The Evolutionary Fates of a Large Segmental Duplication in Mouse

Andrew P. Morgan,* J. Matthew Holt,† Rachel C. McMullan,* Timothy A. Bell,* Amelia M.-F. Clayshulte,* John P. Didion,* Liran Yadgary,* David Thybert,‡,1 Duncan T. Odom,§, Paul Flicek,‡,** Leonard McMillan,† and Fernando Pardo-Manuel de Villena*,2**

*Department of Genetics and Lineberger Comprehensive Cancer Center, and †Department of Computer Science, University of North Carolina, Chapel Hill, North Carolina 27599, §Cancer Research United Kingdom Cambridge Institute, University of Cambridge, CB2 0RE, United Kingdom, ‡European Bioinformatics Institute, European Molecular Biology Laboratory, and **Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, CB10 1SA, United Kingdom

ORCID ID: 0000-0003-1942-4543 (A.P.M.)

ABSTRACT Gene duplication and loss are major sources of genetic polymorphism in populations, and are important forces shaping the evolution of genome content and organization. We have reconstructed the origin and history of a 127-kbp segmental duplication, *R2d*, in the house mouse (*Mus musculus*). *R2d* contains a single protein-coding gene, *Cwc22*. *De novo* assembly of both the ancestral (*R2d1*) and the derived (*R2d2*) copies reveals that they have been subject to nonallelic gene conversion events spanning tens of kilobases. *R2d2* is also a hotspot for structural variation: its diploid copy number ranges from zero in the mouse reference genome to >80 in wild mice sampled from around the globe. Hemizygosity for high copy-number alleles of *R2d2* is associated in *cis* with meiotic drive; suppression of meiotic crossovers; and copy-number instability, with a mutation rate in excess of 1 per 100 transmissions in some laboratory populations. Our results provide a striking example of allelic diversity generated by duplication and demonstrate the value of *de novo* assembly in a phylogenetic context for understanding the mutational processes affecting duplicate genes.

KEYWORDS copy-number variation; meiotic drive; segmental duplications

DUPLICATION is an important force shaping the evolution of plant and animal genomes: it provides a substrate for evolution. Redundancy transiently frees the duplicates from selective constrant (Lynch and Conery 2000). Segmental duplications (SDs), defined as contiguous sequences which map to more than one physical position (Bailey and Eichler 2006), are a common feature of eukaryotic genomes and particularly those of vertebrates.

Like any sequence variant, a duplication first arises in a single individual in a population. The distinction between such copy-number variants (CNVs) and SDs is fluid and somewhat arbitrary: tracts of SDs are highly polymorphic in populations in species from *Drosophila* (Dopman and Hartl 2007) to mouse (She *et al.* 2008) to human (Bailey and Eichler 2006). Studies of parent-offspring transmissions have shown that SDs are prone to recurrent *de novo* mutations including some implicated in human disease (reviewed in Stankiewicz and Lupski 2002). Bursts of SD have preceded dramatic species radiations in primates and, more broadly, blocks of conserved synteny in mammals frequently terminate at SDs (Bailey and Eichler 2006). This suggests that SDs could mediate the chromosomal rearrangements through which karyotypes diverge and reproductive barriers arise.

Notwithstanding their evolutionary importance, SDs are difficult to analyze. Repeated sequences with high pairwise similarity are likely to be collapsed into a single sequence during genome assembly. Efficient and sensitive alignment of high-throughput sequencing reads to duplicated sequence remains challenging (Treangen and Salzberg 2011). Genotyping of sites within SDs is difficult because variants between copies (paralogous variants) are easily confounded with variants within

doi: 10.1534/genetics.116.191007
Manuscript received April 29, 2016; accepted for publication June 27, 2016; published Early Online June 30, 2016.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.191007/-/DC1.
[1]Present address: Earlham Institute, Norwich Research Park, Norwich NR4 7UH, United Kingdom.
[2]Corresponding author: 5049 Genetic Medicine Bldg., Department of Genetics, University of North Carolina, 120 Mason Farm Rd., CB#7264, Chapel Hill, NC 27599-7264. E-mail: fernando@med.unc.edu

copies between individuals at a given copy (allelic variants). Latent paralogous variation may bias interpretations of sequence diversity and haplotype structure (Hurles 2002).

Paralogy also complicates phylogenetic inference. Ancestral duplication followed by differential losses along separate lineages may yield a local phylogeny that is discordant with the genome-wide phylogeny (Goodman *et al.* 1979). Within each duplicate copy, local phylogenies for adjacent intervals may also be discordant due to nonallelic gene conversion between copies (Dover 1982; Nagylaki and Petes 1982).

In this manuscript we present a detailed analysis of *R2d*, a segmental duplication in the house mouse (*Mus musculus*). *R2d* is a 127-kbp unit which contains the protein-coding gene *Cwc22* and flanking intergenic sequence. Although the C57BL/6J reference strain and other classical laboratory strains have a single haploid copy of the *R2d* sequence (in the *R2d1* locus), the wild-derived CAST/EiJ, ZALENDE/EiJ, and WSB/EiJ strains have an additional 1, 16, and 33 haploid copies, respectively, in the *R2d2* locus. *R2d2* is the responder locus in a recently-described meiotic drive system on mouse chromosome *2* (chr2) but is absent from the mouse reference genome (Waterston *et al.* 2002; Didion *et al.* 2015, 2016). We draw on a collection of species from the genus *Mus* sampled from around the globe to reconstruct the sequence of events giving rise to the locus' present structure (Figure 1). Using novel computational tools built around indexes of raw high-throughput sequencing reads, we perform local *de novo* assembly of phased haplotypes and explore patterns of sequence diversity within and between copies of *R2d*.

Both phylogenetic analyses and estimation of mutation rate in laboratory mouse populations reveal that *R2d2* and its surrounding region on chr2 are unstable in copy number. Cycles of duplication, deletion, and nonallelic gene conversion have led to complex phylogenetic patterns discordant with species-level relationships within *Mus*, which cannot be explained by known patterns of introgression between *Mus* species (Bonhomme *et al.* 2007; Yang *et al.* 2011).

Materials and Methods

Mice

Wild *M. musculus* mice used in this study were trapped at a large number of sites across Europe, the United States, the Middle East, northern India, and Taiwan. Trapping was carried out in accordance with local regulations and with the approval of all relevant regulatory bodies for each locality and institution. Trapping locations are listed in Supplemental Material, Table S1. Most samples have been previously published (Didion *et al.* 2016).

Tissue samples from the progenitors of the wild-derived inbred strains ZALENDE/EiJ (*M. m. domesticus*), TIRANO/EiJ (*M. m. domesticus*), and SPRET/EiJ (*M. spretus*) were provided by Muriel Davisson, as described in Didion *et al.* (2016).

Tissue samples from the high running (HR) selection and intercross lines were obtained as described in Didion *et al.* (2016).

Female Diversity Outbred (DO) mice used for estimating mutation rates at *R2d2* were obtained from the Jackson Laboratory and housed with a single FVB/NJ male. Progeny were killed at birth by cervical dislocation in order to obtain tissue for genotyping.

All live laboratory mice were handled in accordance with the Institutional Animal Care and Use Committees protocols of the University of North Carolina at Chapel Hill.

DNA preparation

High molecular weight DNA: High molecular weight DNA was obtained for samples genotyped with the Mouse Diversity Array (MDA) or subject to whole-genome sequencing. Genomic DNA was extracted from tail, liver, or spleen using a standard phenol-chloroform procedure (Sambrook and Russell 2006). High molecular weight DNA for most inbred strains was obtained from the Jackson Laboratory, and the remainder as a generous gift from Francois Bonhomme and the University of Montpellier Wild Mouse Genetic Repository.

Low molecular weight DNA: Low molecular weight DNA was obtained for samples to be genotyped on the Mega Mouse Universal Genotyping Array (MegaMUGA) (see *Microarray genotyping* below). Genomic DNA was isolated from tail, liver, muscle, or spleen using QIAGEN (Valencia, CA) Gentra Puregene or DNeasy Blood and Tissue Kits according to the manufacturer's instructions.

Whole-genome sequencing and variant discovery

Inbred strains: Sequencing data for inbred strains of mice except ZALENDE/EiJ and LEWES/EiJ were obtained from the Sanger Mouse Genomes Project web site (ftp://ftp-mouse.sanger.ac.uk/current_bams) as aligned BAM files. Details of the sequencing pipeline are given in Keane *et al.* (2011). Coverage ranged from ~25 to 50× per sample.

The strains LEWES/EiJ and ZALENDE/EiJ were sequenced at the University of North Carolina High-Throughput Sequencing Facility. Libraries were prepared from high molecular weight DNA using the Illumina (Carlsbad, CA) TruSeq kit and insert size ~250 bp, and 2× 100-bp paired-end reads were generated on an Illumina HiSequation 2000 instrument. LEWES/EiJ was sequenced to ~12× coverage and ZALENDE/EiJ to ~20×. Alignment was performed as in Keane *et al.* (2011).

Wild mice: Whole-genome sequencing data from 26 wild *M. m. domesticus* individuals described in Pezer *et al.* (2015) was downloaded from the European Nucleotide Archive (ENA) under accession number PRJEB9450. Coverage ranged from ~12 to 20× per sample. An additional two wild *M. m. domesticus* individuals, IT175 and ES446, were sequenced at the University of North Carolina to approximate coverage of 8× each. Raw reads from an additional 10 wild *M. m. castaneus* described in Halligan *et al.* (2013), sequenced to ~20× each, were downloaded from ENA under accession number PRJEB2176. Reads for a single *M. caroli* individual sequenced to ~40× were obtained from ENA under accession number

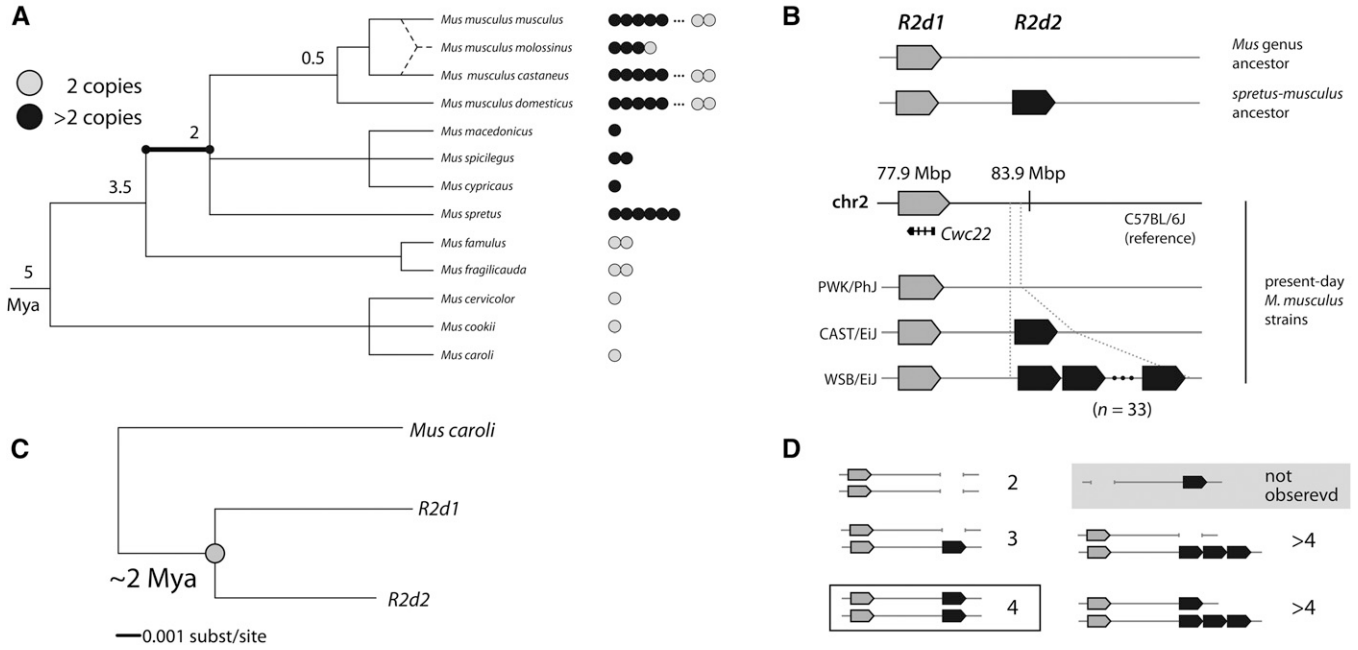


Figure 1 Origin and age of the *R2d2* duplication. (A) *R2d* copy number across the phylogeny of the genus *Mus*. Each dot represents one individual; shaded dots indicate a diploid copy number of two and solid dots a copy number of more than two. The duplication event giving rise to *R2d1* and *R2d2* most likely occurred on the highlighted branch. Approximate divergence times (Suzuki *et al.* 2004) are given in MYA at internal nodes. (B) Schematic structure of the *R2d1-R2d2* locus. The mouse reference genome (strain C57BL/6J, *M. m. domesticus*) contains a single copy of *R2d* at *R2d1*. Wild-derived inbred strains vary in haploid copy number from 1 (PWK/PhJ, *M. m. musculus*) to 2 (CAST/EiJ, *M. m. castaneus*) to 33 (WSB/EiJ, *M. m. domesticus*). *R2d1* is located at ~77.9 Mb and *R2d2* at 83.8 Mb. (C) Concatenated tree constructed from *R2d1* (reference genome) and *de novo* assembled *R2d2* and *M. caroli* sequences assuming a strict molecular clock. The duplication node is indicated by a shaded dot. (D) Relationship between observed *R2d* copy-number states and inferred structure of the *R2d1-R2d2* locus. The configuration of the *M. spretus–M. musculus* common ancestor (four diploid copies) is boxed. We have yet to identify samples with a diploid copy number of two but no *R2d1* (gray shaded box).

PRJEB2188. Reads for each sample were realigned to the mm10 reference using BWA-MEM v0.7.12 with default parameters (Li 2013). Optical duplicates were removed with samblaster (Faust and Hall 2014).

Variant discovery: Polymorphic sites on chr2 in the vicinity of *R2d2* (Figure 4B) were called using Freebayes v0.9.21-19-gc003c1e (Garrison and Marth 2012) with parameters "--standard-filters" using the Sanger Mouse Genomes Project VCF files as a list of known sites (parameter "--@"). Raw calls were filtered to have quality score > 30, root mean square mapping quality > 20 (for both reference and alternate allele calls) and at most two alternate alleles.

Copy-number estimation

R2d copy number was estimated using quantitative PCR (qPCR) as described in Didion *et al.* (2016). Briefly, we used commercial TaqMan assays against intron-exon boundaries in *Cwc22* (Life Technologies assay numbers Mm00644079_cn and Mm00053048_cn) to determine copy number relative to reference genes *Tert* (catalogue number 4458368, for target Mm00644079_cn) or *Tfrc* (catalogue number 4458366, for target Mm00053048_cn). Cycle thresholds for *Cwc22* relative to the reference gene were normalized across assay batches using linear mixed models with batch and target-reference pair treated as random effects. Control samples with known haploid *R2d* copy numbers of 1 (C57BL/6J), 2 (CAST/EiJ), 17 (WSB/EiJxC57BL/6J)F_1, and 34 (WSB/EiJ) were included in each batch.

Samples were classified as having one, two, or more than two haploid copies of *R2d* using linear discriminant analysis. The classifier was trained on the normalized cycle thresholds of the control samples from each plate, whose precise integer copy number is known, and applied to the remaining samples.

Microarray genotyping

Genome-wide genotyping was performed using MegaMUGA (Neogen/GeneSeek, Lincoln, NE) (Morgan *et al.* 2015). Genotypes were called using the GenCall algorithm implemented in the Illumina BeadStudio software. For quality control we computed, for each marker i on the array: $S_i = X_i + Y_i$, where X_i and Y_i are the normalized hybridization intensities for the two alleles. The expected distribution of S_i was computed from a large set of reference samples. We excluded arrays for which the distribution of S_i was substantially shifted from this reference; in practice, failed arrays can be trivially identified in this manner (Morgan *et al.* 2015). Access to MegaMUGA genotypes was provided by partnership between the McMillan and Pardo-Manuel de Villena labs and the University of North Carolina Systems Genetics Core Facility.

Additional genotypes for inbred strains and wild mice from the MDA were obtained from Yang *et al.* (2011).

All microarray genotypes are provided in PLINK format in File S2.

De novo assembly of R2d2

Raw whole-genome sequencing reads for WSB/EiJ from the Sanger Mouse Genomes Project were converted to a multi-string Burrows–Wheeler transform (msBWT) and associated FM-index (Holt and McMillan 2014) using the msbwt v0.1.4 Python package (https://pypi.python.org/pypi/msbwt). The msBWT and FM-index implicitly represent a suffix array to provide efficient queries over arbitrarily large string sets. Given a seed k-mer present in that string set, this property can be exploited to rapidly construct a de Bruijn graph which can in turn be used for local *de novo* assembly of a target sequence (Figure S1A). The edges in that graph can be assigned a weight (corresponding to the number of reads containing the $k+1$-mer implied by the edge) which can be used to evaluate candidate paths when the graph branches (Figure S1B).

R2d2 was seeded with the 30-bp sequence (TCTAGAGCAT GAGCCTCATTTATCATGCCT) at the proximal boundary of *R2d1* in the GRCm38/mm10 reference genome. A single linear contig was assembled by "walking" through the local de Bruijn graph. Because WSB/EiJ has ~33 copies of *R2d2* and a single copy of *R2d1*, any branch point in the graph which represents a paralogous variant should having outgoing edges with weights differing by a factor of ~33. Furthermore, when two (or more) branch points occur within less than the length of a read, it should be possible to "phase" the underlying variants by following single reads through both branch points (Figure S8B). We used these heuristics to assemble the sequence of *R2d2* (corresponding to the higher-weight path through the graph) specifically.

After assembling a chunk of ~500 bp, the contig was checked for colinearity with the reference sequence (*R2d1*) using the BLAST-like alignment tool (BLAT) and ClustalW2 (using the European Molecular Biology Laboratory, European Bioinformatics Institute web server: http://www.ebi.ac.uk/Tools/msa/clustalw2/).

Repetitive elements such as retroviruses are refractory to assembly with our method. Upon traversing into a repetitive element, the total edge weight (total number of reads) and number of branch points (representing possible linear assembled sequences) in the graph become large. It was sometimes possible to assemble a fragment of a repetitive element at its junction with a unique sequence, but it was not possible to assemble unambiguously across the repeat. Such regions were marked with blocks of Ns, and assembly reseeded using a nearby k-mer from the reference sequence. The final contig is provided in FASTA format in File S1.

The final contig was checked against its source msBWT by confirming that each 30-mer in the contig which did not contain an N was supported by at least 60 reads. A total of 16 additional haplotypes in eight regions of *R2d* totaling 16.9 kbp (Table S5) were assembled in a similar fashion, using the WSB *R2d2* contig and the *R2d1* reference sequence as guides. Multiple sequence alignments from these regions are provided in File S1.

Sequence analysis of R2d2 contig

Pairwise alignment of R2d paralogs: The reference *R2d1* sequence and our *R2d2* contig were aligned using LASTZ v1.03.54 (http://www.bx.psu.edu/~rsharris/lastz/) with parameters "--step=10 --seed=match12 --notransition --exact=20 --notrim --identity=95."

Transposable element content: The *R2d2* contig was screened for transposable-element (TE) insertions using the RepeatMasker web server (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) with species set to "mouse" and default settings otherwise. As noted previously, we could not assemble full-length repeats, but the fragments we could assemble at junctions with unique sequence allowed identification of some candidate TEs to the family level. *R2d1*-specific TEs were defined as TEs annotated in the RepeatMasker track at the University of California, Santa Cruz (UCSC) Genome Browser with no evidence (no homologous sequence, and no Ns) at the corresponding position in the *R2d2* contig. Candidate *R2d2*-specific TEs were defined as gaps ≥100 bp in size in the alignment to *R2d1* for which the corresponding *R2d2* sequence was flagged by RepeatMasker.

Gene conversion tracts: To unambiguously define gene conversion events without confounding from paralogous sequences, we examined 15 wild *M. m. domesticus* samples and 37 laboratory strains with evidence of two diploid copies of *R2d*. We confirmed that these copies of *R2d* were located at *R2d1* by finding read pairs spanning the junction between *R2d1* and the neighboring sequence. Gene conversion tracts were delineated as clusters of derived alleles shared with *R2d2*. Using a pairwise of alignment of *R2d2* and *R2d1* we identified single-nucleotide variants between the two sequences, and queried those sites in aligned reads for *M. caroli*. If the *M. caroli* and *R2d1* shared an allele, we recorded the site as a derived allele informative for the presence of *R2d2*. We used the resulting list of 1411 informative sites to query aligned reads for the samples of interest and recorded, for each site and each sample, whether the derived allele (*R2d2*), ancestral allele (*R2d1*), or both alleles were present. Conversion tracts were then identified by manual inspection. Boundaries of conversion tracts were defined at approximately the midpoint between the first *R2d1*- (or *R2d2*-) specific variant and the last *R2d2*- (or *R2d1*-) specific variant.

Sequence diversity in R2d1 and R2d2: Assembling individual copies of *R2d2* is infeasible in high-copy samples. Instead we treated each *R2d* unit as an independent sequence and used the number of segregating sites to estimate sequence diversity. Segregating sites were defined as positions in a collection of alignments (BAM files) with evidence of an alternate allele. To identify segregating sites we used Freebayes v0.9.21-19-gc003c1e (Garrison and Marth 2012) with parameters "-ui -Kp 20 --use-best-n-alleles 2 -m 8." These

parameters treat each sample as having ploidy up to 20, impose an uninformative prior on genotype frequencies, and limit the algorithm to the discovery of atomic variants (single nucleotide variations or short indels, not multi-nucleotide polymorphisms or other complex events) with at most two alleles at each segregating site. Sites in low-complexity sequence (defined as Shannon entropy <1.6 in the 30-bp window centered on the site) or within 10 bp of another variant site were further masked to minimize spurious calls due to ambiguous alignment of indels. To avoid confounding with the retrocopies of *Cwc22* outside *R2d*, coding exons of *Cwc22* were masked. Finally, sites corresponding to an unaligned or gap position in the pairwise alignment between *R2d1* and *R2d2* were masked.

To compute diversity in *R2d1* we counted segregating sites in 12 wild *M. m. domesticus* samples with two diploid copies of *R2d* (total of 24 sequences), confirmed to be in *R2d1* by the presence of read pairs spanning the junction between *R2d1* and a neighboring sequence. To compute diversity in *R2d2*, we counted segregating sites in 14 wild *M. m. domesticus* samples with more than two diploid copies of *R2d* (range 3–83 per sample; total of 406 sequences), but excluded sites corresponding to variants among *R2d1* sequences. Remaining sites were phased to *R2d2* by checking for the presence of a 31-mer containing the site and the nearest *R2d1*-vs-*R2d2* difference in the raw reads for each sample using the corresponding msBWT. Sequence diversity was then computed using Watterson's estimator (Watterson 1975), dividing by the number of alignable bases (128,973) to yield a per-site estimate. Standard errors were estimated by 100 rounds of resampling over the columns in the *R2d1*-vs-*R2d2* alignment.

Analyses of Cwc22 expression

RNA sequencing read alignment: Expression of *Cwc22* was examined in adult whole brain using data from Crowley *et al.* (2015), Sequence Read Archive (SRA) accession number SRP056236. Paired-end reads (2× 100 bp) were obtained from eight replicates for each of three inbred strains: CAST/EiJ, PWK/PhJ, and WSB/EiJ. Raw reads were aligned to the mm10 reference using STAR v2.4.2a (Dobin *et al.* 2012) with default parameters for paired-end reads. Alignments were merged into a single file per strain for further analysis. Expression in adult testis was examined in 23 wild-derived inbred strains from Phifer-Rixey *et al.* (2014), SRA accession number PRJNA252743. Single-end reads (76 bp) were aligned to the mm10 genome with STAR using default parameters for single-end, reads.

Transcript assembly: Read alignments were manually inspected to assess support for *Cwc22* isoforms in Ensembl v83 annotation. To identify novel isoforms in *R2d2*, we applied the Trinity v0.2.6 pipeline (Grabherr *et al.* 2011) to the subset of reads from WSB/EiJ, which could be aligned to *R2d1* plus their mates (a set which represents a mixture of $Cwc22^{R2d1}$ and $Cwc22^{R2d2}$ reads). *De novo* transcripts were aligned both to the mm10 reference and to the *R2d2* contig using BLAT, and were assigned to *R2d1* or *R2d2* based on sequence similarity. Because expression from *R2d2* is high in WSB/EiJ, *R2d2*-derived transcripts dominated the assembled set. Both manual inspection and the Trinity assembly indicated the presence of retained introns and an extra 3′ exon, as described in the *Results*. To obtain a full set of *Cwc22* transcripts including those of both *R2d1* and *R2d2* origin, we supplemented the *Cwc22* transcripts in Ensembl v83 with their paralogs from *R2d2* as determined by a strict BLAT search against the *R2d2* contig. We manually created additional transcripts reflecting intron-retention and 3′ extension events described above, and obtained their sequence from the *R2d2* contig. Transcript coordinates (with respect to *R2d1*) and sequences are provided in File S3.

Abundance estimation: Relative abundance of *Cwc22* paralogs was estimated using kallisto v0.42.3 (Bray *et al.* 2016) with parameters "—bias" (to estimate and correct library-specific, sequence-composition biases). The transcript index used for pseudoalignment and quantification included only the *Cwc22* targets.

Phylogenetic analyses

Tree for R2d: Multiple sequence alignments for eight of the regions in Figure S5 were generated using the MUSCLE software (Edgar 2004) with default parameters. The resulting alignments were manually trimmed and consecutive gaps removed. Phylogenetic trees were inferred with RAxML v8.1.9 (Stamatakis 2014) using the general time reversible (GTR)+gamma model with four rate categories and *M. caroli* as an outgroup. Uncertainty of tree topologies was evaluated using 100 bootstrap replicates.

Divergence time: The time of the split between *R2d1* and *R2d2* was estimated using the Bayesian method implemented in BEAST v1.8.1r6542 (Drummond *et al.* 2012). We assumed a divergence time for *M. caroli* of 5 MYA with a strict molecular clock, and analyzed the concatenated alignment for our *de novo* assembled regions under the GTR+gamma model with four rate categories and allowance for a proportion of invariant sites. The chain was run for 10,000,000 iterations with trees sampled every 1000 iterations.

Local phylogeny around R2d2: Genotypes for 173 SNPs in the region surrounding *R2d2* (chr2: 83–84 Mb) were obtained for 90 individuals representing both laboratory and wild mice genotyped with the MDA (Table S1) (Yang *et al.* 2011). Individuals with evidence of heterozygosity (more than three heterozygous calls) were excluded to avoid ambiguity in phylogenetic inference. A distance matrix for the remaining 62 samples was created by computing the proportion of alleles shared identical by state between each pair of samples. A neighbor-joining tree (Figure 2C) was inferred from the distance matrix and rooted at the most recent common ancestor of the *M. musculus* and non-*M. musculus* samples.

Cwc22 coding sequences: To create the tree of *Cwc22* coding sequences, we first obtained the sequences of all its paralogs in mouse. The coding sequence of *Cwc22*^{R2d1} (RefSeq transcript NM_030560.5) was obtained from the UCSC Genome Browser and aligned to our *R2d2* contig with BLAT to extract the exons of *Cwc22*^{R2d2}. The coding sequence of retro-*Cwc22* [genomic sequence corresponding to GenBank complementary DNA (cDNA) AK145290] was obtained from the UCSC Genome Browser. Coding and protein sequences of CWC22 homologs from non-*M. musculus* species were obtained from Ensembl (Cunningham *et al.* 2014). The sequences were aligned with MUSCLE and manually trimmed, and a phylogenetic tree estimated as described above. Sequence alignments and tree are provided in File S4.

We observed that the branches in the rodent clade of the *Cwc22* tree appeared to be longer than branches for other taxa. We used Phylogenetic Analysis Using Maximum Likelihood (PAML) (Yang *et al.* 2007) to test the hypothesis that *Cwc22* is under relaxed purifying selection in rodents using the branch-site model (null model "model = 2, NSsites = 2, fix_omega = 1"; alternative model "model = 2, NSsites = 2, omega = 1, fix_omega = 1") as described in the PAML manual. This is a test of difference in evolutionary rate on a "foreground" branch (ω_1)—in our case, the rodent clade—relative to the tree-wide "background" rate (ω_0). The distribution of the test statistic is an even mixture of a χ^2 distribution with 1 d.f. and a point mass at zero. To obtain the *P*-value, we calculated the quantile of the χ^2 distribution with 1 d.f. and divided by two.

Analyses of recombination rate at R2d2

To test the effect of *R2d2* copy number on local recombination rate, we examined recombination events accumulated during the first 16 generations of breeding of the DO population, in which the high-copy *R2d2* allele from WSB/EiJ is segregating. Founder haplotype reconstructions were obtained for 4640 DO individuals (reported in Didion *et al.* 2016) and recombination events were identified as junctions between founder haplotypes. We compared the frequency of junctions involving a WSB/EiJ haplotype to junctions not involving a WSB/EiJ haplotype over the 75–90 Mb region of chr2. Within each generation we also tested for differences in the lengths of haplotype blocks overlapping *R2d2* using one-sided Wilcoxon rank-sum tests (alternative hypothesis: WSB/EiJ haplotypes longer than others). Resulting *P*-values were subject to Bonferroni correction: for nominal significance level $\alpha = 0.01$, the corrected threshold is $P = \frac{0.01}{12} = 8.3 \times 10^{-4}$.

We also estimated the difference between observed and expected recombination fraction in 11 experimental crosses, in which one of the parental lines was segregating for a high-copy allele at *R2d2*. We obtained expected recombination fractions from the standard mouse genetic map (Cox *et al.* 2009), which was constructed from crosses between strains lacking *R2d2*^{HC} alleles. Genotype data were obtained from The Jackson Laboratory's Mouse Phenome Database QTL Archive (http://phenome.jax.org/db/q?rtn=qtl/home). Recombination fractions were calculated using R/qtl (http://rqtl.org/). Confidence intervals for difference between observed and expected recombination fractions were calculated by 100 iterations of nonparametric bootstrapping over individuals in each dataset.

Data availability

All *de novo* assemblies used in this study are included in File S1. The data structures on which the assemblies are based, and the interactive computational tools used for assembly, are publicly available at http://www.csbio.unc.edu/CEGSseq/index.py?run=MsbwtTools. Genotype data used for mapping the location of *R2d2* and defining associated haplotypes are provided in File S2. Transcript sequences used for analyses of gene expression are provided in File S3. Nucleotide and amino acid sequences used for phylogenetic analyses of *Cwc22* are provided in File S4.

Results

R2d was duplicated in the common ancestor of M. musculus and M. spretus

In order to determine when the *R2d* CNV arose, we used qPCR and/or depth of coverage in whole-genome sequencing to assay *R2d* copy number in a collection of samples spanning the phylogeny of the genus *Mus*. Samples were classified as having a diploid copy number of two (two chromosomes each with a single copy of *R2d*) or more than two (at least one chromosome with an *R2d* duplication).

We find evidence for more than two diploid copies in representatives of all mouse taxa tested from the Palearctic clade (Figure 1 and Table S1) (Suzuki *et al.* 2004): 236 of 525 *M. musculus*, 1 of 1 *M. macedonicus*, 1 of 1 *M. spicilegus*, 1 of 1 *M. cypriacus*, and 8 of 8 *M. spretus* samples. However, we find no evidence of duplication in species from the southeast Asian clade, which is an outgroup to Palearctic mice: 0 of 2 *M. famulus*, 0 of 2 *M. fragilicauda*, 0 of 1 *M. cervicolor*, 0 of 1 *M. cookii*, and 0 of 1 *M. caroli* samples. Outside the subgenus *Mus*, we found evidence for more than two diploid copies in none of the nine samples tested from subgenus *Pyromys*. We concluded that the *R2d* duplication most likely occurred between the divergence of southeast Asian from Palearctic mice (~3.5 MYA) and the divergence of *M. musculus* from *M. spretus* (~2 MYA) (Suzuki *et al.* 2004; Chevret *et al.* 2005), along the highlighted branch of the phylogeny in Figure 1A. If the *R2d* duplication was fixed in the ancestor of *M. musculus*, then extant lineages of house mice which have only two diploid copies of *R2d*—including the reference strain C57BL/6J (of predominantly *M. musculus domesticus* origin; Yang *et al.* 2007)—represent subsequent losses of an *R2d* copy. Alternatively, the *R2d* duplication may have been polymorphic in the ancestor of *M. musculus* and then continued to segregate in *M. musculus* and *M. spretus*.

Duplication of the ancestral *R2d* sequence resulted in two paralogs residing in loci which we denote *R2d1* and *R2d2* (Figure 1B). Only one of these is present in the mouse reference genome, at chr2: 77.87 Mb; the other copy maps ~6 Mb

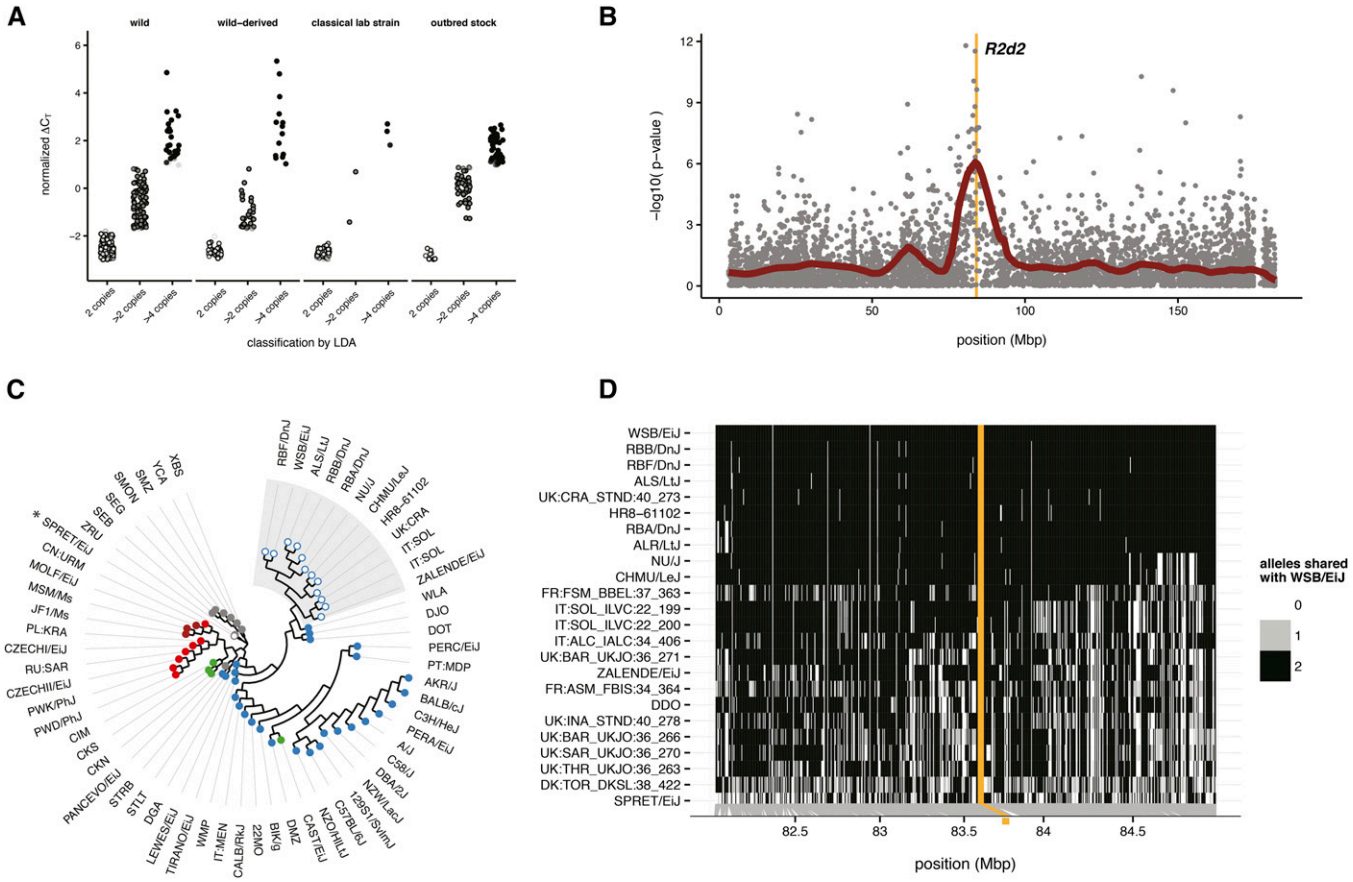


Figure 2 Copy-number variation of *R2d* in mouse populations worldwide. (A) Copy-number variation as measured by qPCR. The normalized ΔCt value is proportional to log2(copy number). Samples are classified as having two diploid copies, more than two copies, or more than four copies of *R2d* using linear discriminant analysis. (B) Fine-mapping the location of *R2d2* in 83 samples genotyped on the MDA. Gray points give nominal *P*-values for association between *R2d* copy number and genotype, red points show a smoothed fit through the underlying points. The candidate interval for *R2d2* from Didion *et al.* (2015), shown as an orange shaded box, coincides with the association peak. (C) Local phylogeny at chr2: 83–84 Mb in 62 wild-caught mice and laboratory strains. Tips are colored by subspecies of origin: blue, *M. m. domesticus*; red, *M. m. musculus*; green, *M. m. castaneus*; maroon, *M. m. molossinus*; and gray, outgroup taxa. Individuals with more than four diploid copies of *R2d* are shown as open circles. (D) Haplotypes of laboratory strains and wild mice sharing a high-copy allele at *R2d2*. All samples share a haplotype over the region shaded in orange.

distal (Didion *et al.* 2015), as we describe in more detail below. The more proximal copy, *R2d1*, lies in a region of conserved synteny with rat, rabbit, chimpanzee, and human (Figure S1) (Muffato *et al.* 2010); we conclude that it is the ancestral copy.

The sequence of the *R2d2* paralog was assembled *de novo* from whole-genome sequence reads (Keane *et al.* 2011) from the strain WSB/EiJ (of pure *M. m. domesticus* origin; Yang *et al.* 2011), which has diploid *R2d* copy number of ∼68 (Didion *et al.* 2015). We exploited the difference in depth of coverage for *R2d1* (one haploid copy) and *R2d2* (33 haploid copies) to assign variants to *R2d1* or *R2d2*. Pairwise alignment of the *R2d2* contig against *R2d1* is shown in Figure S2. The paralogs differ by at least eight TE insertions: seven long interspersed nuclear elements (LINEs) specific to *R2d1* and one endogenous retroviral element (ERV) specific to *R2d2* (Table S2). (Due to the inherent limitations of assembling repetitive elements from short reads, it is likely that we have underestimated the number of young TEs in *R2d2*.) The *R2d1*-specific LINEs are all <2% diverged from the consensus for their respective families in the RepeatMasker database (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker), consistent with insertion within the last 2 MY. The oldest *R2d2*-specific ERV we could detect is 0.7% diverged from its family consensus. TE insertions occurring since the ancestral *R2d* duplication are almost certainly independent, so these data are consistent with duplication <2 MYA. The *R2d* unit, minus paralog-specific TE insertions, is 127 kbp in size. *R2d* units in the *R2d2* locus are capped on both ends by (CTCC)n microsatellite sequences, and no read pairs spanning the breakpoint between *R2d2* and flanking sequence were identified.

In order to obtain a more precise estimate of the molecular age of the duplication event we assembled *de novo* an additional 16.9 kbp of intergenic and intronic sequences in eight regions across the *R2d* unit from diverse samples and constructed phylogenetic trees. The trees cover 17 *R2d1* or *R2d2* haplotypes: 13 from inbred strains, and 4 from wild mice. The sequence of *M. caroli* (diploid copy number of two) is used as an outgroup. A concatenated tree is shown

in Figure 1C. Using 5.0±1.0 MYA as the estimated divergence date for *M. caroli* and *M. musculus* (Suzuki *et al.* 2004; Chevret *et al.* 2005), Bayesian phylogenetic analysis with BEAST v1.8 (Drummond *et al.* 2012) yields 1.6 MYA (95% highest probability density 0.7–5.1 MYA) as the estimated age of the duplication event that gave rise to *R2d1* and *R2d2*. Although the assumption of a uniform molecular clock may not be strictly fulfilled for *R2d1* and *R2d2*, the totality of evidence—from presence/absence data across the mouse phylogeny, paralog-specific TE insertions, and sequence divergence between paralogs—strongly supports the conclusion that *R2d* was first duplicated within the last 2 MY in the common ancestor of *M. musculus* and *M. spretus*. All *de novo* assembled sequences are provided in File S1.

For clarity, Figure 1D illustrates diploid copy-number states that will be referenced in the remainder of the manuscript. Hereafter we refer to diploid copy numbers except when discussing inbred strains (which are effectively haploid).

Copy number at R2d2 is highly polymorphic in M. musculus

We previously demonstrated that haploid copy number of *R2d* ranges from one in the reference strain C57BL/6J and classical inbred strains A/J, 129S1/SvImJ, NOD/ShiLtJ, and NZO/HILtJ; to two in the wild-derived strain CAST/EiJ; to 34 in the wild-derived strain WSB/EiJ. Using linkage mapping in two multi-parental genetic reference populations, the Collaborative Cross (Collaborative Cross Consortium 2012) and DO (Svenson *et al.* 2012), we showed that for the two strains with haploid copy number of more than one, one of the copies maps to *R2d1* while all extra copies map to the *R2d2* locus at chr2: 83 Mb (Didion *et al.* 2015). *Cwc22* was recently reported to have diploid copy number as high as 83 in wild *M. m. domesticus* (Pezer *et al.* 2015). In whole-genome sequence data from >60 mice from both laboratory stocks and natural populations (Table S1), we have observed zero instances in which the *R2d* copy in *R2d1* is lost. We conclude that diploid copy number of more than two indicates at least one copy of *R2d* is present in *R2d2* (Figure 1D).

In order to understand the evolutionary dynamics of copy-number variation at *R2d2*, we investigated the relationship between copy number and the local phylogeny in the *R2d2* candidate region. In particular, we sought evidence for or against a single common origin for each of the copy-number states at *R2d2* which are derived with respect to the *M. spretus*–*M. musculus* common ancestor (Figure 1D). If a derived copy-number state has a single recent origin, it should be associated with a single haplotype at *R2d2*. If a derived copy-number state arises by recurrent mutation, the same copy number should be associated with multiple haplotype backgrounds and possibly in multiple populations.

The extent of *R2d* copy-number variation in *M. musculus*, as estimated on a continuous scale by qPCR, is shown in Figure 2A. (Note that the qPCR readout is proportional to copy number on the log scale. Extrapolation to integer copy number is imprecise for copy numbers of more than approximately six.) We confirmed that *R2d2* maps to chr2: 83 Mb by performing association mapping between SNP genotypes from the MegaMUGA array (Morgan *et al.* 2015) and the qPCR readout (Figure 2B).

We performed a similar analysis to test the hypothesis that *R2d2* alleles with high copy number (diploid copy number more than four, hereafter $R2d2^{HC}$; Figure 1D) have a single origin. First we observed that $R2d2^{HC}$ alleles are confined with few exceptions to *M. m. domesticus* (Table S3). The best-associated SNP on the MegaMUGA array (JAX00494952) only weakly tags copy number (r^2 = 0.137), but ascertainment bias on the MUGA platform (Morgan *et al.* 2015) makes local linkage disequilibrium patterns difficult to interpret. To examine further, we constructed a neighbor-joining phylogenetic tree for the region containing *R2d2* (chr2: 83–84 Mb) using genotypes from the 600K SNP MDA (Yang *et al.* 2011). We restricted our attention to inbred strains or wild mice with homozygous, nonrecombinant haplotypes in the target region. Twelve samples with $R2d2^{HC}$ alleles, both wild mice and laboratory stocks, cluster in a single clade (Figure 2C). (A single *M. spretus* strain, SPRET/EiJ, also carries an $R2d2^{HC}$ allele, but see *Discussion*).

Next we expanded the analysis to include an additional 11 samples with $R2d2^{HC}$ alleles and evidence of heterozygosity around *R2d2*. The total set of 24 samples includes 7 wild-derived laboratory strains (DDO, RBA/DnJ, RBB/DnJ, RBF/DnJ, WSB/EiJ, ZALENDE/EiJ, and SPRET/EiJ), 4 classical inbred strains (ALS/LtJ, ALR/LtJ, CHMU/LeJ and NU/J), 1 line derived from the ICR:HsD outbred stock (HR8; Swallow *et al.* 1998), and 12 wild-caught mice. All 24 samples with $R2d2^{HC}$ alleles share an identical haplotype across a single 21-kbp interval, chr2: 83,896,447–83,917,565 (GRCm38/mm10 coordinates) (Figure 2D). These analyses support a single origin for $R2d2^{HC}$ alleles within *M. m. domesticus*.

To test the hypothesis that losses of *R2d2* (diploid copy number less than four, at least one chromosome with zero copies in *R2d2*; Figure 1D) have a single origin, we examined their distribution across the three well-differentiated subspecies of *M. musculus*. Losses of *R2d2* occur in all subspecies of *M. musculus* in populations that span its geographic range (Table S3). Based on this distribution and a previous observation that no common haplotype is shared in samples with low copy number in *M. m. domesticus* (Didion *et al.* 2016), we reject the hypothesis of single origin and conclude that *R2d2* has been lost multiple times on independent lineages in each subspecies.

Alternatively, we could posit that the *R2d* duplication never fixed in the ancestor of *M. musculus* and that both duplicated and unduplicated alleles have been maintained for 2 MY as balanced polymorphisms in the major lineages in the Palearctic clade of *Mus*. We find this a less-likely scenario given current estimates of effective population size (N_e) in house mice (50,000–250,000; Suzuki *et al.* 2004; Salcedo *et al.* 2007; Halligan *et al.* 2013) and the expected fixation time of a neutral allele ($\sim 4N_e$; Kimura and Ohta 1968).

The genomic region containing R2d2 is structurally unstable but has low sequence diversity

The extent of copy-number polymorphism involving *R2d2* suggests that it is prone to recurrent mutation. Consistent with these observations, we find that the rate of *de novo* copy-number changes at *R2d2* is extremely high in laboratory populations (Figure 3). In 183 mice sampled from the DO population we identified and confirmed through segregation analysis eight new alleles, each with distinct copy number and each occurring in an unrelated haplotype (Table S4). Without complete pedigrees and genetic material from breeders, a direct estimate of the mutation rate in the DO is not straightforward to obtain. However, since the population size is known, we can make an analogy to microsatellite loci (Moran 1975) and estimate the mutation rate via the variance in allele sizes: 3.2 mutations per 100 transmissions (3.2%) (95% bootstrap C.I. 1.1–6.0%).

Structural instability in this region of chr2 extends outside the *R2d2* locus itself. There is another duplication <200 kbp distal to *R2d2* (Figure 4B, gray shaded region)—containing a retrotransposed copy of *Cwc22*—that is present in seven tandem copies in the reference genome. That region, plus a further 80 kbp immediately distal to it, is copy-number polymorphic in wild *M. m. domesticus* and wild *M. m. castaneus* (Figure 4B). Instability of the region over longer timescales is demonstrated by the disruption, just distal to the aforementioned SD, of a syntenic block conserved across all other mammals (Figure S1).

Despite the high mutation rate for structural variants involving *R2d2* and nearby sequences, sequence diversity at the nucleotide level is modestly reduced relative to diversity in *R2d1* and relative to the genome-wide average in *M. m. domesticus*. In a 200-kbp region containing the *R2d2* insertion site at its proximal end, $\hat{\pi}$ (an estimator of average heterozygosity) in *M. m. domesticus* reduced by at least a factor of two from the local average of ~0.3% (which is comparable to previous reports in this subspecies, see Salcedo *et al.* 2007) (Figure 4B). Divergence between *M. musculus* and *M. caroli* is similar to its genome-wide average of ~2.5% over the same region.

Estimation of diversity *within* a duplicated sequence such as *R2d* is complicated by the difficulty of distinguishing allelic from paralogous variation. To circumvent this problem we split our sample of 26 wild *M. m. domesticus* into two groups: those having *R2d1* sequences only, and those having both *R2d1* and *R2d2* sequences. Within each group we counted the number of segregating sites among all *R2d2* copies, using nearby fixed differences between *R2d1* and *R2d2* to phase sites to *R2d2* (see *Materials and Methods* for details), and used Watterson's estimator of theta to calculate nucleotide diversity per site. Among *R2d1* sequences in *M. m. domesticus*, $\theta = 0.09 \pm 0.03\%$ *vs.* $\theta = 0.04 \pm 0.02\%$ among *R2d2* sequences in *M. m. domesticus* (Figure 4C) and $\theta = 0.13 \pm 0.04\%$ among *R2d2* sequences in *M. m. castaneus*.

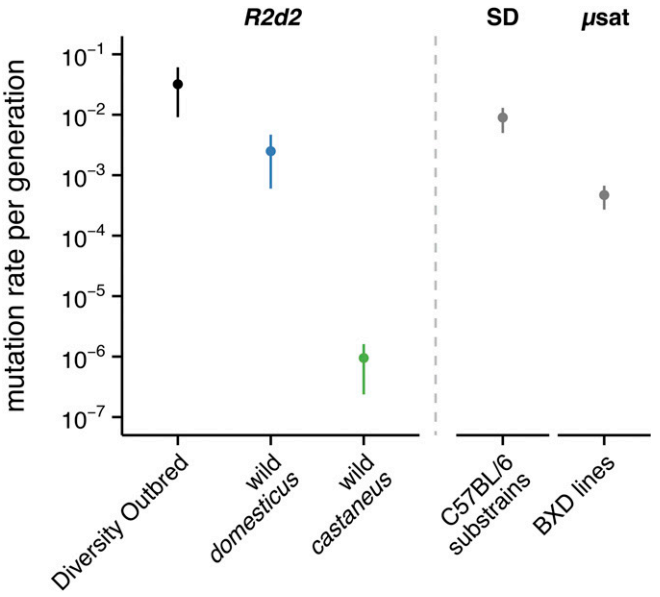


Figure 3 Rate of *de novo* copy-number changes at *R2d2*. Estimates of per-generation mutation rate for CNVs at *R2d2* (±1 bootstrap SE) in the DO population, among wild *M. m. domesticus*, and among wild *M. m. castaneus*. For comparison, mutation rates are shown for the CNV with the highest rate of recurrence in a C57BL/6J pedigree (Egan *et al.* 2007) and for a microsatellite whose mutation rate was estimated in the BXD panel (Dallas 1992).

R2d contains the essential gene Cwc22

The *R2d* unit encompasses one protein-coding gene, *Cwc22*, which encodes an essential (mRNA) splicing factor (Yeh *et al.* 2010). The gene is conserved across eukaryotes and is present in a single copy in most non-rodent species represented in the TreeFam database (http://www.treefam.org/family/TF300510; Li 2006). Five groups of *Cwc22* paralogs are present in mouse genomes: the copies in *R2d1* ($Cwc22^{R2d1}$) and *R2d2* ($Cwc22^{R2d2}$), the retrotransposed copies in one locus at chr2: 83.9 Mb, and at two loci on the X chromosome (chrX) (Figure 5A).

The three retrotransposed copies are located in regions with no sequence similarity to each other, indicating that each represents an independent retrotransposition event. The copy on chr2 was subsequently expanded by further duplication and now exists (in the reference genome) in seven copies with >99.9% mutual similarity. The two retrotransposed copies on chrX are substantially diverged from the parent gene (< 90% sequence similarity), lack intact open reading frames (ORFs), have minimal evidence of expression among GenBank cDNAs, and are annotated as likely pseudogenes (Pei *et al.* 2012). We therefore restricted our analyses to the remaining three groups of *Cwc22* sequences, all on chr2.

The canonical transcript of $Cwc22^{R2d1}$ (ENSMUST00000065889) is encoded by 21 exons on the negative strand. The coding sequence begins in the third exon and ends in the terminal exon (Figure 5B). Six of the seven protein-coding $Cwc22^{R2d1}$ transcripts in Ensembl v83 use this terminal exon, while one transcript (ENSMUST0000011824) uses an alternative

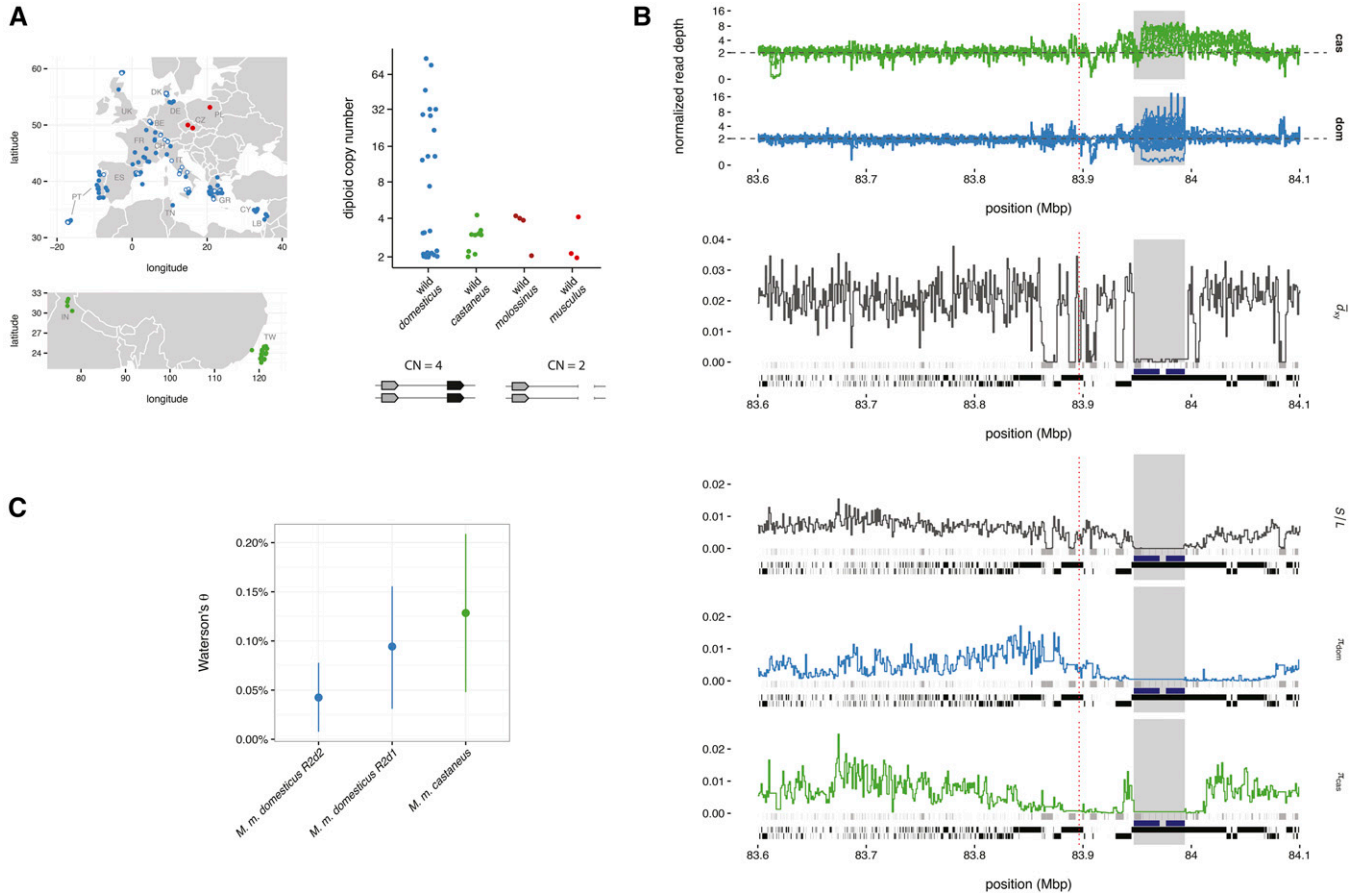


Figure 4 Sequence and structural diversity around *R2d2*. (A) Geographic origin of wild mice used in this study, color-coded by subspecies (blue, *M. m. domesticus*; red, *M. m. musculus*; green, *M. m. castaneus*). Diploid copy number of the *R2d* unit is shown for wild samples for which integer copy-number estimates are available: 26 *M. m. domesticus* and 10 *M. m. castaneus* with whole-genome sequencing data, and representatives from *M. m. molossinus* and *M. m. musculus* for comparison. Schematic shows the *R2d1*/*R2d2* configurations corresponding to diploid copy numbers of two and four. CN, copy number. (B) Profiles of read depth (top two panels), average sequence divergence to outgroup species *M. caroli* (d_{xy}, third panel), number of segregating sites per base (*S*/*L*, fourth panel) and within-population average heterozygosity (π, fifth and sixth panels). The region shown is 500 kbp in size; the insertion site of *R2d2* is indicated by the red dashed line. Gray boxes along baseline show positions of repetitive elements (from UCSC RepeatMasker track); black boxes show nonrecombining haplotype blocks. Blue bars indicate the position of seven tandem duplications in the mm10 reference sequence with >99% mutual identity, each containing a copy of retro-*Cwc22*. Gray shaded region indicates duplicate sequence absent from *M. caroli*. (C) Estimated per-site nucleotide diversity within *M. m. domesticus R2d1*, *M. m. domesticus R2d2*, and *M. m. castaneus R2d2*.

terminal exon. Alignment of the retrogene sequence (ENSMUST00000178960) to the reference genome demonstrates that the retrogene captures the last 19 exons of the canonical transcript—that is, the 19 exons corresponding to the coding sequence of the parent gene.

Cwc22 is intact in and expressed from all R2d paralogs, and fast-evolving in rodents

To identify the coding sequence of $Cwc22^{R2d2}$, we first aligned the annotated transcript sequences of $Cwc22^{R2d1}$ from Ensembl to our *R2d2* contig. All 21 exons present in *R2d1* are present in *R2d2*. We created a multiple sequence alignment and phylogenetic tree of *Cwc22* cDNAs and predicted amino acid sequences from $Cwc22^{R2d1}$, $Cwc22^{R2d2}$, retro-*Cwc22*, and *Cwc22* orthologs in 19 other placental mammals, plus opossum, platypus, and finally chicken as an outgroup (Figure S3). An ORF is maintained in all three *Cwc22* loci in mouse, including the retrogene. Information content of each column along the alignment (Figure S4) reveals that sequence is most conserved in two predicted conserved domains, MIF4G and MA3, required for *Cwc22*'s function in mRNA processing (Yeh *et al.* 2010).

Next we examined public RNA-sequencing (RNA-seq) data from adult brain and testis in inbred strains with one or more copies of *R2d2* for evidence of transcription of each *Cwc22* family member. We identified several novel transcript isoforms specific to *R2d2* arising from two intron-retention events and one novel 3′ exon (Figure 6A). The 18th intron is frequently retained in $Cwc22^{R2d2}$ transcripts, most likely due to an A > G mutation at the 5′ splice donor site of exon 17 in $Cwc22^{R2d2}$. The 12th intron is also frequently retained. While we could not identify any splice-region variants near this intron, it contains an ERV insertion that may interfere with splicing (Figure 6A). Both intron-retention events would

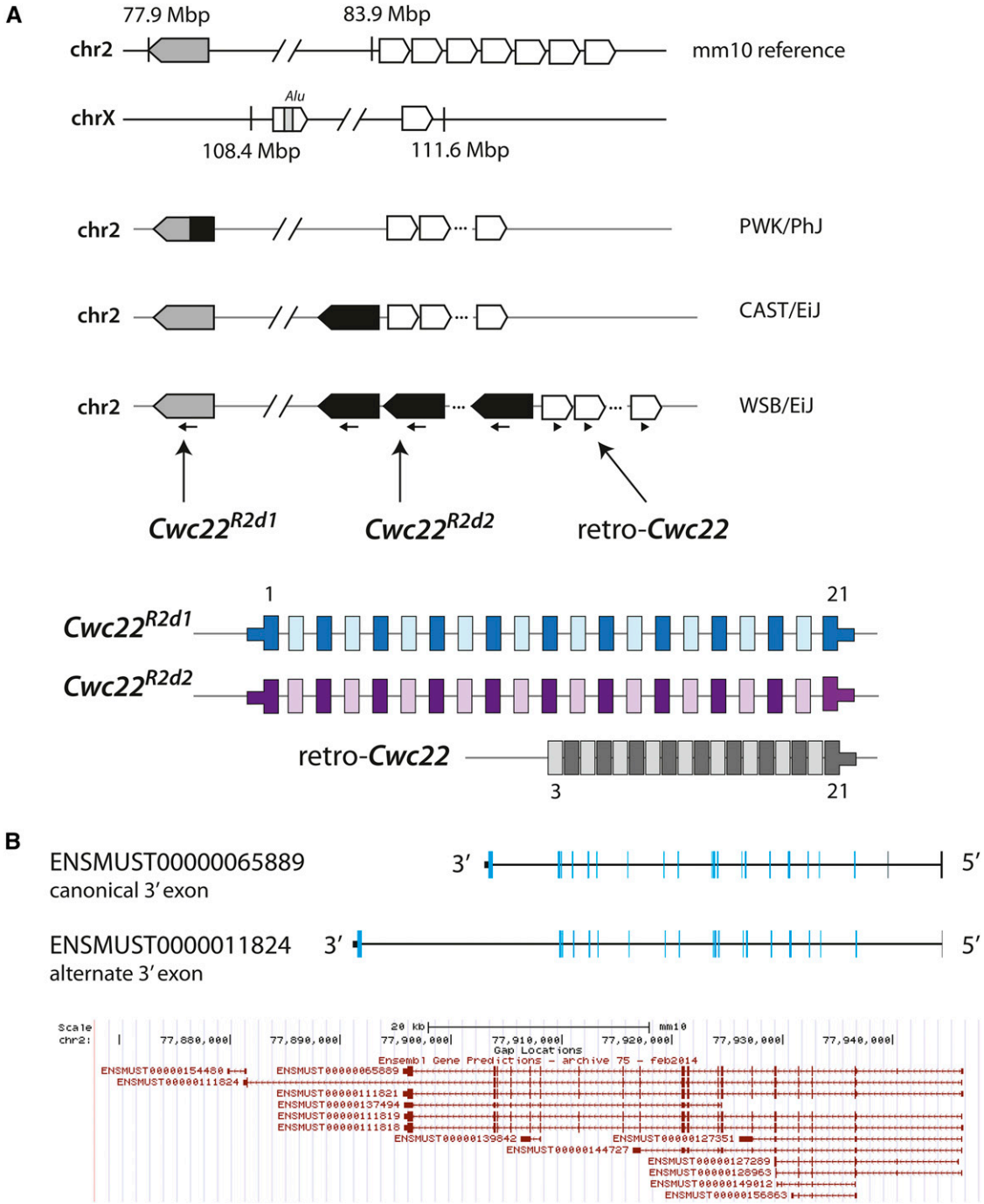


Figure 5 *Cwc22* paralogs in the mouse genome. (A) Location and organization of *Cwc22* gene copies present in mouse genomes. The intact coding sequence of *Cwc22* exists in in both *R2d1* (shaded shapes) and *R2d2* (solid shapes). Retrotransposed copies (open shapes) exist in two loci on chrX and one locus on chr2, immediately adjacent *R2d2*. Among the retrotransposed copies, coding sequence is intact only in the copy on chr2. Exon numbers are shown in gray above transcript models. (B) Alternate transcript forms of *Cwc22*, using different 3′ exons. Coding exons shown in blue and untranslated regions in black. All Ensembl annotated transcripts are shown in the lower panel (from UCSC Genome Browser.)

create an early stop codon. Finally, we find evidence for a novel 3′ exon that extends to the boundary of the *R2d* unit and is used exclusively by *Cwc22*^{R2d2} (Figure 6A).

We estimated the expression of the various isoforms of *Cwc22*^{R2d1}, *Cwc22*^{R2d2}, and retro-*Cwc22* in adult brain and testis. For brain we obtained reads from eight replicates (representing both sexes) on three inbred strains, and for testis a single replicate on 23 inbred strains; and estimated transcript abundance using the kallisto package (Bray *et al.* 2016). Briefly, kallisto uses an expectation-maximization algorithm to accurately estimate the abundance of a set of transcripts by distributing the "weight" of

each read across all isoforms with compatible sequence. *Cwc22* is clearly expressed from all three paralogs in both brain and testis (Figure 6B). However, both the total expression and the pattern of isoform usage differ by tissue and copy number.

Maintenance of an ORF in all *Cwc22* paralogs for >2 MY is evidence of negative selection against disrupting mutations in the coding sequence, but long branches within the rodent clade in Figure S3 suggest that *Cwc22* may also be under relaxed purifying selection or positive selection in rodents. The rate of evolution of *Cwc22* sequences in mouse is faster than in the rest of the tree ($\chi^2 = 4.33$, d.f. = 1, $P = 0.037$).

Phylogenetic discordance in R2d1 is due to nonallelic gene conversion

The topology of trees across *R2d* is generally consistent: a long branch separating the single *M. caroli* sequence from the *M. musculus* sequences, and two clades corresponding to *R2d1*- and *R2d2*-like sequences. However, we observed that the affinities of some *R2d* paralogs change along the sequence (Figure 7A), a signature of nonallelic (*i.e.*, interlocus) gene conversion. In this context, we use "gene conversion" to describe a nonreciprocal "copy-and-paste" transfer of sequence from one donor locus into a different, homologous receptor locus, without reference to a specific molecular mechanism (Chen *et al.* 2007).

To investigate further, we inspected patterns of sequence variation in whole-genome sequencing data from 15 wild-caught mice, 2 wild-derived inbred strains, and 22 classical inbred strains of mice with a diploid *R2d* copy number of two. We first defined 1411 pairwise single-nucleotide differences (1 per 89 bp; Ti:Tv = 1.85) between *R2d2* and *R2d1*, for which *R2d2* has the derived allele with respect to *M. caroli*. We then tested for the presence of the derived allele, ancestral allele, or both at each site in each sample. Finally, we identified conversion tracts by manual inspection as clusters of derived variants shared with *R2d2* (Figure S5).

This analysis revealed nonallelic gene conversion tracts on at least 9 chromosomes out of the small sample of 54 chromosomes examined (Figure 7B). The conversion tracts range in size from ~1.2 to ~119 kbp. The boundaries of several tracts are shared within populations, suggesting that they are identical by descent. We excluded the possibility of complementary losses from *R2d1* and *R2d2* (which would leave similar patterns of sequence variation) by finding read pairs spanning the boundary between *R2d1* and flanking sequence, and between *R2d1*- and *R2d2*-like tracts on the same chromosome (examples shown in Figure 7C).

The conversion tracts we detected are orders of magnitude longer than the 15–750 bp reported in recent studies of allelic gene conversion at recombination hotspots in mouse meiosis (Cole *et al.* 2010, 2014). We require the presence of *R2d2*-diagnostic alleles at two or more consecutive variants to declare a conversion event, and these variants occur at a rate of ~1 per 100 bp, so the smallest conversion tracts we could theoretically detect are on the order of 200 bp in size. Even if we require only a single variant to define a conversion tract, all samples without a long conversion tract share <55 and most <10 (of 1411) derived alleles with *R2d2*, of which all are also shared by multiple other samples from different populations (Figure S5). This pattern indicates that those sites in fact represent either artifacts (from mis-assignment of ancestral and derived alleles) or homoplasy rather than short gene conversions.

Four conversion tracts partially overlap the *Cwc22* gene to create a sequence that is a mosaic of *R2d1*- and *R2d2*-like exons (Figure 7B). Recovery of *Cwc22* mRNA in an inbred strain (PWK/PhJ) indicates that its exons are intact, adjacent, and properly oriented in *cis*. The presence of both *R2d1*- and *R2d2*-like sequence in extant *M. musculus* lineages with two diploid copies of *R2d* further reinforces our conclusion that the duplication is indeed ancestral to the divergence of *M. musculus*.

In addition to exchanges between *R2d1* and *R2d2*, we identified an instance of exchange between *R2d2* and the nearby retrotransposed copy of *Cwc22* in a single *M. m. domesticus* individual from Iran (IR:AHZ_STND:015; Figure S6). This individual carries a rearrangement that has inserted a 30-kbp fragment corresponding to the 3′ half of $Cwc22^{R2d2}$ into the retro-*Cwc22* locus, apparently mediated by homology between the exons of $Cwc22^{R2d2}$ and retro-*Cwc22*.

High copy number at R2d2 suppresses meiotic recombination

The difficulty of fine-mapping *R2d2* in crosses (Didion *et al.* 2015) and patterns associated with selective sweeps for $R2d2^{HC}$ (Didion *et al.* 2016) suggested that recombination is suppressed around *R2d2*. Based on a previous observation that the rate of meiotic recombination is reduced near clusters of SDs (Liu *et al.* 2014), we tested whether the region around *R2d2* has lower recombination when an $R2d2^{HC}$ allele is present. Understanding patterns of recombination at *R2d2* is important for interpreting levels of sequence and haplotype diversity in the surrounding region.

First we analyzed local recombination rate in the DO population. The DO is an outbred stock derived from eight inbred founder strains (including one, WSB/EiJ, with an $R2d2^{HC}$ allele) and maintained by random mating with 175 breeding pairs; at each generation, one male and one female offspring are chosen from each mating and randomly paired with a nonsibling to produce the next generation (Svenson *et al.* 2012). Figure 8A shows the cumulative distribution of 2917 recombination events on central chr2, stratified according to *R2d2* copy number of the participating haplotypes. The recombination map has a pronounced plateau in the region between *R2d1* and ~1 Mb distal to *R2d2* (dashed lines) for $R2d2^{HC}$ haplotypes, but not $R2d2^{LC}$ haplotypes. As a result, $R2d2^{HC}$ haplotype blocks overlapping *R2d2* are significantly longer than $R2d2^{LC}$ haplotype blocks ($p < 0.01$ by Wilcoxon rank-sum tests with Bonferroni

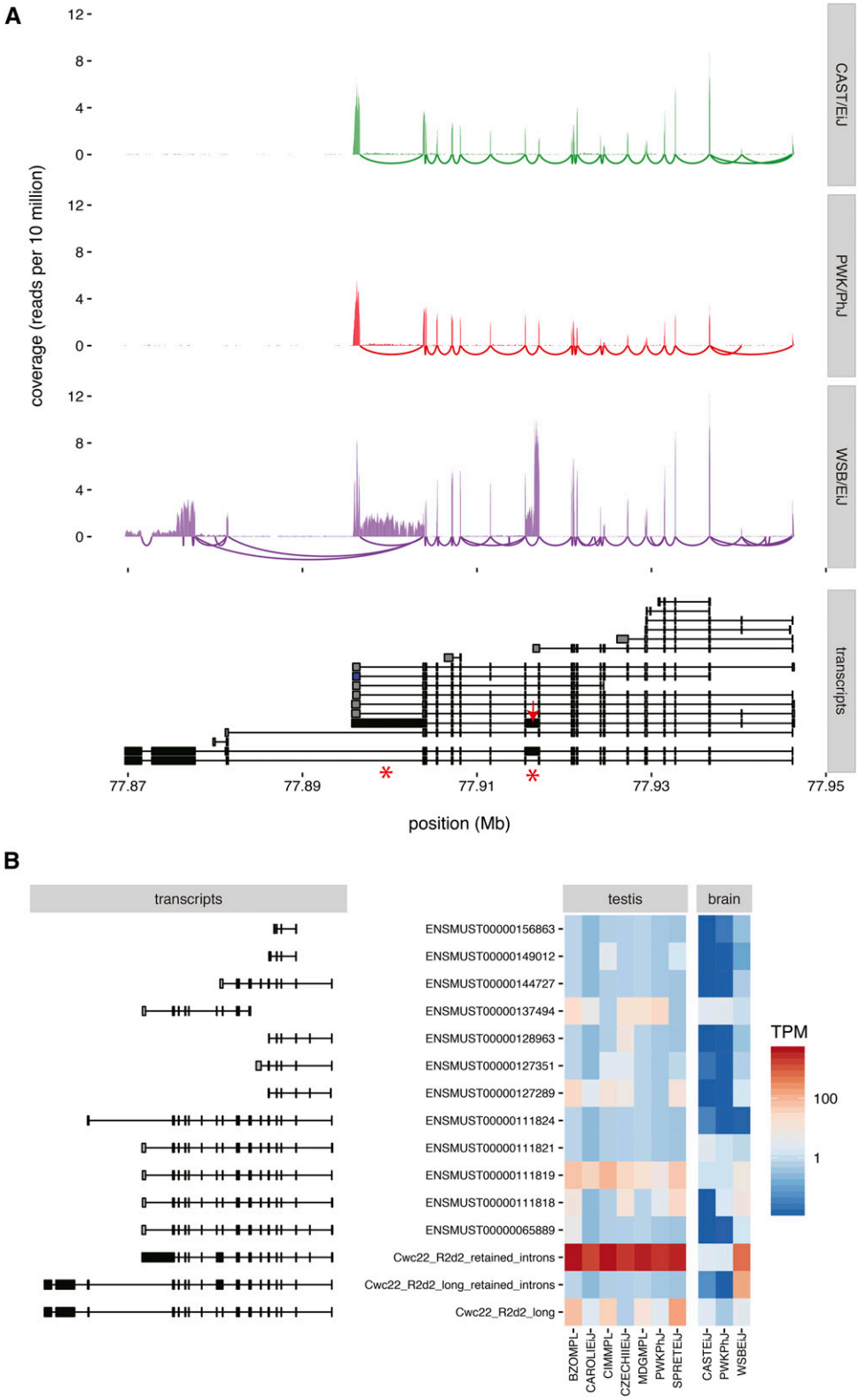


Figure 6 Expression of *Cwc22* isoforms. (A) Read coverage and splicing patterns in *Cwc22* in adult whole brain from three wild-derived inbred strains. Swoops below *x*-axis indicate splicing events supported by five or more split-read alignments. Known transcripts of *Cwc22*^{R2d1} (gray, from Ensembl), inferred transcripts from *Cwc22*^{R2d2} (black), and the sequence of retro-*Cwc22* mapped back to the parent gene (blue) are shown in the bottom panel. Red stars indicate retained introns, red arrow indicates insertion site of an ERV in *R2d2*. (B) Estimated relative expression of *Cwc22* isoforms (*y*-axis) in adult mouse brain and testis in wild-derived inbred strains (*x*-axis). TPM, transcripts per million on log10 scale.

correction) in 8 of the 10 generations sampled (Figure 8B). The difference arose early in the breeding of the DO and persists through the most recent generation for which the randomized breeding scheme was maintained (E. J. Chesler, D. M. Gatti, A. P. Morgan, M. Strobel, L. Trepanier, D. Oberbeck, S. McWeeney, R. Hitzemann, M. Ferris, R. McMullan, A. Clayshulte, T. A. Bell, F. P. M. de Villena, and G. A. Churchill, unpublished data).

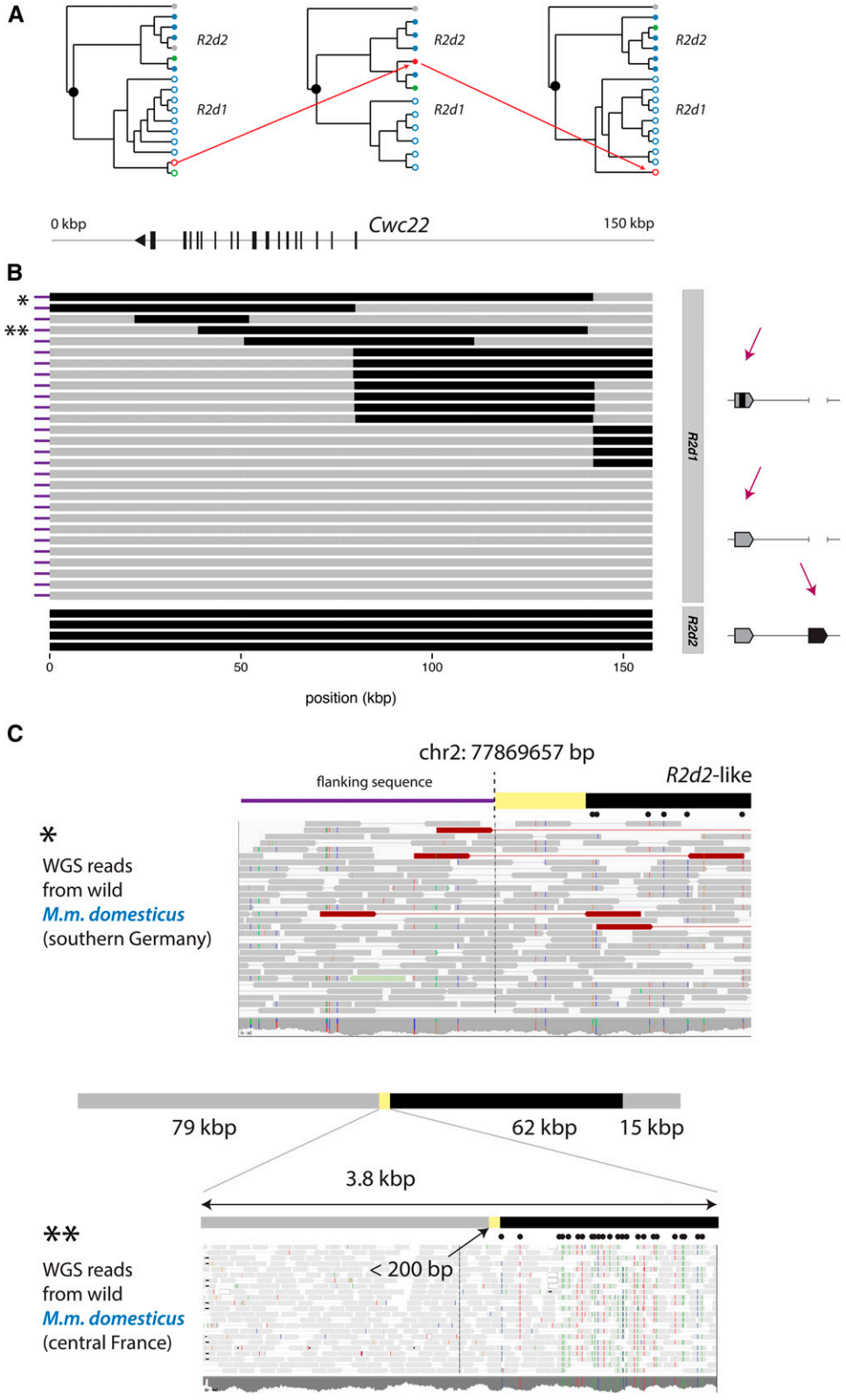


Figure 7 Signatures of nonallelic gene conversion between *R2d1* and *R2d2*. (A) Phylogenetic trees for three representative intervals across *R2d*. Sequences are labeled according to their subspecies of origin using the same color scheme as in Figure 4; open circles are *R2d1*-like sequences and solid circles are *R2d2*-like. Trees are drawn so that *M. caroli*, the outgroup species used to root the trees, is always positioned at the top. The changing affinities of PWK/PhJ (red) and CAST/EiJ (green) along *R2d* are evidence of nonallelic gene conversion. (B) *R2d* sequences from 20 wild-caught mice and 5 laboratory inbred strains. Each track represents a single chromosome; shaded regions are classified as *R2d1*-like based on manual inspection of sequence variants, and solid regions as *R2d2*-like. Top panel shows sequences from samples with a single copy of *R2d*, residing in *R2d1*. Bottom panel shows representative *R2d2* sequences for comparison. * indicates samples for which read alignments are shown in panel C. (C) Top panel: paired-end read alignments (visualized with Integrative Genomics Viewer) across the proximal boundary (dashed line) of *R2d1* in a sample with a conversion tract extending to the boundary. Positions of derived variants shared with *R2d2* are indicated by black dots. Bottom panel: read alignments across the boundary of a nonallelic gene conversion tract. *R2d1* sequence from a single chromosome with a mosaic of *R2d1*-like (gray) and *R2d2*-like (black) segments. A magnified view of read pairs in the 3.8 kbp surrounding the proximal boundary of the tract shows read pairs spanning the junction. Black dots indicate the position of derived alleles diagnostic for *R2d2*. The precise breakpoint lies somewhere in the yellow shaded region between the last *R2d1*-specific variant and the first *R2d2*-specific variant.

Second, we reexamined genotype data from 11 published crosses in which at least one parent was segregating for an $R2d2^{HC}$ allele. Whereas in the DO we used haplotype block length as a proxy for recombination rate, in these F_2 and backcross designs we can directly estimate the recombination fraction across *R2d2* and compare it to its expected value in the absence of an $R2d2^{HC}$ allele (Figure S7). In 9 of 11 crosses examined, the observed recombination

fraction is lower than the expected ($P < 0.032$, one-sided binomial test).

Discussion

In this manuscript we have reconstructed in detail the evolution of a multi-megabase SD in mouse, *R2d2*. Our findings illustrate the challenges involved in accurately interpreting patterns of polymorphism and divergence within duplicated sequence.

SDs are among the most dynamic loci in mammalian genomes. They are foci for copy-number variation in populations, but the sequences of individual duplicates beyond those present in the reference genome are often poorly resolved. Obtaining the sequence of this "missing genome," as we have done for *R2d2*, is an important prerequisite to understanding the evolution of duplicated loci. Since each paralog follows a partially-independent evolutionary trajectory, individuals in a population may vary both quantitatively (in the number of copies) and qualitatively (in which copies are retained). Cycles of duplication and loss may furthermore lead to the fixation of different paralogs along different lineages. This "genomic revolving door" (Demuth *et al.* 2006) leaves a signature of polymorphism far in excess of the genome-wide background, due to coalescence between alleles originating from distinct paralogs.

Accurate deconvolution of recent duplications remains a difficult task that requires painstaking manual effort. Clone-based and/or single-molecule, long-read sequencing remain the gold standard techniques. But short reads at sufficient depth nonetheless contain a great deal of information. We exploited the specific properties of *R2d2* in the WSB/EiJ mouse strain—many highly-similar copies of *R2d2* relative to *R2d1*, with informative paralogous variants every ~100 bp—to obtain a nearly complete assembly of *R2d2* from short reads (Figure S8). With the sequence of both the *R2d1* and *R2d2* paralogs in hand, we were able to recognize several remarkable features of *R2d2* that are discussed in detail below.

Long-tract gene conversion

Previous studies of nonallelic gene conversion in mouse and human have focused either on relatively small (<5 kbp) intervals within species, or have applied phylogenetic methods to multiple paralogs from a single reference genome (Dumont and Eichler 2013). This study is the first, to our knowledge, with the power to resolve large (>5 kbp) nonallelic gene conversion events on an autosome in a population sample. We identify conversion tracts up to 119 kbp in length, orders of magnitude longer than tracts arising from allelic conversion events during meiosis. Gene conversion at this scale can rapidly and dramatically alter paralogous sequences, including (as shown in Figure 7) the sequences of essential protein-coding genes. This process has been implicated as a source of disease alleles in humans (Chen *et al.* 2007).

Importantly, we were able to identify nonallelic exchanges in *R2d1* as such only because we were aware of the existence of *R2d2* in other lineages. In this case the transfer of paralogous *R2d2* sequence into *R2d1* creates the appearance of deep coalescence among *R2d1* sequences. Ignoring the effect of gene conversion would cause us to overestimate the degree of polymorphism at *R2d1* by an order of magnitude, and would bias any related estimates of population-genetic parameters (for instance, of effective population size).

Our data are not sufficient to estimate the rate of nonallelic gene conversion between *R2d2* and homologous loci. At minimum we have observed two distinct events: one from *R2d2* into *R2d1*, and a second from *R2d2* into retro-*Cwc22*. From a single conversion event replacing most of *R2d1* with *R2d2*-like sequence, the remaining shorter conversion tracts could be generated by recombination with *R2d1* sequences. Because we find converted haplotypes in both *M. m. musculus* and *M. m. domesticus*, the single conversion event would have had to occur prior to the divergence of the three *M. musculus* subspecies and subsequently remain polymorphic in the diverged populations. We note that all conversion tracts we observed are polarized: *R2d2* is always the donor.

The other possibility is that nonallelic gene conversion between *R2d* sequences is recurrent. Recurrent gene conversion homogenizes duplicate sequences, coupling their evolutionary trajectories ("concerted evolution", Dover 1982). The absolute sequence divergence (~2%) between *R2d1* and *R2d2* (Figure 1B) argues against the hypothesis that gene conversion has occurred at a uniformly high rate throughout their history. However, we cannot rule out a role for gene conversion in maintaining sequence identity between multiple copies of *R2d* located in *R2d2*. This would help explain the reduced diversity within *R2d2 vs. R2d1* (Figure 4C). There is some direct evidence that the rate of gene conversion is positively correlated with copy number and negatively correlated with physical distance between duplicates (Melamed and Kupiec 1992), so we might expect it to be highest for $R2d2^{HC}$ alleles.

In this respect *R2d2* may be similar to the male-specific region of the *Y* chromosome (chrY) in mouse (Soh *et al.* 2014) and human (Rozen *et al.* 2003). The large palindromic repeats on chrY are homogenized by frequent nonallelic gene conversion (Hallast *et al.* 2013) such that they have retained >99% sequence identity to each other even after millions of years of evolution. Frequent nonallelic gene conversion has also been documented in arrays of U2 small nuclear RNA genes in human (Liao *et al.* 1997), and in ribosomal RNA gene clusters (Eickbush and Eickbush 2007) and centromeric sequences (Schindelhauer 2002; Shi *et al.* 2010) in several species.

Pervasive copy-number variation

Clusters of SDs have long been known to be hotspots of copy-number variation in populations (Bailey and Eichler 2006; She *et al.* 2008) and *de novo* mutations in pedigrees (Egan *et al.* 2007). Recent large-scale sequencing efforts have

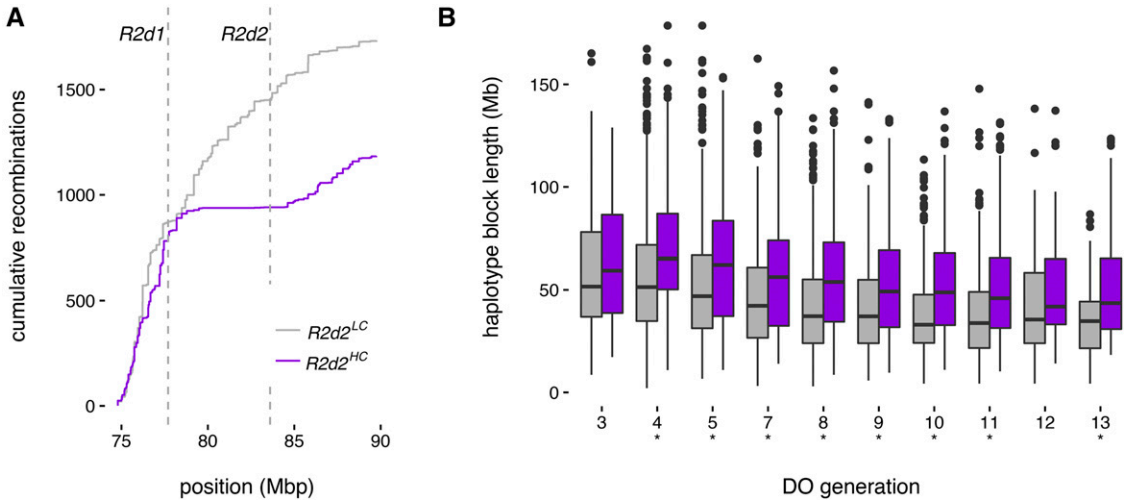


Figure 8 Suppression of crossing-over around *R2d2*. (A) Cumulative number of unique recombination events in the middle region of chr2 in genomes of 4640 DO mice. Recombination events involving the high copy-number WSB/EiJ haplotype are shown in purple and all other events in gray. Dashed vertical lines indicate the position of *R2d1* (left) and *R2d2* (right). (B) Distribution of haplotype block sizes at *R2d2* in selected generations of the DO, for $R2d2^{HC}$ (WSB/EiJ, purple) *vs.* $R2d2^{LC}$ (the other seven founder haplotypes, gray). * indicates generations in which the length distributions are significantly different by Wilcoxon rank-sum test.

revealed the existence of thousands of multi-allelic CNVs segregating in human populations (Handsaker *et al.* 2015).

We have surveyed *R2d2* copy number in a large and diverse sample of laboratory and wild mice, and have shown that it varies from 0 to >80 in certain *M. m. domesticus* populations (Figure 4A). In a cohort of outbred mice expected to be hemizygous for an $R2d2^{HC}$ allele from WSB/EiJ (33 diploid copies), we estimate that large deletions, >2 Mb in size, occur at a rate of 3.2% (95% bootstrap C.I. 1.1–6.0%) per generation. This estimate of the mutation rate for CNVs at *R2d2* should be regarded as a lower bound. The power of our copy-number assay to discriminate between copy numbers >∼25 is low, so the assay is much more sensitive to losses than to gains. Even our lower-bound mutation rate exceeds that of the most common recurrent deletions in human (∼1 per 7000 live births) (Turner *et al.* 2007) and is an order of magnitude higher than the most active CNV hotspots described to date in the mouse (Egan *et al.* 2007).

However, the structural mutation rate appears to depend strongly on the diplotype configuration at *R2d2*. As Figure 1D shows, individuals heterozygous for an $R2d2^{HC}$ haplotype and an *R2d2*-null haplotype are in fact hemizygous for several megabases of DNA in *R2d2*. This has important consequences. High mutation rates are observed only in the context of populations in which hemizygosity for $R2d2^{HC}$ is common (Figure 3): highest in the DO, and to a lesser extent in wild *M. m. domesticus* populations harboring both $R2d2^{HC}$ and *R2d2*-null alleles. Homozygosity for $R2d2^{HC}$ is not associated with mutability: in eight recombinant inbred lines from the Collaborative Cross which are homozygous for an $R2d2^{HC}$ haplotype, we observed zero new mutations in at least 400 meioses, through both the male and female germline (8 lines × 2 meioses/generation × 25 or more generations of inbreeding). Sex also appears to have a role in determining the mutation rate at *R2d2*: in a pedigree in which all females were hemizygous for $R2d2^{HC}$, zero new mutations were observed in 1256 meioses (data not shown).

Taken together, these observations hint at a common structural or epigenetic mechanism affecting the resolution of double-strand breaks in large tracts of unpaired (*i.e.*, hemizygous) DNA during male meiosis. At least one other study in mouse has hinted that hemizygous SDs on the sex chromosomes are unstable in intersubspecific hybrids (Scavetta and Tautz 2010). Both the obligate-hemizygous sex chromosomes and large unpaired segments on autosomes are epigenetically marked for transcriptional silencing during male meiotic prophase (van der Laan *et al.* 2004; Baarends *et al.* 2005), and are physically sequestered into a structure called the sex body (Bhattacharyya *et al.* 2013). Repair of double-strand breaks within the sex body is delayed relative to the autosomes (Mahadevaiah *et al.* 2001) and involves a different suite of proteins (Turner *et al.* 2004). We hypothesize that these male-specific pathway(s) may be error-prone in the presence of nonallelic homologous sequences.

However, we cannot exclude the possibility that large-scale rearrangement (such as an inversion) associated with copy-number expansion at *R2d2* contributes to its instability. Physical mapping of the *R2d2* locus in WSB/EiJ is in progress (T. Keane, personal communication) and will shed light on this question.

Origin and distribution of an allele subject to meiotic drive

Females heterozygous for a high- and low-copy allele at *R2d2* preferentially transmit the high-copy allele to progeny via meiotic drive (Didion *et al.* 2015). Meiotic drive can rapidly alter allele frequencies in laboratory and natural populations (Lindholm *et al.* 2016), and we recently showed that $R2d2^{HC}$

alleles sweep through laboratory and natural populations despite reducing the fitness of heterozygous females (Didion *et al.* 2016). These "selfish sweeps" account for the marked reduction in within-population diversity in the vicinity of *R2d2* (Figure 4B).

The present study sheds additional light on the age, origins, and fate of $R2d2^{HC}$ alleles. We find that $R2d2^{HC}$ alleles have a single origin in *M. m. domesticus*. They are present in several different "chromosomal races"—populations fixed for specific Robertsonian translocations between which gene flow is limited (Hauffe and Searle 1993)—indicating that they were likely present at intermediate frequency prior to the origin of the chromosomal races within the past 6000–10,000 years (Nachman *et al.* 1994) and were dispersed through Europe as mice colonized the continent from the south and east (Boursot *et al.* 1993). The presence of $R2d2^{HC}$ in a non-*M. m. domesticus* sample (SPRET/EiJ, *M. spretus* from Cadiz, Spain) is best explained by recent introgression following secondary contact with *M. m. domesticus* (Bonhomme *et al.* 2007; Yang *et al.* 2011).

Additional members of the Cwc22 family

The duplication that gave rise to *R2d2* also created a new copy of *Cwc22*. Based on our assembly of the *R2d2* sequence, the ORF of $Cwc22^{R2d2}$ is intact and encodes a nearly full-length predicted protein that retains the two key functional domains characteristic of the *Cwc22* family. Inspection of RNA-seq data from samples with high copy number at *R2d2* reveals several novel transcript isoforms whose expression appears to be copy number- and tissue-dependent. In testis, the most abundant isoform retains an intron containing an ERV insertion (red arrow in Figure 6), consistent with the well-known transcriptional promiscuity in this tissue. The most abundant isoforms in adult brain is unusual in that its stop codon is in an internal exon which is followed by a 7-kbp 3′ UTR in the terminal exon. Transcripts with a stop codon in an internal exon are generally subject to nonsense-mediated decay (NMD) triggered by the presence of exon-junction complexes downstream of the stop codon. Curiously, *Cwc22* is itself a member of the exon-junction complex (Steckelberg *et al.* 2012).

That an essential gene involved in such a central biochemical pathway should both escape NMD and be overexpressed >10-fold is surprising. Preliminary data from the DO population shows that the $R2d2^{HC}$ allele is associated with elevated levels of both *Cwc22* transcripts and protein in adult liver (Churchill *et al.* 2016). Further studies will be required to determine the distribution of transcription and translation of *Cwc22* across isoforms, tissues, and developmental stages.

Conclusions and future directions

Our detailed analysis of the evolutionary trajectory of *R2d2* provides insight into the fate of duplicated sequences over short (within-species) timescales. The exceptionally high mutation rate and low recombination associated specifically with hemizygous $R2d2^{HC}$ alleles motivate hypotheses regarding the biochemical mechanisms which contribute to observed patterns of polymorphism at this and similar loci. Finally, the birth of a new member of the deeply conserved *Cwc22* gene family in *R2d2* provides an opportunity to test predictions regarding the evolution of young duplicate gene pairs.

Acknowledgments

We thank all the scientists and personnel who collected and processed the wild mouse samples used in this study. In particular we thank Francois Bonhomme for providing samples from wild-derived inbred strains housed at the University of Montpellier Wild Mouse Genetic Repository, and Ted Garland for providing tissue samples from the "high running" selection lines and related crosses. This work was supported by National Institutes of Health grants P50 GM-076468 (F.P.-M.d.V.), U19 AI-100625 (F.P.-M.d.V. and A.P.M.), F30 MH-103925 (A.P.M.), T32 GM-067553 (J.P.D. and A.P.M.) and R21MH09621 (F.P.-M.d.V.); and the Vaadia-Binational Agricultural Research and Development Fund Postdoctoral Fellowship Award FI-12 478-13 to L.Y. Additional support was provided by Cancer Research United Kingdom, the European Research Council, European Molecular Biology Organization Young Investigator Programme (D.T.O.); the European Molecular Biology Laboratory (D.T.O. and P.F.); the Wellcome Trust (WT095908 to P.F. and WT098051 to P.F. and D.T.O.); and finally by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement HEALTH-F4-2010-241504 (EURATRANS).

Literature Cited

Baarends, W. M., E. Wassenaar, R. van der Laan, J. Hoogerbrugge, E. Sleddens-Linkels *et al.*, 2005 Silencing of unpaired chromatin and histone H2A ubiquitination in mammalian meiosis. Mol. Cell. Biol. 25: 1041–1053.

Bailey, J. A., and E. E. Eichler, 2006 Primate segmental duplications: Crucibles of evolution, diversity and disease. Nat. Rev. Genet. 7: 552–564.

Bhattacharyya, T., S. Gregorova, O. Mihola, M. Anger, J. Sebestova *et al.*, 2013 Mechanistic basis of infertility of mouse intersubspecific hybrids. Proc. Natl. Acad. Sci. USA 110: E468–E477.

Bonhomme, F., E. Rivals, A. Orth, G. R. Grant, A. J. Jeffreys *et al.*, 2007 Species-wide distribution of highly polymorphic minisatellite markers suggests past and present genetic exchanges among house mouse subspecies. Genome Biol. 8: R80.

Boursot, P., J. C. Auffray, J. Britton-Davidian, and F. Bonhomme, 1993 The evolution of house mice. Annu. Rev. Ecol. Syst. 24: 119–152.

Boyden, L. M., J. M. Lewis, S. D. Barbee, A. Bas, M. Girardi *et al.*, 2008 Skint1, the prototype of a newly identified immunoglobulin superfamily gene cluster, positively selects epidermal T cells. Nat. Genet. 40: 656–662.

Bray, N. L., H. Pimentel, P. Melsted, and L. Patcher, 2016 Near-optimal RNA-seq quantification. Nat Biotechnol. 34: 525–527.

Chen, J.-M., D. N. Cooper, N. Chuzhanova, C. Férec, and G. P. Patrinos, 2007 Gene conversion: Mechanisms, evolution and human disease. Nat. Rev. Genet. 8: 762–775.

Chevret, P., F. Veyrunes, and J. Britton-Davidian, 2005 Molecular phylogeny of the genus Mus (Rodentia: Murinae) based on mitochondrial and nuclear data. Biol. J. Linn. Soc. Lond. 84: 417–427.

Chick, J. M., S. C. Munger, P. Simecek, E. L. Huttlin, K. Choi *et al.*, 2016 Defining the consequences of genetic variation on a proteome-wide scale. Nature 534: 500–505.

Cole, F., S. Keeney, and M. Jasin, 2010 Comprehensive, fine-scale dissection of homologous recombination outcomes at a hot spot in mouse meiosis. Mol. Cell 39: 700–710.

Cole, F., F. Baudat, C. Grey, S. Keeney, B. de Massy *et al.*, 2014 Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics. Nat. Genet. 46: 1072–1080.

Collaborative Cross Consortium, 2012 The genome architecture of the Collaborative Cross mouse menetic reference population. Genetics 190: 389–401.

Cox, A., C. L. Ackert-Bicknell, B. L. Dumont, Y. Ding, J. T. Bell *et al.*, 2009 A new standard genetic map for the laboratory mouse. Genetics 182: 1335–1344.

Crowley, J. J., V. Zhabotynsky, W. Sun, S. Huang, I. K. Pakatci *et al.*, 2015 Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nat. Genet. 47: 353–360.

Cunningham, F., M. R. Amode, D. Barrell, K. Beal, K. Billis *et al.*, 2014 Ensembl 2015. Nucleic Acids Res. 43: D662–D669.

Dallas, J. F., 1992 Estimation of microsatellite mutation rates in recombinant inbred strains of mouse. Mamm. Genome 3: 452–456.

Demuth, J. P., T. De Bie, J. E. Stajich, N. Christianini, and M. W. Hahn, 2006 The evolution of mammalian gene families. PLoS One 1: e85.

Didion, J. P., A. P. Morgan, A. M.-F. Clayshulte, R. C. Mcmullan, L. Yadgary *et al.*, 2015 A multi-megabase copy number gain causes maternal transmission ratio distortion on mouse chromosome 2. PLoS Genet. 11: e1004850.

Didion, J. P., A. P. Morgan, L. Yadgary, T. A. Bell, R. C. McMullan *et al.*, 2016 R2d2 drives selfish sweeps in the house mouse. Mol. Biol. Evol. 33: 1381–1395.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2012 STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Dopman, E. B., and D. L. Hartl, 2007 A portrait of copy-number polymorphism in drosophila melanogaster. Proc. Natl. Acad. Sci. USA 104: 19920–19925.

Dover, G., 1982 Molecular drive: A cohesive mode of species evolution. Nature 299: 111–117.

Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut, 2012 Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29: 1969–1973.

Dumont, B. L., and E. E. Eichler, 2013 Signals of historical interlocus gene conversion in human segmental duplications. PLoS One 8: e75949.

Edgar, R. C., 2004 MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792–1797.

Egan, C. M., S. Sridhar, M. Wigler, and I. M. Hall, 2007 Recurrent DNA copy number variation in the laboratory mouse. Nat. Genet. 39: 1384–1389.

Eickbush, T. H., and D. G. Eickbush, 2007 Finely orchestrated movements: Evolution of the ribosomal RNA genes. Genetics 175: 477–485.

Faust, G. G., and I. M. Hall, 2014 SAMBLASTER: Fast duplicate marking and structural variant read extraction. Bioinformatics 30: 2503–2505.

Garrison, E., and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907.

Goodman, M., J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda, 1979 Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Biol. 28: 132–163.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson *et al.*, 2011 Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 29: 644–652.

Hallast, P., P. Balaresque, G. R. Bowden, S. Ballereau, and M. A. Jobling, 2013 Recombination dynamics of a human Y-chromosomal palindrome: Rapid GC-biased gene conversion, multi-kilobase conversion tracts, and rare inversions. PLoS Genet. 9: e1003666.

Halligan, D. L., A. Kousathanas, R. W. Ness, B. Harr, L. Eöry *et al.*, 2013 Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. PLoS Genet. 9: e1003995.

Handsaker, R. E., V. V. Doren, J. R. Berman, G. Genovese, S. Kashin *et al.*, 2015 Large multiallelic copy number variations in humans. Nat. Genet. 47: 296–303.

Hauffe, H. C., and J. B. Searle, 1993 Extreme karyotypic variation in a Mus musculus domesticus hybrid zone: The tobacco mouse story revisited. Evolution 47: 1374–1395.

Holt, J., and L. McMillan, 2014 Merging of multi-string BWTs with applications. Bioinformatics 30: 3524–3531.

Hurles, M., 2002 Are 100,000 "SNPs" useless? Science 298: 1509.

Hurst, J. L., C. E. Payne, C. M. Nevison, A. D. Marie, R. E. Humphries *et al.*, 2001 Individual recognition in mice mediated by major urinary proteins. Nature 414: 631–634.

Keane, T. M., L. Goodstadt, P. Danecek, M. A. White, K. Wong *et al.*, 2011 Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477: 289–294.

Kimura, M., and T. Ohta, 1968 The average number of mutations until fixation of a mutant gene in a finite population. Genetics 61: 763–771.

Koonin, E. V., 2005 Orthologs, paralogs and evolutionary genomics. Annu. Rev. Genet. 39: 309–338.

Lander, E. S., and M. S. Waterman, 1988 Genomic mapping by fingerprinting random clones: A mathematical analysis. Genomics 2: 231–239.

Li, H., 2006 TreeFam: A curated database of phylogenetic trees of animal gene families. Nucleic Acids Res. 34: D572–D580.

Li H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

Liao, D., T. Pavelitz, J. R. Kidd, K. K. Kidd, and A. M. Weiner, 1997 Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the RUN2 locus) involves rapid intrachromosomal homogenization and rare interchromosomal gene conversion. EMBO J. 16: 588–598.

Lindholm, A. K., K. A. Dyer, R. C. Firman, L. Fishman, W. Forstmeier *et al.*, 2016 The ecology and evolutionary dynamics of meiotic drive. Trends Ecol. Evol. 31: 315–326.

Liu, E. Y., A. P. Morgan, E. J. Chesler, W. Wang, G. A. Churchill *et al.*, 2014 High-resolution sex-specific linkage maps of the mouse reveal polarized distribution of crossovers in male germline. Genetics 197: 91–106.

Lynch, M., and J. S. Conery, 2000 The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.

Mahadevaiah, S. K., J. M. Turner, F. Baudat, E. P. Rogakou, P. de Boer *et al.*, 2001 Recombinational DNA double-strand breaks in mice precede synapsis. Nat. Genet. 27: 271–276.

Melamed, C., and M. Kupiec, 1992 Effect of donor copy number on the rate of gene conversion in the yeast Saccharomyces cerevisiae. Mol. Genet. Genomics 235: 97–103.

Moran, P., 1975 Wandering distributions and the electrophoretic profile. Theor. Popul. Biol. 8: 318–330.

Morgan, A. P., C.-P. Fu, C.-Y. Kao, C. E. Welsh, and J. P. Didion *et al.*, 2015 The Mouse Universal Genotyping Array: from Substrains to Subspecies. G3 6: 236–279.

Muffato, M., A. Louis, C. E. Poisnel, and H. R. Crollius, 2010 Genomicus: A database and a browser to study gene synteny in modern and ancestral genomes. Bioinformatics 26: 1119–1121.
Nachman, M. W., S. N. Boyer, J. B. Searle, and C. F. Aquadro, 1994 Mitochondrial DNA variation and the evolution of robertsonian chromosomal races of house mice, Mus domesticus. Genetics 136: 1105–1120.
Nagylaki, T., and T. D. Petes, 1982 Intrachromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. Genetics 100: 315–337.
Pamilo, P., and M. Nei, 1988 Relationships between gene trees and species trees. Mol. Biol. Evol. 5: 568–583.
Pei, B., C. Sisu, A. Frankish, C. Howald, L. Habegger *et al.*, 2012 The GENCODE pseudogene resource. Genome Biol. 13: R51.
Pezer, Ž, B Harr, M Teschke, H Babiker, and D Tautz, 2015 Divergence patterns of genic copy number variation in natural populations of the house mouse (Mus musculus domesticus) reveal three conserved genes with major population-specific expansions. Genome Res. 25: 1114–1124.
Phifer-Rixey, M., M. Bomhoff, and M. W. Nachman, 2014 Genome-wide patterns of differentiation among house mouse subspecies. Genetics 198: 283–297.
Rozen, S., H. Skaletsky, J. D. Marszalek, P. J. Minx, H. S. Cordum *et al.*, 2003 Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature 423: 873–876.
Salcedo, T., A. Geraldes, and M. W. Nachman, 2007 Nucleotide variation in wild and inbred mice. Genetics 177: 2277–2291.
Sambrook, J., and D. W. Russell (Editors), 2006 *Molecular cloning: A laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Scavetta, R. J., and D. Tautz, 2010 Copy number changes of CNV regions in intersubspecific crosses of the house mouse. Mol. Biol. Evol. 27: 1845–1856.
Schindelhauer, D., 2002 Evidence for a fast, intrachromosomal conversion mechanism from mapping of nucleotide variants within a homogeneous alpha-satellite DNA array. Genome Res. 12: 1815–1826.
She, X., Z. Cheng, S. Zöllner, D. M. Church, and E. E. Eichler, 2008 Mouse segmental duplication and copy number variation. Nat. Genet. 40: 909–914.
Shi, J., S. E. Wolf, J. M. Burke, G. G. Presting, J. Ross-Ibarra *et al.*, 2010 Widespread gene conversion in centromere cores. PLoS Biol. 8: e1000327.
Soh, Y. Q., J. Alföldi, T. Pyntikova, L. Brown, and T. Graves *et al.*, 2014 Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. Cell 159: 800–813.
Stamatakis, A., 2014 RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313.
Stankiewicz, P., and J. R. Lupski, 2002 Genome architecture, rearrangements and genomic disorders. Trends Genet. 18: 74–82.
Steckelberg, A.-L., V. Boehm, A. Gromadzka, and N. Gehring, 2012 Cwc22 connects pre-mRNA splicing and exon junction complex assembly. Cell Reports 2: 454–461.
Suzuki, H., T. Shimada, M. Terashima, K. Tsuchiya, and K. Aplin, 2004 Temporal, spatial, and ecological modes of evolution of Eurasian Mus based on mitochondrial and nuclear gene sequences. Mol. Phylogenet. Evol. 33: 626–646.
Svenson, K. L., D. M. Gatti, W. Valdar, C. E. Welsh, R. Cheng *et al.*, 2012 High-resolution genetic mapping using the mouse Diversity Outbred population. Genetics 190: 437–447.
Swallow, J. G., P. A. Carter, and T Garland, Jr., 1998 Artificial selection for increased wheel-running behavior in house mice. Behav. Genet. 28: 227–237.
Treangen, T. J., and S. L. Salzberg, 2011 Repetitive DNA and next-generation sequencing: Computational challenges and solutions. Nat. Rev. Genet. 13: 36–46.
Turner, J. M., O. Aprelikova, X. Xu, R. Wang, S. Kim *et al.*, 2004 BRCA1, histone H2AX phosphorylation, and male meiotic sex chromosome inactivation. Curr. Biol. 14: 2135–2142.
Turner, D. J., M. Miretti, D. Rajan, H. Fiegler, N. P. Carter *et al.*, 2007 Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. Nat. Genet. 40: 90–95.
van der Laan, R., E. J. Uringa, E. Wassenaar, J. W. Hoogerbrugge, E. Sleddens *et al*., 2004 Ubiquitin ligase Rad18Sc localizes to the XY body and to other chromosomal regions that are unpaired and transcriptionally silenced during male meiotic prophase. J. Cell Sci. 117: 5023–5033.
Waterston, R. H., A. T. Chinwalla, L. L. Cook, K. D. Delehaunty, G. A. Fewell *et al.*, 2002 Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.
Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256–276.
White, M. A., C. Ané, C. N. Dewey, B. R. Larget, and B. A. Payseur, 2009 Fine-scale phylogenetic discordance across the house mouse genome. PLoS Genet. 5: e1000729.
Yang, H., T. A. Bell, G. A. Churchill, and F. Pardo-Manuel de Villena, 2007 On the subspecific origin of the laboratory mouse. Nat. Genet. 39: 1100–1107.
Yang, H., J. R. Wang, J. P. Didion, R. J. Buus, T. A. Bell *et al.*, 2011 Subspecific origin and haplotype diversity in the laboratory mouse. Nat. Genet. 43: 648–655.
Yeh, T.-C., H.-L. Liu, C.-S. Chung, N.-Y. Wu, Y.-C. Liu *et al.*, 2010 Splicing factor Cwc22 is required for the function of Prp2 and for the spliceosome to escape from a futile pathway. Mol. Cell. Biol. 31: 43–53.

Communicating editor: B. A. Payseur

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.191007/-/DC1

The Evolutionary Fates of a Large Segmental Duplication in Mouse

Andrew P. Morgan, J. Matthew Holt, Rachel C. McMullan, Timothy A. Bell, Amelia M.-F. Clayshulte, John P. Didion, Liran Yadgary, David Thybert, Duncan T. Odom, Paul Flicek, Leonard McMillan, and Fernando Pardo-Manuel de Villena

DOI: 10.1534/genetics.116.191007

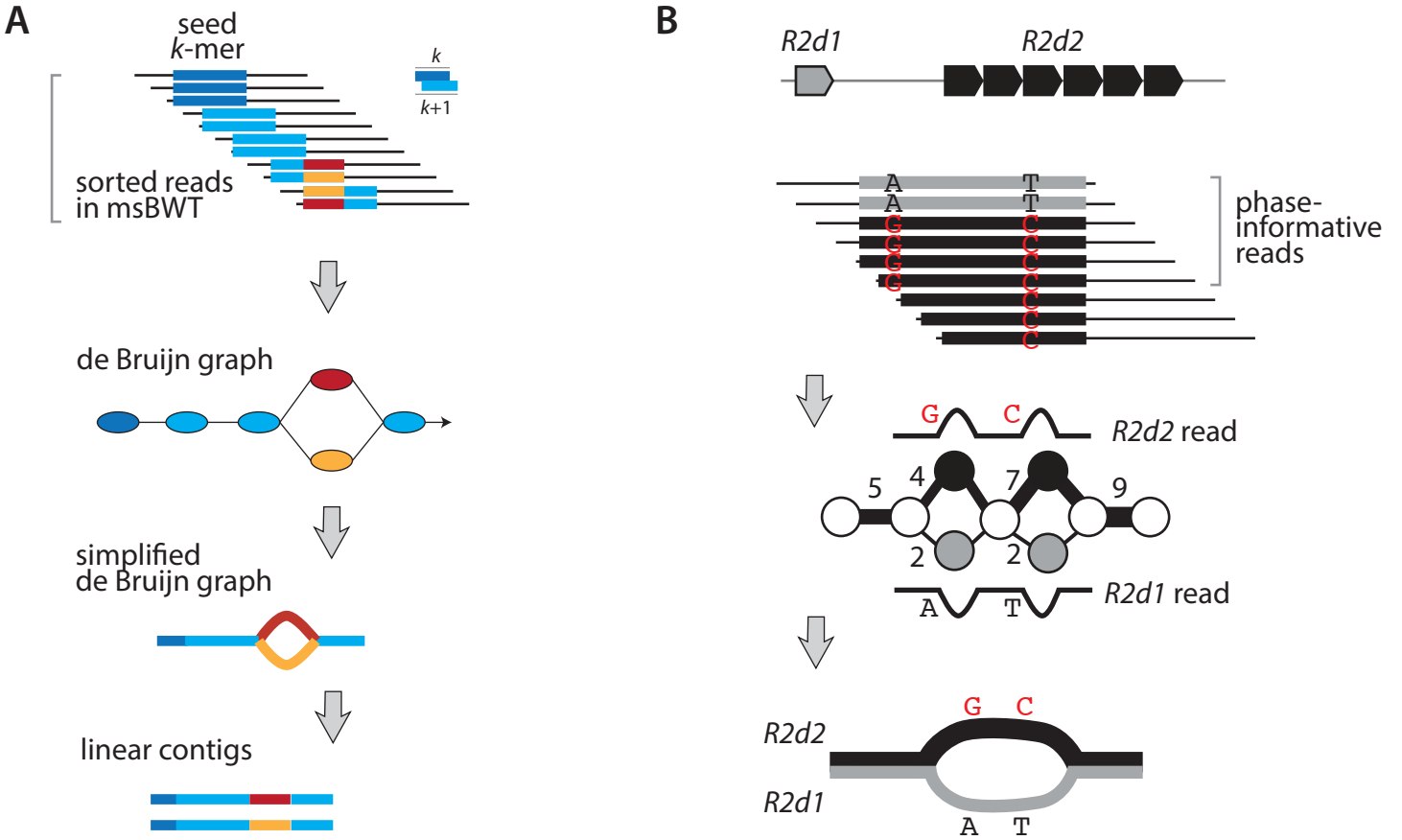
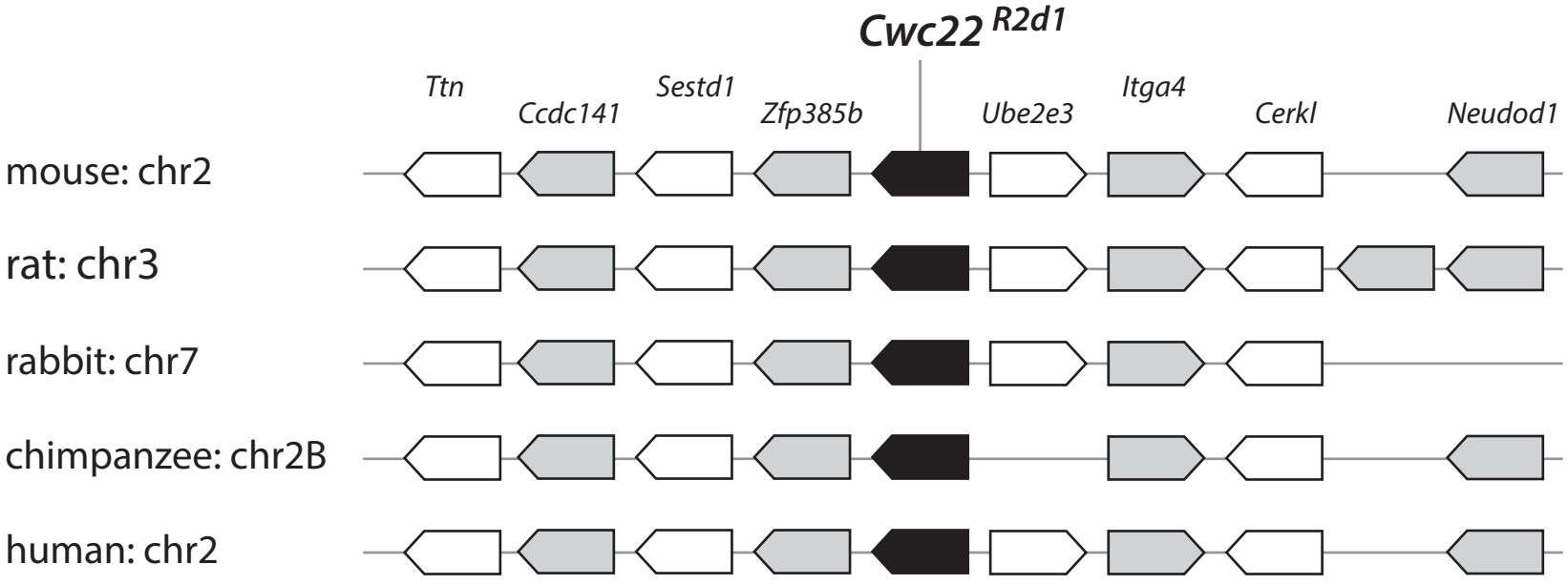

A
seed
k-mer
k
k+1
sorted reads
in msBWT
de Bruijn graph
simplified
de Bruijn graph
linear contigs
B
R2d1
R2d2
phase-
informative
reads
G
C
R2d2 read
5
4
7
9
2
2
A
T
R2d1 read
R2d2
R2d1
G
C
A
T

Figure S1. Targeted de novo assembly using the multi-string Burrows-Wheeler Transform (msBWT). (A) The msBWT and its associated FM-index implicitly represent a suffix array of sequencing reads, such that read suffixes sharing a k-mer prefix are adjacent in the data structure. This allows rapid construction of a local de Bruijn graph starting from a k-mer seed (dark blue) and extending by successive k-mers (light blue) containing the (k-1)-length suffix of the previous k-mer. A (k-1)-length prefix with more than one possible suffix (red and orange) creates a branch point. Adjacent nodes in the graph with in-degree and out-degree one can be collapsed into a single node, yielding a simplified graph, which can then be traversed to obtain linear contig(s). (B) Paralogs of R2d can be disentangled using the local de Bruijn graph by exploiting differences in copy number. Edges in the graph are weighted by read count, and linear contigs for the R2d1 and R2d2 paralogs obtained by traversing the graph in a manner that minimizes the variance in edge weights along possible paths. Phase-informative reads (those overlapping multiple paralogous variants) provide a second source of evidence.


Cwc22 R2d1
Ttn
Ccdc141
Sestd1
Zfp385b
Ube2e3
Itga4
Cerkl
Neudod1
mouse: chr2
rat: chr3
rabbit: chr7
chimpanzee: chr2B
human: chr2

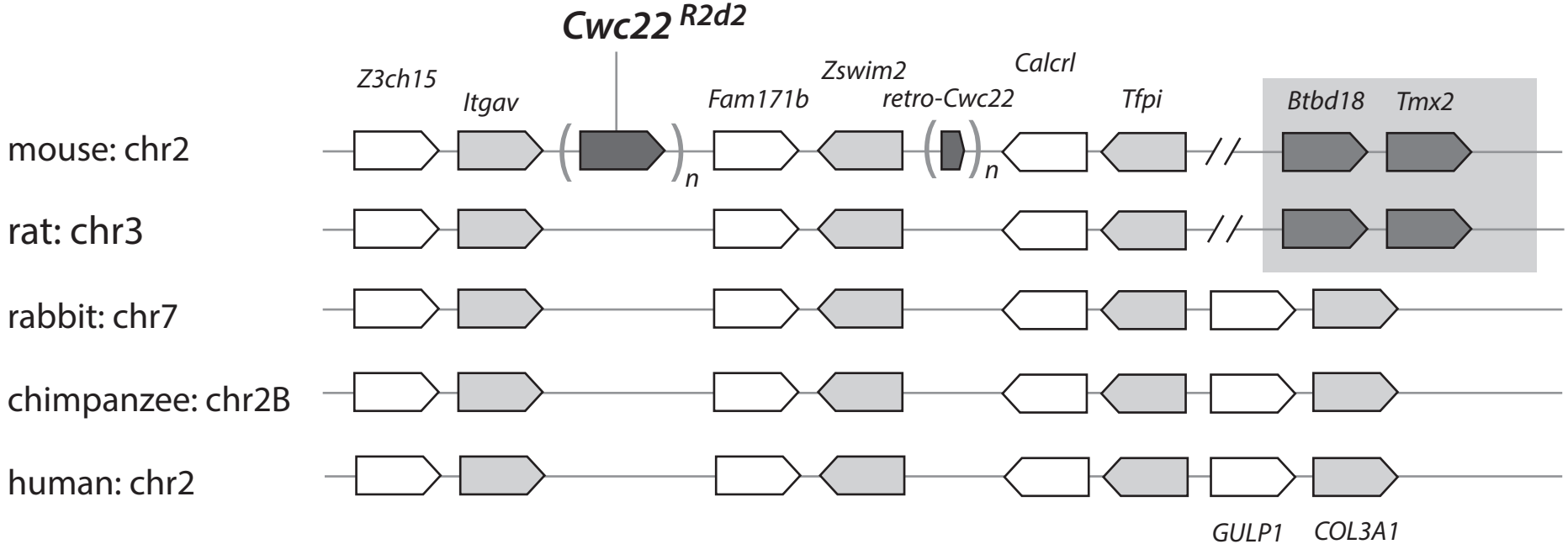

Cwc22 R2d2
Z3ch15
Itgav
Zswim2
Fam171b
retro-Cwc22
Calcrl
Tfpi
Btbd18
Tmx2
n
n
mouse: chr2
rat: chr3
rabbit: chr7
chimpanzee: chr2B
human: chr2
GULP1
COL3A1

Figure S2. Conservation of synteny between mouse and four other mammals around Cwc22R2d1 (upper panel) indicates that the R2d1 sequence remains in its ancestral position. Chevrons represent genes, alternating white and grey, and are oriented according to the strand on which the gene is encoded. Cwc22R2d2 is novel in the mouse but its position relative to genes with conserved order is shown in the lower panel. Note that synteny is disrupted in mouse and rat distal to R2d2.

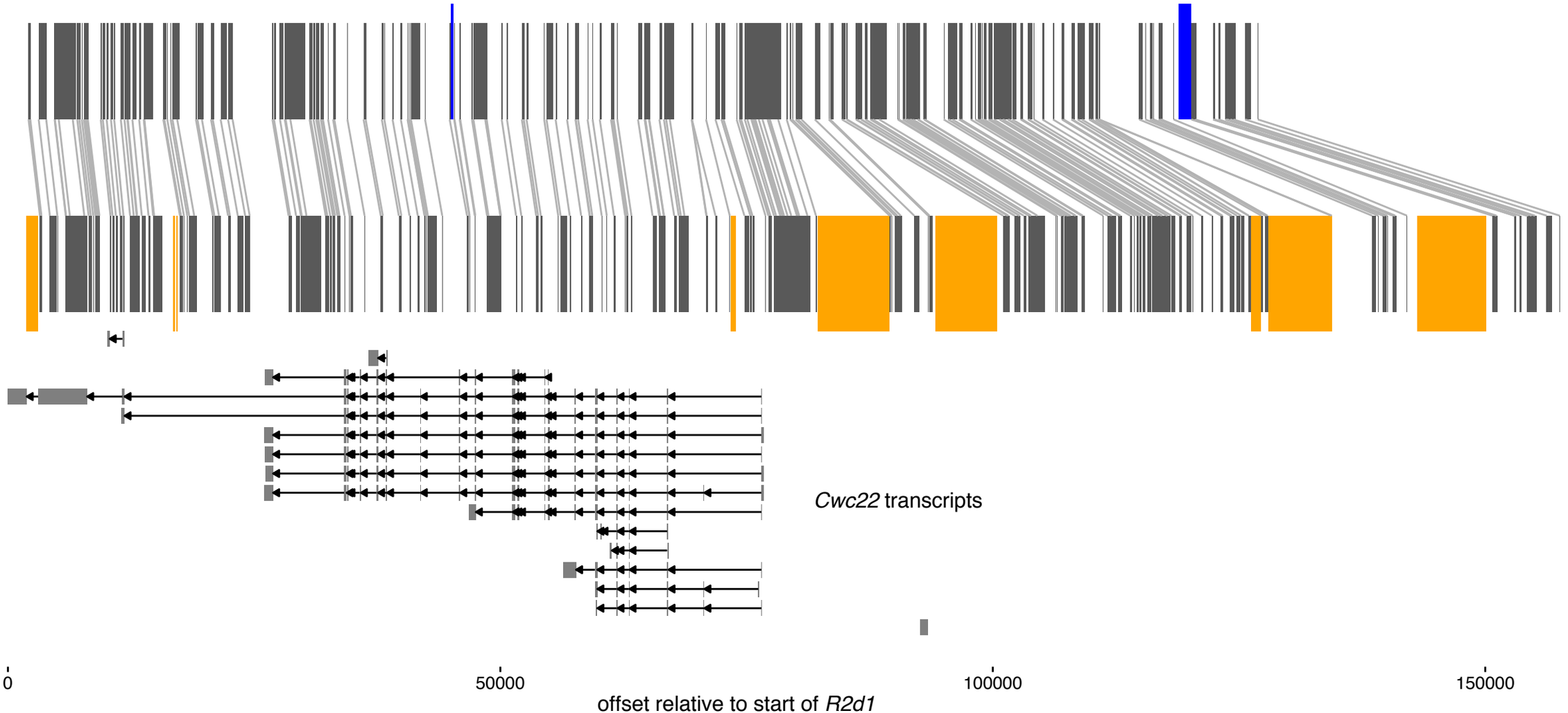

Cwc22 transcripts
0
50000
100000
150000
offset relative to start of R2d1

Figure S3. Pairwise alignment of R2d2 contig (top) to the R2d1 reference sequence (bottom). Dark boxes show position of repetitive elements present in both sequences; syntenic positions are connected by grey anchors, and blank space represents aligned bases in both sequences. Orange boxes indicate position of repetitive elements present in the R2d1 sequence but not detected in R2d2; blue boxes indicate position of elements in R2d2 but not R2d1. Cwc22 transcripts are shown below the alignment.

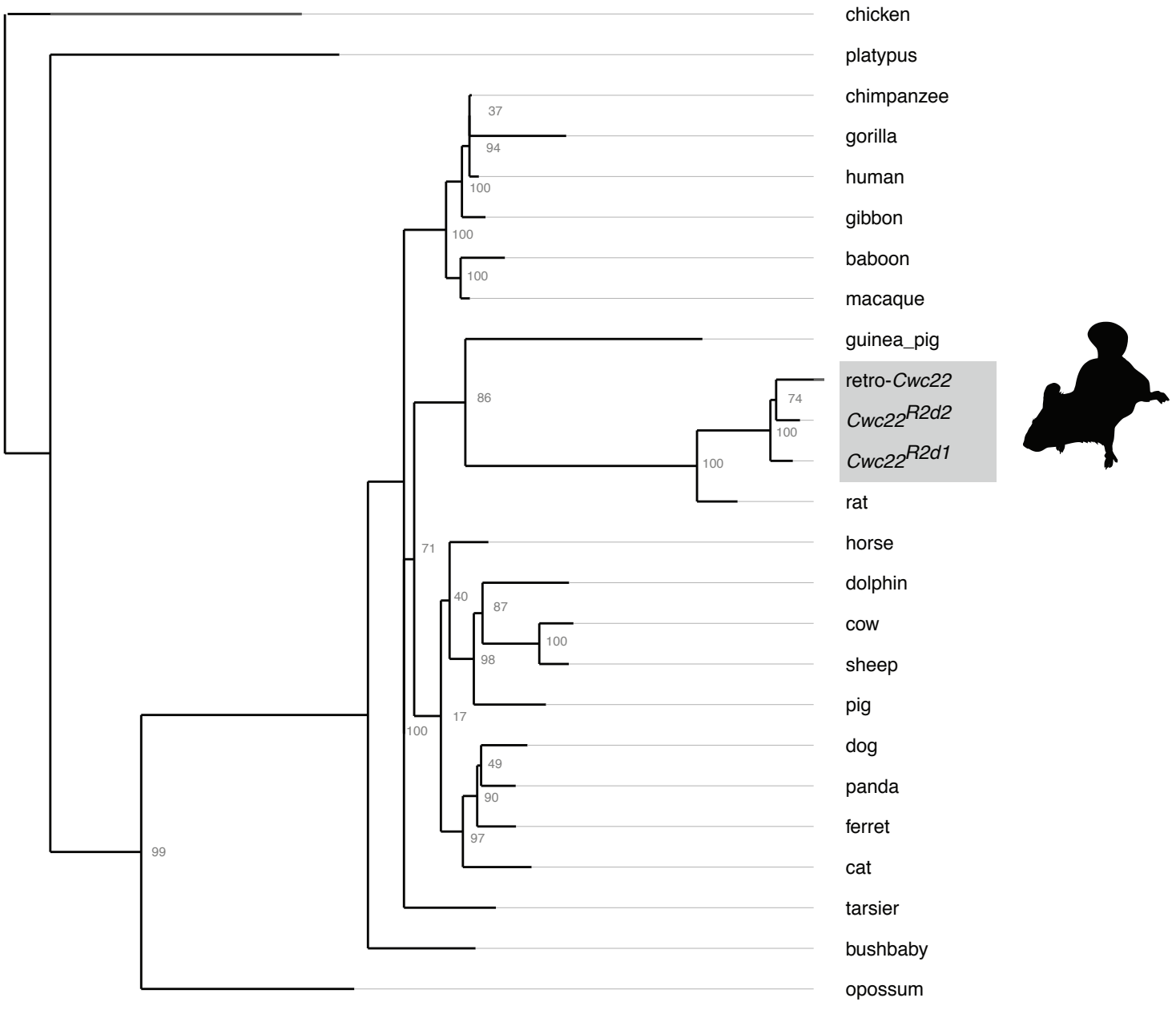

chicken
platypus
chimpanzee
gorilla
human
gibbon
baboon
macaque
guinea_pig
retro-Cwc22
Cwc22R2d2
Cwc22R2d1
rat
horse
dolphin
cow
sheep
pig
dog
panda
ferret
cat
tarsier
bushbaby
opossum
37
94
100
100
100
86
74
100
100
71
40
87
100
98
17
100
49
90
97
99

Figure S4. Phylogenetic tree constructed from amino acid sequences for mammalian Cwc22 homologs (including all three mouse paralogs) with chicken as an outgroup. Node labels indicate support in 100 bootstrap replicates.

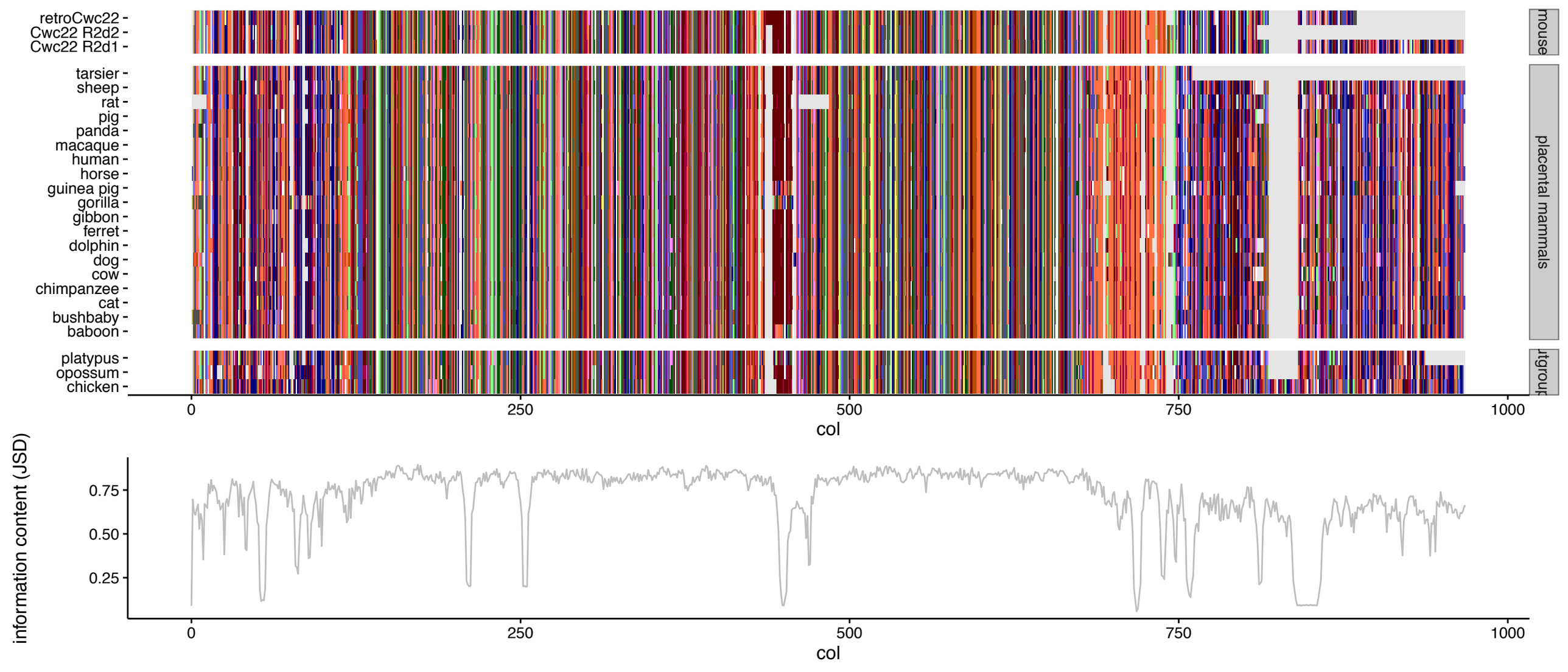

retroCwc22
Cwc22 R2d2
Cwc22 R2d1
tarsier
sheep
rat
pig
panda
macaque
human
horse
guinea pig
gorilla
gibbon
ferret
dolphin
dog
cow
chimpanzee
cat
bushbaby
baboon
platypus
opossum
chicken
mouse
placental mammals
outgroup
0
250
500
750
1000
col
information content (JSD)
0.75
0.50
0.25
0
250
500
750
1000
col

Figure S5. Alignment of amino acid sequences from mouse Cwc22R2d1, Cwc22R2d2 and retro-Cwc22, plus Cwc22 orthologs from 19 other placental mammals plus opossum, platypus and chicken as outgroups. Residues are colored according to biochemical properties and gaps are shown in grey. Information content of each column in the alignment, measured as the Jenson-Shannon divergence, is plotted in the lower panel.

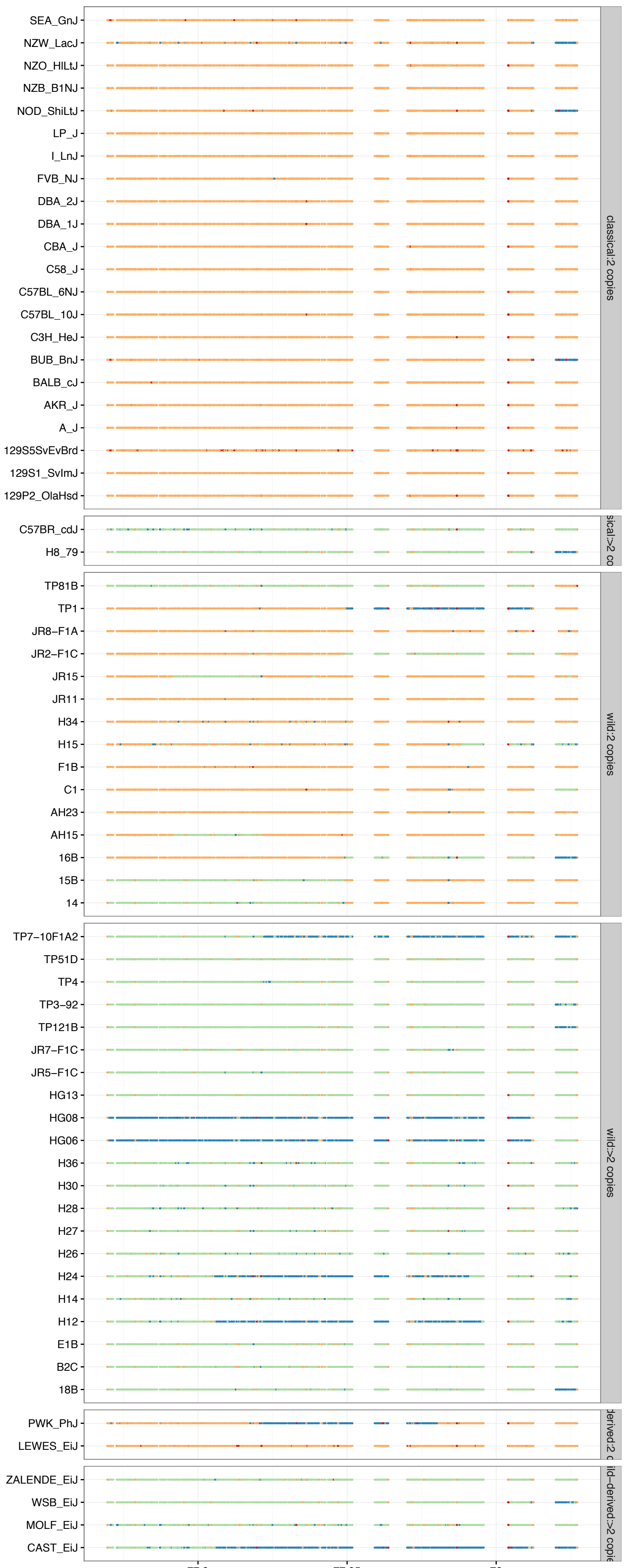

SEA_GnJ
NZW_LacJ
NZO_HlLtJ
NZB_B1NJ
NOD_ShiLtJ
LP_J
I_LnJ
FVB_NJ
DBA_2J
DBA_1J
CBA_J
C58_J
C57BL_6NJ
C57BL_10J
C3H_HeJ
BUB_BnJ
BALB_cJ
AKR_J
A_J
129S5SvEvBrd
129S1_SvImJ
129P2_OlaHsd
classical:2 copies
C57BR_cdJ
H8_79
sical:>2 co
TP81B
TP1
JR8−F1A
JR2−F1C
JR15
JR11
H34
H15
F1B
C1
AH23
AH15
16B
15B
14
wild:2 copies
TP7−10F1A2
TP51D
TP4
TP3−92
TP121B
JR7−F1C
JR5−F1C
HG13
HG08
HG06
H36
H30
H28
H27
H26
H24
H14
H12
E1B
B2C
18B
wild:>2 copies
PWK_PhJ
LEWES_EiJ
derived:2 c
ZALENDE_EiJ
WSB_EiJ
MOLF_EiJ
CAST_EiJ
ild−derived:>2 copie
77.9
77.95
78
allelotype
0:0
1:0
1:1
0:1

Figure S6. Diagnostic variants used to identify gene conversion tracts between R2d2 and R2d1. Each column represents a single variant (total of 1,411) between R2d1 and R2d2 for which R2d2 has the derived allele, and each row a single individual with whole-genome sequence data available (named in Table S1). Points are colored according to the "allelotype" of each variant detected in 8 or more reads in each sample: 0:0 (red), neither R2d1 nor R2d2 allele present; 1:0, R2d1 but not R2d2 (orange); 1:1, both R2d1 and R2d2 (green); 0:1, R2d2 but not R2d1 (blue). Individuals are grouped according to diploid R2d copy number.

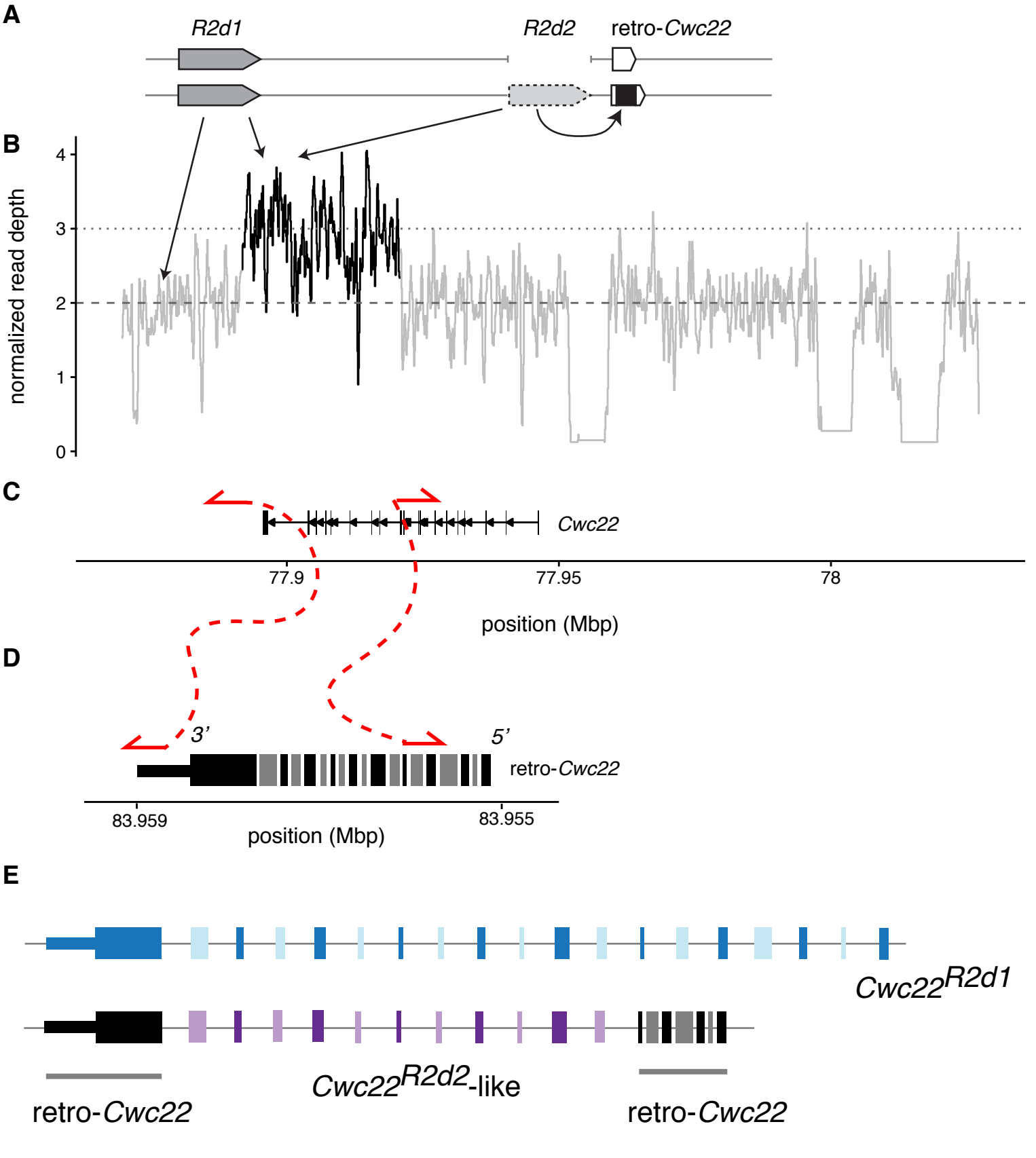

A
R2d1
R2d2
retro-Cwc22
B
normalized read depth
4
3
2
1
0
C
Cwc22
77.9
77.95
78
position (Mbp)
D
3'
5'
retro-Cwc22
83.959
83.955
position (Mbp)
E
Cwc22R2d1
Cwc22R2d2-like
retro-Cwc22
retro-Cwc22

Figure S7. Partial loss of R2d2 with structural rearrangement. (A) Inferred structure of the R2d1-R2d2 region in IR:AHZ_STND:015, a wild M. m. domesticus individual from Iran. R2d1 is present on both chromosomes but only a fragment of R2d2 remains on one chromosome, and it has been transposed into the retro-Cwc22 array. (B) Normalized depth of coverage (2 = normal diploid level) across R2d. Regions in grey represent reads from R2d1 alone, while region in black captures reads from R2d1 and R2d2, as shown by arrows from panel A. (C) Position of read pairs (red; not drawn to scale) with soft-clipped alignments to R2d1. The proximal read aligns in the 3' UTR of Cwc22, and the distal read across an exon-intron boundary within the gene body. Note the "outward"-facing direction of the alignments. (D) Positions of the mates of the reads in panel C. Note that the x-axis is reversed so that the exons of retro-Cwc22 (encoded on the plus strand) parallel those of Cwc22 (encoded on the minus strand). The 3' read maps across the boundary of th 3' UTR of Cwc22 and the ERV mediating the retrotransposition event. The 5' read maps across two exon-exon boundaries in retro-Cwc22, so there is no ambiguity regarding its alignment to the retro-transposed copy. (E) Inferred structure of Cwc22 paralogs in this sample. Note that one of the copies of retro-Cwc22 is now a mosaic of retrotransposed and Cwc22R2d2-derived sequence.

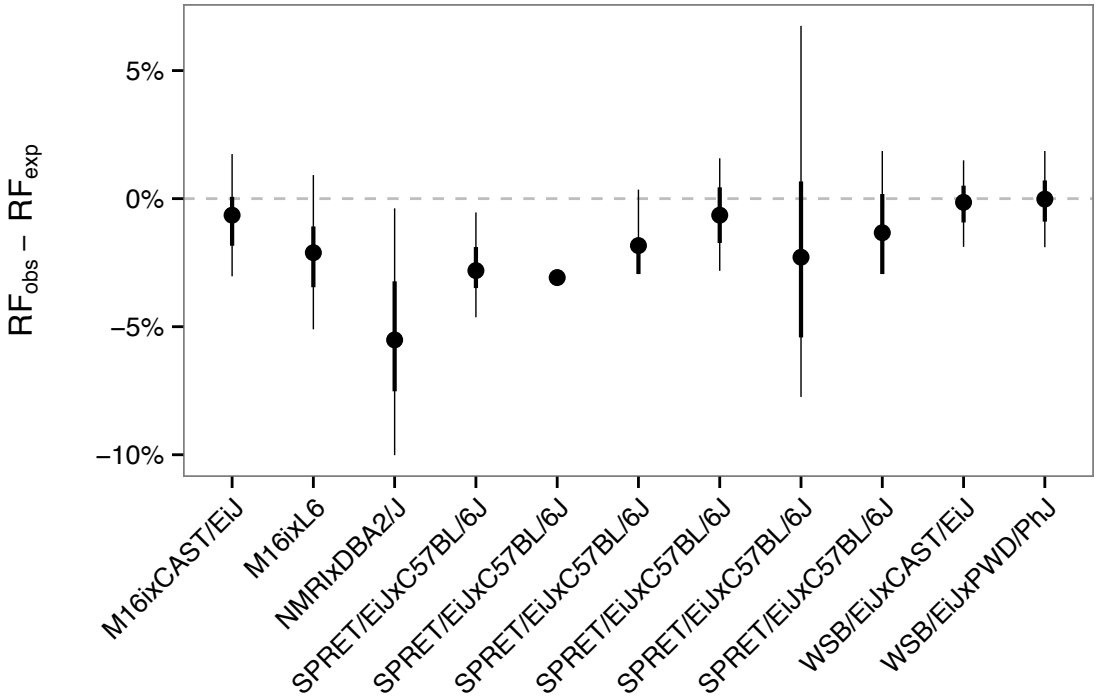

5%
0%
−5%
−10%
RF_obs − RF_exp
M16ixCAST/EiJ
M16ixL6
NMRIxDBA2/J
SPRET/EiJxC57BL/6J
SPRET/EiJxC57BL/6J
SPRET/EiJxC57BL/6J
SPRET/EiJxC57BL/6J
SPRET/EiJxC57BL/6J
SPRET/EiJxC57BL/6J
WSB/EiJxCAST/EiJ
WSB/EiJxPWD/PhJ

Figure S8. Difference between expected and observed recombination fraction between markers flanking R2d2 in experimental crosses in which at least one parent is segregating for a high-copy allele of R2d2. Thick and thin vertical bars show 90% and 95% confidence bounds, respectively, obtained by non-parametric bootstrap.

Table S1. List of mouse samples used in this study, with their taxonomic designation, geographic origin, karyotype (STND, standard: 2n = 40; all others chromosomal races with Robertsonian translocations) and *R2d2* copy-number classification. (.xlsx, 136 KB)

Available for download as a .xlsx file at:

http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.191007/-/DC1/TableS1.xlsx

Table S2. Regions of *R2d* targeted for *de novo* assembly in inbred strains. (.xlsx, 41 KB)

Available for download as a .xlsx file at:

http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.191007/-/DC1/TableS2.xlsx

Table S3. Transposable-element insertions private to *R2d1* or *R2d2*. Coordinates are offsets with respect to the start position of *R2d* (for *R2d1*: chr2: 77,869,657 in the reference genome; for *R2d2*: the beginning of the *de novo* assembled contig in **File S1**.) (.xlsx, 43 KB)

Available for download as a .xlsx file at:

http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.191007/-/DC1/TableS3.xlsx

Table S4. Frequency table of copy-number status by geographic origin for wild-caught and wild-derived Mus musculus individuals used in this study, stratified by subspecies. “Europe/Mediterranean” includes continental Europe, the United Kingdom and countries in the Mediterranean basin (Tunisia, Cyprus, Israel). “Asia” includes Asia, the Middle East and countries in the Indian Ocean basin (Madagascar). (.xlsx, 35 KB)

Available for download as a .xlsx file at:

http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.191007/-/DC1/TableS4.xlsx

Table S5. Individuals from the Diversity Outbred population carrying *de novo* copy-number mutations at *R2d2*. Each was expected to be heterozygous for the WSB/EiJ allele (33 haploid copies). (.xlsx, 39 KB)

Available for download as a .xlsx file at:

http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.191007/-/DC1/TableS5.xlsx

File S1. Compressed archive containining *R2d2* contig (from WSB/EiJ) and multiple sequence alignments from selected regions in **Table S2**. (.zip, 60 KB)

Available for download as a .zip file at:

http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.191007/-/DC1/FileS1.zip

File S2. Genotypes from the Mouse Diversity Array (files mda.*) and MegaMUGA array (files mm.*) in PLINK format. (.zip, 378 KB)

Available for download as a .zip file at:

http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.191007/-/DC1/FileS2.zip

File S3. *Cwc22* transcript models and sequences used in **Figure 6**. (.zip, 16 KB)

Available for download as a .zip file at:

http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.191007/-/DC1/FileS3.zip

File S4. Nucleotide and amino acid sequences from mammalian orthologs of mouse *Cwc22*. (.zip, 20 KB)

Available for download as a .zip file at:

http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.191007/-/DC1/FileS4.zip