Genet. Res., Camb. (2017), vol. 99, e4. © Cambridge University Press 2017
doi:10.1017/S0016672317000027

Fine mapping by composite genome-wide association analysis

JOAQUIM CASELLAS[1]*, JHON JACOBO CAÑAS-ÁLVAREZ[2], MARTA FINA[2], JESÚS PIEDRAFITA[2] AND ALESSIO CECCHINATO[3]

[1] *Grup de Recerca en Millora Genètica Molecular Veterinària, Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain*
[2] *Grup de Recerca en Remugants, Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain*
[3] *Department of Agronomy, Food, Natural Resources, Animals and Environment (DAFNAE), University of Padova, Viale dell'Università 16, 35020 Legnaro, Italy*

(*Received 26 July 2016; revised 22 February 2017; accepted 7 March 2017*)

Summary

Genome-wide association (GWA) studies play a key role in current genetics research, unravelling genomic regions linked to phenotypic traits of interest in multiple species. Nevertheless, the extent of linkage disequilibrium (LD) may provide confounding results when significant genetic markers span along several contiguous cM. In this study, we have adapted the composite interval mapping approach to the GWA framework (composite GWA), in order to evaluate the impact of including competing (possibly linked) genetic markers when testing for the additive allelic effect inherent to a given genetic marker. We tested model performance on simulated data sets under different scenarios (i.e., qualitative trait loci effects, LD between genetic markers and width of the genomic region involved in the analysis). Our results showed that the genomic region had a small impact on the number of competing single nucleotide polymorphisms (SNPs) as well as on the precision of the composite GWA analysis. A similar conclusion was derived from the preferable range of LD between the tested SNP and competing SNPs, although moderate-to-high LD seemed to attenuate the loss of statistical power. The composite GWA improved specificity and reduced the number of significant genetic markers. The composite GWA model contributes a novel point of view for GWA analyses where testing circumscribed to the genomic region flanking each SNP (delimited by the nearest competing SNPs) and conditioning on linked markers increases the precision to locate causal mutations, but possibly at the expense of power.

1. Introduction

Genome-wide association (GWA) analyses are studies where genomic variation measured, often by single nucleotide polymorphism (SNP) markers, is correlated across production, health and other traits of interest to identify candidate loci that regulate them. They must be viewed as the natural evolution of candidate gene association studies (Singer, 2009), although they focus on thousands or millions of SNPs without reference to any particular gene (Benyamin, 2009). Genomic data has been released by the human genome project (International HapMap Consortium, 2005; Sachidanandam *et al.*, 2001), and other livestock (International Chicken Genome Sequencing Consortium, 2004), laboratory (Gibbs *et al.*, 2004) and wild species (Li *et al.*, 2010; Scally *et al.*, 2012) genome projects. Since the first successful GWA study published in 2005 (Klein *et al.*, 2005), this methodology has represented a key tool for the study of common genetic variations in complex traits.

As noted by Wang *et al.* (2010), GWA studies have succeeded in the identification of phenotype-associated genetic markers, but pinpointing causal mutations in subsequent fine-mapping studies remains a challenge. Despite marker SNPs not being the causal mutation (Wang, 2010), GWA methodology relies on the assumption that (*a*) linkage disequilibrium (LD) would enable one or few SNPs to act as surrogate markers for association and (*b*) these markers would be placed near to the causal genetic variant. Nevertheless, the extent of LD in mammalian genomes (Tenesa *et al.*, 2004; Sargolzaei *et al.*, 2008) are used to reveal significant SNPs across several

* Corresponding author: joaquim.casellas@uab.cat

contiguous cM. As a consequence, this extends the genomic region potentially harbouring causal mutations and enlarges the list of candidate genes to be tested. Moreover, current LD between SNPs may lead to marginally associated effects, even when not in direct LD with the causal mutation (He & Lin, 2011). The unprecedented potential for false-positives shown by GWAs (Pearson & Manolio, 2008) must be viewed as a controversial challenge inherent to this methodology.

Analytical approaches for GWA studies must appropriately account for LD among genetic markers. In this article, the methodology developed by Zeng (1994) and Jansen & Stam (1994) for quantitative trait loci (QTLs) mapping has been adapted to improve both the precision and efficiency of GWA studies. The main idea relies on the inclusion of additional (possibly linked) genetic markers when testing for a specific marker; this must benefit from the statistical properties of multiple regression analysis, which were previously reviewed by Rodolphe & Lefort (1993) and Zeng (1993; 1994) within the context of QTL analysis. Nevertheless, dissimilarities between GWA and QTL analyses (Kemper *et al.*, 2012) evidence that previous advantages reported for linkage analysis methods (Zeng, 1994) cannot be directly extrapolated to GWA approaches and statistical properties inherent to our modified GWA approach must be assessed in detail.

This article focuses on two major objectives. First, the multiple regression analysis from Zeng (1994) and Jansen & Stam (1994) has been adapted to the GWA framework. The analytical approach was implemented in Fortran90 programs and is available upon request from the first author of this article (J. Casellas). Second, the statistical performance of this modified GWA methodology has been evaluated on simulated data sets by testing different scenarios; different simulation (e.g., QTL effects and allelic frequencies) and analytical parameters (e.g., LD between competing SNPs and genomic regions involved in the analysis) were evaluated.

2. Materials and methods

(i) *Composite GWA analysis*

Take as a starting point a sample of n individuals with phenotypic information for a given quantitative trait. Moreover, assume that all individuals are genotyped for m biallelic genetic markers, and these markers are more or less evenly distributed across the genome. Under a standard approach, the analysis of the additive association effect inherent to the kth marker can be carried out by the following model:

$$y_i = \mu + \beta_k x_{ik} + e_i$$

where y_i is the phenotypic record collected from the ith individual, μ is the population mean, β_k is the additive association effect of the kth marker, x_{ik} is an indicator variable taking values of -1 (homozygote), 0 (heterozygote) and 1 (opposite homozygote), and e_i is the residual term. Within the context of a composite GWA, previous model generalizes to:

$$y_i = \mu + \beta_k x_{ik} + \Sigma_J \beta_{j*} x_{ij} + e_i$$

where β_{j*} is the partial regression coefficient of the jth marker in set J, and x_{ij} is an indicator variable (see x_{ik}). Focusing on a given marker k, note that β_k and β_{k*} are both regression coefficients, although their interpretation becomes quite different. Whereas β_k estimates the effect of the kth genetic marker on the phenotypic trait and after accounting for the remaining competing markers, β_{k*} must be viewed as a nuisance parameter. From a general point of view, we assume that any marker (i.e., j) included in J must satisfy that (*a*) $j \neq k$, (*b*) marker j is located no farther away from k than δ cM (i.e., analytical window) and (*c*) the LD between markers j and k falls within a range of values with predefined boundaries τ_1 and τ_2 ($0 \leqslant \tau_1 < \tau_2 \leqslant 1$). Despite additional sources of variation in the previous model summarized into the μ term, this model can expand to accommodate additional factors influencing the phenotypic trait.

(ii) *Simulation process*

Each simulated population evolved without selection during 1000 non-overlapping generations with effective population size (N_e) 100. In order to mimic a polygynous-like species, which is common under current livestock practices, generation 1001 expanded up to 1000 individuals, with 200 males and 800 females. Note that this design expanded N_e up to 640 individuals (Wright, 1931). Generation 1001 was randomly mated to obtain 1000 individuals in generation 1002.

Each individual had a 100-cM chromosome with 2000 biallelic SNPs (one SNP each 0·05 cM) and a unique QTL located in cM 50. This initial density of SNPs matched previous research (Habier *et al.*, 2009; Casellas & Varona, 2011) and fell within the range of lower (⩽500 markers/M; Meuwissen *et al.*, 2001; Ødegård *et al.*, 2009) and higher (6000 to ~10 000 SNPs/M; Ibáñez-Escriche *et al.*, 2009; Toosi *et al.*, 2010) SNP densities reported in the scientific literature. Founder individuals were homozygous throughout the whole genome for the wild-type allele (i.e., allele 1), and this switched from allele 1 to 2 (or *vice versa*) by appropriate mutation rates. The QTL was affected by a mutation rate of $2{\cdot}5 \times 10^{-5}$ in all generations (Meuwissen *et al.*, 2001), whereas SNPs had a mutation rate of $2{\cdot}5 \times 10^{-3}$ from generation 1 to 900 (Meuwissen *et al.*, 2001) to guarantee a high percentage of polymorphic markers. This parameter reduced to a more realistic $2{\cdot}5 \times 10^{-8}$ for subsequent generations (Hickey & Gorjanc, 2012). Chromosome recombination was ruled by Kosambi's function (Kosambi, 1943).

Genomic data from all individuals born in generation 1002 were stored and checked. The minimum allele frequency (MAF) was calculated for each marker in order to validate the two following restrictions: (*a*) QTLs with MAF $\geqslant 0{\cdot}25$, and (*b*) 900 to 1100 SNPs with MAF $\geqslant 0{\cdot}05$. Only those populations satisfying these criteria were retained for further analyses. The restriction applied on the QTLs aimed to narrow the impact of the genetic variability contributed by the QTLs on the overall phenotypic variance, whereas the restriction on the number of polymorphic SNPs tried to homogenize the number of potential competing genetic markers across populations (see below). Given that 100 populations were required for each scenario (see below) and the rejection rate could not be anticipated, simulations were performed back to back until 100 valid populations were available. At the end of the simulation process, the rejection rate was 79%. A unique phenotypic record was generated for each individual born in generation 1002. This resulted from the additive allelic effects from the QTLs and a random value sampled from a standard normal distribution with mean 0 and variance 1. Four different additive allelic effects (α) were assumed for the mutant-type allele ($\alpha = 0$, 0·25, 0·5 and 1), whereas the additive effect of the wild-type allele was null. These values were assumed in order to simulate QTLs with small ($h^2 \sim 0{\cdot}03$), moderate ($h^2 \sim 0{\cdot}13$) and large ($h^2 \sim 0{\cdot}33$) contributions to the phenotypic variance.

(iii) *Analytical process*

Different scenarios were generated by combining the additive allelic effect of the QTLs, analytical window and LD range between tested and competing SNPs. The LD between SNPs (r^2) was calculated as the squared correlation of the alleles (Hill & Robertson, 1968). Both analytical window (10, 30 and 50 cM on each side of the tested SNP) and LD range ($0{\cdot}1 \leqslant r^2 \leqslant 0{\cdot}9$, $0{\cdot}1 \leqslant r^2 \leqslant 0{\cdot}5$ and $0{\cdot}5 \leqslant r^2 \leqslant 0{\cdot}9$) had three different values (or ranges), leading to a total of 36 combinations. It is important to note that SNPs with high LD ($r^2 > 0{\cdot}9$) were discarded to prevent identifiability problems in the analytical model, whereas SNPs with low LD ($r^2 < 0{\cdot}1$) were also discarded to restrict the number of competing SNPs included in the model.

Within each population, SNP-by-SNP analyses were performed twice by applying both the standard GWA and the composite GWA models described above. This duplicate analysis aimed to characterize the statistical performance of composite GWA analysis against a well-known analytical approach. Both models were solved by Gauss-Seidel (Mrode, 2005) and significance of β_k was tested by a likelihood ratio test (Neyman & Pearson, 1933) with one degree of freedom. Results were discussed on the basis of three different levels of significance. Although ~1000 SNPs were tested within a population, the standard (and uncorrected for multiple testing) $p < 0{\cdot}05$ was assumed as upper boundary of significance. On the contrary, Bonferroni's (1936) correction assuming 1000 independent tests was applied as lower boundary of significance ($p < 0{\cdot}00005$). Between them, an intermediate significance level was defined by the Benjamini & Hochberg (1995) approach with (on average) $p < 0{\cdot}0005$.

From each simulation scenario, differences between the composite GWA model and the standard GWA model were evaluated in terms of statistical power (i.e., probability of identifying significantly associated SNPs when $\alpha > 0$) and specificity (i.e., probability of no SNPs being significant when $\alpha > 0$) on a chromosomal level, as well as precision. More specifically, precision was evaluated in terms of the total number of significant SNPs, the average absolute distance between significant SNPs and the QTL, and the percentage of significant SNPs located not father than 2·5 cM from the QTL.

3. Results

(i) *Power and specificity*

On the basis of the simulation process described above, the average number of competing SNPs varied depending on LD requirements and slightly decreased with the width of the analytical window where competing SNPs were assessed (Fig. 1). The wider range of LD between the tested and competing SNPs ($0{\cdot}1 \leqslant r^2 \leqslant 0{\cdot}9$) included an average of ~42 competing SNPs into the model, although minimum and maximum estimates for the within-chromosome average number of competing SNPs were 34·0 and 50·3, respectively. On average, ~36 competing SNPs were accounted for in the model when LD was restricted to $0{\cdot}1 \leqslant r^2 \leqslant 0{\cdot}5$, whereas higher LD ($0{\cdot}5 \leqslant r^2 \leqslant 0{\cdot}9$) reduced the average number of competing SNPs to ~6 (Fig. 1). It is important to note that differences on the basis of the width of the genomic window where SNPs were assessed were minimum.

Competing SNPs included in the analytical model must be viewed as a relevant increase of the number of parameters to be inferred; therefore, they may influence both the power (i.e., probability of identifying significantly associated SNPs when $\alpha > 0$) and specificity (i.e., probability of no SNP being significant when $\alpha > 0$) of the test. Power decreased for larger numbers of competing SNPs and, as anticipated, power

. https://doi.org/10.1017/S0016672317000027

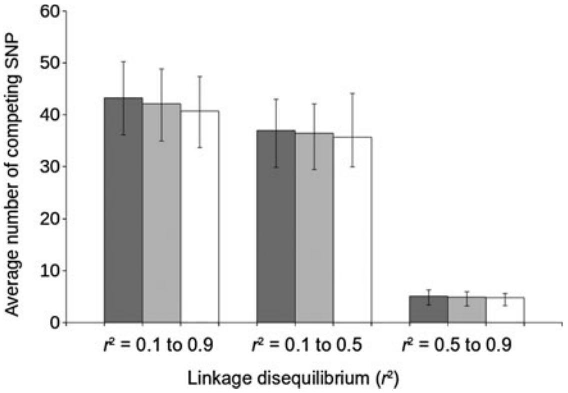

Fig. 1. Average number of competing SNPs included in the composite genome-wide association studies analysis; the whiskers extend the range of the results. Columns are organized in three independent groups depending on the linkage disequilibrium (r^2) between competing SNPs and the QTL; within-group colour differences identify the size of the genomic region where competing SNPs were assessed, this being 10 cM (white), 30 cM (light grey) and 50 cM (dark grey) on each side of the tested SNP.

increased with the magnitude of the QTLs effect (Table 1). Maybe more relevant than these trends, we must put a special emphasis on the comparison between composite GWA and standard GWA. The smallest number of model parameters in standard GWA analyses provided the highest power, and this only failed to reveal significantly ($p < 0{\cdot}0005$) associated SNPs in some simulated populations when $\alpha = 0{\cdot}25$ (Table 1). On the contrary, competing SNPs became a penalization for composite GWA in terms of statistical power, as evidence for small-effect QTL ($\alpha = 0{\cdot}25$). This was attenuated for medium-effect QTLs ($\alpha = 0{\cdot}50$) and composite GWA almost mimicked the statistical power of standard GWA analyses when testing for large-effect QTLs ($\alpha = 1{\cdot}00$) (Table 1).

Specificity was improved under composite GWA, this approach discards significant ($p < 0{\cdot}0005$) associations in all replicates regardless of LD range and analytical window; standard GWA had 94% specificity, whereas it identified one or more significant ($p < 0{\cdot}0005$) SNPs in 6% of the simulated populations under null QTLs effects (Table 1). This was even more drastic if multiple testing correction was not applied. On average, the composite GWA approach reached a ~20% specificity, whereas standard GWA returned significant ($p < 0{\cdot}05$) SNPs from all simulated populations.

(ii) *Refining QTL-associated genomic regions*

The average number of significant ($p < 0{\cdot}0005$) SNPs (lower and upper boundaries) under standard GWA depended on the magnitude of the simulated QTLs, this being 18·6 (1 to 56) for $\alpha = 0{\cdot}25$, 64·0 (17 to 137) for $\alpha = 0{\cdot}50$ and 205·3 (93 to 318) for $\alpha = 1{\cdot}00$ (Fig. 2). These averages drastically reduced under composite GWA; this approach identified a maximum of seven significant ($p < 0{\cdot}0005$) SNPs when $\alpha = 0{\cdot}25$, increasing up to 30 SNPs for $\alpha = 0{\cdot}50$. Under large-effect QTLs, the average number of significant SNPs was less than a third of the number of significant SNPs under standard GWA. Moreover, this scenario revealed remarkable differences depending on the range of LD for competing SNPs. The average number of significant SNPs clearly increased when $0{\cdot}5 \leqslant r^2 \leqslant 0{\cdot}9$ (~63 SNP), whereas $0{\cdot}1 \leqslant r^2 \leqslant 0{\cdot}5$ revealed ~15 significant SNPs and $0{\cdot}1 \leqslant r^2 \leqslant 0{\cdot}9$ showed reductions of up to ~8 significant SNPs (Fig. 2).

Results on the average of the absolute distance between the QTL and every significant SNP is shown in Fig. 3. This was larger for $\alpha = 1$ than for smaller QTLs effects and slightly increased for smaller analytical windows. Differences between standard and composite GWA were almost negligible under $\alpha \leqslant 0{\cdot}5$. Only QTLs with $\alpha = 1$ suggested that standard GWA approaches associated more distant SNPs (from the QTL; 11·8 cM) than composite GWA (~8·8 cM), although lower and upper boundaries overlapped (Fig. 3). A similar pattern was revealed when checking the percentage of significant QTLs located not farther than 2·5 cM from the QTL (Fig. 4). Small- and medium-effect QTLs did not reveal relevant differences between standard and composite GWA (results not shown), whereas simulations under $\alpha = 1$ suggested advantages when applying composite GWA. On average, the standard GWA identified 20·8% of the significant SNPs in the nearest 2·5 cM around the QTL, whereas this average percentage rose to values larger than 30% when applying composite GWA analysis (Fig. 4).

4. Discussion

This research contributes a novel approach for GWA analyses that increases the precision for locating causal mutations but at the expense analytical power. Accurate genome-wide association methodologies are of special relevance in the current genomics era, where large amounts of sequence data are becoming available. The composite GWA approach described in this article focuses on the main idea of including additional (i.e., competing) genetic markers when testing for association effects inherent to a given SNP; although it could be viewed as an over-parameterization of the analytical model, this approach tries to narrow the genomic region where QTL-associated effects can be detected by appropriate SNPs in LD with the causal mutation. Competing SNPs must account for marginally associated effects as was

Table 1. *Percentage of simulated populations without any significant ($p < 0{\cdot}05/p < 0{\cdot}0005/p < 0{\cdot}00005$) SNPs across the whole chromosome.*

			Additive allelic effect (α) for the mutant-type allele of the QTLs (the wild-type allele was assumed with null effect)			
Model[a]	AW[b] (cM)	LD range[c]	$\alpha = 0$	$\alpha = 0{\cdot}25$	$\alpha = 0{\cdot}5$	$\alpha = 1$
GWASc	10	0·1–0·9	19/100/100	1/97/98	0/69/93	0/0/5
GWASc	10	0·1–0·5	19/100/100	0/88/96	0/36/60	0/0/2
GWASc	10	0·5–0·9	18/100/100	0/51/71	0/0/3	0/0/0
GWASc	30	0·1–0·9	22/100/100	3/98/100	0/70/95	0/2/7
GWASc	30	0·1–0·5	20/100/100	2/89/97	0/47/73	0/1/5
GWASc	30	0·5–0·9	18/100/100	0/54/74	0/1/7	0/0/0
GWASc	50	0·1–0·9	23/100/100	4/99/100	0/70/96	0/3/10
GWASc	50	0·1–0·5	22/100/100	2/92/98	0/49/77	0/2/7
GWASc	50	0·5–0·9	19/100/100	1/56/77	0/2/9	0/0/1
GWAS	–[d]	–[d]	0/94/100	0/8/23	0/0/0	0/0/0

[a] GWAS: standard genome-wide association analysis; GWASc: composite genome-wide association analysis.
[b] AW: width of the analytical window where competing SNPs were assessed.
[c] LD range: range of linkage disequilibrium between competing SNPs and the tested SNP.
[d] –: not applicable.

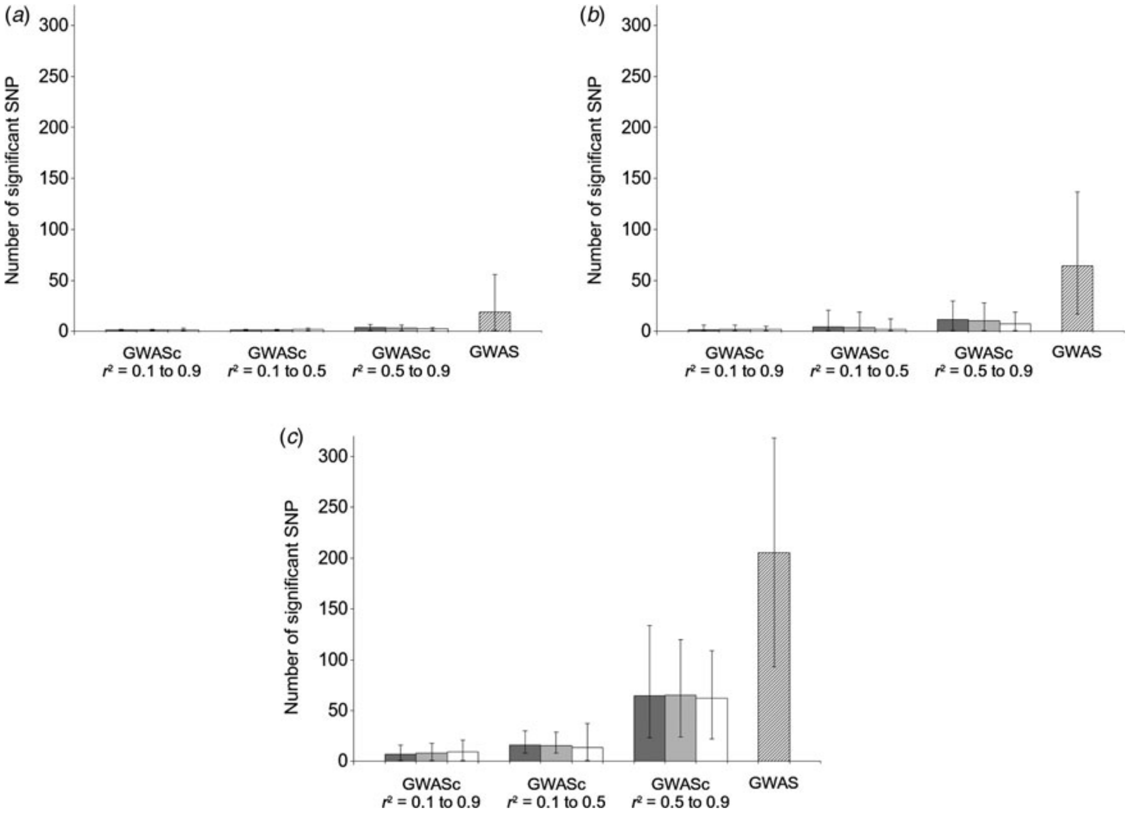

Fig. 2. Average number of significant ($p < 0{\cdot}0005$) SNPs under standard genome-wide association studies (GWAS) analysis and composite GWAS (GWASc) for small-effect QTLs (*a*), medium-effect QTLs (*b*) and large-effect QTLs (*c*); the whiskers extend the range of the results. Columns are organized in four independent groups depending on the analytical approach (GWAS vs. GWASc) and the linkage disequilibrium (r^2) between competing SNPs and the QTL for GWASc analyses; within-group colour differences identify the size of the genomic region where competing SNPs were assessed, this being 10 cM (white), 30 cM (light grey) and 50 cM (dark grey) on each side of the tested SNP. The striped bar corresponds to the standard GWAS approach.

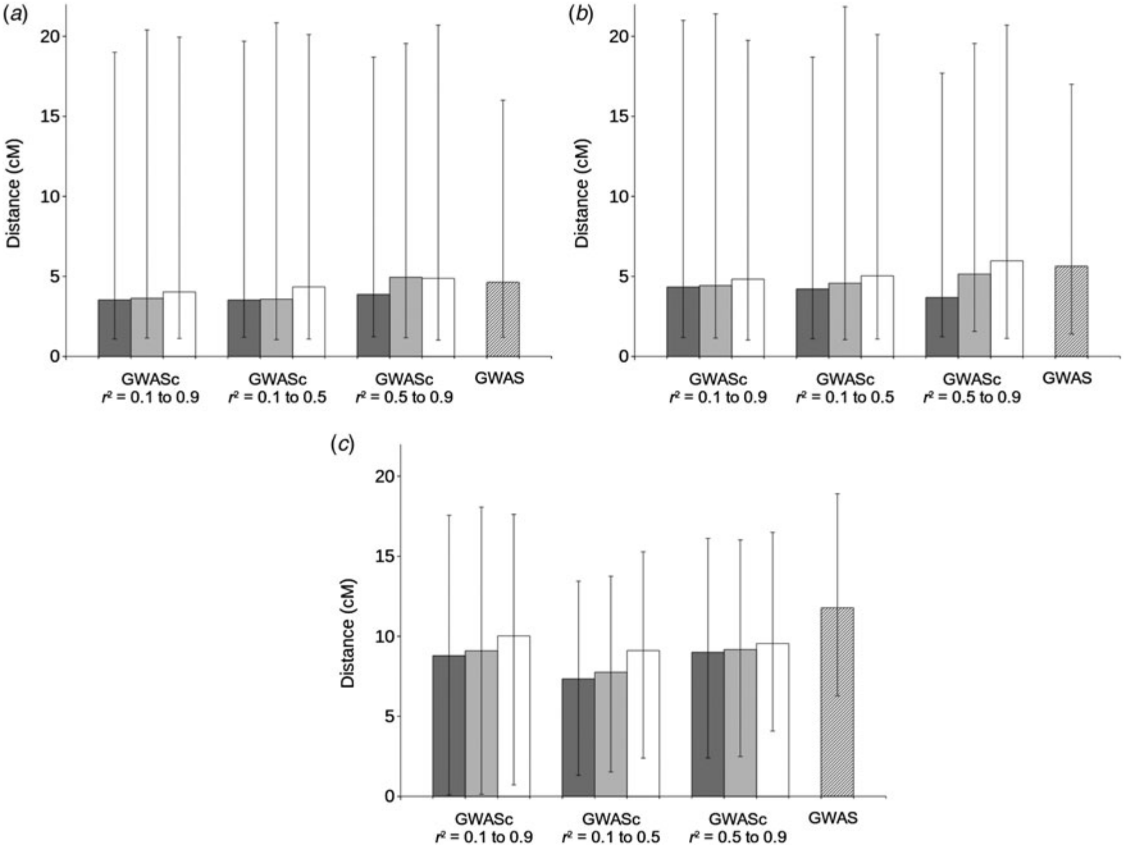

Fig. 3. Average absolute distance between significant ($p < 0{\cdot}0005$) SNPs and the QTL under standard genome-wide association studies (GWAS) analysis and composite GWAS (GWASc) for small-effect QTLs (*a*), medium-effect QTLs (*b*) and large-effect QTLs (*c*); the whiskers extend the range of the results. Columns are organized in four independent groups depending on the analytical approach (GWAS vs. GWASc) and the linkage disequilibrium (r^2) between competing SNPs and the QTL for GWASc analyses; within-group colour differences identify the size of the genomic region where competing SNPs were assessed, this being 10 cM (white), 30 cM (light grey) and 50 cM (dark grey) on each side of the tested SNP. The striped bar corresponds to the standard GWAS approach.

previously shown by Zeng (1994) and Jansen & Stam (1994) within the context of QTLs mapping, this being generalized to genome-wide markers by Bernardo (2013).

The composite GWA approach was developed on the basis of multiple regression; focusing on the GWA scenario, various properties inherent to the multiple regression methodology must be revisited before discussing the results obtained under simulation. As noted by Zeng (1993; 1994) and previously demonstrated by Stam (1991), the expected partial regression coefficient of the analysed trait on the *i*th SNP depends only on those causal mutations that are located on the interval between the neighbouring SNPs *i*-1 and *i* + 1, both of which are accounted for in the analytical model (property 1). This is a very desirable property that characterizes the composite GWA approach as an interval test. Note that standard GWA analyses focusing on SNP-by-SNP approaches are less precise unconditional tests, in which we can only check whether there is one or more causal mutations on a chromosome (Jensen, 1993). Multiple regression analysis allows for conditioning on both unlinked and linked markers, which reduces the sampling variance of the test statistic (property 2) and the chance of interference of possible multiple-linked QTLs (property 3), respectively (Rodolphe & Lefort, 1993; Zeng, 1993; 1994). Property 2 derives from the evidence that unlinked markers can account for some residual genetic variation and, as a consequence, increase the statistical power of the test. Nevertheless, property 3 may counteract the increase in power because of the increase in the sampling variance inherent to conditional testing (Zeng, 1993). Finally, it has been shown that partial regression coefficients on two markers in a multiple regression analysis are generally uncorrelated, unless the two markers are adjacent; even in this case, correlation is usually very small (Zeng, 1993).

As shown by the results obtained on simulated populations, statistical properties of the composite GWA characterized a compromise between precision (properties 1 and 2) and power (property 3) of the association test. Indeed, power loss was quite relevant

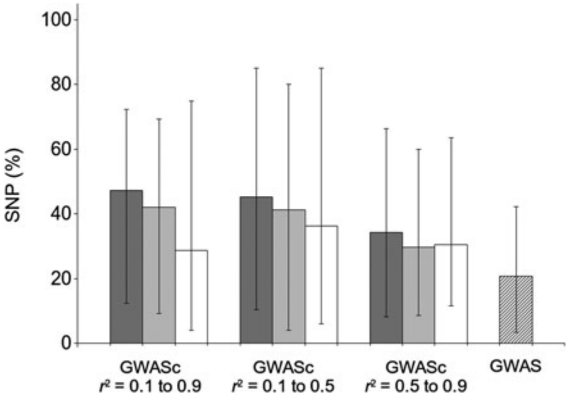

Fig. 4. Average percentage of significant ($p < 0{\cdot}0005$) SNPs located not farther than 2·5 cM from the QTL under standard genome-wide association studies (GWAS) analysis and composite GWAS (GWASc) for large-effect QTLs; the whiskers extend the range of the results. Columns are organized in four independent groups depending on the analytical approach (GWAS vs. GWASc) and the linkage disequilibrium (r^2) between competing SNP and the QTL for GWASc analyses; within-group colour differences identify the size of the genomic region where competing SNP were assessed, this being 10 cM (white), 30 cM (light grey) and 50 cM (dark grey) on each side of the tested SNP. The striped bar corresponds to the standard GWAS approach.

as evidenced by the percentage of simulated populations where composite GWA failed to detect significantly associated SNPs (Table 1). Differences between standard and composite GWA analyses were minimum when checking large-effect QTLs, whereas power loss was faster for composite than for standard GWA when the effect of the QTL decreased (Fig. 5). This was not greater than the evidence that the implementation of a composite GWA approach implies the payment of a particularly high price in terms of power, discouraging the systematic use of composite GWA models if medium- to small-effect QTLs could be anticipated. Indeed, composite GWA studies must be viewed as a refining methodology that must be implemented after confirming the presence of significantly associated SNPs by standard GWA analysis. If not, genomic research could be impaired by a massive incidence of false-negatives due to an excessive zeal to refine the location of causal QTLs before roughly identifying their presence and approximating their additive genetic effect. Far from discouraging the implementation of composite GWA analyses, this conclusion warns future users about the consequences of power loss when screening genomic data for association effects.

The composite GWA model developed above assumed two highly flexible model parameters when selecting competing SNPs. Both the width of the analytical window where competing SNPs were assessed and the LD range between the tested SNPs and competing SNPs could be modified and adapted to different scenarios. This allowed for the evaluation of their impact on model performance as well as for elucidation of preliminary recommendations for further genomic analyses. The analytical window had a small impact on the number of competing SNPs (Fig. 1) as well as on the precision of the composite GWA analysis (Fig. 2 to 4). This could be mainly due to the relatively small extent of LD in mammalian genomes (Tenesa *et al.*, 2004; Sargolzaei *et al.*, 2008), which was mimicked in our simulated chromosomes. Nevertheless, small advantages shown by wider windows would suggest that, if not conflicting with computing requirements, the wider the better. A similar conclusion is derived from the preferable range of

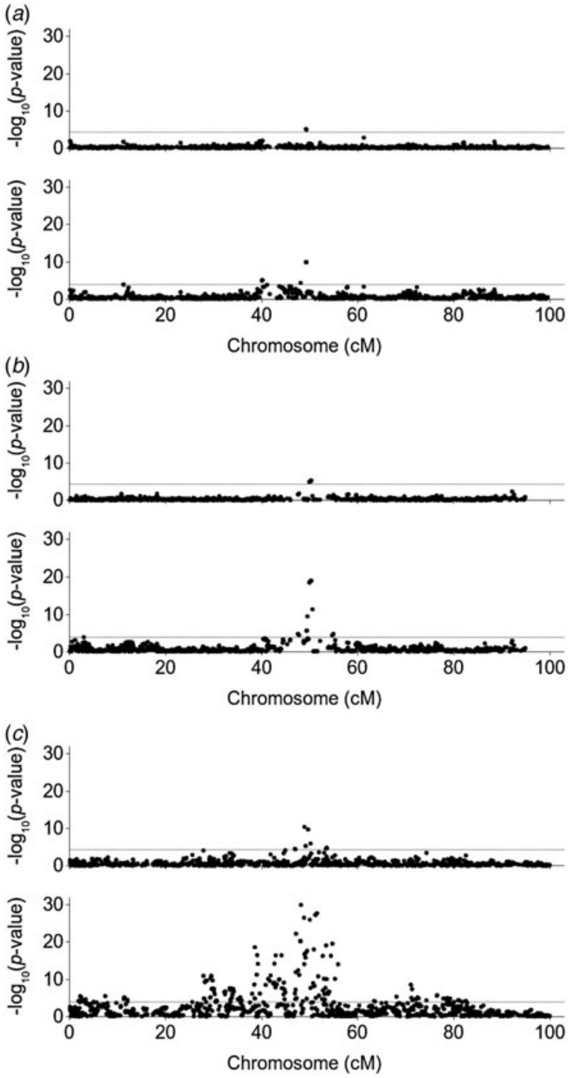

Fig. 5. Representative examples of Manhattan plots from the standard genome-wide association analysis (upper panel) and the composite genome-wide association analysis (lower panel) for populations with small- (*a*), medium- (*b*) and large-effect QTLs (*c*). Competing SNPs for composite genome-wide association analyses were assessed in the whole chromosome and linkage disequilibrium (r^2) with the tested SNP was restricted to $0{\cdot}1 \leqslant r^2 \leqslant 0{\cdot}9$.

LD between the tested SNP and competing SNPs. As suggested by properties 2 and 3, the inclusion of both lowly and highly linked competing SNPs could contribute remarkable advantages (and some disadvantages mainly linked to power loss) to the composite GWA. Nevertheless, and compared with remaining composite GWA parameterizations, the wider interval ($0\cdot 1 \leqslant r^2 \leqslant 0\cdot 9$) neither remarkably reduced the average absolute distance between significant SNPs and the QTL (Fig. 3) nor increased the percentage of significant SNPs located in the nearest 2·5 cM around the QTL (Fig. 4), although this did suffer from larger power loss. A similar pattern was shown by $0\cdot 1 \leqslant r^2 \leqslant 0\cdot 5$. Within this context, the LD interval characterized by $0\cdot 5 \leqslant r^2 \leqslant 0\cdot 9$ could be viewed as an appealing alternative where the loss of statistical power was attenuated.

The composite GWA model contributes a novel point of view for GWA analyses where testing circumscribed to the genomic region flanking each SNP (delimited by the nearest competing SNP) and conditioning on linked markers increases the precision of locating causal mutations, but possibly at the expense of power.

This research was supported by the research project AGL2010-21176/GAN. The research contract of J. Casellas was partially funded by the *Ministerio de Ciencia e Innovación* of Spain's government (reference RYC-2009-04049).

Declaration of interest

None.

References

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**, 289–290.

Benyamin, B., Visscher, P. M. & McRae, A. F. (2009). Family-based genome-wide association studies. *Pharmacogenomics* **10**, 181–190.

Bernardo, R. (2013). Genomewide markers as cofactors for precision mapping of quantitative trait loci. *Theoretical and Applied Genetics* **126**, 999–1009.

Bonferroni, C. E. (1936). Teoria statistica della classi a calcolo della probabilità. *Pubblicasioni del R Instituto Superiore di Scienze Economiche e Commerciali di Firenze* **8**, 3–62.

Casellas, J. & Varona, L. (2011). Effect of mutation age on genomic prediction. *Journal of Dairy Science* **94**, 4224–4229.

Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Munzy, D. M., Sodergren, E. J., Scherer, S., Scott, G., Steffen, D., Worley, K. C., Burch, P. E., Okwuonu, G., Hines, S., Lewis, L., DeRamo, C., Delgado, O., Dugan-Rocha, S., Miner, G., Morgan, M., Hawes, A., Gill, R., Celera, Holt, R. A., Adams, M. D., Amanatides, P. G., Baden-Tillson, H., Barnstead, M., Chin, S., Evans, C. A., Ferriera, S., Fosler, C., Glodek, A., Gu, Z., Jennings, D., Kraft, C. L., Nguyen, T., Pfannkoch, C. M., Sitter, C., Sutton, G. G., Venter, J. C., Woodage, T., Smith, D., Lee, H. M., Gustafson, E., Cahill, P., Kana, A., Doucette-Stamm, L., Weinstock, K., Fechtel, K., Weiss, R. B., Dunn, D. M., Green, E. D., Blakesley, R. W., Bouffard, G. G., De Jong, P. J., Osoegawa, K., Zhu, B., Marra, M., Schein, J., Bosdet, I., Fjell, C., Jones, S., Krzywinski, M., Mathewson, C., Siddiqui, A., Wye, N., McPherson, J., Zhao, S., Fraser, C. M., Shetty, J., Shatsman, S., Geer, K., Chen, Y., Abramzon, S., Nierman, W. C., Havlak, P. H., Chen, R., Durbin, K. J., Egan, A., Ren, Y., Song, X. Z., Li, B., Liu, Y., Qin, X., Cawley, S., Worley, K. C., Cooney, A. J., D'Souza, L. M., Martin, K., Wu, J. Q., Gonzalez-Garay, M. L., Jackson, A. R., Kalafus, K. J., McLeod, M. P., Milosavljevic, A., Virk, D., Volkov, A., Wheeler, D. A., Zhang, Z., Bailey, J. A., Eichler, E. E., Tuzun, E., Birney, E., Mongin, E., Ureta-Vidal, A., Woodwark, C., Zdobnov, E., Bork, P., Suyama, M., Torrents, D., Alexandersson, M., Trask, B. J., Young, J. M., Huang, H., Wang, H., Xing, H., Daniels, S., Gietzen, D., Schmidt, J., Stevens, K., Vitt, U., Wingrove, J., Camara, F., Mar Albà, M., Abril, J. F., Guigo, R., Smit, A., Dubchak, I., Rubin, E. M., Couronne, O., Poliakov, A., Hübner, N., Ganten, D., Goesele, C., Hummel, O., Kreitler, T., Lee, Y. A., Monti, J., Schulz, H., Zimdahl, H., Himmelbauer, H., Lehrach, H., Jacob, H. J., Bromberg, S., Gullings-Handley, J., Jensen-Seaman, M. I., Kwitek, A. E., Lazar, J., Pasko, D., Tonellato, P. J., Twigger, S., Ponting, C. P., Duarte, J. M., Rice, S., Goodstadt, L., Beatson, S. A., Emes, R. D., Winter, E. E., Webber, C., Brandt, P., Nyakatura, G., Adetobi, M., Chiaromonte, F., Elnitski, L., Eswara, P., Hardison, R. C., Hou, M., Kolbe, D., Makova, K., Miller, W., Nekrutenko, A., Riemer, C., Schwartz, S., Taylor, J., Yang, S., Zhang, Y., Lindpaintner, K., Andrews, T. D., Caccamo, M., Clamp, M., Clarke, L., Curwen, V., Durbin, R., Eyras, E., Searle, S. M., Cooper, G. M., Batzoglou, S., Brudno, M., Sidow, A., Stone, E. A., Venter, J. C., Payseur, B. A., Bourque, G., López-Otín, C., Puente, X. S., Chakrabarti, K., Chatterji, S., Dewey, C., Pachter, L., Bray, N., Yap, V. B., Caspi, A., Tesler, G., Pevzner, P. A., Haussler, D., Roskin, K. M., Baertsch, R., Clawson, H., Furey, T. S., Hinrichs, A. S., Karolchik, D., Kent, W. J., Rosenbloom, K. R., Trumbower, H., Weirauch, M., Cooper, D. N., Stenson, P. D., Ma, B., Brent, M., Arumugam, M., Shteynberg, D., Copley, R. R., Taylor, M. S., Riethman, H., Mudunuri, U., Peterson, J., Guyer, M., Felsenfeld, A., Old, S., Mockrin, S., Collins, F & Rat Genome Sequencing Project Consortium (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521.

Habier, D., Fernando, R. L. & Dekkers, J. C. M. (2009). Genomic selection using low-density marker panels. *Genetics* **182**, 343–353.

He, Q. & Lin, D.-Y. (2011). A variable selection method for genome-wide association studies. *Bioinformatics* **27**, 1–8.

Hickey, J. M. & Gorjanc, G. (2012). Simulated data from genomic selection and genome-wide association studies using a combination of coalescent gene drop methods. *G3* **2**, 425–427.

Hill, W. G. & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**, 226–231.

Ibáñez-Escriche, N., Fernando, R. L., Toosi, A. & Dekkers, J. C. M. (2009). Genomic selection of purebred for crossbred performance. *Genetics, Selection, Evolution* **41**, 12.

International Chicken Genome Sequencing Consortium (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716.

International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**, 1299–1320.

Jansen, R. C. & Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**, 1447–1455.

Jensen, R. C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **167**, 1987–2002.

Kemper, K. E., Daetwyler, H. D., Visscher, P. M. & Goddard, M. E. (2012). Comparing linkage and association analyses in sheep points to a better way of doing GWAS. *Genetics Research* **94**, 191–203.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C. & Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389.

Kosambi, D. D. (1943). The estimation of map distances from recombination values. *Annals of Eugenics* **12**, 172–175.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O. A., Leung, F. C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C. C., Lam, T. T., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M. W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T. W., Yiu, S. M., Liu, S., Zhang, H., Li, D., Huang, Y., Wang, X., Yang, G., Jiang, Z., Wang, J., Qin, N., Li, L., Li, J., Bolund, L., Kristiansen, K., Wong, G. K., Olson, M., Zhang, X., Li, S., Yang, H., Wang, J. & Wang, J. (2010). The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317.

Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

Mrode, R. A. (2005). *Linear Models for the Prediction of Animal Breeding Values*. CAB International, Oxon, UK.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transaction of the Royal Society A* **231**, 289–337.

Ødegård, J., Soneson, A. K., Yazdi, M. H. & Meuwissen, T. H. E. (2009). Introgression of a major QTL from an inferior into a superior population using genomic selection. *Genetics, Selection, Evolution* **41**, 38.

Pearson, T. A. & Manolio, T. A. (2008). How to interpret a genome-wide association study. *Journal of the American Medical Association* **19**, 1335–1344.

Rodolphe, F. & Lefort, M. (1993). A multi-marker model for detecting chromosomal segments displaying QTL activity. *Genetics* **134**, 1277–1288.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S., Altshuler, D & International SNP Map Working Group (2001). A map of human genome sequence variation containing 1·42 million single nucleotide polymorphisms. *Nature* **409**, 928–933.

Sargolzaei, M., Schenkel, F. S., Jansen, G. B. & Schaeffer, L. R. (2008). Extent of linkage disequilibrium in Holstein cattle in North America. *Journal of Dairy Science* **91**, 2106–2117.

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S. H., Schwalie, P. C., Tang, Y. A., Ward, M. C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L. N., Ayub, Q., Ball, E. V., Beal, K., Bradley, B. J., Chen, Y., Clee, C. M., Fitzgerald, S., Graves, T. A., Gu, Y., Heath, P., Heger, A., Karakoc, E., Kolb-Kokocinski, A., Laird, G. K., Lunter, G., Meader, S., Mort, M., Mullikin, J. C., Munch, K., O'Connor, T. D., Phillips, A. D., Prado-Martinez, J., Rogers, A. S., Sajjadian, S., Schmidt, D., Shaw, K., Simpson, J. T., Stenson, P. D., Turner, D. J., Vigilant, L., Vilella, A. J., Whitener, W., Zhu, B., Cooper, D. N., de Jong, P., Dermitzakis, E. T., Eichler, E. E., Flicek, P., Goldman, N., Mundy, N. I., Ning, Z., Odom, D. T., Ponting, C. P., Quail, M. A., Ryder, O. A., Searle, S. M., Warren, W. C., Wilson, R. K., Schierup, M. H., Rogers, J., Tyler-Smith, C. & Durbin, R. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175.

Singer, J. B. (2009). Candidate gene association analysis. *Methods in Molecular Biology* **573**, 223–230.

Stam, P. (1991). Some aspects of QTL analysis. In *Proceedings of the Eighth Meeting of the Eucarpia Section Biometrics in Plant Breeding. Brno, Czech Republic, July 1991*. European Association for Research on Plant Breeding (EUCARPIA).

Tenesa, A., Wright, A. F., Knott, S. A., Carothers, A. D., Hayward, C., Angius, A., Maestrale, G., Hastie, N. D., Pirastu, M. & Visscher, P. M. (2004). Extent of linkage disequilibrium in a Sardinian sub-isolate: sampling and methodological considerations. *Human Molecular Genetics* **13**, 25–33.

Toosi, A., Fernando, R. L. & Dekkers, J. C. M. (2010). Genomic selection in admixed and crossbred populations. *Journal of Animal Science* **88**, 32–46.

Wang, K., Dickson, S. P., Stolle, C. A., Krantz, I. D., Goldstein, D. B. & Hakonarson, H. (2010). Interpretation of association signals and identification of causal variants from genome-wide association studies. *American Journal of Human Genetics* **86**, 730–742.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.

Zeng, Z.-B. (1993). Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. *Proceedings of the National Academy of Sciences of USA* **90**, 10972–10976.

Zeng, Z.-B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.