Supplemental Notes

Numerical relationships between the mutations and different categories of variable genes in a genome

We established highly specific and constant numerical relationships between the number of mutations per genome and the number of phase-sensitive genes, and any category of variable genes in general. So the ratio (quotient) of the number of protein-altering mutations per genome (median 2,869 in 1,092 genomes, and between 2,655.5 in EAS and 3,672 in AFR in the four ancestry groups) to the total of variable genes defined by presence of ≥ 1 protein-altering mutations was 0.71, to the genes with = 1 mutation 0.54, to the phase-sensitive genes with ≥ 2 mutations 0.17 and their sub-categories with *cis* and *trans* configurations 0.10 and 0.07, respectively. These ratios, which were calculated from median values, were essentially identical in all ancestry groups and the total set of 1,092 genomes, with minor shifts of 1% in AFR towards phase-sensitive genes and *trans* configurations due to the increased number of mutations in this ancestry group (Supplemental Table S6A). Calculating the ratios individually for each of the EUR and AFR genomes, they were found to vary within a very small range; so did the numbers of mutations per genome and the numbers of variable genes in each of the defined categories. We also refer to these ratios as proportionality constants. Furthermore, also the fractions of mutations that were found distributed to each of these categories were nearly the same across all ancestry groups (Supplemental Table S6B): about 53–54% of the protein-altering mutations per genome generated genes with = 1 mutation and over 46–47% of the mutations were contained in phase-sensitive genes with ≥ 2 mutations; of those, 25–26% were found to reside on the same chromosomal homologue (in *cis* configurations) and the remaining 20–22% were located on both homologues of a gene (in *trans* configurations). Again, minor shifts towards phase-sensitive genes (~1–2%) and *trans* configurations (~3–4%) were observed in AFR. Overall, the number of protein-altering mutations per phase-sensitive gene remained constant at ~2.7.

Highly specific and constant numerical relationships between the numbers of mutations per genome and the different categories of variable genes were also obtained when the entirety of AA exchanges were examined, although with different ratios, as shown in Supplemental Table S6C. Analyzing these ratios again individually across the genomes in EUR and AFR, the individual values were within very narrow corridors, as were the numbers of AA exchanges per genome and the numbers of variable genes in each of the defined categories. In this case, the 2.25-fold (AFR 2.2-fold) higher number of AA exchanges per genome resulted in a disproportionately, 2.7-fold (AFR 2.6-fold) higher number of phase-sensitive genes. Thus, these contained a much higher fraction of AA exchanges, 62% (AFR 64%), with 31–33% (AFR 29%) of the AA exchanges residing in *cis* and 29–31% (AFR 35%) in *trans*, the number of AA exchanges per gene again remaining again constant, 3.2 (Supplemental Table S6D). Thus, these results indicate that the considerably higher load of mutations in this case leads to a disproportionately higher increase of phase-sensitive genes, which accumulate a relatively higher number of mutations per gene.

To explain these internally consistent and highly constant numerical relationships (‘proportionality constants’) between the mutations per genome and the numbers of genes with ≥ 1 mutations, = 1 mutation, ≥ 2 mutations, *cis* and *trans* configurations, we hypothesized that these could be due to the inherent nature of the underlying exome structure. In fact, simulating a random distribution of protein-altering mutations and AA exchanges as a whole, onto existing exome structure as described in Methods, we were able to well approximate the ‘proportionality constants’ for these categories of mutated genes (Supplemental Table S7). So the simulated quotient relating the number of protein-altering mutations per genome to the total of variable genes with ≥ 1 mutations was 0.68 as compared to 0.71, to the genes with = 1 mutation 0.49 as compared to 0.54, and to the phase-sensitive genes with ≥ 2 mutations 0.19 as compared to 0.17. The same was true for the simulated quotients obtained for the entirety of AA exchanges. So the simulated quotient relating the number of AA exchanges per genome to the total of variable genes with ≥ 1 AA exchanges was 0.59 as compared to 0.58, to the genes with = 1 AA exchange 0.37 as compared to 0.38, and to the phase-sensitive genes with ≥ 2 AA exchanges 0.22 as compared to 0.20. This supported our hypothesis that these highly constant numerical relationships between mutations and the different categories of mutated genes could be due to the inherent nature of the underlying exome structure.

In order to simulate the ratios for the numerical relationships between the number of mutations per genome and the numbers of phase-sensitive genes with either *cis* or *trans* configurations, we simulated phase as described in Methods. Importantly, for both protein-altering mutations and the entirety of AA exchanges, the proportionality constants could not be reproduced under the assumption of random distribution of mutations between the two homologues for both *cis* and *trans* forms of the genes. The simulated proportionality constants, hence the ratios for the *cis* and *trans* forms of the phase-sensitive genes were in effect reversed in both cases, corresponding to ~40:60 *cis*/*trans* ratios instead of the ~60:40 ratios observed. This result has been described in main text in context with the simulations of phase for combinations of 2 up to 5 mutations under conditions of random distribution of mutations between the two homologues. Thus, the 'true' proportionality constants, resulting in a 60:40 *cis*/*trans* ratio of the *cis* and *trans* forms, are effectively reversed.

Global sets of variable and phase-sensitive genes: overrepresentation of pathways and GO terms

In this work, we have focused on genes with ≥ 2 mutations to examine the distribution of mutations between the two homologues and global patterns of phase in the diploid human genome. From a biological point of view, however, two different forms of a protein preserving functional flexibility can be generated by any variable gene with at least one mutation encoding a 'molecular diplotype' (Hoehe et al. 2014). Thus, we examined to which extent the over-representation of described functional content is restricted to phase-sensitive genes, or part of a bigger picture. To this end, we extracted analogously a larger global set of 7,524 genes with ≥ 1 mutation, which was found significantly enriched for 138 pathways (P<3.5 x 10^{-45}–9.7 x 10^{-3}) and 177 GO terms (P<1.09 x 10^{-37}–9.6 x 10^{-4}) (Supplemental Table S10A, B). Notably, this set of genes includes the 2,402 phase-sensitive genes that constitute the global set. Subsequently, we determined the relative proportions of genes with = 1 mutation and ≥ 2 mutations, respectively, for each of these pathways and GO terms (separately for 'biological process', 'cellular component' and 'molecular function'; Supplemental Fig. S4A, B). Evidently, the phase-sensitive genes, which represent ~32% of the total set of variable genes, represent a disproportionately large fraction within each pathway and GO term, for instance between 71.4% and 32% in most pathways, supporting their high functional load.

Quantitative characterization of *cis*- and *trans*-abundant genes

To characterize *cis*- and *trans*-abundant genes in more detail, we examined first the 1,227 *cis*- and 786 *trans*-abundant genes identified in the global set of phase-sensitive genes (1000G), and subsequently the subsets of these genes which were shared with the experimentally phased genomes (PGP). *Cis*-abundant genes had ≥ 2 protein-altering mutations in 23% of the 1,092 genomes (1000G) on average (a fraction corresponding well with the spectrum of average mutation frequencies from all pairs of protein-altering mutations described earlier). The mutations were in *cis* configurations on average in 82% of the genes with ≥ 2 protein-altering mutations, with a maximum absolute *cis* count of 769 out of 1,016 *cis* configurations in the gene CDK11B (Supplemental Table S12A). The gene-based *cis* fractions tended to be high, for example, 35% of the *cis*-abundant genes had *cis* fractions ≥ 90% and 57% *cis* fractions ≥ 80% of total configuration count. Correspondingly, the (cross-validated) *cis*-abundant genes in PGP had ≥ 2 protein-altering mutations in 24% of the 184 genomes, of which 90% on average resided in *cis*. Furthermore, *trans*-abundant genes had ≥ 2 protein-altering mutations in 19% of the 1,092 genomes (1000G), of which 80% resided in *trans,* with a maximum *trans* count of 1,092 (out of 1,092 configurations) in the gene MAP2K3 (Supplemental Table S12B). Specifically, 29% of all *trans*-abundant genes had gene-based *trans* fractions ≥ 90%, and 50% *trans* fractions ≥ 80%. Correspondingly, in PGP, *trans*-abundant genes had ≥ 2 protein-altering mutations in 22% of the 184 genomes, of which 86% resided in *trans.* Remarkably, 4% and 8% of the genes in the 1,092 genomes, and 27% and 18% of the genes in the 184 PGP genomes, consisted solely of *cis* or *trans* configurations, respectively.

***Cis*- and *Trans*-abundance: Characteristics of autosomal genes**

We corroborated in an expanded analysis that *cis*- and *trans*-abundance in effect represent a fairly constant characteristic in nearly 90% of the genes, with only 8.7% of all *cis*- and 12.9% of all *trans*-abundant genes changing configuration type in PGP. To this end, we identified first those phase-sensitive genes, which were shared by the global set (1000G) and PGP (1,627 genes representing 68% of the global set), and their configuration types in each of the two sample sets. Then we intersected the genes which were *cis*-abundant in 1000G with the genes *trans*-abundant in PGP and vice versa, the genes *trans*-abundant in 1000G with those *cis*-abundant in PGP. The identification of overlaps of 72 and 71 genes, respectively, indicated that 8.7% of the *cis*- and 12.9% of the *trans*-abundant genes had changed configuration type in PGP. Thus, *cis*- and *trans*-abundance was found constant in nearly 90% of the genes examined, and could therefore represent a characteristic of autosomal genes, confirming earlier results.

Autosomal *cis*/*trans* ratios

Examining *cis* and *trans* configurations of protein-altering mutations per autosome, *cis* fractions ranged between 52.6% and 68.8% (median values), resulting in a median of 60% across all autosomes (Supplemental Table S14A). Assessing autosomal *cis* fractions separately per ancestry group unveiled certain patterns: nearly half of the autosomes exhibited very high *cis* fractions between 60% and 71.4% in EUR, EAS and AMR, with the corresponding fractions in AFR being roughly 5% to 20% lower; several autosomes exhibited high *cis* fractions in all four ancestry groups, between ~60% and 73.5%; and few autosomes such as chromosome 6 harboring MHC had comparatively higher *trans* fractions in all ancestry groups, that is, lower *cis* fractions between 56% and 48.8%, the latter value (in AFR) being the lowest overall. Such patterns were also observed when analyzing the autosomal *cis* fractions based on the entirety of AA exchanges (Supplemental Table S14B). Although the *cis* fractions in this case showed some variation compared to the results obtained from protein-altering mutations, the resulting net global *cis*/*trans* ratios remained, however, nearly the same. Thus, global *cis* abundance is the net result of diverging autosomal *cis*/*trans* ratios, which apparently results from different mixtures of *cis*- and *trans*-abundant genes on each of the autosomes. Also, there is no overall proportional decline of *cis* ratios in AFR. While the majority of autosomes show a decrease of *cis* ratios most possibly due to the erosion of linkage disequilibrium (LD), the net *cis*/*trans* ratios of some autosomes appear to remain unchanged in the population samples, with their *cis*/*trans* ratios being either relatively high or low. Thus, these autosomes specifically may harbor groups of genes with an abundance of either phase configuration, suggesting that the phase of mutations may be of potential biological importance in these cases. The specific distribution of *cis*- and *trans*-abundant genes across all autosomes is shown in Supplemental Fig. S7A, B.