RESEARCH ARTICLE

On the Recombination Rate Estimation in the Presence of Population Substructure

Julian Hecker[1]*, Dmitry Prokopenko[1], Christoph Lange[1,2,3,4], Heide Löhlein Fier[1,2]

1 Institute of Genomic Mathematics, University of Bonn, Bonn, Germany, **2** Department of Biostatistics, Harvard School of Public Health, Boston, United States of America, **3** Channing Laboratory, Brigham and Women's Hospital, Boston, United States of America, **4** German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

* julian.heckerIGM@uni-bonn.de

OPEN ACCESS

Citation: Hecker J, Prokopenko D, Lange C, Fier HL (2015) On the Recombination Rate Estimation in the Presence of Population Substructure. PLoS ONE 10 (12): e0145152. doi:10.1371/journal.pone.0145152

Editor: Francesc Calafell, Universitat Pompeu Fabra, SPAIN

Received: June 22, 2015

Accepted: November 25, 2015

Published: December 30, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: HLF and CL were supported by Cure Alzheimer's fund, Award Number (R01 MH081862, R01 MH087590) from the National Institute of Mental Health. The project described was supported by the National Heart, Lung, and Blood Institute, Award Number (R01 HL089856, R01 HL089897). The research of C. Lange was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014S1A2A2028559).

Competing Interests: The authors have declared that no competing interests exist.

Abstract

As recombination events are not uniformly distributed along the human genome, the estimation of fine-scale recombination maps, e.g. HapMap Project, has been one of the major research endeavors over the last couple of years. For simulation studies, these estimates provide realistic reference scenarios to design future study and to develop novel methodology. To achieve a feasible framework for the estimation of such recombination maps, existing methodology uses sample probabilities for a two-locus model with recombination, with recent advances allowing for computationally fast implementations. In this work, we extend the existing theoretical framework for the recombination rate estimation to the presence of population substructure. We show under which assumptions the existing methodology can still be applied. We illustrate our extension of the methodology by an extensive simulation study.

Introduction

The discovery that recombination events in human genome are not uniformly distributed, but concentrated in specific genomic regions, which are typically referred to as recombination hotspots [1], was one of the driving forces behind the HapMap Project [2]. The characterization and understanding of the local linkage disequilibrium structure between genetic variants is fundamental for the discovery of disease susceptibility loci (DSLs). The peak recombination rate within recombination hotspots can be multiple times higher than the recombination rate outside such genomic areas [3]. To create recombination hotspots in simulated data, commonly used software tools based on coalescent simulations, e.g. cosi, incorporate recombination rates that vary along the chromosome [4]. Such software tools, in order to generate realistic data, require an accurate fine-scale recombination rate map as input parameter. McVean et al. developed the LDhat software which estimates the recombination rate along the genome according to a assumed step-wise constant model [5]. The approach is computationally fast and can be applied at a genome-wide level. In 2007, Auton and McVean extended this approach and incorporated recombination hotspots into the assumed form of the recombination rate-model [6]. Both approaches are based on the so-called composite-likelihood approximation [7] that calculates the data likelihood as the product of pairwise two-locus probabilities. In order to obtain these sample probabilities, an Importance Sampling (IS) scheme by Fearnhead and Donnelly [8] was applied which requires exhaustive lookup tables of probabilities for a wide range of recombination rates. These methods were used to estimate the recombination rate for the HapMap samples CEU, YRI and ASN [2] (resp. later the 1000Genomes [9] samples), separately. As explained in [10], the overall genetic map of recombination rates was produced by comparison of the total map lengths to that estimated by the pedigree method in [11] and averaging these maps. In 2009, Jenkins and Song proposed approaches to calculate the sample probability for a given configuration with an analytic asymptotic formula of order two in the reciprocal recombination rate. Their approach was initially intended for an infinite-allele model [12] and later extended to a finite-allele model [13]. In particular, they showed how to transform the symmetric diallelic mutation model in Fearnhead and Donnelly [8] resp. LDhat to a so-called Parent-Independent-Mutation (PIM) model and were therefore able to present a different approach to calculate the required two-locus probabilities for LDhat. The key aspect they showed was that the calculations were independent from the specific value of the recombination rate. Thus, they were able to evaluate the sample probability instantly for every recombination rate. This substantially reduces the computational burden. In addition, for large recombination rate values, Monte Carlo based methods as the IS scheme by Fearnhead and Donnelly are less efficient, since the number of recombination events increases and the sampled genealogies become very complicated, whereas the asymptotic sampling formula by Jenkins and Song becomes very precise for large values. In 2012, Jenkins and Song derived an arbitrary order asymptotic expansion for the finite-allele model, using the generator of the corresponding Wright-Fisher diffusion [14]. Together with the application of Pade approximations, this provided a way to evaluate two-locus probabilities efficiently for a wide range of recombination rates. The mentioned diffusion generator used in [14] describes the diffusion limit of the Markov Chain explained in [15]. In this communication, we extend this underlying Markov Chain model to the presence of population substructure with a finite number of subpopulations. Under the assumption of strong migration between the subpopulations, we derive the corresponding diffusion limit of a specific weighted mean of subpopulation frequencies. The key result of our work is, that differences between subpopulation frequencies disappear and the diffusion limit has the form as in the panmictic case with rescaled effective population size. This implies the possibility to combine subpopulation samples and evaluate the corresponding sample probability with the existing methodology without any additional computation effort. The advantage for the estimation of recombination rates lies in the increased sample size and the more realistic underlying model.

Methods

We start with a summary of the results in [13]. We denote the sample, i.e. the genetic data at locus A and B for n study subjects, by $\mathbf{s}$. The parameter $\rho = 4N_e r$ is the population-scaled recombination rate. The parameters θ_A and θ_B refer to the mutation rates at locus A and B. Let r_A be the number of possible alleles at the first locus A and r_B be the number of possible alleles at the other locus B. Denote the different alleles by $A_1, \ldots, A_{r_A}$ and $B_1, \ldots, B_{r_B}$. For notational convenience, we introduce $[k] := \{1, \ldots, k\}$. First, we summarize the work in [14] to emphasize the differences to our model later. Write the sample configuration $\mathbf{s}$ for a single population by

$$\mathbf{s} = (\mathbf{a}, \mathbf{b}, \mathbf{c})$$

with

$$\mathbf{c} = (c_{ij})_{i\in[r_A],j\in[r_B]},$$

where c_{ij} gives the number of gametes with allele A_i at locus A and B_j at locus B. Similar,

$$\mathbf{a} = (a_1, \cdots, a_{r_A}),$$

where a_i gives the number of gametes with allele A_i at locus A and unspecified allele at locus B. $\mathbf{b}$ is defined analogously. Further, write $a = \sum_{i=1}^{r_A} a_i$, $b = \sum_{j=1}^{r_B} b_j$, $c = \sum_{i=1}^{r_A}\sum_{j=1}^{r_B} c_{ij}$ and $n = a + b + c$. This notation is also used in [13]. Jenkins and Song utilize the fact that, under a stationary distribution, the expectation of a suitable function, which is applied to the generator (described in [14]), is equal to zero. If $q(\mathbf{s}; \rho)$ describes the sample probability with reference to ρ, they derived that the sample probabilities can be calculated from the following linear system

$$\begin{aligned}
&[n(n-1) + \theta_A(a+c) + \theta_B(b+c) + \rho c]q((\mathbf{a},\mathbf{b},\mathbf{c});\rho) = \\
&\sum_{i=1}^{r_A} a_i(a_i - 1 + 2c_{i\bullet})q((\mathbf{a}-e_i,\mathbf{b},\mathbf{c});\rho) + \sum_{j=1}^{r_B} b_j(b_j - 1 + 2c_{\bullet j})q((\mathbf{a},\mathbf{b}-e_j,\mathbf{c});\rho) \\
&+\sum_{i=1}^{r_A}\sum_{j=1}^{r_B}[c_{ij}(c_{ij}-1)q((\mathbf{a},\mathbf{b},\mathbf{c}-e_{ij});\rho) + 2a_ib_jq((\mathbf{a}-e_i,\mathbf{b}-e_j,\mathbf{c}+e_{ij});\rho)] \\
&+\theta_A\sum_{i=1}^{r_A}\left[\sum_{j=1}^{r_B} c_{ij}\sum_{k=1}^{r_A} P^A_{ki}q((\mathbf{a},\mathbf{b},\mathbf{c}-e_{ij}+e_{kj});\rho) + a_i\sum_{k=1}^{r_A} P^A_{ki}q((\mathbf{a}-e_i+e_k,\mathbf{b},\mathbf{c});\rho)\right] \\
&+\theta_B\sum_{j=1}^{r_B}\left[\sum_{i=1}^{r_A} c_{ij}\sum_{l=1}^{r_B} P^B_{lj}q((\mathbf{a},\mathbf{b},\mathbf{c}-e_{ij}+e_{il});\rho) + b_j\sum_{l=1}^{r_B} P^B_{lj}q((\mathbf{a},\mathbf{b}-e_j+e_l,\mathbf{c});\rho)\right] \\
&+\rho\sum_{i=1}^{r_A}\sum_{j=1}^{r_B} c_{ij}q((\mathbf{a}+e_i,\mathbf{b}+e_j,\mathbf{c}-e_{ij});\rho),
\end{aligned} \tag{1}$$

with boundary conditions

$$q((e_i, 0, 0);\rho) = \pi^A_i \text{ and } q((0, e_j, 0);\rho) = \pi^B_j \text{ for all } i \in [r_A], j \in [r_B],$$

where π^A and π^B denote the stationary distributions of the mutation models for locus A and B. Since this linear system has to be solved for every recombination rate ρ separately and grows rapidly in n, they propose an asymptotic expansion in $\frac{1}{\rho}$

$$q(\mathbf{s};\rho) = q_0(\mathbf{s}) + \frac{q_1(\mathbf{s})}{\rho} + \frac{q_2(\mathbf{s})}{\rho^2} + \cdots.$$

Now, the corresponding recursions are solved only once for a fixed mutation model and the sample probabilities, with reference to several recombination rates, can be evaluated by plugging in the recombination parameter ρ. The estimation procedures in [5] and [6] are based on two-locus probabilities from the model in [8] with respect to some mutation parameter θ_{FD}, recombination rate ρ and a symmetric diallelic mutation model

$$P_{FD} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

As shown in [13], both models are in line if we set $r_A = r_B = 2$, $\theta = 2\theta_{FD}$ and

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

As explained in the introduction, the generator in [14] corresponds to the diffusion limit of the discrete Markov Chain in [15]. In the next section, we introduce the population substructured extension of this discrete Markov Chain, derive the diffusion limit of weighted mean frequencies and show that we can utilize the same idea as in [14] to improve the recombination rate estimation framework.

Model setting

As explained, we consider a two-locus model. As in [14], let r_A be the number of possible alleles at the first locus A and r_B be the number of possible alleles at the other locus B. The generations are supposed to be non-overlapping and the population is monoecious. We suppose the underlying model for the discrete Wright-Fisher model as it is described in [15], which leads to a diffusion approximation, corresponding to the generator in [14]. But in addition, we assume that the population is subdivided into $\Gamma < \infty$ subpopulations. Denote by q'_α, $\alpha = 1, \ldots, \Gamma$, the fractions of the corresponding subpopulation sizes and let N be the overall population size. Thus, $0 < q'_\alpha < 1$ and $\sum_{\alpha=1}^{\Gamma} q'_\alpha = 1$. Set $N_\alpha = q'_\alpha N$ for $\alpha = 1, \ldots, \Gamma$. We model the migration procedure as in the discrete model in [16]. In this context, let $m_{\alpha\beta}$ denote the probability that an individual in subpopulation α was in subpopulation β one generation before. These probabilities are given by the backward migration matrix M. The following diagram summarizes the lifecycle of the model (compare with [16]).

$$\begin{tikzcd}
\text{Adult} \arrow[r, "\text{reproduction}"] & \text{Zygote} \arrow[r, "\text{recombination}"] & \text{Adult} \arrow[r, "\text{migration}"] & \text{Adult} \\
(N_\alpha, (P_{\alpha ij})_{ij})_\alpha & (\infty, (P_{\alpha ij})_{ij})_\alpha & (\infty, (P^{*}_{\alpha ij})_{ij})_\alpha & (\infty, (P^{**}_{\alpha ij})_{ij})_\alpha
\end{tikzcd}$$

$$\begin{tikzcd}
{} \arrow[r, "\text{mutation}"] & \text{Adult} \arrow[r, "\text{regulation}"] & \text{Adult} \\
 & (\infty, (P^{***}_{\alpha ij})_{ij})_\alpha & (\infty, (P'_{\alpha ij})_{ij})_\alpha
\end{tikzcd}$$

The Markov Chain Z^N, which describes the relative frequencies $(P_{\alpha ij})_{\alpha, i, j}$, is completely defined in S1 Appendix. $P_{\alpha ij}$ denotes the relative frequency of type $A_i\, B_j$ in subpopulation α in the model Z^N, we suppress the dependency on N for convenience. Note that the subpopulation sizes are assumed to be effectively infinite except shorlty before and after population regulation [16].

Parameter assumptions

In order to obtain a diffusion limit, we make the following assumptions on the model parameters.

Biological parameters. We use the well established assumptions on the parameter scaling for the biological mechanisms. These conditions depend on the effective population size (section 3.7 in [17]). The effective population size for our model is derived below. We write

$$u^N_{ki} = u^N_A P^A_{ki} > 0, \text{ for all } i, k \in [r_A], i \neq k, \tag{2}$$

and

$$v^N_{jl} = u^N_B P^B_{jl} > 0, \text{ for all } j, l \in [r_B], j \neq l, \tag{3}$$

where u_A and u_B are the mutation rates per individual per generation and P^A as well as P^B are the transition matrices for mutation for location A resp. B with $P^A_{ii} > 0$ and $P^B_{jj} > 0$ as in [14]. See Eq (A.1) and (A.2) in S1 Appendix for the definition of the mutation model. For convenience, we omit technical conditions of the truncation N, since we are only interested in the diffusion limit resp. large N. We assume

$$\lim_{N\to\infty} 4N_e u^N_A = \theta_A$$

resp.

$$\lim_{N\to\infty} 4N_e u^N_B = \theta_B.$$

For the recombination fractions in Eq (A.3) resp. (A.4) in S1 Appendix, we suppose

$$\lim_{N\to\infty} 4N_e r_N = \rho < \infty. \tag{4}$$

Migration. The backward migration matrix M is assumed to be a stochastic matrix, which is irreducible and aperiodic. Migration is independent of time and N_e (see below) resp. N. This is the strong migration assumption. Denote the stationary distribution of M by ξ. This means that migration dominates all other evolutionary forces. This assumption was used in several papers. We adopted it from [16]. Additionally, this suggestion was analyzed in [18], in other papers of [19–21] and [22–24].

Diffusion approximation

The next step is to analyze the diffusion limit of the Markov Chain Z^N. As in [15] and [16], this is essentially an application of the diffusion approximation Theorem in [25] in S1 Appendix, which is also stated as Theorem (A.3). We describe by simple calculation that the Theorem is still applicable for our extended Markov Chain S1 Appendix. Important for the derivation of the diffusion limit is the connection between N and N_e, the effective population size. Nagylaki observed and described this connection in the setting of a subdivided population with strong migration in a one-locus model in [16]. He stated

$$N_e = \delta N,$$

where

$$\delta = \left[\sum_\alpha \xi^2_\alpha / q'_\alpha\right]^{-1} \tag{5}$$

Therefore, $N_e \leq N$ and in particular N_e is equal to N if and only if $\sum_\alpha q'_\alpha m_{\alpha\beta} = q_\beta$. This scenario is called conservative migration. These results are still valid in our two-locus scenario. Analogous to [16], we define

$$X^N := \Phi_N((P_{\alpha ij})_{\alpha ij}) = (P_{ij})_{(i,j)\in J} =: P, \tag{6}$$

where $P_{ij} = \sum_{\alpha} \xi_\alpha P_{\alpha ij}$ and

$$Y^N := \Psi_N((P_{\alpha ij})_{\alpha ij}) = (d_{\alpha ij})_{\alpha,(i,j)\in J} =: d, \tag{7}$$

where $d_{\alpha ij} = P_{\alpha ij} - P_{ij}$. In addition,

$$J := \{(i,j) : i \in \{1,\cdots,r_A\}, j \in \{1,\cdots,r_B\}, (i,j) \neq (r_A, r_B)\}. \tag{8}$$

The idea is to set the two timescales in Theorem (A.3) to $\varepsilon_N = 2N_e$ and $\delta_N = 1$. Recombination and mutation work on the slow timescale ε_N, but the strong migration assumption implies that migration works on the fast timescale δ_N. Then, it follows that the scaled $X^N([2N_e\bullet])$ converges to a diffusion process X with explicitly known generator (A.17). The process Y^N describes the derivation of the frequencies within the subpopulations from the weighted mean frequencies. The result is, that $Y^N([2N_e\, t]])$ converges to 0, for every $t > 0$. The interpretation is that, since migration works faster than the biological mechanisms, the differences in relative frequencies within the subpopulations disappear and approach the mean frequencies. In S1 Appendix, it is shown that the process $X^N[2N_e\bullet]$ converges weakly to a process X with state space (A.16), which is associated to the generator (A.17) resp. (A.23) with state space (A.22). This generator has the same form as the generator which was used in [14]. The difference lies in the scaled effective population size. This observation is in agreement with the one-locus results in [16]. The assumptions about the mutation model (Eqs (2) and (3)) imply an important fact about the discrete Markov Chains.

Lemma 1: The Markov chain Z^N has a unique stationary distribution μ_N.

Proof See S1D Appendix.

Another important observation is that there exists a unique stationary distribution μ for the diffusion process X.

Lemma 2: The diffusion process X has a unique stationary distribution μ.

Proof This is a consequence of the restriction of the results in [26] to a finite number of possible alleles.

Sample configurations

As described in the introduction, the main interest lies in the efficient evaluation of approximate sample probabilities. Therefore, we need to extend the definition of the latter object to incorporate our scenario. A sample configuration for a subdivided population is defined by

$$\mathbf{s} = (\mathbf{s}_1, \cdots, \mathbf{s}_\Gamma),$$

where

$$\mathbf{s}_\alpha = (\mathbf{a}_\alpha, \mathbf{b}_\alpha, \mathbf{c}_\alpha) \text{ for } \alpha \in [\Gamma].$$

Here, we have

$$\mathbf{c}_\alpha = (c_{\alpha ij})_{i\in[r_A], j\in[r_B]},$$

where $c_{\alpha ij}$ gives the number of gametes in subpopulation α with allele A_i at locus A and B_j at locus B. Similar,

$$\mathbf{a}_\alpha = (a_{\alpha 1}, \cdots, a_{\alpha r_A}),$$

where $a_{\alpha i}$ gives the number of gametes in subpopulation α with allele A_i at locus A and unspecified allele at locus B. $\mathbf{b}_\alpha$ is defined analogously. This is the straightforward extension of the definition in [14].

Sample probabilities

Recall the stationary distributions μ_N of the discrete model and define

$$\mathcal{P}_\Gamma((P_{\alpha ij})_{\alpha ij};\mathbf{s}) = \prod_{\alpha=1}^{\Gamma} \mathcal{P}((P_{\alpha ij})_{ij};\mathbf{s}_\alpha)$$

with

$$\mathcal{P}((P_{\alpha ij})_{ij};\mathbf{s}_\alpha) = \left(\prod_i P_{\alpha i\bullet}^{a_{\alpha i}}\right)\left(\prod_j P_{\alpha\bullet j}^{b_{\alpha j}}\right)\left(\prod_{i,j} P_{\alpha ij}^{c_{\alpha ij}}\right), \quad \alpha \in [\Gamma],$$

where $P_{\alpha i\bullet} = \sum_{j=1}^{r_B} P_{\alpha ij}$ and $P_{\alpha\bullet j}$ analogously. The probability of a sample configuration for the model can be expressed as

$$\mu_N\left(\mathcal{P}_\Gamma((P_{\alpha ij})_{\alpha ij};\mathbf{s})\right).$$

If we have Γ = 1, this is in line with the approach in [14]. Based on the observations above, we can derive that we can approximate these sample probabilities by combined samples for the diffusion limit. More precise:

Lemma 3:

$$\lim_{N\to\infty} \mu_N\left(\mathcal{P}_\Gamma((P_{\alpha ij})_{\alpha ij};\mathbf{s})\right) = \mu\left(\mathcal{P}((x_{ij})_{ij};\cup_\alpha \mathbf{s}_\alpha)\right),$$

where $\cup_\alpha \mathbf{s}_\alpha$ is the combined sample for one population with $a_i = \sum_{\alpha=1}^{\Gamma} a_{\alpha i}$, $b_j = \sum_{\alpha=1}^{\Gamma} b_{\alpha j}$ and $c_{ij} = \sum_{\alpha=1}^{\Gamma} c_{\alpha ij}$ for all i, j.

Proof. See S1E Appendix.

μ is the stationary distribution of the diffusion process X and $(x_{ij})_{ij}$ in the state space K_c (A.22). We do not distinguish in the notation between these objects in different characterizations, since they can be identified with each other.

Main result. Thus, our main result is that we can combine the samples over all subpopulations and evaluate the corresponding sample probability for the diffusion process, which is stated above. The advantage is that the computational effort is still the same as in the panmictic case. The only adjustment is done by the rescaling of the effective population size. To clarify this connection, note that the generator for a single population in [14], has the same form and depends on the recombination rate ρ and mutation rates θ_A resp. θ_B. As mentioned above, Jenkins and Song evaluated the expression for the sample probability, given biological parameters, with a sophisticated recursion technique, which was derived by the fact that

$$\mu(\mathcal{L}\mathcal{P}) = 0.$$

Here denotes $\mathcal{L}$ the associated generator of the diffusion process. For our scenario, we are dealing with a rescaled effective population size, conditioned on the migration model, by the factor δ in Eq (5) and can apply the same technique.

Results/Simulation study

In order to give a better impression about the results, we present an empirical example for the theory. We consider a model with two subpopulations. A realistic choice for N is about $N = 10{,}000$, according to [4]. The fractions of subpopulation sizes are denoted by $q_1^{'}$ and $q_2^{'}$. We choose the backward migration matrix to be equal to

$$M = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}.$$

Note that this matrix satisfies the migration assumptions and implies $\xi = \left(\frac{2}{3}, \frac{1}{3}\right)$. We consider a diallelic model for both loci, i.e.

$$r_A = 2, \quad r_B = 2,$$

and assume that the mutation rates are the same for loci A and loci B. Furthermore, we choose

$$P^A = P^B = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix},$$

as in [14]. As it is used in [14] (as explained in the Methods section, this corresponds to the mutation model used in LDhat), reasonable parameter values for θ are $\theta_A = \theta_B = \theta \in [0.001, 0.01]$ and $\rho = 50$. To test our results, we simulate the discrete model Z^N of the relative frequencies in both subpopulations with arbitrary initial distributions according to Eqs (A.4), (A.5), (A.6) and (A.7) in S1 Appendix. For this, we compute the corresponding mutation and recombination rates u_0 and r_0 satisfying

$$\theta = 4N_e u_0$$

and

$$\rho = 4N_e r_0.$$

for $\theta = 0.01$, $\rho = 50$ and $N_e = 10{,}000$, to get realistic parameter values for the Markov Chain. To obtain an estimate of the sample probability under the stationary distribution, we run the model with respect to u_0 and r_0 over 10^{12} generations and estimate the corresponding quantity for some example configurations every 10^5 generations. We do this for two different scenarios of subpopulation sizes. The first choice is $(q_1, q_2) = \xi$ and the second $(q_1, q_2) = \left(\frac{1}{5}, \frac{4}{5}\right)$. In the first case, the condition for conservative migration is satisfied, $\delta_1 = 1$. Therefore, the theory predicts that the effects of the subdivision disappear. In the second scenario, we compute

$$\delta_2 = \left(\frac{\left(\frac{2}{3}\right)^2}{\frac{1}{5}} + \frac{\left(\frac{1}{3}\right)^2}{\frac{4}{5}} \right)^{-1} \approx \frac{1}{2.3611}.$$

Based on these observations, we compare the empirical results for the discrete model with the results from the diffusion approximation. As explained, we have to take care about the corresponding scaled mutation and recombination rates. In the first case, this implies $\theta = 0.01$ and $\rho = 50$, since $\delta_1 = 1$, and in the second case $\theta \approx 0.004235$ and $\rho \approx 21.17$. The sample probabilities for the diffusion approximation can either be obtained by the software package asf [13] or can be calculated by solving the linear system Eq (1). Solving the linear system gives us the exact value for the sample probabilities, the software asf only calculates a second order approximation. We implement a solver for the linear system, similar to the goldings recursion program

Table 1. Comparison of estimated and expected probabilities for the scenario $(q_1, q_2) = (\xi_1, \xi_2)$.

sample configuration	discrete	combined sample	diffusion
((0, 0, 1, 0), (0, 0, 0, 1))	0.00123218	(0, 0, 1, 1)	0.00123137
((0, 0, 1, 1), (0, 0, 1, 0))	0.000615191	(0, 0, 2, 1)	0.00061418
((1, 0, 2, 0), (1, 0, 2, 0))	6.1255e-05	(2, 0, 4, 0)	6.1262e-05
((4, 0, 0, 0), (2, 0, 0, 0))	0.245318	(6, 0, 0, 0)	0.244383
((3, 0, 0, 0), (2, 1, 0, 0))	0.000243998	(5, 1, 0, 0)	0.000244123
((1, 1, 0, 0), (0, 0, 0, 1))	1.46358e-06	(1, 1, 0, 1)	1.50463e-06

6 different example sample configurations. The discrete value corresponds to the estimated probability from the discrete model, the diffusion value to exact solution of the linear system.

doi:10.1371/journal.pone.0145152.t001

Table 2. Comparison of estimated and expected probabilities for the scenario $(q_1, q_2) = (\frac{1}{5}, \frac{4}{5})$.

sample configuration	discrete	combined sample	diffusion
((0, 0, 1, 0), (0, 0, 0, 1))	0.000522292	(0, 0, 1, 1)	0.000525972
((0, 0, 1, 1), (0, 0, 1, 0))	0.000260697	(0, 0, 2, 1)	0.00026271
((1, 0, 2, 0), (1, 0, 2, 0))	2.62979e-05	(2, 0, 4, 0)	2.62440e-05
((4, 0, 0, 0), (2, 0, 0, 0))	0.248091	(6, 0, 0, 0)	0.247599
((3, 0, 0, 0), (2, 1, 0, 0))	0.000104652	(5, 1, 0, 0)	0.000104806
((1, 1, 0, 0), (0, 0, 0, 1))	2.60423e-07	(1, 1, 0, 1)	2.67692e-07

Description: See Table 1.

doi:10.1371/journal.pone.0145152.t002

from Richard R. Hudson [7], and use these results (the results are consistent with the the sample probabilities from asf). The comparison of the values is given in the Tables 1 and 2. We only consider sample configurations with full gamete information, e.g. $a_1 = a_2 = b_1 = b_2 = 0$. As we can see in the Tables 1 and 2, the observations are in line with the theoretical predictions.

Discussion

In their work, Jenkins and Song derived an approximate sampling formula of arbitrary order from the generator of the Wright-Fisher diffusion [14]. Relying on first results in [13] and the detailed study of accuracy in [14], we restate that the application of the approximate sampling formula has huge advantages in relation to the computational demanding Monte Carlo based methods, especially for large values of ρ. These theoretical results can be applied to make the existing recombination rate estimation procedures as implemented in LDhat and rhomap more efficient. In this communication, we go one step further and extend the underlying discrete Wright-Fisher model to the scenario in which the sample population consists of distinct subpopulations. We describe the required assumptions for the biological parameters and the migration model, in order to derive the corresponding diffusion limit of the weighted mean frequencies. The strong migration assumption implies that the form of the generator for the diffusion limit is the same as in the panmictic case. Under our mutation and recombination model, we can show the existence and uniqueness of stationary distributions for the discrete Markov Chains and the diffusion process. We describe the connection between these stationary distributions, which is essentially the motivation for the approach to compute approximate sample probabilities. An empirical example shows that our theoretical derivations are valid. The conclusion is that the probability of a sample containing subsamples from each subpopulation can be calculated as the probability of the combined sample with the original approach of [14]. The difference lies in the rescaled effective population size. The advantage is that one can handle 2-loci data for multiple subpopulations without any additional computational burden.

We believe that the suggested extension can help to achieve a more realistic setting for the recombination rate estimation and increased efficiency due to larger sample sizes.

Supporting Information

S1 Appendix. Theoretical derivations and description. In this file, we describe the underlying Markov Chain, the theoretical tools behind the diffusion limit and the proofs of Lemma 1 and 3.
(PDF)

Acknowledgments

We thank Professor Martin Rumpf and Professor Anton Bovier for helpful comments and suggestions. HLF and CL were supported by Cure Alzheimer's fund, Award Number (R01 MH081862, R01 MH087590) from the National Institute of Mental Health. The project described was supported by the the National Heart, Lung, and Blood Institute, Award Number (R01 HL089856, R01 HL089897). The research of C. Lange was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014S1A2A2028559).

Author Contributions

Conceived and designed the experiments: JH CL HLF. Performed the experiments: JH DP. Analyzed the data: JH DP HLF. Wrote the paper: JH CL HLF.

References

1. Kauppi L, Jeffreys AJ, Keeney S. Where the crossovers are: recombination distributions in mammals. Nat Rev Genet. 2004; 5: 413–424. PMID: 15153994
2. The Internation HapMap Consortium, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467: 52–58. PMID: 20811451
3. Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P. Human recombination hot spots hidden in regions of strong marker association. Nat Genet. 2005; 37: 601–606. PMID: 15880103
4. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 2005; 15: 1576–1583. doi: 10.1101/gr.3709305 PMID: 16251467
5. McVean GA, 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 doi: 10.1126/science.1092500 PMID: 15105499
6. Auton A, McVean G. Recombination rate estimation in the presence of hotspots. Genome Res. 2007; 17: 1219–1227. doi: 10.1101/gr.6386707 PMID: 17623807
7. Hudson RR. Two-locus sampling distributions and their application. Genetics. 2001; 159: 1805–1827. PMID: 11779816
8. Fearnhead P, Donnelly P. Estimating recombination rate from population genetic data. Genetics. 2001; 159: 1299–1318. PMID: 11729171
9. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491: 56–65. doi: 10.1038/nature11632 PMID: 23128226
10. Auton A. The estimation of recombination rates from population genetic data. DPhil Thesis 2007; University of Oxford.

11. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, et al. A high-resolution recombination map of the human genome. Nat Genet. 2002; 31: 241–247. PMID: 12053178

12. Jenkins PA, Song YS, An asymptotic sampling formula for the coalescent with recombination. Ann Appl Probab. 2010; 20: 1005–1028. doi: 10.1214/09-AAP646 PMID: 20671802

13. Jenkins PA, Song YS. Closed-form two-locus sampling distributions: Accuracy and universality. Genetics. 2009; 183: 1087–1103. doi: 10.1534/genetics.109.107995 PMID: 19737744

14. Jenkins PA, Song YS. Pade approximants and exact two-locus sampling distributions. Ann of Appl Probab. 2012; 22: 576–607.

15. Ethier SN, Nagylaki T. Diffusion approximations of the two-locus Wright-Fisher model. J Math Biol. 1989; 27: 17–28. doi: 10.1007/BF00276078 PMID: 2708916

16. Nagylaki T. The strong-migration limit in geographically structured populations. J Math Biol. 1980; 9: 101–114. doi: 10.1007/BF00275916 PMID: 7365330

17. Ewens WJ. Mathematical Population Genetics I. Theoretical Introduction. Second edition, Springer, New York 2004.

18. Bahlo M, Griffiths RC. Coalescence time for two genes from a subdivided population. J Math Biol. 2001; 43: 397–410. doi: 10.1007/s002850100104 PMID: 11767204

19. Nagylaki T. The robustness of neutral models of geographical variation. Theor Popul Biol. 1983; 24: 268–294. doi: 10.1016/0040-5809(83)90029-1

20. Nagylaki T. The expected number of heterozygous sites in a subdivided population. Genetics. 1998; 149: 1599–1604. PMID: 9649546

21. Nordborg M. Structured coalescent processes on different time scales. Genetics. 1997; 146: 1501–1514. PMID: 9258691

22. Notohara M. The strong-migration limit for the genealogical process in geographically structured populations. J Math Biol. 1993; 31: 115–122. doi: 10.1007/BF00171221

23. Notohara M. The number of segregating sites in a sample of DNA sequences from geographically structured population. J Math Biol. 1997; 36: 188–200. PMID: 9440307

24. Notohara M. A perturbation method for the structured coalescent with strong migration. J Appl Probab. 1997; 37: 148–167.

25. Ethier SN, Kurtz TG. Markov Processes: Characterization and Convergence. Wiles Series in Probability and Statistics, John Wiley & Sons, Inc., New York 2005.

26. Ethier SN, Griffiths RC. The two-locus model as a measure-valued diffusion. Adv Appl Probab. 1990; 22: 773–786.