nature genetics

Genes with monoallelic expression contribute disproportionately to genetic diversity in humans

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

An unexpectedly large number of human autosomal genes are subject to monoallelic expression (MAE). Our analysis of 4,227 such genes uncovers surprisingly high genetic variation across human populations. This increased diversity is unlikely to reflect relaxed purifying selection. Remarkably, MAE genes exhibit an elevated recombination rate and an 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: variants in these genes tend to be older and are enriched in polymorphisms shared by Neanderthals and chimpanzees. Both synonymous and nonsynonymous alleles of 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 the generation of cell-to-cell heterogeneity; the increased genetic variation contributes to this heterogeneity.

Among the epigenetic regulatory modes causing unequal allelic transcription of mammalian autosomal genes, by far the most widespread is MAE, with mitotically stable maintenance of the initial random choice of an active allele[1]. Although 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 cell type assessed, 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 at 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 MAE, imprinting, predict that deleterious allelic variation in such genes would be more efficiently removed by purifying selection[12,13]. Similar 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). At the same time, in contrast to imprinted genes, 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 in 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]. Unexpectedly, 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 a somewhat elevated recombination rate and an 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 previously used Encyclopedia of DNA Elements (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 that is capable of reliably identifying MAE in polyclonal cell lines. By choosing this data set as a starting point, we deliberately limit ourselves to mitotically stable MAE (Online Methods).

Because MAE is largely a tissue-specific phenomenon and we are interested in the evolutionary forces acting on the entire organism, we created a unified data set of MAE and BAE genes, with one cell

[1]Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [3]Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [4]Present addresses: Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA (V.S.); Evolutionary Immunogenomics, Department of Evolutionary Ecology, Max Planck Institute for Evolutionary Biology, Plön, Germany (T.L.L.). [5]These authors contributed equally to this work. Correspondence should be addressed to A.A.G. (gimelbrant@mail.dfci.harvard.edu) or S.R.S. (ssunyaev@rics.bwh.harvard.edu).

Received 14 December 2014; accepted 23 December 2015; published online 25 January 2016; doi:10.1038/ng.3493

line representing each of the following six cell types we had previously characterized for the MAE signature: lymphoid cells, myeloid cells, embryonic stem cells, myocytes, and mammary and vascular epithelial cells. Note that the chromatin signature has been demonstrated to be effective outside the lymphoblastoid cell line (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 the Online Methods for details). Specifically, a gene was only included in our MAE data set if it had the MAE chromatin signature in at least one cell type while being expressed at a moderate level or higher (reads per kilobase per million (RPKM) ≥1). For a gene to be included in the BAE data set, it was required to have no MAE signature in any cell type where its expression was detected at any level and to exhibit moderate expression in at least one of the other cell types considered. After applying additional filters (such as excluding olfactory receptor genes and the extended major histocompatibility complex (MHC) region; see the Online Methods for the full description), the resulting high-confidence genome-wide data set contained 10,233 human genes, of which 4,227 had MAE and 6,006 had 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% confidence interval (CI): $5.0 \times 10^{-4} \pm 2.0 \times 10^{-5}$ for MAE genes in the global population and $3.3 \times 10^{-4} \pm 9.3 \times 10^{-6}$ 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 ($1.1 \times 10^{-3} \pm 4.4 \times 10^{-5}$ for MAE genes in the global population and $7.4 \times 10^{-4} \pm 2.7 \times 10^{-5}$ for BAE genes; **Fig. 1b**). This difference in nucleotide diversity was not limited to a particular population: MAE genes showed a similar increase in nucleotide diversity when assessed separately in different populations from the 1000 Genomes Project (**Fig. 1**), as well as in African-American and European-American populations from Exome Sequencing Project (ESP) data[18] (**Supplementary Fig. 2**). Note that nucleotide diversity was robust to the number of cell types with MAE, and the difference between MAE and BAE genes was 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 in MAE genes. However, MAE genes remained more diverse than BAE genes after controlling for the relevant Gene Ontology (GO) categories ($P = 1 \times 10^{-4}$; **Supplementary Fig. 4** and **Supplementary Table 2**).

The observed increase in nucleotide diversity for MAE genes 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 might 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 in 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 set with BAE in all cell types[5]. To assess whether MAE genes, as a group, are less constrained by selection, we asked whether these genes are less likely to result in morbidity when mutated than BAE genes. Using a set of human genes known to be associated with morbidity whose mutation causes Mendelian diseases (extracted from the Online Mendelian Inheritance in Man (OMIM) database; Online Methods), we calculated the representation of these genes in the MAE and BAE gene sets. There was no depletion of the morbidity-associated genes in the MAE set (**Fig. 2a** and **Supplementary Fig. 5**); in fact, there was a slight enrichment for these genes ($P < 1 \times 10^{-3}$).

To further estimate the relative effects of purifying selection in the overall MAE and BAE gene sets, we focused on variation at synonymous fourfold-degenerate sites. In the 1000 Genomes Project data, nucleotide diversity was elevated in MAE genes relative to BAE genes to a similar extent when all sites were assessed and when only nondegenerate sites and fourfold-degenerate sites were assessed ($\pi_{\text{coding}}/\pi_{\text{fourfold degenerate}} = 0.47$ and 0.45 for MAE and BAE genes, respectively, in the global population; **Fig. 1**). Similarly, analysis of sequence substitutions between human and chimpanzee indicates that the strength of purifying selection on MAE and BAE genes has been nearly identical. The reduction in the number of nonsynonymous substitutions per site as compared to the number of synonymous substitutions per site (d_N/d_S) measures the proportion of amino acid–altering mutations

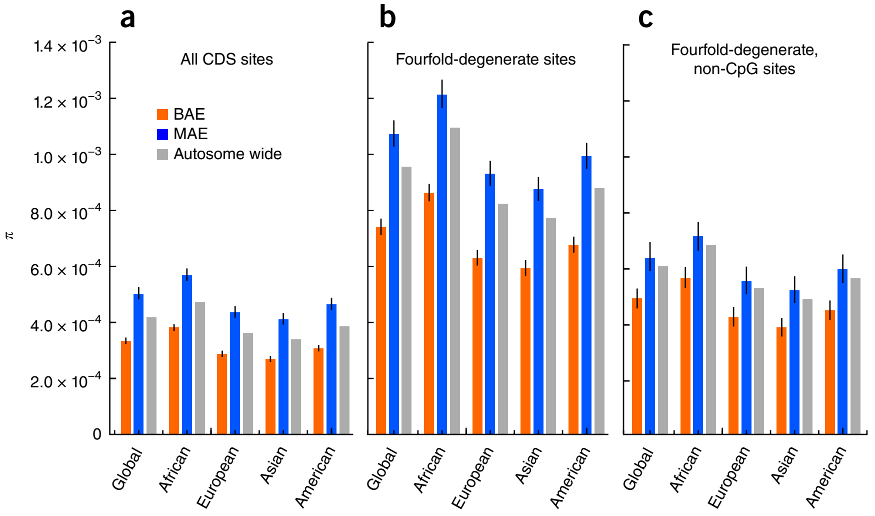

Figure 1 Nucleotide diversity is higher in MAE genes. (**a**) Average nucleotide diversity (π) for MAE and BAE genes in the 1000 Genomes Project data set (global) and all four continental groups: African, European, Asian and American. Nucleotide diversity is calculated for the coding regions (CDS), including all sites. Error bars, 95% confidence intervals calculated by bootstrapping. Orange, BAE genes; blue, MAE genes; gray, all autosomal genes. (**b**) As in **a**, with nucleotide diversity calculated for fourfold-degenerate sites only. (**c**) As in **b**, excluding CpG-prone sites (all sites preceded by a cytosine or followed by a guanine) from the calculation of nucleotide diversity. Nucleotide diversity was adjusted for the difference in mutation rates at non-CpG-prone fourfold-degenerate sites estimated from a mutational model in ref. 20.

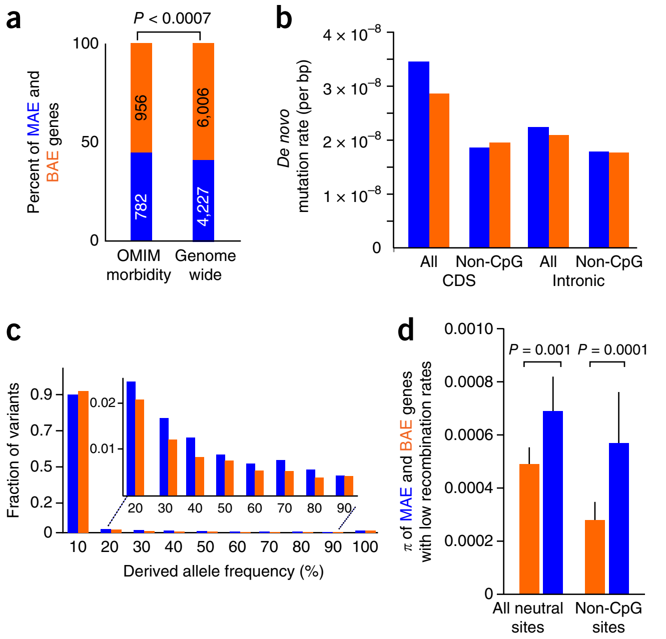

Figure 2 Purifying selection, mutation rate and recombination as potential sources of genetic diversity in MAE genes. (**a**) The proportion of MAE and BAE genes among genes known to cause Mendelian diseases extracted from the OMIM database (OMIM MorbidMap) and across the genome. The numbers of genes in each category are shown. Here and elsewhere, MAE data are shown in blue and BAE data are shown in orange; the *P* value is from a Fisher's exact test. (**b**) Average *de novo* per-base diploid mutation rate for MAE and BAE genes derived from the whole-genome sequences of GoNL parent-child trios. Left, mutation rate estimated from 82 *de novo* mutations in coding regions. Right, mutation rates estimated from 2,272 *de novo* mutations in intronic regions. All, including CpG sites; non-CpG, excluding CpG sites. (**c**) Site frequency spectra (SFS) for derived alleles in MAE and BAE genes in the 1000 Genomes Project data set. Shown is the fraction of variants for neutral (fourfold-degenerate) alleles with a given derived allelic frequency, in bins of 10%. The inset shows a close-up view of high allelic frequencies (10–90%). (**d**) Average nucleotide diversity (π) for the ~1,000 genes (blue, MAE; orange, BAE) with the lowest local recombination rate ($r \leq 0.21$ cM/Mb) using data from the 1000 Genomes Project global population. Left, all neutral (fourfold-degenerate) sites were used for the calculation. Right, CpG-prone sites were excluded from the calculation and the 1.06-fold difference in non-CpG mutation rates was taken into account. Error bars, 95% confidence intervals. Analysis for other ranges of recombination rate can be found in **Supplementary Table 6**.

that are selectively unfavorable and thus prevented from being fixed in a population[24]. The d_N/d_S ratio was not significantly different for MAE and BAE genes as a group (0.21 ± 0.01 for both the MAE and BAE gene sets; **Supplementary Table 3**) nor in per-gene estimates (Wilcoxon rank-sum test, $P = 0.09$; **Supplementary Fig. 6**).

However, as synonymous sites, including fourfold-degenerate sites, have been shown to be under selection to some extent[25], we also compared the frequencies of MAE and BAE genes among genes reported to be under selective constraint, as assessed by depletion of missense SNPs in ESP data[26]. The MAE and BAE genes were equally represented among the 1,003 genes reported to be under the highest selective constraint (6.0% and 6.1%, respectively; Fisher's exact test, $P = 0.87$). Moreover, MAE and BAE genes were identically distributed with respect to all positive values of selective constraint (**Supplementary Table 4**), ruling out the possibility that purifying selection might affect MAE genes differently for weakly constrained genes. Collectively, these observations suggest that, in comparison 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 dinucleotides, the leading factor determining sequence-specific differences in mutation rate. We observed that CpG sites were significantly more frequent ($P < 1 \times 10^{-15}$) in the coding sequences of MAE genes (41.5 CpGs/kb) than in those of BAE genes (27.1 CpGs/kb). To test whether the difference in CpG content translates into a difference in mutation rate, we analyzed the per-gene mutation rate map constructed using both human-chimpanzee divergence and observed patterns of *de novo* mutation in humans[20]. This map confirmed the significant elevation of mutation rates in MAE genes (**Supplementary Table 5**). Synonymous mutations at fourfold-degenerate sites and overall protein-coding mutations occurred at 1.28- and 1.22-fold higher levels in MAE than in BAE genes, respectively ($P < 1 \times 10^{-4}$). This difference appeared fully consistent with the difference in densities of true *de novo* mutations identified in the 250 family trio pedigrees of the Genome of the Netherlands (GoNL) project[20] (**Fig. 2b** and **Supplementary Table 5**). However, the latter analysis lacked power because of the scarcity of *de novo* mutations.

Interestingly, the intronic regions of MAE genes were only slightly more enriched with CpG dinucleotides (11.2 CpGs/kb) than those of BAE genes (10.9 CpGs/kb; **Supplementary Table 5**). Thus, the high CpG density within the 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 indicated that the intronic regions of MAE genes had only a 1.04-fold higher mutation rate than those of BAE genes ($P < 1 \times 10^{-4}$), and the set of true *de novo* mutations from the GoNL pedigrees also suggested a 1.07-fold difference (95% CI = 0.99–1.17; $P = 0.09$).

As the non-CpG mutation rate has been reported to be higher within regions of high CpG density[27], we also examined whether the increased protein-coding mutation rate in MAE genes is entirely driven by hypermutable CpG dinucleotides. When we excluded CpG sites, the per-gene mutation rates derived from divergence data showed statistically significant but small increases of 1.03-fold across coding regions ($P < 1 \times 10^{-4}$) and 1.06-fold at fourfold-degenerate sites ($P < 1 \times 10^{-4}$) (**Fig. 2b** and **Supplementary Table 5**).

To determine whether increased nucleotide diversity in the coding regions of MAE genes can be explained entirely by high CpG content, we compared the nucleotide diversity in non-CpG-prone sites while adjusting for the 1.06-fold difference in non-CpG mutation rates between the MAE and BAE gene sets. The difference between MAE and BAE genes remained highly significant ($\pi = 6.2 \times 10^{-4} \pm 4.8 \times 10^{-5}$ for MAE genes in the global population and $5.1 \times 10^{-4} \pm 3.5 \times 10^{-5}$ for BAE genes; $P < 5 \times 10^{-4}$) (**Fig. 1c**). This suggests that differences in the raw mutation rate 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 the 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 toward alleles that are common in all populations combined ($P < 1 \times 10^{-20}$; **Fig. 2c**), as well as in the individual populations analyzed (**Table 1** and **Supplementary Fig. 7**).

Table 1 Site frequency spectrum is shifted toward common frequency in MAE genes

Population	Synonymous (fourfold degenerate)	Missense damaging	Missense benign
Global	$<1 \times 10^{-20}$	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}

P values from Pearson's χ^2 test for a significant shift toward a frequency corresponding to a common variant in MAE as compared to BAE genes in the global 1000 Genomes Project data set and all four continental groups: African, European, Asian and American.

The shift in allele frequency distribution between MAE and BAE genes persisted in all functional categories of sites, including fourfold-degenerate sites. Notably, 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 observed 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[33], MAE genes tended to be associated with a local recombination rate that was higher than that for BAE genes. This observation held in our much larger MAE and BAE gene sets ($P < 3 \times 10^{-54}$; **Supplementary Fig. 8**). We thus tested whether the differences in 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 equally populated ranges of recombination rate, with ~1,000 genes per bin, and compared the nucleotide diversity 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 $r = 0.11$ cM/Mb for both groups of genes), the difference in nucleotide diversity remained; this difference also remained after additional exclusion of CpG-prone sites and correction for non-CpG mutation rates, showing that it is independent of local mutation rate ($P = 2.2 \times 10^{-4}$; **Fig. 2d**).

Although the difference in nucleotide diversity between MAE and BAE genes appeared to be greater in regions of lower recombination rate, the effect remained in regions with higher recombination rates. To boost statistical power (only 22% of fourfold-degenerate SNPs are not CpG prone), we calculated nucleotide diversity using all SNPs, with CpG and non-CpG SNPs combined, and directly corrected for the mutation rate difference using divergence-based mutation rate estimates. Further, we controlled for underestimation of nucleotide diversity due to lower sequencing depth in MAE genes by analyzing only the sites passing a strict filter on read depth[17]. Fourfold-degenerate sites with lower coverage (average read coverage 50% below the genomic average) were significantly enriched in MAE genes (12.1% and 5.5% for MAE and BAE genes, respectively; $P < 1 \times 10^{-15}$). When we recalculated nucleotide diversity 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, the diversity was significantly elevated for MAE genes across all recombination rate bins ($\Delta\pi = 7.4 \times 10^{-5} \pm 5.2 \times 10^{-5}$; $P = 0.0037$) and remained significant even when we excluded the bin with the lowest recombination rate ($r \geq 0.21$ cM/Mb; $P = 0.011$; **Supplementary Table 7**; see the Online 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. Notably, however, although these factors contribute to the elevated diversity in MAE genes, our data argue that they are not able to fully account for it.

Genetic variation is older in MAE genes

Because MAE genes showed increased nucleotide diversity and a shift in allele frequency in synonymous sites, we examined whether 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 the evaluation of relative allelic ages. To ensure that overall differences in recombination rate and in allele frequency did not have a major role in the comparison, we further refined the set of variants assessed as follows. We performed the analysis conditional on local recombination rate by dividing the variants into decile bins by derived allele frequency and analyzing the genes within each bin separately (**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 ($P < 7.1 \times 10^{-7}$; **Fig. 3a**), indicating that MAE-associated variants are 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 data set[36]. CpG sites were excluded in the 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 at non-CpG sites, safeguarding the analysis from the effect of differences in non-CpG mutation rates[36].

We first confirmed that genetic variation in MAE genes was, on average, older than that in BAE genes, as measured by T_{MRCA} ($P < 2 \times 10^{-16}$;

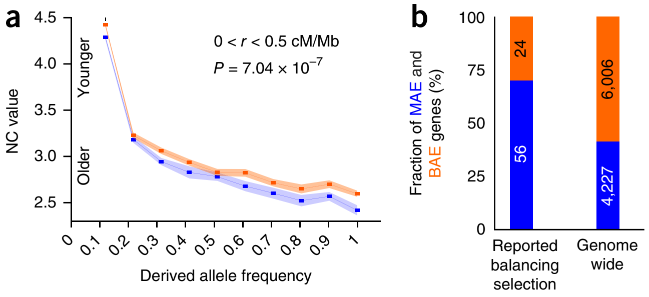

Figure 3 Older variants and genes under balancing selection are enriched among MAE genes. (**a**) Allelic age of synonymous SNPs in MAE and BAE genes estimated by applying the NC method[35] to genomes sequenced by the GoNL project. NC values are plotted for synonymous SNPs in MAE genes (blue) and BAE genes (orange) as a function of derived allele frequency (10% bins). Error bars, s.e.m. Analysis was limited to variants associated with local recombination rates between 0 and 0.5 cM/Mb. For other ranges of recombination rates, see **Supplementary Table 6**. See the Online Methods for details. (**b**) Percentage of MAE (blue) and BAE (orange) genes among genes thought to be under balancing selection (**Supplementary Table 10**) as compared to the genome-wide data set (Pearson's χ^2 test, $P < 3.9 \times 10^{-7}$).

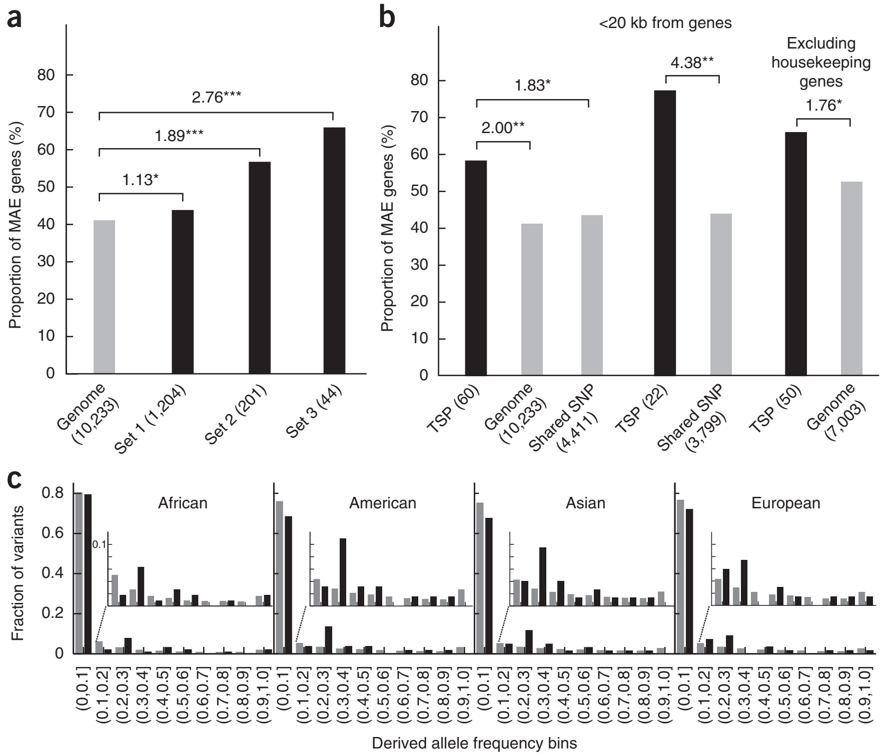

Figure 4 Trans-species polymorphisms are enriched among MAE genes. (**a**) Percentage of MAE genes with application of the Neanderthal filter to candidates for human-chimpanzee TSPs. The candidate gene sets corresponding to TSPs are shown in black: set 1, genes with ancient SNPs shared with Neanderthals—genes harboring at least one ancient protein-coding SNP predating the human-Neanderthal split; set 2, genes with any evidence of human-chimpanzee TSPs[38]; set 3, genes that meet the criteria for both sets 1 and 2. The genome-wide data set is shown in gray. 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**) Percentage of MAE genes among human-chimpanzee trans-species haplotypes (TSP, black) and control data sets (gray), including the genome-wide data set (genome) and the set of genes adjacent to SNPs segregating in both species identically by state (shared SNPs) as a conservative control for the uneven density of recurrent mutations. Left, data for all haplotypes; center, data for haplotypes less than 20 kb from genes; right, data for all haplotypes except those for housekeeping genes, defined by ubiquitous and low-variance expression across tissues[43]. The number of genes per category is shown below the group labels. Odds ratios and their significance levels are reported (*$P < 0.05$, **$P < 0.01$). See the Online Methods for details. (**c**) The SFS for derived alleles in MAE genes that also have human-chimpanzee trans-species haplotypes within 20 kb of the genes (17 genes in total; black) as compared to the SFS for all genes lacking trans-species haplotypes (gray). The insets increase *y*-axis resolution. All SNPs are at fourfold-degenerate sites; allele frequencies are from the 1000 Genomes Project.

Online Methods and **Supplementary Fig. 9**). T_{MRCA} is a direct measure of 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 gene expression across tissues, selective constraint of the gene, gene length and recombination rate, we confirmed that MAE status remained a significant predictor of greater T_{MRCA} ($P = 7.5 \times 10^{-8}$; **Supplementary Fig. 10** and **Supplementary Table 9**). 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 greater T_{MRCA} ($P = 6.8 \times 10^{-7}$), and the regression coefficient decreased only slightly from 0.054 to 0.051 (5.6%).

Our observations suggest that genetic variation is not only greater but is also, on average, older in MAE genes than in 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 (for example, taste receptors and the extended MHC region; Online Methods). We found that genes encoding extracellular 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; $P = 7.5 \times 10^{-33}$) (**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 strong enrichment of genes classified as MAE in this list (1.75-fold; $P < 1 \times 10^{-4}$) (**Fig. 3b**).

Ancient balancing selection can leave a trace in 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. Although most of the polymorphisms are noncoding, they have been associated with specific genes in the human and chimpanzee genomes. We asked whether these TSPs (**Supplementary Table 11**) are differently represented in the MAE and BAE gene sets. We found that the set of TSPs was strongly and significantly enriched in MAE genes (odds ratio (OR) = 1.89; $P = 6.3 \times 10^{-6}$) (**Fig. 4a**). The enrichment was stronger still when we required human-chimpanzee 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 the human population as a polymorphism (OR = 2.76; $P = 8.4 \times 10^{-4}$) (**Fig. 4a**).

One possible confounding factor is that some (perhaps a large) fraction of the TSPs could be due to independent remutations, the occurrence of exactly the same mutation in both lineages after the split. To estimate the contribution of remutations, we assessed the relative enrichment of MAE and BAE genes among trans-species haplotypes, as defined in the same analysis of human-chimpanzee TSPs[38]. The haplotypes, consisting of more than one polymorphism, are fewer in number than the SNPs but are less likely to arise by remutation. We found that the enrichment for MAE genes was even stronger in this analysis, especially when genes within 20 kb of these haplotypes (where the majority of *cis* expression quantitative trait loci (*cis*-eQTLs) are located[39]) were considered (OR = 4.38; $P = 0.0015$) (**Fig. 4b**). Although extracellular matrix genes have been suggested to be a target of balancing selection[37] and are predominantly MAE, this is not attributable to the enrichment of trans-species haplotypes among MAE genes. Only four trans-species haplotypes were located around the genes encoding extracellular matrix, and excluding this category of genes did not affect the association between TSPs and MAE genes (**Supplementary Fig. 4**).

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

DISCUSSION

Using several large data sets characterizing human genetic variation, including the 1000 Genomes Project[17] and ESP[18], we showed that human autosomal genes classified as MAE on the basis of a characteristic gene body chromatin signature[5] have considerably higher nucleotide diversity (π) than BAE genes (**Fig. 1**). Although the chromatin signature shows remarkable consistency across different genetic backgrounds (**Supplementary Fig. 11**), some type I and type II errors are expected. Note that the increase in nucleotide diversity was observed even though identification of MAE and BAE gene sets using 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 stem from relaxed purifying selection. We show that MAE genes have, on average, increased recombination rates and an elevated density of hypermutable contexts, which contribute 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 MAE genes as a group than on BAE genes (**Figs. 3** and **4**). Gene classes thought to evolve under balancing selection are preferentially MAE, the frequency distributions of putatively neutral alleles in MAE genes are shifted toward those consistent with common variation, variation in MAE genes is, on average, older, and TSPs preferentially colocalize with MAE genes. We conclude that 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 refs. 3,14–16). In particular, heterogeneity involving cells of the same type is likely increased in individuals heterozygous for MAE genes whose alleles are functionally distinct (**Supplementary Fig. 12**). Intriguingly, MAE genes are enriched for ones encoding proteins present on the cell surface and 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 the susceptibility of a tissue as a whole to infectious agents. Such a general adaptive role for MAE would be consistent with the increased allelic diversity that is widespread in human populations rather than being limited to particular environments or geographic locations. Because MAE genes are enriched for particular functional categories, high nucleotide diversity and MAE might be two separate but interacting phenomena that 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.

URLs. Online Mendelian Inheritance in Man (OMIM) MorbidMap, http://omim.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D. Balick for useful discussions and I. Adzhubei for help with PolyPhen analysis. This work was supported in part by the following US National Institutes of Health (NIH) awards: R01 GM114864 to A.A.G. and R01 GM078598, GM105857 and MH101244 to S.R.S. A.A.G. was supported in part by the Pew scholar award; T.L.L. was supported by a fellowship from the German Research Foundation (DFG; LE 2593/1-1 and LE 2593/2-1); L.G. was a summer scholar in the Harvard/Massachusetts Institute of Technology BIG program (supported by US NIH award U54 LM008748); and R.B.M. and C.W. were supported by grants from the US NIH/National Institute of General Medical Sciences (R01 GM61936 and 5DP1 GM106412) and Harvard Medical School.

AUTHOR CONTRIBUTIONS

A.A.G. and S.R.S. conceived the study. All authors contributed to data analysis. A.A.G., S.R.S. and V.S. wrote the manuscript with input from S.C., M.S. and R.B.M.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Savova, V., Vigneau, S. & Gimelbrant, A.A. Autosomal monoallelic expression: genetics of epigenetic diversity? *Curr. Opin. Genet. Dev.* **23**, 642–648 (2013).
2. Chess, A., Simon, I., Cedar, H. & Axel, R. Allelic inactivation regulates olfactory receptor gene expression. *Cell* **78**, 823–834 (1994).
3. Gimelbrant, A., Hutchinson, J.N., Thompson, B.R. & Chess, A. Widespread monoallelic expression on human autosomes. *Science* **318**, 1136–1140 (2007).
4. Zwemer, L.M. *et al.* Autosomal monoallelic expression in the mouse. *Genome Biol.* **13**, R10 (2012).
5. Nag, A. *et al.* Chromatin signature of widespread monoallelic expression. *eLife* **2**, e01256 (2013).
6. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).
7. Jeffries, A.R. *et al.* Stochastic choice of allelic expression in human neural stem cells. *Stem Cells* **30**, 1938–1947 (2012).
8. Gendrel, A.V. *et al.* Developmental dynamics and disease potential of random monoallelic gene expression. *Dev. Cell* **28**, 366–380 (2014).
9. Eckersley-Maslin, M.A. *et al.* Random monoallelic gene expression increases upon embryonic stem cell differentiation. *Dev. Cell* **28**, 351–365 (2014).
10. Li, S.M. *et al.* Transcriptome-wide survey of mouse CNS-derived cells reveals monoallelic expression within novel gene families. *PLoS One* **7**, e31751 (2012).
11. Pereira, J.P., Girard, R., Chaby, R., Cumano, A. & Vieira, P. Monoallelic expression of the murine gene encoding Toll-like receptor 4. *Nat. Immunol.* **4**, 464–470 (2003).
12. Spencer, H.G. Population genetics and evolution of genomic imprinting. *Annu. Rev. Genet.* **34**, 457–477 (2000).
13. Wilkins, J.F. & Haig, D. What good is genomic imprinting: the function of parent-specific gene expression. *Nat. Rev. Genet.* **4**, 359–368 (2003).
14. Wu, C.T. & Dunlap, J.C. Homology effects: the difference between 1 and 2. *Adv. Genet.* **46**, xvii–xxiii (2002).
15. Hoehe, M.R. *et al.* Multiple haplotype–resolved genomes reveal population patterns of gene and protein diplotypes. *Nat. Commun.* **5**, 5569 (2014).
16. Chess, A. Mechanisms and consequences of widespread random monoallelic expression. *Nat. Rev. Genet.* **13**, 421–428 (2012).

17. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
18. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
19. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
20. Francioli, L.C. *et al.* Genome-wide patterns and properties of *de novo* mutations in humans. *Nat. Genet.* **47**, 822–826 (2015).
21. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
22. Nag, A., Vigneau, S., Savova, V., Zwemer, L.M. & Gimelbrant, A.A. Chromatin signature identifies monoallelic gene expression across mammalian cell types. *G3 (Bethesda)* **5**, 1713–1720 (2015).
23. Nei, M. & Li, W.H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **76**, 5269–5273 (1979).
24. Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275–276 (1977).
25. Chamary, J.V. & Hurst, L.D. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* **6**, R75 (2005).
26. Samocha, K.E. *et al.* A framework for the interpretation of *de novo* mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
27. Walser, J.C. & Furano, A.V. The mutational spectrum of non-CpG DNA varies with CpG content. *Genome Res.* **20**, 875–882 (2010).
28. Li, W.-H. *Molecular Evolution* (Sinauer Associates, 1997).
29. Begun, D.J. & Aquadro, C.F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519–520 (1992).
30. Charlesworth, B., Morgan, M.T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
31. Smith, J.M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).
32. Hellmann, I. *et al.* Why do human diversity levels vary at a megabase scale? *Genome Res.* **15**, 1222–1231 (2005).
33. Necsulea, A., Sémon, M., Duret, L. & Hurst, L.D. Monoallelic expression and tissue specificity are associated with high crossover rates. *Trends Genet.* **25**, 519–522 (2009).
34. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010).
35. Kiezun, A. *et al.* Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. *PLoS Genet.* **9**, e1003301 (2013).
36. Rasmussen, M.D., Hubisz, M.J., Gronau, I. & Siepel, A. Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* **10**, e1004342 (2014).
37. Andrés, A.M. *et al.* Targets of balancing selection in the human genome. *Mol. Biol. Evol.* **26**, 2755–2764 (2009).
38. Leffler, E.M. *et al.* Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**, 1578–1582 (2013).
39. Veyrieras, J.B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4**, e1000214 (2008).
40. Sellis, D., Callahan, B.J., Petrov, D.A. & Messer, P.W. Heterozygote advantage as a natural consequence of adaptation in diploids. *Proc. Natl. Acad. Sci. USA* **108**, 20666–20671 (2011).
41. DeGiorgio, M., Lohmueller, K.E. & Nielsen, R. A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genet.* **10**, e1004561 (2014).
42. Yang, S. *et al.* Parent-progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* **523**, 463–467 (2015).
43. Eisenberg, E. & Levanon, E.Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).

ONLINE METHODS

Data sets. Genes were classified as MAE or BAE using specific chromatin signature[5] (co-occurrence of the trimethylation of histone H3 at lysine 27 (H3K27me3) silencing mark and the trimethylation of histone H3 at lysine 36 (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), H1 hESCs (embryonic stem cells), HSMM (skeletal muscle myocytes), HUVEC (umbilical vascular epithelial cells) and HMEC/HCC1954 (mammary epithelial cells). To consider a gene as MAE, we required monoallelic status in at least one cell line with an expression level of RPKM ≥1 in that cell line. To consider a gene as BAE, we required the absence of monoallelic status in all cell lines where expression of 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 was ~10%, with values ranging from ~4% to 15% (**Supplementary Fig. 3** and **Supplementary Table 1**).

We excluded genes that did 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 supercontigs. 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 (chr. 6: 28,000,000–34,000,000) because 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 the 'genome-wide data set'.

Analyses of genetic variation were primarily carried out on 1000 Genomes Project Phase 1 data[17], encompassing 1,092 individuals from four superpopulations: African (n = 246), European (n = 379), admixed American (n = 181) and Asian (n = 286). We also examined protein-coding variants in 4,300 European Americans and 2,203 African Americans in the ESP data set (ESP6500SI-V2)[18]. In addition, the GoNL[20] data set was used for *de novo* mutation rate estimation and allelic age analysis. The GoNL data set consists of the phased whole-genome sequences of 250 Dutch parent-child trios and a genome-wide collection of 11,020 *de novo* mutations identified in the offspring.

The candidates for 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 data set also includes the smaller set of mostly noncoding trans-species haplotypes defined by two or more trans-species SNPs within 4 kb of each other and in shared linkage disequilibrium (LD) structure. For intergenic trans-species haplotypes, the authors selected the gene at the 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-chimpanzee TSPs, we used the genome sequence of a single Neanderthal individual from a cave in the Altai mountains[45]. The genome was sequenced at ~52× coverage, with autosomal contamination estimated to be between 0.8 and 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 the YRI population.

For all analyses, ancestral alleles were distinguished from derived alleles on the basis of 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 (ref. 23) for MAE and BAE in humans by analyzing SNPs in neutral sites or all sites of protein-coding regions. The neutral π was calculated in fourfold-degenerate sites with polymorphism data from the 1000 Genomes Project. To rule out the possibility that the biased distribution of hypermutable CpG dinucleotides explains the difference of π in MAE and BAE, we further computed the neutral π value using only non-CpG-prone sites, which are defined by nucleotides that are not preceded by a cytosine or followed by a guanine and therefore do not overlap with a CpG site. For the non-CpG π value, 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, on the basis of divergence-based mutation rates. For all-site π values, all SNPs in protein-coding regions were analyzed using ESP data as well as the 1000 Genomes Project data. For ESP data, we derived all-site π from the per-gene π value and the observed length of each gene. For the 1000 Genomes Project data, we annotated SNPs with the change of amino acids in canonical transcripts using the 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 using 11,020 *de novo* point mutations from 269 GoNL offspring[20] (**Supplementary Table 5**). MAE and BAE genes contained 32 and 50 mutations in protein-coding regions and 1,170 and 1,102 mutations in intronic regions, respectively. To control for local variation in the power to detect *de novo* events, we estimated detection power (between 0 and 1) from simulated positive controls (kindly provided by L.C. Francioli)[20]: 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

$$\Theta_{\mathrm{MAE}} \sim \mathrm{Poisson}\left(\lambda = 269\mu\tau_{\mathrm{MAE}}\ P_{\mathrm{MAE}}\right)$$
$$\text{and } \Theta_{\mathrm{BAE}} \sim \mathrm{Poisson}\left(\lambda = 269\mu\tau_{\mathrm{BAE}}\ P_{\mathrm{BAE}}\right)$$

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 the mutational target and P is the estimated mean detection power. The null model assumes an equal mutation rate μ across MAE and BAE genes. We tested for unequal mutation rates by excluding as well as including CpG dinucleotides in the mutational target because CpGs are more frequent in MAE than in BAE genes.

Because of 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[20]. Briefly, a context-dependent substitution rate matrix was inferred for each 1-Mb genomic block 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 the 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 whether 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 a 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[46]. We annotated r for a total of 3,281 MAE and 4,980 BAE genes and grouped the genes into eight equally sized bins by r. In each bin, we calculated the P value for the test of significant difference in π between MAE and BAE genes 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 values similarly to control for both recombination and mutation rates 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 fourfold-degenerate sites, we used all fourfold-degenerate sites (both CpG prone and not) to estimate neutral π and then cancelled out mutation rate bias by using the divergence-based mutation rate map. The mutational rate bias was calculated for each bin of r separately to account for the variation in sequence composition by r. Furthermore, we excluded fourfold-degenerate sites of too low sequencing coverage as they are enriched in MAE genes (12.1% as compared to 5.5% for BAE genes) and lead to underestimation of π as a result of 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 in π ($\Delta\pi$) and its 95% confidence interval was calculated by variance-normalized meta-analysis across $\mathbf{r}$ bins. The variance of $\Delta\pi$ in each bin was estimated by bootstrapping.

Site frequency spectra. We calculated the derived SFS of SNPs in coding regions of MAE and BAE genes using the 1000 Genomes Project data set. Only SNPs polymorphic in each individual population were used for the analysis. For neutral SFS, we used SNPs in fourfold-degenerate sites, and for all-site SFS we stratified SNPs by amino acid changes and their functional impact predicted by PolyPhen-2 (ref. 47). To test for significant difference in SFS between MAE and BAE genes, we subdivided the SNPs into high- and low-frequency bins, with the cutoff of an 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 data sets: OMIM MorbidMap (see URLs) and selectively constrained genes[26]. The 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 as compared to the expected mutation rates, allows a more comprehensive and unbiased survey of selective constraints on genes, although 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. Of the 1,003 top constrained genes[26], we mapped the RefSeq IDs of 990 genes to Ensembl, excluding genes with missing RefSeq ID 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 was not significantly different across 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 (**Supplementary Table 4**).

To examine subtle difference in selective pressure that is difficult to identify in polymorphism-based data, we compared the groupwise d_N/d_S values[49] between MAE and BAE gene sets. The numbers of nonsynonymous and synonymous substitutions and sites were aggregated over MAE and BAE genes, and the overall nonsynonymous substitutions per nonsynonymous site (d_N) and synonymous substitutions per synonymous site (d_S)[48] were then calculated for MAE and BAE (**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 values for 2,021 MAE and 3,223 BAE genes after excluding genes with no synonymous substitution (Wilcoxon rank-sum test, $P = 0.09$).

Neighborhood-based clock algorithm. To compare the allelic age of synonymous SNPs within MAE and BAE genes, we applied the 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.

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 10-kb windows containing the test SNP on the deCODE sex-averaged genetic map. The window size of 10 kb 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 genes were significantly younger than those in MAE genes 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. We conducted a multivariate regression analysis to study the correlation between a gene's MAE status and its T_{MRCA} in the 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

$$\log(T_{\mathrm{MRCA}}) = \beta_0 + \beta_1 I_{\mathrm{MAE/BAE}} + \beta_2\text{length} + \beta_3 r + \beta_4 \text{ expression level} + \beta_5\text{expression breadth} + \beta_6 Z$$

where $I_{\mathrm{MAE/BAE}}$ 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 410-kb windows), expression level 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 $I_{\mathrm{MAE/BAE}}$)[5], expression breath is the gene expression breadth (scores between 0 and 1 for tissue specificity across 12 human tissues: 0, housekeeping; 1, tissue specific)[49], and Z is the selective constraint[26]. T_{MRCA} is unlikely to be confounded by mutation rate variation because (i) CpG dinucleotides were excluded from the analysis[36] and (ii) 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 whether 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. MAE status remained significantly correlated with older T_{MRCA} ($P = 1.2 \times 10^{-10}$).

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

$$\log(T_{\mathrm{MRCA}}) = \beta_0 + \beta_1 I_{\mathrm{MAE/BAE}} + \beta_m \mathrm{MAE}_m + \text{covariates}$$

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 nonzero ($P = 0.49$), showing that genes that have the MAE signature in multiple tissues do not have larger 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 regions, 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 nonzero ($P = 6.8 \times 10^{-7}$).

Finally, we confirmed that genes that were experimentally established as MAE in human lymphoblasts[3] had significantly greater T_{MRCA} values than BAE genes in the same set (Wilcoxon rank-sum test, $P = 6.3 \times 10^{-6}$).

Trans-species polymorphisms. To enrich for the strongest signals of long-term balancing selection, we intersected the genes identified by human-chimpanzee 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 the YRI population and also present in the Altai Neanderthal genome in one or two copies. To minimize false positives due to remutation, we excluded derived alleles in a CpG context. In total, we collected

3,383 ancient protein-coding SNPs (2,603 genes) predating the Neanderthal split. Among these, 104 genes are also associated with human-chimpanzee 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 remutations using 33,906 SNPs shared by human and chimpanzee across the autosomes ('shared SNPs'). We conservatively assumed that all shared SNPs were false positives due to remutation. 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 testing. Second, because housekeeping 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 housekeeping genes. In total, 549 MAE and 2,681 BAE genes were classified as housekeeping, defined by the ubiquitous presence of transcripts and minimal variation in expression levels across all tissues[43]. Third, the identification of candidate genes for balancing selection based on the closest distance to intergenic trans-species haplotypes can be ambiguous, especially if the haplotypes are distant from the gene. On the basis of a previous observation that the majority of *cis*-eQTLs are located within 20 kb of their target gene[39], we repeated the enrichment test using only trans-species haplotypes within 20 kb of 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 the derived allelic frequencies of fourfold-degenerate variants from the 1000 Genomes Project data. The significant difference in SFS was tested by χ^2 goodness-of-fit test, combining the frequencies above 40% into a single bin because of their small number of observed counts.

44. Borel, C. *et al.* Biased allelic expression in human primary fibroblast single cells. *Am. J. Hum. Genet.* **96**, 70–80 (2015).
45. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
46. Cai, J.J., Macpherson, J.M., Sella, G. & Petrov, D.A. Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genet.* **5**, e1000336 (2009).
47. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
48. Bustamante, C.D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157 (2005).
49. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).

doi:10.1038/ng.3493