www.nature.com/hgv

ARTICLE OPEN

Detecting adaptive changes in gene copy number distribution accompanying the human out-of-Africa expansion

Moritz Otto[1,3], Yichen Zheng[1,3], Paul Grablowitz[2] and Thomas Wiehe[1]✉

Genes with multiple copies are likely to be maintained by stabilizing selection, which puts a bound to unlimited expansion of copy number. We designed a model in which copy number variation is generated by unequal recombination, which fits well with several genes surveyed in three human populations. Based on this theoretical model and computer simulations, we were interested in determining whether the 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 the copy number distribution in most of the analyzed gene clusters can be explained by a bottleneck, such as in the out-of-Africa expansion of Homo sapiens 60–10 kyrs ago, we identified several candidate genes, such as **AMY1A** and **PGA3**, whose copy numbers are likely to differ among African, Asian, and European populations.

Human Genome Variation (2024) 11:1–10; https://doi.org/10.1038/s41439-024-00293-w

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 the human population is the salivary amylase gene, whose variation in the number of copies is hypothesized to correlate with the extent of dietary starch consumption not only in humans but also in other species[6–11].

In general, CNV 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.

Both demographic effects and selection may produce similar patterns in single nucleotides 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[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 CNV data. Therefore, we resort to a more basic approach involving modeling and computer simulations.

We have recently examined the evolutionary dynamics of multicopy 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 play 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 CNV in human populations. Based on empirical data of human gene copy numbers, we identified several candidate genes whose copy numbers are likely to be selected differently among African, Asian, and European populations.

MATERIALS AND METHODS

Gene CNV in humans

We started with the dataset provided by Brahmachary et al.[1]. Using NanoString technology, they estimated the gene copy numbers of 180 gene families in 165 individuals of three populations (60 African Yoruba – YRI, 60 Central Europe – CEU, and 45 Asia – CHB) based on data collected in the framework of the 1000 Genomes Project[17].

While some of these loci presented 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 numbers less than 2, and genes with mean copy numbers (in YRI) <5 or >60. For genes that have two primer sets, only one is used.

The filtering procedure is described in the supplementary material, which resulted in 49 gene families. For these analyses, we used t tests and F tests to select gene families with significant differences in means or standard deviations between the YRI–CHB or YRI–CEU comparisons and removed those that showed no statistical evidence ($\alpha = 5\%$) in any of these comparisons. The remaining 42 gene families are shown in Table 1, and the copy number distributions of four of them are shown in Fig. 1.

[1]Institue for Genetics, University of Cologne, Cologne, Germany. [2]Department of Computer Science, University of Tübingen, Tübingen, Germany. [3]These authors contributed equally: Moritz Otto, Yichen Zheng. ✉email: twiehe@uni-koeln.de

Received: 6 March 2024 Revised: 5 July 2024 Accepted: 22 July 2024
Published online: 23 September 2024

Table 1. Differences in means and variances between populations.

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

Zero indicates no significant change, '+' indicates a significant increase, and '−' indicates a significant decrease (*t* test for the mean and F test for the standard deviation; $\alpha = 0.05$). The four candidate genes shown in Fig. 1 are highlighted with a light gray background.

Unequal recombination model

In a recently developed model, we considered unequal recombination and selection to describe the evolution of tandem gene arrays[16].

We briefly summarize the main findings. Consider two chromosomes with gene arrays of size y_1 and y_2. A recombination event occurs at rate *r* and may produce a gamete of gene array size according to the trapezoidal distribution, such that

$$Prob(y|y_1, y_2) = \frac{1}{y_1 y_2} \begin{cases} 0, & y < 1 \\ y, & 1 \le y < \min(y_1, y_2) \\ \min(y_1, y_2), & \min(y_1, y_2) \le y < \max(y_1, y_2) \\ y_1 + y_2 - y, & \max(y_1, y_2) \le y < y_1 + y_2 - 1 \\ 0, & y \ge y_1 + y_2 \end{cases}$$

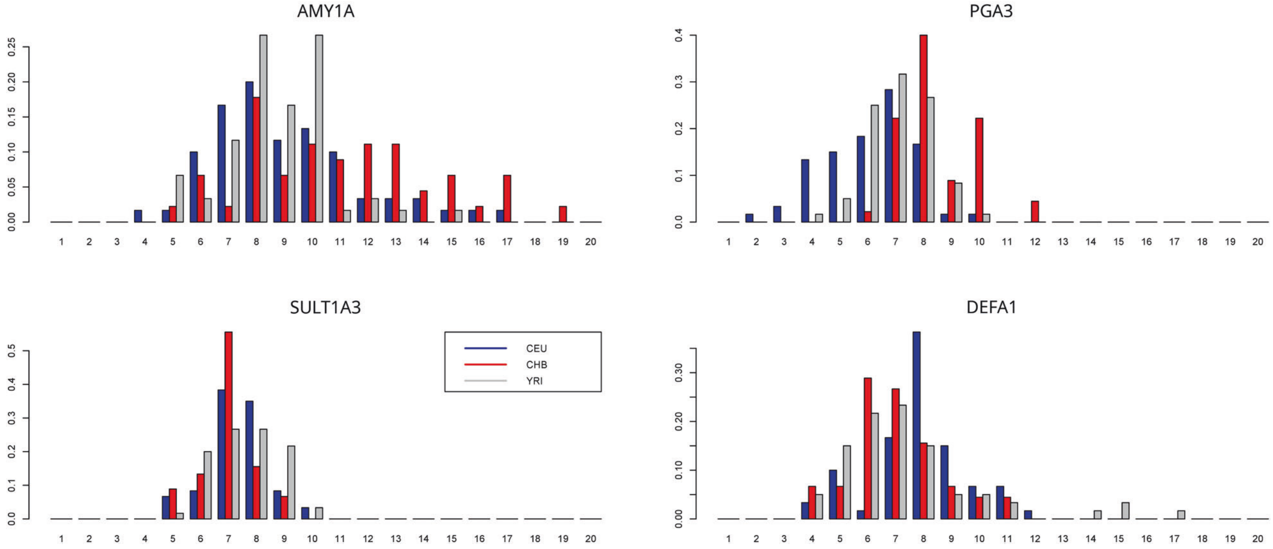

Fig. 1 Gene copy number distribution in four exemplary gene families in three human populations: CEU, CHB, and YRI. The data were adapted from Brahmachary et al.[1]. The *y* axis indicates the frequency of individuals carrying the number of copies shown on the *x* axis.

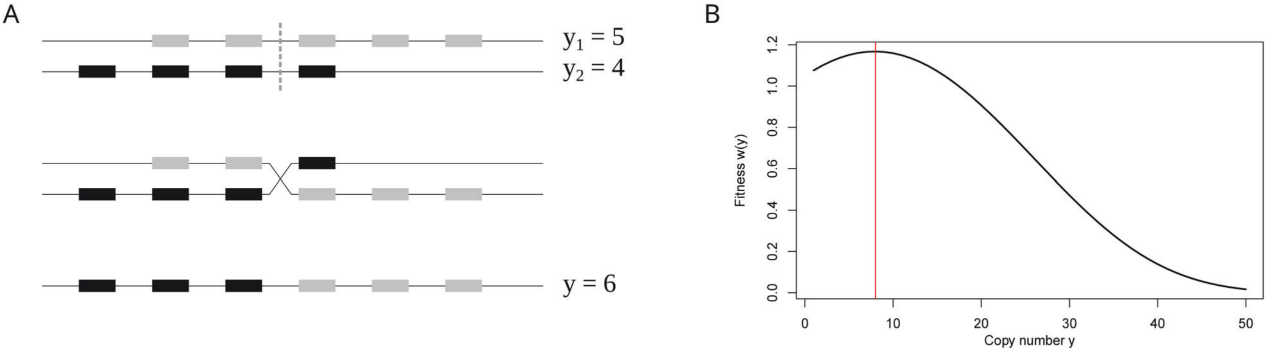

Fig. 2 Recombination process and fitness function. A Sketch of the unequal recombination process. Starting with two chromosomes with $y_1 = 5$ and $y_2 = 4$ gene copies, two break points are chosen. One of the recombinants is then propagated. Its copy number (here $y = 6$) is trapezoidal, as shown in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

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 the assumption of 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 that additional copies are selected with increasing selective disadvantage s_y. This is motivated by the increasing cost of replication, gene processing and 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 trivial long-term evolution equilibrium of one copy, we assume that $s_x > s_y$. Furthermore, we assume that $\varepsilon = 0.05$ is constant. In summary, the fitness of a diploid individual with total gene copy number *y* is given by

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

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

$$y_{opt} = 1 + \frac{\log\left(s_x/s_y\right)}{2 \cdot \varepsilon} \qquad (2)$$

which is determined by the ratio s_x/s_y when ε is fixed. See Fig. 2B for an example. The population is then simulated according to a Wright–Fisher model with nonoverlapping generations and with selection and recombination described above. In the deterministic model, the equilibrium copy number distribution is centered on y_{opt} and is well approximated by a gamma distribution[16]. Furthermore, the coefficient of variation $C_V = \sigma/y$ is correlated with 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. For convenience, 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

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

- $q_R = r/s_x$, the *'recombination/selection ratio'*, which measures the impact of unequal recombination compared with the selective pressure of the fitness function and therefore determines the coefficient of variation $C_V = \sigma/y$ of the equilibrium distribution.

Note that this selection model affects solely the copy number of a gene family, not its sequence. We calculated Tajima's *D* in a 1 Mb interval around each copy of the 42 gene families in sliding windows of 10 kb via data from the 1000 Genomes Project [http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/][18] to test whether selective sweeps may also affect the copy number distribution. Furthermore, we also tested whether these regions are affected by the introgression of archaic hominins. With the data of Browning et al.[19], we counted the number of introgressed SNPs of Neanderthals or Denisovans in an interval of 10 Mb around the candidate genes to test whether admixture affected the copy number distribution in humans.

Regression

We aim to quantify the effect of *(r, s_x, s_y)* on the resulting equilibrium copy number distribution and, vice versa, to estimate the underlying parameter triple for given empirical data. We simulated the population evolution under different parameter settings to analyze the equilibrium distribution of the unequal recombination process under drift. The codes for all the following simulations are available on GitHub https://github.com/y-zheng/gCNV-human. The population size is kept at $N = 5000$ 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 maintain a realistic parameter range.

For each of these 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 *y* and standard deviation *σ*) were recorded every 20,000 generations. Note that in contrast to the deterministic model with an infinitely large population size, the population does not reach a stationary distribution but rather fluctuates around the equilibrium distribution. However, since we included an extensive burn-in phase, we are confident that the population is close to equilibrium.

In total, this results in ≈185,000 data points, which we used to determine the relationships between the input parameters *(r, s_x, s_y)* and the output population statistics *(y, σ)*.

As indicated in Otto et al.[16], we suggest a mean copy number *y* close to its optimal value y_{opt} and a correlation of the C_V to $log(q_R)$. Indeed, with coefficients of determination (r^2) of 0.9842 and 0.9088, we find

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

We calculated the q_S and q_R ratios based on *y* and C_V from gene copy numbers (see Table 1) via regression Formula (3), with four recombination rates, $r = 0.001$, 0.002, 0.005, and 0.01.

The results for the four candidate genes shown in Fig. 1 are given in Table 3.

Table 2. Parameters for regression simulations.

4 recombination rates *r*	0.1%, 0.2%, 0.5%, and 1%
9 recombination/selection 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

Demography simulations

To determine whether significant changes in the mean and variance of the copy number distribution (Table 1) can be explained by the demographic history of human populations, we examined a total of 6 different scenarios (enumerated as I–VI), as shown in 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 inputs to produce an equilibrium population of $N = 10{,}000$ via 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 1000, denoted scenarios I, II, and III, and maintained it for 5000 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 Table 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 resulted in a total of 42 gene families and 95 gene-*r* combinations. For each gene, recombination rate and bottleneck population size combination, 10,000 replicates are produced (from 100 'independent' starting equilibria). We then traced the means and C_V along the recovery phase and compared them with the empirical data from the CHB and CEU populations.

Simulation of human population history A more realistic population history of humans is given by the genetic algorithm for demographic model analysis (GADMA)[20], 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 an ancestral population ($N = 9900$ in GADMA), we used the equilibrium populations ($N = 10{,}000$) from the previous section. Therefore, we started the simulation 5992 generations before the 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 an exponentially increasing population size. In the following, we refer to this simulation as scenario IV.

Table 3. Estimates of selection coefficients s_x, s_y under four recombination rates $r = 0.001\ldots,0.01$ based on regression Eq. (3).

					$r = 0.001$		$r = 0.002$		$r = 0.005$		$r = 0.01$	
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

The displayed gene families are the ones in Fig. 1 for all three populations. The values in parentheses are outside the range of $0.001 < s_x < 0.1$ in YRI and hence are not used in simulations.

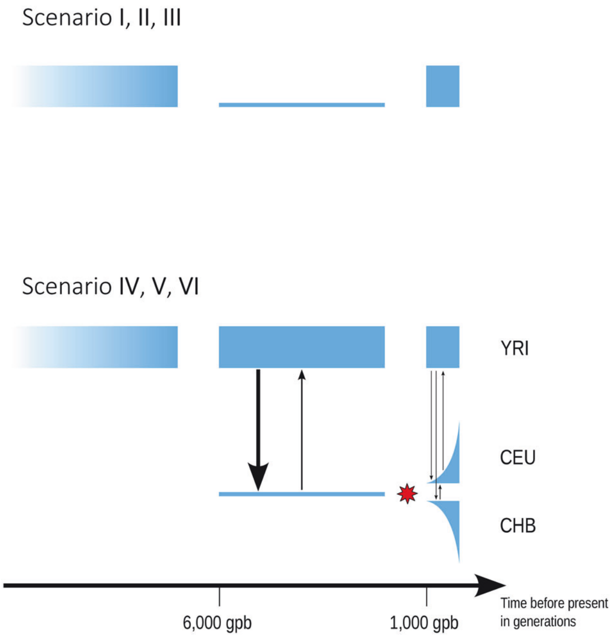

Fig. 3 Illustration of the six different scenarios investigated. Scenarios I–III: simple bottleneck lasting 5000 generations. Reduction to 1% (scenario I), 5% (II), and 10% (III) of its original size ($N = 10{,}000$). Scenarios IV–VI: GADMA model of human demographic history. There was no change in selection intensity (scenario IV), with a change in selection intensity (represented by the red star) 896 generations ago (scenario V) and 500 generations ago (VI). The black arrows indicate the direction and frequency of migration between subpopulations.

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 in 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 the CEU and CHB populations and are estimated from the present CEU/CHB distributions (see Table 3). These simulations are hereafter called scenario V (896 generations of selection change before present) and VI (500 gpb).

Rejecting a purely demographic model

By observing the copy number distribution for a gene family in the ancestral (YRI) population, we seek to answer the question of whether the observed distribution in the derived population (CEU or CHB) can be explained by a purely demographic model (various bottlenecks but keeping selection pressure constant as modeled in scenarios I to IV) or not (demography plus change in selection pressure as modeled in scenarios V or VI). We use the following strategy to decide this.

For each scenario I–VI and each parameter triple estimated from the YRI population, 10,000 replicates were produced. From each resulting equilibrium distribution, we record the mean y and standard deviation σ. This results in a parameter distribution for each scenario. If a chosen empirical dataset has a mean y or standard deviation σ, which are not in the 95% quantile of the 10,000 simulated values, we conclude that this scenario is rejected as a possible explanation of the data. We reject a purely demographic explanation if scenarios I–IV are rejected.

RESULTS AND DISCUSSION

In this study, we conducted an analysis of multicopy gene family evolution via 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 to either constant selective pressure and demographic factors or an adaptive change in selection together with demography. Our findings reveal that the observed CNVs in several genes cannot adequately be explained by demographic processes alone, suggesting a possibly adaptive change in selective pressure in the derived populations.

Based on the data of Brahmachary et al.[1], we chose 42 gene families with intermediate copy numbers that presented significant differences in distribution among different populations (Table 1).

Although the raw data rely on phase I of the 1000 Genomes project, they proved to be most suitable for our analyses. More recent data, for example, from the human pangenome project[21], still lack sufficient coverage of the different subpopulations.

When we compared the copy number distributions of the 42 candidate genes in the Asian and European populations with those in the African population (assumed to be in equilibrium), we observed 61 significant changes in the mean copy number and 29 significant changes in the variance (Table 1), of which only seven showed a decrease in variance (one example was **DEFA1**). Within our model, assuming a constant recombination rate among subpopulations and no demographic changes, a decreased variance (or standard deviation) can be achieved only by an increase in positive selection (s_x), since the C_V is determined by $q_R = r/s_x$; see Eq. (3). However, the most common case is that of a consistent significant shift in the mean in both derived populations without affecting the variance, i.e., either $\mathbf{(++|\ 0\ 0)}$ or $\mathbf{(--|\ 0\ 0)}$, which occurs in 12 of the 42 analyzed genes. Only one gene (**PGA3**) showed opposite significant changes in the mean (increasing in Asia but decreasing in Europe).

In this model from Otto et al.[16], selection does not act on allele sequences, but it acts on the number of gene copies such that an individual with y_{opt} (2) many copies has the highest fitness. Several methods exist to detect sequence traces of recent selective sweeps, ranging from classical statistics, such as Tajima's D[15] or Fu and Way's *H*[22], to recent machine learning methods—see, for example, Lauterbur et al.[23]. It is conceivable that selective sweeps in regions close to CNV loci may, by hitchhiking, also affect the gene copy number distribution. A prominent example is the selective sweep in the Asian population in the **EDAR** gene[24], which is located at ~300 kb to **LIMS1** and ~500 kb to the **LIMS3** gene family from our set. To determine how generally this phenomenon might be and whether selection on allele sequences may correlate, or coincide, with selection on gene copy number, we calculated Tajima's *D* in an area of 1 Mb approximately 33 MANE-selected genes[25]. Our results suggest that these two forces are independent of each other. For example, **AMY1A** shows a negative Tajima's *D* in CHB, positive in CEU and approximately zero in YRI (see Supplementary material), but the estimated selection coefficients s_x for a recombination rate of $r = 0.005$ (see Table 3) are 0.6% in CHB, 0.5% in CEU and 2.4% in YRI. **LIMS3** showed only small differences in estimates of s_x (8.3% in CEU, 7.7% in CHB and 9.1% in YRI for $r = 0.001$), and the strong selective sweep of **EDAR** in CHB seems to have no effect.

However, sequence-based summary statistics (such as Tajima's *D*) calculated across CNV loci need to be interpreted with caution: correct alignment and correct variant calling remain challenging problems, especially when gene copies are highly similar. Biased test statistics are imminently dangerous in such cases[26].

Another force that may affect the copy number distribution is introgression. Derived populations are known to have interbred with archaic hominins (i.e., Neanderthals and Denisovans). If the archaic hominins had a greater copy number than did *Homo sapiens*, this might have had a significant effect on the distribution of the derived population. Hence, we counted the introgressed SNPs of Neanderthals and Denisovans[19] in an area of 10 Mb around the candidate gene families. We find the analyzed genes to seem to be in *introgression deserts*[27], i.e., regions that lack introgression (see Supplementary). However, as a Markovian

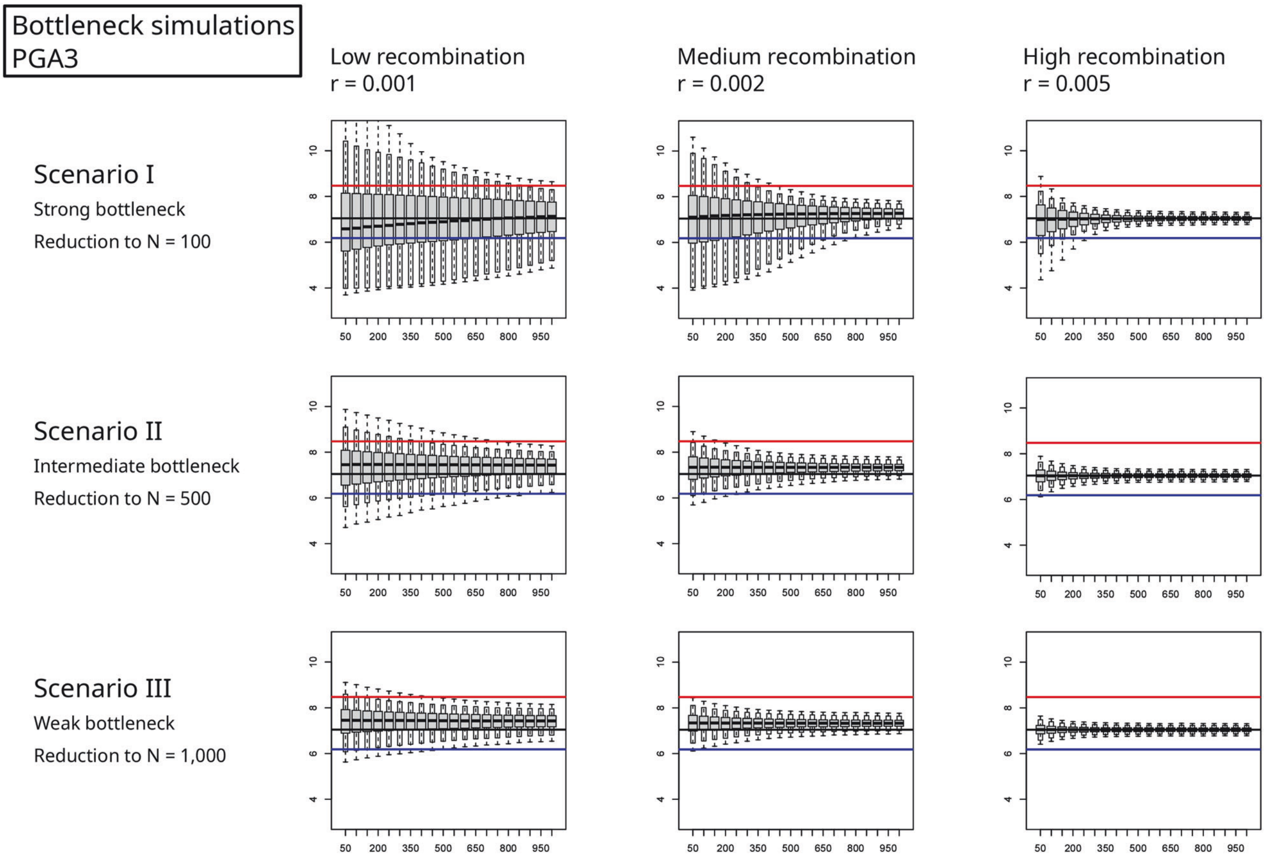

Fig. 4 Mean copy number over time. After population reduction to $N = 100$, 500, or 1000 (top to bottom), we traced the mean value (*y* axis) of 10,000 replicates over time (*x* axis in generations). The input parameters s_x and s_y were estimated for $r = 0.001$, 0.002, and 0.005 (from left to right) from the YRI dataset for the candidate gene **PGA3** (see Table 3) and kept constant over time to determine the effects of the bottleneck and recovery. Whiskers indicate the 99% quantile, and white boxes indicate the 95% quantile. The horizontal lines mark the values from the original dataset of Brahmachary et al.[1] (black: YRI; red: CHB; blue: CEU).

process, the equilibrium distribution of gene copy number is independent of its initial distribution and hence robust against outliers. Therefore, even if an archaic hominins would have introduced a higher (or lower) copy number, the effect would vanish over time as the distribution would return to its derived equilibrium. In addition, the same note of caution toward technical shortcomings, as mentioned above, applies here as well: marker allele mapping to a gene family may be counted only once rather than multiple times, thereby leading to an underestimation of the amount of introgression.

To test whether a change in population size is sufficient to explain these significant differences in copy number statistics (shown in Table 1), we ran simple bottleneck scenarios I–III with 95 parameter combinations *(r, s_x, s_y)* on the basis of the ancestral copy number distribution of YRI and the regression Eq. (3). As an example, for **PGA3** and $r = 0.001$, we ran simulations with selection coefficients of $s_x = 0.0122$ and $s_y = 0.0066$; see Table 3. Fig. 4 shows the mean gene copy number of 10,000 simulated bottleneck populations over time for each recombination strength ($r = 0.001$, 0.002, and 0.005). Note that for this gene, the value of $r = 0.01$ was neglected since the selection strength s_x would exceed the threshold of 0.1. The gray boxes indicate the centered 50% quantile, the white boxes indicate the 95% quantile, and the whiskers indicate the 99% quantile. With a strong bottleneck (reduction to $N = 100$ for 5000 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. A higher *r* and stronger selection result in a mean value close to that of the YRI population, i.e., the value we would expect from the initial parameter estimation.

In this example, the empirical data show a significantly greater mean copy number of **PGA3** in CHB (red line) and a lower mean value in CEU (blue line) than in YRI (black horizontal line). It is the only gene in our set that shows significant shifts in the mean copy number in opposite directions. These changes could be explained only under a strong bottleneck and with low recombination without invoking a change in selection intensity.

The results for all 95 parameter combinations obtained for scenario I (strong bottleneck) are summarized in Table 4. To test whether the observed means or standard deviations can be explained by scenario I, we considered the time point after 1000 generations of recovery (first row of Fig. 4, last boxplot in each panel). If the mean or C_V lies within the 95% quantile, we indicate nonsignificant differences with a value of 0. Significant changes are marked with a single * ($\alpha = 5\%$) or double asterisk ** ($\alpha = 1\%$). Taking again **PGA3** as an example, we find a mean value that is significantly smaller in CEU than in YRI (marked with –). When $r = 0.001$, this might be explained by a bottleneck (denoted by 0), whereas when $r = 0.002$ and $r = 0.005$, we find a significant difference (**) and that the bottleneck explanation is highly unlikely. Higher recombination ($r = 0.01$) led to s_x values >0.1 in CHB and YRI (see Table 3) and hence were omitted.

If we consider a significant difference in the mean of CEU compared with YRI (28 genes; the first column in Table 4 is nonzero), we find that only 17 out of 65 simulated parameter combinations in scenario I can explain these differences. For significant mean changes in CHB (33 genes; Table 4), 22 out of 72 parameter

Table 4. Results of bottleneck simulations.

	CEU mean					CEU sd					CHB mean					CHB sd				
Gene	***t-test***	**r1**	**r2**	**r5**	**r10**	***F-test***	**r1**	**r2**	**r5**	**r10**	***t-test***	**r1**	**r2**	**r5**	**r10**	***F-test***	**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	.	.

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 could be explained by a bottleneck. The blank space indicates that this parameter combination led to an s_x value out of the range of 0.001–0.1; hence, no simulation was run. The columns with 0, + and – indicate whether there is a significant difference from the empirical dataset (see Table 1). The column names r1–r10 indicate recombination rates ranging from 0.001 to 0.01, and a value of 0 in that column indicates that the data can be explained by a bottleneck. * and ** represent significant differences (5% and 1%, respectively) between the simulated and empirical data. The four candidate genes that were used for further simulations are highlighted with a light gray background.

combinations are compatible with the observations, whereas the remaining 50 cannot explain the significant difference. For other examples, **AMY1A** and **PGA3** both presented increased mean values in CHB. In neither case nor for either parameter combination is scenario I sufficient to explain the observation.

From the candidates with a significant difference in mean or variance, we selected well-studied genes with known functions and annotations and chose three genes coding for digestive enzymes, **AMY1A, SULT1A3, PGA3**, and the defense gene **DEFA1**, for a more detailed analysis and tested the GADMA

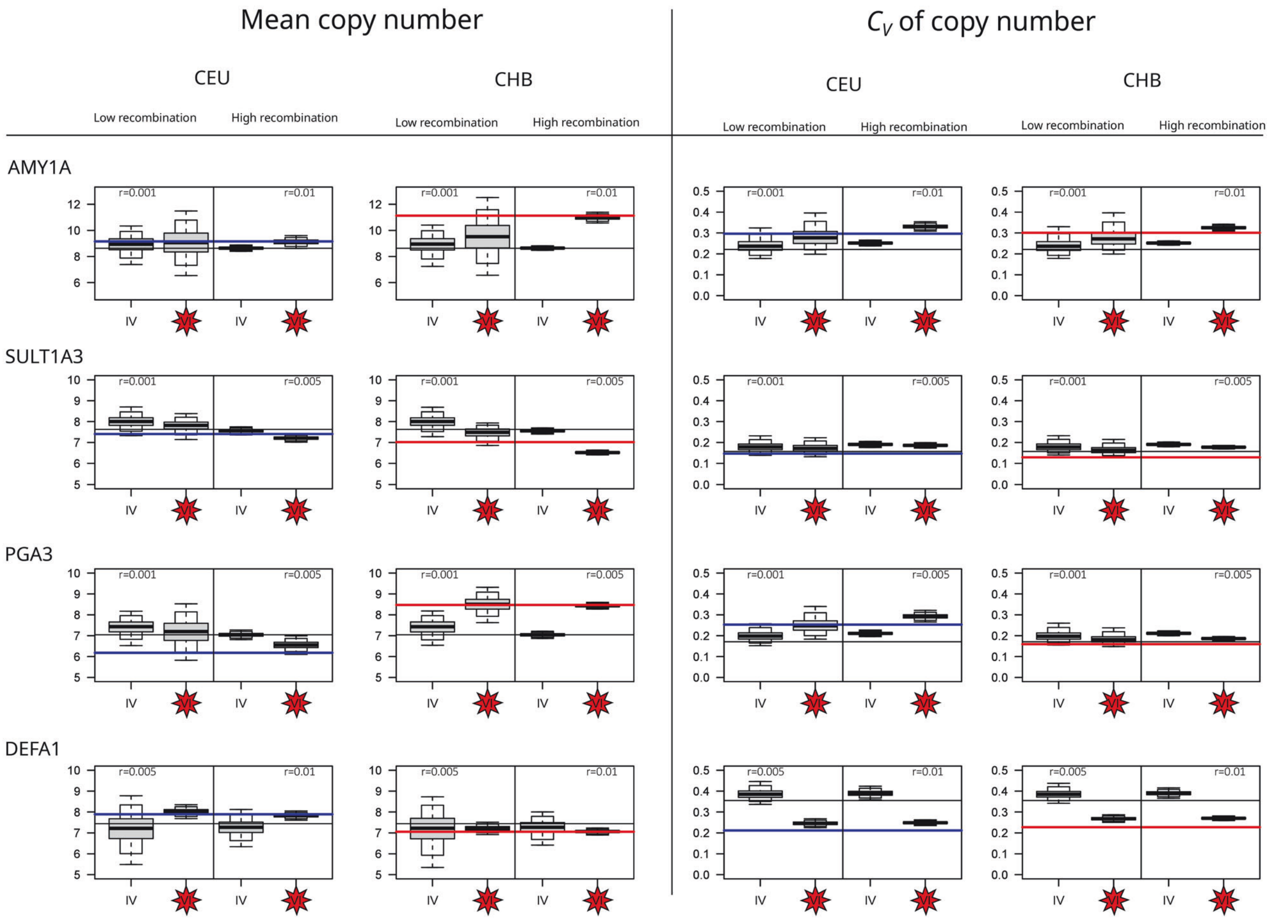

Fig. 5 Mean copy number y and coefficient of variation $C_V = \sigma/y$ for four candidate genes (**AMY1A, SULT1A3, PGA3**, and **DEFA1**). Boxplots: simulation results for the GADMA demographic model without (scenario IV) and with (scenario VI) a change in the selection and for two settings of the recombination rate (low and high; see Table 3). Horizontal lines: means and C_V of the experimental data in YRI (black), CEU (blue), and CHB (red).

model[20] without and with selection changes according to the estimates from regression (scenarios IV–VI).

Figure 5 shows the mean copy number and coefficient of variation (C_V) at present, which are simulated according to scenarios IV and VI for 10,000 replicates each.

As in scenarios I–III, the terminal values generated in scenario IV are close to the initial values of the ancestral YRI dataset (black line). Therefore, even the more realistic GADMA migration model often fails to explain the data found in CEU and CHB when constant selection parameters derived from the ancestral YRI population are considered.

However, when selection strength is allowed to change, as in scenario VI, a different picture emerges: Consider a change in s_x and s_y at 500 generations before being present with respect to the values estimated from Eq. (3), given in Table 3. Then, the simulations return the mean and C_V, which are closer to the values found in the CEU and CHB data. Indeed, the empirical data often lie within the 95%- or 99%-quantiles of the simulated data distributions. We observe no strong difference between the results of scenarios V and VI, suggesting that even 500 generations represents a sufficiently large time span to reach a new equilibrium.

Hence, one possible explanation for the shift in the copy number distribution of the four candidate genes is a change in selection pressure and adaptation.

The **AMY1A** gene, which encodes amylase, an enzyme that breaks down starch, has strongly increased mean and σ values in the Asian population, which is likely linked to adaptations to high grain intake. In the European population, while the variation increased, the change in the mean copy number was small.

These findings agree with the results of several studies that indicate that individuals from populations with high-starch diets have more gene copies on average than those with traditional low-starch diets[6,9,10]. Our model selection strength is relaxed in CEU and CHB with a factor of 4, such that a higher copy number is not selected, and a more widespread distribution of CNVs can evolve. A recent study[28] suggested a more complicated model of Amylase evolution involving two steps: expansion from one to several copies after the human–Neanderthal split but before the separation of modern human populations and a subsequent shift in the optimal gene copy number, independently in different populations. This study also suggested that the increase in **AMY1** copy number occurred in South America even more dramatically than in East Asia, a hypothesis that should be tested in the framework of our model as soon as suitable data become available.

SULT1A3 is a gene in the SULT (sulfotransferase) family that catalyzes the sulfation of a variety of substrates, especially catecholamines, including dopamine and epinephrine[29,30]. Polymorphisms in **SULT1A3** and **SULT1A4** have been shown to affect the metabolism of therapeutic drugs[31,32], and these genes have therefore been studied extensively in the framework of medico- and pharmacogenetics[33,34]. In the dataset analyzed, there was a reduced mean copy number in Asia but not in Europe. The reduced mean (from 7.6 in YRI to 7.0 in CHB) is a significant difference, which cannot be explained by a simple bottleneck scenario with a recombination rate higher than 0.002 (see Table 4). If one considers a change in selection, as in scenarios V and VI, we expect a stronger selective pressure (rising from $s_x = 0.03$ to $s_x = 0.05$ for $r = 0.002$) in CHB. There have been few studies on the copy numbers of **SULT1A3/4** genes. Hildebrandt et al.[33] 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 revealed variable **SULT1A3/4** copy numbers ranging from 1 to 10 and associated its copy number with the risk and onset of neurodegenerative disease[35]. Note that **SULT1A3** and **SULT1A4** are closely related paralogs that are often difficult to distinguish, and studies on copy numbers usually combine them.

PGA3 (Pepsinogen, a precursor for pepsin, an enzyme that breaks down protein into smaller peptides) is associated with prostate-specific antigen production. It is the only gene in our list that has 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. Indeed, the bottleneck simulations shown in Table 4 and simulations V and VI with a change in selection parameters, as shown in Fig. 5, support this hypothesis. When considering the estimates of Table 3, we observe a small increase in s_x in Asia compared with Africa but a strong decrease in both s_x and s_y in Europe to cope with the increased variance in copy number in CEU.

CNV in the pepsinogen (PGA) locus was originally discovered via electrophoresis, and three individual genes (**PGA** 3, 4, 5) were initially identified[36]. Pepsinogen genes have been shown to duplicate and become recurrently lost in vertebrates[37]. The pepsinogen genes were also shown to have variable expression levels in tumor cells, particularly a reduction in PGA expression in specific stomach and thyroid cancers[38]. This could be an additional source of selective pressure in addition to protein metabolism. While the simplest explanation is that dietary differences between Asian and European populations during the spread of agriculture (in the last 5000–10,000 years) are the drivers 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 the YRI population with high copy numbers, which indicates relaxed selective pressure in Africa. With Eq. (3), we find selection coefficients 10-fold smaller in Africa than in Europe and Asia (see Table 3). Alpha defensins are expressed in neutrophil cells and intestinal epithelial cells and act as microbiocidal agents[39–41]. The genes **DEFA1** and **DEFA3** encode some of the alpha-defensins (HNP1/2/3) and appear to be “interchangeable variant cassettes” within a tandem array of 19 kb[42]. CNV of **DEFA1** is present in all apes, including gibbons, but the version identified as **DEFA3** is human-specific; the copy number has also been demonstrated to affect the expression level[42]. A low copy number of **DEFA1/3** has been shown to be associated with hospital-acquired infection[43] as well as kidney diseases[44]. On the other hand, and counterintuitively, a high copy number of **DEFA1/3** may lead to more severe cases of sepsis[45,46] and is associated with Crohn’s disease[47]; thus, this gene was selected against. The trade-off between infective and autoimmune diseases could lead to a selection toward an intermediate copy number of alpha-defensins. Therefore, our results suggest that out-of-Africa expansion may be accompanied by a change in environmental pathogen diversity such that a delicately tuned dosage of defensin is needed. This can be corroborated by the fact that YRI has a few individuals with very high (outliers) copy numbers of **DEFA1**, which cannot 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 manifest not only in the mean copy number but also in the variance or compound statistics, such as the coefficient of variation.

DATA AVAILABILITY

Data and custom codes are available at https://github.com/y-zheng/gCNV-human.

REFERENCES

1. Brahmachary, M. et al. Digital genotyping of macrosatellites and multicopy genes reveals novel biological functions associated with copy number variation of large tandem repeats. *PLoS Genet.* **10**, e1004418 (2014).
2. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).
3. Iskow, R. C., Gokcumen, O. & Lee, C. Exploring the role of copy number variants in human adaptation. *Trends Genet.* **28**, 245–257 (2012).
4. Sebat, J. et al. Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
5. Sudmant P. H. et al. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).
6. Atkinson, F. S., Hancock, D., Petocz, P. & Brand-Miller, J. C. The physiologic and phenotypic significance of variation in human amylase gene copy number. *Am. J. Clin. Nutr.* **108**, 737–748 (2018).
7. Carpenter, D. et al. Obesity, starch digestion and amylase: association between copy number variants at human salivary (AMY1) and pancreatic (AMY2) amylase genes. *Hum. Mol. Genet.* **24**, 3472–3480 (2015).
8. Falchi, M. et al. Low copy number of the salivary amylase gene predisposes to obesity. *Nat. Genet.* **46**, 492–497 (2014).
9. Pajic P. et al. Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *eLife* **8**, e44628 (2019).
10. Perry, G. H. et al. Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2007).
11. Usher, C. L. et al. Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nat. Genet.* **47**, 921–925 (2015).
12. Lohmueller, K. E. The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet.* **10**, e1004379 (2014).
13. Stajich, J. E. Disentangling the effects of demography and selection in human history. *Mol. Biol. Evol.* **22**, 63–73 (2004).
14. Fu, Y. X. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925 (1997).
15. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
16. Otto M., Zheng Y. & Wiehe T. Recombination, selection, and the evolution of tandem gene arrays. *Genetics* **221**, giaa005 (2022).
17. Sudmant, P. H. et al. Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
18. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
19. Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of human sequence data reveals two pulses of archaic denisovan admixture. *Cell* **173**, 53–61.e9 (2018).
20. Noskova E., Ulyantsev V., Koepfli K. P. & O’Brien S. J., Dobrynin P. GADMA: genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data. *GigaScience*. **9**, giaa005 (2020).
21. Liao, W. W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
22. Fay, J. C. & Wu, C. I. Hitchhiking under positive darwinian selection. *Genetics* **155**, 1405–1413 (2000).
23. Lauterbur M. E., Munch K. & Enard D. Versatile detection of diverse selective sweeps with flex-sweep. *Mol. Biol. Evol.* **40**, msad139 (2023).
24. Fujimoto, A. et al. A scan for genetic determinants of human hair morphology: Edar is associated with asian hair thickness. *Hum. Mol. Genet.* **17**, 835–843 (2008).
25. Morales, J. et al. A joint ncbi and embl-ebi transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).

26. Otto, M. & Wiehe, T. The structured coalescent in the context of gene copy number variation. *Theor. Popul. Biol.* **154**, 67–78 (2023).
27. Harris, D. N. et al. Diverse African genomes reveal selection on ancient modern human introgressions in Neanderthals. *Curr. Biol.* **33**, 4905–4916.e5 (2023).
28. Inchley C. E. et al. Selective sweep on human amylase genes postdates the split with Neanderthals. *Sci. Rep.* **6**, 37198 (2016).
29. Brix, L. A., Barnett, A. C., Duggleby, R. G., Leggett, B. & McManus, M. E. Analysis of the substrate specificity of human sulfotransferases SULT1a1 and SULT1a3: site-directed mutagenesis and kinetic studies. *Biochemistry* **38**, 10474–10479 (1999).
30. Dajani, R. et al. Kinetic properties of human dopamine sulfotransferase (SULT1a3) expressed in prokaryotic and eukaryotic systems: comparison with the recombinant enzyme purified from Escherichia coli. *Protein Expr. Purif.* **16**, 11–18 (1999).
31. Bairam, A. F. et al. Impact of SULT1a3/SULT1a4 genetic polymorphisms on the sulfation of phenylephrine and salbutamol by human SULT1a3 allozymes. *Pharmacogenet. Genomics* **29**, 99–105 (2019).
32. Hui, Y. & Liu, M. C. Sulfation of ritodrine by the human cytosolic sulfotransferases (SULTs): effects of SULT1a3 genetic polymorphism. *Eur. J. Pharmacol.* **761**, 125–129 (2015).
33. Hildebrandt, M. A. et al. Human SULT1a3 pharmacogenetics: gene duplication and functional genomic studies. *Biochem. Biophys. Res. Commun.* **321**, 870–878 (2004).
34. Thomae, B. A. et al. Human catecholamine sulfotransferase (SULT1a3) pharmacogenetics: functional genetic polymorphism. *J. Neurochemistry* **87**, 809–819 (2004).
35. Butcher, N. J. et al. Sulfotransferase 1a3/4 copy number variation is associated with neurodegenerative disease. *Pharmacogenomics J.* **18**, 209–214 (2017).
36. Taggart, R. T., Mohandas, T. K., Shows, T. B. & Bell, G. I. Variable numbers of pepsinogen genes are located in the centromeric region of human chromosome 11 and determine the high-frequency electrophoretic polymorphism. *Proc. Natl Acad. Sci.* **82**, 6240–6244 (1985).
37. Castro, L. F. C. et al. Recurrent gene loss correlates with the evolution of stomach phenotypes in gnathostome history. *Proc. R. Soc. B: Biol. Sci.* **281**, 20132669 (2014).
38. Shen, S., Li, H., Liu, J., Sun, L. & Yuan, Y. The panoramic picture of pepsinogen gene family with pan-cancer. *Cancer Med.* **9**, 9064–9080 (2020).
39. Ayabe, T. et al. Secretion of microbicidal α-defensins by intestinal Paneth cells in response to bacteria. *Nat. Immunol.* **1**, 113–118 (2000).
40. Ganz, T. et al. Defensins. natural peptide antibiotics of human neutrophils. *J. Clin. Investig.* **76**, 1427–1435 (1985).
41. Nassar, H. et al. α-defensin: link between inflammation and atherosclerosis. *Atherosclerosis* **194**, 452–457 (2007).
42. Aldred, P. M., Hollox, E. J. & Armour, J. A. Copy number polymorphism and expression level variation of the human α-defensin genes DEFA1 and DEFA3. *Hum. Mol. Genet.* **14**, 2045–2052 (2005).
43. Zhao, J. et al. Low-copy number polymorphism in DEFA1/DEFA3 is associated with susceptibility to hospital-acquired infections in critically ill patients. *Mediators Inflamm.* **2018**, 1–8 (2018).
44. Ai Z. et al. 2016. Low α-defensin gene copy number increases the risk for IgA nephropathy and renal dysfunction. *Sci. Transl. Med.* **8**, 345ra88 (2016).
45. Chen, Q. et al. Increased genomic copy number of DEFA1/DEFA3 is associated with susceptibility to severe sepsis in chinese han population. *Anesthesiology* **112**, 1428–1434 (2010).
46. Chen, Q. et al. Increased gene copy number of DEFA1/DEFA3 worsens sepsis by inducing endothelial pyroptosis. *Proc. Natl. Acad. Sci.* **116**, 3161–3170 (2019).
47. Jespersgaard, C. et al. Alpha-defensin DEFA1a3 gene copy number elevation in danish crohn's disease patients. *Dig. Dis. Sci.* **56**, 3517–3524 (2011).

FUNDING

This work has been funded by grants from the German Research Foundation (DFG SFB-1211, subproject B6) to T.W. Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41439-024-00293-w.

Correspondence and requests for materials should be addressed to Thomas Wiehe.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.