|  |
| --- |
|  |
| Supplementary Figure |
| MAE can lead to co-existence of epigenetically distinct cell subpopulations. |
| A cell (*circle*) is located along the horizontal axis according to relative expression of paternal (Pat) and maternal (Mat) alleles. Cells form a uniform population when assessed with focus on genes with both alleles transcriptionally active (biallelic expression; BAE). By contrast, mitotically stable MAE leads to formation of distinct cell subpopulations, depending on which allele is active (*gray*) and which is downregulated (*x*). Note that MAE genes might be expressed biallelically in some cells, thus it is important to distinguish between biallelic expression in a given cell, which might occur both in MAE and BAE genes, and the capability of a gene to show MAE-based heterogeneity.  Briefly, the properties of MAE can be summed up as follows (reviewed in [[1](#_ENREF_1)]): Extreme allelic bias (10-fold or more) is common for MAE, though RNA-Seq-based approaches also reveal examples of more attenuated biases. MAE is highly mitotically stable over multiple cell divisions. Multiple genes are simultaneously subject to MAE in a given cell and its clonal progeny, and allelic choice at each MAE locus is apparently independent. When the proteins encoded by the two alleles are functionally distinct, MAE can lead to dramatic functional differences between otherwise similar cells of the same type. This can result in astronomical combinatorial diversity in overall allelic expression patterns between clonal lineages of otherwise similar cells. |
|  |
| Supplementary Figure |
| Nucleotide diversity in the Exome Sequencing Project subpopulations. |
| π is calculated for the coding regions (CDS), including all sites. Error bars represent 95% confidence intervals calculated by bootstrapping. Colors as in main figures. Groups: total for all populations (ESP-all), African-American (ESP-AA), American of European descent (ESP-EA). |
|  |
| Supplementary Figure |
| Average nucleotide diversity (π) for MAE and BAE genes in the global 1000 Genomes dataset. |
| 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;  **a.** π per cell line for genes classified MAE and BAE in that cell line;  **b.** π for genes classified MAE with RPKM >1 in only one (MAE 1), two, three or four cell lines, compared to genes classified BAE;  **c.** π for genes experimentally determined MAE (217) or BAE (2412) based on SNP array assays of five clones from GM13130 cell line, as reported in [[3](#_ENREF_3)];  **d**. π for MAE and BAE genes by expression level for low, mid and high expression, as determined in (e);  **e.** definition of low, mid and high expression categories for genes in the genome-wide dataset. RPKM=highest RPKM observed with each gene’s assigned status in the six cell lines; boundaries of the categories are shown in hashed lines.  **f**. Mutation-rate corrected, non-CpG-prone π for MAE and BAE genes by expression level. 95% confidence intervals estimated by bootstrapping. Colors as in panel (a).  **g.** Mutation-rate corrected π for MAE and BAE genes by expression level. 95% confidence intervals estimated by bootstrapping. Colors as in panel (a).  **Note to Supplementary figure 3 (f,g):**  **Nucleotide diversity (π) in expression level bins.** BAE genes are shifted toward much higher mRNA expression levels compared to MAE genes (*blue* – MAE genes, *orange –* BAE genes). Although previous studies have suggested that highly expressed genes are subjected to higher selective pressure than lowly expressed genes (PMID=16176987, 18249460), in our gene sets we do not find strong evidence for negative correlation between **π** and expression levels. Specifically, we stratified MAE and BAE genes into eight equal-sized bins by expression levels in six cell types (log10 (RPKM), see Methods for our definition of expression level) and examined the linear relationship between **π** and expression level (**f** – non-CpG **π**, **g** – overall **π**). The difference in mutation rate was corrected with divergence-based mutation rate map. Expression level is not significantly correlated with non-CpG **π** (p=0.52 for MAE and 0.61 for BAE) or overall **π** (p=0.10 for MAE and 0.07 for BAE). Note that even the marginal correlation between expression level and overall **π** for BAE genes is almost entirely driven by the genes in the highest expression level bin (log10(RPKM) > 2.0), without which the trend becomes flat (p=0.89, solid black line). This most highly expressed group of genes can explain only 9% of the difference in overall **π** (**Δπ)** between MAE and BAE genes.  To make an extremely conservative assumption, one can argue that the insignificant trend of overall **π** for BAE is uniform and holds over lowest expression levels (dashed black line). Even in that case, the potential contribution of expression bias is estimated to explain only 36% of **Δπ**. To estimate this, we extrapolated **π** of hypothetical BAE genes that follow the distribution for gene lengths and expression levels of MAE by:  where and are estimated intercept and slope of BAE trend over expression level (7.8×10-4 and -6.2×10-5, respectively, from dashed black line) and and are the number of 4-fold degenerate sites and expression level of MAE gene *i*. |
|  |
| Supplementary Figure |
| MAE genes and extra-cellular matrix molecules. |
| **a.** Genes encoding for extra-cellular matrix molecules (ECM, Gene Ontology category: GO:0031012) were strongly enriched for MAE genes (Fisher’s Exact Test, odds ratio = 8.09, *P* = 7.5×10-33). The total number of genes (Genome) and the number of genes associated with ECM are given in parenthesis.  **b.** Genes showing signatures of trans-species polymorphic haplotype (TSP) are significantly enriched for MAE genes. This enrichment remained even when excluding genes encoding for extra-cellular matrix molecules (ECM, Gene Ontology category: GO:0031012), a functional category that has previously been associated with balancing selection [[2](#_ENREF_2)]. The total number of genes in each category is given in parenthesis. Odds ratios and their significance level are reported (\* *P*<0.05, \*\* *P*<0.01). |
|  |
| Supplementary Figure |
| Proportions in genome-wide group are similar to proportions in the group that excludes the main assessed gene list. |
| **a.** Percent MAE (*blue*) and BAE (*orange*) genes among genes known to cause Mendelian diseases extracted from the OMIM database (OMIM MorbidMap), genes known to cause Mendelian diseases (Other) and in the genome-wide dataset as a whole. Shown on the colored portions of each bar are actual numbers of genes in each category.  **b.** Proportion of MAE (*blue*) and BAE (*orange*) among genes thought to be under balancing selection (listed in **Supplementary Table 5**) compared to the remaining fraction (other), and the genome-wide dataset. |
|  |
| Supplementary Figure |
| Distribution of dN/dS per gene. |
| The distribution of per-gene ratios of nonsynonymous substitution per nonsynonymous site (*dN*) to the number of synonymous substitution per synonymous site (*dS*) (obtained from [[8](https://mail.google.com/mail/u/0/#150cfa21ff8cfa7a__ENREF_8)]) is presented for 2,006 MAE **(*blue*)**and 3,209 BAE genes**(*orange*)**within the range of 0 <=*dN*/*dS* < 2. |
|  |
| Supplementary Figure |
| Allele frequencies are higher in MAE genes than BAE genes. |
| **a.** Site frequency spectra for variants in BAE and MAE genes by PolyPhen 2 classification for African subpopulation. Top: benign variants; middle: four-fold degenerate synonymous variants; bottom: damaging variants (defined as the union of possibly damaging, probably damaging and nonsense). Plot inserts zoom into derived allele frequencies between 20% and 90%. The x-axis represents derived allele frequencies in bins of ten percent; y-axis – the fraction of variants.  **b.** Low-frequency tail of site frequency spectra showing the fractions of singleton and doubleton sites in the sequences in the 1000 genomes project, African population.  **c.** Low-frequency tail of site frequency spectra for the sequences in the 1000 genomes project, global population. The allele frequency bins on the x-axis correspond to the derived allele counts of 1 to 10. |
|  |
| Supplementary Figure |
| Comparison of local recombination rates for MAE and BAE genes. |
| Red lines represent the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers. The local recombination rate is defined as the average over 410 kb window centered at each gene, based on the deCODE sex-averaged genetic map. |
|  |
| Supplementary Figure |
| Distribution of Time to Most Recent Common Ancestor (TMRCA) for MAE (*blue*) and BAE (*orange*) genes. |
| For each gene, mean TMRCA over the gene length was calculated from genome wide TMRCA estimates generated by running ARGWeaver on Complete Genomics data[9]. |
|  |
| Supplementary Figure |
| Residuals for MAE and BAE genes in a multivariate regression model of TMRCA values as a function of individual variables. |
| For each gene (MAE in *blue,* BAE in *orange)*, mean TMRCA over the gene length was calculated from genome-wide TMRCA estimates generated by running ARGWeaver on Complete Genomics data[9]. The TMRCA estimates were log-transformed for the regression analysis. All confounders except one (plotted on x-axis) were residualized for each panel. The intercept and slope of MAE and BAE trend lines (solid line for MAE and dashed line for BAE) are from the multivariate regression model. The gap between trend lines represents 1.06-fold increase in TMRCA for MAE compared to BAE genes (p=7.5 x10-8). See methods for details on model covariates. |
|  |
| Supplementary Figure |
| Identification of genes as MAE based on their chromatin signature is consistent between unrelated individuals. |
| Comparison of gene sets identified as MAE or BAE based on ChIP-Seq data in different individuals. Chromatin signature analysis was performed using ChIP-Seq data from HapMap lymphoblastoid cell lines. *Left bar* shows comparison of analyses performed on GM12878 (CEU) and GM19239 (YRI). *Center bar* shows comparison of MAE and BAE calls obtained using ChIP-Seq data from GM12878 cells (CEU) in ENCODE [[4](#_ENREF_4)] (same dataset used in [[5](#_ENREF_5)]) and in a biological replicate using the same cell line in another laboratory [[6](#_ENREF_6)] (“rep”). *Right bar* shows the same gene set as in the center, but additionally limited only to genes that show expression level of RPKM>1 in GM12878 cells. Genes BAE in both samples are *orange*; MAE in both samples shown in *blue*; MAE in the first listed sample but BAE in the other shown in *red;* vice versa in *green*. Only genes with calls made for both samples are used in this analysis. Numbers of genes in the major categories are shown. |
|  |
| Supplementary Figure |
| High allelic diversity in the context of MAE leads to increase in functional cell-to-cell heterogeneity. |
|  |
|  |
| Supplementary Figure |
| **Analysis of selective constraint in MAE and BAE genes.** |
| Distribution of signed Z-scores [[7](#_ENREF_7)] of BAE (*orange*) and MAE (*blue*) genes. Z scores are binned into intervals of 0.5 units for 3,609 MAE and 5,191 BAE genes (p-value = 0.003457, Wilcoxon rank-sum test). See also **Supplementary Table 4**. |
| Click inside this box and insert figure. For best results, use Insert menu to select a saved file; do not paste images. Source images must be in RBG color profile, at a resolution of 150—300 dpi. Optimize panel arrangement to a 2:3 height-to-width ratio; maximum online display is 600h x 900w pixels. Reduce empty space between panels and around image. Keep each image to a single page. Delete these instructions before inserting the image. |
| Supplementary Figure |
| Insert figure title here by deleting or overwriting this text; keep title to a single sentence. |
| Insert figure caption here by deleting or overwriting this text; captions may run to a second page if necessary. To ensure accurate appearance in the published version, please use the Symbol font for all symbols and Greek letters. |
| Click inside this box and insert figure. For best results, use Insert menu to select a saved file; do not paste images. Source images must be in RBG color profile, at a resolution of 150—300 dpi. Optimize panel arrangement to a 2:3 height-to-width ratio; maximum online display is 600h x 900w pixels. Reduce empty space between panels and around image. Keep each image to a single page. Delete these instructions before inserting the image. |
| Supplementary Figure |
| Insert figure title here by deleting or overwriting this text; keep title to a single sentence. |
| Insert figure caption here by deleting or overwriting this text; captions may run to a second page if necessary. To ensure accurate appearance in the published version, please use the Symbol font for all symbols and Greek letters. |