The many evolutionary fates of a large segmental duplication in mouse

Andrew P Morgan[1], J Matthew Holt[2], Rachel C McMullan[1], Timothy A Bell[1], Amelia M-F Clayshulte[1], John P Didion[1], Liran Yadgary[1], David Thybert[3], Duncan T Odom[4,5], Paul Flicek[3,5], Leonard McMillan[2], Fernando Pardo-Manuel de Villena[1]

[1] Department of Genetics, Carolina Center for Genome Sciences and Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC

[2] Department of Computer Science, University of North Carolina, Chapel Hill, NC

[3] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

[4] University of Cambridge, Cancer Research UK Cambridge Institute, Robinson Way, Cambridge, CB2 0RE, United Kindgom

[5] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom

Corresponding author:

Fernando Pardo-Manuel de Villena

5049 Genetic Medicine Building

120 Mason Farm Road CB#7264

Chapel Hill, NC 27599-7264

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*). *De novo* assembly of both the ancestral (*R2d1*) and the derived (*R2d2*) copies reveals that they have been subject to non-allelic gene conversion events spanning tens of kilobases. *R2d2* is also a hotspot for structural variation: its diploid copy number ranges from 0 in the mouse reference genome to more than 80 in wild mice sampled from around the globe. Heterozygosity for low- and high-copy 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 generation in laboratory populations. We identify an additional 57 loci, covering 0.8% of the mouse genome, that have characteristics similar to *R2d2*: segmental duplication, phylogenetic discordance and low recombination rate. 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.

INTRODUCTION

Duplication is an important force shaping the evolution of plant and animal genomes: redundant sequence provides both a substrate for evolution and transient relief from selective pressure (Lynch 2000). Like any sequence variant, a duplication first arises in a single individual in a population. Such polymorphisms are commonly called copy-number variants (CNVs) while segregating in the population, and segmental duplications (SDs) once fixed (Chain *et al.* 2014). The distinction between CNVs and SDs is 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), indicating that duplicated regions are hotspots for ongoing copy-number variation (CNV). Mutation rates for CNVs -- both gains and losses -- in SD-rich regions may be orders of magnitude higher than in unique sequence (Egan *et al.* 2007). This inherent instability is associated with speciation: in both the primate (Bailey and Eichler 2006) and mouse lineages, bursts of segmental duplication have preceded dramatic species radiations. Likewise, blocks of conserved synteny in mammals frequently terminate at SDs (Bailey and Eichler 2006), suggesting that SDs mediate the chromosomal rearrangements through which karyotypes diverge and reproductive barriers arise. Characterization of this "churning" in SD-rich regions is key to understanding the evolution of genome structure and function.

Notwithstanding their evolutionary importance, SDs are difficult to analyze. Repeated sequences with period longer than the insert size in a sequencing library and 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 requires specialized software (Treangen and Salzberg 2011). Genotyping of sites within SDs is difficult because variants between copies (that is, paralogous variants) are easily confounded with variants within copies between individuals (that is, allelic variants.) Inability to distinguish allelic from paralogous variation similarly hampers any analysis that requires a positional ordering of sites along the genome -- for instance, delineation of haplotype blocks.

Paralogy also complicates phylogenetic inference. Ancestral duplication followed by differential losses along separate lineages yields 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 non-allelic gene conversion between copies (Nagylaki and Petes 1982) -- despite the fact that such loci have reduced rates of crossing-over (Liu *et al.* 2014). The extremely high rate of gene conversion between palindromic repeats in the male-specific regions of mouse and human Y-chromosomes provide a dramatic example of this phenomenon (Hallast *et al.* 2013). As a result, over some fraction of the genome, individuals from the same species may be more closely related to an outgroup species than they are to each other.

In this manuscript we present a detailed analysis of one such paralogous sequence, *R2d*, in the house mouse *Mus musculus*. *R2d* is a 127 kbp unit which contains the protein-coding gene *Cwc22*. Although the C57BL/6J reference strain and other classical laboratory strains have a single 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 copies respectively in the *R2d2* locus. *R2d2* is the responder locus in a recently-described meiotic drive system on mouse chromosome 2 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 explore patterns of sequence divergence across the locus and perform local *de novo* assembly of phased haplotypes.

Both phylogenetic analyses and estimation of mutation rate in laboratory mouse populations reveal that *R2d2* and its surrounding region on chromsome 2 are exceptionally unstable in copy number. Cycles of duplication, deletion, retrotransposition and non-allelic gene conversion and loss have led to complex phylogenetic patterns discordant with species-level relationships within *Mus*.

Finally, we identify 57 other loci, covering 0.8% of the mouse genome, which share the features of *R2d2*: elevated local sequence divergence; low recombination rate; and enrichment for segmental duplication. Previous studies of sequence variation in the mouse (Keane *et al.* 2011) have attributed this pattern to sorting of alleles segregating in the common ancestor of *M. musculus* and its sister species. We suggest instead that these loci have been subject to independent cycles of duplication and loss along *Mus* lineages. Marked enrichment for odorant, pheromone, and antigen-recognition receptors supports a role for balancing selection on the generation and maintenance of the extreme level of polymorphism observed at these loci.

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 quantitative PCR or depth of coverage in whole-genome sequencing to assay *R2d* copy number in a collection of samples spanning the phylogeny of the subgenus *Mus.* Samples were classified as having haploid copy number 1 (two chromosomes each with a single copy of *R2d*), 2 (at least one chromosome with an *R2d* duplication) or >2 (both chromosomes with an *R2d* duplication).

We find evidence for >1 haploid copy in representatives of all mouse taxa tested from the Palearctic clade (Suzuki *et al.* 2004) (**Figure 1** and **Supplementary Table 1**): 202 of 496 *Mus musculus*, 1 of 1 *M. macedonicus*, 2 of 2 *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 >1 haploid copy in none of the 9 samples tested from subgenus *Pyromys*. We conclude that the *R2d* duplication most likely occurred between the divergence of southeast Asian from Palearctic mice (~3.5 million years ago [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 is ancestral with respect to *M. musculus*, then extant lineages of laboratory mice which have a single haploid copy of *R2d* -- including the reference strain C57BL/6J (of predominantly *M. musculus domesticus* origin (Yang *et al.* 2007)) -- represent subsequent losses of an *R2d* paralog.

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 Mbp; the other copy maps approximately 6 Mbp 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 (Muffato *et al.* 2010) (**Supplementary Figure 1**); 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 haploid *R2d* copy number ~34 (Didion *et al.* 2015). Pairwise alignment of *R2d2* against *R2d1* is shown in **Supplementary Figure 2**. The paralogs differ by at least 8 transposable-element (TE) insertions: 7 LINE

elements specific to *R2d1* and 1 endogenous retroviral element (ERV) specific to *R2d2* (**Supplementary Table 2**). (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 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* a total of 16.9 kbp of intergenic and intronic sequence in 8 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 *Mus caroli* (haploid copy number 1) is used as an outgroup. A representative tree is shown in **Figure 1C**. The *M. musculus* sequences form two clades representing *R2d1*- and *R2d2*-like sequence, respectively. Using 5.0 ± 1.0 million years before present (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 (Drummond *et al.* 2012) yields 1.6 Mya (95% HPD 0.7 − 5.1 Mya) as the estimated age of the duplication event. This estimate is consistent with the conclusion in **Figure 1A**.

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* ($Cwc22^{R2d2}$) plus retrotransposed copies in one locus at chr2: 83.9 Mbp and at two loci on the X chromosome (**Figure 2A**).

The three retrogenes are located in regions with no sequence homology to each other, indicating that each represents an independent retrotransposition event. The copy on chr2 was subsequently expanded by further segmental duplication and now exists (in the reference genome) in 7 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 2B**). Six of the seven protein-coding $Cwc22^{R2d1}$ transcripts in Ensembl v83 use this terminal exon, while one transcript (ENSMUST0000011824) uses an alternative 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.

Copy number at *R2d2* is highly polymorphic in *M. musculus*

We previously demonstrated that haploid copy number of *R2d* ranges from 1 in the reference strain C57BL/6J and classical inbred strains A/J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HILtJ; to 2 in the wild-derived

164 strain CAST/EiJ; to 34 in the wild-derived strain WSB/EiJ. Using linkage mapping in two multiparental
165 genetic reference populations, the Collaborative Cross (Collaborative Cross Consortium 2012) and
166 Diversity Outbred (Svenson *et al.* 2012), we showed that, for the two strains with haploid copy number
167 >1, one copy maps to *R2d1* while all extra copies map to the *R2d2* locus at chr2: 83 Mbp (Didion *et al.*
168 2015). *Cwc22* was recently reported to have copy number as high as 83 in wild *M. m. domesticus* (Pezer *et*
169 *al.* 2015).

170 In order to understand the evolutionary dynamics of copy-number variation at *R2d2*, we investigated the
171 relationship between copy number and phylogeny in the *R2d2* candidate region. In particular, we sought
172 evidence for or against a single common origin for each of the derived copy-number states (zero copies or
173 many copies) at *R2d2*. The extent of copy-number variation in *M. musculus*, as estimated on a continuous
174 scale by qPCR, is shown in **Figure 3A**; samples are colored according to their copy-number classification
175 (1, 2 or >2 haploid copies). (Note that the qPCR readout is proportional to copy number on the log scale.)
176 We confirmed that *R2d2* maps to chr2: 83 Mbp by performing association mapping between SNP
177 genotypes from the MegaMUGA array (Morgan *et al.* 2016) and the qPCR readout (**Figure 3B**).

178 A phylogenetic tree was constructed using SNPs in the region of maximum association (chr2: 82 - 85
179 Mbp) and is shown in **Figure 3C**. The topology of the tree reflects predicted relationships: samples cluster
180 by subspecies and (more loosely) by geography. Samples with a single *R2d* copy are present in all three
181 subspecies and in every major branch within *M. m. domesticus*, supporting the conclusion that the *R2d2*
182 paralog has been lost independently in multiple lineages. Several instances of intersubspecific
183 introgression are evident: *M. m. musculus* into TW:DAJ (predominantly *M. m. castaneus*) and C57BR/cdJ
184 (*M. m. domesticus*), and *M. m. domesticus* into CAST/EiJ (*M. m. castaneus*).

185 With two exceptions (SPRET/EiJ and TW:MEI; see **Discussion**), high-copy haplotypes are restricted to *M.*
186 *m. domesticus*. They are present in classical inbred strains (ALR/LtJ, ALS/LtJ, CHMU/LeJ, and NU/J); in the
187 ICR:HsD and related CD-1 commercial outbred stocks; in two inbred lines derived from ICR:HsD stock,
188 HR8 (Swallow *et al.* 1998) and M16i (Rocha *et al.* 2004); and in wild-derived strains DDO, RBA/DnJ,
189 RBB/DnJ, RBF/DnJ, WSA/EiJ, WSB/EiJ and ZALENDE/EiJ. In the wild, high-copy haplotypes are found in
190 populations from across Europe and in the eastern United States. Although a clade representing mice of
191 western European origin (highlighted in **Figure 3C**) is enriched for high-copy haplotypes, the most recent
192 common ancestor of all samples with high copy number is basal to all *M. m. domesticus* surveyed.

193 The most highly associated SNP marker (JAX00494952 at chr2: 83,673,784) has only a weak correlation
194 ($r^2 = 0.137$) with (log) copy number. High-copy haplotypes are thus poorly tagged in this dataset, and
195 their apparently wide phylogenetic distribution in **Figure 3C** may be artefactual. To investigate further
196 we turned to a partially overlapping dataset from the higher-density Mouse Diversity Array (MDA;
197 (Yang *et al.* 2011)) containing 24 samples with copy number >2 (of which 5 are classical laboratory strains,
198 7 wild-derived strains, and 7 European wild mice; **Supplementary Table 1**). In fact all 24 high-copy
199 samples are identical across a single 21 kbp interval, chr2: 83,896,447 - 83,917,565 (GRCm38/mm10
200 coordinates) (**Figure 3D**).

201 These analyses support a single origin within *Mus* for alleles with >2 copies and multiple origins for
202 alleles with 1 copy.

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

204 To identify the coding sequence of $Cwc22^{R2d2}$ we first aligned the annotated transcript sequences of
205 $Cwc22^{R2d1}$ from Ensembl to our *R2d2* contig. All 21 exons present in *R2d1* are present in *R2d2*. Then we
206 used RNA-seq data from adult brain and testis in inbred strains with one or more copies of *R2d2* to

identify several novel transcript isoforms specific to *R2d2* (**Figure 4A**). First, we find evidence for frequent retention of the 12th and 18th introns in *Cwc22*^{R2d2} transcripts. The latter event can be explained by an A>G mutation at the 5′ splice donor site of exon 17 in *Cwc22*^{R2d2}. No variants were identified in the splice donor or acceptor sites for the 12th intron, but mis-splicing of this intron may be related to the presence of an ERV insertion (**Figure 4A**). Both intron-retention events would create an early stop codon. Second, 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 4A**).

We estimated the expression of the various isoforms of *Cwc22*^{R2d1}, *Cwc22*^{R2d2} and retro-*Cwc22* in adult brain (8 replicates on 3 inbred strains) and testis (a single replicate on 23 inbred strains) using RNA-seq and the kallisto package (Bray *et al.* 2015). Briefly, kallisto uses an expectation-maximization (EM) algorithm to accurately estimate the abundance of a set of transcripts by distributing the "weight" of each read across all isoforms with whose sequence it is compatible. *Cwc22* is clearly expressed from all three paralogs in both brain and testis (**Figure 4B**). However, both the total expression and the pattern of isoform usage differ by tissue and copy number.

Next, 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 (**Supplementary Figure 3**). Ignoring the effects of retained introns in *Cwc22*^{R2d2}, an open reading frame (ORF) is maintained in all three *Cwc22* loci in mouse, including the retrogene. Information content of each column along the alignment (**Supplementary Figure 4**) 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).

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 **Supplementary Figure 3** suggest that *Cwc22* may also be under relaxed purifying selection in rodents. The rate of evolution of *Cwc22* sequences in mouse is faster than in the rest of the tree ($\chi^2 = 4.33$, df = 1, $p = 0.037$ by likelihood ratio test).

Phylogenetic discordance in *R2d1* is due to non-allelic 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 5A**), a signature of non-allelic (*i.e.* inter-locus) gene conversion. To investigate further, we inspected patterns of sequence variation in whole-genome sequencing data from 16 wild-caught mice, 5 wild-derived strains, and 5 classical inbred strains of mice with a single *R2d* copy (at *R2d1*). We identified tracts of *R2d2*-like sequence as clusters of derived variants shared with our *R2d2* contig, and confirmed that they are physically contiguous with neighboring *R2d1* sequence by finding read pairs spanning the junction between *R2d1*- and *R2d2*-like sequence.

This analysis revealed non-allelic gene conversion tracts on at least 9 chromosomes (**Figure 5B**). The conversion tracts range in size from approximately 1.2 kbp to 127 kbp (the full length of the *R2d* unit). We require the presence of *R2d1*- or *R2d2*-diagnostic alleles at two or more consecutive variants to declare a conversion event, and these variants occur at a rate of approximately 1 per 50 bp, so the smallest conversion tracts we could detect are on the order of 100 bp in size. Nonetheless, the conversion tracts we detected are orders of magnitude longer than the 15 to 750 bp reported in recent studies of allelic gene conversion at recombination hotspots in mouse meiosis (Cole *et al.* 2010, 2014).

250 Four conversion tracts partially overlap the *Cwc22* gene to create a sequence that is a mosaic of *R2d1*- and
251 *R2d2*-like exons (**Figure 5B**). Recovery of *Cwc22* mRNA in an inbred strain with a mosaic sequence
252 (PWK/PhJ, see section "*Cwc22* is intact and expressed" below) indicates that transcription remains intact.
253 The presence of both *R2d1*- and *R2d2*-like sequence in extant *M. musculus* lineages with a single copy of
254 *R2d* reinforces our conclusion that the duplication is indeed ancestral to the divergence of *M. musculus*
255 and that the reduction to haploid copy number 1 is a derived state.

256 In addition to exchanges between *R2d1* and *R2d2*, we identified an instance of exchange between *R2d2*
257 and the adjacent retrotransposed copy of *Cwc22* in a single *M. m. domesticus* individual from Iran
258 (IR:AHZ_STND:015; **Supplementary Figure 6**). A 30 kbp fragment corresponding to the 3' half of
259 $Cwc22^{R2d2}$ was transposed into the retro-*Cwc22* locus, apparently mediated by homology between the
260 exons of $Cwc22^{R2d2}$ and retro-*Cwc22*.

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

262 The extent of copy-number polymorphism and the frequency of non-allelic gene conversion involving
263 *R2d2* suggest that it is relatively unstable. Consistent with these observations, we find that the rate of *de*
264 *novo* copy-number changes at *R2d2* is extremely high in laboratory populations (**Figure 6**). In 183 mice
265 sampled from the Diversity Outbred (DO) population we identified and confirmed through segregation
266 analysis 8 new alleles, each with distinct copy number and each occurring in an unrelated haplotype
267 (**Supplementary Table 3**). The DO is an outbred stock derived from eight inbred founder strains and
268 maintained by random mating with 175 breeding pairs; at each generation, one male and one female
269 offspring are sampled from each mating and randomly paired with a non-sibling to produce the next
270 generation (Svenson *et al.* 2012). Without complete pedigrees and genetic material from breeders a direct
271 estimate of the mutation rate in the DO is not straightforward to obtain. However, since the population
272 size is known, we can make an analogy to microsatellite loci (Moran 1975) and estimate the mutation rate
273 via the variance in allele sizes: 3.2% (95% bootstrap CI 1.1% − 6.0%) per generation.

274 Structural instability in this region of chr2 extends outside the *R2d2* locus itself. Less than 200 kbp distal
275 to *R2d2* is another segmental duplication (**Figure 7**) -- containing a retrotransposed copy of *Cwc22* --
276 which is present in 7 tandem copies in the reference genome. That region, plus a further 80 kbp
277 immediately distal to it, is copy-number polymorphic in wild *M. m. domesticus* and wild *M. m. castaneus*
278 (**Figure 7**). Instability of the region over longer timescale is demonstrated by the disruption, just distal to
279 the aforementioned segmental duplication, of a syntenic block conserved across all other mammals
280 (**Supplementary Figure 1**).

281 Despite the high mutation rate for structural variants involving *R2d2* and nearby sequences, sequence
282 diversity at the nucleotide level is modestly reduced relative to diversity in *R2d1* and relative to the
283 genome-wide average in *M. m. domesticus*. In an approximately 200 kbp region containing the *R2d2*
284 insertion site at its proximal end, $\hat{\pi}$ (an estimator of average heterozygosity) in *M. m. domesticus* reduced
285 from approximately 0.3% (comparable to previous reports in this subspecies, (Salcedo *et al.* 2007)) to
286 nearly zero (**Figure 7**). Divergence between *M. musculus* and *M. caroli* is similar to its genome-wide
287 average of ~ 2.5% over the region.

288 Estimation of diversity *within* a duplicated sequence such as *R2d* is complicated by the difficulty of
289 distinguishing allelic from paralogous variation. To circumvent this problem we split our sample of 26
290 wild *M. m. domesticus* into two groups: those having *R2d1* sequences only, and those having both *R2d1*
291 and *R2d2* sequences. Within each group we counted the number of segregating sites among all *R2d2*
292 copies, using nearby fixed differences between *R2d1* and *R2d2* to phase sites to *R2d2* (see **Methods** for

293 details), and used Watterson's estimator to calculate nucleotide diversity per site. Among *R2d1*
294 sequences, $\theta = 0.25\% \pm 0.07\%$ versus $\theta = 0.048\% \pm 0.02\%$ among *R2d2* sequences (**Figure 7**) and
295 $\theta = 0.17\% \pm 0.05\%$ among *R2d2* sequences in *M. m. castaneus*.

296 **Churning of segmental duplications affects 0.8% of the mouse genome**

297 The superposition of segmental duplication, non-allelic gene conversion, and recurrent loss -- which we
298 collectively term "churning" -- of *R2d* paralogs mimics incomplete lineage sorting at *R2d1*. That is, the
299 time to most recent common ancestor of *R2d1* haplotypes within the single-copy *M. musculus* samples is
300 ~2 Mya, much older than the separation of the three *M. musculus* subspecies (**Figure 5**). We hypothesized
301 that churning affects other segmentally duplicated sequences, and sought evidence for it at other loci in
302 the mouse genome. To do so we identified regions where the divergence between sequenced samples and
303 the reference genome assembly substantially exceeds what is predicted by those samples' ancestry.
304 Because churning is likely to involve sequence that, like *R2d2*, is absent from the reference genome
305 assembly, we developed a simple method to estimate divergence from unaligned reads. Briefly, we
306 estimated the proportion of short subsequences (k-mers, with $k = 31$) from windows along the reference
307 assembly for which no evidence exists in the reads generated for a sample. This quantity can be rescaled
308 to approximate the divergence between the reference sequence and the template sequence from which
309 reads were generated (see **Methods**). Applied genome-wide to 31 kbp ($= 1000\times k$) windows, this method
310 captures the distribution of sequence divergence between the reference assembly (chiefly *M. m. domesticus*
311 in origin) in representative samples from the three subspecies of *M. musculus* and the outgroups *M.*
312 *spretus* and *M. caroli* (**Figure 8**).

313 Divergent regions were identified by fitting a hidden Markov model (HMM) to the windowed
314 divergence profiles for 7 wild or wild-derived samples with available whole-genome sequence (**Figure 8**
315 and **Supplementary Table 4**). This analysis revealed a striking pattern: over most of the genome, the
316 divergence estimates hover around the genome-wide expectation, but high-divergence windows are
317 clustered in regions 100 kbp to 5 Mbp in size. Our method does not capture signal from structural
318 variation (besides deletions, see **Methods**) or heterozygosity, and so probably underestimates the true
319 level of sequence divergence. The union of these divergent regions across all 7 *M. musculus* samples
320 analyzed covers 0.82% of the reference genome (mean 0.58% per sample). Divergent regions are enriched
321 for segmental duplications: 39.0% of sequence in divergent regions is comprised of segmental
322 duplications versus a median of 3.4% (central 99% interval $0.2\% - 16.7\%$) in random regions of equal
323 size ($p < 0.001$). Yet divergent regions are more gene-dense than the genomic background: they contain
324 3.6×10^{-2} genes/kbp relative a median 1.6×10^{-2} genes/kbp (central 99% interval 1.1×10^{-2} -- 2.7×10^{-2}
325 genes/kbp) genome-wide ($p = < 0.001$). Divergent regions are strongly enriched for genes related to
326 odorant and pheromone sensing ($p = 1.3\times10^{-8}$) and adaptive immunity ($p = 3.1\times10^{-3}$).

327 As a prototypical example we focus on a divergent region at chr4: 110 - 115 Mbp (**Figure 8**). This region
328 contains the 11 members of the *Skint* family of T-cell-borne antigen receptors. The first member to be
329 described, *Skint1*, functions in negative selection in the thymus of T-cells destined for the epidermis
330 (Boyden *et al.* 2008). Coding sequence from 23 inbred strains (including CAST/EiJ, WSB/EiJ, PWD/PhJ and
331 SPRET/EiJ) was reported in Boyden *et al*. (2008). A phylogenetic tree constructed from those sequences,
332 plus *M. caroli* and rat as outgroups, reveals the expected pattern of "deep coalescence" (**Figure 8**): the *M.*
333 *m. domesticus* sequences are paraphyletic, and some (group A) are more similar to SPRET/EiJ (*M. spretus*)
334 than to their subspecific congeners (group B). Although the level of sequence divergence in the *Skint1*
335 coding sequence was attributed to ancestral polymorphism maintained by balancing selection in the
336 original report, selection would not be expected to maintain diversity in both coding and non-coding
337 sequence equally as we observe in **Figure 8**. The best explanation for the observed pattern of diversity at

338 *Skint1* is therefore that groups A and B represent paralogs descended from an ancestral duplication
339 followed by subsequent deletion of different paralogs along different lineages (**Figure 8**). This conclusion
340 is supported by the structure of the *Skint* region in the mouse reference genome assembly, which reflects
341 the superposition of many duplications and rearrangements (**Figure 8**).

342 ## DISCUSSION

343 In this manuscript we have reconstructed in detail the evolution of a multi-megabase segmental
344 duplication (SD) in mouse, *R2d2*. Our findings demonstrate the challenges involved in accurately
345 interpreting patterns of polymorphism and divergence within duplicated sequence.

346 SDs are among the most dynamic loci in mammalian genomes. They are foci for copy-number variation
347 in populations, but the sequences of individual duplicates beyond those present in the reference genome
348 are often poorly resolved. Obtaining the sequence of this "missing genome," as we have done for *R2d2*, is
349 an important prerequisite to understanding the evolution of duplicated loci. Since each paralog follows a
350 partially independent evolutionary trajectory, individuals in a population may vary both quantitatively
351 (in the number of copies) and qualitatively (in which copies are retained). Cycles of duplication and loss
352 may furthermore lead to the fixation of different paralogs along different lineages. This "churning" leaves
353 a signature of polymorphism far in excess of the genome-wide background, due to coalescence between
354 alleles originating from distinct paralogs. We identify 57 additional regions covering 0.82% of the mouse
355 genome with this property (**Figure 8**). These regions have gene density similar to unique sequence and
356 are strongly enriched for genes involved in odorant sensing, pheromone recognition and immunity that
357 play important roles in social behavior and speciation (Hurst *et al.* 2001). Excess polymorphism at these
358 loci has previously been attributed to some combination of incomplete lineage sorting and diversifying
359 selection (White *et al.* 2009; Keane *et al.* 2011). Our results suggest that inferences regarding the strength of
360 selection on highly polymorphic loci in regions of genomic churn should be treated with caution.

361 Accurate deconvolution of recent duplications remains a difficult task that requires painstaking manual
362 effort. We exploited the specific properties of *R2d2* in the WSB/EiJ mouse strain -- many highly-similar
363 copies of *R2d2* relative to the single divergent *R2d1* copy -- to extract and assemble the sequence of *R2d2*
364 from short reads (**Supplementary Figure 8**). With the sequence of both the *R2d1* and *R2d2* paralogs in
365 hand, we were able to recognize several remarkable features of *R2d2* that are discussed in detail below.

366 **Long-tract gene conversion.**

367 Previous studies of non-allelic gene conversion in mouse and human have focused either on relatively
368 small (<5 kbp) intervals within species, or have applied phylogenetic methods to multiple paralogs from a
369 single reference genome (Dumont and Eichler 2013). This study is the first, to our knowledge, with the
370 power to resolve large (>5 kbp) non-allelic gene conversion events on an autosome in a population
371 sample. We identify conversion tracts up to 127 kbp in length, orders of magnitude longer than tracts
372 arising from allelic conversion events during meiosis. Gene conversion at this scale can rapidly and
373 dramatically alter paralogous sequences, including -- as shown in **Figure 5** -- the sequences of essential
374 protein-coding genes. This process has been implicated as a source of disease alleles in humans (Chen *et*
375 *al.* 2007).

376 Importantly, we were able to identify non-allelic exchanges in *R2d1* as such only because we were aware
377 of the existence of *R2d2* in other lineages. In this case the transfer of paralogous *R2d2* sequence into *R2d1*
378 creates the appearance of deep coalescence among *R2d1* sequences. Ignoring the effect of gene conversion

379 would cause us to overestimate the degree of polymorphism at *R2d1* by an order of magnitude, and
380 would bias any related estimates of population-genetic parameters (for instance, of effective population
381 size).

382 Our data are not sufficient to estimate the rate of non-allelic gene conversion between *R2d2* and other loci.
383 At minimum we have observed two distinct events: one from *R2d2* into *R2d1*, and a second from *R2d2*
384 into retro-*Cwc22*. From a single conversion event replacing all of *R2d1* with *R2d2*-like sequence, the
385 remaining shorter conversion tracts could be generated by recombination with *R2d1* sequences. Because
386 we find converted haplotypes in both *M. m. musculus* and *M. m. domesticus*, the single conversion event
387 would have had to occur prior to the divergence of the three *M. musculus* subspecies and subsequently
388 remain polymorphic in the diverged populations.

389 The other possibility is that non-allelic gene conversion between *R2d* sequences is recurrent. If this is the
390 case, it probably also has a role in maintaining sequence identity between paralogs in *R2d2* -- an example
391 of so-called "concerted evolution" (Dover 1982). Provided the rate of gene conversion is high enough
392 relative to the rate of mutation, gene conversion in multi-copy sequences like *R2d2* tends to slow the
393 accumulation of new mutations (Nagylaki and Petes 1982). New mutations arising in any single copy are
394 prone to loss not only by drift but also by being "pasted-over" by gene conversion from other intact
395 copies. The strength of this effect increases with copy number (Melamed and Kupiec 1992).

396 In this respect, *R2d2* appears similar to the male-specific region of the Y chromosome in mouse (Soh *et al.*
397 2014) and human (Rozen *et al.* 2003). The large palindromic repeats on chrY are homogenized by frequent
398 non-allelic gene conversion (Hallast *et al.* 2013) such that they have retained >99% sequence identity to
399 each other even after millions of years of evolution. Frequent non-allelic gene conversion has also been
400 documented in arrays of U2 snRNA genes in human (Liao 1997), and in rRNA gene clusters (Eickbush
401 and Eickbush 2007) and centromeric sequences (Schindelhauer 2002; Shi *et al.* 2010) in several species.

402 **Pervasive copy-number variation.**

403 Clusters of segmental duplications have long been known to be hotspots of copy-number variation
404 (Bailey and Eichler 2006; Egan *et al.* 2007; She *et al.* 2008). Recent large-scale sequencing efforts have
405 revealed the existence of thousands of multiallelic CNVs segregating in human populations (Handsaker
406 *et al.* 2015), including cases of "runaway duplication" restricted to specific haplotypes. *R2d2* is another
407 example of this phenomenon.

408 We have surveyed *R2d2* copy number in a large and diverse sample of laboratory and wild mice, and
409 have shown that it varies from 0-1 (the ancestral state) to >80 in certain *M. m. domesticus* populations
410 (**Figure 7**). In a cohort of outbred mice expected to be heterozygous for the WSB/EiJ haplotype at *R2d2* (33
411 copies) we estimate that large deletions, >2 Mbp in size, occur at a rate of 3.2% per generation. This
412 estimate of the mutation rate for CNVs at *R2d2* should be regarded as a lower bound. The power of our
413 copy-number assay to discriminate between copy numbers above ~25 is essentially zero, so that the assay
414 is much more sensitive to losses than to gains. Even our lower-bound mutation rate exceeds that of the
415 most common recurrent deletions in human (~1 per 7000 live births) (Turner *et al.* 2007) and is an order of
416 magnitude higher than the most active CNV hotspots described to date in the mouse (Egan *et al.* 2007).
417 While recurrent copy-number changes are often ascribed to non-allelic homologous recombination, the
418 recombination rate in the vicinity of *R2d2* is dramatically reduced in populations segregating for *R2d2*
419 haplotypes with high copy number (**Supplementary Figure 7**).

420 A second key observation is that both the recombination rate and the structural mutation rate at *R2d2*
421 depend on heterozygosity. In contrast to estimates of mutation rate based on the number of alleles in

422 outbred populations (1% − 10% per generation), the copy number of *R2d2* appears to be stable over at
423 least dozens of generations within inbred strains of mice, including in the Collaborative Cross
424 (Collaborative Cross Consortium 2012). Mutation rate further differs by sex: zero new mutations were
425 observed in 1256 progeny of females heterozygous for a high-copy allele at *R2d2* (data not shown).

426 Taken together, these observations hint at a common structural or epigenetic mechanism affecting the
427 resolution of double-strand breaks in large tracts of unpaired (*i.e.* hemizygous) DNA during male
428 meiosis. Both the obligate-hemizygous sex chromosomes and large unpaired segments on autosomes are
429 epigenetically marked for transcriptional silencing during male meiotic prophase (Laan 2004; Baarends *et*
430 *al.* 2005), and are physically sequestered into a structure called the sex body. Repair of double-strand
431 breaks within the sex body is delayed relative to the autosomes (Mahadevaiah *et al.* 2001) and involves a
432 different suite of proteins (Turner *et al.* 2004). We hypothesize that these male-specific pathway(s) are
433 error-prone in the presence of non-allelic homologous sequences.

434 **Origin and distribution of a meiotic driver.**

435 Females heterozygous for a high- and low-copy allele at *R2d2* preferentially transmit the high-copy allele
436 to progeny, a process called meiotic drive (Didion *et al.* 2015). Meiotic drive can rapidly alter allele
437 frequencies in laboratory and natural populations (Lindholm *et al.* 2016), and we recently showed that
438 high-copy alleles of *R2d2* ($R2d2^{HC}$) sweep through laboratory and natural populations despite reducing
439 the fitness of heterozygous females (Didion *et al.* 2016). These "selfish sweeps" account for the marked
440 reduction in within-population diversity in the vicinity of *R2d2* (**Figure 7**).

441 The present study sheds additional light on the age, origins and fate of $R2d2^{HC}$ alleles. We find that $R2d2^{HC}$
442 alleles have a single origin in western Europe. They are present in several different "chromosomal races"
443 -- populations fixed for specific Robertsonian translocations between which gene flow is limited (Hauffe
444 and Searle 1993) -- indicating that they were likely present at intermediate frequency prior to the origin of
445 the chromosomal races within the past 6,000 to 10,000 years (Nachman *et al.* 1994) and were dispersed
446 through Europe as mice colonized the continent from the south and east (Boursot *et al.* 1993). The
447 presence of $R2d2^{HC}$ in non-*M. m. domesticus* samples (SPRET/EiJ, *M. spretus* from Cadiz, Spain; and
448 TW:MEI, *M. m. castaneus* from Taiwan) is best explained by recent introgression following secondary
449 contact with *M. m. domesticus* (Bonhomme *et al.* 2007; Yang *et al.* 2011).

450 **A new member of the *Cwc22* family.**

451 The duplication that gave rise to *R2d2* also created a new copy of *Cwc22*. Based on our assembly of the
452 *R2d2* sequence, the open reading frame of $Cwc22^{R2d2}$ is intact and encodes a nearly full-length predicted
453 protein that retains the two key functional domains characteristic of the *Cwc22* family. Inspection of
454 RNA-seq data from samples with high copy number at *R2d2* reveals several novel transcript isoforms
455 whose expression appears to be copy-number- and tissue-dependent. In testis, the most abundant
456 isoform retains an intron containing an ERV insertion (red arrow in **Figure 4**), consistent with the well-
457 known transcriptional promiscuity in this tissue. The most abundant isoforms in adult brain is unusual in
458 that its stop codon is in an internal exon which is followed by a 7 kbp 3′ UTR in the terminal exon.
459 Transcripts with a stop codon in an internal exon are generally subject to nonsense-mediated decay
460 (NMD) triggered by the presence of exon-junction complexes downstream the stop codon. Curiously,
461 *Cwc22* is itself a member of the exon-junction complex (Steckelberg *et al.* 2012).

462 That an essential gene involved in such a central biochemical pathway should both escape NMD and be
463 overexpressed more than tenfold is surprising. However, it may be the case that standing levels of
464 $Cwc22^{R2d2}$ transcript do not reflect levels of the functional *Cwc22* protein. Either of the two intron-retention

465 events in **Figure 4** would introduce early stop codons and consequently be subject to NMD or produce a
466 truncated and likely nonfunctional protein. Further studies will be required to determine the distribution
467 of expression of *Cwc22* across isoforms, tissues and developmental stages.

468 **Conclusions and future directions**

469 Our detailed analysis of the evolutionary trajectory of *R2d2* provides insight into the fate of duplicated
470 sequences over short (within-species) timescales. The exceptionally high mutation rate and low
471 recombination rate at *R2d2* motivate hypotheses regarding the biochemical mechanisms which contribute
472 to observed patterns of polymorphism at this and similar loci. Finally, the birth of a new member of the
473 deeply conserved *Cwc22* gene family in *R2d2* provides an opportunity to test predictions regarding the
474 evolution of duplicate gene pairs.

475 METHODS

476 **Mice**

477 Wild *M. musculus* mice used in this study were trapped at a large number of sites across Europe, the
478 United States, the Middle East, northern India and Taiwan (**Figure 7A**). Trapping was carried out in
479 accordance with local regulations and with the approval of all relevant regulatory bodies for each locality
480 and institution. Trapping locations are listed in **Supplementary Table 1**. Most samples have been
481 previously published (Didion *et al.* 2016).

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

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

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

490 All live laboratory mice were handled in accordance with the IACUC protocols of the University of North
491 Carolina at Chapel Hill.

492 **DNA preparation**

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

499 *Low molecular weight DNA.* Low molecular weight DNA was obtained for samples to be genotyped on the
500 MegaMUGA array (see "Microarray genotyping" below). Genomic DNA was isolated from tail, liver,
501 muscle or spleen using Qiagen Gentra Puregene or DNeasy Blood & Tissue kits according to the
502 manufacturer's instructions.

503 **Whole-genome sequencing and variant discovery**

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

508 The strains LEWES/EiJ and ZALENDE/EiJ were sequenced at the University of North Carolina High-
509 Throughput Sequencing Facility. Libraries were prepared from high molecular weight DNA using the
510 Illumina TruSeq kit and insert size approximately 250 bp, and 2x100bp paired-end reads were generated
511 on an Illumina HiSeq 2000 instrument. LEWES/EiJ was sequenced to approximately 12X coverage and
512 ZALENDE/EiJ to approximately 20X. Alignment was performed as in Keane *et al.* (2011).

513 *Wild mice.* Whole-genome sequencing data from 26 wild *M. m. domesticus* individuals described in Pezer
514 *et al.* (2015) was downloaded from ENA under accession #PRJEB9450. Coverage ranged from
515 approximately 12X to 20X per sample. An additional two wild *M. m. domesticus* individuals, IT175 and
516 ES446, were sequenced at the University of North Carolina to approximate coverage 8X each. Raw reads
517 from an additional 10 wild *M. m. castaneus* described in Halligan *et al.* (2013), sequenced to approximately
518 20X each, were downloaded from ENA under accession #PRJEB2176. Reads for a single *Mus caroli*
519 individual sequenced to approximately 40X were obtained from ENA under accession #PRJEB2188.
520 Reads for each sample were realigned to the mm10 reference using bwa-mem v0.7.12 with default
521 parameters (Li 2013). Optical duplicates were removed with samblaster (Faust and Hall 2014).

522 *Variant discovery.* Polymorphic sites on chromosome 2 in the vicinity of *R2d2* (**Figure 7**) were called using
523 freebayes v0.9.21-19-gc003c1e (Garrison and Marth 2012) with parameters "--standard-filters" using the
524 Sanger Mouse Genomes Project VCF files as a list of known sites (parameter "--@"). Raw calls were
525 filtered to have quality score > 30, root mean square mapping quality > 20 (for both reference and
526 alternate allele calls) and at most 2 alternate alleles.

527 **Copy-number estimation**

528 *R2d* copy number was estimated using qPCR as described in Didion *et al.* (2016). Briefly, we used
529 commercial TaqMan assays against intron-exon boundaries in *Cwc22* (Life Technologies assay numbers
530 Mm00644079_cn and Mm00053048_cn) to determine copy number relative to reference genes *Tert* (cat. no.
531 4458368, for target Mm00644079_cn) or *Tfrc* (cat. no. 4458366, for target Mm00053048_cn). Cycle
532 thresholds for *Cwc22* relative to the reference gene were normalized across assay batches using linear
533 mixed models with batch and target-reference pair treated as random effects. Control samples with
534 known haploid *R2d* copy numbers of 1 (C57BL/6J), 2 (CAST/EiJ), 17 (WSB/EiJxC57BL/6J)F_1 and 34
535 (WSB/EiJ) were included in each batch.

536 Samples were classified as having 1, 2 or >2 haploid copies of *R2d* using linear discriminant analysis. The
537 classifier was trained on the normalized cycle thresholds of the control samples from each plate, whose
538 precise integer copy number is known, and applied to the remaining samples.

539 **Microarray genotyping**

540 Genome-wide genotyping was performed using MegaMUGA, the second version of the Mouse Universal
541 Genotyping Array platform (Neogen/GeneSeek, Lincoln, NE) (Morgan *et al.* 2016). Genotypes were called
542 using the GenCall algorithm implemented in the Illumina BeadStudio software (Illumina Inc, Carlsbad,
543 CA). 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

544 the normalized hybridization intensities for the two alleles. The expected distribution of Si was computed
545 from a large set of reference samples. We excluded arrays for which the distribution of S_i was
546 substantially shifted from this reference; in practice, failed arrays can be trivially identified in this manner
547 (Morgan *et al.* 2016). Access to MegaMUGA genotypes was provided by partnership between the
548 McMillan and Pardo-Manuel de Villena labs and the UNC Systems Genetics Core Facility.

549 Additional genotypes for inbred strains and wild mice from the Mouse Diversity Array were obtained
550 from Yang *et al.* (2011).

551 ***De novo* assembly of *R2d2***

552 Raw whole-genome sequencing reads for WSB/EiJ from the Sanger Mouse Genomes Project were
553 converted to a multi-string Burrows-Wheeler transform and associated FM-index (msBWT) (Holt and
554 McMillan 2014) using the msbwt v0.1.4 Python package (https://pypi.python.org/pypi/msbwt). The
555 msBWT and FM-index implicitly represent a suffix array of sequencing to provide efficient queries over
556 arbitrarily large string sets. Given a seed k-mer present in that string set, this property can be exploited to
557 rapidly construct a de Bruijn graph which can in turn be used for local *de novo* assembly of a target
558 sequence (**Supplementary Figure 8A**). The edges in that graph can be assigned a weight (corresponding
559 to the number of reads containing the $k + 1$-mer implied by the edge) which can be used to evaluate
560 candidate paths when the graph branches (**Supplementary Figure 8B**).

561 *R2d2* was seeded with the 30 bp sequence (TCTAGAGCATGAGCCTCATTTATCATGCCT) at the
562 proximal boundary of *R2d1* in the GRCm38/mm10 reference genome. A single linear contig was
563 assembled by "walking" through the local de Bruijn graph. Because WSB/EiJ has ~33 copies of *R2d2* and a
564 single copy of *R2d1*, any branch point in the graph which represents a paralogous variant should having
565 outgoing edges with weights differing by a factor of approximately 33. Furthermore, when two (or more)
566 branch points occur within less than the length of a read, it should be possible to "phase" the underlying
567 variants by following single reads through both branch points (**Supplementary Figure 8B**). We used
568 these heuristics to assemble the sequence of *R2d2* (corresponding to the higher-weight path through the
569 graph) specifically.

570 After assembling a chunk of approximately 500 bp the contig was checked for colinearity with the
571 reference sequence (*R2d1*) using BLAT and CLUSTAL-W2 (using the EMBL-EBI web server:
572 http://www.ebi.ac.uk/Tools/msa/clustalw2/).

573 Repetitive elements such as retroviruses are refractory to assembly with our method. Upon traversing
574 into a repetitive element, the total edge weight (total number of reads) and number of branch points
575 (representing possible linear assembled sequences) in the graph become large. It was sometimes possible
576 to assemble a fragment of a repetitive element at its junction with unique sequence but not to assemble
577 unambiguously across the repeat. Regions of unassembleable sequence were marked with blocks of Ns,
578 and assembly re-seeded using a nearby *k*-mer from the reference sequence. The final contig is provided
579 in FASTA format in **Supplementary File 1**.

580 The final contig was checked against its source msBWT by confirming that each 30-mer in the contig
581 which did not contain an N was supported by at least 60 reads. A total of 16 additional haplotypes in 8
582 regions of *R2d* totaling 16.9 kbp (**Supplementary Table 5**) were assembled in a similar fashion, using the
583 WSB *R2d2* contig and the *R2d1* reference sequence as guides. Multiple sequence alignments from these
584 regions are provided in **Supplementary File 1**.

585 **Sequence analysis of *R2d2* contig**

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

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

597 *Gene conversion tracts.* Using multiple sequence alignments (see "Phylogenetic analyses" below) of the 8 *de*
598 *novo* assembled regions in *R2d*, we classified each sequence as *R2d1*-like or *R2d2*-like based on the pattern
599 of shared derived alleles (using *M. caroli* as the outgroup). The result of this analysis is shown in
600 **Supplementary Figure 5**.

601 To unambiguously define gene conversion events in a larger sample, and without confounding from
602 paralogous sequence, we examined 12 wild *M. m. domesticus* samples and 8 laboratory strains with
603 evidence of 2 diploid copies of *R2d*. We first confirmed that these copies of *R2d* were located at *R2d1* by
604 finding read pairs spanning the junction between *R2d1* and neighboring sequence. Gene conversion tracts
605 were delineated by manual inspection of alignments in IGV. Because *R2d1* and *R2d2* diverged
606 approximately 2 Mya, alignment of *R2d2*-like sequence to the *R2d1* reference sequence creates
607 mismatches at a density of approximately 1 variant per 50 bp, approximately fivefold greater than the
608 density of variants in wild *M. m. domesticus* (~ 0.4% or 1 per 250 bp; (Salcedo *et al.* 2007)). Boundaries of
609 conversion tracts were defined at approximately the midpoint between the first *R2d1*- (or *R2d2*-) specific
610 variant and the last *R2d2*- (or *R2d1*-) specific variant.

611 *Sequence diversity in* R2d1 *and* R2d2. Assembling individual copies of *R2d2* is infeasible in high-copy
612 samples. Instead we treated each *R2d* unit as an independent sequence and used the number of
613 segregating sites to estimate sequence diversity. Segregating sites were defined as positions in a collection
614 of alignments (BAM files) with evidence of an alternate allele. To identify segregating sites we used
615 freebayes v0.9.21-19-gc003c1e (Garrison and Marth 2012) with parameters "-ui -Kp 20 --use-best-n-alleles
616 2 -m 8". These parameters treat each sample as having ploidy up to 20, impose an uninformative prior on
617 genotype frequencies, and limit the algorithm to the discovery of atomic variants (SNVs or short indels,
618 not multinucleotide polymorphisms or other complex events) with at most 2 alleles at each segregating
619 site. Sites in low-complexity sequence (defined as Shannon entropy < 1.6 in the 30 bp window centered on
620 the site) or within 10 bp of another variant site were further masked, to minimize spurious calls due to
621 ambiguous alignment of indels. To avoid confounding with the retrocopies of *Cwc22* outside *R2d*, coding
622 exons of *Cwc22* were masked. Finally, sites corresponding to an unaligned or gap position in the pairwise
623 alignment between *R2d1* and *R2d2* were masked.

624 To compute diversity in *R2d1* we counted segregating sites in 12 wild *M. m. domesticus* samples with 2
625 diploid copies of *R2d* (total of 24 sequences), confirmed to be in *R2d1* by the presence of read pairs
626 spanning the junction between *R2d1* and neighboring sequence. To compute diversity in *R2d2*, we
627 counted segregating sites in 14 wild *M. m. domesticus* samples with >2 diploid copies of *R2d* (range 3 -- 83
628 per sample; total of 406 sequences) but excluded sites corresponding to variants among *R2d1* sequences.

629 Remaining sites were phased to *R2d2* by checking for the presence of a 31-mer containing the site and the
630 nearest *R2d1*-vs-*R2d2* difference in the raw reads for each sample using the corresponding msBWT.
631 Sequence diversity was then computed using Watterson's estimator (Watterson 1975), dividing by the
632 number of alignable bases (128973) to yield a per-site estimate. Standard errors were estimated by 100
633 rounds of resampling over the columns in the *R2d1*-vs-*R2d2* alignment.

634 **Analyses of *Cwc22* expression**

635 *RNA-seq read alignment.* Expression of *Cwc22* was examined in adult whole brain using data from
636 Crowley *et al.* (2015), SRA accession #SRP056236. Paired-end reads (2x100bp) were obtained from 8
637 replicates each of 3 inbred strains: CAST/EiJ, PWK/PhJ and WSB/EiJ. Raw reads were aligned to the
638 mm10 reference using STAR v2.4.2a (Dobin *et al.* 2012) with default parameters for paired-end reads.
639 Alignments were merged into a single file per strain for further analysis. Expression in adult testis was
640 examined in 23 wild-derived inbred strains from Phifer-Rixey *et al.* (2014) SRA accession #PRJNA252743.
641 Single-end reads (76bp) were aligned to the mm10 genome with STAR using default parameters for
642 single-end, non-strand-specific reads.

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

655 *Abundance estimation.* Relative abundance of *Cwc22* paralogs was estimated using kallisto v0.42.3 (Bray *et*
656 *al.* 2015) with parameters "--bias" (to estimate and correct library-specific sequence-composition biases).
657 The transcript index used for pseudoalignment and quantification included only the *Cwc22* targets.

658 **Phylogenetic analyses**

659 *Trees.* Multiple sequence alignments for 8 the regions in **Supplementary Figure 5** were generated using
660 MUSCLE (Edgar 2004) with default parameters. The resulting alignments were manually trimmed and
661 consecutive gaps removed. Phylogenetic trees were inferred with RAxML v8.1.9 (Stamatakis 2014) using
662 the GTR+gamma model with 4 rate categories and *M. caroli* as an outgroup. Uncertainty of tree topologies
663 was evaluated using 100 bootstraps replicates.

664 *Divergence time.* The time of the split between *R2d1* and *R2d2* was estimated using the Bayesian method
665 implemented in BEAST v1.8.1r6542 (Drummond *et al.* 2012). We assumed a divergence time for *M. caroli*
666 of 5 Mya and a strict molecular clock, and analyzed the alignment from region A in **Supplementary**
667 **Figure 5** under the GTR+gamma model with 4 rate categories and allowance for a proportion of invariant
668 sites. The chain was run for 10 million iterations with trees sampled every 1000 iterations.

669 *Local phylogeny around* R2d2. Genotypes from the MegaMUGA array at 38 SNPs in the region surrounding
670 *R2d2* (chr2: 82 -- 85 Mb) were obtained for 493 individuals representing both laboratory and wild mice

671 (**Supplementary Table 1**). A distance matrix was created by computing the proportion of alleles shared
672 identical by state between each pair of samples. A neighbor-joining tree was inferred from the distance
673 matrix and rooted at the most recent common ancestor of the *M. musculus*- and non-*M. musculus* samples.
674 **Figure 3** shows a simplified tree with 135 representative samples, including all those with high-copy
675 alleles at *R2d2*.

676 Cwc22 *coding sequences*. To create the tree of Cwc22 coding sequences, we first obtained the sequences of
677 all its paralogs in mouse. The coding sequence of *Cwc22*^{R2d1} (RefSeq transcript NM_030560.5) was
678 obtained from the UCSC Genome Browser and aligned to our *R2d2* contig with BLAT to extract the exons
679 of *Cwc22*^{R2d2}. The coding sequence of retro-*Cwc22* (genomic sequence corresponding to GenBank cDNA
680 AK145290) was obtained from the UCSC Genome Browser. Coding and protein sequences of *Cwc22*
681 homologs from non-*M. musculus* species were obtained from Ensembl (Cunningham *et al.* 2014). The
682 sequences were aligned with MUSCLE and manually trimmed, and a phylogenetic tree estimated as
683 described above.

684 We observed that the branches in the rodent clade of the *Cwc22* tree appeared to be longer than branches
685 for other taxa. We used PAML (Yang *et al.* 2007) to test the hypothesis that *Cwc22* is under relaxed
686 purifying selection in rodents using the branch-site model (null model "model = 2, NSsites = 2, fix_omega
687 = 1"; alternative model "model = 2, NSsites = 2, omega = 1, fix_omega = 1") as described in the PAML
688 manual. This is a test of difference in evolutionary rate on a "foreground" branch (ω_1) -- in our case, the
689 rodent clade -- relative to the tree-wide "background" rate (ω_0). The distribution of the test statistic is an
690 even mixture of a χ^2 distribution with 1 df and a point mass at zero; to obtain the p-value, we calculated
691 the quantile of the χ^2 distribution with 1 df and divided by 2.

692 **Genome-wide sequence divergence**

693 The msBWT of a collection of whole-genome sequencing reads can be used to estimate the divergence
694 between the corresponding template sequence (*i.e.* genome) and a reference sequence as follows. Non-
695 overlapping k-mers from the reference sequence are queried against the msBWT. (The value of k is
696 chosen such that nearly all k-mers drawn from genomic sequence exclusive of repetitive elements.) Let x
697 be the count of reads containing an exact match to the k-mer or its reverse complement. If the template
698 and reference sequence are identical, standard theory for shotgun sequencing (Lander and Waterman
699 1988) holds that $P(x > 0|\lambda) = 1 - e^{-\lambda}$, where λ is the average sequencing coverage. We assume
700 $P(x > 0|\lambda) \approx 1$, which is satisfied in practice for high-coverage sequencing.

701 However, if a haploid template sequence contains at least one variant (versus the reference) within the
702 queried k-mer, it will be the case that $x = 0$. We use this fact and assume that mutations arise along a
703 sequence via a Poisson process to estimate the rate parameter α from the proportion of k-mers that have
704 read count zero. Let m be the number of mutations arising between a target and reference in a window of
705 length L, and y the number of k-mers in that window with nonzero read count. Then $P(m = 0|\alpha) = e^{-\alpha L}$
706 and a simple estimator for α is $\hat{\alpha} = -\log\left(\frac{1-y}{L}\right)$.

707 Interpretation of α is straightforward in the haploid case: it is the per-base rate of sequence divergence
708 between the template sequence and the reference sequence. In the diploid case it represents a lower
709 bound on the sequence divergence of the two homologous chromosomes.

710 We applied this estimator with $k = 31$ and $L = 1000 \times k = 31$ kbp to msBWTs for 7 inbred strains (3 *M. m.*
711 *domesticus*, 1 *M. m. musculus*, 1 *M. m. castaneus*, 1 *M. spretus*, 1 *M. caroli*) and 2 wild *M. m. domesticus*
712 individuals (IT175, ES446) using the GRCm38/mm10 mouse reference sequence as the source of k-mer

queries. As shown in **Figure 8**, the mode of the distribution of divergence values matches what is expected based on the ancestry of the samples with respect to the reference. To identify divergent regions, we fit a discrete-time hidden Markov model (HMM) to the windows divergence values. The HMM had two hidden states: "normal" sequence, with emission distribution $N(0.005,0.005)$ and initial probability 0.99; and "divergent" sequence, with emission distribution $N(0.02,0.005)$ and initial probability 0.01. The transmission probability between states was 1×10^{-5}. Posterior decodings were obtained via the Viterbi algorithm, as implemented in the R package HiddenMarkov (https://cran.r-project.org/package=HiddenMarkov).

Significance tests for overlap with genomic features were performed using the resampling algorithm implemented in the Genomic Association Tester (GAT) package for Python (https://pypi.python.org/pypi/gat). Segmental duplications were obtained from the genomicSuperDups table of the UCSC Genome Browser and genes from Ensembl v83 annotation.

Analyses of recombination rate at *R2d2*

To test the effect of *R2d2* copy number on local recombination rate we estimated the difference between observed and expected (based on the most recent mouse genetic map (Liu *et al.* 2014)) recombination fraction in 11 experimental crosses in which one of the parental lines was segregating for a high-copy allele at *R2d2*. Genotype data was 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.

We also examined recombination events accumulated during the first 18 generations of breeding of the Diversity Outbred (DO) population, in which the high-copy *R2d2* allele from WSB/EiJ is segregating. Founder haplotype reconstructions were obtained for each of the 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 region chr2: 75-90 Mb.

Results of these analyses are presented in **Supplementary Figure 7**.

DATA AVAILABILITY

All *de novo* assemblies used in this study are included in **Supplementary File 1**. 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.

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 HR selection lines and related crosses. This work was supported by National Institutes of Health grants P50GM076468 (FPMdV), U19AI100625 (FPMdV, APM), F30MH103925 (APM), T32GM067553 (JPD, APM), and by Vaadia-BARD Postdoctoral Fellowship Award

752 FI-12 478-13 to LY. Additional support was provided by Cancer Research UK, the European Research
753 Council, EMBO Young Investigator Programme (DTO), European Molecular Biology Laboratory (DTO,
754 PF) and the Wellcome Trust (WT095908) (PF) and (WT098051) (PF, DTO); and finally by the European
755 Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement HEALTH-F4-2010-
756 241504 (EURATRANS).

757 **REFERENCES**

758 Baarends W. M., Wassenaar E., Laan R. van der, Hoogerbrugge J., Sleddens-Linkels E., Hoeijmakers J. H.
759 J., Boer P. de, Grootegoed J. A., 2005 Silencing of unpaired chromatin and histone H2A ubiquitination in
760 mammalian meiosis. Mol Cell Biol **25**: 1041–1053.

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

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

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

768 Boyden L. M., Lewis J. M., Barbee S. D., Bas A., Girardi M., Hayday A. C., Tigelaar R. E., Lifton R. P., 2008
769 Skint1, the prototype of a newly identified immunoglobulin superfamily gene cluster, positively selects
770 epidermal T cells. Nat Genet **40**: 656–662.

771 Bray N., Pimentel H., Melsted P., Patcher L., 2015 Near-optimal RNA-seq quantification. arXiv.

772 Chain F. J. J., Feulner P. G. D., Panchal M., Eizaguirre C., Samonte I. E., Kalbe M., Lenz T. L., Stoll M.,
773 Bornberg-Bauer E., Milinski M., Reusch T. B. H., 2014 Extensive copy-number variation of young genes
774 across stickleback populations. PLoS Genet **10**: e1004830.

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

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

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

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

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

785 Crowley J. J., Zhabotynsky V., Sun W., Huang S., Pakatci I. K., Kim Y., Wang J. R., Morgan A. P., Calaway
786 J. D., Aylor D. L., Yun Z., Bell T. A., Buus R. J., Calaway M. E., Didion J. P., Gooch T. J., Hansen S. D.,
787 Robinson N. N., Shaw G. D., Spence J. S., Quackenbush C. R., Barrick C. J., Nonneman R. J., Kim K.,

788 Xenakis J., Xie Y., Valdar W., Lenarcic A. B., Wang W., Welsh C. E., Fu C.-P., Zhang Z., Holt J., Guo Z.,
789 Threadgill D. W., Tarantino L. M., Miller D. R., Zou F., McMillan L., Sullivan P. F., Villena F. P.-M. de,
790 2015 Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive
791 allelic imbalance. Nat Genet **47**: 353–360.

792 Cunningham F., Amode M. R., Barrell D., Beal K., Billis K., Brent S., Carvalho-Silva D., Clapham P.,
793 Coates G., Fitzgerald S., Gil L., Giron C. G., Gordon L., Hourlier T., Hunt S. E., Janacek S. H., Johnson N.,
794 Juettemann T., Kahari A. K., Keenan S., Martin F. J., Maurel T., McLaren W., Murphy D. N., Nag R.,
795 Overduin B., Parker A., Patricio M., Perry E., Pignatelli M., Riat H. S., Sheppard D., Taylor K., Thormann
796 A., Vullo A., Wilder S. P., Zadissa A., Aken B. L., Birney E., Harrow J., Kinsella R., Muffato M., Ruffier M.,
797 Searle S. M. J., Spudich G., Trevanion S. J., Yates A., Zerbino D. R., Flicek P., 2014 Ensembl 2015. Nucleic
798 Acids Res **43**: D662–D669.

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

801 Didion J. P., Morgan A. P., Clayshulte A. M.-F., Mcmullan R. C., Yadgary L., Petkov P. M., Bell T. A., Gatti
802 D. M., Crowley J. J., Hua K., Aylor D. L., Bai L., Calaway M., Chesler E. J., French J. E., Geiger T. R.,
803 Gooch T. J., Garland T., Harrill A. H., Hunter K., McMillan L., Holt M., Miller D. R., O'Brien D. A., Paigen
804 K., Pan W., Rowe L. B., Shaw G. D., Simecek P., Sullivan P. F., Svenson K. L., Weinstock G. M., Threadgill
805 D. W., Pomp D., Churchill G. A., Villena F. P.-M. de, 2015 A multi-megabase copy number gain causes
806 maternal transmission ratio distortion on mouse chromosome 2. PLoS Genet **11**: e1004850.

807 Didion J. P., Morgan A. P., Yadgary L., Bell T. A., McMullan R. C., Solorzano L. O. de, Britton-Davidian J.,
808 Bult C. J., Campbell K. J., Castiglia R., Ching Y.-H., Chunco A. J., Crowley J. J., Chesler E. J., Förster D. W.,
809 French J. E., Gabriel S. I., Gatti D. M., Garland T., Giagia-Athanasopoulou E. B., Giménez M. D., Grize S.
810 A., Gündüz I., Holmes A., Hauffe H. C., Herman J. S., Holt J. M., Hua K., Jolley W. J., Lindholm A. K.,
811 López-Fuster M. J., Mitsainas G., Luz Mathias M. da, McMillan L., Ramalhinho M. G., Rehermann B.,
812 Rosshart S. P., Searle J. B., Shiao M.-S., Solano E., Svenson K. L., Thomas-Laemont P., Threadgill D. W.,
813 Ventura J., Weinstock G. M., Pomp D., Churchill G. A., Villena F. P.-M. de, 2016 R2d2 drives selfish
814 sweeps in the house mouse. Mol Biol Evol: msw036.

815 Dobin A., Davis C. A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., Gingeras T. R.,
816 2012 STAR: Ultrafast universal RNA-seq aligner. Bioinformatics **29**: 15–21.

817 Dopman E. B., Hartl D. L., 2007 A portrait of copy-number polymorphism in drosophila melanogaster.
818 Proceedings of the National Academy of Sciences **104**: 19920–19925.

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

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

822 Dumont B. L., Eichler E. E., 2013 Signals of historical interlocus gene conversion in human segmental
823 duplications. PLoS ONE **8**: e75949.

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

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

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

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

832 Garrison E., Marth G., 2012 Haplotype-based variant detection from short-read sequencing. arXiv.

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

836 Grabherr M. G., Haas B. J., Yassour M., Levin J. Z., Thompson D. A., Amit I., Adiconis X., Fan L.,
837 Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., Palma F. di, Birren B.
838 W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A., 2011 Full-length transcriptome assembly from
839 RNA-seq data without a reference genome. Nat Biotechnol **29**: 644–652.

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

843 Halligan D. L., Kousathanas A., Ness R. W., Harr B., Eöry L., Keane T. M., Adams D. J., Keightley P. D.,
844 2013 Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid
845 rodents. PLoS Genet **9**: e1003995.

846 Handsaker R. E., Doren V. V., Berman J. R., Genovese G., Kashin S., Boettger L. M., McCarroll S. A., 2015
847 Large multiallelic copy number variations in humans. Nat Genet **47**: 296–303.

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

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

851 Hurst J. L., Payne C. E., Nevison C. M., Marie A. D., Humphries R. E., Robertson D. H. L., Cavaggioni A.,
852 Beynon R. J., 2001 Individual recognition in mice mediated by major urinary proteins. Nature **414**: 631–
853 634.

854 Keane T. M., Goodstadt L., Danecek P., White M. A., Wong K., Yalcin B., Heger A., Agam A., Slater G.,
855 Goodson M., Furlotte N. A., Eskin E., Nellåker C., Whitley H., Cleak J., Janowitz D., Hernandez-Pliego P.,
856 Edwards A., Belgard T. G., Oliver P. L., McIntyre R. E., Bhomra A., Nicod J., Gan X., Yuan W., Weyden L.
857 van der, Steward C. A., Bala S., Stalker J., Mott R., Durbin R., Jackson I. J., Czechanski A., Guerra-
858 Assunção J. A., Donahue L. R., Reinholdt L. G., Payseur B. A., Ponting C. P., Birney E., Flint J., Adams D.
859 J., 2011 Mouse genomic variation and its effect on phenotypes and gene regulation. Nature **477**: 289–294.

860 Laan R. van der, 2004 Ubiquitin ligase Rad18Sc localizes to the XY body and to other chromosomal
861 regions that are unpaired and transcriptionally silenced during male meiotic prophase. J Cell Sci **117**:
862 5023–5033.

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

865 Li H., 2006 TreeFam: A curated database of phylogenetic trees of animal gene families. Nucleic Acids Res
866 **34**: D572–D580.

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

868 Liao D., 1997 Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the
869 RUN2 locus) involves rapid intrachromosomal homogenization and rare interchromosomal gene
870 conversion. EMBO J **16**: 588–598.

871 Lindholm A. K., Dyer K. A., Firman R. C., Fishman L., Forstmeier W., Holman L., Johannesson H., Knief
872 U., Kokko H., Larracuente A. M., Manser A., Montchamp-Moreau C., Petrosyan V. G., Pomiankowski A.,
873 Presgraves D. C., Safronova L. D., Sutter A., Unckless R. L., Verspoor R. L., Wedell N., Wilkinson G. S.,
874 Price T. A., 2016 The ecology and evolutionary dynamics of meiotic drive. Trends Ecol Evol.

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

878 Lynch M., 2000 The evolutionary fate and consequences of duplicate genes. Science **290**: 1151–1155.

879 Mahadevaiah S. K., Turner J. M., Baudat F., Rogakou E. P., Boer P. de, Blanco-Rodríguez J., Jasin M.,
880 Keeney S., Bonner W. M., Burgoyne P. S., 2001 Nat Genet **27**: 271–276.

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

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

884 Morgan A. P., Fu C.-P., Kao C.-Y., Welsh C. E., Didion J. P., Yadgary L., Hyacinth L., Ferris M. T., Bell T.
885 A., Miller D. R., Giusti-Rodriguez P., Nonneman R. J., Cook K. D., Whitmire J. K., Gralinski L. E., Keller
886 M., Attie A. D., Churchill G. A., Petkov P., Sullivan P. F., Brennan J. R., McMillan L., Villena F. P.-M. de,
887 2016 The Mouse Universal Genotyping Array: From substrains to subspecies. G3 **6**.

888 Muffato M., Louis A., Poisnel C. E., Crollius H. R., 2010 Genomicus: A database and a browser to study
889 gene synteny in modern and ancestral genomes. Bioinformatics **26**: 1119–1121.

890 Nachman M. W., Boyer S. N., Searle J. B., Aquadro C. F., 1994 Mitochondrial DNA variation and the
891 evolution of robertsonian chromosomal races of house mice, Mus domesticus. Genetics **136**: 1105–1120.

892 Nagylaki T., Petes T. D., 1982 Intrachromosomal gene conversion and the maintenance of sequence
893 homogeneity among repeated genes. Genetics **100**: 315–337.

894 Pei B., Sisu C., Frankish A., Howald C., Habegger L., Mu X., Harte R., Balasubramanian S., Tanzer A.,
895 Diekhans M., Reymond A., Hubbard T. J., Harrow J., Gerstein M. B., 2012 The GENCODE pseudogene
896 resource. Genome Biol **13**: R51.

897 Pezer, Harr B., Teschke M., Babiker H., Tautz D., 2015 Divergence patterns of genic copy number
898 variation in natural populations of the house mouse (Mus musculus domesticus) reveal three conserved
899 genes with major population-specific expansions. Genome Res **25**: 1114–1124.

900 Phifer-Rixey M., Bomhoff M., Nachman M. W., 2014 Genome-wide patterns of differentiation among
901 house mouse subspecies. Genetics **198**: 283–297.

902 Rocha J. L., Eisen E. J., Vleck L. D. V., Pomp D., 2004 A large-sample QTL study in mice: I. Growth.
903 Mamm Genome **15**: 83–99.

904 Rozen S., Skaletsky H., Marszalek J. D., Minx P. J., Cordum H. S., Waterston R. H., Wilson R. K., Page D.
905 C., 2003 Abundant gene conversion between arms of palindromes in human and ape Y chromosomes.
906 Nature **423**: 873–876.

907 Salcedo T., Geraldes A., Nachman M. W., 2007 Nucleotide variation in wild and inbred mice. Genetics
908 **177**: 2277–2291.

909 Sambrook J., Russell D. W. (Eds.), 2006 *Molecular cloning: A laboratory manual*. Cold Spring Harbor
910 Laboratory Press.

911 Schindelhauer D., 2002 Evidence for a fast, intrachromosomal conversion mechanism from mapping of
912 nucleotide variants within a homogeneous alpha-satellite DNA array. Genome Res **12**: 1815–1826.

913 She X., Cheng Z., Zöllner S., Church D. M., Eichler E. E., 2008 Mouse segmental duplication and copy
914 number variation. Nat Genet **40**: 909–914.

915 Shi J., Wolf S. E., Burke J. M., Presting G. G., Ross-Ibarra J., Dawe R. K., 2010 Widespread gene conversion
916 in centromere cores. PLoS Biol **8**: e1000327.

917 Soh Y., Alföldi J., Pyntikova T., Brown L., Graves T., Minx P., Fulton R., Kremitzki C., Koutseva N.,
918 Mueller J., Rozen S., Hughes J., Owens E., Womack J., Murphy W., Cao Q., de~Jong P., Warren W.,
919 Wilson R., Skaletsky H., Page D., 2014 Sequencing the mouse Y chromosome reveals convergent gene
920 acquisition and amplification on both sex chromosomes. Cell **159**: 800–813.

921 Stamatakis A., 2014 RAxML version 8: A tool for phylogenetic analysis and post-analysis of large
922 phylogenies. Bioinformatics **30**: 1312–1313.

923 Steckelberg A.-L., Boehm V., Gromadzka A., Gehring N., 2012 Cwc22 connects pre-mRNA splicing and
924 exon junction complex assembly. Cell Rep **2**: 454–461.

925 Suzuki H., Shimada T., Terashima M., Tsuchiya K., Aplin K., 2004 Temporal, spatial, and ecological
926 modes of evolution of Eurasian Mus based on mitochondrial and nuclear gene sequences. Mol
927 Phylogenet Evol **33**: 626–646.

928 Svenson K. L., Gatti D. M., Valdar W., Welsh C. E., Cheng R., Chesler E. J., Palmer A. A., McMillan L.,
929 Churchill G. A., 2012 High-resolution genetic mapping using the mouse Diversity Outbred population.
930 Genetics **190**: 437–447.

931 Swallow J. G., Carter P. A., Jr. T. G., 1998 Artificial selection for increased wheel-running behavior in
932 house mice. Behav Genet **28**: 227–237.

933 Treangen T. J., Salzberg S. L., 2011 Repetitive DNA and next-generation sequencing: Computational
934 challenges and solutions. Nat Rev Genet.

935 Turner J. M., Aprelikova O., Xu X., Wang R., Kim S., Chandramouli G. V., Barrett J., Burgoyne P. S., Deng
936 C.-X., 2004 BRCA1, histone H2AX phosphorylation, and male meiotic sex chromosome inactivation. Curr
937 Biol **14**: 2135–2142.

938 Turner D. J., Miretti M., Rajan D., Fiegler H., Carter N. P., Blayney M. L., Beck S., Hurles M. E., 2007
939 Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. Nat
940 Genet **40**: 90–95.

941 Waterston R. H., Chinwalla A. T., Cook L. L., Delehaunty K. D., Fewell G. A., Fulton L. A., Fulton R. S.,
942 Graves T. A., Hillier L. W., Mardis E. R., McPherson J. D., Miner T. L., Nash W. E., Nelson J. O., Nhan M.
943 N., Pepin K. H., Pohl C. S., Ponce T. C., Schultz B., Thompson J., Trevaskis E., Waterston R. H., Wendl M.
944 C., Wilson R. K., Yang S.-P., An P., Berry E., Birren B., Bloom T., Brown D. G., Butler J., Daly M., David R.,
945 Deri J., Dodge S., Foley K., Gage D., Gnerre S., Holzer T., Jaffe D. B., Kamal M., Karlsson E. K., Kells C.,
946 Kirby A., Kulbokas E. J., Lander E. S., Landers T., Leger J. P., Levine R., Lindblad-Toh K., Mauceli E.,
947 Mayer J. H., McCarthy M., Meldrim J., Meldrim J., Mesirov J. P., Nicol R., Nusbaum C., Seaman S., Sharpe
948 T., Sheridan A., Singer J. B., Santos R., Spencer B., Stange-Thomann N., Vinson J. P., Wade C. M.,
949 Wierzbowski J., Wyman D., Zody M. C., Birney E., Goldman N., Kasprzyk A., Mongin E., Rust A. G.,
950 Slater G., Stabenau A., Ureta-Vidal A., Whelan S., Ainscough R., Attwood J., Bailey J., Barlow K., Beck S.,
951 Burton J., Clamp M., Clee C., Coulson A., Cuff J., Curwen V., Cutts T., Davies J., Eyras E., Grafham D.,
952 Gregory S., Hubbard T., Hunt A., Jones M., Joy A., Leonard S., Lloyd C., Matthews L., McLaren S., McLay
953 K., Meredith B., Mullikin J. C., Ning Z., Oliver K., Overton-Larty E., Plumb R., Potter S., Quail M., Rogers
954 J., Scott C., Searle S., Shownkeen R., Sims S., Wall M., West A. P., Willey D., Williams S., Abril J. F., Guigó
955 R., Parra G., Agarwal P., Agarwala R., Church D. M., Hlavina W., Maglott D. R., Sapojnikov V.,
956 Alexandersson M., Pachter L., Antonarakis S. E., Dermitzakis E. T., Reymond A., Ucla C., Baertsch R.,
957 Diekhans M., Furey T. S., Hinrichs A., Hsu F., Karolchik D., Kent W. J., Roskin K. M., Schwartz M. S.,
958 Sugnet C., Weber R. J., Bork P., Letunic I., Suyama M., Torrents D., Zdobnov E. M., Botcherby M., Brown
959 S. D., Campbell R. D., Jackson I., Bray N., Couronne O., Dubchak I., Poliakov A., Rubin E. M., Brent M. R.,
960 Flicek P., Keibler E., Korf I., Batalov S., Bult C., Frankel W. N., Carninci P., Hayashizaki Y., Kawai J.,
961 Okazaki Y., Cawley S., Kulp D., Wheeler R., Chiaromonte F., Collins F. S., Felsenfeld A., Guyer M.,
962 Peterson J., Wetterstrand K., Copley R. R., Mott R., Dewey C., Dickens N. J., Emes R. D., Goodstadt L.,
963 Ponting C. P., Winter E., Dunn D. M., Niederhausern A. C. von, Weiss R. B., Eddy S. R., Johnson L. S.,
964 Jones T. A., Elnitski L., Kolbe D. L., Eswara P., Miller W., O'Connor M. J., Schwartz S., Gibbs R. A.,
965 Muzny D. M., Glusman G., Smit A., Green E. D., Hardison R. C., Yang S., Haussler D., Hua A., Roe B. A.,
966 Kucherlapati R. S., Montgomery K. T., Li J., Li M., Lucas S., Ma B., McCombie W. R., Morgan M., Pevzner
967 P., Tesler G., Schultz J., Smith D. R., Tromp J., Worley K. C., Lander E. S., Abril J. F., Agarwal P.,
968 Alexandersson M., Antonarakis S. E., Baertsch R., Berry E., Birney E., Bork P., Bray N., Brent M. R., Brown
969 D. G., Butler J., Bult C., Chiaromonte F., Chinwalla A. T., Church D. M., Clamp M., Collins F. S., Copley R.
970 R., Couronne O., Cawley S., Cuff J., Curwen V., Cutts T., Daly M., Dermitzakis E. T., Dewey C., Dickens
971 N. J., Diekhans M., Dubchak I., Eddy S. R., Elnitski L., Emes R. D., Eswara P., Eyras E., Felsenfeld A.,
972 Flicek P., Frankel W. N., Fulton L. A., Furey T. S., Gnerre S., Glusman G., Goldman N., Goodstadt L.,
973 Green E. D., Gregory S., Guigó R., Hardison R. C., Haussler D., Hillier L. W., Hinrichs A., Hlavina W.,
974 Hsu F., Hubbard T., Jaffe D. B., Kamal M., Karolchik D., Karlsson E. K., Kasprzyk A., Keibler E., Kent W.
975 J., Kirby A., Kolbe D. L., Korf I., Kulbokas E. J., Kulp D., Lander E. S., Letunic I., Li M., Lindblad-Toh K.,
976 Ma B., Maglott D. R., Mauceli E., Mesirov J. P., Miller W., Mott R., Mullikin J. C., Ning Z., Pachter L.,
977 Parra G., Pevzner P., Poliakov A., Ponting C. P., Potter S., Reymond A., Roskin K. M., Sapojnikov V.,

978 Schultz J., Schwartz M. S., Schwartz S., Searle S., Singer J. B., Slater G., Smit A., Stabenau A., Sugnet C.,
979 Suyama M., Tesler G., Torrents D., Tromp J., Ucla C., Vinson J. P., Wade C. M., Weber R. J., Wheeler R.,
980 Winter E., Yang S.-P., Zdobnov E. M., Whelan S., Worley K. C., Zody M. C., 2002 Initial sequencing and
981 comparative analysis of the mouse genome. Nature **420**: 520–562.

982 Watterson G., 1975 On the number of segregating sites in genetical models without recombination. Theor
983 Popul Biol **7**: 256–276.

984 White M. A., Ané C., Dewey C. N., Larget B. R., Payseur B. A., 2009 Fine-scale phylogenetic discordance
985 across the house mouse genome. PLoS Genet **5**: e1000729.

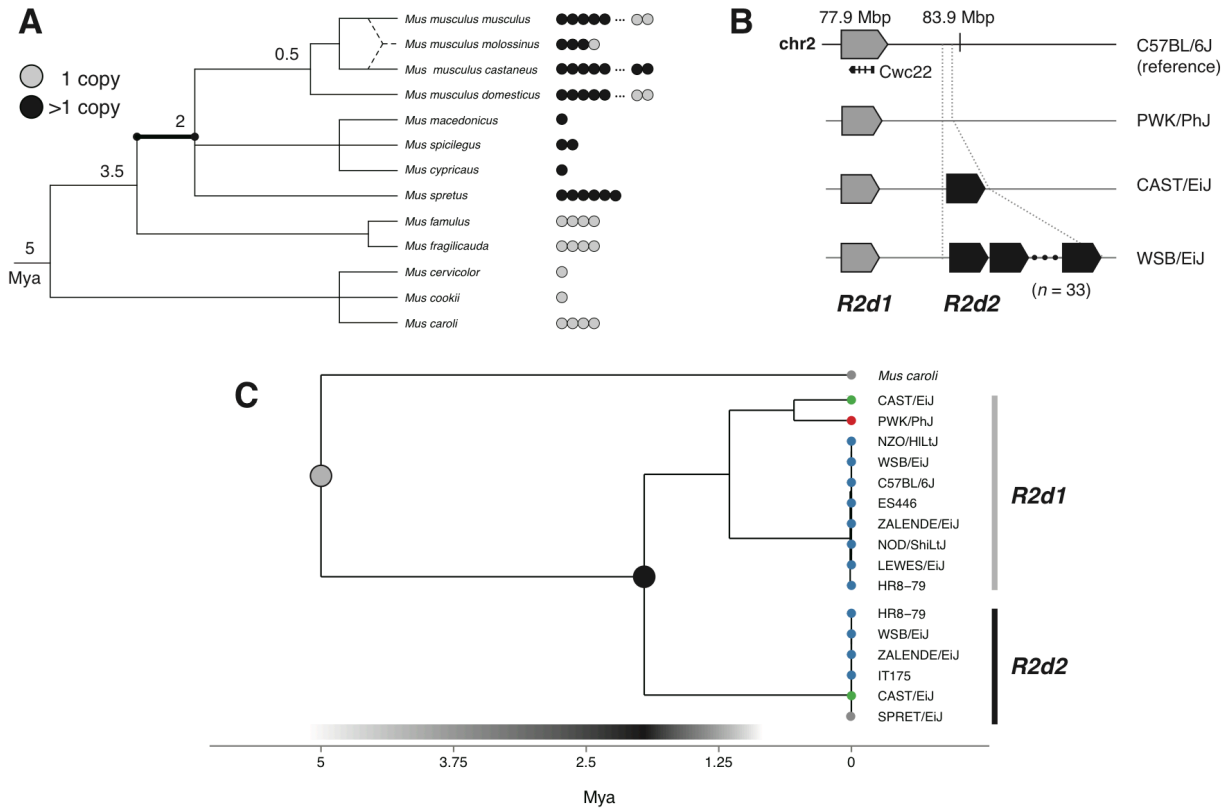
986 Yang H., Bell T. A., Churchill G. A., Villena F. P.-M. de, 2007 On the subspecific origin of the laboratory
987 mouse. Nat Genet **39**: 1100–1107.

988 Yang H., Wang J. R., Didion J. P., Buus R. J., Bell T. A., Welsh C. E., Bonhomme F., Yu A. H.-T., Nachman
989 M. W., Pialek J., Tucker P., Boursot P., McMillan L., Churchill G. A., Villena F. P.-M. de, 2011 Subspecific
990 origin and haplotype diversity in the laboratory mouse. Nat Genet **43**: 648–655.

991 Yeh T.-C., Liu H.-L., Chung C.-S., Wu N.-Y., Liu Y.-C., Cheng S.-C., 2010 Splicing factor Cwc22 is required
992 for the function of Prp2 and for the spliceosome to escape from a futile pathway. Mol Cell Biol **31**: 43–53.

993

994 **FIGURE LEGENDS**



995

996 **Figure 1. Origin and age of the *R2d2* duplication.** (**A**) *R2d* copy number across the phylogeny of the
997 genus *Mus*. Each dot represents one individual; grey dots indicate copy number 1 and black dots copy
998 number >1. The duplication event giving rise to *R2d1* and *R2d2* most likely occurred on the highlighted
999 branch. Approximate divergence times (REF: Suzuki 2004) are given in millions of years ago (Mya) at
1000 internal nodes. (**B**) Schematic structure of the *R2d1-R2d2* locus. The mouse reference genome (strain
1001 C57BL/6J, *M. m. domesticus*) contains a single copy of *R2d* at *R2d1*. Wild-derived inbred strains vary in
1002 copy number from 1 (PWK/PhJ, *M. m. musculus*) to 2 (CAST/EiJ, *M. m. castaneus*) to 33 (WSB/EiJ, *M. m.*
1003 *domesticus*). *R2d1* is located at approximately 77.9 Mbp and *R2d2* at 83.8 Mbp. (**C**) Representative tree
1004 constructed from *de novo* assembled *R2d1* and *R2d2* sequences assuming a strict molecular clock.
1005 Sequences are colored according to their subspecies of origin: *M. m. domesticus*, blue; *M. m. musculus*, red;
1006 *M. m. castaneus*, green; and the outgroup species *M. spretus* in grey. The duplication node is indicated
1007 with a black dot. The 95% HPDI for the age of the duplication event obtained by Bayesian phylogenetic
1008 analysis with BEAST is displayed below the tree.

1009

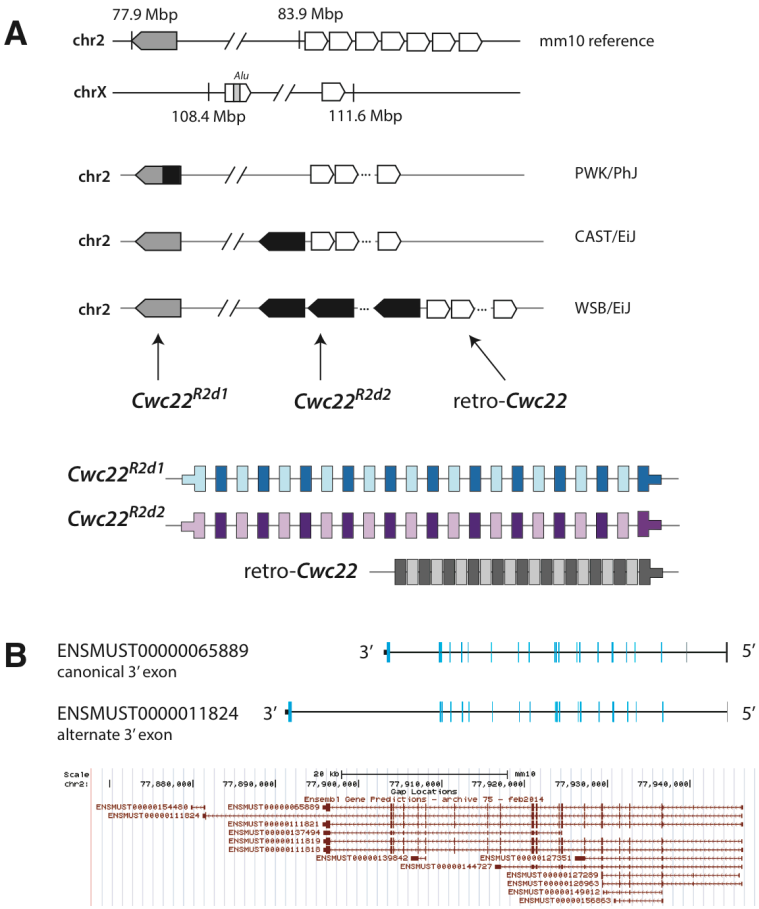


Figure 2. *Cwc22* paralogs in the mouse genome. (**A**) Location and organization of *Cwc22* gene copies present in mouse genomes. The intact coding sequence of *Cwc22* exists in in both *R2d1* (grey shapes) and *R2d2* (black shapes). Retrotransposed copies (empty shapes) exist in two loci on chrX and one locus on chr2, immediately adjacent *R2d2*. Among the retrotransposed copies, coding sequence is intact only in the copy on chr2. (**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.)

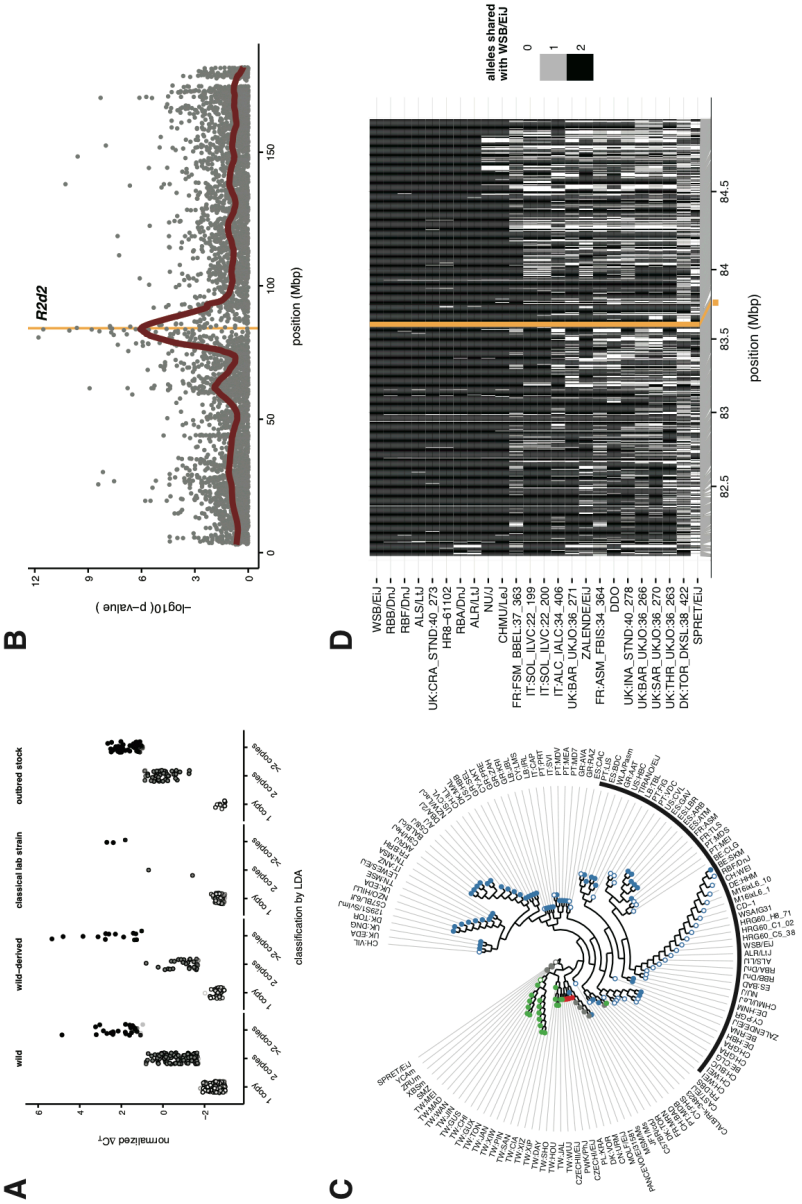


Figure 3. Copy-number variation of *R2d* in mouse populations worldwide. (**A**) Copy-number variation as measured by quantitative PCR. The normalized deltaCt value is proportional to log2(copy number). Samples are classified as having 1 copy, 2 copies or >2 copies of *R2d* using linear discriminant analysis (LDA). (**B**) Fine-mapping the location of *R2d2* in 83 samples genotyped on the Mouse Diversity Array (MDA). Grey 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: 82-85 Mbp in 135 wild-caught mice and wild-derived strains. Tips are colored by subspecies of origin: *M. m. domesticus*, blue; *M. m. musculus*, red; *M. m. castaneus*, green; other taxa, grey. Individuals with >2 copies of *R2d* are shown as open circles. Black arc indicates the portion of the tree enriched for individuals with high copy number. (**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.

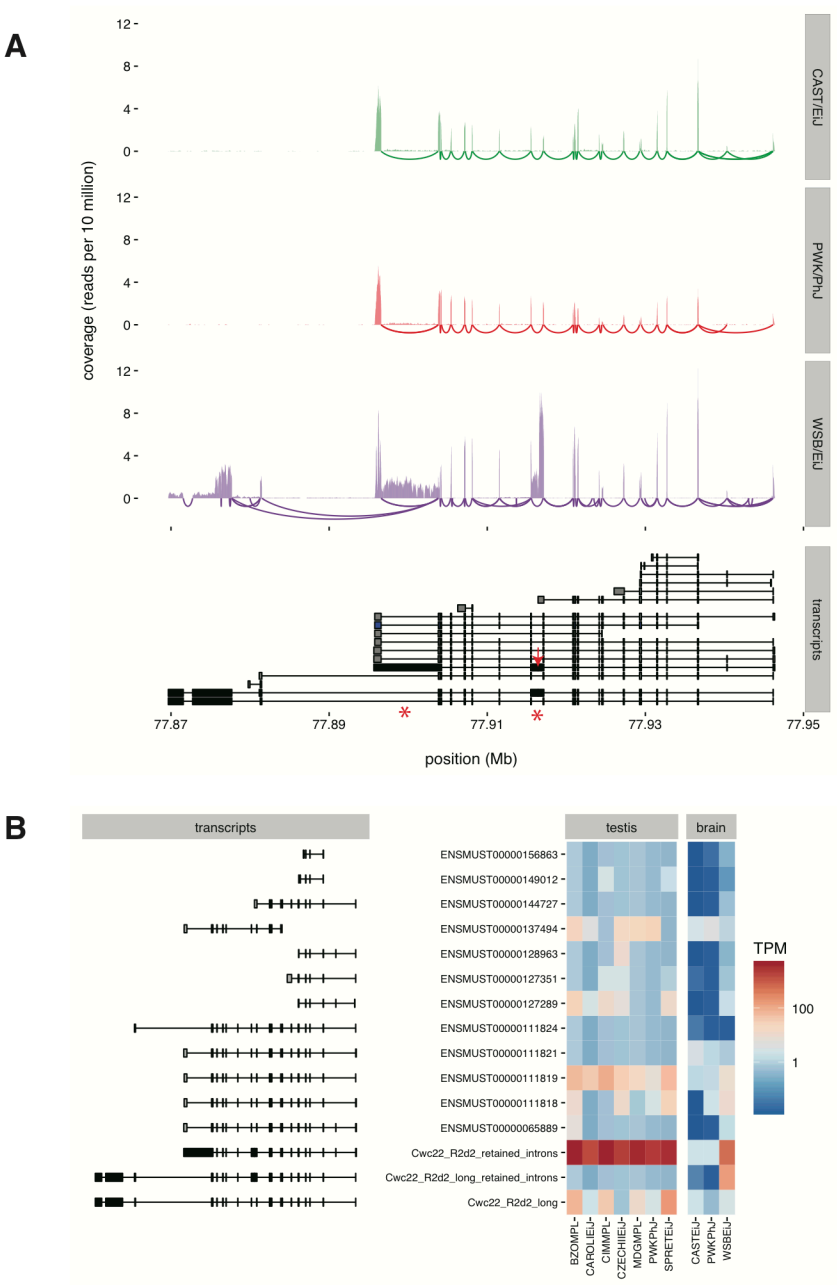


Figure 4. Expression of *Cwc22* isoforms. (**A**) Read coverage and splicing patterns in *Cwc22* in adult mouse brain from three wild-derived inbred strains. Swoops below x-axis indicate splicing events supported by 5 or more split-read alignments. Known transcripts of *Cwc22*^{R2d1} (grey, from Ensembl), inferred transcripts from *Cwc22*^{R2d2} (black) and the sequence of retro-*Cwc22* mapped back to the parent gene (blue) are shown in the lower 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.

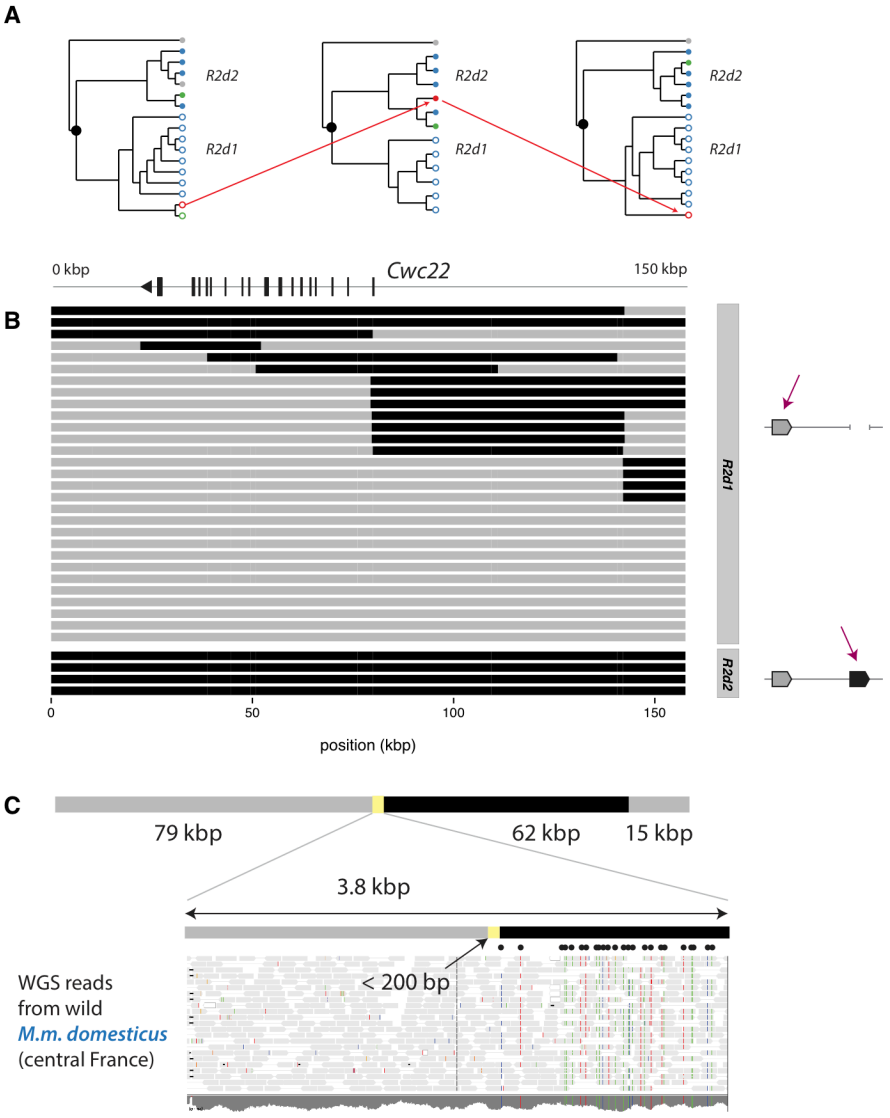
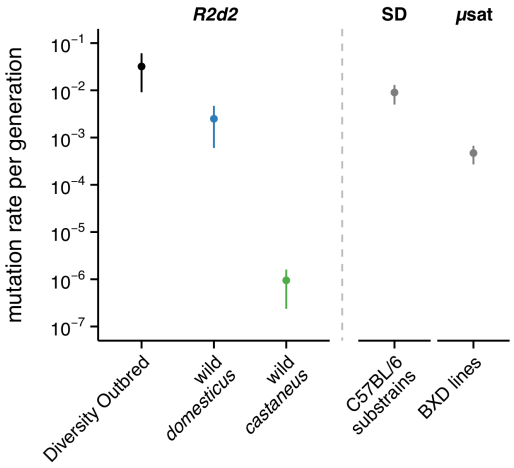


Figure 5. Signatures of non-allelic gene conversion between *R2d1* and *R2d2*. (**A**) Phylogenetic trees for three representative intervals across *R2d*. Sequences are labeled according to their subspecies of origin using the same color scheme as in **Figure 1**; open circles are *R2d1*-like sequences and closed 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 non-allelic gene conversion. (**B**) *R2d* sequences from 20 wild-caught mice and 5 laboratory inbred strains. Each track represents a single chromosome; grey regions are classified as *R2d1*-like based on manual inspection of sequence variants, and black-regions *R2d2*-like. Upper panel shows sequences from samples with a single copy of *R2d*, residing in *R2d1*. Lower panel shows representative *R2d2* sequences for comparison. (**C**) Read alignments at the boundary of a non-allelic gene conversion tract. The *R2d1* sequence from a single chromosome from mouse trapped in central France is a mosaic of *R2d1*-like (grey) and *R2d2*-like (black) segments. A magnified view of read pairs in the 3.8 kbp surrounding the proximal boundary of the tract (generated with IGV) reveals 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.



1058

1059 **Figure 6. Rate of *de novo* copy-number changes at *R2d2*.** Estimates of per-generation mutation rate for
1060 CNVs at *R2d2* (±1 bootstrap SE) in the Diversity Outbred population; among wild *M. m. domesticus;* and
1061 among wild *M. m. castaneus*. For comparison, mutation rates are shown for the CNV with the highest rate
1062 of recurrence in a C57BL/6J pedigree (Egan *et al.* 2007) and for a microsatellite whose mutation rate was
1063 estimated in the BXD panel (Dallas 1992).

1064

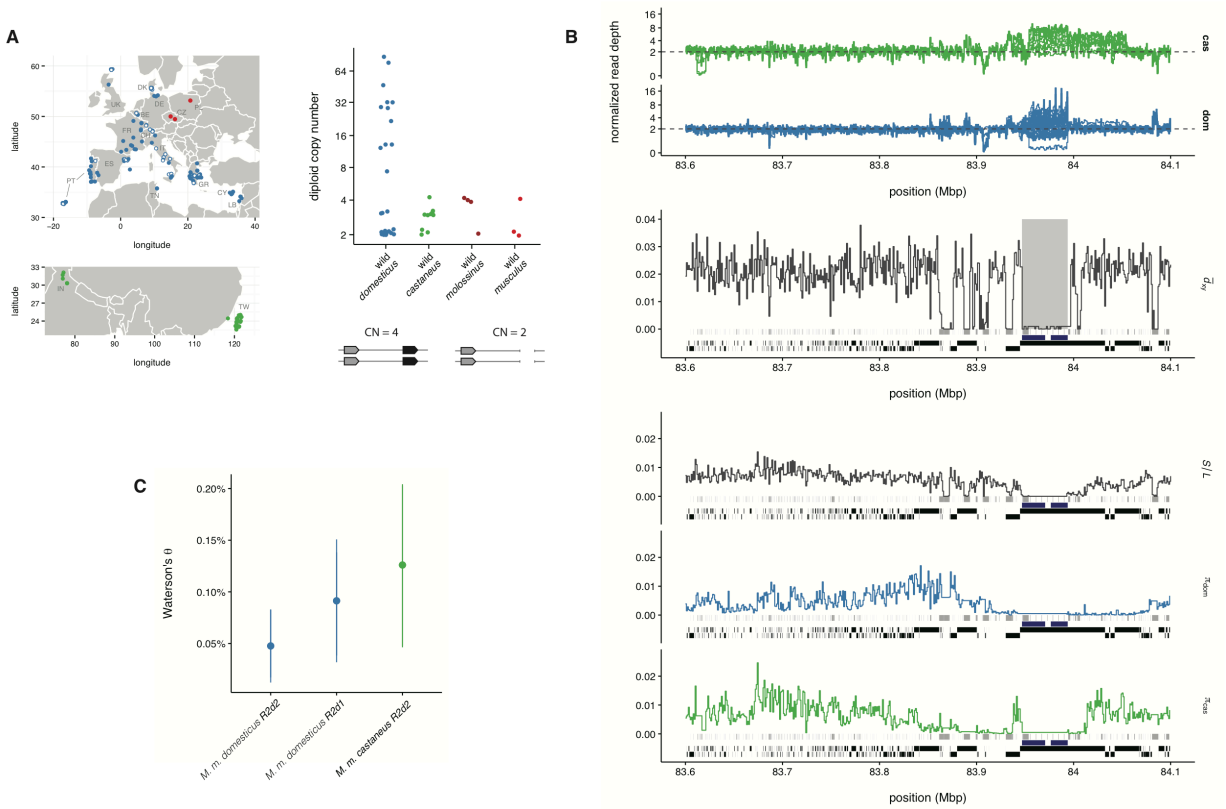


Figure 7. Sequence and structural diversity around *R2d2*. (**A**) Geographic origin of wild mice used in this study, color-coded by subspecies (blue, *M. m. domesticus*; red, *M. m. musculus*; green, *M. m. castaneus*). Diploid copy number of the *R2d* unit is shown for wild samples for which integer 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 2 and 4. (**B**) Profiles of read depth (first panel), average sequence divergence to outgroup species *M. caroli* (d_{xy}, second panel), number of segregating sites per base (S/L, third panel) and within-population average heterozygosity (π, fourth panel). The region shown is 500 kbp in size and centered on the insertion site of *R2d2*. Grey boxes along baseline show positions of repetitive elements (from UCSC RepeatMasker track); black boxes show non-recombining haplotype blocks. Blue bars indicate the position of 7 tandem duplications in the mm10 reference sequence with >99% mutual identity, each containing a copy of retro-*Cwc22*. The duplications are absent in *M. caroli* (indicated by grey shaded box.)

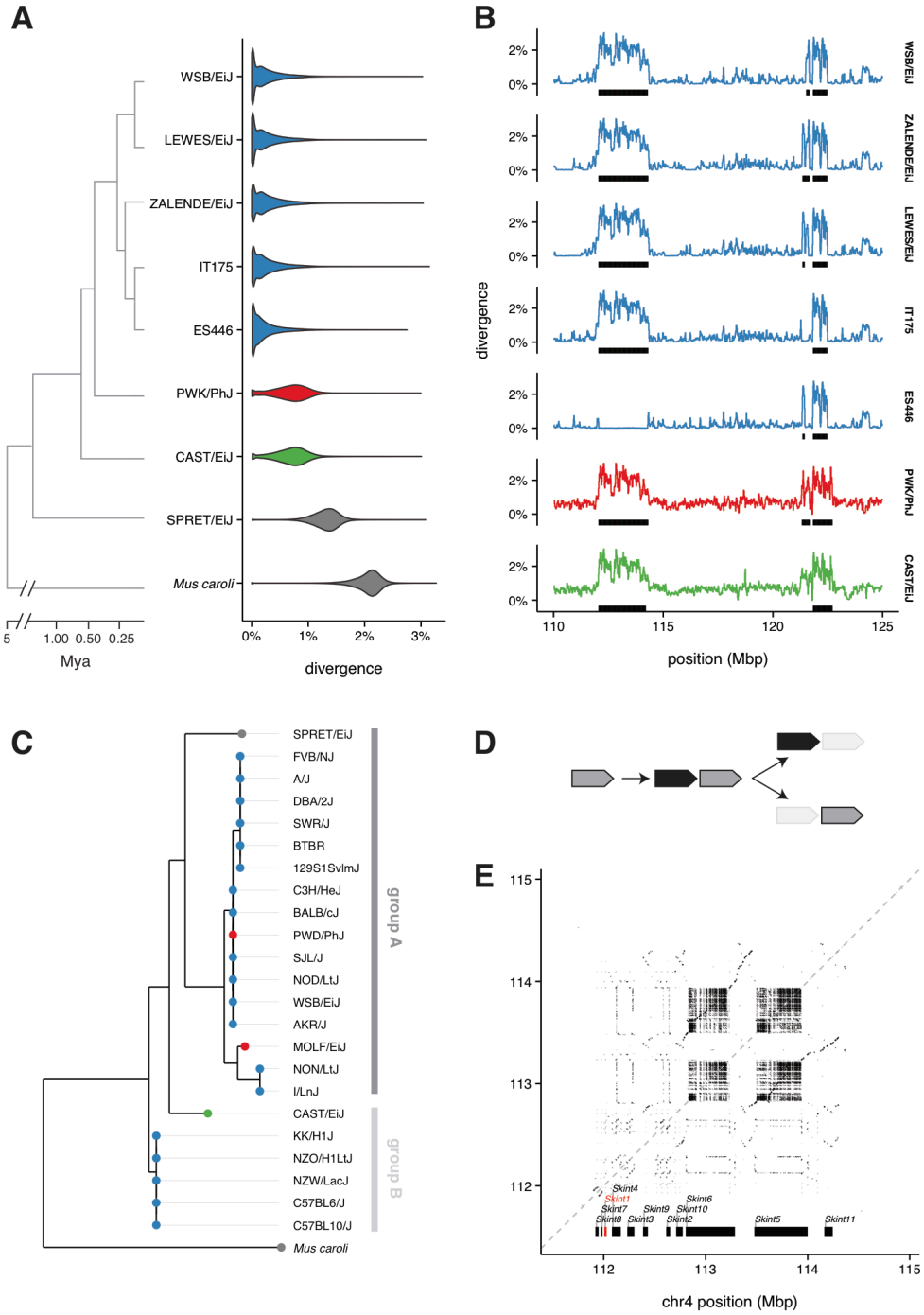


Figure 8. (**A**) Genome-wide sequence divergence estimates for representative samples from the sub-genus *Mus*. (**B**) Estimated sequence divergence in 31 kbp windows across distal chr4 for the samples in panel A. Divergent regions identified by the hidden Markov model (HMM) are indicated with black bars along the horizontal axis. (**C**) Phylogenetic tree constructed from *Skint1* coding sequences reported in Boyden *et al.* (2008) (**D**) Schematic representation of the process of gene duplication, followed by differential loss of paralogs along independent lineages. (**E**) Dotplot of self-alignment of sequence from the region of distal chr4 containing the *Skint* gene family. Positions of *Skint* genes are indicated along the horizontal axis; *Skint1* highlighted in red.

1089 ## SUPPLEMENTARY MATERIAL

1090 **Supplementary Figure 1.** Conservation of synteny between mouse and four other mammals around
1091 *Cwc22*^{R2d1} (upper panel) indicates that the *R2d1* sequence remains in its ancestral position. Chevrons
1092 represent genes, alternating white and grey, and are oriented according to the strand on which the gene is
1093 encoded. *Cwc22*^{R2d2} is novel in the mouse but its position relative to genes with conserved order is shown
1094 in the lower panel. Note that synteny is disrupted in mouse and rat distal to *R2d2*.

1095 **Supplementary Figure 2.** Pairwise alignment of *R2d2* contig (top) to the *R2d1* reference sequence
1096 (bottom). Dark boxes show position of repetitive elements present in both sequences; syntenic positions
1097 are connected by grey anchors, and blank space represents aligned bases in both sequences. Orange boxes
1098 indicate position of repetitive elements present in the *R2d1* sequence but not detected in *R2d2*; blue boxes
1099 indicate position of elements in *R2d2* but not *R2d1*. *Cwc22* transcripts are shown below the alignment.

1100 **Supplementary Figure 3.** Phylogenetic tree constructed from amino acid sequences for mammalian
1101 *Cwc22* homologs (including all three mouse paralogs) with chicken as an outgroup. Node labels indicate
1102 support in 100 bootstrap replicates.

1103 **Supplementary Figure 4.** Alignment of amino acid sequences from mouse *Cwc22*^{R2d1}, *Cwc22*^{R2d2} and retro-
1104 *Cwc22*, plus *Cwc22* orthologs from 19 other placental mammals plus opossum, platypus and chicken as
1105 outgroups. Residues are colored according to biochemical properties and gaps are shown in grey.
1106 Information content of each column in the alignment, measured as the Jenson-Shannon divergence, is
1107 plotted in the lower panel.

1108 **Supplementary Figure 5.** Gene conversion tracts identified by *de novo* assembly. (**A**) Phylogenetic trees
1109 for twelve intervals across *R2d*. Samples are labeled according to their subspecies of origin using the same
1110 color scheme as in **Figure 1.** Trees are drawn so that *M. caroli*, the outgroup species used to root the trees,
1111 is always positioned at the top. The changing affinities of PWK/PhJ (red) and CAST/EiJ (green) along *R2d*
1112 is evidence of non-allelic gene conversion. (**B**) Inspection of derived alleles diagnostic for *R2d1* or *R2d2*
1113 reveals conversion tracts. Each horizontal track represents a haplotype, and each dot a variant site. Filled
1114 dots are fully diagnostic for *R2d1* or *R2d2*; open circles are partially-informative. Positions with *R2d1*-like
1115 sequence are colored grey, and those with *R2d2*-like sequence colored black. Conversion tracts are
1116 indicated by yellow boxes. Physical positions of variant sites within *R2d* are shown with respect to the
1117 *R2d1* sequence present in the mouse reference genome. The *Cwc22* gene spans conversion tracts in
1118 multiple samples.

1119 **Supplementary Figure 6.** Partial loss of *R2d2* with structural rearrangement. (**A**) Inferred structure of the
1120 *R2d1-R2d2* region in IR:AHZ_STND:015, a wild *M. m. domesticus* individual from Iran. *R2d1* is present on
1121 both chromosomes but only a fragment of *R2d2* remains on one chromosome, and it has been transposed
1122 into the retro-*Cwc22* array. (**B**) Normalized depth of coverage (2 = normal diploid level) across *R2d*.
1123 Regions in grey represent reads from *R2d1* alone, while region in black captures reads from *R2d1* and
1124 *R2d2*, as shown by arrows from panel A. (**C**) Position of read pairs (red; not drawn to scale) with soft-
1125 clipped alignments to *R2d1*. The proximal read aligns in the 3′ UTR of *Cwc22*, and the distal read across
1126 an exon-intron boundary within the gene body. Note the "outward"-facing direction of the alignments.
1127 (**D**) Positions of the mates of the reads in panel C. Note that the x-axis is reversed so that the exons of
1128 retro-*Cwc22* (encoded on the plus strand) parallel those of *Cwc22* (encoded on the minus strand). The 3′
1129 read maps across the boundary of th 3′ UTR of *Cwc22* and the ERV mediating the retrotransposition
1130 event. The 5′ read maps across two exon-exon boundaries in retro-*Cwc22*, so there is no ambiguity
1131 regarding its alignment to the retro-transposed copy. (**E**) Inferred structure of *Cwc22* paralogs in this

1132 sample. Note that one of the copies of retro-*Cwc22* is now a mosaic of retrotransposed and *Cwc22*^{R2d2}-
1133 derived sequence.

1134 **Supplementary Figure 7. Suppression of crossing-over around *R2d2*.** (**A**) Difference between expected
1135 and observed recombination fraction between markers flanking *R2d2* in experimental crosses in which at
1136 least one parent is segregating for a high-copy allele of *R2d2*. Thick and thin vertical bars show 90% and
1137 95% confidence bounds, respectively, obtained by non-parametric bootstrap. (**B**) Cumulative
1138 recombination map in the middle region of chr2 obtained from 4,640 Diversity Outbred mice.
1139 Recombination events involving the WSB/EiJ haplotype (*R2d2* copy number 33) are shown in purple and
1140 all other events in grey. Maps are normalized such that they begin and end at the same value.

1141 **Supplementary Figure 8. Targeted *de novo* assembly using the multi-string Burrows-Wheeler**
1142 **Transform (msBWT).** (**A**) The msBWT and its associated FM-index implicitly represent a suffix array of
1143 sequencing reads, such that read suffixes sharing a k-mer prefix are adjacent in the data structure. This
1144 allows rapid construction of a local de Bruijn graph starting from a k-mer seed (dark blue) and extending
1145 by successive k-mers (light blue) containing the $(k-1)$-length suffix of the previous k-mer. A $(k-1)$-
1146 length prefix with more than one possible suffix (red and orange) creates a branch point. Adjacent nodes
1147 in the graph with in-degree and out-degree one can be collapsed into a single node, yielding a simplified
1148 graph, which can then be traversed to obtain linear contig(s). (**B**) Paralogs of *R2d* can be disentangled
1149 using the local de Bruijn graph by exploiting differences in copy number. Edges in the graph are
1150 weighted by read count, and linear contigs for the *R2d1* and *R2d2* paralogs obtained by traversing the
1151 graph in a manner that minimizes the variance in edge weights along possible paths. Phase-informative
1152 reads (those overlapping multiple paralogous variants) provide a second source of evidence.

1153

1154 **Supplementary Table 1.** List of mouse samples used in this study, with their taxonomic designation,
1155 geographic origin and *R2d2* copy-number classification.

1156 **Supplementary Table 2.** Transposable-element insertions private to *R2d1* or *R2d2*. Coordinates are
1157 offsets with respect to the start position of *R2d* (for *R2d1*: chr2: 77,869,657 in the reference genome; for
1158 *R2d2*: the beginning of the *de novo* assembled contig in **Supplementary File 1**.)

1159 **Supplementary Table 3.** Individuals from the Diversity Outbred population carrying *de novo* copy-
1160 number mutations at *R2d2*. Each was expected to be heterozygous for the WSB/EiJ allele (33 haploid
1161 copies).

1162 **Supplementary Table 4.** Regions of excess divergence between wild or wild-derived mice and the mouse
1163 reference genome (GRCm38/mm10 build).

1164 **Supplementary Table 5.** Regions of R2d targeted for *de novo* assembly in inbred strains.

1165 **Supplementary File 1.** Compressed archive containining *R2d2* contig (from WSB/EiJ) and multiple
1166 sequence alignments from selected regions in **Supplementary Table 5**.