Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Genomic Feature Models

P. Sørensen, S.M. Edwards, P. Jensen.
Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University

ABSTRACT: Whole-genome sequences and multiple trait phenotypes from large numbers of individuals will soon be available in many populations. Well established statistical modeling approaches enable the genetic analyses of complex trait phenotypes while accounting for a variety of additive and non-additive genetic mechanisms. These modeling approaches have proven to be highly useful to determine population genetic parameters as well as prediction of genetic risk or value. We present a series of statistical modelling approaches that use prior biological information for evaluating the collective action of sets of genetic variants. We have applied these approaches to whole genome sequences and a complex trait phenotype resistance to starvation collected on inbred lines from the Drosophila Genome Reference Panel population. We identified a number of genomic features classification schemes (e.g. prior QTL regions and gene ontologies) that provide better model fit and increase predictive ability of the statistical model for this trait.

Keywords: Genomic feature models; whole-genome sequences; data integration

Introduction

Whole-genome sequences and multiple trait phenotypes from large numbers of individuals will soon be available in many populations. Simultaneously a large number of genome-wide molecular profiling experiments provide molecular phenotypes (e.g. levels of RNA, protein, metabolites, phosphorylisation, glycosylation, or methylation) that are associated to the trait of interest. Genome-wide molecular interaction maps (e.g. protein-protein, protein-DNA or protein-metabolite interactions) provide insight into the structural and functional organization of the genome. These data should in principal allow a detailed molecular characterization of the genetic variability at the sequence level and should enable us to investigate several fundamental aspects of the genetic architecture of complex traits.

Evidence collected across genome-wide association studies reveals patterns that provide insight into the genetic architecture of complex traits (e.g. Lango et al. 2010). Although many genetic variants with small or moderate effects contribute to the overall genetic variation it appears that the sequence variants associated with trait variation are enriched for genes that are connected in biological pathways. Another important finding is that multiple independently associated variants are located in the same genes and that the associated variants are enriched for likely functional effects on genes such as altered amino-acid structure of proteins and expression levels of nearby genes.

Well established statistical modeling approaches enable the genetic analyses of complex trait phenotypes while accounting for a variety of additive and non-additive genetic mechanisms. These modeling approaches have proven to be highly useful to determine population genetic parameters as well as prediction of genetic risk or potential of complex trait phenotypes. Further research is required to better understand how and to what extent it is useful to use prior biological information for improving these modeling approaches.

In this paper we present a series of statistical modelling approaches that use prior biological information for evaluating the collective action of sets of genetic variants. We have applied these approaches to whole genome sequences (~ 2.5M SNPs) and a complex trait phenotype resistance to starvation collected on ~200 lines from the Drosophila Genome Reference Panel population. In addition we have access to a wealth of annotation data that can be used to link the SNPs to different types of genomic features (e.g. genes, biological pathways, prior QTL regions, gene and sequence ontologies, and so on). Our hypothesis is that there exist genomic features that provide better model fits and better predict the complex trait phenotypes.

Materials and Methods

Data. The phenotypic- and genomic data used originate from a public available reference population, the ***Dro****sophila melanogaster* **G**enetic **R**eference **P**anel (DGRP) (Mackay et al. (2012)). The population was originally caught in Raleigh, North Carolina, USA and consists of 168 fully inbred (F≈1), independent lines, and obtained using 20 generations of full-sib mating.

Initially SNPs were called from raw sequence data (as described in Mackay et al. (2012)) and included with coverage greater than 2X but less than 30X for which the minor allele frequency was present in at least 4 lines and if the SNP was called in minimum 60 lines. Missing genotypes were imputed using Beagle Version 3.3.1 software (Browning and Browning (2009)). The observed SNPs spanned a region of 23.0 Mb on 2L, 21.1 Mb on 2R, 24.5 Mb on 3L, 27.9 Mb on 3R and 22.4 Mb on X. This corresponds to 20.89 SNP pr Kb.

SNP sets were defined based on a number of genomic feature classification schemes including prior QTL information and gene and sequence ontologies.

We used resistance to starvation as phenotypes in our analysis. Resistance to starvation is a measure of how long time it takes before a fly dies due to food deprivation. Ten same sex, 2-days old flies were placed in vials containing a solution of 1.5% agar and 5 ml water to avoid the flies dies of dehydration. Every eight hour the survival rate was scored. Sample size is 17,324 observations with 10 flies in 5 vials/sex/line (Mackay et al. (2012)).

Statistical analyses

Well established statistical modeling approaches enable the genetic analyses of complex trait phenotypes while accounting for a variety of additive and non-additive genetic mechanisms. We apply three statistical modelling approaches that evaluate the collective action of sets of SNPs on the trait phenotypes using genomic features (e.g., genes, QTL regions from previous studies or biological pathways).

Approach 1: In the first approach we initially perform a genome-wide association analysis of single variants using the following linear mixed model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{s}_{\mathrm{i}} + \mathbf{e} \qquad \text{(M1)}$$

where $\mathbf{y}$ is the vector of phenotypic observations, $\mathbf{X}$ and $\mathbf{Z}$ are design matrices linking the fixed effect (sex and replicate) and random genetic effect (line) to the phenotypic records, $\mathbf{b}$, the vector of fixed effects of sex and replicate, $\mathbf{s}_{\mathrm{i}}$ is the additive genetic effect of the i^{th} SNP, the random effect of lines $\mathbf{g}\sim\mathrm{N}(0, \mathbf{I}\sigma_g^2)$, and residuals $\mathbf{e}\sim\mathrm{N}(0, \mathbf{I}\sigma_e^2)$. The additive genetic effect of each SNP was assessed by comparing the full model (M1) to a null model excluding the SNP effect using a likelihood ratio test.

The single variant analyses are followed by multi-variant analyses. This is done by grouping the single variant test statistics in sets using information on genomic features. For each set we construct an appropriate summary statistic that measures the degree of association between the set of variants and the phenotypes. We consider two summary statistics. The first summary statistic is the total number of likelihood ratio tests within a genomic feature that is above a certain threshold. The threshold for all tests was 6.635 corresponding to a marginal p-value of 0.01 assuming that the likelihood ratio test statistic has an approximate χ^2-distribution with 1 degree of freedom. The second summary statistic was the sum of all likelihood ratio test statistics belonging to the same genomic feature.

The observed the summary statistics for a particular SNP set is compared to the empirical distribution for the summary statistic of random samples of sets of SNPs. Test statistics for closely linked SNPs will likely be highly correlated due to linkage disequilibrium. This will affect the distribution of the observed summary statistics. To account for this correlation structure we used the following procedure for obtaining the empirical distribution of the summary statistic. Let the vector of observed test statistics be ordered according to the physical position on the genome for the corresponding SNPs. SNPs are mapped to genes using the coordinates for the physical location of the genes on the genome. Let the elements in this vector be numbered 1,2,…,N. The permutation consists of the following two steps. First, randomly pick an element from this vector. Let this j^{th} test statistic be the first element in the permuted vector and the remaining elements ordered j, j+1, j+2,.,N, 1, 2,…j-1 according to the original numbering. All the elements from the original vector of test statistics are now shifted to a new position. Second, a summary statistic is computed for each SNP set based on the original SNP position in the test statistic vector. In this way the link between the SNPs and the genes is broken while maintaining the correlation structure among the test statistics. These two steps are repeated k=10000 times. From this empirical distribution of the summary statistic for each SNP set a p-value can be obtained. The empirical p-value for a one-

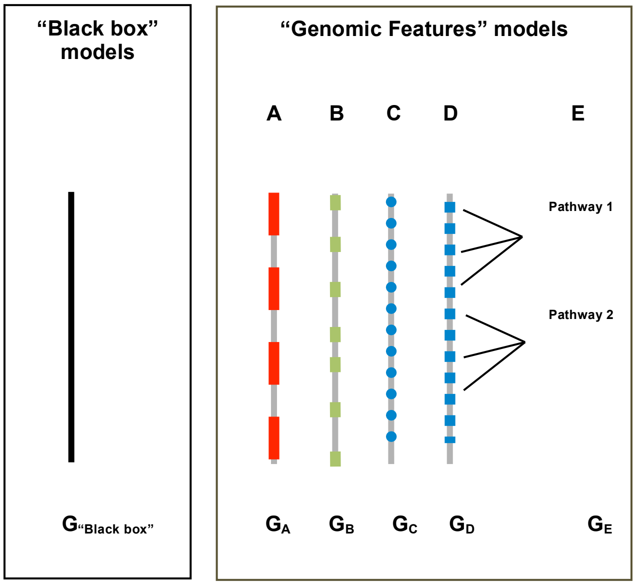

Figure 1. The "Black box" modeling approach works on the individual SNP level and treats all SNPs equally. The "Genomic feature" modeling approach accounts for the correlations among SNPs by grouping them according to genomic features such as A) location in transcriptionally active genomic regions, B) sequence-based prediction of deleteriousness, C) location in genomic regions found to be associated to the complex trait in previous GWAS studies, D) location in coding (exons, introns), non-coding or regulatory (e.g. promoters, transcription factor binding sites) regions, or E) location in genes part of biological complexes (PPI), pathways (KEGG) or modules (co-expression). Genomic parameters (G) such as variances, correlations and heritabilities are estimated for each layer of information using statistical models.

sided test is equal to the proportion of the randomly sampled summary statistic values that are larger than the observed summary statistic.

Approach 2: In the second approach we use a linear mixed model to identify genetic variation in genomic features that is associated to a complex trait phenotype (Figure 1).

This is done by fitting the following linear mixed model:

$$\mathbf{y = Xb + Z_i g_i + Z_{-i} g_{-i} + e} \quad \mathbf{(M2)}$$

where y is the phenotypic observations, X is design matrix linking, b, the vector of fixed effects of sex and replicate to the phenotypic records, $\mathrm{Z_i}$ and $\mathrm{Z_{-i}}$ are design matrices linking phenotypic records to the random genetic effects $\mathrm{g_i}$, for the set of markers belonging to the genomic feature of interest and $\mathrm{g_{-i}}$ for remaining set of markers. In this full model the random genetic effects and the residuals are assumed to be independent normally distributed variables such that $\mathrm{g_i} \sim \mathrm{N}(0, \mathrm{G_i}\sigma_i^2)$, $\mathrm{g_{-i}} \sim \mathrm{N}(0, \mathrm{G_{-i}}\sigma_{-i}^2)$, and residuals $\mathrm{e} \sim \mathrm{N}(0, \mathrm{I}\sigma_e^2)$. The corresponding additive genomic relathionship matrices $\mathrm{G_i}$ and $\mathrm{G_{-i}}$ are constructed from the subset of markers as:

$$\mathbf{G_i} = \mathbf{W_i W}_i' \frac{}{\mathbf{N_i}}$$

where $\mathrm{W_i}$ is the centred and scaled genotype matrix, and $\mathrm{N_i}$ the number of markers for i'th SNP set. The full model is compared to a reduced model where we estimate one genetic variance component defined by all the markers using the following linear mixed model:

$$\mathbf{y = Xb + Zg + e} \quad \mathbf{(M3)}$$

In the reduced model all markers are considered equal with respect to their contribution to the genetic variance. The linear mixed model approach allows us to use a likelihood ratio test to compare different genomic feature-based partitioning of the genomic variance. Likelihood ratios were calculated as twice the difference between the log-transformed restricted likelihood of the simple model (M3) and full model (M2). It has been shown that the likelihood ratio test statistic follows a mixture of $\chi 2$ distributions with one or two degrees of freedom (Self & Liang (1987)). An alternative way to assess the importance of the $\mathrm{i^{th}}$ SNP set is the proportion of explained genetic variance of the $\mathrm{i^{th}}$ set defined as $H_i^2 = \sigma_i^2/(\sigma_i^2 + \sigma_{-i}^2)$. Finally the importance was assessed using a tenfold cross validation procedure.

Approach 3: The third approach builds on the linear mixed model (M3) described in the previous section. The difference is that the genetic parameters are estimated using an iterative REML procedure (Wang et al. (2012)) based on a weighted genomic relationship matrix, G, constructed using all SNP markers as:

$$\mathbf{G = WDW'/N,}$$

where W is the centred and scaled genotype matrix, D is a diagonal matrix containing the weight for each SNP, and N is the sum of the diagonal elements D. The SNP weights were initially set to unity. In subsequent iterations, each SNP was weighted according to its variance contribution equal to the squared SNP effect. The individual SNP effects were obtained from:

$$\mathbf{\hat{b} = DW'(WDW')^{-}\hat{g},}$$

where $\hat{\mathrm{b}}$ is the vector of estimated SNP effects. In each iteration the log-likelihood for the fitted model was determined and this iterative procedure was repeated until we observed a decrease in model fit as determined by a decrease in the log-likelihood. During this process the values of $\hat{\mathrm{b}}$ become more extreme and should result in SNPs that are causative, or highly correlated to the causative genetic variant, having a high weight in the model disregarding whether the effect on the trait is positive or negative. We determined a genetic value for each SNP set defined by the genomic feature using:

$$\mathbf{\hat{g} = \hat{g}_\iota + \hat{g}_{-\mathit{i}} = W_{\mathit{i}}\hat{b} + W_{-\mathit{i}}\hat{b}}$$

where $\hat{\mathrm{g}}_i$ is the genetic value associated to the i'th SNP set and $\hat{\mathrm{g}}_{-i}$ denotes the genetic values associated to the remaining SNPs. From these partitioned genetic values we decomposed the genomic variance using:

$$\mathbf{Var(\hat{g})} = \begin{bmatrix} \mathbf{Var(\hat{g}_i)} & \mathbf{Cov(\hat{g}_i, \hat{g}_{-i})} \\ \mathbf{Cov(\hat{g}_{-i}, \hat{g}_i)} & \mathbf{Var(\hat{g}_{-i})} \end{bmatrix}$$

We determined the relative importance of the SNP set as the average genomic variance explained by each SNP in the set calculated as:

$$\boldsymbol{\gamma}(\mathbf{\hat{g}}_{\mathit{i}}) = \mathbf{Var(\hat{g}_i)/N_i}$$

This approach gives us a framework where we can easily decompose the variance contributed by different types of genomic feature classification schemes.

Implementation. The various genomic feature modelling approaches was implemented in R (R Core Team. (2013)) using the software package DMU (Madsen & Jensen (2012)) as computational engine for some of the variance components analysis.

Results and Discussion

In this paper we have used three statistical modelling approaches that evaluate the collective action of sets of SNPs on the trait phenotypes. Using a linear mixed modeling approach we partitioned he genomic variance into components defined by Sequence Ontology (e.g. intron, exon), degree of association as determined in a traditional genome-wide association analysis (e.g. associated or not associated)

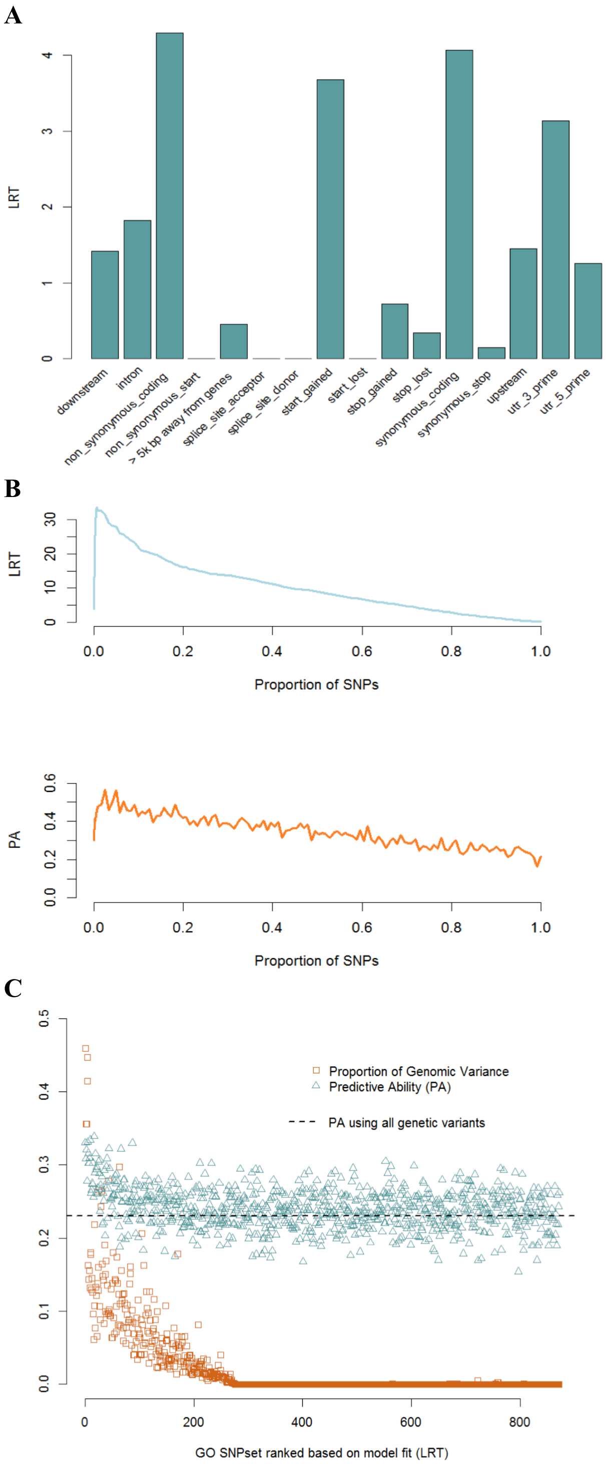

Figure 2. A) A linear mixed model approach for partitioning of genomic variance using prior biological information. A) Partitioning using Sequence Ontology information revealed that non-synonymous genetic variants explained a large proportion of the genomic variance. B) SNPs ranked according to their degree of associated to the complex trait phenotypes in a genome-wide association analysis provide a better model fit (LRT) and increase the predictive ability (PA) of the statistical model. C) Partitioning using Gene Ontology information identified several terms that explained a large proportion of the genomic variance and increased the predictive ability compared to a statistical model that include all genomic variants in one variance component.

Partitioning using Gene Ontology (GO) information identified several genomic features that explained a large proportion of the genomic variance and increased the predictive ability compared to a statistical model that include all genomic variants in one variance component.

The genomic feature model analyses were implemented using standard mixed model methodology and estimation of variance components was done using a Restricted Maximum Likelihood approach. The linear mixed model approach is used to identify genetic variation in gene groups that is associated to a complex trait phenotype. This is done by fitting and comparing two statistical models. In the full model we estimate two genetic variance components; one variance component for the set of markers defined by the random gene group and one variance component for remaining set of markers. In the reduced model we estimate one genetic variance component defined by all the markers. In the reduced model all markers are considered equal with respect to their contribution to the genetic variance whereas in the full model we allow the two markers sets to be weighted differently. The weight is proportional to the proportion of variance explained by the marker set estimated from the data being analyzed. We use a likelihood ratio test to compare these two models. A high likelihood ratio shows that the model with two (different) variance components is better at explaining the observed genetic variance than the simple model with only one variance component. We interpret this as the ability to separate the 'genetic signal' (governing the complex trait) from the background noise of the genome, i.e. a high likelihood ratio has successfully separated the signal from the noise. In our analyses we found in that a partitioning that provided a better model fit also leads to a better predictive ability of the statistical model. This result may depend on the population structure being investigated.

There are several alternative statistical modeling approaches that can be used to evaluate the collective action of multiple genetic variants in genomic features (Wang et al. (2010), Wu et al. (2011), Newton et al. (2007), Jiang & Gentleman (2007), Ehsani et al. (2012), Silver & Montana (2012), Friedlye & Biernacka (2012)). These approaches can be classified according to the input data required, statistical method used and the null hypothesis being tested. We compared our linear mixed model approaches to a commonly used two-step approach and found a relatively good agreement between the approaches (Figure 3). However, the variance decomposition based on the iterative REML procedure requires further investigation.

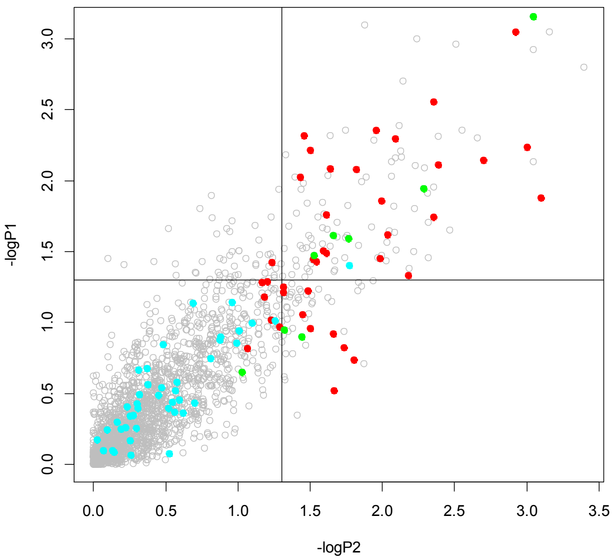

Figure 3. Comparison of test statistics from three statistical modeling approaches for evaluating the collective action of multiple genetic variants in genomic features. For the two step approach there was a high correlation between the minus log of the empirical p-values for summary statistic 1 (-logP1) and for summary statistic 2 (-logP1) using genomic features defined by Gene Ontology. For both summary statistics a marginal p value cutoff of 0.05 is indicated by a horizontal or vertical black line. In comparison we have plotted results based on the linear mixed approaches showing genomic features that 1) provide better model fit (M2; LRT>3.84) (red dots), 2) with the highest average genomic variance explained by each SNP (M3, Top50) (light blue dots), or 3) fulfil both criteria (LRT>3.84 & Top50) (light green dots).

The two-step approach is widely used because it is computationally fast and can easily be combined with association results obtained from previous GWAS. In the first step, a test statistic for the association (e.g. t-statistics, p-values) of individual genetic variants with the trait phenotype is obtained from traditional single-marker or all-marker statistical models. In the second step, for each genomic feature being tested, a summary statistics is obtained. The summary statistics should reflect the degree or level of association. We considered two summary statistics. The first summary statistic was the total number of likelihood ratio tests within a genomic feature that is above a certain threshold. Alternatively a Hyper geometric test can be used to compare the frequency of significantly associated variants located within or outside the genomic feature (Jiang & Gentleman (2007), Newton et al. (2007)). However, there is the arbitrariness of the threshold for determining "significantly associated", no matter how it is chosen and genetic variants whose test statistics differ by a tiny amount may be treated completely differently. By design this test will have high power to detect association if the genomic feature harbor genetic variants with large effects, but it will not detect a situation where there are many genetic variants with small to moderate effects. In this case, it is more powerful to use a summary statistic such as the mean or sum of the test statistic for all genetic variants belonging to the same genomic feature (Newton et al. (2007)).

Conclusion

We identified a number of genomic features classification schemes (e.g. prior QTL regions and gene ontologies) that provide better model fit and better predicted the complex trait phenotype, resistance to starvation, in *Drosophila melanogaster*.

Acknowledgement

The presented work was done as part of the Quantomics a collaborative project under the 7th Framework Programme (contract no. KBBE-2A-222664) and the Danish Strategic Research Council (GenSAP: Centre for Genomic Selection in Animals and Plants, contract no. 12-132452).

Literature Cited

Lango Allen, H., Estrada, K., Lettre, G., et al. (2010). *Nature*. 467: 832–838.

Madsen, P., and Jensen, J. (2012). DMU. Version 6, release 5.1.

Wang, H., Misztal, I., Aguilar, I. et al. (2012). *Genet. Res.* 94: 73–83.

Mackay, T. F. C., Richards, S., Stone, E. A., et al. (2012) *Nature* 482:173–178

Self, S. G. and Liang, K. Y. (1987) *J. Amer. Statist. Assoc.* 82:605-610

Browning, B. L. and Browning, S. R. (2009). *Am. J. Hum. Genet.*, 84(2):210–223.

Carlson, M. (2013). org.Dm.eg.db: Genome wide annotation for Fly. R package version 2.9.0.

Jiang, Z. and Gentleman, R. (2007). *Bioinformatics*, 23(3):306-313.

Newton, M., Quintana, F. A., Boon, J. A., et al (2007). *Ann. Appl. Stat.*, 1(1):85-106.

R Core Team. (2013). Version 3.0.2.

Wang K, Mingyao Li & Hakon Hakonarson (2010). Nature Reviews Genetics 11, 843-854

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. (2011). American Journal of Human Genetics , 89, 82-93.

Newton MA, Quintana FA, Den Boon JA, Sengupta S, Ahlquist P (2007). Journal of Computational and Theoretical Nanoscience, 1: 85-106.

Jiang Z, Gentleman R (2007). Bioinformatics 23: 306-313.

Ehsani A, Sørensen P, Pomp D, Allan M, Janss L (2012). BMC Genomics. 13:456.

Silver M, Montana G. (2012). Stat Appl Genet Mol Biol. 11(1):Article 7.

Fridley BL, Biernacka JM. (2011). Eur J Hum Genet. 19(8):837-43.