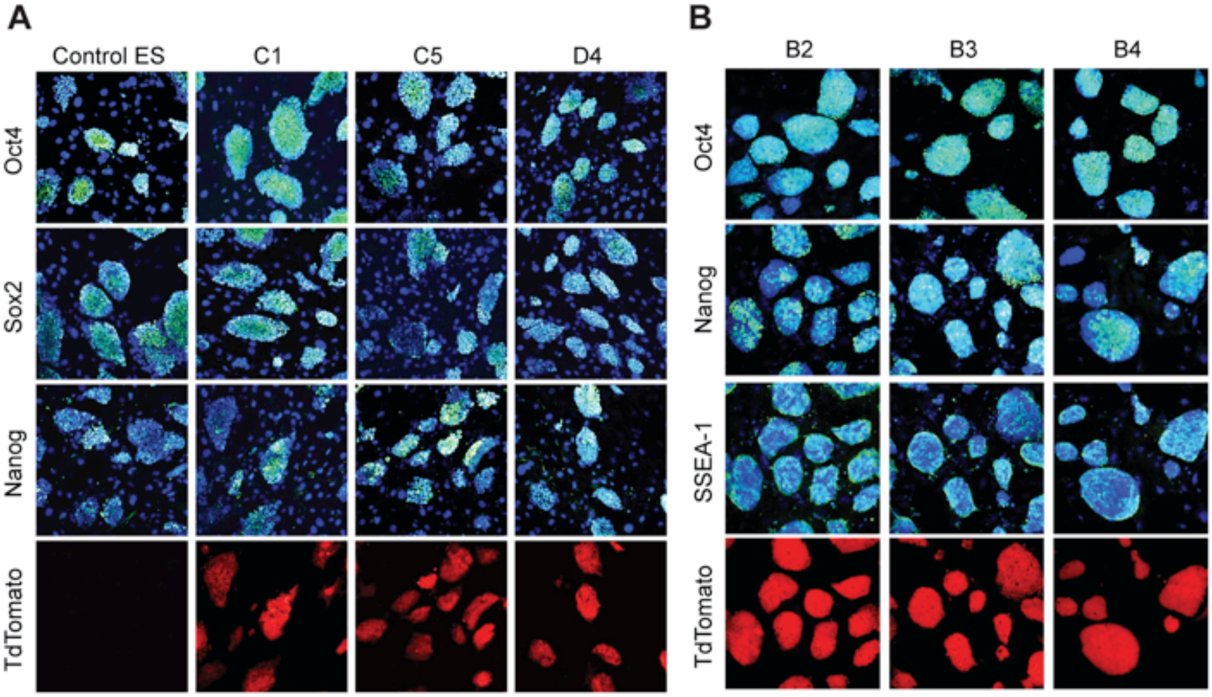
SUPPLEMENTAL DATA

SUPPLEMENTAL FIGURES

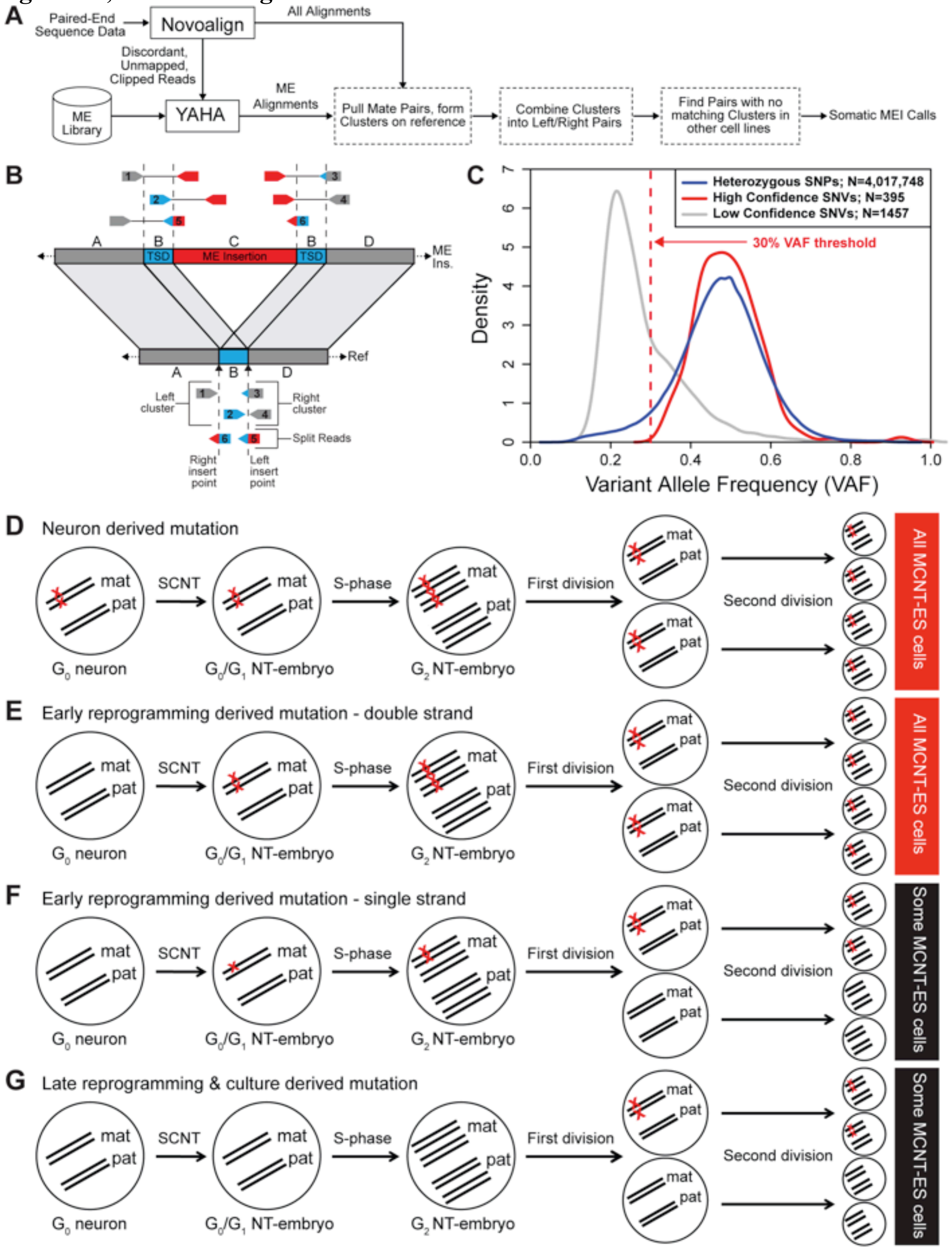
Figure S1, related to Figure 2.



Characterization of ES cells derived from MT neurons.

(A) and (B) MCNT-ES cells display endogenous tdTomato fluorescence (red) and express multiple proteins characteristic of pluripotent stem cells (Oct4, Sox2, Nanog, and SSEA-1; green). Nuclei are stained with DAPI (blue).

Figure S2, related to Figure 2.



Mutation detection and models for developmental timing of mutations. (A) Flowchart depicting processing steps in MEI detection pipeline. (B) Schematic depiction of the structure of a ME insertion event. The ends of paired-end reads that fall within the ME insertion (red) are difficult to map to the reference genome. Therefore, all discordant, unmapped and clipped reads are first aligned to a ME library. The mates of reads that map well to the ME library (1, 2, 3 and

4) are clustered by their reference coordinates. Left/right clusters that form properly oriented pairs define a possible MEI event. Further supporting evidence for the call is gathered from split-reads in which one end of the read maps well to the reference adjoining an insertion point, while the other maps well to the ME library, thereby spanning an insertion breakpoint (5 and 6). In addition, we determine if a target site duplication (TSD) has occurred by checking if the right insertion point falls before the left insertion point on the reference. Such a TSD is further confirming evidence for a MEI event. See Materials and Methods and Table S2 for details. Diagram adapted from a previously published work (Lee et al., 2012). (C) The distribution of Variant Allele Frequency (VAF), defined as the number of reads containing the alternate allele divided by total read depth, for three different categories of single nucleotide mutations. Note that, as expected, heterozygous autosomal SNPs have a VAF distribution that is roughly normal with a mean of 50%. This distribution is very closely matched by the high confidence (HC) SNVs (as defined by GATK) that have an estimated FDR of 0% based on our PCR validation experiments. In contrast, low confidence (LC) SNVs have a much lower mean VAF and the distribution is heavily skewed to the left. This is an indication of possible mutations that arose during clonal expansion, or other contamination, and not from the original neuron used during SCNT. The vertical line at 30% VAF demarcates the threshold we applied to putative SNVs above which they were considered candidate neuronal somatic mutations. This threshold is just over two standard deviations from the SNP and HC SNV 50% mean, and as can be seen from the graph eliminates almost no HC calls, but most of the LC calls. (D-G) Sources of mutations in MCNT-ES cell genomes and their predicted prevalence in MCNT-ES cell populations. (D) Nearly all neuron-derived mutations should be heterozygous and therefore present on only one of two homologous chromosomes (maternally derived, mat, paternally derived, pat). As a result, neuron derived mutations have an expected VAF of ~50% and appear in all subclones generated from that MCNT-ES cell line (Figure 2E). Single-strand mutations occurring during early reprogramming (F), and all mutations occurring during late reprogramming or in culture (G) are present in half or fewer subclones and in one quarter or fewer of homologous chromosomes (VAF <25%). The only non-neuronal mutational category that can pass our calling filters and validation methods are mutations acquired on both strands before the first S-phase following SCNT (E). Such mutations are expected to be extremely rare.

Figure S3, related to Figure 3.

A Coding changes in MT neurons

Type	Neuron	Potency	Gene(s)	Effect
SNV	B4	early gest	Cdc40	Missense:Thr→Ala
SNV	D4	term	Tas2r113	Missense:Phe→Leu
SNV	**D4**	**term**	**Klf16**	**Missense:Gly→Val**
SNV	E1	nd	Dhx37	Missense:Arg→Trp
SNV	**C5**	**full**	**Tekt5**	**Missense:Phe→Tyr**
Indel	C5	full	Gpr44	Codon Deletion: Leu
CGR	**B2**	**full**	**Aven**	**Exon Deletion**
CGR	**B2**	**full**	**Aven**	**Multi Exon Inversion**
Deletion	**C5**	**full**	**Pkd2l2**	**Exon Deletion**
Duplication	E1	nd	Atp10b	Exon Duplication
Deletion	E1	nd	Zic1, Zic4	Gene Deletion

B SNVs and indels in regulatory features

	C5	D4	B2	B3	B4	E1	Total
SNVs	3	0	3	3	4	2	15
Indels	2	1	0	0	1	0	4
Total	**5**	**1**	**3**	**3**	**5**	**2**	**19**

C SNV and indels in introns

MT high expressed genes

	C5	D4	B2	B3	B4	E1	Total
SNVs	11	8	10	4	10	9	52
Indels	4	5	1	1	4	1	16
Total	**15**	**13**	**11**	**5**	**14**	**10**	**68**

MT low expressed genes

	C5	D4	B2	B3	B4	E1	Total
SNVs	32	13	17	14	18	21	115
Indels	6	6	6	3	4	3	28
Total	**38**	**19**	**23**	**17**	**22**	**24**	**143**

D Genes containing intronic indels

High expressed

Pmm1
Dnajc7
Strn3
BC017647
Taf12
Trim44
Tmem108
Stau1
Sh3gl3
BC023829
Exoc4
Rnf220
Ncoa1
Mtf2
1810063B07Rik
Srrm4
D5Ertd579e
Rab11fip2
Mipep
Acvr2a
Nav2
Psme4
Mtss1l
Pacs2
2410137F16Rik
Kif21b
Nr6a1

Low expressed

Lrp4
Grik2
Slc1a6
Kif26a
Epb4.1l4b
Nid2
Slc44a5
Sntb1
Xdh
Gucy2f
Tdo2
Col4a5
Cnn1
Armc4
Cd247
Prl3a1

E Genes containing intronic MEIs

Low expressed

Tenm4
Tubal3
Hpse2

F Genes containing intronic SNVs

High expressed

Rab3b
Syt1
Iqcj-schip1
Nrxn3
Snrk
Psmd7
Cntnap2
Lhfp
Pcdha4-g
Kcnip4
Mpi
Rab37
Dync1i2
Ephx4
Clstn2
Nell1
Ogdh
Rgs7
Deptor
Dnajb4
Cadm1
Ank3
Cadm2
Dpp10
Tmem108
Fhit
Ptpn2
Atxn7l1
Fbxo21
Pknox2
Chn2
Mgat4c
Cand1
Hnrnpm
Pls3
Tbc1d14
Ptpre
Prpf4b
Macf1
Anks1b
Trdn
Rtn4r
Macrod2
Hiatl1
Slc33a1
Cdk9
Sgk1
Nfasc
Inpp4b
Dhrs13
Stx6
Nrxn2
Lnpep
Khdrbs2
Csgalnact1
Rnf111
Trhde
Pkp4
BC018507
Ptplb
9930021J03Rik
Ptprt
Pnldc1
Stxbp5l
Auts2
Lrfn5
Grid1
BC030867
Acad11
Fndc3a
Enox1
Slc20a2
Phip
Zfp810
Syt16
Nkain2
Pex5l
Ankrd26
Frmd4a
Utrn
Serpinb9
9330179D12Rik
Plcl1
Fert2
Cntnap4
Slc4a4
Klf12
Lekr1
Dnm2
Nrxn1
Slc39a11
Erc2
Pip5k1b
Angpt1
2810429I04Rik
2700049A03Rik
Nbeal1
Mmp16
Vat1l
Esrrg
Ptprm
Diap2
Epha6
1700025G04Rik
Sgcd
C130039O16Rik
D430041D05Rik
Bicc1
Trpm3

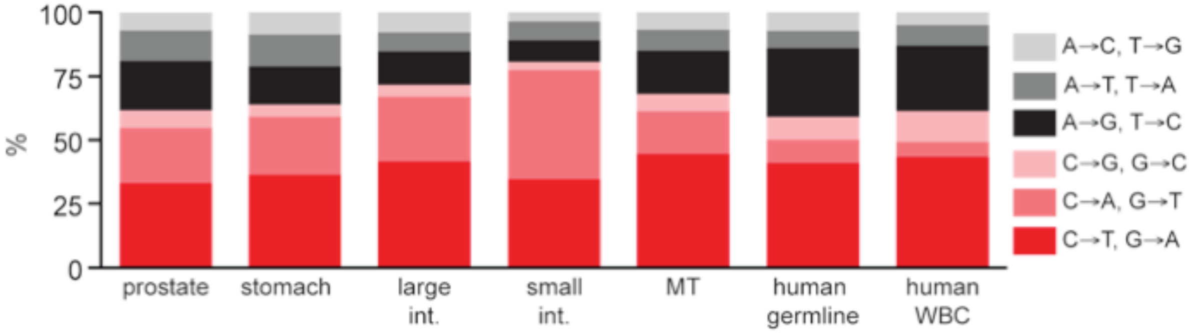
Low Expressed

Slc1a5
Setbp1
Rad54l2
Gm9766
Zcchc24
Pola1
St6galnac5
Osbpl10
Ism1
Csmd3
Palld
Odz2
Pde11a
Etl4
Clnk
Lypd6
Itpr2
Celsr1
Ccrl1
4932425I24Rik
BB123696
Tnni3k
Slc6a4
Zbbx
Gli3
5830432E09Rik
Rgs6
Stab2
Hcls1
Ccdc147
March1
Nkx2-2as
Clrn3
Pik3cg
Gm10556
Ckmt2
Cntnap3
Hnf4a
Sh3tc1
Ptprq
1700012I11Rik
4931428L18Rik
Atp8b4
Esrp1
Gm16294
Gm4850
Gm8633
Lep
Psd4
Sult2a4
Tinag

Features impacted by MT neuron mutations. (A) Mutations that produce coding changes in MT neuron genomes and the developmental potency of the MCNT-ES cells that harbor them, as defined by performance in the TEC assay. Mutations within highly expressed genes are bolded.

(B) Summary of SNV and indel mutations that fall in regulatory features present in e14.5 brain identified by Ensembl regulatory build. Impacted regulatory features include CTCF binding sites, enhancers, open chromatin regions, promoters, promoter flanking regions, and splice regions. (C) Summary of mutations that fall within non-exonic regions of transcripts. (D-F) Genes containing indel, MEI, or SNV mutations in the non-exonic portions of their transcripts. High expression and low expression are defined as the top and bottom 50% of transcripts respectively.

Figure S4, related to Figure 3.

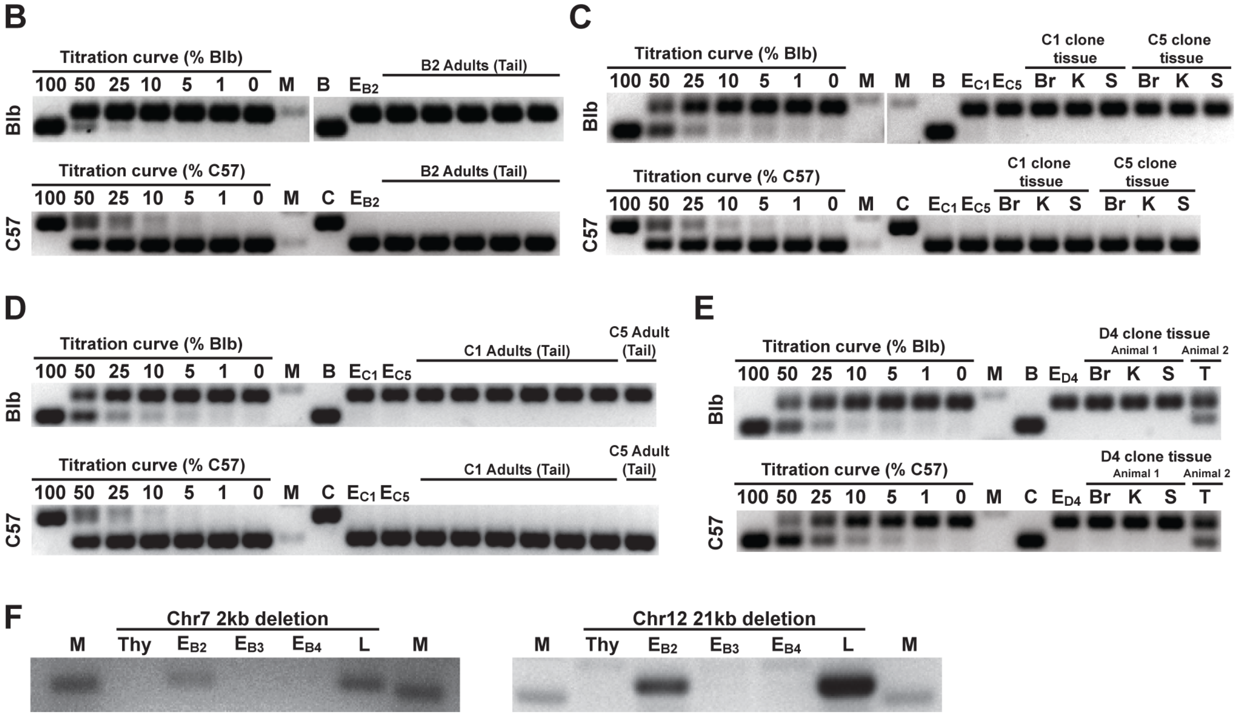


SNV substitutions for MT neurons and other cell types. MT neuron base substitutions and substitutions from other cell types for which comparable genome wide mutation data are available in the literature.

Figure S5, related to Figure 4.

A

MCNT-ES cell line	Blb MS	C57 MS
B2	D17Mit133	D6Mit15
C1&C5	D6Mit102	D6Mit15
D4	D6Mit102	D6Mit102



PCR assays demonstrating absence of TEC host blastocyst contribution and faithful somatic mutation transmission in MCNT-mice. (A) Diagnostic microsatellites used to distinguish MCNT-ES cell DNA from tetraploid host strains Balb/cByJ (Blb) and C57BL/6J-Tyrc-2J (C57). Primary data for mice derived from MCNT-ES cell lines B2 (B), C1 and C5 (C, D), and D4 (E). For each line, analysis was performed on various organs from a newborn animal and on tails from several different adult animals, with the exception of D4, which only produced a single perinatal animal. This single D4 perinatal animal was the only MCNT-mouse to show detectable tetraploid host strain contribution (E), which may explain why it was able to survive longer than a littermate displaying no tetraploid host contribution. For all microsatellite primer pairs, DNA titration curves demonstrate a 5-10% detection limit. M, molecular weight; B, Blb; C, C57; E_{B2}, E_{C1}, E_{C5}, E_{D4}, DNA from B2, C1, C5, and D4 MCNT-ES cells respectively; Br, brain; K, kidney; S, spleen; T, tail. (F) MCNT-ES cell structural variants (SVs) are present in cloned mice. We assayed for two independent SVs from MCNT-ES cell line B2. PCR primers diagnostic for SV breakpoints confirmed the presence of SVs in DNA purified from B2 MCNT-ES cells (E_{B2}) and from B2 clone lung tissue (L) but not in DNA harvested from the thymus of the original Pcdh21/Cre-Ai9 donor animal (Thy) nor in DNA from other MCNT-ES cell lines (E_{B3}, E_{B4}).

SUPPLEMENTAL TABLES

Table S1, related to Figure 2.
Genome sequencing information. Table showing information about the four mice used in this study, along with the tissue source of the control sample, passage number of MCNT-ES cells at the time of Whole Genome Sequencing (WGS), and statistics concerning the WGS runs associated with each sample. The WGS Illumina paired-end sequencing resulted in two paired reads approximately 100bp in length, encompassing an outer template length of ~474bp. The Median Genome Coverage is a measure of the median number of 100bp reads covering each base in the genome, while the Median Physical Coverage also includes those bases in the insert between the two 100mer reads.

Sample	Source	Passage Number	Median Genome Coverage	Median Template Length	Median Outer Span Physical Coverage
C Mouse: 3 week old female					
C0	Spleen	n/a	32	471	78
C1	SCNT	21	34	464	81
C5	SCNT	7	32	479	81
D Mouse: 3 week old male					
D0	Spleen	n/a	34	481	85
D4	SCNT	7	33	486	85
B Mouse: 4.5 month old male					
B1	Thymus	n/a	38	465	90
B2	SCNT	4	59	477	146
B3	SCNT	4	59	464	143
B4	SCNT	4	58	470	149
E Mouse: 6 month old female					
E0	Thymus	n/a	34	469	83
E1	SCNT	7	36	484	88
Range for all samples					
Min		4	34	464	83
Max		7	59	484	149

Table S2, related to Figure 2.
Mobile element library composition. We combined mobile elements from both Repbase and RepeatMasker, as described in Supplemental Experimental Procedures.

Source	Type	Subtype	Count	Minimum Length	Median Length	Mean Length	Maximum Length	Total Bases
Repbase	LTR	ERV1	22	335	502	716	4961	15 752
Repbase	LTR	ERVK	97	292	581	1422	8574	137 934
Repbase	LTR	ERVL	1	479	479	479	479	479
Repbase	SINE	Alu	2	147	148	148	148	296
Repbase	SINE	B2	3	193	193	194	195	582
Repbase	SINE	ID	1	152	152	152	152	152
RepeatMasker	LINE	L1	22 311	150	996	1869	9892	41 691 831
RepeatMasker	LTR	ERV1	1093	166	627	2397	7752	2 619 487
RepeatMasker	LTR	ERVK	7883	150	440	1627	7633	12 822 849
RepeatMasker	LTR	ERVL	3228	152	493	1454	5421	4 692 812
RepeatMasker	LTR	MaLR	7274	150	394	435	1121	3 162 595
RepeatMasker	SINE	Alu	8153	100	145	140	226	1 139 113
RepeatMasker	SINE	B2	5609	100	190	186	228	1 044 940
Total			55 677	100	397	1209	9892	67 328 822

Table S3, related to Table 2. False negative rates (FNRs) for somatic mutation discovery. The FNR is a way to measure the sensitivity of our calling pipelines (Sensitivity = 1-FNR). Note that we also calculate the combined FNR for the B and C mouse. For the other two mice, there is only one MCNT-ES cell line.

Table S4, related to Table 2. False discovery rates (FDRs) for somatic mutation discovery. The FDR is a way to measure the accuracy of our calling pipelines (Accuracy = 1-FDR). For SNVs and indels we tested a subset of calls to estimate the FDR. Since the FDR is so low for HC calls, we use the aggregate value for all mouse lines in Table 2. For SVs and indels, all calls were tested. At the bottom are the statistics for the shared mutation calls.

Table S5, related to Table 2. Somatic single nucleotide variant (SNV) calls. Listed in three groups. The two top groups are for putative mutations that appear in exactly one MCNT-ES cell line. The top most set are high confidence calls, while the second set are low confidence calls. The third set are putative mutations that appear in more than one MCNT-ES cell line from the same mouse. See the bottom of the spreadsheet for a description of the columns (Column Key).

Table S6, related to Table 2. Somatic indel calls. Listed in two groups. The top set are high confidence calls, while the second set are low confidence. See the bottom of the spreadsheet for a description of the columns (Column Key).

Table S7, related to Table 2. Validated somatic SV breakpoints. Column B indicates the MCNT-ES cell line in which the breakpoint occurs. Columns D-N are directly from the output of the lumpy run. Column O provides the result of subclone test to determine if the mutation occurred during reprogramming or clonal expansion. Columns P-U define the architecture of the SV as determined by examination of split-read mapping of the validating PCR product (See Figure 3A-3C). Column V indicates whether the SV was also detected as a Copy Number Variant. Columns W-Z provide the validating primers and their strand orientation. Finally columns AA-AB list the gene effects. The color-coding highlights the breakpoints that comprise the three complex genomic rearrangements as given by the ID in Column C.

Table S8, related to Table 2. Validated somatic Mobile Element Insertions (MEIs). Column B indicates the MCNT-ES cell line in which the mutation occurs. Columns D-N are directly from the output of the MEI calling pipeline (see Figure S2A and S2B). Column O provides the result of subclone test to determine if the mutation occurred during reprogramming or clonal expansion. Columns P-S define the architecture of the MEI as determined by examination of split-read mapping of the validating PCR product and visual inspection of clipped reads in IGV. Column T indicates if the MEI was also detected as a Copy Number Variant. Columns U-X provide the validating primers and their strand orientation. Finally columns Y-AA list the gene effects.

Table S9, related to Figure 3. Summary of genomic enrichment studies. Top: Statistics for the length of the mm9 reference genome in autosomes, as well as the number of bases that fall into gaps or regions in which we have total read depth of < 10 or > 250. Our somatic SNV calling methods exclude such regions. By subtracting this out, we get the number of bases in autosomes in which we might have a SNV call. Middle: Results of Monte Carlo Simulations in nine genomic regions. We calculate the expected number of SNVs in each region if 395 SNVs (our number of HC autosomal SNVs) were randomly distributed throughout the genome, as well as the probability that the number of our actual SNVs that fall in the region is within the 95% confidence interval (by Poisson Test). In addition, the results of simulation runs that count the number of times 395 randomly distributed SNVs fall in the region as shown. From these simulations we can also calculate the probability of seeing our actual number of mutation in the region by chance, giving us two confirming estimates of the enrichment or depletion of our actual SNVs with each genomic feature. Bottom: Sources for the nine genomic regions tested.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Immunohistochemistry

Newborn tissues were dissected, fixed at 4˚C overnight in PBS buffered 4% paraformaldehyde (PFA/PBS). Adult tissues were perfused with PFA/PBS, dissected, and fixed in PFA/PBS for 30 minutes on ice. After fixation, all tissues were sucrose protected in 30% sucrose at 4˚C overnight. Tissues were embedded in OCT and cryosectioned into 15 μ sections using a Leica CM3050S Cryostat. Sections were air-dried on superfrost slides for 40 minutes and fixed in PFA/PBS for 8 minutes. They were stained with primary antibodies against Iba1 (Wako, 019-19741, RRID:AB_839504, 1:1000), Ki67 (Acris, DRM004, RRID:AB_1004358, 1:200), Olig2 (gift of Dr. Charles Stiles, Harvard Medical Center, RRID:AB_2336877, 1:20,000), S100b (Abcam, ab868, RRID:AB_306716, 1:500), Tbr2 (Abcam, ab23345, RRID:AB_778267, 1:500). ES cells were stained as in Boland et. al.(Boland et al., 2009) using primary antibodies against Oct4 (Santa Cruz Biotechnology, sc-5279, RRID:AB_628051, 1:100), SSEA1 (Developmental Studies Hybridoma Bank, MC-480, RRID:AB_528475, 1:500), Nanog (Cosmo Bio Co., REC-RCAB0002P-F, RRID:AB_10706358, 1:50), Sox2 (R&D Systems, MAB2018, RRID:AB_358009, 1:50). Images were collected on a Nikon C2 or Nikon A1 confocal microscope and analyzed in Adobe Photoshop.

Isolation of MT neurons for nuclear transfer

Pcdh21/Cre-Ai9 mice were generated by crossing *Pcdh21*/Cre mice to Ai9 reporter mice (RRID:MGI_MGI:1932557). MT neurons were dissociated and purified as in Brewer and Torricelli (Brewer and Torricelli, 2007) with the following modifications. We found it unnecessary to siliconize Pasteur pipettes to prevent cell loss and chopped olfactory bulbs using a scalpel rather than with a tissue slicer. We also used papain-containing L-cysteine (Worthington Biochemical, PAP2 10 units/ml) as it has higher activity and allows shorter dissociation times (10 minutes total). We found it essential to add small amounts of DNaseI (6.25 μg, Roche 10104159001) during papain treatment to prevent DNA related cell aggregation. After density gradient centrifugation, we found most MT neurons in the cell pellet fraction and the 2 ml fraction immediately above the pellet. Cells from both fractions were combined and washed once in 10 mls of HAGB (Hibernate-A (Gibco A1247501), 1X B-27 supplement (Gibco 12587010), 500 μM GlutaMAX (Gibco 35050061)). After pelleting, cells were resuspended in 1 ml HAGB, transferred to a 1.5 ml centrifuge tube, pelleted again, resuspended in ~30 μls HAGB media, and stored on ice until nuclear transfer.

Derivation of MCNT-ES cell lines

SCNT was performed as in Kishigami et al. (Kishigami et al., 2006). Briefly, oocytes were harvested from superovulated females and metaphase II spindles were removed and replaced with neuronal nuclei using an 8 μ injection pipette. Embryos were then treated with 5 nM Trichostatin A and artificially activated with strontium chloride. We extended the length of treatment with 5 nM Trichostatin A to 16 hours (6 hours during activation, 10 hours overnight) to improve efficiency of blastocyst and NT-ES cell generation (Kang and Roh, 2011). Embryos resulting from NT were cultured to blastocyst stage and zonae pellucida were removed using a piezo-actuated drill needle (Nakayama M.D et al., 1998). ES-cell lines were derived essentially as described in (Meissner et al., 2009), with some modifications in media composition. Briefly, zona-free embryos were cultured for 7-9 days on MEF feeder layer in ES-cell derivation medium. ES-cell derivation medium contained: 500 mls Knockout DMEM (Gibco 10829-018), 80 mls Knockout Serum Replacement (Gibco 10828-028), 6 mls MEM non-essential amino acids (Gibco11140-050), 6 mls Glutamax (Gibco 35050-079), 6 mls Pen/Step (Gibco 15140-122), 6ul B-Mercaptoethanol (Sigma M7522), 50 µm final concentration MEK1 Inhibitor PD98059 (Cell Signaling Technology 9900) and 2000 Units/ml LIF (Chemicon ESG1107). Outgrowths of inner cell mass were picked and dissociated with 0.25% trypsin-EDTA. Cells were then expanded on a MEF feeder layer in ES-cell maintenance medium which contained: 500 mls Knockout DMEM (Gibco 10829-018), 80 mls Knockout Serum Replacement (Gibco 10828-028), 6 mls MEM non-essential amino acids (Gibco 11140-050), 6mls Glutamax (Gibco 35050-079), 6 mls Pen/Step (Gibco 15140-122), 6ul B-Mercaptoethanol (Sigma M7522) and 1000 Units/ml LIF (Chemicon ESG1107).

Microsatellite PCR assay to rule out host blastocyst contribution

This assay was described by us previously (Boland et al., 2009), and relies on the detection of microsatellites that vary in length between MCNT-ES cells and tetraploid embryo cells. In these experiments, tetraploid embryos were F2 (BALB/cByJ X C57BL/6J-Tyr^{c-2j}). Therefore, to rule out trace contribution of tetraploid cells to MCNT-mice, we assayed for differences in microsatellite length diagnostic of both BALB/cByJ and C57BL/6J-Tyr^{c-2j} strains. Microsatellites assayed for each MCNT-ES cell line are listed in Figure S5A. The primers used were:

D17Mit133:
-forward TCTGCTGTGTTCACAGGTGA
-reverse GCCCCTGCTAGATCTGACAG

D6Mit102:
-forward CCATGTGGATATCTTCCCTTG
-reverse GTATACCCAGTTGTAAATCTTGTGTG

D6Mit15:
-forward CACTGACCCTAGCACAGCAG
-reverse TCCTGGCTTCCACAGGTACT

Whole genome sequencing

Prior to sequencing, early passage MCNT-ES cells were separated from feeders by serial pre-plating on gelatin coated tissue culture dishes. DNA was isolated from MCNT-ES cells and thymus or spleen using standard phenol chloroform extraction, ethanol precipitation and RNase A digestion. Samples were sequenced by Beijing Genomics Institute using standard library prep for an Illumina Hi-Seq 2000. During library preparation target template length of approximately 500bp was chosen to give increased physical coverage to aid in accurate structural variant discovery. Each end of the paired-end data was 100bp in length. Quality control was performed on the output of the sequence run to eliminate reads with low base quality (≤5 (“A”-“E”)) over at least 50% of their length as well as reads with unknown nucleotides (“N”) over at least 10% of their length.

Initial alignment and post-processing

In these studies, default parameters were used for all bioinformatics software except as explicitly noted. We refer to an index with word length L and skip distance S as a L/S index. Mouse MCNT-ES cells and thymus/spleen control samples were sequenced using Illumina next-generation whole genome shotgun paired-end sequencing in which each read in the pair was approximately 100bp in length with a template length of approximately 475bp. Each sequencing lane was then separately aligned to the mm9 reference genome (July 2007 NCBI Build 37) using Novoalign v2.08.02 (Hercus) using a 14/1 index (-k 14, -s 1). Repetitive alignments were resolved using the random selection method (-r random).

GATK (v2.5-2-gf57256b) (DePristo et al., 2011) and Picard Tools (v1.92) (Broad) were used to further process alignments. Read group, library, platform, platform unit, and sample name information was added to the above alignments using Picard AddOrReplaceReadGroups. The BAM files for the various sequencing lanes for each cell line were then position sorted and merged using Picard ReorderSam and MergeSamFiles respectively. Duplicates were marked using Picard MarkDuplicates and removed with samtools view (Li et al.) (-F 0x400), resulting in a non-duplicate median per sample read-depth of approximately 32x-39x (Table S1).

SNV and indel Detection

GATK and Picard Tools were further used for single nucleotide variant (SNV) and indel calling following the recommended best practices pipeline for GATK v2.0 (Van der Auwera et al., 2002). Here “indel” refers to any insertion or deletion of consecutive bases of less than 50bp in length. The GATK IndelRealigner was used to realign indel regions identified by RealignTargetCreator. Mate-pair information was cleaned by Picard FixMateInformation. We then used GATK BaseRecalibrator and PrintReads to recalibrate base quality scores. This step takes as input a set of known sites, which we created by selecting those single nucleotide polymorphisms (SNPs) marked as "High Confidence" by the Mouse Genomes Project (MGP) in the 129S1 mouse strain (Keane et al.). The GATK UnifiedGenotyper was then run on all samples combined, calling indels and SNPs together, using per sample read-depth downsampling to a maximum read-depth of 500 (-glm BOTH –dt BY_SAMPLE –dcov 500).

GATK VariantRecalibrator and ApplyRecalibration steps were then run first on SNPs (--mode SNP), then on indels (--mode INDEL), to assign our calls into one of four sensitivity tranches. These steps require SNP and indel “truth” sets that were created as follows. For SNPs, we again started with the high confidence 129S1 SNP calls from the MGP, intersected these with our own autosomal GATK SNP calls from above, and selected the top 1 million such calls as ranked by the MGP variant quality score. For indel variant recalibration, we used all 129S1 indel calls from the MGP.

We then identified putative *de novo* somatic SNV and indel variants private to MCNT-ES cells lines using custom scripts that implement a modified version of the approach used by Kong et al. (Kong et al., 2012) . Although we called variants in each donor mouse separately, we used information from the same locus across all samples to help reduce false positives. For a given mouse, the samples were partitioned into three sets; (1) the “control” sample of the thymus/spleen for that mouse, (2) the “MCNT-ES” cell line(s) for that mouse, and (3) the “other” samples, comprised of all samples from the other mice. In what follows, “RR”, “AR” and “AA” will refer to the genotype of a locus as homozygous for the reference allele (R), heterozygous, or homozygous for the alternate allele (A) respectively. The alternate allele genotype (AAG) of interest for the calling process depends on the chromosome and sex of the mouse. We used AR for all autosomes and for the X chromosome of female mice, and AA for the X/Y chromosomes of male mice. The variant allele frequency (VAF) is defined as the (alternate allele read-depth)/(reference allele read-depth + alternate allele read-depth). Phred likelihood scores for genotypes and per allele read-depth information are provided by GATK in the VCF output file.

To be called a putative somatic SNV in a particular MCNT-ES cell line, a SNV locus/allele pair was required to meet all of the following criteria:

1. The alternate allele is not reported as a variant at the same locus in any inbred mouse strain by the MGP at either high or low confidence.
2. The call appears in one of the 19 autosomes or the X or Y chromosome. No calls are made in “random” or “unknown” scaffolds. Mitochondrial variant calls were also excluded from the analysis because mitochondria in MCNT-ES cell lines are expected to originate from the oocyte used in nuclear transfer, not from the original neuron.
3. The control sample and the MCNT-ES cell line(s) from the mouse of interest each have a total read-depth between 10 and 250.
4. A control RR/AAG ratio of phred likelihood scores >= 10^5, and a control VAF of at most 5%.
5. An MCNT cell line AAG/RR ratio of phred likelihood scores >= 10^{10}, and VAF of at least 30% (95% for X/Y chromosomes in male mice).
6. A RR/AAG ratio of phred likelihood scores >=1, and a VAF of at most 5% for all “other” samples.

Indel calling strategies are known to have higher false positive rates than SNV calling strategies. Therefore, we slightly modified the filtering criteria for indels to be more conservative as follows. In step 1, the variant is eliminated as a somatic call if it overlaps any indel reported by the MGP in an inbred mouse strain regardless of the type and size of the indel. In steps 4 and 6, the VAF for both the control sample and all samples from other mice are held to the stricter criteria that they be equal to zero.

We further categorized our SNV and Indel calls by the GATK VariantRecalibration assigned tranche annotation as high confidence (HC) if they fall into the two lowest sensitivity (highest specificity) tranches with an implied false discovery rate (FDR) threshold for the corresponding truth set of 1%. The remaining calls are categorized as low confidence (LC). As discussed below, our validation rates are markedly higher for the HC calls than for the LC calls. The resulting somatic SNV and indel calls are in Table S5 and Table S6.

Structural variation breakpoint detection

We used Lumpy (Layer et al.) to detect structural variant breakpoints. Here we define a structural variant (SV) as an apparent deletion, tandem duplication or inversion (as defined by relative read-pair orientation) of greater than 50bp in length, or an unexpected juxtaposition in the sample genome of two loci that appear far away from each other (>1 Mb) on the same or different chromosome(s) in the reference genome (which we refer to as "distant" rearrangements). Insertions are not directly detected by Lumpy, but will instead be composed of two of the above event types (one for each of the two insertion breakpoints).

Lumpy can map SV breakpoints using evidence from both discordant paired-end reads ("read-pairs") and split-read mappings from multiple samples to find SVs. Informative discordant read-pairs were extracted from each BAM file as those read-pairs in which both reads were mapped, the mappings were either 1) on different chromosomes, 2) had improper strand orientation, or 3) had a template length that fell outside the mean length +/- 5 standard deviations (STDs). The insert size mean and STD was calculated for each dataset using custom scripts using properly paired alignments (samtools view –F 0x400 –f 0x2). In order to reduce the probability of false positive SV calls, we further filtered the set of input discordant read-pairs as follows. We first located collections of nearly duplicate pairs in which the corresponding mates of each pair mapped to the reference genome within ±3bp of each other. From such collections, we eliminated all but the pair aligned with the least edit distance from the reference genome. Discordant reads were converted to bedpe format using bedtools V2.16.2 (Quinlan and Hall, 2010) bamToBed and pairBedToBedpe, and additional duplicates were removed using dedupDiscordantsMultiPass.py (-s 3).

Separately, we extracted putative split-read alignments that were either unmapped or had a clipped region of >= 20bp on either end of the alignment. These were then realigned using YAHA version 0.1.78 (Faust and Hall) with an 13/1 index (-L 15 –S 1), and default alignment parameters except for maxHits of 2000, and minMatch of 15 (-H 2000 –M 15). From the resulting alignments, we selected for input to Lumpy reads that had a single split alignment (two mappings) in which each aligned portion involved >=20bp of query sequence that was not included as part of the other aligned portion. We also required that split-read alignments suggesting a deletion variant had an implied deletion size >=50bp (our definition of SV).

Lumpy was run on the above-described discordant read-pairs and split-read mappings from all eleven samples, requiring at least 4 confirming reads across 11 samples for a call, and a trimThreshold of 10^{-3} (-mw 4 –tt 1e-3), using a minimum alignment mapping quality of 10, and excluding all genomic regions in which any cell line had an aligned read-depth >500.

The resulting SV calls were filtered to find putative *de novo* somatic variants that appear in a single MCNT-ES cell line. We required such a call to meet all of the following criteria:

1. The SV call had at least 5 supporting discordant read-pairs and/or split-reads from one MCNT-ES cell line, and no supporting reads in any other MCNT-ES or control sample from any mouse.
2. The SV call was not previously reported as a germline polymorphism by MGP for any mouse strain. A LUMPY call was judged to correspond to an MGP call if the two were of the same variant type (e.g., deletion) and were at the same genomic location, as defined by 50% reciprocal overlap (bedtools intersect -r -f 0.5). Distant rearrangement involving >1mb of genomic sequence, or spanning multiple chromosomes, were not filtered in this manner since such variants were not reported by MGP.
3. The call appears in one of the 19 autosomes or the X or Y chromosome. No calls were made in unmapped contigs or in mitochondrial DNA.

Mobile element insertion detection

Mobile element insertions (MEIs) pose a challenge for SV calling algorithms due to several factors including the fact that the mobile element (ME), or "transposon", is itself composed of repetitive sequence. Therefore, we have developed our own MEI calling pipeline based on the strategy used by Lee et al to study somatic retrotransposition in human cancers (Lee et al., 2012).

The general approach is to start with all the reads that the aligner had difficulty aligning to the reference genome, and re-align them to a custom-built library of mobile element sequences. The mates to the reads that map well to this ME library are then used to identify regions of the sample genome in which to search for MEIs. In addition, we look for confirming evidence of MEIs using split-read mappings in which one side of the split maps to the ME library, and the other side to the reference genome next to the predicted ME insertion point (Figure S2A and S2B).

The ME library is formed using both canonical sequences from version 18.02 of RepBase (Jurka et al., 2005) and their variants predicted to appear in the mm9 reference genome by RepeatMasker (Smit et al.) and included in the mm9 UCSC RepeatMasker annotation track (downloaded from http://genome.ucsc.edu/cgi-bin/hgTables/). From RepBase RepeatMaskerLib.embl and mousub.ref (downloaded from http://www.girinst.org/server/RepBase/) we found 120 LTR sequences labeled with "Species: Mus_musculus" and 6 SINE sequences, respectively. From the mm9 RepeatMasker annotation track we selected the genomic regions for all LINEs, SINEs, and LTRs with low sequence divergence (<= 30 millidev) and length of at least 100bp, then extracted the corresponding DNA sequences from the reference genome using bedtools getfasta. We then removed duplicate sequences from the above, and appended multiple "N" bases to the ends of each sequence to aid in alignment. The final ME library contains 51,413 unique sequences. Detailed information about the composition of the ME library can be found in Table S2.

We selected reads to align to the ME library that met any of the following criteria:

1. It was the unmapped read of a pair in which one read is mapped and the other unmapped.
2. It was either read of a discordant read-pair in which either the reads were aligned to separate chromosomes, or the reads were aligned at least 100 kb apart from each other.
3. Any mapped read not in the above two categories whose alignment was clipped by at least 20bp.

The above reads were then aligned to the ME library using YAHA. Since the ME library is highly repetitive, we used very sensitive alignment parameters: an 11/1 index, maxHits of 9000, minMatch of 15, and a maxGap of 20 (-H 9000 –M 15 –G 20).

We then formed clusters separately for each ME subtype as follows. From the ME library alignments, we selected ones matching the current ME type and subtype that were from a discordant or unmapped read and had a good alignment to the ME library, defined as at least 50bp in length and clipped by no more than 3bp on at least on end. We then extracted the aligned coordinates for their mate in the reference genome, and formed clusters from those reads that were aligned to the same strand and fell within the inter-read distance from each other. The inter-read distance is calculated separately for each sample as $\lfloor(\mathrm{ETL} - \mathrm{RL}) \times 2 / 3\rfloor$ where ETL is the extended template length (median template length + 3 STDs) and RL is the read length (100). We then found potential ME insertion points as a pair of such clusters from the same ME type and subtype such that the reference coordinates of a plus strand cluster were 5' of a minus strand cluster within twice the inter-read distance, and had at most 20bp of overlap. In addition, the cluster pairs had to have at least 6 combined supporting reads from the two clusters. These pairs were then filtered to exclude those with at least 25% of their length overlapping an ME of the same type and subtype annotated in the reference genome as defined by the UCSC repeat masker tract ME (bedtools intersect –f 0.25).

Confirming split-read mappings for remaining pairs were found as follows. All unmapped alignments, and any clipped alignments overlapping a pair region were aligned to the reference genome with YAHA using the same parameters as above; an 11/1 index on the mm9 reference genome, and these alignment parameters: (–H 9000 -M 15 –G 20). We then counted as a confirming split-read mapping one in which the type and subtype of the ME matched the one from the cluster pair, the portion of the read aligned to the reference fell within exactly one of the pair clusters, and the two split-read alignments together cover almost the entire read length with at most a few unmapped bps (the alignment mapped to the reference and the alignment mapped to the ME library ended within 3bp of opposite ends of the read, and there was a maximum of 4bp of unaligned sequence between them). We added the count of such split mappings to the total read count of the associated cluster, and kept a list of all of the reference loci for their reference aligned portion nearest to the implied insertion breakpoint to more precisely define where the breakpoint occurred (Figure S2A and S2B).

We then filtered the cluster pairs formed above to find putative *de novo* somatic MEIs in a single MCNT-ES cell line using a similar strategy we used to identify *de novo* somatic SV events. We first eliminated cluster pairs that had fewer than 10 confirming reads. We then eliminated cluster pairs with evidence in other samples as follows. We separately intersected the genomic region of each cluster of a pair with clusters from all other samples that had the same ME type (disregarding subtype) and were on the same strand. We then eliminated all cluster pairs that had any confirming reads from such a matching cluster. Finally, we further filtered the remaining pairs to eliminate any pair that had any overlap with any MGP MEI call from any mouse strain regardless of ME type or subtype.

Copy number variation detection by read-depth analysis

To detect copy number variation (CNV), we used a read-depth strategy very similar to the one described by Malhotra et al. (Malhotra et al., 2013). Assuming that Illumina genome sequencing uniformly samples the source DNA, the DNA copy number within a given genomic region should be directly proportional to the number of sequence reads mapped to the region relative to other regions. However, local read-depth is subject to two major sources of bias that must be overcome to make these calculations more accurate. First, Illumina sequencing exhibits significant GC bias such that local coverage depth falls off at GC content extremes, especially in regions with high GC percentages (Aird et al., 2011). To counteract this bias, we normalize the coverage data within small genomic regions by their percent GC content. The strategy used to do this normalization is based on the observation that the read-depth in regions of similar GC content approximates a normal distribution. Second, repetitive sequences are known to pose difficulties in sequence alignment and assembly, causing potentially large fluctuations in local read-depth mapping to the reference genome. To counteract this bias, we base all of our calculations on read-depth in unique genomic regions.

We therefore start by breaking the reference genome up into regions (“windows”) containing 5kb of unique sequence as defined by a mappability value equal to 1 in the UCSC 100mer mappability track (crgMappabilityAlign100mer). This results in 458,040 windows with a mean and median size of 5796 and 5030 bp, respectively.

We then process each of our cell samples separately as follows. We first count the number of reads mapped to the unique portion of each such 5kb window, then consider as a group those regions with the same percent GC content in 1-3% increments, e.g. (45.0-47.0%] GC. We then use the autosomal windows in each group to calculate the median and median absolute deviation (MAD) of read-depth for the group, and estimate its normal distribution using the MATLAB “normfit” function using all windows in each group that are within ±4 MADs from the median read-depth for the group. This yields a mean and standard deviation (STD) for each group as a whole. For each window within the group we then calculate the normalized read-depth as the raw read-depth for the window divided by the median read-depth for the group and multiply by two (assuming a diploid genome). Similarly we calculate a Z-score for each window as the raw read-depth for the window minus the mean read-depth for the group divided by the STD.

We next combine consecutive windows with similar Z-scores into copy number segments as described in (Malhotra et al., 2013) using the circular binary segmentation function in the DNAcopy package in R (http://cran.r-project.org/) with the following parameters: (undo.splits=”sdundo”, undo.SD=1 and alpha=0.001). For each segment we keep track of the count, mean, STD, median, and MAD of the read-depth values for windows it contains. In addition, for each SCNT-ES cell line, we performed the same segmentation as above based on the $\log_2$ of the ratio of the normalized cell line read-depth divided by the corresponding thymus/spleen control sample read-depth. Such a division is useful for determining somatic CNVs as described below. We also calculated the total dataset median and MAD for each cell line and $\log_2$ ratio dataset separately for autosomes and the X chromosome to account for the expected difference in copy number on the X/Y chromosomes in males.

Finally, we called the somatic CNVs as follows. As CNV calling is fairly error-prone, we chose to use conservative filters that result in a low false positive rate, but potentially lower sensitivity. We find all segments in the $\log_2$ ratio datasets for the MCNT-ES cell lines that are formed from at least 3 windows and have a normalized segment median read-depth that is plus/minus at least 6 MADs above/below the full dataset median normalized read-depth for the corresponding chromosome set (autosomes or X chromosome as appropriate). From these, we remove any segment(s) that overlap with a segment in any of the 4 control samples with a normalized segment median read-depth that is plus/minus at least 6 MADs above/below the full dataset normalized read median read-depth. Together, these filters require a strong signal in one or more of the MCNT-ES cell lines in a genomic region that has no such signal in any control line, indicating a *de novo* somatic variant. Interestingly, this filter criteria results in putative CNV duplication calls in T-cell receptor alpha and/or gamma sites for all MCNT-ES cell lines using thymus as the control sample (B2, B3, B4, and E1). These are actually an artifact of the deletions in these regions in the thymus samples due to V(D)J recombination, and act as a positive control for the calling pipeline. Removing these spurious calls leaves us with four CNV calls all of which are also LUMPY SV breakpoint calls as shown in Table S7. Note that the above calling strategy requires segments of at least three adjacent 5 kb windows and is insensitive to any CNV below ~15 kb is size. We have only five validated LUMPY breakpoint calls that are unbalanced variants of this length. Four of them are found as CNVs by read-depth analysis, and the fifth duplication call falls just below our detection thresholds in a segment three windows in length with a normalized copy number that is 4.7 MADs above the median.

Somatic variant false negative rate estimations

To gauge the sensitivity of our somatic variant calling strategies in the absence of a known set of true positives, we estimate the false negative rate (FNR), and calculate the sensitivity as 1-FNR. The general strategy is to find a set of high confidence germline variants of the variant category of interest, called the gold standard set (GSS), and then count how many of these were detected in our analysis and would pass all relevant MCNT somatic call filters. To eliminate issues regarding sex chromosome differences across datasets from both male and female mice, all of our FNR estimates are based solely on autosomal variants. The detailed calculations of FNR estimates are shown in Table S3.

Single nucleotide variant and indel false negative rate estimation

The GSS set for SNV calls was found on a per mouse basis as follows. We started with the set of all GATK autosomal SNP calls for a given donor mouse, and selected the subset of such calls that were also found as high confidence calls by the Mouse Genomes Project (MGP) in any inbred mouse strain. From this set, we further selected those that were called heterozygous in our data by GATK in at least one sample from the mouse in question. This is an important step, as we expect that, barring some rare event that causes loss of heterozygosity, all *de novo* somatic autosomal variants should be heterozygous. We therefore use solely heterozygous calls in our GSS as they should display similar patterns of variant allele frequencies and associated genotype phred likelihood scores as our sought-after somatic variants.

We then applied all of our MCNT-ES cell line filtering criteria except the “control” and “other” sample filters (filters 4 and 6 from above), and counted the percentage of the GSS calls that are eliminated in each MCNT-ES cell line. We take this as our estimate of overall FNR for that cell line. We then calculated the overall FNR rate for each mouse as the average of the FNRs of the (one or more) MCNT-ES cell line(s) from that mouse. The resulting per-mouse overall FNR estimates for all SNV calls range from 6.7% to 11.1%, and for our high confidence SNV calls from 19.0% to 22.8%.

We estimate the FNR rates for the indels in a similar fashion with one difference. The MGP does not report confidence levels for indels. Therefore, we intersected our per-mouse GATK heterozygous autosomal indels calls with all MGP inbred indel calls to find the per-mouse GSS set. The resulting per-mouse overall FNR estimates range from 22.5% to 27.0%, and for our high confidence calls from 24.3% to 28.6%. Note that the FNR estimates for our high confidence SNV and indel calls are quite similar, while those for all indel calls are significantly higher than for all SNV calls. This is not surprising given the increased difficulty in calling indels vs. SNP and the lower quality “truth” set we had available as input to the GATK tranche calculations, which resulted in GATK placing almost all of the indel calls in the high confidence tranches. See Table S6 for details.

Structural variant and mobile element insertion false negative rate estimation

For our SV and MEI FNR estimates, we also calculate a gold standard set (GSS) on a per mouse basis. To find our GSS set, we started with MGP calls from the 129S1 mouse strain, and found the subset of these that are located in genomic regions that we predict to be in a haplotype block inherited from the 129 strain lineage in the mouse of interest. This is necessary because the different donor mice are mixed 129/Black6 genetic background, but due to their breeding history have inherited distinct 129 haplotype blocks. To find these haplotype regions, we first determined the set of germline SNPs called by GATK in each mouse that are also called by the MGP in the 129S1 mouse strain. We call these the 129S1-SNPs for that mouse.

Deletions are the most numerous and easiest to detect structural variants. We therefore have highest confidence in the deletion call annotations in the MGP. To estimate the FNR of our LUMPY SV breakpoint calls, we restricted our GSS to MGP deletions found in the 129S1 mouse strain. We further restricted the GSS to those calls that have two 129S1-SNPs within 250bp of both sides of the outer span of the call region. This results in per-mouse GSSs with ~2200 calls each. For initial FNR estimates, we counted the percentage of the GSS calls that do not have 50% reciprocal overlap with a LUMPY deletion call in the cell line of interest. The resulting per-mouse FNR estimates range from 38% to 42%. However, this dramatically overestimates the true FNR. Approximately half of the calls in each GSS are small (less than 500 bp). For these deletions, the uncertainty in the breakpoint location calculated by LUMPY is large relative to the size of the call. As a result, approximately 75% of these small calls failed the above test compared to only 6% to 9% for larger deletions. Therefore, for more accurate FNR estimates, we required 25% and 50% reciprocal overlap for the small and large calls respectively, then formed a weighted average of the resulting FNR estimates leading to final per-mouse FNR estimates ranging from 10.2% to 13.5%.

For our MEI calls, we formed an initial GSS in a similar fashion to the deletion calls. We chose those MGP MEI calls from the 129S1 mouse strain that have two 129S1-SNPs within 250bp on both sides of the insertion point, estimated as the midpoint of the insertion call region. We then counted the percentage of these calls that do not intersect the insertion region of any of our MEI calls of the same ME type in the cell line of interest. This results in initial per-mouse FNR estimates ranging from 45.8% to 48.2%. However, it is likely that this is an overestimate of FNR due to false positive MEI calls in the MGP. Therefore, we formed a stricter GSS for each mouse by adding the requirement that we have at least weak evidence for the insertion in our data. Specifically, we required there be at least two reads from any of our clusters from the same ME type that overlap the insertion region of the GSS call. We then again count the percentage of these restricted call set that do not intersect the insertion region of any of our MEI calls of the same ME type. The resulting per-mouse FNR estimates range from 19.4% to 15.5%. These probably underestimate the true FNR rate because we have pre-selected MGP calls that we are likely to find. The real FNR rate probably lies between these two extremes. See Table S3 for details.

Validation of putative somatic SNVs

To test SNV calls, we designed PCR primers to amplify the region of genome containing the predicted SNV. PCR was performed on genomic DNA from MCNT-ES cells and from thymus or spleen of the original donor animal. The resulting PCR product was sequenced by Sanger sequencing, either directly, or after gel extraction if greater than one PCR product was amplified. Most PCR products were sequenced using either the forward or reverse primer from amplification. Sequencing results were aligned to the mouse genome to confirm the intended region was amplified before specifically looking for the presence or absence of the predicted SNV. If the predicted mutation was present in the predicted MCNT-ES cell sample and not in control donor tissue, the SNV was judged to be validated.

Validation of putative somatic indels

Indel validation was essentially identical to SNV validation. However, in SNV detection, single base polymorphisms are visible directly in the sequencing data. In indel validation, longer heterozygous sequences result in a decay of the quality of the sequencing data starting with the first base that differs between the reference and mutant alleles. So, the presence, bounds, and in most cases the actual sequence of the indel were confirmed by Sanger sequencing from both upstream and downstream of the predicted indel. PCR primers are listed in Table S6.

Validation of putative somatic structural variants and MEIs

To validate putative *de novo* somatic SV and MEI breakpoints, PCR was performed on genomic DNA from MCNT-ES cells and donor animal thymus or spleen as control. Primers were designed to flank a putative SV breakpoint to produce a 200-800 bp product for the variant allele, and to produce either no product or a product of significantly different size for the reference allele. Primers were designed to be 18-25 bp in length, with a 57°C-63°C Tm, and 40%-60% GC content. All validating primers are listed with their corresponding variant call descriptions in Table S7 and Table S8. CNV calls were not separately validated, as all somatic CNV calls were redundant with a validated SV call.

If a unique amplified product was present in the predicted MCNT-ES cell line(s) but not the control, the breakpoint was considered validated. If the same product(s) were present in both the predicted MCNT-ES cell line(s) and control DNA, the breakpoint was judged to be a germline variant. If amplified products were absent in all lines, or if the primers were non-specific (i.e., yielded multiple products) a second pair of primers were made. If the second pair of primers also failed to yield specific product(s) then the variant was judged to be a false positive. We note that this could result in a small number of false negatives due to off target amplification at loci that are difficult to amplify cleanly. The unique band produced by validating primers was cut from the gel and sent to GENEWIZ (http://www.genewiz.com) for capillary sequencing of both strands.

To further determine that validated SV and MEI calls were present in the original donor neuron and did not arise either in culture or during reprogramming, PCR was performed with the validating primers on subclones from the relevant MCNT-ES cell line. DNA from MCNT-ES cell subclones was purified in 96-well format using the following protocol. MCNT-ES cell subclones were grown to confluency on MEF feeder cells. They were then washed with PBS and incubated in 50 ul lysis buffer (100 mM Tris pH8.0, 5mM EDTA, 0.2% SDS, 200mM NaCl, 100 µg/ml proteinase K) for 16 hours at 55˚C. To precipitate DNA, lysed cells were incubated in 100 uls of cold 100% ethanol for 30 minutes on an orbital shaker. Supernatant was removed, and precipitated DNA was washed twice with 70% ethanol and air dried for 20 minutes. The resulting DNA was resuspended in 35 µl of TE by incubating overnight at 37 ˚C. PCR was then performed on 1ul of DNA from subclones.

Structural variant and MEI breakpoint determination

SV and MEI call breakpoints were determined to single base pair resolution primarily by split-read mapping of the capillary sequence data of the unique PCR product validating the call. Split-read mapping was done using YAHA with sensitive parameters and a breakpoint penalty neutral to variant length (a 11/1 index, -M 12 -BP 20 -MGDP 1 and –H 2000 for SVs and –H 65525 for MEIs). However, all of the four validated MEI LINE insertions had PCR validation of only one breakpoint due to the difficulty in finding usable primers in poly-a tails. Therefore, these breakpoints were determined by visual inspection of clipped alignments using the Integrative Genomics Viewer (https://www.broadinstitute.org/igv/home). Once the breakpoint locations were determined, we calculated additional breakpoint features by looking for additional features of the split-read mappings. Microhomology for SV breakpoints manifests as overlap of the two split-read alignments on the query, and target-site duplication for MEIs as the distance between the insertion breakpoints on the reference (Figure S2B). The details of the breakpoint architectures of SVs and MEI are provided for each validated call in Table S7 and Table S8. In addition, about half of the SV breakpoints were caused by complex genomic rearrangements as shown in Figure 3.

Detection and validation of shared mutations.

We sought to identify somatic mutations that are shared among multiple MCNT-ES cell lines. Somatic variants that are shared among cell lines derived from a single donor mouse could exist due to clonal mutations that arose early in development, whereas variants that are shared among lines from different donor mice could exist due to recurrent mutation at hotspots, or conceivably due to programmed rearrangement (as in the immune system). Since it has long been hypothesized that recurrent structural mutations might be involved in generating neuronal diversity, we focused our search for mutations shared across different donor mice to SVs and MEIs. We restricted our search for shared SNVs to within-mouse mutations.

Within-mouse shared SNVs are naturally detected by our primary SNV calling procedures outlined above. We identified 13 such SNV calls in the B mouse and 8 in the C mouse (bottom of Table S5). Shared SVs were identified from the primary LUMPY run as before, except that we modified criteria 1 to require at least 5 supporting reads in each of two or more MCNT-ES cell lines. These criteria identified 73 shared SV calls of which 13 well-supported candidates were tested; 10 were shared between two mice, and 3 were shared among MCNT-ES cell lines from the same mouse. Shared MEIs were identified as before except that we selected pairs that had at least 6 overlapping cluster reads in at least one other MCNT-ES cell line. These criteria identified only three shared MEI calls; two within the B mouse, and one shared between two mice. No CNV calls made by read-depth analysis were shared among multiple cell lines.

We attempted to validate putative shared mutation calls using the same methods described above, except that we included all relevant MCNT-ES and control samples during PCR validation and subsequent Sanger sequencing. We were able to successfully make primers that yielded a product that could be sequenced for 10 of the shared SNV calls, all of which showed that the putative mutation was also in the control sample, and thus a germline SNP. It is also worth noting that all of the 21 putative shared SNVs were low confidence SNV calls that are known to have a low validation rate. Thus, our detection of zero high confidence SNVs that are shared among multiple neurons from the same mouse is by itself strong evidence that early-arising clonal mutations are extremely rare. For the 13 shared SV calls and 3 shared MEI calls, all failed validation either because the mutation was discovered in one or more of the control samples, or because we failed twice to successfully make usable primers (Table S4). Overall, we identified no bona-fide shared mutations either among MCNT-ES cell lines within a single mouse, or within different mice.

Analysis of predicted functional consequences of somatic mutations

We first determined how many of the mutations have coding effects. For SNVs and Indels, we used SnpEff (Cingolani et al., 2012) version 3.1m and filter for effects in codons. For SVs and MEIs we determined the coding effects using a combination of feature intersection (bedtools intersect) with RefSeq exome, as well as visual inspection. We identified five SNVs, one indel, and four SVs that disrupt exons in 11 different genes with various levels of predicted severity. Four of the genes involved are highly expressed in MT neurons as determined by our RNA-Seq data (Figure S3). Many of the remaining tests focus on our high confidence SNV calls, as they are the most numerous and have been identified with high accuracy.

The number of regulatory features containing indels or SNVs was determined using the Ensembl Regulatory Build (Zerbino et al., 2015) which uses data from ENCODE. Data was accessed using the Variant Effect Predictor tool on the Ensembl website (http://Jul2015.archive.ensembl.org/info/docs/tools/vep/index.html?redirect=no). The only available brain related track at the time of accession (August 2015) was for embryonic brain (e14.5), therefore the numbers presented are likely an under representation of the number of regulatory features relevant to the development and function of MT neurons. Analysis was performed using default settings on all high confidence SNVs and indels and all validated low confidence SNVs and indels. The regulatory features containing SNVs and indels were; CTCF binding sites, enhancers, open chromatin regions, promoters, promoter flanking regions, and splice regions.

The number of transcripts containing mutations within non-exonic regions was calculated using the SnpEff output (see above). Each SNV:transcript or indel:transcript pair was considered a unique count within the table, so that mutations that fell within more than one gene were counted twice. The exception to this was when multiple isoforms of the same gene were assigned different RefSeq entries, as for *Pcdha* isoforms and for *Iqcj-Schip1* and *Schip1*. In these cases, multiple isoforms were condensed to a single transcript.

Determination of mutational burden

The total number of mutations for cell types other than MT neurons were taken from the literature (Behjati et al., 2014; Kong et al., 2012; Young et al., 2012). For each cell type, the total number of mutations per cell was divided by the total number of megabases in the diploid mouse or human genome (5,600 and 6,600 megabases respectively), except for human germ cell values, which were divided by the number of megabases in the haploid human genome (3,300). For MEF and TTF values, the number of mutations found in each iPSC clone was taken to reflect the mutational burden in the original MEF or TTF prior to reprogramming. For endodermal cell types, only somatic mutations unique to each organoid were included in per cell estimates.

The mean and SEM are plotted in Figure 3D. For TTF, n = 3, for prostate, n = 4, oocytes and sperm, n = 5, for MT neurons and stomach, n = 6, for large intestine and MEF, n = 7, and for small intestine, n = 8. To determine whether the MT neuron mutational burden varied significantly from the mutational burden in other cell types, we performed a 1 way parametric ANOVA testing the null hypothesis that there is no difference in the population average mutational burden between any of the cell types described above. We found the means were significantly different ($p < 0.0001$), and thus reject the null hypothesis. To determine which cell type mutational burdens differed from the MT neuron burden, we performed a post-hoc Dunnett's multiple comparison test. The results are summarized in Figure 3D. We note however, that when a simple unpaired one-tailed t-test is performed analyzing the difference between MT neuron and oocyte mutational burden, the means are found to be significantly different, with a p value of 0.0009.

SNV base conversion profiles

We compared our SNV base conversion profiles to those reported in other studies (Behjati et al., 2014; Holstege et al., 2014; Kong et al., 2012; Young et al., 2012) by strand normalizing the base conversion and counting the number of mutations in each of the 6 possible categories (Figure S4). As is common, we have more C→T conversions than any other base conversions.

We also compared the 3-base context of somatic SNVs and germline SNPs as a possible indicator of mutational process. The germline SNPs were determined for each mouse separately by using the same criteria used to identify somatic SNVs except that all MCNT-ES cell lines and the control sample from the same mouse were all used as sample lines, and no parent or other lines were used. As almost all of the calls occur in all mice, the final germline callset was determined by taking the union of the calls in each mouse. The strand corrected 3-base contexts were identified using bedtools getfasta. We find that the somatic SNVs in MT neurons are enriched in C→T conversions taking place in TpCpN contexts, as compared to germline SNPs using Fisher's Exact Test ($P<0.0001$) (Figure 3E).

Enrichment in genomic features

We next sought to determine if our somatic SNV calls occur randomly throughout the genome, or instead co-locate more or less frequently than chance in certain genomic features. We restricted this study to autosomes to eliminate any issues with the fact that we have both male and female mice in this study. We chose to use nine genomic features that were broadly diverse, of potential functional interest, and that were common enough to have more than 5 somatic SNVs fall within them. The chosen features were 100mer-uniquely-mappable regions, 100mer-unmappable regions, segmental duplications, elements conserved across placental mammals, simple repeats, LINEs, RefSeq exons, RefSeq transcripts, and RefSeq transcripts that are highly expressed in MT neurons. Almost all of these feature tracks were either directly downloaded from the UCSC table browser (http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database/) or readily derived from such a download. See Table S9 for details.

We then calculated SNV enrichment relative to chance using two different strategies. First, we separately ran a 10,000 trial Monte Carlo simulation for each genomic feature. The 395 autosomal high confidence SNVs were distributed randomly throughout the autosomes using bedtools shuffle while excluding all reference genome assembly gaps and regions in which the read-depth in any sample was less than 10 or greater than 250. These latter regions were excluded from the simulations as they were also excluded by our SNV calling strategy. From the simulations, we captured the mean and standard deviation of the number of SNVs that fell in the feature, and the number of trials in which the number of SNVs in the feature was greater than or less than the actual count of the number of somatic SNV calls that fell in the feature. This latter provides an estimate of the p-value of an enrichment or depletion of our SNVs vs. random chance. Second, we calculated the expected value of the number of SNVs that should randomly fall in each feature based on the length of the feature vs. the accessible genome. We then also derived a p-value for the likelihood that our SNV count in the feature falls within the 95% confidence interval given the feature length using the Poisson Test. The simulation means and the expected values based on feature length are in very close agreement. All results are summarized in Table S9.

The mappable and unmappable regions were chosen as controls, as it is easier to correctly align reads and make mutation calls within unique regions of the genome. As expected, our SNV calls are significantly enriched in the mappable regions and depleted in unmappable regions. Also, our SNV calls are significantly depleted in segmental duplications. We estimate that at least half of this effect is due to the fact that only 38% of the segmental duplicates track falls within mappable regions, and 7 of our 9 SNV calls fall within that 38%. Similarly, our SNV calls are mildly depleted in simple repeats (not statistically significant). Interestingly, our MT neuron SNVs are significantly enriched in elements conserved across placental mammals where one might expect the opposite due to cellular selection pressure (Figure 3F). Similarly, our SNVs are enriched in the RefSeq exome, transcripts, and transcripts highly expressed in MT neurons, but these enrichments do not quite reach statistical significance. However, we find this suggestive, and note that our statistical power to make such discriminations is limited by the relatively small number of SNVs found in this study.

Comparison of MT neuron and endodermal SNVs accumulation in genes.

To further explore the enrichment of MT neuron SNVs in genes, we compared our SNVs to those found in a recent study of clonal organoids formed from mouse prostate, stomach, small intestine and bowel (Behjati et al., 2014). We find that the MT neuron SNVs are enriched in genes compared to endodermal SNVs (P=0.0039, Fisher's Exact, Figure 3G).

RNA-Seq and analysis

MT neurons were dissociated from *Pcdh21*/Cre-Ai9 mice as for nuclear transfer and flow sorted using the MoFlo® Astrios™ (Beckman Coulter). Ten minutes prior to sorting, DAPI (1 μM) and DRAQ5 (BioStatus DR50050, 1 μM) were added to the cell suspension. Dead cells and debris were first gated out using side and forward scatter. Objects were identified as cells by positive staining for DRAQ5, and as live cells by the absence of DAPI staining. From this population, MT neurons were identified by tdTomato expression, and sorted directly into TRIzol® LS Reagent (Life Technologies). The following lasers were used: DRAQ5 (642nm laser), DAPI

(405nm laser), tdTomato (561nm laser). Three biological replicates were collected on independent days using this method.

Prior to RNA extraction, all samples were adjusted to 1.75 mL TRIzol® LS and 1 ug of linear acrylamide (Ambion AM9520) was added. RNA was extracted using Direct-zol™ RNA MiniPrep (Zymo Research) using their Zymo-Spin™ IC columns for low amounts of RNA. The optional in-column DNAse treatment was included. RNA was eluted in 10ul of water and RNA quality was assessed using Agilent RNA 6000 Pico Kit. All RNA samples had RIN scores >7.5.

Prior to sequencing, 10ng of RNA from each biological replicate was amplified using SMARTer® Ultra™ Low Input RNA for Illumina® Sequencing – HV (Clontech Laboratories, Inc.). Amplified cDNA was checked for quality using High Sensitivity DNA Kit (Agilent Technologies) and acoustically sheared using the Covaris system. Sequencing libraries were prepped from sheared cDNA using NEBNext® Ultra™ DNA Library Prep Kit for Illumina® and sequenced on an Illumina HiSeq.

We analyzed the RNA-Seq data using TopHat (Trapnell et al., 2009) v2.0.10 (http://tophat.cbcb.umd.edu) and Cufflinks(Trapnell et al.) v2.0.2 (http://cufflinks.cbcb.umd.edu) from the Tuxedo suite. We first created the genome and annotation indexes by downloading the mm9 annotation data (mm9/Mus_musculus_UCSC_mm9.tar.gz) from (ftp://igenome:G3nom3s4u@ussd-ftp.illumina.com) and using gtf_to_fasta. Each of the three MT neuron samples were then aligned separately using bowtie through the tophat interface (-r 160 –libarary-type fr-unstranded –coverage-serach –b2-sensitive). The BAM files for the reads from the three MT neuron samples were then merged using samtools merge. We then assembled the reads and determined expression levels for the combined MT neuron samples using cufflinks (--library-type fr-unstranded --multi-read-correct –max-intron-length 500000). The resulting "genes.fpkm_tracking" file was converted to bed format for further processing. Finally, we considered those genes with greater than the median expression level of ~0.78 to be "highly expressed".

Three RNA-Seq datasets from Lgr5 positive small intestine stem cells (Sheaffer et al., 2014) with accession ids ERX421326, ERX421327 and ERX421329 were downloaded from (http://www.ncbi.nlm.nih.gov/sra/) in SRA format. From these files, fastq files of the RNA-Seq reads were extracted using fastq-dump v2.1.18 (--gzip) from the SRA Toolkit (http://www.ncbi.nlm.nih.gov/Traces/sra/). The reads were processed as above, with a resulting median expression level of ~0.69. Again, we considered those genes with greater than the median expression level of ~0.78 to be "highly expressed".

Comparison of MT neuron and endodermal SNVs accumulation in highly expressed genes. Using the RNA-Seq data described above, we explored how SNVs distribute within highly expressed genes. We first compared how MT neuron and endodermal SNVs distribute relative to one another using Fisher's Exact test. We found that MT neuron SNVs are enriched in genes highly expressed in MT neurons (P=0.025, Fisher's Exact, Figure 3H) when compared to those found in the mouse organoid study. We further found that the SNVs from small intestine organoids are depleted in genes highly expressed in Lgr5 positive small intestine stem cells relative to MT neuron SNVs ($P=7.06 \times 10^{-4}$, Fisher's Exact, Figure 3I).

We also compared MT neuron and small intestine SNVs to chance, within their respective transcriptomes using a one sided Poisson test and the null hypothesis that SNVs are not depleted relative to chance. We defined chance as the percent of the transcriptome length that falls within highly expressed genes and compared this to the percent of SNVs that fall within highly expressed genes. For MT neurons, we found 69.5% of MT neuron SNVs fall within genes highly expressed in MT neurons, versus 66.1% of the transcriptome length (p = 0.85). For small intestine stem cells, we found that 34.9% of small intestine SNVs fall within genes highly expressed in small intestine stem cells, versus 50.2% of the transcriptome length ($p = < 2.2 \times 10^{-16}$).

Estimation of the per cell division mutation rate in MT neuron progenitors.
First, we estimated the number of cell divisions that might have occurred during the developmental window covered by our mutation detection strategy. While the precise number of cell divisions that precede cell cycle exit has not been reported, MT neuron production peaks around embryonic day 11.5 (e11.5) (Imamura et al., 2011), a time at which MT precursors would be predicted to have undergone ~22 cell divisions (Imamura et al., 2011). However, in our mutation detection strategy, we required that putative neuronal mutations be absent from thymus or spleen, which segregate from ectoderm at gastrulation (e6.5) or slightly before (e4.5). A conservative estimate of the number of divisions between e4.5 and e11.5 is 14 (~2 per day). Next, we divided the highest and lowest per neuron estimate of SNVs found within MT neuron genomes (142 SNVs in line C5 and 62 SNVs in line B2) by the estimated number of cell divisions. This resulted in the reported range of ~4-10 SNVs per cell division.

TSRI Institutional Animal Care and Use Committee approved all animal procedures.

SUPPLEMENTAL REFERENCES

Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome biology *12*, R18.

Broad. Picard Tools (http://picard.sourceforge.net).

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

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research *110*, 462-467.

Kang, H., and Roh, S. (2011). Extended Exposure to Trichostatin A after Activation Alters the Expression of Genes Important for Early Development in Nuclear Transfer Murine Embryos. Journal of Veterinary Medical Science *73*, 623-631.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Malhotra, A., Lindberg, M., Faust, G.G., Leibowitz, M.L., Clark, R.A., Layer, R.M., Quinlan, A.R., and Hall, I.M. (2013). Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. Genome research *23*, 762-776.

Meissner, A., Eminli, S., and Jaenisch, R. (2009). Derivation and Manipulation of Murine Embryonic Stem Cells. In Stem Cells in Regenerative Medicine, J. Audet, and W. Stanford, eds. (Humana Press), pp. 3-19.

Nakayama M.D, T., Fujiwara M.D, H., Tastumi M.D, K., Fujita M.D, K., Higuchi M.D, T., and Mori M.D, T. (1998). A New Assisted Hatching Technique Using a Piezo-Micromanipulator. Fertility and Sterility *69*, 784-788.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Smit, A., Hubley, R., and Green, P. (1996-2010). RepeatMasker Open-3.0.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., *et al.* (2002). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In Current Protocols in Bioinformatics (John Wiley & Sons, Inc.).

Zerbino, D., Wilder, S., Johnson, N., Juettemann, T., and Flicek, P. (2015). The Ensembl Regulatory Build. Genome biology *16*, 56.