

# HaploBuild: A Program for Constructing Haplotypes from High Density SNP Genotype Data

Jason M Laramie,<sup>1</sup> Jemma B Wilk,<sup>2</sup> Richard H Myers<sup>3</sup>

**Keywords:** haplotype, single nucleotide polymorphism, candidate region, association

## 1 Introduction

Haplotypes, closely linked alleles on a chromosome that are inherited as a unit, play key roles in deciphering the genetic basis of complex disease. Haplotypes provide information on ancestral chromosome segments that may harbor alleles that influence disease phenotypes. Here we present an algorithm called *HaploBuild* that, in conjunction with the program haplotype FBAT (Family Based Association Tests) [1], can construct nonconsecutive haplotypes of any length that are statistically associated with a trait from a set of high density single nucleotide polymorphisms (SNPs) or microsatellite genetic markers.

## 2 Algorithm and Results

*HaploBuild* finds haplotypes that are strongly associated with a trait using a three step approach. (1) Given a set of genetic markers, all two marker haplotypes within a specified physical distance are tested for association with a trait using haplotype FBAT. Haplotypes showing statistically significant association ( $Z$ -statistic derived  $p \leq .05$ , not adjusted for multiple comparisons) are saved for further analysis. (2) Each two marker haplotype found in step 1 is used to create the root node of a tree and each genetic marker within a specified distance is added one at a time to create a three marker haplotype that is tested for trait association using haplotype FBAT. The three marker haplotypes with haplo-specific  $p$ values lower than their parent node (in this case the root node) are added to the tree. Each round of recursion adds child nodes to the tree increasing the number of markers in a haplotype as long as the  $p$ value is decreased. The completed tree is an directed acyclic graph with nodes containing specific information about the haplotype that generated the node (i.e genetic marker names,  $Z$ -statistic score, etc) and each edge is labeled with the haplo-specific  $p$ value for that node. Therefore, leaf nodes represent haplotypes that contain the maximum number of genetic markers within a specified physical distance that improved trait association (decrease in  $p$ value) and a traversal of the tree from the root to a leaf node shows the construction of a haplotype. (3) After all trees are constructed an empirical  $p$ value is calculated for each haplotype contained in every leaf node. Empirical  $p$ values are calculated by permutating the trait phenotype and then re-testing the association between the haplotype and the permuted phenotype.

*HaploBuild* was used to find statistically significant haplotypes between the phenotype body mass index (BMI) and 458 SNPs and microsatellites densely genotyped on chromosome 7q31-34 in 91 families comprising 742 individuals. To ensure minimal recombination occurred between markers within a haplotype, any two markers within a haplotype had to be  $\leq 50$ kb

---

<sup>1</sup>Department of Bioinformatics, Boston University, Boston MA. E-mail: [laramiej@bu.edu](mailto:laramiej@bu.edu)

<sup>2</sup>Departments of Medicine and Neurology, Boston University School of Medicine, Boston MA. E-mail: [jwilk@bu.edu](mailto:jwilk@bu.edu)

<sup>3</sup>Departments of Medicine and Neurology, Boston University School of Medicine, Boston MA. E-mail: [rmyers@bu.edu](mailto:rmyers@bu.edu)

apart from each other. This restriction resulted in 168 trees being built resulting in 990 haplotypes containing three to nine markers. To test the statistical significance of these 990 haplotypes, 10,000 permutations were conducted to compute an empirical pvalue. Of the 990 haplotypes 824 remained significant with an empirical pvalue  $\leq .05$ . Figure 1 plots the  $-\log(\text{empirical pvalue})$  versus the region 7q31-34. From this figure it is interesting to note the haplotype clusters that result when trees converge to the same haplotypes regardless of the two marker haplotype that was used to initiate the tree. Furthermore, there is a definitive group of statistically significant haplotypes (arrow Figure 1) that were not seen in the linkage [2].

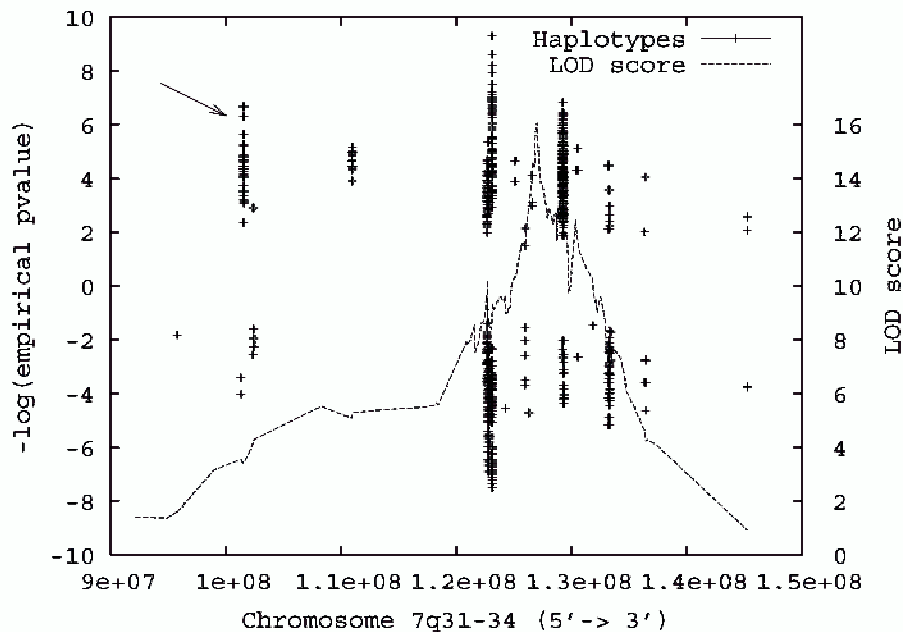


Figure 1: Haplotypes (pluses) are plotted as the  $-\log(\text{empirical pvalue})$  versus their position in the chromosome 7q31-34 region. The sign of the Z-statistic was added to each haplotypes  $-\log(\text{empirical pvalue})$ . Therefore a positive (negative)  $-\log(\text{empirical pvalue})$  is indicative of a high-risk (protective) haplotype. LOD scores using the same markers is plotted as a dashed line.

### 3 References and bibliography.

#### References

- [1] Horvath S, Xu X, Lake S, Silverman E, Weidd S, and Laird N (2004) Tests for Associating Haplotypes with General Phenotype Data: Application to Asthma Genetics. *Genetic Epidemiology* 26:61-69.
- [2] Wilk JB, Jiang Y, Williamson S, Prakash R, DeStefano AL, Ellison RC, Borecki IB, Province MA, Myers RH. Linkage disequilibrium mapper for body mass index on chromosome 7q31-7q34 implicates multiple genes and sex-specific effects: The NHLBI Family Heart Study. *American Society of Human Genetics* October 26-30, 2004 Toronto, Ontario, Canada.