Genome-wide variation in rates of gene conversion and crossover

Garrett Hellenthal & Matthew Stephens

Department of Statistics
University of Washington, Box 354322, Seattle WA 98195-4322, USA
garretth@stat.washington.edu

RECOMBINATION: Crossover vs. Gene Conversion

Crossover and **gene conversion** are thought to be the two possible outcomes of a single biological process known as the **double-strand-break (DSB)** (Fig 1).

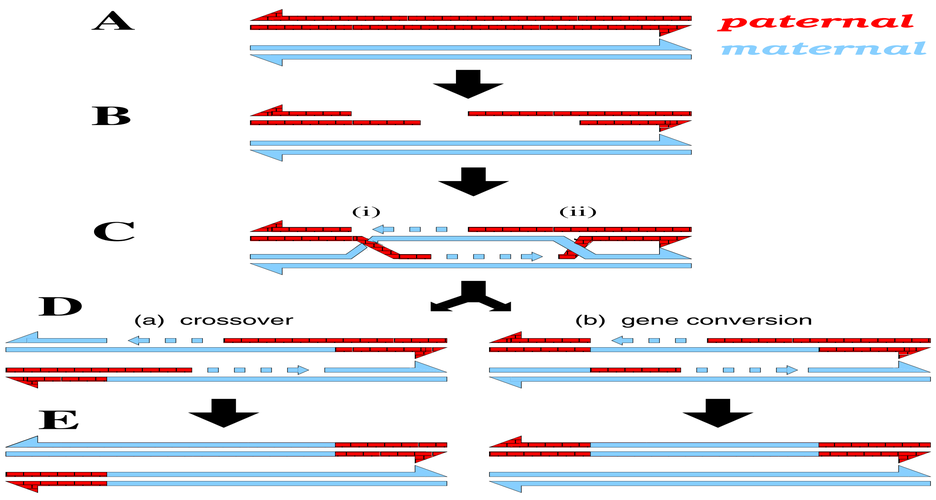

Figure 1: *An example of the DSB model of meiotic recombination. Two homologous chromosomes align (***A***). A double-strand-break occurs in the initiating (here paternal) chromosome (***B***), and is repaired by the formation of Holiday junctions (***C***). These junctions are "resolved" in one of two ways, resulting in either a gene conversion with crossover (***D-a***) or a gene conversion without exchange of flanking markers (***D-b***). Mismatch correction (***E***) completes the process (here in favor of the maternal chromosome). (adapted from Franklin Stahl's webpage)*

From experimental sperm analysis and statistical methods applied to pedigree and population data, rates of crossover are known to vary at both fine and large scales. This variation has been observed to be correlated with sequence factors such as GC-content and SNP density.

It appears there are two main processes behind the formation of crossovers: (1) the double-strand-break (Fig 1-B), and (2) the resolution of the Holliday junction (Fig 1-C, Fig 1-D). Some questions that arise:

1. Is genome-wide variability in crossover rate due more to variability in the rate of double-strand-breaks or to variability in the relative rate in which Holliday junctions are resolved as gene conversions versus crossovers? Or are they roughly equal?
2. Is the observed correlation between crossover rates and any particular sequence factor due to a correlation between DSB rates and that factor or a correlation between relative rates of Holliday junction resolution and that factor? Are these two processes independent?

The prevailing wisdom seems to be that variability in double-strand-break rates predominantly determines variability in crossover rates. While this is probably true in crossover "hotspots," what about outside of hotspots?

Model

We examine these questions by using patterns of LD in population data (i.e. chromosome data on individuals randomly sampled from a population) to estimate rates of crossover and gene conversion (see Fig 2) and thus rates of DSB and the relative rate of gene conversion to crossover (commonly referred to in the literature as f). We estimate f and DSB rates per regions of ≈20-40kb genes spread across the genome, utilizing a Bayesian hierarchical model that borrows strength across regions to improve estimation (in particular, f has proved difficult to estimate per gene with precision in previous research (e.g. Frisse et al 2001, Wall 2004)).

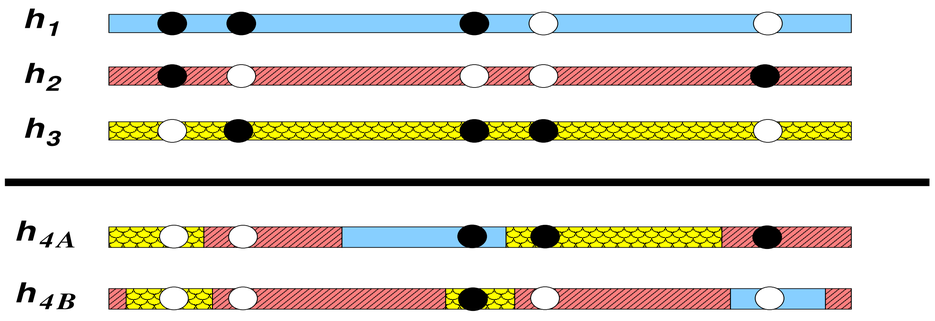

Figure 2: *Capturing effects of crossover and gene conversion via associations in haplotype SNP data. Rectangles represent haplotypes and circles represent biallelic SNPs. The model considers each haplotype in turn and assumes each added haplotype will look like a mosaic of the previously observed ones. Consider observing the top three haplotypes at five SNPs. Two examples of h_4 are given: h_{4A}, h_{4B}. The coloring scheme illustrates which previous haplotype each SNP in h_4 is "most closely related" to. The first example, h_{4A}, is formed out of several "crossover" events, such that each consecutive SNP "copies" from a different haplotype. The second example, h_{4B}, copies chiefly from h_2, but has experienced several "gene conversion" events (each assumed to replace at most 1 SNP), replacing SNPs 1, 3, and 5. Counting the number of such (unobserved) "crossovers" and "gene conversions" provides an estimate for the actual underlying rates of crossover and gene conversion in the region.*

The model makes the following assumptions:

- chromosomes randomly sampled from a constant-sized population with no selection
- both "crossover" events and "gene conversion" events (see Fig 2) occur on a chromosome as independent Poisson processes
- tract length of gene conversions is ≈ 100bp
- no repeat mutation
- f is constant within genes
- there is at most one crossover "hotspot" (in which f is constant) per gene

We concentrate on exploring rates of DSB and f *outside* of any fine-scale recombination "hotspots" (gene conversion is difficult to estimate within hotspots because there are few SNPs in hotspots; therefore within hotspot f estimation is poor).

Application & Results

We applied the model to the *SeattleSNP* dataset, which consists of SNPs sequenced in 24 individuals of African-American ancestry and 23 (CEPH) individuals of European ancestry. We analyzed each group separately, considering 204 African-American genes and 173 CEPH genes.

To summarize variability in estimated rates of DSB and f, we use the ratio of the 95^{th} quantile value and the 5^{th} quantile value of each (Table 1).

Results suggest f is considerably more variable than DSB!

	Af-Amer	European
DSB	4 (3-5)	5 (4-7)
f	40 (28-57)	60 (42-84)

Table 1: *Estimated mean factor by which f and DSB rates vary among genes (parentheses represent 95% credible intervals).*

We performed a multiple linear regression for each of crossover rates, DSB rates, and relative rates of gene conversion to crossover (f) (each on a $\log_{10}$ scale) on GC-content and SNP-density to assess correlations (Table 2, results omitted for crossover rates).

GC-content and SNP density appear to be more correlated with f than with DSB! (Fig 3, Table 2)

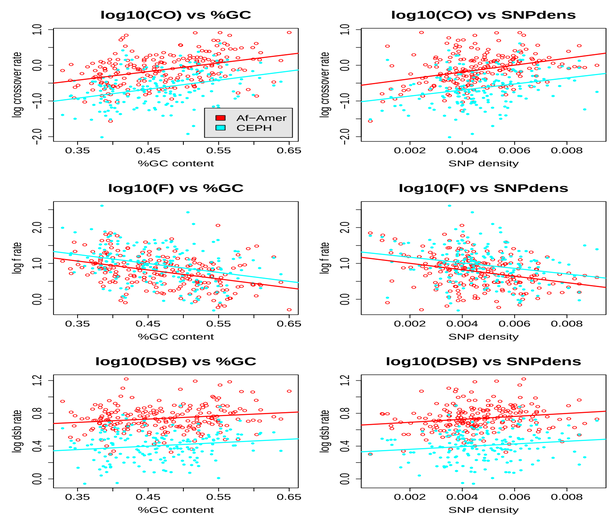

Figure 3: *The top row shows a correlation between estimates of crossover rate (here on a $\log_{10}$ scale) and each of GC-content and SNP-density for both the African-American and European (CEPH) samples, in agreement with previous observations (e.g. Crawford et al. 2004). Breaking the crossover rate into its f and DSB components, it appears this pattern is due more to a correlation with f (second row) than with DSB (third row). This suggests both GC-content and SNP-density may be more correlated with Holliday junction resolution than with double-strand-break initiation (see Table 2 as well).*

	DSB			f		
	Coeff	S.E.	p-val	Coeff	S.E.	p-val
Af-Amer						
Intercept	0.50	0.08	–	2.16	0.19	–
%GC content	0.35	0.16	0.025	-2.23	0.40	< 0.001
SNP density	15.63	6.87	0.024	-74.86	17.65	< 0.001
European						
Intercept	0.17	0.10	–	2.30	0.27	–
%GC content	0.36	0.20	0.075	-2.24	0.56	< 0.001
SNP density	13.73	8.49	0.108	-61.40	23.42	0.009

Table 2: *Summary of linear regressions performed on estimates of $\log_{10}$DSB and $\log_{10} f$ versus GC-content and SNP density, for genes in the SeattleSNP genotype data of African-American and European (CEPH) descent.*

Conclusions

- On average, f appears to be ≈10, suggesting gene conversions occur on average 10 times more often than crossovers across the genome (though with perhaps more variability than previously thought).
- Outside of hotspots, variability in the relative rate of gene conversion to crossover appears to be higher than variability in double-strand-breaks.
- These two biological processes appear to be independently-acting (i.e. f and DSB rates are uncorrelated). (results omitted)
- GC-content and SNP density are more correlated with estimated rates of f than with estimated rates of DSB.

Future Work

- Simulate data with multiple hotspots (crossover and/or gene conversion) to see if this gives, e.g., falsely large variability in rates of f relative to DSB outside of hotspots
- Simulate genotyping error and repeat mutation (either could inflate estimates of gene conversion while presumably keeping estimates of crossover rates relatively unchanged, thus increasing f estimates and perhaps variability)
- Simulate to test robustness of model to violations of other assumptions