Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Genetics & Development

Promises and limitations of hitchhiking mapping

Sergey V Nuzhdin[1] and Thomas L Turner[2]

Building the connection between genetic and phenotypic variation is an important 'work in progress', and one that will enable proactive diagnosis and treatment in medicine, promote development of environment-targeted varieties in agriculture, and clarify the limits of species adaptation to changing environments in conservation. Quantitative trait loci (QTL) mapping and genome wide association (GWA) studies have recently been allied to an additional focus on 'hitchhiking' (HH) mapping — using changes in allele frequency due to artificial or natural selection. This older technique has been popularized by the falling costs of high throughput sequencing. Initial HH-resequensing experiments seem to have found many thousands of polymorphisms responding to selection. We argue that this interpretation appears too optimistic, and that the data might in fact be more consistent with dozens, rather than thousands, of loci under selection. We propose several developments required for sensible data analyses that will fully realize the great power of the HH technique, and outline ways of moving forward.

Addresses
[1] Program in Molecular and Computation Biology, University of Southern California, Los Angeles 90089, United States
[2] Department of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, United States

Corresponding authors: Nuzhdin, Sergey V (snuzhdin@usc.edu)

Current Opinion in Genetics & Development 2013, **23**:694–699

This review comes from a themed issue on **Genetics of system biology**

Edited by **Shamil Sunyaev** and **Fritz Roth**

For a complete overview see the Issue and the Editorial

Available online 12th November 2013

http://dx.doi.org/10.1016/j.gde.2013.10.002

From QTL to hitchhiking mapping

Heritable differences among individuals are abundant in almost all populations and for nearly any phenotype. What kinds of genetic variants underlie this variation? What kinds of genes harbor these variants, and how are their effects linked to phenotype? Efforts to answer these questions date back to the early 20th century [1], and they became mainstream in quantitative genetics after [2] popularized the use of molecular markers. Thousands of QTL mapping projects have contributed superb advances. We now know that genes with major-effect mutations on a phenotype also harbor natural alleles with more moderate effects. We also have a reasonably good idea about the distribution of effects sizes for these mutations, and their role in new or fluctuating environments [3].

Major limitations of QTL mapping have also become clear [4]. These experiments are very tedious and labor intensive, as they require developing, genotyping, and maintaining hundreds of recombinant inbred genotypes or accessions. Because of this limitation, the precision of mapping is frequently limited to large regions of chromosomes rather than individual genes. The majority of experiments have ample ability to roughly map larger-effect QTLs. However, power for identifying alleles contributing to the phenotypic variation in more modest, though still sizable way, is substantially less impressive. Whenever a modest-effect allele is discovered, its contribution is typically overestimated (the so-called ''winners' curse'' [5]). Some limitations have been overcome in simpler models, like yeast [6], but others persist. QTL analysis typically starts from crosses of two, or just a few accessions, thus most of natural variation remains untapped. Phenotypes are sometimes scored in individuals that are largely homozygous, thus causing concerns about effects of life history and behavioral phenotypes that strongly depend on inbreeding. Most of all, the task of moving from a large region to a causal polymorphism remains daunting in most systems.

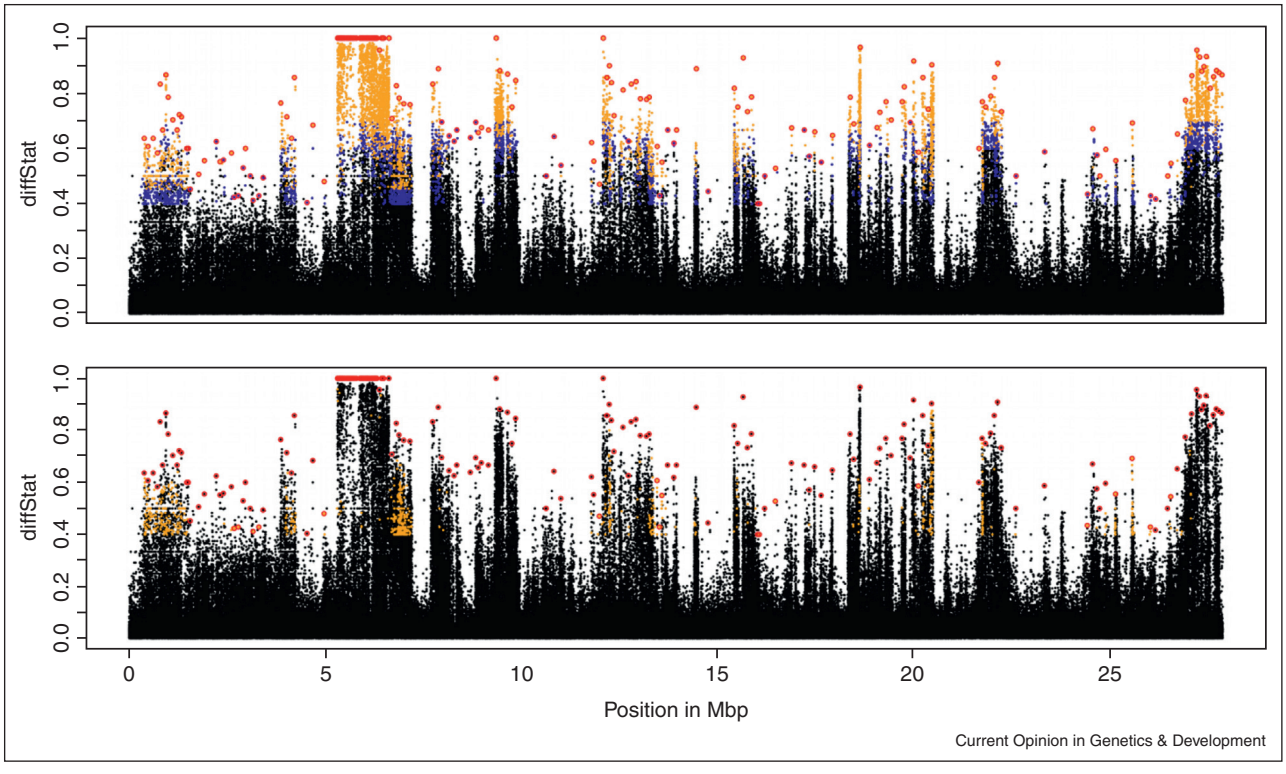
Recently, an alternative mapping technique–to follow frequency changes at marker loci in selected populations–has been gaining popularity. It originally stems from the experiments of Dumouchel and Anderson [7] and Garnett and Falconer [8], and theoretical treatments of Thomson [9] and Thoday [10], but was first formalized as a mapping approach by Lebowitz *et al.* [11]. The idea is that selection changes the frequencies of molecular markers because they hitchhike (HH) with alleles of QTLs of the selected trait [12••], allowing inference of the linkage between the markers and QTLs. This is a very powerful approach, as QTLs with relatively small effects can be detected by genotyping a manageably small number of individuals. Initial experiments had involved crossing two accessions and applying multi-generation selection to their progeny [13,14]; these were then extended to mapping populations originating from a large number of isogenic founders [15] in *Drosophila* [13,16] and mice [14,17]. Those same ideas are applicable to any population under artificial selection, as long as a linkage map is available. Unlike QTL mapping, the technique is applicable to organisms in which controlled crosses are difficult to implement. Additionally, individuals remain largely heterozygous in the multi-founder case, removing the potential confounding of inbreeding depression. This has enabled a genetic dissection of life-history and behavioral characters [18•].

Resequencing and hitch-hiking mapping

Similar to QTL mapping, initial HH mapping relied on recombination breakpoints produced during the experiment, thereby limiting the mapping precision. This was necessary, as only a fraction of the genome could be measured and used as molecular markers, and selection could only be detected if linked to a marker. Recent advances in resequencing technology have greatly reduced this limitation. Researchers can now create a mapping population simply by sampling a large number of individuals in nature. Selection is then applied for several-to-hundreds of generations on replicate populations, and these populations (and preferably the starting population as well) are resequenced. Changes in frequency of most genomic polymorphisms are assayed (some loci and alleles remain difficult to annotate). Turner *et al.* [19] referred to this approach as ''Evolve and Resequence'' (E&R), as (in theory) the detection of neutral hitchhikers is no longer required. Here, we will term this technique HH-resequencing, or HHR, to emphasize the continuity of this approach with previous work. By what ever name, the approach seeks to combine the resolution of population genetic analysis of selection in natural populations (e.g. [20,21]) with the functional precision gained by applying selection to specific characters.

The power of early HHR efforts appears astounding. Turner *et al.* [22•] found 5205 genomic regions putatively responding to selection on body size. Is it possible that selection had acted on just a few polymorphisms, but many regions then appeared differentiated as they were in linkage disequilibrium (LD) with the selected loci? We examine Figure 1 to evaluate such a proposition. The responding polymorphisms appear well-spread over the genome and separated by those not exhibiting large changes in frequency. It thus appears that mapping had high precision, and that a very large number of QTL were found. Given that the extent of strong LD in natural fly populations is on the scale of ~100 base pairs in many genomic regions, such a 'fine-grained' selection response might be possible. However, to be conservative, Turner *et al.* [22•] identified 10 kb windows around strong-responding polymorphisms, and still found that

Figure 1



Differentiated polymorphisms on chromosome arm 3R, reproduced from Turner *et al.* [22•], Figure 4, with a following Figure Legend. The diffStat is shown for each variant that had higher or lower allele frequencies in the large-selected lines compared to the small-selected lines. Above: Color coding indicates significance: black = nonsignificant variants, blue = significant variants at the permissive FDR threshold (FDR < 10%); gold = significant variants at the restrictive FDR threshold (FDR < 5%); red = peak variants. Below: Color coding indicates estimated starting allele frequency: black = all variants, gold = variants with an average control frequency <0.05; red circles indicate peak variants, as in A. When 50 kb regions around strongest selected sites are assumed to be changing in frequency due to local hitch-hiking, the estimate for the number of selected polymorphisms is reduced to ~ 300. http://dx.doi.org/10.1371/journal.pgen.1001336.g004.