Wittenburg *et al. BMC Genetics* (2020) 21:66
https://doi.org/10.1186/s12863-020-00871-1

BMC Genetics

METHODOLOGY ARTICLE

Open Access

Design of experiments for fine-mapping quantitative trait loci in livestock populations

Dörte Wittenburg[1*], Sarah Bonk[2], Michael Doschoris[1] and Henry Reyer[3]

Abstract

Background: Single nucleotide polymorphisms (SNPs) which capture a significant impact on a trait can be identified with genome-wide association studies. High linkage disequilibrium (LD) among SNPs makes it difficult to identify causative variants correctly. Thus, often target regions instead of single SNPs are reported. Sample size has not only a crucial impact on the precision of parameter estimates, it also ensures that a desired level of statistical power can be reached. We study the design of experiments for fine-mapping of signals of a quantitative trait locus in such a target region.

Methods: A multi-locus model allows to identify causative variants simultaneously, to state their positions more precisely and to account for existing dependencies. Based on the commonly applied SNP-BLUP approach, we determine the z-score statistic for locally testing non-zero SNP effects and investigate its distribution under the alternative hypothesis. This quantity employs the theoretical instead of observed dependence between SNPs; it can be set up as a function of paternal and maternal LD for any given population structure.

Results: We simulated multiple paternal half-sib families and considered a target region of 1 Mbp. A bimodal distribution of estimated sample size was observed, particularly if more than two causative variants were assumed. The median of estimates constituted the final proposal of optimal sample size; it was consistently less than sample size estimated from single-SNP investigation which was used as a baseline approach. The second mode pointed to inflated sample sizes and could be explained by blocks of varying linkage phases leading to negative correlations between SNPs. Optimal sample size increased almost linearly with number of signals to be identified but depended much stronger on the assumption on heritability. For instance, three times as many samples were required if heritability was 0.1 compared to 0.3. An R package is provided that comprises all required tools.

Conclusions: Our approach incorporates information about the population structure into the design of experiments. Compared to a conventional method, this leads to a reduced estimate of sample size enabling the resource-saving design of future experiments for fine-mapping of candidate variants.

Keywords: Single nucleotide polymorphism, Statistical power, Target region, SNP-BLUP, Linkage disequilibrium

*Correspondence: wittenburg@fbn-dummerstorf.de
[1] Leibniz Institute for Farm Animal Biology, Institute of Genetics and Biometry, 18196 Dummerstorf, Germany
Full list of author information is available at the end of the article

Background

Genomewide association studies (GWAS) help exploring the relationship between genetic and phenotypic variation. Genetic variation is often expressed in terms of genomic markers such as single nucleotide polymorphisms (SNPs). Identified variants may or may not be part of known genes. In a candidate-gene approach, variants are then assigned to the closest known gene and their functional importance can be studied further (e.g., [1]). The functional meaning of a variant may be differently interpreted if, due to statistical uncertainty, it was identified a few kbp upstream or downstream of its position. In general, it could be a complicated task to detect single loci as reported by Sahana et al. [2] in a study on udder health in dairy cattle. Instead of identifying important SNPs for clinical mastitis, only target regions were found. For instance, a window of about 1 Mbp length was detected on BTA6. A statistical reason for this complication lies in the high multicollinearity among predictor variables due to linkage and linkage disequilibrium (LD) between SNPs (e.g., [3]). Region-based aggregation tests in biologically relevant regions (e.g., genes; [4]) or fine-mapping approaches in independent partitions of the genome [5] have been suggested as powerful options. To eventually unravel which of the variants in a target region might be truly related to the trait, a follow-up experiment is recommended. The experimental design should account for the dependence between SNPs to ensure sufficient statistical power. This will be reflected in the sample size required. Statistical tools for the design of experiments (e.g., QUANTO; [6]) could not provide this until now. However, the denser the SNP chip is, the higher will be the correlation between SNPs. For instance, the target region on BTA6 of Sahana's paper covers 17 SNPs using a 50k SNP panel, 192 SNPs based on a 700k SNP panel and 21 796 SNPs in case of DNA sequence [2, 7].

In theory, it can be determined what sample size is needed for discovering a new variant in a single-locus model at a given power, e.g., 80 %. Such investigations are based on ANOVA (one way classification; [8]) and can also account for a hypothetical degree of LD between causative variant and SNP [9, 10]. Proposals for an optimum experimental design have been made for mapping of a quantitative trait locus (QTL) in different population structures (e.g., F2, backcross or daughter design; [11]). However, it is not clear what sample size is required to distinguish multiple independent signals of a QTL using dense marker data.

Moreover, the power of association analysis depends not only on sample size and population parameters (e.g., heritability) but also on the underlying statistical model. Among myriad options for whole genome regression models, SNP-BLUP is an obvious choice for estimating genetic effects captured by all SNPs simultaneously. Also because of its direct relationship to GBLUP (e.g., [12]), it is widely used in livestock (e.g., [13, 14]) and beyond (e.g., [15, 16]). Being enormously relevant in practice, it has been upgraded to comprise information of individuals with and without genotypic data in the framework of single-step methods [17, 18]. Though directly or indirectly estimated SNP effects are tested for being significantly different from zero [18], reports on statistical power of the underlying study design are lacking.

This paper addresses the question how to design a follow-up experiment based on a SNP-BLUP approach knowing that the predictor variables are so highly correlated. Our objective is the theoretical inference of optimal sample size to fine-map a QTL signal or to find evidence for multiple independent signals in a specified chunk of DNA. Eventually, it should be possible to detect variants at their actual position with high power. This paper concentrates on the case study of paternal half-sib families which is a typical family structure in livestock (e.g., dairy cattle). But the methodology developed enables sample size calculation for any population structure (e.g., full siblings, half siblings, mixture of both, unrelated individuals). Given the number of families, SNPs, signals of QTL and heritability, the optimal sample size is then presented as overall number of progeny. We validated our approach using simulated data. Furthermore, publicly available bovine HD SNP chip data helped verifying that the simulated linkage blocks resemble the genome structure in dairy cattle. A discussion of our achievements complements this study.

Methods

The design of experiment requires a statistical model that combines phenotype with genotype data. Here, we assume a multiple-SNP approach that considers information of as many SNPs as desired simultaneously. For comparing the outcome with a conventionally used approach, a single-SNP model is specified.

Multi-SNP model

For a joint association analysis of p SNPs with additive effects, a regression model is fitted to a phenotype $y = (y_1, \ldots, y_n)'$ of n individuals,

$$y = X\beta + e\,.$$

The $n \times p$ design matrix X contains the genotype codes: $X_{j,k} \in \{1, 0, -1\}$ for $j = 1, \ldots, n$ and $k = 1, \ldots, p$. The columns of X and the vector y are centered within family and scaled afterwards to ensure $\frac{1}{n}X'_{.,k}X_{.,k} = 1\ \forall k$ and $\frac{1}{n}y'y = 1$. This way the model becomes independent of allele frequency. The residual error term is $e \sim N(0, I_n\sigma_e^2)$. Then the coefficient vector β is estimated using a ridge regression approach as

$$\widehat{\beta} = \left(X'X + \lambda I_p\right)^{-1} X'y \,.$$

This step requires a penalty term λ which is practically obtained via cross-validation or REML approach.

Next, we investigate a multiple testing problem. For each SNP k, $k = 1, \ldots, p$, it is tested

$$H_0: \ \beta_k = 0 \text{ vs. } H_A: \ \beta_k \neq 0 \qquad (1)$$

with a suitable test statistic which is defined as the estimator of the k-th regression coefficient over its standard deviation, i.e.,

$$T_k = \frac{\widehat{\beta}_k}{SD(\widehat{\beta}_k)} \,. \qquad (2)$$

The calculation of power π requires the distribution of T_k under H_A, then

$$\pi(\mu_k) = \Pr(T_k \geq q_{1-\alpha/2}) + \Pr(T_k < q_{\alpha/2}) \,,$$

where $q_{1-\alpha/2}$ and $q_{\alpha/2}$ denote the upper and lower threshold, respectively, of the distribution of T_k under H_0 with respect to a type-I error α. Due to the ridge approach, requirements for fulfilling a t distribution do not hold ([19], p. 57). Hence the distribution of T_k is approximated as normal with mean μ_k and variance 1. The distribution mean μ_k is obtained from the expectation and variance of the estimator $\widehat{\beta}_k$. The moments are

$$E(\widehat{\beta}) = \left(X'X + \lambda I_p\right)^{-1} X'X\beta \,,$$
$$V(\widehat{\beta}) = \left(X'X + \lambda I_p\right)^{-1} X'X \left(X'X + \lambda I_p\right)^{-1} \sigma_e^2 \,.$$

The central point of our investigation is to substitute the correlation matrix $\frac{1}{n}X'X$ to be observed in the progeny generation by the theoretical correlation matrix R. For any SNP pair $k, l \in \{1, \ldots, p\}$, $\frac{1}{n}X'_{.,k}X_{.,l}$ is a plausible approximation to its expectation $E(X_{j,k}X_{j,l}) = \operatorname{cor}(X_{j,k}, X_{j,l})$ because of centered and scaled genotype codes. The derivation of R is shown in the Appendix; it requires a genetic map and genetic information of parents.

Then the mean of the test statistic becomes

$$\mu_k = \frac{\{E(\widehat{\beta})\}_k}{\sqrt{\{V(\widehat{\beta})\}_{k,k}}} = \frac{\sqrt{n}}{\sigma_e} \frac{\left\{(R + \frac{\lambda}{n}I_p)^{-1}R\beta\right\}_k}{\sqrt{\left\{(R + \frac{\lambda}{n}I_p)^{-1}R(R + \frac{\lambda}{n}I_p)^{-1}\right\}_{k,k}}} \,. \qquad (3)$$

Under H_0, $\mu_k = 0$. In order to calculate the optimal sample size, the experimenter has to specify a set of parameters: number of SNPs (p) in the investigated window of DNA, number of QTL signals to be detected (κ), proportion of variance explained by the QTL signals in that window (h^2) and number of families (e.g., N sires). The input parameters for statistical power calculation are inferred from this experimental set-up:

1. R requires haplotypes of N sires (plus genetic map and maternal LD in general).
2. We assume that all variants corresponding to the QTL signals contribute equally to the genetic variance. Hence the relative effect size is determined at κ QTL signals as

$$\frac{\beta_l}{\sigma_e} = \sqrt{\frac{h^2}{\kappa(1-h^2)}} \quad \text{for } l \text{ in the set of QTL signals} \,. \qquad (4)$$

   The remaining β's are 0.
3. The shrinkage parameter is derived corresponding to Hoerl et al. [20],

$$\lambda = p\frac{\sigma_e^2}{\beta'\beta} = p\frac{1-h^2}{h^2} \,.$$

   This is a rough approximation assuming linkage equilibrium between variants corresponding to the QTL signals.

We circumvent doing any assumption about the unknown positions of QTL signals by taking a random sample of κ positions. Then the optimal sample size is calculated over a range of n's (e.g., $1 - 5\,000$) employing the method of bisection. The minimum n that exceeds the given power is selected as "optimal" and denoted as n_{opt}. Here, we considered a power level of 80 % which is arbitrary but often used for statistical analysis (e.g., [21]). In order to get a reliable estimate of optimal sample size, sampling is repeated 100 times, and the median of n_{opt} is suggested as final n^*_{opt}. The overall type-I error was $\alpha = 0.01$.

Single-SNP model

For comparison, we consider a single SNP $k \in \{1, \ldots, p\}$ in a sliding window over the target region. Using the parameter definitions as above, the linear model in its simplest form is

$$y = X_k\beta_k + e \,.$$

Then the regression coefficient is estimated via ordinary least squares as

$$\widehat{\beta}_k = \left(X'_kX_k\right)^{-1} X'_ky \,.$$

The null hypothesis testing problem (1) and the corresponding test statistic (2) also apply in the single-SNP analysis. The test statistic is t-distributed with $n - 1$ degrees of freedom and non-centrality parameter δ_k ([19], pp. 110),

$$\delta_k = \frac{\beta_k}{\sigma_e}\sqrt{n} \,.$$

This approach neglects any impact of the other SNPs in the target region on y. Thus, a reduced pointwise error level (α_k) has to be employed to keep the overall type-I error at α. Knowing the effective number of independent tests (M_{eff}), a suitable type-I-error correction is

$$\alpha_k = \alpha / M_{\text{eff}} \,.$$

In accordance with the simple$\mathcal{M}$ method of Gao et al. [22], we suggest using R instead of $\frac{1}{n}X'X$ for calculating M_{eff}. More precisely, the number of eigenvalues of R that contribute at least 99.5 % to the sum of all eigenvalues yields M_{eff}.

Data and validation study

The software R version 3.6.1 [23] was employed in this study. Unless otherwise stated, we implemented own R scripts.

Population genetic data were simulated using the R package AlphaSimR version 0.11.0 [24]. In total, 300 SNPs were uniformly spread in a chunk of DNA of 1 cM length corresponding approximately to 1 Mbp. The SNP density roughly resembled HD data. Five traits were simulated simultaneously, one for each number of QTL signals affecting the trait, $\kappa = 1, \ldots, 5$. The founder population comprised 2 000 individuals (gender ratio 1:1) and constituted the parent generation. Other population parameters were kept at default settings (e.g., effective population size 100, mutation rate $2.5 \cdot 10^{-8}$). Using this information, the coalescent simulation program MaCS [25] was internally called: it generated parental haplotypes with realistic amount of LD. As the data simulation yielded no consistent pattern of SNP dependence, the simulation of the parent population was repeated 100 times. The maternal LD in terms of r^2 between adjacent SNPs was on average 0.45 and reflected high multicollinearity. In each repetition, $N = 10$ sires were selected with best phenotypes with respect to $\kappa = 1$ and mated with 1 000 dams. In order to resemble dairy cattle, one progeny per cross was simulated yielding 100 half-siblings per family. At each SNP, the major allele was coded as reference. Then, haplotypes of all selected sires, or a subset thereof if $N = 1$ or $N = 5$, and maternal haplotypes of 1 000 progeny were used to set up the R matrix. Few loci with no variation were disregarded. Optimal sample size was estimated based on R. Separately for each κ but using the same parent generation, N males were selected based on their phenotype as sires of half-siblings in the progeny generation. The number of dams was determined according to optimal sample size required and assuming balanced family sizes. The simulation of the progeny generation was also repeated 100 times to estimate and test SNP effects for validation purposes; this yielded 100 × 100 data sets in total. Heritability was $h^2 \in \{0.1, 0.2, 0.3\}$ which was partitioned into κ QTL effects of equal size. QTL positions were drawn at random out of the segregating sites. For each h^2, data sets were simulated independently.

Additionally, to explore a direct relationship between positions of QTL signals and n_{opt}, we selected arbitrarily a single repetition of simulation with $h^2 = 0.1$ and $N = 10$. For this particular data set, we determined n_{opt} for each SNP position (i.e., assuming one QTL signal) and for all possible SNP pairs (i.e., assuming two QTL signals).

The R package asreml version 3.0 [26] was used for association analysis. Other suitable R packages, such as rrBLUP [27] or ridge [28], had difficulties to converge or produced almost zero variance components due to the high multicollinearity of predictor variables. The multi-SNP model was applied to all simulated scenarios as described in Multi-SNP model section. Unlike in Single-SNP model section, the single-SNP model considered an additional factor $u \sim N(0, A\sigma_a^2)$ that accounts for background genetic effects due to the relationship between individuals. This was modeled similarly, e.g., in EMMAX [29] but we used the numerator relationship matrix A for computational convenience. The pointwise testing of SNP effects was followed by p-value correction according to Benjamini & Hochberg [30]. P-values from the multi-SNP model were not altered. The outcome was used to assess sensitivity and specificity of the multi-SNP and single-SNP model. For this, a window of 0.01 cM to both sides of a QTL signal (covering 2-3 SNPs) was specified in order to accept a significant SNP as a true positive result. Then, the true-positive rate (TPR) reflected sensitivity. Specificity was obtained as 1− the false-positive rate (FPR), and ROC curves were produced from TPR and FPR.

To evaluate how realistically the simulation of genetic data worked, empirical HD SNP chip data from the Dryad repository have been used [31]. These data included 1 151 dairy cows with no pedigree specification. We selected an arbitrary window on BTA7 comprising 300 SNPs on 1.16 Mbp and phased haplotypes of all animals using AlphaPhase [32]. We selected randomly 10 animals and marked them as sires in order to set up a matrix R. Because of the high SNP density, genetic distances were approximated linearly, i.e., 1 Mbp $\approx$ 1 cM. Maternal LD was roughly approximated from haplotype frequencies of all animals. Furthermore, this R matrix was used for the inspection of optimal sample size assuming $\kappa = 1, \ldots, 5$ QTL signals and $h^2 = 0.1$ and following the workflow of Multi-SNP model section.

Results

The optimal sample size suggested by the single-SNP model required the effective number of independent tests which was on average $M_{\text{eff}} = 53$ if $h^2 = 0.1$ and rather constant for R set up from $N = 1, 5$ or 10 sires ($h^2 = 0.2$: $M_{\text{eff}} = 54$; $h^2 = 0.3$: $M_{\text{eff}} = 56$). Hence results are reported for M_{eff} based on $N = 10$. Table 1 presents the

Table 1 Median of optimal sample size for detecting different number of QTL signals from 100 repetitions of simulations

QTL	$h^2 = 0.1$				$h^2 = 0.2$				$h^2 = 0.3$			
	$N = 1$	$N = 5$	$N = 10$	Single	$N = 1$	$N = 5$	$N = 10$	Single	$N = 1$	$N = 5$	$N = 10$	Single
1	128	126	127	195	57	58	57	91	34	33	34	56
2	275	269	273	382	125	126	122	175	73	70	73	106
3	421	426	436	569	214	201	205	259	126	120	120	155
4	613	540	584	756	291	288	281	342	177	170	170	204
5	763	713	685	943	385	349	344	426	228	208	207	253

Results are based on the multi-SNP approach ($N = 1, 5, 10$ families) or single-SNP approach. In each repetition, sample size was repeatedly determined for randomly drawn QTL positions and the median was calculated

median of n^*_{opt} from 100 repetitions of simulation. The median increased almost linearly with number of QTL signals but reduced with increasing heritability, and it was rather unaffected by the number of families. As an example, 127 individuals were required to fine-map a single QTL signal based on the multi-SNP model if $h^2 = 0.1$. Almost twice as much were required to distinguish two signals if $h^2 = 0.1$ or only 34 individuals were required to detect a single signal correctly when $h^2 = 0.3$ instead of $h^2 = 0.1$. Optimal sample size suggested by the multi-SNP model was 17 % to 39 % less than estimated from the single-SNP model. Figure S.1 (Additional file 1) visualizes the dependence of optimal sample size estimated from the single-SNP model on heritability. It also shows that a much larger sample was required if QTL heritability was less than 0.2.

In case of $h^2 = 0.1$, the distribution of n_{opt} is represented in Fig. 1; a separate panel is shown for each number of QTL signals to be detected. Based on 100 × 100 estimates of n_{opt}, we derived a bimodal distribution of optimal sample size in the multi-SNP model. The median of n_{opt} was consistently less than sample size estimated from single-SNP investigations. With increasing heritability, the first mode approached the median of n_{opt} but was still less than optimal sample size based on the single-SNP model, see Figures S.2 ($h^2 = 0.2$) and S.3 ($h^2 = 0.3$) (Additional file 1). The second mode appeared due to strong negative correlations between SNPs. Particularly this outcome was observed when all possible pairs of SNPs were evaluated for detecting two QTL signals in a single repetition of simulation. Figure 2a shows the correlation matrix for a single data set. Those entries of R have been selected that belonged to 10 % of the highest estimates of sample size, i.e., $n_{\text{opt}} \geq 864$ ($h^2 = 0.1$). Correspondingly, Fig. 2b indicates that, with few exceptions, negative correlations caused this outcome. The separation of SNP dependence into maternal and paternal contribution revealed further insight, and most often negative maternal LD was the driving term (Fig. S.4, Additional file 1). The distance between two QTL signals hardly influenced n_{opt} (Fig S.5, Additional file 1); any possible association was overlaid by the strong impact of correlation between loci. An additional inspection of the relationship between position of a single QTL signal and n_{opt} was not conclusive. Neither extreme maternal allele frequency nor missing sire heterozygosity led to obviously increased n_{opt} for detecting one QTL signal (Fig. S.6, Additional file 1).

The association analysis of data sets of optimal sample size was validated in terms of sensitivity and specificity of testing SNP effects. The shape of ROC curves was similar for all investigated simulation scenarios. As an example, if $N = 10$ and $\kappa = 2$, the median of n^*_{opt} was 273, and the outcome is displayed in Fig. 3. The analysis showed superiority of the multi-SNP model over the single-SNP model. In general, it was observed that the smaller n^*_{opt} was estimated, the larger both TPR and FPR turned out for the single-SNP model. The multi-SNP model performed rather robust against changes in sample size. However, the flat appearance of the ROC curve complicates fine-mapping of QTL signals based on the suggested multi-SNP approach. For instance, a TPR of 80 % is accompanied with a FPR larger than 20 %.

Blocks of varying linkage phases, as shown in Fig. 2, might be an artifact of data simulation. Based on empirical bovine HD SNP chip data, a possible R matrix was set up, see Fig. 4. The blocking structure was less pronounced. Using this R for estimating n^*_{opt} led to results being similar to the simulation study for one and two QTL signals but larger samples were required to detect more QTL signals: n^*_{opt} was 123 (1 signal), 288 (2 signals), 516 (3 signals), 800 (4 signals) and 1 342 (5 signals) if $h^2 = 0.1$. The number of repetitions of randomly drawing the positions of QTL signals did not substantially affect the final n^*_{opt}. For instance, the median deviated less than 4 % if n_{opt} was calculated 1 000 instead of 100 times.

Discussion

Our investigation contributes to the design of powerful experiments for fine-mapping of causative variant(s) in a genomic target region. We incorporated the expected dependence among SNPs in this region and estimated

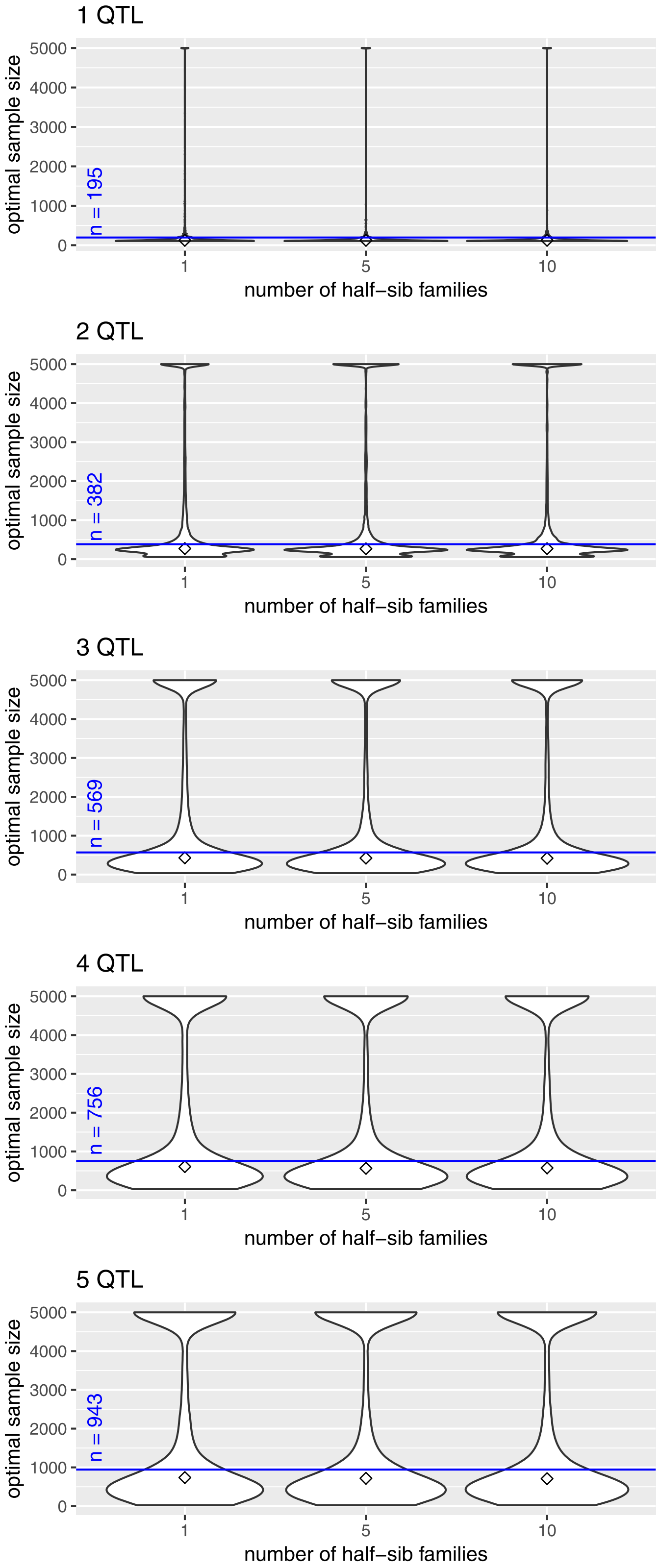

Fig. 1 Distribution of optimal sample size. Violinplot of n_{opt} vs. number of half-sib families for different numbers of QTL signals in a multi-SNP model. The parent generation was simulated 100 times and 100 random draws of positions of QTL signals were analyzed in each run, $h^2 = 0.1$. The diamond indicates the median of n_{opt} and the blue line marks the results based on a single-SNP model

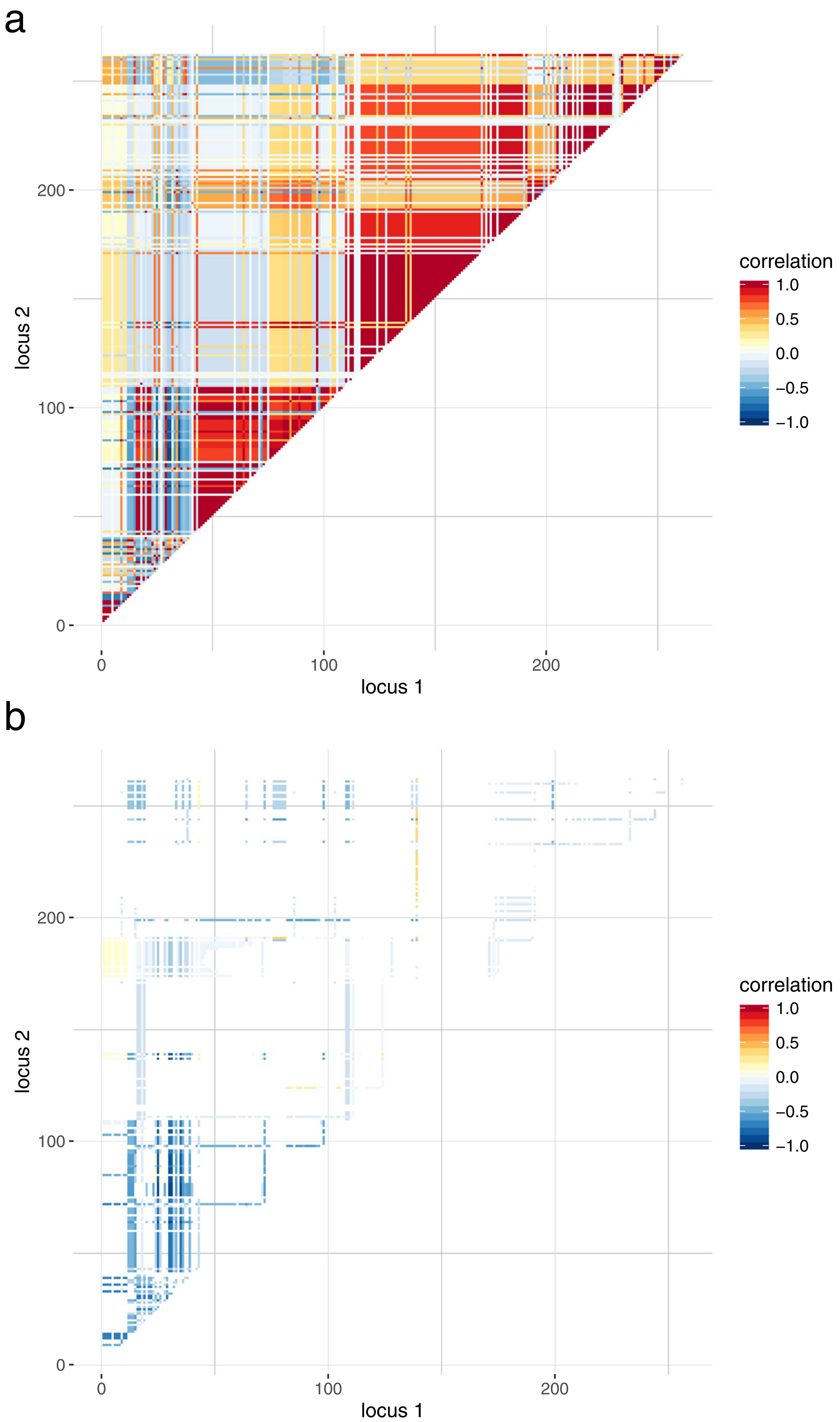

Fig. 2 Dependence between SNPs in a single simulated data set with $N = 10$ sires. **a** Correlation matrix *R*, **b** entries selected from *R* which belong to 10 % highest sample size ($n_{\text{opt}} \geq 864$). All possible SNP pairs were evaluated to detect two QTL signals ($h^2 = 0.1$)

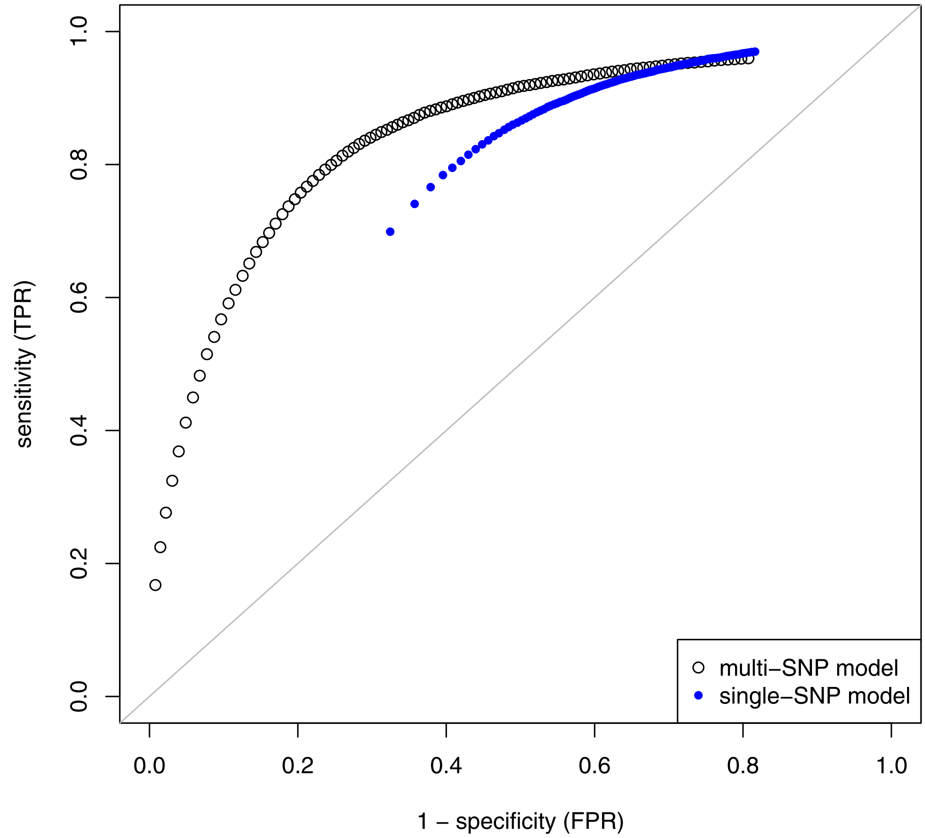

Fig. 3 Sensitivity and specificity of testing SNP effects. ROC curve is based on 100 ×100 repeated simulations of genotypes and phenotypes in progeny generation comprising $N = 10$ half-sib families (two QTL signals, $h^2 = 0.1$). Optimal sample size suggested by the multi-SNP model was considered for setting up the progeny generation

optimal sample size based on a SNP-BLUP approach. The outcome was compared to a single-SNP model. Negative correlations between SNPs, which were mainly due to negative maternal LD, caused essentially inflated sample size estimates. In case of positive correlations, the majority of sample size estimates was less than sample size estimated from the single-SNP approach. The less the heritability, the higher the deviation between models was.

Population parameters

Our approach is applicable to any population structure. The matrix K of covariance between SNPs can be set up for any kind of family stratification by adapting the derivations of the Appendix or, in case of unrelated individuals, by using population LD in K.

Due to the way of model parametrization (columns X_k have been scaled), the dependence on allele frequency has been excluded. For instance, in a random mating population, the column-scaling term is $\sqrt{2p_k(1-p_k)}$ with allele frequency p_k at SNP k. Likewise, a scaling term can be derived for half-sib families as the square root of Eq. (7) in the Appendix by investigating maternally and paternally inherited SNP alleles separately. Results of our association analyses suggested that there was no clear relationship between high n_{opt} and maternal allele frequency or sire heterozygosity (Fig. S.6, Additional file 1). However, regions with large or low variation have to be taken into account when selecting sires for fine-mapping of QTL signals in a follow-up experiment. The lower sire heterozygosity or maternal minor allele frequency is, the lower the effect size β_k on the model scale will be and, consequently, higher n^*_{opt} is required in order to detect QTL signals. To investigate this, we employed the relationship $X_k\beta_k = X_k^{(o)}\beta_k^{(o)}$ at SNP k. Here $X_{j,k}^{(o)}$ is the allele count at SNP k for individual j, and $\beta_k^{(o)}$ is the coefficient on the observed genotype scale, i.e., $\beta_k^{(o)}s_k = \beta_k$ with scaling term $s_k = \sqrt{V(X_{j,k})}$. The relationship between allele frequency and optimal sample size for detecting one QTL signal based on the single-SNP model is presented in Figure S.7 (Additional file 1). Extreme alleles, roughly

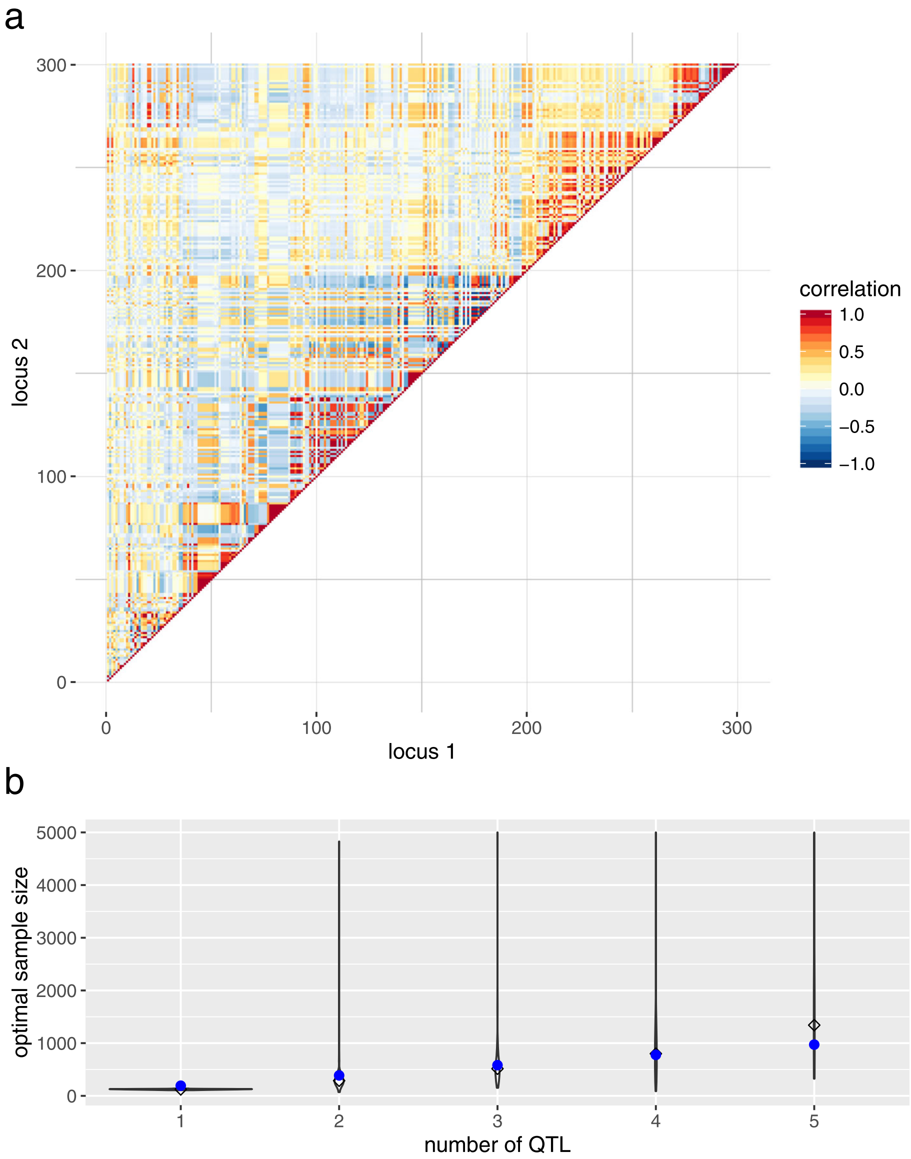

Fig. 4 Empirical bovine HD SNP chip data. **a** Correlation matrix for a randomly selected window containing 300 SNPs on BTA7. **b** Violinplot of n_{opt} vs. number of QTL signals to be detected. The diamond indicates the median of n_{opt} and the blue dots mark the results based on a single-SNP model, $N = 10$ and $h^2 = 0.1$

spoken with major allele frequency > 0.95, require drastically increased sample size.

In the simulation study, equal effects of QTL signals were simulated on the observed genotype scale. For comparison with relative effect sizes derived from heritability (Eqn. 4), allele frequencies in the (random mating) founder population and the residual variance component were used to calculate the corresponding relative effect sizes: $\beta_k^{(o)}\sqrt{2p_k(1-p_k)}/\sigma_e$. Figure 5 shows the effect sizes of all repetitions of simulation with $h^2 = 0.1$ separated by the number of simulated QTL signals (κ). As expected, the relative effect size decreased with increasing κ but a high fluctuation has been observed which was due to the high variation of allele frequencies in the simulated data. This observation underlines the difficulty of detecting multiple QTL signals at given h^2 – the lower the effect size, the higher n_{opt} required.

The suggested optimal sample size is divided into N sires which are selected for most heterogeneity in the target region. The actual number of sires is of minor importance. The choice of individuals depends on the objectives of the follow-up study. Sires can be chosen independently

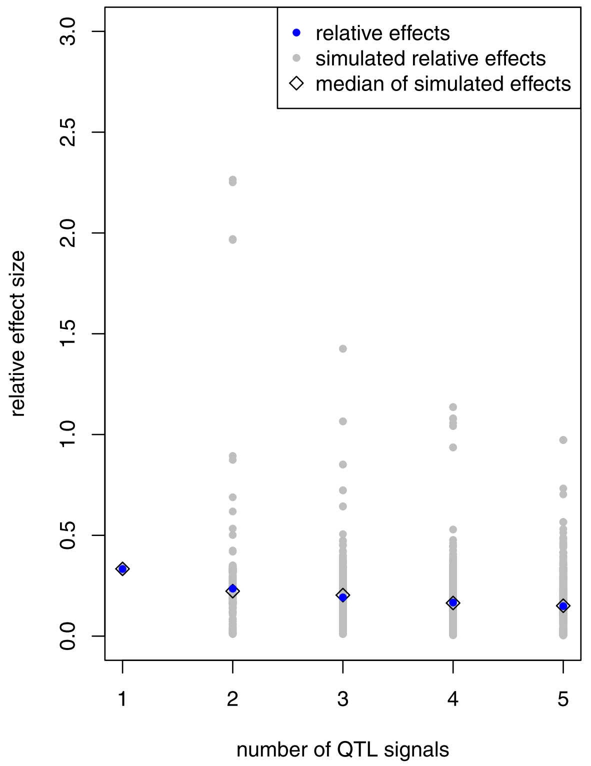

Fig. 5 Relative effect size depending on number of QTL signals. The relative effects (blue dots) were calculated based on the assumption of heritability and number of QTL signals. The simulated relative effects (grey dots) were derived from simulated QTL effects on the observed genotype scale, allele frequencies at positions of QTL signals in the founder population and residual variance component. Simulation was based on $h^2 = 0.1$

from the GWAS population in order to confirm and fine-map QTL signal(s). However, if the initial study indicated the presence of rare variants, sires under suspicion should be re-used. Selective genotyping is an option to increase power [11] but this might have negative impact on reproducibility of the study design [4]. In our investigation of paternal half-sib families, mothers are treated as random samples from a dam population. Thus, the choice of dams for future matings is not addressed here but is definitely an issue for other family designs.

Being equally important for fine-mapping of QTL signals is the positive correlation between SNPs. Positive correlatedness is a matter of genotype coding. Coding has to be consistent throughout the target region to avoid unnecessary sign changes in correlation. We employed coding in terms of counting the major allele in the population. But in regions of intermediate frequency, the coding might not be appropriate and hence a dynamic approach of coding the SNP alleles can circumvent negative correlations. A strategy on this is worth further investigation.

Power calculations are needed to quantify and judge the prospects of identifying causative variants with a hypothesized effect size in a particular population. In practice, however, experiments are usually not planned to obtain maximum power but data are regularly collected for purposes of breeding as a standard routine. Thus, experimental designs being theoretically optimal could be compared with available field data to understand the possible shortcomings of such data and to understand differences between theoretical/expected and actually achieved power. Based on the results, decisions can be made whether the amount of data is sufficient or, in case of underpowered experiments, more data should be acquired.

Necessity of fine-mapping of QTL signals using an appropriate design

The QTL databases of livestock species [33] contain information on several thousands QTL for a wide range of traits. This shows that the variability of most of the traits studied has a polygenic origin, with multiple QTL contributing to the overall genetic variance. Despite the number of QTL, only a handful of causal mutations could be detected and verified in the different livestock species [34]. This is partly due to the fact that GWAS show considerable weaknesses in the fine-mapping of QTL signals which are related to the SNP panel requirements for a genomewide distribution and high LD to neighboring markers [5]. Accordingly, these SNPs are usually indicative of a large genomic region that likely comprises the unmeasured causal SNP but does not provide information about the causal variant itself. Statistical methods for fine-mapping have been designed to overcome these issues and perform fine-mapping using the available SNP information from a SNP-chip or GBS (summarized by [5]). However, even these methods require a high SNP density in the region of interest, which favors a targeted sequencing strategy that enable the dissection of QTL regions and increase the chance of detecting causal variants [35]. Major factors to be considered for designing a targeted sequencing study are effect size, the number of causal SNPs, local LD structure and sample size [5]. The approach proposed in this study incorporates information on κ, h^2 and R derived from the data to estimate the optimal sample size (n^*_{opt}) and thus provides all the information needed to design a fine-mapping experiment.

Currently, several fine-mapping studies are based on imputation strategies or the integration of results with functional enrichment analysis to identify promising candidate genes and QTNs (e.g., [36–38]). These approaches largely depend on imputation accuracy and the status of genome annotation, thus limiting the ability to detect causal variants, especially those with a low minor allele frequency [39]. Specific examples for the fine-mapping of important genomic regions with a resequencing strategy are still rare nowadays. Fraser et al. [40] focused in their study on collagenous lectins in horses by resequencing 658 kb DNA consisting of different candidate genes and regulatory regions. Therefore, a case-control design with pooled samples was used and with this approach 113 variants were identified, which differed between the groups. Although the results are promising, the authors concluded that a large-scale genotyping of individual samples is necessary for deeper insights. In this context, and considering that targeted sequencing for a reasonable set of samples is becoming increasingly affordable, an accurate estimate of sample size is advisable.

Other random effects

Association analysis of empirical data with certain pedigree structure requires an additional model term to account for genetic effects beyond the target window (Zu). Then $u = (u_1, \ldots, u_n)$, $u \sim N(0, G\sigma_u^2)$, is the vector of individual genetic effects with suitable relationship matrix G. The calculation of optimal sample size should consider the presence of additional random effects (genetic or environmental) for the design of experiments. For instance, the coefficients of the single-SNP model could be estimated via BLUE. This affects sample size calculation because the variance of the estimator $\widehat{\beta}_k$,

$$V(\widehat{\beta}_k) = \left(X_k' V^{-1} X_k\right)^{-1} \quad \text{with} \quad V = ZZ'\sigma_u^2 + I_n\sigma_e^2 \,,$$

has impact on the distribution of the test statistic. Accounting for V in the denominator of test statistic increases the denominator of non-centrality parameter. However, in order to keep it simple, it would be sufficient to increase σ_e^2 or reduce h^2 appropriately without any other alterations.

It is possible to consider other kinds of genetic effects with the proposed methods. For instance, exploring dominance genetic effects requires only one modification. Instead of coding SNP genotypes for additive effects via $X_{j,k} \in \{1, 0, -1\}$, genotypes can be coded as $X_{j,k} \in \{0, 1, 0\}$ to account for dominance effects. The covariance between dominance effects has been worked out by Bonk et al. [41]. Feeding the Equation (3) with the corresponding dominance correlation matrix will provide estimates of optimal sample size to fine-map QTL signals with dominance effect.

Conclusion

For planning the design of experiment, we recommend a multi-SNP approach which considers the expected dependence among SNPs. Compared to a conventional approach, this leads to a reduced estimate of sample size and thus promises a more efficient use of animal resources. The benefit depends strongly on heritability: the lower heritability, the more resources can be saved. In general, optimal sample size increases almost linearly with the number of QTL signals to be detected. This study constitutes a frequentist framework for the design of experiments in specific populations that may be characterized by family stratification. It will help differentiating independent signals in QTL regions that can be further examined for cellular and molecular properties.

Appendix: Derivation of correlation matrix

We study the dependence between pairs of SNPs, each with two alleles A and B, in a population consisting of N paternal half-sib families. Let $X_{j,k}$ be the genotype code at SNP $k \in \{1, \ldots, p\}$ of individual $j \in \{1, \ldots, n\}$ being progeny of sire s and dam d. Homozygous genotypes A/A and B/B are coded as 1 and -1, respectively, and the heterozygous genotype A/B is indicated as 0. The family-specific (i.e., sire-specific) covariance between SNPs k and l of individual j is, according to Bonk et al. [41] and Wittenburg et al. [42],

$$\begin{aligned} K_{k,l}^s &= E(X_{j,k}X_{j,l}|\mathcal{S}_s) - E(X_{j,k}|\mathcal{S}_s)E(X_{j,l}|\mathcal{S}_s) \\ &= D_{k,l}^d + D_{k,l}^s \end{aligned}$$

a function of maternal and paternal contribution and depends on the sire diplotype $\mathcal{S}_s$. The $D_{k,l}^d$ denotes the LD of maternal gametes in a dam population. The sire term depends on the phase of paternal haplotypes and recombination rate ($\theta_{k,l}$). It is determined as

$$D_{k,l}^s = \begin{cases} \frac{1}{4}(1-2\theta_{k,l})\,, & \text{for sire with haplotypes A-A and B-B} \\ -\frac{1}{4}(1-2\theta_{k,l})\,, & \text{for sire with haplotypes A-B and B-A} \\ 0\,, & \text{else}\,. \end{cases} \quad (5)$$

To achieve the covariance between a pair of SNPs, we employ conditioning on families,

$$\begin{aligned} E(X_{j,k}X_{j,l}) &= \sum_{s=1}^{N} \Pr(\mathcal{S}_s)E(X_{j,k}X_{j,l}|\mathcal{S}_s) \\ E(X_{j,k}) &= \sum_{s=1}^{N} \Pr(\mathcal{S}_s)E(X_{j,k}|\mathcal{S}_s) \\ \operatorname{cov}(X_{j,k}, X_{j,l}) &= \sum_{s=1}^{N} w_s E(X_{j,k}X_{j,l}|\mathcal{S}_s) \\ &\quad - \sum_{s=1}^{N} w_s E(X_{j,k}|\mathcal{S}_s) \sum_{s=1}^{N} w_s E(X_{j,l}|\mathcal{S}_s) \end{aligned}$$

and approximate $\Pr(\mathcal{S}_s)$ by family weights $w_s = \frac{n_s}{n}$ with $\sum_{s=1}^{N} n_s = n$. The aim is now to derive an expression that depends on already known terms. For instance, using

$$E(X_{j,k}X_{j,l}|\mathcal{S}_s) = K_{k,l}^s + E(X_{j,k}|\mathcal{S}_s)E(X_{j,l}|\mathcal{S}_s)$$

yields

$$E(X_{j,k}X_{j,l}) = \sum_{s=1}^{N} w_s \left(K^s_{k,l} + E(X_{j,k}|\mathcal{S}_s)E(X_{j,l}|\mathcal{S}_s)\right) .$$

We exploit the separation into independently inherited maternal and paternal SNP alleles: $X_{j,k} = X_{j,k,s} + X_{j,k,d}$, where $X_{j,k,s}$ and $X_{j,k,d}$ take a value of $\frac{1}{2}$ if the A allele was inherited but $-\frac{1}{2}$ otherwise. Then

$$\begin{aligned} E(X_{j,k}|\mathcal{S}_s) &= E(X_{j,k,d}|\mathcal{S}_s) + E(X_{j,k,s}|\mathcal{S}_s) \\ E(X_{j,k,d}|\mathcal{S}_s) &= E(X_{j,k,d}) \;=\; p_k - \frac{1}{2} , \end{aligned}$$

where p_k denotes the maternal allele frequency at SNP k. Furthermore,

$$E(X_{j,k,s}|\mathcal{S}_s) = \begin{cases} \frac{1}{2} , & \text{for sire genotype A/A} \\ 0 , & \text{for sire genotype A/B} \\ -\frac{1}{2} , & \text{for sire genotype B/B} . \end{cases} \qquad (6)$$

Putting it all together,

$$\begin{aligned} K_{k,l} = & \sum_{s=1}^{N} w_s (D^d_{k,l} + D^s_{k,l}) \\ & + \sum_{s=1}^{N} w_s \left[\left(p_k - \frac{1}{2}\right) + E(X_{j,k,s}|\mathcal{S}_s)\right] \\ & \left[\left(p_l - \frac{1}{2}\right) + E(X_{j,l,s}|\mathcal{S}_s)\right] \\ & - \sum_{s=1}^{N} w_s \left[\left(p_k - \frac{1}{2}\right) + E(X_{j,k,s}|\mathcal{S}_s)\right] \\ & \sum_{s=1}^{N} w_s \left[\left(p_l - \frac{1}{2}\right) + E(X_{j,l,s}|\mathcal{S}_s)\right] . \end{aligned}$$

This reduces to

$$\begin{aligned} K_{k,l} = & D^d_{k,l} + \sum_{s=1}^{N} w_s D^s_{k,l} + \sum_{s=1}^{N} w_s E(X_{j,k,s}|\mathcal{S}_s) E(X_{j,l,s}|\mathcal{S}_s) \\ & - \sum_{s=1}^{N} w_s E(X_{j,k,s}|\mathcal{S}_s) \sum_{s=1}^{N} w_s E(X_{j,l,s}|\mathcal{S}_s) , \end{aligned}$$

and this is evaluated using the sire-specific terms in (5) and (6).

Now the variance of genotype codes at SNP k is derived explicitly – it also serves as a scaling term in a regression model for association analysis. The second moment of the paternally inherited SNP allele is constant $E(X^2_{j,k,s}|\mathcal{S}_s) = \frac{1}{4}$ for all s. Hence

$$V(X_{j,k,s}) = \frac{1}{4} - \left(\sum_{s=1}^{N} w_s E(X_{j,k,s}|\mathcal{S}_s)\right)^2 .$$

Then, the variance at SNP k is

$$\begin{aligned} V(X_{j,k}) &= V(X_{j,k,d}) + V(X_{j,k,s}) \\ &= p_k(1-p_k) + \frac{1}{4} - \left(\sum_{s=1}^{N} w_s E(X_{j,k,s}|\mathcal{S}_s)\right)^2 \;=\; K_{k,k} . \end{aligned} \qquad (7)$$

Finally, the correlation matrix $R = \{R_{k,l}\}_{k,l=1,\ldots,p}$ is calculated by scaling the entries correspondingly,

$$R_{k,l} = \frac{K_{k,l}}{\sqrt{K_{k,k}K_{l,l}}} .$$

Note that the covariance based on non-centered genotype codes (as derived above) is identical to the one based on centered genotype codes (as used in Methods section). Centering is used to study within-family genetic effects, and it allows the direct estimation of allele substitution effects [43].

The R package `hscovar` for the calculation of K and R is provided at CRAN repository.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12863-020-00871-1.

Additional file 1: PDF file containing additional figures.

Additional file 2: R script for data simulation and analysis. Running this script performs the complete data analysis and evaluation. It calls functions from Additional file 3.

Additional file 3: R script including functions used for data simulation and analysis.

Abbreviations

ANOVA: Analysis of variance; BLUE: Best linear unbiased estimation; BTA: Bos taurus autosome; cM: CentiMorgan; DNA: Deoxyribonucleic acid; FPR: False positive rate; (G)BLUP: (Genomic) best linear unbiased prediction; GBS: Genotyping by sequencing; GWAS: Genomewide association study; HD: High density; LD: Linkage disequilibrium; Mbp: Mega base pairs; QTL: Quantitative trait locus; QTN: Quantitative trait nucleotide; ROC: Receiver operating characteristic; SNP: Single nucleotide polymorphism; TPR: True positive rate

Acknowledgments

Special thanks are given to N. Reinsch (Leibniz Institute for Farm Animal Biology, Dummerstorf), J. Hartung (University of Hohenheim, Stuttgart) and V. Liebscher (University of Greifswald), who contributed invaluable ideas to the project, and to C. Gaynor (Roslin Institute, Edingburgh) for giving handy insight into `AlphaSimR`. We also thank the reviewers for their helpful comments.

Authors' contributions

DW developed the theory, implemented the statistical methods, performed the analysis, and wrote the manuscript. SB contributed to the research on covariance between SNPs, MD was involved in theoretical investigations. HR contributed to the discussion. All authors have read and approved the final manuscript.

Funding

The project was funded by the German Research Foundation (DFG, WI 4450/1-1). The funder had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

R scripts used to simulate and analyze data are available in Additional files 2 and 3. However, an exact reproduction of the results of our simulation study is not possible because `AlphaSimR` relies on random number generation which cannot be initialized yet. The R package `hscovar` version 0.2.1 is available at CRAN; it provides tools for setting up the covariance or correlation matrix as well as for performing power calculations. The empirical bovine HD SNP chip data are accessible through the Dryad repository https://doi.org/10.5061/dryad.519bm [31].

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Leibniz Institute for Farm Animal Biology, Institute of Genetics and Biometry, 18196 Dummerstorf, Germany. [2]University Medicine Greifswald, Department of Psychiatry and Psychotherapy, 17475 Greifswald, Germany. [3]Leibniz Institute for Farm Animal Biology, Institute of Genome Biology, 18196 Dummerstorf, Germany.

Received: 21 February 2020 Accepted: 9 June 2020
Published online: 29 June 2020

References

1. Reyer H, Hawken R, Murani E, Ponsuksili S, Wimmers K. The genetics of feed conversion efficiency traits in a commercial broiler line. Sci Rep. 2015;5:16387.
2. Sahana G, Guldbrandtsen B, Thomsen B, Holm LE, Panitz F, Brøndum RF, et al. Genome-wide association study using high-density single nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle. J Dairy Sci. 2014;97(11):7258–75.
3. Hampel A, Teuscher F, Gomez-Raya L, Doschoris M, Wittenburg D. Estimation of recombination rate and maternal linkage disequilibrium in half-sibs. Front Genet. 2018;9:186.
4. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95(1):5–23.
5. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. 2018;19(8):491–504.
6. Gauderman J, Morrison J. QUANTO Version 1.2. 2007. Retrieved June 10, 2015. Available from: http://biostats.usc.edu/Quanto.html.
7. Schnabel R. ARS-UCD1.2 Cow Genome Assembly: Mapping of all existing variants. 2018. Retrieved Sep 21, 2018. Available from: https://www.animalgenome.org/repository/cattle/UMC_bovine_coordinates/.
8. Luo Z. Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. Heredity. 1998;80(2):198.
9. Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. Am J Hum Genet. 2001;69(1):1–14.
10. Khatkar MS, Nicholas FW, Collins AR, Zenger KR, Cavanagh JA, Barris W, et al. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. BMC Genomics. 2008;9(1):187.
11. Weller J. Quantitative trait loci analysis in animals: CABI Publishing; 2001. https://doi.org/10.1079/9781845934675.0000.
12. Gualdrón Duarte JL, Cantet RJ, Bates RO, Ernst CW, Raney NE, Steibel JP. Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. BMC Bioinf. 2014;15(1):246. Available from: https://doi.org/10.1186/1471-2105-15-246.
13. Koivula M, Strandén I, Su G, Mäntysaari EA. Different methods to calculate genomic predictions—Comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). J Dairy Sci. 2012;95(7):4065–73.
14. Mucha S, Mrode R, MacLaren-Lee I, Coffey M, Conington J. Estimation of genomic breeding values for milk yield in UK dairy goats. J Dairy Sci. 2015;98(11):8201–8.
15. Maier R, Moser G, Chen GB, Ripke S, Absher D, Agartz I, et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am J Hum Genet. 2015;96(2):283–94.
16. Kristensen PS, Jahoor A, Andersen JR, Cericola F, Orabi J, Janss LL, et al. Genome-wide association studies and comparison of models and cross-validation strategies for genomic prediction of quality traits in advanced winter wheat breeding lines. Front Plant Sci. 2018;9:69.
17. Taskinen M, Mäntysaari EA, Strandén I. Single-step SNP-BLUP with on-the-fly imputed genotypes and residual polygenic effects. Genet Sel Evol. 2017;49(1):36.
18. Aguilar I, Legarra A, Cardoso F, Masuda Y, Lourenco D, Misztal I. Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. Genet Sel Evol. 2019;51(1):28.
19. Searle S. Linear models. New York: Wiley; 1971.
20. Hoerl AE, Kennard RW, Baldwin KF. Ridge regression: some simulations. Commun Stat Theor M. 1975;4(2):105–23.
21. Cohen J. Statistical power analysis for the social sciences. Hillsdale: Erlbaum; 1988.
22. Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet Epidemiol. 2008 May;32:361–9.
23. R Core Team. R: A Language and Environment for Statistical Computing. Vienna; 2019. Retrieved Dec 16, 2019. Available from: https://www.R-project.org/.
24. Faux AM, Gorjanc G, Gaynor RC, Battagin M, Edwards SM, Wilson DL, et al. AlphaSim: software for breeding program simulation. Plant Genome. 2016;9(3):1–14. Available from: https://doi.org/10.3835/plantgenome2016.02.0013.
25. Chen GK, Marjoram P, Wall JD. Fast and flexible simulation of DNA sequence data. Genome Res. 2009;19(1):136–42.
26. Butler D, Cullis BR, Gilmour A, Gogel B. ASReml-R reference manual. Brisbane: The State of Queensland, Department of Primary Industries and Fisheries; 2009.
27. Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome. 2011;4(3):250–5.
28. Cule E, Vineis P, De Iorio M. Significance testing in ridge regression for genetic data. BMC Bioinf. 2011;12:372.
29. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong Sy, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42(4):348.
30. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B. 1995;57(1):289–300.
31. Bermingham ML, Bishop SC, Woolliams JA, Pong-Wong R, Allen AR, McBride SH, et al. Data from: Genome-wide association study identifies novel loci associated with resistance to bovine tuberculosis. Dryad, Dataset. 2013. Available from: https://doi.org/10.5061/dryad.519bm.
32. Hickey JM, Kinghorn BP, Tier B, Wilson JF, Dunstan N, van der Werf JH. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet Sel Evol. 2011;43(1):12.
33. Hu ZL, Park CA, Wu XL, Reecy JM. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Res. 2012;41(D1):D871—9.
34. Andersson L, Georges M. Domestic-animal genomics: deciphering the genetics of complex traits. Nat Rev Genet. 2004;5(3):202.
35. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010;7(2):111.
36. Jiang J, Cole JB, Freebern E, Da Y, VanRaden PM, Ma L. Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls. Commun Biol. 2019;2(1):212.
37. Cai Z, Guldbrandtsen B, Lund MS, Sahana G. Weighting sequence variants based on their annotation increases the power of genome-wide association studies in dairy cattle. Genet Sel Evol. 2019;51(1):20.
38. Liu Z, Wang T, Pryce JE, MacLeod IM, Hayes BJ, Chamberlain AJ, et al. Fine-mapping sequence mutations with a major effect on oligosaccharide content in bovine milk. Sci Rep. 2019;9(1):2137.
39. Dadaev T, Saunders EJ, Newcombe PJ, Anokian E, Leongamornlert DA, Brook MN, et al. Fine-mapping of prostate cancer susceptibility loci in a large meta-analysis identifies candidate causal variants. Nat Commun. 2018;9(1):2256.

40. Fraser RS, Arroyo LG, Meyer A, Lillie BN. Identification of genetic variation in equine collagenous lectins using targeted resequencing. Vet Immunol Immunopathol. 2018;202:153–63.
41. Bonk S, Reichelt M, Teuscher F, Segelke D, Reinsch N. Mendelian sampling covariability of marker effects and genetic values. Genet Sel Evol. 2016;48(1):36.
42. Wittenburg D, Teuscher F, Klosa J, Reinsch N. Covariance between genotypic effects and its use for genomic inference in half-sib families. G3 Genes Genom Genet. 2016;6:2761–72.
43. Abecasis GR, Cardon LR, Cookson W. A general test of association for quantitative traits in nuclear families. Am J Hum Genet. 2000;66(1):279–92.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.