Use Hidden Markov Model to Infer Haplotype Structure from Mouse SNPs

Jin P. Szatkiewicz, Glen L. Beane, Gary A. Churchill

The Jackson Laboratory , Bar Harbor, Maine 04609 USA

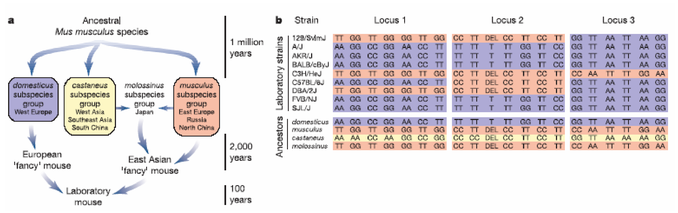
ABSTRACT

Recent studies of sequence variation using whole genome SNP (single nucleotide polymorphism) data suggest that genetic variation is organized in haplotype blocks. In particular, it has been suggested that the mouse genome has a mosaic structure of polymorphisms attributable to recent descent from a limited number of genetically diverse founders. Understanding the haplotype structure of the laboratory mouse will accelerate the identification of genes associated with complex phenotypes. However, missing data, genotyping errors, mutations and the absence of clear haplotype structure in some genomic regions present analytic challenges that must be addressed. We have developed a Hidden Markov Model and a software tool that assigns individual strains to local haplotypes and imputes missing SNP alleles. An ad-hoc state-trimming approach improves the performance of the model, resulting in a more concise and interpretable haplotype reconstruction. We illustrate the method and its applications using a publicly available dataset of 130,000 SNPs collected on 42 inbred mouse strains.

BACKGROUND

The genome of classical inbred strains derives from a handful of progenitors (Ferris *et al.* 1982 *Nature*).

The genome of classical inbred strains represents a mosaic with unequal contributions of several *Mus musculus* subspecies (Wade *et al.* 2002 *Nature*).

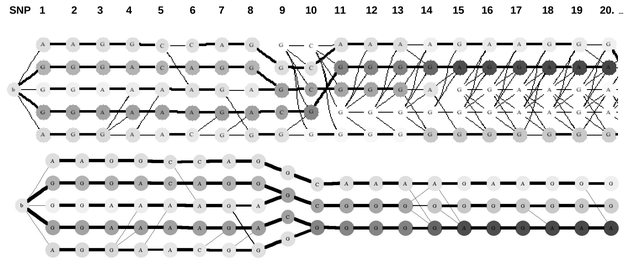


SUMMARY OF METHOD

- A left-right Hidden Markov Model
  - State: unobserved haplotype assignment
  - Output: observed SNP genotypes
- Parameter estimation
  - Expectation-Maximization method in an incomplete data context
  - Maximum Likelihood training, Dirichlet Priors
  - Hidden state reconstructed by the Smoothing algorithm (Churchill 1989)
- Haplotype state path is decoded by the Viterbi algorithm (Viterbi 1967).
- Structural learning
  - The number of states varies along the mouse genome
  - Initialize a “full” model and iteratively remove excess states
- Software
  - Algorithm is implemented in C language, fast & user friendly
  - A JAVA tool for viewing high density SNPs annotated by haplotype states

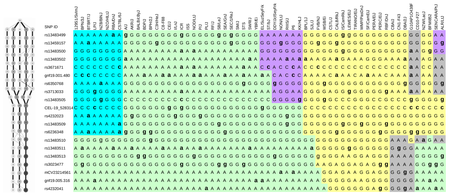
A TRAINED HMM SEGMENT

Figure 1: Graphical representation of a trained HMM segment. Each node represents a haplotype state and is labeled by the name of the nucleotide that is more likely to be emitted. Node probabilities are color coded according to the marginal probability with darker color indicate more frequent haplotype. The edges represent state transitions with edge thickness proportional to the probability of transition.



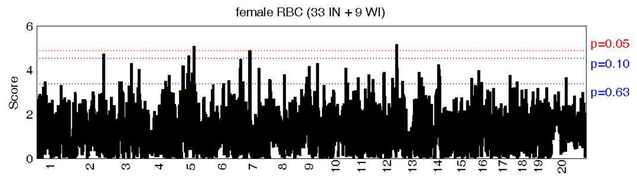
APPLICATION: Imputing missing genotypes

Figure 2: Illustration of missing data imputation. The original dataset consists of 67 inbred strains including both classical and wild-derived and a total of 270 SNPs. To evaluate the HMM method, 10% missing spots were generated randomly within the original data. The haplotype assignments are labeled by distinct colors. The imputed SNPs are indicated by lowercase letters. Approximately 90% “missing” SNPs were accurately recovered.



APPLICATION: Haplotype association mapping

Figure 3: An association mapping scan score plot. A total of 42 inbred female mouse strains were typed at ~130,000 SNP markers and their red blood cell counts were measured as a quantitative trait. A genome wide scan was conducted to assess the association of phenotype and the inferred haplotype using HMM. Significance threshold was established by permutation. Considerably fewer false positives were observed with the highest peak identified on Chromosome 12.



APPLICATION: Genome of RILs

Figure 4: A recombinant inbred line (RIL) is formed by crossing two inbred strains followed by repeated sibling matting to create a panel of new inbred lines whose genomes are a mosaic of the two parental strains. RILs are powerful for genetic mapping (Broman 2005).

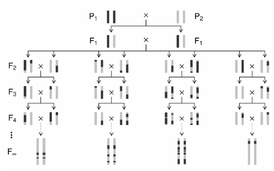
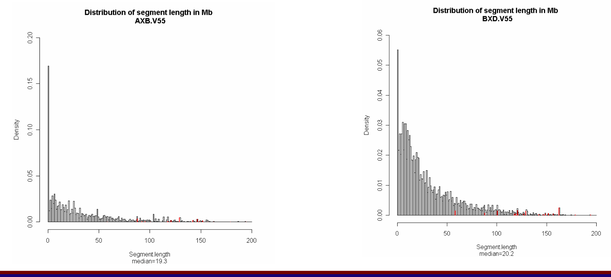


Figure 5: A specialized HMM is developed to study the genome of RILs. The AXB panel includes 20 AXB RI strains and the BXD panel includes 89 BXD RI strains. The distribution of total number of recombination segments based on 5,600 SNP markers are displayed, where each bin represents 1Mb and red lines indicates the chromosomes inherited intact. The results closely resemble the theoretical prediction by Broman (2005).



DISCUSSION

HMM offers an efficient tool for haplotype inference using high density SNPs. State trimming facilitates structural learning of HMM in particular for mouse genome.

Method described here is implemented in a fast and user friendly package.

Understanding haplotype structure will accelerate the identification of genes associated with complex phenotypes.

Future prospects:

- Improve ML parameter estimation and state trimming approach
- Investigate more advanced missing data algorithm
- Extend model to other type of polymorphism
- Explore other applications of HMM outputs

REFERENCES

Churchill GA (1989) Stochastic models for heterogeneous DNA sequences. Bull Math Biol. 51(1):79-94.

Kimmel G, Shamir R. (2005) A block-free hidden Markov model for genotypes and its application to disease association. Comput Biol. 2005 12(10):1243-60.A

Scheet P, Stephens M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 78(4):629-44.

Broman KW (2005) The genomes of recombinant inbred lines. Genetics. 169(2):1133-46

ACKNOLEDGEMENTS

The authors acknowledge support from NIH/NIGMS grant GM076468, The Center for Genome Dynamics, and The Department of Computational Sciences of The Jackson Laboratory; Shirng-Wern Tsaih, Sc.D., for proving genome scan plot.