Sampling ancestries at a genomic location conditional on data from surrounding genetic markers

K.M. Burkett, J. Graham, and B. McNeney

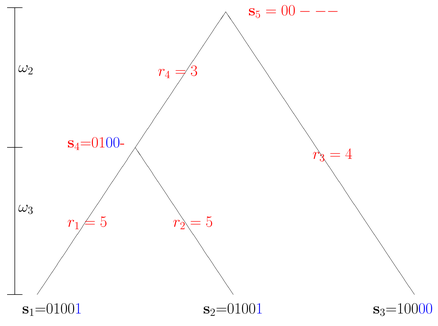
Statistics and Actuarial Science, Simon Fraser University, Burnaby Canada, kburkett@sfu.ca, http://statgen.stat.sfu.ca

Introduction

- We propose that an association between genetic variability and disease outcomes reflects the latent genetic ancestries giving rise to the sample's genetic variability.
  - These ancestries contain information about which sequences carry disease-predisposing variants, but they are typically not known.
- However, the observed genetic marker data $\mathbf{G}$ gives us information about the unknown ancestry $\mathbf{T}$.
- We sample ancestries from $\Pr(\mathbf{T}|\mathbf{G})$, which we know only up to a normalizing constant, to gain insight about association from their structure.
- This poster summarizes
  - our implementation of a Markov Chain Monte Carlo (MCMC) sampler, outlined in [1], that samples ancestries compatible with sequence data from unrelated individuals.
  - an extension of the sampler to unphased genotype data
  - an application of the sampler to data from an association study

The MCMC sampler for sequence data

Figure 1
Notation defining the ancestry



- Let $\mathbf{H} = (\mathbf{s}_1, \ldots \mathbf{s}_n)$ be sequence data available at present. Sample ancestry $\mathbf{T}$ from $\Pr(\mathbf{T}|\mathbf{H})$:
  - Each sampled genealogy $\mathbf{T}$ is for a specified focal point along the sequence.
  - $\mathbf{T}$ consists of the labelled topology of the tree τ and the coalescence times $\mathbf{\Omega} = (\omega_2, \omega_3 \ldots \omega_{2n-1})$ of internal nodes
  - Define latent variables $\mathbf{S} = (\mathbf{s}_{n+1}, \ldots \mathbf{s}_{2n-1})$ and $\mathbf{R} = (r_1, \ldots, r_{2n-2})$ for the sequences and location of recombination events along the sequence at internal nodes, and mutation and recombination rates θ and ρ.
  - Let $\mathbf{A} = (\mathbf{\Omega}, \tau, \theta, \rho, \mathbf{R}, \mathbf{S})$ be the augmented data.
- To sample $\mathbf{T}$ we actually sample the augmented data from $\Pr(\mathbf{A}|\mathbf{H})$ using an MCMC approach.
  - At the t^{th} step, $\tilde{\mathbf{A}}$ is proposed from distribution $Q(\mathbf{A}|\mathbf{A}^{t-1})$. It is accepted as $\mathbf{A}^t$ according to the Metropolis-Hasting probability $\alpha(\mathbf{A}^{t-1}, \tilde{\mathbf{A}})$, which is a function of $\Pr(\mathbf{A}|\mathbf{H})$ and Q [2].
  - $\Pr(\mathbf{A}|\mathbf{H})$ computed using standard population genetic theory.
  - Proposals from different distributions $Q_i(\mathbf{A}|\mathbf{A}^{(t)})$ modifying subsets i of $\mathbf{A}$ that include the sequence, recombination break-points, branch lengths of internal nodes and the topology of the tree.

Extension to unphased genotype data

- Sequences within an individual are now a latent variable. Introduce a proposal distribution to randomly update them. Updates will induce a topology change and change of associated recombination, branch length and sequence variables.
- Sequence that recombines in to the tree around the focal point is assumed to be drawn from a specified haplotype frequency distribution, which is estimated from the data.

Validation with real data

- In type 1 diabetes patients, IA-2 autoantibodies are associated with the 3'-UTR of the GIMAP5 gene [3]. Do the ancestries give any insight into this association?
- Haplotype data [4] from 56 SNPs in a 469Kb region spanning GIMAP5 for 125 patients with very-high IA-2 antibody status (vh+) and 125 patients without (vh-).
- Two subsets of the 56 SNPs were chosen, with one focal point in each subset
  - **(A)** First 15 SNPs; not near association signal. Focal point between 5th and 6th SNP
  - **(B)** Last 15 SNPs; includes most significantly associated SNP. Focal point beside most significant SNP.

Figure 2
No Association (A)

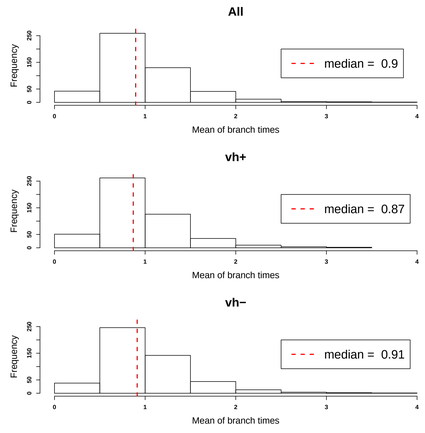
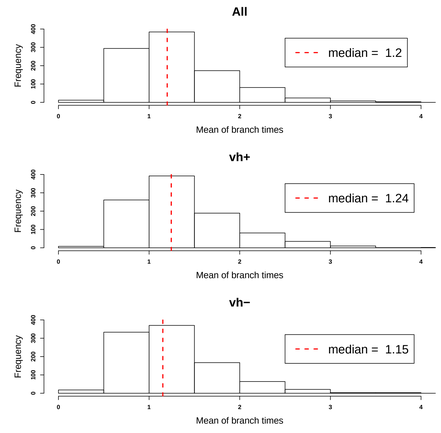


Figure 3
Association Region (B)



- Histograms in figures (2) and (3) display the distribution across 1000 sampled ancestries of the averages of the pairwise branch times between all pairs of 500 sequences and subsets of the 500.
  - Three histograms are for averages of all pairs of 500 (top), averages of all pairs of 250 sequences with very high IA2A antibody status (middle) and averages of all pairs of 250 sequences without very high IA2A antibody status (bottom).
- In region (A) (figure 2), the mean and median are slightly lower for those with very high IA2A status however histograms appear to be quite similar.
- In region (B) (figure 3), the distribution of averages of the branch times for those without very high IA2A status is shifted to the left relative to those with very high IA2A status. Both mean and median are lower for those without very high IA2A status, indicating lower pairwise branch times
- Future work includes determining significance of such differences in distribution.

References

[1] Zöllner S. and Pritchard J.K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. Genetics 169:1071-1092.

[2] Hastings W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97-109.

[3] Shin J-H, Janer M., McNeney B., Blay S., Deutsch K., Sanjeevi C.B., Kockum I., Lernmark ÅA. and Graham J. on behalf of the Swedish Childhood Diabetes Study Group and the Diabetes Incidence in Sweden Study Group (2007). IA-2 autoantibodies in incident type I diabetes patients are associated with a polyadenylation signal polymorphism in GIMAP5. Genes and Immunity, 8:503-512.

[4] Scheet P. and Stephens M. (2006). A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. American Journal of Human Genetics, 78:629-644.

This research was supported by CIHR Doctoral Research Award (KB) and operating grants (BM, JG); MSFHR Senior Graduate Trainee (KB) and Scholar (JG) Awards; and by NSERC and MITACS.