit. The identification of QTLs can in principle estimate the number of genes responsible for quantitative variation and the size of their effects, but in practice the majority of current approaches carry significant problems. First, the vast majority of study designs are underpowered for detecting polygenes, and thus show an ascertainment bias towards large-effect loci [9–12]. Second, spurious QTLs and skewed effect sizes occur due to non-representative allele frequencies in the mapping population (e.g., few founders), population stratification (e.g., caused by population structure or family structure), or to low environmental

Trends

Understanding the genetic basis of organismal form and function is fundamental in evolutionary biology.

Theoretical work supports models of polygenic evolution, but years of underpowered mapping analyses have biased the literature in favor of large-effect QTLs.

The disconnect between theoretical models and empirical data is troublesome because it distorts our understanding of the molecular targets of selection.

Recent methodological advancements, and improvements in statistics and experimental designs, promise a less-biased empirical evaluation of the causal variants of phenotypic evolution.

Despite these advancements, the application of new methods has been slow, and empirical data powerful enough to genetically dissect polygenic traits are only starting to emerge.

[1]Department of Biology, Lund University, Lund, Sweden
[2]Institute for Plant and Food Research, Nelson, New Zealand

*Correspondence:
maren.wellenreuther@biol.lu.se (M. Wellenreuther) and
bengt.hansson@biol.lu.se (B. Hansson).

variance in laboratory experiments (e.g., leading to an overestimation of the genetic components) [13]. Third, many of the commonly employed experimental designs use **candidate gene** approaches that effectively target large-effect variants *a priori* [14].

In the past years, more powerful methods have been developed that can potentially overcome many pitfalls inherent in the traditional approaches. These approaches, coupled with advances in **next-generation sequencing** (NGS), allows the generation of thousands of markers in any organism, in more detail, and at lower cost than ever before [15], and hold immense promise to obtain a less-biased empirical evaluation of the causal molecular variants of phenotypic evolution. Despite this appealing promise, applications of the new polygenic frameworks are still in their infancy in evolutionary biology, but are already being increasingly applied in the fields of human medicine and agriculture [16–18]. We review here the methods available to generate genotype–phenotype maps by (i) briefly outlining the traditional approaches and discussing their underlying problems and bias towards the detection of the types of genes underlying phenotypic evolution. Then, we (ii) turn to the very recent methodological developments and statistical models that now allow a more-powerful dissection of polygenic evolution. Finally, we outline (iii) how these new developments can be applied to detect polygenic evolution in evolutionary biology, and highlight areas where conceptual uncertainties remain that require further development.

Traditional Approaches and their Problems and Pitfalls

In the pre- and early-**genomics** era, the mapping of genes underlying phenotypic traits employed different approaches that can be coarsely classified as either forward genetics ('top-down') or reverse genetics ('bottom-up'). We do not intend to provide a comprehensive review of the statistical frameworks and assumptions of the methods here, which can be found elsewhere [19–23], but instead we aim to briefly discuss their inherent biases and how these may impact on their suitability to detect genomic regions that correspond to phenotypic traits.

Forward-Genetics Approaches

Forward-genetics approaches start with the measurement of a phenotype followed by associating markers and phenotypic variation to detect causative genes or chromosome regions. The two main procedures for phenotype-driven mapping approaches are (i) **QTL mapping** analysis and (ii) **genome-wide association study** (GWAS) [13,24]. These approaches depend on the existence of a positioned genome-wide marker map, but differ in how the association to the phenotype variation is modeled. QTL mapping measures, loosely speaking, the correlation between marker and phenotype variation in experimental crosses or pedigrees with related, but phenotypically variable, individuals [23]. GWAS aims at obtaining statistical genotype–phenotype associations with physically positioned markers in a set of phenotypically variable but typically unrelated individuals (e.g., in humans [25]). Thus, the main distinction between QTL mapping and GWAS is that the former examines genotype–phenotype associations within controlled crosses or wild pedigrees, and therefore exploits recent recombination events, whereas the latter detects such associations in populations with an old history of recombination and thus with low levels of **linkage disequilibrium** (LD). As a consequence, QTL mapping has low precision but requires fewer markers (one marker every ~1–10 cM), whereas GWAS has higher precision but requires much denser marker maps.

In addition to this main distinction, the two approaches also differ in their power to detect QTLs and in their experimental design flexibility. For example, the power of QTL mapping studies ultimately relies on large families, and these can be difficult to obtain (i.e., mammals often take several years to reach sexual maturity, and then only produce a small number of recombinant offspring [26]). With small family sizes, the power of detecting small- to medium-effect QTLs is limited, which is corroborated by empirical data showing that QTL mapping studies generally

Glossary

Background selection: a process in which weakly deleterious mutations drift to low frequencies and are then purged from the population by negative selection, which causes decreased genetic diversity at linked loci in general and around conserved genes in particular.
Candidate gene: a gene of hypothesized relevance to the studied phenotype. This could be a gene involved in a pathway affecting a phenotype or a gene that has been implicated with the trait in previous studies. Sequencing the gene in individuals with divergent phenotypes can identify mutations which are associated with adaptive variation.
Genome-wide association study (GWAS): also known as association mapping, a trait-mapping approach where polymorphisms across the whole genome are screened for an association with a trait in multiple individuals. Statistical associations between genotype and phenotype only arise when the marker and the causative locus are in strong LD. It relies on historical recombination in the mapping population and has therefore relatively high precision.
Genomics: study of the function and structure of genomes.
Linkage disequilibrium (LD): non-random association of alleles at different loci (often, but not necessarily, in close genomic proximity).
Mendelian trait: a trait controlled by a single locus that is inherited according to Mendel's laws.
Next-generation sequencing (NGS): several different types of high-throughput DNA sequencing methods where hundreds of thousands or millions of reads (sequences) are produced simultaneously.
Omics: a study that targets everything of something. For example, genomics targets all genes in the genome, and transcriptomics targets all expressed gene in the genome.
Polygenic evolution: a process in which adaptation occurs by simultaneous selection operating on variants at many loci (perhaps tens or hundreds or more). A common scenario of polygenic evolution would be that there is a shift in the optimal phenotype for a quantitative trait that is affected by hundreds of alleles of