Received Date : 06-Jun-2016
Revised Date : 23-Aug-2016
Accepted Date : 02-Sep-2016
Article type : Opinion

BREAKING RAD: AN EVALUATION OF THE UTILITY OF RESTRICTION SITE ASSOCIATED DNA SEQUENCING FOR GENOME SCANS OF ADAPTATION

David B. Lowry[a,b*], Sean Hoban[c,d], Joanna L. Kelley[e], Katie E. Lotterhos[f], Laura K. Reed[g], Michael F. Antolin[h], Andrew Storfer[e]

[a]Department of Plant Biology, Michigan State University, East Lansing MI 48824, USA

[b]Program in Ecology, Evolutionary Biology, and Behavior, Michigan State University, East Lansing, MI 48824, USA

[c]The Morton Arboretum, Lisle, IL, USA

[d]National Institute for Mathematical and Biological Synthesis (NIMBioS), Knoxville, TN, USA

[e]School of Biological Sciences, Washington State University, Pullman, WA 99164, USA

[f]Department of Marine and Environmental Sciences, Northeastern University Marine Science Center, 430 Nahant Rd., Nahant, MA 01908, USA

[g]Department of Biological Sciences, University of Alabama, Tuscaloosa, AL, 35406, USA

[h]Department of Biology, Colorado State University, Fort Collins, CO 80523-1878, USA

Key words: F_{ST}, Genotyping by sequencing, genome-environment association, genome scan, local adaptation, outlier analysis

[*]To whom correspondence should be addressed: Department of Plant Biology, Michigan State University, Plant Biology Laboratories, 612 Wilson Road, Room 166, East Lansing MI 48824, USA

Tel: 517-432-4882

E-mail: dlowry@msu.edu

Running title: Breaking RAD

ABSTRACT

Understanding how and why populations evolve is of fundamental importance to molecular ecology. RADseq (Restriction site-Associated DNA sequencing), a popular reduced representation method, has ushered in a new era of genome-scale research for assessing population structure, hybridization, demographic history, phylogeography, and migration. RADseq has also been widely used to conduct genome scans to detect loci involved in adaptive divergence among natural populations. Here, we examine the capacity of those RADseq-based genome scan studies to detect loci involved in local adaptation. To understand what proportion of the genome is missed by RADseq studies, we developed a simple model using different numbers of RAD-tags, genome sizes, and extents of linkage disequilibrium (length of haplotype blocks). We then surveyed recent studies that have used RADseq for genome scans and found that that the median density of RADseq markers across these studies was one marker per 3.96 megabases. Given that the length of linkage disequilibrium is often orders of magnitude less than a megabase, we conclude that genome scans based on RADseq data alone are unlikely to advance our understanding of molecular ecology or evolutionary genetics for most systems.

"The moral of the story is: I chose a half measure, when I should have gone all the way. I'll never make that mistake again. No more half measures, Walter."

-Mike Ehrmantraut, *Breaking Bad*

Understanding how and why populations evolve is of fundamental importance to molecular ecology. Restriction site-Associated DNA sequencing (RAD-seq) has dramatically decreased the cost of generating large numbers of polymorphic markers and has ushered in a new era of genome-scale research on the evolution of populations across a broad range of organisms. Several genotyping

methods (RAD, ddRAD, GBS, MSG, etc.) developed in recent years, for convenience, can collectively be referred to as RADseq. The details of each of these methods have been reviewed recently by Andrews et al. (2016). Generally, RADseq methods produce DNA libraries for high-throughput sequencing by using restriction enzymes that cut at specific motifs throughout the genome. RADseq markers come in the form of RAD-tags, which are short-read sequences adjacent to restriction enzyme cut sites. Because many polymorphic markers are produced by RADseq, it has frequently been used for population genetic analyses, including assessment of population structure, hybridization, demographic history, phylogeography, and migration (Catchen et al. 2013; Cavender-Bares et al 2015; Combasch and Vollmer 2015; Qi et al. 2015). Markers generated by RADseq have also been useful for constructing linkage maps and identifying quantitative trait loci (QTL; Pfender et al. 2011; Houston et al. 2012; Weber et al 2013; Laporte et al. 2015; Lowry et al. 2015).

Many recent studies have used RADseq for genome scans to detect locally adapted loci (reviewed in Arnold et al. 2013; Tiffin & Ross-Ibarra 2014; Andrews et al. 2016). However, RADseq presents a major challenge for genome scan studies because the approach usually samples only a small proportion of the genome, despite generating hundreds to thousands of polymorphic markers (Figure 1). This problem is particularly acute for organisms with large genome sizes and/or low levels of linkage disequilibrium (LD). While others have articulated the challenges that RADseq poses for genome scans (Arnold et al. 2013; Davey et al. 2013; Tiffin & Ross-Ibarra 2014), RADseq studies are still relatively common despite potentially major pitfalls.

Here, we analyze how much of the genome will be missed by genome scan studies that utilize RADseq as the source of genetic markers for analyses. We also survey the recent literature to establish how RADseq is being used for contemporary genome scan studies. Finally, we discuss

potential alternative approaches for genome scans, including exome capture, RNA sequencing, pooled sequencing, and whole genome sequencing.

RAD APPROACHES FOR IDENTIFYING CANDIDATE LOCI UNDERLYING LOCAL ADAPTATION

A major goal of molecular ecology is to identify the loci underlying local adaptation to important environmental factors in wild populations (Hereford 2009; Anderson et al. 2011; Barrett & Hoekstra 2011; Jones et al. 2012; Savolainen et al. 2013; Rausher & Delph 2015; Rellstab et al. 2015; Hoban et al. 2016). The two predominant genome scan approaches for identifying candidate loci underlying local adaptation are: 1) differentiation outlier studies that scan the genome for differences in allele frequencies between locally adapted populations, and 2) genetic-environment association studies that evaluate associations between environmental variables and allele frequencies across a set of populations spread over geographic space (reviewed in Rellstab et al. 2015; Hoban et al. 2016). Both approaches require data for many loci across the genome. RADseq has recently been used in genome scans for adaptation loci because it is a cost-effective method for acquiring large numbers of polymorphic markers.

While RADseq datasets comprise several orders of magnitude more polymorphic markers than microsatellite datasets, RADseq will still not adequately sample haplotype blocks with enough markers to provide genome-wide coverage (Figure 1). It should be noted that even though multiple single nucleotide polymorphisms (SNPs) within a single RAD-tag can be detected, those SNPs are usually redundant with other SNPs in the same haplotype block. Further, not all SNPs will be in complete linkage disequilibrium ($R^2 = 1$) within haplotype blocks because new mutations occur at different times over the course of the evolution of haplotypes and because recombination does occur within blocks. Thus, studies that lack full genome sequencing coverage will miss potentially

important adaptive SNPs because they are not in complete linkage disequilibrium with SNPs markers detected by RADseq genotyping. In the next section, we quantify the portion of the genome that will be overlooked when surveyed by RADseq.

HOW BAD IS RAD FOR GENOMIC SCANS OF ADAPTATION?

To understand potential limitations for RADseq, we developed a simple model using different numbers of RAD-tags, genome sizes, and extents of LD (length of haplotype blocks) to calculate the percentage of the genome that could be studied using RADseq. The model depends on a number of assumptions: (i) recombination hotspots are equally distributed across the genome, and thus that the genome is divided into equal-length haplotype blocks; (ii) the optimal situation in which RAD-tag markers are distributed equally among these blocks; (iii) sequencing depth is uniform across RAD-tags; (iv) no more than one RAD-tag marker occurs on each linkage block; (v) RAD-tags are in perfect linkage disequilibrium with adaptive SNPs; (vi) all RAD-tags are polymorphic; (vii) if an adaptive locus is not linked to a polymorphic RAD-tag, genome scans would not be able to detect that locus as having a potential role in adaptation. Further, our analyses were conducted assuming the methodology of the original RAD protocol (Baird et al. 2008, Miller et al. 2007). Because this protocol digests DNA with a single enzyme, shears it, and then attaches adaptors (Davey et al. 2013), it is the RAD method that will return the highest density of RAD-tags (Puritz et al. 2014). Finally, we assumed 50% GC content in the genome and thus, for an enzyme cut site of length *n* the frequency of cut sites could be estimated as 4^n (higher or lower GC content will result in a difference in cut frequency; Schweyen et al. 2014). Note that all of the assumptions described above are for a best-case scenario because they are virtually never met in reality. Violations of any of these assumptions will make the problem of genome scans based on RADseq even worse.

It should be noted that for detecting signatures of selection, genomic regions affected by strong selection are expected to have higher LD, as neutral loci hitchhike along with a selected locus as it sweeps to fixation. Our assumption of equal-sized haplotype blocks does not account for this scenario, which would intuitively increase the ability of RADseq-based genome scans to detect selective sweeps. However, Tiffin & Ross-Ibarra (2014) demonstrated that this is actually not often the case by calculating the probability of detecting recent hard-sweeps as a function of SNP density, recombination rates, and the strength of selection. Overall, they showed that only with a fairly high RAD-tag density (1 polymorphic RAD-tag 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.

Using our best-case scenario assumptions, we evaluated the relationship between different haplotype block sizes and the theoretical maximum proportion of the genome that would be sampled by RADseq (Figure 2). We conducted the analysis for evenly spaced 6 base pair (bp) (every 4^6 = 4096 base pairs) and 8-bp (every 4^8 = 65,536 base pairs) restriction enzyme cutters, which are both commonly used in RADseq studies (Figure 2). For a species with low LD (less than 1000 base pairs (bp), such as in insects, some marine species, and many trees; Table 1), a 6-bp cutter would cover less than 25% of the genome, and an 8-bp cutter would cover less than 2% of the genome even in this best-case scenario (Figure 2; R code provided in Supplemental Materials).

No RADseq study achieves the best-case assumptions. In reality, RAD-tags are distributed non-uniformly across the genome and LD actually varies throughout the genome because of recombination hotspots, proximity to centromeres, and chromosomal rearrangements (Ortiz-Barrientos et al. 2016). Therefore, the ability to detect associations with linked markers with adaptive loci will vary depending on the genomic region analyzed. Finally, any RADseq method that

uses size selection on non-sheared DNA (MSG: Andolfatto et al. 2011; ddRAD: Peterson et al. 2012; RRLs: Greminger et al. 2014) or preferentially amplifies short fragments (GBS: Elshire et al. 2011; SBG: Truong et al. 2012) will further decrease the density of RAD-tags across the genome and therefore decrease the probability of detecting selection (Davey et al. 2013).

To more realistically assess the potential of RADseq, we quantified how different numbers of informative RAD-tags will influence genome scan studies. It is important to consider the number of informative RAD-tags generated in a study, as many tags will lack sufficient polymorphism for genome scans or be lost during bioinformatic filtering. Our analysis calculated how many haplotype blocks contained at least one informative RAD-tag for different genome sizes, lengths of LD, and total numbers of informative RAD-tags (Figure 3; R code provided in Supplemental Materials). The results of this analysis should have utility for researchers who wish to assess the potential of a proposed genome scan study. For example, a species with a 3 gigabase pair (Gbp) genome (roughly the size of the human genome) and haplotype block lengths of 10 kilobase (kb) (moderate linkage), 20,000 informative RAD-tags will only cover ~7% of the genome—even if they are evenly distributed. Thus, at best, ~93% of the genome will not contain an informative RAD-tag and effectively be ignored by a genome scan study with these parameters. Note the 'fan' pattern that we observe in Figure 3 underscores the multiplicative effect of genome size and length of haplotype blocks. Overall, these results demonstrate that RADseq-based approaches will often only be able to evaluate a small portion of the genome for signals of local adaptation.

RECENT RAD GENOME SCANS

Based on the results of our theoretical analyses of the efficacy of RADseq for genome scans, we were interested in evaluating how much of the genome is sampled by RADseq for contemporary

genome scan studies. To identify recent studies using RADseq for genome scans of local adaptation, we conducted a literature survey of articles from January 2015 to April 2016 (Table S1). We identified articles in two steps: 1) We examined the contents of the journals *Molecular Ecology*, *Conservation Genetics*, *Genetics*, *Ecology and Evolution*, *Molecular Biology & Evolution*, and *Tree Genetics and Genomics*, 2) We also examined all articles citing the major RADseq methods papers (Miller et al. 2007; Baird et al. 2008; Andolfatto et al. 2011; Elshire et al. 2011; Peterson et al. 2012) that were published since January 2015. In total, we identified 27 articles (with 3 studies using multiple species) that fit the following criteria: (i) the study used a genome scan to infer the genetic basis of adaptation in natural populations, (ii) the study used at least 1,000 SNPs in the genome scan, and (iii) the study was not based on an *a priori* list of candidate genes. Of those studies, 19 were conducted with single digest RAD, and 8 with double-digest RAD. For the following calculations, if the study examined multiple closely-related species, the average values for the two species were used for the study. The distribution of the total number of individuals sampled was skewed, with most studies using fewer than 200 individuals, a few using from 200 to 700, and one study with nearly 2000 individuals (median=144). There was a strong taxonomic bias among studies: 16 fish, 2 birds, 5 invertebrates, 3 mammals, and 1 plant. The genome sizes of studied species were slightly skewed, with most between 0.5 and about 1 Gbp, and fewer species of 3 to 5 Gbp (median=1.26 Gbp). Note that we used the species genome size, if available, or used the genome size of the closely related species to which reads were aligned. To gauge the breadth of possible values we compiled the genome sizes and estimates of the length of linkage disequilibrium for a wide diversity of species (Table 1).

From our survey, we conclude that recent RADseq studies are typically too sparse to cover an appreciable amount of the genome and that they likely miss the vast majority of loci underlying adaptation. To estimate the number of markers per megabase pair (Mbp), we used the number of

RAD-tags rather than number of SNPs, except in four cases, where the number of RAD-tags was not reported. The number of markers was skewed with most studies having fewer than 20,000, two studies with ~70,000 markers, and one study with more than 160,000 makers (median=9,000). Excluding those species with no estimates of genome size, the number of markers per Mbp was also noticeably skewed (median of 1 RAD-tag marker per 3.96 Mbp), with most studies having 5 or fewer RAD-tags per Mbp, a few having between 6 and 20 per Mbp, three having up to 110 per Mbp, and only one study having more than a few hundred (362) RAD-tags per Mbp. Given that RAD-tags and recombination rates are not evenly spaced throughout the genome, and not all RAD-tags are informative, the extent of linkage disequilibrium would have to be multiple megabases in length for recent RADseq-based genome scan studies to be effective (Table S1). However, for most organisms, LD is far less than 100 kb (Table 1). Even for species with large LD blocks, such as humans (50 kb; Pritchard & Przeworski 2001), this would be one marker per 76 haplotype blocks and thus only samples ~1.3% of the genome. For some plants and ocean animals having large population sizes and high levels of gene flow, where haplotype blocks are between 250bp and 1kb, the proportion of the genome sampled will be very small. Thus, it appears that many recent RADseq studies have only scanned a small fraction of the genome.

BEYOND RAD

Although RADseq is an efficient and economical approach for generating genetic makers, our results suggest that other genotyping methods should be considered when conducting genome scan studies. We advise that researchers planning to use reduced representation approaches for genome scans to target genic regions, since genic regions are likely to be the location of much of the functional genetic changes involved in adaptation (Hoekstra & Coyne 2007; Stern & Orgogozo 2008). For example, transcriptome sequencing and exome capture can be used to target genic regions when no reference genome is available (Gugger et al 2016). Genic regions captured by these

methods are often in linkage with gene promoters, which also are likely locations of functional adaptive variation (Wray 2007; Stern & Orgogozo 2008). However, transcriptome sequencing will miss many genes because they have tissue- or condition-specific expression. Further, important genes underlying adaptation may be expressed at low levels, thus requiring more sequencing at deeper coverage. Allele-specific expression (Chen et al. 2016) can also lead to misinterpretation of allele frequency differences among populations. In contrast, exome capture sequencing of genomic DNA has a greater potential to genotype a larger set of genic regions, provided that the capture method is designed from a reference genome (Jones & Good 2016). Even so, both transcriptome sequencing and exon capture will likely fail to detect many polymorphisms in distant regulatory regions, such as in enhancers and insulators (Pennacchio et al. 2013).

Although still not feasible for large genomes, whole genome sequencing is by far the best approach for genome scans because SNPs are quantified in most, if not all, haplotype blocks across the genome. An economical alternative to whole genome sequencing of individuals is population pooled-sequencing (pool-seq; Schlötterer et al. 2014). Pool-seq studies combine dozens to hundreds of individuals prior to sequencing in order to calculate allele frequencies of populations and conduct genome scans (Schlötterer et al. 2014; Wright et al. 2015; Kapun et al. 2016). It should be noted that genotype information generated from whole-genome sequencing and from pool-seq are no panacea and will contain significant levels of missing data and data artifacts. Some missing data and artifacts are caused by structural differences (insertions, deletions, inversions, etc.) between re-sequenced individuals and the reference genome to which they are aligned (Tiffin & Ross-Ibarra 2014, Hoban et al. 2016). Finally, it is important that researchers retain invariant sites in their datasets to accurately calculate population genetic summary statistics and demographic parameters. Invariant sites are often removed as part of bioinformatics processing of sequence data, but are necessary for coalescent analyses of demographic history.

We evaluated the relative sequencing costs for these different approaches by calculating the number of individuals that could be multiplexed per 250 million Illumina paired-end reads (assuming a read length of 100 and with 20% of reads removed due to contamination, PCR duplicates, and other quality control filtering). Currently, this corresponds to a single lane of Illumina HiSeq 2500 which yields ~250 million reads (04/2016 http://www.illumina.com). For RADseq, we assumed a 6 base-pair cutter (cut site 50% GC content) is used to digest the genome (50% GC content) and that a 20x depth of read coverage is needed per RAD-tag. Given these assumptions, an expected 40 individuals can be sequenced in a lane for genome sizes greater than 500 Mbp (e.g. stickleback fish), and an expected 10 individuals can be sequenced in a lane for genome sizes greater than 2 Gbp (e.g. bottle nose dolphin, Figure 4 top). For transcriptome sequencing and exome capture, we assumed 30,000 genes with an average of 7 exons per gene, one isoform, and 150-bp per exon. For transcriptome sequencing, we also assumed that 20-30 million raw reads were needed per individual to ascertain genes expressed at low levels (The ENCODE Consortium). For sequence capture, we assumed that capture probes were designed near exon boundaries and would target the exon in addition to 100 bp of DNA on either side of the exon boundary, with a desired coverage of 20x per sequence. Based on these assumptions, the number of individuals multiplexed in a lane for RADseq is similar to exome capture for genome sizes in the range of 500 Mbp to 2 Gbp, and for transcriptome sequencing in the range of 1.5 to 2.5 Gbp (Figure 4 top). For genomes larger than 2.5 Gbp, more individuals can be multiplexed per lane for transcriptome sequencing and exome capture than for RADseq.

We also conducted calculations for whole-genome sequencing of individuals (assuming a 20x depth of read coverage per individual) and whole-genome pool-seq of populations (assuming 100x coverage per population, Schlötterer et al 2014). For genome sizes in the range of 500 Mbp to 2 Gbp, only a few individuals can be sequenced per lane at 20x coverage, which may be outside the budget

for many laboratories (Figure 4 bottom). Pool-seq appears to be a cost-effective alternative to sequencing individuals (Figure 4 bottom). For investigators wishing to explore this parameter space further, we have included the R scripts used to generate the figures in the supplementary information online.

IF YOU RAD

There are some situations where RADseq can have utility for genome scans of local adaptation. For example, RADseq is very useful for studies of local adaptation where candidate genes have already been identified because RAD markers can be used to generate a null distribution of loci across the genome (Luikart et al. 2003). Natarajan et al. (2015) recently conducted an excellent study of this sort with candidate alpha- and beta-globin genes in ducks. They compared allele frequencies of these candidate genes to a null distribution generated by RADseq data for high and low altitude population pairs of three Andean duck species. For all three species, F_{ST} of the globin genes was greatly elevated above the null F_{ST} distribution generated by RADseq markers. We anticipate that many more researchers will use this approach in the near future.

RADseq might also be useful in some cases for genome scans that do not focus on candidate genes identified *a priori*, depending on the nature of the genome of the study organism. There are a few examples of RADseq genome scans recovering loci previously known to be involved in adaptive divergence (Hohenlohe et al. 2010; Nadeau et al. 2014). Further, some methylation-sensitive restriction enzymes do preferentially target genic regions of the genome (Pegadaraju et al. 2013). However, researchers thinking of using RADseq for naïve genome scans should first establish how much of the genome they can realistically expect to evaluate and make sure to report those estimates in publications. Pilot RADseq experiments with different enzymes can be used to estimate

the number of informative RAD-tag markers that will be generated. Researchers should also try to establish the genome size of their focal organisms. Genome sizes for many organisms are listed the Animal Genome Size Database (http://www.genomesize.com/), Plant C Values Database (http://data.kew.org/cvalues/), or other resources (Garcia et al 2014). If genome size is unknown, it can readily be established by flow cytometry. Unfortunately, it is more difficult to estimate the extent of LD for a given species. Rough estimates of LD can be made by comparisons to species with known LD estimates, although some closely related species may have very different genome sizes (e.g. Plethodontid salamanders; Sessions & Larson 1987). Additionally, LD is highly variable and even varies greatly among populations within species (e.g. haplotype blocks ranged from 8,800 - 25,200 bp among human populations in one study; Hinds et al. 2005). Once the number of informative RAD-tags, genome size, and LD have been estimated, researchers can conduct similar calculations to those that we described above to determine the maximum amount of the genome they can reasonably expect to evaluate in a genome scan study. Based on our survey of the literature, it appears that genome scans based upon RADseq will be informative for a minority of species.

CONCLUSIONS AND CLOSING REMARKS

The problem with the low density of markers produced by RADseq is just one of many challenges for studies aiming to identify loci involved in local adaptation. Other reviews have outlined these problems, which include the confounding effects of population structure, demographic history, and issues with analytical methods (Tiffin & Ross-Ibarra 2014; Lotterhos & Whitlock 2014; Rellstab et al. 2015; Hoban et al. 2016). However, the problem with RADseq is with the marker datasets themselves, which are often far too sparse to have a reasonable chance of detecting loci under selection. The issue of marker sparseness will be true for any other sort of genomic scan conducted without genome resequencing, including AFLPs and microsatellites (Nosil et al. 2009; Fischer et al.

2011). Marker sparseness is also an issue for Genome-Wide Association Studies (GWAS) and caution must be taken for those studies as well when using RADseq (e.g. Stanton-Geddes et al. 2013).

Overall, RADseq should be viewed as one of many important tools to study local adaptation. Field experimentation, linkage and association mapping, forward and reverse genetics, candidate gene approaches, functional molecular genetics, and modeling are also crucial approaches for understanding the genetic basis of local adaptation (Kawecki & Ebert 2004; Hereford 2009; Anderson et al. 2011; Savolainen et al. 2013; Pardo-Diaz et al. 2015). RADseq can be very useful for estimating important parameters that may affect interpretation of how adaptation occurs, including population structure, gene flow, and demography. RADseq can also be very useful for genomic scans of local adaptation when candidate genes have been identified *a priori* (e.g. Natarajan et al. 2015). However, genome scans based on RADseq data alone are very unlikely to advance our general understanding of molecular ecology or evolutionary genetics for most systems.

ACKNOWLEDGMENTS

Comments from Jon Puritz, Kevin Wright, Joel Sharbrough, Dan Sloan, members of the Lowry Lab group, and two anonymous reviewers helped to improve this manuscript. This work was conducted as a part of the Computational Landscape Genomic working group at the National Institute for Mathematical and Biological Synthesis, sponsored by the National Science Foundation through NSF Award #DBI-1300426, with additional support from The University of Tennessee, Knoxville. In addition, SH was supported during this working group as a postdoctoral fellow at NIMBioS. Partial support for individuals provided by NSF DEB-1316549 to AS; NIH R01GM098856 to LKR; Washington State University, Army Research Office W911NF-15-1-0175, and NSF IOS-1557795 to JLK; Michigan State University to DBL; and by Northeastern University to KEL.

AUTHOR CONTRIBUTIONS

The idea for the manuscript was conceived collectively by all authors during an NSF National Institute for Mathematical and Biological Synthesis (NIMBioS) working group. All authors contributed to the writing of the manuscript.

FIGURE LEGENDS

Figure 1: Diagram depicting theoretical best case scenario for RADseq, where all RAD-tags (red line) are polymorphic and evenly distributed across the genome among linkage disequilibrium (LD) blocks (grey and white). In the case illustrated, RAD-tags are not in LD with 75% of the genome and thus, RADseq will miss many SNPs involved with local adaptation.

Figure 2: The theoretical maximum proportion of a genome that may be ascertained by RAD-tags from digestion with enzymes that recognize either a 6-bp or 8-bp recognition site, given the size of linkage disequilibrium blocks across the genome and assuming 50% GC content in the cut site and across the genome. Organisms with LD on the scale of: A) a few hundred base pairs include honey bee *Apis mellifera*, mosquito *Anopheles gambiae*, purple sea urchin *Strongylocentrotus purpuratus*, Zebra finch *Taeniopygia guttata*, diverse maize *Zea mays*, wild tomato *Solanum peruvianum*, white spruce *Picea glauca*, and European Aspen *Populus tremula L.*, Salicaceae; B) 1 Kb include mosquito *Anopheles arabiensis*, fruit fly *Drosphila melanogaster,* zebra fish *Danio rerio* in the wild, and coastal Douglas-fir *Pseudotsuga menziesii var. menziesii,* C) a few Kb include loblolly pine *Pinus taeda*, some populations of three-spined stickleback *Gasterosteus aculeatus,* and round worm *Caenorhabditis remanei*; D) greater than 10 Kb include Collared Flycatcher *Ficedula hypoleuca*, Siberian jay *Perisoreus infaustus,* some populations of three spined stickleback *Gasterosteus aculeatus,* European eel *Anguilla anguilla,* wild mouse *Mus musculus domesticus,* big horn

sheep *Ovis canadensis, Arabidopsis thaliana*, and wild yeast *Saccharomyces paradoxus* (Supplementary Table LD).

Figure 3: The proportion of the genome that will be sampled by RADseq (y-axes) for different numbers of RAD-tags (x-axes), genome sizes (line types), and lengths of linkage disequilibrium blocks.

Figure 4: The number of individuals (or populations) that can be sequenced as a function of genome size. It is assumed that a 6-bp cutter will cut on average every 4096 base pairs (assuming 50% GC content in cut site and in genome). For RADseq, exome capture, and whole genome sequencing we assume an average of 20x coverage. For transcriptome sequencing we assume 20-30 million reads per individual, and for whole genome pooled sequencing we assume 100x coverage. See also supplementary R script. Shown on the figure are some species for which RAD-seq has been applied to identify genomic signatures of selection. For RADseq, CI were based on a more frequent cutsite every 4096*0.6= 2,457.6 bp and less frequent cutsite 4096*1.4=5,734.4 bp. For exome capture, CI were based on half as many sequences captures and twice as many sequences captured. For transcriptome, we assume 25 million raw reads per individual, with confidence intervals from 20-30 million raw reads per individual. For whole genome data, we assume 20x coverage per individual, with confidence intervals from 10x to 30x coverage. For whole-genome pool-seq, we assume 100x coverage per pool and CI based on 80-120x coverage per pool.

TABLES

Table 1: Estimated length of average linkage disequilibrium for biological organisms.

Group	Species	Average LD	Reference
Invertebrates			
insect	fruit fly, *Drosophila melanogaster*	<1 Kb	Long et al., 1998
insect	mosquito, *Anopheles arabiensis*	1 Kb	Marsden et al., 2014
insect	mosquito, *Anopheles gambiae*	<< 1Kb	Harris et al., 2010
insect	honey bee, *Apis mellifera*	500 bp	Wallberg et al., 2014
insect	bumble bee, *Bombus impatiens*	< 10 Kb	Sadd et al., 2015
marine invertebrate	purple sea urchin, *Strongylocentrotus purpuratus*	very low to none	Pespeni et al., 2010
worm	round worm, *Caenorhabditis elegans*	very large	Cutter, 2006
worm	round worm, *Caenorhabditis remanei*	1-2 Kb	Cutter et al., 2006
Vertebrates			
bird	collared Flycatcher, *Ficedula hypoleuca*	400 Kb	Backström et al., 2006
bird	siberian jays, *Perisoreus infaustus*	6.28 Mb	Li & Backström, 2010
bird	zebra finch, *Taeniopygia guttata*	300 bp	Balakrishnan & Edwards, 2009
bird	black-footed Albatrosses, *Phoebastria nigripes*	75 bp	Dierickx et al., 2015
fish	three spined stickelback, *Gasterosteus aculeatus*	1 Kb	Roesti et al. 2015
fish	european eel, *Anguilla anguilla*	10-20 Kb	Pujolar et al., 2014
fish	zebra fish (wild), *Danio rerio*	<1 Kb	Whiteley et al., 2011
mammal	humans, *Homo sapiens*	50 Kb	Pritchard & Przeworski, 2001
mammal	wild mice, *Mus musculus domesticus*	50 Kb	Laurie et al., 2007
mammal	big horn sheep, *Ovis canadensis*	4 Mb	Miller et al., 2010
Plants			
flowering plant	*Arabidopsis thaliana*	10 Kb	Kim et al., 2007
flowering plant	diverse maize, *Zea mays*	400 bp	Tenaillon et al., 2001
flowering plant	wild tomato, *Solanum peruvianum*	<150 bp	Arunyawat et al., 2007
flowering plant	wild tomato, *Solanum chilense*	<750 bp	Arunyawat et al., 2007

flowering plant	chickpea, Cicer arietinum	> 1 Mb	Kujur et al, 2015
flowering plant	barrelclover, *Medicago truncatula*	5-10 Kb	Branca et al., 2011
flowering plant	wild soybean, *Glycine soya*	<500 Kb	Wang et al., 2015
flowering plant	wild rice, *Oryza rufipogon*	<<40 Kb	Mather et al., 2007
tree	scots pine, *Pinus sylvestris*	300 bp	Pyhäjärvi et al., 2007
tree	coastal Douglas-fir, *Pseudotsuga menziesii*	1.5 Kb	Eckert et al., 2009
tree	european aspen, *Populus tremula*	<500 bp	Ingvarsson, 2005
Fungi			
fungus	wild yeast, *Saccharomyces paradoxus*	50-100 Kb	Tsai et al., 2008

SUPPORTING INFORMATION

Table S1: Recent (January 2015 to April 2016) genome scan studies, which used RAD-seq for genotyping.

REFERENCES

Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama-Morita T, Stern DL (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, **21**, 610-617.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, **17**, 81-92.

Arunyawat U, Stephan W, Städler T (2007) Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Molecular Biology and Evolution*, **24**, 2310-2322

Arnold B, Corbett-Detig RB, Hartl D, K Bomblies (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179-3190.

Backström N, Qvarnström A, Gustafsson L, Ellegren H (2006) Levels of linkage disequilibrium in a wild bird population. *Biology Letters*, **2**, 435-438.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**, e3376.

Balakrishnan C, Edwards S (2009) Nucleotide variation, linkage disequilibrium and founder-facilitated speciation in wild populations of the zebra finch (Taeniopygia guttata). *Genetics*, **181**, 645–660.

Barrett RD, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics*, **12**, 767-780.

Benestan L, Gosselin T, Perrier C, Sainte-Marie B, Rochette R, Bernatchez L (2015) RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Molecular Ecology*, **24**, 3299-3315.

Bhakta MS, Jones VA, Vallejos CE (2015) Punctuated distribution of recombination hotspots and demarcation of pericentromeric regions in *Phaseolus vulgaris L*. *PLoS One*, **10**, e0116822.

Buerkle AC, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, **22**, 3028–3035.

Branca A, Paape TD, Zhou P, Briskine R, Farmer AD, Mudge J, Bharti AK, Woodward JE, May GD, Gentzbittel L, Ben C, Denny R, Sadowsky MJ, Ronfort J, Bataillon T, Young ND, Tiffin P (2011) Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula. Proceedings of the National Academy of Sciences USA*, **108**, E864–E870.

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proceedings of the National Academy of Sciences USA*, **101**, 15255–15260.

Catchen J, Bassham S, Wilson T, Currey M, O'Brien C, Yeates Q, Cresko WA (2013) The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. *Molecular Ecology,* **22**, 2864-2883.

Cavender-Bares J, González-Rodríguez A, Eaton DA, Hipp AA, Beulke A, Manos PS (2015) Phylogeny and biogeography of the American live oaks (*Quercus subsection* Virentes): A genomic and population genetics approach. *Molecular Ecology*, **24**, 3668-3687.

Combosch DJ, Vollmer SV (2015) Trans-Pacific RAD-Seq population genomics confirms introgressive hybridization in Eastern Pacific *Pocillopora* corals. *Molecular Phylogenetics and Evolution*, **88**, 154-162.

Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research*, **70**, 155-174.

Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, Abyzov A, Kong Y, Regan L, Gerstein M (2016) A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nature Communications*, **7**, 11101.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133-3157.

Cutter A (2006) Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer *Caenorhabditis elegans*. *Genetics*, **172**, 171–184

Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML (2013) Special features of RAD sequencing data: implications for genotyping. *Molecular Ecology*, **22**, 3151-3164.

Dierickx EG, Shultz AJ, Sato F, Hiraoka T, Edwards SV (2015) Morphological and genomic comparisons of Hawaiian and Japanese Black-footed Albatrosses (*Phoebastria nigripes*) using double digest RADseq: implications for conservation. *Evolutionary Applications*, **8**, 662-678.

Dufresne F, Jeffery N (2011) A guided tour of large genome size in animals: what we know and where we are heading. *Chromosome Research*, **19**, 925-938.

Eckert AJ, Wegrzyn JL, Pande B, Jermstad KD, Lee JM, Liechty JD, Tearse BR, Krutovsky KV, Neale DB (2009) Multilocus patterns of nucleotide diversity and divergence reveal positive selection at candidate genes related to cold hardiness in coastal Douglas Fir (*Pseudotsuga menziesii var. menziesii*). *Genetics*, **183**, 289-298.

Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, González-Martínez SC, Neale DB (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda L*., Pinaceae). *Genetics*, **185**, 969-982.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379.

Fischer MC, Foll M, Excoffier L, Heckel G (2011) Enhanced AFLP genome scans detect local adaptation in high-altitude populations of a small rodent (*Microtus arvalis*). *Molecular Ecology*, **20**, 1450-1462.

Garcia S, Leitch IJ, Anadon-Rosell A, Canela MÁ, Gálvez F, Garnatje T, Gras A, Hidalgo O, Johnston E, de Xaxars GM, Pellicer J (2014) Recent updates and developments to plant genome size databases. *Nucleic Acids Research*, **42**, D1159-D1166.

González-Martínez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB (2006) DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda L*. *Genetics*, **172**, 1915-1926.

Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH (2009) A first-generation haplotype map of maize. *Science*, **326**, 1115-1117.

Greminger MP, Stölting KN, Nater A, Goossens B, Arora N, Bruggmann R, Patrignani A, Nussberger B, Sharma R, Kraus RH, Ambu LN (2014) Generation of SNP datasets for orangutan population genomics using improved reduced-representation sequencing and direct comparisons of SNP calling algorithms. *BMC Genomics*, **15**, 16.

Guo Y, Yuan H, Fang D, Song L, Liu Y, Liu Y, Wu L, Yu J, Li Z, Xu X, Zhang H (2014) An improved 2b-RAD approach (I2b-RAD) offering genotyping tested by a rice (*Oryza sativa L*.) F2 population. *BMC Genomics*, **15**, 956.

Gugger PF, Cokus SJ, Sork VL (2016) Association of transcriptome-wide sequence variation with climate gradients in valley oak (*Quercus lobata*). *Tree Genetics & Genomes*, **12**,1-4.

Harris C, Rousset F, Morlais I, Fontenille D, Cohuet A (2010) Low linkage disequilibrium in wild *Anopheles gambiae s.l.* populations. *BMC Genetics*, **11**, 81.

Hereford J (2009) A quantitative survey of local adaptation and fitness trade-offs. *The American Naturalist*, **173***,* 579-588.

Heuertz M, De Paoli E, Källman T, Larsson H, Jurman I, Morgante M, Lascoux M, Gyllenstrand N. (2006) Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. *Genetics*, **174**, 2095-2105

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072-1079.

Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, Poss ML, Reed LK, Storfer A, Whitlock MC (2016) Finding the genomic basis of local adaptation in non-model organisms: pitfalls, practical solutions and future directions. *The American Naturalist*, *in press.*

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **367**, 395–408.

Hoekstra HE, Coyne JA (2007) The locus of evolution: evo devo and the genetics of adaptation. *Evolution*, **61**, 995-1016.

Houston RD, Davey JW, Bishop SC, Lowe NR, Mota-Velasco JC, Hamilton A, Guy DR, Tinch AE, Thomson ML, Blaxter ML, Gharbi K (2012) Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. *BMC Genomics*, **13**, 244.

Ingvarsson PK. (2005) Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). *Genetics*, **169**, 945-953.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, Birney E (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55-61.

Jones MR, Good JM (2015) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, **25**, 185–202.

Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. *Ecology Letters*, **7**, 1225-1241.

Kapun M, Fabian DK, Goudet J, Flatt T (2016) Genomic evidence for adaptive inversion clines in *Drosophila melanogaster*. Molecular Biology and Evolution *in press.*

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics*, **39**, 1151-1155

Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation. *Genetics*, **173**, 419-434.

Kujur A, Bajaj D, Upadhyaya HD, Das S, Ranjan R, Shree T, Saxena MS, Badoni S, Kumar V, Tripathi S, Gowda CL, Sharma S, Singh S, Tyagi AK, Parida SK (2015) Employing genome-wide SNP discovery and genotyping strategy to extrapolate the natural allelic diversity and domestication patterns in chickpea. *Frontiers in Plant Science*, **6**, 162.

Laporte M, Rogers SM, Dion-Côté AM, Normandeau E, Gagnaire PA, Dalziel AC, Chebib J, Bernatchez L (2015) RAD-QTL Mapping reveals both genome-level parallelism and different genetic architecture underlying the evolution of body shape in Lake Whitefish (*Coregonus clupeaformis*) species pairs. *G3: Genes, Genomes, Genetics*, **5**, 1481-1491.

Laurie CC, Nickerson DA, Anderson AD, Weir BS, Livingston RJ, Dean MD, Smith KL, Schadt EE, Nachman MW (2007) Linkage disequilibrium in wild mice. *PLoS Genetics*, **3**, e144.

Li MH, Merila J (2010) Extensive linkage disequilibrium in a wild bird population. *Heredity*, **104**, 600–610.

Long AD, Lyman RF, Langley CH, Mackay TF (1998) Two sites in the Delta gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. *Genetics*, **149**, 999–1017.

Lowry DB, Hernandez K, Taylor SH, Meyer E, Logan TL, Chapman JA, Rokhsar DS, Schmutz J, Juenger TE. 2015. The genetics of divergence and reproductive isolation between ecotypes of *Panicum hallii*. *New Phytologist*, **205**, 402-414.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, **23**, 2178-2192.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981-994.

Marsden CD, Lee Y, Kreppel K, Weakley A, Cornel A, Ferguson HM, Eskin E, Lanzaro GC (2014) Diversity, differentiation, and linkage disequilibrium: prospects for association mapping in the malaria vector *Anopheles arabiensis*. G3, 4,121-31.

Mather KA, Caicedo AL, Polato NR, Olsen KM, McCouch S, Purugganan MD (2007) The Extent of Linkage Disequilibrium in Rice (*Oryza sativa* L.). *Genetics*, **177**, 2223-2232.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240-248.

Miller JM, Poissant J, Kijas JW, Coltman DW, International Sheep Genomics Consortium (2011) A genome-wide set of SNPs detects population substructure and long range linkage disequilibrium in wild sheep. *Molecular Ecology Resources*, **11**, 314-322

Nachman MW, Payseur BA (2012) Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **367**, 409-421.

Nadeau NJ , Ruiz M, Salazar P, Counterman B, Medina JA, Ortiz-Zuazaga H, Morrison A, McMillan WO, Jiggins CD, Papa R (2014) Population genomics of parallel hybrid zones in the mimetic butterflies, H. melpomene and H. erato.. *Genome Research* **24**, 1316-1333.

Natarajan C, Projecto-Garcia J, Moriyama H, Weber RE, Muñoz-Fuentes V, Green AJ, Kopuchian C, Tubaro PL, Alza L, Bulgarella M, Smith MM (2015) Convergent evolution of hemoglobin function in high-altitude Andean waterfowl involves limited parallelism at the molecular sequence level. *PLoS Genetics*, **11**, e1005681.

Nosil P, Funk DJ, Ortiz-Barrientos (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375-402.

Ortiz-Barrientos D, Engelstädter J, Rieseberg LH (2016) Recombination rate evolution and the origin of species. *Trends in Ecology & Evolution*, **31**, 226-236.

Palazzo AF, Gregory TR (2014) The case for junk DNA. *PLoS Genetics*, **10**, e1004351.

Pardo-Diaz C, Salazar C, Jiggins CD (2015) Towards the identification of the loci of adaptive evolution. *Methods in Ecology and Evolution*, **6**, 445-464.

Pavy N, Namroud MC, Gagnon F, Isabel N, Bousquet J (2012) The heterogeneous levels of linkage disequilibrium in white spruce genes and comparative analysis with other conifers. *Heredity*, **108**, 273–284.

Pegadaraju V, Nipper R, Hulke B, Qi L, Schultz Q (2013) De novo sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach. *BMC Genomics*, **14**, 556.

Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions (2013) *Nature Reviews Genetics*, **14**, 288-295.

Pespeni MH, Oliver TA, Manier MK, Palumbi SR (2010) Restriction Site Tiling Analysis: accurate discovery and quantitative genotyping of genome-wide polymorphisms using nucleotide arrays. *Genome Biology*, **11**, R44.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Pfender WF, Saha MC, Johnson EA, Slabaugh MB (2011) Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*. *Theoretical and Applied Genetics*, **122**, 1467-1480.

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics*, **69**, 1-14.

Pujolar JM, Jacobsen MW, Als TD, Frydenberg J, Munch K, Jónsson B, Jian JB, Cheng L, Maes GE, Bernatchez L, Hansen MM (2014) Genome-wide single-generation signatures of local selection in the panmictic European eel. *Molecular Ecology*, **23**, 2514-2528.

Puritz JB, Matz M V, Toonen RJ, Weber JN, Bolnick DI, Bird CE (2014) Demystifying the RAD fad. *Molecular Ecology*, **23**, 5937–5342

Pyhäjärvi T, García-Gil MR, Knürr T, Mikkonen M, Wachowiak W, Savolainen O (2007) Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. *Genetics*, **177**, 1713-1724.

Qi ZC, Yu Y, Liu X, Pais A, Ranney T, Whetten R, Xiang QYJ (2015) Phylogenomics of polyploid *Fothergilla* (Hamamelidaceae) by RAD-tag based GBS—insights into species origin and effects of software pipelines. *Journal of Systematics and Evolution*, **53**, 432-447.

Rausher MD, Delph LF (2015) Commentary: When does understanding phenotypic evolution require identification of the underlying genes? *Evolution*, **69**, 1655-1664.

Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R (2015) A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, **24**, 4348-4370.

Roesti M, Kueng B, Moser D, Berner D (2015) The genomics of ecological vicariance in threespine stickleback fish. *Nature Communications*, **6**, 8767.

Ruegg K, Anderson EC, Boone J, Pouls J, Smoth TB (2014) A role for migration-linked genes and genomic islands in divergence of a songbird. *Molecular Ecology*, **23**, 4757-4769.

Sadd BM, Barribeau SM, Bloch G, de Graaf DC, Dearden P (2015) The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biology*, **16**, 76.

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics*, **14**, 807-820.

Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, **15**, 749–763

Schwander T, Libbrecht R, Keller L (2014) Supergenes and complex phenotypes. *Current Biology*, **24**, R288-R294.

Sessions SK, Larson A (1987) Developmental correlates of genome size in plethodontid salamanders and their implications for genome evolution. *Evolution*, **41**, 1239-1251.

Stanton-Geddes J, Paape T, Epstein B, Briskine R, Yoder J, Mudge J, Bharti AK, Farmer AD, Zhou P, Denny R, May GD, Eriandson S, Yakub M, Sugawara M, Sadowsky MJ, Young ND, Tiffin P (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**, e65688.

Stern DL, Orgogozo V (2008) The loci of evolution: how predictable is genetic evolution? *Evolution*, **62**, 2155-2177.

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, et al. (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays ssp. mays* L.). *Proceedings of the National Academy of Science USA*, **98**, 9161–9166.

The ENCODE Consortium. *Standards, Guidelines, and Best Practices for RNA-seq*. 2011; Available from:
https://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf.

Tiffin P, Ross-Ibarra J (2014) Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution*, **29**, 673-680.

Toonen RJ, Puritz JB, Forsman ZH, Whitney JL, Fernandez-Silva I, Andrews KR, Bird CE (2013) ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, **1**, e203.

Truong HT, Ramos AM, Yalcin F, de Ruiter M, van der Poel HJ, Huvenaars KH, Hogers RC, van Enckevort LJ, Janssen A, van Orsouw NJ, van Eijk MJ. Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLoS One*, **7**, e37565.

Tsai IJ, Bensasson D, Burt A, Koufopanou V (2008) Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. *Proceedings of the National Academy of Science USA*, **105**, 4957–4962

Van Orsouw NJ, Hogers RC, Janssen A, Yalcin F, Snoeijers S, Verstege E, Schneiders H, van der Poel H, Van Oeveren J, Verstegen H, Van Eijk MJ (2007) Complexity reduction of polymorphic sequences (CRoPS™): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One*, **2**, e1172.

Van Tassell CP, Smith TP, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, **5**, 247-252.

Wallberg A, Han F, Wellhagen G, Dahle B, Kawata M, Haddad N, Simões ZL, Allsopp MH, Kandemir I, De la Rúa P, Pirk CW, Webster MT (2014) A worldwide survey of genome sequence variation

provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nature Genetics*, **46**, 1081–1088.

Wang Y, Shahid MQ, Huang H, Wang Y (2015) Nucleotide diversity patterns of three divergent soybean populations: evidences for population-dependent linkage disequilibrium and taxonomic status of *Glycine gracilis*. *Ecology and Evolution*. **5**, 3969-3978.

Weber JN, Peterson BK, Hoekstra HE (2013) Discrete genetic modules are responsible for complex burrow evolution in *Peromyscus* mice. *Nature*, **493**, 402-405.

Whiteley AR, Bhat A, Martins EP, Mayden RL, Arunachalam M,Uusi-Heikkilä S, Ahmed ATA, Shrestha J, Clark M, Stemple D, Bernatchez L (2011) Population genomics of wild and laboratory zebrafish (*Danio rerio*). *Molecular Ecology*, **20**, 4259–4276.

Wray GA (2007) The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, **8**, 206-216.

Wright KM, Arnold B, Xue K, Šurinová M, O'Connell J, Bomblies K (2015) Selection on meiosis genes in diploid and tetraploid *Arabidopsis arenosa*. *Molecular Biology and Evolution*, **32**, 944-955.