Emergence and influence of sequence bias in evolutionarily malleable, mammalian tandem arrays

Margarita V Brovkina[1,*], Margaret A. Chapman[2,*], Matthew L. Holding[3], and E. Josephine Clowney[4,5,#]

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. Life Sciences Institute, University of Michigan, Ann Arbor, Michigan, United States of America

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

5. 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. Genes in these "outward-looking" families are allelically diverse among humans and exhibit tissue-specific and sometimes stochastic expression. Here, we show that outward-looking genes are clustered in tandem arrays, enriched in AT-biased isochores, and lack CpG islands in their promoters. Models of mammalian genome evolution have not incorporated the sharply different functions and transcriptional patterns of genes in AT- versus GC-biased regions. To examine the relationship between gene family expansion, sequence content, and functional diversification, we use population genetic data and comparative analysis. First, we find that AT bias can emerge with gene family expansion *in cis*. Second, human genes in AT-biased isochores or with GC-poor promoters experience relatively low rates of *de novo* point mutation today but are enriched for functional variants. Finally, we find that isochores containing gene clusters exhibit low rates of recombination. We hypothesize that the depletion of GC bases in outward-facing gene clusters results from tolerance of sequence variation and low recombination. In turn, high AT content 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 (1–3). Gene blooms are expansions *in cis* that result in arrays of dozens or even hundreds of genes. During gametogenesis, these tandem duplication events are thought to arise via incorrect crossovers between paralogues or via non-homologous repair of chromosome breaks (4). 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 response in bats, keratins for whale baleen, venom production in snakes, and amylase copy number for starch consumption in modern humans (1,2,5–11). The definitive gene family of mammals, the caseins, arose through local duplication of enamel genes (12). In the mouse, we have shown that genes in copy-number-variable blooms exhibit extremely high AT content in their promoters and are often located in AT-biased regions of the genome (13).

GC content in mammalian genomes varies markedly at the megabase scale (14). 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 (15). Banding patterns were found to reflect local variation in AT/GC content: Giemsa-staining "G bands" are AT-biased, and Quinacrine-staining "Q bands" are GC-biased (15–18). 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" (15,19). Moreover, human-chimp divergence rates were found to be higher in AT-rich G-bands than in GC-rich Q bands (20). In the genome sequencing era, 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 (21,22). 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 (22–25).

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 (26,27), leading to adaptationist speculation that isochore structure serves a function related to endothermy (28). However, consensus has emerged that GC-biased gene conversion (gBGC) following meiotic recombination, which occurs in most eukaryotes, is one important contributor to isochore emergence in mammals (26,29). 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 by Pouyet 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" (30,31). In this model, the isochores themselves do not serve an adaptive function, but rather have emerged due to the molecular genetic ("neutral") forces of meiosis. Over evolutionary time, the GC-increasing effect of recombination is counteracted by the AT-increasing effect of point mutation due to mutation of fragile cytosines to thymines (32–37). As the rate of recombination and cytosine loss may themselves be influenced by sequence context (i.e. recombination more likely in GC-rich regions, cytosine mutation more common in AT-rich regions), positive feedback could have caused large genomic regions to diverge (20,30,32,38). However, the influence of these neutral forces on the genes contained within AT- versus GC-biased isochores has not been described.

Here, we examine the local and isochore-level AT/GC content of human protein-coding genes. Genes located in AT-rich regions of the genome have unique and consistent characteristics: They are copy-number-variable families located in tandem arrays, are expressed in terminally differentiated cells, are cell surface or secreted proteins, lack CpG islands in their promoters, 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. 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 (16,39–41). 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.

Next, we ask how mammalian genes with outward-looking functions came to be located in AT-rich regions of the genome. By comparing gene blooms of different sizes within the human genome and across mammalian species, we find that AT content is not necessarily inherited from the ancestral species but can emerge with cluster expansion. Many of the mutagenic and repair processes that contribute to the neutral mutation spectrum are likely to differ in strength across AT-rich and GC-rich regions (reviewed in (42)). Using human 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 paralogous clusters are subject to less recombination than genes located near non-paralogues. Recombination may be dangerous in gene clusters due to the potential for chromosome rearrangements or within-cluster ectopic exchange that leads to duplications or deletions. It could also separate genes in large families from locus control regions they depend on for expression.

In addition, we find that while genes in AT-biased isochores have high sequence diversity among humans and divergence across species, they currently exhibit relatively low rates of *de-novo* point mutation, as expected from the neutral mutation spectrum. Instead, they appear to have accumulated sequence variants over evolutionary time. These genes also lack CpG islands in their promoters. We hypothesize that the functional roles of genes whose protein products interface with unpredictable and rapidly changing molecules in the environment ("outward-looking genes") make them particularly likely to tolerate point mutations. Lack of purifying selection on these genes is expected to shift GC content down over evolutionary time due to deamination of cytosine leading to C->T transitions (42); intolerance of recombination would prevent gBGC from shifting GC content back up (37). We propose a model in which reduced recombination and relaxation of selection on point mutations act together to strand arrays of outward-looking paralogues in wells of low GC content (Figure 1). Loss of CpG islands and residence in AT-rich genomic regions predisposes these genes to exotic forms of highly tissue-specific transcriptional regulation (43–47).

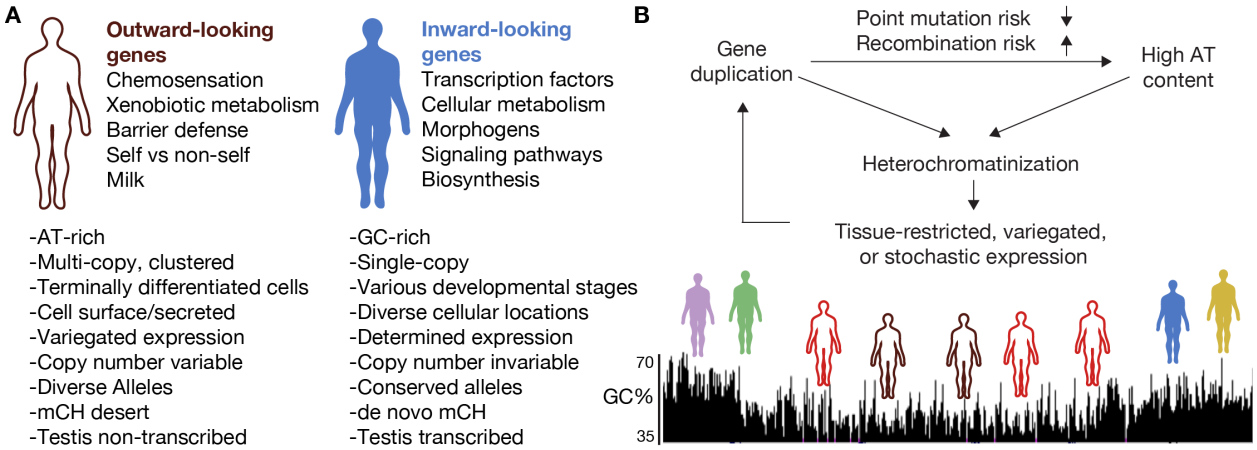

Figure 1: Outward- versus inward- looking genes (A) Summary of inward-looking and outward-looking gene families and their genomic distinctions. This table is inspired by (48). Descriptions derive from our analyses here and from references cited throughout the text. (B) Model of possible relationships between expansion of gene families in cis, genomic architecture, selective forces, and mode of expression.

Results

Characterizing a set of human isochores and their gene contents

We first comprehensively characterized the types of genes located in AT- versus GC-biased isochores and assessed if genes in AT-biased isochores are more likely to have numerous paralogues as neighbors. Isochore boundaries have been computed for previous genome assemblies, including for hg38 (22–24,49). While previous methods of isochore annotation binned the genome into 100kb pieces or used manual inspection to annotate isochore ends, we used a segmentation algorithm which detects transitions in GC content and created a UCSC Genome Browser track for easy visualization of isochore assignments versus other genomic elements (50). Cozzi et al compare various methods of isochore assignment, including GC Profile, showing that differences are subtle and are most prevalent in the mid-ranges of GC

155 content. We tested three resolutions and selected a map that visually matched 100kb-Mb
156 transitions in GC content using the %GC track on the UCSC Genome Browser (Fig 2A,B). At
157 this resolution, we call 4328 isochores; these range from ~30-70% GC and most are between
158 100kb and 5Mb (Figure 1C, D, S2A-C, Supplemental Tables 1 and 2).
159 We next defined the home isochore of each gene in the NCBI MANE set. The MANE set
160 includes one promoter and splice isoform for each intact, protein-coding gene and omits
161 pseudogenes and complex "gene parts" such as V, D, and J segments of the B and T cell
162 receptors (51). On average, higher-GC isochores were more likely to contain protein-coding
163 genes and had higher gene density (Figure 2C, Figure S2A, B) (15). To compare features across
164 isochores, we ordered isochores by GC% and divided them into ten groups (deciles) of ~400. To
165 test whether tandemly arrayed "gene blooms" were associated with AT-rich isochores, we used
166 Shannon's H to measure gene name prefix diversity in isochores with at least ten genes (Figure
167 2D). Gene names were least diverse in AT-rich isochores, consistent with the presence of tandem
168 arrays in these isochores. We extended our analysis to isochores with fewer genes (at least five)
169 and again found that gene arrays were less common in high-GC isochores (Figure 2E, S2H,I).
170 While AT-rich isochores are longer, gene diversity is lower in AT-rich isochores across the
171 length distribution and for isochores with different numbers of genes (Figure S2F, G). Arrays in
172 AT-rich isochores contained more paralogues, and the kinds of proteins present in AT-rich arrays
173 versus GC-rich arrays were different (Figure 2E, S2F): Arrays in AT-rich isochores often served
174 sensory, digestive, or barrier functions, while arrays in higher-GC isochores most often
175 contained *HOX* and *ZNF* transcription factors or arrays of histones (Figure 2E, S2I).
176 Only 25% of genes are located in isochores <40% GC. What sorts of gene families have
177 bloomed in these isochores? As GO analysis is biased by the annotations that are available,
178 especially for extremely tissue-specific genes, we simply searched for common gene name
179 prefixes in isochores <40% GC (Figure 2F). Gene families with at least four members in high-
180 AT isochores were overwhelmingly involved in chemosensation (*OR, TAAR, TAS2R, MS4*);
181 xenobiotic metabolism (e.g. *AMY1, UGT2, ADH*); and immune, defense, and barrier functions
182 (e.g. *KLR, IFN, DSC/DSG*). Human ORs were also shown previously to be located in AT-biased
183 isochores (52). Examples of high-GC isochores with diverse gene members and high-AT
184 isochores with repetitive gene members are shown in Figure 2A, B and Figure S2D,E. While
185 immunoglobulin parts do not appear in the MANE gene set, arrays of immunoglobulin V regions
186 are also highly AT-biased (Figure S2G). These analyses demonstrate that AT-biased regions of
187 the human genome contain tandem arrays of genes with outward-looking functions. We invite
188 readers to explore these patterns further on our TrackHub:
189 http://genome.ucsc.edu/s/mbrovkin/hg38.

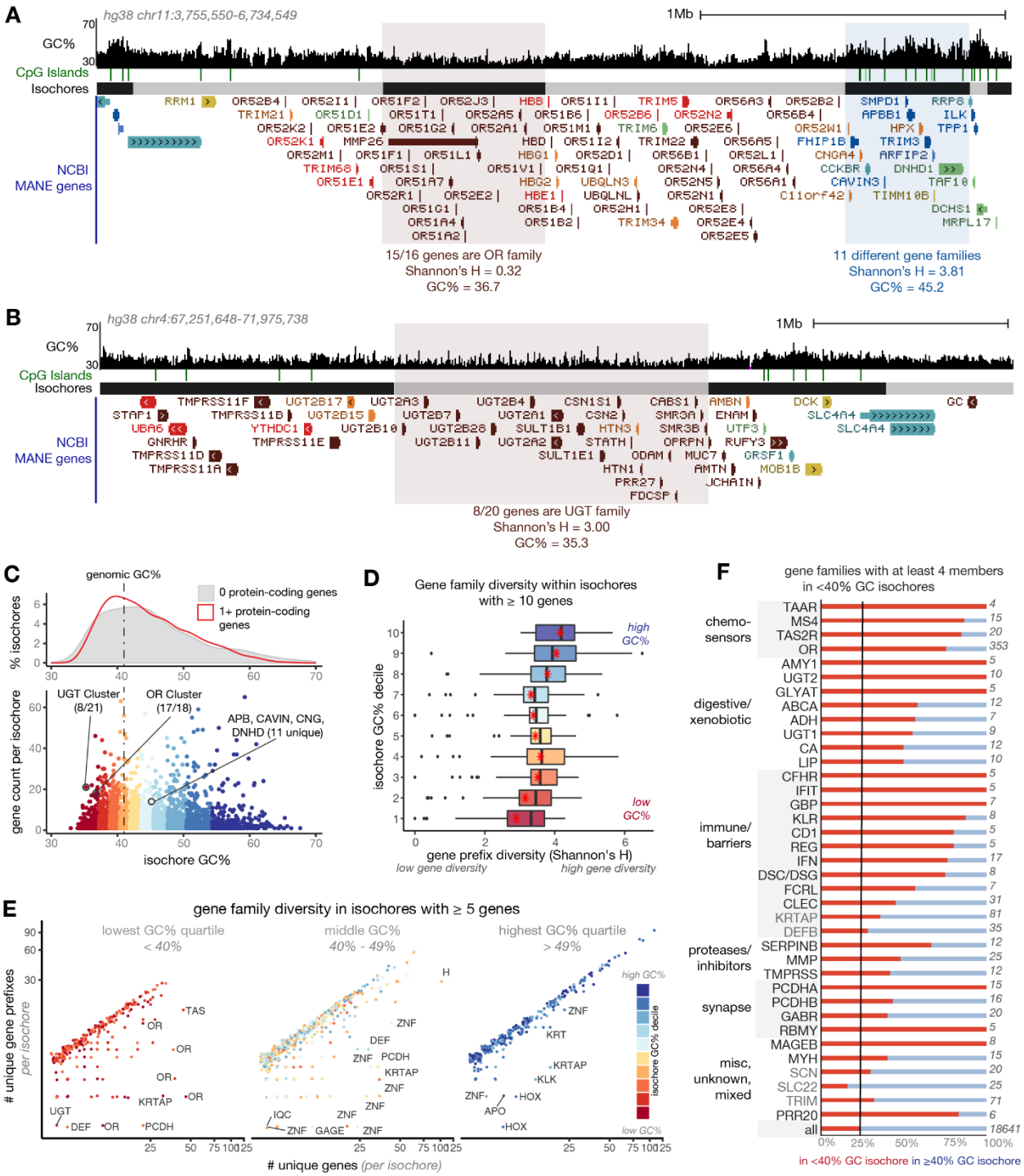

Figure 2: AT-rich isochores in the human genome contain tandem arrays of genes with outward-looking functions (A, B) UCSC Genome Browser screenshots showing GC% trajectory (53), CpG islands (54), our isochore calls, and simplified gene models. Gene models are colored according to k-means clusters described below. Gene name prefix diversity (Shannon's H) is shown for the highlighted isochores. (C) Relationship between isochore GC% and gene contents. Isochores shown in A, B are highlighted. Colors show how we binned isochores into deciles according to GC%. Dashed line shows the mean GC content of the human genome (41%). (D) Boxplots representing gene name diversity (Shannon's H) of gene-rich

199 isochores across GC% deciles shown in (C). Red points indicate the mean prefix diversity of the
200 decile; black lines show medians. Here and throughout, most groups are statistically different
201 from one another, except for adjoining groups; full statistical comparisons are presented in
202 Supplemental Table 3. (E) Comparison of number of genes in an isochore versus number of
203 different gene name prefixes represented in that isochore, for isochores of at least five genes.
204 Each dot is an isochore; isochores falling below the trend line contain clusters of genes with the
205 same prefix. Isochores are labeled by the most common gene name prefix in that isochore. (F)
206 Gene families with at least four members in isochores <40% GC. 25% of all genes are in
207 isochores less than 40% GC (black line). Red bars depict proportions of genes with that prefix
208 that are located in <40% GC isochores. Gene family prefixes shown in black text are enriched in
209 AT-rich isochores, while those shown in grey text (e.g. *KRTAP*, *TRIM*, *SCN*) have multiple
210 family members in AT-rich isochores but are not enriched there. Functions of these gene families
211 are marked at left, and total number of genes with that prefix in the MANE set are shown at
212 right.

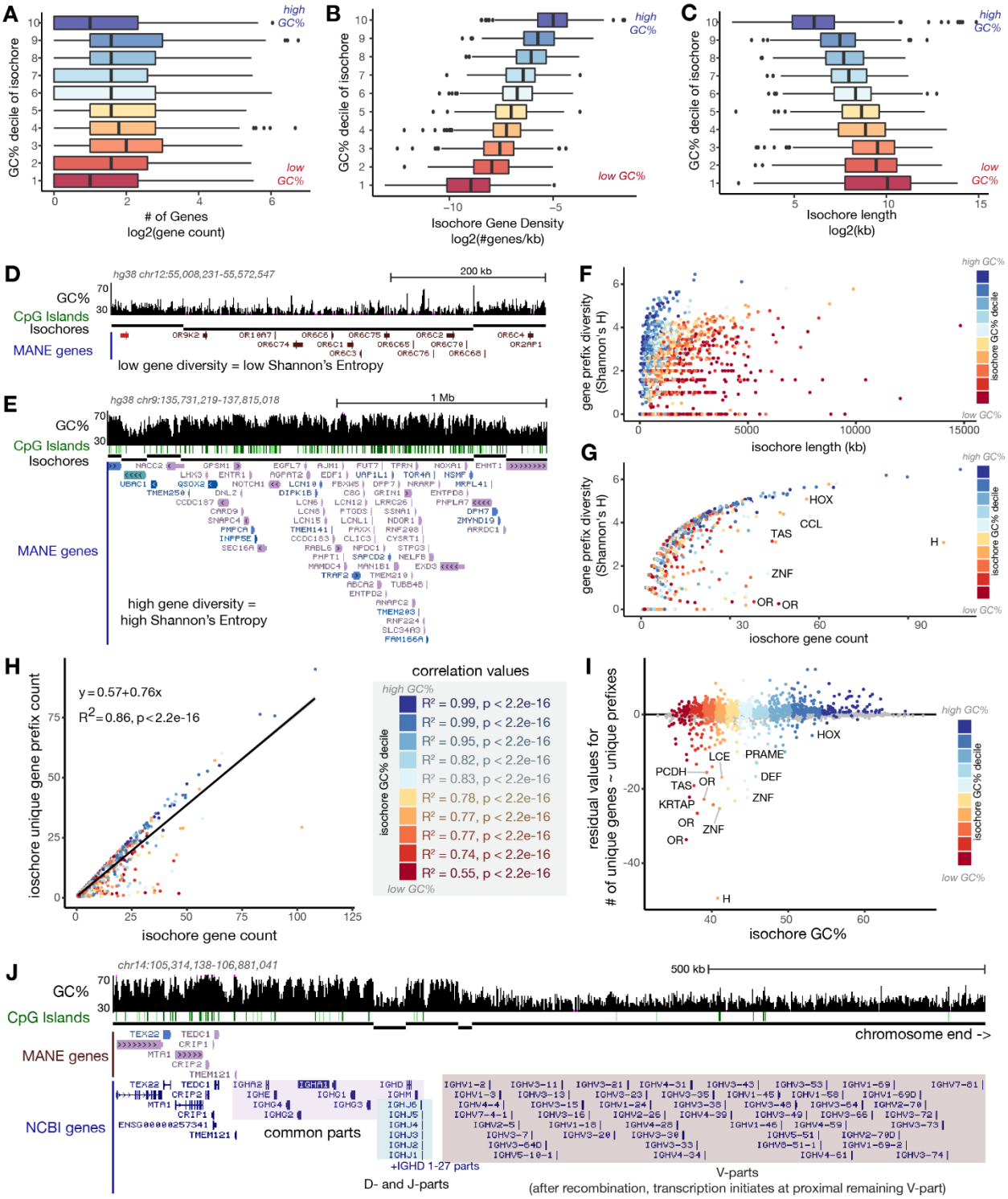

Figure S2 Characterization of human genomic isochores. (A-C) Box and whisker plots describe the relationship between isochore GC% and gene number (A), gene density (B) and isochore length (C). Isochores are binned into deciles according to GC content with decile 10 representing high GC and decile 1 representing high AT. (D,E) UCSC Genome Browser screenshots showing additional example isochores. Gene models are colored according to k-means clusters described below. (D) 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. (E) This isochore is GC-rich and contains genes from a variety of families, so diversity is high. (F,G) Relationship between isochore length (F) and gene count (G) and gene prefix diversity. Isochores are labeled by most common occurring gene prefix. (H) Full graph in Figure 2E including all gene-containing isochores. Spearmann correlation values per isochore deciles are shown on the right. (I) Residuals (i.e. distance of each point from regression line) of graphs in Figure 2E, S2H. (J) 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 a single 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 are enriched in AT-rich regions of the human genome, as they are in mouse (13). We sought next to ask whether the regulatory and transcribed regions of genes, which occupy a fraction of genome space, also vary in GC content across different kinds of genes, and whether GC content of local gene features follow that of the isochore context. The null expectation of this analysis is different from a statistical genetic versus a molecular point of view. From a statistical standpoint, the expectation would be that GC% would co-vary between gene parts (e.g, flanking, promoter, coding sequence) and their isochore context. From a molecular point of view, transcribed units and regulatory regions would be expected to have a GC content aligned with their regulatory or protein-coding function, and would not necessarily match their genomic location. 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 (51,53). Our analysis here includes introns, but clustering on exons and flanking regions produced similar results (Figure S3B). We used iterative k-means clustering to group the 18,640 MANE genes into 3x3 sets (Figure 3A, S3A). The top-level clusters (1-, 2-, 3-) reflect overall differences in AT content in different genes (Figure S3A), while the subclusters (1.1, 1.2, 1.3 etc, Figure 3A) 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 Tables 4 and 5.

The power of this approach is that it captures patterns of feature GC% in relation to one another: For most gene categories, a sharp rise in GC content marks the approach of the TSS, while the transcribed region and the region 3' of the TES share lower GC content. This GC rise at the TSS clearly corresponds to the promoter. In this context, the paltry GC enrichment at the promoters of genes in cluster 3.3 (and to a lesser extent 3.1) is extremely stark (Figure 3A).

Based on high-confidence annotation of transcription start sites, we showed previously that mouse olfactory receptor promoters share this GC-poor pattern (13). At that time, the TSS's of other highly tissue-specific genes had not been mapped. Current human annotations in the MANE set are high-confidence, curated gene models. Nevertheless, many of the genes in cluster 3.3 are extremely tissue-specific (see below) and have less supporting mRNA data than more widely expressed genes. We identified 240 genes in the MANE set (1.3%) that lack annotated 5' UTRs, i.e. where the annotated transcription start site and translation start site are the same. While these were indeed mostly contained in cluster 3.3 (150 of 1636 genes in 3.3, 9%), there

266 was no difference in promoter GC content between these suspect gene models and other genes in
267 cluster 3.3 (data not shown). Indeed, our manual inspection of available RNA data for a subset of
268 these genes suggests that the transcription start sites are correct, but that translation likely
269 initiates at a downstream ATG.
270 We next examined how a gene's isochore GC context relates to the sequence content of
271 its promoter and coding region. We found that patterns of local GC content of genes predicted
272 the GC content of their home isochore, consistent with the statistical genetic null hypothesis, but
273 very surprising considering the functional implications (Figure 3B). The GC-content of a gene
274 with its flanking regions (gene extent with 25kb on each side) correlated closely with the GC
275 content of its whole isochore (Figure S3C). Individually, promoter and coding sequence GC%
276 were also positively correlated with isochore sequence content, but the correlation coefficients
277 were weaker: A subset of genes in AT-rich isochores have GC-rich promoters or GC-rich coding
278 sequences (Figure S3D, E).
279 Genes in clusters 3.1 and 3.3, lacking GC enrichment in their promoters, were highly
280 enriched for the same functional categories as were genes in AT-rich isochores: chemosensation,
281 xenobiosis, and defense/barriers. Indeed, as can be seen in Figure 1A, B and Figure S2D,
282 sometimes entire arrays were members of cluster 3.3 (brown color). To systematically test this,
283 we plotted the promoter GC content distribution of genes in categories we term "outward-
284 looking" (chemosensation, defense, xenobiosis, barriers) versus "inward-looking" (e.g.
285 transcription, kinase function, morphogens). Outward-looking genes have AT-rich promoters
286 while inward-looking genes have GC-rich or average promoters (Figure 3C). We manually
287 annotated common prefixes and enrichment of genes in cluster 3.3 (Figure 3D): This group
288 included all the chemosensory families, many sets of digestive and detoxifying enzymes, and
289 several receptor arrays in the immune system and skin. It also included clustered protocadherins,
290 which share transcriptional regulation patterns with chemosensors. In accordance with the
291 preponderance of tandemly arrayed genes found in cluster 3.3, we found that genes in this cluster
292 were housed in fewer unique isochores and had lower name diversity than those in the other k-
293 means clusters (Figure 3E, S3C).
294 Finally, we asked whether being located near paralogues could predict local sequence
295 features (Figure 3F). Indeed, genes near 1-4 neighbors with the same prefix had more AT-rich
296 promoters than genes not located near paralogous genes, and genes with more than four same-
297 prefix neighbors had AT-elevated promoters, coding regions, and flanking regions compared to
298 both genes in small clusters and singletons. The striking coordination of isochore, promoter, and
299 coding sequence AT content in tandemly arrayed "outward-looking" gene families prompted us
300 to investigate how this pattern relates to the evolution of tandem arrays.

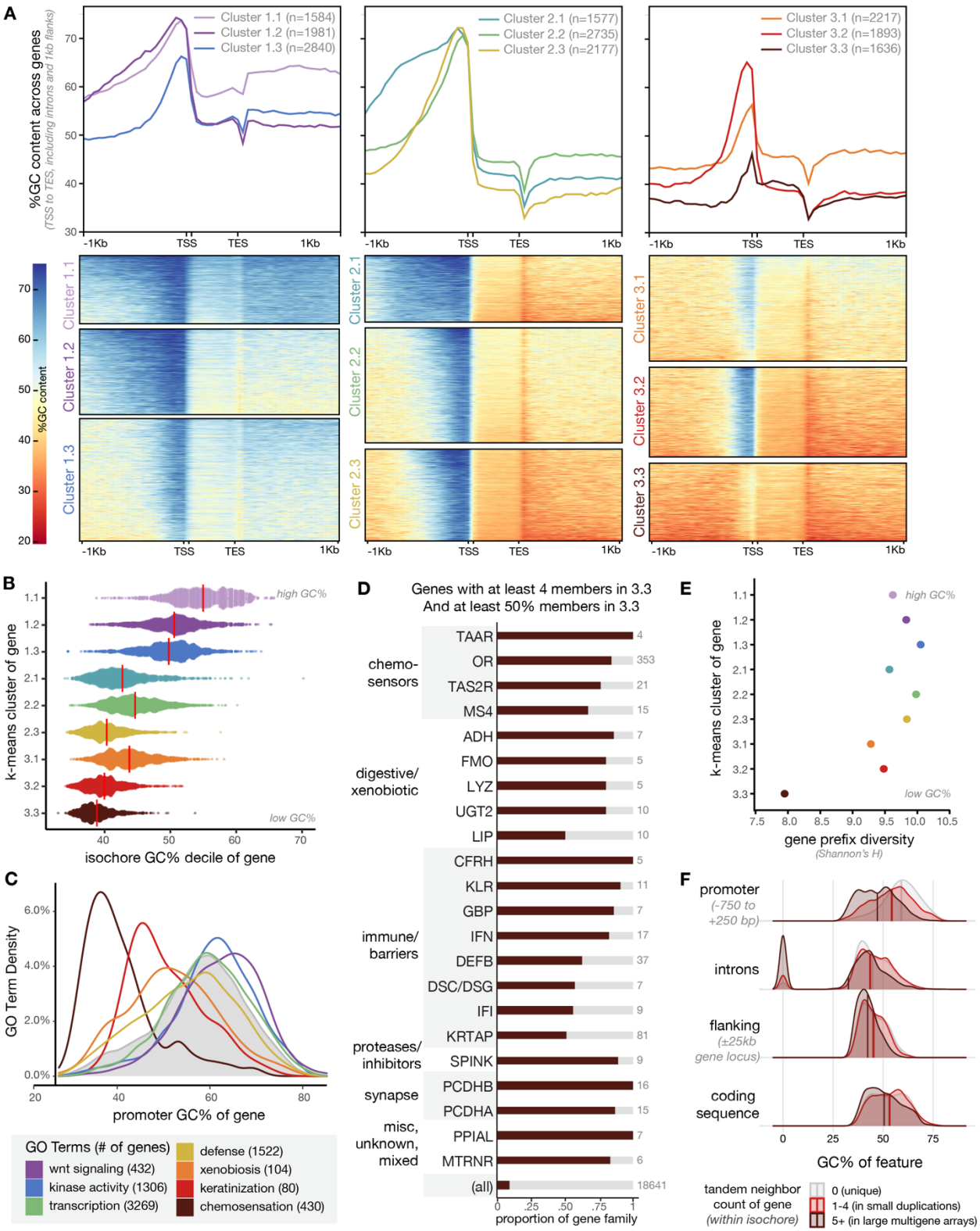

Figure 3: Genes with outward-looking functions have high local AT content. (A) GC content trajectory for human protein-coding genes in the MANE set. Genes were subdivided by iterative k-means clustering. At top, the average GC content trajectory for each k-means cluster is shown

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 immune, barrier, chemosensory, and xenobiotic functions have AT-skewed promoters. Genes with developmental and intracellular functions have GC-skewed promoters. 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. (F) Distribution of promoter, intron, flanking sequence, and coding sequence GC% across genes which do not have paralogous neighbors in their home isochore (gray, genes with 0 tandem neighbors), genes which exist in small local duplications (red, genes with 1-4 tandem neighbors), and genes which exist in large multigene arrays (brown, genes with 5 or more tandem neighbors). A subset of genes without introns appear as 0's.

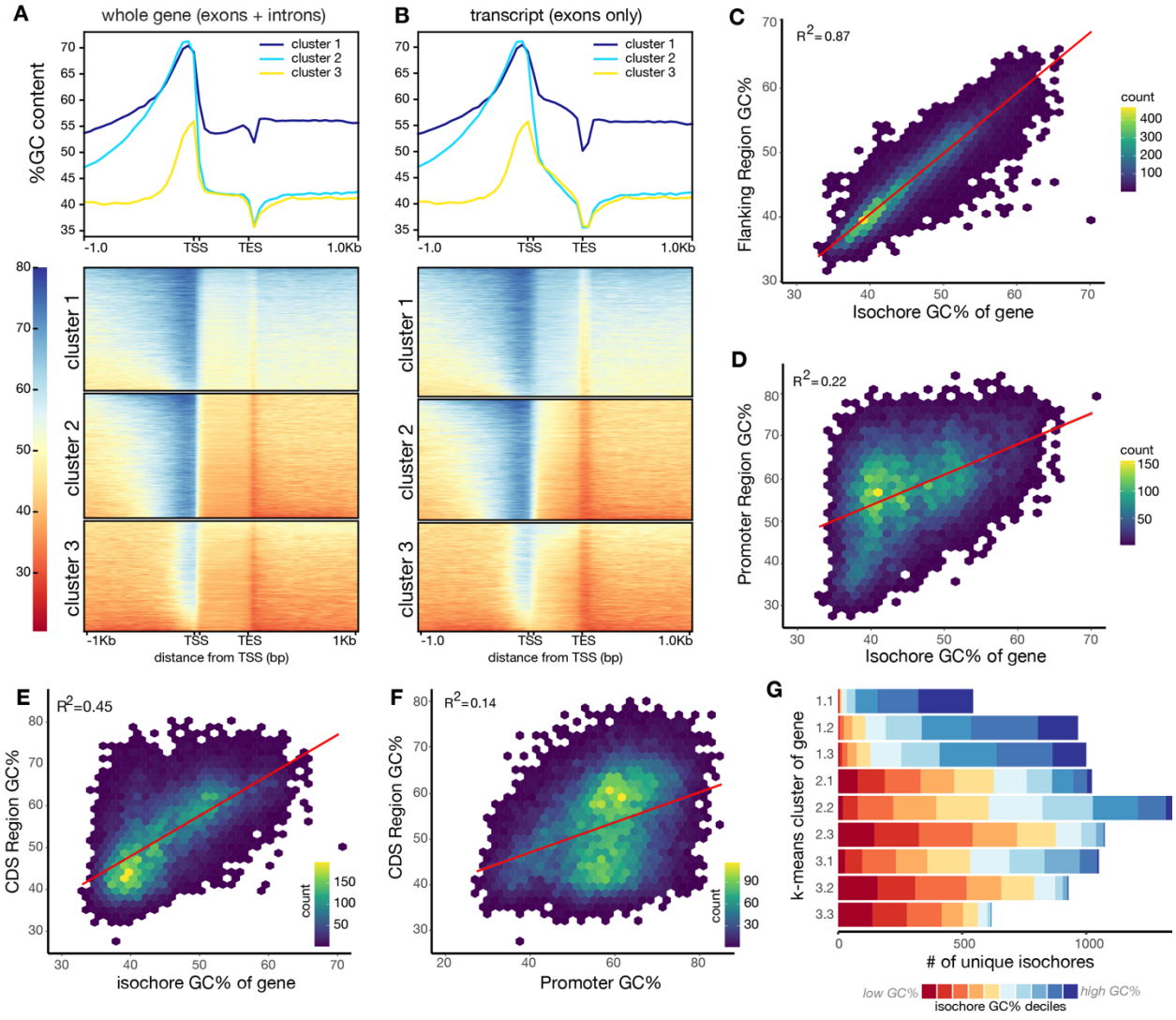

Figure S3: Relationship between GC content of local gene features and the home isochore. (A-B) Heatmaps displaying GC% over 50bp sliding windows calculated across genes. Graphs in (A) include introns, graphs in (B) exclude introns. Clusters were determined by k-means. Summary line plots depict mean GC% across clusters. <10 genes switched clusters when introns were excluded. (C-F) 2D histograms depicting the relationship between (C) flanking, (D) promoter, (E) coding region, and GC percent of "home" isochore for each gene in the MANE set. (F) Relationship between promoter GC percent and coding sequence GC percent. "Flanking region" spans from 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.

Increasing AT content during evolutionary expansion of tandem arrays

As shown above, we find that both the regional and local AT content is high in tandemly arrayed gene clusters in human. AT bias could have emerged as gene families expanded or could have been pre-existing and supported molecular mechanisms of gene duplication. To ask whether AT content rises with the number of paralogous genes in a tandem array, we sought to track copy-number-variable tandem arrays across mammals. Assessing the evolution of tandem arrays is difficult, as large proportions of local paralogues can be species-specific duplications (Figure S4C and (55–58)). Therefore, instead of seeking to identify true homologous genes across different mammalian species, we used synteny analysis to follow whole arrays over evolutionary time. We searched for large gene arrays in human that could be identified across diverse mammals using microsynteny, where conserved heterologous genes serve as "bookends" bounding the ends of the paralgous tandem array.

While copy-number-invariable gene arrays, such as the HOX clusters, were easy to track across all mammals, as expected, the copy-number-variable gene arrays that would be informative for this analysis showed frequent assembly errors, micro-inversions, and array invasions by other heterologous genes. For example, while we observed that the expanded KLR array of NK cell receptors in the fruit bat *Rousettus aegyptiacus* and pheromone-associated MUP array in *Mus musculus* each had sharply higher AT content than surrounding genomic regions, each array had a large assembly gap (9,59). While this limited the number (and, indeed, the extremity) of arrays that we were able to track, we identified six copy-number-variable arrays that were suitable for analysis: An OR array adjacent to the hemoglobin beta genes (60); the Cyp2ABGFST cluster of xenobiotic enzymes; the SERPINA cluster of defensive protease inhibitors; the "epidermal differentiation complex" (EDC), containing skin proteins in the LCE, S100, and SPRR families; and two clusters of Keratin-associated proteins (KRTAPs) which form epidermal appendages such as hair, nails, and claws. These six clusters have diverse median GC content as can be seen in Figure 4A-H. We also included the copy-number-invariable HOXA cluster for comparison.

For each of these clusters, we defined bookend genes as indicators of synteny that must be present for a particular species to be included in analysis. In mammals for which we could find these bookends on the same scaffold or chromosome, we computed the number of genes between bookends, spot-checking species whose counts were markedly higher than those of other species. The number of paralogues in each cluster is plotted across the mammalian clade in Figure S4C, showing the frequency of lineage-specific cluster expansions and contractions. Our full annotations of the contents of these clusters are provided in supplemental Table 6.

368 For each cluster in each species, we calculated the GC% between the bookend genes;
369 relationships between GC% and gene count appeared to be log-linear. We performed
370 phylogenetic least squares (PGLS) analysis to test for a relationship between GC content and
371 log(paralog number) while controlling for phylogenetic relatedness in the multi-species dataset
372 (Figure 4A-H). For the OR, EDC, and SERPINA clusters, AT content rose significantly with
373 paralogue number (PGLS $P < 0.05$), suggesting that AT content can rise as the number of
374 paralogs in tandem arrays grows. Certain arrays in certain species will have latent assembly or
375 annotation errors; however, for the HBB-OR array, we manually counted a subset of species in
376 parallel and saw the same trend (Figure S4A). The HOXA cluster showed variation in GC%
377 despite maintaining the same number of paralogs in the cluster over evolutionary time; this
378 variation appears to be predicted by evolutionary variation in the length of the cluster (i.e.
379 number of base pairs between bookends, not shown). We were able to follow the HBB-OR
380 cluster to non-mammalian amniotes, and found a sharp rise in GC content as the number of ORs
381 in the region melted to 0 (Figure S4A, B).
382 The raw GC% of each cluster was poorly correlated with variation in genome-wide GC
383 content across species (Figure 4I, J). However, each cluster type maintained a consistent
384 relationship to the overall GC% of the genome: Across diverse mammalian species, the OR
385 cluster and KRTAP cluster were almost always more AT-rich than the genome as a whole, while
386 the other five clusters were almost always more GC-rich than the genome as a whole (not
387 shown).
388 What portions of these genomic regions drive the variation in local GC content over
389 evolutionary time? To assess this, we developed a series of scripts we call TandemClipR (insert
390 github link once available), which divided each cluster in each species into CDS, intron,
391 promoter, and intergenic regions. To avoid the "false-TSS" problem described above, we
392 included only genes with annotated 5' UTRs in our promoter analyses. We found that the GC%
393 of each sub-region was positively correlated with the overall cluster GC% (Figure 4H-J). As we
394 described above, this is expected from a statistical point of view, but surprising from a molecular
395 point of view given that coding regions and promoters perform important molecular work. For
396 example, only some amino acids can be coded with high-GC codons. Variation in intergenic
397 regions contributed the most to cluster GC%, as expected based on their comprising a
398 preponderance of the sequence. Despite varying with cluster GC%, coding regions, introns, and
399 promoters had consistently higher GC content than the cluster as a whole, consistent with their
400 functional constraints (Figure 4H).
401 Finally, we asked how the GC% of sub-regions of the cluster were related to the number
402 of paralogues in the cluster (Figure S4D,E). For the clusters whose AT content rose with local
403 paralogue number, the AT content of each substituent portion of the sequence (promoters, coding
404 regions, introns, intergenic regions) also rose. The particular feature that correlated best with
405 paralogue number varied across gene families, and no individual feature was consistently more
406 correlated with paralog number than GC% of the whole cluster. In sum, we find that cluster AT
407 content can rise as paralogous clusters bloom over evolutionary time; remarkably, these trends
408 affect all the sequence components of the cluster, suggesting neutral or selective mechanisms
409 that act cluster-wide. We hypothesize that local gene contents, particularly paralogous clusters of
410 genes, can influence the emergence of isochores differing in GC content. Next, we provide a
411 model for how this could have come about.

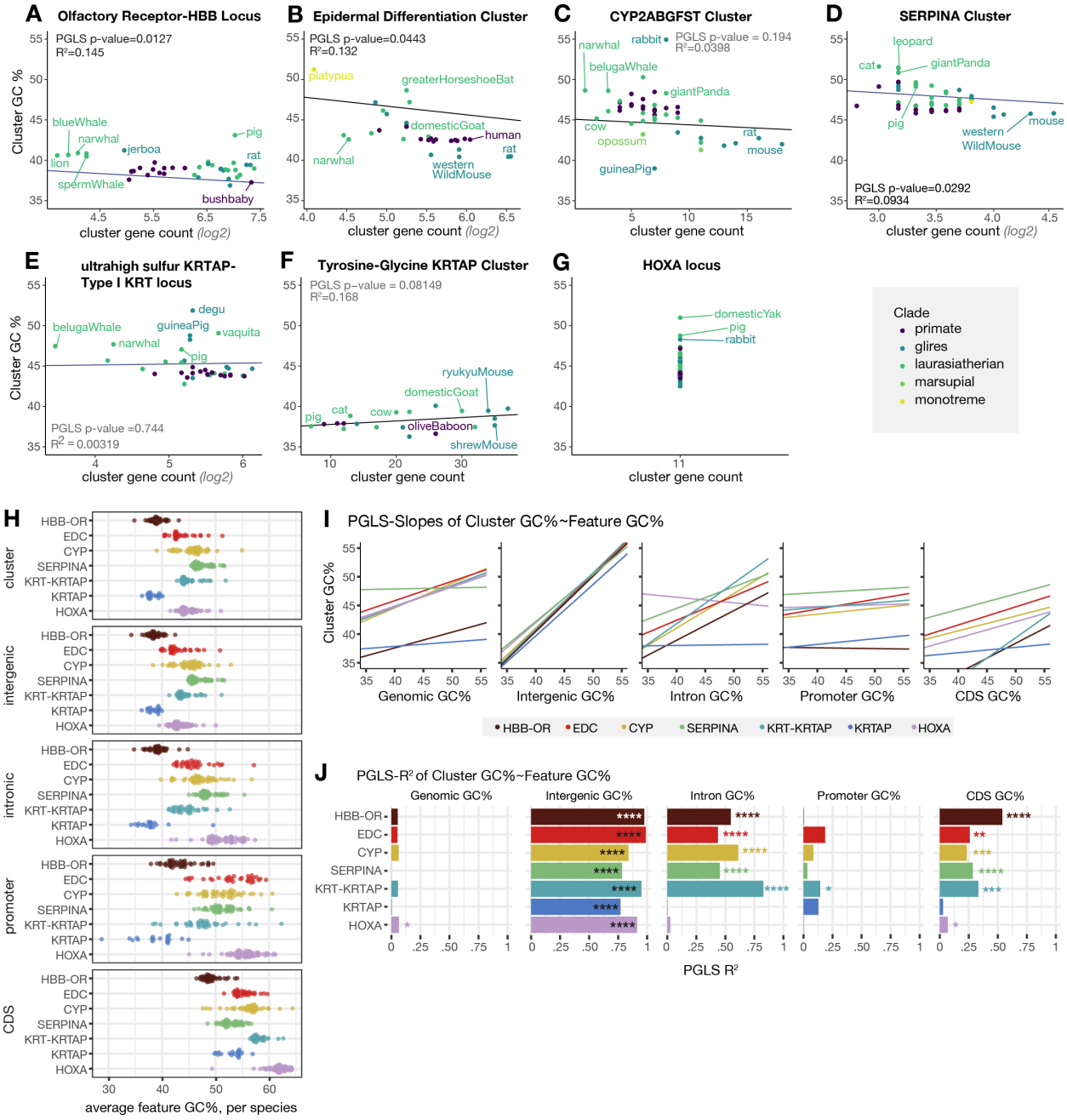

Figure 4: AT content is correlated with tandem array expansion. (A-G) Comparison of number of (A) OR genes in the Hemoglobin β (*HBB*) cluster, (B) LCE, SP100, and SRR genes in the Epidermal Differentiation cluster, (C) CYP2A genes in the CYP2A-GFST cluster, (D) SERPINA genes, (E) KRTAPs and KRTs in the Type I KRT locus, (F) KRTAPs in the high tyrosine-glycine KRTAP cluster, and (G) the HOXA locus with cluster GC% across mammals. All p-values report results of phylogenetic least squares analysis (PGLS). (H) Average GC% across the cluster, intergenic, intronic, promoter, and coding sequence of each tandem array. Each point represents one mammal species from panels A-G. (I) PGLS slope values and (J) R-squared values of phylogeny-corrected correlation of feature GC% to cluster GC%. Asterixes represent magnitude of PGLS p-values for each relationship, i.e. one asterixes represents a p-value between 0.05 and 0.01, two represent a p-value between 0.01 and 0.001.

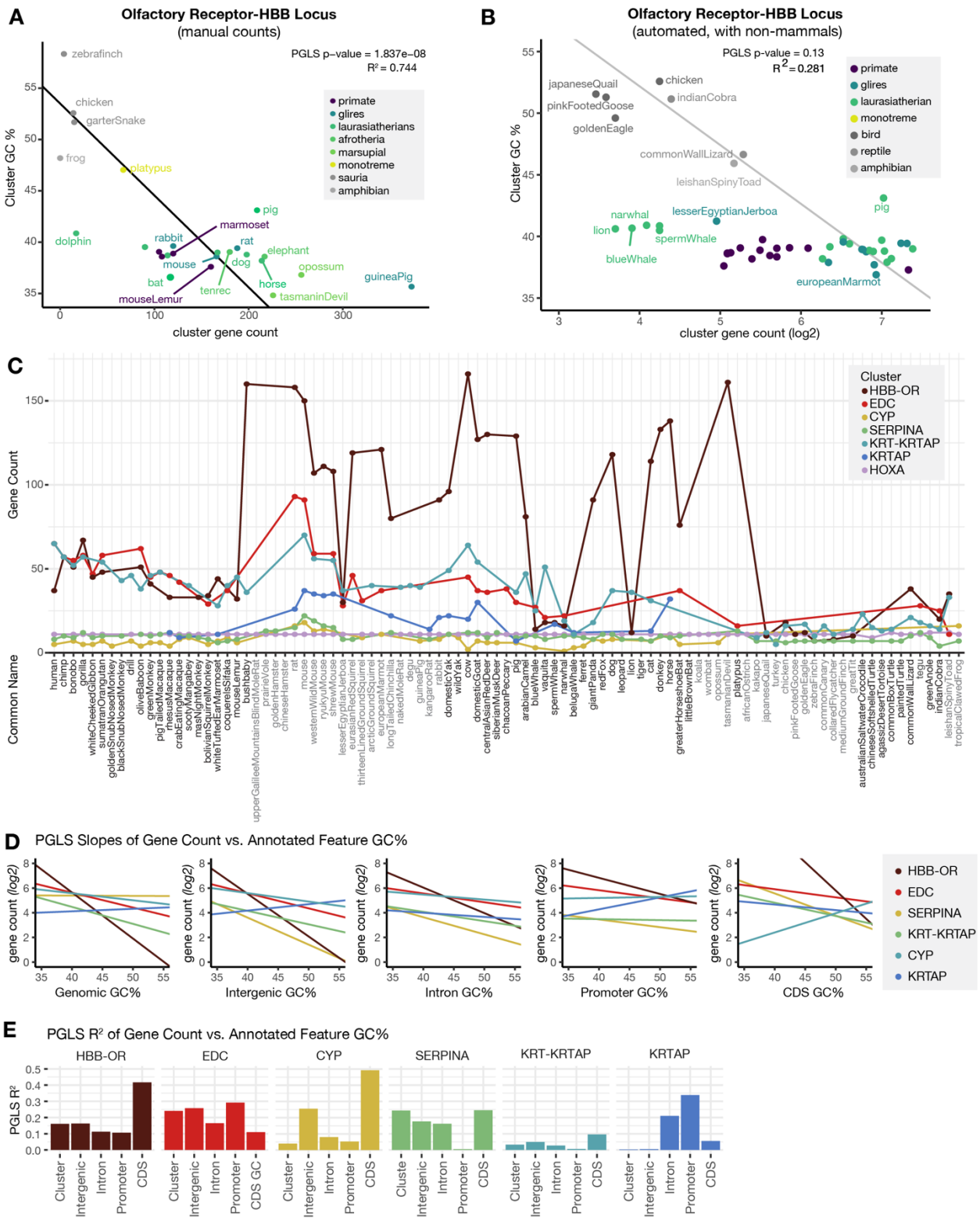

Figure S4: Data related to Figure 4 (A) Counts of gene number for each paralogous cluster across amniotes analyzed (B) UCSC Genome Browser screenshot showing the hemoglobin β cluster on human chromosome 11. As described previously, the *HBB* cluster is flanked by olfactory receptor genes and these are bracketed by conserved single-copy genes, including

429 *STIM1*, *RRM1* and *FHIP1B* shown here, that are syntenic with the *HBB* cluster since before
430 mammals branched (60). (C) Analysis of gene number and GC% in the HBB-OR cluster across
431 mammals, reptiles, and birds. (D) R^2 values for the relationships between the GC% of each
432 cluster and the GC% of its constituent parts across mammalian species. (E) R^2 values for the
433 relationship between cluster or constituent part GC and the number of genes in the cluster across
434 mammalian species.
435
436 *Accumulation of Functional Diversity and Divergence*
437 As we show above, many types of outward-facing genes have extreme variation in
438 paralogue number across mammalian species (Figure S4C). Anecdotally, genes in these families
439 also exhibit extreme allelic diversity and copy number variations among humans, and
440 polymorphisms in these genes underlie human phenotypic variation in drug metabolism, sensory
441 perception, and immune response (1,11,55,61–66). Colloquially, outward-looking genes are so
442 diverse in copy number and sequence that a first step in GWAS is often to "throw out the ORs."
443 Previous reports, including ours, have speculated that partitioning inward- and outward-
444 looking genes into different parts of the genome could enable a higher ongoing mutation rate in
445 AT-rich, outward-looking genes; however, point mutation rate is in general positively correlated
446 with GC content because cytosines are especially mutation-prone (13,42,67,68). In the first
447 model, AT content would facilitate mutagenesis and diversification of genes in these families. In
448 the second model, relatively high point mutation drift in large gene families could have
449 contributed to their shift to higher AT content over evolutionary time.
450 To test between these possibilities, we first systematically examined the degree of coding
451 sequence variation in human genes grouped by AT/GC content or by degree of local tandem
452 gene duplication, using the gnomAD dataset of rare single nucleotide variants ascertained from
453 whole exome sequencing of >100,000 unrelated people (gnomAD v2.1.1, Figure 5A, B) (69).
454 The ratio of protein-altering versus synonymous variants is positively correlated with AT
455 content, with genes in AT-rich isochores and k-means cluster 3.3 highly enriched for potentially
456 functional variants (Figure 5A,B). Moreover, genes near paralogous neighbors had higher levels
457 of functional variants (Figure 5E,F). We note that use of rare variants profoundly understates the
458 allelic variety in outward-looking genes, which exhibit radical common variation and high rates
459 of copy number and structural variation. For example, any two humans are estimated to have
460 function-changing variation in 30% of their olfactory receptor genes (70,71).
461 Does the high polymorphism of genes in outward-looking tandem arrays result from
462 differential mutation or selection versus inward-looking genes? To examine mutations directly,
463 we used synonymous variants from gnomAD as a proxy. We see that AT-rich genes have *fewer*
464 synonymous variants across unrelated people than do GC-rich genes (Figure 5C, D, S5A-D).
465 This is consistent with point mutation rate being grossly driven by deamination of cytosine,
466 especially in the $C^{me}G$ context: AT-biased genes have essentially run low on CpG dinucleotides
467 to mutate (32–36,72,73). Therefore, we reject the hypothesis that AT-biased genes are
468 "prioritized for mutation;" rather, they are enriched for functional diversity relative to a low
469 ongoing mutation rate. We observe that olfactory receptor genes have higher rates of variant
470 calls than do other AT-rich genes (Figure S5F). This inflated rate could reflect an unknown
471 mutagenic process but could also result from incomplete knowledge of the full human "OR-
472 ome" and incorrect assignment of homology relationships.
473 The enrichment of functional variation in outward-looking genes relative to low levels of
474 synonymous variation could result from different selective processes acting on single- versus

475 multi-copy genes. Single-copy genes are uniquely responsible for a particular biological process,
476 while tandem arrays may distribute a particular process over a large set of family members (i.e.
477 subfunctionalization). Deleterious mutations to single members of such large gene families are
478 therefore more likely to be of scant phenotypic consequence. Indeed, we find that genes with
479 more tandem neighbors have higher rates of function-changing versus synonymous variants
480 (Figure 5E,F, S5E). To test how local and isochore GC content relate to the functional
481 dispensability of genes, we used the gnomAD metric of "loss-of-function intolerance" (pLI):
482 genes have high pLI if deleterious mutations are depleted from the population (69,74). Genes
483 predicted to be loss-of-function intolerant were enriched in GC-rich isochores, had GC-rich
484 promoters, and were absent from cluster 3.3 (Figure 5G,H, S5G); AT-skewed, outward-looking
485 genes had lower probabilities of being loss-of-function intolerant, i.e. they are relatively
486 dispensable.
487 To examine modern mutagenesis patterns in a different way, we sought to measure the
488 rate of *de novo* mutations that occur in genes in AT- versus GC-biased regions of the human
489 genome. Whole genome sequencing of two parents and a child (trios) enables detection of *de*
490 *novo* mutations (DNMs). We used a recent dataset that compiles DNMs from >11,000 trios,
491 nearly all those who have been sequenced to date (75). While these data remain sparse relative to
492 the size of the genome (~700,000 total DNMs), DNMs approximate a record of mutagenesis that
493 has yet to be operated on by external selection. Pooling DNMs across each isochore and across
494 transcriptional units (TSS to TES) within that isochore, we found that both genic and isochore-
495 wide DNMs were more common in higher-GC isochores (Figure 5I,J, S5H-I). This is consistent
496 with sequence-based predictions, prior findings, and our analysis of gnomAD synonymous
497 variants within genes (76,77). Together, comparison of *de novo* mutations in meiosis and single
498 nucleotide variants in unrelated humans both support the conclusion that despite their high
499 divergence and diversity, AT-biased genes experience fewer contemporary mutations than do
500 GC-biased genes, allowing us to reject the hypothesis that quarantining outward-looking genes in
501 AT-rich regions of the genome facilitates higher mutation rates. Instead, lack of purifying
502 selection on outward-looking gene blooms could be one factor that increases their AT content.

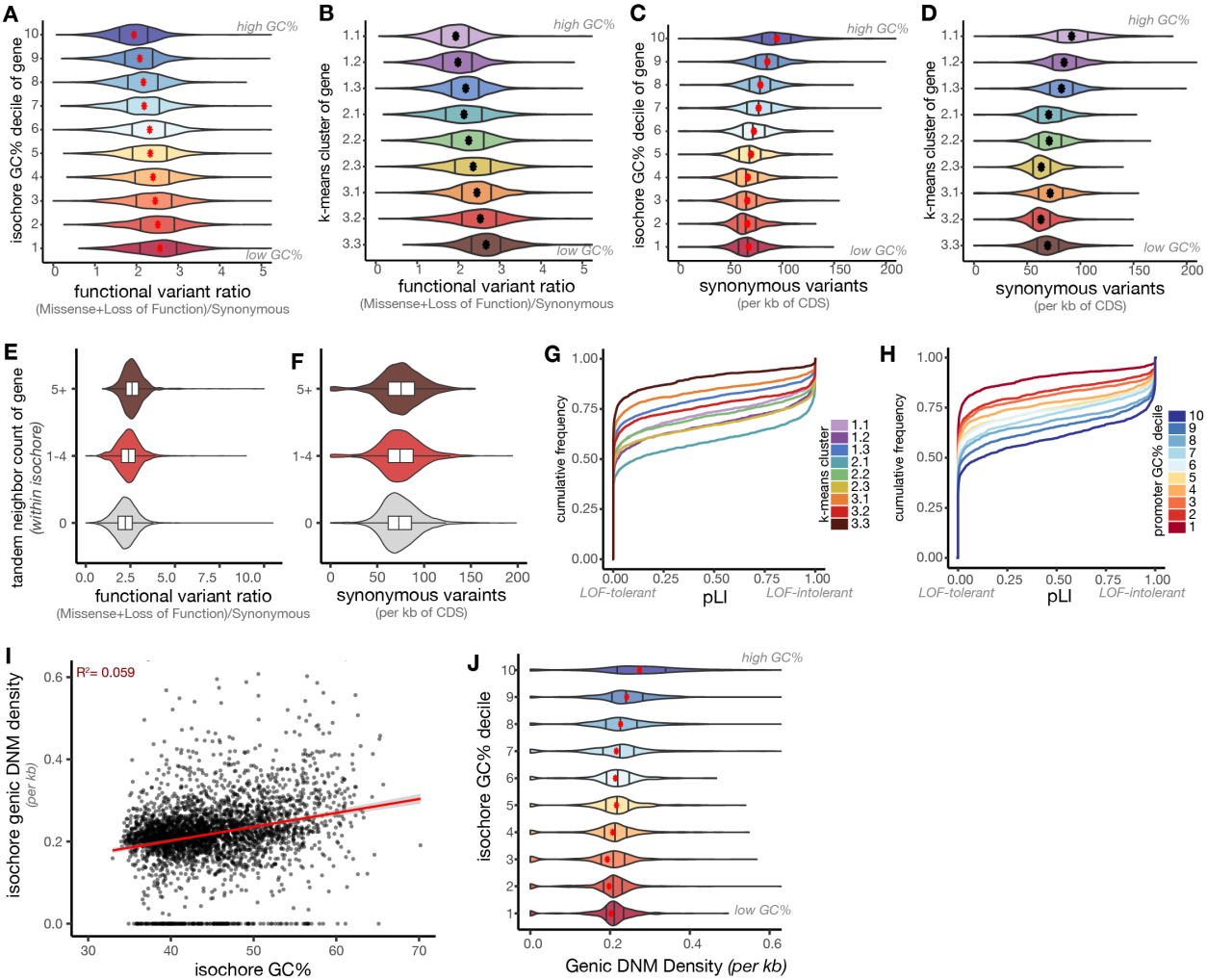

Figure 5: AT-rich genes have high functional diversity despite experiencing moderate mutation rates in the present. (A,B) Ratio of functional (missense plus loss of function) versus synonymous rare variants in the MANE gene set identified in gnomAD v2.1.1 exome sequencing of >100,000 unrelated individuals (69). 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,F) Functional (E) and synonymous (F) variant rates from gnomAD v2.1.1 for genes with or without paralogous neighbors. Box plots show median and mid-quartile distribution. (G,H) Cumulative frequency distribution plots of gnomAD pLI (likelihood that a gene is loss-of-function intolerant) relative to a gene's k-means cluster assignment (G) or promoter GC% (H) (69). (I,J) Number of *de novo* point mutations observed per kb across the genes (TSS to TES) within an isochore relative to isochore GC% (I) and isochore GC% binned by decile (J). ~700,000 DNM calls are pooled from all ~11,000 trios sequenced to date (75).

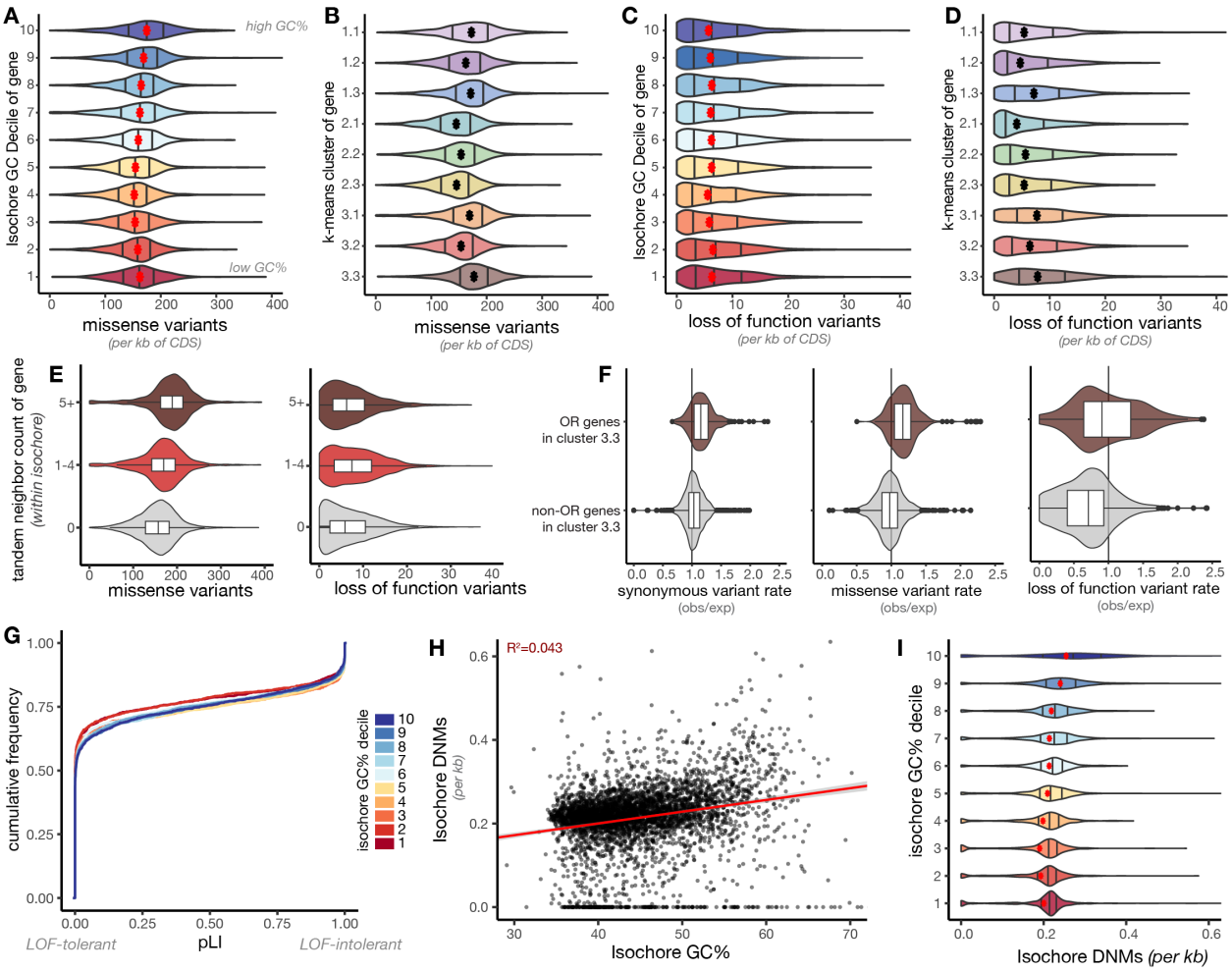

Figure S5: Raw rates of loss of function and missense variation. Raw gnomAD v2.1.1 counts of missense (A, B) and loss-of-function (C, D) SNVs across MANE genes binned by isochore GC% decile (A, C) or k-means cluster (B, D). (E) Missense and loss of function mutation accumulation in genes of differing degrees of local tandem duplication. (F) Ratio of observed over expected rates of synonymous variants, missense variants, and loss of function variants of genes in k-means cluster 3.3, split by OR and non-OR genes. Observed/expected values are calculated relative to the neutral mutation spectrum by gnomAD. (G) Cumulative frequency of pLI scores for genes binned by their home isochore GC% decile. (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 (75).

Recombination and PRDM9 Binding

Recombination may be the primary influence on large-scale patterns of AT/GC content due to GC-biased gene conversion, which elevates GC content. AT-biased chromosomal bands are described as experiencing less recombination than GC-biased bands (20,26,30). We sought to test how recombination rates relate to GC content of inward- versus outward-looking genes. We therefore used whole genome sequencing data from human trios to measure recombination in isochores of different GC contents and in their constituent genes (78).

538 Crossovers appeared rare within gene blooms (Figure 6A, S6B). We calculated a relative
539 crossover rate for each isochore and found that AT-rich isochores experienced less maternal and
540 paternal crossovers than GC-rich isochores, as has been previously observed, though maternal
541 crossovers were sharply diminished in the highest-GC isochores (Figure 6C,D) (15,78–80). We
542 noticed that these low-crossover, high-GC isochores were often at chromosome ends, where
543 maternal recombination has been shown to be low (81). To systematically examine
544 recombination relative to each gene along the chromosome, we generated a Manhattan plot of
545 crossover rate for each gene and its flanking regions (Figure 6B, S6A). This highlights the higher
546 recombination of genes located in GC-rich isochores, except for maternal recombination at
547 chromosome ends.
548 Recombination in paralogous clusters can induce dramatic insertions, deletions, and
549 chromosome rearrangements if paralogues errantly pair with one another (82). While clusters of
550 outward-looking paralogues are enriched in AT-rich regions, the conserved clusters of
551 paralogues located in GC-rich regions, such as Hox clusters, may also be risky to recombine. To
552 separate the influence of AT/GC content versus clustered-ness, we compared recombination rates
553 across clusters of different sizes (Figure 6F). However, most large clusters of genes were AT-
554 rich and had little recombination, making it difficult for us to separate which variable dominates.
555 While recombination is generally directed away from genes, some genes experienced crossovers.
556 We plotted the isochore and k-means cluster distribution of genes with the 10% highest internal
557 crossover rate (TSS-TES, Figure S6D, E). These genes were in GC-rich isochores and excluded
558 from AT-rich k-means cluster 3.3. Together, these analyses suggest that gene blooms located in
559 AT-rich regions of the genome experience low current crossover rates.
560 In many vertebrates, including humans, crossovers are seeded by binding of PRDM9 to
561 its target site (83–85). To test whether variation in observed recombination across AT/GC
562 categories is due to differential seeding, we examined PRDM9 binding data from human cells
563 (Figure 6E,G, S6F,G)(86). We observed a striking depletion of PRDM9 binding from AT-biased
564 gene clusters and isochores, in line with previous analyses (79). We note that many animals have
565 lost PRDM9, and in the absence of PRDM9, recombination is often seeded at CpG islands (87).
566 As described below, AT-biased gene families also lack CpG islands. Taken together, these
567 results suggest that AT-biased, tandemly arrayed gene clusters experience low rates of
568 recombination, and that this likely results from low rates of crossover initiation.

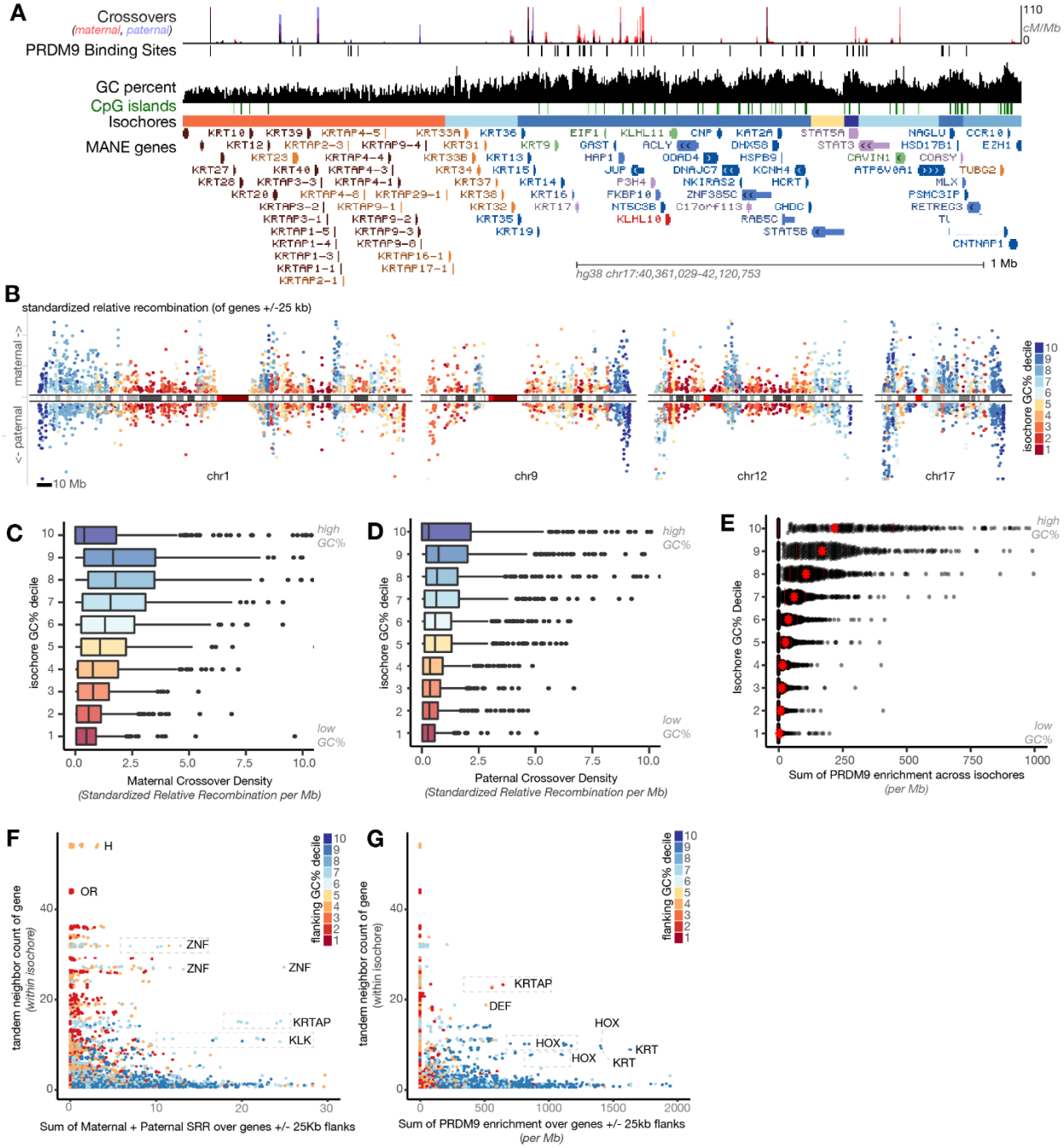

Figure 6: Crossovers are directed away from AT-rich isochores and genes. (A) UCSC genome browser screenshot of human KRTAP cluster showing recombination rates (maternal meiosis in red, paternal meiosis in blue) and PRDM9 binding calls (black). (B) Manhattan plot of standardized maternal (above 0) and 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). Scale bars: 10Mb. Chromosome ideograms reflect centromeres (bright red), gaps (dark red) and Giemsa bands (grays). Recombination rates are low for genes located in AT-rich isochores, except at chromosome ends, which are depleted for maternal recombination. (C) Maternal and (D) paternal crossover rates in isochores binned by GC percent. Crossover calls are from deCODE (78). (E) PRDM9 peak enrichment across isochores binned by

GC%. (F) Scatterplot of tandem neighbor counts per genes within the same isochore versus (F) the sum of maternal and paternal standardized relative recombination and (G) sum PRDM9 enrichment.

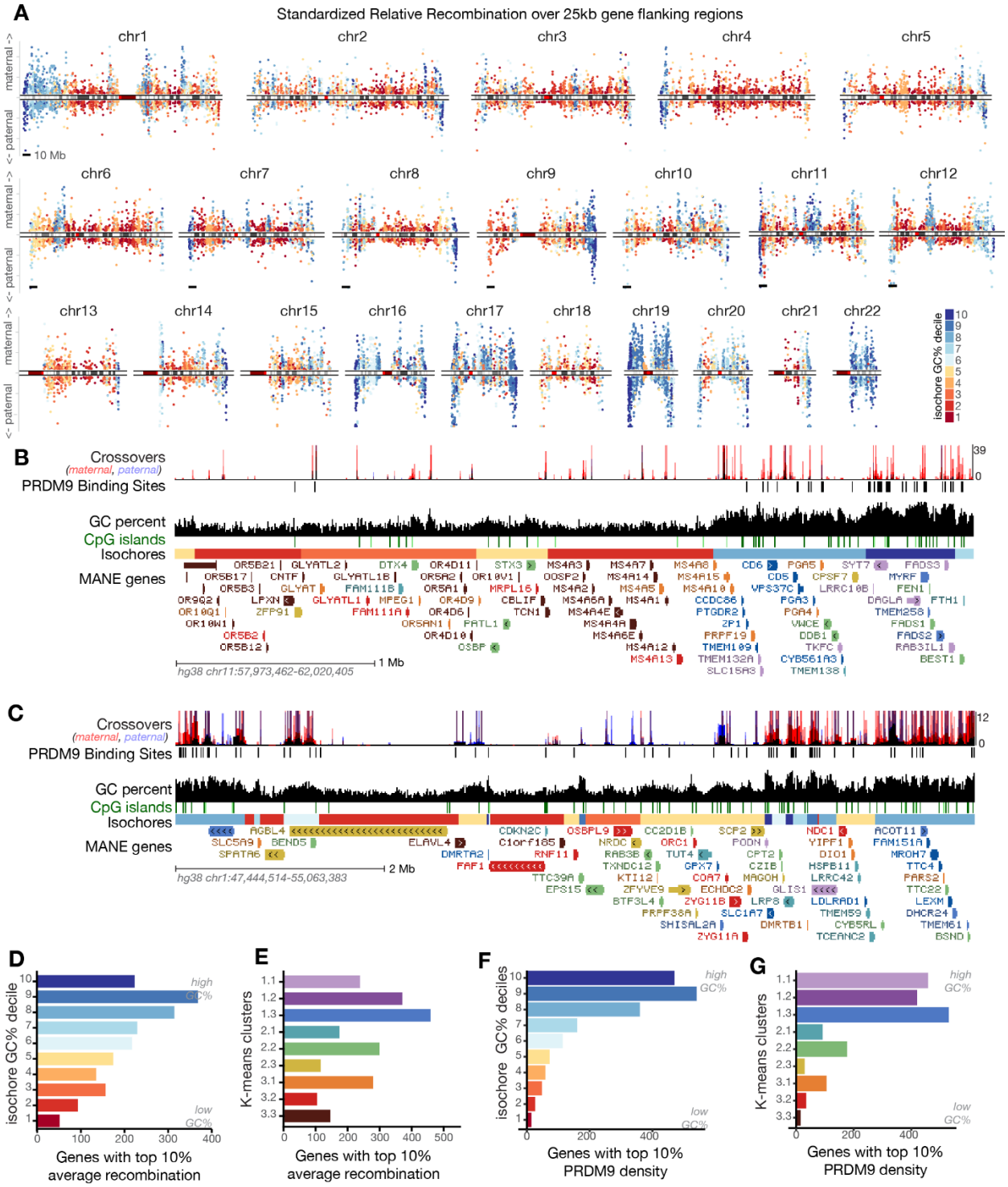

Figure S6: (A) Manhattan plots of maternal and paternal standardized relative recombination rates for all chromosomes, as described in Figure 6B. (B-C) UCSC Genome Browser screenshots of (B) OR and MS4A loci, flanked by GC-rich isochores and (C) interspersed GC and AT-rich

587 isochores. (D,E) Counts of (D) isochore decile and (E) k-means cluster distribution of genes with
588 10% highest rate of maternal and paternal within-gene crossovers. (F,G) Counts of (F) isochore
589 decile and (G) k-means cluster distribution of genes with 10% highest rate of within-gene
590 PRDM9 binding.
591
592 *Transcriptional regulation of outward- versus inward-looking genes*
593 Our analyses above lead to the hypothesis that clusters of paralogues that are not subject
594 to purifying selection (i.e. "outward-looking genes") lose GC content as they bloom due to
595 reduced recombination and increased tolerance of point mutations. We further hypothesize that
596 the sequence content and context of these genes has been evolutionarily co-opted to produce
597 extreme tissue-specificity and/or variegation of their transcription. While we expect that
598 isochore-level AT/GC content influences genomic organization and histone mark flavors, which
599 would influence transcription indirectly, we also noticed that paralogous clusters of genes lacked
600 annotated CpG islands in their promoters (see CpG Island track, Figure 2A, B, Figure S2D)(54).
601 CpG dinucleotides are depleted from vertebrate genomes due to the mutability of methylated
602 cytosine; nevertheless, CpGs are relatively enriched in vertebrate promoters, and these "CpG
603 islands" frequently remain unmethylated (88). Previous analyses suggested that 50-70% of
604 mammalian genes have CpG island promoters (19,89,90). Nevertheless, by considering each
605 gene in its sequence context, we estimate that 90% of protein-coding genes have GC enrichment
606 directly upstream of the TSS (Figure 3A).
607 While CpG islands are calculated relative to the GC content of the region, we sought to
608 ask if it is even *possible* to have a CpG island in an AT-rich region. As in Figure 2, we separated
609 isochores of at least five genes by GC content, gene number, and prefix diversity, and then
610 calculated CpG island scores on a per-gene basis (Figure 7A,B). Surprisingly, genes located in
611 all types of isochores *could* have CpG islands in their promoters, but genes in large paralogous
612 arrays often did not. Highly conserved genes in arrays, like Hox and Histone arrays, retained
613 their islands regardless of the GC content of the isochores. Gene blooms lacking CpG islands
614 tended to be outward-looking, and almost all large, island-poor arrays were in AT-rich isochores.
615 We also noticed that the shape of islands around the TSS varied across our k-means clusters,
616 with some clusters having islands that were symmetrical relative to the TSS, and others having
617 islands that were polarized to the 3' side of the TSS (Figure S7A).
618 Because CpGs are only protected from methylation when they are located near other
619 CpG's, i.e. in islands, and because methylated CpG's are extremely mutation-prone, CpG loss
620 can be precipitous once it begins (32). We calculated observed promoter CpG sites compared to
621 the expected rate based on local sequence content and observed a bimodal distribution across
622 genes of all isochores, though the observed/expected CpG's elsewhere in the isochore were
623 predicted by GC content (Figure 7C,D, Figure S7A,B). Thus, though CpG-less promoters are
624 more common in high-AT isochores, islands are compatible with genes in any isochore context;
625 island loss appears specific to paralogous gene blooms (Figure 7E). Across a range of isochore
626 GC contents, genes near tandem neighbors had CpG-poor promoters compared to singletons
627 (Figure S7E). This promoter state also drives k-means clustering of genes in 3.1 and 3.3 (Figure
628 7D). Many genes in 3.1 and 3.3 have no more CpG's in their promoters than the genome-wide
629 average (~25% of expected CpGs). Further separating these k-means clustered genes into those
630 with paralogous neighbors and those without did not stratify promoter architecture, suggesting
631 that the k-means categories are already sensitive to the AT/GC sequence features delineating
632 genes in paralogous clusters (Figure S7D).

633 As we and others suggested previously in the mouse, CpG-less promoters are likely to be
634 regulated by non-canonical mechanisms that are independent of TATA Binding Protein (TBP)
635 (13,91). This could allow the unique and rare expression of these genes relative to their CpG-
636 containing brethren: their ground state is to be "off forever." While hemoglobin beta and
637 amylase genes, which we now know to be small clusters of paralogues, were shown to lack CpG
638 islands in their promoters in the 1980's, measurements of the relationships between tissue-
639 specificity and promoter architecture were performed before the transcription start sites and
640 promoters of tissue-specific genes were mapped (15,19,92). To test how AT/GC distribution in
641 genes relates to patterns of gene expression, we used GTEx data, which measures gene
642 expression in 54 tissue types taken from adult human donor cadavers (93). We set a threshold
643 (RPKM of 5) to binarize this quantitative data to "expression" or "no expression." In agreement
644 with Bird's hypothesis, this simple metric varies sharply across genes of different k-means
645 clusters or with differing promoter or home isochore GC content (Figure 7F-H)(13,15,19,92).
646 Genes that are GC-rich are often expressed in many or most tissues tested, while AT-biased
647 genes most often appear to be expressed nowhere or in one tissue. We note the many genes with
648 "no" expression in GTEx are specific to tissues not sampled by GTEx (e.g., olfactory receptor
649 genes in the olfactory epithelium). We infer that genes not detected in any tissue in GTEx data—
650 i.e. the preponderance of AT-rich genes—are highly tissue-, cell type-, time-, or condition-
651 dependent in their expression. As we have shown throughout this study, these exotically
652 transcribed, CpG-less genes are located in evolutionary dynamic clusters of paralogues and tend
653 to be housed in AT-rich isochores.
654 In the longest-lived cells in the body, post-mitotic neurons, tandemly arrayed gene
655 families have been shown to be clustered with one another in nuclear space and to be uniquely
656 protected from accumulation of CpH (Cp-nonG) methylation (47,94,95). These findings,
657 together with the overwhelming transcriptional repression of these gene families, suggest that
658 they might be sequestered in nuclear space away from the transcriptional machinery. Indeed, we
659 found that across 21 tissues sampled by Hi-C, AT-rich genes and genes located in AT-rich
660 isochores were likely to be located in transcription-suppressing "B" compartments (Figure S7F,
661 G)(96). Previous analyses have demonstrated that variation in GC content also predicts local
662 chromatin looping and association with the lamina and other nuclear structures (16,39,49,97).

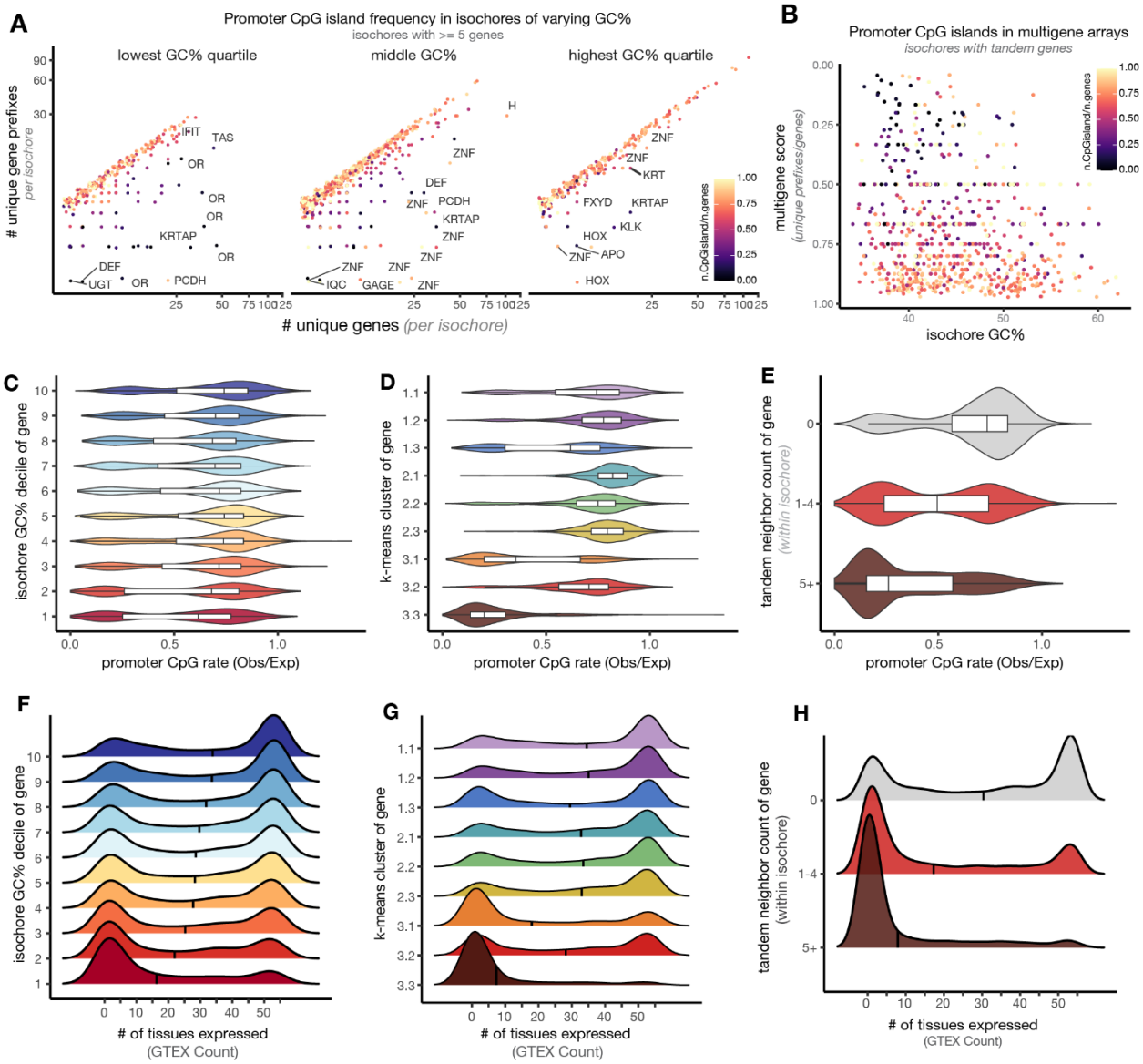

Figure 7: Lack of CpG islands and restricted expression is a feature of tandemly arrayed AT-rich genes. (A) Promoter CpG island frequency (from (54)) in isochores of varying gene content. Same plots as in Figure 1E, recolored to highlight promoter characteristics. Most common gene name prefixes in the isochore are labeled. (B) Multigene score versus isochore GC% for isochores containing tandem duplicated genes. Color represents fraction of genes in the isochore with promoter CpG islands. (C-E) Distribution of promoter CpG rate for genes binned by (C) isochore GC% decile, (D) k-means cluster assignment, or (E) tandem duplication score. We counted CpG's from -750 to +250bp around the gene TSS and compared to the expected rate for a sequence of that GC content. (F-H) Distribution of tissue-level gene expression for genes binned by (F) isochore GC% decile, (G) their k-means cluster assignment, or (H) tandem duplication score. GTEx data from 54 human tissues was binarized to "expression" or "no expression" based on RPKM of 5. AT-rich genes and those in paralogous arrays are detected in few or no sampled GTEx tissues, while GC-rich genes are frequently detected in all sampled tissues. Black bars depict medians.

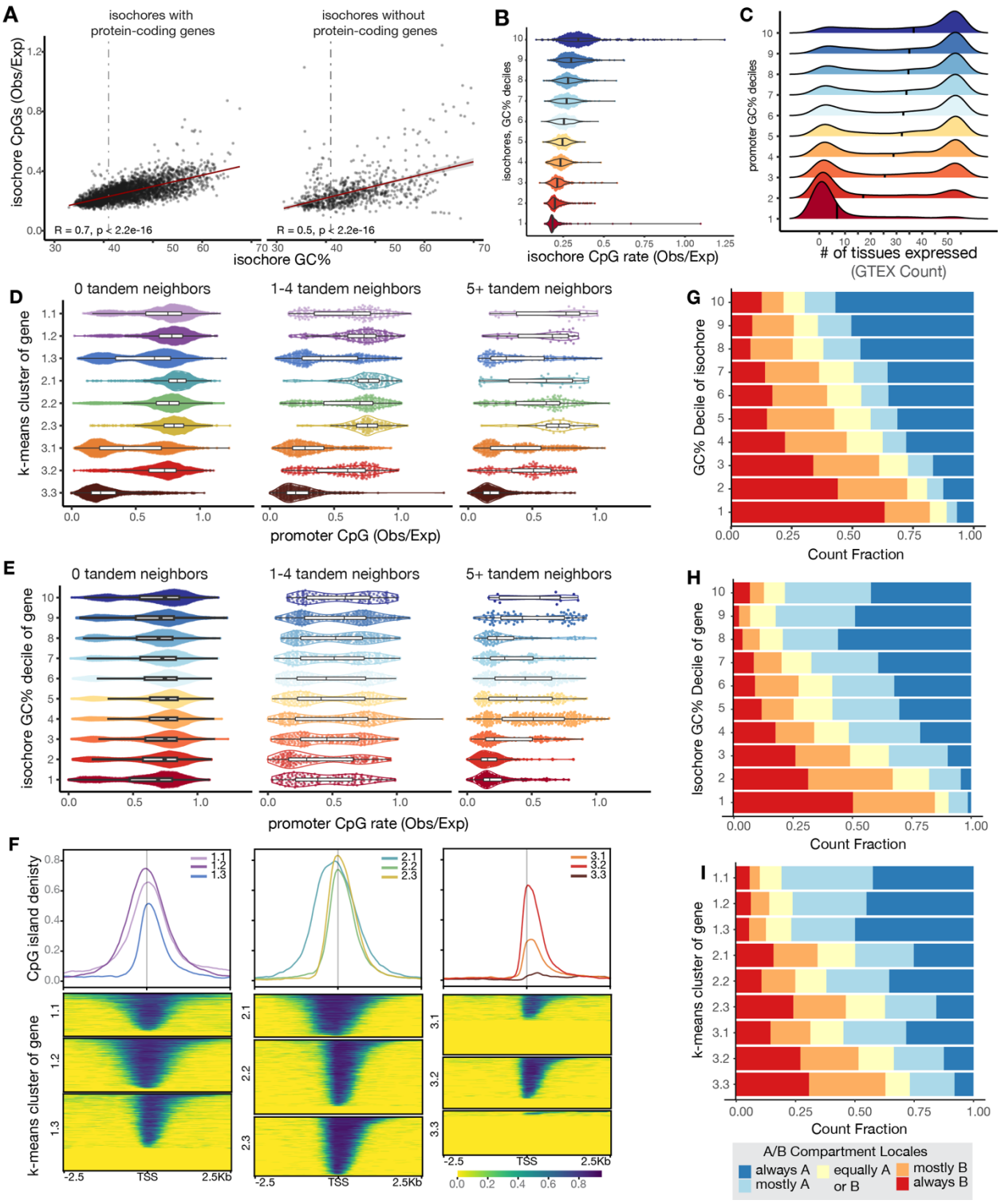

Figure S7: (A) Total CpG dinucleotide rate across isochores by their GC% content, split by isochores containing protein-coding genes and isochores without protein-coding genes. (B) Violin plots displaying total CpG rate across isochores, binned by decile. (C) Ridgeline plots displaying GTEX tissue expression distribution for genes binned by promoter GC% decile. (D) Distribution of promoter CpG rate of genes binned by k-means cluster assignment, split by

686 degree of tandem duplication. (E) Distribution of promoter CpG rate of genes binned by isochore
687 GC% decile, split by degree of tandem duplication. (F) CpG island density around the TSS (grey
688 line) for genes in each k-means cluster. CpG island density reflects strength of called CpG
689 islands from USCS unmasked CpG islands (54). Mean CpG rates across gene TSS in each
690 cluster are summarized in line plots, while heatmaps represent island calls around the TSS for
691 each gene. (G-I) Hi-C compartment assignment across 21 tissues (96) for isochores (G), or genes
692 in each isochore (H) or k-means cluster (I). "Always A" and "always B" means the gene was
693 assigned to that compartment in every sampled tissue.
694
695 **Discussion**
696 Animals make extensive and varied contacts with the external environment, both
697 engaging with foreign molecules and producing and excreting their own substances.
698 Specialization of these input-output functions plays a definitive role in animal lifestyle, and often
699 occurs in mammals via amplification and diversification of tandemly arrayed gene families
700 (10,12,13,98,99). Extensive gene losses are also common—for example, just as the vomeronasal
701 organ is vestigial in humans, human genes for vomeronasal receptors are no longer functional
702 (100,101). Here, we define a common genomic architecture in mammals for genes whose
703 products engage the external world: they are in tandem arrays, are found in AT-biased isochores,
704 and lack CpG islands in their promoters (Figure 1) (13,44). In humans today, we find that genes
705 in AT-rich tandem arrays experience low rates of point mutation and recombination, and tend
706 toward tissue-specific expression patterns.
707
708 *Modeling the emergence of AT/GC sequence bias in outward-looking tandem arrays*
709 The rapid turnover of genes in tandem arrays by birth-and-death evolution can complicate
710 efforts to reconstruct their evolutionary history. Using synteny, we were able to track paralogue
711 number changes for a set of polymorphic arrays in mammals (Figure 4). For arrays of olfactory
712 receptors, skin proteins, and defensive protease inhibitors, AT content increases with copy
713 number across mammals. This analysis suggests that high AT content need not be an ancestral
714 feature of gene blooms, but can emerge during array expansion. Using population genetic data
715 from humans, we test whether elevated rates of allelic diversity in outward-looking genes results
716 from distinct mutational or selective effects. We accept the argument of Lynch that AT/GC
717 content itself is unlikely to be directly operated on by selection, and instead focus on neutral and
718 selective events affecting the genes contained in sequence-biased ischores (27).
719 We find that genes in AT-biased tandem arrays experience low ongoing rates of point
720 mutation (Figure 5) and low rates of recombination (Figure 6). We suggest that excessive allelic
721 diversity in these regions could have arisen due to weakened selection on historical point
722 mutations. Because the most common point mutations replace C or G bases with T or A bases,
723 tolerance of point mutations in expanding tandem arrays could shift GC content down; low rates
724 of point mutation in the present would be due to the scarcity of mutable CpG dinucleotides
725 remaining in these clusters (38). Tandem arrays also appear depleted for recombination events
726 that can shift GC content back up (Figure 6). Together, tolerance of historical point mutation and
727 strengthened intolerance of recombination as gene families expand would tend to bias expanding
728 arrays toward higher AT content (Figure 1).
729 While deamination of cytosine in the CpG context dominates mutation rates, many other
730 neutral processes contribute to the overall mutation spectrum, and these too operate differently in
731 copy-number-variable tandem arrays. For example, AT-biased regions of the genome are late-

732 replicating, which is associated with higher rates of germline mutations, and genes in tandem
733 arrays tend not to be transcribed in the testis, which would exempt them from transcription-
734 coupled repair (42,102). Excess allelic variation could also result from paralogous gene
735 conversion. While paralogous gene conversion can be measured for paralogous pairs, in
736 extended arrays this phenomenon would be difficult to distinguish from copy number variation
737 without long-read data spanning the region (103). Similarly, genes in tandem arrays do not have
738 one-to-one homology relationships across species. Trivially, this means that their interspecific
739 divergence is high, but they would be left out of gene-wise measures of divergence that rely on
740 clear homology assignments.
741 Other mutagenic and selective forces may also contribute to the evolution of AT-
742 sequence bias. For example, the repetitive nature of these regions of the genome predisposes
743 them to problems, such as replication fork slippage, that require non-homologous end joining
744 (NHEJ) to resolve (104,105). RCC4-DNA ligase IV is primary ligase complex for NHEJ, and it
745 preferentially ligates poly-dT single-stranded DNA and long dT overhangs (106). Polymerase
746 mu, which is also involved in NHEJ, shares a similar bias towards both dT and dC, though
747 preferentially towards dT (107). These biases could encourage these regions of the genome to
748 adopting AT-rich sequences with each duplication, and thus each addition of repetitive material.
749
750 *Implications for gene regulation*
751 Genes located in AT-biased tandem arrays are typically silenced almost everywhere in
752 the body and expressed at extremely high levels at a specific place and time. Where and when
753 these genes are expressed, they perform the definitive work of the cell type. Many of these
754 outward-looking gene families exhibit some kind of exclusive expression, from the fetal to adult
755 switch in hemoglobin β expression (i.e. exclusion over time) to the one-receptor-per-neuron
756 pattern of olfactory receptor expression (i.e. exclusion over space). We expect that genes located
757 in high-AT isochores that lack CpG islands in their promoters are durably silenced by particular
758 mechanisms when they are not expressed (i.e. almost everywhere), and that their expression will
759 be activated by non-canonical mechanisms in the single condition where each is expressed. As
760 we and others suggested previously for mouse ORs, CpG-less promoters are likely to be
761 regulated by non-canonical mechanisms that are independent of TATA Binding Protein (TBP)
762 and the recently discovered basal CpG-island-binding factors BANP and BEND3
763 (13,91,108,109). As Bird hypothesized presciently in 1986, "When activated, they appear to be
764 bound to a complex of tissue specific factors which presumably accomplish what…islands can
765 achieve using ubiquitous, non-tissue specific factors. Given that there are usually rather few
766 CpGs near tissue specific genes…one would not expect to find CpG or methylation built into the
767 activation mechanism of genes of this type. (92)" One could argue that these genes do not have
768 promoters at all and rely completely on locus control regions to concentrate and deliver
769 transcription factors to the TSS (45,110,111). This atypical architecture appears to predispose
770 these genes to restricted expression relative to their CpG-containing brethren: their ground state
771 is to be "off forever."
772 Many or most molecular genetic events are sensitive to variation in AT/GC distribution:
773 AT content predicts compartmentalization of the genome in 3D space, replication timing, and
774 patterns of histone marks (41,112–116). Typically, AT-biased sequence is packaged as
775 heterochromatin and silenced. Work on olfactory receptors, clustered protocadherins, and
776 secreted liver proteins suggest that these gene families are expressed from the context of
777 constitutive heterochromatin, which appears to be present prior to expression and to be retained

778 on family members that are not expressed (117–121). CpG islands also function as molecular
779 beacons: they mark transcription start sites, serve as recombination hotspots in the absence of
780 PRDM9, and act as replication origins in meiosis (87,112,122). The accrual of high AT content
781 in gene arrays and the lack of CpG islands is therefore likely to exert a strong effect on the
782 molecular regulation of these genes. In ectodermal development, single-copy genes accrue CpH
783 methylation, perhaps passively, while AT-rich gene arrays remain devoid of this modification
784 (94). This suggests that AT-rich gene arrays are locked away from the ambient molecular stew of
785 the nucleus, perhaps over very long developmental time periods. Indeed, tandem arrays clump
786 together in the nucleus in post-mitotic neurons (47,95).
787 In addition to the extreme time and/or tissue specificity of most outward-looking gene
788 families, a fraction of these families exhibit stochastic expression such that each cell expresses
789 just one or a sparse subset of family members. Chemosensors, B- and T- cell receptors, NK cell
790 receptors, and clustered protocadherins all exhibit this restricted expression (119,123,124). As
791 we argued recently, sparse cell-wise expression patterns compartmentalize the effects of
792 mutations (119). These mechanisms are also likely to result in insensitivity to copy number
793 variation, as each cell chooses its own dose of family members for expression. Feedback
794 mechanisms that ensure cells can "choose again" if they originally pick a pseudogene further
795 buffer potential deleterious effects of mutations (125,126). Combined with the sheer numbers of
796 family members that a particular gene function is distributed across, these expression
797 mechanisms could predispose these genes to selective drift.
798
799 *Role for recombination in diversification versus maintenance of gene arrays*
800 Local sequence diversity and recombination rates are usually positively correlated
801 (reviewed in (127)). In contrast, we find that AT-biased gene families exhibit high diversity and
802 divergence despite low recombination. Comparison of the human and chimpanzee genomes
803 found two distinct types of high-divergence regions: Genome-wide, chromosome ends are GC-
804 rich and have high divergence and recombination; in chromosome interiors, however,
805 recombination is more modest overall, and divergence is higher in G-bands, which are AT-rich
806 (20). As Holmquist argued previously based on much more limited data, we show here that these
807 G-bands contain very different types of genes than do Q-bands (15,48).
808 Overall GC content is theorized to be a record of past gene conversion, with high-GC
809 regions having experienced high historical recombination (31). If this is the case, then AT-bias in
810 tandem arrays could reflect historical depletion of recombination in these regions, while also
811 dampening recombination rates in the present (Figure 6). Modeling suggests that once variation
812 in AT/GC content starts to emerge, it can be self-reinforcing via positive feedback (32). There
813 remains conflict between the mode by which these gene arrays are thought to have bloomed—i.e.
814 via gene duplication through ectopic exchange during recombination—and their current
815 depletion for recombination events. Other modes of duplication, including replication slippage
816 and transposition, may also be at work in expanding these arrays, and duplication mechanisms
817 could themselves influence GC content (128).
818 Ectopic exchange in repetitive gene regions can have benign or catastrophic
819 consequences. Induction of copy number variation within a gene array may be of small
820 phenotypic consequence, as the jobs these genes perform are by nature distributed across many
821 family members. In contrast, ectopic exchange that deletes a cluster or induces recombination
822 between clusters can result in catastrophic chromosome rearrangements (82). Retrotransposition
823 or ectopic exchange mediated by repetitive elements can seed new gene clusters elsewhere in the

genome (129,130), and gene families with more than one cluster genome-wide are likely to be particularly dangerous for genome stability. Indeed, mammalian chromosome evolution appears to have been shaped by ectopic exchange between OR clusters, to the extent that ORs are often positioned near chromosome ends (131–139). Finally, even if structural variation in outward-looking tandem arrays is benign within an individual, it can lead to hybrid incompatibility and can initiate or reinforce reproductive isolation that leads to speciation (140–142). Recent modeling work has sought to characterize the tradeoffs between the structural fragility of gene blooms and the potential positive effects of allelic diversification (143).

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 (12,144). Astonishingly, these genes have remained syntenic. Why 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 (145–147). Recent research on enhancer evolution in animals suggests that enhancer tangling can result in the preservation of synteny over ~700 million years (148). 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 (111,111,149–152). This mutual dependence would again increase the phenotypic consequences of recombination events that break synteny by separating genes in large families from locus control regions they depend on for expression.

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 could cause 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 could allow products of these meiosis to be preserved in the population. As gene arrays get larger, point mutation tolerance could shift their GC content downward, putting the brakes on recombination as they become ever more unwieldy. Overall, recombination in these regions would be dampened, 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 (153). 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 (138,154). 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

870 are observed at chromosome ends across eukaryotes (155). In the modern human population,
871 unequal crossovers between OR clusters are a source of recurrent and pathological
872 rearrangement hotspots (156).
873 While we also observe these AT-rich isochores *near* chromosome ends, terminal
874 isochores are often some of the most GC-rich in the genome and can contain heterogeneous
875 single-copy genes (157). Indeed, the largest OR cluster at the end of the q-arm of chromosome 1
876 in humans is followed by a higher GC% isochore containing ZNF genes. This strong end-GC%
877 accumulation appears to arise paternally: genes in high GC% isochores at chromosome ends are
878 enriched for paternal crossovers and relatively depleted of maternal crossovers. Overall, paternal
879 crossovers are biased towards chromosome ends (81,158). Chromatin organization of pachytene
880 spermatocytes is implicated in this phenomenon. Spermatocytes have shorter synaptonemal
881 complexes compared to oocytes, and, futher, subtelomeric regions do not require PRDM9 for
882 crossovers, however, how these factors contribute to paternal crossover end-bias remains
883 incompletely understood (159,160). Potentially, recombination-based alternative lengthening of
884 telomeres (ALT) in spermatocytes biases hotspots towards chromosome ends (161).
885 Over evolutionary time, as ectopic recombination places high-AT tandem arrays at
886 chromosome ends, high paternal rates of gBGC at the ends of chromosomes would generate new
887 isochores of increasing GC% comprising newly evolving genes (162,163).
888
889 *Is mutation biased or random with respect to gene function?*
890 Recent mutation accumulation studies have suggested that *de novo* mutations could occur
891 with different frequencies in different kinds of genes or in genic versus non-genic locations
892 (164). We and others also argued previously that segregation of mutation-tolerant versus
893 mutation-intolerant genes into AT- versus GC-biased regions of the genome could allow
894 differential mutation rates on different classes of genes (13,67,68). However, our analyses of
895 synonymous versus functional variant rate and of *de novo* mutation rate in AT- versus GC-biased
896 genes suggest the opposite: that AT-biased genes experience fewer mutations in living humans
897 than do GC-biased genes. This is in line with neutral, sequence-based expectations (42). While
898 CpG prevalence dominates the mutation spectrum, other mutational and repair processes that
899 differentially affect AT- versus GC-biased regions of the genome almost certainly contribute to
900 the overall pattern of diversity in inward- versus outward-looking genes. In particular, AT-rich
901 regions of the genome are late-replicating, and tandemly arrayed genes are excluded from
902 transcription-coupled repair during spermatogenesis. Each of these patterns could increase the
903 mutation rate in outward-looking genes. Late replication can worsen the loss of cytosines, while
904 lack of TCR would be expected to increase the rate of A>G (T>C) mutations (42,102).
905 While mutagenic or repair processes specific to gene blooms could contribute to their
906 overall mutation spectrum, we expect that for most, allelic diversity and differential AT/GC
907 content in inward- versus outward-looking genes result from differential selection trajectory over
908 evolutionary time. One attractive model is that as a gene cluster expands in size and the function
909 of that gene family is partitioned over more and more members, purifying selection becomes
910 weaker and weaker on individual family members (Figure 1). Because the most common point
911 mutations are C->T and especially CpG->TpG, evolutionary tolerance of point mutations would
912 shift GC content down. On the other hand, purifying selection and higher recombination rate
913 would both preserve the GC content of singleton genes.
914
915 *Matching sequence content for genes and their isochore environments*

916 Throughout this study, we show that the local AT/GC content of genes correlates closely
917 with that of their isochore environment (e.g. Figure 2B, S2B-G). We found this pattern
918 perplexing. For example, while higher purifying selection on single-copy genes can help to
919 explain their higher GC content, we would expect most variants in the non-coding portions of an
920 isochore to be phenotypically neutral. Because gene conversion acts over longer chromosomal
921 distances, differential recombination could affect the sequence content of genes and their
922 environment together. This pattern could also result from background/linked selection, which is
923 known to vary in strength across the genome (165,166). Because function-changing variants are
924 physically linked to local variants that don't alter phenotype, differential purifying selection on
925 mutations in single-copy versus multi-copy genes would affect the strength of background
926 selection in the neighborhood and inheritance of linked neutral variants. Background selection
927 will operate over shorter distances in regions with higher recombination (i.e. GC-rich isochores),
928 but because multiple inward-looking, single-copy genes are co-located in the same isochore,
929 there may be little nearby that is exempt from linked selection. In this way, purifying selection
930 that results in maintenance of high GC content in single-copy genes would also preserve the high
931 GC content of their isochore neighborhood. Finally, variation in repetitive element distribution
932 could also contribute to the AT/GC content of isochores housing outward-looking versus inward-
933 looking genes. Indeed, increased LINE density in chemosensory gene families has been proposed
934 to contribute to their regulation (167).
935 The matching sequence content of genes and their isochore neighborhoods would seem to
936 facilitate coherent patterns of histone marks across entire isochore units and their organization in
937 nuclear space. On the other hand, genes with AT- versus GC- rich coding sequences use distinct
938 codons, which could affect their translation rates given variation in the prevalence of various
939 tRNAs. Further analysis will be required to assess whether amino acid distribution varies for
940 proteins coded by AT-rich versus GC-rich sequences.
941 Finally, while we can find genes with GC-rich promoters in all kinds of isochores (Figure
942 S3F), we never find genes with AT-rich promoters in GC-rich isochores. While we favor the
943 model shown in Figure 1, where isochore-level AT content rises during gene blooms due to the
944 parallel actions of protein coding drift and suppressed recombination, we cannot exclude (and are
945 indeed fascinated by) the possibility that the distinct regulatory characteristics that allow LCR-
946 mediated expression could be the keystone factor tying outward-looking genes to AT-rich
947 isochores. However, in the absence of other evidence, we think that LCR-driven transcription is
948 a kludge that allows expression of paralogous clusters that failed to maintain their CpG's, rather
949 than CpG-less promoters emerging via positive selection to allow regulation by LCRs.
950
951 *Is this genomic architecture specific to mammals?*
952 While isochore structure is not unique to mammals, it is not a universal feature across
953 animal clades, and the AT/GC variation observed in mammals is extreme (27). We are curious
954 whether stem mammals evolved molecular mechanisms that facilitated the evolution of gene
955 arrays. These could include both systems that maintain these arrays as constitutive
956 heterochromatin when they are not being expressed and unique transcriptional mechanisms that
957 activate them, often in a stochastic or highly restricted manner, in their target tissues. One
958 candidate factor that mediates long-range enhancer-promoter interactions in multiple arrayed
959 families is Ldb1 (45,168). Social insects have also massively expanded their olfactory receptor
960 gene repertoire in cis; in the ant, this is accompanied by increased AT content (169). 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 (170–173). 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 (174). Trypanosome arrays of surface VSGs are located in subtelomeric regions (175). For mammals, the "isochore solution" balances diversity in gene arrays with genomic integrity.

Author Contributions

Conceptualization, MVB, MAC, EJC; Methodology, MVB, MAC, MLH, EJC; Investigation, MVB, MAC, MLH, EJC; Formal Analysis, MVB, MAC, MLH; Visualization, MVB; Data curation, MAC; Writing – Original Draft, MVB, MAC, EJC; Writing – Review & Editing, MVB, MAC, MLH, EJC; Funding Acquisition, EJC; Supervision, EJC.

Acknowledgements

Thanks to Rachel Duffié, Sebastian Zöllner, David Lyons, 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.

Declaration of interests

The authors declare no competing interests.

Descriptions of supplemental files

Supplemental Table 1: Isochore calls in the hg38 human genome assembly.
Supplemental Table 2: Isochore calls in the hg19 human genome assembly.
Supplemental Table 3: Results of all statistical analyses presented here.
Supplemental Table 4: Genewise data, including isochore assignment, k-means cluster, local sequence characteristics, variant rates.
Supplemental Table 5: Description of data in Supplemental Table 4.
Supplemental Table 6: GC contents of selected gene families and their annotated features (ie: CDS, introns, intergenic regions).
Supplemental Table 7: List of genomes searched with Tandem ClipR.

Methods

Describing isochores

To call isochores, we implemented a genomic segmentation algorithm called GC-Profile (50) 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

1006 rank 1 having the highest and 4328 having the lowest. We also performed this analysis in hg19
1007 (Supplemental Table 2). Isochores are reported in Supplemental Tables 1-2.
1008
1009 *Statistical analyses*
1010 We performed Kruskal-Wallis non-parametric ANOVA for each group of comparisons
1011 (Supplemental Table 3). We then used Dunn's pairwise analysis to compare individual groups
1012 with one another (Supplemental Tables 3). We use phylogenetic generalized least squares
1013 (PGLS) regression as implemented in caper (176) to test for correlations between genomic
1014 characteristics across vertebrate species while controlling for the evolutionary non-independence
1015 of the multi-species dataset. We used timetree.org (177) to produce our species phylogeny for
1016 these analyses, where the tree was imported and pruned to the subset of species for a given
1017 comparison using the ape package (178). In PGLS, lambda was optimized by 0 and 1 by
1018 maximum likelihood.
1019
1020 *GC content calculations*
1021 Genes from the Matched Annotation dataset from the NCBI and EMBL-EBI (MANE)
1022 Select dataset (51) were downloaded from the UCSC Genome Browser. Isochores were matched
1023 to genes using the coordinates of the transcription start site. GC content across gene features,
1024 including promoters (-750 to +250bp flanking TSS), flanking regions (+/- 25kb), coding exons,
1025 exons and UTRs, and introns were separately calculated from FASTA sequences using bedTools
1026 (179).
1027 To generate 9 k-means clustesr, we used gc5BaseBw from the UCSC Genome Browser
1028 (53) to calculate GC% scores across MANE genes with +/- 1kb flanks. We generated 3-kmeans
1029 clusters of genes, which were further clustered into 3-kmeans clusters each using deepTools
1030 plotHeatmap (180). Cluster assignment and quantification of other parameters for each gene are
1031 reported in Supplemental Table 4.
1032
1033 *Characterizing types of genes*
1034 To characterize the types of genes residing in isochores of varying GC, we used 2
1035 categories of descriptors: GO terms and gene prefixes. To identify GO terms associated with
1036 genes in each isochore GC decile, we used the R package, clusterProfiler (version 4.2.2)(181).
1037 This helped us streamline identification of key terms that appeared in each decile. With this list,
1038 we identified GO terms that were most significantly enriched in each decile with a depth of at
1039 least 30 genes. Using AmiGO (182), an online database of GO identifiers, we pulled the list of
1040 genes associated with our selected group of significant GO terms and plotted GC content across
1041 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)(183) 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 we have the most up-to-date names for each of our genes (ie: some which may have been labeled as 'FAM' may have a new symbol to go with the rest of the gene family).
    - Match names in the MANE set to names labeled "Approved symbols" by HGNC, and replace those symbols with the HUGO names.
2. Replace numbers with "_". We can't remove all numbers because there are several genes that have more letters after numbers that aren't important for our purposes (ie: CSN2A will become CSN_A).
3. Remove anything after the first instance of "_" (ie: CSN_A will become CSN). The goal of this step is to keep the first part of the prefix, removing letters and numbers that indicate subfamilies.
4. While it isn't common, some genes require us to know those numbers to know what they do (most commonly, enzymes involved in modifying carbohydrates). Largely, these genes start with a single letter, followed by numbers, then more letters. Thus, to fix these genes, we pull out the genes that have 1 letter after steps 4 and 5.
5. Look through each of those genes that start with only one letter, then decide how best to group them.
6. View gene prefixes in alphabetical order and search for prefixes that are likely to be families, then rename (ie: KCNT and KCNQ are both potassium channels, so we grouped these together).

Once we had a list of gene prefixes, we calculated a Shannon's H diversity metric for each isochore based on the prefix probabilities in each isochore (proportions + log2(1/proportions) = diversity metric). Larger values are indicative of more diversity. Similarly, we calculated a Shannon's H diversity metric for each k-means cluster.

De novo *mutations*

De novo mutations (DNMs) were compiled by (75) from seven family-based whole genome sequencing (WGS) datasets, encompassing a total of 679,547 single nucleotide variants (SNVs), which comprise data from both neurotypical and neurodivergent individuals. We remapped the dataset to hg38 using LiftOver in UCSC Genome Browser. To calculate genomic DNM density, we counted the number of DNMs occurring within the coordinates listed in the GC calculation section above. To calculate DNM density, we pooled genic DNMs within each isochore and divided by the sum of the region of interest's size, i.e. we identified all the genes in

1084 an isochore, summed the DNMs between their transcription start and termination sites, then
1085 divided by the summed length of those genic regions.
1086
1087 *Allelic variants*
1088 We used the gnomAD v2.1.1 dataset of single nucleotide allelic variants (69). The
1089 authors defined rare single nucleotide variants (<0.1% allele frequency) from 125,748 exomes
1090 and 15,708 whole genomes and predicted whether variants within coding regions are likely to be
1091 functionally synonymous, missense, or loss-of-function. Here, we used observed synonymous,
1092 missense, and loss-of-function mutation rates. We ported variant calls to MANE genes in hg38
1093 using the gene symbol and Ensembl transcript IDs. In Figure 5, we also use the calculated pLI
1094 score from gnomAD, which describes the likelihood that a gene is loss-of-function intolerant in
1095 humans.
1096
1097 *Recombination*
1098 Crossover data for hg38 was acquired from deCODE where the authors used whole-
1099 genome sequence (WGS) of trios and were able to refine crossover boundaries for 247,942
1100 crossovers in 9423 paternal meioses and 514,039 crossovers in 11,750 maternal meioses (78). Of
1101 note, the data we used here is restricted to autosomes. To calculate crossover density, we
1102 assigned crossovers to a region of interest based on the median of the crossover coordinates. We
1103 normalized counts within a region by dividing by the genomic average for that sex. PRDM9
1104 binding data from HEK293T cells transfected with the PRDM9 reference allele was acquired
1105 from (86). We selected the top 10% of PRDM9 peaks based on enrichment scores to account for
1106 weak PRDM9 binding sites associated with overexpression in the system, as noted by the
1107 authors. Like with crossovers, we calculated PRDM9 binding site density across genes as the
1108 summed enrichment scores across genic regions within an isochore, mapping by the midpoint of
1109 the binding coordinates.
1110
1111 *Gene Regulatory Information*
1112 To determine tissue specificity of gene expression, RNA-sequencing data was sourced
1113 from Genotype-Tissue Expression (GTEx) project (V8, released in August 2019), containing
1114 17,382 samples collected from 54 tissues from 948 donors (93). For each gene in the MANE set,
1115 we counted the number of tissues in which expression was at least 5 transcripts per million
1116 (TPM).
1117 To measure A/B compartment occupancy of genes across tissues, AB compartments were
1118 sourced from published Hi-C data from 21 tissues and cell types (96). MANE genes were lifted
1119 over into hg19 to match A/B compartment domain calls in hg19. Isochores called in hg19 were
1120 assigned to a compartment by matching the isochore's midpoint to the midpoint of the closest
1121 compartment. Genes were assigned to a compartment by matching the transcription start site to
1122 the midpoint of the nearest compartment, as most genes did not fall into a single compartment
1123 (~90%). Further, we counted the occurrences of compartments A and B for each isochore and
1124 gene. These counts were binned into always A (21 counts of A), mostly A (14-20 counts of A or
1125 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
1126 counts of B), and always B (21 counts of B).
1127 To identify genes with CpG islands in promoter regions, we downloaded the CpG Island
1128 track (unmasked) from the UCSC Genome Browser (54). The ratio of observed vs. expected
1129 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.

Gene Bloom Evolution with Tandem ClipR

To access and analyze syntenic tandem gene blooms across species, we created a pipeline we call TandemClipR that defines tandem arrays based on orthologous "bookend" genes. Bookends define syntenic array bounding positions and are not members of the bloomed family. First, we used the biomaRt R package (184) to define the human genome as the reference mart, and then created a table of bookending ortholog chromosome names, starts, and stops using the 'getLDS' function and the dec2021 ensembl archive as a source. We tabulated these ortholog positions in all 193 vertebrate species in the archive (See Supplemental Tables 6 and 7). If both orthologs were on the same scaffold, indicating a contiguous assembly of the tandem array, we used the retrieved coordinates to subset the gff3 annotation file to the focal region using the subsetByOverlaps() function in the GenomicRanges package (185). Filters were applied when imported annotations to retain only "gene", "mRNA","CDS","exon","five_prime_UTR", and "three_prime_UTR" annotations.

The resulting gff3 files were used to analyze feature GC content across each tandem array based on sequences in the Ensembl "toplevel" assemblies. Namely, we used command-line tools to subset the gff3 annotations into focal features (genes, exons, five_prime_UTR), and used while awk and bedtools merge and subtract (179) were used to produced genic, intergenic, exonic, intronic, and whole cluster bedfiles defining these regions. Additionally, bedfiles defining the promoter region of each gene were produced by defining a window 750bp upstream and 250bp downstream of the transcription start site. We fed these bedfiles to bedtools nuc to calculate GC content for each feature, which were counted and aggregated into average GC% for final analyses. We spot-checked gene counts in NCBI's Genome Data Viewer and noticed that there were a few cases in which counts varied between recent assemblies. For example, the lion HBB-OR cluster is much shorter in the assembly we used compared to the most recent assembly. However, our manual counts of the HBB-OR array in a subset of species produced similar trends as the automated annotation. To further verify that one of these assembly/annotation errors did not sway the overall trend, we ran each PGLS analysis iteratively, removing one species at a time, and found that PGLS p-values of the whole were largely reflected in each iteration.

References

1. Charkoftaki G, Wang Y, McAndrews M, Bruford EA, Thompson DC, Vasiliou V, et al. Update on the human and mouse lipocalin (LCN) gene family, including evidence the mouse Mup cluster is result of an "evolutionary bloom." Human Genomics. 2019 Feb 19;13(1):11.

2. Feyereisen R. Evolution of insect P450. Biochem Soc Trans. 2006 Dec;34(Pt 6):1252–5.

3. Nelson DR, Goldstone JV, Stegeman JJ. The cytochrome P450 genesis locus: the origin and evolution of animal cytochrome P450s. Philos Trans R Soc Lond B Biol Sci. 2013 Feb 19;368(1612):20120474.

4. Ohno, Susumu. Evolution by gene duplication. [Internet]. Springer; 1970 [cited 2021 Nov 9]. Available from:

1173 https://www.ncbi.nlm.nih.gov/nlmcatalog/?term=evolution%20by%20gene%20duplication
1174 %20ohno%201970

1175 5. Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW. The evolution of mammalian
1176 gene families. PLoS One. 2006 Dec 20;1:e85.

1177 6. Giorgianni MW, Dowell NL, Griffin S, Kassner VA, Selegue JE, Carroll SB. The origin and
1178 diversification of a novel protein family in venomous snakes. PNAS. 2020 May
1179 19;117(20):10911–20.

1180 7. Holding ML, Strickland JL, Rautsaw RM, Hofmann EP, Mason AJ, Hogan MP, et al.
1181 Phylogenetically diverse diets favor more complex venoms in North American pitvipers.
1182 Proc Natl Acad Sci U S A. 2021 Apr 27;118(17):e2015579118.

1183 8. Johnson RN, O'Meally D, Chen Z, Etherington GJ, Ho SYW, Nash WJ, et al. Adaptation
1184 and conservation insights from the koala genome. Nat Genet. 2018 Aug;50(8):1102–11.

1185 9. Pavlovich SS, Lovett SP, Koroleva G, Guito JC, Arnold CE, Nagle ER, et al. The Egyptian
1186 Rousette Genome Reveals Unexpected Features of Bat Antiviral Immunity. Cell. 2018 May
1187 17;173(5):1098-1110.e18.

1188 10. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution
1189 of human amylase gene copy number variation. Nat Genet. 2007 Oct;39(10):1256–60.

1190 11. Sun X, Zhang Z, Sun Y, Li J, Xu S, Yang G. Comparative genomics analyses of alpha-
1191 keratins reveal insights into evolutionary adaptation of marine mammals. Frontiers in
1192 Zoology. 2017 Aug 2;14(1):41.

1193 12. Kawasaki K, Lafont AG, Sire JY. The evolution of milk casein genes from tooth genes
1194 before the origin of mammals. Mol Biol Evol. 2011 Jul;28(7):2053–61.

1195 13. Clowney EJ, Magklara A, Colquitt BM, Pathak N, Lane RP, Lomvardas S. High-throughput
1196 mapping of the promoters of the mouse olfactory receptor genes reveals a new type of
1197 mammalian promoter and provides insight into olfactory receptor gene regulation. Genome
1198 Res. 2011 Aug;21(8):1249–59.

1199 14. Corneo G, Ginelli E, Soave C, Bernardi G. Isolation and characterization of mouse and
1200 guinea pig satellite deoxyribonucleic acids. Biochemistry. 1968 Dec;7(12):4373–9.

1201 15. Holmquist GP. Chromosome bands, their chromatin flavors, and their functional features.
1202 Am J Hum Genet. 1992 Jul;51(1):17–37.

1203 16. Bickmore WA. Patterns in the genome. Heredity. 2019 Jul;123(1):50–7.

1204 17. Filipski J. Evolution of DNA Sequence Contributions of Mutational Bias and Selection to the
1205 Origin of Chromosomal Compartments. In: Obe G, editor. Advances in Mutagenesis
1206 Research [Internet]. Berlin, Heidelberg: Springer; 1990 [cited 2021 Feb 26]. p. 1–54.

1207 (Advances in Mutagenesis Research). Available from: https://doi.org/10.1007/978-3-642-
1208 75599-6_1

1209 18. Korenberg JR, Rykowski MC. Human genome organization: Alu, lines, and the molecular
1210 structure of metaphase chromosome bands. Cell. 1988 May 6;53(3):391–400.

1211 19. Schug J, Schuller WP, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ. Promoter features
1212 related to tissue specificity as measured by Shannon entropy. Genome Biology. 2005 Mar
1213 29;6(4):R33.

1214 20. Waterson RH, Lander ES, Wilson RK, The Chimpanzee Sequencing and Analysis
1215 Consortium. Initial sequence of the chimpanzee genome and comparison with the human
1216 genome. Nature. 2005 Sep;437(7055):69–87.

1217 21. Niimura Y, Gojobori T. In silico chromosome staining: Reconstruction of Giemsa bands
1218 from the whole human genome sequence. Proc Natl Acad Sci U S A. 2002 Jan
1219 22;99(2):797–802.

1220 22. Costantini M, Clay O, Auletta F, Bernardi G. An isochore map of human chromosomes.
1221 Genome Res. 2006 Apr 1;16(4):536–41.

1222 23. Cohen N, Dagan T, Stone L, Graur D. GC Composition of the Human Genome: In Search of
1223 Isochores. Molecular Biology and Evolution. 2005 May 1;22(5):1260–72.

1224 24. Cozzi P, Milanesi L, Bernardi G. Segmenting the Human Genome into Isochores. Evol
1225 Bioinform Online. 2015 Nov 25;11:253–61.

1226 25. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing
1227 and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860–921.

1228 26. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic
1229 landscapes. Annu Rev Genomics Hum Genet. 2009;10:285–311.

1230 27. Lynch, Michael. The origins of genome architecture [Internet]. Indiana University Press;
1231 2007 [cited 2021 Nov 9]. Available from:
1232 https://www.ncbi.nlm.nih.gov/nlmcatalog/101296442

1233 28. Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, et al. The mosaic genome of
1234 warm-blooded vertebrates. Science. 1985 May 24;228(4702):953–8.

1235 29. Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. Quantification of GC-biased
1236 gene conversion in the human genome. Genome Res. 2015 Aug 1;25(8):1215–28.

1237 30. Duret L, Arndt PF. The Impact of Recombination on Nucleotide Substitutions in the Human
1238 Genome. PLOS Genetics. 2008 May 9;4(5):e1000071.

1239 31. Pouyet F, Mouchiroud D, Duret L, Sémon M. Recombination, meiotic expression and human
1240 codon usage. Przeworski M, editor. eLife. 2017 Aug 15;6:e27344.

1241 32. Fryxell KJ, Zuckerkandl E. Cytosine deamination plays a primary role in the evolution of
1242 mammalian isochores. Mol Biol Evol. 2000 Sep;17(9):1371–83.

1243 33. Hershberg R, Petrov DA. Evidence that mutation is universally biased towards AT in
1244 bacteria. PLoS Genet. 2010 Sep 9;6(9):e1001115.

1245 34. Hildebrand F, Meyer A, Eyre-Walker A. Evidence of selection upon genomic GC-content in
1246 bacteria. PLoS Genet. 2010 Sep 9;6(9):e1001107.

1247 35. Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proceedings of
1248 the National Academy of Sciences. 2010 Jan 19;107(3):961–8.

1249 36. Simmen MW. Genome-scale relationships between cytosine methylation and dinucleotide
1250 abundances in animals. Genomics. 2008 Jul 1;92(1):33–40.

1251 37. Mugal CF, Arndt PF, Holm L, Ellegren H. Evolutionary Consequences of DNA Methylation
1252 on the GC Content in Vertebrate Genomes. G3 (Bethesda). 2015 Jan 15;5(3):441–7.

1253 38. Fryxell KJ, Moon WJ. CpG Mutation Rates in the Human Genome Are Highly Dependent
1254 on Local GC Content. Molecular Biology and Evolution. 2005 Mar 1;22(3):650–8.

1255 39. Jabbari K, Chakraborty M, Wiehe T. DNA sequence-dependent chromatin architecture and
1256 nuclear hubs formation. Sci Rep. 2019 Oct 10;9(1):14646.

1257 40. Ramani V, Shendure J, Duan Z. Understanding Spatial Genome Organization: Methods and
1258 Insights. Genomics, Proteomics & Bioinformatics. 2016 Feb 1;14(1):7–20.

1259 41. Woodfine K, Fiegler H, Beare DM, Collins JE, McCann OT, Young BD, et al. Replication
1260 timing of the human genome. Hum Mol Genet. 2004 Jan 15;13(2):191–202.

1261 42. Ségurel L, Wyman MJ, Przeworski M. Determinants of Mutation Rate Variation in the
1262 Human Germline. Annual Review of Genomics and Human Genetics. 2014;15(1):47–70.

1263 43. Armelin-Correa LM, Gutiyama LM, Brandt DYC, Malnic B. Nuclear compartmentalization
1264 of odorant receptor genes. Proc Natl Acad Sci U S A. 2014 Feb 18;111(7):2782–7.

1265 44. Clowney EJ, LeGros MA, Mosley CP, Clowney FG, Markenskoff-Papadimitriou EC, Myllys
1266 M, et al. Nuclear aggregation of olfactory receptor genes governs their monogenic
1267 expression. Cell. 2012 Nov 9;151(4):724–37.

1268 45. Monahan K, Horta A, Lomvardas S. LHX2- and LDB1-mediated trans interactions regulate
1269 olfactory receptor choice. Nature. 2019;565(7740):448–53.

1270 46. Tan L, Xing D, Daley N, Xie XS. Three-dimensional genome structures of single sensory
1271 neurons in mouse visual and olfactory systems. Nature Structural & Molecular Biology.
1272 2019 Apr;26(4):297–307.

1273 47. Tan L, Ma W, Wu H, Zheng Y, Xing D, Chen R, et al. Changes in genome architecture and
1274 transcriptional dynamics progress independently of sensory experience during post-natal
1275 brain development. Cell. 2021 Feb 4;184(3):741-758.e17.

1276 48. Holmquist GP, Filipski J. Organization of mutations along the genome: a prime determinant
1277 of genome evolution. Trends Ecol Evol. 1994 Feb;9(2):65–9.

1278 49. Jabbari K, Bernardi G. An Isochore Framework Underlies Chromatin Architecture. PLOS
1279 ONE. 2017 Jan 6;12(1):e0168023.

1280 50. Gao F, Zhang CT. GC-Profile: a web-based tool for visualizing and analyzing the variation
1281 of GC content in genomic sequences. Nucleic Acids Res. 2006 Jul 1;34(Web Server
1282 issue):W686–91.

1283 51. Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, et al. A joint NCBI and
1284 EMBL-EBI transcript set for clinical genomics and research. Nature. 2022
1285 Apr;604(7905):310–5.

1286 52. Glusman G, Yanai I, Rubin I, Lancet D. The complete human olfactory subgenome. Genome
1287 Res. 2001 May;11(5):685–702.

1288 53. Clawson H. GC Percent in 5-Base Windows (gc5BaseBw). Unpublished.
1289 https://genome.ucsc.edu/cgi-
1290 bin/hgc?hgsid=950293831_VmX9SYAwpTKoTme1yOSQX58aR0aR&c=chr16&l=48043
1291 557&r=48045592&o=48043557&t=48045592&g=gc5BaseBw&i=gc5BaseBw. Accessed
1292 6/21/22 [Internet]. 2018 [cited 2022 Jun 21]. Available from: https://genome.ucsc.edu/cgi-
1293 bin/hgc?hgsid=950293831_VmX9SYAwpTKoTme1yOSQX58aR0aR&c=chr16&l=48043
1294 557&r=48045592&o=48043557&t=48045592&g=gc5BaseBw&i=gc5BaseBw

1295 54. Micklem G, Hillier LW. CpG Islands. Unpublished.
1296 http://genomewiki.ucsc.edu/index.php/CpG_Islands. Accessed 6/21/22. [Internet]. 2006
1297 [cited 2022 Jun 21]. Available from: http://genomewiki.ucsc.edu/index.php/CpG_Islands

1298 55. Niimura Y, Nei M. Extensive Gains and Losses of Olfactory Receptor Genes in Mammalian
1299 Evolution. PLOS ONE. 2007 Aug 8;2(8):e708.

1300 56. Niimura Y, Matsui A, Touhara K. Extreme expansion of the olfactory receptor gene
1301 repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13
1302 placental mammals. Genome Res. 2014 Sep 1;24(9):1485–96.

1303 57. Khan I, Maldonado E, Vasconcelos V, O'Brien SJ, Johnson WE, Antunes A. Mammalian
1304 keratin associated proteins (KRTAPs) subgenomes: disentangling hair diversity and
1305 adaptation to terrestrial and aquatic environments. BMC Genomics. 2014 Sep 10;15(1):779.

1306 58. Walker MB, King BL, Paigen K. Clusters of Ancestrally Related Genes That Show Paralogy
1307 in Whole or in Part Are a Major Feature of the Genomes of Humans and Other Species.
1308 PLoS One. 2012 Apr 26;7(4):e35274.

1309 59. Sheehan MJ, Campbell P, Miller CH. Evolutionary patterns of major urinary protein scent
1310 signals in house mice and relatives. Molecular Ecology. 2019;28(15):3587–601.

1311 60. Hardison RC. Evolution of Hemoglobin and Its Genes. Cold Spring Harb Perspect Med.
1312 2012 Dec;2(12):a011627.

1313 61. Thomas JH. Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in
1314 vertebrates. PLoS Genet. 2007 May 11;3(5):e67.

1315 62. Tan HM, Low WY. Rapid birth-death evolution and positive selection in detoxification-type
1316 glutathione S-transferases in mammals. PLOS ONE. 2018 Dec 26;13(12):e0209336.

1317 63. Nei M, Niimura Y, Nozawa M. The evolution of animal chemosensory receptor gene
1318 repertoires: roles of chance and necessity. Nature Reviews Genetics. 2008 Dec;9(12):951–
1319 63.

1320 64. Schwartz JC, Gibson MS, Heimeier D, Koren S, Phillippy AM, Bickhart DM, et al. The
1321 evolution of the natural killer complex; a comparison between mammals using new high-
1322 quality genome assemblies and targeted annotation. Immunogenetics. 2017 Apr;69(4):255–
1323 69.

1324 65. Semple F, Dorin JR. β-Defensins: Multifunctional Modulators of Infection, Inflammation
1325 and More? JIN. 2012;4(4):337–48.

1326 66. Shelton JF, Shastri AJ, Fletez-Brant K, Stella Aslibekyan, Auton A. The UGT2A1/UGT2A2
1327 locus is associated with COVID-19-related loss of smell or taste. Nat Genet. 2022
1328 Feb;54(2):121–4.

1329 67. Chuang JH, Li H. Functional bias and spatial organization of genes in mutational hot and
1330 cold regions in the human genome. PLoS Biol. 2004 Feb;2(2):E29.

1331 68. Grimwood J, Gordon LA, Olsen A, Terry A, Schmutz J, Lamerdin J, et al. The DNA
1332 sequence and biology of human chromosome 19. Nature. 2004 Apr 1;428(6982):529–35.

1333 69. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The
1334 mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020
1335 May;581(7809):434–43.

1336 70. Mainland JD, Keller A, Li YR, Zhou T, Trimmer C, Snyder LL, et al. The missense of smell:
1337 functional variability in the human odorant receptor repertoire. Nat Neurosci. 2014
1338 Jan;17(1):114–20.

1339 71. Trimmer C, Keller A, Murphy NR, Snyder LL, Willer JR, Nagai MH, et al. Genetic variation
1340 across the human olfactory receptor repertoire alters odor perception. PNAS. 2019 May
1341 7;116(19):9475–80.

1342 72. Nachman MW, Crowell SL. Estimate of the Mutation Rate per Nucleotide in Humans.
1343 Genetics. 2000 Sep 1;156(1):297–304.

1344 73. Sved J, Bird A. The expected equilibrium of the CpG dinucleotide in vertebrate genomes
1345 under a mutation model. Proc Natl Acad Sci U S A. 1990 Jun;87(12):4692–6.

1346 74. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of
1347 protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug;536(7616):285–91.

1348 75. Rodriguez-Galindo M, Casillas S, Weghorn D, Barbadilla A. Germline de novo mutation
1349 rates on exons versus introns in humans. Nat Commun. 2020 Jul 3;11(1):3304.

1350 76. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, et al. Genome-wide
1351 patterns and properties of *de novo* mutations in humans. Nature Genetics. 2015
1352 Jul;47(7):822–6.

1353 77. Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, et al. Parental
1354 influence on human germline de novo mutations in 1,548 trios from Iceland. Nature. 2017
1355 Sep 28;549(7673):519–22.

1356 78. Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, et
1357 al. Characterizing mutagenic effects of recombination through a sequence-level genetic
1358 map. Science. 2019 Jan 25;363(6425).

1359 79. Jabbari K, Wirtz J, Rauscher M, Wiehe T. A common genomic code for chromatin
1360 architecture and recombination landscape. PLOS ONE. 2019 Mar 13;14(3):e0213278.

1361 80. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, et al.
1362 Fine-scale recombination rate differences between sexes, populations and individuals.
1363 Nature. 2010 Oct;467(7319):1099–103.

1364 81. Lee YS, Chao A, Chen CH, Chou T, Wang SYM, Wang TH. Analysis of human meiotic
1365 recombination events with a parent-sibling tracing approach. BMC Genomics. 2011 Aug
1366 26;12(1):434.

1367 82. Uhrberg M. The KIR gene family: life in the fast lane of evolution. European Journal of
1368 Immunology. 2005;35(1):10–5.

1369 83. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, et al. PRDM9 is a major
1370 determinant of meiotic recombination hotspots in humans and mice. Science. 2010 Feb
1371 12;327(5967):836–40.

1372 84. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, et al. Drive against
1373 hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science.
1374 2010 Feb 12;327(5967):876–9.

1375 85. Parvanov ED, Petkov PM, Paigen K. Prdm9 controls activation of mammalian
1376 recombination hotspots. Science. 2010 Feb 12;327(5967):835.

1377 86. Altemose N, Noor N, Bitoun E, Tumian A, Imbeault M, Chapman JR, et al. A map of human
1378 PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger
1379 proteins in meiosis. Przeworski M, editor. eLife. 2017 Oct 26;6:e28383.

1380 87. Baker Z, Schumer M, Haba Y, Bashkirova L, Holland C, Rosenthal GG, et al. Repeated
1381 losses of PRDM9-directed recombination despite the conservation of PRDM9 across
1382 vertebrates. de Massy B, editor. eLife. 2017 Jun 6;6:e24133.

1383 88. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet.
1384 2013 Mar;14(3):204–20.

1385 89. Mohn F, Schübeler D. Genetics and epigenetics: stability and plasticity during cellular
1386 differentiation. Trends Genet. 2009 Mar;25(3):129–36.

1387 90. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011 May
1388 15;25(10):1010–22.

1389 91. Michaloski JS, Galante PAF, Malnic B. Identification of potential regulatory motifs in
1390 odorant receptor genes by analysis of promoter sequences. Genome Res. 2006
1391 Sep;16(9):1091–8.

1392 92. Bird AP. CpG-rich islands and the function of DNA methylation. Nature. 1986 May
1393 15;321(6067):209–13.

1394 93. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013
1395 Jun;45(6):580–5.

1396 94. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, et al. Global
1397 Epigenomic Reconfiguration During Mammalian Brain Development. Science. 2013 Aug
1398 9;341(6146):1237905.

1399 95. Winick-Ng W, Kukalev A, Harabula I, Zea-Redondo L, Szabó D, Meijer M, et al. Cell-type
1400 specialization is encoded by specific chromatin topologies. Nature. 2021
1401 Nov;599(7886):684–91.

1402 96. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A Compendium of Chromatin
1403 Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Reports. 2016
1404 Nov 15;17(8):2042–59.

1405 97. Naughton C, Avlonitis N, Corless S, Prendergast JG, Mati IK, Eijk PP, et al. Transcription
1406 forms and remodels supercoiling domains unfolding large-scale chromatin structures. Nat
1407 Struct Mol Biol. 2013 Mar;20(3):387–95.

1408 98. Christmas MJ, Kaplow IM, Genereux DP, Dong MX, Hughes GM, Li X, et al. Evolutionary
1409 constraint and innovation across hundreds of placental mammals [Internet]. bioRxiv; 2023
1410 [cited 2023 Mar 22]. p. 2023.03.09.531574. Available from:
1411 https://www.biorxiv.org/content/10.1101/2023.03.09.531574v1

1412 99. Hughes GM, Boston ESM, Finarelli JA, Murphy WJ, Higgins DG, Teeling EC. The Birth
1413 and Death of Olfactory Receptor Gene Families in Mammalian Niche Adaptation.
1414 Molecular Biology and Evolution. 2018 Jun 1;35(6):1390–406.

1415 100. Witt M, Hummel T. Vomeronasal Versus Olfactory Epithelium: Is There a Cellular Basis
1416 for Human Vomeronasal Perception? In: International Review of Cytology [Internet].
1417 Academic Press; 2006 [cited 2021 Aug 11]. p. 209–59. Available from:
1418 https://www.sciencedirect.com/science/article/pii/S0074769606480049

1419 101. Zhang J, Webb DM. Evolutionary deterioration of the vomeronasal pheromone transduction
1420 pathway in catarrhine primates. PNAS. 2003 Jul 8;100(14):8337–41.

1421 102. Xia B, Yan Y, Baron M, Wagner F, Barkley D, Chiodin M, et al. Widespread
1422 Transcriptional Scanning in the Testis Modulates Gene Evolution Rates. Cell. 2020 Jan
1423 23;180(2):248-262.e21.

1424 103. Harpak A, Lan X, Gao Z, Pritchard JK. Frequent nonallelic gene conversion on the human
1425 lineage and its effect on the divergence of gene duplicates. PNAS. 2017 Nov
1426 28;114(48):12779–84.

1427 104. Caridi PC, Delabaere L, Zapotoczny G, Chiolo I. And yet, it moves: nuclear and chromatin
1428 dynamics of a heterochromatic double-strand break. Philosophical Transactions of the
1429 Royal Society B: Biological Sciences. 2017 Aug 28;372(1731):20160291.

1430 105. Jimeno S, Mejías-Navarro F, Prados-Carvajal R, Huertas P. Chapter Four - Controlling the
1431 balance between chromosome break repair pathways. In: Donev R, editor. Advances in
1432 Protein Chemistry and Structural Biology [Internet]. Academic Press; 2019 [cited 2023 Mar
1433 9]. p. 95–134. (DNA Repair; vol. 115). Available from:
1434 https://www.sciencedirect.com/science/article/pii/S187616231830066X

1435 106. Gu J, Lu H, Tsai AG, Schwarz K, Lieber MR. Single-stranded DNA ligation and XLF-
1436 stimulated incompatible DNA end ligation by the XRCC4-DNA ligase IV complex:
1437 influence of terminal DNA sequence. Nucleic Acids Res. 2007;35(17):5755–62.

1438 107. Davis BJ, Havener JM, Ramsden DA. End-bridging is required for pol μ to efficiently
1439 promote repair of noncomplementary ends by nonhomologous end joining. Nucleic Acids
1440 Res. 2008 May;36(9):3085–94.

1441 108. Grand RS, Burger L, Gräwe C, Michael AK, Isbel L, Hess D, et al. BANP opens chromatin
1442 and activates CpG-island-regulated genes. Nature. 2021 Jul 7;1–5.

1443 109. Zhang J, Zhang Y, You Q, Huang C, Zhang T, Wang M, et al. Highly enriched BEND3
1444 prevents the premature activation of bivalent genes during differentiation. Science. 2022
1445 Mar 4;375(6584):1053–8.

1446 110. Monahan K, Schieren I, Cheung J, Mumbey-Wafula A, Monuki ES, Lomvardas S.
1447 Cooperative interactions enable singular olfactory receptor expression in mouse olfactory
1448 neurons. Elife. 2017 21;6.

1449 111. Li Q, Peterson KR, Fang X, Stamatoyannopoulos G. Locus control regions. Blood. 2002
1450 Nov 1;100(9):3077–86.

1451 112. Pratto F, Brick K, Cheng G, Lam KWG, Cloutier JM, Dahiya D, et al. Meiotic
1452 recombination mirrors patterns of germline replication in mice and humans. Cell. 2021 Aug
1453 5;184(16):4251-4267.e20.

1454 113. Xie WJ, Meng L, Liu S, Zhang L, Cai X, Gao YQ. Structural Modeling of Chromatin
1455 Integrates Genome Features and Reveals Chromosome Folding Principle. Sci Rep
1456 [Internet]. 2017 Jun 6 [cited 2018 Jul 27];7. Available from:
1457 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5460185/

1458 114. Costantini M, Musto H. The Isochores as a Fundamental Level of Genome Structure and
1459 Organization: A General Overview. J Mol Evol. 2017 Mar 1;84(2):93–103.

1460 115. Dekker J. GC- and AT-rich chromatin domains differ in conformation and histone
1461 modification status and are differentially modulated by Rpd3p. Genome Biology. 2007 Jun
1462 18;8(6):R116.

1463 116. Wang Z, Willard HF. Evidence for sequence biases associated with patterns of histone
1464 methylation. BMC Genomics. 2012 Aug 2;13(1):367.

1465 117. Magklara A, Yen A, Colquitt BM, Clowney EJ, Allen W, Markenscoff-Papadimitriou E, et
1466 al. An epigenetic signature for monoallelic olfactory receptor expression. Cell. 2011 May
1467 13;145(4):555–70.

1468 118. Toyoda S, Kawaguchi M, Kobayashi T, Tarusawa E, Toyama T, Okano M, et al.
1469 Developmental Epigenetic Modification Regulates Stochastic Expression of Clustered
1470 Protocadherin Genes, Generating Single Neuron Diversity. Neuron. 2014 Apr 2;82(1):94–
1471 108.

1472 119. Williams DL, Sikora VM, Hammer MA, Amin S, Brinjikji T, Brumley EK, et al. May the
1473 Odds Be Ever in Your Favor: Non-deterministic Mechanisms Diversifying Cell Surface
1474 Molecule Expression. Front Cell Dev Biol. 2021;9:720798.

1475 120. Balan S, Iwayama Y, Ohnishi T, Fukuda M, Shirai A, Yamada A, et al. A loss-of-function
1476 variant in SUV39H2 identified in autism-spectrum disorder causes altered H3K9
1477 trimethylation and dysregulation of protocadherin β-cluster genes in the developing brain.
1478 Mol Psychiatry. 2021 Dec;26(12):7550–9.

1479 121. Nicetto D, Donahue G, Jain T, Peng T, Sidoli S, Sheng L, et al. H3K9me3-heterochromatin
1480 loss at protein-coding genes enables developmental lineage specification. Science. 2019 Jan
1481 18;363(6424):294–7.

1482 122. Antequera F, Bird A. CpG islands as genomic footprints of promoters that are associated
1483 with replication origins. Current Biology. 1999 Sep 9;9(17):R661–7.

1484 123. Horowitz A, Strauss-Albee DM, Leipold M, Kubo J, Nemat-Gorgani N, Dogan OC, et al.
1485 Genetic and environmental determinants of human NK cell diversity revealed by mass
1486 cytometry. Sci Transl Med. 2013 Oct 23;5(208):208ra145.

1487 124. Wilk AJ, Blish CA. Diversification of human NK cells: lessons from deep profiling. J
1488 Leukoc Biol. 2018 Apr;103(4):629–41.

1489 125. Dalton RP, Lyons DB, Lomvardas S. Co-opting the unfolded protein response to elicit
1490 olfactory receptor feedback. Cell. 2013 Oct 10;155(2):321–32.

1491 126. Hetz C, Zhang K, Kaufman RJ. Mechanisms, regulation and functions of the unfolded
1492 protein response. Nature Reviews Molecular Cell Biology. 2020 Aug;21(8):421–38.

1493 127. Smukowski CS, Noor M a. F. Recombination rate variation in closely related species.
1494 Heredity. 2011 Dec;107(6):496–508.

1495 128. Jeffreys AJ, Cotton VE, Neumann R, Lam KWG. Recombination regulator PRDM9
1496 influences the instability of its own coding sequence in humans. Proc Natl Acad Sci U S A.
1497 2013 Jan 8;110(2):600–5.

1498 129. Casola C, Betrán E. The Genomic Impact of Gene Retrocopies: What Have We Learned
1499 from Comparative Genomics, Population Genomics, and Transcriptomic Analyses?
1500 Genome Biology and Evolution. 2017 Jun 1;9(6):1351–73.

1501 130. Lane RP, Young J, Newman T, Trask BJ. Species specificity in rodent pheromone receptor
1502 repertoires. Genome Res. 2004 Apr;14(4):603–8.

1503 131. Trask BJ, Friedman C, Martin-Gallardo A, Rowen L, Akinbami C, Blankenship J, et al.
1504 Members of the olfactory receptor gene family are contained in large blocks of DNA
1505 duplicated polymorphically near the ends of human chromosomes. Human Molecular
1506 Genetics. 1998 Jan 1;7(1):13–26.

1507 132. Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. Human
1508 subtelomeres are hot spots of interchromosomal recombination and segmental duplication.
1509 Nature. 2005 Sep;437(7055):94–100.

1510 133. Mefford HC, Linardopoulou E, Coil D, van den Engh G, Trask BJ. Comparative
1511 sequencing of a multicopy subtelomeric region containing olfactory receptor genes reveals
1512 multiple interactions between non-homologous chromosomes. Hum Mol Genet. 2001 Oct
1513 1;10(21):2363–72.

1514 134. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy
1515 number variation: new insights in genome diversity. Genome Res. 2006 Aug;16(8):949–61.

1516 135. Newman T, Trask BJ. Complex Evolution of 7E Olfactory Receptor Genes in Segmental
1517 Duplications. Genome Res. 2003 May 1;13(5):781–93.

1518 136. Kim J, Farré M, Auvil L, Capitanu B, Larkin DM, Ma J, et al. Reconstruction and
1519 evolutionary history of eutherian chromosomes. PNAS. 2017 Jul 3;114(27):E5379–88.

1520 137. Rouquier S, Taviaux S, Trask BJ, Brand-Arpon V, van den Engh G, Demaille J, et al.
1521 Distribution of olfactory receptor genes in the human genome. Nat Genet. 1998
1522 Mar;18(3):243–50.

1523 138. Yue Y, Haaf T. 7E olfactory receptor gene clusters and evolutionary chromosome
1524 rearrangements. Cytogenet Genome Res. 2006;112(1–2):6–10.

1525 139. Young JM, Endicott RM, Parghi SS, Walker M, Kidd JM, Trask BJ. Extensive copy-
1526 number variation of the human olfactory receptor gene family. Am J Hum Genet. 2008
1527 Aug;83(2):228–42.

1528 140. Paudel Y, Madsen O, Megens HJ, Frantz LAF, Bosse M, Crooijmans RPMA, et al. Copy
1529 number variation in the speciation of pigs: a possible prominent role for olfactory receptors.
1530 BMC Genomics. 2015 Apr 22;16(1):330.

1531 141. Rogers RL. Chromosomal Rearrangements as Barriers to Genetic Homogenization between
1532 Archaic and Modern Humans. Molecular Biology and Evolution. 2015 Dec 1;32(12):3064–
1533 78.

1534 142. North HL, Caminade P, Severac D, Belkhir K, Smadja CM. The role of copy-number
1535 variation in the reinforcement of sexual isolation between the two European subspecies of
1536 the house mouse. Philosophical Transactions of the Royal Society B: Biological Sciences.
1537 2020 Aug 31;375(1806):20190540.

1538 143. Otto M, Zheng Y, Wiehe T. Recombination, selection, and the evolution of tandem gene
1539 arrays. Genetics. 2022 Apr 23;iyac052.

1540 144. Qu Q, Haitina T, Zhu M, Ahlberg PE. New genomic and fossil data illuminate the origin of
1541 enamel. Nature. 2015 Oct 1;526(7571):108–11.

1542 145. Mann RS. Why are Hox genes clustered? Bioessays. 1997 Aug;19(8):661–4.

1543 146. Darbellay F, Bochaton C, Lopez-Delisle L, Mascrez B, Tschopp P, Delpretti S, et al. The
1544 constrained architecture of mammalian Hox gene clusters. Proceedings of the National
1545 Academy of Sciences. 2019 Jul 2;116(27):13424–33.

1546 147. Montavon T, Soshnikova N, Mascrez B, Joye E, Thevenet L, Splinter E, et al. A regulatory
1547 archipelago controls Hox genes transcription in digits. Cell. 2011 Nov 23;147(5):1132–45.

1548 148. Wong ES, Zheng D, Tan SZ, Bower NL, Garside V, Vanwalleghem G, et al. Deep
1549 conservation of the enhancer regulatory code in animals. Science. 2020 Nov
1550 6;370(6517):eaax8137.

1551 149. Markenscoff-Papadimitriou E, Allen WE, Colquitt BM, Goh T, Murphy KK, Monahan K,
1552 et al. Enhancer interaction networks as a means for singular olfactory receptor expression.
1553 Cell. 2014 Oct 23;159(3):543–57.

1554 150. Roy AL, Sen R, Roeder RG. Enhancer–promoter communication and transcriptional
1555 regulation of Igh. Trends in Immunology. 2011 Nov 1;32(11):532–9.

1556 151. Ribich S, Tasic B, Maniatis T. Identification of long-range regulatory elements in the
1557 protocadherin-α gene cluster. PNAS. 2006 Dec 26;103(52):19719–24.

1558 152. Yokota S, Hirayama T, Hirano K, Kaneko R, Toyoda S, Kawamura Y, et al. Identification
1559 of the cluster control region for the protocadherin-beta genes located beyond the
1560 protocadherin-gamma cluster. J Biol Chem. 2011 Sep 9;286(36):31885–95.

1561 153. de Lange T. A loopy view of telomere evolution. Frontiers in Genetics [Internet]. 2015
1562 [cited 2022 Jun 20];6. Available from:
1563 https://www.frontiersin.org/article/10.3389/fgene.2015.00321

1564 154. Zody MC, Garber M, Adams DJ, Sharpe T, Harrow J, Lupski JR, et al. DNA sequence of
1565 human chromosome 17 and analysis of rearrangement in the human lineage. Nature. 2006
1566 Apr;440(7087):1045–9.

1567 155. Riethman H, Ambrosini A, Castaneda C, Finklestein J, Hu XL, Mudunuri U, et al. Mapping
1568 and Initial Analysis of Human Subtelomeric Sequence Assemblies. Genome Res. 2004 Jan
1569 1;14(1):18–28.

1570 156. Giglio S, Broman KW, Matsumoto N, Calvari V, Gimelli G, Neumann T, et al. Olfactory
1571 Receptor–Gene Clusters, Genomic-Inversion Polymorphisms, and Common Chromosome
1572 Rearrangements. Am J Hum Genet. 2001 Apr;68(4):874–83.

1573 157. Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, et al. Comparative
1574 Recombination Rates in the Rat, Mouse, and Human Genomes. Genome Res. 2004 Apr
1575 1;14(4):528–38.

1576 158. Hultén M. Chiasma distribution at diakinesis in the normal human male. Hereditas.
1577 1974;76(1):55–78.

1578 159. Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD.
1579 Recombination initiation maps of individual human genomes. Science. 2014 Nov
1580 14;346(6211):1256442.

1581 160. Tease C, Hultén MA. Inter-sex variation in synaptonemal complex lengths largely
1582 determine the different recombination rates in male and female germ cells. CGR.
1583 2004;107(3–4):208–15.

1584 161. Antunes DMF, Kalmbach KH, Wang F, Dracxler RC, Seth-Smith ML, Kramer Y, et al. A
1585 single-cell assay for telomere DNA content shows increasing telomere length

1586 heterogeneity, as well as increasing mean telomere length in human spermatozoa with
1587 advancing age. J Assist Reprod Genet. 2015 Nov 1;32(11):1685–90.

1588 162. Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A. A model-based analysis of GC-
1589 biased gene conversion in the human and chimpanzee genomes. PLoS Genet.
1590 2013;9(8):e1003684.

1591 163. Huttener R, Thorrez L, in't Veld T, Granvik M, Snoeck L, Van Lommel L, et al. GC
1592 content of vertebrate exome landscapes reveal areas of accelerated protein evolution. BMC
1593 Evolutionary Biology. 2019 Jul 16;19(1):144.

1594 164. Monroe JG, Srikant T, Carbonell-Bejerano P, Becker C, Lensink M, Exposito-Alonso M, et
1595 al. Mutation bias reflects natural selection in Arabidopsis thaliana. Nature. 2022
1596 Feb;602(7895):101–5.

1597 165. McVicker G, Gordon D, Davis C, Green P. Widespread Genomic Signatures of Natural
1598 Selection in Hominid Evolution. PLOS Genetics. 2009 May 8;5(5):e1000471.

1599 166. Murphy D, Elyashiv E, Amster G, Sella G. Broad-scale variation in human genetic diversity
1600 levels is predicted by purifying selection on coding and non-coding elements [Internet].
1601 bioRxiv; 2021 [cited 2022 Sep 1]. p. 2021.07.02.450762. Available from:
1602 https://www.biorxiv.org/content/10.1101/2021.07.02.450762v2

1603 167. Kambere MB, Lane RP. Exceptional LINE density at V1R loci: the Lyon repeat hypothesis
1604 revisited on autosomes. J Mol Evol. 2009 Feb;68(2):145–59.

1605 168. Schoenfelder S, Fraser P. Long-range enhancer–promoter contacts in gene expression
1606 control. Nat Rev Genet. 2019 Aug;20(8):437–55.

1607 169. McKenzie SK, Fetter-Pruneda I, Ruta V, Kronauer DJC. Transcriptomics and
1608 neuroanatomy of the clonal raider ant implicate an expanded clade of odorant receptors in
1609 chemical communication. Proc Natl Acad Sci USA. 2016 Dec 6;113(49):14091–6.

1610 170. Armitage SAO, Freiburg RY, Kurtz J, Bravo IG. The evolution of Dscam genes across the
1611 arthropods. BMC Evol Biol. 2012 Apr 13;12:53.

1612 171. Labrador M, Corces VG. Extensive Exon Reshuffling Over Evolutionary Time Coupled to
1613 Trans-Splicing in Drosophila. Genome Res. 2003 Oct 1;13(10):2220–8.

1614 172. Goeke S, Greene EA, Grant PK, Gates MA, Crowner D, Aigaki T, et al. Alternative
1615 splicing of lola generates 19 transcription factors controlling axon guidance in Drosophila.
1616 Nat Neurosci. 2003 Sep;6(9):917–24.

1617 173. Venables JP, Tazi J, Juge F. Regulated functional alternative splicing in Drosophila.
1618 Nucleic Acids Res. 2012 Jan;40(1):1–10.

1619 174. Schield DR, Card DC, Hales NR, Perry BW, Pasquesi GM, Blackmon H, et al. The origins
1620 and evolution of chromosomes, dosage compensation, and mechanisms underlying venom
1621 regulation in snakes. Genome Res. 2019 Apr;29(4):590–601.

1622 175. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al.
1623 The genome of the African trypanosome Trypanosoma brucei. Science. 2005 Jul
1624 15;309(5733):416–22.

1625 176. Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, et al. caper: Comparative
1626 Analyses of Phylogenetics and Evolution in R [Internet]. 2018 [cited 2023 Mar 22].
1627 Available from: https://cran.r-project.org/web/packages/caper/index.html

1628 177. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines,
1629 Timetrees, and Divergence Times. Molecular Biology and Evolution. 2017 Jul
1630 1;34(7):1812–9.

1631 178. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary
1632 analyses in R. Bioinformatics. 2019 Feb 1;35(3):526–8.

1633 179. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
1634 features. Bioinformatics. 2010 Mar 15;26(6):841–2.

1635 180. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a
1636 next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016 Jul
1637 8;44(Web Server issue):W160–5.

1638 181. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment
1639 tool for interpreting omics data. The Innovation. 2021 Aug 28;2(3):100141.

1640 182. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access
1641 to ontology and annotation data. Bioinformatics. 2009 Jan 15;25(2):288–9.

1642 183. Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, et al. Genenames.org: the
1643 HGNC and VGNC resources in 2021. Nucleic Acids Research. 2021 Jan 8;49(D1):D939–
1644 46.

1645 184. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of
1646 genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4(8):1184–
1647 91.

1648 185. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for
1649 Computing and Annotating Genomic Ranges. PLOS Computational Biology. 2013 Aug
1650 8;9(8):e1003118.

1651
1652