bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Purging of deleterious variants in Italian founder populations with extended autozygosity

Massimiliano Cocca[1], Marc Pybus[2], Pier Francesco Palamara[3], Erik Garrison[4], Michela Traglia[6], Cinzia F Sala[6], Sheila Uiivi[5], Yasin Memari[4], Anja Kolb-Kokocinski[4], Richard Durbin[4], Paolo Gasparini[1,5], Daniela Toniolo[6], Nicole Soranzo [4,7], Vincenza Colonna[8]*

1. Department of Medical, Surgical and Health Sciences, University of Trieste, 34100 Trieste, Italy
2. Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Barcelona, Spain, 08003
3. Harvard T. H. Chan School of Public Health, 02115 Boston, MA.
4. Human Genetics, Wellcome Trust Sanger Institute, Genome Campus, Hinxton, CB10 1HH
5. Medical Genetics, Institute for Maternal and Child Health IRCCS "Burlo Garofolo", 34100 Trieste, Italy
6. Division of Genetics and Cell Biology, San Raffaele Scientific Institute, Milano,Italy
7. Department of Haematology, University of Cambridge, Hills Rd, Cambridge CB2 0AH
8. National Research Council, Institute of Genetics and Biophysics, Naples, Italy

*Correspondence to:

Vincenza Colonna
Consiglio Nazionale delle Ricerche
Istituto di Genetica e Biofisica "Adriano Buzzati-Traverso"
Via Pietro Castellino 111
80131 Napoli - IT

tel +39 081 6132 254
fax +39 081 6132 706
email vincenza.colonna@igb.cnr.it

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Purging through inbreeding defines the process through which deleterious alleles can be removed from populations by natural selection when exposed in homozygosis through the occurrence of consanguineous marriage. In this study we carried out low-read depth (4-10x) whole-genome sequencing in 568 individuals from three Italian founder populations, and compared it to data from other Italian and European populations from the 1000 Genomes Project. We show depletion of homozygous genotypes at potentially detrimental sites in the founder populations compared to outbred populations and observe patterns consistent with consanguinity driving the accelerated purging of highly deleterious mutations.

Introduction

Population genetics theory and empirical evidence from model organisms predict that deleterious variants with dominant effect are quickly purged by selection, at a rate proportional to the effective population size and their fitness effect [1,2,3,4,5,6,7] . In a strictly recessive model, deleterious variants carried in heterozygous state are expected to conform to neutral modes of evolution. Therefore, for equal effect sizes recessive models allow for higher frequencies of deleterious variants compared to dominant ones. In other words, recessive variants are less easily purged by selection and remain in the populations at low frequency (mostly in heterozygosis). These recessive variants provide a significant contribution to the genetic load of the population as there is less opportunity for a detrimental phenotype caused by recessive mutations to manifest, compared to dominant phenotypes.

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Empirical and theoretical evidence from recent studies suggest that deleterious variants, which are rare in the general population, are enriched in human populations that experienced a founder effect[7,8,9,10]. However a debate exists about whether this is due to reduced efficacy of natural selection or other causes [11,12] and little is known about how consanguinity affects population genetic load. Consanguineous marriages result in substantial deviation from Hardy-Weinberg equilibrium due to increased autozygosity, which can cause deleterious recessive variants to be present in a homozygous state. While autozygosity is associated with higher rate of disorders (congenital, late onset, infertility, miscarriage, infant mortality and morbidity) in consanguineous populations [13,14,15,16,17], it also provides an opprtunity for selection to act by purging exposed deleterious phenotypes, eventually reducing the fraction of homozygous deleterious genotypes [18]. Purging causes less fit individuals to be eliminated, ultimately increasing the population's fitness. Indeed, in some species inbreeding caused by a bottleneck has been shown to have beneficial effects, for instance promoting species invasiveness in ladybirds[19] or removing highly deleterious variants in mountain gorillas[20]. In humans, consanguineous couples experience higher fertility than non-consanguineous couples[15,21,22], but this has been associated with causes other than purging, such as social factors (e.g. younger maternal age at first birth), reproductive compensation[15,23], negative effects of outbreeding (break-up of co-adapted gene complexes, and the increase of maternal–foetal incompatibilities)[22]. Little is known about how inbreeding might play a role in the improved fecundity of consanguineous couples.

In this study we investigated whether homozygosis has had a role in the purging of deleterious mutations in three Italian founder populations with different demographic histories [23,24,25] by comparing them with with outbred Italian and

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

European populations. We show evidence for a depletion of homozygous genotypes at putatively detrimental sites in all the three populations. The presence of this effect in multiple founder populations with high rates of consanguinity suggests that inbreeding might have had a role in accelerating purging of deleterious mutations. We conclude that purging through inbreeding is possible in human populations, and this study demonstrates it for the first time.

Results

Whole genome sequence of isolates identified 21M variable sites, one seventh of which are novel

We carried out low-read depth whole-genome sequencing (average 6x, Table S1) in 568 uselected individuals from three populations from the Italian network of genetic isolates (INGI), namely Val Borbera (VBI), Carlantino (CAR) and Friuli Venezia Giulia (FVG). Of these, FVG is further subdivided in 4 different villages, namely Erto (FVG-E), Illegio (FVG-I), Resia (FVG-R) and Sauris (FVG-S) (Figure 1a and Table S2). After stringent quality control (Methods), we identified a total of 21,244,190 variable sites. We further compared genetic variability of the isolates with populations representative of the general Italian and European population, namely 98 Tuscans (TSI) and 85 Utah Residents with Northern and Western European ancestry (CEU) from the 1000 Genomes Project (TGP)[26,27] Phase 1 data set (Figure 1a). The isolate call sets were generated using the same variant calling and QC pipelines employed by the TGP project, so the two variant datasets were merged to generate a final set of 46,281,641 variable sites including sites variable both in isolates and TGP. To verify the validity of this approach, we estimated concordance of genotypes at 298k variable sites on chromosome 22 within the merged call set with those obtained by simultaneously calling isolates and the TGP. We found a median non-reference

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genotype discordance of 1.55% and 3.8% for SNPs and INDELs, respectively (Table S3), suggesting that merging the two call sets is a viable approach.

Variable sites discovered in the three isolates were classified as 'private' if variable in at least one isolates but not variable in both CEU and TSI, and 'shared' if variable in at least one isolate and CEU or TSI. In a set of 46 random individuals per populations private sites sites represent on average the 13% of the variability in the isolates (Figure 1b, Table S4). The high incidence of newly-discovered ('private') variants highlights the importance of carrying out variant discovery in different populations for comprehensively characterising human genetic variability. Novel private variants tend to be enriched for missense mutations in all isolates (Figure 1c, two-sample Kolmogorov-Smirnov test p-value <10^e-11, Table S5): in many individuals the missense/synonymous ratio is >1 at private sites, whereas at shared sites this ratio has very small variance and is close to one. This trend is expected in expanding populations[28] and has been observed in other founder populations[7].

The incidence of rare variants at private sites (Minor Allele Frequency, MAF<0.03, first two bins in Figure 1d and Figure S1) was at least three times higher compared compared to shared sites in all populations considered (odds ratio>3, p-values < 2.2×10^{-16}, Table S6). When comparing incidence of rare private variants between isolates and reference populations, we observed a trend of lower incidence of rare variants in FVG villages compared to CEU and TSI (Wilcoxon rank-sum test p-value < 2.2×10^{-16}) suggesting a loss of alleles in FVG due to downsampling during a recent bottleneck followed by isolation. In contrast, the rate of rare private variants in VBI and CAR is is significantly higher than in CEU and TSI (Wilcoxon rank-sum test p-value < 2.2×10^{-16}), suggesting a recent population expansion in these two populations.

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We further investigated genetic relationship among individuals of the three isolates, TSI and CEU using a random set of 281k variable chromosome 22 sites spanning all frequency ranges. A previous exploration of VBI and FVG was limited to low-resolution SNP array data (not informative for rare, private and novel genetic variation) 23,24, while the genetic structure of CAR has not been described before. In a principal component analysis, two FVG villages (FVG-R and FVG-S) appeared to be distinct from the other populations along the first three components. Further, VBI and CAR separated from CEU and TSI on the fourth component (Figure S2). In admixture analyses, FVG-R and FVG-S separated from other populations already at K=4 clusters, although the most likely number of clusters was 10 (Figure S3). These results confirm earlier results based on SNP array data from the entire FVG sample[24], and are thus not related to the sampling strategy used in this study. Because of this sub-structuring, when relevant FVG villages were considered separately for subsequent analyses. CAR, described here for the first time, clusters with VBI and other reference populations (TSI and CEU) along the first two components and with VBI along the 3rd and 4th component (Figures S2 and S3).

Long term isolation and diffused homozygosity in isolated populations

The historical isolation experienced by the INGI, FVG and CAR populations is expected to result in greater genetic homogeneity and consistent homozygosity, as the result of systematic inbreeding, compared to reference populations. We investigated this hypothesis using several approaches.

First, we calculated relatedness between pairs of individuals using two independent measures, namely pedigree-based and genomic-based kinship. Pedigree based kinship represents the expected fraction of the genome shared that is identical by descent (IBD) and was calculated from genealogical records dating

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

back to the 17th century. Genomic kinship evaluates the probability of sharing genome identical by descent and was calculated using genetic data. Pedigree- and genomic-based kinship values were highly correlated (Figure S4). In FVG villages the resolution of pedigree-based kinship was poor because genealogies are incomplete, therefore we concentrated on genomic based-kinship and referred to pedigree-based kinship to confirm first-degree relatedness of couples that were excluded from the analyses. Genomic-based kinship values tended to be significantly higher in isolates compared to reference populations (Figure 2 dark blue, FigureS5 and Table S7; two-sample Kolmogorov-Smirnov test p-values <0.0003, details in Table S8). The kinship distribution were highly skewed towards low values in all cases, however maximum values were in the range of 1st degree relationships for isolates, compared to <3rd degree in reference populations. When removing one individual per pair from pairs related to 1st degree according to pedigree-based kinship, we still observed genomic-kinship values in the range of 1st degree in isolates (Figure 2, light blue). Overall these results not only show that individuals from isolates tend to be more consanguineous than reference populations, but also suggest that consanguinity has systematically occurred in these populations in historical times, and ancestors of 2nd and 3rd degree relatives were in turn the result of consanguineous relationships. Because of this extended consanguinity, in subsequent analyses we used a set of 500 individuals where one random individual from pairs with 1st degree relationships has been retained (Table S2. Figure S4).

Having assessed diffused consanguinity in the isolates, we next quantified genome segments which were IBD among individuals within populations in 46 samples per population. We assessed 46 samples per population, where we identified IBD segments between pairs of individuals. In Figure 3a we report the distributions of total extension of IBD genome (i.e. the sum of single segments).

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

While there was no significant difference between isolates and CEU or TSI in the total IBD distributions (two-sample Kolmogorov-Smirnov test p-value >0.01, Table S8) FVG villages displayed higher total IBD sharing on average (Figure S6 and Table S9). These differences were more evident when contrasting cumulative distributions of total IBD (Figure 3b, figures in Tables S9): in FVG villages 95% of the individuals shared 337-461(+/-85-123)Mb compared to 122-157(+/-8-45)Mb in other isolates and reference populations. Considering an accessible genome size of 2.5Gb27, this would be equivalent to say that individuals in FVG villages share between 13% and 18% of their genomes, *versus* 5-6% in other populations, corroborating evidence for extended consanguinity in the isolates.

Finally, we investigated allele sharing within an individual by identifying genomic segments with contiguous genotypes in homozygosity (Runs Of Homozygosity, ROH) in each individual. We report in Figure 3c the distributions of the total length of ROH (i.e. the sum of single segments) per individual. With the exception of CAR, ROH extension was significantly higher in isolates compared to CEU and TSI (see two-sample Kolmogorov-Smirnov test p-values given in Table S9), suggesting increased homozygosity in isolates. This is also evident when comparing cumulative distributions of total ROH in 46 randomly selected individuals per population (Figure 3d, figures in Table S9): 95% of individuals in FVG villages and VBI had between 163-209(+/-30-43)Mb of genome in homozygosys compared to less than <100(+/-10-15)Mb in the outbred reference populations. These figures translate respectively in 5-8% of the genome in homozygosity in isolates (except CAR) compared to 3-4% in CEU, TSI.

When ROHs single segments were classified according to their length[29], we observed a trend of prevalence of medium and long ROHs in isolates (and especially

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FVG-R and FVG-S) compared to CEU and TSI (Figure S7). This condition is consistent with the presence of extended consanguinity or very recent admixture; both cases apply to the isolates under study here, for which isolation (either geographical or linguistic) was a past condition that is no longer met.

Selection against highly deleterious mutations has been more effective in isolates compared to reference populations

To test the hypothesis that autozygosity contributes to the purging of deleterious mutations through exposure of detrimental alleles to purifying selection, we determined the incidence of homozygous detrimental genotypes and measured the efficiency of purifying selection in isolates compared to reference populations.

For this analysis we used derived alleles information in 46 samples at three set of sites where the derived allele causes missense, synonymous, or loss of function (LoF)[30] consequences. Alleles determining missense, and in minor measure synonymous[31], consequences introduce changes to the coded protein that might alter the protein structure, whereas LoF mutation completely disrupts the protein function. Only shared sites were considered for this analysis both to avoid technical bias due to variant calling and because private sites are generally too young to observe the effect of selection. Rather than assessing the overall burden of isolates, to which private site will be major contributors, we are indeed interested in asking if purging due to inbreeding has occurred in isolates.

We observed significantly fewer derived alleles counts per individual in FVG and VBI compared to reference populations at all sets of variants (Figure 4a, Wilcoxon rank sum test p-values in Table S10), suggesting that this might be a general effect of the demography, independent from putative detrimentality of alleles and

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

compatible with the general loss of rare alleles due to bottlenecks and genetic drift. We than evaluated how often putatively detrimental alleles are found in homozygosis by counting the number of homozygous genotypes per individuals at missense, synonymous and LoF sites, normalized for the number of homozygous genotypes at intergenic sites to avoid population bias. Isolates had significantly higher rates of homozygous genotypes at mildly deleterious sites (missense and synonymous, Figure 4b) compared to reference populations, in accordance with the general homozygosity trend reported in the previous section. However, at highly deleterious sites (LoF) the rate of homozygosity was comparable between isolates and reference populations (with the exception of FVG-E and FVG-R) indicative of an overall depletion of homozygous genotypes at these sites in isolates. This depletion of deleterious alleles in homozygosis is compatible with the hypothesis of purging due to inbreeding in isolates.

We therefore used a recently proposed metric ($R_{X/Y}$)[8,32] to assess whether selection has been more effective in isolates compared to reference populations. The metric $R_{X/Y}$ contrasts frequencies of allele between pairs of populations and measures their accumulation in one population compared to the other: a value of 1 indicates that the two populations have comparable number of mutations per individual, whereas values lower than 1 indicate a depletion in population X and values higher than 1 indicate a depletion in population Y. We compared $R_{X/Y}$ at missense, synonymous and LoF variants in isolates and reference populations. The load of moderately deleterious (missense and non-synonymous) mutations in isolates was higher than in reference populations, consistent with less efficient selection and stronger drift in the former (Figure 5). However, the opposite was true of variation at LoF sites (with the exception of FVG-S), with a possible cause being

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the purging effect of inbreeding alongside strong selection coefficients for highly disruptive LoF mutations.

Simulations support the observed depletion of highly deleterious homozygous genotypes

We evaluated expectations to observe purging in simulated data under a range of demographic scenarios, natural selection intensities and models. A population isolate was simulated to arise from an European-like population at 160, 60 or 20 generations ago (generation time 25 years), and to evolve to present time under five demographic models (continued bottleneck, serial founder effect, bottleneck plus linear expansion, bottleneck plus instantaneous expansion and bottleneck plus linear reduction; Figure S8) and three different values of average kinship (0, 0.002 and 0.1). We forward-in-time simulated genomic data by mimicking features of exomes spanning 115 kbp (Methods). Mutations were assumed to be deleterious ¾ of the time in exons and ½ in UTRs under six selection coefficients (s = -0.2, -0.1, -0.05, -0.01, -0.005 and -0.001).Each scenario was replicated 500 times. We found purging to be more effective in the bottleneck plus expansion (Figure S9) and thus for this model we further explored a range bottleneck intensities (effective population size, Ne, during the bottleneck = 15, 25 and 50 and bottleneck duration 1 to 25 generations in steps of 5, Table S11). Demographic parameter choice was a compromise between computational feasibility and similarity to the demography of the isolates under study, its main purpose being to explore properties of purging rather than to perfectly fit isolate demography.

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Purging was calculated in a sample of 50 individuals as the fold change between present and past (at founding event) number of recessive deleterious homozygous genotypes per individual. A negative fold change described a reduction of deleterious homozygous genotypes expected in the case of purging. Results in Figure 6 (also see Figure S10) show that: (i) under neutrality (s = 0) no purging is seen in all three models; (ii) for mildly deleterious mutations (s = 0.01) purging is almost complete under dominance, moderate for additive models but not happening under recessive models and (iii) for highly deleterious mutations (s = 0.1) purging of homozygous recessive deleterious genotypes is seen also under recessive models.

Overall, these simulations suggest that the selection coefficient and dominance model have a major effect on purging and that purging is possible for highly deleterious recessive mutations. While it is difficult to make assumptions on the inheritance model of missense and synonymous variants, a recessive model and a strong selection coefficient are highly plausible for LoF[30]. We therefore conclude that simulations support the observed depletion of highly deleterious homozygous genotypes observed at LoF sites. Nevertheless, our set of simulations does not explain the role of consanguinity as the effect of kinship is negligible.(Figure 6). This may be due to the imperfect fit of the demographic parameters used for simulations to the isolates studied here. A more comprehensive analysis will be required to fully understand this aspect.

Some isolates experienced genetic drift and a recent decrease of the effective population size

A decrease in the number of deleterious mutation could more generally be the result of loss of rare variants following a bottleneck, a founder event, or isolation. We thus investigated to what extent genetic drift acted on the populations in this study.

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We used a method based on inference from genomic segments IBD that enables exploring very recent fluctuations in effective population size[33] . Low coverage and computational phasing are likely to affect the accuracy of IBD detection methods, occasionally breaking down long IBD segments into shorter ones and thus potentially creating the artefactual effect of a population expansion. Despite this potential confounding, our analysis displayed clear patterns of overall lower effective population size estimates in isolates compared to reference populations (Figure 7). Indeed in the last fifty generations (last ~1500 years in Figure 7) we observed a population size contraction in FVG villages compared to CEU and TSI, which appear unaffected by the very recent trends towards expansion of other European populations [28,34]. Such abundant sharing of long haplotypes co-inherited from recent common ancestors during a recent population contraction provides further evidence for genetic drift that extended until the very recent centuries in these isolated groups compared to the reference populations. The same trend of contraction is not apparent in CAR and VBI, that follow TSI and CEU, however with smaller effective population sizes. We therefore conclude that genetic drift might have contributed to reducing deleterious mutations in FVG villages, while this has not been the case in VBI and CAR.

Discussion

Inbreeding consists of mating between closely related individuals. Chromosome pairs of inbred offspring share segments that are identical by descent more often than expected by chance, and consequently large portions of the genome in inbred individuals are in homozygosis. Inbred offspring undergo inbreeding depression, the phenomenon by which detrimental mutations increase in frequency in a population, with consequent reduction in overall fitness. However, in the long term the rise of detrimental mutations frequency is offset by an increased efficiency of selection in

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

removing individuals carrying these mutations. Therefore, in the long term inbreeding can have beneficial effects.

In this study we demonstrate that consanguinity results in purging of deleterious mutations in isolated human populations. Analysis of whole genome data from three isolated populations demonstrated the occurrence of (i) diffused consanguinity and extended regions of autozygosity; (ii) small effective population sizes until recent times; (iii) depletion of homozygous genotypes of putatively deleterious mutations and (iv) more effective selection against detrimental mutations in comparison to outbred Italian and European populations.

This last observation is in contrast with theoretical prediction, as selection is less efficient in populations with small effective population size. We thus investigated plausible explanations underlying the observed purging. Our simulations shows that a reduction of highly deleterious mutations in homozygosis is possible under a range of demographic and evolutionary conditions even if they do not fully clarify the role of consanguinity. Genetic drift as a consequence of founder effects and reduced effective population size is another possible cause. Bottlenecks cause a reduction of segregating sites and a loss of rare variants, including deleterious mutations, as they tend to be rare. According to our data, genetic drift has certainly played a major role in FVG as the effective population size of villages is very small, but this does not hold for other isolates and therefore the patterns observed cannot be explained by selection and drift alone. All isolates have extended regions of the genome in homozygosity as a result of isolation and kin marriages, and we cannot exclude that consanguinity is likely to have played a major role in accelerating the process of selection in FVG and VBI through the exposure of the deleterious mutations in homozygosis.

Our observations are compatible with the effect of purging of deleterious mutations carried in homozygosis and our conclusions are supported by the observation reported in literature that in

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cases of systematic inbreeding due to non-random mating (as is the case of small consanguineous populations), purging can work even for mild deleterious alleles as opposed to cases of panmictic inbreeding (i.e. inbreeding due to finite size of the populations) where purging is effective only for lethal or semilethal alleles [35] .

Consanguineous populations represent a model for the effects of systematic inbreeding. Beside social constrains, consanguinity can naturally occur for several reasons, including geographical, linguistic and religious isolation. Recently there has been a great interest in exploring the genetic load in populations with different demographic history. Consanguineous populations in this study complement the range of demographic scenarios and here we demonstrate that they provide an excellent model study to understand the interplay between selection, genetic drift and homozygosis in determining the genetic load of human populations. .

Methods

Samples individuals and data generation

Low coverage whole genome sequence was generated in for 568 individuals belonging to the INGI network. Individuals are from Italian populations form North-West (Val Borbera[25], 225 samples), North-East (Friuli Venezia Giulia[24], 250 samples) and South-West (Carlantino[24], 93 samples). Sequencing was carried out using Illumina technology (Genome Analyzer and HiSeq 2000) at the Wellcome Trust Sanger Institute and Beijing Genomics Institute (54 samples from FVG cohort). Data coverage was 4-10X (Table S2).

The study was reviewed and approved by the following Ethical Committees: Ethical committee of the San Raffaele Hospital and of the Piemonte Region (VBI),

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ethics approval was obtained from the Ethics Committee of the Burlo Garofolo children hospital in Trieste (FVG), the local administration of Carlantino, the Health Service of Foggia Province, Italy, and ethical committee of the IRCCS Burlo-Garofolo of Trieste (CAR). Written informed consent was obtained from every participant to the study and all methods were carried out in accordance with the approved guidelines (Declaration of Helsinki of the World Medical Association).

Data processing, variant calling

Genotype calls were produced for each population separately using the following pipeline. Samtools mpileup (v. 0.1.19)[36] was used for multisample genotype calling (parameter set: -EDVSp -C50 -m3 -F0.2 -d 40). Variant Quality Score Recalibrator (VQSR) filtering was applied to the raw call data with GATK v.2.5[37] through VariantRecalibrator module separately for SNVs and INDELs. The filter creates a Gaussian Mixture Model by looking at annotations values over a high quality subset of the input call set and then evaluate all input variants. For SNVs we used the following parameters: I) Annotations: QD, DP, FS, HaplotypeScore, MQRankSum, ReadPosRankSum, InbreedingCoeff II) Training set: HapMap 3.3, Omni 2.5M chip, 1000 Genomes Phase I; III) Truth set: HapMap 3.3 ,Omni 2.5M chip; IV) Known set: dbSNP build 138. For INDELs we selected: I) Annotations: DP,FS,ReadPosRankSum,MQRankSum; II) Training set: Mills-Devine, 1000 Genomes Phase I, DbSnp v138; III) Truth set: Mills-Devine; IV) Known set: Mills-Devine, dbSNP build 138. For each population the lowest VQSLOD threshold has been chosen according to the output produced by VariantRecalibrator to select the best cut off in terms of specificity and sensitivity of the trained model. For SNPs the minimum VQSLOD values selected are 3.016 (97.3% truth sensitivity threshold), 2.8309 (97.82% truth sensitivity threshold), 1.4512 (98.5% truth sensitivity threshold) for

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FVG, VBI and CARL cohort respectively. Since INDELs calling and alignment is still more prone to error we used a conservative approach, selecting a sensitivity threshold of 90% for each population. The filter has been applied using GATK's Apply Recalibration module.

Because of low coverage we performed several genotype refinement steps on the filtered data: 1)we used BEAGLEv4.r1230[38] to assign posterior probabilities to all remaining genotypes. 2) SHAPEITv2[39] has been used to phase all genotypes call and 3) IMPUTEv2[38] to perform internal imputation in order to correct genotyping errors. Information about Ancestral Allele and allele frequencies from TGP populations have been retrieved from dbSNP v.138[40]. The Variant Effect Predictor v.74[41] provided all consequence annotation as well as Polyphen,Sift and GERP informations.

We merged genotype call set of the INGI populations with low coverage data available from Phase1 of the 1000 Genomes Project[27] in a union using set using bcftools merge[36]. To check if this approach would introduce bias in downstream analyses, we evaluated genotype concordance on chromosome 22 between the union set and a genotype call set in which variants were called simultaneously in all INGI, CEU and TSI populations using the same pipeline described above.

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Population genetic structure

Shared ancestry between populations was evaluated using ADMIXTURE v 1.22 and the number of cluster that better represent the data was established by cross-validation as described in[42]. Each ADMIXTURE run was replicated 5 time using different random seeds. Principal Component Analysis (PCA) analysis was performed using EIGENSOFT [43].

Kinship, identical by descent state and runs of homozygosity calculations

Pairwise genomic kinship was calculated using KING[44] (options: --kinship --ibs), while pedigree-based kinship was calculated using the R package Kinship2[45]. Pedigree-based kinship represent mostly an expected value indicative of the amount of genome shared identical by descent, however recombinations and mutation introduce variability around the expected values. Genomic-based kinship catches this variability providing estimates in a range rather than pedigree-based point estimates (e.g first-degree kinship vary between 0.25+/-variability). A negative genomic kinship coefficient indicates an unrelated relationship and for practical purpose it is safe to rescale negative kinship values to zero [44]. All analyses except kinship distributions were performed on individuals related at most to 2nd degree according to pedigree-based kinship.

Segments of identity by descent (IBD) and runs of homozigosity (ROH) were identified using the refined IBD algorithm implemented in Beagle v4.r1274[46] (parameters used: ibd=true, window=24000, overlap=7000). All variants with uncalled genotypes were removed, leaving 7,502,857 sites for further analyses. After discovery we retained IDB and ROH segments with LOD score >5 as threshold to define a true IBD/ROH segment (default value LOD >= 3). When comparing IBD

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sharing, ROH data and allele frequencies populations were randomly down-sampled to match the size of the smallest village size (FVG-I, 46 unrelated samples) in order to reduce bias from different samples size.

RX/Y

The metric $R_{X/Y}$ was calculated implementing the formula described in [32,8] and [20] using the information on derived alleles counts per individual in 46 samples per population. Variance of $R_{X/Y}$ was obtained by block jackknife procedure as described in [47]. To avoid biases due to lack of calling and imputation we considered only sites called as variable in both X and Y populations (shared sites). Information on the ancestral state of the allele were taken from 1000 Genomes annotations (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/). We used four sets of variants with information on functional consequences from Ensembl database v. 78 (www.ensembl.org): (i) 136,805 variants annotated as intergenic, i.e. “in region containing or overlapping no genes that is bounded on either side by a gene, or bounded by a gene and the end of the chromosome” according to Sequence Ontology definition (S.O., http://www.sequenceontology.org/); (ii) 25,773 missense variants, i.e. “A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved” (S.O.); (iii) 23,242 synonymous, i.e. “A sequence variant where there is no resulting change to the encoded amino acid” (S.O.), and (iv) 2,804 loss of function variants described in[30]

Simulations

We used SLiM[48] to perform forward-in-time simulation of purging of deleterious recessive alleles in a small size human isolated population. SLiM and allow

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

simulations of specific genomic structures with different selective pressures acting on them. We simulated a genomic region of 115 kbp containing fifty gene-like structures representative of the average human gene[49] each composed of 8 exons of 150 bp surrounded by 2 UTR regions of 550 bp. Recombination rate was set to 1.6e-8 recombinations/bp/generation, mutation rate to 1.2e-8 mutations/bp/generation. Generation time was set to 25 years. Mutations were assumed to be deleterious ¾ of the time in exons and ½ in UTRs. Six deleterious selection coefficients were tested independently (s = -0.2, -0.1, -0.05, -0.01, -0.005 and -0.001) being the strongest one the upper limit estimated for human species. Deleterious mutations were assumed to be fully recessive additive or dominant.

As for the demographic model we simulated an isolated population of European ancestry undergoing a founder event of different intensity in terms of number of individuals and length of the bottleneck (see simulation scheme in Figure S7) and then evolving under an specific demography. Three different split times of the isolate form Europe (160, 60 and 20 generations ago), and five demographies were tested: continued bottleneck, serial founder effect, bottleneck plus linear expansion, bottleneck plus instantaneous expansion and bottleneck plus linear reduction). To mimic isolation we assumed no migration between the isolate and the European population after the split. The European population was simulated under the model described in[50] with a minor modifications: because SLiM does not simulate exponential growth we extrapolated the trajectory of the exponential growth found in Out-of-Africa populations from[50] and increased population effective size every few generations. Finally, in our model the European population itself separate from Africa as described in[50]. To mimic relatedeness we let inbreeding starts one generation after the split between Europe and the isolate using two inbreeding coefficients: one similar to estimates from empirical data (averages among all isolates, f = 0.002,

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S7) and a second one five times bigger (f = 0.01). To use these inbreeding coefficient in the forward simulator (SLiM), we converted them into selfing coefficient using this formula described in (http://darwin.eeb.uconn.edu/eeb348/lecturenotes/inbreeding.pdf).

Each set of parameters and demographies was simulated 500 times and, additionally, we considered groups of 10 replicates together as they were independent chromosomes in order to reduce the variance in the individual recessive load calculation. From every simulated replica we sampled 50 individuals in each of the three populations. In simulations with a final population effective size of less than 50 individuals the whole population was sampled. To speed up computation we used two strategies. First, we removed from simulations the neutrally evolving intergenic and intronic regions, thus mimicking exomic data. Secondly, to avoid repeating the burn-in process, we simulated 500 replicates for Africans and Europeans and saved all the individual genotypes at 200 generations after the beginning of the simulation (5000 years ago) and we used these population snapshots as genetic pools for the founding nucleus of the isolate while testing different demographic models. We believe that this strategy ensures enough variation for testing properly the different population sizes for the ISO population without having to use a new simulations run from the scratch.

Effective population size

Recent effective population size for each population was estimated using the fraction of genome shared IBD by pairs of individuals[33]. After computational phasing, IBD sharing was computed as described in[51]. We initially removed all variants with minor allele frequency less than 1%. We used a publicly available genetic map (b37, Web Resources), using linear interpolation to infer the genetic position of variants

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

that were not found in the map. We then used the GERMLINE[52] software to infer IBD sharing across all pairs of samples, using parameters "-min_m 1 -err_hom 2 -err_het 0 -bits 75 –h_extend -homoz". We computed the density of IBD sharing per genomic region using windows of 0.5 centimorgans (cM), and selected for downstream analysis regions within 5 standard deviations from the genome-wide average, removing regions of unusual sharing, likely due to artifacts or the presence of underlying structural variation. We further imposed a minimum length of 45cM per genomic region to avoid introducing biases due to boundary effects in very short regions. We obtained 27 regions, for a total of ~2050 cM. We computed the fraction of genome shared by the average pair of haploid individuals in each analysed group through segments of length at least 6 cM. We then used the formula: N=(1-f+sqrt(1-f))/2uf derived in[33], where *f* is the observed fraction of genome shared IBD and *u* is the minimum segment length (in Morgans) to infer effective population size (assuming a constant size). We report diploid effective sizes. Standard errors were computed using the weighted jackknife method[52], using the 27 genomic regions. To infer an approximate correspondence between IBD segment cut-off and time, we considered the distribution of IBD segments age in a constant population size. Assuming a generation time of 30 years[53] a cutoff of 6 cM roughly corresponds to IBD segments transmitted by common ancestors living 30*(100/6)=500 years before present.

References

1. Kimura, M. *Population Genetics, Molecular Evolution, and the Neutral Theory*. (University of Chicago Press, 1995). at <http://www.press.uchicago.edu/ucp/books/book/chicago/P/bo3645416.html>

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2. Robertson, A. The Effect of Inbreeding on the Variation Due to Recessive Genes. *Genetics* **37,** 189–207 (1952).

3. Wang, J., Caballero, A., Keightley, P. D. & Hill, W. G. Bottleneck effect on genetic variance. A theoretical investigation of the role of dominance. *Genetics* **150,** 435–447 (1998).

4. Wang, J., Hill, W. G., Charlesworth, D. & Charlesworth, B. Dynamics of inbreeding depression due to deleterious mutations in small populations: mutation parameters and inbreeding rate. *Genet. Res.* **74,** 165–178 (1999).

5. Whitlock, M. C. Selection, load and inbreeding depression in a large metapopulation. *Genetics* **160,** 1191–1202 (2002).

6. García-Dorado, A. A simple method to account for natural selection when predicting inbreeding depression. *Genetics* **180,** 1559–1566 (2008).

7. Casals, F. *et al.* Whole-Exome Sequencing Reveals a Rapid Change in the Frequency of Rare Functional Variants in a Founding Population of Humans. *PLoS Genet* **9,** e1003815 (2013).

8. Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat. Genet.* **47,** 126–131 (2015).

9. Lohmueller, K. E. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451,** 994–997 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

10. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat. Genet.* **46,** 220–224 (2014).

11. Gravel, S. When is selection effective? *bioRxiv* 010934 (2014). doi:10.1101/010934

12. Lohmueller, K. E. The distribution of deleterious genetic variation in human populations. *Curr. Opin. Genet. Dev.* **29,** 139–146 (2014).

13. Bittles, A. H. & Neel, J. V. The costs of human inbreeding and their implications for variations at the DNA level. *Nat. Genet.* **8,** 117–121 (1994).

14. Jorde, L. B. Consanguinity and prereproductive mortality in the Utah Mormon population. *Hum. Hered.* **52,** 61–65 (2001).

15. Ober, C., Hyslop, T. & Hauck, W. W. Inbreeding effects on fertility in humans: evidence for reproductive compensation. *Am. J. Hum. Genet.* **64,** 225–231 (1999).

16. Postma, E., Martini, L. & Martini, P. Inbred women in a small and isolated Swiss village have fewer children. *J. Evol. Biol.* **23,** 1468–1474 (2010).

17. Rudan, I. *et al.* Inbreeding and risk of late onset complex disease. *J. Med. Genet.* **40,** 925–932 (2003).

18. Kirkpatrick, null & Jarne, null. The Effects of a Bottleneck on Inbreeding Depression and the Genetic Load. *Am. Nat.* **155,** 154–167 (2000).

19. Facon, B. *et al.* Inbreeding depression is purged in the invasive insect Harmonia axyridis. *Curr. Biol. CB* **21,** 424–427 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

20. Xue, Y. *et al.* Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science* **348,** 242–245 (2015).

21. Bittles, A. H., Mason, W. M., Greene, J. & Rao, N. A. Reproductive behavior and health in consanguineous marriages. *Science* **252,** 789–794 (1991).

22. Helgason, A., Pálsson, S., Gudbjartsson, D. F., Kristjánsson, T. & Stefánsson, K. An association between the kinship and fertility of human couples. *Science* **319,** 813–816 (2008).

23. Colonna, V. *et al.* Small effective population size and genetic homogeneity in the Val Borbera isolate. *Eur. J. Hum. Genet.* **21,** 89–94 (2013).

24. Esko, T. *et al.* Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *Eur. J. Hum. Genet.* **21,** 659–665 (2013).

25. Traglia, M. *et al.* Heritability and Demographic Analyses in the Large Isolated Population of Val Borbera Suggest Advantages in Mapping Complex Traits Genes. *PLoS ONE* **4,** e7554 (2009).

26. Consortium, T. 1000 G. P. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

27. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

28. Keinan, A. & Clark, A. G. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science* **336,** 740–743 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

29. Pemberton, T. J. *et al.* Genomic patterns of homozygosity in worldwide human populations. *Am. J. Hum. Genet.* **91,** 275–292 (2012).

30. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335,** 823–828 (2012).

31. Lawrie, D. S., Messer, P. W., Hershberg, R. & Petrov, D. A. Strong purifying selection at synonymous sites in D. melanogaster. *PLoS Genet.* **9,** e1003527 (2013).

32. Balick, D. J., Do, R., Reich, D. & Sunyaev, S. R. Response to a population bottleneck can be used to infer recessive selection. *bioRxiv* 003491 (2014). doi:10.1101/003491

33. Palamara, P. F., Lencz, T., Darvasi, A. & Pe'er, I. Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History. *Am. J. Hum. Genet.* **91,** 809–822 (2012).

34. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* **1,** 131 (2010).

35. Glémin, S. How are deleterious mutations purged? Drift versus nonrandom mating. *Evol. Int. J. Org. Evol.* **57,** 2678–2687 (2003).

36. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25,** 2078–2079 (2009).

37. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

38. Browning, S. R. & Browning, B. L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am. J. Hum. Genet.* **81,** 1084–1097 (2007).

39. Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F. & Marchini, J. Haplotype Estimation Using Sequencing Reads. *Am. J. Hum. Genet.* **93,** 687–696 (2013).

40. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).

41. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinforma. Oxf. Engl.* **26,** 2069–2070 (2010).

42. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).

43. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).

44. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).

45. Sinnwell, J. P., Therneau, T. M. & Schaid, D. J. The kinship2 R package for pedigree data. *Hum. Hered.* **78,** 91–93 (2014).

46. Browning, B. L. & Browning, S. R. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics* **194,** 459–471 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

47. Busing, F. M. T. A., Meijer, E. & Leeden, R. V. D. Delete-m Jackknife for Unequal m. *Stat. Comput.* **9,** 3–8 (1999).

48. Messer, P. W. SLiM: simulating evolution with selection and linkage. *Genetics* **194,** 1037–1039 (2013).

49. Sakharkar, M. K., Chow, V. T. K. & Kangueane, P. Distributions of exons and introns in the human genome. *In Silico Biol.* **4,** 387–393 (2004).

50. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 11983–11988 (2011).

51. The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** 818–825 (2014).

52. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19,** 318–326 (2009).

53. Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128,** 415–423 (2005).

Acknowledgements

We acknowledge the contribution of people from villages. This study was supported by the European Commission (FP7/2007-2013, under grant agreement number no. 262055 (ESGI), as a Transnational Access project of the European Sequencing and Genotyping Infrastructure"; Itailan Ministero della Salute - Ricerca Finalizzata PE-

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2011-02347500 to PG and DT; Italian Ministero della Salute – Public Health Genomics 2010 and Fondazione Cariplo to DT; the Wellcome Trust (Grant Codes WT098051 and WT091310), the NIHR and the EU FP7 (EPIGENESYS Grant Code 257082 and BLUEPRINT Grant Code HEALTH-F5-2011-282510) to NS.

Author contributions

MC, MP, PFP, EG and VC designed and performed analyses. MT, CFS, SU performed genotype quality controls. YM, RD generated and analysed exome data. AKK coordinated exome sequencing/data generation. PG, DT discussed the analyses' results. NS and VC wrote the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Figure captions

Figure 1 Overview of genomic variation in isolates **(a)** Geographic location of the populations in this study. Circles are proportional to sample sizes. VBI=Val Borbera, CAR= Carlantino, FVG = Friuli Venezia Giulia, FVG-E = Erto, FVG-I = Illegio, FVG-R= Resia and FVG-S = Sauris **(b)** Genomic sites variable in isolates were classified as shared if variable in both in CEU and/or TSI and isolates and private if variable only in isolates. When considering 46 samples per population about 13% of variants are private to isolates (yellow) **(c)** Boxplots summarize the ratio of missense/synonymous counts per individual. Private variants have larger variance and and are significantly enrichment for missense mutations compared to shared

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ones. **(d)** Genome wide spectrum of minor allele allele frequencies (MAF) from 46 samples per population. For simplicity only MAF< 23 is shown, full spectra in Figure S1. First two bins include rare variants (MAF< 0.3%). Private sites are significantly enriched for rare variants compared to shared sites (Table S6).

Figure 2 Systematic consanguinity in the isolated populations. Violin plots represent genomic-based kinship distribution before (dark blue) and after (light blue) removing 1st degree relatives according to pedigree. Grey dotted lines represent reference values for 1st-, 2nd- and 3rd-degree relationships. People in isolates tend to be more consanguineous, i.e. they have more recent common ancestors than reference populations. In FVG-R and VBI relatedness to 1st degree persists even after removing 1st degree relatives according to pedigree data. The explanation is that ancestors of some 2nd and 3rd degree relatives were in turn the results of consanguineous relationship, showing that consanguinity in these populations has occurred on a long time scale

Figure 3 Genome sharing between individuals and within individual. Summary statistics **(a and b)** and cumulative frequency **(c and d)** of the total length of segments identical by descent per pairs of individuals and total length of runs of homozygosity per single individuals. In both **c** and **d** the dashed line indicates the 95% percentile

Figure 4 Depletion in the isolates of homozygous genotypes at loss of function (LoF) variants. **(a)** Counts of derived allele per individual at variable sites with missense synonymous and LoF consequences sites. Dark and light asterisks indicate statistically significant (p-value<0.001, Table S9) differences from TSI and CEU, respectively. Overall in isolates individuals have less derived alleles compared to reference populations. **(b)** Fraction of genotypes homozygous for the derived

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

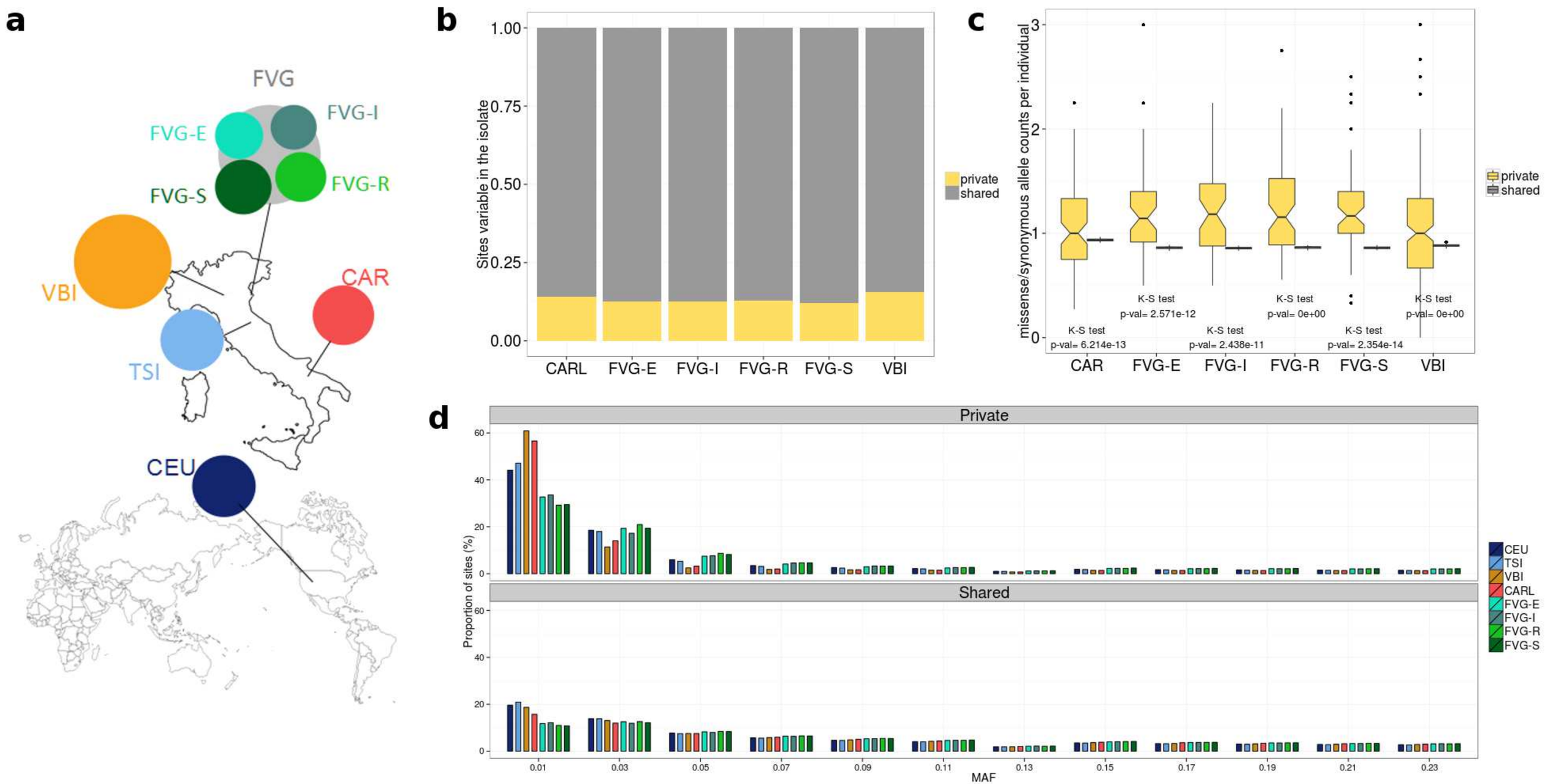
allele normalized by homozygous derived genotypes at intergenic sites. At mildly deleterious sites (missense and synonymous) isolates have significantly higher rates of homozygosity compared to reference (Table S9) whereas at LoF sites rates are comparable, suggesting a depletion of homozygous deleterious genotypes in isolates.

Figure 5 Effectiveness of selection Pairwise comparison of isolates with CEU **(a)** and TSI **(b)** at missense, synonymous and LoF sites. $R_{X/Y}$ is a measure of the accumulation of mutations in population X respect to population Y; values lower than 1 indicate a depletion in population X and *viceversa*. Bars represent variance obtained from block jackknifes. While for mildly deleterious mutations $R_{X/Y}$ follows a trend compatible with less efficient selection and stronger drift, $R_{X/Y}$ at LoF shows more efficient selection in isolates compared to reference populations.

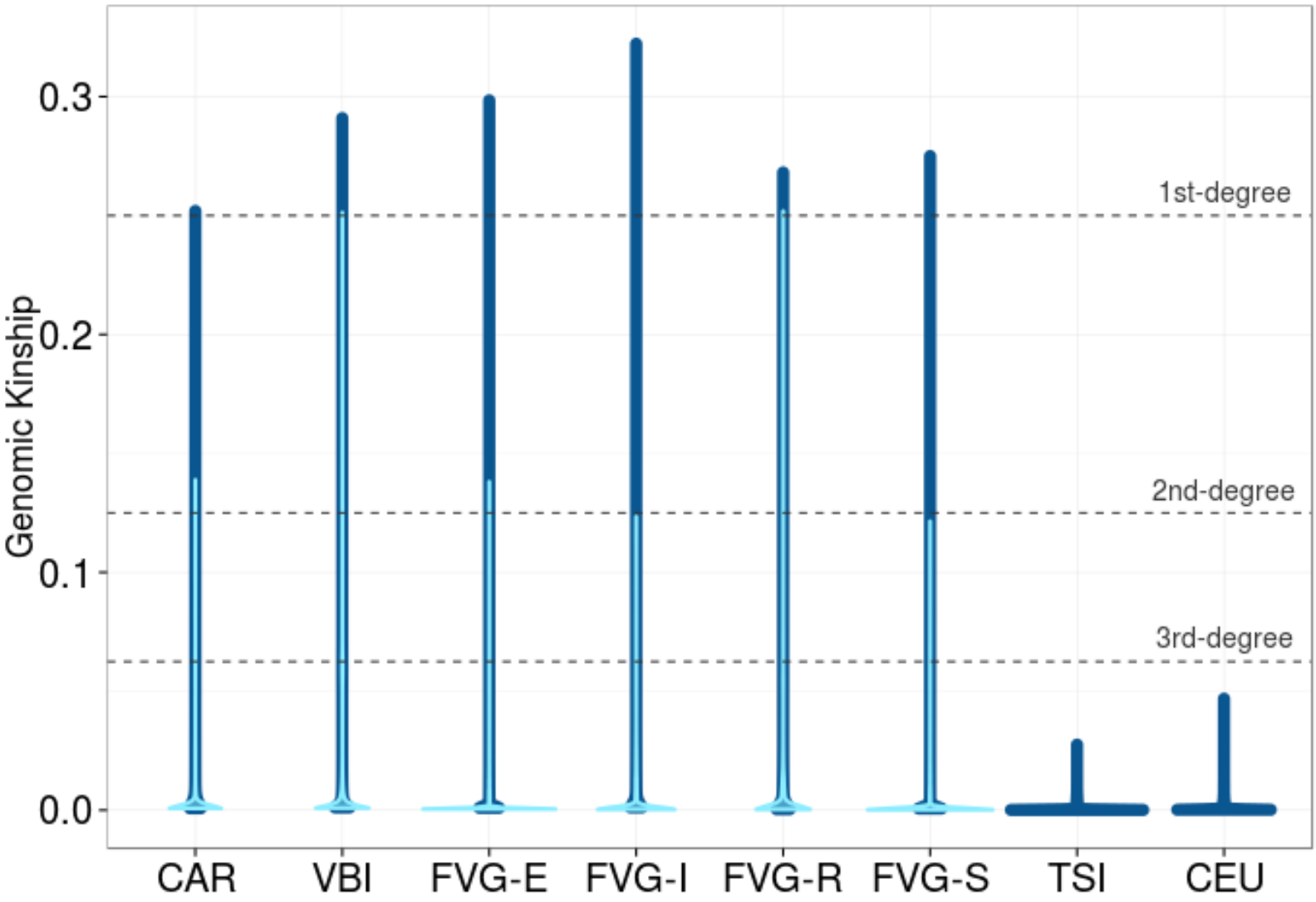
Figure 6 Expectations to observe purging. Purging is measured as the fold change between present and past (at founding event) average number of recessive deleterious homozygous genotypes per individual. If this number was less than one the (negative) reciprocal is listed. A negative fold change is expected in case of purging. Boxplots summarize results of 100 replicates, each replicate being the average over 50 individuals. Dominant, additive and recessive models are shown for neutral (selection coefficient, s=0), mildly deleterious (s=0.01) and highly deleterious (s=0.1) variants in three populations with different average kinship. Ne=effective population size.

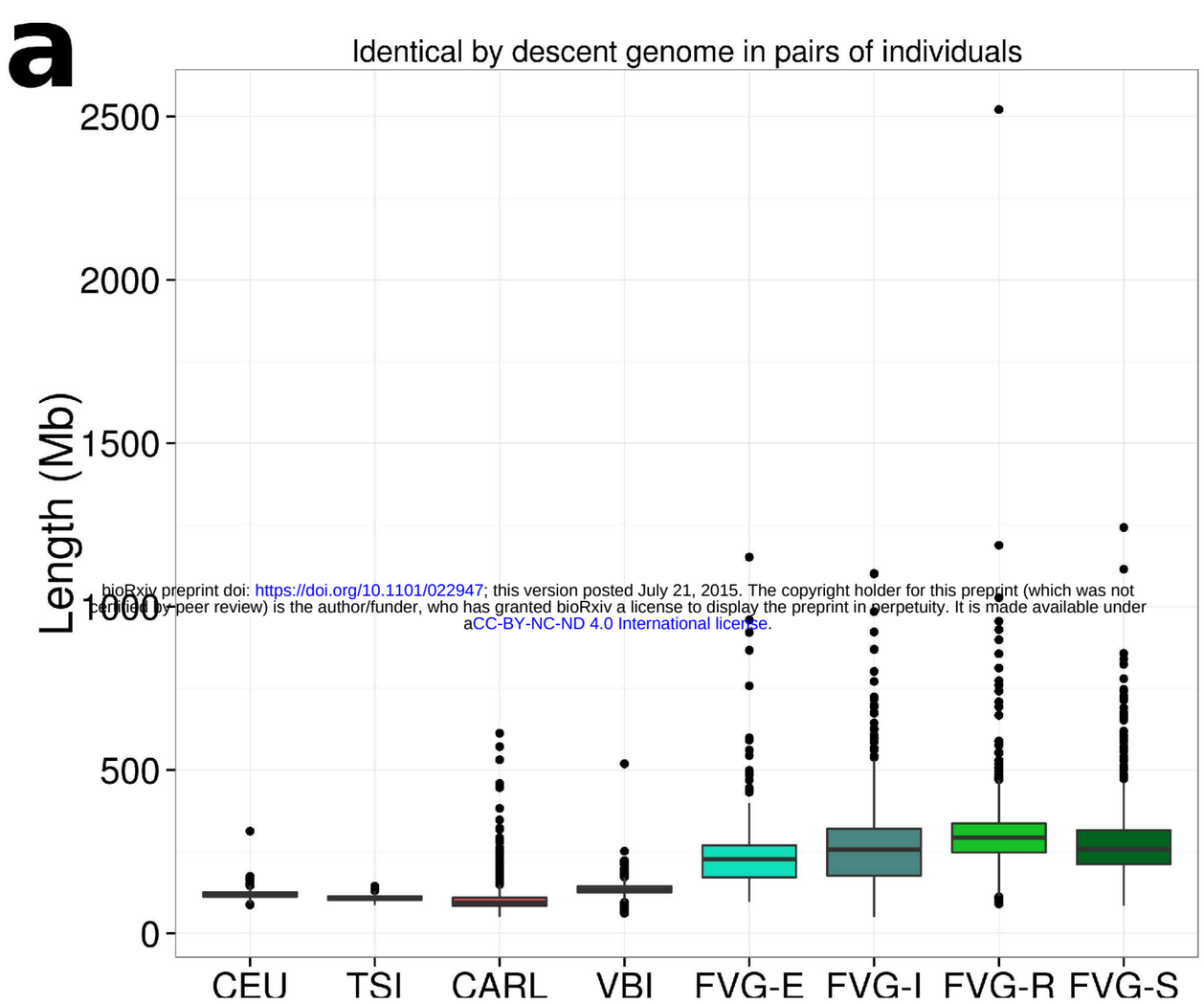
Figure 7 Effective population size Skyline plot of effective population size in very recent times inferred from identical by descent genome of individuals within populations

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

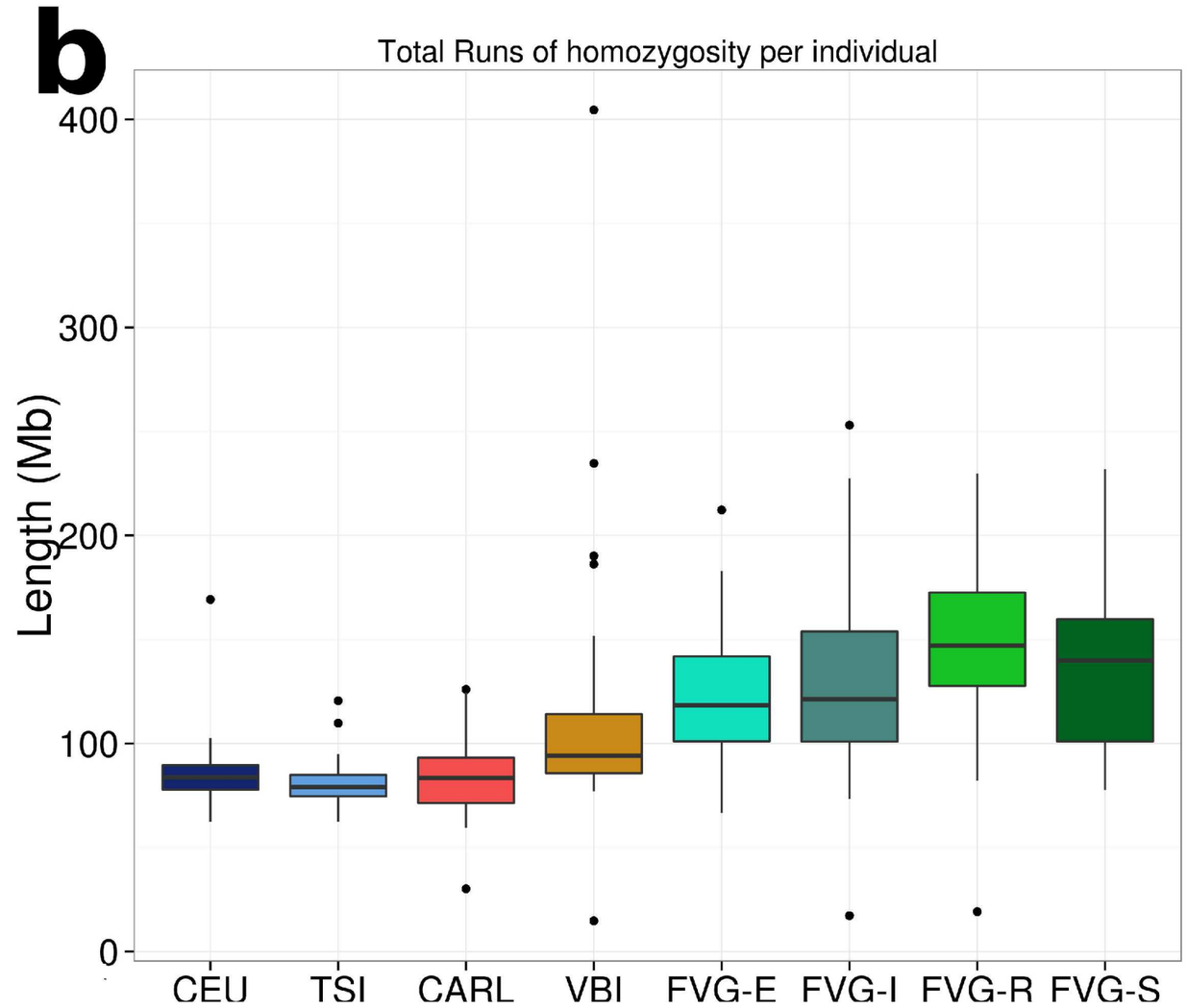

a
FVG
FVG-I
FVG-E
FVG-S
FVG-R
VBI
CAR
TSI
CEU
b
Sites variable in the isolate
1.00
0.75
0.50
0.25
0.00
CARL
FVG-E
FVG-I
FVG-R
FVG-S
VBI
private
shared
c
missense/synonymous allele counts per individual
K-S test
p-val= 6.214e-13
K-S test
p-val= 2.571e-12
K-S test
p-val= 2.438e-11
K-S test
p-val= 0e+00
K-S test
p-val= 2.354e-14
K-S test
p-val= 0e+00
CAR
FVG-E
FVG-I
FVG-R
FVG-S
VBI
private
shared
d
Private
Shared
Proportion of sites (%)
MAF
CEU
TSI
VBI
CARL
FVG-E
FVG-I
FVG-R
FVG-S

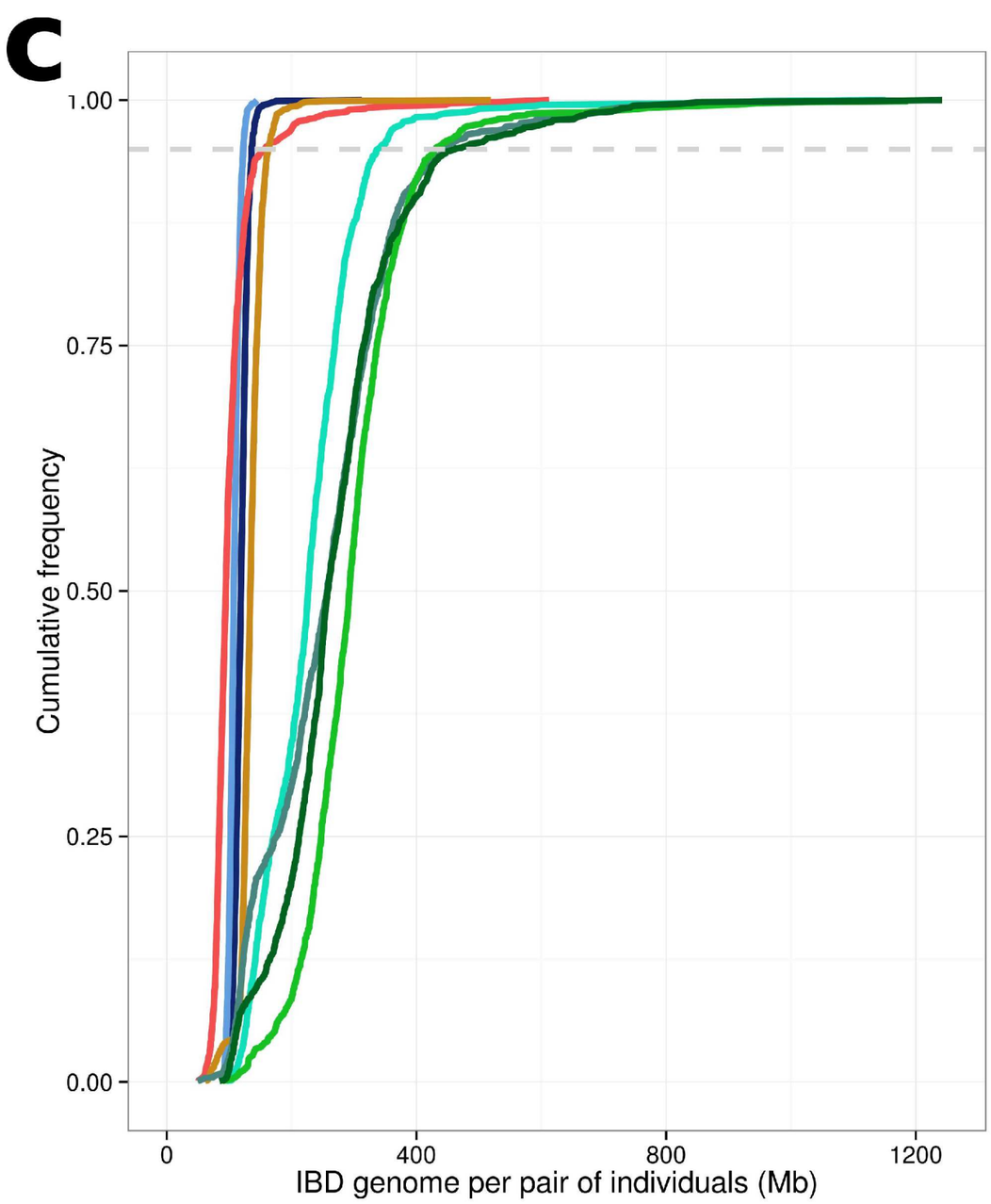
bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

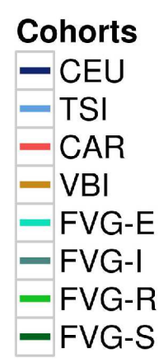


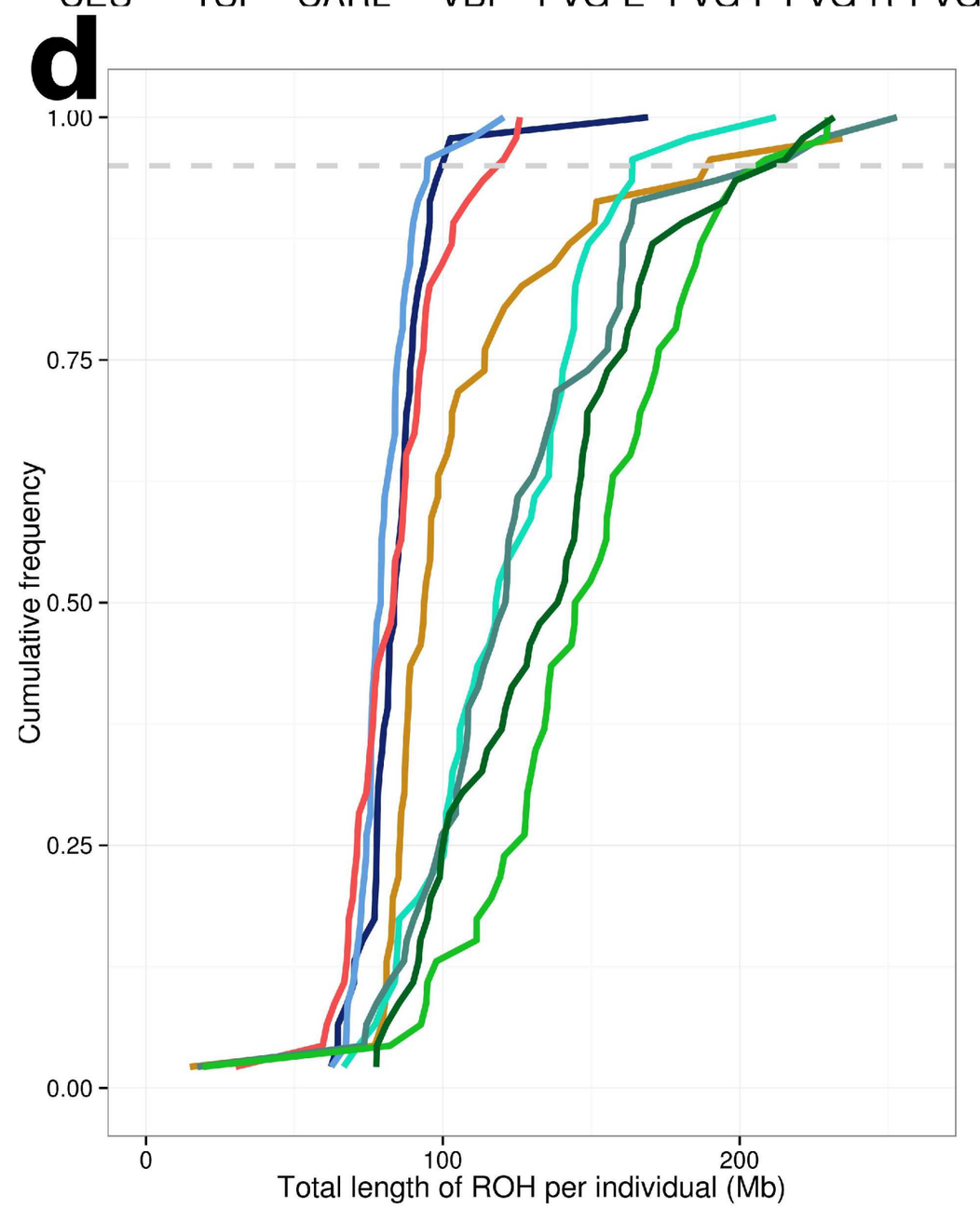


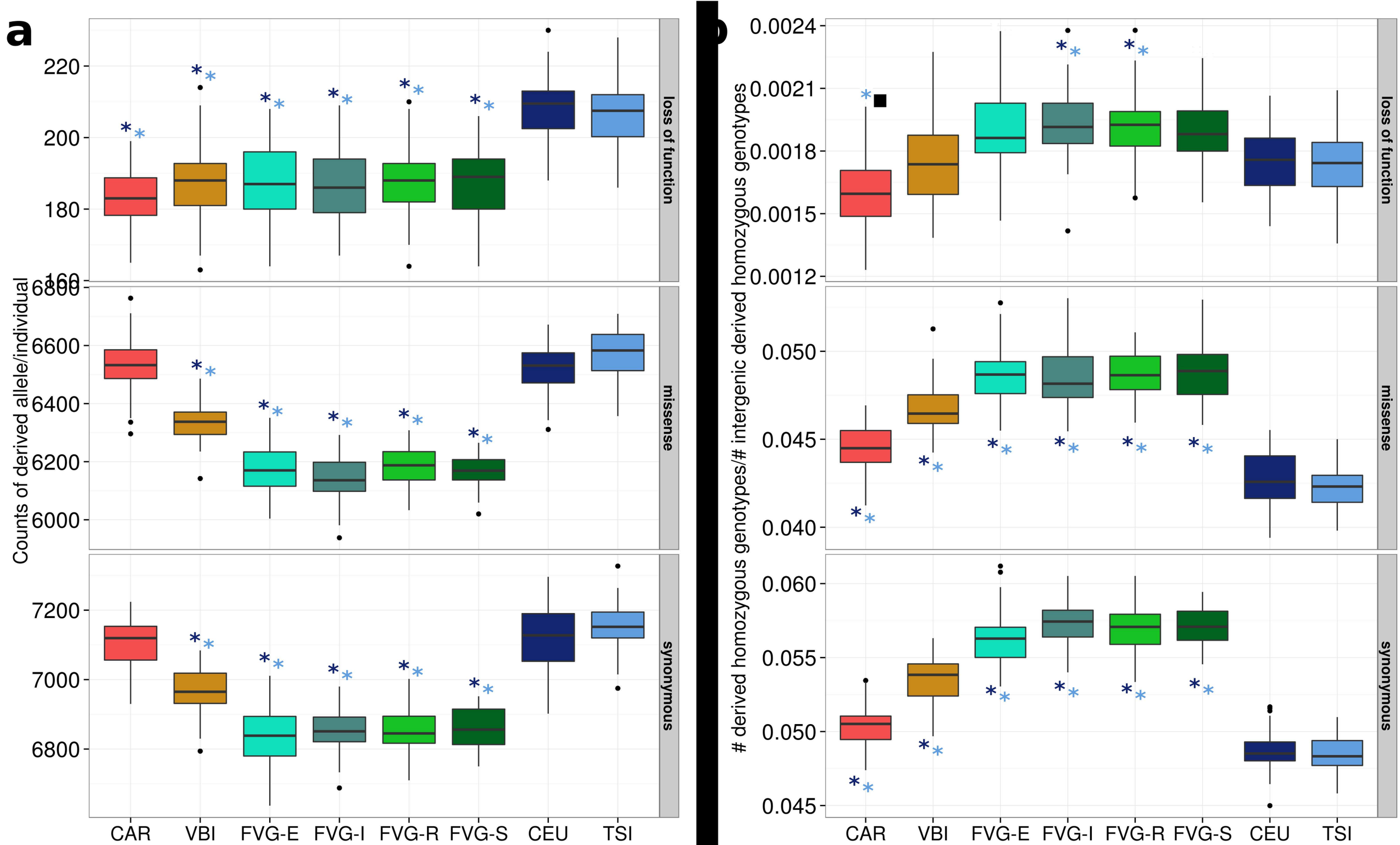
bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

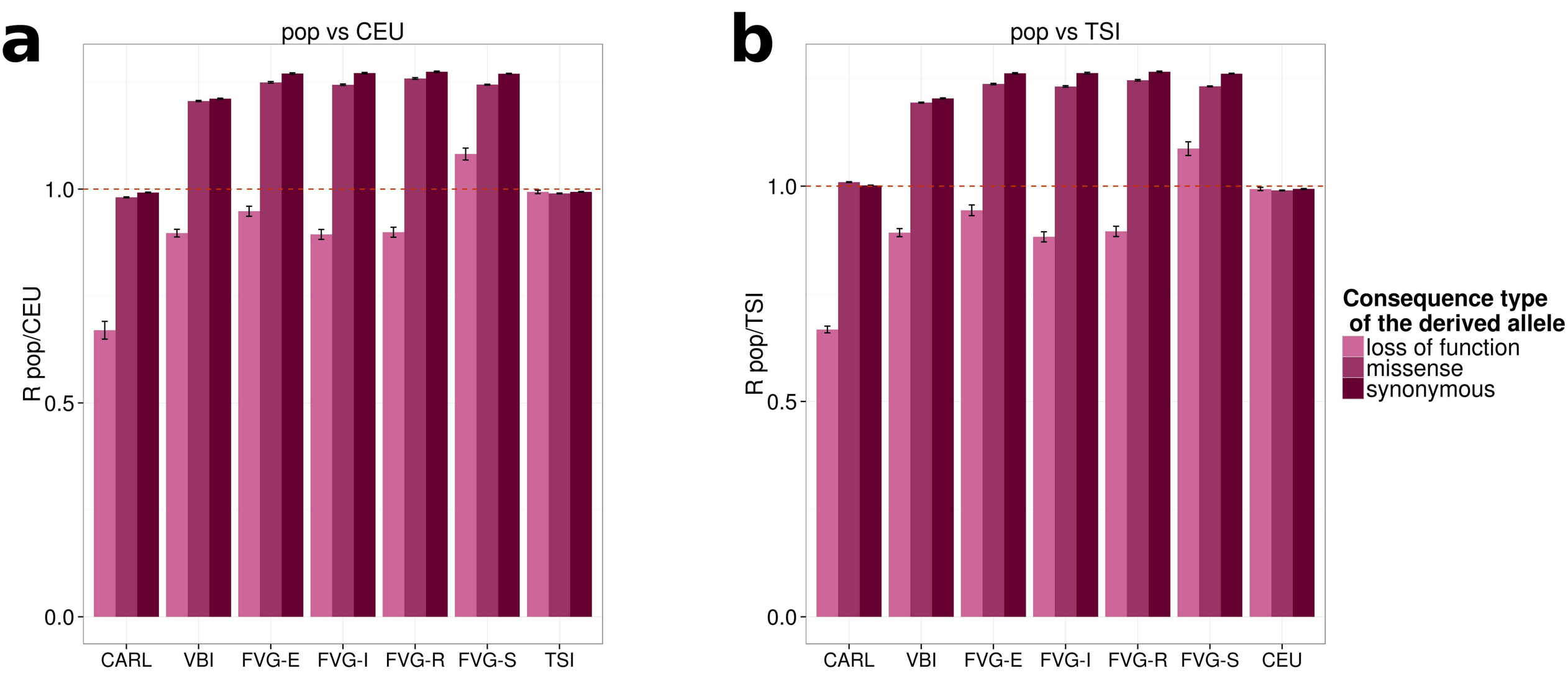




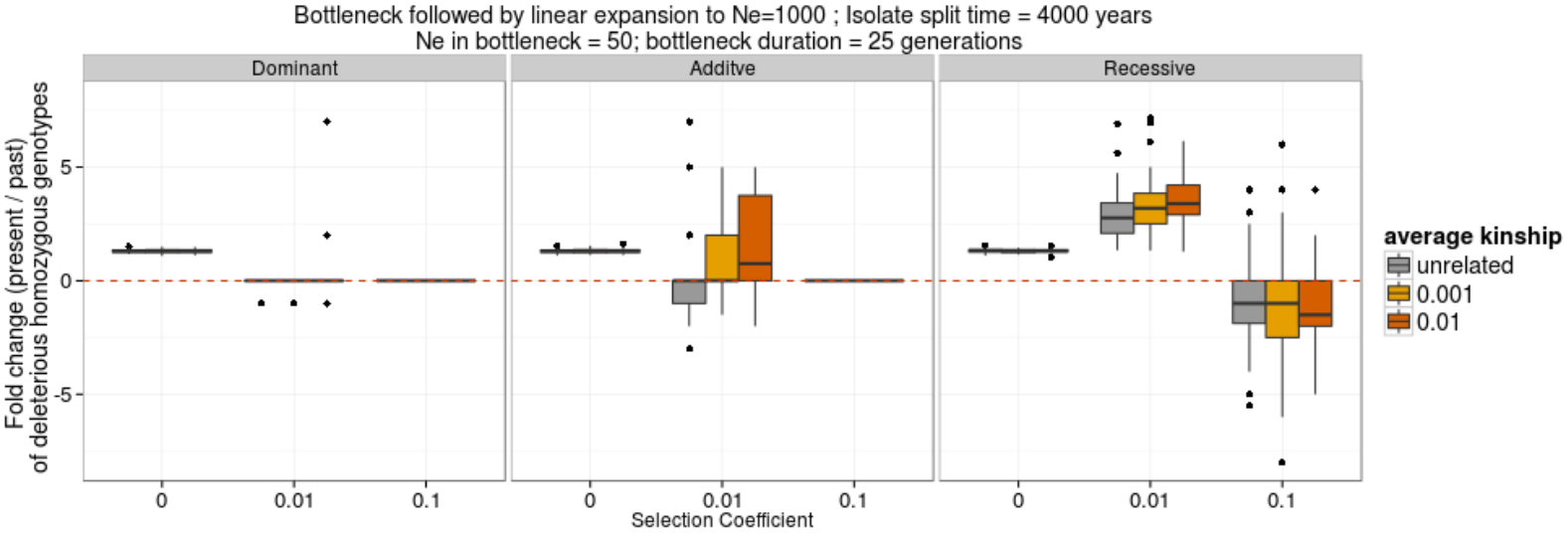





a
b
Counts of derived allele/individual
derived homozygous genotypes/# intergenic derived homozygous genotypes
loss of function
missense
synonymous
CAR
VBI
FVG-E
FVG-I
FVG-R
FVG-S
CEU
TSI


a
pop vs CEU
R pop/CEU
1.0
0.5
0.0
CARL
VBI
FVG-E
FVG-I
FVG-R
FVG-S
TSI
b
pop vs TSI
R pop/TSI
1.0
0.5
0.0
CARL
VBI
FVG-E
FVG-I
FVG-R
FVG-S
CEU
Consequence type
of the derived allele
loss of function
missense
synonymous

bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/022947; this version posted July 21, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

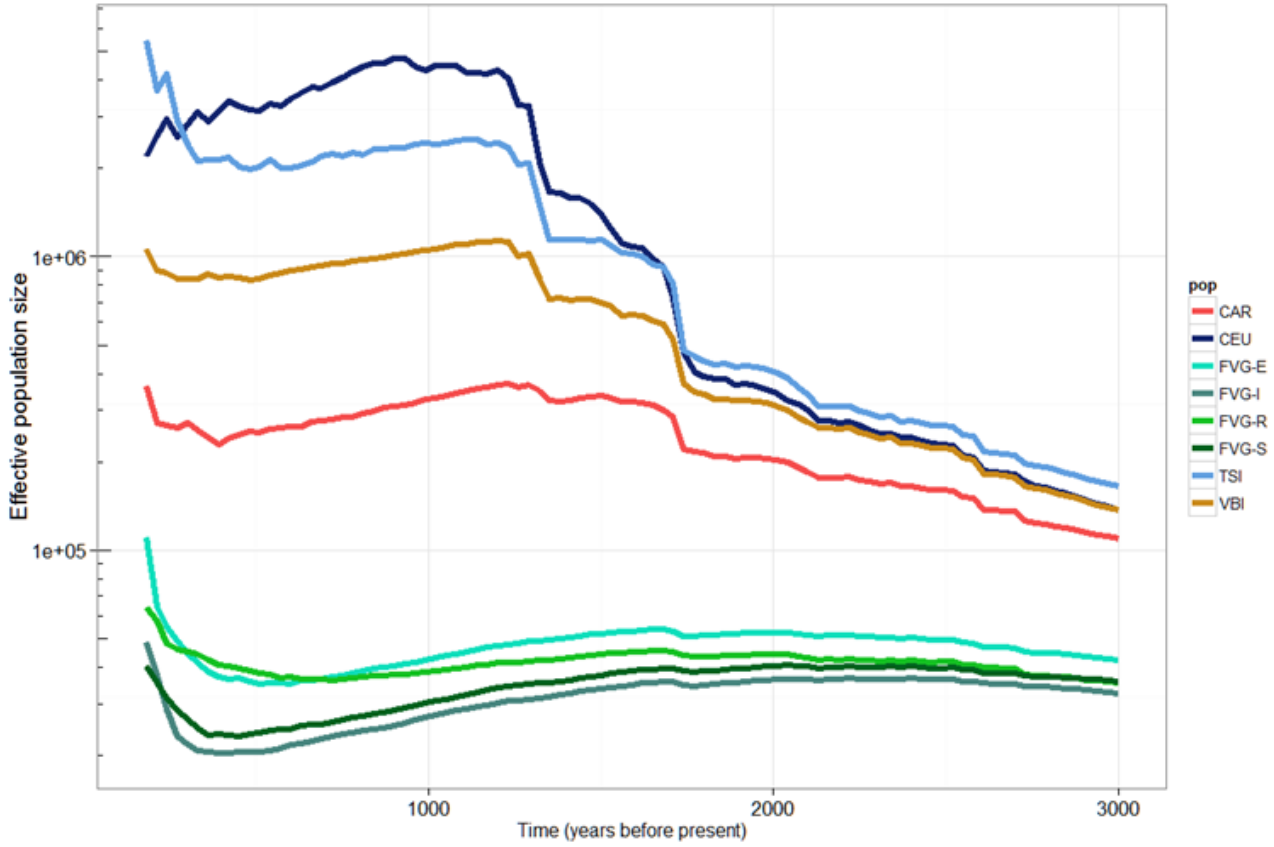



Figure-7_Cocca_etal