Note

On the Effect of Fluctuating Recombination Rates on the Decorrelation of Gene Histories in the Human Genome

A. Eriksson and B. Mehlig¹

Physics and Engineering Physics, Gothenburg University/Chalmers, 41296 Gothenburg, Sweden

Manuscript received May 27, 2003

Accepted for publication November 2, 2004

ABSTRACT

We show how to incorporate fluctuations of the recombination rate along the chromosome into standard gene-genealogical models for the decorrelation of gene histories. This enables us to determine how small-scale fluctuations (Poissonian hot-spot model) and large-scale variations (KONG *et al.* 2002) of the recombination rate influence this decorrelation. We find that the empirically determined large-scale variations of the recombination rate give rise to a significantly slower decay of correlations compared to the standard, unstructured gene-genealogical model assuming constant recombination rate. A model with long-range recombination-rate variations and with demographic structure (divergent population) is found to be consistent with the empirically observed slow decorrelation of gene histories. Conversely, we show that small-scale recombination-rate fluctuations do not alter the large-scale decorrelation of gene histories.

GENOME-WIDE variation and decorrelation of gene histories are reflected in patterns of linkage disequilibrium, which in turn shape the genetic variation observed on the molecular level. Recently REICH *et al.* (2002) reported on the first genome-wide measurement of correlations of human gene histories. REICH *et al.* (2002) show that their data are inconsistent with standard gene-genealogical models allowing for nontrivial population structures and demographic schemes, but assuming a constant recombination rate over the genome. The question is thus: Can fluctuations of the recombination rate along the chromosome explain the slow correlation decay of gene histories?

Empirical results (JEFFREYS *et al.* 2001; KONG *et al.* 2002) indicate that the recombination rate is not constant along the chromosome. It was observed (JEFFREYS *et al.* 2001) that, at certain locations, an appreciable fraction of recombination events are concentrated in short regions (~ 1 kb long and spaced 60–90 kb apart), so-called hot spots. At least locally this implies small-scale (< 100 kb) variations of the recombination rate along the chromosome. Genome-wide, long-range fluctuations of the recombination rate for humans have been empirically determined by KONG *et al.* (2002). It is thus necessary to incorporate the effect of fluctuating recombination rates into the standard gene-genealogi-

cal model (GRIFFITHS 1981; HUDSON 1983; TAVARÉ 1984; KAPLAN and HUDSON 1985; HUDSON 1990; NORD-BORG and TAVARÉ 2002). More generally, it is necessary to determine: On which length scales do recombination-rate fluctuations at a certain scale influence the decorrelation function of gene histories most significantly? It has been argued (REICH *et al.* 2002) that small-scale recombination-rate fluctuations (< 100 kb) related to hot spots are an important if not the main feature determining the slow decorrelation of gene histories (assuming that hot spots are to be found genome-wide).

Here we derive an expression for the correlation of gene histories in neutral gene-genealogical models allowing for fluctuating recombination rates. This enables us to explain and quantitatively describe the influence of recombination-rate fluctuations on the correlation of gene histories. We find that large-scale fluctuations of empirically determined recombination rates (KONG *et al.* 2002) give rise to a significantly slower decay of correlations compared to the standard, unstructured, constant population-size gene-genealogical model assuming constant recombination rate. Furthermore, a model with large-scale recombination-rate fluctuations and with demographic structure (divergent population, see EYRE-WALKER *et al.* 1998; REICH *et al.* 2002; TESHIMA and TAJIMA 2002; and references cited therein) is found to be consistent with the empirically observed decorrelation of gene histories. It is not necessary to invoke hot spots.

In a neutral model, KAPLAN and HUDSON (1985; see also GRIFFITHS 1981) have derived a relation between

¹Corresponding author: Physics and Engineering Physics, Gothenburg University/Chalmers, 41296 Gothenburg, Sweden.
E-mail: mehlig@fy.chalmers.se

FIGURE 1.—Illustration of models. (a) Poisson-distributed hot spots (model I). (b) Regions of Poisson-distributed hot spots, separated by regions of uniform recombination (model II). (c) The distribution $p_X(C)$ of model III, for chromosome 5: $X = 200$ kb (red) and 1 Mb (blue), estimated from KONG *et al.* (2002). (d) Demographic structure (divergent population): The population was of constant size N until τ_0 generations ago, when it split into two fractions of size γN and $(1 - \gamma)N$. The two subpopulations remained separate until a recent merging.

the correlation function ρ_{τ_x, τ_y} of the times τ_x and τ_y to the most recent common ancestors of two loci x and y and the amount C of recombination between these two loci. The result of KAPLAN and HUDSON (1985) for the unstructured, constant population-size model is exact for sample size $n = 2$; for large n it is a very good approximation. We observe that their result depends on the total amount of recombination between the two loci, but not on the distribution of recombination events between these loci. Moreover, this is still true when population structure is taken into account. The expected correlation $\rho_{\tau_x, \tau_y}^{\text{exp}}$ is obtained by averaging with a sliding window of length $|y - x|$ along the chromosome. Thus, if $p_X(C)$ is the genome-wide distribution of recombination intensity C in bins of lengths $X = |y - x|$, the expected correlation is

$$\rho_{\tau_x, \tau_y}^{\text{exp}} = \int dC p_X(C) \rho_{\tau_x, \tau_y}(C). \quad (1)$$

It also follows that small-scale fluctuations of the recombination rate on length scales much smaller than X are irrelevant to the decay of correlations on scales of the order of X . In particular, fluctuations due to hot spots at small scales cannot change the decorrelation of gene histories at much larger scales.

Using (1) we have computed $\rho_{\tau_x, \tau_y}^{\text{exp}}$ in four models (illustrated in Figure 1): assuming small-scale variation of the recombination rate (model I); incorporating, in addition, large-scale variation (model II); estimating $p_X(C)$ from the empirical data of KONG *et al.* (2002) (model III); and, in addition, taking into account demographic population structure (model IV).

Model I is the Poisson-hot-spot model of REICH *et*

al. (2002). Recombination events occur at hot spots (of zero width) with rate $R_1 = R/\lambda$. Nearest-neighbor distances between hot spots are exponentially distributed with expected value λ^{-1} (*cf.* Figure 1a). Using (1), the expected correlation within this model can be derived explicitly. We obtain the following expression for $\rho_{\tau_x, \tau_y}^{\text{exp}}$:

$$\rho_{\tau_x, \tau_y}^{\text{exp}} = e^{-\lambda X} \sum_{k=0}^{\infty} \frac{(\lambda X)^k}{k!} \frac{R_1 k + 18}{(R_1 k)^2 + 13 R_1 k + 18}. \quad (2)$$

It is exact for sample size $n = 2$. The expected correlation according to (2) is shown in Figure 2a. In agreement with REICH *et al.* (2002) we find that on distances of the order of the hot-spot spacing, correlations are larger than those in a constant recombination-rate model. However, no choice of parameters could explain the empirically observed decorrelation function (*cf.* data in REICH *et al.* 2002, Figure 6a, reproduced in Figure 2b here). In particular, no significant increase in correlations on length scales $\gg \lambda^{-1}$ is observed, as expected from Equation 1. REICH *et al.* (2002) have fitted an “arbitrary mixed model” to their empirical data. To obtain these results it is necessary to introduce large-scale variations of the recombination rate, on a scale $L \gg X \sim 1$ Mb.

One possibility (model II) is to assume that hot spots occur in clusters, with long ($\gg 1$ Mb) regions of low recombination intensity between them: Hot spots occur in clusters of size pL , separated by empty regions of length $(1 - p)L$, where $0 \leq p \leq 1$ and L is a typical length scale (of the order of several megabases). Within each cluster, the number of hot spots is Poisson distrib-

FIGURE 2.—Decorrelation of gene histories $\rho_{\tau_x, \tau_y}^{\text{exp}}$. (a) Model I (red lines), $R = 4Nr$, $N = 10^4$, and $r = 1.2$ cM/Mb, and for different values of λ as indicated. The dashed line corresponds to constant recombination rate. (b) Comparison between the models and empirical data (upper and lower confidence limits, shown as open and solid squares, respectively, from REICH *et al.* 2002). The parameters of the models were chosen to be consistent with the empirically estimated time to the most recent common ancestor and its coefficient of variation (REICH *et al.* 2002). The lines are: model II (blue line) with parameters $p = 0.55$, $\lambda^{-1} = 50$ kb, $r = 1.2$ cM/Mb, and $L \gg X$; model III (red line), no fitting parameters, based on the data of KONG *et al.* (2002); and model IV (green line) with $\gamma = 0.3$ (asymmetric split). Also shown is the curve for the standard model with constant recombination rate (dashed line).

uted (*cf.* Figure 1b). This model provides a better fit to the empirical data (Figure 2b) than model I, indicating that large-scale fluctuations of the recombination rate are important. We note that assuming $p_X(C) = (1 - p) \cdot \delta(C - R_0X) + p \cdot \delta(C - R_1X)$ can produce an equally good fit to the data (*e.g.*, for $p = 0.55$, $R_0 = 1.2$ cM/Mb and $R_1 = 0.02$ cM/Mb, not shown). In this model the recombination rate is constant on large scales (≥ 1 Mb) and alternates between two values R_0 and R_1 .

We have estimated the genome-wide distribution $p_X(C)$ from empirical recombination data (model III) by sampling $C = g(x + X) - g(x)$, where $g(x)$ is the cumulative genetic distance, at uniformly distributed positions x in the genome. We have obtained $g(x)$ from the pedigree data of KONG *et al.* (2002) as follows: We have taken columns 1 (physical distance) and 3 (sex-

averaged genetic distance) in the web supplement NG917-S13 to KONG *et al.* (2002), ignored entries labeled “NA,” and shifted the origin of both physical and genetic distances so that $g(0) = 0$. The resulting $p_X(C)$ for chromosome 5 is shown in Figure 1c, and the gene-history correlation results for the genome-wide distribution of $p_X(C)$ are shown in Figure 2b. We find that the empirically determined large-scale fluctuations of the recombination rate give rise to significantly enhanced correlations (compared to the standard model assuming constant recombination rate), especially for large distances.

It is expected that population structure can increase the correlations of gene histories at large distances. We have considered the effect of large-scale recombination-rate fluctuations within a well-established model of demographic structure (model IV): the population was of constant size N until τ_0 generations ago, when it split into two fractions of size γN and $(1 - \gamma)N$. The two subpopulations remained separate until a recent merging (see, for instance, EYRE-WALKER *et al.* 1998; REICH *et al.* 2002; TESHIMA and TAJIMA 2002; and references therein). We have calculated ρ_{τ_x, τ_y} explicitly in this model (ERIKSSON and MEHLIG 2004). Without recombination-rate fluctuations, demographic models do not describe the empirically observed correlation of gene histories (REICH *et al.* 2002). We have determined the effect of large-scale recombination-rate fluctuations (KONG *et al.* 2002) on the correlation of gene histories in this model, using Equation 1 together with the appropriate expression for $\rho_{\tau_x, \tau_y}(C)$ derived in ERIKSSON and MEHLIG (2004). The parameters of the model (τ_0 and N) were chosen to be consistent with the empirically estimated time to the most recent common ancestor and its coefficient of variation (REICH *et al.* 2002). The parameter γ was set to 0.3. The resulting correlation function matches the empirical data reasonably well. Decreasing γ gives rise to decreased correlations ($\gamma = 0$ corresponds to the model with fluctuating recombination rates but without demographic structure).

In summary we have determined the influence of recombination-rate fluctuations on the decorrelation of gene histories. We find that small-scale fluctuations are irrelevant to long-range correlation decay. Empirically determined large-scale fluctuations of the recombination rate, however, are found to significantly increase the correlations. Within a model with demographic structure, large-scale fluctuations of empirically determined recombination rates significantly contribute to the empirically observed slow decay of correlations.

We conclude by discussing the implications of our results for the study of genome-wide variability as reflected in single-nucleotide polymorphism (SNP) statistics. Equation 1 determines the effect of recombination-rate fluctuations on $\rho_{\tau_x, \tau_y}^{\text{exp}}$. This quantity, in turn, determines the genome-wide statistics of SNP locations: the variance of the number of SNPs in bins of lengths l along the

chromosomes is determined by the integral of $\rho_{\tau_x, \tau_y}^{\text{exp}}$ over x and y from 0 to l , *i.e.*, by how fast the correlations decay on scales of length l (HUDSON 1990). The INTERNATIONAL SNP MAP WORKING GROUP (2001) has empirically determined the variance of the number of SNPs in short reads (of average length 500 bp); the result was found to be consistent with the standard, unstructured gene-genealogical model assuming a constant recombination rate. This is consistent with our results (Figure 2b): On scales of the order of 500 bp, the recombination-rate fluctuations have little effect on the correlation function. We expect, however, that to understand the statistics of SNP counts in longer bins, it will be necessary to account for long-range recombination-rate fluctuations.

LITERATURE CITED

- ERIKSSON, A., and B. MEHLIG, 2004 Gene-history correlation and population structure. *Phys. Biol.* **1**: 220–228.
- EYRE-WALKER, A., R. L. GAUT, H. HILTON, D. L. FELDMAN and B. S. GAUT, 1998 Investigation of the bottleneck leading to the domestication of maize. *Proc. Natl. Acad. Sci. USA* **95**: 4441–4446.
- GRIFFITHS, R. C., 1981 Neutral 2-locus multiple allele models with recombination. *Theor. Popul. Biol.* **19**: 169–186.
- HUDSON, R. R., 1983 Properties of a neutral allele model with intra-genic recombination. *Theor. Popul. Biol.* **23**: 183–201.
- HUDSON, R. R., 1990 Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7**: 1–44.
- INTERNATIONAL SNP MAP WORKING GROUP, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928–933.
- JEFFREYS, A. J., L. KAUPPI and R. NEUMANN, 2001 Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29**: 217–222.
- KAPLAN, N., and R. R. HUDSON, 1985 The use of sample genealogies for studying a selectively neutral m -loci model with recombination. *Theor. Popul. Biol.* **28**: 382–396.
- KONG, A., D. F. GUDBJARTSSON, J. SAINZ, G. M. JONSDOTTIR, S. A. GUDJONSSON *et al.*, 2002 A high-resolution recombination map of the human genome. *Nat. Genet.* **31**: 241–247.
- NORDBORG, M., and S. TAVARÉ, 2002 Linkage disequilibrium: what history has to tell us. *Trends Genet.* **18**: 83–85.
- REICH, D. E., S. F. SCHAFFNER, M. J. DALY, G. McVEAN, J. C. MULLIKIN *et al.*, 2002 Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.* **32**: 135–142.
- TAVARÉ, S., 1984 Lines of descent and genealogical processes, and their application in population genetics models. *Theor. Popul. Biol.* **26**: 119–164.
- TESHIMA, K., and F. TAJIMA, 2002 The effect of migration during the divergence. *Theor. Popul. Biol.* **62**: 81–95.

Communicating editor: S. TAVARÉ