Single-library chromosome-scale diploid assemblies of vole genomes resolve a species-specific duplication implicated in pair bonding

Mohamed Abuelanin[1], Gulhan Kaya[1], Juniper A. Lake[2], Christine Lambert[2], Melody V. Wu[3], Kristen M. Berendzen[8, 9], Ksenia Krasheninnikova[4], Jonathan M.D. Wood[4], Nancy G. Solomon[5], Zoe R. Donaldson[6], Karen L. Bales[7], Kerstin Howe[4], Jonas Korlach[2], Devanand Manoli[8, 9], Jessica Tollkuhn[3], Megan Y. Dennis[1†].

Running title: Single-library vole assemblies

[1] Genome Center, MIND Institute, and Department of Biochemistry & Molecular Medicine, University of California, Davis, CA 95616, USA.

[2] Pacific Biosciences, Menlo Park, CA 94025, USA.

[3] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA.

[4] Wellcome Sanger Institute, Hinxton CB10 1SA, UK.

[5] Department of Biology, Miami University of Ohio, Miami, OH, USA.

[6] Department of Molecular, Cellular, and Developmental Biology, Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, CO, USA.

[7] Departments of Psychology; Neurobiology, Physiology, and Behavior; and California National Primate Research Center, University of California, Davis, Davis, CA, USA.

[8] Department of Psychiatry and Behavioral Sciences, [9]Weill Institute for Neuroscience, University of California, San Francisco, San Francisco, CA, USA.

[†]Corresponding author:
Megan Y. Dennis, Ph.D.
University of California, Davis, School of Medicine
One Shields Avenue
Genome Center, 4303 GBSF
Davis, CA 95616
Email: mydennis@ucdavis.edu

Keywords: Long-read sequencing, Chromosome conformation capture, 3C, HiFi, CiFi, genome assembly, genome scaffolding, prairie vole, meadow vole, mate-pair bonding, vasopressin

SUMMARY

High-quality reference genomes are essential to effectively characterize genomic drivers of speciation, phenotypic diversity, and disease causality. Larger complex genomes often require integration of long-read DNA sequencing with additional genomic data, such as chromosome conformation capture (Hi-C or CiFi) to generate phased chromosome-scale assemblies, however this requires multiple sequencing platforms (in the case of Hi-C) or the construction of multiple long-read sequencing libraries. Here, we devise a strategy that combines PacBio HiFi and CiFi sequencing in a single library and run to efficiently produce high-quality contiguous chromosome-scale diploid genome assemblies. We apply this approach to liver tissue from single individuals of prairie vole (*Microtus ochrogaster*) and meadow vole (*Microtus pennsylvanicus*), generating haplotype-resolved, chromosome-scale 2.3 Gbp genomes with QV~62, and 99.3% BUSCO completeness. Comparing the two new genomes identifies complex structural changes impacting *Avpr1a*, previously implicated in pair bonding, including a species-specific duplication missing from the existing prairie vole reference genome. These divergent genomic features offer new avenues of investigation related to behavioral divergence between prairie and meadow voles. This single-library approach facilitates a simplified and more affordable assembly workflow, producing near-complete genomes of diverse species using one sequencing platform.

INTRODUCTION

High-quality reference genomes are foundational for characterizing the genomic basis of speciation, phenotypic diversity, and disease susceptibility [1]. However, repetitive and structurally complex regions, including segmental duplications (SDs) and centromeres, have historically been underrepresented or absent from reference assemblies. These regions are recognized as a primary source of new gene functions for adaptive evolution [2], yet their accurate assembly remains technically challenging. Recent advances in long-read sequencing, particularly PacBio HiFi technology, have transformed *de novo* genome assembly by producing highly accurate reads exceeding 10 kbp [3]. When combined with trio-binning approaches that leverage parental short-read data, genome-assembly approaches can generate fully phased, haplotype-resolved assemblies [4]. Achieving chromosome-scale contiguity, however, typically requires chromosome conformation capture (3C) data. Genome-wide 3C (Hi-C) depends on short-read sequencing, which maps poorly to repetitive regions and necessitates multi-platform library preparations [5,6]. CiFi addresses these limitations by integrating 3C libraries with PacBio HiFi long-read sequencing, validated on assembling the Mediterranean fruit fly genome (~600 Mbp) and demonstrating effective scaffolding and improved signal-to-noise ratio compared with Hi-C [7].

Here, we extend this combined HiFi-CiFi assembly approach to a further optimization of the workflow and two mammalian genomes approximately four-fold larger (~2.3 Gbp): prairie vole (*Microtus ochrogaster*) and meadow vole (*Microtus pennsylvanicus*). Prairie voles are among the few mammalian species exhibiting social monogamy [8,9]. They form enduring social attachments to their partners, known as pair bonds [10], and both parents care for offspring. In contrast, closely related meadow voles, which diverged from a common ancestor with prairie voles <4 million years ago (mya) [11], do not show social monogamy, and only mothers care for offspring. Thus, the neurogenetic differences between vole species offer a powerful system to examine the mechanistic basis of complex social behaviors. Across taxa, neuropeptide hormones represent a critical substrate in the evolution of social behaviors, modulating the neural circuits and underlying physiology that control social interactions [12,13]. Seminal comparative

behavioral and pharmacologic studies have implicated the nonapeptide hormones oxytocin and vasopressin in prairie vole social attachment [14–17]. Oxytocin receptor (*Oxtr*) and vasopressin receptor 1a (*Avpr1a*) exhibit divergent brain expression patterns in prairie voles relative to other vole species, suggesting substantial regulatory divergence at these loci [18,19].

The existing prairie vole reference (MicOch1.0; Broad Institute, 2012) predates modern long-read technologies—generated from a 94× coverage Illumina sequence of a female individual—and lacks chromosome-scale contiguity and haplotype resolution. Notably, a >105 kbp prairie-vole duplication of *Avpr1a*, identified by targeted BAC sequencing [20], was not present in MicOch1.0. Subsequent work on the *Avpr1a* paralog has been hampered, including verification of its existence in prairie vole or other related species, due to the complex nature of this locus. A chromosome-scale meadow vole assembly was recently generated by the Vertebrate Genomes Project (VGP mMicPen1 [1]) but has not been compared to the prairie vole in the context of social behavior genetics. Using a new workflow that combines HiFi and CiFi DNA for a single library preparation and sequencing run per vole species, we produce haplotype-resolved, chromosome-scale references enabling a systematic assessment of species divergence across the *Avpr1a* locus.

RESULTS

HiFi-CiFi single-library sequencing and genome assembly

To generate HiFi and CiFi data from a single PacBio sequencing library, we isolated high-molecular-weight genomic DNA (gDNA) and fixed chromatin from liver tissue of an individual male prairie vole (*Microtus ochrogaster*) and male meadow vole (*Microtus pennsylvanicus*), respectively (Figure 1A). HiFi gDNA was sheared to a target average fragment size of 16 kbp, while CiFi DNA was prepared using the restriction enzyme HindIII and standard protocol [7], yielding average fragment sizes of ~8 kbp (Figure S1). The two preparations were pooled at a 90:10 molar ratio (HiFi:CiFi), constructed into a single library, and sequenced on one Revio SMRT Cell per species, producing 102.7 Gbp (prairie vole) and 98.2 Gbp (meadow vole) of data at a median read quality of QV38.

HiFi (88%) and CiFi (12%) reads were segregated using unique Ampli-Fi adapter sequences flanking CiFi fragments (Methods, Figure 1A). HiFi reads showed median lengths of 15.6 kbp (prairie vole) and 15.2 kbp (meadow vole) at 35–40× predicted coverage, while CiFi reads were shorter, as expected, at 7.9 kbp (prairie vole) and 7.2 kbp (meadow vole) at 5× predicted coverage (Figure 1B, Table S1). CiFi reads were then converted from multi-segment concatemers into chromatin-contact pairs: 1.37 million CiFi reads yielded 23.4 million contact pairs for prairie vole (median segment size 1.4 kbp), and 1.54 million reads yielded 8.1 million pairs for meadow vole (median segment size 1.7 kbp). The greater contact-pair yield for prairie vole reflects slightly smaller segment sizes and longer CiFi read lengths. Together, these results demonstrate that a combined HiFi and CiFi library produce high-quality data.

In order to assess CiFi as a single-platform alternative to Hi-C for *de novo* genome assembly, we generated parallel short-read Hi-C data from the same liver tissues using the standard DpnII restriction enzyme (~49 Gbp, 20× for prairie vole; ~73 Gbp, 30× for meadow vole). HiFi reads with either CiFi-derived contact pairs or Hi-C read pairs were used to generate phased diploid contig assemblies with hifiasm [4], then scaffolded with YAHS [21] (Figure 1C). CiFi-phased contigs were longer in three of four

haplotypes and consolidated into fewer final scaffolds, approaching expected chromosome numbers more closely than Hi-C (Figures S2 and S3, Table S2). For example, prairie vole Hap1 yielded 63 scaffolds with CiFi versus 171 with Hi-C (Figure S3). Hi-C also accumulated 1.7–2× more scaffold joins across YAHS rounds yet produced more final scaffolds, and introduced approximately twice the gap sequence across all four assemblies (Figure S4), suggesting that many Hi-C joins did not persist in the final assembly. Where Hi-C showed a higher scaffold N50 in the meadow vole, this was accompanied by more gaps. Overall, CiFi achieved better or equivalent scaffold consolidation, supporting its utility as a single-platform approach for chromosome-resolved diploid *de novo* assembly.

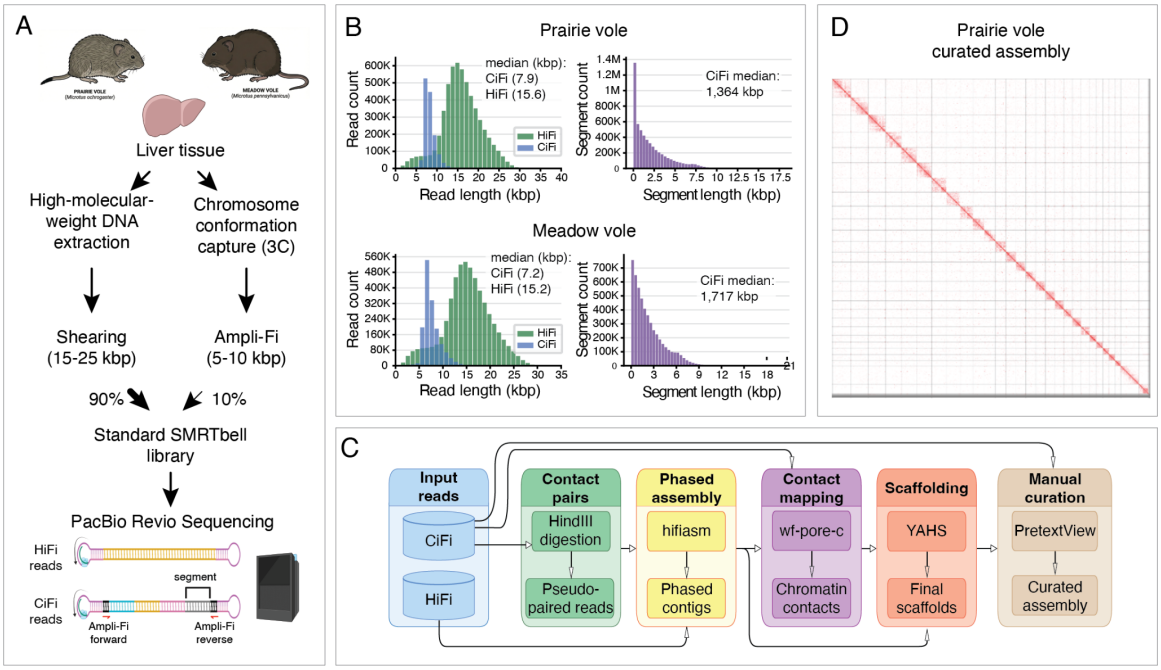

Figure 1. Single-library HiFi-CiFi workflow and sequencing stats for workflow prairie and meadow vole. (A) Library workflow from liver tissue: high-molecular-weight DNA is used to generate HiFi whole-genome reads, and a 3C library generates CiFi reads; both are combined in a single SMRTbell library (90:10 HiFi:CiFi). **(B)** Distribution of read length (HiFi and CiFi) and segment length (CiFi) for prairie (top) and meadow (bottom) voles. **(C)** Computational pipeline for HiFi-CiFi-based genome assembly and scaffolding. **(D)** CiFi contact map of the prairie vole curated Hap1 assembly. See Figures S3, S4, and S5 for contact maps of pre- and post-curated assemblies for both haplotypes and vole species.

We next downsampled CiFi reads from 100% (5× coverage) to 1% in each species to determine the minimum input required to generate contiguous genomes (Figure S5). Assembly sizes remained stable across all conditions and haplotypes, while scaffold N50 and counts were both sensitive to reduced CiFi depth. All four haplotypes maintained near-optimal scaffold N50 at 40% input. Together, these results indicate that approximately 5 Gbp of CiFi data (~2× genome coverage) is sufficient for chromosome-scale scaffolding in mammalian genomes of this size and complexity and achievable at even lower input depending on the species [7].

Final curated vole assemblies

Assemblies produced with HiFi and CiFi required minimal correction during manual curation (see Methods), including five scaffold breaks and four joins for prairie vole and, similarly, five scaffold breaks and ten joins for meadow vole (Figures 1D, S2, S3, and S6). The sex chromosomes were identified for

each species (represented in Hap1), with the remaining autosomes named according to descending size. Assessment of curated assemblies show both haplotypes of each diploid vole assembly to be contiguous (contig N50 > 39 Mbp; scaffold N50 > 89 Mbp), complete (genomic BUSCO ≥99%), and high quality [22] (QV ~62) (Tables 1 and S3, Figure S7).

Table 1. Metric summary of prairie and meadow vole curated genomes

Metric	Prairie vole (this study)		Prairie vole (MicOch1.0)	Meadow vole (this study)		Meadow vole (mMicPen1)	
Haplotype	**Hap1***	**Hap2**	**merged**	**Hap1***	**Hap2**	**Hap1***	**Hap2**
Total size	2.53 Gbp	2.30 Gbp	2.29 Gbp	2.36 Gbp	2.15 Gbp	2.37 Gbp	2.16 Gbp
No. contigs	235	156	187,012	200	134	240	196
Contig N50	39.2 Mbp	51.0 Mbp	29.2 kbp	46.5 Mbp	43.6 Mbp	47.2 Mbp	63.6 Mbp
No. scaffolds	97	62	6,335	105	52	180	143
Scaffold N50	89.2 Mbp	89.4 Mbp	61.8 Mbp	125.8 Mbp	115.6 Mbp	125.3 Mbp	118.4 Mbp
auN	98.6 Mbp	97.5 Mbp	56.0 Mbp	119.3 Mbp	120.1 Mbp	118.7 Mbp	120.3 Mbp
Scaffold L50	11	10	14	8	8	8	8
Scaffold L90	21	22	70	18	21	18	21
Gaps (%)	0.073%	0.039%	8.001%	0.006%	0.173%	< 0.001%	< 0.001%
Genome BUSCO	99.7%	97.2%	98.9%	99.7%	97.0%	99.7%	97.1%

* Primary assemblies with both sex chromosomes

The prairie vole diploid assembly represents a complete karyotype (2n = 54) [23,24] with 52 autosomes, chrX, and chrY (Total: 4.78 Gbp, Hap1: 2.50 Gbp). Telomeres were detected on all but three chromosomes (mean 3.5 gaps/chr). Of these, 31 represent T2T scaffolds and three gapless contigs (overall 63%) (Figure 2, Table S4). In contrast, the current prairie vole reference (MicOch1.0)—a haploid assembly using short-read sequencing of a female individual—comprises 18 chromosomes (17 autosomes and chrX), 10 linkage groups, and four unlocalized scaffolds (1.66 Gbp). 632 Mbp of sequence is represented as 6,303 unplaced scaffolds versus ~30 Mbp in 69 unplaced scaffolds in prairie-vole Hap1 in this study. Comparing the assemblies also revealed major improvements in contiguity (e.g., contig N50: 29.2 kbp MicOch1.0 vs. 39.2 Mbp Hap1) and reduced gap content (8.0% vs. 0.07%), with 238 previously unplaced MicOch1.0 scaffolds (520.8 Mbp) now incorporated in Hap1 chromosome-scale scaffolds (Figure S8, Tables S5 and S6).

Similarly, the meadow vole assembly matches the expected karyotype (2n=46) [25,26] across 44 autosomes, chrX, and chrY, representing 4.47 Gbp total sequence (Hap1: 2.34 Gbp). All but two chromosomes have at least one telomere detected (mean 3.4 gaps/chr), with 17 T2T scaffolds and six gapless contigs (overall 50%; Figure 2, Table S4). A majority of the meadow vole autosomes are telocentric (42/44), with the two sex chromosomes classified as subtelocentric [25,26], likely contributing to the higher proportion of single telomeres detected in the meadow versus prairie vole. This assembly is on par with the recent long-read VGP meadow vole reference (mMicPen1), which has all expected chromosomes at near equal contiguity (scaffold Hap1 N50: 125.8 Mbp vs. this study 125.3 Mbp) but a larger number of unanchored sequences (n=156 at 46.1 Mbp vs. this study n=81 at 19.7 Mbp; Figure S9).

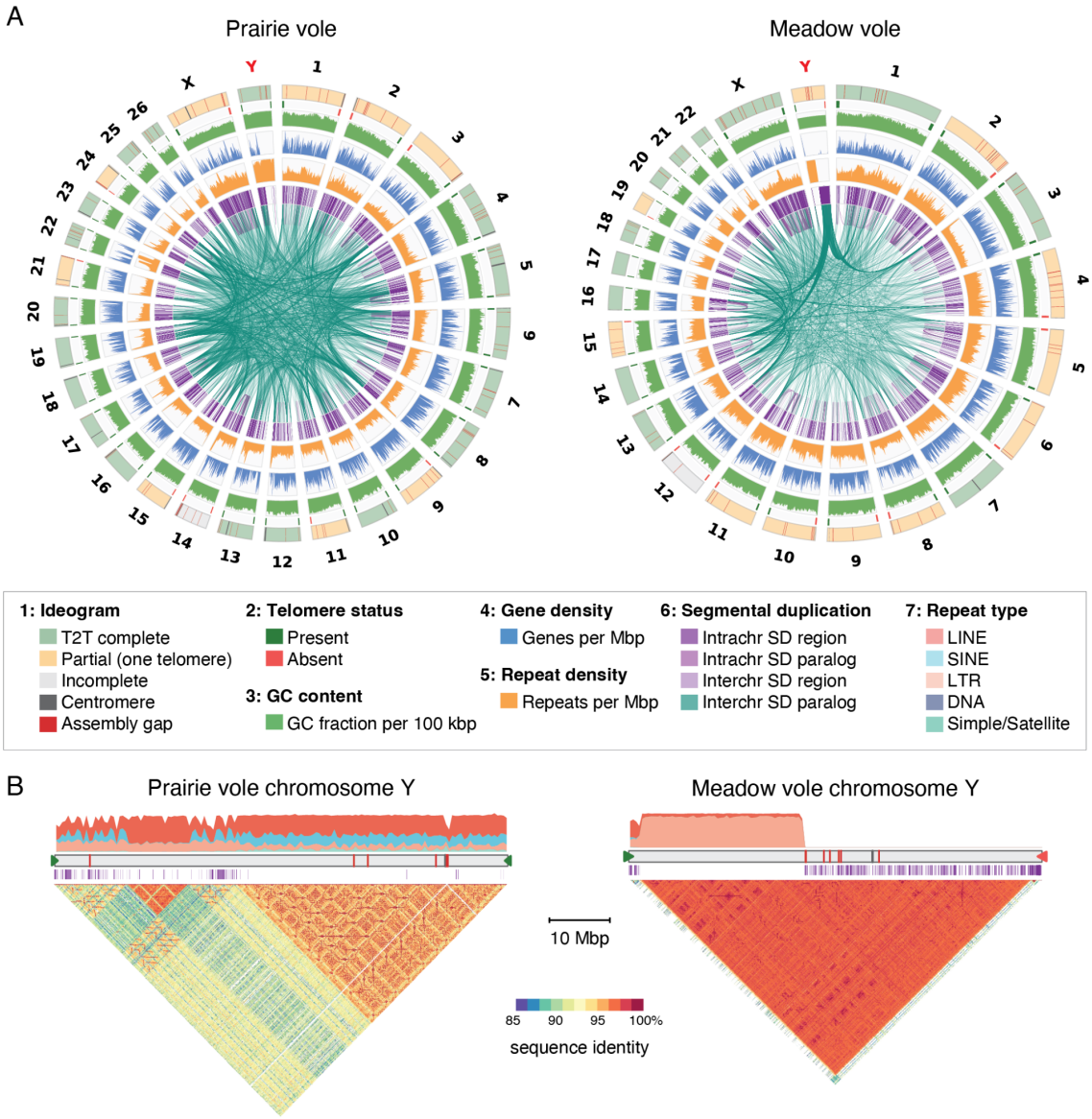

Figure 2. Prairie and meadow vole genome assembly characteristics. (A) Circos plots depicting chromosome characteristics and content of Hap1 curated assemblies for prairie (left) and meadow (right) vole assemblies. Tracks are described in the legend from outer ring (1: Ideogram) to inner ring (6: Segmental duplication (SD)). **(B)** Repeat and SD distribution of both genomes are depicted above and below the Hap1 chromosome Y ideogram for each vole species, respectively. Annotations on the ideogram use colors depicted in the legend from (A), including presence of centromere and assembly gap (legend 1), telomere status (legend 2), SD (purple; legend 6), and repeat types (legend 7). Below the ideogram are dot plots for each species' chromosome Y, with colors representing percent identities. Annotated ideograms of all Hap1 chromosomes are shown in Figures S10 and S11.

Moving forward, we characterized primary Hap1 chromosomes, comprising single sets of autosomes and both sex chromosomes. Using RNA-seq data from liver, brain, and testes of adults and e11.5 (prairie vole) or e13.5 (meadow vole) embryos, we annotated ~20,500 protein-coding genes per species (Methods, Figure 2, Table S7). Repeat annotation revealed an increased proportion of interspersed repeats in prairie vole (e.g., L1 elements representing 16.7% vs. 13.0% of the genome), satellite repeats (8.62 Mbp vs. 511 kbp), and unclassified repeats (>40 Mbp more than meadow vole) (Figures 2A, S10, S11,

Table S8). Near-equal numbers of segmental duplications (SDs; regions >1 kbp at >90% sequence identity [27,28]) were identified in both genomes (~5.3K), with 72–79% intersecting an annotated gene (prairie vole: 1,194 genes; meadow vole: 1,151 genes). Despite this similarity, prairie vole harbored nearly double the duplicated sequence (54 Mbp, 2.15% of genome) compared to meadow vole (28 Mbp, 1.2%) (Figures 2A, S10, S11, Table S9). This expanded SD content is driven by larger duplicons (15.2 kbp vs. 8.21 kbp) and a greatly expanded interchromosomal SD fraction (34.7 Mbp, 63.7% of pairs vs. 9.7 Mbp, 46.2%).

We next examined the sex chromosomes given their known rapid and repeated structural remodeling in *Microtus* [29]. The most notable differences are evident on the Y chromosome (Figure 2B): both species showed gene depletion and likely constitutive heterochromatin spanning either half (prairie vole) or the entirety (meadow vole) of the chromosome. The remaining half of the prairie vole chrY comprises tandemly arrayed SDs of a *Usp9y* mouse homolog, while the meadow vole chrY also carries considerable SD content. Direct sequence comparison reveals no orthology (Figure 3A), indicating that homologous Y chromosomes can reach complete sequence divergence over remarkably short evolutionary timescales.

Comparison of vole genomes identifies *Avpr1a* prairie-vole specific duplication

To better match orthologous regions, we re-oriented Hap1 for both vole assemblies based on synteny to the prairie vole reference (MicOch1.0) and linkage mapping (Tables S5, S10, and S11). Alignment between the two species shows ~98% similarity with notable cytogenetic differences, including several fusions and fissions, impacting 14 chromosomes in each genome (Figure 3A). Large-scale inversions are evident on homologous chromosomes 1 and X, respectively. Mapping single-nucleotide, indel, and structural variants identified ~462K coding variants using prairie vole as a reference [30], of which 17,466 were annotated as likely gene disrupting (LGD) impacting 2,418 genes (Table S12). Focusing on candidate genes implicated in behavior—including those known to interact with nonapeptide hormones oxytocin and vasopressin—show global reduced rates of substitutions relative to synonymous mutations (Ka/Ks << 1; Figure S12) and no LGD variants obviously impacting function, suggesting purifying selection (Tables S13 and S14).

Understanding that gene duplication is a common mechanism of trait innovation across the animal kingdom [31], we identified 306 and 389 gene duplications unique to prairie or meadow vole, respectively (Table S15). Intersecting with our candidate genes list, we identified a prairie-vole-specific duplication of *Avpr1a*, encoding vasopressin V1A receptor, a transmembrane G-protein coupled receptor that binds arginine vasopressin. While this duplication was characterized via BAC sequencing nearly 15 years ago [20], both *Avpr1a* paralogs are missing in the current prairie vole reference (MicOch1.0) impeding genomic analyses of these genes. Each prairie vole haplotype shows *Avpr1a* paralogs residing ~900 kbp apart on chromosome 26, separated proximally by a large stretch of nearly identical satellite repeats not present at the syntenic chr20 locus of the meadow vole (see chain alignment for Meadow vole Hap1; Figures 3B–3D). The *Avpr1a* paralogs flank the chr26 annotated metacentric centromere, operationally defined here as the largest stretch of satellite sequence per chromosome [32], which is not observed in chromosome 20 of the meadow vole (annotated as telocentric). A search for this sequence identifies matching satellite repeat sequences on chromosomes 4 and X (~500 bp to 3 kbp in size at <93% sequence identity) for both Hap1 and Hap2 meadow vole assemblies but is found only at the chr26 *Avpr1a* locus in both prairie vole haplotypes (547 kbp and 517 kbp in size, respectively).

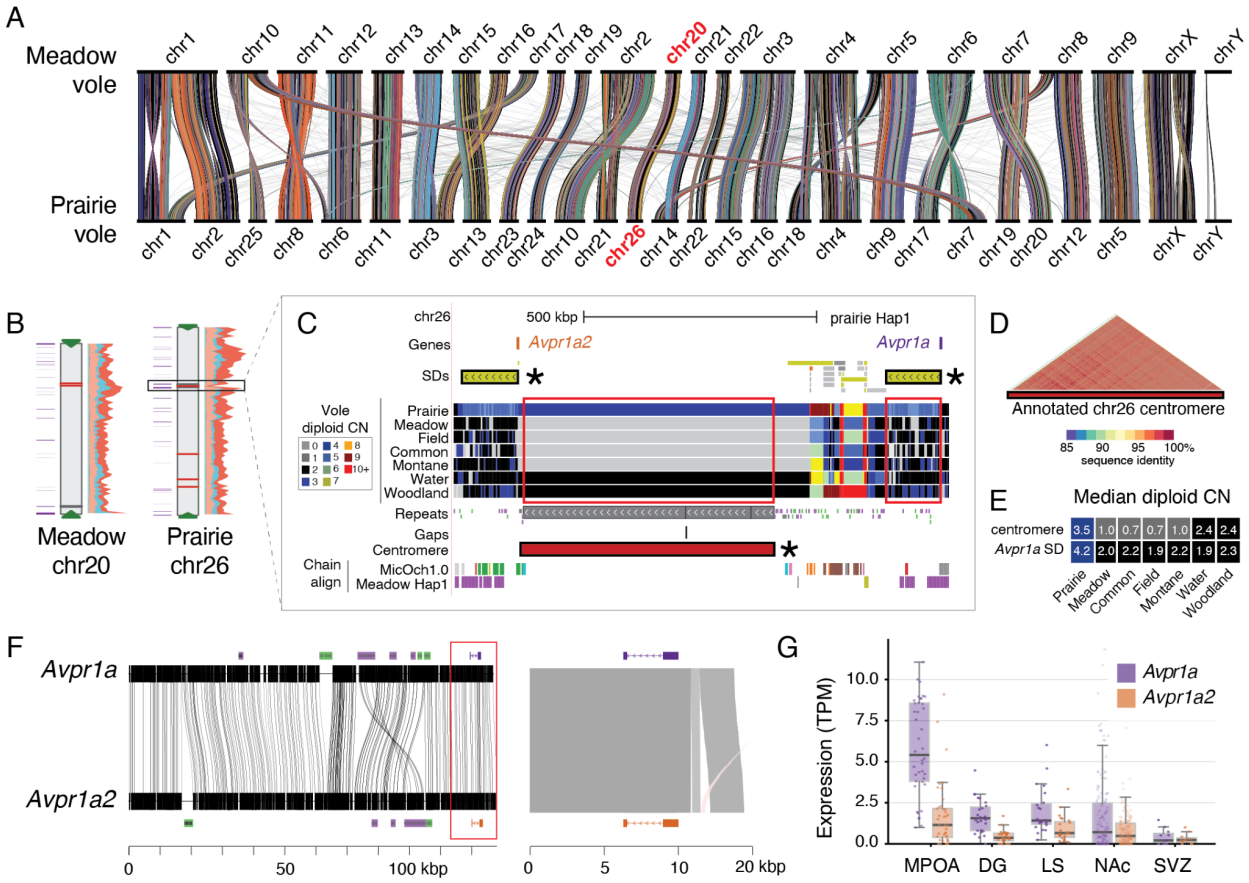

Figure 3. Prairie versus meadow vole cross-species comparison (**A**) Assembly comparisons of synteny between prairie and meadow voles depicted as ribbon plots [33] colored by ancestral linkage groups (ALGs [34]). (**B**) Repeat and segmental duplications (SDs) of *Avpr1a* flanking the prairie vole chr26 annotated centromeric region compared with meadow vole chr20. Color scheme described in Figure 2B. (**C**) A UCSC Genome Browser screenshot of the prairie vole Hap1 *Avpr1a* locus (chr26:13.0 Mbp–14.1 Mbp), with SDs of the paralogs and intervening centromeric repeat element highlighted with an asterisk. Windowed copy numbers (CN) of *Microtus* vole species are depicted as colors. Alignment chains for the current prairie vole reference (MicOch1.0) and the meadow vole Hap1 assembly (this study) are depicted at the bottom. (**D**) A dotplot of the annotated centromere with color representing sequence identities. (**E**) Both the annotated centromere and *Avpr1a* are at increased diploid CN uniquely in prairie vole. The two genotyped regions are highlighted as red boxes and colors match the legend in (C). **(F)** Hap1 genomic alignments [35] of prairie vole *Avpr1a* (purple) and *Avpr1a2* (red) including repeat annotations. (**G**) Transcriptomic analysis showing expression of both *Avrp1a* paralogs in medial preoptic area (MPOA), dentate gyrus (DG), lateral septum (LS), nucleus accumbens (NAc), and subventricular zone (SVZ). No expression was detected for either paralog in amygdala, hypothalamus, and ventral pallidum (not shown). TPM, transcripts per million reads.

We next computed diploid copy number (CN) genome-wide from Illumina read depth [36] using data generated in this study from prairie vole, meadow vole, and woodland vole (*M. pinetorum*, also known as pine vole, also exhibiting monogamous pair bonding [37,38]) alongside four publicly available *Microtus* genomes [39,40] representing species with more promiscuous behaviors [41–46]. While CN-estimates tend to be unreliable across repetitive loci due to repeat masking, prairie vole showed elevated CN across the chr26 centromere annotation relative to all other voles, suggesting this region is prairie-vole specific (Figure 3E). Also querying the CN of the 118-kbp SD spanning the complete *Avpr1a* gene and adjacent regions

confirmed this duplication to be prairie-vole specific with diploid CN of 4 compared to CN of 2 for all other tested vole genomes.

Both curated prairie vole references show 97% nucleotide identity between *Avpr1a* paralogs, with *Avpr1a2* harboring frameshift variants consistent with a truncated protein (218 aa vs. 420 aa full-length). This result was supported by both Hap1 and Hap2 assemblies (Figure S13 and Table S16) and is consistent with prior reports [19,20]. If translated, the transcript would produce a truncated protein (218 amino acids (aa) versus 420 aa full length), sharing the first identical 199 aa with Avpr1a followed by novel 19 aa, that includes the extracellular vasopressin-binding domain and the first four of seven transmembrane domains but lacks the intracellular G-protein binding region [47]. Comparing transcript abundances using published RNA-seq data [48–53] reveals consistent expression of both paralogs across five of eight tested brain regions, with highest expression in the medial preoptic area, albeit reduced for *Avpr1a2* (~1–4× relative to *Avpr1a*) (Figures 3G and S14); this could be influenced by a 600-bp indel ~1.8 kbp upstream of *Avpr1a* paralogs (Figure 3F) and altered chromatin interactions outside of the duplicated region [54], with the SD breakpoint only 4 kbp upstream. Even if *Avpr1a2* ultimately proves to be a pseudogene, the corrected assembly enables expression and epigenomic analysis of the ancestral full-length *Avpr1a* and its *cis*-regulatory landscape, previously absent from the prairie vole reference genome entirely.

DISCUSSION

Assemblies enable robust and complete genome comparisons of variation contributing to divergence, diversity, and disease. Most recent efforts to build gigabase-sized chromosome-scale genomes require multiple technologies [55], typically comprising highly accurate HiFi long reads (~15-20 kbp at 30–60× coverage), ultralong nanopore reads (>100 kbp at ~30× coverage), and short-read long-range information to phase and scaffold at chromosome scale [56]; even higher coverage is necessary to achieve complete diploid T2T genomes [57]. This “ideal” recipe is inaccessible to many researchers, requiring large amounts of starting material, multiple libraries, and three sequencing platforms. Here, we generated ~100 Gbp of HiFi and CiFi data from a single library and sequencing, producing contiguous (scaffold N50 90–125 Mbp), high-quality diploid assemblies (QV>60). The experimental approach is conceivably scalable to tens of thousands of cells and nanograms of DNA/chromatin [7]. Based on CiFi downsampling (Figure S5), a 2× CiFi and 30× HiFi coverage ratio is theoretically sufficient to assemble diploid genomes as large as ~3.1 Gbp (human sized) on a single SMRT Cell. Compared with the “ideal” multi-platform recipe, this approach reduces costs by at least threefold while still producing over half T2T-scale chromosomes.

Beyond introducing a simplified assembly approach, we provide a long-awaited resource expanding the genomic toolkit for prairie voles, an emerging model organism for studying complex social behaviors relevant to humans [58]. This includes high-quality annotated assemblies for both prairie and meadow voles, repeat and SD annotations, CN maps across additional vole species, chain alignments facilitating genome liftover, and named gene orthologs enabling transcriptomic comparisons, all publicly accessible through a UCSC Genome Browser hub (see Data Availability). These resources complement ongoing neurobiological studies examining how social relationships are encoded in the brain.

Neuropeptide hormone pathways represent an important substrate through which neural circuits mediating innate behaviors can rapidly evolve and diversify [12,13]. Hence, divergence of the genes encoding neuropeptide receptors or their regulatory regions may underlie species differences in behavior. The highly contiguous assemblies generated here enabled discovery of a complex >350-kbp satellite repeat element with centromere annotation and confirmed the presence of a >100 kbp SD impacting *Avpr1a* (Figure 3). Neither feature is present in any of the four meadow vole haplotypes examined (this study and mMicPen1), nor was elevated CN detected at these regions in the five other vole species tested, including woodland vole, which also exhibits pair bonding. While this specificity to prairie vole argues against these variants as cross-species drivers of pair bonding, they remain compelling candidates for prairie-vole-specific function. Notably, if epigenetic signatures confirm the satellite repeat as a true centromere, its position ~350 kbp downstream of *Avpr1a* would place the gene within the pericentromeric region. Centromerization has been shown to result in increased H3K27me3-mediated repression, altered chromatin accessibility, and elevated genomic instability [59]—consequences that may alter vasopressin signaling dynamics and could have contributed to the birth of the prairie-vole-specific *Avpr1a2* paralog. This is a particularly compelling finding given that prairie voles exhibit elevated *Avpr1a* expression in the ventral pallidum while meadow voles show higher expression in the lateral septum [19]. These assemblies will serve as an important resource for continued investigation of *Avpr1a* paralog functions and divergent brain expression patterns.

Our assessment of candidate behavioral genes (Tables S13 and S14) identified no obvious protein-coding differences between prairie and meadow voles, suggesting that regulatory divergence, rather than coding sequence change, plays an outsized role in behavioral differences between these species. These near-complete genomes will enable comparative, genome-wide analyses of chromatin structure and gene regulation, providing a framework for examining regulatory mechanisms underlying bond formation and social memory. Ultimately, connecting vole genomics to human genetic studies offers a path toward understanding how social behavior is encoded in the brain, and how its disruption contributes to neuropsychiatric risk.

Limitations of the study

A notable limitation is that current assembly and scaffolding tools, including hifiasm and YAHS, do not leverage the multi-contact information inherent to CiFi concatemers, instead requiring reduction into Hi-C-like pairs. Development of algorithms designed to leverage these higher-order chromatin contacts will likely yield further improvements in phasing, with CiFi showing an >8-fold greater ability to infer haplotypes versus Hi-C when considering multicontacts [7], as well as scaffolding, particularly in complex genomic regions. Additionally, use of a single restriction enzyme (HindIII) introduces the possibility of sequence-biased coverage (also inherent in short-read Hi-C); adoption of alternative fragmentation approaches such as Omni-C could mitigate this in future implementations. Gene annotations might be further refined by incorporating long-read isoform sequencing, enabling a fully integrated workflow for assembly, scaffolding, and annotation from a single sequencing technology. Finally, although two haplotypes were resolved per species through diploid assembly, single-individual sampling limits the ability to distinguish fixed from polymorphic differences. Broader sampling within and across *Microtus* species using a phylogenetic framework will be necessary to determine whether identified variants segregate with behavioral phenotypes, an effort complicated by the uncertain evolutionary relationships within this large clade of >60 species that has undergone rapid radiation over the last ~2 million years [60].

In summary, we present a combined HiFi and CiFi sequencing strategy and simplified bioinformatic workflow enabling accurate, contiguous, chromosome-scale genome assembly from a single sequencing platform and experiment. Applying this approach to prairie and meadow voles, we generate high-quality diploid assemblies that reveal genome-wide differences between these behaviorally divergent species, with particular focus on loci implicated in social behavior. Ultimately, the simplicity, affordability, and minimal sample requirements of the combined HiFi–CiFi approach position it as a broadly accessible tool for comparative genomics, with the future potential to make chromosome-scale assembly tractable for rare, difficult-to-sample, and non-model organisms alike.

STAR METHODS

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Vole procedures

Adult prairie vole (*M. ochrogaster*) and meadow vole (*M. pennsylvanicus*) individuals were maintained under protocols approved by the Institutional Animal Care and Use Committee (IACUC) at the University of California, San Francisco and Cold Spring Harbor Laboratories. Animals were euthanized according to institutional guidelines, and tissues were rapidly dissected, flash-frozen in liquid nitrogen, and stored at −80 °C until DNA or RNA extraction.

METHOD DETAILS

HiFi and CiFi library preparations and sequencing of prairie and meadow vole liver samples

For whole-genome HiFi sequencing, high molecular weight (HMW) genomic DNA was extracted from frozen liver tissue using the Monarch HMW DNA Extraction Kit for Tissue (New England Biolabs) according to the manufacturer's protocol. DNA quality and fragment size distribution were assessed by Qubit fluorometry and Femto Pulse analysis prior to downstream HiFi library preparation.
For CiFi library preparation, frozen liver tissue (~100 mg) was pulverized under liquid nitrogen and crosslinked, followed by processing according to CiFi protocol Part 1 (HindIII restriction digestion and proximity ligation) [7]. Crosslinks were reversed, and proximity-ligated 3C DNA was purified by phenol–chloroform extraction prior to downstream size selection and amplification.

A starting amount of 2.2 µg of HMW DNA was sheared on a Hamilton NGS STAR MOA system and size-selected on the Pippin HT with 10 kbp cut off. The pre-PCR CiFi DNA (1.5 µg) was size-selected on the Pippin HT with 6.5 kbp cutoff and PCR amplified (50 ng input, 8 cycles) using the Ampli-Fi protocol (PacBio, 103-648-000). The post-PCR CiFi DNA was size-selected on the Pippin HT with 5.5 kbp cutoff. They were subsequently mixed at a molar ratio of 90% sheared DNA for HiFi and 10% CiFi DNA (translating to 650 ng and 35 ng, respectively, accounting for the size differences), followed by standard library preparation using SMRTbell prep kit 3.0 (PacBio, 102-182-700). The SMRTbell library was cleaned up with 1X SMRTbell cleanup beads. One SMRT Cell per species was run on the Revio system at 250 pM on-plate loading concentration with SPRQ chemistry and 30-hours acquisition.

HiFi and CiFi data from the sequencing run were segregated and adapters were trimmed with lima v2.14.0 (https://github.com/PacificBiosciences/barcoding) using a three-step process. First, the CiFi reads

were identified by their unique dual index adapters and separated from HiFi reads using very relaxed lima settings (`--ccs --min-passes 0 --min-end-score 0 --min-score 5 --min-ref-span 0.2 --min-score-lead 0`) to ensure that no CiFi reads remained in the HiFi data. Adapters were then trimmed from each dataset using the recommended settings: `--hifi-preset SYMMETRIC` for the HiFi reads and `--hifi-preset ASYMMETRIC --neighbors` for the CiFi reads. PCR duplicates were then removed and duplication rates assessed for the CiFi reads using pbmarkdup v1.1.0 (https://github.com/PacificBiosciences/pbmarkdup).

Illumina sequencing vole genomic samples

Chromatin isolation and Hi-C library preparation was performed by Phase Genomics (WA, USA), using 100 mg flash-frozen liver tissue from the same individual male prairie and meadow voles as were used for HiFi and CiFi library preparation. The Proximo Hi-C protocol (Phase Genomics, WA, USA) [61] was used to prepare the proximity ligation library and process it into an Illumina-compatible sequencing library. Hi-C libraries were sequenced with 150-bp paired-end reads on an Illumina NextSeq500 at the CSHL Next Generation Sequencing Core Facility.

gDNA was extracted from dried museum pelt tissue of the woodland vole (*M. pinetorum*) using the DNeasy Blood & Tissue Kit (Qiagen) following manufacturer instructions with minor modifications to reduce PCR inhibitors. Extracted DNA was treated with RNase A and further purified by ethanol precipitation. Sequencing libraries were prepared for short-read sequencing using prepared genomic DNA for prairie vole, meadow vole, and woodland vole. Prepared libraries were sequenced (2×150 bp) by Novogene on a NovaSeq 6000 platform to produce approximately 60 Gbp of raw data (~30× genome coverage) for each species.

Genome assembly and scaffolding

HiFi reads were extracted from unaligned BAM files using samtools v1.21 [62] (`samtools fasta`). CiFi reads were converted from BAM to FASTQ (`samtools collate -O -u | samtools fastq`) and then processed into Hi-C-like paired-end reads using the cifi-toolkit (https://github.com/mr-eyes/cifi-toolkit) that performs in silico HindIII digestion on each CiFi concatemer read, extracting the outermost restriction fragments as R1 and R2 paired-end reads. Phased diploid assembly was performed with hifiasm v0.19.8 [4] using the `--dual-scaf` mode, which integrated either CiFi (input as paired-end reads) or Hi-C contact information during graph resolution to produce haplotype-resolved primary contig graphs for each haplotype (Hap1 and Hap2). HiFi FASTA and derived R1/R2 FASTQ files were provided as input (`--h1`, `--h2`). GFA contig graphs were converted to FASTA using gfatools v0.5 (https://github.com/lh3/gfatools).

CiFi and Hi-C contact maps were generated by aligning the full CiFi BAM to each haplotype assembly using the epi2me-labs/wf-pore-c Nextflow pipeline v1.3.0 (https://github.com/epi2me-labs/wf-pore-c) with minimap2 in `-ax map-hifi` mode, HindIII as the restriction enzyme, and `--paired_end` enabled to produce BED contact files. Contig scaffolding was performed with YAHS v1.2a.2 [21] using the porec BED contact file as input, with contig error correction disabled (`--no-contig-ec`).

To evaluate the minimum CiFi sequencing depth required for chromosome-scale scaffolding, the CiFi BAM was subsampled at 1%, 10%, 15%, 20%, 25%, 40%, 60%, 80%, and 100% of the original read depth using `samtools view -s` with a fixed random seed. At each fraction, the full assembly and

scaffolding pipeline described above was executed independently, and scaffold N50 and auN were compared across titration points.

Genome assembly manual curation

To facilitate manual assembly curation for each genome of the Prairie and the Meadow voles the two scaffolded haplotypes were combined together and the corresponding CiFi datasets were mapped onto the assemblies using wf-pore-c pipeline with parameters `--paired_end --cutter HindIII --paired_end_minimum_distance 100 --paired_end_maximum_distance 200`. Hi-C maps were further generated for the combined haplotypes of each genome with PretextMap using the mock paired-end BAM files as input and retaining non-uniquely mapping reads in the contact maps with --mapq 0. Supplementary analysis were also embedded in the Pretext file using the Tree of Life curationpretext pipeline - telomeres with standard vertebrate motif TTAGGG, N base scaffold gaps and mapped PacBio long-read coverage. An AGP of the corrected contact maps were exported from PretextView and curated haplotype assemblies generated using pretext-to-asm.

PretextView https://github.com/sanger-tol/PretextView
PretextMap https://github.com/sanger-tol/PretextMap
Curationpretext https://github.com/sanger-tol/curationpretext
Pretext-to-asm https://github.com/sanger-tol/agp-tpf-utils

Telomere-to-telomere (T2T) completeness and gap assessment

To evaluate the contiguity and completeness of our assemblies, we assessed telomere presence and sequence gaps across all chromosome-scale scaffolds for both haplotypes of each species. Gaps (runs of N bases) were identified using seqtk gap (https://github.com/lh3/seqtk). Telomeric repeat sequences were detected using tidk v0.2 (Telomere Identification Toolkit; [63]), which scans for the canonical vertebrate telomere motif (TTAGGG) in sliding windows across each scaffold. We searched for TTAGGG repeats in 10 kbp windows and considered a telomere present at a chromosome terminus when the terminal window contained at least 10 repeat units (forward and reverse strands combined). Chromosomes were classified as T2T (telomeric repeats at both ends), partial (one end only), or incomplete (neither end). A chromosome was considered fully T2T-resolved only when both telomere ends were detected and no sequence gaps were present.

Centromere annotation

Centromeric regions were identified on each hap1 assembly using centroAnno v1.0.2 [32] in assembly annotation mode (`-x anno-asm`). Each chromosome was processed individually; centroAnno decomposed tandem satellite repeats into monomer units and reported their genomic coordinates. Contiguous monomer annotations separated by fewer than 10 kbp were merged into repeat regions, and the largest repeat region per chromosome was designated as the putative centromere. Centromeres were detected on all chromosomes of both species.

Segmental duplication and repeat annotations

Species-specific *de novo* repeat libraries were constructed for each vole species using RepeatModeler v2.0.7 [64] with default parameters. A RepeatModeler database was built from each hap1 primary assembly, and *de novo* repeat family consensus sequences were identified. The resulting species-specific libraries

were combined with the Dfam database and used as input for RepeatMasker v4.2.2 [65] with the RMBLAST search engine. RepeatMasker was executed via the Dfam TE Tools container (https://hub.docker.com/r/dfam/tetools; RepeatModeler v2.0.7, RepeatMasker v4.2.1, RMBLAST v2.14.1+). Segmental duplications were identified using BISER v1.4 [28] with default parameters. BISER was run on the soft-masked hap1 assemblies for each species, and output was generated in BEDPE format.

Transcriptome analysis of vole tissues

Total RNA was prepared from flash frozen prairie and meadow vole tissues: liver, brain, and testes from adults and e11.5 (prairie) or e13.5 (meadow) embryos. 50–75mg of tissue per sample was homogenized in 500 µl Trizol (Life Technologies) on ice using a Kimble Kontes Disposable Pellet Pestle (VWR). An additional 500 µl Trizol was added followed by further homogenization using an 18-gauge needle on a 1ml syringe before proceeding to RNA extraction according to the Trizol protocol. The subsequent RNA was DNase treated with a TURBO DNA-free kit (Life Technologies). 150 ng of total RNA was used as input for library preparation with Encore Complete RNA-seq kits (NuGen), using ten cycles of amplification. Multiplexed libraries were sequenced with 76-bp paired-end reads on the Illumina NextSeq 500 at the CSHL Next Generation Sequencing Core Facility.

Gene annotation: Illumina RNA-seq data from the four samples per species were used in the NCBI Eukaryotic Genome Annotation Pipeline (EGAPx) v0.4.1-alpha [66] to perform gene annotation for bothHap1 assemblies of both species.

Gene expression analysis: Transcript quantification was performed using Salmon v1.10.3 [67] in mapping-based mode. For each species, *Avpr1a* transcript-level expression was first quantified across four in-house paired-end Illumina RNA-seq libraries (brain, liver, testes, and embryo) using a targeted Salmon index containing the *Avpr1a* coding sequences with the hap1 genome assembly as a decoy. For the prairie vole, the index included both *Avpr1a* paralogs to enable paralog-specific quantification. To contextualize *Avpr1a* expression within the broader transcriptome, we performed transcriptome-wide Salmon quantification for the prairie vole. The full EGAPx-annotated transcriptome was extracted from the prairie hap1 assembly using gffread, and a decoy-aware Salmon index was built by concatenating the transcriptome with the genome assembly. We quantified 453 publicly available prairie vole brain RNA-seq libraries from previously published studies [48–53] spanning eight brain regions—amygdala (AMY), dentate gyrus (DG), hypothalamus (HT), lateral septum (LS), medial preoptic area (MPOA), nucleus accumbens (NAc), subventricular zone (SVZ), and ventral pallidum (VP)—under BioProject accessions PRJNA428754, PRJNA682808, PRJNA631040, PRJNA786347, PRJNA887096, PRJNA792575, PRJNA1005323, and PRJEB89367. After excluding 147 technical replicates, 306 samples were retained for analysis. Salmon was run with `--validateMappings`, `--gcBias`, and `--seqBias` correction, with library type automatically inferred. Differential expression analysis across brain regions was performed using pyDESeq2 [68].

Copy-number analysis

Copy number (CN) was estimated using the FastCN pipeline [36] with mrsFAST v3.4.2. All species were mapped to the prairie vole hap1 assembly as a common reference to enable direct cross-species comparison at the same genomic coordinates. A four-layer masked reference was constructed from the

prairie vole hap1 assembly. Repetitive elements were masked using the RepeatMasker annotations described above, tandem repeats were identified with Tandem Repeats Finder [69] run per chromosome, and low-complexity regions were masked with WindowMasker/DUST [70]. The masked genome was then indexed with mrsFAST and all 50-mers were extracted and searched back against the genome; positions where a 50-mer aligned 20 or more times were additionally masked (K50 masking). Gaps in the final masked reference were extended by 36 bp on each side to account for the read-length shadow effect. Control regions expected to be diploid (CN=2) were defined by excluding segmental duplications (identified by BISER), centromeric regions, and target gene loci from the genome-wide window set.

For each species, the first 36 bp were extracted from each mate of the paired-end whole-genome sequencing reads and mapped to the masked reference using mrsFAST with up to two mismatches allowed. GC-corrected read depth was computed in 1 kb windows across the genome. Copy number was calculated as CN = 2 x (window depth / autosomal control mean). To prevent inflated copy number values from masked regions with zero depth falling within control windows, we excluded zero-depth windows from the control mean calculation. Control regions were further refined by retaining only windows where all samples showed consistent CN near 2.0 (coefficient of variation < 0.2), and a final normalization ensured the median CN at refined control windows equaled 2.0.

Whole-genome sequencing reads from seven *Microtus* species —prairie vole (*M. ochrogaster*, this study), meadow vole (*M. pennsylvanicus*, this study), woodland vole (pine vole, *M. pinetorum*, this study), montane vole (*M. montanus*; SRR12966109), common vole (*M. arvalis*; ERR3427942), field vole (*M. agrestis*; SRR2167807), North American water vole (*M. richardsoni*; SRR12963053)— were mapped to the prairie vole reference. Each species was then normalized independently by scaling CN values so the median at refined diploid autosomal control regions equals 2.0, correcting for sequencing depth differences between species.Per-gene CN was calculated as the mean CN across all non-zero 1 kb windows overlapping each target locus. CN was evaluated at 15 target loci spanning the vasopressin/oxytocin system (*Avpr1a*, *Avpr1b*, *Avp*, *Oxt*, and *Oxtr*), dopamine system (*Drd2*), estrogen system (*Esr1*, *Esr2*, and *Cyp19a1*), stress system (*Crh*, *Crhr1*, and *Crhr2*), neural development genes (*Chd8* and *Shank3*), and the androgen receptor (*Ar*).

Dot plots were generated using ODP v0.3.3 [33], circular genome visualizations using pyCirclize v1.9.1(https://github.com/moshi4/pyCirclize), and linear karyotype ideograms using matplotlib v3.10.8. Assembly accuracy was evaluated using Yak [4] by concatenating both haplotypes to calculate a combined genome-wide adjusted QV against HiFi reads, while k-mer completeness was assessed by providing both haplotypes in diploid mode to Merqury [22].

Comparative assembly and paralog analyses

Ortholog groups were identified using Orthofinder (v2.5.5) [71] with default parameters, using protein sequences from both species. Orthogroups were classified as one-to-one, expanded in one of the species, multi-copy in both, or species-specific. For the one-to-one ortholog pairs, we aligned CDS using parasail (v2.6.2) [72], then variants were called from the pairwise alignments. SNPs and indels were identified by parsing the alignment CIGAR string. Variants were classified by comparing reference and alternate codons using the standard genetic code: synonymous if both codons encode the same amino acid, missense if different, and nonsense if the alternate codon is a stop codon. For genes with multiple

annotated isoforms, the longest one was selected as a representative. Ka/Ks (dN/dS) were calculated using the Nei-Gojobori method implemented in BioPython v1.83 [73]. CDS length ratio <0.8 were excluded as unreliable. Structural differences between ortholog CDS pairs were detected using BLASTN megablast alignments (BLAST+ v2.16.0) [74]. Frameshifts were defined as insertions/deletions events in the CDS alignments with length not divisible by 3. Gene-level inversions were identified when BLASTN returned alignments on opposite strands. Exon structure changes were detected by comparing the number, boundaries, and sizes of exons between ortholog pairs in the EGAPx annotations. Truncations were flagged when the CDS length ratio between species was < 0.8. *Avpr1a* coordinates for Hap2 assemblies were obtained by aligning Hap1 regions ± 50 kbp to Hap2 with minimap2 v2.30 with -x asm5, followed by annotation transfer with miniprot v0.18. CDS and protein MSAs (6 sequences: 2 species × 2 haplotypes, plus *Avpr1a2* × 2 haplotypes) were generated with MACSE v2.07 [75], which handles the +1 G frameshift in *Avpr1a2* without breaking the reading frame. Gene DNA MSAs were generated with MAFFT v7.525 (L-INS-i) [76]. Pairwise SNPs were classified as synonymous, nonsynonymous, or intronic using bcftools csq v1.22 [77].

Pairwise comparisons and visualization of SD sequence containing *Avpr1a* (prairie vole Hap1 chr26:13970803-14087830) and Avpr1a2 (prairie vole Hap1 chr26:13058418-13176477) was performed using Micropeats [35]. The annotated centromeric sequence (prairie vole Hap1 chr26:13182318-13728901) was queried against both Hap1 and Hap2 of the prairie and meadow vole assemblies (this study and mMicPen1) using minimap2 (v2.26) [78] selecting for matches 90% or higher. Dot plots generated using nf-core/pairgenomealign [79].

Scaffold orientation

Prairie vole: To ensure compatibility with existing prairie vole genomic resources and maintain consistent chromosome orientation conventions, scaffold orientation in the *de novo* haplotype-resolved assemblies was standardized relative to the MicOch1.0 reference genome (GCF_000317375.1). The assembly retains its original chromosome numbering (SUPER_1 through SUPER_26, SUPER_X, SUPER_Y, renamed to chr1-chr26, chrX, chrY). The MicOch1.0 alignment was used exclusively to determine whether each scaffold required reverse-complementation, without reassigning chromosome identities.

Whole-genome alignments were conducted between each query haplotype assembly and MicOch1.0 using minimap2 v2.28 [78] with the asm5 preset. Alignments were generated in both directions (query-to-reference and reference-to-query) to facilitate orientation determination and to identify previously unplaced MicOch1.0 scaffolds incorporated into chromosome-scale scaffolds in the new assembly. Alignments shorter than 10 kb were excluded to minimize spurious matches in repetitive regions.

For each query scaffold, orientation was determined by quantifying the total aligned bases on the forward (+) and reverse (-) strands relative to the best-matching MicOch1.0 chromosome or linkage group. Scaffolds were designated for reverse-complementation if the majority of aligned bases mapped to the reverse strand; otherwise, the original orientation was retained. Confidence was classified as high (>95% of aligned bases on the dominant strand), medium (80-95%), or low/mixed (<80%, indicating mixed strand orientations typical of alignments involving fragmented reference regions). As MicOch1.0 was derived from a female individual, the Y chromosome scaffold (chrY) could not be oriented by alignment

and was therefore retained in its original orientation. Alignment-based orientation calls were independently cross-validated using a radiation hybrid (RH) linkage map for the prairie vole [24]. For each linkage group, RH marker sequences were mapped to the *de novo* assembly using BLASTn, and the Spearman rank correlation (ρ) between marker centimorgan position and physical position on the scaffold was calculated. A positive ρ indicates concordance with the primary scaffold orientation, whereas a negative ρ indicates the scaffold is in the reverse orientation relative to the genetic map. For scaffolds with low alignment-based confidence (<80% dominant strand) and a strong linkage map orientation signal ($|\rho| \geq 0.70$, ≥5 mapped markers), the linkage map call was prioritized as the primary evidence. Cross-validation results are presented in Table S5.

Corrected assemblies were produced by reverse-complementing the designated scaffolds and renaming all scaffolds according to a standardized nomenclature (chr1-chr26, chrX, chrY for chromosome-scale scaffolds). A UCSC liftOver chain file was generated to facilitate coordinate conversion of annotation files from the original to the corrected assembly. The complete set of orientation decisions, alignment statistics, and confidence classifications for all 97 scaffolds is provided in Table S10.

The prairie vole Hap1 assembly consists of 97 scaffolds, including 28 chromosome-scale scaffolds (chr1-chr26, chrX, chrY) and 69 minor unplaced scaffolds. All 26 autosomal scaffolds and chrX were assigned to corresponding MicOch1.0 chromosomes or linkage groups, with alignment coverage ranging from 32.6% to 78.4%. As expected, chrY showed no alignment to the female-derived MicOch1.0 reference (Table S10). Alignment-based orientation analysis revealed that 15 of 28 chromosome-scale scaffolds were concordant with the MicOch1.0 convention (forward orientation), while 13 required reverse complementation. Alignment confidence was classified as high for 7 scaffolds, medium for 9, low/mixed for 11, and not assessable for 1 (chrY). The high frequency of low/mixed confidence values reflects the substantial structural divergence between a chromosome-scale long-read assembly and a fragmented short-read reference containing 6,336 sequences and approximately 631 Mbp of unplaced sequence.

Cross-validation with the RH linkage map was feasible for 26 of 28 chromosome-scale scaffolds (chrY and chr22 lacked linkage markers), resulting in 29 scaffold-linkage group comparisons. Of these, 27 out of 29 (93.1%) demonstrated concordant orientation calls between the two independent methods (Table S5). Two comparisons were discordant: (i) chr10 was concordant with its primary linkage group (LG14; $\rho = +0.41$, 11 markers) but discordant with a secondary linkage group (LGLG8; $\rho = -0.98$, 5 markers), indicating that this scaffold incorporates sequences from multiple MicOch1.0 linkage groups in opposite orientations; and (ii) chrX exhibited a weak linkage signal ($\rho = +0.60$, 11 markers) discordant with the alignment call, consistent with limited recombination on the sex chromosome reducing the resolution of marker-order correlations. For chr8, where alignment confidence was low/mixed (63.7% forward strand), the RH linkage map provided strong evidence for reverse-complementation ($\rho = -0.93$, 13 markers), which was used as the primary evidence source.

In the reference-to-query analysis, 238 previously unplaced MicOch1.0 scaffolds totaling 520.8 Mbp were mapped within chromosome-scale scaffolds of the new assembly, with alignment coverage ranging from 50.0% to 284.3% (median 78.8%) (Table S6). These resolved scaffolds were distributed across all 28 chromosomes, with the largest contributions to chrX (55.8 Mbp from 27 scaffolds), chr12 (78.8 Mb from 10 scaffolds), and chr6 (46.0 Mb from 16 scaffolds). Among the 238 resolved scaffolds, 103

exceeded 1 Mb in length, and 149 exceeded 100 kb, indicating that the new assembly anchors a substantial amount of previously unplaced sequence within a chromosomal context. An additional 69 minor scaffolds (hap1_scaffold_27 through hap1_scaffold_95; collectively 29.1 Mb) were evaluated. Of these, 28 could be oriented by alignment to MicOch1.0 (16 reverse-complemented, 12 retained), while 41 showed no meaningful alignment and were retained in their original orientation. The complete scaffold-to-chromosome assignments, orientation decisions, and validation results for all 97 scaffolds are provided in Table S10.

Meadow Vole: Scaffold orientation in the meadow vole Hap1 assembly was standardized using the same methodology described above, with the orientation-corrected prairie vole Hap1 assembly as the reference. This ensures a consistent orientation convention across both species in this study. The minimap2 asm10 preset was used for cross-species alignment. The meadow vole Hap1 assembly comprises 105 scaffolds: 24 chromosome-scale (chr1-chr22, chrX, chrY; 2.34 Gb) and 81 minor unplaced (19.7 Mb). Alignment coverage for the 23 evaluable chromosome-scale scaffolds ranged from 43.4% to 87.9%; chrY showed no cross-species alignment and was retained in its original orientation. Of the 24 chromosome-scale scaffolds, 12 were concordant and 12 required reverse-complementation, with confidence classified as high (13), medium (4), low/mixed (6), or not assessable (1, chrY). The 6 low/mixed scaffolds reflect cross-species rearrangements where a single meadow vole chromosome aligns to multiple prairie vole chromosomes in mixed orientations (meadow 2n=46 vs. prairie 2n=54). Post-correction validation showed 13 of 24 chromosome-scale scaffolds with ≥80% forward-strand alignment; the remainder are attributable to chromosomal fusions. Among the 81 minor scaffolds, 53 were oriented by alignment (29 reverse-complemented, 24 retained), and 28 showed no cross-species alignment (Table S11).

DATA AVAILABILITY

All data generated as part of this project can be found at the European Nucleotide Archive and NCBI GenBank Project accession PRJEB108798. Previously published datasets have GenBank accessions listed below.

Prairie and meadow vole reference genomes:

- mMicPen1 reference: GCF_037038515.1
- MicOch1.0 reference: GCF_000317375.1

RNA-seq of prairie vole brain regions: PRJNA428754, PRJNA682808, PRJNA631040, PRJNA786347, PRJNA887096, PRJNA792575, PRJNA1005323, PRJEB89367

Illumina WGS sequence data of *Microtus* species: SRR12966109, ERR3427942, SRR2167807, SRR12963053

Temporary hubs for UCSC Genome browser resources

- Prairie vole: https://genome.ucsc.edu/s/mabuelanin/prairie_vole_cifi_hifi_hap1
- Meadow vole: https://genome.ucsc.edu/s/mabuelanin/meadow_vole_cifi_hifi_hap1

CODE AVAILABILITY

Methods used for genome curation are available at https://github.com/sanger-tol/. A Zenodo doi will be generated for code related to this manuscript at the time of publication. All source code and workflows can be accessed through https://github.com/mydennislab/2026-voles-assembly.

ACKNOWLEDGEMENTS

We thank Dr. Aaron Wenger and Jacob Brandenburg for coordination support in generating and analyzing the HiFi-CiFi datasets. Also thanks to Drs. Michael Schatz and Daniela Soto for assistance with earlier iterations of vole genome assemblies, and Michael Sherman for help with animal husbandry. This work was supported, in part, by the U.S. National Science Foundation (CAREER 2145885 to M.Y.D.; CAREER IOS-2045348 to Z.R.D.) and the National Institutes of Health (NIH) grants from the National Institute of Mental Health (R01MH132818 to M.Y.D. and DP2MH119427 to Z.R.D.). K.K., J.W., and K.H. are supported by Wellcome through the 220540/Z/20/A award that supports the Wellcome Sanger Institute. Voles and tissue samples were supported by NIMH (R01MH123513) and NSF (1556974) grants to D.S.M. CSHL voles studies were supported by an institutional grant to J.T. Woodland vole analysis was supported by a UC-Davis Interdisciplinary Catalyst Award to K.L.B. and M.Y.D. Illumina sequencing of RNA-seq and Hi-C libraries was performed at the Cold Spring Harbor Next Generation Sequencing Shared Resource, which is supported by NIH Cancer Center Support Grant 5P30CA045508 Some images were created using BioRender.

AUTHOR CONTRIBUTIONS

G.K., C.L., M.V.W., N.G.S., Z.R.D., K.L.B., and J.T. generated biospecimens and/or data for the project. M.A., J.L., K.K., J.W. K.H., and M.Y.D. performed bioinformatic and genomic analysis. K.L.B., D.M. and J.T. provided biomaterials and provided vole expertise. J.K., D.M., J.T., and M.Y.D. devised the project. M.A., J.K., D.M., and J.T., and M.Y.D. wrote the manuscript and all authors edited and approved the manuscript.

COMPETING INTERESTS

J.L. & C.L. are employees and shareholders and J.K. is a consultant and shareholder of Pacific Biosciences, a company developing single-molecule sequencing technologies.

REFERENCES

1. Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature *592*, 737–746.

2. Dennis, M.Y., and Eichler, E.E. (2016). Human adaptation and evolution by segmental duplication. Curr. Opin. Genet. Dev. *41*, 44–52.

3. Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. *37*,

1155–1162.

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

5. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

6. Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. *31*, 1119–1125.

7. McGinty, S.P., Kaya, G., Sim, S.B., Makunin, A., Corpuz, R.L., Quail, M.A., Abuelanin, M., Lawniczak, M.K.N., Geib, S.M., Korlach, J., et al. (2025). CiFi: accurate long-read chromosome conformation capture with low-input requirements. Nat. Commun. *17*, 215.

8. Kleiman, D.G. (1977). Monogamy in mammals. Q Rev Biol *52*, 39–69.

9. Lukas, D., and Clutton-Brock, T.H. (2013). The evolution of social monogamy in mammals. Science *341*, 526–530.

10. Bales, K.L., Ardekani, C.S., Baxter, A., Karaskiewicz, C.L., Kuske, J.X., Lau, A.R., Savidge, L.E., Sayler, K.R., and Witczak, L.R. (2021). What is a pair bond? Horm. Behav. *136*, 105062.

11. Steppan, S.J., and Schenk, J.J. (2017). Muroid rodent phylogenetics: 900-species tree reveals increasing diversification rates. PLoS One *12*, e0183070.

12. Robinson, K.J., Bosch, O.J., Levkowitz, G., Busch, K.E., Jarman, A.P., and Ludwig, M. (2019). Social creatures: Model animal systems for studying the neuroendocrine mechanisms of social behaviour. J Neuroendocrinol *31*, e12807.

13. Elphick, M.R., Mirabeau, O., and Larhammar, D. (2018). Evolution of neuropeptide signalling systems. J Exp Biol *221*. https://doi.org/10.1242/jeb.151092.

14. Winslow, J.T., Hastings, N., Carter, C.S., Harbaugh, C.R., and Insel, T.R. (1993). A role for central vasopressin in pair bonding in monogamous prairie voles. Nature *365*, 545–548.

15. Lim, M.M., Wang, Z., Olazábal, D.E., Ren, X., Terwilliger, E.F., and Young, L.J. (2004). Enhanced partner preference in a promiscuous species by manipulating the expression of a single gene. Nature *429*, 754–757.

16. Carter, C.S., DeVries, A.C., and Getz, L.L. (1995). Physiological substrates of mammalian monogamy: the prairie vole model. Neurosci Biobehav Rev *19*, 303–314.

17. Sadino, J.M., and Donaldson, Z.R. (2018). Prairie voles as a model for understanding the genetic and epigenetic regulation of attachment behaviors. ACS Chem. Neurosci. *9*, 1939–1950.

18. Shapiro, L.E., and Insel, T.R. (1990). Infant's response to social separation reflects adult differences in affiliative behavior: a comparative developmental study in prairie and montane voles. Dev. Psychobiol. *23*, 375–393.

19. Young, L.J., Nilsen, R., Waymire, K.G., MacGregor, G.R., and Insel, T.R. (1999). Increased

affiliative response to vasopressin in mice expressing the V1a receptor from a monogamous vole. Nature *400*, 766–768.

20. McGraw, L.A., Davis, J.K., Thomas, P.J., NISC Comparative Sequencing Program, Young, L.J., and Thomas, J.W. (2012). BAC-based sequencing of behaviorally-relevant genes in the prairie vole. PLoS One *7*, e29345.

21. Zhou, C., McCarthy, S.A., and Durbin, R. (2023). YaHS: yet another Hi-C scaffolding tool. Bioinformatics *39*, btac808.

22. Rhie, A., Walenz, B.P., Koren, S., and Phillippy, A.M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol *21*, 245.

23. Hartke, G.T., Leipold, H.W., Huston, K., Cook, J.E., and Saperstein, G. (1974). Three mutations and the karyotype of the prairie vole. White spotting, polydipsia, and muscular dystrophy in Microtus ochrogaster. J Hered *65*, 301–307.

24. McGraw, L.A., Davis, J.K., Young, L.J., and Thomas, J.W. (2011). A genetic linkage map and comparative mapping of the prairie vole (Microtus ochrogaster) genome. BMC Genet. *12*, 60.

25. Reich, L.M. (1981). Microtus pennsylvanicus. Mammalian Species, 1.

26. Schmid, W., and Leppert, M.F. (1968). [Karyotype, heterochromatin and DNA-content in 13 species of voles (Microtinae, Mammalia-Rodentia)]. Arch Julius Klaus Stift Vererbungsforsch Sozialanthropol Rassenhyg *43*, suppl 88–91.

27. Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. (2002). Recent segmental duplications in the human genome. Science *297*, 1003–1007.

28. Išerić, H., Alkan, C., Hach, F., and Numanagić, I. (2022). Fast characterization of segmental duplication structure in multiple genome assemblies. Algorithms Mol Biol *17*, 4.

29. Fredga, K. (1988). Aberrant chromosomal sex-determining mechanisms in mammals, with special reference to species with XY females. Philos Trans R Soc Lond B Biol Sci *322*, 83–95.

30. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

31. Ohno, S. (1970). Evolution by gene duplication (Allen & Unwin; Springer-Verlag).

32. Qi, J., Ma, J., Han, Z., Han, R., Yu, T., and Li, G. (2025). De novo annotation of centromere with centroAnno. bioRxiv. https://doi.org/10.1101/2025.02.19.639205.

33. Schultz, D.T., Haddock, S.H.D., Bredeson, J.V., Green, R.E., Simakov, O., and Rokhsar, D.S. (2023). Ancient gene linkages support ctenophores as sister to other animals. Nature *618*, 110–117.

34. Simakov, O., Bredeson, J., Berkoff, K., Marletaz, F., Mitros, T., Schultz, D.T., O'Connell, B.L., Dear, P., Martinez, D.E., Steele, R.E., et al. (2022). Deeply conserved synteny and the evolution of metazoan chromosomes. Sci. Adv. *8*, eabi5884.

35. Parsons, J.D. (1995). Miropeats: graphical DNA sequence comparisons. Comput Appl Biosci *11*, 615–619.

820 36. Pendleton, A.L., Shen, F., Taravella, A.M., Emery, S., Veeramah, K.R., Boyko, A.R., and Kidd, J.M.
821 (2018). Comparison of village dog and wolf genomes highlights the role of the neural crest in dog
822 domestication. BMC Biol *16*, 64.

823 37. FitzGerald, R.W., and Madison, D.M. (1983). Social organization of a free-ranging population of
824 pine voles, Microtus pinetorum. Behav. Ecol. Sociobiol. *13*, 183–187.

825 38. Oliveras, D., and Novak, M. (1986). A comparison of paternal behaviour in the meadow vole
826 Microtus pennsylvanicus, the pine vole M. pinetorum and the prairie vole M. cchrogaster. Anim.
827 Behav. *34*, 519–526.

828 39. Duckett, D.J., Sullivan, J., Pirro, S., and Carstens, B.C. (2021). Genomic Resources for the North
829 American Water Vole () and the Montane Vole (). GigaByte *2021*, gigabyte19.

830 40. Gouy, A., Wang, X., Kapopoulou, A., Neuenschwander, S., Schmid, E., Excoffier, L., and Heckel,
831 G. (2024). Genomes of Microtus Rodents Highlight the Importance of Olfactory and Immune
832 Systems in Their Fast Radiation. Genome Biol Evol *16*. https://doi.org/10.1093/gbe/evae233.

833 41. Wang, Z., Young, L.J., De Vries, G.J., and Insel, T.R. (1998). Voles and vasopressin: a review of
834 molecular, cellular, and behavioral studies of pair bonding and paternal behaviors. Prog Brain Res
835 *119*, 483–499.

836 42. Jannett, F.J. (1982). Nesting patterns of adult voles, Microtus montanus, in field populations. J.
837 Mammal. *63*, 495–498.

838 43. Madison, D.M., and Mcshea, W.J. (1987). Seasonal changes in reproductive tolerance, spacing, and
839 social organization in meadow voles: A microtine model. Am. Zool. *27*, 899–908.

840 44. Agrell, J. (1995). A shift in female social organization independent of relatedness: an experimental
841 study on the field vole (Microtus agrestis). Behav. Ecol. *6*, 182–191.

842 45. Schweizer, M., Excoffier, L., and Heckel, G. (2007). Fine-scale genetic structure and dispersal in the
843 common vole (Microtus arvalis). Mol Ecol *16*, 2463–2473.

844 46. Jeppsson, B. (1990). Effects of density and resources on the social system of water voles. In Social
845 Systems and Population Cycles in Voles (Birkhäuser Basel), pp. 213–226.

846 47. Fink, S., Excoffier, L., and Heckel, G. (2007). High variability and non-neutral evolution of the
847 mammalian avpr1a gene. BMC Evol Biol *7*, 176.

848 48. Duclot, F., Liu, Y., Saland, S.K., Wang, Z., and Kabbaj, M. (2022). Transcriptomic analysis of
849 paternal behaviors in prairie voles. BMC Genomics *23*, 679.

850 49. Tripp, J.A., Berrio, A., McGraw, L.A., Matz, M.V., Davis, J.K., Inoue, K., Thomas, J.W., Young,
851 L.J., and Phelps, S.M. (2021). Comparative neurotranscriptomics reveal widespread species
852 differences associated with bonding. BMC Genomics *22*, 399.

853 50. Duclot, F., Sailer, L., Koutakis, P., Wang, Z., and Kabbaj, M. (2022). Transcriptomic Regulations
854 Underlying Pair-bond Formation and Maintenance in the Socially Monogamous Male and Female
855 Prairie Vole. Biol Psychiatry *91*, 141–151.

856 51. Waddell, N.J., Liu, Y., Chitaman, J.M., Kaplan, G.J., Wang, Z., and Feng, J. (2023). Transcription
857 and DNA methylation signatures of paternal behavior in hippocampal dentate gyrus of prairie voles.

Sci Rep *13*, 11020.

52. Sadino, J.M., Bradeen, X.G., Kelly, C.J., Brusman, L.E., Walker, D.M., and Donaldson, Z.R. (2023). Prolonged partner separation erodes nucleus accumbens transcriptional signatures of pair bonding in male prairie voles. Elife *12*. https://doi.org/10.7554/eLife.80517.

53. Danoff, J.S., Carter, C.S., Gordevičius, J., Milčiūtė, M., Brooke, R.T., Connelly, J.J., and Perkeybile, A.M. (2024). Maternal oxytocin treatment at birth increases epigenetic age in male offspring. Dev Psychobiol *66*. https://doi.org/10.1002/dev.22452.

54. Karageorgiou, C., Gokcumen, O., and Dennis, M.Y. (2024). Deciphering the role of structural variation in human evolution: a functional perspective. Curr Opin Genet Dev *88*, 102240.

55. Li, H., and Durbin, R. (2024). Genome assembly in the telomere-to-telomere era. Nat Rev Genet *25*, 658–670.

56. Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H.A., Lucas, J.K., Phillippy, A.M., Popejoy, A.B., Asri, M., Carson, C., Chaisson, M.J.P., et al. (2022). The Human Pangenome Project: a global resource to map genomic diversity. Nature *604*, 437–446.

57. Hansen, N.F., Dwarshuis, N., Ji, H.J., Rhie, A., Loucks, H., Logsdon, G.A., Vollger, M.R., Storer, J.M., Kim, J., Adam, E., et al. (2025). A complete diploid human genome benchmark for personalized genomics. bioRxiv. https://doi.org/10.1101/2025.09.21.677443.

58. Berendzen, K.M., and Manoli, D.S. (2022). Rethinking the Architecture of Attachment: New Insights into the Role for Oxytocin Signaling. Affect Sci *3*, 734–748.

59. Naughton, C., Huidobro, C., Catacchio, C.R., Buckle, A., Grimes, G.R., Nozawa, R.-S., Purgato, S., Rocchi, M., and Gilbert, N. (2022). Human centromere repositioning activates transcription and opens chromatin fibre structure. Nat Commun *13*, 5609.

60. Barbosa, S., Paupério, J., Pavlova, S.V., Alves, P.C., and Searle, J.B. (2018). The Microtus voles: Resolving the phylogeny of one of the most speciose mammalian genera using genomics. Mol Phylogenet Evol *125*, 85–92.

61. Liu, Z., Roesti, M., Marques, D., Hiltbrunner, M., Saladin, V., and Peichel, C.L. (2022). Chromosomal fusions facilitate adaptation to divergent environments in threespine stickleback. Mol. Biol. Evol. *39*. https://doi.org/10.1093/molbev/msab358.

62. Bonfield, J.K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., Keane, T., and Davies, R.M. (2021). HTSlib: C library for reading/writing high-throughput sequencing data. Gigascience *10*, giab007.

63. Brown, M.R., Manuel Gonzalez de La Rosa, P., and Blaxter, M. (2025). Tidk: A toolkit to rapidly identify telomeric repeats from genomic datasets. Bioinformatics *41*, btaf049.

64. Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., and Smit, A.F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. U. S. A. *117*, 9451–9457.

65. Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics *Chapter 4*, 4.10.1–4.10.14.

66. Thibaud-Nissen, F., DiCuccio, M., Hlavina, W., Kimchi, A., Kitts, P.A., Murphy, T.D., Pruitt, K.D., and Souvorov, A. (2016). P8008 The NCBI Eukaryotic Genome Annotation Pipeline. J. Anim. Sci. *94*, 184–184.

67. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419.

68. Muzellec, B., Teleńczuk, M., Cabeli, V., and Andreux, M. (2023). PyDESeq2: a python package for bulk RNA-seq differential expression analysis. Bioinformatics *39*, btad547.

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

70. Morgulis, A., Gertz, E.M., Schäffer, A.A., and Agarwala, R. (2006). WindowMasker: window-based masker for sequenced genomes. Bioinformatics *22*, 134–141.

71. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238.

72. Daily, J. (2016). Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. BMC Bioinformatics *17*, 81.

73. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

74. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

75. Ranwez, V., Douzery, E.J.P., Cambon, C., Chantret, N., and Delsuc, F. (2018). MACSE v2: Toolkit for the alignment of Coding Sequences accounting for frameshifts and stop codons. Mol. Biol. Evol. *35*, 2582–2584.

76. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

77. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*, giab008.

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

79. Plessy, C., Mansfield, M.J., Bliznina, A., Masunaga, A., West, C., Tan, Y., Liu, A.W., Grašič, J., Del Río Pisula, M.S., Sánchez-Serna, G., et al. (2024). Extreme genome scrambling in marine planktonic Oikopleura dioica cryptic species. Genome Res. *34*, 426–440.