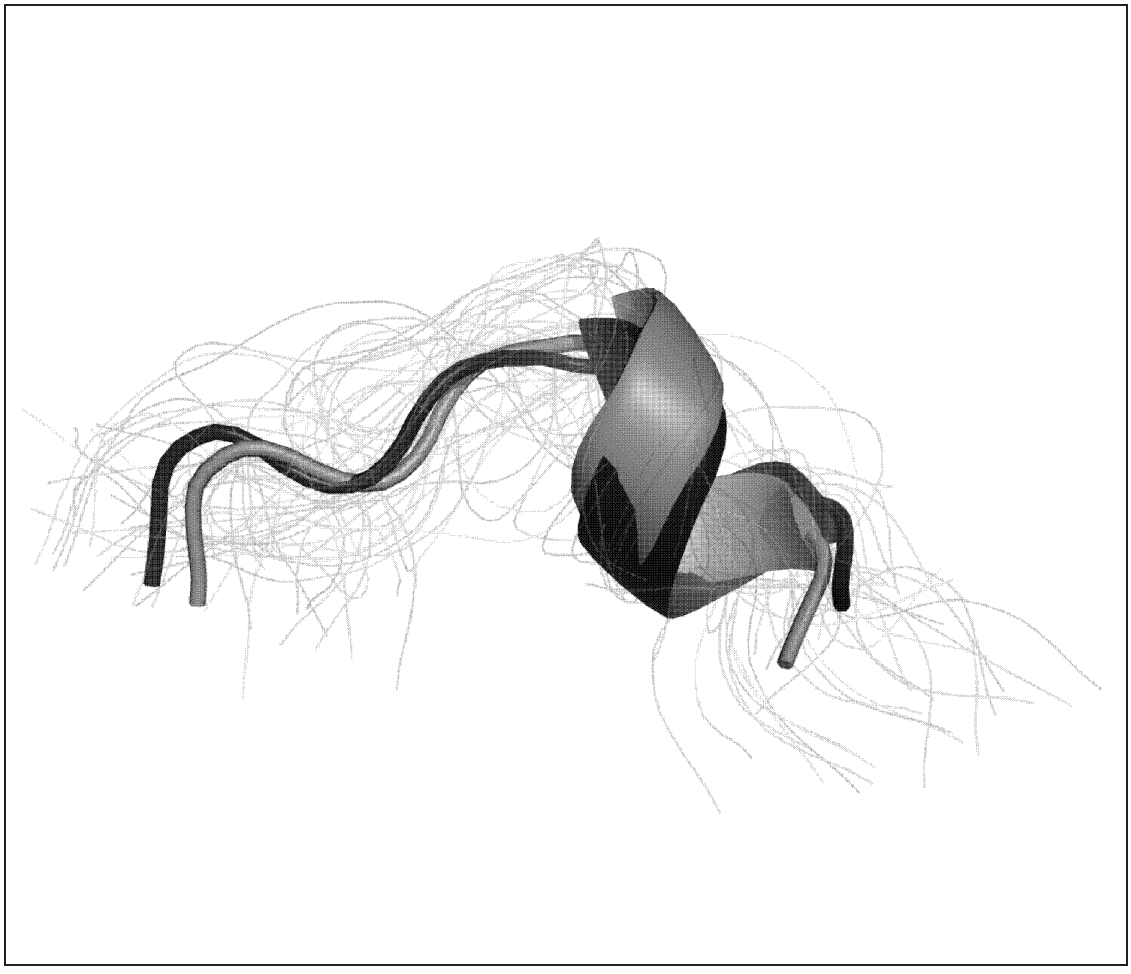
LASR 2007 — Systems Biology & Statistical Bioinformatics

4th to 6th July 2007



Programme and abstracts

Edited by S. Barber, P.D. Baxter, & K.V. Mardia

A probabilistic model of protein structure generates near-native local conformations. See p. 14 for more details. Image by: Wouter Boomsma.

ISBN **978 0 85316 263 6**

Systems Biology & Statistical Bioinformatics

International Conference, held in Leeds, UK, 4-6 July 2007,

The 26th Leeds Annual Statistical Research (LASR) Workshop

Sponsored and organized by the Department of Statistics, University of Leeds.

Sponsored by GlaxoSmithKline, `http://www.gsk.com/`

Proceedings edited by

S. Barber, P.D. Baxter, & K.V. Mardia

Department of Statistics,

University of Leeds, Leeds LS2 9JT.

Conference Organisers

S. Barber, P.D. Baxter, W.R. Gilks, R.M. Jackson, J.T. Kent,

K.V. Mardia, J.E. Shuttleworth, & D.R. Westhead

ISBN **978 0 85316 263 6**

Mapping multiple QTLs of geometric shape of the mouse mandible

Nicolas Navarro* & Christian Peter Klingenberg

Faculty of Life Sciences, University of Manchester

1 Introduction

High resolution genome scans from complex, advanced genetic designs yield many potential QTLs (Navarro & Klingenberg 2006, Valdar *et al.* 2006). However, the complex structure of the linkage disequilibrium between molecular markers and the unbalance of the population (i.e. family structure) lead to the inflation of the genetic association and the results from the initial genome scan produce many false positives. The highly structured genetic relatedness of individuals in these designs makes usual methods inapplicable. Building models incorporating multiple QTLs is therefore a necessary step in order to control for the linkage disequilibrium and haplotype structures of the population.

The problem can be stated as choosing a subset of the p candidates explaining a large amount of the genetic variance present in the q-variate phenotype, and can be formulated into a multivariate multiple regression framework. However, uncertainties in the model selection need to be incorporated (Buckland *et al.* 1996). Bayesian approaches have been proposed for these variable selection and model choice problems (e.g., Brown *et al.* 1998). Frequentist alternatives based on bootstrap aggregation (bagging, Breiman 1996), subsample aggregation (subagging, Bühlmann & Yu 2002) or bootstrap model averaging (Augustin *et al.* 2005) have been proposed or applied in a univariate multiple regression setting.

The integration of model uncertainties provides the opportunity to estimate the likelihood of a locus to be genuine QTL (see Valdar *et al.* 2006). These authors used a bagging approach and forward selection on univariate phenotypes from 8-way heterogeneous stock mice. However, redundancy in bootstrap samples implied model instability when the dimensionality of the model becomes high as we expected with shape. In this paper, we extended their approach to multivariate, highly polygenic phenotype. We stabilised the approach using subsampling instead of bootstrapping, and we proposed to average models based on parameter estimates obtained from the complete sample instead of aggregating them. We applied the proposed approach for high resolution mapping of QTLs of the mandible shape of 8-way heterogeneous stock mice where $p = 258$ potential QTLs were previously selected over the 12,092 marker intervals.

2 Building multiple QTLs models

The problem of mapping multiple QTLs can be expressed as the following multivariate linear model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{1}$$

with a $(n \times q)$ centered (for convenience) phenotypic matrix $\mathbf{Y}$, a $(n \times gp)$ design matrix $\mathbf{X}$ of g founder genotypes of the p loci, the $(gp \times q)$ matrix of genetic effects $\mathbf{B}$, and the $(n \times q)$ residual matrix $\mathbf{E}$. We set a binary p-vector γ such as its lth element is either 1 or 0 depending if the locus l is selected in the model or not. Therefore, considering this model vector, the equation

(1) can be re-written as

$$\mathbf{Y} = \mathbf{X}_{(\gamma)}\mathbf{B}_{\gamma} + \mathbf{E}, \tag{2}$$

with $\mathbf{X}_{(\gamma)}$ the (gL) subset of the columns of $\mathbf{X}$ for which $\gamma_l = 1$ and $L = |\gamma|$ the number of selected loci. The corresponding $(gL \times q)$ matrix of genetic effects is $\mathbf{B}_{\gamma}$. The least squares estimates of these effects are then

$$\hat{\mathbf{B}}_{\gamma} = (\mathbf{X}'_{(\gamma)}\mathbf{X}_{(\gamma)})^{-}\mathbf{X}'_{(\gamma)}\mathbf{Y}, \tag{3}$$

where $(\mathbf{X}'\mathbf{X})^{-}$ is the Moore-Penrose generalised inverse of the cross-product of $\mathbf{X}$.

Computational effectiveness and model dimensionality made forward selection the most suitable approach to build the model from our type of data. According to a current model state γ (e.g., without any locus incorporated, $\forall l, \gamma_l = 0$), we screened independently all remaining loci with $\gamma_l = 0$ by setting one at the time their $\gamma_l = 1$. We estimated the residual sum of squares and cross product (SSCP) matrix given this new model state $\gamma^* = \{\gamma, \gamma_l\}$ that incorporated all loci previously selected and the locus l under consideration. This new residual SSCP matrix is

$$\mathbf{E}_{\gamma^*} = \mathbf{Y}'\mathbf{Y} - \mathbf{B}'_{\gamma^*}\mathbf{X}'_{(\gamma^*)}\mathbf{X}_{(\gamma^*)}\mathbf{B}_{\gamma^*}. \tag{4}$$

The presence of a QTL in the marker interval l is evaluated using the Bartlett's approximation of the Wilk's Λ statistic

$$-(n - r_{\gamma^*} - \frac{1}{2}(d - (r_{\gamma^*} - r_{\gamma}) + 1)) \log\left(\frac{|\mathbf{E}_{\gamma^*}|}{|\mathbf{E}_{\gamma}|}\right) \sim \chi^2_{d(r_{\gamma^*} - r_{\gamma})}, \tag{5}$$

with n the sample size, d the dimensionality of the shape space, r_{γ} the rank of the cross-product of $\mathbf{X}_{(\gamma)}$ which is equal to its number of non-zero eigenvalues. The determinants of the residual SSCP matrices $|.|$ are calculated as the product of non-zero eigenvalues of the matrix. The new model state γ after this screening is chosen as the candidate model γ^* having maximal $-\log_{10}$ of the p-value ($LogP$) from the equation (5). The locus selection is stopped when this $LogP$ of the best candidate model γ^* is lower than a predefined threshold.

Given the p initial candidate loci, possible models γ are a $\{0, 1\}^p$ space (Brown*et al.* 1998). The number of possible γ-vectors is therefore 2^p which in our application (section 3) with $p = 258$ is $2^{258} \approx 4 \times 10^{77}$. We use a Monte Carlo approach based on a subsampling without replacement of $0.63 \times n$ of the original sample in order to explore this model space conditionally on the forward selection. This reduced sample size has been chosen given the expected number of unique observations in a bootstrap sample (method previously used for multiple QTL model construction with a 8-way cross, see Valdar *et al.* (2006)). According to the M γ-vectors, the selection frequency of the locus l is $h(l) = \sum_{m=1}^{M} \gamma_l^{(m)}$. This frequency $h(l)$ gives support for the candidate locus l to be an actual QTL based on its consistency in models.

According to a model averaging approach, we re-computed all γ-models using the complete sample size n before averaging genetic parameters θ according to $\bar{\theta} = \sum_{m=1}^{M} \theta_{\gamma^{(m)}} w_m$ instead to use averages of the estimates obtained from the resampling. These parameters include the allelic effect $\mathbf{B}$ but also the partial $LogP$ of each locus in the γ-model. We don't eliminate any locus with a $h(l)$ lower than some threshold and therefore give an equal weight $w_m = M^{-1}$ to each model in the averaging step on contrary to the approach proposed for bootstrap averaging that incorporate an elimination step and a second resampling procedure in order to derive model weights (see Augustin *et al.* 2005). Although these steps are straightforward to implement, loci with small selection frequency are likely to account for background genetic effects linked to the population structure.

3 Application: Mandible shape of heterogeneous stock mice

In this study, we used mice derived from the 50^{th} generation of crossing eight inbred strains. In the final generation, each chromosome is a fine grained mosaic of these eight founders (Mott *et al.* 2000). Mice were genotyped at ~ 15K single nucleotide polymorphisms (SNP) from which 12K were informative (Valdar *et al* 2006).

An interval-wide probability of the QTL alleles was obtained from a multipoint dynamic programming algorithm using the HAPPY R package (Mott *et al.* 2000). This interval-wide probability $F_{Li}(s,t)$ is the probability that the individual i descended from the founder strains s and t at marker interval l (Mott *et al.* 2000). We consider here only an additive genetic model. Therefore, we use the expected number of alleles of each ancestral strain at each locus as explanatory variables in our regression setting. This expected number of alleles from the founder s for the ith individual at the locus l is $x_i^{(s)} = \sum_{t=1}^{8} F_{li}(s,t) + F_{li}(t,s)$.

The mandible shape was described by 15 landmarks in 2 dimensions. As far as possible left and right mandibles were digitised, yielding $2,053$ individuals with one or both mandibles, of which $1,697$ individuals were genotyped. A full generalised Procrustes analysis including a reflection step to take into account the matching symmetry was performed (Dryden & Mardia 1998; Klingenberg & McIntyre 1998). The tangent coordinates were averaged over the two sides in order to remove variation due to asymmetry.

Firstly, we ran an initial genome scan using multivariable multiple regression with gender and centroid size as covariates and incorporating only one locus at a time. The association along each chromosome between the founder haplotypes and shape was returned as the $LogP$ obtained from the equation (5) with a candidate model γ^* incorporating only one locus ($L = 1$) against a reduced model γ without any genetic effect ($L = 0$). In preliminary analyses, we found massive multicollinearity problems in $\mathbf{X}_{(\gamma^*)}$. These multicollinearities are due to uncertainties in the founder probabilities within a locus arising from genetic linkage and sharing of haplotype blocks between founder strains. Therefore, we adjusted the way to compute generalised inverses of $\mathbf{X}'_{(\gamma^*)}\mathbf{X}_{(\gamma^*)}$ by relaxing the threshold on eigenvalues equal to zero.

Then, we selected locations of potential QTLs using the known linkage disequilibrium in the population (see Valdar *et al.* 2006). We defined a candidate as a local maximum $LogP$ in a window of 2Mb to the left and right, but at least 4Mb distant from other selected peaks. This candidate also has to exceed a threshold corresponding to the expected value of association between founder probabilities and shape given an infinitesimal model (i.e. a model incorporating a multitude of loci but not localisable and of small effects but yielding to the observed components of variance-covariance in the population, and keeping the pedigree structure of this population). This primo selection yielded 258 potential loci on the 19 autosomes.

On these candidates, we applied the multiple QTLs approach described above. Gender and centroid size were always incorporated as covariates in the models. We ran $1,000$ model searches using a sample size of $1,070$. We stopped the forward selection when the $LogP$ of the best candidate model γ^* (i.e. the best locus to enter) was lower than a threshold corresponding to the genome-wide null hypothesis of no genetic and no family effects. This threshold was constructed by reshuffling the phenotypic data, recording the maximum $LogP$ in the genomes and taking the 5th upper-quantile of the maximum $LogP$ distribution from $1,000$ reshufflings (Churchill & Doerge 1994). Selection frequencies of the loci $h(l)$ range from 0 to 1 with a high frequency lower than 0.1. Considering an arbitrary threshold of $h(l) \geq 0.25$, 52 loci can be considered as actual QTLs. Nevertheless, calibration of this threshold given its expected number of false positive QTLs is required and will have to be done according to the structure of genotypic and phenotypic data.

4 Conclusion

Using complex genetic designs to discover the genetic basis of quantitative traits is a challenging task. Here, we formalised and stabilised the approach for high dimensional models and multivariate traits using subsampling model averaging. Based on consistencies found in locus selection, this approach was able to identify 52 loci likely to be genuine QTLs from the initial set of 258 candidates. Therefore, this approach seems to be an alternative and promising tool compared to usual approaches that are highly confounded by the complex, unbalanced genetic relatedness of individuals in these new complex genetic crosses.

Acknowledgments

This work is funded by the Wellcome Trust. We would like to thank Jonathan Flint and Richard Mott for kindly provide the HS mice and their genotypes, and Kirsty Steward and Simon Harold for the preparation of the mandibles.

References

Augustin, N., Sauerbrei, W. & Schumacher, M. (2005). The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistical Modelling*, **5**, 95-118.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123-140.

Bühlmann, P. & Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, **30**, 927-961.

Brown, P.J., Vannucci, M. & Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society* B, **60**, 627-641.

Churchill, G. & Doerge, R. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963-971.

Dryden, I.L. & Mardia, K.V. (1998). *Statistical Shape Analysis*. Chichester, Wiley.

Klingenberg, C.P. & McIntyre, G.S. (1998). Geometric morphometrics of developmental instability: Analysing patterns of fluctuating asymmetry with Procrustes methods. *Evolution*, **52**, 1363-1375.

Mott, R., Talbot, C.J., Turri, M.G., Collins, A.C. & Flint, J. (2000). A method for fine mapping quantitative trait loci in outbred animals stocks. *Proceedings of the National Academy of Sciences USA*, **97**, 12649-12654.

Navarro, N. & Klingenberg, C.P. (2006). Combining geometric morphometrics and genomics data: The genetics of mouse mandible shape. Interdisciplinary Statistics and Bioinformatics, 63-66. Edited by S. Barber, P.D. Baxter, K.V. Mardia & R.E. Walls. Leeds University Press.

Valdar, W., Solberg, L.C., Gauguier, D., Burnett, S., Klenerman, P., Cookson, W.O., Taylor, M.S., Rawlins, J.N.P., Mott, R. & Flint, J. (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics*, **38**, 879-887.