Natural history of a copy-number variant in mouse

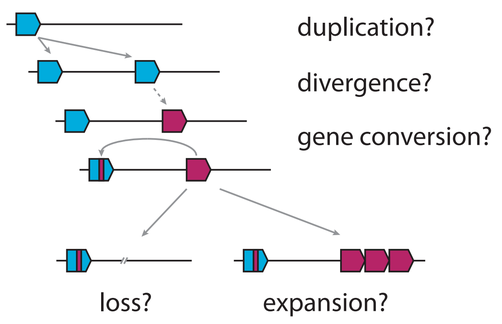
Andrew P Morgan[1], Rachel C McMullan[1], John P Didion[1], Timothy A Bell[1], J Matthew Holt[2], Leonard McMillan[2], Fernando Pardo-Manuel de Villena[1]

[1]Carolina Center for Genome Sciences and Department of Genetics; [2]Department of Computer Science, University of North Carolina at Chapel Hill

Background

Departures from expected Mendelian segregation (transmission ratio distortion, TRD) on mouse chr2 have been reported in several independent crosses. In all cases, TRD is in favor of alleles with a copy-number gain at chr2: 77.9 Mbp. We have recently shown that this due, in part, to female meiotic drive[1]. High-copy alleles have been subject to rapid selective sweeps in both laboratory and wild populations (see poster **P-14**).

In order to understand and correctly interpret extant patterns of sequence variation around this locus, we aimed to reconstruct its evolutionary history using whole-genome sequence from samples across the phylogeny of *Mus*. Specifically, we sought to account for the following events:



msBWT: new tool for sequence analysis

The multi-string Burrows-Wheeler transform (`msBWT`) is a compressed representation of unaligned reads with an associated index which allows efficient search[2]. Queries for sequences of length k take $O(k)$ time, regardless of the size of the dataset.

```
                    ???:CTGGCCTGTCCACAGTGTC
------CATGAGCCTCATTTATCATGCCTTCTGGCCTGTCCACAGTGTCTAACATA----------------------
----------AGCCTCATTTATCATGCCTTCTGGCCTGTCCACAGTGTCTAACATAGTAC----------------
----------GCCTCATTTATCATGCCTTCTGGCCTGTCCACAGTGTCTAACATAGTACT----------------
--------------CATTTATCATGCCTTCTGGCCTGTCCACAGTGTCTAACATAGTACTAGAT-----------
--------------------TATCATGCCTTCTGGCCTGTCCACAGTGTCTAATATAGTACTAGATACAC-------
--------------------TATCATGCCTTCTGGCCTGTCCACAGTGTCTAATATAGTACTAGATACAC-------
--------------------TATCATGCCTTCTGGCCTGTCCACAGTGTCTAATATAGTACTAGATACAC-------
----------------------CATGCCTTCTGGCCTGTCCACAGTGTCTAATATAGTACTAGATACACAAT----
```

The `msBWT` enables targeted *de novo* assembly and variant discovery, even in repetitive sequence.

Acknowledgements

Whole-genome sequence for CAROLI/EiJ was obtained on pre-publication release from the Wellcome Trust Sanger Institute's Caroli Genome Project (NCBI BioProject `PRJEB2188`). Wild mouse samples were contributed to JPD by Jeremy Searle and: Francois Bonhomme, Pierre Boursot, Janice Britton-Davidian, Ricardo Castiglia, Eva Giagia-Athan-asopolou, Sofia Gabriel, Silvia Garagna, Sofia Grize, Isla Gündüz, Bettina Harr, Heidi Hauffe, Jeremy Herman, Leon Kontrimavicius, Anna Lindholm, Maria de Luz Mathias, George Mistainas, Jaroslav Pialek, Priscilla Tucker, Jacint Ventura and Jan Wojcik.

This work was supported by the the Jackson Lab Center for Genome Dynamics (P50 GM076468). APM is supported by T32 GM067553 and F30 MH103925.

References

[1] Didion JP *et al.* (2014) *PLoS Genet*, in review.

[2] Holt JM, McMillan L (2014) *Bioinformatics* epub 25 Aug 2014.

[3] Keane TM *et al.* (2011) *Nature* **477**: 289-294.

[4] Suzuki H *et al.* (2004) *Mol Phylogenet Evol* **33**: 626-646.

[5] Didion JP *et al.* Wild Mouse Genetic Survey, in preparation.

[6] Nachman MW *et al.* (1994) *Genetics* **163**: 1105-1120.

[7] Nagylaki T (2014) *Genetics* **106**: 529-548.

Conclusions

Using whole-genome sequence from three mouse species, we have shown that the multiple copies of *R2d* present in several classical and wild-derived laboratory strains represent a segmental duplication ancestral to the divergence of *Mus musculus*. The resulting duplicates in the *R2d1* and *R2d2* loci have undergone multiple independent gene conversion events. Evidence that copies of *R2d* in the distal locus are unstable is twofold. First, multiple independent copy-number losses have occurred within the *M. m. musculus* and *M. m. domesticus* lineages. Second, although expansion alleles at *R2d2* arose recently on a single haplotype, copy number at *R2d2* is highly polymorphic in wild *M. m. domesticus*.

Evidence for a CNV

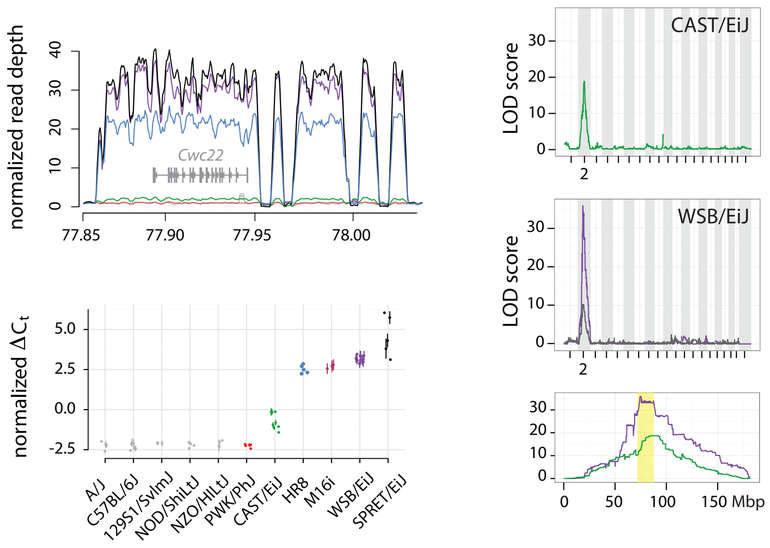


Figure 1. Copy-number gain over 150 kbp at chr2: 77.9 Mbp in four inbred strains, revealed by sequencing[3] and confirmed by qPCR. The extra copies are not in tandem: they map ~ 6 Mbp away[1].

Structure of the locus

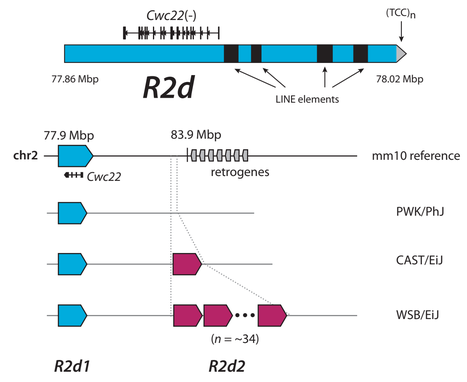


Figure 2. A single copy of the 150 kbp duplicated sequence (named "responder to (meiotic) drive", *R2d*) is located chr2: 77.9 Mbp (locus *R2d1*) in the reference genome. Additional copies are located at *R2d2* in other strains.

The *R2d* unit contains a single protein-coding gene, *Cwc22*. It encodes a spliceosome protein which is essential for development in mouse. Four recent LINE insertions are present in the reference genome, and the unit terminates in a $(\mathrm{TCC})_n$ microsatellite.

Phylogenetic context

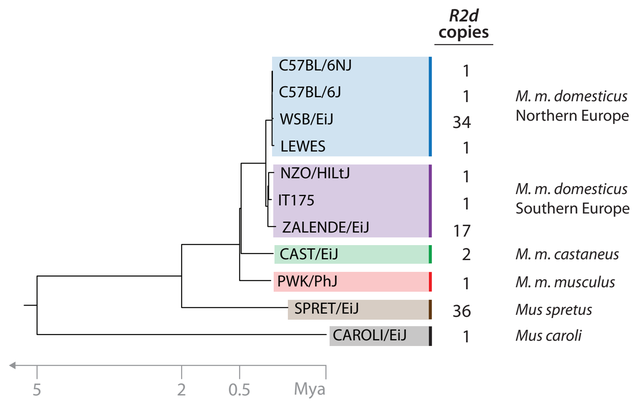


Figure 3. Mitochondrial phylogeny of subgenus *Mus* for samples used in this study (dates from [4]), with copy number in *R2d1* + *R2d2*.

We assembled *R2d* sequence(s) in 2 classical and 8 wild-derived inbred strains, plus 1 wild-caught mouse (IT175). The samples span all 3 subspecies of *M. musculus* as well as outgroups *M. spretus* and *M. caroli*.

Ancestral duplication of *R2d*

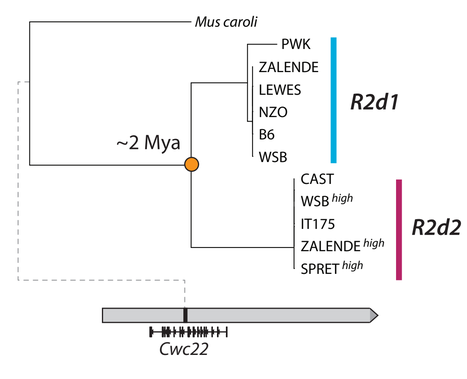


Figure 4. Maximum-likelihood phylogeny built from 1557 bp of non-coding sequence in *R2d*, assembled from Illumina sequence reads using the `msBWT` toolkit[2].

R2d was duplicated ~ 2 million years ago (orange node); since then, the copies in *R2d1* and *R2d2* have diverged by $\sim 2.5\%$. Sequence diversity within *R2d2* is very low.

R2d2 expansion alleles of recent European origin

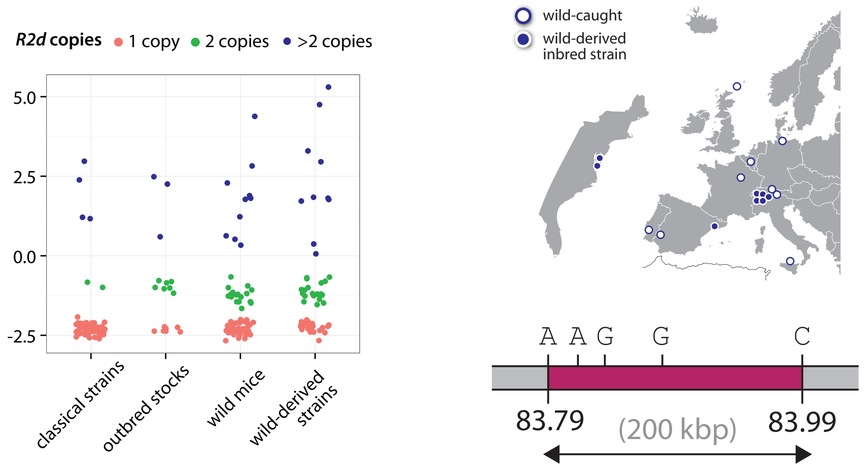


Figure 5. *R2d* copy number varies widely in laboratory and wild mice of *M. m. domesticus* origin (left). Expansion alleles – those with copy number > 2 – are widespread in Europe.

We surveyed *R2d* copy number by qPCR in a large panel of laboratory and wild mice[5]. We found that expansion alleles (copy number > 2) are common in mice of *M. m. domesticus* origin (MAF = 0.21 in the wild) and are associated exclusively with a single 200 kbp haplotype.

This haplotype is present both in populations with Robertsonian karyotypes and those with standard karyotypes, suggesting that it arose before the separation of the chromosomal races ($\sim 5 - 10$ kya[6]).

Inter-locus gene conversion between *R2d1* and *R2d2*

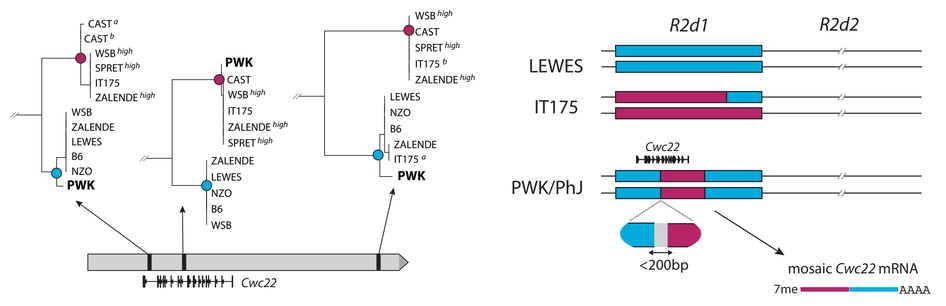


Figure 6. Changes in tree topology along *R2d* (left) are a signature of inter-locus gene conversion[7]. Resulting mosaic *R2d1* alleles are shown at right.

The presence of *R2d2*-like sequence within *R2d1* in extant mouse lineages with a single *R2d* copy is evidence of inter-locus gene conversion prior to the loss of copies at *R2d2*. PWK/PhJ shows evidence of a conversion event between exons of *Cwc22*, such that transcripts are a chimera of two sequences separated by 2 million years of evolution.