OXFORD GENETICS

GENETICS, 2022, 221(3), iyac052
https://doi.org/10.1093/genetics/iyac052
Advance Access Publication Date: 23 April 2022
Investigation

Recombination, selection, and the evolution of tandem gene arrays

Moritz Otto ,† Yichen Zheng ,† Thomas Wiehe *

Institut für Genetik, Universität zu Köln, 50674 Köln, Germany

*Corresponding author: Institut für Genetik, Universität zu Köln, Zülpicher Straße 47a, 50674 Köln, Germany. Email: twiehe@uni-koeln.de
†These authors contributed equally to this work.

Abstract

Multigene families—immunity genes or sensory receptors, for instance—are often subject to diversifying selection. Allelic diversity may be favored not only through balancing or frequency-dependent selection at individual loci but also by associating different alleles in multicopy gene families. Using a combination of analytical calculations and simulations, we explored a population genetic model of epistatic selection and unequal recombination, where a trade-off exists between the benefit of allelic diversity and the cost of copy abundance. Starting from the neutral case, where we showed that gene copy number is Gamma distributed at equilibrium, we derived also the mean and shape of the limiting distribution under selection. Considering a more general model, which includes variable population size and population substructure, we explored by simulations mean fitness and some summary statistics of the copy number distribution. We determined the relative effects of selection, recombination, and demographic parameters in maintaining allelic diversity and shaping the mean fitness of a population. One way to control the variance of copy number is by lowering the rate of unequal recombination. Indeed, when encoding recombination by a rate modifier locus, we observe exactly this prediction. Finally, we analyzed the empirical copy number distribution of 3 genes in human and estimated recombination and selection parameters of our model.

Keywords: gene families; unequal recombination; epistasis; balancing selection; immune genes

Introduction

Multigene families occur in most, if not all, genomes of eukaryotes—in metazoans as well as in plants. They may be conserved across large evolutionary distances, such as the histones or tRNA gene families, or rapidly diversify in single species, such as the nucleotide binding leucine rich domain (*NLR*) genes in *Danio rerio* (Howe *et al.* 2016) or the leucine rich repeat (*LRR*) genes in *Arabidopsis thaliana* (de Weyer *et al.* 2019).

Interspecies comparison of gene families derived from whole-genome duplication has been used, for instance, to estimate relative rates of gene loss and functional divergence (Nadeau and Sankoff 1997). On a shorter time scale, segmental duplication and unequal recombination are perhaps the more important mechanisms to explain gene family size differences between species, populations, and individuals. Modeling gene family evolution has a quite long history (Smith 1974; Demuth and Hahn 2009; Innan 2009; Liu *et al.* 2011). The roadmap in a population genetic framework was laid out in a series of contributions by Ohta (1976, 1979, 1984, 1987, 1988, 2000). These models typically include forces such as selection and unequal recombination or gene conversion. To describe the dynamics of copy number variation (CNV) generated by unequal recombination Takahata (1981) introduced a general model based on the work of Krüger and Vogel (1975). Fostered especially by the human genome diversity projects, leading to the realization that structural variation is more than abundant in human populations and observing genome size differences between individuals of 100 Mb and more (Tuzun *et al.* 2005; Redon *et al.* 2006; Eichler 2008), we are witnessing revived interest in modeling and analyzing the evolution of gene families and of the forces and mechanisms driving copy number polymorphisms.

Tandem gene duplication may happen due to some form of replication error, mispairing or segregation anomaly, notably unequal or—less frequently—nonhomologous recombination (Silver 2001). A duplicated gene initially arises in a single individual, very much like a base mutation, and may be lost by drift or be propagated to the offspring in subsequent generations. On its way to fixation, or loss, such a duplication manifests itself as CNV in a given population and—given sufficiently large populations—is sensed by the filter of natural selection. When beneficial, directional selection will accelerate its fixation and subsequent purifying selection will prevent it from loss. Alternatively, when beneficial only in conjunction with other alleles or other copies, balancing selection may force it to remain at intermediate frequency. The best-known examples are perhaps the alleles of pathogen receptors and immune genes, such as those of the MHC complex in vertebrates. Balancing selection, however, comes with a fitness cost in terms of segregation load. Haldane (1937) had suggested that this effect may be alleviated

Received: January 26, 2022. **Accepted:** March 17, 2022

or avoided when overdominant alleles are arrayed in tandem on the same chromosome rather than be combined on homologous chromosomes. Only recently, this fundamental idea has been experimentally tested—and confirmed—in populations of the mosquito *Culex pipiens* (Milesi *et al.* 2017).

Here, we designed a model of tandemly arrayed genes whose evolution is driven by unequal recombination together with a mixture of diversifying and negative selection. More precisely, negative selection will keep copy number in check, while allelic diversity is positively selected. We implement this via a product of 2 multiplicative fitness components: one of them is decreasing with copy number and the other one is increasing with allele number [see equation (1)]. In its structure, this fitness function is an old acquaintance. Very similar versions feature in the classical model of Muller's ratchet (Haigh 1978) and its epistatic relatives (Kondrashov 1982; Chao 1988).

We discovered the following: first, in the absence of selection, i.e. when diversity of alleles does not confer any fitness benefit and additional copies do not provide any cost, the distribution of copy numbers can be analytically expressed. It is a Gamma distribution with shape $\alpha = 4$ and with a scale, which depends only on the mean copy number of the initial distribution. With selection, the limiting distribution is still well approximated by a Gamma distribution, but depends on the combination of selection coefficients and recombination rate, and *not* on the initial distribution. Second, population size can have a stronger effect on mean fitness and allelic diversity than the strength of selection itself. Third, low recombination rates may be favorable to maintain allelic diversity. Consistent with this, when recombination rates are coded as alleles at a modifier locus and are allowed to evolve over time, we observe a tendency toward recombination rate reduction.

Taken together, our model captures essential aspects of a multigene family driven by a force of increasing allelic diversity and, at the same time, an opposing force of maintaining genome and chromosome integrity and of limiting both segregation and recombination loads.

Based on the empirical copy number distribution in a set of 3 exemplary gene families in human, we estimated the strengths of selection and (unequal) recombination rates in a natural population.

Methods

Model

We consider a *compound* model in which the number of copies (y) of a certain gene per individual, as well as the number of alleles (x), is variable. When alleles are all considered distinct (but without labeling their identities) and copy numbers remain variable, we call this the *y-only model*.

In a diploid population of effective size $N \leq \infty$ let individual i, $1 \leq i \leq N$, carry $y_i = y_i^{m\prime} + y_i^{p\prime}$ copies of a particular gene on its maternal (m) and paternal (p) chromosomes. We use the notation y' for the number of copies per chromosome when neither the individual nor the parental status matter. If copies are distinguishable, we call them *alleles* and let x', $1 \leq x' \leq y'$, be the number of different alleles on a chromosome with y' copies. By extension, individual i has $x_i \leq x_i^{m\prime} + x_i^{p\prime}$ alleles (Fig. 1c, alleles indicated by different colors). Fitness ω_i of individual i is determined by both copy and allele numbers: $\omega_i = \omega_i(x_i, y_i)$. We assume that increasing the number of copies incurs a fitness *cost*, representing adverse effects to genomic structure and integrity, while increasing the number of alleles incurs a fitness *benefit*, representing improved function such as recognition of a wider range of pathogens or stimuli. To fix ideas, we consider the following fitness function:

$$\omega_i = \omega(x_i, y_i) = (1+s_x)^{\left(\sum_{k=0}^{x_i-1}\beta_x^k\right)} \times (1-s_y)^{\left(\sum_{k=0}^{y_i-3}\beta_y^k\right)}. \quad (1)$$

The cost is only counted from the third copy, since the ground state is a single-copy gene with exactly one copy on each chromosome. The selection coefficients $0 < s_x, s_y \ll 1$ are positive and the epistasis parameters $0 < \beta_x \leq 1 \leq \beta_y$ are independent of i. In the following, we omit index i unless required for clarity. The way we define epistasis reflects the classical concepts of diminishing returns (β_x) and synergistic epistasis (β_y): the benefit of adding new alleles decreases with the number of already existing alleles. Think of the physiological limit preventing perfect recognition of an infinite number of possible pathogens or sensory stimuli in nature. In contrast, the cost of adding more copies increases with the number of already existing copies. This reflects the growing threat to genome integrity by inserting more and more copies.

For any fixed copy number y, fitness is maximized when $x = y$, i.e. when every copy is a different allele (which is an assumption in the y-only model). Whether fitness is maximized for small or for large y depends on the relative magnitudes of s_x and s_y: assuming $x = y$ and $s_x \leq s_y$, maximum fitness is achieved at the lowest possible copy number, $y = 2$. Arguably, this situation represents the standard scenario for single-copy genes in nature: the cost of adding copies would outweigh its benefit. In contrast, when $s_x > s_y$, maximum fitness may be attained at values $y > 2$. Without epistasis, and as a function of y, fitness is monotonically increasing, with lowest fitness at $y = 2$. With epistasis, fitness has a nontrivial maximum at y^* (Fig. 1a). In this case, we have (see Appendix):

$$y^* = \frac{1}{\ln\left(\frac{\beta_y}{\beta_x}\right)}\left(2\ln(\beta_y) + \ln\left(\frac{\beta_y - 1}{1-\beta_x}\right) + \ln\left(-\frac{\ln(1+s_x)}{\ln(1-s_y)}\right) + \ln\left(-\frac{\ln(\beta_x)}{\ln(\beta_y)}\right)\right). \quad (2)$$

Assuming further $\beta_x = 1-\varepsilon$ and $\beta_y = 1+\varepsilon$ for small $\varepsilon > 0$, and using $\ln(1+\varepsilon) \approx \varepsilon$, y^* simplifies to

$$y^* \approx 1 + \frac{\ln(s_x) - \ln(s_y)}{2\varepsilon} = 1 + \frac{\ln\left(\frac{s_x}{s_y}\right)}{2\varepsilon}. \quad (3)$$

In finite populations, alleles are lost by drift. Although new alleles are introduced by mutation, one generally has $x < y$ at mutation-drift equilibrium. We employ an infinite alleles model: mutation occurs with rate μ per copy per individual per generation and turns a given allele into a new, previously nonexisting one. The more copies an individual has, the more likely a new allele will be generated. Note that mutation does not change y or y', but it may increase x and x'. The y-only model can be interpreted as the limiting scenario for large mutation rates such that any 2 copies are different. Therefore, mutation is explicitly required only in the simulations of the compound model, but not for the analytical results of the y-only model.

In both the compound and the y-only models, recombination may be nonhomologous or *unequal*. As a consequence, copy number may change across generations. It is implemented as follows (Fig. 1c): first, choose a pair of chromosomes and decide whether recombination occurs (probability r) or not ($1-r$). In the first

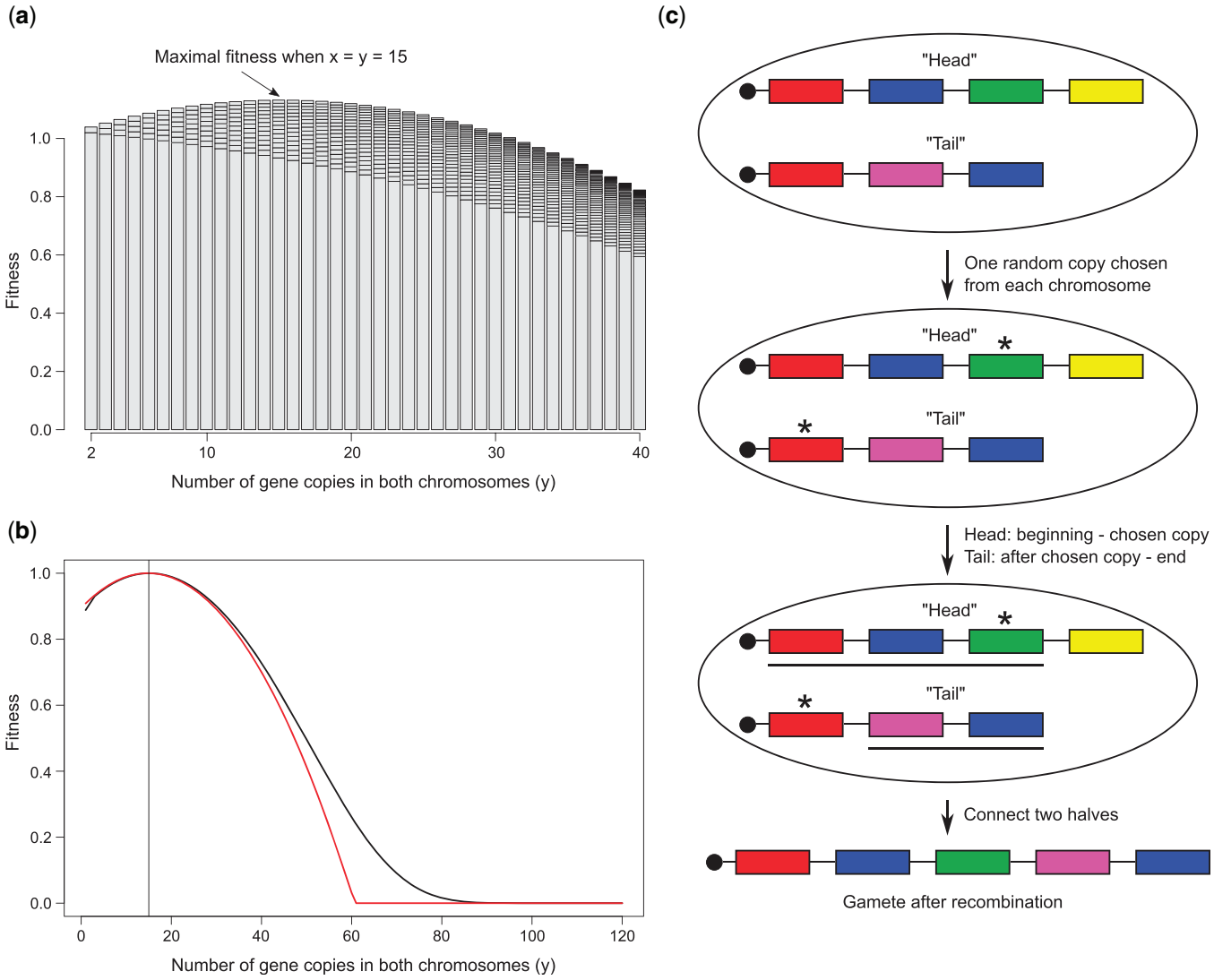

Fig. 1. a) Fitness of an individual as a function of x (stacks) and y (bars). Parameters: $s_x = 0.02$, $s_y = 0.005$, $\beta_x = 0.95$, $\beta_y = 1.05$. Each bar represents one value of y with stacked fitness "layers" for $x = 1$ to $x = y$. b) Normalized fitness of an individual in the y-only model. Parameters: $s_x = 0.02$, $s_y = 0.005$, $\varepsilon = 0.05$ (black) and its Taylor-approximated version $\mathcal{T}(y) = 1 - \tilde{s}(y - y^*)^2$, with $\tilde{s} \approx 0.00047$ (red). The vertical line marks $y^* \approx 14.86$. c) Illustration of individual genotype unequal recombination. Recombination occurs in an individual with $y = 7 = 4 + 3$ gene copies and $x = 5 < 4 + 3$ different alleles (colors). The black bullet on each chromosome represents the RRM locus (see text).

case, randomly mark a gene copy on both chromosomes. Then, the "upstream" fragment *in*cluding the marked copy of chromosome m ("head"), say, is fused with the "downstream" fragment *ex*cluding the marked copy of chromosome p ("tail"). For simplicity, we assume recombination break points to lie outside of genes and exclude the possibility that genes may be disrupted by recombination. Only one recombination product is considered further. If the last copy was marked on the tail chromosome, no copy is added to the head fragment. Starting from 2 chromosomes with $y^{m\prime}$ and y'^{p} copies, copy number in the offspring gamete can range between 1 and $y^{m\prime} + y'^{p} - 1$. More precisely, copy number in the offspring chromosome is a sum of uniform random variables with

$$Y' = B_1 + B_2 - 1,$$

where $B_1 \sim U(y^{m\prime})$, $B_2 \sim U(y'^{p})$ are uniform on the integers $\{1, \ldots, y^{m\prime}\}$ and $\{1, \ldots, y'^{p}\}$, respectively. The sum Y' is trapezoidal with

$$P[Y' = y' \mid y^{m\prime}, y'^{p}] = T(y', y^{m\prime}, y'^{p})$$
$$= \frac{1}{y^{m\prime} \cdot y'^{p}} \begin{cases} 0, & y' \le 0 \\ y', & 1 \le y' \le (y^{m\prime} \wedge y'^{p}) \\ (y^{m\prime} \wedge y'^{p}), & (y^{m\prime} \wedge y'^{p}) \le y' \le (y^{m\prime} \vee y'^{p}) \\ y^{m\prime} + y'^{p} - y', & (y^{m\prime} \vee y'^{p}) \le y' \le y^{m\prime} + y'^{p} - 1 \\ 0, & y' \ge y^{m\prime} + y'^{p} \end{cases},$$

where $\wedge$ denotes the minimum and $\vee$ the maximum. When no recombination occurs, only one of the 2 parental chromosomes is propagated.

We also consider a version with recombination rate variation: assume that each chromosome carries a recombination rate modifier (RRM) locus, which encodes a chromosome-specific recombination rate. For a pair of chromosomes m and p, a recombination event occurs with rate $r = r_o\sqrt{(\rho_m \rho_p)}$ for modifier "alleles" ρ_m, $\rho_p > 0$, which are multipliers of the base recombination rate r_o. The modifier allele inherited to the recombination product is the geometric mean $\sqrt{\rho_m \rho_p}$. Note that selection, operating on the

genotype, exerts an indirect force on the recombination rate. Symbolically, the modifier locus is represented by a black bullet in Fig. 1c. It is itself not subject to recombination, but attached to the first gene copy. We set $r_o = 0.01$ in all simulations.

Simulations

For all simulations, we used an in-house developed R program (https://github.com/y-zheng/Recombination-gene-family) implementing a Wright–Fisher type model with discrete generations and multinomial sampling of gametes. Simulation raw data can be downloaded from the same repository. Simulations consisted of a burn-in phase and an observation phase in which the statistics shown in Table 1 were recorded at certain time intervals. We considered 4 basic scenarios:

a) single population with constant size N;
b) single population with bottleneck;
c) two subpopulations with reciprocal migration; and
d) single population of constant size with RRM.

Simulations for scenario (a) were started with $y = 10$ and $x = 1$ for all i and run for an initial burn-in phase of 20,000 generations. A run was restarted in case it entered the (absorbing) state $y = 2$ during burn-in, i.e. when all individuals have only a single copy on each chromosome. To start simulations in scenarios (b)–(d), we used the final state, which was reached at the end of scenario (a). To reduce standard error of the mean of this final sampling point, we ran 500 replicates for scenario (a) and 200 replicates for scenarios (b)–(d). For the simulations, we selected parameter ranges which we considered realistic and which turned out to be compatible with the estimates for s_x, s_y, and r and the mean copy number obtained from empirical data (see below). The parameters used in the different scenarios are listed in Table A1 in the Appendix.

Empirical data

Based on data from the pilot phase of the 1000 Genomes project, Brahmachary *et al.* (2014) analyzed CNV in 193 gene families and microsatellite loci in 3 human populations (CEU, CHB, and YRI). We chose 3 representative examples [pregnancy-specific glycoprotein 3 (*PSG3*), Mucin 12 (*MUC12*), and proline-rich protein 20A (*PRR20A*)], which satisfied the following criteria:

- genes tandemly arrayed;
- genes autosomal;
- mean copy number between 10 and 20; and
- one example each with small, intermediate and large copy number variance.

PSG3 is located on the long arm of the particularly gene-rich chromosome 19 (Grimwood *et al.* 2004). It is a member of the carcinoembryonic antigen gene family and of the immunoglobulin superfamily and is involved in pregnancy maintenance. *MUC12* is a membrane glycoprotein of the mucin family. Mucins are involved in mucous protection, epithelial cell differentiation, and intracellular signaling and have been recognized having similar evolutionary features as *HLA* genes (Vahdati and Wagner 2016). *PRR20A* is a predicted gene located on the long arm of chromosome 13. It has low Uniprot annotation score with experimental evidence only at transcript level (https://www.uniprot.org/uniprot/P86496).

The available empirical data from this data set can be analyzed in the context of the y-only model. To estimate the underlying parameters (s_x, s_y, and r) of the y-only model that best describe the empirical copy number distribution, we implemented an *EM*-like grid search as follows: we use the data from the African (YRI) population, assuming that it is closest to recombination-selection-drift equilibrium and least affected by a recent population bottleneck (e.g. Rafajlović *et al.* 2014; Schiffels and Durbin 2014; Spence and Song 2019). Individual copy numbers are derived from the data published by Brahmachary *et al.* (2014) and calculated by dividing the individual read ("nanostring," in the authors' terminology) counts by the average read count per copy (https://github.com/y-zheng/Recombination-gene-family). This way, we found for *MUC12*, *PSG3*, and *PRR20A* mean numbers of, respectively, 11.85, 14.94, and 19.85 copies per individual in the YRI population (diploid sample size $n = 60$). To compare these results with our model, we uniformly sampled 5,000 parameter combinations of independently chosen s_x, s_y, and r from the product of initial intervals $[1e-6, 5e-2]^3$. For each parameter combination, we calculate the Gamma approximation of the equilibrium distribution of the y-only model (see *Results*) and use the Kolmogorov–Smirnov (KS) test to calculate the probability that the data are sampled from this distribution. We choose the top 100 (= 2%) parameter combinations to define the range of the new parameter intervals to sample from. In each

Table 1. Summary statistics recorded in simulations.

	Mean[a]	Std. Dev.	Min.	Max.
Individual statistics				
Copies	$\overline{\mathbf{y}} = (\sum_i y_i)/N_e$	$\sigma_{\mathbf{y}}$	$\min_{\mathbf{y}}$	$\max_{\mathbf{y}}$
Alleles	$\overline{\mathbf{x}} = (\sum_i x_i)/N_e$	$\sigma_{\mathbf{x}}$	$\min_{\mathbf{x}}$	$\max_{\mathbf{x}}$
Ratio	$\overline{\mathbf{x}/\mathbf{y}} = (\sum_i \frac{x_i}{y_i})/N_e$	$\sigma_{\mathbf{x}/\mathbf{y}}$	$\min_{\mathbf{x}/\mathbf{y}}$	$\max_{\mathbf{x}/\mathbf{y}}$
Fitness	$\overline{\boldsymbol{\omega}} = (\sum_i \omega_i)/N_e$	$\sigma_{\boldsymbol{\omega}}$	$\min_{\boldsymbol{\omega}}$	$\max_{\boldsymbol{\omega}}$
Population statistics				
Total number of copies in population[b]	$\lvert\mathbf{y}\rvert$			
Total number of *different alleles*[b]	$\lvert\mathbf{x}\rvert$			
Absolute frequency of alleles[c]	$m_j, j = 1, \ldots, \lvert\mathbf{x}\rvert$			
Relative frequency of alleles	$\xi_j = \frac{m_j}{2N_e}, j = 1, \ldots, \lvert\mathbf{x}\rvert$			
Effective number of alleles[d]	$\lvert\mathbf{x}\rvert_{\text{eff}} = \left(\sum_{j=1}^{\lvert\mathbf{x}\rvert} \left(\frac{m_j}{\lvert\mathbf{y}\rvert}\right)^2\right)^{-1}$			

[a] Sums are taken across all individuals $i = 1, \ldots, N_e$.
[b] Note that $\lvert\mathbf{y}\rvert = N_e\overline{\mathbf{y}} = \sum_i y_i$. In contrast, $\lvert\mathbf{x}\rvert \leq N_e\overline{\mathbf{x}} = \sum_i x_i$. The inequality is strict as soon as different individuals share alleles.
[c] That is, m_j is the number of occurrences of allele j in the entire population. We assume that alleles are labeled in decreasing frequency: $m_j \geq m_k$ for all $j < k$.
[d] Note that $\lvert\mathbf{x}\rvert_{\text{eff}}$ is the inverse Simpson index of diversity.

iteration, parameter intervals are shrinking and we terminate this process after 10 iterations to obtain a possibly small range of the final parameter combinations with highest KS P-value. We then chose the best parameter combinations for further analysis. The range of these parameters is shown in Supplementary Fig. 1. Epistasis is kept fixed at $\varepsilon = 0.05$ during the entire search.

Results

y-only model

Consider first the y-only model. Each copy is considered a unique and distinct allele. Therefore, at any time, $x_i = y_i \,\forall i$, and fitness of an individual is a function only of y:

$$\omega = \omega(y_i) = (1+s_x)^{\left(\sum_{k=0}^{y_i-1}\beta_x^k\right)} \times (1-s_y)^{\left(\sum_{k=0}^{y_i-3}\beta_y^k\right)}.$$

for all individuals i.

Let y' be the number of gene copies on a single chromosome, without regard of parental status, and let $p_t(y')$ be the frequency of chromosomes with y' copies in an infinitely large population in generation t.

Choosing parental chromosomes according to their fitness $\omega(y = y^{m\prime} + y'^{p})$, the frequency of y' changes to

$$p_{t+1}(y') = (1-r)\sum_{y'^p} q_t(y', y'^p) + r\sum_{y^{m\prime}, y'^p} q_t(y^{m\prime}, y'^p)T(y', y^{m\prime}, y'^p), \quad (4)$$

where T denotes the trapezoidal distribution and

$$q_t(y^{m\prime}, y'^p) = \frac{p_t(y^{m\prime})p_t(y'^p)\cdot\omega(y^{m\prime}+y'^p)}{\overline{\omega}_t}$$

is the frequency of the pair $(y^{m\prime}, y'^p)$ after selection. In the last equation, $\overline{\omega}_t$ is mean population fitness at time t, i.e.

$$\overline{\omega}_t = \sum_{y^{m\prime}, y'^p} p_t(y^{m\prime})p_t(y'^p)\cdot\omega(y^{m\prime}+y'^p),$$

where the sum runs over all possible pairs $(y^{m\prime}, y'^p) \in \mathbb{N}\times\mathbb{N}$. Therefore, this process can be thought of as an irreducible aperiodic Markov chain on the state space $\{1, 2, \ldots\}$, which converges to its unique stationary distribution. Under neutrality ($\omega \equiv 1$), this simplifies to

Proposition 1. *Under (unequal) recombination and under neutrality it holds that*

- *the expected value of copy number remains constant over time, i.e.* $\forall t$

$$\sum_{y'=1}^{\infty} y'\cdot p_{t+1}(y') = \sum_{y'=1}^{\infty} y'\cdot p_t(y') = \cdots = \sum_{y'=1}^{\infty} y'\cdot p_0(y') =: E_{Y'}$$

- *the stationary distribution is given by the discrete kernel of the Gamma distribution with shape parameter* $\alpha = 2$ *and expected value* $E_{Y'}$*, i.e.*

$$p_{stat}(y') = y'\cdot\exp\left\{-\frac{2}{E_{Y'}}y'\right\}\cdot\frac{1}{Z}, \quad (5)$$

where Z *is the normalization constant given by*

$$Z = \sum_{y'} y'\cdot\exp\left\{-\frac{2}{E_{Y'}}y'\right\} = \exp\frac{\left\{\frac{2}{E_{Y'}}\right\}}{\left(\exp\left\{\frac{2}{E_{Y'}}\right\}-1\right)^2}.$$

The proof is given in the Appendix.

Hence, the neutral equilibrium distribution of copy numbers on individuals is given by the convolution

$$\begin{aligned}\tilde{p}_{stat}(y) &= \sum_{y'_1+y'_2=y} p_{stat}(y'_1)p_{stat}(y'_2)\\ &= \frac{1}{6}(y^3-y)\exp\left\{-\frac{1}{E_Y}y\right\}\cdot\frac{1}{Z^2},\end{aligned}$$

which is the discrete kernel of the Gamma distribution with shape parameter $\alpha = 4$ and expected value $E_Y = 2E_{Y'}$.

Adding selection to the process makes the analysis less straightforward. We note that the process described by equation (4) is still an irreducible Markov chain, which has a stationary distribution. However, determining a closed formula of p_{stat} is not easily feasible and we resorted to the following approximation.

We choose ω as defined in equation (1), assume that $|\mathbf{x}| = |\mathbf{y}|$ (y-only model) and that $\beta_x = 1-\varepsilon$ and $\beta_y = 1+\varepsilon$ for some $\varepsilon > 0$. Thus, the fitness function simplifies to

$$\begin{aligned}&\omega(y) = \exp\{f(y)\}, \text{ where}\\ &f(y) = \frac{s_x+s_y}{\varepsilon} - \frac{s_x}{\varepsilon}\cdot e^{-\varepsilon y} - \frac{s_y}{\varepsilon}\cdot e^{\varepsilon(y-2)}.\end{aligned} \quad (6)$$

The Taylor expansion up to order 2, evaluated at y^* and scaled with $\omega(y^*)^{-1}$ is

$$\begin{aligned}\mathcal{T}(f(y)) &= \frac{1}{\omega(y^*)}\left(\omega(y^*) + \frac{\mathrm{d}}{\mathrm{d}y}\omega(y^*)(y-y^*) + \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}y^2}\omega(y^*)(y-y^*)^2\right)\\ &= 1 + \frac{1}{2}\frac{\mathrm{d}^2 f}{\mathrm{d}y^2}(y^*)\cdot(y-y^*)^2\\ &= 1 - \varepsilon e^{-\varepsilon}\sqrt{s_x s_y}\cdot(y-y^*)^2.\end{aligned}$$

Note, that this coincides with the fitness function introduced by Krüger and Vogel (1975)

$$\tilde{\omega}(y) = 1 - \tilde{s}(y-y^*)^2, \quad (7)$$

when substituting

$$\tilde{s} = \varepsilon e^{-\varepsilon}\sqrt{s_x s_y}.$$

Hence, the quadratic distance of y from the optimal copy number y^* determines fitness. It fits well with our definition of synergistic epistasis when y is not too far from y^* (see Fig. 1b) and yields a threshold $y^o = y^* + 1/\sqrt{\tilde{s}}$ with $\mathcal{T}(f(y)) < 0$ for $y > y^o$.

Therefore, with this quadratic approximation of the fitness function, equation (4) becomes a finite system of equations, which can be numerically solved with standard iteration algorithms. Starting with an arbitrary initial distribution we iterate until

$$||p_{t+1} - p_t||_{TV} := \sum_{y'} |p_{t+1}(y') - p_t(y')| < 0.001,$$

where $||\cdot||_{TV}$ denotes the *total variation* and the sum runs from 1 to the maximal y' given by y^o. After convergence, we calculate the

copy number distribution on *individuals* as convolution of the copy number distribution on chromosomes.

For fixed parameters, the process converges to the same limiting distribution, independently of initial conditions. Varying the recombination rate leads to different limiting distributions: it is close to the neutral stationary distribution when r is large; it is sharply peaked, and centered at y^*, when r is small. The variance is almost vanishing when $r < 0.01\,s_x$. Increasing selection shifts $\bar{y}$ toward y^*. Generally, the stationary distribution is determined by a balance of recombination and selection and the relative magnitudes of r, s_x, and s_y. Visual inspection of the limiting distribution for various parameter choices suggests that it is well approximated by a Gamma distribution also in the non-neutral case (see, for instance, the 3 examples shown in Fig. 3, lines in blue). We estimate its parameters as follows.

We numerically solved the system of equations [equation (4)] for about 50,000 random parameter combinations. We kept $\varepsilon = 0.05$ constant and chose $r \in [0, 0.01]$, $s_x \in [0, 0.05]$ and s_y such that $s_x/s_y \in [2.5, 18]$, producing an optimal copy number y^* between 10 and 30. Then, we calculated mean and variance of the equilibrium distribution for all parameter combinations. Assuming that the expectation (E_Y) of the limiting Gamma distribution is determined by equation (3), we set

$$\hat{E}_Y = y^* = \frac{\ln(s_x) - \ln(s_y)}{2 \cdot 0.05} + 1.$$

Assuming $r > 0.01 \cdot s_x$ and that its standard deviation scaled by the mean (σ/E_Y) depends on recombination–selection balance, $\ln(r/s_x)$, we obtain by linear fitting (Fig. 2a):

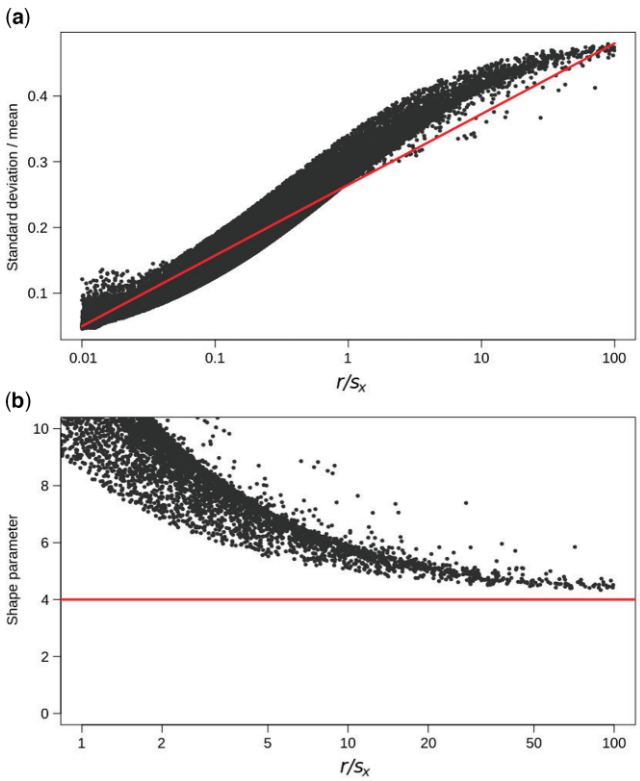

Fig. 2. a) Linear fit of σ/E_Y on $\ln(r/s_x)$ (for details see text). Note the strong correlation of $\ln(r/s_x)$ and σ/E_Y, with a Pearson correlation coefficient of $\rho = 0.97$. The estimated regression line $\sigma/y^* = 0.046 \cdot \ln(r/s_x) + 0.26$ is shown in red. b) Convergence of the Gamma shape parameter $\alpha = (E_Y/\sigma)^2$ toward the value $\alpha = 4$, expected under neutrality, when r is increasing or s_x is decreasing.

$$\hat{\sigma} = y^* \cdot (0.046 \cdot \ln(r/s_x) + 0.26).$$

Furthermore, $(E_Y/\sigma)^2$ converges toward the shape parameter ($\alpha = 4$) of the Gamma distribution under neutrality, when selection becomes small or recombination becomes large (Fig. 2b). Therefore, for given parameters r, s_x, s_y, and $\varepsilon = 0.05$, we use the discrete kernel of the Gamma distribution with shape parameter $\alpha = (y^*/\hat{\sigma})^2$ and expected value y^* as an approximation of the equilibrium distribution of the y-only process with selection. Note that the distribution is uniquely determined by its shape and mean.

Application of the y-only model to empirical data

To estimate selection coefficients and rates of unequal recombination for the 3 gene families *PSG3*, *MUC12*, and *PRR20A*, we used the *EM*-like grid search described above. We calculated the KS-test *P*-value for 3 distributions: (1) a neutral equilibrium distribution $\tilde{p}_{stat}$ with mean value given by the arithmetic mean of the data, (2) one of the best-fitting Gamma distributions with parameters given by the *EM*-like grid search, and (3) the equilibrium distribution of the y-only process with the same recombination and selection coefficients as obtained from the grid search. Sufficiently, small *P*-values indicate a significant difference from any of the 3 models, whereas a *P*-value close to one can be interpreted as a good approximation of the data. The results are given in Table 2 and Fig. 3. Distributions of the 100 best parameter combinations for each gene are shown in Supplementary Fig. 1. For *PSG3*, the empirical distribution of copy numbers (histogram in Fig. 3, top) is well approximated by a Gamma distribution (red line) yielding a KS-test *P*-value of 0.99. The limiting distribution under the y-only model still fits fairly well with $P = 0.82$ (blue line). In contrast, the hypothesis of neutrality can be clearly rejected: the neutral Gamma distribution [equation (5)] produces a *P*-value of 1.4e − 9 (black line). The parameter estimates suggest a small recombination rate of about 0.1% per generation per gamete and strong selection ($s_x = 0.04$ and $s_y = 0.01$), maintaining copy number close to its optimal value. Although the gene family *PRR20A* is much more variable than *MUC12* (Fig. 3, middle and bottom), we estimate the same recombination rate of about 0.8% for both families. However, the difference in their distributions can be explained by different selection strengths. The estimates in *MUC12* are $s_x = 0.017$ and $s_y = 0.006$—about half as strong as in *PSG3*. In contrast, the estimates in *PRR20A* are $s_x = 0.001$ and $s_y = 0.00028$, lower by roughly a factor of 40 than in *PSG3*. While neutrality can still be clearly rejected in *MUC12* ($P = 0.0012$), it cannot be rejected in *PRR20A*. Still, also for this gene family, pure neutrality has a much lower explanatory power than do have models with selection ($P = 0.217$ vs $P = 0.98$). One should keep in mind, however, that the above estimates depend on our choice of the epistasis parameter $\varepsilon = 0.05$. From equation (3), it is clear that the ratios s_x/s_y and ε are inversely related. In work dedicated to data analysis, rather than model development, one may want to include ε (or even β_x and β_y separately) among the parameters to be estimated.

Simulation results of the compound model

In scenario (a), we analyzed the effect of different population sizes, selection strengths (a1) and recombination rates (a2) on the statistics of Table 1 at equilibrium. In scenario (a1), we used $s_x = 0.01$, 0.02, 0.04 (weak, medium, and strong selection), with $s_x/s_y = 4$ and $\varepsilon = 0.05$. These parameters were chosen such that the optimal genotype for an individual is $x = y = 15$ in all 3 selection regimes. Population size varied from $N_e = 500$, 1,000, 2,000 to

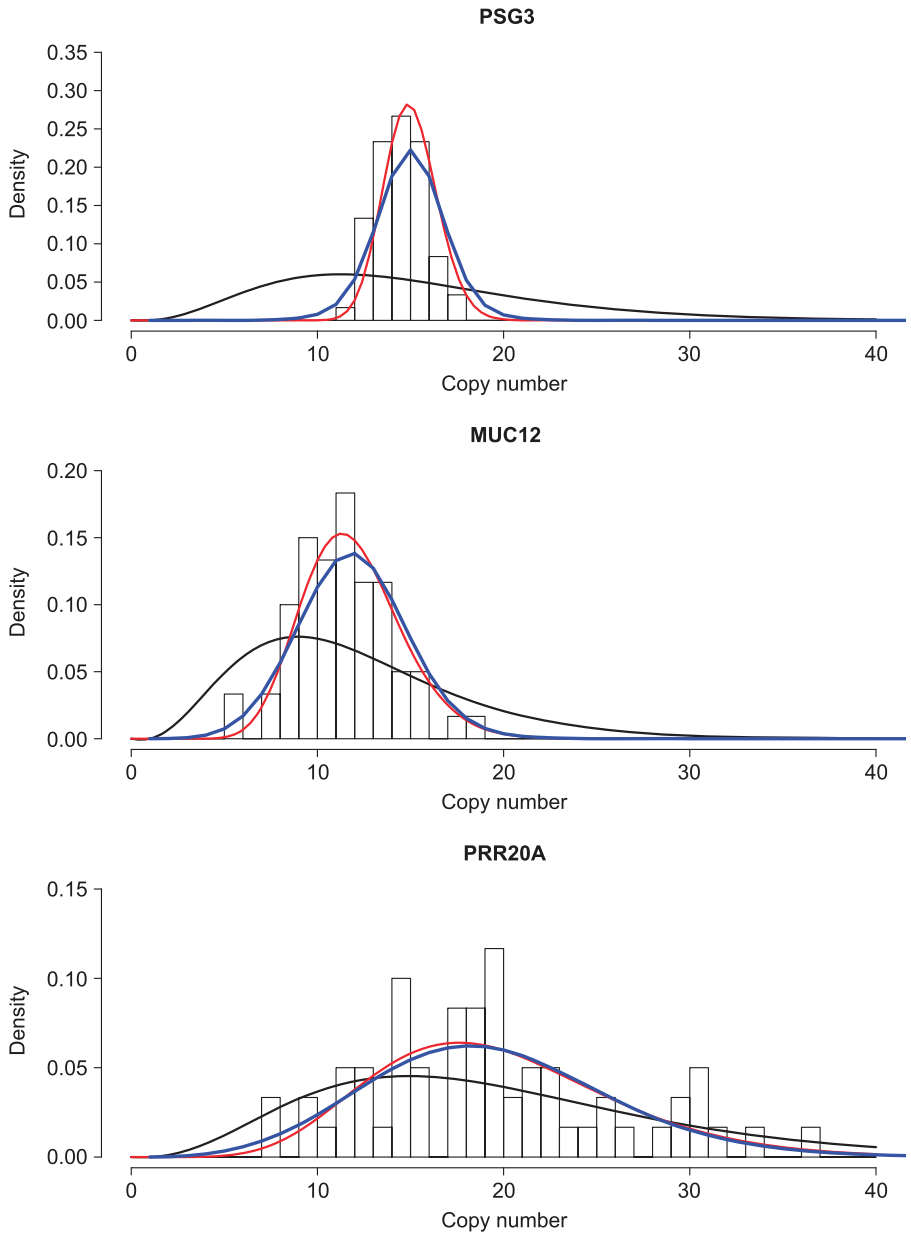

Fig. 3. Copy number distribution of 3 different human genes and their approximations. Black: Copy number distribution under neutrality $\tilde{p}_{stat}$ with E_Y =14.94, 11.85, and 19.85 for *PSG3*, *MUC12*, and *PRR20A*, respectively. Red: Gamma distribution with parameters given in Table 2, resulting in best KS-test P-value. Blue: Equilibrium distribution of the y-only model generated from equation (4) with parameters as in Table 2.

Table 2. Parameter estimates for empirical data obtained by *EM* grid search, with fixed $\varepsilon = 0.05$, that returned the best KS P-value for the Gamma approximation.

Gene family	Estimated parameters	P-value of KS-test		
		Neutral	Gamma	y-only
PSG3	$r = 0.001$ $s_x = 0.04$ $s_y = 0.01$	1.4e − 9	0.99	0.82
MUC12	$r = 0.008$ $s_x = 0.017$ $s_y = 0.006$	0.0012	0.99	0.98
PRR20A	$r = 0.008$ $s_x = 0.001$ $s_y = 0.00028$	0.217	0.98	0.98

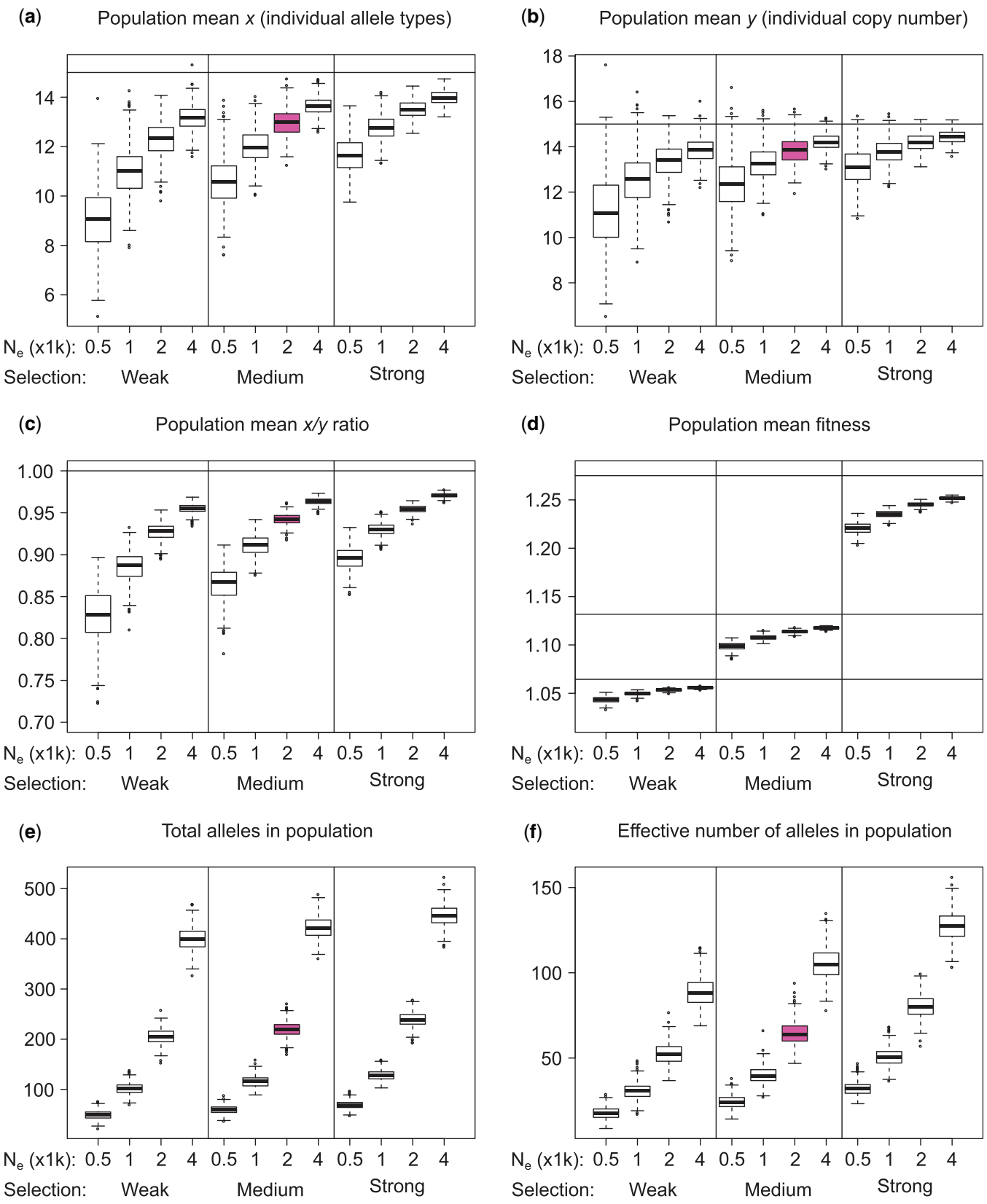

Fig. 4. Scenario (a1)—constant population size. Population statistics at equilibrium: population mean $\overline{\mathbf{x}}$ (a); population mean $\overline{\mathbf{y}}$ (b); $\overline{\mathbf{x}}/\overline{\mathbf{y}}$ ratio (c); population mean fitness (d); total number (e), and effective number of alleles $|\mathbf{x}|_{\text{eff}}$ (f). Varying parameters: population size N_e and selection coefficient s_x. Mutation ($\mu = 0.0005$) and recombination rate ($r = 0.01$) are kept fixed. Boxplots based on 500 independent replicates. Box colored in purple indicates a parameter combination ($N_e = 2{,}000$, $r = 0.01$, $s_x = 0.02$, $s_y = 0.005$) shared by scenarios (a), (b), (c), and (d). Horizontal lines in a–c indicate the optimal copy number in the y-only model. Horizontal lines in D indicate optimal fitness.

4,000 and recombination rate was kept constant at $r = 0.01$. Results are shown in Fig. 4.

Both larger population sizes and stronger selection lead to an increase in population means $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ (Fig. 4, a and b). Note, that the demographic effect (decrease of drift by increase of population size) on these quantities is much stronger than the effect by increasing selection. Both $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ are always below the optimal value of 15. However, doubling N_e has a stronger effect than doubling selection strength in bringing the population closer to the optimal value. Essentially the same pattern is observed for the ratio $\overline{\mathbf{x}/\mathbf{y}}$ (Fig. 4c). For example, $N_e = 1{,}000$, 2,000, 4,000 with low selection leads to a higher ratio $\overline{\mathbf{x}/\mathbf{y}}$ than $N_e = 500$, 1,000, 2,000 with intermediate selection. The total (Fig. 4e) and the effective (Fig. 4f) number of alleles scale roughly linearly with N_e. Again, both quantities depend more strongly on population size than on selection strength. This effect is more pronounced in the total number of alleles than in $|\mathbf{x}|_{\text{eff}}$, which is explained by drift: alleles at low frequency, in particular newly generated alleles ($N_e\mu\overline{\mathbf{y}}$ per generation), are prone to loss when drift is strong. They count for the total number, but contribute little to $|\mathbf{x}|_{\text{eff}}$. In contrast, mean fitness is more affected by the strength of selection than by N_e. This is because mean fitness depends on 2 ingredients: the equilibrium distribution y itself and the weights ω_i of its components. Both are altered by selection. Finally, the frequencies of the most common alleles (Supplementary Fig. 2) are negatively correlated both with N_e and s_x. In summary, allelic diversity at population scale appears to be driven mainly by N_e.

In scenario (a2), we kept selection at intermediate level ($s_x = 0.02, s_y = 0.005$) and varied the rate of (unequal) recombination from $r = 0.002$ to 0.05. Results are shown in Fig. 5. Increasing recombination decreases $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$, as well as the ratio $\overline{\mathbf{x}/\mathbf{y}}$. Therefore, it also decreases mean fitness $\overline{\omega}$. Recombination acts here in a similar way as drift: doubling the recombination rate has the same effect on fitness as halving the population size. This observation can be interpreted as a recombination load: frequent recombination can generate chromosomes whose copy number is far away from the optimum. Deviation from the optimal copy number has an asymmetric effect because of epistasis: a surplus of copies is more harmful than a deficit (Fig. 1b), explaining the somewhat counter-intuitive effect that increasing the recombination rate decreases both total and effective number of alleles in the population.

In scenario (b), we explored the impact of a single instantaneous and short bottleneck. Starting with an equilibrated panmictic population of constant size $N = 2{,}000$, population size was reduced to 1% (= 20) for 5, 10, or 20 generations, then restored to its original value N and the generation counter reset to $t = 0$. After that, simulations are carried on for another 10,000 generations during which the recovery process of the 6 summary statistics mentioned above is recorded. Results for different selection strengths are summarized in Fig. 6. A longer period of population size reduction results in populations with lower $\overline{\mathbf{x}}$ and lower $\overline{\omega}$. In contrast, length of the reduction period hardly affects $\overline{\mathbf{y}}$. Recovery time correlates positively with the length of the reduction period.

We observed that $\overline{\mathbf{y}}$ and, to a lesser extent, $\overline{\mathbf{x}}$ experience a decrease *after* the restoration of population size, and before it returns to its constant equilibrium value. Furthermore, the total number of alleles recovers much faster than the effective number. The reason is that new alleles are quickly created by mutation, but—while rare—they continue to bias the effective number of alleles, before equilibrium frequencies are restored. By segmental regression, we found that mean fitness recovers faster than $|\mathbf{x}|_{\text{eff}}$ (Supplementary Fig. 3, a and b). Furthermore, populations under stronger selection recover faster. The variation of these statistics among replicates is shown in Supplementary Fig. 4. Except for total and effective number of alleles, all other statistics show little among-replicate-variation after about 500 to 1,000 generations after the bottleneck. Variation of the total number of alleles reaches a plateau and then gradually decreases, while among-replicate-variation of $|\mathbf{x}|_{\text{eff}}$ is generally small.

In scenario (c), we studied the effect of population subdivision and migration. We simulated reciprocal migration with 2 subpopulations of equal size, small ($N = 500$) and intermediate ($N = 1{,}000$), starting from pairs of independent equilibrated replicates from scenario (a). Then, time was reset to $t = 0$ and migration was turned on with rates $Nm = 0.1$, 1, or 10 individuals per generation per direction. Summary statistics $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$, mean fitness $\overline{\omega}$, total number of alleles, and $|\mathbf{x}|_{\text{eff}}$ in the combined superpopulation were recorded over time. After about 1,500 to 2,000 generations, these statistics approached a migration-drift-selection equilibrium, which is between the means for the panmictic populations of size $N_e = 1{,}000$ and $N_e = 2{,}000$. While the scenario with high migration ($Nm = 10$) is almost indistinguishable from the panmictic population with respect to $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$ and $\overline{\omega}$ (Fig. 7, a–d), there is still a clear deficit in the total and effective number of alleles compared to the panmictic population, even when the migration rate is high (Fig. 7, e and f). Note also in this case, the initial overshooting of the panmictic equilibrium in the statistics $\overline{\mathbf{x}/\mathbf{y}}$, $\overline{\omega}$ and $|\mathbf{x}|_{\text{eff}}$ at about 100–200 generations, which is reminiscent of transient "hybrid vigour." Variation of these statistics among population replicates does not change appreciably with time (Supplementary Fig. 5). Similar results are observed for small populations $N_e = 500$ (Supplementary Figs. 6 and 7).

In scenario (a2), we observed that lower recombination rates lead to an equilibrium of $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ which are closer to the optimum. A natural question to ask is whether the recombination rate itself maybe subject to selection. Therefore, in scenario (d), an RRM was added to the simple model. Given an equilibrated population which was reached with $r = 0.01$ as described in scenario (a), recombination rate modification was switched on, and time reset to $t = 0$. Recombination rate was coded by an RRM allele, which can increase or decrease the current recombination rate by a factor $e^{\pm 0.05}$ when mutated. Modification happens per chromosome per generation each with probability $P = 0.002$ for increase or for decrease. The RRM locus is thought to reside on the tip of a chromosome without itself being affected by recombination (Fig. 1). Simulations were carried on for 50,000 generations and runs for each parameter setting of (s_x and s_y) were replicated 200 times. The results show that the mean recombination rate (average across all RRM alleles in the population) is continuously decreasing (Fig. 8). It decreases more and faster when selection (s_x and s_y) is strong. When simulations terminated, the recombination rate was reduced—on average—to 56%, 41%, and 31% of its original value ($r = 0.01$) and it showed a strongly negative correlation with population mean fitness (Pearson's $r = -0.75$, –0.83, –0.78) for weak, intermediate, and strong selection, respectively.

Discussion

We considered here a model in which 2 mechanisms, unequal recombination and mutation, may generate chromosomal diversity. While mutation leads to genetic diversity sensu strictu, by unequal recombination a chromosome may receive additional, or lose existing gene copies. Therefore, it is similar, but not identical, to segmental duplication or loss: copies gained by unequal

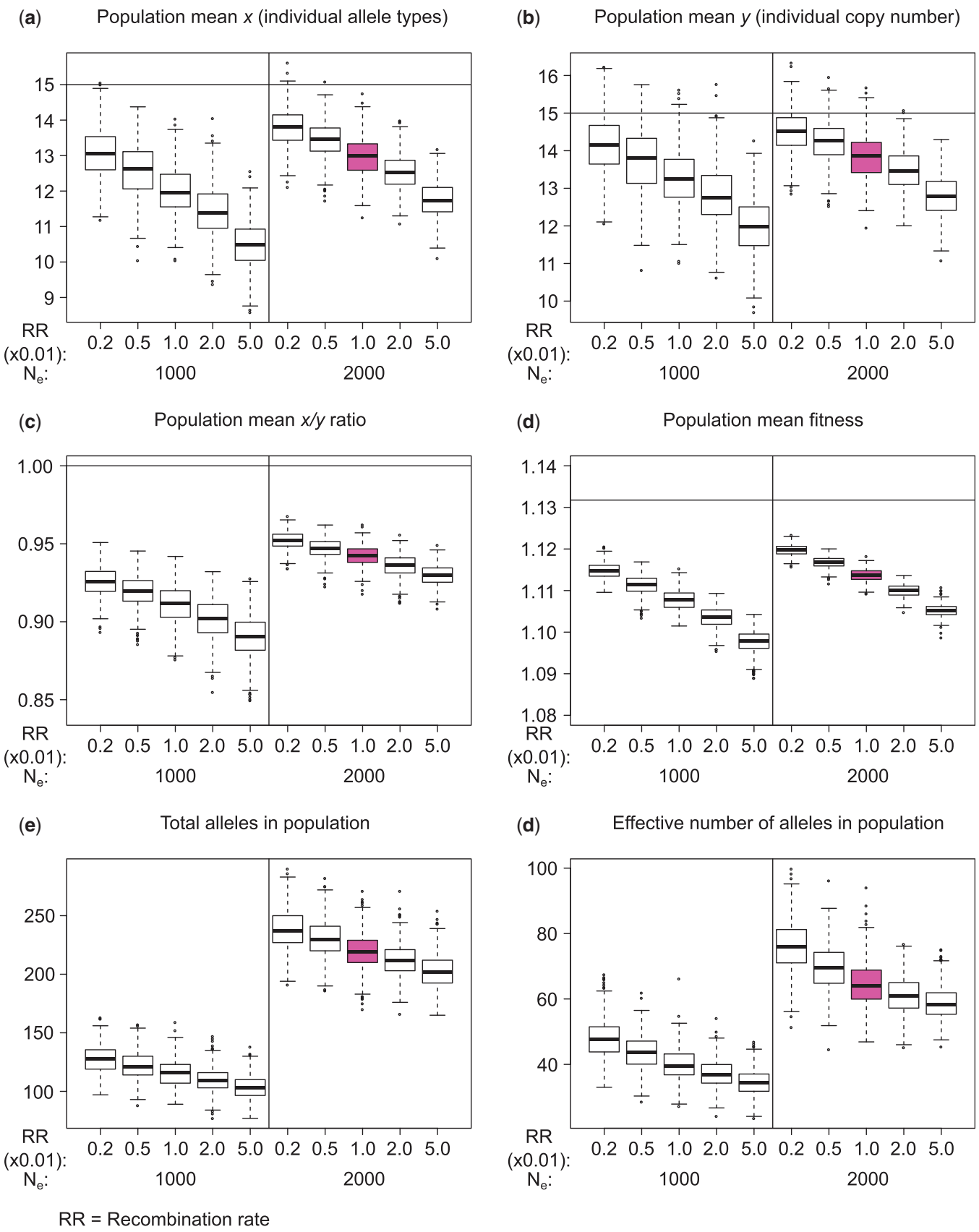

Fig. 5. Scenario (a2)—constant population size. Population statistics at equilibrium: population mean $\overline{\mathbf{x}}$ (a); population mean $\overline{\mathbf{y}}$ (b); $\overline{\mathbf{x}/\mathbf{y}}$ ratio (c); population mean fitness (d); total (e), and effective number $|\mathbf{x}|_{\text{eff}}$ (f) of alleles. Varying parameters: population size $N_e = 1{,}000, 2{,}000$ and recombination rate ($r = 0.01$ times the factor indicated on the abscissa). Mutation rate ($\mu = 0.0005$) and selection strength $((s_x, s_y) = (0.02, 0.005))$ are kept fixed. Boxplots based on 500 independent replicates. Box colored in purple indicates the parameter combination (see Fig 4) shared by scenarios (a), (b), (c), and (d). Horizontal lines as explained in Fig. 4.

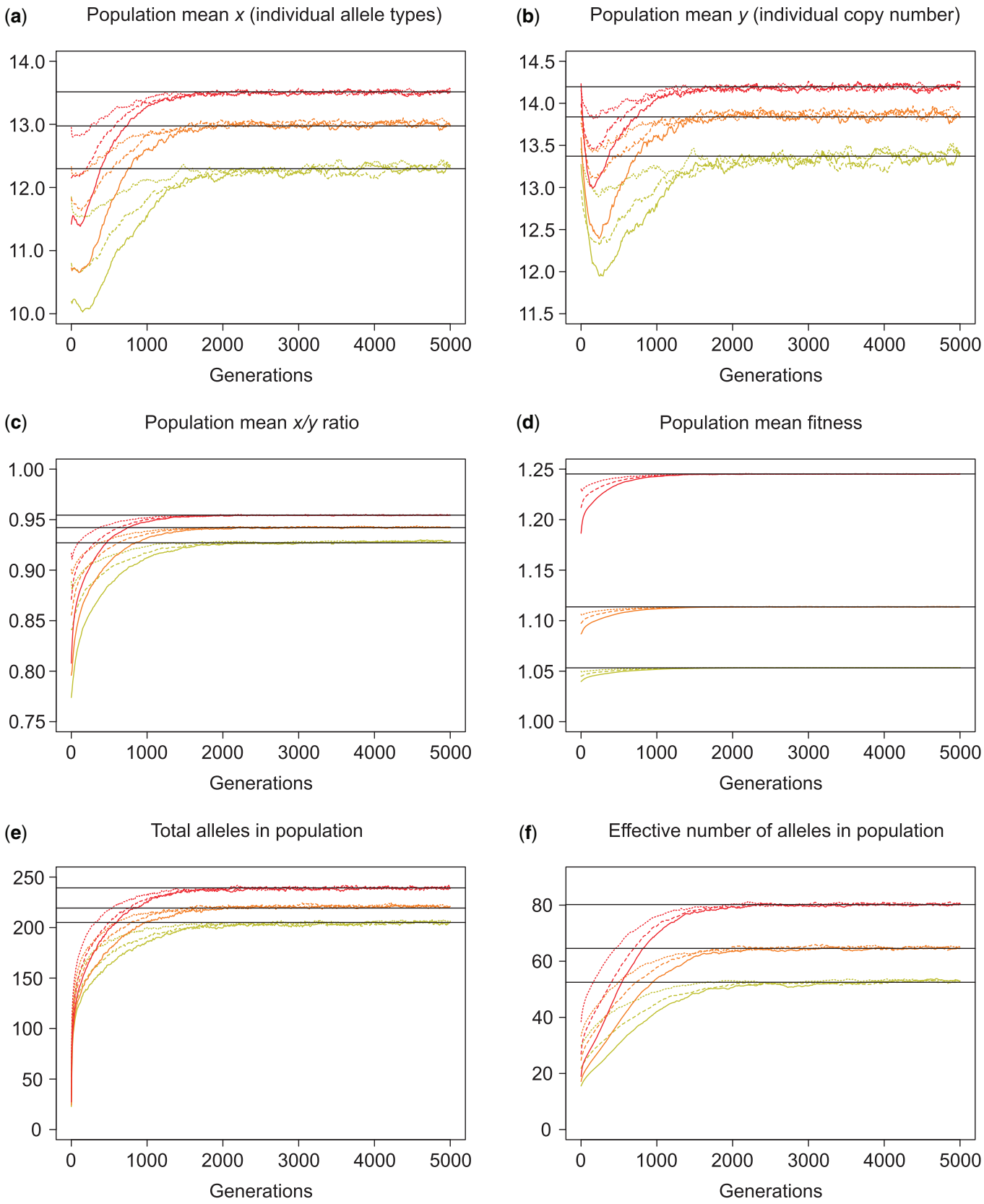

Fig. 6. Scenario (b)—recovery after a bottleneck. Equilibrium populations with $N = 2{,}000$ are reduced to $N_{red} = 20$ for a period of 5, 10, or 20 generations and then restored. During recovery, 6 statistics are traced. a) population mean $\overline{\mathbf{x}}$; (b) population mean $\overline{\mathbf{y}}$; (c) ratio $\overline{\mathbf{x}/\mathbf{y}}$; (d) mean fitness $\overline{\omega}$; (e) total number of alleles; and (f) $|\mathbf{x}|_{eff}$. Red, orange, and yellow indicate strong, intermediate, and weak selection. Solid, dashed, and dotted lines indicate bottleneck durations of 5, 10, and 20 generations. Each curve is an average across 200 replicates. Horizontal black lines are equilibria under constant population size.

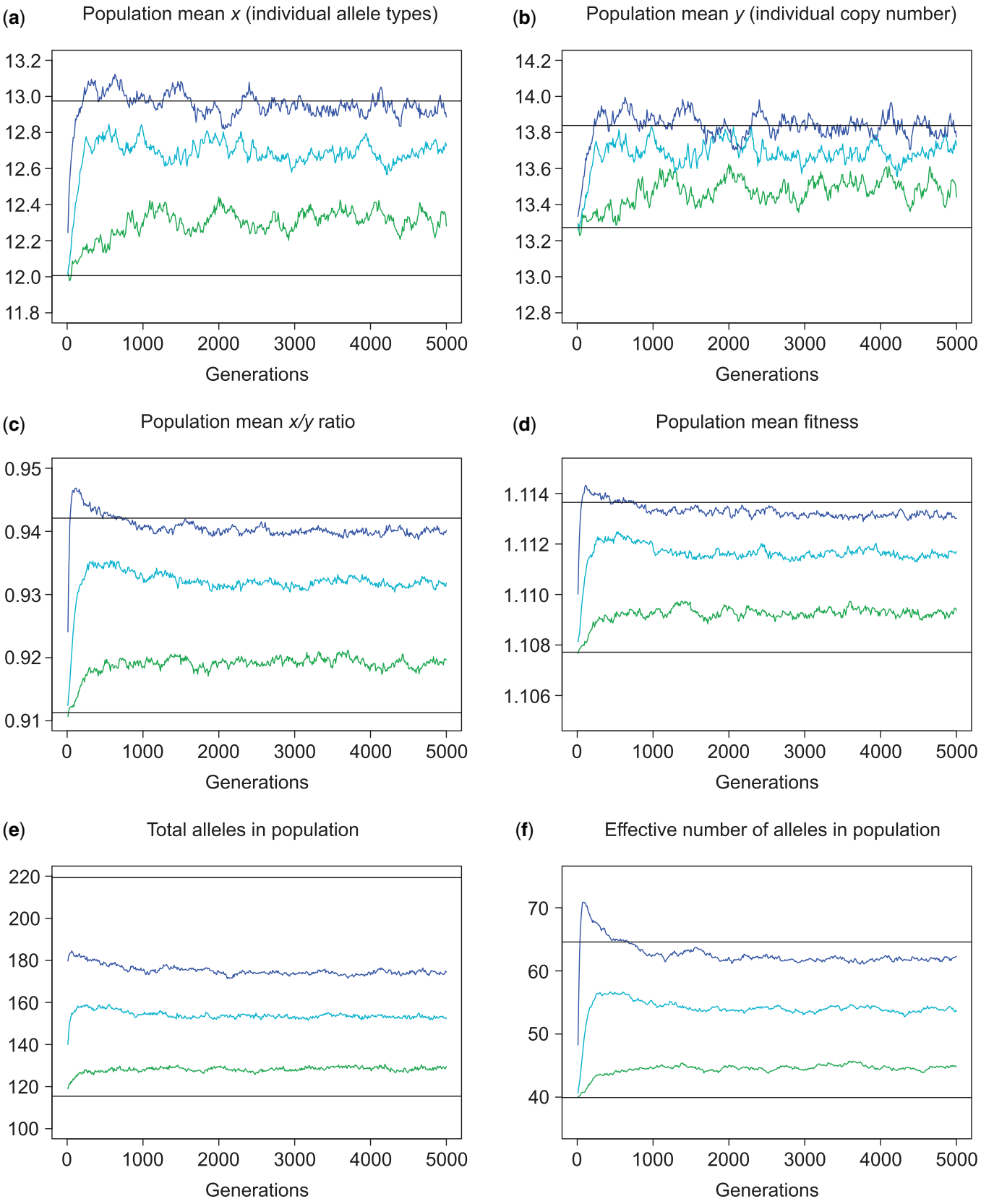

Fig. 7. Scenario (c)—migration. Two separated and equilibrated subpopulations of size $N = 1{,}000$ start to exchange migrants at time $t = 0$. Medium strength of selection ($s_x = 0.02, s_y = 0.005$). Migration rate: $2Nm = 0.1$ (green), 1 (cyan), or 10 (blue) migrants per generation in each direction. (a) population mean $\overline{\mathbf{x}}$; (b) population mean $\overline{\mathbf{y}}$; (c) ratio $\overline{\mathbf{x}}/\overline{\mathbf{y}}$; (d) population mean fitness $\overline{w}$; (e) total, and (f) effective number of alleles in the combined super-population. Shown are mean values across 100 replicates. Black lines indicate mean values (across 500 replicates) in panmictic populations of size $N_e = 1{,}000$ (lower line) and $N_e = 2{,}000$ (upper line).

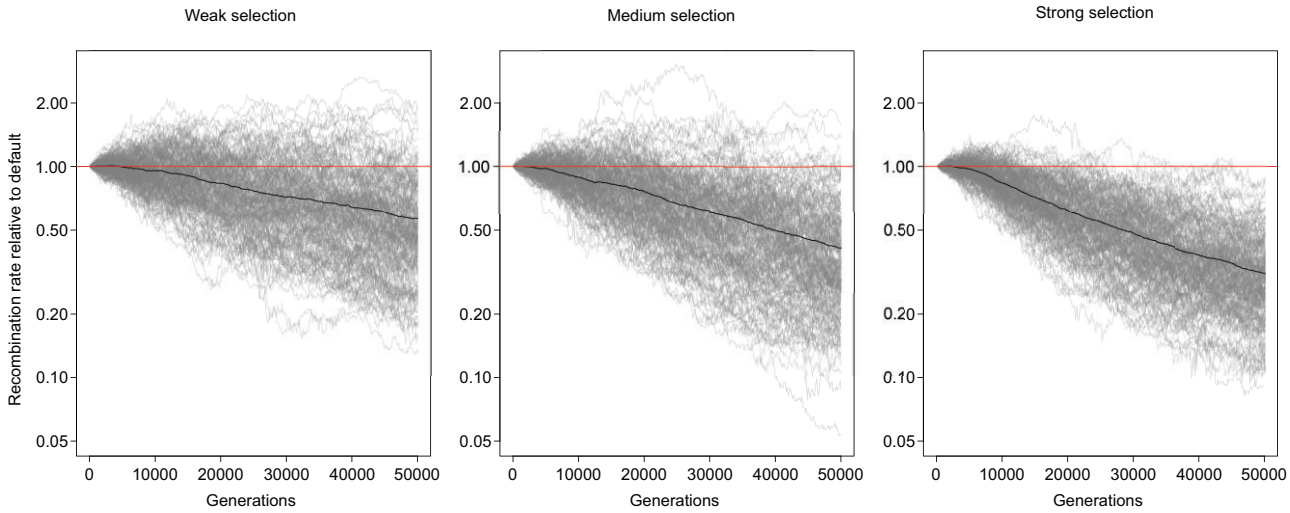

Fig. 8. Scenario (d)—RRM: recombination rate modification. Populations, which have reached equilibrium without RRM, are carried on for 50,000 generations during which the recombination rate, encoded at a modifier locus, may change under the influence of selection. For all iterations: $N_e = 2{,}000$, $r = 0.01$. Left: weak $(s_x, s_y) = (0.01, 0.0025)$; middle: intermediate $(0.02, 0.005)$; right: strong selection $(0.04, 0.01)$. Shown are trajectories of the recombination rate (in percentage of its original value $r = 0.01$) for 200 replicates each. The mean across all 200 replicates is shown as a black line.

recombination have their origin in a pairing haplotype, hence may be genetically diverse upon arrival, while those gained by duplication have their origin in the same haplotype, hence are genetically identical upon arrival. However, this distinction is negligible, since a single mutation event already suffices to make 2 identical copies distinct from each other when working in the context of the infinite alleles model. Another feature of our model is the 2 overlaid components of the fitness function: it decreases with copy number, but increases with allele number, entailing a subtle and very interesting interaction of recombination and selection.

To gain some analytical insight into copy number dynamics under recombination, we first considered the neutral case in an infinitely large population. We find copy number of individuals to be distributed according to the discrete kernel of a Gamma distribution with an equilibrium mean which is identical to the initial mean at time $t = 0$ and remains constant over time. The limiting shape parameter is $\alpha = 4$, which is identical for all initial configurations. These 2 properties together uniquely determine the limiting distribution, which is independent of the shape of the initial distribution and of the recombination and mutation rates.

Adding selection changes the game. The limiting distribution becomes dependent on both the recombination rate and the strength of selection, but independent from the initial configuration. Still, it is well approximated by a Gamma distribution. The distribution that results from low selection strength or high recombination converges to the neutral equilibrium.

We inferred selection and recombination parameters for 3 different human genes, under the assumption of fixed epistasis $\varepsilon = 0.05$. Our analysis shows that observed copy number distributions can be well approximated within the framework of our model. Different means and variances of the distributions can be explained in terms of higher or lower recombination rates and stronger or weaker selection.

Note, that compound fitness, in which allele diversity is credited, contains a component of balancing selection: an individual which is heterozygous at any given locus has a higher fitness than one which is homozygous at the same locus. An important difference between the model considered here and one-locus models of balancing selection is the existence of gene CNV and unequal recombination. Note that allelic diversity in the population can be stably maintained even in the case of allele fixation at single loci. The possibility to maintain allelic diversity through gene duplication, or unequal recombination, has been suggested by Haldane (1937). It is somewhat surprising that Haldane's idea has received only little attention in classical population genetics theory nor in experimental work. To our knowledge, tests confirming Haldane's hypothesis were conducted only a few years ago (Milesi *et al.* 2017).

We have shown that a high recombination rate has a negative effect on allelic diversity and resultant mean fitness. There are two reasons: (1) a higher rate of unequal recombination produces individuals with much higher or lower copy number than the optimum, which have reduced fitness; (2) low recombination increases the likelihood for highly unfit homozygotes to appear, thus improving the efficiency of selection.

Populations that experienced strong bottlenecks are at risk of inbreeding depression, and loci under balancing selection are particularly affected (Frankham *et al.* 2014). Random loss of alleles increases homozygosity and consequently reduces fitness. This can affect and delay the recovery of genetic diversity even after population size has recovered (Miller and Lambert 2004). In this study, we explored the effect of some parameters on the speed and process of bottleneck recovery at loci under diversifying selection. Both selection strength and bottleneck length influence the process. Relatively, longer bottlenecks produce a temporary reduction in $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$ and mean fitness. The most likely reason is that high homozygosity results in selection toward haplotypes with fewer copies. Selection is more powerful after, than during, the bottleneck, when population size has recovered, but copy number recovery may lag behind. However, this somewhat paradoxical effect of fitness reduction at the initial phase of bottleneck recovery is only a short-term effect, and—at least in part—due to the instantaneous, rather than gradual, restoration of population size in our model. Compared to fitness, $|\mathbf{x}|_{\text{eff}}$ is recovering even more slowly: for fitness to recover it suffices that new alleles appear and survive. But $|\mathbf{x}|_{\text{eff}}$ has recovered only when allele frequencies have reached their equilibrium values.

Therefore, $|\mathbf{x}|_{\text{eff}}$ is a more sensitive statistic to test for deviation from equilibrium.

Simulations of scenario (c) show that fitness under population subdivision with moderate migration reaches an equilibrium that is intermediate between those under panmixis on the one hand and complete isolation on the other. While a short boost of hybrid vigor exists, we do not see a positive effect from limiting migration compared to panmixis. An earlier simulation study (Schierup *et al.* 2000) showed that the allelic diversity is largely insensitive to migration rates, but low-migration scenarios result in alleles with more divergent sequences. Additionally, balancing selection in the form of heterosis could increase the effective migration rate because migrant haplotypes are more likely to be successful in this case than under neutrality (Ingvarsson and Whitlock 2000). Diversifying selection on MHC alleles has been shown to increase divergence between subpopulations, while diversity within subpopulations is still mostly governed by drift (Herdegen *et al.* 2014). MHC alleles and genes are also known to be shared among species through introgression, leading to restoration of diversity previously lost by drift (Dudek *et al.* 2019). In addition to generic balancing selection also local adaptation, i.e. the fixation of alleles that are adapted to specific subpopulations, may increase allelic diversity between populations (Ekblom *et al.* 2007). However, this effect is not considered in the model presented here, where selection operates only on the number of distinct alleles.

When the recombination rate is allowed to change over time, we observe a trend toward lower rates. It is driven by selection and happens on a realistic population genetic timescale of some thousand generations. However, there is little empirical knowledge about (unequal) recombination rates in multigene families. For example, in the human MHC locus, the recombination rate is only about a third of the average genomic background rate (de Bakker *et al.* 2006; Traherne 2008). On the other hand, studies on bovids (Schaschl *et al.* 2006) and horse (Beeson *et al.* 2019) show the opposite: high recombination in the MHC and olfactory receptor loci. In contrast again, the values reported for chicken seem to depend on mapping methodology (Fulton *et al.* 2016). Results from sheep (Petit *et al.* 2017) suggest a high "historical" (estimated from population data), but a low "meiotic" (from pedigree data) recombination rate, which suggests a recent change in time. From humans again, it is well known that recombination hotspots have a very fast turn-over time and are distinct in different subpopulations (Lam *et al.* 2013). Also, recombination rates may substantially differ in females and males—one example is the long arm of human chromosome 19 (Grimwood *et al.* 2004). Additionally, the presence of gene conversion makes the estimation of (reciprocal) recombination rates difficult (Martinsohn *et al.* 1999; Hosomichi *et al.* 2008). Anyway, current experimental results do not reveal a consistent picture as to whether there is a benefit, or trend, to suppress recombination in large multigene families.

Caveats and future direction

While our model has incorporated multiple genetic processes, it is likely still far away from the details of how multigene families evolve in real-life populations. One issue, not considered here, is gene conversion where an allele, or a fragment thereof, overwrites another one in a pairing chromosome. For example, gene conversion is known to play an important role in maintaining MHC diversity (Högstrand and Böhme 1999; Martinsohn *et al.* 1999; Wiehe *et al.* 2000; Bahr and Wilson 2012).

Also, our selection model assumes time-independent fitness and each allele provides the same selective benefit. This corresponds to an ideal situation where external factors are ubiquitous and stable. In practice, however, the selective benefits of certain alleles do change together with a changing environment. Evolving pathogens, for instance, leads to arbitrarily complex coevolution dynamics (Ejsmond and Radwan 2011; Tellier *et al.* 2014). Furthermore, population structure may interact with diversifying selection in a complex or even counter-intuitive way. In humans, it is known that different populations harbor different MHC alleles, likely driven by pathogen diversity (Manczinger *et al.* 2019). A hypothesis is that multiple subpopulations act as reservoirs of alleles and backups for each other, allowing for quick response against new pathogens (Lenz *et al.* 2009; Linnenbrink *et al.* 2018). Interaction between population structure and local adaptation needs to take into account subpopulation sizes and migration networks. For instance, it was shown that subpopulation sizes can affect local allelic diversity (Mason *et al.* 2011).

Finally, and perhaps most importantly, gene function decides on fitness. On population genetic time scales pseudogenization plays an important role for the evolution of multigene families (Hess 2000; Menashe *et al.* 2006). Although eventually removed by selection, pseudogenes can persist in real-life populations with high frequency. Conditions under which pseudogenes appear and persist can be identified in accordingly modified models. Structural and functional aspects being included together with gene conversion, temporally or locally varying selection strengths into theoretical models will help to address open questions, but remains to be considered in future work.

Data availability

Results from simulation experiments, as well as copy number counts in empirical data, are available at https://github.com/y-zheng/Recombination-gene-family.

Supplemental material is available at *GENETICS* online.

Acknowledgments

The authors would like to thank 2 anonymous reviewers for their detailed and constructive comments on an earlier draft of this manuscript.

Funding

This work has been funded by grants from the German Research Foundation (DFG SPP-1590 and DFG SFB-1211/B6) to TW.

Conflicts of interest

None declared.

Literature cited

Bahr A, Wilson AB. The evolution of MHC diversity: evidence of intralocus gene conversion and recombination in a single-locus system. Gene. 2012;497(1):52–57.

Beeson SK, Mickelson JR, McCue ME. Exploration of fine-scale recombination rate variation in the domestic horse. Genome Res. 2019; 29(10):1744–1752.

Brahmachary M, Guilmatre A, Quilez J, Hasson D, Borel C, Warburton P, Sharp AJ. Digital genotyping of macrosatellites and multicopy genes reveals novel biological functions associated with copy number variation of large tandem repeats. PLoS Genet. 2014;10(6):e1004418.

Chao L. Evolution of sex in RNA viruses. J Theor Biol. 1988;133(1): 99–112.

de Bakker PIW, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M, *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet. 2006;38(10): 1166–1172.

de Weyer ALV, Monteiro F, Furzer OJ, Nishimura MT, Cevik V, Witek K, Jones JD, Dangl JL, Weigel D, Bemm F. A species-wide inventory of NLR genes and alleles in *Arabidopsis thaliana*. Cell. 2019;178(5): 1260–1272.e14.

Demuth JP, Hahn MW. The life and death of gene families. Bioessays. 2009;31(1):29–39.

Dudek K, Gaczorek TS, Zieliński P, Babik W. Massive introgression of major histocompatibility complex (MHC) genes in newt hybrid zones. Mol Ecol. 2019;28(21):4798–4810.

Eichler EE. Copy number variation and human disease. Nat Educ. 2008;1:1.

Ejsmond MJ, Radwan J. MHC diversity in bottlenecked populations: a simulation model. Conserv Genet. 2011;12(1):129–137.

Ekblom R, Saether SA, Jacobsson P, Fiske P, Sahlman T, Grahn M, Kålås JA, Höglund J. Spatial pattern of MHC class II variation in the great snipe (*Gallinago media*). Mol Ecol. 2007;16(7):1439–1451.

Frankham R, Bradshaw CJ, Brook BW. Genetics in conservation management: revised recommendations for the 50/500 rules, red list criteria and population viability analyses. Biol Conserv. 2014;170: 56–63.

Fulton JE, McCarron AM, Lund AR, Pinegar KN, Wolc A, Chazara O, Bed'Hom B, Berres M, Miller MM. A high-density SNP panel reveals extensive diversity, frequent recombination and multiple recombination hotspots within the chicken major histocompatibility complex b region between BG2 and CD1a1. Genet Sel Evol. 2016;48(1):1–15.

Grimwood J, Gordon LA, Olsen A, Terry A, Schmutz J, Lamerdin J, Hellsten U, Goodstein D, Couronne O, Tran-Gyamfi M, *et al.* The DNA sequence and biology of human chromosome 19. Nature. 2004;428(6982):529–535.

Haigh J. The accumulation of deleterious genes in a population—muller's ratchet. Theor Popul Biol. 1978;14(2):251–267.

Haldane J. The effect of variation of fitness. Am Nat. 1937;71(735): 337–349.

Herdegen M, Babik W, Radwan J. Selective pressures on MHC class II genes in the guppy (*Poecilia reticulata*) as inferred by hierarchical analysis of population structure. J Evol Biol. 2014;27(11): 2347–2359.

Hess CM, Gasper J, Hoekstra HE, Hill CE, Edwards SV. MHC class II pseudogene and genomic signature of a 32-kb cosmid in the house finch (*Carpodacus mexicanus*). Genome Res. 2000;10(5): 613–623.

Högstrand K, Böhme J. Gene conversion can create new MHC allees. Immunol Rev. 1999;167:305–317.

Hosomichi K, Miller MM, Goto RM, Wang Y, Suzuki S, Kulski JK, Nishibori M, Inoko H, Hanzawa K, Shiina T. Contribution of mutation, recombination, and gene conversion to chicken MHC-B haplotype diversity. J Immunol. 2008;181(5):3393–3399.

Howe K, Schiffer PH, Zielinski J, Wiehe T, Laird GK, Marioni JC, Soylemez O, Kondrashov F, Leptin M. Structure and evolutionary history of a large family of NLR proteins in the zebrafish. Open Biol. 2016;6(4):160009.

Ingvarsson PK, Whitlock MC. Heterosis increases the effective migration rate. Proc Biol Sci. 2000;267(1450):1321–1326.

Innan H. Population genetic models of duplicated genes. Genetica. 2009;137(1):19–37.

Kondrashov AS. Selection against harmful mutations in large sexual and asexual populations. Genet Res. 1982;40(3):325–332.

Krüger J, Vogel F. Population genetics of unequal crossing over. J Mol Evol. 1975;4(3):201–247.

Lam TH, Shen M, Chia JM, Chan SH, Ren EC. Population-specific recombination sites within the human MHC region. Heredity (Edinb). 2013;111(2):131–138.

Lenz TL, Wells K, Pfeiffer M, Sommer S. Diverse MHC IIB allele repertoire increases parasite resistance and body condition in the long-tailed giant rat (*Leopoldamys sabanus*). BMC Evol Biol. 2009;9:269.

Linnenbrink M, Teschke M, Montero I, Vallier M, Tautz D. Meta-populational demes constitute a reservoir for large MHC allele diversity in wild house mice (*Mus musculus*). Front Zool. 2018;15:15.

Liu L, Yu L, Kalavacharla V, Liu Z. A Bayesian model for gene family evolution. BMC Bioinformatics. 2011;12:426.

Manczinger M, Boross G, Kemény L, Müller V, Lenz TL, Papp B, Pál C. Pathogen diversity drives the evolution of generalist MHC-II alleles in human populations. PLoS Biol. 2019;17(1):e3000131.

Martinsohn JT, Sousa AB, Guethlein LA, Howard JC. The gene conversion hypothesis of MHC evolution: a review. Immunogenetics. 1999;50(3–4):168–200.

Mason RAB, Browning TL, Eldridge MDB. Reduced MHC class II diversity in island compared to mainland populations of the black-footed rock-wallaby (*Petrogale lateralis lateralis*). Conserv Genet. 2011;12(1):91–103.

Menashe I, Aloni R, Lancet D. A probabilistic classifier for olfactory receptor pseudogenes. BMC Bioinformatics. 2006;7:393.

Milesi P, Weill M, Lenormand T, Labbé P. Heterogeneous gene duplications can be adaptive because they permanently associate overdominant alleles. Evol Lett. 2017;1(3):169–180.

Miller HC, Lambert DM. Genetic drift outweighs balancing selection in shaping post-bottleneck major histocompatibility complex variation in New Zealand robins (Petroicidae). Mol Ecol. 2004; 13(12):3709–3721.

Nadeau JH, Sankoff D. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics. 1997;147(3):1259–1266.

Ohta T. An extension of a model for the evolution of multigene families by unequal crossing over. Genetics. 1979;91(3):591–607.

Ohta T. Evolution of gene families. Gene. 2000;259(1–2):45–52.

Ohta T. Further simulation studies on evolution by gene duplication. Evolution. 1988;42(2):375–386.

Ohta T. Multigene Families and Their Implications for Evolutionary Theory. Berlin, Heidelberg: Springer; 1984. p. 133–139.

Ohta T. Simple model for treating evolution of multigene families. Nature. 1976;263(5572):74–76.

Ohta T. Simulating evolution by gene duplication. Genetics. 1987; 115(1):207–213.

Petit M, Astruc JM, Sarry J, Drouilhet L, Fabre S, Moreno CR, Servin B. Variation in recombination rate and its genetic determinism in sheep populations. Genetics. 2017;207(2):767–784.

Rafajlović M, Klassmann A, Eriksson A, Wiehe T, Mehlig B. Demography-adjusted tests of neutrality based on genome-wide SNP data. Theor Popul Biol. 2014;95:1–12.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, *et al.* Global variation in copy number in the human genome. Nature. 2006;444(7118):444–454.

Schaschl H, Wandeler P, Suchentrunk F, Obexer-Ruff G, Goodman SJ. Selection and recombination drive the evolution of MHC class II DRB diversity in ungulates. Heredity (Edinb). 2006;97(6):427–437.

Schierup MH, Vekemans X, Charlesworth D. The effect of subdivision on variation at multi-allelic loci under balancing selection. Genet Res. 2000;76(1):51–62.

Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014; 46(8):919–925.

Silver L. Evolution of gene families. In: S Brenner, JH Miller, editors. Encyclopedia of Genetics. New York (NY): Academic Press; 2001. p. 666–669.

Smith GP. Unequal crossover and the evolution of multigene families. Cold Spring Harb Symp Quant Biol. 1974;38:507–513.

Spence JP, Song YS. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. Sci Adv. 2019;5(10):eaaw9206.

Takahata N. A mathematical study on the distribution of the number of repeated genes per chromosome. Genet Res. 1981;38(1):97–102.

Tellier A, Moreno-Gámez S, Stephan W. Speed of adaptation and genomic footprints of host-parasite coevolution under arms race and trench warfare dynamics. Evolution. 2014;68: 2211–2224.

Traherne JA. Human MHC architecture and evolution: implications for disease association studies. Int J Immunogenet. 2008;35(3): 179–192.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, *et al.* Fine-scale structural variation of the human genome. Nat Genet. 2005; 37(7):727–732.

Vahdati AR, Wagner A. Parallel or convergent evolution in human population genomic data revealed by genotype networks. BMC Evol Biol. 2016;16(1):1–19.

Wiehe T, Mountain J, Parham P, Slatkin M. Distinguishing recombination and intragenic gene conversion by linkage disequilibrium patterns. Genet Res. 2000;75(1):61–73.

Communicating editor: A. Tellier

Appendix

Proof of (2). Using the closed-form formula of the geometric series and the fact that $x = y$, we can write the fitness function $\omega = (1+s_x)^{\left(\sum_{i=0}^{x-1}\beta_x^i\right)} \times (1-s_y)^{\left(\sum_{j=0}^{y-3}\beta_y^j\right)}$ as a function of y that equals

$$f(y) = (1+s_x)^{\frac{1-\beta_x^y}{1-\beta_x}} \times (1-s_y)^{\frac{1-\beta_y^{y-2}}{1-\beta_y}}.$$

Defining

$$a := (1+s_x)^{\frac{1}{1-\beta_x}}, \quad b := (1-s_y)^{\frac{1}{1-\beta_y}},$$

we find that

$$f'(y) = -\left(\ln(a)\ln(\beta_x)\cdot\beta_x^y + \ln(b)\ln(\beta_y)\cdot\beta_y^{y-2}\right)\cdot a^{1-\beta_x^y}\cdot b^{1-\beta_y^{y-2}}.$$

Setting $f'(y^*) = 0$ leads us to

$$\underbrace{-\ln(a)\ln(\beta_x)}_{:=p_1}\cdot\beta_x^{y^*} = \underbrace{\ln(b)\ln(\beta_y)\frac{1}{\beta_y^2}}_{:=p_2}\cdot\beta_y^{y^*}$$

$$\Rightarrow\ p_1\beta_x^{y^*} = p_2\beta_y^{y^*}$$

$$\Rightarrow\ y^* = \frac{\ln(p_1)-\ln(p_2)}{\ln(\beta_y)-\ln(\beta_x)},$$

and inserting the expressions for p_1, p_2, a, b gives the result. □

Proof of Proposition 1. We note that the parental status of the chromosomes does not matter in the following calculations. Therefore, we use the notation $y'_{(\cdot)}$ instead of $y^{m\prime}$ and $y^{p\prime}$. Since the T describes the distribution of the sum of 2 uniform random variables, we observe that the expected value is given by

$$\sum_{y'} y'\cdot T(y', y'_1, y'_2) = \mathbb{E}[B_1 + B_2 - 1] = \frac{y'_1+1}{2} + \frac{y'_2+1}{2} - 1 = \frac{y'_1+y'_2}{2},$$

and therefore conclude that

$$\begin{aligned}
&\textstyle\sum_{y'} y'\cdot p_{t+1}(y') \\
&= (1-r)\textstyle\sum_{y'} y'\cdot p_t(y') + r\sum_{y'} y'\cdot\sum_{y'_1,y'_2} p_t(y'_1)p_t(y'_2)T(y', y'_1, y'_2) \\
&= (1-r)\textstyle\sum_{y'} y'\cdot p_t(y') + r\sum_{y'_1,y'_2} p_t(y'_1)p_t(y'_2)\frac{y'_1+y'_2}{2} \\
&= (1-r)\textstyle\sum_{y'} y'\cdot p_t(y') + r\sum_{y'_1}\frac{y'_1}{2}p_t(y'_1)\underbrace{\sum_{y'_2} p_t(y'_2)}_{=1} + r\sum_{y'_2}\frac{y'_2}{2}p_t(y'_2)\underbrace{\sum_{y'_1} p_t(y'_1)}_{=1} \\
&= \textstyle\sum_{y'} y'\cdot p_t(y').
\end{aligned}$$

We define $a = 2/E_{Y'}$ and note that the stationary distribution is independent from the recombination rate $r > 0$, i.e.

$$\begin{aligned}
& p_{stat}(y') = (1-r)p_{stat}(y') + r\cdot\textstyle\sum_{y'_1,y'_2} p_{stat}(y'_1)p_{stat}(y'_2)T(y', y'_1, y'_2) \\
\iff\quad & p_{stat}(y') = \textstyle\sum_{y'_1,y'_2} p_{stat}(y'_1)p_{stat}(y'_2)T(y', y'_1, y'_2).
\end{aligned}$$

Therefore, we find that

$$\begin{aligned}
&\textstyle\sum_{y'_1,y'_2} p_{stat}(y'_1)p_{stat}(y'_2)T(y', y'_1, y'_2) \\
=\ & \left(\tfrac{1}{Z}\right)^2\cdot\Big(\sum_{y'_2=1}^{y'} y'_2\cdot e^{-ay'_2}\sum_{y'_1=y'+1}^{\infty} e^{-ay'_1} + \sum_{y'_2=y'+1}^{\infty} e^{-ay'_2}\sum_{y'_1=1}^{y'} y'_1\cdot e^{-ay'_1} \\
& + y'\cdot\sum_{y'_2=y'+1}^{\infty} e^{-ay'_2}\sum_{y'_1=y'+1}^{\infty} e^{-ay'_1} + \sum_{y'_2=1}^{y'} e^{-ay'_2}\sum_{y'_1=y'-y'_2}^{y'_2}(y'_1+y'_2-y')e^{-ay'_1}\Big) \\
=\ & \dots(*)\dots \\
=\ & y'\cdot e^{-ay'}\cdot\frac{(e^a-1)^2}{e^a} \\
=\ & p_{stat}(y'),
\end{aligned}$$

where the detailed calculations of $(*)$ are shown below. □

Proof of (*). Using the substitution $k = (y'_1 + y'_2 - y')$ we find that

$$\begin{aligned}
&\textstyle\sum_{y'_1,y'_2} p_{stat}(y'_1)p_{stat}(y'_2)T(y', y'_1, y'_2) \\
=\ & \frac{(e^a-1)^4}{e^{2a}}\cdot\Big[\sum_{y'_2=1}^{y'} y'_2\cdot e^{-ay'_2}\sum_{y'_1=y'+1}^{\infty} e^{-ay'_1} + \sum_{y'_2=y'+1}^{\infty} e^{-ay'_2}\sum_{y'_1=1}^{y'} y'_1\cdot e^{-ay'_1} \\
& y'\cdot\sum_{y'_2=y'+1}^{\infty} e^{-ay'_2}\sum_{y'_1=y'+1}^{\infty} e^{-ay'_1} + \sum_{y'_2=1}^{y'} e^{-ay'_2}\sum_{y'_1=y'-y'_2}^{y'_2}(y'_1+y'_2-y')e^{-ay'_1}\Big] \\
=\ & \frac{(e^a-1)^4}{e^{2a}}\cdot\Big[2\frac{e^{-a(y'+1)}}{1-e^{-a}}\cdot\left(-\frac{\partial}{\partial a}\right)\left(\frac{1-e^{-a(y'+1)}}{1-e^{-a}}-1\right) + y'\cdot\left(\frac{e^{-a(y'+1)}}{1-e^{-a}}\right)^2 \\
& + \sum_{y'_2=1}^{y'} e^{-ay'_2}\sum_{k=0}^{y'_2} k\cdot e^{-a(k+y'-y'_2)}\Big] \\
=\ & \frac{(e^a-1)^4}{e^{2a}}\cdot\Big[2\frac{e^{-a(y'+1)}}{1-e^{-a}}\cdot\frac{e^{-ay'}\left(e^{a(y'+1)}-((y'+1)e^a+y')\right)}{(e^a-1)^2} + y'\cdot\left(\frac{e^{-a(y'+1)}}{1-e^{-a}}\right)^2 \\
& + e^{-ay'}\sum_{y'_2=1}^{y'}\left(-\frac{\partial}{\partial a}\right)\left(\frac{1-e^{-a(y'_2+1)}}{1-e^{-a}}-1\right)\Big] \\
=\ & \frac{(e^a-1)^4}{e^{2a}}\cdot\Big[\frac{e^{-2ay'}\left(2e^{a(y'+1)}-(y'+2)e^a+y'\right)}{(e^a-1)^3} \\
& + e^{-ay'}\cdot\left(-\frac{\partial}{\partial a}\right)\left(\frac{y'-e^{-a}\left(\frac{1-e^{-a(y'+1)}}{1-e^{-a}}-1\right)}{1-e^{-a}}\right)\Big] \\
=\ & \frac{(e^a-1)^4}{e^{2a}}\cdot\Big[\frac{e^{-2ay'}\left(2e^{a(y'+1)}-(y'+2)e^a+y'\right)}{(e^a-1)^3} \\
& + \frac{e^{-2ay'}\left(y'e^{a(y'+2)}+(y'+2)e^a-(y'+2)e^{a(y'+1)}-y'\right)}{(e^a-1)^3}\Big] \\
=\ & \frac{(e^a-1)\cdot e^{-2ay'}\cdot\left(y'e^{a(y'+2)}-y'e^{a(y'+1)}\right)}{e^{2a}} \\
=\ & y'\cdot e^{-ay'}\cdot\frac{(e^a-1)^2}{e^a}.
\end{aligned}$$

□

Table A1. Parameters used in simulations of the compound model.

Scenario (a) Single population of constant size N_e		
	N_e	500, 1,000, 2,000, 4,000
	μ	0.0005
(a1)	(s_x, s_y) :	$(0.01, 0.0025)$, $(0.02, 0.005)$, $(0.04, 0.01)$[a]
	r :	0.01
(a2)	(s_x, s_y) :	$(0.02, 0.005)$
	r :	$0.002, 0.005, 0.01, 0.02, 0.05$
	Replicates	500 per parameter combination
	Recording	every 100-th for 20,000 generations
Scenario (b) Instantaneous bottleneck		
N_0^b		1,000, 2,000
N_b^c :		20
duration :		5, 10, 20 generations
μ		0.0005
(s_x, s_y)		$(0.01, 0.0025)$, $(0.02, 0.005)$, $(0.04, 0.01)$
r		0.01
Replicates		200 per parameter combination
Recording		Every 10-th for 5,000 generations after bottleneck
Scenario (c) Two populations of constant size N_e with 2-way migration[d]		
N_e		500, 1,000
$N_e m$		0.1, 1, 10
μ		0.0005
(s_x, s_y)		$(0.01, 0.0025)$, $(0.02, 0.005)$, $(0.04, 0.01)$
r		0.01
Replicates		100 pairs per parameter combination
Recording		Every 10-th for 2,000 generations
Scenario (d) Single population of constant size N_e with recomb. rate modifier ρ		
N_e		1,000, 2,000
μ		0.0005
(s_x, s_y)		$(0.01, 0.0025)$, $(0.02, 0.005)$, $(0.04, 0.01)$
Base rate r_o		0.01
Initial ρ_0		1 for all chromosomes
Modification[e] of $r = r_o \cdot \rho$ according to		$\begin{cases} \rho_{t+1} = \rho_t \ (p = 0.996) \\ \rho_{t+1} = \rho_t \cdot e^{0.05} \ (p = 0.002) \\ \rho_{t+1} = \rho_t \cdot e^{-0.05} \ (p = 0.002) \end{cases}$
Replicates		200 per parameter combination
Recording		Every 100-th for 50,000 generations

[a] The 3 levels of selection strengths are referred to as "weak," "intermediate," and "strong" in the text.
[b] Population size before and after bottleneck.
[c] Population size during bottleneck.
[d] At rate m per individual per generation per direction.
[e] ρ Changes from ρ_t to ρ_{t+1} per generation per chromosome with probability P.