Optimizing the Power to Identify the Genetic Basis of Complex Traits with Evolve and Resequence Studies

Christos Vlachos[1,2] and Robert Kofler *,[1]

[1]Institute für Populationsgenetik, Vetmeduni Vienna, Wien, Austria

[2]Vienna Graduate School of Population Genetics, Wien, Austria

***Corresponding author:** E-mail: rokofler@gmail.com.

Associate editor: John True

Abstract

Evolve and resequence (E&R) studies are frequently used to dissect the genetic basis of quantitative traits. By subjecting a population to truncating selection for several generations and estimating the allele frequency differences between selected and nonselected populations using next-generation sequencing (NGS), the loci contributing to the selected trait may be identified. The role of different parameters, such as, the population size or the number of replicate populations has been examined in previous works. However, the influence of the selection regime, that is the strength of truncating selection during the experiment, remains little explored. Using whole genome, individual based forward simulations of E&R studies, we found that the power to identify the causative alleles may be maximized by gradually increasing the strength of truncating selection during the experiment. Notably, such an optimal selection regime comes at no or little additional cost in terms of sequencing effort and experimental time. Interestingly, we also found that a selection regime which optimizes the power to identify the causative loci is not necessarily identical to a regime that maximizes the phenotypic response. Finally, our simulations suggest that an E&R study with an optimized selection regime may have a higher power to identify the genetic basis of quantitative traits than a genome-wide association study, highlighting that E&R is a powerful approach for finding the loci underlying complex traits.

Key words: **experimental evolution, quantitative genetics, GWAS, population genetics, forward simulations, evolve and resequence.**

Introduction

Most variation of traits important in agriculture, medicine, ecology, and evolution is quantitative (Mackay 2001). Variation in such quantitative traits (or complex traits) is usually due to multiple segregating loci (Mackay 2001). For these quantitative traits the simple Mendelian correspondence between genotype and phenotype breaks down, such that one particular phenotype may be due to several distinct genotypes (Lander and Schork 1994). Unraveling the genetic basis of quantitative traits will be crucial for improving crop yield, leveraging personalized medicine and shedding light on poorly understood evolutionary processes such as extinctions, rapid adaptation, and canalization. It has even been argued that identifying the genetic basis of quantitative traits will be the key challenge for biology in the 21st century (Mackay 2001; Stapley et al. 2010; Losos et al. 2013).

Due to this wide interest many approaches for identifying the genetic basis of complex traits have been developed, such as quantitative trait locus (QTL) mapping and genome-wide association studies (GWAS) (Mackay 2001; Korte and Farlow 2013). These methods suffer from some limitations. QTL studies only capture a limited amount of the variation present in natural populations and the resolution of QTL studies is usually low (Mackay 2001). Identifying the causative quantitative trait nucleotide (QTN) is thus rarely achieved with QTL studies (Rockman 2012). GWAS however capture more natural variation and have a higher resolution than QTL studies, frequently enabling the identification of some QTNs. GWAS achieve this high resolution by utilizing historical recombination events rather than recombination events occurring within QTL mapping populations (Mackay et al. 2009). However, GWAS have also some limitations. With GWAS it is difficult to identify rare variants and variants of small effect size (Marchini et al. 2004; Korte and Farlow 2013). Hence alternative approaches for identifying the QTNs are of wide interest.

The advent of NGS made it feasible to monitor adaptation at the genomic level with an approach termed Evolve and resequence (E&R [Long et al. 2015; Schlötterer et al. 2015]). A base population, that usually captures a substantial amount of the variation of a natural population, is subject to some selective pressure over multiple generations and allele frequency changes are monitored by sequencing the experimental populations. The selective pressure may be either natural, when a population is exposed to a defined environment, or artificial, when a specific phenotype is selected (Garland and Rose 2009; Schlötterer et al. 2015). Since the selected traits are usually not known with natural adaptation, E&R studies relying on natural selection are mostly used to study the dynamics of adaptation rather than the genetic basis of complex traits. However E&R studies with artificial selection may be a powerful approach for identifying the QTNs, especially since E&R studies rely on both, historical recombination events (base population) and recombination events occurring during the experiment. Using computer simulations several theoretical studies found that E&R studies have sufficient power to identify selected loci, provided that a powerful experimental design is used (Baldwin-Brown et al. 2014; Kofler and Schlötterer 2014; Kessner and Novembre 2015). Such a powerful design usually requires large population sizes (>1,000), several replicates (>5), and multiple generations of selection (>90). Hence E&R studies are mostly suitable for small organisms having short generation times such as fruit flies (Turner et al. 2011; Orozco-Terwengel et al. 2012; Turner and Miller 2012; Tobler et al. 2014), nematodes (Teotonio et al. 2012), yeast (Kosheleva and Desai 2018), bacteria (Wannier et al. 2018), and mice (Keightley and Bulfield 1993). But even for model organisms, E&R studies come at a considerable cost in terms of time and sequencing effort. It is thus important to ensure that the invested resources are optimally utilized by maximizing the power to identify the QTNs. One promising but little explored approach for increasing the performance of E&R studies is the selection regime, that is the number of selected individuals during the experiment. This approach is especially promising as an optimized selection regime comes at no, or only little, additional cost (time, sequencing, and phenotyping). As a major challenge, an optimal selection regime needs to strike a balance between too weak and too strong selection. With weak selection it will be difficult to distinguish QTNs from neutral loci subject to genetic drift, whereas with strong selection the QTNs may not be distinguished from vast amounts of hitchhikers.

Using genome-wide forward simulations of populations under truncating selection we show that the performance of E&R studies may be maximized by gradually increasing the strength of selection during the experiment, that is decreasing the number of selected individuals with time. This approach reduces hitchhiking associated with strongly selected loci but nevertheless generates a noticeable response of weakly selected loci. Interestingly, we found that an E&R study with an optimized selection regime may have a higher power to identify QTNs than a GWAS involving several thousands of individuals. A suboptimal selection regime, such as constant strong selection, however results in a poor performance. Our results highlight that the selection regime is a crucial factor determining the success of E&R studies.

Results

To test if the selection regime has an influence on the performance of E&R studies we performed genome-wide forward simulations with MimicrEE2 (Vlachos and Kofler 2018). MimicrEE2 is a versatile tool that allows the simulation of temporally variable truncating selection with a quantitative trait. We aimed to capture the genomic landscape of *Drosophila melanogaster*, a commonly used model organism for E&R studies (Long et al. 2015; Schlötterer et al. 2015). We used the recombination rate estimates of Comeron et al. (2012) for windows of 100 kb and a base population consisting of 1,000 diploid and homozygous genomes that reproduce the pattern of natural variation found in a *D. melanogaster* population from Vienna (Bastide et al. 2013; Kofler and Schlötterer 2014) (supplementary fig. 1, Supplementary Material online). Simulations were performed for the major autosomes, where low recombining regions were excluded as they inflate the false-positive rate (FPR) (Kofler and Schlötterer 2014) (supplementary fig. 1, Supplementary Material online). Based on the recommendations of Kofler and Schlötterer (2014) we simulated an E&R study with a population size of 1,000, 10 replicates and 90 generations of selection (table 1). We simulated a quantitative trait model, where 100 randomly selected loci contribute to a trait (100 QTNs). Only loci with frequencies between 5% and 95% were selected, which ensures that all selected single nucleotide polymorphisms (SNPs) contribute at least moderately to the genotypic variance (Falconer and Mackay 1960).

Table 1. Overview of the Default Parameters Used for the Simulations.

Parameter	Default Value
Population size (N)	1,000
Number of causative loci	100
Number of generations	90
Replicates	10
Distribution of effect sizes	Gamma with shape = 0.42 and scale = 1
Heritability	1 (genotype to phenotype mapping = 1:1)
Recombination map	Comeron et al. (2012)
Repetitions	10 (using different causative loci and effect sizes)

The effect sizes of the QTNs followed a gamma distribution that captures the distribution of effect sizes found with QTL studies (Hayes and Goddard 2001; Meuwissen et al. 2001) (table 1). We thus simulated quantitative traits with few large effect loci and many weak effect loci. The sign of the effect (a vs. $-a$) was randomly chosen and a heritability of $h^2 = 1$ was initially used. Simulations for each experimental design were repeated ten times using different sets of randomly drawn QTNs (random position and effect size) (table 1). The significance of the allele frequency differences between the base population and the evolved populations was estimated with the Cochran–Mantel–Haenszel (CMH) test (Landis et al. 1978). This test takes replicates into account and has a good performance with E&R studies (Kofler and Schlötterer 2014; Schlötterer et al. 2015). The power of the different selection regimes was assessed using Receiver Operating Characteristic (ROC) curves (Hastie et al. 2009), which relate the true-positive rate (TPR) to the FPR). The TPR can be calculated as TP/(TP + FN), where TP stands for true positives and FN for false negatives. The FPR can be calculated as FP/(TN + FP), where FP refers to false positives and to TN true negatives. A ROC curve having a TPR of 1.0 with a FPR of 0.0 indicates the best possible performance. We displayed average ROC curves based on the ten different sets of QTNs. As we are mostly interested in identifying QTNs at a low FPR we used a FPR threshold of 0.01 and computed the area under the partial

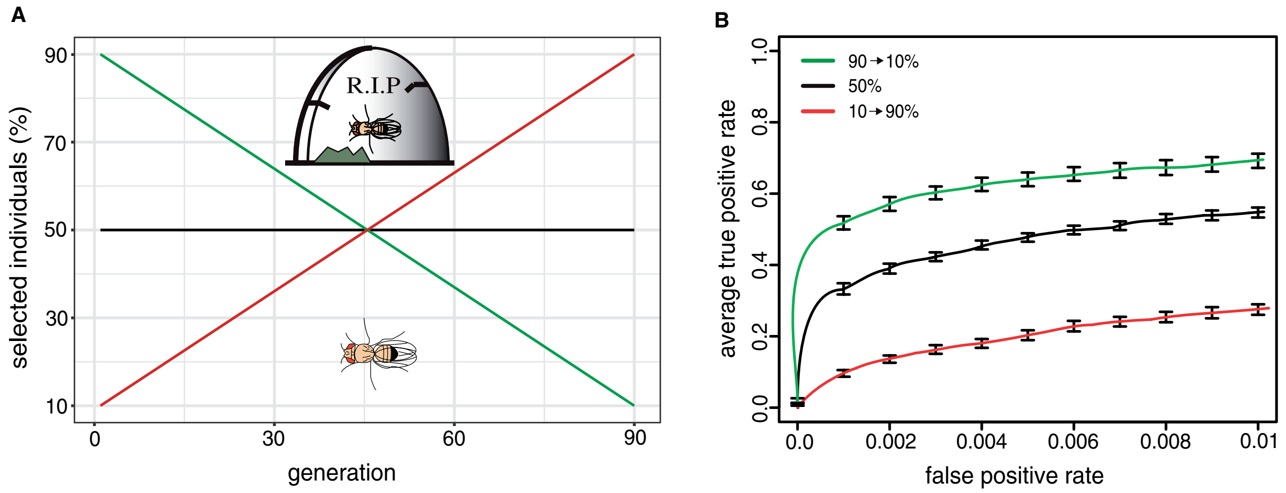

Fig. 1. The selection regime has a significant influence on the performance of E&R studies. (*A*) We simulated three truncating selection regimes. The strength of selection increased (green), remained constant (black), or decreased (red) during the experiment. Note that the total number of selected individuals is identical for the three selection regimes. (*B*) ROC curves showing the performance of the three selection regimes. The increasing selection regime has the best performance.

ROC curve (pAUC $= \int_0^{0.01} \mathrm{ROC}(f)\mathrm{d}f$) to assess the performance of a selection regime. For an overview of the simulation pipeline see supplementary figure 2, Supplementary Material online.

The Selection Regime Influences the Performance of E&R Studies

We first tested the hypothesis that selection regimes have a significant influence on the performance of E&R studies. We generated different selection regimes by varying the strength of truncating selection throughout the experiment. With truncating selection the individuals having the most pronounced phenotypes are selected and allowed to mate. The offspring of the selected individuals will constitute the next generation. We evaluated the performance of three different selection regimes: 1) constant strength of selection, where 50% of the individuals are selected at each generation (50%); 2) linearly increasing strength of selection, where 90% of the individuals are selected at the beginning of the experiment and 10% at the end (90 → 10%); and 3) linearly decreasing strength of selection, where 10% of the individuals are selected at the beginning of the experiment and 90% at the end (10 → 90%, fig. 1*A*). Note that the sum of individuals selected over the 90 generations is identical for the three selection regimes.

We found that the selection regime has a significant influence on the power to identify the causative loci (Kruskal–Wallis rank-sum test with pAUC; $P = 2.8e{-}06$) where linearly increasing strength of selection (henceforth "increasing regime") had the best performance (fig. 1*B*). A constant strength of selection had an intermediate performance (henceforth "constant regime") and decreasing the strength of selection (henceforth "decreasing regime") had the worst performance (fig. 1*B*).

This raises the question why the increasing regime performed better than the constant and the decreasing regime. Apart from technical problems (e.g. sequencing) an E&R study has two sources of noise: genetic drift and hitchhiking of neutral alleles linked to selected loci. If selection is weak, allele frequency changes of neutral loci subject to genetic drift may be more pronounced than the response to selection of the QTNs. However if selection is strong, alleles linked to selected loci will have little opportunity to recombine to neutral haplotypes. These hitchhikers will thus show a significant response to selection. An optimal selection regime must thus aim to minimize both sources of noise, hitchhiking, and drift.

To identify possible causes for the performance differences among selection regimes we investigated the trajectories of strong (effect size >1), weak (effect size ≤1), and neutral alleles in a single replicate of each selection regime (supplementary fig. 3, Supplementary Material online). Most selected loci got fixed (frequency of 1.0, polarized to selected allele) in all three selection regimes, but fixation of selected alleles appears to be delayed in the increasing regime relative to the other regimes (supplementary fig. 3, Supplementary Material online). Delayed fixation of selected alleles could lead to fewer hitchhikers and thus account for the performance differences among selection regimes (supplementary fig. 4, Supplementary Material online). To test this hypothesis we quantified the response to selection of strong and weak effect loci in the selection regimes using 50 simulations with different random sets of QTNs (50 sets of QTNs; 1 replicate; supplementary fig. 5, Supplementary Material online). Fixation of both strong and weak effect loci was significantly delayed in the increasing regime compared with the constant and the decreasing regime (Wilcoxon rank-sum test; $P < 2.2e{-}16$). Furthermore, fixation of strong effect loci was more delayed than fixation of weak effect loci (strong: incr./cons. = 1.9×, incr./decr. = 5.2×; weak: incr./cons. = 1.4×, incr./decr. = 3.3×; supplementary fig. 5, Supplementary Material online). This suggests that the increasing regime affords more time to neutral alleles to recombine away from selected haplotypes.

Next, we aimed to quantify the extent of hitchhiking for the three selection regimes. Replicated E&R studies enable to roughly distinguish between hitchhikers and alleles subject to genetic drift by the consistency of the allele frequency change among replicates. For example, alleles that increase in frequency in some replicates but decrease in frequency in others are likely subject to drift, whereas neutral alleles that increase in frequency in all replicates are likely hitchhikers. We thus classified loci as hitchhikers when the allele frequency consistently changed in the same direction in all 10 replicates (excluding QTNs). In agreement with our hypothesis we found fewer hitchhikers in the increasing regime than in the other two regimes (10 set of QTNs; 10 replicates; supplementary fig. 6, Supplementary Material online). Furthermore, the allele frequency change of hitchhikers was least pronounced in the increasing regime (supplementary fig. 6, Supplementary Material online). However, most loci subject to genetic drift (consistent allele frequency change in 4–6 replicates) were found for the increasing regime, but the allele frequency change due to drift was least pronounced for the constant regime (supplementary fig. 6, Supplementary Material online). Of course, constant weak selection where for example 90% of the individuals are selected may reduce hitchhiking even more than the tested increasing regime (supplementary fig. 5, Supplementary Material online). However, constant weak selection results in an overall reduced response to selection (supplementary fig. 5, Supplementary Material online) such that the noise generated by genetic drift may dominate the weak signal of selected loci. In summary, we propose that the increasing regime has a high performance because it initially delays the fixation of strong effect alleles, allowing hitchhikers to recombine out of selected haplotypes, but amplifies the response of weak effect loci at the end of the experiment.

So far, we have evaluated the performance of three different selection regimes having an identical total number of selected individuals. However many more different selection regimes, with varying amounts of selected individuals are feasible. We thus carried out additional simulations to test if increasing regimes outperform other selection regimes.

Linearly Increasing the Strength of Selection Maximizes the Power to Identify QTNs

Among all feasible linear selection regimes (increasing, constant, and decreasing) we aimed to identify the regime that results in the highest power to identify the QTNs. Since genome-wide forward simulations are computationally demanding we needed to limit the number of necessary simulations. As the decreasing regime had a poor performance we solely considered increasing and constant regimes (fig. 1; supplementary figs. 5 and 6, Supplementary Material online). Furthermore, we evaluated the performance of selection regimes in steps of 10% selected individuals (fig. 2A and B). To identify the best increasing regime we thus evaluated the performance of 36 different regimes (90 → 80%, 90 → 70%, . . ., 90 → 10%, . . ., 20 → 10%; fig. 2A, left panel). Based on the area under the partial ROC curve (pAUC) we found that the increasing regime where 90% of the individuals are selected at the beginning of the experiment and 20% at the end had the best performance (90 → 20% increasing regime; fig. 2A). Out of the constant regimes, however, selection of 80% of the individuals resulted in the highest power to identify the QTNs (fig. 2B). This is in agreement with the results of Kessner and Novembre (2015) who evaluated the performance of multiple constant regimes (20%, 40%, 60%, and 80%) under a QTL model with truncating selection and found that selection of 80% of individuals performed best. However, we found that the best increasing regime had a higher power to identify QTNs than any of the evaluated constant regimes (fig. 2B; Wilcoxon rank-sum test with pAUC; 90 → 20% vs. each constant regime; $P \leq 0.0002$).

Influence of the Experimental Design

Depending on the experimental organism, the default of 90 generations of selection may be quite time consuming (e.g. 3 years in *Drosophila*). We thus asked whether a high performance may also be achieved with shorter experiments if an optimized selection regime is used. We evaluated the performance of different selection regimes for 20, 45, and 90 generations of selection (fig. 3). Simulations were performed for all 45 combinations of increasing and constant regimes (9 constant and 36 increasing regimes, as shown in fig. 2A and C).

In agreement with previous works we found that the performance of E&R studies increases with the number of generations of selection (Baldwin-Brown et al. 2014; Kofler and Schlötterer 2014; Kessner and Novembre 2015) (fig. 3). Of the constant regimes, selection of 80% of the individuals consistently had the best performance, irrespective of the length of the experiment (fig. 3). This is again in agreement with Kessner and Novembre (2015). With 90 and 45 generations of selection the best increasing regime significantly outperformed the constant regimes (Wilcoxon rank-sum test with pAUC, $p_{45} = 0.014$ for 16% increased performance; $p_{90} = 0.0002$ for 22.4% increased performance). With 20 generations of selection, however, the performance of the best increasing and constant regime was quite similar (Wilcoxon rank-sum test with pAUC; $p_{20} = 0.68$ for 2% increased performance). We also noticed that the optimal increasing regime changed from 90 → 20% with 90 generations of selection to 90 → 70% with 20 generations. With 45 generations of selection the performance of the 80 → 10% increasing regime was not significantly different from the 90 → 20% increasing regime (Wilcoxon rank-sum test, $P = 1$). For short experiments the performance of the best increasing regime thus approaches the performance of the best constant regime. Short experiments may not provide sufficient time for the benefit of increasing regimes, such as the delayed fixation of large effect loci, to take effect and thus influence the performance. We conclude that an optimized selection regime is not able to compensate for the loss of performance incurred by reducing the generations of selection. In fact the advantage of an optimized selection regime, such as the increasing regime, is most pronounced for long experiments.

Next, we assessed the effect of the number of replicate populations on the shape of the optimal selection regime. We evaluated the performance of the different selection regimes with 3, 5, and 10 replicates (10 is the default). Our results

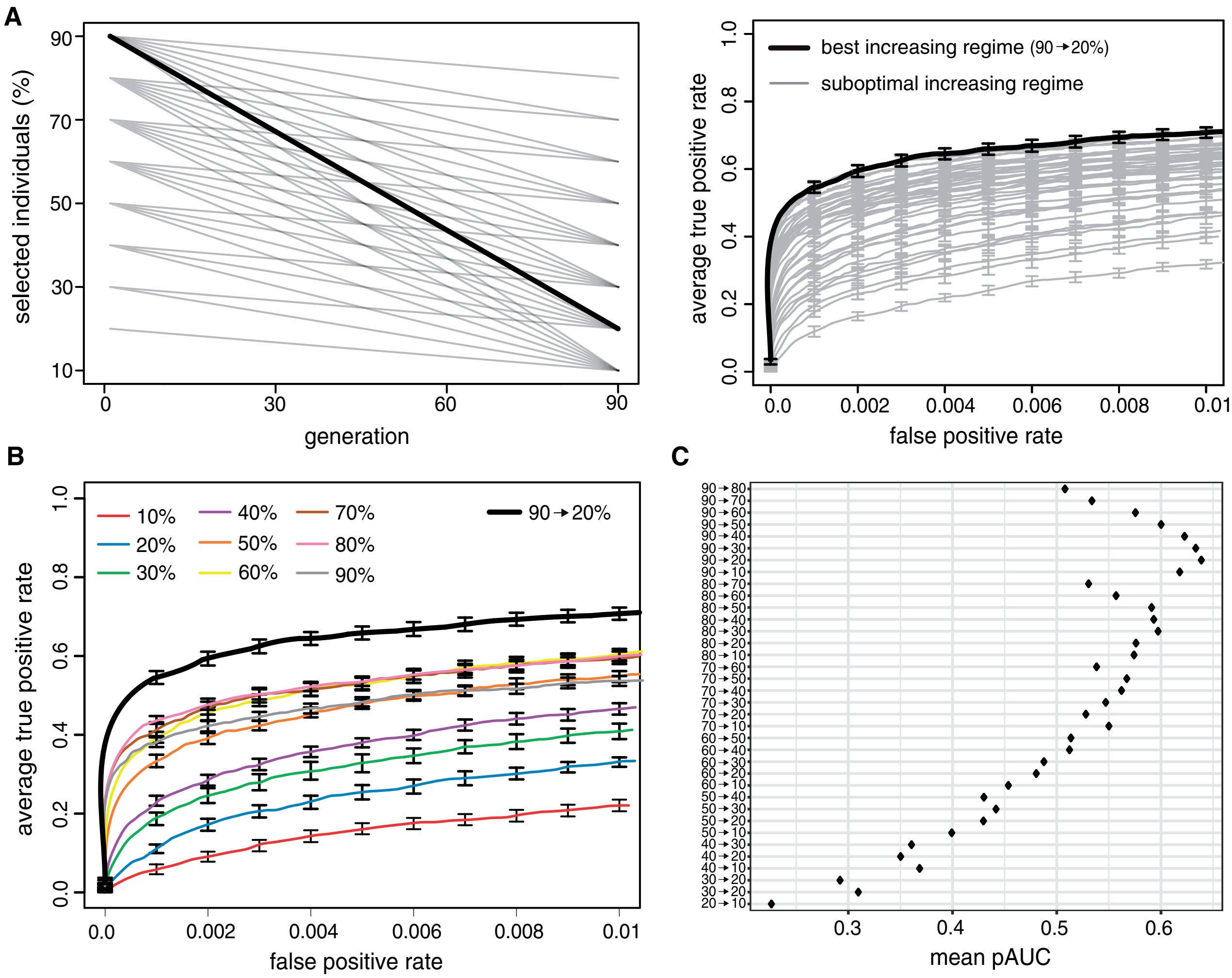

FIG. 2. Increasing the strength of selection during an E&R study enhances the power to identify QTNs. (*A*) Approach for identifying the best increasing regime. We evaluated the performance of all possible increasing regimes using steps of 10% (left panel). The best selection regime (bold) has the largest AUC (right panel). (*B*) The best increasing regime outperforms the evaluated constant regimes (steps of 10%). (*C*) Performance of each evaluated increasing regime (pAUC).

suggest that the best increasing regime significantly outperforms the best constant regime when 10 or 5 replicates are used (Wilcoxon rank-sum test with pAUC; $p_5 = 0.0004$ for 16% increased performance; $p_{10} = 0.0002$ for 22.4% increased performance; supplementary fig. 7, Supplementary Material online). For 3 replicates the performance of the best increasing regime is not significantly different from the best constant regime (Wilcoxon rank-sum test with pAUC; $p_3 = 0.5$ for 4.9% increased performance). The advantage of an increasing regime is thus most pronounced when many replicates are used. Interestingly, also the shape of the optimal selection regime depends on the number of replicates, where strong selection at the end of the experiment is especially beneficial when many replicates are used (supplementary fig. 7, Supplementary Material online).

This raises the question why replication influences the shape of the optimal selection regime. For loci mostly subject to genetic drift the direction of the allele frequency change will vary across replicates. As high CMH-scores require consistent allele frequency changes across replicates, it will be easier to distinguish between selected loci and loci subject to drift when many replicates are used. An elevated strength of selection at the end of highly replicated E&R studies may thus boost the response to selection of weak effect loci without incurring excessive additional noise from genetic drift caused by the population size reduction at the end of the experiment.

Finally, we investigated the influence of the population size (supplementary fig. 8, Supplementary Material online). We found that an increasing regime consistently performed better than the constant regimes where especially the 90 → 20% increasing regime had a high performance with the evaluated population sizes (supplementary fig. 8, Supplementary Material online).

To summarize, in all tested experimental designs an increasing regime outperformed the constant regimes. However, the slope and intercept of the optimal increasing regime depends on the experimental design where especially the number of replicates and the length of the experiment had a noticeable influence.

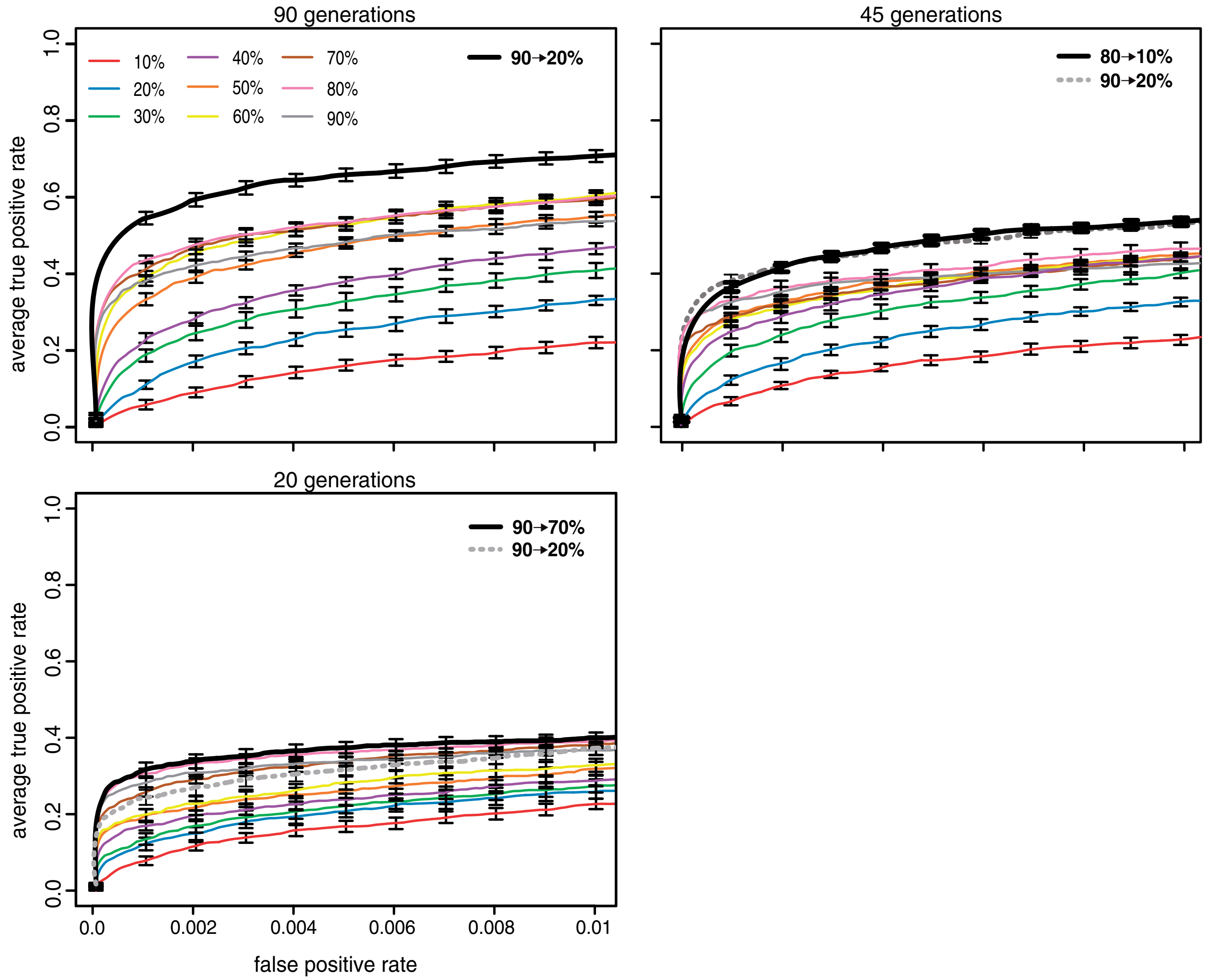

Fig. 3. Influence of the number of generations (*default* = 90). The performance of the best increasing regime (black, bold), the 90 → 20% increasing regime (gray, dashed), and the constant regimes (different colors) is shown. Although increasing regimes consistently perform best, the advantage of an increasing regime is most pronounced for long E&R studies.

Influence of the Trait Architecture

We next asked if the optimal selection regime depends on the architecture of a quantitative trait. We investigated the influence of the number of QTNs, the heritability and the effect size distribution of the QTNs.

First, we considered the influence of the number of QTNs. We simulated E&R studies with 25 and 1,000 QTNs in addition to the default of 100 QTNs. In agreement with previous works we found that the performance of E&R studies is weak when the number of QTNs is large (supplementary fig. 9, Supplementary Material online [Kofler and Schlötterer 2014; Kessner and Novembre 2015]). A large number of QTNs results in widespread interference among selected loci, which cannot (or only very slowly) be resolved by recombination events arising during the experiment. Regardless of the number of QTNs, the best constant selection regime was 80%. This is again in agreement with previous works (Kessner and Novembre 2015). Although an increasing regime consistently performed better than the constant regimes (supplementary fig. 9, Supplementary Material online) the advantage of an increasing regime was most pronounced for intermediate numbers of QTNs (supplementary fig. 9, Supplementary Material online; 2% increased performance for 25 QTNs; 22.4% increased performance for 100 QTNs; 6% increased performance for 1,000 QTNs). We noticed that the 90 → 20% increasing regime resulted in a good performance for diverse numbers of QTNs (supplementary fig. 9, Supplementary Material online).

The heritability, that is the proportion of the phenotypic variance that is due to the genotype, varies among environments, populations and traits (Falconer 1992; Visscher et al. 2008). To explore the influence of the heritability we simulated E&R studies with heritabilities of $h^2 = 0.3$ and $h^2 = 0.6$ in addition to the default of $h^2 = 1.0$ (supplementary fig. 10, Supplementary Material online). We found that an increasing regime consistently outperformed the constant regimes (supplementary fig. 10, Supplementary Material online) (Wilcoxon rank-sum test; $p_{0.3} = 0.035$ for 9% increased performance; $p_{0.6}$

$= 0.035$ for 7.5% increased performance; $p_{1.0} = 0.0002$ for 22.4% increased performance). Especially the 90 → 20% increasing regime had a high performance across different heritabilities. The advantage of an increasing regime was however most pronounced for a high heritability ($h^2 = 1$; supplementary fig. 10, Supplementary Material online). We also noticed that the influence of the selection regime diminishes with decreasing heritability (supplementary fig. 10, Supplementary Material online).

Finally, we evaluated the influence of the effect size distribution of the QTNs. Per default we used a distribution that captures the effect sizes found in QTL studies (gamma distribution with shape 0.42) (Hayes and Goddard 2001; Meuwissen et al. 2001). However, the effect size distribution may vary among traits, populations and even environments (El-Soda et al. 2014; Dittmar et al. 2016). To evaluate the influence of the effect size distribution we simulated E&R studies with QTNs drawn from different gamma distributions with shape parameters ranging from 0.1 to 1.0. Furthermore, we simulated one distribution where all loci had identical effect sizes. Note that the absolute value of the effect size is not important when truncating selection (i.e. soft selection) is used and that effect sizes are getting more similar with an increasing shape parameter (e.g. ratio between the 10% largest and smallest effect sizes $\Gamma_{0.1} = 602, 622$; $\Gamma_{1.0} = 62$). Interestingly, we found that an increasing regime consistently performed best when effect sizes followed a gamma distribution (Wilcoxon rank-sum test with pAUC; $p_{0.1} = 4.33e{-}05$ for 29% increased performance; $p_{0.42} = 2e{-}04$ for 22.4% increased performance; $p_{0.7} = 3.2e{-}04$ for 12.3% increased performance; $p_{1.0} = 0.035$ for 7% increased performance; fig. 4). However, when effect sizes were identical constant selection of 90% of the individuals performed best. For gamma distributed effect sizes the 90 → 20% increasing regime consistently had a high performance. Generally, we note that increasing regimes perform best when effect sizes are highly unequally distributed (e.g. gamma with shape = 0.1). This is in agreement with our proposed explanation for the good performance of increasing regimes, that is an initially delayed fixation of large effect loci combined with an encouraged fixation of small effect loci at later generations. Such dynamics will not be beneficial when all loci have identical effect sizes.

Finally, we asked if selection regimes may influence our ability to estimate the effect size distribution of QTNs. We investigated the effect sizes of QTNs among the 2,000 most significant SNPs in the simulations with gamma distributed effect sizes (supplementary fig. 11, Supplementary Material online). For both increasing and constant regimes, the 2,000 most significant SNPs only contained a fraction of the QTNs (supplementary fig. 11, Supplementary Material online). However, the best increasing regime allowed the recovery of a higher fraction of the QTNs than the best constant regime (increasing regimes 42.3%, constant regimes 29.2%). Especially, loci with weak effect sizes were more readily identified with an increasing regime (supplementary fig. 11, Supplementary Material online). Hence, increasing regimes may enable us to more accurately recover the effect size distribution of QTNs than constant regimes.

To summarize, with the exception of a trait architecture where QTNs have identical effects, increasing regimes outperformed constant regimes over a wide range of different trait architectures.

Selection That Optimizes the Power to Identify QTNs Does Not Necessarily Maximize the Phenotypic Response

Before the advent of E&R studies truncating selection was used to change a phenotype of interest. In a classic example the oil content of maize was raised from 5% to about 20% by continuously selecting the individuals with the highest oil content (Dudley and Lambert 2010). We were interested whether selection regimes that aim to optimize the power to identify QTNs (henceforth “QTN regime”) are identical to selection regimes that aim to maximize the phenotypic response (henceforth “phenotype regime”). To address this question we simulated multiple truncating selection regimes in steps of 10% selected individuals and identified 1) the best QTN regime (see above) and 2) the best phenotype regime, that is the regime that maximizes the phenotypic difference between the base and the evolved population (R).

We found that a selection regime that maximizes the phenotypic response to selection does not necessarily have a high power to identify QTNs (fig. 5). This discrepancy is especially pronounced for short experiments (fig. 5; 20 generations; Wilcoxon rank-sum test with pAUC; 90 → 70% vs. 20 → 10%; $P = 1.083e{-}05$).

To maximize the phenotypic response strong selection is optimal for short experiments whereas weaker selection is best for long experiments (fig. 5). This is in agreement with previous theoretical works which found that the optimal percentage of selected individuals increases with the length of the experiment (Robertson 1970a). For very long experiments (infinite generations) and in the absence of linkage the optimal phenotype regime approaches constant selection of 50% of the individuals (Robertson 1960). With linkage, as in our simulations, a slightly larger fraction of selected individuals is optimal (Robertson 1970b). Previous works also found that for finite experiments an increasing regime (albeit a sigmoid increase in the strength of selection) yields the largest phenotypic response to selection (Robertson 1970a). Interestingly with QTN regimes the situation is reversed. Here, weak selection is optimal for short experiments whereas stronger selection, especially at the end of the experiment, is best for long E&R studies (fig. 5). Due to these two contrasting trends the optimal QTN and phenotype regime are very similar at intermediate generations of selection (45 generations; fig. 5A).

What is responsible for the discrepancy between optimal QTN and phenotype regimes? We think that two factors are likely important: hitchhiking and replication. Neutral hitchhikers are a major source of noise for QTN studies but do not impede selection of phenotypes. Hence, for short studies, phenotype regimes may benefit more from strong selection than QTN regimes. However, the phenotypic response to selection is not affected by the number of replicates but the power to identify QTNs increases with replication

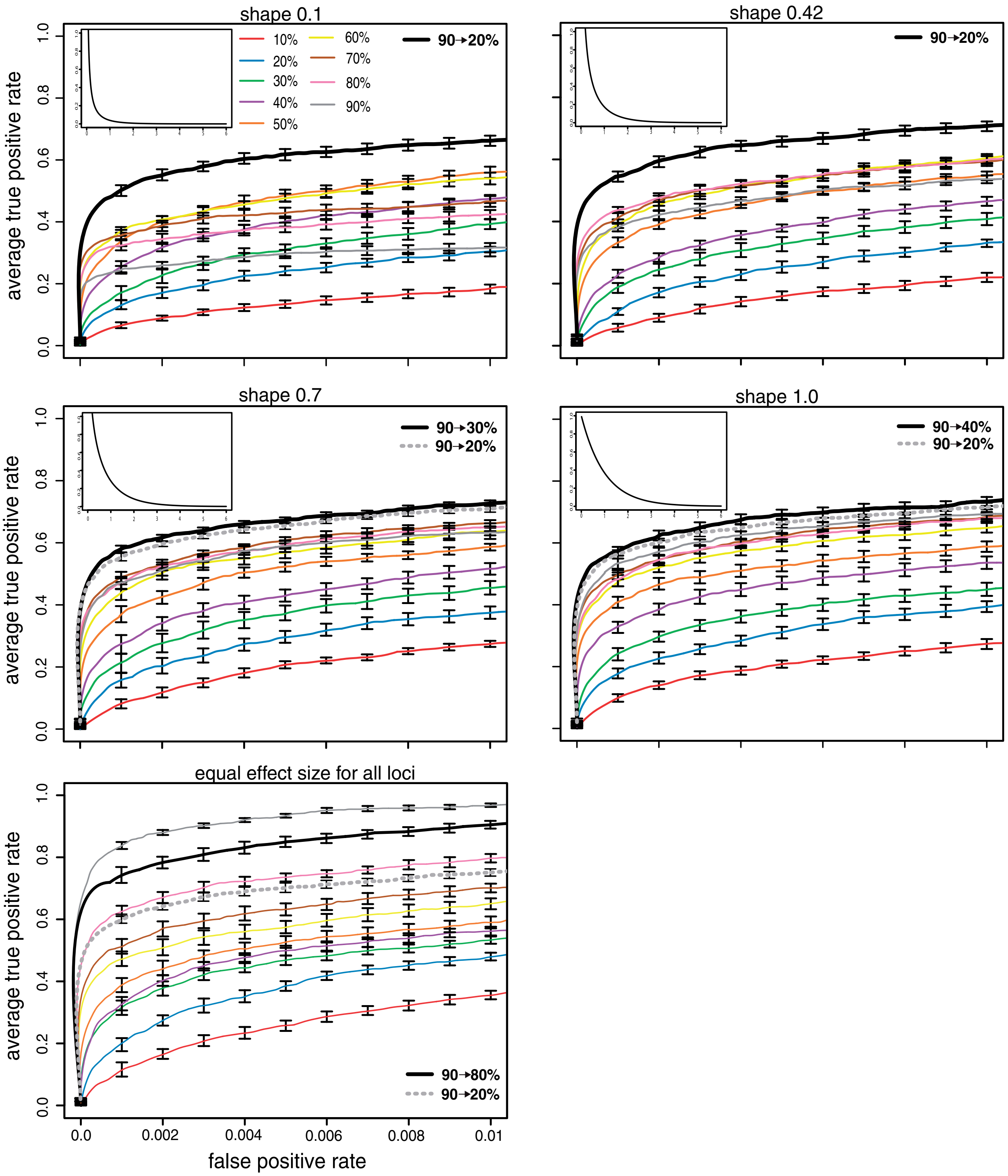

Fig. 4. Influence of the effect size distribution. We evaluated the performance of selection regimes with different effect size distributions. In addition to different gamma distributions we included a scenario where all effect sizes were identical (default is gamma with shape = 0.42). Note that effect sizes become more similar to an increasing shape parameter and that the advantage of the increasing regime is most pronounced for highly unequal effect sizes (e.g. *shape* = 0.1).

(supplementary fig. 7, Supplementary Material online). With QTN studies, strong selection at the end seems to be especially beneficial when many replicates are used (supplementary fig. 7, Supplementary Material online). This influence of replication may explain why strong selection at the end is more beneficial for QTN regimes than for phenotype regimes.

We however also made the observation that an optimized selection regime seems to be more important for QTN

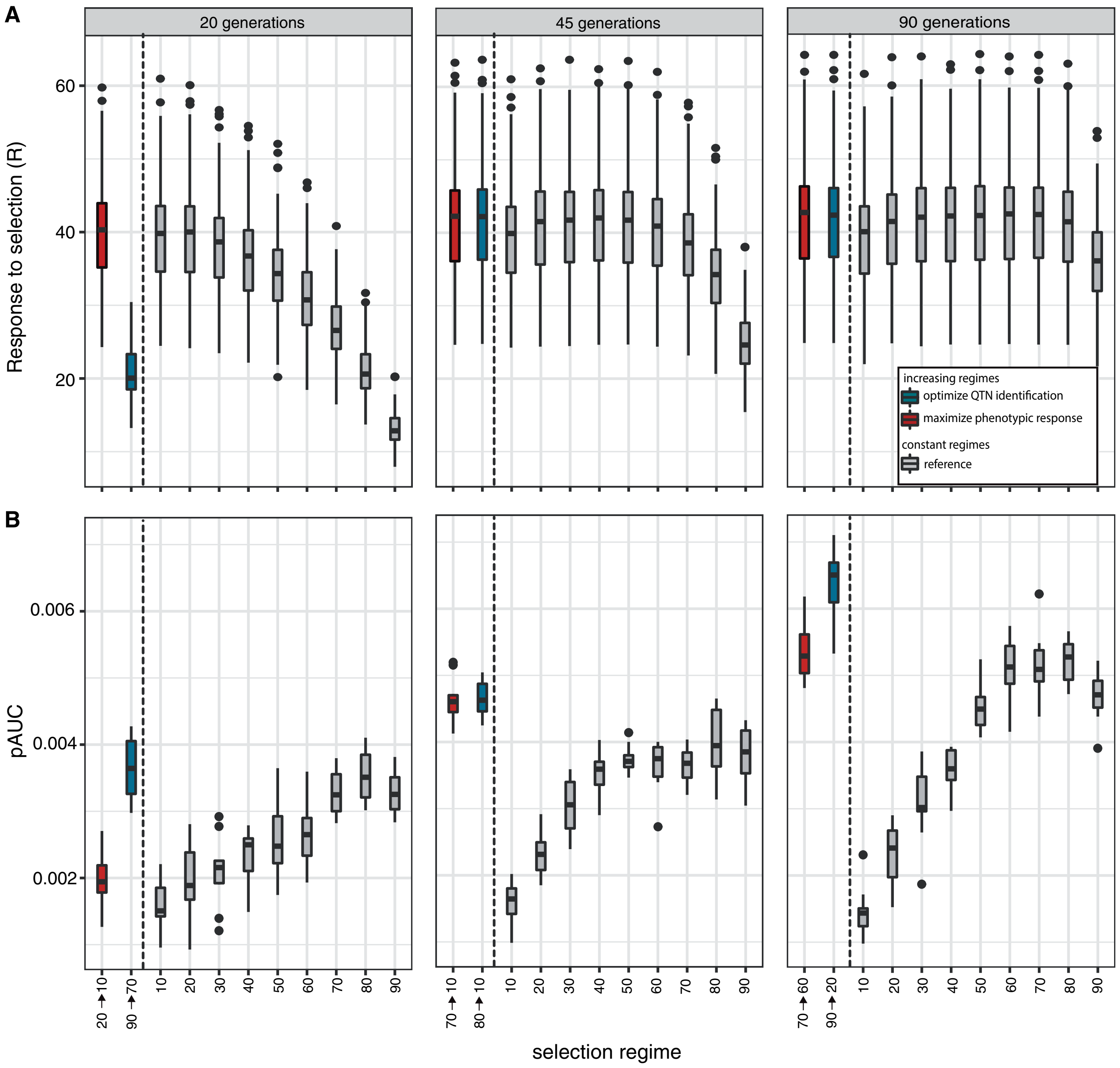

FIG. 5. Selection regimes that maximize the power to identify QTNs (blue) are not necessarily identical to regimes that maximize the phenotypic response to selection (red). As a reference the performance of constant selection regimes (gray) is shown. (*A*) Phenotypic response to selection. (*B*) Power to identify the QTNs assessed by the partial AUC (highest possible performance pAUC = 0.01).

identification than for phenotype selection. Even suboptimal phenotype regimes yield a substantial phenotypic response to selection (fig. 5A; most linear regimes yield a similar *R*). This is in agreement with previous works which found that the phenotypic response to selection is quite robust over many different selection regimes (Robertson 1960). However the same observation does not hold for QTN regimes, where deviations from the best regime lead to a noticeable drop in the power to identify QTNs (fig. 5B; constant regimes lead to dissimilar pAUC values).

We conclude that selection regimes that maximize the phenotypic response to selection are not necessarily identical to regimes that have a high power to identify QTNs. Furthermore, optimizing the selection regime is more important for QTN identification than for phenotype selection.

E&R versus GWAS

One of the most widely used approaches for identifying the genetic basis of quantitative traits are GWAS (Visscher et al. 2012). They have for example been used to shed light on the genetic basis of schizophrenia in humans (Ripke et al. 2014) and starvation resistance in *Drosophila* (Mackay et al. 2012). GWAS allow to more accurately pinpoint the location of the causative variants than the widely used QTL studies (Mackay 2001). GWAS achieve this high resolution by utilizing historical recombination events whereas QTL studies solely rely on recombination events occurring in the mapping populations (Mackay et al. 2009). Since E&R studies utilize both historical recombination events and recombination events occurring in the experimental populations we hypothesized that E&R may offer a higher power to identify QTNs than GWAS.

Ideally, one would compare the performance of a GWAS to an E&R study requiring identical effort, in terms of phenotyping, time and sequencing. Such a comparison is however difficult to accomplish. Although GWAS typically require sequencing and phenotyping of each strain/genotype separately, shortcuts may be used for E&R studies. For example, most E&R studies solely require estimates of allele frequencies which can be readily obtained by sequencing the populations as pools (e.g. with 10 replicates only 20 sequencing libraries are necessary). With E&R studies shortcuts may also be used for phenotyping. For example, Turner et al. (2011) performed an E&R study selecting for increased body size in *Drosophila* using a sieving apparatus.

We therefore decided to compare the performance of an E&R study (using default parameters; table 1) to multiple GWAS having different population sizes (ranging from 500 to 8,000). By iteratively sampling haplotypes of a small population from the next larger population we ensured that SNPs segregating in small populations are a subset of SNPs segregating in each of the larger populations. Furthermore, solely SNPs with a frequency between 5% and 95% in each population were picked as QTNs (supplementary fig. 12, Supplementary Material online). GWAS was performed with the widely used tool SNPtest (Marchini et al. 2007) and the performance of the two approaches was evaluated with ROC curves (fig. 6). Note that ROC curves avoid the problem of picking arbitrary significance thresholds for GWAS and E&R.

As expected, the power of GWAS increased with the population size (fig. 6A) (Gibson 2018). Interestingly, an E&R study with an optimized regime (90 → 10%) had a higher power to identify QTNs than a GWAS with 8,000 individuals (fig. 6A); (Wilcoxon rank-sum test; $E\&R_{1,000}$ vs. $GWAS_{8,000}$; $P = 2.16e{-}05$). This is also evident from the Manhattan plots, where most peaks in the GWAS are due to large effect loci whereas also many small effect loci generate peaks in the E&R study (fig. 6B and C). However, the power of E&R drops dramatically when a suboptimal selection regime is used (fig. 6C; red, 20% constant selection), highlighting that the selection regime is a crucial factor determining the performance of an E&R study.

Next, we tested if the performance differences between E&R studies and GWAS depend on the architecture of a trait. With a reduced heritability ($h^2 = 0.5$) the E&R study also had a higher performance than the GWAS (Wilcoxon rank-sum test; $E\&R_{1,000}$ vs. $GWAS_{8,000}$; $P = 0.023$; supplementary fig. 13, Supplementary Material online). Interestingly, when all loci have identical effect sizes, the GWAS had a higher performance than the E&R study (Wilcoxon rank-sum test; $E\&R_{1,000}$ vs. $GWAS_{8,000}$; $P = 1.08e{-}05$; supplementary fig. 13, Supplementary Material online). One complication however arises from the fact that large populations have more polymorphism than small populations. Thus for a given FPR threshold different numbers of false positive SNPs will be compared. We therefore repeated this analysis using absolute numbers of false positive SNPs, but obtained largely similar results (supplementary fig. 14, Supplementary Material online).

It has been shown that GWAS have a low power with rare alleles and alleles of small effect (Gibson 2012; Visscher et al. 2012). We were thus interested if E&R studies suffer from the same or similar weaknesses and investigated the effect sizes and allele frequencies of QTNs among the 2,000 most significant SNPs identified with either approach.

Consistent with expectations, the GWAS ($N = 8{,}000$) identified all large effect loci (>1: 100%) but only few of the small effect loci (≤1: 30%; fig. 6D). An E&R study with an optimized design had a worse performance with large effect loci (>1: 64%) but a higher performance with small effect loci (≤1: 50%; fig. 6D). Since small effect loci were more abundant than large effect loci, the E&R study allowed to identify significantly more QTNs than the GWAS (Wilcoxon rank-sum test; $E\&R_{1,000}$ vs. $GWAS_{8,000}$; $P = 0.0008$).

Furthermore, the GWAS identified most of the loci with intermediate allele frequencies ($0.2 \geq f \geq 0.8$: 46%) but few of the alleles with low or high frequencies ($f < 0.2$: 32%, $f > 0.8$: 36%; supplementary fig. 15, Supplementary Material online). In contrast, the E&R study allowed to identify most of the loci with low and medium frequencies in the base population ($f < 0.2$: 90%, $0.2 \geq f \geq 0.8$: 57%) but few of the loci having a high frequency ($f > 0.8$: 0%; supplementary fig. 15, Supplementary Material online). This is due to the fact that selected loci already starting at a high frequency may only exhibit a small allele frequency change, which will usually result in insignificant *P* values (low signal). So far, we simulated an E&R study with a powerful design ($N = 1{,}000$, 10 replicate, 90 generations). Although feasible with organisms such as *Drosophila* (Graves et al. 2017; Barghi et al. 2019), this design requires a substantial effort and may therefore be out of reach for many research questions. We were thus interested in the performance of E&R studies with a less powerful design ($N = 500$, 5 replicates, 45 generations). This low-budget design still resulted in a considerable power to identify QTNs, with a performance comparable with a GWAS with about 1,000–2,000 individuals (supplementary fig. 16, Supplementary Material online).

We conclude that an optimized E&R study may provide a higher power to identify QTNs than a GWAS. E&R studies avoid some problems of GWAS, such as a low power with rare alleles and alleles of weak effect, but have other weaknesses in turn, such as a low power with alleles starting at high frequency and a sensitivity to suboptimal selection regimes.

Discussion

We showed that E&R is a powerful approach for dissecting the genetic basis of complex traits and that the performance of E&R can be optimized by gradually increasing the strength of selection during the experiment. In contrast to previous works which showed that the performance of E&R studies may be improved by increasing the number of

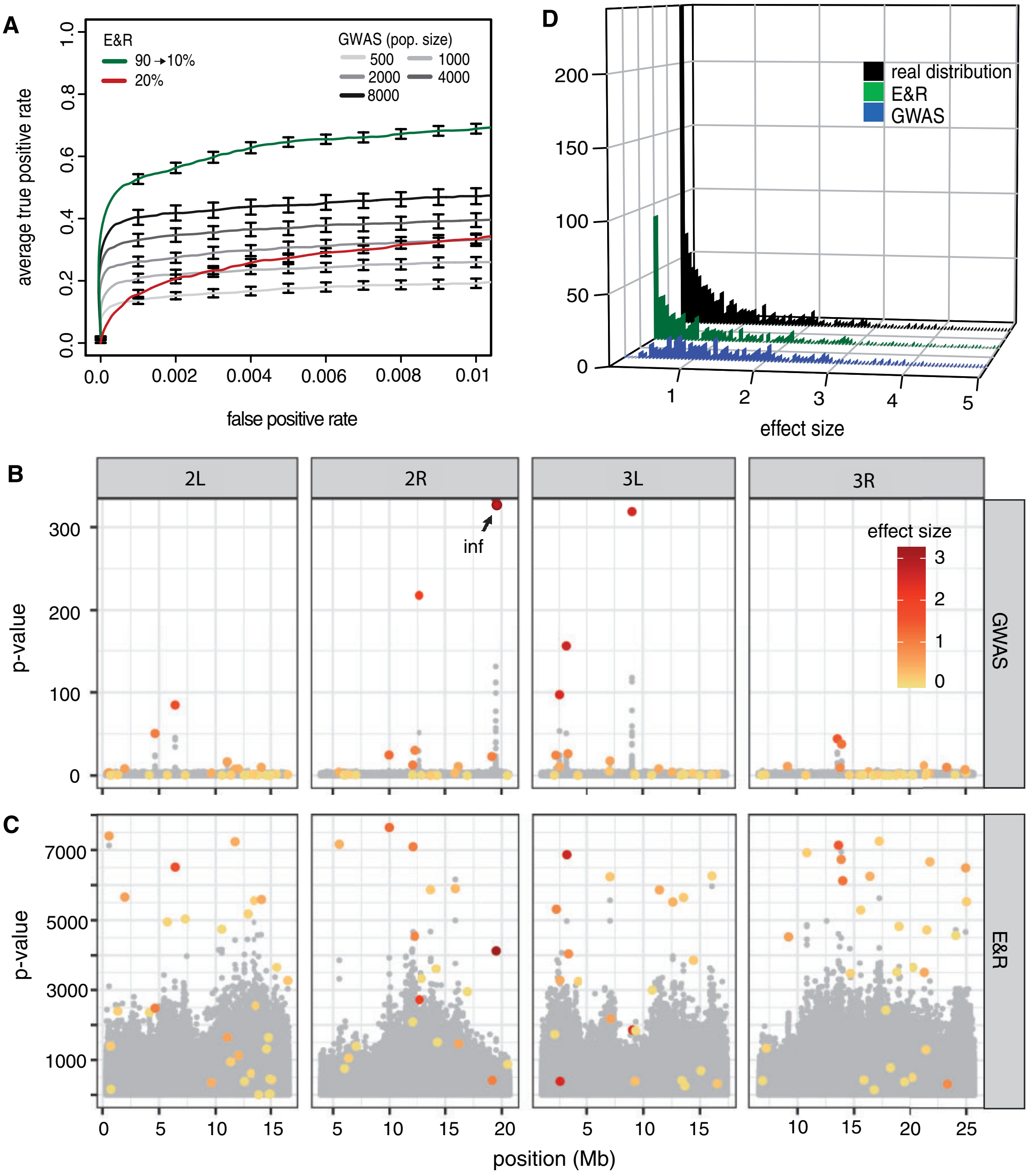

Fig. 6. Performance of E&R and GWAS. (*A*) Performance of GWAS with different population sizes and E&R studies using a powerful (green; 90 → 10%) and a suboptimal selection regime (red; constant 20%). The E&R study with an optimized selection regime has a higher power to identify QTNs than the evaluated GWAS. (*B*) Manhattan plot for a GWAS ($N = 8{,}000$). The effect sizes of QTNs are shown in a color gradient, where large effect loci are red. (*C*) Manhattan plot for an E&R study ($N = 1{,}000$, powerful selection regime). (*D*) Histogram of the effect sizes recovered with GWAS and E&R. The 2,000 most significant SNPs were used for each approach. The GWAS recovered mostly loci of large effect size whereas the E&R also recovered many loci of small effect. A summary over ten experiments is shown.

replicates, the length of the experiment and the population size (Baldwin-Brown et al. 2014; Kofler and Schlötterer 2014; Kessner and Novembre 2015), an optimized selection regime as suggested in this work comes at no, or only little, additional cost.

All approaches for identifying the genetic basis of complex traits rely on a crucial assumption about the distribution of effect sizes. A classic model proposed by Fisher et al. holds that an infinite number of loci with equal and small effects contribute to a quantitative trait (Fisher 1930; Barton et al. 2017). If this model is correct any attempts to identify the QTNs, irrespective of the used method (e.g. GWAS or E&R), are hopeless (Mackay 2001). Alternatively, Robertson (1970a) and others suggested that the distribution of QTN effects may resemble an exponential distribution, with few loci having large effects and many loci having small effects (Mackay 2001). In this case it should be feasible to identify at least a fraction of the QTNs (Mackay 2001). Although there is evidence that the infinitesimal model is a good approximation for many traits, there is also substantial evidence that effect sizes of many traits follow a more or less exponential distribution, with a few QTNs of large and many QTNs of small effect. (Hayes and Goddard 2001; Mackay 2001). For these traits it should be feasible to identify QTNs, at least QTNs with an appreciable effect size. We thus assumed a finite architecture of quantitative traits (10–1,000 QTNs; mostly using gamma distributed effect sizes) throughout the manuscript. The assumption of a limited number of QTNs also explains why we did not simulate de novo mutations. Under an infinitesimal model most mutations will hit a QTN and thus affect the quantitative trait. Hence, de novo mutations will generate some genetic variation at each generation (Barton et al. 2017). Under the alternative assumption of a finite trait architecture de novo mutations will rarely hit one of the few QTNs and thus only generate a limited amount of genetic variation. Furthermore, even if de novo mutations hit a QTN, the mutation will be restricted to a single replicate and thus only have a minor influence on the dynamics of highly replicated E&R studies as simulated in this work. For these reasons we did not consider de novo mutations.

Here, we aimed to identify the selection regime that maximizes the power to identify QTNs. Since genome-wide forward simulations are computationally demanding we simulated increasing regimes using steps of 10% selected individuals. Due to this discrete sampling of selection regimes we likely missed the absolutely best increasing regime. Nevertheless we think that our approach allowed us to obtain a reasonable approximation of the absolutely best regime as, for example, the three regimes with the highest performance in our simulations consistently have a very similar performance, slope and intercept (supplementary table 1, Supplementary Material online). It is however feasible that nonlinear selection regimes achieve a better performance than linear regimes. For example Robertson (1970b) found that the best selection regime for maximizing the phenotypic response to selection has a sigmoid shape. Mostly for computational reasons we did not consider nonlinear regimes. Evaluating the different linear regimes already required about 378,000 CPU hours and specifying the shape of nonlinear regimes will at least require one additional parameter which would substantially increase the number of necessary simulations.

We assessed the significance of the response to selection using the CMH test which contrasts, for each SNP, the allele frequency of the base and the evolved populations (Landis et al. 1978). The CMH test is fast, implemented in user-friendly software and, so far, has the best performance among tests that rely on allele frequency estimates for two time-points (Kofler et al. 2011; Kofler and Schlötterer 2014; Schlötterer et al. 2015). For theses reasons the CMH test is widely used in E&R studies (Orozco-Terwengel et al. 2012; Martins et al. 2014; Tobler et al. 2014; Phillips et al. 2018; Barghi et al. 2019; Kelly and Hughes 2019). Recently however several test-statistics became available that utilize time-series data, that is allele frequencies estimates for multiple time points (>2) during the experiment (Topa et al. 2015; Iranmehr et al. 2017; Spitzer et al. 2019). We were interested if our conclusion, that an increasing regime enhances the power to identify QTNs, also holds when a time-series based test-statistic is used. We evaluated an adaptation of the CMH test to E&R studies, which utilizes time-series data and takes the over-dispersion resulting from drift and pooled sequencing into account (Spitzer et al. 2019). In addition, we evaluated CLEAR, a composite likelihood based approach for detecting selected regions with E&R studies. With both test statistics the selection regime had a significant influence on the power to identify the QTNs and the increasing regime outperformed the constant regime (supplementary figs. 17 and 18, Supplementary Material online). Throughout this work we assumed that allele frequencies were accurately estimated, which usually requires sequencing all individuals in a population separately. As this approach is prohibitively costly most E&R studies rely on Pool-Seq, that is sequencing populations as pools, to obtain allele frequency estimates (Schlötterer et al. 2014; Long et al. 2015). To evaluate the influence of the coverage, a crucial parameter determining the accuracy of allele frequency estimates with Pool-Seq, we performed binomial sampling of allele frequencies to different coverages. An increasing regime outperformed the constant regime, irrespective of the coverage (supplementary fig. 19, Supplementary Material online).

We found that the optimal selection regime depends on the experimental setup and the trait architecture. Unfortunately the trait architecture is usually not known at the onset of an E&R study. In fact shedding light on the architecture of a trait of interest may be the aim of an E&R study. We however noticed that the 90 → 20 increasing regime, where 90% of the individuals are selected at the beginning and 20% at the end of the experiment, shows a good performance over a wide range of parameters. The only exceptions are short experiments and traits with QTNs of identical effect sizes, where constant weak selection (e.g. 80% selected individuals) achieves the best results. Furthermore, we noticed that the advantage of an increasing regime is most pronounced for traits having a high heritability and intermediate numbers of QTNs as well as E&R studies with many replicates. In any case a selection regime where the strength of selection decreases with time performed worst and we thus do not recommend to use it.

A scan of previously used selection regimes showed that many E&R studies applied very strong selection: Turner et al. (2011) selected the ≈9% largest flies for over 100 generations; Turner and Miller (2012) selected 20% of the flies with the longest pause in courtship song for 14 generations; Hardy et al. (2018) selected 20% of the most starvation resistant flies for 80 generations; Griffin et al. (2017) selected the 10% most desiccation resistant flies for 20 generations; Castro et al. (2018) selected the 20% mice with the longest legs for more than 20 generations; Although the impact of the selection regime is weaker with low heritabilities as expected for many traits used in real E&R studies ($h^2 = 0.3$–0.6 [Visscher et al. 2008]; supplementary fig. 10, Supplementary Material online), this work and previous works suggest that such strong selection likely results in a suboptimal power to identify QTNs (Kessner and Novembre 2015). We speculate that the main motivation for choosing strong selection is the concern about an insufficient response to selection of the QTNs with weak selection. With a weak response to selection it may not be possible to distinguish the QTNs from noise generated by genetic drift. We however show that hitchhikers generated by strong selection may also lead to a substantial amount of noise, thus reducing the power to identify QTNs. We thus argue that an ideal selection regime needs to strike a balance between too strong and too weak selection. This also explains why the performance of E&R studies is very sensitive to the selection regime, where a suboptimal regime may result in a dramatically reduced power to identify the QTNs. The 90 → 20 increasing regime seems to provide a good compromise between the two opposing sources of noise, drift and hitchhiking, over a wide range of parameters.

Finally, we found that E&R studies may have a higher power to identify QTNs than GWAS. Ideally the performance of these two approaches would be compared at an identical effort, in terms of sequencing, phenotyping, and time required. This is however difficult due to several reasons. In terms of sequencing an E&R study clearly requires less effort than a GWAS. Even for a powerful E&R study involving 10 replicates, solely 20 sequencing samples are necessary, whereas several hundreds (or thousands) sequencing samples are necessary for a powerful GWAS. However for GWAS with a reference panel sequencing of each strain is solely performed once and many GWAS using different traits may be performed (Mackay et al. 2012; Schlötterer et al. 2015). Thus after an initial investment, all further GWAS carried out with the reference panel will not incur additional sequencing cost. Also the phenotyping effort is difficult to compare. For example, the low-budget E&R study ($N = 500$, 5 replicates, 45 generations) had a similar performance to a GWAS with about 1,500 individuals. Hence, 112,500 (500 * 5 * 45) individuals need to be phenotyped with the E&R study to achieve a similar performance to a GWAS with 1,500 phenotyped individuals. In case all individuals need to be phenotyped separately GWAS thus clearly requires less phenotyping effort than an E&R study (unless GWAS is performed with a reference panel where each strain may be phenotyped multiple times for the same trait). However, with E&R studies, shortcuts that allow bulked phenotyping of populations are frequently used. For example Turner et al. (2011) selected for large flies using a sieving apparatus. Griffin et al. (2017) selected for desiccation resistance by exposing fly populations to dry conditions until 90% of the flies died. Hardy et al. (2018) selected for starvation resistance by depriving flies of food until 80–90% died. Other examples of bulked phenotyping, that could be used in E&R studies, are selection for flight speed using wind tunnels (Weber 1996) and selection for pathogen resistance by breeding survivors of infections (Kraaijeveld and Godfray 2008). In terms of our previous example, only 225 (5 * 45) bulked phenotypings of populations need to be performed for the E&R study compared with 1,500 phenotypings for the GWAS. When bulked phenotyping is feasible an E&R study may thus require less phenotyping effort than a GWAS. However many traits not amenable to bulked phenotyping, like pigmentation in *Drosophila*, may only become accessible to E&R studies when phenotyping can be automated, for example with devices such as the FlySorter (Zucker and Zucker 2017). An E&R study is usually much more time consuming than a GWAS, as E&R typically requires multiple generations of selection. This may take several years, depending on the organism used. An exception is a GWAS with a reference panel, where establishment of the highly inbred lines requires many generations. For example the *Drosophila* Genetic Reference panel was inbred for 20 generations (Mackay et al. 2012). This is however again a one-time investment, that is paid off by every GWAS performed with the reference panel.

For these reasons it is difficult to compare the required effort between GWAS and E&R. As a very rough guide to feasibility of an experiment, we may ask which experimental designs have been used so far. For example in *Drosophila* powerful E&R studies were already used: Barghi et al. (2019) used 10 replicates, $N = 1{,}250$ and 68 generations of adaptation. To our knowledge the largest GWAS in *Drosophila* used several 100 individuals (Mackay et al. 2012); not considering Pool-GWAS (Bastide et al. 2013). Based solely on the experimental designs that have been used so far in *Drosophila*, E&R may thus be a more powerful approach to identify QTNs than GWAS. However, GWAS involving several thousands of individuals are regularly performed in humans (Visscher et al. 2017) where E&R is not feasible. The optimal approach for identifying QTNs will thus depend on the organism and the phenotype (e.g. if bulked phenotyping is feasible).

Our results suggest that E&R studies suffer to a lesser extent from some problems of GWAS, like the weak performance with rare alleles and alleles of small effect. With an optimal selection regime E&R studies can identify many rare alleles and weak effect loci. However, E&R studies suffer from their own limitations. Most notably E&R studies have difficulties identifying alleles starting at a high frequency and E&R studies are highly sensitive to the selection regime. A suboptimal regime may result in a dramatically reduced power to identify QTNs, which makes E&R a more risky approach than GWAS. Moreover, with E&R studies it is not feasible to estimate the fraction of the genetic variation explained by the identified QTNs, a crucial benchmark for GWAS (Yang et al. 2011; Segura et al. 2012). Finally, due to the requirement for large populations and many generations of selection E&R will only be an option for small organisms with a short generation time such as yeast, *Drosophila* and *Caenorhabditis* (Long et al. 2015). Therefore, we do not view E&R as an alternative to GWAS but rather as a complementary approach with its own strengths and weaknesses.

Materials and Methods

Forward Simulations

All simulations were performed with the software MimicrEE2 (Vlachos and Kofler 2018). Briefly, MimicrEE2 is able to perform genome-wide forward simulations of evolving populations. It uses nonoverlapping generations and supports simulation of temporally varying truncating selection, that is different numbers of individuals may be selected at each generation (Vlachos and Kofler 2018). As not-evolved base population we obtained haplotypes that capture the pattern of natural variation of a *D. melanogaster* population from Vienna (2010) (Bastide et al. 2013; Kofler and Schlötterer 2014). We used the recombination rate estimates for *D. melanogaster* of Comeron et al. (2012). Recombination rate estimates were obtained for 100 kb windows from the RRC webpage (Version 2.3) (Fiston-Lavier et al. 2010). Simulations were performed for chromosomes 2L, 2R, 3L, and 3R. Hence, sex chromosomes were excluded. Low recombining regions, including the entire chromosome 4, were excluded from the analysis, as these regions inflate the FPR (Kofler and Schlötterer 2014). De novo mutations were not considered since we are mostly interested in adaptation from standing genetic variation. We simulated populations of hermaphrodites. Because males do not recombine in *D. melanogaster* we divided the recombination rate estimates by two.

If not mentioned otherwise we simulated an E&R study with a population size of $N = 1{,}000$, 10 replicates, and 90 generations of selection. We randomly picked 100 QTNs where effect sizes were drawn from a gamma distribution with shape $= 0.42$. The sign of the effect (a vs. $-a$) was randomly chosen and a heritability of $h^2 = 1$ was initially used. Only QTNs with allele frequencies between 5% and 95% were considered (default parameters used for the simulations are shown in table 1).

The QTN effects were additive (no dominance or epistasis was simulated). All simulations were repeated ten times with independent sets of randomly drawn QTNs.

Statistical Analysis

We used the CMH test (Landis et al. 1978) implemented in PoPoolation2 (Kofler et al. 2011) to identify selected loci. Previous works showed that the CMH test has a high power to identify selected loci in E&R studies (Kofler and Schlötterer 2014; Vlachos and Kofler 2018). The CMH test is based on a meta-analysis of a $2 * 2 * k$ contingency table. This contingency table contains for each replicate (k) the counts of the major and the minor allele (2), for the base and the evolved population (2). The null hypothesis is the absence of differentiation between base and evolved populations. In addition, we evaluated the performance of two time-series bases approaches: CLEAR and an adaptation of the CMH test to E&R studies (Iranmehr et al. 2017; Spitzer et al. 2019). To obtain time-series data we performed simulations with the default parameters, requesting an output each 10th generation (10 time points in total). We provided the harmonic mean of the population size (i.e. the number of selected individuals) as estimate of *Ne* required by the adapted CMH test. As CLEAR is very slow, we solely analyzed the data for a single chromosome arm (2L). Finally, we used the programming language R (R Core Team 2014) and the library ROCR to generate ROC curves and to compute the area under the ROC curve (AUC) (Sing et al. 2005).

Maximizing the Phenotypic Response

Truncating selection may be performed either to identify the QTNs or to maximize the phenotypic response. To compare the performance of selection regimes that are best suited for these two tasks we simulated truncating selection for 90 generations. The selection regime with the highest power to identify QTNs was identified as described above (E&R study with $N = 1{,}000$, 10 replicates; 10 independent sets of QTNs). To identify the regime which maximizes the phenotypic value we computed the response to selection (R: phenotypic difference between evolved and base population) for all increasing regimes using steps of 10% selected individuals (E&R study with $N = 1{,}000$, 1 replicate; 100 independent sets of QTNs). Finally we picked the regime with the largest average response.

Genome-Wide Association Studies

All GWAS were performed with the software SNPtest (Marchini et al. 2007). We used an additive model and raw phenotypic data for a quantitative trait (parameters: -frequentist 1 -method expected -use_raw_phenotypes). All simulations were performed with ten independent sets of SNPs.

Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgments

We thank Christian Schlötterer, Nick Barton, and Ilse Höllinger for comments, and Rupert Mazzucco for help with the computer cluster. The results have been partly obtained using the Vienna Scientific Cluster (VSC). We thank all members of the Institute of Population Genetics for feedback and support. This work was supported by an Austrian Science Foundation (FWF) grant (P29016-B25 to R.K.).

References

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(4):1040–1055.

Barghi N, Nolte V, Taus T, Jakšić AM, Tobler R, Mallard F, Dolezal M, Schlötterer C, Kofler R, Otte KA. 2019. Genetic redundancy fuels polygenic adaptation in *Drosophila*. *PLoS Biol.* 17(2):e3000128.

Barton NH, Etheridge AM, Véber A. 2017. The infinitesimal model: definition, derivation, and implications. *Theor Popul Biol.* 118:50–73.

Bastide H, Betancourt A, Nolte V, Tobler R, Stöbe P, Futschik A, Schlötterer C. 2013. A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*. *PLoS Genet.* 9(6):e1003534.

Castro J, Yancoskie MN, Marchini M, Belohlavy S, Beluch WH, Naumann R, Skuplik I, Cobb J, Nick H, Rolian C, et al. 2018. An integrative genomic analysis of the Longshanks selection experiment for longer limbs in mice. *bioRxiv.* 8:e42014.

Comeron J, Ratnappan R, Bailin S. 2012. The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genet.* 8(10):e1002905.

Dittmar EL, Oakley CG, Conner JK, Gould BA, Schemske DW. 2016. Factors influencing the effect size distribution of adaptive substitutions. *Proc R Soc B: Biol Sci.* 283(1828):20153065.

Dudley JW, Lambert RJ. 2004. *100 generations of selection for oil and protein in corn.* In: Janick J, editor. Plant Breeding Reviews, volume

24, part 1: long-term selection: maize. Hoboken: John Wiley and Sons. 79–110.

El-Soda M, Malosetti M, Zwaan BJ, Koornneef M, Aarts MG. 2014. Genotype–environment interaction QTL mapping in plants: lessons from *Arabidopsis*. *Trends Plant Sci*. 19(6):390–398.

Falconer D. 1992. Early selection experiments. *Annu Rev Genet*. 26(1):1–14.

Falconer D, Mackay T. 1960. Introduction to quantitative genetics. Harlow (United Kingdom): Pearson, Prentice Hall.

Fisher R. 1930. The genetical theory of natural selection. Oxford: Oxford Univ. Press.

Fiston-Lavier A-S, Singh ND, Lipatov M, Petrov DA. 2010. *Drosophila melanogaster* recombination rate calculator. *Gene* 463(1–2):18–20.

Garland T, Rose MR. 2009. Experimental evolution: concepts, methods, and applications of selection experiments. Berkeley (CA): University of California Press.

Gibson G. 2012. Rare and common variants: twenty arguments. *Nat Rev Genet*. 13(2):135–145.

Gibson G. 2018. Population genetics and GWAS: a primer. *PLoS Biol*. 16(3):e2005485.

Graves JL, Hertweck KL, Phillips MA, Han MV, Cabral LG, Barter TT, Greer LF, Burke MK, Mueller LD, Rose MR, et al. 2017. Genomics of parallel experimental evolution in *Drosophila*. *Mol Biol Evol*. 34(4):831–842.

Griffin PC, Hangartner SB, Fournier-Level A, Hoffmann AA. 2017. Genomic trajectories to desiccation resistance: convergence and divergence among replicate selected *Drosophila* lines. *Genetics* 205(2):871–890.

Hardy C, Burke M, Everett L, Han M, Lantz K, Gibbs A. 2018. Genome-wide analysis of starvation-selected *Drosophila melanogaster*—a genetic model of obesity. *Mol Biol Evol*. 35(1):50–65.

Hastie T, Tibshirani R, Friedman J. 2009. The elements of statistical learning: data mining, inference, and prediction. Springer series in statistics. New York: Springer-Verlag.

Hayes B, Goddard ME. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. *Genet Sel Evol*. 33(3):209–229.

Iranmehr A, Akbari A, Schlötterer C, Bafna V. 2017. CLEAR: composition of likelihoods for evolve and resequence experiments. *Genetics* 206(2):1011–1023.

Keightley PD, Bulfield G. 1993. Detection of quantitative trait loci from frequency changes of marker alleles under selection. *Genet Res*. 62(3):195.

Kelly JK, Hughes KA. 2019. Pervasive linked selection and intermediate-frequency alleles are implicated in an evolve-and-resequencing experiment of *Drosophila simulans*. *Genetics* 211(3):943–961.

Kessner D, Novembre J. 2015. Power analysis of artificial selection experiments using efficient whole genome simulation of quantitative traits. *Genetics* 199(4):991–1005.

Kofler R, Pandey RV, Schlötterer C. 2011. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* 27(24):3435–3436.

Kofler R, Schlötterer C. 2014. A guide for the design of evolve and resequencing studies. *Mol Biol Evol*. 31(2):474–483.

Korte A, Farlow A. 2013. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9(1):29.

Kosheleva K, Desai MM. 2018. Recombination alters the dynamics of adaptation on standing variation in laboratory yeast populations. *Mol Biol Evol*. 35(1):180–201.

Kraaijeveld A, Godfray H. 2008. Selection for resistance to a fungal pathogen in *Drosophila melanogaster*. *Heredity* 100(4):400.

Lander ES, Schork NJ. 1994. Genetic dissection of complex traits. *Science* 265(5181):2037–2048.

Landis JR, Heyman ER, Koch GG. 1978. Average partial association in three-way contingency tables: a review and discussion of alternative tests. *Int Stat Rev*. 46(3):237–254.

Long A, Liti G, Luptak A, Tenaillon O. 2015. Elucidating the molecular architecture of adaptation via evolve and resequence experiments. *Nat Rev Genet*. 16(10):567–582.

Losos JB, Arnold SJ, Bejerano G, Brodie ED III, Hibbett D, Hoekstra HE, Mindell DP, Monteiro A, Moritz C, Orr HA, et al. 2013. Evolutionary biology for the 21st century. *PLoS Biol*. 11(1):e1001466.

Mackay TF. 2001. The genetic architecture of quantitative traits. *Annu Rev Genet*. 35(1):303–339.

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The *Drosophila melanogaster* genetic reference panel. *Nature* 482(7384):173–178.

Mackay TFC, Stone EA, Ayroles JF. 2009. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet*. 10(8):565–577.

Marchini J, Cardon L, Phillips M, Donnelly P. 2004. The effects of human population structure on large genetic association studies. *Nat Genet*. 36(5):512–517.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*. 39(7):906–913.

Martins NE, Faria VG, Nolte V, Schlötterer C, Teixeira L, Sucena E, Magalhães S. 2014. Host adaptation to viruses relies on few genes with different cross-resistance properties. *Proc Nat Acad Sci*. 111(16):5938–15597.

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157(4):1819–1829.

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(20):4931–4941.

Phillips MA, Rutledge GA, Kezos JN, Greenspan ZS, Matty S, Arain H, Mueller LD, Talbott A, Rose MR, Shahrestani P. 2018. Effects of evolutionary history on genome wide and phenotypic convergence in *Drosophila* populations. *BMC Genomics* 19(1):1–17.

R Core Team. 2014. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Ripke S, Neale BM, Corvin A, Walters JT, Farh KH, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang H, et al. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511(7510):421–427.

Robertson A. 1960. A theory of limits in artificial selection. *Genetics* 144(95):234–249.

Robertson A. 1970a. Some optimum problems in individual selection. *Theor Popul Biol*. 1(1):120–127.

Robertson A. 1970b. *A theory of limits in artificial selection with many linked loci*. In: Kojima K, editor. Mathematical Topics in Population Genetics. Berlin: Springer-Verlag. p. 246–288.

Rockman M. 2012. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution* 66(1):1–17.

Schlötterer C, Kofler R, Versace E, Tobler R, Franssen S. 2015. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. *Heredity* 114(5):431–440.

Schlötterer C, Tobler R, Kofler R, Nolte V. 2014. Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. *Nat Rev Genet*. 15(11):749–763.

Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet*. 44(7):825–830.

Sing T, Sander O, Beerenwinkel N, Lengauer T. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics* 21(20):3940–3941.

Spitzer K, Pelizzola M, Futschik A. 2019. Modifying the Chi-square and the CMH test for population genetic inference: adapting to overdispersion. arXiv. https://arxiv.org/abs/1902.08127

Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP, Slate J. 2010. Adaptation genomics: the next generation. *Trends Ecol Evol*. 25(12):705–712.

Teotonio H, Carvalho S, Manoel D, Roque M, Chelo IM. 2012. Evolution of outcrossing in experimental populations of *Caenorhabditis elegans*. *PLoS One* 7(4):e35811.

Tobler R, Franssen SU, Kofler R, Orozco-Terwengel P, Nolte V, Hermisson J, Schlötterer C. 2014. Massive habitat-specific genomic response in *D. melanogaster* populations during experimental evolution in hot and cold environments. *Mol Biol Evol*. 31(2):364–375.

Topa H, Jónás Á, Kofler R, Kosiol C, Honkela A. 2015. Gaussian process test for high-throughput sequencing time series: application to experimental evolution. *Bioinformatics* 31(11):1762–1770.

Turner TDA, Andrew S, Fields T, 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(3):e1001336.

Turner T, Miller P. 2012. Investigating natural variation in *Drosophila* courtship song by the evolve and resequence approach. *Genetics* 191(2):633–642.

Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. *Am J Hum Genet*. 90(1):7–24.

Visscher PM, Hill WG, Wray NR. 2008. Heritability in the genomics era—concepts and misconceptions. *Nat Rev Genet*. 9(4):255–266.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 Years of GWAS discovery: biology, function, and translation. *Am J Hum Genet*. 101(1):5–22.

Vlachos C, Kofler R. 2018. MimicrEE2: genome-wide forward simulations of evolve and resequencing studies. *PLoS Comput Biol*. 14(8):e1006413.

Wannier TM, Kunjapur AM, Rice DP, McDonald MJ, Desai MM, Church GM. 2018. Adaptive evolution of genomically recoded *Escherichia coli*. *Proc Natl Acad Sci U S A*. 115(12):3090–3095.

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(1):205–213.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 88(1):76–82.

Zucker DS, Zucker MA. 2017. Insect singulating device. Seattle, WA: United States Flysorter, LLC. 20170071164. http://www.freepatentsonline.com/y2017/0071164.html.