Assessing orthogonality and statical properties of linear regression methods for interval mapping with partial information

Carl Nettelblad[1] , Örjan Carlborg[2,3] and José M. Álvarez-Castro[*2,3,4]

[1]Department of Information Technology, Box 337, Uppsala University, SE-75105, Uppsala, Sweden
[2]Linnaeus Centre for Bioinformatics, Box 598, Uppsala University, SE-75124, Uppsala, Sweden
[3]Department of Animal Breeding and Genetics, Box 7023, Swedish University of Agricultural Sciences, SE-75007, Uppsala, Sweden
[4]Department of Genetics, University of Santiago de Compostela, ES-27002 Lugo, Spain

Email: Carl Nettelblad - carl.nettelblad@it.uu.se; Örjan Carlborg - orjan.carlborg@hgen.slu.se; José M. Álvarez-Castro*- jose.alvarez.castro@usc.es;

*Corresponding author

Abstract

Background: Mapping quantitative trait loci (QTL) has become a widely used tool in genetical research. In such experiments, it is desired to obtain orthogonal estimates of genetic effects for a number of reasons concerning both the biological meaning of the estimated locations and effects, and making the statistical analysis clearer and more robust. The currently used statistical methods, however, are not optimized for orthogonality, especially in cases involving interval mapping between markers and/or in incomplete datasets. This is an adverse limitation for the application of such methods for QTL scans involving model selection over putative complex gene networks.

Results: We describe how deviations from orthogonality arise in currently used methods. We demonstrate one option for obtaining orthogonal estimates of genetic effects using multiple imputations per individual in an otherwise unchanged regression context. Our proposed IRIM method avoids inflated values for explainable variance and genetic effect variables, while showing a clear preference for marker locations in a fine mapping context. Despite possible shortcomings, similar results to linear regression are demonstrated for our proposed approach (IRIM) in an experimental dataset.

Conclusions: Imputation-based methods can be used to enhance the statistical dissectability of effects, as well as computational performance. We exemplify how Haley-Knott regression is not only distorting the explainable variance, but also point out how the estimated phenotype values between classes, and the resulting effects, become dependent. This illustrates the need for a more radical departure in the approach chosen in order to achieve orthogonality.

Background

When the field of genetics was first initiated, the traits studied were frequently binary in nature, e.g. simple phenotypes of color or shape. In addition, the traits successfully studied could often be linked to specific genetic loci, hence called monogenic traits, following the deterministic rules of Mendelian genetics. Many phenotypes of importance in animal breeding, medicine and molecular biology are, however, quantitative, rather than binary. The weight of an individual, the age at disease onset or disease propensity, and gene expression, are all examples of variables that are inherently quantitative in nature, requiring a more elaborate statistical approach for finding the underlying genetic architecture [1]. While heritable components are present, they are also all characterized by a significant environmental component, which might be modeled as a normal distribution, or described in further detail by covariates [2]. A quantitative trait locus (QTL) is a locus in the genome of a population under study where allelic variation shows a correlation with the variation in one or more quantitative traits. The process of identifying such loci is called QTL mapping, which is the first step of QTL analysis, which should also include an assessment of the phenotype effects of each locus.

One of the key issues in QTL analysis, and specifically in approaches based on model selection, is to obtain orthogonal estimates of genetic effects in the sample of individuals under study. This specific problem has motivated theoretical studies to develop orthogonal genetic models for use in QTL detection. Orthogonality in this context refers to the parameter estimates of the model being statistically independent, which is directly related to the property of mathematical orthogonality, applied to a linear regression design matrix. In genetics, the history of such models goes back to the one-dimensional F_2 case as described by [1]. There are a number of reasons why orthogonality is a desired property for models of genetic effects in QTL

analysis [3–5]. In practice, orthogonal estimates of genetic effects have a defined biological meaning as average effects of allele substitutions in populations and directly lead to a proper decomposition of the genetic variance in the population or sample under study. Furthermore, orthogonal models of genetic effects considerably facilitate model selection in QTL analysis, since orthogonal estimates and their variances remain constant in models with a reduced parameter set. In fact, when using an orthogonal model, a single linear regression including all variables in a full model directly provides estimates for any subset of those variables, thus enabling routine comparisons of alternative models of genetic effects (effects of a set of loci and a particular set of interactions among them) underlying the chosen phenotype.

Using orthogonal models does not preclude obtaining estimates fitting alternative, opportune interpretations. Indeed, change of reference tools have been developed to give different interpretations of QTL estimates. These transformations are useful to inspect evolutionary properties of a genetic system, such as domestication processes and Dobzhansky-Muller incompatibilities [6]. In any case, the use of a model of genetic effects for routine analyses of real data also requires a statistical framework suitable to perform interval mapping (IM, [7]), in order to deal with missing data and to evaluate the effects of putative QTL inside marker intervals. The Haley-Knott regression (HKR, also frequently called REG for its use of linear regression) is a frequently used method to approximate IM that allows simple and rapid estimation of genetic effects [8, 9] and is the method used in several user-friendly and readily available software packages [10, 11]. Some disadvantages of the HKR have been pointed out, especially concerning a bias of the residual variance, which increases with epistasis and linkage disequilibrium [12, 13]. [14] have developed an improvement of the HKR – the estimating equation method – that reduces this bias with the drawback of higher computational requirements.

In addition, the HKR was not developed taking orthogonality into consideration, but originally implemented with the F_∞ model of genetic effects, which is not orthogonal for the populations commonly used in QTL analysis. This is not a major problem since the statistical framework of the HKR is independent of the genetic model used. Thus, it can be implemented with other models of genetic effects, including the F_2 model, which is orthogonal in ideal F_2 populations [3] or with NOIA [6, 15], a model framework which is orthogonal even under departures from the Hardy-Weinberg proportions. However, even with an adapted model, the statistical framework of the HKR itself is not designed to maintain orthogonality. This observation also applies to the estimating equation method of [14].

In this communication we discuss details of HKR, as it is an established norm in the field. We show that this method, even if implemented with the F_2 or another orthogonal model for the population under study,

introduces departures from orthogonality during the statistical estimation of genetic effects. Furthermore, we show that it is possible to overcome this problem by modifying the system of equations from HKR to perform imputations. Imputations have previously been suggested for QTL mapping in the context of general expectation maximization [16], or Monte Carlo [17], frameworks. For neither of these approaches, the possible orthogonal properties of the methods, or their results, have been discussed or shown. The computational requirements of the methods involved are also significant due to the repeated function evaluation procedures that form their basis. In this communication, we consider a computationally efficient linear regression imputation method, which we call Imputation Regression Interval Mapping (IRIM).
We use a simple numerical example and reanalyze an experimental dataset to illustrate the features of orthogonality and discuss the other properties of the IRIM method. In particular, we point out that the IRIM method does not compromise the detection power in our practical experiments and that it presents an advantage in terms of computational requirements, leading to more efficient and faster analyses due to the orthogonality. In addition, we provide simulations with varied results, concluding that understanding of the differences between IRIM and HKR can be a step towards workable orthogonal estimates, with or without a linear regression approach, in the case of partial information.

Methods

The following linear system of equations is used in a linear regression to obtain estimates of genetic effects:

$$\mathbf{G}^* = \mathbf{X} \cdot \mathbf{E} + \varepsilon \tag{1}$$

In this equation, ε is the vector of errors, $\mathbf{E}$ is a vector of genetic effects, and $\mathbf{G}^*$ is the vector of phenotypes of the individuals in the analysed population. $\mathbf{X} = \mathbf{Z} \cdot \mathbf{S}$, where $\mathbf{S}$ is the genetic-effects design matrix of the statistical formulation of the model of genetic effects, and $\mathbf{Z}$ is the matrix linking the observed individual phenotypes to their corresponding genotypes [5, 15].
Concerning first the matrix $\mathbf{S}$, a frequently used example of an orthogonal model comes from a single diallelic locus in an ideal F_2 population. This model was already described by [1], and it can be expressed as [5, 15]:

$$\mathbf{S}_{F_2} = \begin{pmatrix} 1 & -1 & -0.5 \\ 1 & 0 & 0.5 \\ 1 & 1 & -0.5 \end{pmatrix} \tag{2}$$

[15] provide a general statistical orthogonal model for any number of diallelic loci under linkage equilibrium backcross or multi-generational intercrosses. With genotype frequencies p_{ij}, where ij the allele

combination, and allele frequencies p_i, this model is:

$$\mathbf{S} = \begin{pmatrix} 1 & -2p_2 & -\frac{p_{12}p_{22}}{2p_1p_2-0.5p_{12}} \\ 1 & p_1 - p_2 & \frac{p_{11}p_{22}}{p_1p_2-0.25p_{12}} \\ 1 & 2p_1 & -\frac{p_{11}p_{12}}{2p_1p_2-0.5p_{12}} \end{pmatrix} \tag{3}$$

In an ideal F_2 population, the model will coincide with equation 2 above, but it is also applicable to any diallelic design (including a backcross or multi-generational intercrosses). For the case of two loci, the $9 * 9$ **S** matrix is the Kronecker product of the **S** matrices of the two loci.

Concerning now the indicator matrix **Z**, for a multilocus case it can be constructed using a Kronecker product from the indicator matrices of the individual loci, assuming no linkage. In practice, this means that if we denote the genotypes in individual loci 11, 12, 22, the horizontal ordering in a combined **Z** matrix for two loci will consist of indicator probabilities ordered as

$(11, 11); (12, 11); (22, 11); (11, 12); (12, 12); (22, 12); (11, 22);$

$(12, 22); (22, 22)$. The general IRIM expansion results in each individual (observation) being expanded to nine rows tracking the diagonal, with the weights assigned accordingly. For details, we refer to the illustration of a numerical example for a single locus, and our suggested efficient evaluation method that avoids increases in computational load (Section 0.8).

Regardless of the number of loci involved, when fitting a genetic model at a fully informative marker, all rows in **Z** have a single unitary element ("1") in the column corresponding to the genotype of the individual, and zeros in all other columns. Here, $\mathbf{Z}^T\mathbf{Z}$ is a diagonal matrix, which is a key feature for $\mathbf{X}^T\mathbf{X}$ being diagonal and hence **X** being orthogonal [15]. However, when there is missing information for a marker (or equivalently when inspecting locations inside marker intervals where the genotype is estimated from flanking markers) the HKR proposes to build rows in **Z**, where the columns corresponding to particular genotypes contain the probabilities for that genotype given data at flanking markers [8]. This implies having more than one non-zero value in many rows of **Z** and consequently compromising orthogonality. Indeed, as we will demonstrate later on using an example, this causes $\mathbf{Z}^T\mathbf{Z}$ not to be diagonal and consequently the system not to be orthogonal. Hereafter we propose one method to overcome this problem, including how this method relates to other approaches.

0.1 The Imputation Regression method for IM

To achieve orthogonality, it is thus desirable to build **Z** matrices with a single non-zero element per row, which can be done by imputations [16]. The procedure involves introducing multiple rows per individual,

where the number of rows is chosen to result in a proportion of ones ("1") for each genotype of an individual that approximates the probability of the individual having that genotype.

Parameter estimates can then be obtained using imputation-based weighted or unweighted regression, what we call the IRIM method. Multiple identical rows in an unweighted regression are conceptually equivalent to a single row with an appropriate weight. Therefore, weighted imputations conveniently do not increase computational requirements. The weights can be introduced as either an additional diagonal matrix within the normal equations for the regression, or equivalently by multiplying the left-hand side and right-hand side of each row with the square root of the intended weight. The latter approach can be useful if the computations are made with a numerical library or in an existing code structure where weighted regression is not explicitly available.

Most other imputation-based approaches have, in addition to using imputations, deviated, in one way or another, from the simplicity of the single-pass computation of HKR. These approaches tend to perform less favorably concerning either the statistical interpretation or their applicability to large-scale datasets due to additional computational load.

0.2 Comparison of likelihoods for estimation procedures

The original presentation of IM [7] posed the phenotype likelihood at a locus for a single individual as a mixture of several normal distributions, weighted by the genotype probabilities at that locus, as determined by the marker values:

$$\mathcal{L}_i^{IM} = \sum_j p_{ij} N(\mu_j, \sigma_e) \tag{4}$$

As noted in [13], posterior probabilities π_{ij} can be inferred for the genotypes from this distribution. Furthermore, deviations from the linear regression expressions were at a maximum if $\pi_{ij} \neq p_{ij}$. The likelihood for Haley-Knott regression can be expressed as:

$$\mathcal{L}_i^{REG} = N\left(\sum_j p_{ij}\mu_j, \sigma_e\right) \tag{5}$$

IRIM uses several rows with weights. This results in a likelihood where the components are multiplied, rather than added:

$$\mathcal{L}_i^{IRIM} = \prod_j N(\mu_j, \sigma_e)^{p_{ij}} \tag{6}$$

As can be seen, given a set of μ_j estimates, the IM and IRIM likelihood formulas are similar in structure, representing an arithmetical, and a geometrical, mean, respectively. In Figures 0.11 and 0.11 are

illustrating to cases with $p_{i1} = 0.75, \mu_1 = -1, p_{i2} = 0.25, \mu_2 = 1$, but with varying standard deviations. Which curve constitutes the best match to IR is not a clear-cut issue.

0.3 Comparison of residual contributions from HKR and IRIM

In addition to our main motivation of this communication–using the proper statistical tools for obtaining orthogonal estimates of genetic effects in real experiments–we note that IRIM also have additional convenient statistical properties when compared to HKR. To show these properties, we consider the contributions of the individuals to the residual sum of squares (RSS) resulting in the linear regression. In the HKR this contribution is:

$$\sigma_i^2(\mathbf{Z}_{HK}) = (y_i - \mathbf{G}_1^* p_{i1} - \mathbf{G}_2^* p_{i2})^2 \tag{7}$$

where $\mathbf{G}_j^*$ indicates the phenotype estimate, and p_{ij} is the probability for individual i, both for genotype j, in a simple backcross. As can be seen, HKR will only detect differences between y_i, and an artificial average of the two genotype states, an average with no biological interpretation. In IRIM, the corresponding quantity is:

$$\sigma_i^2(\mathbf{Z}_{IRIM}, W_i) = W_i((y_i - \mathbf{G}_1^*)^2 p_{i1} + (y_i - \mathbf{G}_2^*)^2 p_{i2}) \tag{8}$$

In this case, we compare against two separate normal distributions, corresponding to the two actually possible genotypes, rather than the shifted-mean distribution used in HKR. From these expressions, the difference between them ((8) - (7)) is:

$$\begin{gathered}\sigma_i^2(\mathbf{Z}_{IRIM}) - \sigma_i^2(\mathbf{Z}_{HK}) = \\ (\mathbf{G}_1^*)^2 p_{i1} + (\mathbf{G}_2^*)^2 p_{i2} - (y_i - \mathbf{G}_1^* p_{i1} - \mathbf{G}_2^* p_{i2})^2\end{gathered} \tag{9}$$

Using basic calculus it can be shown that this expression has zeros as minima at fully informative locations. Therefore, both methods produce identical results at those minima. With missing data, the RSS from the imputation method will, however, always be higher than the one derived using HKR (see e.g. σ_{res}^2 in Table 0.11), i.e. the LOD score function of the IRIM method will be lower than the one of the HKR. The difference is also scaling to the difference $\mathbf{G}_1^* - \mathbf{G}_2^*$, the two functions actually sharing identical shape with only a constant vertical translation. However, this does not seem to lead to a lower power of detection of QTL by IRIM in practice, as we show in an analysis of experimental data. The matching of results to HKR (as well as IM with full expectation maximization, EM) in areas of full information is convenient, giving a stable point of reference. In addition, the resulting genotype-class averages with IRIM will coincide with the phenotype of each individual, distributed by the class probabilities. A shift in

probabilities for a single individual gives an immediately determinable shift in the averages, which can also be used to compute the RSS, which we are employing in our efficient algorithm described in section 0.8.

0.4 Estimates of genetic effects

The indicator matrix $\mathbf{Z}$ used in IRIM contains a single non-zero element per row. As a direct consequence, the estimates of per-genotype phenotype values (the genotype-phenotype map) can be done independently for each genotype. Furthermore, this estimate is simply the weighted arithmetical mean of the observed phenotypes for all individuals, with accompanying weights.

A non-informative individual will contribute to the means for all classes between which there is uncertainty, i.e. the expectation of the difference between the class means will shrink, compared to a case with full information for that individual. In a two-class system, the per-individual probability variances can be used to "correct" these estimates, translating IRIM estimates into HKR estimates. This is a special case of a (symmetric) linear system between all class estimates. The unknowns would represent the HKR estimates. The coefficients in each row i are then weighted sums over all rows in $\mathbf{Z}$, weighted by Z_{ji}, or in other words $\mathbf{Z}^T\mathbf{Z}$. This is a contrived way of constructing the normal equation system.

The multiple imputation approach for QTL mapping used in [17] will not render identical effect estimates for two reasons. One is the simple fact that our approach represents the results with an infinite number of imputations by using weights, while [17] are using unweighted linear regression for each imputation. The other important source of differences is that the other imputation approach uses the resulting residuals as a source for posterior probabilities for the genotype realizations (pseudomarkers) that form the basis for their imputation scheme. Due to that weighting, realizations that result in stronger phenotype differences between genotype classes will be preferred.

As linear regression provides an extremely fast framework that is familiar to many users, with an extensive base of knowledge and existing source code, we choose to investigate the possibilities of using imputations without the weighting approach of different replicates.

0.5 QTL in different information environments

Many of the recent analyses of QTL detection methods [13, 14] simulate QTL between markers. Methods are also motivated based on how they handle detecting a QTL with partial information. If there is a marker of full information with no informative surrounding markers, all individuals will show an identical decay towards the population mean. This is only a linear transformation of the design matrix, something

which will not affect the portion of explainable variance, but rather inflate the estimated effect values. If all individuals which has a value of 1 for some indicator column at the marker has a value of 0.75 instead at some position x, due to the effects of the mapping function, the estimated effect value for the corresponding non-confounded genotype will be appropriately scaled. Under these conditions, there is no actual information lost from the perspective of HKR. Information, in the sense of an increasing residual in the model, is only possibly lost when different individuals show different recombination rates. This loss can be due to another marker being present, with variations in information content, or variations in indicated genotype compared to the first marker. This latter type of variations will result in a steep change in genotype probabilities for some individuals, while others will be essentially unchanged. The result is a plateau around an informative marker, for the full region where the only significant linkage is to that marker. Further markers with limited information content, or only weak linkage to the original marker, can then shift the plateau in either direction. IM will not perform in an identical manner. The explainable variance will demonstrate identical behavior (a solid plateau), but the phenotype class means will not be inflated in the same manner. Furthermore, and more importantly, the IM method makes a distinction between explainable variance and the total likelihood, the environmental variance just being one of parameters fitted in the EM algorithm.

If the true QTL is instead located at some other position within the possible plateau region, than the informative marker, this does not affect the shape of the plateau. The marker location and the true QTL location will both be considered as equally good fits. However, the expectation value is lower, due to the confounding effect [12, 14, 18]. IM will instead find a different posterior distribution of genotype probabilities. Especially for low broad-sense heritability, there is a risk of overfitting. For e.g. a dataset from a single phenotype distribution with no genotype dependency (especially if this distribution is non-normal), the distribution can be decomposed into multiple normal components with lower standard deviation, thus inflating explainable variance.

In a specific marker, the pattern of information can also be dependent on the actual genotype at the marker, e.g. making homozygotes with alleles from one founder line identifiable in several sibsets, while there would be a lack of discriminatory power between heterozygotes and homozygotes from the other founder line. Depending on the nature of the true genotype-phenotype map, the consequences can be similar to those of selective genotyping, which can cause HKR to severely over-estimate the explainable variance [14].

The explainable variance from IRIM will always be reduced in low information environments. If there is a

QTL at a marker, the QTL will also be detected in that position, which should be contrasted to the tendency of HKR to consider a non-marker position in some configurations equally likely. This is more consistent with the IM likelihood. Detection power is reduced, compared to IM, at non-marker locations, as the posterior distribution will not adapt in the same manner. However, the same is true, to a similar degree, for HKR [13].

0.6 Information-based weighting

The intended structure of a weighting matrix presented so far, as also demonstrated in the example in eq. 12, results in all individuals being given identical weight. For the reasons already presented, this provides full orthogonality as well as simple to understand averages or effect estimates. The total phenotypic variance is also independent of the location analyzed, as the total variance is by definition independent of the distribution between the genotype classes of the model and all individuals, as already stated, share identical weight.

In some cases, however, it might be relevant to consider to what degree a specific individual can contribute to the information at a locus. As can be seen in Section 0.3, lack of information will reduce the contribution of phenotype information to the correct position, but also influence the averages for those genotypes/classes that do not reflect the true genotype of the individual. Rather than weighting each imputation by p_{ij}, we can instead choose to weight it by p_{ij}^2, where the first factor is based on the probability distribution of only the model, and the other the expectation of the real event (also estimated by our model, in lack of any other external indicator).

Furthermore, the "unexplainable" variance, due to the uncertainty in genotypes, should be accounted for, with a motivation similar to the one given for the estimation function methodology in [14]. In our case, the procedure consists of adjusting the explained variance by the aggregated information content in the population (in the loci considered). In the case of no information, the theoretical maximum explainable variance is 0. If some individuals are not fully informative, we cannot expect our model to explain the full genetic effect. When scanning for a putative QTL, one might prefer to compensate for this bias against informativeness. The IRIM method, with or without the weighting given above, will select a locus with high certainty, and so is good for reliable detection of the presence of a QTL, but the indicated location can be heavily distorted, giving preference to loci of high information content (while linked to the true QTL).

In a simple two-class case, like a typical backcross, and reasonably well for a full biallelic single-locus case, like an F_2 population, the maximum explainable variance can be expressed as the same probabilities

chosen to be our weights. This is natural, as the weights are motivated by the variations in individual locus-specific information content. The proper correction factor, with which to divide the "raw" explained variance is then a weighted average of the weights themselves, i.e. $\frac{\sum_i \left(\sum p_{ij}^2\right)^2}{\sum_i \sum_j p_{ij}^2}$. The similarities to [18] are noteworthy, although the regressions are different in structure, and we aim for a non-iterative method. The contributions from specific individuals are no longer constant in IRWIM, which is our chosen designation for this weighted individual imputation regression for interval mapping. The possibility of spurious results due to a small number outliers, deemed to be informative by the model, increases. IRWIM still allows an orthogonal decomposition of variance and effects, given a matching $\mathbf{S}$ matrix (e.g. based on the NOIA model structure), making IRWIM a superior choice over HKR. However, in the presence of segregation distortion (including apparent segregation distortion due to non-random information content patterns and sampling from finite populations), removing a locus completely from the model, rather than just removing its parameters, will modify the weights for all individuals. This changes the virtual population studied and so the effect values will be changed. The compensation for the non-explainable portion of the variance is also at this point only a heuristic.

0.7 Simulations

Simulations have been carried out to assess the performance of the different methods suggested. These were based on the marker map and population structure from an existing experimental cross. This data is based on F_2 experimental cross between Red Jungelfowl and a White Leghorn chicken line with thyroditis ("Obese strain", OS), with approximately 400 markers, only a handful of which are not biallelic SNPs (L. Andersson, personal communication). The population structure consists of nearly 800 F_2 individuals, originating from a total of 13 founders, with about 40 individuals in the F_1 generation. The marker map is relatively dense, and most missing data consists of a few randomly distributed genotypes, thus fulfilling the requirements for the HKR to be a good approximation of IM [13, 14]. For simplicity, simulations were done on the 74 markers located on chromosome 1.

From this experimental dataset, haplotypes were inferred in F_0 individuals using an internally developed and publicly released tool [19]. Based on these haplotype estimates, repeated simulated crossings were created with random cross-over events based on the exponential Haldane model. In this manner, we can simulate arbitrary genetic traits, expressed as mappings between allele line origin and resulting phenotypes, as the full simulated genotype and inheritance pathway will be known in every locus. At the same time, the use of an authentic marker map allows us to reproduce most of the statistic effects from

non-homogenous marker spacing and partially uninformative markers. As the shortcomings we are addressing in HKR are related to the handling of non-unitary genotype probabilities, any simulation needs to include these effects to allow a thorough comparison.

Simulations were done with a Gaussian environmental component and a single genetic effect, with an underdominant nature (genotype-phenotype map per locus: $11 = 0, 12 = -1, 22 = 1$), while varying the heritability h^2 with values from the set $\{0.03, 0.05, 0.07\}$. The loci with these effects were chosen randomly from the full autosomal map, with equal probability per cM. A new sampling of the recombination process giving rise to the population was performed for each replication. For each heritability value, $10,000$ such simulations were performed. In addition, the same number of null cases were tested, with identical methodology, only including the environmental component, in order to assess the behavior regarding false positives. Identical data was introduced in each run to HKR as well as IRIM, and the optimum QTL location identified. An optimum was considered a positive match if the true position was less than 25 cM upstream or downstream.

A Mersenne-Twister as implemented in the `boost` library (http://www.boost.org), with known seed for reproducibility, was used in all simulations, to avoid biases due to poor random number distributions.

Results

0.8 Algorithm

In this description, we focus on how to compute the RSS, or equivalently the residual variance, efficiently for a (W)IRIM model. General algorithms for performing QTL scans, whether by forward selection, an exhaustive search, or some more general global optimization method, is outside the scope of this communication, as we consider the evaluation of a model and an indicator of model fitness at an individual set of loci, or a single locus for a one-dimensional model.

Linear regression generally has a computational complexity on the order of $O(n^3)$. The least-squares problem can sometimes be solved more efficiently, due to a specific structure or repetitive behavior. In the context of QTL scans, for example, updateable QR factorization has been employed to reduce the actual complexity [20]. However, such updates are most efficient for continuous scans, or in cases where a high number of covariates are included.

Another option for performing the least-squares regression is to consider the separation of observations into categories, or classes. If the application only requires a test statistic based on the residual sum of squares, which is routinely done in QTL scans, it is sufficient to find the within-class inner variances. In a

non-orthogonal structure, a class decomposition of observations in this manner might be impossible. Therefore, the PERF algorithm [21], which we have adopted, was originally proposed only for a small set of cases with limited experimental relevance.

However, in the approach presented in this communication, the indicator structure of the $\mathbf{Z}$ matrices allows a decomposition of individual imputations into classes (each corresponding to a single indicator), and thus the PERF algorithm is applicable. This means that the residual sum of squares for the full model can be computed by looping over all individuals and all genotypes, resulting in $O(mn)$ complexity, where m is the number of genotypes, and n the number of individuals. More specifically, in a full model the estimate for each indicator is optimal (i.e., the arithmetic mean), so that the total unexplained variance is the sum of the variances for each set of identical imputations. The computation of variance for each indicator is completely independent, which is important for efficient parallelization, an issue of increasing importance in modern computer architectures, including multi-core and GPGPU (general-purpose computing on graphics processing units) approaches.

For reduced models, the RSS can be computed by finding the specific optimal system to solve [21] and then using that system as the basis for the classwise sums. This is highly efficient, but also cumbersome to implement. It is frequently simpler to compute the within-class variances, followed by solving a reduced-size $m \times m$ system through traditional methods. Such an approach will maintain most of the performance advantage, as frequently $m << n$, and that is the specific implementation we have chosen. Hence, each row in the new system will represent the average of all observed imputed rows for that class, and the $m \times m$ system will give the full variance between classes. This value will be translatable into the full RSS when combined with the within-class variance figures computable through the loop over classes and individuals, as described above. Thus, the full complexity will be approximately $O(mn) + O(m^3)$. While iterative solvers can frequently state a $O(n^2)$ complexity, the limited size of the system necessitates taking the constant factors for those methods in account. Furthermore, the structure of the QTL regression (as expressed in HKR) is different from those problems to which actual iterative method implementations are most commonly optimized.

0.9 Implementation

We present a short numerical example, illustrating the source of departure from orthogonality in the HKR, as well as the differences when using the same genetic model in an HKR and IRIM setting, respectively. We also present a more concrete re-analysis of existing data, showing an improvement in detection power,

and the results from our repeated simulations based on another experimental dataset.

0.9.1 Numerical example

Consider one locus with two alleles, A_1 and A_2, in a sample population of seven individuals where individual 1, 2, and 3 lack marker information. Individual 1 could be either A_1A_1 or A_1A_2 with probabilities 0.75 and 0.25, individual 2 could either be A_1A_2 or A_2A_2, with probabilities 0.75 and 0.25, and individual 3 could be A_1A_2 or A_2A_2, with equal probability. Individuals 4, 5, 6, 7 have genotypes A_1A_1, A_1A_2, A_1A_2, and A_2A_2, respectively.

The $\mathbf{Z}$ matrix in HKR is an $n \times n$ matrix, where n is the number of individuals and m the number of genotypes. In this population, this would be:

$$\mathbf{Z}_{HK} = \begin{pmatrix} 0.75 & 0.25 & 0 \\ 0 & 0.75 & 0.25 \\ 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{10}$$

Note that the first three rows contain multiple non-zero elements. The column averages of this matrix are exactly $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$. Thus, the population is an ideal F_2 population, whose genetic-effect design matrix can be written as presented in equation 2:

¿From matrices (10), (2) we can compute $\mathbf{X}^T\mathbf{X}$:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 7 & 0 & 0 \\ 0 & 2.875 & 0.25 \\ 0 & 0.25 & 1.125 \end{pmatrix} \tag{11}$$

Since $\mathbf{X}^T\mathbf{X}$ is not diagonal, the system in (1) is not orthogonal. If the marker was fully informative, then each row of $\mathbf{Z}_{HK}$ (10) would contain a single non-zero value, and so $\mathbf{X}^T\mathbf{X}$ would be diagonal as shown by [15].

By using imputations a $\mathbf{Z}$ matrix can be constructed where each row contains exactly one non-zero value. In this example, each row in 10 is split into three rows [16], and an accompanying diagonal weight matrix $\mathbf{W}$. New unitary indicators are introduced in each set of three rows (which corresponds to an individual) to result in a column frequency that equals the probability of the individual's genotype, when taking the weights into account. In this example, the resulting $\mathbf{Z}$ matrix and its accompanying weight vector $\mathbf{W}$ for

weighted imputations $(\mathbf{Z}_{IRIM}, \mathbf{W}_{IRIM})$ read:

$$\mathbf{Z}_{IRIM} = \left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right) \quad \mathbf{W}_{IRIM} = \text{diag} \left(\begin{array}{c} 0.75 \\ 0.25 \\ 0 \\ \hline 0 \\ 0.75 \\ 0.25 \\ \hline 0 \\ 0.5 \\ 0.5 \\ \hline 1 \\ 0 \\ 0 \\ \hline 0 \\ 1 \\ 0 \\ \hline 0 \\ 1 \\ 0 \\ \hline 0 \\ 0 \\ 1 \end{array}\right) \tag{12}$$

To compute $\mathbf{X}^T\mathbf{X}$ in the weighted definition, we can use the fact that what is generally expressed $\mathbf{Z}^T\mathbf{Z}$ in an unweighted ordinary least-squares regression. In the case of weighted regression, the analogous expression becomes $\mathbf{Z}^T\mathbf{W}\mathbf{Z}$. The resulting, diagonal $\mathbf{X}^T\mathbf{X}$ is:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 7 & 0 & 0 \\ 0 & 3.5 & 0 \\ 0 & 0 & 1.75 \end{pmatrix} \tag{13}$$

We can now perform two independent regressions (based on (1), (10), (12)) using $\mathbf{Z}_{HK}$ and $\mathbf{Z}_{IRIM}, \mathbf{W}_{IRIM}$, respectively, to obtain genetic estimates for the F_2 model. First, regressions are performed for the full model, describing the genotypic value of genotype "ij", G_{ij}, in terms of additive, α, and dominance, δ, deviances from the population mean μ:

$$G_{ij} = \mu + \alpha_i + \alpha_j + \delta_{ij} \tag{14}$$

We also evaluate the estimates for two reduced models:

$$G_{ij} = \mu + \alpha_i + \alpha_j \tag{15}$$

$$G_{ij} = \mu + \delta_{ij} \tag{16}$$

The models are reduced by removing the appropriate column of the $\mathbf{S}$ matrix (2). In both cases $\mathbf{G}^* = (5, 8, 8, 4, 6, 6, 9)^T$.

The estimates for both regressions are shown in Table 0.11. First, we note that the parameter estimates for the full model differ depending on whether $\mathbf{Z}_{HK}$ or $\mathbf{Z}_{IRIM}$ is used. The estimates of the additive effects differ slightly, and the dominance effects even have opposite signs. Second, the estimates of the genetic parameters are identical for full and reduced models when regressions are based on the matrix $\mathbf{Z}_{IRIM}$ (i.e. they are orthogonal), which is not the case for estimates obtained using the matrix $\mathbf{Z}_{HK}$, where the estimates of the additive and dominance effects change radically in the reduced models.

Orthogonality also makes the variances of the estimates consistent in both full and reduced models. This is why the explained variance of the full model equals the sum of the variances explained by each parameter in the complete model or the sum of the variances of the parameters in the reduced models, when computing orthogonal estimates using IRIM (see Table 0.11). These properties, which facilitate model selection, do not hold when using HKR. The estimated genetic effects are presented in Table 0.11. The estimated effect for genotype 22 is also higher than any observed phenotype value. No full-information realization based on the probabilities defined in the $\mathbf{Z}_{HK}$ matrix would render such an estimate. The estimate arises in the fully parameterized model, so it is not dependent on the $\mathbf{S}$ model matrix used.

0.10 Re-analysis of experimental data

Several QTL with different size effects on bodyweight in an intercross between Red Junglefowl and White Leghorn chicken at 112 days of age has previously been evidenced on GGA1 [22] and this data set is thus a good choice for illustrating the properties of HKR and the IRIM when applied to experimental data with actual phenotypes, as a complement to the detection performance from the more theoretical simulations described below.

Figure 0.11 shows that HKR and IRIM, as expected, produce similar LOD-score profiles in a chromosome scan of GGA1 for QTL affecting body-weight at 112 days, when adjusting the score based on a generalized extreme-value distribution derived from $1,000$ permutations of the dataset. The two curves are very similar especially for chromosomal regions where the LOD scores ranges between zero and five, where small differences in height of peaks of the LOD score function can determine whether QTL are significant or not. The IRIM shows higher peaks at marker locations close to putative QTL as discussed below. The implementation for HKR was based on mature code for the purpose, and the IRIM implementation was a simple rewrite of that code, but adapted to benefit from class-wise sums, although in a very simple and

non-optimized implementation. On the same Intel Core 2-based system, calculations for $1,000$ permutations of the range studied took 100.6 seconds with IRIM and 197.3 seconds with HKR.

0.11 Testing of detection performance in simulations

The performance of the methods has been assessed by simulations based on a real case, as explained in the methods section. In order to do this, the resulting regression scores from HKR, IRIM, and IRWIM were sorted separately. From these sorted lists, it is possible to choose any top fraction out of the total number of $10,000$ runs as accepted QTL candidates, resulting in different possible tradeoffs between the number of true positives (i.e. correctly identified loci), and the number of errors, in the known presence of a QTL in the data (incorrect positives). Furthermore, the same threshold level as determined by the fraction chosen can be used to assess false detection of QTL (false positives) in a set of data with no simulated genetic effects. The results from these two test modes are shown in Figures 0.11 and 0.11. Choosing a specific accept rate for the sorted scores is equivalent to specifying a score threshold, which just as well could be derived by e.g. permutation testing [23].

For all heritability values, IRWIM is superior to HKR, for any chosen thresholds, with a few exceptions within limited regions, while IRIM is in most regions less powerful than HKR and IRWIM. A closer study reveals that most of these errors are related to a very wide region in the start of the chromosome, with a low-information marker at the start, followed by a region about 50 cM in length with no markers at all. IRIM results will favor marker positions, resulting in a heavy over-representation of putative QTL at the first informative marker. As the distance between markers exceeds 25 cM, QTL simulated in this region are frequently classified as incorrect. Furthermore, the resulting scores for these cases are sometimes low (as the evidence of the QTL itself is limited), resulting in a more generous threshold when the threshold is based on an accept rate within a datset, and thus a higher false positive rate as well. The mean explainable variance resulting from the models were 0.0244, 0.0276, 0.0297 for IRIM, HKR, and IRWIM, respectively, in the $h^2 = 0.03$ case, illustrating the reduced "observed" variance due to the lack of information with IRIM, as well as a near-perfect assessment of the true underlying genetic variance with IRWIM.

The correct detection condition could be changed to accept any position within the same intra-marker interval as the simulated QTL, rather than the condition of a maximum distance of 25 cM. Such an analysis might improve the IRIM results in Figure 0.11, but the false positive rates presented in 0.11 would not improve. It should also be noted that all methods tend to identify the correct interval relatively frequently. When this is done, and the interval is wide, a method showing preference for center positions

within the interval, might seem superior with some evaluation criteria, as the maximum distance to any position within the interval (and, hence, the maximum distance to any likely correct QTL position) will be minimized.

Discussion

We have discussed analyzed the well-established HKR method for QTL analysis, pointing out possible issues regarding non-orthogonality and incorrect explainable variance, issues that have not previously been connected to the property of orthogonality. We have also explored a regression method based on imputations, IRIM, which can be used to obtain a form of orthogonal estimates of genetic effects in a QTL mapping experiment with missing marker data. By construction, this method can be conceptually interpreted in close relation to the classical IM [7], in which each individual's contribution to the likelihood is an aritmethic mean of the deviation from the average in the observation, conditioned over each possible genotype realization. It can be shown that the RSS contributions of the IRIM (8) can be translated into similar expressions, with the only difference from actual EM interval mapping being that the contributions from different genotype cases are slightly modified, with a geometric rather than arithmetic mean. From this, bounds can be derived on the maximum deviation from an EM IM result in different settings, based on well-known algebraic relations between the different classes of means. The relation between IM and traditional HKR is far less straightforward, although we have also discussed the link that the individual log-likelihood functions in HKR and IRIM differ by a constant (given identical phenotype estimates per genotype class).

Furthermore, IRIM comprises a better management of the uncertainity than the HKR. Indeed, expression (7) shows that the HKR assumes that the phenotype for an individual with incomplete information is normally distributed around a linear mixture of the genotypic effects for the pure genotypes. An observation for an individual with uncertain information will therefore only affect the regression with regard to the value of the weighted mean of the genotypic effects, independently of their variance. The IRIM method, on the other hand, deals with uncertainty in a more reasonable way, by making contributions to the RSS depend on the variance of the genotypic effects of individuals with uncertain genotypes. In addition, this guarantees the estimated class means of the IRIM to be the (probability) weighted averages of genotypes across individuals (which is not the case for the HKR, see Table 0.11). However, the estimated phenotype values for the classes will be compressed towards the overall average in IRIM, in cases of limited information. In HKR, the effects can be unchanged or even inflated, depending

on the specific nature of the information pattern. This is why the LOD score function of the IRIM method does not introduce artificial "bumps" in intervals between markers in Figure 0.11 (e.g. in the segments flanking the QTL peak at about 490 cM) and moreover, it improves the detection of QTL in high information regions as they will show a more distinct contrast when the extreme-value distribution (from permutations test or similar) is taken into account.

Using expression (9), we have shown that the reported unexplained variance is always higher for IRIM, relative to HKR, in cases of incomplete information. This will not, by necessity, make the IRIM more conservative in detecting QTL, since the empirical significance threshold of the HKR will also be higher. This is why the IRIM shows slightly higher peaks than the HKR in Figure 0.11, where the empirical thresholds are accounted for. This is an area which should be explored further, including the design of possible compensation schemes.

In practice, we have shown that in addition to providing orthogonal estimates and being noticeably faster, IRIM performs at least as good as HKR. We have illustrated this achievement in Table 0.11 and Figure 0.11, where IRIM achieves orthogonality in the parameter estimates of the genetic model without loss in statistical power in detection of QTL as compared to classic HKR. As IRIM preserves independent parameters, it facilitates development of statistically sound and more efficient model selection strategies for QTL detection than what is possible in approaches based on HKR.

We have further improved the IRIM to take into account the certainity of the locus-specific genotype probabilities of each individual. The resulting IRWIM enables comprenhensive fine mapping, since it further tunes the penalization of the LOD score function in the way mentioned above, which is actually analogous to the estimating equation method [14]. There is a confounding effect between environmental and genetic variance, which is not completely solved by either HKR or IRIM. A modified weighting achieves consistent estimates and a test statistic that rapidly finds likely QTL positions based on the information available. When doing optimizations to search for putative QTL, a monotonous test statistic, without any unwarranted inflections or "bumps" at all, is preferable. Such a modified statistic might not always get closer to replicate conventional EM-based interval mapping results.

For assessing the performance of the methods by simulations, we have chosen a case entailing the worst possible scenario for IRIM in relation to HKR, a map where the randomly placed QTL will frequently be found inside a region with low information content. Interestingly, however, even under this scenario, IRWIM makes less position misdetection errors than HKR for all situations checked (different heritabilities and acceptance rate). It is also worth noting that the methods are essentially equivalent in the simulations

with regard to the null error rate, the main motivation for the IRWIM is instead the fact that those QTL that are identified are located at their correct positions. IRWIM needs to be put on a more solid theoretical basis, to allow for a straightforward extension into multiple loci. While the current approach allows different models to be fitted, the total information content in all loci involved is controlling the population structure that is actually modelled, and hence influencing the estimates. In a population with none of these effects, incidentally the conditions where a multi-locus NOIA model is also orthogonal, IRWIM is a more attractive option.

We have additionally assessed the performance of the different methods using locus-specific (rather than experiment-wide) thresholds based on permutation testing [23]. This approach resulted in a slight improvement for low heritability (results not shown). However, the computational cost was prohibitive in comparison, and the relative difference between HKR and IRWIM remained. IRIM performs more favorably in this setting, while still not completely approaching the fine-mapping performance of HKR. Similar results were also generally achieved with two-locus configurations (given appropriate adjustments of the IRWIM weighting for the higher number of classes), as well as different genotype-phenotype maps, but the specific genetic structure might certainly influence the results. There is no specific reason for why the dataset used as the basis for the simulations would inherently give preference to either method, apart from the information content issues described. In line with this, short verification runs on other datasets reveal no radical differences.

In the light of the latest ongoing progress made in experimental techniques, in particular the advent of increasingly dense marker maps based on single-nucleotide polymorphisms (SNPs), it could be interpreted that the concept of IM is rapidly becoming obsolete. In some cross structures, this might be true, if local marker windows (or single markers) can consistently be used to conclusively decide allele origin, i.e. a very high information content. However, if the founder individuals share a relatively recent common ancestor (e.g. due to artificial breeding or population bottlenecks), SNP maps intended for the species at large might fail to accurately identify polymorphic regions directly. A stretch of 10s of cM in length can lack data to make a conclusive discrimination of alleles, as it is possible that the genetic polymorphisms underlying the phenotypic variation are more recent than the SNPs used for mapping. It is well-known that e.g. short tandem repeats frequently demonstrate a mutation rate multiple orders of magnitude higher than that of SNPs [24], and several quantitative traits have been linked directly to repeat-count variability. In these settings, an interval mapping approach where the allele origin is assessed based on a model of recombination and the closest linked markers with discriminating power, will be superior to simpler

window-based trait association methods.

Conclusions

In conclusion, we find that one of the possible benefits of a simple linear regression approach, in the form of orthogonal estimates, is not realized in the real-world conditions of partial information with the well-known HKR approach. We suggest an approach, IRIM, which solves this problem, but with a possibility of lower detection power in regions of low information. In a re-analysis of experimental results, we do not observe this and it might be possible to generalize a correction, which further improves on IRIM when compared to HKR.

Some of the problems arising are of a more fundamental nature. HKR and IRIM are both assuming the prior probabilities to represent posterior probabilities for genotypes. IM will in fact have different posterior probabilities, but only one such realization is considered. A more fully Bayesian approach, where the different unknown factors (genotypes per individual, effects and standard deviations per genotype class), should probably be more successful. Treating the problem as a problem of ensembles of possible realizations, weighted by their support, seems to have the most sound basis in theory, i.e. the pseudomarker approach [17]. However, the model fitted still needs to be orthogonal. Within each single-locus replicate, finding an orthogonal model will be trivial thanks to the theory of recent years [15]. The future challenge will be finding efficient and theoretically sound methods to represent multi-locus models with replicate ensembles in a biologically relevant parameterization.

Acknowledgements

Leif Andersson and Per Jensen are acknowledged for sharing experimental chicken data and Lars Rönnegård for useful technical and methodological advice regarding extreme-value distributions.

References

1. Fisher RA: **The correlation between relatives on the supposition of Mendelian inheritance**. *Trans. R. Soc. Edinburgh* 1918, (52):399–433.
2. Lynch M, Walsh B: *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, 1 edition 1998, [http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0878934812].
3. Kao CH, Zeng ZB: **Modeling epistasis of quantiative trait loci using Cockerham's model**. *Genetics* 2002, **160**:1243–1261.
4. Yang R: **Epistatis of quantitative trait loci under different gene action models**. *Genetics* 2004, **167**:1493–1505.
5. Zeng Z, Wang T, Zou W: **Modeling quantitative trait Loci and interpretation of models**. *Genetics* 2005, **169**:1711–1725.

6. Álvarez-Castro JM, Le Rouzic A, Carlborg O: **How to perform meaningful estimates of genetic effects**. *PLoS Genetics* 2008.

7. Lander ES, Botstein D: **Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps**. *Genetics* 1989, **121**:185–199, [http://www.genetics.org/cgi/content/abstract/121/1/185].

8. Haley CS, Knott SA: **A simple regression method for mapping quantitative trait loci in line crosses using flanking markers.** *Heredity* 1992, **69**(4):315–24.

9. Martínez O, Curnow R: **Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers**. *Theoretical and Applied Genetics* 1992, **85**:480–488.

10. Broman KW, Wu H, Sen S, Churchill GA: **R/qtl: QTL mapping in experimental crosses**. *Bioinformatics* 2003, **19**(7):889–890, [http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/7/889].

11. Seaton G, Haley CS, Knott SA, Kearsey M, Visscher PM: **QTL Express: mapping quantitative trait loci in simple and complex pedigrees** . *Bioinformatics* 2002, **18**(2):339–340, [http://bioinformatics.oxfordjournals.org/cgi/content/abstract/18/2/339].

12. Xu S: **A comment on the simple regression method for interval mapping**. *Genetics* 1995, **141**:1657–1659.

13. Kao CH: **On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci**. *Genetics* 2000, **156**:855–865.

14. Feenstra B, Skovgaard I, Broman KW: **Mapping quantitative trait loci by an extension of the Haley-Knott regression method using estimation equations**. *Genetics* 2006, **173**:2269–2282.

15. Álvarez-Castro JM, Carlborg O: **A Unified Model for Functional and Statistical Epistasis and Its Application in Quantitative Trait Loci Analysis**. *Genetics* 2007, **176**(2):1151–1167, [http://www.genetics.org/cgi/content/abstract/176/2/1151].

16. Jansen RC: **Maximum Likelihood in a Generalized Linear Finite Mixture Model by Using the EM Algorithm**. *Biometrics* 1993, **49**:227–231, [http://www.jstor.org/stable/2532616].

17. Sen S, Churchill GA: **A Statistical Framework for Quantitative Trait Mapping**. *Genetics* 2001, **159**:371–387.

18. Xu S: **Further investigation on the regression method of mapping quantitative trait loci**. *Heredity* 1998, **80**:364–373.

19. Nettelblad C, Holmgren S, Crooks L, Carlborg O: **cnF2freq: Efficient Determination of Genotype and Haplotype Probabilities in Outbred Populations Using Markov Models**. In *BICoB '09: Proceedings of the 1st International Conference on Bioinformatics and Computational Biology*, Berlin, Heidelberg: Springer-Verlag 2009:307–319.

20. Ljungberg K, Holmgren S, Carlborg O: **Efficient Algorithms for Quantitative Trait Loci Mapping Problems**. *Journal of Computational Biology* 2002, **9**(6):793–804.

21. Ljungberg K: **Efficient evaluation of the residual sum of squares for quantitative trait locus models in the case of complete marker genotype information**. Technical Report 2005-033, Department of Information Technology, Uppsala University 2005.

22. Kerje S, Carlborg O, Schütz K, Hartmann C, Jensen P, Andersson L: **The twofold difference in adult size between the red junglefowl and White Leghorn chickens is largely explained by a limited number of QTLs**. *Animal Genetics* 2003, **34**(4):264–274.

23. Churchill GA, Doerge RW: **Empirical Threshold Values for Quantitative Triat Mapping**. *Genetics* 1994, **138**(3):963–971, [http://www.genetics.org/cgi/content/abstract/138/3/963].

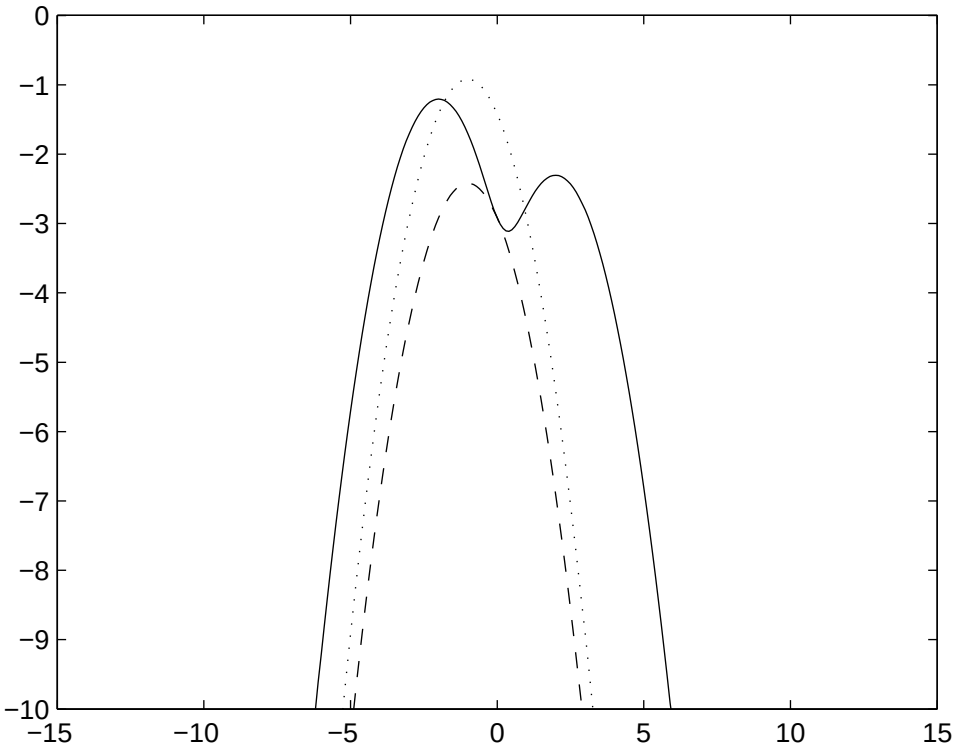
24. Amorim A, Pereira L: **Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs.** *Forensic Sci Int* 2005, **150**:17–21.

25. Rönnegård L, Besnier F, Carlborg O: **An improved method for quantitative trait loci detection of within-line segregation in F2 intercross designs**. *Genetics* 2008.

26. Valdar W, Flint J, Mott R: **Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice**. *Genetics* 2008, **172**:1783–1797.

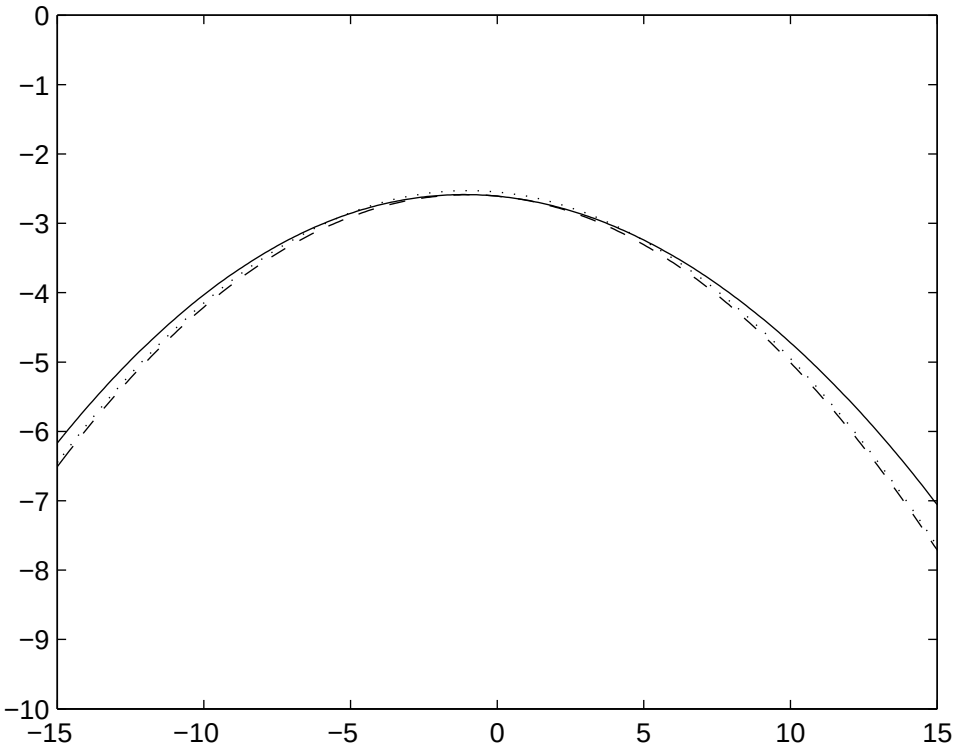
Figures

Figure 1 - Comparison of IR, HKR, and IRIM, sd = 1



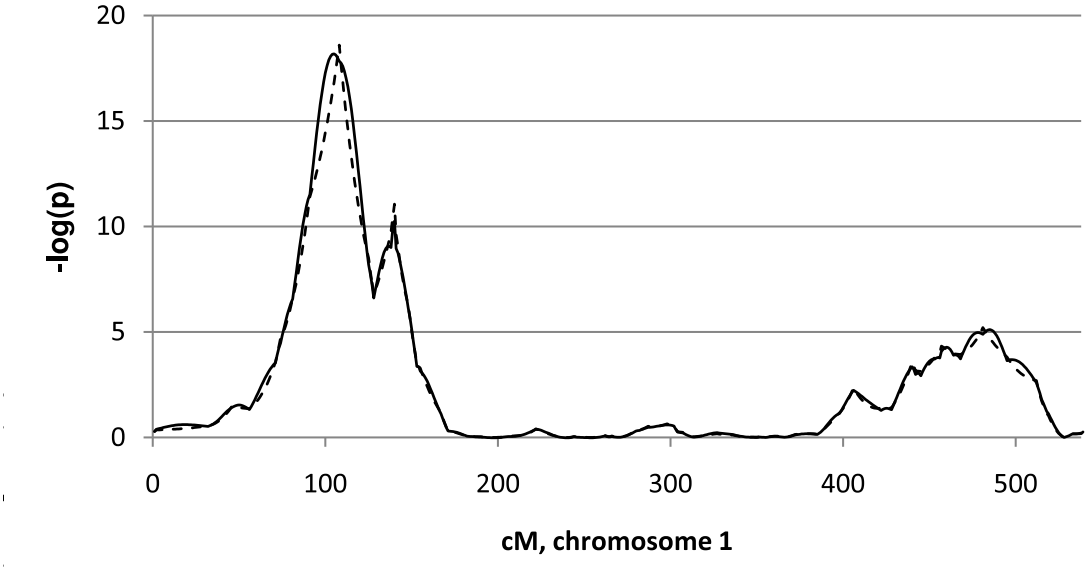
Comparison of logged likelihood contributions for a single individual with $p_{i1} = 0.75, \mu_1 = -1, p_{i2} = 0.25, \mu_2 = 1$, for IM (solid), HKR (dotted), and IRIM (dashed), with a standard deviation of 1 for all normal components. HKR can exceed the IM score in some region, while IRIM is tangent to the IM curve in the same region. The double peak in the IM curve are clearly visible.

Figure 2 - Comparison of IR, HKR, and IRIM, sd = 5



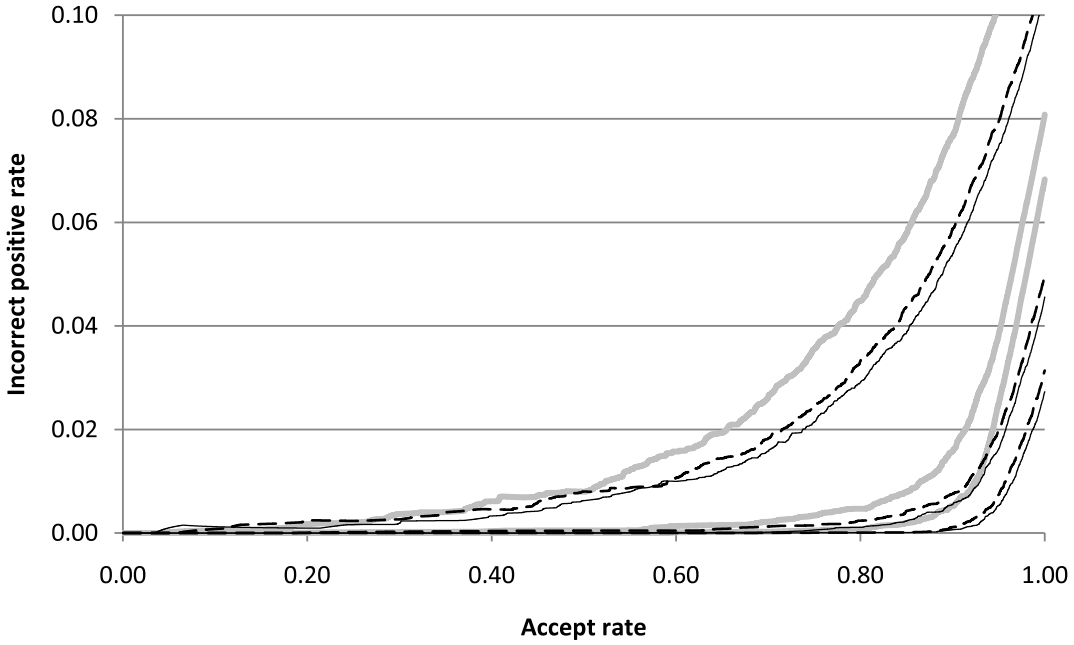
Comparison of logged likelihood contributions for a single individual with
$p_{i1} = 0.75, \mu_1 = -1, p_{i2} = 0.25, \mu_2 = 1$, for IM (solid), HKR (dotted), and IRIM (dashed), with a standard deviation of 5 for all normal components, corresponding to a low-heritability situation. HKR is exceeding IM in the center, while IRIM and IR coincide in that region.

Figure 3 - Comparison of IRIM vs. HKR for experimental data



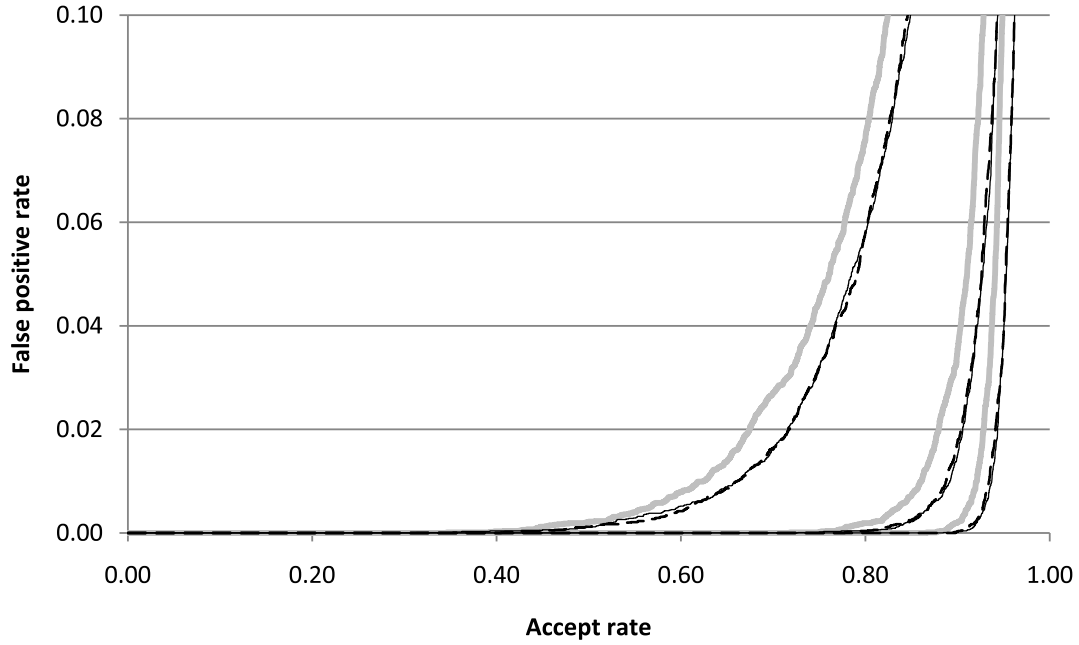
Log probability scores for IRIM (dashed line) and HKR (solid line) for a scan of GGA 1 for QTL affecting body weight 112 days after hatching in an F_2 chicken intercross between White Leghorn and Red Junglefowl where several QTL have previously been identified [22]. The statistical significance is reported as $-\log(p)$ derived from an extreme value distribution based on 1000 permutated datasets for each method [25, 26].

Figure 4 - Incorrectly identified QTL in simulation



Proportion of incorrectly identified QTL from set with one simulated pair in each run, with $h^2 = 0.03, 0.05, 0.07$ from left to right (Imputation Regression grey, Haley-Knott regression dashed, Weighted individual Imputation Regression solid). Accept rate corresponds to threshold for accepting a scan optimum, giving different tradeoffs between number of true QTL found, and controlling the error rates. Detection was considered incorrect if any locus deviated by more than 25 cM from true position.

Figure 5 - False positives in null simulations



Proportion of "detected" QTL from a dataset with no actual simulated genetic effects, for the thresholds derived from different accept rates in Figure 0.11, with $h^2 = 0.03, 0.05, 0, 07$. Imputation Regression in grey lines, Haley-Knott regression dashed, Weighted individual Imputation Regression solid.

Tables

Table 1 - Genetic effects in numerical example

Estimates and variance components of genetic effects obtained by regression using $\mathbf{Z}$ matrices based on the genotype probabilities ($\mathbf{Z}_{HK}$) and imputations ($\mathbf{Z}_{IRIM}$) for full and reduced models including additive (Reduced α) and dominance (Reduced δ) effects only. μ, α, and δ are the mean of the population and the additive and dominance effects, respectively. σ^2_{res} is the residual variance. σ^2_{sum} is the total explained variance for all parameters, either computed from the full model or as sums of the variances in the reduced models. σ^2_{expl} is the explained variance and equals the phenotypic variance (2.8163) minus σ^2_{res} .

	$\mathbf{Z}_{HK}$			$\mathbf{Z}_{IRIM}$		
	Full	Red. α	Red. δ	Full	Red. α	Red. δ
μ	6.57	6.57	6.57	6.57	6.57	6.57
α	2.55	2.52	-	2.07	2.07	-
δ	-0.34	-	0.22	0.14	-	0.14
σ^2_{res}	0.1858	0.2045	2.8084	0.6658	0.6709	2.8112
σ^2_{expl}	2.6305	2.6118	0.0079	2.1505	2.1454	0.0051
σ^2_{sum}	2.6305	2.6197		2.1505	2.1505	

Table 2 - Genotypic effects estimated by the two models

The estimated genotypic effects obtained by HKR and IRIM. The effects for $\mathbf{Z}_{IRIM}$ correspond to an arithmetic mean of the observations within each genotype class, as opposed to the effects determined using $\mathbf{Z}_{HK}$.

	$\mathbf{Z}_{HK}$	$\mathbf{Z}_{IRIM}$
11	4.19	4.43
12	6.40	6.64
22	9.30	8.57