Published in final edited form as:

Genet Epidemiol. 2018 February ; 42(1): 123–126. doi:10.1002/gepi.22094.

Letter to the Editor: Family-Based Tests for Associating Haplotypes With General Phenotype Data: Application to Asthma Genetics:

Improving the FBAT haplotype algorithm

Julian Hecker^{1,2}, Xin Xu³, F. William Townes¹, Heide Loehlein Fier^{1,2}, Chris Corcoran⁴, Nan Laird¹, and Christoph Lange^{1,5}

¹Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

²Department of Genomic Mathematics, University of Bonn, Bonn, Germany

³Nanfeng Hospital, Southern Medical University, Guangzhou, China

⁴Department of Mathematics and Statistics, Utah State University, Logan, UT, USA

⁵Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA

Abstract

For family-based association studies, Horvath et al. (2004) proposed an algorithm for the association analysis between haplotypes and arbitrary phenotypes when the phase on the haplotypes is unknown, i.e. genotype data is given. Their approach to haplotype analysis maintains the original features of the TDT/FBAT-approach, i.e. complete robustness against genetic confounding and misspecification of the phenotype. The algorithm has been implemented in the FBAT and PBAT software package, and has been used in numerous substantive manuscripts.

Here, we propose a simplification of the original algorithm that maintains the original approach, but reduces the computational burden of the approach substantially and gives valuable insights regarding the conditional distribution. With the modified algorithm, the application to whole-genome sequencing studies becomes feasible, e.g. in sliding window approaches (Beissinger, Rosa, Kaeppler, Gianola, & de Leon, 2015; Sung, Korthauer, Swartz, & Engelman, 2014) or spatial-clustering approaches (Loehlein Fier et al., 2017). The reduction of the computational burden that our modification provides is especially dramatic when both parental genotypes are missing. For example, for 8 variants along 441 nuclear families, with mostly offspring-only families, in a WGS study at the APOE locus, the running time decreased from approximately 21 hours for the original algorithm to 0.11 seconds after our modification.

Keywords

FBAT; candidate region; admixture; whole-genome sequencing

The FBAT-haplotype algorithm by Horvath et al

First, we recall and describe the original FBAT-haplotype algorithm by Horvath et al. (2004). After we derive our proposed modification, we illustrate the differences based on small examples and a running time comparison. We use the same notation as Horvath et al. (2004).

Consider a nuclear family and a region of interest consisting of tightly linked markers. Throughout the following derivations, we denote the offspring genotype information along these markers by g . The observed offspring genotypes are denoted by g^{obs} . In addition, we introduce the mating type m . In m , the paternal, the maternal or both parental genotypes can be missing. The mating type m can be unphased or phased.

We adopt the convention that phased offspring genotypes are denoted by uppercase G and parental mating types, where mother and father are phased, by uppercase M . For example, $(1/2, 2/2)$ denotes an unphased 2 marker genotype along two bi-allelic markers with alleles 1 and 2. The genotype for the first marker consists of both alleles, the genotype for the second marker has twice the allele 2. The phase is unknown. $M = (2, 2)/(2, 2) \times (2, 1)/(2, 2)$ is an example for a phased parental mating type consisting of the haplotypes $(2, 2)$ and $(2, 1)$. The notation will be helpful in the examples below.

The algorithm in Horvath et al. proceeds in five steps to compute the conditional distribution of the vectors of offspring genotypes under the null hypothesis, where the conditioning is on the sufficient statistic for all nuisance parameters in the model.

These five steps are derived by the characterization of the minimal sufficient statistic as described in Rabinowitz and Laird (2000). This concept is rather abstract and we refer to Rabinowitz and Laird (2000) as well as Horvath et al. (2004) for a more comprehensive introduction and detailed example computations.

Original haplotype algorithm

In the first step, all phased mating types that are compatible with g^{obs} and m are constructed. Note that a compatible phased parental mating type can contain haplotypes that are not compatible with g^{obs} . Step 2 identifies the set of offspring genotypes that are consistent with all phased mating types from Step 1. The phase information for each possible offspring genotype in this set is recorded. In addition, the algorithm finds, based on the genotypes in this set, all possible corresponding vectors of offspring genotypes that lead to the same set of phased compatible mating types as g^{obs} . Then, as Step 3, the conditional probability of all these offspring genotype vectors from Step 2, given each phased parental mating type, is computed. In Step 4, the algorithm identifies all offspring genotype vectors that have a constant ratio of conditional probabilities over the parental mating types, compared to g^{obs} . All other genotype vectors are discarded. In the last Step (5), these remaining offspring genotype vectors are used to compute the conditional probability of each (permuted) possible configuration of offspring genotypes. This characterizes the conditional distribution of the vectors of offspring genotypes under the null hypothesis, where the conditioning is on the sufficient statistic for all nuisance parameters in the model.

In addition, the FBAT software estimates haplotype frequencies from the observed data using an EM algorithm. This allows the use of weights based on haplotype frequencies for the test statistic, as described in Horvath et al. (2004).

The computational burden of the haplotype algorithm is mostly attributable to the construction of all compatible phased parental mating types and performing corresponding computational steps within each nuclear family.

In the following, we show that it is sufficient to construct only compatible phased parental mating types within each nuclear family that are completed by haplotypes that are compatible with the observed offspring genotypes g^{obs} . A more detailed definition of these phased mating types can be found below and in the Appendix. Thereby, we reduce substantially the number of mating types that must be evaluated and achieve the savings in computation time, as described above.

Modification of the FBAT-haplotype algorithm

To reduce the set of parental mating types that must be evaluated in the algorithm, we modify the following components.

The derivation of the conditional distribution of the vectors of offspring genotypes

Given the observed offspring genotypes g^{obs} , we can identify the set of all haplotypes h_{off} that are compatible with these offspring genotypes. We propose to replace the construction of the set of all compatible phased parental mating types by the construction of all compatible phased parental mating types that are completed by haplotypes in h_{off} only. If both parental genotypes are missing in m , this means that we use only haplotypes from the set h_{off} . If one parental genotype is observed, we use only haplotypes from the set h_{off} for the missing parental genotype. See the Appendix for more details. This replacement affects the construction of mating types in Step 1 and the constructions of mating types in Step 2. In the appendix (Claim 1–3), we show that indeed all related objects and sets are invariant to this replacement. That means that we obtain the same output from the modified algorithm in terms of the conditional distribution of the vectors of offspring genotypes under the null hypothesis.

The estimation of haplotype frequencies and test statistic weights

From our reduced set of phased parental mating types, as described in the modification of the algorithm, it is straightforward to conclude the set of all compatible phased compatible mating types, as used in the original version of the algorithm. Due to the implementation details of the haplotype frequency estimation and EM-algorithm calculations, it is not necessary to construct this set. It is sufficient to count the number of mating type one would add for each mating type in the reduced set and this number can be computed by combinatorial arguments. But due to the heuristic motivation and the insensitivity of the test statistic weights, our modified implementation in FBAT performs the same computation as described in Horvath et al., but using the reduced set of phased parental mating types. We also plan to include an option to reconstruct the original weights.

Advantage of the modified algorithm in practice

In WGS studies, a large proportion of the variants have a low minor allele frequency. This leads to the observation that, within a nuclear family, a notable number of variants are monomorphic, i.e. only the major allele is observed. This implies that the set h_{off} is much smaller than the set of all possible haplotypes along the markers. Therefore, our simplification leads to a substantial reduction of computational time, i.e. several magnitudes (see Table 1).

When phased haplotype data will be available in WGS studies, the haplotype analysis scenario can be interpreted as the multiallelic single variant FBAT test. In this case, our result shows that we can utilize the standard single variant FBAT approach with a maximum of four possible alleles instead of considering the setting with an exponential sized number of possible alleles/haplotypes.

We updated the haplotype analysis algorithms in the FBAT package. The latest version will be made available on the FBAT homepage. Both packages, FBAT and PBAT, break extended pedigrees into nuclear families for the haplotype analysis. The modified version of PBAT is under development and will also be made available.

We would like to thank the anonymous referees for their helpful comments and suggestions that helped us to improve the manuscript. The project was supported by Cure Alzheimer's fund.

We illustrate the modification and differences to the original algorithm by the following examples.

Example 1

Consider a nuclear family where two bi-allelic markers have been genotyped in three offspring. The genotypes for mother and father are missing. The observed unphased offspring genotypes are given by $g^{obs} = ((2/2, 2/2), (2/2, 2/1), (2/2, 2/1))$. Since both parental genotypes are missing, $m = missing \times missing$, our modification implies that we only must consider phased parental mating types that consist of haplotypes in h_{off} . Given the observed offspring genotypes, these haplotypes are (2,2) and (2,1). The original algorithm by Horvath et al. constructs four compatible phased mating types: $M_1 = (2,2)/(2,2) \times (2,1)/(2,2)$, $M_2 = (2,2)/(2,1) \times (2,2)/(2,1)$, $M_3 = (2,1)/(2,2) \times (2,2)/(1,2)$, $M_4 = (2,2)/(2,1) \times (2,2)/(1,1)$. Our modified algorithm constructs only the two compatible mating types M_1 and M_2 , since M_3 and M_4 contain a haplotype that is not compatible with the offspring genotypes. Performing all steps of the algorithm, both versions of the algorithm compute that the conditional distribution is given by $P_{cond}[2 \times g_1, 1 \times g_2] = 1.0$, where $g_1 = (2/2, 2/1)$ and $g_2 = (2/2, 2/2)$.

If we add additional four markers where all genotypes of the offspring are (2/2, 2/2), there are still only two haplotypes in h_{off} . Our modified algorithm constructs only two compatible phased parental mating types, whereas the number of compatible phased parental mating types constructed by the original algorithm increases to 64. This illustrates the advantage of

our modification and why the reduction of running time can be large in the presence of monomorphic markers in the nuclear family.

Example 2

Consider a nuclear family where two bi-allelic markers have been genotyped in two offspring and the father. The genotype of the mother is missing. The observed unphased offspring genotypes are given by $g^{obs} = ((1/2, 2/2), (2/2, 2/2))$. The genotype of the father is given by $(1/2, 2/2)$. The original algorithm by Horvath et al. constructs four compatible phased mating types: $M_1 = (1,2)/(2,2) \times (2,2)/(2,2)$, $M_2 = (2,2)/(1,2) \times (2,2)/(2,1)$, $M_3 = (1,2)/(2,2) \times (2,2)/(1,2)$, $M_4 = (2,2)/(1,2) \times (2,2)/(1,1)$. Our modified algorithm constructs only the two mating types M_1 and M_3 . Both versions of the algorithm compute that the conditional distribution is given by $P_{cond}[1 \times g_1, 1 \times g_2] = 1.0$, where $g_1 = (2/2, 2/2)$, and $g_2 = (1/2, 2/2)$.

Running time

To demonstrate the reduction of running time, we performed a comparison between the original algorithm and the modified version for different sets of variants in the APOE locus in a real data set. This data set consists of 441 nuclear families. 421 nuclear families have between two and five offspring but no parental genotype information. For 18 nuclear families, one parental genotype is observed. We chose 13 different set of consecutive variants, where the number of variants ranged between 5 and 8. In Table 1, we provide the corresponding results.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Beissinger TM, Rosa GJ, Kaeppler SM, Gianola D, de Leon N. Defining window-boundaries for genomic analyses using smoothing spline techniques. *Genetics, Selection, Evolution : GSE*. 2015; 47(1) <https://doi.org/10.1186/s12711-015-0105-9>.
- Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM. Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genetic Epidemiology*. 2004; 26(1):61–69. <https://doi.org/10.1002/gepi.10295>. [PubMed: 14691957]
- Loehlein Fier H, Prokopenko D, Hecker J, Cho MH, Silverman EK, Weiss ST, Lange C. On the association analysis of genome-sequencing data: A spatial clustering approach for partitioning the entire genome into nonoverlapping windows. *Genetic Epidemiology*. 2017; 41(4):332–340. <https://doi.org/10.1002/gepi.22040>. [PubMed: 28318110]
- Rabinowitz D, Laird N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human Heredity*. 2000; 50(4):211–223. <https://doi.org/22918>. [PubMed: 10782012]
- Sung YJ, Korthauer KD, Swartz MD, Engelman CD. Methods for Collapsing Multiple Rare Variants in Whole-Genome Sequence Data. *Genetic Epidemiology*. 2014; 38(01):S13–S20. <https://doi.org/10.1002/gepi.21820>. [PubMed: 25112183]

Table 1

Comparison of running time between the original and the modified version of the algorithm.

set	Number of variants	Original version	Modified version
1	5	8.18sec	0.04sec
2	5	9.89sec	0.05sec
3	5	5.2sec	0.02sec
4	5	5.63sec	0.02sec
5	6	230.76sec	0.04sec
6	6	191.15sec	0.14sec
7	6	182.32sec	0.04sec
8	6	94.07sec	0.02sec
9	7	43 min	0.06sec
10	7	28 min	2.3sec
11	7	27 min	0.04sec
12	7	14 min	0.02sec
13	8	~21 hrs	0.11sec