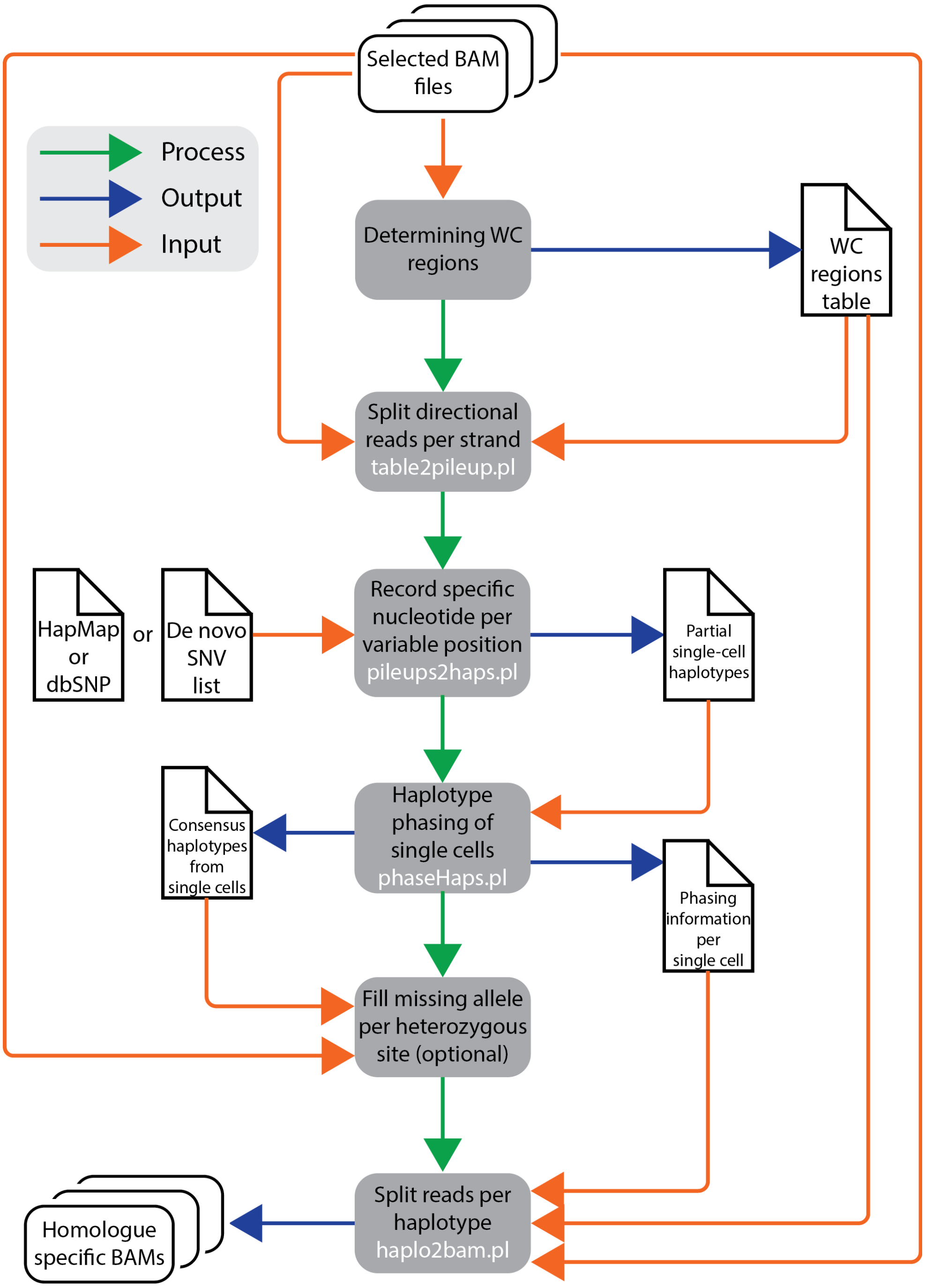
# Supplemental Materials

## Direct chromosome-length haplotyping by single cell sequencing

David Porubsky[1], Ashley D. Sanders[2], Niek van Wietmarschen[1], Ester Falconer[2], Mark Hills[2], Diana C.J. Spierings[1], Marianna R. Bevova[1] , Victor Guryev[1], Peter M. Lansdorp[1,2,3]

1. *European Research Institute for the Biology of Ageing, University Medical Center Groningen, University of Groningen, 9713 AV Groningen, The Netherlands*
2. *Terry Fox Laboratory, BC Cancer Agency, Vancouver, BC V5Z 1L3, Canada*
3. *Division of Hematology, Department of Medicine, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada*


Selected BAM files
Process
Output
Input
Determining WC regions
WC regions table
Split directional reads per strand
table2pileup.pl
HapMap or dbSNP
or
De novo SNV list
Record specific nucleotide per variable position
pileups2haps.pl
Partial single-cell haplotypes
Consensus haplotypes from single cells
Haplotype phasing of single cells
phaseHaps.pl
Phasing information per single cell
Fill missing allele per heterozygous site (optional)
Homologue specific BAMs
Split reads per haplotype
haplo2bam.pl

**Supplemental Figure S1: Bioinformatics pipeline for building haplotypes from Strand-seq data.**

Flow chart of the computational steps executed by our custom analysis pipeline StrandPhase. First, Strand-seq libraries are preselected based on quality criteria (**Supplemental Fig. S2**). Next, WC regions are localized in selected BAM files for every chromosome in each single cell (**see Methods, Section 3**). A list of all genomic regions found to have a WC inheritance pattern is generated. Across these regions, all variants (called *de novo*, or retrieved from a publicly-available database, such as dbSNP or the HapMap project) are recorded separately for Watson and Crick reads. This generates low-density haplotypes for all WC regions in each individual cell. The single cell haplotypes serve as an input for our phasing algorithm to build higher density consensus haplotypes (H1 and H2) of each chromosome. These consensus haplotypes are generated for each chromosome, and together represent a whole genome haplotype for a given individual.

**A**

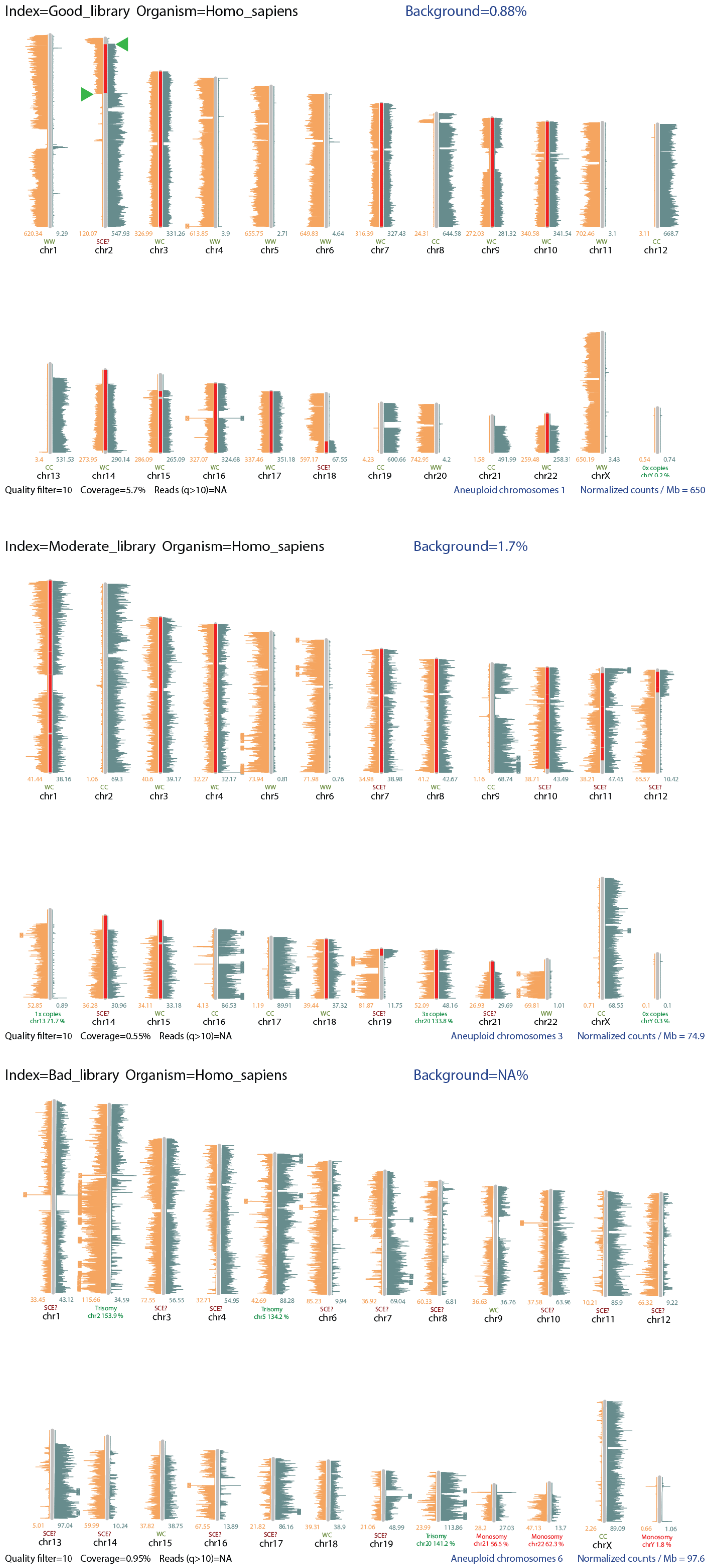
Good Strand-seq library: High reads/Mb, even coverage profile, low background reads, no structural rearrangements like CNVs or aneuploidy.

**B**

Moderate Strand-seq library: Lower reads/Mb, less even coverage profile, low background reads, no structural rearrangements like CNVs or aneuploidy.

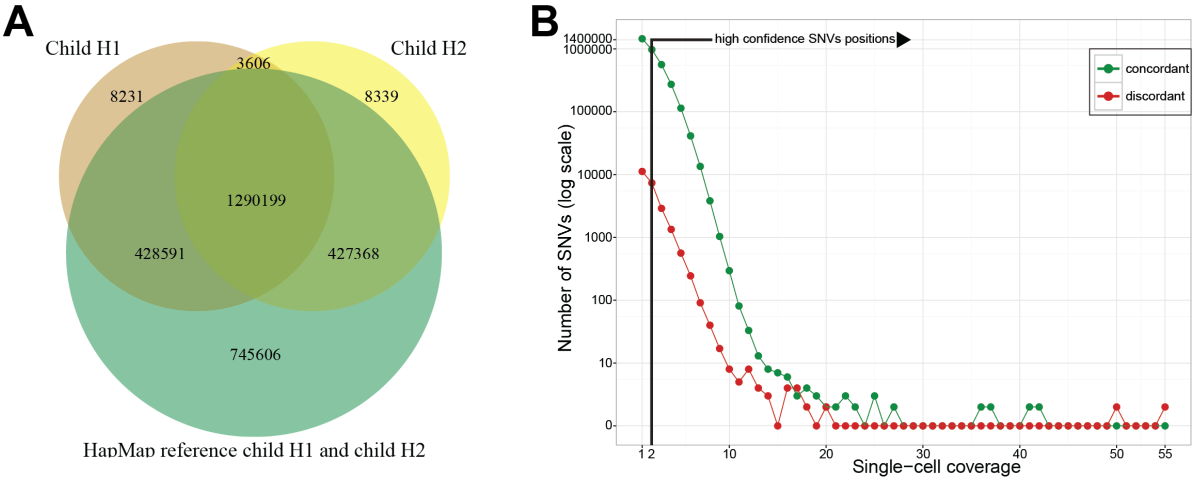
**C**

Bad Strand-seq library: Low reads/Mb, uneven coverage profile, high background reads, presence of structural rerrangements like large segments of CNVs.



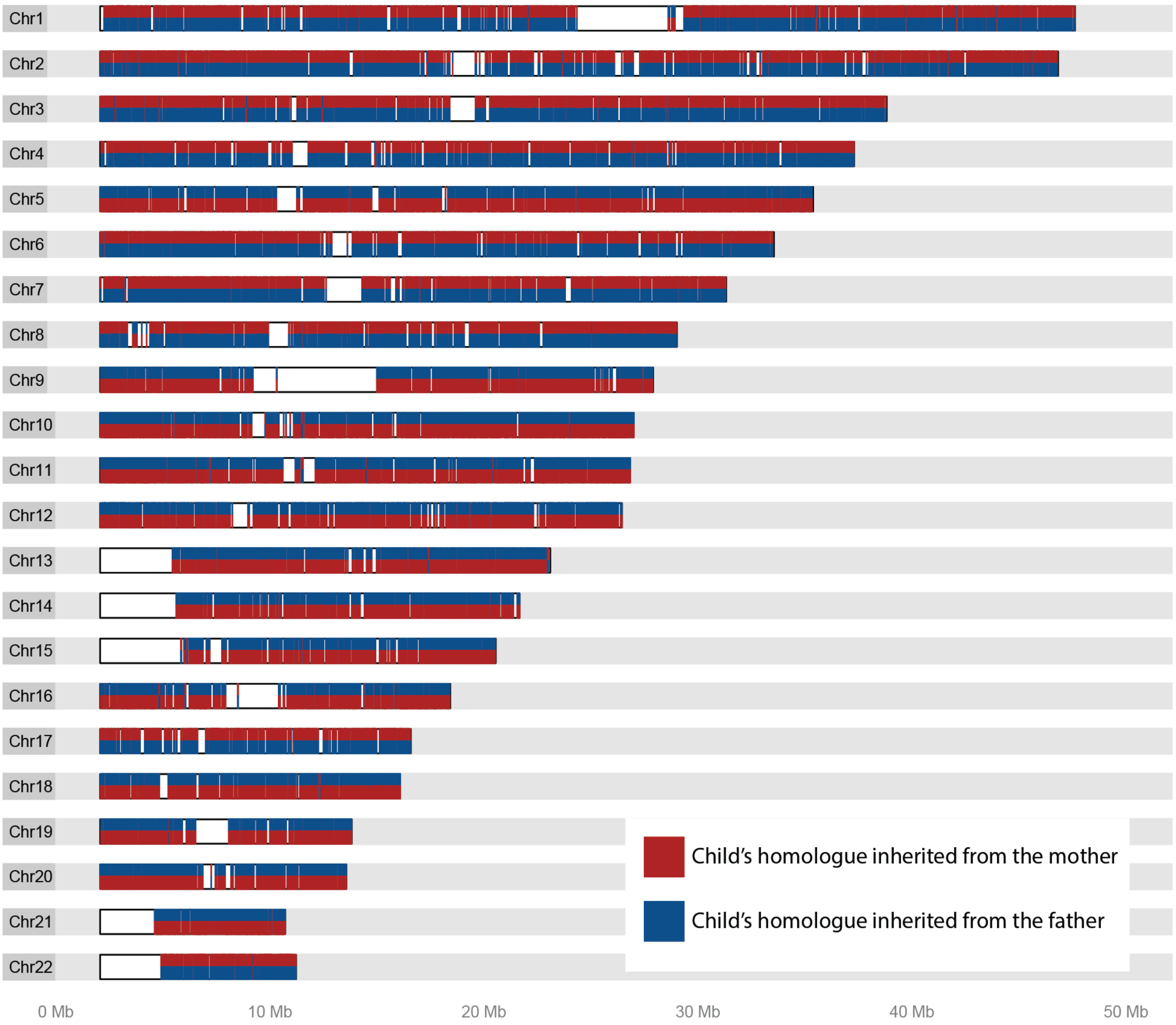
## Supplemental Figure S2: Quality criteria for single cell Strand-seq library

To avoid phasing errors introduced by low quality libraries, we performed a preliminary screen of Strand-seq libraries to select only those suitable for haplotype assembly. Shown are examples of BAIT (Hills et al. 2013) ideograms of libraries categorized by quality. **A**) Good quality Strand-seq libraries have high (> 200) reads/Mb, an even read coverage profile, low background reads (i.e. reads mapped to opposite direction on chromosomes expected to have unidirectional reads), and no obvious structural rearrangements like copy number changes or aneuploidy events. **B**) Moderate quality Strand-seq libraries have lower (50-200) reads/Mb, less even coverage profile, low background reads, and no structural rearrangements. **C**) Poor quality Strand-seq libraries have either low (< 50) reads/Mb, or an uneven coverage profile, high background reads (>5%), or obvious structural rearrangements. Poor libraries were excluded from our analysis. Within high and moderate quality libraries, chromosomes were interrogated for WC inheritance (**see Methods, section 3**). Chromosomal regions highlighted in red were picked for the haplotype assembly, since in these regions we can separate reads mapping to the plus and minus strand of the reference genome. Note, sometimes only a portion of a chromosome exhibited WC inheritance pattern, visible as a template strand state switch from WC to CC or WW (**A, green arrowheads**). This occurs when a double strand break is repaired by homologous recombination during DNA replication, resulting in a sister chromatid exchange event. These WC portions were also selected for analysis.



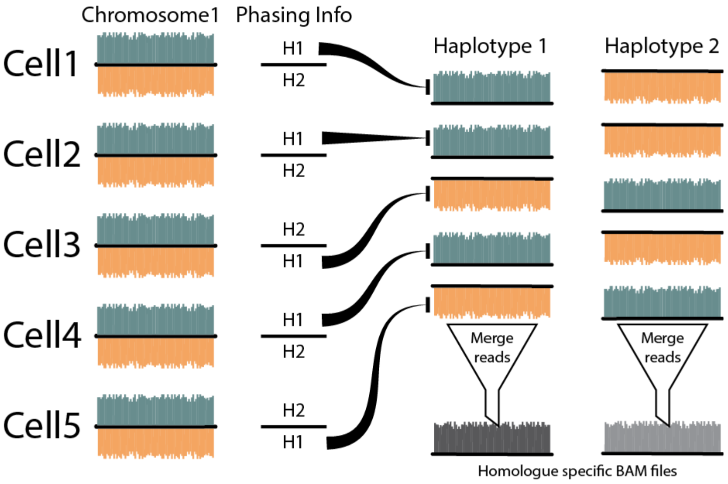
## Supplemental Figure S3: Summary of matches and mismatches with HapMap reference and single cell coverage distribution for SNV positions

**A**) Venn diagram summarizing the total number of SNVs found in Strand-seq data in comparison to the HapMap reference. Brown and yellow circles; haplotypes assembled from the Strand-seq data, green circle; HapMap reference SNVs used for validation. Overlaps with green circle shows number of concordant reads in comparison to the HapMap reference. For example, there are 1,290,199 concordant SNV positions covered on both haplotypes, Child H1 and H2. **B**) All SNV positions found in our Strand-seq haplotypes are plotted by their single cell coverage, which represents the total number of independent cells that supported the variant position. SNVs covered by more than one cell are considered high confidence (black arrow). The SNVs we identified that agree with the variant listed in the HapMap reference are shown in green, and the discordant SNVs (i.e. mismatches) are shown in red. The mismatching SNV positions that are high confidence may represent errors in the HapMap reference or possible *de novo* mutations in our cell sample.



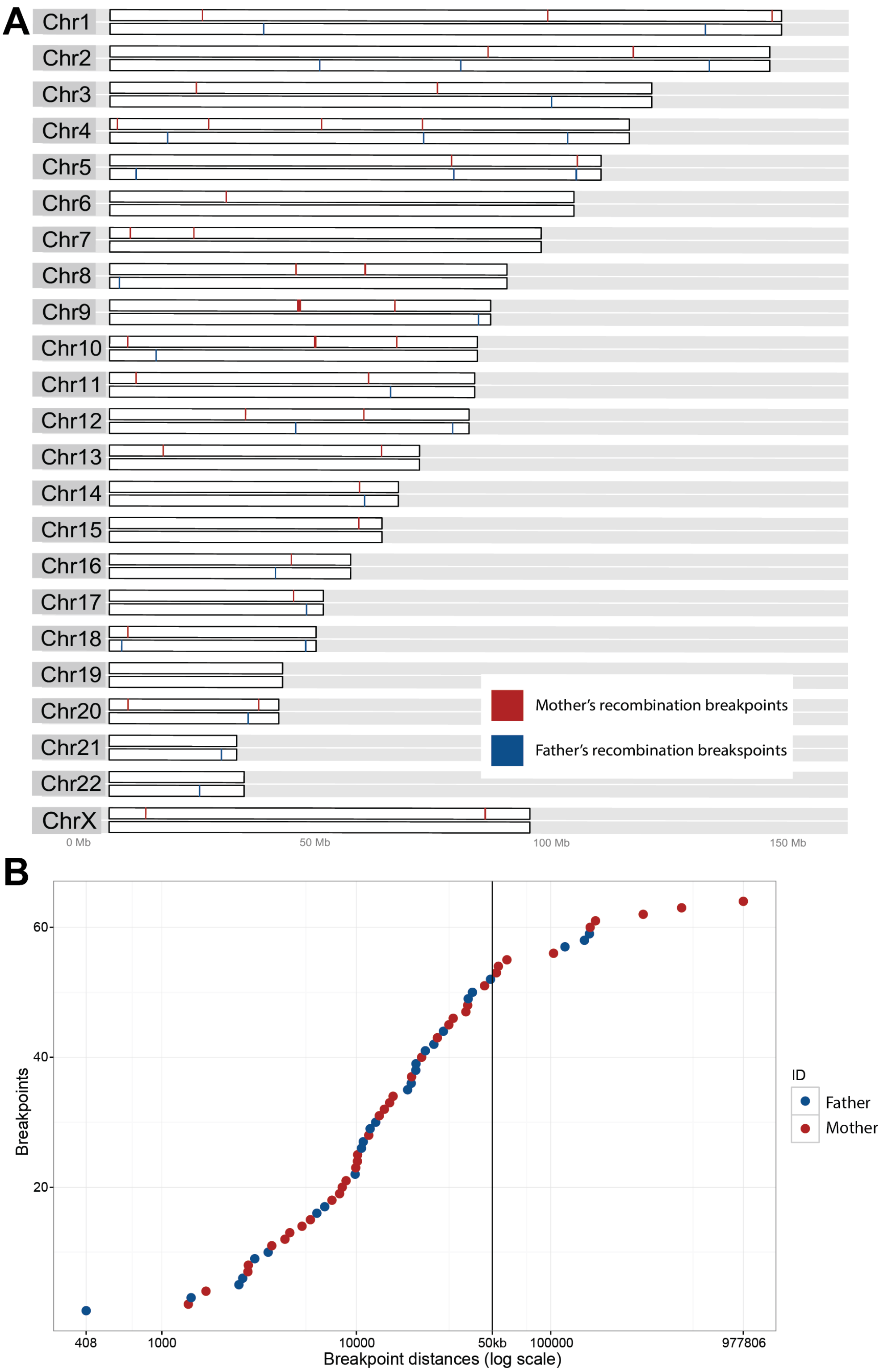
**Supplemental Figure S4: Comparison of Strand-seq child's haplotypes with Strand-seq parental haplotypes**

To unambiguously assign the parental origin of each allele in the child, we assessed only high confidence SNV positions (i.e. present in > 2 cells) that were heterozygous in the child. In addition, such positions had to be homozygous in at least one parent and the other parent had at least one variant phased. Each horizontal ideogram represents the two haplotypes of a chromosome, and each SNV is represented as a vertical line in the ideogram, with the colour denoting the parental homologue they match. The child's haplotypes were either of paternal (blue) or maternal (red) origin.



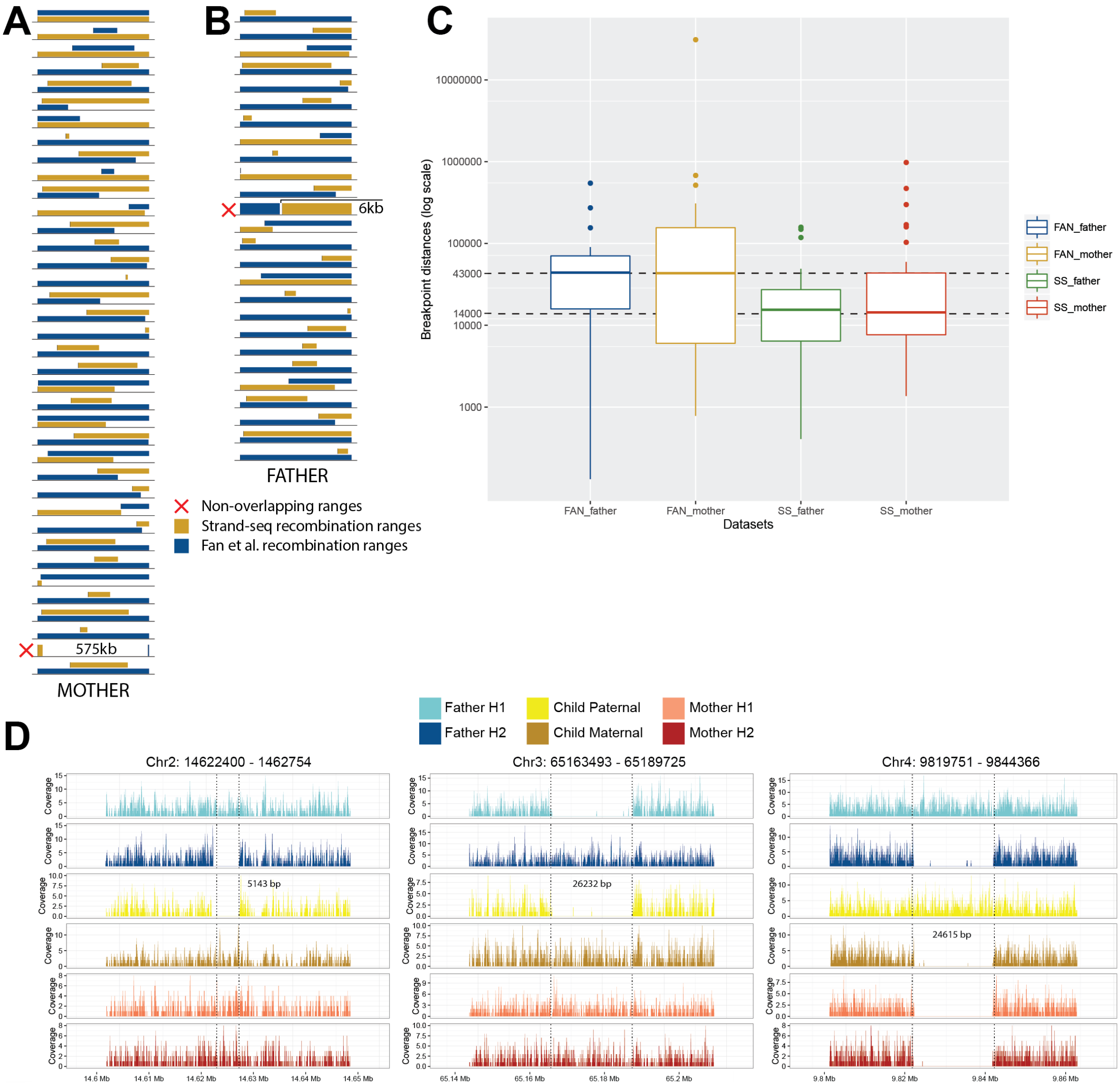
**Supplemental Figure S5: Generation of homologue specific BAM file**

Homologue specific BAM files were created for each phased homologue using SAMtools (Li et al. 2009). For this, the sequencing reads in every cell were split based on directionality (Crick shown in blue, Watson shown in orange) and assigned to their respective haplotype using our phasing algorithm (**see Methods, Section 3**). All phased reads were then merged together into a high-density homologue specific BAM file representing consensus haplotype 1 (H1), or haplotype 2 (H2). Two BAM files were generated for every chromosome.



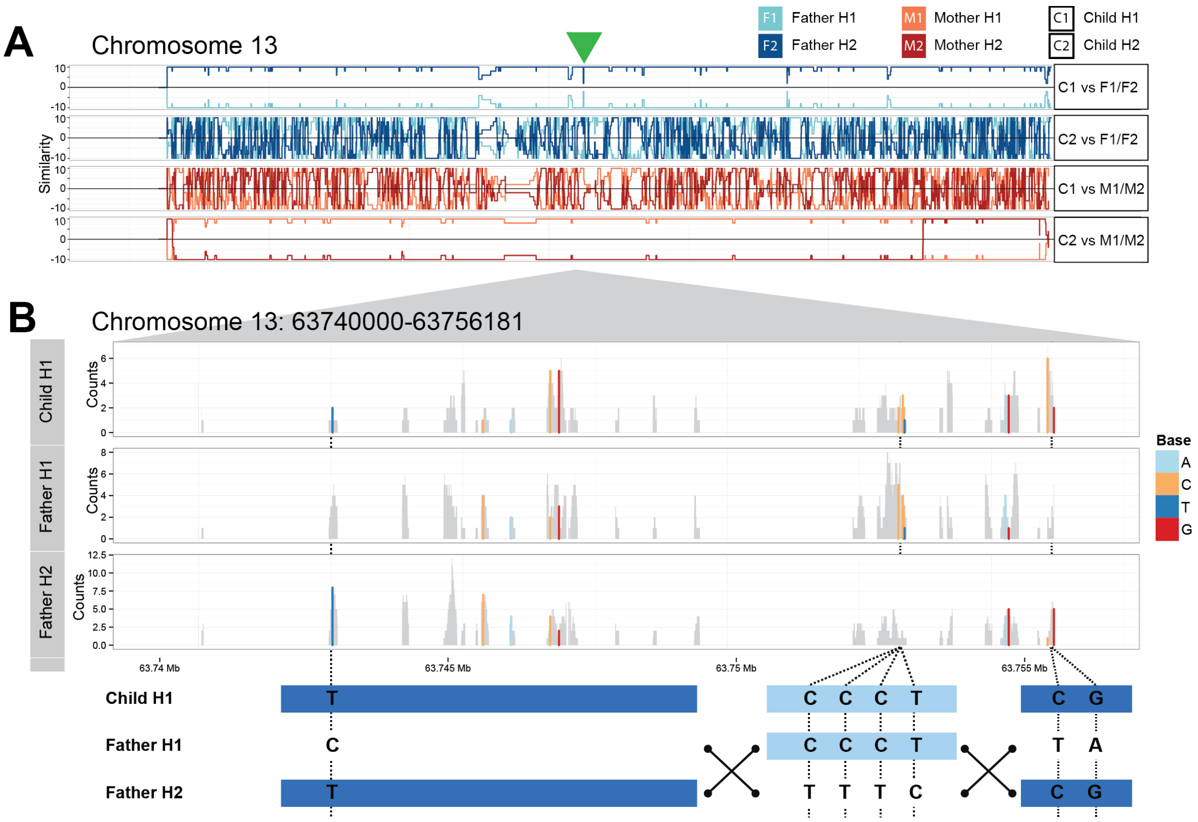
## Supplemental Figure S6: Genome-wide map of recombination breakpoints

**A**) The genomic locations of maternal (red) and paternal (blue) meiotic recombination events, plotted for each homologue of the child. The width of each vertical bar represents the length of the region where recombination event was mapped. **B**) The size distribution of all mapped recombination breakpoints. Vertical line shows that the majority of breakpoints were mapped to a region less than 50 kb in size. The outliers arise within centromeres, where precise breakpoint mapping is challenged by reference assembly gaps and/or only a small number of reads mapping uniquely.



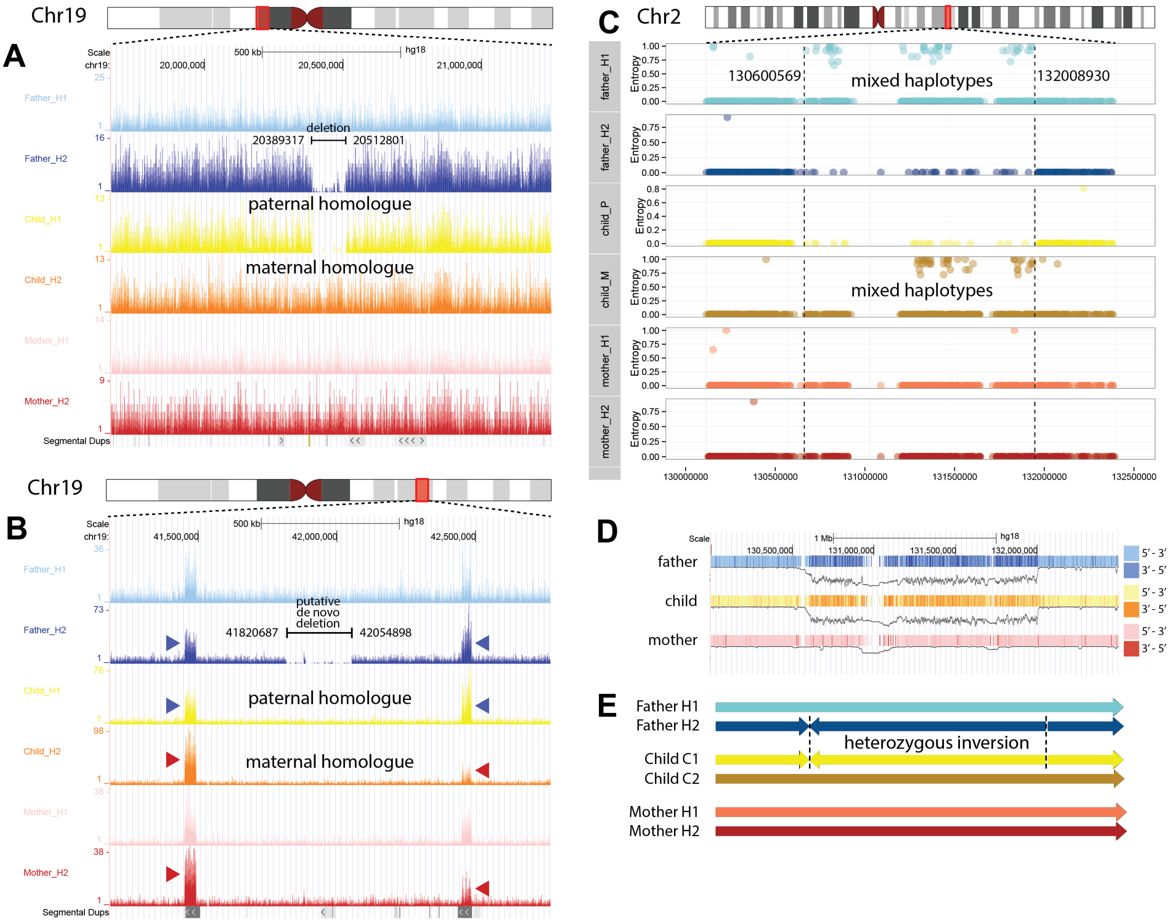
**Supplemental Figure S7: Comparison of phasing accuracy between Strand-seq and data from Fan et al. (2011)**

**A**) Overlap of localized meiotic recombination breakpoints between Strand-seq and Fan *et al.* for the mother. Each horizontal black line underlines one recombination event with yellow and blue rectangles showing region of meiotic event localized by Strand-seq and Fan *et al.*, respectively. Red cross point to the recombination event where Strand-seq and Fan *et al.* do not overlap with corresponding distance between localized recombination events. **B**) Overlap of localized meiotic recombination breakpoints between Strand-seq and Fan *et al.* for the father. **C**) Boxplot comparing size distribution of localized meiotic breakpoints using Strand-seq (SS) and by Fan *et al.* (FAN). **D**) Example of three heterozygous deletions from Fan *et al.* validated by Strand-seq (complete set in **Supplemental Table S7**). Horizontal panels represents separate homologues of each individual in the trio. Vertical colored lines represent read coverage in homologue specific BAM files (**Supplemental Fig. S5**). Dotted lines shows boundaries of heterozygous deletions with breakpoint coordinates at the top.



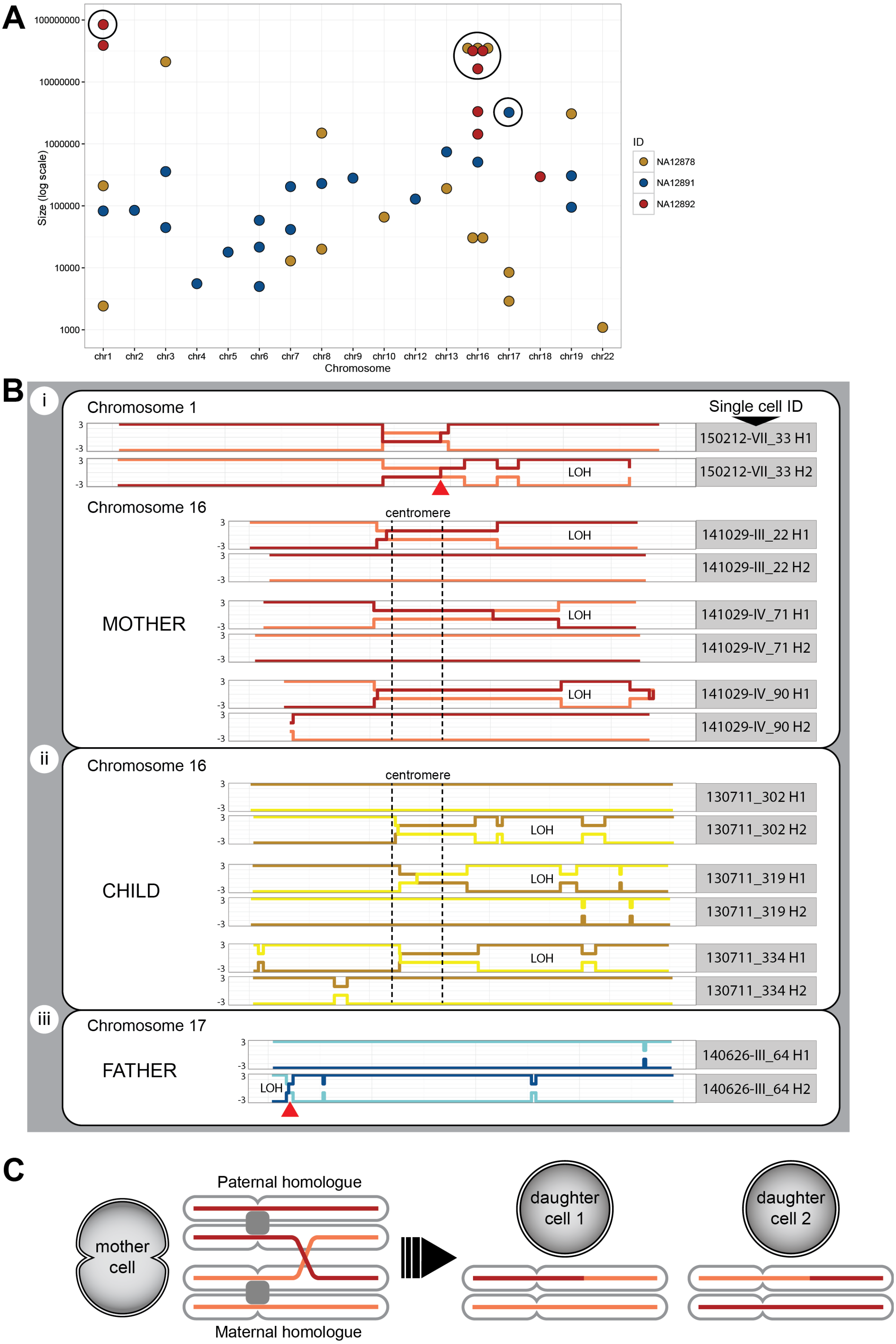
**Supplemental Figure S8: Location of putative gene conversion event.**

**A**) Similarity plot for Chromosome 13 depicting pairwise comparison of each child homologue with both parental homologues. Green arrowhead points to a short region where similarity of Child H1 and Father H1 decreases. This presents a putative meiotic event resolved as a gene conversion. **B**) Enlarged region on Chromosome 13 of the child's homologue inherited from the father. Along each homologue (child H1, father H1 and H2) we plot read coverage (gray) with differing nucleotides highlighted (see legend). In this short region 4 consecutive heterozygous SNVs (in light blue) are switched.



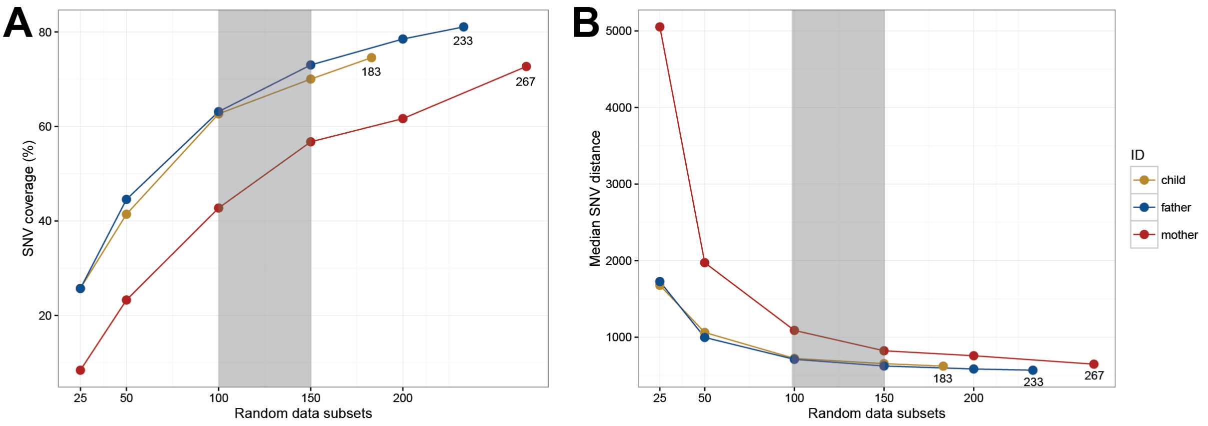
## Supplemental Figure S9: Phasing of structural variants on Chromosome 2 and Chromosome 19

**A,B**) UCSC Genome Browser view of reads from all single cells aligned to a single individual's homologue (**Supplemental Fig. S5**) in a zoomed region of Chromosome 19 (Chr19). **A**) The disruption in read density illustrates a heterozygous deletion found on Father H2 and inherited in the child (Child H1). **B**) A second deletion downstream on Chr19, which is flanked by segmental duplications. Here reads weren’t filtered by mapping quality because reads mapping to segmental duplications are assigned low mapping quality. The absence of the Father H2 deletion in the child lineage suggests the variant arose *de novo* in the father cell line. In addition, we can see two copy number variants overlapping with known segmental duplications in this region. Read coverage of these regions suggests that copy number holds for corresponding paternal (blue arrowheads) and maternal (red arrowheads) homologues inherited in the child. **C**) Horizontal panels represent entropy values for every SNV in a single individual's homologue (H1 or H2) of a zoomed region on Chromosome 2 (Chr2). High values of entropy reflect the presence of more than one allele at the variable site as a result of mixed haplotype structures at the locus. We can see mixed haplotypes (more than one allele) in the father H1 and child H2. Breakpoints of this region are drawn in dashed line. **D**) UCSC Genome Browser view showing Strand-seq reads in the corresponding region for each individual, with the colour denoting the directionality of reads aligned to the reference. The underlying Invert.R histogram (Sanders et al. 2016) shows the mixed representation of directional reads aligned to the plus and minus strand of the reference genome in the father and the child, indicative of a heterozygous inversion at the locus. **E**) Schematic representation of each homologue per individual illustrating the phase of the inversion (arrow), which is placed to Father H2 and Child H1.


A
100000000
10000000
1000000
100000
10000
1000
Size (log scale)
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr12
chr13
chr16
chr17
chr18
chr19
chr22
Chromosome
ID
NA12878
NA12891
NA12892
B
i
Chromosome 1
Single cell ID
150212-VII_33 H1
150212-VII_33 H2
LOH
Chromosome 16
centromere
MOTHER
141029-III_22 H1
141029-III_22 H2
141029-IV_71 H1
141029-IV_71 H2
141029-IV_90 H1
141029-IV_90 H2
ii
Chromosome 16
centromere
CHILD
130711_302 H1
130711_302 H2
130711_319 H1
130711_319 H2
130711_334 H1
130711_334 H2
iii
Chromosome 17
FATHER
140626-III_64 H1
140626-III_64 H2
3
-3
C
mother cell
Paternal homologue
Maternal homologue
daughter cell 1
daughter cell 2

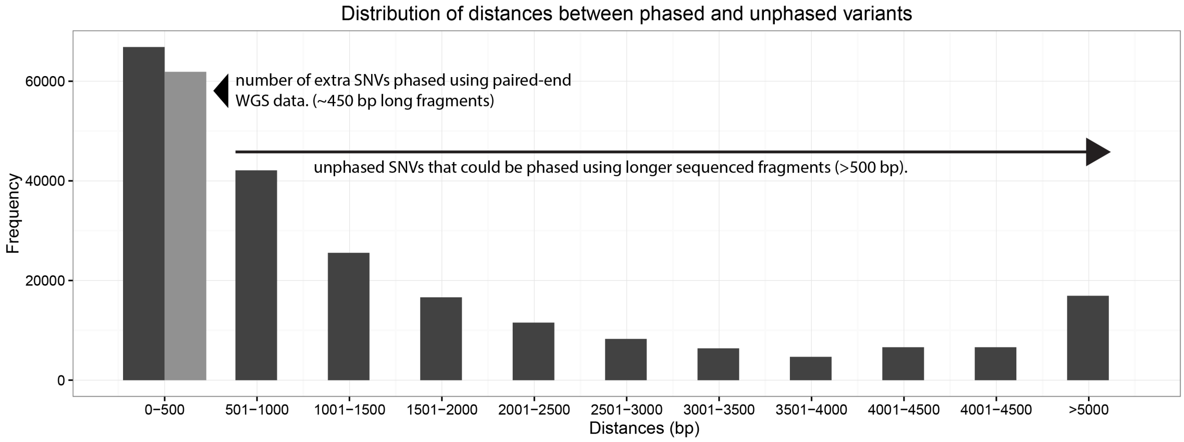
## Supplemental Figure S10: Mapping regional changes in haplotypes at the single cell level

**A**) Size distribution of all loss of heterozygosity (LOH) regions located within single cell Strand-seq libraries, plotted for each individual by chromosome and colored based on the family member. Black circles mark LOH regions encompassing a whole chromosomal arm, some of which occurred near the same genomic location in multiple single cells, and within different individuals (as exampled for Chromosome 16 in B). **B**) Detailed analysis of LOH. For the chromosome arms outlined in a), a comparison was made between haplotypes. Each single-cell identifier is assigned as H1 or H2 based on the consensus haplotype it belongs to. The y-axis represents similarity values (+3, -3) of this homologue, in comparison to the consensus haplotypes (**see Methods, Section 6**), with the x-axis representing the position along the single cell homologue (H1 and H2). Red arrows points to positions where single cell haplotype starts to match opposing consensus haplotype. For Chromosome 16 of the mother and the child we predict that recombination occurred in the centromeric region (dashed line). Note that the observed LOH only occurs on one of the two homologues. **C**) Model for observed LOH, where double strand break repair in mitotic cells results in homologous mitotic recombination between homologues.



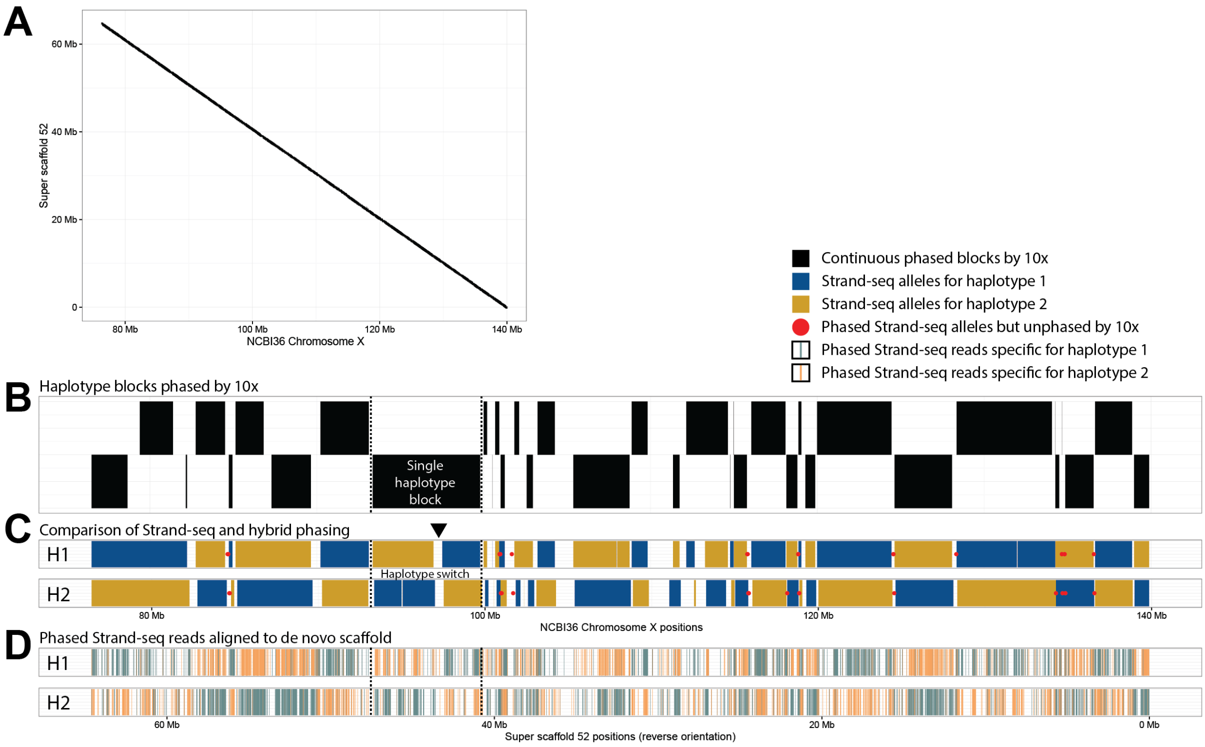
## Supplemental Figure S11: Evaluation of SNV coverage and SNV density in various subsets of Strand-seq libraries

To determine the number of Strand-seq libraries required to build accurate whole genome haplotypes, we down-sampled our datasets and assessed the SNV coverage and density of the resultant haplotypes. Subsets of single cell libraries (between 25-200 cells) were randomly selected and haplotypes built for each (**see Methods, Section 3**). **A**) The percentage of covered SNVs is plotted for each subset and for each individual as a separate line. As expected, we see a positive correlation between number of cells and percentage of SNVs covered. We observed that the increase of covered SNVs is less prominent at higher number of cells. **B**) The median distance between neighboring SNVs is plotted for each subset and for each individual as a separate line. Here increasing number of cells is negatively correlated with decreasing distance between neighboring SNVs. From this data, we concluded that ~100-150 cells (shaded gray region) are optimal to reach informative haplotypes at a reasonable cost.



## Supplemental Figure S12: Improvement of phased SNV coverage by combining Strand-seq and WGS data

In dark gray bars the frequency of distances between any phased heterozygous variants and the closest unphased heterozygous variant from the HapMap reference is plotted. To estimate how many additional HapMap reference variants can be phased we used publicly available WGS data for NA12878 (SRR1910366 – NCBI SRA archive). This dataset contains 250 bp long paired-end reads sequenced on Illumina 2500 platform. We aligned these data to the reference genome NCBI36 using the Bowtie 2 (Langmead and Salzberg 2012) aligner. Subsequently we searched for the read pairs for which at least one mate of the pair overlapped with phased heterozygous SNV in our data (findOverlaps function from R package Granges). Since read pairs originate from the same fragment of DNA (haplotype), every mate of the pair overlapping with a phased SNV can be used as an anchor to phase the other mate of the read pair. In the light gray bar we show the estimated number of additional heterozygous variants listed in the HapMap reference phased using this approach. Since the average fragment size of the WGS data was 450 bp, we assume most of these additionally phased variants were within 500 bp from the closest phased variant. As you can see almost all (92.5%) unphased variants that were at the distance lower than 500 bp were phased using WGS data. We anticipate that using paired-end reads with longer fragment sizes or long-read sequencing data can be used to phase other variants listed in the HapMap reference.



**Supplemental Figure S13: Comparison of phasing accuracy between Strand-seq and hybrid phasing (Mostovoy et al. 2016)**

**A**) Dotplot visualizing the alignment of Super scaffold 52 to NCBI36 Chromosome X. **B**) Blocks of haplotypes phased by 10x Genomics plotted alternatively below and above the midline for better resolution. **C**) Strand-seq haplotypes (see the legend) compared to phased haplotype blocks from B. Each horizontal panel represents a single Strand-seq haplotype (H1 or H2) compared separately to phased blocks from B. Red dots represent alleles phased by Strand-seq but unphased by 10x. **D**) Phased Strand-seq reads colored by haplotype (see the legend) aligned to Super scaffold 52. Each horizontal panel represents reads aligned to the haplotype specific sequence of Super scaffold 52 (H1 or H2).

**Dotted line –** shows haplotype block where Strand-seq and 10x phasing disagree.

**Black arrowhead** – points to putative haplotype switch error

# Supplemental Tables 1-9

| | NA12878 | | NA12891 | | NA12892 | |
|---|---|---|---|---|---|---|
| Total number of libraries | 183 | | 233 | | 267 | |
| | SE 51 | PE 132 | SE 44 | PE 189 | SE 209 | PE 58 |
| Average genome coverage per cell | 2.5% | | 2.4% | | 1.4% | |
| Genome coverage in merged cells | 79.4% | | 86.5% | | 80.9% | |
| Depth in merged cells | 5.7 | | 7.9 | | 4 | |
| Genome coverage per haplotype | 55.1% | | 60.8% | | 50.2% | |
| Depth per haplotype | 1.36 | | 1.82 | | 0.98 | |
| HapMap SNVs covered | 74.6% | | 82.5% | | 72.7% | |

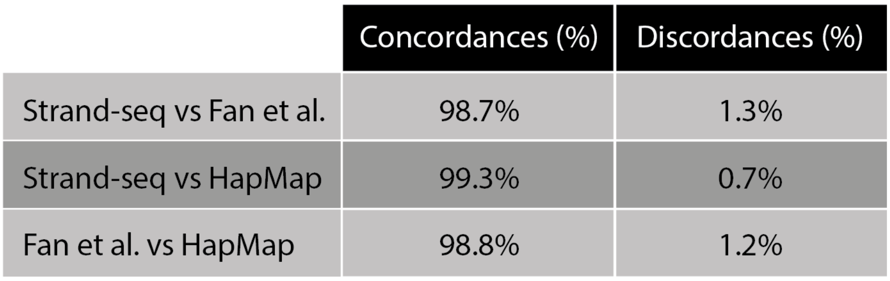
**Supplemental Table S1: Summary of sequencing data for each individual in the HapMap family trio**

Total number of sequenced libraries for the child (NA12878), father (NA12891) and mother (NA12892) of the family trio analyzed in this study. The number of libraries sequenced as single-end (SE) or paired-end (PE) reads are listed. Genome coverage was calculated per mappable genome (mappability file obtained from the UCSC Genome Browser database - /gbdb/hg18/bbi/wgEncodeCrgMapabilityAlign50mer.bw) and represents the percentage of genomic positions covered by sequencing reads. Depth of coverage represents the average amount of bases sequenced per genomic position. Finally, the percentage of HapMap reference SNVs covered per individual is shown.

| | NA12878 | NA12891 | NA12892 |
|---|---|---|---|
| **Consistent with Strand-seq (%)** | **99,2%** | **98,9%** | **99,8%** |
| **Consistent with HapMap (%)** | **94,7%** | **94,8%** | **93,9%** |

**Supplemental Table S2: Comparison of Pacbio data to Strand-seq haplotypes**

We performed a direct comparison of our Strand-seq haplotypes with long-range Pacbio RNA-seq reads as an additional test that our haplotypes are correct. Our validation is based on the fact that any PacBio read overlapping at least two heterozygous positions represents a phased “mini” haplotype. Therefore, only PacBio reads that overlapped with at least two heterozygous alleles (phased using Strand-seq) were included in the analysis. The percentage of consistent and inconsistent PacBio reads was calculated as a fraction of all PacBio reads overlapping with Strand-seq haplotype backbone and passing filtering criteria.



| | Concordances (%) | Discordances (%) |
|---|---|---|
| Strand-seq vs Fan et al. | 98.7% | 1.3% |
| Strand-seq vs HapMap | 99.3% | 0.7% |
| Fan et al. vs HapMap | 98.8% | 1.2% |

**Supplemental Table S3: Comparison of whole genome haplotypes between Strand-seq and Fan et al.**

To directly compare phasing performance of Strand-seq with other single-cell based phasing approach we chose study by Fan *et al.* (2011). Both techniques can achieve chromosome length haplotypes with the ability to map meiotic recombination breakpoints within a family trio. To evaluate these two techniques, we performed three-way comparison of the child (NA12878) between Strand-seq, Fan et al. and HapMap reference haplotypes. We have observed slightly better concordance between Strand-seq and HapMap (99.3%) than between Fan *et al.* and HapMap (98.8%). Concordance between Strand-seq and Fan *et al.* was 98.7%. Comparison of parental haplotypes (NA12891 and NA12892) between Strand-seq and Fan *et al.* scored equally well as in the case of child's haplotypes with overall concordance 98.7%. These results demonstrate the high accuracy of both techniques.

**Supplemental Table S4: List of phased germline *de novo* mutations of the child**

(provided as separate xls file as a part of Supplemental dataset)

**Supplemental Table S5: Coordinates of mapped meiotic recombination events**

(provided as separate xls file as a part of Supplemental dataset)

**Supplemental Table S6: Coordinates of short switch events**

(provided as separate xls file as a part of Supplemental dataset)

**Supplemental Table S7: Comparison of phased deletions between Strand-seq and Fan et al.**

(provided as separate xls file as a part of Supplemental dataset)

**Supplemental Table S8: Phased deletions >1000kb from 1000 Genomes Project**

(provided as separate xls file as a part of Supplemental dataset)

**Supplemental Table S9: Coordinates of LOH events in single cells**

(provided as separate xls file as a part of Supplemental dataset)

# Supplemental Methods

## Comparison of Strand-seq phasing with data from Fan et al.

We compared Strand-seq based phasing with data obtained from Fan *et al.* (2011). First we compared overlap of meiotic recombination events localized by Strand-seq (**Supplemental Table S5**) and Fan *et al.*. Comparison was visualized using R packages ggbio and ggplot2. Overlaps between recombination ranges were summarized using R function findOverlaps from Genomic Ranges package. Next, we compared the phasing of heterozygous deletions described in Fan *et al.* with Strand-seq phasing. For this analysis we used homologue specific BAM files created by Strand-seq for each individual in the trio (NA12878, NA12891 and NA12892, **Supplemental Fig. S5**). Using a custom PERL script and SAMtools we counted the number of homologue specific reads in regions of heterozygous deletions obtained from Fan *et al.* Such read counts were corrected for the size of the deletion and normalized per 1kb [ (readCount/deletionSize)*1000 ]. Such normalized read counts were compared with the deletion profiles described by Fan *et al.* Lastly we compared whole genome haplotypes for all family members between Strand-seq and Fan *et al.* (**Supplemental Table S3**).

## Comparison of Strand-seq phasing with *de novo* genome assembly based phasing

Strand-seq phasing for an individual (NA12878) was compared with *de novo* genome assembly based phasing from Mostovoy et al (Mostovoy et al. 2016). Data necessary for comparison (assembled contigs in FASTA file and phased VCF file) were downloaded from http://kwoklab.ucsf.edu/resources/. First we aligned Super-scaffold 52 to the reference Chromosome X (NCBI36 build) using Lastz (Harris R. S. 2007) with the parameters used by Mostovoy et al., except that we used rdotplot as the output format (--format=rdotplot --ungapped --notransition --maxwordcount=90% --exact=500 --identity=95 --seed=match15 --ambiguous=iupac --match=1,5 --twins=1..100). The resulting file contained coordinates of each mapped part of the contig relative to the Chromosome X. Using custom PERL script we transferred the contig specific coordinates of phased SNVs from VCF file into Chromosome X (NCBI 36) specific coordinates to make them comparable with phased Strand-seq data. Next phased SNVs from Mostovoy *et al.* were compared to phased SNVs covered in Strand-seq dataset. To exclude possible errors caused by alignment of Super-scaffold to the reference Chromosome X we decided to align Strand-seq phased reads to the Super-scaffold 52. For this we converted homologue specific BAM files for the Chromosome X into a single FASTQ file for each homologue using the bedtools <bamtofastq> function (bedtools v2.17.0). Before alignment we created two haplotype specific references for Super-scaffold 52 by substituting every SNV position with haplotype specific allele from previously downloaded VCF file. Such haplotype specific reference were merged into a single FASTA file and indexed using Bowtie 2 (v2.1.0). Subsequently homologue specific reads with unique ID were merged into a single FASTQ file and were aligned using Bowtie 2 to the haplotype specific reference for Super-scaffold 52. The resulting SAM file was converted into a BAM file using SAMtools (v0.1.19-44428cd). Data were plotted using ggplot and filtered for mapping quality of 30 and duplicate reads.

## Strand-seq phasing of structural variants from 1000 Genomes Project

In order to prove that Strand-seq can be used as a tool to phase structural variants (SV) of various sizes we explored previously mapped and phased variants from 1000 Genomes Project (Sudmant et al. 2015). VCF file with phased SV for this study was downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/ALL.wgs.integrated_sv_map_v2.20130502.svs.genotypes.vcf.gz.

For our analysis we filtered out deletions smaller than 1kb. We extracted coordinates of each deletion from the VCF file and transferred them from the reference genome hg19 to hg18 using UCSC's liftover tool. We decided to phase this set of deletion for whole trio to be able to see inheritance patterns as well. For this analysis we used homologue specific BAM files (**Supplemental Fig. S5**) created by StrandPhase. Next we simply counted the number of reads within the boundaries of each deletion for all homologue specific BAM files using a custom PERL script and SAMtools. Additionally, we excluded deletions with read count lower than 50 reads across all homologue specific BAM files. We manually genotyped every deletion for all three individuals. Results are summarized in **Supplemental Table S8**. Next we attempted to phase smaller SV (<1kb) like indels as well. List of mapped indels for NA12878 was obtained from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. First we have transferred coordinates of all SVs in the list from the reference genome hg19 to hg18 using overlapping rs IDs from dbSNP138. List of SVs from the VCF file was then genotyped in both homologue specific BAM files (**Supplemental Fig. S5**) using GATK's HaplotypeCaller with default settings (McKenna et al. 2010). Resulting VCF file is available upon request.

# Supplemental Note

**Comparison of Strand-seq phasing with hybrid phasing described by Mostovoy et al.**

Recent hybrid phasing strategies, such as that presented by Mostovoy *et al.,* integrate short Illumina reads, linked-reads from 10x Genomics and BioNano Genomics optical data to generate haplotype-aware *de novo* genome assemblies. This approach aims to be less biased and more accurate than phasing strategies that rely on read alignment to reference genomes. To test how a hybrid approach compares with Strand-seq phasing we compared the phasing of the large 64 Mb long 'Super-scaffold' 52 assembled by Mostovoy *et al.* To translate the coordinates of the contigs to our reference assembly we aligned Super-scaffold 52 to NCBI 36 Chromosome X (**Supplemental Fig. S13a, see Supplemental Methods**). This revealed Super-scaffold 52 was composed of shorter haplotype blocks that were not linked continuously from start to end, (**Supplemental Figure S13b**). This is reflected in our comparison where we see smaller haplotype blocks matching between long range Strand-seq haplotypes and shorter 10x Genomics derived haplotypes (**Supplemental Fig. S13c, blue and yellow rectangles**). Despite this, the concordance within each haplotype block between Strand-seq and hybrid phasing was impressive, at 99.9 % and 99.7% for haplotype 1 and haplotype 2, respectively.

In contrast, phased block number 9 (**Supplemental Fig. S13c, dotted lines**) did not agree with the phasing obtained from Strand-seq, where a large switch error is evident in the middle of the phased block (**Supplemental Fig. S13c, black arrowhead)**. To test whether this reflects an error in the reference assembly used for Strand-seq phasing we aligned Strand-seq reads directly to the haplotypes *de novo* assembled for Super-scaffold 52 (**see Supplemental Methods**). The alignment of Strand-seq sequencing reads (hap1 – blue, hap2 – orange) to the Super-scaffold 52 supported the phasing observed in the Strand-seq data (**Supplemental Fig. S13d)**, suggesting that phased block number 9 was incorrectly phased using the hybrid approach. This comparison of Strand-seq phasing with hybrid phasing suggests no substantial bias was introduced by mapping reads to the reference genome assembly. On the contrary, our results suggest that integrating Strand-seq may help better refine *de novo* assemblies to build the most accurate haplotypes for an individual.

**Cost and labor requirements on Strand-seq library production**

Strand-seq currently costs approximately $2,500 per 96 cells ($ 26/cell), not including sequencing costs. It takes ~ 3 working days to prepare 192 cells from sorting to sequence-ready Strand-seq libraries. To obtain enough coverage, 192 barcoded Strand-seq libraries are then pooled together and sequenced on a single lane of Illumina HiSeq 2500. In case of paired-end protocol using rapid run 100 b.p. we obtain on average of 2,029,799 uniquely mapped reads/library. Sequencing cost may vary depending on the platform used for sequencing (MiSeq, HiSeq or specialized sequencing service), country or special discounts.

## References

Harris, R.S. 2007. Improved Pairwise Alignment of Genomic DNA PhD thesis, Pennsylvania State Univ.