university of groningen

University Medical Center Groningen

University of Groningen

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):
Porubský, D. (2017). *Haplotype resolved genomes: Computational challenges and applications*. University of Groningen.

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.

Download date: 15-05-2021

CHAPTER 6
General discussion

Basis for review:

Genetic variation of diploid genomes unveiled.

David Porubsky[1], Ashley D. Sanders[2], Victor Guryev[1], Peter M. Lansdorp[1,3,4]

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

Undoubtedly, completion of the human reference genome has helped to better understand the molecular basis of many human traits and diseases. While it is now relatively easy to chart unique variation of individual human genomes using state-of-the-art sequencing technologies, it remains challenging to distinguish what portion of this variation belongs to the maternal and paternal copies of the genome.

Haplotypes-resolved genomes hold invaluable information applicable in many areas of population and clinical genetics. For instance, haplotypes helped us to better understand the relationship between genetic variants and diseases (Tewhey et al. 2011). From a clinical perspective, haplotype-resolved personal genomes are important to assess the context of mutations on homologous chromosomes (cis – on the same homologue, trans – on different homologues) in order to predict the effect of such potentially disease-causing mutations in offspring (Kitzman et al. 2011; Hoehe et al. 2014; Roach et al. 2010). Furthermore, haplotypes are required to determine the parental origin of *de novo* mutations in a child (Conrad et al. 2011; Kloosterman et al. 2015) and map meiotic recombination events. Lastly, haplotypes are important to study loss of heterozygosity in cancer (Huang et al. 2007), as well as allele-specific events like gene expression (Kasowski et al. 2010).

Due to the above-mentioned reasons our goal should be, for every genome sequenced, to assign maternally and paternally inherited genetic variation to the corresponding homologous chromosomes in the form of haplotypes. Ideally, one would like to obtain highly accurate, genome-wide haplotypes for every single homologous chromosome. However, available phasing methods differ widely in the completeness, accuracy and the length of provided haplotypes (**see also Chapter 1**). Especially, phasing of alleles across the entire length of all chromosomes is currently very challenging unless both parents of the individual are also sequenced (Kitzman et al. 2011; Amini et al. 2014) or specialized experimental techniques are applied (Ma et al. 2010; Brown et al. 2012; Fan et al. 2011). In this thesis, I described a genome-wide phasing technique based on single cell sequencing of inherited template strands called Strand-seq (Porubsky et al. 2016).

Strand-seq is a sequencing technique in which parental DNA template strands from single cells are sequenced and thus the structure and parental identity of individual homologues is preserved. The technique was originally developed to track the inheritance of sister chromatids in daughter cells after one cell division (Falconer et al. 2012). The power of Strand-seq lies in its ability to distinguish parental homologues based on the directionality of sequencing reads mapped to the reference genome. Assuming random segregation of sister chromatids following

cell division, around 50% of single daughter cells will generate reads that map in both directions ('+' - Crick and '-' - Watson) of the reference genome for a given chromosome. In such cases, all SNVs present in reads mapping to either the Watson or the Crick strand of the reference genome will, in theory, be derived from either parent. Unfortunately, current bioinformatics pipelines are not suited to fully exploit template strand directionality to phase diploid genomes and map genomic rearrangements. The design and development of new bioinformatic analytical tools, capable to handle the specific nuances of Strand-seq data, has been the most important objective and challenge of my doctoral work.

In order to track template strand changes in single cells we have developed BreakPointR, a software package written in R, specifically tailored to handle directional reads stored in aligned BAM files. Furthermore, I have implemented phasing pipelines (StrandPhase and StrandPhaseR), which are able to exploit the haplotype information present in every single cell Strand-seq library. We believe that these tools will be indispensable for future high-throughput analysis of Strand-seq data (**see Chapter 2**). Future plans are focused on the implementation of the analytical tools we have developed into a single toolbox. This toolbox, named SingleCellToolkit will bring to potential users an easy to use environment to run their own single cell analysis pipelines.

We have applied these tools (**see Chapter 2**) in order to assemble haplotypes from multiple Strand-seq libraries, and we were able to reconstruct whole-genome haplotypes of a single individual (NA12878) without the parental information or statistical inference. We recapitulated reference haplotypes from HapMap project with high precision (concordance 99.3%) along the whole length of all homologous chromosomes for the whole family trio (child-NA12878, father-NA12891, mother-NA12892). This allowed us to map all meiotic recombination events in the child of this family with high resolution (median range ~14kb), what was 3-fold better than in other single cell phasing study (Wang et al. 2012). In total, we mapped 64 recombination events, from which 38 on the maternal and 26 on the paternal homologues of the child. These numbers are in close agreement with results from previous studies (Fan et al. 2011) (**see Chapter 3**).

Besides SNVs, larger structural variants like deletions, insertions as well as balanced rearrangements like inversions can also be phased. To map such genetic variants we exploit the capability of Strand-seq to phase the directional reads for specific chromosomes in every single cell and split them accordingly into two separate high density haplotypes. Such reads can then be stored in two separate BAM files

what allows an easy detection of larger structural variants (> 5kb) using BreakPointR or other copy number detection tools. Importantly, balanced rearrangements which are known to be notoriously difficult to map using current technologies, can be directly visualized, mapped and phased using Strand-seq and our analysis pipelines, a major interest for basic and clinical research (**see Chapter 4**).

Unlike other haplotyping techniques, Strand-seq retains the ability to track haplotype differences at the single cell level. This is a valuable feature to study heterogeneous cell populations, like cancer. We have observed a number of localized regions with one haplotype being converted to the other, resulting in loss of heterozygosity (LOH) at a specific genomic region within a single cell. Notably, this loss was not due to a deletion, since the read depth analysis supported diploid status of this locus. The observed LOH patterns suggest that mitotic recombination events might be common between homologous chromosomes (Moynahan and Jasin 2010) at a frequency of ~ 0.06 events per cell. The possibility to explore haplotype differences at the single cell level will be a significant advantage for studies of heterogeneous cell populations, such as cancer cells (**see Chapter 4**).

Another important advantage of Strand-seq is that whole genome amplification (WGA) prior to library preparation is avoided. As a result, genome coverage bias and allelic drop-out introduced by PCR amplification are minimized allowing assembly of highly accurate haplotypes. Of note, each SNV is independently sampled in multiple single cell libraries, allowing us to directly cross-validate obtained variant calls and build highly accurate consensus haplotypes.

Needless to say, like every haplotyping technique, also Strand-seq has its limitations that have to be taken into account. One of them is the requirement for BrdU incorporation in dividing cells as the input for library preparation in order to remove only newly synthesized strands. Another limitation of Strand-seq libraries is the low and non-uniform genome coverage. One factor that plays a role in coverage non-uniformity is MNase digestion during library preparation. Not all genomic DNA is bound to nucleosomes and the effectiveness of MNase digestion may also vary depending on chromatin accessibility across the genome. To account for this and to lower the effect of coverage inconsistencies throughout the genome we have developed and implemented read-based genome binning strategies in BreakPointR package. Moreover, low genomic coverage of single cell libraries results in incomplete set of alleles phased in final consensus haplotypes assembled from single cells.

Additionally, we have shown that low-density haplotypes obtained by

Strand-seq can be augmented by other sequencing methods, such as short- and long-read whole genome sequencing (WGS) technologies. We demonstrated that usage of alternative data decrease the number of Strand-seq libraries needed for phasing while achieving long-range haplotypes and high density of phased alleles. In **Chapter 5** we have demonstrated that a combination of 10 Strand-seq libraries with 10x PacBio coverage is sufficient to phase the majority of alleles into highly accurate and chromosome-length haplotypes. A similar outcome was achieved with 30x coverage of short Illumina reads in combination with ~ 40 Strand-seq libraries. We argue that such integrative phasing approaches will lower the costs and labor requirements needed to phase diploid genomes in the future and, in turn, will bring haplotype-resolved genomes closer to routine practice. We propose a prominent role for Strand-seq in this task since this technique can provide a global haplotype scaffold for other sequencing technologies.

The last limitation and a source of biases is the reference genome itself. The 'reference' genome is used as a scaffold to align directional Strand-seq reads. Therefore, any imperfections in the reference genome assembly will result in a range of mapping artifacts that could be wrongly interpreted as real structural variants. The typical example of such artifacts are misplaced contigs (Falconer et al. 2012) and mapping biases caused by low complexity regions in the reference genome. Such cases have to be treated with special attention and tools for the processing of Strand-seq data have to be able to localize them. One way to uncover such biases is to look at the frequency of such events in the population of cells (Sanders et al. 2016). We have proved that accuracy of long-range phasing using Strand-seq is not compromised despite its dependency on a reference genome assembly (**see Chapter 3**). However, this might not apply for shorter haplotype stretches or complex genomic variants that lie in the vicinity of repetitive regions of the genome, such as segmental duplications.

Nevertheless, most of the re-sequencing projects are based on the alignment to the reference genome assembly which does not represent genome of the individual in question but rather an 'patchwork' genome obtained by sequencing hundreds of individuals. This might conceal important portion of individual's genome variation and results in an inadequate understanding of the genome architecture. The biggest obstacle to assemble genomes *de novo* is the abundant repetitive sequence present in the genome. Especially short sequencing reads are unable to resolve repeats that are longer than the read itself. In contrast longer reads (PacBio or Oxford NanoPore) have the capabilities to resolve more repetitive regions (Chaisson, Wilson, and Eichler 2015). The ultimate challenge is *de novo* assembly of haplotype-resolved individual

genomes, what will open up new opportunities to study personal structural variation across diverse human populations. As long-read sequencing technologies continue to mature, and as low-input amplification methods improve, we anticipate that these direct haplotyping methods will be refined over the next few years, becoming more cost-effective and increasingly adoptable to automation, multiplexing and routine use.

To bring haplotype-resolved *de novo* assembly of diploid genomes into practice, new algorithmic solutions together with efficient data structures have to be developed. Most genome assembly programs internally use a graph representation to build the assembly, but ultimately produce a flattened structure for use by downstream tools (Zerbino and Birney 2008; Schatz, Delcher, and Salzberg 2010). Currently a graph structure is a natural way to represent a population-based genome assembly, with branches in the graph representing all variation found within the individual genomes (Church et al. 2015). Recently, formal proposals for representing a population-based reference graph have been described (Marcus, Lee, and Schatz 2014; Dilthey et al. 2015; Paten, Novak, and Haussler). The established Global Alliance for Genomics and Health (GA4GH) together with Pangenome consortium (Marschall et al. 2016) is leading an effort to formalize data structures for graph-based reference assemblies, but it will likely take years to develop the infrastructure and analysis tools needed to support these new structures and see their widespread adoption across the biological and clinical research communities (Church et al. 2015).

Our better understanding of human genome variation should also be reflected in updated models and structures currently used to represent haplotype-resolved individual genomes. Essentially, haplotype-resolved genome of an individual can be viewed as the smallest pan-genome structure comprised from two haploid genomes one inherited from the mother and the other from the father. Such structure is meant to provide a catalogue of physiologically and pathologically occurring variations within a single individual. This will allow unbiased comparison of genomes at the population scale and will ensure that unique set of variants of each haploid genome will be considered.

We conclude that Strand-seq is a unique and powerful approach to completely phase individual genomes and map inheritance patterns in families, while preserving haplotype differences between single cells. We expect that future studies on haplotypes will benefit from the combination of Strand-seq and long-read sequencing technologies to assemble complete and chromosome-length haplotypes.

As single cell sequencing becomes more and more accessible, we anticipate that Strand-seq haplotyping will have an important contribution to *de novo* assembly of haplotype-resolved personal genomes and thereby greatly facilitate studies of genomic variants in human health and disease.