`FastCMH`: Genome-wide genetic heterogeneity discovery with categorical covariates

Felipe Llinares-Lopez ^{1,2,†,*}, Laetitia Papaxanthos ^{1,2,†,*}, Dean Bodenham ^{1,2},
Damian Roqueiro ^{1,2}, COPDGene Investigators ^{3}, Karsten Borgwardt ^{1,2,*}

^{1} Machine Learning and Computational Biology Lab, D-BSSE, ETH Zurich, Switzerland
^{2} SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland, ^{3} COPDGene® Study
† Equally contributing authors.
* To whom correspondence should be addressed.

Univariate Genome-Wide Association Studies (GWASs) are a widely used approach to retrieve Single-Nucleotide Polymorphisms (SNPs) significantly associated with a phenotype of interest, such as the presence or absence of a disease, eye color or the size of bones. However, the phenotypic variance explained by findings of univariate GWASs is typically considerably lower than the estimated heritability of the corresponding trait, leading to the well know "missing heritability" problem. A partial explanation to this phenomenon is the existence of genetic heterogeneity, by which several SNPs have a common but weak effect on the phenotype of interest.

In our recently published paper *Genome-wide genetic heterogeneity discovery with categorical covariates* Llinares-Lopez *et al.* (2017), we propose `FastCMH`, an algorithm that allows to detect statistically significant associations between genomic regions and the phenotype of interest, while correcting for categorical confounders. Our method uses an state-of-the-art pattern mining approach to test *all* genomic regions, without imposing a maximum region length, while being computationally efficient and showing a high statistical power.

1 Introduction

The association between single SNPs and a phenotype of interest are mainly assessed through Genome-wide Association Studies (GWASs) (Wellcome Trust Case Control (2007)). However, the signal carried by single SNPs is often not strong enough to be detected by univariate studies. Genetic heterogeneity is a concept that can partly explain the existence of weak signals: it assumes that several nearby SNPs have a joint effect on the trait of interest (Burrell *et al.* (2013)). However, studying the association of genomic regions or sets of contiguous SNPs can be computationally expensive and statistically challenging because of the resulting high number of regions to test. To give a sense of scale, a typical association mapping contains 500, 000 SNPs, which leads to a total number of sets of contiguous SNPs in the order of 500 billion. Therefore, existing methods restrain the search to specific biological regions, such as exons or genes, or fixed-size genomic regions (Lee *et al.* (2014)).

In a recent study, Llinares-López *et al.* (2015a) presented `FAIS`, an algorithm that allows to test for all genomic regions in GWASs and that overcomes the runtime complexity and statistical challenges. In order to gain efficiency, the algorithm uses the key concept of *testability*, introduced by (Tarone (1990)), which differentiates hypotheses that cannot be significant -*untestable* hypotheses- from others -*testable* hypotheses- before computing the corresponding statistical test. The concept of testability, first, allows to reduce the runtime complexity by pruning untestable hypotheses and, second, leads to a better statistical power as the untestable hypotheses will not be taken into account in the multiple hypothesis testing correction, without leading to any additional false positives.

However, FAIS does not allow to correct for covariates, such as sex, age or population stratification, which are the cause of many spurious associations in GWASs, therefore limiting its applicability to datasets for which confounding effects can be rule out a priori. A recently published method (Papaxanthos *et al.* (2016)), used the Cochran-Mantel-Haenszel (CMH) together with Tarone's concept of testability, to correct for categorical covariates when testing for *all* combinations of features. However, this setup is not applicable to GWASs as it would not scale to hundreds of thousands of SNPs.

Here, we present `FastCMH`, a novel method that combines the pattern mining search strategy of `FAIS` with the correction for categorical covariates of (Papaxanthos et al. *(2016)), without compromising the scalability to genome-wide association mapping.*

2 Detection of significant genomic regions in the presence of covariates

2.1 Notation

We aim at discovering genomic regions that are significantly associated with a binary phenotype of interest y. Consider a genomic dataset of interest with n individuals split as n_1 cases and $n_2 = n - n_1$ controls. Each individual i is represented by an ordered genomic sequence g_i of d genomic markers $g_i = (g_i[1], \ldots, g_i[d])$ with $g_i[1] \in \{0, 1\}$. The genomic markers can be SNPs that have been encoded with a dominant or recessive model or can represent the absence or presence of a genomic function. Additionally, we assume that for each individual i, c_i records a categorical covariate with k states.

2.2 Accounting for genetic heterogeneity

Under the assumption of genetic heterogeneity, several contiguous SNPs might jointly influence the phenotype of interest in the same manner, while presenting rather weak signals when studied separately. Exploiting this phenomenon could help explain a part of the the missing heritability described in most GWASs. Unlike existing approaches, such as burden tests, we aim to test *all* possible genomic regions for association, without prior assumptions on location or size.

Let $[\![t_s, t_e]\!]$ be a genomic region with starting and ending positions t_s and t_e, hence having length $m = t_e - t_s + 1$. For each individual i, we define the *meta-marker* $\tilde{g}_i([\![t_s, t_e]\!])$ of region $[\![t_s, t_e]\!]$ as $\tilde{g}_i([\![t_s, t_e]\!]) = 1$ if any of the markers within the region is encoded as a 1. On the contrary, the meta-marker $\tilde{g}_i([\![t_s, t_e]\!]) = 0$ if all markers in the region are encoded as 0. By pooling together SNPs in this manner, the meta-marker $\tilde{g}([\![t_s, t_e]\!])$ might exhibit a stronger associative signal with the phenotype than each of the single SNPs in region $[\![t_s, t_e]\!]$ separately.

2.3 Correcting for covariates with the Cochran-Mantel-Haenszel (CMH) test

For each genomic region $[\![t_s, t_e]\!]$, we need to test if its meta-marker $\tilde{g}([\![t_s, t_e]\!])$ and the phenotype y are statistically associated given the covariate c.

In order to do so we use the Cochran-Mantel-Haenszel (CMH) test (Cochran, 1954; Mantel and Haenszel, 1959) that is based on contingency tables. As it is a conditional test, it builds one contingency table per category of the covariate, unlike Fisher's Exact Test (Fisher, 1922) and Pearson's χ^2 test (Pearson, 1900) which need a unique table. For each 2×2 contingency table h, with $h = 1, \ldots, k$, cell counts are computed based on all individuals for which $c_i = h$:

Variables	$\tilde{g}_i([\![t_s, t_e]\!]) = 1$	$\tilde{g}_i([\![t_s, t_e]\!]) = 0$	**Row totals**
$y =$ case	a_h	$n_{1,h} - a_h$	$n_{1,h}$
$y =$ control	$x_h - a_h$	$n_{2,h} - x_h + a_h$	$n_{2,h}$
Col. totals	x_h	$n_h - x_h$	n_h

Here n_h is the number of individuals with $c_i = h$, divided into $n_{1,h}$ cases and $n_{2,h}$ controls. Similarly, x_h is the number of individuals with $c_i = h$ for which the meta-marker $\tilde{g}_i([\![t_s, t_e]\!])$ takes value 1, a_h of which are cases and $x_h - a_h$ controls. Using the cell counts $\{n_h, n_{1,h}, x_h, a_h\}_{h=1}^{k}$, we can compute the p-value $p([\![t_s, t_e]\!])$ for genomic region $[\![t_s, t_e]\!]$ under the CMH test as explained in Llinares-Lopez *et al.* (2017). A genomic region $[\![t_s, t_e]\!]$ is found to be significantly associated with the phenotype y given the covariate c if $p([\![t_s, t_e]\!]) \leq \delta$, where δ is the adjusted significance threshold.

2.4 The concept of *testability* and of *minimum attainable P-value*

Tarone's concept of *testability* relies in the ability to compute a minimum attainable P-value p_{min} for each hypothesis as a function of the margins of the contingency table. If an association shows a minimum attainable P-value smaller than the adjusted significance threshold, the hypothesis is considered *testable*, if it is larger it is considered *untestable* as the actual P-value of the hypothesis –larger than p_{min}– can never be smaller than δ. It has been shown that p_{min} can be written in closed-form for the CMH test. More details are available in Llinares-Lopez *et al.* (2017).

3 Description of FastCMH

3.1 FastCMH is organized in three main steps

The high-level pseudo-code is shown below in Algorithm 1.

Algorithm 1 FastCMH

```
Input: Dataset $\mathcal{D} = \{\mathbf{g}_i, y_i, c_i\}_{i=1}^{n}$, desired FWER $\alpha$
Output: Set of non-overlapping conditionally associated genomic regions $\mathcal{R}_{sig,filt} = \{[\![t_s, t_e]\!] \,|\, p([\![t_s, t_e]\!]) \leq \delta_{tar}\}$
1: $(\delta_{tar}, \mathcal{R}_{\mathcal{T}}(\delta_{tar})) \leftarrow$ get_testable_regions($\mathcal{G}, \alpha$)
2: $\mathcal{R}_{sig,raw} \leftarrow \{[\![t_s, t_e]\!] \in \mathcal{R}_{\mathcal{T}}(\delta_{tar}) \,|\, p([\![t_s, t_e]\!]) \leq \delta_{tar}\}$
3: $\mathcal{R}_{sig,filt} \leftarrow$ filter_overlapping_regions($\mathcal{R}_{sig,raw}$)
4: Return $\mathcal{R}_{sig,filt}$
```

The algorithm takes as input the genomic sequence $\mathbf{g}_i$, binary phenotype y_i and categorical covariate c_i for $i = 1, \ldots, n$ individuals, as well as the desired FWER α.

The main routine of FastCMH is get_testable_regions, Line 1. It screens the search space for testable regions of increasing length, under the CMH test. To do so efficiently, it iteratively updates the significance threshold δ from 1 to δ_{tar} the final significance threshold. If we denote $\mathcal{R}_{\mathcal{T}}(\delta)$ the set of testable regions at the significance threshold δ, this routine solves the following optimization problem: $\delta_{tar} = \max\{\delta | \delta < \alpha / |\mathcal{R}_{\mathcal{T}}(\delta)|\}$. It ensures that the FWER α is controlled as closely as possible. get_testable_regions outputs δ_{tar} and the set of testable regions $\mathcal{R}_{\mathcal{T}}(\delta_{tar})$. The routine get_testable_regions is described in detail in Section 3.2 below.

Once δ_{tar} is obtained, Line 2 of the algorithm evaluates the association of the testable regions $[\![t_s, t_e]\!] \in \mathcal{R}_{\mathcal{T}}(\delta)$ with the phenotype of interest under the CMH test. It returns the set of statistically significant regions $\mathcal{R}_{sig,raw}$ and their corresponding P-values $p(\llbracket t_s, t_e \rrbracket)$ for $\llbracket t_s, t_e \rrbracket \in \mathcal{R}_{sig,raw}$. Only computing the P-values of the testable regions is key to reduce the computational runtime without causing additional false-negatives.

The set of statistically significant genomic regions $\mathcal{R}_{sig,raw}$ is often composed of several overlapping regions, due to the use of exhaustive search and to linkage disequilibrium (LD) in genomic data. In order to correct for the redundancy of the results, the routine `filter_overlapping_regions`, Line 3, clusters significant regions in $\mathcal{R}_{sig,raw}$ that overlap and selects, as a representative of each cluster, the most significant region. As a consequence, it outputs the set of non-overlapping conditionally associated genomic regions $\mathcal{R}_{sig,filt}$ to be used for further analysis.

3.2 Detailed description of `get_testable_regions`

The routine `get_testable_regions` of `FastCMH` combines a branch-and-bound approach used in `FAIS` with the implementation of a lower envelop to the minimum attainable p-values of the CMH test, designed in Papaxanthos *et al.* (2016), that allows an efficient pruning. This reduces strongly the search space of candidate regions $\mathcal{R}_{cand}$. For example, a naive implementation would compute the P-values of all genomic regions, i.e. performing $\frac{d(d-1)}{2} = O(d^2)$ tests, which is hardly feasible in the GWAS setting.

Algorithm 2 `get_testable_regions`

```
Input: Dataset $\mathcal{G} = \{\mathbf{g}_i, y_i, c_i\}_{i=1}^n$, desired FWER $\alpha$
Output: Tarone's adjusted significance threshold $\delta_{tar}$ and set of testable genomic regions $\mathcal{R}_T(\delta_{tar})$
 1: $\delta \leftarrow 1$, $\mathcal{R}_T(\delta) \leftarrow \emptyset$
 2: $\mathcal{R}_{cand} \leftarrow \{\llbracket t_s, t_e \rrbracket \,|\, 1 \leq t_s \leq t_e \leq l\}$
 3: for $\llbracket t_s, t_e \rrbracket \in \mathcal{R}_{cand}$ do ▷ Regions in $\mathcal{R}_{cand}$ enumerated firstly in increasing order of length and then starting position
 4:     if $p_{min}(\llbracket t_s, t_e \rrbracket) \leq \delta$ then
 5:         $\mathcal{R}_T(\delta) \leftarrow \mathcal{R}_T(\delta) \cup \{\llbracket t_s, t_e \rrbracket\}$
 6:         while $\delta > \alpha / |\mathcal{R}_T(\delta)|$ do
 7:             Decrease $\delta$
 8:             $\mathcal{P} \leftarrow \{\llbracket t_s, t_e \rrbracket \in \mathcal{R}_T(\delta) \,|\, p_{min}(\llbracket t_s, t_e \rrbracket) > \delta\}$
 9:             $\mathcal{R}_T(\delta) \leftarrow \mathcal{R}_T(\delta) \setminus \mathcal{P}$
10:     if pruning_condition($\llbracket t_s, t_e \rrbracket$) then
11:         Remove all $\llbracket t'_s, t'_e \rrbracket \supset \llbracket t_s, t_e \rrbracket$ from $\mathcal{R}_{cand}$
12: Return $\delta_{tar} \leftarrow \delta$ and $\mathcal{R}_T(\delta_{tar}) = \mathcal{R}_T(\delta)$
```

In Lines 1 and 2, the routine `get_testable_regions` initializes a) the adjusted significance threshold δ to 1, the largest value it could possibly attain, b) the set of testable genomic regions $\mathcal{R}_T(\delta)$ to the empty set and c) the search space of genomic regions $\mathcal{R}_{cand}$ to contain all possible candidate genomic regions, i.e. $\mathcal{R}_{cand} = \{\llbracket t_s, t_e \rrbracket \,|\, 1 \leq t_s \leq t_e \leq l\}$.

After initialization, in Line 3 the algorithm enumerates the genomic regions as described in Algorithm 2. For each genomic region $\llbracket t_s, t_e \rrbracket$ being processed, we perform the steps described below.

Firstly, in Line 4, we compute the minimum attainable p-value for the CMH test, $p_{min}(\llbracket t_s, t_e \rrbracket)$, using the closed-form expression shown in Papaxanthos *et al.* (2016). Then, we check the testability criterion $p_{min}(\llbracket t_s, t_e \rrbracket) \leq \delta$ at the current threshold δ. If it is True, the region is added to the set of testable regions $\mathcal{R}_T(\delta)$ in Line 5 and Tarone's condition $\delta \leq \alpha / |\mathcal{R}_T(\delta)|$ is checked in the following line. If the condition is False, it means that the current significance threshold δ is too large and must be decreased (Line 7). By decreasing δ, some already processed genomic regions which were found to be testable, i.e. $p_{min}(\llbracket t_s, t_e \rrbracket) \leq \delta$ for a larger value of δ, might now become untestable. Those genomic regions are retrieved and removed from $\mathcal{R}_T(\delta)$ in Lines 8 and 9, an operation which can be implemented in $O(1)$ time if an appropriate data structure is used for storing $\mathcal{R}_T(\delta)$ in memory.

The last step in processing a candidate genomic region $\llbracket t_s, t_e \rrbracket$ is the evaluation of a pruning criterion Line 10. If the pruning condition evaluates to True, it becomes possible to discard from the set of candidate regions $\mathcal{R}_{cand}$ all regions that contain the current one. This step can dramatically reduce the size of $\mathcal{R}_{cand}$ and make the method multiple orders of magnitude faster compared to a naive approach. The details about the pruning condition and its efficient computation in $O(k \log k)$, where k is the number of categories of the covariate, can be found in Llinares-Lopez *et al.* (2017).

The routine `get_testable_regions` ends when all candidate regions in $\mathcal{R}_{cand}$ have either been pruned or processed. At that point, the algorithm has converged and we can return δ_{tar} and $\mathcal{R}_T(\delta_{tar})$ as the final values of δ and $\mathcal{R}_T(\delta)$.

4 Simulated experiments and application to GWAS

We first compared the statistical power of `FastCMH` and of burden tests on a set of simulated datasets. Then we applied our method to find significant genomic regions in a Chronic Obstructive Pulmonary Disease GWAS dataset.

4.1 `FastCMH` outperforms burden tests in simulated studies in terms of statistical power

As `FastCMH`, burden tests aim at discovering genomic regions that are associated with a phenotype of interest. However, they control the number of regions to test by predefining them based on biological domain knowledge or by fixing the region size a priori. We performed simulations using burden tests with a genome-wide scan window-based approach. We conducted experiments on *sliding* windows, with a shift of 1 marker, and on *disjoint* windows (Schmid and Yang (2008), Lee *et al.* (2014)). We showed that our method `FastCMH` outperforms the burden tests in terms of statistical power for different window-sizes (see Fig. 1). It can be explained by the inability of the burden tests to test all genomic regions, making then sensitive to misspecification of the size of the associated regions. Additional experimental results and details about the setup can be found in Llinares-Lopez *et al.* (2017).

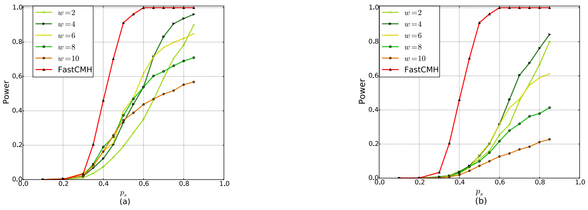

Figure 1: A comparison of the power between `FastCMH` and several burden tests with (a) non-overlapping and (b) sliding windows. The burden tests were performed for various windows sizes (w) and used the encoding that counts all minor alleles in the window.

4.2 `FastCMH` corrects for confounders efficiently in a Pulmonary Disease GWAS dataset

We applied `FastCMH` to several GWAS datasets (Llinares-Lopez *et al.* (2017)), including a Chronic Obstructive Pulmonary Disease (COPD) GWAS dataset from the COPDGene study (Regan *et al.* (2011)). The COPD GWAS dataset contains 615,906 SNPs and 7,993 individuals of which 3,633 are cases and 4,360 are controls. The population is stratified into two categories: African-American and non-Hispanic whites. The SNPs were binarized according to a dominant encoding. For each genomic region, the meta-marker of individual i was set to 1 if at least one SNP in the region was a minor allele. We compared the performance of several algorithms including `FastCMH`, several window-based burden test, `FAIS` (which does not correct for population stratification) and univariate testing. `FastCMH` found a cluster of three genes CHRNA-CHRNA3-CHRNB4 located on Chromosome 15, which is supported by the literature (Cho (2014); Cho *et al.* (2010)). Neither univariate testing nor the burden tests could not retrieve those three genes. `FAIS` retrieves $88,403$ hits and shows a genomic inflation factor of 16.70, which indicates the existence of confounding due to population structure. On the contrary, `FastCMH` finds 3 hits with a genomic inflation factor of 1.05. From this last experiment, we can deduce that `FastCMH` corrects for population structure, resulting in a smaller false-positive rate.

Conclusion

`FastCMH`, presented in this article, is the first algorithm able to detect genetic heterogeneity genome-wide while correcting for covariates. The state-of-the-art pattern mining approach combined with Tarone's concept of testability allows `FastCMH` to scan all genomic regions for association with a phenotype of interest, while being computationally efficient and keeping a high detection performance. `FastCMH` outperforms burden tests and non conditional approaches in several simulation experiments and GWAS dataset analysis.

References

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.

Cho, M. H. (2014). Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *The Lancet Respiratory Medicine*, **2**(3), 214–225.

Cho, M. H., Boutaoui, N., Klanderman, B. J., Sylvia, J. S., Ziniti, J. P., Hersh, C. P., DeMeo, D. L., Hunninghake, G. M., Litonjua, A. A., Sparrow, D., Lange, C., Won, S., Murphy, J. R., Beaty, T. H., Regan, E. A., Make, B. J., Hokanson, J. E., Crapo, J. D., Kong, X., Anderson, W. H., Tal-Singer, R., Lomas, D. A., Bakke, P., Gulsvik, A., Pillai, S. G., and Silverman, E. K. (2010). Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat Genet*, **42**(3), 200–202.

Cochran, W. G. (1954). Some methods for strengthening the common chi2 tests. *Biometrics*, **10**(4), 417–451.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, **85**(1), 87–94.

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, **95**(1), 5–23.

Llinares-López, F., Grimm, D. G., Bodenham, D. A., Gieraths, U., Sugiyama, M., Rowan, B., and Borgwardt, K. (2015a). Genome-wide detection of intervals of genetic heterogeneity associated with complex traits. *Bioinformatics*, **31**(12), i240–i249.

Llinares-Lopez, F., Papaxanthos, L., Bodenham, D., Roqueiro, D., Investigators, C., and Borgwardt, K. (2017). Genome-wide genetic heterogeneity discovery with categorical covariates. *Bioinformatics*.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**(4), 719.

Papaxanthos, L., Llinares-López, F., Bodenham, D. A., and Borgwardt, K. (2016). Finding significant combinations of features in the presence of categorical covariates. In *Advances in Neural Information Processing Systems 2016. Print in process.*

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonable be supposed to have arisen from random sampling. *Philosophical Magazine*, **50**, 157–175.

Regan, E. A., Hokanson, J. E., Murphy, J. R., Make, B., Lynch, D. A., Beaty, T. H., Curran-Everett, D., Silverman, E. K., and Crapo, J. D. (2011). Genetic epidemiology of copd (copdgene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, **7**(1), 32–43.

Schmid, K. and Yang, Z. (2008). The trouble with sliding windows and the selective pressure in brca1. *PLoS One*, **3**(11), e3746.

Tarone, R. E. (1990). A Modified Bonferroni Method for Discrete Data. *Biometrics*, **46**(2), 515.

Wellcome Trust Case Control, C. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**(7145), 661–678.