Chapter 15

Analysis of 1276 Haplotype-Resolved Genomes Allows Characterization of *Cis*- and *Trans*-Abundant Genes

Margret R. Hoehe and Ralf Herwig

Abstract

Many methods for haplotyping have materialized, but their application on a significant scale has been rare to date. Here we summarize analyses that were carried out in 1092 genomes from the 1000 Genomes Consortium and validated in an unprecedented number of 184 PGP genomes that have been experimentally haplotype-resolved by application of the Long-Fragment Read (LFR) technology. These analyses provided first insights into the diplotypic nature of human genomes and its potential functional implications. Thus, protein-changing variants were not randomly distributed between the two homologues of 18,121 autosomal protein-coding genes but occurred significantly more frequently in *cis* than in *trans* configurations in virtually each of the 1276 phased genomes. This resulted in global *cis/trans* ratios of ~60: 40, establishing "*cis* abundance" as a universal characteristic of diploid human genomes. This phenomenon was based on two different classes of genes, a larger one exhibiting *cis* configurations of protein-changing variants in excess, so-called "*cis*-abundant" genes, and a smaller one of "*trans*-abundant" genes. These two gene classes, which together constitute a common diplotypic exome, were further functionally distinguished by means of gene ontology (GO) and pathway enrichment analysis. Moreover, they were distinguishable in terms of their effects on the human interactome, where they constitute distinct *cis* and *trans* modules, as shown with network propagation on a large integrated protein–protein interaction network. These analyses, recently performed with updated database and analysis tools, further consolidated the characterization of *cis*- and *trans*-abundant genes while expanding previous results. In this chapter, we present the key results along with the materials and methods to motivate readers to investigate these findings independently and gain further insights into the diplotypic nature of genes and genomes.

Key words Haplotypes, Diplotypes, Phase, Haplotype-resolving genomes, Long Fragment Read (LFR) technology, *Cis* abundance, *Cis*- and *trans*-abundant genes, Common diplotypic exome, Network propagation

Supplementary Information The online version contains supplementary material available at [https://doi.org/10.1007/978-1-0716-2819-5_15].

Brock A. Peters and Radoje Drmanac (eds.), *Haplotyping: Methods and Protocols*,
Methods in Molecular Biology, vol. 2590, https://doi.org/10.1007/978-1-0716-2819-5_15,

1 Introduction

"Much of biology revolves around the number 2, or more precisely, the repercussion of > 1. The number 1, by itself, establishes existence, whereas 2 and all other numbers > 1 suggest persistence" (Wu C-T and Dunlap JC (2002), Homology effects: The difference between 1 and 2) [1].

1.1 Moving from a "World of 1" to a "World of 2"

Human genomes are diploid by nature. Thus, as an indispensable prerequisite for *eukaryotic* evolution and bisexual reproduction, every individual has two genomes, that is, two sets of chromosomes, one from the father and one from the mother. This number two guarantees the survival of the human species, the situation of homology allowing evolution through the generation of diversity and the regulation as well as maintenance of function. Where the two homologues of a gene or any functional genomic unit are not the same, but *different*, evolution chooses *different* from time to time allowing adaptation to changing environments. The existence of two (different) homologues of a gene not only facilitates a high functional flexibility of genes and their products but also represents a fail-safe system in the event that one of the two homologues should be perturbed, preserving the function of the gene.

To study the human genome in its diploid nature, in an ideal world with all technologies at hand, one would simply split open a cell, separate all chromosomes from one another, and sequence each one from end to end. This would allow a direct determination of the specific combinations of genetic variants as they exist on each of the two homologous chromosomes, also defined as *haplotypes*, which as a pair constitute a *diplotype*. It is becoming increasingly clear that the *phase* of the variants, that is, the way in which the variants are distributed between the two homologous chromosomes, is of key importance [2–4]; whether variants reside on the same chromosome, in *cis*, or on both chromosomes, in *trans*, can critically impact gene/protein function and phenotype. Thus, Seymour Benzer showed as early as 1957 that two null mutations in a *cis* configuration leave at least the second form of the gene intact, while in the case of a *trans* configuration, both forms of the gene are defect, changing phenotype completely [5]. As an example of a more complex scenario, Drysdale and colleagues [6] demonstrated that different pairs of promoter and coding region β_2-adrenergic receptor gene (*ADRB2*) haplotypes elicit significantly divergent in vitro and in vivo responses to β agonist in asthmatics, that is, induce different biologic and therapeutic phenotypes. Concerning the interaction between genes, e.g., when a mutation in the cell essential gene *Rpa1* and a closely linked mutant allele of the tumor suppressor gene *Trp53* were in *cis*, the tumor phenotype was attenuated and survival prolonged, whereas in the *trans* configuration, *Rpa1*^{L230P} significantly enhanced tumorigenesis and reduced survival [7]. Established phenomena that attest to the importance of phase include moreover compound heterozygosity in monogenic disorders as well as numerous other specific (complex) diplotypes in common diseases and pharmacogenetic phenotypes [3]. Furthermore, phase information is crucial where diploid organisms express but one of the two homologues of a gene. This is for instance the case in phenomena as widespread as allele-specific expression (ASE), allele-specific methylation (ASM)/genomic imprinting and monoallelic expression (MAE), and the loss of heterozygosity. Thus, in differentially expressed transcriptomes and proteomes, the specific combinations of variants situated on the *expressed* homologues will be the ones that actually exert function. Thus, phase information is essential to understand the diploid biology of genes and genomes and establish meaningful relationships between DNA sequence, gene/protein function, and phenotype. Ultimately, DNA sequence and its variation can only be understood in phase context.

Although the diploid nature of the human genome is obvious, it has been much too often ignored. Current genome biology still relates both empirically and conceptually mostly to a "World of 1." The approaches to genome sequence analysis, documentation, description, and annotation routinely rely on *one* single DNA sequence readout per individual, ultimately the result of a key strategic decision due to financial and technical limitations at the dawn of the Human Genome Project. The DNA templates routinely prepared for sequence analysis actually represent a mixture of both paternal and maternal chromosomes, resulting in the generation of "mixed diploid" sequence, that is, a composite of both haploid parental sequences, instead of a *separate* readout of the two parental homologues.

1.2 Many Haplotyping Methods, Not Much Application

To date, an estimated hundreds of thousands [8] to 30 million [9] human genomes have been routinely generated as single readouts, as have roughly about 700,000 exomes, with the largest studies including 125,748 [10] and 454,787 exomes [11]. In stark contrast, at most about 300 (partially) haplotype-resolved genomes have materialized since 2007 [12, 13]. Notably, the first direct experimental haplotyping methods at the genome scale were described in 2011 including fosmid pool-based next-generation sequencing approaches [4, 14] or the use of a microfluidic device [15]. The Long-Fragment Read (LFR) technology was reported soon thereafter [16]. Many haplotyping methods or variations thereof have followed since [17, 18], with the most recent one producing a fully phased de novo genome assembly using single-cell strand sequencing and long reads [19]. The majority of the studies typically described one or a few genomes to illustrate a new or improved haplotyping method. However, while each new method has been hailed, too little attention has been paid to the application of these methods at any significant scale. In only a few studies, an experimental method has been applied to haplotype-resolve a sizeable number of genomes in order to be able to address obvious questions concerning the diplotypic nature of human genomes or their structural variation. Thus, we have validated and applied a fosmid pool-based next-generation sequencing approach [4, 20, 21] to haplotype-resolve and comprehensively analyze a set of 14 human genomes in 2014 [22], the largest data set produced at the time. Over 100 personal genomes that were experimentally phased by application of the LFR technology were reported in 2016 [23], and analyses including a total of 184 LFR haplotype-resolved genomes (generously provided by Brock Peters and Radoje Drmanac) were described by us in 2019 [24]. Finally, 32 diverse human genomes that were phased by means of a combination of long-read and strand-specific sequencing technologies [19] to perform integrated analysis of structural variation were reported in 2021 [25]. Thus, the stage is set to move from a "World of 1" toward a "World of 2."

1.3 From Methods to Insight

This requires, in addition to appropriate technologies, a reevaluation of previously prevailing conceptual approaches to haplotype analysis. With the development of experimental haplotyping methods, the obvious biological interpretation of haplotypes as the molecular equivalents of the two homologues of a diploid gene (or any functional region) has finally been gaining momentum. This is worth mentioning because haplotypes have been conceived for most of the time as genetic markers, particularly as the research objects of the major human genome-related initiatives "International HapMap Project" and "1000 Genomes Project." Accordingly, a genome-wide map of "haplotypes," that is, "blocks," or combinations of common SNPs in linkage disequilibrium (LD), was developed to allow efficient inference by means of LD of unobserved disease variants. A biological conception of haplotypes far beyond the mainstream, the author (MH) has early on focused on "gene-based functional haplotypes," because it is "essential in diploid organisms to determine the specific combinations of *all* given gene sequence variants for each of the (2) chromosomes defined as haplotypes" [2, 26]. "The correct determination of the molecular haplotypes underlying each genotype ... is essential to make conclusions on the 'functionality' of both forms of the gene and establish relationships between gene variation and gene function..." [2]. In 2011, Tewhey and colleagues [3] elaborated the importance of phase information in remarkable detail, particularly at the example of settings in which the characterization of haplotypes is "essential for understanding phenotypic expression... and disease states." At the same time, we illustrated many of these considerations using "Max Planck 1" (MP1) then the most comprehensively haplotype-resolved and annotated individual human genome [4].

Researchers have since addressed an increasingly broad spectrum of specific instances, such as disease-related phenotypes, where phase matters. Of particular importance in this context are studies on clusters of two or more nearby variants that exist on the same haplotype in an individual, so-called multinucleotide polymorphisms (MNPs) [27, 28], mutations (MNMs) [29–31], or variants (MNVs) [10, 32]. Combinations of variants within a codon can have different functional consequences than the individual variants. Because currently available tools do not correctly classify MNVs, misannotations are common, resulting in missed diagnoses or false-positive pathogenic candidates [28, 31]. In a first large-scale study of MNVs, Wang et al. [10] examined the functional impact of SNP pairs within 2 bp distance of each other in 125,748 exomes with local phase information. In total, 18,756 MNVs showed a novel, combined effect on protein sequence including 407 gained nonsense and 1821 rescued nonsense mutations, further underscoring the value of haplotype-aware annotation. Thus, the "World of 1" consists of many exceptions already that reflect a "World of 2."

To move toward a more coherent, holistic view of the "World of 2," we first wanted to learn more about the diplotypic nature and architecture of the human genome. To this end, we addressed the following questions: (i) How are protein-coding variants, especially variants of potential functional significance, distributed between the two homologues of the autosomal genes? Can we distinguish certain patterns based on the distribution of *cis* and *trans* configurations as the two major categories of (diplotypic) genetic variation? For instance, could *cis* configurations, leaving one form of the gene intact, be expected to occur more frequently in human genomes than *trans* configurations to preserve organismal function? (ii) Moving from the analysis of whole genomes to the genes, can we identify a specific, common subset of genes that preferentially have *cis* or *trans* configurations, that is, that preferentially encode two potentially functionally different homologues, diplotypes? If so, does this concern specific functional classes of genes? (iii) Analyzing then the rates of *cis* and *trans* configurations per gene, can we distinguish diplotypic genes further by an excess of either configuration? (iv) If so, can we further functionally distinguish these classes of genes at different levels of molecular organization such as functional annotation terms, pathways, or protein–protein interaction networks?

In this chapter, we first start with an overview of the terms and definitions that we have developed and applied to assess the "World of 2." Subsequently, we summarize the key results providing first insights into the diplotypic nature and architecture of (diploid) human genomes, as obtained in earlier studies [22, 24]; these include global *cis* abundance as a universal characteristic of diploid human genomes and the classification and characterization of *cis*-

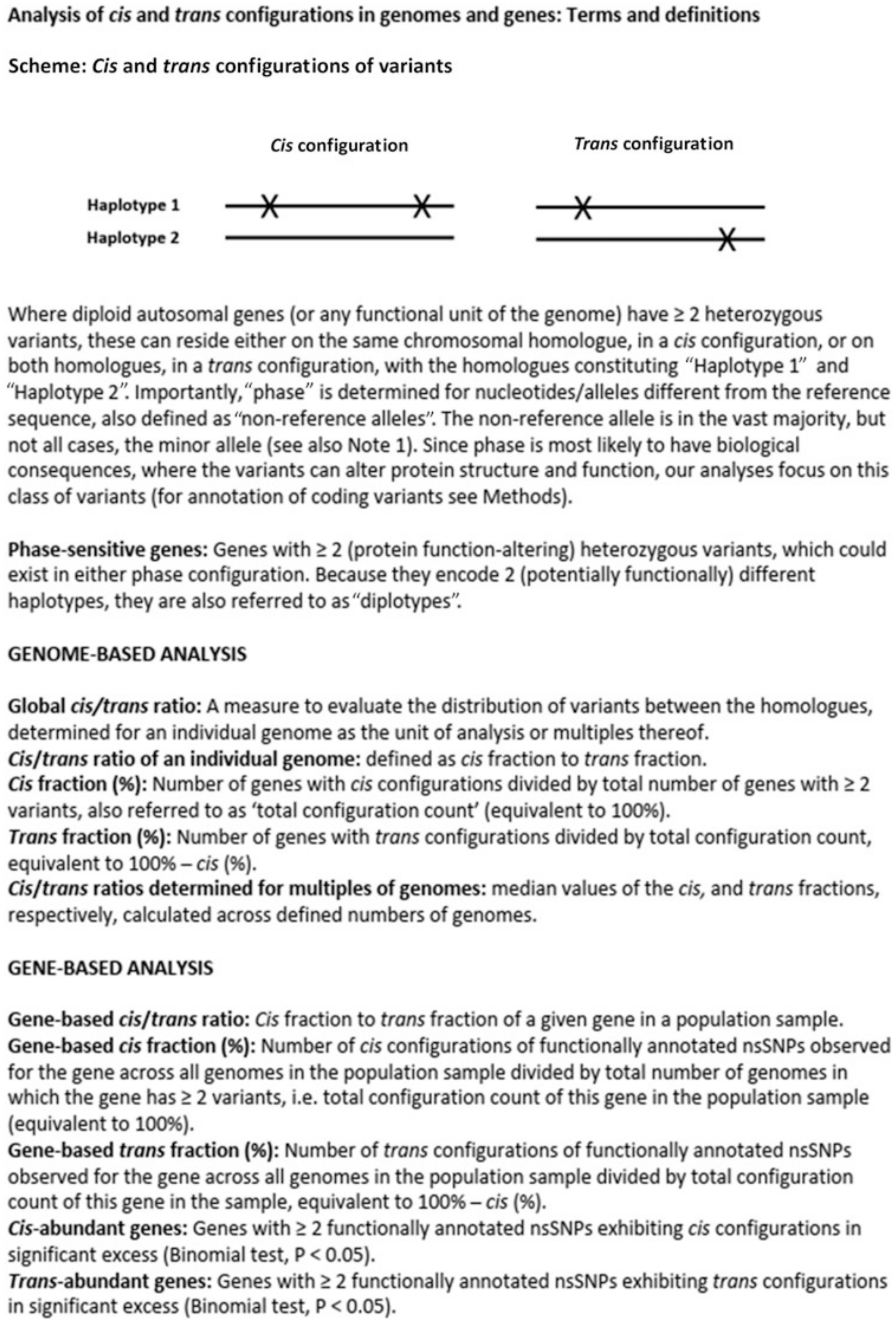
and *trans*-abundant genes. (In this context, we would like to refer the reader to these publications for further detail.) These results have recently been consolidated and expanded using an updated database and advanced analysis tools. Thus, we were able to substantiate the potential functional importance of these two gene classes also at the higher hierarchical level of protein–protein interaction networks (PPI). Altogether, these results were obtained through comprehensive bioinformatic analyses of multiple haplotype-resolved genomes, generated by application of three different experimental approaches, two of which were presented in this series earlier [21, 33]. Here, we present the methods and approaches we have developed and applied to analyze the large numbers of phased genomes, along with lists of (cross-validated) *cis*- and *trans*-abundant genes and information on their functional annotation. This should enable researchers to conduct independent investigations within different and much larger data sets, thus substantiating results and refining approaches that reveal more about the "World of 2" that is the diploid human genome.

2 Key Results

2.1 Genome-Based Analysis of Cis and Trans Configurations of Coding Variants

2.1.1 Global Abundance of Cis Configurations

To test our hypothesis of a nonrandom distribution of protein-changing variants in 1092 genomes (Materials), we assessed the *cis/trans* ratios of predicted protein function-altering non-synonymous SNPs (PFA-nsSNPs) across 18,121 autosomal protein-coding genes (primary transcripts) for each of the genomes and derived the median of these ratios (see Fig. 1 and Methods). Indeed, *cis* configurations of PFA-nsSNPs occurred significantly more frequently than *trans* configurations, with a global *cis/trans* ratio of 59.6:40.4 ($P < 3.53E\text{-}21$). Significant *cis* excess existed moreover in each of the four ancestry groups that were contained in the 1092 genomes, with *cis* fractions between 61.2% and 59.5% in EUR, AMR, and EAS ($P < 2.25E\text{-}21$–$1.46E\text{-}17$) and 54.7% in AFR ($P < 1.66E\text{-}14$). The same was true for each of the 14 populations contained in these four ancestry groups. When examining in addition the entirety of nsSNPS and sSNPs existing within the coding sequences, the results were again almost identical. Significant *cis* abundance was strongly corroborated by analysis of the 184 experimentally haplotype-resolved PGP genomes (Materials), with ratios of 60.4:39.6 ($P < 1.66E\text{-}16$) for PFA-nsSNPs, and slightly higher ratios for nsSNPs and sSNPs, respectively. Thus, protein-changing variants, and coding variants as a whole, are not distributed randomly between the two homologues of a gene but occur significantly more frequently on the same homologue, with ~60% of the phase-sensitive genes carrying their coding variants in *cis*.



Analysis of *cis* and *trans* configurations in genomes and genes: Terms and definitions

Scheme: *Cis* and *trans* configurations of variants

Cis configuration
Trans configuration
Haplotype 1
Haplotype 2

Where diploid autosomal genes (or any functional unit of the genome) have ≥ 2 heterozygous variants, these can reside either on the same chromosomal homologue, in a *cis* configuration, or on both homologues, in a *trans* configuration, with the homologues constituting "Haplotype 1" and "Haplotype 2". Importantly, "phase" is determined for nucleotides/alleles different from the reference sequence, also defined as "non-reference alleles". The non-reference allele is in the vast majority, but not all cases, the minor allele (see also Note 1). Since phase is most likely to have biological consequences, where the variants can alter protein structure and function, our analyses focus on this class of variants (for annotation of coding variants see Methods).

Phase-sensitive genes: Genes with ≥ 2 (protein function-altering) heterozygous variants, which could exist in either phase configuration. Because they encode 2 (potentially functionally) different haplotypes, they are also referred to as "diplotypes".

GENOME-BASED ANALYSIS

Global *cis/trans* ratio: A measure to evaluate the distribution of variants between the homologues, determined for an individual genome as the unit of analysis or multiples thereof.
***Cis/trans* ratio of an individual genome:** defined as *cis* fraction to *trans* fraction.
***Cis* fraction (%):** Number of genes with *cis* configurations divided by total number of genes with ≥ 2 variants, also referred to as 'total configuration count' (equivalent to 100%).
***Trans* fraction (%):** Number of genes with *trans* configurations divided by total configuration count, equivalent to 100% – *cis* (%).
***Cis/trans* ratios determined for multiples of genomes:** median values of the *cis*, and *trans* fractions, respectively, calculated across defined numbers of genomes.

GENE-BASED ANALYSIS

Gene-based *cis/trans* ratio: *Cis* fraction to *trans* fraction of a given gene in a population sample.
Gene-based *cis* fraction (%): Number of *cis* configurations of functionally annotated nsSNPs observed for the gene across all genomes in the population sample divided by total number of genomes in which the gene has ≥ 2 variants, i.e. total configuration count of this gene in the population sample (equivalent to 100%).
Gene-based *trans* fraction (%): Number of *trans* configurations of functionally annotated nsSNPs observed for the gene across all genomes in the population sample divided by total configuration count of this gene in the sample, equivalent to 100% – *cis* (%).
***Cis*-abundant genes:** Genes with ≥ 2 functionally annotated nsSNPs exhibiting *cis* configurations in significant excess (Binomial test, P < 0.05).
***Trans*-abundant genes:** Genes with ≥ 2 functionally annotated nsSNPs exhibiting *trans* configurations in significant excess (Binomial test, P < 0.05).

Fig. 1 Toward a "World of 2." Analysis of *cis* and *trans* configurations in genomes and genes: terms and definitions

2.1.2 Significant Cis Abundance in Each Individual Genome

When each of the 1092 genomes was examined individually, 99.7% had more *cis* than *trans* configurations of PFA-nsSNPs (Fig. 2a). Individual *cis* fractions varied within a limited corridor, e.g., between 55.7% and 67.1% in EUR and between 49.7% and 63.1% in AFR. Importantly, every single genome had significantly larger *cis* fractions than would be expected if the variants were distributed randomly between the two homologues of a gene, with P < 2.55E-29–2.30E-10 in EUR, EAS, and AMR and P < 8.83E-23–1.32E-07 in AFR (see derivation of significance values below). The same applied to the individual *cis* fractions of coding nsSNPs (Fig. 2b) and sSNPs, their noticeably lower variance obviously due to the ~three-fold higher number of phase-sensitive genes per genome in these cases. Likewise, each of the 184 experimentally phased PGP genomes showed highly significant *cis* abundance (P < 2.54E-26–7.99E-06), with *cis* fractions for PFA-nsSNPs between 49.4% and 70.8% (Fig. 2c). Significant *cis* abundance was also observed when nsSNPs (Fig. 2d), sSNPs, and all classes of coding variants were analyzed together as they co-occur in many genes. Thus, significant abundance of *cis* configurations of coding variants could represent a universal characteristic of diploid human genomes. As indicated by the decay of *cis* fractions in AFR, this phenomenon most likely reflects a manifestation of LD.

2.1.3 Significant Cis Abundance Driven by Pairs of Coding Variants

Cis/trans ratios actually represent composite ratios. This is because the genes have different numbers of variants, and the probability of variants to occur in a *trans* configuration increases with the number of variants in a gene. Thus, we calculated the *cis/trans* ratios in the 1092 genomes separately for configurations with 2 up to 5 PFA-nsSNPs, which comprised 95.7% of all configurations. The vast majority of these, 66.7%, consisted of pairs of variants of which 65.8% existed in *cis*. About 18.6% of the configurations consisted of combinations of 3 PFA-nsSNPs, with a ~50:50 ratio of *cis* to *trans*. Comparatively small proportions of the configurations, 7.3% and 3.1%, respectively, had 4 and 5 PFA-nsSNPs, which as expected exhibited larger *trans* than *cis* fractions, 53.4% and 60.3%, respectively. Similar results were obtained when the composite *cis/trans* ratios were dissected in each of the four ancestry groups. The same applied to the 184 experimentally phased PGP genomes; again, pairs of PFA-nsSNPs accounted for roughly two-thirds of the configurations of which two-thirds resided in *cis*. Corresponding analyses of the *cis/trans* ratios calculated from the entirety of nsSNPs and sSNPs, as well as of all types of coding variants together, also confirmed these results. Thus, the excess of *cis* configurations is mainly due to genes with pairs of coding variants that are predominantly in *cis*.

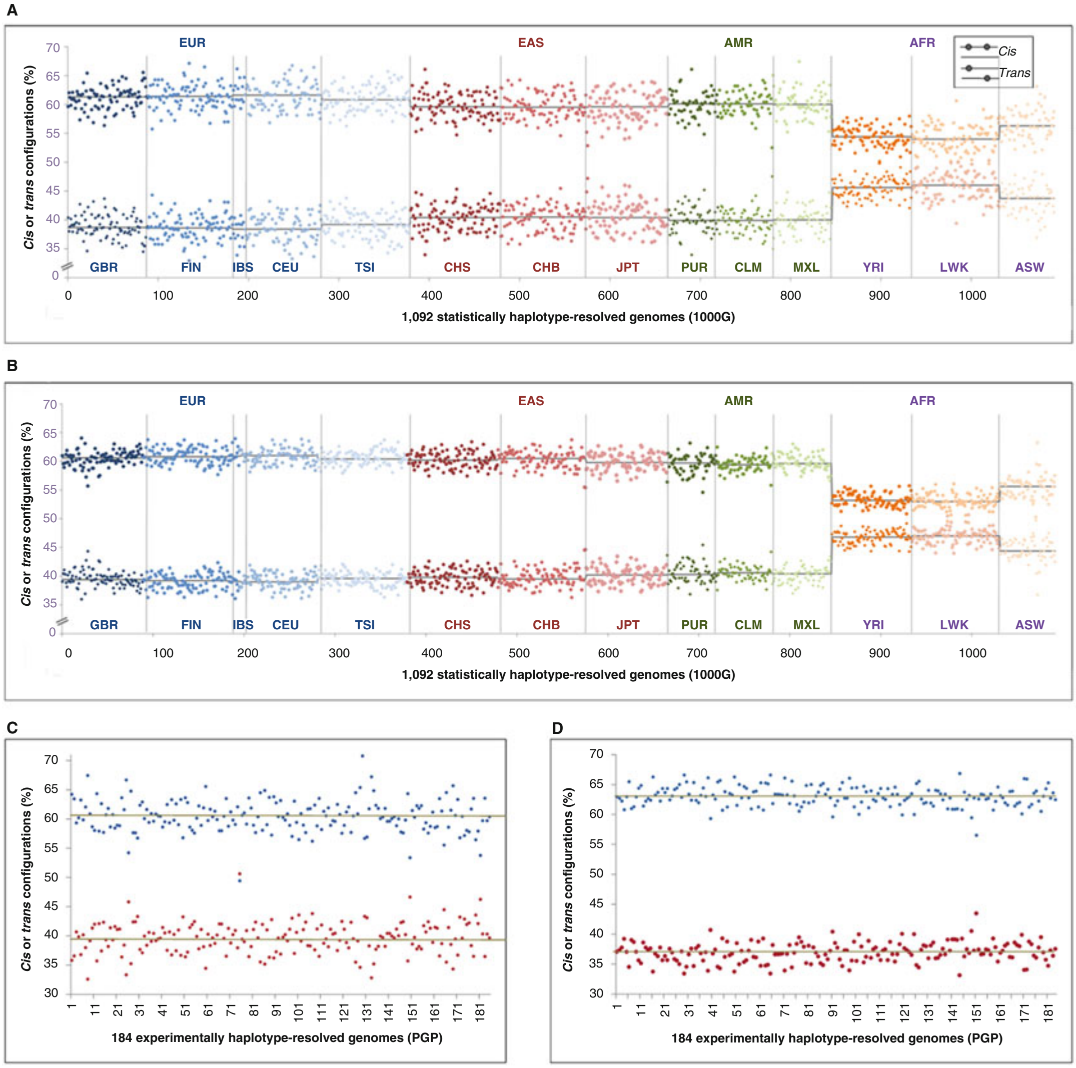


Fig. 2 Individual fractions of *cis* and *trans* configurations of coding variants in 1000 Genomes and PGP. (**a**) Results shown for predicted protein function-altering nsSNPs (PFA-nsSNPs). The fractions of *cis* configurations (%) (number of *cis* configurations divided by total configuration count per genome) (*y*-axis) are presented in the upper half, the complementary *trans* fractions (100% – *cis* fraction (%)) in the lower half. Results are shown for each of the 1092 statistically haplotype-resolved genomes from the 1000 Genomes (1000G) database (*x*-axis), ordered by ancestry group (as indicated on top; color-coded), and further subdivided into different populations as indicated at the bottom (separated by vertical lines); horizontal black lines indicate median values for each of these populations. (**b**) Correspondingly, individual fractions of *cis* and *trans* configurations from the analysis of all nsSNPs. (**c**) Individual fractions of *cis* (blue) and *trans* (red) configurations of PFA-nsSNPs shown for each of the 184 experimentally haplotype-resolved PGP genomes. (**d**) Correspondingly, PGP results from the analysis of all nsSNPs. This figure has previously been published in reference [24]

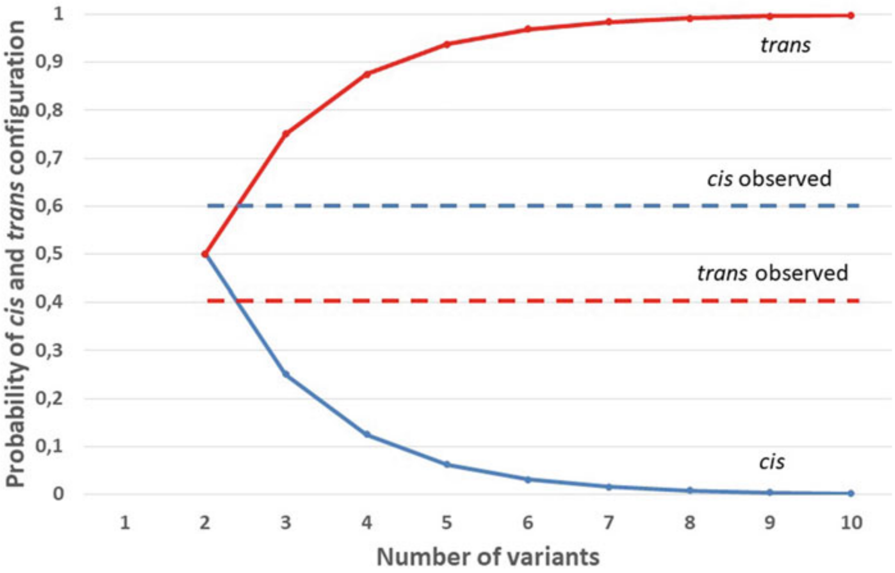


Fig. 3 *Cis* and *trans* configuration probabilities under random assumptions (*y*-axis) in relation to the number of variants (*x*-axis). *Cis* fraction: blue; *trans* fraction: red. Dashed lines refer to the observed composite *cis*/*trans* ratio of 60:40

2.1.4 *Expected Versus Observed* Cis *Fractions*

Importantly, the dissection of the composite *cis/trans* ratios by number of variants enabled a more precise evaluation of the significance of *cis* excess by comparing theoretically expected to observed *cis/trans* ratios. So if the variants in a gene were randomly distributed between the homologues, that is, the chance for each variant to occur on either homologue is equal, then the expected fraction of *cis* configurations is $1/2^{n-1}$, with n being the number of variants (Fig. 3). Accordingly, the expected *cis* fraction for pairs of variants would be 50%, compared to the observed *cis* fractions of 66–70%; for combinations of 3 variants expected 25% versus observed 50–54%; for combinations of 4 variants expected 12.5% versus observed 45–47%, and for 5 variants 6.25% versus 39–41%. This corresponds to a 1.32–1.4- up to 6.24–6.56-fold enrichment of *cis* fractions.

The expected composite probability of *cis* configurations to occur in the genes was modeled using a Bernoulli experiment, resulting in an expected probability of ~0.4 (see Methods). The significance of a *cis/trans* ratio observed in an individual genome was then computed with an exact Binomial test with $P = 0.4$. Thus, even where *cis* fractions were minimally below 50%, which was the case in two of the 1092 genomes and one of the 184 PGP genomes (Fig. 2a, c), they are still significantly larger than would be expected by chance. To assess the significance values for *cis/trans* ratios, which were calculated for defined population sample sets, we derived the median values for both *cis* and *trans* fractions across all genomes. This "median genome" was then treated as an individual genome as described above. Thus, the significance values estimated for global *cis/trans* ratios most likely represent an underestimation.

In order to corroborate the expected composite probability of a *cis* configuration to occur under random conditions, we simulated 1092 phased genomes, assigning to each variant in a protein-coding gene a 50:50 chance to exist on either homologue (Methods). After our simulations proved to be valid, we were able to derive confidently the expected composite *cis* fractions, ~39% for PFA-nsSNPs, ~37% for nsSNPs and sSNPs, respectively, and 33% for combining all types of coding variants. Thus, our simulation studies resulted in expected composite *cis* ratios of approximately 40% and lower versus observed *cis* ratios of ~60% (*see* also Fig. 3).

2.1.5 Pairs of Variants in Cis *more Closely Spaced than Pairs in* Trans

As outlined above, significant *cis* abundance is primarily due to an excess of pairs of PFA-nsSNPs in *cis*. These pairs were found to be much closer together than pairs in *trans*, with inter-mutation genome distances of 1607 bp (median) as opposed to 5125 bp in the 1092 genomes (1570 bp versus 5290 bp in EUR, 1830 versus 4771 bp in AFR) and 2584 versus 5984 bp in PGP. These distances were inversely related to the *cis/trans* ratios; the smaller the distance, the larger the *cis* fraction. To this end, the *cis* and *trans* configurations, e.g., identified in EUR (83,432 *cis*, 39,687 *trans*), were sorted by inter-mutation genome distance, binned per 6000 configurations, and for each bin, an average distance calculated together with its corresponding *cis/trans* ratio. Accordingly, the smallest average distance between the pairs of PFA-nsSNPs in EUR, 11 bp, corresponded to the highest *cis/trans* ratio, 77:23. Average distances of 25, 47 and 86 bp, respectively, were calculated for the adjacent bins and a distance of 1017 bp for the tenth bin, with a corresponding *cis/trans* ratio of 70:30. This ratio declined to 51:49 at an average distance of 54,281 bp and was ~49:51 at the largest distance calculated, 81,099 bp, where a cumulative *cis* fraction of 67.8% was reached (for more information, see Hoehe et al. [24]). Thus, *cis* abundance is largely driven by distance.

2.1.6 Pairs of Variants in Cis *More Frequent than Pairs in* Trans

Furthermore, the pairs of PFA-nsSNPs in *cis* were more frequent than the pairs in *trans*. To this end, we first calculated the average minor allele frequencies (MAFs) for each pair of variants in EUR and AFR. Notably, the PFA-nsSNPs that occurred together had very similar MAFs (as derived from the 1000G database), and combinations of such common PFA-nsSNPs with singletons were very rare. In brief, the mean values of the average MAFs of the pairs of PFA-nsSNPs in *cis* and *trans,* respectively, were 0.18 and 0.16 in EUR and 0.143 and 0.128 in AFR. In the MAF spectra, the average MAFs of the variant pairs in either configuration peaked between 0.1 and 0.25 (with a larger upper tail of *cis* compared to *trans* configurations), a stark contrast to the MAF spectra derived from the entirety of PFA-nsSNPs in the genome, with 56% in EUR and

45% in AFR having a MAF ≤ 0.01. "Taken together, these findings could reflect the result of ancestral admixture as a potential underlying mechanism. Accordingly, the observed, significant *cis* abundance results primarily from pairs of protein function-altering variants and coding variants as a whole that are closely spaced and have therefore been inherited together until present. Thus, they are more common than pairs of coding variants in *trans* which, much farther apart, have been subject to recombination, but also may be due to evolutionary forces other than recombination, such as genetic mutation and positive selection. So pairs of co-occurring protein function-altering variants in *cis*, which are even closer together than other pairs of coding variants in *cis*, may represent ancestral signals of potential functional significance and mark small ancestry segments in the 'mosaic that is the human genome' [34]" (verbatim taken from [24]).

2.2 Gene-Based Analysis of Cis and Trans Configurations of Coding Variants

2.2.1 A Global Set of Phase-Sensitive Genes

To investigate the potential functional implications of *cis* abundance, we moved from whole genome analysis to the genes underlying this phenomenon. A first inspection showed that the numbers of phase-sensitive genes with ≥2 PFA-nsSNPs, where phase most likely has an impact, were very similar between the genomes. The same was true for their *cis* and *trans* forms (for details, see Hoehe et al. [24]). Thus, the questions arose: Is there a common, shared set of phase-sensitive genes that underlie the phenomenon of *cis* abundance? Furthermore, are there common subsets of genes that preferentially have *cis*, or *trans* configurations? If so, which functional gene classes are concerned?

In fact, we were able to identify a set of 2402 phase-sensitive genes, which were common to all ancestry groups (Table S1a). To this end, we intersected the genes with ≥2 PFA-nsSNPs in each of these groups, 4000 in EUR, 3357 in EAS, 4005 in AMR, and 5217 in AFR, any two of which shared 80–87%. This global set of phase-sensitive genes, which we also refer to as a common diplotypic exome, showed a highly significant overrepresentation of 94 pathways (P < 1.24E-48–0.01; Table S1b) (see also Methods for use of an updated database and analysis tools as described by Kamburov and Herwig [35]). These pathways included, for instance, "olfactory transduction" (P = 1.24E-48), "sensory perception" (P = 9.13E-40), "extracellular matrix organization" (1.39E-09), metabolic processes such as "xenobiotics metabolism" (P = 1.19E-06), "androgen and estrogen biosynthesis" (P = 2.16E-05), "C21-steroid hormone biosynthesis" (P = 3.19E-05), and infection/immune response pathways such as "antigen processing and presentation" (P = 7.23E-07), "graft-versus-host disease" (P = 5.90E-06), and "interferon gamma signaling" (P = 4.26E-05). These results were complemented by a highly significant enrichment of manifold GO terms (120 GO categories with P < 9.21E-37–0.001; Table S1c). These included, for instance,

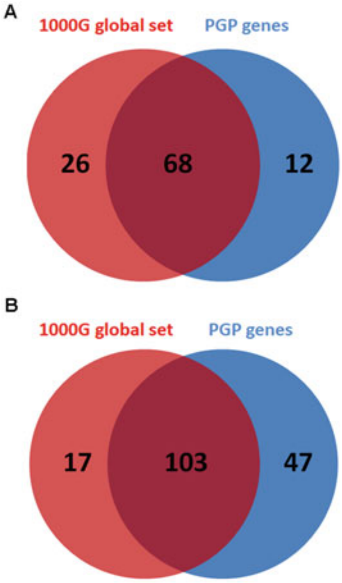


Fig. 4 Pathways and GO terms shared between the 1000G global set and PGP set of phase-sensitive genes. (**a**) VENN diagram showing the overlap of pathways, which were significantly enriched ($P < 0.01$) in the global set of 2402 phase-sensitive genes (1000G) (red circle) and the set of 1627 phase-sensitive genes ($P < 0.01$), which PGP shared with 1000G (blue circle). (**b**) VENN diagram showing the overlap of GO terms, which were significantly enriched ($P < 0.001$) in the global set of 2402 phase-sensitive genes (1000G) (red circle) and the set of 1627 phase-sensitive genes ($P < 0.001$), which PGP shared with 1000G (blue circle). The overlap of the sets can be quantified with Sorensen's similarity index, $S = \frac{2ab}{a+b}$, where *a* is the number of genes in the first set, *b* the number of genes in the second set, and *ab* the number of genes shared by the two sets. This results in $S = 0.78$ for the similarity between pathways (~78%) and $S = 0.76$ for the similarity between GO terms (~76%)

"detection of (chemical) stimulus" (P = 3.61E-43), "transmembrane signaling receptor activity" (P = 4.36E-28), "olfactory receptor activity" (P = 3.29E-46) and "G protein-coupled receptor activity" (P = 9.42E-23), "extracellular structure organization" (P = 5.29E-10), and "nervous system process" (P = 1.64E-20).

A fraction of 68% of the phase-sensitive genes (1627) contained in this global set were shared by the 184 experimentally phased PGP genomes (at about one-sixth of the population size). This set of 1627 genes shared 78% of the overrepresented pathways (P < 2.50E-42–0.01) (Fig. 4a; Table S1d) and 76% of the overrepresented GO terms (P < 1.18E-41–0.001) (Fig. 4b; Table S1e) with the global set, allowing extraction of very similar functional content. Thus, there exists a common diplotypic exome encoding two potentially functionally different homologues, which may modulate cell–environment interactions, cell–cell communication, immune response, metabolism and biosynthesis, and the development of disease.

Remarkably, this common diplotypic exome was significantly enriched with genes thought to be under balancing selection (BS), as well as with monoallelically expressed (MAE) compared to biallelically expressed (BAE) genes [36]. Like MAE, this gene set was significantly enriched with genes with evidence of human–chimpanzee trans-species polymorphisms or haplotypes (TSPs) and with genes harboring at least one ancient protein-coding SNP or haplotype predating the human-Neanderthal split (HNS), underscoring its old allelic age. The overrepresentation of evolutionarily significant gene sets was even more pronounced in the 1627 phase-sensitive genes cross-validated by PGP (see also Hoehe et al. [24]). This suggests that this common set of diplotypes may play an important role in preserving functional flexibility in evolutionary processes. Having confirmed existence of a common set of diplotypic genes underlying *cis* abundance, can we further distinguish within this set two groups of genes that preferentially have either *cis* or *trans* configurations of PFA-nsSNPs?

2.2.2 Classification of Cis- and Trans-Abundant Genes

Analyzing the gene-based *cis/trans* ratios (see Fig. 1) for each of the 2402 phase-sensitive genes across the 1092 genomes, we were in fact able to identify a subset of 1227 genes that showed *cis* configurations significantly more often ($P < 1.05E\text{-}63$–0.048) and were therefore defined as "*cis*-abundant genes" (Table S2a). A subset of 786 genes had significantly more frequently *trans* configurations ($P < 1.78E\text{-}15$–0.049) and were therefore defined as "*trans*-abundant genes" (Table S2b). Thus, 2013 of the 2402 autosomal genes with ≥2 PFA-nsSNPs (84%) could be classified into the two major categories *cis*- and *trans*-abundant genes, while the remaining 385 genes had almost the same proportions of both configurations.

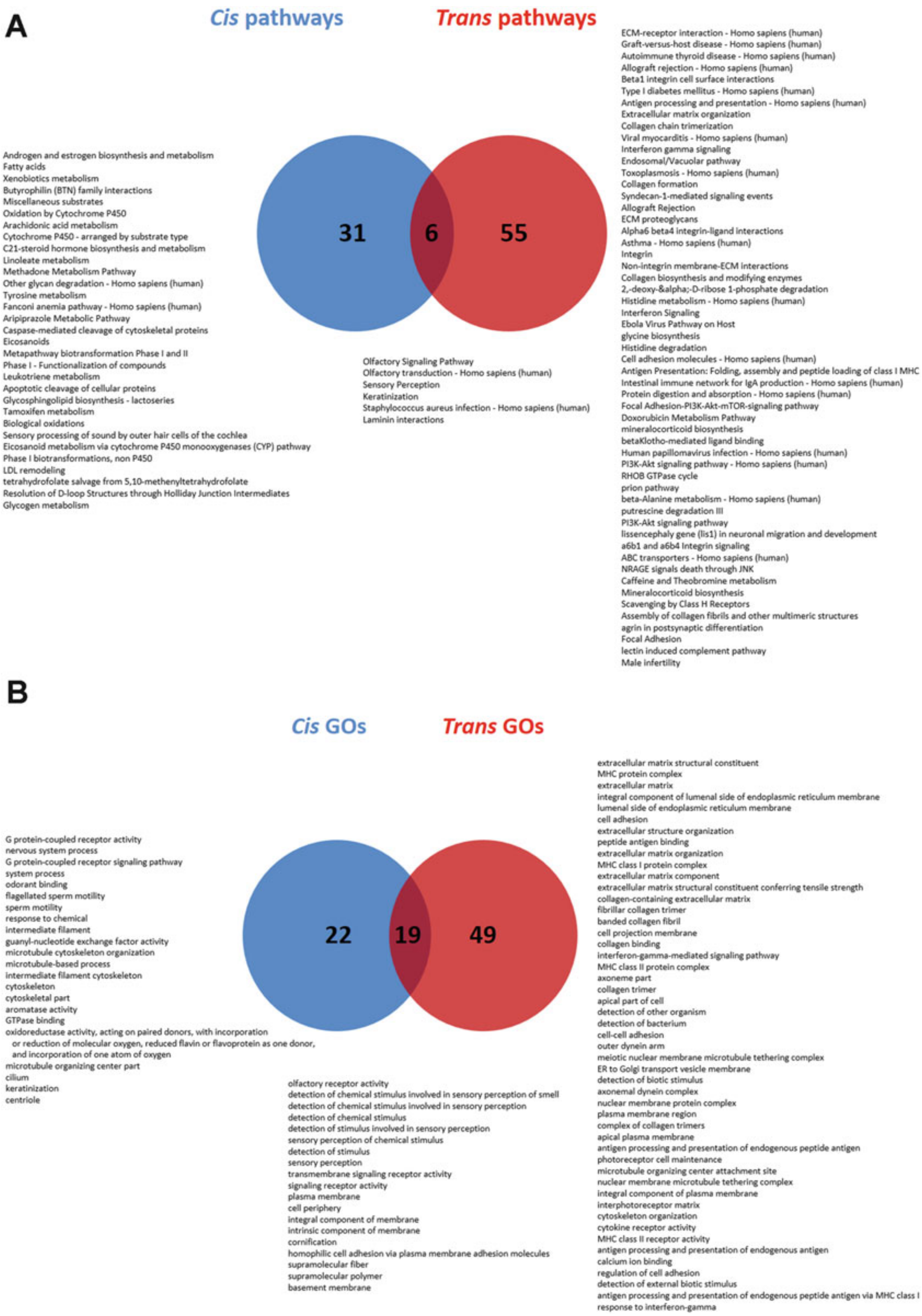
This classification of *cis*- and *trans*-abundant genes was derived from the global set of 2402 phase-sensitive genes, which, viewed closely, represents 46–72% of all autosomal protein-coding genes in each of the intersected ancestry groups. To test whether this classification also applies to each of these populations as a whole, we examined the ancestry groups individually. Thus, 1173 significantly *cis*- and 670 *trans*-abundant genes were observed in EUR, 966 and 590, respectively, in EAS, 981 and 497 in AMR, and 1265 and 817 in AFR, accounting for 78–88% of all autosomal phase-sensitive genes. Subsequent analysis of the 184 experimentally haplotype-resolved PGP genomes further validated this classification; accordingly, 83.5% of the phase-sensitive genes in this set of genomes were grouped into 778 *cis*- and 436 *trans*-abundant genes. Taken together, 78–88% of all phase-sensitive genes were classifiable into *cis*- and *trans*-abundant genes in each of the sample sets examined, underscoring the importance of these two gene classes as major categories of variable autosomal genes. Obviously, the group of *cis*-abundant genes was always larger than that of *trans*-abundant genes, with ratios between 1.55:1 and ~2:1. Thus, global *cis* abundance is the net result of these two groups. *Cis*- and *trans*-abundant genes were found distributed in different proportions across all autosomes, with chromosomes 14, 20, and 22 showing the largest fractions of genes with *cis* configurations (up to 73.5%) and chromosomes 6, 8, and 10 the largest fractions of genes with *trans* configurations (up to 51.2%).

In this context, the question arises as to whether, or to what extent, the "configuration type" of a gene represents a constant characteristic. Thus, we examined whether the "configuration type" of genes as determined in the 1092 genomes was the same in the set of PGP genomes. An expanded analysis (*see* **Note 2**) showed that 8.7% of the *cis*- and 12.9% of the *trans*-abundant genes in the 1092 genomes had a different configuration type in PGP. Thus, *cis* and *trans* abundance represents a constant characteristic in approximately 90% of the autosomal genes.

Finally, we examined the relationship between configuration status and gene length, motivated by the marked differences in inter-mutation genome distance observed between *cis* and *trans* configurations (see above). While *trans*-abundant genes are on average longer than *cis*-abundant genes (primary transcript lengths 35,969 versus 25,859 bp, P = 3.9E-08; protein lengths 1012 versus 875 amino acids; P = 0.001), the overall correlations between gene-based *trans* fractions and primary transcript (c = 0.07) as well as protein length (c = 0.01) were not significant. This can be illustrated at the example of the relatively short HLA genes, which were classified as members of a predominantly *trans*-abundant gene family. Furthermore, although there is an overall tendency for genes to reside in *trans* with increasing numbers of variants, the opposite was true, e.g., for the highly diverse OR genes, which were largely *cis*-abundant. In sum, configuration status overall is not significantly influenced by either gene length or the number of coding variants.

2.2.3 Functional Enrichment Differs Between Cis- and Trans-Abundant Genes

We then analyzed the functional information contained in these two gene classes for further characterization. Using the updated content of the ConsensusPathDB [35], we were able to confirm the recently reported enrichment of pathways and GO categories [24] that suggested different functional roles of these two gene classes. Thus, *cis*-abundant genes were found to enrich 37 pathways that were annotated in different pathway databases (P < 3.96E-49–0.01), while *trans*-abundant genes enriched 61 pathways (P < 3.16E-08–0.01); both gene classes showed only little overlap in six pathways and were differentially enriched for 93% of the pathways (Table S2c, d; Fig. 5a). Furthermore, the two gene classes enriched different GO terms, *cis*-abundant genes 41 GO terms (P < 1.15E-46–0.001) and *trans*-abundant genes 68 GO terms (P < 1.21E-10–0.001), with an overlap of 19 terms, indicating fairly distinct functional classes (Table S2e, f; Fig. 5b).



Androgen and estrogen biosynthesis and metabolism
Fatty acids
Xenobiotics metabolism
Butyrophilin (BTN) family interactions
Miscellaneous substrates
Oxidation by Cytochrome P450
Arachidonic acid metabolism
Cytochrome P450 - arranged by substrate type
C21-steroid hormone biosynthesis and metabolism
Linoleate metabolism
Methadone Metabolism Pathway
Other glycan degradation - Homo sapiens (human)
Tyrosine metabolism
Fanconi anemia pathway - Homo sapiens (human)
Aripiprazole Metabolic Pathway
Caspase-mediated cleavage of cytoskeletal proteins
Eicosanoids
Metapathway biotransformation Phase I and II
Phase I - Functionalization of compounds
Leukotriene metabolism
Apoptotic cleavage of cellular proteins
Glycosphingolipid biosynthesis - lactoseries
Tamoxifen metabolism
Biological oxidations
Sensory processing of sound by outer hair cells of the cochlea
Eicosanoid metabolism via cytochrome P450 monooxygenases (CYP) pathway
Phase I biotransformations, non P450
LDL remodeling
tetrahydrofolate salvage from 5,10-methenyltetrahydrofolate
Resolution of D-loop Structures through Holliday Junction Intermediates
Glycogen metabolism

Olfactory Signaling Pathway
Olfactory transduction - Homo sapiens (human)
Sensory Perception
Keratinization
Staphylococcus aureus infection - Homo sapiens (human)
Laminin interactions

ECM-receptor interaction - Homo sapiens (human)
Graft-versus-host disease - Homo sapiens (human)
Autoimmune thyroid disease - Homo sapiens (human)
Allograft rejection - Homo sapiens (human)
Beta1 integrin cell surface interactions
Type I diabetes mellitus - Homo sapiens (human)
Antigen processing and presentation - Homo sapiens (human)
Extracellular matrix organization
Collagen chain trimerization
Viral myocarditis - Homo sapiens (human)
Interferon gamma signaling
Endosomal/Vacuolar pathway
Toxoplasmosis - Homo sapiens (human)
Collagen formation
Syndecan-1-mediated signaling events
Allograft Rejection
ECM proteoglycans
Alpha6 beta4 integrin-ligand interactions
Asthma - Homo sapiens (human)
Integrin
Non-integrin membrane-ECM interactions
Collagen biosynthesis and modifying enzymes
2,-deoxy-&alpha;-D-ribose 1-phosphate degradation
Histidine metabolism - Homo sapiens (human)
Interferon Signaling
Ebola Virus Pathway on Host
glycine biosynthesis
Histidine degradation
Cell adhesion molecules - Homo sapiens (human)
Antigen Presentation: Folding, assembly and peptide loading of class I MHC
Intestinal immune network for IgA production - Homo sapiens (human)
Protein digestion and absorption - Homo sapiens (human)
Focal Adhesion-PI3K-Akt-mTOR-signaling pathway
Doxorubicin Metabolism Pathway
mineralocorticoid biosynthesis
betaKlotho-mediated ligand binding
Human papillomavirus infection - Homo sapiens (human)
PI3K-Akt signaling pathway - Homo sapiens (human)
RHOB GTPase cycle
prion pathway
beta-Alanine metabolism - Homo sapiens (human)
putrescine degradation III
PI3K-Akt signaling pathway
lissencephaly gene (lis1) in neuronal migration and development
a6b1 and a6b4 Integrin signaling
ABC transporters - Homo sapiens (human)
NRAGE signals death through JNK
Caffeine and Theobromine metabolism
Mineralocorticoid biosynthesis
Scavenging by Class H Receptors
Assembly of collagen fibrils and other multimeric structures
agrin in postsynaptic differentiation
Focal Adhesion
lectin induced complement pathway
Male infertility

G protein-coupled receptor activity
nervous system process
G protein-coupled receptor signaling pathway
system process
odorant binding
flagellated sperm motility
sperm motility
response to chemical
intermediate filament
guanyl-nucleotide exchange factor activity
microtubule cytoskeleton organization
microtubule-based process
intermediate filament cytoskeleton
cytoskeleton
cytoskeletal part
aromatase activity
GTPase binding
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen
microtubule organizing center part
cilium
keratinization
centriole

olfactory receptor activity
detection of chemical stimulus involved in sensory perception of smell
detection of chemical stimulus involved in sensory perception
detection of chemical stimulus
detection of stimulus involved in sensory perception
sensory perception of chemical stimulus
detection of stimulus
sensory perception
transmembrane signaling receptor activity
signaling receptor activity
plasma membrane
cell periphery
integral component of membrane
intrinsic component of membrane
cornification
homophilic cell adhesion via plasma membrane adhesion molecules
supramolecular fiber
supramolecular polymer
basement membrane

extracellular matrix structural constituent
MHC protein complex
extracellular matrix
integral component of lumenal side of endoplasmic reticulum membrane
lumenal side of endoplasmic reticulum membrane
cell adhesion
extracellular structure organization
peptide antigen binding
extracellular matrix organization
MHC class I protein complex
extracellular matrix component
extracellular matrix structural constituent conferring tensile strength
collagen-containing extracellular matrix
fibrillar collagen trimer
banded collagen fibril
cell projection membrane
collagen binding
interferon-gamma-mediated signaling pathway
MHC class II protein complex
axoneme part
collagen trimer
apical part of cell
detection of other organism
detection of bacterium
cell-cell adhesion
outer dynein arm
meiotic nuclear membrane microtubule tethering complex
ER to Golgi transport vesicle membrane
detection of biotic stimulus
axonemal dynein complex
nuclear membrane protein complex
plasma membrane region
complex of collagen trimers
apical plasma membrane
antigen processing and presentation of endogenous peptide antigen
photoreceptor cell maintenance
microtubule organizing center attachment site
nuclear membrane microtubule tethering complex
integral component of plasma membrane
interphotoreceptor matrix
cytoskeleton organization
cytokine receptor activity
MHC class II receptor activity
antigen processing and presentation of endogenous antigen
calcium ion binding
regulation of cell adhesion
detection of external biotic stimulus
antigen processing and presentation of endogenous peptide antigen via MHC class I
response to interferon-gamma

Fig. 5 Overrepresentation of pathways and GO terms in *cis*- and *trans*-abundant genes. (**a**) VENN diagram showing the differential enrichment of pathways, which are overrepresented ($P < 0.01$) in either *cis*- (left) or *trans*-abundant genes (right); pathways listed in between, the numbers of which are indicated in the overlap, are enriched in both gene categories. (**b**) VENN diagram illustrating the differential enrichment of GO terms, which are overrepresented ($P < 0.001$) in either *cis*- (left) or *trans*-abundant genes (right); GO terms listed in between are enriched in both gene categories

Whereas *cis*-abundant genes were overrepresented mostly in metabolic pathways, e.g., "xenobiotics metabolism" (1.03E-04) including "cytochrome P450-mediated oxidation" (6.28E-04) and other phase 1-related processes, *trans*-abundant genes were enriched for numerous immune response-related processes and diseases as well as autoimmune diseases, viral and infectious diseases, cell surface interactions, ECM–receptor interaction, signaling pathways, and drug transport (Table S2c, d). The two gene classes were furthermore differentially enriched for 79% of the GO terms, evaluated separately for their taxonomies: *cis*-abundant genes, for instance, for GPCR activity and signaling, odorant binding, and activities involved in the metabolism of substrates; *trans*-abundant genes, for instance, for MHC proteins/receptors, antigen binding, ECM structural constituent and organization, and cell adhesion (Table S2e, f). Thus, *cis* and *trans*-abundant genes may be differentially involved in gene functions and pathways, which could indicate different mechanisms for exerting gene functions.

2.2.4 Different Distribution Patterns of Variants in Cross-Validated Cis- *and* Trans-*Abundant Genes*

In the next step, we comparatively evaluated those genes that had been classified as *cis*- or *trans*-abundant in both the 1092 statistically phased and the 184 experimentally phased PGP genomes. Intersection of the corresponding data sets resulted in overlaps of 322 *cis*-abundant and 153 *trans*-abundant genes. Figure 6 presents examples of these cross-validated genes. These examples highlight that the two different methodological approaches lead to very similar results. Furthermore, they illustrate that autosomal genes with ≥2 PFA-nsSNPs can have *cis* or *trans* configurations in highly significant excess. Notable fractions of *cis*- and *trans*-abundant genes had solely *cis* or *trans* configurations, respectively, such as *PON2* and *OR8D2* with *cis* fractions of 100% and *MT1A* with a *trans* fraction of 100% (Fig. 6).

Examples of *cis*-abundant genes (Fig. 6) include members of pathways and GOs that were found to be significantly overrepresented in this gene category, such as *CYP2D6*, *GSTZ1*, *MAP4,* and *SULT1C3* involved in xenobiotics metabolism or *OR8D2* and *OR2M7* involved in olfactory transduction. In addition, examples include disease genes such as *BRCA1*, *LRRK2*, *ADAM33*, *ADD1*, or *PON2*. *Trans*-abundant genes (Fig. 6) include genes involved in immune response; immune and autoimmune diseases; viral and infectious diseases, such as *HLA-E*, *KIR3DL2*, *TPO,* and *DDX58*; and genes involved in cancer such as *MTA1* and *NAT2.*

Further inspection of some of these cross-validated genes revealed different, potentially functional relevant distribution patterns of PFA-nsSNPs in *cis*- and *trans*-abundant genes. Thus, *cis*-abundant genes appeared to be characterized by presence of a closely spaced pair of PFA-nsSNPs that occurred in numerous individuals, which we termed a "major configuration." In contrast, two or more less frequent and more distant pairs of PFA-nsSNPs

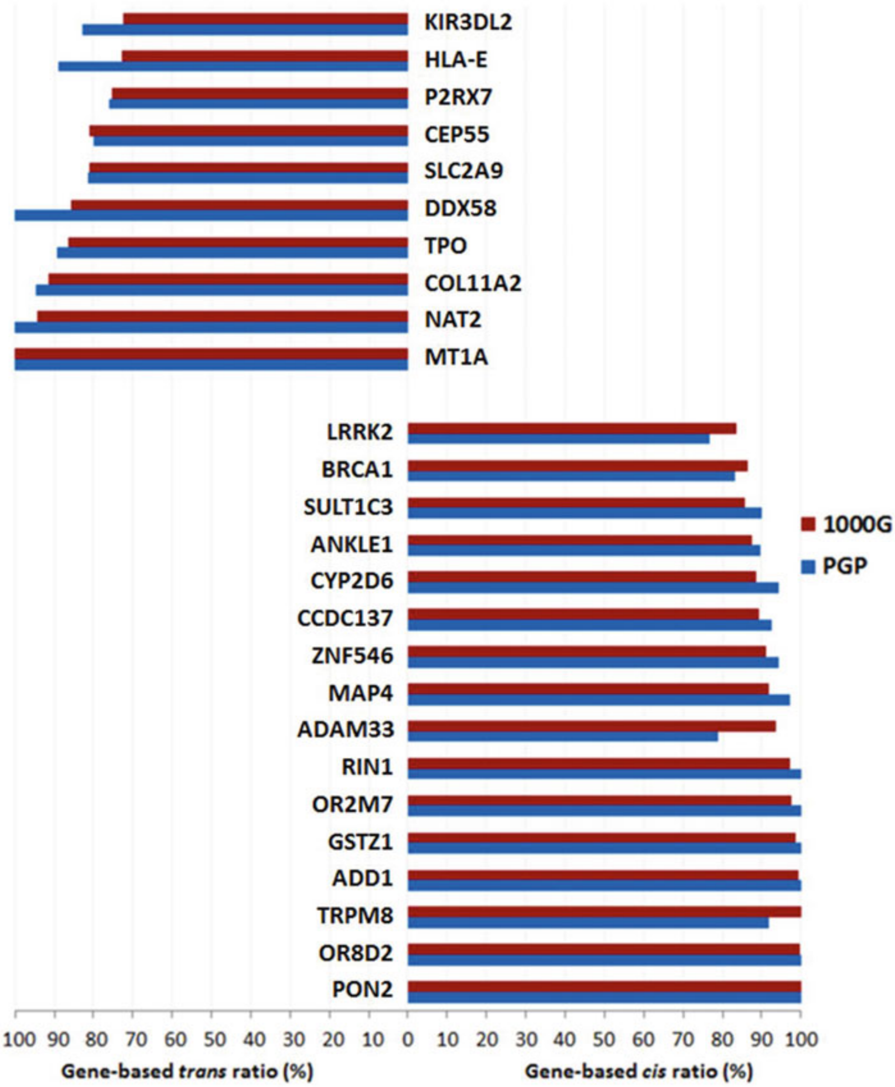


Fig. 6 Examples of *cis*- and *trans*-abundant genes. Genes were selected from the cross-validated sets of 322 *cis*- and 153 *trans*-abundant genes, which were shared by both 1000 Genomes (1000G) and PGP. Top left *trans*-abundant genes, lower right *cis*-abundant genes; blue bars indicate the gene-based *cis*- and *trans* fractions (%) derived from the 184 experimentally phased PGP genomes, and red bars the corresponding fractions derived from the 1092 statistically phased genomes (1000G). This figure has previously been published in reference [24]

were observed in *trans*-abundant genes. Subsequent systematic analyses (**Note 3**) in both the cross-validated sets and the larger sets of 1227 *cis*- and 786 *trans*-abundant genes derived from the 1092 genomes confirmed that "major configurations" were typical for substantial fractions of *cis*-abundant genes. Thus, over 40% of this gene class had a major configuration that accounted for >90 up to 100% of the total number of *cis* configurations scored for a gene in PGP, while less than 13% of the cross-validated *trans*-abundant genes in PGP showed such a configuration. Correspondingly, 31.2% of the *cis*-abundant genes in the 1092 genomes had such a major configuration, in contrast to less than 13% of the *trans*-abundant genes. Thus, significant *cis* abundance seems to be related to a substantial fraction of *cis*-abundant genes that have highly frequent pairs of closely spaced, protein function-altering variants co-occurring on the same homologue (**Note 4**). In addition, the different distribution patterns in *cis*- and *trans*-abundant genes could indicate further functional differences between the two gene classes.

2.2.5 Cis- and Trans-Abundant Genes Affect Different Parts of the Human Interactome

Potential functional differences between *cis*- and *trans*-abundant genes can be further investigated at the level of their gene products, the proteins. Because proteins interact with other proteins in networks, it is through these interactions that they perform their functions. Therefore, as a complement to pathway and GO enrichment, we wanted to investigate whether *cis*- and *trans*-abundant genes map to different parts of the human interactome. To this end, we initially weighted the proteins according to their probability of belonging to *cis*- and *trans*-abundant gene classes, and we then propagated these weights through a large protein–protein interaction network integrated from different resources (Methods). Such network propagation results in subgraphs that agglomerate high *cis* or *trans* load, and these subgraphs, or modules, represent parts of the interactome that are affected by either *cis*- or *trans*-abundant genes.

We performed network propagation with the P-value scores derived from the Binomial test for *cis*- and *trans*-abundance, respectively (Methods). We used the program NetCore to infer a protein–protein interaction network consisting of 10,560 genes and 139,267 interactions integrated from 18 different resources (Methods). This network proved to be a good approximation to the human interactome and is specifically suited for network inferences [37]. Network propagation of *cis*-abundant genes resulted in 20 modules (subgraphs with high *cis* load; Table S3a) with a total of 112 genes (Fig. 7a) and *trans*-abundant gene network propagation resulted in nine modules with a total of 110 genes (Table S3b; Fig. 7b). The largest modules, of size 55 (*cis*) and 91 (*trans*), are shown in Fig. 7c, d. It is evident that the *trans*-abundant genes are more densely connected than the *cis*-abundant genes, resulting not only in a larger major module but also in a lesser number of smaller modules. This is because the underlying protein–protein interaction (PPI) network tends to reflect signaling events that are dominant in the *trans*-abundant genes such as immune-related signaling pathways, compared to the *cis*-abundant genes that were considerably enriched for metabolic reactions.

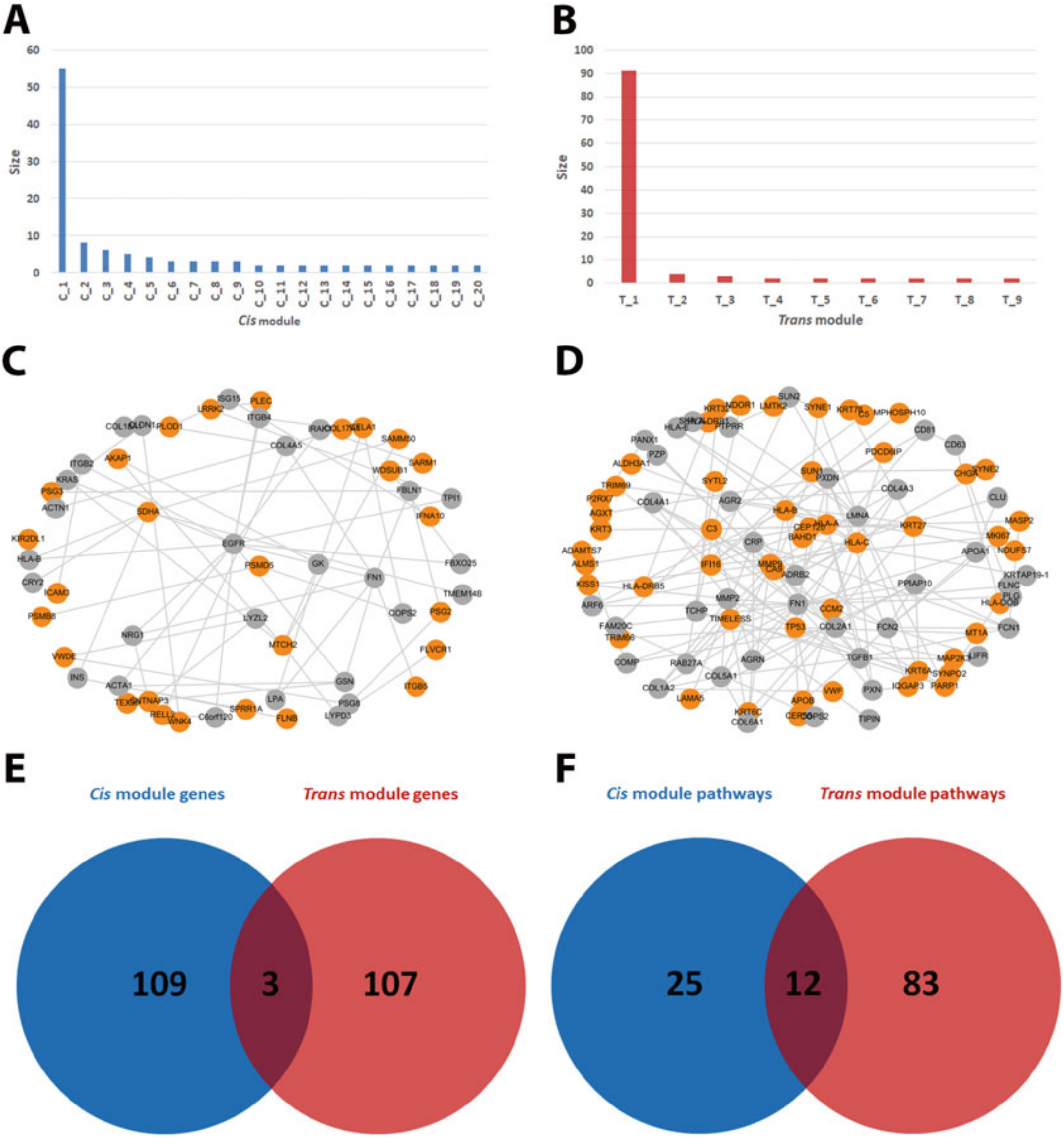


Fig. 7 Network propagation results with *cis*- and *trans*-abundant genes. (**a**) Histogram of computed *cis*-modules; *x*-axis: module Id; *y*-axis: size of module. (**b**) Histogram of computed *trans*-modules; *x*-axis: module Id; *y*-axis: size of module. (**c**) Largest *cis*-module of size 55. Orange nodes correspond to seed genes that were among the 100 most significant *cis*-abundant genes, while genes that were inferred during network propagation are colored in gray. (**d**) Largest *trans*-module of size 91. Orange nodes correspond to seed genes that were among the 100 most significant *trans*-abundant genes, while genes that were inferred during network propagation are colored in gray. (**e**) VENN diagram showing the overlap of *cis*-module genes (112 genes, blue circle) and *trans*-module genes (110 genes, red circle). Similarity index $S = 0.027$. (**f**) VENN diagram showing the overlap of highly significantly enriched pathways ($P < 0.001$) with *cis*-module genes (37 pathways, blue circle) and *trans*-module genes (95 pathways, red circle). Similarity index $S = 0.18$

In addition to the functional enrichment results described above (Sect. 2.2.3), we can show that also at the interactome level, *cis*- and *trans*-abundant modules are very distinct both in gene content (99% different genes; Fig. 7e) as well as pathway enrichment of the inferred module genes (90% different enriched pathways; Fig. 7f). In particular, *trans*-abundant module genes enrich more than 3x as many pathways than *cis*-abundant module genes, in particular immune-related processes such as "allograft rejection" (P = 4.63E-11), the "complement system" (P = 1.78E-08), or "antigen processing and presentation" (P = 1.24E-06) (Table S3c, d). Thus, this additional functional (interaction) layer highlights the fact that the two gene classes map to distinct parts of the interactome and that in particular, *trans*-abundant genes are connected in network modules, with multiple functional consequences for human signaling pathways.

3 Materials

3.1 Data Sets and Phasing Quality

3.1.1 Haplotype Data from 1092 Genomes Generated by the 1000 Genomes (1000G) Consortium

These genomes [38] were generated by routine second-generation sequencing techniques, statistically haplotype-resolved and originally downloadable from ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis results/shapeit2 phased haplotypes/. The total set of 1092 genomes consisted of four different ancestry groups, European (EUR), n = 379; East Asian (EAS), n = 286; American (AMR), n = 181; and African (AFR), n = 246, which in turn consisted of a total of 14 different populations (see also Fig. 2). Thus, *cis* and *trans* configurations were evaluated comparatively in the total set and in each of the ancestry groups as well as the different populations.

Phasing Quality Phased data were available for all 1092 genomes, with "no call" rates between 2.1 and 6% and routine use of imputation in the case of missing data. Concerning the accuracy of the inferred haplotypes, a phasing switch error every 300–400 kb on average has been estimated [38]. Importantly, exome data had a very high coverage due to the addition of deep (50–100×) exome sequence and dense SNP genotype data, enhancing accuracy (see Abecasis et al. [38]).

3.1.2 Data from 184 Experimentally Phased Genomes from the Personal Genome Project (PGP)

The haplotype data from an unprecedented set of 184 experimentally haplotype-resolved genomes from the Personal Genome Project (PGP) database [39] were generated by application of the Long-Fragment Read (LFR) technology [16] and generously provided by Brock A. Peters and Radoje Drmanac, Complete Genomics Inc. & BGI-Shenzen. The individuals were ascertained as part of PGP [23, 39]. They gave full consent to have their genotypic and detailed phenotypic data as well as self-reported ethnicity made freely and publicly available. The documents were reviewed and signed by each participant and can be found at http://www.personalgenomes.org/harvard/signup.

Phasing Quality The LFR technology enabled >98% of all heterozygous SNPs to be placed into contigs with an average N50 of 800 kb [23]. Phasing error rates of this technology were shown to be exceedingly low [16], with average short and long switch error rates (SERs) of 0.00068 and 0.00051, respectively. These SERs were obtained from the analysis of overlapping blocks between the (independent) replicate samples from 35 individual genomes, with ~86% of the overlapping blocks being completely error-free [23], and were much lower than those determined for other experimental and computational phasing approaches [18]. Importantly, a contig filter was applied to the analyses in our study to ensure that phase configurations were only determined for those coding variants that were contained within the same contig. For additional analyses of phasing quality, see Subheading 4.1.

3.2 Gene Sets

(i) A set of 226 genes reported to evolve under balancing selection (BS); (ii) a set of 60 genes with evidence of human–chimpanzee trans-species polymorphisms or haplotypes (TSPs); (iii) a set of 104 genes harboring at least one ancient protein-coding SNP or haplotype shared between humans and Neanderthals (HNS). These gene sets had been described and used for analysis by Savova et al. [36], which we used as the source.

4 Methods

4.1 Evaluation of Phasing Quality of Experimentally Haplotype-Resolved Genome Data

We performed additional analyses to control that the level of polymorphism within the genes does not affect the phasing accuracy to an extent that could bias the *cis/trans* ratio results. To this end, we identified the (heterozygous) positions of switch errors from the 35 PGP samples with replicates [23]. Then we mapped these positions onto the subsets of those *cis*- and *tran*s-abundant genes that are difficult to phase, that is, where switch errors are likely to occur. Thus, we extracted from the total of *cis*- and *trans*-abundant genes those genes that were among the top and bottom 10% with the highest and lowest density of PFA-nsSNPs. Where we identified switch errors within these genes, we checked a possible influence on the configuration type. For the reader's information, we identified switch errors in 6.9% of the *cis*- and *trans*-abundant genes with the highest PFA-nsSNP density; these were found to change configuration type in one case, that is, in ~1% of these genes. Examining the extreme group of *cis*- and *trans*-abundant genes with the lowest PFA-nsSNP density resulted in switch errors affecting 8.9% of these genes, also changing configuration type in one case. Examining the genes that did not belong to either of the two extreme groups uncovered switch errors in 6.8% of the genes. Thus, the switch errors that occurred in particularly difficult-to-phase genes did not appear to affect the key results in any way; that is, they affected neither the global *cis/trans* ratios nor the distinction of *cis*- and *trans*-abundant genes.

4.2 Annotation of Coding Variants

First, RefSeq genes were downloaded from UCSC table browser (Hg19). All transcripts that belong to an autosomal gene were merged and the coordinates that mark the entire gene region extracted. This resulted in a final set of 18,121 autosomal protein-coding genes in our study. It is worth mentioning that we were able to confirm that pseudogenes were indeed excluded from analysis. With this, we wanted to refute the assumption that small imperfections in the pseudogene detection algorithm could account for false significances, for instance, attributable to the olfactory receptors (*see* **Note 5**).

For analysis of the 1092 genomes, annotated potentially protein function-altering nsSNPs (PFA-nsSNPs) as well as nsSNPs and sSNPs, respectively, were available from the 1000G database. To predict a potentially significant effect of nsSNPs on protein structure and function in PGP, a combination of PolyPhen-2 [40] and SIFT [41, 42] (see also [24]) as well as GERP conservation scores [43] were applied to ensure comparability with the 1000G annotations. Default threshold values for PolyPhen-2 and SIFT, and GERP scores >2, were used. The annotation of nsSNPs and sSNPs in PGP data was provided by the PGP database [23]. Importantly, the annotation of the variants and thus the determination of their *cis* and *trans* configurations always relate to the minor allele (see Fig. 1 and **Note 1**). Furthermore, our approach implies that each of the non-reference alleles of each autosomal gene with ≥2 coding variants is scored independently of its allele frequency in the population. Thus, none of the minor alleles were ignored and no frequency cut-off filter was applied.

4.3 Scoring Cis and Trans Configurations

In the first step, those genes that have PFA-nsSNPs are extracted from each of the 1092 genomes, and 1092 intermediate data output files are prepared. It is important that these are organized in table format as follows: "Haplotype 1" and "Haplotype 2" are contained in two adjacent columns and together represent an individual diploid genome. The rows represent the heterozygous genomic positions, with their corresponding gene IDs assigned; the heterozygous positions are sorted 5′ to 3′ in a gene. (Homozygous variants are excluded from the analysis.) Within each row, the two neighboring cells in the columns "Haplotype 1" and "Haplotype 2" contain the two alleles at the heterozygous position of a gene as assigned to the two haplotypes 1 and 2, respectively. Thus, one of the cells contains the reference allele designated "0," and the other the non-reference allele designated "1." Thus, each of the two columns/haplotypes consists of a unique combination of "1's" and "0's". Both haplotype columns have as many rows per gene ID as there are heterozygous positions (i.e., in this case heterozygous PFA-nsSNPs) in this gene. It is important to note that although a gene can have many heterozygous positions in a *population*, the number of its heterozygous positions in an *individual genome* is usually limited to a few. For a more detailed information, *see* **Note 6**.

For the analysis of the PGP haplotype data, PolyPhen-2 and SIFT in combination with GERP (as described above) are applied to the PGP output files generated from each of the 184 experimentally phased genomes, and intermediate output files prepared accordingly (*see* **Note 7**).

In the next step, all genes that have only one row assigned, that is, only one heterozygous position, are removed from the intermediate output files to ensure that only genes with ≥2 PFA-nsSNPs are present in these files. Then the alleles, which are assigned to the same gene ID and sorted 5′ to 3′ within both "Haplotype 1" and "Haplotype 2" columns, are stored as units. These units are then subjected to the assessment of phase configurations. Importantly, while it is a *pair* of haplotypes that underlie a *cis* or *trans* configuration, it is *sufficient to inspect only one haplotype*, per definition "Haplotype 1," to be able to score a gene "*cis*" or "*trans.*" Thus, a gene is classified "*cis*" if every allele in "Haplotype 1" is either 1 or 0 (non-reference or reference), otherwise "*trans.*" This will allow immediate calculation of global as well as gene-based *cis/trans* ratios (see below and Fig. 1). It is needless to say that information on these key bioinformatic steps provides the necessary basis for automation to enable large-scale analyses.

4.4 Calculation of Global and Gene-Based Cis/Trans Ratios

4.4.1 Global Cis/Trans Ratios

Global *cis/trans* ratios refer to one or multiples of diploid genomes. The global *cis/trans* ratio of an individual genome as the unit of analysis is defined as the ratio of *cis* fraction to *trans* fraction (see also Fig. 1). The *cis* fraction (%) is calculated as the number of (autosomal) genes with *cis* configurations divided by total number of genes with ≥2 variants, also referred to as "total configuration count" (equivalent to 100%); the *trans* fraction is calculated analogously and equivalent to 100% – *cis* (%).

To determine the *cis/trans* ratios for many genomes, for instance, the 1092 or 184 genomes, or any subsets of these, the median values of the *cis* and *trans* fractions are calculated across a defined number of genomes.

4.4.2 Gene-Based Cis/Trans Ratios

Gene-based *cis/trans* ratios are determined to identify *cis*- and *trans*-abundant genes, that is, genes with ≥2 PFA-nsSNPs that exhibit either *cis* or *trans* configurations in significant excess. The gene-based *cis/trans* ratio is defined as the ratio of *cis* fraction to *trans* fraction of a given gene in a defined population sample. The gene-based *cis* fraction (%) is calculated as follows: number of *cis* configurations of functionally annotated nsSNPs observed for the gene across all genomes in the population sample divided by total number of genomes in which the gene has ≥2 variants, that is, total configuration count of this gene in the population sample (equivalent to 100%). The gene-based *trans* fraction (%) is calculated analogously and equivalent to 100% – *cis* (%).

The **distinction of *cis*- and *trans*-abundant genes** was carried out by demonstrating a significant abundance of one of the two configurations with a Binomial test ($P < 0.05$) (see also Fig. 1).

4.5 Modeling the Expected Composite Probability of a Cis Configuration

The probability of an observed *cis* or *trans* configuration in a gene with i variants can be modelled with a Bernoulli experiment, $P_i(X = 1)$ and $P_i(X = 0)$, where $X = 1$ denotes a *cis* configuration and $X = 0$ a *trans* configuration. Thus, we have $P_i(X = 1) = \frac{1}{2^{i-1}}$ and $P_i(X = 0) = 1 - P_i(X = 1)$ taking into account the genomic order of the variants. Among the number of genes with ≥2 variants that have either *cis* or *trans* configurations, let w_i be the relative frequency of genes with exactly i variants. Thus, we have $\Sigma_{i \geq 2} w_i = 1$. The probability of observing a *cis* configuration among all phase-sensitive genes in a genome is then given by the weighted sum of the above defined Bernoulli probabilities: P-$(X = 1) = \Sigma_{i \geq 2} w_i P_i(X = 1)$. Thus, inserting the observed relative frequencies, w_i, from the 1000G data yielded an expected probability of ~0.4 for a *cis* configuration to occur.

As outlined earlier, the significance of a *cis/trans* ratio calculated for an individual genome is computed with an exact Binomial test with $P = 0.4$. To assess the significance values for *cis/trans* ratios calculated for population samples, the median values for both *cis* and *trans* fractions across all genomes are derived and the “median genome” treated as an individual genome.

4.6 Simulating Phased Genomes Under Random Assumptions to Derive the Expected Composite Cis/Trans Ratios

To corroborate the theoretical assumptions on the composite probability of a *cis* or *trans* configuration as modeled by a Bernoulli experiment, simulations of phase were performed in addition, assuming that the variants are distributed randomly between the two homologues of a gene. Thus, we simulated diploid genomes as follows: In the first step, a virtual set of 1092 phased genomes was generated. For each virtual genome, random numbers of PFA-nsSNPs were drawn within the range observed in the 1092 genomes data set (~2500–3500) [38]; the PFA-nsSNPs were sampled from the total of ~300,000 PFA-nsSNPs annotated in this data set. Then phase was simulated by assigning a 50:50 chance to each individual SNP in a gene to exist on either homologue. In practice, this was achieved by randomly drawing a phase, that is, “homologue 1” or “homologue 2” attached, from the 1000G database with each individual PFA-nsSNP. Accordingly, a second virtual set of 1092 phased genomes was generated, simulating a random distribution of all nsSNPs between the two homologues. To this end, between ~5500 and ~7500 nsSNPs with either homologue 1 or 2 attached were drawn from the entire pool of ~1.5 million nsSNPs annotated in the 1092 "real" genomes. Two additional virtual sets of 1092 phased genomes were generated, simulating analogously a random distribution of all sSNPs, and of all nsSNPs and sSNPs combined. Subsequently, the *cis/trans* ratios were calculated for these simulated phased genomes as described above.

To test the validity of our approach to simulate phase under random assumptions, we assessed the *cis/trans* ratios separately for two up to five variants in all virtual data sets. Then we compared these ratios to the probabilities P for these numbers of variants to occur in *cis* under conditions of random distribution, which is $1/2^{n-1}$, with n the number of variants. The comparative evaluation showed that the *cis/trans* ratios, which were generated for defined numbers of variants by simulation, were virtually identical to those expected. Accordingly, the simulated *cis* fraction for pairs of variants was approximately 50%, for combinations of three variants ~25%, for combinations of four variants ~12.5%, and for five variants ~6.25%. These results validated our simulation studies, and we were able to derive confidently the expected composite *cis* ratios: ~39% for PFA-nsSNPs; ~37% for nsSNPs and sSNPs, respectively; and 33% for combining all types of coding variants. Overall, these proved to be in excellent agreement with the theoretically derived composite probability of a *cis* or *trans* configuration to occur. In sum, our simulation studies resulted in expected composite *cis* ratios of approximately 40% and lower versus observed *cis* ratios of about 60% (*see* also graph in Fig. 3).

4.7 Gene Set Enrichment and Protein–Protein Interaction Integration with ConsensusPathDB

ConsensusPathDB is a meta-database for molecular interactions and pathways that currently integrates 31 public resources (*see* **Note 8**). With more than 850,000 different, experimentally derived molecular interactions, it represents one of the most comprehensive models of the human interactome [44]. ConsensusPathDB is accessible via a Web server (http://consensuspathDB.org) with functionality for characterizing genes, proteins, metabolites, and other types of biomolecules at the level of pathways and interaction networks. ConsensusPathDB is used in this study to compute the overrepresentation of pathways and GO categories for the classes of *cis*- and *trans*-abundant genes using Fisher's exact test. Test p-values are adjusted for multiple testing with the Benjamini–Hochberg correction.

Besides information on 5578 biological pathways, the ConsensusPathDB holds an integrated network of molecular interactions that was constructed from 19 different databases comprising more than 522,618 binary protein–protein interactions (PPIs) [35]. The integrated ConsensusPathDB PPI network is available from the download section of the Web server (http://cpdb.molgen.mpg.de/download/ConsensusPathDB_human_PPI.gz).

Since PPI networks can contain a large share of false-positive interactions, the quality assessment of individual PPIs is particularly critical. In ConsensusPathDB, we provide a quality check for PPIs consisting of different methods [45]. Only interactions with a high-quality score >0.95 were used in this study, which results in an interaction network comprising 10,560 genes and 139,267 binary interactions.

4.8 Network Propagation

Network propagation is a theoretical framework for network analyses. It describes a set of analysis tools that use experimental data such as genotype data, expression data, or categorical data to initialize node weights and subsequently distribute these weights simultaneously to the network neighborhoods of the nodes [46]. This process converges to a steady state and leads to a re-ranking of the original network nodes. This re-ranking typically amplifies functional associations and is used to identify hot-spot subnetworks that agglomerate much of the experimental weights and can be associated with specific biological pathways or parts thereof. Typical applications are to draw inference on genotype–phenotype relations from mutation data [47] or to identify functional networks from gene and protein expression data [48]. The integrated ConsensusPathDB PPI described above has been used in the past as a resource for network propagation [49], and it has been evaluated as one of the best performing networks for disease gene identification in an independent benchmark comparison among 21 publicly available networks [37].

Network propagation has been carried out with the tool NetCore [50]. This is a semi-supervised workflow, which applies random walk with restart on a robust node metric (node core) in order to derive a re-ranking of the network nodes and, in a second step, the identification of subnetworks agglomerating highly re-ranked nodes with seed genes supplied by the user. A documentation of the algorithm can be found at https://github.molgen.mpg.de/barel/NetCore (**Note 9**).

We applied network propagation separately with the list of *cis*- and *trans*-abundant genes. Gene node scores, s_i were initialized as follows

$$s_i = -\, log_{10} p_i,$$

where p_i is the p-value of the Binomial test for testing *cis*- and *trans*-abundant genes (Sect. 2.2.2). All other nodes were set to zero. After the re-ranking step, subnetworks were derived by connecting the significantly re-ranked nodes with the top 100 *cis*- and *trans*-abundant genes, respectively.

5 Conclusions and Outlook

Our work represents an analysis of the largest body of experimentally haplotype-resolved genomes to date, complementing the analysis of 1000 Genomes statistical haplotype data. We were able to show that the systematic investigation of *cis* and *trans* configurations as two major categories of *diplotypic* genetic variation can lead to biologically relevant results that provide first insights into the nature and architecture of diploid human genomes. Thus, we have distinguished and further functionally characterized two classes of autosomal protein-coding genes, so-called *cis*- and *trans*-abundant genes, which together constitute a "common diplotypic exome." We have thus described those parts of the human exome that preferentially exist as diplotypes, in contrast to a set of invariable genes under strong mutational constraint ([51]; *see* **Notes 10, 11**). These diplotypes may encode two functionally different homologues of the genes and play an important role in the diploid biology of genes and genomes. Our results have implications for the conceptual and functional characterization of autosomal genes in the context of a diploid biology and highlight the importance of phase information for the interpretation of protein-coding genetic variation. (A detailed discussion of our results and their implications can be found in Hoehe et al. [24]).

For our analyses, we used, on the one hand, established bioinformatic tools, such as for the prediction of potentially protein-altering coding variants, and on the other hand, we have developed first approaches to the large-scale analysis of experimentally haplotype-resolved genomes. These include the efficient scoring of *cis*- and *trans* configurations and the simulation of phased genomes to allow evaluation of composite *cis/trans* ratios under random assumptions. This served to assess the significance of the observed *cis/trans* ratios. Moreover, for advanced functional annotation, we applied a most recently updated database, which allowed expanding the functional analyses onto a higher hierarchical level of genome organization, protein–protein interactions, while substantiating the previous results. Because there were no precedents for some of our research approaches, we have attempted to develop new terms and approaches to describe and quantify the distribution of *cis* and *trans* configurations, such as the global and gene-based "*cis/trans* ratio."

The available extensive, cross-validated phase-information on protein-coding genetic variation in a sizeable number of genomes allowed a more confident, accurate assessment of the phase-sensitive part of diploid human exomes. Thus, more than one-third (6287) of the 18,121 autosomal protein-coding genes showed ≥2 protein-changing variants in fractions of the 1092 genomes and therefore could be either *cis*- or *trans*-abundant.

Almost half (47%) of all protein-changing variants were found to exist in phase, 25% in a *cis* and 22% in a *trans* configuration; correspondingly, 62% of the entirety of nsSNPs existed in phase, half each in *cis* and *trans* (see also **Note 11**). Finally and most importantly, the analysis of extensive phase information enabled the detection of *cis*- and *trans*-abundant genes and thus the common diplotypic exome as characteristic features of diploid genome organization. In contrast, only few exome studies have addressed phase issues, mainly related to MNVs, defined as two or more nearby variants existing on the same haplotype. These included variants within 2 bp, that is, within a codon [10, 32], up to 10–100 bp distance [27, 29–31] of each other, identified mostly by use of read-based phasing, that is, local phase information. In the largest study of MNVs in 125,748 exomes, a total of 18,756 MNVs with functional impact were described [10], corresponding to 0.15 MNVs per genome. In our work, we surveyed protein-changing variants in phase up to observed maximum distances of ~81 kb, with an average of >2.7 protein-changing variants per gene. Thus, our work represents an advance in the phase-informed analysis of protein-coding genetic variation.

Importantly, comprehensive phase information allowed classification and further functional characterization of *cis*- and *trans*-abundant genes. For this, we applied functional enrichment to distinguish these gene classes in terms of general functions and different levels of organization such as gene pathways and interaction networks. Thus, *cis*-abundant genes were, for instance, involved in numerous metabolic functions, and *trans*-abundant genes in many immune response-related processes and diseases. The enriched functions overall appear to concern cell–environment and cell–cell communication, membrane-related processes, metabolism and biosynthesis, and (disease-related) immune processes. By encoding two potentially functionally different homologues, *cis*- and *trans*-abundant genes can exert great functional flexibility and thus modulate these functions, allowing adaptation of the organism to external and internal stimuli. Their potential role as modulators and adaptive agents is supported by the observation of substantial overlap with monoallelically expressed (MAE) genes; this also refers to the functional classes and evolutionarily significant gene sets that were overrepresented, as well as the shift in allele frequency distributions toward those consistent with common variation (i.e., greater allelic age on average). These results support the hypothesis that *cis*- and *trans*-abundant genes, in constituting a common diplotypic exome, may play an important role in adaptive and evolutionary processes and generate widespread cell-to-cell, organismal, and phenotypic diversity.

Furthermore, *cis*-abundant genes were characterized by frequently occurring pairs of protein-altering variants that were close together. Thus, these could represent "evolutionary signals" going back to ancient populations. Reasons for their preservation could be epistatic or compensatory interactions between their corresponding amino acid substitutions that maintain or enhance the functionality of the protein [52–54], coevolution [55], or hitchhiking effects [56]. Further examination of these results could provide valuable information for studying protein evolution, structure, and functionality. Finally, *cis*-abundant genes were 1.6- to two-fold more common than *trans*-abundant genes, resulting overall in a significant, global *cis*-abundance at a ~60:40 *cis/trans* ratio as a universal characteristic of diploid human genomes. Thus, the apparently old, co-occurring protein-changing coding variants characterizing *cis*-abundant genes underlie this phenomenon to a considerable extent.

These findings open up many avenues of research for further exploration. To begin with, studies in much larger samples and different populations are desirable to further substantiate and refine these findings. Under this precondition, an in-depth examination of the entirety of protein-changing variants that (in combination) constitute the *cis* and *trans* configurations will be of great interest. For instance, it can be clarified to what extent co-occurring variants can modify each other and/or cause *cis*-suppression of pathogenic variants. Then there is a need for population and evolutionary genetics studies to gain a better understanding of the likely common underlying mechanisms of our findings, which were observed equally in all populations studied. These could involve processes of ancestral admixture and ancient (balancing) selection as well as other evolutionary forces. The common pairs of protein-altering variants that characterize *cis* abundance and possibly represent "ancestral signals" may help trace the ancient origins of the phenomena described. Another important issue is which of the two homologues of *cis*- and *trans*-abundant genes are active in differentially expressed transcriptomes and proteomes as well as higher (omics) levels of genome organization. Overall, future questions concern the role of diplotypic genes in transcriptome, proteome, and phenotype diversity within and between cells, tissues, individuals, populations, and species, at various developmental and physiological stages, as well as health and disease [24]. Finally, as far as the functional annotation of diplotypic genes is concerned, future approaches to the analysis of sequence–structure–function relationships must focus on combinations of variants, that is, haplotypes, both in silico and in vitro. Thus, these approaches must account for three potentially different states per gene, that is, express and characterize each of the two molecular haplotypes of a gene separately, and together as a pair [2, 4, 6, 22]. Overall, these are just a few of many important questions arising from our attempt to capture the diplotypic genome architecture as the basis for a diploid biology that is inherently allelically biased at all levels of genome organization. Our work raises even more intriguing questions for a more distant future, such as whether and how biology and phenotype might change with *cis/trans* ratios or whether *cis* abundance might represent a general phenomenon across all diploid species.

6 Notes

1. For the reader's information, the fraction of non-reference alleles that have an allele frequency >0.5 and therefore are not minor alleles has been estimated genome-wide to about 2.5% (https://www.biostars.org/p/119420/). Thus, the fraction of such alleles in the coding exome seems to be extremely small. The given link provides further information on the correspondence of non-reference and reference alleles with minor and major alleles.
2. We proceeded as follows: First, we identified those phase-sensitive genes, which were shared by the global set (1000G) and PGP, resulting in 1627 phase-sensitive genes. Then we determined the configuration types in this overlap separately for 1000G and PGP. 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 71 and 72 genes, respectively, indicated that 8.7% of the *cis*- and 12.9% of the *trans*-abundant genes had changed configuration type in PGP.
3. We examined the question, whether our initial case observations could represent a more general picture, as follows: First, we identified the "major configuration," that is, the most frequently occurring pair of functionally annotated nsSNPs, for each of the 1227 *cis*- and 786 *trans*-abundant genes. We then calculated for each gene the "major configuration frequency" (MCF) (%) defined as follows: number of times the major configuration is observed in a population sample divided by the total number of *cis* configurations counted for this gene in the sample. The MCFs were calculated separately (i) for the 1227 *cis*- and 786 *trans*-abundant genes, (ii) for the cross-validated 322 *cis*-abundant and 153 *trans*-abundant genes from 1000G, and (iii) for the cross-validated 322 *cis*-abundant and 153 *trans*-abundant genes from PGP. Subsequently, the genes were binned by MCF into 10% intervals up to a MCF of 100%, which were then related to the number of genes (see also Hoehe et al. [24]).
4. Referring to the recent reports of MNVs as outlined in the Introduction, our findings could raise questions about the presence of MNVs of functional impact in our experimentally

haplotype-resolved genomes. Referring to the report by Wang et al. [10], we expected that the number of MNVs likely to be found in the 184 PGP genomes would be very small. Given that a total of 18,756 MNVs (pairs of SNPs within 2 bp distance of each other) with a novel, combined effect on protein sequence were identified in 125,748 exomes, equivalent to 0.15 MNVs per genome, a total of ~28 MNVs could be expected in our sample. Thus, our sample, although the largest set of experimentally haplotype-resolved genomes generated to date, does not promise to provide new insights into this issue, apart from the fact, that appropriate phenotypic data are not readily available, and our capacities limited.

5. Since pseudogenes could interfere with the analysis of the derived set of autosomal protein-coding genes, we have excluded them from analysis. To this end, we have downloaded from BioMart the two types of human pseudogenes, that is, "processed" (retro-transposed) pseudogenes and "unprocessed" pseudogenes generated through imperfect duplication. For the reader's information, BioMart identified 10,440 processed and 2937 unprocessed pseudogenes in the human genome build at the time of our analyses. Both "processed" and "unprocessed" pseudogenes generated minimal overlap with all 18,121 autosomal protein-coding genes in study. Furthermore, we were able to exclude an impact of pseudogenes on olfactory receptors (ORs), which could account for the significance attributed to the olfactory receptors, by intersection. Accordingly, only 8 of the total 372 OR genes that have been analyzed in our study were found to overlap with the 364 Ors comprised in the unprocessed pseudogenes. Thus, these pseudogenes did not contribute to the statistics for the OR gene family in our approach.

6. For the reader's information, the number of heterozygous PFA-nsSNPs per gene, averaged over all 1092 genomes, was 2.73, with a maximum of 26 PFA-nsSNPs per gene. If all heterozygous nsSNPs in the protein-coding regions of the autosomal genes were taken into account, the average was 3.2 nsSNPs per gene.

7. PGP intermediate output files should be consistent with 1000G output file format.

8. Pathway enrichment in this study is performed with the ConsensusPathDB resource, which integrates 31 human resources and pathways annotated from multiple resources such as KEGG, Reactome, or WikiPathways. Similarly, we have integrated protein–protein interactions from multiple resources such as BioGrid, Intact, or DIP. We find in all cases that working with multiple resources is beneficial since the

pathway content in different databases is still largely complementary. Thus, to explore the full functional information inherent in a gene list, integration of pathway and interactome information is important.

9. Network propagation was carried out with the tool NetCore. There are many alternatives for these kinds of analyses; however, most tools consist of two steps, node re-ranking and module identification. In NetCore, the node re-ranking step is done with a random walk with restart procedure in which initial node weights are distributed across the network until a convergence state is reached. At the end of this step, each node contains a re-ranked weight and the significance of this weight is assessed with random graph models. In contrast to other tools that utilize node degree in the propagation step, NetCore relies on the cores of the nodes. Node core is a measure for the density of the node neighborhood and has been shown to be more robust against degree bias in protein–protein interactions.

10. A total of 5040 genes did not contain any potentially protein-changing variants at all in the 1092 genomes and were significantly enriched for numerous essential cellular functions that may not tolerate variability ($P < 6.15E\text{-}25$–$9.51E\text{-}11$ for the top 50 pathways; $P < 2.35E\text{-}24$–$2.12E\text{-}14$ for the top 50 GOs). This finding is in agreement with other reports on mutational constraint that classify autosomal protein-coding genes along a spectrum of tolerance to inactivation [51].

11. In our results section, we have not included analyses of genes with 1 PFA-nsSNPs. Although these are not phase-sensitive, they encode two potentially functional different homologues of a gene and thus represent biologically molecular diplotypes. We have included these “molecular diplotypes” in our expanded analyses reported recently [24].

References

1. Wu C-T, Dunlap JC (2002) Homology effects: the difference between 1 and 2. In: Dunlap JC, Wu C-T (eds) Advances in genetics, vol 46, 1st edn. Academic Press, pp xvii–xxiii
2. Hoehe MR (2003) Haplotypes and the systematic analysis of genetic variation in genes and genomes. Pharmacogenomics 4(5):547–570. https://doi.org/10.2217/14622416.4.5.547
3. Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ (2011) The importance of phase information for human genomics. Nat Rev Genet 12(3):215–223. https://doi.org/10.1038/nrg2950
4. Suk EK, McEwen GK, Duitama J, Nowick K, Schulz S, Palczewski S, Schreiber S, Holloway DT, McLaughlin S, Peckham H et al (2011) A comprehensively molecular haplotype-resolved genome of a European individual. Genome Res 21(10):1672–1685. https://doi.org/10.1101/gr.125047.111
5. Benzer S (1957) The elementary units of heredity. In: McElroy WD, Glass B (eds) The chemical basis of heredity. Johns Hopkins University Press, Baltimore, pp 70–93
6. Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K,

Arnold K, Ruano G, Liggett SB (2000) Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proc Natl Acad Sci U S A 97(19): 10483–10488. https://doi.org/10.1073/pnas.97.19.10483

7. Wang Y, Zhang W, Edelmann L, Kolodner RD, Kucherlapati R, Edelmann W (2010) Cis lethal genetic interactions attenuate and alter p53 tumorigenesis. Proc Natl Acad Sci U S A 107(12):5511–5515. https://doi.org/10.1073/pnas.1001223107

8. Smith J (2021) The next 20 years of human genomics must be more equitable and more open. Nature 590(7845):183–184. https://doi.org/10.1038/d41586-021-00328-0

9. Crespi S (2021, February 4) Looking back at 20 years of human genome sequencing. Podcast Science. https://www.science.org/content/podcast/looking-back-20-years-human-genome-sequencing

10. Wang Q, Pierce-Hoffman E, Cummings BB, Alföldi J, Francioli LC, Gauthier LD, Hill AJ, O'Donnell-Luria AH, Genome Aggregation Database Production Team; Genome Aggregation Database Consortium et al (2020) Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. Nat Commun 11(1):2539. https://doi.org/10.1038/s41467-019-12438-5

11. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, Benner C, Liu D, Locke AE, Balasubramanian S et al (2021) Exome sequencing and analysis of 454,787 UK biobank participants. Nature 599(7886): 628–634. https://doi.org/10.1038/s41586-021-04103-z

12. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G (2007) The diploid genome sequence of an individual human. PLoS Biol 5(10):e254. https://doi.org/10.1371/journal.pbio.0050254

13. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J et al (2008) The diploid genome sequence of an Asian individual. Nature 456(7218):60–65. https://doi.org/10.1038/nature07484

14. Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, Ng SB, Alkan C, Qiu R, Eichler EE et al (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nat Biotechnol 29(1): 59–63. https://doi.org/10.1038/nbt.1740

15. Fan HC, Wang J, Potanina A, Quake SR (2011) Whole-genome molecular haplotyping of single cells. Nat Biotechnol 29(1):51–57. https://doi.org/10.1038/nbt.1739

16. Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, Jiang Y, Dahl F, Tang YT, Haas J et al (2012) Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature 487(7406):190–195. https://doi.org/10.1038/nature11236

17. Huang M, Tu J, Lu Z (2017) Recent advances in experimental whole genome haplotyping methods. Int J Mol Sci 18(9):1944. https://doi.org/10.3390/ijms18091944

18. Choi Y, Chan AP, Kirkness E, Telenti A, Schork NJ (2018) Comparison of phasing strategies for whole human genomes. PLoS Genet 14(4):e1007308. https://doi.org/10.1371/journal.pgen.1007308

19. Porubsky D, Ebert P, Audano PA, Vollger MR, Harvey WT, Marijon P, Ebler J, Munson KM, Sorensen M, Sulovari A et al (2021) Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. Nat Biotechnol 39(3): 302–308. https://doi.org/10.1038/s41587-020-0719-5

20. Duitama J, McEwen GK, Huebsch T, Palczewski S, Schulz S, Verstrepen K, Suk EK, Hoehe MR (2012) Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of single individual haplotyping techniques. Nucleic Acids Res 40(5): 2041–2053. https://doi.org/10.1093/nar/gkr1042

21. Suk EK, Schulz S, Mentrup B, Huebsch T, Duitama J, Hoehe MR (2017) A Fosmid Pool-based next generation sequencing approach to haplotype-resolve whole genomes. Methods Mol Biol 1551:223–269. https://doi.org/10.1007/978-1-4939-6750-6_13

22. Hoehe MR, Church GM, Lehrach H, Kroslak T, Palczewski S, Nowick K, Schulz S, Suk EK, Huebsch T (2014) Multiple haplotype-resolved genomes reveal population patterns of gene and protein diplotypes. Nat Commun 5:5569. https://doi.org/10.1038/ncomms6569

23. Mao Q, Ciotlos S, Zhang RY, Ball MP, Chin R, Carnevali P, Barua N, Nguyen S, Agarwal MR, Clegg T et al (2016) The whole genome sequences and experimentally phased haplotypes of over 100 personal genomes. Gigascience 5(1):42. https://doi.org/10.1186/s13742-016-0148-z

24. Hoehe MR, Herwig R, Mao Q, Peters BA, Drmanac R, Church GM, Huebsch T (2019) Significant abundance of *cis* configurations of coding variants in diploid human genomes.

Nucleic Acids Res 47(6):2981–2995. https://doi.org/10.1093/nar/gkz031

25. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R et al (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 372(6537): eabf7117. https://doi.org/10.1126/science.abf7117
26. Hoehe MR, Köpke K, Wendel B, Rohde K, Flachmeier C, Kidd KK, Berrettini WH, Church GM (2000) Sequence variability and candidate gene analysis in complex disease: association of mu opioid receptor gene variation with substance dependence. Hum Mol Genet 9(19):2895–908. https://doi.org/10.1093/hmg/9.19.2895
27. Rosenfeld JA, Malhotra AK, Lencz T (2010) Novel multi-nucleotide polymorphisms in the human genome characterized by whole genome and exome sequencing. Nucleic Acids Res 38(18):6102–11. https://doi.org/10.1093/nar/gkq408
28. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536(7616):285–91. https://doi.org/10.1038/nature19057
29. Schrider DR, Hourmozdi JN, Hahn MW (2011) Pervasive multinucleotide mutational events in eukaryotes. Curr Biol 21(12):1051–4. https://doi.org/10.1093/hmg/9.19.2895
30. Besenbacher S, Sulem P, Helgason A, Helgason H, Kristjansson H, Jonasdottir A, Jonasdottir A, Magnusson OT, Thorsteinsdottir U, Masson G et al (2016) Multi-nucleotide de novo Mutations in Humans. PLoS Genet 12 (11):e1006315. https://doi.org/10.1371/journal.pgen.1006315
31. Kaplanis J, Akawi N, Gallone G, McRae JF, Prigmore E, Wright CF, Fitzpatrick DR, Firth HV, Barrett JC, Hurles ME; Deciphering Developmental Disorders study (2019) Exome-wide assessment of the functional impact and pathogenicity of multinucleotide mutations. Genome Res 29(7):1047–1056. https://doi.org/10.1101/gr.239756.118
32. Degalez F, Jehl F, Muret K, Bernard M, Lecerf F, Lagoutte L, Désert C, Pitel F, Klopp C, Lagarrigue S (2021) Watch Out for a Second SNP: Focus on Multi-Nucleotide Variants in Coding Regions and Rescued Stop-Gained. Front Genet 12:659287. https://doi.org/10.3389/fgene.2021.659287
33. McElwain MA, Zhang RY, Drmanac R, Peters BA (2017) Long Fragment Read (LFR) technology: cost-effective, high-quality genome-wide molecular haplotyping. Methods Mol Biol 1551:191–205. https://doi.org/10.1007/978-1-4939-6750-6_11
34. Pääbo S (2003) The mosaic that is our genome. Nature 421:409–412. https://doi.org/10.1038/nature01400
35. Kamburov A, Herwig R (2022) ConsensusPathDB 2022: molecular interactions update as a resource for network biology. Nucleic Acids Res 50(D1):D587–D595. https://doi.org/10.1093/nar/gkab1128
36. Savova V, Chun S, Sohail M, McCole RB, Witwicki R, Gai L, Lenz TL, Wu CT, Sunyaev SR, Gimelbrant AA (2016) Genes with monoallelic expression contribute disproportionately to genetic diversity in humans. Nat Genet 48(3):231–237. https://doi.org/10.1038/ng.3493
37. Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P, Ideker T (2018) Systematic evaluation of molecular networks for discovery of disease genes. Cell Syst 6(4): 484–495.e5. https://doi.org/10.1016/j.cels.2018.03.001
38. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65. https://doi.org/10.1038/nature11632
39. Ball MP, Thakuria JV, Zaranek AW, Clegg T, Rosenbaum AM, Wu X, Angrist M, Bhak J, Bobe J, Callow MJ, Cano C, Chou MF et al (2012) A public resource facilitating clinical use of genomes. Proc Natl Acad Sci U S A 109(30):11920–11927. https://doi.org/10.1073/pnas.1201904109
40. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7(4):248–249. https://doi.org/10.1038/nmeth0410-248
41. Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11(5):863–874. https://doi.org/10.1101/gr.176601
42. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4(7):1073–1081. https://doi.org/10.1038/nprot.2009.86
43. Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A (2005) Distribution

and intensity of constraint in mammalian genomic sequence. Genome Res 15(7):901–913. https://doi.org/10.1101/gr.3577405

44. Herwig R, Hardt C, Lienhard M, Kamburov A (2016) Analyzing and interpreting genome data at the network level with ConsensusPathDB. Nat Protoc 11(10):1889–907. https://doi.org/10.1038/nprot.2016.117

45. Kamburov A, Stelzl U, Herwig R (2012) Int Score: a web tool for confidence scoring of biological interactions. Nucleic Acids Res 40 (Web Server issue):W140–W146. https://doi.org/10.1093/nar/gks492

46. Cowen L, Ideker T, Raphael BJ, Sharan R (2017) Network propagation: a universal amplifier of genetic associations. Nat Rev Genet 18(9):551–562. https://doi.org/10.1038/nrg.2017.38

47. Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M et al (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet 47(2):106–114. https://doi.org/10.1038/ng.3168

48. Drake JM, Paull EO, Graham NA, Lee JK, Smith BA, Titz B, Stoyanova T, Faltermeier CM, Uzunangelov V, Carlin DE et al (2016) Phosphoproteome integration reveals patient-specific networks in prostate cancer. Cell 166(4):1041–1054. https://doi.org/10.1016/j.cell.2016.07.007

49. Selevsek N, Caiment F, Nudischer R, Gmuender H, Agarkova I, Atkinson FL, Bachmann I, Baier V, Barel G, Bauer C et al (2020) Network integration and modelling of dynamic drug responses at multi-omics levels. Commun Biol 3:573. https://doi.org/10.1038/s42003-020-01302-8

50. Barel G, Herwig R (2020) NetCore: a network propagation approach using node coreness. Nucleic Acids Res 48(17):e98. https://doi.org/10.1093/nar/gkaa639

51. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581(7809):434–443. https://doi.org/10.1038/s41586-020-2308-7

52. DePristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet 6(9):678–687. https://doi.org/10.1038/nrg1672

53. Ferrer-Costa C, Orozco M, de la Cruz X (2007) Characterization of compensated mutations in terms of structural and physico-chemical properties. J Mol Biol 365(1):249–256. https://doi.org/10.1016/j.jmb.2006.09.053

54. Baresić A, Hopcroft LE, Rogers HH, Hurst JM, Martin AC (2010) Compensated pathogenic deviations: analysis of structural effects. J Mol Biol 396(1):19–30. https://doi.org/10.1016/j.jmb.2009.11.002

55. Anishchenko I, Ovchinnikov S, Kamisetty H, Baker D (2017) Origins of coevolution between residues distant in protein 3D structures. Proc Natl Acad Sci U S A 114(34):9122–9127. https://doi.org/10.1073/pnas.1702664114

56. Slatkin M (2008) Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9(6):477–485. https://doi.org/10.1038/nrg2361