Introduction to NGS Alignment

Presented by
Paula Moolhuijzen| Centre for Crop Disease Management (CCDM), Curtin University, Perth

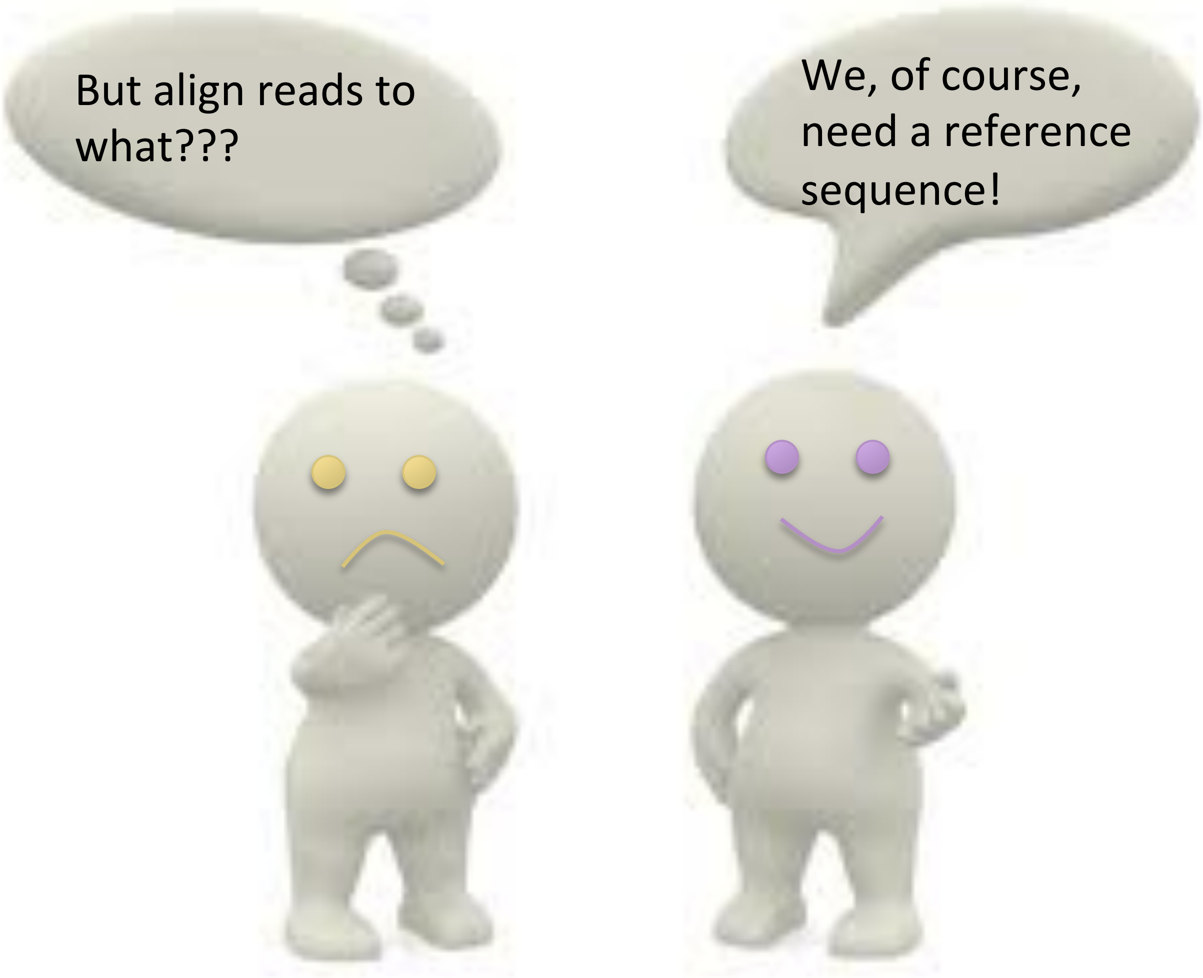
Contributors:
Trainers BPA-CSIRO training platform and EMBL-EBI

Outline

- What is short read alignment?
- Keep in mind
  - **Different Sequencing Purposes and alignment**
- Formats & tools
- Understand current challenges
- Hands-on session on short read alignment

Short Read Alignment (I)

From FASTQ format to meaningful alignment


But align reads to what???
We, of course, need a reference sequence!

Short Read Alignment (II)

GOAL: Given a reference sequence and a set of short reads, align each read to the reference sequence

Reference Sequence

```
GCTGATGTGCCGCCTCACTTCGGTGG
```

Short-reads

```
 CTGATGTGCCGCCTCACTTCGGTGGT
  TGATGTGCCGCCTCACTACGGTGGTG
   GATGTGCCGCCTCACTTCGGTGGTGA
GCTGATGTGCCGCCTCACTACGGTG
GCTGATGTGCCGCCTCACTACGGTG
```

Short Read Alignment (II)

GOAL: Given a reference sequence and a set of short reads, align each read to the reference sequence.

Reference Sequence

```
GCTGATGTGCCGCCTCACTTCGGTGG
```

Short-reads

```
 CTGATGTGCCGCCTCACTTCGGTGGT
  TGATGTGCCGCCTCACTACGGTGGTG
   GATGTGCCGCCTCACTTCGGTGGTGA
GCTGATGTGCCGCCTCACTACGGTG
GCTGATGTGCCGCCTCACTACGGTG
```

Reference sequence availability (?)

No -> *de novo* assembly (see Day 3)

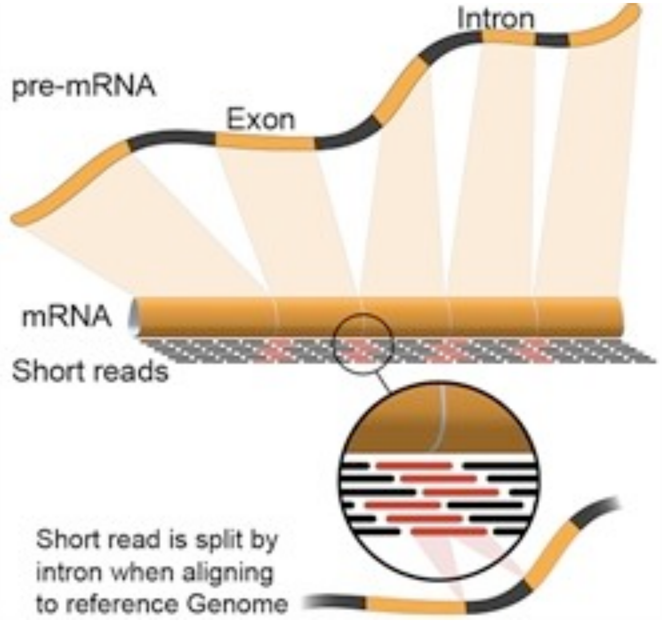
Yes -> use available reference sequences

Different Sequencing Purposes and alignment

- Whole genome sequencing/re-sequencing
  - Align genomic DNA to reference genome
  - often no reference sequence – *de novo* assembly (Day 3)

- ChIP-Seq (protein-DNA associations)
  - Aligning genomic DNA to a reference genome

- RNA-Seq (Transcriptome sequencing)
  - Can align RNA sequence to a reference genome (spliced alignment, Day 2) or a reference transcriptome


Intron
pre-mRNA
Exon
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

Keep in mind

- Allow for mismatches when aligning reads to a reference sequence
  - Number of expected mismatches (1~2 per read)
    - Sequencing machines are not infallible
    - Species polymorphism
  - Distinguish between SNPs and sequencing errors

- Different types of sequencing errors across multiple platforms
  - Insertion and deletion errors at homopolymers (454)
  - Unpredictable distributions of low quality calls (Illumina)

Short-read Aligners

PROGRAM	ALGORITHM	LONG READ	GAPPED	PAIR- END	SPLICED
BOWTIE	BWT	NO	NO	YES	NO
BWA	BWT	YES	YES	YES	NO
MAQ	HASH (read)	NO	NO	YES	NO
SOAP	HASH (ref.)	NO	YES	YES	NO
TopHat	BWT	YES	YES	YES	YES
GSNAP	HASH (read)	YES	YES	YES	YES

STAR fast RNA-Seq aligner

https://omictools.com/read-alignment-category

Alignment Data Formats

- Alignment inputs
  - FASTA format (Reference) *.fa
  - FASTQ format (Raw Read Sequence) *.fq.gz
- Alignment outputs
  - SAM format (Alignment, text) *.sam
  - BAM format (SAM alignment, compressed binary) *.bam

http://samtools.sourceforge.net/SAM1.pdf

SAM a tab-delimited text format

The Sequence Alignment/Map (SAM) format is a generic nucleotide alignment format that describes the alignment of query sequences or sequencing reads to a reference sequence or assembly.

- Flexible store information (default format for aligners)
- Simple to generate or convert different formats
- Compact in file size;
- Works on streaming - Memory
- Allows indexing by genomic position to efficiently retrieve all reads aligning to a locus.

SAM is a bit slow to parse; so there is a binary equivalent to SAM, called BAM.

SAM format – header section

- SAM file header lines start with @
- @ is followed by TAGs of Header fields in TYPE:VALUE pairs

@RG **ID**:RUN_LANE **CN**:Institute **LB**:LibraryName
PL:Technology **PU**:RunName **SM**:Sample

Example:

@RG **ID**:61DP1AAXX_1 **CN**:AGRF **LB**:Rameses
PL:ILLUMINA
PU: 61DP1AAXX.1 **SM**: HOLAUSM000A00009637

SAM format – Alignment section

Alignment section- 11 tab separated mandatory fields

```
HWI-HI83:6:1101:1210:1974#0/1 99 chr20 287833 30 10M1D25M = 287993 195 \
ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA ?8?D?DDDDD8DDDE?E2:<A4CFC?CFB3A?F?C
```

1. QNAME: Query name of the read or the read pair	HWI-HI83:6:1101:1210:
2. FLAG: Bitwise flag (pairing, strand, mate strand, etc.)	99
3. RNAME: Reference sequence name	chr20
4. POS: 1-Based leftmost position of clipped alignment	287833
5. MAPQ: Mapping quality (Phred-scaled)	30
6. CIGAR: Extended CIGAR string (operations: MIDNSHP)	10M1D25M
7. MRNM: Mate reference name (‘=’ if same as RNAME)	=
8. MPOS: 1-based leftmost mate position	287993
9. ISIZE: Inferred insert size	195
10. SEQQuery: Sequence on the same strand as the reference	ACCTATATCTTGGCCTTGGCC
11. QUAL: Query quality (ASCII-33 = Phred base quality)	?8?D?DDDDD8DDDE?E2:<A

Bitwise flag http://ppotato.files.wordpress.com/2010/08/sam_output.pdf

SAM/BAM format

CIGAR operators

- M: match/mismatch
- I: insertion
- D: deletion
- S: softclip
- H: hardclip
- P: padding
- N: skip

```
Ref: GCATTCAGATGCAGTACGC
Read:  ccTCAG--GCAGTAgtg
POS  CIGAR
5    2S4M2D6M3S
```

BAM

- Binary compressed version of SAM
- About 1/3 – 1/5 the storage requirements of SAM

http://samtools.sourceforge.net/

SAM/BAM tools

- Well defined specifications for SAM/BAM

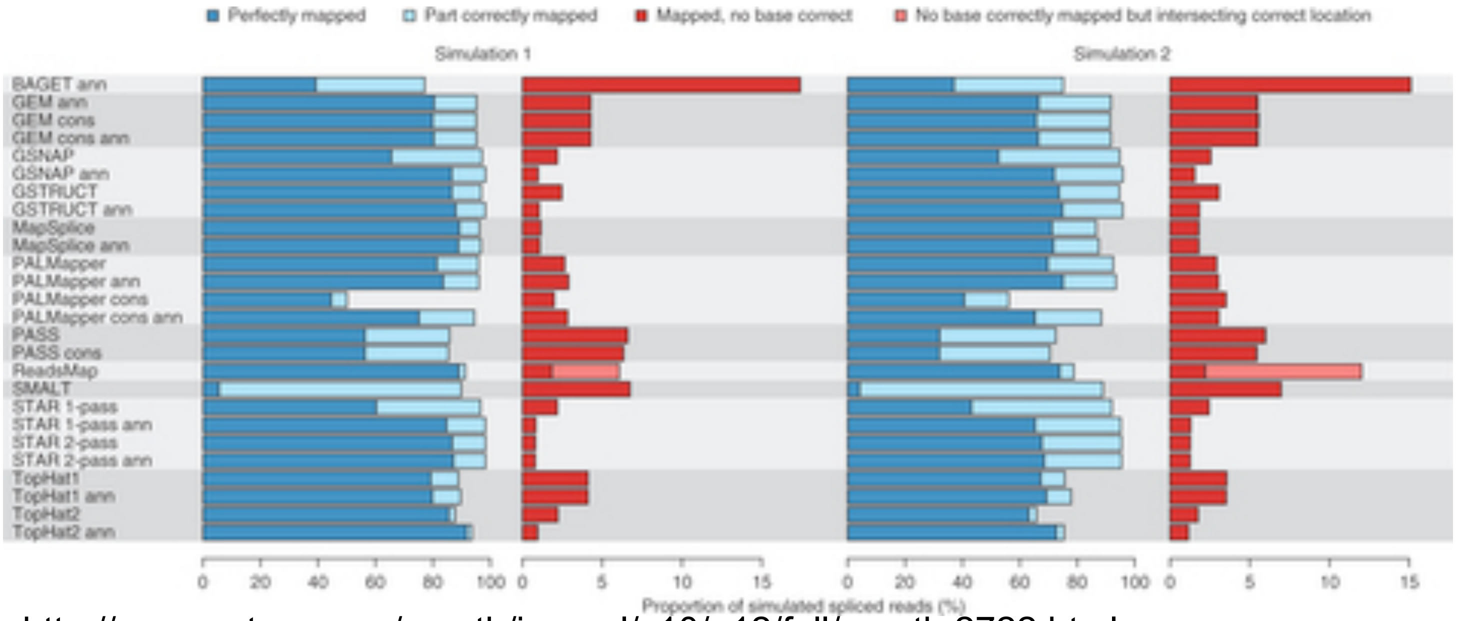
- Advanced interacting programs
  - Samtools – by Sanger (http://samtools.sourceforge.net)
    - Command-line tool
    - Packages a number of utilities to access the information stored in SAM/BAM file
    - e.g. sort reads based on the mapping position in reference:
      - samtools sort aln.bam aln_sorted.bam

  - Picard – By Broad Institute (http://picard.sourceforge.net)
    - Command-line tool, required Java 1.6
    - Designed to run in 2GB of JVM (Xmx2g is recommended)
    - MarkDuplicates, CollectAlignmentSummaryMetrics, SamToFastq
    - More advanced options than Samtools, Running time consuming

  - Bio-SamTool (http://search.cpan.org/~lds/Bio-SamTools/)

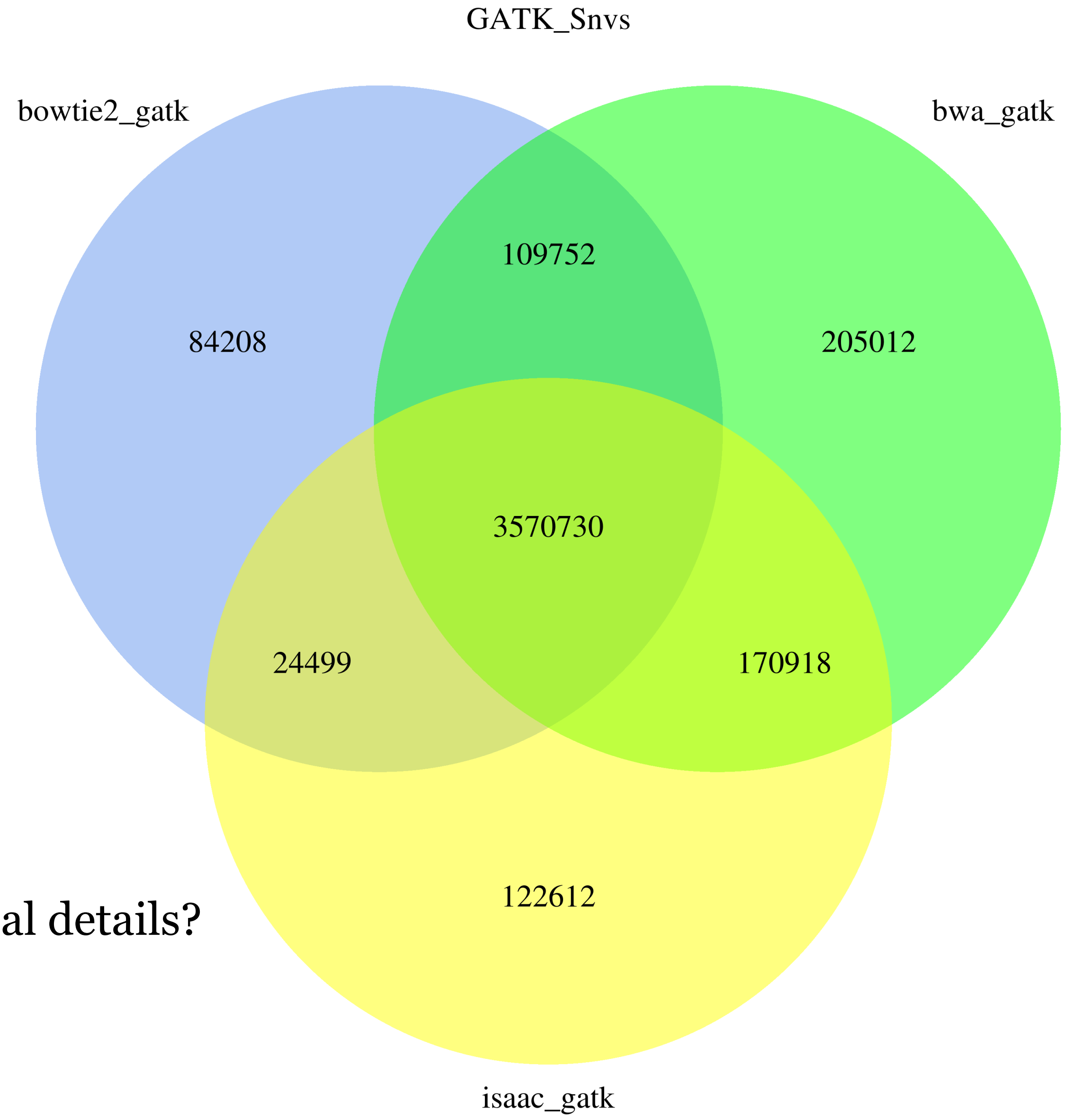
Alignment challenges

- Aligners need to be fast and accurate
  - Trade-off between speed and sensitivity
  - Running time with the growing sequence capacity
    - Illumina HiSeq produces at the moment up to 200m reads per lane
- Gold standard aligner?
  - Evaluate new methods
    - benchmarking of RNAseq by the RGASP project
    - http://www.gencodegenes.org/rgasp/



http://www.nature.com/nmeth/journal/v10/n12/full/nmeth.2722.html

Aligner Choice Effect on Variant Calls



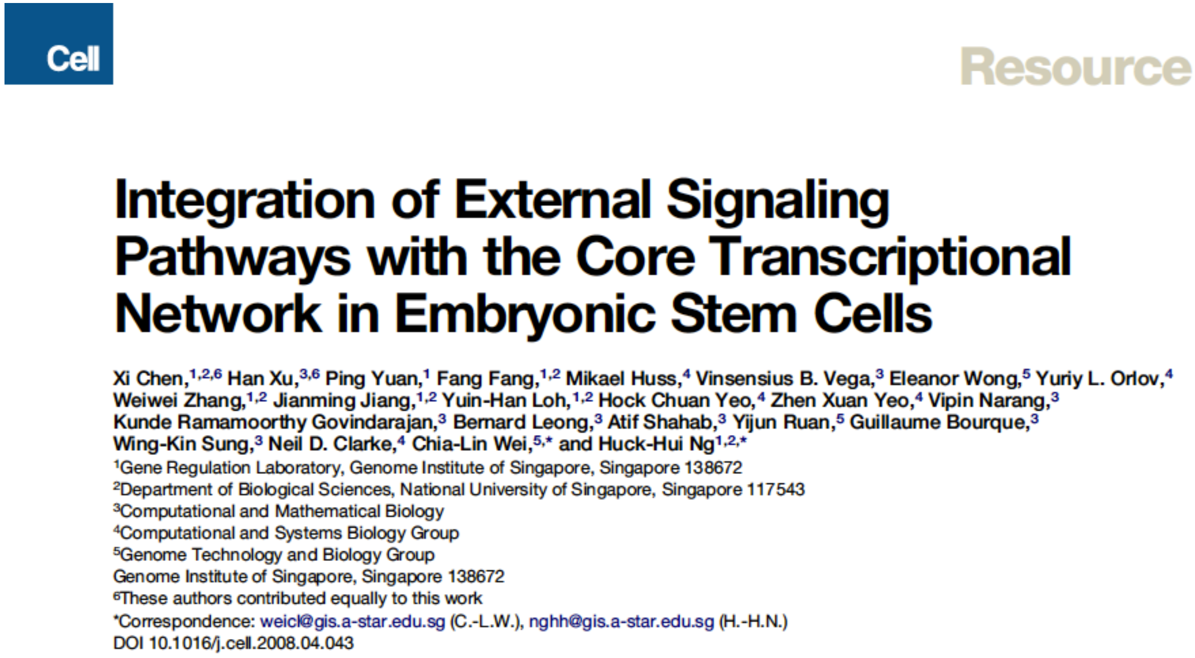
Computational details?

Hands-on session on short read alignment

Practical steps

- ***Index Mouse genome (Chr1)***

1. Align ChIP-Seq samples to the index Mouse Chr1 -> SAM
2. Convert SAM alignment to a sorted BAM file
3. View alignments in Genome viewer IGV



Cell

Resource

Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells

Xi Chen,[1,2,6] Han Xu,[3,6] Ping Yuan,[1] Fang Fang,[1,2] Mikael Huss,[4] Vinsensius B. Vega,[3] Eleanor Wong,[5] Yuriy L. Orlov,[4] Weiwei Zhang,[1,2] Jianming Jiang,[1,2] Yuin-Han Loh,[1,2] Hock Chuan Yeo,[4] Zhen Xuan Yeo,[4] Vipin Narang,[3] Kunde Ramamoorthy Govindarajan,[3] Bernard Leong,[3] Atif Shahab,[3] Yijun Ruan,[5] Guillaume Bourque,[3] Wing-Kin Sung,[3] Neil D. Clarke,[4] Chia-Lin Wei,[5,*] and Huck-Hui Ng[1,2,*]

[1]Gene Regulation Laboratory, Genome Institute of Singapore, Singapore 138672
[2]Department of Biological Sciences, National University of Singapore, Singapore 117543
[3]Computational and Mathematical Biology
[4]Computational and Systems Biology Group
[5]Genome Technology and Biology Group
Genome Institute of Singapore, Singapore 138672
[6]These authors contributed equally to this work
*Correspondence: weicl@gis.a-star.edu.sg (C.-L.W.), nghh@gis.a-star.edu.sg (H.-H.N.)
DOI 10.1016/j.cell.2008.04.043

Thank you