Canadian Bioinformatics Workshops

www.bioinformatics.ca

Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu

creative commons

Attribution-Share Alike 2.5 Canada

You are free:

to Share — to copy, distribute and transmit the work

to Remix — to adapt the work

Free Cultural Works — APPROVED FOR

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.
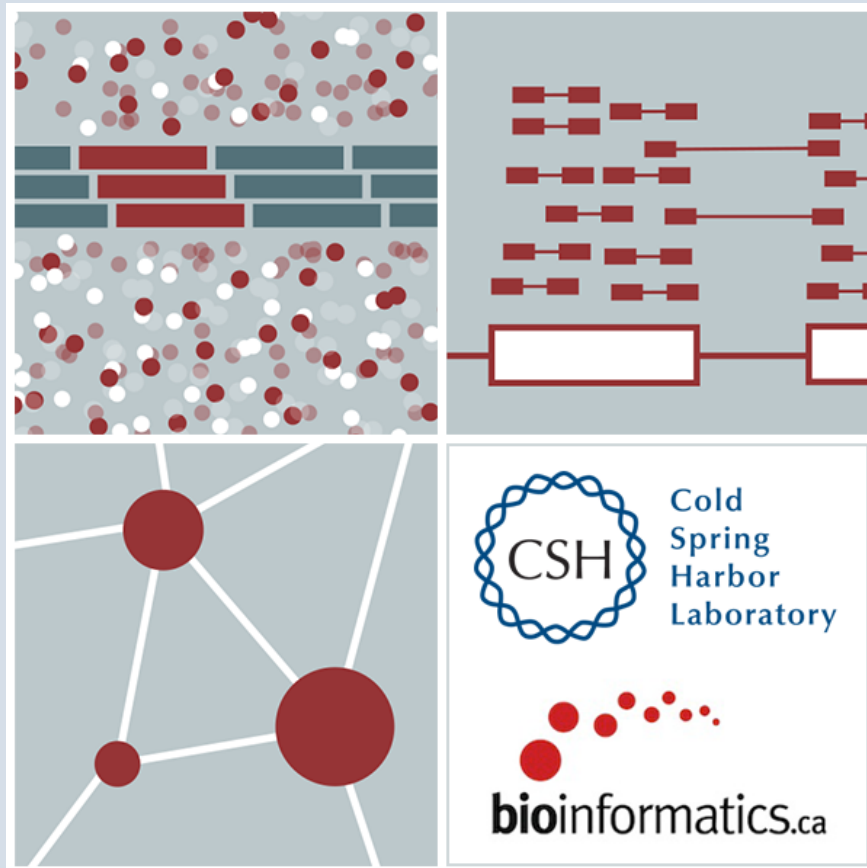
Disclaimer

Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French

Learn how to distribute your work using this licence

Learning objectives of the course

- Module 0: Introduction to Cloud Computing
- Module 1: Introduction to RNA Sequencing
- **Module 2: Alignment and Visualization**
- Module 3: Expression and Differential Expression
- Module 4: Alignment Free Expression Estimation
- Module 5: Isoform Discovery and Alternative Expression

- Tutorials
  - Provide a working example of an RNA-seq analysis pipeline
  - Run in a 'reasonable' amount of time with modest computer resources
  - Self contained, self explanatory, portable

Learning objectives of module 2

- RNA-seq alignment challenges and common questions

- Alignment strategies

- HISAT2

- Introduction to the BAM and BED formats

- Basic manipulation of BAMs

- Visualization of RNA-seq alignments in IGV

- Alignment QC Assessment

- BAM read counting and determination of variant allele expression status

RNA-seq alignment challenges

- Computational cost
  - 100's of millions of reads

- Introns!
  - Spliced vs. unspliced alignments

- Can I just align my data once using one approach and be done with it?
  - Unfortunately probably not

- Is HISAT2 the only mapper to consider for RNA-seq data?
  - http://www.biostars.org/p/60478/

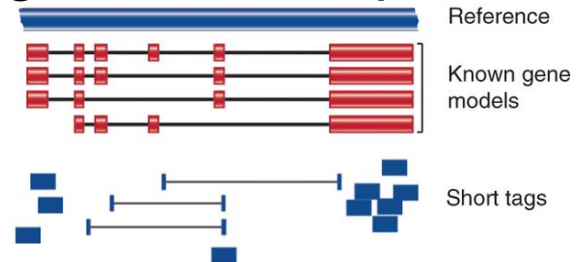# Three RNA-seq mapping strategies

De novo assembly

Assemble transcripts from overlapping tags

Optional: align to genome to get exon structure

Align to transcriptome

Use known and/or predicted gene models to examine individual features

Align to reference genome

Infer possible transcripts and abundance

Diagrams from Cloonan & Grimmond, Nature Methods 2010

Which alignment strategy is best?

- De novo assembly
  - If a reference genome does not exist for the species being studied
  - If complex polymorphisms/mutations/haplotypes might be missed by comparing to the reference genome

- Align to transcriptome
  - If you have short reads (< 50bp)

- Align to reference genome
  - All other cases

- Each strategy involves different alignment/assembly tools
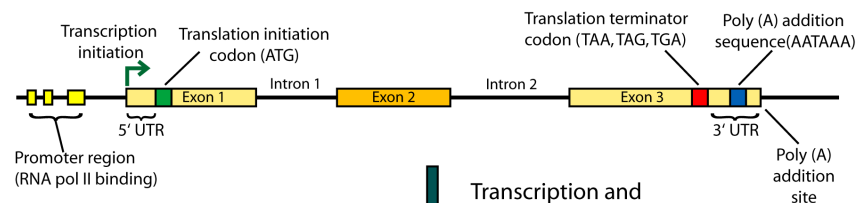
Which read aligner should I use?

RNA
Bisulfite
DNA
microRNA
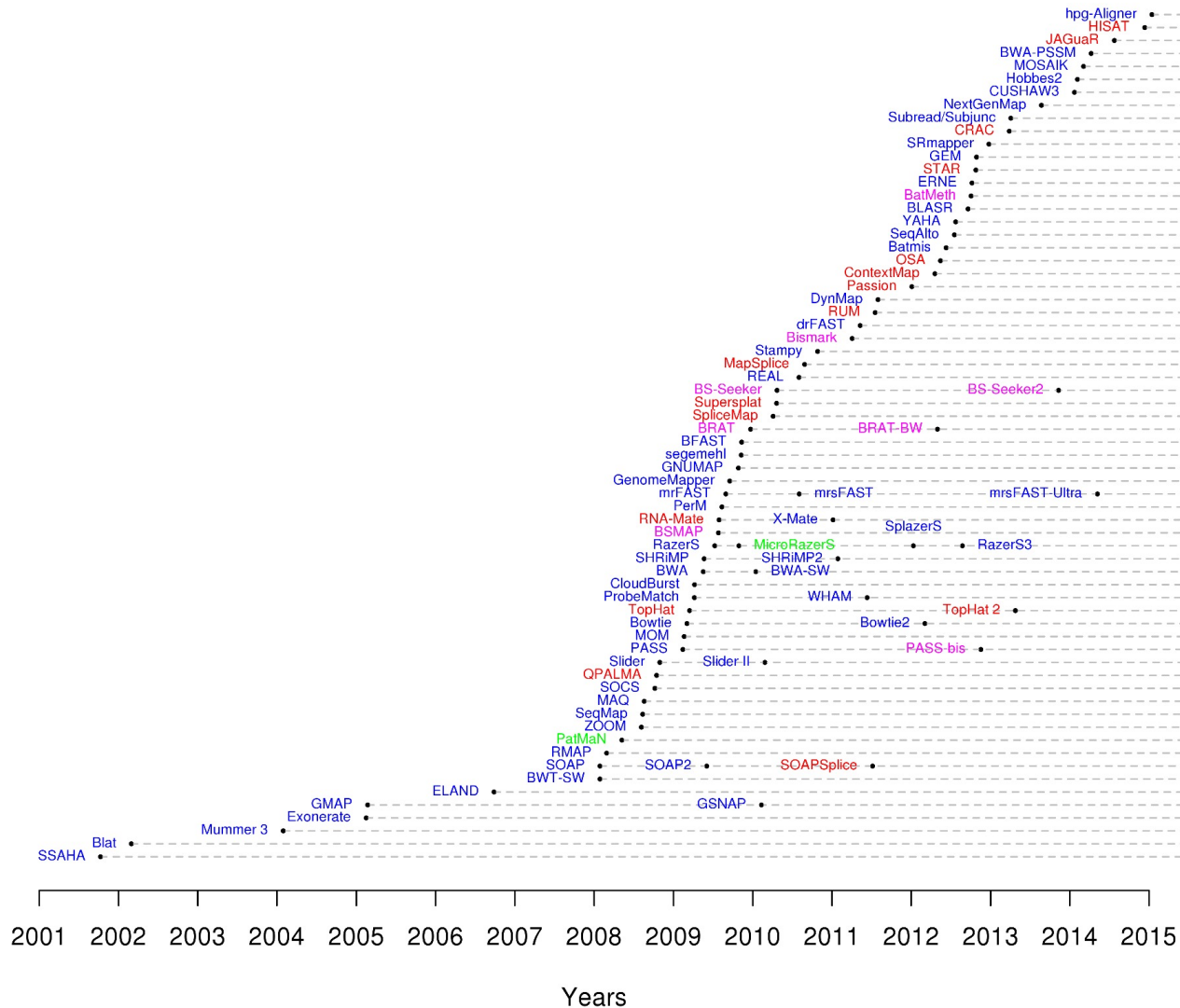
Years
http://wwwdev.ebi.ac.uk/fg/hts_mappers/

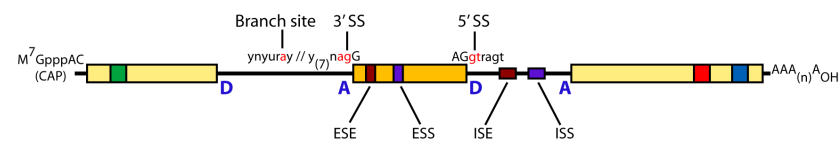# Should I use a splice-aware or unspliced mapper

- RNA-seq reads may span large introns
- The fragments being sequenced in RNA-seq represent mRNA and therefore the introns are removed
- But we are usually aligning these reads back to the reference genome
- Unless your reads are short (<50bp) you should use a splice-aware aligner
  - HISAT2, STAR, MapSplice, etc.
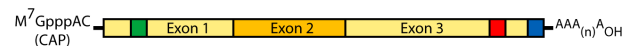

Double-stranded genomic DNA template

Transcription initiation

Translation initiation codon (ATG)

Translation terminator codon (TAA, TAG, TGA)

Poly (A) addition sequence(AATAAA)

Exon 1   Intron 1   Exon 2   Intron 2   Exon 3

5' UTR

Promoter region (RNA pol II binding)

3' UTR

Poly (A) addition site

Transcription and polyadenylation

Single-stranded pre-mRNA (nuclear RNA)

Branch site   3' SS   5' SS

ynyuray // y(7)nagG   AGgtragt

M^7GpppAC (CAP)

D   A   D   A

ESE   ESS   ISE   ISS

$AAA_{(n)}A_{OH}$

RNA processing

Mature mRNA

M^7GpppAC (CAP)

Exon 1   Exon 2   Exon 3

$AAA_{(n)}A_{OH}$
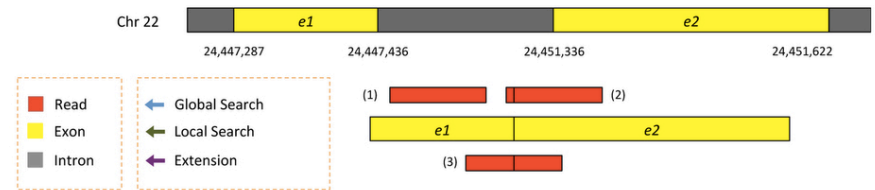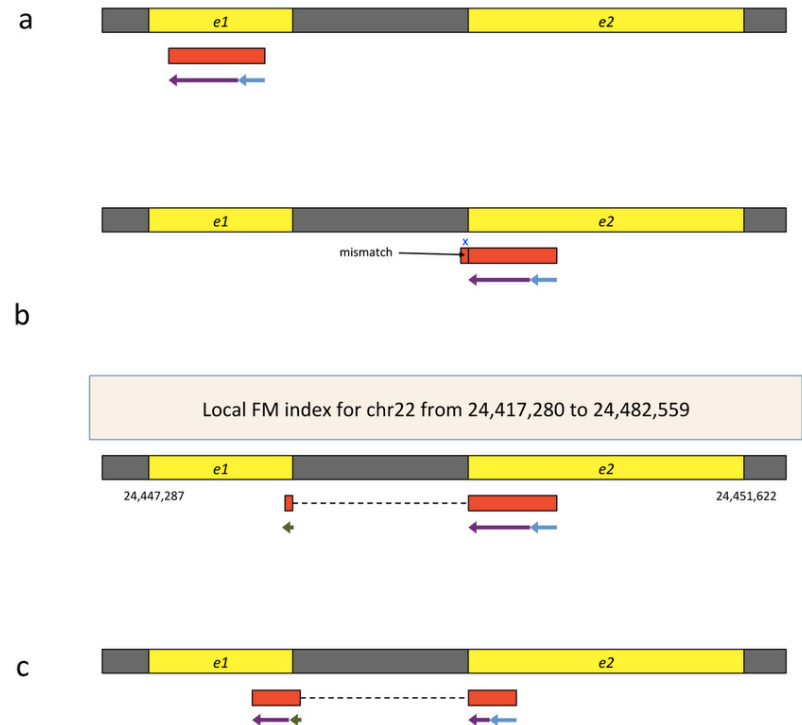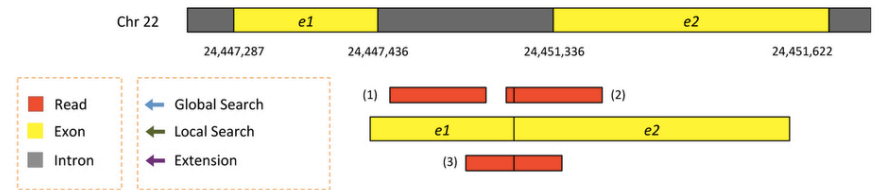
HISAT/HISAT2

- HISAT is a 'splice-aware' RNA-seq read aligner

- Requires a reference genome

- Very fast

- Uses an indexing scheme based on the Burrows-Wheeler transform and the Ferragina-Manzini (FM) index

- Multiple types of indexes for alignment

  – a whole-genome FM index to anchor each alignment

  – numerous local FM indexes for very rapid extensions of these alignments.

  – Whole-genome indices with SNPs and known transcript structures accounted for

Kim et al. 2015. Nat Methods 12:357–360

bioinformatics.ca

HISAT/HISAT2

- Uses hierarchical indexing algorithm and several adaptive strategies, based on the position of a read with respect to splice sites

- First tries to find candidate locations across the target genome from which the read may have originated by mapping part of each read using the global FM index, which in most cases identifies one or a small number of candidates.

- Then selects one of ~48,000 local indexes for each candidate and uses it to align the remainder of the read.

- For paired reads, each mate is separately aligned and the alignments of both mates are combined.
    - If a read fails to align, then the alignments of its mate are used as anchors to map the unaligned mate

Kim et al. 2015. Nat Methods 12:357–360

HISAT/HISAT2
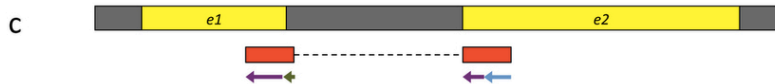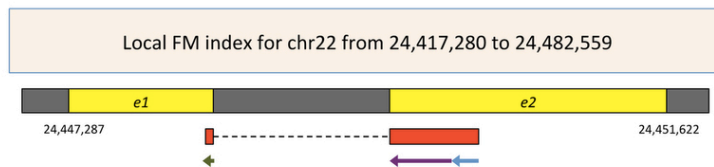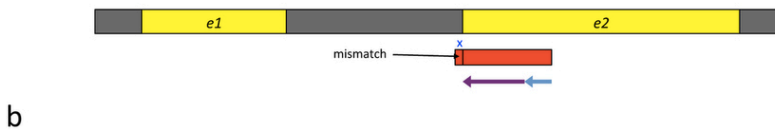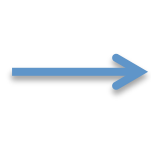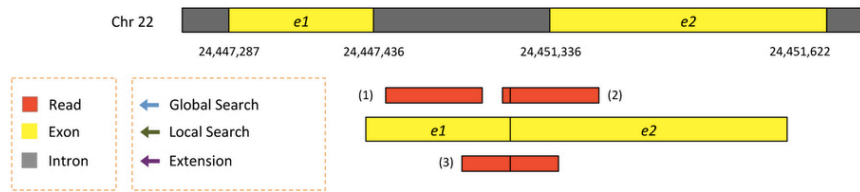
- First align read with global index (slower)
- Once at least 28bp and exactly one location switch to extension mode against reference genome (faster)

HISAT/HISAT2
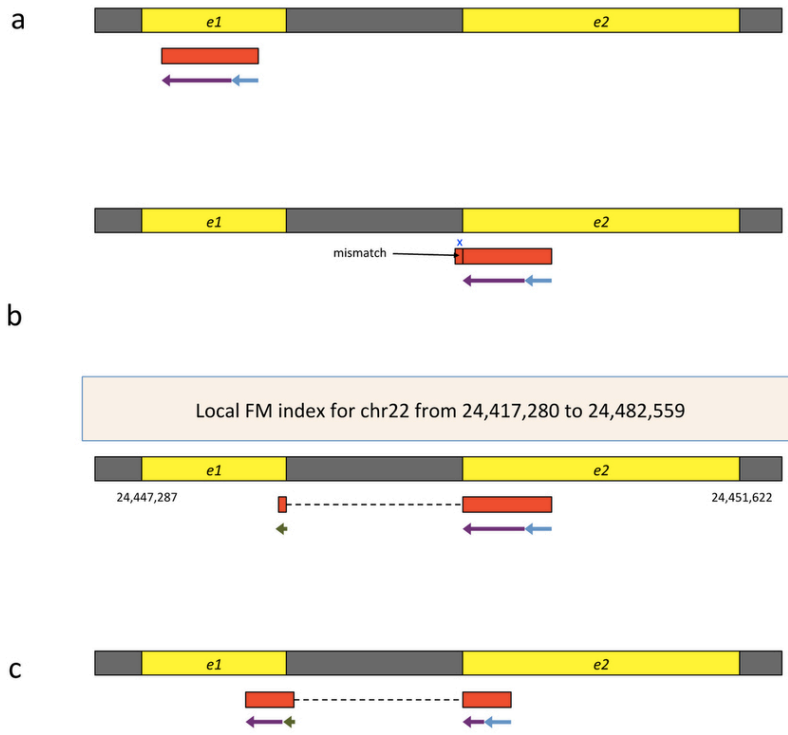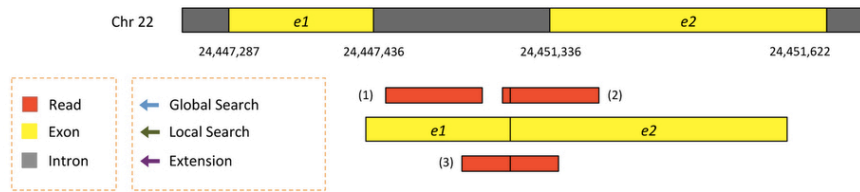
- Again use global search until exactly one match of at least 28bp (slower)
- Extend as before until mismatch at 93bp (faster)
- Switch to local FM index to align remaining 8bp
  - Because the index covers only a small region, in this case we find just one match for the 8-bp segment.
- Check for compatibility and combine into single spliced alignment

HISAT/HISAT2
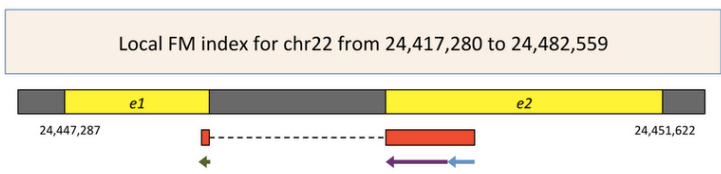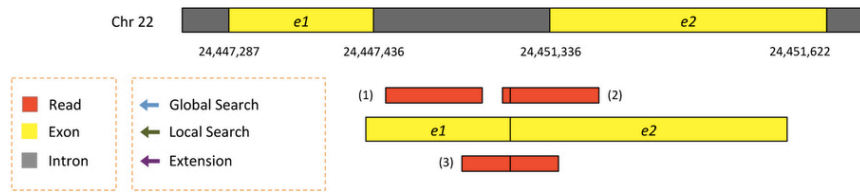
- Again use global search until exactly one match of at least 28bp (slower)
- Extend as before until mismatch at 51bp (faster)
- Switch to local FM index to align first 8bp of remaining read
  - If too many matches increase prefix size
- Extend again
- Check for compatibility and combine into single spliced alignment
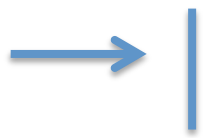
Should I allow 'multi-mapped' reads?

- Depends on the application

- In *DNA* analysis it is common to use a mapper to randomly select alignments from a series of equally good alignments

- In *RNA* analysis this is less common
  - Perhaps disallow multi-mapped reads if you are variant calling
  - Definitely should allow multi-mapped reads for expression analysis with Cufflinks (and StringTie?)
  - Definitely should allow multi-mapped reads for gene fusion discovery

bioinformatics.ca

What is the output of HISAT2?

- A SAM/BAM file
  - SAM stands for Sequence Alignment/Map format
  - BAM is the binary version of a SAM file

- Remember, compressed files require special handling compared to plain text files

- How can I convert BAM to SAM?
  - http://www.biostars.org/p/1701/

bioinformatics.ca

Example of SAM/BAM file format

Example SAM/BAM header section (abbreviated)

Example SAM/BAM alignment section (only 10 alignments shown)

Introduction to the SAM/BAM format

- The specification
  - http://samtools.sourceforge.net/SAM1.pdf
- The SAM format consists of two sections:
  - Header section
    - Used to describe source of data, reference sequence, method of alignment, etc.
  - Alignment section
    - Used to describe the read, quality of the read, and nature alignment of the read to a region of the genome
- BAM is a compressed version of SAM
  - Compressed using lossless BGZF format
  - Other BAM compression strategies are a subject of research. See 'CRAM' format for example
- BAM files are usually 'indexed'
  - A '.bai' file will be found beside the '.bam' file
  - Indexing aims to achieve fast retrieval of alignments overlapping a specified region without going through the whole alignments. BAM must be sorted by the reference ID and then the leftmost coordinate before indexing

SAM/BAM header section

- Used to describe source of data, reference sequence, method of alignment, etc.
- Each section begins with character '@' followed by a two-letter record type code. These are followed by two-letter tags and values
  - @HD The header line
    - VN: format version
    - SO: Sorting order of alignments
  - @SQ Reference sequence dictionary
    - SN: reference sequence name
    - LN: reference sequence length
    - SP: species
  - @RG Read group
    - ID: read group identifier
    - CN: name of sequencing center
    - SM: sample name
  - @PG Program
    - PN: program name
    - VN: program version

SAM/BAM alignment section

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
★ 2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
★ 6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next segment
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next segment
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Example values

```
1       QNAME   e.g.   HWI-ST495_129147882:1:2302:10269:12362 (QNAME)
2       FLAG    e.g.   99
3       RNAME   e.g.   1
4       POS     e.g.   11623
5       MAPQ    e.g.   3
6       CIGAR   e.g.   100M
7       RNEXT   e.g.   =
8       PNEXT   e.g.   11740
9       TLEN    e.g.   217
10      SEQ     e.g.   CCTGTTTCTCCACAAAGTGTTTACTTTTGGATTTTTGCCAGTCTAACAGGTGAAGCCCTGGAGATTCTTATTAGTGATTTGGGCTGGGGCCTGGCCATGT
11      QUAL    e.g.   CCCFFFFFHHHHHJJIJFIJJJJJJJJJJJHIJJJJJJJJIJJJJJGGHIJHIJJJJJJJJJGHGGIJJJJJJIJEEHHHHFFFFCDCDDDDDDB@ACDD
```

SAM/BAM flags explained

- http://broadinstitute.github.io/picard/explain-flags.html
- 12 bitwise flags describing the alignment
- These flags are stored as a binary string of length 11 instead of 11 columns of data
- Value of '1' indicates the flag is set. e.g. 00100000000
- All combinations can be represented as a number from 1 to 2048 (i.e. 2^{11}-1). This number is used in the BAM/SAM file. You can specify 'required' or 'filter' flags in samtools view using the '-f' and '-F' options respectively

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	**SEQ** being reverse complemented
32	0x20	**SEQ** of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Note that to maximize confusion, each bit is described in the SAM specification using its hexadecimal representation (i.e., '0x10' = 16 and '0x40' = 64).

CIGAR strings explained

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

- The CIGAR string is a sequence of base lengths and associated 'operations' that are used to indicate which bases align to the reference (either a match or mismatch), are deleted, are inserted, represent introns, etc.

- e.g. 81M859N19M
  - A 100 bp read consists of: 81 bases of alignment to reference, 859 bases skipped (an intron), 19 bases of alignment

bioinformatics.ca

Introduction to the BED format

- When working with BAM files, it is very common to want to examine a focused subset of the reference genome
  - e.g. the exons of a gene
- These subsets are commonly specified in 'BED' files
  - https://genome.ucsc.edu/FAQ/FAQformat.html#format1
- Many BAM manipulation tools accept regions of interest in BED format
- Basic BED format (tab separated):
  - Chromosome name, start position, end position
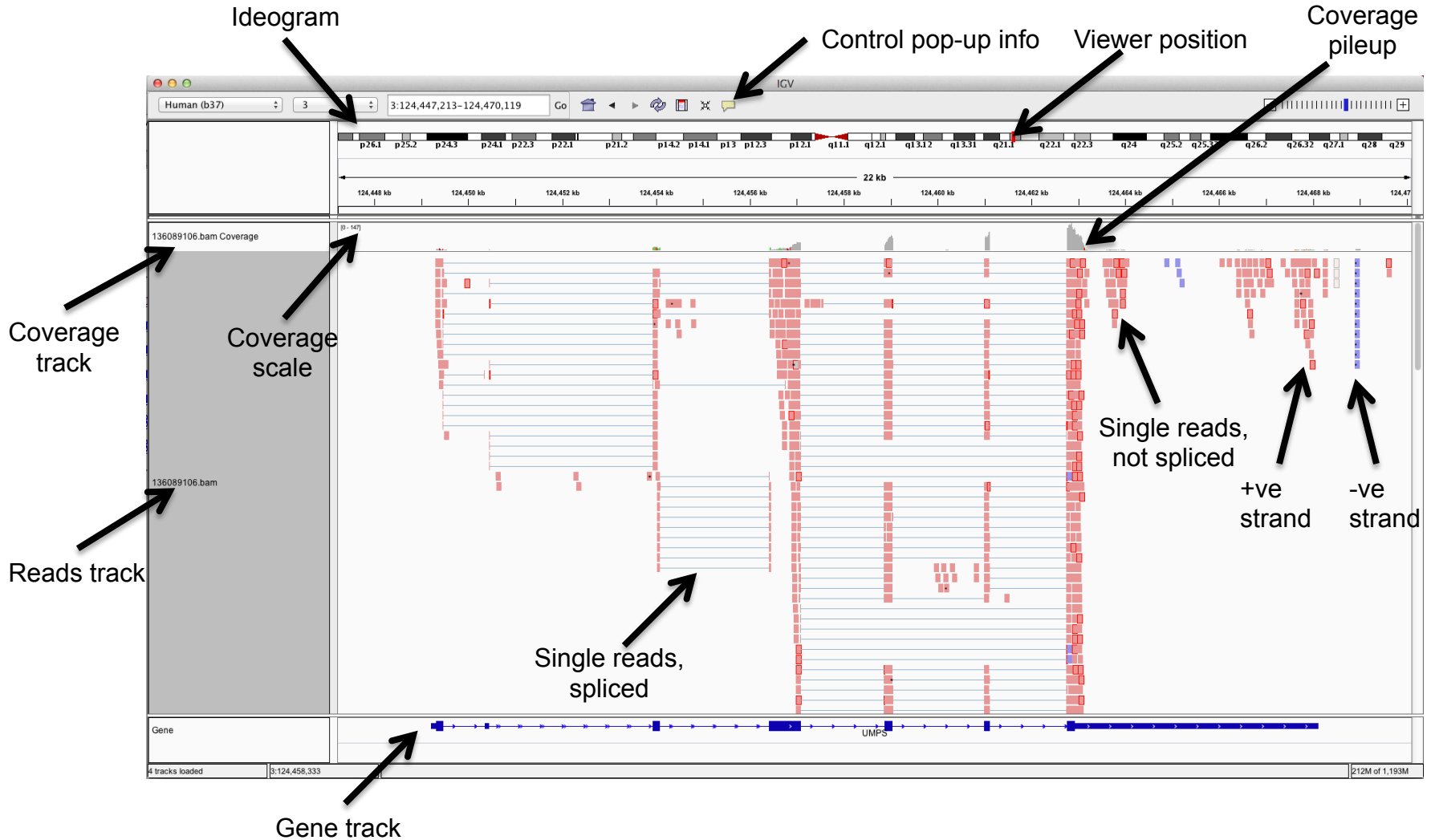 - Coordinates in BED format are 0 based

Manipulation of SAM/BAM and BED files

- Several tools are used ubiquitously in sequence analysis to manipulate these files

- SAM/BAM files
  - samtools
  - bamtools
  - picard

- BED files
  - bedtools
  - bedops

bioinformatics.ca

How should I sort my SAM/BAM file?

- Generally BAM files are sorted by <u>position</u>
  - This is for performance reasons
    - When sorted and indexed, arbitrary positions in a massive BAM file can be accessed rapidly
- Certain tools require a BAM sorted by <u>read name</u>
  - Usually this is when we need to easily identify both reads of a pair
    - The insert size between two reads may be large
    - In fusion detection we are interested in read pairs that map to different chromosomes…

Visualization of RNA-seq alignments in IGV browser

Ideogram

Control pop-up info

Viewer position

Coverage pileup

Coverage track

Coverage scale

Single reads, not spliced

+ve strand

-ve strand

Reads track

Single reads, spliced

Gene track

Alternative viewers to IGV

- Alternative viewers to IGV
  - http://www.biostars.org/p/12752/
  - http://www.biostars.org/p/71300/
- Artemis, BamView, Chipster, gbrowse2, GenoViewer, MagicViewer, **Savant**, Tablet, tview

Alignment QC Assessment

- 3' and 5' Bias

- Nucleotide Content

- Base/Read Quality

- PCR Artifact

- Sequencing Depth

- Base Distribution

- Insert Size Distribution

bioinformatics.ca

Alignment QC: 3' & 5' Bias

BED file
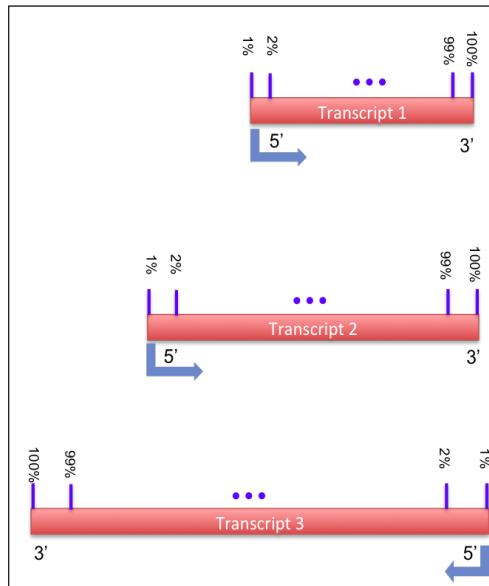
BAM file

Take 100 quantiles from each transcripts in BED file
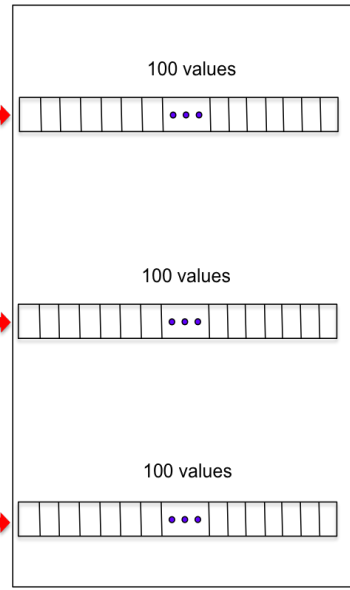
Extract coverage signals from BAM file

http://rseqc.sourceforge.net/
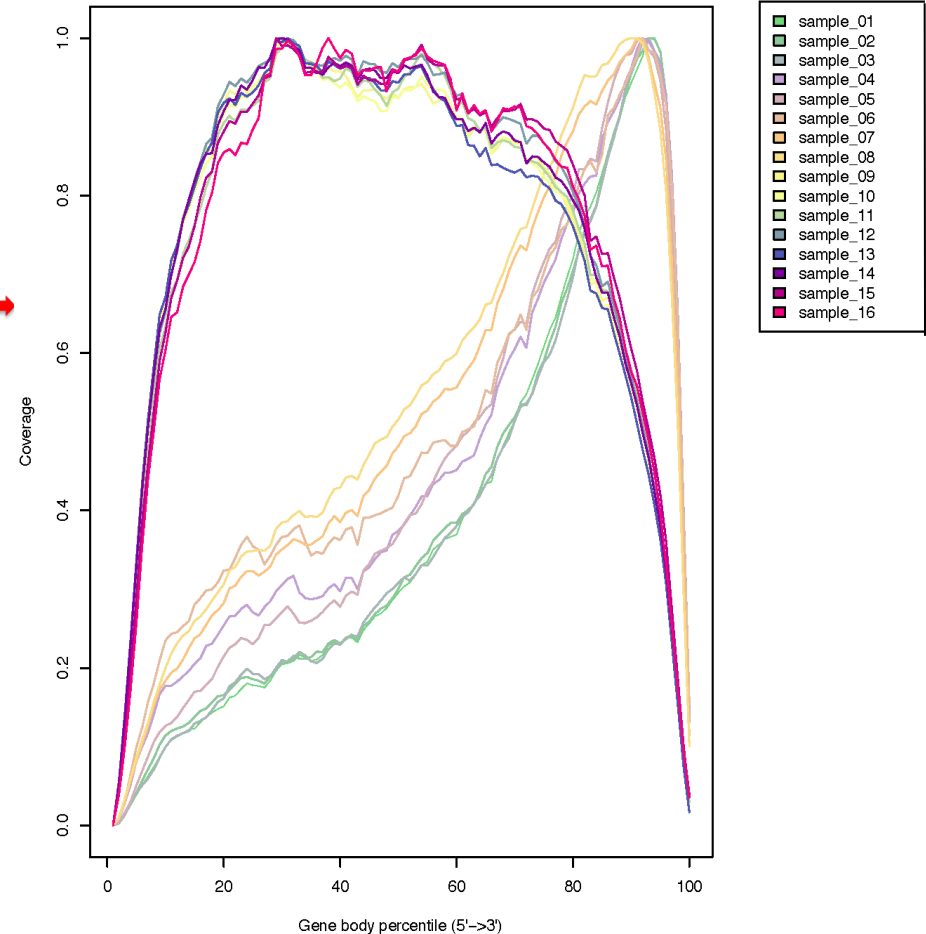
Alignment QC: Nucleotide Content

- **Random primers** are used to reverse transcribe RNA fragments into double-stranded complementary DNA (dscDNA)
- Causes certain patterns to be over represented at the beginning (5'end) of reads
- Deviation from expected A%=C%=G%=T%=25%

Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen,[1,*] Steven E. Brenner,[2] and Sandrine Dudoit[1,3]

Author information ▸ Article notes ▸ Copyright and License information ▸

http://rseqc.sourceforge.net/
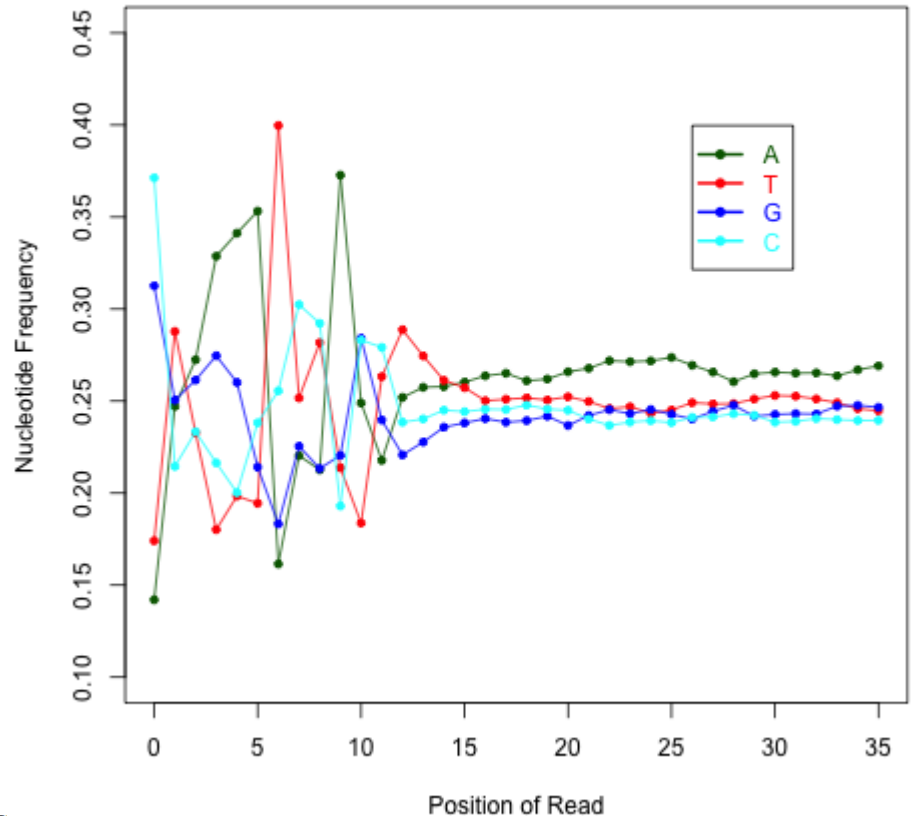
Alignment QC: Quality Distribution

- Phred quality score is widely used to characterize the quality of base-calling

- Phred quality score = -10xlog(10)P, here P is probability that base-calling is wrong

- Phred score of 30 means there is 1/1000 chance that the base-calling is wrong

- The quality of the bases tend to drop at the end of the read, a pattern observed in sequencing by synthesis techniques
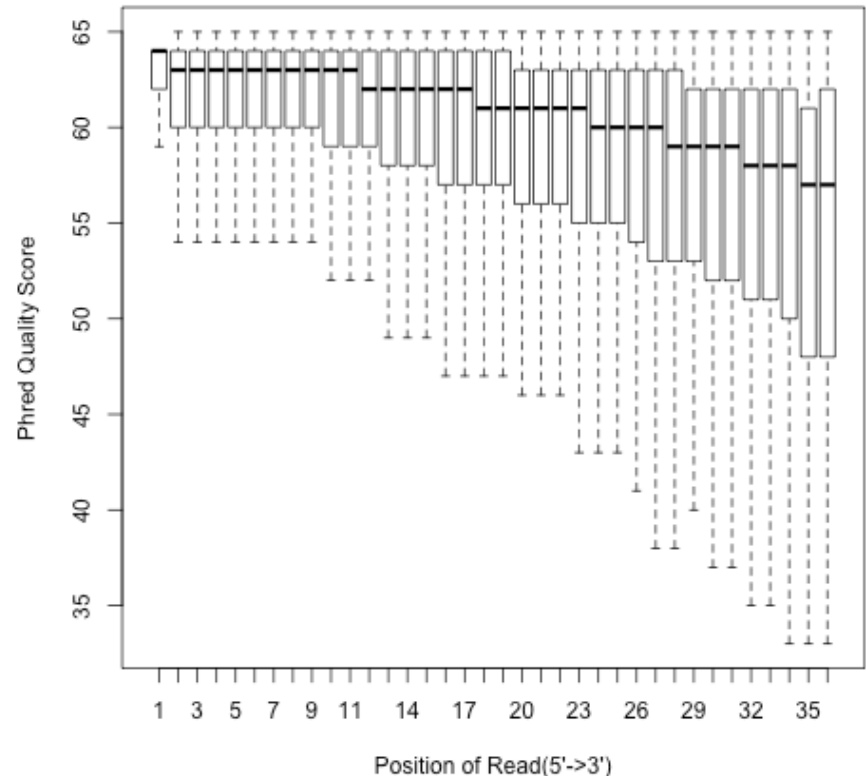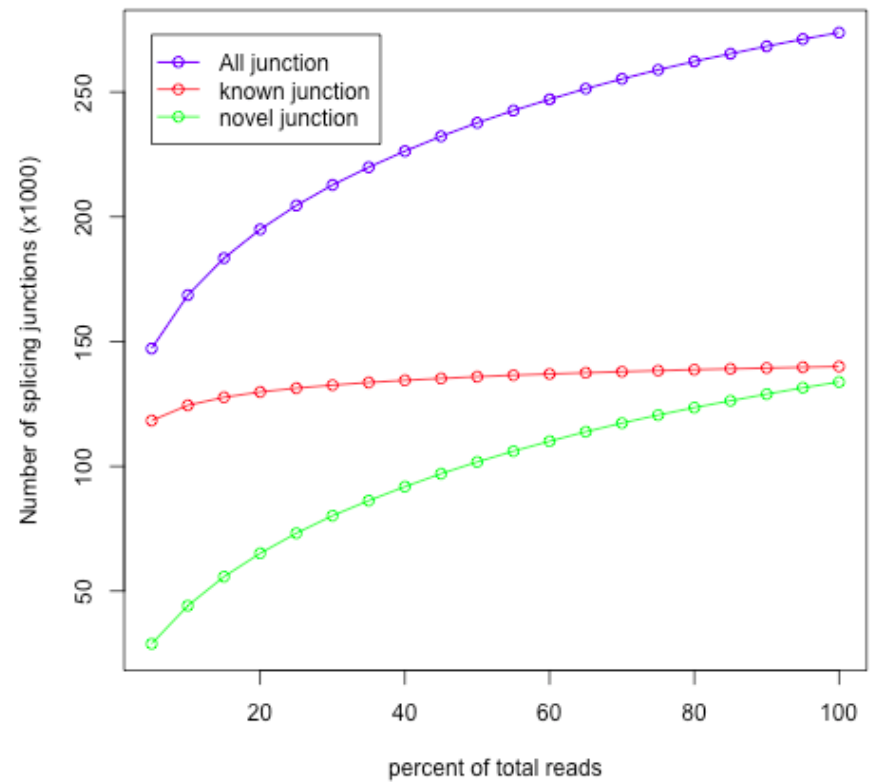
Alignment QC: PCR Duplication

- Duplicate reads are reads that have the same start/end positions and same exact sequence

- In DNA-seq, reads/start point is used as a metric to assess PCR duplication rate

- In DNA-seq, duplicate reads are collapsed using tools such as picard

- How is RNA-seq different from DNA-seq?


http://rseqc.sourceforge.net/
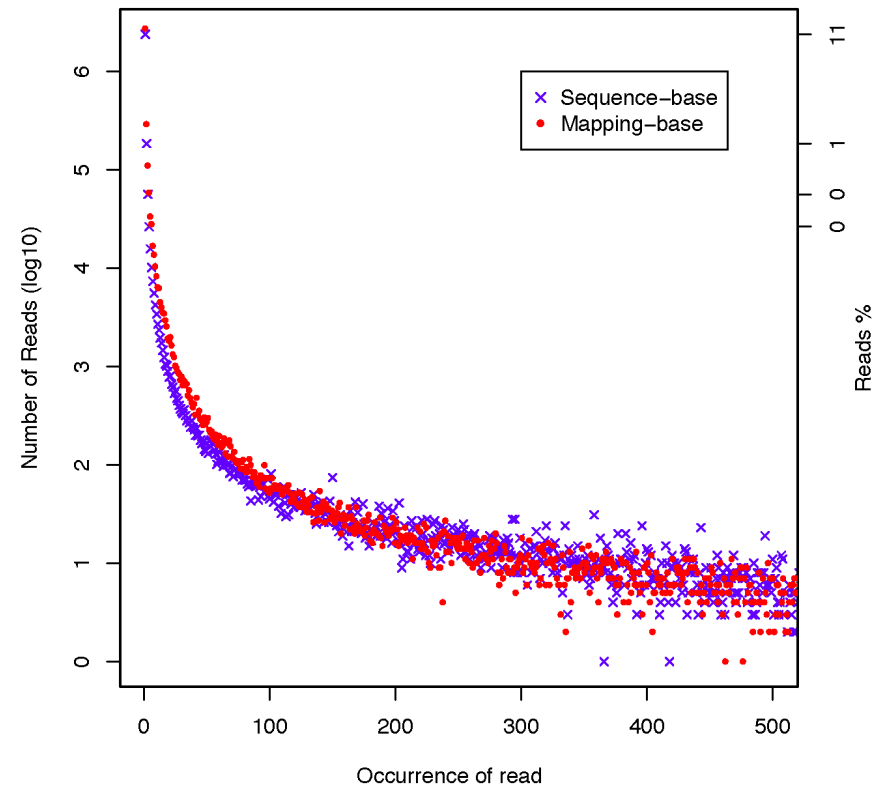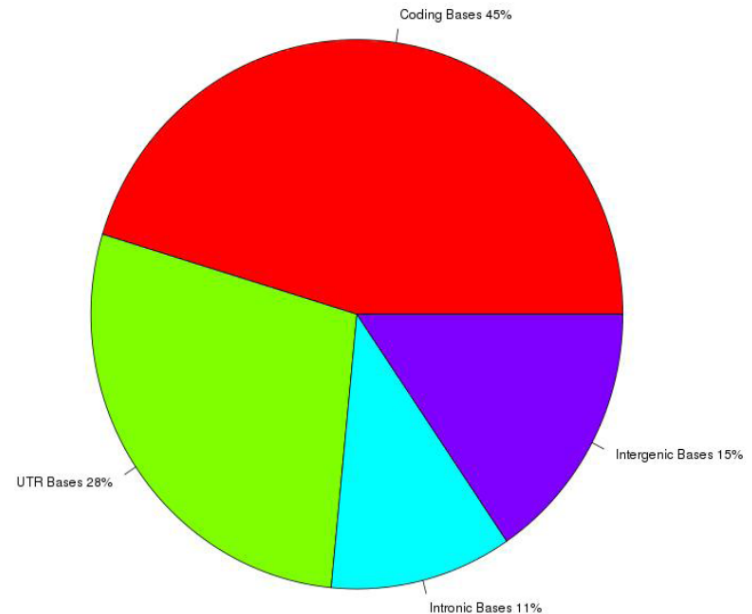
Alignment QC: Sequencing Depth

- **Have we sequenced deep enough?**

- In DNA-seq, we can determine this by looking at the average coverage over the sequenced region. Is it above a certain threshold?

- In RNA-seq, this is a challenge due to the variability in gene abundance

- Use splice junctions detection rate as a way to identify desired sequencing depth

- Check for saturation by resampling 5%, 10%, 15%, ..., 95% of total alignments from aligned file, and then detect splice junctions from each subset and compare to reference gene model.

- This method ensures that you have sufficient coverage to perform alternative splicing analyses

Alignment QC: Base Distribution

UTR Bases 20%
Coding Bases 27%
Intronic Bases 30%
Intergenic Bases 23%
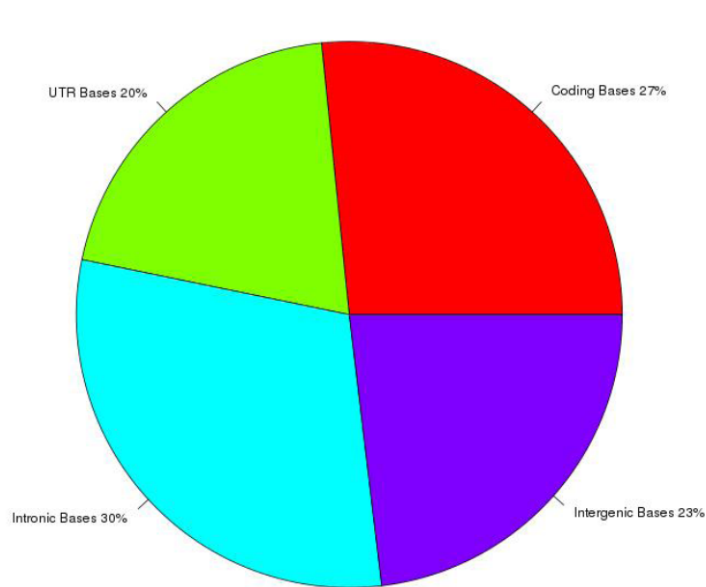
Whole Transcriptome Library

Coding Bases 45%
UTR Bases 28%
Intronic Bases 11%
Intergenic Bases 15%
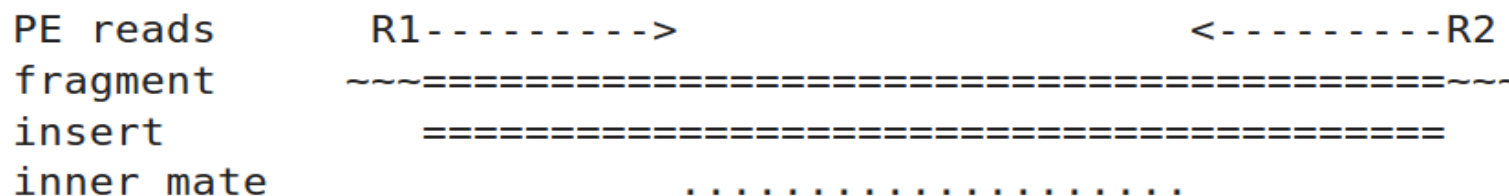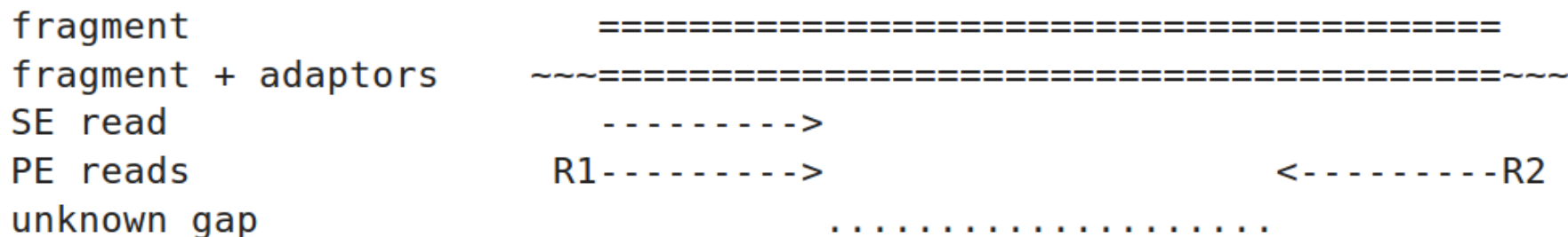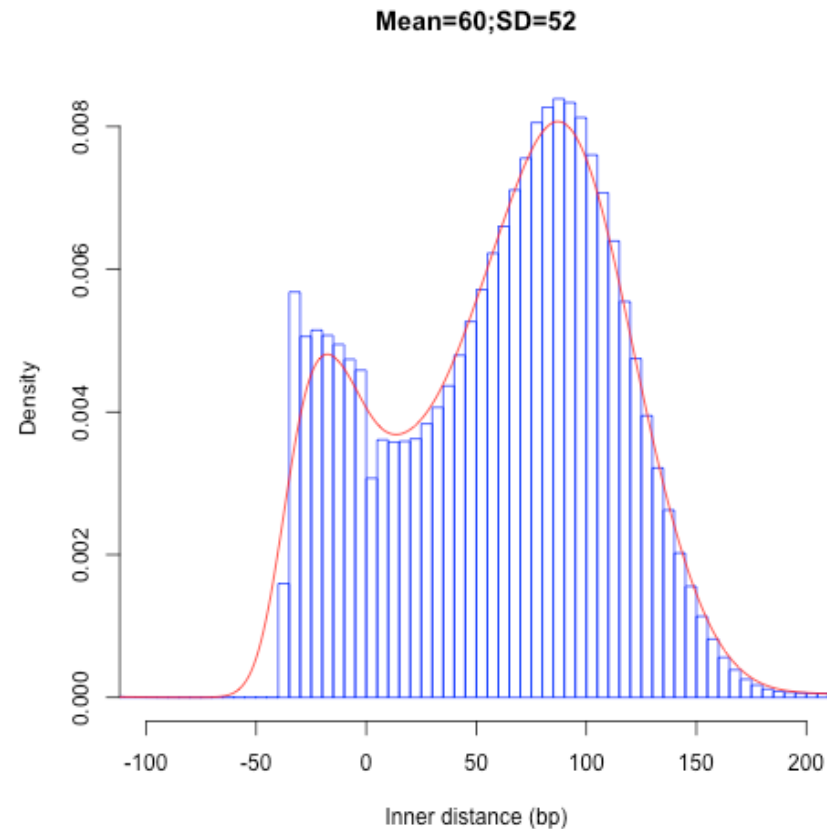
PolyA mRNA library

- Your sequenced bases distribution will depend on the library preparation protocol selected

Alignment QC: Insert Size

```
fragment                          =========================================
fragment + adaptors          ~~~=========================================~~~
SE read                          --------->
PE reads                    R1--------->                      <---------R2
unknown gap                                   .....................
```

```
PE reads            R1--------->                      <---------R2
fragment            ~~~=========================================~~~
insert                 =========================================
inner mate                       .....................
```

http://thegenomefactory.blogspot.ca/2013/08/paired-end-read-confusion-library.html

Alignment QC: Insert Size

Consistent with library size selection?

http://rseqc.sourceforge.net/

BAM read counting and variant allele expression status

- A variant C->T is observed in 12 of 25 reads covering this position. Variant allele frequency (VAF) 12/25 = 48%.
- Both alleles appear to be expressed equally (not always the case) -> heterozygous, no allele specific expression
- How can we determine variant read counts, depth of coverage, and VAF without manually viewing in IGV?

RNA sequencing and analysis

bioinformatics.ca

Introduction to tutorial
(Module 2)

HISAT2/StringTie/Ballgown
RNA-seq Pipeline
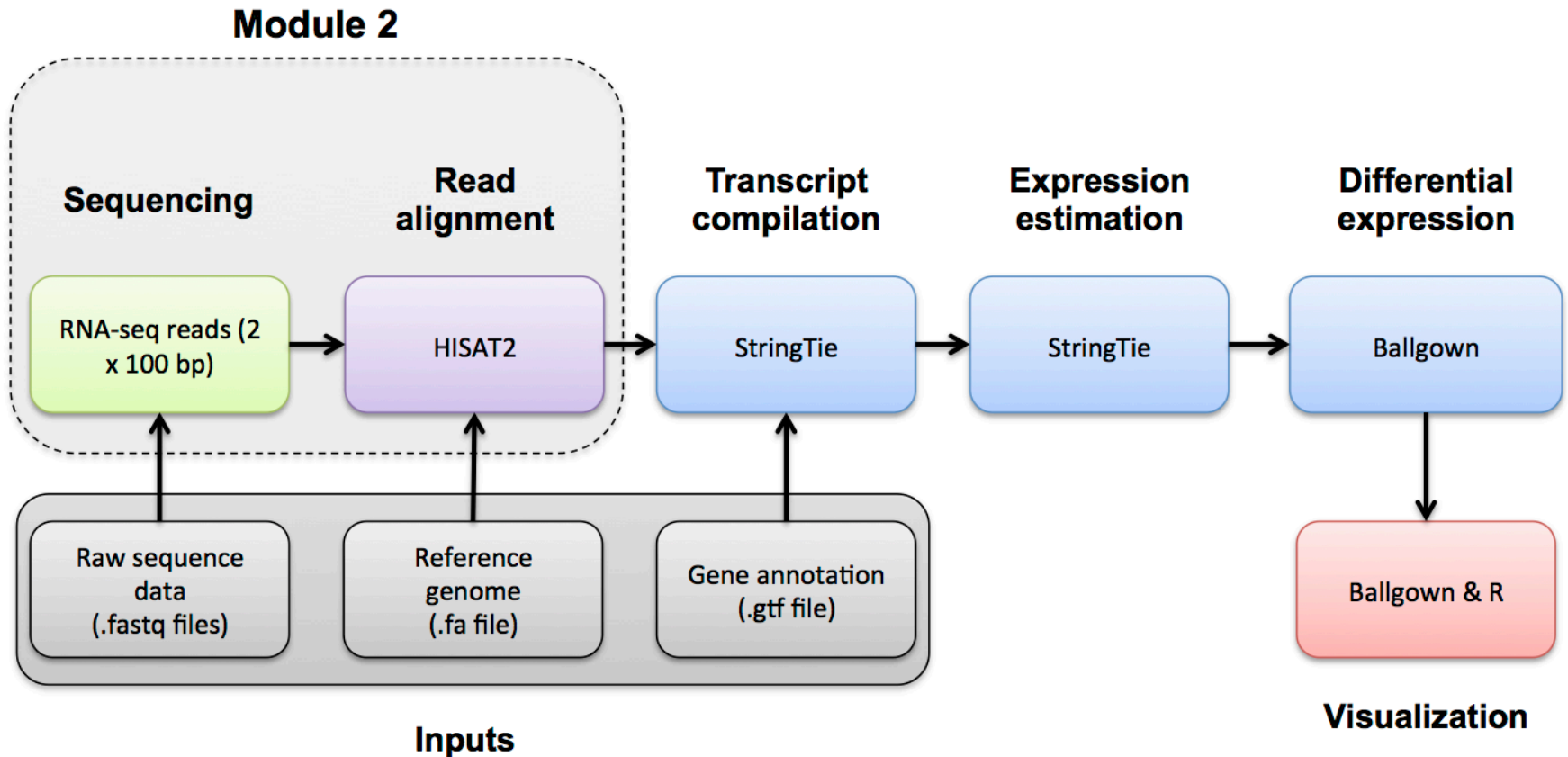
We are on a Coffee Break & Networking Session