European Journal of Human Genetics (2013), 1–8

www.nature.com/ejhg

ARTICLE

Meiotic gene-conversion rate and tract length variation in the human genome

Badri Padhukasahasram*[,1,2] and Bruce Rannala[2]

Meiotic recombination occurs in the form of two different mechanisms called crossing-over and gene-conversion and both processes have an important role in shaping genetic variation in populations. Although variation in crossing-over rates has been studied extensively using sperm-typing experiments, pedigree studies and population genetic approaches, our knowledge of variation in gene-conversion parameters (ie, rates and mean tract lengths) remains far from complete. To explore variability in population gene-conversion rates and its relationship to crossing-over rate variation patterns, we have developed and validated using coalescent simulations a comprehensive Bayesian full-likelihood method that can jointly infer crossing-over and gene-conversion rates as well as tract lengths from population genomic data under general variable rate models with recombination hotspots. Here, we apply this new method to SNP data from multiple human populations and attempt to characterize for the first time the fine-scale variation in gene-conversion parameters along the human genome. We find that the estimated ratio of gene-conversion to crossing-over rates varies considerably across genomic regions as well as between populations. However, there is a great degree of uncertainty associated with such estimates. We also find substantial evidence for variation in the mean conversion tract length. The estimated tract lengths did not show any negative relationship with the local heterozygosity levels in our analysis.
European Journal of Human Genetics advance online publication, 27 February 2013; doi:10.1038/ejhg.2013.30

Keywords: Bayesian; crossing-over; gene-conversion; MCMC; heterozygosity

INTRODUCTION

Meiotic recombination is an important determinant of linkage disequilibrium (LD; ie, the non-random associations between alleles) in population genomic data. This process shuffles material between homologous chromosomes and creates mosaic chromosomes during the first meiotic division. Current models allow for two different mechanisms of genetic exchange called crossing-over (possibly accompanied by gene-conversion) and gene-conversion (without crossing-over). Crossing-over involves the reciprocal exchange of large chromosomal segments via double-stranded DNA breaks while gene-conversion involves the transfer of short-tracts. Note that conversion events accompanying crossing-over cannot be detected by population based methods, and so here gene-conversion refers to only events not accompanied by crossing-over. Both mechanisms contribute to genetic diversity by breaking down allelic associations and lead to the decay of LD over time.

Many population genetic studies have ignored the effects of gene-conversion although crossing-over and gene-conversion events have qualitatively different effects on the evolutionary history of chromosomes and leave different traces in genomic polymorphism data. Although crossing-over rates are known to vary tremendously, by several orders of magnitude at the fine-scale (kb),[1–9] results are limited concerning the homologous gene-conversion process and the extent to which meiotic conversion rates and tract lengths vary along the human genome.[10–15] Gene-conversion rates have been estimated by experimental studies in yeast and fruit flies[16–20] and tract lengths are estimated to be in the range of 350–2000 bp in these organisms. A basic knowledge of gene-conversion rate and tract length variation will further our understanding of the recombination process, refine the design of association mapping studies and help fine-tune methods for inferring demographic parameters and natural selection along the human genome. Using sperm typing,[12] estimated gene-conversion rates at three known human crossing-over hotspots and found that all three regions showed evidence of being gene-conversion hot spots as well. They estimated that gene-conversion occurs at a rate of 4–15 times the crossing-over rate and mean tract lengths are in the range of 54–132 bp. In all cases, the peaks of gene-conversion rates coincided with the peaks of crossing-over rates, which suggests that the molecular mechanisms generating most cross-overs and gene-conversion events are related.[21]

Currently, there exist several statistical methods designed to jointly estimate the crossing-over and gene-conversion parameters from population genetic data. Methods developed by (refs 11, 22, 23) generalize the composite likelihood approach first proposed in Hudson[24] (also see McVean *et al*[25]). These approaches divide the data into small subsets (pairs or triplets of segregating sites), calculate likelihoods for these subsets and multiply them together. This is equivalent to assuming that they are independent. The likelihood thus obtained is called the composite likelihood. Composite likelihood methods use pre-computed likelihood look-up tables for all the possible configurations of the subsets and are typically fast. However, because the subsets are not actually independent of one another, they

[1]Center for Health Policy and Health Services Research, Henry Ford Health System, Detroit, MI, USA; [2]Genome Center and Department of Evolution and Ecology, University of California, Davis, Davis, CA, USA
*Correspondence: Dr B Padhukasahasram, Center for Health Policy and Health Services Research, Henry Ford Health System, One Ford Place 3A, Detroit, MI 48202, USA. Tel: +1 3138747051; Fax: +1 3138747137; E-mail: pkbadri@yahoo.com
Received 17 September 2012; revised 17 December 2012; accepted 10 January 2013

only approximate the likelihood of the data. One consequence is that correct confidence intervals for composite likelihood estimates can only be obtained by using simulations. Another consequence is that information is potentially discarded because higher-order multilocus associations are ignored.

Padhukasahasram *et al*[14] described a rejection-sampling method that simultaneously utilizes informative long-range and short-range summary statistics to infer the recombination parameters. This approach also ignores some of the information available in the data for the sake of computational efficiency. However, it has the advantage that confidence intervals may be directly calculated from the likelihood surface without simulations.

Another approximate likelihood method for estimating crossing-over rates (called the product of approximate conditionals, or PACs) that was proposed in Li and Stephens[26] has also been extended by several subsequent studies to include gene-conversion. Briefly, this method infers recombination parameters under a heuristic model using all the information available in the data and is computationally efficient. However, inference is currently restricted to the constant population size Wright–Fisher model only. Hellenthal[27] uses a PAC model where the conversion tract can include at most one marker. Gay *et al*[15] improved on their work to allow for arbitrary gene-conversion tract lengths and this method can be used for co-estimating crossing-over and gene-conversion rates as well as tract lengths. One simplification of their model was that they disallowed overlapping gene-conversion events. Yin *et al*[28] further generalized this work to allow for overlapping events, leading to a more accurate PAC-based method for jointly estimating crossing-over, gene-conversion and the mean conversion tract length.

Here, we further develop a recent Bayesian population genomic approach we proposed to study gene-conversion rate variation and its relationship to crossing-over. Padhukasahasram and Rannala[29] introduced a coalescent-based Markov Chain Monte Carlo (MCMC) method for jointly estimating crossing-over, gene-conversion rates and mean conversion tract lengths from population genomic data. In this article, we further develop and extend this full-likelihood methodology to infer all three recombination parameters under models with recombination hotspots. Full-likelihood methods such as ours are appealing from a statistical perspective because they perform model-based inference based on the exact configuration of the observed sample of haplotypes or genotypes and thus their efficiency is expected to be optimal in theory. In contrast to the recombination model used previously, here we allow the relative rates of gene-conversion to crossing-over (f) to take different possible values depending on whether we are located within a recombination hotspot or outside and include many novel proposal moves designed to improve the efficiency of the MCMC algorithm. We first validate the new method using coalescent simulations and then apply it to two human data sets genotyped in the human leukocyte antigen (HLA) region of chromosome 6 and the MS32 region on chromosome 1 and estimate the conversion parameters for these loci. Next, we use it to analyze SNP data from three different human populations (Northwest European descent residing in Utah (CEU), Yoruba from Ibadan, Nigeria (YRI) and Han Chinese in Beijing (CHB)) genotyped as part of the HapMap project. One goal is to address several longstanding open questions about the process of meiotic gene-conversion and, in particular, how rate parameters vary across the human genome. We also explore how gene-conversion rates vary in relation to crossing-over rates and check whether these patterns are conserved between populations. Finally, we examine whether mean conversion tract lengths are variable along the genome or between populations and test whether gene-conversion parameters show any systematic relationship with the local heterozygosity levels.

MATERIALS AND METHODS

Our recombination inference method is based on the retrospective coalescent framework in which the genealogy of a sample of sequences is approximated as a graph called the ancestral recombination graph (ARG).[30–32] In particular, the method uses the coalescent with gene-conversion as described in Wiuf and Hein.[33] We use a variable recombination rate model with background rate variation and hotspots as part of our inference procedure. In this model, the background crossing-over rate follows a gamma distribution with shape and scale parameters and includes recombination hotspots that arise according to the process described in Wang and Rannala[34]. The extension of the Wang and Rannala model described in Padhukasahasram and Rannala[29] assumed that gene-conversion and crossing-over rates vary in an identical pattern such that the ratio of rates (f) is constant along the sequence. Thus, all crossing-over hotspots are also gene-conversion hotspots and both parameters can take different values in each of the marker intervals. Transitions from hotspots to non-hotspots or *vice versa* occur at points referred to as change points. We propose various modifications to these change points in the MCMC chain to modify the locations of hotspots along the sequence.

The model we use here is a variation of the recombination model described in Padhukasahasram and Rannala.[29] In contrast to the previous model, the present model allows the ratio of gene-conversion to crossing-over rates to vary along the sequence, taking one of two possible values, f_1 or f_2, depending on whether or not the sequence is within a hotspot, respectively. The MCMC program we have developed, attempts to jointly estimate both f_1 and f_2 along with other parameters of interest from the data. In addition, we perform two different kinds of inferences here based on models with hotspots where: (i) the mean tract length (m) is fixed to some reasonable value (eg, 125 bp); and (ii) the mean tract length is jointly estimated along with all other recombination parameters.

Recombination hotspot model

Following Wang and Rannala,[34] it is assumed that the distribution of recombination hotspots along chromosomes follows a Markov process. Hotspots arise with an instantaneous rate of λ_1 and revert with an instantaneous rate of λ_2. The waiting distance until the occurrence of a hotspot is exponentially distributed with parameter λ_1 and the distance till the loss of a hotspot is also exponentially distributed with a rate of λ_2. The parameters $1/\lambda_1$ and $1/\lambda_2$ represent the average distance between hotspots and the average width of a hotspot, respectively. The probability that a sequence starts with a hotspot is $\lambda_1/(\lambda_1+\lambda_2)$. In Appendix A1–A3 (Supplementary Information), we present some useful theoretical results for this recombination model. Three variables are associated with each hotspot, denoted by X_1, X_2 and Z, and represent the starting location, the ending location and the intensity of the hotspot, respectively. Variable Z has a prior distribution that is log-normal with parameters $\mu_z=9$ and $\sigma_z=1$.

Background rate variation

The prior distribution of background recombination rates between SNPs is assumed to follow a Γ distribution with shape (sh) and scale (sc) parameters. The shape parameter is fixed to 100, while the scale parameter is estimated via the MCMC algorithm.

Let X denote a sample of haplotypes or genotypes and let G_S denote an ARG consistent with the data. Many of the moves in the MCMC chain in the new version are similar to those previously described in Padhukasahasram and Rannala.[29] The moves involving changes in hotspot locations, widths, etc are identical to those in Wang and Rannala,[34] except that the prior probability of a genealogy involves terms ρ, f_1 and f_2 and mean tract length m. Note that the ρ vector, f_1 and f_2 together determine the vector of γ values (ie, the gene-conversion rates for marker intervals) in the variable recombination rate model. Let H denote the set of hotspots and H_i represent the *ith* hotspot in the sequence. Each H_i is a vector with three variables the start (X_1), end (X_2) and intensity (Z) of the hotspot. Let $f(\cdot)$ denote priors, $Q(\cdot)$ denote proposal distributions and let $l(X|G_S, \theta)$ be the likelihood of the data given an ARG,

G_S, and the population mutation rate, θ. Variables followed by a prime sign denote the proposals. In Appendix A4, we have described in detail the moves used in the new model (see references 35 and 36 for running multiple chains).

RESULTS

Checking the MCMC program

We first ran the MCMC program without any data and compared the posterior distributions of various quantities of interest with their prior distribution. Supplementary Figures S1–S4 show the results of this analysis. We find that when the likelihood ratio is set equal to 1 and the chain is run without data, the posterior distributions for the variables match their priors as expected. The priors used for f_1 and f_2 in these tests were uniform between 0 and 5. For hotspot intensity and hotspot width, the priors are described in the Materials and methods section and in Appendix A1 (Supplementary Information).

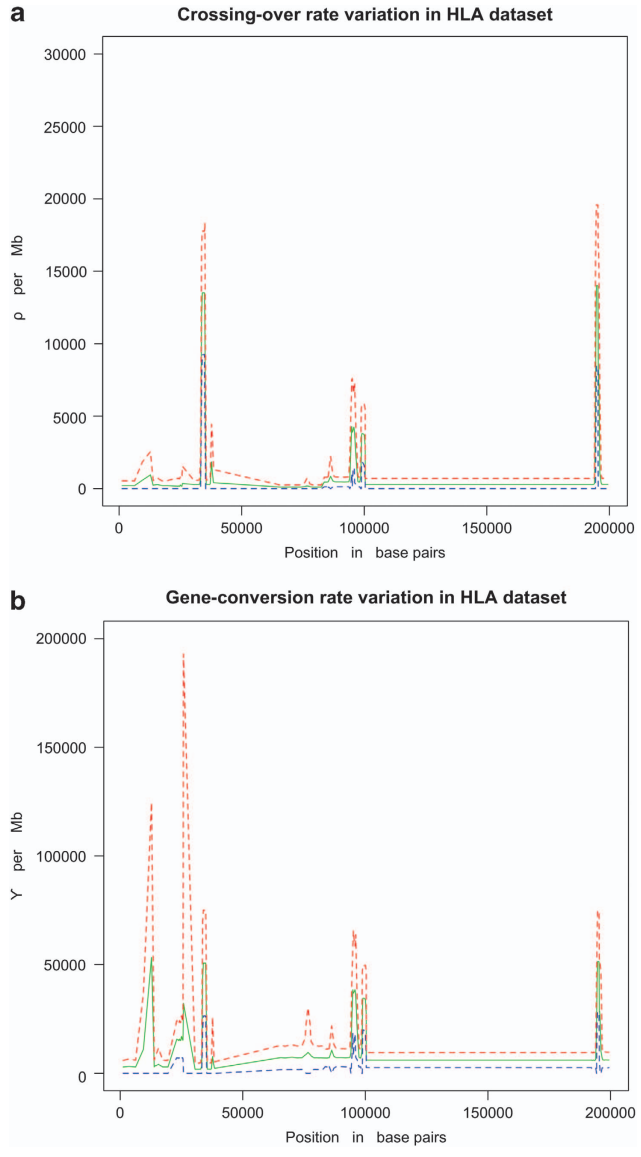

Figure 1 Crossing-over (ρ) and gene-conversion (γ) rate variation in the HLA data set assuming $m = 125$ bp. The solid green curve shows the posterior means of the estimated rates while the dashed blue and red lines represent the lower and upper bounds of the 95% credible intervals, respectively.

Test runs on simulated data

To check the correctness of our MCMC algorithm, we also tested the method on data sets simulated with both crossing-over and gene-conversion hotspots. We simulated 10 independent data sets of 20 kb sequences, 50 samples with a single recombination hotspot of width 2 kb at the center of the sequence, $\theta = 20.0$ and varying values of gene-conversion tract length m. Supplementary Figure S5 and Supplementary Table S1 show the results obtained for this analysis. We find that the estimated locations of recombination hotspots, intensities and mean tract lengths are consistent with the values used in simulations.

Analysis of data from MHC and MS32 regions in humans

We applied the new method to two data sets from the HLA and MS32 regions that have been previously studied by sperm typing.[3,37]

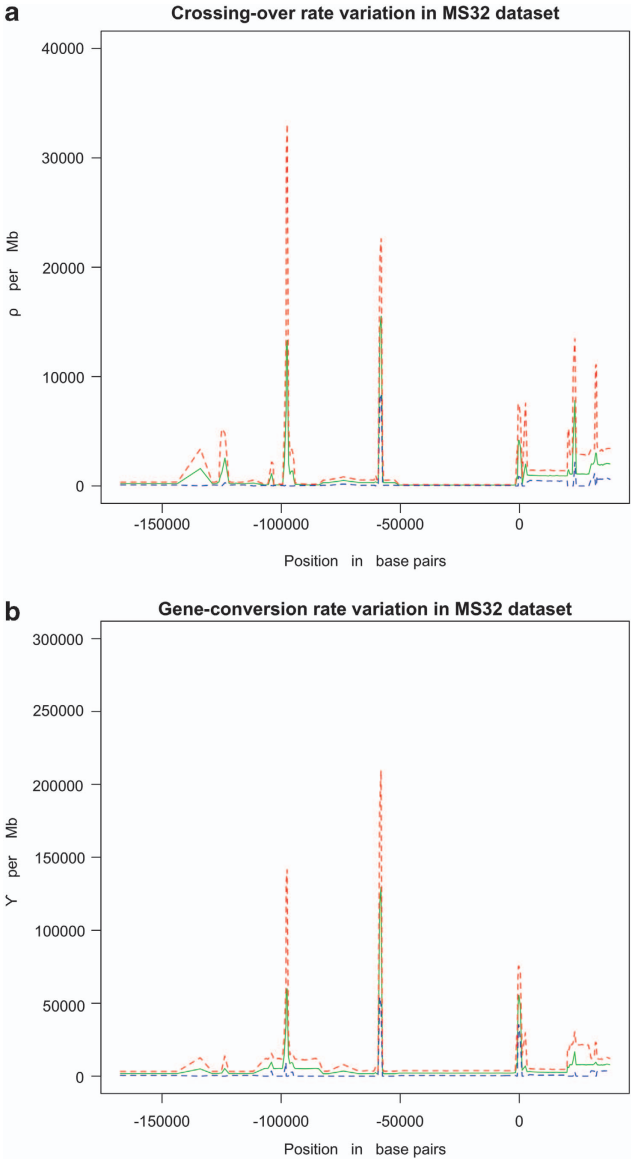

Figure 2 Crossing-over (ρ) and gene-conversion (γ) rate variation in the MS32 data set assuming $m = 125$ bp. The solid green curve shows the posterior means of the estimated rates while the dashed blue and red lines represent the lower and upper bounds of the 95% credible intervals, respectively.

The HLA data set consists of 274 SNPs distributed across a 0.216 Mb region, sampled from 50 unrelated individuals. Six hotspots were revealed in the sperm-typing study[3] and the data have been previously analyzed using a composite likelihood approach.[6] Jeffreys *et al*[37] investigated recombination rates in the MS32 and surrounding regions by both sperm-typing and coalescent analysis of genotypes (recombination rate estimated using the programs LDhat and PHASE). The MS32 data set consists of 206 SNPs sampled from 80 individuals and distributed across a 0.206 Mb region. For both data sets, we considered subsets of 20 consecutive markers and applied our inference algorithm. Figures 1–4 and Supplementary Figures S9 and S10 show the results obtained from this analysis. For Figures 3 and 4, we estimated the parameter f_1 jointly with all other parameters including the mean tract length m. In the analyses corresponding to Supplementary Figures S9 and S10, we fixed $m = 125$ bp (estimates obtained in Jeffreys and May[12]) and inferred all other parameters (ie, ρ, γ, f_1, f_2, etc) jointly. We can see that patterns of crossing-over and gene-conversion rate variation are quite similar in both MS32 and HLA. We can clearly see that peaks of crossing-over are also regions of high population gene-conversion rates. f_1 estimates vary across different hotspots and the posterior distributions include 1 in many cases. The estimates are also dependent on assumptions about tract length (compare Figure 3 *vs* Supplementary Figure S9, and Figure 4 *vs* Supplementary Figure S10).

Analysis of data from chromosome 19 for CEU, YRI and CHB HapMap populations

We applied our inference procedure to 25 genomic windows of 20 markers from human chromosome 19 genotyped in 3 different

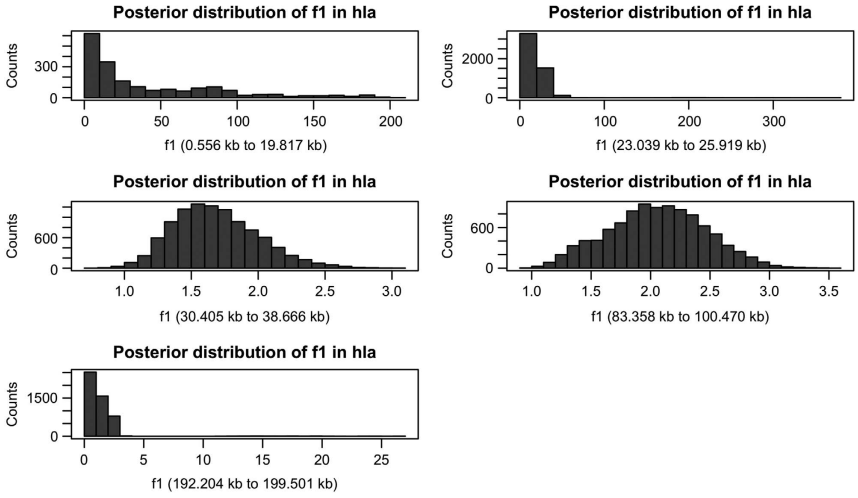

Figure 3 Posterior distribution of f_1 in HLA data set. Mean tract length m was jointly estimated along with all other parameters for 20 marker windows. f_1 denotes the relative rate of gene-conversion to crossing-over within a inferred hotspot.

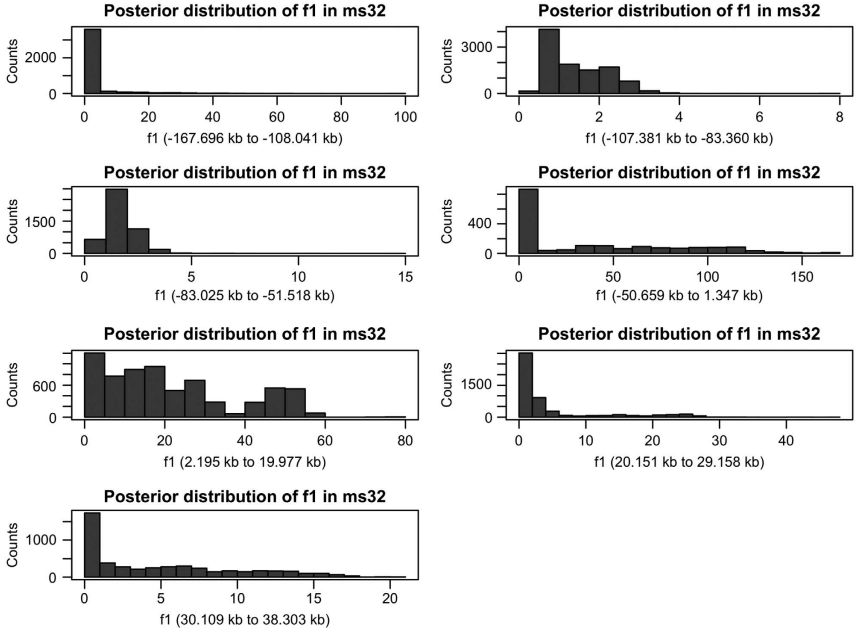

Figure 4 Posterior distribution of f_1 in MS32 data set. Mean tract length m was jointly estimated along with all other parameters for 20 marker windows. f_1 denotes the relative rate of gene-conversion to crossing-over within a inferred hotspot.

Table 1 Estimated values of f_1 in chromosome 19 in HapMap data sets

CEU			*YRI*			*CHB*		
Position (Mb)	f_1	*CI*	*Position (Mb)*	f_1	*CI*	*Position (Mb)*	f_1	*CI*
0.046–0.258	0.5	0–11	0.195–0.267	104.5	0–273	0.195–0.262	2.5	0–103.024
0.262–0.318	12.5	0–163	0.270–0.324	12.5	0–63	0.267–0.323	1.5	0–32.625
0.323–0.367	5.5	0–25	0.327–0.382	8.5	0–22	0.324–0.376	3.5	0–85.648
0.370–0.423	1.5	0–138	0.382–0.445	6.5	0–160	0.379–0.436	1.5	0.013–15.09
0.429–0.480	1.5	0–21	0.446–0.498	23.5	2.6–60.3	0.442–0.495	6.5	0–112.104
0.484–0.526	17.5	0–324	—	—	—	—	—	—
0.527–0.580	0.5	0–146	0.500–0.565	5.5	0–49.7	0.498–0.552	17.5	0–47.734
0.589–0.636	0.5	0–66	0.569–0.636	7.5	0–131.7	0.556–0.626	8.5	0–74.24
0.638–0.702	60.5	1.1–99	0.638–0.708	5.5	0–62.7	0.626–0.694	1.5	0–263.804
0.703–0.757	8.5	0–128	0.708–0.781	37.5	0–127.6	0.694–0.768	17.5	0–51.324
0.763–0.817	0.5	0–57	0.785–0.840	120.5	0–120.9	0.771–0.820	6.5	0–107.277
0.819–0.874	6.5	1.1–13	0.840–0.927	34.5	0–243.2	0.822–0.901	14.5	0.532–44.80
0.878–0.953	0.5	0–181	—	—	—	0.906–0.975	1.5	0–12.008
0.954–1.007	24.5	0–146	0.935–0.992	17.5	0–67.1	0.982–1.025	18.5	0–81.972
1.007–1.058	2.5	0–28	0.993–1.052	3.5	0–40.7	1.026–1.087	1.5	0–8.317
1.061–1.116	0.5	0–99	1.058–1.116	1.5	0–216.1	1.087–1.162	2.5	0–70.347
1.121–1.189	0.5	0.18–58	1.121–1.190	3.5	0–80.3	—	—	—
1.190–1.248	0.5	0–185	1.193–1.255	3.5	0–113.1	1.165–1.228	0.5	0–92.995
1.253–1.292	1.5	0–59	1.256–1.301	35.5	0–107.8	1.234–1.279	14.5	0–56.752
1.299–1.342	0.5	0–140	1.308–1.366	101.5	0.5–190	1.280–1.335	1.5	0–66.052
1.347–1.393	27.5	0–86	1.368–1.411	4.5	0–50.2	1.335–1.395	1.5	0–133.106
1.395–1.446	67.5	0–106	—	—	—	1.395–1.447	7.5	0–292.278
1.447–1.484	0.5	0–107	1.422–1.477	28.5	0–64.7	1.456–1.490	9.5	0–151.625
1.486–1.540	4.5	0–109	1.477–1.537	0.5	0–74.6	1.491–1.564	11.5	0–113.341
1.544–1.590	3.5	0–146	1.540–1.605	27.5	0–142.9	1.566–1.624	20.5	0–244.858
—	—		1.609–1.667	5.5	0–208.9	1.627–1.686	2.5	1.745–107.5
—	—		1.674–1.753	5.5	0–77.1	1.690–1.762	3.5	0–182.81
—	—		1.756–1.805	9.5	0–170.4	—	—	—

Abbreviation: CI, 95% credible intervals.

populations CEU, YRI and CHB as part of the HapMap project. We considered 60, 60 and 45 unrelated individuals, respectively, from these 3 populations in this analysis. We summarize the results obtained to address several questions below.

Is f_1 variable along the genome or among populations?

Table 1 shows the posterior modes of f_1 estimated for 25 genomic windows analyzed for each of the 3 HapMap populations. We find that estimates of f_1 vary substantially across the genome. However, large uncertainties are associated with these estimates and consequently credible intervals are wide. For many windows, there is considerable overlap between 95% credible intervals. We also find that for comparable locations in the genome estimated f_1's are variable across populations (see Table 1).

Is m variable along the genome or among populations?

Table 2 shows the posterior means for the mean conversion tract lengths and 95% credible intervals for 25 genomic windows analyzed for each of the 3 populations. As before, we see that there is great uncertainty in the estimated tract lengths and credible intervals are wide. There is evidence that m varies across genomic regions as well as between populations for comparable regions (also see results in Supplementary Figures S6 and S7).

Is f correlated with crossing-over rates?

Previous studies in Yeast and *Drosophila melanogaster* have suggested that f may be higher in regions with reduced crossing-over rates.[38–40] To explore this issue, we looked at the distribution of the ratio f_2/f_1 in each of the HapMap populations for the 25 windows. Table 3 shows the estimates (and 95% credible intervals) for this joint parameter. If there were a systematic difference in f that depended on the levels of crossing-over, we would expect this variable to assume high values (eg, >1).

Are gene-conversion parameters correlated with local heterozygosity levels?

Previous studies have also suggested that m may vary with levels of heterozygosity in a genomic region.[38–40] To further examine this question, we first calculated the average heterozygosity levels for the regions corresponding to the 20 marker windows of HapMap using data from the 1000 Genomes Project.[41] We estimated heterozygosity levels using all the individuals in the EUR, AFR and ASN groups of this data set. Then, we made scatterplots of the m values estimated in CEU, YRI and CHB of HapMap versus average heterozygosity values as calculated in the three groups (ie, EUR, AFR and ASN, respectively) from the 1000 Genomes Project. These results are shown in Supplementary Figure S11. The Spearman's rank correlation between these two variables was 0.34 (P-value = 0.09677) for CEU, 0.322 (P-value = 0.1163) for YRI, 0.572 (P-value = 0.003288) for CHB

Table 2 Estimated values of mean tract length (*m*) in chromosome 19 in HapMap data sets

CEU			*YRI*			*CHB*		
Position (Mb)	*m*	*CI*	*Position (Mb)*	*m*	*CI*	*Position (Mb)*	*m*	*CI*
0.046–0.258	682.782	376–989	0.195–0.267	76.900	0–427	0.195–0.262	101.133	0–482
0.262–0.318	306.136	0–978	0.270–0.324	532.06	135–928	0.267–0.323	273.065	0–893
0.323–0.367	481.574	0–972	0.327–0.382	839.011	678–999	0.324–0.376	369.153	0–967
0.37–0.423	71.397	0–310	0.382–0.445	758.658	517–999	0.379–0.436	827.947	655–999
0.429–0.48	871.548	743–999	0.446–0.498	714.213	428–999	0.442–0.495	463.396	0–971
0.484–0.526	271.017	0–879	—	—	—	—	—	—
0.527–0.58	732.477	465–999	0.500–0.565	841.073	682–999	0.498–0.552	397.629	0–954
0.589–0.636	831.686	663–999	0.569–0.636	610.861	222–999	0.556–0.626	765.853	531–999
0.638–0.702	763.92	527–999	0.638–0.708	930.193	860–999	0.626–0.694	300.969	0–710
0.703–0.757	859.09	718–999	0.708–0.781	837.227	674–999	0.694–0.768	533.389	249–817
0.763–0.817	873.258	746–999	0.785–0.840	153.213	20–285	0.771–0.820	554.327	109–998
0.819–0.874	660.092	321–999	0.840–0.927	339.305	0–955	0.822–0.901	349.159	0–737
0.878–0.953	198.099	0–800	—	—	—	0.906–0.975	744.217	489–998
0.954–1.007	271.689	0–835	0.935–0.992	432.540	0–885	0.982–1.025	743.638	487–999
1.007–1.058	837.089	674–999	0.993–1.052	901.336	802–999	1.026–1.087	855.013	710–999
1.061–1.116	443.335	0–966	1.058–1.116	53.583	0–264	1.087–1.162	289.055	0–922
1.121–1.189	315.011	49–580	1.121–1.190	647.166	294–999	—	—	—
1.19–1.248	57.6284	0–334	1.193–1.255	740.960	481–999	1.165–1.228	168.509	0–831
1.253–1.292	113.828	0–252	1.256–1.301	814.298	628–999	1.234–1.279	242.181	0–837
1.299–1.342	723.158	446–999	1.308–1.366	168.048	0–791	1.280–1.335	596.641	193–999
1.347–1.393	294.551	0–865	1.368–1.411	678.923	360–997	1.335–1.395	159.323	0–795
1.395–1.446	230.284	0–899	—	—	—	1.395–1.447	652.857	306–999
1.447–1.484	672.412	345–999	1.422–1.477	856.071	712–999	1.456–1.490	711.126	426–995
1.486–1.540	887.049	774–999	1.477–1.537	650.601	301–999	1.491–1.564	94.1493	0–275
1.544–1.59	173.743	0–628	1.540–1.605	133.782	0–648	1.566–1.624	92.608	0–192
—	—	—	1.609–1.667	130.556	0–267	1.627–1.686	895.63	791–999
—	—	—	1.674–1.753	549.284	99–999	1.690–1.762	615.025	230–999
—	—	—	1.756–1.805	865.987	732–999	—	—	—

Abbreviation: CI, 95% credible intervals.

and 0.34 (P-value = 0.002971) considering all the 75 windows. In addition, we also determined the average rate of gene-conversion ($\gamma_{average}$) for all the 25 windows in these 3 populations as inferred by our program and calculated the Spearman's rank correlation with heterozygosity estimates obtained from 1000 genomes project data. The corresponding values are −0.12 (P-value = 0.5663) for CEU, −0.028 (P-value = 0.8961) for YRI, −0.501 (P-value = 0.011) for CHB and −0.163 (P-value = 0.1624) considering all 75 windows. Thus, we do not see any clear relationship between gene-conversion parameters and levels of heterozygosity for these data.

DISCUSSION

We have developed and validated a powerful full-likelihood MCMC method for inferring recombination parameters from population genomic data under a Bayesian framework. This method is based on the approach originally proposed in Wang and Rannala[34,42] and can jointly infer the three fundamental parameters of recombination (ie, crossing-over, gene-conversion and tract length) under variable rate models that include recombination hotspots. The MCMC algorithm has been implemented as an updated version of the software package InferRho (see Wang and Rannala[42]). In addition to the moves already described in Padhukasahasram and Rannala[29], the current program uses a new recombination model and implements many novel proposal schemes for updating the locations of recombination hotspots and the gene-conversion parameters. The method is also more general than before and allows the ratio of gene-conversion to crossing-over (f) to assume two possible values depending on whether we are within a hotspot or outside. We applied the new method to data from MHC and MS32 regions of the human genome, which are known to harbor many recombination hotspots. We found that InferRho identifies most of the major hotspots already known in these data sets (as in Wang and Rannala[42]). In addition, we also find that regions with elevated crossing-over rates are regions of higher gene-conversion rates as well which is consistent with the experimental findings of Jeffreys and May.[12]

To explore variability in gene-conversion parameters, we also applied our method to 25 windows of 20 markers each on chromosome 19 from 3 human populations CEU, YRI and CHB of HapMap. Owing to computational constraints imposed by this more comprehensive full-likelihood method, we did not attempt to carry out an extensive analysis over the entire genome. We can draw several conclusions as a result of this analysis. We find that the uncertainty in the estimates of f_1 are high for data sets of this size. Indeed, for many windows the 95% credible intervals overlap with each other and most of these intervals include 1.0. The great uncertainty in f_1 reflects the limited information about conversion in population genomic data combined with confounding when attempting to infer gene-conversion rates and tract lengths jointly. We notice considerable differences in the estimated values of the parameters f_1 and m across different windows as well as across populations for comparable positions in the genome. As recombination patterns and rates can vary between individuals and also evolve with time, the differences between

Table 3 Estimated values of f_2/f_1 in chromosome 19 in HapMap data sets

CEU			*YRI*			*CHB*		
Position (Mb)	f_2/f_1	*CI*	*Position (Mb)*	f_2/f_1	*CI*	*Position (Mb)*	f_2/f_1	*CI*
0.046–0.258	40.558	0–196.1	0.195–0.267	0.872	0–1.9	0.195–0.262	0.967	0–1.9
0.262–0.318	1.370	0–3.0	0.270–0.324	0.909	0–2.7	0.267–0.323	5.764	0–17.0
0.323–0.367	5.029	0.04–10	0.327–0.382	7.140	0–14	0.324–0.376	1.672	0–4.8
0.37–0.423	11.674	0–45.2	0.382–0.445	3.798	0–7.7	0.379–0.436	27.95	0–89.5
0.429–0.48	1.889	0–4.1	0.446–0.498	1.209	0–2.7	0.442–0.495	25.49	0–52.2
0.484–0.526	1.656	0–3.9	—	—	—	—	—	—
0.527–0.58	97.243	0–206	0.500–0.565	1.563	0–4.7	0.498–0.552	2.266	0–4.9
0.589–0.636	137.804	0–305	0.569–0.636	0.358	0–0.8	0.556–0.626	0.833	0–3.7
0.638–0.702	0.385	0–0.77	0.638–0.708	1.372	0–3.6	0.626–0.694	1.804	0–3.6
0.703–0.757	0.531	0–1.1	0.708–0.781	0.883	0.05–1.7	0.694–0.768	0.493	0–1.1
0.763–0.817	3.721	0.06–7.3	0.785–0.840	2.963	0–7.3	0.771–0.820	1.897	0–3.8
0.819–0.874	1.949	0–4.4	0.840–0.927	0.712	0–1.6	0.822–0.901	1.631	0–4.6
0.878–0.953	35.105	0–71.6	—	—	—	0.906–0.975	9.177	0–24.3
0.954–1.007	1.026	0–2.9	0.935–0.992	1.407	0–3.7	0.982–1.025	3.451	0–6.9
1.007–1.058	1.344	0–4.2	0.993–1.052	3.664	0–10.2	1.026–1.087	5.293	0.17–10.4
1.061–1.116	21.588	0–71.8	1.058–1.116	2.108	0–6.5	1.087–1.162	5.404	0–13.7
1.121–1.189	8.069	0–22.4	1.121–1.190	1.429	0–3.7	—	—	—
1.19–1.248	20.158	0–40.6	1.193–1.255	0.876	0–2.7	1.165–1.228	18.56	0–37.2
1.253–1.292	8.001	0–26.8	1.256–1.301	1.860	0–4.5	1.234–1.279	3.095	0–6.4
1.299–1.342	51.474	0–130.5	1.308–1.366	1.675	0.04–3.3	1.280–1.335	7.252	0–15.1
1.347–1.393	11.324	0–23.0	1.368–1.411	3.866	0–7.9	1.335–1.395	7.226	0–16.3
1.395–1.446	2.009	0–4.2	—	—	—	1.395–1.447	4.204	0–8.4
1.447–1.484	12.910	0–26.1	1.422–1.477	4.912	0–10.6	1.456–1.490	0.515	0–1.3
1.486–1.540	0.447	0–2.1	1.477–1.537	84.21	0–170.4	1.491–1.564	3.670	0–7.9
1.544–1.59	3.333	0–6.9	1.540–1.605	1.163	0–2.7	1.566–1.624	0.560	0–2.2
—	—	—	1.609–1.667	3.656	0–10.0	1.627–1.686	3.342	0–7.2
—	—	—	1.674–1.753	2.013	0–5.0	1.690–1.762	3.582	0–10.8
—	—	—	1.756–1.805	0.708	0–2.5	—	—	—

Abbreviation: CI, 95% credible intervals.

populations might be reflecting the average of the differences between the individuals in different ancestral groups. However, given the high degree of uncertainty in the relative rate estimates, it is difficult to draw strong conclusions.

Previous studies in Yeast and *Drosophila melanogaster* have suggested that f may be higher in regions with reduced crossing-over rates.[38–40] It has also been speculated that the rate of gene-conversion may go up as the rate of crossing-over goes down, that is, there is shunting of incipient cross-overs toward conversions.[37] In our analysis, we found that f_2/f_1 estimates tend to be >1 although credible intervals are wide and include 1 for most windows. Furthermore, it has also been suggested that mean conversion tract lengths may vary with levels of heterozygosity in a region. If a heterozygosity-dependent form of gene-conversion were to be operating,[39,40] regions with reduced variability are expected to have higher gene-conversion rates and longer tract lengths on average. Our results using the three populations of HapMap did not indicate any negative correlations between heterozygosity estimates and mean tract length estimates. In addition, we do not see any clear relationship between conversion rates and heterozygosity in our data. Future work, with extensive estimates using denser SNP data (eg, from the 1000 genomes project) and much larger number of windows across the entire genome is likely to give us a better picture of the relationship between these variables and resolve these various empirical questions about gene-conversion more definitively.

In Supplementary Figure S8 and Supplementary Table S2, we show the effect of increasing SNP density on the posterior distribution of recombination parameters for simulated data sets. We can see that for the same set of recombination parameters, increasing θ by threefold (ie, from 10.0 to 30.0 for the 10 kb sequences), generally decreases the variance of the posterior distributions of f_1 and m. This implies that we expect to obtain tighter credible intervals in denser polymorphism data. With denser SNP data such as from 1000 genomes, we would apply InferRho to many more (100s of) shorter windows (5–20 kb) spread across the entire genome and omit variants with minor allele frequency $<10\%$ from all analyses. Doing this kind of large-scale analysis will also necessitate modifications to the program as well as access to greater computing resources, so that many more jobs can be run in parallel. We are planning to implement several optimizations to boost the speed of the new version of MCMC algorithm. For example, InferRho currently considers crossing-over and gene-conversion events that occur in regions that have reached their marginal MRCAs. However, a subset of such events does not affect the sample configuration at all and can be omitted. In addition, ARGs can be based on recently developed approximations that closely mimic the standard coalescent process.[43,44] Both these steps will help to make ARGs more compact on average. In future versions of InferRho, we also plan to use pre-computed look-up tables, when calculating the rates of informative recombination events,[29] which are used repeatedly in acceptance probability calculations and encode

haplotypes in the graph nodes in bitwise notation rather than as character arrays. A combination of such strategies can help to further reduce the computational burden of full-likelihood inference.

WEB RESOURCES

InferRho program and binaries are freely available from http://www.rannala.org.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This research project was supported by NHGRI grant HG01988 to Bruce Rannala.

1 Fullerton SM, Harding RM, Boyce AJ, Clegg JB: Molecular and population genetic analysis of allelic sequence diversity at the human beta-globin locus. *Proc Natl Acad Sci USA* 1994; **91**: 1805–1809.
2 Dunham I, Shimizu N, Roe BA *et al*: The DNA sequence of human chromosome 22. *Nature* 1999; **402**: 489–495.
3 Jeffreys AJ, Kauppi L, Neumann R: Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 2001; **29**: 217–222.
4 Innan H, Padhukasahasram B, Nordborg M: The pattern of polymorphism on human chromosome 21. *Genome Res* 2003; **13**: 1158–1168.
5 Crawford DC, Bhangale T, Li N *et al*: Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat Genet* 2004; **36**: 700–706.
6 McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P: The fine-scale structure of recombination rate variation in the human genome. *Science* 2004; **304**: 581–584.
7 International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
8 Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: A fine-scale map of recombination rates and hotspots across the human genome. *Science* 2005; **310**: 321–324.
9 Tiemann-Boege I, Calabrese P, Cochran DM, Sokol R, Arnheim N: High resolution recombination patterns in a region of human chromosome 21 measured by sperm typing. *PLoS Genet* 2006; **2**: e70.
10 Zangenberg G, Huang MM, Arnheim N, Erlich H: New HLA-DPB1 alleles generated by interallelic gene conversion detected by analysis of sperm. *Nat Genet* 1995; **10**: 407–414.
11 Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A: Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am J Hum Genet* 2001; **69**: 831–843.
12 Jeffreys AJ, May CA: Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet* 2004; **36**: 151–156.
13 Padhukasahasram B, Marjoram P, Nordborg M: Estimating the rate of gene conversion on human chromosome 21. *Am J Hum Genet* 2004; **75**: 386–397.
14 Padhukasahasram B, Wall JD, Marjoram P, Nordborg M: Estimating recombination rates from single-nucleotide polymorphisms using summary statistics. *Genetics* 2006; **174**: 1517–1528.
15 Gay J, Myers S, McVean G: Estimating meiotic gene conversion rates from population genetic data. *Genetics* 2007; **177**: 881–894.
16 Fogel S, Mortimer RK, Lusnak K: *Meiotic Gene Conversion in Yeast: Molecular and Experimental Perspectives*. New York, NY, USA: Springer-Verlag, 1983.
17 Hilliker AJ, Clark SH, Chovnick A: The effect of DNA sequence polymorphisms on intragenic recombination in the rosy locus of *Drosophila melanogaster*. *Genetics* 1991; **129**: 779–781.
18 Hilliker AJ, Harauz G, Reaume AG, Gray M, Clark SH, Chovnick A: Meiotic gene conversion tract length distribution within the rosy locus of *Drosophila melanogaster*. *Genetics* 1994; **137**: 1019–1026.
19 Paques F, Haber JE: Multiple pathways of recombination induced by doublestrand breaks in *Saccharomyces cerevisiae*. *Microbiol Mol Biol Rev* 1999; **63**: 349–404.
20 Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM: High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 2008; **454**: 479–485.
21 Wall JD: Close look at gene conversion hotspots. *Nat Genet* 2004; **36**: 114–115.
22 Ptak SE, Voelpel K, Przeworski M: Insights into recombination from patterns of linkage disequilibrium in humans. *Genetics* 2004; **167**: 387–397.
23 Wall JD: Estimating recombination rates using three-site likelihoods. *Genetics* 2004; **167**: 1461–1473.
24 Hudson RR: Two-locus sampling distributions and their application. *Genetics* 2001; **159**: 1805–1817.
25 McVean G, Awadalla P, Fearnhead P: A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 2002; **160**: 1231–1241.
26 Li N, Stephens M: Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 2003; **165**: 2213–2233.
27 Hellenthal G: *Exploring Rates and Patterns of Variability in Gene Conversion and Crossover in the Human Genome*. PhD thesis University of Washington: Seattle, 2006.
28 Yin J, Jordan MI, Song YS: Joint estimation of gene conversion rates and mean conversion tract lengths from population SNP data. *Bioinformatics* 2009; **25**: i231–i239.
29 Padhukasahasram B, Rannala B: Bayesian population genomic inference of crossing over and gene conversion. *Genetics* 2011; **189**: 607–619.
30 Kingman JFC: The coalescent. *Stochast Proc Appl* 1982; **13**: 235–248.
31 Hudson RR: Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol* 1983; **23**: 183–201.
32 Griffiths RC, Marjoram P: Ancestral inference from samples of DNA sequences with recombination. *J Comput Biol* 1996; **3**: 479–502.
33 Wiuf C, Hein J: The coalescent with gene conversion. *Genetics* 2000; **155**: 451–462.
34 Wang Y, Rannala B: Population genomic inference of recombination rates and hotspots. *Proc Natl Acad Sci USA* 2009; **106**: 6215–6219.
35 Geyer CJ: Markov chain monte carlo maximum likelihood; in Keramides E (eds). *Computing Science and Statistics*. Fairfax Station, VA, USA: Interface Foundation, 1991; pp 156–163.
36 Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F: Parallel metropolis coupled MCMC for Bayesian phylogenetic inference. *Bioinformatics* 2004; **20**: 407–415.
37 Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P: Human recombination hotspots hidden in regions of strong marker association. *Nat Genet* 2005; **37**: 601–606.
38 Borts RH, Haber JE: Meiotic recombination in yeast: alteration by multiple heterozygosities. *Science* 1987; **237**: 1459–1465.
39 Langley CH, Lazzaro BP, Phillips W, Heikkinen E, Braverman JM: Linkage disequilibria and the site frequency spectra in the su(s) and su(w^a) regions of the *Drosophila melanogaster* X chromosome. *Genetics* 2000; **156**: 1837–1852.
40 Andolfatto P, Wall JD: Linkage disequilibrium patterns across a recombination gradient in african *Drosophila melanogaster*. *Genetics* 2003; **165**: 1289–1305.
41 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. *Nature* 2010; **467**: 1061–1073.
42 Wang Y, Rannala B: Bayesian inference of fine-scale recombination rates using population genomic data. *Philos Trans R Soc Lond B Biol Sci* 2008; **363**: 3921–3930.
43 McVean GA, Cardin NJ: Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci* 2005; **360**: 1387–1393.
44 Chen GK, Marjoram P, Wall JG: Fast and flexible simulation of DNA sequence data. *Genome Res* 2009; **19**: 136–142.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)