Haplotype resolved genomes
Porubský, David

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2017

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
CHAPTER 3
Direct chromosome-length haplotyping by single cell sequencing – PART 1: Method validation

David Porubsky¹, Ashley D. Sanders²,³, Niek van Wietmarschen¹, Ester Falconer², Mark Hills², Diana C.J. Spierings¹, Marianna R. Bevova¹, Victor Guryev¹, Peter M. Lansdorp¹,²,⁴

1. European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, 9713 AV Groningen, The Netherlands
2. Terry Fox Laboratory, BC Cancer Agency, Vancouver, BC V5Z 1L3, Canada
3. Current address: European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
4. Department of Medical Genetics, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada

Adapted from Genome Research 2016; 10.1101/gr.209841.116
CHAPTER 3

ABSTRACT

Haplotypes are fundamental to fully characterize the diploid genome of an individual, yet methods to directly chart the unique genetic makeup of each parental chromosome are lacking. Here we introduce single cell DNA template strand sequencing (Strand-seq) as a novel approach to phase diploid genomes along the entire length of all chromosomes. We demonstrate this by building a complete haplotype for HapMap individual (NA12878) at high accuracy (concordance 99.3%), without using generational information or statistical inference. High accuracy and robustness of our phasing approach was further highlighted by the comparison to PacBio long reads as well as other haplotyping studies. Therefore we conclude that Strand-seq offers a novel haplotyping approach able to reliably phase diploid genomes along the whole length of all chromosomes.

INTRODUCTION

Diploid organisms, like humans, contain two homologous copies of each chromosome, one inherited from the mother and one from the father. Despite being highly similar, each homologous chromosome harbors a unique set of genetic variants, ranging from single nucleotide variants (SNVs), insertions, and deletions, to large polymorphic inversions. The collection of genetic variants along a single chromosome is called a haplotype, and the process of assigning variants to corresponding haplotypes is referred to as phasing.

Haplotype-resolved genomes are important in many areas of personalized medicine and genetics, ranging from variant-disease associations (Glusman, Cox, and Roach 2014), mapping regions with loss of heterozygosity (LOH) (Huang et al. 2007), to studies of inheritance patterns in pedigrees and populations (Tewhey et al. 2011). To phase genetic variants (alleles) into haplotypes, both computational and experimental approaches have been developed (Browning and Browning 2011). Currently, massively parallel sequencing provides the most complete set of alleles of an individual. Unfortunately, phasing these variants across the length of a chromosome is currently very challenging unless the parents of the individual are also sequenced (Kitzman et al. 2011; Amini et al. 2014). To overcome this limitation, whole chromosome sorting (Ma et al. 2010; Brown et al. 2012; Fan et al. 2011) and chromatin capture techniques (Selvaraj et al. 2013) have been developed. However, such techniques are labor- and time-consuming and have not been widely adopted in practice. To overcome these limitations linked-read sequencing (Zheng et al. 2016) was recently proposed to deliver long-range haplotypes. However, using this method...
it is not yet possible to phase genetic variants across whole chromosomes.

Here we introduce Strand-seq (Falconer et al. 2012) together with a custom bioinformatics pipeline as a novel, direct approach for haplotyping variants along the entire length of the chromosome. While our approach requires preparation of single cell libraries, it circumvents the need for generational information and rapidly builds accurate whole chromosome haplotypes. We demonstrated high accuracy of our phasing method by comparison with various independent data sources like HapMap project, PacBio RNA-seq as well as other haplotyping studies (Fan et al. 2011; Mostovoy et al. 2016).

RESULTS
Phasing using single cell template strand sequencing

Strand-seq is a single cell sequencing technique in which only one strand of DNA of each chromosome is sequenced, allowing individual homologues to be distinguished as either Watson (W, reverse strand), or Crick (C, forward strand) based on read alignment to the reference genome (Fig. 1A, i). The principle of Strand-seq is based on template strand identity of sister chromatids generated during DNA replication. During mitosis, each daughter cell inherits one sister chromatid from each parental homologue (Fig. 1A, ii). By sequencing only the original template strand of the inherited chromatids, we can distinguish both homologues in a single cell as either two Crick template strands (CC), two Watson templates (WW) or a combination of Watson and Crick templates (WC) (Falconer et al. 2012; Hills et al. 2013; Sanders et al. 2016) (Fig. 1A, iii). Consequently, when a cell inherits a chromosome as WC, the parental haplotypes for that chromosome can be readily distinguished (Fig. 1A, iv). This allows the variant alleles found in short sequencing reads of Strand-seq libraries to be phased along entire chromosomes, generating haplotypes that span centromeres, reference genome gaps, and regions of homozygosity. By pooling data of multiple Strand-seq libraries from cells that inherited a chromosome as WC, accurate and dense linkage maps of the two parental haplotypes for that chromosome can be achieved.
Figure 1: Direct whole-chromosome haplotyping using single cell template strand sequencing (Strand-seq).

A) (i) Two homologous chromosomes, one originating from the mother (light red) and one from the father (light blue), are shown. Each homologue is composed of a positive template strand (Crick, blue) and a negative template strand (Watson, orange) strand. (ii) Cells incorporate BrdU during DNA replication generating hemi-substituted sister chromatids containing one BrdU-negative template strand (solid line) and one BrdU-positive newly-synthesized strand (dashed line). (iii) Segregation of sister chromatids in two daughter cells follows the depicted combinations of maternal and paternal template strands. The newly-formed DNA strands containing BrdU are selectively removed in daughter cells during library preparation, such that only the original template DNA strands are sequenced. Read density along a chromosome is plotted as horizontal bars. (iv) When daughter cells inherit one Crick and one Watson template strand for a particular chromosome, we can use strand directionality to directly assign all reads to separate haplotypes.

B) Example of a single cell Strand-seq library, generated from HapMap cell line NA12878. Each chromosome is represented as a vertical ideogram and the distribution of directional sequencing reads is represented as horizontal lines along each chromosome, with Watson in orange, and Crick in blue. WC regions that were selected for haplotype phasing are highlighted by red bars.

To evaluate haplotype phasing using Strand-seq, we generated sequencing libraries from an extensively studied HapMap family trio (The International HapMap Consortium 2007; The International HapMap 3 Consortium 2010) (see Methods, Section 1). We selected the child (NA12878) for our initial analysis because this individual was previously phased using parental genotype information and can therefore serve as a reference to assess the validity and precision of our approach. The Strand-seq library for a single NA12878 cell is illustrated in Figure 1B. Within this single cell, reads that aligned to the reference assembly (see Methods, Section 2) covered ~5% of the genome and half of the genome was inherited as WC and thus suitable for phasing (Fig. 1B, red bars). Using single nucleotide variants (SNVs) listed in the HapMap reference for NA12878, we phased 77,717 variant alleles in
this single cell (1.34% of reference SNVs), with 99.3% of the phased SNVs matching the reference haplotypes. This result illustrates that Strand-seq can be used to rapidly generate highly accurate chromosome-length haplotypes from single cells.

**Building whole genome haplotypes from multiple single cell Strand-seq libraries**

In order to build more complete whole genome haplotypes, Strand-seq data from multiple cells were combined (**Fig. 2, i**). Single cell haplotypes can easily be established by separating W and C specific alleles (**Fig. 2, ii**). Each single cell library samples the genome in a random fashion. By combining Strand-seq data from multiple cells, subsets of phased SNVs can be compiled into a dense consensus haplotype (**Fig. 2, iii and iv**). For this purpose, we developed a Strand-seq phasing algorithm and analysis pipeline, called ‘StrandPhase’ (**see Chapter 2**). Briefly, all WC regions are first identified within each individual cell, and SNVs present on each template strand are phased to build single cell haplotypes. Then, StrandPhase iteratively adds the phased variants from each single cell into two consensus haplotypes based on the best concordance. Accordingly, our algorithm concatenates haplotype information from multiple single cells, reinforcing and validating the phased variants in a consensus haplotype for each homologue (**Fig. 2, v**).

![Figure 2: Phasing of multiple single cell Strand-seq libraries.](image)

(i) Multiple single cells are sorted and processed by our library preparation pipeline to prepare Strand-seq libraries. (ii) In every single cell WC chromosomal regions are identified and homologue specific alleles are recorded. (iii) Single cell haplotypes for a single chromosomes serves as an input for phasing pipeline. (iv) Single cell haplotypes are concatenated together based on the best overlap of haplotype specific SNVs. (v) Consensus haplotypes are reported as the best consensus from all single cells.
To evaluate the performance of our analysis pipeline, we selected 183 Strand-seq libraries derived from NA12878. All Strand-seq libraries were preselected based on read depth and coverage distribution in order to avoid phasing errors introduced by low quality libraries (Fig. 3). Using StrandPhase this data was used to build two consensus haplotypes, each representing a phased parental homologue inherited by the child (NA12878) who had previously been extensively phased using parental information as a part of a known HapMap family trio. In a family trio, the child can be unambiguously phased under the assumption that at least one parent is homozygous for a given variable site. Therefore, we used the reference HapMap haplotypes of a child as a gold-standard to assess the validity and precision of our approach.

<table>
<thead>
<tr>
<th>A</th>
<th>Good Strand-seq library: High reads/Mb, even coverage profile, low background reads, no structural rearrangements like CNVs or aneuploidy.</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Moderate Strand-seq library: Lower reads/Mb, less even coverage profile, low background reads, no structural rearrangements like CNVs or aneuploidy.</td>
</tr>
<tr>
<td>C</td>
<td>Bad Strand-seq library: Low reads/Mb, uneven coverage profile, high background reads, presence of structural rearrangements like large segments of CNVs.</td>
</tr>
</tbody>
</table>
Figure 3: Quality criteria for single cell Strand-seq library.
Criteria for 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 Chapter 2, BreakPointR). 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. Such WC portions were also selected for our analysis.

Across all 183 libraries, the aligned reads covered a total of 2,156,208 SNV positions, representing 74.6% of the variants listed in the HapMap reference (Table 1). Of the all identified variants, 1,730,627 SNV alleles were assigned to consensus haplotype 1 (Child H1) and 1,729,512 SNV alleles to consensus haplotype 2 (Child H2) (Fig. 4A), yielding a median distance between all phased alleles of 622 bp (1309 bp for heterozygous alleles). As we increased the number of cells analyzed, SNV coverage increased and distance between subsequent SNVs decreased (Fig. 4C inset), eventually reaching saturation. Next, we compared our haplotypes to the HapMap reference and found 99.3% of our phased SNV alleles concordant with the reported haplotypes (Fig. 4C). The long-range information of Strand-seq data generated haplotypes spanning centromeres and reference assembly gaps. In addition to continuous stretches of haplotypes, we also observed smaller haplotype switches (Fig. 4C, black asterisks). These switches most likely represent homozygous inversions in these regions (Sanders et al. 2016).

Despite the accurate phasing of SNVs spanning every chromosome in the genome, we found 23,782 alleles (0.7%) that were discordant to the HapMap reference. Strikingly, 52.9% of these discordanences were observed in more than one cell in our dataset, supporting the confidence of our allele phasing (Fig. 4B). Because the likelihood of random PCR or sequencing errors occurring at the same genomic position in the same homologue in multiple independent libraries is very low, we propose that discordant phasing at these SNV positions represent either errors in the HapMap reference, polymorphic inversions or somatic mutations in the HapMap cell lines.
Table 1: Summary of sequencing data for each individual sequenced using Strand-seq.
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.
Figure 4: Accurate and dense whole-genome haplotypes are built from multiple single cell Strand-seq libraries.

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. C) Assembled haplotypes of the child derived from 183 Strand-seq libraries. Chromosome ideograms illustrate 151,700 high confidence (covered in more than 1 cell) heterozygous SNV positions phased from Strand-seq data and compared to the HapMap reference. The consensus haplotypes determined by Strand-seq, are depicted for each chromosome, with each SNV represented by a vertical line and color-coded based on whether it matched the child’s reference homologue 1 (brown) or homologue 2 (yellow) listed in the HapMap reference. The contiguous haplotypes extend the whole length of each chromosome, spanning centromeres and reference assembly gaps (white blocks). Discordant alleles that did not match either reference haplotype are shown in red. Asterisks – points to short localized switches in haplotypes that were confirmed as homozygous inversions. Inset (Black line) The percentage of HapMap reference SNVs covered; and (Red line) the median distance between these SNVs is plotted for various numbers of libraries (25, 50, 100, 150), randomly sampled from the entire data set of 183 cells.

Secondary validations of Strand-seq haplotypes

To further confirm the precision of haplotype reconstruction using Strand-seq, we tested haplotyping discordances between Strand-seq and HapMap phasing using publicly-available long-read PacBio RNA-seq data from the same (NA12878) individual (The International HapMap Consortium 2007). We chose PacBio data from RNA-seq because sequenced transcripts holds long linkage information spanning multiple exons within a single PacBio read. This is because, often long, intronic regions are not part of sequenced transcripts (Fig. 5A). For this analysis we cross-referenced the alleles segregating together on each transcript (cDNA molecule) with both the Strand-seq and HapMap-derived haplotypes (Fig. 5B, see Methods Section 3). We found nearly perfect concordance (99.2%) of the PacBio dataset to our haplotypes while its concordance to HapMap reference was only 94.7%. Similar levels of concordance between Strand-seq and PacBio reads were also found for parental haplotypes (NA12981 and NA12892) (Table 2). These results provide an additional level of evidence that our child’s haplotypes are phased at high accuracy.
Figure 5: Validation of Strand-seq haplotypes using PacBio reads.
A) Exons (brown and green rectangles) of hypothetical genes A and B are shown to illustrate long linkage information embedded in long PacBio reads from RNA-seq. B) Illustrates how transcriptome PacBio data was overlaid on the top of the Strand-seq haplotype backbone. Every PacBio read matching Strand-seq haplotypes was assigned as concordant and vice versa every read not matching Strand-seq haplotypes as discordant.

Table 2: 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.

<table>
<thead>
<tr>
<th></th>
<th>NA12878</th>
<th>NA12891</th>
<th>NA12892</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistent with Strand-seq (%)</td>
<td>99,2%</td>
<td>98,9%</td>
<td>99,8%</td>
</tr>
<tr>
<td>Consistent with HapMap (%)</td>
<td>94,7%</td>
<td>94,8%</td>
<td>93,9%</td>
</tr>
</tbody>
</table>

In addition to PacBio comparison we set to compare our haplotypes to principally different single cell haplotyping study (Fan et al. 2011). In this study, Fan et al. developed a new haplotyping approach termed direct deterministic phasing. With this approach single homologous chromosomes originating from single cells can be captured using microfluidic device, amplified by multiple displacement amplification (MDA) and then sequenced as separate entities (Fan et al. 2011). This way genetic variants (alleles) residing on a single homologue can be sampled. However, chromosome pre-amplification step using MDA is known to be error-prone
and therefore can be source of errors in resulting haplotypes (De Bourcy et al. 2014). Three way comparison between Strand-seq, HapMap and Fan et al. Haplotypes is summarized in Table 3. Overall our results shows high agreement between all three datasets. Of note our results show slightly better agreement between Strand-seq haplotypes and HapMap reference haplotypes than in comparison to Fan et al (2011). This slightly higher level of disagreement for Fan et al. (2011) haplotypes might be result of biases introduced during genome pre-amplification step using MDA.

<table>
<thead>
<tr>
<th></th>
<th>Concordances (%)</th>
<th>Discordances (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strand-seq vs Fan et al.</td>
<td>98.7%</td>
<td>1.3%</td>
</tr>
<tr>
<td>Strand-seq vs HapMap</td>
<td>99.3%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Fan et al. vs HapMap</td>
<td>98.8%</td>
<td>1.2%</td>
</tr>
</tbody>
</table>

Table 3: 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.

Besides SNVs we evaluated specificity of meiotic recombination localized by Strand-seq in comparison to Fan et al (2011). We have found exactly the same number of meiotic recombination events for maternal (38) and paternal homologues (26), inherited in the child, as in Fan et al (2011). (Fig. 6A,B). All, but two, defined meiotic recombinations regions overlapped with regions defined in Fan et al (2011). (Fig. 6A,B). Non-overlapping meiotic recombinations regions were 575kb and mere 6kb apart from each other (Fig. 6A,B red crosses). The larger 575kb difference was caused by low density of heterozygous SNVs in a given region of meiotic recombination. Details on mapping meiotic recombination events using Strand-seq can be found in Chapter 4.

At last, we set to compare phasing accuracy for selected deletions phased as a part of Fan et al. study (Table 4, Fig. 6D). All deletion phased in Fan et al. (2011) were successfully localized in Strand-seq homologue specific read densities. Phase of all deletions matched the phase reported in Fan et al. (2011) (see Methods, Section...
Taken together, these results yet again confirm that we can generate accurate haplotypes in the absence of generational (parental or population) information, which represents a major advance in the field.

Figure 6: 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 Table 4). Horizontal panels represent separate homologues of each individual in the trio. Vertical colored lines represent read coverage in homologue specific BAM files (see Chapter 2, Fig 6A). Dotted lines shows boundaries of heterozygous deletions with breakpoint coordinates at the top.
Comparison of Strand-seq phasing with hybrid phasing described by Mostovoy et al.

Lastly, it is also important to note that Strand-seq phasing relies on a reference genome to map directional reads, and therefore alleles that are not represented in the reference genome, including new duplications, may not be phased. Moreover, balanced rearrangements like inversions cause directional reads to map in opposite directions to the reference genome and are visible as switches in resulting haplotypes (Fig. 4C, black asterisks). To overcome this, others have used hybrid phasing approaches based on de novo assembly to improve haplotype accuracy (Pendleton et al. 2015; Mostovoy et al. 2016). 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.

To explore how Strand-seq relates to hybrid phasing, we compared our phasing with the large 64Mb scaffold assembled for Chromosome X by Mostovoy et. al. (2016). 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 64Mb 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 (Fig. 7A, see Methods, section 5). This revealed Super-scaffold 52 was composed of shorter haplotype blocks that were not linked continuously from start to end, (Fig. 7B). 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 (Fig. 7C, 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.

Figure 7: 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 Genomics. 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 Genomics phasing disagree. Black arrowhead – points to putative haplotype switch error.
In contrast, phased block number 9 (Fig. 7C, 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 (Fig. 7C, 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 Methods, section 5). 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 (Fig. 7D), suggesting that phased block number 9 was incorrectly phased using the hybrid phasing 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 to better refine de novo assemblies to build the most accurate haplotypes for an individual. Moreover we showed that despite relying on the reference genome assembly our phasing approach keeps high phasing accuracy.

METHODS
1. Raw data production
Cells and cell culture: Epstein-Barr virus (EBV) transformed B-lymphocyte cell lines GM12878, GM12891, and GM12892 were obtained from the Coriell Institute for Medical Research (Camden, NJ, USA). The pedigree of all cell lines is UTAH/MORMON from USA, which were part of the International HapMap Project (The International HapMap Consortium 2007; The International HapMap 3 Consortium 2010). Cells were cultured in RPMI 1640 medium (Gibco) supplemented with 15% FBS (Sigma Aldrich) in 37°C at 5% CO₂. For Strand-seq, BrdU (40 or 100 µM, final) was added to exponentially growing cells for 24 hours.

Single cell sorting: Cells were harvested, and nuclei isolated by resuspending in nuclear isolation buffer (100mM Tris-HCl pH7.4, 150 mM NaCl, 1 mM CaCl₂, 0.5 mM MgCl₂, 0.1% NP-40, 0.2% BSA). In each sample, cells cultured without BrdU were added as an internal control for Hoechst fluorescence. Nuclei were stained with Hoechst-33258 and Propidium iodide (PI) by adding both to the isolation buffer at final concentration of 10 µg/mL and incubating on ice for 30 minutes. Nuclei of cells that underwent a cell division in the presence of BrdU were sorted based on low Hoechst fluorescence (quenched by BrdU in DNA) and PI (gated on G1 phase), using a MoFLo Atrios cell sorter (Beckman Coulter) and deposited into 96-well skirted PCR plates (4Titude) containing 5µL/well freeze medium (Pro-Freeze CDM Freeze Medium (Lonza) containing 15% DMSO).
CHAPTER 3

Library construction: Library preparation was performed using modified versions of the previously described protocol (Falconer et al. 2012). To scale for production on a Bravo Automated Liquid Handling Platform (Agilent) the enzymatic reactions were performed in smaller volumes, while keeping buffer and enzyme concentrations at the same levels. DNA clean-up steps were performed using AMPure XP paramagnetic beads (Agencourt AMPure, Beckman Coulter). After adapter ligation and 17 PCR cycles, two consecutive AMPure bead clean-ups were performed using a 1.2x bead volume.

Next generation sequencing: Libraries were pooled for sequencing and 250 to 300 bp size range fragments were purified using 2% E-Gel Agarose EX-Gels (Invitrogen). DNA quality was assessed on a High Sensitivity dsDNA kit (Agilent) using the 2100 Bio-Analyzer (Agilent), and DNA was quantified on the Qubit 2.0 Fluorometer (Life Technologies). For sequencing, clusters were generated on the cBot and paired-end 100 bp or single-end 50 bp long reads were generated using the HiSeq2500 sequencing platform (Illumina) following the manufacturer’s instructions. For 50 bp and 100 bp long reads, 192 and 96 single cell libraries were pooled together, respectively and sequenced in one lane of the rapid run flow cell. Each plate included two 10-cell controls and two zero-cell controls.

2. Raw data processing

The single cell raw sequencing data were demultiplexed based on the library-specific barcodes and converted to FASTQ files using Illumina standard software (bcl2fastq, version 1.8.4). The resulting reads were mapped to the human reference genome NCBI36/hg18 using Bowtie 2 aligner (Langmead and Salzberg 2012) (version 2.2.4). After alignment, reads were sorted using SAMtools (Li et al. 2009) (version 0.1.19) and duplicate reads were marked using BamUtil (version 1.0.3). All Strand-seq libraries were pre-filtered to avoid haplotype errors arising from low quality data. For this, we excluded libraries with less than 50 reads/Mb, more than 5% level of background reads and/or exhibiting excessive genomic rearrangements, aneuploidy events or uneven coverage (Fig. 3). BAM files passing our quality criteria served as an input for our haplotyping pipeline.

3. PacBio and Strand-seq cross-validation

We incorporated PacBio data using a three-stage approach. First we mapped PacBio reads to the human transcriptome (NCBI36/hg18, Ensembl release 54) using bwasw module implemented in BWA aligner (version 0.7.12.) (Li and Durbin 2010).
Second, for every PacBio read we recorded the specific variant at each position listed in the HapMap reference. Lastly, we added strand information to each allele based on the mapping directionality. To directly compare our haplotypes with the PacBio dataset we selected all PacBio reads that overlapped with at least two heterozygous positions in our Strand-seq haplotypes (Fig. 7). We filtered out reads containing SNVs with a base quality less than 20. Next we calculated the percent of phased PacBio reads that matched the phase we found for our haplotypes, to test the level of concordance between these datasets. To assess non-random concordance, we randomly shuffled the SNVs between the H1 and H2 Strand-seq haplotypes and counted the number of concordant and discordant reads again. Reshuffling eliminated the concordance between Strand-seq and PacBio data. PacBio data used for this analysis were downloaded from the SRA database. Accession numbers: SRR1163655 (NA12878), SRR1163657 (NA12891), SRR1163658 (NA12892) (Tilgner et al. 2014).

4. 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 (see Chapter 4 Table 1) 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, see Chapter 2 Fig. 6A). 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. (Table 3).

5. 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.

**DISCUSSION**

The results presented here show that Strand-seq, together with StrandPhase, is a novel single cell haplotyping method that retains linkage information along whole chromosomes. Because Strand-seq does not involve whole genome amplification (WGA) prior to library preparation, the sequence bias and allelic drop-out introduced by PCR amplification are reduced allowing extraction of highly accurate phase information from single cells. By compiling SNVs across multiple Strand-seq libraries, we were able to reconstruct whole genome haplotypes without generational information. Each SNV is independently sampled in multiple single cell libraries, allowing us to directly cross-validate variant calls made in a sample and rapidly build highly accurate consensus haplotypes. Highlighting this, our results recapitulate
the HapMap project reference haplotypes without statistical inference, population or pedigree data, demonstrating the strength of our approach for clinical studies. Additionally we have confirmed high accuracy of our approach in comparison to various data sources including PacBio sequencing as well as published haplotypes from other single cell study. Importantly, we have proved that accuracy phasing using Strand-seq, despite its dependency on a reference genome assembly, is not compromised by imperfections in the reference genome assembly.

Taken together, we propose that Strand-seq is a unique tool to completely phase individual genomes, map meiotic recombination events in family trios and explore haplotype structure in single cells. By avoiding pre-amplification Strand-seq offers unmatched accuracy over other sequencing-based phasing techniques.

Data access
Strand-seq libraries selected for this study have been submitted to the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under the accession number: PRJEB14185.

Acknowledgements
We thank Nancy Halsema, Inge Kazemier and Karina Hoekstra-Wakker for technical help. Financial support for these studies was provided by a European Research Council Advanced grant (ROOTS-Grant Agreement n. 294740) to P.M.L.

Author contributions
D.P. performed data analysis and implemented the phasing algorithm. D.P. and A.S. wrote the manuscript, N.v.W. cultured cells and prepared Strand-seq libraries, A.S., E.F., M.H. and V.G. helped with data analysis and development of bioinformatics approaches, D.S. and M.B helped with sequencing of Strand-seq libraries, M.B., V.G. and P.M.L. designed experiments and helped with writing of the manuscript.