Distinguishing the roles of adaptation and demography in gene copy number changes in human populations

Moritz Otto[1◑], Yichen Zheng[1◑], Paul Grablowitz[2], Thomas Wiehe[1*]

1 Institute for Genetics, University of Cologne, Germany

2 Department of Computer Science, University of Tübingen, Germany

◑These authors contributed equally to this work.

* twiehe@uni-koeln.de

Abstract

Genes with multiple copies are likely to be maintained by stabilizing selection that puts a bound to unlimited expansion of copy number. We designed a model where copy number variation is generated by unequal recombination, which fits well to a number of genes, surveyed in three human populations. Based on this theoretical model and on computer simulations, we were interested in answering the question whether gene copy number distribution in the derived European and Asian populations can be explained by a purely demographic scenario or whether shifts in the distribution are signatures of adaptation. Although copy number distribution in most of the analyzed gene clusters can be explained by a bottleneck as in the out of Africa expansion of homo sapiens 60-10kyrs ago, we identified several candidate genes, for instance `AMY1A` and `PGA3`, whose copy numbers are likely to be selected differently among African, Asian and European populations.

Introduction

Gene copy number variation (CNV) refers to the presence of multiple copies of a gene family within a genome, resulting from duplications, deletions, or rearrangements.

Combined with their high mutation rate CNVs constitute a significant driver of genomic variability that allows for rapid adaptive evolution in response to environmental changes [1–5].

A well studied example of CNV within human population is provided by the salivary amylase gene, whose variations in the number of copies are hypothesized to correlate with the extent of dietary starch consumption not only in human but also in other species [6–11].

In general, copy number variation may result from different evolutionary forces acting upon them. Demographic events, such as population migrations and expansions, can lead to changes in gene frequencies and distributions over time. Simultaneously, natural selection acts on genetic variations, favoring advantageous alleles and promoting their proliferation within populations.

It is known that both demographic effects and selection may produce similar patterns in single nucleotide as well as in structural variants, making it difficult to disentangle these forces [12, 13]. For SNP or allele frequency data, there have been well-developed statistics (e.g., [14, 15]) that are "standardized " so that a genomic baseline can be established, from which loci under selection may be detected. However, such a genomic baseline is not available for gene copy number variation data. Therefore, we resorted to a more basic approach involving modelling and computer simulations.

We have recently examined the evolutionary dynamics of multi-copy gene families with respect to selective pressure and unequal recombination [16]. This study focused on analyzing the impact of stabilizing selection on gene copy numbers, while considering the role of recombination as a randomizing mechanism that introduces variability within the population.

By expanding this model, we aimed to assess whether gene copy number alterations observed within human populations could be solely attributed to demographic events or whether selective pressures have played a role in shaping these variations.

In this study, we conducted extensive simulations under various scenarios of human demography and selective changes. By disentangling the effects of these two forces, we sought to gain a deeper understanding of the evolutionary processes driving gene copy number variation in human populations. Based on empirical data of human gene copy numbers we identified several candidate genes, whose copy numbers are likely to be

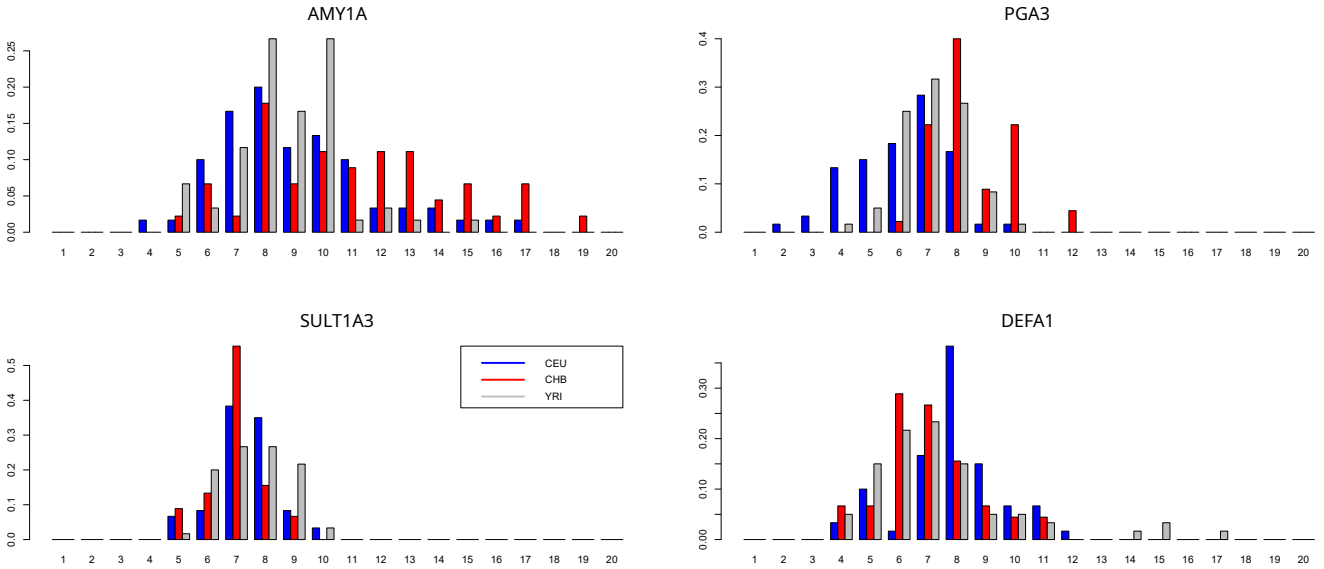

Fig 1. Gene copy number distribution in four exemplary gene famiies in three human populations, CEU, CHB, YRI. Data adapted from [2].

selected differently among African, Asian and European populations.

Materials and Methods

Gene copy number variation in human

We started with the dataset provided by [2]. Using Nanostring technology they estimated gene copy numbers of 180 gene-families in 165 individuals of three populations (60 African Yuroba - YRI, 60 Central Europe - CEU and 45 Asia - CHB) based on data collected in the framework of the 1,000 Genomes Project [17]. While some of these loci showed copy numbers of > 100 copies (`DUX4` even up to 600), we focused on intermediate copy numbers and removed all satellite loci, genes on sex chromosomes, genes with minimum copy number below 2, and genes with mean copy number (in YRI) below 5 or above 60. For genes that have two primer sets, only one is taken. We used t-test and f-test statistics to select gene families with significant differences in mean and standard deviation between either YRI-CHB or YRI-CEU comparisons and removed those that showed no statistical evidence in any of these. This resulted in 42 gene families, see Table 1. An example of the copy number distribution of four gene families is shown in Fig 1.

Table 1. Differences in mean and variance between populations. 0 indicates no significant change, + a significant increase and − a significant decrease (t- and f-test, $\alpha = 0.05$).

Gene	YRI mean	CEU mean		CHB mean		YRI sd	CEU sd		CHB sd	
AMY1A	8.628	0	9.151	+	11.128	1.907	+	2.713	+	3.35
ANKRD20A3	26.414	+	29.148	+	29.904	1.804	0	1.718	0	1.698
BOLA2B	7.235	-	6.679	-	6.49	1.012	0	0.914	0	0.839
CBWD3	12.146	0	12.374	+	13.068	0.995	0	0.949	+	1.81
CDC37P1	14.941	+	19.977	0	16.582	4.11	+	5.619	+	5.579
CLEC18A	7.799	+	8.362	0	7.932	1.331	0	1.216	0	1.392
CSH	6.738	+	7.182	+	7.474	0.497	0	0.555	0	0.575
DEFA1	7.442	0	7.891	0	7.056	2.643	-	1.671	-	1.604
DEFB130	5.081	+	5.315	0	5.243	0.562	0	0.532	0	0.462
FAM72A	6.914	+	7.573	+	7.561	0.617	+	0.86	0	0.651
FAM75A1	11.859	0	11.972	+	13.362	1.473	0	1.391	+	2.019
FAM75A5	11.693	0	11.522	+	12.533	1.115	0	1.197	+	1.751
FCGBP	5.282	+	5.693	+	5.79	1.291	-	0.678	0	1.046
FOXD4L2	13.013	+	13.694	+	14.55	1.015	0	0.994	+	1.877
GOLGA6L9	27.683	0	28.586	+	29.181	2.615	0	2.532	0	2.59
GOLGA8G	29.209	+	31.641	+	30.37	3.065	0	2.783	0	2.35
GUSBP1	12.95	+	15.886	+	13.987	2.249	0	2.585	0	2.213
HIST2	8.436	+	8.709	+	8.894	0.528	0	0.673	0	0.644
LIMS3	5.829	-	5.408	-	5.661	0.346	0	0.354	0	0.39
LOC23117	50.194	0	50.304	-	48.639	3.685	0	2.963	0	2.789
LOC653606	6.56	0	6.403	-	5.999	0.486	0	0.621	+	0.917
MUC12	11.845	+	14.098	0	12.123	2.586	0	2.01	-	1.803
NBPF11	49.963	-	48.002	0	48.68	4.203	-	3.114	0	3.311
NBPF16	45.25	0	46.436	+	47.006	4.706	0	5.023	0	3.988
NPIP	51.171	-	49.488	-	48.938	2.16	0	2.327	0	2.224
PGA3	7.044	-	6.181	+	8.473	1.205	+	1.565	0	1.353
PPIAP21	43.141	+	48.632	+	49.493	3.765	0	4.315	0	3.881
PRAMEF14	10.516	+	11.835	+	11.888	1.295	+	2.246	+	1.937
PRAMEF20	7.253	0	7.415	+	7.576	0.566	0	0.723	+	0.924
PRAMEF5	17.844	-	16.475	-	15.804	1.721	+	2.386	+	2.578
PRAMEF8	5.919	0	5.787	0	5.842	0.652	+	1.281	0	0.819
PRR11	6.868	+	8.298	+	8.305	0.923	0	0.965	0	0.708
PRR20A	20.639	-	17.284	-	14.85	6.903	-	5.288	0	5.584
PSG3	14.943	+	15.624	0	15.087	1.314	0	1.238	+	1.843
RGPD1	13.959	0	14.037	0	14.151	0.791	+	1.309	+	1.266
SPDYE3	34.611	-	31.656	-	32.828	2.836	-	2.105	0	2.617
SULT1A3	7.627	0	7.406	-	7.017	1.197	0	1.087	0	0.904
TBC1D3	45.515	-	33.191	-	39.306	6.337	0	6.888	+	8.381
TCEB3C	33.02	-	28.574	-	25.895	7	0	7.383	0	6.299
TP53TG3	9.172	0	8.904	-	6.735	1.825	0	2.08	0	1.666
TRIM49L1	12.353	+	14.078	+	14.112	1.664	0	2.06	0	1.874
ZNF658B	5.544	+	6.273	+	6.647	0.727	0	0.827	+	1.01

Unequal recombination model

In a recently developed model we considered unequal recombination and selection to describe the evolution of tandem gene arrays [16]. We shortly summarize the main findings. Consider two chromosomes with gene arrays of size y_1 and y_2. A recombination event happens at rate r and may produce a gamete of gene array size according to the trapezodial distribution, such that

$$Prob[y|y_1, y_2] = \frac{1}{y_1 y_2} \begin{cases} 0, & y < 1 \\ y, & 1 \leq y < \min(y_1, y_2) \\ \min(y_1, y_2), & \min(y_1, y_2) \leq y < \max(y_1, y_2) \\ y_1 + y_2 - y, & \max(y_1, y_2) \leq y < y_1 + y_2 - 1 \\ 0, & y \geq y_1 + y_2 \end{cases} .$$

See Fig 2A for an illustration. We apply a fitness function, where each newly arising copy has a positive, yet decreasing benefit s_x. This is motivated by assuming a beneficial effect, yet with diminishing returns, either of increased gene dosage or of increased allelic diversity within an individual [16]. At the same time, we assume additional copies to be selected against with an increasing selective disadvantage s_y. This is motivated by assuming an increasing cost of replication, of gene processing and of maintaining genome integrity. Both effects are cast in a double-epistatic fitness function with two selection coefficients (s_x, s_y), governed by a single epistasis parameter (ε). To avoid the trivial long-term evolution equilibrium of one copy, we assume $s_x > s_y$. Furthermore, we assume $\varepsilon = 0.05$ to be constant in the following. Summarizing, fitness of a diploid individual with total gene copy number y is given by

$$\omega(y) = \exp\left\{\frac{1}{\varepsilon}\left((s_x + s_y) - s_x \cdot e^{-\varepsilon y} - s_y \cdot e^{\varepsilon(y-2)}\right)\right\} \tag{1}$$

This leads to an optimal copy number y_{opt} of

$$y_{\text{opt}} = 1 + \frac{\log(s_x/s_y)}{2\varepsilon}, \tag{2}$$

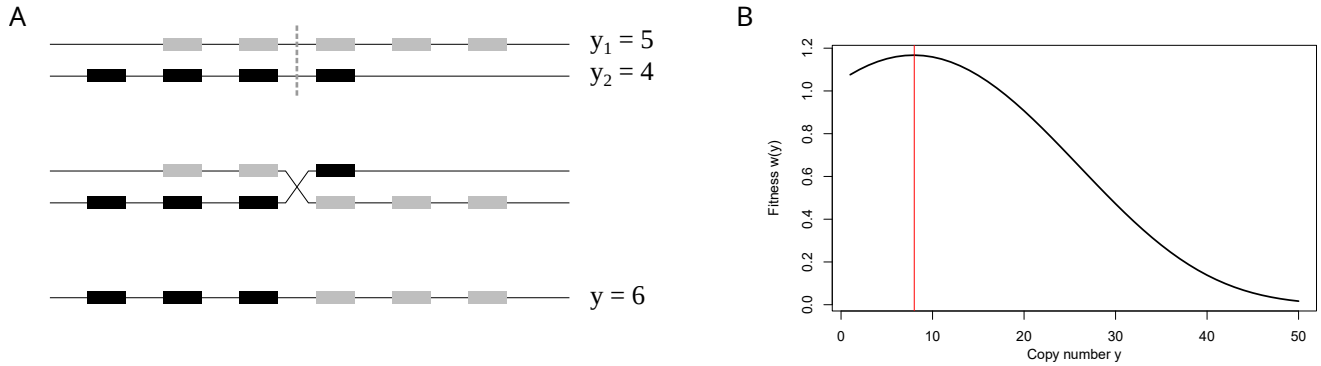

Fig 2. A. Sketch of the unequal recombination process. Starting with two chromosomes with $y_1 = 5$ and $y_2 = 4$ gene copies, two breaking points are chosen. One of the recombinants is then propagated. Its copy number $y = 6$ is Trapezodial as shown in [16]. B. Example of the fitness function $\omega(y)$ eq (1) with $\varepsilon = 0.05, s_x = 0.05, s_y = 0.0025$, which leads to an optimal copy number $y_{\mathbf{opt}} \approx 8$ copies.

which is determined by the ratio s_x/s_y when ε is kept fixed. See Fig 2B for an example. The population is then simulated according to a Wright-Fisher model with non-overlapping generations and with selection and recombination described as above. It was shown, that in the deterministic model the equilibrium copy number distribution is centered around y_{opt} and is well approximated by a Gamma distribution [16]. Furthermore, it holds that the coefficient of variation $\mathcal{C}_V = \sigma/\bar{y}$ is correlated to the logarithm of the recombination - selection ratio $\log(r/s_x)$. With strong selection and low recombination the distribution is tightly distributed around the optimal value, whereas higher r and lower s_x lead to a widespread distribution. Therefore, we introduce two new parameters:

- $q_S = s_x/s_y$, the *'selection ratio'*, which determines the optimal copy number, such that for $\varepsilon = 0.05$ we find

$$y_{\text{opt}} = y_{\text{opt}}(q_S) = 1 + 10 \cdot \log(q_S)$$

- $q_R = r/s_x$, the *'recombination/selection ratio'*, which measures the randomness produced by the unequal recombination process versus the selective pressure of the fitness function and therefore determines the coefficient of variation $\mathcal{C}_V = \sigma/\bar{y}$ of the equilibrium distribution.

Table 2. Parameters for regression simulations.

4 recombination rates r	0.1%, 0.2%, 0.5% and 1%
9 recombination ratios $q_R = r/s_x$	0.01, 0.02, 0.05, 0.1, 0.5, 1.0, 2.0, 5.0
9 optimal copy number values y_{opt}	10, 15, 20, 25, 30, 35, 40, 45, 50

Regression

To analyze the equilibrium distribution of the unequal recombination process under drift, we simulated the population evolution under different parameter settings. Population size is kept at $N = 5,000$ and assumed to be at an initial state of 5 copies on each chromosome. The different input parameters are given in Table 2. Together, they define 324 triples r, s_x, s_y. Additionally, we generated 160 random pairs such that q_R is between 0.01 and 5 and y_{opt} is between 4 and 60 and combined them with the four recombination rates, leading to a total parameter set of 964 combinations, where we disregarded those triples with selective strengths $s_x > 0.1$ to keep a realistic parameter range.

For each of this parameter combinations, we evolve the population under the given selection scheme for 5 million generations. The first 200,000 generations were discarded as burn-in and the population statistics (mean copy number $\bar{y}$ and standard deviation σ) are recorded every 20,000 generations.

In total, this results in $\approx 185,000$ data points, which we used to determine the relationship of input parameters (r, s_x, s_y) and output population statistics $(\bar{y}, \sigma)$.

As indicated in [16] we suggest a mean copy number $\bar{y}$ close to its optimal value y_{opt} and a correlation of $\mathcal{C}_V$ to $\log(q_R)$. Indeed, with r^2-values of 0.9842 and 0.9088 we find

$$\begin{aligned} \bar{y} &= 0.0379 + 0.983 \cdot y_{\text{opt}} \\ \mathcal{C}_V = \frac{\sigma}{\bar{y}} &= 0.323 + 0.0566 \cdot \log(q_R) - 0.00152 \cdot y_{\text{opt}} - 0.000036 \cdot \log(q_R) \cdot y_{\text{opt}} \end{aligned} \tag{3}$$

We calculated the q_S and q_R ratios based on $\bar{y}$ and $\mathcal{C}_V$ from gene copy numbers (see Tab 1) using the regression formula (3) with four recombination rates $r = 0.001$, 0.002, 0.005 and 0.01. Results for the four candidate genes shown in Fig 1 are given in Table 3.

Table 3. Estimates of selection coefficients s_x, s_y under four recombination rates $r = 0.001, ..., 0.01$ based on regression eq (3). The displayed gene families are the ones of Fig 1 for all three populations. Values in brackets are out of the range $0.001 < s_x < 0.1$ in YRI and hence not used in simulations.

					$r = 0.001$		$r = 0.002$		$r = 0.005$		$r = 0.01$	
Gene	Pop	Mean	SD	y_{opt}	s_x	s_y	s_x	s_y	s_x	s_y	s_x	s_y
AMY1A	CEU	9.1511	2.7133	9.2708	0.0012	0.0005	0.0025	0.0011	0.0062	0.0027	0.0125	0.0055
	CHB	11.128	3.3503	11.282	0.0011	0.0004	0.0022	0.0008	0.0054	0.0019	0.0109	0.0039
	YRI	8.6279	1.9074	8.7386	0.0048	0.0022	0.0097	0.0045	0.0242	0.0111	0.0483	0.0223
PGA3	CEU	6.1808	1.5646	6.2491	0.0029	0.0017	0.0058	0.0035	0.0146	0.0086	(0.0292)	(0.0173)
	CHB	8.4731	1.3526	8.5811	0.0144	0.0068	0.0289	0.0135	0.0722	0.0338	(0.1445)	(0.0677)
	YRI	7.0444	1.2053	7.1277	0.0122	0.0066	0.0245	0.0133	0.0611	0.0331	(0.1223)	(0.0663)
SULT1A3	CEU	7.4058	1.0872	7.4953	0.0186	0.0097	0.0373	0.0195	0.0932	0.0487	(0.1865)	(0.0974)
	CHB	7.0165	0.9041	7.0993	0.0259	0.0141	0.0518	0.0282	0.1295	0.0704	(0.2591)	(0.1408)
	YRI	7.6269	1.1971	7.7202	0.0155	0.0079	0.0311	0.0158	0.0774	0.0395	(0.1548)	(0.0791)
DEFA1	CEU	7.8911	1.6708	7.9889	(0.0058)	(0.0029)	(0.0116)	(0.0058)	0.0291	0.0145	0.0581	0.0289
	CHB	7.0561	1.6041	7.1396	(0.0045)	(0.0024)	(0.0091)	(0.0049)	0.0225	0.0122	0.0451	0.0244
	YRI	7.4421	2.6428	7.5321	(0.0005)	(0.0002)	(0.0009)	(0.0005)	0.0023	0.0012	0.0046	0.0024

Demography simulations

To determine whether significant changes of mean and variance of the copy number distribution (Table 1) can be explained by demographic history of human populations, we examined in total 6 different scenarios (enumerated as I - VI), see Fig 3.

Simulation of the bottleneck model First, we ran a simple bottleneck model of three different population reductions. Each is divided into three phases: (1) Burn-in phase. For each gene we used the estimated (r, s_x, s_y)-triple based on the dataset from YRI. These parameters were chosen as input to produce an equilibrium population of $N = 10,000$ by a burn-in process of 200,000 generations. 'Independent' equilibrium populations are produced by recording the population state every 20,000 generations. (2) Bottleneck. From equilibrium we reduced the population size to $N = 100,500$ or $1,000$, denoted scenario I, II and III, and kept it such for 5,000 generations. (3) Recovery phase. At the end of the bottleneck, the population is reset to $N = 10,000$ and the copy number distribution is recorded every 50 generations until generation 1000 after the bottleneck. We ran the bottleneck simulations I - III on all gene families given in Tab 1, with recombination rates $r = 0.001$, 0.002, 0.005 and 0.01, and discarded parameter combinations with s_x outside the interval [0.001, 0.1] in YRI. This gives a final total of 42 gene families and 95 gene-r combinations. For each gene, recombination rate and bottleneck population size combinations, 10,000 replicates are produced (from 100 'independent' starting equilibria).

We then traced mean and $\mathcal{C}_V$ along the recovery phase and compared these with the

empirical data from CHB and CEU populations.

Simulation of the human population history A more realistic population history of human is given by the *Genetic Algorithm for Demographic Model Analysis* (GADMA) [18], which also includes migration between subpopulations. We ran simulations on four candidate genes (`AMY1A, PGA3, SULT1A3, DEFA1`) with the following modification of the GADMA-demography: As ancestral population ($N = 9,900$ in GADMA), we used the equilibrium populations ($N = 10,000$) from the previous section. Therefore, we started the simulation 5992 generations before present, roughly corresponding to the onset of the 'out-of-Africa' expansion, when the Eurasian population split from the ancestral population and experienced a sharp bottleneck. To reduce computation time, we did not simulate the continued evolution of the African (YRI) population, since we assumed it to be in equilibrium; for migration from YRI to Eurasian populations, we drew samples from the ancestral population. At 896 generations before present, CEU and CHB split from each other and started to evolve including reciprocal migration and exponentially increasing population size. In the following, we refer to this simulation as scenario IV. At 'present', copy number distributions (mean and variance) were recorded. For each gene and recombination rate combination, 10,000 replicates were produced.

We also ran the same population model with a change of the selection parameter either at 500 generations or 896 generations before present (the latter being the CEU/CHB split time). The new selection parameters (s_x and s_y) are different for CEU and CHB populations, and are estimated from present CEU/CHB distributions (see Table 3). These simulations are hereafter called scenario V (selection change 500 generations before present) and VI (896 gpb).

Results and Discussion

In this study, we conducted an analysis of multicopy gene family evolution using a model that incorporates unequal recombination and selection. Our investigation aimed to examine the copy number changes observed in subpopulations of Europe, Asia, and Africa and to determine whether these changes could be attributed solely to

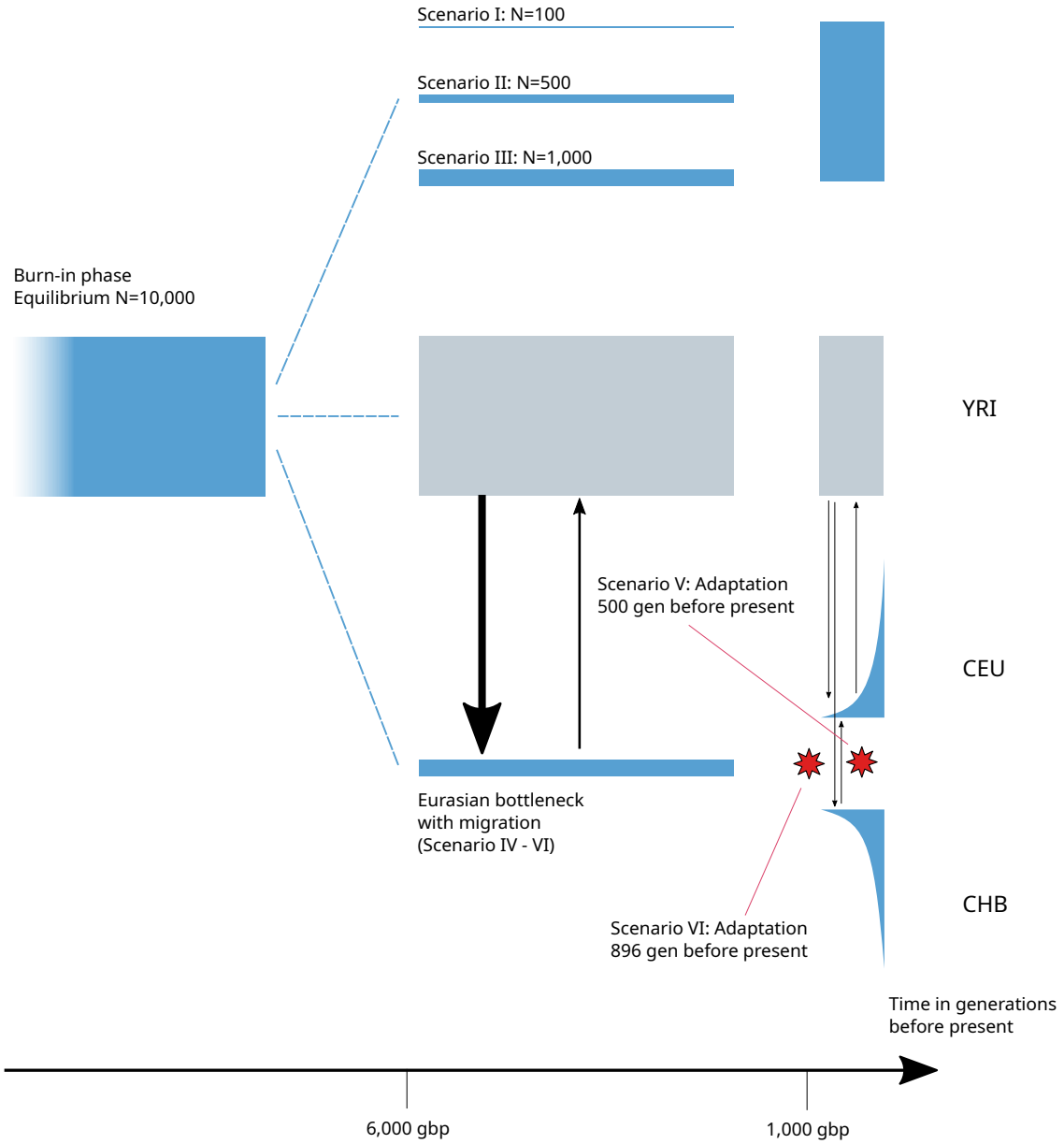

Fig 3. Illustration of simulated demography scenarios. The height of the blue boxes denote the population size. Scenario I-III cover a simple bottleneck lasting 5,000 generations, whereas scenario IV-VI are modifications of the GADMA-model with migration between subpopulations and change of selection parameters (indicated by red stars) in scenario V and VI. The gray color in the YRI population indicates, that it is at equilibrium.

demographic factors or if additionally selective pressures played a significant role. Our findings reveal that the observed copy number variations in several genes cannot adequately be explained by demographic processes alone, suggesting the involvement of selection.

Based on the data of [2], we chose 42 gene families of intermediate copy number that show significant differences in their distribution among different populations (Table 1). This dataset relies on phase I of the 1000 Genomes project and – despite of been published almost a decade ago – turned out to be most suitable for our analysis. More recent data from the human pangenome project still lack phasing and a sufficiently large sample size across different populations [19].

Using eq (3), we ran the simple bottleneck scenarios I-III of in total 95 parameter combinations (r, s_x, s_y), to see whether population size changes can explain the differences in copy number distribution. Fig 4 shows mean gene copy number of 10,000 simulated bottleneck populations over time for one chosen gene (here: `PGA3`). Gray boxes indicate the centerd 50% quantile, white the 95% and the whiskers 99% quantiles. With strong bottleneck (reduction to $N = 100$ for 5,000 generations) and under low recombination and hence weak selection ($r = 0.001$, and $q_R = r/s_x$, $q_S = s_x/s_y$ constant) we find the widest variation among the 10,000 replicates. Higher r and stronger selection shift the mean value back to the one of YRI population, which was the basis of the parameter estimates. In this particular example, we find compared to YRI (black horizontal line) a higher mean copy number of `PGA3` in CHB (red line) and a lower mean value in CEU (blue line). It is the only gene in our chosen set, that shows opposite direction of significant mean copy number change. Only under strong bottleneck and low recombination these changes can be explained with same selection parameters as in the ancestral population and lay within the 99% quantile.

An overview of all 95 bottleneck results is given in Table 4. We consider scenario I – the strongest bottleneck – and the time point after 1,000 generations of recovery (i.e. first row of Fig 4, last boxplot of each column). If the mean or resp. $\mathcal{C}_V$ lays within the 95% quantile, we indicate non-significant differences with a 0. Significant changes are marked with a single $*$ ($\alpha = 5\%$) or double asterisk $**$ ($\alpha = 1\%$). As an example, we find for `PGA3` a mean value which is significantly smaller in CEU than in YRI (marked with –). With $r = 0.001$, this might be explained by a bottleneck (denoted by 0),

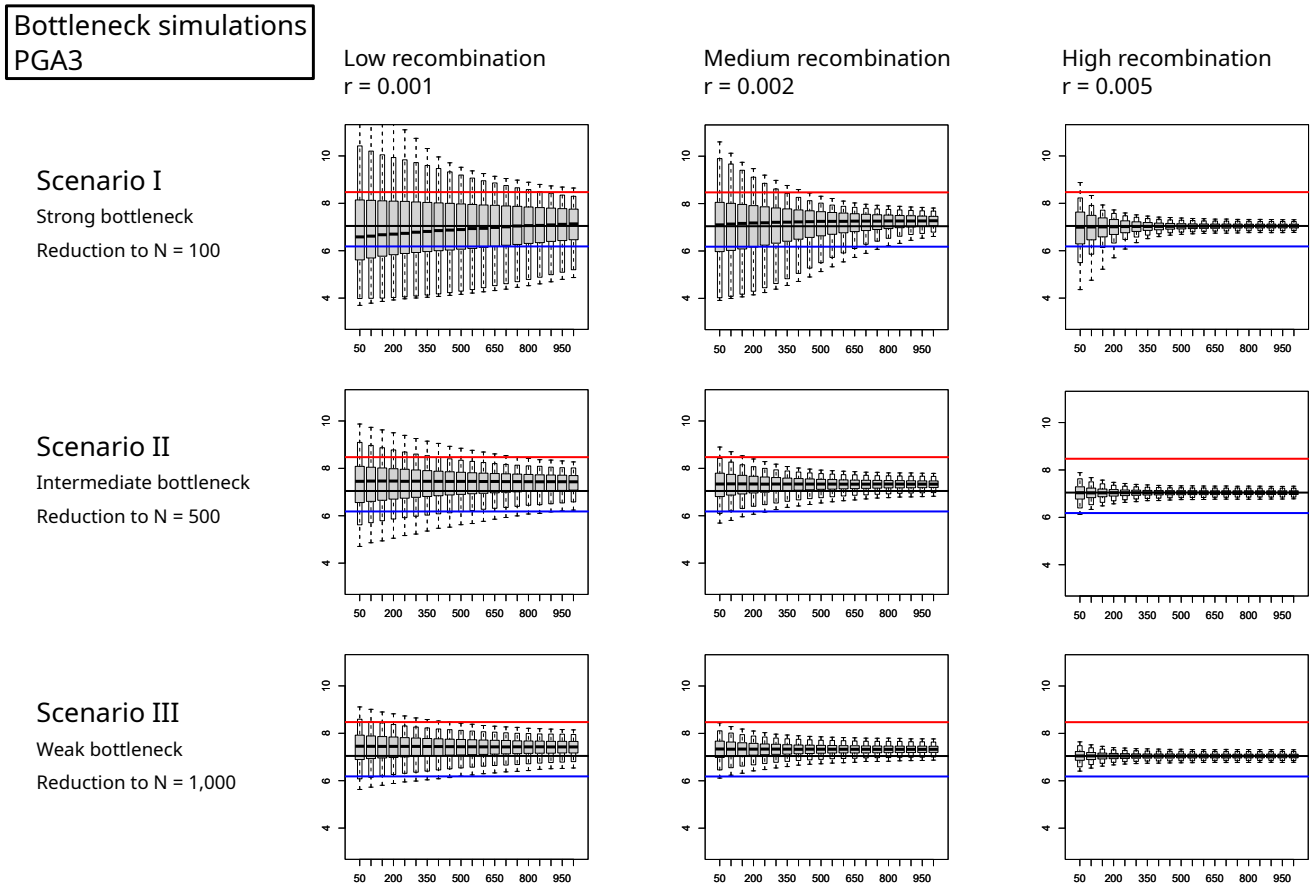

Fig 4. Mean copy number over time. After population reduction to $N = 100, 500, 1000$ (top to bottom) we traced the mean-value of 10,000 replicates over time (x-axis in generations). The input parameters s_x, s_y were estimated for $r = 0.001, 0.002, 0.005$ (left to right) from the YRI-data set for the candidate gene `PGA3` (see Tab 3) and kept constant over time, to see the effect of the bottleneck and recovery. Whiskers mark the 99% quantile, the white box the 95% quantile. Horizonal lines mark the values from the original data set of [2] (black: YRI, red: CHB, blue: CEU).

whereas for $r = 0.002$ and $r = 0.005$ we find a significant difference ($**$) and the bottleneck explanation to be highly unlikely. Higher recombination $r = 0.01$ led to s_x values greater than 0.1 in CHB and YRI (see Tab 3) and hence was omitted.

From the candidates with a significant difference in mean or variance we selected the three genes coding for digestive enzymes, `AMY1A`, `SULT1A3`, `PGA3`, and the defense gene `DEFA1` for a more detailed analysis and tested the GADMA demography without and with selection change according to the estimates from regression (scenario IV-VI).

Fig 5 shows mean copy number and coefficient of variation $\mathcal{C}_V$ at present, simulated according to scenarios I,IV,V and VI for 10,000 replicates each. As in the simplified bottleneck scenario, scenario IV with subpopulation migration returns to values of basis

Table 4. Results of bottleneck simulations. We ran simulations of scenario I (the strongest bottleneck with a reduction to $N = 100$) with parameters (r, s_x, s_y) estimated from the YRI-data and tested, whether after 1000 generations of recovery the mean and standard deviation σ of the CEU and CHB-data can be explained by a bottleneck. Blank space indicates that this parameter combination led to an s_x value out of the range of 0.001 and 0.1, and hence no simulation was run. The columns with 0,+ and − indicate whether there is a significant difference to the empirical data set (see Tab 1). In the r1 - r10 columns, a 0 indicates that the data can be explained by a bottleneck. * and ** show significant differences (5% and 1%) of the simulated and empirical data.

	CEU mean					CEU sd					CHB mean					CHB sd				
Gene		r1	r2	r5	r10		r1	r2	r5	r10		r1	r2	r5	r10		r1	r2	r5	r10
AMY1A	0	0	0	0	*	+	0	*	*	**	+	*	**	**	**	+	*	**	**	**
ANKRD20A3	+	**	**	.	.	0	0	0	.	.	+	**	**	.	.	0	0	0	.	.
BOLA2B	-	0	**	.	.	0	*	**	.	.	-	0	**	.	.	0	*	**	.	.
CBWD3	0	0	.	.	.	0	0	.	.	.	+	0	.	.	.	+	*	.	.	.
CDC37P1	+	**	**	**	**	+	*	**	**	**	0	0	0	*	**	+	*	**	**	**
CLEC18A	+	0	0	*	.	0	0	*	*	.	0	0	0	0	.	0	0	0	0	.
CSH	+	0	.	.	.	0	**	.	.	.	+	**	.	.	.	0	**	0	0	0
DEFA1	0	.	.	0	0	-	.	.	0	**	0	.	.	0	0	-	.	.	0	**
DEFB130	+	0	*	.	.	0	**	**	.	.	0	0	0	.	.	0	**	**	.	.
FAM72A	+	0	.	.	.	+	0	.	.	.	+	0	.	.	.	0	**	.	.	.
FAM75A1	0	0	0	.	.	0	0	0	.	.	+	*	**	.	.	+	0	**	.	.
FAM75A5	0	0	0	.	.	0	0	0	.	.	+	0	**	.	.	+	0	**	.	.
FCGBP	+	0	0	0	0	-	*	**	**	**	+	0	0	0	0	0	0	0	**	**
FOXD4L2	+	0	.	.	.	0	0	.	.	.	+	**	.	.	.	+	*	.	.	.
GOLGA6L9	0	0	0	.	.	0	0	0	.	.	+	0	*	.	.	0	0	0	.	.
GOLGA8G	+	*	**	.	.	0	0	0	.	.	+	0	0	.	.	0	0	0	.	.
GSUBP1	+	**	**	**	.	0	0	0	0	.	+	0	0	**	.	0	0	0	0	.
HIST2	+	0	.	.	.	0	0	.	.	.	+	*	.	.	.	0	0	.	.	.
LIMS3	-	*	.	.	.	0	**	.	.	.	-	0	.	.	.	0	**	.	.	.
LOC23117	0	0	0	.	.	0	*	*	.	.	-	*	**	.	.	0	*	**	.	.
LOC653606	0	0	.	.	.	0	0	.	.	.	-	**	.	.	.	+	0	.	.	.
MUC12	+	*	**	**	**	0	0	*	**	**	0	0	0	0	0	-	0	**	**	**
NBPF11	-	*	**	**	.	-	*	**	**	.	0	0	**	**	.	0	*	*	*	.
NBPF16	0	0	0	0	.	0	0	0	0	.	+	0	0	0	.	0	0	0	0	.
NPIP	-	**	**	.	.	0	*	*	.	.	-	**	**	.	.	0	*	*	.	.
PGA3	-	0	**	**	.	+	0	0	0	.	+	*	**	**	.	0	0	0	0	.
PPIAP21	+	**	**	.	.	0	0	0	.	.	+	**	**	.	.	0	0	0	.	.
PRAMEF14	+	*	**	.	.	+	**	**	.	.	+	**	**	.	.	+	0	**	.	.
PRAMEF20	0	0	.	.	.	0	0	.	.	.	+	0	.	.	.	+	0	.	.	.
PRAMEF5	-	**	**	.	.	+	*	*	.	.	-	**	**	.	.	+	*	**	.	.
PRAMEF8	0	0	0	.	.	+	0	**	.	.	0	0	0	.	.	0	0	0	.	.
PRR11	+	**	**	.	.	0	0	*	.	.	+	**	**	.	.	0	**	**	.	.
PRR20A	-	.	.	0	**	-	.	.	0	*	-	.	.	0	**	0	.	.	0	0
PSG3	+	0	*	.	.	0	0	0	.	.	0	0	0	.	.	+	0	*	.	.
RGPD1	0	0	.	.	.	+	0	.	.	.	0	0	.	.	.	+	0	.	.	.
SPYDE3	-	**	**	.	.	-	0	0	.	.	-	**	**	.	.	0	0	0	.	.
SULT1A3	0	0	0	0	.	0	0	*	*	.	-	0	**	**	.	0	**	**	**	.
TBC1D3	-	**	**	**	**	0	0	0	0	0	-	**	**	**	**	+	0	*	**	**
TCEB3C	-	0	**	**	**	0	0	0	0	0	-	0	**	**	**	0	0	0	0	0
TP53TG3	0	0	0	0	0	0	0	0	0	0	-	0	**	**	**	0	0	0	*	*
TRIM49L1	+	**	**	.	.	0	0	0	.	.	+	**	**	.	.	0	0	0	.	.
ZNF658B	+	0	**	.	.	0	0	0	.	.	+	**	**	.	.	+	0	0	.	.

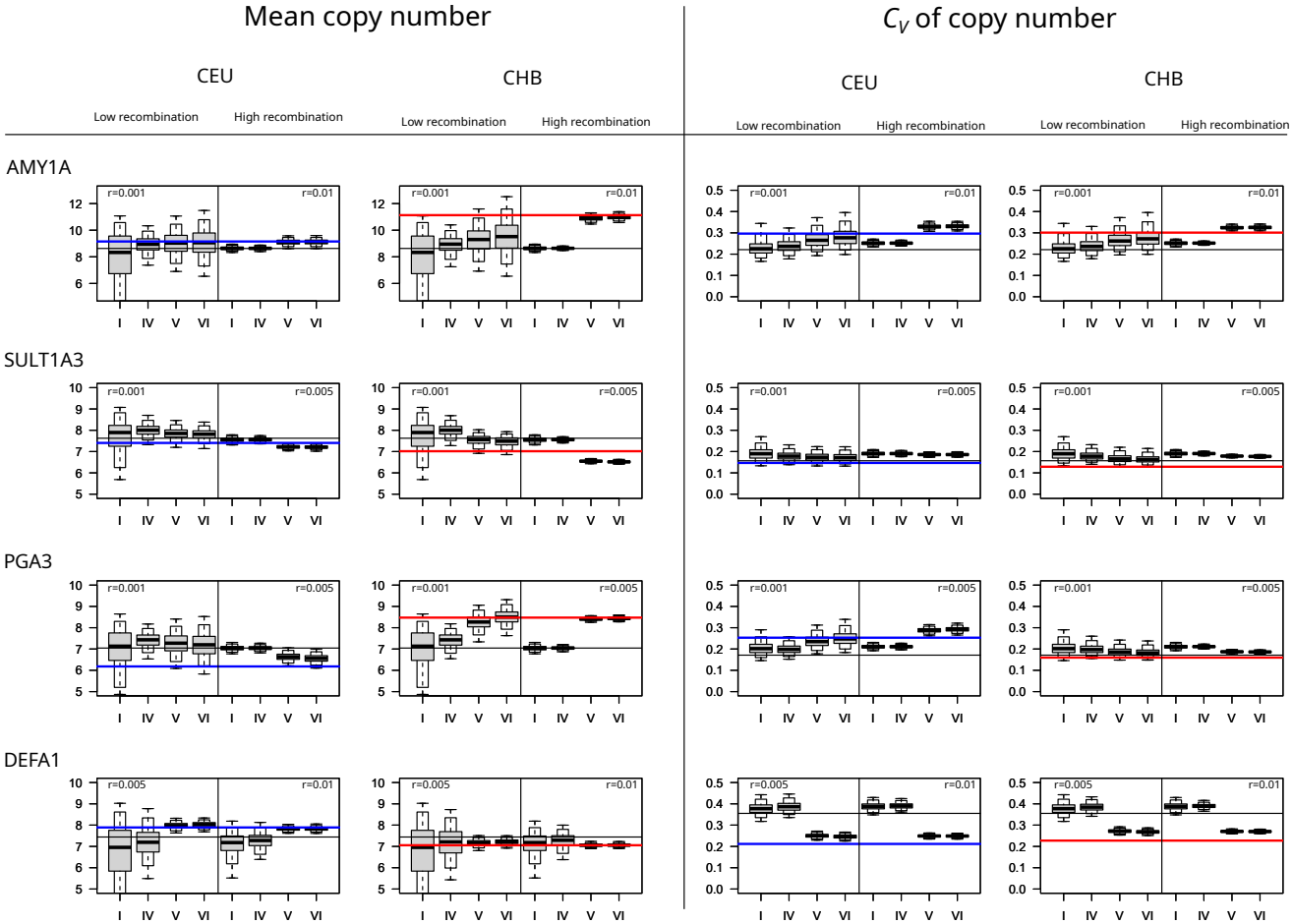

Fig 5. Comparison of mean copy number $\bar{y}$, and coefficient of variation $\mathcal{C}_V = \sigma/\bar{y}$ in four scenarios I,IV,V,VI for candidate genes `AMY1A`, `SULT1A3`, `PGA3`, `DEFA1`. In scenario V and VI the initial selection coefficients of YRI (see Table 3) were changed 500 resp. 986 generations before present. Simulation results are shown for lowest and highest recombination rate. Black lines refer to mean and $\mathcal{C}_V$ of the experimental data in YRI, blue to CEU and red to CHB.

data set YRI. Hence, under high recombination and strong selection the mean and standard deviation values of CHB and CEU show significant differences to the simulations.

However, with a change of selection the data of scenario V and VI show a different pattern. With new s_x and s_y estimated for subpopulations according to eq (3), the mean and $\mathcal{C}_V$ are shifted and the empirical data lies often within the 95%- and 99%-quantiles of the simulated data distributions. Deviations are a result of the sensitivity of the logarithmic regression. We observe no strong difference between V and VI, suggesting that 500 generations represent a sufficiently large time span to reach a new equilibrium.

Hence, the differences of the four candidate genes may be explained by adaptive processes and different selective pressures.

The AMY1A gene, which encodes amylase, an enzyme that breaks down starch, has strongly increased mean and σ in the Asian population, likely linked to adaptations to high grain intake. In the European population, while the variation is increased, the change in mean copy number is small. These findings are in agreement with results of several studies that indicate that individuals from populations with high-starch diets have, on average, more gene copies than those with traditionally low-starch diets [6, 7, 11]. Under our model selection strength is relaxed in CEU and CHB, such that a higher copy number is not selected against and a more widespread distribution of CNV can evolve. A recent study [20] has suggested a more complicated model of Amylase evolution, involving two steps: an expansion from one to several copies after the human-Neanderthal split, but before separation of modern human populations, and a subsequent shift of the optimal gene copy number, independently in different populations. This study also suggests that increase of AMY1 copy number occurred in South America even more dramatically than in East Asia, a hypothesis which should be tested in the framework of our model as soon as suitable data become available.

SULT1A3 is a gene in the SULT (sulfotransferases) family, which catalyze sulfation of a variety of substrates, especially catecholamines including dopamine and epinephrine [21, 22]. Polymorphisms in SULT1A3 and SULT1A4 have been shown to affect metabolism of therapeutic drugs (eg., [23, 24]), and these genes have therefore been studied extensively in the framework of medico- and pharmacogenetics [25, 26]. In the analyzed dataset, it has reduced mean copy number in Asia but not in Europe. There have been a few studies on copy numbers of SULT1A3/4 genes. Hildebrandt et al. (2004) [26] first noted possible duplication of SULT1A3 and identified a duplicated copy in all four different human populations. More recently, a study of 172 human individuals discovered variable SULT1A3/4 copy numbers from 1 to 10, and associated its copy number with risk and onset of neurodegenerative disease [27]. Note that SULT1A3 and SULT1A4 are closely related paralogs that are often difficult to distinguish, and studies on copy number usually put them together.

PGA3 (Pepsinogen, precursor for pepsin, an enzyme that breaks down protein to smaller peptides) is associated with prostate-specific antigen production. It is the only

gene in our list to have opposite changes in two derived populations: its mean copy number increases in Asia and decreases in Europe. As Asian and European humans share most of the same bottleneck period, the diverging copy number distribution is highly unlikely to be a demographic effect, and complex selection patterns are needed to explain the data. The copy number variation on the Pepsinogen (PGA) locus was originally discoverd with electrophoresis and three individual genes (named `PGA 3, 4, 5`) were initially found [28]. Pepsinogen genes have been shown to duplicate and become lost recurrently in vertebrates [29]. The pepsinogen genes were also shown to have variable expression level in tumor cells, particularly a reduction of PGA expression in esphagael, stomach and thyroid cancers [30]. This could be an additional source of selective pressure besides protein metabolism. While the simplest explanation is that dietary differences between Asian and European populations during the spread of of agriculture (in the last 5000-10000 years) is the driver of `PGA` copy number changes, alternative hypotheses involving tumor suppression or interaction with other enzymes must be considered.

Finally, we analyzed the immune gene Alpha-defensin `DEFA1`. It codes for defensins, proteins that are involved in innate (non-learned) immunity, specifically in antimicrobial defense against a broad spectrum of microorganisms, including bacteria, fungi, and viruses. `DEFA1` shows a decrease in variance in both Asia and Europe, indicating stronger selective pressures. More precisely, when considering the distribution in Fig 1, one observes four individuals in YRI population with high copy number which indicates a relaxed selective pressure in Africa. Alpha-defensins are expressed in neutrophil cells and intestinal epithelial cells, acting as microbiocidal agents [31–33]. The genes `DEFA1` and `DEFA3` code for some of the Alpha-defensins (HNP1/2/3), and appear to be "interchangeable variant cassettes" within a tandem array of 19kb [34]. Copy number variation of `DEFA1` is present in all apes including gibbon, but the version identified as `DEFA3` is human-specific; the copy number is also demonstrated to affect expression level [34]. Low copy number of `DEFA1/3` is shown to be associated with hospital-acquired infection [35] as well as kidney diseases [36]. On the other hand and counterintuitively, a high copy number of `DEFA1/3` may lead to more severe cases of sepsis [37, 38] and is associated with Crohn's Disease [39], and thus selected against. The trade-off between infective and autoimmune diseases could lead to selection towards

an intermediate copy number of Alpha-defensins. Therefore, our results suggest a possibility that the out-of-Africa expansion is accompanied by such a change in environmental pathogen diversity that a delicately tuned dosage of defensin is required. This can be corroborated by the fact that YRI has a few individuals with very high (outliers) copy numbers of `DEFA1`, which can not be found in CHB or CEU.

In conclusion, while both demographic effects and shifts in selection schemes can result in changes in copy number distributions, in some of our candidate genes the former is not sufficient to explain the observation. Adaptive processes can induce new relationships between copy number and fitness, and impact the resulting copy number distribution. Importantly, changes in the strength, or direction of selection may become manifest not only in mean copy number, but also in the variance or compound statistics, such as the coefficient of variation.

Acknowledgment

Funding

This work has been funded by grants from the German Research Foundation (DFG SFB-1211, subproject B6) to TW.

Conflict of interest

The authors declare no conflict of interest.

References

1. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity, population stratification, and selection of human copy-number variation. Science. 2015;349(6253). doi:10.1126/science.aab3761.

2. Brahmachary M, Guilmatre A, Quilez J, Hasson D, Borel C, Warburton P, et al. Digital Genotyping of Macrosatellites and Multicopy Genes Reveals Novel

Biological Functions Associated with Copy Number Variation of Large Tandem Repeats. PLoS Genetics. 2014;10(6):e1004418. doi:10.1371/journal.pgen.1004418.

3. Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nature Reviews Genetics. 2016;17(4):224–238. doi:10.1038/nrg.2015.25.

4. Iskow RC, Gokcumen O, Lee C. Exploring the role of copy number variants in human adaptation. Trends in Genetics. 2012;28(6):245–257. doi:10.1016/j.tig.2012.03.002.

5. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-Scale Copy Number Polymorphism in the Human Genome. Science. 2004;305(5683):525–528. doi:10.1126/science.1098918.

6. Pajic P, Pavlidis P, Dean K, Neznanova L, Romano RA, Garneau D, et al. Independent amylase gene copy number bursts correlate with dietary preferences in mammals. eLife. 2019;8. doi:10.7554/elife.44628.

7. Atkinson FS, Hancock D, Petocz P, Brand-Miller JC. The physiologic and phenotypic significance of variation in human amylase gene copy number. The American Journal of Clinical Nutrition. 2018;108(4):737–748. doi:10.1093/ajcn/nqy164.

8. Carpenter D, Dhar S, Mitchell LM, Fu B, Tyson J, Shwan NAA, et al. Obesity, starch digestion and amylase: association between copy number variants at human salivary (AMY1) and pancreatic (AMY2) amylase genes. Human Molecular Genetics. 2015;24(12):3472–3480. doi:10.1093/hmg/ddv098.

9. Usher CL, Handsaker RE, Esko T, Tuke MA, Weedon MN, Hastie AR, et al. Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. Nature Genetics. 2015;47(8):921–925. doi:10.1038/ng.3340.

10. Falchi M, Moustafa JSES, Takousis P, Pesce F, Bonnefond A, Andersson-Assarsson JC, et al. Low copy number of the salivary amylase gene predisposes to obesity. Nature Genetics. 2014;46(5):492–497. doi:10.1038/ng.2939.

11. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. Nature Genetics. 2007;39(10):1256–1260. doi:10.1038/ng2123.

12. Lohmueller KE. The Impact of Population Demography and Selection on the Genetic Architecture of Complex Traits. PLoS Genetics. 2014;10(5):e1004379. doi:10.1371/journal.pgen.1004379.

13. Stajich JE. Disentangling the Effects of Demography and Selection in Human History. Molecular Biology and Evolution. 2004;22(1):63–73. doi:10.1093/molbev/msh252.

14. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123(3):585–595. doi:10.1093/genetics/123.3.585.

15. Fu YX. Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. Genetics. 1997;147(2):915–925. doi:10.1093/genetics/147.2.915.

16. Otto M, Zheng Y, Wiehe T. Recombination, selection, and the evolution of tandem gene arrays. Genetics. 2022;221(3). doi:10.1093/genetics/iyac052.

17. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, et al. Diversity of Human Copy Number Variation and Multicopy Genes. Science. 2010;330(6004):641–646. doi:10.1126/science.1197005.

18. Noskova E, Ulyantsev V, Koepfli KP, O'Brien SJ, Dobrynin P. GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data. GigaScience. 2020;9(3). doi:10.1093/gigascience/giaa005.

19. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023;617(7960):312–324. doi:10.1038/s41586-023-05896-x.

20. Inchley CE, Larbey CDA, Shwan NAA, Pagani L, Saag L, Antao T, et al. Selective sweep on human amylase genes postdates the split with Neanderthals. Scientific Reports. 2016;6(1). doi:10.1038/srep37198.

21. Brix LA, Barnett AC, Duggleby RG, Leggett B, McManus ME. Analysis of the Substrate Specificity of Human Sulfotransferases SULT1A1 and SULT1A3: Site-Directed Mutagenesis and Kinetic Studies. Biochemistry. 1999;38(32):10474–10479. doi:10.1021/bi990795q.

22. Dajani R, Sharp S, Graham S, Bethell SS, Cooke RM, Jamieson DJ, et al. Kinetic Properties of Human Dopamine Sulfotransferase (SULT1A3) Expressed in Prokaryotic and Eukaryotic Systems: Comparison with the Recombinant Enzyme Purified fromEscherichia coli. Protein Expression and Purification. 1999;16(1):11–18. doi:10.1006/prep.1999.1030.

23. Hui Y, Liu MC. Sulfation of ritodrine by the human cytosolic sulfotransferases (SULTs): Effects of SULT1A3 genetic polymorphism. European Journal of Pharmacology. 2015;761:125–129. doi:10.1016/j.ejphar.2015.04.039.

24. Bairam AF, Rasool MI, Alherz FA, Abunnaja MS, Daibani AAE, Gohal SA, et al. Impact of SULT1A3/SULT1A4 genetic polymorphisms on the sulfation of phenylephrine and salbutamol by human SULT1A3 allozymes. Pharmacogenetics and Genomics. 2019;29(5):99–105. doi:10.1097/fpc.0000000000000371.

25. Thomae BA, Rifki OF, Theobald MA, Eckloff BW, Wieben ED, Weinshilboum RM. Human catecholamine sulfotransferase (SULT1A3) pharmacogenetics: functional genetic polymorphism. Journal of Neurochemistry. 2004;87(4):809–819. doi:10.1046/j.1471-4159.2003.02027.x.

26. Hildebrandt MAT, Salavaggione OE, Martin YN, Flynn HC, Jalal S, Wieben ED, et al. Human SULT1A3 pharmacogenetics: gene duplication and functional genomic studies. Biochemical and Biophysical Research Communications. 2004;321(4):870–878. doi:10.1016/j.bbrc.2004.07.038.

27. Butcher NJ, , Horne MK, Mellick GD, Fowler CJ, Masters CL, et al. Sulfotransferase 1A3/4 copy number variation is associated with neurodegenerative disease. The Pharmacogenomics Journal. 2017;18(2):209–214. doi:10.1038/tpj.2017.4.

28. Taggart RT, Mohandas TK, Shows TB, Bell GI. Variable numbers of pepsinogen genes are located in the centromeric region of human chromosome 11 and

determine the high-frequency electrophoretic polymorphism. Proceedings of the National Academy of Sciences. 1985;82(18):6240–6244. doi:10.1073/pnas.82.18.6240.

29. Castro LFC, Goncalves O, Mazan S, Tay BH, Venkatesh B, Wilson JM. Recurrent gene loss correlates with the evolution of stomach phenotypes in gnathostome history. Proceedings of the Royal Society B: Biological Sciences. 2014;281(1775):20132669. doi:10.1098/rspb.2013.2669.

30. Shen S, Li H, Liu J, Sun L, Yuan Y. The panoramic picture of pepsinogen gene family with pan-cancer. Cancer Medicine. 2020;9(23):9064–9080. doi:10.1002/cam4.3489.

31. Ganz T, Selsted ME, Szklarek D, Harwig SS, Daher K, Bainton DF, et al. Defensins. Natural peptide antibiotics of human neutrophils. Journal of Clinical Investigation. 1985;76(4):1427–1435. doi:10.1172/jci112120.

32. Ayabe T, Satchell DP, Wilson CL, Parks WC, Selsted ME, Ouellette AJ. Secretion of microbicidal α-defensins by intestinal Paneth cells in response to bacteria. Nature Immunology. 2000;1(2):113–118. doi:10.1038/77783.

33. Nassar H, Lavi E, Akkawi S, Bdeir K, Heyman SN, Raghunath PN, et al. α-Defensin: Link between inflammation and atherosclerosis. Atherosclerosis. 2007;194(2):452–457. doi:10.1016/j.atherosclerosis.2006.08.046.

34. Aldred PMR, Hollox EJ, Armour JAL. Copy number polymorphism and expression level variation of the human α-defensin genes DEFA1 and DEFA3. Human Molecular Genetics. 2005;14(14):2045–2052. doi:10.1093/hmg/ddi209.

35. Zhao J, Gu Q, Wang L, Xu W, Chu L, Wang Y, et al. Low-Copy Number Polymorphism in DEFA1/DEFA3 Is Associated with Susceptibility to Hospital-Acquired Infections in Critically Ill Patients. Mediators of Inflammation. 2018;2018:1–8. doi:10.1155/2018/2152650.

36. Ai Z, Li M, Liu W, Foo JN, Mansouri O, Yin P, et al. Low α-defensin gene copy number increases the risk for IgA nephropathy and renal dysfunction. Science Translational Medicine. 2016;8(345). doi:10.1126/scitranslmed.aaf2106.

37. Chen Q, Hakimi M, Wu S, Jin Y, Cheng B, Wang H, et al. Increased Genomic Copy Number of DEFA1/DEFA3 Is Associated with Susceptibility to Severe Sepsis in Chinese Han Population. Anesthesiology. 2010;112(6):1428–1434. doi:10.1097/aln.0b013e3181d968eb.

38. Chen Q, Yang Y, Hou J, Shu Q, Yin Y, Fu W, et al. Increased gene copy number of DEFA1/DEFA3 worsens sepsis by inducing endothelial pyroptosis. Proceedings of the National Academy of Sciences. 2019;116(8):3161–3170. doi:10.1073/pnas.1812947116.

39. Jespersgaard C, Fode P, Dybdahl M, Vind I, Nielsen OH, Csillag C, et al. Alpha-Defensin DEFA1A3 Gene Copy Number Elevation in Danish Crohn's Disease Patients. Digestive Diseases and Sciences. 2011;56(12):3517–3524. doi:10.1007/s10620-011-1794-8.

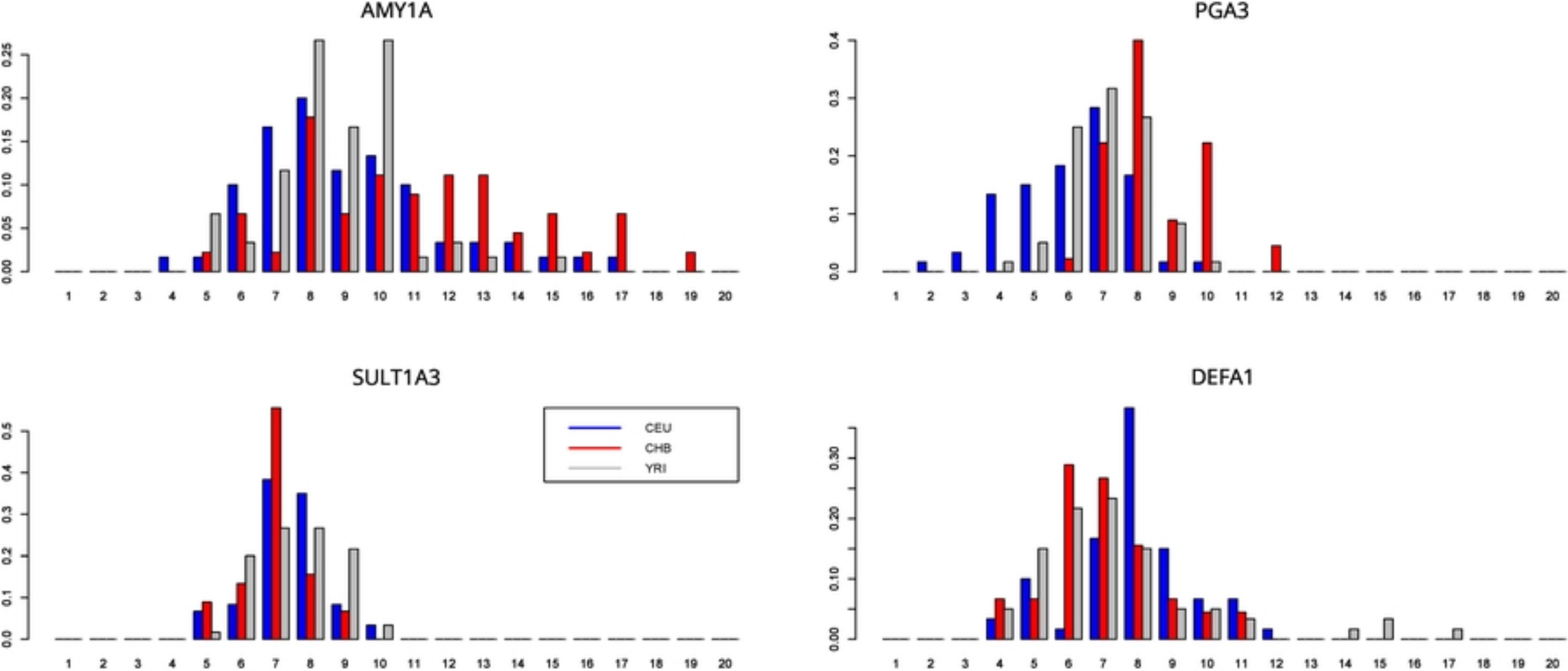

Figure-1

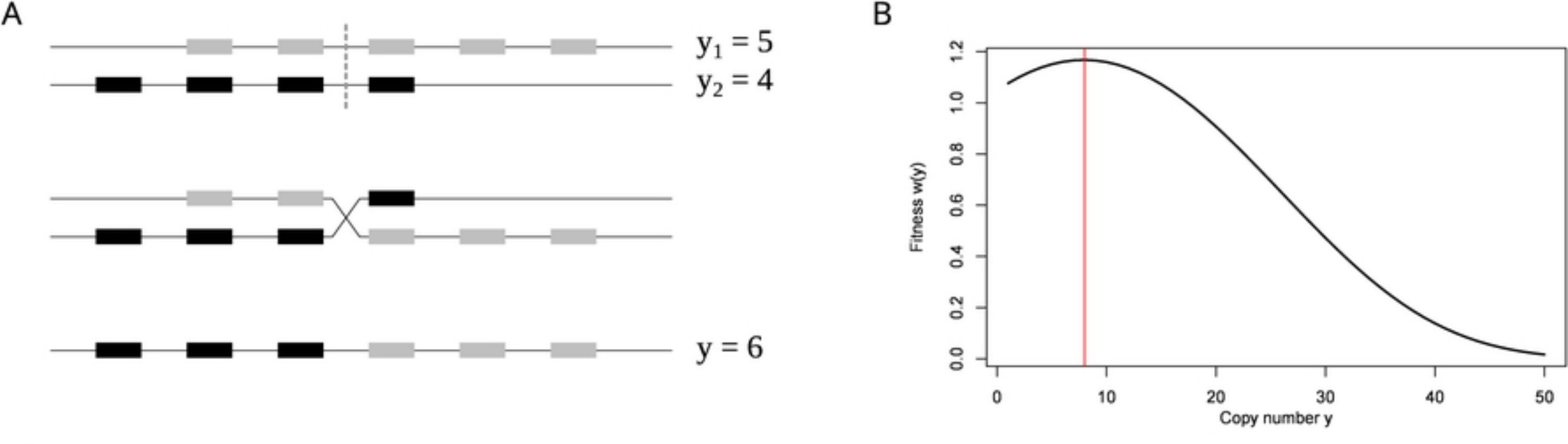
A
$y_1 = 5$
$y_2 = 4$
$y = 6$
B
Fitness w(y)
Copy number y
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
10
20
30
40
50

Figure-2

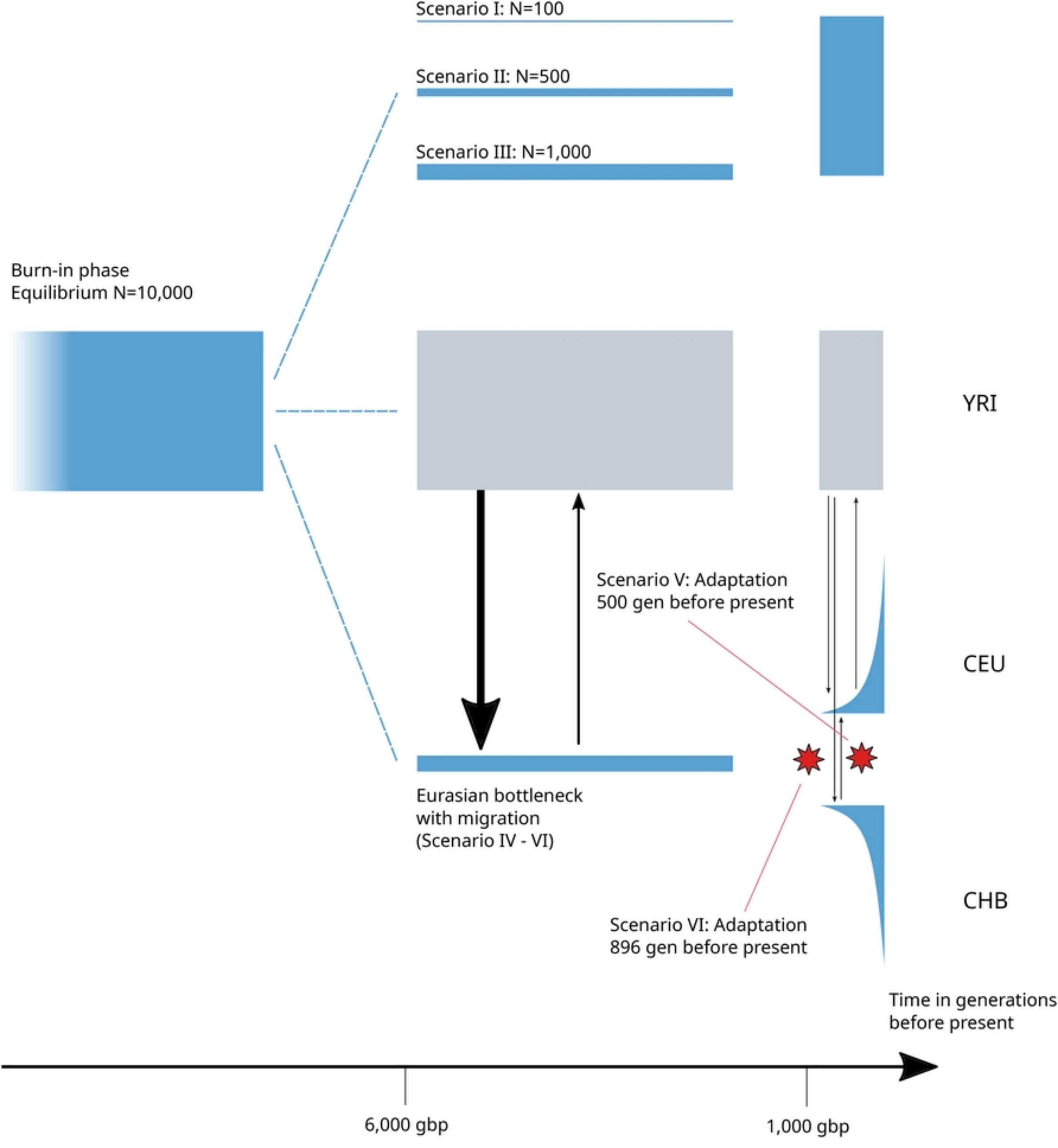
Scenario I: N=100
Scenario II: N=500
Scenario III: N=1,000
Burn-in phase
Equilibrium N=10,000
YRI
Scenario V: Adaptation
500 gen before present
CEU
Eurasian bottleneck
with migration
(Scenario IV - VI)
CHB
Scenario VI: Adaptation
896 gen before present
Time in generations
before present
6,000 gbp
1,000 gbp

Figure-3

Bottleneck simulations
PGA3

Low recombination
r = 0.001

Medium recombination
r = 0.002

High recombination
r = 0.005

Scenario I
Strong bottleneck
Reduction to N = 100

Scenario II
Intermediate bottleneck
Reduction to N = 500

Scenario III
Weak bottleneck
Reduction to N = 1,000

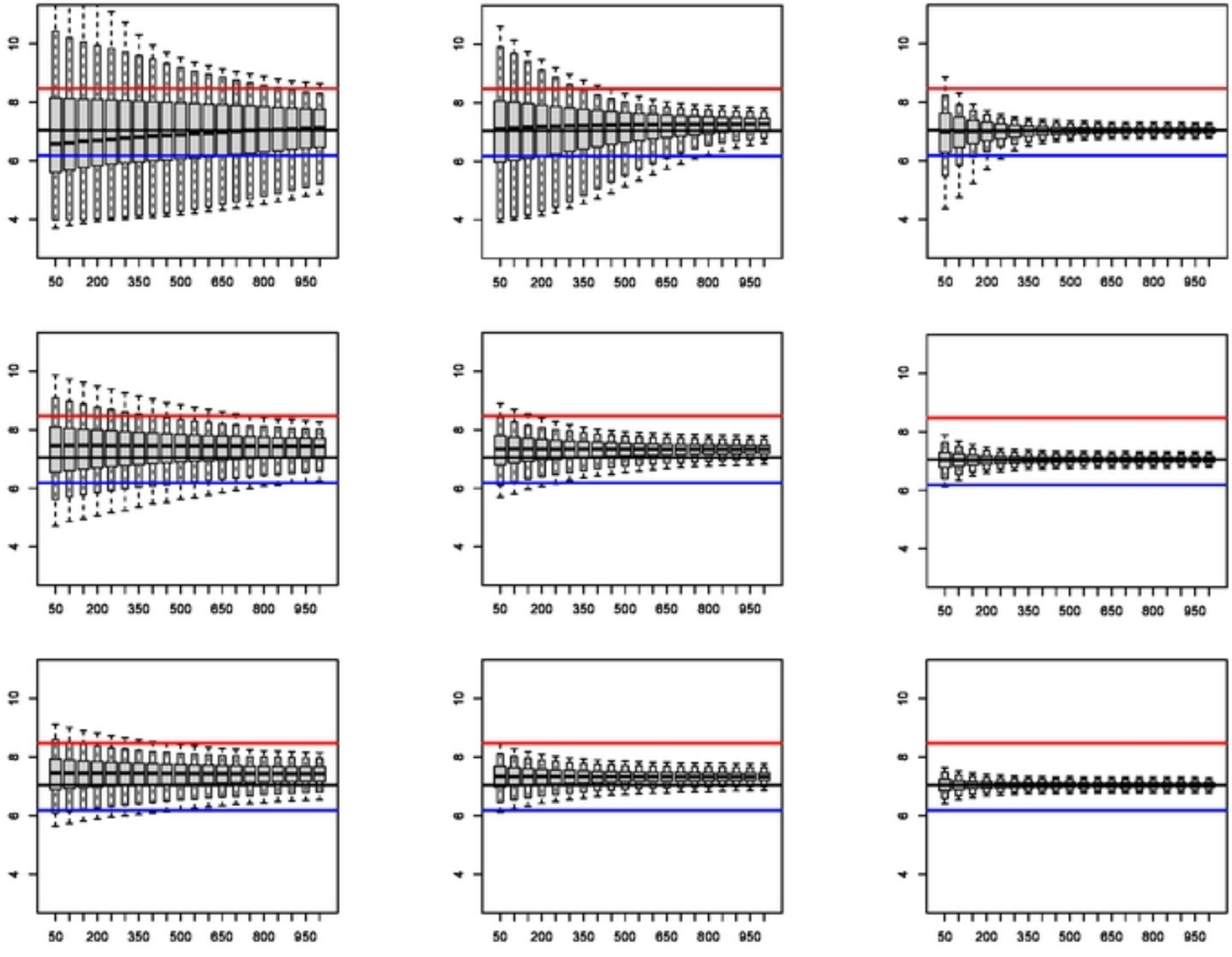

Figure-4

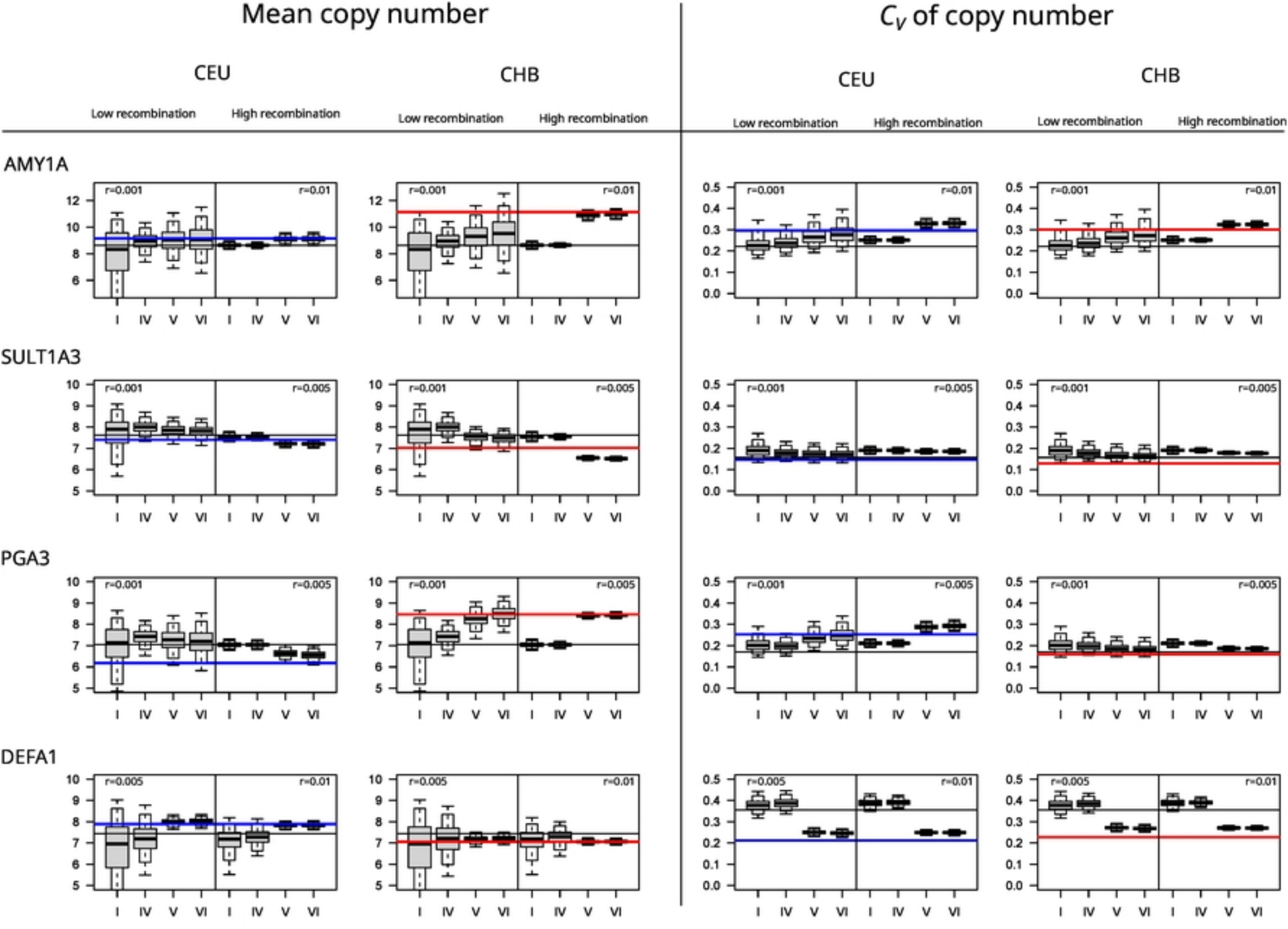
Mean copy number
C_V of copy number
CEU
CHB
Low recombination
High recombination
AMY1A
SULT1A3
PGA3
DEFA1
r=0.001
r=0.01
r=0.005
I
IV
V
VI

Figure-5