Recombination, and haplotype structure

Simon Myers, Gil McVean
Department of Statistics, Oxford

The starting point

• We have a genome’s worth of data on genetic variation

• We wish to understand why the haplotype structure looks how it does
  – Differences between regions, populations

Where do haplotypes come from?

• In the absence of recombination, the most natural way to think about haplotypes is in terms of the genealogical tree representing the history of the chromosomes

• Tree affects mutation patterns
• Mutation patterns give information on tree

What determines the shape of the tree?
Ancestry of current population

The coalescent: a model of genealogies

Simulating histories with the coalescent
Simulating data with the coalescent

Haplotype structure in the absence of recombination

- In the absence of recombination, the shape of the tree and where mutations fall on it determine patterns of haplotype structure
- Two mutations on the same branch will be in complete association, mutations on different branches will have lower and often low association

\[ r^2 = 1 \]
\[ r^2 = 0.04 \]

Haplotypes when there is recombination

- When there is no recombination, haplotype structure reflects the age distribution of mutations and the shape of the underlying tree
- When there is some recombination, every nucleotide position has a tree, but the tree changes along the chromosome at a rate determined by the local recombination landscape
- By using SNP information to inform us about the trees, we can learn about how quickly the trees changes
  - This relates to the recombination rate

A bit of recombination ‘shuffles’ genetic variation
Lots of recombination does lots of shuffling

Recombination and haplotype diversity

- Without recombination, a new mutation can create at most one new haplotype
  - Any two mutations delineate at most 3 haplotypes in total (ancestral, plus two new types)
- With recombination, this mutation can spread onto every existing haplotype background, creating the potential for more haplotypes
- For a given number of SNPs a region with recombination will tend to have (in comparison to a region with no recombination)
  - More haplotypes
  - Less variance in the pairwise differences between haplotypes
  - Less skewed haplotype frequencies

The ancestral recombination graph

- The combined history of recombination, mutation and coalescence is described by the ancestral recombination graph

In humans, recombination is not uniformly distributed

- Most recombination occurs in recombination hotspots – short (1-2kb) regions every 50-100kb that occupy at most 3% of the genome but probably account for 90% or more of the recombination
- This means that haplotype structure in humans is an interesting hybrid between the no recombination and lots of recombination situations
Learning about recombination

- Just like there is a true genealogy underlying a sample of sequences without recombination, there is a true ARG underlying samples of sequences with recombination.

- We can consider nonparametric and parametric ways of learning about recombination.

- There are useful nonparametric ways of learning about recombination which we will consider first.
  - These really only apply to species, such as humans, where we can be fairly sure that most SNPs are the result of a single ancestral mutation event.

The signal of recombination?

Detecting recombination from DNA sequence data

- Look for all pairs of “incompatible” sites.

- Find minimum number of intervals in which recombination events must have occurred (Hudson and Kaplan 1985): $R_m$.

Improving the detection algorithm

- $R_m$ greatly underestimates the amount of recombination in the history of a set of sequences.

- Myers and Griffiths (2003) developed an improved way of detecting recombination events.
  - Without recombination, every new mutation can create only a single new haplotype.
  - With recombination, mutations can be shuffled between haplotype background, generating haplotype diversity.
  - Each recombination makes at most one new haplotype.
  - If I see $H$ haplotypes with $S$ segregating sites, at least $H-S-1$ recombination events must have occurred.

- This offers potential to identify many more recombination events.
  - Carefully combine bounds from different collection of sites.
  - Dynamic programming algorithm makes computation extremely fast.
  - Better (sometimes slower) algorithms developed recently.
**Problems with ‘counting’ recombination events**

A tree-pair where we could see recombination events, but don’t

Tree-pairs where we cannot see recombination events

---

**Modelling recombination**

- Model-based approaches to learning about recombination allow us to ask more detailed questions than nonparametric approaches
  - What is the rate of recombination (as opposed to just the number of events)
  - Does gene A have a higher recombination rate than gene B?
  - Is the rate of recombination across a region constant?
  - Where are the recombination hotspots?

- We can use coalescent model approaches (approximations) to calculating the likelihood of arbitrary recombination maps given observed data

---

**Fitting a variable recombination rate**

- Use a reversible-jump MCMC approach (Green 1995)

---

**Acceptance rates**

\[
\alpha(\psi', \psi) = \min \left[ 1, \frac{\ell_c(\psi')}{\ell_c(\psi)} \times \frac{\pi(\psi')}{\pi(\psi)} \times \frac{q(\psi', \psi)}{q(\psi, \psi')} \times \left| \frac{\partial(\psi', u')}{\partial(\psi, u)} \right| \right]
\]

Composite likelihood ratio

Ratio of priors

Hastings ratio

Jacobian of partial derivatives relating changes in dimension to sampled random numbers

- Include a prior on the number of change points that encourages smoothing
Strong concordance between fine-scale rate estimates from sperm and genetic variation

Rates estimated from genetic variation

Rates estimated from sperm
Jeffreys et al (2001)

Inferring hotspots

- We perform a statistical test for hotspot presence
- Based on an approximation to the coalescent similar to that used for rate estimation
- All previously identified hotspots are 1-2kb in size
  - At a position in genome, consider where 2kb hotspot might be present
  - Fit a model with hotspot
  - Fit one without
  - Compare in terms of (approximate) likelihood ratio test
  - Evaluate significance via simulation
  - When p-value below threshold, declare a hotspot

Rates and hotspots across the human genome

From Myers et al. (2005)

Applications of recombination approaches to real data

- Rates and hotspots across the human genome (Myers et al. 2005)
  - Previously, no understanding of why hotspots localise where they do
  - Can 35,000 hotspots, accounting for >50% of human recombination, help?
- Comparison of recombination rates (Winckler et al. 2004, Ptak et al. 2005)
  - Between humans and chimpanzees
  - At individual recombination hotspots
- Understanding genomic rearrangements (Myers et al., submitted!)
  - Cause a number of “genomic disorders”
  - Relationship to recombination hotspots
32,996 Phase II HapMap hotspots

Estimated 50-70% of all human recombination
Hotspots on all chromosomes, including X

- \( \text{THE1B} \): Found in 1196 hotspots versus 606 coldspots \( (p<10^{-20}) \)
- \( \text{AluY} \): Found in 3635 hotspots versus 3262 coldspots \( (p=7 \times 10^{-5}) \)

~20,000 hotspots localised to within 5kb

THE1B: (LTR of retrotransposon)

Human hotspot motifs

- In humans, specific words produce recombination hotspot activity
  - Hotspot motif CCTCCCTNCCAC \( (p<10^{-23}) \)
    - Raises probability of a hotspot across genetic backgrounds
    - Degenerate versions CCNCCNTNNCCNC and truncated CCTCCCT also raise probability, to lesser extent
    - Motif explains ~40% of human hotspots
    - Operates in both sexes
    - We don’t know, very clearly, which hotspots
    - On THE1 background, hotspot 70-80% of time!
  - Biology not clearly understood
  - We identified a second, different hotspot motif (the best 9bp motif), CCCCACC

Variation in individual hotspots

Sequence variation affects recombination at DNA2 (Jeffreys and Neumann, Nature Genetics 2002)
SNPs disrupting hotspots

- **DNA2:**

  **Hot** AAAAAAGACAGCCCTCCCTGTGGCTGC
  **Cold** AAAAAAGACAGCCCTCCCTGTGGCTGC

- **NID1:**
  - Hot CACC CCCACCCCCACCCCCAACATA
  - Cold CACC TCCCCACCCCCACCCCCAACATA

Disruption of CCCCACCCC, best 9bp motif

Role of motif in X-linked ichthyosis

- Many other diseases are caused by recombination-mediated deletions and duplications (NAHR)
  - Smith-Magenis syndrome (hotspot)
  - CMT1A (hotspot)
  - NF1 microdeletion syndrome (hotspot)
  - DiGeorge syndrome....

- Two recent studies suggest normal hotspots and hotspots of disease-causing deletion may coincide
  - de Raadt, Stephens et al. (Nature Genetics, 2006)
  - Two NF1 deletion hotspots both likely to coincide with crossover hotspots
  - Lindsay et al. (ASHG, 2006)
  - CMT1A deletion hotspot associated with crossover hotspot