Sequence Alignment with BWA

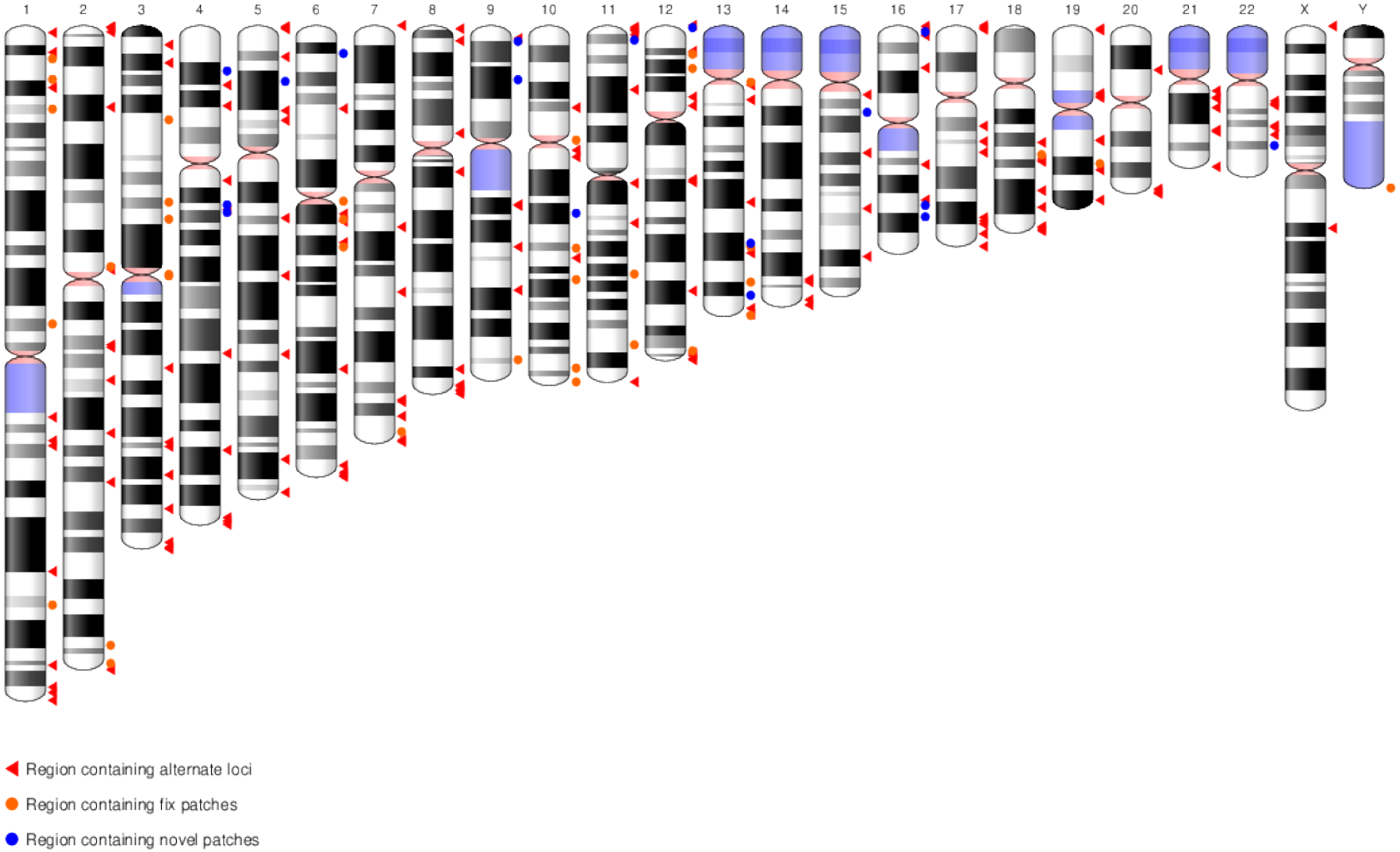
Shamith Samarajiwa
Integrative Systems Biomedicine Group
MRC Cancer Unit
University of Cambridge

27th July 2014, CRUK Bioinformatics Summer School
CRUK Cambridge Institute

Reference Genomes

- A haploid representation of a species genome.
- The human genome is a haploid mosaic derived from 13 volunteer donors from Buffalo, NY.
- For regions where there is known large scale variation, sets of alternate loci (178 in GRCh38) are assembled alongside the reference locus.
- The current build has around 500 gaps, whereas the first version had ~150,000 gaps

GRCh 38


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
Region containing alternate loci
Region containing fix patches
Region containing novel patches

Genome Reference Consortium

- http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/
- The original model for representing the genome assemblies was to use a single, preferred tiling path to produce a single consensus representation of the genome.
- Subsequent analysis has shown that for most mammalian genomes a single tiling path is insufficient to represent a genome in regions with complex allelic diversity.
- GRC routinely releases patches and corrections.

- GRCh37 =hg19
- GRCh38 =hg38 released in early 2014
- GRCm38 =mm10

The Genome Reference Consortium consists of:

The Wellcome Trust Sanger Institute

The Genome Institute at Washington University

The European Bioinformatics Institute

The National Center for Biotechnology Information

BWA

- BWA can map low-divergent sequences against a large reference genome, such as the human genome.
- It consists of three algorithms:
  1. BWA-backtrack (Illumina sequence reads up to 100bp)
  2. BWA-SW
  3. BWA-MEM
- BWA SW and MEM can map longer sequences (70bp to 1Mbp) and share similar features such as long-read support and split alignment, but BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate.
- BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.

Download and install BWA

http://sourceforge.net/projects/bio-bwa/files/

```
tar xvfj bwa-0.7.12.tar.bz2     # x extracts, v is verbose (details of what it is doing), f skips prompting for each individual file, and j tells it to unzip .bz2 files
cd bwa-0.7.12
make
```

```
export PATH=$PATH:/path/to/bwa-0.7.12    # Add bwa to your PATH by editing ~/.bashrc file (or .bash_profile or .profile file)
# /path/to/ is an placeholder. Replace with real path to BWA on your machine

source ~/.bashrc
```

Download Reference Genome

```
# download hg19 chromosome fasta files
wget http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz

# unzip and concatenate chromosome and contig fasta files
tar zvfx chromFa.tar.gz
cat *.fa > hg19.fa
rm chr*.fa
```

Create Reference Index

bwa index [-a bwtsw|is] **index_prefix** reference.fasta

bwa index -p hg19bwaidx -a bwtsw hg19.fa

-p index name (change this to whatever you want)

-a index algorithm (**bwtsw** for long genomes and **is** for short genomes)

Align to Reference Genome

```
# aligning single end reads
bwa aln -t 4 hg19bwaidx  sequence1.fq.gz >  sequence1.bwa
bwa samse hg19bwaidx sequence1.bwa sequence1.fq.gz> sequence1_se.sam
```

```
# aligning paired end reads
bwa aln -t 4 hg19bwaidx  sequence1.fq.gz >  sequence1.sai
bwa aln -t 4 hg19bwaidx  sequence2.fq.gz >  sequence2.sai
bwa sampe hg19bwaidx sequence1.sai sequence2.sai sequence1.fq.gz sequence2.fq.
gz > sequence12_pe.sam
```

Generate BAM files

samtools view -bT hg19.fa sequence1.sam > sequence1.bam # when no header

samtools view -bS sequence1.sam > sequence1.bam # when SAM header present

samtools sort -O bam -o sequence1.sorted.bam -T temp sequence1.bam # sort by coordinate to streamline data processing

samtools index sequence1.sorted.bam # a position-sorted BAM file can also be indexed

Acknowledgments

CRUK CI

MRC Cancer Unit