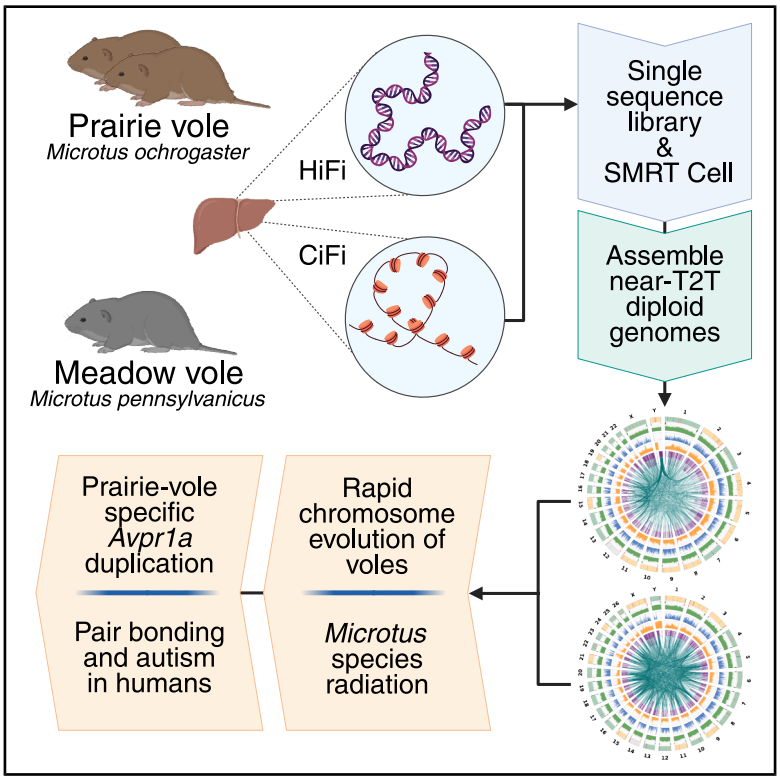
Cell Genomics

Short article

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

Graphical abstract



Highlights

- Single run and library via PacBio HiFi-CiFi produce near T2T diploid assemblies
- Vole genomes are highly structurally divergent despite ~2 million years of separation
- Resolved *Avpr1a* prairie-vole-specific duplication implicated in pair bonding
- Evidence of rapid chromosomal evolution in the drastic radiation of *Microtus* voles

Authors

Mohamed Abuelanin, Gulhan Kaya, Juniper A. Lake, ..., Devanand Manoli, Jessica Tollkuhn, Megan Y. Dennis

Correspondence

mydennis@ucdavis.edu

In brief

Abuelanin et al. combine PacBio HiFi with low-coverage CiFi proximity-ligation reads in a single library and sequencing run to generate contiguous, near-telomere-to-telomere, accurately phased diploid genomes of prairie and meadow voles. These assemblies reveal strikingly rapid genome evolution, including near-complete Y chromosome divergence and a prairie-vole-specific *Avpr1a* duplication implicated in pair bonding.

Abuelanin et al., 2026, Cell Genomics 6, 101336
October 14, 2026
https://doi.org/10.1016/j.xgen.2026.101336

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Short article

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,[4,5] Jonathan Wood,[6] Ksenia Krasheninnikova,[6] Nancy G. Solomon,[7] Zoe R. Donaldson,[8] Karen L. Bales,[9] Kerstin Howe,[6] Jonas Korlach,[2] Devanand Manoli,[4,5] Jessica Tollkuhn,[3] and Megan Y. Dennis[1,10,*]

[1]Genome Center, MIND Institute, and Department of Biochemistry & Molecular Medicine, University of California, Davis, Davis, CA 95616, USA
[2]Pacific Biosciences, Menlo Park, CA 94025, USA
[3]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
[4]Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, San Francisco, CA, USA
[5]Weill Institute for Neuroscience, University of California, San Francisco, San Francisco, CA, USA
[6]Wellcome Sanger Institute, Hinxton CB10 1SA, UK
[7]Department of Biology, Miami University of Ohio, Miami, OH, USA
[8]Department of Molecular, Cellular, and Developmental Biology, Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, CO, USA
[9]Departments of Psychology, Neurobiology, Physiology, and Behavior and California National Primate Research Center, University of California, Davis, Davis, CA, USA
[10]Lead contact
*Correspondence: mydennis@ucdavis.edu
https://doi.org/10.1016/j.xgen.2026.101336

SUMMARY

Comparing prairie voles (*Microtus ochrogaster*), a species that forms lasting pair bonds and exhibits biparental care, with meadow voles (*Microtus pennsylvanicus*), a non-monogamous species showing maternal care only, provides a framework to examine the genomic basis of behavioral divergence over short evolutionary time. Here, we develop a simplified assembly approach combining PacBio HiFi and CiFi sequencing in a single library and run, producing high-quality contiguous chromosome-scale diploid prairie and meadow vole genomes (2.3 Gbp). Comparative analysis reveals substantial differences, including near-complete Y chromosome divergence and a prairie-vole-specific duplication of *Avpr1a*, a vasopressin receptor gene implicated in social bonding and autism in humans. These extensive genomic changes suggest rapid chromosome evolution as a driver of the dramatic *Microtus* radiation, generating ~60 vole species in >2 million years. 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] They provide the substrate to identify functional genes and regulatory regions and to overlay variants that impact them. 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] are vital for basic cellular functions (centromeres), and, when mutated, can lead to human disease. 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 chromosome conformation capture (3C) data, chromosome-scale, fully phased, haplotype-resolved assemblies can be generated.[4] 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,[7] validated on a Mediterranean fruit fly genome assembly (~600 Mbp) and requiring lower coverage compared with Hi-C for effective scaffolding. Since both DNA products can be sequenced on a uniform platform and separated bioinformatically based on distinct sizes/adapters, there is an opportunity to combine HiFi/CiFi into one library preparation and optimized workflow that reduces expense and technical complexity.

Here, we demonstrate a single-library HiFi-CiFi assembly approach for two mammalian genomes approximately 4-fold

Please cite this article in press as: Abuelanin et al., Single-library chromosome-scale diploid assemblies of vole genomes resolve a species-specific duplication implicated in pair bonding, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101336

larger (~2.3 Gbp) than the previously sequenced fruit fly: prairie vole (*Microtus ochrogaster*) and meadow vole (*Microtus pennsylvanicus*). Voles diverged from the murine lineage, including mice and rats, approximately 20 million years ago (mya).[8] Mice and rats (within subfamily *Murinae*), both extensively used as model organisms, share a common ancestor estimated at ~12–23 mya, while the *Arvicolinae* subfamily comprising voles originated ~6–7 mya, with its most dramatic radiation occurring within the genus *Microtus* ~2 mya.[9,10] Similar to humans, prairie voles are among the few mammalian species exhibiting social monogamy.[11,12] They form enduring social attachments to their partners, known as pair bonds,[13] and both parents care for offspring. In contrast, closely related meadow voles 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. Further, prairie vole behaviors better parallel aspects of human interactions compared with mice or rats, making them a valuable model for understanding the neural basis of human 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.[14,15] Seminal comparative behavioral and pharmacologic studies have implicated the nonapeptide hormones oxytocin and vasopressin in prairie vole social attachment.[16–19] 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.[20,21]

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 bacterial artificial chromosome (BAC) sequencing,[22] was not present in MicOch1.0. Subsequent work on the *Avpr1a* paralog, including verification of its existence in prairie vole or other related species, has been hampered 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.

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 and meadow vole, 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 QV 38.

HiFi (88%) and CiFi (12%) reads were segregated using unique Ampli-Fi adapter sequences flanking CiFi fragments (STAR 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, producing 5× predicted coverage (Figure 1B; Table S1). CiFi reads were then converted from multisegment concatemers into chromatin-contact pairs: 1.37 million CiFi reads yielded 23.4 million contact pairs for prairie vole, and 1.54 million reads yielded 8.1 million pairs for meadow vole. The greater contact-pair yield for prairie vole reflects slightly smaller segment sizes (median 1.4 vs. 1.7 kbp) and longer CiFi read lengths (median 7.9 vs. 7.2 kbp). Together, these results demonstrate that a combined HiFi and CiFi library produces 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[23] (Figure 1C). CiFi-phased contigs were comparable or longer in three of four haplotypes and achieved better or equivalent scaffold consolidations, approaching expected chromosome numbers more closely than Hi-C (Figures S2–S4; Table S2; Methods S1). Downsampling CiFi reads, we observed near-optimal scaffold N50 of all four haplotypes at 40% input or 5 Gbp of data (~2× genome coverage) (Figure S5). We next extended this test to a human genome (3.2 Gbp) where 2× CiFi and 30× HiFi are achievable on a single SMRT Cell. Comparing a hifiasm assembly generated with 2× CiFi[7] against pedigree-phased NA12878 variants,[24] we observed highly accurate haplotype phasing: the median per-chromosome switch error was 0.18% vs. 0.38% for HiFi only and essentially equal to standard 30× Hi-C (0.16%) (Figure S6; Table S3; Methods S1). Together, this indicates that 2× CiFi is suitable to achieve equal-to-better phasing and contiguity assembly metrics compared to a standard ~30× Hi-C.

Final curated vole assemblies

Assemblies produced with HiFi and CiFi required minimal correction during manual curation (see STAR 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 S7). 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[25] (diploid QV 50.9–51.9), with diploid

Please cite this article in press as: Abuelanin et al., Single-library chromosome-scale diploid assemblies of vole genomes resolve a species-specific duplication implicated in pair bonding, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101336

CellPress
OPEN ACCESS

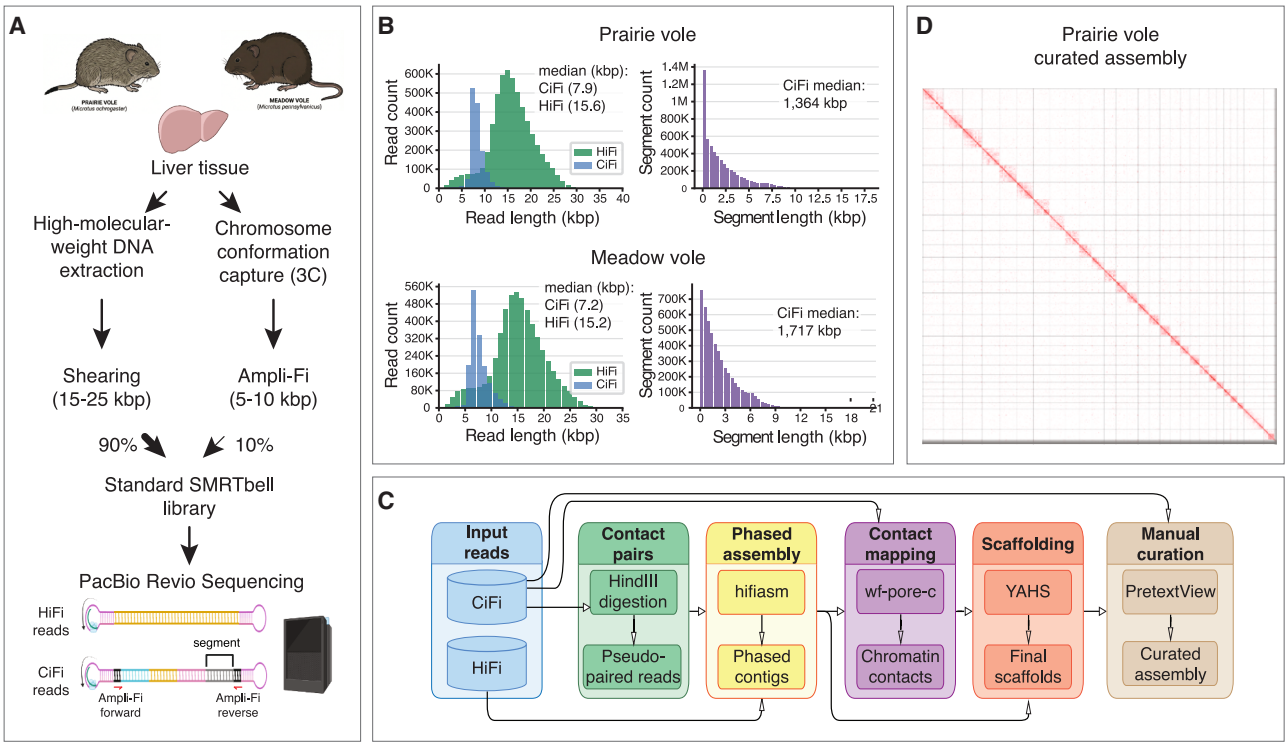


Figure 1. Prairie and meadow vole single-library HiFi-CiFi workflow and sequencing statistics

(A) Library preparation from liver tissue using DNA for HiFi (left) and chromatin for CiFi (right) combined (90:10) to construct a single SMRTbell library.

(B) Read-length (HiFi and CiFi) and segment-length (CiFi) distributions.

(C) Computational pipeline for HiFi-CiFi-based genome assembly and scaffolding.

(D) CiFi contact map of the curated prairie vole Hap1 assembly. Uncurated contact maps for both haplotypes and species are in Figures S2 and S3.

See also Figures S1 and S4–S7; Tables S1, S2, S3 and S4; and Methods S1.

heterozygosities of 0.99% (prairie vole) and 1.14% (meadow vole) (Tables 1 and S4; Figure S8).

The prairie vole diploid assembly represents a complete karyotype (2n = 54)[26,27] with 52 autosomes, chrX, and chrY (total: 4.83 Gbp; Hap1: 2.53 Gbp). Telomeres were detected on all but three chromosomes (mean 3.5 gaps/chromosome). Of these, 31 represent telomere-to-telomere (T2T) scaffolds and three gapless contigs (overall 63%) (Figure 2; Table S5). 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 vs. ~30 Mbp in 69 unplaced scaffolds in prairie vole Hap1 from 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 S9; Tables S6 and S7).

Similarly, the meadow vole assembly matches the expected karyotype (2n = 46)[28,29] across 44 autosomes, chrX, and chrY, representing 4.51 Gbp total sequence (Hap1: 2.36 Gbp). All but two chromosomes have at least one telomere detected (mean 3.4 gaps/chromosome), with 17 T2T scaffolds and six gapless contigs (overall 50%; Figure 2; Table S5). A majority of the meadow vole autosomes are telocentric (42/44), with the two sex chromosomes classified as subtelocentric,[28,29] likely contributing to the higher proportion of single telomeres detected in the meadow vole vs. the 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 S10).

Moving forward, we characterized primary Hap1 chromosomes, comprising single sets of autosomes and both sex chromosomes. Using RNA sequencing (RNA-seq) data from the liver, brain, and testes of adults and embryonic day 11.5 (E11.5; prairie vole) or E13.5 (meadow vole) embryos, we annotated ~20,500 protein-coding genes per species (STAR Methods; Figure 2; Table S8). 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, S11, and S12; Table S9). Near-equal numbers of SDs (regions >1 kbp at >90% sequence identity[30,31]) were identified in both genomes (~5.3K), with 72%–79% intersecting an annotated gene (prairie vole: 1,151 genes; meadow vole: 1,194 genes). Despite this similarity, prairie vole

Please cite this article in press as: Abuelanin et al., Single-library chromosome-scale diploid assemblies of vole genomes resolve a species-specific duplication implicated in pair bonding, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101336

Table 1. Summary of prairie- and meadow-vole-curated genomes

Metric	Prairie vole (this study)	Prairie vole (this study)	Prairie vole (MicOch1.0)	Meadow vole (this study)	Meadow vole (this study)	Meadow vole (mMicPen1)	Meadow vole (mMicPen1)
Haplotype	Hap1[a]	Hap2	merged	Hap1[a]	Hap2	Hap1[a]	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%
Diploid QV	51.9	51.9	–	50.9	50.9	–	–
Diploid k-mer completeness	99.5%	99.5%	–	99.7%	99.7%	–	–
Heterozygosity	0.98%–1.01%	0.98%–1.01%	–	1.13%–1.15%	1.13%–1.15%	–	–

See also Figures S8–S10 and Tables S4, S5, S6, and S7.
[a]Primary assemblies with both sex chromosomes.

harbored nearly double the duplicated sequence (54 Mbp, 2.15% of genome) compared to meadow vole (28 Mbp, 1.2%) (Figures 2A, S11, and S12; Table S10). 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*.[32,33] The most notable differences are evident on the chrY (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) (Tables S6, S11, and S12). 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 several chromosomes, including homologous chr1 and chrX (Figure S13). Mapping single-nucleotide, indel, and structural variants identified ~415K coding variants using prairie vole as a reference[37]; of these, 839 (VEP) and 1,043 (SnpEff) were annotated as likely gene disrupting (LGD), impacting a consensus set of 253 single-copy 1:1 ortholog genes (Tables S13 and S14). Focusing on candidate genes implicated in behavior—including those known to interact with nonapeptide hormones oxytocin and vasopressin—shows global reduced rates of substitutions relative to synonymous mutations (Ka/Ks << 1; Figure S14) and no LGD variants obviously impacting function, suggesting purifying selection (Tables S15 and S16).

Understanding that gene duplication is a common mechanism of trait innovation across the animal kingdom,[38] we identified 306 and 389 gene duplications unique to prairie or meadow vole, respectively (Table S17). 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,[22] both *Avpr1a* paralogs are missing in the current prairie vole reference (MicOch1.0), impeding analyses of these genes. Each prairie vole haplotype shows *Avpr1a* paralogs residing ~900 kbp apart on chr26, separated proximally by a large stretch of nearly identical satellite repeats (547 kbp and 517 kbp) 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, defined as the largest stretch of satellite sequence per chromosome[39] and including a >100 kbp hypomethylated region, not observed adjacent to *Avpr1a* in chr20 of the meadow vole (instead annotated as telocentric) or at any other regions in either species' genomes. Computing diploid copy number (CN) from Illumina read depth[40] showed an elevated CN across the chr26 centromere annotation and a CN of 4 unique to the *Avpr1a*-containing SD (Figure 3E) in prairie vole vs. CN of 2 in meadow vole, woodland vole (*M. pinetorum*, also known as pine vole, a species exhibiting monogamous pair bonding[41,42] sequenced as part of this study), and an additional four publicly available *Microtus* genomes[43,44] representing species with more promiscuous behaviors.[45–50]

Both curated prairie vole references show 97% nucleotide identity between *Avpr1a* paralogs, with *Avpr1a2* harboring

Please cite this article in press as: Abuelanin et al., Single-library chromosome-scale diploid assemblies of vole genomes resolve a species-specific duplication implicated in pair bonding, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101336

CellPress
OPEN ACCESS

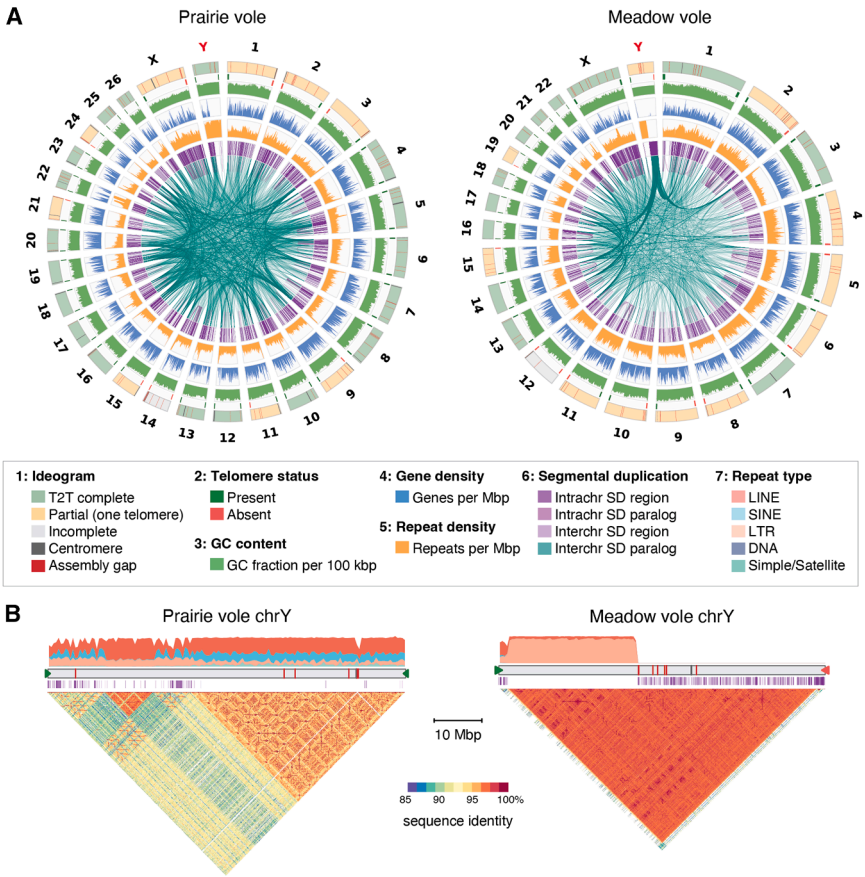


Figure 2. Prairie and meadow vole assembly characteristics

(A) Circos plots depicting chromosome characteristics of Hap1 assemblies for prairie (left) and meadow (right) vole. Tracks are described in the legend from outer to inner ring (1: Ideogram; 6: segmental duplication).

(B) Repeat and SD distributions are shown above and below the chrY ideogram for each species. Ideogram annotations follow the color scheme in (A), indicating centromere and assembly gap (legend 1), telomere status (legend 2), SDs (purple; legend 6), and repeat types (legend 7). Dotplots below each ideogram show chrY sequence identity. Annotated ideograms of all Hap1 chromosomes are shown in Figures S11 and S12.

See also Figures S8–S10 and Tables S8–S10.

frameshift variants encoding a truncated protein (218 aa vs. 420 aa full length). This result was supported by both Hap1 and Hap2 assemblies (Figure S15; Table S18) and is consistent with prior reports.[21,22] If translated, the first 199 aa would be identical between AVPR1A paralogs, encoding the extracellular vasopressin-binding domain and the first four of seven transmembrane domains but lacking the intracellular G-protein binding region,[51] with the terminating 19 aa unique to the prairie-vole-specific AVPR1A2. Quantifying transcript abundances using published RNA-seq data[52–57] reveals consistent expression of both paralogs across five of eight tested brain regions, with the highest expression in the medial preoptic area, albeit reduced for *Avpr1a2* (~1×–4× relative to *Avpr1a*) (Figures 3G and S16). Divergent expression could be influenced by a 636-bp indel ~1.6 kbp upstream of the *Avpr1a* start codons (Figure 3F) directly adjacent to a polymorphic microsatellite repeat previously associated with differences in *Avpr1a* expression and socio-behavioral traits in prairie voles.[35,36] Further, divergent chromatin interactions outside of the duplicated region likely occur between paralogs,[58] with the SD breakpoint only 4 kbp upstream. Even if *Avpr1a2* ultimately proves to be a non-functional 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

Most recent efforts to build gigabase-sized chromosome-scale genomes require multiple technologies,[59] 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[60]; even higher coverage is necessary to achieve complete diploid T2T genomes.[61] 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 run, producing contiguous (scaffold N50 90–125 Mbp) and high-quality (QV > 50) diploid assemblies. While large amounts of input were available here, experimental approaches are scalable to smaller amounts, as we previously demonstrated.[7] This is particularly important when studying rare, endangered, or difficult-to-sample species and tissues, where collecting large DNA quantities might be impractical. CiFi downsampling shows that 2× CiFi plus 30× HiFi coverage, achievable on a single SMRT Cell, is sufficient to accurately assemble a larger, more homogeneous (heterozygosity ~0.1% vs. 1% for voles) human diploid genome (3.2 Gbp) (Figures S5 and S6; Table S3). Relative to the ideal multiplatform recipe, the single-library approach reduces cost at least 3-fold while still resolving more than half of chromosomes at T2T scale. In effect, this lowers technical and cost barriers, enabling institutions with only a single sequencing platform to produce high-quality and contiguous genomes.

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.[62] This includes high-quality annotated assemblies for both prairie and meadow voles, repeat and SD annotations, CN maps across additional

Please cite this article in press as: Abuelanin et al., Single-library chromosome-scale diploid assemblies of vole genomes resolve a species-specific duplication implicated in pair bonding, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101336

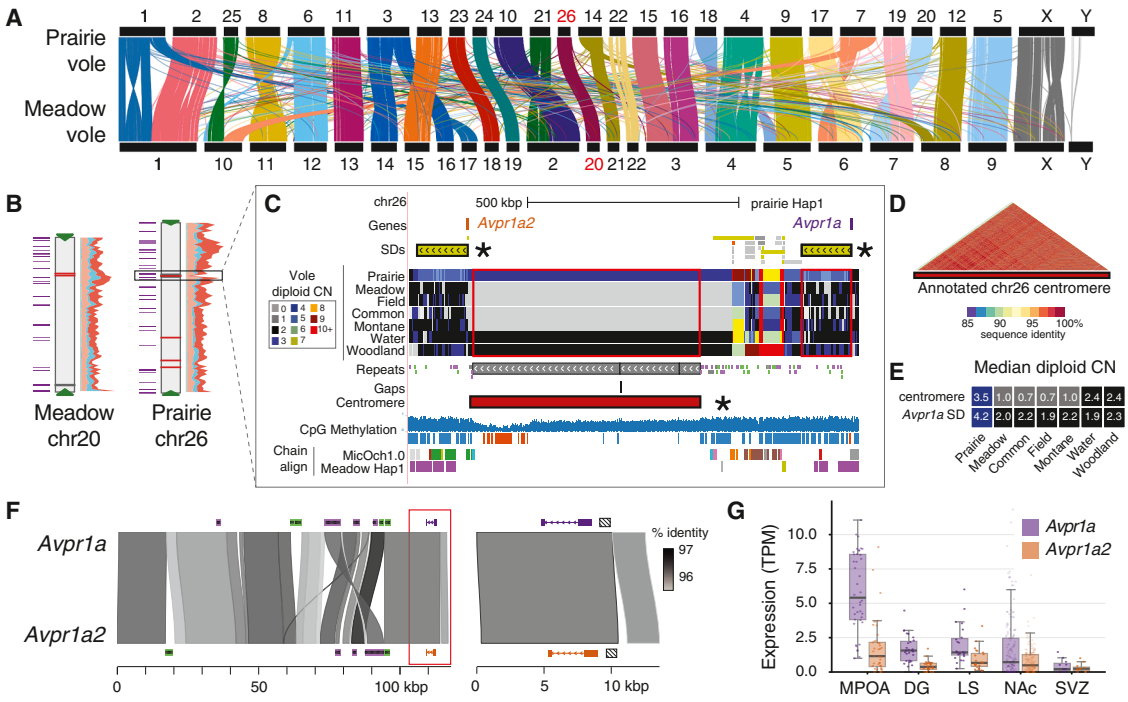


Figure 3. Prairie and meadow vole cross-species comparison

(A) Ribbon plots depicting synteny between voles[34] (colored by homologous chromosome) with chromosome numbers indicated above and below.

(B) Repeat (right) and SD (left) annotations of the homologous chromosomes carrying *Avpr1a*. Color scheme as in Figure 2B.

(C) UCSC Genome Browser view of the prairie vole Hap1 *Avpr1a* locus (chr26:13.0 Mbp–14.1 Mbp). SDs of paralogs and intervening centromeric repeat are highlighted with an asterisk. Windowed copy numbers (CNs) of vole species are depicted by colors. CpG methylation is shown as the proportion of methylated HiFi reads (0–1); methylated (blue) and hypomethylated (red) regions are annotated below. Alignment chains for MicOch1.0 and the meadow vole Hap1 assembly (this study) are shown at the bottom.

(D) Dotplot of the annotated centromere.

(E) Median CN of the two regions highlighted with red boxes; colors match the legend in (C).

(F) Hap1 genomic alignments of the prairie vole SD comprising *Avpr1a* (purple) and *Avpr1a2* (orange) with repeat annotations; the red box indicates the region enlarged on the right. The hashed box upstream of both paralogs marks a microsatellite associated with altered *Avpr1a* expression.[35,36]

(G) RNA-seq expression of *Avpr1a* paralogs in the MPOA, DG, LS, NAc, and SVZ. Boxplots show median (center line), IQR (box), and 1.5× IQR whiskers; dots are individual samples. MPOA, medial preoptic area; DG, dentate gyrus; LS, lateral septum; NAc, nucleus accumbens; SVZ, subventricular zone; IQR, interquartile range; TPM, transcripts per million reads.

See also Figures S11–S16 and Tables S11–S18.

vole species, chain alignments facilitating genome liftover between species/builds (e.g., VGP mMicPen1), and named gene orthologs enabling transcriptomic comparisons, all publicly accessible through a UCSC Genome Browser hub (see STAR Methods). 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.[14,15] 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 functions. Hypo-methylation across satellite repeats strengthens annotation of the prairie vole chr26 centromere and places *Avpr1a* approximately 350 kbp downstream, within the pericentromeric region. This positioning may expose the gene to centromerization-associated effects, including increased H3K27me3-mediated repression, altered chromatin accessibility, and elevated genomic instability[63]—consequences that would alter vasopressin signaling dynamics and likely facilitated the emergence of the prairie-vole-specific *Avpr1a2* paralog. This is a compelling finding given that prairie voles exhibit elevated *Avpr1a* expression in the ventral pallidum while meadow voles show higher expression in the lateral septum.[21] 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 identified no obvious amino acid divergence between prairie and meadow

voles (Figure S14; Tables S15 and S16), suggesting that regulatory alterations, rather than coding sequence change, play an outsized role in behavioral differences between these species. These near-complete genomes will enable comprehensive comparisons of altered chromatin structure and gene regulation underlying bond formation and social memory. Ultimately, integration with human genetic studies, such as *AVPR1A* variants linked to autism spectrum disorder,[64–69] provides a framework for understanding both how social behavior is encoded neurobiologically and how genetic disruptions to these pathways contribute 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 multicontact 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 a >8-fold greater ability to infer haplotypes vs. 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, 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. Such efforts would also support our observation of near-complete chrY divergence between prairie and meadow voles, in alignment with past work that suggests rapid chromosome evolution as a possible driver of the *Microtus* radiation,[70] one of the most dramatic speciation events recorded in vertebrates.[71]

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.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by lead contact, Megan Y. Dennis (mydennis@ucdavis.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All raw sequencing data generated from this study have been deposited to ENA and NCBI (ENA/GenBank: PRJEB108798). Accession numbers of data used from published sources can be found in the key resources table.
- All source code, workflows, and processed data associated with this study are publicly available through https://github.com/mydennislab/2026-voles-assembly and https://github.com/sanger-tol/ and has been deposited at Zenodo (https://doi.org/10.5281/zenodo.20501386[72]).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank Dr. Aaron Wenger and Jacob Brandenburg for coordination support in generating and analyzing the HiFi-CiFi datasets. We also thank 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 US National Science Foundation (NSF) (CAREER 2145885 to M.Y.D.and CAREER IOS-2045348 to Z.R.D.) and National Institutes of Health (NIH) grants from the National Institute of Mental Health (NIMH) (RF1MH132818 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. Cold Spring Harbor Laboratory vole 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., K.M.B., N.G.S., Z.R.D., K.L.B., and J.T. generated biospecimens and/or data for the project. M.A., J.A.L., K.K., J.W., K.H., and M.Y.D. performed bioinformatic and genomic analysis. N.G.S., Z.R.D., K.L.B., D.M., and J.T. provided biomaterials and vole expertise. J.K., D.M., J.T., and M.Y.D. devised the project. M.A., J.K., D.M., J.T., and M.Y.D. wrote the manuscript, and all authors edited and approved the manuscript.

DECLARATION OF INTERESTS

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

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used Claude AI to copyedit portions of the text. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Vole procedures

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2026.101336.

Received: March 6, 2026
Revised: June 2, 2026
Accepted: July 23, 2026

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. https://doi.org/10.1038/s41586-021-03451-0.
2. Dennis, M.Y., and Eichler, E.E. (2016). Human adaptation and evolution by segmental duplication. Curr. Opin. Genet. Dev. *41*, 44–52. https://doi.org/10.1016/j.gde.2016.08.001.
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. https://doi.org/10.1038/s41587-019-0217-9.
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. https://doi.org/10.1038/s41592-020-01056-5.
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. https://doi.org/10.1126/science.1181369.
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. https://doi.org/10.1038/nbt.2727.
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., and Dennis, M.Y. (2025). CiFi: accurate long-read chromosome conformation capture with low-input requirements. Nat. Commun. *17*, 215. https://doi.org/10.1038/s41467-025-66918-y.
8. Steppan, S.J., and Schenk, J.J. (2017). Muroid rodent phylogenetics: 900-species tree reveals increasing diversification rates. PLoS One *12*, e0183070. https://doi.org/10.1371/journal.pone.0183070.
9. Galewski, T., Tilak, M.-K., Sanchez, S., Chevret, P., Paradis, E., and Douzery, E.J.P. (2006). The evolutionary radiation of Arvicolinae rodents (voles and lemmings): relative contribution of nuclear and mitochondrial DNA phylogenies. BMC Evol. Biol. *6*, 80. https://doi.org/10.1186/1471-2148-6-80.
10. Conroy, C.J., and Cook, J.A. (2000). Molecular Systematics of a Holarctic Rodent (Microtus: Muridae). J. Mammal. *81*, 344–359. https://doi.org/10.1093/jmammal/81.2.344.
11. Kleiman, D.G. (1977). Monogamy in mammals. Q. Rev. Biol. *52*, 39–69. https://doi.org/10.1086/409721.
12. Lukas, D., and Clutton-Brock, T.H. (2013). The evolution of social monogamy in mammals. Science *341*, 526–530. https://doi.org/10.1126/science.1238677.
13. 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. https://doi.org/10.1016/j.yhbeh.2021.105062.
14. 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. https://doi.org/10.1111/jne.12807.
15. Elphick, M.R., Mirabeau, O., and Larhammar, D. (2018). Evolution of neuropeptide signalling systems. J. Exp. Biol. *221*, jeb151092. https://doi.org/10.1242/jeb.151092.
16. 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. https://doi.org/10.1038/365545a0.
17. 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. https://doi.org/10.1038/nature02539.
18. 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. https://doi.org/10.1016/0149-7634(94)00070-h.
19. 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. https://doi.org/10.1021/acschemneuro.7b00475.
20. 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. https://doi.org/10.1002/dev.420230502.
21. 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. https://doi.org/10.1038/23475.
22. 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. https://doi.org/10.1371/journal.pone.0029345.
23. Zhou, C., McCarthy, S.A., and Durbin, R. (2023). YaHS: yet another Hi-C scaffolding tool. Bioinformatics *39*, btac808. https://doi.org/10.1093/bioinformatics/btac808.
24. Kronenberg, Z., Nolan, C., Porubsky, D., Mokveld, T., Rowell, W.J., Lee, S., Dolzhenko, E., Chang, P.-C., Holt, J.M., Saunders, C.T., et al. (2025). The Platinum Pedigree: a long-read benchmark for genetic variants. Nat. Methods *22*, 1669–1676. https://doi.org/10.1038/s41592-025-02750-y.
25. 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. https://doi.org/10.1186/s13059-020-02134-9.

26. 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. https://doi.org/10.1093/oxfordjournals.jhered.a108533.
27. 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. https://doi.org/10.1186/1471-2156-12-60.
28. Reich, L.M. (1981). Microtus pennsylvanicus. Mammalian Species *159*, 1–8. https://doi.org/10.2307/3503976.
29. 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*, 88–91.
30. 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. https://doi.org/10.1126/science.1072047.
31. 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. https://doi.org/10.1186/s13015-022-00210-2.
32. 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. https://doi.org/10.1098/rstb.1988.0116.
33. Couger, M.B., Roy, S.W., Anderson, N., Gozashti, L., Pirro, S., Millward, L.S., Kim, M., Kilburn, D., Liu, K.J., Wilson, T.M., et al. (2021). Sex chromosome transformation and the origin of a male-specific X chromosome in the creeping vole. Science *372*, 592–600. https://doi.org/10.1126/science.abg7019.
34. 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. https://doi.org/10.1038/s41586-023-05936-6.
35. Hammock, E.A.D., and Young, L.J. (2005). Microsatellite instability generates diversity in brain and sociobehavioral traits. Science *308*, 1630–1634. https://doi.org/10.1126/science.1111427.
36. Donaldson, Z.R., and Young, L.J. (2013). The relative contribution of proximal 5' flanking sequence and microsatellite variation on brain vasopressin 1a receptor (Avpr1a) gene expression and behavior. PLoS Genet. *9*, e1003729. https://doi.org/10.1371/journal.pgen.1003729.
37. 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. https://doi.org/10.1186/s13059-016-0974-4.
38. Ohno, S. (1970). Evolution by Gene Duplication (Allen & Unwin; Springer-Verlag).
39. Qi, J., Ma, J., Han, Z., Han, R., Yu, T., and Li, G. (2025). *De novo* annotation of centromere with centroAnno. Preprint at bioRxiv. https://doi.org/10.1101/2025.02.19.639205.
40. Pendleton, A.L., Shen, F., Taravella, A.M., Emery, S., Veeramah, K.R., Boyko, A.R., and Kidd, J.M. (2018). Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. BMC Biol. *16*, 64. https://doi.org/10.1186/s12915-018-0535-2.
41. FitzGerald, R.W., and Madison, D.M. (1983). Social organization of a free-ranging population of pine voles, Microtus pinetorum. Behav. Ecol. Sociobiol. *13*, 183–187. https://doi.org/10.1007/bf00299921.
42. Oliveras, D., and Novak, M. (1986). A comparison of paternal behaviour in the meadow vole Microtus pennsylvanicus, the pine vole M. pinetorum and the prairie vole M. cchrogaster. Anim. Behav. *34*, 519–526. https://doi.org/10.1016/s0003-3472(86)80120-8.
43. Duckett, D.J., Sullivan, J., Pirro, S., and Carstens, B.C. (2021). Genomic Resources for the North American Water Vole (Microtus richardsoni) and the Montane Vole (Microtus montanus). GigaByte *2021*, gigabyte19. https://doi.org/10.46471/gigabyte.19.
44. Gouy, A., Wang, X., Kapopoulou, A., Neuenschwander, S., Schmid, E., Excoffier, L., and Heckel, G. (2024). Genomes of Microtus Rodents Highlight the Importance of Olfactory and Immune Systems in Their Fast Radiation. Genome Biol. Evol. *16*, evae233. https://doi.org/10.1093/gbe/evae233.
45. Wang, Z., Young, L.J., De Vries, G.J., and Insel, T.R. (1998). Voles and vasopressin: a review of molecular, cellular, and behavioral studies of pair bonding and paternal behaviors. Prog. Brain Res. *119*, 483–499. https://doi.org/10.1016/s0079-6123(08)61589-7.
46. Jannett, F.J. (1982). Nesting patterns of adult voles, Microtus montanus, in field populations. J. Mammal. *63*, 495–498. https://doi.org/10.2307/1380450.
47. Madison, D.M., and Mcshea, W.J. (1987). Seasonal changes in reproductive tolerance, spacing, and social organization in meadow voles: A microtine model. Am. Zool. *27*, 899–908. https://doi.org/10.1093/icb/27.3.899.
48. Agrell, J. (1995). A shift in female social organization independent of relatedness: an experimental study on the field vole (*Microtus agrestis*). Behav. Ecol. *6*, 182–191. https://doi.org/10.1093/beheco/6.2.182.
49. Schweizer, M., Excoffier, L., and Heckel, G. (2007). Fine-scale genetic structure and dispersal in the common vole (Microtus arvalis). Mol. Ecol. *16*, 2463–2473. https://doi.org/10.1111/j.1365-294X.2007.03284.x.
50. Jeppsson, B. (1990). Effects of density and resources on the social system of water voles. In Social Systems and Population Cycles in Voles (Birkhäuser Basel), pp. 213–226.
51. Fink, S., Excoffier, L., and Heckel, G. (2007). High variability and non-neutral evolution of the mammalian avpr1a gene. BMC Evol. Biol. *7*, 176. https://doi.org/10.1186/1471-2148-7-176.
52. Duclot, F., Liu, Y., Saland, S.K., Wang, Z., and Kabbaj, M. (2022). Transcriptomic analysis of paternal behaviors in prairie voles. BMC Genom. *23*, 679. https://doi.org/10.1186/s12864-022-08912-y.
53. Tripp, J.A., Berrio, A., McGraw, L.A., Matz, M.V., Davis, J.K., Inoue, K., Thomas, J.W., Young, L.J., and Phelps, S.M. (2021). Comparative neurotranscriptomics reveal widespread species differences associated with bonding. BMC Genom. *22*, 399. https://doi.org/10.1186/s12864-021-07720-0.
54. Duclot, F., Sailer, L., Koutakis, P., Wang, Z., and Kabbaj, M. (2022). Transcriptomic Regulations Underlying Pair-bond Formation and Maintenance in the Socially Monogamous Male and Female Prairie Vole. Biol. Psychiatry *91*, 141–151. https://doi.org/10.1016/j.biopsych.2020.11.022.
55. Waddell, N.J., Liu, Y., Chitaman, J.M., Kaplan, G.J., Wang, Z., and Feng, J. (2023). Transcription and DNA methylation signatures of paternal behavior in hippocampal dentate gyrus of prairie voles. Sci. Rep. *13*, 11020. https://doi.org/10.1038/s41598-023-37521-2.
56. 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*, e80517. https://doi.org/10.7554/eLife.80517.
57. 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*, e22452. https://doi.org/10.1002/dev.22452.
58. 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. https://doi.org/10.1016/j.gde.2024.102240.

59. Li, H., and Durbin, R. (2024). Genome assembly in the telomere-to-telomere era. Nat. Rev. Genet. *25*, 658–670. https://doi.org/10.1038/s41576-024-00718-w.

60. 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. https://doi.org/10.1038/s41586-022-04601-8.

61. 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. (2026). A complete diploid human genome benchmark for personalized genomics. Cell *189*. https://doi.org/10.1016/j.cell.2026.06.016.

62. 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. https://doi.org/10.1007/s42761-022-00142-5.

63. 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. https://doi.org/10.1038/s41467-022-33426-2.

64. Ebstein, R.P., Israel, S., Lerer, E., Uzefovsky, F., Shalev, I., Gritsenko, I., Riebold, M., Salomon, S., and Yirmiya, N. (2009). Arginine Vasopressin and Oxytocin Modulate Human Social Behavior. Ann. N.Y. Acad. Sci. *1167*, 87–102. https://doi.org/10.1111/j.1749-6632.2009.04541.x.

65. Francis, S.M., Kim, S.-J., Kistner-Griffin, E., Guter, S., Cook, E.H., and Jacob, S. (2016). ASD and Genetic Associations with Receptors for Oxytocin and Vasopressin—*AVPR1A*, *AVPR1B*, and *OXTR*. Front. Neurosci. *10*, 516. https://doi.org/10.3389/fnins.2016.00516.

66. Yang, S.Y., Kim, S.A., Hur, G.M., Park, M., Park, J.-E., and Yoo, H.J. (2017). Replicative genetic association study between functional polymorphisms in *AVPR1A* and social behavior scales of autism spectrum disorder in the Korean population. Mol. Autism *8*, 44. https://doi.org/10.1186/s13229-017-0161-9.

67. Yirmiya, N., Rosenberg, C., Levi, S., Salomon, S., Shulman, C., Nemanov, L., Dina, C., and Ebstein, R.P. (2006). Association between the arginine vasopressin 1a receptor (AVPR1a) gene and autism in a family-based study: mediation by socialization skills. Mol. Psychiatry *11*, 488–494. https://doi.org/10.1038/sj.mp.4001812.

68. Wassink, T.H., Piven, J., Vieland, V.J., Pietila, J., Goedken, R.J., Folstein, S.E., and Sheffield, V.C. (2004). Examination of AVPR1a as an autism susceptibility gene. Mol. Psychiatry *9*, 968–972. https://doi.org/10.1038/sj.mp.4001503.

69. Kim, S.-J., Young, L.J., Gonen, D., Veenstra-VanderWeele, J., Courchesne, R., Courchesne, E., Lord, C., Leventhal, B.L., Cook, E.H., Jr., and Insel, T.R. (2002). Transmission disequilibrium testing of arginine vasopressin receptor 1A (AVPR1A) polymorphisms in autism. Mol. Psychiatry *7*, 503–507. https://doi.org/10.1038/sj.mp.4001125.

70. Lemskaya, N.A., Romanenko, S.A., Golenishchev, F.N., Rubtsova, N.V., Sablina, O.V., Serdukova, N.A., O'Brien, P.C.M., Fu, B., Yiğit, N., Ferguson-Smith, M.A., et al. (2010). Chromosomal evolution of Arvicolinae (Cricetidae, Rodentia). III. Karyotype relationships of ten Microtus species. Chromosome Res. *18*, 459–471. https://doi.org/10.1007/s10577-010-9124-0.

71. 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. https://doi.org/10.1016/j.ympev.2018.03.017.

72. Abuelanin, M. (2026). Source code of Single-library chromosome-scale diploid assemblies of vole genomes resolve a species-specific duplication implicated in pair bonding. Zenodo. https://doi.org/10.5281/ZENODO.20501386.

73. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021.

74. Moshi, Dyachkoova, M., and Weijers, J., moshi/pyGenomeViz: A genome visualization python package for comparative genomics (GitHub). https://github.com/moshi4/pyGenomeViz?tab=MIT-1-ov-file.

75. Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. PLoS Comput. Biol. *14*, e1005944. https://doi.org/10.1371/journal.pcbi.1005944.

76. 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. https://doi.org/10.1093/gigascience/giab007.

77. Li, H., Feng, X., and Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. Genome Biol. *21*, 265. https://doi.org/10.1186/s13059-020-02168-z.

78. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

79. Griffiths, S., Wright, C., Nicholls, S., Parker, M., Harrington, E., Horner, N., Hesketh, S., and Garcia Garcia. N wf-pore-c at v1.3.0 (GitHub). https://github.com/epi2me-labs/wf-pore-c.

80. Li, H. seqtk: Toolkit for processing sequences in FASTA/Q formats (GitHub). https://github.com/lh3/seqtk.

81. 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. https://doi.org/10.1093/bioinformatics/btaf049.

82. Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun. *11*, 1432. https://doi.org/10.1038/s41467-020-14998-3.

83. Li, H. yak: Yet another k-mer analyzer (GitHub). https://github.com/lh3/yak.

84. Ewels, P.A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M.U., Di Tommaso, P., and Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. Nat. Biotechnol. *38*, 276–278. https://doi.org/10.1038/s41587-020-0439-x.

85. Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics *34*, 867–868. https://doi.org/10.1093/bioinformatics/btx699.

86. Casper, J., Speir, M.L., Raney, B.J., Perez, G., Nassar, L.R., Lee, C.M., Hinrichs, A.S., Gonzalez, J.N., Fischer, C., Diekhans, M., et al. (2026). The UCSC Genome Browser database: 2026 update. Nucleic Acids Res. *54*, D1331–D1335. https://doi.org/10.1093/nar/gkaf1250.

87. 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. USA *117*, 9451–9457. https://doi.org/10.1073/pnas.1921046117.

88. Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics *4*, 4.10.1–4.10.14. https://doi.org/10.1002/0471250953.bi0410s25.

89. Strope, P., Tvedte, E., Sweeney, D., and Sylla, P. egapx: Eukaryotic Genome Annotation Pipeline-External caller scripts and documentation (GitHub). https://github.com/ncbi/egapx.

90. 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. https://doi.org/10.1038/nmeth.4197.

91. Pertea, G. gffread: GFF/GTF utility providing format conversions, region filtering, FASTA sequence extraction and more (GitHub). https://github.com/gpertea/gffread.

92. Kidd, J., Shen, J., fastCN (GitHub). https://github.com/KiddLab/fastCN.

93. Hach, F., Hormozdiari, F., Alkan, C., Hormozdiari, F., Birol, I., Eichler, E.E., and Sahinalp, S.C. (2010). mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat. Methods *7*, 576–577. https://doi.org/10.1038/nmeth0810-576.

94. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580. https://doi.org/10.1093/nar/27.2.573.

95. Morgulis, A., Gertz, E.M., Schäffer, A.A., and Agarwala, R. (2006). WindowMasker: window-based masker for sequenced genomes. Bioinformatics *22*, 134–141. https://doi.org/10.1093/bioinformatics/bti774.

96. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92. https://doi.org/10.4161/fly.19695.

97. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238. https://doi.org/10.1186/s13059-019-1832-y.

98. 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., and de Hoon, M.J.L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.

99. pyCirclize, Moshi. Circular visualization in Python (Circos Plot, Chord Diagram, Radar Chart) (GitHub). https://github.com/moshi4/pyCirclize.

100. Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. *9*, 90–95. https://doi.org/10.1109/mcse.2007.55.

101. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinf. *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

102. 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. https://doi.org/10.1093/molbev/mst010.

103. Li, H. (2023). Protein-to-genome alignment with miniprot. Bioinformatics *39*, btad014. https://doi.org/10.1093/bioinformatics/btad014.

104. Li, H., Bloom, J.M., Farjoun, Y., Fleharty, M., Gauthier, L., Neale, B., and MacArthur, D. (2018). A synthetic-diploid benchmark for accurate variant-calling evaluation. Nat. Methods *15*, 595–597. https://doi.org/10.1038/s41592-018-0054-7.

105. Martin, M., Patterson, M., Garg, S., Fischer, S.O., Pisanti, N., Klau, G.W., Schöenhuth, A., and Marschall, T. (2016). WhatsHap: fast and accurate read-based phasing. Preprint at bioRxiv. https://doi.org/10.1101/085050.

106. 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*, msab358. https://doi.org/10.1093/molbev/msab358.

107. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

108. Tegenfeldt, F., Kuznetsov, D., Manni, M., Berkeley, M., Zdobnov, E.M., and Kriventseva, E.V. (2025). OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. Nucleic Acids Res. *53*, D516–D522. https://doi.org/10.1093/nar/gkae987.

109. YiJin-Xiong, J., Bot, N.-C., and Fellows Yates, J.A. (2025). nf-core/methylong: (Zenodo) https://doi.org/10.5281/ZENODO.17666904.

110. 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. https://doi.org/10.2527/jas2016.94supplement4184x.

111. Okamura, Y. transanno: accurate LiftOver tool for new genome assemblies (GitHub). https://github.com/informationsea/transanno.

112. 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. https://doi.org/10.1093/molbev/msy159.

113. Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. *3*, 418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410.

114. 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. https://doi.org/10.1101/gr.278295.123.

Please cite this article in press as: Abuelanin et al., Single-library chromosome-scale diploid assemblies of vole genomes resolve a species-specific duplication implicated in pair bonding, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101336

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Prairie vole (*Microtus ochrogaster*) liver tissue, adult male, UCDWGSCIFI-01	This paper	BioSample: SAMEA122714324
Meadow vole (*Microtus pennsylvanicus*) liver tissue, adult male, UCDWGSCIFI-03	This paper	BioSample: SAMEA122714325
Prairie vole (*Microtus ochrogaster*) brain tissue for RNA-seq, 2017 sample	This paper	BioSample: SAMEA122715132
Prairie vole (*Microtus ochrogaster*) liver tissue for RNA-seq, 2017 sample	This paper	BioSample: SAMEA122715133
Prairie vole (*Microtus ochrogaster*) testes tissue for RNA-seq, 2017 sample	This paper	BioSample: SAMEA122715134
Prairie vole (*Microtus ochrogaster*) e11.5 embryo tissue for RNA-seq, 2017 sample	This paper	BioSample: SAMEA122715135
Meadow vole (*Microtus pennsylvanicus*) brain tissue for RNA-seq, 2017 sample	This paper	BioSample: SAMEA122715136
Meadow vole (*Microtus pennsylvanicus*) liver tissue for RNA-seq, 2017 sample	This paper	BioSample: SAMEA122715137
Meadow vole (*Microtus pennsylvanicus*) testes tissue for RNA-seq, 2017 sample	This paper	BioSample: SAMEA122715138
Meadow vole (*Microtus pennsylvanicus*) e13.5 embryo tissue for RNA-seq, 2017 sample	This paper	BioSample: SAMEA122715139
Woodland vole/pine vole (*Microtus pinetorum*) dried museum pelt tissue	This paper	BioSample: SAMEA122737214
Chemicals, peptides, and recombinant proteins		
UltraPure Phenol:Chloroform:Isoamyl Alcohol (25:24:1, v/v)	Thermo Fisher Scientific	Cat#15593031
Critical commercial assays		
Monarch HMW DNA Extraction Kit for Tissue	New England Biolabs (NEB)	Cat# T3050
SMRTbell Prep Kit 3.0	Pacific Biosciences	Cat# 102-182-700
HindIII Restriction Enzyme	New England Biolabs (NEB)	Cat# R0104
Qubit dsDNA High Sensitivity Assay Kit	Thermo Fisher Scientific	Cat# Q33231
AMPure PB bead size selection kit	Pacific Biosciences	Cat# 102-182-500
KOD Xtreme Hot Start DNA polymerase	Millipore Sigma	Cat# 71975-3
Revio SPRQ Polymerase Kit	Pacific Biosciences	Cat# 103-520-100
Revio SMRT Cell Tray	Pacific Biosciences	Cat# 102-202-200
Deposited data		
Vole sequencing and assembly project generated in this study	This paper	BioProject: PRJEB108798
Prairie vole genome assembly mMicOch_UCD_v1 haplotype 1	This paper	ENA/GenBank: ERZ29502723
Prairie vole genome assembly mMicOch_UCD_v1 haplotype 2	This paper	ENA/GenBank: ERZ29502724
Meadow vole genome assembly mMicPen_UCD_v1 haplotype 1	This paper	ENA/GenBank: ERZ29502725

(*Continued on next page*)

Please cite this article in press as: Abuelanin et al., Single-library chromosome-scale diploid assemblies of vole genomes resolve a species-specific duplication implicated in pair bonding, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101336

CellPress
OPEN ACCESS

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Meadow vole genome assembly mMicPen_UCD_v1 haplotype 2	This paper	ENA/GenBank: ERZ29502726
Prairie vole PacBio HiFi reads	This paper	ENA/GenBank: ERX16712379
Prairie vole PacBio CiFi HindIII reads	This paper	ENA/GenBank: ERX16712347
Prairie vole Hi-C reads	This paper	ENA/GenBank: ERX16712351
Prairie vole Illumina whole-genome sequencing reads	This paper	ENA/GenBank: ERX16712349
Meadow vole PacBio HiFi reads	This paper	ENA/GenBank: ERX16712352
Meadow vole PacBio CiFi HindIII reads	This paper	ENA/GenBank: ERX16712346
Meadow vole Hi-C reads	This paper	ENA/GenBank: ERX16712350
Meadow vole Illumina whole-genome sequencing reads	This paper	ENA/GenBank: ERX16712348
Prairie vole brain RNA-seq reads	This paper	ENA/GenBank: ERX16712717
Prairie vole liver RNA-seq reads	This paper	ENA/GenBank: ERX16712718
Prairie vole testes RNA-seq reads	This paper	ENA/GenBank: ERX16712719
Prairie vole E11 embryo RNA-seq reads	This paper	ENA/GenBank: ERX16712720
Meadow vole brain RNA-seq reads	This paper	ENA/GenBank: ERX16712721
Meadow vole liver RNA-seq reads	This paper	ENA/GenBank: ERX16712722
Meadow vole testes RNA-seq reads	This paper	ENA/GenBank: ERX16712723
Meadow vole E13 embryo RNA-seq reads	This paper	ENA/GenBank: ERX16712724
Prairie vole reference genome MicOch1.0	Broad Institute; NCBI Assembly	NCBI Assembly: GCF_000317375.1
Meadow vole reference genome mMicPen1	Vertebrate Genomes Project; NCBI Assembly	NCBI Assembly: GCF_037038515.1
Prairie vole brain RNA-seq dataset (paternal behaviors)	University of California, Davis	BioProject: PRJNA428754
Prairie vole brain RNA-seq dataset (comparative neurotranscriptomics)	Tripp et al.[53]	BioProject: PRJNA682808
Prairie vole brain RNA-seq dataset (pair-bond formation/maintenance)	Duclot et al.[54]	BioProject: PRJNA631040
Prairie vole dentate gyrus RNA-seq dataset	Duclot et al.[52]	BioProject: PRJNA786347
Prairie vole nucleus accumbens RNA-seq dataset (partner separation)	Waddell et al.[55]	BioProject: PRJNA887096
Prairie vole brain RNA-seq dataset	Sadino et al.[56]	BioProject: PRJNA792575
Prairie vole RNA-seq dataset (maternal oxytocin)	Danoff et al.[57]	BioProject: PRJNA1005323
Prairie vole RNA-seq dataset	EBI	BioProject: PRJEB89367
Montane vole (*Microtus montanus*) Illumina whole-genome sequencing reads	The Ohio State University	SRA: SRR12966109
Common vole (*Microtus arvalis*) Illumina whole-genome sequencing reads	Norwegian University of Life Sciences	SRA: ERR3427942
Field vole (*Microtus agrestis*) Illumina whole-genome sequencing reads	University of Liverpool	SRA: SRR2167807
North American water vole (*Microtus richardsoni*) Illumina whole-genome sequencing reads	The Ohio State University	SRA: SRR12963053
NA12878/HG001 PacBio HiFi reads used for human phasing benchmark	Kronenberg et al.[24]	SRA: SRX5780566
GM12878 CiFi reads used for human phasing benchmark	McGinty et al.[7]	BioProject: PRJEB83708
NA12878 Hi-C reads used for human phasing benchmark	Rao et al.[73]	SRA: SRR1658583, SRR1658588, and SRR1658596

(*Continued on next page*)

Please cite this article in press as: Abuelanin et al., Single-library chromosome-scale diploid assemblies of vole genomes resolve a species-specific duplication implicated in pair bonding, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101336

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GRCh38 human reference genome used for human phasing benchmark	Genome Reference Consortium	NCBI Assembly: GCA_000001405.15
Experimental models: Organisms/strains		
Prairie vole (*Microtus ochrogaster*): adult male	This paper	RRID:NCBITaxon_79684
Meadow vole (*Microtus pennsylvanicus*): adult male	This paper	RRID:NCBITaxon_159058
Woodland vole/pine vole (*Microtus pinetorum*): dried museum pelt tissue	This paper	RRID:NCBITaxon_267028
Software and algorithms		
lima v2.14.0	Pacific Biosciences	https://github.com/PacificBiosciences/barcoding
Pygenomeviz v1.6.1	Moshi et al.[74]	https://github.com/moshi4/pyGenomeViz
Mummer 4.0.1	Marçais et al.[75]	https://github.com/mummer4/mummer
pbmarkdup v1.1.0	Pacific Biosciences	https://github.com/PacificBiosciences/pbmarkdup
samtools v1.21	Bonfield et al.[76]	https://www.htslib.org/
cifi-toolkit	This paper	https://github.com/mydennislab/cifi-toolkit
hifiasm v0.19.8 and v0.25.0	Cheng et al.[4]	https://github.com/chhylp123/hifiasm
gfatools v0.5	Li et al.[77]	https://github.com/lh3/gfatools
minimap2 v2.26 and v2.30	Li et al.[78]	https://github.com/lh3/minimap2
paftools.js	Li et al.[78]	https://github.com/lh3/minimap2
epi2me-labs/wf-pore-c v1.3.0	Griffiths et al.[79]	https://github.com/epi2me-labs/wf-pore-c
YaHS v1.2a.2	Zhou et al.[23]	https://github.com/c-zhou/yahs
PretextView	Wellcome Sanger Tree of Life	https://github.com/sanger-tol/PretextView
PretextMap	Wellcome Sanger Tree of Life	https://github.com/sanger-tol/PretextMap
curationpretext	Wellcome Sanger Tree of Life	https://github.com/sanger-tol/curationpretext
pretext-to-asm/agp-tpf-utils	Wellcome Sanger Tree of Life	https://github.com/sanger-tol/agp-tpf-utils
seqtk	Li[80]	https://github.com/lh3/seqtk
tidk v0.2	Brown et al.[81]	https://github.com/tolkit/telomeric-identifier
Meryl v1.3	Rhie et al.[25]	https://github.com/marbl/meryl
GenomeScope2 v2.1.0	Ranallo-Benavidez et al.[82]	https://github.com/tbenavi1/genomescope2.0
Merqury v1.3	Rhie et al.[25]	https://github.com/marbl/merqury
Yak	Li[83]	https://github.com/lh3/yak
centroAnno v1.0.2	Qi et al.[39]	https://github.com/CL-CHEN-Lab/CentroAnno
nf-core/methylong v2.0.0	Ewels et al.[84]	https://nf-co.re/methylong
mosdepth	Pedersen et al.[85]	https://github.com/brentp/mosdepth
bedGraphToBigWig	Casper et al.[86]	https://hgdownload.soe.ucsc.edu/admin/exe/
RepeatModeler v2.0.7	Flynn et al.[87]	https://github.com/Dfam-consortium/RepeatModeler
RepeatMasker v4.2.1	Tarailo-Graovac et al.[88]	https://www.repeatmasker.org/
Dfam TE Tools container v1.95	Dfam Consortium	https://hub.docker.com/r/dfam/tetools
BISER v1.4	Išerić et al.[31]	https://github.com/0xTCG/biser
EGAPx v0.4.1-alpha	Strope et al.[89]	https://github.com/ncbi/egapx
Salmon v1.10.3	Patro et al.[90]	https://combine-lab.github.io/salmon/
gffread	Pertea et al.[91]	https://github.com/gpertea/gffread

(*Continued on next page*)

Please cite this article in press as: Abuelanin et al., Single-library chromosome-scale diploid assemblies of vole genomes resolve a species-specific duplication implicated in pair bonding, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101336

CellPress
OPEN ACCESS

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
FastCN	Kidd and Shen[92]	https://github.com/KiddLab/fastCN
mrsFAST v3.4.2	Hach et al.[93]	https://github.com/sfu-compbio/mrsfast
Tandem Repeats Finder 4.09	Benson et al.[94]	https://tandem.bu.edu/trf/trf.html
WindowMasker/DUST	Morgulis et al.[95]	NCBI C++ Toolkit
Ensembl Variant Effect Predictor (VEP) v114.1	McLaren et al.[37]	https://github.com/Ensembl/ensembl-vep
SnpEff v5.4c	Cingolani et al.[96]	https://pcingola.github.io/SnpEff/
bcftools v1.22	Bonfield et al.[76]	https://samtools.github.io/bcftools/
OrthoFinder v2.5.5	Emms et al.[97]	https://github.com/davidemms/OrthoFinder
Biopython v1.86	Cock et al.[98]	https://biopython.org/
ODP v0.3.3	Schultz et al.[34]	https://github.com/conchoecia/odp
pyCirclize v1.9.1	Moshi[99]	https://github.com/moshi4/pyCirclize
matplotlib v3.10.8	Hunter et al.[100]	https://matplotlib.org/
BLAST+	Camacho et al.[101]	https://blast.ncbi.nlm.nih.gov/
MAFFT v7	Katoh et al.[102]	https://mafft.cbrc.jp/alignment/software/
miniprot 0.18-r281	Li et al.[103]	https://github.com/lh3/miniprot
dipcall v0.3	Li et al.[104]	https://github.com/lh3/dipcall
WhatsHap v2.7	Martin et al.[105]	https://whatshap.readthedocs.io/
Vole assembly analysis source code and workflows	This paper	https://github.com/mydennislab/2026-voles-assembly
Other		
PacBio Revio sequencing system	Pacific Biosciences	N/A
Illumina NextSeq 500 sequencing system	Illumina	N/A
Illumina NovaSeq 6000 sequencing system	Illumina	N/A
Hamilton NGS STAR MOA system	Hamilton	N/A
Pippin HT pulsed-field size-selection system	Sage Science	N/A
Femto Pulse system	Agilent	N/A
Qubit Fluorometer	Thermo Fisher Scientific	N/A

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Vole procedures

The male 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 cutoff. 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 1× 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-h 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)[106] was used to prepare the proximity ligation library and processed 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 gDNA 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[76] ('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/mydennislab/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[23] using the Pore-C 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.

NA12878 CiFi haplotype phasing benchmark

Because no phased truth set exists for the vole species sequenced in this study, phasing accuracy was benchmarked on the human GIAB HG001/NA12878 sample. Public PacBio HiFi reads from NA12878[24] were paired with CiFi reads (prepared using HindIII) from the matched GM12878 lymphoblastoid line.[7] HiFi reads were downsampled to 30× on GRCh38, and CiFi reads were downsampled to 0, 2, 3.5, 5, 7.5, 10, and 15× and also evaluated at full depth. A matched Hi-C analysis used a public GM12878 Hi-C dataset (prepared with MboI)[73] at the same coverage (2–30×). CiFi reads were converted to paired-end concatemer segments by in silico digestion ('cifi digest'). For each condition, hifiasm v0.25.0 was run on the 30× HiFi reads with paired-end CiFi or Hi-C reads supplied via '–h1'/'–h2' ('–dual-scaf –telo-m CCCTAAA'); the 0×-CiFi baseline used hifiasm on HiFi alone. Each phased haplotype was aligned to GRCh38 with dipcall v0.3, and 'whatshap compare' (v2.7) was run on the dipcall × GIAB confident intersection (HG001 v4.2.1, autosomes chr1–chr22) to compute per-chromosome switch error and blockwise Hamming error rates and against the pedigree-phased NA12878 call set (Platinum Pedigree[24]).

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'. Chromatin contact 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 was 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. AGP files of the corrected contact maps were exported from PretextView and curated haplotype assemblies generated using pretext-to-asm.

Completeness and contiguity 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. Telomeric repeat sequences were detected using tidk v0.2 (Telomere Identification Toolkit;[81]), 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 T2T scaffolds when both telomere ends were detected but with sequence gaps were present. A fully T2T assembly comprised both telomere ends and a single contig with no gaps. Genome completeness of each assembly was also assessed with BUSCO v6.0.0[107] in genome mode against the glires_odb12 lineage (12,556 orthologs).[108]

Heterozygosity and consensus quality estimation

Per-individual heterozygosity and consensus quality (QV) were estimated from Illumina paired-end whole-genome sequencing of gDNA prepared from the same individuals used for HiFi sequencing. Per-species k-mer databases were built with Meryl v1.3[25] using k = 21. Heterozygosity was estimated by extracting the k-mer frequency histograms from the Meryl databases (meryl histogram) and fitting a diploid mixture model with GenomeScope2 v2.1.0[82] and k-mer completeness were computed with Merqury v1.3 using k = 31, with diploid values represented in Table 1. Haplotype-specific analyses provided similar QV (prairie vole 51.9 (Hap1) and 52.0 (Hap2); meadow vole 50.0 (Hap 1) and 52.1 (Hap2)) and k-mer completeness estimates (prairie vole 83.3% (Hap1) and 78.6% (Hap2); meadow vole 81.2% (Hap 1) and 76.6% (Hap2)).

Centromere annotation

Centromeric regions were identified on each Hap1 primary assembly using centroAnno v1.0.2[39] 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. The sequence comprising the chr26 centromere using minimap2 (default parameters) matched 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 chromosome 26 *Avpr1a* locus in both prairie vole haplotypes.

HiFi read mapping and methylation analysis

PacBio Revio HiFi reads carrying per-base 5mC calls (MM/ML BAM tags from the Jasmine primary-analysis basecaller) were processed with nf-core/methylong v2.0.0[109] against the prairie and meadow vole references. Per-CpG bedMethyl outputs were converted to per-site and 1-kbp-windowed bigWig tracks with UCSC bedGraphToBigWig. Region-scale methylation segmentation into high- and low-methylated blocks was performed with MethBat as part of the nf-core/methylong workflow and rendered as a bigBed track; these segments classify each block as methylated or unmethylated within a sample. Coverage tracks were generated with mosdepth[85] (1-kbp windows) on the HiFi methylation BAMs and on CiFi alignments produced by minimap2 v2.30 (map-hifi preset).

Segmental duplication and repeat annotations

Species-specific *de novo* repeat libraries were constructed for each vole species using RepeatModeler v2.0.7[87] 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 v4.2.1[88] with the RMBLAST search engine. RepeatMasker was executed via the Dfam TE Tools docker container; RepeatModeler v2.0.7, RepeatMasker v4.2.1, RMBLAST v2.14.1+). SDs were identified using BISER v1.4[31] with default parameters. BISER was run on the soft-masked Hap1 assemblies for each species, and output was generated in BEDPE format.

Dot plots were generated using ODP v0.3.3,[34] circular genome visualizations using pyCirclize v1.9.1, and linear karyotype ideograms using matplotlib v3.10.8.

RNA-seq 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. 50–75 mg 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 1-mL 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[110] to perform gene annotation for both Hap1 assemblies of both species.

Gene expression analysis

Transcript quantification was performed using Salmon v1.10.3[90] 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 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[52–57] 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.

Scaffold orientation

To standardize chromosome naming, chromosome-scale scaffolds in the prairie Hap1 assemblies were renamed from SUPER_1–SUPER_26, SUPER_X, and SUPER_Y to chr1–chr26, chrX, and chrY, respectively. Scaffold numbering was retained from the *de novo* assembly. Scaffold orientation was then evaluated relative to the MicOch1.0 prairie vole reference assembly (GCF_000317375.1). Whole-genome alignments were generated with minimap2 v2.30 using -cx asm5 –cs in both query-to-reference and reference-to-query directions, and alignment blocks shorter than 10 kbp were excluded. For each scaffold, orientation was inferred from the dominant aligned strand based on aligned bases; scaffolds with >50% of aligned bases on the reverse strand were reverse-complemented. Orientation confidence was classified as high when $\geq$95% of aligned bases supported the dominant strand, medium when 80–95% supported the dominant strand, and low/mixed when <80% supported the dominant strand. Because MicOch1.0 was generated from a female individual and lacks chrY, the prairie chrY scaffold was retained in its native orientation.

Prairie scaffold orientations were cross-validated using the published prairie vole genetic linkage map.[27] Marker flanking sequences were aligned to the assembly with BLASTn using an e-value threshold of 1×10^{-10}, retaining the single best hit per marker. For each linkage group, the Spearman rank correlation between genetic position in centimorgans and physical position in the assembly was calculated. Linkage-map evidence was used to override the alignment-based orientation only when the alignment call had low/mixed confidence, $|\rho| \geq 0.70$, and $\geq$5 markers were mapped to the scaffold. Discordant cases are reported in Table S6.

The corrected prairie Hap1 assembly was generated by reverse-complementing scaffolds designated by this procedure. Corrected scaffolds were then re-aligned to MicOch1.0 to record post-correction forward-strand percentages. Scaffolds with persistently low forward-strand alignment after correction were flagged in Tables S6 and S11, consistent with mosaic content or internal inversions that cannot be resolved by a single scaffold-level reverse-complement. A UCSC liftOver chain was generated to convert coordinates between the original and orientation-corrected assemblies.

To facilitate cross-species comparative analysis, meadow vole Hap1 scaffolds were oriented against the orientation-corrected prairie vole Hap1 assembly using the same alignment-based procedure, with minimap2 run using the cross-species preset -cx asm10. Meadow chrY had no reliable cross-species alignment and was retained in its native orientation. Minor meadow scaffolds were oriented when sufficient cross-species alignment evidence was available; scaffolds without informative alignments were retained in their native orientation. Per-scaffold orientation calls, alignment statistics, linkage-map validation results, and coordinate-conversion resources are provided in Tables S6, S7, S11, and S12.

To enable coordinate conversion from our meadow vole Hap1 assembly to the published VGP meadow vole reference (mMicPen1.hap1; GCF_037038515.1), we generated a UCSC liftOver chain file by aligning the two meadow vole assemblies. Our Hap1 assembly scaffolds were aligned to mMicPen1.hap1 using minimap2 v2.30 ('-cx asm5 –cs'). The resulting alignment was converted to UCSC chain format with transanno v0.4.5 'minimap2chain'[111] and sorted with UCSC 'chainSort'.[86] The liftOver chain file is provided as Supplementary Data and is compatible with UCSC 'liftOver' and related coordinate-conversion tools.

AVPR1A amino acid alignment

Coding sequences for the prairie and meadow vole *Avpr1a* homologs were compared across both haplotypes: six sequences in total, comprising prairie *Avpr1a* and *Avpr1a2* and meadow *Avpr1a*, each from Hap1 and Hap2. Coding sequences were taken from the EGAPx gene annotations[89] of each assembly; because the Hap2 assemblies were unannotated, the hap1 *Avpr1a* proteins were aligned to the Hap2 contigs with miniprot[103] and the coding sequences extracted from the resulting gene models. The six coding

sequences were aligned with MACSE v2,[112] a codon-aware aligner that accommodates the frameshift and internal stop codons of the *Avpr1a2* paralog and produces a corresponding amino acid alignment.

Comparative assembly and paralog analyses

One-to-one orthologs between meadow and prairie voles were identified with OrthoFinder v2.5.5[97] from the species' EGAPx proteomes. The meadow vole Hap1 assembly was aligned to the prairie vole Hap1 reference with minimap2 v2.30[78] (asm10 preset), and variants were called with paftools.js (alignment blocks $\geq$10 kb, MAPQ $\geq$30); calls were restricted to autosomal coding sequence and normalized with bcftools v1.22.[76] Coding variants were annotated independently with SnpEff v5.4c[96] and Ensembl VEP v114.1,[37] and were retained as candidate gene-disrupting if any consequence carried IMPACT = HIGH and matched the high-impact Sequence Ontology set (stop_gained, frameshift_variant, splice_acceptor_variant, splice_donor_variant, start_lost, stop_lost, transcript_ablation, exon_loss_variant). Indels within 1 bp of a reference mononucleotide run of $\geq$6 bp were flagged as homopolymer artifacts and excluded from the per-gene evidence count; candidates were further restricted to 1:1 ortholog groups with a canonical gene symbol, each retained group required at least two distinct non-homopolymer HIGH-impact variants, and genes called by both annotators defined the final candidate-LoF set. Pairwise meadow-prairie Ka/Ks was computed in BioPython v1.86[98] for every 1:1 ortholog carrying at least one coding SNV by applying the VEP-annotated substitutions to the prairie reference coding sequence and scoring each gene by the Nei-Gojobori (1986) method.[113]

Pairwise comparisons and visualization of homologous chromosomes, as well as SD sequence containing *Avpr1a* (prairie vole Hap1 chr26:13,970,803-14,087,830) and *Avpr1a2* (prairie vole Hap1 chr26:13,058,418-13,176,477) were performed using MUMmer4 alignments[75] and visualized with pyGenomeViz v1.6.1.[74] The annotated centromeric sequence (prairie vole Hap1 chr26:13,182,318-13,728,901) 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 for chromosomes Y (both prairie and meadow voles) and 23 (prairie vole) were generated using nf-core/pairgenomealign.[114]

Copy number analysis

Copy number (CN) was estimated using the FastCN pipeline[40] with mrsFAST v3.4.2. All species were mapped to the prairie vole Hap1 primary 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[94] per chromosome, and low-complexity regions were masked with WindowMasker/DUST.[95] 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 SDs (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-kbp windows across the genome. CN was calculated as CN = 2 x (window depth/autosomal control mean). To prevent inflated CN 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-kbp windows overlapping each target locus. CN evaluated for the putative centromere (prairie vole Hap 1, chr26:13,182,318-13,728,901), where estimates tend to be unreliable due to high repetitive content producing repeat masking, showed elevated CN in prairie vole across the chr26 centromere annotation relative to all other voles, suggesting this region is prairie-vole specific. CN estimation of the *Avpr1a* 118-kbp SD region (prairie vole Hap 1, chr26:13,970,803-14,087,830) confirmed the duplication to be prairie-vole specific with diploid CN of 4 compared to CN of 2 for all other tested vole genomes.

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantification procedures are described in the corresponding Method details sections. Analyses were primarily descriptive and are reported as counts, proportions, medians, or genome-wide/interval-level summary metrics unless otherwise stated. No additional formal statistical hypothesis testing was performed beyond the quantitative metrics described in the STAR methods. For RNA-seq boxplots, center lines indicate medians, boxes indicate interquartile ranges, whiskers indicate 1.5$\times$ the interquartile range, and points indicate individual samples.

ADDITIONAL RESOURCES

- UCSC Genome Browser hub for prairie vole: https://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=mabuelanin&hgS_otherUserSessionName=prairie_vole_cifi_hifi_hap1
- UCSC Genome Browser hub for meadow vole: https://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=mabuelanin&hgS_otherUserSessionName=meadow_vole_cifi_hifi_hap1
- General information and resources related to this project: dennislab.org/voles