DNA template strand sequencing of single cells maps genomic rearrangements at high resolution

Ester Falconer, Mark Hills, Ulrike Naumann, Steven S.S. Poon, Elizabeth A. Chavez, Ashley D. Sanders, Yongjun Zhao, Martin Hirst and Peter M. Lansdorp

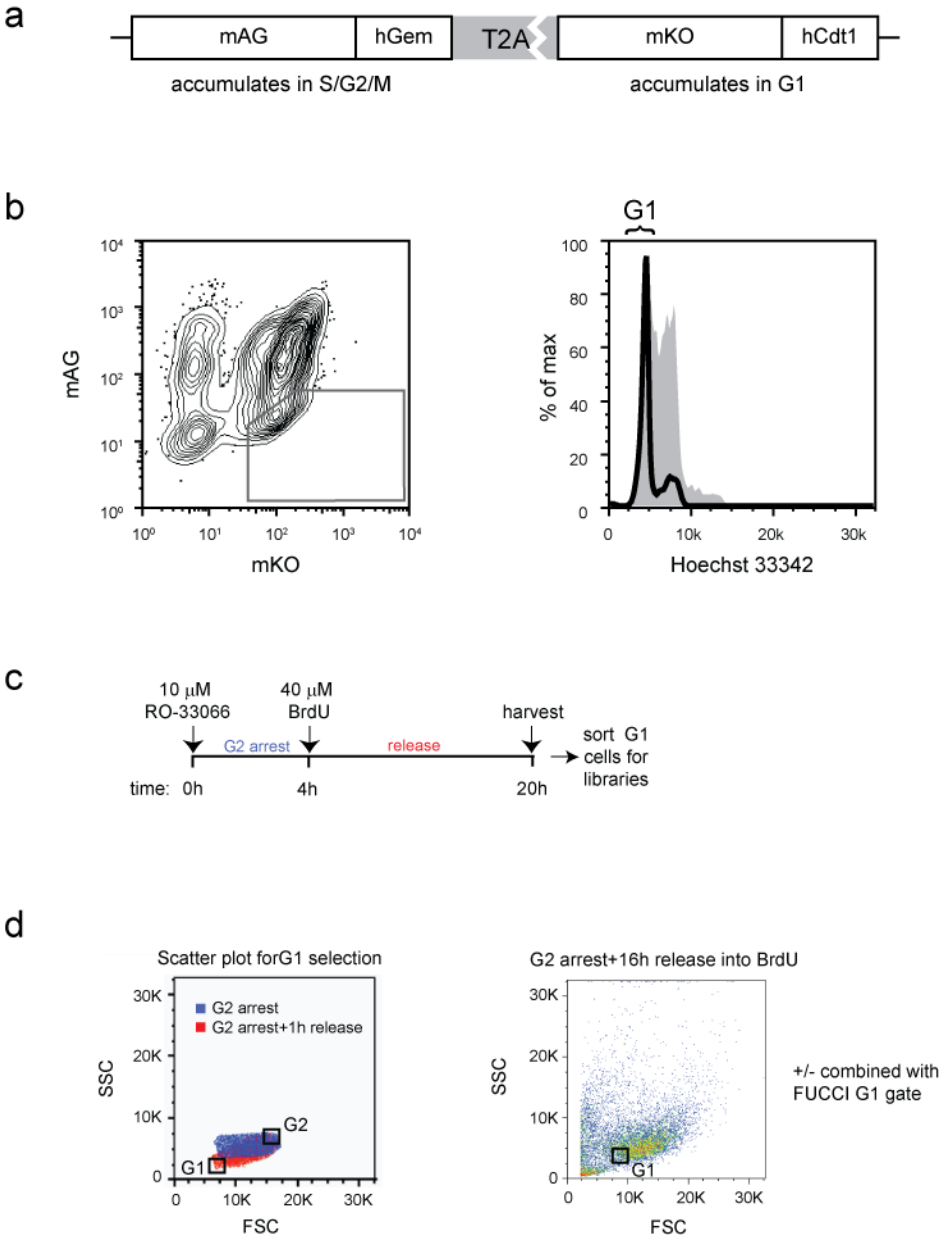
Supplementary Figures 1-14

Supplementary Tables 1-2

Note: Supplementary Data is available on the Nature Methods website.

Nature Methods: doi:10.1038/nmeth.2206

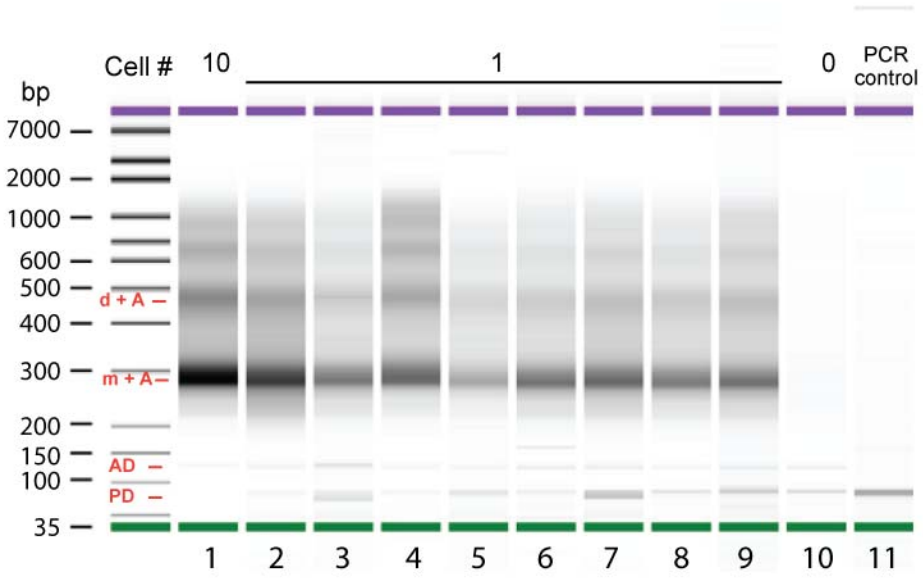
Supplementary Figure 1. Sorting strategy for isolating C2 murine embryonic stem (mES) cells in G1 of the cell cycle



Supplementary Figure 1. (**a**) Diagram of Fucci cell cycle reporter in which the green S/G2/M mAG-Geminin reporter is fused in frame with the orange G1 mKO-Cdt1 reporter via a T2A linker. Post-translational cleavage of the T2A peptide ensures equimolar expression of both fluorescent reporters from a single construct. (**b**) FACS profile of a cloned C2 ES cell line expressing the modified Fucci cell cycle reporter and stained with Hoechst 33342. Gating on mKO single positive cells as indicated (left panel) enriches for G1 cells with 2N DNA content (open histogram, right panel) as compared to the total cell population (filled gray histogram). (**c**) Some cells were arrested at G2 using the Cdk1 inhibitor RO-33066 and released into medium containing BrdU. (**d**) Cells were subsequently sorted at G1 on scatter profiles, Fucci cell cycle reporters, or a combination of both, which yielded G1 stage cells with 2N DNA content with hemi-substituted DNA strands as evidenced by Strand-Seq libraries (see **Supplementary Data**).

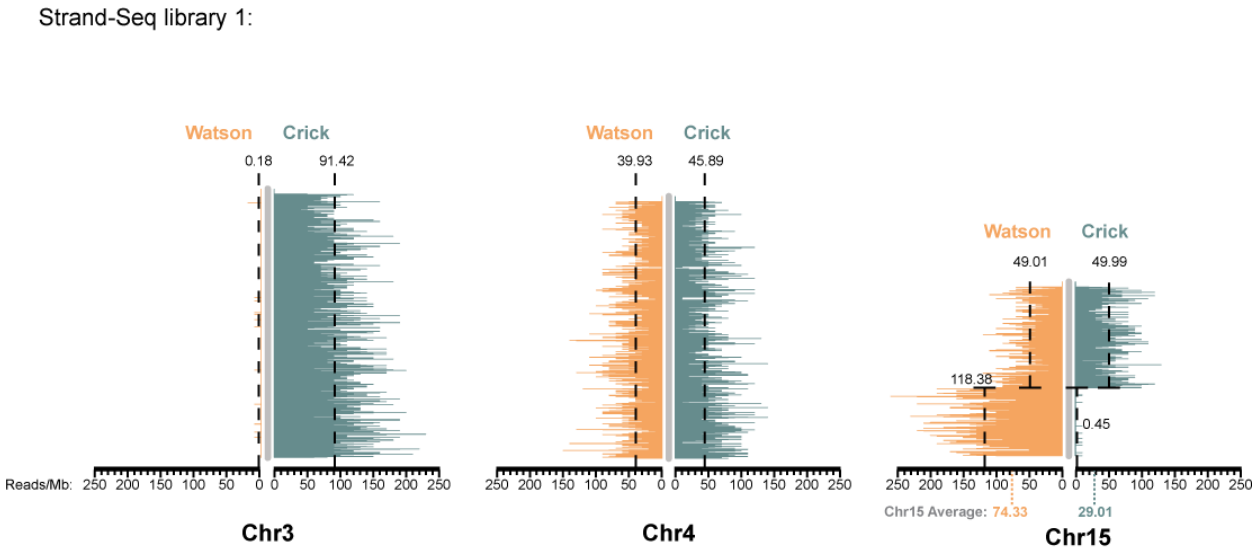
Nature Methods: doi:10.1038/nmeth.2206

Supplementary Figure 2. Size distribution of DNA in typical single cell Strand-Seq libraries.



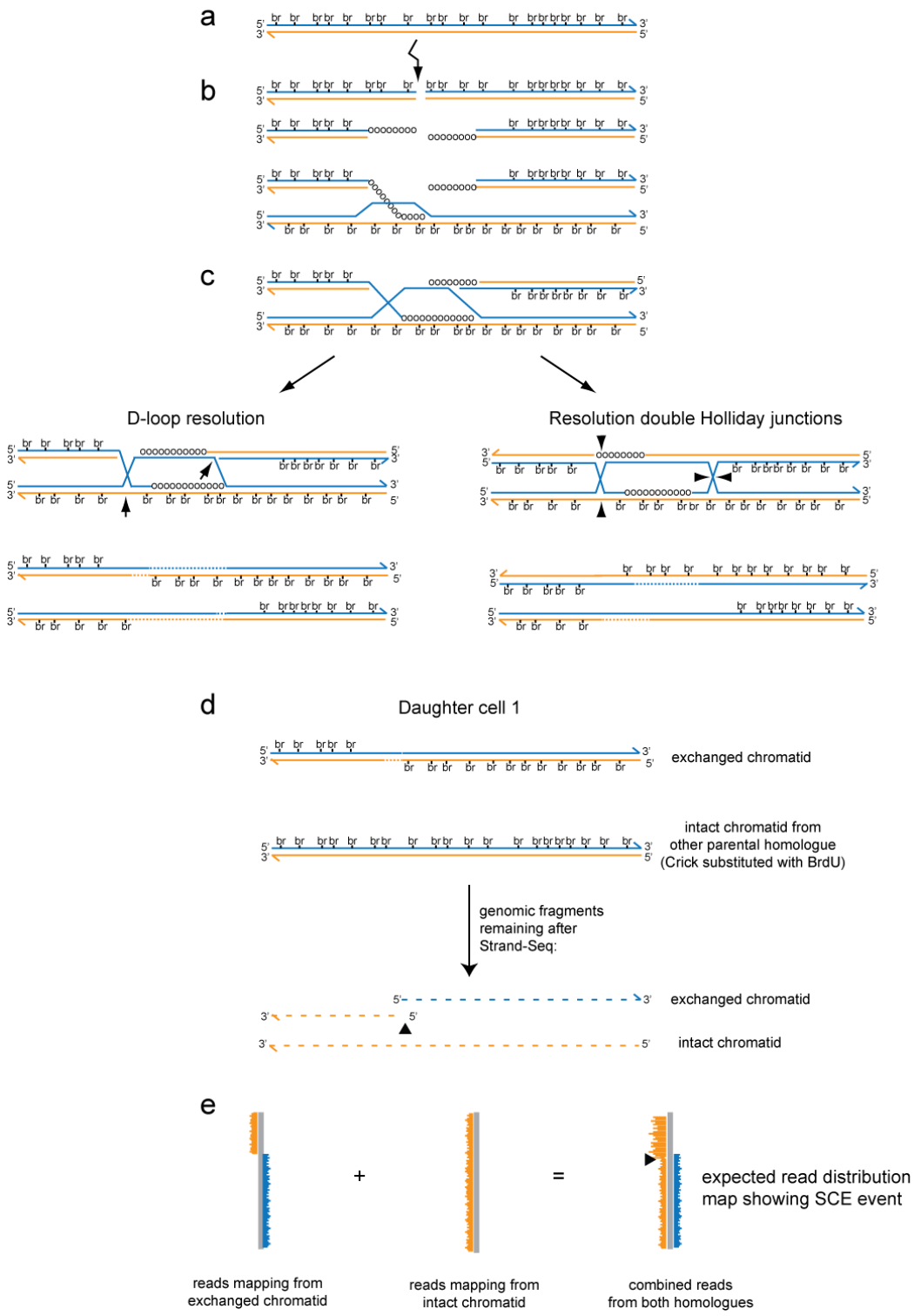
Supplementary Figure 2. 1 µl (9%) of total library yield was analyzed on an Agilent 2100 Bioanalyzer. Remarkably, libraries derived from single cells, without whole genome (pre-)amplification, show products of expected sizes after all the enzymatic and cleanup steps in the library construction protocol. The densitometry plot from an Agilent run shows that the predominant product in single cell libraries runs at ~ 270 bp (labeled m + A) corresponding to the expected size of mononucleosomes (m = 150 bp) plus adapter and primer sequences (A = 120 bp). Dinucleosomal library fragments (labeled d + A) also run at the expected size of 470 bp (corresponding to 2 × 150 bp nucleosomal DNA plus 50 bp linker DNA and 120 bp adapter and primer sequences) indicating that higher molecular weight library fragments are not chimeric ligation products of several mononucleosomal DNA fragments during the adapter ligation step. Single cell libraries (lanes 2-9) show a product that is easily observed whereas libraries made without input DNA (but having undergone all steps of library construction) show no product (lane 10). A final gel-based size selection of pooled libraries (data not shown) selects the mononucleosomal library fragments for paired-end sequencing, and removes the higher molecular weight library fragments as well as the remaining adapter dimers (AD) and primer dimers (PD).

Supplementary Figure 3. High resolution analysis of read depth demonstrates that reads are proportional to template strand inheritance.



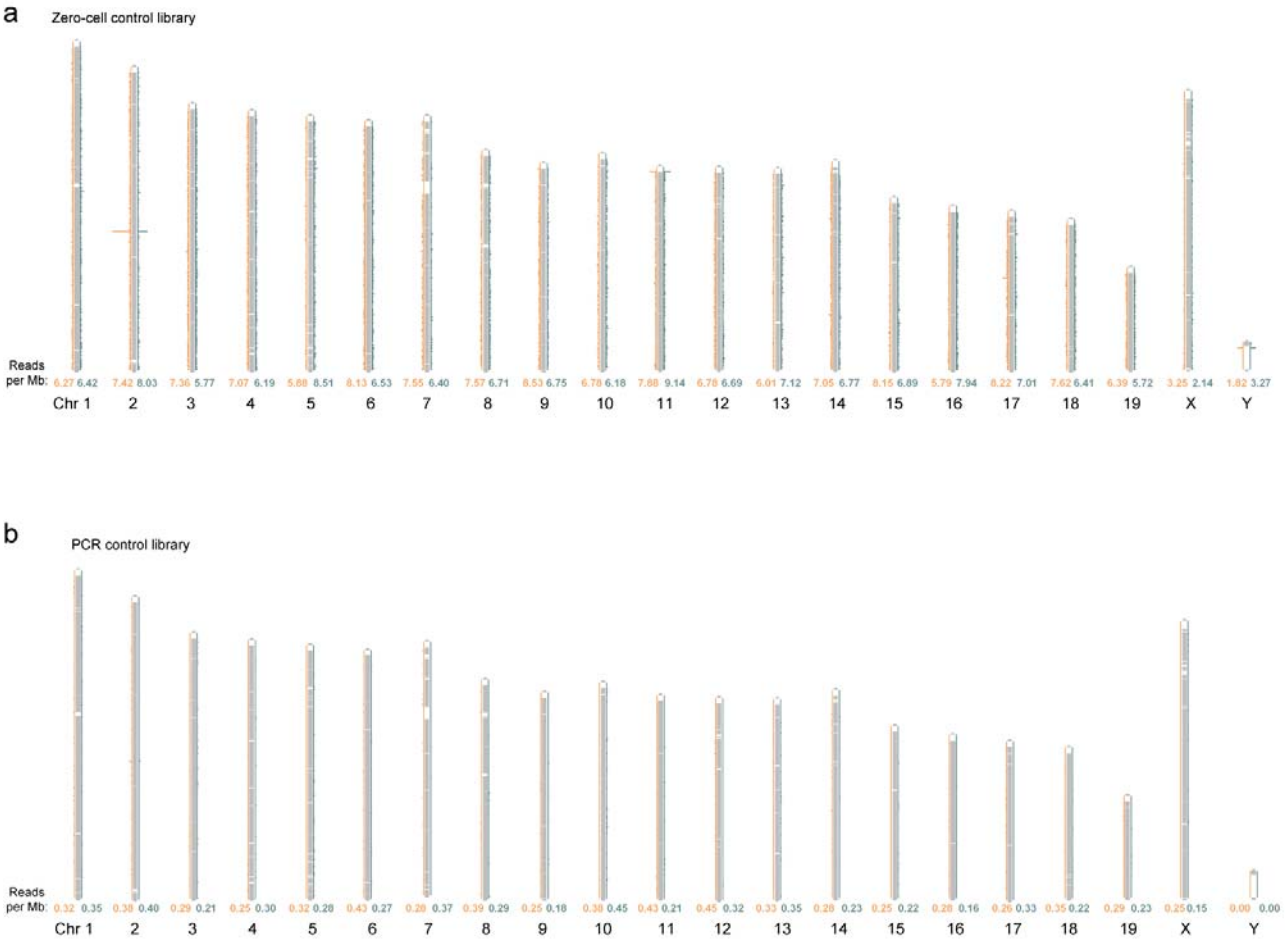
Supplementary Figure 3. Reads from chromosomes 3, 4 and 15 from Strand-Seq library 1 were enlarged to show the absolute number of reads per Mb (X-axis). The average read depth per Mb for both Watson (orange) and Crick (blue) reads is shown by the dashed vertical lines. While the sum of average reads/Mb is similar between these chromosomes (0.18 + 91.42 is roughly equal to 39.93 + 45.89 and 74.33 + 29.01), the read depth of Crick strand (blue) is approximately twice in chromosome 3 than in chromosome 4 (91.42 reads per Mb vs. 45.89 reads per Mb), indicating that this cell inherited two Crick template strands for chromosome 3 and only one for chromosome 4. Chromosome 15 shows a clear SCE event at 62.9 Mb, with the depth of reads changing at either side of the interval. The Watson reads roughly double depth after the SCE event (from 49.01 to 118.38 reads per Mb) while correspondingly the Crick read depth changes from 49.99 reads per Mb to close to zero. This indicates that the cell inherited both Watson template strands for the fragment between 62.9 Mb and the end of the chromosome. Note that the sum of the average reads per Mb for chromosome 15 is similar before and after the SCE event (49.01 + 49.99 is roughly equal to 118.38 + 0.45). These data reinforce the notion that Strand-Seq can accurately assign the inheritance of template strands in cells with minimal PCR bias. Moreover these findings support the use of read depth analysis in Strand-Seq libraries to identify gains and losses of chromosomes within single cells and to identify SCE events (see also **Supplementary Fig. 8**).

Supplementary Figure 4. Strand-Seq can be used to enumerate and map sister chromatid exchange events at high resolution.



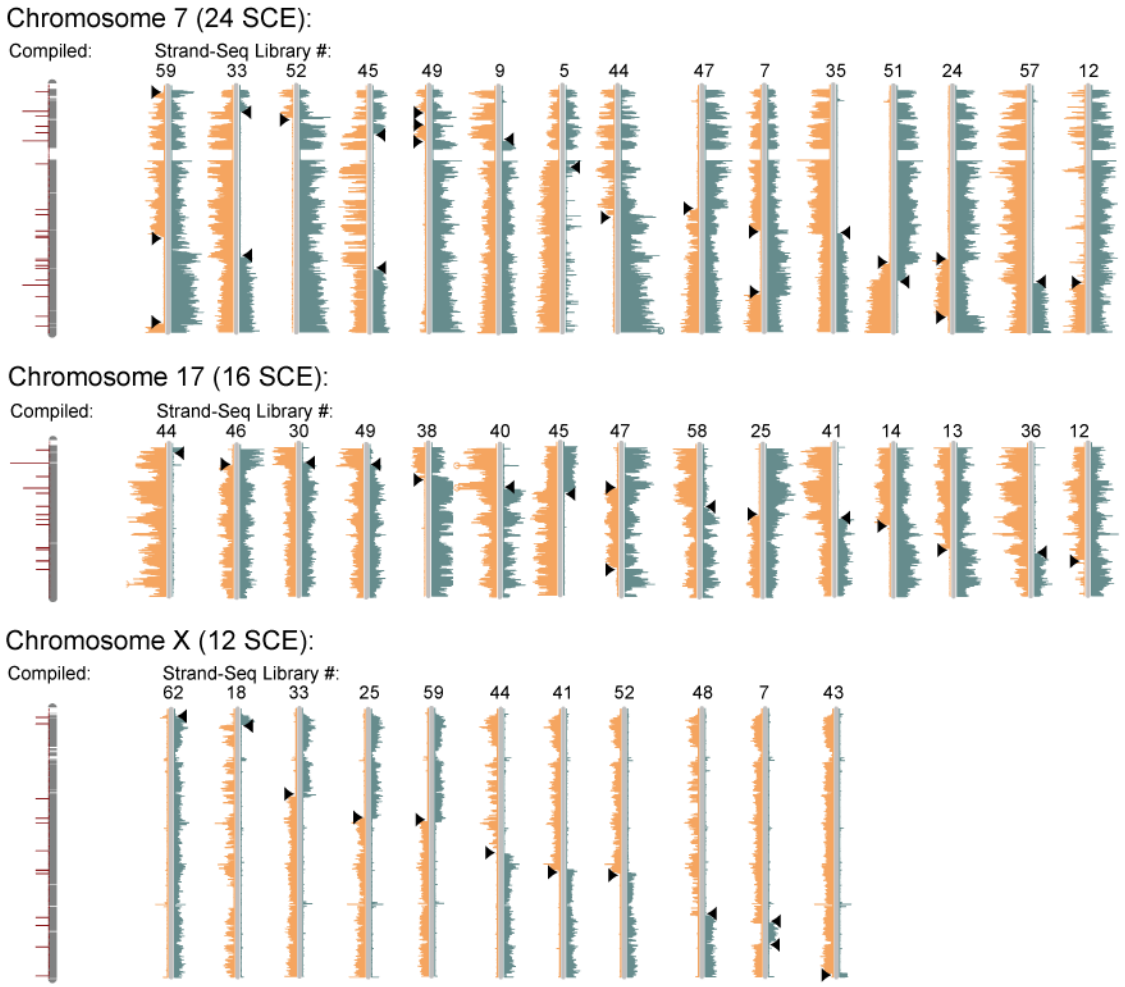
Supplementary Figure 4. (**a**) Example of a sister chromatid with the Crick (non-template strand, blue) substituted with BrdU (br). (**b**) Repair of a double strand break (arrow) is assumed to involve resection of the 5' ends at the site of the break, coating of 3' single strand fragments with Rad51 (circles) and invasion of the Rad51 filament into the unbroken sister chromatid, forming a D-loop at homologous sequences. (**c**) The second (homologous) strand is captured by the other Rad51-coated filament. A sister chromatid exchange results when D-loop resolution involves recombination (left panel), or an event similar to the resolution of double Holliday junctions with recombination between sister chromatids (right panel). (**d**) Upon mitosis each daughter cell will inherit one sister chromatid from each parental homologue: one with exchanged template strands and one intact chromatid from the other parental homologue (only one possible combination is shown for clarity). Strand-Seq library construction will remove all BrdU-substituted genomic fragments, leaving only the original unsubstituted template strands. (**e**) The expected read distribution ideogram clearly shows the location of the SCE event, which can be mapped at high resolution (see **Fig. 2b**).

Supplementary Figure 5. Control libraries.



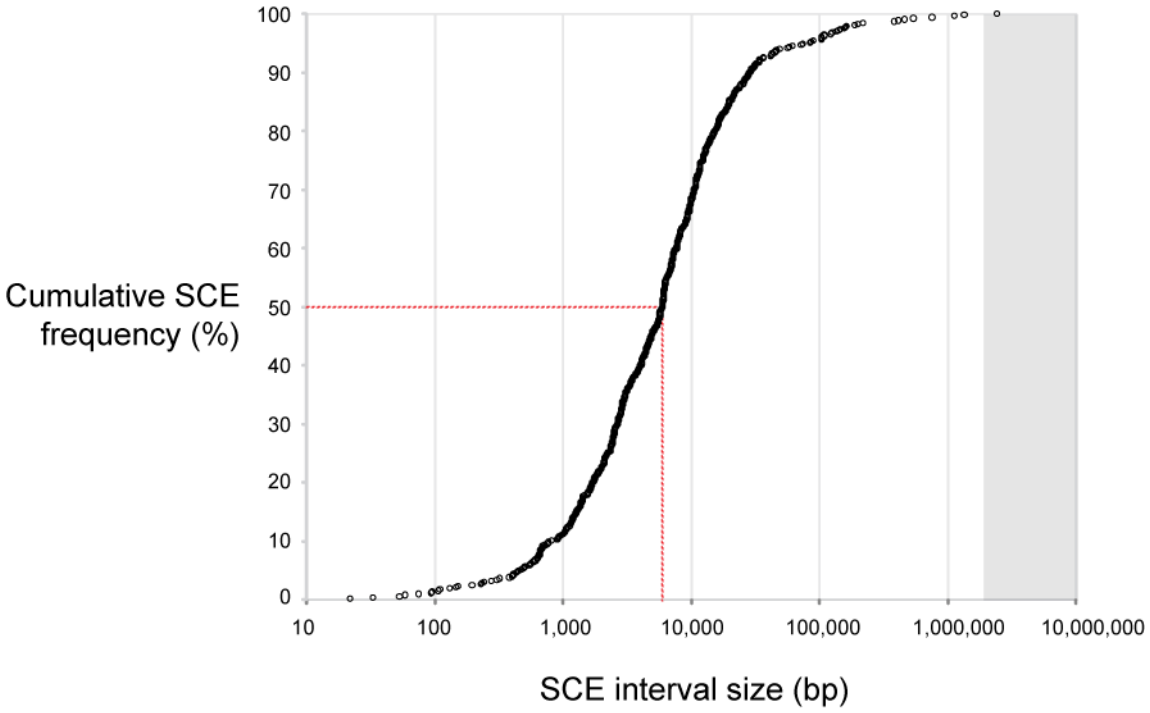
Supplementary Figure 5. Libraries from zero cell controls (**a**) and PCR controls (**b**) were sequenced on the Illumina HiSeq2000 sequencing platform. The very low background reads we observed indicate that the reads from our single cell libraries were not due to contamination. Compared to these controls, the average read depth in our single cell libraries is at least two orders of magnitude higher (see **Supplementary Data**).

Supplementary Figure 6. Compilations of sister chromatid exchanges.



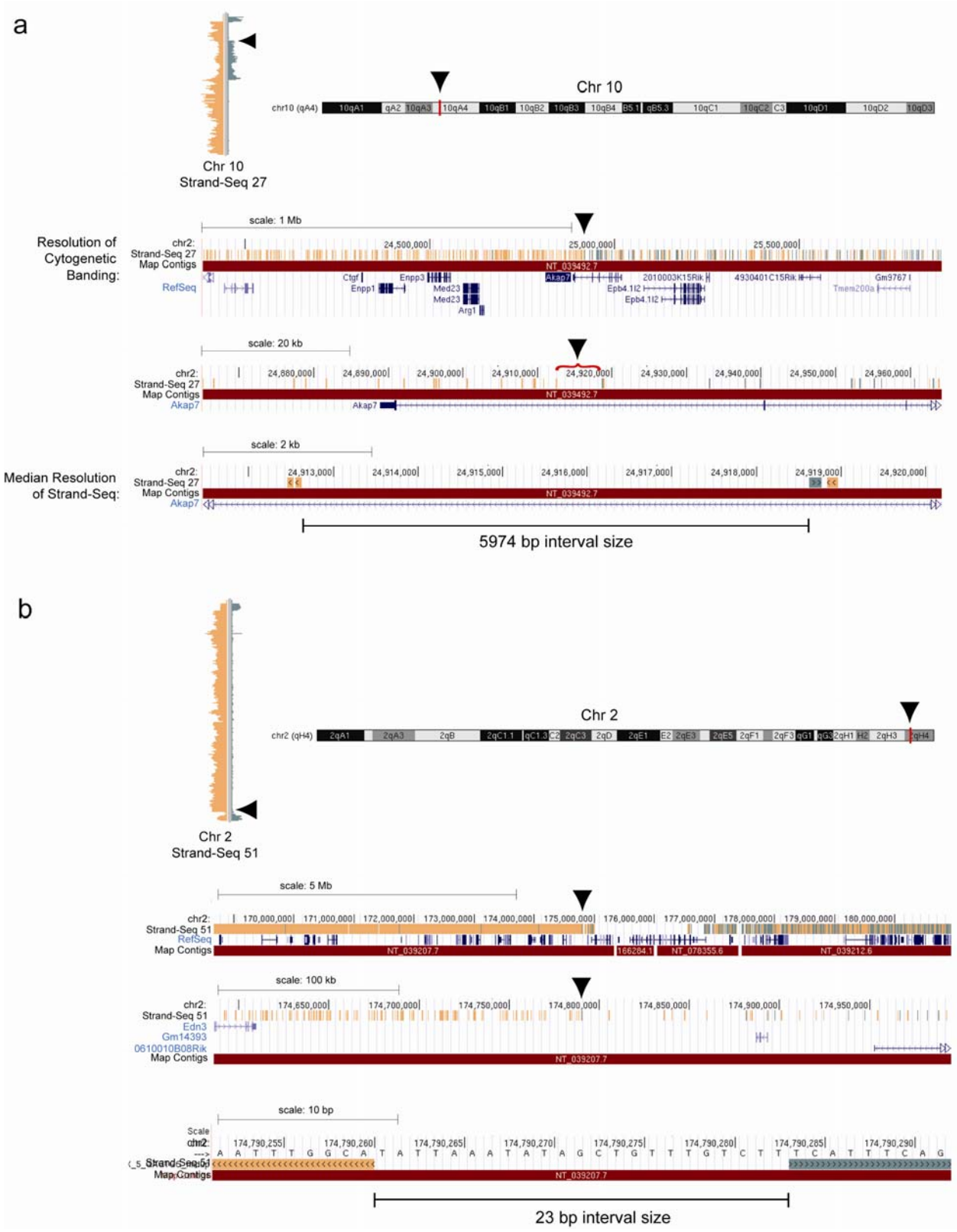
Supplementary Figure 6. The location of the SCE intervals in three different chromosomes from 62 Strand-Seq libraries (right) are combined in 1 Mb bins and mapped to ideograms of each chromosome (left, red lines). SCEs from the X chromosome appear as complete switches from Watson to Crick since there is only one copy of X in these male mES cells (see also **Supplementary Fig. 4**). The genomic location of the SCE for this analysis is determined by the first base pair of the interval between the Watson and Crick reads flanking the SCE. The actual genomic interval of the exchange event depends on the distance between the last read mapping to one template strand and the first read mapping to the opposite template strand, and is therefore defined by the depth of sequencing at that location. This depth depends on many factors: which fragments survived library construction protocols, which subset of fragments was sampled for sequencing from the total pooled libraries, as well as the sequencing platform and chemistry version used. It is also dependent on the gaps in the build of the reference genome and reads from repetitive regions, which map poorly. The genomic distance between the reads that map SCE events varies because the overall number of reads in each library depends on the sequencing platform used. However, there is sufficient coverage to assign the inherited template strand for each chromosome and map sister chromatid exchanges with much higher resolution than conventional methods.

Supplementary Figure 7. Cumulative distribution of SCE interval size.



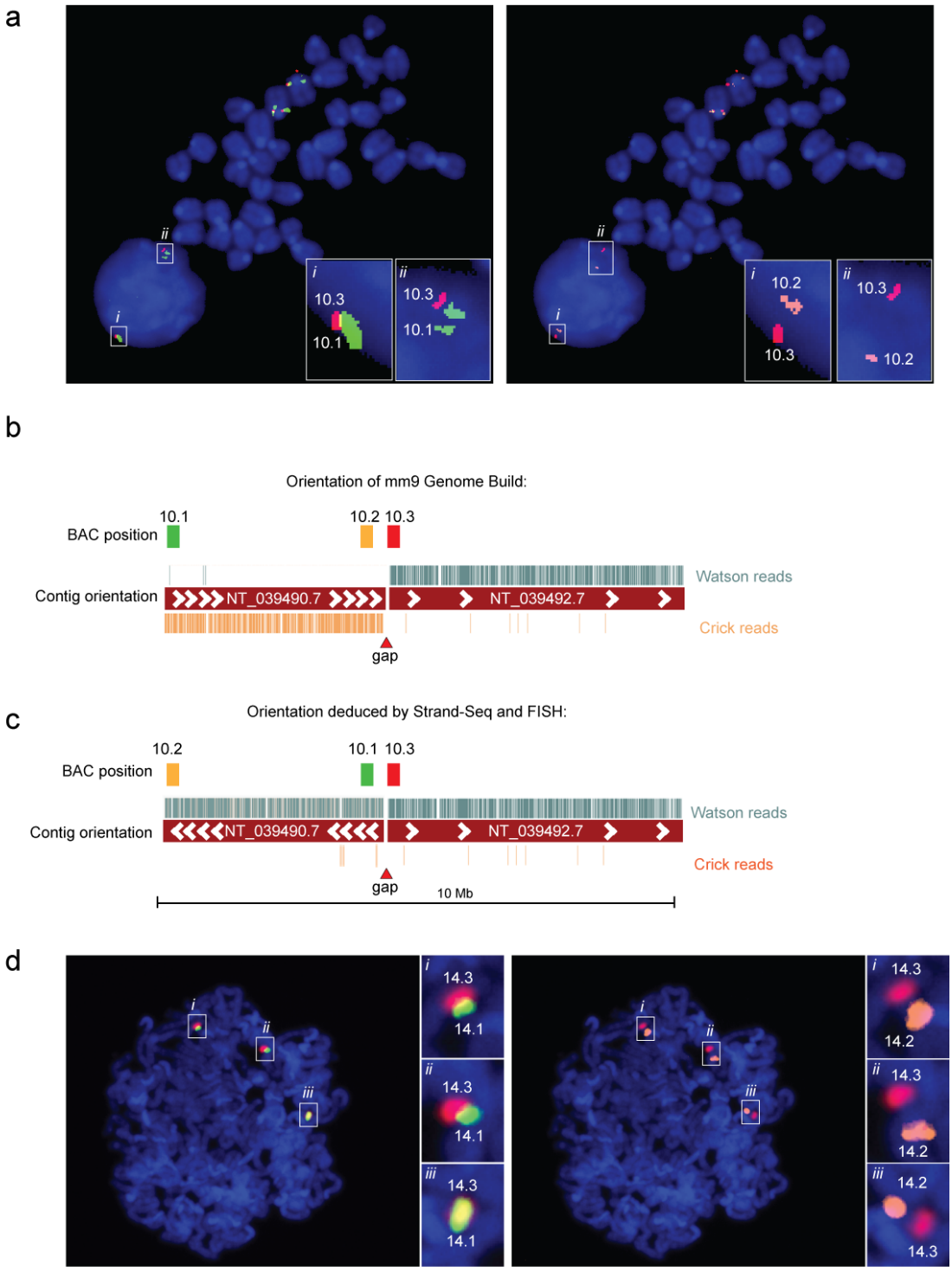
Supplementary Figure 7. The interval size of all 529 SCEs in 62 Strand-Seq libraries was calculated in bp, and plotted as a cumulative frequency distribution. The median interval size of 5.97 kb is shown (red dashed lines). Note that only one Strand-Seq library has an SCE interval range over 2 Mb, which is the typical resolution expected when using traditional cytogenetic techniques (gray shading).

Supplementary Figure 8. Strand-Seq allows very high resolution mapping of SCE intervals.



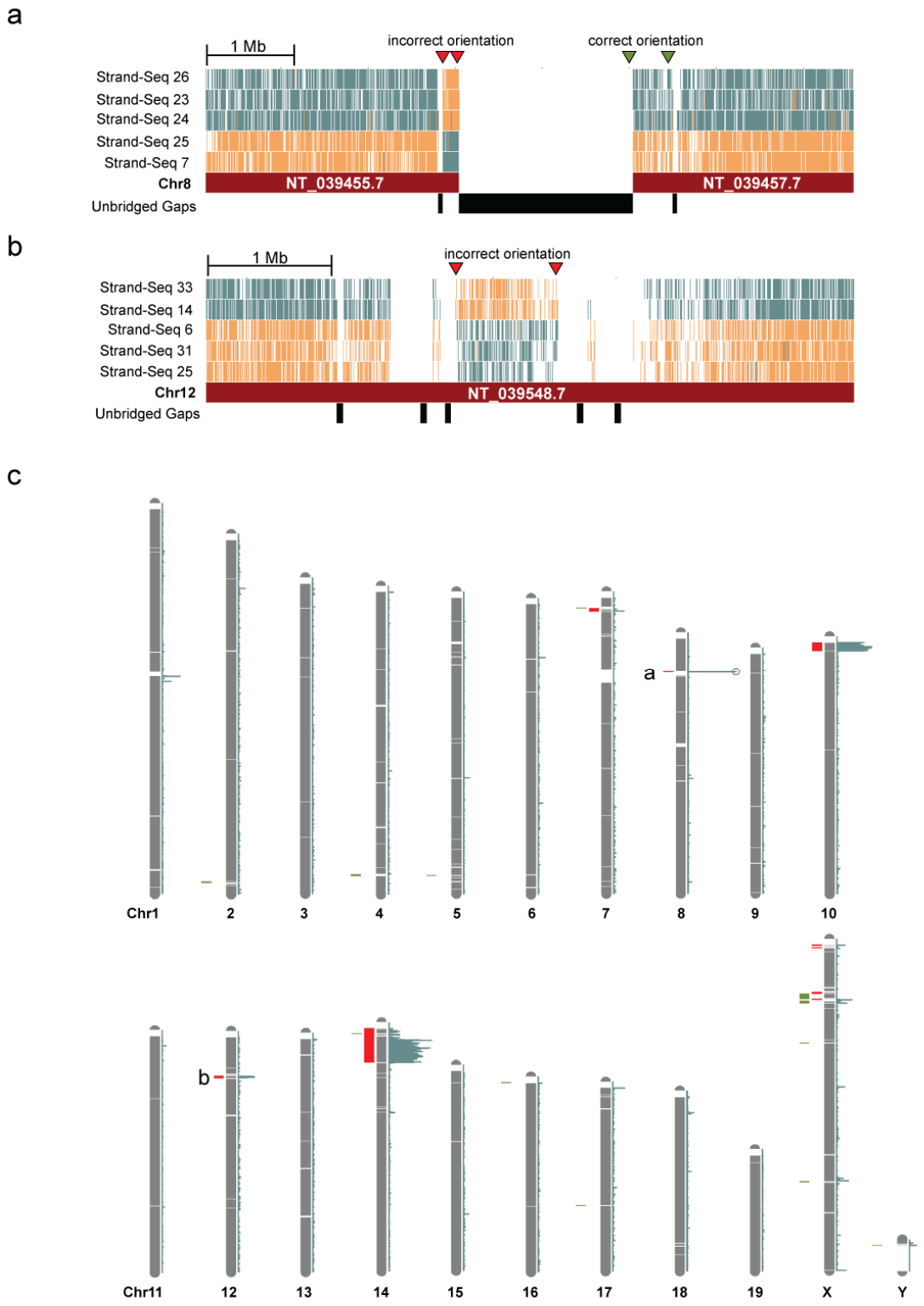
Supplementary Figure 8. (**a**) An example of an SCE interval corresponding to the median interval size shown in **Supplementary Fig. 7**. The ideogram of chromosome 10 from Strand-Seq library 27 (See **Supplementary Data**) is shown with screenshots from the UCSC genome browser at increasing levels of zoom. The top of the three zoom levels is a window encompassing 2 Mb of sequence, which is the resolution of current cytogenetic banding techniques. With Strand-Seq we can narrow the exchange interval size (black arrowheads) to 5974 bp (bottom panel), which falls within a large intron of the A-kinase anchor protein 7 gene (*Akap7*). This interval size represents the median interval size of all 527 SCEs mapped in this study. (**b**) The closest SCE interval we detected (23 bp) is mapped to chromosome 2 of Strand-Seq library 51. There are no genes detected within this interval.

Supplementary Figure 9. Strand-Seq can be used to correct ambiguities in genomic maps and help build better reference genomes.



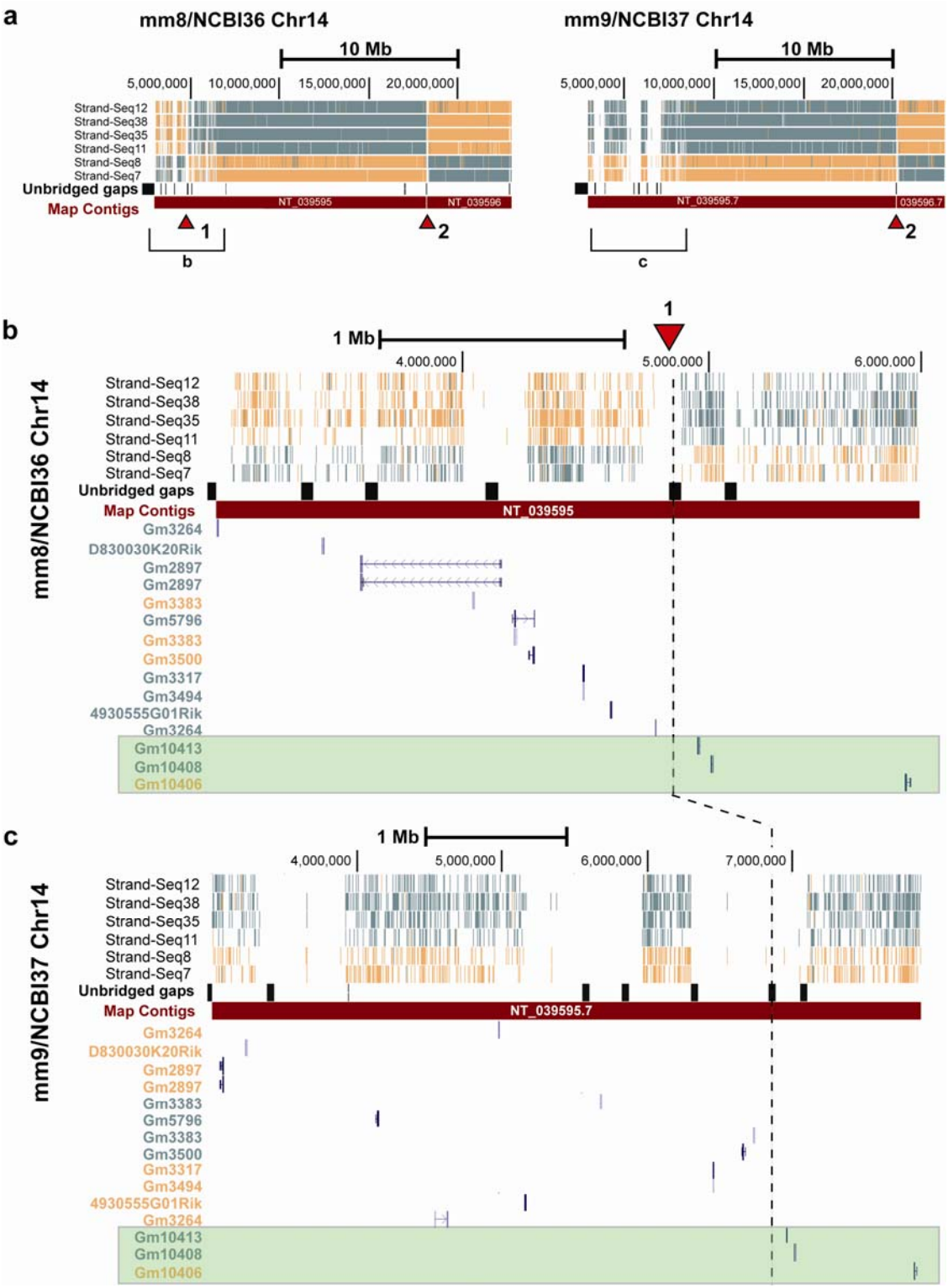
Supplementary Figure 9. Analysis of Strand-Seq libraries revealed a complete template switch in chromosome 10 in more than one cell, which mapped to a gap in the reference genome (see **Fig. 2**). (**a,b**) To test the hypothesis that these switches could be explained by an error of the orientation of contig NT_039490.7 in the reference genome, we performed fluorescence *in situ* hybridization of C2 murine embryonic stem cells (C57BL/6NTac background) simultaneously with 3 BAC probes: 10.1 (RP23-38N9 labeled with Spectrum-Green), 10.2 (RP23-128M21 labeled with Spectrum-Orange) and 10.3 (RP24-258P4 labeled with Spectrum red-dUTP). The expected genomic location of these probes relative to the gap in mm9 on chromosome 10 (red arrowhead) is shown in (**b)**. Probes 10.1 (green) and 10.3 (red) are in closer proximity to each other (**a**, left panel) than are 10.1 and 10.2 (**a**, right panel) indicating that the orientation of contig NT_039490.7 should be reversed in the reference genome (**c**). (**d**) Using the chromosome 14 probes as in **Fig. 3**, a murine 3T6 fibroblast cell line (ATCC, Swiss albino genetic background) shows identical results, indicating that our FISH results are independent of cell type or genetic background.

Supplementary Figure 10. Orientation of contigs and contig fragments in the mouse mm9 genome assembly as deduced by Strand-Seq.



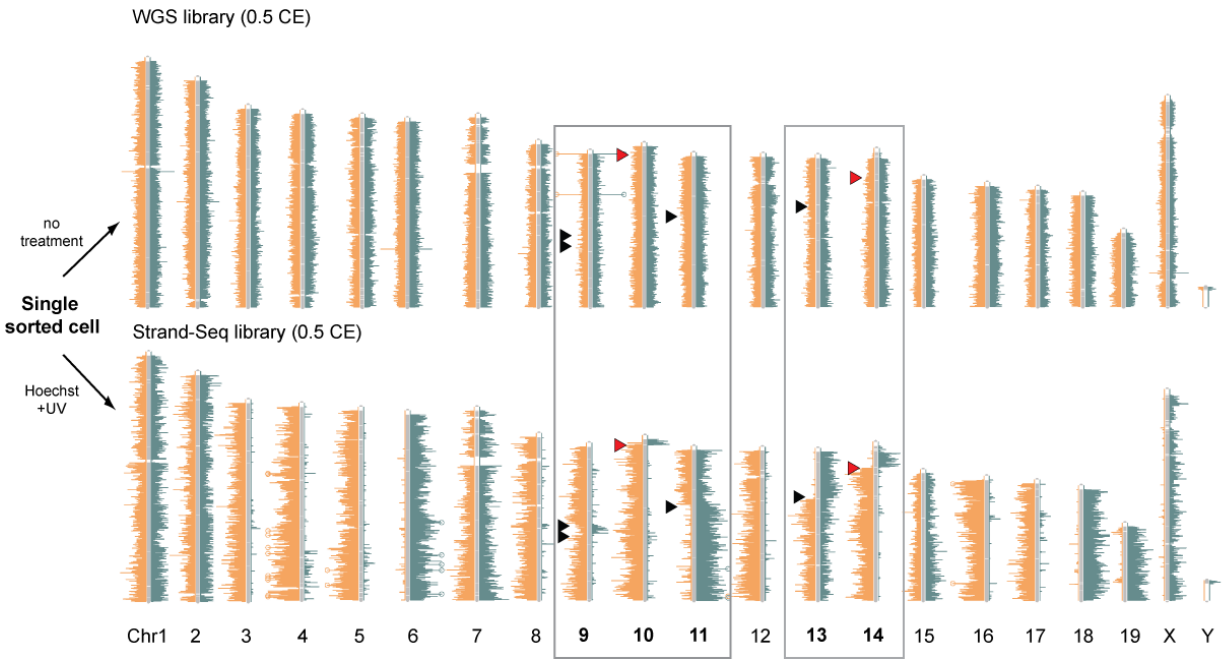
Supplementary Figure 10. The mm9 build of the mouse genome contains multiple gaps, and in many cases the orientation of adjacent contigs is unknown and results in regions of the genome which are mis-oriented (see **Fig. 3**). While chromosome 10 and chromosome 14 involve comparatively large mis-oriented regions that are obvious in Strand-Seq ideograms, it is important to consider the orientation of smaller fragments. (**a**) Example of an incorrectly oriented fragment on chromosome 8. A fragment at the end of contig NT_039455.7 shows a complete template switch in all 31 Strand-Seq libraries we can analyze (5 libraries shown). Note that an adjacent fragment on contig NT_039457.7 is in the correct orientation in all libraries we can analyze. (**b**) Mis-oriented fragment identified on chromosome 12 within a contig (NT_039548.7). A complete template switch can be seen in the central fragment in all 25 Strand-Seq libraries we can analyze (5 libraries shown), while the 5 additional fragments present at this resolution are all correctly oriented. (**c**) Visualisation of all mis-oriented contigs. The ideogram represents a composite of all chromosomes that inherited both Watson templates. For these chromosomes, all opposite-strand Crick reads are shown to the right of the ideograms. The pileup of these reads represents the incorrectly oriented fragments of the genome, which are shown as red bars to the left of the ideograms. The green bars show regions of unknown orientation (complete list in **Supplementary Table 2**). The mis-oriented regions span both larger regions on chromosomes 10 and 14 which were confirmed by FISH, as well as the smaller regions such as those shown for chromosomes 8 and 12 in (**a**) and (**b**), respectively. Note that the orientation of fragments can be determined only for the chromosomes that inherit either both Watson or both Crick templates and not those that inherit one Watson and one Crick strand.

Supplementary Figure 11. Strand-Seq identifies mis-oriented contigs in early genome builds that are corrected in later genome builds.



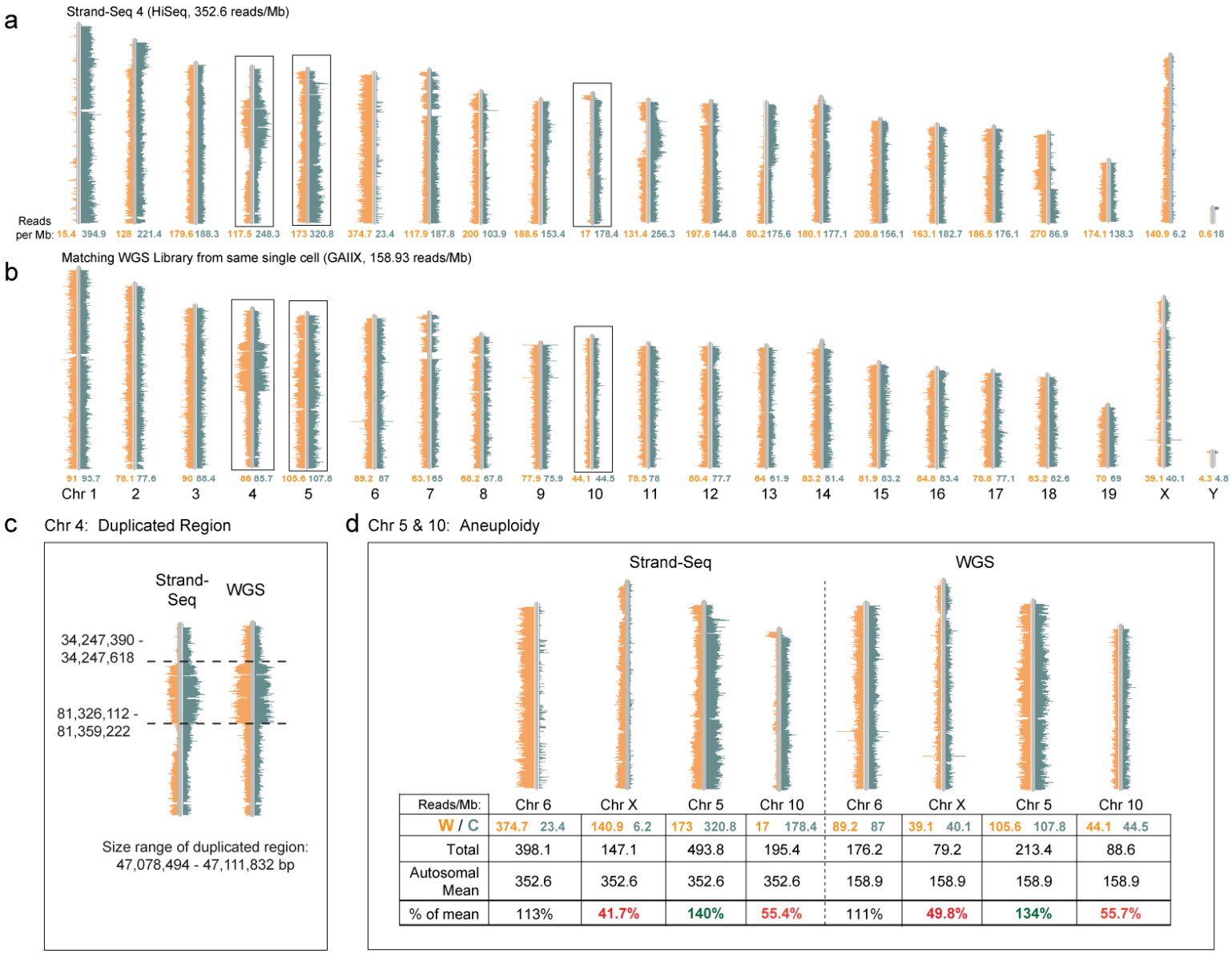
Supplementary Figure 11. In order to demonstrate the ability of Strand-Seq to correctly orient contig fragments, libraries were aligned to the 2006 mouse genome build mm8 (NCBI36) and compared to the current genome build (mm9/NCBI37). While comparison is impeded due to the changes in size and location of contigs between builds, we discovered 4 clear examples of regions that were mis-oriented as determined by Strand-Seq in mm8 and that had been corrected independently in mm9. (**a**) Example of mis-oriented contigs between genome builds. While the first 20 Mb of chromosome 14 is mis-oriented in both builds of the mouse genome (the switch region is denoted by the red arrowhead labeled 2), the first 5 Mb was also mis-oriented in mm8 but subsequently corrected in mm9 (switch region denoted by red arrowhead labeled 1). (**b**) Magnification of the first 6 Mb of Strand-Seq libraries aligned to mm8. Red arrowhead highlights the region that is mis-oriented with respect to the adjacent contig. UCSC RefSeq gene tracks are shown, with genes present on the + strand denoted in blue, and those on the – strand denoted in orange. Black dashed line serves as a reference point to anchor the same unbridged gap between genome builds. (**c**) Corresponding magnification of Strand-Seq libraries aligned to mm9. Note that all UCSC RefSeq genes in mm9 are listed in the reverse orientation with respect to mm8, and the gene order has been shuffled. Conversely, genes present outside of the mis-oriented region (green shading) are in the same orientation and order between builds.

Supplementary Figure 12. Strand-Seq reveals genomic rearrangements and errors in the reference genome otherwise masked by single cell libraries.



Supplementary Figure 12. Strand-Seq reveals genomic rearrangements and errors in the reference genome otherwise masked by single cell libraries. A library derived from a single sorted cell was divided in two half-cell equivalents (0.5 CE), prior to PCR amplification of adaptor-ligated fragments (see **Fig. 1**). One half of the library received no treatment prior to PCR (Whole Genome Shotgun (WGS) library, top panel) while the other half was treated with Hoechst and UV photolysis to maintain genomic directionality in the final amplified library (Strand-Seq library, bottom panel). Strand-Seq libraries, unlike WGS libraries, provide unique information including the inheritance of template strands. Most striking are the putative sister chromatid exchange events (black arrowheads) and errors in the reference genome (red arrowheads) which are otherwise masked in the untreated control WGS library (see main text).

Supplementary Figure 13. Single cell WGS and Strand-Seq libraries can identify numerical chromosomal abnormalities.

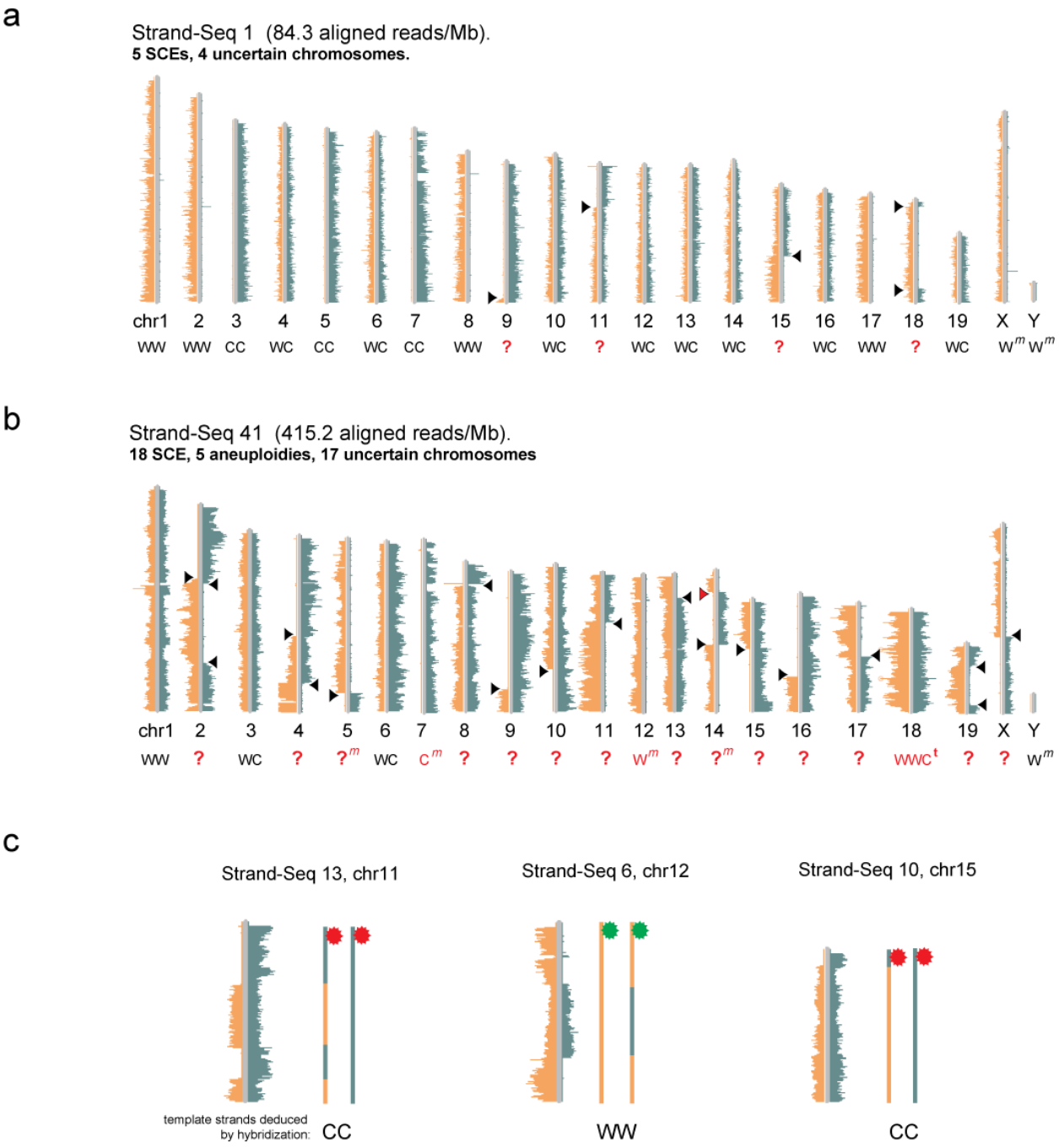


Reads/Mb:	Chr 6	Chr X	Chr 5	Chr 10	Chr 6	Chr X	Chr 5	Chr 10
W / C	374.7 23.4	140.9 6.2	173 320.8	17 178.4	89.2 87	39.1 40.1	105.6 107.8	44.1 44.5
Total	398.1	147.1	493.8	195.4	176.2	79.2	213.4	88.6
Autosomal Mean	352.6	352.6	352.6	352.6	158.9	158.9	158.9	158.9
% of mean	113%	41.7%	140%	55.4%	111%	49.8%	134%	55.7%

Nature Methods: doi:10.1038/nmeth.2206

Supplementary Figure 13. Genomic DNA from a single ES cell was split in half to make a Strand-Seq library (**a**) and a WGS library that was not treated with UV and therefore contains both template and newly-formed DNA (**b**). The average number of reads per Mb across each chromosome is indicated underneath the ideograms, with reads corresponding to Watson strands shaded orange, and Crick strands shaded blue. The average read number is similar across all chromosomes with the exception of a portion of chromosome 4, and the entire length of chromosomes 5 and 10 (boxed). (**c**) A 47 Mb region of chromosome 4 has roughly twice the average read depth as the rest of the chromosome, indicating a duplicated region. The Strand-Seq library shows an amplification of both Watson and Crick reads, suggesting that this portion of chromosome 4 was duplicated prior to BrdU incorporation. Note that while a genomic amplification can be identified and mapped at a high resolution, the location of the amplified chromosome 4 fragment cannot be deduced from analysis of this single cell. However, for heritable copy number variations in cells, including head to tail, head to head duplications and translocations between chromosomes, Strand-Seq libraries are expected to be informative. (**d**) Chromosome 5 contains roughly 50% more reads than the average diploid chromosome (represented by chromosome 6) indicating a trisomy, while chromosome 10 contains roughly 50% less reads, and is closer to the read depth of the single-copy X chromosome, indicating a monosomy. Note that both Strand-Seq and WGS libraries show numerical chromosome abnormalities, without the need for WGA to amplify input material. However, only Strand-Seq libraries allow the combining of this type of analysis with SCEs.

Supplementary Figure 14. Sister chromatid exchanges and aneuploidy confound template strand inheritance analysis in mES cells.



Supplementary Figure 14. (**a**) The cell in Strand-Seq library 1 contains five SCEs on four chromosomes, which must be excluded from template strand analysis since it is not possible to classify the inherited template strands as either Watson or Crick (red question marks). (**b**) Strand-Seq library 41 contains 18 SCEs and five aneuploid chromosomes (m, monosomy; t, trisomy) resulting in 17 out of 21 chromosomes for which we cannot assign Watson or Crick template strands. (**c**) Chromosome ideograms and template strands for several chromosomes which have undergone SCEs illustrate that SCEs must be taken into account when assaying for Watson or Crick template strand inheritance. Probe hybridization protocols (red and green pericentric probes shown) would assign WW or CC designations to these chromosomes, though the actual DNA strands inherited are a mixture of template strands and newly-formed DNA.

Supplementary Table 1. Identification of mis-oriented and repetitive regions in the mouse genome.

a. Location of mis-oriented regions in the mouse genome

Contig	Gap Type	Chr	Start	Finish	Size (kb)	Proportion of WW or CC strands showing mis-orientation
NT_039413.7	Fragment	7	8443310	8671340	**228.03**	15/15
NT_039413.7	Fragment	7	8721340	9452281	**730.94**	33/33
NT_039413.7	Fragment	7	9502281	10171903	**669.62**	31/31
NT_039455.7	Fragment	8	19853128	20038713	**185.59**	31/31
NT_039490.7	Contig	10	3000000	7246542	**4,246.54**	23/23
NT_039548.7	Fragment	12	22658993	23684469	**1025.48**	27/27
NT_039595.7	Fragment	14	3000000	3381821	**381.82**	26/26
NT_039595.7	Fragment	14	3431821	5557431	**2,125.61**	26/26
NT_039595.7	Fragment	14	5878008	6303768	**425.76**	25/25
NT_039595.7	Fragment	14	6353768	6842153	**488.39**	14/14*
NT_039595.7	Fragment	14	6892153	7058960	**166.81**	12/12*
NT_039595.7	Contig	14	7108960	20238926	**13,129.97**	27/27
NT_039699.7	Fragment	X	3000000	3316790	**316.79**	53/53
NT_039699.7	Fragment	X	4388241	4606212	**217.97**	37/37
NT_097750.2	Contig	X	26532881	26826796	**293.92**	19/19*
NT_111905.2	Contig	X	26876796	27634477	**757.68**	40/40*
NT_166332.1	Fragment	X	30212328	30391205	**178.88**	52/52
				Total (kb):	**25,776.03**	

b. Location of contigs of unknown orientation

Contig	Gap Type	Chr	Start	Finish	Size (kb)
NT_166284.1	Contig	2	175374802	175992526	**617.72**
NT_039268.5	Fragment	4	145539878	145771462	**231.58**
NT_039268.5	Fragment	4	145821462	145967376	**145.91**
NT_039268.5	Fragment	4	146017376	146260876	**243.5**
NT_166301.1	Contig	5	143269871	143311751	**41.88**
NT_081055.6	Fragment	5	143411751	143455620	**43.87**
NT_039413.7	Fragment	7	8231954	8393310	**161.36**
NT_039595.7	Fragment	14	5607431	5828008	**220.58**
NT_039624.7	Fragment	16	3000000	3181574	**181.57**
NT_039649.7	Fragment	17	62167794	62277689	**109.9**
NT_097749.2	Contig	X	27684477	29598808	**1914.33**
NT_111907.2	Contig	X	29648808	30162328	**513.52**
NT_166332.1	Fragment	X	31187882	31480348	**292.47**
NT_166332.1	Fragment	X	31530348	32204481	**674.13**
NT_111911.1	Contig	X	52181231	52365893	**184.66**
NT_166343.1	Fragment	X	121836353	122102747	**266.39**
NT_166332.1	Fragment	X	30441205	30647460	**206.26**
NT_166343.1	Fragment	Y	2847387	2902555	**55.17**
				Total (kb):	**6,104.8**

c. Regions of strand-independent mapping excluded from analysis

Contig	Chr	Start	Finish
NT_039173.7	1	84825357	85413400
NT_039207.7	2	98502526	98507587
NT_162143.3	3	25417704	25459010
NT_039268.5	4	145550273	147208659
NT_039299.7	5	14871715	14889560
NT_165760.2	5	15112964	15143941
NT_039515.6	11	3019750	3100760
NT_039649.7	17	39979964	39985812
NT_039713.7	X	121300929	121315872
NT_039713.7	X	120761558	120812521
NT_039713.7	X	120112573	120129750
NT_166343.1	Y	2780268	2903191

Supplementary Table 1. (**a**) **The location and size of mis-oriented regions in the mouse genome.** Genomic location of all unbridged gaps, defined as gaps in the genome build where the relative order and orientation of the contigs on either side of the gap is unknown, were downloaded from the UCSC table browser. These regions split the chromosomes into 183 genomic fragments where the orientation between fragments is unknown. Mis-oriented fragments were identified in regions where there was a complete switch in the direction of reads (see **Fig. 3**, **Supplementary Figs. 8,9**). The fragments are categorized by chromosome (Chr) and the fragment start and end location, as well as the size of the fragment in kb. The gap type refers to whether the gap occurs within a contig (Fragment) or between adjacent contigs (Contig). For all chromosomes that inherit both Watson or both Crick template strands, the mis-orientation is expected to be visible from Strand-Seq data, depending on the size of the fragment and the coverage in the Strand-Seq library. The proportion of libraries which show these mis-oriented regions out of the expected total is shown in the last column. All libraries with Watson/Watson or Crick/Crick inheritance patterns at these regions unambiguously showed the misorientation with no discrepancies, even in cases where a proportion of libraries had low read numbers in this region (asterisks). (**b**) **Genomic regions in which there is insufficient data to assess orientation status.** Sequence coverage over some fragments was very low, and other fragments contained segmental duplications and repetitive DNA that did not allow strand direction to be determined. With the exception of these unknown fragments, and the mis-oriented regions listed, every single other fragment within the current build of the mouse genome (mm9) is correctly oriented. (**c**) **Regions of strand-independent mapping excluded from analysis.** Taking all 62 Strand-Seq libraries together, several genomic regions were identified in which both Watson and Crick reads always mapped. These regions are rich in segmental duplications, and consistently map a large number of reads in both orientations. Since no Strand-Seq could be detected in these regions in all 62 libraries, they were excluded from the analysis**.**

Supplementary Table 2. Watson/Crick template strand inheritance patterns in non-SCE chromosomes.

	Strand frequency				Aneuploidy	
Chromosome	**WW**	**WC**	**CC**	**SCE**	**Monosomy**	**Trisomy**
1	10	23	9	20	0	0
2	8	11	6	37	0	0
3	15	23	4	20	0	0
4	9	17	11	25	2	0
5	9	17	8	28	4	3
6	13	19	2	28	0	0
7	14	25	7	16	1	0
8	12	20	11	19	0	0
9	10	20	13	19	2	1
10	7	20	7	28	1	0
11	8	17	9	28	0	5
12	11	25	6	20	1	0
13	7	25	8	22	1	0
14	10	24	6	22	1	0
15	10	22	8	22	0	0
16	13	20	12	17	1	0
17	14	22	11	15	0	0
18	12	18	12	20	2	1
19	10	29	12	11	0	1
X	30	0	23	9	62*	0
Y	35	0	23	4	62*	0
Total	267	397	208	430	16	11

Supplementary Table 2. The prevalence of sister chromatid exchanges and aneuploid chromosomes complicates simple analysis of Watson-Crick template strand inheritance. For chromosome 7, the distribution of Watson and Crick template strands in intact, diploid autosomes (and haploid sex chromosomes) without SCE do not differ from expected random frequencies of Watson or Crick template strands as judged by Chi-Squared analysis with a Bonferroni adjustment (data not shown).