A New Chromosome-Assigned Mongolian Gerbil Genome Allows Characterization of Complete Centromeres and a Fully Heterochromatic Chromosome

Thomas D. Brekke[1], Alexander S.T. Papadopulos[1], Eva Julià[2], Oscar Fornas[2,3], Beiyuan Fu[4], Fengtang Yang[5], Roberto de la Fuente[6], Jesus Page[7], Tobias Baril[8], Alexander Hayward[8], and John F. Mulley[1,*]

[1]School of Natural Sciences, Bangor University, Gwynedd, United Kingdom

[2]Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology (BIST), Barcelona, Spain

[3]Flow Cytometry Unit, Pompeu Fabra University, Barcelona, Spain

[4]Cambridge Epigenetix, Cambridge, United Kingdom

[5]Present address: School of Life Sciences and Medicine, Shandong University of Technology, Zibo, China

[6]Department of Experimental Embryology, Institute of Genetics and Animal Biotechnology of the Polish Academy of Sciences, Magdalenka, Poland

[7]Departamento de Biología, Facultad de Ciencias, Universidad Autónoma de Madrid, Madrid, Spain

[8]Centre for Ecology and Conservation, University of Exeter, Cornwall, United Kingdom

***Corresponding author:** E-mail: j.mulley@bangor.ac.uk.

Associate editor: Amanda Larracuente

Abstract

Chromosome-scale genome assemblies based on ultralong-read sequencing technologies are able to illuminate previously intractable aspects of genome biology such as fine-scale centromere structure and large-scale variation in genome features such as heterochromatin, GC content, recombination rate, and gene content. We present here a new chromosome-scale genome of the Mongolian gerbil (*Meriones unguiculatus*), which includes the complete sequence of all centromeres. Gerbils are thus the one of the first vertebrates to have their centromeres completely sequenced. Gerbil centromeres are composed of four different repeats of length 6, 37, 127, or 1,747 bp, which occur in simple alternating arrays and span 1–6 Mb. Gerbil genomes have both an extensive set of GC-rich genes and chromosomes strikingly enriched for constitutive heterochromatin. We sought to determine if there was a link between these two phenomena and found that the two heterochromatic chromosomes of the Mongolian gerbil have distinct underpinnings: Chromosome 5 has a large block of intraarm heterochromatin as the result of a massive expansion of centromeric repeats, while chromosome 13 is comprised of extremely large (>150 kb) repeated sequences. In addition to characterizing centromeres, our results demonstrate the importance of including karyotypic features such as chromosome number and the locations of centromeres in the interpretation of genome sequence data and highlight novel patterns involved in the evolution of chromosomes.

***Key words: Meriones*, genome, karyotype, centromeres, chromosome evolution.**

Article

Introduction

Understanding the organization and function of genomes and how they vary has been an important goal in the field of biology since at least the 1950s. The new and relatively inexpensive long-range sequencing technologies such as PacBio HiFi and Oxford Nanopore are facilitating the sequencing and chromosome-scale assembly of the genomes of many new species (Jayakumar and Sakakibara 2019). Such high-quality genomes are an important tool to address long-standing questions about variation in the structure and function of genomes across the tree of life. Such questions include the following: What is the nucleotide sequence and structure of centromeres in nonmodel species? What is the recombination landscape, and how does it influence nucleotide content variation in genes and along chromosomes? In addition and often overlooked, what new insights can be gleaned when we reinterpret cytological data, such as the banding patterns of chromosomes in a karyotype, in light of chromosome-scale assemblies?

Centromeres are crucially important during mitosis and meiosis. Functionally, they are the binding site of centromere-specific histones and other proteins that facilitate their binding to the spindle apparatus (McKinley and Cheeseman 2016). They are visible in karyotypes as constrictions in the chromosome, which stain very darkly under different chemical treatments (Willard 1990). They are characterized by arrays of various repeated sequences of DNA of various lengths, where the sequence of the repeat is species specific (Talbert and Henikoff 2020). Due to their size and repetitive nature, they have proven intractable to assembly by all but the most recent of long-range sequencing technologies; indeed, it is only within the last year that human centromeres have been completely sequenced and annotated (Altemose et al. 2022). An immense amount of work has gone into studying centromeres at the functional level using visualization techniques, but very little is known about the specific sequence of centromeres in most species. Sequencing and characterizing centromeres in various nonmodel species are and will be an important addition to understanding the variation and function of centromeres across the tree of life.

The nucleotide composition of genomes is not homogenous; it varies along chromosome arms and between chromosomes, individuals, populations, and species (Eyre-Walker and Hurst 2001). Variation in the distribution of guanine (G) and cytosine (C) bases is heavily determined by the recombination-associated process of GC-biased gene conversion (gBGC), which favors fixation of G and C over adenine (A) and thymine (T) (Lamb 1984; Arbeithuber et al. 2015). Over evolutionary time, this process results in a GC bias around recombination hotspots (Galtier et al. 2001). Gerbils and their relatives have multiple extensive regions of extremely high GC bias within their genomes, higher than that seen in any other mammal (Hargreaves et al. 2017; Dai et al. 2020; Pracana et al. 2020). Historically, this has complicated attempts to obtain high-quality contiguous gerbil genome assemblies (Leibowitz et al. 2001; Gustavsen et al. 2008). Intriguingly, there appear to be two distinct patterns of GC skew in gerbils: 1) a region associated with the ParaHox cluster and the surrounding genes, where virtually all genes in this region have very high mutation rates and an extreme GC bias, and 2) a further set of 17 large clusters of GC-rich genes also with high mutation rates (Pracana et al. 2020). These intriguing characteristics of gerbil genomes make them an ideal system in which to examine the association between gBGC and the organization of eukaryotic genomes.

Chromatin state is an important mechanism for the regulation of gene activity. Facultative heterochromatin is cell type specific and may be converted to open, active euchromatin during gene regulatory processes. In contrast, constitutive heterochromatin is marked by trimethylation of histone H3 at the lysine 9 residue (H3K9me3) (Saksouk et al. 2015) and comprises densely compacted, gene-poor inactive regions of the genome, which are condensed in all cell types at all developmental stages, such as centromeres and telomeres (Saksouk et al. 2015; Penagos-Puig and Furlan-Magaril 2020). Many gerbil species (family Gerbillidae) have chromosomes with high levels of constitutive heterochromatin, though the specific chromosome and extent of heterochromatin vary by species. Mongolian gerbils possess distinctive karyotypic features (supplementary fig. S1, Supplementary Material online): Nearly a third of chromosome 5 and all of chromosome 13 appear to be composed of constitutive heterochromatin by multiple different assays: Chromosome 13 stains entirely dark in C-banding stains (Gamperl and Vistorin 1980) and is completely coated by heterochromatin histone marks in immunofluorescence assays (Gamperl and Vistorin 1980; de la Fuente et al. 2014) The genomes of the North African Gerbil (*Gerbillus campestris*), the hairy-footed gerbil (*Gerbilliscus paeba*), and the fat sand rat (*Psammomys obesus*) all contain a single heterochromatic chromosome (Solari and Ashley 1977; Gamperl and Vistorin 1980; Knight et al. 2013).

The heterochromatic chromosomes in gerbils are present in all individuals examined to date and do not meet the criteria for classification as B chromosomes, that is, they are not nonessential, and do not vary in copy number among individuals and tissues without an adverse impact on fitness (Ahmad and Martins 2019). These chromosomes therefore provide a unique system to examine the impact of their heterochromatic state on genic evolution and particularly whether it is linked to the extensive number of GC-rich genes in gerbil genomes. Heterochromatin is typically gene poor (Dimitri et al. 2005) and transcriptionally repressed (Grewal and Moazed 2003; Dillon 2004). This makes it unlikely that entire heterochromatic chromosomes would be maintained and transmitted across generations for millions of years if they did not encode any genes or are entirely selfish independent genetic elements. High GC% in certain gerbil genes could be an adaptation to a transcriptionally repressive environment. Genes with high GC% in their coding regions and adjacent regions of DNA and especially those with high GC% in the third codon position (GC_3) can show elevated expression (Lercher et al. 2002; Vinogradov 2005). Conversely, since gBGC is a recombination-dependent process and since all chromosomes must undergo at least one reciprocal recombination event (crossover) with their homologue during meiosis (Lydall et al. 1996), an alternative hypothesis is that the extreme GC% present in some gerbil genes is a consequence their becoming entrapped in or near a recombination hotspot. If the bulk of the extensive heterochromatin observed on these gerbil chromosomes is nonpermissive to recombination, then genes in those regions where recombination can occur will become increasingly GC rich because of continual exposure to gBGC. We may therefore reasonably expect a link between GC-rich genes and these unusual gerbil chromosomes.

A key question is how did fully heterochromatic chromosomes in gerbils arise? They may once have been “normal” chromosomes that have degenerated into gene-poor, nonfunctional, or silenced chromosomes by accumulation of repetitive DNA. Alternatively, they may have formed from heterochromatic pieces that broke off from other chromosomes, in the same way that the neochromosomes of tumors (Garsed et al. 2009, 2014) and some B chromosomes (Camacho et al. 2000; Dhar et al. 2002) develop from fragments of other chromosomes. Alternatively, they could be the duplicate of another chromosome, which condensed into heterochromatin a mechanism of dosage compensation in the same way that additional copies of X chromosomes are inactivated in female mammals (Lyon 1962). Finally, they may potentially have grown from a smaller chromosomal "seed," which broke off from another chromosome and subsequently grew by repeated segmental duplication.

Until very recently, questions such as those posed above could not be addressed in a nonmodel system for several key reasons. A particularly important issue was the difficulties that short read-based genome sequencing approaches face regarding the assembly of GC%-rich regions (Hron et al. 2015; Bornelöv et al. 2017; Botero-Castro et al. 2017; Tilak et al. 2018; Yin et al. 2019). Meanwhile, the current trend toward the generation of chromosome-scale assemblies has perhaps lost sight of the importance of an understanding of the karyotype of the species being studied and of physically linking genome sequence to identified chromosomes.

Using a new chromosome-scale genome assembly for the Mongolian gerbil and methods enabling us to assign the genomic scaffolds to physical chromosomes, we first characterize gerbil centromeres and then test 1) whether GC-rich gene clusters correlate with recombination hotspots and 2) if those genes are associated with a single heterochromatic chromosome. Our approach allows us to examine the origin and propose a new hypothesis for the evolution of some unusual and possibly unique, heterochromatic gerbil chromosomes.

Results and Discussion

Gerbil Genome: Approach and Summary Statistics

We sequenced and assembled the genome of the Mongolian gerbil, *Meriones unguiculatus*, into 245 contigs using PacBio HiFi reads (2,699,742,000 total bases, $N_{50} = 58{,}726{,}396$, $L_{50} = 16$, $N_{90} = 15{,}971{,}047$, $L_{90} = 48$). We scaffolded the contigs with OmniC chromatin conformation capture data (supplementary fig. S2, Supplementary Material online), Oxford Nanopore long and ultralong-read sequence data, a genetic map (supplementary table S1, Supplementary Material online) (Brekke et al. 2019), and BioNano optical mapping. We assigned scaffolds to chromosomes by flow-sorting chromosomes into pools. Each pool was sequenced with Illumina short reads, and these reads used to determine which scaffolds associated with each pool. The sorted pools were also made into fluorescence in-situ hybridization (FISH) paint probes to identify which physical chromosome from the karyotype associated with each pool. This approach linked the physical chromosomes with sequenced scaffolds (full methods are in supplementary material S1, figs. S1 and S3–S7, and tables S2 and S3, Supplementary Material online). The final genome assembly contains 194 scaffolds spanning 21 autosomes, the X and Y sex chromosomes, and the mitochondrial genome (supplementary table S4, Supplementary Material online). For 20 of the 23 chromosomes, a single large scaffold contains over 94% (often over 99%) of all the sequence assigned to that chromosome (fig. 1A). Only chromosome 13, with 121 scaffolds, and the X and Y chromosomes, each with six scaffolds, are appreciably fragmented, and there are only 30 unassigned scaffolds making up 0.066% of the sequenced bases (fig. 1B). The assembly was annotated using RNAseq data and is 92% complete based on a BUSCO analysis (complete: 92.3% [single-copy: 91.7%; duplicated: 0.6%]; fragmented: 1.7%; missing: 6.0%; *n*: 13,798) (Manni et al. 2021). We used the program NeSSie (Berselli et al. 2018) to calculate two measures of sequence complexity (entropy and linguistic complexity) in sliding windows across every chromosome. The complexity metrics revealed two chromosomes with unusual features: chromosome 5 that has an extensive region where both entropy and linguistic complexity are very low and chromosome 13 that shows a marked homogeneity in its entropy across the length of the chromosome (fig. 2 and supplementary fig. S8, Supplementary Material online). A chromosome-by-chromosome summary of all data is found in figure 3, and a high-resolution version is in supplementary material S2, Supplementary Material online.

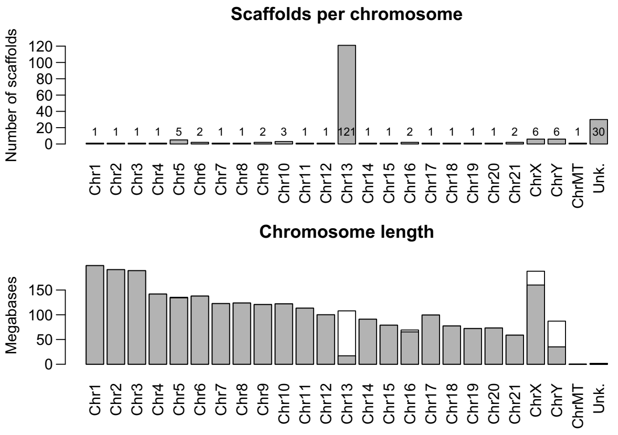

Fig. 1. Summary statistics for the Mongolian gerbil (*M. unguiculatus*) genome assembly. Top: The number of scaffolds assigned to each chromosome, the mitochondrial genome, and the "unknown" category. Most chromosomes are assembled into one or two scaffolds, while chromosome 13 is in 121 pieces. Bottom: The number of bases assigned to each chromosome with the single longest scaffold shaded in gray. The total amount of DNA sequence assigned to chromosome 13 is about what would be expected, showing that we are not missing data, and that the large number of scaffolds is not an artifact.

Two *M. unguiculatus* genome sequences have been previously published, based on short-read sequence data (Cheng et al. 2019; Zorio et al. 2019); both contain hundreds of thousands of contigs and equally large numbers of scaffolds (supplementary table S4, Supplementary Material online). One of these has recently been improved with Hi-C data (www.DNAZoo.org) into 22 chromosome-length scaffolds and ~300,000 additional scaffolds (Cheng

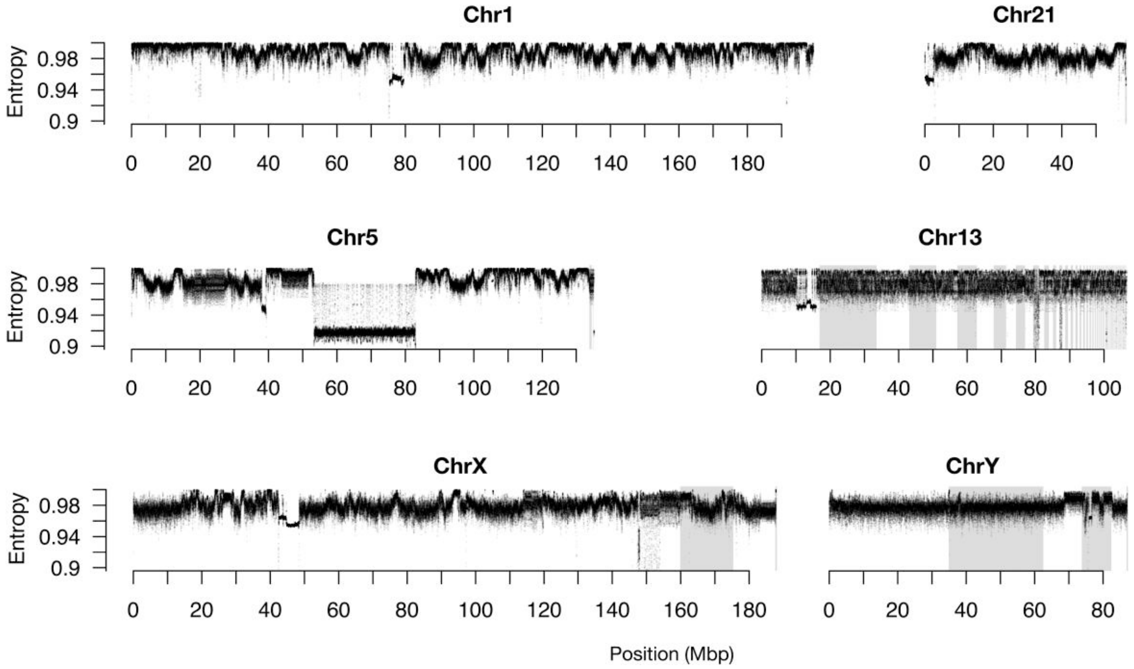

Fig. 2. Entropy plots for a selection of chromosomes including the morphologically standard autosomes 1 and 21, the unusual autosomes 5 and 13, and both sex chromosomes. The unordered scaffolds within a chromosome are shaded alternately white and gray. Centromeres are apparent at ~75–80 Mbp in Chr1, ~0–5 Mbp in Chr21, ~35–40 Mbp in Chr5, ~14–17 Mbp in Chr13, ~45–50 Mbp in ChrX, and ~75 Mbp in ChrY. Note the spatial heterogeneity in chromosomes 1 and 21 that is absent in chromosome 13 and the Y. Indeed, chromosome 13 is the most homogenous chromosome in the gerbil. Entropy plots for every chromosome, as well as GC content, gene density, and linguistic complexity can be found in supplementary figure S2, Supplementary Material online.

et al. 2019). Full-genome alignments between our genome assembly and this Hi-C assembly (supplementary fig. S9, Supplementary Material online) showed that most scaffolds are collinear between the assemblies but that the "improved" Cheng et al. (2019) assembly lacks chromosome 13 entirely, hence only 22 chromosome-scale scaffolds for this species with 23 unique chromosomes (21 autosomes, an X and a Y; supplementary fig. S1, Supplementary Material online). Our highly contiguous and physically associated assembly provides the foundation for all subsequent analyses.

Characterization of Gerbil Centromeres

Relatively little is known about centromere organization in nonmodel species, as centromeres are comprised of extensive runs of repeated sequences, which short-read technologies (and even Sanger sequencing) have struggled to cross. It is only this year that full coverage of human centromeres was obtained, from a mixture of long-read sequencing approaches applied to the genome of a hydatidiform mole cell line by the telomere-to-telomere (T2T) consortium (Altemose et al. 2022). Our high-quality PacBio HiFi-derived sequence data resulted in a single large scaffold per chromosome (for all but a few chromsomes), which spanned from telomere to telomere (fig. 1). Such completeness suggested that we sequenced through the centromeres of all *M. unguiculatus* chromosomes and so we set about bioinformatically identifying centromeres. Centromeres are known to be highly repetitive, occur once on each chromosome, are visually apparent as a constriction in the karyotype, and are typically on the order of a few megabases long (Talbert and Henikoff 2020). We used the entropy and linguistic complexity metrics (measures of sequence repetitiveness) to reveal a region of each chromosome that matched the above predictions: Every chromosome has a single highly repetitive region ranging from ~1 to 10 Mbp long, which lines up with the constriction in the karyotype (figs. 2 and 3). As we did no molecular assay for centromere function, we submit these as "putative centromeres," though for brevity, we hereafter refer to them simply as "centromeres."

To further characterize the gerbil centromeres, we used the program NTRprism (Altemose et al. 2022), which identified four different simple repeat sequences (supplementary fig. S10, Supplementary Material online). We have named these "MsatA" (for Meriones satellite A), "MsatB," "MsatC," and "MsatD" (fig. 3A): MsatA is 6 bp long and has the sequence TTAGGG, which is the same simple sequence repeat found in telomeres, MsatB is 37 bp long, MsatC is 127 bp long, and MsatD is 1,747 bp long and is only found on the Y chromosome. A representative sequence of each Msat can be found in the legend of supplementary figure S10, Supplementary Material online. At the time of writing, MsatB and MsatC return no Blast hits from NCBI's "nt" library (update 2023/01/12) and MsatD returns a single 32-bp run of identity (out of 1,748 bp) with *Acomys russatus* suggesting

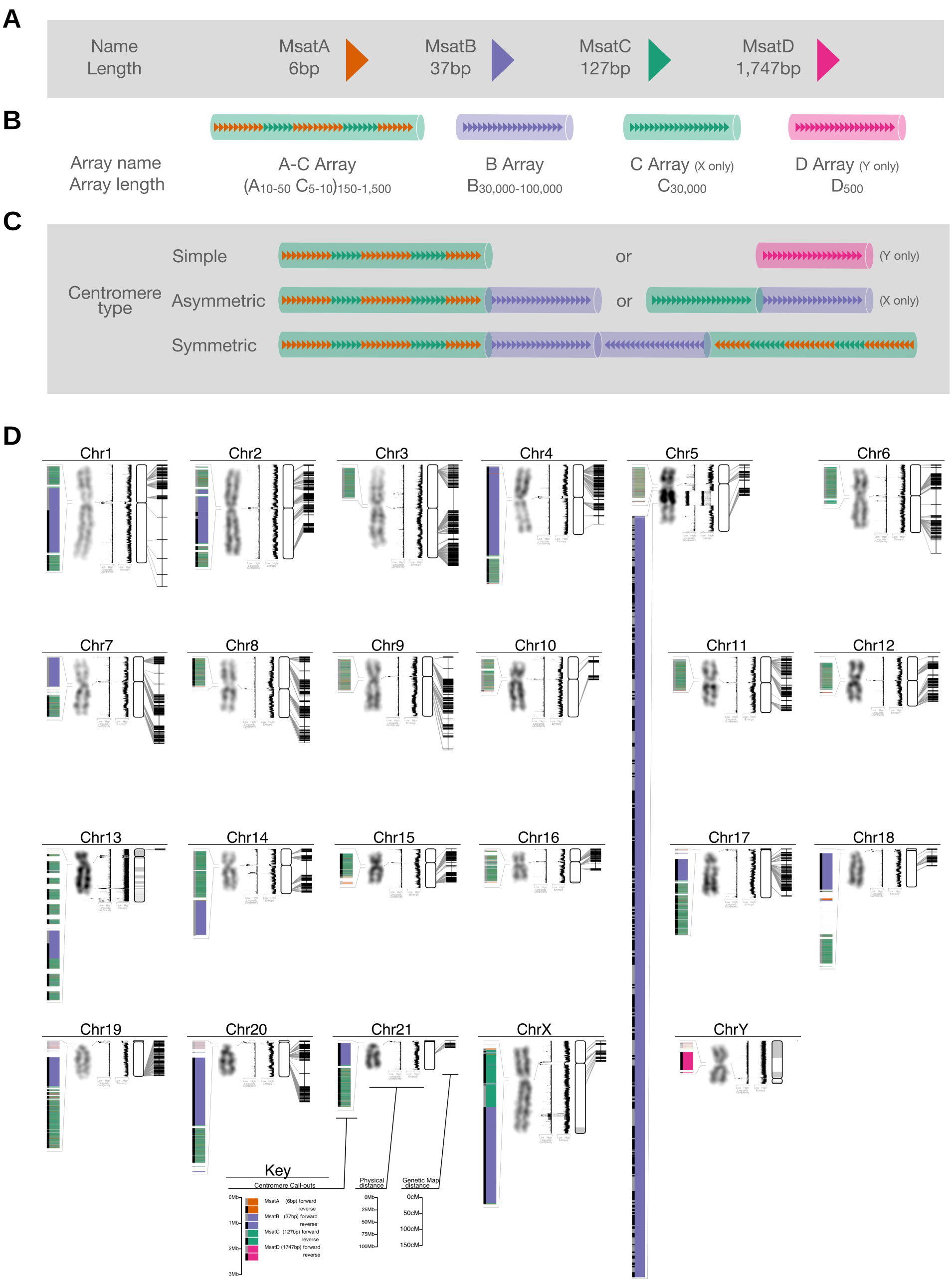

Fig. 3. The Mongolian gerbil (*M. unguiculatus*) genome. Gerbil centromere types. (*A*) There are four different repeat types in gerbil centromeres: MsatA (6 bp), MsatB (37 bp), MsatC (127 bp), and MsatD (1,747 bp). (*B*) These repeats appear in one of four repeat arrays. The A–C array consists of 10–50 copies of MsatA alternating with 5–10 copies of MsatC, all of which is repeated 150–1,500 times. The B, C, and D arrays contain only multiple copies of their respective repeat. Repeat units within an array most often occur in the same orientation. In some chromosomes, however, both orientations occur within a single array, in which case hundreds of repeat units in the forward orientation are followed by hundreds of units in the reverse orientation (e.g., the B array of chromosomes 1, 2, 13 and the X). (*C*) Centromeres consist of between one and three repeat arrays and are classed as either "simple," "asymmetric," or "symmetric." Simple centromeres have a single array type, either an A–C array as in the autosomes or a D array as on the Y chromosome. Asymmetric centromeres have two arrays: either an A–C array and a B array (for the autosomes) or a C array and a B array (for the X chromosome). Symmetric centromeres consist of three arrays, a B array sandwiched between two A–C arrays that typically appear in opposite orientation to each other. (*D*) Genome schematic, for each chromosome we show, from left to right: (1) centromere organization, with repeats of different lengths in different colors and the orientation of the repeat array denoted by a gray or black bar on the left. Chromosome 5 has a large expansion of centromeric repeats in the long arm. All callouts are drawn to the same scale. (2) The DAPI-banding karyotype image, showing the intraarm heterochromatin on chromosome 5, and the entirely dark staining on chromosome 13. (3) Linguistic complexity and (4) entropy, both measured in overlapping sliding 10-kb windows with a step size of 1 kb. For both metrics, a low value indicates highly repetitive or predictable sequence as are characteristics of centromeres while high values indicate more complex sequence as may be found in gene-rich regions. (5) A depiction of the physical map with unplaced scaffolds organized by length and shaded alternately white and gray, and (6) a depiction of the genetic map with links between the genetic markers and their physical location. Thin gray lines link the location of similar features on adjacent plots (i.e., centromere callout to karyotype; centromere location in the karyotype to centromere in the linguistic complexity plot; and genetic markers to their physical location). A high-resolution copy of panel *D* can be found in the supplementary material S2, Supplementary Material online.

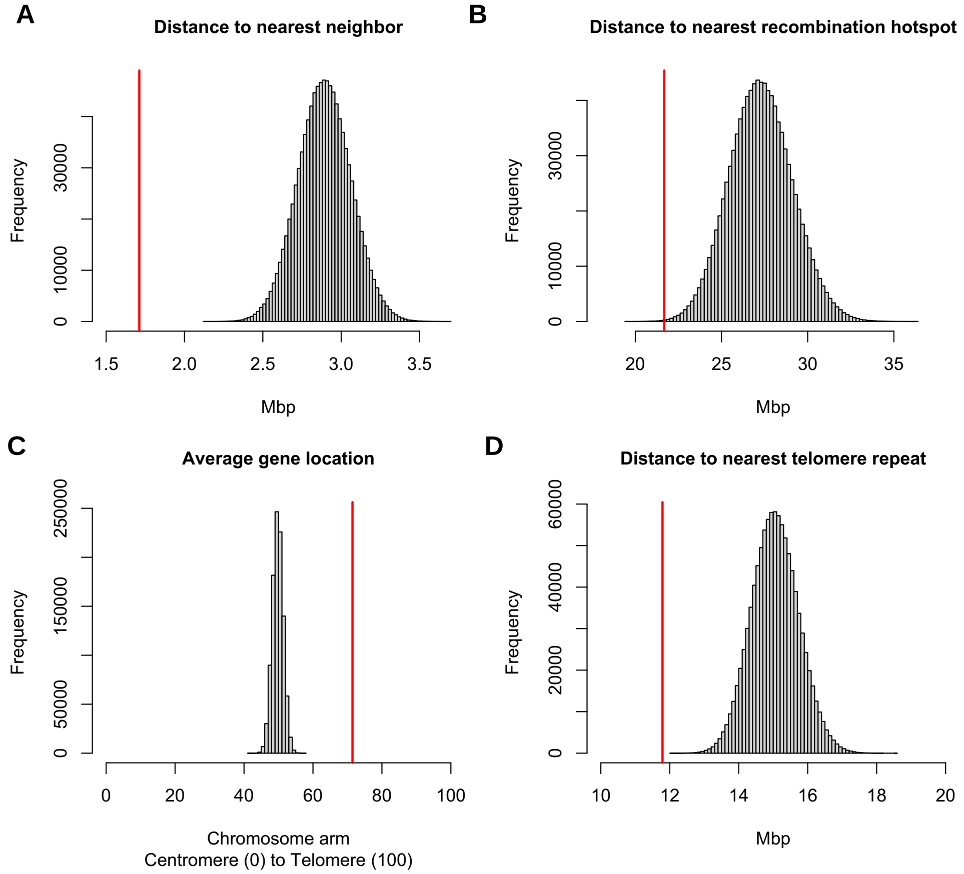

Fig. 4. GC-rich genes are nonrandomly distributed in the *M. unguiculatus* genome. We compared the location of the 410 GC-rich genes (Pracana et al. 2020) in relation to each other, the nearest recombination hotspot, their location along the chromosome arm, and their proximity with telomere repeats both interstitial and at the ends of chromosome arms. These comparisons were done once against the entire gene set (supplementary figs. S14, S15, and S17, Supplementary Material online) and here using a permutation test with 1,000,000 draws of a random set of 410 genes where the vertical (red) line indicates the observed value. (*A*) GC-rich genes are clustered in the genome. The observed distance between each outlier gene and its nearest outlier gene neighbor is significantly shorter than those distances between a random group of genes (observed = 1.71 Mbp, mean = 2.89 Mbp, *n* = 1,000,000 permutations, $P < 0.000001$). (*B*) GC-rich genes occur closer to recombination hotspots than expected by chance (observed = 21.68 Mbp, mean = 27.29 Mbp, *n* = 1,000,000, $P < 0.00058$). (*C*) GC-rich genes are found closer to the telomere end of chromosome arms than expected by chance (observed = 71.46%, mean = 49.78%, *n* = 1,000,000, $P < 0.000001$). (*D*) GC-rich genes are clustered nearer telomere repeats (interstitial or otherwise) than expected by chance (observed = 11.79 Mbp, mean = 15.06 Mbp, *n* = 1,000,000, $P < 0.000001$).

that these Msats are new sequences not previously identified.

Copies of Msats are arranged into one of four variant arrays that define an intermediate-order structure in the centromeres (fig. 3*B*). "B arrays" are formed from copies of MsatB and range in size from 1 to 3 Mbp long (∼30,000–100,000 copies). Similarly, the Y chromosome centromere is a "D array" comprised of ∼500 copies of MsatD spanning slightly less than a megabase. MsatA and MsatC repeats are rarely found alone, tending instead to intersperse with each other to form "A–C arrays." Typically 10–50 copies of MsatA will alternate with 5–10 copies of an MsatC unit, and this alternating pattern will extend for between 100 kb and 1 Mb depending on the chromosome. The only place that MsatC is found without interspersed copies of MsatA is on the X chromosome in what we term a "C array." While not interspersed with MsatC, there are a number of MsatA repeats that do appear at both ends of the X centromere and are detectable by FISH (de la Fuente et al. 2014). The orientation of the Msat repeats is typically consistent across an array; however, some arrays are composed of blocks of Msat repeats in alternating orientations with many copies of repeat in the forward orientation followed by many copies in the reverse orientation.

The highest level of centromere organization is characterized by groups of between one and three arrays that fall into one of a few patterns which we term "simple," "asymmetric," or "symmetric" (fig. 3C). Simple centromeres are comprised of a single A–C array and are present in ten of the smaller metacentric chromosomes (see chromosomes 3, 5, 6, 8–12, 15, and 16 in fig. 3*D*). The metacentric Y chromosome also has a simple centromere, though with a D array instead of the A–C array. Asymmetric centromeres are comprised of two arrays, one of which is always a B array and the other is typically an A–C array. Eight autosomes fall into this category including all four of the small telocentric chromosomes (chromosomes 18–21 in fig. 3D), three of the metacentric chromosomes (chromosomes 4, 7, and 14 in fig. 3D), and one acrocentric chromosome (chromosome 17 in fig. 3D). The metacentric X chromosome also has an asymmetric centromere but is the only location in the genome where a pure C array exists. Finally, symmetric centromeres are comprised of three arrays: A C array is sandwiched between two A–C arrays and is found in the metacentric chromosomes 1 and 2 and the acrocentric chromosome 13. Many centromeres also contain 10–50 kbp blocks of nonrepetitive, complex DNA both between and within the various arrays (see fig. 3D).

The Location of GC-Rich Genes

A set of over 380 genes with extreme GC content clustered in the genomes of sand rats and gerbils has previously been identified (Pracana et al 2020). It has been hypothesized that biased gene conversion has driven their GC content to extraordinary levels since they are near recombination hotspots (Pracana et al. 2020), but the resources to test this were not available so mouse gene locations had been used as an evolutionarily informed proxy for the location of those genes in gerbils. Here, we use our newly generated chromosome-scale assembly to explicitly test how these GC-rich genes are distributed across gerbil chromosomes. We used a permutation test to show that GC-rich genes are clustered together more than is expected by chance (fig. 4A; observed = 1.71 Mbp, mean = 2.89 Mbp, $n = 1{,}000{,}000$ permutations, $P < 0.000001$). We used our genetic map (Brekke et al. 2019) to locate recombination hotspots that were defined as regions with 5× higher recombination rate than the genome average (as per Katzer et al. 2011). Hotspots were found on 18 of 22 chromosomes (21 autosomes and the X chromosome, we omit the Y chromosome here as it does not recombine) with 2.4 ± 2.2 (SD) hotspots per chromosome (supplementary figs. S11–S13, Supplementary Material online). Chromosomes 2, 18, 21, and the X lack recombination hotspots. We tested proximity of GC-rich genes to hotspots in two ways, first by comparing the GC-rich genes with the entire gene set (supplementary fig. S14, Supplementary Material online) and second by performing a permutation test (fig. 4B). In both cases, GC outlier genes were found to lie significantly closer to recombination hotspots than expected by chance (fig. 4B; observed = 21.68 Mbp, mean = 27.29 Mbp, $n = 1{,}000{,}000$, $P < 0.00058$). These results demonstrate a clear association of GC-rich gene clusters with recombination hotspots as expected under gBGC.

While a genetic map shows the location of current recombination hotspots, hotspots move through evolutionary time due to large-scale chromosomal rearrangements and the mutational load caused by crossing over (Paigen and Petkov 2010; Tiemann-Boege et al. 2017). Consequently, we next tested whether GC outliers are associated with proxies of ancestral hotspots. Recombination rate is not uniform across a chromosome and is typically higher near the telomeres (Nachman 2002; Martinez-Perez and Colaiácovo 2009); thus, we tested whether GC outliers are correlated with position along the chromosome arm. We found that whether considering the full distribution of gene locations (supplementary fig. S15, Supplementary Material online) or 1,000,000 draws of the same number of random genes in a permutation test (fig. 4C), the GC outliers are found to lie much closer to the telomere than expected by chance (fig. 4C; observed = 71.46%, mean = 49.78%, $n = 1{,}000{,}000$, $P < 0.000001$). Furthermore, gerbils have many interstitial telomere sites (de la Fuente et al. 2014) that are caused by chromosomal fusions embedding what was an ancestral telomere within a chromosome arm, typically near the centromere. Thus, interstitial telomere repeats are proxies for the ends of ancestral chromosomes and their associated ancient recombination hotspots. We identified interstitial telomere sites as arrays of the 6 bp "MsatA" with at least 70 tandem copies (supplementary fig. S16, Supplementary Material online). We therefore tested whether GC outlier genes are closer to these telomere repeats (which could be interstitial or otherwise) than expected by chance and found that they are (fig. 4D; observed = 11.79 Mbp, mean = 15.06 Mbp, $n = 1{,}000{,}000$, $P < 0.000001$; supplementary fig. S17, Supplementary Material online). In short, GC outlier genes are found in clusters across the genome and are nearer to recombination hotspots (current or ancient) and telomere/interstitial telomere sites than expected by chance, strongly supporting the hypothesis that gBGC is driving the extreme GC content of these genes. Supplementary figure S18, Supplementary Material online shows the distribution of centromeres, recombination hotspots, high GC genes, and telomere sites that were used in these analyses.

However, we did not find that all GC-rich genes are located on heterochromatic chromosomes and find instead that they are distributed on the order of 19.5 ± 13.7 GC-rich genes per chromosome across the genome. The tendency for genes to become highly GC rich in and around recombination hotspots in gerbils therefore seems unrelated to their unusual chromosomes and may instead be the result of greater recombination hotspot stability, where hotspots stay in one place for longer in gerbils compared with other species. Similarly stable hotspot location has previously been reported for birds (Singhal et al. 2015) though in birds the absence of PRDM9 correlates with greater hotspot stability. The gerbil genome encodes a full-length *Prdm9* gene on chromosome 20, and so this hotspot stability in gerbils must arise via some other mechanism.

We next sought to understand the genomic basis of the heterochromatic appearance of the chromosomes 5 and 13 in *M. unguiculatus*.

Chromosome 5: The Relevance of Centromeric Drive

Chromosome 5 is characterized by an enormous centromeric repeat expansion that is visible as a dark band on the q arm (fig. 3D). Our data show that the repeat expansion is a 29-Mb-long B array, which comprises ∼22% of the entire chromosome. This repeat expansion is distinct from the centromere that is a simple A–C array 1.5 Mb long. In

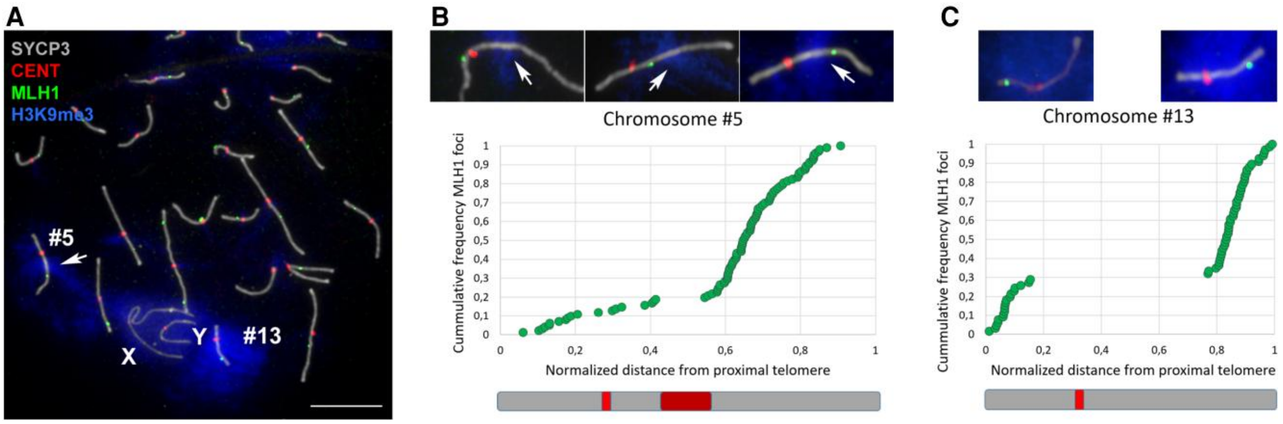

Fig. 5. Distribution of recombination events in gerbil spermatocytes. Scale bar is 10 *μ*m. (*A*) Immunolocalization of SYCP3 protein (gray) on meiotic chromosomes marks the trajectory of the synaptonemal complex along bivalents; trimethylation of H3K9me3 (blue) marks heterochromatin; CENT (red) stains centromeres; and MLH1 (green) marks the sites of crossovers. H3K9me3 is associated with the entirety of chromosome 13 (#13), a large intraarm region of chromosome 5 (#5), and, to a lesser extent, the X and Y. The anti-CENT antibody (red) stains centromeres on all chromosomes but is not specifically associated with the large centromeric expansion of the long arm of chromosome 5. MLH1 foci can be located proximally, interstitially, or distally along bivalent 5 (central details, selected from three different spermatocytes), but they are never found within the centromere repeat expansion on this chromosome. Chromosome 13 shows either proximal or distal location of MLH1 foci (details on the right). (*B* and *C*) Graphs of MLH1 frequency against distance from the nearest telomere for bivalents 5 and 13, respectively. Each dot represents the location of the MLH1 focus along the synaptonemal complex on a single spermatocyte. The locations of centromeres and the chromosome 5 expansion are indicated as red and maroon boxes, respectively, on the schematic chromosomes below each graph. The graphs and drawings preserve the relative size of both chromosomes. For chromosome 5, most crossovers (over 80%) are located from the heterochromatic expansion toward the distal end. For chromosome 13, MLH1 foci are conspicuously accumulated toward the chromosomal ends, with an approximate 70:30 distribution on the long and short arms, respectively.

contrast to the B arrays in the centromeres of other chromosomes, the orientation of MsatB repeats on chromosome 5 switches far more frequently. With a few exceptions, B arrays in centromeres maintain their orientation across the entire array, or in the case of the symmetric centromeres, have a few large blocks in opposite orientations; the centromeric B arrays maintain orientation for 1–3 Mb. Repeats in the chromosome 5 expansion, however, switch orientation over 200 times across the 29 Mb, so the average block length is just 140 kb.

There is a similar large expansion of a centromeric repeat found in human chromosome 9 (Altemose et al. 2022). However, although it is similar in size to the expansion on gerbil chromosome 5, the human expansion is polymorphic in the population (Craig-Holmes and Shaw 1971). The dark band on the q arm of gerbil chromosome 5 is visible in all published karyotypes dating back to the 1960s, which derive from many different individuals and laboratory colonies (Pakes 1969; Weiss et al. 1970; Gamperl and Vistorin 1980) suggesting that in contrast, the gerbil expansion is fixed at this massive size.

The repeat expansion is absent in karyotypes of many closely related Gerbillinae species, including representatives from the genera *Desmodillus*, *Gerbillurus*, *Gerbillus*, *Tatera*, and *Taterillus*, and is even absent in other species of *Meriones*. (Gamperl and Vistorin 1980; Benazzou et al. 1982, 1984; Qumsiyeh 1986a, 1986b; Dobigny et al. 2002; Aniskin et al. 2006; Volobouev et al. 2007; Gauthier et al. 2010). The expansion is also absent in the sequenced genome assemblies of the closely related fat sand rat (*P. obesus*) and fat-tailed gerbil (*Pachyuromys duprasi*). Alignment with the *Psammomys* genome assembly shows that the location of the repeat expansion on *M. unguiculatus* chromosome 5 is homologous to the *Psammomys* chromosome 10 centromere (supplementary fig. S19, Supplementary Material online), suggesting that the region in *M. unguiculatus* is an ancestral centromere that has expanded. The centromere-driven hypothesis (Malik 2009) may explain the distribution of array types in the autosomal centromeres under the following model: The ancestral gerbil centromeres were predominately B arrays and at some point after the *Meriones*–*Psammomys* split, centromeric drive triggered a massive repeat expansion of the B array on what would become *Meriones* chromosome 5. This runaway expansion was the catalyst for genome-wide centromere turnover, where A–C arrays replaced B arrays as the new functional centromeres and many B arrays were evolutionarily lost, with those that remained being nonfunctional relics. Indeed, the centromere expansion on chromosome 5 does not bind CENT proteins, although it preserves other heterochromatic marks (such as H3K9me3) and excludes recombination events, as assessed in male meiosis by the localization of the recombination marker MLH1 (fig. 5). While the heterochromatic state of a large portion of chromosome 5 can therefore be explained by the massive expansion of a centromeric repeat, this is not the case for chromosome 13.

Chromosome 13: Origin of a New Autosome

Chromosome 13 is the most unusual chromosome in the gerbil genome for a variety of reasons. Karyotypically,

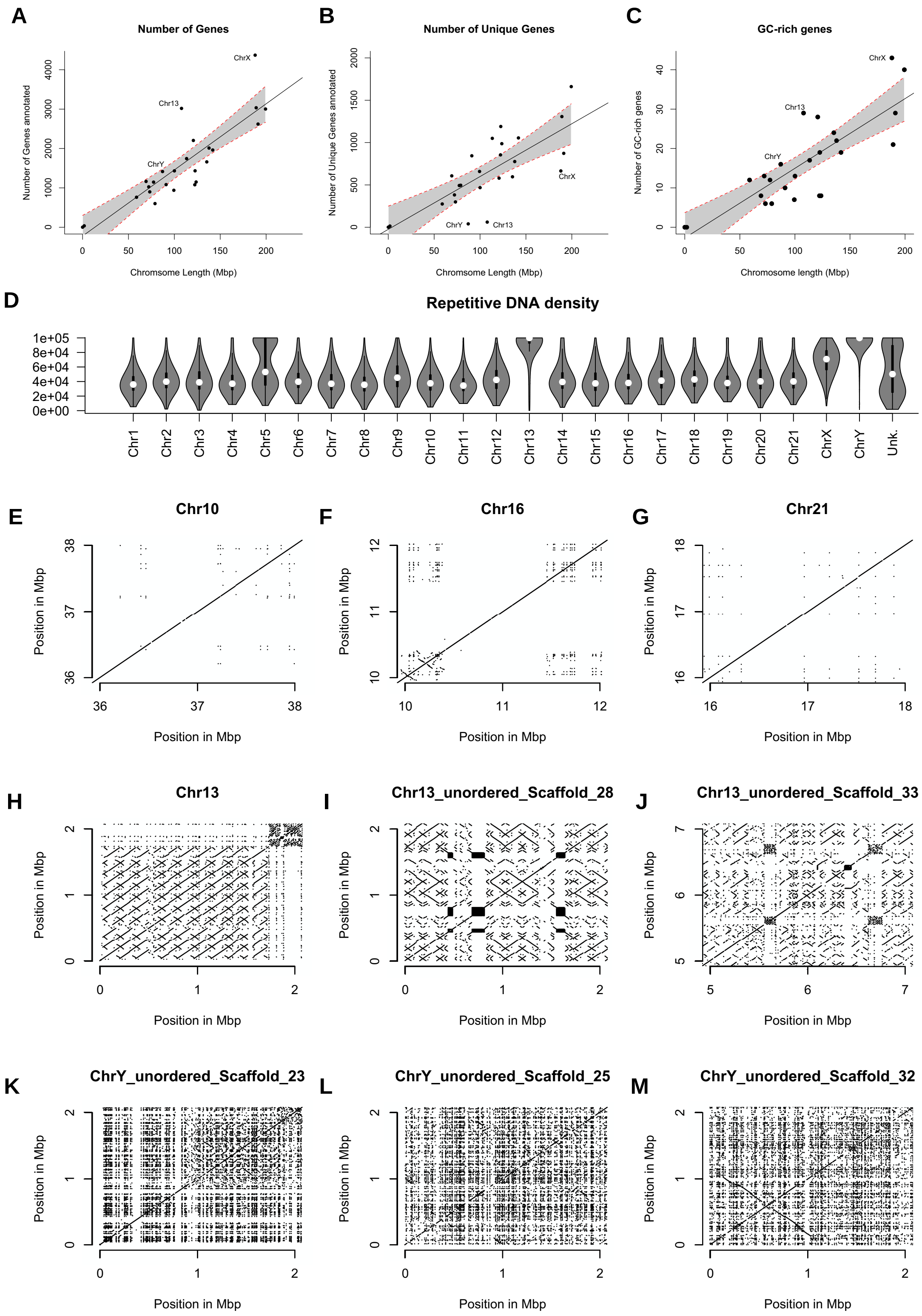

Fig. 6. Chromosome 13 is unusual in terms of gene content and repetitive DNA density. (*A*) There is a strong relationship between chromosome length and gene number, but both chromosome 13 and the X have more genes than expected for their length. (*B*) When duplicate genes are removed, chromosome 13 and both sex chromosomes have far fewer genes than expected based on their length (error bars show the 95% confidence interval). (*C*) Chromosome 13 is enriched for GC-rich genes. (*D*) Chromosome 13 has far higher repetitive DNA content than the other autosomes and is rivaled only by the Y. Panels *E*–*M* show a self-alignment of a selection of "typical" chromosomes (*E*, Chr10; *F*, Chr16; *G*, Chr21), as well as three of the longer scaffolds from the highly repetitive chromosome 13 (*H*, *I*, and *J*) and the Y (*K*, *L*, and *M*). Each panel shows a 2-Mbp section of chromosome, and only alignments longer than 1,000 bases are plotted. The primary alignments are clearly visible as diagonal lines at $y = x$. All alignments off of the 1:1 line are repetitive sequence. The prevalence of repetitive sequence on chromosome 13 is much higher than other autosomes and is most similar to the situation on the Y chromosome (*D*). However, repeats on chromosome 13 (*H*, *I*, and *J*) are much longer than those on the Y (*K*, *L*, and *M*), as expected based on their fundamentally different evolutionary history.

it stains very dark and appears heterochromatic in G (fig. 3*D*) and C-banding images (Gamperl et al. 1977; Gamperl and Vistorin 1980). It also displays delayed synapsis during the first meiotic prophase, when compared with all other chromosomes (de la Fuente et al. 2007, 2014). On a technical level, it is the only chromosome that failed to assemble into a single chromosome-length scaffold (fig. 1), and even optical mapping was unable to improve the assembly. In a phylogenetic context, there is no ortholog of chromosome 13 in mouse and rat, but similarity in G-banding patterns suggests that it may share ancestry with chromosome 14 in the fat sand rat (*P. obesus*). Short reads assigned to chromosome 13 have very low mapping quality as they map to multiple locations. As a result, chromosome 13 has very few genetic markers and a very short relative genetic map length compared with the other chromosomes (supplementary table S1, Supplementary Material online), and we suspect this is what prevented the OmniC data and HiRise pipeline from successfully assembling this chromosome. The centromere of chromosome 13 is unique in that the A–C arrays have more nonrepetitive blocks interspersed within them than the other chromosomes (fig. 3*D*), and in terms of sequence complexity, there is no fine-scale variation in entropy across the chromosome (fig. 3 and supplementary fig. S8, Supplementary Material online) as on the other autosomes, suggesting very low sequence diversity. Indeed, the entropy of chromosome 13 appears even more homogenous than that of the Y chromosome (fig. 3 and supplementary fig. S8, Supplementary Material online). Chromosome 13 has more than the expected number of genes based on its size (fig. 6*A*) but far fewer unique genes (fig. 6*B*), demonstrating high levels of gene duplication: Of the 1,990 genes on chromosome 13 annotated as something other than “protein of unknown function,” 566 are copies of a viral *pol* protein (and so represent either endogenous retrovirus or LET retrotransposon sequences), 406 are Vmn2r (olfactory receptor) genes (of which 337 are copies of *Vmn2r116*), and 331 are Znf (Zinc finger) genes (257 of which are *Znf431*). There are more GC-rich genes located on chromosome 13 than expected based on its size (fig. 6*C*), and chromosome 13 houses the original high-GC cluster (including the ParaHox genes) identified by (Hargreaves et al. 2017; Pracana et al. 2020). Chromosome 13 has a far higher repetitive sequence content (fig. 6*D*), as measured by the EarlGrey pipeline (Baril et al. 2022), which is clearly visible in comparison with other chromosomes in a self-alignment plot (fig. 5*E*–*M*). In fact, after filtering out alignments under 1,000 bp, over 93% of bases on chromosome 13 are found in multiple copies on the chromosome, compared with ~10% on other autosomes (e.g., 11.5%, 8.2%, and 12.7% on the similarly sized chromosomes 10, 11, and 12, respectively). The bulk of chromosome 13 consists of around 400 copies of a block of DNA 170 kb long, the periodicity and variable orientation of which can easily be seen in figure 5*H*–*J*. Although we find no evidence of a link between high GC% genes and this chromosome generally, chromosome 13 does encode the set of genes previously identified as being the most extreme outliers in gerbil and sand rat genomes (Pracana et al. 2020). These genes surrounding the ParaHox gene cluster include *Pdx1*, *Cdx2*, *Brca2*, and others crucial for proper embryogenesis and cell function (Withers et al. 1998). The cluster is contained within an ancient genomic regulatory block (Kikuta et al. 2007), where genes are locked together by the presence of overlapping regulatory elements. The presence of the most unusual genes on the most unusual chromosome is very interesting and is consistent with a model where the selective pressure to keep this block of genes intact may have had a role in the formation of the chromosome.

We propose the following model to explain the origin of chromosome 13: A chromosomal fragment ~5 million bases long that included the ParaHox cluster (Hargreaves et al. 2017) broke off from an ancestral chromosome, perhaps during a genome rearrangement. The ParaHox genes and many of their neighbors are crucially important during development and so could not be lost altogether. For example: *Pdx1*−/− mice die shortly after birth (Jonsson et al. 1994; Offield et al. 1996) as do those lacking *Brca2* (Evers and Jonkers 2006), *Insr* (Accili et al. 1996), or *Hmgb1* (Calogero et al. 1999) function; *Cdx2*−/− mice die within the first 5 or 6 days of development (Chawengsaksophak et al. 1997); 75% of *Gsx1*−/− mice die within 4 weeks of birth, and none live beyond 18 weeks (Li et al. 1996); and *Flt1*−/− mice die in utero (Fong et al. 1995). These are just a small selection of genes in this region, but they demonstrate the selective pressure(s) that must exist for its maintenance within the genome. Although the simplest option might have been for this fragment to have joined onto or into another chromosome, this does not appear to have happened, and instead we propose that this chromosomal fragment became the seed for the growth of an entirely new chromosome. In some species, the evolutionary fate of such a fragment may be long-term persistence as a microchromosome: a small, gene-dense, repeat-poor, GC-rich chromosome of ≤30 Mb with a high recombination rate. But while microchromosomes are common in birds, reptiles, and fish, they do not persist in mammals over evolutionarily time (Srikulnath et al. 2021; Waters et al. 2021). Efficient transmission of mammalian chromosomes between generations and into daughter cells therefore seems to require a minimum size, and in the case of *M. unguiculatus* chromosome 13, we propose the hypothesis in which the fragment grew rapidly via a breakage–fusion–bridge mechanism (McClintock 1938, 1941; Bignell et al. 2007; Campbell et al. 2010; Greenman et al. 2012), where the chromatid ends without a telomere fuse, and then is pulled apart at anaphase, breaking randomly and resulting in long inverted repeats. The patterns apparent in the chromosome 13 self-alignments (fig. 6*G*–*I*) are consistent with what would be expected under this model and may explain how a 170-kb region at the end of the chromosome was repeatedly duplicated, at multiple scales, until a 107-Mb chromosome was formed. The high similarity of these duplicated regions explains our difficulty in assembling this chromosome, the multimapping of short reads, and the failure of BioNano optical mapping to improve our assembly.

Although the repetitive nature of chromosome 13 is consistent with it arising and evolving under the model described above, that does not explain why chromosome 13 houses genes with the most extreme GC content. Previous authors (Gamperl and Vistorin 1980) have described that chromosome 13 forms ring-like structures during meiosis, suggesting that the bulk of the heterochromatic material on this chromosome does not, or possibly cannot, form chiasma, and therefore cannot undergo recombination. However, based on localization of the recombination marker MLH1, we have found evidence of recombination during male meiosis (fig. 5). Bivalent chromosome 13 presents a recombination event in most spermatocytes, although a small proportion (around 23%) lacks MLH1 foci. Strikingly, MLH1 are not evenly distributed along this chromosome, as previously reported for other chromosomes (de la Fuente et al. 2014). Instead, recombination events are strongly concentrated at the chromosome ends. We therefore propose that the extreme GC skew of the ParaHox-associated genes in gerbils is the result of the inability of recombination hotspots to move out of this genomic region, leading to runaway GC bias.

Conclusion

The two heterochromatin-rich chromosomes of Mongolian gerbils have distinct origins. Chromosome 5 has undergone a massive expansion of a centromeric repeat, most likely as a result of meiotic drive, and chromosome 13 has likely arisen de novo from an initially small seed via multiple breakage–fusion–bridge cycles. In general, these results show the importance of karyotypic knowledge of study species and serve as a warning for large-scale genome sequencing programs such as the Vertebrate Genomes Project (VGP) or the Darwin Tree of Life Project (DToL) that we must not neglect knowledge of chromosome number and morphology. Had we not known the diploid chromosome number for *M. unguiculatus* and had we not performed chromosome sorting and FISH, we likely would have binned the 121 fragments corresponding to chromosome 13 into the "unknown" category and may have even deduced that gerbils had one fewer chromosome than they actually have. We applied what are becoming the standard approaches for genome sequencing and assembly to the *M. unguiculatus* genome (PacBio HiFi, chromatin conformation capture, Oxford Nanopore long reads, and Bionano optical mapping) and incorporated chromosome sorting, FISH, and a single-nucleotide polymorphism (SNP)-based linkage map and were still unable to assemble chromosome 13 into a single scaffold. The huge size and high similarity of the chromosome 13 repeats suggest that only ultralong Oxford Nanopore reads, on the order of several hundred kilobases, might be able to achieve the T2T coverage of this enigmatic chromosome.

Materials and Methods

The complete details of the methods are available at the end of supplementary material S1, Supplementary Material online; here follows a very brief overview.

For sequencing and assembly, we extracted DNA from gerbil liver and sequenced to a depth of 34× using PacBio HiFi technology. Genome assembly was done with the program HiFiAsm (Cheng et al. 2021). Scaffolding was done using a combination of Dovetail OmniC, Oxford Nanopore Ultra-long sequencing, Bioano Optical Mapping, and a genetic map from Brekke et al. (2019). The genome was annotated using RNAseq from kidney and testis from three individuals. Repeats were annotated using the EarlGrey pipeline (Baril et al. 2022).

For cell culture, chromosome sorting, and FISH, we cultured cells from the gerbil fibroma cell line IMR-33 and extracted chromosomes for cell sorting after being arrested in mitosis. Chromosome sorting was done with a BD Influx Cell sorter into the 17 pools containing one or two chromosomes. Each pool was sequenced with Illumina MiSeq. FISH paints were made from each pool as well.

For mitotic chromosome preparation and FISH, we cultured fresh spleen cells that were then arrested in mitosis for chromosome spreads. These were stained with DAPI and the FISH probes derived from the chromosome sorting and visualized on a confocal microscope.

For meiotic chromosome preparation and immunofluorescence, we extracted meiotic cells from fresh testis and processed them for spreads and immunofluorescence. Slides were incubated with the primary antibodies goat anti-SYCP3 to mark the synaptonemal complex, rabbit antihistone H3 trimethylated at lysine 9 to mark heterchromatin, human anticentromere, and mouse anti-MLH1 to mark meiotic crossovers. Then, slides were incubated with secondary antibodies and visualized with an Olympus BX61 microscope equipped with appropriate fluorescent filters and an Olympus DP72 digital camera.

We assigned the sequenced scaffolds with the chromosomes in the karyotype by aligning the reads from the sequenced pools to the scaffolds and identifying which pools' reads most often aligned to each scaffold. Then, we linked the pools to the karyotype by staining mitotic chromosome spreads with the FISH probes derived from each pool.

We calculated GC content and gene density for each chromosome in sliding windows of size 1 kb and 1 Mb respectively with step size 1 kb. We calculated recombination rate with a sliding window of eight markers with a step of one marker and regressed marker position against physical position. Hotspots were identified as a region whose recombination rate was 5× the genome average. Entropy and linguistic complexity were calculated with the program NeSSie (Berselli et al. 2018) using a sliding window of size 10 kb with a step of 1 kb.

Centromeres were located at the trough of the linguistic complexity plot and the fine-scale structure was analyzed with NTRprism (Altemose et al. 2022) and TandemRepeatFinder (Benson 1999). Interstitial telomeres were identified as those with >70 copies of the telomere sequence in the TandemRepeatfinder data. Self-alignments were done with mummer (Kurtz et al. 2004).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors would like to thank two anonymous reviewers and the editor for their comments as well as Aaron Comeault, Martin Swain, Yichen Dai, Adam Hargreaves, Peter Holland, and Roddy Pracana for helpful discussions pertaining to the project, and Rebecca Snell for help with animal care, and also, David Thybert for providing the *Psammomys* genome assembly. T.D.B. would like to thank Kris Crandell. This work was supported by the Leverhulme Trust grant entitled “Decoding Dark DNA” (grant number RPG-2018-433) and by the National Environmental Research Council of the UK (grant number NE/R001081/1 to A.S.T.P) and by grant CGL2014-53106-P from Ministerio de Economía y Competitividad (Spain to J.P.). Unpublished genome assemblies for *M. unguiculatus* are used with permission from the DNA Zoo Consortium (dnazoo.org).

Author Contributions

O.F. and E.J.—chromosome sorting and editing the manuscript. F.Y. and B.F.—FISH and editing the manuscript. T.B. and A.H.—EarlGrey, repeat annotations, and editing the manuscript. J.P and R. d. l. F.—recombination/histone analyses and editing the manuscript. T.D.B., J. F. M., and A. S. T. P.—conceived the study, genome sequencing, assembly, analysis, overall coordination, and writing and editing the manuscript.

Data Availability

All sequencing data and the genome are available under SRA BioProject PRJNA397533. Specific accession numbers can be found in supplementary material S1, Supplementary Material online. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAODIK000000000. The version described in this paper is version JAODIK010000000. The genetic map, a vcf of the genetic markers and their genotypes in the mapping panel, the gff of the gene annotations, the gff of the repetitive element annotations, and “Supplemental_Material 3_codebase.zip”, can be found in the Dryad repository here: Brekke, Thomas D. (2022), Data for “The origin of a new chromosome in gerbils”, Dryad, Dataset, https://doi.org/10.5061/dryad.1vhhmgqws.

Conflict of interest statement: The authors declare no competing interests.

References

Accili D, Drago J, Lee EJ, Johnson MD, Cool MH, Salvatore P, Asico LD, José PA, Taylor SI, Westphal H. 1996. Early neonatal death in mice homozygous for a null allele of the insulin receptor gene. *Nat Genet.* **12**:106–109.

Ahmad S, Martins C. 2019. The modern view of B chromosomes under the impact of high scale omics analyses. *Cells* **8**:156–126.

Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, Hoyt SJ, Uralsky L, Ryabov FD, Shew CJ, *et al.* 2022. Complete genomic and epigenetic maps of human centromeres. *Science* **376**:eabl4178.

Aniskin VM, Benazzou T, Biltueva L, Dobigny G, Granjon L, Volobouev V. 2006. Unusually extensive karyotype reorganization in four congeneric *Gerbillus* species (Muridae: Gerbillinae). *Cytogenet Genome Res.* **112**:131–140.

Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc Natl Acad Sci U S A.* **112**:2109–2114.

Baril T, Imrie RM, Hayward A. 2022. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. Biorxiv 2022.06.30.498289.

Benazzou T, Viegas-Pequignot E, Petter F, Dutrillaux B. 1982. Chromosomal phylogeny of four *Meriones* (Rodentia, Gerbillidae) species. *Ann Genet.* **25**:19–24.

Benazzou T, Viegas-Pequignot E, Prod’Homme M, Lombard M, Petter F, Dutrillaux B. 1984. Chromosomal phylogeny of Gerbillidae. III. Species study of the genera *Tatera*, *Taterillus*, *Psammomys*, and *Pachyuromys*. *Ann Genet.* **27**:17–26.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**:573–580.

Berselli M, Lavezzo E, Toppo S. 2018. NeSSie: a tool for the identification of approximate DNA sequence symmetries. *Bioinformatics* **34**:2503–2505.

Bignell GR, Santarius T, Pole JCM, Butler AP, Perry J, Pleasance E, Greenman C, Menzies A, Taylor S, Edkins S, *et al.* 2007. Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res.* **17**:1296–1303.

Bornelöv S, Seroussi E, Yosefi S, Pendavis K, Burgess SC, Grabherr M, Friedman-Einat M, Andersson L. 2017. Correspondence on Lovell et al.: identification of chicken genes previously assumed to be evolutionarily lost. *Genome Biol.* **18**:1–4.

Botero-Castro F, Figuet E, Tilak M-K, Nabholz B, Galtier N. 2017. Avian genomes revisited: hidden genes uncovered and the rates versus traits paradox in birds. *Mol Biol Evol.* **34**:3123–3131.

Brekke TD, Supriya S, Denver MG, Thom A, Steele KA, Mulley JF. 2019. A high-density genetic map and molecular sex-typing assay for gerbils. *Mamm Genome.* **30**:63–70.

Calogero S, Grassi F, Aguzzi A, Voigtländer T, Ferrier P, Ferrari S, Bianchi ME. 1999. The lack of chromosomal protein Hmg1 does not disrupt cell growth but causes lethal hypoglycaemia in newborn mice. *Nat Genet.* **22**:276–280.

Camacho JP, Sharbel TF, Beukeboom LW. 2000. B-chromosome evolution. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences.* **355**:163–178.

Campbell PJ, Yachida S, Mudie LJ, Stephens PJ, Pleasance ED, Stebbings LA, Morsberger LA, Latimer C, McLaren S, Lin M-L, *et al.* 2010. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**:1109–1113.

Chawengsaksophak K, James R, Hammond VE, Köntgen F, Beck F. 1997. Homeosis and intestinal tumours in *Cdx2* mutant mice. *Nature* **386**:84–87.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. **18**(2):170–175.

Cheng S, Fu Y, Zhang Y, Xian W, Wang H, Grothe B, Liu X, Xu X, Klug A, McCullagh EA. 2019. De novo assembly of the Mongolian gerbil genome and transcriptome. Biorxiv 522516.

Craig-Holmes AP, Shaw MW. 1971. Polymorphism of human constitutive heterochromatin. *Science* **174**:702–704.

Dai Y, Pracana R, Holland PWH. 2020. Divergent genes in gerbils: prevalence, relation to GC-biased substitution, and phenotypic relevance. *BMC Evol Biol*. **20**:134.

de la Fuente R, Manterola M, Viera A, Parra MT, Alsheimer M, Rufas JS, Page J. 2014. Chromatin organization and remodeling of interstitial telomeric sites during meiosis in the Mongolian gerbil (*Meriones unguiculatus*). *Genetics* **197**:1137–1151.

de la Fuente R, Parra MT, Viera A, Calvente A, Gómez R, Suja JÁ, Rufas JS, Page J. 2007. Meiotic pairing and segregation of achiasmate sex chromosomes in eutherian mammals: the role of SYCP3 protein. *PLoS Genet*. **3**:e198–e112.

Dhar MK, Friebe B, Koul AK, Gill BS. 2002. Origin of an apparent B chromosome by mutation, chromosome fragmentation and specific DNA sequence amplification. *Chromosoma* **111**:332–340.

Dillon N. 2004. Heterochromatin structure and function. *Biol Cell*. **96**: 631–637.

Dimitri P, Corradini N, Rossi F, Vernì F. 2005. The paradox of functional heterochromatin. *Bioessays* **27**:29–41.

Dobigny G, Aniskin V, Volobouev V. 2002. Explosive chromosome evolution and speciation in the gerbil genus *Taterillus* (Rodentia, Gerbillinae): a case of two new cryptic species. *Cytogenet Genome Res*. **96**:117–124.

Evers B, Jonkers J. 2006. Mouse models of BRCA1 and BRCA2 deficiency: past lessons, current understanding and future prospects. *Oncogene* **25**:5885–5897.

Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. *Nat Rev Genet*. **2**:549–555.

Fong G-H, Rossant J, Gertsenstein M, Breitman ML. 1995. Role of the Flt-1 receptor tyrosine kinase in regulating the assembly of vascular endothelium. *Nature* **376**:66–70.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* **159**:907–911.

Gamperl R, Vistorin G. 1980. Comparative study of G- and C-banded chromosomes of *Gerbillus campestris* and *Meriones unguiculatus* (Rodentia, Gerbillinae). *Genetica* **52–53**:93–97.

Gamperl R, Vistorin G, Rosenkranz W. 1977. New observations on the karyotype of the Djungarian hamster, *Phodopus sungorus*. *Experientia* **33**:1020–1021.

Garsed DW, Holloway AJ, Thomas DM. 2009. Cancer-associated neochromosomes: a novel mechanism of oncogenesis. *Bioessays* **31**: 1191–1200.

Garsed DW, Marshall OJ, Corbin VDA, Hsu A, Di Stefano L, Schröder J, Li J, Feng Z-P, Kim BW, Kowarsky M, *et al*. 2014. The architecture and evolution of cancer neochromosomes. *Cancer Cell* **26**:653–667.

Gauthier P, Hima K, Dobigny G. 2010. Robertsonian fusions, pericentromeric repeat organization and evolution: a case study within a highly polymorphic rodent species, *Gerbillus nigeriae*. *Chromosome Res*. **18**:473–486.

Greenman C, Cooke S, Marshall J, Stratton M, and Campbell P. 2012. Modelling breakage-fusion-bridge cycles as a stochastic paper folding process. Arxiv.

Grewal SIS, Moazed D. 2003. Heterochromatin and epigenetic control of gene expression. *Science* **301**:798–802.

Gustavsen CR, Chevret P, Krasnov B, Mowlavi G, Madsen OD, Heller RS. 2008. The morphology of islets of Langerhans is only mildly affected by the lack of Pdx-1 in the pancreas of adult *Meriones jirds*. *Gen Comp Endocr*. **159**:241–249.

Hargreaves AD, Zhou L, Christensen J, taz FM, Liu S, Li F, Jansen PG, Spiga E, Hansen MT, Pedersen SVH, *et al*. 2017. Genome sequence of a diabetes-prone rodent reveals a mutation hotspot around the ParaHox gene cluster. *PNAS* **12**:201702930–6.

Hron T, Pajer P, Pačes J, Bartůněk P, Elleder D. 2015. Hidden genes in birds. *Genome Biol*. **16**:164.

Jayakumar V, Sakakibara Y. 2019. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. *Brief Bioinform*. **20**:866–876.

Jonsson J, Carlsson L, Edlund T, Edlund H. 1994. Insulin-promoter-factor 1 is required for pancreas development in mice. *Nature* **371**:606–609.

Katzer F, Lizundia R, Ngugi D, Blake D, McKeever D. 2011. Construction of a genetic map for *Theileria parva*: identification of hotspots of recombination. *Int J Parasitol*. **41**:669–675.

Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engstrom PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K, *et al*. 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res*. **17**:545–555.

Knight LI, Ng BL, Cheng W, Fu B, Yang F, Rambau RV. 2013. Tracking chromosome evolution in southern African gerbils using flow-sorted chromosome paints. *Cytogenet Genome Res*. **139**:267–275.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol*. **5**:R12–R19.

Lamb BC. 1984. The properties of meiotic gene conversion important in its effects on evolution. *Heredity (Edinb)*. **53**:113–138.

Leibowitz G, Ferber S, Apelqvist A, Edlund H, Gross DJ, Cerasi E, Melloul D, Kaiser N. 2001. IPF1/PDX1 deficiency and β-cell dysfunction in *Psammomys obesus*, an animal with type 2 diabetes. *Diabetes* **50**:1799–1806.

Lercher MJ, Smith NGC, Eyre-Walker A, Hurst LD. 2002. The evolution of isochores: evidence from SNP frequency distributions. *Genetics* **162**:1805–1810.

Li H, Zeitler PS, Valerius MT, Small K, Potter SS. 1996. Gsh-1, an orphan hox gene, is required for normal pituitary development. *Embo J*. **15**:714–724.

Lydall D, Nikolsky Y, Bishop DK, Weinert T. 1996. A meiotic recombination checkpoint controlled by mitotic checkpoint genes. *Nature* **383**:840–843.

Lyon MF. 1962. Sex chromatin and gene action in the mammalian X-chromosome. *Am J Hum Genet*. **14**:135–148.

Malik HS. 2009. The centromere-drive hypothesis: a simple basis for centromere complexity. In: Ugarkovic D, editor. *Centromere. Progress in molecular and subcellular biology*. Vol. 48. Berlin, Heidelberg: Springer.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol*. **38**:4647–4654.

Martinez-Perez E, Colaiácovo MP. 2009. Distribution of meiotic recombination events: talking to your neighbors. *Curr Opin Genet Dev*. **19**:105–112.

McClintock B. 1938. The production of homozygous deficient tissues with mutant characteristics by means of the aberrant mitotic behavior of ring-shaped chromosomes. *Genetics* **23**:315–376.

McClintock B. 1941. The stability of broken ends of chromosomes in *Zea mays*. *Genetics* **26**:234–282.

McKinley KL, Cheeseman IM. 2016. The molecular basis for centromere identity and function. *Nat Rev Mol Cell Bio*. **17**:16–29.

Nachman MW. 2002. Variation in recombination rate across the genome: evidence and implications. *Curr Opin Genet Dev*. **12**:657–663.

Offield MF, Jetton TL, Labosky PA, Ray M, Stein RW, Magnuson MA, Hogan BL, Wright CV. 1996. PDX-1 is required for pancreatic outgrowth and differentiation of the rostral duodenum. *Development* **122**:983–995.

Paigen K, Petkov P. 2010. Mammalian recombination hot spots: properties, control and evolution. *Nat Rev Genet*. **11**:221–233.

Pakes SP. 1969. *The somatic chromosomes of the Mongolian gerbil (Meriones unguiculatus)*. NAMI-1056. NASA Contract Rep NASA CR. 1969 Jan:1–8.

Penagos-Puig A, Furlan-Magaril M. 2020. Heterochromatin as an important driver of genome organization. *Frontiers Cell Dev Biology*. **8**:579137.

Pracana R, Hargreaves AD, Mulley JF, Holland PWH. 2020. Runaway GC evolution in gerbil genomes. *Mol Biol Evol*. **37**:2197–2210.

Qumsiyeh MB. 1986a. Phylogenetic studies of the rodent family Gerbillidae: i. Chromosomal evolution in the Southern African complex. *JMamm*. **67**:680–692.

Qumsiyeh MBH. 1986b. Chromosomal evolution in the rodent family Gerbillidae. Texas Tech University Thesis.

Saksouk N, Simboeck E, Déjardin J. 2015. Constitutive heterochromatin formation and transcription in mammals. *Epigenet Chromatin*. **8**:3.

Singhal S, Leffler EM, Sannareddy K, Turner I, Venn O, Hooper DM, Strand AI, Li Q, Raney B, Balakrishnan CN, *et al.* 2015. Stable recombination hotspots in birds. *Science* **350**:928–932.

Solari AJ, Ashley T. 1977. Ultrastructure and behavior of the achiasmatic, telosynaptic XY pair of the sand rat (*Psammomys obesus*). *Chromosoma* **62**:319–336.

Srikulnath K, Ahmad SF, Singchat W, Panthum T. 2021. Why do some vertebrates have microchromosomes? *Cells* **10**:2182.

Talbert PB, Henikoff S. 2020. What makes a centromere? *Exp Cell Res*. **389**:111895.

Tiemann-Boege I, Schwarz T, Striedner Y, Heissl A. 2017. The consequences of sequence erosion in the evolution of recombination hotspots. *Philos Trans R Soc B Biol Sci* . **372**:20160462.

Tilak M-K, Botero-Castro F, Galtier N, Nabholz B. 2018. Illumina library preparation for sequencing the GC-rich fraction of heterogeneous genomic DNA. *Genome Biol Evol*. **10**: 616–622.

Vinogradov AE. 2005. Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth. *Trends Genet*. **21**:639–643.

Volobouev V, Aniskin VM, Sicard B, Dobigny G, Granjon L. 2007. Systematics and phylogeny of West African gerbils of the genus Gerbilliscus (Muridae: Gerbillinae) inferred from comparative G- and C-banding chromosomal analyses. *Cytogenet Genome Res*. **116**:269–281.

Waters PD, Patel HR, Ruiz-Herrera A, Álvarez-González L, Lister NC, Simakov O, Ezaz T, Kaur P, Frere C, Grützner F, *et al.* 2021. Microchromosomes are building blocks of bird, reptile, and mammal chromosomes. *Proc Natl Acad Sci U S A*. **118**: e2112494118.

Weiss L, Mayeda K, Dully M. 1970. The karyotype of the Mongolian gerbil, *Meriones unguiculatus*. *Cytologia (Tokyo)* **35**:102–106.

Willard HF. 1990. Centromeres of mammalian chromosomes. *Trends Genet*. **6**:410–416.

Withers DJ, Gutierrez JS, Towery H, Burks DJ, Ren J-M, Previs S, Zhang Y, Bernal D, Pons S, Shulman GI, *et al.* 1998. Disruption of IRS-2 causes type 2 diabetes in mice. *Nature* **391**:900–904.

Yin Z-T, Zhu F, Lin F-B, Jia T, Wang Z, Sun D-T, Li G-S, Zhang C-L, Smith J, Yang N, *et al.* 2019. Revisiting avian 'missing' genes from de novo assembled transcripts. *BMC Genomics* **20**:4.

Zorio DAR, Monsma S, Sanes DH, Golding NL, Rubel EW, Wang Y. 2019. De novo sequencing and initial annotation of the Mongolian gerbil (*Meriones unguiculatus*) genome. *Genomics* **111**:441–449.