A BIOINFORMATIC WORKFLOW TO ANALYZE SINGLE CELL TEMPLATE STRAND SEQUENCING DATA

by

Carl-Adam Mattsson

B.Sc., The University of Kristianstad, 2014

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE
in
THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

April 2020

© Carl-Adam Mattsson, 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, a thesis entitled:

A Bioinformatic Workflow For Analyzing Single Cell Template Strand Sequencing Data

submitted by	Carl-Adam Mattsson	In partial fulfillment of the requirements for
the degree of	Master of Science	
in	Bioinformatics	

Examining Committee:

Peter Lansdorp, Department of Medical Genetics, UBC

Co-supervisor

Martin Hirst, Microbiology & Immunology, UBC

Co-supervisor

Intan Schrader, Department of Medical Genetics, UBC

Supervisory Committee Member

Inanc Birol, Medical Genetics, UBC Genome Sciences Centre, BCCA, UBC

Additional Examiner

Abstract

Structural variants (SVs) contribute greater diversity at the nucleotide level between two human genomes than any other form of genetic variation and are three-fold more likely to correlate in genome-wide association studies (GWAS) than single nucleotide variants (SNVs). Using short-read, high-throughput sequencing technologies to uncover such variation has proven to be troublesome and the methods to detect SVs depend on indirect inferences. However, while larger (>5kb) copy number variations (CNVs) could be characterized using read-depth-based algorithms, this approach often fails for smaller and balanced events. Another fundamental problem for detection of SVs from short-read sequencing is inherent to the predominant data type and typical SV detection algorithm that is effective in unique sequences often fails within complex genomic regions, which have been proven to be highly enriched for SVs. In addition, most SV discovery methods do not indicate the haplotype-origin for a given SV and require parental sequencing for this information. For a more complete description and interpretation of human genomic information in relation to phenotypes such as e.g. cancer predisposition and response to therapies, it will, therefore, be necessary to arrange sequence data into parental haplotypes and ascertain polymorphic inversions with respect to such haplotypes. All this can be achieved using Strand-seq. Strand-seq complements other sequencing approaches by providing crucial information about the genetic make-up of individuals that cannot be obtained in any other way. To make Strand-seq available for human studies worldwide is an immense challenge. Library construction, as well as data analysis, needs to be further developed, integrated and made user-friendly to allow accurate and rapid interpretation of results. Here we present a custom bioinformatics pipeline for analyzing Strand-seq data that streamlines the workflow of raw sequence read alignment, putative variant calling, variant call refinement and haplotype assembly by integrating current available Strand-seq specific tools. In addition, relevant metric data are compiled and visualized, ensuring and reinforcing the potential of Strand-seq as a robust sequencing method for uncovering clinically significant SVs and the assembly of WGH without additional parental genomic data.

Lay Summary

Given the implication of genetic variation in relation to disease predisposition and progression, the importance of such variation should not be overlooked. Recent high-throughput sequencing technologies have shown to be successful in uncovering such variation, particularly larger variants. However, our ability to decipher smaller, more complex variants remains limited. Highlighting the need for accurate and robust technologies/workflows to expose such variation. By implementing a single-cell sequencing technology called Strand-seq, we can improve our understanding in relation to such genetic variation. Here we present a custom bioinformatic pipeline for analyzing Strand-seq sequence data. Developed pipeline integrates current available Strand-seq specific tools in a user-friendly environment to allow accurate, rapid and reproducible analysis of structural genetic variation. In addition, relevant metric data are compiled and visualized in a standardized way, ensuring and reinforcing the potential of Strand-seq as a robust sequencing method for uncovering clinically significant genetic variation.

Preface

This thesis comprises unpublished work performed by the author. All analyses were performed by the author as was the development of the bioinformatics pipeline. Previously sequenced Strand-seq data, used to verify the performance of our pipeline, was obtained from European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession number PRJEB14185. Pre-assembled 10x Genomics haplotypes produced on the Chromium platform with Chromium Genome v1 reagents, sequenced on an Illumina HiSeq X Ten and processed with LongRanger 2.1.0 were downloaded from 10X genomics website (https://support.10xgenomics.com/genome exome/datasets/NA12878_WGS_210) and filtered for heterozygous SNVs. Interpretation of structural variant analysis results were the collaborative effort of Peter Lansdorp, Diana Spierings, Victor Guryev. Figure 1.1 was adapted from *Genome Research*, 26(11), A. D. Sanders, M. Hills, D. Porubský, V. Guryev, E. Falconer, and P. M. Lansdorp, p 3. Figure 1.2 was reproduced from *Genome Med,* 5(9), M. Hills, K. O'Neil, E. Falconer, R. Brinkman, and P. M. Lansdorp, p1. Figure 1.3 was adapted from *Genome Research*, 26(11), A. D. Sanders, M. Hills, D. Porubský, V. Guryev, E. Falconer, and P. M. Lansdorp, p 4.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AS denovo assembly

BAM binary alignment map

BP breakpoint

C crick

CNV copy number variation

GWAS genome-wide association study

HGP Human Genome Project

HMM hidden Markov model

HTS high-throughput sequencing

LOH loss of heterozygosity

ORCA omics research container architecture

RD read depth

ROI region of interest

RP read pair

SAM sequence alignment map

SCE sister chromatid exchange

SMRT single-molecule real-time

SNP single nucleotide polymorphism

SNV single nucleotide variant

SR split reads

SV structural variant

T-ALL t-cell acute lymphoblastic leukaemia

VCF variant call format

W watson

WGH whole-genome haplotype

Acknowledgements

I would like to thank my supervisor Peter Lansdorp for his intransigent intellect, insightful advice, academic support and hospitality. Also, I would like to express my gratitude towards past and present members of the Lansdorp lab and Diana Spierings and her group at ERIBA, Groningen, for providing numerous insightful thoughts throughout my work. Thank you Intan Schrader for the many interesting discussions and for making me feel welcome in Vancouver. Thank you to my committee members, Inanc Birol and Martin Hirst for taking the time to advise me on my thesis. Lastly, I would like to express my sincere gratitude towards Lisa Salomonsson for uncompromised support, guidance and love.

1. Introduction

1.1 Overview

1.1.1 The Human Genome

The completion of the Human Genome Project (HGP) in 2003 was an enormous leap forward in terms of understanding the human genome and how it relates to genotypic and phenotypic variation [1, 3, 5]. In the latter part of the last century sequencing an entire human genome was considered to be a daunting task. This is reflected by the limited number of catalogued organisms with published genome sequences that were available at that time. In general, these organisms harbored relatively small and simple genomes. The challenge when it comes to characterizing the human genome mainly reflects the presence of long repetitive sequences and the fact that the human genome is roughly an order of magnitude larger than the total of all previously studied genomes [4]. However, the first draft of the human genome, published in 2001, was highly imperfect [1, 3]. Only ~90% of the euchromatic genome was covered and multiple errors in the nucleotide sequence and over 250,000 gaps were present [4]. After the publication of the initial draft, our understanding of the human genome has rapidly improved. Already in 2004, the HGP consortium published an updated reference sequence. This time, ~99,7% of the euchromatic genome was represented, interrupted by some 300 gaps and only one nucleotide error per 100,000 bases was present. Interestingly, the gaps consisted of ~28 megabases (Mb) of euchromatic sequence representing large repetitive regions that could not be reliably cloned or assembled, again highlighting the fact that this characteristic of the human genome is a fundamental problem. However, this improved reference material allowed for more accurate inference of gene structure and detection of polymorphisms and mutations [2]. Interestingly, the long regions of repetitive DNA were shown to have unexpectedly high rates of copy-number variation (CNV) within and between species and were frequently involved in large-scale chromosomal rearrangement. Overall, the human genome was found to be way more complex than was previously imagined [2, 3, 4, 6, 7]. Since the completion of HGP many new sequencing technologies, computational methods and increased computational resources have paved the way for an improved reference

sequence which in turn has improved our understanding of the human genome and how it relates to genetic and phenotypic diversity within our species.

1.1.2 Human Genetic Variation

Human genetic variation is typically defined as genetic differences in and among populations and it is understood that no two human genomes are genetically identical [2]. There are many different forms of genetic variation. For example, polymorphism refers to the presence of multiple variants of any given gene in the human population and mutations are describing a change in the DNA sequence that deviates from what is considered normal [9, 2]. In more detail, for a variant to be classified as a polymorphism this variant allele, has to be present in 1 percent (or more) of the population; if the observed allele frequency is <1% the allele is regarded as a mutation. The magnitude of human genetic variation is highly variable, spanning from larger alterations at the karyotype level to single nucleotide changes (i.e. single nucleotide polymorphism (SNP) and single nucleotide variants (SNVs)). In the light of diagnostics and precision medicine, the study of human genetic variation is important for identifying disease-causing alleles and how genetic variation correlates to disease progression and treatment [10, 11].

1.1.3 Single Nucleotide Polymorphism

SNPs are typically defined as a difference in a single nucleotide among species that occurs in at least 1% of the population [9, 2]. To give some perspective on the frequency and implications of SNPs in human populations, the 2,504 individuals characterized by the 1000 Genomes Project, harbored 84,7 million SNPs among them and SNPs are the most common type of sequence variation, occurring every 100 to 300 bases [12]. However, only 3-5% of human SNPs are considered to be functional, meaning causing phenotypic differences between members of the species [13]. Before the completion of HGP, SNPs were considered to be the main source of genetic variation. This turned out to be wrong, as the reference material improved. The consensus in the scientific community now is that structural variants (SVs) are the main contributor to genetic diversity [1, 2, 5, 6, 10, 11].

1.2 Structural Variants

1.2.1 Overview

A structural variant is usually defined as a variant affecting a sequence length of > 50 bp and structural variants include both microscopic and submicroscopic types such as deletions, inversions, copy-number variants, insertions, duplications and translocations [14, 15, 16]. Many of these SVs are directly linked with genetic diseases and some are not [16]. For example, certain variations in the tumour suppressor gene *BRCA1* were shown to lead to an increased risk for developing breast cancer [21, 22]. In general, it's more difficult to detect and characterize SVs compared to SNVs [18, 19, 20]. This is mainly due to current SV detection methods being highly dependent on indirect references such as read-depth and discordant read-pair mapping. Hence, smaller (< 5 kb) SVs and balanced events such as inversions or translocations remain poorly ascertained in most current sequence reports including whole genome sequence data [10, 12, 17].

1.2.2 Types

There are many different types of structural variations, relevant events are described in the next section.

1.2.2.1 Indels

The term indel describes an **in**sertion and/or a **del**etion of nucleotide content in the genome and thus, indels are examples of CNVs [23, 29]. An insertion refers to the addition of one or more nucleotides into a DNA sequence, frequently observed in microsatellite regions due to the DNA polymerase slipping during replication. A deletion describes a mutation where one or more nucleotides are lost during DNA replication or repair. A deletion can span from one nucleotide up to whole chromosomes. If the length of an indel is multiple of 3, it will produce a frameshift mutation (resulting in changing the reading frame). For example, a common micro-indel which results in a frameshift is known for causing Bloom syndrome in the Jewish or Japanese population [25, 26].

1.2.2.2 Inversions and Polymorphic Inversions

An inversion describes a chromosome rearrangement were a segment of a chromosome is reversed end to end due in a single chromosome following breakage and rearrangement within itself [27, 29]. Typically, there are two types of inversions, paracentric that do not include the centromere with both breaks occurring in one arm of the chromosome and pericentric inversions that span the centromere with breaks in each chromosome arm. Inversions are examples of balanced events, were no loss of genetic material is observed, but rather a reorientation. Large inversions (>1 Mb) are typically within the detection-range using cytogenetic approaches, submicroscopic and polymorphic inversions, however, are ill-defined by current human genome sequencing as the majority of these events are flanked by highly repetitive, duplicated sequences. Recent studies show that on average, each genome harbours ~156 inversions, constituting around 23 Mb of inverted genome [10].

1.2.2.3 Duplications

An SV of subtype duplication describes an event where a particular locus is amplified, usually due to errors in the DNA repair and replication machinery as well as through fortuitous capture by selfish genetic elements [28, 29].

1.2.2.4 Translocations

Translocations are another example of a balanced chromosomal abnormality in which a chromosome breaks and a resulting fragment is attached to a different chromosome that also underwent a break event [28, 29].

1.2.3 Methods of Detection

The methods used to detect SVs are typically categorized into array-based (including microarray comparative genome-hybridization) and sequencing-based computational methods [29, 30, 31]. Each detection method has its advantages and disadvantages. For example, Array-based approaches are advantageous for high-throughput analysis but limited by only capturing certain types of SVs, low sensitivity

for small SVs and less accurate breakpoint (BP) calling. To allow for the detection of a broad range of SVs, it is therefore necessary to adopt sequencing-based methods. Sequenced based methods can be further sub-categorized into Read Pair (RP), Read Depth (RD), Split Reads (SR) and De Novo Assembly (AS) approaches. RD (or Read Count) based methods assume a random Poisson distribution of reads from sequencing and SVs are called in regions where divergence from this distribution is observed. For example, genomic regions subject to duplication will have a higher read depth compared to other loci, while deletions would result in lower than average read depth. However, RD based detection methods do not take into account the copy neutral state of balanced SVs such as inversions and translocations [10, 11]. RP based detection methods utilize similar discordant alignment features to detect SVs as RD based approaches, where length and orientation of paired-end-reads are used to infer SVs. For example, if read pairs are further apart than what is expected this would indicate a deletion in that locus. SR approaches allow for SV detection down to base-pair resolution (including mobile element insertions) and uses a split (soft-clipped) alignment feature of single-end or paired-end reads that span a BP of an SV. In de novo AS approaches, sequence reads are assembled as contigs and are, therefore, much dependent on the size and continuity of the contigs [32].

1.2.4 Data Types

Regardless of the sequencing strategy, all SV callers suffer from a high rate of miscalling of SVs, due to errors in base-calls, alignment or assembly. This is most apparent in repetitive regions of the genome, where short-reads are failing to span regions with SVs [31]. Studies have shown that short-read methods have particularly poor sensitivity for such regions, especially for smaller SVs (<10 Kb). To overcome the limitations of short-read sequencing and to capture the full spectrum of human genetic variation, single-molecule sequencing technologies producing long reads (>10 Kb) have been developed, i.e. single-molecule real-time (SMRT), sequencing by Pacific Biosciences (PacBio) and Nanopore by Oxford Nanopore [33, 34, 35]. Sequencing reads produced by these technologies shows many advantages over

short-read sequencing methods. The nature of long reads allows for mapping with higher accuracy, which allows the sequencing of repetitive and low-complexity regions [36].

1.2.5 Difficulties with SV Detection

However, high cost, sequencing errors and relatively low throughput of the current long-read producing methods are limiting its general use. Navigating through the ever-expanding catalogue of available bioinformatics tools for SV calling can be challenge too. Many comparative studies to evaluate available tools have been performed [37, 38, 39, 40, 47]. All show specific advantages and disadvantages when it comes to SV detection. Certain tools may be biased to a specific type of SV, genomic context, size etc.. Kosugi et. al., recently evaluated the performance of 69 existing SV detection algorithms [41]. Only a subset of algorithms accurately called SVs, depending on specific types and size ranges of the SVs. This result highlights the potential of producing high confidence call-sets by using a combination of specific available tools. Given that each approach and data type has its advantages and caveats, recent SV detection efforts have been able to produce high confidence call-sets by combining multiple sequencing technologies and computational approaches, such as the Human Genetic Structural Variation (HGSV) consortium and the Genome In A Bottle (GIAB) consortium [11, 42]. Given the current predominant data types, SV detection in unique regions with low mappability can be quite effective but tend to break down within regions of repetitive DNA (which are highly enriched for SVs). Another fundamental problem is that most SV detection methods do not indicate the haplotype origin of a given SV resides or require parental genomic information to infer this information [10, 11, 43]. The importance of SVs related to human genetic variation can not be overlooked. For instance, SVs are three-fold more likely to associate with a genome-wide association study signal (GWAS) than SNVs. and larger SVs (> 20 Kb) are up to 50-fold more likely to affect gene expression compared to SNVs. Thus, variants that are currently escaping detection are likely to represent important disease-causing genetic variations in unsolved Mendelian disorders and important contributors to “missing heritability” in

genome wide association studies (GWAS) of complex disorders [10]. By implementing a thorough understanding of available data types, computational resources, it should be possible to develop accurate bioinformatic pipelines to allow for accurate whole genome SV detection.

1.3 Bench-marking

A map of every individual's genome will soon be possible, but how will we know if it is correct? To estimate the correctness of any sequenced genome and/or computational performance, it is crucial to develop well-characterized benchmark data-sets that can be used to evaluate and compare results from different sequencing and analysis platforms. Particularly, the importance of accurate genome maps cannot be overestimated when it comes to diagnostics and precision medicine. Well-studied reference materials will allow for development of technical infrastructure that can be used as a standard for development, optimization and demonstration of new tools and technologies [42]. Besides the 1000 genomes project, there are quite a few recent and ongoing initiatives, set out to provide such well-characterized reference materials. This next section will describe one of them, GIAB.

1.3.1 Genome In A Bottle

GIAB was initiated in 2011 and is a public academic consortium hosted by the U.S. National Institute of Statistics and Technology (NIST), focused on developing reference methods, reference materials and reference data to enable the translation of whole human genome sequencing to clinical practice [42, 44, 45, 46]. The priority of GIAB is an authoritative characterization of human genomes for use in analytical validation and technology development, optimization and demonstration. Rather than studying a large set of genomes, GIAB is currently focused on a few sets of human genomes that are subjected to deep characterization, using multiple technologies and methodologies. The data described in this consortium includes BioNano Genomics, Complete Genomics paired-end and LFR, Ion Proton exome, Pacific Biosciences, SOLid, 10x Genomics GemCodeTM WGS and Illumina paired-

end, mate-pair and synthetic long read. Samples included in GIAB are visualized in table 1.1.

Coriell ID	Description	Gender	Race	Ethnicity	Relationship
NA12878	CEPH/UTAH	Female	Caucasian	Utah/Mormon	Mother
NA24385	PGP	Male	Caucasian	Ashkenazi Jewish	Son
NA24149	PGP	Male	Caucasian	Ashkenazi Jewish	Father
NA24143	PGP	Female	Caucasian	Ashkenazi Jewish	Mother
NA24632	PGP	Male	Asian	Chinese	Son
NA24694	PGP	Male	Asian	Chinese	Father
NA24695	PGP	Female	Asian	Chinese	Mother

Table 1.1 GIAB samples: Reference materials from samples have been developed by extracting DNA from a large homogenized growth of B lymphoblastoid cell lines from the Human Genetic Cell Repository at Coriell Institute for Medical Research (NIGMS).

GIAB's current main focus is to benchmark large indels and structural variant calls, including calls in difficult genomic regions (e.g. homopolymers/tandem repeats and difficult-to-map regions like pseudogenes and segmental duplications). As of today, the current benchmark provided by GIAB includes; simple SNPs, simple indels, small indels (15-50 bp) and indels > 50 bp. A description of variants in relation to their genomic context and the fraction of such variants that are called relative to the number of variants that are missing is shown in table 1.2.

Type of variant	Genome context	Fraction of variants called	Number of variants missing
Simple SNPs	Non-repetitive	~97%	>100k
Simple indels	Non-repetitive	~93%	>10k
All variants	Low mappability	<30%	>170k
All variants	Regions not in GRCh37/38	0	>>100k
Small indels	Tandem repeats and homopolymers	<50%	>200k
Indels 15-50bp	All	<25%	>30k
Indels >50bp	All	<1%	>20k

Table 1.2 GIAB benchmark: As of today, we have a considerably good understanding, looking at simple SNPs/indels in non-repetitive regions of the genome but when including all variant types (especially balanced events) throughout the whole genome, a substantial, large fraction of variants are missed using standard sequencing approaches.

Based on these sobering numbers new approaches to uncover more complex variants residing in difficult-to-map regions are clearly needed. As of today, only <1% of larger indels throughout the whole genome can accurately be interrogated, leaving approximately 20 thousand variants undiscovered. Given the significance of SVs in relation to the genotypic and phenotypic variation, this current knowledge gap and its implications are highlighted in this thesis, especially when it comes to finding disease-causing variants or other significant clinical variants.

1.3.2 Challenges in Benchmarking Variant Calling

Generally, it is difficult to achieve robust benchmarking of tests designed to detect many analytes (variants) and efforts tend to be biased towards regions that are more easily analyzed [11, 45, 46]. Another fundamental problem for SV detection is to

resolve events that are being flanked by virtually identical duplicated sequences and copy-neutral, balanced events (e.g. inversions and translocations) which preclude detection by read-depth analysis [10].

1.4 Single Cell Template Strand Sequencing

1.4.1 Technology

Strand-seq is a single-cell sequencing technique that identifies the original parental DNA template strands in daughter cells following cell division [48]. This method leverages the directionality of single-stranded DNA molecules, which can be distinguished as either Crick (C; forward or plus strand) or Watson (W; reverse or minus strand) based on their 5'-3' orientation. Following division after one DNA replication round in the presence of the thymidine analog 5-Bromo-2'-deoxyuridine (BrdU), one strand will have randomly incorporated BrdU. Strands with BrdU can be selectively removed by treating samples with UV and the DNA dye Hoechst 33258 (figure 1.1), ensuring that only BrdU negative, parental DNA template strands are sequenced for each chromosome in each single cell. This approach maintains directionality within sequenced libraries and provides crucial information on the genetic makeup that cannot be obtained in another way.

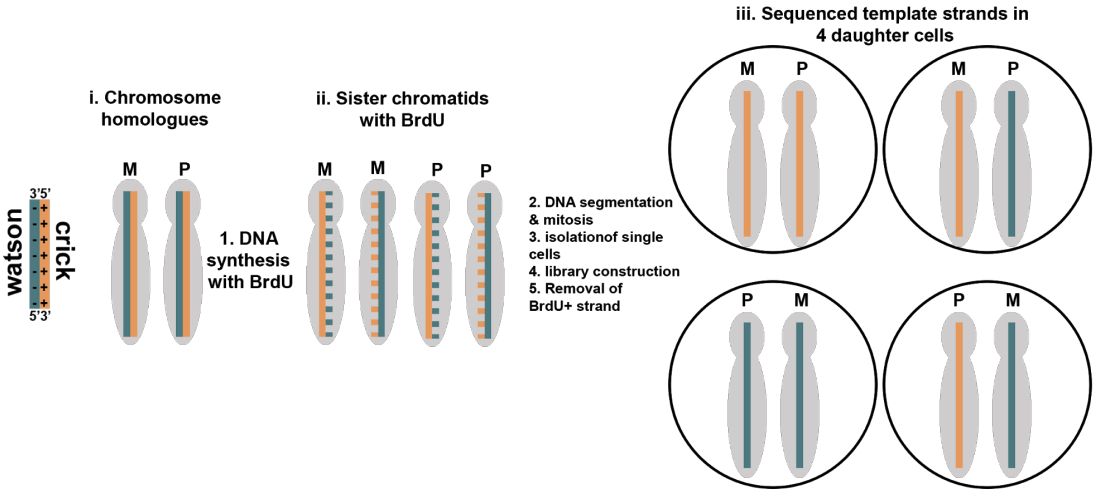


Figure 1.1 Single cell template strand sequencing (Strand-seq): Newly formed DNA strands containing BrdU (dashed lines) in parental cells are selectively removed in Strand-seq libraries, allowing only the original parental DNA template-strand to be sequenced (solid lines, right panels). Right panel show all possible chromatid segregation patterns.

1.4.2 Available Strand-seq Bioinformatics Tools

After library construction and high-throughput sequencing (HTS) sequence reads are aligned either to the Crick or Watson strand of the reference genome using the software package BAIT and DNA template strand inheritance patterns are determined for each chromosome within the cell [49]. Strand-seq has multiple applications. For example, it allows for comprehensive studies of DNA repair and replication, fine mapping of sister chromatid exchange events (SCE), refinement and assembly of reference genomes, de novo haplotyping (without relying on parental sequence data), mapping of polymorphic inversions, mapping of chromosomal copy number variations, analysis of chromosomal abnormalities (including translocations) and analysis of loss of heterozygosity (LOH) at the level of single cells [10, 11, 50, 51, 52, 53]. Strand-seq is also a powerful application for identifying clinically relevant variants that typically escape detection using read-depth based approaches. For example, recent studies show that analysis of Strand-seq data reveals a de novo somatic inversion on chromosome 17, involving breakpoints subject to a translocation between chromosome 15 and 17, resulting in a clinically significant fusion between *TP53/NTRK3* [58]. In addition, recent efforts successfully assemble karyotypes in PDX-derived T-cell acute lymphoblastic leukemia (T-ALL) as well as identifying novel SVs, demonstrating the diagnostic value of Strand-seq as a technology. Strand-seq has the potential to complement other sequencing technologies by providing crucial information about genetic makeup. In the next section, currently available bioinformatics tools for analyzing Strand-seq data are outlined.

1.4.2.1 BAIT

BAIT was introduced in 2013 as the first software package to analyze Strand-seq data [49]. BAIT assigns templates and identifies and localizes SCEs as well as counts aneuploid chromosomes or fragments. This is achieved by organizing resultant directional reads into chromosome dependent ideograms (igure 1.2). If a chromosome has reads mapping solely to the Watson strand, the cell has inherited a Watson template strand from each of the parental homologues and if the reads are

mapping to both the Watson and the Crick template, the cell has inherited one Watson and one Crick homologue (figure 1.1, right panels). This ability to distinguish which template strand was inherited by dividing cells allows for high-resolution of SCEs, genomic rearrangements and more recently, refinement of reference assemblies. SCEs are visualized on chromosome ideograms as regions where reads switch from a homozygous template state (WW or CC) to a heterozygous template state (WC) and BAIT identifies these regions by calculating the ratio [(W-C)/(W+C)] of Watson and Crick reads within a user-defined bin and normalizing to the nearest integer. When all reads map to the Watson strand a ratio of 1 is returned, for reads mapping solely to the Crick strand the ratio is -1 and for reads that map to both strands the ratio is 0. A change in this ratio corresponds to the SCE localized to neighboring bins. For example, a template-switch from CC in one bin (default 200kb) to a WC template strand in a neighboring bin indicates that an SCE occurred somewhere within the 400 kb interval encompassing those two bins. Previous studies have shown that Strand-seq identifies over 20 Mb of the MGSCv37/mm9 *Mus Musculus* reference assembly as miss-oriented and that a large number of these orientations can be corrected using Strand-seq.

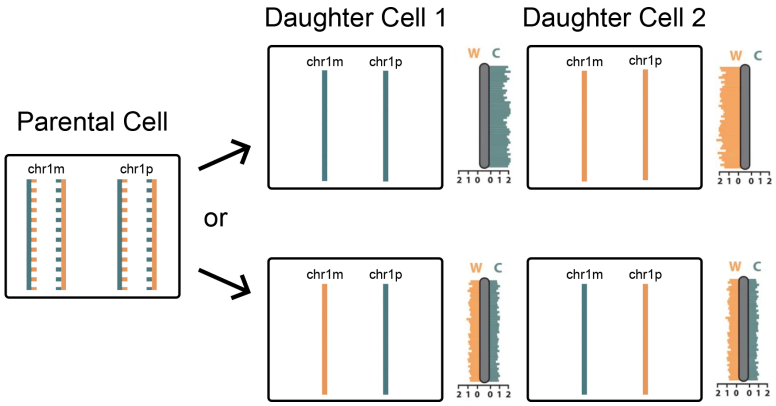


Figure 1.2 BAIT generated karyograms: Daughter cells inherit two template strands, since there is a maternal (m) and paternal (p) copy of each chromosome. Chromatids segregate either with both Crick strands inherited into one daughter (daughter cell 1 top), both Watson strands in the other (daughter cell 2 top) or with one Watson and one Crick strand in each daughter cell (bottom panels). Sequence read density is plotted onto ideograms (gray bars).

1.4.2.2 Aneufinder

Aneufinder was developed to streamline quality control and to facilitate copy number calling in single-cell sequencing data [54]. This is achieved by counting reads in non-overlapping bins of variable size, based on mappability and correction for GC content. In the next step, a Hidden Markov model (HMM) is integrated into the binned sequence data, assuming potential copy number states, from nullisomy up to decasomy. A negative binomial distribution is used to model all states (except for nullisomy which is modelled by a delta distribution). To obtain the best fit for distribution parameters, transition probabilities and posterior probabilities the Bauman-Welch algorithm is implemented. Lastly, each bin is assigned the copy number state with the highest posterior probability.

1.4.2.3 InvertR

As previously stated, the neutral copy number state of inversions, in general, makes structural variations of this type generally more difficult to detect/characterize using available read-depth based bioinformatics algorithms. This is reflected by the fact that few human inversions have been studied comprehensively to date and thus, the phenotypic consequences and clinical relevance of most remain uncertain. Inversions are identified in Strand-seq sequence data as a localized reorientation (with respect to the reference assembly) in the Watson-Crick state along the DNA strand of a chromosome and are distinguished from sporadic SCE by requiring reoccurring changes at the same loci in multiple cells [55]. Each localized inversion can be further genotyped if one or both chromosome homologs show a localized strand reorientation. Heterozygous inversions are visualized as a single homolog being inverted with respect to the reference assembly (WW or CC state switches to a mixed WC state) or a homozygous rearrangement where both homologues are inverted and template strands completely switch from WW to CC or vice versa. To allow for robust characterization of inversions in Strand-seq libraries, InvertR was developed. InvertR is an R-based pipeline that localizes and genotypes putative inversion calls based on read alignment. More in detail, InvertR implements a read-based sliding window to calculate the ratio of W/C reads across each chromosome.

It then plots these ratios as a histogram, visualizing changes in template strand orientation. Putative inversions are called in regions where the W/C ratio drops below a user-defined threshold before returning to the normal ratio. This approach allows for accurate genotyping of putative inversions, by interrogating the overall change in W/C ratio. A homozygous inversion is classified as a complete change in template state with respect to the genomic context and gives a W/C ratio close to 1.0, and a heterozygous inversion is classified as a change in template states from either W or C to a mixture of both reads. This gives a W/C ratio close to 0.5.

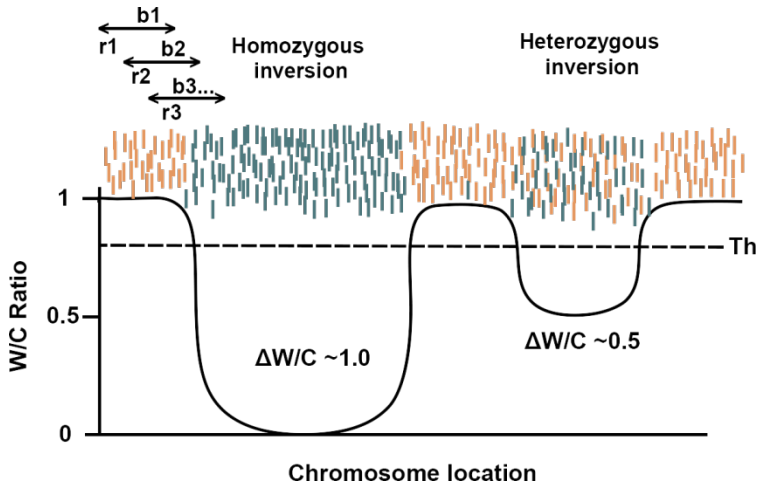


Figure 1.3 InvertR algorithm: By stepping along reads in a user-defined bin this algorithm calculates the proportion of Watson and Crick reads. The bin moves sequentially (b1 to bn) along every read (r1 to rn), and the W/C ratio calculation is repeated and plotted as a histogram. Putative inversions are localized to the genomic region where the W/C ratio passes a threshold and genotyped based on the magnitude of change at the inverted segment.

1.4.2.4 StrandphaseR

The phasing of SVs across the length of a chromosome without parental sequence data is currently very challenging [10, 56]. To overcome these challenges, a few recent efforts have been made such as whole-chromosome sorting, chromatin capture techniques and linked-read sequencing [11]. All these techniques are labour/ time consuming and fail to phase genetic variants across whole chromosomes. StrandphaseR was developed to allow for accurate direct chromosome-length

haplotyping using Strand-seq data [56]. Given the template states (WW, CC or WC) parental haplotypes can be readily distinguished when a cell inherits sequence data that maps to both the W and the C template strand for a particular chromosome. By combining multiple Stand-seq libraries with regions inherited as WC for that chromosome, phasing of variant alleles along the entire chromosome is possible. This approach allows the construction of haplotypes spanning centromeres, sequence gaps and regions of homozygosity. More in detail, the StrandphaseR algorithm first identifies WC regions in multiple single-cell Strand-seq libraries. Identified regions are selected as input for haplotype phasing and together with SNVs identified on each template strand, single-cell haplotypes can be derived. StrandphaseR first establishes anchor haplotypes by pairing single-cell haplotypes showing the highest number of overlapping heterozygous SNVs. Anchored haplotypes are implemented to initialize the consensus haplotypes (H1 and H2) and are improved after each iteration. In the next iteration, the second most-dense haplotype is considered and compared to established consensus haplotypes, adding any new SNVs to the consensus haplotype showing best concordance. For each iteration, consensus haplotypes are extended until no additional single-cell haplotype can be assigned to one of the consensus haplotypes.

1.4.2.5 BreakpointR

This algorithm was designed to allow for robust interrogation of template strand changes in Strand-seq data, using bi-directional read-based binning (scales each bin size dynamically) [57]. BreakpointR then records read directionality in each dynamically scaled bin to identify points where directionality of sequenced template strand changes. In similarity to BAIT and InvertR, BrekapointR allows for fine-mapping of SCEs, germ-line inversions.

1.5 Thesis Rationale and Objective

SVs contribute greater diversity at the nucleotide level between two human genomes than any other form of genetic variation and are three-fold more likely to correlate with a genome-wide association study signal than SNVs (Chiang et al. 2017). Using short-read, high-throughput sequencing technologies to uncover such variation has proven to be troublesome and the methods to detect SVs are independent on indirect inferences (e.g read-depth and discordant read-pair mapping). However, while larger (>5kb) CNVs could be characterized by using read-depth-based algorithms, this approach often fails for smaller and balanced events, such as inversions. Another fundamental problem for SV detection from short-read sequencing is inherent to the predominant data type and the typical SV detection algorithm can thus be effective in unique sequences but fails within complex genomic regions (e.g repetitive and low map-ability regions), which are highly enriched for SVs. In addition, most SV discovery methods do not indicate the haplotype-origin for a given SV or requires parental sequencing information to achieve this. For a more complete description and interpretation of human genomic information in relation to phenotypes and e.g. cancer predisposition and response to therapies, it will, therefore, be necessary to arrange sequence data into parental haplotypes and ascertain polymorphic inversions with respect to such haplotypes. All this can be achieved using Strand-seq. Strand-seq complements other sequencing approaches by providing crucial information about the genetic make-up of individuals that cannot be obtained in any other way. To make Strand-seq available for human studies worldwide is an immense challenge. Library construction, as well as data analysis, needs to be further developed, integrated and made user-friendly to allow accurate and rapid interpretation of results. Here we present a custom bioinformatics pipeline for analyzing Strand-seq data that streamlines the workflow of raw sequence read alignment and putative variant calling to allow for accurate whole-genome haplotype (WGH) assembly and high confidence variant call-sets. In addition, relevant metric data are compiled and visualized, ensuring and reinforcing the potential of Strand-seq as a robust sequencing method for uncovering clinically significant SVs and the assembly of WGH without additional parental genomic data.

2. Results

The result section consists of three major parts. First, design-decisions associated with the pipeline development are described in detail together with an overview of the proposed workflow and its processes. Next, suggested workflows are tested using previously sequenced Strand-seq data and the results are presented, ensuring the performance of developed pipelines. Lastly, an inversion validation assay was designed to experimentally assess the accuracy of the developed inversion calling pipeline.

2.1 Pipeline Development

2.1.1 Sequencing, Alignment

The goal of the first pipeline is to align raw read data to a reference sequence in a reproducible manner. This will allow for transparent comparisons of different sequencing efforts as well as data drawn from various library preparation protocols. In addition to a streamlined alignment approach, the workflow also compiles performance metrics into a standardized quality report that will aid performance evaluation. The workflow utilizes a variety of different well-established bioinformatics tools to build a comprehensive and widely customizable quality report. In the first step, the pipeline takes raw reads in fastq format. Two scripts are available to handle each underlying data type (paired-end or single-end) and align raw reads to a reference sequence using Bowtie2. The output of the read alignment data is in the Sequence Alignment Map (SAM) format. The workflow then converts read data in SAM format to Binary Alignment Map (BAM) format using Samtools, to minimize the computational footprint. An additional script is included to trim each single-cell library to only include chromosome 1-22, X and Y thus reducing the computational footprint even further. Since the alignments that are produced are in random order with respect to their position in the reference genome, all alignments are sorted using the sort function from Samtools to ensure that all alignments are ordered positionally based upon their chromosomal alignment coordinates. Next, duplicate reads in processed bam files are located using the MarkDuplicateReads function from Picard Tools. Duplicate reads are defined as reads originating from a single fragment of

DNA. The tools accurately interrogate duplicate reads by comparing sequences in the 5’ positions of both reads and read-pairs in the bam file. By implementing an algorithm that ranks reads based on the sum their base-quality score, primary and duplicate reads can be distinguished. Besides from accurately marking duplicate reads throughout the alignment data, the script also produces a metrics file that will be used for compiling an informative quality report. Lastly, sorted and duplicate marked reads are indexed using Samtools.

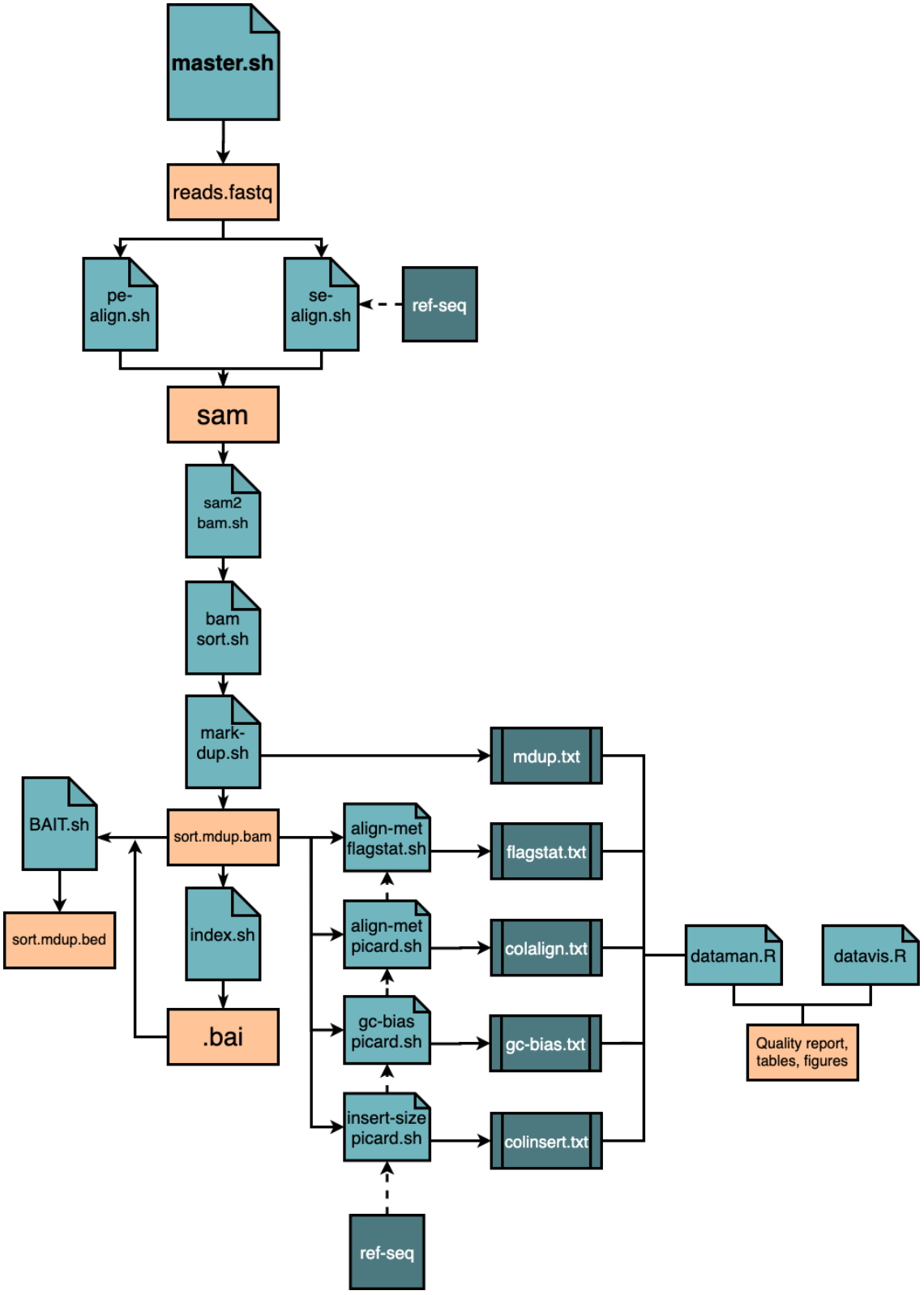
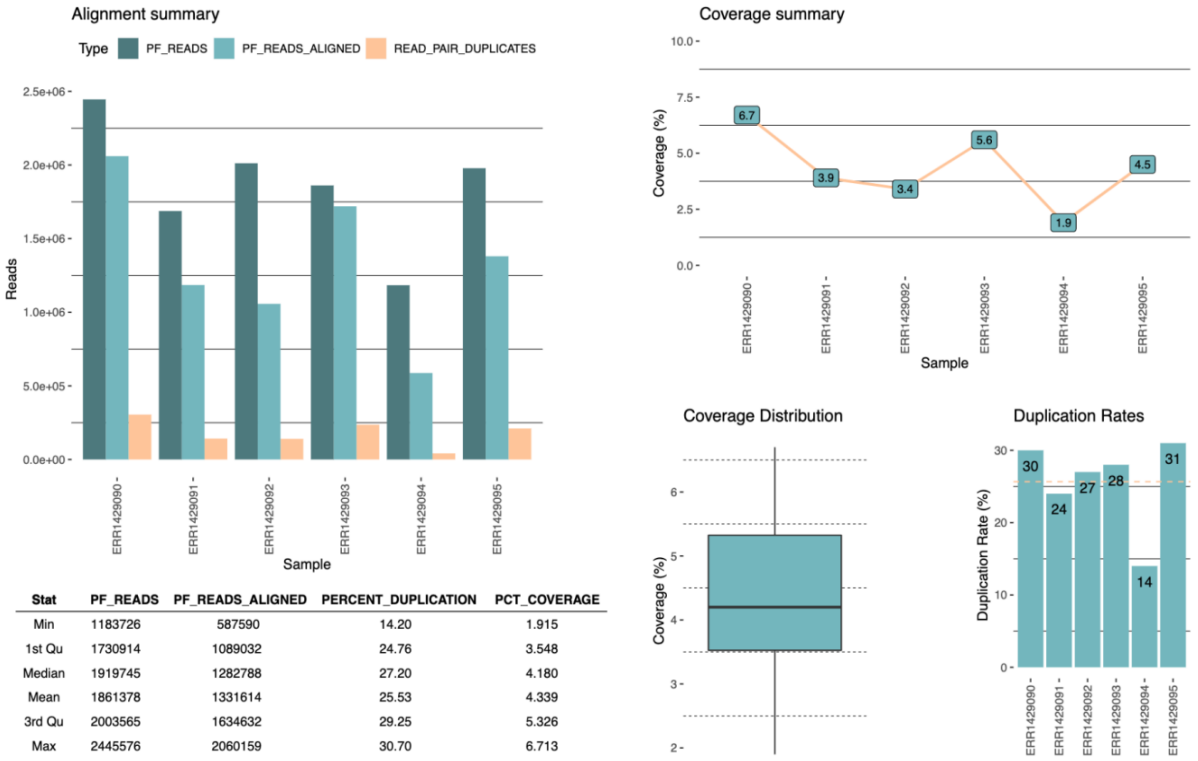


Figure 2.1 Pipeline: Alignment and quality control: illustrates the bioinformatic workflow created to align multiple Strand-seq libraries to the reference sequence with multiple additional steps, as well as compiling necessary alignment metrics to construct a standardized quality report. Orange boxes are annotating data input/output, light-blue are highlighting data scripts (bash and R), dark-blue boxes show dependencies and metrics. The complete workflow is executed by running the master.sh script.

2.1.2 Alignment Metrics and Quality Report

To enable the construction of an informative quality report, the pipeline implements a variety of functions from different packages. Besides compiling metrics regarding duplicate reads for each single cell library, the pipeline also gathers relevant alignment metrics to allow for an easy and readable assessment of sequencing performance as well as quality evaluation of each individual single cell library. CollectAlignmentMetrics from Picard tools and flagstat from Samtools is implemented to provide information on total reads (n), n reads that can be aligned to the reference sequence, n reads that are passing a user-defined quality threshold as well as the percentage of adapter dimers in every single cell. CollectGCBiasMetric function is called from Picard tools to interrogate potential GC bias. The genome is then binned into windows of 1 Mb and the produced metric file is used to calculate a standardized coverage for each bin. Lastly, insert-size distribution is addressed calling InsertSizeMetric from Picard tools. Custom R-scripts are then implemented to concatenate relevant information from each output and construct a quality report in pdf format together with filtered tables, containing information for every single cell included in the analysis allowing for a transparent and reproducible interrogation of the sequencing performance.



Stat	PF_READS	PF_READS_ALIGNED	PERCENT_DUPLICATION	PCT_COVERAGE
Min	1183726	587590	14.20	1.915
1st Qu	1730914	1089032	24.76	3.548
Median	1919745	1282788	27.20	4.180
Mean	1861378	1331614	25.53	4.339
3rd Qu	2003565	1634632	29.25	5.326
Max	2445576	2060159	30.70	6.713

Figure 2.2 Quality report: Quality report example from executing described pipeline on experimental data (6 random samples). The alignment summary visualizes the number of reads that successfully can be aligned to the reference sequence as well as the fraction of those reads that are passing the Illumina quality filter and number of duplicate reads in relation to the overall number of reads. Coverage summary shows the distribution of standardized coverage for every single cell. Coverage distribution highlights the different quantiles for coverage of all included single cells. Duplication rates visualize the number of duplicate reads as a fraction of the total number of reads. Alignment table, provides additional statistical information on aligned reads, high-quality reads, duplications and coverage.

2.1.3 Locating Putative Inversions and Refinement of Variant Calls

The aim for this second pipeline is to accurately call putative inversions in multiple single-cell libraries; to refine and genotype putative inversion calls and to output final high-confidence call-set in a format that allows for a user-friendly inspection of calls and underlying data to the same calls. This workflow makes use of the previously described algorithm, invertR. Inversions are identified in Strand-seq sequence data as a localized reorientation (with respect to the reference assembly) and are distinguished from SCEs with the requirement of recurring in multiple libraries. InvertR creates a variety of outputs such as; whole chromosome overlay figures, allele frequencies, library-specific bed files, WC libraries and chromosome-specific region of interest files (ROI). The latter holds information that annotates each putative inversion call with; library, chromosome, calling threshold, ROI start, ROI end, ROI size, delta WC and ROI reads. After invertR is executed on selected single-cell libraries together with a region file (tab-separated) stating what genomic regions that are to be analyzed, the workflow concatenates chromosome-specific ROI files into one ROI spanning all chromosomes. Putative inversions calls found in the newly constructed ROI file are further manipulated with custom R-scripts, filtering out any events larger than 4 Mb (default) and appropriate variables are selected to produce an initial bed file. Next, the workflow sorts all retained putative variant calls on the outer left-most coordinate implementing sort function from bedtools. Overlapping sorted putative calls are then merged, implementing a confidence interval (CI) of 10KB. The number of cells in agreement for any particular call is annotated as a new variable in the bed file. Calls with < 2 occurrences are filtered out using custom R-script to ensure only recurring events are retained and SCEs are filtered out. Final refined call-sets are saved in bed format, allowing for an easy visual inspection using preferred genome-browser (i.e IGV or UCSC). Lastly, breakpointR is executed on selected libraries to allow for further confirmation of refined inversion call-set as well as preparing data for phasing structural variants using the next pipeline.

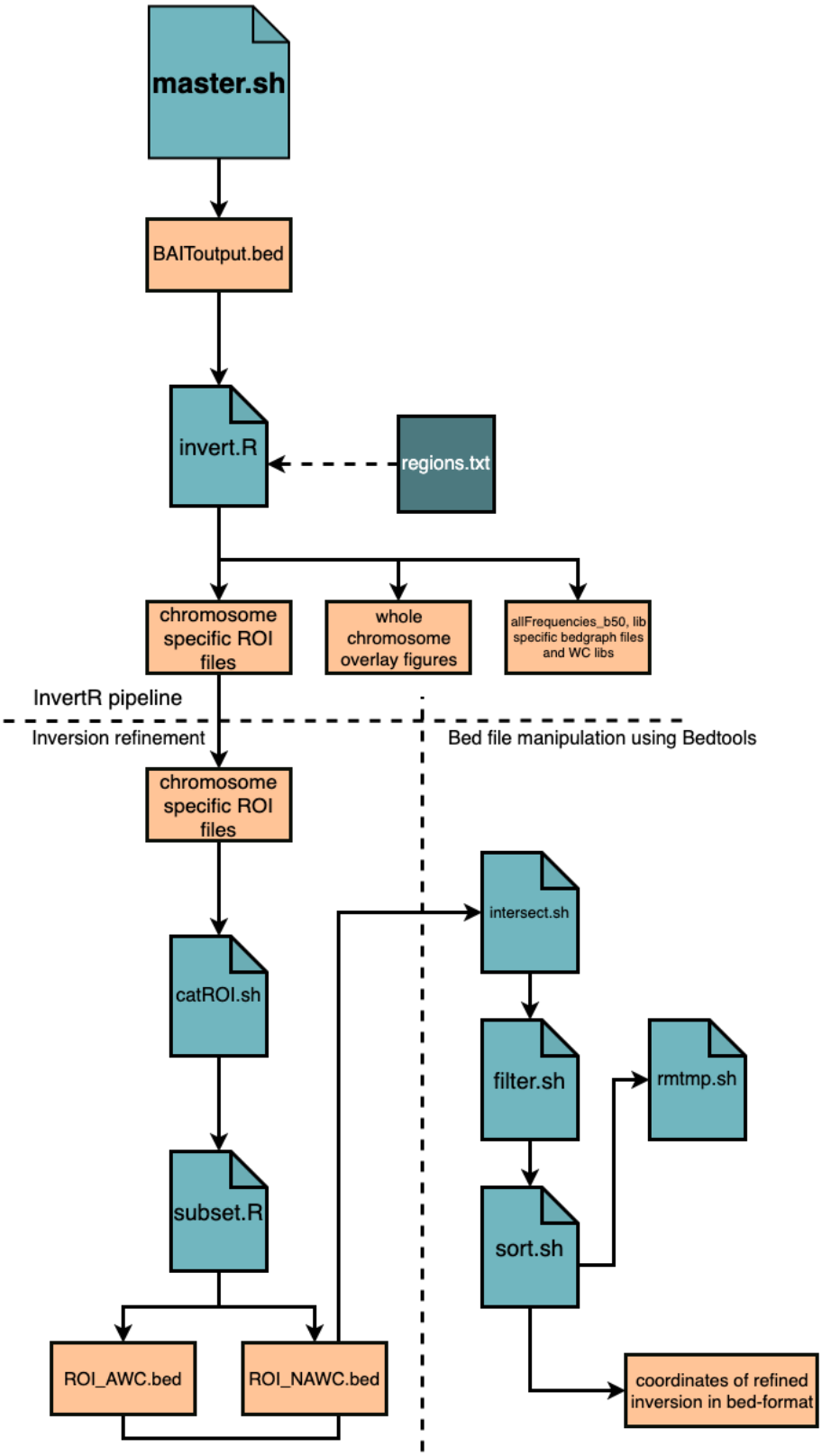


Figure 2.3 Pipeline: inversion calling: Flowchart describing the bioinformatic workflow for locating putative inversions and refinement of putative inversion calls in multiple Strand-seq libraries. Orange boxes are annotating the main data input/output. Custom designed scripts are highlighted in light-blue and dependencies are annotated in dark-blue. Workflow is executed by running the master.sh script.

2.1.4 Haplotype Assignment

The goal with this pipeline is to accurately and in a reproducible manner assign haplotypes to an internal or external set of variants (both SNPs and SNVs) and assemble haplotypes or phase DNA without relying on parental sequencing data. Given the sparseness reflected on the current Strand-seq library (average coverage 1-5%) the phasing of SV call sets generated using Strand-seq data alone is not recommended. Rather an external set of variants coming from any other long-read technology such as PacBio or Illumina is preferred to accurately assign a haplotype to known variants. The workflow first interrogates Rdata files generated using breakpointR to locate WC regions throughout all single cells using a custom-designed R-script. Identified WC regions for each library will be subject to construct anchor haplotypes (see algorithm explanation in section 1.4.2.4). A custom R-script is then implemented to filter the external variant call-set to match the required format for the workflow (removing the header and maintaining chromosome and position for each variant). An additional script is included to filter variants to reflect either indels or SNPs. The pipeline then feeds identified WC regions together with corresponding bam files, reference genome and filtered external set of variants (to be phased) into the strandphaseR algorithm. The workflow outputs homologue specific bam files as well as chromosome-specific variants in variant call format (VCF). The workflow then concatenates and sorts chromosome-specific VCF files into one VCF file spanning all chromosomes. To ensure and evaluate phasing performance, an additional script is included to compare generated VCF file with a truth/benchmark VCF file and performance metrics is calculated. Lastly, an additional script for performing integrative phasing (using WhatsHapp) with additional sequencing technologies is included. If parental sequencing data is available, the pipeline can be used to assess meiotic breakpoints by performing a pairwise comparison of homologue-specific bam-files.

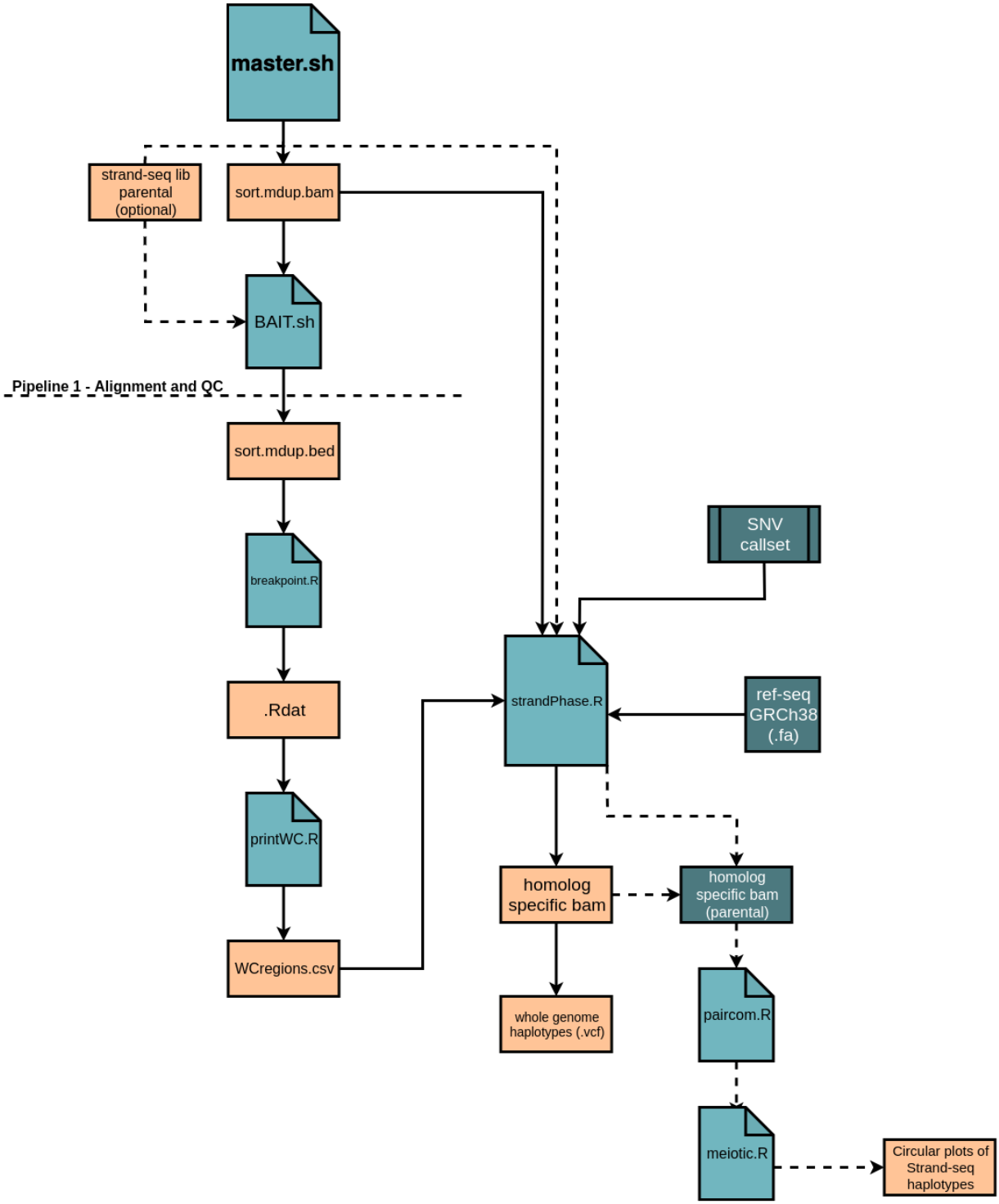


Figure 2.4 Pipeline: phasing: Workflow describing the bioinformatic processes for performing phasing of structural variants. Orange boxes are annotating the main data input/output. Custom designed scripts are highlighted in light-blue and dependencies are annotated in dark-blue. Dotted lines are describing the workflow for performing a pairwise comparison of homologue-specific bam files for parental sequencing data for locating meiotic breakpoints. Complete workflow is executed by running master.sh script.

2.2 Pipeline Performance

A bioinformatics workflow was created to allow for a streamlined and transparent approach for analyzing multiple single cell (Strand-seq) libraries. The workflow first takes raw read data and aligns it to a reference sequence with additional steps such as; sorting reads, mark for duplicate reads, index reads, and file conversions. Sequencing and alignment metrics such as; the number of reads, reads aligned, unique reads, high-quality reads, coverage, standardized coverage, GC bias, insert size and duplication rates are calculated and formatted into a quality report, to allow for a standardized way of comparing sequencing and alignment performance. The developed pipeline also allows for accurate inversion calling, and haplotype assembly (without parental sequence data). The workflow was executed on a selection of previously sequenced Strand-seq libraries for NA12878 and in the results are described in detail below.

2.2.1 Sequencing, Alignment

On average, selected libraries showed a total read-count of around 3.1 million reads and 600,000 unmapped reads (20%). Duplication rates ranged from 7 to 31% with a mean value of 22% and mean coverage of 8%.

	Total Reads	Unmapped Reads	PF_reads	PF_reads aligned (%)	Duplication (%)	Coverage (%)
min	735456	122804	735456	29	7	1
median	3184839	606896	3184839	79	25	8
mean	3168164	661974	3168164	76	22	8
max	6982882	1551504	6982882	92	31	19

Table 2.1 Alignment summary: Summary of alignment performance metrics for 150 single cell Strand-seq libraries.

2.2.2 Alignment Comparison

To evaluate alignment performance on selected libraries and to provide support for alignment-choice, a comparison of three well-established aligners (Bowtie2, Hisat2, and Minimap2) were implemented. Bowtie2 is the one aligner that captures the highest standardized coverage of all evaluated aligners (figure 2.6). When comparing other alignment metrics, such as; uniquely mapped reads, duplicate reads, estimated library size and aligned reads that are passing Illumina's quality filter, both Bowtie2 and Minimap2 produces similar results. However, Bowtie2 was chosen as the best aligner mainly because it captures the highest standardized coverage compared to all other evaluated aligners.

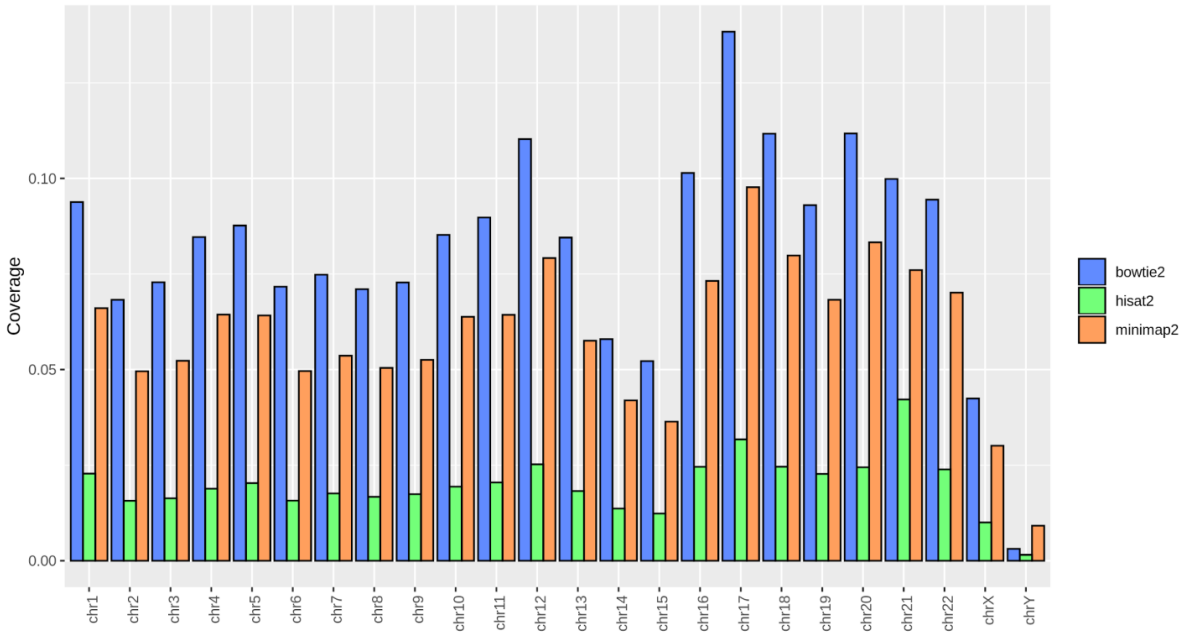


Figure 2.5 Alignment comparison: The standardized coverage for all evaluated aligners (Bowtie2, Hisat2, Minimap2) is shown. Standardized coverage was retrieved by binning each chromosome in windows of 1 Mb and standardized coverage was calculated for each bin. For every chromosome, average standardized coverage for each bin was visualized as a histogram.

	Bowtie2	Hisat2	Minimap2
Total reads	4099292	4099292	4099292
Unmapped reads	1073521	1200668	1033223
PF reads	4099292	4099292	4099292
PF reads aligned	3025770	2898624	3066068
Estimated library size	901167	847575	925329
Unique reads	1460581	1386890	1494157
Coverage (duplicates removed)	0.047	0.045	0.048
Coverage (duplicates included)	0.098	0.0944	0.099
PCT adapter	1.6	1.6	1.5
Percent duplications	52	52	51

Table 2.2 Metric comparison of alignment performance: Metrics associated with alignment and how metrics from each individual evaluated aligner compare to each other.

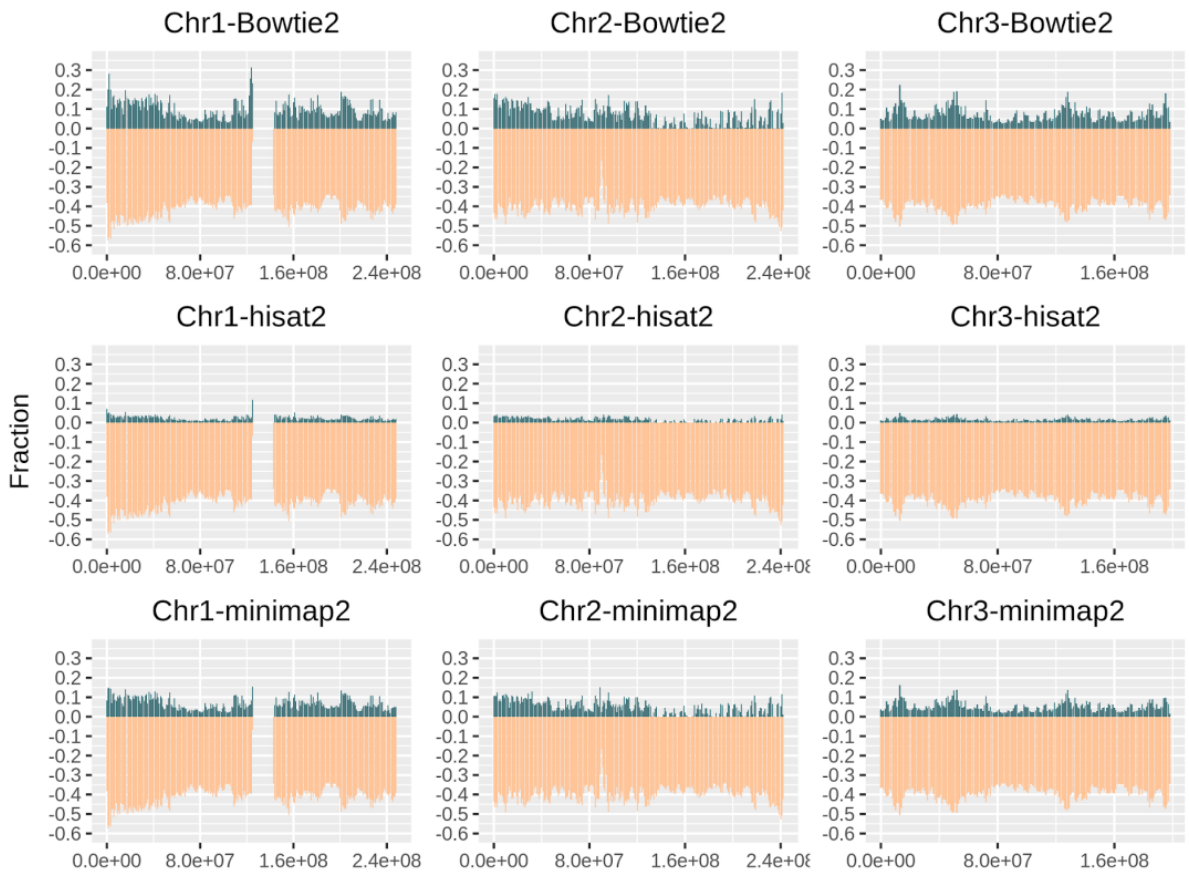


Figure 2.6 Standardized coverage comparison: In-depth view of standardized coverage for each evaluated aligner on a subset for chromosome 1-3, using one high-quality library (ERR1429117).

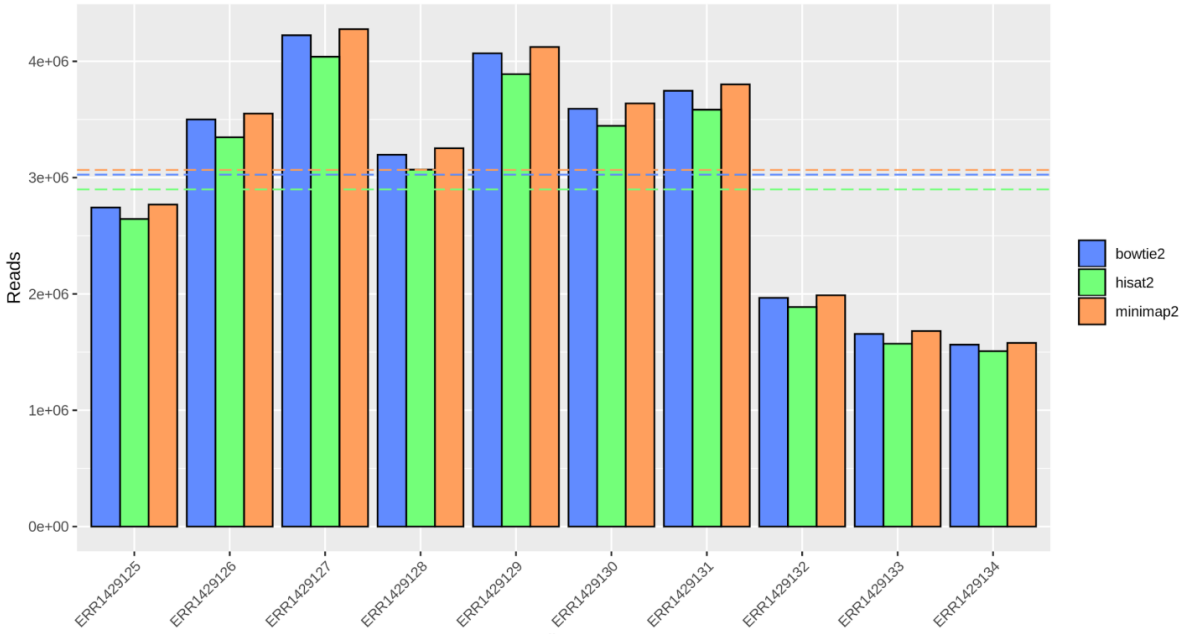


Figure 2.7 Alignment comparison: PF reads aligned: The percentage of PF reads that successfully can be aligned to the reference sequence. Mean values for each aligner are annotated as dotted lines. Data is drawn from a subset of Strand-seq samples that are passing the quality evaluation criteria previously stated.

2.2.3 Locating Putative Inversions and Refinement of Variant Calls

Similarly, the developed inversion calling pipeline InvertR was executed on selected libraries to produce high-confident variant call-set of Strand-seq resolved inversions. InvertR was implemented with the following parameters; Bin-size was set to 50, W/C cutoff of 0.75, read quality criteria of 10, minimum depth and minimum reads were set to 20. Created pipeline initially identifies 3531 ROIs resulting in 457 inversions after merging overlapping events. Putative inversions constitute 1500 Mb inverted genome with a mean inversions size of 5.9 Mb with a size-range of 117 997 - 108 582 279 bp. After refinement of identified putative inversion, data shows 167 high-confidence Strand-seq resolved inversions, where 12 inversion calls are genotyped as homozygous and 155 as heterozygous. Refined inversions constitute 22 Mb of inverted genome with a median size of 1.3 Mb with a size-range of 33 385 - 8 202 914 bp.

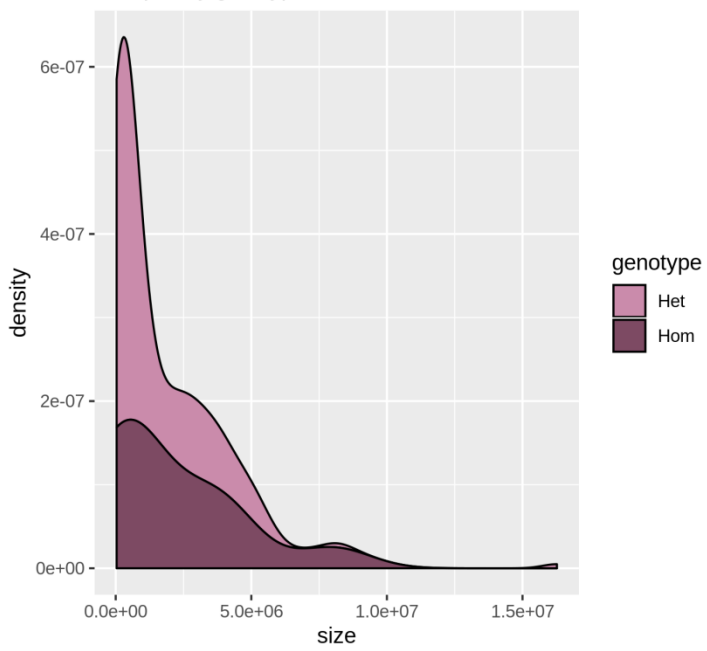


Figure 2.8 Inversion genotype distribution for NA12878: Distribution of refined Strand-seq resolved inversions grouped by genotype. Developed pipeline identifies 167 high-confidence resolved inversions, genotyped as 12 heterozygous and 155 homozygous inversions. The majority of Strand-seq resolved inversions are in the size-range of 1 Mb (for both genotypes).

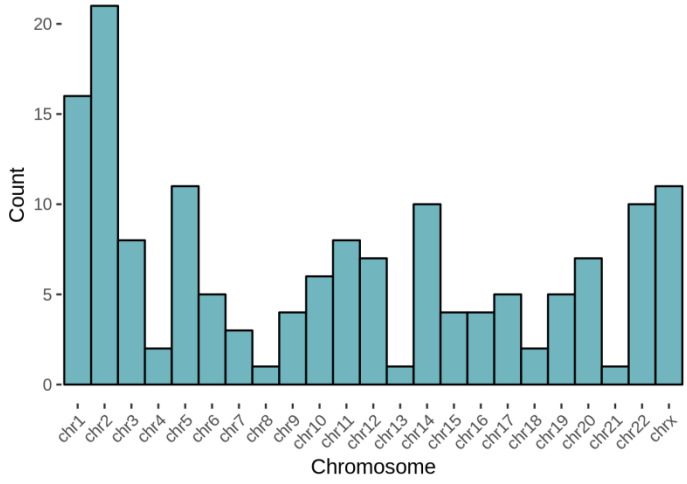


Figure 2.9 Inversion chromosome distribution for NA12878: Inversion-count distribution sorted dependent on chromosomes. Visualizing chromosome distribution for all (167) Strand-seq resolved inversions with average inversions per chromosome count of 7.2.

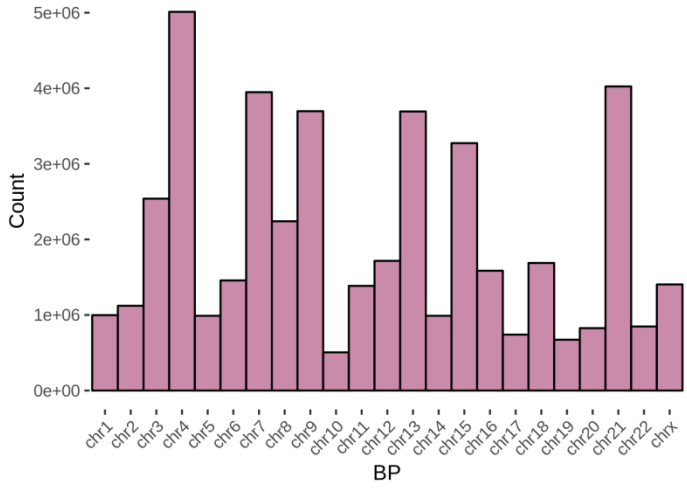


Figure 2.10 Inversion size for NA12878: Median inversion-size grouped by chromosome. Visualizing median inversion size for each chromosome for all Strand-seq resolved inversions with a median size of 1.3 Mb.

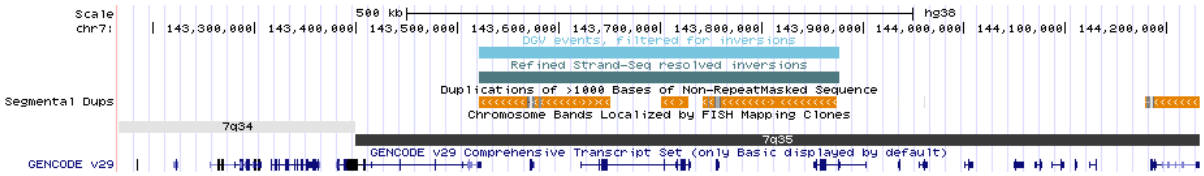


Figure 2.11 NA12878 inversion IGV genome browser example: 355 Kb homozygous inversion (dark blue track) on chromosome 7q35 shown with corresponding inversion call (light-blue track) from Database of Genomic Variants (DGV). Breakpoints for Strand-seq resolved inversion are in 99% concordance with DGV call. Both breakpoints are flanked with regions of segmental duplications (orange track).

2.2.4 Haplotype Assignment

To determine the performance of the developed phasing pipeline, the workflow was executed on previously described Strand-seq libraries. The data consists of 100 PE libraries and an external set of SNPs (www.illumina.com/platinumgenomes.html) containing a total of 2 149 538 SNPs was phased. Out of all variants, 1 341 438 (62%) was successfully phased. Coordinates of phased variants showed to be in 99.2% concordance with documented SNP positions. In order to investigate how the various number of single cells relates to fraction of successfully phased variants, previously sequenced Strand-seq data for three trios (HG00512, HG00513, HG00514, HG00731, HG00732, HG00733, NA19238, NA19239, NA19240) was randomly subsetted into sample groups of 25, 50, 100 and 150 samples. The developed pipeline was executed on all sample groups for all included samples and SNV coverage was interrogated.

	Phased variants	Unphased variants	SNV coverage	HapMap concordance
NA12878	1 341 438	808 100	62	99.2

Table 2.3 Phasing metrics: Associated phasing metrics. Phased variants describe the total number of variants that successfully could be phased. SNV coverage is the ratio of phased variants relative to unphased (x 100). HapMap concordance describes coordinate concordance with previously assembled haplotype for NA12878.

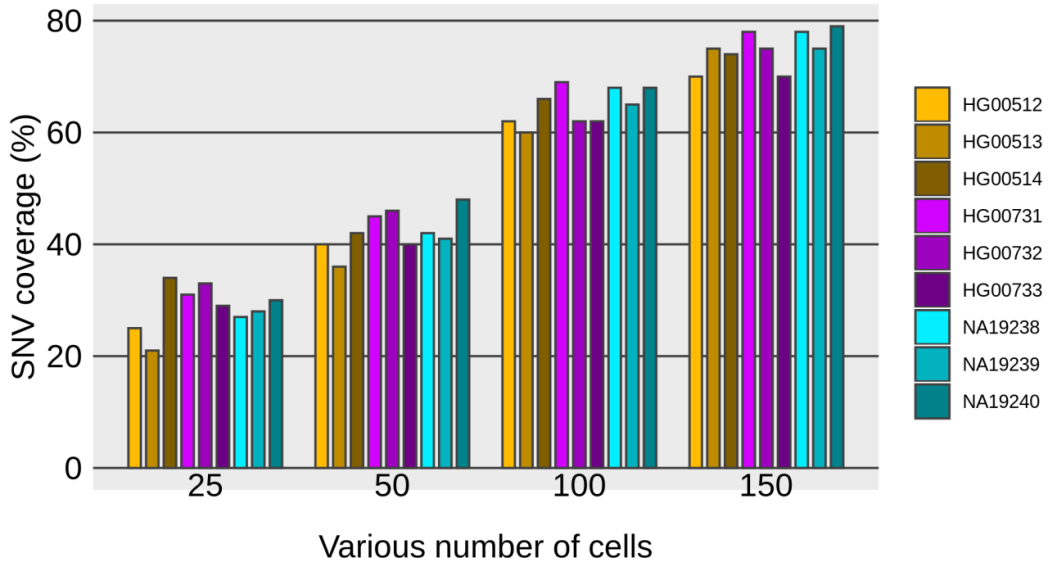


Figure 2.12 Fraction of phased variants vs various number of single cells: Phasing of a random subset of samples (25, 50, 100 and 150 samples) were performed to investigate how the number of cells correlates to the fraction of variants that can successfully be phased using developed pipeline.

2.2.5 Inversion Validation Experiment

To experimentally validate Strand-seq resolved inversions and how identified breakpoints relate to known variants for the same individual, custom bed tracks for each call-set was uploaded to IGV genome browser and breakpoints where compared. Calls from different technologies with breakpoints discordance of 500 bp were considered to belong to the same event. Surprisingly, only 2 calls are in concordance between all data-sets, whereas Strand-seq data-set showed the highest number of uniquely resolved inversions. The largest overlap (14) between two different technologies belonged to Strand-seq and Kidd et al.

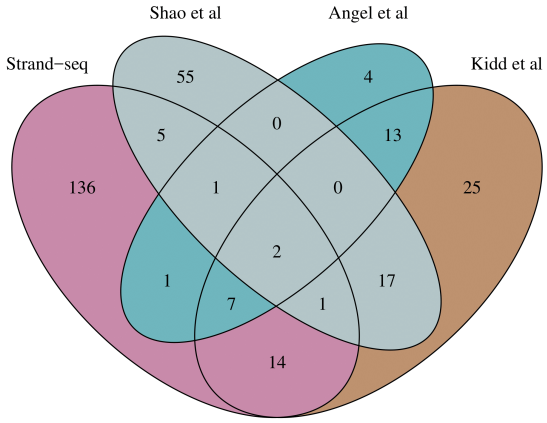


Figure 2.13 Venn diagram of overlapping inversion calls for NA12878: Unscaled Venn diagram visualizing inversion calls in concordance between validated datasets.

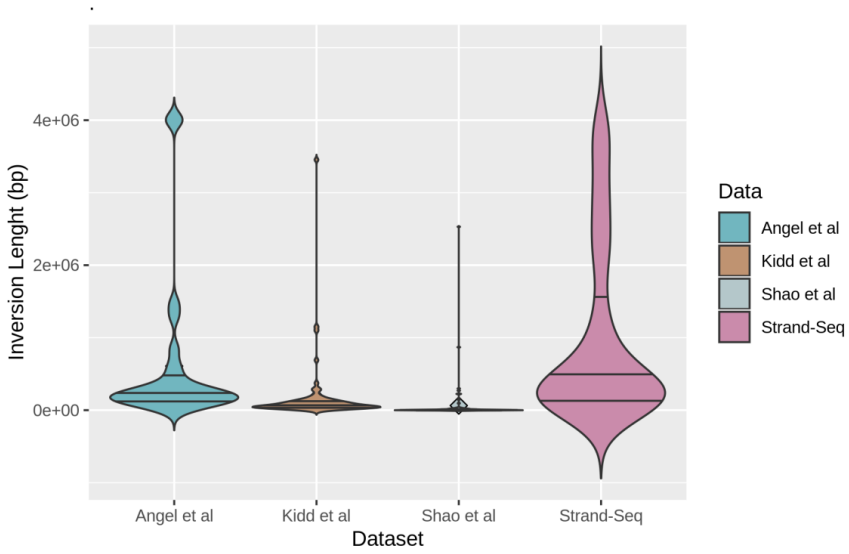


Figure 2.14 Violin plot of size distributions for each inversion call-set: Visualizing the inversion size density for each dataset.

3. Discussion

3.1 Overview

The aim of this thesis was to develop a bioinformatics workflow, with emphasis on reproducibility and transparency, for analyzing Strand-seq data. More specifically, the objectives were to develop a workflow that; 1) aligns raw read data to a reference sequence and outputs relevant alignment metrics formatted into a quality report, 2) identifies and refines putative inversions, 3) phases an external set of variants to assemble WGS haplotypes. To evaluate the performance of the pipeline that was developed we benchmarked our results against previously sequenced Strand-seq data for NA12878. To ensure reproducibility, a test-data set was compiled and included in the main repository. This test-dataset consists of a human genomic subset (chr16) and contains all the necessary files/dependencies to execute in any computational environment. Lastly, a validation assay was developed to compare Strand-seq resolved inversions with inversions for the same individual drawn from different technologies.

3.2 Pipeline Development

This section discusses and explains component and module decisions in the construction of presented pipelines.

3.2.1 Sequencing, Alignment

This first pipeline was developed with a strong emphasis on user-friendliness and customization. The workflow is easily executed by running the associated master script. The design of this program allows the user to customize the workflow to match, not only datatype but also omit/include additional scripts in a user-friendly manner. The steps taken in this pipeline are typical for any standard alignment approach, using common file-formats that allows for further manipulation of read data using a variety of different bioinformatics tools. The pipeline first aligns reads to a reference sequence, using Bowtie2. Alignment read data is saved in SAM and the converted to BAM, to minimize computational footprint. Reads are then sorted and

marked for duplicate reads. The reason for marking duplicate reads (and not removing them) is solely based around archiving purposes and the fact that there is no significant difference in run-time when marking/removing duplicate reads. In regards to the additional script that trims read data to only include chromosome 1-22, X and Y. This was an optional step to reduce the computational footprint and to remove contigs that would complicate the downstream visualization and GC bias interrogation. However, concerning variant calling, it is recommended not to omit data from the primary assembly such as; unlocalized, unplaced and alternative sequences. Particularly alt sequences, since they represent important human variation within a specific region that can not be represented by a single sequence. Alt sequences together with standard reference sequence have been proven to be crucial for constructing graph-based genome assemblies [59]. All together, graph-based alignment has the potential to allow for a more detailed and accurate variant analysis compared to alignment-based approaches [60, 61]. This suggests a future potential application for Strand-seq using graph-based genome assemblies for developing even better reference materials and more accurate variant discovery. Lastly, custom R-scripts for visualization of sequencing metrics are written in an easily readable language to allow for further customization and additional plotting, besides the standardized figures. For example, one could subset different metrics and compile informative tables for a subset of selected libraries and these tables could be further visualized using additional R-scripts.

3.2.2 Alignment Metrics and Quality Report

Besides a standardized quality report, all individual underlying metrics (for each single-cell library) are readily available in the corresponding metric folder. This allows for an even more thorough performance assessment at the single-cell level as well as aiding the researcher in selecting libraries of “good quality” for further downstream analysis. This is an important step since low-quality libraries have the potential to skew breakpoints of putative inversions and in the end, increase the number of false positives in the final refined inversion call-set.

3.2.3 Alignment Comparison

To advocate the choice of alignment approach (Bowtie2), three different alignment algorithms were validated and compared. This analysis produced comparable alignment metrics for two of the chosen aligners (Bowtie2 and Minimap2). Alignment with Hisat2 showed significantly lower alignment metrics compared to the other evaluated aligners. This might be a reflection of the intended data type for each aligner. Hisat2 makes use of a graph-based alignment and is generally considered to be the better alignment choice when working with split reads and RNA sequencing data. Hisat2 also has additional optimization options compared to Bowtie2 and Minimap2. Since all aligners were implemented using the recommended parameters, it is possible that the alignment performance of Hisat2 could be improved if available parameters were adjusted with respect to the characteristics of Strand-seq data. Of all evaluated aligners, Bowtie2 and Minimap2 produced similar performance metrics. When comparing metrics such as; PF reads aligned, estimated library size, unique reads, and unmapped reads, Minimap2 exceeds Bowtie2. However, when standardized coverage is calculated Bowtie2 exceeds Minimap2. Since this metric is in direct relation to developed pipeline's ability to capture putative inversions and haplotype assembly, Bowtie2 was chosen as the best aligner.

3.2.4 Locating Putative Inversions and Refinement of Variant Calls

After the execution of the second pipeline, a variety of outputs are generated. Besides chromosome-specific ROI files (used for calling putative inversions), InvertR provides whole-chromosome overlay figures of observed WC ratios for each single-cell library. These figures might be informative in distinguishing inversions from SCEs. Besides the mentioned figures, invertR also generates bedgraph and allele frequencies files for each call. All of which can be investigated to ensure putative inversions calls. For example, bedgraph files can be uploaded to any genome browser (i.e IGV and UCSC) to allow for visualization of reads and read orientation for any inversion call. In order to accurately and further distinguish inversions from

SCE, custom scripts are implemented, requiring calls to be present in multiple cells within the stated CI (500 bp by default). This CI can easily be changed to match any specific intervals. However, as CI is increased the number of unique calls will decrease and if a higher CI is used, manual inspection of retained calls is recommended prior to refinement, avoiding inaccurately merging of individual variant calls. The format (bed) of refined inversion calls allows for straightforward visualization of metrics associated with retained calls such as inversion-chromosome distribution, median inversion size, genotype distribution etc.

3.2.5 Haplotype Assignment

This workflow implements BreakpointR to identify genomic regions sequenced as WC. This is crucial to construct anchor haplotypes and accurately assess the haplotype origin of any given SV. This highlights one of the current limitations of Strand-seq. Since segregation patterns in observed Strand-seq libraries are completely random, 50% of sequenced regions will show a segregation pattern where reads mapping to both strands, thus requires multiple single-cells to be sequenced in order to for the genome to be represented as WC. Another fundamental problem is reflected by the sparseness of current Strand-seq data. In order to successfully assign a haplotype to a large set of SVs, multiple cells are required (figure 2.12), since the number of single-cells is in direct correlation with SNV coverage. However, recent sequencing efforts and new library preparation protocols, yielding higher genomic coverage, has the potential to substantially decrease the dependence on the number of single-cells required to accurately phase SVs. This is crucial in making Strand-seq available world-wide since a higher coverage per single-cell would allow for; lower cost per sample, faster turnaround, the potential of sequencing multiple individuals in a multiplex situation.

3.3 Pipeline Performance

This section discusses the experimental validation results from benchmarking developed pipelines.

3.3.1 Sequencing, Alignment

Developed alignment pipeline produces results comparable with previously published data [4, 6, 7, 8]. However, hyper-variable metrics exist between individual single-cells once again suggesting the need for an improved library preparation protocol. For example, out of 200 sequenced single-cells only approximately 100 cells are passing the stated good-quality criterium, resulting in discarding a large number of cells and corresponding sequence data. The need for standardized and unified definitions of metrics, such as coverage, can not be overstated. Previous sequencing efforts, all with its own coverage definitions, makes the comparison of different library preparation protocols troublesome. For example, BAIT implements a coverage calculation that excludes/masks highly mappable genomic regions, resulting in a lower coverage compared to coverage calculations from different tools, such as Aneufinder [4, 6]. Developed pipeline utilizes a standardized calculation for coverage calculation, allowing evaluation and comparison of performance using different library preparation protocols.

3.3.2 Locating Putative Inversions and Refinement of Variant Calls

Inversion analysis reveals 167 inversions constituting 22 Mb of inverted genome. This is a 480-fold increase of inverted bases per individual when compared to the 1000 genomes project [12]. This is in concordance with recent studies, showing that our genome is substantially more inverted than what previously thought [4]. In addition, it has been shown that Strand-seq can aid in resolving errors in the current version of the human reference genome, once again highlighting the importance of inversions and how this SV subtype has previously escaped detection [5, 6]. An interesting discovery is that larger chromosomes don’t necessarily harbour more inversions and that certain chromosomes harbour fewer inversions, but with a larger median inversion size. The genomic and phenotypic implications of complex SVs residing in low mappable regions could potentially be huge. Given the current knowledge gap in relation to such SVs, it is important to accurately call variants residing in these complex regions (such as segmental duplications). Developed

pipeline shows promising performance in resolving balanced inversions within such genomic context, which not only allows for a better and more accurate variant calling but could also help us improve currently available benchmarks such as GIAB and HGSV in relation to such variants.

3.3.3 Haplotype Assignment

Developed pipeline shows good phasing performance and given the average depth of sequenced libraries, using 100 single-cell libraries, 62% of all identified variants could be assigned a haplotype. Previous studies have shown that by integrating multiple data types (such as PacBio and Oxford Nanopore) with Strand-seq, it is possible to successfully phase all reported variants, without relying on parental sequence data [15,16]. Again, highlighting the potential benefits of integrating multiple technologies. Given the spareness of current sequenced Strand-seq libraries, the pipeline depends on an external set of variants to phase. However, if average coverage could be increased by designing more efficient library preparation protocols, not only would the required number of sequenced cells decrease, but also, initial SVs and SNVs could be called using Strand-seq alone. This would truly reinforce Strand-seq as a powerful tool for accurate discovery and phasing of a large variety of variants in multiple genomic contexts.

3.3.4 Inversion Validation

For all compared datasets, the proposed workflow identifies the highest number of novel inversion candidates, once again demonstrating the superiority of Strand-seq compared to other technologies. Surprisingly, only 2 variants were in concordance between all data sets. Potentially, increasing the CI this overlap could be expanded. In addition, it is possible that some discrepancy is related to lift-over complications when creating validation datasets. Inversion breakpoints for all compared studies were in reference to GRCh37 and consequently, had to be lifted to GRCh38 to match coordinates of identified breakpoints for the Strand-seq call-set. GRCh38 was chosen as the best version of the reference genome for this study since previous

studies related to segmental duplications and low-mappability regions are all implementing this version [20, 21, 22]. Strand-seq shows outstanding performance in resolving larger inversions (>1 Mb) compared to all other technologies and the median size of Strand-seq resolved inversions are in agreement with previous studies [18]. Novel Strand-seq resolved variants for NA12878 could be further verified by comparing retained call-set with high confidence data such as the Illumina Platinum high-confidence variant call set for NA1278 or any of the published GIAB reference materials. For Strand-seq studies involving non-reference samples, a subset of candidate SVs could be experimentally validated using technologies such as PCR or NGS [15].

3.4 Conclusions and Further Directions

3.4.1 Summary

This thesis presents the development of a custom bioinformatic workflow, designed to analyze Strand-seq sequencing data. The pipeline streamlines bioinformatic processes concerning sequence read alignment, variant calling and haplotype assembly, as well as outputs relevant metrics in a standardized way, allowing for reproducibility and transparency. The overall design of developed workflows (individual scripts for each process executed with pipeline specific master scripts) allows for an easy and intuitive way for users to customize specific processes of the workflow as well as allowing for the incorporation of additional bioinformatic processes. These design decisions, together with implemented programming language allows for a future-proof, highly customizable workflow. Besides the development of this bioinformatic workflow, suggested pipeline was also benchmarked using previously sequenced Strand-seq data. The results showed good performance both for sequence alignment and variant calling. Identifying 136 novel candidate inversions and phasing 62% of all variants present in the platinum genome benchmark dataset, using only 100 single-cell libraries.

3.4.2 Further Directions

Proposed pipelines were developed with user-friendliness in mind and to further emphasize this, the developed pipeline could be written as a snakemake workflow. Allowing for an even higher level of reproducible and scalable analysis. A more comprehensive validation assay for identified novel inversions should also be designed, enforcing the performance inversion calling pipeline.

3.4.3 Conclusions

The bioinformatics workflow described in this paper is the first effort to unify available Strand-seq specific tools into one streamlined workflow. Developed pipelines have been proven to be crucial in assessing the performance of different sequencing efforts (such as different library preparation protocols). Workflows are all implemented on the Omics Research Container Architecture (ORCA) platform and are currently being used as the standard workflow for genomic analysis in the Lansdorp lab at BCCRC. This workflow consists of three major blocks; alignment, inversion calling, and phasing. The advantages of a standardized approach for alignment and variant calling are many and while Strand-seq has the potential to revolutionize the genomic field as we know it today, bioinformatics workflows with emphasis on availability, reproducibility and transparency are crucial for unlocking the full potential of Strand-seq.

Bibliography

1. L. Hood, and L. Rowen. The Human Genome Project: big science transforms biology and medicine. Genome Med. 5(9):79, September 2013.
2. L. Feuk, A. R. Carson, S. W. Scherer. Structural variation in the human genome. Nature Reviews Genetics. 7:85-97, February 2006.
3. L. L. Cavalli-Sforza. The Human Genome Diversity Project: past, present and future. Nature Reviews Genetics. 6:333-340, April 2005.
4. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 431:931-945, October 2004.
5. J. Shmutz, J. Wheeler, J. Grimwood, M. Dickson, J. Yang, C. Caolie, E. Bajorek, S. Black, Y. M. Chan, M. Denys, J. Escobar, D. Flowers, D. Fotopulos, C. Garcia, M. Gomez, E. Gonzales, L. Haydu, F. Lopez, L. Ramirez, J. Retterer, A. Rodriguez, S. Rogers, A. Salazar, M. Tsai, and R. M. Myers. Quality assessment of the human genome sequence. Nature. 429:365-368. May 2004.
6. J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. Russo Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman0, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X.

Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. Science. 291(5507):1304-1351, February 2001.

7. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 409:860-921, February 2001.

8. H. E. Wyandt, G. N. Wilson, and V. S. Tonk. Human Chromosome Variation: Heteromorphism, Polymorphism and Pathogenesis. Springer, Singapore. 2017.
9. R. Karki, D. Pandya, R. C. Elston, and C. Ferlini. Defining “mutation” and “polymorphism” in the era of personal genomics. BMC Med Genomics. 8:37, July 2015.
10. M. J. P. Chaisson, A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E. J. Gardner, O. L. Rodriguez, L. Guo, R. L. Collins, X. Fan, J. Wen, R. E. Handsaker, S. Fairley, Z. N. Kronenberg, X. Kong, F. Hormozdiari, D. Lee, A. M. Wenger, A. R. Hastie, D. Antaki, T. Anantharaman, P. A. Audano, H. Brand, S. Cantsilieris, H. Cao, E. Cerveira, C. Chen, X. Chen, C. Chin, Z. Chong, N. T. Chuang, C. C. Lambert, D. M. Church, L. Clarke, A. Farrell, J. Flores, T. Galeev, D. U. Gorkin, M. Gujral, V. Guryev, W. H. Heaton, J. Korlach, S. Kumar, J. Y. Kwon, E. T. Lam, J. Eun Lee, J. Lee, W. Lee, S. P. Lee, S. Li, Patrick Marks, Karine Viaud-Martinez, Sascha Meiers, Katherine M. Munson, F. C. P. Navarro, B. J. Nelson, C. Nodzak, A. Noor, S. Kyriazopoulou-Panagiotopoulou, A. W. C. Pang, Y. Qiu, G. Rosanio, M. Ryan, A. Stütz, D. C. J. Spierings, A. Ward, A. E. Welch, M. Xiao, W. Xu, C. Zhang, Q. Zhu, X. Zheng-Bradley, E. Lowy, S. Yakneen, S. McCarroll, G. Jun, L. Ding, C. L. Koh, B. Ren, P. Flicek, K. Chen, M. B. Gerstein, P. Kwok, P. M. Lansdorp, G. T. Marth, J. Sebat, X. Shi, A. Bashir, K. Ye, S. E. Devine, M. E. Talkowski, R. E. Mills, T. Marschall, J. O. Korbel, E. E. Eichler, and C. Lee. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nature Communications. 10(1784), April 2019.
11. A. D. Sanders, S. Meiers, M. Ghareghani, D. Porubsky, H. Jeong, M. A. C. C. van Vliet, T. Rausch, P. Richter-Pechańska, J. B. Kunz, S. Jenni, B. Raeder, V. Kinanen, J. Zimmermann, V. Benes, M. Schrappe, B. R. Mardin, A. Kulozik, B. Bornhauser, J. Bourquin, T. Marschall, and J. O. Korbel. Single cell tri-channel-processing reveals structural variation landscapes and complex rearrangement processes. bioRxiv. November 2019.

12. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 526:68-74, September 2015.
13. J. Arensbergen, L. Pagie, V. D. FitzPatrick, M. de Haas, M. P. Baltissen, F. Comoglio, R. H. van der Weide, H. Teunissen, U. Võsa, L. Franke, E. de Wit, M. Vermeulen, H. J. Bussemaker, and B. van Steensel. High-throughput identification of human SNPs affecting regulatory element activity. Nature Genetics. 51:1160-1169, June 2019.
14. S. S. Ho, A. E. Urban, and R. E. Mills. Structural variation in the sequencing era. Nature Reviews Genetics. 21:171-189, November 2019.
15. P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. Hsi-Yang Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, The 1000 Genomes Project Consortium, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, and J. O. Korbel. An integrated map if structural variation in 2,504 human genomes. Nature. 526:75-81. September 2015.
16. C. Alkan, B. P. Coe, and E. E. Eichler. Genome structural variation discovery and genotyping. Nature reviews Genetics. 12(5):363-373, March 2011.
17. A. D. Sanders, M. Hills, D. Porubsky, V. Guryev, E. Falconer, and P. M. Lansdorp. Characterizing polymorphic inversions in human genomes by single-cell sequencing. Genome Research. 26:1575-1587, July 2016.

18. E. Tuzun, A. J. Sharp, J. A. Bailey, R. Kaul, V. A. Morrison, L. M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, M. V. Olson, and E. E. Eichler. Fine-scale structural variation of the human genome. Nature Genetics. 37(7):727-732, May 2005.
19. A. J. Sharp, D. P. Locke, S. D. McGrath, Z. Cheng, J. A. Bailey, R. U. Vallente, R. A. Clark, S. Schwartz, R. Segraves, V. V. Oseroff, D. G. Albertson, D. Pinkel, and E. E. Eichler. Segmental duplications and copy-number variation in the human genome. American Journal of Human Genetics. 77(1):78-88, July 2005.
20. P. J. Hastings, J. R. Lupski, S. M. Rosenberg, and G. Ira. Mechanisms of change in gene copy-number. Nature Reviews Genetics. 10(8):551-564, August 2009.
21. D. F. Easton, D. Ford, and D. T. Bishop. Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. American Journal of Human Genetics. 56(1):265-271, January 1995.
22. Y. Miki, J. Swensen, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L. M. Bennett, W. Ding, R. Bell, J. Rosenthal, C. Hussey, T. Tran, M. McClure, C. Frye, T. Hattier, R. Phelps, A. Haugen-Strano, H. Katcher, K. Yakumo, Z. Gholami, D. Shaffer, S. Stone, S. Bayer, C. Wray, R. Bogden, P. Dayananth, J. Ward, P. Tonin, S. Narod, P. K. Bristow, F. H. Norris, L. Helvering, P. Morrison, P. Rosteck, M. Lai, J. C. Barrett, C. Lewis, S. Neuhausen, L. Cannon-Albright, D. Goldgar, R. Wiseman, A. Kamb, and M. H. Skolnick. A strong candidate for breast and ovarian cancer susceptibility gene BRCA1. Science. 266(5182):66-71, October 1994.
23. M. Lin, S. Whitmire, J. Chen, A. Ferrel, X. Shi, and J. Guo. Effects of short indels on protein structure and function in human genomes. Scientific Reports. 7:9313, August, 2017.
24. A. J. F. Griffiths, W. M. Gelbart, J. H. Miller, and R. C. Lewontin. Modern Genetic Analysis. New York, W. H. Freeman, 1999.
25. H. Kaneko, K. Isogai, T. Fukao, E. Matsui, K. Kasahara, A. Yachie, H. Seki, S. Koizumi, M. Arai, J. Utunomiya, Y. Miki, and N. Kondo. Relatively common

mutations of the Bloom syndrome gene in the Japanese population. International Journal of Molecular Medicine. 14(3):439-442, September 2014.

26. T. Kaneo, S. Tahara, and M. Matsuo. Non-linear accumulation of 8-hydroxy-2′-deoxyguanosine, a marker of oxidized DNA damage, during aging. Mutation Research. 316(5-6):277-285, May 1996.
27. C. Giner-Delgado, S. Villatoro, J. Lerga-Jaso, M. Gayà-Vidal, M. Oliva, D. Castellano, L. Pantano, B. D. Bitarello, D. Izquierdo, I. Noguera, I. Olalde, A. Delprat, A. Blancher, C. Lalueza-Fox, T. Esko, P. F. O'Reilly, A. M. Andrés, L. Ferretti, M. Puig, and M. Cáceres. Evolutionary and functional impact of common polymorphic inversions in the human genome. Nature Communications. 10(4222), September 2019.
28. K. Suzuki. Nature of Mutations in genetic Disorders. In: Siegel GJ, Agranoff BW, Albers RW, et al., editors. Basic Neurochemistry: Molecular, Cellular and Medical Aspects. 6th edition. Philadelphia: Lippincott-Raven; 1999.
29. R. E. Mills, C. T. Luttig, C. E. Larkins, A. Beauchamp, C. Tsui, W. S. Pittard, and S. E. Devine. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Research. 16(9):1182-1190, September 2006.
30. G. Escaramis, E. Docampo, and R. Rabionet. A decade of structural variants: description, history and methods to detect structural variation. Brief Functional Genomics. 14(5):305-314, September 2015.
31. W. De Coster and C. van Broeckhoven. Newest Methods for Detecting Structural Variations. Trends Biotechnology. 37(9):973-982, September 2019.
32. L. Zhang, X. Zhou, Z. Weng, and A. Sidow. De novo diploid genome assembly for genome-wide structural variant detection. NAR Genomics and Bioinformatics. 2(1), March 2020.
33. S. Ardui, A. Ameur, J. R. Vermeesch, and M. S. Hestand. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acid Research. 46(5): 2159-2168, March 2018.
34. Oxford Nanopore bests PacBio. Nature Biotechnology. 37(336), April 2019.
35. M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J.

O'Grady, H. E Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan, T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, and M. Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology. 36:338-345, January 2018.

36. A. M. Wenger, P. Peluso, W. J. Rowell, P. Chang, R. J. Hall, G. T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N. D. Olson, A. Töpfer, M. Alonge, M. Mahmoud, Y. Qian, C. Chin, A. M. Phillippy, M. C. Schatz, G. Myers, M. A. DePristo, J. Ruan, T. Marschall, F. J. Sedlazeck, J. M. Zook, H. Li, S. Koren, A. Carroll, D. R. Rank, and M. W. Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology. 37:1155-1162, August 2019.
37. M. Zhao, Q. Wang, P. Jia, and Z. Zhao. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 14(S1), September 2013.
38. F. Zare, M. Dow, N. Monteleone, A. Hosny, and S. Nabavi. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. BMC Bioinformatics. 18(286), May 2017.
39. M. Pirooznia, F. S. Goes, and P. P. Zandi, Whole-genome CNV analysis: advances in computational approaches. Front Genet. 6(138), April 2015.
40. L. Zhang, W. Bai, N. Yuan, and Z. Du. Comprehensively benchmarking applications for detecting copy number variation. PLoS Comput Biol. 15(5):e1007069, May 2019.
41. S. Kosugi, Y. Momozawa, X. Liu, C. Terao, M. Kubo, and Y. Kamatani. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. 20(1):117, June 2019.
42. J. M. Zook, D. Catoe, J. McDaniel, L. Vang, N. Spies, A. Sidow, Z. Weng, Y. Liu, C. E. Mason, N. Alexander, E. Henaff, A. B. R. McIntyre, D. Chandramohan, F. Chen, E. Jaeger, A. Moshrefi, K. Pham, W. Stedman, T. Liang, M. Saghbini, Z. Dzakula, A. Hastie, H. Cao, G. Deikus, E. Schadt, R. Sebra, A. Bashir, R. M. Truty, C. C. Chang, N. Gulbahce, K. Zhao, S. Ghosh,

F. Hyland, Y. Fu, M. Chaisson, C. Xiao, J. Trow, S. T. Sherry, A. W. Zaranek, M. Ball, J. Bobe, P. Estep, G. M. Church, P. Marks, S. Kyriazopoulou-Panagiotopoulou, G. X.Y. Zheng, M. Schnall-Levin, H. S. Ordonez, P.A. Mudivarti, K. Giorda, Y. Sheng, K. Bjarnesdatter Rypdal, and M. Salit. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Scientific Data. 3(160025), June 2016.

43. D. Porubsky, S. Garg, A. D. Sanders, J. O. Jorbel, V. Guryev, P. M. Lansdorp, and T. Marshall. Dense and accurate whole-chromosome haplotyping of individual genomes. Nature Communications. 8(1293), November 2017.
44. J. M. Zook, J. McDaniel, N. D. Olson, J. Wagner, H. Parikh, H. Heaton, S. A. Irvine, L. Trigg, R. Truty, C. Y. McLean, F. M. De La Vega, C. Xiao, S. Sherry, and M. Salit. An open resource for accurately benchmarking small variant and reference calls. National Biotechnology. 37(5):561-566, May 2019.
45. J. M. Zook, B. Chapman, J. Wang, D. Mittelman, O. Hofmann, W. Hide, and M. Salit. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nature Biotechnology. 32:246-251, February 2014.
46. P. Krusche, L. Trigg, P. C. Boutros, C. E. Mason, F. M. De La Vega, B. L. Moore, M. Gonzalez-Porta, M. A. Eberle, Z. Tezak, S. Lababidi, R. Truty, G. Asimenos, B. Funke, M. Fleharty, B. A. Chapman, M. Salit, and J. M. Zook. Best practices for benchmarking germline small-variant calls in human genomes. Nature Biotechnology. 37(5):555-560, May 2019.
47. H. Parikh, M. Mohiyuddin, H. Y. K. Lam, H. Iyer, D. Chen, M. Pratt, G. Bartha, N. Spies, W. Losert, J. M. Zook, and M. Salit. svclassify: a method to establish benchmark structural variant calls. BMC Genomics. 17(64), January 2016.
48. E. Falconer, M. Hills, U. Naumann, S. Poon, E. A. Chavez, A. D. Sanders, Y. Zhao, M. Hirst, and P. M. Lansdorp. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. Nature Methods. 9(11):1107-1112, November 2012.

49. M. Hills, K. O'Neill, E. Falconer, R. Brinkman, and P. M. Lansdorp. BAIT: Organizing genomes and mapping rearrangements in single cells. Genome Med. 5(9):82, September 2013.
50. A. D. Sanders, E. Falconer, M. Hills, D. C. J. Spierings, and P. M. Lansdorp. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. Nature Protocols. 12:1151-1176, May 2017.
51. M. Ghareghani, D. Porubskỳ, A. D. Sanders, S. Meiers, E. E. Eichler, J. O. Korbel, and T. Marschall. Strand-seq enables reliable separation of long reads by chromosome via expectation maximization. Bioinformatics. 34(13):i115-i123, July 2018.
52. N. van Wietmarschen, S. Merzouk, N. Halsema, D. C. J. Spierings, V. Guryev, and P. M. Lansdorp. BLM helicase suppresses recombination at G-quadruplex motifs in transcribed genes. Nature Communications. 9(1):271, January 2018.
53. C. Claussin, D. Porubský, D. C. J. Spierings, N. Halsema, S. Rentas, V. Guryev, P. M. Lansdorp, and M. Chang. Genome-wide mapping of sister chromatid exchange events in single yeast cells using Strand-seq. Elife. 6, December 2017.
54. B. Bakker, A. Taudt, M. E. Belderbos, D. Porubsky, D. C. J. Spierings, T. V. de Jong, N. Halsema, H. G. Kazemier, K. Hoekstra-Wakker, A. Bradley, E. S. J. M. de Bont, A. van den Berg, V. Guryev, P. M. Lansdorp, M. Colomé-Tatché, and F. Foijer. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. Genome Biology. 17(115), May 2016.
55. A. D. Sanders, M. Hills, D. Porubský, V. Guryev, E. Falconer, and P. M. Lansdorp. Characterizing polymorphic inversions in human genomes by single-cell sequencing. Genome Research. 26(11): 1575-1587, November 2016.
56. D. Porubský, A. D. Sanders, N. van Wietmarschen, E. Falconer, M. Hills, D. C. J. Spierings, M. R. Bevova, V. Guryev, and P. M. Lansdorp. Direct

chromosome-length haplotyping by single-cell sequencing. Genome Research. 26(11): 1565-1574, November 2016.

57. D. Porubsky, A. D. Sanders, and A. Taudt. (2019). breakpointR: Find breakpoints in Strand-seq data. R package version 1.4.0, https://github.com/daewoooo/BreakPointR
58. A. D. Sanders, S. Meiers, M. Ghareghani, D. Porubsky, H. Jeong, M. A. C.C. van Vliet, T. Rausch, P. Richter-Pechańska, J. B. Kunz, S. Jenni, B. Raeder, V. Kinanen, J. Zimmermann, V. Benes, M. Schrappe, B. R. Mardin, A. Kulozik, B. Bornhauser, J. P. Bourquin, T. Marschall, J. O. Korbel. (2020). Single cell tri-channel-processing reveals structural variation landscapes and complex rearrangement processes. bioRxiv 849604; doi: https://doi.org/10.1101/849604
59. J. Pritt, N. Chen, and B. Langmead. FORGe: prioritizing variants for graph genomes. Genome Biology. 19:220, December 2018.
60. K. Daehwan, J. M. Channhee Park, C. Bennet, and S. L. Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology. 37, 907-915, August 2019.
61. S. Garg, M. Rautiainen, A. M. Novak, E. Garrison, R. Durbin, and T. Marchall. A graph-based approach to diploid genome assembly. Bioinformatics. 34:13, July 2018.

Appendices

A. Methods

Library Selection Criterium

To avoid errors introduced by low-quality libraries, the following quality criteria were developed. Good quality libraries are defined as having > 200 reads/Mb, even coverage profile, low background reads, no structural rearrangements (i.e copy number changes or aneuploidy events).

Library Preparation and Sequencing

To experimentally assess the performance of developed pipelines, previously sequenced Strand-seq data for NA12878 was used (100 libraries). The next section outlines in detail how this data was generated.

Cells and Cell Culture

Cell line GM12878 (Epstein-Barr virus-transformed B-lymphocyte cell line) was acquired from Coriell Institute of Medical Research. Cells were cultured in RPMI 1640 medium (Gibco) supplemented with 15% FBS (Stigma Aldrich) in 37*C at 5% CO2. BrdU was added to exponentially growing cells for 24h.

Single Cell Sorting

After harvesting cells, nuclei were isolated by resuspension in nuclear isolation buffer (100 mM Tris-HCl at pH 7.4, 150 mM NaCl, 1 mM CaCl2, 0.5 mM MgCl2, 0.1% NP-40, 0.2% BSA). Cells cultured without BrdU were added as an internal control for Hoechst fluorescence. Hoechst-33258 and propidium iodide was used to stain nuclei and was incubated on ice for 30 min. Based on low Hoechst fluorescence (quenched by BrdU in DNA) and PI, Nuclei of cells that endured cell division in the presence of BrdU were sorted out using a MoFLo Atrios cell sorter

(Beckman Coulter) and dispensed into 96-well skirted PCR plates (4Titude) containing 5 ul/well freeze medium (pro-freeze CDM freeze medium containing 15% DMSO)

Library Construction

Modified versions of a previously described protocol (Falconer et al. 2012) were used for library construction. Enzymatic reactions were performed in smaller volumes and buffer/enzyme concentrations were kept at the same levels to scale for production on a Bravo automated liquid handling platform (Agilent). AMPure XP paramagnetic beads (Agencourt AMPure, Beckman Coulter) was used for DNA clean-up. After adapter ligation and PCR (17 cycles), two AMPure bead clean-ups were carried out using 1.2x bead volume.

Next-generation Sequencing

250- to 300-bp size-range fragments were purified for pooled libraries using 2% E-gel Agarose EX-gels (Invitrogen). 2100 Bioanalyzer (Agilent) was used to assess DNA quality. DNA was then quantified on the Qubit 2.0 fluorometer (Life Technologies). Clusters were generated on the cBot and reads 100 bp (paired-end) were generated using the HiSeq 2500 sequencing platform (Illumina) according to the manufacturer's instructions. 96 single-cell libraries were pooled together and sequence in one lane of the rapid run flow cell. Each plate included two 10-cell controls and two zero-cell controls.

B. Test Data-set

To evaluate the performance of developed pipeline as well as ensuring all required dependencies are installed correctly (pipeline runs without errors), a custom-designed test data-set is included. This data-set consists of a subset of five randomly selected previously sequenced Strand-seq libraries for NA12878 filtered to only include chromosome 16. All selected libraries are passing the stated quality evaluation criterium. Chromosome 16 was chosen due to the magnitude of complexity (i.e segmental duplications, homopolymers, etc) and therefore are highlighting the true potential of Strand-seq when it comes to accurately call variants/phase variants residing in these regions. All necessary files, such as subsetted reference genome (both in Fasta format as well as Bowtie2 index reference genome) for the chosen region, region table for inversion calling and SNPs call-set filtered from Illuminas Platinum call-set are included. For information on how to execute, see section 2.4. Note, Due to the sparseness of sequenced Strand-seq libraries, the phasing pipeline will not run on selected libraries (currently relying on multiple libraries to successfully identify WC regions and construct anchor haplotypes). For more info on how multiple single cells and coverage related to phasing performance, see figure 2.12.

C. Inversion Validation Experiment

To evaluate the performance of developed inversion pipeline, retained high confidence refined inversion call-set was compared to previously published data for the same individual (Kidd et al. 2008, Angel et al. 2016, Shao et al. 2018). Published inversion call-sets were lifted from GRCh37 to GRCh38 using a custom script. Lifted call-sets were then converted to bed files to allow for easy genome browser inspection of retained calls. Strand-seq resolved inversions were then uploaded together with described call-sets to IGV genome browser for manual inspection of concordance between calls for each methodology. A confidence interval of 500 bp was used for breakpoint comparison. Table C.1 describes the characteristics of each validation data-set.

Dataset	Inversions (n)	Inverted genome (Mb)	Mean Inversion Size (Kb)	Technology	Validation
Kidd et al.	79	26.8	339	Fosmid insert sequencing	NGS/PCR
Angel et al.	27	16.4	607	Bionano	NGS
Shao et al.	81	10.7	132	PacBio, Bionano, Long-read sub-alignment	PCR

Table C.1 Inversion validation data-set overview: Overview of included data-sets used for inversion validation.

D. How to Execute

Clone the repository (https://github.com/mattsada/sspipe) and cd main directory (sspipe-master). Put raw reads (fastq or compressed fastq files in gzip) in the corresponding subfolder. Install dependencies accordingly. Download the reference genome (scripts provided). Each pipeline is executed with its own master script (master.sh).

E. Dependencies and Maintaining

All needed dependencies are listed below. Note, Picard jar file is already included in the pipeline (sspipe-master/packages/). To acquire all listed dependencies one could execute the “install-dep.sh” listed in the main folder.

Package	Version	Environment
samtools	0.1.19-44428cd	bash
bowtie2	2.2.3	bash
picard	2.18.11	bash/java
BAIT	1.0	bash
breakpointR	3.9	R
dplyr	0.7.8	R
tidyverse	1.2.0	R
ggplot2	3.1.0	R
invertR	0.1	R
bedtools	2.17.0	bash
strandphaseR	0.1	R
DNAcopy	3.9	R
changepoint	2.2.2	R
genomicRanges	3.9	R
VCF tools	0.15	R
BCF tools	1.9	bash
htslib	1.9	R

Table E.1 List of dependencies: Complete list of dependencies required to execute developed pipelines. To acquire all listed dependencies one could execute the “install-dep.sh” listed in the main folder.