bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io

CC BY-SA 4.0 DEED

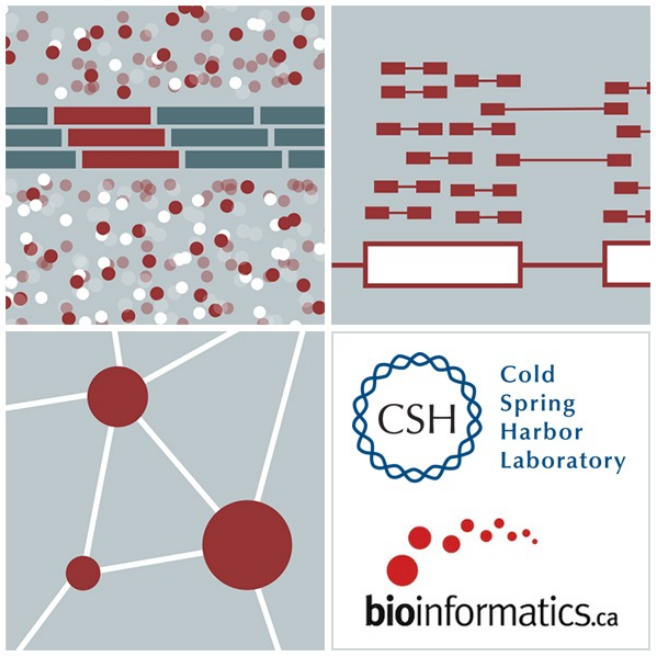
Attribution-ShareAlike 4.0 International

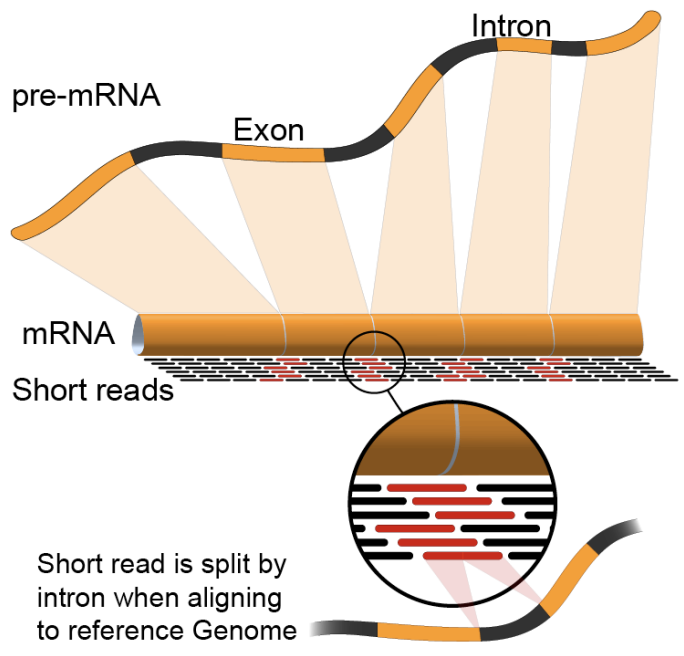
RNA-Seq Module 2: Alignment QC

Malachi Griffith, Obi Griffith, Isabel Risch, Vida Talebian

RNA-seq Analysis 2024. June 17-19, 2024

bioinformatics.ca



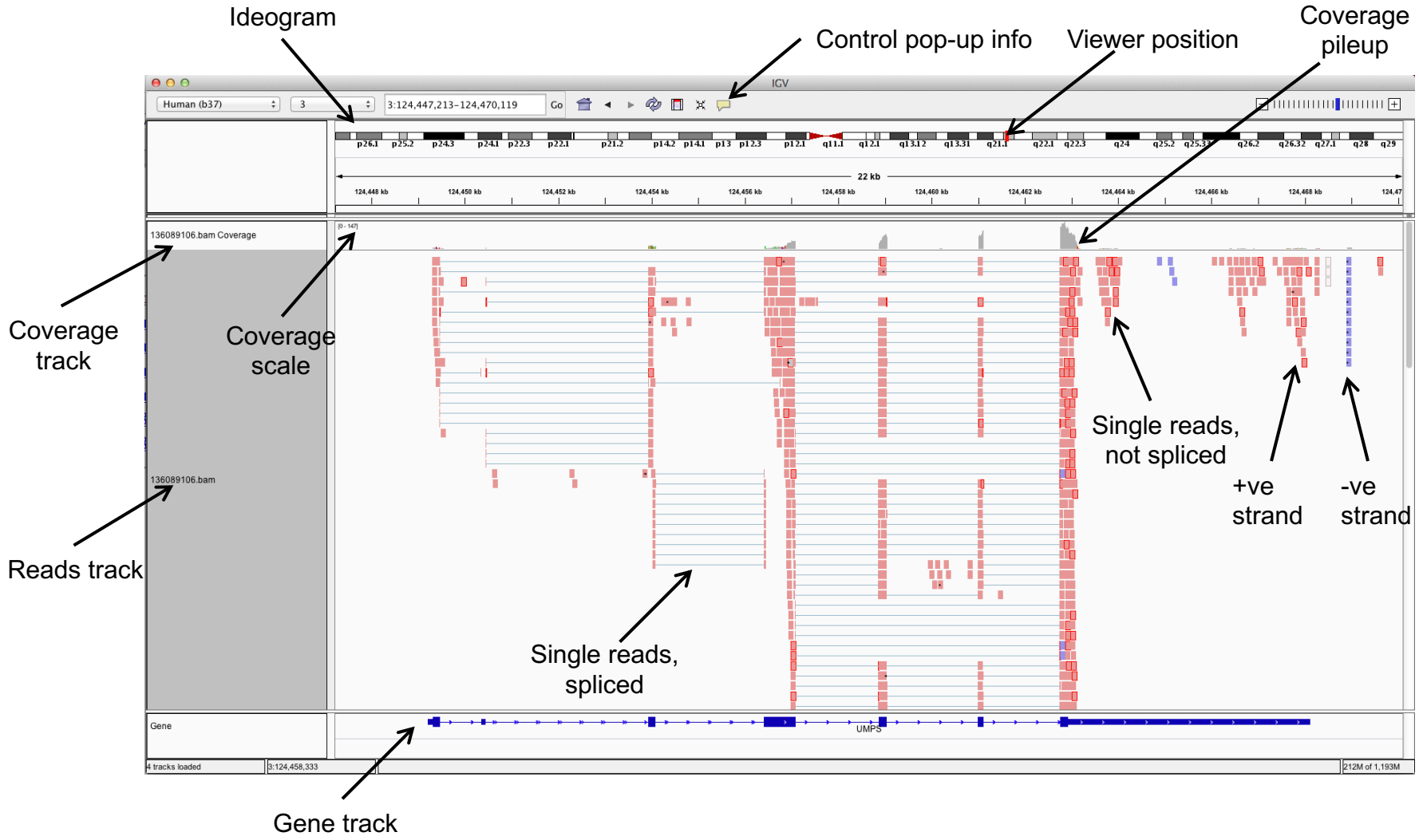


Washington University in St. Louis
School of Medicine

Learning objectives of module 3

- Visualization of RNA-seq alignments in IGV
- Alignment QC Assessment
- BAM read counting and determination of variant allele expression status

Visualization of RNA-seq alignments in IGV browser



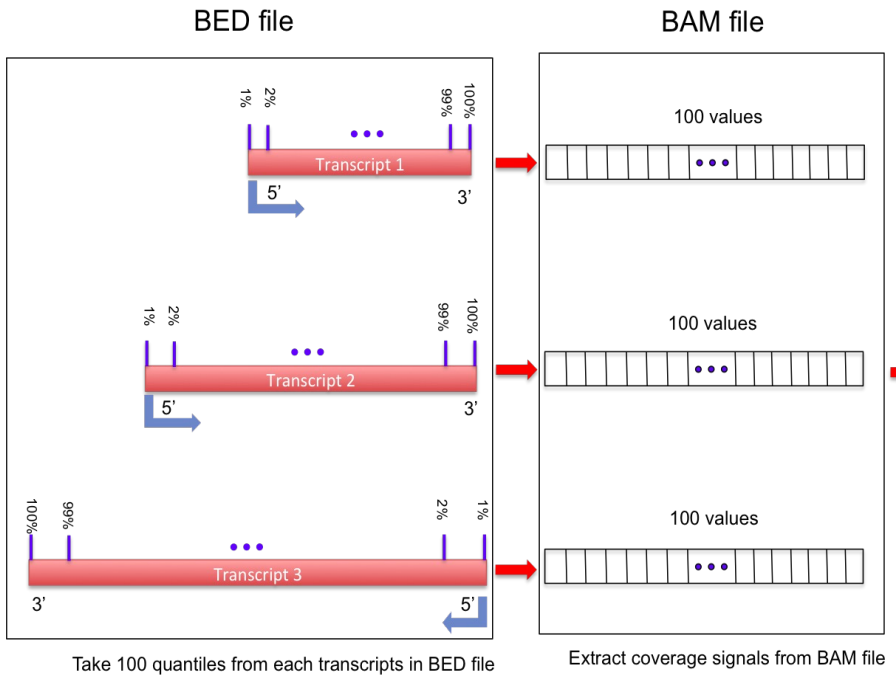
Alternative viewers to IGV

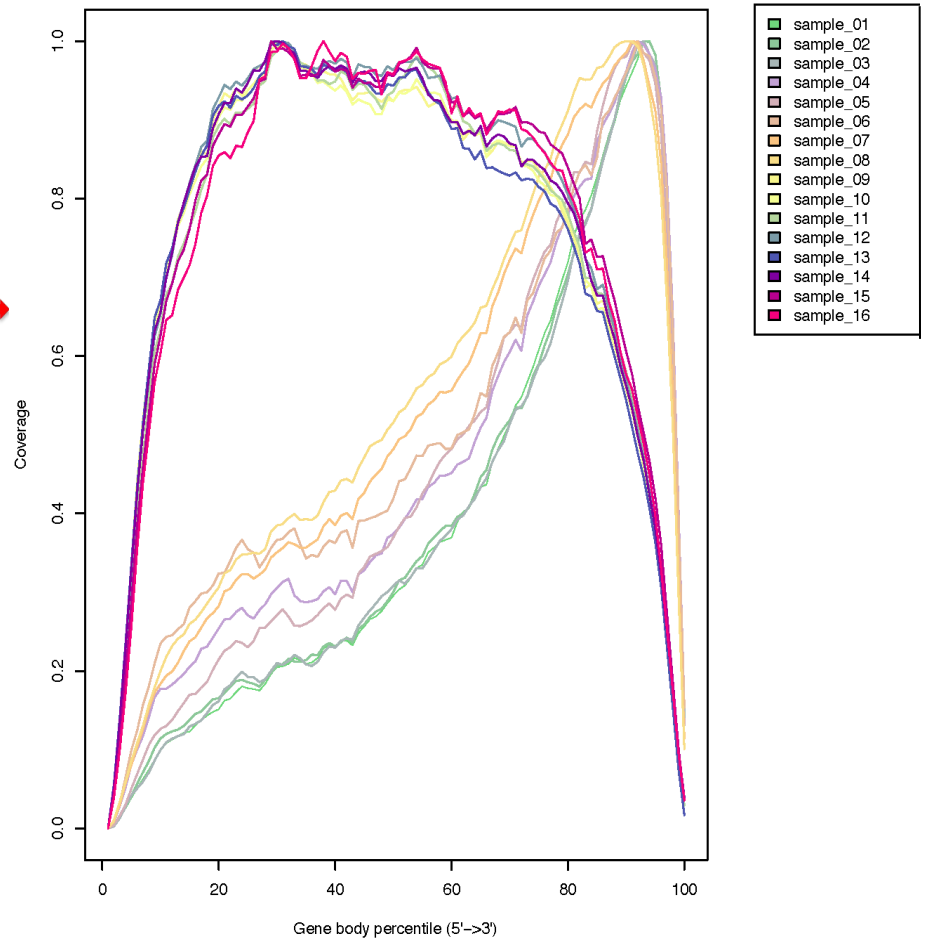
- Alternative viewers to IGV
  - http://www.biostars.org/p/12752/
  - http://www.biostars.org/p/71300/
- Artemis, BamView, Chipster, gbrowse2, GenoViewer, MagicViewer, **Savant**, Tablet, tview

Alignment QC Assessment

- 3' and 5' Bias
- Nucleotide Content
- Base/Read Quality
- PCR Artifact
- Sequencing Depth
- Base Distribution
- Insert Size Distribution

Alignment QC: 3' & 5' Bias

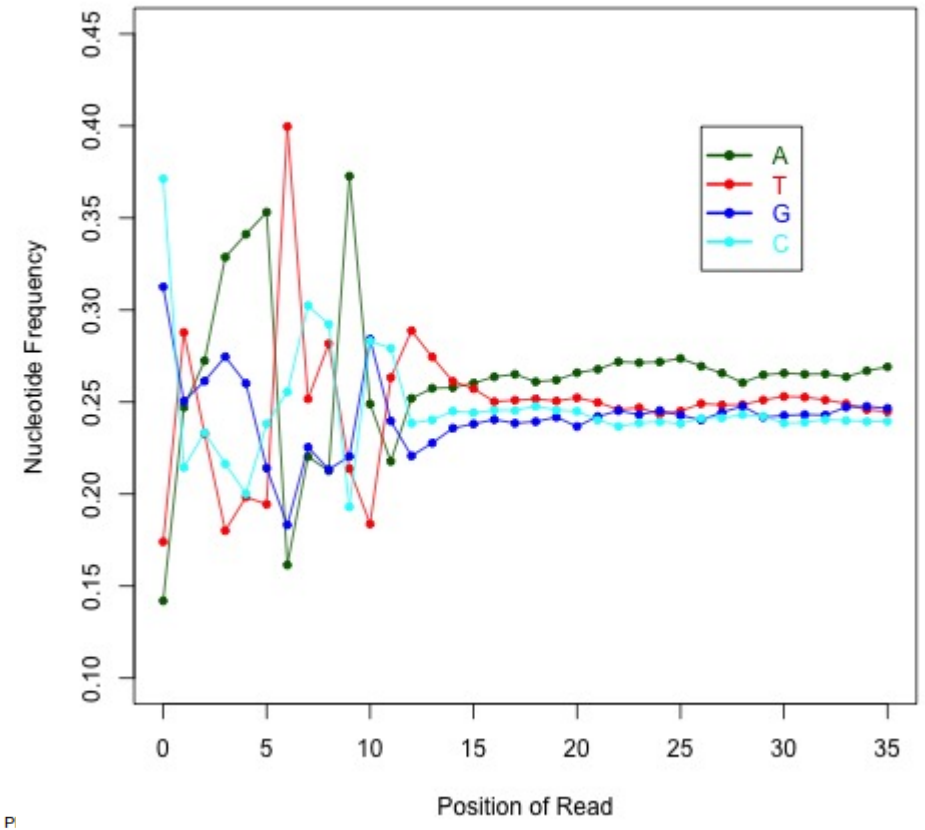




http://rseqc.sourceforge.net/

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%



Journal List › Nucleic Acids Res › v.38(12); 2010 Jul › PMC2896536

Nucleic Acids Research

Nucleic Acids Res. 2010 Jul; 38(12): e131.
Published online 2010 Apr 14. doi: 10.1093/nar/gkq224

Biases in Illumina transcriptome sequencing caused by random hexamer priming

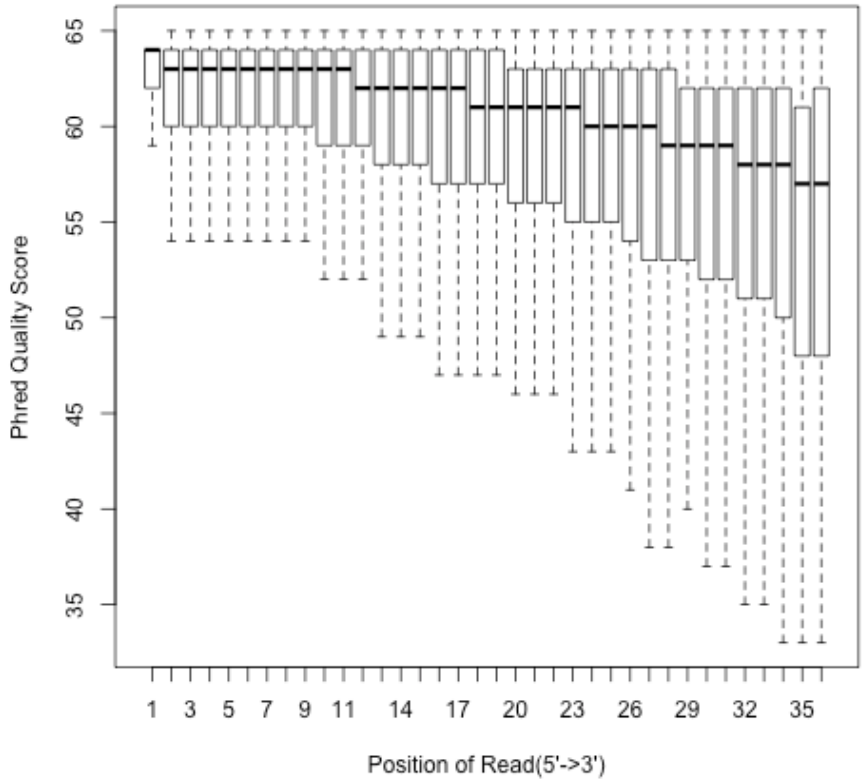
Kasper D. Hansen,[1,*] Steven E. Brenner,[2] and Sandrine Dudoit[1,3]

Author information ► Article notes ► Copyright and License information ►

http://rseqc.sourceforge.net/

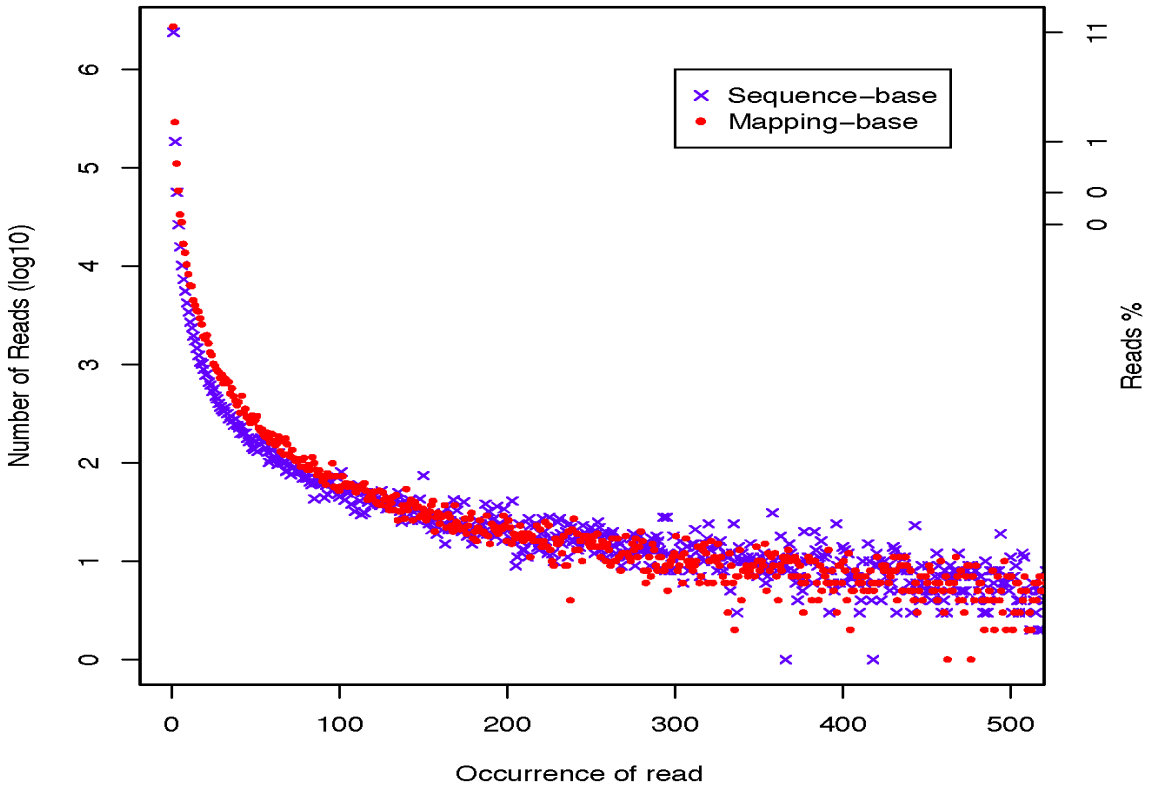
Alignment QC: Quality Distribution

- 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

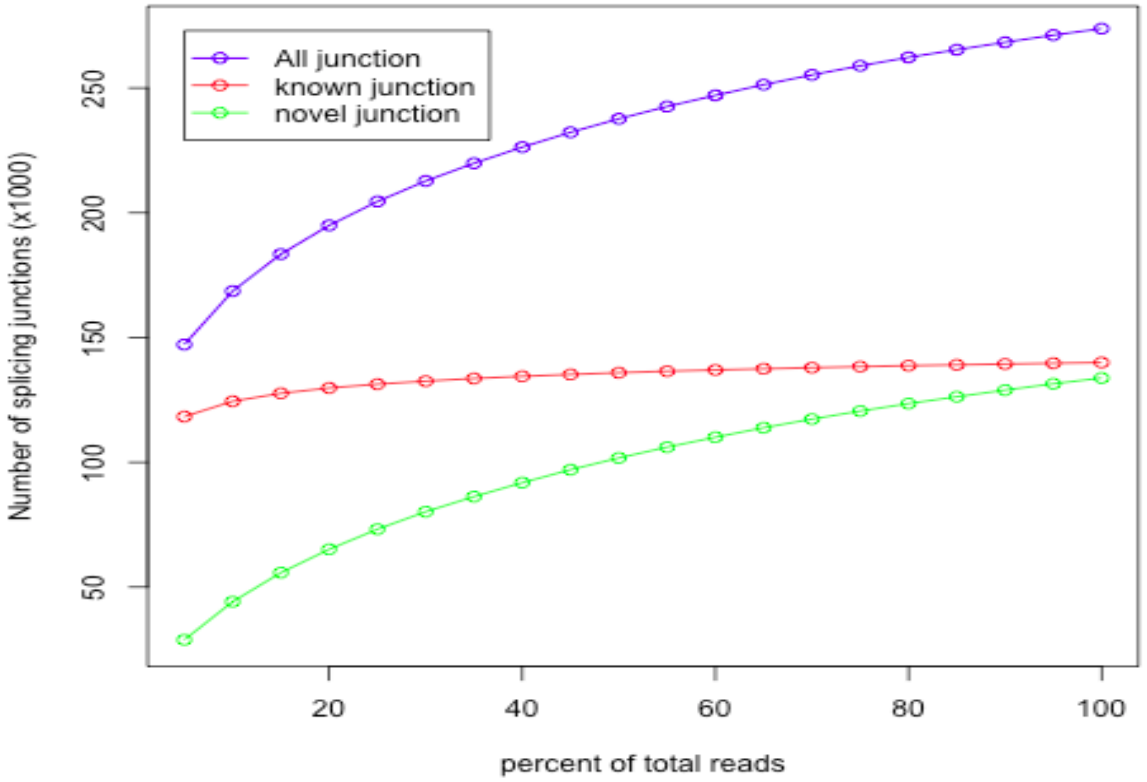
- Duplicate reads are reads that have the same start/end positions and same exact sequence
- In DNA-seq, reads/start point is used as a metric to assess PCR duplication rate
- In DNA-seq, duplicate reads are collapsed using tools such as picard
- How is RNA-seq different from DNA-seq?



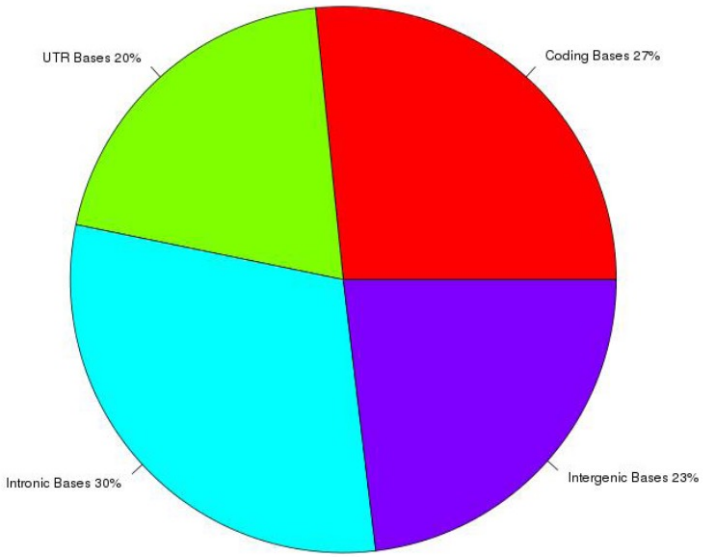
http://rseqc.sourceforge.net/

Alignment QC: Sequencing Depth

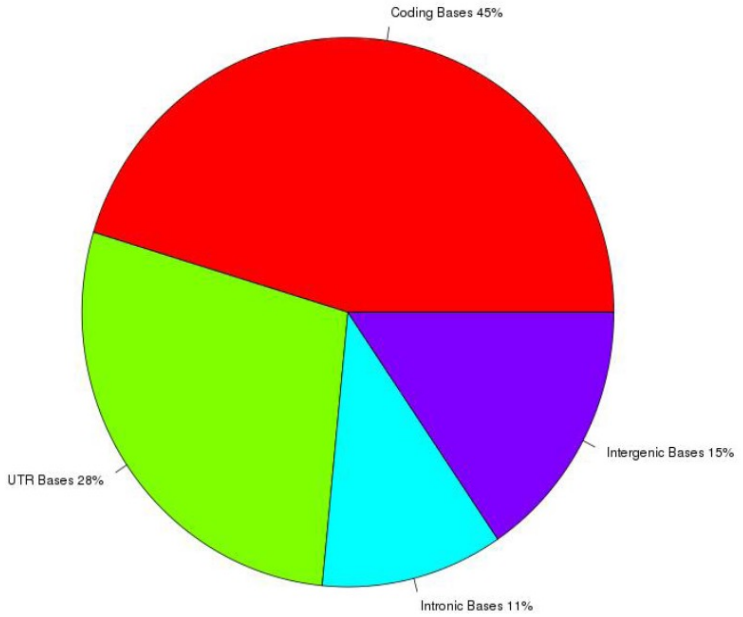
- **Have we sequenced deep enough?**
- In DNA-seq, we can determine this by looking at the average coverage over the sequenced region. Is it above a certain threshold?
- In RNA-seq, this is a challenge due to the variability in gene abundance
- Use splice junctions detection rate as a way to identify 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



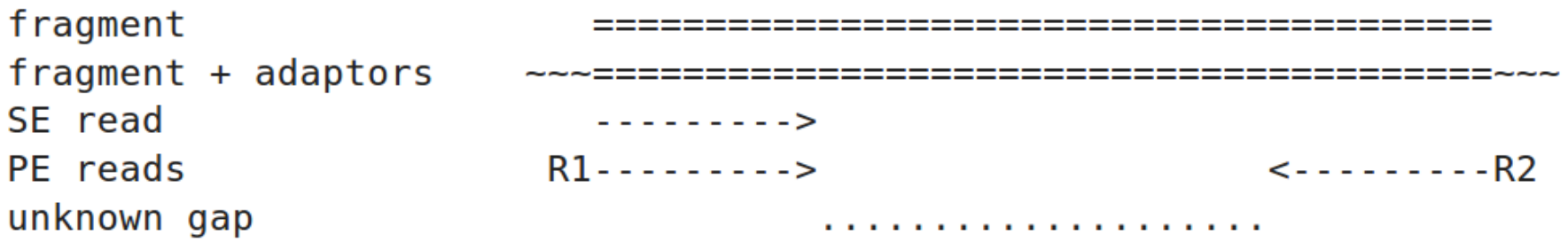
Whole Transcriptome Library

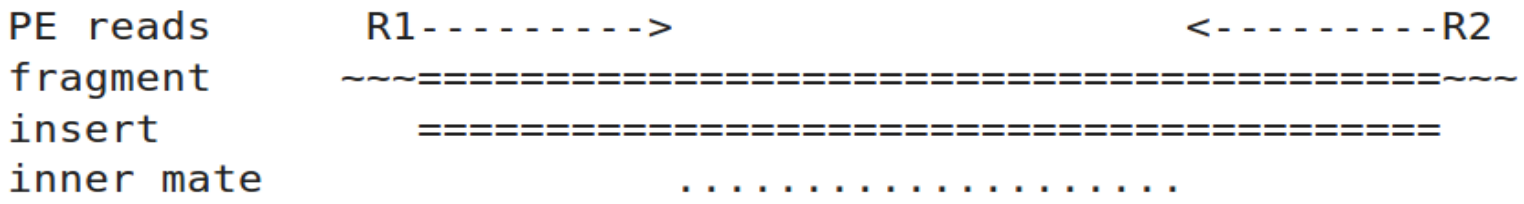


PolyA mRNA library

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

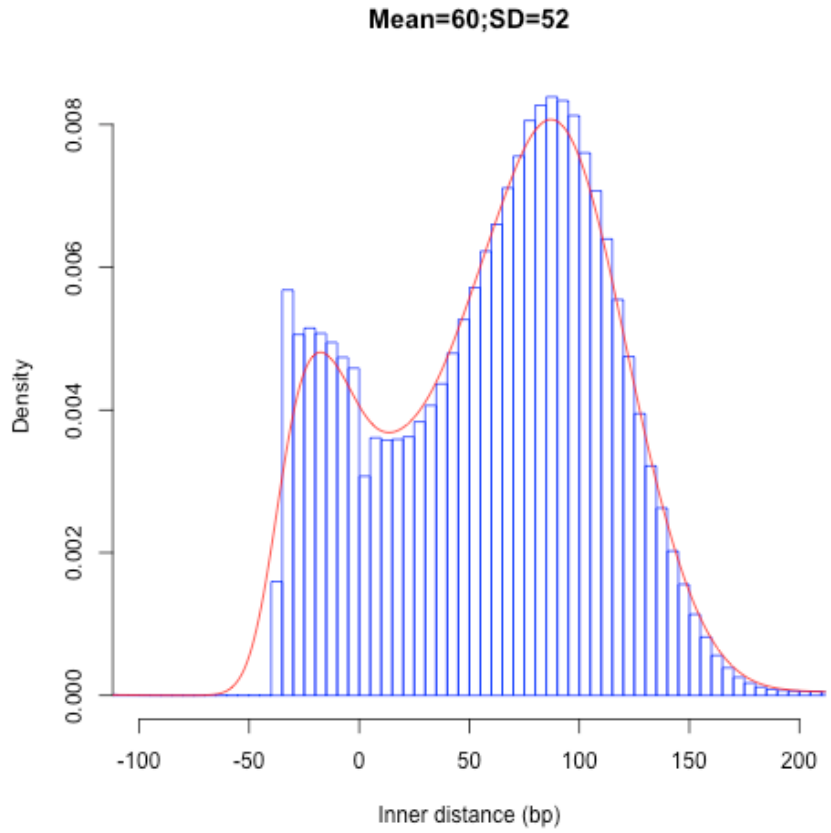
Alignment QC: Insert Size





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?

We are on a Coffee Break & Networking Session

Workshop Sponsors:

Canadian Centre for Computational Genomics