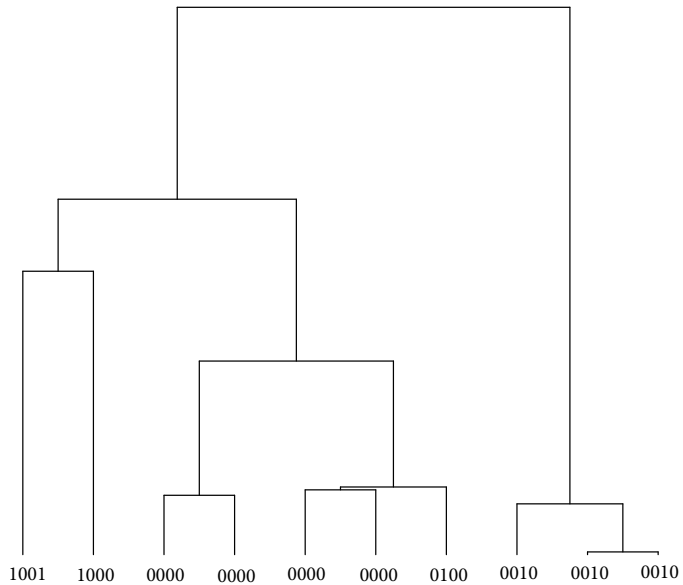
Sampling gene genealogies conditional on genotype data from trios

Kelly Burkett

August, 2018

Gene genealogy or Ancestral tree

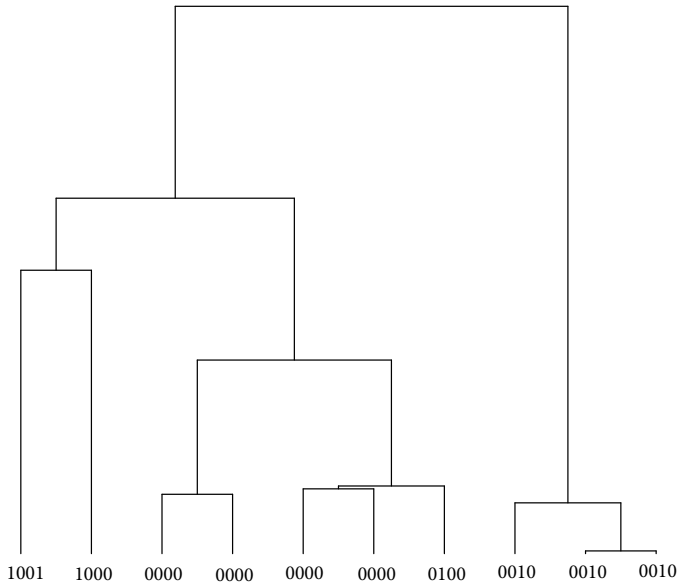
- Tree describing the relationships among sequences sampled from unrelated individuals in a population
  - Tips of the tree correspond to the observed sequences
  - Root of the tree is common ancestor of all sequences



- Due to recombination, ancestral tree differs along a genomic sequence
  - Restrict attention to ancestry of a single focal point
- True tree can not be known

Gene genealogy or Ancestral tree

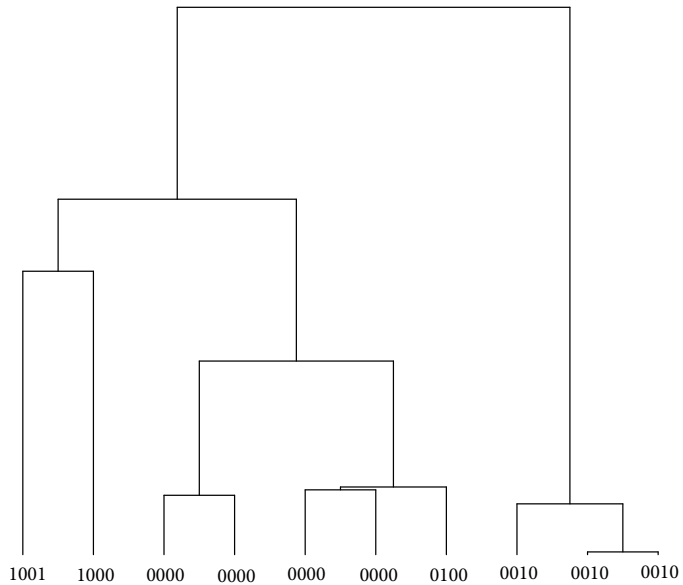
- Tree describing the relationships among sequences sampled from unrelated individuals in a population
  - Tips of the tree correspond to the observed sequences
  - Root of the tree is common ancestor of all sequences



- Due to recombination, ancestral tree differs along a genomic sequence
  - Restrict attention to ancestry of a single focal point
- True tree can not be known

Gene genealogy or Ancestral tree

- Tree describing the relationships among sequences sampled from unrelated individuals in a population
  - Tips of the tree correspond to the observed sequences
  - Root of the tree is common ancestor of all sequences

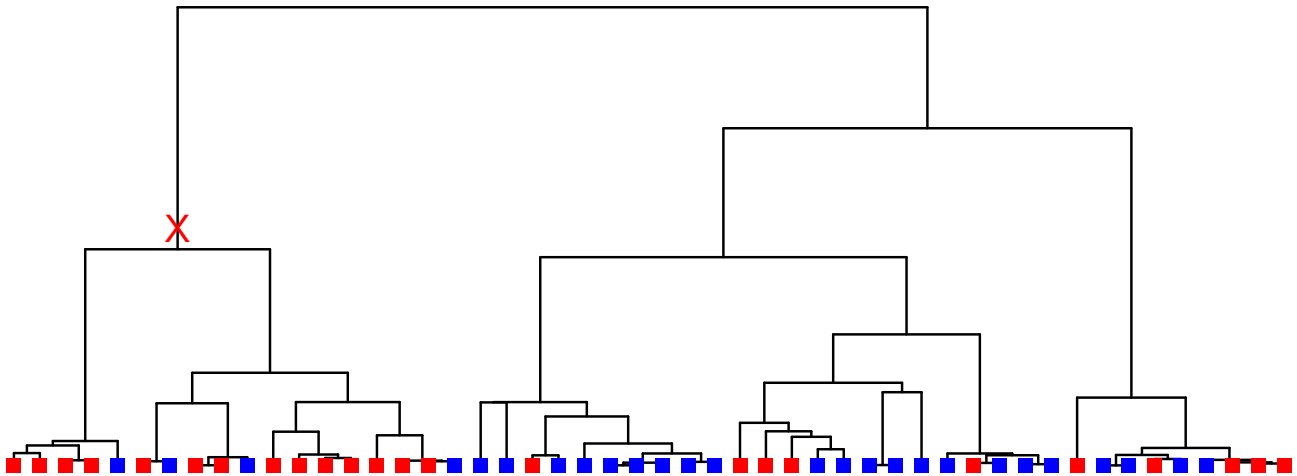


- Due to recombination, ancestral tree differs along a genomic sequence
  - Restrict attention to ancestry of a single focal point

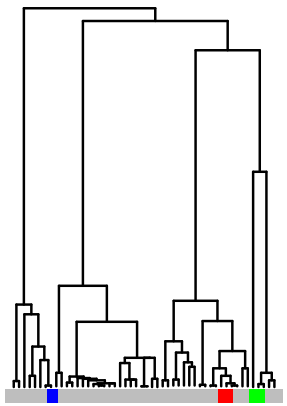
- True tree can not be known

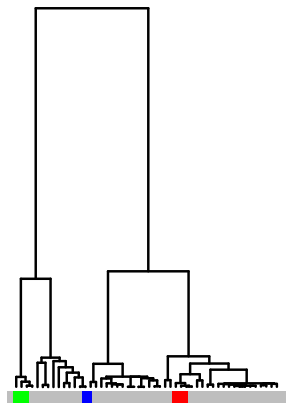
Disease mutations on the ancestral tree

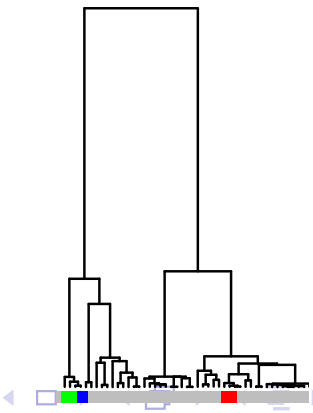
Common variant:



Rare variants:

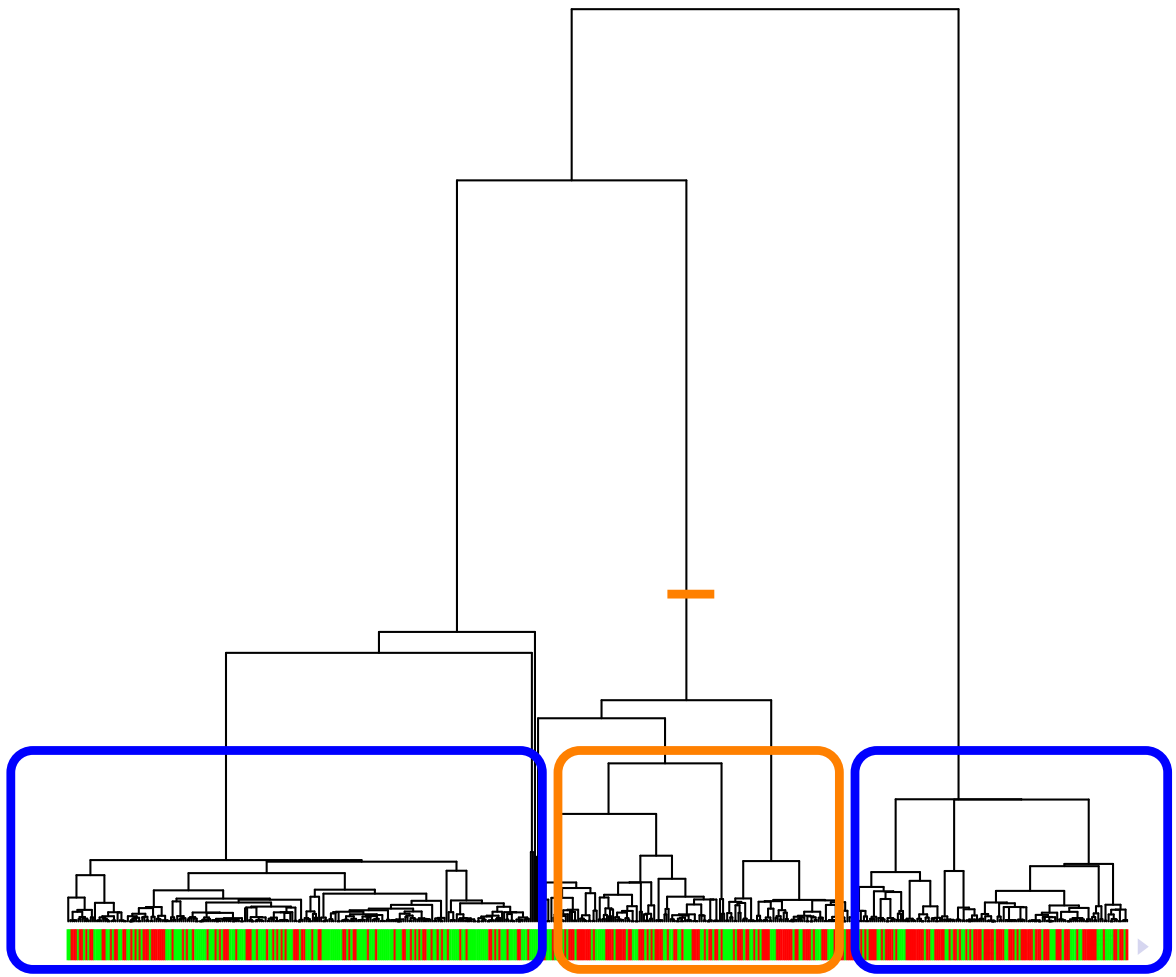






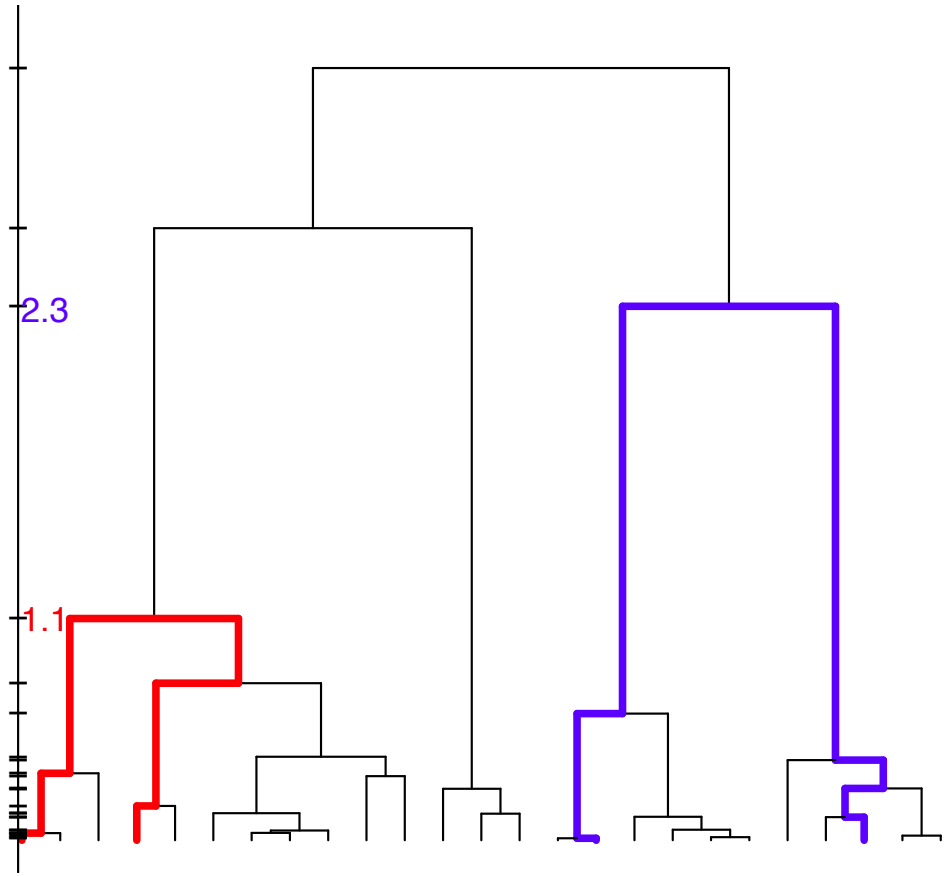
Bipartition clustering (e.g. Minichiello and Durbin, 2006)

- Each internal branch partitions tips into two groups
- Tree defines a factor with two levels



Using the tree as a measure of similarity/distance

$$d_{ij}(T) = f(\text{time to first common ancestor})$$



-Correlate tree distance with distance in phenotype using Mantel test (see Burkett KM, McNeney B, Graham J, Greenwood CMT, 2014.)

Handling uncertainty of the ancestral tree

- True gene genealogy not known; genetic data contains information about the unknown tree.

- Use statistical/phylogenetic methods to reconstruct a tree
  - Hierarchical clustering, minimal trees, ...
  - Reconstructing a single tree doesn't account for tree uncertainty.

- Alternatively, sample multiple trees from distribution that depends on observed data
  - Incorporate population genetic models
  - Use MCMC for sampling

Handling uncertainty of the ancestral tree

- True gene genealogy not known; genetic data contains information about the unknown tree.
- Use statistical/phylogenetic methods to reconstruct a tree
  - Hierarchical clustering, minimal trees, ...
  - Reconstructing a single tree doesn't account for tree uncertainty.
- Alternatively, sample multiple trees from distribution that depends on observed data
  - Incorporate population genetic models
  - Use MCMC for sampling

Handling uncertainty of the ancestral tree

- True gene genealogy not known; genetic data contains information about the unknown tree.
- Use statistical/phylogenetic methods to reconstruct a tree
  - Hierarchical clustering, minimal trees, ...
  - Reconstructing a single tree doesn't account for tree uncertainty.
- Alternatively, sample multiple trees from distribution that depends on observed data
  - Incorporate population genetic models
  - Use MCMC for sampling

Sampling trees conditional on haplotype data

- **T** is the tree topology and node times; **H** vector of observed haplotypes:

Haplotype	Count
10000110000101000100010001001	15
10000110000101000100010000000	3
10000110010101011100011000000	1
⋮	⋮
00011000100010100011100110110	3
00111000100010100011100100000	4

- `sampletrees` (Burkett, McNeney, Graham 2013a,b and 2016) builds on an approach outlined in Zöllner and Pritchard (2005) to sample from $f(\mathbf{T}|\mathbf{H})$
  - Include latent variables for states of internal nodes: recombination variables, **R**, and sequence at internal nodes, **S**.
  - Sample from the augmented distribution, $f(\mathbf{T}, \mathbf{R}, \mathbf{S}|\mathbf{H})$; interested mainly in **T**.

Sampling trees conditional on haplotype data

- **T** is the tree topology and node times; **H** vector of observed haplotypes:

Haplotype	Count
10000110000101000100010001001	15
10000110000101000100010000000	3
10000110010101011100011000000	1
⋮	⋮
00011000100010100011100110110	3
00111000100010100011100100000	4

- `sampletrees` (Burkett, McNeney, Graham 2013a,b and 2016) builds on an approach outlined in Zöllner and Pritchard (2005) to sample from $f(\mathbf{T}|\mathbf{H})$
  - Include latent variables for states of internal nodes: recombination variables, **R**, and sequence at internal nodes, **S**.
  - Sample from the augmented distribution, $f(\mathbf{T}, \mathbf{R}, \mathbf{S}|\mathbf{H})$; interested mainly in **T**.

`sampletrees`: Sampling conditional on haplotypes

- $f(\mathbf{T}, \mathbf{R}, \mathbf{S}|\mathbf{H}) \propto f(\mathbf{H}, \mathbf{S}|\mathbf{R}, \mathbf{T}, \theta) f(\mathbf{R}|\mathbf{T}, \rho) f(\mathbf{T}) f(\theta) f(\rho)$ modelled using population genetic models:
  - $f(\mathbf{T})$: coalescent model
  - $f(\mathbf{H}, \mathbf{S}|\mathbf{R}, \mathbf{T}, \theta)$ and $f(\mathbf{R}|\mathbf{T}, \rho)$: Mutation/recombination events on the branches of the tree assumed Poisson distributed with rates θ and ρ

- MCMC sampling requires proposal distributions to update components of $\mathbf{T}, \mathbf{R}, \mathbf{S}$.
  - Five proposal distributions to update state of internal nodes, topology, rate parameters.
  - At a step, a random choice is made to determine which proposal distribution is used.
  - Update is either accepted or rejected with probability determined by Metropolis-Hastings acceptance ratio.

`sampletrees`: Sampling conditional on haplotypes

- $f(\mathbf{T}, \mathbf{R}, \mathbf{S}|\mathbf{H}) \propto f(\mathbf{H}, \mathbf{S}|\mathbf{R}, \mathbf{T}, \theta) f(\mathbf{R}|\mathbf{T}, \rho) f(\mathbf{T}) f(\theta) f(\rho)$ modelled using population genetic models:
  - $f(\mathbf{T})$: coalescent model
  - $f(\mathbf{H}, \mathbf{S}|\mathbf{R}, \mathbf{T}, \theta)$ and $f(\mathbf{R}|\mathbf{T}, \rho)$: Mutation/recombination events on the branches of the tree assumed Poisson distributed with rates θ and ρ

- MCMC sampling requires proposal distributions to update components of $\mathbf{T}, \mathbf{R}, \mathbf{S}$.
  - Five proposal distributions to update state of internal nodes, topology, rate parameters.
  - At a step, a random choice is made to determine which proposal distribution is used.
  - Update is either accepted or rejected with probability determined by Metropolis-Hastings acceptance ratio.

Example: Applying sampler to 'haplotype' data

Data: Trios with Crohn's disease; 103 markers in 5q31 (Rioux et al. 2001; gap R package)

Procedure:

1. Impute haplotype data based on family relationships and genotypes (Beagle; Browning and Browning, 2009)
2. For each of $K = 100$ focal points, sample M trees $\{\mathbf{T}_{k,1}, \ldots \mathbf{T}_{k,M}\}$ from $f(\mathbf{T}_k|\mathbf{H})$ using `sampletrees`
3. Compute a tree-based statistic, $S(\mathbf{T}, D)$, on each tree to get $S(\mathbf{T}_{k,1}, D), \ldots, S(\mathbf{T}_{k,M}, D)$
   - D - Haplotype is transmitted or not.
   - $S(\cdot)$ - measures whether transmitted haplotypes are more closely related. E.g correlation between transmission status and clusters defined on the tree.
4. For each focal point, summarize the distribution of $\mathbf{S}_k$

Example: Applying sampler to 'haplotype' data

Data: Trios with Crohn's disease; 103 markers in 5q31 (Rioux et al. 2001; gap R package)

Procedure:

1. Impute haplotype data based on family relationships and genotypes (Beagle; Browning and Browning, 2009)
2. For each of $K = 100$ focal points, sample M trees $\{\mathbf{T}_{k,1}, \ldots \mathbf{T}_{k,M}\}$ from $f(\mathbf{T}_k|\mathbf{H})$ using `sampletrees`
3. Compute a tree-based statistic, $S(\mathbf{T}, D)$, on each tree to get $S(\mathbf{T}_{k,1}, D), \ldots, S(\mathbf{T}_{k,M}, D)$
   - D - Haplotype is transmitted or not.
   - $S(\cdot)$ - measures whether transmitted haplotypes are more closely related. E.g correlation between transmission status and clusters defined on the tree.
4. For each focal point, summarize the distribution of $\mathbf{S}_k$

Example: Applying sampler to 'haplotype' data

Data: Trios with Crohn's disease; 103 markers in 5q31 (Rioux et al. 2001; gap R package)

Procedure:

1. Impute haplotype data based on family relationships and genotypes (Beagle; Browning and Browning, 2009)
2. For each of $K = 100$ focal points, sample M trees $\{\mathbf{T}_{k,1}, \ldots \mathbf{T}_{k,M}\}$ from $f(\mathbf{T}_k|\mathbf{H})$ using `sampletrees`
3. Compute a tree-based statistic, $S(\mathbf{T}, D)$, on each tree to get $S(\mathbf{T}_{k,1}, D), \ldots, S(\mathbf{T}_{k,M}, D)$
   - D - Haplotype is transmitted or not.
   - $S(\cdot)$ - measures whether transmitted haplotypes are more closely related. E.g correlation between transmission status and clusters defined on the tree.
4. For each focal point, summarize the distribution of $\mathbf{S}_k$

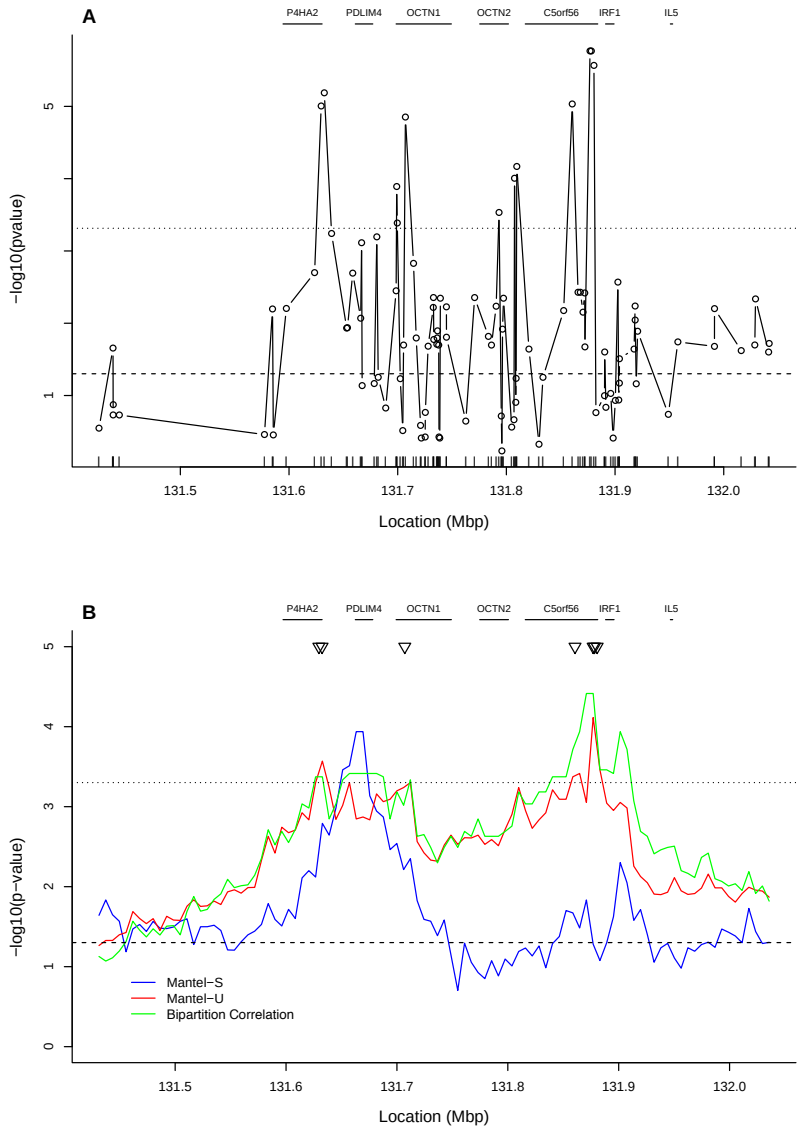
Example: Applying sampler to 'haplotype' data

Data: Trios with Crohn's disease; 103 markers in 5q31 (Rioux et al. 2001; gap R package)

Procedure:

1. Impute haplotype data based on family relationships and genotypes (Beagle; Browning and Browning, 2009)
2. For each of $K = 100$ focal points, sample M trees $\{\mathbf{T}_{k,1}, \ldots \mathbf{T}_{k,M}\}$ from $f(\mathbf{T}_k|\mathbf{H})$ using `sampletrees`
3. Compute a tree-based statistic, $S(\mathbf{T}, D)$, on each tree to get $S(\mathbf{T}_{k,1}, D), \ldots, S(\mathbf{T}_{k,M}, D)$
   - D - Haplotype is transmitted or not.
   - $S(\cdot)$ - measures whether transmitted haplotypes are more closely related. E.g correlation between transmission status and clusters defined on the tree.
4. For each focal point, summarize the distribution of $\mathbf{S}_k$

Example: Applying sampler to ‘haplotype’ data



(green - Burkett KM, Greenwood CMT, McNeney B, Graham J, 2013c).

Sampling trees conditional on trio data (joint work with Marie-Hélène Roy-Gagnon)

Motivation:

1. With trio data, child's genotypes provide information about the phase of the parental haplotypes
   - In example shown, we imputed haplotype phase and treated it as known when sampling genealogies.
   - Conditioning on the child's genotypes would be a better use of the data.

2. Does inclusion of additional phase information from other sources improve sampling when phase is not known?
   - Sampling conditional on unphased genotypes performs poorly.
   - The child's genotype limits the parental phase configurations
   - In theory, could improve sampling when phase is not known.

Sampling trees conditional on trio data (joint work with Marie-Hélène Roy-Gagnon)

Motivation:

1. With trio data, child's genotypes provide information about the phase of the parental haplotypes
   - In example shown, we imputed haplotype phase and treated it as known when sampling genealogies.
   - Conditioning on the child's genotypes would be a better use of the data.

2. Does inclusion of additional phase information from other sources improve sampling when phase is not known?
   - Sampling conditional on unphased genotypes performs poorly.
   - The child's genotype limits the parental phase configurations
   - In theory, could improve sampling when phase is not known.

Gene genealogy of parental sequences

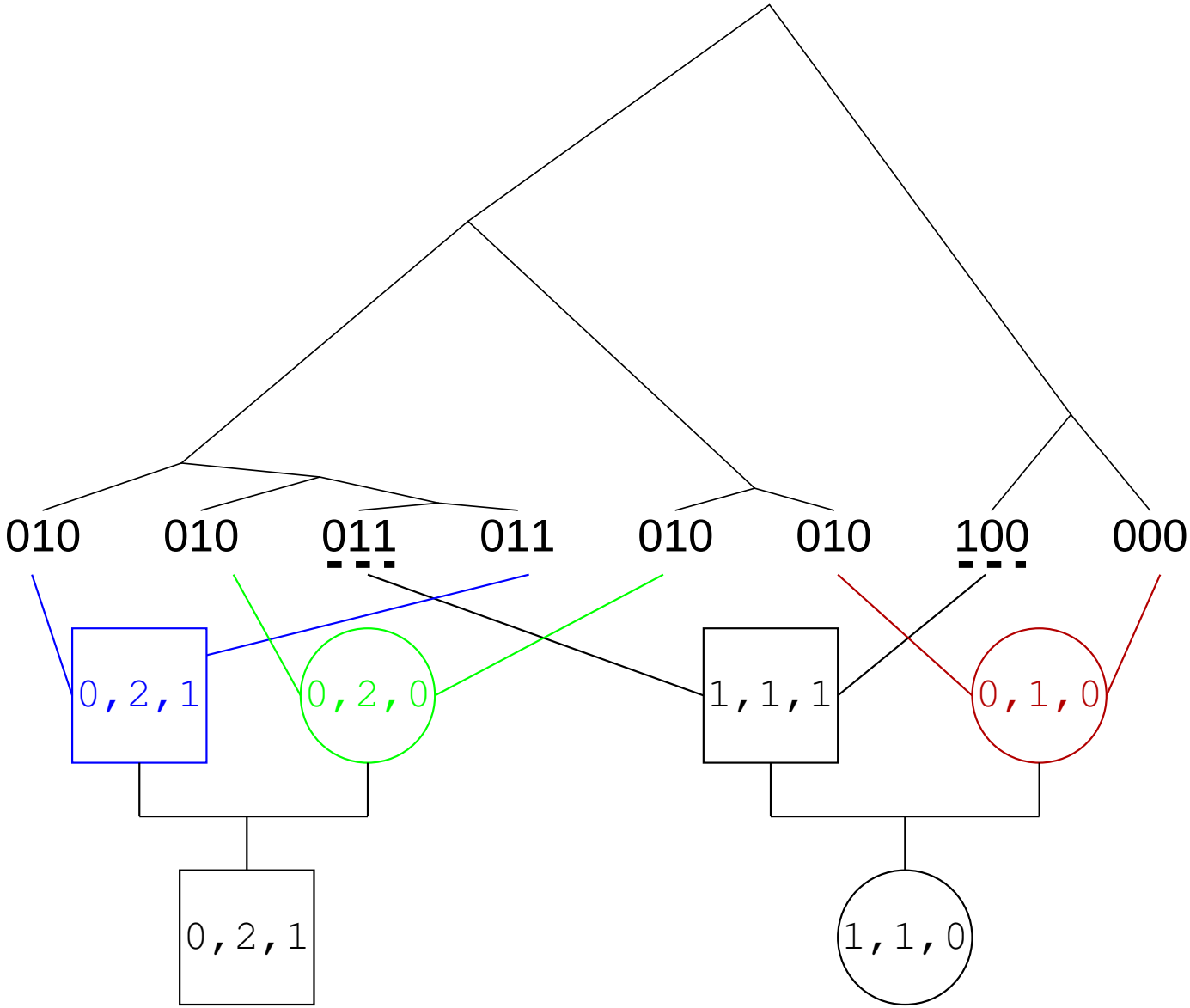


Table of possible genotype values for a trio

Mother	Father 0	Father 1	Father 2
0	0 (0/0)	0 (0/0) 1 (0/1)	1 (0/1)
1	0 (0/0) 1 (1/0)	0 (0/0) 1 2 (1/1)	1 (0/1) 2 (1/1)
2	1 (1/0)	1 (1/0) 2 (1/1)	2 (1/1)

- Each internal cell of the table contains the child's possible genotype.
- Brackets denotes the maternally inherited allele on the left and paternally inherited allele on the right.

Genotypes/haplotypes at 14 loci for an example trio

Genotype data

Individual	1	2	3	4	5	6	7	8	9	10	11	12	13	14
m	1	1	0	1	1	0	1	0	1	0	1	1	1	0
f	1	1	1	1	1	0	1	0	2	0	1	1	1	0
c	1	0	0	1	1	0	1	0	1	0	1	0	1	0

Haplotype data

Sequence	1	2	3	4	5	6	7	8	9	10	11	12	13	14
m_t	?	0	0	?	?	0	?	0	0	0	?	0	?	0
m_u	?	1	0	?	?	0	?	0	1	0	?	1	?	0
f_t	?	0	0	?	?	0	?	0	1	0	?	0	?	0
f_u	?	1	1	?	?	0	?	0	1	0	?	1	?	0

- m=Mother, f=Father, c=Child
- m_t, m_u = Mother's transmitted and untransmitted haplotypes, respectively
- f_t, f_u = Father's transmitted and untransmitted haplotypes, respectively
- '?' if alleles transmitted from the parents are not known

Extending model for genotype data on trios

- Now assume that **H** are set of (unknown) parental sequences. **G** are the observed genotypes for the trio.
- Let $\mathcal{F}$ denote the familial relationships present amongst members of the n trios. We are now interested in sampling from $f(\mathbf{T}|\mathbf{G}, \mathcal{F})$
- As before, model by augmenting the data with latent variables and sequentially conditioning on parental nodes:

$$f(\mathbf{T}, \mathbf{R}, \mathbf{S}, \mathbf{H}|\mathbf{G}, \mathcal{F}) \propto f(\mathbf{T}, \mathbf{R}, \mathbf{S}, \mathbf{H}, \mathbf{G}, |\mathcal{F}) = f(\mathbf{G}|\mathbf{H}, \mathcal{F}) f(\mathbf{T}, \mathbf{R}, \mathbf{S}, \mathbf{H}).$$

  - Assume no recombination/mutation in trios
  - Assuming compatility of genotypes with haplotypes/family structure, $\Pr(\mathbf{G}|\mathbf{H}, \mathcal{F}) = 1$
  - $f(\mathbf{T}, \mathbf{R}, \mathbf{S}, \mathbf{H})$ as previously defined

Extending model for genotype data on trios

- Now assume that **H** are set of (unknown) parental sequences. **G** are the observed genotypes for the trio.
- Let $\mathcal{F}$ denote the familial relationships present amongst members of the n trios. We are now interested in sampling from $f(\mathbf{T}|\mathbf{G}, \mathcal{F})$
- As before, model by augmenting the data with latent variables and sequentially conditioning on parental nodes:

$$f(\mathbf{T}, \mathbf{R}, \mathbf{S}, \mathbf{H}|\mathbf{G}, \mathcal{F}) \propto f(\mathbf{T}, \mathbf{R}, \mathbf{S}, \mathbf{H}, \mathbf{G}, |\mathcal{F}) = f(\mathbf{G}|\mathbf{H}, \mathcal{F}) f(\mathbf{T}, \mathbf{R}, \mathbf{S}, \mathbf{H}).$$

  - Assume no recombination/mutation in trios
  - Assuming compatility of genotypes with haplotypes/family structure, $\Pr(\mathbf{G}|\mathbf{H}, \mathcal{F}) = 1$
  - $f(\mathbf{T}, \mathbf{R}, \mathbf{S}, \mathbf{H})$ as previously defined

Extending model for genotype data on trios

- Now assume that **H** are set of (unknown) parental sequences. **G** are the observed genotypes for the trio.
- Let $\mathcal{F}$ denote the familial relationships present amongst members of the n trios. We are now interested in sampling from $f(\mathbf{T}|\mathbf{G}, \mathcal{F})$
- As before, model by augmenting the data with latent variables and sequentially conditioning on parental nodes:

$$f(\mathbf{T}, \mathbf{R}, \mathbf{S}, \mathbf{H}|\mathbf{G}, \mathcal{F}) \propto f(\mathbf{T}, \mathbf{R}, \mathbf{S}, \mathbf{H}, \mathbf{G}, |\mathcal{F}) = f(\mathbf{G}|\mathbf{H}, \mathcal{F}) f(\mathbf{T}, \mathbf{R}, \mathbf{S}, \mathbf{H}).$$

  - Assume no recombination/mutation in trios
  - Assuming compatility of genotypes with haplotypes/family structure, $\Pr(\mathbf{G}|\mathbf{H}, \mathcal{F}) = 1$
  - $f(\mathbf{T}, \mathbf{R}, \mathbf{S}, \mathbf{H})$ as previously defined

Implementing trio-based tree sampler

1. Determine as much of the phase as possible from the trio genotypes alone
   - For any loci with missing transmission, we know that:

$$(m_t, m_u, f_t, f_u) = (1, 0, 0, 1) \text{ or } (0, 1, 1, 0)$$

2. Choose initial values for loci with missing transmission
   - Sample from the two configurations above; condition on surrounding loci

3. Proposal distribution to propose new transmission of alleles at the loci with missing information.
   - Allele swap
   - Similar to proposal distribution used in our genotype-based sampler
   - Randomly sample a locus with uncertain phase. Assume the chosen locus is the l^{th} for trio i.
   - Swap the alleles at that locus in both parents.

$$(m_t, m_u, f_t, f_u) = (1, 0, 0, 1) \rightarrow (0, 1, 1, 0)$$

uOttawa

Implementing trio-based tree sampler

1. Determine as much of the phase as possible from the trio genotypes alone
   - For any loci with missing transmission, we know that:

$$(m_t, m_u, f_t, f_u) = (1, 0, 0, 1) \text{ or } (0, 1, 1, 0)$$

2. Choose initial values for loci with missing transmission
   - Sample from the two configurations above; condition on surrounding loci

3. Proposal distribution to propose new transmission of alleles at the loci with missing information.
   - Allele swap
   - Similar to proposal distribution used in our genotype-based sampler
   - Randomly sample a locus with uncertain phase. Assume the chosen locus is the l^{th} for trio i.
   - Swap the alleles at that locus in both parents.

$$(m_t, m_u, f_t, f_u) = (1, 0, 0, 1) \rightarrow (0, 1, 1, 0)$$

uOttawa

Implementing trio-based tree sampler

1. Determine as much of the phase as possible from the trio genotypes alone
   - For any loci with missing transmission, we know that:

$$(m_t, m_u, f_t, f_u) = (1, 0, 0, 1) \text{ or } (0, 1, 1, 0)$$

2. Choose initial values for loci with missing transmission
   - Sample from the two configurations above; condition on surrounding loci

3. Proposal distribution to propose new transmission of alleles at the loci with missing information.
   - Allele swap
   - Similar to proposal distribution used in our genotype-based sampler
   - Randomly sample a locus with uncertain phase. Assume the chosen locus is the l^{th} for trio i.
   - Swap the alleles at that locus in both parents.

$$(m_t, m_u, f_t, f_u) = (1, 0, 0, 1) \rightarrow (0, 1, 1, 0)$$

Example: Application to the trio data

- Ran the sampler on the Crohn's dataset
  - Couldn't run all focal points; first 20 only.
  - 9-10 million iterations per focal point. Need longer run lengths.
- Acceptance rate for new proposal distribution is very low (average of 0.1%)
  - Expected based on my experience with sampling conditional on genotypes. Need longer run lengths.
- Average mutation rate/recombination rate estimates are similar whether haplotypes are imputed or not:

	Mutation Rate		Recombination Rate	
Focal Point	Trio-based	Imputed	Trio-based	Imputed
2	2.66	2.58	0.00017	0.00014
5	2.55	2.55	0.00019	0.00015
7	3.03	2.71	0.00020	0.00015
16	3.20	3.01	0.00035	0.00022
20	2.96	2.79	0.00084	0.00073

Example: Application to the trio data

- Ran the sampler on the Crohn's dataset
  - Couldn't run all focal points; first 20 only.
  - 9-10 million iterations per focal point. Need longer run lengths.
- Acceptance rate for new proposal distribution is very low (average of 0.1%)
  - Expected based on my experience with sampling conditional on genotypes. Need longer run lengths.
- Average mutation rate/recombination rate estimates are similar whether haplotypes are imputed or not:

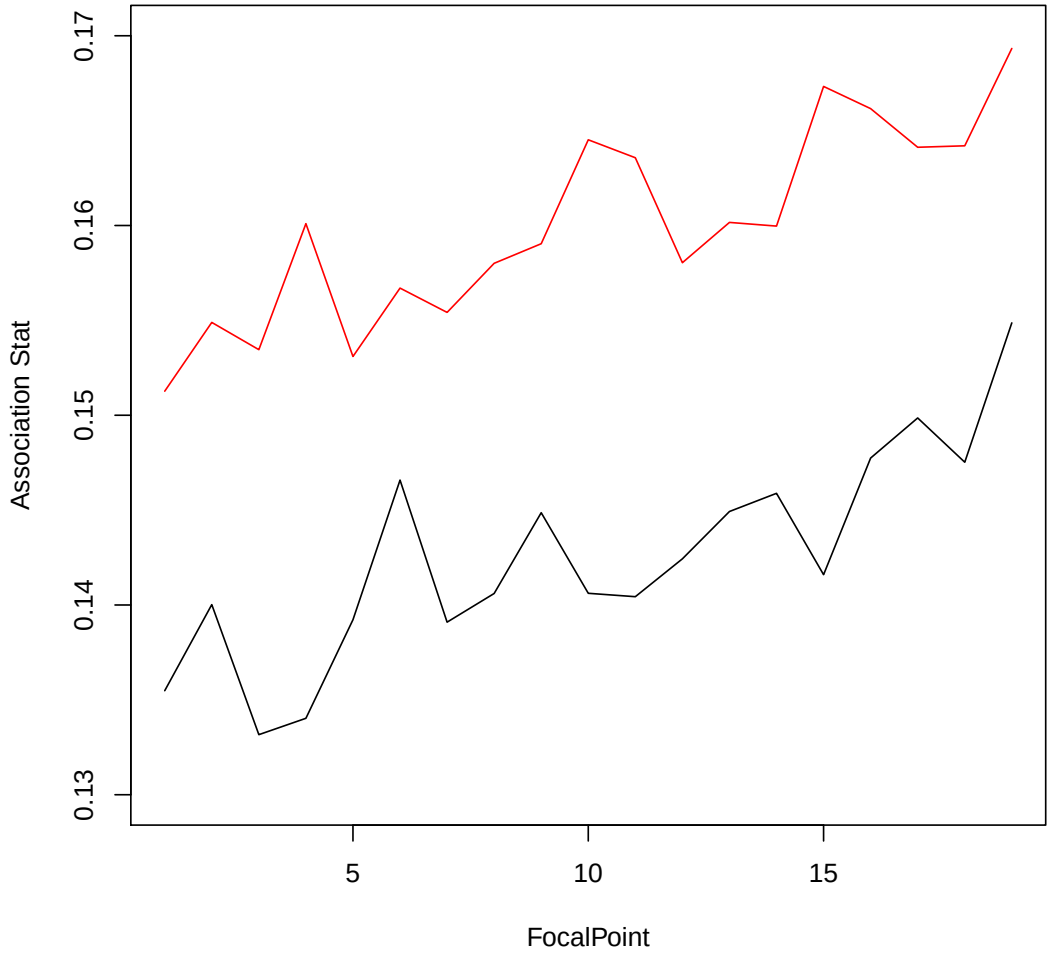
	Mutation Rate		Recombination Rate	
Focal Point	Trio-based	Imputed	Trio-based	Imputed
2	2.66	2.58	0.00017	0.00014
5	2.55	2.55	0.00019	0.00015
7	3.03	2.71	0.00020	0.00015
16	3.20	3.01	0.00035	0.00022
20	2.96	2.79	0.00084	0.00073

Example: Application to the trio data

- Ran the sampler on the Crohn's dataset
  - Couldn't run all focal points; first 20 only.
  - 9-10 million iterations per focal point. Need longer run lengths.
- Acceptance rate for new proposal distribution is very low (average of 0.1%)
  - Expected based on my experience with sampling conditional on genotypes. Need longer run lengths.
- Average mutation rate/recombination rate estimates are similar whether haplotypes are imputed or not:

	Mutation Rate		Recombination Rate	
Focal Point	Trio-based	Imputed	Trio-based	Imputed
2	2.66	2.58	0.00017	0.00014
5	2.55	2.55	0.00019	0.00015
7	3.03	2.71	0.00020	0.00015
16	3.20	3.01	0.00035	0.00022
20	2.96	2.79	0.00084	0.00073

Example: Application to the trio data



Red - Imputed data; Black - Trio data

Summary

Some thoughts based on my experiences with trio data:

- Conceptually ‘easy’ to extend previous work to trios.
  - Tips of the genealogy are the parents haplotypes.
  - Don’t need to model the history within the families.
- Need to improve the sampling at loci with missing transmission
  1. Better initialization of tip sequences
  2. Add topology change after a swap.
- Missing data
- Bigger families

Acknowledgements and References

Thanks to Bryan Paget for programming help

Some of the work presented was joint work with Jinko Graham, Brad McNeney, Celia Greenwood:

Burkett KM, McNeney B, Graham J. Sampletrees and Rsampletrees: sampling gene genealogies conditional on SNP genotype data. Bioinformatics, 32:1580-2, 2016.

Burkett KM, McNeney B, Graham J, Greenwood CMT. Using gene genealogies to detect rare variants associated with complex traits. Human Heredity, 78:117-130, 2014.

Burkett KM, Greenwood CMT, McNeney B, Graham J. Gene genealogies for genetic association mapping, with application to Crohn's disease. Frontiers in Statistical Genetics and Methodology, 4: article 260, 2013c.

Burkett KM, McNeney B, Graham, J. Markov chain Monte Carlo sampling of gene genealogies conditional on unphased SNP genotype data. Statistical Applications in Genetics and Molecular Biology, 12: 559-581, 2013b.

Burkett KM, McNeney B, Graham J. A Markov chain Monte Carlo sampler for gene genealogies conditional on haplotype data. In Chaubney, Y., editor, Some Recent Advances in Mathematics and Statistics, Proceedings of Statistics 2011 Canada/IMST 2011-FIM XX, Montreal, July 2011. Singapore, World Scientific Publishing, 2013, 29-44a.

Browning BL and Browning S. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. American Journal of Human Genetics, 84:210–223, 2009.

Rioux, J. D. at al. Genetic variation in 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nature Genetics, 29:223–228, 2001.

Zöllner, S. and Pritchard, J. K. Coalescent-based association mapping and fine mapping of complex trait loci. Genetics, 169:1071–1092, 2005.