Advances and limits of using population genetics to understand local adaptation

Peter Tiffin[1] and Jeffrey Ross-Ibarra[2]

[1] Department of Plant Biology, University of Minnesota, 250 Biosciences, Saint Paul, MN 55108
ptiffin@umn.edu
612-614-7406

[2] Department of Plant Sciences, 262 Robbins Hall, Mail Stop 4, University of California
One Shields Ave, Davis, CA 95616
rossibarra@ucdavis.edu
530-752-1152

Corresponding author: Peter Tiffin ptiffin@umn.edu

Keywords: Clinal adaptation, selection, genomics.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

Abstract

Local adaptation is an important process shaping within species diversity. In recent years, population genetic analyses, which complement organismal approaches in advancing our understanding of local adaptation have become widespread. Here we focus on using population genetics to address some key questions in local adaptation: what traits are involved? what environmental variables are most important? Does local adaptation target the same genes in related species? do loci responsible for local adaptation exhibit tradeoffs across environments? After discussing these questions we highlight important limitations to population genetic analyses including challenges with obtaining high quality data, deciding which loci are targets of selection, and limits to identifying the genetic basis of local adaptation.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

Local adaptation results when populations of a species evolve in response to geographically variable selection. Decades of field studies and manipulative experiments have established local adaptation as extremely common [1] and central to understanding the role of adaptation in shaping species diversity. Local adaptation may also contribute to the maintenance of genetic variation, be a stepping stone to ecological speciation, and facilitate species range expansion (reviewed in [2]).

Local adaptation has been an area of active study by evolutionary ecologists since Turresson [3] first defined the concept of ecotypes and Clausen, Keck and Hiesey [4] established the use of reciprocal transplant and common garden experiments to investigate the role of habitat in driving population divergence. Even earlier, forest tree biologists were using provenance tests to identify phenotypic differences among trees from different geographic or climatic regions (reviewed in [5]). Field studies are powerful for identifying locally adapted traits, identifying the ecological forces that drive selection, and predicting short term response to selection. Organismal perspectives are also necessary for interpreting results from population genetic analyses in an ecologically meaningful context. These approaches are limited, however, in that they provide no direct insight into evolutionary processes at the molecular level and because they reflect selection over fairly short periods of time that may not be representative of historical conditions.

Population genetic approaches that explore adaptation based on sampling DNA sequences from multiple individuals offer a temporal and genetic perspective that complement organism-based approaches. Moreover, because population genetic analyses are not constrained by logistical difficulties of caring for, growing, or handling live organisms they can be used to investigate local and clinal adaptation when reciprocal transplant, common garden, or phenotypic selection analyses are not feasible due to logistical (e.g. organism size, life-span) or ethical (e.g. humans) concerns. Here we focus on recent empirical population genetic studies that have furthered our understanding of local adaptation. We first discuss some basic questions of local adaptation and then review important challenges and limitations of population genetic approaches to studying local adaptation. For discussion of other topics related to the ecological genetics of local adaptation including theory, field experiments, and statistical tests we refer readers to an excellent review by Savolainen et al. [2].

What traits are locally adapted?

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

Until recently, most population genetic analyses of local adaptation have focused on candidate genes chosen because of their putative phenotypic effects (e.g. coat color in mice [6, 7], flowering time in annual plants [8], immunity in plants and animals, [9, 10], high altitude adaptation [11]). These studies have targeted traits already thought to have been targets of local adaptation. Therefore the value of these studies has been in both providing confirmatory evidence of adaptation, elucidating the molecular mechanisms of adaptation, and identifying which of the many genes that can affect a phenotype in the laboratory are responsible for local adaptation.

As genomic data have become more available, genome scans of local adaptation have become more commonplace than candidate gene studies. One promise of genomic scans is their potential to discover genes that have been subject to local adaptation without identifying loci of interest *a priori*. Once genes subject to selection have been identified, the phenotypes upon which selection has acted can potentially be inferred on the basis of gene function (a bottom-up [12] or reverse-ecology [13] approach). For example, a transcriptome scan for local adaptation in *Neurospora crassa* identified not only several genes affecting temperature-dependent growth but also a gene involved in circadian oscillation, suggesting a role for circadian cycles in latitudinal-related adaptation in this species [14]. Genome scans also may lead to a refinement of the phenotype responsible for local adaptation. While water-use efficiency is of clear importance in plant adaptation, it is a complex trait consisting of a multitude of physiological processes. Recent genome-wide scans in wide scans *Medicago truncatula* [15] and *A. thaliana* [16] however, found evidence of selection along precipitation gradients for genes affecting stomatal closure or photosynthetic capacity relating to the proportion of time stomata are open.

The promise of reverse-ecology approaches to identify selectively important traits is limited by knowledge of gene function. Such information is generally restricted to coding regions and dependent upon annotation information derived from mutational screens of model species in laboratory environments. Genes exhibiting genotype by environment effects or having minor effects on phenotypes are likely to be missed in these screens, and phenotypes that have not been subject of functional genetic analyses will also be missed. These factors may limit the utility of reverse-ecology approaches by focusing the results of genome scans on well-studied phenotypes. Limited information on gene function may also lead to over-interpretation, as it is often easy for biologists to find biologically interesting genes that can be interpreted in the light of known selective pressures [17]. Incomplete

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

knowledge of gene function thus serves to unjustifiably reinforce preconceived ideas of the traits and selective forces driving local adaptation [18].

The vast majority of both candidate gene and genomic scan studies have relied on analyses that treat each locus independently. However, most ecologically important traits are quantitative with variation potentially affected by many, perhaps even hundreds or thousands of loci [19]. The molecular evidence of selection acting on quantitative traits is expected to be weak because the signal of selection is distributed across many loci [20-22]. Therefore, the signal of selection acting on quantitative traits may not be revealed via standard genome scans. Approaches that investigate the signal of selection aggregated across loci, however, show promise in identifying selection on quantitative traits using genomic data. Berg and Coop [23], for example, have developed methods using candidate loci from genome-wide association mapping results to identify selection on quantitative phenotypes such as height and skin color.

What environmental variables are most important in structuring population differences?

Just as population genetic approaches promise to help identify locally adapted traits, they also may be used to identifying the ecological variables most important in driving adaptation. One way to achieve this goal is shown by Fumagalli et al. [24] who linked annotated gene functions to the strength of gene-environment correlations to identify pathogens as a major driver of local adaptation in humans. Taking advantage of gene-level sequence data and detailed functional information, Fumagalli et al. identified ~ 100 genes with unexpectedly strong correlations to pathogen environment but none that were strongly correlated with climate or diet. A potential limitation of this approach is that it requires both gene level data and information on the functional role of individual genes, which is limited for most species.

Population genetic approaches are also powerful for identifying the relative importance of geographic distance and different environmental variables in structuring populations, i.e. for asking whether spatial patterns of genetic diversity are structured more by geographic distance (Isolation by distance or IBD) or the environment (isolation by environment, or IBE). A recent meta-analysis of population-genetic studies [25] revealed that both IBD and IBE are important in structuring population genetic diversity, but that overall IBE may be more important. However, studies of plants species revealed more statistically significant evidence for IBD than IBE. Recently developed statistical frameworks including a Bayesian model by Bradburd et al. [26] and structural equation modeling by Wang et al. [27],

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

provide formal means to move beyond simply asking whether IBE is statistically significant and to ask more interesting questions such as the relative contributions of IBD and IBE and comparing IBE among different environmental factors. Applications of these approaches should allow for not only the identification of the major environmental factors structuring population differentiation, but also relating the relative strengths of IBD and IBE to species characteristics and across different spatial scales [28].

Is local adaptation convergent at the molecular level?

As the number of population genetic studies identifying genes responsible for adaptation has grown, there is increasing interest in determining whether independent bouts of adaptation are achieved through the same molecular targets. Finding evidence for parallel or convergent adaptation provides strong evidence for a gene's adaptive importance, informs our understanding of constraints to adaptive evolution, and may improve our ability to predict adaptive responses to selection [29, 30]. Population genetic approaches have revealed several examples of convergent local adaptation at the molecular level: within the threespine stickleback (*Gaterosteus aculeatus*) the EBT locus has been involved in repeated adaptation to freshwater habitats [31], two species of European spruce harbor adaptive latitudinal clines at Ft2 and GI genes [32], and EPAS1 and HBB genes show patterns of high altitude adaptation in both dogs [33] and Tibetan humans [34].

Although there are examples of convergent adaptation at the gene level, there are far more examples of adaptation occurring via different genes. Even the systems cited in the paragraph above provide many examples in which loci are putatively adaptive in only one population or species: in sticklebacks Deagle et al. [35] report 9 instances of repeated adaptation of stream-lake differences but also 64 loci targeted in only a single watershed; Chen et al. [32, 36] identified a total of 18 genes harboring SNPs with signatures of latitudinal adaptation in the two species but only 2 genes were detected in both species, and Wang et al. [33] identified 14 genes with strong signatures of high altitude adaptation in dogs but only 2 targeted in both dogs and humans. Moreover, the EPAS1 locus does not appear to have been the target of selection in Andean populations [37]. The geographic distribution of GWAS candidates underlying local adaptation in *A. thaliana* is also indicative that local adaptation largely occurs through different genes in different parts of the species range [38].

In the end, it is difficult to know whether convergent evolution happens more or less than expected by chance, both because we seldom know the number of possible paths by

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

which these adaptations could evolve and because reports of parallel adaptation are given greater attention than cases in which adaptation has involved separate genes. The need for null expectations of convergent adaptation suggests the need for further theoretical work (e.g. [39]) on the topic.

Do local adaptation loci exhibit tradeoffs or conditional neutrality?

Local adaptation may result from alleles with environment-dependent fitness tradeoffs (antagonistic pleiotropy) or alleles that confer a selective advantage in one environment but are neutral in others (conditional neutrality, [40]). Of these two possibilities, only antagonistic pleiotropy is expected to lead to strong balancing selection and the maintenance of genetic variation. While antagonistic pleiotropy is found in some studies [41], field experiments indicate that conditional neutrality may be far more common than antagonistic pleiotropy [42]. Population genetic characterization of candidate loci identified through genome-wide association mapping also suggest that conditional neutrality may be widespread Fournier-Level et al. [43]. Interestingly, the analyses of Fournier-Level et al. also indicate that for many loci the environment in which alleles are selected upon is not one in which they are favored, but rather one in which they are deleterious.

Understanding the extent to which locally adapted loci exhibit conditional neutrality or antagonistic pleiotropy not only advances our understanding of genetic mechanisms, but also is important for evaluating the power of statistical tests used to identify genes responsible for local adaptation. Despite empirical support for conditional neutrality, simulation studies evaluating the power of statistical approaches for identifying targets of local adaptation [44-46] have assumed adaptive loci exhibit antagonistic pleiotropy. These tests may, however, have much less power to identify loci that are conditionally neutral than antagonistically pleiotropic (Figure 1).

Challenges with Population Genetic Analyses

Obtaining high-quality data

The quality of the inferences made from population genetic analyses depend on the quality of the data. In the era of PCR amplification and Sanger sequencing obtaining high quality data was not always easy, but ensuring data quality with high-throughput sequencing

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

and genome-scale data acquisition is even more difficult. While genomes can be relatively easily sequenced, going from a collection of short sequence reads to accurately representing genomic variation segregating within and among species is a challenge that has yet to be completely resolved.

The standard practice for making sense of sequence reads is to align them to a single reference genome or transcriptome. Because a single reference genome will not capture all the variation segregating within a species, alignment of reads to a reference genome will result in poor coverage and thus missing sequence information in highly diverged genomic regions including many intergenic regions, presence-absence variants (PAV), or copy number variants (CNV) [47]. Intergenic variation, PAV (including transposable elements), and CNV all may play important roles in phenotypic variation and adaptation [e.g. [48-50]]. Developing pan-genomes from de-novo assemblies as references for aligning sequences will alleviate some limitations of aligning reads to a single reference [e.g. [51]], but the data and bioinformatic demands are such that wide-spread de novo assemblies will be limited to relatively few systems at least for the near future.

The challenges with genome scale sequence data do not end with mapping reads. Once reads are mapped it is necessary to identify the segregating variants (i.e. SNPs, short indels) that form the basis for statistical analyses of local adaptation (reviewed in [2]). The standard approach for calling variants has been to align reads from each individual to a reference genome and then used computational tools (e.g. GATK or SAMTools) to identify the allelic state in each individual. However, due to sequencing errors and misalignments, this call-based approach can lead to considerable error, especially in calling rare variants and when sequence coverage is <10X (i.e. an average of 10 reads aligned to each base of the reference) [52]. Coverage of 10X is is considerably higher than what has been used in many population genomic studies. Much lower error rates are associated with direct estimation approaches, such as that implemented in ANGSD [53], which make sample-wide inferences of variant frequencies without calling genotypes for each sampled individual. An additional advantage of these approaches is that they can be easily applied to the sequencing of pools of individuals which can greatly reduce the costs of data collection.

The expense and complexity of handling full-genome sequence data has led to considerable interest in the use of reduced-representation data (GBS, RAD-tag, gene capture, SNP chips, RNA-seq), especially in non-model systems for which reference genomes are not available. Although reduced representation approaches are attractive for characterizing population structure, and useful for IBD vs IBE comparisons (e.g. [26]), see above) it is

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

important to remember that they sample only a small portion of the genome and thus cannot be expected to identify a meaningful portion of the genes underlying local adaptation or phenotypic variation (Box 1). While reduced representation data may be powerful for identifying structural variants, such as large inversion polymorphisms associated with local adaptation [54, 55], they are poorly suited for evaluating the relative importance of chromosomal rearrangements and single nucleotide variants.

Challenge of identifying adaptive loci

Genome scans, as well as candidate-gene studies, can also be vexed by difficulty in identifying which genes have evolved in response to geographically variable selection. One approach for identifying selected loci is to look for those with extreme values of a trait relative to expectations under a demographic model. However, the robustness of these inferences can be highly dependent on the assumed demographic model (Box 1, [46, 56-58]).

An alternative to trying to model demographic history is to compare statistics calculated at loci of interest (candidate loci) to statistics found at non-candidate or reference loci. Under the assumption that reference loci provide an estimate of the expected distribution of the statistic in the absence of adaptation, then the distribution of these values provides a metric to estimate the probability of obtaining a statistical value at candidate genes (e.g. [23, 59]). A potential drawback of this approach is that it requires *a priori* separation of reference and genes of interest. Care is also needed in the selection of reference loci – references should be draw from a portion of the genome with similar levels of recombination, background selection, and hitchhiking as candidate loci.

A third approach to identify genes of interest is to rank loci based on P-values or posterior probabilities; loci with the lowest (highest) values are those most likely to be targets of selection (e.g. [15, 43]). Given the sample sizes of many local adaptation studies and the fact that p-values are dependent on sample sizes, the use of ranks may capture functionally important loci that would not be captured if formal probability-based rejection of a null model is required. The exclusion of these loci has the potential to lead to a problem of "missing local adaptation" analogous to the association genetics problem of "missing heritability" (although of course nothing is actually missing, it has just not been detected with the statistical tools applied). Ranking loci to identify candidates is clearly a heuristic approach and may lead to many false positives, especially when data is incomplete (Box 2) or populations have not diverged in response to differential local selection). Nevertheless, if

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

genome scans are viewed as exploratory analyses for identifying candidates for subsequent analyses then simple ranking has merits.

Regardless of the statistical approach used to identify candidates, independent validation of functional importance is an important step in elevating the status of genes from candidates to causative. For candidate gene studies, functional information is generally central to the selection of candidates. For genome scans, however, this is not the case. Support for a candidate's importance can be obtained from follow-up functional analyses (e.g. [60, 61]) or bi-parental QTL mapping (e.g. [62]). Identifying the same gene in independent occurrences of local adaptation is also strong evidence of a functionally important role [29], but because of the potential for independent paths of adaptation, the repeatability of a candidate should not be viewed as a necessary step to establishing importance. An alternative to focusing on individual loci is to test whether candidate genes, as a group, predict performance in an independent sample. This approach was used by Fournier-Level et al. [43] who tested for correlations between the frequency of putatively adaptive alleles and the geographic distance from the experimental location at which those alleles were identified. Yoder et al. [43] inverted this approach, they first identified candidates genes underlying clinal adaptation to high precipitation environments, and then found a positive correlation between growth rate and the predicted genotypic effect using a set of accessions not used in the initial analyses.

Limits in identifying the genetic basis of local adaptation

Many of the important questions in local adaptation being pursued with population genetics approaches begin —rather than end — with identifying loci responsible for variation. It is therefore important to realize that a full accounting of local adaptation at the molecular level goes beyond having high quality data to analyze and statistical methods to identify causative genes. The crux of the challenge is that most ecologically important traits responsible for local adaptation are quantitative and identifying all of the genes responsible for variation in quantitative traits is likely not possible. Association genetic analyses provide some insight into this problem. The cumulative explanatory power of individual loci identified in human genotype-phenotype association studies, which often involve tens-of-thousands of individuals, is generally only a small percentage of the phenotypic variation [63].

Although some association analyses of traits implicated in local adaptation have reported explaining a high proportion of phenotypic variance, these studies have generally used the same data to identify causative loci and predict the proportion of variance explained

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

by those loci. The potential inflationary bias of using the same data for both gene identification and phenotypic prediction is illustrated by Stanton-Geddes et al. [64], who show that when the same data are used to identify causative variants with association approaches and estimate the amount of phenotypic variation explained by those SNPs, on average more than 60% of the phenotypic variance could be explained even when phenotypes were randomly assigned to genotypes and no causation was present.

Conclusions

Population genetic analyses of local adaptation have come a long way since Lewontin and Krakauer [65] first used F_{ST} to investigate local adaptation at the molecular level. With high-throughput sequencing now commonplace and greater interest in using population genetics to understand geographically variable selection, empirical population genetic analyses now promise to greatly advance our understanding of local adaptation. To do this it will be important for researchers to move beyond simply searching for genes — after all we already know genetic variation contributes to variation in many ecologically relevant phenotypes. Rather, the challenge is to use population genetic data to advance our understanding of local adaptation as a process.

In this review we have focused on some of the local adaptation questions that population genetic analyses seem both appropriate for and have attempted to address. Certainly population genetic approaches are well suited for informing other aspects of the patterns and process of local adaptation, including the genetic architecture of local adaptation and the strength and stability of selection. Improved understanding of local adaptation as a process is also likely to require advances in theory that provide empiricists with testable predictions, and analytical tools that incorporate quantitative traits, non-equilibrium conditions, and simultaneous estimation of gene-flow and selection. Finally, a stronger link between population genetic analyses and organism-based studies is likely a fruitful direction. Not only can organismal approaches be used to validate the putative importance of genes identified through genome scans, but organism-based follow-up analyses of pop-gen studies serves to inform questions related to evolutionary process.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

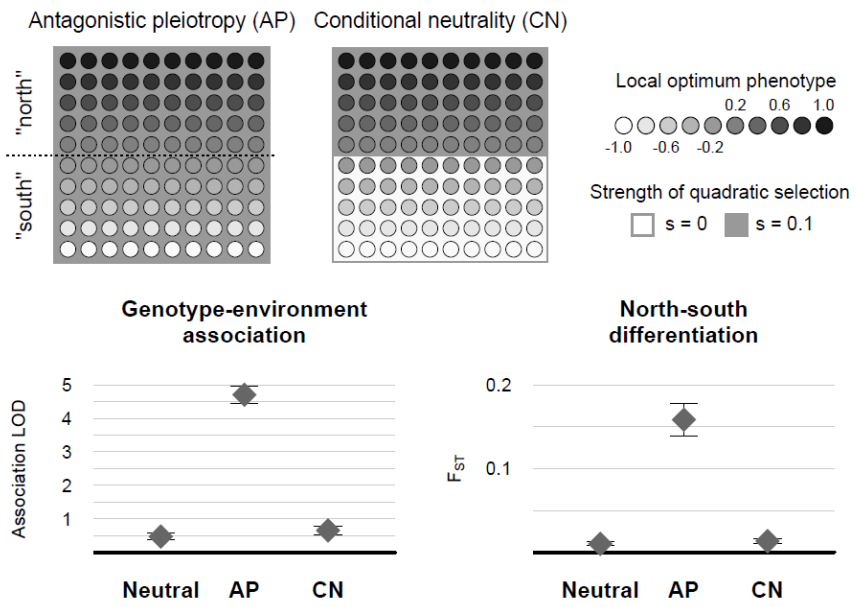

Figure 1. The power of statistical tests to identify targets of local adaptation depends on gene action with both linear-models (genotype-environment associations) and Fst tests having greater power to identify alleles showing antagonistic pleiotropy (AP) than conditional neutrality (CN). Results shown in the bottom two panels are the mean and 95% CI of the LOD score or Fst values for neutral, AP, and CN loci from 100 simulations. Fst was calculated between north and south populations. The upper part of the figure shows the optimal phenotype at each of 100 populations and the strength of quadratic selection acting around that optimum. For AP one allele at a biallelic locus is favored in the north and the other allele is favored in the south. For CN one allele is favored in the north but is neutral in the south. Results are from sampling 100 individuals sampled 2,000 generations after the start of selection. Figure and simulations by J.B. Yoder.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

Box 1: How many sweeps in Sweden?

Two recent papers searching for evidence of local adaptation in Swedish populations of *A. thaliana* highlight how our ability to identify local adaptation depends on our understanding of demographic history. In the first of these papers, Long *et al.* [69] used full genome sequence data from 180 individuals to search for evidence of recent selective sweeps in northern (50 individuals) and southern (130 individuals) Sweden. Based on previous demographic analyses of European populations of *A. thaliana*, Long et al. treated the northern and southern populations separately and applied SweepFinder, an algorithm designed to detect recent sweeps based on the site frequency spectrum [70] to each population. SweepFinder identified 22 strong signals of adaptation in the northern sample, but only a single signal in the southern sample. This difference is not expected due to in statistical power since more than twice as many individuals were sampled from the south than from the north.

Huber *et al.* [71] used data from Long et al. and extensive simulations implemented in *∂a∂i*, which infers demography based on a diffusion approximation to the site frequency spectrum [72] to ask whether the observed asymmetry in the number of sweeps may result from complex demographic history. They find the data are most consistent with a secondary contact model in which northern and southern populations separated more than 100 kya and then within the past 40 ky started exchanging migrants again. Once migration started, gene flow was asymmetric with more gene flow from the north to the south. Simulations showed that the critical value of the CLR test statistic used in SweepFinder is strongly skewed under a model of secondary contact. Using the simulations to determine critical values, Huber *et al.* identified only 6 loci in the north and 10 loci in the south that have experienced local sweeps – one third less and 10 times more than what was found using a naïve demographic model. Huber *et al.* also estimated that the strength of selection was much greater (>4X) in the north than the south. Potential explanations include the northern population being farther from its selective optimum due to its relatively recent arrival, or the existence of more heterogeneous environment in the south. Two lines of evidence – a higher frequency of long haplotypes and a higher ratio of non-synonymous to synonymous site diversity, both of which can be indicative of partial selective sweeps – indicate that the southern population was drawn from a more heterogeneous environment. This highlights that care must be taken assigning individuals to populations when conducting tests of local adaptation; geography-based assignment may not

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

validly capture the selective environments, and ignorance of demography can greatly skew estimates of the loci under selection and the strength of that selection.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

Box 2: Incomplete data produce an incomplete picture

There are a variety of approaches that can be used to sample a large number of genetic markers distributed across a genome without requiring whole-genome sequencing (e.g. Genotype by sequencing [66], RADseq [67], multiplexed genotype sequencing [68], sequence capture). These technologies are powerful for collecting data for mapping genes in genetic crosses and characterizing population structure. However, because they sample only a small portion of the genome they have very limited power to identify loci responsible for phenotypic variation in association analyses or in identifying targets of local adaptation. To illustrate this limited power we calculated the expected probability of detecting recent hard sweeps (in which selection drives a new mutation to fixation) as a function of SNP density, recombination rates, and the strength of selection (Figure 1). The results show that even for hard sweeps, the easiest form of selection to detect, the great majority of targets of selection will be missed: only with fairly high SNP density (1 SNP per 5 kb), a large number of sweeps (100), and low to normal levels of recombination is there a high probability of detecting even a single sweep.

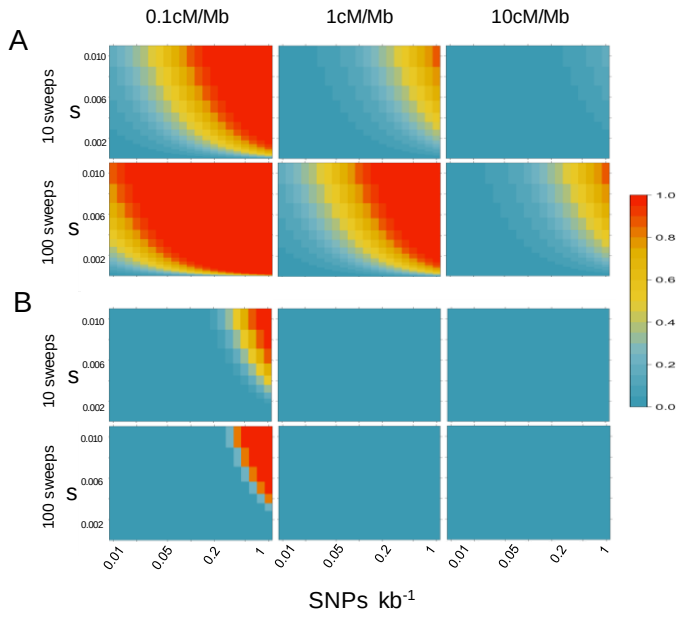

Figure 1. Probability of detecting (A) 1 or more or (B) > 50% of hard sweeps as a function of the number of sweeps in the genome , the strength of selection and number of SNPs kb^{-1} for three rates of recombination.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

Glossary Box

Local adaptation: adaptation in response to selection that varies among geographically distinct populations.

Clinal adaptation: a form of local adaptation in which the mean phenotype of populations change gradually across an environmental or geographic gradient.

Isolation by distance: A negative relationship between the genetic similarity of populations and geographic distance.

Isolation by environment: A positive relationship between the genetic similarity of populations and the similarity of the environments in which populations are found. Can be caused by selection or spatial autocorrelation.

Reference genome: An assembly of a genome that is representative of a species and used to align sequencing reads for population-genomic studies. Depending on the species the reference can be based on a single individual or a collection of individuals. Reference genomes do not capture the full extent of nucleotide or structural variation segregating within a species.

Acknowledgements

We thank Amanda Gorton, Jeremy Yoder, and members of the Ross-Ibarra lab for comments and discussion and NSF awards (IOS-1238014 and 1237993) for financial support.

References

1 Hereford, J. (2009) A Quantitative Survey of Local Adaptation and Fitness Trade-Offs. *Am Nat* 173, 579-588

2 Savolainen, O. *et al*. (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics* 14, 807-820

3 Turresson, G. (1922) The genotypical response of the plant species to the habitat. . *Hereditas* 3, 211-350

4 Clausen, J.:.K., D.D. and Hiesey, W. (1940) Experimental studies on the nature of species. I. Effects of varied environments on western North American plants. .

5 Langlet, O. (1971) Two hundred years of genecology. *Taxon* 20, 653-722

6 Nachman, M. *et al*. (2003) The genetic basis of adaptive melanism in pocket mice. *Proc Natl Acad Sci U S A* 100, 5268-5273

7 Linnen, C.R. *et al*. (2009) On the Origin and Spread of an Adaptive Allele in Deer Mice. *Science* 325, 1095-1098

8 Le Corre, V. *et al*. (2002) DNA polymorphism at the FRIGIDA gene in Arabidopsis thaliana: Extensive nonsynonymous variation is consistent with local selection for flowering time. *Mol Biol Evol* 19, 1261-1271

9 Moeller, D.A. and Tiffin, P. (2008) Geographic Variation in Adaptation at the Molecular Level: a Case Study of Plant Immunity Genes. *Evolution* 62, 3069-3081

10 Prugnolle, F. *et al*. (2005) Pathogen-driven selection and worldwide HLA class I diversity. *Current Biology* 15, 1022-1027

11 Storz, J.F. *et al*. (2007) The molecular basis of high-altitude adaptation in deer mice. *Plos Genetics* 3, 448-459

12 Ross-Ibarra, J. *et al*. (2007) Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc Natl Acad Sci U S A* 104, 8641-8648

13 Li, Y.F. *et al*. (2008) "Reverse Ecology" and the Power of Population Genomics. *Evolution* 62, 2984-2994

14 Ellison, C.E. *et al*. (2011) Population genomics and local adaptation in wild isolates of a model microbial eukaryote. *Proc Natl Acad Sci U S A* 108, 2831-2836

15 Yoder, J.B. *et al*. (2014) Genomic Signature of Adaptation to Climate in Medicago truncatula. *Genetics* 196, 1263-+

16 Hancock, A.M. *et al*. (2011) Adaptation to Climate Across the Arabidopsis thaliana Genome. *Science* 334, 83-86

17 Pavlidis, P. *et al*. (2012) A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans. *Mol Biol Evol* 29, 3237-3248

18 Barrett, R.D.H. and Hoekstra, H.E. (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics* 12, 767-780

19 Rockman, M.V. (2012) The Qtn Program and the Alleles that Matter for Evolution: all That's Gold does Not Glitter. *Evolution* 66, 1-17

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

20 Kelly, J. (2006) Geographical variation in selection, from phenotypes to molecules. *Am Nat* 167, 481-495

21 Kemper, K.E. *et al*. (2014) Selection for complex traits leaves little or no classic signatures of selection. *BMC Genomics* 15, 246

22 Le Corre, V. and Kremer, A. (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Mol Ecol* 21, 1548-1566

23 Berg, J. and Coop, G. (2014) A population genetic signal of polygenic adaptation. *PLoS Genetics* 10, e1004412

24 Fumagalli, M. *et al*. (2011) Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. *Plos Genetics* 7, e1002355

25 Sexton, J.P. *et al*. (2014) Genetic Isolation by Environment Or Distance: which Pattern of Gene Flow is most Common? *Evolution* 68, 1-15

26 Bradburd, G.S. *et al*. (2013) Disentangling the Effects of Geographic and Ecological Isolation on Genetic Differentiation. *Evolution* 67, 3258-3273

27 Wang, I.J. *et al*. (2013) Quantifying the roles of ecology and geography in spatial genetic divergence. *Ecol Lett* 16, 175-182

28 Richardson, J.L. *et al*. (2014) Microgeographic adaptation and the spatial scale of evolution. *Trends in Ecology & Evolution* 29, 165-176

29 Conte, G.L. *et al*. (2012) The probability of genetic parallelism and convergence in natural populations. *Proceedings of the Royal Society B-Biological Sciences* 279, 5039-5047

30 Streisfeld, M.A. and Rausher, M.D. (2011) Population Genetics, Pleiotropy, and the Preferential Fixation of Mutations during Adaptive Evolution. *Evolution* 65, 629-642

31 Jones, F.C. *et al*. (2012) A Genome-wide SNP Genotyping Array Reveals Patterns of Global and Repeated Species-Pair Divergence in Sticklebacks. *Current Biology* 22, 83-90

32 Chen, J. *et al*. (2014) Clinal Variation at Phenology-Related Genes in Spruce: Parallel Evolution in FTL2 and Gigantea? *Genetics* 197, 1025-+

33 Wang, G. *et al*. (2014) Genetic convergence in the adaptation of dogs and humans to the high altitude environment of the Tibetan plateau. *Genome Biology and Evolution* in press,

34 Yi, X. *et al*. (2010) Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science* 329, 75-78

35 Deagle, B.E. *et al*. (2012) Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. *Proceedings of the Royal Society B-Biological Sciences* 279, 1277-1286

36 Chen, J. *et al*. (2012) Disentangling the Roles of History and Local Selection in Shaping Clinal Variation of Allele Frequencies and Gene Expression in Norway Spruce (Picea abies). *Genetics* 191, 865-U377

37 Bigham, A. *et al*. (2010) Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data. *Plos Genetics* 6, e1001116

38 Fournier-Level, A. *et al*. (2013) Paths to selection on life history loci in different natural environments across the native range of Arabidopsis thaliana. *Mol Ecol* 22, 3552-3566

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

39 Ralph, P.L. and Coop, G. (2014) Convergent Evolution During Local Adaptaiton to Patchy Landscapes. *Biorxiv*

40 Schnee, F. and Thompson, J. (1984) Conditional Neutrality of Polygene Effects. *Evolution* 38, 42-46

41 Agren, J. *et al*. (2013) Genetic mapping of adaptation reveals fitness tradeoffs in Arabidopsis thaliana. *Proc Natl Acad Sci U S A* 110, 21077-21082

42 Anderson, J.T. *et al*. (2011) Evolutionary genetics of plant adaptation. *Trends in Genetics* 27, 258-266

43 Fournier-Level, A. *et al*. (2011) A Map of Local Adaptation in Arabidopsis thaliana. *Science* 334, 86-89

44 De Mita, S. *et al*. (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Mol Ecol* 22, 1383-1399

45 Jones, M.R. *et al*. (2013) Integrating Landscape Genomics and Spatially Explicit Approaches to Detect Loci Under Selection in Clinal Populations. *Evolution* 67, 3455-3468

46 Lotterhos, K.E. and Whitlock, M.C. (2014) Evaluation of demographic history and neutral parameterization on the performance of F-ST outlier tests. *Mol Ecol* 23, 2178-2192

47 Sims, D. *et al*. (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* 15, 121-132

48 Demuth, J.P. *et al*. (2006) The Evolution of Mammalian Gene Families. *Plos One* 1, e85

49 Hanikenne, M. *et al*. (2013) Hard Selective Sweep and Ectopic Gene Conversion in a Gene Cluster Affording Environmental Adaptation. *Plos Genetics* 9, e1003707

50 Fischer, I. *et al*. (2011) Adaptation to drought in two wild tomato species: the evolution of the Asr gene family. *New Phytol* 190, 1032-1044

51 Gan, X. *et al*. (2011) Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* 477, 419-423

52 Han, E. *et al*. (2014) Characterizing Bias in Population Genetic Inferences from Low-Coverage Sequencing Data. *Mol Biol Evol* 31, 723-735

53 Nielsen, R. *et al*. (2012) SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *Plos One* 7, e37558

54 Pyhajarvi, T. *et al*. (2013) Complex Patterns of Local Adaptation in Teosinte. *Genome Biology and Evolution* 5, 1594-1609

55 Hoffmann, A. *et al*. (2004) Chromosomal inversion polymorphisms and adaptation. *Trends in Ecology & Evolution* 19, 482-488

56 Excoffier, L. *et al*. (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* 103, 285-298

57 Excoffier, L. *et al*. (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* 103, 285-298

58 de Villemereuil, P. *et al*. (2014) Genome scan methods against more complex models: when and how much should we trust them? *Mol Ecol* 23, 2006-2019

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014

59 Wright, S. and Charlesworth, B. (2004) The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model. *Genetics* 168, 1071-1076

60 Carneiro, M. *et al*. (2014) Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science* 345, 1074-1079

61 Prasad, K.V.S.K. *et al*. (2012) A Gain-of-Function Polymorphism Controlling Complex Traits and Fitness in Nature. *Science* 337, 1081-1084

62 Huang, X. *et al*. (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490, 497-+

63 Visscher, P.M. *et al*. (2012) Five Years of GWAS Discovery. *Am J Hum Genet* 90, 7-24

64 Stanton-Geddes, J. *et al*. (2013) Candidate Genes and Genetic Architecture of Symbiotic and Agronomic Traits Revealed by Whole-Genome, Sequence-Based Association Genetics in Medicago truncatula. *Plos One* 8,

65 Lewontin, R. and Krakauer, J. (1973) Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms. *Genetics* 74, 175-195

66 Elshire, R.J. *et al*. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *Plos One* 6, e19379

67 Baird, N.A. *et al*. (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *Plos One* 3, e3376

68 Andolfatto, P. *et al*. (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res* 21, 610-617

69 Long, Q. *et al*. (2013) Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat Genet* 45, 884-U218

70 Nielsen, R. *et al*. (2005) Genomic scans for selective sweeps using SNP data. *Genome Res* 15, 1566-1575

71 Huber, C.D. *et al*. (2014) Keeping it Local: Evidence for Positive Selection in Swedish *Arabidopsis thaliana*. *Molecular Biology and Evolution*

72 Gutenkunst, R.N. *et al*. (2009) Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *Plos Genetics* 5, e1000695

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.488v1 | CC-BY 4.0 Open Access | rec: 5 Sep 2014, publ: 5 Sep 2014