A common genomic architecture for interacting with the external world

Margarita V Brovkina[1,*], Margaret A. Chapman[2,*], and E. Josephine Clowney[3,4,#]

1. Graduate Program in Cellular and Molecular Biology, University of Michigan Medical School, Ann Arbor, Michigan, United States of America

2. Neurosciences Graduate Program, University of Michigan Medical School, Ann Arbor, Michigan, United States of America

3. Department of Molecular, Cellular, and Developmental Biology, The University of Michigan, Ann Arbor, Michigan, United States of America

4. Michigan Neuroscience Institute, University of Michigan, Ann Arbor, Michigan, United States of America

*These authors contributed equally to this work.

To whom correspondence should be addressed:
jclowney@umich.edu

Abstract

The radiation of mammals at the extinction of the dinosaurs produced a plethora of new forms—as diverse as bats, dolphins, and elephants—in only 10-20 million years. Behind the scenes, adaptation to new niches is accompanied by extensive innovation in large families of genes that allow animals to contact the environment, including chemosensors, xenobiotic enzymes, and immune and barrier proteins. These large gene families share a common genomic organization and are often characterized by unusual modes of transcriptional regulation: they are clustered in tandem arrays in AT-biased isochores and exhibit tissue-specific and sometimes stochastic expression. Here, we use population genetic data and evolutionary analysis to examine the relationship between gene family diversification and genomic organization in mammals. First, we find that AT bias emerges as gene families expand in cis. Second, AT-biased, clustered gene families experience relatively low rates of *de novo* point mutation, and we suggest that multi-copy gene families have accrued high AT content due to relaxed selection compared to single-copy genes. Finally, we find that AT-biased, clustered gene families exhibit low rates of recombination and are depleted for binding of the recombination-seeding factor PRDM9. We posit that tolerance of point mutation and intolerance of recombination together result in depressed GC content of multi-copy versus single-copy genes. In turn, differential sequence content of gene blooms exerts a profound effect on their chromatin organization and transcriptional regulation.

Introduction

Reports of newly sequenced genomes frequently describe gene families that have "bloomed," undergoing explosive diversification in the focal species (Charkoftaki et al., 2019; Feyereisen, 2006; Nelson et al., 2013). Species-specific gene blooms generally occur *in cis*, resulting in large genomic arrays of dozens or even hundreds of genes. These tandem duplication events are thought to arise during gametogenesis via incorrect crossovers between paralogues or via non-homologous repair of chromosome breaks (Ohno, Susumu, 1970). The resulting expansions can confer unique life history traits recognized as definitive characteristics of the species: Examples include Cytochrome p450 genes for plant detoxification in koala and insects, lipocalins for pheromone communication in mouse, NK cell receptors for viral defense in bats, keratins for whale baleen, venom production in snakes, and amylase copy number for starch consumption in modern humans (Charkoftaki et al., 2019; Demuth et al., 2006; Feyereisen, 2006; Giorgianni et al., 2020; Holding et al., 2021; Johnson et al., 2018; Pavlovich et al., 2018; Perry et al., 2007; Sun et al., 2017). The definitive gene family of mammals, the caseins, arose through local duplication of enamel genes (Kawasaki et al., 2011).

Retrotransposition or ectopic exchange mediated by repetitive elements can seed new gene clusters elsewhere in the genome (Casola and Betrán, 2017; Lane et al., 2004). Ectopic exchange between different clusters during meiosis can have profound effects on genome structure (Freeman et al., 2006; Young et al., 2008). In most mammals, olfactory receptors are the largest gene family and often comprise 5% of protein-coding genes. Ectopic crossovers between distant olfactory receptor gene clusters have shaped mammalian chromosome evolution to the extent that ORs are often positioned at chromosome ends (Kim et al., 2017; Linardopoulou et al., 2005; Mefford et al., 2001; Newman and Trask, 2003; Rouquier et al., 1998; Trask et al., 1998; Yue and Haaf, 2006). In the mouse, we have shown that genes in tandemly arrayed families exhibit extremely high AT content in their promoters and are often located in AT-biased regions of the genome (Clowney et al., 2011).

GC content in mammalian genomes varies markedly at the megabase scale (Corneo et al., 1968). Since the earliest days of cytology, variation in staining patterns of DNA-binding dyes were apparent across the nucleus (heterochromatin and euchromatin) or along chromosomes (banding patterns). Banding patterns served as the original genetic maps and allowed scientists to link genetic phenotypes to physical positions in DNA (Holmquist, 1992). Banding patterns were found to reflect local variation in AT/GC content; Giemsa stains AT-biased regions of the genome, or G-bands, and Quinacrine stains GC-biased regions, or Q-bands (Bickmore, 2019; Filipski, 1990; Holmquist, 1992; Korenberg and Rykowski, 1988). Early reports suggested that G-bands were depleted for genes; that genes in G-bands tended to be "tissue-specific;" and that genes in Q-bands tended to be "housekeeping genes" (Clowney et al., 2011; Holmquist, 1992; Schug et al., 2005). After the human genome was sequenced, breaks between bands were found to correspond to local transitions in GC content, and bands were found to be composed of smaller "isochores" with locally consistent GC content (Costantini et al., 2006; Niimura and Gojobori, 2002). While isochore definition has been debated, a representative classification breaks the human genome up into ~3000 isochores of 100kb-5Mb that range from 35-58% GC (Cohen et al., 2005; Costantini et al., 2006; Lander et al., 2001).

The variation in GC content along the chromosome that is observed in mammals is not a general feature of metazoan, animal, or even vertebrate genomes. Both average GC content and the amount of local variation show wide divergence across clades (Duret and Galtier, 2009; Lynch, Michael, 2007), leading to adaptationist speculation that isochore structure serves a function related to warm-bloodedness (Bernardi et al., 1985). However, consensus has emerged that isochore structure results from GC-biased gene conversion (gBGC) following meiotic recombination (Duret and Galtier, 2009; Glémin et al., 2015). In this process, crossovers are statistically more likely to result in gene conversion towards more GC-rich sequences, resulting in a higher likelihood of inheriting higher-GC alleles. As stated in a recent paper by Sémon and colleagues, "The gBGC model predicts that the GC-content of a given genomic segment should reflect its average long-term recombination rate over tens of million years" (Duret and Arndt, 2008; Pouyet et al., 2017). In this model, the isochores themselves do not serve an adaptive function, but rather have emerged due to the molecular genetic forces of meiosis. Over evolutionary time, the GC-increasing effect of recombination is balanced by the AT-increasing effect of point mutation due to mutation of fragile cytosines to thymines (Fryxell and Zuckerkandl, 2000; Hershberg and Petrov, 2010; Hildebrand et al., 2010; Lynch, 2010; Simmen, 2008).

Genes with the highest AT content in their promoters tend to be located in AT-rich regions of the genome and have unique and consistent characteristics (Clowney et al., 2011). They are multi-copy families located in tandem arrays, are expressed in terminally differentiated cells, are cell surface or secreted proteins, and often have stochastic or variegated expression. These protein families are overwhelmingly involved in the "input-output" functions of an organism: sensation of the environment, protection from the environment, consumption of the environment, and production of bodily fluids. Human ORs were also shown previously to be located in AT-biased isochores (Glusman et al., 2001). AT- versus GC-skewed isochores differ in their replication timing and histone marks, they associate in nuclear space with other isochores of the same type, and they occupy different domains within the nucleus (Bickmore, 2019; Jabbari et al., 2019a; Ramani et al., 2016; Woodfine et al., 2004). The distinct treatment of AT- versus GC-rich isochores by the molecular machinery of the mammalian cell means that the *genes*

located in AT-rich isochores must experience distinct molecular events from those located in GC-rich isochores.

Here, we ask how genes with outward- versus inward-looking functions came to be partitioned into AT- versus GC-rich regions of the mammalian genome. We find that AT content rises as gene clusters expand *in cis*. Using population genetic data, we analyze allelic variation, patterns of point mutation, and recombination in human genes located in AT- versus GC-biased isochores. We find that genes in AT-biased isochores exhibit relatively low rates of *de-novo* point mutation but have accumulated point mutations over evolutionary time. These gene clusters also exhibit low rates of recombination, low binding of the recombination-seeding factor PRDM9, and a lack of CpG islands. Tolerance of point mutation is expected to shift GC content down over evolutionary time due to deamination of cytosine (Fryxell and Zuckerkandl, 2000; Hershberg and Petrov, 2010; Hildebrand et al., 2010; Lynch, 2010; Simmen, 2008). Intolerance of recombination would prevent gBGC from shifting GC content back up. We propose that recombination is deprecated in large gene families due to the dangers of chromosome rearrangements and because it could separate genes in large families from locus control regions they depend on for expression. Reduced recombination and relaxation of selection on point mutations strands multi-copy gene blooms in a well of low GC content. This sequence context supports exotic forms of highly tissue-specific transcriptional regulation (Armelin-Correa et al., 2014; Clowney et al., 2012; Monahan et al., 2019; Tan et al., 2019, 2021).

Results

Characterizing a set of human isochores and their gene contents

In human, isochores are sub-band-level structures on the order of hundreds of kilobases (kb) defined by transitions in local GC content (Figure 1A, B). Previous isochore annotations were performed on earlier genome assemblies, broke the genome into chunks of arbitrary size, and/or used manual inspection to define isochore ends (Cohen et al., 2005; Costantini et al., 2006; Jabbari and Bernardi, 2017). To call a robust reference set of isochores for the human hg38 assembly, we used a segmentation algorithm which detects transitions in GC content to call isochore boundaries (Gao and Zhang, 2006). The 4328 resulting isochores ranged from ~30-70% GC and most were between 100kb and 5Mb (Figure 1C, D, Supplemental Tables 1 and 2). To define the large-scale sequence context of each gene, we annotated the home isochore of each gene in the NCBI MANE set of intact, protein-coding genes. The MANE set includes one promoter and splice isoform per gene and omits pseudogenes and complex "gene parts" such as V, D, and J segments of the B and T cell receptors (Morales et al., 2022). On average, higher-GC isochores were more likely to contain protein-coding genes and had higher gene density (Figure 1C, D, Figure S1A, B) (Holmquist, 1992). On the other hand, AT-rich isochores often contain tandemly arrayed gene blooms.

What sorts of gene families have bloomed in AT-rich isochores? We selected genes located in isochores <40% GC (~25% of genes) and searched for common prefixes. Gene families with at least four members in high-AT isochores were overwhelmingly involved in chemosensation (e.g. *OR, TAAR, TAS2R*); xenobiotic metabolism (e.g. *AMY1, UGT2, ADH*); and immune, defense, and barrier functions (e.g. *KLR, IFN, DSC/DSG*). Moreover, members of these gene families tended to be sharply enriched in high-AT isochores (Figure 1E). While immunoglobulin parts do not appear in the MANE gene set, arrays of immunoglobulin V regions are also highly AT-biased (Figure S1E).

To compare features across isochores, we ordered isochores by GC% and divided them into ten groups of ~400 (deciles). The variation in gene density across isochore GC% deciles can be seen in Figure 1F. To test if high-AT isochores were statistically more likely to contain tandem arrays, we used Shannon's H to calculate the diversity of gene names in isochores with at least 10 genes (Figure 1A, B, G, Figure S1C, D). Indeed, genes located in the highest-AT isochores (decile 1) were less diverse than those located in high-GC isochores (decile 10), suggesting that gene-rich, high-AT isochores house tandem arrays (Figure 1G). Examples of high-GC isochores with diverse gene members and high-AT isochores with repetitive gene members are shown in Figure 1A, B and Figure S1C-E.

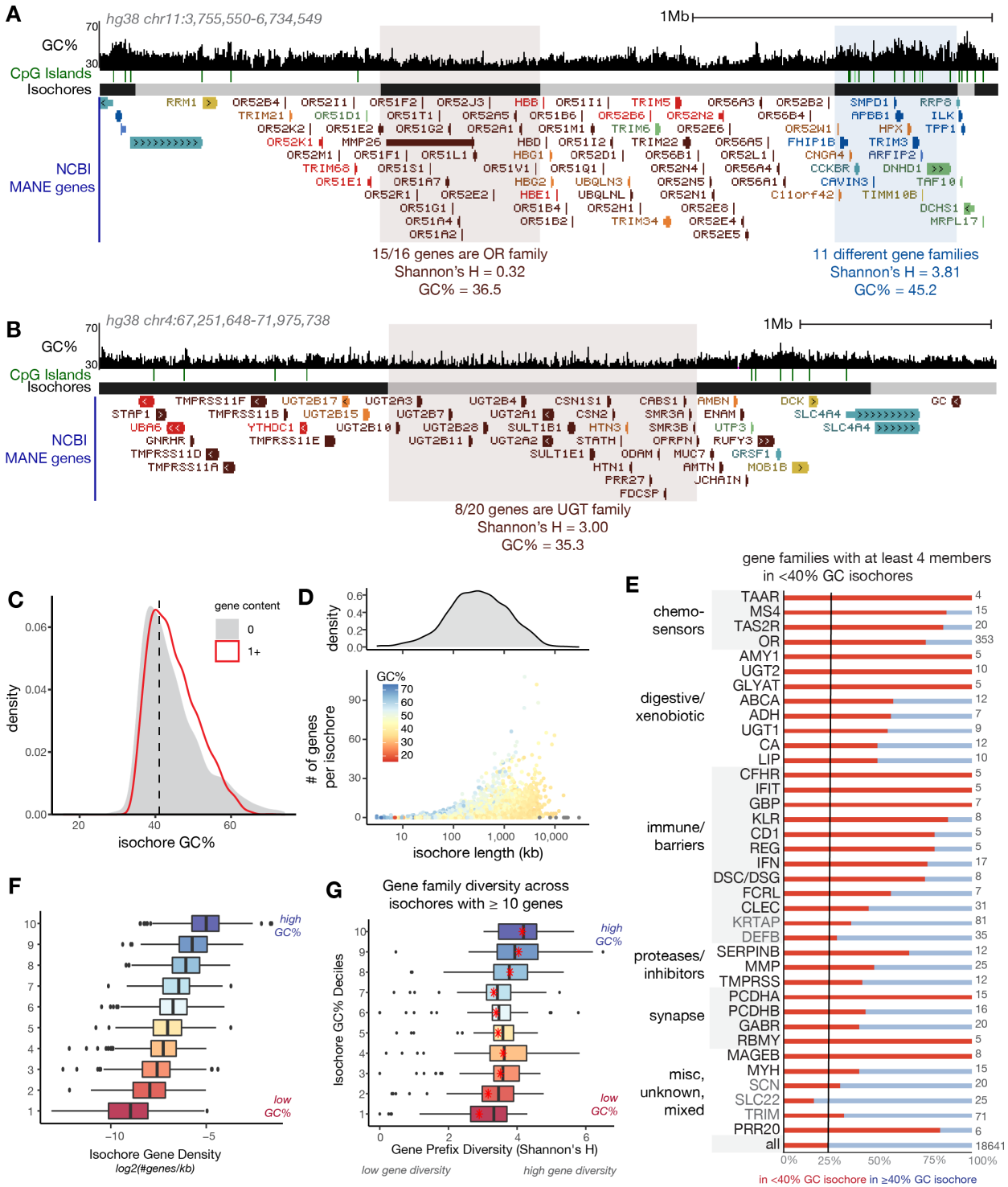


Figure 1: AT-rich isochores in the human genome contain tandem arrays of genes with outward-looking functions (A, B) UCSC Genome Browser screenshots showing GC% trajectory (Clawson, 2018), CpG islands (Micklem and Hillier, 2006), our isochore calls, and simplified gene models. Gene models are colored according to k-means clusters described below. Gene name prefix diversity (Shannon's H) is shown for the highlighted isochores. (C) Distribution of GC% for isochores with and without genes. (D) Relationship between isochore

length, gene content, and GC%. (E) Gene families with at least four members in <40% GC isochores are shown. Red bars depict portions of genes with that prefix that are located in <40% GC isochores. Gene family prefixes shown in black text are enriched in AT-rich isochores, while those shown in grey text (e.g. KRTAP, TRIM, SCN) have multiple family members in AT-rich isochores but are not enriched there. Functions of these gene families are marked at left, and number of genes with that prefix in the MANE set are shown at right. (F) Boxplots representing gene density across isochores binned by GC%. Black lines show medians. (G) Boxplots representing gene name diversity (Shannon's H) of gene-rich isochores. Isochores are binned into deciles according to GC%. Red points indicate the mean prefix diversity of the decile; black lines show medians. Here and throughout, most groups are statistically different from one another, except for adjoining groups; full statistical comparisons are presented in Supplemental Tables 3-5.

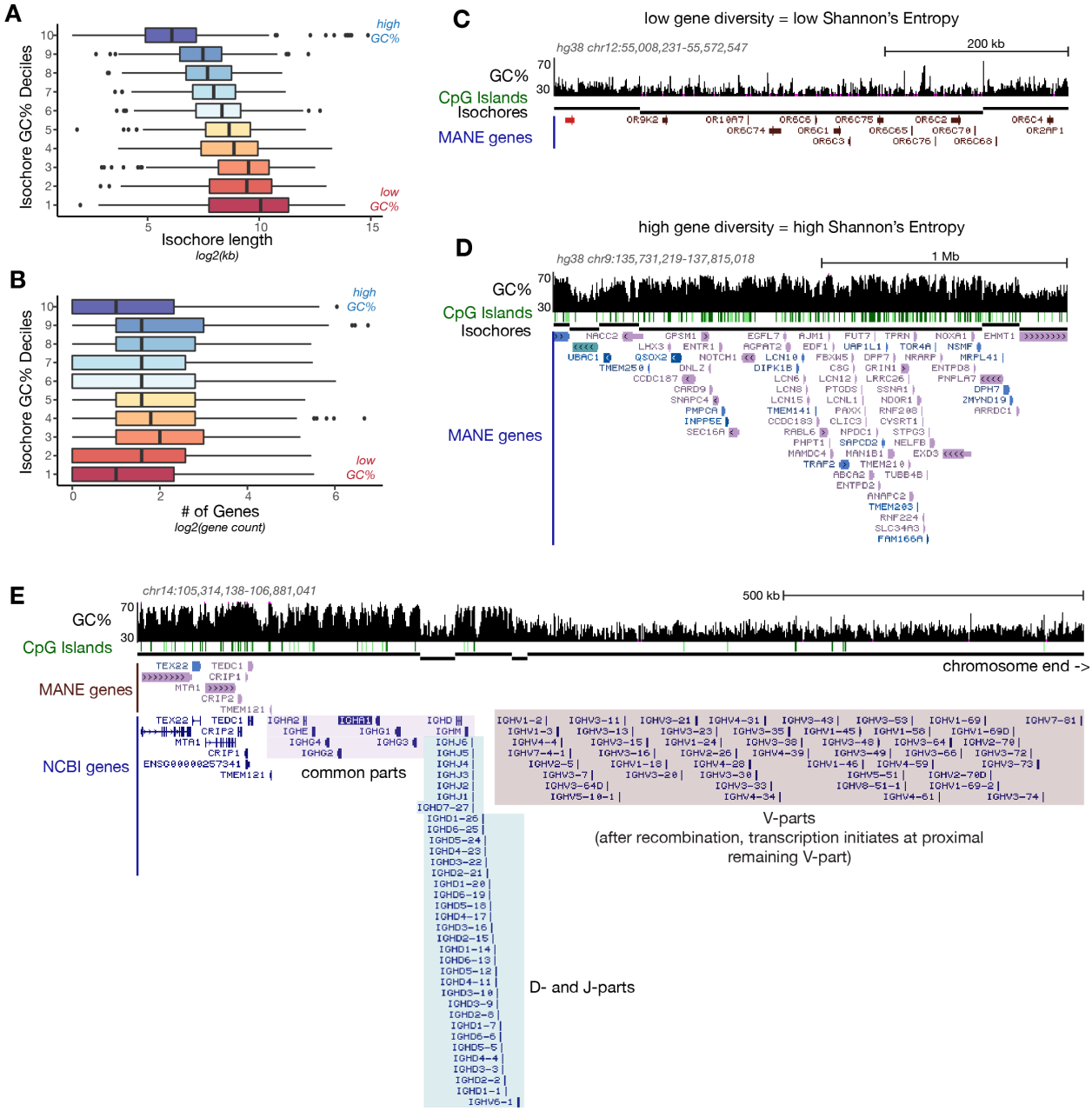


Figure S1 Characterization of human genomic isochores. (A-B) Box and whisker plots describing relationship between isochore GC% and isochore length (A) and gene number (B). Isochores are binned into deciles according to GC content with decile 10 representing high GC and decile 1 representing high AT. (C-E) UCSC Genome Browser screenshots showing additional example isochores. Gene models are colored according to k-means clusters described below. (C) This isochore is AT-rich and contains a gene bloom; all the genes in this isochore have the same prefix, so prefix diversity (Shannon's H) is low. (D) This isochore is GC-rich and contains genes from a variety of families, so diversity is high. (E) The *IGH* locus, containing V, D, J, and common regions of human IgH. V, D, and J regions are classified as "gene parts" and are not represented in the MANE set; however, the V repeats, where transcription initiates, are in an AT-rich isochore and lack CpG islands.

Categorizing human genes according to local patterns of AT/GC content

We show above that tandemly arrayed genes serving outward-looking functions in the human are located in AT-rich isochores, as they are in rodents (Clowney et al., 2011). Because regulatory and transcribed regions comprise a small fraction of the genome and could diverge from the sequence content of the isochore, we also sought to examine the local AT/GC sequence features of individual human genes. We calculated GC% in 50bp sliding windows along the transcriptional unit (transcription start site to transcription end site, TSS-TES) and 1kb flanking regions for genes in the MANE set (Clawson, 2018; Morales et al., 2022). We then used iterative k-means clustering to group the 18,641 MANE genes into 3x3 sets (Figure 2A, S2A). Our analysis here includes introns, but clustering on exons and flanking regions produced similar results (Figure S2B). The top-level clusters (1-, 2-, 3-) reflect overall differences in AT content in different genes (Figure S2A), while the subclusters (1.1, 1.2, 1.3 etc, Figure 2A) reflect variation in AT content of the promoter, transcriptional unit, and 3' region. To capture both the broad isochore context of genes and their local sequence features, we use both the isochore AT/GC metric and the local sequence-based k-means clustering throughout this study; each gene in the MANE set is assigned uniquely to one home isochore and one k-means cluster (isochore deciles 1-10, red-blue palette; k-means clusters 1.1-3.3, rainbow palette). Cluster and isochore assignments and other gene-linked data are provided in Supplemental Table 6.

While many interesting patterns emerge in this analysis, the lack of GC enrichment at the promoters of genes in cluster 3.3 (and to a lesser extent 3.1) is particularly stark (Figure 2A). Based on high-confidence annotation of transcription start sites, we showed previously that mouse olfactory receptor genes share this GC-poor pattern (Clowney et al., 2011). At that time, the TSS's of other highly tissue-specific genes had not been mapped. Current human annotations in the MANE set capture true TSS's for most genes, thus we are confident that we are not missing promoter GC enrichment due to incomplete annotations.

We next examined how the isochore GC content of a gene relates to the sequence content of its promoter and coding region. We found that local GC content of genes predicted the GC content of their home isochore (Figure 2B). Isochore GC% also correlated closely with the GC-content of a gene's flanking regions (gene extent with 25kb on each side) (Figure S2C). Individually, promoter and coding sequence GC% were also positively correlated with isochore sequence content, but the correlation coefficients were weaker: A subset of genes in AT-rich isochores have GC-rich promoters or GC-rich coding sequences (Figure S2D-F).

Genes in clusters 3.1 and 3.3, lacking GC enrichment in their promoters, were highly enriched for the same functional categories as were genes in AT-rich isochores: chemosensation, xenobiosis, and defense/barriers. Indeed, as can be seen in Figure 1A, B and Figure S1D, sometimes entire arrays were members of cluster 3.3 (brown color). To systematically test this, we plotted the promoter GC content distribution of genes we term "outward-looking" (chemosensation, defense, xenobiosis, barriers) versus "inward-looking" (e.g. transcription, kinase function, morphogens). Outward-looking genes have AT-rich promoters while inward-looking genes have GC-rich or average promoters (Figure 2C). To comprehensively describe the functions of these gene families, we manually annotated common prefixes and enrichment of genes in cluster 3.3 (Figure 2D). This group included all the chemosensory families, many sets of digestive and detoxifying enzymes, and several receptor arrays in the immune system and skin. It also included clustered protocadherins, which share transcriptional regulation patterns with chemosensors. Finally, in accordance with the preponderance of tandemly arrayed genes found

in cluster 3.3, we found that genes in this cluster were housed in fewer unique isochores and had lower prefix diversity than those in the other clusters (Figure 2E, S2G).

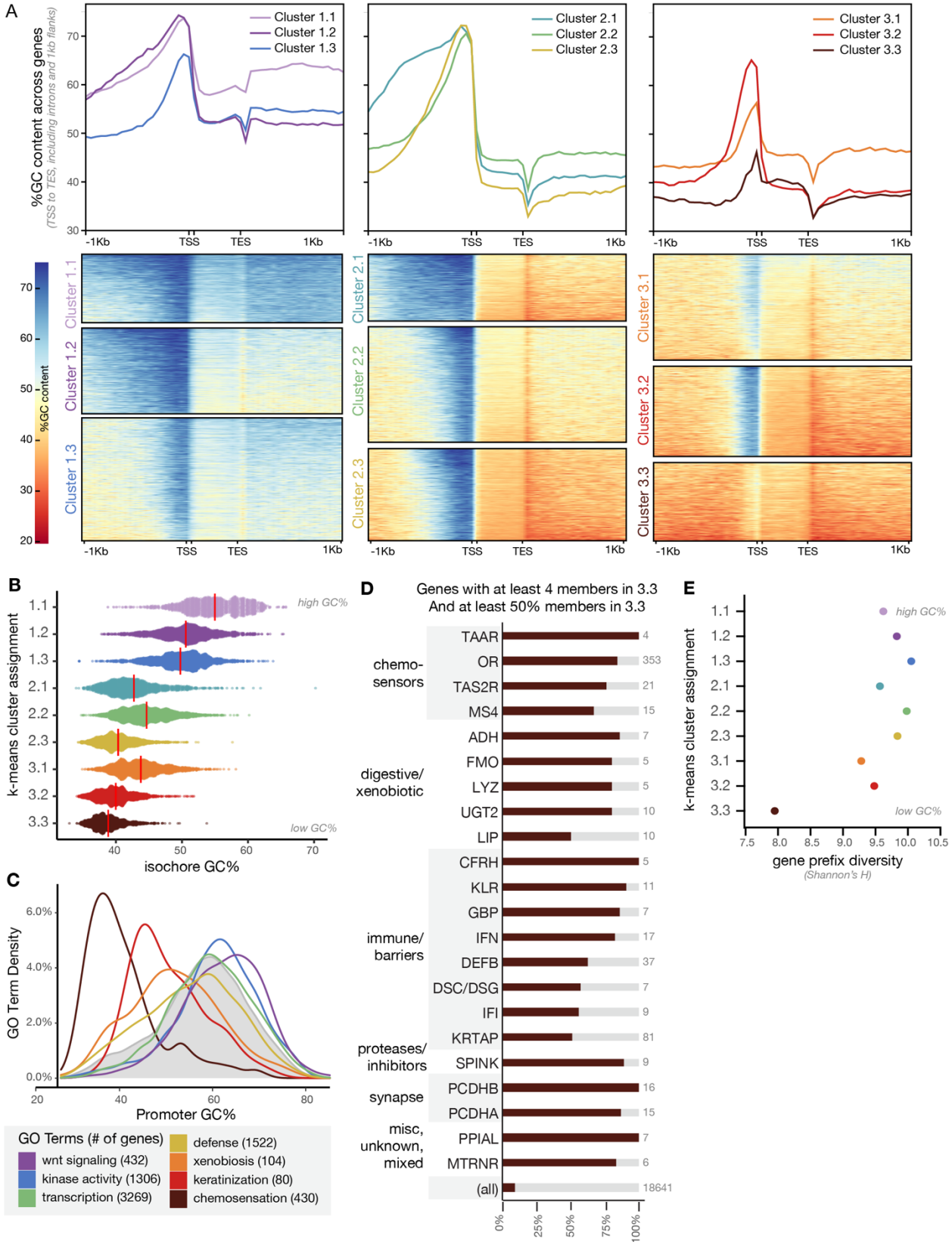


Figure 2: Genes with outward-looking functions have high local AT content. (A) GC content trajectory for human protein-coding genes in the MANE set. Genes were subdivided by iterative k-means clustering. At top, the average GC content trajectory for each k-means cluster is shown as a line graph. At bottom, each gene is a row and GC content across the transcriptional unit and flanking regions is depicted from red (high AT) to blue (high GC). Rainbow colors assigned to each k-means cluster here will be used throughout. (B) Relationship between k-means cluster assignment and home isochore GC% for each gene in the MANE set. Red lines depict medians. (C) GO term distribution by promoter GC content for genes in the MANE set. Genes with AT-rich promoters are overwhelmingly enriched for immune, barrier, chemosensory, and xenobiotic functions. Genes with GC-rich promoters are enriched for developmental and intracellular functions. Grey shading shows promoter GC content distribution of the whole MANE set. (D) Gene name prefixes enriched in cluster 3.3. Families shown have at least four members in cluster 3.3; proportion of the family located in this cluster is depicted in brown bars. Less than 10% of genes are in cluster 3.3 ("all"). (E) Gene prefix diversity (Shannon's H) is lowest in cluster 3.3.

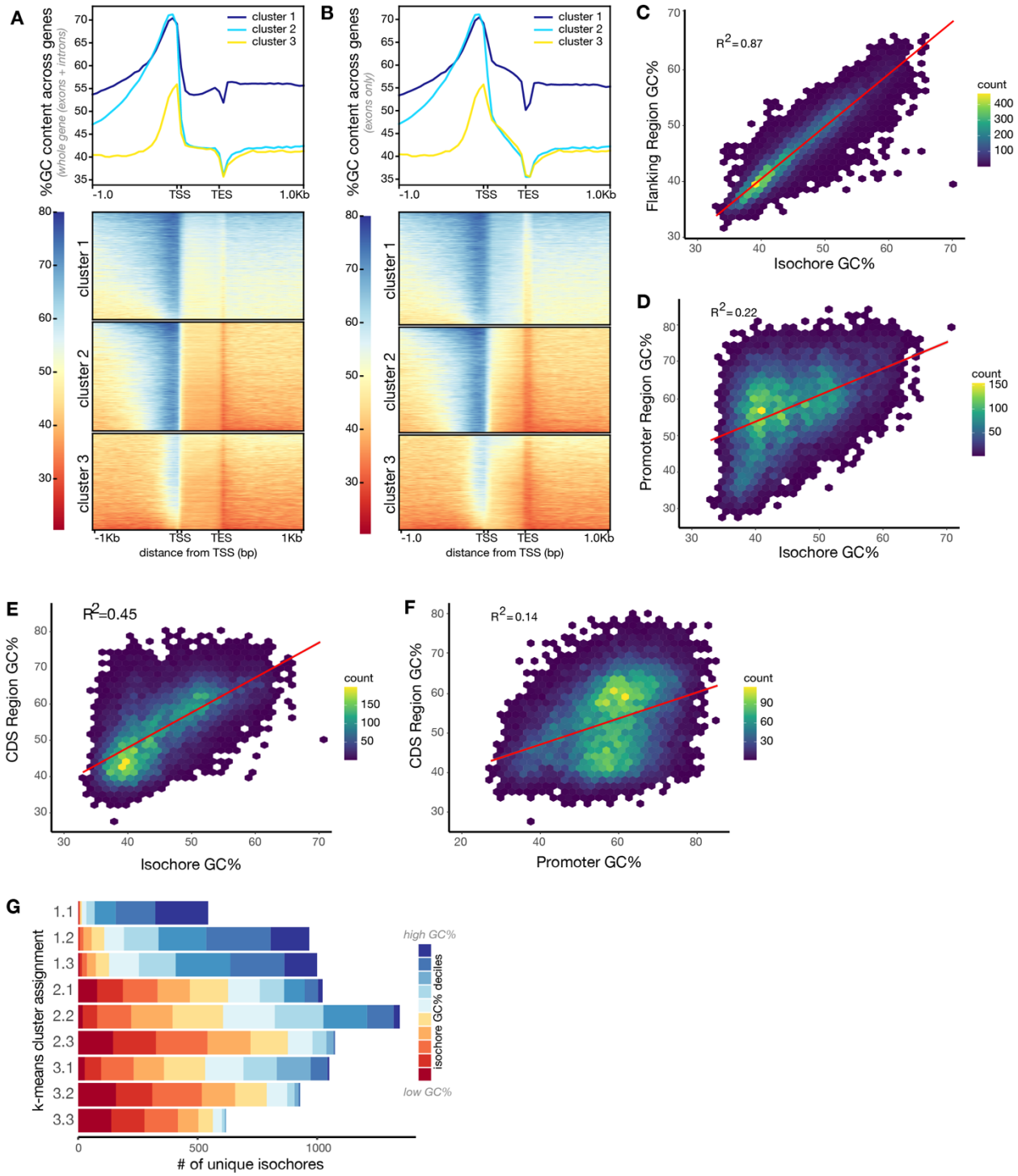


Figure S2: Relationship between GC content of local gene features and the home isochore. (A-B) Heatmaps displaying GC% over 50bp sliding windows calculated across genes with (A) and without (B) introns. Clusters were determined by k-means. Summary line plots depict mean GC% across clusters. <10 genes switched clusters when introns were excluded. (C-F) 2D histograms depicting the relationship between promoter, coding region, flanking region, and isochore GC percent for each gene in the MANE set. "Flanking region" is 25kb upstream of TSS to 25kb downstream of TES. Correlation coefficient (R^2) and trend (red line) are shown. (G)

Number of unique isochores housing genes in each k-means cluster. Genes in the more extreme k-means clusters are contributed by fewer isochores.

Emergence of high AT content during evolutionary expansion of tandem arrays

AT bias could have emerged as gene families expanded or could have been pre-existing and supported molecular mechanisms of gene duplication. To assess this, we examined sequence context of related genes across extant vertebrates. We took advantage of the emergence of a large olfactory receptor cluster near the hemoglobin β gene cluster on human chromosome 11 (Figure 1A). The hemoglobin β genes are thought to have been relocated by transposition in an amniote ancestor of reptiles and mammals to a region near *DHCS1*, *STIM1*, *RRM1*, and *FHIP1B* named "DS" (Hardison, 2012). In many species, a cluster of olfactory receptors is observed in this region, and their synteny with the *HBB* genes and DS genes allows us to track this cluster across evolution. In mammalian outgroups, this region contains few or no ORs and has higher GC content than the genomic average (Figure 3A, B, Figure S3). In mammalian genomes, there are dozens or hundreds of OR genes in this region, GC content is the same or lower than the genome-wide average, and GC content is negatively correlated with OR number and cluster length (Figure 3A-D, Figure S3). This suggests that the high AT content observed in mammals is not ancestral, but rather it emerged as the gene family bloomed.

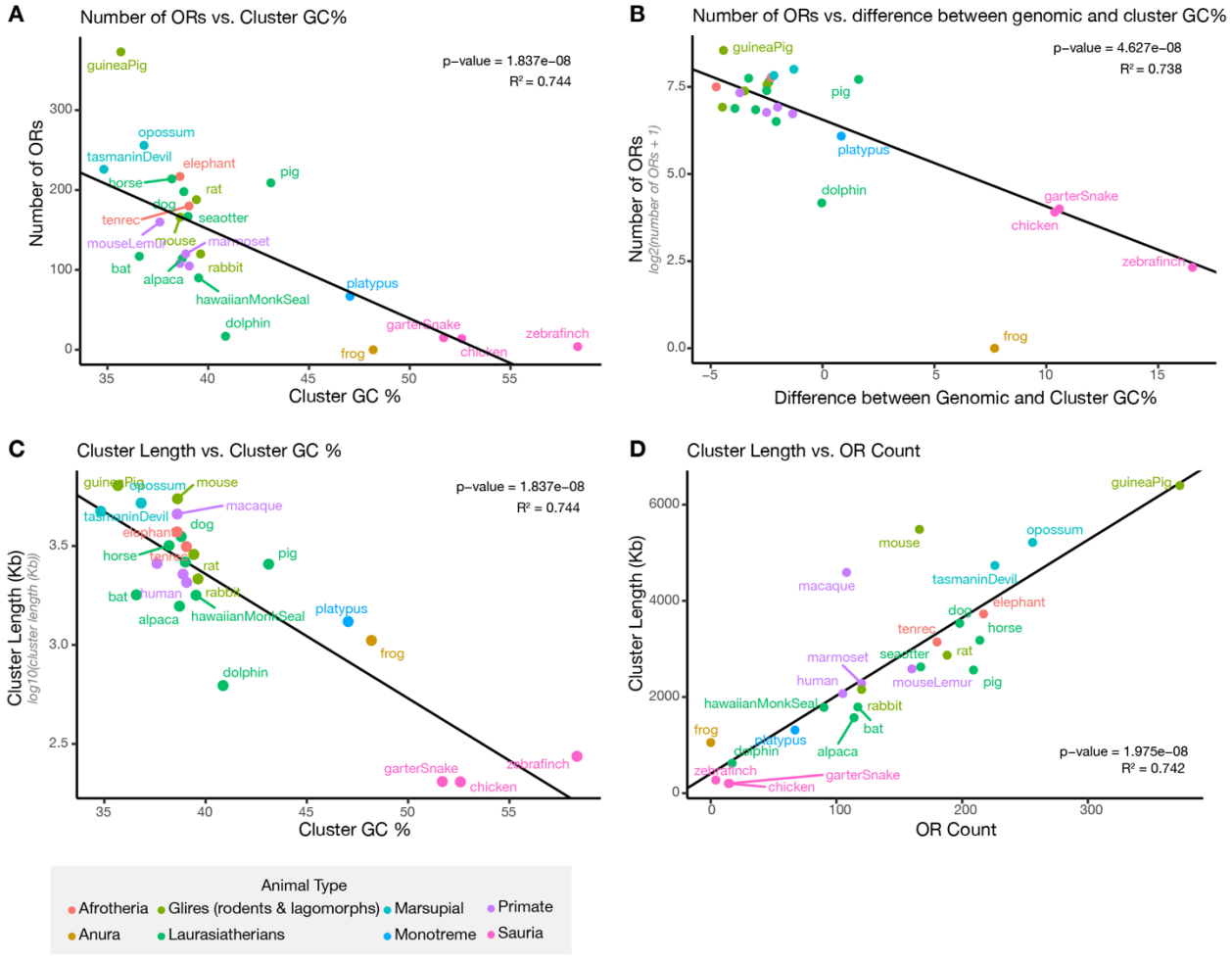


Figure 3: AT content rises during tandem array expansion. (A, B) Comparison of olfactory receptor number in the Hemoglobin β (*HBB*) cluster with raw cluster GC% (A) or GC% difference from the genomic average (B). (C) Comparison of cluster length and GC%. (D) Comparison of OR count and cluster length. Throughout this figure, OR count includes intact genes as well as low-quality or pseudogenes. Cluster GC% is calculated from the transcription termination site of *RRM1* to the transcription start site of *FHIP1B*. p-values report results of phylogenetic least squares analysis.

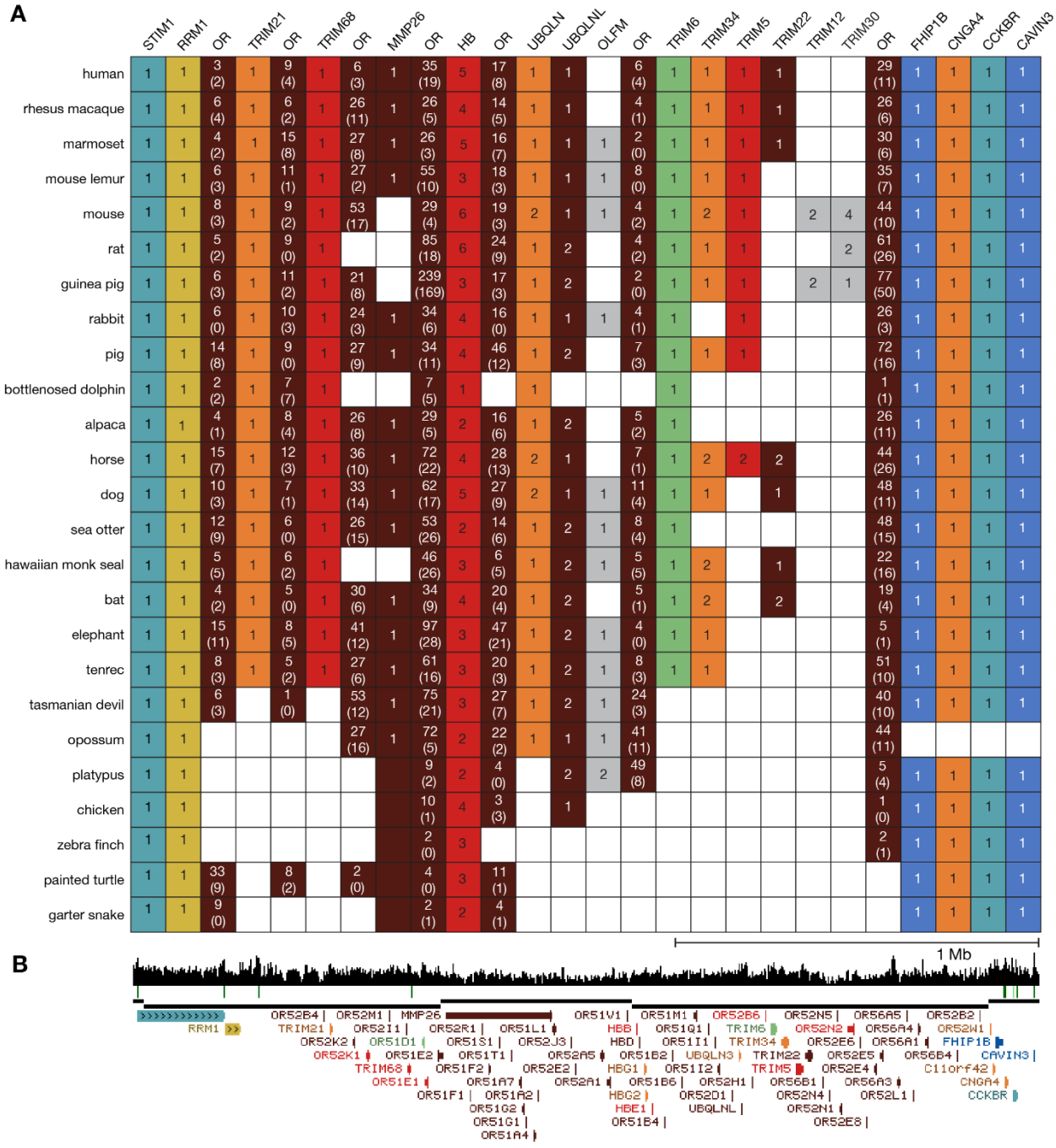


Figure S3: Distribution of olfactory receptor genes and pseudogenes flanking the *HBB* cluster in mammals and non-mammalian vertebrates. (A) Each row depicts genes in the *HBB*

cluster environs of a particular vertebrate; species are ordered by distance from human. Different genes or families shown at top are depicted in different colors (corresponding to hg38 k-means clusters from Figure 2), with the total number of local sequences with homology to that gene family marked. Subset of hits which are flagged by ENSEMBL as low-quality or pseudogenes are in parentheses. (B) UCSF Genome Browser screenshot showing the hemoglobin β cluster on human chromosome 11. As described previously, the *HBB* cluster is flanked by olfactory receptor genes and these are bracketed by conserved single-copy genes, including *STIM1*, *RRM1* and *FHIP1B* shown here, that are syntenic with the *HBB* cluster since before mammals branched (Hardison, 2012). *DCHS1*, not shown here, is distal to *CAVIN3*.

Allelic Variation

Many types of outward-facing genes have been reported to exhibit extreme allelic diversity or rapid divergence across species, and polymorphisms in these genes underlie human phenotypic variation in drug metabolism, sensory perception, and immune response (Charkoftaki et al., 2019; Nei et al., 2008; Niimura and Nei, 2007; Schwartz et al., 2017; Semple and Dorin, 2012; Shelton et al., 2022; Sun et al., 2017; Tan and Low, 2018; Thomas, 2007). Colloquially, outward-looking genes are so diverse in copy number and sequence that a first step in GWAS is often to "throw out the ORs." To systematically examine the degree of coding sequence variation in human genes grouped by AT/GC content, we used a dataset of rare single nucleotide variants ascertained from whole exome sequencing of >100,000 unrelated people (gnomAD v2.1.1, Figure 4A, B) (Karczewski et al., 2020). The ratio of protein-altering versus synonymous variants is positively correlated with AT content, with genes in AT-rich isochores and k-means cluster 3.3 highly enriched for potentially functional variants. We note that use of rare variants profoundly understates the allelic variety in outward-looking genes, which exhibit radical common variation as well. For example, any two humans are estimated to have function-changing variation in 30% of their olfactory receptor genes (Mainland et al., 2014; Trimmer et al., 2019).

Previous reports, including ours, have speculated that partitioning inward- and outward-looking genes into different parts of the genome could enable a higher ongoing mutation rate in outward-looking genes (Chuang and Li, 2004; Clowney et al., 2011; Grimwood et al., 2004). Recent studies of mutation accumulation in isogenic *Arabidopsis* lines have also suggested that mutation rate is biased by gene features and by chromatin context in the gamete progenitors (Monroe et al., 2022). Is the enhanced functional allele diversity of AT-rich genes due to distinct patterns of mutation or selection?

To examine modern patterns of mutagenesis in genes relative to AT/GC content, we examined synonymous variants from gnomAD. We see that AT-rich genes have *fewer* synonymous variants across unrelated people than do GC-rich genes, refuting the idea that as a group, outward-looking genes experience a higher rate of ongoing mutation (Figure 4C, D). Rather, the combination of enhanced functional variation combined with low synonymous variation suggests that the alleleic diversity in AT-biased genes arose due to tolerance of historic mutations, rather than increased mutation rates—that is, functional alterations of these genes are not deleterious like they might be in a single copy gene and are not removed from the population (Figure S4A-D). This is consistent with point mutation rate being grossly driven by deamination of cytosine, especially in the $C^{me}G$ context, and the lack of remaining CpG sites in AT-biased genes (Fryxell and Zuckerkandl, 2000; Hershberg and Petrov, 2010; Hildebrand et al., 2010; Lynch, 2010; Simmen, 2008; Sved and Bird, 1990). Mutation rates in AT-biased genes could

still be subtly skewed relative to sequence-based expectations. Indeed, we see that olfactory receptor genes have higher rates of variant calls than do other AT-rich genes, and that this rate is higher than predicted from sequence alone (Figure S4E-G). This inflated rate could reflect an unknown, active mutagenic process but could also result from incomplete knowledge of the full human "OR-ome" and incorrect assignment of homology relationships.

Finally, we used the gnomAD metric of "loss of function intolerance" (pLI) as a measure of purifying selection—deleterious mutations in these genes are depleted from the population (Karczewski et al., 2020; Lek et al., 2016). Genes predicted to be "loss of function intolerant" were enriched in GC-rich isochores, had GC-rich promoters, and were absent from cluster 3.3 (Figure 4E-G); AT-skewed, outward-looking genes are relatively loss of function tolerant.

Our analysis of functional and synonymous variation in unrelated humans suggests that AT-biased genes are subject to weaker selection than GC-biased genes, not higher levels of mutation. To test this in a different way, we sought to measure the rate of *de novo* mutations that occur in genes in AT- versus GC-biased regions of the human genome. Whole genome sequencing of two parents and a child (trios) enables detection of *de novo* mutations (DNMs). We used a recent dataset that compiles DNMs from >11,000 trios, nearly all those who have been sequenced to date (Rodriguez-Galindo et al., 2020). While these data remain sparse relative to the size of the genome (~700,000 total DNMs), DNMs approximate a record of mutagenesis that has yet to be operated on by selection. Pooling DNMs across each isochore and across transcriptional units (TSS to TES) within that isochore, we found that both genic and isochore-wide DNMs were more common in higher-GC isochores (Figure 4H-I, S4H-I). This is consistent with sequence-based predictions, prior findings, and our analysis of gnomAD synonymous variants within genes (Francioli et al., 2015; Jónsson et al., 2017).

Together, comparison of *de novo* mutations in meiosis and single nucleotide variants in unrelated humans both support the conclusion that AT-biased genes experience fewer contemporary mutations and at the same time better tolerate the mutations that do occur. We conclude that the allelic diversity in clustered, AT-biased genes is due to relaxed selection on these genes, rather than excess mutations. AT content would therefore have increased over time in these gene families due to evolutionary tolerance of C->T and especially CpG->TpG mutations. 8-oxoguanine, another source of point mutations, would also preferentially affect GC-rich sequences (Ohno et al., 2006). One attractive model is that as a gene cluster expands in size and the function of that gene family is partitioned over more and more members, selection becomes weaker and weaker on individual family members, allowing these clusters to attain higher AT content due to drift.

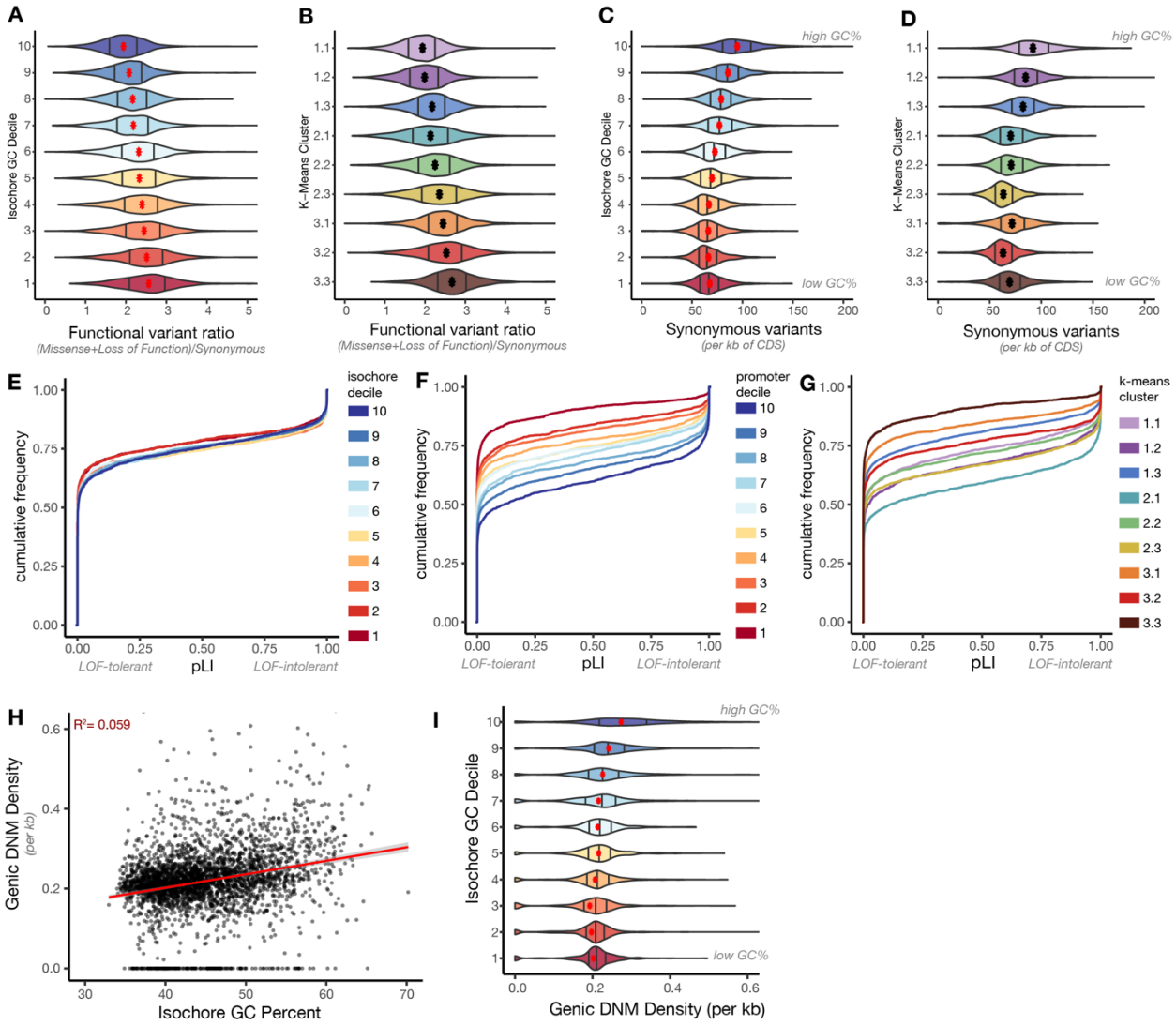


Figure 4: AT-rich genes have high functional diversity despite experiencing moderate mutation rates in the present. (A,B) Ratio of functional (missense plus loss of function) versus synonymous rare variants in the MANE gene set identified in gnomAD v2.1.1 exome sequencing of >100,000 unrelated individuals (Karczewski et al., 2020). Genes are binned by isochore decile (A) or k-means cluster (B), and dots indicate means. gnomAD rare variants are defined by < 0.1% allele frequency. (C, D) Raw counts of rare synonymous variants per gene in gnomAD v2.1.1 binned by isochore decile (C) or k-means cluster (D). (E-G) Cumulative frequency distribution plots of gnomAD pLI (likelihood that a gene is loss-of-function intolerant) relative to a gene's isochore GC% (E), promoter GC%(F), or k-means cluster assignment (G) (Karczewski et al., 2020). (H,I) Number of *de novo* point mutations observed per kb across the genes (TSS to TES) within an isochore relative to isochore GC% (H) and isochore GC% binned by decile (I). ~700,000 DNM calls are pooled from all ~11,000 trios sequenced to date (Rodriguez-Galindo et al., 2020).

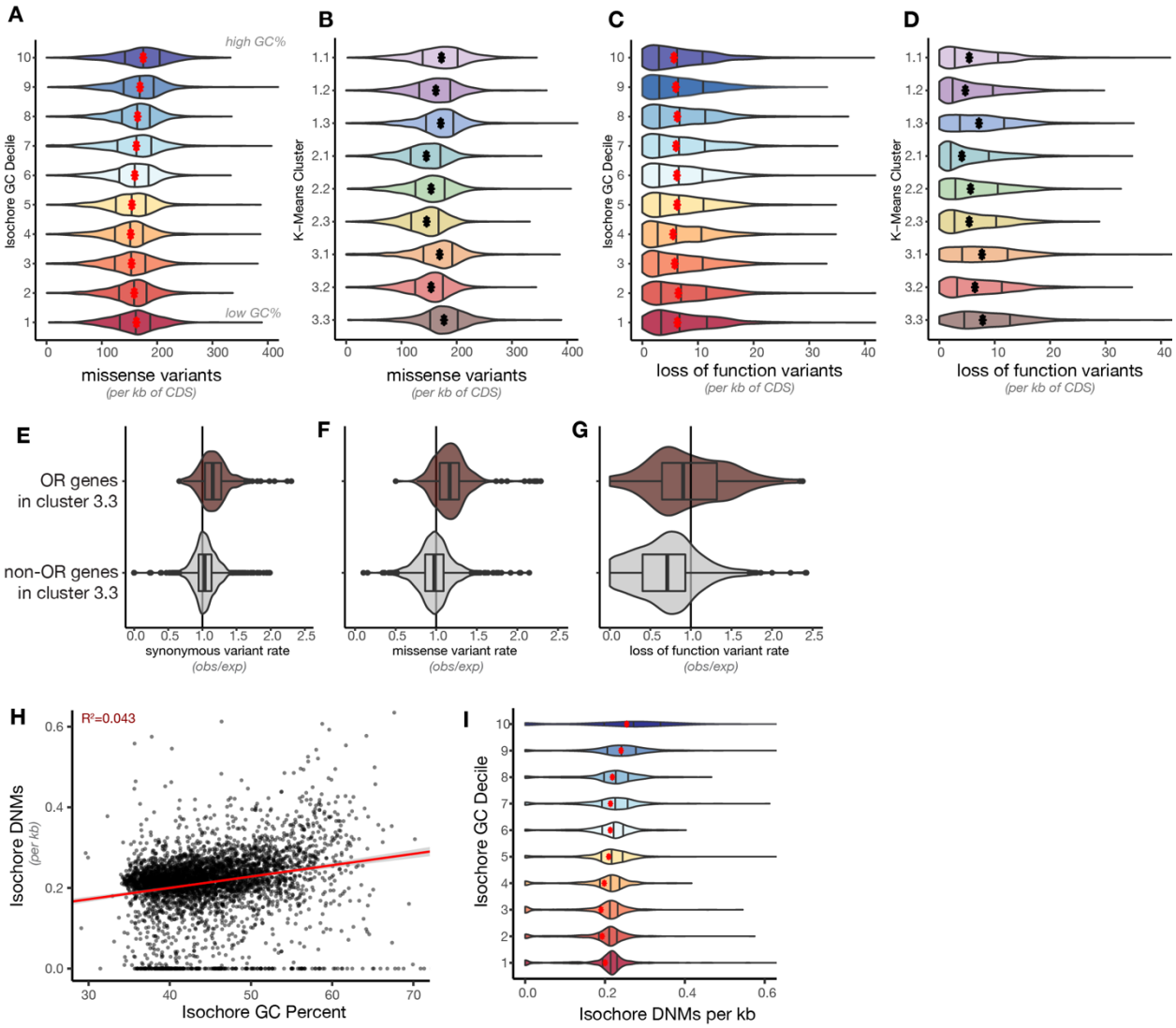


Figure S4: Raw rates of loss of function and missense variation. Raw gnomAD v2.1.1 counts of missense (A, B) and loss-of-function (C, D) SNVs across MANE genes binned by isochore GC% decile (A, C) or k-means cluster (B, D). (E-G) Ratio of observed over expected rates of synonymous variants (E), missense variants (F) and loss of function variants (G) of genes in k-means cluster 3.3, split by OR and non-OR genes. (H,I) Number of *de novo* point mutations observed per kb in each isochore plotted relative to isochore GC% (H) and isochore GC% binned by decile (I). ~700,000 DNM calls are pooled from all ~11,000 trios sequenced to date (Rodriguez-Galindo et al., 2020).

Recombination and PRDM9 Binding

While point mutations reduce GC content, meiotic recombination increases GC content due to GC-biased gene conversion, in which recombination is statistically more likely to resolve to the higher-GC allele (Duret and Arndt, 2008; Duret and Galtier, 2009). We therefore examined recombination patterns from whole genome sequencing of trios with respect to AT/GC content (Halldorsson et al., 2019).

Crossovers appeared rare within gene blooms (Figure 5A, S5B). We calculated a relative crossover rate for each isochore and found that AT-rich isochores experienced less maternal and

paternal crossovers than GC-rich isochores, as has been previously observed, though maternal crossovers were sharply diminished in the highest-GC isochores (Figure 5B) (Halldorsson et al., 2019; Holmquist, 1992; Jabbari et al., 2019b; Kong et al., 2010). We noticed that these low-crossover, high-GC isochores were often at chromosome ends, where maternal recombination has been shown to be low (Lee et al., 2011). To systematically examine recombination relative to each gene along the chromosome, we generated a Manhattan plot of crossover rate for each gene and its flanking regions (Figure 5C, S5A). This highlights the higher recombination of genes located in GC-rich isochores, except for maternal recombination at chromosome ends.

While recombination is generally directed away from genes, some genes experienced crossovers. We plotted the isochore and k-means cluster distribution of genes with the 10% highest internal crossover rate (TSS-TES, Figure 5E, S5D, E). These genes were in GC-rich isochores and excluded from AT-rich k-means cluster 3.3. Together, these analyses suggest that gene blooms located in AT-rich regions of the genome experience low current crossover rates. As predicted by the gBGC theory, the high AT content of these gene blooms can also be considered to reflect low historical rates of recombination (Pouyet et al., 2017). Previous modeling suggests that once variation in AT/GC content starts to emerge, it can be self-reinforcing via positive feedback (Fryxell and Zuckerkandl, 2000).

In many vertebrates, including humans, crossovers are seeded by binding of PRDM9 to its target site (Baudat et al., 2010; Myers et al., 2010; Parvanov et al., 2010). To test whether variation in observed recombination across AT/GC categories is due to differential seeding, we examined PRDM9 binding data from human cells (Figure 5D, F)(Altemose et al., 2017). We observed a striking depletion of PRDM9 binding from AT-biased genes and isochores, in line with its GC-rich DNA binding motif and with previous analyses (Jabbari et al., 2019b). This suggests that recombination is less likely to initiate in AT-biased regions of the genome. We note that many animals have lost PRDM9, and in the absence of PRDM9, recombination is often seeded at CpG islands (Baker et al., 2017). As described below, AT-biased gene families also lack CpG islands, and would thus experience low recombination seeding with and without PRDM9. These results suggest that AT-bias in tandemly arrayed gene clusters has emerged due to either low historical recombination or selective intolerance of recombination, and that these gene clusters have evolved an avoidance of recombination seeding.

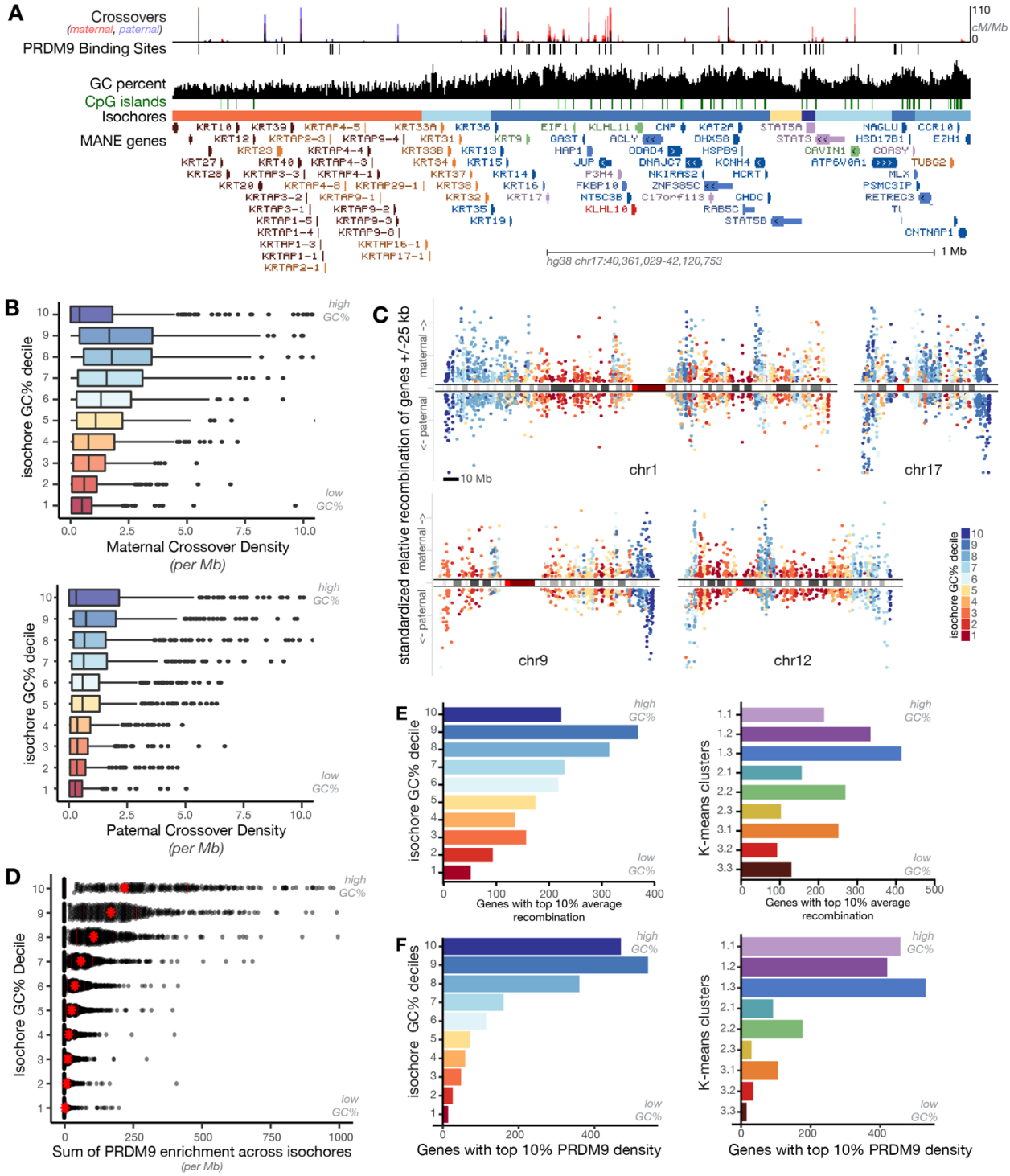


Figure 5: Crossovers are directed away from AT-rich isochores and genes. (A) UCSC genome browser screenshot of human KRTAP cluster showing recombination rates (maternal meiosis in red, paternal meiosis in blue) and PRDM9 binding calls (black). (B) Maternal (top) and paternal (bottom) crossover rates in isochores binned by GC percent. Crossover calls are from deCODE (Halldorsson et al., 2019). (C) Manhattan plot of standardized maternal (above 0) and paternal (below 0) recombination rate for each gene with its 25kb flanking regions. Genes are colored according to the GC content of their home isochore (Red: high AT. Blue: high GC).

458 Scale bars: 10Mb Chromosome ideograms reflect centromeres (bright red), gaps (dark red) and
459 Giemsa bands (grays). Recombination rates are low for genes located in AT-rich isochores,
460 except at chromosome ends, which are depleted for maternal recombination. (D) PRDM9 peak
461 enrichment across isochores binned by GC%. (E) Isochore (left) and k-means (right) distribution
462 of genes with the 10% highest rate of within-gene crossovers. (F) Isochore (left) and k-means
463 (right) distribution of genes with the 10% highest strength of within-gene PRDM9 binding.
464

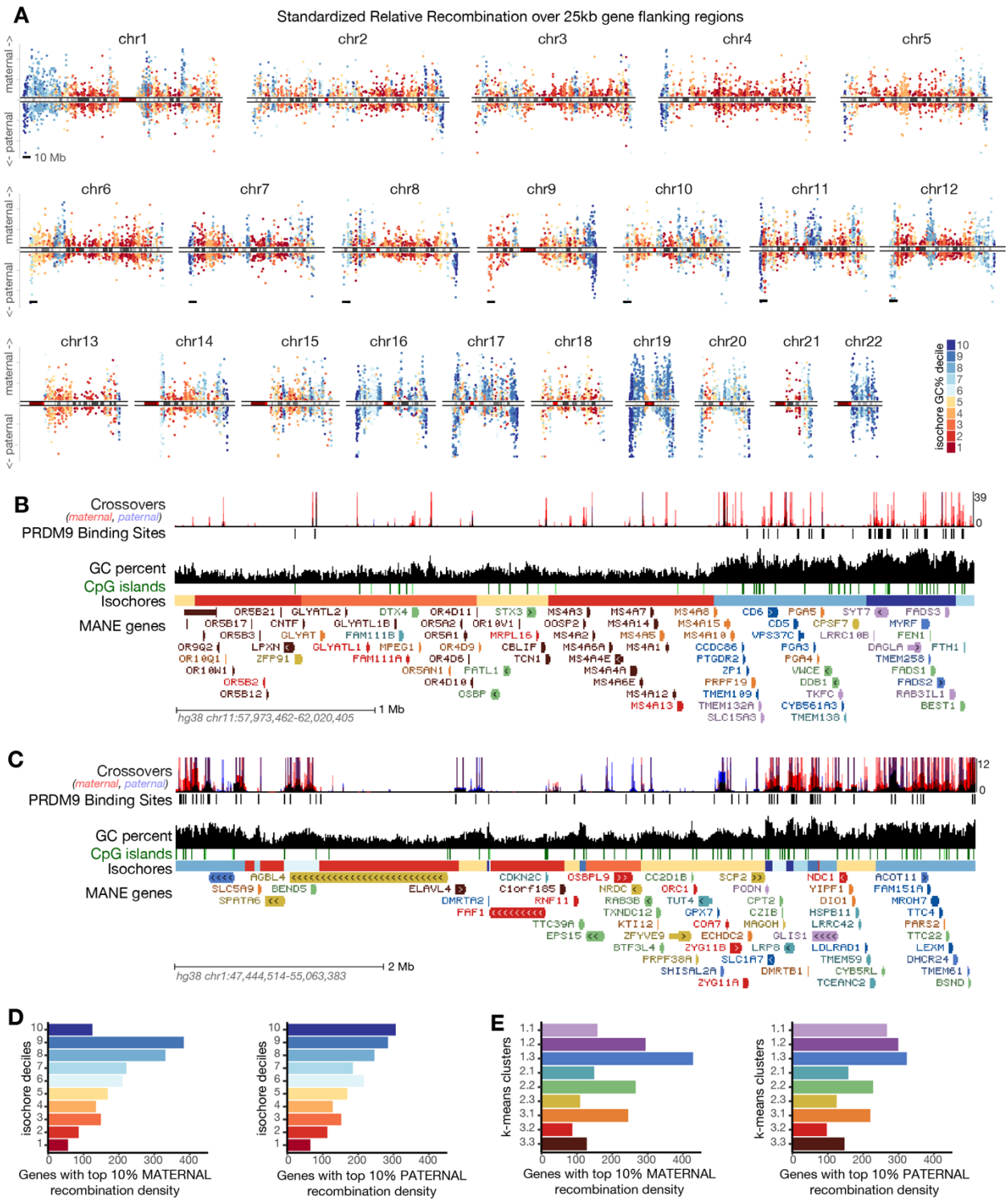


Figure S5: (A) Manhattan plots of maternal and paternal standardized relative recombination rates for all chromosomes, as described in Figure 5C. (B-C) UCSC Genome Browser screenshots of (B) OR and MS4A loci, flanked by GC-rich isochores and (C) interspersed GC and AT-rich isochores. (D-E) Counts of (D) isochore decile and (E) k-means cluster distribution of genes with 10% highest rate of maternal and paternal within-gene crossovers.

Gene expression

To test how AT/GC distribution in genes relates to patterns of gene expression, we used GTEx data, which measures gene expression in 54 tissue types taken from adult human donor cadavers (GTEx Consortium, 2013). We set a threshold (RPKM of 5) to binarize this quantitative data to "expression" or "no expression." In agreement with past reports that AT-biased genes tend to be more "tissue-specific" in their expression, while GC-biased genes tend to be "housekeeping genes," this simple metric varies sharply across genes of different k-means clusters or with differing promoter or home isochore GC content (Figure 6A-C)(Clowney et al., 2011; Holmquist, 1992; Schug et al., 2005). Genes that are GC-rich are often expressed in many or most tissues tested, while AT-biased genes most often appear to be expressed nowhere or in one tissue. We note the many genes with "no" expression in GTEx are specific to tissues not sampled by GTEx (e.g., olfactory receptor genes in the olfactory epithelium). We infer that genes not detected in any tissue in GTEx data—i.e.the preponderance of AT-rich genes—are highly tissue-, cell type-, time-, or condition-dependent in their expression.

CpG dinucleotides are depleted from vertebrate genomes due to the mutability of methylated cytosine; nevertheless, CpGs are relatively enriched in vertebrate promoters, and these "CpG islands" frequently remain unmethylated (Smith and Meissner, 2013). Examining gene blooms in the UCSC browser, we found entire isochores that lacked CpG islands, as calculated by the "CpG Island" track (e.g. see figure 1A, B, Figure S1D)(Micklem and Hillier, 2006). Previous analyses suggested that 50-70% of mammalian genes have CpG island promoters (Deaton and Bird, 2011; Mohn and Schübeler, 2009; Schug et al., 2005). Nevertheless, by considering each gene in its sequence context, we estimate that 90% of protein-coding genes have GC enrichment directly upstream of the TSS (Figure 2A). To test if this GC enrichment reflects CpG islands, we plotted island strength around the TSS across k-means clusters (Figure 6D, E). As predicted by overall patterns of GC content, genes in cluster 3.3 completely lacked CpG islands. As we and others suggested previously in the mouse, CpG-less promoters are likely to be regulated by non-canonical mechanisms that are independent of TATA Binding Protein (TBP)(Clowney et al., 2011; Michaloski et al., 2006). This likely allows the unique and rare expression of these genes relative to their CpG-containing brethren: their ground state is to be "off forever."

In the longest-lived cells in the body, post-mitotic neurons, tandemly arrayed gene families have been shown to be clustered with one another in nuclear space and to be uniquely protected from accumulation of CpH (Cp-nonG) methylation (Lister et al., 2013; Tan et al., 2021). These findings, together with the overwhelming transcriptional repression of these gene families, suggest that they might be sequestered in nuclear space away from the transcriptional machinery. Indeed, we found that across 21 tissues sampled by Hi-C, AT-rich genes and genes located in AT-rich isochores were likely to be located in transcription-suppressing "B" compartments (Figure 6F, G)(Schmitt et al., 2016). Previous analyses have demonstrated that variation in GC content also predicts local chromatin looping and association with the lamina and other nuclear structures (Jabbari and Bernardi, 2017; Jabbari et al., 2019a; Naughton et al., 2013).

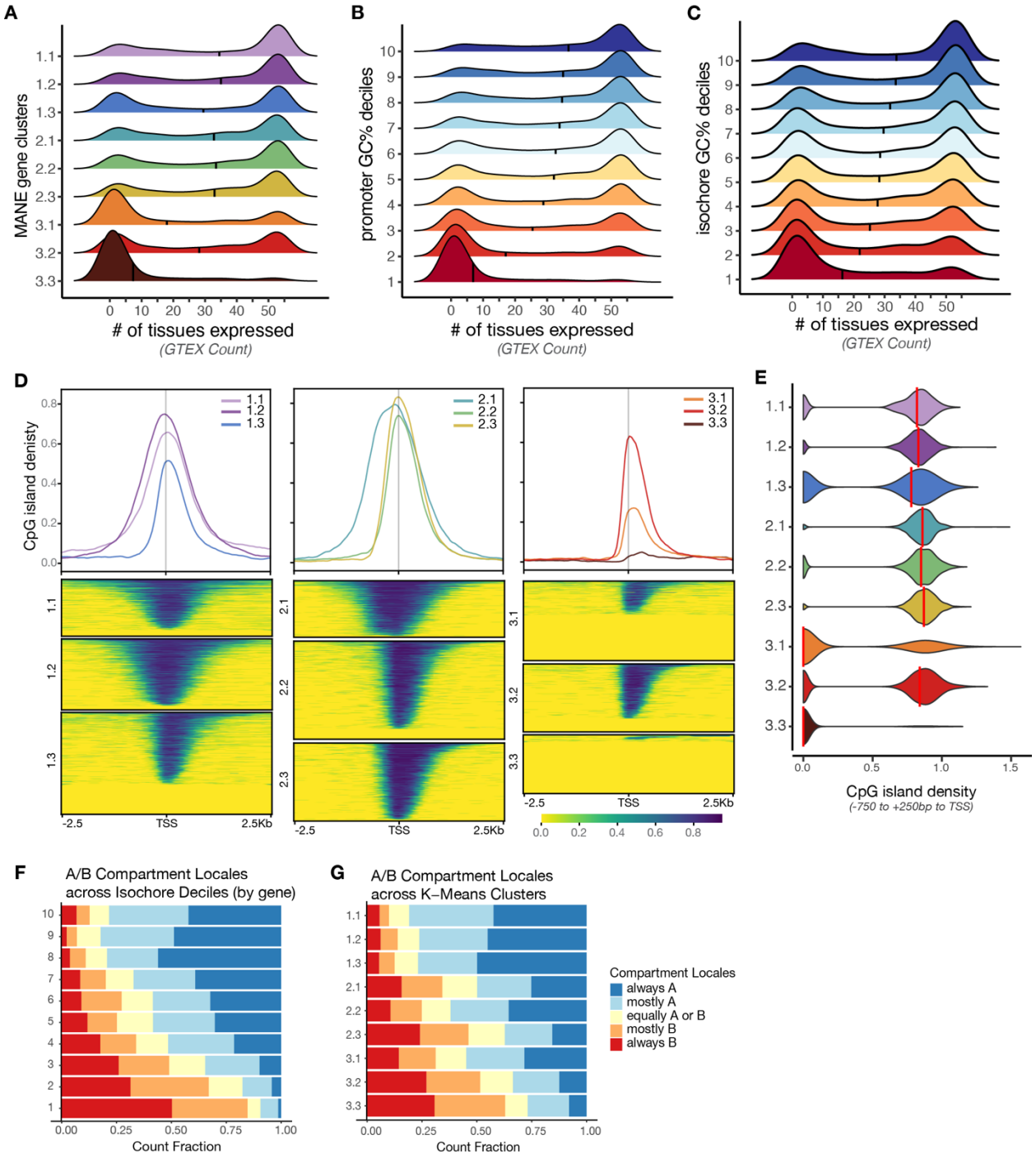
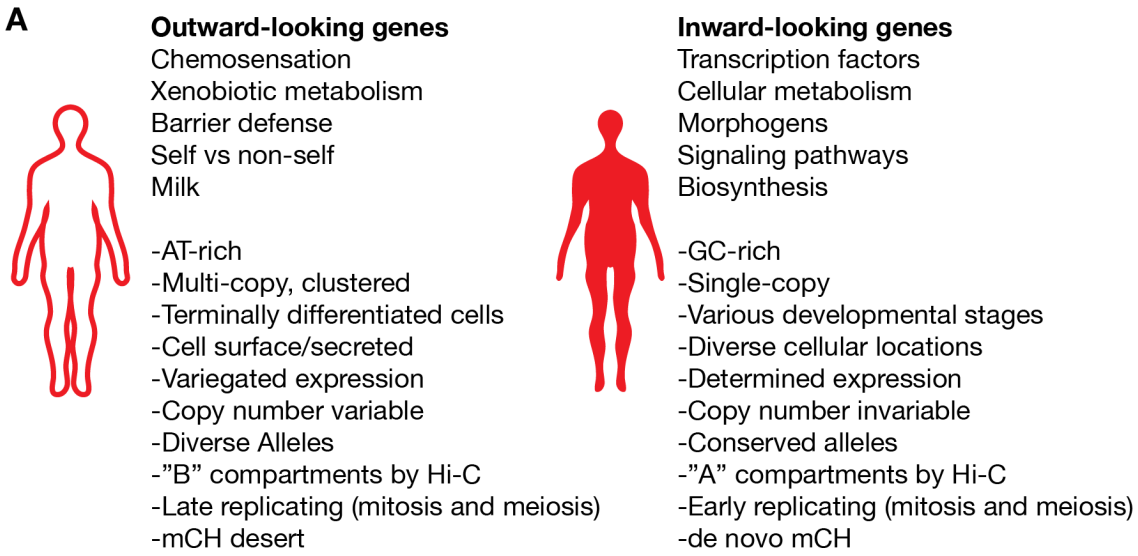


Figure 6: AT-rich genes have restricted expression and lack CG islands. (A-C) Distribution of tissue-level gene expression for genes binned by k-means cluster (A), promoter GC content (B), or home isochore GC content (C). GTEx data from 54 human tissues was binarized to "expression" or "no expression" based on RPKM of 5. AT-rich genes are detected in few or no sampled GTEx tissues, while GC-rich genes are frequently detected in all sampled tissues. Black bars depict median. (D) CpG island density around the TSS (grey line) for genes in each k-means cluster. CpG island calls reflect rate of observed CpG dinucleotides relative to the expected rate for a sequence of that GC content (Micklem and Hillier, 2006). CpG island rates across gene

TSS in each cluster are summarized in line plots, while heatmaps represent island calls around the TSS for each gene. (E) Violin plots showing CpG islands within -750 to +250 bp of TSS for genes in each k-means cluster. Red line depicts median. (F-H) Hi-C compartment assignment across 21 tissues (Schmitt et al., 2016) for genes in each isochore GC% decile (F) and k-means cluster (G). “Always A” and “always B” means the gene was assigned to that compartment in every sampled tissue.

Discussion

Animals make extensive and diverse contacts with the external environment, both engaging with foreign molecules and producing and excreting their own substances. Specialization of these input-output functions plays a definitive role in animal lifestyle, and often occurs in mammals via amplification and diversification of tandemly arrayed gene families (Clowney et al., 2011; Kawasaki et al., 2011; Perry et al., 2007). Extensive gene losses are also common—just as the vomeronasal organ is vestigial in humans, human genes for vomeronasal receptors are no longer functional (Witt and Hummel, 2006; Zhang and Webb, 2003). Here, we expand on our previous work in the mouse to define a common genomic architecture in human for genes whose products engage the external world: these genes are in tandem arrays, are found in AT-biased isochores, and lack CpG islands in their promoters (Figure 7A) (Clowney et al., 2011, 2012). Regions containing AT-skewed gene clusters in mammals are not AT-skewed in mammalian outgroups (Figure 3), suggesting that AT bias emerged as these gene families expanded. Using population genetic data from humans, we test whether elevated rates of allelic diversity in outward-looking genes results from distinct mutational or selective effects. We find that genes in AT-biased tandem arrays experience low ongoing rates of point mutation (Figure 4) and low rates of recombination (Figure 5). We suggest that excessive allelic diversity in these regions is due to weakened selection on historical point mutations, and that low rates of point mutation in the present are due to the scarcity of mutable CpG dinucleotides remaining in these clusters. Together, tolerance of historical point mutation and strengthened intolerance of recombination as gene families expand can explain the high AT content of outward-looking genes in tandem arrays.



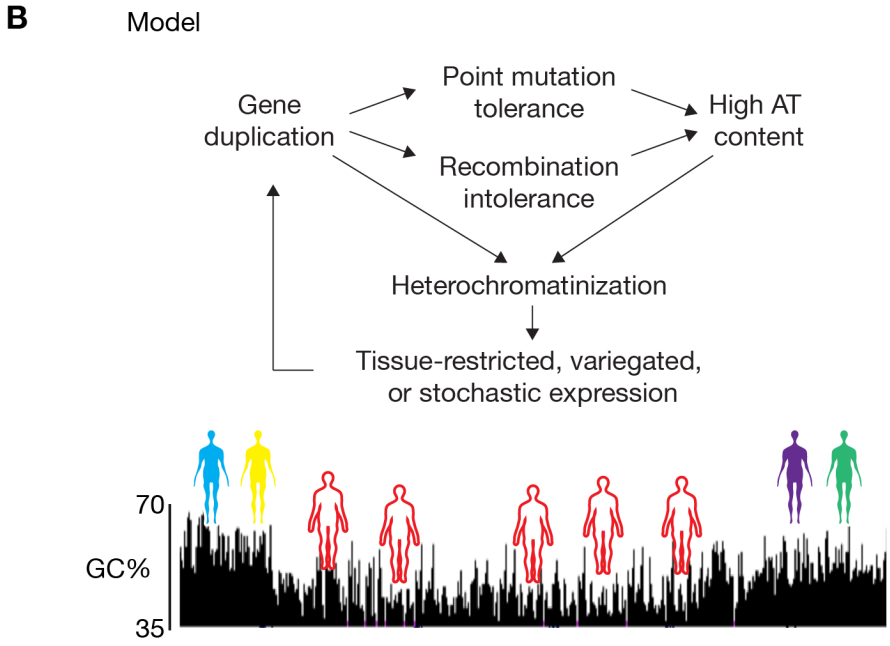


Figure 7: As gene family size grows in cis, selection on point mutations weakens while selection on recombination grows (A) Summary of inward-looking and outward-looking gene families and their genomic distinctions. This table is inspired by (Holmquist and Filipski, 1994). (B) Model of the relationship between gene family expansion, point mutation tolerance, recombination intolerance, and mode of expression.

Implications for gene regulation

Genes located in AT-biased tandem arrays are typically durably silenced almost everywhere in the body and expressed at extremely high levels at a specific place and time. Where and when these genes are expressed, they perform the definitive work of the cell type. Many of these gene families exhibit some kind of exclusive expression, from the fetal to adult switch in hemoglobin β expression (i.e. exclusion over time) to the one-receptor-per-neuron

pattern of olfactory receptor expression (i.e. exclusion over space). We expect that genes located in high-AT isochores that lack CpG islands in their promoters are silenced by particular mechanisms when they are not expressed (i.e. almost everywhere), and that their expression will be activated by non-canonical mechanisms in the single condition where each is expressed. As we have argued previously for ORs, transcription of these genes is likely to be initiated independent of TATA binding protein and likely excludes the recently discovered basal CpG-island-binding factors BANP and BEND3 (Grand et al., 2021; Zhang et al., 2022). One could argue that these genes do not have promoters at all and rely completely on locus control regions to concentrate and deliver transcription factors to the TSS (Monahan et al., 2017, 2019).

Many or most molecular genetic events are sensitive to variation in AT/GC distribution: AT content predicts compartmentalization of the genome in 3D space, replication timing, and patterns of histone marks (Costantini and Musto, 2017; Dekker, 2007; Pratto et al., 2021; Wang and Willard, 2012; Woodfine et al., 2004; Xie et al., 2017). Typically, AT-biased sequence is packaged as heterochromatin and silenced. Work on olfactory receptors, clustered protocadherins, and secreted liver proteins suggest that these gene families are expressed from the context of constitutive heterochromatin, which appears to be present prior to expression and to be retained on family members that are not expressed (Balan et al., 2021; Magklara et al., 2011; Nicetto et al., 2019; Toyoda et al., 2014; Williams et al., 2021). CpG islands also function as molecular beacons: they mark transcription start sites, serve as recombination hotspots in the absence of PRDM9, and act as replication origins in meiosis (Antequera and Bird, 1999; Baker et al., 2017; Pratto et al., 2021). The accrual of high AT content in gene arrays and the lack of CpG islands is therefore likely to exert a strong effect on the molecular regulation of these genes. In ectodermal development, single-copy genes accrue CpH methylation, perhaps passively, while AT-rich gene arrays remain devoid of this modification (Lister et al., 2013). This suggests that AT-rich gene arrays are locked away from the ambient molecular stew of the nucleus, perhaps over very long developmental time periods. Indeed, tandem arrays cluster together in the nucleus in post-mitotic neurons (Tan et al., 2021).

In addition to the extreme time and/or tissue specificity of most outward-looking gene families, a fraction of these families exhibit stochastic expression such that each cell expresses just one or a sparse subset of family members. Chemosensors, B- and T- cell receptors, and clustered protocadherins all exhibit this restricted expression (Williams et al., 2021). As we argued recently, sparse cell-wise expression patterns compartmentalize the effects of mutations (Williams et al., 2021). These mechanisms are also likely to result in insensitivity to copy number variation, as each cell chooses its own dose of family members for expression. Feedback mechanisms that ensure cells can "choose again" if they originally pick a pseudogene further buffer potential deleterious effects of mutations (Dalton et al., 2013; Hetz et al., 2020).

Role for recombination in diversification versus maintenance of gene arrays

A canonical rule in evolutionary biology holds that recombination increases allelic variation (Begun and Aquadro, 1992). While this relationship likely holds among single-copy genes, our analysis of multicopy genes stands in contrast to this trend: AT-biased gene families exhibit extraordinary allelic diversity despite low recombination. We argue that this lack of recombination is both historical and ongoing—if indeed GC content is a record of past gene conversion (high-GC regions are thought to have experienced high historical recombination) (Pouyet et al., 2017), then AT-biased arrays would have arisen due to historical depletion of recombination in these regions. In the present, lower-GC regions experience lower rates of

crossovers as well, as shown in Figure 5. There remains conflict between the mode by which these gene arrays are thought to have bloomed—i.e. via gene duplication through ectopic exchange during recombination—and their current depletion for recombination events. Other modes of duplication, including replication slippage and transposition, may also be at work in expanding these arrays.

Ectopic exchange in repetitive gene regions can have benign or catastrophic consequences. Induction of copy number variation within a gene array may be of small phenotypic consequence, as the jobs these genes perform are by nature distributed across many family members. In contrast, ectopic exchange that deletes a cluster or induces recombination between clusters can result in catastrophic chromosome rearrangements. Indeed, mammalian chromosome evolution appears to have been shaped by ectopic exchange between OR clusters (Kim et al., 2017; Linardopoulou et al., 2005; Mefford et al., 2001; Newman and Trask, 2003; Rouquier et al., 1998; Trask et al., 1998; Yue and Haaf, 2006). Finally, even if structural variation in an outward-looking tandem array is benign within an individual, it can lead to hybrid incompatibility and can initiate or reinforce reproductive isolation that leads to speciation (North et al., 2020; Paudel et al., 2015; Rogers, 2015). Recent modeling work has sought to characterize the tradeoffs between the structural fragility of gene blooms and the potential positive effects of allelic diversification (Otto et al., 2022).

Given the genomic danger of these tandem arrays, why have gene family members remained *in cis* with one another? An extreme example is the "milk and teeth" locus on human chromosome 4. The casein genes in this locus evolved via tandem duplication of enamel genes at the root of the mammalian tree; the enamel genes themselves evolved from *follicular dendritic cell secreted protein* in bony fish (Kawasaki et al., 2011; Qu et al., 2015). Astonishingly, all these genes have remained syntenic. Why on earth would this be the case, given that they're expressed in three separate body systems and that such tandem arrays are genomically dangerous? We propose that as in the case of maintenance of hox gene synteny, the regulatory elements of these genes remain tangled with one another, such that relocation of array members elsewhere in the genome would divorce them from cis-regulatory elements that they depend on for expression (Darbellay et al., 2019; Mann, 1997; Montavon et al., 2011). Recent research on enhancer evolution in animals suggests that enhancer tangling can result in the preservation of synteny over ~700 million years (Wong et al., 2020). In other cases, as in B Cell Receptor, hemoglobin, clustered protocadherin, interferon, and chemosensor arrays, family members share and compete for the same regulatory elements (Li et al., 2002; Markenscoff-Papadimitriou et al., 2014; Ribich et al., 2006; Roy et al., 2011; Yokota et al., 2011). This mutual dependence would again increase the phenotypic consequences of recombination events that break synteny.

Array incompatibility between individuals of a species and the necessity of remaining co-located with regulatory elements that may be tangled with or shared by other gene family members would lead tandem arrays to behave like supergenes—multigene regions inherited as an allelic unit. We suspect that depletion of CpG islands and PRDM9 sites from tandemly arrayed genes protects the genome from the danger of errantly recombining these duplicative regions. Nevertheless, recombination and gene duplication or deletion still sometimes occur in these regions—their crossover rate even today is non-zero—and the marginal fitness effects of resulting copy number variants allow products of these meioses to be preserved in the population. As gene arrays get larger, point mutation tolerance shifts their GC content downward, putting the brakes on recombination as they become ever more unwieldy. Overall

recombination in these regions is therefore suppressed, while differential tolerance of local duplications versus gross rearrangements could allow an increase in local allelic diversity.

Implications for chromosome organization

Repetitive elements have shaped chromosomal evolution since the dawn of eukaryotes. The linear genome is proposed to have arisen from erroneous meiotic recombination between Group II introns which invaded the circular genome to create the t-loop precursors to stable telomeres (de Lange, 2015). Similarly, dispersion and expansion of ORs and other large tandem gene arrays have shaped mammalian chromosome evolution. Tandem arrays of ORs represent ancestral breakpoints of chromosomal synteny between mice, rats, and humans (Yue and Haaf, 2006; Zody et al., 2006). A large OR cluster is found at the end of the q-arm of chr1 in humans but not in mice. In addition to ORs, large gene families including zinc finger (*ZNF*) and immunoglobulin heavy chain (*IGH*) genes are observed at chromosome ends across eukaryotes (Riethman et al., 2004). In the modern human population, unequal crossovers between OR clusters are a source of recurrent and pathological rearrangement hotspots (Giglio et al., 2001).

While we also observe these AT-rich isochores at chromosome ends, we predominately find isochores at chromosome ends to be the most GC-rich across the genome with high gene diversity (i.e. many single-copy genes)(Jensen-Seaman et al., 2004). Indeed, the largest OR cluster at the end of the q-arm of chromosome 1 in humans is followed by a higher GC% isochore containing ZNF genes. This strong end-GC% accumulation arises paternally: genes in high GC% isochores at chromosome ends are enriched for paternal crossovers and relatively depleted of maternal crossovers. Overall, paternal crossovers are biased towards chromosome ends (Hultén, 1974; Lee et al., 2011). Chromatin organization of pachytene spermatocytes is implicated in this phenomenon, including synaptonemal complex length and lack of PRDM9 requirement for crossovers in subtelomeric regions, however, the precise mechanism underlying it is unknown (Pratto et al., 2014; Tease and Hultén, 2004). Potentially, recombination-based alternative lengthening of telomeres (ALT) in spermatocytes biases hotspots towards chromosome ends (Antunes et al., 2015).

Over evolutionary time, as ectopic recombination places high AT% tandem arrays at chromosome ends, high paternal rates of gBGC at the ends of chromosomes would generate new isochores of increasing GC% and comprising newly evolving genes (Capra et al., 2013; Huttener et al., 2019).

Is mutation biased or random with respect to gene function?

Recent mutation accumulation studies have suggested that *de novo* mutations could occur with different frequencies in different kinds of genes or in genic versus non-genic locations (Monroe et al., 2022). We and others also argued previously that segregation of mutation-tolerant versus mutation-intolerant genes into AT- versus GC-biased regions of the genome could allow differential mutation rates on different classes of genes (Chuang and Li, 2004; Clowney et al., 2011; Grimwood et al., 2004). However, our analyses of synonymous versus functional variant rate and of *de novo* mutation rate in AT- versus GC-biased genes suggest the opposite: that AT-biased genes experience fewer mutations in living humans than do GC-biased genes. As loss-of-function-intolerant genes tend to be GC-rich, depletion of mutations with strong fitness effects from the pool of living humans whose genomes have been sequenced would only weaken the trend towards higher mutation rates in GC-biased genes.

While active mutagenic processes specific to gene blooms remain possible, overall mutation rates are higher in GC-rich sequences. Therefore, we expect that differential AT/GC content in inward- versus outward-looking genes in the present is the result of differential selection trajectory over evolutionary time. We conclude that AT-biased genes have attained that high AT content mostly due to drift, while purifying selection in GC-biased genes has combined with higher recombination rates to help to preserve their high GC content. Evolutionary depletion of GC bases from outward-looking genes lowers present *de novo* mutation rates due to lack of remaining mutable cytosines.

Is this genomic architecture specific to mammals?

While isochore structure is not unique to mammals, it is not a universal feature across animal clades, and the AT/GC variation observed in mammals is extreme (Lynch, Michael, 2007). We are curious whether stem mammals evolved molecular mechanisms that facilitated the evolution of gene arrays. These could include both systems that maintain these arrays as constitutive heterochromatin when they are not being expressed and unique transcriptional mechanisms that activate them, often in a stochastic or highly restricted manner, in their target tissues. One candidate factor that mediates long-range enhancer-promoter interactions in multiple arrayed families is Ldb1 (Monahan et al., 2019; Schoenfelder and Fraser, 2019). Social insects have also massively expanded their olfactory receptor gene repertoire in cis; in the ant, this is accompanied by increased AT content (McKenzie et al., 2016). Have convergent mechanisms for stochastic expression facilitated olfactory receptor repertoire expansion in insects?

Other clades may have evolved distinct mechanisms to organize repetitive genes or gene pieces: In *Diptera*, repetitive arrays are often organized as alternative splicing hubs (Armitage et al., 2012; Goeke et al., 2003; Labrador and Corces, 2003; Venables et al., 2012). Reptiles and birds exhibit "microchromosomes" which have distinct GC content from the rest of the genome and can house arrays of rapidly evolving, outward-looking genes such as venoms (Schield et al., 2019). Trypanosome arrays of surface VSGs are located in subtelomeric regions (Berriman et al., 2005). For mammals, the "isochore solution" balances diversity in gene arrays with genomic integrity.

Methods

Describing isochores

To call isochores, we implemented a genomic segmentation algorithm called GC-Profile (Gao and Zhang, 2006) using halting parameter (number of segmentation iterations) of t_0 275 and minimum segment length of 3000 bp. Gaps less than 1% of the input sequence were filtered out, generating 4328 distinct isochores in hg38 (Supplemental Table 1). Isochores were ranked by average GC%, with rank 1 having the highest and 4328 having the lowest. We also performed this analysis in hg19 (Supplemental Table 2). Isochores are reported in Supplemental Tables 1-2.

Statistical analyses

We performed Kruskal-Wallis non-parametric ANOVA for each group of comparisons (Supplemental Table 3). We then used Dunn's pairwise analysis to compare individual groups with one another (Supplemental Tables 4 and 5).

GC content calculations

Genes from the Matched Annotation dataset from the NCBI and EMBL-EBI (MANE) Select dataset (Morales et al., 2022) were downloaded from the UCSC Genome Browser. Isochores were matched to genes using the coordinates of the transcription start site. GC content across gene features, including promoters (-750 to +250bp flanking TSS), flanking regions (+/- 25kb), coding exons, exons and UTRs, and introns were separately calculated from FASTA sequences using bedTools (Quinlan and Hall, 2010).

To generate 9 k-means clustesr, we used gc5BaseBw from the UCSC Genome Browser (Clawson, 2018) to calculate GC% scores across MANE genes with +/- 1kb flanks. We generated 3-kmeans clusters of genes, which were further clustered into 3-kmeans clusters each using deepTools plotHeatmap (Ramírez et al., 2016). Cluster assignment and quantification of other parameters for each gene are reported in Supplemental Table 6.

Characterizing types of genes

To characterize the types of genes residing in isochores of varying GC, we used 2 categories of descriptors: GO terms and gene prefixes. To identify GO terms associated with genes in each isochore GC decile, we used the R package, clusterProfiler (version 4.2.2)(Wu et al., 2021). This helped us streamline identification of key terms that appeared in each decile. With this list, we identified GO terms that were most significantly enriched in each decile with a depth of at least 30 genes. Using AmiGO (Carbon et al., 2009), an online database of GO identifiers, we pulled the list of genes associated with our selected group of significant GO terms and plotted GC content across each term. The terms we chose are listed in the table below.

Shortened Term (from Fig 2)	Full GO Term Description	GO ID
wnt signaling	wnt signaling pathway	GO:0016055
kinase activity	kinase activity	GO:0016301
transcription	transcription, DNA-templated	GO:0006351
defense	immune response	GO:0006955
xenobiosis	xenobiotic metabolic process	GO:0006805
keratinization	keratinization	GO:0031424
chemosensation	detection of a chemical stimulus	GO:0009593

We wanted an alternative to GO analysis for assessing diversity across isochores and k-means gene clusters. Since the prefixes of well-annotated genes (like the ones from the MANE dataset) are shared across genes within the same gene family, we used this as a means of assessing diversity with more specificity than one would achieve through GO analysis. The process of assigning gene prefixes is as follows:

1. Convert old names into new nomenclature.
    - Go to the HUGO Gene Nomenclature Committee's (HGNC)(Tweedie et al., 2021) website and the list of gene symbols from the MANE set into their "Multi-symbol checker" (the link provided will take you there directly). This will ensure

780 we have the most up-to-date names for each of our genes (ie: some which may
781 have been labeled as 'FAM' may have a new symbol to go with the rest of the
782 gene family).
783 • Match names in the MANE set to names labeled "Approved symbols" by HGNC,
784 and replace those symbols with the HUGO names.
785 2. Replace numbers with "_". We can't remove all numbers because there are several genes
786 that have more letters after numbers that aren't important for our purposes (ie: CSN2A
787 will become CSN_A).
788 3. Remove anything after the first instance of "_" (ie: CSN_A will become CSN). The goal
789 of this step is to keep the first part of the prefix, removing letters and numbers that
790 indicate subfamilies.
791 4. While it isn't common, some genes require us to know those numbers to know what they
792 do (most commonly, enzymes involved in modifying carbohydrates). Largely, these
793 genes start with a single letter, followed by numbers, then more letters. Thus, to fix these
794 genes, we pull out the genes that have 1 letter after steps 4 and 5.
795 5. Look through each of those genes that start with only one letter, then decide how best to
796 group them.
797 6. View gene prefixes in alphabetical order and search for prefixes that are likely to be
798 families, then rename (ie: KCNT and KCNQ are both potassium channels, so we grouped
799 these together).
800 Once we had a list of gene prefixes, we calculated a Shannon's H diversity metric for each
801 isochore based on the prefix probabilities in each isochore (proportions + log2(1/proportions) =
802 diversity metric). Larger values are indicative of more diversity. Similarly, we calculated a
803 Shannon's H diversity metric for each k-means cluster.
804
805 De novo *mutations*
806 *De novo* mutations (DNMs) were compiled by (Rodriguez-Galindo et al., 2020) from
807 seven family-based whole genome sequencing (WGS) datasets, encompassing a total of 679,547
808 single nucleotide variants (SNVs), which comprise data from both neurotypical and
809 neurodivergent individuals. We remapped the dataset to hg38 using LiftOver in UCSC Genome
810 Browser. To calculate genomic DNM density, we counted the number of DNMs occurring
811 within the coordinates listed in the GC calculation section above. To calculate DNM density, we
812 pooled genic DNMs within each isochore and divided by the sum of the region of interest's size,
813 i.e. we identified all the genes in an isochore, summed the DNMs between their transcription
814 start and termination sites, then divided by the summed length of those genic regions.
815
816 *Allelic variants*
817 We used the gnomAD v2.1.1 dataset of single nucleotide allelic variants (Karczewski et
818 al., 2020). The authors defined rare single nucleotide variants (<0.1% allele frequency) from
819 125,748 exomes and 15,708 whole genomes and predicted whether variants within coding
820 regions are likely to be functionally synonymous, missense, or loss-of-function. Here, we used
821 observed synonymous, missense, and loss-of-function mutation rates. We ported variant calls to
822 MANE genes in hg38 using the gene symbol and Ensembl transcript IDs. In Figure 5, we also
823 use the calculated pLI score from gnomAD, which describes the likelihood that a gene is loss-of-
824 function intolerant in humans.
825

Recombination

Crossover data for hg38 was acquired from deCODE where the authors used whole-genome sequence (WGS) of trios and were able to refine crossover boundaries for 247,942 crossovers in 9423 paternal meioses and 514,039 crossovers in 11,750 maternal meioses (Halldorsson et al., 2019). Of note, the data we used here is restricted to autosomes. To calculate crossover density, we assigned crossovers to a region of interest based on the median of the crossover coordinates. We normalized counts within a region by dividing by the genomic average for that sex. PRDM9 binding data from HEK293T cells transfected with the PRDM9 reference allele was acquired from (Altemose et al., 2017). We selected the top 10% of PRDM9 peaks based on enrichment scores to account for weak PRDM9 binding sites associated with overexpression in the system, as noted by the authors. Like with crossovers, we calculated PRDM9 binding site density across genes as the summed enrichment scores across genic regions within an isochore, mapping by the midpoint of the binding coordinates.

Gene Regulatory Information

To determine tissue specificity of gene expression, RNA-sequencing data was sourced from Genotype-Tissue Expression (GTEx) project (V8, released in August 2019), containing 17,382 samples collected from 54 tissues from 948 donors (GTEx Consortium, 2013). For each gene in the MANE set, we counted the number of tissues in which expression was at least 5 transcripts per million (TPM).

To measure A/B compartment occupancy of genes across tissues, AB compartments were sourced from published Hi-C data from 21 tissues and cell types (Schmitt et al., 2016). MANE genes were lifted over into hg19 to match A/B compartment domain calls in hg19. Isochores called in hg19 were assigned to a compartment by matching the isochore's midpoint to the midpoint of the closest compartment. Genes were assigned to a compartment by matching the transcription start site to the midpoint of the nearest compartment, as most genes did not fall into a single compartment (~90%). Further, we counted the occurrences of compartments A and B for each isochore and gene. These counts were binned into always A (21 counts of A), mostly A (14-20 counts of A or 0-6 counts of B), equally A or B (7-13 counts of A or B), mostly B (0-6 counts of A or 14-20 counts of B), and always B (21 counts of B).

To identify genes with CpG islands in promoter regions, we downloaded the CpG Island track (unmasked) from the UCSC Genome Browser (Micklem and Hillier, 2006). The ratio of observed vs. expected CpG dinucleotides was converted to a bigwig coverage file and plotted across gene TSS's (+/-2.5kb) in 9 k-means clusters using deepTools plotHeatmap. The average score of CpG islands within -750bp and +250 bp of a gene TSS were calculated using bedTools.

*OR expansion around the hemoglobin β (*HBB*) locus*

To measure olfactory receptor expansions flanking the hemoglobin β (*HBB*) locus, olfactory receptors across representative species of reptiles, monotremes, and mammals were counted using the NCBI Genome Data Viewer. We hand-counted ORs, OR-like genes, and OR pseudogenes between genes *STIM1* and *RRM1* and *FHIP1B*, *CNGA4*, *CCKBR*, and *CAVIN3* to mark the ends of the tandem OR array surrounding HBB.

Common Name	Assembly	Coordinates
Human	hg38	Chr11: 4,138,933 - 6,211,337

Rhesus macaque	Mmul_10	Chr14: 59,743,017 - 64,335,040
Marmoset	Callithrix_jacchus_cj1700_1.1	Chr11: 67,513,744 - 69,794,829
Mouse lemur	Mmur3.0	Chr5: 53,853,980 - 56,438,887
Rat	mRatBN7.2	Chr1: 156,848,263 - 159,718,678
Guinea pig	Cavpor3.0	Un|NT_176316.1: 3,962,532 - 4,637,412 Un|NT_176348.1: 3,258,992 - 8,982,342
Rabbit	OryCun2.0	Chr1: 145,054,257 - 147,211,267
Pig	Sscrofa11.1	Chr9: 3,464,238 - 6,026,150
Bottlenose dolphin	mTurTru.mat.Y	Chr8: 59,408,982 - 60,032,380
Alpaca	VicPac3.1	10|NW_021964172.1: 23,413,706 - 24,984,720
Horse	EquCab3.0	10|NW_021964172.1: 23,413,706 - 24,984,720
Dog	ROS_Cfam_1.0	Chr21: 26,840,937 - 30,373,644
Sea otter	ASM2288905.1	Un|NW_019154152.1: 7,935,810 - 10,562,113
Hawaiian monk seal	ASM220157v2	Chr11: 49,856,742 - 51,641,650
Bat	Pvam_2.0	NW_011889092.1: 2,496,882 - 2,724,376 Un|NW_011889212.1: 1 - 760,764 Un|NW_011889285.1: 1 - 679,978 Un|NW_011889452.1: 390,672 - 514,816
Elephant	Loxafr3.0	Un|NW_003573499.1: 8,339,565 - 8,685,383 Un|NW_003573536.1: 1 - 3,251,147 Un|NW_003573441.1: 44,869,48 - 45,000,610
Tenrec	ASM31398v2	Un|NW_022105611.1: 22,229,370 - 22,436,814 Un|NW_022103939.1: 675,963 - 3,612,721
Tasmanian devil	mSarHar1.11	Chr3: 531,712,487 - 536,448,100
Opossum	monDom5	Chr4: 349,070,556 - 354,286,059
Platypus	mOrnAna1.pri.v4	Chr2: 138,334,962 - 139,648,670
Chicken	GCF_16699485.2	Chr1: 195,900,384 - 196,103,623
Zebra finch	bTaeGut1.4.pri	Chr1: 113,563,080 - 113,836,654 Un|NW_003573499.1: 8,299,414 - 8,339,565
Painted turtle	Chrysemys_picta_BioNan-3.0.4	NW_024919015.1: 1 - 1,282,942
Garter snake	rTHAEle1.pri	Chr6: 317,034 - 521,038

868
869

Acknowledgements

Thanks to Rachel Duffié, David Lyons, Matthew Holding and members of the Clowney lab for discussion and comments on the manuscript. This work was supported by the Rita Allen Foundation Milton Cassel Scholarship and the Alfred P Sloan Research Scholarship in Neuroscience to EJC. EJC is a McKnight Scholar and a Pew Biomedical Scholar. MVB was supported by the NIH Cellular and Molecular Biology Training Grant T32-GM007315. MAC was supported by the NIH Early Stage Training in the Neurosciences Training Grant T32-NS076401.

References

Altemose, N., Noor, N., Bitoun, E., Tumian, A., Imbeault, M., Chapman, J.R., Aricescu, A.R., and Myers, S.R. (2017). A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis. ELife *6*, e28383. https://doi.org/10.7554/eLife.28383.

Antequera, F., and Bird, A. (1999). CpG islands as genomic footprints of promoters that are associated with replication origins. Current Biology *9*, R661–R667. https://doi.org/10.1016/S0960-9822(99)80418-7.

Antunes, D.M.F., Kalmbach, K.H., Wang, F., Dracxler, R.C., Seth-Smith, M.L., Kramer, Y., Buldo-Licciardi, J., Kohlrausch, F.B., and Keefe, D.L. (2015). A single-cell assay for telomere DNA content shows increasing telomere length heterogeneity, as well as increasing mean telomere length in human spermatozoa with advancing age. J Assist Reprod Genet *32*, 1685–1690. https://doi.org/10.1007/s10815-015-0574-3.

Armelin-Correa, L.M., Gutiyama, L.M., Brandt, D.Y.C., and Malnic, B. (2014). Nuclear compartmentalization of odorant receptor genes. Proc Natl Acad Sci U S A *111*, 2782–2787. https://doi.org/10.1073/pnas.1317036111.

Armitage, S.A.O., Freiburg, R.Y., Kurtz, J., and Bravo, I.G. (2012). The evolution of Dscam genes across the arthropods. BMC Evol Biol *12*, 53. https://doi.org/10.1186/1471-2148-12-53.

Baker, Z., Schumer, M., Haba, Y., Bashkirova, L., Holland, C., Rosenthal, G.G., and Przeworski, M. (2017). Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. ELife *6*, e24133. https://doi.org/10.7554/eLife.24133.

Balan, S., Iwayama, Y., Ohnishi, T., Fukuda, M., Shirai, A., Yamada, A., Weirich, S., Schuhmacher, M.K., Dileep, K.V., Endo, T., et al. (2021). A loss-of-function variant in SUV39H2 identified in autism-spectrum disorder causes altered H3K9 trimethylation and dysregulation of protocadherin β-cluster genes in the developing brain. Mol Psychiatry *26*, 7550–7559. https://doi.org/10.1038/s41380-021-01199-7.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science *327*, 836–840. https://doi.org/10.1126/science.1183439.

908 Begun, D.J., and Aquadro, C.F. (1992). Levels of naturally occurring DNA polymorphism
909 correlate with recombination rates in D. melanogaster. Nature *356*, 519–520.
910 https://doi.org/10.1038/356519a0.

911 Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M.,
912 and Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. Science *228*, 953–958.
913 https://doi.org/10.1126/science.4001930.

914 Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C.,
915 Lennard, N.J., Caler, E., Hamlin, N.E., Haas, B., et al. (2005). The genome of the African
916 trypanosome Trypanosoma brucei. Science *309*, 416–422.
917 https://doi.org/10.1126/science.1112642.

918 Bickmore, W.A. (2019). Patterns in the genome. Heredity *123*, 50–57.
919 https://doi.org/10.1038/s41437-019-0220-4.

920 Capra, J.A., Hubisz, M.J., Kostka, D., Pollard, K.S., and Siepel, A. (2013). A model-based
921 analysis of GC-biased gene conversion in the human and chimpanzee genomes. PLoS Genet *9*,
922 e1003684. https://doi.org/10.1371/journal.pgen.1003684.

923 Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., Lewis, S., AmiGO Hub, and Web
924 Presence Working Group (2009). AmiGO: online access to ontology and annotation data.
925 Bioinformatics *25*, 288–289. https://doi.org/10.1093/bioinformatics/btn615.

926 Casola, C., and Betrán, E. (2017). The Genomic Impact of Gene Retrocopies: What Have We
927 Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses?
928 Genome Biology and Evolution *9*, 1351–1373. https://doi.org/10.1093/gbe/evx081.

929 Charkoftaki, G., Wang, Y., McAndrews, M., Bruford, E.A., Thompson, D.C., Vasiliou, V., and
930 Nebert, D.W. (2019). Update on the human and mouse lipocalin (LCN) gene family, including
931 evidence the mouse Mup cluster is result of an "evolutionary bloom." Human Genomics *13*, 11.
932 https://doi.org/10.1186/s40246-019-0191-9.

933 Chuang, J.H., and Li, H. (2004). Functional bias and spatial organization of genes in mutational
934 hot and cold regions in the human genome. PLoS Biol *2*, E29.
935 https://doi.org/10.1371/journal.pbio.0020029.

936 Clawson, H. (2018). GC Percent in 5-Base Windows (gc5BaseBw). Unpublished.
937 https://genome.ucsc.edu/cgi-
938 bin/hgc?hgsid=950293831_VmX9SYAwpTKoTme1yOSQX58aR0aR&c=chr16&l=48043557&
939 r=48045592&o=48043557&t=48045592&g=gc5BaseBw&i=gc5BaseBw. Accessed 6/21/22.

940 Clowney, E.J., Magklara, A., Colquitt, B.M., Pathak, N., Lane, R.P., and Lomvardas, S. (2011).
941 High-throughput mapping of the promoters of the mouse olfactory receptor genes reveals a new
942 type of mammalian promoter and provides insight into olfactory receptor gene regulation.
943 Genome Res. *21*, 1249–1259. https://doi.org/10.1101/gr.120162.110.

944 Clowney, E.J., LeGros, M.A., Mosley, C.P., Clowney, F.G., Markenskoff-Papadimitriou, E.C.,
945 Myllys, M., Barnea, G., Larabell, C.A., and Lomvardas, S. (2012). Nuclear aggregation of
946 olfactory receptor genes governs their monogenic expression. Cell *151*, 724–737.
947 https://doi.org/10.1016/j.cell.2012.09.043.

948 Cohen, N., Dagan, T., Stone, L., and Graur, D. (2005). GC Composition of the Human Genome:
949 In Search of Isochores. Molecular Biology and Evolution *22*, 1260–1272.
950 https://doi.org/10.1093/molbev/msi115.

951 Corneo, G., Ginelli, E., Soave, C., and Bernardi, G. (1968). Isolation and characterization of
952 mouse and guinea pig satellite deoxyribonucleic acids. Biochemistry *7*, 4373–4379.
953 https://doi.org/10.1021/bi00852a033.

954 Costantini, M., and Musto, H. (2017). The Isochores as a Fundamental Level of Genome
955 Structure and Organization: A General Overview. J Mol Evol *84*, 93–103.
956 https://doi.org/10.1007/s00239-017-9785-9.

957 Costantini, M., Clay, O., Auletta, F., and Bernardi, G. (2006). An isochore map of human
958 chromosomes. Genome Res. *16*, 536–541. https://doi.org/10.1101/gr.4910606.

959 Dalton, R.P., Lyons, D.B., and Lomvardas, S. (2013). Co-opting the unfolded protein response to
960 elicit olfactory receptor feedback. Cell *155*, 321–332. https://doi.org/10.1016/j.cell.2013.09.033.

961 Darbellay, F., Bochaton, C., Lopez-Delisle, L., Mascrez, B., Tschopp, P., Delpretti, S., Zakany,
962 J., and Duboule, D. (2019). The constrained architecture of mammalian Hox gene clusters.
963 Proceedings of the National Academy of Sciences *116*, 13424–13433.
964 https://doi.org/10.1073/pnas.1904602116.

965 Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev
966 *25*, 1010–1022. https://doi.org/10.1101/gad.2037511.

967 Dekker, J. (2007). GC- and AT-rich chromatin domains differ in conformation and histone
968 modification status and are differentially modulated by Rpd3p. Genome Biology *8*, R116.
969 https://doi.org/10.1186/gb-2007-8-6-r116.

970 Demuth, J.P., De Bie, T., Stajich, J.E., Cristianini, N., and Hahn, M.W. (2006). The evolution of
971 mammalian gene families. PLoS One *1*, e85. https://doi.org/10.1371/journal.pone.0000085.

972 Duret, L., and Arndt, P.F. (2008). The Impact of Recombination on Nucleotide Substitutions in
973 the Human Genome. PLOS Genetics *4*, e1000071. https://doi.org/10.1371/journal.pgen.1000071.

974 Duret, L., and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian
975 genomic landscapes. Annu Rev Genomics Hum Genet *10*, 285–311.
976 https://doi.org/10.1146/annurev-genom-082908-150001.

977 Feyereisen, R. (2006). Evolution of insect P450. Biochem Soc Trans *34*, 1252–1255.
978 https://doi.org/10.1042/BST0341252.

Filipski, J. (1990). Evolution of DNA Sequence Contributions of Mutational Bias and Selection to the Origin of Chromosomal Compartments. In Advances in Mutagenesis Research, G. Obe, ed. (Berlin, Heidelberg: Springer), pp. 1–54.

Francioli, L.C., Polak, P.P., Koren, A., Menelaou, A., Chun, S., Renkens, I., Consortium, G. of the N., Duijn, C.M. van, Swertz, M., Wijmenga, C., et al. (2015). Genome-wide patterns and properties of *de novo* mutations in humans. Nature Genetics *47*, 822–826. https://doi.org/10.1038/ng.3292.

Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., et al. (2006). Copy number variation: new insights in genome diversity. Genome Res *16*, 949–961. https://doi.org/10.1101/gr.3677206.

Fryxell, K.J., and Zuckerkandl, E. (2000). Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol Biol Evol *17*, 1371–1383. https://doi.org/10.1093/oxfordjournals.molbev.a026420.

Gao, F., and Zhang, C.-T. (2006). GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. Nucleic Acids Res *34*, W686–W691. https://doi.org/10.1093/nar/gkl040.

Giglio, S., Broman, K.W., Matsumoto, N., Calvari, V., Gimelli, G., Neumann, T., Ohashi, H., Voullaire, L., Larizza, D., Giorda, R., et al. (2001). Olfactory Receptor–Gene Clusters, Genomic-Inversion Polymorphisms, and Common Chromosome Rearrangements. Am J Hum Genet *68*, 874–883. .

Giorgianni, M.W., Dowell, N.L., Griffin, S., Kassner, V.A., Selegue, J.E., and Carroll, S.B. (2020). The origin and diversification of a novel protein family in venomous snakes. PNAS *117*, 10911–10920. https://doi.org/10.1073/pnas.1920011117.

Glémin, S., Arndt, P.F., Messer, P.W., Petrov, D., Galtier, N., and Duret, L. (2015). Quantification of GC-biased gene conversion in the human genome. Genome Res. *25*, 1215–1228. https://doi.org/10.1101/gr.185488.114.

Glusman, G., Yanai, I., Rubin, I., and Lancet, D. (2001). The complete human olfactory subgenome. Genome Res *11*, 685–702. https://doi.org/10.1101/gr.171001.

Goeke, S., Greene, E.A., Grant, P.K., Gates, M.A., Crowner, D., Aigaki, T., and Giniger, E. (2003). Alternative splicing of lola generates 19 transcription factors controlling axon guidance in Drosophila. Nat Neurosci *6*, 917–924. https://doi.org/10.1038/nn1105.

Grand, R.S., Burger, L., Gräwe, C., Michael, A.K., Isbel, L., Hess, D., Hoerner, L., Iesmantavicius, V., Durdu, S., Pregnolato, M., et al. (2021). BANP opens chromatin and activates CpG-island-regulated genes. Nature 1–5. https://doi.org/10.1038/s41586-021-03689-8.

Grimwood, J., Gordon, L.A., Olsen, A., Terry, A., Schmutz, J., Lamerdin, J., Hellsten, U., Goodstein, D., Couronne, O., Tran-Gyamfi, M., et al. (2004). The DNA sequence and biology of human chromosome 19. Nature *428*, 529–535. https://doi.org/10.1038/nature02399.

1016 GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 580–
1017 585. https://doi.org/10.1038/ng.2653.

1018 Halldorsson, B.V., Palsson, G., Stefansson, O.A., Jonsson, H., Hardarson, M.T., Eggertsson,
1019 H.P., Gunnarsson, B., Oddsson, A., Halldorsson, G.H., Zink, F., et al. (2019). Characterizing
1020 mutagenic effects of recombination through a sequence-level genetic map. Science *363*.
1021 https://doi.org/10.1126/science.aau1043.

1022 Hardison, R.C. (2012). Evolution of Hemoglobin and Its Genes. Cold Spring Harb Perspect Med
1023 *2*, a011627. https://doi.org/10.1101/cshperspect.a011627.

1024 Hershberg, R., and Petrov, D.A. (2010). Evidence that mutation is universally biased towards AT
1025 in bacteria. PLoS Genet *6*, e1001115. https://doi.org/10.1371/journal.pgen.1001115.

1026 Hetz, C., Zhang, K., and Kaufman, R.J. (2020). Mechanisms, regulation and functions of the
1027 unfolded protein response. Nature Reviews Molecular Cell Biology *21*, 421–438.
1028 https://doi.org/10.1038/s41580-020-0250-z.

1029 Hildebrand, F., Meyer, A., and Eyre-Walker, A. (2010). Evidence of selection upon genomic
1030 GC-content in bacteria. PLoS Genet *6*, e1001107. https://doi.org/10.1371/journal.pgen.1001107.

1031 Holding, M.L., Strickland, J.L., Rautsaw, R.M., Hofmann, E.P., Mason, A.J., Hogan, M.P.,
1032 Nystrom, G.S., Ellsworth, S.A., Colston, T.J., Borja, M., et al. (2021). Phylogenetically diverse
1033 diets favor more complex venoms in North American pitvipers. Proc Natl Acad Sci U S A *118*,
1034 e2015579118. https://doi.org/10.1073/pnas.2015579118.

1035 Holmquist, G.P. (1992). Chromosome bands, their chromatin flavors, and their functional
1036 features. Am J Hum Genet *51*, 17–37. .

1037 Holmquist, G.P., and Filipski, J. (1994). Organization of mutations along the genome: a prime
1038 determinant of genome evolution. Trends Ecol Evol *9*, 65–69. https://doi.org/10.1016/0169-
1039 5347(94)90277-1.

1040 Hultén, M. (1974). Chiasma distribution at diakinesis in the normal human male. Hereditas *76*,
1041 55–78. https://doi.org/10.1111/j.1601-5223.1974.tb01177.x.

1042 Huttener, R., Thorrez, L., in't Veld, T., Granvik, M., Snoeck, L., Van Lommel, L., and Schuit, F.
1043 (2019). GC content of vertebrate exome landscapes reveal areas of accelerated protein evolution.
1044 BMC Evolutionary Biology *19*, 144. https://doi.org/10.1186/s12862-019-1469-1.

1045 Jabbari, K., and Bernardi, G. (2017). An Isochore Framework Underlies Chromatin Architecture.
1046 PLOS ONE *12*, e0168023. https://doi.org/10.1371/journal.pone.0168023.

1047 Jabbari, K., Chakraborty, M., and Wiehe, T. (2019a). DNA sequence-dependent chromatin
1048 architecture and nuclear hubs formation. Sci Rep *9*, 14646. https://doi.org/10.1038/s41598-019-
1049 51036-9.

1050 Jabbari, K., Wirtz, J., Rauscher, M., and Wiehe, T. (2019b). A common genomic code for
1051 chromatin architecture and recombination landscape. PLOS ONE *14*, e0213278.
1052 https://doi.org/10.1371/journal.pone.0213278.

1053 Jensen-Seaman, M.I., Furey, T.S., Payseur, B.A., Lu, Y., Roskin, K.M., Chen, C.-F., Thomas,
1054 M.A., Haussler, D., and Jacob, H.J. (2004). Comparative Recombination Rates in the Rat,
1055 Mouse, and Human Genomes. Genome Res. *14*, 528–538. https://doi.org/10.1101/gr.1970304.

1056 Johnson, R.N., O'Meally, D., Chen, Z., Etherington, G.J., Ho, S.Y.W., Nash, W.J., Grueber,
1057 C.E., Cheng, Y., Whittington, C.M., Dennison, S., et al. (2018). Adaptation and conservation
1058 insights from the koala genome. Nat Genet *50*, 1102–1111. https://doi.org/10.1038/s41588-018-
1059 0153-5.

1060 Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M.T.,
1061 Hjorleifsson, K.E., Eggertsson, H.P., Gudjonsson, S.A., et al. (2017). Parental influence on
1062 human germline de novo mutations in 1,548 trios from Iceland. Nature *549*, 519–522.
1063 https://doi.org/10.1038/nature24018.

1064 Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins,
1065 R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint
1066 spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.
1067 https://doi.org/10.1038/s41586-020-2308-7.

1068 Kawasaki, K., Lafont, A.-G., and Sire, J.-Y. (2011). The evolution of milk casein genes from
1069 tooth genes before the origin of mammals. Mol Biol Evol *28*, 2053–2061.
1070 https://doi.org/10.1093/molbev/msr020.

1071 Kim, J., Farré, M., Auvil, L., Capitanu, B., Larkin, D.M., Ma, J., and Lewin, H.A. (2017).
1072 Reconstruction and evolutionary history of eutherian chromosomes. PNAS *114*, E5379–E5388. .

1073 Kong, A., Thorleifsson, G., Gudbjartsson, D.F., Masson, G., Sigurdsson, A., Jonasdottir, A.,
1074 Walters, G.B., Jonasdottir, A., Gylfason, A., Kristinsson, K.T., et al. (2010). Fine-scale
1075 recombination rate differences between sexes, populations and individuals. Nature *467*, 1099–
1076 1103. https://doi.org/10.1038/nature09525.

1077 Korenberg, J.R., and Rykowski, M.C. (1988). Human genome organization: Alu, lines, and the
1078 molecular structure of metaphase chromosome bands. Cell *53*, 391–400.
1079 https://doi.org/10.1016/0092-8674(88)90159-6.

1080 Labrador, M., and Corces, V.G. (2003). Extensive Exon Reshuffling Over Evolutionary Time
1081 Coupled to Trans-Splicing in Drosophila. Genome Res. *13*, 2220–2228.
1082 https://doi.org/10.1101/gr.1440703.

1083 Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K.,
1084 Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human
1085 genome. Nature *409*, 860–921. https://doi.org/10.1038/35057062.

1086 Lane, R.P., Young, J., Newman, T., and Trask, B.J. (2004). Species specificity in rodent
1087 pheromone receptor repertoires. Genome Res *14*, 603–608. https://doi.org/10.1101/gr.2117004.

1088 de Lange, T. (2015). A loopy view of telomere evolution. Frontiers in Genetics *6*. .

1089 Lee, Y.-S., Chao, A., Chen, C.-H., Chou, T., Wang, S.-Y.M., and Wang, T.-H. (2011). Analysis
1090 of human meiotic recombination events with a parent-sibling tracing approach. BMC Genomics
1091 *12*, 434. https://doi.org/10.1186/1471-2164-12-434.

1092 Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-
1093 Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding
1094 genetic variation in 60,706 humans. Nature *536*, 285–291. https://doi.org/10.1038/nature19057.

1095 Li, Q., Peterson, K.R., Fang, X., and Stamatoyannopoulos, G. (2002). Locus control regions.
1096 Blood *100*, 3077–3086. https://doi.org/10.1182/blood-2002-04-1104.

1097 Linardopoulou, E.V., Williams, E.M., Fan, Y., Friedman, C., Young, J.M., and Trask, B.J.
1098 (2005). Human subtelomeres are hot spots of interchromosomal recombination and segmental
1099 duplication. Nature *437*, 94–100. https://doi.org/10.1038/nature04029.

1100 Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J.,
1101 Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013). Global Epigenomic Reconfiguration
1102 During Mammalian Brain Development. Science *341*, 1237905.
1103 https://doi.org/10.1126/science.1237905.

1104 Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. Proceedings
1105 of the National Academy of Sciences *107*, 961–968. https://doi.org/10.1073/pnas.0912629107.

1106 Lynch, Michael (2007). The origins of genome architecture (Indiana University Press).

1107 Magklara, A., Yen, A., Colquitt, B.M., Clowney, E.J., Allen, W., Markenscoff-Papadimitriou,
1108 E., Evans, Z.A., Kheradpour, P., Mountoufaris, G., Carey, C., et al. (2011). An epigenetic
1109 signature for monoallelic olfactory receptor expression. Cell *145*, 555–570.
1110 https://doi.org/10.1016/j.cell.2011.03.040.

1111 Mainland, J.D., Keller, A., Li, Y.R., Zhou, T., Trimmer, C., Snyder, L.L., Moberly, A.H.,
1112 Adipietro, K.A., Liu, W.L.L., Zhuang, H., et al. (2014). The missense of smell: functional
1113 variability in the human odorant receptor repertoire. Nat Neurosci *17*, 114–120.
1114 https://doi.org/10.1038/nn.3598.

1115 Mann, R.S. (1997). Why are Hox genes clustered? Bioessays *19*, 661–664.
1116 https://doi.org/10.1002/bies.950190804.

1117 Markenscoff-Papadimitriou, E., Allen, W.E., Colquitt, B.M., Goh, T., Murphy, K.K., Monahan,
1118 K., Mosley, C.P., Ahituv, N., and Lomvardas, S. (2014). Enhancer interaction networks as a
1119 means for singular olfactory receptor expression. Cell *159*, 543–557.
1120 https://doi.org/10.1016/j.cell.2014.09.033.

McKenzie, S.K., Fetter-Pruneda, I., Ruta, V., and Kronauer, D.J.C. (2016). Transcriptomics and neuroanatomy of the clonal raider ant implicate an expanded clade of odorant receptors in chemical communication. Proc. Natl. Acad. Sci. U.S.A. *113*, 14091–14096. https://doi.org/10.1073/pnas.1610800113.

Mefford, H.C., Linardopoulou, E., Coil, D., van den Engh, G., and Trask, B.J. (2001). Comparative sequencing of a multicopy subtelomeric region containing olfactory receptor genes reveals multiple interactions between non-homologous chromosomes. Hum Mol Genet *10*, 2363–2372. https://doi.org/10.1093/hmg/10.21.2363.

Michaloski, J.S., Galante, P.A.F., and Malnic, B. (2006). Identification of potential regulatory motifs in odorant receptor genes by analysis of promoter sequences. Genome Res *16*, 1091–1098. https://doi.org/10.1101/gr.5185406.

Micklem, G., and Hillier, L.W. (2006). CpG Islands. Unpublished. http://genomewiki.ucsc.edu/index.php/CpG_Islands. Accessed 6/21/22.

Mohn, F., and Schübeler, D. (2009). Genetics and epigenetics: stability and plasticity during cellular differentiation. Trends Genet *25*, 129–136. https://doi.org/10.1016/j.tig.2008.12.005.

Monahan, K., Schieren, I., Cheung, J., Mumbey-Wafula, A., Monuki, E.S., and Lomvardas, S. (2017). Cooperative interactions enable singular olfactory receptor expression in mouse olfactory neurons. Elife *6*. https://doi.org/10.7554/eLife.28620.

Monahan, K., Horta, A., and Lomvardas, S. (2019). LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. Nature *565*, 448–453. https://doi.org/10.1038/s41586-018-0845-0.

Monroe, J.G., Srikant, T., Carbonell-Bejerano, P., Becker, C., Lensink, M., Exposito-Alonso, M., Klein, M., Hildebrandt, J., Neumann, M., Kliebenstein, D., et al. (2022). Mutation bias reflects natural selection in Arabidopsis thaliana. Nature *602*, 101–105. https://doi.org/10.1038/s41586-021-04269-6.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., de Laat, W., Spitz, F., and Duboule, D. (2011). A regulatory archipelago controls Hox genes transcription in digits. Cell *147*, 1132–1145. https://doi.org/10.1016/j.cell.2011.10.023.

Morales, J., Pujar, S., Loveland, J.E., Astashyn, A., Bennett, R., Berry, A., Cox, E., Davidson, C., Ermolaeva, O., Farrell, C.M., et al. (2022). A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature *604*, 310–315. https://doi.org/10.1038/s41586-022-04558-8.

Myers, S., Bowden, R., Tumian, A., Bontrop, R.E., Freeman, C., MacFie, T.S., McVean, G., and Donnelly, P. (2010). Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science *327*, 876–879. https://doi.org/10.1126/science.1182363.

Naughton, C., Avlonitis, N., Corless, S., Prendergast, J.G., Mati, I.K., Eijk, P.P., Cockroft, S.L., Bradley, M., Ylstra, B., and Gilbert, N. (2013). Transcription forms and remodels supercoiling

1158 domains unfolding large-scale chromatin structures. Nat Struct Mol Biol *20*, 387–395.
1159 https://doi.org/10.1038/nsmb.2509.

1160 Nei, M., Niimura, Y., and Nozawa, M. (2008). The evolution of animal chemosensory receptor
1161 gene repertoires: roles of chance and necessity. Nature Reviews Genetics *9*, 951–963.
1162 https://doi.org/10.1038/nrg2480.

1163 Nelson, D.R., Goldstone, J.V., and Stegeman, J.J. (2013). The cytochrome P450 genesis locus:
1164 the origin and evolution of animal cytochrome P450s. Philos Trans R Soc Lond B Biol Sci *368*,
1165 20120474. https://doi.org/10.1098/rstb.2012.0474.

1166 Newman, T., and Trask, B.J. (2003). Complex Evolution of 7E Olfactory Receptor Genes in
1167 Segmental Duplications. Genome Res. *13*, 781–793. https://doi.org/10.1101/gr.769003.

1168 Nicetto, D., Donahue, G., Jain, T., Peng, T., Sidoli, S., Sheng, L., Montavon, T., Becker, J.S.,
1169 Grindheim, J.M., Blahnik, K., et al. (2019). H3K9me3-heterochromatin loss at protein-coding
1170 genes enables developmental lineage specification. Science *363*, 294–297.
1171 https://doi.org/10.1126/science.aau0583.

1172 Niimura, Y., and Gojobori, T. (2002). In silico chromosome staining: Reconstruction of Giemsa
1173 bands from the whole human genome sequence. Proc Natl Acad Sci U S A *99*, 797–802.
1174 https://doi.org/10.1073/pnas.022437999.

1175 Niimura, Y., and Nei, M. (2007). Extensive Gains and Losses of Olfactory Receptor Genes in
1176 Mammalian Evolution. PLOS ONE *2*, e708. https://doi.org/10.1371/journal.pone.0000708.

1177 North, H.L., Caminade, P., Severac, D., Belkhir, K., and Smadja, C.M. (2020). The role of copy-
1178 number variation in the reinforcement of sexual isolation between the two European subspecies
1179 of the house mouse. Philosophical Transactions of the Royal Society B: Biological Sciences *375*,
1180 20190540. https://doi.org/10.1098/rstb.2019.0540.

1181 Ohno, M., Miura, T., Furuichi, M., Tominaga, Y., Tsuchimoto, D., Sakumi, K., and Nakabeppu,
1182 Y. (2006). A genome-wide distribution of 8-oxoguanine correlates with the preferred regions for
1183 recombination and single nucleotide polymorphism in the human genome. Genome Res *16*, 567–
1184 575. https://doi.org/10.1101/gr.4769606.

1185 Ohno, Susumu (1970). Evolution by gene duplication. (Springer).

1186 Otto, M., Zheng, Y., and Wiehe, T. (2022). Recombination, selection, and the evolution of
1187 tandem gene arrays. Genetics iyac052. https://doi.org/10.1093/genetics/iyac052.

1188 Parvanov, E.D., Petkov, P.M., and Paigen, K. (2010). Prdm9 controls activation of mammalian
1189 recombination hotspots. Science *327*, 835. https://doi.org/10.1126/science.1181495.

1190 Paudel, Y., Madsen, O., Megens, H.-J., Frantz, L.A.F., Bosse, M., Crooijmans, R.P.M.A., and
1191 Groenen, M.A.M. (2015). Copy number variation in the speciation of pigs: a possible prominent
1192 role for olfactory receptors. BMC Genomics *16*, 330. https://doi.org/10.1186/s12864-015-1449-
1193 9.

1194 Pavlovich, S.S., Lovett, S.P., Koroleva, G., Guito, J.C., Arnold, C.E., Nagle, E.R., Kulcsar, K.,
1195 Lee, A., Thibaud-Nissen, F., Hume, A.J., et al. (2018). The Egyptian Rousette Genome Reveals
1196 Unexpected Features of Bat Antiviral Immunity. Cell *173*, 1098-1110.e18.
1197 https://doi.org/10.1016/j.cell.2018.03.070.

1198 Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea,
1199 F.A., Mountain, J.L., Misra, R., et al. (2007). Diet and the evolution of human amylase gene
1200 copy number variation. Nat Genet *39*, 1256–1260. https://doi.org/10.1038/ng2123.

1201 Pouyet, F., Mouchiroud, D., Duret, L., and Sémon, M. (2017). Recombination, meiotic
1202 expression and human codon usage. ELife *6*, e27344. https://doi.org/10.7554/eLife.27344.

1203 Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G.V., and Camerini-Otero, R.D.
1204 (2014). Recombination initiation maps of individual human genomes. Science *346*, 1256442.
1205 https://doi.org/10.1126/science.1256442.

1206 Pratto, F., Brick, K., Cheng, G., Lam, K.-W.G., Cloutier, J.M., Dahiya, D., Wellard, S.R.,
1207 Jordan, P.W., and Camerini-Otero, R.D. (2021). Meiotic recombination mirrors patterns of
1208 germline replication in mice and humans. Cell *184*, 4251-4267.e20.
1209 https://doi.org/10.1016/j.cell.2021.06.025.

1210 Qu, Q., Haitina, T., Zhu, M., and Ahlberg, P.E. (2015). New genomic and fossil data illuminate
1211 the origin of enamel. Nature *526*, 108–111. https://doi.org/10.1038/nature15259.

1212 Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing
1213 genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

1214 Ramani, V., Shendure, J., and Duan, Z. (2016). Understanding Spatial Genome Organization:
1215 Methods and Insights. Genomics, Proteomics & Bioinformatics *14*, 7–20.
1216 https://doi.org/10.1016/j.gpb.2016.01.002.

1217 Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S.,
1218 Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-
1219 sequencing data analysis. Nucleic Acids Res *44*, W160–W165.
1220 https://doi.org/10.1093/nar/gkw257.

1221 Ribich, S., Tasic, B., and Maniatis, T. (2006). Identification of long-range regulatory elements in
1222 the protocadherin-α gene cluster. PNAS *103*, 19719–19724.
1223 https://doi.org/10.1073/pnas.0609445104.

1224 Riethman, H., Ambrosini, A., Castaneda, C., Finklestein, J., Hu, X.-L., Mudunuri, U., Paul, S.,
1225 and Wei, J. (2004). Mapping and Initial Analysis of Human Subtelomeric Sequence Assemblies.
1226 Genome Res. *14*, 18–28. https://doi.org/10.1101/gr.1245004.

1227 Rodriguez-Galindo, M., Casillas, S., Weghorn, D., and Barbadilla, A. (2020). Germline de novo
1228 mutation rates on exons versus introns in humans. Nat Commun *11*, 3304.
1229 https://doi.org/10.1038/s41467-020-17162-z.

1230 Rogers, R.L. (2015). Chromosomal Rearrangements as Barriers to Genetic Homogenization
1231 between Archaic and Modern Humans. Molecular Biology and Evolution *32*, 3064–3078.
1232 https://doi.org/10.1093/molbev/msv204.

1233 Rouquier, S., Taviaux, S., Trask, B.J., Brand-Arpon, V., van den Engh, G., Demaille, J., and
1234 Giorgi, D. (1998). Distribution of olfactory receptor genes in the human genome. Nat Genet *18*,
1235 243–250. https://doi.org/10.1038/ng0398-243.

1236 Roy, A.L., Sen, R., and Roeder, R.G. (2011). Enhancer–promoter communication and
1237 transcriptional regulation of Igh. Trends in Immunology *32*, 532–539.
1238 https://doi.org/10.1016/j.it.2011.06.012.

1239 Schield, D.R., Card, D.C., Hales, N.R., Perry, B.W., Pasquesi, G.M., Blackmon, H., Adams,
1240 R.H., Corbin, A.B., Smith, C.F., Ramesh, B., et al. (2019). The origins and evolution of
1241 chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes.
1242 Genome Res *29*, 590–601. https://doi.org/10.1101/gr.240952.118.

1243 Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L., et
1244 al. (2016). A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the
1245 Human Genome. Cell Reports *17*, 2042–2059. https://doi.org/10.1016/j.celrep.2016.10.061.

1246 Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer–promoter contacts in gene
1247 expression control. Nat Rev Genet *20*, 437–455. https://doi.org/10.1038/s41576-019-0128-0.

1248 Schug, J., Schuller, W.-P., Kappen, C., Salbaum, J.M., Bucan, M., and Stoeckert, C.J. (2005).
1249 Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biology
1250 *6*, R33. https://doi.org/10.1186/gb-2005-6-4-r33.

1251 Schwartz, J.C., Gibson, M.S., Heimeier, D., Koren, S., Phillippy, A.M., Bickhart, D.M., Smith,
1252 T.P.L., Medrano, J.F., and Hammond, J.A. (2017). The evolution of the natural killer complex; a
1253 comparison between mammals using new high-quality genome assemblies and targeted
1254 annotation. Immunogenetics *69*, 255–269. https://doi.org/10.1007/s00251-017-0973-y.

1255 Semple, F., and Dorin, J.R. (2012). β-Defensins: Multifunctional Modulators of Infection,
1256 Inflammation and More? JIN *4*, 337–348. https://doi.org/10.1159/000336619.

1257 Shelton, J.F., Shastri, A.J., Fletez-Brant, K., Stella Aslibekyan, and Auton, A. (2022). The
1258 UGT2A1/UGT2A2 locus is associated with COVID-19-related loss of smell or taste. Nat Genet
1259 *54*, 121–124. https://doi.org/10.1038/s41588-021-00986-w.

1260 Simmen, M.W. (2008). Genome-scale relationships between cytosine methylation and
1261 dinucleotide abundances in animals. Genomics *92*, 33–40.
1262 https://doi.org/10.1016/j.ygeno.2008.03.009.

1263 Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. Nat
1264 Rev Genet *14*, 204–220. https://doi.org/10.1038/nrg3354.

1265 Sun, X., Zhang, Z., Sun, Y., Li, J., Xu, S., and Yang, G. (2017). Comparative genomics analyses
1266 of alpha-keratins reveal insights into evolutionary adaptation of marine mammals. Frontiers in
1267 Zoology *14*, 41. https://doi.org/10.1186/s12983-017-0225-x.

1268 Sved, J., and Bird, A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate
1269 genomes under a mutation model. Proc Natl Acad Sci U S A *87*, 4692–4696.
1270 https://doi.org/10.1073/pnas.87.12.4692.

1271 Tan, H.M., and Low, W.Y. (2018). Rapid birth-death evolution and positive selection in
1272 detoxification-type glutathione S-transferases in mammals. PLOS ONE *13*, e0209336.
1273 https://doi.org/10.1371/journal.pone.0209336.

1274 Tan, L., Xing, D., Daley, N., and Xie, X.S. (2019). Three-dimensional genome structures of
1275 single sensory neurons in mouse visual and olfactory systems. Nature Structural & Molecular
1276 Biology *26*, 297–307. https://doi.org/10.1038/s41594-019-0205-2.

1277 Tan, L., Ma, W., Wu, H., Zheng, Y., Xing, D., Chen, R., Li, X., Daley, N., Deisseroth, K., and
1278 Xie, X.S. (2021). Changes in genome architecture and transcriptional dynamics progress
1279 independently of sensory experience during post-natal brain development. Cell *184*, 741-
1280 758.e17. https://doi.org/10.1016/j.cell.2020.12.032.

1281 Tease, C., and Hultén, M.A. (2004). Inter-sex variation in synaptonemal complex lengths largely
1282 determine the different recombination rates in male and female germ cells. CGR *107*, 208–215.
1283 https://doi.org/10.1159/000080599.

1284 Thomas, J.H. (2007). Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes
1285 in vertebrates. PLoS Genet. *3*, e67. https://doi.org/10.1371/journal.pgen.0030067.

1286 Toyoda, S., Kawaguchi, M., Kobayashi, T., Tarusawa, E., Toyama, T., Okano, M., Oda, M.,
1287 Nakauchi, H., Yoshimura, Y., Sanbo, M., et al. (2014). Developmental Epigenetic Modification
1288 Regulates Stochastic Expression of Clustered Protocadherin Genes, Generating Single Neuron
1289 Diversity. Neuron *82*, 94–108. https://doi.org/10.1016/j.neuron.2014.02.005.

1290 Trask, B.J., Friedman, C., Martin-Gallardo, A., Rowen, L., Akinbami, C., Blankenship, J.,
1291 Collins, C., Giorgi, D., Iadonato, S., Johnson, F., et al. (1998). Members of the olfactory receptor
1292 gene family are contained in large blocks of DNA duplicated polymorphically near the ends of
1293 human chromosomes. Human Molecular Genetics *7*, 13–26. https://doi.org/10.1093/hmg/7.1.13.

1294 Trimmer, C., Keller, A., Murphy, N.R., Snyder, L.L., Willer, J.R., Nagai, M.H., Katsanis, N.,
1295 Vosshall, L.B., Matsunami, H., and Mainland, J.D. (2019). Genetic variation across the human
1296 olfactory receptor repertoire alters odor perception. PNAS *116*, 9475–9480.
1297 https://doi.org/10.1073/pnas.1804106115.

1298 Tweedie, S., Braschi, B., Gray, K., Jones, T.E.M., Seal, R.L., Yates, B., and Bruford, E.A.
1299 (2021). Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Research *49*,
1300 D939–D946. https://doi.org/10.1093/nar/gkaa980.

1301 Venables, J.P., Tazi, J., and Juge, F. (2012). Regulated functional alternative splicing in
1302 Drosophila. Nucleic Acids Res *40*, 1–10. https://doi.org/10.1093/nar/gkr648.

1303 Wang, Z., and Willard, H.F. (2012). Evidence for sequence biases associated with patterns of
1304 histone methylation. BMC Genomics *13*, 367. https://doi.org/10.1186/1471-2164-13-367.

1305 Williams, D.L., Sikora, V.M., Hammer, M.A., Amin, S., Brinjikji, T., Brumley, E.K., Burrows,
1306 C.J., Carrillo, P.M., Cromer, K., Edwards, S.J., et al. (2021). May the Odds Be Ever in Your
1307 Favor: Non-deterministic Mechanisms Diversifying Cell Surface Molecule Expression. Front
1308 Cell Dev Biol *9*, 720798. https://doi.org/10.3389/fcell.2021.720798.

1309 Witt, M., and Hummel, T. (2006). Vomeronasal Versus Olfactory Epithelium: Is There a
1310 Cellular Basis for Human Vomeronasal Perception? In International Review of Cytology,
1311 (Academic Press), pp. 209–259.

1312 Wong, E.S., Zheng, D., Tan, S.Z., Bower, N.L., Garside, V., Vanwalleghem, G., Gaiti, F., Scott,
1313 E., Hogan, B.M., Kikuchi, K., et al. (2020). Deep conservation of the enhancer regulatory code
1314 in animals. Science *370*, eaax8137. https://doi.org/10.1126/science.aax8137.

1315 Woodfine, K., Fiegler, H., Beare, D.M., Collins, J.E., McCann, O.T., Young, B.D., Debernardi,
1316 S., Mott, R., Dunham, I., and Carter, N.P. (2004). Replication timing of the human genome. Hum
1317 Mol Genet *13*, 191–202. https://doi.org/10.1093/hmg/ddh016.

1318 Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al.
1319 (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The
1320 Innovation *2*, 100141. https://doi.org/10.1016/j.xinn.2021.100141.

1321 Xie, W.J., Meng, L., Liu, S., Zhang, L., Cai, X., and Gao, Y.Q. (2017). Structural Modeling of
1322 Chromatin Integrates Genome Features and Reveals Chromosome Folding Principle. Sci Rep *7*.
1323 https://doi.org/10.1038/s41598-017-02923-6.

1324 Yokota, S., Hirayama, T., Hirano, K., Kaneko, R., Toyoda, S., Kawamura, Y., Hirabayashi, M.,
1325 Hirabayashi, T., and Yagi, T. (2011). Identification of the cluster control region for the
1326 protocadherin-beta genes located beyond the protocadherin-gamma cluster. J Biol Chem *286*,
1327 31885–31895. https://doi.org/10.1074/jbc.M111.245605.

1328 Young, J.M., Endicott, R.M., Parghi, S.S., Walker, M., Kidd, J.M., and Trask, B.J. (2008).
1329 Extensive copy-number variation of the human olfactory receptor gene family. Am J Hum Genet
1330 *83*, 228–242. https://doi.org/10.1016/j.ajhg.2008.07.005.

1331 Yue, Y., and Haaf, T. (2006). 7E olfactory receptor gene clusters and evolutionary chromosome
1332 rearrangements. Cytogenet Genome Res *112*, 6–10. https://doi.org/10.1159/000087507.

1333 Zhang, J., and Webb, D.M. (2003). Evolutionary deterioration of the vomeronasal pheromone
1334 transduction pathway in catarrhine primates. PNAS *100*, 8337–8341.
1335 https://doi.org/10.1073/pnas.1331721100.

Zhang, J., Zhang, Y., You, Q., Huang, C., Zhang, T., Wang, M., Zhang, T., Yang, X., Xiong, J., Li, Y., et al. (2022). Highly enriched BEND3 prevents the premature activation of bivalent genes during differentiation. Science *375*, 1053–1058. https://doi.org/10.1126/science.abm0730.

Zody, M.C., Garber, M., Adams, D.J., Sharpe, T., Harrow, J., Lupski, J.R., Nicholson, C., Searle, S.M., Wilming, L., Young, S.K., et al. (2006). DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. Nature *440*, 1045–1049. https://doi.org/10.1038/nature04689.