Genome-wide detection of intervals of genetic heterogeneity associated with complex traits*

Felipe Llinares López [1], Dominik G. Grimm [1], Dean A. Bodenham [1], Udo Gieraths [1], Mahito Sugiyama [2,3], Beth Rowan [4], Karsten M. Borgwardt [1]

[1]*Machine Learning and Computational Biology Lab, Department of Biosystems Science and Engineering, ETH Zurich, Switzerland* [2]*The Institute of Scientific and Industrial Research, Osaka University, Japan* [3]*JST, PRESTO, Japan* [4]*Department of Molecular Biology, Max Planck Institute for Developmental Biology, Germany*

contact email address: udo.gieraths@bsse.ethz.ch

*accepted for oral presentation at ISMB 2015

ABSTRACT Genetic heterogeneity is the phenomenon that several distinct sequence variants may give rise to the same phenotype (Burrell *et al.*, 2013). This phenomenom is of the utmost importance to the exploration of the genetic basis of complex phenotypes, as most of them have been found to be affected by numerous loci, rather than a single locus (McClellan and King, 2010).

Current approaches for finding regions in the genome that exhibit genetic heterogeneity suffer from at least one of two shortcomings: 1) they require the definition of an exact interval in the genome that is to be tested for genetic heterogeneity, potentially missing intervals of high relevance, or 2) they suffer from an enormous multiple hypothesis testing problem due to the large number of potential candidate intervals being tested, which results in either many false positive findings or a lack of power to detect true intervals.

To illustrate the scale of this multiple testing problem in genetic heterogeneity search: When one considers all possible intervals in a genome in a dataset with 10^6 SNPs, the number of tests one performs is quadratic in the number of SNPs, that is approximately $5 \cdot 10^{11}$ candidate intervals. When ignoring the multiple testing problem, one will obtain billions of false positives. If one performs the standard Bonferroni correction (Bonferroni, 1936), which divides the significance threshold α (typically 0.05 or 0.01) by the number of tests, then the corrected threshold will be so low that hardly any finding will be statistically significant.

We propose an algorithm for genome-wide detection of contiguous intervals that may exhibit genetic heterogeneity with respect to a given binary phenotype. More specifically, we search for genomic intervals in which the occurrence of at least one type of sequence variant (e.g. a point mutation or minority allele) is significantly more frequent in one of the two phenotypic classes. Figure 1 illustrates this matter.

Our algorithm, Fast Automatic Interval Search (`FAIS`), automatically finds the starting and end positions of these intervals, while properly correcting for multiple hypothesis testing and preserving statistical power. Central to this algorithm is an approach by Tarone (Tarone, 1990), which allows one to reduce the Bonferroni correction factor for multiple hypothesis testing. Additionally we extended `FAIS` to a Westfall-Young permutation based version called `FAIS-WY`. In practice, `FAIS-WY` is more computationally demanding than `FAIS` but has increased statistical power.

We employ our novel algorithms on simulated data as well as on *Arabidopsis thaliana* GWAS data. In the simulations our algorithms outperform in terms of

power the brute force approach using Bonferroni correction as well as an approach using univariate Fisher's Exact Test (UFE) that only checks for a significant difference in single SNPs. For the *Arabidopsis thaliana* GWAS data, out of 21 binary phenotypes we were able to discover intervals of SNPs that are associated with 14 of these phenotypes, but could not be found with previous methods. The comparison is done to the univariate Fisher's Exact Test (UFE) and a state-of-the-art linear mixed model (LMM) to account for confounding due to population structure (Lippert *et al.*, 2011). The Proportion of novel intervals among all intervals found by `FAIS-WY`, across all phenotypes is visualized in figure 2.

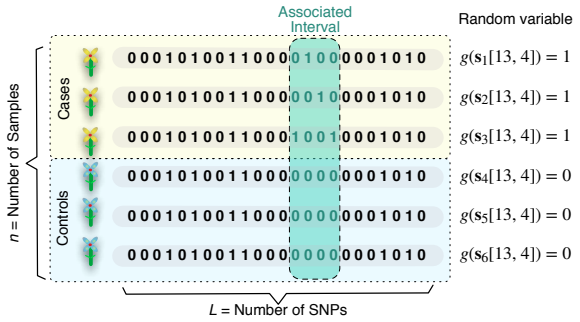

Figure 1: Schematic illustration of the problem of detecting genomic intervals that may exhibit genetic heterogeneity. $\mathbf{s}_i[\tau; l]$: interval of length l, starting at index τ of the i-th genomic binary sequence , $g(\mathbf{s}_i[\tau; l]) = \mathbf{s}_i[\tau] \vee \mathbf{s}_i[\tau+1] \vee \ldots \vee \mathbf{s}_i[\tau+l-1]$, where $\vee$ denotes the binary OR operator. The problem to solve is that of finding all intervals (τ, l) with $l = 1, \ldots, L$ and $\tau = 0, \ldots, L-l$ such that the random variable $g(\mathbf{s}[\tau; l])$ is statistically associated with the phenotype $y \in \{\text{Cases}, \text{Controls}\}$ after correction for multiple hypothesis testing.

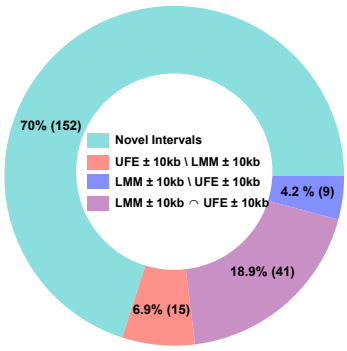

Figure 2: Proportion of novel intervals among all intervals found by `FAIS-WY`, across all phenotypes. The green part shows the proportion of novel intervals found by `FAIS-WY`. The red part (UFE ± 10kb \LMM ± 10kb) are intervals containing an UFE hit or are in close proximity (± 10kb) to one and the hit could not be found with a LMM. The blue part (LMM ± 10kb \UFE ± 10kb) are intervals containing a LMM hit or are in close proximity (± 10kb) to one and the hit could not be found with an UFE. The purple part (LMM ± 10kb ∩ UFE ± 10kb) are intervals that contain both, a hit (± 10kb) found with an UFE and a LMM.

References:

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, **501**(7467), 338–345.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat Meth*, **8**(10).

McClellan, J. and King, M.-C. (2010). Genetic heterogeneity in human disease. *Cell*, **141**(2), 210–217.

Tarone, R. E. (1990). A modified bonferroni method for discrete data. *Biometrics*, **46**(2), 515–522.