

# Canadian Bioinformatics Workshops

[www.bioinformatics.ca](http://www.bioinformatics.ca)

[bioinformaticsdotca.github.io](https://bioinformaticsdotca.github.io)

Creative Commons

This page is available in the following languages:

Afrikaans Български Català Dansk Deutsch Ελληνικά English (CA) English (US) Esperanto  
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)  
Euskara Suomiksi français français (CA) Galego മലബാറി ഹാവാസി Magyar Italiano 日本語 한국어 Macedonian Melayu  
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски (ћирилица) Sotho svenska  
中文 華語 (台灣) isiZulu



Attribution-Share Alike 2.5 Canada

You are free:



to Share — to copy, distribute and transmit the work



to Remix — to adapt the work



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.

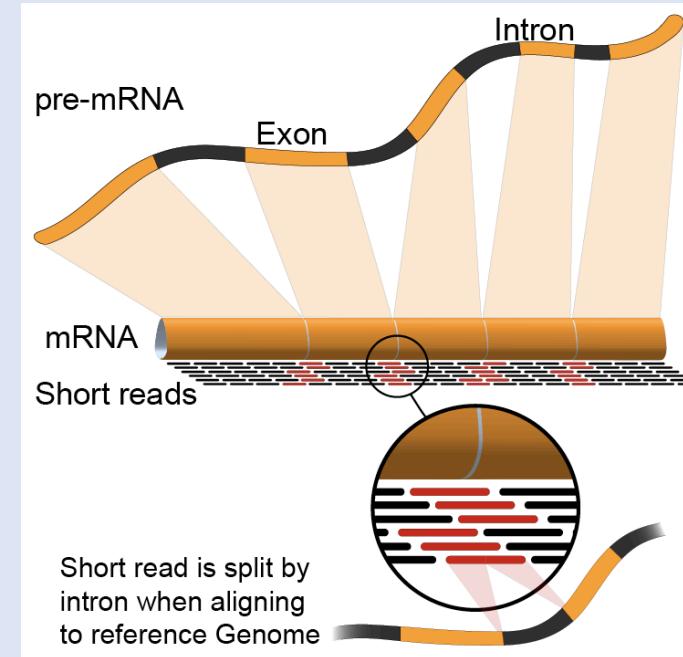
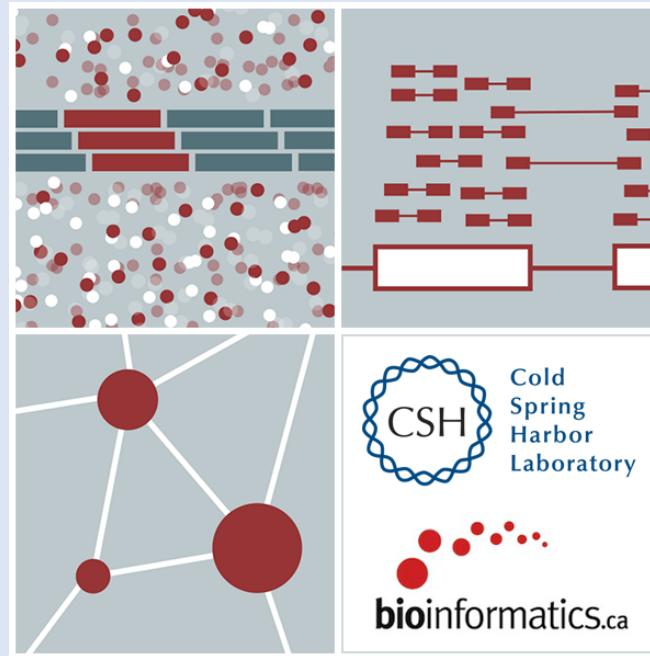
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:  
[English](#) [French](#)

[Learn how to distribute your work using this licence](#)

# RNA-Seq Module 2: Alignment QC



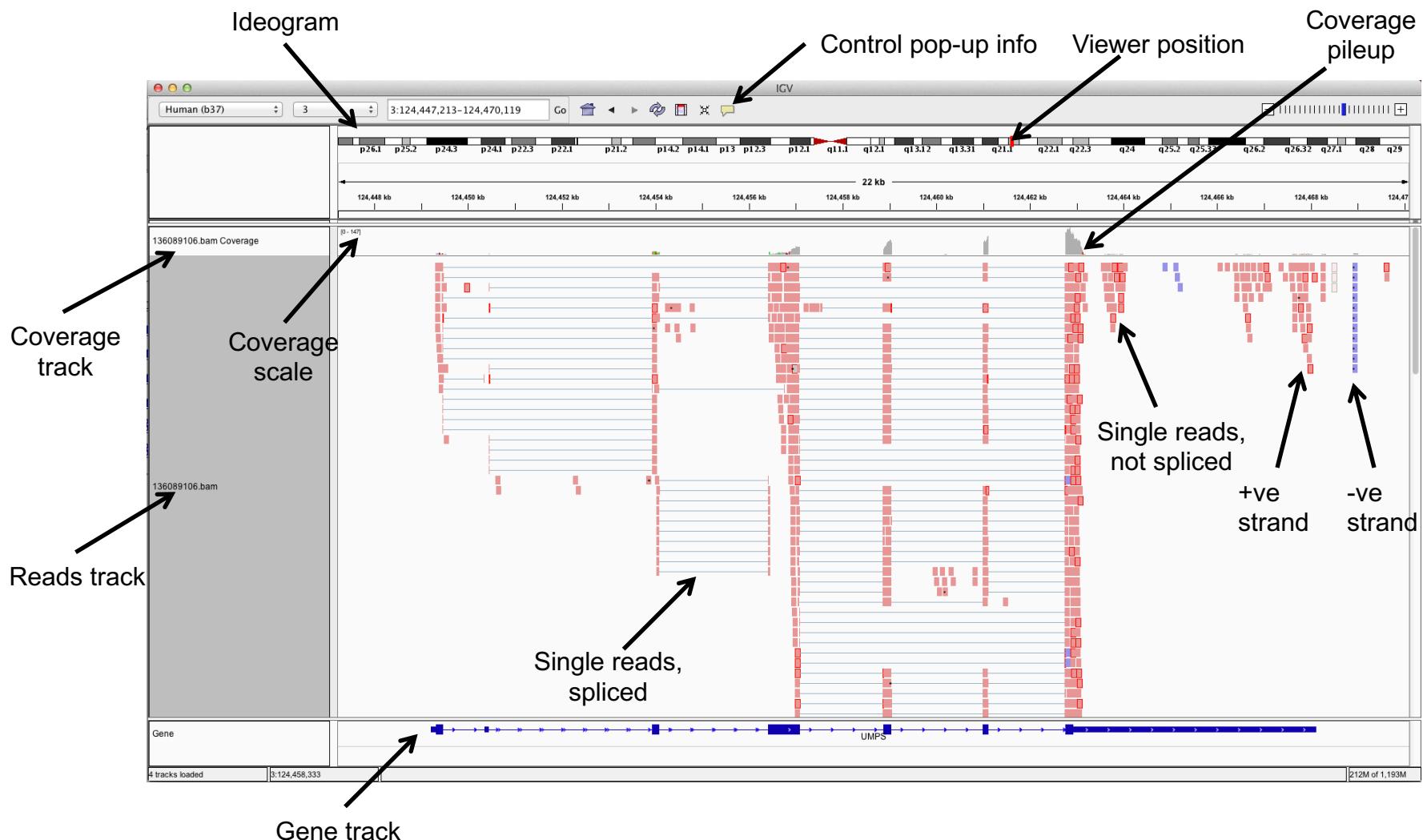
Obi Griffith and Malachi Griffith  
RNA-seq Analysis 2023. July 17-19, 2023



# 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



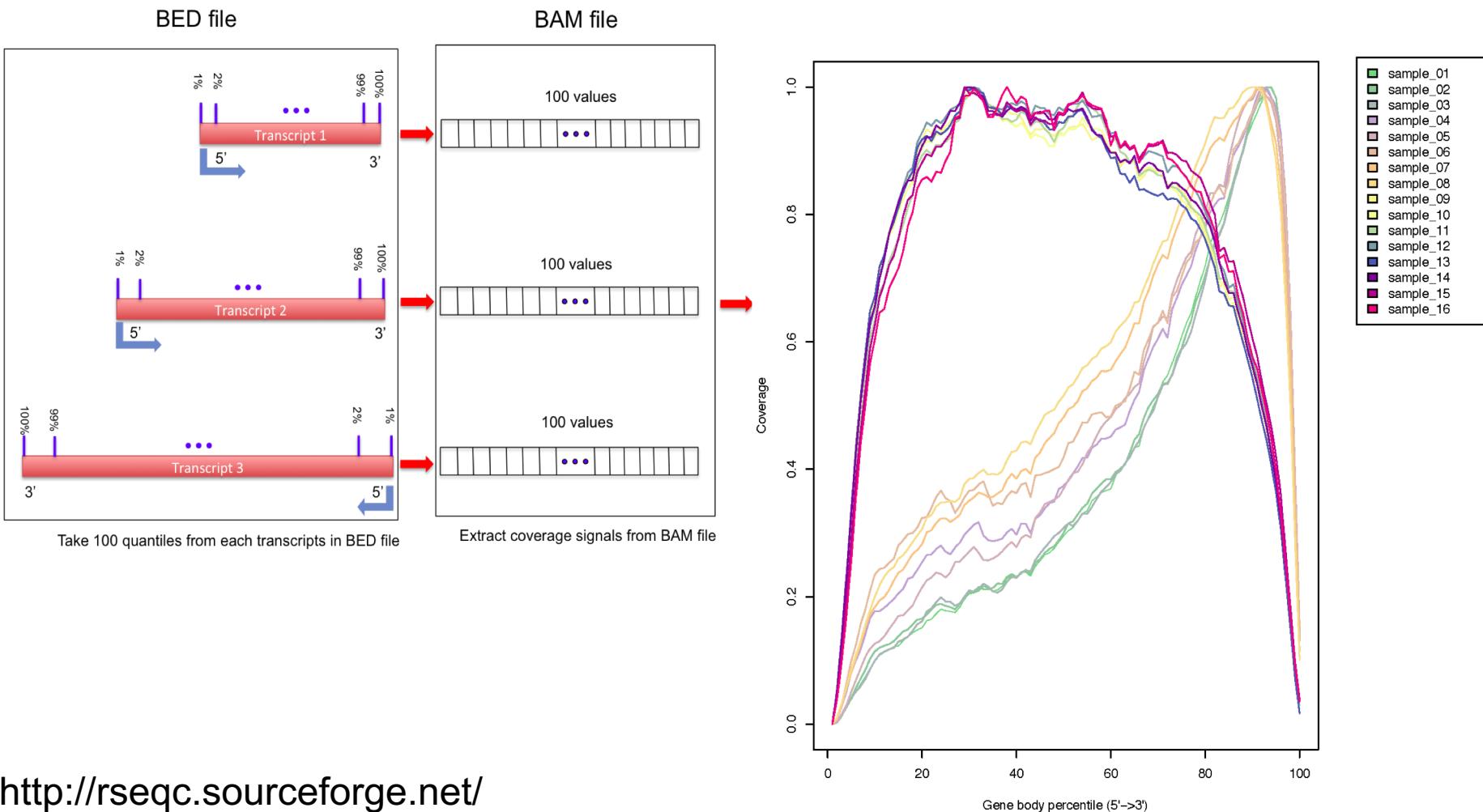
# Alternative viewers to IGV

- 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



# Alignment QC: Nucleotide Content

- Random primers are used to reverse transcribe RNA fragments into double-stranded complementary DNA (dsDNA)
- 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

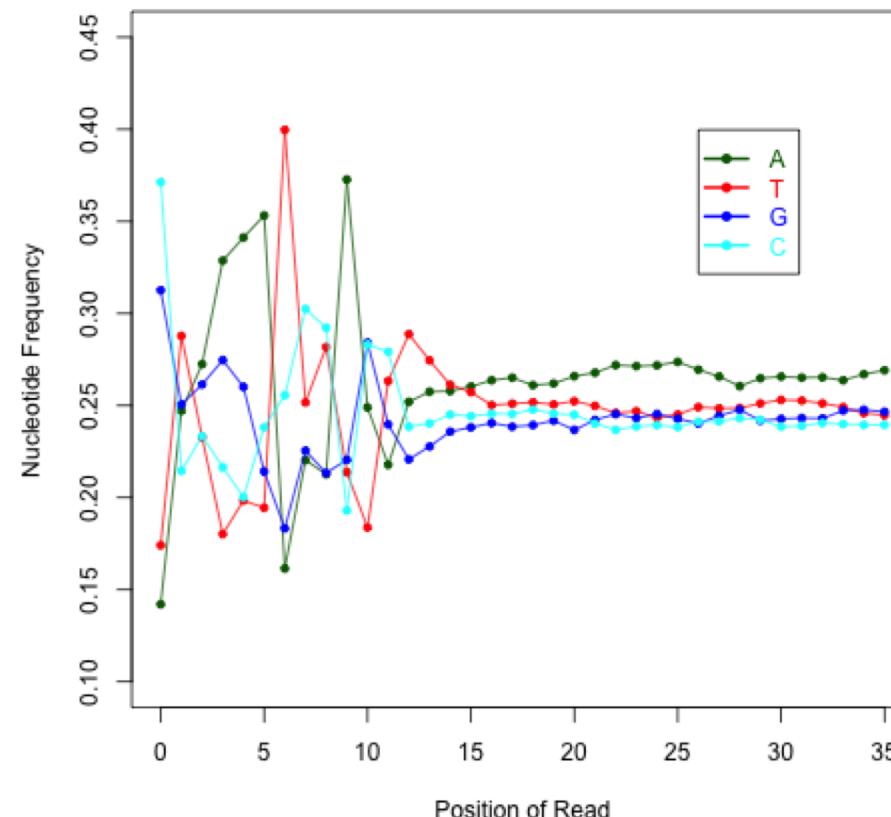
Nucleic Acids Res. 2010 Jul; 38(12): e131.  
Published online 2010 Apr 14. doi: [10.1093/nar/gkq224](https://doi.org/10.1093/nar/gkq224)

### Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen,<sup>1,\*</sup> Steven E. Brenner,<sup>2</sup> and Sandrine Dudoit<sup>1,3</sup>

[Author information ▾](#) [Article notes ▾](#) [Copyright and License information ▾](#)

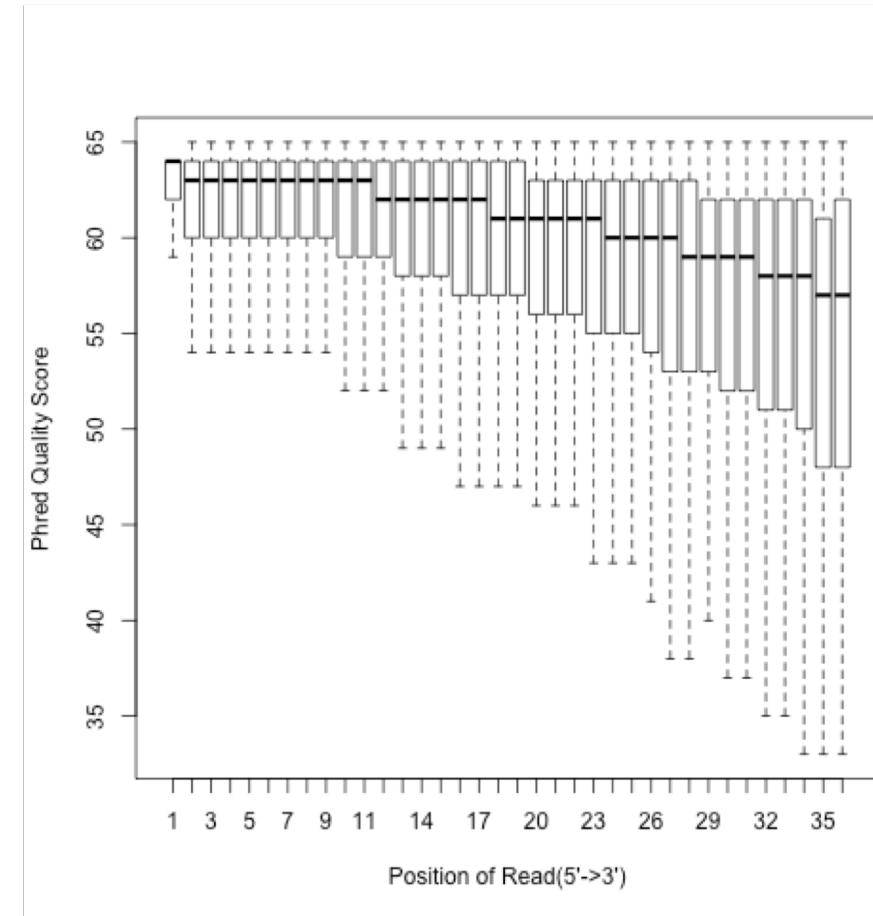
This article has been cited by other articles in PMC.



<http://rseqc.sourceforge.net/>

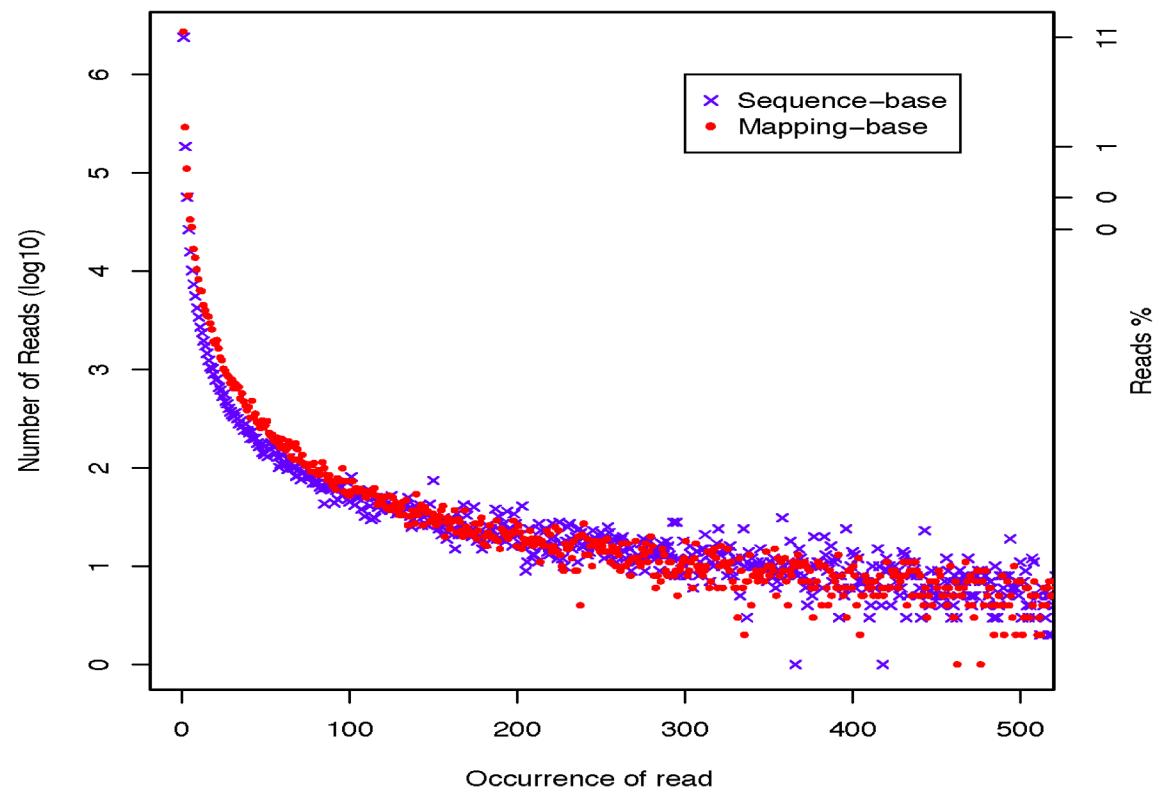
# Alignment QC: Quality Distribution

- Phred quality score is widely used to characterize the quality of base-calling
- Phred quality score =  $-10 \times \log(10)P$ , 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

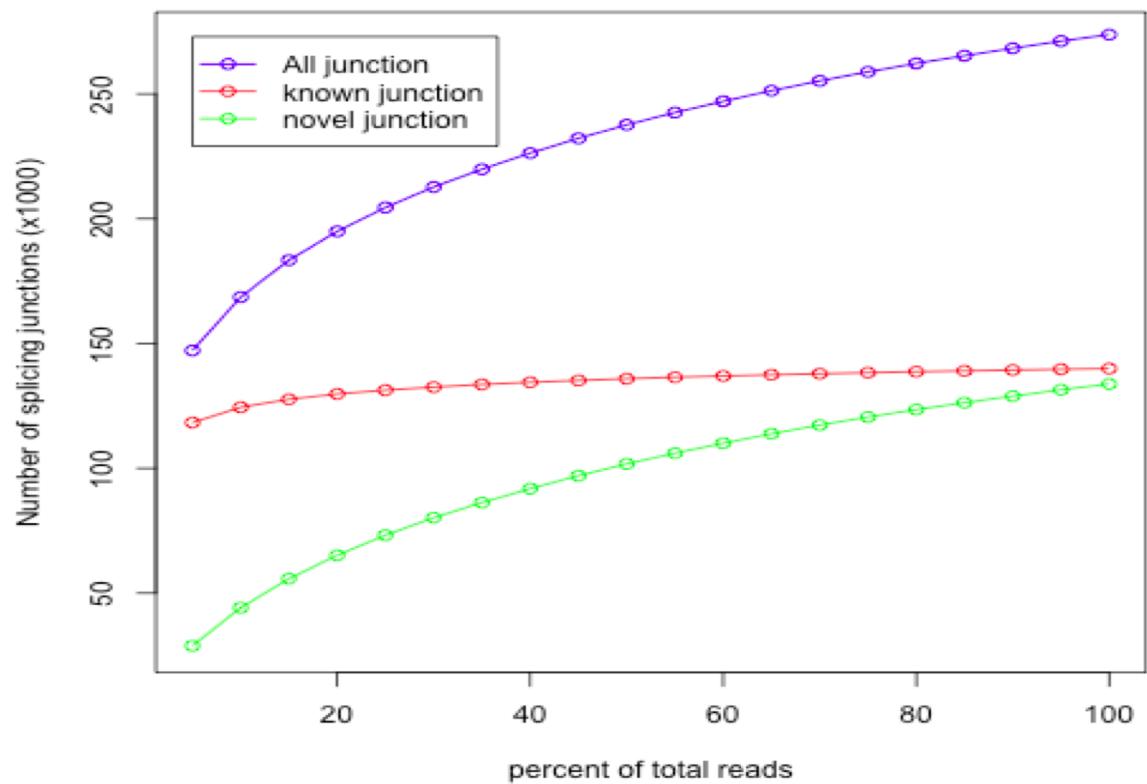
- Duplicate reads are reads that have the same start/end positions and same exact sequence
- In DNA-seq, reads/start point is used as a metric to assess PCR duplication rate
- In DNA-seq, duplicate reads are collapsed using tools such as picard
- How is RNA-seq different from DNA-seq?



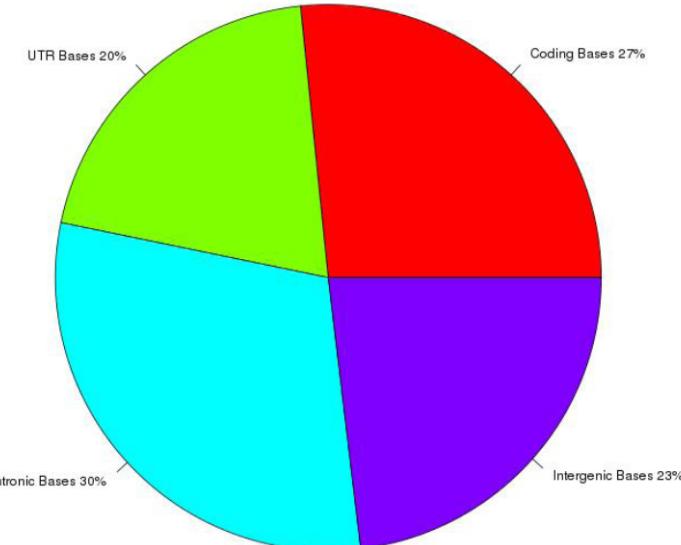
<http://rseqc.sourceforge.net/>

# Alignment QC: Sequencing Depth

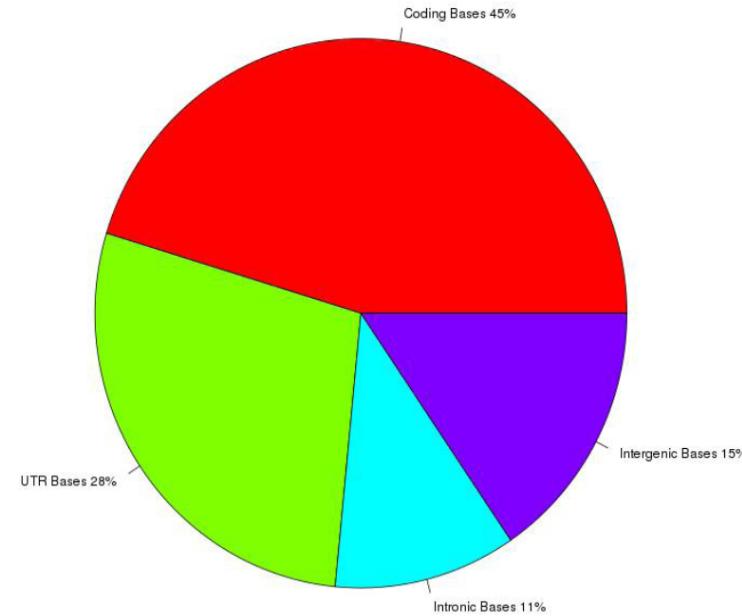
- **Have we sequenced deep enough?**
- In DNA-seq, we can determine this by looking at the average coverage over the sequenced region. Is it above a certain threshold?
- In RNA-seq, this is a challenge due to the variability in gene abundance
- Use splice junctions detection rate as a way to identify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



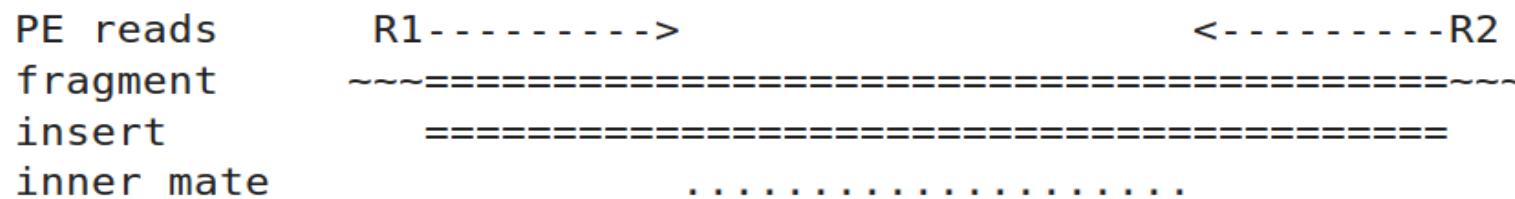
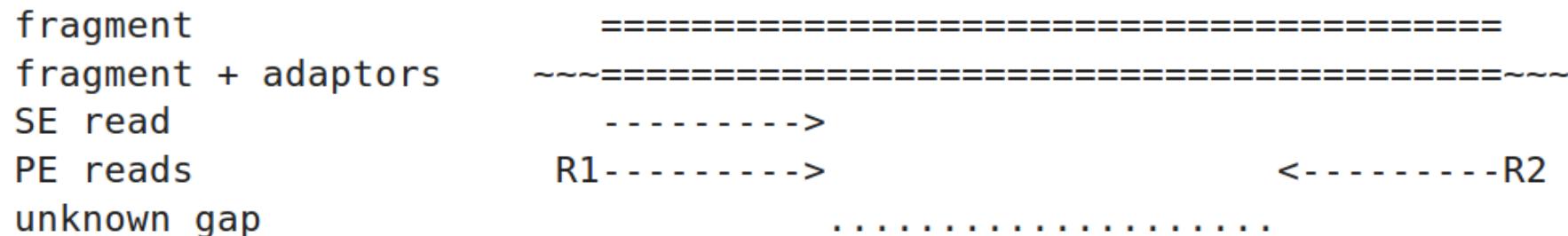
Whole Transcriptome Library



PolyA mRNA library

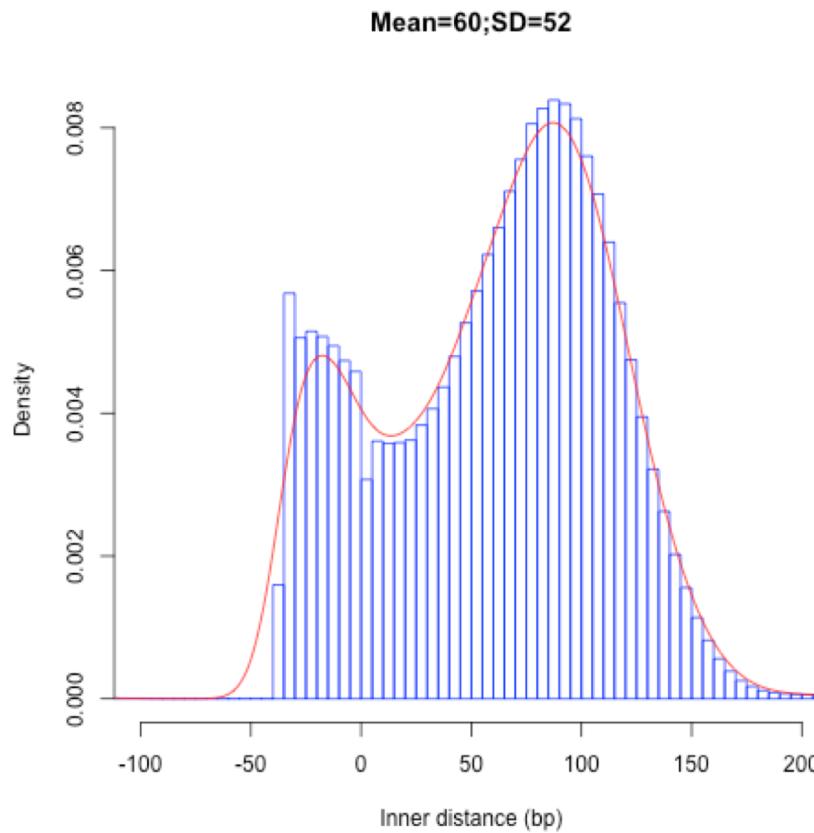
- Your sequenced bases distribution will depend on the library preparation protocol selected

# Alignment QC: Insert Size



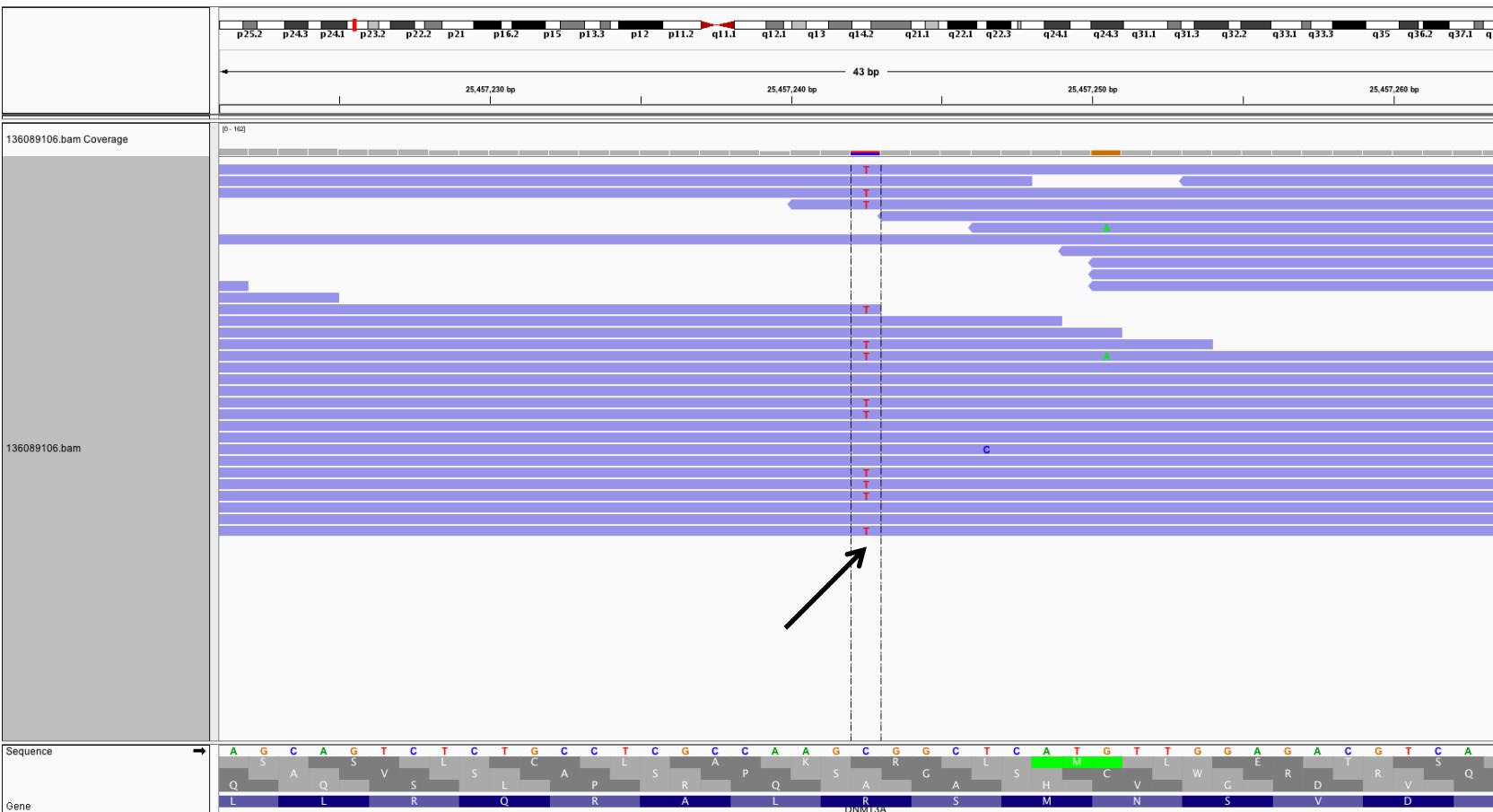
<http://thegenomefactory.blogspot.ca/2013/08/paired-end-read-confusion-library.html>

# Alignment QC: Insert Size



Consistent with library size selection?

# 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

