Exploiting Public Human Genome NGS Datasets to Characterize Repetitive DNA and Recover Assembly gaps

Thesis submitted for the degree of

Doctor of Philosophy

At the University of Leicester

Denye Nathaniel Ogeh

Department of Genetics and Genome Biology

University of Leicester

May, 2018

Exploiting Public Human Genome NGS Datasets to Characterize Repetitive DNA and Recover Assembly gaps

Denye Nathaniel Ogeh

With the advent of Next Generation Sequencing (NGS), we have witnessed the generation of enormous volumes of short read sequence data, cheaply and on short time scales. Nevertheless, the quality of genome assemblies generated using NGS technologies has been greatly affected by this innovation, compared to those generated using Sanger DNA sequencing. This is largely due to the inability of short read sequence data alone to scaffold repetitive structures, creating gaps, inversions and rearrangements and ultimately resulting in assemblies that are, at best, draft forms (by draft we mean, assembly that is only a preliminary result that will require more work to be done to make it a more complete and accurate representation of the genome).

Single molecule long-read sequencing (SMS) technologies on the other hand, address this challenge by generating sequences with greatly increased read lengths, offering the prospect to better recover these complex repetitive structures, concomitantly improving assembly quality.

Following this development, we evaluate the ability of SMS data (specifically Pacific Biosciences SMRT data and Oxford Nanopore MinION data from human genomes) to recover poorly represented repetitive sequences (specifically, GC-rich human minisatellites), identify novel transposable element insertions and enable the closing of gapped regions.

Our results show that by using single molecule sequencing and long read technology, poorly represented repetitive sequences (specifically, minisatellites and L1s) and other missing elements in published human genome assemblies can be characterized by developing custom software, scalable for the analysis of single molecule long-reads (particularly, Pacific Biosciences' SMRT technology). The tool designed is cross-platform, thus, giving computational and non-computational biologists a straightforward approach and less technical platform for local analysis of specific poorly characterized DNA sequences.

Acknowledgements

I acknowledge the grace and hand of God upon my life. If it were not for him, I would not have come this far. To him alone be praise and glory. Amen.

I want to acknowledge Richard Badge for his fatherly support and encouragement throughout my study years at Leicester. I am grateful for his assistance with my PCR analysis.

To all G18/19 lab members, I owe you my gratitude.

To my family here and overseas, I am immensely grateful for the prayers and encouragements received all these years.

And to my lovely wife, daughter and son, I cannot express how fulfilled I am knowing that you were part of this journey. Together we have made it. I will forever be grateful to you all.

Table of contents

Index of tables

Index of figures

Abbreviations

BACs:	Bacteria Artificial Chromosome
BLASR:	Basic Local Alignment with Successive Refinement
BLAST:	Basic Alignment and Search Tool
bp:	Basepair
BWA:	Burrows-Wheeler Aligner
CA:	Celera Assembler
contigs:	Contiguous sequence
cDNA:	Coding DNA
CGI:	Common Gateway Interface
CPU:	Central Processing Unit
DAS:	Direct attached server
DBG:	de-Bruijn graph
DNA:	Deoxyribonucleic acid
ERVs:	Endogenous retroviruses
euL1db:	European database of human-specific L1
FDR:	False Discovery Rate
GAIIx:	Genome AnalyserIIx
GPU:	General Purpose Computing
GS:	Genome sequencer
GUI:	Graphical User Interface
HGP:	Human genome project
HPC:	High Performance Computing
IGV:	Integrative Genomics Viewer
IHGSC:	The International Human Genome Consortium
IRCAN:	The Institute for Research on Cancer & Ageing of Nice
IT:	Information Technology
LINE:	Long Interspersed Nuclear Elements
L1/LINE-1:	Long Interspersed Nuclear Elements 1
LTR:	Long Terminal Repeats
Mb:	Megabase
MPSS:	Massively Parallel Signature Sequencing
mRNA:	Messenger ribonucleic acid
MVR-PCR:	Minisatellite Variant Repeat mapping by PCR

ncDNA:	Non-coding DNA
NGS:	Next Generation Sequencing
OLC:	Overlap Layout and Consensus
ONT:	Oxford Nanopore Technologies
ORF:	Open reading frame
OS:	Operating System
PacBio:	Pacific Biosciences
PCR:	Polymerase chain reaction
PERL:	Practical Extraction and Report Language
PRDM9:	PR domain-containing 9
RAM:	Random Access Memory
RNA:	Ribonucleic acid
ROI:	Region of interest
Samtools:	Sequence Alignment/Map Tools
SBS:	Sequencing by synthesis
SINES:	Short Interspersed Nuclear Elements
SMS:	Single Molecule Sequencing
SMRT:	Single Molecule Real Time
SNPs:	Single Nucleotide Polymorphisms
SOLiD:	Sequencing by Oligonucleotide and Ligation Detection
SRA:	Sequence Read Archive
SSRs:	Simple sequence repeats
TIGR:	The Institute of Genomic Research
tRNA:	Transfer ribonucleic acid
tSMS:	True single-molecule-sequencing technology
UTR:	Untranslated region
UCSC:	University of California Santa Cruz
VNTR:	Variable Number Tandem Repeat
WGS:	Whole genome shotgun
WGSA:	Whole genome shotgun assembly
ZMW:	Zero mode waveguides

1. Introduction

1.1 History of genome sequencing

Genome Sequencing is the process of determining the exact order and organization of nucleotides in a genome. The desire to perform genome sequencing dates back to the 1950's when Watson and Crick first described the structure of the DNA helix and noted its potential for stable replication of information (Watson and Crick, 1953). This discovery opened a new window in modern genomics, paving the way for the elucidation of DNA replication, gene expression, and protein synthesis. Holley and co-workers (Holley *et al.,* 1965) sequenced the first nucleic acid molecule (the *Escherichia coli* alanine tRNA) using ribonuclease digestion. This led to the determination of model structures for tRNA's under the assumption of base pairing analogous to that found in the DNA double helix (Holley *et al*., 1965). In the year 1970, a new technique by which large DNA molecules were fragmented into pieces was introduced following the discovery of Type II restriction enzymes by Hamilton and co-workers (Smith and Wilcox, 1970; Kelly and Smith, 1970). The fragmented DNA had termini of defined sequence and so could function as starting points for today's sequencing techniques (Smith and Wilcox, 1970; Kelly and Smith, 1970). Various separation methods were employed including 2D chromatography electrophoresis (Danna and Nathans, 1971) and in principle, incomplete digestion of a small fragment of DNA from one end by an exonuclease could lead to the determination of the sequence, by either analysis of the terminal base of each partial digestion product, or nucleotide-specific shifts in the location of fragment following 2D separation of the products (Danna and Nathans, 1971). Although these efforts did not successfully determine the complete sequences of any gene, certain useful regulatory signals were sequenced, such as operator sequences from the *E.coli lac* operon (Gilbert and Maxam, 1973) and mutant operators from phage lambda (Maniatis *et al*., 1974). Over the period 1975 -1976, sequencing of DNA molecules was done using the Sanger-Coulson and Maxam-Gilbert methods. Sanger and Coulson (Sanger

and Coulson, 1975) developed a sequencing technique called the 'plus and minus' technique which formed the basis upon which modern sequencing technologies have been built for the past 30 years (Hutchinson, 2007). The plus and minus technique used DNA polymerase to synthesize from a primer, adding radiolabelled nucleotides prior to performing two second polymerization reactions: a '+' reaction, where only a single type of nucleotide is present, resulting in all extensions ending with that base. And a '-' reaction where three other types of nucleotides are used, generating sequences up to the position before the next missing nucleotide. The position of nucleotides at each position in the covered sequence is determined by running the products on a polyacrylamide gel and comparing between the eight lanes (Sanger and Coulson, 1975). This method enabled the determination of the order of bases in single-stranded DNA molecules; it was successfully applied in determining two sequences of bacteriophage φX174 DNA using a synthetic decanucleotide, and a restriction enzyme digestion product as primers. However, establishing the right sequence from this technique sometimes required further evidence (Sanger and Coulson, 1975)), and the technique was limited to sequencing single stranded DNA molecules.

In a bid to overcome the limitation posed by the Sanger-Coulson's plus and minus method, Maxam and Gilbert in 1976 (Maxam and Gilbert, 1977) developed a new technique that enabled the sequencing of both double and single stranded DNA molecules. This was a chemically driven method for determining the nucleotide sequence of a terminally radio-labelled DNA molecule by partially fragmenting it at the occurrence of single bases and combination of bases. With four different chemical reactions cleaving DNA preferentially at A or G, G, C, and C or T residues, the single-stranded radioactive fragments generated were resolved by polyacrylamide gel electrophoresis. The length of each fragment showed the position of the base combinations and, the DNA sequence could be read by from the pattern of the radioactive bands (Maxam and Gilbert, 1977). Even though this method could sequence about 100 bases, it was still limited by the resolving power of

polyacrylamide gels (Maxam and Gilbert, 1977). The technique gained popularity in the late 70's by allowing direct use of purified DNA for sequencing, rather than requiring cloning, as well as for being largely independent of DNA secondary structure. The need for strand separation, use of hazardous chemical reagents, and scalability issues, caused this technique to fall out of use (Sanger, Nicklen and Coulson, 1977). The routine use of the Maxam-Gilbert technique was also frustrated by difficulties in making self-contained DNA sequencing kits.

As speed and accuracy became crucial factors in sequencing reactions, Sanger and Coulson in 1977 developed the dideoxy / chain termination method (Sanger, Nicklen and Coulson, 1977). Similar in principle to the plus and minus method, the dideoxy technique, made use of chain-terminating analogues of the normal deoxynucleoside triphosphates (Sanger, Nicklen and Coulson, 1977). The absence of a 3' hydroxyl group in the deoxynucleoside triphosphate, when incorporated into a growing DNA chain, prevented further addition of nucleotides to the chain hence, the term chain-termination sequencing (Sanger, Nicklen and Coulson, 1977). The complete sequencing of the 5385 base pair (bp) bacteriophage φX174 genome and the bacteriophage T4 genome were carried out using this technique (Sanger *et al.*, 1977). The technique represented improvement in the number of nucleotides that could be determined (about 300 bases) from the 3' end of the primer, when compared to the Maxam-Gilbert technique, which achieved around 100 bases (Sanger, Nicklen and Coulson, 1977). Further applications of this technique include the sequencing of the mammalian mitochondria DNA (Barrell *et al.,* 1980). Although achieving longer read lengths than Maxam-Gilbert, the early dideoxy read length still constrained the sequencing of longer DNA molecules. To overcome this constraint, Sanger, in 1977, developed the shotgun sequencing technique (Sanger *et al*., 1977), which involved the random fragmentation of DNA from the target genome into smaller pieces. The pieces were then ligated into plasmid cloning vectors that could be propagated in bacteria. The vector provided a means to amplify one DNA molecule (by growing huge numbers of bacteria)

and also as the vector was of known sequence, a primer binding site in the vector could be used to sequence the cloned unknown fragment. Assembly of cloned fragments was done based on overlapping regions to form contiguous sequences (contigs), representing a reconstruction of the original DNA. Sequencing more clones, could bridge gaps between contigs, thus giving the complete sequence. In 1982, Sanger utilized this technique in the sequencing of the 48.5 kilobase (thousand basepair, kb) phage lambda (λ) genome (Sanger *et al*., 1982). The shotgun technique became the accepted standard for genome sequencing as well as opening up a window for subsequent sequencing of other viral and organellar genomes using similar methods. Examples of sequencing landmarks carried out with shotgun-related techniques were: the sequencing of the 192kb vaccinia genome (Goebel *et al*., 1990), the 229kb human cytomegalovirus genome (Bankier *et al*., 1991), the 186kb genome of the smallpox virus (Massung *et al*., 1994), the 187kb mitochondrial genome of a liverwort (*Marchantiophyta*) (Ohyama, 1996), as well as the 121kb chloroplast genome of the same species (Ohyama, 1996).

In an attempt to find homologs of human genes of interest, Andre Goffeau and others in 1989, set up a European Consortium for the sequencing of the budding yeast genome (*Saccharomyces cerevisiae*) (Goffeau *et al*., 1996). The 12 megabase (million bp, Mb) budding yeast genome was the first Eukaryote genome to be sequenced (Goffeau *et al*., 1996). Sequencing the *S.cerevisiae* genome was at the time very important to researchers since this eukaryote was a model organism, had a relatively small genome, and its successful sequencing could provide lessons for the eventual sequencing of larger eukaryotic genomes such as *Homo sapiens* (estimated at 3000Mb) (Goffeau *et al*., 1996).

Another striking breakthrough was the complete sequencing of the genome of the free-living bacterium *Haemophilus influenzae* by a team of researchers from The Institute of Genomic Research (TIGR) and Hamilton Smith of Johns Hopkins University. This 1.8Mb bacterial genome was sequenced and assembled using new computational techniques developed at TIGR

(Fleischmann *et al.,* 1995), a research institute that was founded by Craig Venter. Venter's approach involved whole genome shotgunning (WGS); a strategy to reduce the cost of subcloning whole genomes into progressively smaller vectors prior to sequencing. By introducing a single shotgun step the process was greatly simplified but faced the computational challenge of assembling millions of short fragments on feasible timescales. Using this new technique, the entire *H.influenzae* 1.8Mb genome was fragmented into smaller pieces and the fragments sequenced and then computationally assembled (Fleischmann *et al*., 1995).

The Hierarchical shotgun sequencing (hereafter HS) technique is an approach that fragments DNA into pieces of about 150kb. These fragments are inserted into high capacity cloning vectors (initially Bacterial Artificial Chromosomes (BACs), and later P1 Artificial Chromosomes (PAC) or Yeast Artificial Chromosome (YAC)) vectors, and then transformed into *E.coli* (or *S.cerevisiae*) where they are replicated and stored. The isolation of the BAC inserts and their subsequent mapping determines the location and order of each cloned fragments within the target genome. The sequencing of the yeast genome was performed using the HS technique, which was particularly amenable to the chromosome-by-chromosome approach adopted by the *S.cerevisiae* genome sequencing consortium (Goffeau *et al*,.1996).

Precisely a year after Goffeau and colleagues set up the worldwide collaboration for the sequencing of the first eukaryote, a three-step program to generate genetic maps, physical maps and ultimately, the complete DNA sequence map of the human chromosomes, was established as a collaboration between the Department of Energy and the National Institute of Health in the United States of America (Goffeau *et al*., 1996). However, working drafts of the human genome were however not made public until 2000, by both the International Human Genome Sequencing Consortium (Lander *et al*., 2001) and by Celera Genomics (Venter *et al*., 2001). The draft genomes as reported by both groups had many shortcomings such as gaps and assembly ambiguities.

The HS approach is less error prone when assembling shotgun fragments into contigs because the size and chromosomal location for each BAC is known, and because there are fewer fragments to assemble. However, it is time consuming and very expensive to implement. WGS on the other hand, though less expensive and faster than the HS, is prone to more errors due to mis-assembly of the finished sequence. Using WGS to sequence repeat rich regions (telomeres, centromeres, and satellite arrays) in a genome poses a challenge for assembly (in the absence of a genetic map) as it is difficult to infer the relative position of such highly similar reads within the genome. Because WGS involves the fragmentation of DNA into smaller pieces, it is possible to generate sequence reads that do not span long repetitive sequences. Also, the loss of spanning read information means that assembly programs are unable to scaffold reads into contiguous sequences. With multiple copies of reads being nearly identical, the tendency is high for assembly programs to assemble the reads into single, collapsed contigs.

In 1996, another breakthrough aimed at overcoming many of the limitations posed by the Sanger sequencing method was developed. This technique relied on the detection of pyrophosphate release when a nucleotide is incorporated in a sequencing reaction. The DNA sequence is determined based on the emitted fluorescence upon nucleotide incorporation, guided by the fact that only one out of the possible bases (A, G, C, or T) is available for incorporation at any time. Pyrosequencing served as an alternative technique for a detailed characterization of nucleic acids and provided a highly cost-effective sequencing strategy when compared to Sanger's chain termination method. As the nucleotide chain is built while the base identity is determined, the pyrosequencing method is termed a form of sequencing by synthesis (SBS). In 2004, 454 Life sciences introduced a massively parallel version of pyrosequencing, which reduced the sequencing cost by 6-fold when compared to the prevailing Sanger method (Barba, Czosnek and Hadidi, 2014). In 2005, 454 technical developments led to the GS 20 sequencer, which produced 20 million bases per run. Over the years since, Roche 454 continued to develop

the sequencing platform to operate more quickly and generate more accurate sequences.

In 2005, Solexa released the Genome Analyser which implemented SBS but using reversible dye-terminator chemistry. This chemistry enabled controlled incorporation, base interrogation and terminator removal. The GAIIx sequencer could generate up to 85 billion base pairs of sequencing data in a single run (Barba, Czosnek and Hadidi, 2014). In 2007, Illumina acquired Solexa and continued to develop their technology. The current HiSeq (HiSeq 4000) and MiSeq series of instruments can produce up to 1500Gbp and 1.5Gbp of sequence data, respectively, per run (Barba, Czosnek and Hadidi, 2014). Illumina sequencers generate very short reads with lengths in the range of 75bp – 100bp, whereas, 454 was able to generate much longer reads (>400bp and <=1000bp), close to that produced by Sanger sequencing.

SOLiD (Sequencing by Oligonucleotide and Ligation Detection) is another extant sequencing strategy developed and sold by Life Technologies since 2006 (Barba, Czosnek and Hadidi, 2014). The SOLiD system can generate sequences with length and throughput comparable to Illumina sequencing. SOLiD generates sequences with up to 99.94% accuracy owing to its high-fidelity ligase enzymology, primer reset functionality, and two-base encoding technology (Pandey, Nutter and Prediger, 2008).

1.1.1 First single molecule sequencers

All the previously discussed systems rely upon the clonal amplification of single DNA molecules, to yield enough signal for base discrimination by fluorescence / chemiluminescent detection. However, confocal microscopy is at least in principle able to visualize single fluorophores. In 2009, the Helicos Genetic Analysis System platform was the first commercially available NGS system using single molecule fluorescent sequencing. Single molecule fluorescent sequencing involves the imaging of individual fluorophore molecules, with each matching to a base. In this scheme, hybridization of DNA fragments was first carried out on disposable glass flow cells (Thompson and Steinmann, 2010). The addition of fluorescent bases was carried out one after the other with a

terminating base used to pause the process until an image has been captured. Analysing the image enables determination of one base from each DNA sequence. After imaging, the fluorescent terminator label was cleaved and a repetition of the process carried out until all bases of the DNA or RNA fragment have been sequenced (Thompson and Steinmann, 2010).

1.1.2 Semiconductor sequencing

In 2011, a new dimension to sequencing was developed by Ion Torrent Systems, which was later acquired by Life technologies, but now known as Thermofisher. This technology is based on a semiconductor chip's ability to convert chemically encoded incorporation events (addition of A, C, G, or T to a template strand) into digital information (0, 1) (http://www.thermofisher.com/uk/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html#). The ion torrent technology is based on the principle that the incorporation of a base into the strand of DNA by a polymerase causes the release of a hydrogen ion as a by-product, changing the pH of the solution, which can be detected by the proprietary semiconductor ion sensor.

1.1.3 Return of single molecule sequencing

Pacific Biosciences (also in 2011) developed a long-read sequencing strategy called Single Molecule, Real-Time Technology (SMRT). This technology utilizes novel, single molecule sequencing techniques and advanced real-time analytics (Fatih Ozsolak, 2013). The technology is centred on SMRT cells, each patterned with 150,000 zero mode waveguides (ZMW), each of which contains a single DNA polymerase providing an opportunity to observe DNA sequencing in real-time (Fatih Ozsolak, 2013). The Pacific Biosciences (PacBio, hereafter) RS II system observes the ZMWs at once, thereby doubling its throughput per experiment. As the sequencing reaction progresses, the software analyses the dwell time of fluorescently labelled nucleotides in the illuminated region at the bottom of the ZMW, where the polymerase is localised. Diffusing nucleotides move very quickly and are only briefly illuminated, whereas nucleotides in the process of being incorporated by the polymerase dwell (sit still) for a (relatively)

long time. Incorporation cleaves off the fluorophore and the cycle starts again – hence real time (Fatih Ozsolak, 2013). In real time, these light pulses are forwarded to the primary analysis pipeline where proprietary algorithms convert these signals into either of A, C, G, or T base calls, along with each event's unique set of quality measures (Fatih Ozsolak, 2013). Once a base call is completed, further secondary analysis can be carried out on the data. This technology is very useful for sequencing aimed at comprehensive characterization of genetic variation, identification, annotation and full deciphering of genomic structures, as well as the identification of base modifications to assist in characterizing epigenetic regulation and DNA damage (Fatih Ozsolak, 2013).

1.1.4 Nanopore sequencing

DNA sequencing using protein nanopores has been under development since 1995 and was made available for testing by users in 2014. This technique was first practically implemented by Oxford Nanopore Technologies (ONT) and works on the principle that when voltage is applied across a non-conducting membrane bridged by proteins containing nanopores (a hole with internal diameter of ~1 nanometer) immersed in a conducting fluid, current is discharged as a result of the conduction of ions through the nanopore (Kasianowicz *et al*., 1996). The amount of current discharged is associated with the size and shape of the pore and is in turn affected by the type of base (A, C, G, or T), strand of DNA, or any other molecule that passes through the pore (Kasianowicz *et al*., 1996). This change in current presents an opportunity for direct sensing of the DNA sequence. The development of the nanopore sequencing technique has enabled genome sequencing to proceed without the need for a PCR amplification phase or a chemical labelling phase (Kasianowicz *et al*., 1996).

In summary, recent developments in single molecule sequencing have improved sequenced read lengths by more than 100% when compared to those generated from other NGS technologies. With higher quality reads and depth, complete genome assemblies which once could not be achieved, have now

become commonplace (Peng *et al*., 2016; Pendleton *et al*., 2015; Jain, Koren *et al*., 2018).

1.2 Strategies for genome sequencing

In the recent past, great effort has been made by biologists and other researchers towards improving the techniques for analysing DNA on a large scale, generating descriptive information about the genomes of many organisms, as well as establishing advanced experimental and computationally driven methods for understanding genome structure and function (Green, 2001). This effort has yielded profound results including the sequencing of the genomes of model organisms such as *Saccharomyces cerevisiae* (*S. cerevisiae*) ('The yeast genome directory', 1997)*, Caenorhabditis elegans* (*C.elegans*) ('Genome sequence of the nematode C. elegans: A platform for investigating biology', 1998)*, Drosophila melanogaster* (*D. melanogaster*) (Adams, 2000)*, Arabidopsis thaliana* (*A. thaliana*) (The Arabidopsis Genome Initiative, 2000)*,* and also *Homo sapiens* (*H. sapiens*) (Venter *et al*., 2001; Lander *et al*., 2001)*.* All these genome sequences were determined using variations of the sequencing technology developed by Sanger (Sanger *et al*., 1977). Many other techniques have been developed for the sequencing of genomes. To place the technologies underpinning this thesis in context, Sanger sequencing (refer to page 3) and these other techniques are discussed briefly below.

1.2.1 Sanger's shotgun fragmentation

This technique served as the framework for the sequencing of the human genome. Figures 1 – 3 show a typical implementation of the shotgun technique. In Figure 1, the target DNA is randomly fragmented into pieces of about 2 – 3kb in size. These random fragments are then cloned and sequenced. In Figure 2, the sequenced reads are assembled into a set of contiguous reads (contigs) based on read overlap. Finally, Figure 3 shows the finishing stage involving the generation of the final sequence, free of any gaps.

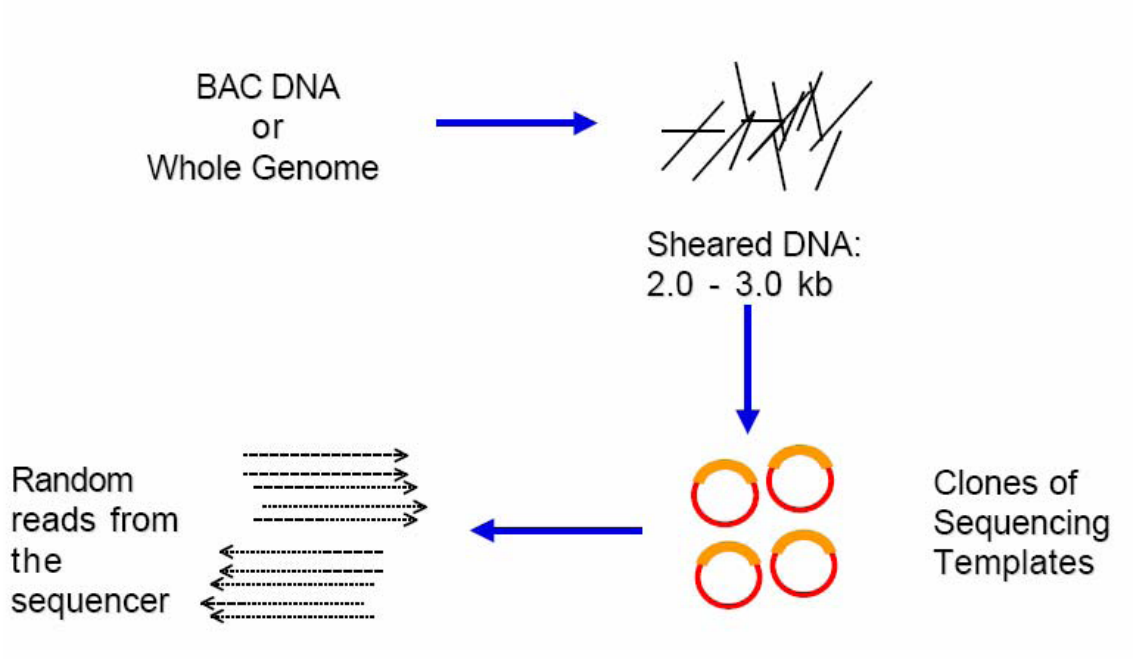

Figure 1: Random phase of DNA fragmentation

Source: Strategies for the systematic sequencing of complex genomes (Green, 2001)

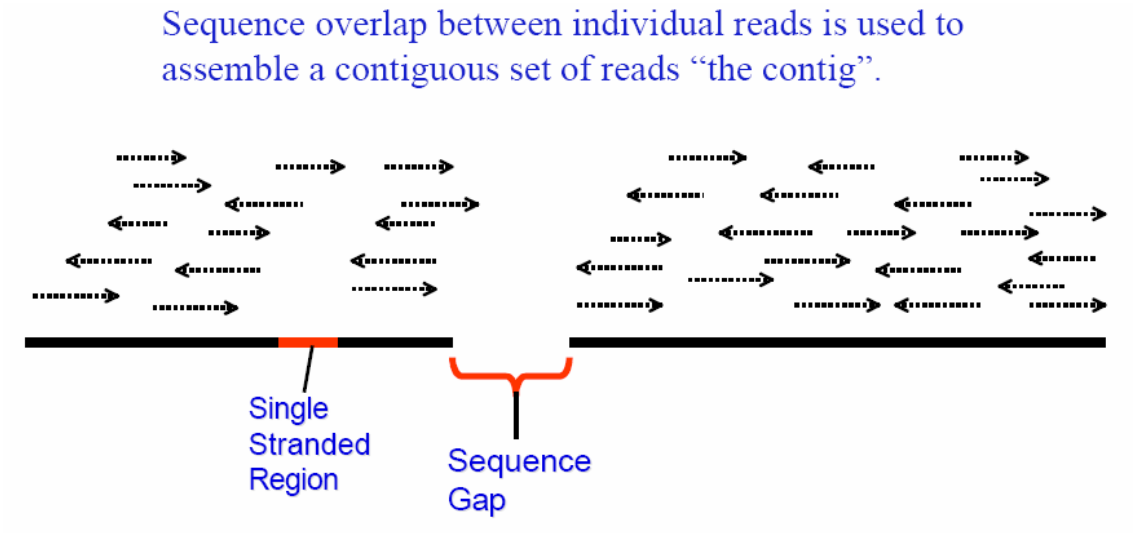

Figure 2: Assembling of overlapping reads

Source: Strategies for the systematic sequencing of complex genomes (Green, 2001)

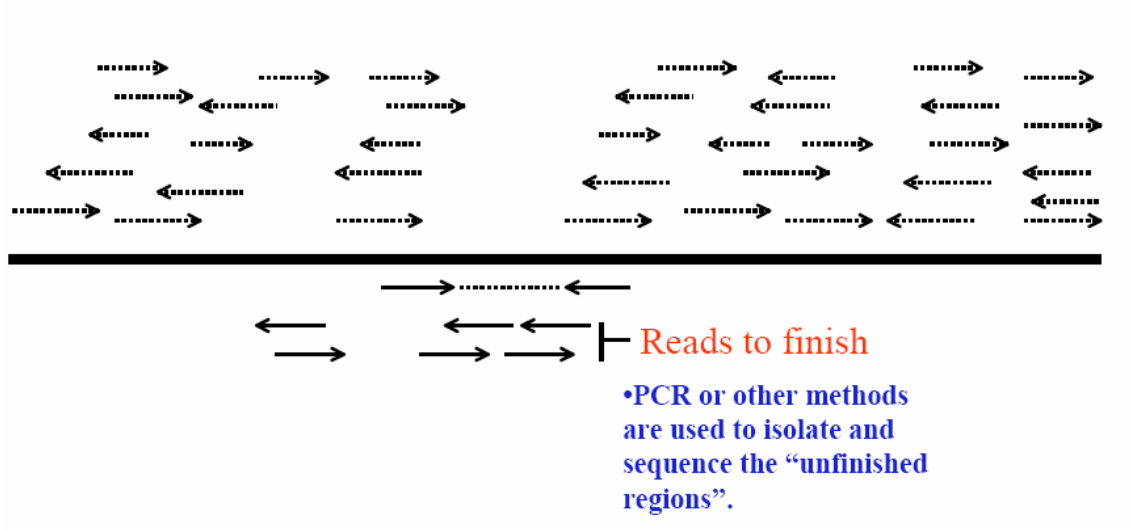

Figure 3: Sequence finishing stage.

Source: Strategies for the systematic sequencing of complex genomes (Green, 2001)

So far, the shotgun technique has been applied to genome sequencing using three different approaches. These are hierarchical shotgun sequencing (clone-by-clone sequencing or map-based shotgun sequencing), whole genome shotgun sequencing, and hybrid shotgun sequencing (see section 1.4, page 21-22). We mentioned these approaches (see section 1.4, page 21 - 22) to acknowledge their implementation.

1.2.2 Massively Parallel Sequencing

An improvement to the standard shotgun sequencing approach, was more scalable sequencing techniques that excluded the gel electrophoresis phase in the sequencing of DNA (Brenner *et al*., 2000). Massively parallel sequencing works by combining non-gel-based sequencing with *in vitro* cloning of DNA templates on microbeads. In this method, it becomes possible to concurrently carry out multiple runs of a ligation-based DNA sequencing strategy on a million microbeads, each carrying clonal copies of a single template in order to generate millions of signature sequences (Brenner *et al*., 2000). This approach was first used to examine changes in gene expression, as an alternative to microarrays. See Figures 4 and 5 below for a detailed representation of the process.

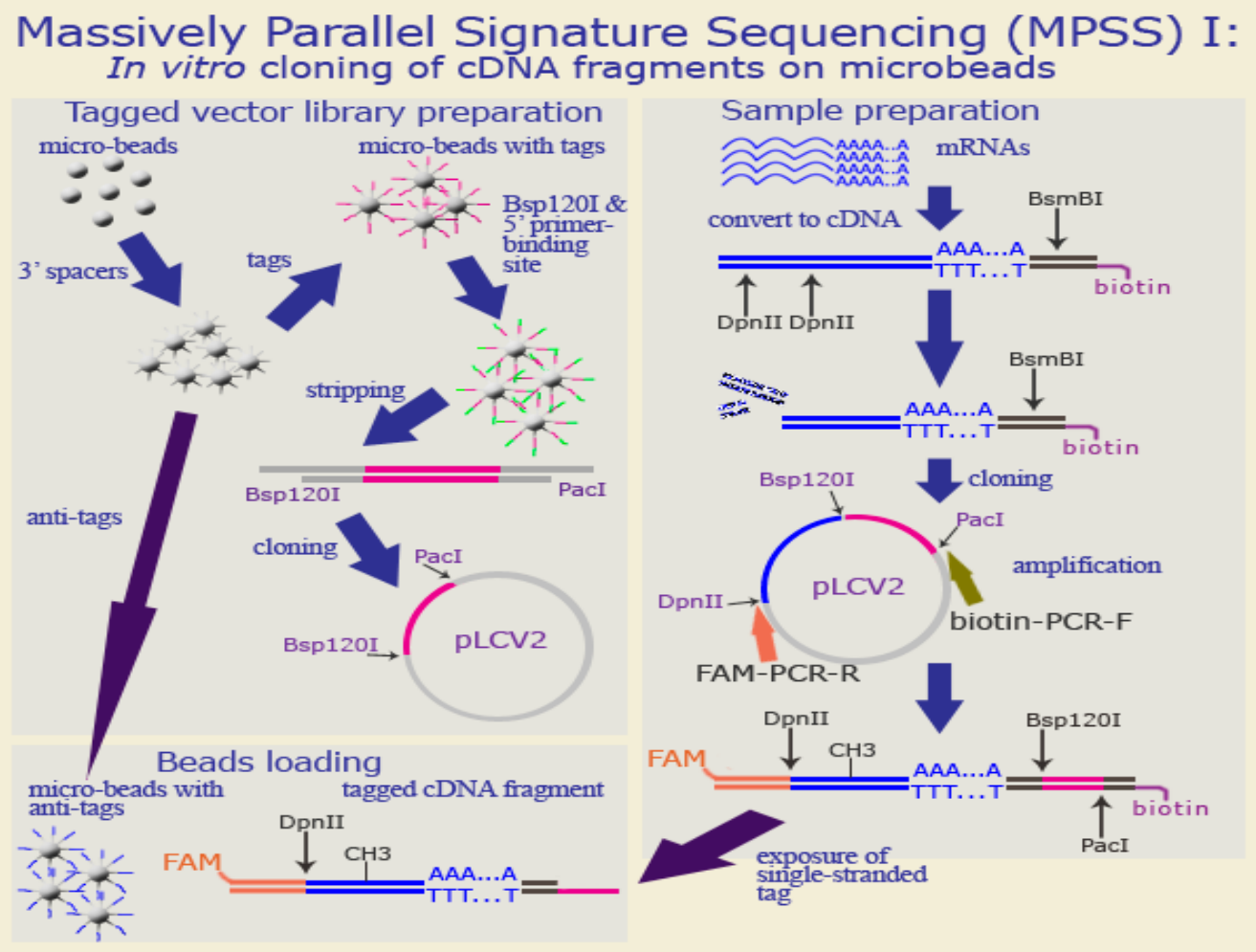

Figure 4: In vitro cloning of cDNA fragments on microbeads

Source: http://www.ncbi.nlm.nih.gov/projects/genome/probe/doc/TechMPSS.shtml

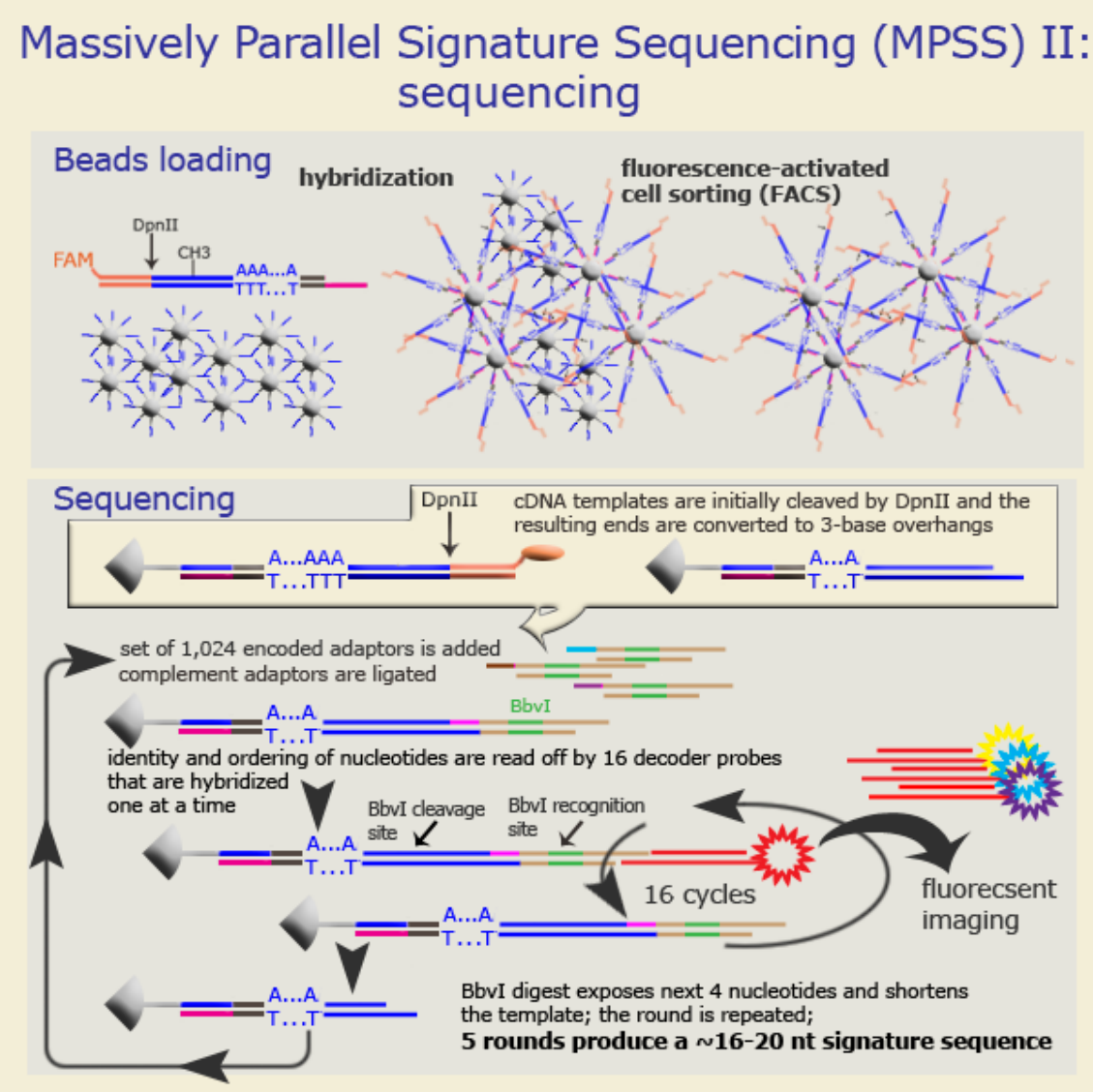

Figure 5: Sequencing process

Source: http://www.ncbi.nlm.nih.gov/projects/genome/probe/doc/TechMPSS.shtml

1.2.3 Genome sequencing by hybridization

This strategy developed in 1988, is based on the principle that the differential hybridization of labelled nucleic acid segments to an array of oligonucleotides (a microarray), can allow for the unique identification of primary DNA sequences and variant positions (Shendure and Ji, 2008). Affymetrix and Perlegen utilised this strategy in their discovery of Single Nucleotide Polymorphisms (SNPs) in the human (Patil *et al*., 2001), mouse (Frazer *et al*., 2007), and yeast (Gresham

et al., 2006) genomes. Even though microarrays are considered useful and cost effective for transcriptomics re-sequencing, sequence verification and analysis of SNPs (Gresham, Dunham and Botstein, 2008), they have many limitations that have made them highly unpopular for genome sequencing. Some of these limitations are: (i) difficulty in interrogating sequences with high repetitive DNA contents; (ii) difficulty in performing *de novo* sequencing (iii) the potential for false positives in the absence of careful data analysis as well as the uncertainty on how to get the equivalent of redundant coverage which is obtainable with Sanger and cyclic array sequencing methods; (iv) can only sequence what is already known. Thus, microarray sequencing will not be considered further.

1.2.4 Single molecule sequencing

Single molecule sequencing handles DNA sequencing by directly sequencing individual DNA molecules rather than involving amplification prior to sequencing (Henson, Tischler and Ning, 2012). The ability to sequence directly from biological samples makes this strategy well suited for diagnostics and clinical applications (Thompson and Milos, 2011). The Helicos true single-molecule-sequencing technology (tSMS); PacBio SMRT and ONT, as well as the recent return of Helicos technology in the form of SeqII (http://seqII.com), are the current technologies implementing this strategy. Refer to page for further details on the technologies implementing this strategy.

1.3 Sequencing and assembling the human genome

History reports the sequencing and assembling of the human genome more than a decade ago, as a product of research by two study groups - the International Human Genome Consortium (IHGSC) and Celera Genomics, leading to the release of the first drafts (Venter *et al*., 2001; Lander *et al*., 2001) and later complete (Collins *et al*., 2004) sequence of the human genome by both study groups in 2001 and 2003 respectively.

The draft genome was a sequence dataset covering about 90% of the human genome at an error rate of one in 1,000 base pairs, contained many gaps (>150,000 gaps), with only 28% of the genome at “finished” status: that is, having less or no gaps and more accuracy with an error rate of less than one error in every 10,000 base pairs (http://www.genome.gov/11006943).

1.3.1 The Human genome project’s (HGP) strategy

Efforts to sequence the human genome began in 1990 as a public funded project with the aim of determining the DNA sequence of the euchromatic human genome within a 15-year period (Chial, 2008). The HGP implemented the hierarchical shotgun sequencing (HS) strategy. The collaborative effort of researchers from about 18 different countries formed the International Human Genome Sequencing Consortium (IHGSC) – a consortium that was expected to make available freely to the public all human genome sequence information within 24 hours of assembly (Chial, 2008). Two goals were set by the IHGSC – the first was the building of genetic and physical maps of the human and mouse genomes (Donis-Keller *et al*., 1987; Gyapay *et al*., 1994; Hudson *et al*., 1995; Dietrich *et al*., 1996; Nusbaum *et al*., 1999) and the second was sequencing smaller genomes like worm (Oliver *et al*., 1992). The success recorded in these initial tests, led to the full-scale sequencing of the human genome (Chial, 2008). Figure 6 shows a diagrammatic representation of the HGP process.

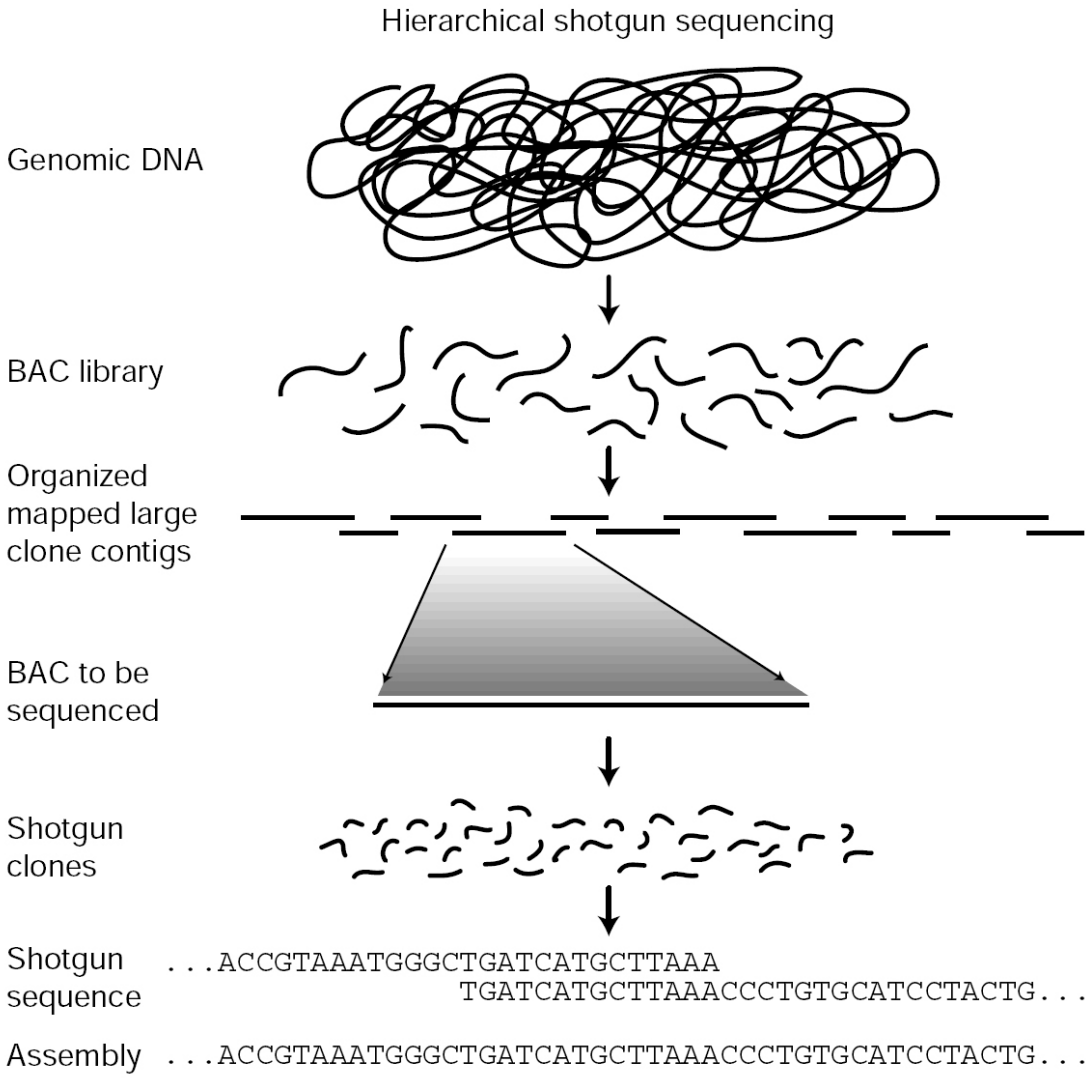

Figure 6: Idealized representation of the hierarchical shotgun sequencing strategy.
Source: Lander *et al*., 2001

In the figure above, the genomic DNA is represented by a long black line that looks like a tangled string. The genomic DNA is cut, and the pieces are cloned into bacterial artificial chromosome (BAC) vectors to create a BAC library. The BAC library is represented by short, disordered, squiggly black line segments. Next, the clones are organized and mapped into overlapping large clone contigs. One of the BAC clones is randomly chosen for sequencing. It is fragmented into small pieces, which are sub-cloned into vectors to generate

shotgun clones. These clones are then sequenced. Overlapping portions of the shotgun sequences are assembled to determine the genomic sequence.

1.3.2 The Whole genome shotgun assembly (WGSA)

On September 8, 1999, nine years after the commencement of the publicly funded HGP project, Celera Genomics; a privately-owned biotechnology company joined the race to sequence the human genome. Led by Craig Venter, the group used two independent approaches (whole genome assembly and regional chromosome assembly) to determine the sequence of the human genome (Venter *et al*., 2001). Figure 8 shows the computational processes in the architecture of Celera's two pronged-assembly strategy pipeline as ovals, with labels describing the function of the processes. The labels seen on arcs between the ovals show the type of objects generated from each computational process. Celera utilized its proprietary shotgun technique (Whole Genome Shotgun) to generate the first set of sequence data from the DNA of five individuals. This data consisted of over 27 million sequences, each with an average length of 543 base pairs (Venter *et al*., 2001). The second data set utilized came from the HGP and was derived from the published BAC contigs. Using WGSA, Celera randomly fragmented the HGP DNA sequence into 550-base-pair sequence reads representing a total of 16.05 million sequence reads (Chial, 2008). The final assembly of the human genome was done using whole genome assembly and regional chromosome assembly methods (Venter *et al*., 2001).

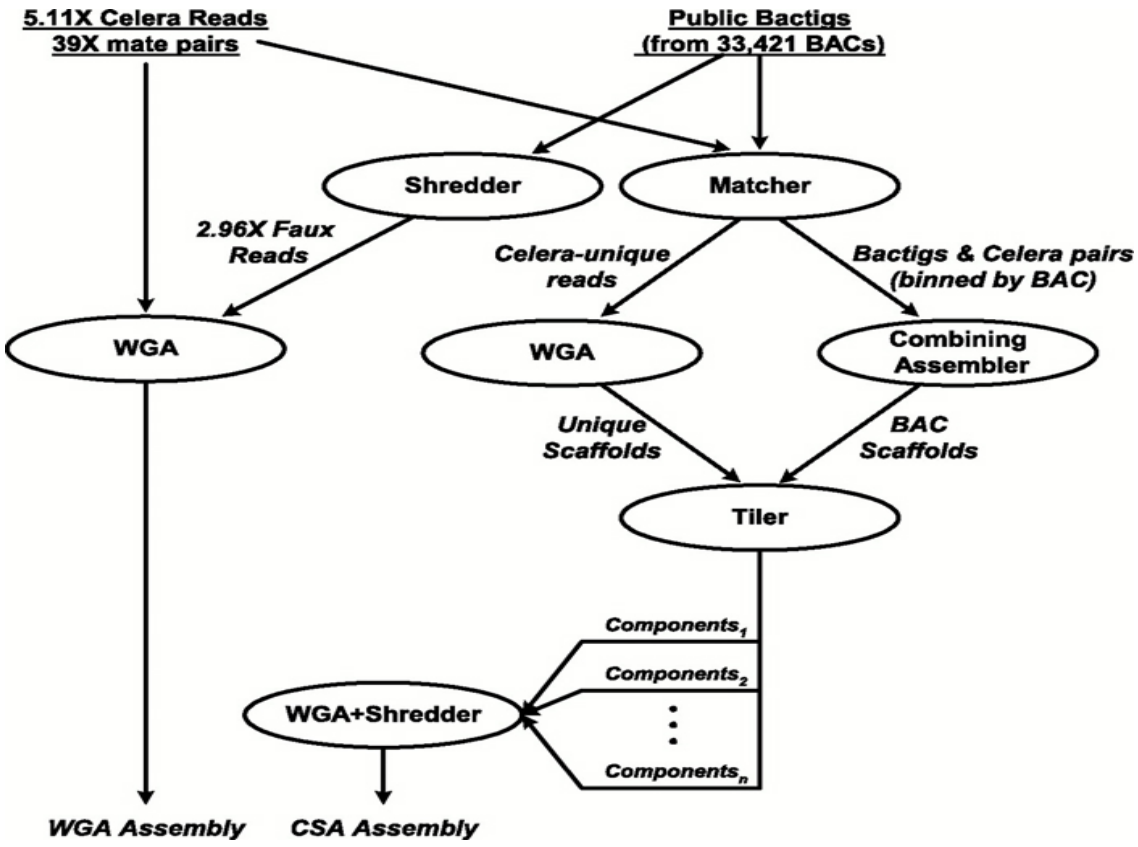

Figure 7: Architecture of Celera's two-pronged strategy

Each oval indicates a computation process with its function indicated by the label. The labels on arrows between the ovals describes the nature of the objects produced by the process.

Source: Venter *et al*., 2001

Although the two approaches (HGP and WGSA) discussed above were not completely independent of each other, it was possible to perform a direct comparison of the generated data. The regional chromosome assembly technique proved to have generated a slightly more consistent assembly than the whole-genome assembly technique. Despite this, the use of both complementary approaches afforded Celera the chance to generate data that was closely matched with data from the IHGSC (Chial, 2008), as well as

reducing the time it took to finish the sequence and assembly of the human genome (to less than a year).

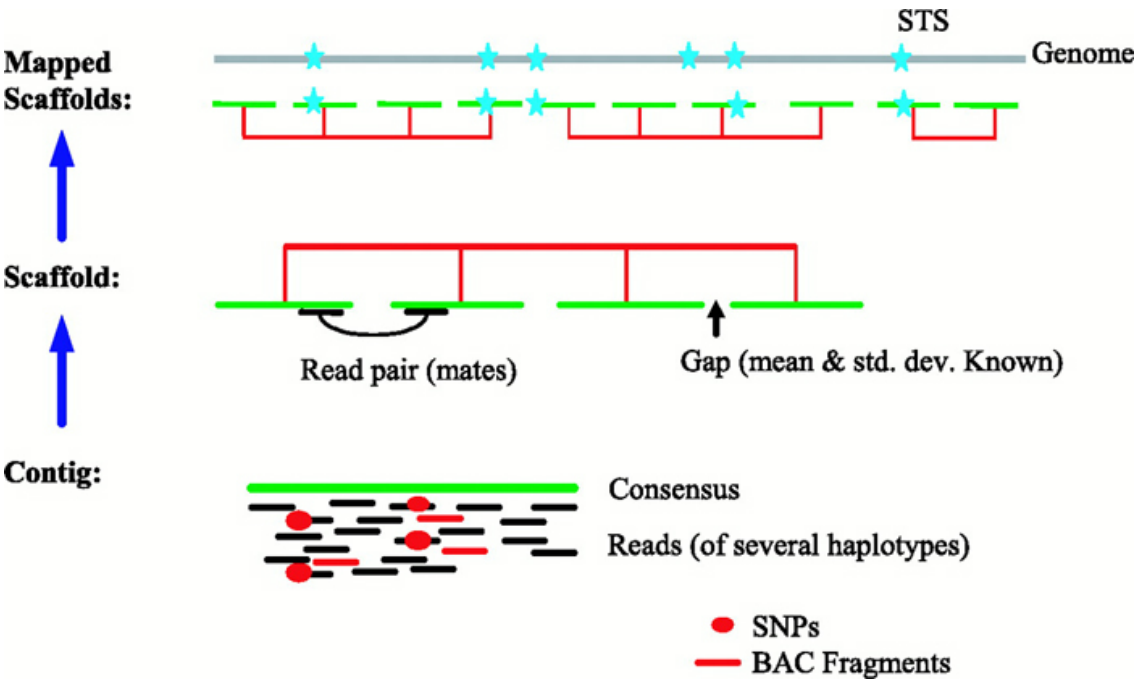

Figure 8: Anatomy of the whole-genome assembly

Red lines indicate the BAC fragments and the black lines show the reads from the five individuals are combined to generate a contig and a consensus sequence indicated by the green line. Scaffolds are made by joining contigs (in red lines) by mate pairing. Size is defined by the presence of a gap between consecutive contigs. Using STS data (as in seen in blue stars), scaffolds are then mapped to the genome (shown as grey lines).

Source: Venter *et al*., 2001

1.4 Assembly strategies and algorithms

The coming of advanced sequencing technologies have led to reduced cost, increased speed and larger volumes of data being generated from whole genome sequencing. *De novo* assembly of these sequence data is a first step for any form of genome analysis. However, existing assembly algorithms face the challenge of assembling shotgun sequence data generated by short read (between 50 – 150 bp in length) NGS technologies in the absence of accurate long-range linking information (Salzberg *et al*., 2012). Scientists however, have developed various sequencing strategies and accompanying algorithms in order to address the problems that come with shotgun short read sequence data. These strategies have been used in the resequencing of many genomes, including that of human (Venter *et al*., 2001). Fundamentally, the presence of repeats poses a major problem for long-range assembly. Three different strategies for shotgun sequencing have been utilized in the sequencing of DNA so far and they are:

1.4.1 Whole-genome shotgun sequencing: This strategy employs the shotgun technique across the entire genome. Refer to section 1.3.2 on page 18 for further details. This strategy does bring the benefit of preparing the shotgun data in just a single step across the whole genome but requires substantial computational resource to assemble large genomes. Celera Genomics utilized this process for the assembly of the human genome (Venter *et al*., 2001).

1.4.2 The Hierarchical shotgun (HS) approach: In this approach, the need for a high-resolution genetic map prior to the whole-genome assembly and low-resolution physical map is a limitation (Lander *et al*., 2001). The strategy offers a chance to generate more accurate genome sequences as it depends on the library hierarchy information for generating target sequences. In comparison to the whole-genome approach used by Celera Genomics, this is a more time consuming and costly approach. However, the strategy was employed by the IHGSC for the HGP (Lander *et al*., 2001).

1.4.3 The Hybrid approach: Pioneered at the Baylor College of Medicine, this strategy utilized both the hierarchical and whole-genome shotgun

strategies, excluding the use of genetic and physical mapping information. With one set of shotgun reads from the whole-genome shotgun, and the other set of reads from the sequencing of individual BACs, the strategy enabled the determination of a minimum tiling path covering the entire genome. A selection of BACs from the tiling path was subjected to shotgun sequencing at low coverage providing sufficient data to determine the complete sequence of the genome. The sequencing of the Drosophila genome (Adams *et al*., 2000; Myers *et al*., 2000; Hoskins *et al*., 2000) and the brown Norway rat genome (Gibbs *et al*., 2004) were carried out using this strategy.

1.5 Assembly algorithms

The construction of a contiguous genome sequence can be very computationally demanding especially when dealing with very large genomes, short reads from NGS platforms, and/or genomes rich in repetitive sequences. It becomes very important to find ways to perform accurate assembly from these kinds of situations. Today, three fundamental approaches have been utilized in most assembling programmes. They are greedy algorithms, overlap, layout and consensus algorithms (OLC), and de-Bruijn graph-based (DBG) algorithms.

For the benefit of this thesis, we will focus only on OLC algorithms as these are implemented in the assembly programmes used throughout this thesis.

1.5.1 OLC algorithms: This class of algorithm fundamentally uses a three-step approach in determining the contiguous sequence from a set of reads. These steps come in the order of overlap (O), layout (L), and Consensus (C). Developed by Staden in 1980 (Staden, 1980), the algorithm works on an intuitive basis and has over the years been improved by different scientists in order to enable it handle the challenges that come with assembling larger genomes and the frequent presence of repeats in these genomes.

1.5.1.1 Overlap (O): This first step in the OLC algorithm compares the sequence of each read with the rest of the other reads, both in the forward and

reverse complement directions (Staden, 1980). This leads to a very time intensive process, especially when very large numbers of reads are considered. In finding the overlap between sequences of reads, there is the possibility of a true overlap or a repeat overlap. A true overlap as seen in Figure 9A below occurs when sequences from two reads (S and T) truly overlap one locus in the genome, whereas, a repeat overlap occurs as a result of a repeat at either end of the two reads (S and T as seen by the orange line in Figure 9B below).

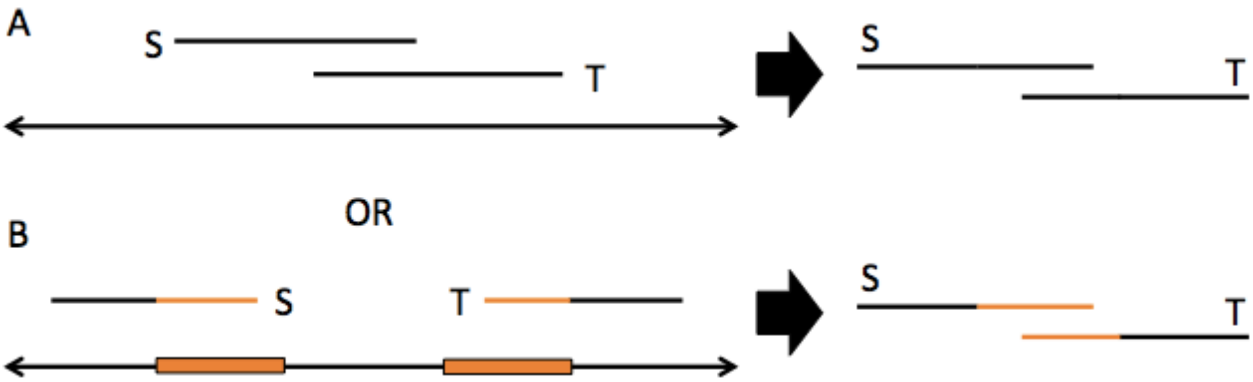

Figure 9: Overlap between reads S and T

The arrows indicate the genome, and orange segments showing the repeats in the genome. A shows true overlap and B shows a repeat induced overlap.

Source: http://gcat.davidson.edu/phast/olc.html. Accessed 02/09/2014

In building the assembly graph, consideration is given to both forms of overlap as no clear distinction can be made between the two. Thus, nodes on the graph (see Figure 10) depict the actual reads while the edges of the assembly graph depict the overlap between the reads. The final determination of the genome sequence is achieved by searching for the path that traverses the graph by visiting each node just once (Staden, 1980).

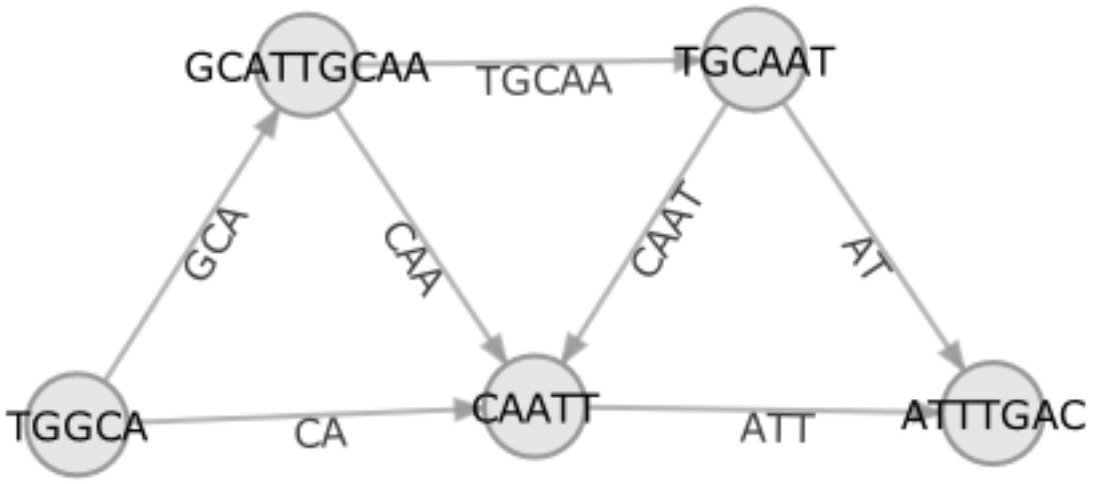

Figure 10: Assembly graph of an OLC algorithm

In this assembly graph from an OLC, nodes represent complete reads while edges link reads that overlap. Source: http://gcat.davidson.edu/phast/olc.html. Accessed 02/09/2014

1.5.1.2 Layout (L): Due to the computational challenge involved in finding the Hamiltonian path within a large OLC assembly graph, the layout stage attempts to decrease the size of the graph by compressing segments of the assembly graph into contigs. These contigs could be a subgraph or a group of nodes with several links among each other as they all overlap with each other and point to the same sequence (see label A and B in Figure 11). Within the contig subgraph, just the outer beginning and end nodes link to nodes external to the graph because the compression of nodes is only done until a fork is reached (see label C in Figure 11).

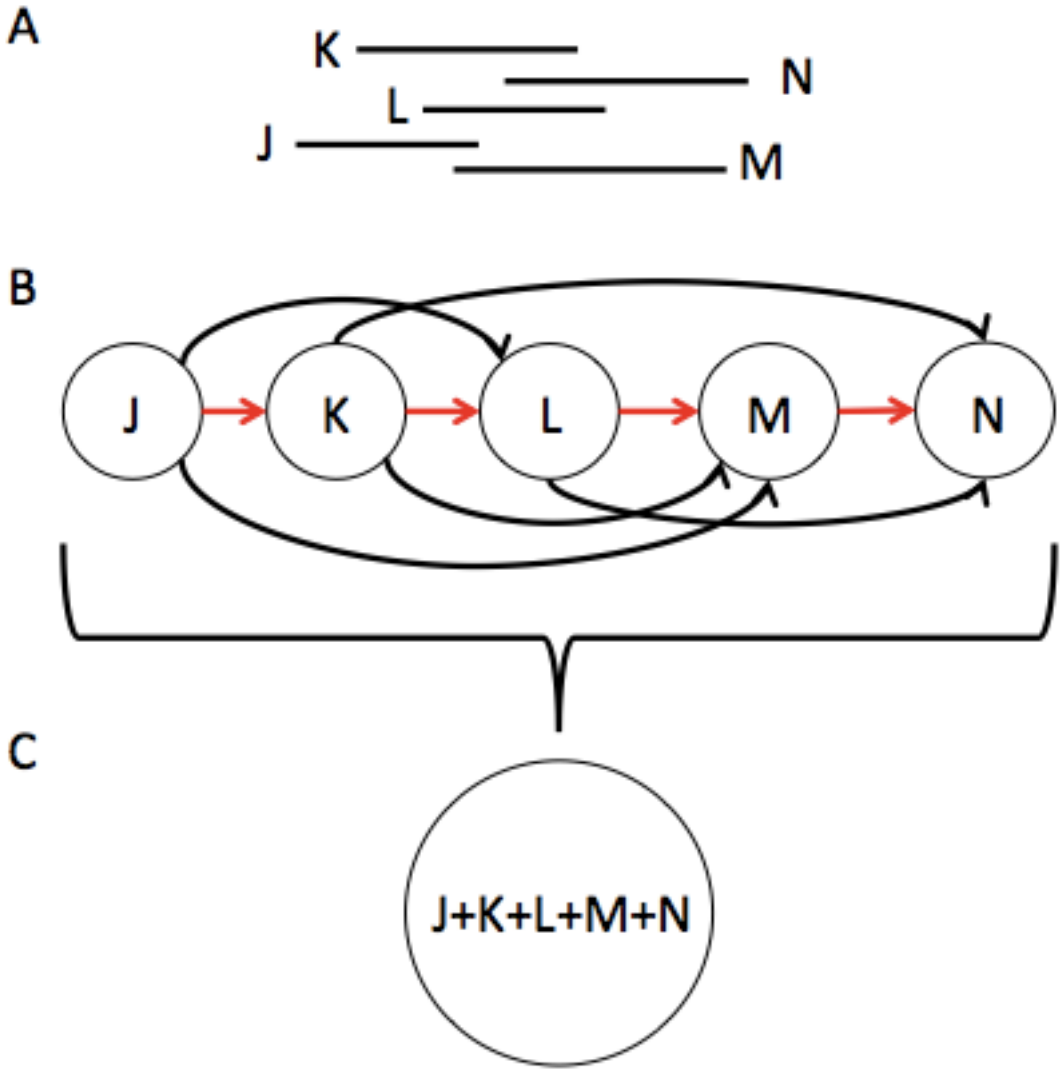

Figure 11: The generation of a unique contig during layout.

In 3A, we see the layout of the reads J, K, L, M, and N. In 3B, the assembly graph is shown. Here, the path shown in red arrows, visits every read exactly once through the graph. The compression of the reads J, K, L, M, and N, generates the unique contig J+K+L+M+N which gives us a simplified graph.
Source: http://gcat.davidson.edu/phast/olc.html. Accessed 02/09/2014

Using overlap information, scaffolds are built from unique contigs. Also, with the use of repetitive contigs, the OLC algorithm does attempt to fill gaps within the scaffolds (Jorde, 2008).

1.5.1.3 Consensus: This last stage involves inferring the final sequence from the assembly process. Ideally, just a single scaffold (represented by a single node) is expected to suffice from the collapsing of all previous nodes. The consensus of all the reads making a scaffold is computed by the OLC algorithm by looking at the left most read of each scaffold (Staden, 1980). The presence of insufficient mate-pair information or repeat contig information at the consensus phase can lead to gaps in the genome, thus generating a fragmented genome made up of many scaffolds that cannot be joined due to the presence of gaps.

The following assemblers were built on the OLC algorithm: Celera Assembler (Myers *et al*., 2000), CAP3 (Huang and Madan, 1999), PCAP (Huang and Yang, 2005), Phrap (de la Bastide and McCombie, 2007), Phusion (Mullikin, 2002), ARACHNE (Batzoglou *et al*., 2002), and Newbler (Margulies *et al*., 2005).

Having discussed the assembly algorithms implemented in the programs used in this thesis, we move on to look at the composition of the human genome in relation to repeat content.

1.6 Repeat content of the human genome

The early days of molecular biology witnessed a startling observation: that the size of a genome did not correlate with organismal complexity (Li, 1997; Gregory and Hebert, 1999). The recognition of the presence of large amounts of repetitive sequences in genomes, exceeding those associated with protein coding genes offered an explanation for this earlier observation (known as the C-value paradox) (Thomas, 1971). The C-value refers to the amount of DNA in the haploid genome of an organism (www.sciencedirect.com/topics/neuroscience/c-value).

In this section, I consider the sequence content of the human genome with particular attention to repeats, which constitute over 50% of our DNA (Lander *et al*., 2001; Treangen and Salzberg, 2012).

The results reported in the draft human genome sequence publication (Lander *et al*., 2001) opened a new perspective on the composition and organization of

the human genome. Even though repeats have often been referred to as “junk DNA” and considered uninteresting, it has however subsequently become clear how much biological information they hold (Lander *et al*., 2001). Repeats give clues to evolutionary events and forces by acting as passive markers – providing assays for studying selection and mutation processes. Repeats can also act as active agents in the reshaping of genomes by initiating ectopic rearrangements, contributing to overall GC content change, changing existing genes, making entirely new genes, shedding light on the structure of chromosomes and their dynamics, and enabling the generation of tools for carrying out medical and population studies (Lander *et al*., 2001).

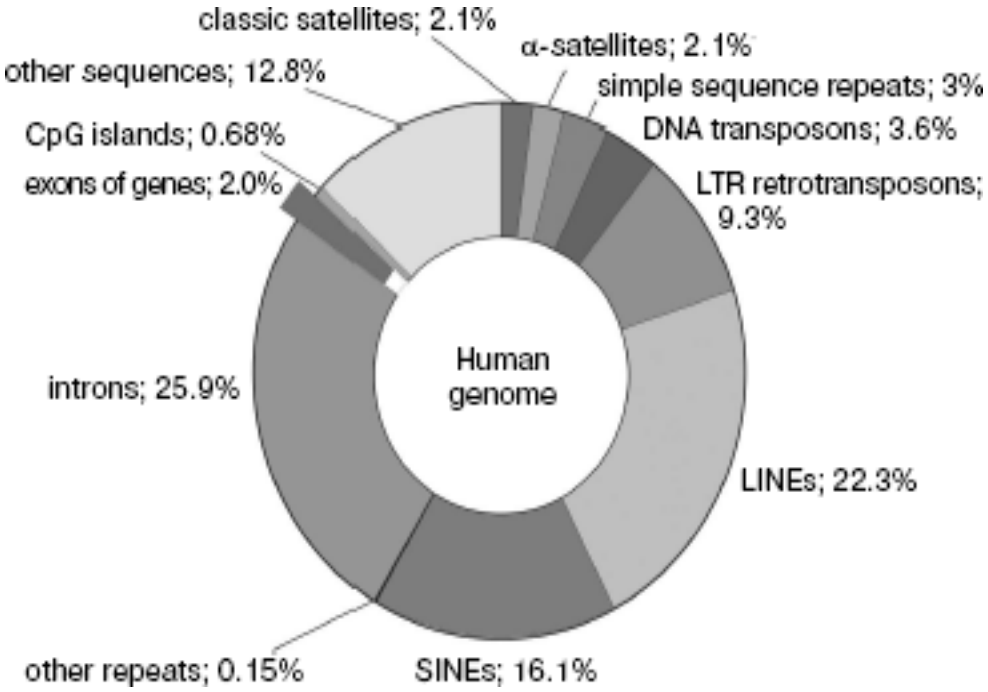

Figure 12: Composition of the Human Genome

Source: (Figure 21: Lander *et al*., 2001)

The human genome can be said to have two categories of sequences - coding and non-coding sequences.

1.6.1 Coding sequences (protein-coding genes)

DNA sequences that are transcribed into messenger RNA (mRNA) and then translated into proteins during the human life cycle are called coding DNAs. These coding DNA sequences account for <2% of the human genome (Lander *et al*., 2001). About 22,500 protein-coding genes have been identified in the human genome, with the possibility of a larger number of distinct protein products being generated from alternative splicing events (Lander *et al*., 2001). Despite the small proportion of these genes in human DNA, they however, represent the most extensively studied and well-understood component of the human genome, as they are responsible for the production of all human proteins.

1.6.2 Non-coding sequences (ncDNA)

ncDNA can be defined as all DNA sequences that are never represented in the amino acid sequences of expressed proteins. 98% of the human genome is composed of these sequences (Lander *et al*., 2001). ncDNA can be divided into the following groups: (a) genes for non-coding RNA, (b) pseudogenes, (c) introns, (d) untranslated regions of mRNA, (e) repetitive DNA sequences, (f) regulatory DNA sequences, (g) sequences associated with mobile genetic elements (though some parts of mobile genetic elements are / were coding).
In the following section, the focus will be on repetitive DNA sequences. Repetitive DNA sequences are stretches of DNA sequences are repeated many times throughout the genome, occurring either as tandem arrays of identical or similar repeats or individual copies interspersed through the genome.

1.6.2.1 Tandem repeats

These are repeat sequences of two or more repeated DNA base pairs, arranged such that on the chromosome, the repeats lie adjacent to each other. These kind of repeats are more likely seen in non-coding DNA, but can occur within genes and some are associated with disease states (such as PolyQ tracts in Huntingtons disease) (Orr and Zoghbi, 2007). Tandem repeats can be classified based on their repeat unit size into microsatellites (that is, short repeats of about 2 -10bp in length), minisatellites (longer repeat units than

microsatellites, usually 10bp - 50bp), and satellite DNA (with repeat units of hundreds of base pairs that can extend over hundreds of kilobases of DNA in length) (Ahmed and Liang, 2012). Over the years, minisatellites and microsatellites have been used as markers in population genetic studies (Armour *et al*., 1996), DNA fingerprinting (Jeffreys, Wilson and Thein, 1985; Tamaki *et al*., 1995; Spurr *et al*., 1994; Jeffreys and Pena, 1993), as well as contributing to intraspecies genetic diversity (Ahmed and Liang, 2012).

1.6.2.2 Transposon Derived Repeats

These repeat sequences, usually occurring as interspersed repeats, comprise about 45% of the human genome (Lander *et al*., 2001). Transposable elements are the source of most human repeat sequences (Smit, Hubley and Green, 2013; Prak and Kazazian, 2000). These interspersed repeats can be classified into four groups, as seen in Figure 13. These groups are Long Interspersed Nuclear Elements (LINES), Short Interspersed Nuclear Elements (SINES), Long Terminal Repeats (LTR) retrotransposons, and DNA transposons. While DNA transposons transpose directly via DNA, LTR retrotransposons, SINES, and LINES however, transpose via an RNA intermediate (Lander *et al*., 2001). Either autonomously or non-autonomously, some of these repeats can currently replicate and integrate into new genomic locations on the chromosomes.

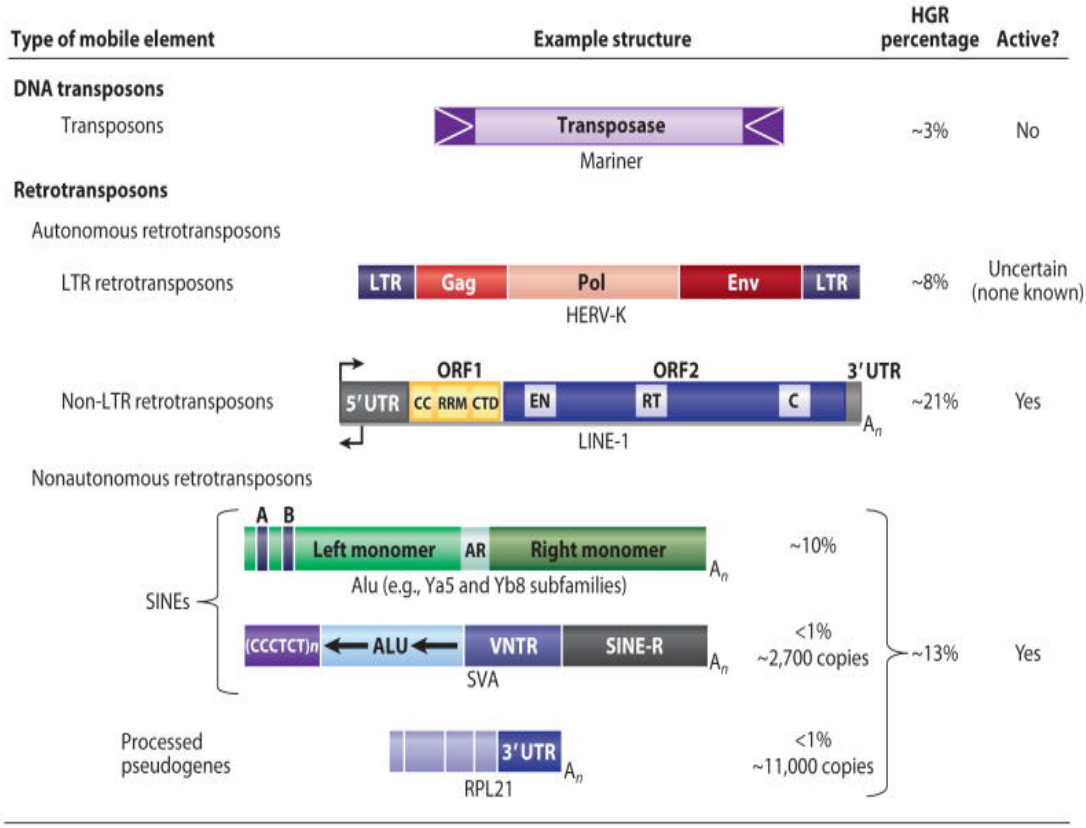

Figure 13: Classes of Interspersed repeat in the human genome.

The content distribution of interspersed repeats in the human genome, showing the type of repeat, the structure of representative elements, the percentage of each element in the human genome and the activity status of each element.

Source: (Figure 1: Beck *et al*., 2011)

1.6.2.2.1 LINES

LINEs form a group of genetic elements with a large presence in some eukaryotic genomes (Singer, 1982). In the human genome, LINEs contribute 21% out of the total 45% repeat content in the human genome. LINEs are the only known autonomous non-LTR elements in the human genome. Full length elements possess two open reading frames (ORFs) (see Figure 13) that encode the functions (reverse transcriptase, endonuclease activity, and nucleic acid binding properties required to form a ribonucleoprotein particle) needed for their retrotransposition (Yadav *et al*., 2009). LINEs comprise three sub-families: LINE-1, LINE-2, and LINE-3. LINE-1 elements make up 17% of the human genome and are the second most numerous repeat sequence (Cordaux and

Batzer, 2009). LINE-2 and LINE-3 though now inactive (Lander *et al*., 2001), make up 3% and <1% of the human genome respectively.

A LINE-1 element's structure starts with an untranslated region (UTR) containing an RNA polymerase II promoter, two non-overlapping ORFs (ORF1 and ORF2), and terminates with another UTR (Doucet *et al*., 2010). An RNA binding protein is encoded by ORF1, while a protein carrying endonuclease and reverse transcriptase activities is encoded by ORF2. The DNA copy of the protein encoding RNA is integrated into the genome at a new site due to the reverse transcriptase's higher specificity for the LINE RNA as compared with other RNAs (Ohshima and Okada, 2005). In the 3' UTR of the LINE-1 sequence, there is a polyadenlylation signal (AATAAA) and a poly-A tail, whereas, the 5' UTR carries a promoter sequence (Deininger and Batzer, 2002). The manner in which LINES replicate (copy and paste rather than cut and paste) is inherently replicative contributing to human genome expansion. Thus, today the human genome contains about 500,000 LINES, with ~7000 of them being full length (>6kb) and a small subset still being capable of retrotransposition (Griffiths, 2007; Rangwala, Zhang and Kazazian., 2009).

1.6.2.2.2 SINES

These are short DNA sequences usually between 30 - 500bp in length (Griffiths, 2008), containing an internal Polymerase III promoter, and encoding no proteins (Lander *et al*., 2001). SINEs make up about 11% of the human genome with about 1,500,000 copies (Cordaux and Batzer, 2009). As opposed to the autonomous LINES, SINEs rely entirely on the LINE machinery for transposition. The sharing of the 3' end of a SINE with a resident LINE element guarantees the survival of most SINEs (Okada *et al*., 1997). However, the only active SINE in the human genome: the Alu element, does not share its 3' end with a LINE element. There are three distinct monophyletic families of SINEs in the human genome. They include: the active Alu, the inactive MIR, and Ther2/MIR3 (Lander *et al*., 2001). The survival of SINEs over evolution has been due to vertical transmission within host genomes.

Although, earlier research considered SINEs as junk DNA, current research has shown that in special cases, the incorporation of both LINEs and SINEs into genes has contributed to the evolution of new functionality (Santangelo *et al*., 2007).

1.6.2.2.3 LTR retrotransposons

These retrotransposons feature long direct sequence repeats flanking the internal coding region, containing genes encoding both structural and enzymatic proteins (Havecker *et al*., 2004). Reverse transcription occurs inside a virus-like particle that is formed when the *gag* gene is expressed as it encodes the structural proteins. The *pol* gene however, does encode several enzymatic functions such as a protease activity for cleaving the Pol polyprotein, a reverse transcriptase for copying the retrotransposon's RNA into a cDNA, and finally, integrating the cDNA into the genome using an integrase activity (Havecker *et al*., 2004).

In the mammalian genome, only the vertebrate-specific endogenous retroviruses (ERVs) seem to be active, even though, a number of LTR retrotransposons are intact, and potentially active *in silico* (Lander *et al*., 2001). LTRs ensure their evolutionary survival using both vertical and horizontal transmission strategies within their host genome. However, there is currently no evidence of horizontal transfer of LTR retrotransposons in humans.

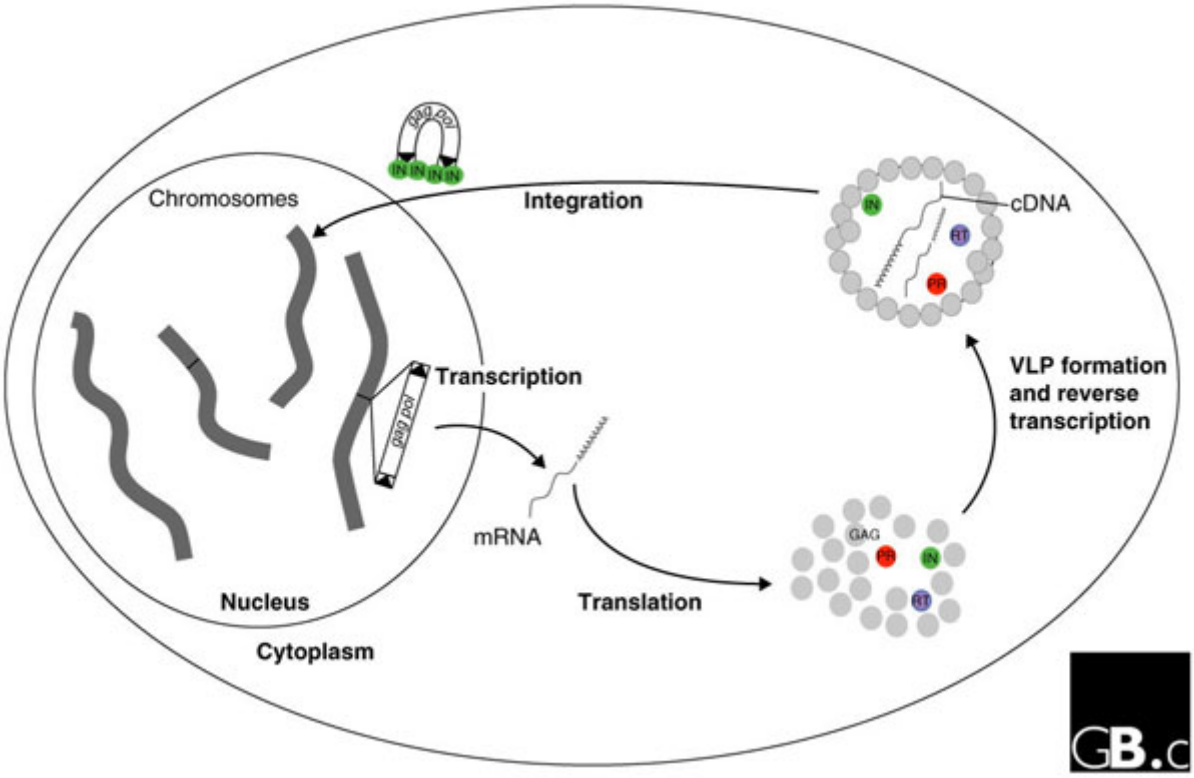

Figure 14: The cycle of LTR retrotransposons

IN indicates integrase; PR indicates protease; RT indicates reverse transcriptase; VLP indicates virus-like particle, and Black triangles indicate the LTRs.

Source: (Figure 23: Lander *et al*., 2001)

1.6.2.2.4 DNA transposons

In contrast to LINEs and SINEs, DNA transposons replicate in the genome by a cut and paste mechanism (Lander *et al*., 2001). That is, their replication, involves a complete change in location of their DNA. DNA transposons generally comprise terminal inverted repeats flanking a sequence that encodes a transposase protein, which binds near the inverted repeats. At least seven major classes of DNA transposons with subdivisions into families with independent origins, exist in the human genome (Smit, Hubley and Green, 2003). The process of transposition by DNA transposons is *cis*-preference deficient. That is, the DNA transposase is unable to differentiate between active and inactive elements when it returns to the nucleus, having been translated in

the cytoplasm. Thus, the accumulation of inactive copies in the genome reduces the efficiency of transposition. This restricts the growth of any DNA transposon family, leading to their gradual extinction (Lander *et al*., 2001). Nonetheless, DNA transposons can survive by moving by horizontal transfer to unclonised genomes (Clark and Kidwell, 1997; Koga *et al*., 2000; Robertson and Lampe, 1995).

1.6.2.3 Pseudogenes

Processed pseudogenes are inactive retroposed copies of cellular genes often generated by retrotransposon reverse transcriptases, whereas, ordinary pseudogenes arise from mutation within the gene. Pseudogenes generally, are considered to have lost their protein coding ability or lack the ability to be expressed in cells (Vanin, 1984). Often, the generation of ordinary pseudogenes come from the accumulation of several mutations within a gene where the product of the gene is inconsequential to the survival of the organism (Poliseno *et al*., 2010).

1.6.2.4 Simple sequence repeats (SSRs)

These sequences are also called microsatellites. They are repeats comprised of direct repetitions of short k-mers like $(A)_n$, $(CA)_n$, or $(CGG)_n$. SSRs are generated by slippage during DNA replication (Kruglyak *et al*., 1998; Tóth *et al*., 2000). SSRs contribution to the human genome is about 3% with a 0.5% singular contribution coming from dinucleotide repeats (Lander *et al*., 2001). AC and AT are the most frequently occurring dinucleotide repeats with a 50% and 35% occurrence rate, respectively. However, AG dinucleotide repeats have a frequency of 15%, GC having the least frequency of 0.1% - being greatly under-represented (Lander *et al*., 2001).

The constant slippage by DNA polymerase during replication has influenced the degree of length polymorphism at SSRs in the human population, thus enhancing the use of SSRs in genetic studies (Lander *et al.,* 2001). Many human disease mapping studies have relied on genetic markers based on SSRs – specifically, $(CA)_n$ repeats (Dib *et al*., 1996; Broman *et al*., 1998).

1.6.2.5 Segmental duplications

These are blocks of sequences around 10 – 300kb in length that have been duplicated in one or more regions of the genome (Lander *et al*., 2001). High sequence identity (at least 90%) is required for two repeats to be considered segmental duplications (Lander *et al*., 2001). These kinds of repeats are recent evolutionary occurrences as observed by their high sequence identity and by the fact that they are not present in closely related species (Lander *et al*., 2001). Segmental duplications can either involve inter-chromosomal duplication whereby segments are duplicated among non-homologous chromosomes. On the other hand, segmental duplication occurring on particular chromosomes or chromosomal arms are referred to as intra-chromosomal duplications (Lander *et al*., 2001). Low copy number repeat sequences mediating recurrent chromosomal structural rearrangements linked with genetic diseases can be considered as inter-chromosomal duplications (Ji *et al*., 2000; Mazzarella and Schlessinger, 1998).

At peri-centromeric and sub-telomeric regions, segmental duplication frequencies are 3 - 4 folds above the genome average. For known genes and Ensembl genes, Zhang (Zhang *et al*., 2005) reported the proportion of segmental duplications having complete genes to be 3.4% and 10.4% respectively. This suggests that segmental duplication could be responsible for a significant amount of gene copy number change. The availability of the complete human genomic sequence now gives us a systematic way of exploring the nature of segmental duplications (Lander *et al*., 2001).

1.7 The Next Generation Sequencing Revolution (NGS)

The sequencing of large genomes has witnessed a significant improvement in terms of time and cost, since the advent of Next Generation Sequencing (NGS) technologies. However, it is widely considered that the quality of assemblies has also been affected by wider use of NGS (Henson et al., 2012). In 2001, the application of WGS to large and complex genomes contributed to the completion of the draft human genome sequence (Lander *et a*l., 2001). Using sequence data generated from NGS systems, the genomes of the giant panda (using illumina data only) (Li *et al*., 2009) and turkey (a combination of illumina and 454 data) (Dalloul *et al*., 2010) have been assembled *de novo,* and these approaches have also been used for various human genome re-sequencing projects (Wang *et al*., 2008; Schuster *et al*., 2010; Ju *et al*., 2011). However, with NGS technologies generating large data sets of short sequence reads and in some cases, data with high error rates, the determination of the sequence of genomes faces a significant challenge at the assembly stage (Henson, Tischler and Ning, 2012).

In Table 1, we see a list of the most common NGS technologies and sequencing platforms. In the table, we observe the striking differences between technologies in relation to read length, coverage bias, and their respective error profiles. Of significance is the high error profile of PacBio despite its high read length and throughput (~5 – 10 Gb per SMRT cell).

Table 1: Properties of currently available NGS technologies

Platform	Read length (bp)	Throughput/run	Approximate time/run	Instrument cost (US$)	Primary error mode	Basic error rates
HiSeq™ 2000	100	600 Gb	11 days	690,000	Substitution	~1-2% over 100 bp
SOLiD™ 4	75	100 Gb	12 days	475,000	A-T bias	0.06%
SOLiD™ 4hq	75	300 Gb	14 days	595,000	A-T bias	0.01%
SOLiD™ PI	75	77 Gb	8 days	349,000	A-T bias	0.01%
454GSFLX TitaniumXL+	700 mean <= 1000	700 Mb	23 h	500,000	Indel	0.5%
IonTorrent™ PGM™ 316	200	100 Mb	~2 h	50,000	Indel	1.2% over 150 bp
IonTorrent™ PGM™ 318	200	1 Gb	~2 h	50,000	Indel	1.2% over 150 bp
MiSeq®	150	>1 Gb	27 h	125,000	Substitution	~1-2% over 100 bp
454 GS Junior	400 mean	35 Mb	12 h	108,000	Indel	1.00%
Helicos Genetic Analyzer	35 median	1Gb per hour	2.5 – 8 days	1.35 million	Substitution Indel	1-3%
MinION	Up to 100kb	1 Gb per flow cell	5 days	500 - 900 per flow cell	Indel	15-40%
PacBio RS	2700 mean <= 5000	90 Mb per cell	<1 day (?)	695,000	GC deletion	13.00%
PacBio RS II	> 2000	500 Mb – 1 Gb per SMRT cell	~ 6hrs per SMRT cell	~750,000	GC deletion	13.00%
PacBio Sequel	> 2000	~5 – 10 Gb per SMRT cell	~ 6hrs per SMRT cell	350,000	GC deletion	13.00%

Source: Adapted from (Glenn, 2011), www.454.com, www.illumina.com, www.appliedbiosystems.com, www.pacificbiosciences.com, www.iontorrent.com

1.8 Single molecule sequencing technologies – reality and prospects

Here we briefly discuss the various technologies that have and / or currently implementing single molecule sequencing strategy.

1.8.1 Helicos tSMS

Based on the work of Braslavsky (Braslavsky *et al*., 2003), Helicos tSMS was the first sequencing technology to avoid the problems associated with de-phasing - where certain members of the group of templates being sequenced fail to incorporate a nucleotide at a given cycle, hence, lagging behind the others as seen with 454, Ilumina, and SOLiD. This is because, no amplification of the DNA molecule occurs before sequencing is carried out, thus enabling long-read sequencers offer simple and rapid library preparation. In the case of Helicos, this is just the addition of a poly-A tail and a fluorescent label. Though not much information is available as to the error rate of this technology, it suffices to say that sequencing by tSMS can lead to sensitivity problems (Turner *et al*., 2009). The presence of long homopolymers can cause errors in the sequencing. Unfortunately, this technology does not permit paired end sequencing (Turner *et al*., 2009). Helicos Biosciences filed for bankruptcy in 2012, having only sold a handful of instruments, but has recently been apparently resurrected in the form of SeqII, although details of the technology are scant at the moment.

1.8.2 Pacific Biosciences RS

This is a sequencing method in which DNA is sequenced without the need for an amplification step. Using a zero-mode waveguide (ZMWs) as proposed by Webb and Craighead (Levene *et al*., 2003) Pacific Biosciences, developed parallelized single molecule DNA sequencing by synthesis approach. The idea for using ZMWs was to completely remove the background noise introduced by the fluorescent nucleotides required for the sequencing process (McCarthy, 2010). The SMRT technology uses a polymerase enzyme, which is affixed at the bottom of a ZMW well. With a single DNA molecule acting as template, the polymerase reads through the template incorporating fluorescently labelled

nucleotides (McCarthy, 2010). The emitted signal from the ZMW is unique since each base carries a different fluorescent dye. The fluorescent signal is read by a detector, which calls the nucleotide based on the colour of the fluorescence (McCarthy, 2010). On addition of the base, the polymerase cleaves the fluorescent tag from the nucleotide. The first application of SMRT to generate data was in 2009 (Eid *et al*., 2009).

SMRT technology enables the monitoring of DNA synthesis as it occurs in real-time. It can sequence single molecules of DNA using a far lower consumption of reagents in comparison to other NGS technologies – thus making the process much more affordable apart from the initial instrument cost. The process utilizes a simple preparation procedure and is very fast compared to other NGS techniques. While other NGS methods involve complicated cyclical steps in which several minutes pass for each nucleotide addition to occur, SMRT technology, reads sequences by watching a single strand of DNA being replicated in real-time, at a rate of about five nucleotides per second.

The PacBio RS system is able to generate 90 megabases of data per SMRT® cell, having a mean read length of 1500 bp. PacBio's advance to the C3 chemistry generated reads with an average length of 8.5kb, with its longest read over 30kb in length (www.pacificbiosciences.com). As an SBS based technique, an error rate of ~13–15% in raw data exists. However, by means of computational processes or by reading the sequences more than once and deriving a circular consensus, the high error rate can be traded in for long read lengths (Henson, Tischler and Ning., 2012).

In 2013, PacBio advanced their sequencing technology to the RS II. With this advancement came a run-time flexibility from 0.5 – 6hrs per SMRT cell, higher throughput (500 Mb – 1 Gb per SMRT cell), longer read lengths (average read >20 Kb), a lower degree of bias, simultaneous epigenetic characterization and highest (99.99%) consensus accuracy (www.pacb.com/products-and-services/PacBio-systems/rsii). The RS II is applicable to whole genome sequencing of small genomes, targeted sequencing, complex population analysis, microbial genetics and RNA-seq of targeted transcripts.

The most recent advance in the PacBio's line of research was the development of the Sequel System. Comparing the Sequel to the RS II, PacBio now offers a high-throughput, cost-effective access to SMRT sequencing that generates about 7X more reads with one million ZMWs per SMRT cell (www.pacb.com/products-and-services/PacBio-systems/sequel). Due to Sequel's high-throughput and long read lengths (up to 8.5kb), SMRT has found application in the following genomics research areas: (i) *De novo* sequencing, (ii) Re-sequencing, (iii) Full isoform sequencing, and (iv) *In vitro* diagnostics (www.pacb.com/products-and-services/pacbio-systems/sequel).

1.8.3 Oxford Nanopore Technologies

ONT allows for real-time, scalable and direct DNA sequencing. And for the first time, ONT now provides direct RNA sequencing, as well as PCR or PCR-free cDNA sequencing (www.nanoporetech.com). With ONT, users select fragment size while the nanopore sequences the entire fragments; generating reads up to 1mb. From the release of the MinION device in 2014 by ONT (The MinION Access Program), through to the benchtop GridION and now the high throughput, high-sample number PromethION devices, we have witnessed a new dimension to genome sequencing. This means that sequencing of single DNA molecules can be carried out in the absence of PCR amplification or a chemically labelling process which requires the use of an optical instrument to identify the chemical label. Due to its small size and relatively low equipment cost, the MinION sequencer finds good applications use in sequencing small genomes but is limited by throughput. Figure 15 describes the protein pore (a laboratory-evolved *E.coli* CsgG mutant named R.4) utilized by ONT.

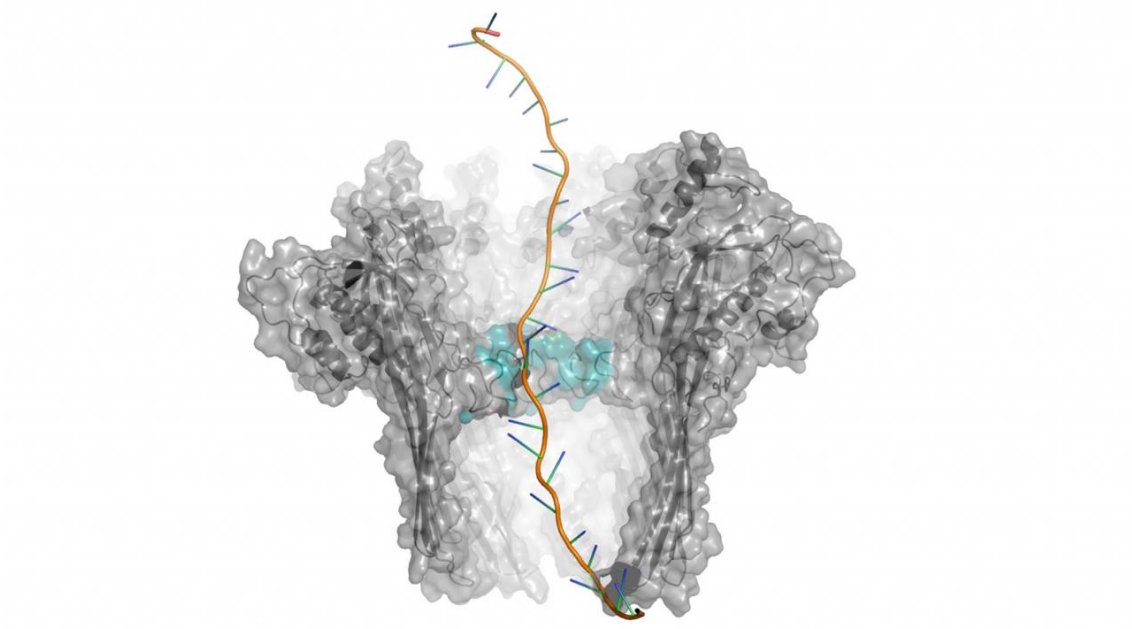

Figure 15: CsgG microscopic pore

The figure above describes nanopore's R9 – a membrane protein derived from *E.coli*, but present in many species

Source: http://www.sciencemag.org/news/2016/03/amid-patent-lawsuit-genetic-sequencing-upstart-unveils-new-technology

Here, there is the continuous application of ionic flow while the single stranded DNA passes through the pore. Using standard electrophysiological techniques, disruption to current flow is detected.

Improvements in library preparation techniques coupled with sequencing speed (450 bases/s) have led to increased throughput (Jain, Koren *et al.,* 2018). The platform has achieved a number of milestones in a rapid succession: successful bacterial genome sequencing using this platform was reported in 2014 (Quick, Quinlan and Loman, 2014). Also reported was the sequencing of the lambda phage genome (Mikheyev and Tin, 2014). ONT have sequenced human genomes, using many flowcells in parallel (being relatively insensitive to cost). And recently, sequenced data for the NA12878 individual using ONT was released (Jain, Koren *et al*, 2018).

1.9 Research goals

The work in this thesis focuses on the *de novo* reconstruction of hard-to-sequence/complex regions and/or missing elements (gaps) in the human reference genome assembly via a local assembly approach. Below, is a brief summary of the rationale for this research, and in chapters 3, 4, and 5, I present full details of the results achieved while addressing these challenges.

Genome sequencing (especially, of large genomes) has improved greatly in terms of speed and cost as a result of the NGS revolution (Henson, Tischler and Ning, 2012). However, despite the successful re-sequencing of the human genome (Wang *et al.,* 2008; Schuster *et al*., 2010; Ju *et al*., 2011) and the *de novo* assemblies of the Panda (Li *et al.,* 2010) and Turkey (Dalloul *et al*., 2010) genomes by purely NGS approaches, the quality of genome assemblies is still greatly affected by short read length and the errors generated by these technologies (Henson, Tischler and Ning, 2012). The presence of high copy number repeats in the genome DNA sequence, severely limits the ability of assembly software (hereafter, assemblers) to infer the relative positions of reads in the genome (Henson, Tischler and Ning, 2012). This effect is particularly acute for very short reads (<100bp) and highly repetitive genomes (Henson, Tischler and Ning, 2012). Thus, there is a need for assemblers that implement novel strategies for dealing with such difficulties in NGS generated data. In addition to repeats, the possibility of systematically incorrect base calling errors can lead to reads being more similar to the wrong location in the genome, reducing assembly contiguity (Schatz *et al*., 2010; Salzberg *et al*., 2012). Although NGS technologies now generate large datasets of short sequence reads with higher coverage to compensate for the reduced connectivity between reads, and to improve assembly, repetitive sequences that are longer than NGS read lengths still cannot be resolved solely by higher coverage, resulting in gaps in assemblies being biased towards repetitive regions (Schatz *et al*., 2010; Salzberg *et al*., 2012). The challenge of assembling genomic sequences from NGS data at particular regions means that single molecule sequencing (SMS) technologies (particularly, Pacific

Biosciences SMRT) with much longer read lengths become highly attractive. Recent studies have reported the use of long reads in improving and validating assemblies (Huddleston *et al*., 2014; English *et al*., 2012; McCoy *et al*., 2014) and at high coverage (>90-fold), high quality assemblies of moderate size genomes have been generated from single-molecule long reads using pre-assembly error correction methods (Huddleston *et al*., 2014; English *et al*., 2012; McCoy *et al*., 2014). Despite the high indel error rate associated with PacBio SMRT technology, these data have been shown to be effective in traversing common repeats during assembly (Huddleston *et al*., 2011).
In order to address these challenges, I have developed a pipeline for the local reconstruction of minisatellite alleles *de novo.* Also achieved, were pipelines for the characterisation of potentially novel L1 elements and the partial recovery of five (5) assembly gap elements from GRCh38.

2. Materials and Methods

This section describes the methods developed and other third-party tools utilized in this thesis to evaluate the effectiveness of Single Molecule Long-read technology in reconstructing coding and non-coding minisatellites DNA, discovery of novel L1 sequences and recovery of missing/gap elements in the human reference genome *de novo*. The tools used for this project are categorized into two groups namely; external components (tools developed by others) and internal (tools written for this project)

2.1 External components

2.1.1 Compute resources

The analysis pipeline in this study was built and implemented on the High-Performance Computing (HPC) cluster (ALICE2) hosted by IT Services of the University of Leicester. ALICE2 runs the Scientific Linux Operating System (OS) (version 6; a clone of RedHat Enterprise Linux) – giving both staff and research students access to a command line and graphical user interface (GUI) environment for executing tasks.

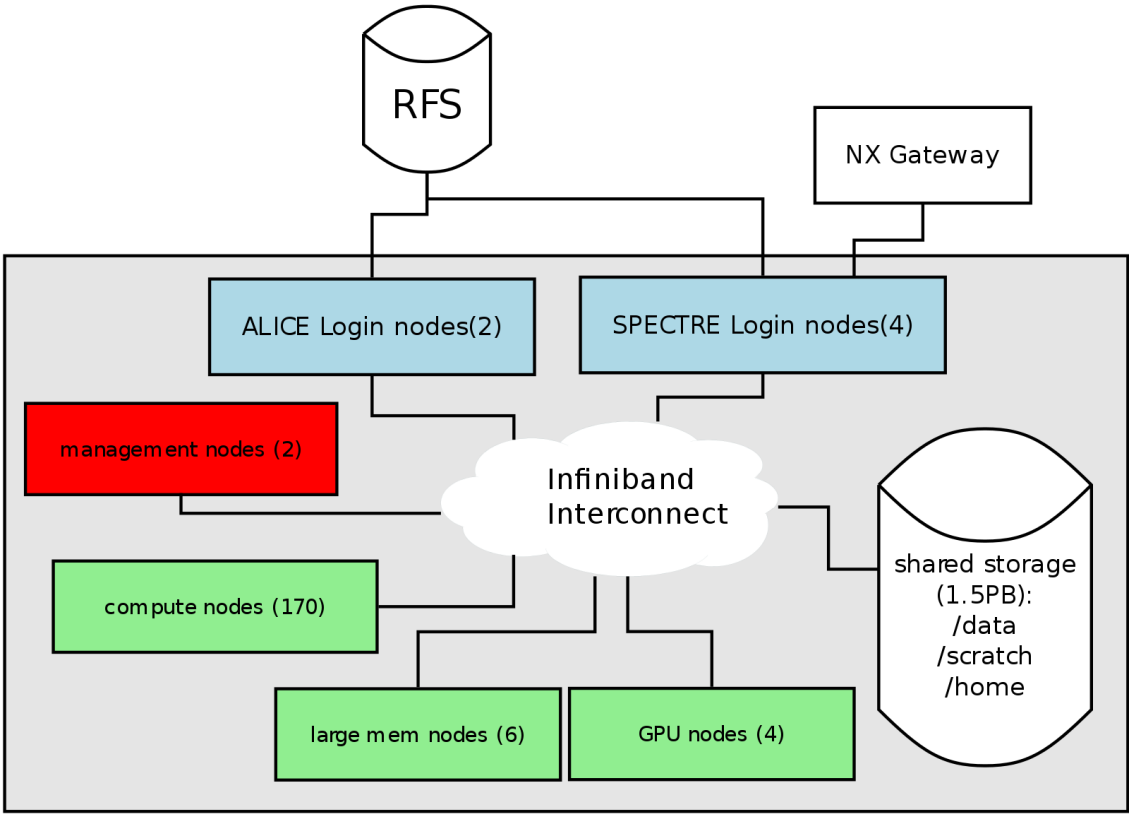

Figure 16: Architecture of ALICE2

(a) Login nodes: provides command line access to the cluster, (b) Management nodes run the job scheduler and management tools of the cluster, (c) Shared storage: this is a high performance Panasas storage system that provides the user with ~1.5PB for data storage, (d) Compute nodes, large mem nodes and GPU nodes: these are nodes available for job execution, (e) SPECTRE login nodes: these provide ALICE users with a remote desktop connection via an NX client, (f) Interconnect: allows for interconnection of each component to all others via a high-bandwidth low latency FDR Infiniband Interconnect.
Source: http://www2.le.ac.uk/offices/ithelp/services/hpc/alice/architecture

With ALICE, parallelization of jobs can be achieved via script-directed job submissions to any of the available or specifically requested compute nodes. The scheduling of jobs on ALICE is managed by the scheduler (Moab version 7 developed by Cluster Resources, Inc.), which queues jobs and then executes them based on the job's requirements and available resources. In this study, we performed our data analysis on ALICE by using the job scheduling function (i.e., qsub from Torque 6.0.2). Upon completion of tasks, the results were returned from the compute nodes to the login node. ALICE currently

has 2 head nodes (for running essential services of the cluster), 2 login nodes (providing an interface for the user), 900TB of storage space and 180 compute nodes (nodes for which tasks are executed). The 180 compute nodes on ALICE comprise 170 standard compute nodes available just for job execution. Each of these standard nodes are made up of a pair of 14-core Intel Xeon skylake CPUs running at 2.6GHz, having a Random Access Memory (RAM) of 128GB, and a storage space of about 428GB. There are another 6 large-memory compute nodes, each having 48 cores, running at 3.00GHz and 1TB of RAM. Lastly, there are four graphical processing unit (GPU) nodes enabling faster calculations on the cluster. Each of these comprises a 64GB RAM, 2 Ivy Bridge CPUs at 2.5GHz and 2 X NVIDIA Tesla K40m GPU cards. These statistics represent the recently upgraded HPC (https://www2.le.ac.uk/offices/itservices/ithelp/services/hpc/alice/architecture).

2.1.2 Programming languages

2.1.2.1 Practical Extraction and Report Language (Perl)

Perl (http://www.perl.org) is a language designed purposely for string manipulation. Created by Larry Wall in 1986, as a result of the limitations that programs like sed, awk, C, and Bourne shell offered in the manipulation of texts. Perl has undergone several improvements, moving from version 1.000 to the stable version 5.26.1 as at May 2017. In this project however, I have used the stable version 5.20.0 (http://www.cpan.org/src/5.0/perl-5.20.0.tar.gz)
Perl finds strength in its ability to interact with its environment, as well as its ability to open and manipulate files efficiently. Beyond string manipulation, Perl has found great use in server-side scripting, database management, system administration, and development of web-based applications via its Common Gateway Interface (CGI) module. Today, Perl5 is able to handle regular expressions for pattern matching operations, deal with nested data structures, allowing for modularity and reusability. Thus, the choice of Perl as the language for this analysis. Perl5 is now Object-Oriented-allowing for class definitions and modelling, POSIX compliant (that is, the language maintains compatibility between operating systems) and can handle multiple concurrent database management software implementations. Most importantly, Perl is a free

program with a large community of developers constantly improving the language. Perl is readily available for installation on any major Operating System (OS), thus allowing for widespread use and portability of code.
In this work, Perl was used in developing custom scripts for parsing, processing and analysing large biological data sets as well as for implementing custom algorithms developed to manipulate the data sets.
Table 2 below presents a list of non-standard Perl modules used in this research.

Table 2: Non-standard Perl modules used

Module	Function
Text::Levenshtein qw(distance);	Used to calculate the Levenshtein edit distance between two strings.
LWP::Simple	Provides a simplified view of the libwww-perl library.
Getopt::Long	Implements an extended getopt function called GetOptions(). This module allows program options that need to have long names instead of single letters and can be introduced by a double dash –.

These modules can be downloaded via the Comprehensive Perl Archive Network (CPAN) http://www/cpan.org

2.1.2.2 Python

This is a general purpose, object-oriented interpreted language that is used for scripting applications and a number of standalone projects. The language emphasizes code readability and enables programming in fewer lines of code.
In this study, we used python language as a tool for manipulating fastq formatted files. The version used in this study was 2.7.6

2.1.3 Aligners

2.1.3.1 Basic Local Alignment with Successive Refinement (BLASR)

BLASR is a PacBio proprietary alignment program that performs rapid mapping of Single Molecule Sequence (SMS) reads to a genome by searching for the highest scoring local alignment or set of local alignments between the reads and the genome (Chaisson and Tesler, 2012). The program is designed with focus on PacBio's extraordinarily long reads, which are dominated by insertion

and deletion errors (Chaisson and Tesler, 2012). With quality values, BLASR does rapid read mapping of reads to genome with high accuracy.
In this study, we used BLASR as one of the alignment programs for mapping our reads to the reference. The BLASR program was obtained in its source code form from https://github.com/PacificBiosciences/BLASR, which was then compiled to run on the ALICE2 HPC. The version used was 1.3.1

2.1.3.2 Burrows-Wheeler Aligner (BWA)

BWA is a software package designed to map low-divergence sequences against large reference genomes (Li and Durbin, 2009). The aligner comprises three algorithms, each designed with particular kind of read in mind. BWA-Backtrack (the first algorithm) is built to handle short illumina reads (about 100 bp in length), while BWA-SW and BWA-MEM are built with support for long reads and split alignment (Li and Durbin, 2010). However, BWA-MEM, the most recent of the three algorithms, is highly recommended for high quality queries (having low error profile and containing fewer gaps when aligned against a reference) because it is faster and more accurate. In comparison to BWA-Backtrack, BWA-MEM is more efficient in dealing with illumina reads between 70bp – 100bp long.
We used the BWA-MEM algorithm as an alternative to BLASR for mapping SMS reads against a selected reference for the sake of comparison and evaluation of BLASR. BWA was downloaded in its source form from http://sourceforge.net/projects/bio-bwa/files/, and then compiled on the HPC. The version used in this study was the bwa 0.7.8

2.1.2.3 Local Alignment and Search Tool (LAST)

LAST is a modified form of the seed and extend approach adopted by alignment programs like Basic Alignment and Search Tool (BLAST) (Chaisson and Tesler, 2012). The principle distinction in LAST is the use of adaptive seeds (Kiełbasa *et al*., 2011). In using LAST to align reads against a reference, the choice of matches is based on their rareness rather than on a fixed-length of match (Kiełbasa *et al*., 2011). This ensures that the number of matches, as well as the running time, scales in a linear manner rather than in a quadractic fashion, based on sequence length (Kiełbasa *et al*., 2011). As a result, LAST is

a fast, memory efficient aligner of large sequences with arbitrarily non-uniform composition (Kiełbasa *et al.*, 2011). The aligner is an open source package downloadable from http://last.cbrc.jp/ as a compressed zip file. The source file was compiled on the HPC by Liam Gretton. In this study, LAST was used to perform pairwise alignment of reads against the unmasked (custom) reference in order to generate a score distribution profile. The version used in this study was LAST-4.60

2.1.2.4 LALIGN

This is a pairwise alignment program built to find internal duplications by computing non-intersecting local alignments of nucleotide or protein sequences. LALIGN was developed by William Pearson as an implementation of Huang and Miller's time-efficient, linear-space local similarity algorithm (Huang and Miller, 1991). In this study, we used LALIGN as an alternative aligner to LAST to perform pairwise alignment of reads against the custom reference. We used the standalone version of LALIGN, which is part of the FASTA package of sequence analysis program and downloadable from ftp://ftp.ebi.ac.uk/pub/software/unix/fasta/. The version used in this study was fasta36.3.6

2.1.4 Data visualization tools

2.1.4.1 Integrative Genomics Viewer (IGV)

IGV is a visualization tool that allows for the interactive exploration of large and integrated genomic datasets. With support for an array of data types, IGV enables the graphical view of array-based and next generation sequence data, as well as genomic annotations (Robinson *et al*., 2011; Thorvaldsdóttir, Robinson and Mesirov, 2013). In this study, we used IGV to view the result of our alignments, allowing a pictorial representation of how well our reads mapped to the reference sequence, to examine the consensus sequence, to see the error distribution at a glance, as well as to identify sequence variants. In this study, we used the standalone binary version of IGV, which was obtained from http://www.broadinstitute.org/software/igv/download. The version of IGV used was 2.3

2.1.4.2 Jalview

Jalview2 offers an interactive system for visualizing and manipulating multiple sequence alignments. The desktop version was installed for this project to enable access to web services for sequence alignment and retrieval of alignments, sequences and annotation from public databases and any DAS 1.53 complaint sequence or annotation server (Waterhouse *et al*., 2009).

2.1.4.3 L1Xplorer

L1Xplorer is the web interface for the dedicated database (L1base) (Penzkofer, Dandekar and Zemojtel, 2005). We used this interface to query L1base for the structure of our reconstructed L1 sequences.

2.1.5 Sequence analysis and manipulation tools

2.1.5.1 Sequence Alignment/Map Tools (Samtools)

Samtools is a software package with utilities for performing post-processing of alignments in the Sequence Alignment Map (SAM) format (Li *et al*., 2009). Sequence Alignment/Map is a generic format used in storing alignment of reads against reference sequences. It supports both short and long reads from varying sequencing technologies. With its flexibility in style, compactness in file size, and efficiency in random access, it has become the most popular file format for NGS read alignments for most alignment programs. Samtools was downloaded from
http://sourceforge.net/projects/samtools/files/ as a binary file.
In this study, we used the following utilities within the Samtools package:
a. Samtools index: this utility enables the indexing of alignments by genomic position so as to improve the retrieval time of reads aligned to a particular locus.
b. Samtools view: this utility enabled the conversion of alignments from text (sam format) to binary (bam format) in order to reduce file size and to have a format acceptable in IGV.
c. Samtools sort: this utility enabled the sorting of the bam file based on start positions. This is a requirement for viewing alignments in IGV.
The version of Samtools used in this research was 0.1.9

2.1.5.2 Bioperl

This is an ensemble of reusable Perl modules offering generalized routines centred on life science information (Stajich *et al*., 2002). It is widely used in bioinformatics applications particularly in parsing blast search output files, carrying out multiple sequence alignments using ClustalW or Muscle and giving access to GenBank and SwissProt files. The version of Bioperl used in this project is 1.6.1. Listed in Table 4 with a brief description are the Bioperl modules used in this project. The following Bioperl modules were used in this project:

Bio::SeqIO - This is a biological file handle for getting at format objects in SeqIO

Bio::DB::Fasta - This offers indexed access to fasta files, several files or a directory of files

2.1.6 Assemblers

2.1.6.1 MIRA

Mira is a sequence assembler and sequence mapper for whole-genome shotgun and Expressed Sequenced Tag (EST) or RNA-Seq sequencing data (www.chevreux.org). However, it can be used for assembling sequence data from varying sequencing technologies such as Sanger, 454, Illumina, IonTorrent, and PacBio. Having started as a PhD project in 1997, Mira has now developed into a full assembly program which has been found useful in assembly projects filled with repetitive sequences. We downloaded Mira as a pre-compiled binary setup from http://sourceforge.net/projects/mira-assembler/, which we used initially in performing *de novo* and mapping sequence assemblies. The version of Mira used was MIRA 4.

2.1.6.2 Celera Assembler (CA)

This assembler is a *de novo* based WGS assembler developed at Celera Genomics in 1999. However, it was only made public in 2004 after its successful use in the sequencing of *Drosophila melanogaster* (Adams *et al*., 2000). CA can re-assemble long sequences of genomic DNA from fragmented data generated by whole-genome shotgun sequencing. CA accepts reads from a wide variety of sequencing techniques such as Sanger dideoxy, Pyrosequencing, SBS, and SMS. In this study, we initially used CA side-by-side

with MIRA to perform *de novo* assembly of reads selected to have mapped to our region of interest (ROI). However, as CA became unsupported, coupled with the difficulty in MIRA processing very long reads (>29kb), our reported reconstructions were assembled using Canu. CA was downloaded from http://sourceforge.net/projects/wgs-assembler/files/wgs-assembler/wgs-8.1/ as a source file, which was then built on the HPC. The version used in this study was wgs-8.1

2.1.5.3 Canu

Canu is a fork of CA built particularly for high-noise SMS data. The Canu pipeline comprises many different executable programs and Perl driver scripts. Canu implements new overlapping and assembly algorithms such as an adaptive overlapping strategy based on tf-idf weighted MinHash and a sparse assembly graph construction that avoids collapsing diverged repeats and haplotypes (Koren *et al*., 2017). The version used in this study was 1.0

2.1.6.4 Falcon

This is an experimental diploid assembler developed by PacBio in 2014 as a set of tools for fast aligning long reads for consensus and assembly. The version used in this work is FALCON-integrate v0.3.0

2.1.7 Data Sources

2.1.7.1 UCSC Genome Browser

This browser was developed and is maintained by the Genome Bioinformatics Group at UC Santa Cruz Genomic Institute and the Center for Biomolecular Science and Engineering at the University of California Santa Cruz. UCSC Genome browser (UCSC-GB) stores reference sequences and draft assemblies for a large collection of genomes. In this study, we retrieved from the UCSC-GB data page (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/chromosomes/), our reference genome sequence chromosome by chromosome (GRCh38/hg38 version of the human genome) and the repeat masker track for full-length L1 Human Specific (L1HS) elements for each Chromosome.

2.1.7.2 NCBI Sequence Read Archive (SRA)

As the name implies, this is an online service that offers the research community access to biological sequence data so as to enhance reproducibility and offer prospects for new discoveries by comparing data sets.

We installed the SRA toolkit and configured it to run from user's path before performing sequence download**.** See below a summary of the datasets used throughout this thesis.

Table 3: Summary of datasets used in this project

Dataset	**Description**	**NCBI/ENA Study Accession**
CHM1 structural variation from single-molecule sequencing	This data was generated using a 30kb sample prep protocol on the SMRT sequencing pipeline. Sequenced to ~60x coverage, this CHM1 data generated using the P6-C4 chemistry with its improved read lengths and high accuracy, enhanced the potential for resolving gaps and generating error free assemblies. A total of 304 PacBio SMRT cell runs produced 420.72G bases.	SRP044331
Assembly and diploid architecture of an individual human genome	Using genomic DNA, PacBio sequenced the individual (NA12878) via a WGS strategy using its P5-C3 chemistry. The data was generated from 26 PacBio SMRT cell runs with a total of 197.1G bases.	SRX627421
Nanopore sequencing and assembly of a human genome with ultra-long reads	ONT sequenced the same individual as PacBio above, but with Oxford Nanopore MinION using 1D ligation kits (450 bp/s) using R9.4 chemistry (FLO-MIN106). A combination of the full datasets in release 3 (generated from 39 flowcells at 30x coverage) plus ultra-long reads generated from additional 14 flowcells at 5x coverage (release 4) were utilized in this project. In total, the analysed dataset comprised of 114,380,310,980 bases and 15,599,452 reads.	PRJEB23027

2.1.7.3 Github

This is an open source development platform that allows developers to host and share code, track projects, and build software alongside a community of developers (www.github.com). We accessed the Nanopore dataset for the sequencing of the individual (NA12878) on github (https://github.com/nanopore-wgs-consortium/NA12878). Also, the modules that run the pipeline can be found at https://github.com/ndliberial/smrt_pipeline

2.1.7.4 The European database of human-specific L1 (L1-HS) retrotransposon insertions in humans (euL1db)

euL1db was developed as a free resource to allow its users search, view and download data on known published L1-HS insertions (Mir, Philippe and Cristofari, 2015). This database of L1-HS insertions is based on the GRCh37 reference assembly and is currently maintained by the Institute for Research on Cancer & Ageing of Nice (IRCAN), France.

2.1.7.5 L1Base

This is a dedicated database which contains putatively active L1 insertions in human and rodent genomes. To be considered as an insertion in L1base, the active L1 must be full-length (>6kb), intact in the two open reading frames (ORFs) and intact ORF2 but disrupted ORF1 (Penzkofer, Dandekar and Zemojtel 2005).

2.2 Internal components

2.2.1 Scripts for Assembly and Analysis Pipeline

The Perl scripts written for use in the pipeline is a collection of functions designed to handle specific tasks within the pipeline. I discuss the scripts relative to the functions below. You can get access to the scripts listed below in our github repository (https://github.com/ndliberial/smrt_pipeline).

2.2.1.1 get_raw_reads

This function handles the retrieval of raw sequence reads from the SRA. To successfully download sequence reads, it takes the SRR number of the first read alongside the total number of reads in the experiment.

2.2.1.2 custom_extract

This function extracts sequences from a defined region from any given sequence.

2.2.1.3 mask_ref

This ro utine masks reference sequences by replacing the ROI with Ns. That way it is certain that any mapping done with the raw reads against the masked reference would avoid reads mapping to the ROI.

2.2.1.4 alignment

This routine presents the choice of locally mapping raw reads against reference sequence or remotely on the cluster. Reads could be mapped either with BLASR or BWA-MEM depending on user preference. For Blasr, default parameter settings were utilized, however, with BWA-MEM, we introduced the -x flag accordingly. See be the syntax and parameters used for each technology.

PacBio: bwa mem -x pacbio $query $reference > $output

Nanopore: bwa mem -x ont2d $query $reference > $output

2.2.1.5 flank_reads

On completion of mapping, the coordinates and chromosome identifier (id) of the ROI are passed as arguments to this function to extract read ids mapping to a 5kb or 10kb flank of the ROI.

2.1.1.6 mapped_read_extracts

The read identifiers extracted above are passed to this function which then retrieves the corresponding reads from the equivalent raw sequence file. All extracted sequences are joined to form a single mapped read file.

2.1.1.7 pairwise_alignment

This function is designed to handle the pairwise alignment of mapped reads with the custom reference. The user can choose what type of alignment program (LAST or LALIGN) to use for the pairwise alignment. A tab-delimited file containing the read identifier and alignment score is generated for reads with alignment score > 0. Here are the commands.

LAST: lastal -k1 -T0 -m10 -w0 -g1.0 $database $query > $output

LALIGN: lalign36 -T 16 -Q $query $reference -O $output

2.2.1.8 filter_reads

This function goes on to categorize reads into three groups based on a score profile. Using the output from the pairwise alignment, a score distribution profile that splits the mapped reads into a 90%, 98% and 100% proportion is implemented. The idea of splitting reads into arbitrary bins was to class reads into some form of low (100%), medium (98%) and higher (90%) identity groups. More details on this grouping is given in section 3.3.1, page 61.

2.2.1.9 assembly

Based on user preference, the categorized mapped reads can now be assembled using either CA, Canu or MIRA. Here are the commands and parameters used during assembly.

Pacbio: canu -d $output_dir -p $output_prefix -genomeSize=12.1m -corMhapSensitivity=low -corMinCoverage=2 -errorRate=0.25 -pacbio-raw $query -userGrid=false

Nanopore: canu -d $output_dir -p $output_prefix -genomeSize=12.1m -corMhapSensitivity=low -corMinCoverage=2 -errorRate=0.04 -nanopore-raw $query -userGrid=false

2.2.1.10 edit_distance

This function allows us to class each repeat unit identified in the minisatellite allele using standard MVR-PCR coding letters. Also, where the identified repeat unit fails to match a known repeat type, we apply the edit distance function to find the nearest repeat type match. In this instance, the derived repeat type is displayed in lowercase.

3. Recovery of Minisatellite Alleles

This chapter reports the *de novo* local assembly and structural recovery of both coding and non-coding minisatellite alleles from unprocessed SMS data, released by PacBio (http://www.ncbi.nlm.nih.gov/Traces/sra/?study=SRP044331) and (https://www.ncbi.nlm.nih.gov/sra?term=SRX627421) and Nanopore (https://github.com/nanopore-wgs-consortium/NA12878).

3.1 Minisatellites

Minisatellites or variable number tandem repeats (VNTRs) are regions of DNA containing tandem repeats of short nucleotide sequences (5 – 50 bp), which can extend over many kilobases (Vergnaud and Denoeud, 2000). They show high variation in the number and types of repeats, thereby resulting in multi-allelic variation and high degrees of heterozygosity (Jeffreys, 1987). The high level of variation shown by minisatellites makes them important as genetic markers (Nakamura *et al*., 1987). The role played by VNTRs in solving problems in human genetics includes individual identification in forensic settings (Gill, Jeffreys and Werrett, 1985; Wong *et al*., 1987), linkage analysis (Nakamura *et al*., 1987), and the determination of family relationships (Jeffreys *et al*., 1985). Minisatellites are generally GC rich and some can contain a core unit of nucleotides which is thought to serve as a recombination signal to enhance unequal crossing over at minisatellites (Jeffreys, Wilson and Thein,1985; Steinmetz, Uematsu and Lindahi, 1987; Jeffreys, 1987). Most minisatellites are non-coding (MS1, MS205, and MS32), but a few occur within protein coding regions (such as PRDM9 and ZNF93, studied in this thesis) (Berg *et al.,* 2011; Jacobs *et al*., 2014).

Here, I present the recovery of the structures and sequences of three (3) non-coding minisatellites and two (2) coding minsatellites from PacBio's shotgun sequence dataset of unprocessed long reads generated from SMS of the human cell line CHM1htert (http://www.pacb.com/blog/data-release-54x-long-read-coverage-for/). This recovery process uses software embedded within the analysis and assembly pipeline illustrated in Figure 17. The pipeline is described in detail in section 3.3.1 on minisatellite recovery.

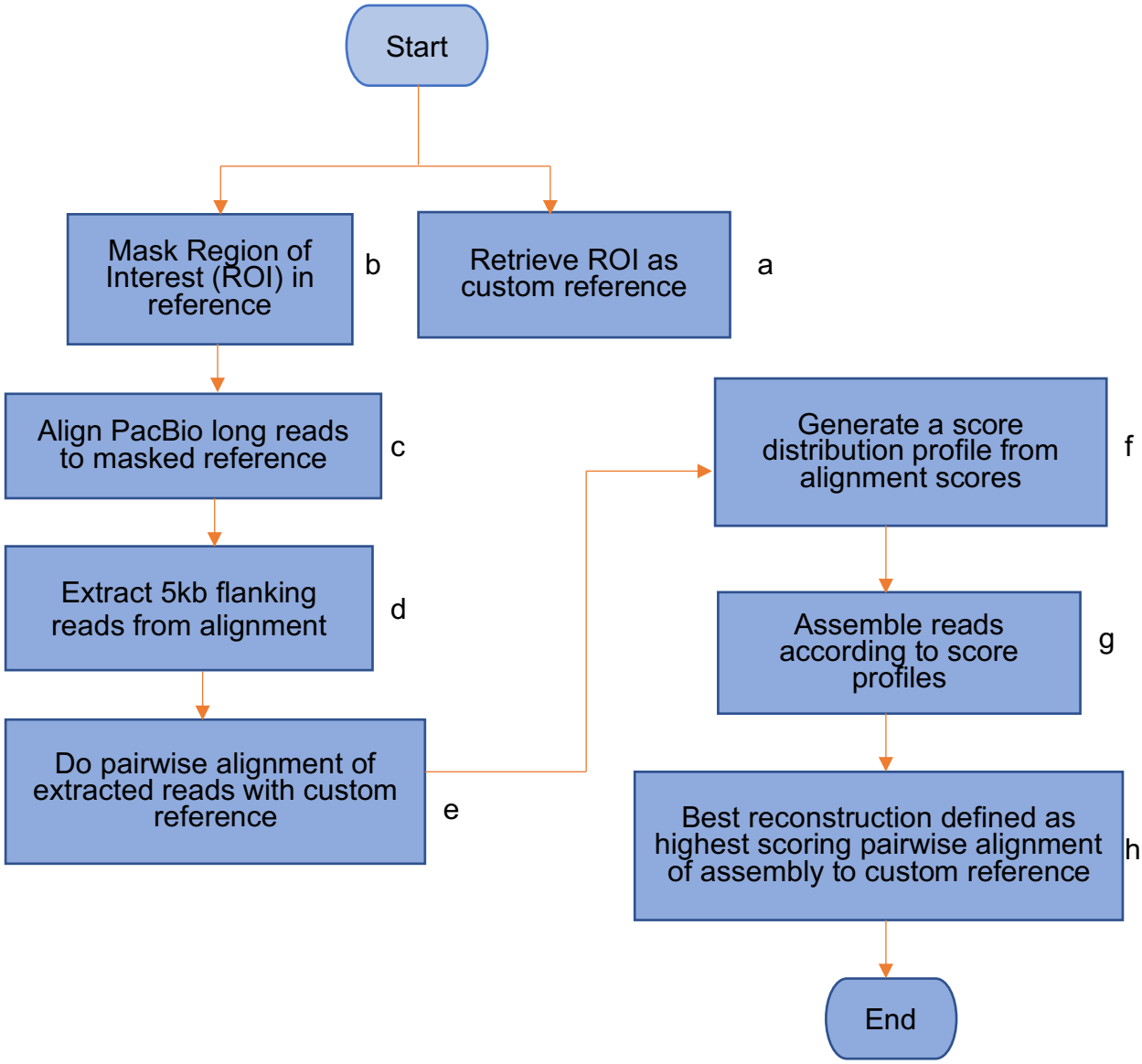

Figure 17: Analysis and assembly pipeline workflow
Orange arrows indicate progression from the start to the end of the analysis and assembly process. Each stage indicated by a blue rectangle must be completed in order for the next stage to commence.
ROI refers to the DNA sequence of the "region of interest" to be recovered.

3.2 Data Source

The data for this experiment is a shotgun sequence dataset of long PacBio reads from the genome of an immortalized human cell line. The data was generated from ~54X SMRT sequencing of genomic DNA from the human cell line (CHM1 htert). This is a haploid cell line, derived from a complete hydatidiform mole, immortalized by expression of the human telomerase reverse transcriptase (hTERT) (Teague *et al.*, 2010). With this long-read data,

PacBio demonstrated the utility of long reads for generating high-quality de *novo* assemblies of increasing size and complexity. This was the first ever *de novo* human genome assembly generated from PacBio sequence reads alone. Following on from this study, PacBio recently released a highly contiguous diploid human genome assembly derived from sequencing of the well-studied NA12878 lymphoblastoid cell line genome using a combination of SMRT sequencing and optical mapping (by generating single molecule genome maps with nicking enzymes) (Pendleton *et al*., 2015).

3.2.1 Properties of the CHM1 dataset

Table 4: Salient statistics of the CHM1 dataset

Total number of reads	Total number of post filtered bases	Average throughput/SMRT cell	Longest DNA insert sequenced	Average read length
21,856,161	167,851,128,644	608 Mb	42,774 bp	7,680 bp

Post-filtered bases refer to nucleotides that meet filtering criteria set by the user.

Figure 18 shows the read length distribution of sequence reads generated from the SMRT sequencing the haploid cell line.

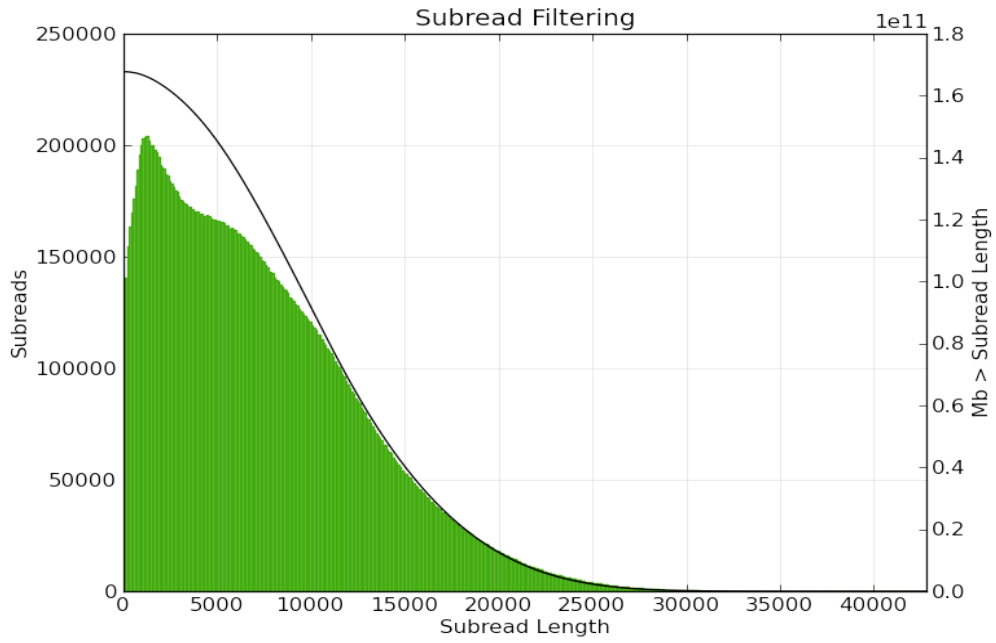

Figure 18: Subread length distribution

The left y axis shows in green, the frequency of subreads with the length on the x axis. A subread is a DNA insert sequenced between two SMRTbell TM hairpin adapters. The solid black line (right y axis) denote the amount of sequenced bases greater than a given subread length (x axis)

Source: (Figure 2 from http://www.pacb.com/blog/data-release-54x-long-read-coverage-for/)

3.3 Reconstructing Repetitive DNA

To test the ability of PacBio's long reads to reconstruct long, highly repetitious human repeats, three candidate non-coding minisatellites (MS1, M32, and MS205) and two candidate coding minisatellites (ZNF93 and PRDM9) were selected. NB: the approaches and a proportion of the results presented in this section were published in Ogeh and Badge, 2017.

3.3.1 MS1

This is the most unstable, and variable, human minisatellite isolated to date (Jeffreys *et al*., 1988). Its spontaneous germline mutation rate to new length alleles is approximately 0.05 per gamete (Jeffreys *et al*., 1988). MS1 is located at chromosomal coordinates chr1: 31904516-31905248 on GRCh38 and consists of a 9bp tandem repeat unit with copy number ranging between 140 to 2500 and having an allele-length heterozygosity predicted to be higher than 99% (Wong *et al*., 1987; Jeffreys *et al*., 1988). Despite the detailed knowledge of this minisatellites properties, the allele represented in GRCh38 is inconsistent with this. Even though the reference allele consists of a 9bp tandem repeat structure, its repeat copy number is only 52.

Using primers 5'-GCTTTTCTGTGATGAGCCTTGATG-3' and 5'-AGAAGCATATGCAACCCATGAGG-3' for MS1 (Gray and Jeffreys, 1991), the ROI was extracted from the genome reference assembly (GRCh38) using *In-Silico* PCR (http://rohsdb.cmb.usc.edu/GBshape/cgi-bin/hgPcr) to generate custom reference “bait sequences” (Figure 17a). We then masked (Quinlan and Hall, 2010) all repeats present in our ROI (Figure 17b) to generate a masked reference. We aligned the long PacBio® reads to the masked reference (Figure 17c) using BWA-MEM and BLASR. On completion of alignment, we extracted 10kb, 5kb, and 2.5kb flanking reads across the ROI (Figure 17d) to generate spanning reads which potentially held information that would enable the reconstruction of the repeat array. Considering the amount of reads in each collection, and the possibility of reads being unrelated (because they were far from the ROI) to our ROI, we proceeded with the 5kb collection of flanking reads. Using LAST (Kiełbasa *et al*., 2011) to align reads to the custom reference, a score distribution profile of reads with >90% identity to the region was generated (Figure 17e, 18f). The reads in each score distribution (90%,

98% and 100%) bin were assembled (Figure 17g) using Celera Assembler 8.1 (Myers *et al.,* 2000) and Canu 1.0 (Koren *et al*., 2017). Using a score distribution bin allowed for retention of long, lower identity alignments that provide contiguity and at sufficient coverage are expected to yield accurate consensus sequences. This is in light of the knowledge that long reads and the majority indel random error mode of the PacBio system means, high identity short alignments can contain less information about repeat structures, than long lower identity alignments. Also, it was hoped that reads that introduced assembly noise would be efficiently removed, while maximizing the recovery of informative reads. LAST was chosen for its use of adaptive seeds (Kiełbasa *et al*., 2011) when performing alignment, thereby guaranteeing that the number of matches and the running time scales in a linear fashion, rather than quadratic, based on the sequence length. Also, the LAST aligner is relatively fast, memory efficient, and suitable for large sequences. In addition, LAST can also be tuned to optimize for long, weak alignments. The highest scoring pairwise alignment of assembly to the reference was used to recover the tandem repeat array. (Figure 17h).

To illustrate that our local assembly pipeline integrates multiple array spanning reads as well as reads that terminate within the MS1 array we visualized their mapping using IGV (Robinson *et al*., 2013). 82 individual reads contribute to the assembly shown in Figure 19. The red coloured region indicates the location of the minisatellite repeat array.

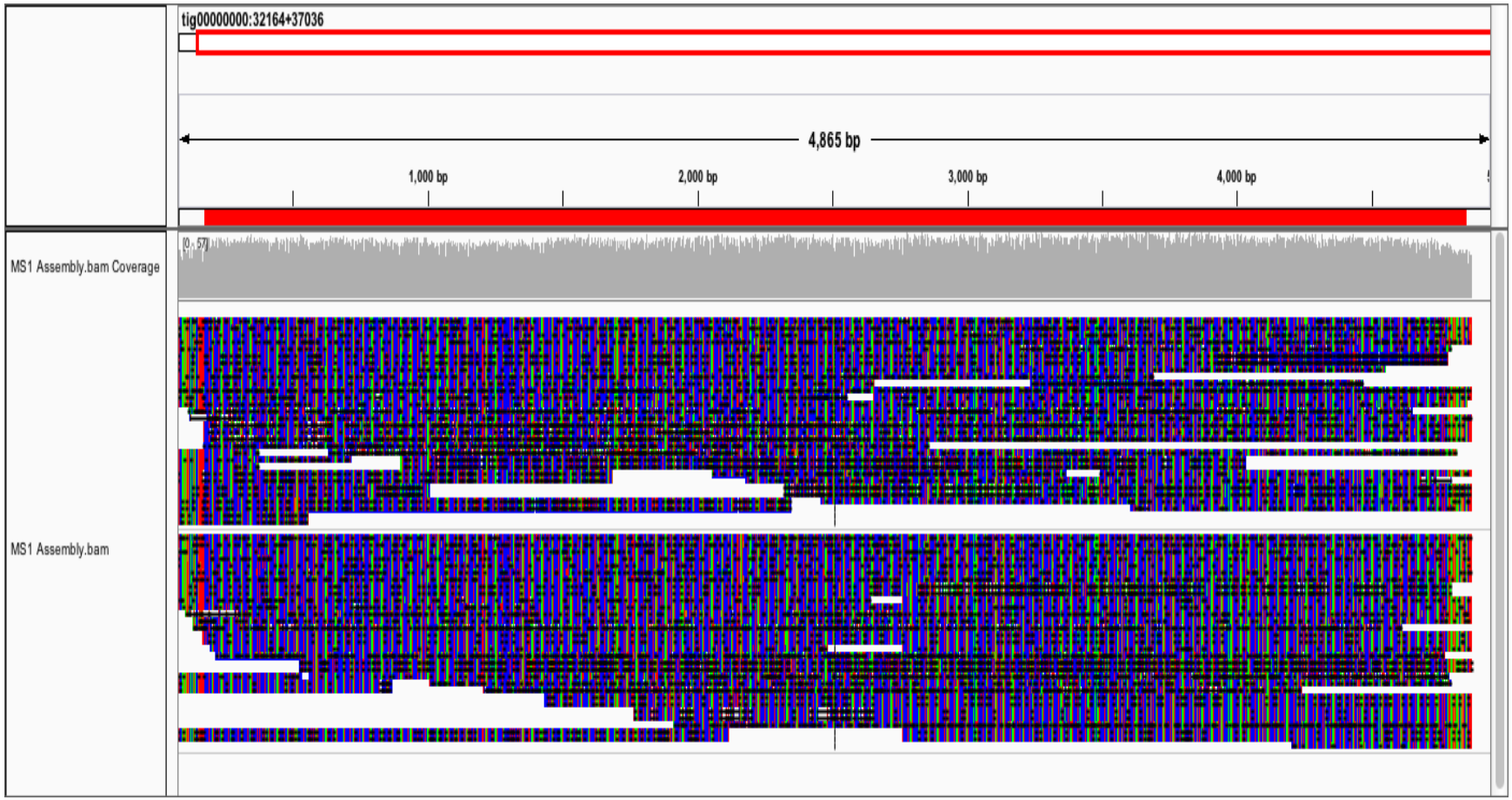

Figure 19: IGV visualization showing mapping of assembly contributing reads to the assembly consensus.

The image shows the alignment of 82 individual reads that make up the assembly, aligned with the assembly consensus sequence. The idea is to show that the minisatellite array is an assembly from overlapping reads. The minisatellite array (red) extends through the majority of the assembly.

As seen in Table 5, the reconstructed sequence showed a 98% sequence identity to the reference allele, with a MS1 repeat number of 513.8 (see Table 5) as reported by Tandem Repeat Finder (Benson, 1999). PCR analysis measured an allele size of ~4.2kb (amplified from a MS1 containing BAC, obtained from a genomic DNA library of CHM1htert, R. Badge, *pers comm*), but our assembly shows a size of 5.4kb consistent in size with PacBio's global genome assembly. The difference in size could result from minisatellite sequence instability in the BACs generated from the cell line (Song *et al*., 2001).

Table 5: Summary statistics of assembled minisatellites

Repeat	Copy number (Ref.)	Copy number (Ass.)	Copy number (PacBio)	Identity of assembly to reference (%)	Identity of assembly to PacBio (%)	*In-Silico* PCR size (Ref.) kb	*In-Silico* PCR size (Ass.) kb	*In-Silico* PCR size (PacBio) kb	PCR size (BAC)
MS1	53.8	513.8	513.8	98.0	99.9	1.2	5.4	4.8	4.2
MS32	10.6	263.6	263.6	93.8	99.9	0.6	8	8	8
MS205	11.9	40.3	40.3	95.4	99.9	0.8	2.7	2.7	3.0
PRDM9	12.5	12.5	12.5	96.8	91.6	1.9	1.9	1.9	2
ZNF93	16.2	16.2	16.2	99.9	79.8	4.3	4.3	4.3	4.4

Ref refers to the GRCh38 reference allele

Ass refers to our locally assembled allele

PacBio refers to PacBio's globally assembled allele

A further check on the size estimates was performed on genomic DNA (CHM1) for MS1 and this showed a consistency of 5.4 kb with assembled allele (R. Badge, *pers comm)*. It is likely that the BAC allele could have collapsed on propagation in bacteria. We also note the difference in size between our assembled allele, the genomic DNA and PacBio's allele which is likely to have been as a result of PacBio's assembly process.

Table 6: TRF report on reference, global and local assemblies of MS1

Seq. Type	Period Size	Copy Number	Consensus Size	% Matches	% Indels	Score	A	C	G	T
Reference	9	53.8	9	93	0	689	19	55	3	21
PacBio	9	513.8	9	92	0	5990	16	58	2	22
Assembly	9	513.8	9	92	0	5963	16	58	2	22

Seq. Type refers to the source of the allele being described

Period Size refers to the size of the repeat units in the array

Copy Number refers to the number of repeat units aligned to the consensus sequence

Consensus Size is the size of the most commonly encountered nucleotide at a specific location in the DNA

% Matches refers to the degree of similarity or identity between adjacent copies overall

% Indels refers to the degree of similarity or identity between adjacent copies overall

Score means alignment score

The next four (4) columns refer to the percentage composition for each of the nucleotides

3.3.1.1 MS1 array structure

Previous work done by Gray and Jeffreys (Gray and Jeffreys, 1991) studying variants within the MS1 9bp repeat sequence identified 19 variations occurring as a result of base changes from the consensus repeat unit sequence. These 19 variant types were coded A to S, following the scheme developed for MVR-PCR mapping (Gray and Jeffreys, 1991). Using a custom perl script (ms1_repeat_finder.pl) for coding individual repeat units based on the variant repeat type sequences, allelic structures at MS1 for PacBio's global assembly and our local assembly were generated (see Figure 20).

3.3.1.2 Repeat type assignment

In order to present repeat units in a form close to MVR-PCR mapping structures for known minisatellite alleles, our workflow (Figure 21), based on minimum edit distance was implemented in the custom perl script (ms1_repeat_finder.pl). Individual repeat units were read from a text file (Figure 21a). Standard repeat types for the minisatellite were retrieved via text file (Figure 21b). Next a check for exactly matching repeat types was made (Figure 21c). Where the repeat unit matched a known type, the corresponding repeat type was stored for that repeat unit (Figure 21d). If the repeat unit did not match a known type, the minimum edit distance for that repeat unit is calculated (Figure 21e), and the known type with the least score was assigned to the repeat unit (Figure 21f). Where more than one known type is reported with an equal least score, the

repeat unit type was considered an unknown, indicated with an asterisk (*) symbol.

Following the repeat type assignment, Figure 21 shows the consistency in repeat types order between our locally assembled allele and the PacBio's globally assembled allele. This again, demonstrates the validity of our assembly and the utility of the recovery pipeline. For the complete sequence data and allele structure of our locally assembled MS1, kindly refer to our github repository (https://github.com/ndliberial/smrt_pipeline/sequences).

```
>recovered_assembly
ABCDEACAA-AAABBB*AAABBAACCBBBBAA---CCCKAAAKABBBBB--BACCCAAKKCABBAAKc-AAAAAAKAAAAAKMKKJBB-
>pacbio_assembly
ABCDEACAACAAABBBBAAABBAACCBBBBA----CCCKAAAKABBBBBbbBACCCAAKKCABBAAKCCAAAAAAKAAAAAKMKKJBB*
```

Figure 20: A snapshot of the internal structures of our locally assembled allele and PacBio's globally assembled allele.

The letters (**ABCDEAC)** indicate a highly stable 5' cluster of 7 repeat units**.** The letters (**AKMKKJBB**) also indicate a highly stable 3' end structure. Both 5' and 3' ends of the local and global assemblies show consistency with literature (Gray and Jeffreys, 1991).

Lower case letters represent repeat types assigned via the edit distance algorithm. Upper case letters represent exact repeat type matches to the A-S MVR types.

* means multiple equally likely (based on edit distance) repeat type matches were found, hence no type was called for the repeat unit.

3.3.1.3 Minimum edit distance algorithm

The minimum edit distance algorithm implemented in this thesis (see figure 21) is the Levenshtein Distance (Levenshtein, 1966). This is a measure of the similarity between two strings. The distance is calculated as the number of insertions, deletions or substitutions needed to change one string into another. And in this implementation, we focused on the transformation with the least cost of transformation.

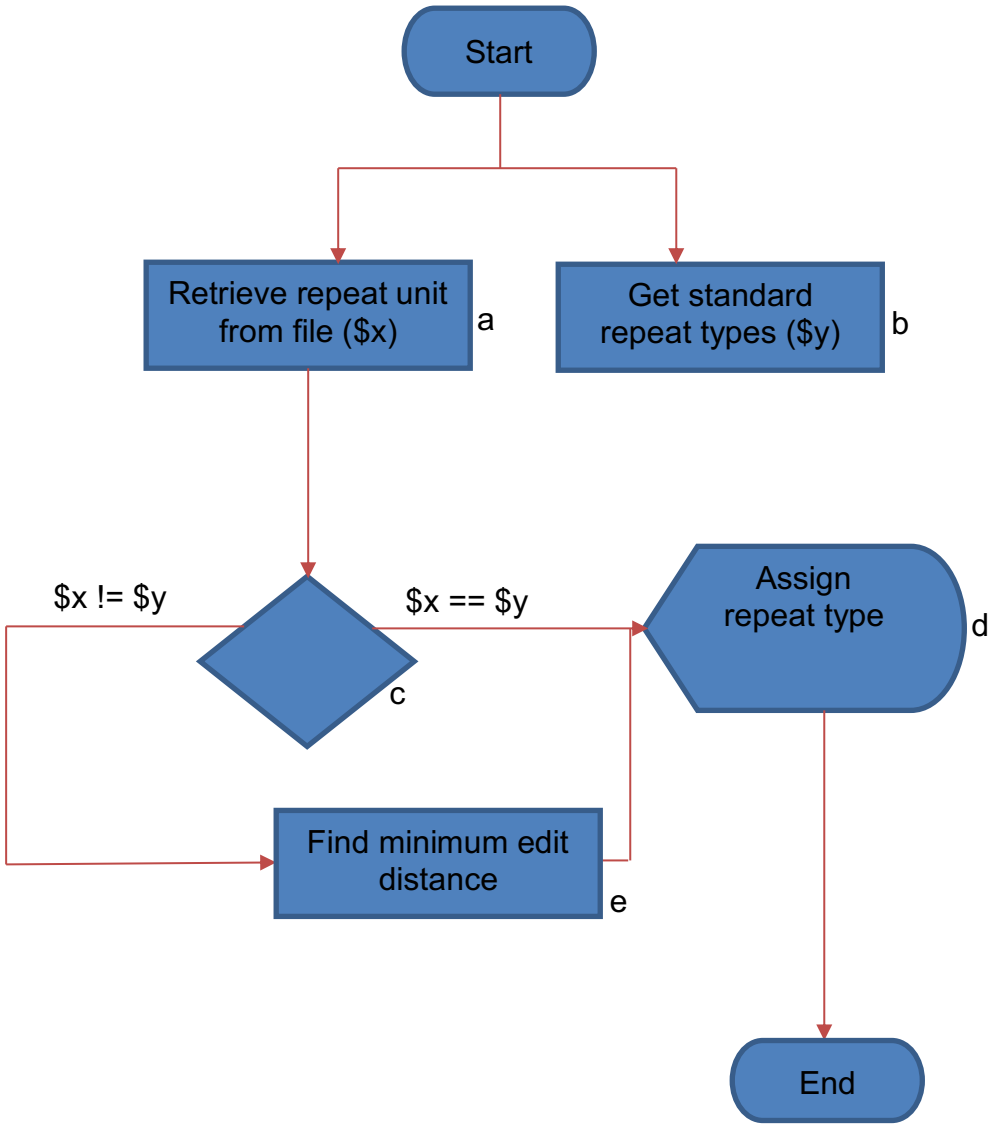

Figure 21: The sequence processing workflow implemented in the script ms1_repeat_finder.pl

Orange arrows indicate progression from one stage to another within the workflow. Each stage indicated by the blue object must be completed in order for the next stage to commence.

3.3.2 MS32

This highly variable minisatellite is located interstitially at chromosomal coordinate chr1: 236260023-236260914 on GRCh38. Having expanded from within a retroviral long terminal repeat-like element, this VNTR consists of a 29 bp repeat sequence (Wong *et al*., 1987; Armour *et al*., 1989). MS32 alleles contain between 12 and 600 repeat units with an allele-length heterozygosity estimated at 97.5% (Armour, Wong, *et al*., 1989). In humans, almost all repeat units are identical except for a T - G transversion that occurs in roughly two thirds of the repeat units (Wong *et al.*, 1987; Jeffreys, Neumann and Wilson, 1990). Previous work by Jeffreys (Jeffreys, Neumann and Wilson, 1990) showed that MS32 alleles in humans mainly comprise two variant repeat types letter-coded as A and T, and only differ by a single base substitution. Also, allelic diversity has been found to be greatest at the 3' end of the array (Gray and Jeffreys, 1991).

Using primers 5'-TCACCGGTGAATTCCACAGACACT-3' and 5'-AAGCTCTCCATTTCCAGTTTCTGG-3' (Gray and Jeffreys, 1991) to extract the ROI via *In-Silico* PCR, the analysis and assembly pipeline was implemented as previously described (see section on MS1).

From Table 5 (on page 63), the reconstructed sequence shows a 93.8% sequence identity to the reference with a repeat copy number of 263.6. Table 7 gives a summary report of 3 alleles as reported by TRF. The recovered allele size is predicted to be 8kb. The allele size from genomic DNA could not be determined at the time of writing (using PCR).

Table 7: TRF report on reference, global and local assemblies of MS32

Seq. Type	Period Size	Copy Number	Consensus Size	% Matches	% Indel	Score	A	C	G	T
Reference	29	10.6	29	97	0	533	25	18	43	12
PacBio	29	263.6	29	96	0	13515	25	18	43	12
Assembly	29	263.6	29	96	0	13496	25	18	43	12

Seq. Type refers to the source of the allele being described

Period Size refers to the size of the repeat units in the array

Copy Number refers to the number of repeat units aligned to the consensus sequence

Consensus Size is the size of the most commonly encountered nucleotide at a specific location in the DNA

% Matches refers to the degree of similarity or identity between adjacent copies overall

% Indels refers to the degree of similarity or identity between adjacent copies overall

Score means alignment score

The next four (4) columns refer to the percentage composition for each of the nucleotides

Using the same letter coding scheme reported in MS32 MVR-PCR, we computationally determined the allelic structures of our local and PacBio's global genome assembly alleles. The complete sequence and allele structure of our locally assembled MS32 array can be found in our github repository (https://github.com/ndliberial/smrt_pipeline/sequences).

Repeat unit types were assigned using the perl script (as in section 3.3.1, page 64) and our locally assembled MS32 allele sequence. Again, we observed a consistency (>90%) in structure between our local assembly and the globally generated assembly of PacBio, which further validates our recovery algorithm.

Local : aaaaaaAaAaaaaAaaAAAAAAAAAAAAAAAAaAtT
Pacbio: aaaaaaAaAaaaaAaaAAAAAAAAAAAAAAAAaAtT

Local : TTTTTTTtAaATTTTTTTtTTTTTTtATtAaAAAa**T**
Pacbio: TTTTTTTtAaATTTTTTTtTTTTTTtATtAaAAAa**t**

Local : aTaTaTaAAaTAaaaTAaaaTaaaaAATTTATaAAA
Pacbio: aTaTaTaAAaTAaaaTAaaaTaaaaAATTTATaAAA

Local : AaAAaAAaATTATTAaTaAaTaTTAaTaAaTaTaAT
Pacbio: AaAAaAAaATTATTAaTaAaTaTTAaTaAaTaTaAT

Local : aTAtTTTTTTtATtaAAAaTaTaTaAaAAAaTaTaT
Pacbio: aTAtTTTTTTtATtaAAAaTaTaTaAaAAAaTaTaT

Local : aTaAAaTAaaaTAaaaTaaaaAATTTATaAAAAaAA
Pacbio: aTaAAaTAaaaTAaaaTaaaaAATTTATaAAAAaAA

Local : aAAaATTATTAaTaAaTaTTAaTaAaTaTaATaTaT
Pacbio: aaAaATTATTAaTaAaTaTTAaTaAaTaTaATaTaT

Local : AaTTTaTAAAaa
Pacbio: AaTTTaTAAAaa

Figure 22: A side by side layout of the internal structures of our locally assembled allele and PacBio's globally assembled allele.

The above structures were determined by computationally assigning repeat types corresponding to individual repeat units based on the MVR-PCR scheme.
Lower case letters represent repeat types assigned via the edit distance algorithm. Upper case letters represent exact repeat type matches to the A or T MVR-PCR types.
The bold letters mark areas where repeat type is assigned differently.

3.3.3 MS205

MS205 is a highly polymorphic minisatellite that maps to the short arm of chromosome 16. It is located at chromosomal region chr16: 1235539-1236362 on GRCh38, and has a heterozygosity estimated at 97% from Southern blots profiles of *Hin*fI-digested genomic DNA (Royle *et al*., 1992). The studies of

Armour (Armour *et al*., 1993) revealed extensive repeat unit sequence variation as well as variation in the length of repeat units (between 49 and 54bp). Using MVR-PCR mapping, repeats were classed into 'A' or 'T' type repeats. Previous work by Armour et al (Armour *et al*.,1993) also revealed variation in the internal map structure of alleles and a predicted true heterozygosity of 99.6%. Just like MS32, allelic diversity was greatest at the 3' end of the array, thus suggesting that almost all recently generated variation stems from one end of the tandem array (Armour *et al*., 1993).

Using primers 5'-GGAGATGCTCAGAAAGCCTAGTGG-3' and 5'-CACCACTAAGGAAACAGGCTCTCC-3' (Armour *et al*., 1993) to extract the ROI via *In-Silico* PCR, the analysis and assembly pipeline was implemented as previously described (see section on MS1, page 60).

From Table 4 (see page 64), the reconstructed sequence shows a 96.8% sequence identity to the reference with repeat copy number of 40.3, as reported by TRF (Table 8). PCR analysis of genomic DNA reveals an allele size of ~3kb (see Table 4) which is similar to the assembled size of 2.7kb.

Table 8: TRF report on reference, global and local assemblies of MS205

Seq. Type	Period Size	Copy Number	Consensus Size	% Matches	% Indels	Score	A	C	G	T
Reference	54	11.9	50	91	5	803	14	56	16	12
PacBio	53	40.3	53	89	3	2940	12	56	17	13
Assembly	53	40.3	53	89	3	2922	12	56	17	13

Seq. Type refers to the source of the allele being described

Period Size refers to the size of the repeat units in the array

Copy Number refers to the number of repeat units aligned to the consensus sequence

Consensus Size is the size of the most commonly encountered nucleotide at a specific location in the DNA

% Matches refers to the degree of similarity or identity between adjacent copies overall

% Indels refers to the degree of similarity or identity between adjacent copies overall

Score means alignment score

The next four (4) columns refer to the percentage composition for each of the nucleotides

Using same letter coding scheme as used in MVR-PCR to determine repeat types, the allelic structures of the local assembly and PacBio's global were determined. We observed that none of the derived structures conformed with any of the published alleles. It is not clear as to what might be happening here, however, with further investigation, we hope we might find an answer. Just like previous recoveries, you can find the complete sequence and derived allele

structures for MS205 in the github repository (https://github.com/ndliberial/smrt_pipeline/sequences).

3.3.4 PR domain-containing 9 (PRDM9)

PRDM9 is a meiosis-specific histone H3 methyltransferease with a C-terminal tandem repeat C2H2 zinc finger (ZnF) domain encoded by a minisatellite (Hayashi, Yoshida and Matsui, 2005). It has recently been found to be a likely trans-regulator of meiotic recombination hot spots in human and mouse (Baudat *et al*., 2010; Parvanov, Petkov and Paigen, 2010). Bioinformatic analyses have shown that PRDM9 is the ZnF protein in humans that is most likely to recognize the “Myers” motif CCNCCNTNNCCNC, which could serve as the binding site for a protein involved in the modulation of recombination intensity (Myers *et al*., 2010; Frazer *et al*., 2007). Previous work by Baudat *et al* (Baudat *et al*., 2010) reported two variant repeat types and five different alleles while Berg et al (Berg *et al*., 2010), identified an additional eighteen (18) repeat types, and twenty-four new alleles. PRDM9 alleles comprise 8-18 repeat units of 84bp, with allele A being the most common. Allele A has been shown to bind to the “Myers” motif *in vitro* (Baudat *et al.*, 2010). PRDM9 is located at chromosomal coordinates chr5: 23526096-23527995 in the GRCh38 assembly.

Using primers 5'-TGAGGTTACCTAGTCTGGCA-3' and 5'-ATAAGGGGTCAGCAGACTTC-3' (Berg *et al*., 2010) to extract the ROI via *In-Silico* PCR, the analysis and assembly pipeline was implemented as previously described (see section on MS1, page 61). As reported by TRF (Table 9 below), both reference and assembly have copy numbers of 12.5 each.

From Table 5 (see page 63), the reconstructed sequence shows a 96.8% sequence identity to the reference with same repeat unit copy number as reference. PCR analysis revealed an allele size of ~1.9 kb which matches assembly size of 1.9kb.

Table 9: TRF report on reference, global and local assemblies of PRDM9

Seq. Type	Period Size	Copy Number	Consensus Size	% Matches	% Indels	Score	A	C	G	T
Reference	84	12.5	84	94	1	1535	24	26	32	16
PacBio	84	12.5	84	94	1	1553	24	26	32	16
Assembly	84	12.5	84	90	5	1399	24	26	32	16

Seq. Type refers to the source of the allele being described

Period Size refers to the size of the repeat units in the array

Copy Number refers to the number of repeat units aligned to the consensus sequence

Consensus Size is the size of the most commonly encountered nucleotide at a specific location in the DNA

% Matches refers to the degree of similarity or identity between adjacent copies overall

% Indels refers to the degree of similarity or identity between adjacent copies overall

Score means alignment score

The next four (4) columns refer to the percentage composition for each of the nucleotides

Using same letter coding scheme as used in MVR-PCR to determine repeat types, the allelic structure of our local assembly, PacBio's global and the reference is as shown (Figure 23).

L9: ABCDDECF**GP**FQJ

L20: ABCDDECF**GK**FQJ

L24: ABCDDECF**TP**FQJ

PacBio: aBCDDECF**GH**FQJ

Assembly: aBCDDECF**GH**FQJ

Reference: aBCDD**C**CFGHF**I**J

Figure 23: Comparison of the internal structures of known (L9, 20 and 24) (Berg *et al.*, 2010) PRDM9 alleles, the GRCh38 (reference) allele, our locally recovered (assembly) allele, and PacBio's globally assembled allele.

The red coloured letters indicate variation in allele structure.

Lower case letters refer to repeat type classified using the minimum edit distance algorithm

3.3.5 ZNF93

This is a protein containing a repeating Zinc Finger motif that is involved in binding (and repressing) the L1 promoter – it was apparently so successful in this role that L1PA3 and younger elements lack the sequence that this protein binds (Jacobs *et al*., 2014). This suggests that this protein may have played a role in the evolutionary arms race between L1 and host repression systems, although its current role is unclear (Jacobs *et al*., 2014). ZNF93 is located at chromosomal region chr19: 20043417-20047776 on GRCh38 (Jacobs *et al*.,

2014). As in PRDM9, this zinc finger array has an 84bp repeat encoded by a minisatellite.

Using primers 5'-CCTAGTGGCTTGCAGTAACAA-3'and 5'-GCTCTCACAAAGGGGCATCT-3' [R. Badge *pers comm*] to extract the region of interest (ROI) via *In-Silico* PCR, the analysis and assembly pipeline was implemented as previously described (see section on MS1). As reported by TRF (Table 10 below), both reference and assembly have 16.2 repeat units each.

From Table 5 (see page 63), the reconstructed sequence shows 99.9% sequence identity to the GRCh38 reference. PCR analysis reveals allele size of ~4.4kb which matches the assembled size of 4.4kb.

Table 10: TRF report on reference, global and local assemblies of ZNF93

Seq. Type	Period Size	Copy Number	Consensus Size	% Matches	% Indels	Score	A	C	G	T
Reference	84	16.2	83	83	3	1390	36	20	17	24
PacBio	84	16.2	83	83	3	1390	36	20	17	24
Assembly	84	16.2	83	83	3	1390	36	20	17	25

Seq. Type refers to the source of the allele being described

Period Size refers to the size of the repeat units in the array

Copy Number refers to the number of repeat units aligned to the consensus sequence

Consensus Size is the size of the most commonly encountered nucleotide at a specific location in the DNA

% Matches refers to the degree of similarity or identity between adjacent copies overall

% Indels refers to the degree of similarity or identity between adjacent copies overall

Score means alignment score

The next four (4) columns refer to the percentage composition for each of the nucleotides

In general, reconstruction of minisatellite alleles using our local assembly pipeline recovers biologically feasible alleles (Figure 20, Figures 22 - 23), which have been validated by the PacBio global genome assembly. The benefit this brings is that computational biologist can relatively analyse and assemble sequences of interests locally with or without access to large compute resources. Also, answers to specific biological questions can be found quickly.

3.4 Finding novel repeats and alleles

The results from MVR analysis on MS1, MS32, and PRDM9 revealed the possibility of the existence of novel repeats and alleles respectively. On the contrary, in the case of ZNF93, no extensive MVR-PCR analysis has been

carried out, hence the inability to speculate on the idea of novelty. However, for MS205, because of the variability in length of repeat units, determining whether a non-matched repeat unit is novel or not becomes more complicated. We however can say that, local assembly of allele structures from SMS data could potentially be a good route to revealing allelic diversity.

Table 11 below gives a detailed description of potentially novel repeat units and alleles discovered via our pipeline.

Table 11: Summary statistics of Novel repeat units and alleles

Minisatellite	No. of known repeats	No. of transformed repeats	No. of novel repeats	Is this a novel allele?
MS1	500	5 (cctctcca (2), cctatcca (2), cctgtcca)	3 (ccctctgca, ccttaacca, ccctatcta)	Yes
MS32	249	12 (ccccggtcacctgctccattctgagtca (11), cccctggcacacctgctc)	3 (cccttggccgcctgctccattctgagtca, cccctggtctccacctgctccattctgagtca, ccccggtcacctgctccattctgagtcac)	Yes
PRDM9	12	1 (accaaaggacacatacaggggagaagctctacgtctgcagggag)q2	Nil	Yes

Known repeats refer to assembled repeat units that match reported repeat types

Transformed repeats refer to recovered repeat units that do not directly match known repeat types but when transformed using the minimum edit distance algorithm, we are able match them to known types.

No. of novel repeats refers to repeats that are sufficiently different (by one nucleotide change from known repeat types) to be assigned as potentially novel repeat unit types.

From the table above, we see that these repeats are mostly 1 bp short of the standard repeat length. We assume this could be a function of the sequencing technology (PacBio is known to generate high levels of indels) (Carneiro *et al.*, 2012).

Novel repeats refer to repeat units identified as not been reported anywhere in literature. These repeats conform in length to standard repeat types.

For the case of PRDM9, no new repeat type was found, however, a new allele structure was discovered based on the sequence of the repeat types.

3.5 Diploid minisatellite recovery

Further validation of assembly structure using alternative single molecule sequencing technology such as Oxford Nanopore's MinION system could not be performed, as there currently exists no data for this haploid cell line. However, sequence data does exist for the individual NA12878 and now made available by PacBio (Pendleton *et al.,* 2015) and Jain et al (Jain, Koren *et al*.,

2018) respectively. Using primer sequences as in earlier sections, we attempted to recover the ROIs for all minisatellites *In-Silico.* Results from PacBio revealed an 8kb size estimate for MS1, 1.8kb size estimate for PRDM9 and 4.5kb size estimate for ZNF93. There were no size estimates for MS32 and MS205 because we could not recover the ROI from the assembly. A further look at the *In-Silico* PCR product showed the presence of single alleles for both PRDM9 and ZNF93. The MS1 product contained two different small sized alleles with copy number 12.8 and 20. In the middle of both alleles, more than 90% of the entire PCR product comprised of unknown sequences represented by N's. From the Nanopore assembly, no *In-silico* PCR products could be generated. The primers used unfortunately were unable to find a binding site. Application of our pipeline to the ~60X and ~30X coverage datasets from PacBio and Nanopore respectively showed the possibility of presence of both alleles for all minisatellites in the PacBio data. Still, our local assemblies appeared to be fragmented in relation to the recovery of alleles. It is likely that a lack of haplotype resolution in the data may have led to the fragmented assemblies.

3.6 Recovery of Telomeres

We attempted the recovery of telomeric sequences (12q and the XpYp chromosomes) from the haploid data using the pipeline. 100 sequences were recovered for the ROI, however, the structure of these sequences were inconsistent with known telomeric types based on TVR-PCR mapping scheme. We speculate that the high dependence of PacBio library preparation systems on double stranded DNA (http://dnatech.genomecenter.ucdavis.edu/wp-content/uploads/2014/07/PacBio-Guidelines-SMRTbell-Libraries-v1.0.pdf), in the light of telomeres being partly single stranded DNA might have been responsible for the inability of the sequencing technology to sequence through these large repetitive DNA structures or perhaps the PacBio sequences are completely different from the reference sequences..

3.7 Discussion

In an effort to reconstruct repeat sequences that have been, hitherto, difficult to sequence and assemble with Sanger and short read NGS technologies, we

developed software embedded within an analysis and assembly pipeline for the acquisition, filtering, and assembly of single molecule long-read sequencing reads (particularly, PacBio). Application of this pipeline to example minisatellites from coding (PRDM9, ZNF93) and non-coding (MS1, MS32, MS205) DNA, showed that the approach was effective in recovering minisatellite alleles with over 95% identity to reference, where aligned, and enabled the recovery of internal repeat variant interspersion patterns by sequencing (Figure 20, Figures 22 – 23). Importantly, where internal repeat variant interspersion patterns have only been inferred (MS1 / MS32) from MVR-PCR methodologies, that are dependent upon partial sequencing of variant repeat units, our bioinformatic approach allows base level sequence assembly, potentially revealing novel repeat types and structures, at nucleotide resolution.

3.7.1 Application of MVR-PCR coding scheme

Using the MVR-PCR coding scheme and a custom Perl script to implement edit distance assignment of sequences to repeat types, we showed a 5' and 3' consistency in allelic structure between all three (3) minisatellite assemblies. Whilst the reference (GRCh38) contained a gap, our local assembly and PacBio's global genome assembly, both showed a recovery of missing and potentially novel repeat units for MS1 (Figure 20). At MS1, the vast majority of repeat units exactly matched the MVR-PCR coding – this, in addition to the consistency with PacBio's global genome assembly, suggests that the assembly accuracy achieved is high. Unlike MS1 and MS32, the allele structures for all three assemblies of MS205 were derived via edit distance calculations. We are not sure why this was the case for this minisatellite allele. Perhaps, it might have been as a result of the hypervariability at this locus. The inability of the edit distance algorithm to make a repeat call for certain repeat units, thereby assigning the asterisk character (*) may likely be a sequencing error or an assembly error. Our speculation is based on the fact that the repeat call from PacBio's global assembly (potentially of higher quality) at the position matches one of the multiple repeat units identified in our assembly. Also, the structure of the recovered PRDM9 allele showed consistency in both the 3' and 5' ends with known human-specific alleles (Figure 23) (Berg *et al*., 2010). The variation between the known human-specific alleles (L9, L20, and L24), the GRCh38 allele, our locally assembled

allele and PacBio's globally assembled allele is consistent with the reported ethnic variation in PRDM9 alleles (Berg *et al*., 2010; Baudat *et al*., 2010). As all known human-specific alleles and the CHM1 assembled alleles were derived from Caucasian samples (Berg *et al*., 2010; Kersbergen *et al*., 2009), we may have discovered a novel allele. We suggest further validation of these discoveries (novel repeats and alleles) be carried out via MVR-PCR on CHM1htert genomic DNA.

3.7.2 Benefit of the Algorithm

Given the accurate representation of coding minisatellites in the assemblies, and the consistency seen at both the 5' and 3' ends of non-coding minisatellites with known allele structures, as described by MVR-PCR mapping, this analysis suggests that our algorithm could be used for the characterization of repetitive sequences that are collapsed or entirely missing in human genome reference sequences consistent with the literature (Huddleston *et al.,* 2014). Another benefit of our approach is that local assembly of ROI is computationally much less resource intensive than whole genome assembly, and thus should be accessible to more researchers.

4. Characterization of Active Transposons

This chapter focuses on the *de novo* local assembly and characterization of Long Interspersed Nuclear Elements (LINEs) from SMS data generated from the following studies (Chaisson *et al*., 2015; Jain, Koren *et al*., 2018). Here are the links to the corresponding datasets http://www.ncbi.nlm.nih.gov/Traces/sra/?study=SRP044331 and https://www.ebi.ac.uk/ena/data/view/PRJEB23027.

4.1 LINE-1 elements

LINE-1 (L1) elements are the only known actively mobile autonomous member of the non-LTR retrotransposons in humans. Refer to section 1.6.2.2.1 for more details on L1 element biology. These elements at full-length (~ 6kb) are large enough to exceed the length of any single Sanger read thus affecting the contiguity of genome assemblies.

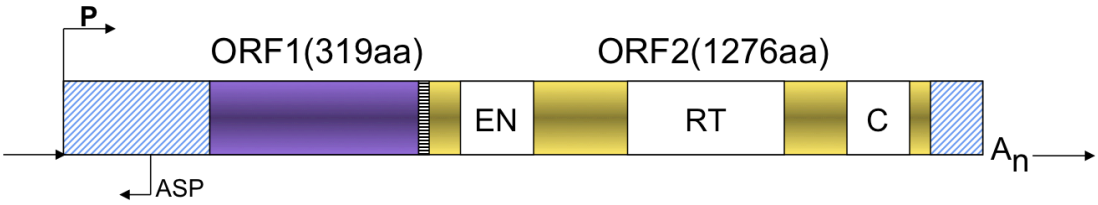

Figure 25: Structure of L1

Here, I present the recovery from PacBio's shotgun sequence dataset of long reads generated from SMS of the human cell line CHM1htert, the structure and sequence of a candidate L1 insertion on the X-chromosome (contained within accession AC002980), which was used as an assembly test case. Also, I present novel L1 insertions computationally discovered by analysis of two separate studies (Pendleton *et al*., 2015; Jain, Koren *et al*., 2018). Our recovery process uses software embedded within the analysis and assembly pipeline illustrated in Figure 17 and Figure 34. The pipeline is described in section 4.5, page 81. The scripts implementing the pipeline are located in our github repository https://github.com/ndliberial/smrt_pipeline.

4.2 Data sources

The data for these analyses were obtained from three (3) sources namely:

- 61.0X SMRT® Sequencing coverage of the CHM1 cell line (http://www.ncbi.nlm.nih.gov/Traces/sra/?study=SRP044331) for the characterization of the candidate L1 insertion on the X chromosome.

- Nanopore sequencing and assembly of a human genome with ultra-long reads (https://www.ebi.ac.uk/ena/data/view/PRJEB23027) for the discovery of novel L1 insertions.
- Assembly and Diploid Architecture of an individual Human Genome via Single Molecule Technologies (https://www.ncbi.nlm.nih.gov/sra?term=SRX627421) for the discovery of novel L1 insertions.

Table 11: Size comparison of all three datasets analyzed

S/n	Dataset	No. of Sequences	Base pair count	Size (Gb)
1	CHM1	171743491	46823141977079	320
2	NA12878 (PacBio)	51707765	33894331557295	452
3	NA12878 (Nanopore)	15599452	3347350953133	120

Dataset describes the experiment and the corresponding data generated

No. of sequences refers to total number of reads processed

Base pair count refers to total number of bases read

Size (Gb) refers to storage requirement for all subreads

4.3 Characterization of a candidate L1 insertion

Following the successful recovery and assembly of minisatellite alleles *de novo* from PacBio's SMS data (see chapter 3), we attempted the reconstruction of the candidate L1 insertion on the X chromosome (contained within accession AC002980). This element is one of the most active known in cell culture retrotransposition assays (Seleme *et al.*, 2006) and has been re-sequenced extensively in studies of allelic diversity (Seleme *et al.*, 2006) and also in our laboratory (R. Badge *pers comm*). The element is a member of an actively expanding lineage and carries a 174bp transduction (process by which retrotransposons carry their flanking sequences during mobilization) sequence flanked by 2 poly A tails (Macfarlane *et al.,* 2013). This structure makes the insertion challenging to sequence, even with Sanger technology, as the poly A tails induce polymerase slippage.

4.4 The structure of L1 at AC002980 when sequenced using Sanger sequencing

In Figure 26, it is evident that Sanger sequencing chemistry poorly represents

poly A tails as the exact sequence between the two poly A tails cannot be deduced, indicated by N's.

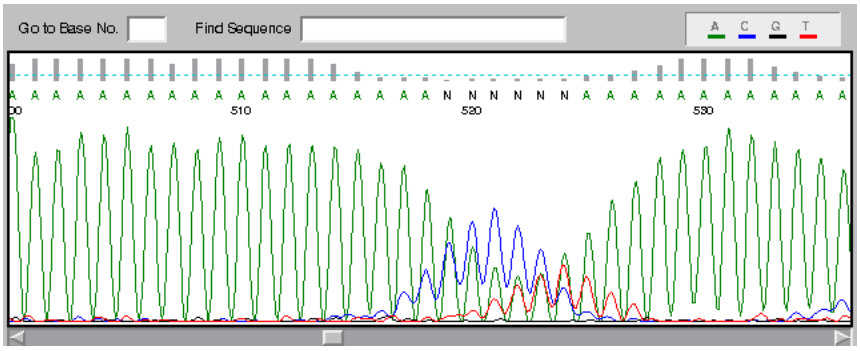

Figure 26: Chromatogram of Sanger sequencing poly A tails

The chromatogram as viewed in FinchTV, shows the result of Sanger sequencing the L1 located in accession AC002980. The region above represented as Ns indicate the region of transduction in an L1 which is usually flanked by polyA tails. Because Sanger sequencing breaks when it encounters these polyA tails, we get a bunch of Ns representing the region.

4.5 L1 recovery

In contrast to the sequence complexity of minisatellite repeat units, homopolymeric sequences such as L1s, while simple are similarly poorly represented in the human genome reference sequence. To reconstruct the sequence of the L1 in accession AC002980, we followed our earlier described process (see section 3.3.1 on MS1 recovery for details). Since the X chromosome is rich in repeat content (about 21%, 26% L1), we retrieved the Repeat Masker (Smit, Hubley and Green, 2013) track for chromosome X (on GRCh38) from the UCSC Table Browser for the purpose of identifying the locations of L1s within the reference. Using the repeat masker annotation, full-length L1s were extracted. The masked reference used for read mapping contained all full-length L1s as masked. The reason for masking all full-length L1's was to restrict read mapping to non-L1 regions. Mapping was carried out using BLASR and BWA-MEM (refer to page 48 for parameters used). These aligners were considered for the sole reason that BLASR – a PacBio proprietary mapper was designed for extraordinarily long reads which are dominated by indel errors. In contrast, BWA was designed for aligning low divergence reads against a large reference genome. Particularly, we adopted the BWA-MEM algorithm (with parameters "-x pacbio" and "-x ont2d") as it offered long-read support and split alignment. It is also highly memory efficient, faster and more

accurate than previous versions of BWA (http://bio-bwa.sourceforge.net/). It is hoped that by using different mappers, we would be able to evaluate the mapper that was best suited for our pipeline. The results of the mapping were stored in a binary form (sorted bam). Next was the extraction of reads aligning to a region 5kb flanking the ROI. This was done to enable the identification of reads that spanned the L1 (since all L1's had been masked). Identification of such reads meant they potentially carried sufficient information needed for the reconstruction of the L1 *de novo*. We performed pairwise alignment of all extracted reads against the ROI (hereafter, custom reference) to enable us place reads into groups based on a score distribution profile. Each collection of reads was then assembled. The assembly with highest pairwise alignment score to the custom reference was determined as the best reconstruction of the ROI. As with the constant improvements that come with long read lengths, BWA-MEM is now deprecated for the processing of PacBio and ONT long reads.

4.6 The effect of alignment programs on sequence alignments

Despite the recommendation to use BLASR for aligning PacBio reads (Chaisson and Tesler, 2012), we also know that BWA-MEM could potentially align long reads at a much faster rate (http://bio-bwa.sourceforge.net/), thus, our choice of both alignment programs. As illustrated in Figure 29, BLASR in contrast to BWA-MEM mapped a higher proportion of the CHM1 raw reads to the masked reference. This behaviour is expected as it accords with the design principle of BLASR. BLASR was designed with extraordinarily long reads in mind and to support long gapped alignments (Chaisson and Tesler, 2012). By contrast, BWA-MEM, though memory efficient for long reads, was originally designed for short reads and does not support long gapped alignments. A further consideration was the number of reads mapped by both programs, which was investigated by comparing the read identifiers of mapped reads from both programs (see Figure 30). BLASR uniquely maps 39% of reads whereas, BWA-MEM reports only 3% of uniquely mapped reads. Be that as it may, both algorithms mapped 58% of reads. These figures indicate that our ROI comprises extremely long reads for which BWA-MEM fails to produce significant alignments. BWA-MEM however, is now deprecated for long read mapping.

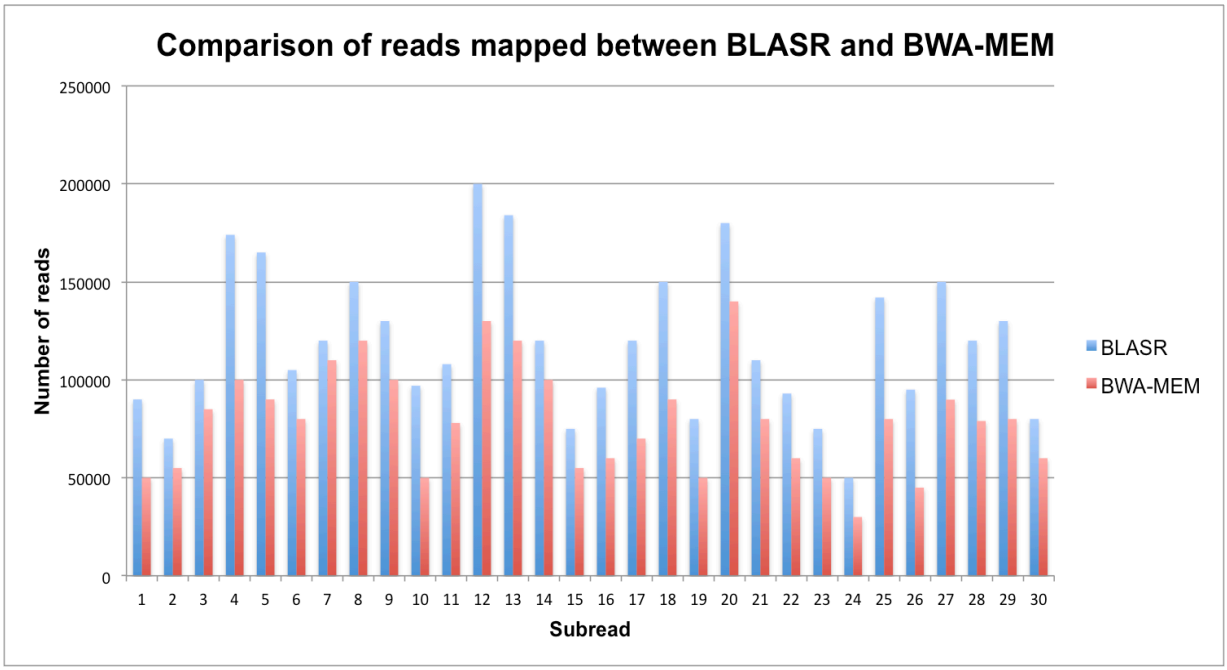

Figure 27: Comparison of reads mapped between BLASR and BWA-MEM aligners

Red bars indicate number of reads mapped using BLASR. Blue bars indicate number of reads mapped using BWA-MEM

4.7 The effect of low-scoring alignments on assembly

From the output of the pairwise alignment of each mapped read against the custom reference, we observed alignment scores resulting from weak, strong and very strong alignments. If these reads, having been reported to span the ROI, would align differently to the custom reference, perhaps, we could limit the number of reads passed to assembly. As illustrated in Figures 30 and 31, over 50% of reads known to have mapped to our ROI were excluded from the assembly process both by LAST and LALIGN since they fell within the lowest 2% of the score distribution – further indicating very poor alignments. If these reads were included in the assembly, contigs with many gaps would likely have been generated.

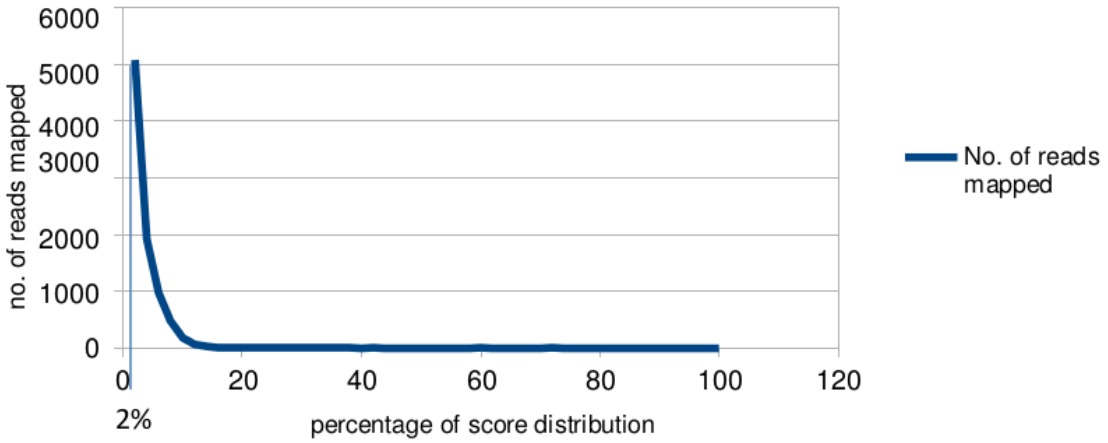

Figure 28: Score distribution of pairwise alignment of reads to regional reference using LAST

Thick blue curve indicate the number of reads mapped based on a score distribution of pairwise alignments using LAST

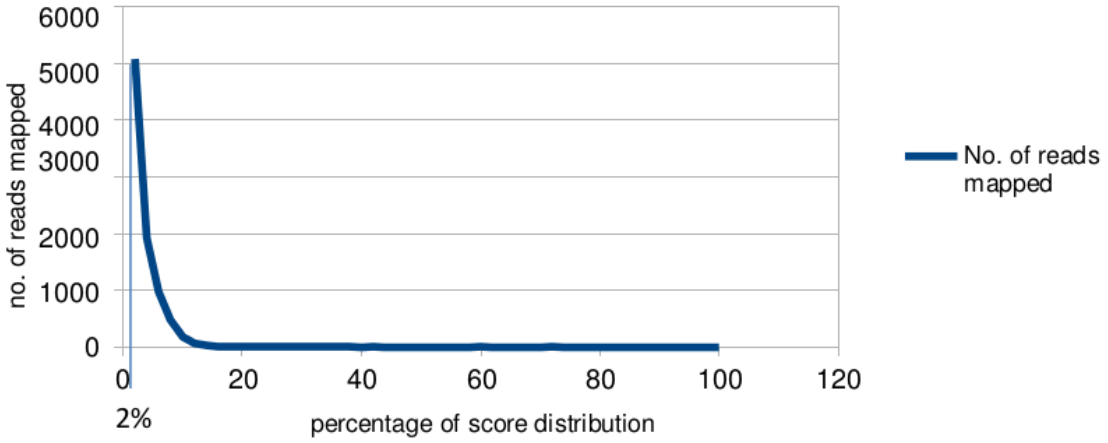

Figure 29: Score distribution of pairwise alignment of reads to regional reference using LALIGN

Thick blue curve indicates the number of reads mapped based on a score distribution of pairwise alignments using LALIGN

4.8 Percentage identity of assembly to reference

In order to recover the ROI, we aligned the assembly to the custom reference. The idea was to ensure that our assembly was as close as possible to the ROI. As illustrated in both Figure 30, the best contig as measured by percentage identity to the reference was derived from reads within score distribution profile 2 - the top 98% of the score distribution (excluding the 2% of weak alignments).

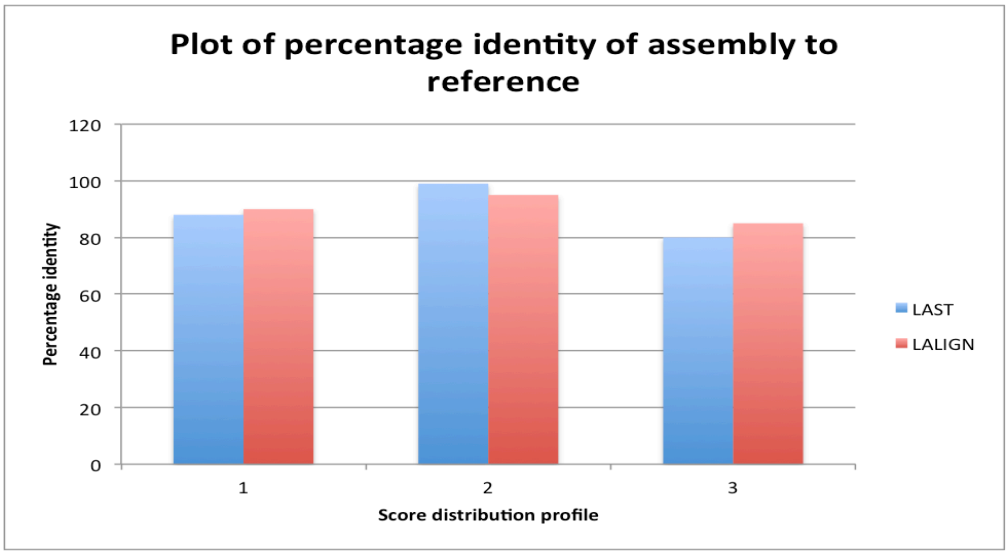

Figure 30: Percentage identity of assembly (based on score distribution) to reference (LAST)

Score distribution profile 1, 2 and 3 comprise reads whose pairwise alignment scores fall within the top 90%, 98% and 100% of the score distribution profile.

Red coloured bars indicate percentage identity for assemblies derived from reads obtained via LALIGN

Blue coloured bars indicate percentage identity for assemblies derived from reads obtained via LAST

4.9 The structure of the reconstructed AC002980 sequence

As illustrated in Figure 31, long stretches of identical sequences to the reference clearly point out the potential structure of the AC002980 L1. Even though few gaps exist, the quality of the assembled sequence is good.

Figure 31: Structure of reconstructed L1 sequence

Using Jalview to display the pairwise alignment of our reconstructed L1 and the reference sequence, the dark boxes indicate regions sequence match. These long stretches show across the region is a proof of assembly contiguity.

4.10 Recovery of novel L1 elements

As a further test of the utility of our pipeline, we explored the possibility of novel L1 insertions being present in other SMS datasets. Publicly available SMS data on the diploid individual (NA12878) from two separate technologies (PacBio and ONT) were evaluated for their content, and the ability of the pipeline to capture, novel (defined as being absent from the human genome reference sequence GRCh37) L1 elements. The algorithm for this analysis is shown in Figure 31.

The recovery process starts with running a BLAST search for the ROI (L19088, a reference human specific L1 element, known to be active in cell culture) on the GRCh37 reference assembly, PacBio and Nanopore's global genome (both polished and unpolished) assemblies (Figure 32a) for the NA12878 individual. Highly identical (>98%) and full-length (>6kb) insertions of our ROI were

extracted from both PacBio (433) and Nanopore (66) assemblies (Figure 32b). The reason for using a cutoff of 98% sequence identity was to ensure that only human-specific insertions were considered. Flanking DNA sequences (1kb on either end) for each full-length L1 were aligned (using blat) to the reference (Figure 32c). Hits spanning more than 2kb (389 from PacBio and 65 from Nanopore) meant the L1 insertion was present in the reference, whereas, hits with less than a 2kb (28 from PacBio and 1 from Nanopore) span meant the reference lacked the L1 insertion. These non-reference insertions were potentially novel. We then parsed these potentially novel insertions (28 from PacBio and 1 from Nanopore) through euL1db (Mir, Philippe and Cristofari, 2015) to further determine the novelty of these insertions, compared to a curated database of human specific L1 insertions, which includes insertions known to be polymorphic in humans (Figure 32d). The results confirmed the previous annotation of 13 PacBio-discovered and 1 Nanopore-discovered insertions. To test whether the remaining 15 potentially novel insertions from the PacBio assembly were real and not an artefact of assembly, we processed all insertions from the raw PacBio reads via our modified pipeline (see Figure 32e - g). Parsing the output from our pipeline through L1Base / L1Xplorer (Penzkofer, 2004) revealed, structural information for all insertions that was consistent with reports from PacBio's global genome assembly (Figure 32h).

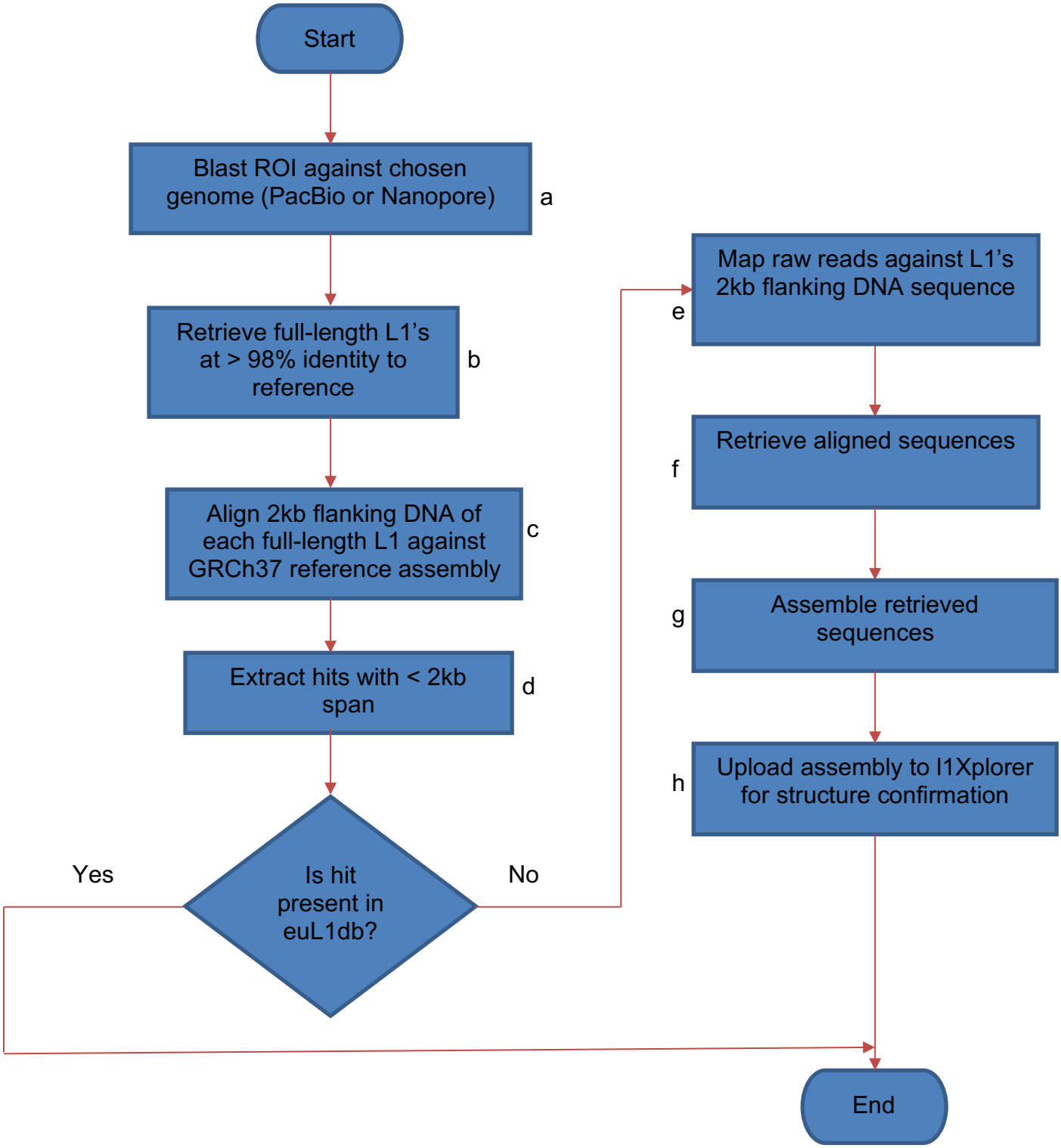

Figure 32: Assembly and recovery workflow for novel L1 insertions

Orange arrows indicate progression from one stage to another within the workflow. Each stage indicated by the blue object must be completed in order for the next stage to commence.

Table 12: PacBio's globally assembled L1's vs PacBio's locally assembled L1's

s/n	Global assembly statistics		Local assembly statistics		Region	% Identity
	Length (bp)	Intactness score	Length (bp)	Intactness score		
1	6052	24	6129	23	Chr11: 67598065-67600051	94.8
2	6030	25	6023	19	ChrX: 58569296-58569464	98.2
3	6042	25	6129	21	Chr5: 109479276-109481254	94.1
4	6034	23	6021	22	Chr5: 21206733-21208313	99.1
5	6051	23	5972	25	Chr5: 89449783-89451766	97.2
6	6040	25	6002	23	Chr3: 38625087-38627020	98.5
7	6050	25	6038	25	Chr3: 55787586-55789566	99.8
8	6050	25	6022	25	Chr3: 139043441-139045430	98.3
9	6042	24	6037	21	Chr3: 151147560-151149537	97.4
10	6050	23	6130	11	Chr3: 79215324-79217302	88.0
11	6046	22	6018	22	Chr10: 25706699-25708685	98.8
12	6050	24	5216	17	Chr1: 200882010-200883966	93.0
13	6050	25	6047	19	Chr1: 179574412-179576361	94.8
14	6039	23	6018	19	Chr6: 142449110-142451099	96.8
15	6057	25	6014	23	ChrX: 58569385-58569466	98.1

Note: all L1's were being assembled from PacBio's data

We compared our locally recovered L1s with its globally assembled equivalent. Length refers to the size of the L1. Intactness score is a measure of conservation of sequence features for an L1 element. This is derived from awarding one point to every conserved sequence feature that is known to affect the transcriptional and/or translational activity of the L1.

A total of 28 non-reference insertions were identified from the PacBio's global genome assembly. From these 28 new insertions, euL1db already contained 13, leaving a total of 15 potentially novel insertions. Processing all 15 insertions from the raw PacBio reads via the modified pipeline (Figure 32) revealed consistency (>90% sequence identity and intactness score difference of about 2) with reports from PacBio's global genome assembly.

Table 13: Locally recovered L1s from Nanopore's raw reads

s/n	Nanopore local assembly statistics		Region	% Identity
	Length (bp)	Intactness score		
1	5865	16	Chr11: 67598065-67600051	90.2
2	5800	19	ChrX: 58569296-58569464	94.3
3	5797	19	Chr5: 109479276-109481254	95.5
4	5832	19	Chr5: 21206733-21208313	96.1
5	5781	20	Chr5: 89449783-89451766	94.3
6	5982	16	Chr3: 38625087-38627020	86.7
7	5651	19	Chr3: 55787586-55789566	93.0
8	5671	19	Chr3: 139043441-139045430	90.7
9	5749	16	Chr3: 79215324-79217302	92.8
10	5571	19	Chr3: 151147560-151149537	90.0
11	5863	19	Chr10: 25706699-25708685	96.8
12	5734	18	Chr1: 200882010-200883966	93.6
13	5839	20	Chr1: 179574412-179576361	92.1
14	5674	19	Chr6: 142449110-142451099	94.6

Note: L1's were being assembled from Nanopore's raw reads based on flanking DNA from the 15 PacBio novel insertions. Length refers to the size of the L1 recovered. Intactness score is a measure of conservation of sequence features for an L1 element based on L1Xplorer analysis (Penzkofer, 2004). This is derived from awarding one point to every conserved sequence feature that is known to affect the transcriptional and/or translational activity of the L1. Here, only 14 L1s were recovered from Nanopore's data instead of all 15 candidates as seen in Table 12.

In the case of Nanopore, only one (1) non-reference insertion was found, and euL1db already contained it. Although, no novel insertion was found in the Nanopore's global genome assembly, yet, analysing the data through our pipeline revealed the existence of most of these potential novel insertions. We hypothesized that an alternate sequencing technology, applied to the same individual could serve as a further means of validating the novelty of these insertions.

Surprisingly the 14 NA12878 specific L1 insertions from the PacBio data were recovered from the Nanopore raw data, by our analysis (see Table 13). This suggests that the absence of these insertions from the global assembly reflects the assembly process, rather than any deficiency in the nanopore data.

4.11 Biological and evolutionary significance of the recovered L1s?

If, as our analyses suggested, the NA12878 specific L1 insertions recovered by PacBio (and nanopore) sequencing belonged to young active lineages, we would expect them to be phylogenetically closely related to known active elements. To test this hypothesis, a phylogenetic analysis on the recovered L1s compared to known and active full-length L1s was performed. With the help of Richard Badge, all 15 L1s (indicated as D_Xn in Figure 35) sequences together with the 42 L1s (extracted from Beck *et al.,* 2010) with intact ORFs and that showed variable activity in cell culture were aligned, relative to these sequences were aligned using ClustalW2 (Larkin et al., 2007) in a multi-sequence alignment. Our goal was to find the closest match if any, to a known element. Initial thoughts on origin of novel L1s based on L1Xplorer, suggested a close match to polymorphic Ta class elements (Penzkofer, 2004). We uploaded the output of the multiple sequence alignment generated by ClustalW2 to Phylemon (http://phylemon.bioinfo.cipf.es/index.html). These sequences were then loaded for analysis with PhyML-Best-AIC-tree – this python script runs PhyML under all possible nucleotide substitution models, calculates the Akaike Information Criteria (AIC) (Akaike, 1974) of each and picks the most informative parameter set. The script then re-runs with the model parameters estimated from data. PhyML calculates alternative log ratios for each node, to give a measure of support for the tree topology.

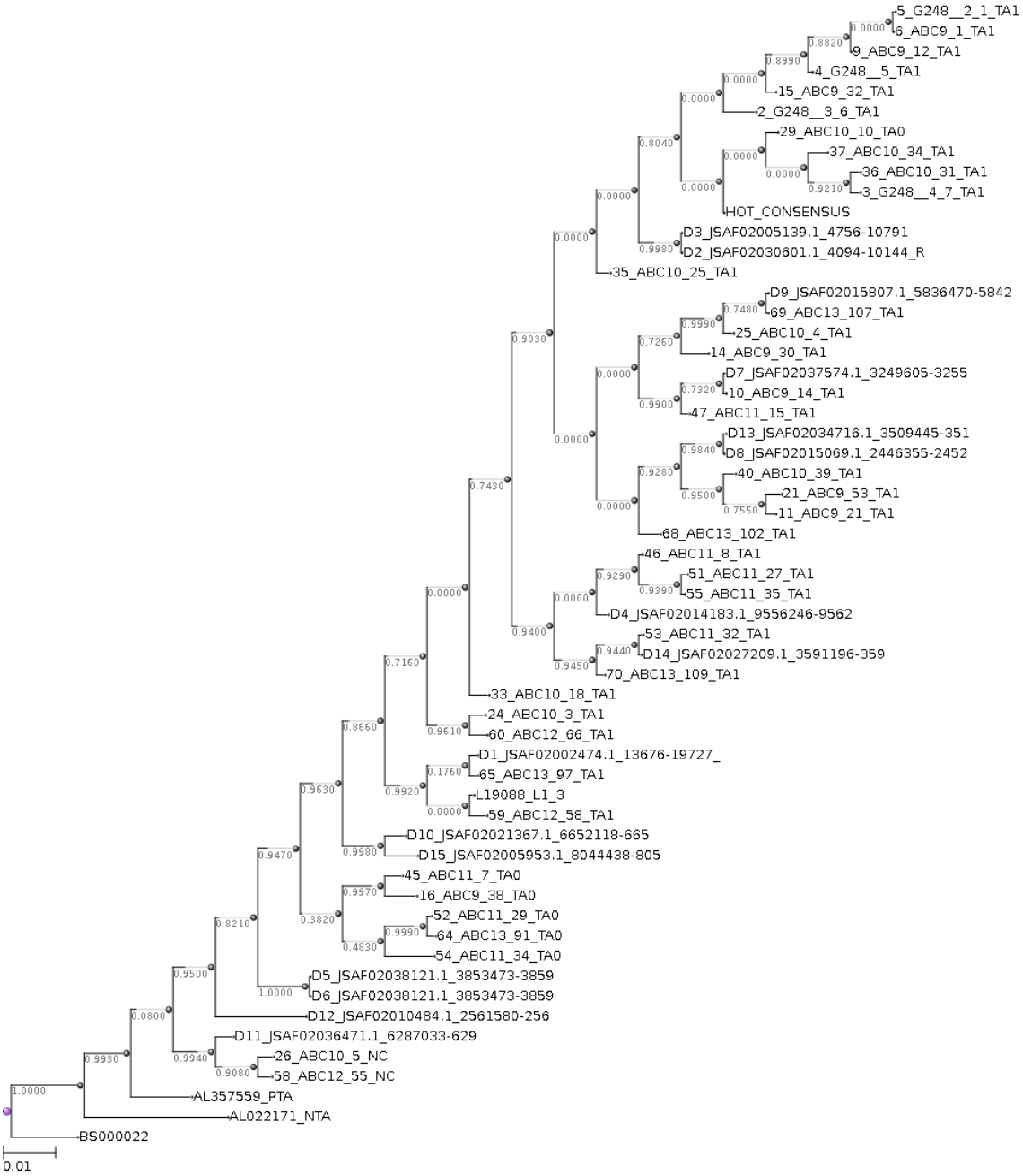

Figure 33: Phylogenetic tree of 15 novel L1s and 42 active L1s

Elements prefixed with D indicate recovered L1s while elements prefixed with numbers (1-42) indicate active L1s extracted from the phylogenetic tree in the Beck *et al*., 2010 paper.

The tree is rooted on a Chimpanzee specific full length L1 (BS000022). AL022171_NTA and AL357559_PTA are examples with intact ORFs from older subfamilies of L1HS elements, while “HOT CONSENSUS” is a synthetic consensus of the L1 elements most active in cell culture. Numbers at nodes are the log-likelihood ratios of the topology.

Notably, one element (D4_JSAF02014183.1_9556246-9562) clusters with high confidence with three closely related intact L1s, that are known to be highly active (79-191% of the activity of the L1.3 reference element) in cell culture. This clustering suggests that this element is also likely to be retrotransposition competent. In addition, this NA12878 specific L1 has been validated as a heterozygous insertion polymorphism, by locus specific PCR assays in NA12878 genomic DNA (R. Badge *pers comms*).

4.12 Discussion

Following the successful implementation and recovery of hard to sequence minisatellite alleles (see chapter 3) from our pipeline, we now discuss the application of an improved version of the pipeline to an example full-length L1 insertion, containing a particularly difficult to sequence region. Our results showed that the approach was effective, recovering the double poly A tail structure of the insertion and the intervening transduced sequence. Also, the pipeline enabled the discovery of new insertions of L1 elements that have not been reported in the human reference genome. This analysis shows that, in principle, this algorithm could be used for gap closure in reference sequences using high coverage long read data, as well as the characterization of repetitive sequences that are inaccurately represented in the human genome reference sequence.

4.12.1 The role of alignment programs on sequence alignments

The next issue we looked at was determining the best alignment program to use in mapping reads to our ROI. The type of alignment program used for sequence alignment obviously affects the quality of alignment and the number of reads aligned. In our analysis, we directly compared the proportion of reads mapped using BLASR and BWA-MEM. We can see that BLASR, in contrast to BWA-MEM, identifies a higher proportion of reads mapping to the reference (see Figure 27). This validates BLASR's support for long gapped alignments, as opposed to BWA-MEM, which was designed for short reads. Based on this, we used BLASR as the preferred mapping tool when acquiring reads for subsequent assembly. There are however, non-PacBio proprietary alignment programs such as LAST, BLAT, YASS (Noé and Kucherov, 2005), and LASTZ (Harris, 2007) which can also map query sequences to a reference. Large-

Scale Genome Alignment Tools (LASTZ) is another fast and efficient pairwise alignment tool. It performs pairwise alignment by using seeds to identify likely similarity regions, and then tries to extend them to local alignments. Unlike BLAT and YASS, LASTZ was designed specifically for aligning genomic sequences of millions of nucleotides in length. In this analysis, we chose LAST and BLAT over LASTZ and YASS as LAST is better able to deal with repetitive sequences while BLAT quickly finds regions in the genome likely homologous to the query sequence.

4.12.2 Low-scoring alignments and their effect on the assembly

We also looked at the quality of the assembly based on the input reads as we know from literature that the quality of an alignment is dependent on the percentage identity and the length of the alignment to the reference (Salmela and Rivals, 2014). However, we see from Figures 28 - 29, an assembly whose quality was affected by the presence of weak alignments (indicated as gaps in Figure 31) even when the reads passed to the assembly pipeline were known to map to our ROI. This could have been caused by the high error rates (>15%) of PacBio reads, thus blurring the signal in the alignment by the introduction of mismatches (Koren *et al*., 2012). To improve the assembly structure, we filtered reads passed to the assembly based on alignment score. A highly identical to the reference reconstructed contig was generated from reads in the top 98% (excluding the 2% of low-scoring alignments) of the pairwise alignment score distribution. This validates the use of higher identity reads in assembly.

4.12.3 Percentage identity of reconstructed contig to reference

In order to determine whether the reconstructed contig closely matched the reference, we compared the reconstructed sequences with the reference. The best contigs as seen in Figures 30 – 31 were derived from the top 98% of the score distribution for both LALIGN and LAST based assemblies. Using LALIGN, a contig with 95% identity to the reference and having a length of 20383bp was generated. This is likely to have been caused by the use of fixed seed lengths by the alignment program. On the other hand, LAST, which uses adaptive seeds for read mapping and allows for weaker alignments, gives us a contig with 99% identity to the reference and a length of 20453bp. This further shows the effect of long-spanning reads that though, displaying weak alignments, still

carry useful information that improves the quality of the reconstructed sequences.

4.12.4 Can long reads determine the structure of difficult to sequence regions?

The high cost of obtaining high-quality sequence contiguity of highly-repetitive regions of genomes and the present abandonment of Sanger sequencing for more modern NGS technologies for sequencing genomes, have left completion of most genomes in draft forms (Huddleston *et al*., 2014). This is due largely because short sequence read data are unable to scaffold across repetitive structures, thus leading to more gaps, missing data, and incomplete reference assemblies. Following from what we already know from the literature on the possibility of reconstructing sequences of complex regions using long reads (Huddleston *et al*., 2014), we decided to look at a full-length L1 insertion with a known difficult to sequence region within PacBio's single molecule long reads. As illustrated in Figure 26 and Figure 31, the sequence of the L1 which is poorly represented by Sanger sequencing experiments (as seen in Figure 26, page 81) was well Characterized when sequenced with single molecule long read technology. Our pipeline successfully determined the structure and sequence of the L1 with a 93% transduced (transfer of flanking sequence) sequence identity to reference (see Figure 31) which is consistent with the literature (Huddleston *et al*., 2014).

4.12.5 Does higher coverage mean anything?

The question above is an important one for discussion following the results of the structural analysis (see Tables 12 and Tables 13, pages 89-90) of the locally assembled L1s. Local assemblies from PacBio's 61.0X coverage datasets showed an average percentage identity >96% to L1's in the PacBio's global assembly. Also, the intactness scores – which is a function of the quality of the structure of the L1, closely matches those of the corresponding global assemblies. However, the intactness score for element 10 is much lower than the corresponding global assembly. Based on the structural analysis for element 10, we suspect the high number of frameshifts in open reading frames (ORF) 1 and 2 coupled with the many mutations that have occurred in sequence, might have been responsible for the low intactness score. Parsing a multi sequence alignment of element 10 (assembly) with reference sequence

shows a total of 139 gaps and 349 mutations. These structural variations further validate the low intactness score of the element.

4.12.6 Why do we observe incomplete recovery of L1s in the Nanopore's local assembly?

Having recovered 14 out of 15 new insertions in the PacBio data from its Nanopore equivalent, we ask the question – what is the problem with the last insertion at ChrX: 58569385-58569466 (GRCh37)? Is it possible that this missing element is contained within a gene? Does it have a long Poly A tail? Could the flanking sequence be very repetitive, or maybe it has a very high GC content? We sought to answer these questions by analysing the sequence and flanking sequences of the L1 from PacBio's global genome assembly. Despite the low repetitiveness in the flanks and the lack of a long Poly A tail in the L1 sequence (although this may not have been accurately reflected in the reference, or PacBio), the high GC (>60%) content in the flanks is weak evidence for gene enrichment. A further look on *Ensembl* (http://www.ensembl.org/) to see if the L1 was linked to a gene revealed several genes overlapping the L1. With the average percentage identity for local assemblies from Nanopore's 30X coverage datasets standing at <87%, we speculate that the low coverage and the inability of the technology to sequence through high GC content regions, might have affected the assembly and thus, contributed to the low intactness scores and the inability of our pipeline to recover all 15 insertions. An alternative test would be to perform PCR on the ROI of interest.

4.12.7 Is the PacBio global genome assembly reliable?

From the results of our analysis so far, we identified 15 new insertions of full-length L1 elements in the genome of the NA12878 individual. Even though competing SMS technologies like Oxford Nanopore have sequenced and released a genome assembly for the same individual, no novel insertions were found in their assembly of the same genome. Certainly, a higher coverage dataset with fewer errors will guarantee a better assembly. Be that as it may, local assembly via our pipeline revealed the presence of (though of less quality than what was obtained from PacBio's assembly) 14 out of 15 insertions within the raw data released by Oxford Nanopore technologies. We conclude that the sequencing strategy adopted by Nanopore is effective but question the

representation of these insertions in the global assembly. It is likely that these misrepresentations are a function of the assembly process. Also, the Nanopore datasets utilized were of lower coverage compared to PacBio.

With regard to the biological significance of these novel L1s, our phylogenetic tree (see Figure 35) shows how these elements spread throughout the tree with most around the root and within the clusters of active elements – suggesting that these elements are closely related to extant L1s, rather than assembly artefacts or chimeras.

5. Gap closure and recovery of missing elements

This chapter discusses our method and the results obtained via our pipeline in recovering missing reference elements *de novo.* By so doing; we potentially have computationally closed two (2) reference gap elements. Our analysis utilized the high coverage (61.0X) CHM1 dataset as released by PacBio (http://www.ncbi.nlm.nih.gov/Traces/sra/?study=SRP044331). For further details of this dataset, refer to Tables 3 and 11 on page 60 and 80 respectively. The scripts for the pipeline are accessible via our github repository (https://github.com/ndliberial/smrt_pipeline).

5.1 Gaps and the GRCh38 reference assembly

The lack of uniformity in sequence coverage, presence of repetitive sequences of varying length, copy number and sequence makes the genome assembly process difficult. This can lead to the introduction of gaps following the inability to correctly merge sequence reads in certain regions of the genome (Chaisson *et al*., 2015). It is worth mentioning that some gaps are as a result of polymorphisms. Even with the advancement in sequencing technologies and assembly algorithms, the latest build of the human reference genome (GRCh38) still contains 819 gaps.

We discuss briefly four (4) types of gaps as seen in Figure 34.

5.1.1 Sequence-coverage gaps: This kind of gap occurs in the absence of sampled sequence reads due to sequencing biases for a specific region in the genome (see Figure 35a). This usually generates dropouts in areas where assembled sequence is incomplete. Though very common with assemblies generated from low-coverage datasets (such as Sanger), with improved coverage, such gaps can be remedied.

5.1.2 Segmental duplication-associated gaps: These types of gaps (see Figure 35b) result from the existence of large tracts of duplicated DNA in the genome. Accordingly, over one third of the euchromatic gaps in the GRCh38 reference genome are flanked by large, highly identical segmental duplications.

5.1.3 Satellite associated gaps: These are gap regions dominated by short tandem repeats (STRs), macrosatellites and centromeric satellite repeats. These gaps exist as a result of the difficulty associated with assembling these

types of sequences because read overlaps are consistent with varying copy numbers of tandem repeats (see Figure 35c).

5.1.4 Muted gaps: These types of gaps (see Figure 35d) refer to regions in the genome that appear to have been closed, however, in a vast majority of individuals, we see additional or different sequences (Chaisson *et al*., 2015). This could be as a result of a very rare deletion variant in the individual whose genome was assembled, but mostly, it is usually due to assembly-based errors.

5.2 Why close gaps?

The presence of gaps makes it difficult to identify disease-causing mutations, thus perhaps compounding the problem of missing heritability (Manolio *et al*., 2009). Not only do gaps leave assemblies in draft forms, they also limit the reliability of such assemblies. The key to understanding genetic variation lies in the accurate assembly of genomes. This means with a more accurate reference genome, we can effectively align read data and interpret functional importance, leading to better annotation with less genotyping error genome-wide (1000 Genomes Project Consortium *et al.*, 2012; Li and Wren, 2014). With improvements in sequencing technology and assembly algorithms, efforts are constantly being made at resolving complex regions in the human reference assembly, thereby closing gaps.

Using SMS data, studies (Chaisson, Wilson and Eichler, 2015, Berlin *et al.,* 2015) have shown the possibility of gap closure based on the CHM1 data. It is in this direction, we considered the possibility of recovering any of the missing sequences in the GRCh38 reference assembly.

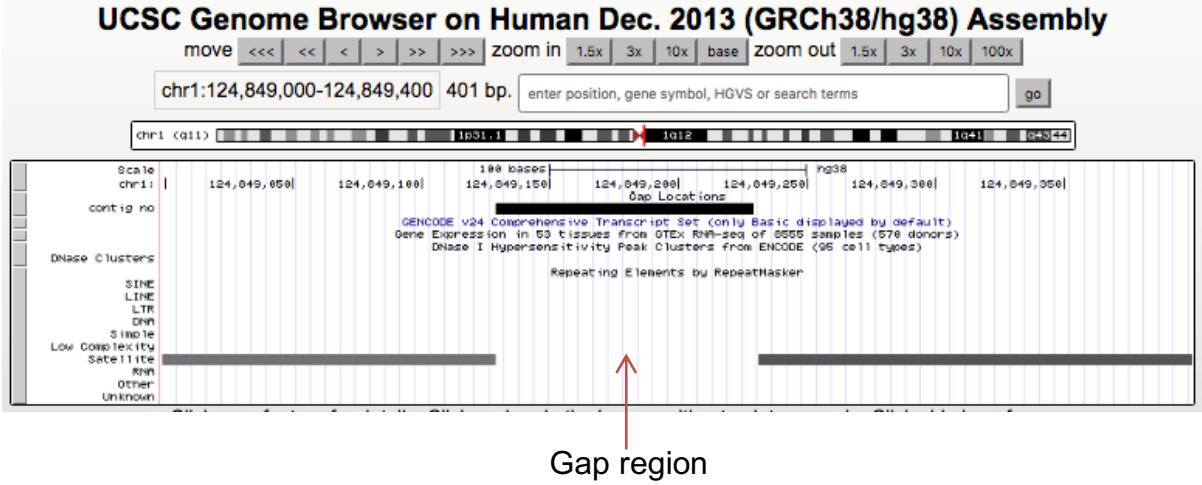

Figure 34: Screenshot of a gap region in GRCh38 reference assembly

The open region referred to by the red coloured arrow reflects a region within GRCh38 that is currently represented with Ns – a feature of most gap regions. Screenshot was generated from UCSC genome browser.

5.3 Recovering missing sequences in the GRCh38 reference assembly

Based on our previous pipelines (see chapters 3 and 4), we developed another pipeline (see Figure 36) for the *de novo* recovery of missing sequences (gaps) in the GRCh38 reference assembly. The pipeline starts by retrieving the summary list of genomic gaps from the NCBI genomes download page (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/ (Figure 36a). For all (819) gaps reported, 1kb of flanking DNA was extracted from both 3' and 5' ends of each gap region (Figure 36b). Using Repeat Masker, we masked out (using N's) all repeat sequences in the flanks (Figure 36c). The reason for this was to filter out gap regions with highly repetitive flanks, which would be challenging to assemble. Next, we selected gap regions (82 in total) with <=20% repetitive content in the flanks (Figure 36d) with the hope that this small amount of repetition would have little or no effect on the mapping and subsequent assembling of reads. Raw reads were mapped to the flanking DNA sequences of each selected gap region (Figure 36e). From the mapping results, all reads reported to have mapped to the flanks were retrieved for onward assembly (Figure 36f). Using Canu, we assembled each individual collection of reads per gap region (Figure 36g). Using blat, we then mapped the assembly back to flanking DNA (Figure 36h). Missing sequence

was recovered as the sequence positioned between highly (>90%) identical flanks (Figure 36i).

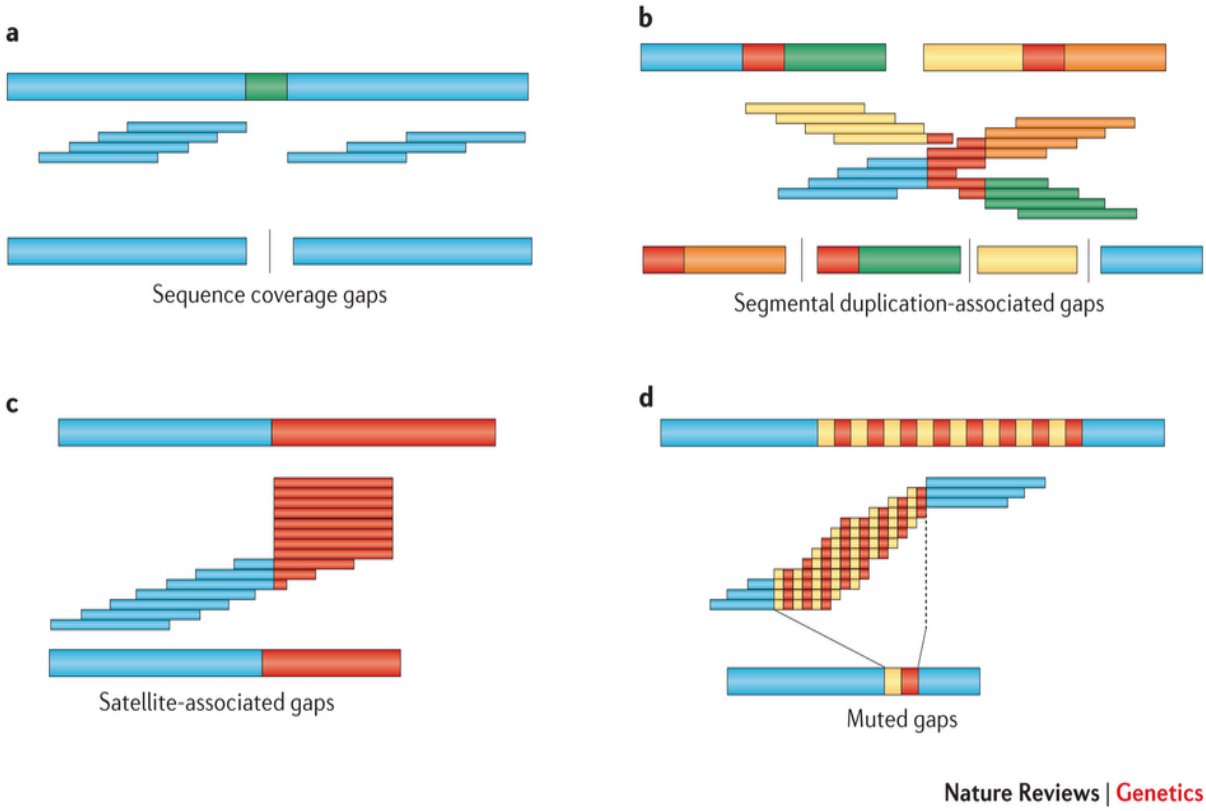

Figure 35: Types of genome assembly gaps

The genome architecture being resolved is shown at the top of each figure part as thick bars. Repetitive sequences are shown in red. Read overlaps are illustrated below the genome as thin bars (middle of each figure part), with regions overlapping repeats filled as red. The resulting assembly contigs are shown below (bottom of each figure part). Gaps are shown as vertical bars separating contigs to indicate unresolved sequences

Source: Chaisson *et al.,* 2015

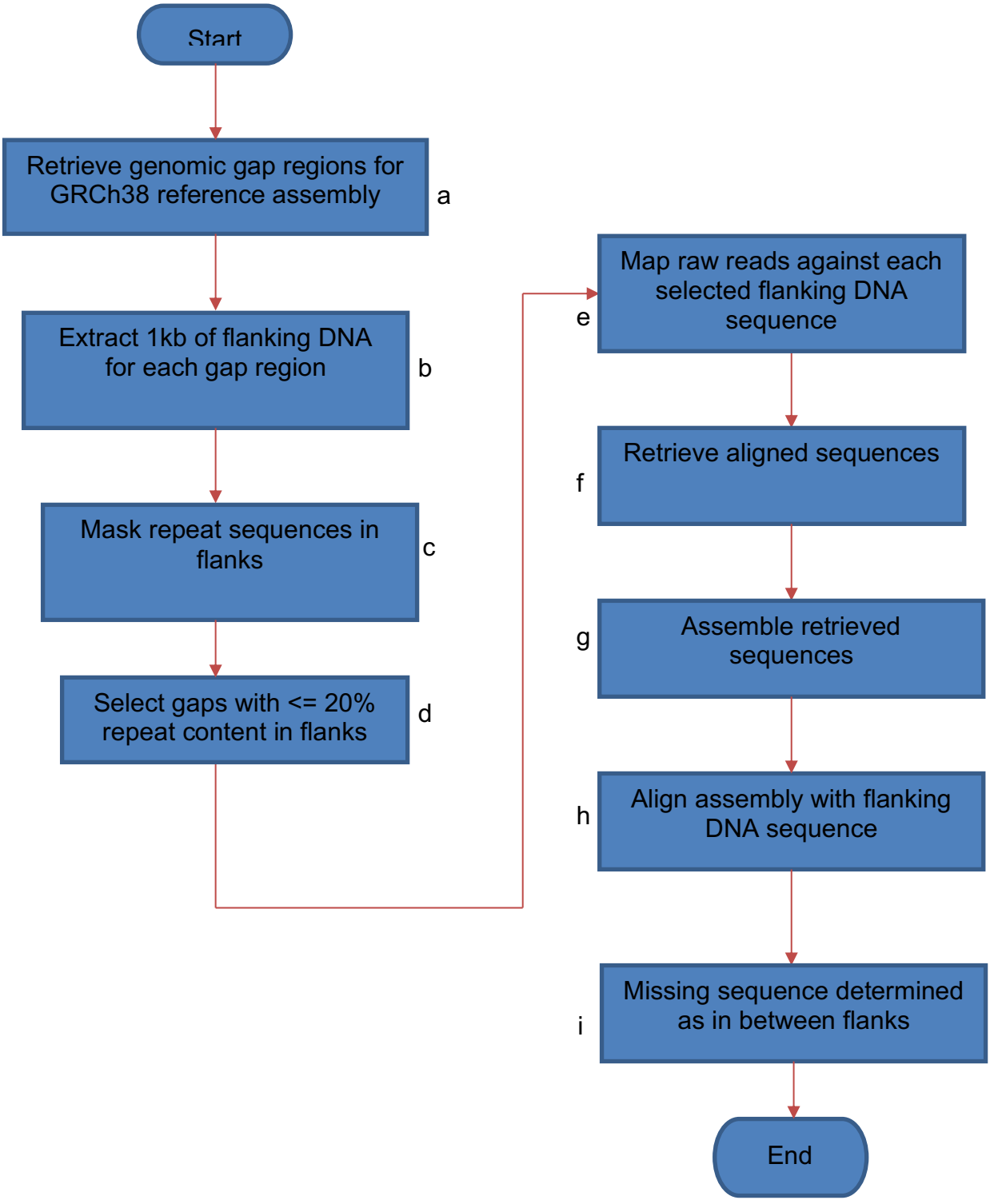

Figure 36: Assembly and recovery workflow for gap sequences
Orange arrows indicate progression from one stage to another within the workflow. Each stage indicated by the blue object must be completed in order for the next stage to commence.

Of the 82 gaps meeting our selection criteria, we compare the gap sequences present in PacBio's global assembly and our locally recovered sequences. We also report percentage identity scores where possible between corresponding gap sequences. All locally recovered gap sequences can be found in https://github.com/ndliberial/smrt_pipeline/sequences

Table 14: Locally vs. globally recovered gap sequences

S/n	Gap region	Estimated size	Globally recovered (bp)	Locally recovered (bp)	% pairwise identity
1	chr1:125171347-125173583	2236	No	Yes (1718)	NA
2	chr1:125184587-143184587	1800000	No	Yes (250)	NA
3	chr18:54536574-54537528	954	Yes (919)	Yes (919)	100
4	chr18:46969912-47019912	50000	Yes (13386)	Yes (1591)	100 over 1591 bp
5	chr18_GL383571v1_alt:44225-94225	50000	3' only, no match for 5'	Yes (1316)	NA
6	chr4:1429358-1434206	4848	Flanks overlapping	Yes (175)	NA

Gap region refers to the GRCh38 genomic coordinates with missing sequence

Estimated size refers to the length of missing sequence as reported in the assembly gap file

Globally recovered is the reported length of missing sequence present in PacBio's globally assembled genome for CHM1

Locally recovered is the length of the missing sequence reconstructed *de novo*

% pairwise Identity is the measure of similarity between the globally recovered gap sequence and the locally recovered gap sequence

Table 15: Summary statistics for some of the assembled gap sequences

S/n	Repeat types					Source Assembly
	SINES	LINES	LTR elements	DNA elements	Simple repeats	
1	0	1 (164 bp)	1 (127 bp)	0	1 (39 bp)	Local
2	0	0	0)	0	1 (199 bp)	Local
3	0	0	0	0	2 (129 bp)	PacBio/Local
4	5 (1463 bp)	0	0	8 (1950 bp)	10 (641 bp)	PacBio
4	1 (289 bp)	0	0	1 (55 bp)	2 (169 bp)	Local

Table data was derived from repeat masker and categorizes the repeat content of the reconstructed gap sequences. It reports the amount of repeat in each class of repeat element found, with the length of each repeat in bracket. The S/n column is equivalent to the S/n column in Table 14. The repeat composition is the same across all assemblies – no distinct repeat type was found. The assembly column identifies the source of the assembly.

In figure(s) 37-38, we see a pictorial representation of a global alignment of a segment of one of our locally recovered gap sequence with the flanks. The alignment shows matching flanking sequences and the recovered gap sequence as the region over Ns.

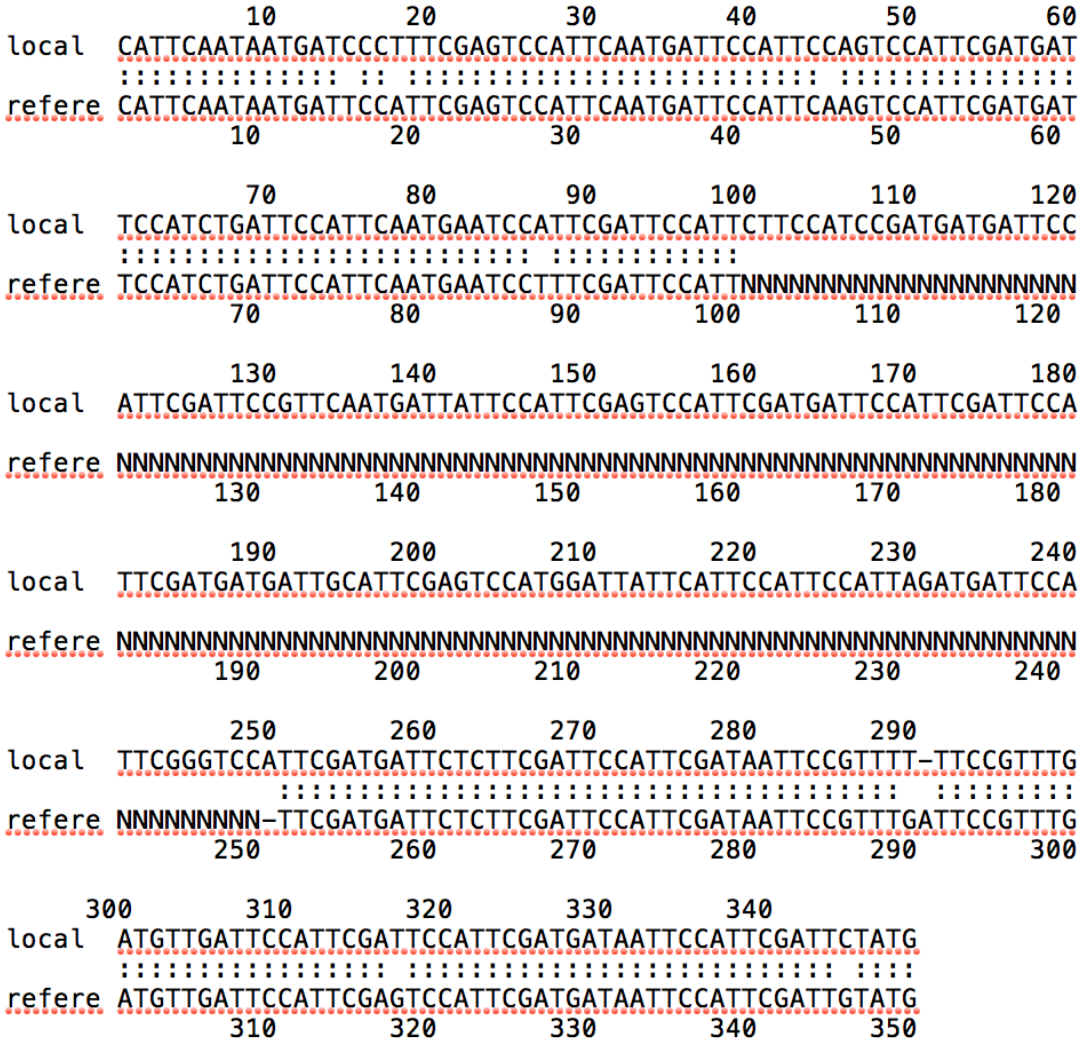

Figure 37: Alignment of a segment of the reconstructed gap sequence 1 and its corresponding reference (GRCh38) region.

The image above is derived from a pairwise alignment of the recovered gap sequence 1 plus 100 bp of flanking sequences with its corresponding reference sequence. Letters above Ns represent the locally reconstructed gap sequence.

A single white space character indicates a mismatch.

A colon (☺ indicates a match

A dash (–) indicates an indel

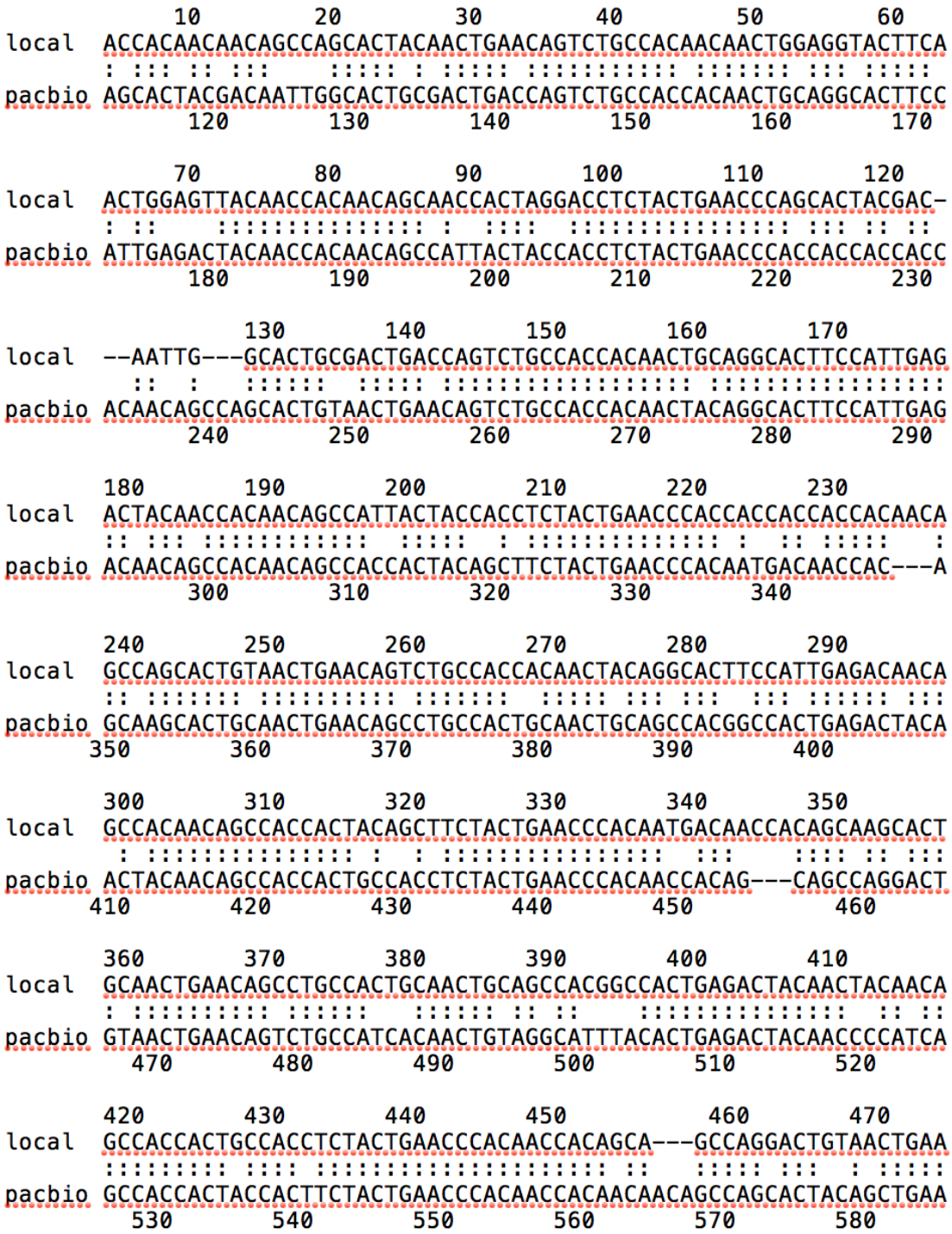

Figure 38: Pairwise alignment of a segment of our locally assembled gap sequence 3 and its PacBio's equivalent

A single white space indicates a sequence mismatch.

A colon (☺ indicates a sequence match

A dash (–) indicates an indel

5.4 Discussion

Based on evidence from the literature (Chaisson *et al*., 2015, Berlin *et al*., 2015) and our previously reported results (see chapters 3 and 4), we discuss the potential closure of six (6) gap elements in the GRCh38 reference assembly.

5.4.1 Recovery of novel sequences

As seen in Table 14 page 103, our pipeline in comparison to PacBio's global genome assembly for the same cell line, we achieved the reconstruction of two (2) novel sequences (sequence 1 and 2) which are completely absent in the PacBio's global assembly.

5.4.2 How reliable are the local assemblies?

For all 6 local assemblies (see Table 15, page 104) reported, only 1 (sequence 3) was fully assembled (i.e. its size agreed with the estimated size) in the PacBio's global genome assembly. Though sequence 4 was partially assembled, sequence 1 and 2 were missing, while sequence 5 and 6 appear to be fragmented in PacBio's global assembly. However, our local assembly shows partial recoveries for 5 (sequence 1, 2, 4, 5, and 6) gap elements and 1 full recovery (sequence 3). Mapping our locally recovered sequence 3 with PacBio's equivalent, revealed a 100% identity between sequences (see Figure 38). Compared to PacBio's, our local recovery for sequence 4 is 11,795 bp shorter (perhaps due to assembly errors), nevertheless, the sequences both share the same features – SINES, LINES, DNA elements, LTR elements and Simple repeats (see Table 15).

5.4.3 What is the biology of these local assemblies?

To gain some insight into the biology of these local recoveries, we provide a summary of 4 gap assemblies (see Table 15) showing the composition of these reconstructed gap sequences. We see an average GC level of 55% GC content and a content of over 35% of repeat sequences (simple repeats, LINEs, SINEs and DNA elements). The sequence composition of these partially recovered regions clearly typifies the basic Characteristic features of complex and difficult to sequence regions. We thus classify these regions as repetitive gaps. The inability to fully sequence through these regions has led to gaps in the genome assembly. It is not that reads covering the region are absent, we just cannot unambiguously map them to the gaps.

6. Discussion

6.1 Final comments

In this study, we have shown that by using single molecule sequencing and long read technology, poorly represented repetitive sequences (specifically, minisatellites and L1s) as well as gapped elements, can be better characterized by developing custom software, scalable for the analysis of single molecule long reads (particularly, PacBio's). As technology advances in the area of SMS, with alternative sequencing strategies (such as Oxford Nanopore) evolving, coupled with the frequent release of good quality and higher coverage reference datasets, it is envisaged that the robustness of these pipelines may make them useful to biologists and geneticists focused on local assemblies. This is particularly significant because of the level of novel discoveries that can be achieved from existing data. The MS1 allele and novel L1s recovered showed >95% identity to the globally recovered equivalent where aligned. Using phylogenetic analysis, we have shown that these sequences are likely real insertions as they are distributed throughout the phylogenetic tree, and show high similarity to known active elements.

The recovery of L1s from Nanopore data (combining both release3 and release4 datasets for NA12878), which hitherto, was missing in the global assembly, further, validates the software as a multi-platform analysis tool. In the absence of PCR, our pipeline can be used for *in-silico* confirmation of sequence existence as well as finding specific novel insertions in future assemblies.

With a pipeline as described in this thesis, insertions due to activation and translocation of genes in leukaemia and cancer could be identified, at least in principle. Using one of the tools for recovering gap elements, we have shown the potential of closing (computationally) two (2) gap regions. No doubt, further validation of the recovered gap sequences by PCR is required, but for the lack of time, we have only reported the computational findings. We hope that these findings would give non-specialist biologists and researchers with less computational resources an alternative and reliable assembly strategy.

6.2 Future directions

Briefly, I present a discussion of the importance of this study to scientific effort directed at understanding both the human and genomes of other model organisms.

6.2.1 Sequencing and assembling of centromeres: Because human centromeres are rich in dispersed repetitive DNA as well as classical satellite sequences, there is currently a lack of understanding of centromere organization and function in the reference genome (which was generated from short reads – mainly Sanger reads) as a result of the absence of a contiguous alpha satellite DNA sequence. Nevertheless, long-read sequencing has shown its potential in overcoming the challenges posed by short read sequencing (Chaisson *et al*., 2015). As the length of reads generated (up to 20 kb on average) spans many of the repetitive regions, which presumably is the cause of lack of contiguity in short read assemblies, we now have studies (Melters *et al*., 2013; Vembar *et al*., 2016) that have utilized PacBio long reads in discovering and mapping of centromeric tandem repeats in a number of organisms. Only recently, using a nanopore long read sequencing strategy, the first ever sequence assembly and characterization of the centromeric region of a human Y chromosome was achieved (Jain et al 2018b). Focusing on the haploid satellite array that spans the Y chromosome (DZY3), the study utilized a transposase-based method to generate high-read coverage of full-length BAC DNA with nanopore sequencing. That way, the circular BAC could be linearized with a single cut-site, followed by the addition of the necessary sequencing adaptors (Jain, Olsen *et al*., 2018). Further studies (Wolfgruber *et al*., 2016, Vanburren *et al*., 2015) on plant genomes have also produced more tractable assemblies of centromeric regions, though these are still challenging to validate. No doubt, the sequencing and assembling of centromeric regions has greatly improved with the latest sequencing technology, but more improvements are needed to generate highly contiguous and accurate assemblies for these complex regions.

6.2.2 Sequencing and assembling of telomeres: As with centromeres, efforts have been made to resolve complex telomeric regions in both human and other genomes. Such efforts involve the development of telomere specific PCR based methods, such as single telomere length analysis (STELA) (Hills *et al*., 2009) and telomere variant repeat mapping by PCR (TVR-PCR) (Baird *et al*., 2000). These methods were developed and implemented to Characterize repeat variant interspersion patterns, in place of Characterization by sanger

sequencing. Nevertheless, due to the complexity of these telomeric regions, they remain unresolved and / or poorly annotated in most reference assemblies. However, with PacBio's SMS, repetitive heterochromatic and telomeric transition sequences in human (Chaisson *et al.,* 2015), *D.melanogaster* (Kim *et al.*, 2014; Berlin *et al*., 2015), and *S.cerevisiae* (Berlin et al., 2015) have been resolved. The genome of *P.falciparum* has been difficult to estimate genetic diversity due to the skewed AT-richness (~80.6%) in the genome. Yet, only recently, using SMRT generated reads from *P.falciparum*, genomic DNA, all 14 chromosomes were resolved telomere-to-telomere (Vembar *et al*., 2016). In addition, ~90-99% of *P. falciparum* centromeric regions were completely resolved in the genome. Again, efforts at resolving telomere repeat lengths have recently become evident with the use of ultra-long reads (Jain, Koren *et al*., 2018). This study showed evidence for telomeric arrays spanning 2-11 kb within 14 subtelomeric regions for the NA12878 individual.

6.2.3 The challenge of nanopore sequencing and assembly: As revealed from our local assemblies derived from nanopore data, we conclude that the inaccurate representation of both L1 sequences and minisatellites in nanopore's (both polished and unpolished) global assembly reflects the need for an improvement in the assembly pipeline. We can see the presence of the repetitive DNA sequences within the read data, but, the assembly pipeline fails to reconstruct the regions accurately. It is possible the assembly could be improved with a higher coverage dataset, close to the levels of coverage achieved in PacBio's datasets. Also, it is apparent (from publicly available data) that using a hybrid assembly approach https://github.com/nanoporetech/ont-assembly-polish) can lead to better assembly results.

6.2.4 SMS and haplotype resolution: The challenge of sequencing and assembling diploid genomes persists, as a result of their varying levels of heterozygosity. This is because heterozygous genomes normally generate more fragmented assemblies than haploid or homozygous genomes of similar sizes. However, with advances in sequencing, SMS long-reads carry the required information to phase haplotypes over multiple kilobase distances (Minio *et al*., 2017). As SMS normally generates low accuracy reads (Continuous Long Reads (CLR)), newly developed software such as FALCON-unzip (Chin *et al*., 2016) can be used in assembling deployed genomes into

highly contiguous and correctly phased genomes (Minio *et al*., 2017). This was used in the assembly of the highly heterozygous diploid genome of the Cabernet Sauvignon grape variety (Chin *et al*., 2016). It may be possible to phase haplotypes by using higher quality (CCS) reads during assembly.

6.2.5 Limit of coverage needed to correct for random errors: Though the error rate of SMS technologies is high (~12-15%), we do know that these errors are generated randomly in CLR (Giordano *et al*., 2017). Thus, we can minimize the impact of these errors by generating CCS reads with sufficient passes. With a coverage of 15 passes, studies report that >99% accuracy can be achieved (Eid *et al*., 2009). This means that we can generate much more reliable local assemblies *de novo*.

Even with the challenges earlier on discussed with SMRT sequencing, we can at least say, the technology as pioneered by PacBio has improved greatly over the years. Longer (>40kb) read lengths, higher quality reads and sequencing depth have all contributed to the generation of better quality genome assemblies for both human and other model organisms. With continuous improvements in read lengths up to 800kb (Nanopore's ultra-long reads), it is promising to say that more and more better assemblies are likely in the future.

7. References

1000 Genomes Project Consortium, T. 1000 G. P. et al. (2012) 'An integrated map of genetic variation from 1,092 human genomes.', Nature, 491(7422), pp. 56–65. doi: 10.1038/nature11632.

Adams, M. D. (2000) 'The Genome Sequence of Drosophila melanogaster', Science, 287(5461), pp. 2185–2195. doi: 10.1126/science.287.5461.2185.

Ahmed, M. and Liang, P. (2012) 'Transposable elements are a significant contributor to tandem repeats in the human genome', Comparative and Functional Genomics, 2012. doi: 10.1155/2012/947089.

Akaike, H. (1974) 'A new look at the statistical model identification', IEEE Transactionson Automatic Control, 19(6), pp. 716–723. doi: 10.1109/TAC.1974.1100705.

Armour, J. A. L., Patel, I., et al. (1989) 'Analysis of somatic mutations at human minisatellite loci in tumors and cell lines', Genomics, 4(3), pp. 328–334. doi: 10.1016/0888-7543(89)90338-8.

Armour, J. A. L., Wong, Z., et al. (1989) 'Sequences flanking the repeat arrays of human minlsatellites: Association with tandem and dispersed repeat elements', Nucleic Acids Research, 17(13), pp. 4925–4936. doi: 10.1093/nar/17.13.4925.

Armour, J. A. L. et al. (1996) 'Minisatellite diversity supports a recent African origin for modern humans', Nature Genetics, 13(2), pp. 154–160. doi: 10.1038/ng0696-154.

Baird, D. M. et al. (2000) 'High Levels of Sequence Polymorphism and Linkage Disequilibrium at the Telomere of 12q: Implications for Telomere Biology and Human Evolution', The American Journal of Human Genetics, 66(1), pp. 235–250. doi: 10.1086/302721.

Bankier, A. T. et al. (1991) 'The DNA sequence of the human cytomegalovirus genome.', DNA sequence : the journal of DNA sequencing and mapping, 2(1), pp. 1–12. doi: 10.3109/10425179109008433.

Barba, M., Czosnek, H. and Hadidi, A. (2014) 'Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology', Viruses, 6(1), pp. 106–136. doi: 10.3390/v6010106.

References

Barrell, B. G. et al. (1980) ‘Sequence of Mammalian Mitochondrial DNA’, in Biological Chemistry of Organelle Formation. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 11–25. doi: 10.1007/978-3-642-81557-7_2.

Batzoglou, S. et al. (2002) ‘ARACHNE: A whole-genome shotgun assembler’, Genome Research, 12(1), pp. 177–189. doi: 10.1101/gr.208902.

Baudat, F. et al. (2010) ‘PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice’, Science, 327(5967), pp. 836–840. doi: 10.1126/science.1183439.

Beck, C. R. et al. (2011) ‘LINE-1 Elements in Structural Variation and Disease’, Annual Review of Genomics and Human Genetics, 12(1), pp. 187–215. doi: 10.1146/annurev-genom-082509-141802.

Benson, G. (1999) ‘Tandem repeats finder: A program to analyze DNA sequences’, Nucleic Acids Research, 27(2), pp. 573–580. doi: 10.1093/nar/27.2.573.

Berg, I. L. et al. (2010) ‘PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans’, Nature Genetics, 42(10), pp. 859–863. doi: 10.1038/ng.658.

Berg, I. L. et al. (2011) ‘Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations’, Proceedings of the National Academy of Sciences, 108(30), pp. 12378–12383. doi: 10.1073/pnas.1109531108.

Berlin, K. et al. (2015) ‘Assembling large genomes with single-molecule sequencing and locality-sensitive hashing’, Nature Biotechnology, 33(6), pp. 623–630. doi: 10.1038/nbt.3238.

Braslavsky, I. et al. (2003) ‘Sequence information can be obtained from single DNA molecules.’, Proceedings of the National Academy of Sciences of the United States of America, pp. 3960–4. doi: 10.1073/pnas.0230489100.

Brenner, S. et al. (2000) ‘Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays’, Nature Biotechnology, 18(6), pp. 630–634. doi: 10.1038/76469.

Broman, K. W. et al. (1998) ‘Comprehensive Human Genetic Maps: Individual and Sex-Specific Variation in Recombination’, The American Journal of Human Genetics, 63(3), pp. 861–869. doi: 10.1086/302011.

References

Carneiro, M. O. et al. (2012) ‘Pacific biosciences sequencing technology for genotyping and variation discovery in human data’, BMC Genomics, 13(1), p. 375. doi: 10.1186/1471-2164-13-375.

Chaisson, M. J. P. et al. (2015) ‘Resolving the complexity of the human genome using single-molecule sequencing’, Nature, 517(7536), pp. 608–611. doi: 10.1038/nature13907.

Chaisson, M. J. P., Wilson, R. K. and Eichler, E. E. (2015) ‘Genetic variation and the de novo assembly of human genomes’, Nature Reviews Genetics, 16(11), pp. 627–640. doi: 10.1038/nrg3933.

Chaisson, M. J. and Tesler, G. (2012) ‘Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory’, BMC Bioinformatics, 13(1). doi: 10.1186/1471-2105-13-238.

Chial, H. (2008) ‘DNA sequencing technologies key to the Human Genome Project’, Nature Education, 1(1), p. 219. Available at: http://www.nature.com/scitable/topicpage/DNA-Sequencing-Technologies-Key-to-the-Human-828?auTags=.

Chin, C.-S. et al. (2016) Phased Diploid Genome Assembly with Single Molecule Real-Time\nSequencing, bioRxiv. doi: 10.1101/056887.

Clark, J. B. and Kidwell, M. G. (1997) ‘A phylogenetic perspective on P transposable element evolution in Drosophila’, Proceedings of the National Academy of Sciences, 94(21), pp. 11428–11433. doi: 10.1073/pnas.94.21.11428.

Collins, F. S. et al. (2004) ‘Finishing the euchromatic sequence of the human genome’, Nature, 431(7011), pp. 931–945. doi: 10.1038/nature03001.

Cordaux, R. and Batzer, M. A. (2009) ‘The impact of retrotransposons on human genome evolution’, Nature Reviews Genetics, pp. 691–703. doi: 10.1038/nrg2640.

Dalloul, R. A. et al. (2010) ‘Multi-platform next-generation sequencing of the domestic Turkey (Meleagris gallopavo): Genome assembly and analysis’, PLoS Biology, 8(9). doi: 10.1371/journal.pbio.1000475.

Danna, K. and Nathans, D. (1971) ‘Specific cleavage of simian virus 40 DNA by restriction endonuclease of Hemophilus influenzae’, Proceedings of the National Academy of Sciences, 68(12), pp. 2913–2917. doi: 10.1073/pnas.68.12.2913.

Deininger, P. L. and Batzer, M. A. (2002) 'Mammalian retroelements', Genome Research, pp. 1455–1465. doi: 10.1101/gr.282402.

Dib, C. et al. (1996) 'A comprehensive genetic map of the human genome based on 5,264 microsatellites', Nature, 380(6570), pp. 152–154. doi: 10.1038/380152a0.

Dietrich, W. F. et al. (1996) 'A comprehensive genetic map of the mouse genome', Nature, 380(6570), pp. 149–152. doi: 10.1038/380149a0.

Donis-Keller, H. et al. (1987) 'A genetic linkage map of the human genome.', Cell, 51(2), pp. 319–37. doi: 10.1016/0092-8674(87)90158-9.

Doucet, A. J. et al. (2010) 'Characterization of LINE-1 ribonucleoprotein particles', PLoS Genetics, 6(10), pp. 1–19. doi: 10.1371/journal.pgen.1001150.

Eid, J. et al. (2009) 'Real-time DNA sequencing from single polymerase molecules', Science, 323(5910), pp. 133–138. doi: 10.1126/science.1162986.

English, A. C. et al. (2012) 'Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology', PLoS ONE, 7(11). doi: 10.1371/journal.pone.0047768.

Fatih Ozsolak (2013) 'Third Generation Sequencing Techniques and Applications to Drug Discovery', Expert Opin Drug Discov, 7(3), pp. 231–243. doi: 10.1517/17460441.2012.660145.Third.

Fleischmann, R. et al. (1995) 'Whole-genome random sequencing and assembly of Haemophilus influenzae Rd', Science, 269(5223), pp. 496–512. doi: 10.1126/science.7542800.

Frazer, K. A., Ballinger, D. G., et al. (2007) 'A second generation human haplotype map of over 3.1 million SNPs', Nature, 449(7164), pp. 851–861. doi: 10.1038/nature06258.

Frazer, K. A., Eskin, E., et al. (2007) 'A sequence-based variation map of 8.27 million SNPs in inbred mouse strains', Nature, 448(7157), pp. 1050–1053. doi: 10.1038/nature06067.

Gray, I. and Jeffreys, A. (1991). Evolutionary Transience of Hypervariable Minisatellites in Man and the Primates. Proceedings of the Royal Society B: Biological Sciences, 243(1308), pp.241-253.

'Genome sequence of the nematode C. elegans: A platform for investigating biology' (1998) Science, pp. 2012–2018. doi: 10.1126/science.282.5396.2012.

Gibbs, R. A. et al. (2004) 'Genome sequence of the Brown Norway rat yields insights into mammalian evolution', Nature, 428(6982), pp. 493–520. doi: 10.1038/nature02426.

Gilbert, W. and Maxam, A. (1973) 'The Nucleotide Sequence of the lac Operator', Proceedings of the National Academy of Sciences, 70(12), pp. 3581–3584. doi: 10.1073/pnas.70.12.3581.

Gill, P., Jeffreys, A. and Werrett, D. (1985). Forensic application of DNA 'fingerprints'. Nature, 318(6046), pp.577-579.

Giordano, F. et al. (2017) 'De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms', Scientific Reports, 7(1). doi: 10.1038/s41598-017-03996-z.

Goebel, S. J. et al. (1990) 'The complete DNA sequence of vaccinia virus', Virology, 179(1), pp. 247–266. doi: 10.1016/0042-6822(90)90294-2.

Goffeau, A. et al. (1996) 'Life with 6000 genes', Science, 274(5287), pp. 546–567. doi: 10.1126/science.274.5287.546.

Gregory, T. R. and Hebert, P. D. N. (1999) 'The modulation of DNA content: Proximate causes and ultimate consequences', Genome Research, pp. 317–324. doi: 10.1101/gr.9.4.317.

Gresham, D. et al. (2006) 'Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA Microarray', Science, 311(5769), pp. 1932–1936. doi: 10.1126/science.1123726.

Gresham, D., Dunham, M. J. and Botstein, D. (2008) 'Comparing whole genomes using DNA microarrays', Nat Rev Genet, 9(4), pp. 291–302. doi: nrg2335 [pii] 10.1038/nrg2335.

Griffiths, A. et al. (2007) 'Introduction to genetic analysis', Biofutur, April, p. 707. doi: 10.1016/S0294-3506(00)90098-X.

Gyapay, G. et al. (1994) 'The 1993-94 Généthon human genetic linkage map', Nature Genetics, 7(2 Spec No), pp. 246–339. doi: 10.1038/ng0694supp-246.

Harris, R. S. (2007) IMPROVED PAIRWISE ALIGNMENT OF GENOMIC DNA. Available at: https://etda.libraries.psu.edu/files/final_submissions/5299 (Accessed: 20 March 2018).

Havecker, E. R., Gao, X. and Voytas, D. F. (2004) 'The diversity of LTR retrotransposons', Genome Biology. doi: 10.1186/gb-2004-5-6-225.

Hayashi, K., Yoshida, K. and Matsui, Y. (2005) 'A histone H3 methyltransferase controls epigenetic events required for meiotic prophase', Nature, 438(7066), pp. 374–378. doi: 10.1038/nature04112.

Henson, J., Tischler, G. and Ning, Z. (2012) 'Next-generation sequencing and large genome assemblies', Pharmacogenomics, 13(8), pp. 901–915. doi: 10.2217/pgs.12.72.

Hills, M. et al. (2009) 'Probing the mitotic history and developmental stage of hematopoietic cells using single telomere length analysis (STELA)', Blood, 113(23), pp. 5765–5775. doi: 10.1182/blood-2009-01-198374.

Holley, R. W. et al. (1965) 'Structure of a Ribonucleic Acid', Science, 147(3664), pp. 1462–1465. doi: 10.1126/science.147.3664.1462.

Hoskins, R. A. et al. (2000) 'A BAC-based physical map of the major autosomes of Drosophila melanogaster', Science, 287(5461), pp. 2271–2274. doi: 10.1126/science.287.5461.2271.

Huang, X. and Madan, A. (1999) 'CAP3: A DNA sequence assembly program', Genome Research, 9(9), pp. 868–877. doi: 10.1101/gr.9.9.868.

Huang, X. and Miller, W. (1991) 'A time-efficient, linear-space local similarity algorithm', Advances in Applied Mathematics, 12(3), pp. 337–357. doi: 10.1016/0196-8858(91)90017-D.

Huang, X. and Yang, S.-P. (2005) 'Generating a genome assembly with PCAP.', Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.], Chapter 11, p. Unit11.3. doi: 10.1002/0471250953.bi1103s11.

Huddleston, J. et al. (2014) 'Reconstructing complex regions of genomes using long-read sequencing technology', Genome Research, 24(4), pp. 688–696. doi: 10.1101/gr.168450.113.

Hudson, T. J. et al. (1995) 'An STS-Based Map of the Human Genome', Science, 270(5244), pp. 1945–1954. doi: 10.1126/science.270.5244.1945.

Hutchison, C. A. (2007) 'DNA sequencing: Bench to bedside and beyond', Nucleic Acids Research, 35(18), pp. 6227–6237. doi: 10.1093/nar/gkm688.

Jacobs, F. M. J. et al. (2014) 'An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons', Nature, 516(7530), pp. 242–245. doi: 10.1038/nature13760.

Jain, M. et al. (2015) 'Improved data analysis for the MinION nanopore sequencer', Nature Methods, 12(4), pp. 351–356. doi: 10.1038/nmeth.3290.

Jain, M., Olsen, H. E., et al. (2018) 'Linear Assembly of a Human Y Centromere', bioRxiv, p. 170373. doi: 10.1101/170373.

Jain, M., Koren, S., et al. (2018) 'Nanopore sequencing and assembly of a human genome with ultra-long reads', Nature Biotechnology. doi: 10.1038/nbt.4060.

Jeffreys, A. (1987). Highly variable minisatellites and DNA fingerprints. Biochemical Society Transactions, 15(3), pp.309-317.

Jeffreys, A. J., Neumann, R. and Wilson, V. (1990) 'Repeat unit sequence variation in minisatellites: A novel source of DNA polymorphism for studying variation and mutation by single molecule analysis', Cell, 60(3), pp. 473–485. doi: 10.1016/0092-8674(90)90598-9.

Jeffreys, A. J. and Pena, S. D. J. (1993) 'Brief introduction to human DNA fingerprinting', in DNA Fingerprinting: State of the Science. Basel: Birkhäuser Basel, pp. 1–20. doi: 10.1007/978-3-0348-8583-6_1.

Jeffreys, A., Royle, N., Wilson, V. and Wong, Z. (1988). Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. Nature, 332(6161), pp.278-281.

Jeffreys, A. J., Wilson, V. and Thein, S. L. (1985) 'Individual-specific "fingerprints" of human DNA', Nature, 316(6023), pp. 76–79. doi: 10.1038/316076a0.

Ji, Y. et al. (2000) 'Structure of chromosomal duplicons and their role in mediating human genomic disorders', Genome Research, pp. 597–610. doi: 10.1101/gr.10.5.597.

Jorde, L. (2008). Encyclopedia of genetics, genomics, proteomics, and bioinformatics. Chichester: John Wiley & Sons Ltd

Ju, Y. S. et al. (2011) 'Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals', Nature Genetics, 43(8), pp. 745–752. doi: 10.1038/ng.872.

Kasianowicz, J. J. et al. (1996) 'Characterization of individual polynucleotide molecules using a membrane channel.', Proceedings of the National Academy of Sciences of the United States of America, 93(24), pp. 13770–3. doi: 10.1073/pnas.93.24.13770.

Kelly, T. J. and Smith, H. O. (1970) 'A restriction enzyme from Hemophilus influenzae II', Journal of Molecular Biology, 51(2), pp. 393–409. doi: 10.1016/0022-2836(70)90150-6.

Kersbergen, P. et al. (2009) 'Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans', BMC Genetics, 10, p. 69. doi: 10.1186/1471-2156-10-69.

Kiełbasa, S. M. et al. (2011) 'Adaptive seeds tame genomic sequence comparison', Genome Research, 21(3), pp. 487–493. doi: 10.1101/gr.113985.110.

Kim, K. E. et al. (2014) 'Long-read, whole-genome shotgun sequence data for five model organisms', Scientific Data, 1. doi: 10.1038/sdata.2014.45.

Koga, A. et al. (2000) 'Evidence for recent invasion of the medaka fish genome by the Tol2 transposable element', Genetics, 155(1), pp. 273–281.

Koren, S. et al. (2012) 'Hybrid error correction and de novo assembly of single-molecule sequencing reads', Nature Biotechnology, 30(7), pp. 693–700. doi: 10.1038/nbt.2280.

Koren, S., Walenz, B., Berlin, K., Miller, J., Bergman, N. and Phillippy, A. (2017). Canu: scalable and accurate long-read assembly via adaptivek-mer weighting and repeat separation. Genome Research, 27(5), pp.722-736.

Kruglyak, S. et al. (1998) 'Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations', Proceedings of the National Academy of Sciences, 95(18), pp. 10774–10778. doi: 10.1073/pnas.95.18.10774.

de la Bastide, M. and McCombie, W. R. (2007) 'Assembling Genomic DNA Sequences with PHRAP', in Current Protocols in Bioinformatics. doi: 10.1002/0471250953.bi1104s17.

Lampe, D. J. et al. (2003) 'Recent horizontal transfer of mellifera subfamily mariner transposons into insect lineages representing four different orders shows that selection acts only during horizontal transfer', Molecular Biology and Evolution, 20(4), pp. 554–562. doi: 10.1093/molbev/msg069.

Larkin, M. A. et al. (2007) 'Clustal W and Clustal X version 2.0', Bioinformatics, 23(21), pp. 2947–2948. doi: 10.1093/bioinformatics/btm404.

Levene, M. J. et al. (2003) 'Zero-mode waveguides for single-molecule analysis at high concentrations.', Science (New York, N.Y.), 299(5607), pp. 682–6. doi: 10.1126/science.1079700.

Levenshtein, V. I. (1966) 'Levenshtein', Soviet Physics Doklady, 10, pp. 707–710.

Li, H. et al. (2009) 'The Sequence Alignment/Map format and SAMtools', Bioinformatics, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Li, H. and Durbin, R. (2009) ‘Fast and accurate short read alignment with Burrows-Wheeler transform’, Bioinformatics, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.

Li, H. and Durbin, R. (2010) ‘Fast and accurate long-read alignment with Burrows-Wheeler transform’, Bioinformatics, 26(5), pp. 589–595. doi: 10.1093/bioinformatics/btp698.

Li, H. and Wren, J. (2014) ‘Toward better understanding of artifacts in variant calling from high-coverage samples’, Bioinformatics, pp. 2843–2851. doi: 10.1093/bioinformatics/btu356.

Li, R. et al. (2010) ‘The sequence and de novo assembly of the giant panda genome’, Nature, 463(7279), pp. 311–317. doi: 10.1038/nature08696.

Li, W.-H. (1997) Molecular evolution. Sinauer Associates.

Macfarlane, C. M. et al. (2013) ‘Transduction-Specific ATLAS Reveals a Cohort of Highly Active L1 Retrotransposons in Human Populations’, Human Mutation, 34(7), pp. 974–985. doi: 10.1002/humu.22327.

Maniatis, T. et al. (1974) ‘Sequence of a represser-binding site in the DNA of bacteriophage λ’, Nature, 250(5465), pp. 394–397. doi: 10.1038/250394a0.

Manolio, T. A. et al. (2009) ‘Finding the missing heritability of complex diseases’, Nature, pp. 747–753. doi: 10.1038/nature08494.

Margulies, M. et al. (2006) ‘Genome Sequencing in Open Microfabricated High Density Picoliter Reactors’, Nature biotechnology, 437(7057), pp. 376–380. doi: 10.1038/nature03959.Copyright.

Massung, R. F. et al. (1994) ‘Analysis of the complete genome of smallpox variola major virus strain Bangladesh-1975’, Virology, 201(2), pp. 215–240. doi: 10.1006/viro.1994.1288.

Maxam, A. and Gilbert, W. (1977). A new method for sequencing DNA. Proceedings of the National Academy of Sciences, 74(2), pp.560-564.

Mazzarella, R. and Schlessinger, D. (1998) ‘Pathological consequences of sequence duplications in the human genome’, Genome Research, pp. 1007–1021. doi: 10.1101/gr.8.10.1007.

Mccarthy, A. (2010) ‘Third generation DNA sequencing: Pacific biosciences’ single molecule real time technology’, Chemistry and Biology, 17(7), pp. 675–676. doi: 10.1016/j.chembiol.2010.07.004.

McCoy, R. C. et al. (2014) 'Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements', PLoS ONE, 9(9). doi: 10.1371/journal.pone.0106689.

Melters, D. P. et al. (2013) 'Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution', Genome Biology, 14(1). doi: 10.1186/gb-2013-14-1-r10.

Mikheyev, A. S. and Tin, M. M. Y. (2014) 'A first look at the Oxford Nanopore MinION sequencer', Molecular Ecology Resources, 14(6), pp. 1097–1102. doi: 10.1111/1755-0998.12324.

Minio, A. et al. (2017) 'How Single Molecule Real-Time Sequencing and Haplotype Phasing Have Enabled Reference-Grade Diploid Genome Assembly of Wine Grapes', Frontiers in Plant Science, 8. doi: 10.3389/fpls.2017.00826.

Mir, A. A., Philippe, C. and Cristofari, G. (2015) 'euL1db: The European database of L1HS retrotransposon insertions in humans', Nucleic Acids Research, 43(D1), pp. D43–D47. doi: 10.1093/nar/gku1043.

Mullikin, J. (2002). The Phusion Assembler. Genome Research, 13(1), pp.81-90.

Myers, E. W. et al. (2000) 'A whole-genome assembly of Drosophila', Science, pp. 2196–2204. doi: 10.1126/science.287.5461.2196.

Nakamura, Y. et al. (1987) 'Variable number of tandem repeat (VNTR) markers for human gene mapping.', Science (New York, N.Y.), 235(4796), pp. 1616–22. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3029872.

Noé, L. and Kucherov, G. (2005) 'YASS: Enhancing the sensitivity of DNA similarity search', Nucleic Acids Research, 33(SUPPL. 2). doi: 10.1093/nar/gki478.

Nusbaum, C. et al. (1999) 'A YAC-based physical map of the mouse genome', Nature Genetics, 22(4), pp. 388–393. doi: 10.1038/11967.

Ogeh, D. and Badge, R. (2017) 'A pipeline for local assembly of minisatellite alleles from single-molecule sequencing data', Bioinformatics, 33(5). doi: 10.1093/bioinformatics/btw687.

Ohshima, K. and Okada, N. (2005) 'SINEs and LINEs: Symbionts of eukaryotic genomes with a common tail', Cytogenetic and Genome Research, pp. 475–490. doi: 10.1159/000084981.

Ohyama, K. (1996) 'Chloroplast and mitochondrial genomes from a liverwort, Marchantia polymorpha--gene organization and molecular evolution.',

Bioscience, biotechnology, and biochemistry, 60(1), pp. 16–24. doi: 10.1271/bbb.60.16.

Okada, N. et al. (1997) 'SINEs and LINEs share common 3' sequences: A review', in Gene, pp. 229–243. doi: 10.1016/S0378-1119(97)00409-5.

Oliver, S. G. et al. (1992) 'The complete DNA sequence of yeast chromosome III', Nature, 357(6373), pp. 38–46. doi: 10.1038/357038a0.

Orr, H. T. and Zoghbi, H. Y. (2007) 'Trinucleotide Repeat Disorders - annurev.neuro.29.051605.113042', Annual Review of Neuroscience, 30, pp. 575–623. doi: 10.1146/annurev.neuro.29.051605.113042.

Pandey, V., Nutter, R. C. and Prediger, E. (2008) 'Applied Biosystems SOLiD™ System: Ligation-Based Sequencing', in Next Generation Genome Sequencing: Towards Personalized Medicine, pp. 29–42. doi: 10.1002/9783527625130.ch3.

Parvanov, E. D., Petkov, P. M. and Paigen, K. (2010) 'Prdm9 controls activation of mammalian recombination hotspots', Science, p. 835. doi: 10.1126/science.1181495.

Patil, N. et al. (2001) 'Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21', Science, 294(5547), pp. 1719–1723. doi: 10.1126/science.1065573.

Pendleton, M. et al. (2015) 'Assembly and diploid architecture of an individual human genome via single-molecule technologies', Nature Methods, 12(8), pp. 780–786. doi: 10.1038/nmeth.3454.

Peng, Z. et al. (2016) 'Long read and single molecule DNA sequencing simplifies genome assembly and TAL effector gene analysis of Xanthomonas translucens', BMC Genomics, 17(1). doi: 10.1186/s12864-015-2348-9.

Penzkofer, T., Dandekar, T. and Zemojtel, T. (2005) 'L1Base: From functional annotation to prediction of active LINE-1 elements', Nucleic Acids Research, 33(DATABASE ISS.). doi: 10.1093/nar/gki044.

Poliseno, L. et al. (2010) 'A coding-independent function of gene and pseudogene mRNAs regulates tumour biology', Nature, 465(7301), pp. 1033–1038. doi: 10.1038/nature09144.

Prak, E. T. and Kazazian, H. H. (2000) 'Mobile elements and the human genome.', Nature reviews. Genetics, 1(2), pp. 134–44. doi: 10.1038/35038572.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: A flexible suite of utilities for comparing genomic features', Bioinformatics, 26(6), pp. 841–842. doi: 10.1093/bioinformatics/btq033.

Rangwala, S. H., Zhang, L. and Kazazian, H. H. (2009) 'Many LINE1 elements contribute to the transcriptome of human somatic cells', Genome Biology, 10(9). doi: 10.1186/gb-2009-10-9-r100.

Robinson, J. T. et al. (2011) 'Integrative genomics viewer', Nature Biotechnology, pp. 24–26. doi: 10.1038/nbt.1754.

Royle, N. J. et al. (1992) 'A hypervariable locus D16S309 located at the distal end of 16p', Nucleic Acids Research, p. 1164. doi: 10.1093/nar/20.5.1164.

Salmela, L. and Rivals, E. (2014) 'LoRDEC: Accurate and efficient long read error correction', Bioinformatics, 30(24), pp. 3506–3514. doi: 10.1093/bioinformatics/btu538.

Salzberg, S. L. et al. (2012) 'GAGE: A critical evaluation of genome assemblies and assembly algorithms', Genome Research, 22(3), pp. 557–567. doi: 10.1101/gr.131383.111.

Sanger, F. et al. (1977) 'Nucleotide sequence of bacteriophage ΦX147 DNA.', Nature, 265, pp. 687–695. doi: 10.1038/276236a0.

Sanger, F. et al. (1982) 'Nucleotide sequence of bacteriophage lambda DNA.', Journal of molecular biology, pp. 729–773. doi: 10.1016/0022-2836(82)90546-0.

Sanger, F. and Coulson, A. R. (1975) 'A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase', Journal of Molecular Biology, 94(3). doi: 10.1016/0022-2836(75)90213-2.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors', Proceedings of the National Academy of Sciences, 74(12), pp. 5463–5467. doi: 10.1073/pnas.74.12.5463.

Santangelo, A. M. et al. (2007) 'Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene', PLoS Genetics, 3(10), pp. 1813–1826. doi: 10.1371/journal.pgen.0030166.

Schatz, M. C., Delcher, A. L. and Salzberg, S. L. (2010) 'Assembly of large genomes using second-generation sequencing', Genome Research, pp. 1165–1173. doi: 10.1101/gr.101360.109.

Schuster, S. C. et al. (2010) ‘Complete Khoisan and Bantu genomes from southern Africa’, Nature, 463(7283), pp. 943–947. doi: 10.1038/nature08795.

Seleme, M. d. C. et al. (2006) ‘Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity’, Proceedings of the National Academy of Sciences, 103(17), pp. 6611–6616. doi: 10.1073/pnas.0601324103.

Shendure, J. and Ji, H. (2008) ‘Next-generation DNA se1. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26(10):1135–45. quencing.’, Nature biotechnology, 26(10), pp. 1135–1145. doi: 10.1038/nbt1486.

Smit, A., Hubley, R. and Green, P. (2013) ‘RepeatMasker Open-4.0. 2013-2015 .’, http://www.repeatmasker.org. Available at: http://repeatmasker.org.

Smith, H. O. and Wilcox, K. W. (1970) ‘A Restriction enzyme from Hemophilus influenzae’, Journal of Molecular Biology, 51(2), pp. 379–391. doi: 10.1016/0022-2836(70)90149-X.

Song, J. et al. (2001) ‘Instability of bacterial artificial chromosome (BAC) clones containing tandemly repeated DNA sequences’, Genome, 44(3), pp. 463–469. doi: 10.1139/gen-44-3-463.

Spurr, N. K. et al. (1994) ‘European Gene Mapping Project (EUROGEM): genetic maps based on the CEPH reference families.’, European journal of human genetics : EJHG, 2(3), pp. 193–203. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7834280 (Accessed: 20 March 2018).

Stajich, J. E. et al. (2002) ‘The Bioperl toolkit: Perl modules for the life sciences’, Genome Research, 12(10), pp. 1611–1618. doi: 10.1101/gr.361602.

Steinmetz, M., Uematsu, Y. and Lindahl, K. (1987). Hotspots of homologous recombination in mammalian genomes. Trends in Genetics, 3, pp.7-10

Tamaki, K. et al. (1995) ‘Applications of minisatellite variant repeat (MVR) mapping for maternal identification from remains of an infant and placenta.’, Journal of forensic sciences, 40(4), pp. 695–700. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7595313.

Teague, B. et al. (2010) ‘High-resolution human genome structure by single-molecule analysis.’, Proc. Natl. Acad. Sci. (U. S. A.), 107(24), pp. 10848–53. doi: 10.1073/pnas.0914638107.

The Arabidopsis Genome Initiative (2000) ‘Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.’, Nature, 408(6814), pp. 796–815. doi: 10.1038/35048692.

‘The yeast genome directory’ (1997) Nature. Macmillan Magazines Ltd., 389, p. 412. Available at: http://dx.doi.org/10.1038/38784.

Thomas, C. A. (1971) ‘The Genetic Organization of Chromosomes’, Annual Review of Genetics, 5(1), pp. 237–256. doi: 10.1146/annurev.ge.05.120171.001321.

Thompson, J. F. and Milos, P. M. (2011) ‘The properties and applications of single-molecule DNA sequencing’, Genome Biology. doi: 10.1186/gb-2011-12-2-217.

Thompson, J. F. and Steinmann, K. E. (2010) ‘Single molecule sequencing with a HeliScope genetic analysis system’, Current Protocols in Molecular Biology. doi: 10.1002/0471142727.mb0710s92.

Thorvaldsdóttir, H., Robinson, J. T. and Mesirov, J. P. (2013) ‘Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration’, Briefings in Bioinformatics, 14(2), pp. 178–192. doi: 10.1093/bib/bbs017.

Tóth, G. et al. (2000) ‘Microsatellites in different eukaryotic genomes: survey and analysis.’, Genome research, 10(7), pp. 967–981. doi: 10.1101/gr.10.7.967.

Treangen, T. J. and Salzberg, S. L. (2012) ‘Repetitive DNA and next-generation sequencing: Computational challenges and solutions’, Nature Reviews Genetics, 13(1), pp. 36–46. doi: 10.1038/nrg3117.

Turner, D. J. et al. (2009) ‘Next-generation sequencing of vertebrate experimental organisms’, Mammalian Genome, pp. 327–338. doi: 10.1007/s00335-009-9187-4.

Vanburen, R. et al. (2015) ‘Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum’, Nature, 527(7579), pp. 508–511. doi: 10.1038/nature15714.

Vanin, E. F. (1984) ‘Processed pseudogenes. Characteristics and evolution’, BBA - Gene Structure and Expression, pp. 231–241. doi: 10.1016/0167-4781(84)90057-5.

Vembar, S. S. et al. (2016) ‘Complete telomere-to-telomere de novo assembly of the Plasmodium falciparum genome through long-read (>11 kb), single molecule, real-time sequencing.’, DNA research : an international journal for

rapid publication of reports on genes and genomes, 23(4), pp. 339–51. doi: 10.1093/dnares/dsw022.

Vergnaud, G. and Denoeud, F. (2000) 'Minisatellites: Mutability and genome architecture', Genome Research, 10(7), pp. 899–907. doi: 10.1101/gr.10.7.899.

Wang, J. et al. (2008) 'The diploid genome sequence of an Asian individual', Nature, 456(7218), pp. 60–65. doi: 10.1038/nature07484.

Waterhouse, A. M. et al. (2009) 'Jalview Version 2-A multiple sequence alignment editor and analysis workbench', Bioinformatics, 25(9), pp. 1189–1191. doi: 10.1093/bioinformatics/btp033.

Watson, J. and Crick, F. (1953) 'Molecular structure of nucleic acids', Nature, 171, pp. 737–738. doi: 10.1038/171737a0.

Wolfgruber, T. K. et al. (2016) 'High Quality Maize Centromere 10 Sequence Reveals Evidence of Frequent Recombination Events', Frontiers in Plant Science, 7. doi: 10.3389/fpls.2016.00308.

Wong, Z., Wilson, V., Patel, I., Povey, S. And Jeffreys, A. (1987). Characterization of a panel of highly variable minisatellites cloned from human DNA. Annals of Human Genetics, 51(4), pp.269-288.

Yadav, V. P. et al. (2009) 'Characterization of the restriction enzyme-like endonuclease encoded by the Entamoeba histolytica non-long terminal repeat retrotransposon EhLINE1', FEBS Journal, 276(23), pp. 7070–7082. doi: 10.1111/j.1742-4658.2009.07419.x.

Zhang, L. et al. (2005) 'Patterns of segmental duplication in the human genome', Molecular Biology and Evolution, 22(1), pp. 135–141. doi: 10.1093/molbev/msh262.