OPEN

Heredity (2015) 114, 431–440

www.nature.com/hdy

REVIEW

Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation

C Schlötterer[1], R Kofler[1], E Versace[1,2], R Tobler[1,3] and SU Franssen[1]

Evolve and resequence (E&R) is a new approach to investigate the genomic responses to selection during experimental evolution. By using whole genome sequencing of pools of individuals (Pool-Seq), this method can identify selected variants in controlled and replicable experimental settings. Reviewing the current state of the field, we show that E&R can be powerful enough to identify causative genes and possibly even single-nucleotide polymorphisms. We also discuss how the experimental design and the complexity of the trait could result in a large number of false positive candidates. We suggest experimental and analytical strategies to maximize the power of E&R to uncover the genotype–phenotype link and serve as an important research tool for a broad range of evolutionary questions.
Heredity (2015) **114,** 431–440; doi:10.1038/hdy.2014.86; published online 1 October 2014

Experimental evolution has a long tradition in biology (Garland and Rose, 2009). By exposing an evolving population to conditions chosen by the researcher, it is possible to study the response to this selection regime. A recent review highlighted the broad range of applications that have been investigated with this methodology and concluded that the breadth of research questions is only limited by the creativity of the experimenter (Kawecki *et al.*, 2012). In addition to the great diversity of experimental designs, experimental evolution provides a unique advantage compared with other evolutionary analyses: the ability to replicate an experiment under identical conditions. Through this replication, experimenters are able to distinguish between stochastic and deterministic effects. Until recently, experimental evolution has mainly focused on phenotypes, sometimes combined with the analysis of a small number of markers (see, for example, Nuzhdin *et al.*, 1993; Teotonio *et al.*, 2009). In the wake of the latest sequencing technologies and the ongoing drop in DNA sequencing costs, however, the ultimate goal to connect the phenotypic response to the underlying genetic changes during an experimental evolution study has now come within reach.

Depending on the starting population, two conceptually different approaches of experimental evolution can be distinguished. Either the experiment starts from a genetically homogeneous (invariable) population or from a polymorphic population. In the first approach, adaptation occurs through the accumulation of new beneficial mutations during the experiment (Elena and Lenski, 2003). These experiments therefore require very large population sizes and many generations to ensure a sufficient mutation supply and are thus largely restricted to microorganisms. Alternatively, experiments starting with a polymorphic population do not require novel mutations as selection can act on beneficial alleles that are already present at the beginning of the experiment. Given the massive genetic variation that is present in the starting population, the key challenge for this approach is distinguishing between selected and neutral variants. Neither randomly selected markers nor whole genome sequencing of a few representative individuals can provide sufficient information about the true target(s) of selection. Rather, genome-wide polymorphism data are needed.

As whole genome sequencing is still not feasible for large numbers of individuals, experimental evolution studies starting from polymorphic base populations rely on a modified next-generation sequencing approach. Rather than sequencing individuals separately, DNA of multiple individuals from a population are sequenced together (Pool-Seq). This method is more cost effective than sequencing of individuals (Futschik and Schlötterer, 2010) and yields highly accurate genome-wide allele frequency estimates (reviewed in Rellstab *et al.*, 2013; Schlötterer *et al.*, 2014). The combination of experimental evolution with Pool-Seq is also known as Evolve and Resequence (E&R; Turner *et al.*, 2011; Figure 1). Here, we review the state of the art of whole genome polymorphism analysis in experimental evolution studies relying primarily on segregating variation in the starting population.

In many experimental evolution studies, researchers select for a well-defined trait in a controlled environment. This assures that both the phenotypic and the underlying genomic response are triggered either directly or indirectly by the selection regime applied during the experiment. Thus, E&R studies provide a complementary approach to genome-wide association studies (GWASs) and linkage mapping experiments as strategies to connect genotype and phenotype.

[1]Institut für Populationsgenetik, Vetmeduni Vienna, Vienna, Austria; [2]Center for Mind/Brain Sciences, University of Trento, Rovereto, Italy and [3]Vienna Graduate School of Population Genetics, Vienna, Austria
Correspondence: Professor C Schlötterer, Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, 1210 Vienna, Austria.
E-mail: christian.schloetterer@vetmeduni.ac.at
Received 8 April 2014; revised 1 July 2014; accepted 14 July 2014; published online 1 October 2014

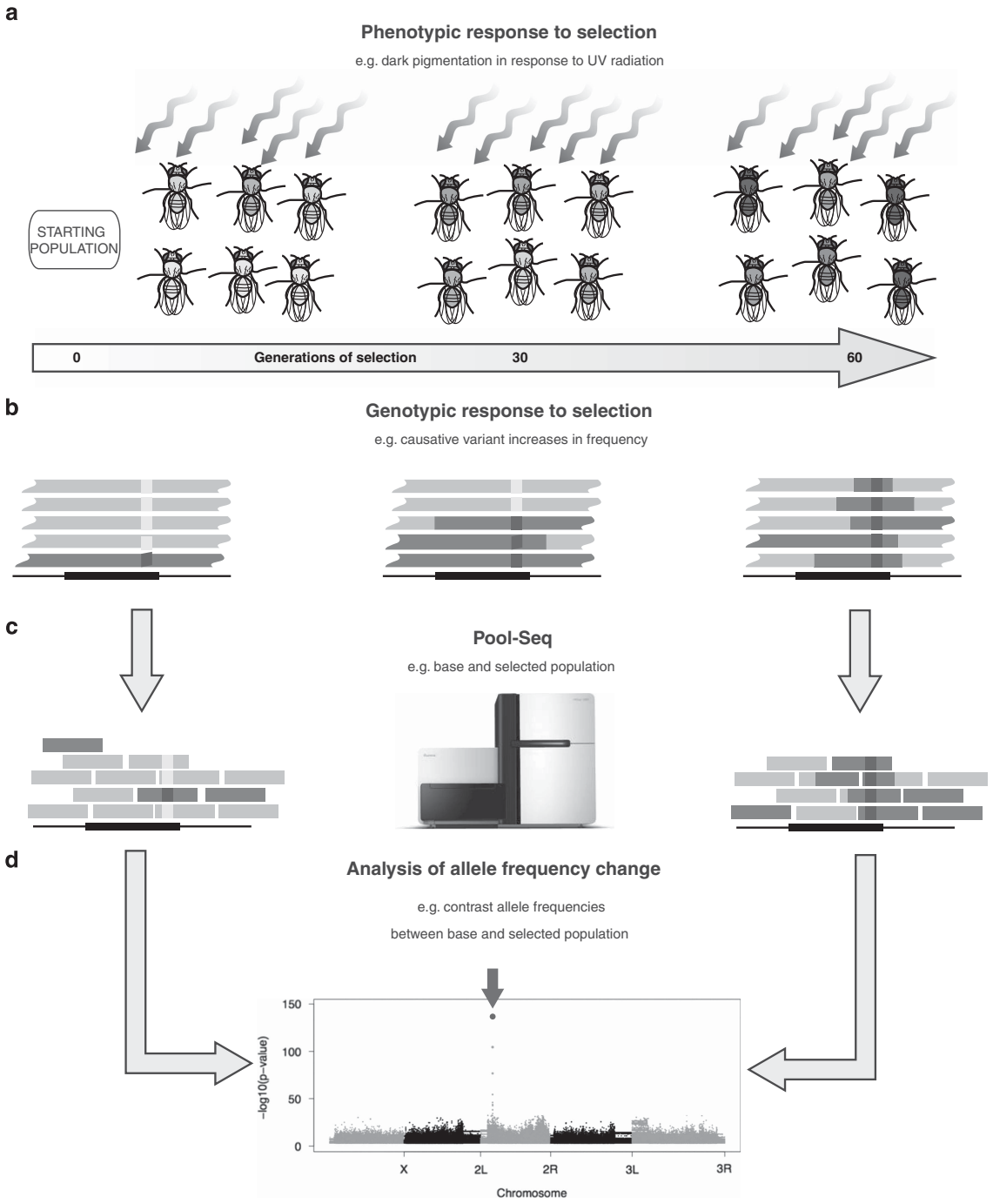

Figure 1 Overview of E&R studies. (**a**) A population of flies is exposed for 60 generations to ultraviolet (UV) radiation (purple arrows). We assume here, for the sake of illustration, that darker pigmentation is beneficial in high UV environments, whereby darker flies will increase in frequency. (**b**) At the genotypic level, the allele frequency of the causative allele (dark brown) will increase, more so than hitchhiking variants (dark gray background) that will be recombined onto other backgrounds (breaks between dark and light gray background). (**c**) The allele frequencies of the starting population and the selected population are measured with Pool-Seq. (**d**) Causative variants can be identified by contrasting the allele frequencies between base and selected population and visualized with Manhattan plots. A full color version of this figure is available at the *Heredity* journal online.

SUPPORT FOR EXPERIMENTAL EVOLUTION RESULTS BY INDEPENDENT METHODS

Despite its conceptual appeal, E&R studies face a lot of uncharted territory. For example, guidelines for experimental design and data analysis (Box 1) were not available for the first studies. Therefore, several E&R studies used additional techniques to provide independent evidence to support the E&R results. In the following, we will highlight the results of E&R studies based on truncating selection and accompanying validation strategies adopted by some of these studies.

Hypoxia tolerance

Having selected a *Drosophila melanogaster* population for an increased ability to tolerate low oxygen concentrations over 200 generations, Zhou *et al.* (2011) identified 188 candidate genes located in genomic

Box 1 Recommendations for experimental design

In the following, two types of genetic loci are referred to as strong effect loci (SEL) and weak effect loci (WEL) that have selection coefficients of 0.1 and $\leqslant 0.25$ respectively.

The experimental population

Duration of experiment: minimum of 10–20 generations for SEL and >100 for WEL.

Experimental population size: minimum of 500 individuals for SEL and >1000 for WEL.

Number of replicates: minimum of 3 for SEL and >10 for WEL.

Number of distinct chromosomes in the starting population: the larger the better.

General remark: given budget constraints, it is difficult to increase both the population size and the number of replicates. In this case, we favor increased replication over larger population sizes, because this increases the power to identify strongly selected loci while leaving weakly selected loci unaffected. Moreover, replication acts as a buffer against accidental population loss, for example because of a bottleneck or viral infection.

Sequencing of pooled individuals (Pool-Seq)

Pool size: maximize the number of individuals in the Pool.

Coverage: minimum of $50\times$ for SEL and $>200\times$ for WEL.

Read type: only paired-end with at least 100 bp read length.

Recommendations for data analysis

Trimming: reads should be trimmed to remove poor-quality bases (base quality <20).

Reference genome: use the conspecific reference genome or that of a closely related species; note however that even closely related species may cause considerable biases.

Mapping: allow for gaps, avoid seeding, align the entire read (semiglobal alignment) and take advantage of the second read by realigning unmapped mates.

Filtering: remove broken pairs, only use reads with mapping quality >20, remove positions flanking indels, remove duplicate reads, mask repetitive sequences and potential copy number variants and treat low recombining regions separately.

Test statistic: properties of novel test statistics should be tested by computer simulations using software tools like MimicrEE (Kofler and Schlötterer, 2014) or forqs (Kessner and Novembre, 2014a).

regions that responded to the selection regime. Of these genes, 28 were previously implicated in hypoxia or similar phenotypes and 12 were linked to the *Notch* pathway. As previous gene expression studies had associated the *Notch* pathway with hypoxia, the authors concluded that they had successfully identified genes involved in this trait. We note, however, that the genomic regions reported to be responding to selection were rather large, probably because of a small number of founder haplotypes (see below). Hence, it is not clear how many of the candidate genes were actually selected in the population during the study.

Genetic basis of aging

Experimental evolution has a long tradition in the study of aging and other life history traits. Remolina *et al.* (2012) selected long-lived flies for 50 generations and compared them with unselected controls propagated in a similar manner. Contrary to many other studies that inferred selection on the basis of individual single-nucleotide polymorphisms (SNPs), this study searched for regions of reduced variability in 1 kb windows. In total, 156 genes were identified to show the signature of adaptive divergence between selected and unselected lines. To validate these candidates, the authors measured gene expression divergence between both groups. Using a false discovery rate of <0.1, 25 candidate genes were found to be differentially expressed.

Parasitoid resistance

Following just five generations of selection for resistance against the parasitoid *Asobara tabida*, a consistent level of resistance was seen across all four replicates of a *D. melanogaster* population (Jalvingh *et al.*, 2014). Contrasting allele frequencies in selected and control flies using a similar window-based approach as Remolina *et al.* (2012), the authors concluded that $<5\%$ of the genome was influenced by selection. Among 345 genes located in the selected regions, 91 could be linked to pathways associated with immune response. Using two independent expression analyses related to *A. tabida* resistance, the authors found that some differentially expressed genes were located in selected regions, but no significant overlap between their data set and either of the expression analyses could be detected.

Courtship song in *Drosophila*

The vibration of wings is an important courtship signal in fruit flies, and has species-specific features. Male courtship song is characterized by several features, including the length of the interpulse interval (IPI). After demonstrating variation for this trait among *D. melanogaster* strains, Turner and Miller (2012) selected from a polymorphic base population for short and long IPI. After 14 generations, the experimental populations had diverged for IPI and were subjected to Pool-Seq. Despite a low empirical false discovery rate of 0.42%, $>13\,000$ variants were significantly differentiated between short and long IPI flies. A significant under-representation of candidate SNPs on the X chromosome was also observed. In a subsequent study, the authors performed a GWAS based on the same set of lines that were used to generate the starting population for the experimental evolution study (Turner *et al.*, 2013). Although no SNP was significant after correction for multiple testing, SNPs with a high differentiation in the E&R study also tended to have low *P*-values in the GWAS. Conversely, none of the most significant SNPs in the GWAS were found among the 13 000 most differentiated ones in the experimental evolution study. This was taken as evidence for variation in IPI being caused by many loci, rather than a small number of large effect genes. Validation of two candidate genes by quantitative complementation tests revealed that one of them, *Syntropin-like 1*, had a small, but significant, effect on IPI.

Drosophila C virus resistance

The *Drosophila* C virus (DCV) is one of the best-studied pathogens of *D. melanogaster*. Exposing a natural population sample for 20 generations to DCV resulted in an increased survival after infection, suggesting that resistance alleles increased in frequency (Martins *et al.*, 2014). By applying Pool-Seq in four replicate populations, Martins *et al.* (2014) identified two genomic regions where a variant increased in frequency in the selected populations. Interestingly, one of these regions had also been identified in a previous GWAS (Magwire *et al.*, 2012) and in both studies the same SNP in the gene *pastrel* was identified as the most likely target of selection. In addition, the involvement of *Ubc-E2H* in the second candidate region was validated with RNA interference. With the same SNP being indentified with E&R and GWAS, the DCV resistance is probably the most convincing example of E&R having identified the causative link between genotype and phenotype.

EVOLUTIONARY INFERENCES OF SELECTION TRAJECTORIES

The E&R method can potentially offer much more than measuring differentiated allele frequencies between two selection regimes or between selected and control populations. By sampling evolving populations at multiple time points, it is also possible to study the trajectories of the selected alleles and thus elucidate their evolutionary dynamics. Such trajectories have been largely studied in a modified experimental evolution design termed laboratory natural selection (Garland and Rose, 2009). Rather than selecting for a specific phenotype, this approach exposes populations to a defined environment where, as in nature, better adapted individuals have a higher reproductive success. Surprisingly, all studies of allele frequency dynamics detected a similar behavior of selected alleles.

Parts *et al.* (2011) generated a polymorphic population of recombinant cells from two diverged yeast strains and then subjected it to high temperature for up to 12 generations. By following allele frequency changes during adaptation, the authors were not only able to pinpoint 21 selected genomic regions, but they also captured interesting dynamics for these loci that were not compatible with classic directional selection. Rather than increasing in frequency until becoming fixed, most favored alleles plateaued at intermediate frequencies. This reduction in the selection coefficient was later confirmed via an elegant population genetic model (Illingworth *et al.*, 2012).

A similar pattern has also been reported for two *D. melanogaster* E&R studies. The first study exposed a *D. melanogaster* population to a novel high-temperature environment and sampled allele frequency changes at two different time points, specifically after 15 and 37 generations (Orozco-terWengel *et al.*, 2012). Among all SNPs with allele frequency changes greater than expected under genetic drift during the experiment, the authors focused on the 2000 SNPs showing the most significant change across all three replicates. Although in the first 15 generations the majority of the candidate SNPs experienced a frequency increase of ~28%, in subsequent generations the allele frequencies had plateaued without becoming fixed. Most importantly, the authors also ruled out the possibility that this pattern was an analytical artifact (Orozco-terWengel *et al.*, 2012). The second *D. melanogaster* study, which did not analyze time series data, reported a large genomic response in flies selected for accelerated development over 600 generations, but found little support for selective sweeps resulting in their fixation (Burke *et al.*, 2010).

The reason for these puzzling dynamics is not yet understood. In experiments with changing environments, such as the fluctuating temperature used by Orozco-terWengel *et al.* (2012), marginal overdominance may explain the plateauing of selected alleles. In the other two experiments where selection was constant, thus marginal overdominance cannot serve as a universal explanation. Alternatively, recessive deleterious alleles or heterozygous advantage may explain the plateaus in the *Drosophila* data, but not in the haploid yeast strains. Finally, it has been proposed that the observed pattern could be explained by selection on a complex trait with several contributing loci: pronounced allele frequency changes are expected as long as the trait is far away from the fitness optimum, but slows down as the optimum is being approached (Chevin and Hospital, 2008). Further empirical testing is required to distinguish between the different explanations for the plateauing of putatively selected SNPs.

THE MYSTERY OF THE LARGE NUMBER OF CANDIDATE SNPS

One common observation in all E&R studies is that a massive number of candidate SNPs are identified, even after rigorous correction for multiple testing. Importantly, such large numbers of selected SNPs are not compatible with the observed large frequency changes, wherein 30% increases for selected alleles are not uncommon (Smith, 1968; Nuzhdin and Turner, 2013). One apparent explanation for the large number of candidate SNPs is that selection may act on a moderate number of loci that drag along many linked neutral variants, a phenomenon known as hitchhiking. In particular, studies that rely on either small experimental population sizes or have starting populations with high levels of linkage disequilibrium are expected to show in a selection signature comprising broad regions of adjacent SNPs.

Although this pattern can be clearly seen in some E&R studies, it is not sufficient to explain the patterns observed for populations with large population sizes. For example, two studies on flies exposed to new thermal environments (Orozco-terWengel *et al.*, 2012; Tobler *et al.*, 2013) tested explicitly whether narrow-range linkage could explain the excess of significant SNPs, but found that only SNPs within ± 200 bp of the focal SNPs were affected, ruling this out as a general explanation. Tobler *et al.* (2013) further investigated this question by comparing independent sets of replicates that were started from the same base population and had subsequently evolved independently from each other under the same selection regime. The authors found a very good general concordance between replicates for SNPs with allele frequency changes deviating from neutral expectations. However, this concordance was also apparent when only short introns were analyzed. Because short introns are, to a good approximation, evolving neutrally in *D. melanogaster*, no significant concordance is expected among sites located in these regions. The authors concluded that instead long-range linkage disequilibrium with selected sites may be responsible for the correlated response of SNPs located in short introns. Such long-range linkage disequilibrium could result from segregating chromosomal inversions that are common in *D. melanogaster*. In fact, using inversion-specific SNP markers, it has been shown that some inversion frequency changes in these experimental populations were probably driven by selection (Kapun *et al.*, 2014). Inversions are not the only cause of long-range linkage disequilibrium, however. Beneficial alleles occurring at a low frequency in the starting population will, by chance, have an association with all SNPs private to the haplotypes upon which the beneficial variant occurs. The lower the starting frequency of the beneficial allele, the more spurious long-range associations will be generated. By using individual-based computer simulations, Tobler *et al.* (2013) demonstrated that such long-range linkage disequilibrium does result in many false positives because of linkage extending over several megabases. Consistent with this idea, a 1-Mb genomic region on chromosome 3R was found to harbor a large number of candidate SNPs with many putatively selected alleles in this region having risen from low frequencies in the starting population to high frequencies in the evolved populations (Orozco-terWengel *et al.*, 2012; Tobler *et al.*, 2013).

THE FUTURE OF E&R

Based on the results of recent E&R studies, it is apparent that E&R could be a powerful method to complement ongoing linkage mapping and GWAS approaches (Table 1). This has been demonstrated by a recent E&R study that identified the causative SNP for at least one gene determining a trait with a simple genetic basis (DCV resistance) (Martins *et al.*, 2014), whereas complementary results were obtained for a GWAS and an E&R study on a more complex trait (see, for example, Turner *et al.*, 2013). Nevertheless, is has also become clear that E&R faces its own specific challenges that need to be considered when interpreting the data, some of them are discussed below.

Table 1 Features of different approaches aiming to link genotype and phenotype

	E&R	*Classic GWAS*	*GWAS in reference panel*	*Pool-GWAS*	*Linkage mapping*
Analysis of heterozygous individuals	+	+	−	+	+
Repeated phenotyping	Every generation	−	+	−	−
Sensitivity to environmental noise	Low because of repeated phenotyping in every generation and replication	High	Low because of repeated phenotyping of identical genotypes	Moderate because of replication	High
Well-established analysis strategies	−	+	+	−	+
Mapping resolution	High	High	High	High	Moderate (cost effective), high (expensive)
Genetic diversity analyzed	High	High	Moderate–high	High	Limited to parental genotypes
Inference of effect size	Selection, coefficient	+	+	−	+
Randomized genetic background	+, in starting population and repeated mixing by sexual reproduction during the experiment	+, but sensitive to population structure that can be accounted for in analysis	+, but sensitive to population structure that can be accounted for in analysis given a sufficient sample size	+, but sensitive to population structure that can be accounted for in analysis	+
Genotyping/sequencing costs	Low because of Pool-Seq	High for establishment, no costs for follow-up experiments	High for establishment, no costs for follow-up experiments	Low because of Pool-Seq	Low because of the use of genetic markers (Rad-Tag sequencing)
Sampling effort	High because of maintenance of experimental populations	Depends on species	High for establishment, low later on	Moderate	Moderate
Analysis of multiple traits from the same genotypes	−	+	+	−	+
Replication	Yes, is common practice	Only across different populations	Requires an independent reference panel	Yes, it is common practice. Easy to expand to multiple populations	Requires independent mapping families
Influence of allele frequency (conditional on presence in the sample)	Low power for high-frequency alleles, low-frequency alleles are often lost	Yes	Yes	Yes	No
Trajectories of selected variants	+	−	−	−	−
Identification of adaptive variants in a defined environment	+	−	−	−	−

Abbreviations: E&R, Evolve and resequence; GWAS, genome-wide association study; Pool-Seq, Pool-sequencing.

Towards improved experimental designs

Current E&R studies employ a diverse array of experimental designs, but until very recently no guidance was available on how to optimize the power of these designs to detect selected loci. Three forward simulation studies (Baldwin-Brown *et al.*, 2014; Kofler and Schlötterer, 2014; Kessner and Novembre, 2014b) have explored the most important factors for an optimal experimental design. All three studies showed that increasing the number of replicates and experimental population size resulted in a higher power to detect selected loci. The strength of selection was also found to have a major impact, with both very strong and very weak selection being problematic (Kofler and Schlötterer, 2014). Although weakly selected sites failed to show a detectable allele frequency change, strong selection caused the fixation of many linked neutral variants, precluding the identification of the causative SNP. Furthermore, the detection of selected alleles becomes more difficult as the experiment continues, as causative SNPs eventually become fixed in the population while drift gradually reduces the signal-to-noise ratio.

The history of the starting population also has an important influence on the power of the study. The results showed that the amount of variation in the starting population is key. Experiments with starting populations using as many independent lines as possible had the highest power because of the low level of linkage disequilibrium. Another approach to reduce linkage disequilibrium is pre-experiment cultivation of the starting population in the laboratory. Computer simulations showed that this experimental approach resulted in the loss of favorable alleles and that the increase in power was moderate with strongly selected alleles benefitting most from this strategy (Kofler and Schlötterer, 2014). Importantly, laboratory adaptation during the pre-experiment cultivation probably does not confound the subsequent analysis: in recent E&R study, the same starting population was selected in two different environmental conditions, and very few SNPs appeared to be selected in both treatments, and this is unexpected if laboratory adaptation is important (Tobler *et al.*, 2013).

One further factor influencing the power of a study is the sequence coverage. Although for strongly selected sites a coverage of 50× is fully sufficient, weakly selected sites require a substantially higher coverage (up to 200×) to estimate the allele frequencies to a level of precision that permits the reliable detection of small frequency differences (Kofler and Schlötterer, 2014) (Box 1). Nevertheless, even when a large number of loci are selected, almost 60% of the target sites can be identified using an appropriate experimental design (Kofler and Schlötterer, 2014). This aptly demonstrates the enormous potential of experimental evolution to identify the target(s) of selection. One further strategy to improve the performance of E&R studies is to include haplotype information (Kessner *et al.*, 2013). Current methods require knowledge of the haplotypes in the starting population, however, that will become increasingly difficult with larger number of founder chromosomes.

Because of its compact genome, high-quality reference genomic sequence, short generation time and ease of cultivation, *D. melanogaster* has been frequently used for E&R studies. However, *D. melanogaster* harbors many segregating inversions that could negatively affect the power of experimental evolution. Therefore, we suggest that future experiments make use of *D. simulans*, a close relative of *D. melanogaster*, that is almost free of segregating inversions (Aulard *et al.*, 2004). Adding to the attraction of *D. simulans*, it has a substantially improved reference genome (Hu *et al.*, 2013; Palmieri *et al.*, 2014), and with latest advances in genome editing (Liu *et al.*, 2013; Terns and Terns, 2014) rigorous functional testing can also be applied outside of genetic model organisms. In addition to changing the focal species, we strongly recommend increasing the number of replicates, number of founder chromosomes and the experimental population size. Although population sizes of ~1000 individuals are currently at the upper end of *Drosophila* E&R studies, it is possible to increase this number by an order of magnitude. Not only is this expected to have an impact on the dynamics of phenotypic change (Weber, 1996), but also on the accuracy of the identification of targets of selection (Baldwin-Brown *et al.*, 2014; Kofler and Schlötterer, 2014). In order to identify causative variants with E&R, several test statistics have been developed (Box 2), some of which show remarkable differences in statistical power under a given evolutionary scenario (Figure 2). Furthermore, new statistical approaches that take full advantage of trajectories from multiple time points and across several replicates have the potential to increase the power of E&R studies substantially (Terhorst and Song, 2014; Topa *et al.*, 2014).

For the long-term success of experimental evolution studies of adaptation from standing variation, it would be helpful to introduce other models that have short generation times, can be cultivated at large effective population sizes and have high recombination rates to uncouple linked sites. Notably, the widely used model organisms yeast and *Caenorhabditis elegans* are not optimal for this purpose, as recombination is possible only under restricted conditions for these species (see, for example, Parts *et al.*, 2011; Teotonio *et al.*, 2012). One possible model, however, may be *Caenorhabditis remanei* that is obligate sexually reproducing and can be cultivated at large population sizes. Furthermore, natural populations appear to harbor substantial levels of natural variation (Cutter *et al.*, 2006).

Validating candidates from E&R studies

It is important to distinguish between validating allele frequency estimates obtained from Pool-Seq and validating candidate loci identified in E&R studies. Because Pool-Seq has been shown to obtain reliable allele frequency estimates when some minimum quality criteria are met (Rellstab *et al.*, 2013; Schlötterer *et al.*, 2014), we will focus here on the second aspect of validation.

Traits with a simple genetic basis, such as DCV resistance, are best validated by functional analysis of the identified genes and variants. Whether the preferred approach is knockdown of the identified genes by RNA interference, quantitative complementation tests or allelic replacements of candidate SNPs depends on the trait of interest. The validation of candidates for complex traits, however, is a notoriously challenging enterprise as the effect sizes of individual mutations tend to be very small. Confirming the predicted effects of candidate variants by another method such as GWAS provides another feasible strategy (Turner *et al.*, 2013). Lack of replication does not necessarily indicate lack of an effect, however; for example, if different populations or samples are used for GWAS and E&R, the validation is complicated by allele frequency variation and possible epistatic interactions. Reversing the selection regime may be a particularly appealing validation approach for some E&R experimental designs. Populations that have been selected to drive a trait in one direction could subsequently be selected in the opposite direction, for example, by moving a population from a high to a low temperature regime. Previously, it has been shown that these reverse selection schemes can change the phenotype and allele frequencies at SNP markers in the opposite direction (Teotonio *et al.*, 2009). Nevertheless, reverse selection will address the problem of linkage between selected and neutral sites only to a moderate extent. One further possibility to validate candidate loci with small effects is via experimental evolution with competing genotypes that differ only in the allele(s) of interest. In these competition assays,

Box 2 Overview of the different statistical approaches to infer targets of selection in E&R studies

In these studies, Pool-Seq is performed for at least two treatments, the population evolved under the selection regime of interest and a control (or ancestral) population, ideally for several replicate pairs. Typically, allele counts or frequencies are determined for each individual SNP or sliding windows along the genome based on the Pool-Seq data. Tests are performed on each SNP (window) individually.

Test statistics based on allele frequency differences between two populations from two different treatments (for example, selected vs control or ancestral). The performance of test statistics allowing for replicates is shown in Figure 2.

Fisher's exact test: This statistical test generally operates on contingency tables. It can be applied to allele counts of biallelic SNPs between selected and control populations in the absence of replication. Fisher's exact test was used in the first genome-wide E&R study in *D. melanogaster* selected for developmental time (Burke *et al.*, 2010). Genetic drift during the experiment violates the null model of Fisher's exact test (and CMH test), and thus in the absence of an empirical false discovery rate (FDR), these tests can be only used to rank candidates.

CMH: The Cochran–Mantel–Haenszel statistics can be used on data arranged in multiple, associated 2×2 contingency tables. The null hypothesis of the test assumes independence of treatment levels in each table, thereby accounting for multiple replicates. It was first employed by Orozco-terWengel *et al.* (2012) to investigate the genomic response in *D. melanogaster* adapting to a novel temperature environment.

S_f statistic: $S_f(C_i, T_i)$ is the log ratio of control and treatment scaled mutation rates of replicate pair *i*, which is a measure comparing the effective population sizes. It was introduced by Zhou *et al.* (2011) to investigate the genetic basis of hypoxia tolerance in *D. melanogaster*, where the statistic was calculated for overlapping 50 kb windows. Final candidate regions were defined by overlapping candidate regions between replicates.

SFselect: This method is based on supervised learning to differentiate features of the site frequency spectrum that best separate different types of selective sweeps from neutrality (Ronen *et al.*, 2013). XP-SFselect, an extended approach for cross-population testing, was applied to the data from Zhou *et al.* (2011) in a window-based approach without explicit use of replicate populations and was found to be more robust than the S_f statistic. The test is designed to detect selection on new mutations/singletons (hard sweeps).

diffStat: The diffStat statistic only considers SNPs for which all replicates of one treatment show higher (or lower) allele frequencies than all replicates of the control. On this subset of SNPs, diffStat is determined as the minimum allele frequency difference between all possible replicate pairs between treatment and control. This method was introduced in Turner *et al.* (2011) investigating the genetic basis of body size in *D. melanogaster*.

Association statistic: The association statistic calculates the absolute value of summed up frequency differences between all replicate pairs, that is, abs(Σ (control – selected)). It was introduced by Turner and Miller (2012), investigating the genetic basis in length of the interpulse interval courtship song of *D. melanogaster*.

Hs/D: The statistic describes the scaled heterozygosity of selected populations (Hs) relative to the divergence between selected and control populations D (calculated pairwise). This statistic aims to detect classical sweep patterns and was suggested along with the Hs/Hc statistic (see below) by Remolina *et al.* (2012). Both statistics were used for 1 kb non-overlapping windows to investigate the genomic response to divergent selection for lifespan and late-age fertility in *D. melanogaster*.

Hs/Hc: This statistic is used analogously to Hs/D but scaled by the heterozygosity in the control population instead in order to capture signals of incomplete sweeps (Remolina *et al.*, 2012).

F_{ST}: The population genetic parameter F_{ST} between selected and control populations obtained by pooling allele counts within a treatment level over replicates was used by Remolina *et al.* (2012).

MAF S-C: This measure simply describes the difference in the major allele frequency between the selected and the control population. It is comparable to the association statistic; here, however, it was calculated by pooling allele counts between replicates (Remolina *et al.*, 2012).

Model of divergence: The statistics developed by Kelly *et al.*, (2013) can be divided into two parts. First, the alleles are arcsin square root transformed to allow variance inflation due to successive sampling events (sampling of individuals, library construction, sequencing coverage and drift) to be modeled. Then, divergence is estimated (using an F_{ST} analog) from the transformed data and a nonparametric method is applied to identify significant outliers (Kelly *et al.*, 2013). The method was used for single SNPs or SNP-windows without replicate information to investigate the genetic basis of corolla (lower flower lip) width in *Mimulus guttatus*.

Usage of replication to identify targets of selection

Generally, biological replicates are crucial to differentiate targets of selection from neutral hitchhikers or genetic drift. It is therefore important how different statistics deal with present replication. For example, whereas the CMH test explicitly models replication, Fisher's exact test is not designed for experiments with replication. Test statistics that do not explicitly model replicates have incorporated replication in different ways: (1) usage of overlapping candidate regions between replicates as done for the $S_f(C_i, T_i)$ statistic in Zhou *et al.* (2011), (2) a conservative definition via the least extreme changes between all selected–control comparisons as done in Turner *et al.* (2011), (3) pooling of allele counts within treatments before calculating the test statistic as done for the four test statistics in Remolina *et al.* (2012) or (4) calculation of a composite log likelihood statistic on results of single replicates introduced in Remolina *et al.* (2012) for their test statistics.

Single SNPs versus window-based approaches

Most of the above statistics can be applied for SNP and window-based approaches. Although some tests as the $S_f(C_i, T_i)$, Hs/D, Hs/Hc and XP-SFselect were originally used for window-based approaches, others have primarily been applied to single SNPs. Note that window-based approaches, in particular large windows, implicitly test for classic selective sweeps, rather than sweeps caused by selection on sites segregating at intermediate frequencies in the starting population.

Statistics that have been designed to identify targets of selection via changes in allele frequency from time series data during experimental evolution

Driver and passenger model: This model is based on population genetic theory, designed to identify selected alleles from time series allele frequency data of replicated evolving outbred population under selection (Illingworth *et al.*, 2012). It was specifically designed for the analysis of time series data in which a large pool of recombinants from two yeast strains was initially created and subsequently evolved asexually under heat stress. Because of the large population size with respect to the estimated number of generations in the experiment, drift was not included in the model.

BBGP: This method is based on a β-binomial Gaussian process model designed to rank SNPs with significant changes in allele frequency over time (Topa *et al.*, 2014). The model consists of two parts: a β-binomial model is included to capture uncertainties in frequency estimates due to limited sequencing coverage and the Gaussian process models the time-dependent behavior that indicates selection and an error due to genetic drift. Replicates are explicitly included in the model. The method was found to outperform the CMH test on simulated whole genome data. Furthermore, it was used to reanalyze the time series data from Orozco-terWengel *et al.* (2012).

Multilocus analysis: Using a Gaussian process approximation to the multilocus Wright–Fisher process with selection, the method models multiple linked sites during a time series (Terhorst and Song, 2014). It directly incorporates replicates and sampling coverage and can be used to estimate population genetic parameters, that is, the selection coefficient, dominance, recombination rates and effective population size. The current implementation of the method requires information about all founder haplotypes and is limited to a single selected site in a given genomic region, and thus it is not directly suited to identify targets of selection on a genome-wide scale.

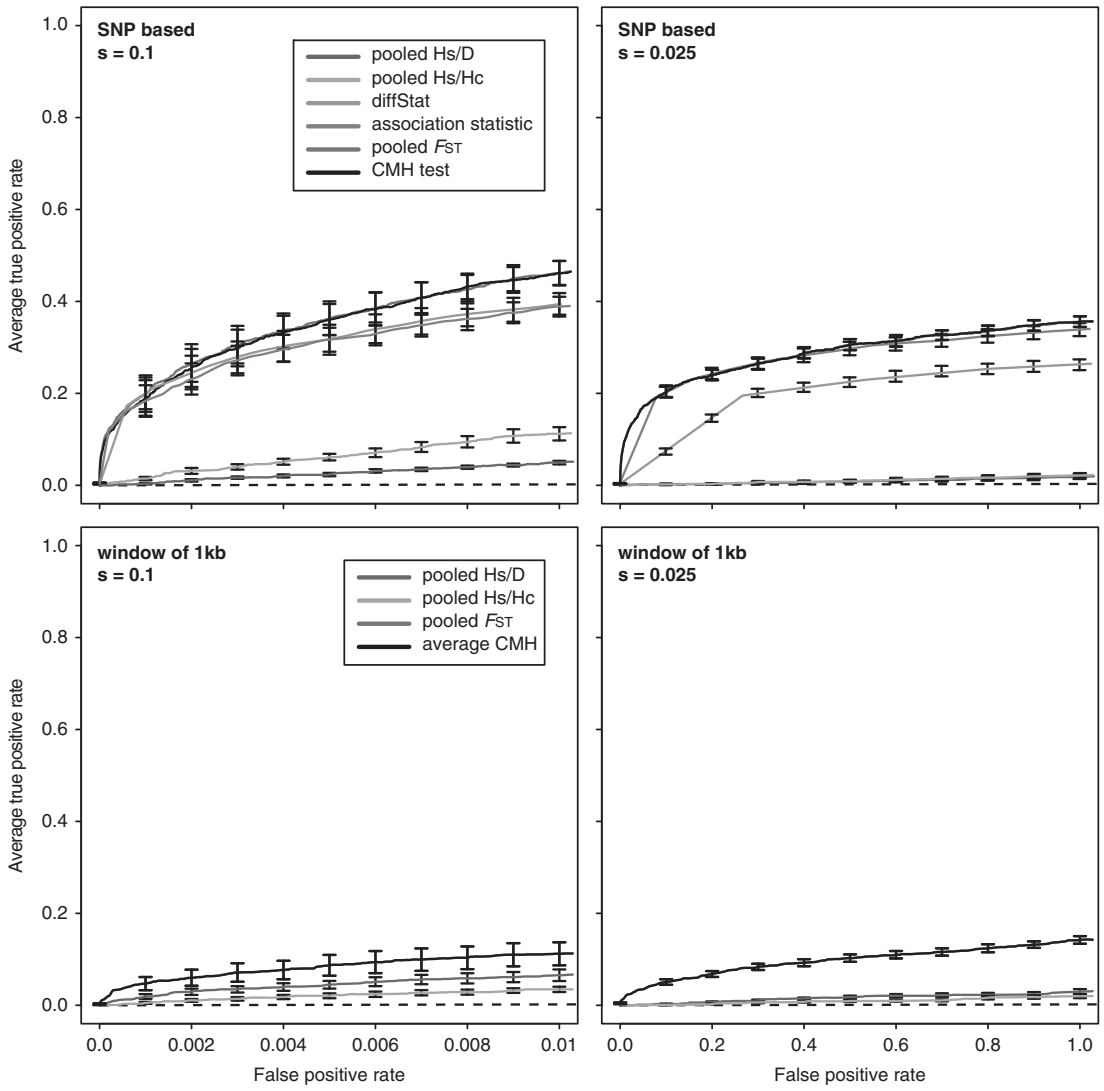

Figure 2 Performance of different test statistics used in E&R studies. Receiver operator characteristic (ROC) curves that contrast the true positive rate with the false positive rate. We extended the results of Kofler and Schlötterer (2014) by including the pooled Hs/D test, the pooled Hs/Hc test and the pooled F_{ST} test (Remolina *et al.*, 2012). Briefly, Kofler and Schlötterer (2014) simulated E&R with a base population that captures the pattern of polymorphism in a natural *D. melanogaster* population. They simulated 60 generations of selection with a population size of 1000 and 3 replicates. Results are shown for SNP-based analysis (top graphs) and for a window-based (bottom graphs) analyses using either 150 strongly (left graphs) or 150 weakly (right graphs) selected loci. The behavior of the Cochran–Mantel–Haenszel (CMH) and pooled F_{ST} tests are very similar, resulting in largely overlapping curves. We note that this comparison is mainly for illustrative purpose and it may be that different evolutionary scenarios change the behavior of the test statistics. A full color version of this figure is available at the *Heredity* journal online.

continued selection over multiple generations may validate even small functional differences between alleles.

Although gene expression analysis has also been used previously to validate E&R results, we caution that in the absence of a good understanding of how expression differences could affect phenotypes, the interpretation of expression data may be too complex to serve as a stringent validation of candidate genes/SNPs.

Finally, we end this section with a cautionary note on the ability to functionally validate candidate SNPs identified in E&R studies. One implicit assumption is that a larger number of generations will increase the power to detect functionally important loci, but qualitatively similar results are obtained independently of the generation at which the tests are being performed (ignoring variation in selection coefficients among loci). Nevertheless, a recent trajectory analysis found almost entirely different sets of candidates depending on whether generation 15 or 37 was compared with the starting population (Orozco-terWengel *et al.*, 2012). The reason for this surprising observation is that some alleles increased rapidly early on, but then did not change thereafter, resulting in a frequency plateau between generations 15 and 37. In contrast, many other alleles increased more slowly, but continuously, achieving a higher frequency change by generation 37 than the plateauing alleles. Hence, we suggest that functional validations may benefit from the inclusion of the selection trajectories of candidate loci when comparing E&R results with GWAS or linkage mapping studies.

E&R unlimited?

We anticipate that as our ability to reliably interpret E&R results continues to improve, there will be an increasing number of studies that will apply this method to a substantially broader range of taxa and

species. So far, most E&R studies focused on *Drosophila*, but this approach has also been successfully applied to species with longer generation times and smaller population sizes (Johansson *et al.*, 2010; Rubin *et al.*, 2010; Kelly *et al.*, 2013; Beissinger *et al.*, 2014). Until now, E&R has been used for rather simple research questions, largely concerned with linking genotypes with phenotypes. But, given the inherent flexibility of experimental evolution framework (Kawecki *et al.*, 2012), we anticipate a broader use of E&R studies in the future. This could include investigating the impact of migration, different combinations of selective environmental conditions, fluctuating environments or the influence of genetic composition of the starting population, among many others.

E&R studies may be further expanded to study not only the response of the host genome, but also the dynamics of pathogens or endosymbionts during the experiment. One nice example for the potential of this approach comes from the analysis of *Wolbachia* strains in an experiment that was designed to identify the genomic response of adaptation of *D. melanogaster* to novel environments (Orozco-terWengel *et al.*, 2012; Tobler *et al.*, 2013). After exposing *D. melanogaster* and its *Wolbachia* endosymbiont to two different temperature regimes, temperature-dependent differences in the dynamics of *Wolbachia* strains were uncovered. Although the frequencies of three different *Wolbachia* clades remained stable in the hot environment, one clade increased from ~25% to ~80% in <15 generations in the cold environment (Versace *et al.*, 2014).

The E&R approach can also be extended beyond DNA polymorphism to incorporate gene expression levels (Yampolsky *et al.*, 2012; Hollis *et al.*, 2014). Contrasting gene expression levels of differentially selected populations provides a powerful and complementary approach to elucidate the selective response. A good example comes from a recent study that compared gene expression response in experimental *D. melanogaster* populations where either a monogamous or polygamous mating system was enforced over 65 generations (Hollis *et al.*, 2014). Theory suggests that genes showing sex-biased expression levels due to sexually antagonistic selection—that is, genes affecting traits with different fitness optima between the sexes—should evolve female-like expression levels in a monogamous system, where selection on male traits is relaxed. The experimental results fit this expectation, showing that genes known to have sex-biased expression had feminized their expression in the monogamous, but not the polygamous, population by the end of the study.

Finally, we anticipate that the combination of allele frequency changes with gene expression dynamics will not only assist in the identification of causative variants, but will also provide a more complete picture of the selected trait, in particular when combined with time series analyses.

DATA ARCHIVING

There were no data to deposit.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We are thankful to members of the Institut für Populationsgenetik, in particular A Betancourt, for continuous feedback and discussion. A Betancourt, A Kalinka, C Kosiol and S Nuzhdin provided helpful feedback on earlier versions of this manuscript. This work has been supported by grants of the Austrian Science Funds (FWF, W1225) and the ERC (ArchAdapt).

GLOSSARY

Candidate SNP: single-nucleotide polymorphism that shows a significant response to selection.

Chromosomal inversion: a chromosomal rearrangement that produces a region with a reversed orientation in different individuals. Chromosomal inversions typically suppress recombination between inverted and noninverted chromosomes. In E&R experiments and other mapping studies, chromosomal inversions often hinder fine mapping of causative variants.

Evolve and resequence (E&R): approach in experimental evolution that measures genome-wide allele frequency changes in an experimental evolution study through Pool-Seq.

Experimental evolution: experimental approach in which the phenotypic and/or genomic change is monitored over subsequent generations during which experimenters apply a predetermined selective pressure under controlled demographic conditions.

Genome-wide association study (GWAS): a trait mapping approach relying on a statistical test for an association between sequence variants and a given phenotype.

Laboratory natural selection: experimental evolution method in which experimenters impose a selection regime and differential reproductive success of individuals results in a selective response. The other widely used method is truncating selection.

Linkage disequilibrium (LD): nonrandom association between alleles of two loci in a population.

Linkage mapping: framework to test for statistical associations between genotype and phenotype based on either F2 individuals or backcrosses. Although initially only a few markers were used, Rad-Tag sequencing combines a high marker density with the ability to analyze a moderate number of individuals. Sometimes also referred to as quantitative trait locus mapping.

Long-range linkage disequilibrium: linkage disequilibrium between distantly located sites on the same chromosome.

Manhattan plot: a scatterplot used in genome surveys to display the *P*-values of individual SNPs (*y* axis). The *x* axis indicates the chromosomal position of the corresponding SNP. Significant loci often show a characteristic chimney structure because of partial association of neutral linked SNPs.

Narrow-range linkage disequilibrium: linkage disequilibrium between proximate sites caused by low recombination rates. In natural *Drosophila melanogaster* populations, narrow-range LD is well below 1 kb.

Pool-sequencing (Pool-Seq): sequencing method in which either DNA is extracted from multiple individuals at once or where DNA from multiple individuals is pooled before preparing the sequencing library.

Reverse selection: selection toward an ancestral state. As long as initially favored alleles are not fixed, the alternate allele is expected to increase in frequency if reverse selection is imposed.

Receiver operator characteristic curve: graphical representation of the sensitivity (true positives on the *y* axis) and specificity (false positives on the *x* axis) for a given test statistic. Receiver operator characteristic curves are used to evaluate the performance of classifiers. The best possible method, detects all true positives without incurring any false positives and yields a curve leading through the most upper left corner.

Sequence coverage: the number of reads that align to a specific genomic region.

Sexually antagonistic selection: selection in which males and females with different fitness optima produce a conflict between the sexes.

Standing genetic variation: all polymorphisms segregating in a population. Thereby, adaptation from standing genetic variation emphasizes that selection is operating on alleles present at the onset of the experiment. Selection for variants present in a population at intermediate/high frequencies generates a genomic signature (that is, soft sweep) that differs from when selection acts on a new mutation (that is, hard sweep).

Starting population: also called base population, the population that initiates the selection regime at generation 0. In *Drosophila*, starting populations are frequently derived from isofemale lines from which either multiple individuals are used directly, or which have been kept at a large population size to allow for recombination among them before the beginning of the experiment. The

number of individuals and the level of linkage disequilibrium in the starting population are key factors determining the success of an E&R study.

Truncating selection: selection regime under which the experimenter selects in each generation for phenotypes that exceed a certain phenotypic threshold (for example, all individuals above a certain size) and only those contribute to the next generation. It is a frequently used approach in E&R.

Window-based approach: analysis based on contiguous genomic regions (windows). Compared with the analysis of single nucleotides, window-based approaches have a lower variance because test statistics of interest are calculated across all sites within the window. As the outcome depends on the window size, window-based tests are better suited for exploratory analyses.

Aulard S, Monti L, Chaminade N, Lemeunier F (2004). Mitotic and polytene chromosomes: comparisons between *Drosophila melanogaster* and *Drosophila simulans*. *Genetica* **120**: 137–150.

Baldwin-Brown JG, Long AD, Thornton KR (2014). The power to detect quantitative trait loci using resequenced, experimentally evolved populations of diploid, sexual organisms. *Mol Biol Evol* **31**: 1040–1055.

Beissinger TM, Hirsch CN, Vaillancourt B, Deshpande S, Barry K, Buell CR *et al.* (2014). A genome-wide scan for evidence of selection in a maize population under long-term artificial selection for ear number. *Genetics* **196**: 829–840.

Burke MK, Dunham JP, Shahrestani P, Thornton KR, Rose MR, Long AD (2010). Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature* **467**: 587–590.

Chevin LM, Hospital F (2008). Selective sweep at a quantitative trait locus in the presence of background genetic variation. *Genetics* **180**: 1645–1660.

Cutter AD, Baird SE, Charlesworth D (2006). High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*. *Genetics* **174**: 901–913.

Elena SF, Lenski RE (2003). Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat Rev Genet* **4**: 457–469.

Futschik A, Schlötterer C (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* **186**: 207–218.

Garland T, Rose MR (2009). *Experimental Evolution: Concepts, Methods, and Applications of Selection Experiments*. University of California Press: Berkeley.

Hollis B, Houle D, Yan Z, Kawecki TJ, Keller L (2014). Evolution under monogamy feminizes gene expression in *Drosophila melanogaster*. *Nat Commun* **5**: 3482.

Hu TT, Eisen MB, Thornton KR, Andolfatto P (2013). A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. *Genome Res* **23**: 89–98.

Illingworth CJ, Parts L, Schiffels S, Liti G, Mustonen V (2012). Quantifying selection acting on a complex trait using allele frequency time series data. *Mol Biol Evol* **29**: 1187–1197.

Jalvingh KM, Chang PL, Nuzhdin SV, Wertheim B (2014). Genomic changes under rapid evolution: selection for parasitoid resistance. *Proc Biol Sci R Soc* **281**: 20132303.

Johansson AM, Pettersson ME, Siegel PB, Carlborg O (2010). Genome-wide effects of long-term divergent selection. *PLoS Genet* **6**: e1001188.

Kapun M, van Schalkwyk H, McAllister B, Flatt T, Schlötterer C (2014). Inference of chromosomal inversion dynamics from Pool-Seq data in natural and laboratory populations of *Drosophila melanogaster*. *Mol Ecol* **23**: 1813–1827.

Kawecki TJ, Lenski RE, Ebert D, Hollis B, Olivieri I, Whitlock MC (2012). Experimental evolution. *Trends Ecol Evol* **27**: 547–560.

Kelly JK, Koseva B, Mojica JP (2013). The genomic signal of partial sweeps in *Mimulus guttatus*. *Genome Biol Evol* **5**: 1457–1469.

Kessner D, Novembre J (2014a). forqs: forward-in-time simulation of recombination, quantitative traits and selection. *Bioinformatics* **30**: 576–577.

Kessner D, Novembre J (2014b). Power analysis of artificial selection experiments using efficient whole genome simulation of quantitative traits. *bioRxiv* http://dx.doi.org/10.1101/005892.

Kessner D, Turner TL, Novembre J (2013). Maximum likelihood estimation of frequencies of known haplotypes from pooled sequence data. *Mol Biol Evol* **30**: 1145–1158.

Kofler R, Schlötterer C (2014). A guide for the design of evolve and resequencing studies. *Mol Biol Evol* **31**: 474–483.

Liu W, Yuan JS, Stewart CN Jr. (2013). Advanced genetic tools for plant biotechnology. *Nat Rev Genet* **14**: 781–793.

Magwire MM, Fabian DK, Schweyen H, Cao C, Longdon B, Bayer F *et al.* (2012). Genome-wide association studies reveal a simple genetic basis of resistance to naturally coevolving viruses in *Drosophila melanogaster*. *PLoS Genet* **8**: e1003057.

Martins N, Faria V, Nolte V, Schlötterer C, Teixeira L, Sucena E *et al.* (2014). Host adaptation to viruses relies on few genes with different cross-resistance properties. *Proc Natl Acad Sci USA* **111**: 5938–5943.

Nuzhdin SV, Keightley PD, Pasyukova EG (1993). The use of retrotransposons as markers for mapping genes responsible for fitness differences between related *Drosophila melanogaster* strains. *Genet Res* **62**: 125–131.

Nuzhdin SV, Turner TL (2013). Promises and limitations of hitchhiking mapping. *Curr Opin Genet Dev* **23**: 694–699.

Orozco-terWengel P, Kapun M, Nolte V, Kofler R, Flatt T, Schlötterer C (2012). Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Mol Ecol* **21**: 4931–4941.

Palmieri N, Nolte V, Chen J, Schlötterer C (2014). Genome assembly and annotation of a *Drosophila simulans* strain from Madagascar. *Mol Ecol Resour* (e-pub ahead of print 24 June 2014; doi:10.1111/1755-0998.12297).

Parts L, Cubillos FA, Warringer J, Jain K, Salinas F, Bumpstead SJ *et al.* (2011). Revealing the genetic structure of a trait by sequencing a population under selection. *Genome Res* **21**: 1131–1138.

Rellstab C, Zoller S, Tedder A, Gugerli F, Fischer MC (2013). Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. *PLoS ONE* **8**: e80422.

Remolina SC, Chang PL, Leips J, Nuzhdin SV, Hughes KA (2012). Genomic basis of aging and life-history evolution in *Drosophila melanogaster*. *Evolution* **66**: 3390–3403.

Ronen R, Udpa N, Halperin E, Bafna V (2013). Learning natural selection from the site frequency spectrum. *Genetics* **195**: 181–193.

Rubin CJ, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT *et al.* (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**: 587–591.

Schlötterer C, Tobler R, Kofler R, Nolte V (2014). Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. *Nature Rev. Genet.* http://dx.doi.org/10.1038/nrg3803.

Smith JM (1968). "Haldane's dilemma" and the rate of evolution. *Nature* **219**: 1114–1116.

Teotonio H, Carvalho S, Manoel D, Roque M, Chelo IM (2012). Evolution of outcrossing in experimental populations of *Caenorhabditis elegans*. *PLoS ONE* **7**: e35811.

Teotonio H, Chelo IM, Bradic M, Rose MR, Long AD (2009). Experimental evolution reveals natural selection on standing genetic variation. *Nat Genet* **41**: 251–257.

Terhorst J, Song YS (2014). Multi-locus analysis of genomic time series data from experimental evolution. *bioRxiv* http://dx.doi.org/10.1101/006734.

Terns RM, Terns MP (2014). CRISPR-based technologies: prokaryotic defense weapons repurposed. *Trends Genet* **30**: 111–118.

Tobler R, Franssen SU, Kofler R, Orozco-terWengel P, Nolte V, Hermisson J *et al.* (2013). Massive habitat-specific genomic response in D. melanogaster populations during experimental evolution in hot and cold environments. *Mol Biol Evol* **31**: 364–375.

Topa H, Jónás Á, Kofler R, Kosiol C, Honkela A (2014). Gaussian process test for high-throughput sequencing time series: application to experimental evolution. Preprint at http://arXiv.org/1403.4086.

Turner TL, Miller PM (2012). Investigating natural variation in *Drosophila* courtship song by the evolve and resequence approach. *Genetics* **191**: 633–642.

Turner TL, Miller PM, Cochrane VA (2013). Combining genome-wide methods to investigate the genetic complexity of courtship song variation in *Drosophila melanogaster*. *Mol Biol Evol* **30**: 2113–2120.

Turner TL, Stewart AD, Fields AT, Rice WR, Tarone AM (2011). Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. *PLoS Genet* **7**: e1001336.

Versace E, Nolte V, Pandey RV, Tobler R, Schlötterer C (2014). Experimental evolution reveals habitat-specific fitness dynamics among *Wolbachia* clades in *Drosophila melanogaster*. *Mol Ecol* **23**: 802–814.

Weber KE (1996). Large genetic change at small fitness cost in large populations of *Drosophila melanogaster* selected for wind tunnel flight: rethinking fitness surfaces. *Genetics* **144**: 205–213.

Yampolsky LY, Glazko GV, Fry JD (2012). Evolution of gene expression and expression plasticity in long-term experimental populations of *Drosophila melanogaster* maintained under constant and variable ethanol stress. *Mol Ecol* **21**: 4287–4299.

Zhou D, Udpa N, Gersten M, Visk DW, Bashir A, Xue J *et al.* (2011). Experimental selection of hypoxia-tolerant *Drosophila melanogaster*. *Proc Natl Acad Sci USA* **108**: 2349–2354.