SUPPLEMENTARY MATERIAL

The rate and spectrum of new mutations in mice inferred by long-read sequencing

Appendix S1. Chromosome-level genome assemblies

Quality of assemblies

The software hifiasm 0.16.1 (Cheng et al. 2021, 2022) was the main tool used to build *de novo* assemblies for all samples from their PacBio HiFi reads. In the case of the founders of the mutation accumulation (MA) experiment, the pooled reads from the male and female founders from each strain were used to build a highly contiguous assembly, here referred to as the pooled founders' assembly. This assembly from pooled reads was then used as a draft for a chromosome-level assembly (see below).

Primary assemblies were always considered when measuring the assembly quality parameters, reported in Table S1. Coverage values were obtained from the peak of the *k*-mer spectrum generated by hifiasm. The contiguity of the assemblies was measured via the N50 value using the software quast 4.4 (Gurevich et al. 2013). The percentage of assembly completeness was assessed with the software BUSCO 5.4.3 (Manni et al. 2021) by searching single-copy orthologs against the glires genome database ("-m genome -l glires_odb10"). BUSCO analyses typically returned >96% completeness, with only ~2.8% missing and 0.6% fragmented orthologs for the chromosome-level assemblies. To estimate the error rate of the assemblies at the nucleotide level (QV), we first built a database of *k*-mers of length 21 bp using the software meryl 1.3 (Rhie et al. 2020) with the Illumina data from our MA experiment as input (López-Cortegano et al. 2024). QV was then estimated with merqury 1.3 (Rhie et al. 2020), showing values generally over 60 on the Phred scale across assemblies. As shown in Table S1, the read coverage had a high impact on the quality of the assemblies.

Table S1. Assembly quality parameters measured for MA samples and founders, including the pooled founders. Chromosome-level reference genomes are named after their corresponding strain. All parameters are measured at the contig level. The following parameters are presented: read coverage, assembly length (in Gb), N50 value, error rate (QV, in Phred scale), and proportion of complete BUSCOs. It is assumed that chromosome-level assemblies have the same read coverage as the pooled founders' assemblies from which they are derived.

Strain	Sample	Coverage	Length (Gb)	N. contigs	N50 (Mb)	QV	BUSCO (% complete)
C3H/HeNRj (C3H)	C3H/HeNRj v1	69×	2.72	154	65.51	63.06	96.6%
	Pooled founders	69×	2.90	345	63.97	62.21	92.1%
	Female founder	34×	2.75	404	46.40	61.03	96.4%
	Male founder	33×	2.78	212	37.44	60.68	96.4%
	1 (MA)	29×	2.78	202	44.29	61.08	96.3%
	2 (MA)	28×	2.79	219	35.44	59.37	96.4%
	3 (MA)	27×	2.77	334	26.45	59.09	96.3%
	4 (MA)	31×	2.78	198	45.42	61.38	96.4%

C57BL/6JRj (BL6)	C57BL/6JRj v1	52×	2.85	1,224	65.51	56.88	96.5%
	Pooled founders	52×	2.88	1,767	45.94	56.21	96.4%
	Female founder	26×	2.62	984	31.15	54.73	96.3%
	Male founder	25×	2.73	1,038	27.52	55.13	96.3%
	2 (MA)	15×	2.67	1,785	8.70	48.37	95.0%
	3 (MA)	20×	2.75	772	16.50	53.47	96.1%
	4 (MA)	15×	2.68	1,631	9.12	49.44	95.5%
	5 (MA)	14×	2.68	1,714	6.85	48.31	94.2%
FVB/NRj (FVB)	FVB/NRj v1	45×	2.88	857	49.79	57.08	96.6%
	Pooled founders	45×	2.89	1,046	49.79	56.68	96.5%
	Female founder	22×	2.68	1,145	25.08	54.24	96.2%
	Male founder	22×	2.73	1,160	32.90	53.92	96.4%
	1 (MA)	20×	2.73	1,006	24.46	54.02	95.9%
	2 (MA)	17×	2.68	1,530	17.48	51.62	96.0%
	3 (MA)	17×	2.72	1,273	10.70	50.18	96.1%
	4 (MA)	12×	2.62	2,650	3.85	45.50	94.2%

Manual scaffolding

To build chromosome-level reference genomes, the pooled founders' *de novo* assemblies were filtered to retain only contigs longer than 30 kb with an average read depth of more than 5 reads. Next, the assemblies were scaffolded to the chromosome level by mapping them against the GRCm39 reference genome. The mapping files were generated with minimap2 (Li 2018), considering only entries with mapping lengths longer than 30 kb and high mapping quality (MQ ≥ 60). The mapping of contigs was validated with Mashmap 3.1.2 (Jain et al. 2018). Assembly gaps were filled with unknown 'N' nucleotides, their length determined by mapping. In ambiguous cases, 100 'N' residues were added, such as in highly repetitive regions where the alignment of two contigs overlapped. Three instances of C3H contig overlap were merged in non-repetitive regions (with lengths from 88 bp to 22,274 bp) by removing the overlapping sequence from one of the contigs. Read and assembly alignments confirmed that these contig breaks were not mistakenly closed. Circos plots in Figures S1-S3 show the chromosome-level assemblies along with parameters related to the 'callability' of the genome, which are discussed below.

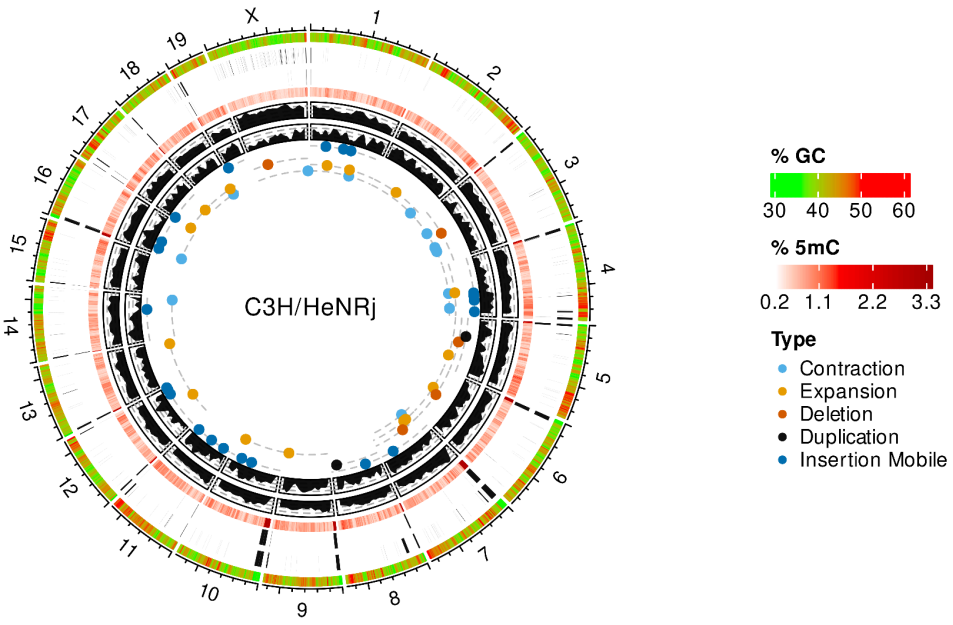

Figure S1. Circos plot (Krzywinski et al. 2009) of the C3H/HeNRj genome, generated using the R package circlize 0.4.15 (Gu et al. 2014). Each sector represents a chromosome, with the outermost track indicating the length of each chromosome (20 Mb per tick). Inner tracks other than genome annotation showing parameters calculated over 300 kb windows. Moving inward from the outermost track, the plot shows: % GC content; uncallable annotation; annotation of tandem repeats longer than 30 kb; CpG methylation status (% of 5mC sites); density of mobile repeat annotations; density of gene annotations; annotation of *de novo* SMs, with points coloured according to SM type.

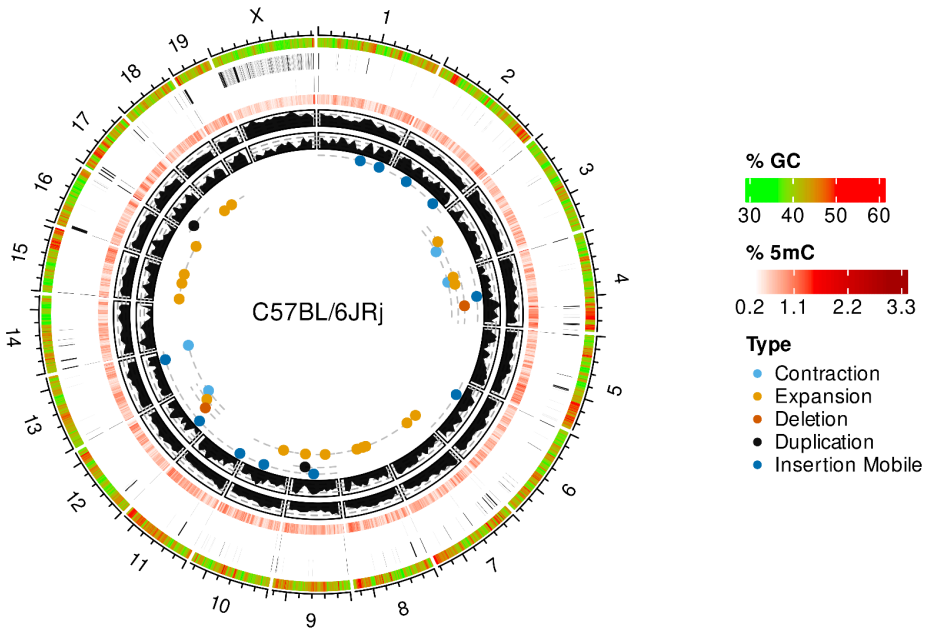

Figure S2. Circos plot (Krzywinski et al. 2009) of the C57BL/6JRj genome, generated using the R package circlize 0.4.15 (Gu et al. 2014). Each sector represents a chromosome, with the outermost track indicating the length of each chromosome (20 Mb per tick). Inner tracks other than genome annotation showing parameters calculated over 300 kb windows. Moving inward from the outermost track, the plot shows: % GC content; uncallable annotation; annotation of tandem repeats longer than 30 kb; CpG methylation status (% of 5mC sites); density of mobile repeat annotations; density of gene annotations; annotation of *de novo* SMs, with points coloured according to SM type.

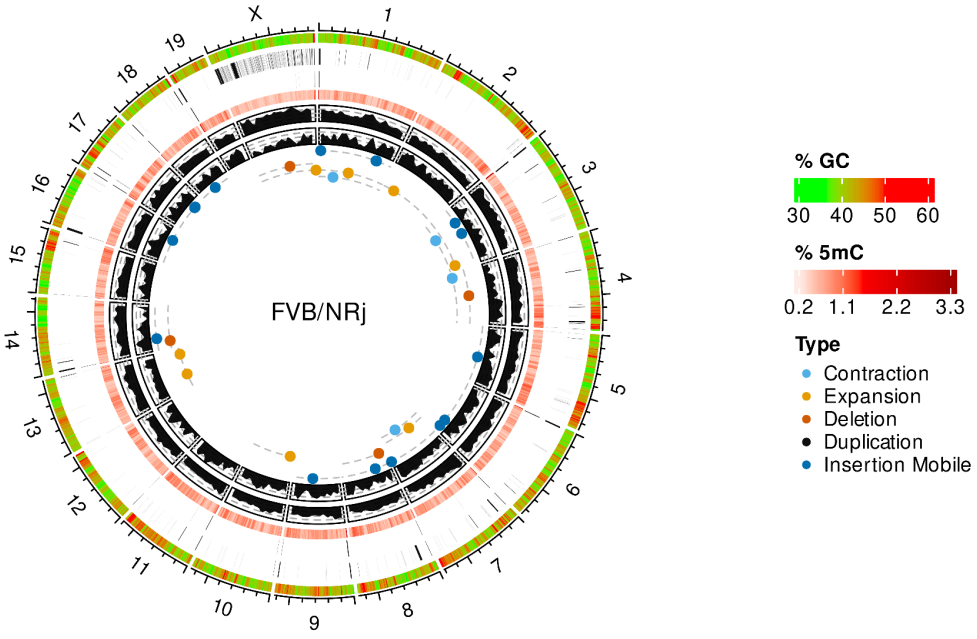

Figure S3. Circos plot (Krzywinski et al. 2009) of the FVB/NRj genome, generated using the R package circlize 0.4.15 (Gu et al. 2014). Each sector represents a chromosome, with the outermost track indicating the length of each chromosome (20 Mb per tick). Inner tracks other than genome annotation showing parameters calculated over 300 kb windows. Moving inward from the outermost track, the plot shows: % GC content; uncallable annotation; annotation of tandem repeats longer than 30 kb; CpG methylation status (% of 5mC sites); density of mobile repeat annotations; density of gene annotations; annotation of *de novo* SMs, with points coloured according to SM type.

Chromosome-level assemblies were validated by aligning each strain's assembly to its respective reference genome using nucmer 4.0.0 (Marçais et al. 2018) to generate whole-genome alignments. The C3H/HeNRj v1 assembly was aligned against the C3H/HeJ v3 reference (GenBank assembly accession GCA_921997125.2), the C57BL/6JRj v1 assembly against the GRCm39 reference (GCA_000001635.9), and the FVB/NRj v1 assembly against the FVB/NJ v3 reference (GCA_921998635.2). These alignments were visualised in dot plots using the genome alignment viewer 'dot' (https://dot.sandbox.bio/). Dot plots showed high collinearity for all the genome-genome alignments, with only the C3H strain showing a large structural change (Figure S4). We visualised assembly and read alignments using the Integrative Genomics Viewer (IGV, Robinson et al. 2011) to confirm that this structural variant, an inversion, was not due to an assembly error. In fact, this inversion in chromosome 6 had been previously identified as In(6)1J (Akeson et al. 2006).

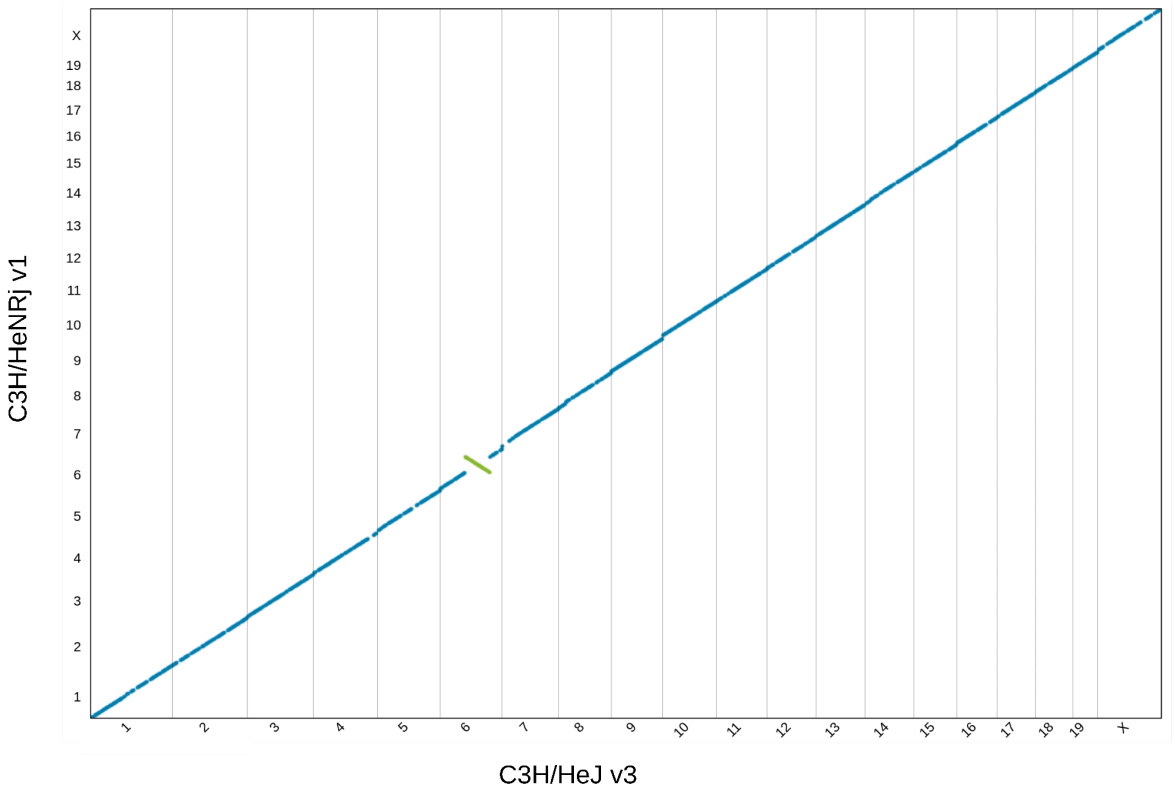

Figure S4. Dot plot showing the genome-genome alignment between the reference C3H/HeJ v3 and our *de novo* C3H/HeNRj v1 assembly. The alignment includes chromosomes identified by numbers on the label axis. Unique forward alignments are coloured in blue, unique reverse alignments (denoting chromosome inversions) are coloured in green, and repetitive alignments are coloured in orange.

The hifiasm assemblies successfully reconstructed full chromosomes from a few contigs. The ratio of contig to chromosome counts was 3.0 for C3H, 4.2 for BL6 and 3.7 for FVB. However, Chromosome Y was not assembled for any strain. Mapping the pooled founders' assembly against the GRCm39 reference genome resulted in a highly fragmented map of Chromosome Y, likely due to its high repetitiveness (Rhie et al. 2023). This issue persisted even when using the male founder assembly alone. Other published assemblies such as the C3H/HeJ and FVB/NJ reference genomes lack a Chromosome Y, so we could not use those for scaffolding.

The final C3H chromosome-level assembly was 2,720,445,931 bp long, in reasonable agreement with the length of the GRCm39 reference genome. In comparison, the BL6 (2,885,702,212 bp) and FVB (2,895,133,978 bp) assemblies were longer. This excess length was largely due to the presence of unmapped contigs, as the total length of chromosomal sequences was below 2.72 Gb for these two strains: 2.58 Gb for BL6 and 2.60 Gb for FVB. The number of unplaced contigs was 93 for C3H, 1,140 for BL6, and 783 for FVB. Unplaced contigs were named numerically based on their length (e.g., ctg_01 for the longest), with a suffix indicating a chromosome name if the assembly graph suggested unambiguous contiguity between the contig and the chromosome. However, unplaced contigs linked to the Chromosome Y were not named. The total number of mismatches, defined as the combined length of unknown nucleotides, was 40,111 bp in C3H, 2,059,354 bp in BL6, and 6,928,223 bp in FVB.

Callable sites

Only fully assembled chromosomes were considered for calling mutations. About 2.5 Gb of the genome was deemed 'callable' across strains, meaning it had sufficient quality for mutation calling. Relative to the total length of the chromosome-level genome, the proportions of callable sites were 95.0% for C3H, 95.9% for BL6, and 95.7% for FVB.

The non-callable fraction of the genome was largely attributed to the presence of repeat content, predominantly large satellite sequences longer than the PacBio reads. Combining all sources of repeat sequences annotated by RepeatMasker (https://www.repeatmasker.org) and Tandem Repeats Finder (Benson 1999), approximately 42% of the genome was repetitive. Among repeats, a small fraction (< 18%) were tandem repeats, mainly microsatellite and satellite sequences, while the majority (> 85.1%) were of transposable origin. The proportion of repeated content per chromosome was significantly and negatively correlated with the proportion of callable sites (t_{58} = -8.49, r = -0.74, $P = 9.32 \times 10^{-12}$). Tandem repeats longer than 30 kb were the main driver of uncallability in C3H (t_{18} = -5.04, r = -0.76, $P = 8.42 \times 10^{-5}$), and uncallable sites usually colocalized with clusters of large tandem repeats in the centromeres for all strains (Figures S1-S3; note that centromeres are acrocentric in mice). It should be noted that our callability criterion excluded tandem repeats larger than 30 kb (see Methods), but even after excluding this criterion, we observed the colocalization between uncallable sites and large tandem repeats due to frequent assembly gaps in these highly repeated regions. In Chromosome X, more scattered uncallable sites were observed than in other chromosomes, which we attribute to its high density of transposable elements (Figures S1-S3). Mobile repeat annotation was the main driver of uncallability for BL6 and FVB (t_{18} < -6.2, r = -0.82, $P = 8.3 \times 10^{-6}$), presumably because their lower coverage compared to C3H led to worse contiguity, particularly for Chromosome X (Table S2).

Table S2. Assembly callability and repeat content. The following proportions are shown for the average autosome and the Chromosome X: callable sites, all repeat sequences, repeats longer than 30 kb, tandem repeats, and mobile repeats.

Strain	Chromosome	Callable sites	All repeats	Large repeats	Tandem repeats	Mobile repeats
C3H/HeNRj (C3H)	Autosomes	95.0%	42.2%	2.61%	7.53%	35.7%
	Chromosome X	93.0%	55.9%	0.44%	3.69%	52.3%
C57BL/6JRj (BL6)	Autosomes	97.0%	40.9%	0.384%	5.40%	36.4%
	Chromosome X	75.3%	56.0%	0.33%	4.63%	52.5%
FVB/NRj (FVB)	Autosomes	97.1%	41.1%	0.67%	5.66%	36.3%
	Chromosome X	72.3 %	54.2%	0.30%	4.51%	50.7%

Appendix S2. Mutation rates

A total of 7,111 mutations were found in our study, including 2,981 single nucleotide mutations (SNMs, 41.9%), insertions and deletions shorter than 50 bp (indels, 56.2%), and structural mutations (SMs, 1.9%). These mutations are provided as tables in separate files: SNMs in Table S3, indels in Table S4, and SMs in Table S5. Their mutation rates per MA sample are provided in Table S6 below.

Table S6. Mutation rates per strain, sample, and mutation type. Rates are given in two scales: per haploid genome per generation (M), and per site per generation (μ). Mutation types included are single nucleotide mutations (SNMs), insertions and deletions shorter than 50 bp (indels), and structural mutations (SMs).

Strain	MA sample	M			μ ($\times 10^{-9}$)		
		SNM	indel	SM	SNM	indel	SM
C3H/HeNRj (C3H)	1	23.3	43.7	1.27	9.29	17.4	0.505
	2	24.6	45.6	0.89	9.82	18.2	0.353
	3	22.1	39.0	0.95	8.83	15.6	0.379
	4	22.8	46.3	0.76	9.11	18.5	0.303
C57BL/6JRj (BL6)	2	17.3	15.3	1.66	6.97	6.18	0.668
	3	18.5	17.5	1.35	7.47	7.06	0.543
	4	16.3	16.7	0.62	6.56	6.72	0.251
	5	14.8	14.1	0.41	5.97	5.68	0.167
FVB/NRj (FVB)	1	22.5	19.6	0.96	9.05	7.86	0.384
	2	18.9	13.8	0.78	7.58	5.56	0.314
	3	18.3	12.1	0.43	7.37	4.86	0.175
	4	15.6	8.6	0.69	6.29	3.46	0.280

Appendix S3. The spectrum of SNMs

Figure S5 shows the distribution of frequencies for different types of SNMs, i.e., the SNM spectrum. Overall, C→T mutations were approximately three times more frequent than expected, based on equal mutation rates for all types of single nucleotide change and a genomic GC content of 41.7%. FVB samples showed a substantial bias toward C→A transversions, approximately double the bias observed in other strains (2.3 vs. ~1). Using a large number of samples sequenced with Illumina short-reads, the impact of sequence context on these mutations was described in López-Cortegano et al. (2024). The SNM spectrum in Figure S5 includes information on changes in methylation for 5-methylcytosine (5mC) at CpG sites, which were visualised from PacBio read alignments in IGV.

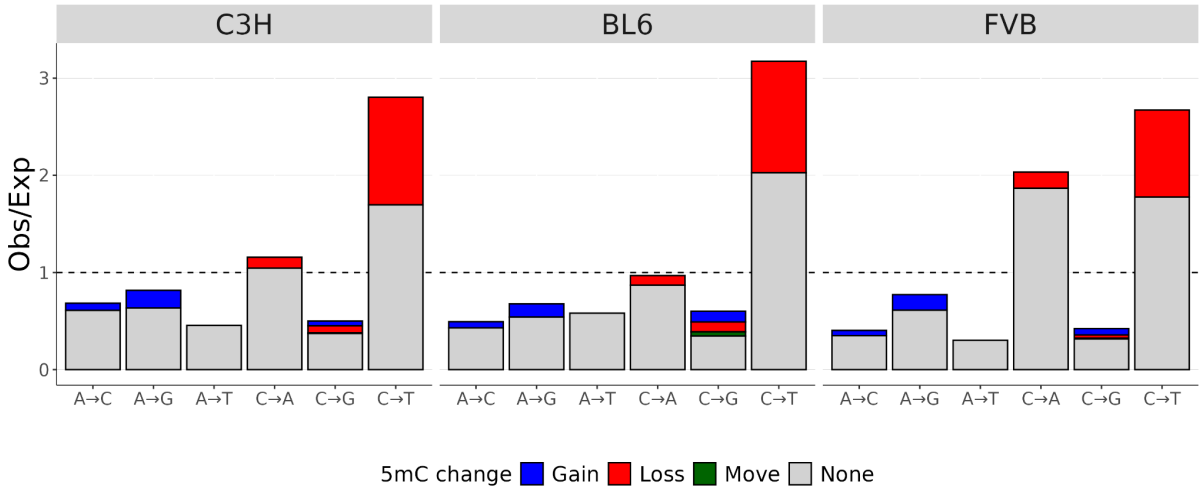

Figure S5. Spectra of SNMs across strains. Each type of SNM is represented with the reference site being either an adenine (A) or a cytosine (C). The y-axis shows the ratio of observed to expected SNM changes for each SNM type. Expectations are calculated based on a genomic GC content of 41.7%. Bars are coloured according to their associated changes in 5-methylcytosine (5mC) marks at CpG sites: blue for 5mC gain, red for 5mC loss, green for change in 5mC position without net gain or loss, and grey for no change.

To predict 5mC sites, HiFi reads generated by the PacBio ccs tool (see Methods) were further processed with dedicated tools from the PacBio official GitHub repository (https://github.com/PacificBiosciences/). C3H samples were processed with primrose 1.3.0, and BL6 and FVB samples were processed with jasmine 2.0.0. From the aligned HiFi read files, 5mC sites were called with the tool "aligned_bam_to_cpg_scores from" v.2.3.1 from PacBio (https://github.com/PacificBiosciences/pb-CpG-tools), using the default recommended built-in model-based approach to score 5mC sites ("-p model").

Appendix S4. Sequence complexity and mutation rate in *Chlamydomonas*

To further investigate the association between DNA sequence repetitiveness and the rate of new SNMs and INDELs on a broad evolutionary scale, we analysed data from the green algae *Chlamydomonas reinhardtii* (López-Cortegano et al. 2023) similarly as was done for Figure 2 of the main text (Figure S6). For this analysis, only data from MA samples of the *C. reinhardtii* CC-2931 strain that were sequenced with PacBio HiFi technology were considered. As in Figure 2, a 10-fold linear increase in mutation rate was observed for SNMs (Linear regression, $R^2 = 0.8$, $F_{1,42} = 173.4$, $P < 2.2 \times 10^{-16}$). For indels, this increase was asymptotic and higher in magnitude, from $\mu \approx 3.2 \times 10^{-10}$ (i.e., $\mu \approx 10^{-9.5}$ in Figure S6) in genomic regions with average repetitiveness to $\mu \approx 3.2 \times 10^{-8}$ (i.e., $\mu \approx 10^{-7.5}$) in the most repetitive sequences. Additionally, the data showed a lower rate for insertions compared to deletions in low repetitive sequences with near-zero scaled linguistic complexity, though the rate increased more steeply for insertions. The resemblance of the pattern observed in mice and *Chlamydomonas* suggests that the relationship between the rate of indels and sequence complexity be broadly present across the tree of life.

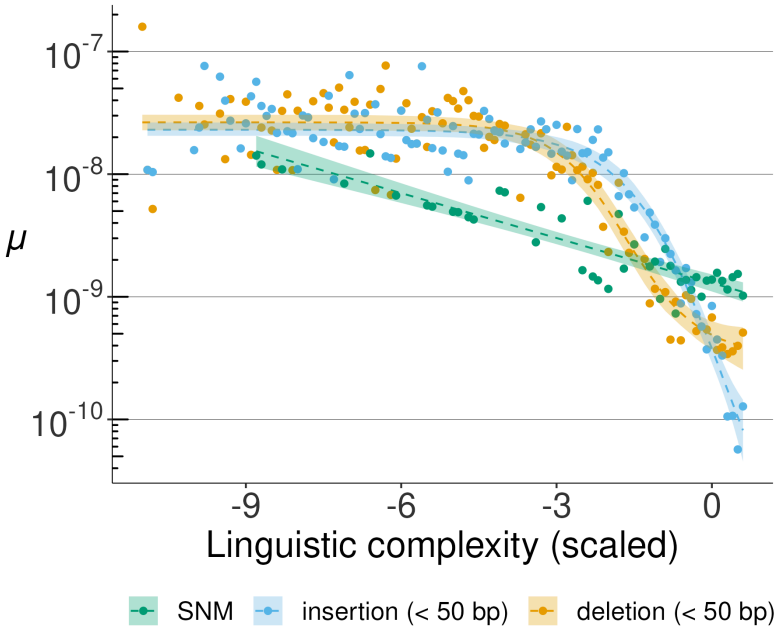

Figure S6. Association of SNM and indel rates (in $\log_{10}$ scale) with the repetitiveness of the sequence in *C. reinhardtii* CC-2931. Sequence repetitiveness was measured as the linguistic complexity of 101 bp genomic windows, and then scaled to standard deviation units with a mean of zero. The observed rates of different types of mutations (in colours) are shown as data points. Dashed lines represent regression model fits for the change in mutation rate with linguistic complexity. For SNM data, a linear model was fitted, while indel data best fit a logistic regression model. 95% Confidence intervals are presented for all regression models.

Appendix S5. The Spectrum of indels

As described in the main text, most indels occurred in microsatellite sequences. Figure S7A shows the spectra of indels across mouse strains for indels located in regions with and without microsatellite annotation. For microsatellite regions, the spectra were markedly biassed towards 4 bp contractions and expansions of microsatellites with 4 bp motifs. Figure S7B shows the minimum number of microsatellite motifs that together explain at least 50% of the observed indels. Among these motifs, 70% correspond to different annotations of the microsatellite motif "AGAT". For example, the repeated motif "AGAT" is equivalent to "ATAG", and is also equivalent to the reverse complement "TATC" (Figure S7B). Most indels were indeed contractions and expansions of one or several microsatellite motifs (Figure S7C). The different patterns observed for deletions and insertions in Figure S7C could be due to deletions generally being longer than insertions (Figure S7A).

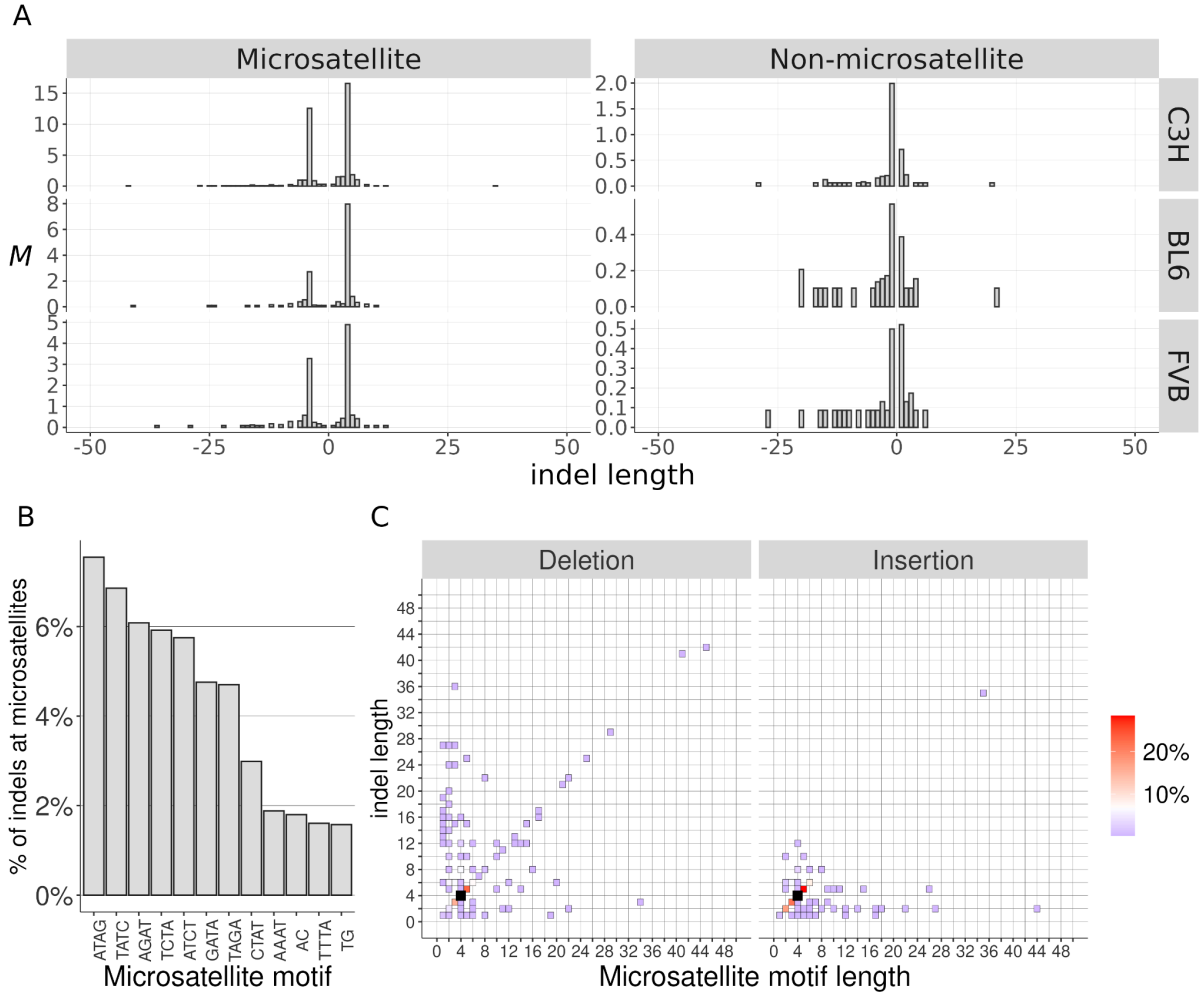

Figure S7. Spectrum of new indels. A) The distribution of indels by rate (*M*), length (in bp), and strain. Indels are categorised by their genomic location within or outside microsatellites. B) Proportion of indels in microsatellites that occurred in specific microsatellite motifs. Only the 12 motifs with the highest proportions are shown. C) Distribution of indels in microsatellites by their length and the length of their respective microsatellite motif. Only indels and microsatellite motifs up to 50 bp are shown. Black dows represent 4 bp indels in 4 bp microsatellites, while other colours indicate the proportion of insertions and deletions.

Appendix S6. Massive repeat expansion

In MA sample BL6 2, a large expansion event of an approximately 3 kb "TCTTCT" satellite sequence on Chromosome 12 was identified, with genomic coordinates 79439755-79442748 in the BL6 reference. The expansion event was evident from long-read alignments (Figure S8), but was collapsed in the MA sample assembly due to its large size. Based on the increase in coverage observed in read alignments, we estimated this expansion to be longer than 100 kb. The coverage measured in IGV for this expansion was 287×, from which we subtracted the average diploid coverage of 15× for this sample (Table S1), resulting in an expanded coverage of 272×. Given that the expansion was heterozygous, we calculated an approximately 36-fold increase in copy number for this satellite, corresponding to an expansion of 108 kb. Analysing soft-clipped single reads, we confirmed at least 20 kb and 17 kb of expansion 5' and 3' to the satellite, respectively. The longest read containing satellite sequence only was 24 kb in length.

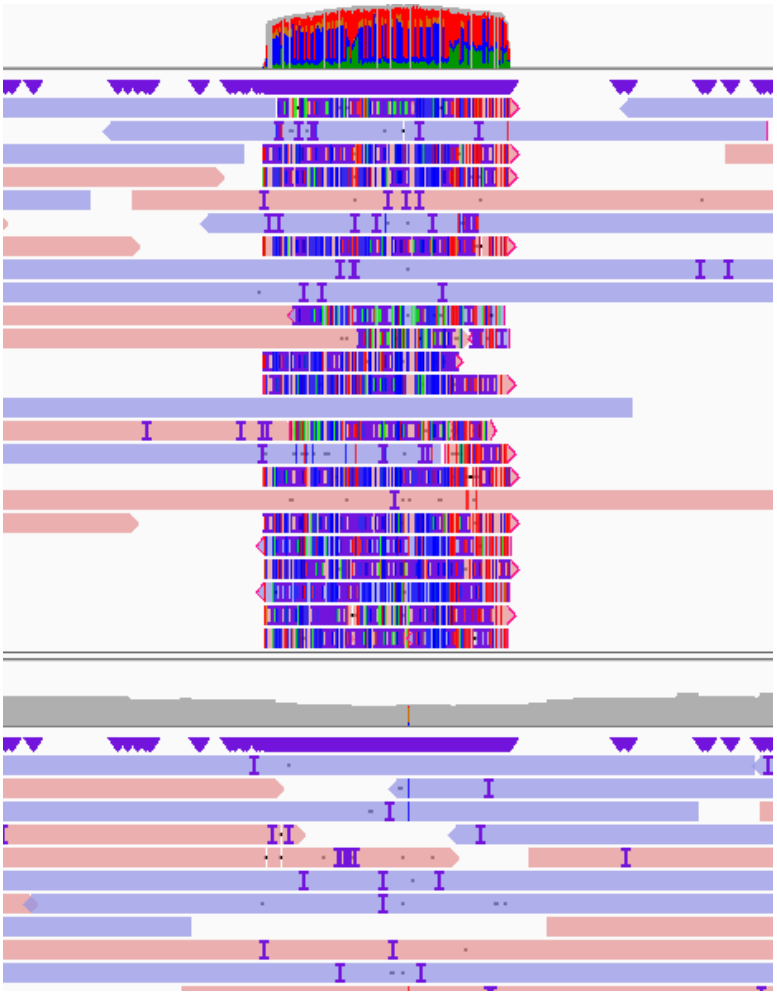

Figure S8. IGV visualisation of DNA satellite expansion in MA sample BL6 2. The top panel shows the read alignment for the MA sample, showing the expansion (note the increase in coverage, measured at 287×). The bottom panel shows the alignment of the female founder, without the mutation.

Appendix S7. Homology-mediated structural mutations

Several deletions and duplications in our data set could be explained by known homology-mediated DNA repair mechanisms. Detailed examples are provided below.

Three deletions were attributed to nonallelic homologous recombination events involving sequence homology longer than 20 bp. Two of these were associated with transposable element (TE) sequences at the breakpoints (Figure S9A), while the third involved the deletion of a previously duplicated sequence (Figure S9B), which had LINE-1 annotation. Six deletions were putatively involved in microhomology-mediated end-joining (MMEJ). These deletions showed homology shorter than 10 bp and, in some cases, were clustered with other mutations, possibly due to the high mutagenicity of MMEJ (Sinha et al. 2017). An example involving two clustered deletions is shown in Figure S9C. Two other deletions attributed to MMEJ were clustered with a duplication and two SNMs, respectively.

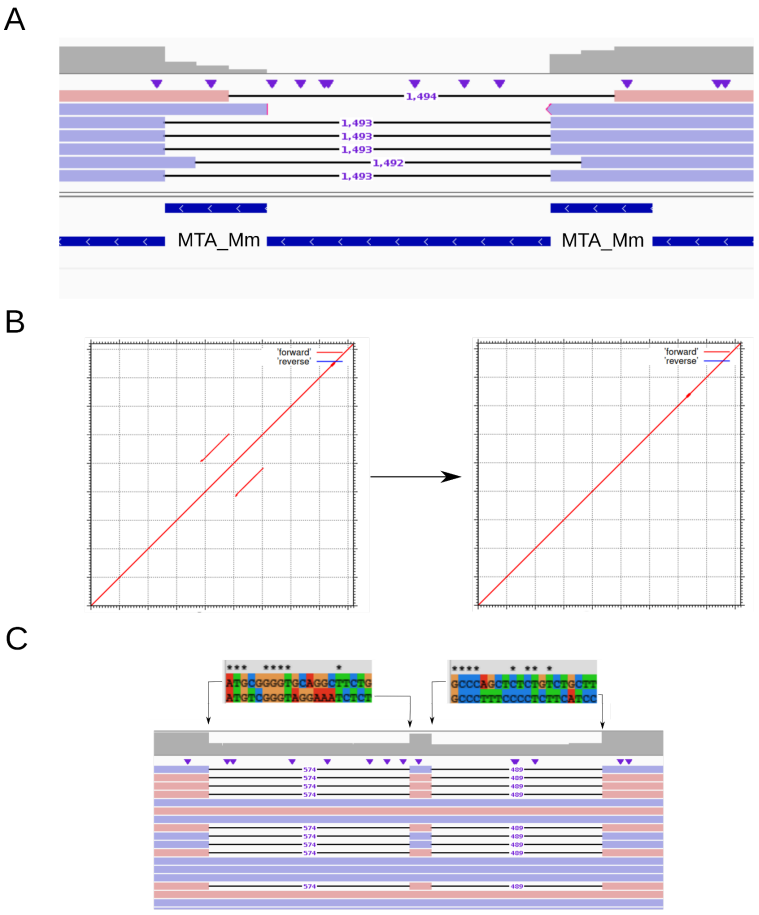

Figure S9. Examples of homology-mediated deletions. A) IGV visualisation of a 1,493 bp deletion in MA sample FVB 4, involving the TE annotation "MTA_Mm". B) Self-to-self dotplot generated with MAFFT of the region including the 596 bp deletion in C3H MA sample 3 (left: the founders; right: the MA sample). Axis lines are in units of 500 bp. C) Alignment and visualisation of sequences putatively involved in MMEJ with Clustal X (top) and IGV (bottom). For each ClustalX alignment, the top sequence corresponds to the starting sequence of the deletion and the bottom to the sequence downstream of the deletion. Only 20 bp of alignment are shown.

The three duplications longer than 150 bp in our data set could be attributed to homology-mediated mechanisms. For example, one 434 bp duplication was likely mediated by homology between sequences near the site of the mutation (Figure S10A). Indeed, the first 100 bp of the duplication perfectly aligned with the 100 bp following the duplication (Figure S10B). Another example was the tandem duplication of an intracisternal A-particle, likely mediated by the homology of the two flanking long terminal repeats (LTRs) of the transposon (Figure S10C). This duplication event resulted in a new 6,762 bp tandem copy of the transposon, deficient in one of the LTRs.

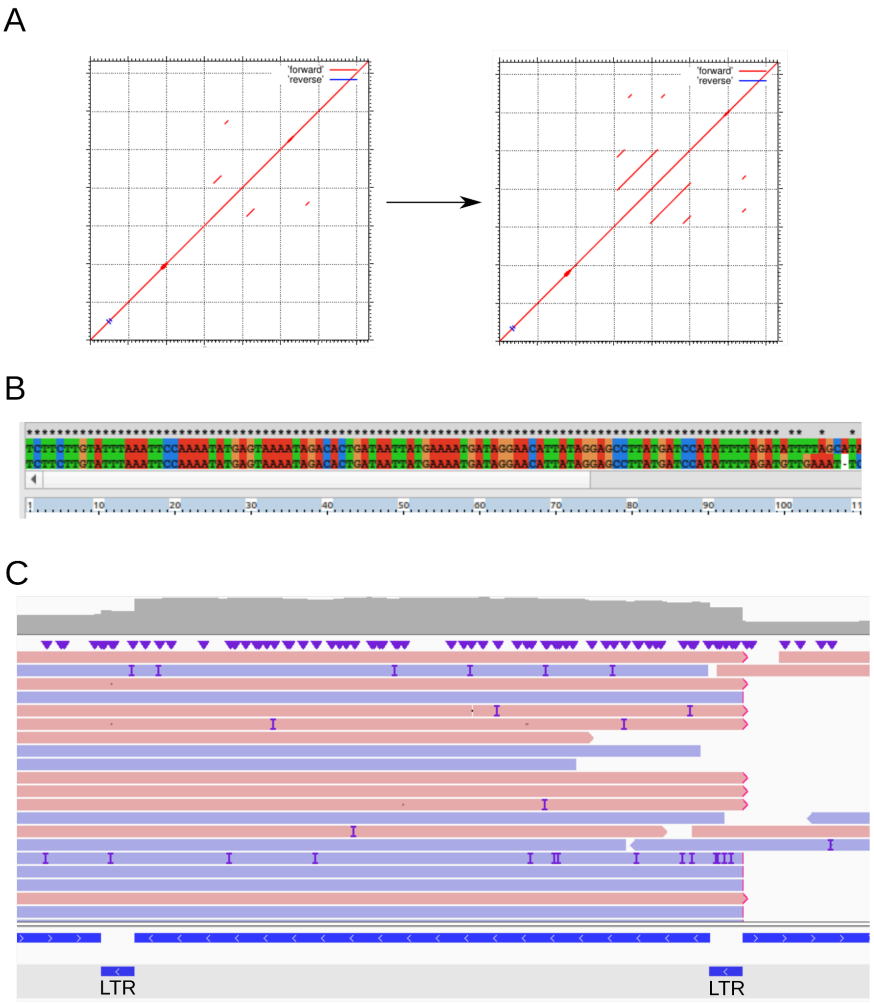

Figure S10. Examples of homology-mediated duplications. A) Self-to-self dotplot generated with MAFFT of the region including the 434 bp duplication in C3H MA sample 1. On the left, the ancestral sequence in the founders' genome. On the right, the derived sequence after duplication in the MA sample genome assembly. Axis lines are in units of 500 bp. B) Alignment and visualisation with Clustal X (Larkin et al. 2017) of the 434 bp duplicated sequence in C3H MA sample 1, and the 434 bp sequence immediately downstream. Only the first 110 bp of the alignment are shown. C) IGV visualisation of a 6,762 bp duplication in MA sample BL6 2, involving a transposable element sequence. On the bottom, two blue tracts are annotated with "LTR" indicating the long terminal repeats of the transposon, which has annotation IAPEz-int for the tract between the two LTRs (LTRs were named "IAPLTR4_I" in the annotation generated by RepeatMasker).

Appendix S8. Mother copies of IAP insertions

Table S7. Insertions of intracisternal A-particles in C3H MA samples, their insertion length, and their best match to the C3H founder genome using the megablast algorithm (Camacho et al. 2009). Only the matches with the highest BLAST score are shown (more than one if multiple matches share the highest score). All results had an Expected (E) value of zero.

IAP insertion (coordinates)	MA Sample	Length (bp)	Match (coordinates)	Identities	Gaps
1:41,288,714	1	5,307	19:26,477,497-26,472,191 14:28,769,477-28,764,171 9:83,065,308-83,070,614 7:87,944,619-87,949,925 4:131,573,235-131,578,541	5296/5307 (99%)	0/5307 (0%)
12:98,353,195		5,485	18:54,970,306-54,964,822	5481/5485 (99%)	0/5485 (0%)
17:115,086,796		5,305	11:100,672,809-100,667,507	5295/5305 (99%)	2/5305 (0%)
7:115,086,796	2	5,273	3:144,732,012-144,726,740	5268/5273 (99%)	0/5273 (0%)
10:108,583,464		5,306	9:69,798,708-69,804,013 5:113,186,412-113,191,717	5306/5306 (100%)	0/5306 (0%)
12:118,272,492		5,311	16:20,446,358-20,441,048	5311/5311 (100%)	0/5311 (0%)
1:89,621,682	3	5,400	16:89,015,380-89,009,981	5395/5400 (99%)	0/5400 (0%)
1:110,513,778		5,281	13:52,071,482-52,076,744	5192/5287 (98%)	30/5287 (1%)
12:111,230,905		5,277	3:140,776,867-140,782,143	5263/5277 (99%)	0/5277 (0%)
16:15,557,201		5,305	8:45,713,279-45,707,975	5301/5305 (99%)	0/5305 (0%)
11:57,382,641	4	5,331	X:86,579,668-86,574,338	5314/5331 (99%)	0/5331 (0%)
11:101,822,282		5,460	18:54,964,822-54,970,305	5438/5490 (99%)	36/5490 (1%)
14:84,502,838		5,306	9:69,804,014-69,798,709 5:113,191,718-113,186,413	5306/5306 (100%)	0/5306 (0%)

Appendix S9. Germline expression of retrocopied genes

Gene retrocopies are RNA-mediated gene duplications originating from the mRNA of genes undergoing expression that are captured by the retrotransposition machinery of LINE retrotransposons. To further confirm that the retrocopies observed in our MA experiment belong to genes expressed in the mouse germ line, we used the EBI Gene Expression Atlas (Moreno et al. 2021, accessed on 1st February 2024, with version numbers Ensembl 104, Ensembl Genomes 51, WormBase ParaSite 15, and EFO 3.10.0). The Ensembl identifiers for the *Amd1* (ENSMUSG00000075232), *Brd2* (ENSMUSG00000024335), *Rpl12* (ENSMUSG00000038900), and *Smn1* (ENSMUSG00000021645) genes were queried against the *Mus musculus* database, and results were filtered to retain information only for the expression in testis and ovaries from transcriptomic studies. As shown in Figure S11, not only was expression typically higher in testis than in ovaries, but also more studies detected expression in testis.

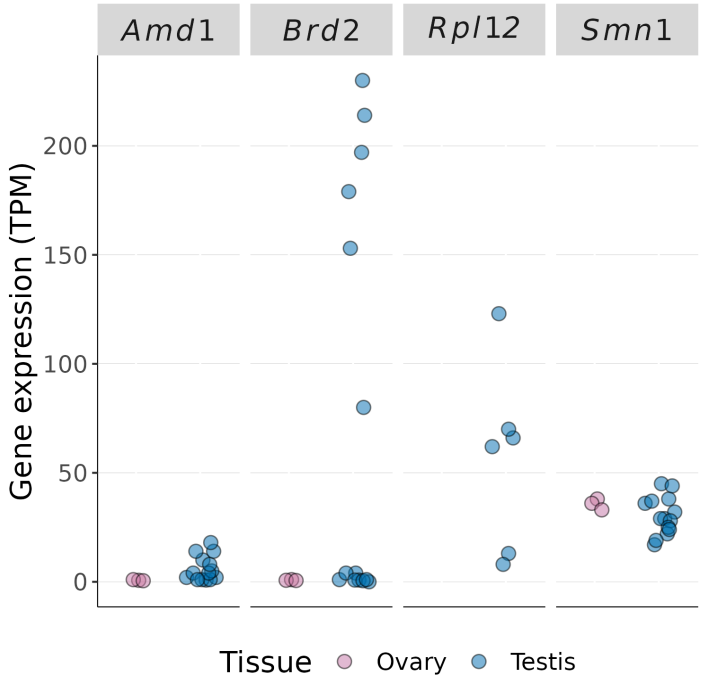

Figure S11. Expression of *Amd1*, *Brd2*, *Rpl12*, and *Smn1* genes in ovaries and testis, measured in units of transcripts per million (TPM), as collected from the EBI Gene Expression Atlas (GXA). Each data point corresponds to one scientific article referenced to by the GXA. Only transcriptomic studies were considered.

Appendix S10. Landscapes of TEs

As shown in the main text, the C3H, BL6 and FVM MA lines differed in their spectra of active TEs (Figure 4). To investigate whether these differences could be attributed to differences in the proportions of near complete, functional TE sequences, we used the scripts "calcDivergenceFromAlign.pl" and "createRepeatLandscape.pl" from the RepeatMasker suite (https://www.repeatmasker.org) to generate "TE landscapes". Such landscapes plot the percentage of the genome annotated with different types of TEs, and in addition, the divergence rates TEs from their respective consensus sequences. Lower divergence rates (measured with the Kimura substitution level) suggest recent TE insertions, whereas higher divergence rates suggest older TEs that have accumulated mutations over time.

The TE landscapes of the three strains studied here were remarkably similar to each other (Figure S12) and to the TE landscape of the mm10 (i.e., GRCm38) mouse reference genome (https://www.repeatmasker.org/species/mm.html). All these landscapes have in common a recent invasion of L1 (LINE) elements and to a lesser extent from endogenous retroviruses (ERV) or LTR retrotransposons. The proportions of L1 elements with near-zero divergence from the consensus were 1.86%, 1.78% and 1.79% for C3H, BL6 and FVB, respectively. For LTR retrotransposons, these proportions were 0.28%, 0.28% and 0.29% for C3H, BL6 and FVB, respectively. Overall, these results suggest that the differences observed in the spectra of active TE among strains are not due to differences in the abundance of these TEs in the genome or to differences in their evolutionary ages.

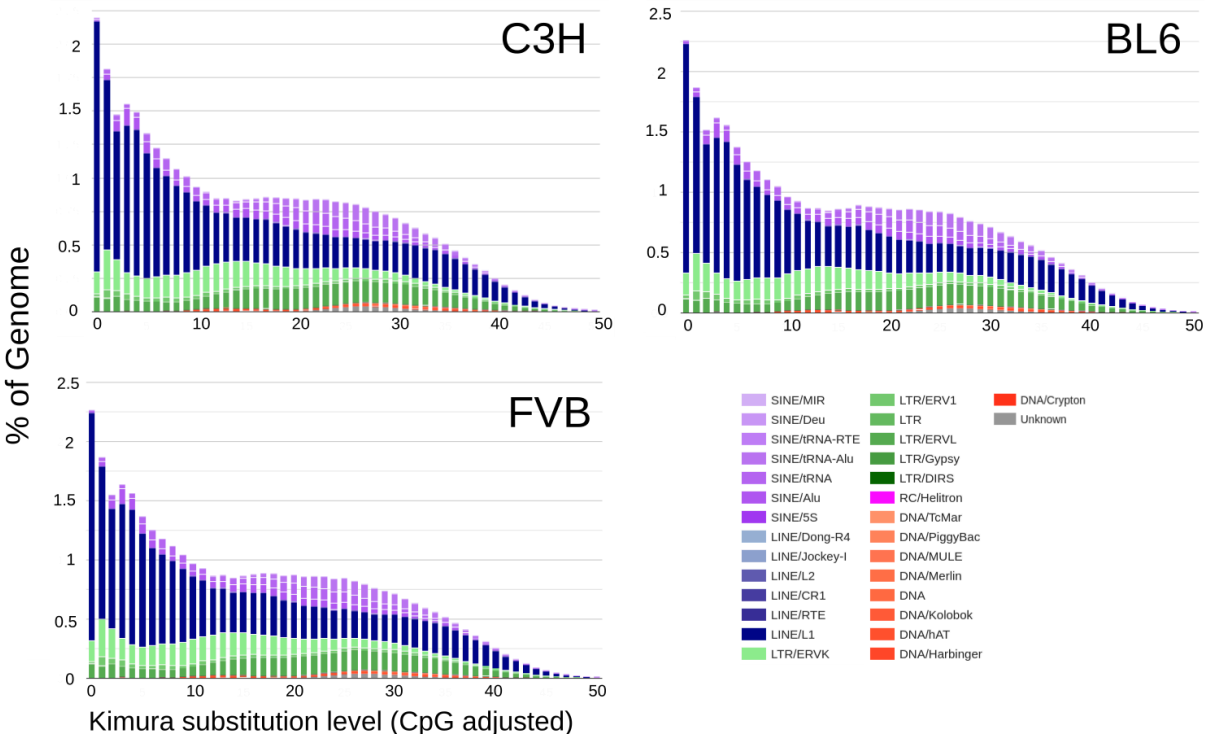

Figure S12. TE landscapes from RepeatMasker for the C3H, BL6 and FVB genomes. TE families are shown in colours, with LINE (blue) and LTR retrotransposons (green) as the most abundant. The proportion of the genome represented by each type of TE is shown in the y-axis, while the evolutionary age of each element, measured with the Kimura substitution level, is shown in the x-axis.

Appendix S11. KRAB-ZFP clusters

Krüppel-associated box zinc finger proteins (KRAB-ZFPs) have previously been described to bind to TE sequences and suppress them (Wolf et al. 2020). Therefore, it is possible that variation in KRAB-ZFPs among strains is responsible for the differences in the spectra of active TEs observed (Figure 4), especially given that the clusters encoding these KRAB zinc finger genes are highly copy number variable in natural populations of mice (Pezer et al. 2015 - see also CNV annotation tracks in the public session "wildmouse" at the UCSC genome browser (Harr et al. 2016)), which are the ultimate source of the inbred strains. To explore this possibility, we used data from three major KRAB-ZFP clusters described in laboratory mice by Wolf et al. (2020), and searched for genes with sequence similarity in the C3H, Bl6 and FVB genomes by using the software platform GENEIOUS (https://www.geneious.com/; Kearse et al. 2012) with a cutoff of 60% sequence similarity. The results (Figure S13) suggest substantial variation in the composition of these clusters, including both changes in their length and composition. While our data does not allow us to confirm this hypothesis, variation in genes responsible for TE repression such as those encoding KRAB-ZFPs offer a more plausible explanation for the differences in rate and spectra of TEs than variation in TE sequences (Figure S12).

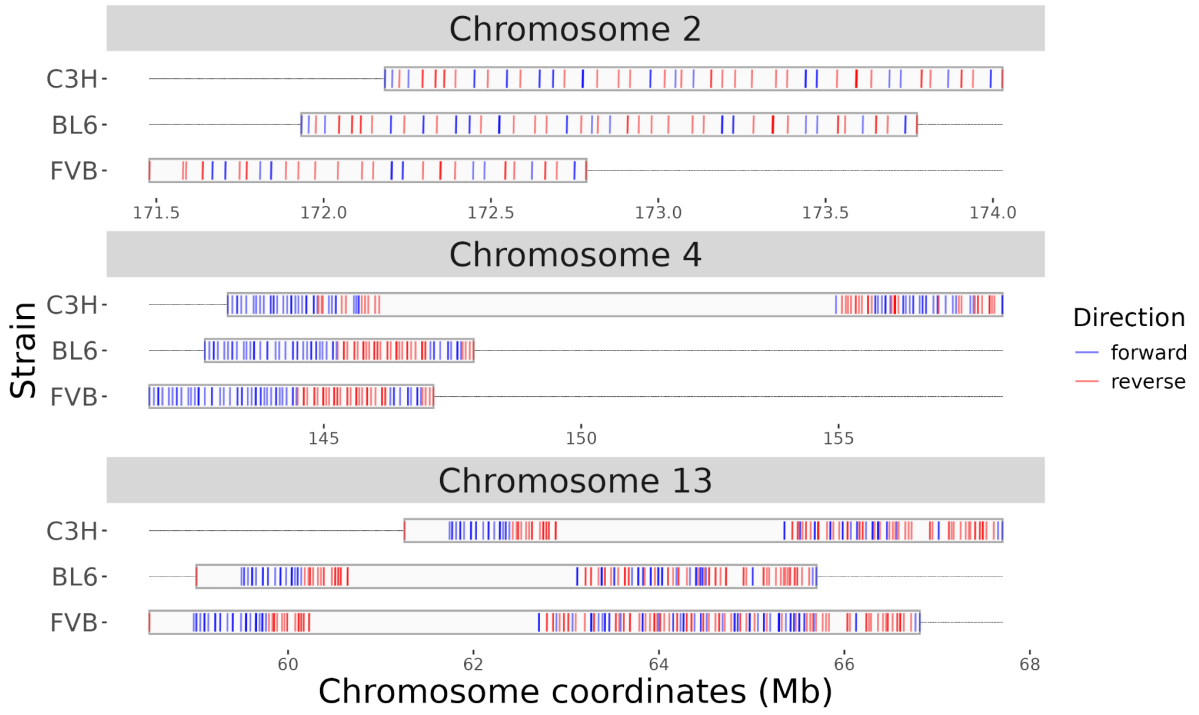

Figure S13. Copy number variation of genes encoding KRAB-ZFPs among strains. Individual genes shown in forward (blue) and reverse (red) orientation at known clusters in Chromosomes 2, 4, and 13. (Wolf et al. 2020).

Appendix S12. Examples of false positive calls

All accepted mutations were manually inspected using IGV. Occasionally, SNM and indel calls were rejected as false positives, while a substantial proportion of SM calls were rejected. Some representative examples are given below.

The most frequent reasons for removing SNM and indel calls included: 1) sites with coverage exceeding 60 reads, suspected to represent paralogous sites, 2) the presence of more than four reads supporting the alternate allele for the two founders, 3) the presence of more than four reads supporting the alternate allele in MA samples other than the one with the mutation call, and 4) lack of support in both the read and assembly alignments. Specifically, in the case of indels called at microsatellite annotation, it was notoriously difficult to determine whether a mutation was unique or not to a MA sample, since partial support for the mutation could be observed in other samples (e.g., Figure S14A). Of course, it is possible that highly mutable microsatellites accumulated mutations independently in more than one sample in our experiment. As a consequence, mutation rates at repetitive regions could be underestimated in our study, as noted in the main text.

The same causes for rejecting SNM and indel calls were repeated for SMs, and most frequently SM calls were rejected when called at highly repetitive regions. Rejected SMs were typically called as copy number variants, and were concentrated in hotspots within large tandem repeat sequences (e.g., Figure S14B). In such hotspots, it was never possible to accurately determine whether an individual call was genuine and unique to a single MA sample. The boundaries of assembly gaps and callable sites were also hotspots for false positive calls (e.g., Figure S14C), again due to the involvement of large tandem repeat sequences. Similar conclusions were reached in a previous study on *de novo* SMs in the green algae *Chlamydomonas* (López-Cortegano et al. 2023). Our results support the need for extensive manual work curating mutation calls at highly repetitive regions with the current state of technology, and highlight the need for better and improved methods of read alignment and genomic analysis of repetitive sequences.

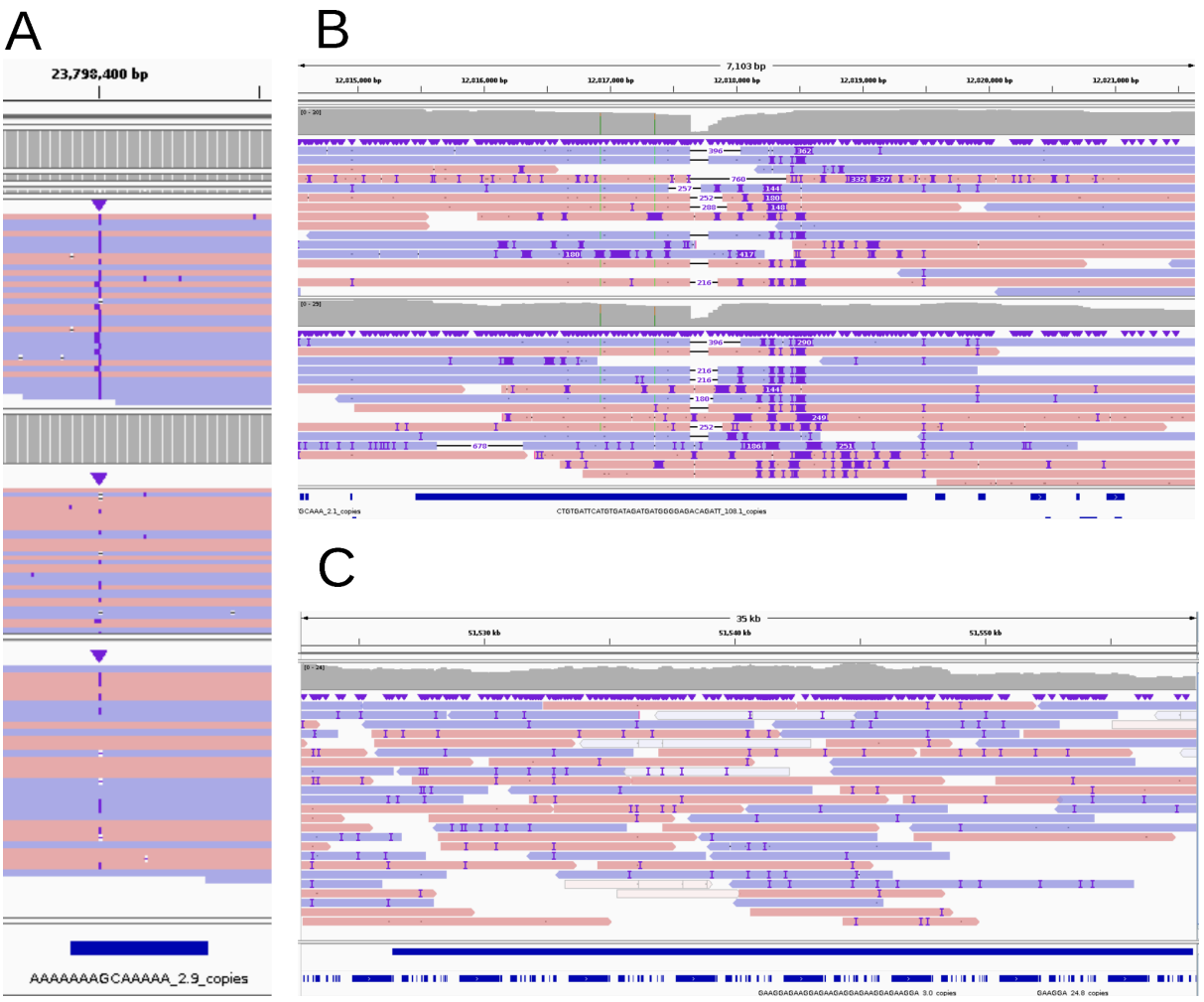

Figure S14. Examples of false positive variant calls visualised in IGV. A) C →CA insertion C3H MA sample 4 (chromosome_03:23798400). While the homozygous mutation call looks genuine in MA sample 4 (top track), it reached the threshold to be accepted as heterozygous mutation in other samples, including the founders (middle and bottom tracks), failing our criterion to be accepted as a new mutation. B) Hotspot of SM calls rejected as false positives. Distinct SM were called for different samples and coordinates, but none of these calls could be clearly confirmed from the reads, and the overall pattern of variation in the reads does not support unique mutation events in single MA samples but rather a problem derived from the read alignments or the genome assembly. In the bottom of the figure, a blue track identifies a large tandem repeat annotation. C) 54 bp insertion called in C3H MA sample 3 (chromosome_06:51540540). The mutation call cannot be confirmed from the reads. It is noticeable the presence of reads with low mapping quality (in white). In the bottom of the figure, the top track in blue represents callable sites, showing the proximity to an uncallable region approximately 15 kb upstream from the coordinates where the mutation was called. The bottom blue track represents tandem repeat annotation, which noticeably is itself tandemly repeated.

References

Akeson EC, Donahue LR, Beamer WG, Shultz KL, Ackert-Bicknell C, et al. 2006. Chromosomal inversion discovered in C3H/HeJ mice. Genomics 87(2): 311-313.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27(2): 573–580.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods 18(2):170-175.

Cheng H, Jarvis ED, Fedrigo O, Koepfli KP, Urban L, et al. 2022. Haplotype-resolved assembly of diploid genomes without parental data. Nat. Biotechnol. 40: 1332–1335.

Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. Circlize implements and enhances circular visualization in R. Bioinformatics 30(19): 2811-2812.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29(8): 1072-1075.

Harr B, Karakoc E, Neme R, Teschke M, Pfeifle C, et al. 2016. Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. Scientific Data 3: 160075.

Jain C, Koren S, Dilthey A, Phillippy AM, Aluru S. 2018. A fast adaptive algorithm for computing whole-genome homology maps. Bioinformatics 34(17): i748-i756.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, et al. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28(12): 1647-1649.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res 19(9): 1639-1645.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. 2017. Clustal W and Clustal X version 2.0. Bioinformatics 23(1): 2947-2948.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34(18): 3094–3100.

López-Cortegano E, Craig RJ, Chebib J, Balogun EJ, Keightley PD. 2023. Rates and spectra of de novo structural mutation in Chlamydomonas reinhardtii. Genome Res. 33(1): 45-60.

López-Cortegano E, Chebib J, Jonas A, Vock A, Künzel S, et al. 2024. Variation in the spectrum of new mutations among inbred strains of mice. Mol. Biol. Evol. doi: 10.1093/molbev/msae163.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol. Biol. Evol. 38(10): 4647-4654.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, et al. 2018. MUMmer 4: a fast and versatile genome alignment system. PLoS Computat. Biol. 14(1): e1005944.

Moreno P, Fexova S, George N, Manning JR, Miao Z, et al. 2021. Expression Atlas update: gene and protein expression in multiple species. Nucl. Acids. Res. 50(D1): D129-D140.

Pezer Ž, Harr B, Teschke M, Babiker H, Tautz D. 2015. Divergence patterns of genic copy number variation in natural populations of the house mouse (Mus musculus domesticus) reveal three conserved genes with major population-specific expansions. Genome Res. 25(8): 1114-1124.

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genom. Biol. 21: 245.

Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, et al. 2023. The complete sequence of a human Y chromosome. Nature 621(7978): 344-354.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, et al. 2011. Integrative genomics viewer. Nat Biotechnol 29(1): 24–26.

Sinha S, Li F, Villarreal D, Shim JH, Yoon S, et al. 2017. Microhomology-mediated end joining induces hypermutagenesis at breakpoint junctions. PLoS Genet. 13(4): e1006714.

Wolf G, de Laco A, Sun MA, Bruno M, Tinkham M, et al. 2020. KRAB-zinc finger protein gene expansion in response to active retrotransposons in the murine lineage. eLife 9: e56337.