HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2016 August 10.

Published in final edited form as:

Nat Genet. 2016 March ; 48(3): 231–237. doi:10.1038/ng.3493.

Genes with monoallelic expression contribute disproportionately to genetic diversity in humans

Virginia Savova[#1,2,4], **Sung Chun**[#3], **Mashaal Sohail**[#3], **Ruth B. McCole**[2], **Robert Witwicki**[1], **Lisa Gai**[1], **Tobias L. Lenz**[3,5], **C.-ting Wu**[2], **Shamil R. Sunyaev**[3], and **Alexander A. Gimelbrant**[1,2]

[1] Dana-Farber Cancer Institute, Boston, USA

[2] Department of Genetics, Harvard Medical School, Boston, USA

[3] Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA

[#] These authors contributed equally to this work.

Abstract

An unexpectedly large number of human autosomal genes are subject to monoallelic expression (MAE). Our analysis of 4,227 such genes reveals surprisingly high genetic variation across human populations. This increased diversity is unlikely to reflect relaxed purifying selection. Remarkably, MAE genes exhibit elevated recombination rate and increased density of hypermutable sequence contexts. However, these factors do not fully account for the increased diversity. We find that the elevated nucleotide diversity of MAE genes is also associated with greater allelic age: their variants tend to be older and are enriched in polymorphisms shared with Neanderthals and chimpanzees. Both synonymous and nonsynonymous alleles in MAE genes have elevated average population frequencies. We also observed strong enrichment of the MAE signature among genes reported to evolve under balancing selection. We propose that an important biological function of widespread MAE might be generation of cell-to-cell heterogeneity; the increased genetic variation contributes to this heterogeneity.

Introduction

Among the epigenetic regulatory modes causing unequal allelic transcription of the mammalian autosomal genes, by far the most widespread is monoallelic expression (MAE), with mitotically stable maint enance of the initial random choice of an active allele[1]. While

Corresponding authors: Alexander Gimelbrant – gimelbrant@mail.dfci.harvard.edu, Shamil Sunyaev – ssunyaev@rics.bwh.harvard.edu.

[4]Current address: Department of Systems Biology, Harvard Medical School, Boston, USA

[5]Current address: Evolutionary Immunogenomics, Department of Evolutionary Ecology, Max Planck Institute for Evolutionary Biology, Plön, Germany

Author Contributions: AAG and SRS conceived the study; all authors contributed to data analysis; AAG, SRS and VS wrote the manuscript with input from SC, MS, and RBM.

Authors declare no competing interests.

individual examples of MAE genes have been known for decades[2], recent developments in transcriptome-wide analysis of allele-specific expression led to a surprising discovery: in every assessed cell type, between 10 and 25% of human and mouse autosomal genes can be subject to MAE in multiple cell types[3-10]. MAE has been directly observed in peripheral blood and derived cell lines, as well as in human placenta[3], mouse lymphoid cells and fibroblasts[4], and mouse neuroprogenitor cells[8]. How gene function and evolution are affected by separate allelic regulation in the same cell nucleus remains a mystery.

The question of allelic diversity is particularly important for understanding the biology of MAE genes. Heterozygosity in an MAE locus may lead to extensive cell-to-cell heterogeneity within tissues (**Supplementary Fig.1**), with potentially dramatic functional differences between otherwise similar cells of the same type[11].

Quantitative models of the evolution of genes with another kind of monoallelic expression, imprinting, predict that deleterious allelic variation in such genes would be more efficiently removed by purifying selection[12,13]. Similarly to imprinted genes, MAE genes as a group could also experience more efficient purifying selection and thus exhibit lower levels of polymorphism than genes showing consistent biallelic expression (BAE genes). At the same time, in contrast to imprinting, MAE genes have both alleles expressed in a tissue as a whole, which might lead to distinct evolutionary consequences, including positive selection for variants that would otherwise be masked[14-16]).

Here, we report the first systematic assessment of the extent and nature of genetic variation of human MAE genes, using several recent large studies of genetic variation in human populations[17-20] and the greatly expanded number of human MAE genes identified on the basis of a distinctive chromatin signature[5]. Stunningly, we find that human genes showing the MAE signature are more genetically variable than BAE genes, substantially increasing the potential for cell-to-cell variability within an individual.

We consider several probable mechanisms that may be responsible for the increased genetic diversity in MAE genes. In addition to somewhat elevated recombination rate and increased density of hypermutable contexts, MAE genes exhibit patterns associated with balancing selection. This suggests a possible evolutionary link between MAE and heterozygote advantage.

Results

Nucleotide diversity is elevated in MAE genes

We have previously used ENCODE chromatin data[21] to identify genes with a specific chromatin signature of MAE in multiple cell types, followed by experimental validation of this classification using allele-specific transcriptome sequencing of clonal cell lines[5]. This is the only high-throughput method so far which is capable in reliably identifying MAE in polyclonal cell lines. By choosing this dataset as a starting point, we deliberately limit ourselves to mitotically stable MAE (see Methods).

Since MAE is largely a tissue-specific phenomenon, and we are interested in evolutionary forces acting on the entire organism, we created a unified dataset of MAE and BAE genes, with one cell line representing each of the following six cell types we had previously characterized for the MAE signature: lymphoid, myeloid, embryonic stem, myocytes, and mammary and vascular epithelia. Note that the chromatin signature has been demonstrated to be effective outside the LCL cell type[22]. To enhance the functional appropriateness of the gene set, we applied several filters to the baseline catalog of genes with the MAE signature[5] (see Methods for details). Specifically, a gene was only included in our MAE dataset if it had the MAE chromatin signature in at least one cell type while being expressed at a moderate or higher level (reads per kilobase per million (RPKM) ≥1). For a gene to be included in the BAE dataset, it should have no MAE signature in any cell type where its expression was detected at any level, and to exhibit moderate expression in atleast one of the other cell types considered. After applying additional filters (such as excluding olfactory receptor genes and the extended MHC region; see Methods for full description), the resulting high-confidence genome-wide dataset contained 10,233 human genes, of which 4,227 were MAE and 6,006 were BAE (**Supplementary Table 1**).

To compare the extent of genetic variation in MAE and BAE genes, we calculated nucleotide diversity (π; ref.[23]) from the sequencing data generated by the 1000 Genomes Project[17]. Surprisingly, nucleotide diversity in coding sequences appeared to be substantially higher in MAE genes than in BAE genes (mean ± 95% CI: $5.0\times10^{-4}\pm2.0\times10^{-5}$ for MAE genes in the global population, $3.3\times10^{-4} \pm 9.3\times10^{-6}$ for BAE genes; **Fig.1a**). High nucleotide diversity in MAE genes was not limited to any functional category of sites and was apparent even in fourfold degenerate sites, where all possible nucleotide changes are synonymous **(**$1.1\times10^{-3}\pm4.4\times10^{-5}$ for MAE genes in the global population, $7.4\times10^{-4}\pm2.7\times10^{-5}$ for BAE genes; **Fig.1b)**. This difference in nucleotide diversity was not limited to a particular population: MAE genes showed a similar increase in π when assessed separately in different populations in the 1000 Genomes Project (**Fig.1**), as well as in African-American and European-American populations in the Exome Sequencing Project dat a[18] (**Supplementary Fig.2)**. Note that nucleotide diversity is robust to the number of cell types with MAE, and the difference between MAE and BAE genes is not diminished when comparing only genes with higher expression levels (**Supplementary Fig.3**). As MAE genes have previously been shown to be enriched for functional categories related to the extracellular matrix and cellular interactions[5], we tested whether these categories could explain the elevated diversity of MAE genes. However, MAE genes remained more diverse than BAE genes after controlling for the relevant Gene Ontology categories ($p = 1\times10^{-4}$; **Supplementary Table 2, Supplementary Fig.4**).

The observed increase in π could be due to a combination of several factors, whose relative contributions might reflect different underlying biological and evolutionary processes. For example, the higher level of nucleotide diversity may reflect relaxed purifying selection if MAE genes were less important for overall fitness. It could also be due to an increased mutation rate. We thus set out to evaluate the roles of different factors in the increase of nucleotide diversity in human MAE genes.

Purifying selection similarly affects MAE and BAE genes

The possibility that weaker purifying selection explains the elevated nucleotide diversity of MAE genes seems consistent with the observation that housekeeping genes, which are likely to be highly constrained, tend to belong to the BAE set in all cell types[5]. To assess whether MAE genes, as a group, are less constrained by selection, we asked if MAE genes are less likely to be morbid than BAE genes. Using a set of known human morbid genes causing Mendelian diseases (extracted from OMIM database, see Methods), we calculated their representation in the MAE and BAE gene sets. There was no depletion of the morbid genes in the MAE set (**Fig.2a; Supplementary Fig.5**); indeed, there was a slight enrichment ($p<10^{-3}$).

To further estimate the relative effects of purifying selection in the overall MAE and BAE gene sets, we focused on variation in synonymous four-fold degenerate sites. In the 1000 Genomes Project data, nucleotide diversity remains elevated in MAE genes relative to BAE genes to a similar extent when all sites are assessed, as well as when only non-degenerate sites and four-fold degenerate sites are assessed (π_{cds}/π_{ffd}=0.47 and 0.45 for MAE and BAE, respectively, in the global population, **Fig.1**). Similarly, sequence substitutions between human and chimpanzee indicate that the strength of purifying selection on MAE and BAE has been nearly identical. The reduction in nonsynonymous substitution per site compared to synonymous substitution per site (d_N/d_S) measures the proportion of amino acid altering mutations that are selectively unfavorable thus prevented from being fixed in a population[24]. d_N/d_S is not significantly different between MAE and BAE genes as a group (0.21±0.01for both MAE and BAE, **Supplementary Table 3**) and also as per-gene estimates (Wilcoxon rank sum test, p = 0.09, **Supplementary Fig. 6**).

However, as synonymous sites including four-fold degenerate sites have been shown to be under selection to some extent[25], we also compared the frequency of MAE and BAE genes among genes reported to be under selective constraint as assessed by depletion of missense SNPs in the ESP data[26]. Both MAE and BAE genes were equally present in the 1,003 genes reported to be under the highest selective constraints (6.0% and 6.1%, respectively; Fisher's exact p = 0.87). Moreover, MAE and BAE genes are identically distributed with respect to all positive values of selective constraint (**Supplementary Table 4**), ruling out the possibility that purifying selection may affect MAE genes differently in weakly constrained genes. Collectively, these observations suggest that compared to BAE genes, MAE genes do not perform less vital functions and are therefore not expected to be less constrained.

Mutation and recombination rates in MAE genes

To test whether the increased diversity in MAE genes is caused by systematic differences in local mutation rates, we examined the density of hypermutable CpG di-nucleotides, the leading fact or determining sequence-specific differences in mutation rates. We observed that CpG sites are significantly more frequent ($p < 10^{-15}$) in coding sequences of MAE (41.5 CpG per kb) compared to BAE (27.1 CpG per kb). To test whether the difference in CpG content does indeed translate into a difference in mutation rates, we analyzed the per-gene mutation rate map constructed using both human-chimpanzee divergence and observed patterns of *de novo* mutations in humans[20]. This map confirmed the significant elevation of

mutation rates in MAE (**Supplementary Table 5**). Both 4-fold degenerate synonymous mutations and overall protein-coding mutations are 1.28 and 1.22-fold higher in MAE than BAE, respectively ($p<10^{-4}$). This difference appeared fully consistent with the difference in densities of true *de novo* mutations identified in the 250 trio pedigrees of the Genome of the Netherlands (GoNL) project[20] (**Fig.2b** and **Supplementary Table 5)**. However, the latter analysis lacked power due to scarcity of *de novo* mutations.

Interestingly, intronic regions of MAE genes are only slightly more enriched with CpG di-nucleotides (11.2 CpG per kb) compared to BAE (10.9 CpG per kb, **Supplementary Table 5**). Thus, the high CpG density within coding regions of MAE genes is not a consequence of broader regional sequence context. In line with CpG density, the divergence-based mutation rate map indicates that intronic regions of MAE genes have only 1.04-fold higher mutation rate than those of BAE ($p<10^{-4}$), and the set of true *de novo* mutations from GoNL pedigrees also suggests 1.07-fold difference (95% CI = [0.99–1.17], p = 0.09).

Since the non-CpG mutation rate was reported to be higher within regions of high CpG density[27], we also examined if the increased protein-coding mutation rate in MAE genes is entirely driven by hypermutable CpG di-nucleotides. When we exclude CpG sites, per-gene mutation rates derived from divergence data show statistically significant but small increases of 1.03-fold across coding regions ($p<10^{-4}$) and 1.06-fold in 4-fold degenerate sites ($p<10^{-4}$) (**Fig.2b** and **Supplementary Table 5**).

To determine whether increased nucleotide diversity in coding regions of MAE genes can be explained entirely by high CpG content, we compared $\boldsymbol{\pi}$ values in non-CpG-prone sites adjusting for 1.06-fold difference in non-CpG mutation rates. The difference between MAE and BAE genes remained highly significant ($\boldsymbol{\pi} = 6.2\times10^{-4}\pm4.8\times10^{-5}$ for MAE genes in the global population, $5.1\times10^{-4}\pm3.5\times10^{-5}$ for BAE genes, $p < 5\times10^{-4}$, see **Fig.1c**). This suggests that differences in raw mutation rat e are not sufficient to account for the observed differences in nucleotide diversity.

As an additional gauge of the role of mutation rate in the increased variation in MAE genes, we assessed allele frequency distributions for SNPs in MAE and BAE coding sequences. By dividing variants into decile bins of allele frequency and noting the fraction of each decile representing neutral alleles, we found that MAE genes showed a shift of allele frequency distribution towards common alleles in all populations combined ($p < 10^{-20}$**, Fig.2c**), as well as in individual analyzed populations (**Table 1** and **Supplementary Fig.7**). The shift in allele frequency distribution between MAE and BAE genes persisted in all functional categories of sites, including four-fold degenerate sites. Importantly, it is well established[28] that a difference in mutation rates cannot lead to a shift in the distribution of derived allele frequencies, as we observe for the MAE and BAE genes.

We next specifically assessed the contribution of local recombination rate. Nucleotide diversity is correlated with local recombination rate (Begun-Aquadro effect[29]). The proposed explanations for the effect include background selection[30], hitchhiking events[31], and a direct mutagenic effect of recombination[32]. As reported earlier, MAE genes tend to be associated with a local recombination rate that is higher than that of BAE genes[33]. This

observation holds in our much larger MAE and BAE gene sets ($p < 3\times10^{-54}$, **Supplementary Fig.8**). We thus tested if the differences in the local recombination rate (***r***) could explain the increased nucleotide diversity in MAE genes. Using the deCODE pedigree-based recombination map[34], we divided 8,261 informative genes into eight equipopulated ranges of ***r***, with ~1,000 genes per bin, and compared π for MAE and BAE genes within the same range (**Supplementary Table 6**). For 291 MAE genes and 742 BAE genes in the bin with the lowest recombination rate ($r \leq 0.21$ cM/Mb; mean r=0.11 for either group of genes), the difference in π remains, and it also remains after additional exclusion of CpG-prone sites and correction for non-CpG mutation rates, showing that it is also independent of local mutation rate ($p = 2.2\times10^{-4}$, **Fig.2d**).

While the difference in π between MAE and BAE genes appears to be greater in the regions of lower recombination rate, the effect remains at higher recombination rates. In order to boost stat istical power (only 22% of four-fold degenerate SNPs are not CpG-prone), we calculated π using all SNPs, with CpG and non-CpG SNPs combined, and directly corrected for mutation rate difference using divergence-based mutation rate estimates. Further, we controlled for underestimation of π due to lower sequencing depth in MAE by analyzing only the sites passing the strict filter on read depth[17]. Four-fold degenerate sites with lower coverage (average read coverage of 50% below the genomic average) were significant ly enriched in MAE genes (12.1 and 5.5% for MAE and BAE, respectively, $p < 10^{-15}$). When we recalculated π in a subset of sites passing the strict filter on read depth and directly controlled for mutational biases over CpG and non-CpG SNPs combined, π was significantly elevated for MAE over all ***r*** bins ($\Delta\pi = 7.4\times10^{-5} \pm 5.2\times10^{-5}$, p = 0.0037) and remained significant even when we excluded the lowest ***r*** bin ($r \geq 0.21$, p = 0.011, **Supplementary Table 7**, see Methods for details).

In sum, we observe that MAE and BAE genes systematically differ in recombination rate and in CpG content, leading to differences in mutation rate. Importantly, however, while these factors contribute to the elevated diversity of MAE genes, our data argue that they are not able to fully account for it.

Genetic variation is older in MAE genes

Since MAE genes showed increased nucleotide diversity and a shift in allele frequencies in synonymous sites, we examined whet her variation in MAE genes is likely to be, on average, older. We assessed the relative ages of the variants associated with MAE and BAE genes, using Neighborhood-based Clock (NC) analysis[35] on the GoNL data. This analysis is independent of the shift in allele frequency distribution and provides a complementary statistic for evaluation of relative allelic ages. To ensure that overall differences in recombination rates and in allele frequencies did not play a major role in the comparison, we further refined the set of assessed variants as follows. We performed the analysis conditional on local recombination rate by dividing the variants into decile bins by derived allele frequency and separately analyzing genes within each bin (**Supplementary Table 8**). For a modest local recombination rate (less than 0.5 cM/Mb), we found that in every derived allele frequency bin, the NC scores of MAE-associated variants were lower than those of BAE-associated variants ($p < 7.1\times10^{-7}$, **Fig.3a)**, indicating that MAE variants were older in age.

Mindful that allelic age analysis can be confounded by systematic differences between MAE and BAE genes in CpG content, we also used locus-specific estimates of time to the most recent common ancestor (T_{MRCA}) to directly address the question of the age of the variation. T_{MRCA} estimates were obtained by computing the Ancestral Recombination Graph (ARG) on the Complete Genomics dataset[36]. CpG sites were excluded in calculation of T_{MRCA}, safeguarding the analysis from the effect of differences in CpG content. Moreover, variability in mutation and recombination rates between loci was also accounted for in the ARG analysis, including in non-CpG sites, safeguarding the analysis from effect of differences in non-CpG mutation rates[36].

We first confirmed that genetic variation in MAE genes is, on average, older than in BAE genes, as measured by T_{MRCA} (**Supplementary Fig.9**, $p < 2\times10^{-16}$; see Methods). T_{MRCA} is a direct measure of the locus age that allows us to assess the effect of potential confounders. Using T_{MRCA} as the outcome variable, we were able to simultaneously incorporate the effects of multiple confounding variables in a multivariate regression model. Controlling for the level of gene expression, breadth of expression across tissues, selective constraint of the gene, gene length and recombination rate, we confirmed that MAE status remained a significant predictor of older T_{MRCA} ($p = 7.5\times10^{-8}$, **Supplementary Table 9, Supplementary Fig.10**). To be conservative, we also tested the effect of adding divergence-based local non-CpG mutation rates to the regression model as a covariate. Predicted MAE status remained significantly correlated with older T_{MRCA} ($p = 6.8\times10^{-7}$), and the regression coefficient decreased only slightly from 0.054 to 0.051 (5.6%).

Our observations suggest that genetic variation is not only higher, but on average also older in MAE genes compared to BAE genes.

Indications of balancing selection among MAE genes

One of the evolutionary mechanisms that maintain long-term genetic diversity is balancing selection. We thus next examined whether genes thought to be under balancing selection are preferentially MAE. The MAE and BAE gene sets we assessed excluded some well-known examples of such genes (e.g. taste receptors and the extended MHC region; see Methods). We found that genes encoding for extra-cellular matrix molecules, a functional category that has previously been reported to be associated with balancing selection[37], were very strongly enriched for MAE genes (8.1-fold, $p = 7.5\times10^{-33}$; **Supplementary Fig.4a**). In addition, we found that our main gene sets included 80 other genes reported to be under balancing selection (**Supplementary Table 10**). We detected a strong enrichment of genes classified as MAE in this list (1.75-fold, $p < 10^{-4}$; **Fig.3b**).

Ancient balancing selection can leave a trace in the genomes in the form of trans-species polymorphisms (TSPs). A recent analysis[38] suggested that some of the polymorphic variants segregating in both human and chimpanzee populations may evolve under strong long-term balancing selection. While most are noncoding, they have been associated with specific genes in human and chimp genomes. We asked if these TSPs (**Supplementary Table 11**) are differently represented in the MAE and BAE gene sets. We found that the set of trans-species polymorphisms is strongly and significantly enriched in MAE genes (OR = 1.89, $p = 6.3\times10^{-6}$, **Fig.4a**). The enrichment was stronger st ill when we required human-chimp TSPs

to be present in the same gene with an old derived allele predating the human-Neanderthal split (**Supplementary Table 12**) but still segregating in human populations as polymorphism (OR = 2.76, p = 8.4×10^{-4}, **Fig.4a**).

One possible confounding factor is that some (perhaps large) fraction of the TSPs could be due to independent re-mutations. To estimate the contribution of re-mutations, we assessed relative enrichment of MAE and BAE genes among trans-species haplotypes, as defined in the same analysis of human-chimp TSPs[38]. The haplotypes, consisting of more than one polymorphism, are fewer in number than are SNPs, but less likely to arise by re-mutation. We found that the enrichment with MAE genes is even stronger, especially when genes less than 20kb away from these haplotypes (where the majority of *cis*-eQTLs are located[39]) were considered (OR = 4.38, p = 0.0015**, Fig.4b**). Although ext ra-cellular matrix proteins have been suggested as a target of balancing selection[37] and are predominantly MAE, that is not attributable for the enrichment of trans-species haplotypes among MAE genes. Only four trans-species haplotypes are around the genes encoding for extra-cellular matrix, and excluding this category of genes did not affect the association between TSPs and MAE (**Supplementary Fig.4**).

Finally, we asked if the putative ancient alleles are likely to be maintained at intermediate allelic frequencies (see Methods for details). Seventeen genes showed the chromatin signature of MAE and evidence of trans-species haplotypes between human and chimpanzee within 20 kb distance from the gene. Strikingly, derived allele frequency spectra at neutral sites in these genes showed a pronounced enrichment at int ermediate frequencies (AFR p = 0.047, ASN = 0.012, AMR = 0.0045, EUR = 0.12, **Fig.4c**), which is consistent with long-term balancing selection.

Discussion

Using several large datasets characterizing human genetic variation, including the 1000 Genomes Project[17] and Exome Sequencing Project[18], we showed that human autosomal genes classified as MAE on the basis of a characterist ic gene-body chromatin signature[5] have considerably higher nucleotide diversity (π) than do biallelic genes (see **Fig.1**). While the chromatin signature shows remarkable consistency across different genetic backgrounds (**Supplementary Fig.11**), some type I and type II errors are expected. Note that this increase was observed even though the identification of the MAE and BAE gene sets on the basis of the chromatin signatures is subject to occasional misclassification of individual genes.

We examined several possible explanations for the higher nucleotide diversity observed in the MAE gene set. Our results indicate that it does not appear to result from relaxed purifying selection. We show that MAE genes have, on average, increased recombination rates and elevated density of hypermutable contexts contributing to the higher allelic diversity. However, these factors alone do not provide a sufficient explanation. Intriguingly, several lines of evidence from our studies point to the greater overall influence of balancing selection on the MAE genes as a group than on BAE genes (**Fig.3,4**). Gene classes thought to evolve under balancing selection are preferentially MAE; frequency distributions of putatively neutral alleles in MAE genes are shifted towards common variation; variation in

MAE genes is, on average, older; and trans-species polymorphisms preferentially co-localize with MAE genes. We conclude that monoallelic expression (MAE) is associated with higher population genetic diversity, mediated by increased mutation and recombination rates and, for a fraction of MAE genes, by balancing selection.

In this context, we speculate that heterozygote advantage might be associated with MAE (also see[3,14-16]). In particular, heterogeneity involving cells of the same type, likely increased in individuals heterozygous for MAE genes whose alleles are functionally distinct (**Supplementary Fig.12**). Intriguingly, MAE genes are enriched for proteins present on the cell surface and are responsible for interactions between the cell and its environment, which includes other cells, signaling molecules, and pathogens. Elevated cell-to-cell diversity is the opposite of the uniformity of a “monoculture”; it should, for example, reduce susceptibility of a tissue as a whole to infectious agents. Such a general adaptive role for MAE would be consistent with increased allelic diversity that is widespread in human populations rather than limited to particular environments or geographical locations. Since MAE genes are enriched with particular functional categories, high nucleotide diversity and MAE might be two separate but interacting phenomena which jointly affect cell diversity within a tissue by targeting the same molecular components.

Recently, theoretical models and genome-scale data analyses have revived a dormant interest in balancing selection and in the issue of overdominance and dominance generally[38,40-42]. The findings we report here support the idea that balancing selection can have a discernible effect on a large group of genes.

Methods

Datasets

Genes were classified as MAE or BAE using specific chromatin signature[5] (co-occurrence of H3K27me3 silencing mark and H3K36me3 active mark on the gene body). Note that we focus on mitotically stable MAE, likely observable in fewer genes than stochastic transcription bursts that could be detected by single-cell RNA sequencing[6,44]. We used the following: GM12878 (lymphoblastoid cells), K562 (myeloid cells), H1ESC (embryonic stem cells), HSMM (skeletal muscle myocytes), HUVEC (umbilical vascular epithelium) and HMEC/HCC1954 (mammary epithelium). To consider a gene MAE, we required monoallelic status in at least one cell line with expression level of RPKM≥1 in that cell line. To consider a gene BAE, we required the absence of monoallelic status in all cell lines where the gene was detected (with RPKM>0). For example, if a gene was monoallelic only at RPKM<1, it was not included in the MAE set, nor was it considered to be positively BAE, and was therefore excluded from consideration. Genes not sharing the MAE chromatin signature, were not counted as MAE. Note that the average fraction of MAE genes per cell line is ~10%, with values ranging from ~4% to 15% (**Supplementary Table 1; Suppl. Fig. 3**).

We excluded: genes that do not uniquely map by name to known Ensembl protein-coding genes (v74); microRNA genes, because the chromatin signature is known to be less accurate for shorter genes[22]; pseudogenes and genes that do not map to primary autosomal super-

contigs. Further, we excluded olfactory receptors, taste receptors, toll-like receptors and HLA genes, which are already known to exhibit both high genetic diversity and MAE. We also eliminated the entire MHC region surrounding the HLA genes (chr6:28,000,000-34,000,000) since the signature of long-term balancing selection extends over neighboring genes. The resulting gene set contained 10,233 genes of which 4,227 were MAE and 6,006 were BAE (**Supplementary Table 1**). We refer to that filtered set of genes as "genome-wide dataset".

Analyses of genetic variation were primarily done on the 1000 Genomes Project Phase 1 data[17], encompassing 1,092 individuals from four super-populations: African (AFR, N=246), European (EUR, N=379), Admixed American (AMR, N=181) and Asian (ASN, N=286). We also examined protein-coding variants in 4,300 European Americans and 2,203 African Americans in the Exome Sequencing Project dataset (ESP6500SI-V2)[18]. In addition, the Genome of the Netherlands (GoNL)[45] dataset was used for *de novo* mutation rate estimation and allelic age analysis. The GoNL dataset consists of phased whole-genome sequences of 250 Dutch parent-child trios and genome-wide collection of 11,020 *de novo* mutations identified in the offspring.

The candidates for trans-species polymorphisms (TSPs) were obtained from published data[38]. Briefly, this is a set of protein-coding SNPs observed in both sub-Saharan African humans and Western chimpanzees. This dataset also includes the smaller set of mostly noncoding trans-species haplotypes defined by two or more trans-species SNPs within 4kb distance and in shared linkage disequilibrium (LD) structure. For intergenic trans-species haplotypes, the authors selected the gene at closest distance from the haplotype as a probable target of balancing selection. Of the genes predicted as either MAE or BAE, 60 genes were identified by trans-species haplotypes, and an additional 141 genes were identified by protein-coding TSPs.

To enrich for true human-chimp TSPs, we utilized the genome sequence of a single Neanderthal individual from a cave in the Altai mountains[46]. The genome was sequenced at ~52x, with autosomal contamination estimated between 0.8–1.2%. We used a population of sub-Saharan Africans (YRI, N=88) to identify ancient genetic variation predating the human-Neanderthal split. There is no evidence of gene flow between Neanderthals and YRI.

For all analyses, ancestral alleles were distinguished from derived alleles based on EPO multiple-sequence alignments (available from the 1000 Genomes project). Only the SNPs with high confidence on predicted ancestral alleles were analyzed.

Nucleotide diversity

We estimated the nucleotide diversity π[23] for MAE and BAE in humans by analyzing single nucleotide polymorphisms (SNPs) in neutral or all sites of protein-coding regions. The neutral π was calculated in four-fold degenerate sites with polymorphism data from the 1000 Genomes. To rule out the possibility that the biased distribution of hyper-mutable CpG di-nucleotides explain the difference of π in MAE and BAE, we further computed the neutral π using only non-CpG-prone sites, which are defined by the nucleotides that are not preceded by C or followed by G therefore do not overlap with CpG. For the non-CpG π, we

always report π adjusted for the variation in non-CpG mutation rates across the genome; non-CpG π was scaled down by 1.02-fold and up by 1.04-fold for MAE and BAE, respectively, based on divergence-based mutation rates. For all-site π, all SNPs in protein-coding regions were analyzed using ESP data as well as the 1000 Genomes. For ESP data, we derived all-site π from per-gene π and observed length of each gene. For the 1000 Genomes, we annotated SNPs with the change of amino acids in canonical transcripts using Variant Effect Predictor. The canonical transcript was defined as the transcript producing the longest known protein-coding sequence. We inferred the 95% confidence intervals of π by bootstrap sampling of genes (N=10,000).

Mutation rate

We computed the mutation rates in protein-coding and intronic regions of MAE and BAE genes utilizing 11,020 *de novo* point mutations from 269 GoNL offspring[45] (**Supplementary Table 5**). MAE and BAE genes contain 32 and 50 mutations in protein-coding regions and 1,170 and 1,102 in intronic regions, respectively. In order to control for local variation of the power to detect *de novo* events, we estimated detection power (between 0 and 1) from simulated positive controls (kindly provided by Laurent C. Francioli)[45]: sets of artificial *de novo* mutations spiked in at 1,811 protein-coding and 58,329 intronic sites randomly sampled from our genic regions, processed by the identical *de novo* mutation detection software.

Assuming that *de novo* mutation events follow a Poisson process, we tested the following null hypothesis using the Exact Poisson Test:

$$\Theta_{MAE} \sim Poisson\,(\lambda{=}269 \quad \mu \quad \tau_{MAE} \quad P_{MAE}) \quad and \quad \Theta_{BAE} \sim Poisson\,(\lambda{=}269 \quad \mu \quad \tau_{BAE} \quad P_{BAE})$$

where 269 is the number of offspring, Θ is the number of *de novo* mutation events observed, μ is the diploid mutation rate per generation per nucleotide, τ is the length of mutational target, and P is the estimated mean detection power. The null model assumes the equal mutation rate μ across MAE and BAE. We tested for the unequal mutation rates excluding as well as including CpG di-nucleotides in the mutational target since CpGs are more frequent in MAE than in BAE genes.

Due to the small number of observed *de novo* mutations in GoNL, we further examined the mutation rate map constructed from human-chimpanzee divergence and observed patterns of *de novo* mutations in GoNL[45]. Briefly, a context-dependent substitution rate matrix was inferred for each of 1Mb genomic blocks from human-chimpanzee sequence alignments. Then, we corrected for deviation of substitution rates from the patterns of observed *de novo* mutations, specifically the biases due to local recombination rates, types of mutations, and transcription strand. Using this local mutation rate map, we derived mutation rates of protein-coding regions and introns and tested for the difference in mutation rates between MAE and BAE regions by bootstrap resampling of MAE and BAE genes (N=10,000).

Recombination rate

To test if the higher neutral π in MAE is due to the difference in local recombination rates (***r***), we examined the recombination rates around MAE and BAE genes on the latest pedigree-based genetic map of the Icelandic population (sex-averaged deCODE map)[34]. For each gene, ***r*** was defined by an average rate across 410-kb region centered at the midpoint of the gene. The window size was chosen as in a previous study of the Begun-Aquadro effect in humans[47]. We annotated ***r*** for a total of 3,281 MAE and 4,980 BAE genes, and grouped genes into eight equal-sized bins by ***r***. In each bin, we calculated the p-value of test for significant difference in π between MAE and BAE by bootstrap sampling of genes (N=100,000); see **Supplementary Table 6**. The per-bin P-values were combined by Fisher's method. We analyzed non-CpG π similarly in order to control for both recombination and mutation rat es at the same time.

To improve statistical power for the analysis of π, we tried an alternative strategy to correct for the variation of mutation rates. Instead of estimating π only in non-CpG-prone sites, which constitute only 22% of four-fold degenerate sites, we used all 4FD sites (both CpG-prone and not) to estimate neutral π and then cancelled out mut ation rate bias by utilizing the divergence-based mutation rate map. The mutational rate bias was calculated for each bin of ***r*** separately to account for the variation of sequence composition by ***r***. Furthermore, we excluded four-fold sites of too low sequencing coverage as they are enriched in MAE genes (12.1% compared to 5.5% of BAE) and lead to underestimation of π due to diminished SNP detection power. Specifically, we used only the whole genome sequencing data of the 1000 genomes project and applied the “strict mask” filter on sequencing depth[17]. The overall difference of π ($\Delta\pi$) and its 95% confidence interval were calculated by variance-normalized meta-analysis across ***r*** bins. The variance of $\Delta\pi$ in each bin was estimated by bootstrapping.

Site frequency spectra

We calculated derived site frequency spectra (SFS) of SNPs in coding regions of MAE and BAE genes using the 1000 Genomes dataset. Only SNPs polymorphic in each individual population were used for the analysis. For neutral SFS, we used SNPs in four-fold degenerate sites, and for all-site SFS, we stratified SNPs by amino acid changes and their functional impact predicted by PolyPhen-2[48]. To test for the significant difference in SFS between MAE and BAE, we subdivided the SNPs into high and low frequency bins, which were cut at the allele frequency of 10%, and applied a χ^2 two-proportion test. Frequencies approaching fixation (>90%) were excluded from the analysis.

Purifying selection

The strength of purifying selection on MAE and BAE was compared using two gene-level datasets: OMIM MobidMap (http://omim.org) and selectively constrained genes[26]. MorbidMap provides a list of genes that are known to cause Mendelian genetic disorders in humans whereas the constrained gene set, which is defined by the depletion of missense polymorphism in ESP compared to the expected mutation rates, allows a more comprehensive and unbiased survey of selective constraints on genes although the selective pressure does not necessarily imply severe morbidity. For MorbidMap, we associated 3,037

autosomal genes with MIM disorder IDs by matching gene names between Ensembl and MorbidMap. Out of the 1,003 top constrained genes[26], we mapped RefSeq IDs of 990 genes to Ensembl, excluding genes with missing RefSeq or incongruent chromosome. We used Fisher's exact test to compare the difference in strength of purifying selection on MAE and BAE. For the top constrained genes, we further confirmed that the degree of selective constraints is not significantly different between 252 constrained MAE and 364 constrained BAE genes by comparing the distribution of signed Z scores[26] (Wilcoxon rank-sum test, P=0.65). To compare the selective constraints in more weakly constrained MAE and BAE genes, we subdivided 3,609 MAE and 5,191 BAE genes annotated with signed Z scores into eight Z score bins and tested for the relative enrichment of MAE genes in each bin by Fisher's exact test (see **Supplementary Table 4**).

To examine the subtle difference of selective pressure that is difficult to identify in polymorphism-based data, we compared the group-wise d_N/d_S[49] between MAE and BAE gene sets. The numbers of nonsynonymous and synonymous substitutions and sites were aggregated over MAE and BAE genes, and then, the overall nonsynonymous substitutions per nonsynonymous site (d_N) and synonymous substitutions per synonymous site (d_S)[49] were calculated for MAE and BAE (see **Supplementary Table 3**). 95% Confidence intervals were computed by bootstrapping (N=1,000). We also compared the distribution of per-gene d_N/d_S for 2,021 MAE and 3,223 BAE genes after excluding genes with no synonymous substitution (Wilcoxon rank-sum test, p = 0.09).

NC-clock

To compare the allelic age of synonymous SNPs within MAE and BAE genes, we applied the Neighborhood-based Clock (NC) algorithm[35] to 498 unrelated GoNL samples. Briefly, the NC test statistic estimates the allelic age of each variant by computing the physical distance to the closest recombination or fully linked mutation event. Only non-singleton variants were analyzed, and SNPs with unphased genotypes were excluded from the analysis.

In order to control for the effect of variation in local recombination rates, we grouped MAE and BAE genes into recombination rate intervals (**Supplementary Table 6**). Here, the local recombination rates were defined by the rate across 10kb windows containing the test SNP on the deCODE sex-averaged genetic map. The window size of 10kb was selected to match the scale of NC estimates for common tested SNPs. For each recombination rate bin, variants were further binned by derived allele frequency (in 10% intervals). For each bin, we tested whether synonymous SNPs in BAE were significantly younger than those in MAE by one-sided Wilcoxon rank-sum test. P-values were combined across allele frequency bins by meta-analysis using Stouffer's Z-score method, weighted by sample size (**Supplementary Fig.13 and Supplementary Table 8**).

Time to most recent common ancestor (T_{MRCA})

We conducted a multivariate regression analysis to study the correlation between a gene's monoallelic expression status and its T_{MRCA} in presence of confounding variables. For each gene, mean T_{MRCA} over the entire transcribed region was calculated from genome-wide

T_{MRCA} estimates generated by running ARGWeaver on Complete Genomics data[36]. The log-transformed T_{MRCA} was regressed under the following model:

$$log\,(T_{MRCA}) = \beta_0 + \beta_1\, I_{MAE/BAE} + \beta_2\, length + \beta_3\, r + \beta_4\, exprlevel + \beta_5\, exprbreadth + \beta_6 Z$$

where $I_{MAE/BAE}$ is an indicator variable for MAE (MAE=1 and BAE=0), *length* is the length of the canonical transcript, r is the local recombination rate (based on the deCode sex-averaged map, averaged over 410kb windows), *exprlevel* is the gene expression level (taken as the highest expression level of the gene as measured by its RPKM value in cell types with expression as indicated by $I_{MAE/BAE}$)[5], *exprbreath* is the gene expression breadth (scores between 0 and 1 for tissue specificity across 12 human tissues; 0=house-keeping, 1=tissue-specific)[50], and Z is the selective constraint[26]. T_{MRCA} is unlikely to be confounded by mutation rate variation since 1) CpG di-nucleotides were excluded from the analysis[36] and 2) ARGWeaver accounted for local variation in non-CpG mutation and recombination rates. The transcript-specific expression breadth score was summarized into a gene-level score by choosing the breadth of the most ubiquitously expressed alternative transcript. However, our results are robust to alternative measures: expression breath of the least ubiquitously expressed transcript and the mean breadth across all alternative transcripts.

To examine if the signal is only coming from a small number of genes annotated with the lowest recombination rates (***r*** < 0.21 cM/Mb), we also tested our model after excluding genes in that bin. The MAE status remains significantly correlated with older T_{MRCA} ($p=1.2\times10^{-10}$).

To test whether there is any additional signal when MAE is detected in multiple tissues, we added one more variable MAE_m to the model:

$$log\,(T_{MRCA}) = \beta_0 + \beta_1\, I_{MAE/BAE} + \beta_m\, MAE_m + covariates$$

where MAE_m is defined as the number of MAE tissues minus 1 if MAE was detected in multiple tissues and 0 otherwise. We found that the coefficient of MAE_m was not significantly non-zero (p = 0.49), showing that genes that have MAE signature in multiple tissues do not have longer T_{MRCA} values than genes with the MAE signature in only one tissue.

To ensure that T_{MRCA} is not confounded by the variation in non-CpG mutation rates, we added non-CpG mutation rates across transcribed region, estimated from the divergence-based mutation rate map, as a covariate to the multivariate regression model. The regression coefficient β_1 of MAE status decreased only slightly from 0.054 to 0.051 (5.6%), and β_1 remained significantly non-zero ($p = 6.8\times10^{-7}$).

Finally, we confirmed that genes that were experimentally established as MAE in human lymphoblasts[3] have significantly older T_{MRCA} values compared to BAE genes in the same set (Wilcoxon rank-sum test, $p = 6.3\times10^{-6}$).

Trans-species polymorphisms

To enrich for the strongest signals of long-term balancing selection, we intersected the genes identified by human-chimp TSPs[38] with genes containing ancient SNPs predating the human-Neanderthal split. Long-term balancing selection acting on TSPs is expected to increase the coalescent time of nearby polymorphisms. Specifically, we looked for derived alleles that are polymorphic in YRI and also present in the Altai Neanderthal genome in one or two copies. To minimize false positives due to re-mutation, we excluded derived alleles in CpG context. In total, we collected 3,383 ancient protein-coding SNPs (2,603 genes) predating the Neanderthal split. Among those, 104 genes are also associated with human-chimp TSPs, forming the strongest candidates for long-term balancing selection, and 44 of these 104 genes have the chromatin signature of either MAE or BAE (**Supplementary Table 9**).

Next, we examined the influence of three potential confounders on the enrichment of MAE among the genes identified by trans-species haplotypes. First, we controlled for uneven genome-wide distribution of re-mutations using 33,906 SNPs shared between human and chimpanzee across the autosomes ("shared SNPs"). We conservatively assumed that all shared SNPs were false positives due to re-mutation. For this, we downloaded the coordinates of shared SNPs from the authors' website and identified the nearest protein-coding genes (GENCODE-12) to these SNPs as in Leffler et al[38]. Then, the genes identified by shared SNPs were used as the baseline for enrichment test. Second, since house-keeping genes are biased toward BAE and may evolve under distinct regulatory and evolutionary constraints, we re-examined the enrichment of MAE genes in trans-species haplotypes after excluding house-keeping genes. 549 MAE and 2,681 BAE genes were classified as house-keeping, defined by the ubiquitous presence of transcripts and minimal variation of expression levels across all tissues[43]. Third, the identification of candidate genes for balancing selection based on closest-distance to intergenic trans-species haplotypes can be ambiguous, especially if the haplotypes are distant from the gene. Based on a previous observation that the majority of *cis*-eQTLs are located within 20kb from the gene[39], we repeated the enrichment test using only trans-species haplotypes within 20kb from genes ("proximal trans-species haplotypes").

For the 17 MAE genes identified by proximal trans-species haplotypes, we could detect the shift in SFS toward intermediate allelic frequencies. The neutral SFS was compared between the 17 MAE genes and genes lacking proximal trans-species haplotypes. The neutral SFS was generated from derived allelic frequencies of four-fold degenerate variants from the 1000 Genomes data. The significant difference of SFS was tested by χ^2 goodness-of-fit test with combining the frequency above 40% into a single bin due to their small observed counts.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dan Balick for useful discussions and Ivan Adzhubei for help with PolyPhen analysis.

Support: This work was supported in part by the following NIH awards: R01 GM114864 to AAG, and R01 GM078598, GM105857 and MH101244 to SRS. AAG was supported in part by the Pew scholar award; TLL was supported by a fellowship from the German Research Foundation (DFG, LE 2593/1-1 & 2-1); LG was a summer scholar in the Harvard/MIT BIG program (supported by the NIH award U54LM008748); RBM and CtW were supported by grants from the NIH/NIGMS (R01GM61936, 5DP1GM106412) and Harvard Medical School.

References

1. Savova V, Vigneau S, Gimelbrant AA. Autosomal monoallelic expression: genetics of epigenetic diversity? Curr Opin Genet Dev. 2013; 23:642–8. [PubMed: 24075575]
2. Chess A, Simon I, Cedar H, Axel R. Allelic inactivation regulates olfactory receptor gene expression. Cell. 1994; 78:823–34. [PubMed: 8087849]
3. Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. Widespread monoallelic expression on human autosomes. Science. 2007; 318:1136–40. [PubMed: 18006746]
4. Zwemer LM, et al. Autosomal monoallelic expression in the mouse. Genome Biol. 2012; 13:R10. [PubMed: 22348269]
5. Nag A, et al. Chromatin signature of widespread monoallelic expression. Elife. 2013; 2:e01256. [PubMed: 24381246]
6. Deng Q, Ramskold D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science. 2014; 343:193–6. [PubMed: 24408435]
7. Jeffries AR, et al. Stochastic choice of allelic expression in human neural stem cells. Stem Cells. 2012; 30:1938–47. [PubMed: 22714879]
8. Gendrel AV, et al. Developmental dynamics and disease potential of random monoallelic gene expression. Dev Cell. 2014; 28:366–80. [PubMed: 24576422]
9. Eckersley-Maslin MA, et al. Random Monoallelic Gene Expression Increases upon Embryonic Stem Cell Differentiation. Dev Cell. 2014; 28:351–65. [PubMed: 24576421]
10. Li SM, et al. Transcriptome-wide survey of mouse CNS-derived cells reveals monoallelic expression within novel gene families. PLoS One. 2012; 7:e31751. [PubMed: 22384067]
11. Pereira JP, Girard R, Chaby R, Cumano A, Vieira P. Monoallelic expression of the murine gene encoding Toll-like receptor 4. Nat Immunol. 2003; 4:464–70. [PubMed: 12665857]
12. Spencer HG. Population genetics and evolution of genomic imprinting. Annu Rev Genet. 2000; 34:457–477. [PubMed: 11092835]
13. Wilkins JF, Haig D. What good is genomic imprinting: the function of parent- specific gene expression. Nat Rev Genet. 2003; 4:359–68. [PubMed: 12728278]
14. Wu CT, Dunlap JC. Homology effects: the difference between 1 and 2. Adv Genet. 2002; 46:xvii–xxiii. [PubMed: 12136788]
15. Hoehe MR, et al. Multiple haplotype-resolved genomes reveal population patterns of gene and protein diplotypes. Nat Commun. 2014; 5:5569. [PubMed: 25424553]
16. Chess A. Mechanisms and consequences of widespread random monoallelic expression. Nat Rev Genet. 2012; 13:421–8. [PubMed: 22585065]
17. Project Consortium G. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–73. [PubMed: 20981092]
18. Tennessen JA, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012; 337:64–9. [PubMed: 22604720]
19. Drmanac R, et al. Human genome sequencing using unchained base reads on self- assembling DNA nanoarrays. Science. 2010; 327:78–81. [PubMed: 19892942]
20. Francioli LC, et al. Genome-wide patterns and properties of de novo mutations in humans. Nat Genet. 2015; 47:822–6. [PubMed: 25985141]
21. Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

22. Nag A, Vigneau S, Savova V, Zwemer LM, Gimelbrant AA. Chromatin Signature Identifies Monoallelic Gene Expression Across Mammalian Cell Types. G3 (Bethesda). 2015; 5:1713–20. [PubMed: 26092837]
23. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A. 1979; 76:5269–73. [PubMed: 291943]
24. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature. 1977; 267:275–6. [PubMed: 865622]
25. Chamary JV, Hurst LD. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol. 2005; 6:R75. [PubMed: 16168082]
26. Samocha KE, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014; 46:944–50. [PubMed: 25086666]
27. Walser JC, Furano AV. The mutational spectrum of non-CpG DNA varies with CpG content. Genome Res. 2010; 20:875–82. [PubMed: 20498119]
28. Li, W-H. Molecular evolution. Sinauer Associates; Sunderland, MA: 1997.
29. Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature. 1992; 356:519–20. [PubMed: 1560824]
30. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993; 134:1289–303. [PubMed: 8375663]
31. Maynard-Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974; 23:23–35. [PubMed: 4407212]
32. Hellmann I, et al. Why do human diversity levels vary at a megabase scale? Genome Res. 2005; 15:1222–31. [PubMed: 16140990]
33. Necsulea A, Semon M, Duret L, Hurst LD. Monoallelic expression and tissue specificity are associated with high crossover rates. Trends Genet. 2009; 25:519–22. [PubMed: 19850368]
34. Kong A, et al. Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 2010; 467:1099–103. [PubMed: 20981099]
35. Kiezun A, et al. Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. PLoS Genet. 2013; 9:e1003301. [PubMed: 23468643]
36. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-wide inference of ancestral recombination graphs. PLoS Genet. 2014; 10:e1004342. [PubMed: 24831947]
37. Andres AM, et al. Targets of balancing selection in the human genome. Mol Biol Evol. 2009; 26:2755–64. [PubMed: 19713326]
38. Leffler EM, et al. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science. 2013; 339:1578–82. [PubMed: 23413192]
39. Veyrieras JB, et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. 2008; 4:e1000214. [PubMed: 18846210]
40. Sellis D, Callahan BJ, Petrov DA, Messer PW. Heterozygote advantage as a natural consequence of adaptation in diploids. Proc Natl Acad Sci U S A. 2011; 108:20666–71. [PubMed: 22143780]
41. DeGiorgio M, Lohmueller KE, Nielsen R. A model-based approach for identifying signatures of ancient balancing selection in genetic data. PLoS Genet. 2014; 10:e1004561. [PubMed: 25144706]
42. Yang S, et al. Parent-progeny sequencing indicates higher mutation rates in heterozygotes. Nature. 2015; 523:463–7. [PubMed: 26176923]
43. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet. 2013; 29:569–74. [PubMed: 23810203]

References for Methods

44. Borel C, et al. Biased allelic expression in human primary fibroblast single cells. Am J Hum Genet. 2015; 96:70–80. [PubMed: 25557783]
45. Francioli LC, et al. Genome-wide pat terns and properties of de novo mut ations in humans. Nat Genet. 2015
46. Prufer K, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014; 505:43–9. [PubMed: 24352235]

47. Cai JJ, Macpherson JM, Sella G, Pet rov DA. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS Genet. 2009; 5:e1000336. [PubMed: 19148272]

48. Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–9. [PubMed: 20354512]

49. Bustamante CD, et al. Natural selection on protein-coding genes in the human genome. Nature. 2005; 437:1153–7. [PubMed: 16237444]

50. Yanai I, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 2005; 21:650–9. [PubMed: 15388519]

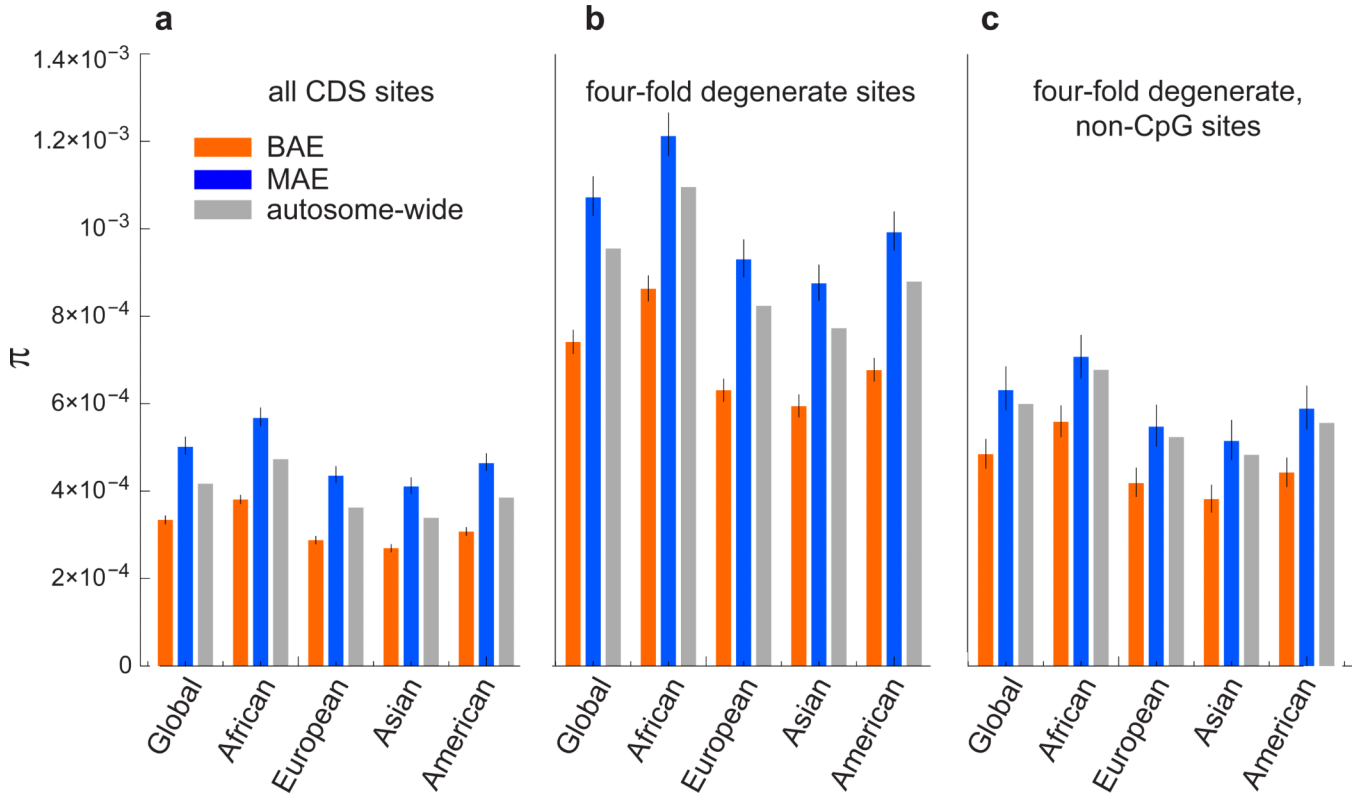

Figure 1. Nucleotide diversity is higher in MAE genes

a. Average nucleotide diversity (***π***) for MAE and BAE genes in the 1,000 Genomes dataset (global), and all four continental groups: African, European, Asian and American. ***π*** is calculated for the coding regions (CDS), including all sites. Error bars represent 95% confidence intervals calculated by bootstrapping. *Orange* - BAE genes; *blue* - MAE genes; *grey* shows data calculated for all autosomal genes.

b. As in (a), calculated on four-fold degenerate sites only.

c. As in (b), excluding CpG-prone sites (all sites preceded by C or followed by G) from the calculation of nucleotide diversity. ***π*** was adjusted for the difference in mutation rates on non-CpG-prone four-fold degenerate sites estimated from a mutational model in ref.[20].

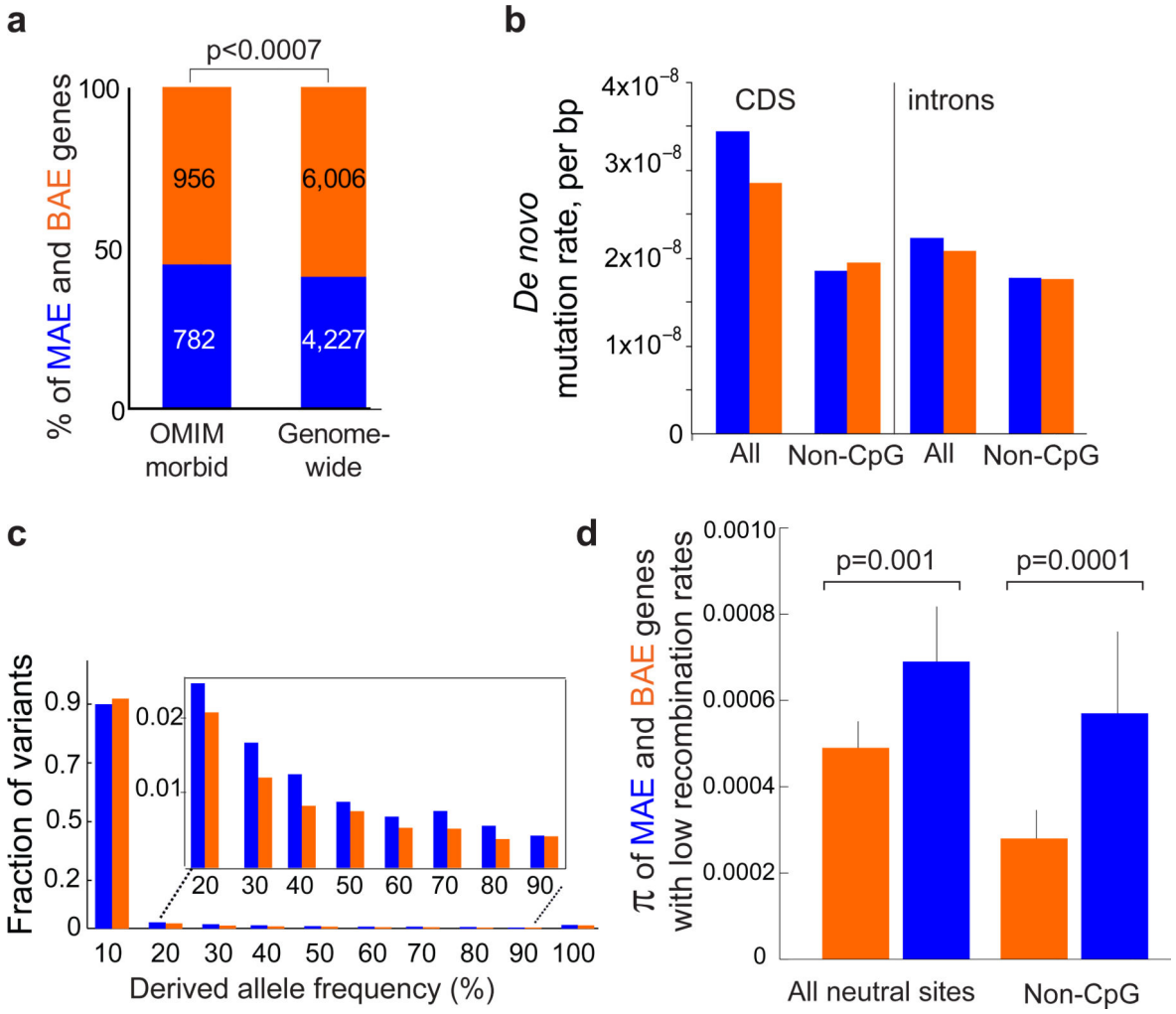

Figure 2. Purifying selection, mutation rates and recombination as potential sources of genetic diversity in MAE genes

a. MAE and BAE genes among genes known to cause Mendelian diseases extracted from the OMIM database (OMIM MorbidMap) and genome-wide. Within each bar numbers of genes in each category are shown. Here and elsewhere, MAE data shown in *blue*, BAE data in *orange*; p-value from Fisher's exact test.

b. Average *de novo* per-base diploid mutation rate for MAE and BAE genes from whole-genome sequences of GoNL parent -child trios. Left: mutation rate estimated from 82 *de novo* mutations in the coding regions: (*all*) including CpG sites; (*non-CpG*) excluding CpG sites. Right: same, est imated from 2,272 *de novo* mutations in intronic regions.

c. Site frequency spectrum for derived alleles in MAE and BAE genes in the 1000 Genomes dataset. Shown is fraction of variants for neutral (four-fold degenerate) alleles with a given derived allelic frequency, in bins of ten. Inset shows a close-up of high allelic frequencies (between 10% and 90%).

d. Average nucleotide diversity (***π***) for ~1,000 genes [both MAE (*blue*) and BAE (*orange*)] with the lowest local recombination rate (***r*** ≤ 0.21 cM/Mb) using data from the 1000 Genomes global population. *Left* - all neutral (four-fold degenerate) sites were used for

calculation. *Right* - same, but excluding CpG-prone sites from the calculation of nucleotide diversity and accounting for 1.06-fold difference in non-CpG mutation rates. Error bars show 95% CI. Analysis for other ranges of ***r*** can be found in **Supplementary Table 6**.

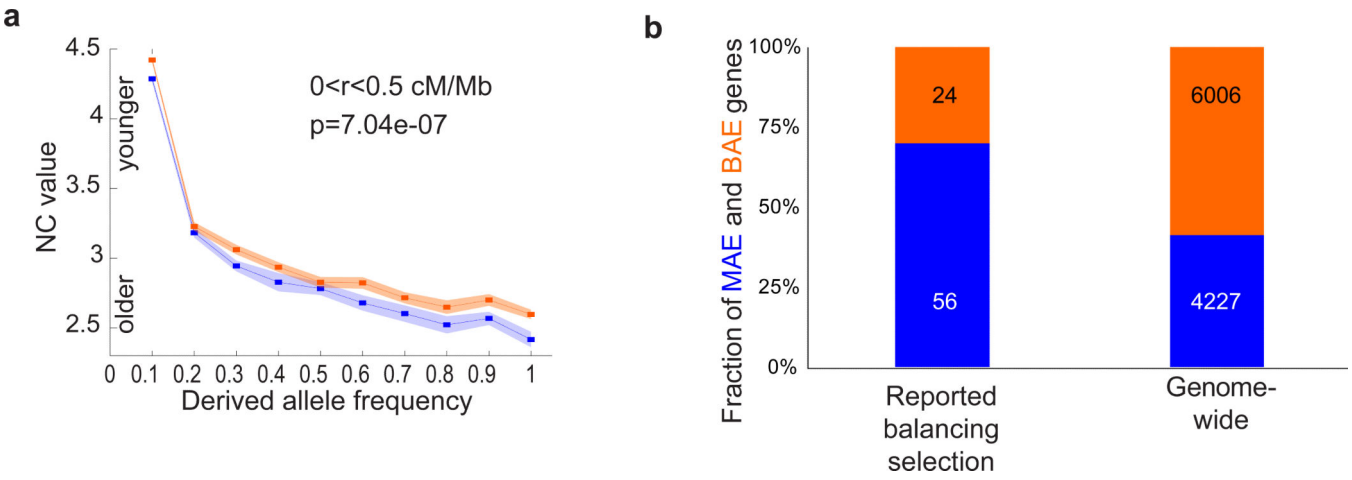

Figure 3. Ancient variants and genes under balancing selection are enriched among MAE genes

a. Percent MAE (*blue*) and BAE (*orange*) among genes thought to be under balancing selection (see **Supplementary Table 10**) compared to the genome-wide dataset (Pearson's χ^2 test , $p < 3.9\times10^{-7}$).

b. Allelic age of synonymous SNPs in MAE and BAE genes estimated by applying the Neighborhood-based Clock (NC) method[35] to genomes sequenced by the GoNL project. NC values are plotted for synonymous SNPs in MAE genes (*blue*) and BAE genes (*orange*) as a function of derived allele frequency (10% bins). Error bars show standard error of the mean. Analysis was limited to variants associated with local recombination rat es between 0 and 0.5 cM/Mb. For other ranges of recombination rates, see **Supplementary Table 6**. Combined P-value calculated across all derived allele frequency bins is reported. See Methods for details.

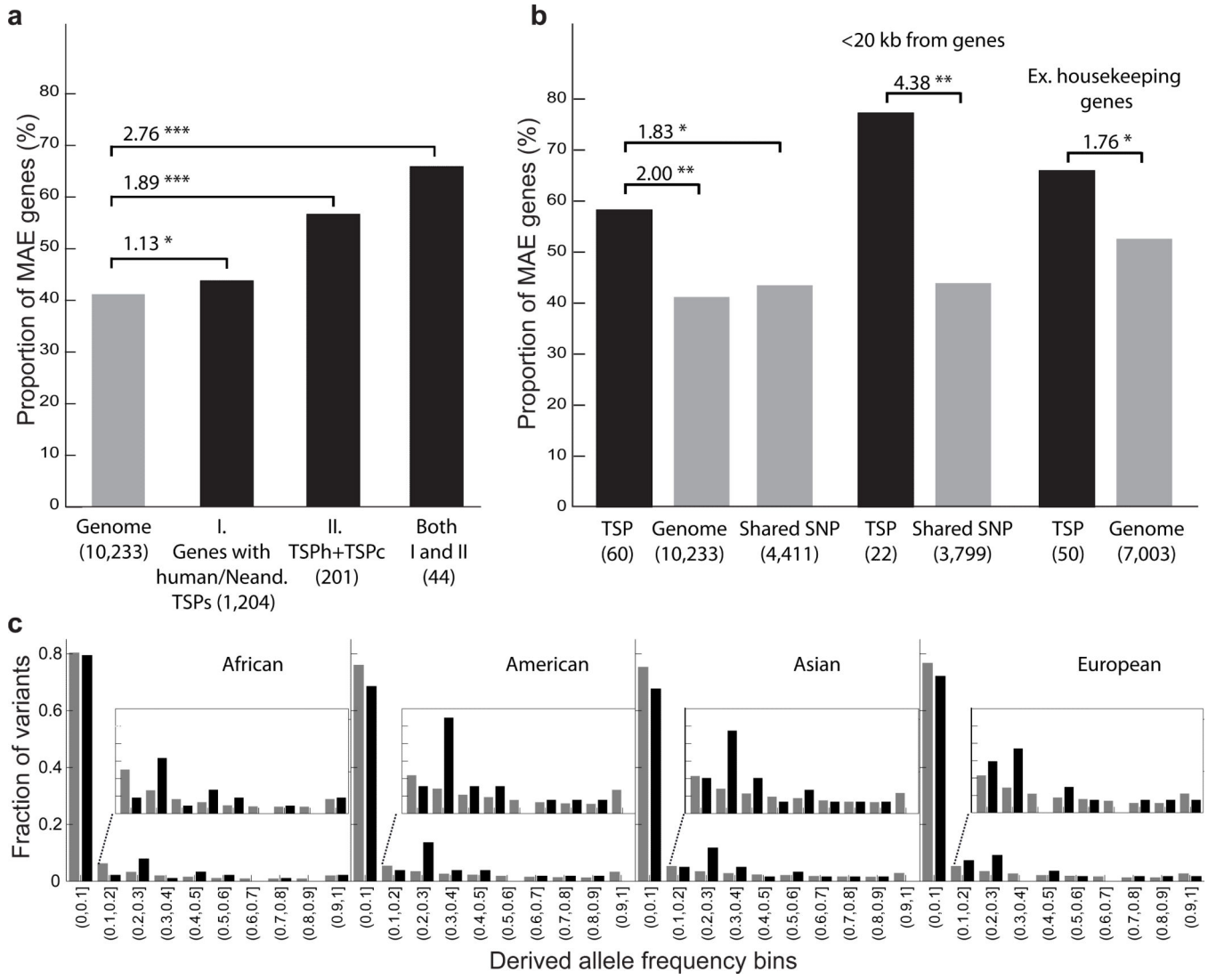

Figure 4. Trans-species polymorphisms are enriched among MAE genes

a. Percent MAE genes with application of the Neanderthal filter to candidates of human-chimpanzee trans-species polymorphisms. *Gray*, the genome-wide dat aset; *Black*: (I): Genes with ancient SNPs shared with Neanderthals – genes harboring at least one ancient protein-coding SNP predating the human-Neanderthal split. (II): TSP_h+TSP_c – genes with any evidence of trans-species polymorphisms between human and chimpanzee[38]. (III): Both I and II: i.e., as in II, after applying the Neanderthal filter. The number of genes per category is shown below each group label. Odds ratios and their significance levels are reported (*p<0.05, ***p<0.001).

b. Percent MAE genes among human-chimpanzee trans-species haplotypes (TSP, *black*) and control datasets (*gray*): the genome-wide dat aset (Genome) and the set of genes adjacent to SNPs segregating in both species identically by state (Shared SNP) as a control for uneven density of recurrent mut ations. Left: Data for all haplotypes; Center: data for haplotypes less than 20kb from genes; Right: as left, excluding housekeeping genes, defined by ubiquitous and low-variance expression across tissues[43]. The number of genes per category

is shown below group label. Odds ratios and their significance levels are reported (*p<0.05, **p<0.01). See Methods for details.

c. Site frequency spectra for derived alleles in MAE genes which also have trans-species haplotypes between human and chimpanzee within 20kb distance from the gene (17 genes total, *black*) compared to all genes lacking trans-species haplotypes (*gray*). Insets increase vertical axis resolution. All SNPs are at 4-fold degenerate sites; allele frequencies from the 1,000 Genomes project.

Table 1

P-values in Pearson's χ^2 test for a significant shift toward common frequency in MAE compared to BAE in the global 1,000 Genomes dataset, and all four continental groups: African, European, Asian and American.

Population	Synonymous (4FD)	Missense damaging	Missense benign
Global	$< 10^{-20}$	3.0×10^{-5}	1.4×10^{-11}
African	4.2×10^{-8}	2.7×10^{-2}	4.8×10^{-11}
American	2.7×10^{-11}	6.5×10^{-4}	8.0×10^{-4}
European	1.1×10^{-13}	3.4×10^{-3}	6.6×10^{-7}
Asian	2.2×10^{-16}	2.1×10^{-5}	8.2×10^{-9}