Research article

Simple models of genomic variation in human SNP density

Raazesh Sainudiin*[1,2], Andrew G Clark[3] and Richard T Durrett[2]

Address: [1]Department of Statistics, University of Oxford, Oxford, OX1 3TG, UK, [2]Department of Mathematics, Cornell University, Ithaca, New York 14853, USA and [3]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA

Email: Raazesh Sainudiin* - sainudii@stats.ox.ac.uk; Andrew G Clark - ac347@cornell.edu; Richard T Durrett - rtd1@cornell.edu

* Corresponding author

Abstract

Background: Descriptive hierarchical Poisson models and population-genetic coalescent mixture models are used to describe the observed variation in single-nucleotide polymorphism (SNP) density from samples of size two across the human genome.

Results: Using empirical estimates of recombination rate across the human genome and the observed SNP density distribution, we produce a maximum likelihood estimate of the genomic heterogeneity in the scaled mutation rate θ. Such models produce significantly better fits to the observed SNP density distribution than those that ignore the empirically observed recombinational heterogeneities.

Conclusion: Accounting for mutational and recombinational heterogeneities can allow for empirically sound null distributions in genome scans for "outliers", when the alternative hypotheses include fundamentally historical and unobserved phenomena.

Background

Understanding the population-genetic forces behind the observed variation among human genome sequences is vital to deciphering the genetic causes of phenotypic variation among humans. The phenomena that influence the density of human SNPs include (1) variation-introducing events that are empirically observable, such as, point-mutations, recombinations, and activities of various transposable elements that may result from the counteraction of various DNA damage and repair pathways [[1], for e.g.], as well as (2) genealogy-affecting events that are historical and generally unobserved, such as population dynamics, population structure, and natural selection. A biological understanding of the observed genomic variation in SNP density, by means of explicit population-genetic models of coalescence in the presence of recombination and mutation, must incorporate any interplay

among the heterogeneities in the above phenomena. Here we strive for an empirically sound understanding of the observed human SNP density, as determined by a genome-wide alignment of two different consensus sequences, by accounting for the empirically observable mutational and recombinational heterogeneities under the simplest model of population history (selectively-neutral, constant-sized, random-mating). The two sequences are the NCBI human genome sequence and the sequence produced by Celera Genomics [2]. Our SNP density data were obtained from first aligning the Celera consensus sequence to the NCBI assembly and then counting the number of SNPs in bins of 100 kb (100,000 base pairs), as was done in section 6 of the above study [2]. Next, we build simple models for the distribution of SNP density from random samples of size 2 from a locus that is 100 kb in length. Our objective is to explain as much of this sim-

ple measure of diversity as possible, under empirically sound null hypotheses that include coarse-grained, genome-wide measurements of recombinational variation.

Results

Two approaches toward modeling are taken. The first approach is descriptive and employs hierarchical Poisson models to obtain better fits than the homogeneous Poisson distribution used earlier [2]. Insights gained from the first approach inform the second approach. The second approach is non-descriptive and population-genetic with biologically interpretable parameters. It employs mixtures of SNP densities simulated under the coalescent with different mutation and recombination rates to obtain a better fit to the observed SNP density distribution. This approach introduces heterogeneity into the coalescent-based simulation of SNP density that was shown to produce a poor fit under the assumptions of genome-wide homogeneity and equality of mutation and recombination rates [2]. The simple closed-form expressions used in the paper are elementary results in coalescent theory [3,4].
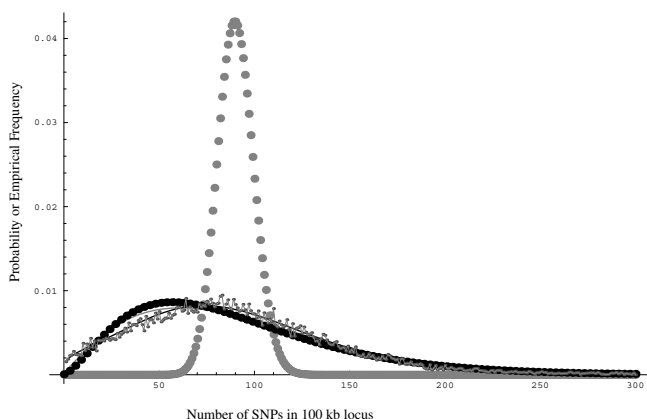
Descriptive Hierarchical Poisson Models

Let Λ and T be the parameters in the mass function of a Poisson distribution given by $\Pr(X = x|\Lambda T) = e^{-\Lambda T}(\Lambda T)^x/x!$. The random variables Λ and T are generally *proxies* for relative mutation rate and the sum of branch lengths of the coalescent trees for all the non-recombining segment(s) of the 100 kb locus, respectively. In other words, T is a proxy for the sum of the branch lengths of the ancestral recombination graph (ARG-size) of our sample of size 2 at a locus that is 100 kb long. The random variable X represents the count of SNPs in contiguous 100 kb intervals from an alignment of two human genomes. In this hierarchical scheme, heterogeneities are modeled by the Gamma and Beta probability density functions (*PDF*s), $G(\gamma_1, \gamma_2)$ and $B(\beta_1, \beta_2)$, respectively, as described in Methods. We chose the Gamma distribution $G(\gamma_1, \gamma_2)$ to model T for the following reasons. When there is no recombination, the depth of the coalescent tree of two samples is exponentially distributed with rate parameter 1, i.e., $G(1,1)$. And when there are n sites with free recombination in between them, the sum of the n independent and exponentially distributed depths is $G(n, 1)$. Thus, T is only a mathematically convenient proxy for the ARG-size of our sample of size 2, since $T \sim G(\gamma_1, \gamma_2)$ does not explicitly capture the distribution of ARG-size for intermediate levels of intra-locus recombination among sites at our locus. We use the relatively flexible Beta family on $[0, 1]$ to model, Λ which is a proxy for relative mutation rate. The Poisson distribution for SNP density follows from the assumption of the infinitely-many-sites mutational model under selective neutrality, where mutations hit a site at most once according to the product of the total

length of the site-specific coalescent tree and the site-specific relative mutation rate. Therefore, such hierarchical Poisson models are merely descriptive, as they are built via mathematically convenient Beta and Gamma distributed random variables Λ and T that act as proxies for the relative mutation rate and the ARG-size, respectively.

The likelihood function for each of the following hierarchical Poisson models was maximized with the Newton's method from several random initial conditions. We use the Akaike information criterion (*AIC*) [5] to make model comparisons. For a given model $AIC = -2 \log(ML) + 2K$, where ML is the maximum likelihood value and K is the number of parameters in the model. In the hierarchical Poisson model A, we allow $T \sim G(\gamma_1, \gamma_2)$, while Λ is fixed at 1. The fit to the data (Figure 1) improved in comparison to the homogeneous Poisson fit which completely ignores the underlying ancestral recombination process. Thus, when the Gamma distribution is used to approximate the distribution of the sum of all branch lengths of the ancestral recombination graph (ARG-size) of a locus, the observed variance is better explained. Model A is mutationally homogeneous as Λ, the proxy for mutation rate, is fixed. In order to allow variation, a hierarchical Poisson model A' that restricts T to a constant parameter λ while allowing Λ to be Beta distributed ($\Lambda \sim B(\beta_1, \beta_2)$) was fit to the data. The fit was significantly better than that of model A. Thus, modeling heterogeneity in mutation rates, via the Beta distributed proxy Λ, across the different 100 kb loci gives better fits to the SNP density distribution. When we allowed both Λ to be Beta distributed and T to be Gamma distributed, we get the hierarchical Poisson model B. As shown in Figure 1, the fit is significantly better to the observed data when heterogeneities in both mutation and recombination are approximately accounted for through the proxies in model B. The results of the maximum likelihood (ML) analysis of these four Poisson models are summarized in Table 1. The first and second moments ($\hat{\mu}$, and $\widehat{\sigma^2}$) under the maximum likelihood estimates are also shown for each model in the Table. Note that the means are almost the same but the variances vary considerably. If one wants a data-descriptive fit to the SNP density distribution, then Model B is a good candidate. With the arrival of more refined data (with counts in low-density regions as discussed later) one could consider further generalizations of such hierarchical Poisson models along the zero-inflated class [6], for instance, to obtain better descriptive fits. Unfortunately, the best-fitted parameters of such descriptive models lack any explicit biological

Figure 1
Fits of the homogeneous Poisson model (large gray dots), hierarchical Poisson model A (black dots) with $T \sim G(\gamma_1, \gamma_2)$ hierarchical Poisson model A' (gray line) with $T = \lambda$ and $\Lambda \sim B(\beta_1, \beta_2)$, and hierarchical Poisson model B (black line) with $T \sim G(\beta_1, \beta_2)$ and $\Lambda \sim B(\beta_1, \beta_2)$ to the observed SNP density distribution (joined gray dots).

interpretability, in terms of standard population-genetic models of reproduction. Guided by insights from these descriptive hierarchical Poisson models, we analyze the simplest population-genetic model of the neutral coalescent with an explicit accounting for heterogeneities in both mutation and recombination rates. We use a simu-

lated maximum likelihood framework [7] for parameter estimation.

Population-Genetic Coalescent Mixture Models

A panmictic, Wright-Fisher, neutral coalescent model with a constant effective population size of 10,000 diploid individuals was assumed to simulate the distribution of the number of segregating sites at a locus of 100 kb evolving under an infinitely-many-sites mutation model using the C program ms [8]. The scaled product of the effective population size (N_e) and the mutation rate per locus per generation (μ) is denoted by $\theta = 4N_e\mu$. The recombination rate r is the probability of cross-over per generation between the ends of the locus being simulated and its scaled product with N_e is denoted by $\rho = 4N_e r$.

In the absence of recombination and with constant mutation rates, the distribution of SNPs is known to have an explicit form. The coalescent tree is identical for every nucleotide site in the locus in any given realization of the coalescent process of two samples. Since the rescaled time to the coalescent event and the mutation event are exponentially distributed with rates 1 and θ, respectively, the probability of a mutation event before the coalescent event is $\theta/(1 + \theta)$. Thus, the probability of observing x mutations at our locus before the coalescent event is $(\theta/(1 + \theta))^x 1/(1 + \theta)$. In other words, the probability of observing x SNPs at a locus when $r = 0$ is geometrically distributed with parameter $1/(1 + \theta)$.

Table 1: Maximum likelihood analysis and comparison of Poisson models

Model	T	Λ	Maximum Likelihood Estimates		ML	AIC*
Poisson	λ	1	$\hat{\lambda}$ = 90.2, $\hat{\mu}$ = 90.2, $\widehat{\sigma^2}$ = 90.2		-616497	861964
A	$G(\gamma_1, \gamma_2)$	1	$\widehat{\gamma_1}$ = 2.7, $\widehat{\gamma_2}$ = 32.9, $\hat{\mu}$ = 90.2,		-186348	1670
			$\widehat{\sigma^2}$ = 3049.7			
A'	λ	$B(\beta_1, \beta_2)$	$\hat{\lambda}$ = 387.6, $\widehat{\beta_1}$ = 2.17, $\widehat{\beta_2}$ = 7.16,		-185869	714
			$\hat{\mu}$ = 90.1, $\widehat{\sigma^2}$ = 2683.9			
B	$G(\gamma_1, \gamma_2)$	$B(\beta_1, \beta_2)$	$\widehat{\gamma_1}$ = 6.4, $\widehat{\gamma_2}$ = 19.0, $\widehat{\beta_1}$ = 1.3,		-185511	0
			$\widehat{\beta_2}$ = 0.46, $\hat{\mu}$ = 90.1, $\widehat{\sigma^2}$ = 2538.2			

The last two columns give the maximum log likelihood values and the translated Akaike information criterion, *AIC* = AIC* - 371034.

It is also known that as the recombination rate at our locus approaches infinity, the distribution of SNPs approaches a Poisson distribution with parameter θ. This can be seen from the following argument. High levels of recombination assures that the coalescent tree at each site is independent of those at other sites. Thus, for a locus with n sites, the probability of observing x SNPs is

$\binom{n}{x}\left(\frac{\theta}{n}\bigg/\left(1+\frac{\theta}{n}\right)\right)^{x}\left(1\bigg/\left(1+\frac{\theta}{n}\right)\right)^{n-x}$. For large loci, this

binomial mass function is known to approximate $e^{-\theta}\theta^{x}/x!$, the Poisson mass function, as $n \rightarrow \infty$ and

$n\frac{\theta}{n}\bigg/\left(1+\frac{\theta}{n}\right) \rightarrow \theta$.

However, when the recombination rate is some intermediate value between the above two extremes no explicit forms are known for the SNP density. We use empirical estimates of the SNP density from a large number of simulations (typically 100,000). Figure 2 shows how the distribution of SNP density under our assumptions morphs from the geometric distribution (black dots) towards the Poisson distribution (grey dots) as the scaled recombination rate ρ increases from 0 to 1000 in decreasing shades of grey. This behavior is identical for any fixed value of θ except for a scale change.

The empirical estimates of the sex-averaged human recombination rates in 1 Mbp intervals based on Genethon [9], Marshfield [10] and deCODE [11] maps were downloaded from [12]. We intrapolated to obtain the estimates over 100 kb segments by assuming rate constancy over the 10 consecutive 100 kb segments that constitute the 1 Mbp segment for which an empirical estimate of the recombination rate were available. The empirical distribution of the sex-averaged human recombination rate in 100 kb intervals, based on Genethon map, as shown in Figure 3, is denoted by $\hat{R}$. The strategy described in Methods was used to obtain a simulation-based empirical estimate of the SNP density distribution for each scaled mutation rate

$$\theta_i \in \Theta = \{\theta_1, \cup, \theta_{304}\}$$
$$= \{0.001, 0.01, 0.1, 0.5, 1, 2, 3, 4, 5, \cup, 298, 299, 300\},$$

when the recombination rate was assumed to be distributed according to $\hat{R}$. We denote this simulation-based estimate of the SNP density distribution for each $\theta_i \in \Theta$ by $\hat{S}_{\hat{R},\theta_i}$. Note that $\hat{S}_{\hat{R},\theta_i} \rightarrow S_{\hat{R},\theta_i}$, the true SNP density distri-

bution, as the number of replicates (N) used to estimate it grows large. In practice, N was set at 100,000. A discretized and rescaled Beta density with parameters α and β was used to find the mixing weights for each $\theta_i \in \Theta$. Thus, for every ordered pair (α, β), the shape of the Beta density specified the mixing weights, as follows:

$$w_{\theta_i}^{(\alpha,\beta)} = \int_{\frac{i-1}{|\Theta|}}^{\frac{i}{|\Theta|}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \lambda^{\alpha-1}(1-\lambda)^{\beta-1} d\lambda$$

where, $|\Theta| = 304$ is the cardinality or size of the set Θ. Such (α, β)-specified $w_{\theta_i}^{(\alpha,\beta)}$'s were used to weigh the corresponding $\hat{S}_{\hat{R},\theta_i}$'s in order to obtain a finite mixture of the form $\sum_{\theta_i \in \Theta} w_{\theta_i}^{(\alpha,\beta)} S_{R,\theta_i}$. A simulated likelihood function of α and β was thus constructed for the given SNP data $X = (x_0, \in, x_n)$, as follows,

$$\prod_{j=0}^{n} \sum_{\theta_i \in \Theta} w_{\theta_i}^{(\alpha,\beta)} S_{R,\theta_i}(x_j)$$
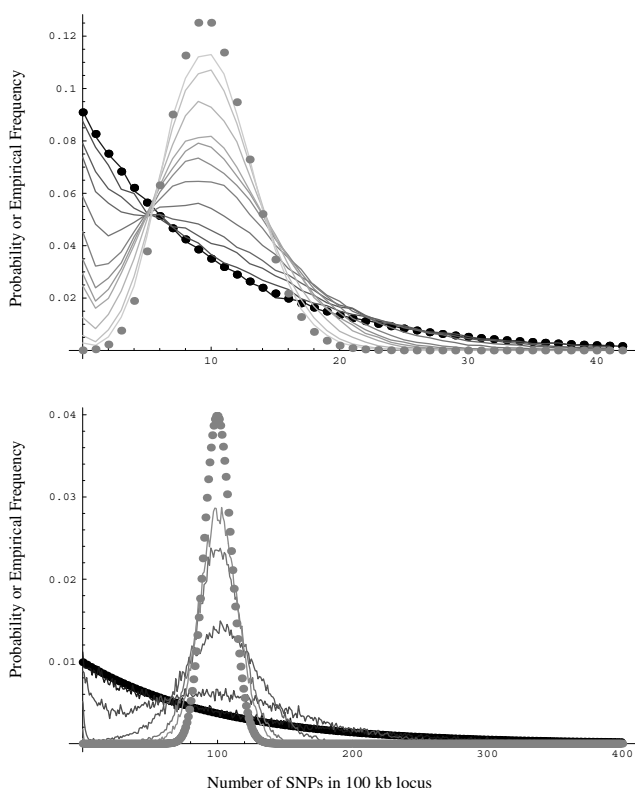
We used the Newton's method to find the maximum simulated likelihood (*MSL*) estimates $\hat{\alpha} = 6.7$ and $\hat{\beta} = 14.9$ (*MSL* = -185555). We also did a least-squares fit of the observed to the predicted densities and found comparable estimates. Empirical estimates of the sex-averaged recombination rates from deCODE, and Marshfield maps were also used in a similar analysis. Comparable estimates were obtained under a reasonably good fit (*MSL* = -185558) with the deCODE map whose empirical CDF resembles that of the Genethon Map. However, an analysis with the Marshfield map yielded a poorer fit (*MSL* = -186007). Figure 4 summarizes the fits to the observed SNP data while Figure 3 shows the marginal density of ρ from the Genethon map and the marginal density of θ under the maximum simulated likelihood estimates ($\hat{\alpha} = 6.7$, $\hat{\beta} = 14.9$) with mean, variance, and standard deviation given by 90.7, 876.1, and 29.6, respectively. Among the three coarse-scaled maps of the *empirical* estimates of the sex-averaged human recombination rates, the Genethon map gave the best fit to our observed SNP density distribution data.

Discussion
Another study [13] claimed to have achieved a good fit to single reads with 0, 1, 2, 3, or 4 SNPs, by accounting for
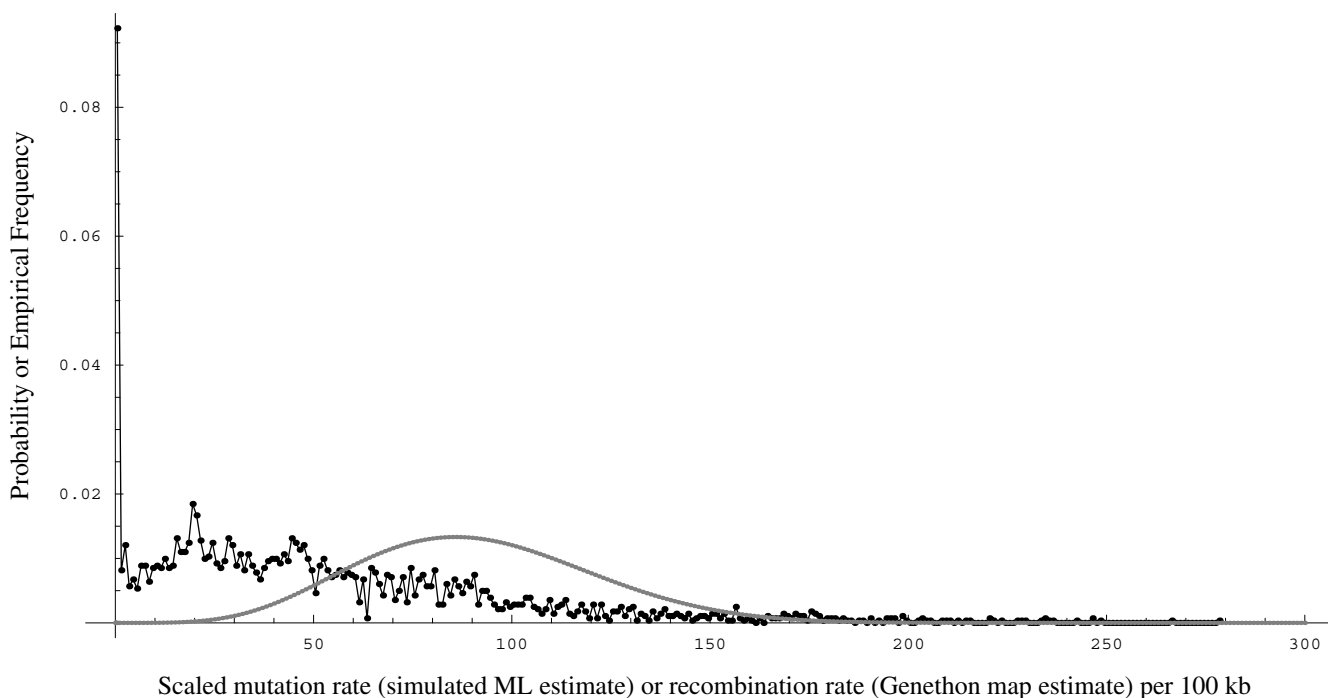
mutational heterogeneity and genealogical variability in a different manner. They partitioned the genome into 200 kb bins, and selected a single read from each bin. They calculated the observed GC content of the bin, and from a regression of GC content on nucleotide diversity across the whole genome, they calculated an expected diversity given the local GC content of each bin induced by the exponentially distributed coalescent time for samples of size 2 in the absence of recombination. However, when the full bin size of 100 kb were used [2], the SNP count ranged to more than 100 per bin. Because many neighboring reads have shared genealogies, the magnitude of variability from bin to bin is much greater, and the power to detect this heterogeneity is far greater. Thus, the latter study [2] found that the coalescent in the presence of recombination fit the observed SNP density better than the coalescent without recombination. The model employed in the former study [13] fits without recombination only because the power is so low to detect a departure and because there are correspondingly fewer recombination events expected within single reads vs. 100

kb bins. Using the data of SNP counts in 100 kb bins in this study, we find that the coalescent with heterogeneities in recombination as well as mutation gives substantially better fits than the coalescent with a constant rate of recombination and mutation. We have shown that by invoking heterogeneities in mutation and recombination rates, one can better explain the observed variation in SNP density across two randomly sampled 100 kb segments of human chromosomes. Descriptive fits by means of hierarchical Poisson models, as well as population-genetic fits by means of coalescent mixture models, significantly improved when heterogeneities in recombination as well as mutation rates were accounted for. The coalescent mixture model does not completely fit the data in the most interesting region, namely, the segments with the least SNP density. This is partly due to the filtering strategy used to obtain the data. Since there were considerable gaps in the alignments for several bins, there was an overestimation of bins with 0 SNPs. Thus, these bins were ignored from the analysis. Were low SNP counts from such currently ignored bins made available from a high-resolution alignment, a similar analysis would reveal the poorer fits of the descriptive hierarchical Poisson models employed here, unless they are further generalized to allow for a larger mass at 0 through the zero-inflated class [6], for instance. If one's objective is to produce a descriptive fit to our observed SNP density distribution, then the hierarchical model B is clearly preferable to all the models considered in this study due to its strikingly high likelihood value. However, if one wanted a population-genetic model with biologically interpretable parameters to fit the same data, then the best fitted coalescent mixture model with the Genethon recombination map is preferable.

It is important to bear in mind that the distribution of T will be affected not only by recombination rate but also by population structure and demography. Likewise, the distribution of Λ and T will be affected by the complex interaction between various DNA damage and repair pathways that ultimately lead to various types of mutational and recombinational events [[1], for e.g.]. Moreover, the action of selection will simultaneously affect both the distribution of T and Λ about the selected site(s). However, since only a small percentage of the genome is expected to be affected by recent selective sweeps, the overall SNP density distribution should not be significantly affected by such selective events. Thus, our MSL estimate of the genomic variation in θ, based on the Genethon map, is under the standard neutral coalescent that allows for recombinational and mutational rate heterogeneity across the genome. The true genomic variation in θ can also be affected by several other confounded historical factors including selection, population structure, and demography, besides genomic variation in mutation rate. All these confounded historical factors can be seen as alternative
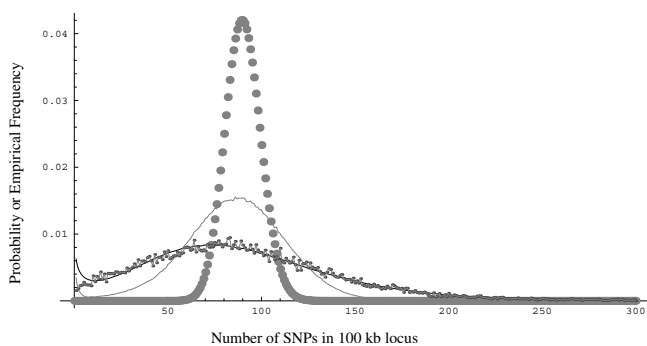

Figure 2
The distribution of SNP density in 100 kb morphs from the geometric distribution (black dots) towards the Poisson distribution (gray dots) as the scaled recombination rate ρ increases from 0 to 1000 in decreasing shades of gray for θ = 10 (top) and θ = 100 (bottom).

Figure 3
The distribution of the empirical estimates of the sex-averaged recombination rate in 100 kb segments of the human genome from the Genethon map (joined black dots) and $w_{\theta_i}^{(6.7,14.9)}$, the maximum simulated likelihood estimate of the weights on $\theta_i \in$ Θ (gray line) for the coalescent mixture model.

hypotheses to the null hypothesis of our coalescent mixture model for the SNP density distribution, i.e., the standard neutral coalescent with genomic heterogeneity in recombination and mutation rates.

Figure 4
The SNP density distribution (joined gray dots), Poisson distribution with mean 90 (large gray dots), simulated distribution of SNPs with $\rho = \theta = 90$ (gray line), and the Maximum Simulated Likelihood estimate from the coalescent simulations with $\rho \sim \hat{R}$ $\theta_i \sim w_{\theta_i}^{(6.7,14.9)}$ and (black line).

Conclusion
As high resolution data for larger samples become available at a genomic scale, one can use such simulated ML methods (with appropriate sample sizes) to get the null distributions of various test statistics while accounting for heterogeneities in recombination rates (from empirical maps or finer-scaled inferred maps) and mutation rates (from the informative phylogenomic constraints imposed by additional ape genomes). Such empirically observable phenomena should be incorporated into the null hypothesis when more complex models with unobserved historical phenomena, such as population dynamics, population structure, and/or natural selection are tested in humans at the genomic scale. Current scans of the human genome tend to underestimate the costs of ignoring the empirically observable heterogeneities under the null hypothesis.

Methods
In the hierarchical Poisson scheme, heterogeneities are modeled by the following Gamma and Beta probability density functions (*PDF*s),

$$T \sim G(\gamma_1, \gamma_2),$$

$$\text{where,} \, PDF(t) = \frac{1}{\Gamma(\gamma_1)\gamma_2^{\gamma_1}} t^{\gamma_1 - 1} \exp\left(-\frac{t}{\gamma_2}\right),$$

$$0 \le t < \infty, \gamma_1, \gamma_2 > 0,$$

$$\Lambda \sim B(\beta_1, \beta_2),$$

$$\text{where,} \, PDF(\lambda) = \frac{\Gamma(\beta_1 + \beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)} \lambda^{\beta_1 - 1}(1 - \lambda)^{\beta_2 - 1},$$

$$0 \le \lambda \le 1, \beta_1, \beta_2 > 0.$$

The following strategy was used to obtain a simulation-based empirical estimate of the SNP density distribution for each scaled mutation rate

$$\theta_i \in \Theta = \{\theta_1, \cup, \theta_{304}\} =$$
$$\{0.001, 0.01, 0.1, 0.5, 1, 2, 3, 4, 5, \cup, 298, 299, 300\},$$

when the recombination rate was assumed to be distributed according to

$$\hat{R}$$

.

1. for each $\theta_i \in \Theta$, repeat N times:

(a) sample a ρ according to

$$\hat{R}$$

(b) simulate the coalescent according to ρ and θ_i [4,8]

(c) record the number of SNPs

2. Obtain the empirical distribution of SNP density for the given θ_i when $\rho \sim$

$$\hat{R}$$

Authors' contributions
AGC posed the question, RS and RTD made simple models, RS implemented the models and all three authors edited the manuscript.

References
1. Schärer OD: **Chemistry and biology of DNA repair.** *Angew Chem Int Ed* 2003, **42**:2946-2974.
2. Venter JC, *et al.*: **The sequence of the human genome.** *Science* 2001, **291**:1304-1351.
3. Kingman JFC: **The coalescent.** *Stochastic Process Appl* 1982, **13**:235-248.
4. Hudson RR: **Properties of a neutral allele model with intragenic recombination.** *Theoretical Population Biology* 1983, **23**:183-201.
5. Akaike H: **A new look at the statistical model identification.** *IEEE Trans Autom Control* 1974, **19**:716-723.
6. Bhöning D: **Zero-Inflated Poisson models and C.A.MAN: A tutorial collection of evidence.** *Biometrical Journal* 1998, **40**:833-843.
7. Pakes A, Pollard D: **Simulation and the Asymptotics of Optimization Estimators.** *Econometrica* 1989, **57**:1027-1057.
8. Hudson RR: **Generating samples under a Wright-Fisher neutral model of genetic variation.** *Bioinformatics* 2002, **18**:337-338.
9. Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal , Millasseau P, Marc S, Kazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J: **A comprehensive genetic map of the human genome based on 5,264 microsatellites.** *Nature* 1996, **380**:152-154.
10. Broman KW, Murray JC, Sheffeld VC, White RL, Weber JL: **Comprehensive human genetic map: individual and sex-specific variation in recombination.** *Am J Hum Genet* 1998, **63**:861-869.
11. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir SJB, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K: **A high-resolution recombination map of the human genome.** *Nature Genetics* 2002, **31**:241-247.
12. **UCSC Genome Annotation Database** [http://genome.ucsc.edu/goldenPath/gbdDescriptions.html]
13. The international SNP map working group: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.** *Nature* 2001, **409**:928-933.