Title:

The origin of a new chromosome in gerbils

Authors:

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

Affiliations:

1. School of Natural Sciences, Bangor University, Bangor, Gwynedd, LL57 2DG, United Kingdom;

2. Centre for Genomic Regulation (CRG), Barcelona, Spain

3. Pompeu Fabra University (UPF), Barcelona, Spain.

4. Wellcome Sanger Institute, Hinxton, Cambridge, UK;

5. Current Address: School of Life Sciences and Medicine, Shandong University of Technology, Zibo, Shandong, China

6. Department of Experimental Embryology, Institute of Genetics and Animal Biotechnology of the Polish Academy of Sciences, Jastrzębiec, 05-552 Magdalenka, Poland;

7. Departamento de Biología, Facultad de Ciencias, Universidad Autónoma de Madrid, 28049, Madrid, Spain;

8. University of Exeter, Penryn Campus, Cornwall, TR10 9FE, United Kingdom.

To whom correspondence should be addressed: j.mulley@bangor.ac.uk

Keywords: Meriones, genome, karyotype, GC biased gene conversion, chromosome evolution

Abstract

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 (*Meriones unguiculatus*) have distinct underpinnings: chromosome 5 has a large block of intra-arm heterochromatin as the result of a massive expansion of centromeric repeats (probably due to centromeric drive); while chromosome 13 is comprised of extremely large (>150kb) repeated sequences. We suggest that chromosome 13 originated when a functionally important 'seed' broke off from another chromosome and underwent multiple breakage-fusion-bridge cycles. Genes with the most extreme GC skew are encoded on this chromosome, most likely due to the restriction of recombination to a narrow permissive region (since GC bias is linked with recombination-associated processes). 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.

Introduction

The nucleotide composition of genomes is not homogenous; varying 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 favours fixation of guanine and cytosine 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, that is not driven by selection (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: (i) a region associated with the ParaHox cluster and the surrounding genes, where virtually all genes in this region have extremely high mutation rates and a marked bias towards weak-to-strong A/T to G/C substitutions, and (ii) a further set of 17 large clusters of GC-rich genes also with high mutation rates that are slightly biased towards weak-to-strong A/T to G/C substitutions (Pracana et al. 2020). These intriguing characteristics of gerbil genomes make them an ideal system in which to examine the association between GC biased gene conversion and the organization of eukaryotic genomes.

In addition to these GC-skewed regions, gerbil genomes possess distinctive karyotypic features where one or more chromosomes with extensive regions of heterochromatin or that appear fully heterochromatic, showing as entirely dark in C-banding stains (Gamperl and Vistorin 1980) or coated by heterochromatin histone marks in immunofluorescence assays (de la Fuente et al. 2014). 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 tri-methylation 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. In Mongolian gerbils (*Meriones unguiculatus*), nearly a third of chromosome 5 and all of chromosome 13 appear to be composed of constitutive heterochromatin (Gamperl and Vistorin 1980). The genomes of the North African Gerbil (*Gerbillus campestris*), the hairy-footed gerbil (*Gerbilliscus paeba*), and the fat sandrat (*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: i.e.: they are not non-essential, 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 3rd 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 non-permissive 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 heterochomatic chromosomes in gerbils arise? They may once have been “normal” chromosomes that have degenerated into gene-poor, non-functional, 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 tumours (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. Finally, they may potentially perhaps 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 have be addressed in a non-model system for a 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 towards 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 test (i) whether GC-rich gene clusters correlate with recombination hotspots and (ii) if those genes are associated with a single heterochromatic chromosome. Our approach allows us to examine the origin and propose a new hypotheses 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 and scaffolded OmniC chromatin conformation capture data (Figure S1), Oxford Nanopore long and ultra-long read sequence data, a genetic map (Table S1) (Brekke et al. 2019), and BioNano optical mapping. We assigned scaffolds to chromosomes by flow-sorting chromosomes into pools and using these pools to generate short-read sequences and paint probes for FISH. Alignments of short reads from pools to scaffolds and FISH probes were used to link scaffolds with physical (Supplemental Material 1). The final genome assembly contains 194 scaffolds spanning 21 autosomes and the X and Y sex chromosomes, and the mitochondrial genome. For 20 of the 23 chromosomes, a single large scaffold contains over 94% (often over 99%) of all the sequence assigned to that chromosome. Only chromosome 13, with 121 scaffolds, and the X and Y chromosomes, each with 6 scaffolds, are appreciably fragmented (Figure 1). 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:13798) (Manni et al. 2021). Entropy and linguistic complexity identified centromeres and two outlier chromosomes: one with an extensive region of low complexity, and one with an overall homogenous level of complexity (Figure 3).

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 (Table S2). 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 et al. 2019). Full-genome alignments between our genome assembly and the Hi-C assembly (Figure S3) showed that most scaffolds are colinear between the assemblies but that the improved Cheng et al. (Cheng et al. 2019) assembly does not include chromosome 13. Our highly contiguous and physically associated assembly provides the foundation for all subsequent analyses.

The location of GC-rich genes

A set of over 380 genes with extreme GC content clustered in the genomes of sandrats and gerbils have previously been identified. 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 (Figure 4A; permutation test, n=1,000,000 permutations, $p<0.000001$). We used our genetic map (Brekke et al. 2019) to locate recombination hotspots which were defined as regions with 5x 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 (Figure 4B, Figure S4, S5). Chromosomes 2, 18, 21, and the X chromosome were found to 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 (Figure 4C) and secondly by performing a permutation test (Figure 4D). In both cases, GC outlier genes were found to lie significantly closer to recombination hotspots than expected by chance (Figure 4C: t-test, df = 383.2, t = 2.585, p = 0.01012; Figure 4D: permutation test, n = 1,000,000, p < 0.000001). 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 (Figures 4E and 4F). We found that whether considering the full distribution of gene locations or 1,000,000 draws of the same number of random genes in a permutation test, the GC outliers are found to lie much closer to the telomere than expected by chance (Figure 4E: t-test, df = 418, t = -14.26, p < 0.000001; Figure 4F: permutation test, n=1,000,000, p < 0.000001). Furthermore, gerbils have many interstitial telomere sites (de la Fuente et al. 2014) which 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 therefore

tested whether GC outlier genes are closer to telomere repeats (interstitial or otherwise) than expected by chance and found that they are (Figure 4G: t-test, df = 418.6, t = 7.876, p=0; Figure 4H: permutation test, n=1,000,000, p=0). 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 GC-biased gene conversion is driving the extreme GC content of these genes.

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 to 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

Relatively little is known about centromere organization in non-model 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 we finally obtained full coverage of human centromeres, 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 allowed us to fully span the centromeres of all *M. unguiculatus* chromosomes, which in all cases correlated with regions of low complexity in linguistic complexity plots. We used NTRprism (Altemose et al. 2022) to identify four different simple repeat sequences found in gerbil centromeres (Figure S6) which we have termed MsatA (for *Meriones* satellite A) , MsatB, MsatC, and MsatD (Figure 2A): MsatA is 6bp long and has the sequence TTAGGG which is the same simple sequence repeat found in telomeres, MsatB is 37bp long, MsatC is 127bp long, and MsatD is 1,747bp long and is only found on the Y chromosome.

Copies of Msats are arranged into one of four variant arrays which define an intermediate-order structure in the centromeres (Figure 2B). 'B arrays' are formed from copies of MsatB and range in size from 1Mbp to 3Mbp long (~30,000 to ~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 100Kb and 1Mb depending on the chromosome. The only place that MsatC are 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 (Figure 2D, for a high resolution image see Supplemental Material 3) 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 which fall into one of a few patterns which we term ‘simple’, ‘asymmetric’, or ‘symmetric’ (Figure 2C). 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, Figure 2D). 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, Figure 2D), three of the metacentric chromosomes (chromosomes 4, 7, and 14, Figure 2D), and one acrocentric chromosome (chromosome 17, Figure 2D). 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 sandwiched between two A-C arrays and are found in the metacentric chromosomes 1, and 2, and the acrocentric chromosome 13. Many centromeres also contain 10Kbp–50Kbp blocks of non-repetitive, complex DNA both between and within the various arrays (see Figure 2D).

Chromosome 5 is characterized by an enormous centromeric repeat expansion which is visible as a dark band on the q arm (Figure 2D). Our data shows that the repeat expansion is a 29Mb long B array, which comprises approximately 22% of the entire chromosome. This repeat expansion is distinct from the centromere which is a simple A-C array 1.5Mb long. In 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 1-3Mb. Repeats in the chromosome 5 expansion however, switch orientation over 200 times across the 29Mb, so the average block length is just 140Kb.

There is a similar large expansion of a centromeric repeat found in human chromosome 9 (Altemose et al. 2022). However, while 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 *Desmodilus, Gerbillurus, Gerbillus, Tatera,* and *Taterillus,* and is even absent in other species of *Meriones.* (Gamperl and Vistorin 1980; Benazzou et al. 1982, 1984; Qumsiyeh 1986b,a; 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 sandrat (*Psammomys 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 (Figure S7), suggesting that the region in *M. unguiculatus* is an ancestral centromere that has expanded. The centromere-drive 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 non-functional 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 (Figure 6). 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 (Figure 3).

Chromosome 13: origin of a new autosome

Chromosome 13 is the most unusual chromosome in the gerbil genome for a variety of reasons. Karyotypically, it stains very dark and appears heterochromatic in G (Figure 2D) and C-banding images (Gamperl and Vistorin 1980). It also displays delayed synapsis during the first meiotic prophase, when compared to 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 (Figure 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 sandrat (*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 to the other chromosomes (Table S1) 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 non-repetitive blocks interspersed within them than the other chromosomes (Figure 2D), and in terms of sequence complexity, there is no fine-scale variation in entropy across the chromosome (Figure 3) 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 (Figure 3). Chromosome 13 has more than the expected number of genes based on its size (Figure 5A), but far fewer unique genes (Figure 5B),

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 (Figure 5C) 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 (Figure 5D), as measured by the EarlGrey pipeline (Baril et al. 2022) which is clearly visible in comparison with other chromosomes in a self-alignment plot (Figure 5E-M). In fact, after filtering out alignments under 1,000bp, 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 170kb long, the periodicity and variable orientation of which can easily be seen in Figures 5H, 5I, and 5J. While we find no evidence of a link between high GC% genes and this chromosome generally, chromosome 13 does encode the set of genes with the highest substitution and weak-to-strong mutation rates that were previously identified as being the most extreme outliers in gerbil and sandrat genomes (Pracana et al. 2020). These are a set of linked genes surrounding the ParaHox gene cluster, including *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. While there is a 1 in 23 possibility of these genes being on this chromosome by chance (1 in 21 for autosomes), the presence of the most unusual genes on the most unusual chromosome is highly suggestive of a causative link.

We propose the following model to explain the origin of chromosome 13 and the extreme GC-skew of many of its genes: a chromosomal fragment approximately 5 million bases long which included the ParaHox cluster (Hargreaves et al. 2017) broke off from an ancestral chromosome perhaps during a genome rearrangement. The ParaHox and neighboring genes 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. While 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 ≤30Mb 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 suggest that 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 are pulled apart at anaphase, breaking randomly and resulting in long inverted repeats, as apparent in our chromosome 13 self-alignments (Figure 5G, 5H, 5I). In this way, a 170kb region at the end of the chromosome was repeatedly duplicated, at multiple scales, until a 107Mb 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. 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 (Figure 6). Bivalent chromosome 13 presents a recombination event in most spermatocytes, although a small proportion (around 23%) lack 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 arisen *de novo* from an initially small seed via multiple breakage-fusion-bridge cycles. 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 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 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 ultra-long Oxford Nanopore reads, on the order of several hundred kilobases, might be able to achieve the telomere-to-telomere coverage of this enigmatic chromosome.

Methods

Animal care and tissue collection

Male Tumblebrook Farm strain Mongolian gerbils from the colony at Bangor University (Brekke et al. 2018) were euthanised using a Schedule 1 method in accordance with EU Directive 2010/63/EU and the Animals (Scientific Procedures) Act 1986. Animal use was reviewed and approved by the Bangor University Animal Welfare and Ethical Review Board. Fresh liver, kidney, and testis were dissected and snap-frozen in liquid nitrogen, then stored at -80°C and whole spleens were dissected into pre-warmed RPMI 1640 buffer (ThermoFisher) for cell culture. Whole blood was collected into EDTA tubes BD Sciences 366450) and stored at 4°C until shipping.

Cell culture and chromosome sorting

For complete experimental details and methods see “Supplemental Material A Chromosome Sorting and FISH”. In brief, chromosomes were harvested from the gerbil fibroma cell line IMR-33 (ECACC General Cell Collection catalogue number 96020931) after being arrested in mitosis with Colcemid. The cells were lysed in a hypotonic solution and the chromosomes in suspension were sorted using a BD Influx cell sorter (Becton Dickinson, San Jose, CA) as described by (Kuderna et al. 2019). 17 separate clusters were identified and the used for sorting resulting in 6 pools with two chromosomes each and 11 pools with a single chromosome. After chromosomes were sorted, we dialysed each pool using a Pur-A-Lyzer Maxi Dialysis kit (Sigma, PURX50005-1KT) following the manufacturer's instructions to remove any dye that remained bound to the DNA from sorting.

DNA and RNA Sequencing

PacBio HiFi sequencing was performed by the Earlham Institute (Norwich, UK), using DNA extracted from frozen liver tissue. High molecular weight DNA was extracted with the

Nanobind Tissue kit (Circulomics), with quality assessed using an Agilent FEMTO-Pulse, with size selection of fragments between 18Kb and 19Kb in size with a SageELF. HiFi sequencing used 3 SMRT cells on a PacBio Sequel II, analysed with the CCS analysis pipeline (SRR18362962). We sequenced 88,071,091,902 base pairs and generated 4,814,749 CCS reads with a quality greater than or equal to Q20. These had an average length of 18,291bp and amount to 34x coverage.

OmniC sequencing was performed by Dovetail Genomics (California, USA), using DNA derived from frozen liver samples. 262,974,243 pairs of 151bp OmniC reads were sequenced (SRR18362944) corresponding to 38x coverage.

Whole fresh blood was sent to the DeepSeq facility (Nottingham, UK) for Oxford Nanopore PromethION sequencing and BioNano Optical Mapping. DNA was extracted in parallel with the Circulomics UHMW DNA extraction protocol (EXT-BLU-001) for PromethION sequencing and the Circulomics HMW DNA Extraction Protocol (EXT-BLH-001) for BioNano optical mapping. We sequenced 1,210,550 ultra-long reads of average length 39387bp and 2,550,000 long reads averaging 19,850bp via the PromethION (SRR18362939).

The sorted chromosomes were sequenced with Illumina MiSeq at Bangor University, and we generated 19,764,484 read pairs of 151bp paired end reads (SRR18362948, SRR18362952, SRR18362951, SRR18362947, SRR18362946, SRR18362945, SRR18362940, SRR18362958, SRR18362956, SRR18362955, SRR18362954, SRR18362949, SRR18362957, SRR18362953, SRR18362960, SRR18362959, SRR18362941).

RNA extraction and sequencing was done by GeneWiz (New Jersey, USA), using kidney and testis from three individuals. We received 177,016,012 151bp paired-end reads (SRR18362961, SRR18362937, SRR18362950, SRR18362943, SRR18362942, SRR18362938).

459

460 Genome assembly and annotation

461 We used the HiFiASM assembly software (Cheng et al. 2020) to build the first iteration
462 genome using the PacBio HiFi reads with the flag -l 1 to lightly purge duplicates. This initial
463 assembly was then scaffolded using the OmniC read-pairs and the HiRise pipeline. To the
464 OmniC-scaffolded assembly we aligned the raw reads from the genetic map (Brekke et al.
465 2019) using bwa (Li and Durbin 2009) and ran the Stacks2 (Rochette et al. 2019) pipeline
466 followed by R/qtl (Broman et al. 2003; Arends et al. 2014) to build a genome-guided genetic
467 map. This map informed additional merges of scaffolds which we did using the custom script
468 assemble_genome_and_recoordinate_gff.py (Supplemental Material 4). At this point there was
469 a single linkage group per chromosome (omitting the Y) and so the scaffold designation was
470 dropped from the name of those fasta entries. Thus anything named, for instance, simply
471 “Chr13” is tied to a linkage group whereas the scaffolds without genetic markers kept the longer
472 names of the form “Chr13_unplaced_Scaffold_28”. To further assemble the reference, we
473 aligned the Oxford Nanopore ultra-long reads using minimap2 (v2.17) (Li 2018) which also
474 suggested additional merges. We built a hybrid assembly using the Bionano optical mapping
475 data but it did not suggest any further merges and was not used in any analysis.

476 Our annotation was done on the Omni-C scaffolded version of the genome using the
477 RNAseq data from GeneWiz (New Jersey, USA). Repeat families found in the genome
478 assemblies of *Meriones unguiculatus* were identified de novo and classified using the software
479 package RepeatModeler (v2.0.1) (Flynn et al. 2020). RepeatModeler depends on the programs
480 RECON (v1.08) (Bao and Eddy 2002) and RepeatScout (v1.0.6) (Price et al. 2005) for the *de*
481 *novo* identification of repeats within the genome. The custom repeat library obtained from
482 RepeatModeler were used to discover, identify, and mask the repeats in the assembly file using
483 RepeatMasker (v4.1.0) (Tarailo-Graovac and Chen 2009). Coding sequences from *Meriones*

unguiculatus, Psammomys obesus, *Mus musculus*, *Rattus norvegicus* and *Peromyscus maniculatus* were used to train the initial *ab initio* model for *Meriones unguiculatus* using the AUGUSTUS software (v2.5.5) (Keller et al. 2011). Six rounds of prediction optimisation were done with the software package provided by AUGUSTUS. The same coding sequences were also used to train a separate *ab initio* model for *Meriones unguiculatus* using SNAP (v2006-07-28) (Korf 2004). RNAseq reads were mapped onto the genome using the STAR aligner software (v2.7) (Dobin et al. 2013) and intron hints generated with the bam2hints tools within the AUGUSTUS software. MAKER, SNAP and AUGUSTUS (with intron-exon boundary hints provided from RNA-Seq) were then used to predict for genes in the repeat-masked reference genome. To help guide the prediction process, Swiss-Prot peptide sequences from the UniProt database were downloaded and used in conjunction with the protein sequences from *Meriones unguiculatus, Psammomys obesus, Mus musculus, Rattus norvegicus* and *Peromyscus maniculatus* to generate peptide evidence in the Maker pipeline. Only genes that were predicted by both SNAP and AUGUSTUS softwares were retained in the final gene sets. To help assess the quality of the gene prediction, AED scores were generated for each of the predicted genes as part of the MAKER pipeline. Genes were further characterised for their putative function by performing a BLAST search of the peptide sequences against the UniProt database. tRNA were predicted using the software tRNAscan-SE (v2.05) (Chan and Lowe 2019). The gff annotation file was then re-coordinated along with each subsequent merging of scaffolds through the remaining assembly steps using the custom python script assemble_genome_and_recoordinate_gff.py (Supplemental Material 4).

The annotation pipeline described above did a perfunctory identification of repetitive elements as a step towards finding a high-quality set of genes, but to assemble and curate a high-quality list of repetitive sequences, we used the EarlGrey pipeline (Baril et al. 2022). This was configured with Dfam (version 3.4) (Hubley et al. 2016) and RepBase (release 20181026)

(Jurka et al. 2005; Kapitonov and Jurka 2008), specifying known repeats from *Rodentia* (-r rodentia).

Chromosome Assignment

Scaffolds from the final genome were assigned to chromosomes by the parallel approaches of sequencing each sorted chromosome pool and also using each pool as a FISH probe. Illumina reads from the pool sequencing were aligned to the assembly with bwa (Li and Durbin 2009). For the alignment of each chromosome pool, we counted the number of reads mapping to each scaffold and calculated the reads mapped per scaffold length. Each scaffold then has a read-mapping density from each pool making it possible to associate every scaffold with the pool to which it belonged. We calculated the 99.99% confidence interval of the read mapping density and the scaffold was assigned to the pool that fell outside the confidence interval. In 30 of the 194 cases, a scaffold assigned to either no pool or to multiple pools and we marked these as 'unknown'. Unknown scaffolds comprise 1,588,872 bases, 0.06% of the total genome.

The final link in the chain connecting the karyotype with the chromosomes was made using cross-species FISH. The chromosomes in the gerbil karyotype were named by Weiss (Weiss et al. 1970) who defined the pattern of G-bands on each chromosome. Spleen cells were cultured as per (Yang et al. 2017) in complete RPMI (i.e. RPMI with fetal calf serum, penicillin, and streptomycin) with added EPS to stimulate immune cell growth at 37C and 5% CO_2 for 48 hours after which colchicine was added to arrest the cells in metaphase. After an hour, the cells were spun down and resuspended in fixative and stored at -20C. Metaphase spreads were made by dropping 14 µl of cell suspension on a glass slide and drying at high humidity while floating in a 55C water bath. We created FISH probes from each chromosome pool and hybridized them to a chromosome spread, thereby linking the banding pattern of each

chromosome with a pool. FISH probes were made and hybridized to the metaphase spreads following (Murchison et al. 2012)

Male gerbils have 23 unique chromosomes (21 autosomes, an X, and a Y) and so we expected six of the 17 pools to have multiple chromosomes which is what we found. Eleven of the pools had a single chromosome while six included multiple chromosomes. For those pools with multiple chromosomes, we extracted the genes from the annotation file of the gerbil scaffolds and queried BioMart for the chromosome locations of those genes in mouse. From this it was clear which mouse chromosome corresponded with each gerbil scaffold thus allowing us to infer which gerbil banding pattern is associated with each sequence record in the gerbil genome. For further details, please see Supplemental Material 1.

Thus, the final version of gerbil genome presented here is the result of PacBio HiFi reads assembled with HiFiAsm (Cheng et al. 2020), scaffolded with OmniC paired reads, annotated with kidney and liver RNAseq data, further scaffolded with a high-density SNP-based genetic linkage map and then Oxford Nanopore ultra-long reads, and assigned to physical chromosomes using chromosome sorting and chromosomal FISH.

Genome Analysis

GC content and gene density were analyzed for every scaffold in the genome in sliding windows using the custom script Calc_R_GC_Gene_density.py (Supplemental Material 4). The window size for GC content is 1,000bp while the window size for gene density is 1Mb in both cases the windows progressed by 1kb across each scaffold. Recombination rate was estimated by taking a sliding window of 8 genetic markers and calculating the slope of the regression of their genetic position against their physical position. As the sliding window progressed by a single marker, each inter-marker region had eight rates associated with it and these were

557 averaged to get the recombination rate of each inter-marker region. The genome-wide
558 recombination rate was calculated by taking the mean of the rate of each region weighted by the
559 length of region. Recombination hotspots were identified by identifying every region whose rate
560 was greater than 5 times the genome average and adjacent regions were merged. Entropy and
561 Linguistic complexity were calculated using the program NeSSie (Berselli et al. 2018) using a
562 sliding window with size 10k and a step of 1k as recommended.

563 We identified centromere location by visually identifying the trough in the linguistic
564 complexity data of each chromosome. To understand the fine-scale structure of each
565 centromere, we extracted the region and used the program NTRprism (Altemose et al. 2022) to
566 identify the lengths of the different repeats and TandemRepeatFinder (Benson 1999) to identify
567 the sequence of repeats of each length, the data from which forms the basis of the coloured
568 centromere call-outs in Figure 2. TandemRepeatFinder also served to identify the location of
569 telomere repeats along the length of the chromosome. We identified interstitial telomere sites as
570 those with at least 70 tandem copies of the telomere repeat.

571 The locations of the 387 GC outlier genes in *Meriones* identified by (Pracana et al. 2020)
572 were extracted from the annotation file. Due to some gene duplications this resulted in 410
573 genomic locations in our assembly. We tested whether these locations were more clustered
574 than chance by drawing 410 random genes from the genome 1,000,000 times and calculating
575 the average that separated each from its closest neighbor in the set, and calculated a p-value
576 by taking the proportion of the permutations that had a lower average distance than the
577 observed set. We calculated how close they were to recombination in two ways, first by a t-test
578 comparing the distribution of all genes' proximity to recombination hotspots with the distribution
579 of the outlier genes and second by randomly drawing 410 genes 1,000,000 times and
580 calculating the average distance to the nearest hotspot for the draw. A p-value was calculated
581 for the permutation test as described above. A similar pair of tests was done to evaluate

whether the outlier genes were non-randomly placed along a chromosome arm and whether they were more closely located to telomere repeats (interstitial and normal) than expected by chance. For the location along a chromosome arm we transformed physical position of each gene in to a percentage going from the centromere at 0 to the telomere at 100 in order to compare chromosome arms of different sizes.

A self alignment was made for each scaffold that assigned to chromosome 13 as well as a few selected autosomes (10, 16, and 21) and the Y chromosome. The self-alignment was done with mummer4 (Kurtz et al. 2004) using the “maxmatch” and “nosimplify” parameters to identify repetitive elements. Mummer was also used to compare our genome with both the other chromosome-scale *Meriones* assembly (Cheng et al. 2019)and an unpublished chromosome-scale *Psammomys obesus* assembly provided by David Thybert using the “mum” parameter.

Meiotic chromosome preparation and immunofluorescence

We obtained preparations of male meiotic chromosomes as previously described (Peters et al. 1997; de la Fuente et al. 2007) Briefly, a cell suspension was made in PBS from whole testicles by rubbing seminiferous tubules with the help of two forceps. Then, cells were transferred to a 10mM sucrose solution and spread over glass slides previously covered with paraformaldehyde 1% in distilled water (pH 9,5) containing 0.15% of Triton X-100. After two hours standing horizontally on a humid chamber, slides were washed in distilled water with 0.04% Photoflo, air dried and stored at -80ºC until use. For immunofluorescence, slides were incubated overnight at 4ºC with the following primary antibodies diluted 1:100 in PBS: goat anti-SYCP3 protein of the synaptonemal complex (Santa Cruz 20845); rabbit anti histone H3 trimethylated at lysine 9 (H3K9me3) (Abcam 8898), as a marker of heterochromatin; human anti-centromere (Antibodies Incorporated 15-234); and mouse anti-MLH1 (Pharmingen 550838), as a marker of meiotic

crossovers. After washing three times in PBS slides were incubated for one hour at room temperature with secondary antibodies conjugated with Alexafluor 350, Alexafluor 488, Cy3 or Cy5, diluted 1:100 in PBS, all of them from Jackson ImmunoResearch Laboratories. After three washes in PBS slides were mounted with Vectashield. Observations were made in an Olympus BX61 microscope equipped with appropriate fluorescence filters and an Olympus DP72 digital camera. Them, images were treated with Adobe Photoshop (Adobe) and Image J.

Bivalents 5 and 13 were identified in pachytene spermatocytes owing to the presence of an interstitial H3K9me3 region or by a complete labeling with this histone marks, respectively. To assess the position of MLH1 foci along the bivalents, we measured the length of the synaptonemal complex of these two bivalents using the free hand tool in ImageJ. The distance of centromeres and MLH1 foci from the tip of the short arm of the bivalents were recorded in the same way. Then the position of MLH1 foci was normalized against the length of the corresponding bivalent, yielding a position between 0 (the proximal end) and 1 (the distal end). The position of all foci was presented in cumulative frequency chart. A total of 83 spermatocytes from two different individuals were scored.

List of Supplementary materials:

Supplemental Material 1: The detailed methods of the chromosome sorting and the logic linking the fasta records to the karyptype.

Supplemental Material 2: Supplemental Tables and Figures including:

Table S1: Genetic map statistics.

Table S2: Comparison of published gerbil genomes.

Figure S1: Dovetail OmniC contact map.

Figure S2: GC content, gene density, entropy, and linguistic complexity in sliding windows across each chromosome.

Figure S3: Whole-genome alignment of the genome presented here and Cheng et al (2019)'s HiC scaffolded version.

Figure S4: Recombination rates for each chromosome showing hotspots.

Figure S5: Marey maps for each chromosome showing hotspots.

Figure S6: Repeat frequency spectra for the centromeric region of each chromosome.

Figure S7: A whole genome alignment of *Meriones unguiculatus* chromosome 5 with *Psammomys obesus* chromosome 10 showing the expansion of *Meriones unguiculatus* Chr5.

Supplemental Material 3: A very high-resolution image of Figure 2D to facilitate close inspection of chromosomal features.

Supplemental Material 4: Code base. All in-house scripts. See do_it_all_genome_polish_v6.sh for a step-by-step call sequence. Also includes some metadata files needed to run the pipeline and some Rdata packets to skip some time-consuming analyses.

647

Acknowledgements

The authors would like to thank Aaron Comeault, Martin Swain, Yichen Dai, Adam Hargreaves, Peter Holland, and Roddy Pracana for helpful discussions pertaining to the project, and Becca Snell for help with animal care. TDB 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 *Meriones unguiculatus* are used with permission from the DNA Zoo Consortium (dnazoo.org).

Authors contributions

O.F. and E.J.– chromosome sorting, editing manuscript

F.Y. and B.F. – FISH, editing manuscript

T.B. and A.H. – EarlGrey, repeat annotations, editing manuscript

J.P and R. d. l. F. – recombination/histone analyses, editing manuscript

T.D.B., J. F. M., and A. S. T. P. – conceived study, genome sequencing, assembly, analysis, overall coordination, writing and editing manuscript

Competing interests

The authors declare no competing interests.

Data and materials availability

All sequencing data and the genome is available under SRA BioProject PRJNA397533.

PacBio HiFi: SRR18362962.

Illumina OmniC: SRR18362944.

Oxford Nanopore PromethION: SRR18362939.

Illumina MiSeq sorted chromosome sequencing: SRR18362948, SRR18362952, SRR18362951, SRR18362947, SRR18362946, SRR18362945, SRR18362940, SRR18362958, SRR18362956, SRR18362955, SRR18362954, SRR18362949, SRR18362957, SRR18362953, SRR18362960, SRR18362959, SRR18362941.

Illumina RNAseq from testis and kidney: SRR18362961, SRR18362937, SRR18362950, SRR18362943, SRR18362942, SRR18362938.

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, and the gff of the repetitive element annotations 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. Reviewers may find these data files ahead of publication here: https://datadryad.org/stash/share/R5vtycW8DE6euNZJEe26JvJIVTmCEaJVo9SQpfXWAJk

Tables and Figures

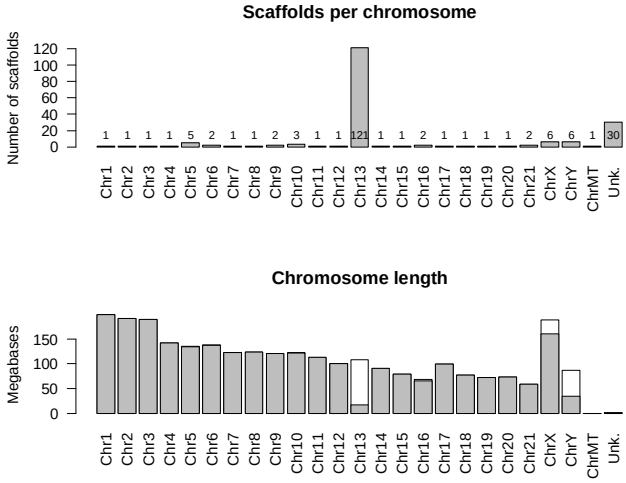


Figure 1: Summary statistics for the Mongolian gerbil (*Meriones unguiculatus*) genome assembly. Top: The number of scaffolds assigned to each chromosome, the mitochondrial genome, and the 'unknown' category. Most chromosomes are assembled into 1 or 2 scaffolds, while chromosome 13 is in 121 pieces. Bottom: The number of bases assigned to each chromosome with the single longest scaffold shaded in grey. 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 artefact.

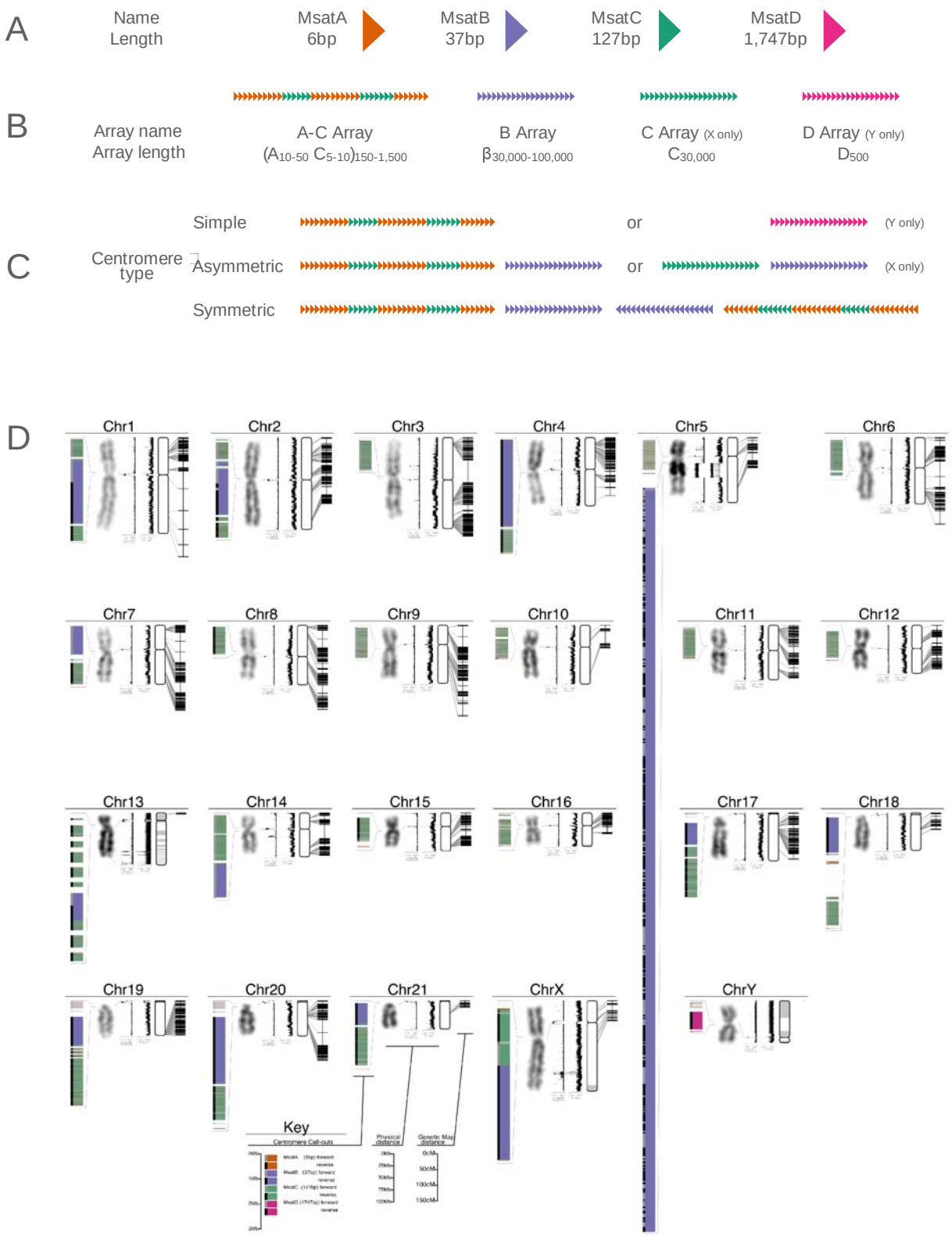


Figure 2: The Mongolian gerbil (*Meriones unguiculatus*) genome. Gerbil centromere types. (A) There are four different repeat types in gerbil centromeres: MsatA (6bp), MsatB (37bp), MsatC

(127bp), 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 Chromosome 2 in Figure 2). (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 which 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 grey or black bar on the left. Chromosome 5 has a large expansion of centromeric repeats in the long arm. All call-outs are drawn to the same scale. (2) The DAPI-banding karyotype image, showing the intra-arm heterochromatin on chromosome 5, and the entirely dark staining on chromosome 13. (3) Linguistic complexity and (4) entropy, both measured in overlapping sliding 10kb windows with a step size of 1kb. For both metrics, a low value indicates highly repetitive or predictable sequence as are characteristic 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 scaffolds shaded alternately white and grey, and (6) a depiction of the genetic map with links between the genetic markers and their physical location. Thin grey lines link the location of similar features on adjacent plots (i.e. centromere callout to karyotype; centromere location in

729 the karyotype to centromere in the linguistic complexity plot; genetic markers to their physical

730 location). A high-resolution copy of panel D can be found in the supplement (Figure SC)

731

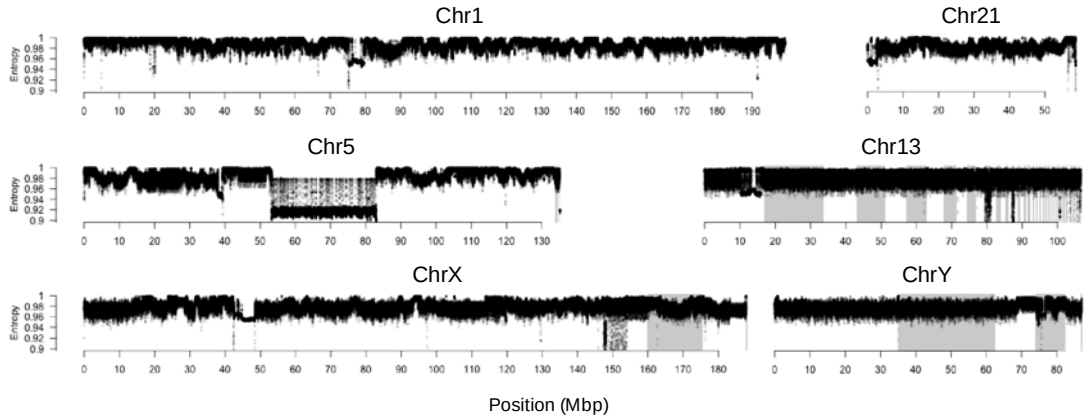


Figure 3: Entropy plots for a selection of chromosomes including the “normal” autosomes 1 and 21, the unusual autosomes 5 and 13, and both sex chromosomes. The unordered scaffolds within a chromosome are shaded alternately white and grey. Note the spatial heterogeneity in chromosomes 1 and 21 that is absent in chromosome 13 and the Y. Indeed, chromosome 13 is the most homogenous chromosome in the gerbil. Plots for every chromosome, as well as GC content, gene density, and linguistic complexity can be found in Figure SC.

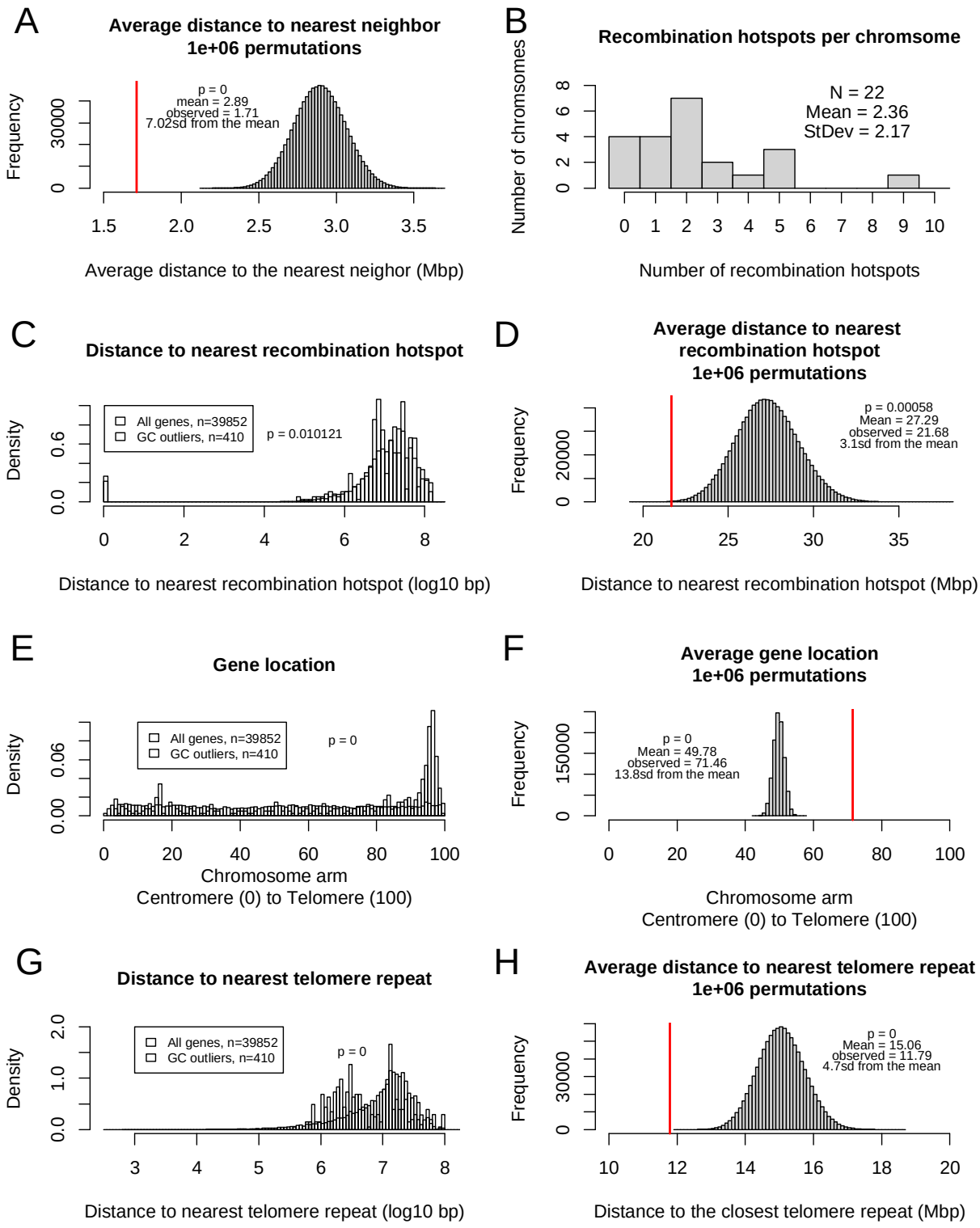
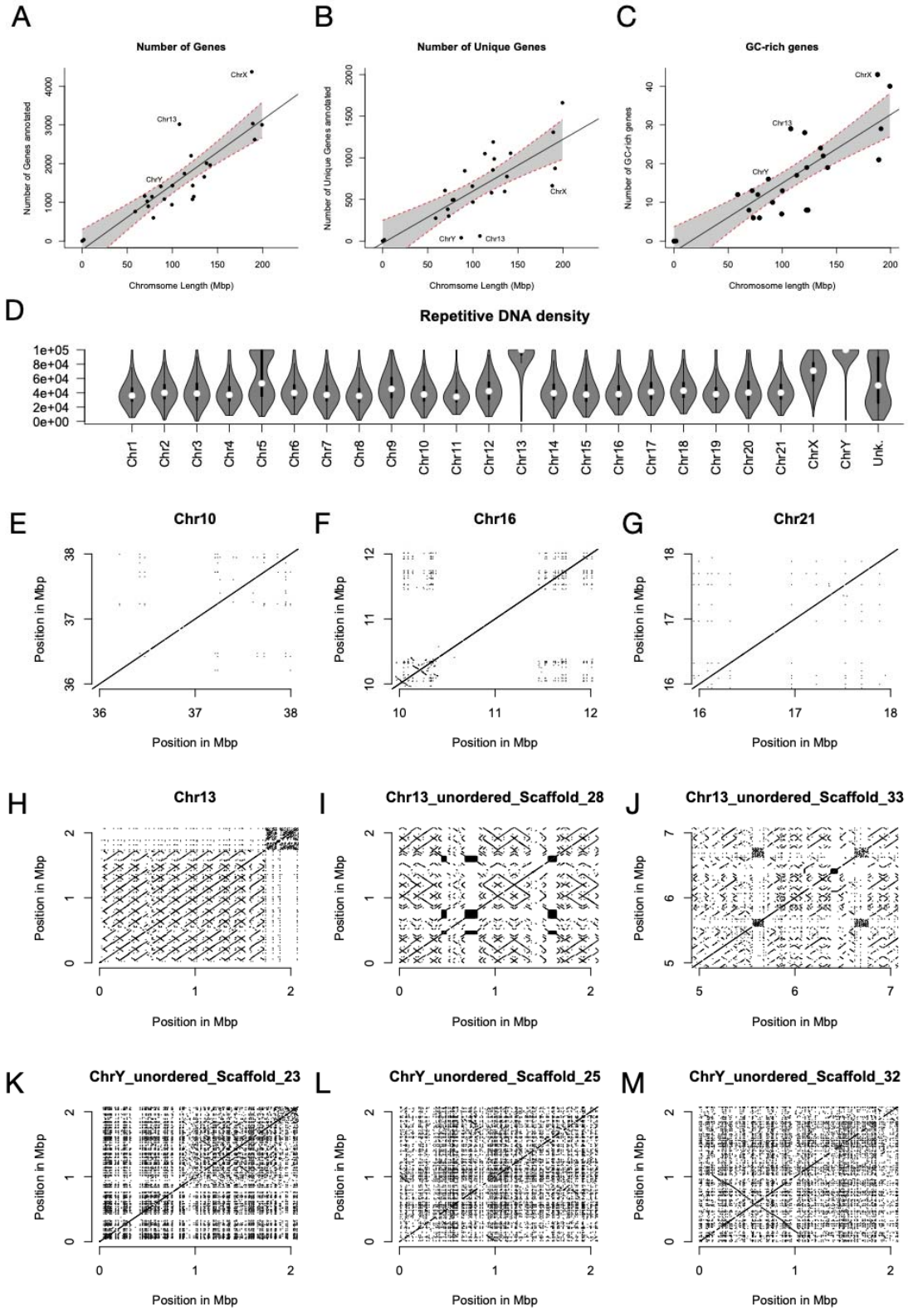


Figure 4: GC-rich genes are nonrandomly distributed in the *M. unguiculatus* genome. We compared the location of the 410 GC rich genes (Pracana et al. 2020)in relation to each other,

743 the nearest recombination hotspot, their location along the chromosome arm, and their proximity
744 to telomere repeats both interstitial and at the ends of chromosome arms. These comparisons
745 were done once against the entire gene set (C, E, G) and again using a permutation test with
746 1,000,000 draws of a random set of 410 genes (A, D, F, H). (A) GC rich genes are clustered in
747 the genome. The observed distance between each outlier gene and its nearest outlier gene
748 neighbor is significantly shorter than those distances between a random group of genes
749 (permutation test, n=1,000,000 permutations, p=0). (B) We identified 52 recombination hotspots
750 spread across 18 of the 22 chromosomes (average = 2.36 hotspots per chromosome). (C, D)
751 GC-rich genes occur closer to recombination hotspots than expected by chance (C: t-test, df =
752 383.2, t = 2.585, p = 0.01012; D: permutation test, n = 1,000,000, p = 0). (E, F) GC rich genes
753 are found closer to the telomere end of chromosome arms than expected by chance (E: t-test,
754 df = 418, t = -14.26, p = 0; F: permutation test, n=1,000,000, p=0). (G, H). GC-rich genes are
755 clustered nearer telomere repeats (interstitial or otherwise) than expected by chance (G: t-test,
756 df = 418.6, t = 7.876, p=0; H: permutation test, n=1,000,000, p=0).

757



758

Figure 5. 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, J) and the Y (K, L, M). Each panel shows a 2Mbp 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, J) are much longer than those on the Y (K, L, M), as expected based on their fundamentally different evolutionary history.

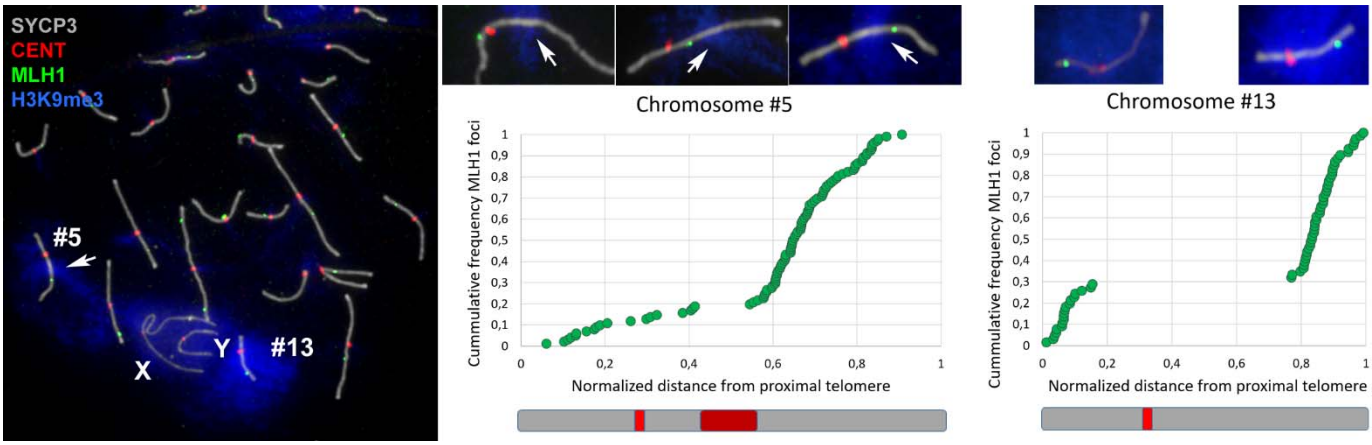


Figure 6 Figure 6. Distribution of recombination events in gerbil spermatocytes. (A) Immunolocalization of SYCP3 protein (grey) on meiotic chromosomes marks the trajectory of the synaptonemal complex along bivalents; trimethylation of histone H3 at lysine 9 (H3K9me3, blue) marks heterochromatin; CENT (red) stains centromeres; and MLH1 (green) marks the sites of crossovers. H3K9me3 is associated with the entirety of chromosome 13 (#13), a large intra-arm 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 purple boxes, respectively, on the schematic chromosomes above 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 towards the distal end. For chromosome 13, MLH1 foci are

conspicuously accumulated towards the chromosomal ends, with an approximate 70:30 distribution on the long and short arms respectively.

References

Accili, D., J. Drago, E. J. Lee, M. D. Johnson, M. H. Cool, P. Salvatore, L. D. Asico, P. A. José, S. I. Taylor, and H. Westphal. 1996. Early neonatal death in mice homozygous for a null allele of the insulin receptor gene. 12:106–109.

Ahmad, S., and C. Martins. 2019. The Modern View of B Chromosomes Under the Impact of High Scale Omics Analyses. Cells 8:156–26.

Altemose, N., G. A. Logsdon, A. V. Bzikadze, P. Sidhwani, S. A. Langley, G. V. Caldas, S. J. Hoyt, L. Uralsky, F. D. Ryabov, C. J. Shew, M. E. G. Sauria, M. Borchers, A. Gershman, A. Mikheenko, V. A. Shepelev, T. Dvorkina, O. Kunyavskaya, M. R. Vollger, A. Rhie, A. M. McCartney, M. Asri, R. Lorig-Roach, K. Shafin, J. K. Lucas, S. Aganezov, D. Olson, L. G. de Lima, T. Potapova, G. A. Hartley, M. Haukness, P. Kerpedjiev, F. Gusev, K. Tigyi, S. Brooks, A. Young, S. Nurk, S. Koren, S. R. Salama, B. Paten, E. I. Rogaev, A. Streets, G. H. Karpen, A. F. Dernburg, B. A. Sullivan, A. F. Straight, T. J. Wheeler, J. L. Gerton, E. E. Eichler, A. M. Phillippy, W. Timp, M. Y. Dennis, R. J. O'Neill, J. M. Zook, M. C. Schatz, P. A. Pevzner, M. Diekhans, C. H. Langley, I. A. Alexandrov, and K. H. Miga. 2022. Complete genomic and epigenetic maps of human centromeres. Science 376:eabl4178.

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

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

Arends, D., P. Prins, R. C. Jansen, and K. W. Broman. 2014. R/qtl: high-throughput multiple QTL mapping. Bioinformatics 26:2990–2992.

Bao, Z., and S. R. Eddy. 2002. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. Genome Res 12:1269–1276.

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

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

Benazzou, T., E. Viegas-Pequignot, M. Prod'Homme, M. Lombard, F. Petter, and B. Dutrillaux. 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., E. Lavezzo, and S. Toppo. 2018. NeSSie: a tool for the identification of approximate DNA sequence symmetries. Bioinformatics 34:2503–2505.

Bignell, G. R., T. Santarius, J. C. M. Pole, A. P. Butler, J. Perry, E. Pleasance, C. Greenman, A. Menzies, S. Taylor, S. Edkins, P. Campbell, M. Quail, B. Plumb, L. Matthews, K. McLay, P. A. W. Edwards, J. Rogers, R. Wooster, P. A. Futreal, and M. R. Stratton. 2007. Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. Genome Res 17:1296–1303.

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

Botero-Castro, F., E. Figuet, M.-K. Tilak, B. Nabholz, and N. Galtier. 2017. Avian Genomes Revisited: Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds. MBE 34:3123–3131.

Brekke, T. D., K. A. Steele, and J. F. Mulley. 2018. Inbred or Outbred? Genetic diversity in laboratory rodent colonies. G3 8:679–686.

Brekke, T. D., S. Supriya, M. G. Denver, A. Thom, K. A. Steele, and J. F. Mulley. 2019. A high-density genetic map and molecular sex-typing assay for gerbils. Mamm Genome 30:63–70.

Broman, K. W., H. Wu, S. Sen, and G. A. Churchill. 2003. R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890.

Calogero, S., F. Grassi, A. Aguzzi, T. Voigtländer, P. Ferrier, S. Ferrari, and M. E. Bianchi. 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, J. P., T. F. Sharbel, and L. W. Beukeboom. 2000. B-chromosome evolution. Philosophical Transactions of the Royal Society B: Biological Sciences 355:163–178.

Campbell, P. J., S. Yachida, L. J. Mudie, P. J. Stephens, E. D. Pleasance, L. A. Stebbings, L. A. Morsberger, C. Latimer, S. McLaren, M.-L. Lin, D. J. McBride, I. Varela, S. A. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A. P. Butler, J. W. Teague, C. A. Griffin, J. Burton, H. Swerdlow, M. A. Quail, M. R. Stratton, C. Iacobuzio-Donahue, and P. A. Futreal. 2010. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature 467:1109–1113.

Chan, P. P., and T. M. Lowe. 2019. Gene Prediction, Methods and Protocols. Methods Mol Biology 1962:1–14.

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

Cheng, H., G. T. Concepcion, X. Feng, H. Zhang, and H. Li. 2020. Haplotype-resolved de novo assembly with phased assembly graphs.

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

880 Craig-Holmes, A. P., and M. W. Shaw. 1971. Polymorphism of Human Constitutive Heterochromatin.
881 Science 174:702–704.

882 Dai, Y., R. Pracana, and P. W. H. Holland. 2020. Divergent genes in gerbils: prevalence, relation to GC-
883 biased substitution, and phenotypic relevance. BMC Evolutionary Biology 1–15.

884 Dhar, M. K., B. Friebe, A. K. Koul, and B. S. Gill. 2002. Origin of an apparent B chromosome by
885 mutation, chromosome fragmentation and specific DNA sequence amplification. Chromosoma
886 111:332–340.

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

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

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

893 Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R.
894 Gingeras. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21.

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

897 Eyre-Walker, A., and L. D. Hurst. 2001. The evolution of isochores. Nat Rev Genet 2:549–555.

898 Flynn, J. M., R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, and A. F. Smit. 2020.
899 RepeatModeler2 for automated genomic discovery of transposable element families. Proc National
900 Acad Sci 117:9451–9457.

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

903 de la Fuente, R. de la, M. Manterola, A. Viera, M. T. Parra, M. Alsheimer, J. S. Rufas, and J. Page. 2014.
904 Chromatin Organization and Remodeling of Interstitial Telomeric Sites During Meiosis in the
905 Mongolian Gerbil (*Meriones unguiculatus*). Genetics 197:1137–1151.

906 de la Fuente, R. de la, M. T. Parra, A. Viera, A. Calvente, R. Gómez, J. Á. Suja, J. S. Rufas, and J. Page.
907 2007. Meiotic Pairing and Segregation of Achiasmate Sex Chromosomes in Eutherian Mammals: The
908 Role of SYCP3 Protein. PLoS Genet 3:e198-12.

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

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

913 Garsed, D. W., A. J. Holloway, and D. M. Thomas. 2009. Cancer□associated neochromosomes: a novel
914 mechanism of oncogenesis. Bioessays 31:1191–1200.

915 Garsed, D. W., O. J. Marshall, V. D. A. Corbin, A. Hsu, L. Di Stefano, J. Schröder, J. Li, Z.-P. Feng, B.
916 W. Kim, M. Kowarsky, B. Lansdell, R. Brookwell, O. Myklebost, L. Meza-Zepeda, A. J. Holloway,
917 F. Pedeutour, K. H. A. Choo, M. A. Damore, A. J. Deans, A. T. Papenfuss, and D. M. Thomas. 2014.
918 The Architecture and Evolution of Cancer Neochromosomes. Cancer Cell 26:653–667.

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

922 Greenman, C., S. Cooke, J. Marshall, M. Stratton, and P. Campbell. 2012. Modelling Breakage-Fusion-
923 Bridge Cycles as a Stochastic Paper Folding Process. Arxiv.

924 Grewal, S. I. S., and D. Moazed. 2003. Heterochromatin and Epigenetic Control of Gene Expression.
925 Science 301:798–802.

926 Gustavsen, C. R., P. Chevret, B. Krasnov, G. Mowlavi, O. D. Madsen, and R. S. Heller. 2008. The
927 morphology of islets of Langerhans is only mildly affected by the lack of Pdx-1 in the pancreas of
928 adult *Meriones* jirds. Gen Comp Endocr 159:241–249.

929 Hargreaves, A. D., L. Zhou, J. Christensen, F. M. taz, S. Liu, F. Li, P. G. Jansen, E. Spiga, M. T. Hansen,
930 S. V. H. Pedersen, S. Biswas, K. Serikawa, B. A. Fox, W. R. Taylor, J. F. Mulley, G. Zhang, R. S.
931 Heller, and P. W. H. Holland. 2017. Genome sequence of a diabetes-prone rodent reveals a mutation
932 hotspot around the ParaHox gene cluster. PNAS 12:201702930–6.

933 Hron, T., P. Pajer, J. Pačes, P. Bartůněk, and D. Elleder. 2015. Hidden genes in birds. Genome Biol
934 16:164.

935 Hubley, R., R. D. Finn, J. Clements, S. R. Eddy, T. A. Jones, W. Bao, A. F. A. Smit, and T. J. Wheeler.
936 2016. The Dfam database of repetitive DNA families. Nucleic Acids Res 44:D81–D89.

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

939 Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. Repbase
940 Update, a database of eukaryotic repetitive elements. 110:462–467.

941 Kapitonov, V. V., and J. Jurka. 2008. A universal classification of eukaryotic transposable elements
942 implemented in Repbase. Nat Rev Genet 9:411–412.

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

945 Keller, O., M. Kollmar, M. Stanke, and S. Waack. 2011. A novel hybrid gene prediction method
946 employing protein multiple sequence alignments. Bioinformatics 27:757–763.

Kikuta, H., M. Laplante, P. Navratilova, A. Z. Komisarczuk, P. G. Engstrom, D. Fredman, A. Akalin, M. Caccamo, I. Sealy, K. Howe, J. Ghislain, G. Pezeron, P. Mourrain, S. Ellingsen, A. C. Oates, C. Thisse, B. Thisse, I. Foucher, B. Adolf, A. Geling, B. Lenhard, and T. S. Becker. 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res. 17:545–555.

Knight, L. I., B. L. Ng, W. Cheng, B. Fu, F. Yang, and R. V. Rambau. 2013. Tracking chromosome evolution in southern African gerbils using flow-sorted chromosome paints. Cytogenet Genome Res 139:267–275.

Korf, I. 2004. Gene finding in novel genomes. Bmc Bioinformatics 5:59–59.

Kuderna, L. F. K., E. Lizano, E. Julià, J. Gomez-Garrido, A. Serres-Armero, M. Kuhlwilm, R. A. Alandes, M. Alvarez-Estape, D. Juan, H. Simon, T. Alioto, M. Gut, I. Gut, M. H. Schierup, O. Fornas, and T. Marques-Bonet. 2019. Selective single molecule sequencing and assembly of a human Y chromosome of African origin. Nat Comm 10:1–8.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. 2004. Versatile and open software for comparing large genomes. Genome Biol 5:R12-9.

Lamb, B. C. 1984. The properties of meiotic gene conversion important in its effects on evolution. Heredity 53:113–138.

Leibowitz, G., S. Ferber, A. Apelqvist, H. Edlund, D. J. Gross, E. Cerasi, D. Melloul, and N. Kaiser. 2001. IPF1/PDX1 Deficiency and β-Cell Dysfunction in *Psammomys obesus*, an Animal With Type 2 Diabetes. Diabetes 50:1799–1806.

Lercher, M. J., N. G. C. Smith, A. Eyre-Walker, and L. D. Hurst. 2002. The Evolution of Isochores: Evidence From SNP Frequency Distributions. Genetics 162:1805–1810.

Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. 25:1754–1760.

Li, H., P. S. Zeitler, M. T. Valerius, K. Small, and S. S. Potter. 1996. Gsh□1, an orphan Hox gene, is required for normal pituitary development. Embo J 15:714–724.

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

Malik, H. S. 2009. The Centromere-Drive Hypothesis: A Simple Basis for Centromere Complexity. Pp. 33–52 *in* Đ. Ugarković, ed. Centromere, Progress in Molecular and Subcellular Biology.

Manni, M., M. R. Berkeley, M. Seppey, F. A. Simão, and E. M. Zdobnov. 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., and M. P. Colaiácovo. 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.

Murchison, E. P., O. B. Schulz-Trieglaff, Z. Ning, L. B. Alexandrov, M. J. Bauer, B. Fu, M. Hims, Z. Ding, S. Ivakhno, C. Stewart, B. L. Ng, W. Wong, B. Aken, S. White, A. Alsop, J. Becq, G. R. Bignell, R. K. Cheetham, W. Cheng, T. R. Connor, A. J. Cox, Z.-P. Feng, Y. Gu, R. J. Grocock, S. R. Harris, I. Khrebtukova, Z. Kingsbury, M. Kowarsky, A. Kreiss, S. Luo, J. Marshall, D. J. McBride, L. Murray, A.-M. Pearse, K. Raine, I. Rasolonjatovo, R. Shaw, P. Tedder, C. Tregidgo, A. J. Vilella, D. C. Wedge, G. M. Woods, N. Gormley, S. Humphray, G. Schroth, G. Smith, K. Hall, S. M. J. Searle, N. P. Carter, A. T. Papenfuss, P. A. Futreal, P. J. Campbell, F. Yang, D. R. Bentley, D. J. Evers, and M. R. Stratton. 2012. Genome Sequencing and Analysis of the Tasmanian Devil and Its Transmissible Cancer. Cell 148:780–791.

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

Offield, M. F., T. L. Jetton, P. A. Labosky, M. Ray, R. W. Stein, M. A. Magnuson, B. L. Hogan, and C. V. Wright. 1996. PDX-1 is required for pancreatic outgrowth and differentiation of the rostral duodenum. Development 122:983–995.

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

Pakes, S. P. 1969. The somatic chromosomes of the Mongolian gerbil (*Meriones unguiculatus*). Naval Aerospace Medical Institute, Naval Aerospace Medial Center Vol 1056.

Penagos-Puig, A., and M. Furlan-Magaril. 2020. Heterochromatin as an Important Driver of Genome Organization. Frontiers Cell Dev Biology 8:579137.

Peters, A. H. F. M., A. W. Plug, M. J. van Vugt, and P. de Boer. 1997. A drying-down technique for the spreading of mammalian meiocytes from the male and female germline. Chromosome Res 5:66–68.

Pracana, R., A. D. Hargreaves, J. F. Mulley, and P. W. H. Holland. 2020. Runaway GC Evolution in Gerbil Genomes. MBE 37:2197–2210.

Price, A. L., N. C. Jones, and P. A. Pevzner. 2005. De novo identification of repeat families in large genomes. Bioinformatics 21:i351–i358.

Qumsiyeh, M. B. 1986a. Phylogenetic Studies of the Rodent Family Gerbillidae: I. Chromosomal Evolution in the Southern African Complex. JMamm 67:680–692.

Qumsiyeh, M. B. H. 1986b. Chromosomal Evolution in the rodent family Gerbillidae.

1017 Rochette, N. C., A. G. Rivera-Colón, and J. M. Catchen. 2019. Stacks 2: Analytical Methods for Paired-
1018 end Sequencing Improve RADseq-based Population Genomics. bioRxiv 32:314–37.

1019 Saksouk, N., E. Simboeck, and J. Déjardin. 2015. Constitutive heterochromatin formation and
1020 transcription in mammals. Epigenet Chromatin 8:3.

1021 Singhal, S., E. M. Leffler, K. Sannareddy, I. Turner, O. Venn, D. M. Hooper, A. I. Strand, Q. Li, B.
1022 Raney, C. N. Balakrishnan, S. C. Griffith, G. McVean, and M. Przeworski. 2015. Stable
1023 recombination hotspots in birds. Science 350:928–932.

1024 Solari, A. J., and T. Ashley. 1977. Ultrastructure and behavior of the achiasmatic, telosynaptic XY pair of
1025 the sand rat (*Psammomys obesus*). Chromosoma 62:319–336.

1026 Srikulnath, K., S. F. Ahmad, W. Singchat, and T. Panthum. 2021. Why Do Some Vertebrates Have
1027 Microchromosomes? Cells 10:2182.

1028 Tarailo□Graovac, M., and N. Chen. 2009. Using RepeatMasker to Identify Repetitive Elements in
1029 Genomic Sequences. Curr Protoc Bioinform 25:4.10.1-4.10.14.

1030 Tiemann-Boege, I., T. Schwarz, Y. Striedner, and A. Heissl. 2017. The consequences of sequence erosion
1031 in the evolution of recombination hotspots. Philosophical Transactions Royal Soc B Biological Sci
1032 372:20160462.

1033 Tilak, M.-K., F. Botero-Castro, N. Galtier, and B. Nabholz. 2018. Illumina Library Preparation for
1034 Sequencing the GC-Rich Fraction of Heterogeneous Genomic DNA. Genome Biology and Evolution
1035 10:616–622.

1036 Vinogradov, A. E. 2005. Dualism of gene GC content and CpG pattern in regard to expression in the
1037 human genome: magnitude versus breadth. Trends Genet 21:639–643.

1038 Volobouev, V., V. M. Aniskin, B. Sicard, G. Dobigny, and L. Granjon. 2007. Systematics and phylogeny
1039 of West African gerbils of the genus *Gerbilliscus* (Muridae: Gerbillinae) inferred from comparative G-
1040 and C-banding chromosomal analyses. Cytogenet Genome Res 116:269–281.

1041 Waters, P. D., H. R. Patel, A. Ruiz-Herrera, L. Álvarez-González, N. C. Lister, O. Simakov, T. Ezaz, P.
1042 Kaur, C. Frere, F. Grützner, A. Georges, and J. A. M. Graves. 2021. Microchromosomes are building
1043 blocks of bird, reptile, and mammal chromosomes. P Natl Acad Sci Usa 118:e2112494118.

1044 Weiss, L., K. Mayeda, and M. Dully. 1970. The Karyotype of the Mongolian Gerbil, *Meriones*
1045 *unguiculatus*. Cytologia 35:102–106.

1046 Withers, D. J., J. S. Gutierrez, H. Towery, D. J. Burks, J.-M. Ren, S. Previs, Y. Zhang, D. Bernal, S. Pons,
1047 G. I. Shulman, S. Bonner-Weir, and M. F. White. 1998. Disruption of IRS-2 causes type 2 diabetes in
1048 mice. Nature 391:900–904.

1049 Yang, F., V. Trifonov, B. L. Ng, N. Kosyakova, and N. P. Carter. 2017. Generation of Paint Probes by
1050 Flow-Sorted and Microdissected Chromosomes. Pp. 35–52 *in* Fluorescence In Situ Hybridization
1051 (FISH)—Application Guide. Springer.

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

Zorio, D. A. R., S. Monsma, D. H. Sanes, N. L. Golding, E. W. Rubel, and Y. Wang. 2019. De novo sequencing and initial annotation of the Mongolian gerbil (*Meriones unguiculatus*) genome. Genomics 111:441–449.