Whole genome resequencing of extreme phenotypes in collared flycatchers highlights the difficulty of detecting quantitative trait loci in natural populations

Marty Kardos[1*], Arild Husby[2,3], S. Eryn McFarlane[4], Anna Qvarnström[4], Hans Ellegren[1*]

[1] Department of Evolutionary Biology, Evolutionary Biology Centre (EBC), Uppsala University, Uppsala, Sweden

[2] Department of Biosciences, University of Helsinki, Helsinki, Finland

[3] Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway

[4] Department of Animal Ecology, Evolutionary Biology Centre (EBC), Uppsala University, Uppsala, Sweden

[*] Address correspondence to H.E. and M.K. at hans.ellegren@ebc.uu.se and martykardos@gmail.com

Keywords: genome-wide association analysis, sexual selection, fitness, condition-dependent trait

Running title: Detecting QTL in natural populations

Abstract

Dissecting the genetic basis of phenotypic variation in natural populations is a long-standing goal in evolutionary biology. One open question is whether quantitative traits are determined only by large numbers of genes with small effects, or if variation also exists in large-effect loci. We conducted genome-wide association analyses of forehead patch size (a sexually selected trait) on 81 whole-genome-resequenced male collared flycatchers with extreme phenotypes, and on 415 males sampled independent of patch size and genotyped with a 50K SNP chip. No SNPs were genome-wide statistically significantly associated with patch size. Simulation-based power analyses suggest that the power to detect large-effect loci responsible for 10% of phenotypic variance was <0.5 in the genome resequencing analysis, and <0.1 in the SNP chip analysis. Reducing the recombination by two thirds relative to collared flycatchers modestly increased power. Tripling sample size increased power to >0.8 for resequencing of extreme phenotypes (N=243), but power remained <0.2 for the 50K SNP chip analysis (N=1,245). At least 1 million SNPs were necessary to achieve power >0.8 when analyzing 415 randomly sampled phenotypes. However, power of the 50K SNP chip to detect large-effect loci was nearly 0.8 in simulations with a small effective populations size of 1,500. These results suggest that reliably detecting large-effect trait loci in large natural populations will often require thousands of individuals and near complete sampling of the genome. Encouragingly, far fewer individuals and loci will often be sufficient to reliably detect large-effect loci in small populations with widespread strong linkage disequilibrium.

Introduction

Understanding the genetic basis of traits that contribute to fitness differences among individuals in natural populations is a long standing goal in evolutionary biology (Ellegren & Sheldon 2008; Stinchcombe & Hoekstra 2007). Identifying genetic variants associated with fitness traits, and the distribution of their effect sizes, provides clues about the content of the standing genetic variation available for selection to act on. Additionally, identifying associated genes or regulatory sequences with known functions can help to pinpoint the developmental, biochemical, and physiological pathways through which selection acting on phenotypes translates into genomic changes over time (Schielzeth & Husby 2014).

The genetic basis of traits closely associated with reproductive performance such as sexually selected traits is of particular interest because such traits may often be influenced by a combination of genes directly involved in trait expression and genes indirectly involved via their effects on condition. Theoretical predictions suggest that variation in condition-dependent, sexually-selected traits is likely to be highly polygenic, because of the large number of genes that could affect condition (Rowe & Houle 1996). Empirical data suggest that many sexually selected traits are indeed polygenic, but the distribution of effect sizes among contributing loci is not well described and seems to be highly variable (Chenoweth & McGuigan 2010; Santure *et al.* 2013). However, relatively few loci have explained a large proportion of variation in some sexually selected traits including eye-stalk length in stalk-eyed flies (Johns *et al.* 2005) and horn size in Soay sheep (Johnston *et al.* 2011). Thus, a general understanding of the genetic architecture of sexually selected traits is lacking in natural populations.

Until recently, discovering variants contributing to phenotypic variation and fitness (i.e., quantitative trait loci or 'QTL') in natural populations has been hindered by the lack of large-scale genomic data on wild individuals. However, genome sequences are accumulating rapidly for non-model species (Ellegren 2014), and technologies such as single nucleotide polymorphism (SNP) genotyping arrays, and genotyping by sequencing have made it possible to type many thousands of markers in any species. Yet, attempts to detect QTL in natural populations – where controlled crosses or breeding in captivity are usually not possible – have yielded mixed results, with some studies identifying QTL via genome-wide association (GWA) analysis (Comeault *et al.* 2014; Husby *et al.* 2015; Johnston *et al.* 2011, 2014; Parchman *et al.* 2012) or pedigree-based linkage mapping (Poissant, Johnston), and others failing to detect candidate causal variants despite moderately high heritability (Santure *et al.* 2013). Small sample sizes and small QTL effects sizes are likely reasons for the failure to detect QTL. An additional reason for the failure to detect QTL using association mapping is likely to be that strong linkage disequilibrium (LD) often extends only over short chromosome distances (Figure 1) (Slatkin 2008), which means the chances of strong associations between marker and trait loci – and thus between marker and phenotype – might be small even when large numbers of genetic markers are used.

Whole genome resequencing has recently become a realistic and increasingly used approach in population genomics (e.g., Ellegren 2014; Lamichhaney *et al*. 2015). Inspired by this progress, we reasoned that whole genome resequencing could potentially offer a novel means for mapping trait loci in natural populations. It would imply unprecedented genomic resolution in the search for loci contributing to phenotypic variation in the wild because, in contrast to approaches based on

genotyping of even very large sets of SNPs, the problem of low or no LD between the typed markers and causal variants is essentially eliminated by typing nearly all variable sites in the genome. Although large sample sizes may still be necessary to detect candidate loci for traits with polygenic genetic architectures (e.g., Allen *et al*. 2010) this could potentially be alleviated by using the experimental design of sequencing extreme phenotypes. Sampling extreme phenotypes can dramatically reduce the number of individuals necessary to achieve high power relative to studies that sample randomly from the phenotype distribution (Barnett *et al.* 2013; Emond *et al.* 2012; Gurwitz & McLeod 2013; Li *et al.* 2011; Perez-Gracia *et al.* 2002). For example, Emond *et al*. (2013) identified a modifier of chronic infection in cystic fibrosis patients using the exome sequences of only 91 individuals with extreme phenotypes. Basically, this approach entails sequencing two groups of individuals – each representing the respective lower and upper tail from the phenotypic distribution, thereby maximizing the phenotypic and genetic variance of the trait of interest among the sampled individuals.

Conducting GWA analyses on whole genome resequencing data by no means ensures that segregating large effect QTL will be detected (King & Nicolae 2014). The mean minor allele frequency (MAF) is often considerably lower in resequencing data compared to data from SNP chips where the mean MAF is often quite high (Kawakami *et al*. 2014). The power to detect phenotypic effects is lower at loci with low MAF (King & Nicolae 2014), so increasing the number of loci by adding large numbers of loci with low MAF might not translate into substantially increased power. Additionally, thresholds for statistical significance of course become more stringent as the number of loci increases, regardless of the MAF of the additional loci. Lastly, sample sizes tend to be limited in whole genome resequencing studies compared to

studies using less expensive SNP arrays. Lower sample sizes for resequencing-based GWAS means that the power to detect strong QTL effects may be low, even though causal variants are likely to be directly screened for phenotypic effects.

Collared flycatchers (*Ficedula albicollis*) are a cavity-nesting, sexually dichromatic species, which breeds in central and eastern Europe, and winters in Southern Africa. They have a mating system where males defend territories and males with higher quality territories tend to have higher reproductive success (Pärt 1994). Males have a sexually-selected white forehead patch which is used as an honest signal of quality in male-male competition for territories (Qvarnström 1997). Males with large patches tend to be in better condition, win territorial disputes, and to produce more offspring than males with smaller patches (Gustafsson *et al.* 1995; Pärt & Qvarnström 1997).

Our objective was to test whether sequencing of extreme phenotypes in collared flycatchers could identify loci contributing to variation in forehead patch size in this species. Specifically, we conducted a genome-wide association (GWA) analysis of forehead patch size based on whole genome resequencing of 81 male collared flycatchers sampled from the extreme ends of the phenotypic distribution. We also tested whether we could detect loci associated with forehead patch size using genotypes of 415 males from a custom 50K SNP chip for the collared flycatcher.

To our knowledge, this is the first study to use whole genome resequencing in conjunction with extreme phenotype sampling to study the genetic basis of phenotypic variation in a natural population. Having found no genome-wide statistically significantly associated SNPs, we used coalescent simulated genomic data to evaluate the power to detect loci with large phenotypic effects in this study. We also evaluated statistical power of association analyses when using larger sample

sizes, and in populations with either a lower recombination rate or smaller effective population size (N_e). Previous work has evaluated the power of pedigree-based QTL linkage mapping methods (Slate 2013). Additionally, the power of GWA analysis for study designs typical of human research has been assessed using simulations (Spencer *et al*., 2009). Our simulations are motivated by the need to evaluate whether association mapping is likely to detect large effect QTL given a range of sample sizes, and genomic and demographic characteristics typical of studies in natural populations.

Materials and Methods

Sampled individuals and forehead patch size measurements

Individuals included in this study were part of a long term study (2002-2012) on the Baltic island of Öland (57° 10′ N, 16° 58′ E), where the first breeding pair of collared flycatchers was observed in the 1960s (Qvarnström *et al.* 2009), although an earlier colonization cannot be excluded. We sampled 81 individuals for GWA analysis based on whole genome resequencing, and 415 separate individuals for GWA analysis using a 50K SNP. To select individuals for sequencing, we first calculated the mean patch size (patch height times patch width (mm)) among all yearly measurements on 819 adult males who were not involved in manipulative experiments that had the potential to influence patch size (e.g., brood size manipulations). We then preferentially selected males from the extreme upper and lower ends of the distribution of mean patch size for sequencing in order to maximize the phenotypic and genetic variance for patch size among the sequenced individuals. The distribution of patch size measurements is shown for the resequencing and SNP chip typing data sets in Figure S1. The mean number of yearly patch size observations per individual among the 81 resequenced males was 2.1 (min. = 1, max. = 5).

415 male collared flycatchers were selected independent of patch size and genotyped with a custom-made Illumina 50K SNP chip (Kawakami *et al.* 2014a). The mean number of patch size observations across years among these 415 males was 2.2 (min = 1, max = 7).

Whole genome resequencing, variant calling and filtering

The 81 males selected with extreme phenotypes were subjected to 100 base pair paired-end whole genome resequencing on an Illumina HiSeq instrument. Sequence reads were aligned to the collared flycatcher reference genome assembly version FicAlb1.5 (Kawakami *et al*. 2014b) using the Burrows-Wheeler Aligner (BWA) (Li & Durbin 2009). We used the Unified Genotyper in the Genome Analysis Toolkit (GATK, McKenna *et al.* 2010) to identify single nucleotide polymorphisms (SNPs) among the whole genome resequenced individuals. We applied variant quality score recalibration (VQSR) in GATK, with the top 20% scoring variants used as a training set for quality score recalibration of the remaining variants. We applied a strict tranche sensitivity threshold of 90% when filtering SNPs after VQSR. Filtering loci based on the strict 90% tranche sensitivity threshold selectively removed low MAF SNPs (see results below). Repeating the analyses while applying a less stringent tranche sensitivity threshold of 99% did not substantively affect the results (data not shown).

We used VCFtools (Danecek *et al.* 2011) for post variant calling SNP filtering. First, we discarded all genotypes with a genotype quality score ≤ 20. We then removed SNPs where the minor allele was observed only once, genotypes were missing for ≥ 4 individuals (i.e., missing in more than approximately 5% of individuals), genotypes deviated significantly ($\alpha = 0.01$, exact test) from Hardy

Weinberg proportions, or where more than two alleles were present. Sampling extreme phenotypes may enrich the data set for large effect loci being out of Hardy-Weinberg proportions, where a large number of homozygotes for different alleles could be found in the large- and small-patch samples of individuals. Repeating the analyses without filtering SNPs based on conformation to Hardy-Weinberg proportions did not qualitatively change the results (data not shown).

50K SNP chip genotyping was conducted at the SNP & Seq Technology Platform at Uppsala University (http://www.molmed.medsci.uu.se/SNP+SEQ+Technology+Platform/) on an Illumina iScan instrument. We discarded 50K SNP chip loci with minor allele frequency (MAF) ≤ 0.01, genotyping rate of < 95%, and loci failing a test for Hardy-Weinberg proportions ($\alpha = 1.0\times10^{-5}$). After these filtering steps, 37,803 out of 45,183 SNP loci (84%) remained and were used in the GWA analysis.

GWA analyses

For GWA analyses, we used linear mixed effects models included as an add-on (RepeatABEL; Husby *et al*. 2015) to the GenABEL package (Aulchenko *et al*. 2007) for the program R (R Core Team 2015). The mixed effect models account for both repeated measurements within individuals and relatedness among individuals. Specifically, we fitted a linear mixed effects model of the form

$$Y \sim X\beta + X_{SNP}\beta_{SNP} + Zg + Wp + e$$

where X is the design matrix for non-genetic fixed effects (age and year of sampling) and β are the corresponding fixed effects. X_{SNP} is the design matrix for the SNP

genotype predictor (coded 0, 1, or 2) and β_{SNP} are the corresponding SNP effects. g is a random genetic effect, p is a permanent environmental effect for each individual, and e is the error term. Patch size has been shown to increase with age (Pärt & Qvarnström 1997), and age was thus included as a fixed effect in our GWA analyses. Year of sampling was also included as a fixed effect to account for the effects of temporal environmental fluctuations on patch size (Figure S2).

Simultaneously including SNPs with large phenotypic effects in the fixed effects (i.e., to estimate the individual SNP effects) and in the random effects by including them when estimating the genetic relatedness matrix can result in reduced power to detect effects of individual SNPs (Yang *et al.* 2014). Therefore, we repeated the GWA analyses and ran the analyses separately for each chromosome. For GWA analysis of each chromosome, we estimated the genetic relatedness matrix using SNPs on all of the chromosomes other than the chromosome included in the GWA analysis. The analysis did not substantively affect the results (data not shown).

The *P*-values reported from GWA analyses are from Wald tests and are corrected for relatedness among individuals, repeated measurements, and genomic inflation (see below). We estimated the narrow sense heritability (h^2) of patch size from GWA analyses as

$$h^2 = \frac{V_a}{V_a + V_{pe} + V_e}$$

where V_a is the additive genetic variance, V_{pe} is the permanent among-individual variance due to environmental differences, and V_e is residual error variance.

253 *Correcting for multiple tests*

254 Stringent thresholds of statistical significance are necessary in order to control the
255 probability of false positive genotype-phenotype associations. The simplest
256 approaches to correct for multiple tests such as Bonferroni correction and false
257 discovery rate techniques are overly conservative in GWA studies such as this one
258 where SNP density is high and many SNPs are in substantial LD. Genotype-
259 phenotype tests at closely linked SNPs are in such cases non-independent (Clarke *et*
260 *al.* 2011). Therefore, in order to determine if the chosen type of statistical correction
261 affected the results, we used both a Bonferroni correction and a permutation approach
262 that accounts for LD among closely linked loci (Clarke *et al.* 2011) and controls the
263 probability of a single false positive occurring. For the permutation approach, we
264 repeated the GWA analysis as described above 1,000 times, each time after randomly
265 reassigning patch size measurements among individuals. We saved the *P*-value from
266 each of the 2,039,641 genotype-phenotype association tests on each randomized data
267 set. This was done in order to derive the distribution of *P*-values expected when no
268 SNPs are truly related to patch size. We then identified the statistical significance
269 threshold for the empirical GWA analysis as the *P*-value below which a single false
270 positive was identified in 5% or fewer of the randomized data sets. The threshold of
271 statistical significance determined with permutation for the GWA analyses of the
272 whole genome resequenced individuals was $P = 1.002\times10^{-7}$.

273 Many of the 50K SNP chip loci were in substantial LD in our study population
274 (Kawakami *et al.* 2014a). Thus, we used the same permutation approach as described
275 above as well as a Bonferroni correction to control for multiple testing in the GWA
276 analysis of these data. Here, the threshold of statistical significance determined by
277 permutation was $P = 2.18\times10^{-6}$.

278

279 *Controlling for genomic inflation*

280 We corrected *P*-values from GWA analyses for genomic inflation by dividing the test
281 statistic (χ^2) by the genomic inflation factor (λ). λ was estimated as the slope from a
282 regression of observed χ^2 versus expected χ^2 assuming that patch size was not affected
283 by variation at any loci. This approach is conservative because genomic inflation is
284 expected in GWA studies involving highly polygenic traits (Yang *et al.* 2011), which
285 is very likely the case for sexually selected traits (Rowe & Houle 1996) such as patch
286 size in collared flycatchers. However, repeating the analyses without correcting for
287 genomic inflation did not qualitatively change the results (data not shown).

288

289 *Simulations to evaluate power to detect QTL for patch size*

290 We used coalescent simulations to evaluate the power to detect loci with large effects
291 on patch size. We simulated genomic data with fastsimcoal2 v. 2.5.1 (Excoffier *et al.*
292 2013). We tested a range of values for the simulated N_e, and recombination and
293 mutation rates in order to identify a set of parameter values that resulted in a LD
294 pattern similar to the empirical data (Figure S3). We simulated populations with
295 constant diploid N_e = 37,500. Recombination and mutation rates were set to 3.1×10^{-8}
296 (Kawakami *et al.* 2014b), and 5.0×10^{-9} (Ellegren 2007) per base pair per generation,
297 respectively. For computational efficiency, we simulated only two single 200 kb
298 chromosomes in each population. Chromosomes of this size are sufficient because LD
299 is substantially greater than the genomic background level only for markers separated
300 by less than approximately 10-20 kb in our study population (Figure 1; Kawakami *et*
301 *al.* 2014a) and in the simulated data (Figure S3). We generated 1,500 diploid

302 individuals from each simulation repetition by randomly pairing 3,000 simulated
303 haploid chromosomes.
304 We used RepeatABEL to simulate a normally distributed quantitative trait
305 associated with SNP variation among the 1,500 individuals sampled from each
306 simulated population. The number of simulated repeated measurements per individual
307 was selected randomly from the empirical distribution of measurements among the 81
308 whole genome resequenced individuals. The simulations assumed the variance
309 components estimated from the SNP chip-based GWA analysis (Table 1). We used
310 these variance components because RepeatABEL simulates a normally distributed
311 quantitative trait and the distribution of the phenotypes in the SNP chip-based
312 empirical analysis was approximately normal. For each simulated population, the
313 phenotype was associated with a single SNP having a MAF of at least 0.2 as close as
314 possible to the physical center of the first chromosome. Our simulations therefore
315 assume that QTL effects are due to common variants, which are more easily detected
316 than rare variants. We varied the genotypic effect of the simulated QTL (a = half the
317 expected phenotypic difference between homozygotes) so that the additive genetic
318 variance attributed to the QTL (V_{qtl}) was equal to 5, 10, 15, and 20% of the total
319 phenotypic variance (V_p, which was set equal to the total phenotype variance from the
320 empirical SNP chip based GWA analysis described above), respectively. We
321 determined values of a by solving the expression $V_{qtl} = 2pqa^2$ (where p and q are the
322 frequencies of the minor and major allele, respectively, Lynch & Walsh (1998)) for a,
323 after setting V_{qtl} equal to 0.05, 0.1, 0.15, or 0.2 times V_p. For each simulation, the
324 simulated polygenic additive genetic variance (V_a*) was set to $V_a^* = V_a - V_{qtl}$, where
325 V_a is the empirical estimate of the polygenic additive genetic variance, so that the total
326 and additive genetic components of variance in the simulated phenotype was

representative of the empirical data, and held constant across all of the simulations. We consider these simulated effect sizes to be large, as they account for approximately 1/6 to 2/3 of the total heritability of the simulated phenotype.

To evaluate the power to detect QTL in the sample of individuals with extreme phenotypes, we randomly sampled 81 individuals from the upper (46 individuals) and lower (35 individuals) 10% quantiles of mean simulated patch size. We randomly sub-sampled SNPs on the first simulated chromosome (where the simulated QTL was located) without including singletons, so that SNP density was equal to the empirical whole genome resequencing data after filtering steps. The loci were randomly selected using the *sample* function in R, with the probability of a SNP being selected weighted by the squared MAF so that the least variable loci would be selectively removed as in our analysis of the empirical resequencing data. We defined power as the proportion of simulations where any SNP located on the first chromosome was statistically significantly associated with phenotype after correcting for multiple tests (see below).

We used the same simulated populations as above to evaluate the power to detect QTL in the 415 males with 50K SNP chip genotypes. Here, we randomly selected 415 simulated diploid individuals for analysis of each simulation. For the GWA analyses of these simulated data, we randomly subsampled SNPs so that marker density was as close as possible to the average density observed in the empirical GWA analyses based on the SNP chip. The SNP chip loci had a relatively high mean MAF of 0.28 (s.d. = 0.13). Therefore, we preferentially selected high MAF loci from the simulated data to maximize the power of SNP chip GWA analysis of the simulated data. We achieved this in the same way as above for the simulated analyses of the resequencing data by using the *sample* function in R to randomly select SNPs

after weighting the selection probabilities by the squared MAF. We repeated the GWA analyses of simulations of the 415 males after subsampling SNPs as described above so that the marker densities were equivalent to 50K, 100K, 250K, 500K, 1 million, and 2 million SNPs (nearly equivalent to SNP density in our whole genome resequencing data set) in the flycatcher genome assembly to evaluate the effects of marker density on power to detect QTL. We again defined power as the proportion of simulations where one or more SNPs located on chromosome one were statistically significantly associated with the simulated phenotype after correcting for multiple tests.

We ran the GWA analyses on the simulated data for power analysis as described above for the empirical data. However, there were a few necessary exceptions. First, age and year of sampling were not simulated, and were therefore not included as fixed effects in GWA analyses of simulated data for power analysis. Additionally, initial testing showed that the power to detect simulated large effect QTL was reduced when the QTL and linked SNPs were used to estimate the GRM (consistent with the findings of Yang *et al*. (2014)). Therefore, in the analyses presented below, we estimated the GRM using only the loci on the second chromosome (i.e., only using SNPs that were not linked to the simulated QTL). We used all of the simulated SNPs on the second chromosome (690 loci on average) to estimate the GRM. This approximates the general scenario where a candidate QTL region is excluded from estimation of the GRM, and a large number of loci unlinked to the candidate region are used to estimate the GRM and thus to account for polygenic effects and relatedness among the sampled individuals.

We evaluated statistical power using statistical significance thresholds ranging from very conservative to very liberal. Using permutation on each simulated dataset

was unreasonable due to the enormous computational requirements. We determined 'conservative' adjusted α values by applying the same Bonferroni-corrected statistical significance thresholds as in the empirical analyses so that the power estimates would reflect the probability of detecting large effect QTL in our empirical data. The Bonferroni adjusted α values were determined by dividing 0.05 by 2,039,641 when evaluating the power of the analysis of whole genome resequenced individuals with extreme phenotypes. We divided 0.05 by each of the number of loci of interest (37.8K, 50K, 100K, 250K, 500K, 1 million, and 2 million SNPs) when evaluating the power of the SNP chip-based GWA analysis of patch size. To derive 'moderate', and 'liberal' adjusted statistical significance threshold values, we multiplied the Bonferroni corrected statistical significance thresholds by 5 and 10, respectively. We then estimated statistical power using the conservative, moderate, and liberal adjusted statistical significance thresholds.

Effects of sample size, recombination rate, and effective population size on power of GWA analyses

The power of GWA analyses is expected to be higher with increased sample size and in populations where strong LD extends over longer chromosomal distances. To extend inferences related to power beyond our empirical study, we evaluated power to detect QTL with GWA analyses in simulated populations with a lower recombination rate, and small N_e. We also evaluated the effects on power of increasing the sample size by 3 times compared to our empirical study.

The pattern of LD in collared flycatchers (Figure 1) is not representative of all populations where GWA analyses will be done in the future. The relatively large N_e and a high recombination rate (3.1 cM/Mb on average, (Kawakami *et al.* 2014b)) in

the collared flycatcher act to reduce the chromosomal distance over which strong LD extends compared to smaller populations, or populations with lower recombination rates. Note, however, that LD is expected to extend even shorter distances in populations with larger N_e (e.g., as in some invertebrates). We conducted power analyses using simulated populations with recombination rate of 1.03 cM/Mb (1/3 times the recombination rate in collared flycatchers, Kawakami *et al.* 2014b), which is typical of the distribution of recombination rates among mammals (Dumont & Payseur 2008). We also ran power analyses on simulated populations with diploid N_e = 1,500, assuming a higher mutation rate of $\mu = 2\times10^{-7}$ to ensure enough polymorphic sites for the GWA analyses of these data to be comparable to the analyses of simulated populations with larger N_e. The physical size of the simulated chromosomes was increased to 600 kb for simulations with low recombination rate and for simulations with small N_e in order to accommodate the increased extent of strong LD in these scenarios (Figure S3). We held all other simulation parameters the same as before. We ran the GWA and power analyses on the simulated populations with lower recombination or smaller N_e as described above.

Results

Whole genome resequencing

81 male collared flycatchers were resequenced to a mean genome-wide coverage of 18.4 X (range 9.6 - 27.0 X) and mapped to a repeat-masked version of the 1.1 Gb reference assembly of the collared flycatcher genome. We applied highly stringent filters for inclusion of sites in variant identification by requesting that a genotype had been called for ≥95% of all individuals. After VQSR (with tranche sensitivity threshold of 90%) and discarding singletons, 2,039,641 SNPs remained (1,928,286

427 SNPs on autosomes and 111,373 SNPs on the Z chromosome) with mean MAF of
428 0.29 (s.d. = 0.13). Analysis of the resequencing data was repeated using a less
429 stringent tranche sensitivity threshold of 99% in the VQSR; this less stringent filtering
430 resulted in 4,376,065 SNPs remaining with mean MAF = 0.20 (s.d. = 0.13). We also
431 repeated the analysis of the resequencing data using SNPs called for ≥90% of all
432 individuals; this filtering resulted in 2,777,500 SNPs remaining after filtering. The
433 repeated analyses applying tranche sensitivity threshold of 99%, and retaining SNPs
434 called in ≥90% of all individuals did not substantively affect the results (data not
435 shown). The strict tranche sensitivity threshold of 90% preferentially removed low
436 MAF SNPs from the data set. The resulting high mean MAF means that the power to
437 detect QTL was maximized for the given number of SNPs distributed across the
438 genome.
439
440 *GWA analysis of whole genome resequencing data*
441 No SNP was genome-wide statistically significantly associated with patch size in the
442 analysis of whole genome resequenced males. The two SNPs with the lowest *P*-values
443 were located 14 base pairs apart within a large intergenic region located on
444 chromosome 2 (Figures 2 & S4). The closest annotated gene to these SNPs (*CSMD3*)
445 was located ~300 kb away. The GWA analysis indicated genomic inflation of the test
446 statistic ($\lambda = 1.11$; Figure S5). The genomic estimate of patch size heritability in the
447 whole genome resequenced males was $h^2 = 0.48$.
448
449 *GWA analysis of 50K SNP chip data*
450 There were also no SNPs genome-wide statistically significantly associated with
451 patch size in the analysis of SNP chip genotyped males (Figure 2). There was little

genomic inflation in this analysis ($\lambda = 1.02$; Figure S5). The genomic estimate of heritability was $h^2 = 0.32$. The values of h^2 from analyses of whole genome resequenced and SNP chip genotyped males are in the range of estimates based on previous pedigree-based analyses (Qvarnström 1999). There were no SNPs that simultaneously had exceptionally low *P*-values in the GWA analyses of whole genome resequencing and SNP chip data sets (Figure S6). Further, the *P*-values from the SNP chip- and resequencing-based GWA analyses were not even weakly correlated ($r^2 < 0.01$). Variance component estimates with 95% confidence intervals from each analysis are shown in Table 1.

Power analyses

Our simulations suggested that the power to detect QTL with large effects was insufficient in both of the empirical analyses (Figures 3a, S7, and S8). The results presented here in the main text are from analyses that used a medium statistical significance threshold. The results from analyses that applied Bonferroni and liberal statistical significance thresholds are presented in detail in Figures S7 and S8. Power was high (> 0.8) only in the case of extreme phenotype sampling with whole genome resequencing when V_{qtl}/V_p was ≥0.15. The power to detect QTL with effect sizes of $V_{qtl}/V_p = 0.05$ and $V_{qtl}/V_p = 0.10$ in our whole genome resequencing analysis was then 0.06 and 0.44, respectively. The power to detect QTL with effect sizes of $V_{qtl}/V_p = 0.05$ and $V_{qtl}/V_p = 0.2$ in our analysis of the 415 SNP chip genotyped individuals sampled independent of phenotype was 0.06 and 0.13, respectively.

The relationship between GWA analysis *P*-values and physical distance from the simulated QTL with effect size of $V_{qtl}/V_p = 0.05$ is shown in Figure 4a. The great majority of SNPs closely linked to the simulated QTL were not close to being

statistically significantly associated with phenotype. For example, the median *P*-value for SNPs located within 1 kb of the simulated QTL among 1,000 simulations of 81 whole genome resequenced individuals with extreme phenotypes was $P = 0.15$, which is eight orders of magnitude larger than the Bonferroni corrected threshold of statistical significance for the GWA analyses of resequenced individuals (2.5×10^{-8}). The median *P*-value for SNPs within 1 kb of the QTL in simulations of 415 individuals sampled independent of patch size was $P =$ was 0.084, which is four orders of magnitude larger than the Bonferroni corrected threshold of statistical significance for the SNP chip-based GWA analysis (1.3×10^{-6}).

Effects of number of loci, sample size, recombination rate, and N_e on power to detect QTL

The number of loci used in the SNP chip scenario had a strong effect on statistical power. However, using hundreds of thousands or millions of SNPs did not ensure that power was high (Figure 3a). For example, power with 100K and 500K SNPs was less than 0.3 and 0.75, respectively, for all simulated QTL effect sizes. Power to detect QTL with effect size of $V_{qtl}/V_p = 0.05$ was <0.5 when using 2 million SNPs. At least 1 million SNPs were necessary to achieve power of 0.8 or higher for QTL with effect sizes of $V_{qtl}/V_p = 0.15$ or 0.2 in this scenario (N = 415 individuals, Figure 3a).

Tripling the sample size (N = 243) substantially increased the power to detect large effect QTL in the simulations of whole genome resequencing of extreme phenotypes (Figure 3b). Specifically, power was > 0.8 for all simulated effect sizes, and 100% of QTL with effect size of $V_{qtl}/V_p \geq 0.1$ were detected. Power was >0.9 for all QTL effect sizes in the simulations of the SNP chip scenario when 1-2 million SNPs were used and the sample sizes were tripled to N=1,245 (Figure 3b). However,

power was quite low (<0.4) for all QTL effect sizes when 100 K or fewer SNPs were used and the sample sizes were tripled.

Next we evaluated the effects on power of increased LD due to lower recombination rate (1.03 cM/Mb instead of 3.1 cM/Mb) in the simulated populations. The recombination rate did not strongly affect the power to detect QTL with whole genome resequencing of extreme phenotypes or when 2 million SNPs were used in the SNP chip typing of randomly sampled phenotypes scenario (Figure 3c). However, the lower recombination rate resulted in moderately higher power to detect QTL in the SNP chip scenario when relatively few loci were used and the QTL effect size was large. For example, the power to detect QTL with $V_{qtl}/V_p = 0.2$ with 100K SNPs was 0.48 in simulations with a low recombination rate, and 0.27 in simulations with a high recombination rate. The general effect of a lower recombination rate in the SNP chip scenario was that power was closer to that of the whole genome sequence for any given number of SNPs used for GWA analysis (Figure 3c).

Reducing the N_e of the simulated populations substantially increased power of the GWA analyses for all numbers of loci considered (Figure 3d). Increased power with smaller N_e was limited to relatively small effect size QTL in the whole genome resequencing or extreme phenotypes scenario, and in the SNP chip scenario when 2 million SNPs were used in the GWA analyses (Figure 3d). Power increased dramatically with smaller N_e in the SNP chip scenario for analyses based on relatively few SNPs. For example, power to detect QTL with effect size of $V_{qtl}/V_p = 0.1$ using 37.8K SNPs was 0.08 in populations with large N_e, and 0.76 in populations with small N_e (Figure 3). Power was ≥ 0.67 for all QTL effect sizes and numbers of loci in the SNP chip scenario with small N_e (Figure 3d).

Discussion

Our analyses revealed no genome-wide significant loci for variation in male forehead patch size, despite moderate narrow sense heritability and typing nearly all polymorphic sites in the genome in 81 individuals with extreme phenotypes. This finding suggests that the additive genetic component of the variance in patch size is determined by a large number of loci with individually small effects (i.e., that patch size is polygenic), and that large effect loci for patch size do not exist. This is consistent with previous studies showing that patch size is condition-dependent (Gustafsson *et al.* 1995) and with previous suggestions that condition-dependent, sexually selected traits are likely to be governed by a large number of loci with individual small effects (Rowe & Houle 1996), largely due to the potentially huge number of genes affecting condition. However, as discussed below, low power of the GWA analyses of patch size means we cannot confidently conclude that large effect loci for patch size were not present.

Patch size heritability

Our results suggest that patch size had moderately high heritability. We found higher estimated narrow sense heritability of patch size in the analysis of whole genome resequenced individuals with extreme phenotypes ($h^2 = 0.48$) than for SNP chip genotyped individuals sampled independent of patch size ($h^2 = 0.31$). This difference in h^2 between analyses is likely due to an enrichment of the genetic variance in patch size in the group of sampled individuals with extreme phenotypes. Thus, the strategy of sampling extreme phenotypes appears to have been successful in maximizing the additive genetic variance for the trait, and thus increased the power of this analysis relative to analyses based on individuals sampled randomly with respect to patch size.

552 However, this difference could be due to a lower number of whole genome
553 resequenced males (81) compared to SNP chip genotyped males (415). Indeed, the
554 number of samples has been found to have stronger effects than the number of typed
555 sites on the precision of h^2 estimates (Stanton-Geddes *et al.* 2013). Nevertheless, both
556 estimates of heritability suggest that patch size is considerably heritable.
557
558 *Power to detect large effect QTL*
559 The inference of the absence of large effect SNPs for patch size assumes that we
560 would have detected loci with large effects on patch size if they existed. We sampled
561 81 males from the ends of the distribution of patch size distribution to enrich for total
562 and additive genetic variance of patch size, thus maximizing power given the number
563 of individuals available for sequencing (Gurwitz & McLeod 2013). Whole genome
564 sequencing of these males means that essentially all SNPs in the genome were
565 screened, thereby nearly eliminating the problem of low or no LD between the typed
566 SNPs and causal loci. SNP chip typing a larger number of samples (415) as in the
567 analysis of males selected independent of patch size is expected to reduce the
568 sampling error of estimated SNP effects at the typed loci. However, as previously
569 noted, random sampling with respect to the phenotype reduces the phenotypic and
570 additive genetic variance for the trait compared to when samples are selected from the
571 phenotypic extremes. Additionally, using a low density SNP chip means that only a
572 very small fraction of the genome was effectively screened for phenotypic effects
573 because strong LD extended less than 10-20 kb in our study population (Figure 1).
574 Our power analyses suggest that the power to detect large effect QTL was low in all
575 cases, even when the causal loci were directly screened for genotype-phenotype
576 associations (Figures 3a and 4a). Thus, the power analyses suggest that our empirical

577 data were not sufficient to confidently determine whether SNPs with large effects on
578 patch sized segregated in the study population.

579 The power analysis results presented here should be useful to future attempts
580 to dissect the genetic basis of complex traits in natural populations. First, the problem
581 of very low LD between typed markers and the great majority of functional positions
582 in the genomes is likely to be characteristic of many studies on natural populations in
583 the future. The distance over which strong LD persists is determined by N_e (the
584 strength of genetic drift), historical fluctuations in population size, population
585 subdivision, population admixture, and the recombination rate (Slatkin 2008). Thus,
586 genomic patterns of LD vary considerably among species and populations. For
587 example, strong LD extends over large chromosomal distances in humans (Reich *et*
588 *al.* 2001), domesticated sheep and cattle (McKay *et al.* 2007; Meadows *et al.* 2008),
589 and three-spined sticklebacks (Hohenlohe *et al.* 2012). However, LD decays much
590 more rapidly in the collared flycatcher (Kawakami *et al.* 2014a) (Figure 1), and
591 invertebrates such as the nematode *Caenorhabditis remanei* (Cutter *et al.* 2006), the
592 fruitfly *Drosophila melanogaster* (Mackay *et al.* 2012) and the mosquito *Anopheles*
593 *arabiensis* (Marsden *et al.* 2014). Rapid decay of LD with increasing chromosomal
594 distance means that QTL are more difficult to detect via association analyses of linked
595 SNPs. However, the flipside of this problem is that causal variants are more difficult
596 to pinpoint within QTL regions in species with low recombination rates or small N_e
597 where genotype-phenotype correlations may be due to causal variants located far
598 away from genotyped loci (Figure 4b).

599 Our results suggest that it will often be necessary to have many SNPs very
600 closely linked to a QTL with large effects to reliably detect its phenotypic effects
601 (Figures 3 and 4), particularly in populations where strong LD extends over only short

602 distances. We suggest that extremely high marker density (approaching whole
603 genome sequence) and very large samples will often be necessary to reliably detect
604 QTLs in populations with weak LD (e.g., due to high recombination rates or large N_e).
605 However, investing in whole genome resequencing will result in smaller increases in
606 power to detect QTL compared to very high density SNP genotyping approaches in
607 study systems with low recombination rate and/or small N_e and thus strong LD
608 extending over larger chromosome segments (Figures 3 and 4).
609 A notable result of the simulation-based power analyses is that the power to
610 detect a large effect QTL (e.g., $V_{qtl}/V_p = 0.1$) can be low when the causal SNP itself is
611 directly screened for a genotype-phenotype association (Figure 4). Clearly, this is
612 caused by relatively small sample sizes and adjusting statistical significance
613 thresholds to correct for multiple testing. For example, *P*-values smaller than 2.5×10^{-8}
614 are necessary to identify candidate QTL when 2 million loci are used in a GWA
615 analysis and a standard Bonferroni correction is applied along with an α value of 0.05.
616 Thus, having every SNP in the genome genotyped means that the sample sizes may
617 often need to be very large for large effect QTLs to consistently surpass reasonable
618 statistical significance thresholds. However, as our simulations (Figure 3) and other
619 results from humans (Barnett *et al.* 2013; Emond *et al.* 2012; Gurwitz & McLeod
620 2013; Li *et al.* 2011; Perez-Gracia *et al.* 2002) demonstrate, sampling from the ends
621 of the distribution of phenotypes can dramatically decrease the number of individuals
622 necessary to achieve high power. Resequencing of samples from the ends of the
623 phenotype distribution is therefore a promising approach to identify the genetic basis
624 of phenotypic and fitness variation in natural populations where budgets and sample
625 sizes are often small.
626

627 *QTL mapping prospects in natural populations*

628 Several affordable and relatively large-scale genotyping technologies including
629 genotyping by sequencing and SNP genotyping arrays have emerged in the last
630 several years, making it possible to genotype thousands to hundreds of thousands of
631 SNPs in any organism (Allendorf *et al.* 2010; Davey *et al.* 2011). There has been
632 much excitement about the potential for new genotyping or genotyping-by-
633 sequencing technologies to help elucidate the genetic basis of phenotypic and fitness
634 variation in natural populations (Slate *et al.* 2009; Stapley *et al.* 2010). However, the
635 simulations here along with previous results (Spencer *et al*. 2009) suggest that reliable
636 detection of QTL with large effect sizes will often require on the order of several
637 hundred thousand SNPs or whole genome sequence, along with very large sample
638 sizes to reliably detect large effect size QTL with GWA analyses. Thus it may be the
639 case that sub-genome scale genomic data will be insufficient to reliably detect large
640 effect QTLs in many other study systems where LD decays rapidly.

641 A notable result from our simulations is the dramatically higher power of
642 GWA analyses based on relatively few loci (e.g., 50K- 100K SNPs) in populations
643 with small N_e (Figure 3d and 4b). This suggests that the prospects are good for
644 detecting large effect QTL in populations with small N_e where LD is likely to extend
645 over very large distances (e.g., in long term studies of isolated populations on habitat
646 islands). Our simulations of small populations assumed $N_e = 1500$ and a
647 recombination rate of 3.1 cM/Mb. Populations with smaller N_e and/or lower
648 recombination rates are expected to have strong LD extending over longer distances
649 than in these simulations. Therefore power to detect large effect QTL with GWA
650 analyses based on tens of thousands of SNPs and sample sizes only in the hundreds
651 may be quite high in some study populations.

652 The genomic pattern of LD has been described in detail in relatively few non-
653 model species. Given its importance for the development of efficient tools for the
654 detection of the genomic basis of phenotypic and fitness variation, describing LD in
655 detail in taxa where fitness and phenotypic data are accumulating will greatly aid in
656 the efforts to identify QTLs in these species. We suggest that future GWA studies
657 should report the genomic pattern of LD and estimates of power to detect large effect
658 QTL, and interpret results in light of whether power is likely to be high or low given
659 the observed pattern of LD and the sampling design.
660 Clearly, if LD is weak and few SNPs are used for GWA analyses only a small
661 fraction of the genome can be effectively scanned for QTL and therefore QTL with
662 even very large phenotypic effects will frequently be missed. Nevertheless, QTL have
663 been detected via GWA analysis in natural populations using small numbers of loci.
664 For example, a recent GWA analysis of parasite burden in red grouse (*Lagopus*
665 *lagopus scotica*) based on only 271 SNPs identified 5 genome-wide statistically
666 significant QTL (Wenzel *et al.* 2015). How can the low power of GWA association
667 analyses using sparse SNPs be reconciled with the successful identification of
668 candidate QTL in such studies? One possibility is of course that many of the QTL
669 reported in highly underpowered studies (e.g., where only hundreds to a few thousand
670 SNPs are typed in large genomes) represent false positives, because very low power
671 means that a large proportion of positive results are expected to be false (Christley
672 2010). Ideally, reported QTL should be replicated in an independent sample(s),
673 though this is not always possible in studies on natural populations. As a result of low
674 power in combination with a possible bias towards publication of positive results,
675 false positives could be overrepresented in the literature. Alternatively, underpowered
676 GWA analyses may frequently detect a small number of true QTL (usually

677 overestimating their effect sizes, Göring *et al*. (2001)) due to the presence of a large
678 number of QTL with individually small effects if the trait is polygenic. In either case,
679 both of these scenarios will not substantially advance our understanding of the genetic
680 basis of quantitative traits in natural populations. Identifying a handful of QTL that
681 together explain a tiny fraction of trait heritability is of limited use because many
682 other undetected genes and biochemical pathways are involved but overlooked.
683 Indeed, focusing interpretation of results on the functions of a few small effect QTL
684 that happen to reach statistical significance is likely to provide a biased view of the
685 genetic and biochemical mechanisms underpinning trait variation.

686 Insufficient power to detect large effect QTL has other important implications
687 for investigations into the genetic basis of phenotypic and fitness variation in natural
688 populations. For example, one question of great interest in evolutionary biology is
689 whether quantitative traits are generally governed by a very large number of genes
690 with individually small effects or whether a substantial proportion of variation is due
691 to large effects of a small number of genes. As demonstrated here, caution is required
692 in interpreting the apparent absence of large effect loci as evidence for a polygenic
693 architecture of quantitative trait variation if power to detect QTL is low. Low power
694 to detect large effect QTL also makes it difficult to rigorously compare the genetic
695 architecture of different traits within natural populations. Describing architectural
696 differences in such traits is important for our understanding of how standing genetic
697 variation affects different traits and how these traits might respond to selection. Large
698 effect loci may be detected for some traits while QTL with similarly large effects are
699 not detected for other traits of interest. We emphasize that it should clearly be
700 acknowledged that apparent differences in the genetic architectures of different traits
701 may be caused by low power to detect large effect QTL, rather than differences in the

distribution of effect sizes of QTL among different phenotypic characteristics or fitness components.

Pedigree-based QTL linkage mapping studies might often have higher power than GWA approaches to detect large effect QTL because LD will obviously extend over longer chromosomal distances within families than in samples of unrelated individuals typical of GWA studies (Schielzeth & Husby 2014). Indeed, candidate QTL have been detected via linkage mapping (e.g., Poissant *et al*. 2012; Johnston *et al*. 2010), with some studies involving controlled crosses (Laporte *et al*. 2015) or breeding in captivity (Knief *et al*. 2012; Schielzeth *et al*. 2012). However, previous simulations suggest that the power to detect large effect QTL has been quite low (e.g., power was estimated at 0.33 for QTL explaining >10% of phenotypic variance) in some of the most powerful linkage mapping studies carried out to date (Slate 2013). Another limiting factor is that multiple generation pedigrees are only available in few study systems, thus limiting the usefulness of pedigree-based QTL mapping to relatively few species and phenotypic traits. Potentially increased power due to longer range LD is balanced by decreased precision in pinpointing causal loci among those linked to identified QTL. A more general understanding of the genetic basis of phenotypic and fitness variation in natural populations will likely require application of very large-scale genotyping or sequencing technologies in GWA studies of large samples of individuals in many natural populations representing a broad diversity of taxa and evolutionary histories. Fortunately, this goal is becoming within reach as the repertoire of genomic resources available for non-model organisms expands rapidly (Ellegren 2014).

Acknowledgements

Funding was provided by the European Research Council (HE), Knut and Alice Wallenberg foundation (HE), Swedish Research Council (HE and AQ), Stiftelsen Olle Engkvist Byggmästare (AQ), National Sciences and Engineering Research Council of Canada (SEM), and the Norwegian Research Council (AH). We are thankful to many fieldworkers who have collected phenotypic data from collared flycatchers on Öland. We thank Pall Olason for running the variant quality score recalibration on the resequencing data. Sequencing and SNP chip genotyping were performed at the SNP & SEQ Technology Platform in Uppsala, Science for Life Laboratory, Uppsala University.

References

Allen HL, Estrada K, Lettre G*, et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-838.

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics* **11**, 697-709.

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294-1296.

Barnett IJ, Lee S, Lin X (2013) Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genetic epidemiology* **37**, 142-151.

Chenoweth SF, McGuigan K (2010) The genetic basis of sexually selected variation. *Annual Review of Ecology, Evolution, and Systematics* **41**, 81-101.

Christley R (2010) Power and error: increased risk of false positive results in underpowered studies. *Open Epidemiology Journal* **3**, 16-19.

Clarke GM, Anderson CA, Pettersson FH*, et al.* (2011) Basic statistical analysis in genetic case-control studies. *Nature protocols* **6**, 121-133.

Comeault AA, Soria-Carrasco V, Gompert Z*, et al.* (2014) Genome-wide association mapping of phenotypic traits subject to a range of intensities of natural selection in *Timema cristinae*. *The American Naturalist* **183**, 711-727.

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.

Danecek P, Auton A, Abecasis G*, et al.* (2011) The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158.

Davey JW, Hohenlohe PA, Etter PD*, et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499-510.

Dumont B, Payseur B (2008) Evolution of the genomic rate of recombination in mammals. *Evolution* **62**, 276 - 294.

Ellegren H (2007) Molecular evolutionary genomics of birds. *Cytogenetic and Genome Research* **117**, 120-130.

768 Ellegren H (2014) Genome sequencing and population genomics in non-model organisms.
769 *Trends in Ecology & Evolution* **29**, 51-63.
770 Ellegren H, Sheldon BC (2008) Genetic basis of fitness differences in natural populations.
771 *Nature* **452**, 169-175.
772 Emond MJ, Louie T, Emerson J*, et al.* (2012) Exome sequencing of extreme phenotypes
773 identifies DCTN4 as a modifier of chronic *Pseudomonas aeruginosa* infection in
774 cystic fibrosis. *Nature genetics* **44**, 886-889.
775 Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic
776 inference from genomic and SNP data. *PLOS Genetics* **9**, e1003905.
777 Gurwitz D, McLeod HL (2013) Genome-wide studies in pharmacogenomics: harnessing the
778 power of extreme phenotypes. *Pharmacogenomics* **14**, 337.
779 Gustafsson L, Qvarnström A, Sheldon BC (1995) Trade-offs between life-history traits and a
780 secondary sexual character in male collared flycatchers. *Nature* **375**, 311-313.
781 Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium
782 and parallel adaptive divergence across threespine stickleback genomes.
783 *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 395-408.
784 Husby A, Kawakami T, Rönnegård L*, et al.* (2015) Genome-wide association mapping in a
785 wild avian population identifies a link between genetic and phenotypic variation in a
786 life-history trait. *Proceedings of the Royal Society of London B: Biological Sciences*
787 **282**, 20150156.
788 Johns PM, Wolfenbarger LL, Wilkinson GS (2005) Genetic linkage between a sexually
789 selected trait and X chromosome meiotic drive. *Proceedings of the Royal Society B:*
790 *Biological Sciences* **272**, 2097-2103.
791 Johnston SE, McEWAN J, Pickering NK*, et al.* (2011) Genome-wide association mapping
792 identifies the genetic basis of discrete and quantitative variation in sexual weaponry
793 in a wild sheep population. *Molecular Ecology* **20**, 2555-2566.
794 Johnston SE, Orell P, Pritchard VL*, et al.* (2014) Genome-wide SNP analysis reveals a
795 genetic basis for sea-age variation in a wild population of Atlantic salmon (*Salmo*
796 *salar*). *Molecular Ecology* **23**, 3452-3468.
797 Kawakami T, Backström N, Burri R*, et al.* (2014a) Estimation of linkage disequilibrium and
798 interspecific gene flow in *Ficedula* flycatchers by a newly developed 50k single-
799 nucleotide polymorphism array. *Molecular Ecology Resources* **14**, 1248-1260.
800 Kawakami T, Smeds L, Backström N*, et al.* (2014b) A high-density linkage map enables a
801 second-generation collared flycatcher genome assembly and reveals the patterns of
802 avian recombination rate variation and chromosomal evolution. *Molecular Ecology*
803 **23**, 4035-4058.
804 King CR, Nicolae DL (2014) GWAS to sequencing: divergence in study design and
805 analysis. *Genes* **5**, 460-476.
806 Knief U, Schielzeth H, Kempenaers B, Ellegren H, Forstmeier W (2012) QTL and
807 quantitative genetic analysis of beak morphology reveals patterns of standing genetic
808 variation in an Estrildid finch. *Molecular Ecology* **21**, 3704-3717.
809 Laporte M, Rogers SM, Dion-Côté A-M*, et al.* (2015) RAD-QTL mapping reveals both
810 genome-level parallelism and different genetic architecture underlying the evolution
811 of body shape in lake whitefish (*Coregonus clupeaformis*) species pairs. *G3 Genes|*
812 *Genomes| Genetics* **5**, 1481-1491.
813 Li D, Lewinger JP, Gauderman WJ, Murcray CE, Conti D (2011) Using extreme phenotype
814 sampling to identify the rare causal variants of quantitative traits in association
815 studies. *Genetic epidemiology* **35**, 790-799.
816 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler
817 transform. *Bioinformatics* **25**, 1754-1760.
818 Lynch M, Walsh B (1998) *Genetics and the analysis of quantitative traits*. Sinauer
819 Associates, Inc., Sunderland, MA.
820 Mackay TF, Richards S, Stone EA*, et al.* (2012) The *Drosophila melanogaster* genetic
821 reference panel. *Nature* **482**, 173-178.

Marsden CD, Lee Y, Kreppel K*, et al.* (2014) Diversity, differentiation, and linkage disequilibrium: Prospects for association mapping in the malaria vector *Anopheles arabiensis*. *G3: Genes| Genomes| Genetics* **4**, 121-131.
McKay SD, Schnabel RD, Murdoch BM*, et al.* (2007) Whole genome linkage disequilibrium maps in cattle. *BMC genetics* **8**, 74.
McKenna A, Hanna M, Banks E*, et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303.
Meadows JR, Chan EK, Kijas JW (2008) Linkage disequilibrium compared between five populations of domestic sheep. *BMC Genetics* **9**, 61.
Parchman TL, Gompert Z, Mudge J*, et al.* (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology* **21**, 2991-3005.
Pärt T (1994) Male philopatry confers a mating advantage in the migratory collared flycatcher, Ficedula albicollis. *Animal Behaviour* **48**, 401-409.
Pärt T, Qvarnström A (1997) Badge size in collared flycatchers predicts outcome of male competition over territories. *Animal Behaviour* **54**, 893-899.
Perez-Gracia JL, Gloria R-IM, Garcia-Ribas I, Maria CE (2002) The role of extreme phenotype selection studies in the identification of clinically relevant genotypes in cancer research. *Cancer* **95**, 1605-1610.
Poissant J, Davis C, Malenfant R, Hogg J, Coltman D (2012) QTL mapping for sexually dimorphic fitness-related traits in wild bighorn sheep. *Heredity* **108**, 256-263.
Purcell S, Neale B, Todd-Brown K*, et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559-575.
Qvarnström A (1999) Genotype-by-environment interactions in the determination of the size of a secondary sexual character in the collared flycatcher (*Ficedula albicollis*). *Evolution* **53**, 1564-1572.
Qvarnström A (1997) Experimentally increased badge size increases male competition and reduces male parental care in the collared flycatcher. *Proceedings of the Royal Society of London B: Biological Sciences* **264**, 1225-1231.
Qvarnström A, Wiley C, Svedin N, Vallin N (2009) Life-history divergence facilitates regional coexistence of competing *Ficedula* flycatchers. *Ecology* **90**, 1948-1957.
R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
Reich DE, Cargill M, Bolk S*, et al.* (2001) Linkage disequilibrium in the human genome. *Nature* **411**, 199-204.
Robinson MR, Santure AW, DeCauwer I, Sheldon BC, Slate J (2013) Partitioning of genetic variation across the genome using multimarker methods in a wild bird population. *Molecular Ecology* **22**, 3963-3980.
Rowe L, Houle D (1996) The lek paradox and the capture of genetic variance by condition dependent traits. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **263**, 1415-1421.
Santure AW, Cauwer I, Robinson MR*, et al.* (2013) Genomic dissection of variation in clutch size and egg mass in a wild great tit (*Parus major*) population. *Molecular Ecology* **22**, 3949-3962.
Schielzeth H, Husby A (2014) Challenges and prospects in genome-wide quantitative trait loci mapping of standing genetic variation in natural populations. *Annals of the New York Academy of Sciences* **1320**, 35-57.
Schielzeth H, Kempenaers B, Ellegren H, Forstmeier W (2012) QTL linkage mapping of zebra finch beak color shows an oligogenic control of a sexually selected trait. *Evolution* **66**, 18-30.
Slate J, Gratten J, Beraldi D*, et al.* (2009) Gene mapping in the wild with SNPs: guidelines and future directions. *Genetica* **136**, 97-107.

Slate J (2013) From Beavis to beak color: a simulation study to examine how much QTL mapping can reveal about the genetic architecture of quantitative traits. *Evolution* **67**, 1251-1262.

Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* **9**, 477-485.

Spencer CC, Su Z, Donnelly P, Marchini J (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLOS Genetics* **5**, e1000477.

Stanton-Geddes J, Yoder JB, Briskine R, Young ND, Tiffin P (2013) Estimating heritability using genomic data. *Methods in Ecology and Evolution* **4**, 1151-1158.

Stapley J, Reger J, Feulner PG*, et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology & Evolution* **25**, 705-712.

Stinchcombe J, Hoekstra H (2007) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* **100**, 158 - 170.

Wenzel MA, James MC, Douglas A, Piertney SB (2015) Genome-wide association and genome partitioning reveal novel genomic regions underlying variation in gastrointestinal nematode burden in a wild bird. *Molecular Ecology*. doi: 10.1111/mec.13313

Yang J, Weedon MN, Purcell S*, et al.* (2011) Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics* **19**, 807-812.

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics* **46**, 100-106.

Data accessibility: The data will be made available as follows.

- DNA sequences: NCBI SRA: **XXXX**
- SNP chip genotypes: Dryad doi:10.5061/dryad.v0v83
- Phenotypic and year of sampling data: Dryad doi: **XXXX**

Table 1. Variance component estimates from GWA analyses. 95% confidence intervals calculated in RepeatABEL are provided in parentheses.

Sample	V_a	V_{pe}	V_e
Whole Genome (N = 81)	251.67(161.4, 392.5)	171.91(102.1,289.5)	101.82(75.6, 137.1)
50K SNP Chip (N = 415)	69.82(55.8, 87.4)	57.43(45.4, 72.7)	88.45(79.2, 98.8)

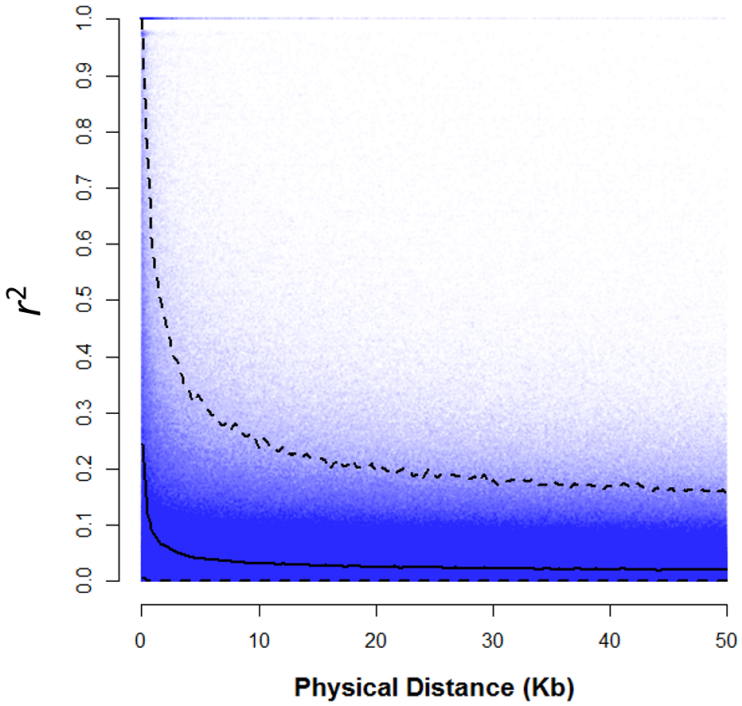

Figure 1. The relationship between the strength of linkage disequilibrium (r^2) and physical distance in 81 whole genome resequenced collared flycatcher males The data shown are from 250,000 randomly selected SNPs from the 81 whole genome resequenced collared flycatchers. r^2 was calculated using the --r2 function in PLINK (Purcell *et al.* 2007), and is shown for each pair of SNPs separated by 50 or fewer kb. The solid line represents a loess function fitted to the rolling mean of r^2 calculated in non-overlapping windows of 100 base pairs. The dashed lines represent loess functions fitted to the rolling 5% and 95% quantiles of r^2 in the same non-overlapping 100 base pair windows.

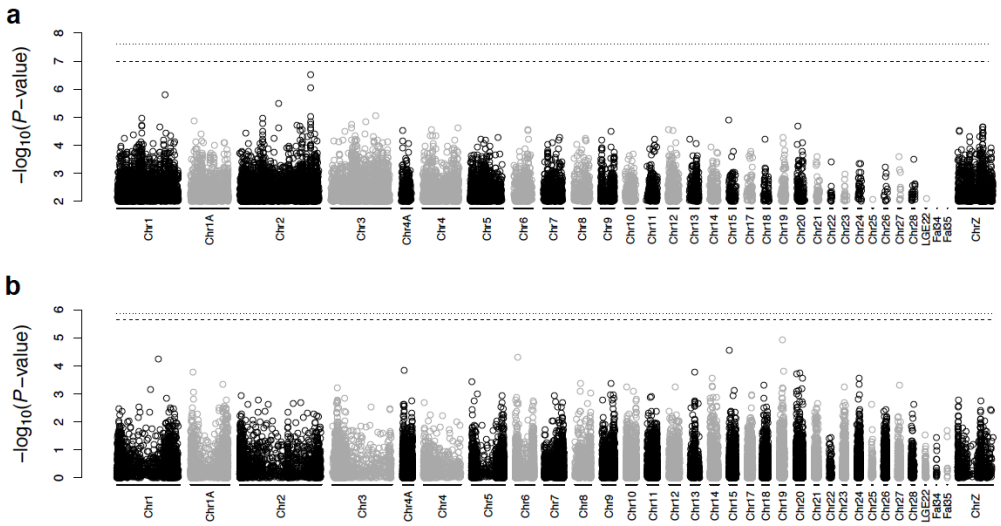

Figure 2. Manhattan plots of -$\log_{10}$(*P*-value) from GWA analyses of patch size based on whole genome resequencing of 81 males (**a**) and 50K SNP chip genotypes from 415 males (**b**). Chromosome identity is shown on the x-axis, and the *P*-values are arranged according to physical SNP positions on each chromosome (assuming 5 kb gaps between adjacent scaffolds). Dashed lines are permutation-based statistical significance thresholds, and the dotted lines are the Bonferroni statistical significance thresholds of statistical significance. Only SNPs with $-\log_{10}$(*P*-value) ≥ 2 are shown for clarity in **a**.

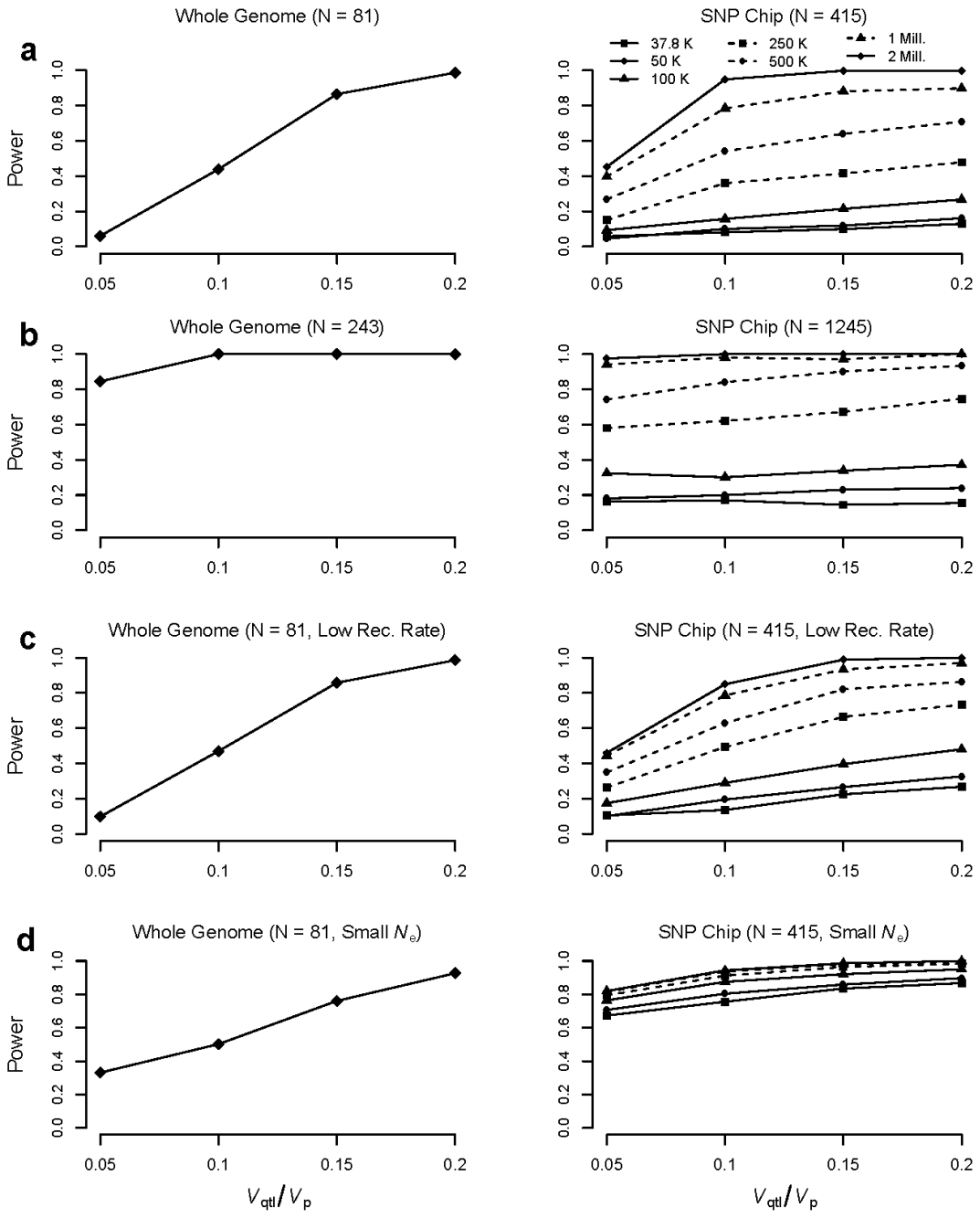

Figure 3. Influence of QTL effect size, sample size, recombination rate, and N_e on statistical power in GWA analyses. Results are shown from simulations with samples sizes equal to the empirical GWA analyses (**a**), when the simulated sample sizes are tripled (**b**), when the recombination rate was low (1.03 cM/Mb instead of 3.1 cM/Mb) (**c**), and when the simulated populations had N_e = 1,500 instead of N_e = 37,500 (**d**). Left panels show results from GWA analyses of whole genome resequenced individuals sampled with extreme phenotypes. Right panels show results from GWA analyses of individuals sampled independent of phenotype and genotyped with a 50K SNP chip. The results shown here are from analyses using the medium statistical significance threshold as described in the methods. Results from analyses of the same simulated data using conservative and liberal statistical significance thresholds are shown in Figures S7 and S8, Supplementary Materials.

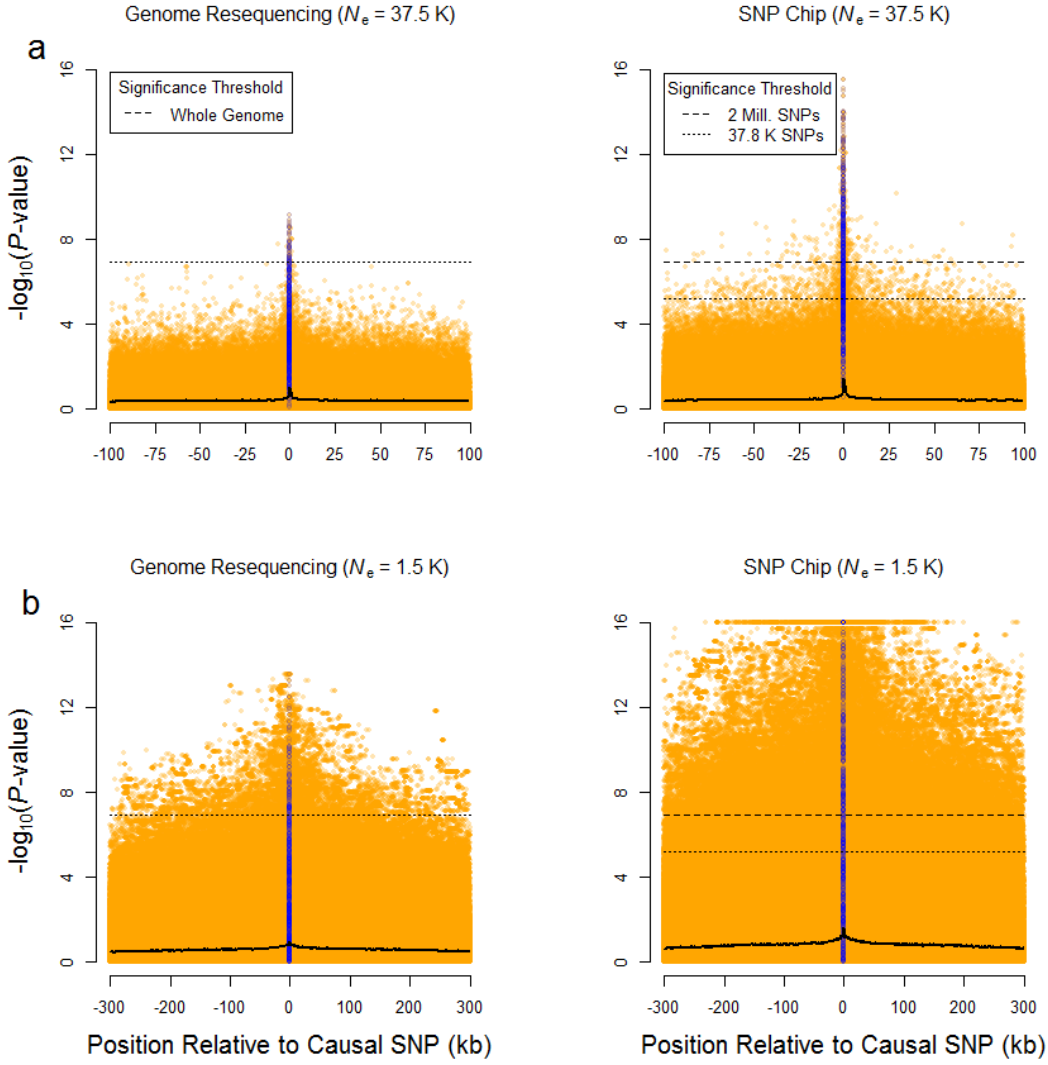

Figure 4. Effects of physical distance from a QTL on *P*-values from GWA analyses of simulated data. Results are shown from simulations where the effect size of the QTL was $V_{qtl}/V_p = 0.05$ in populations with $N_e = 37{,}500$ (**a**) and $N_e = 1500$ (**b**). The *P*-values are from every SNP in 1000 simulations mimicking our GWA analyses of 81 whole genome resequenced individuals with extreme phenotypes (left panels) and GWA analyses of 415 males sampled independent of patch size (right panels). The solid lines represent the median *P*-value calculated in 1 kb windows across the simulated chromosome. The broken lines represent the 'medium' statistical significance thresholds as indicated in the legends. Blue points at position zero on the x axis represent *P*-values from the simulated causal SNPs, and orange points represent *P*-values from SNPs linked to the causal locus. Note the range of the x axis is different in **a** and **b**.