Introduction to RNASeq Data Analysis (Part 1)

Peter FitzGerald, PhD

Head Genome Analysis Unit
Director of BTEP
CCR, NCI

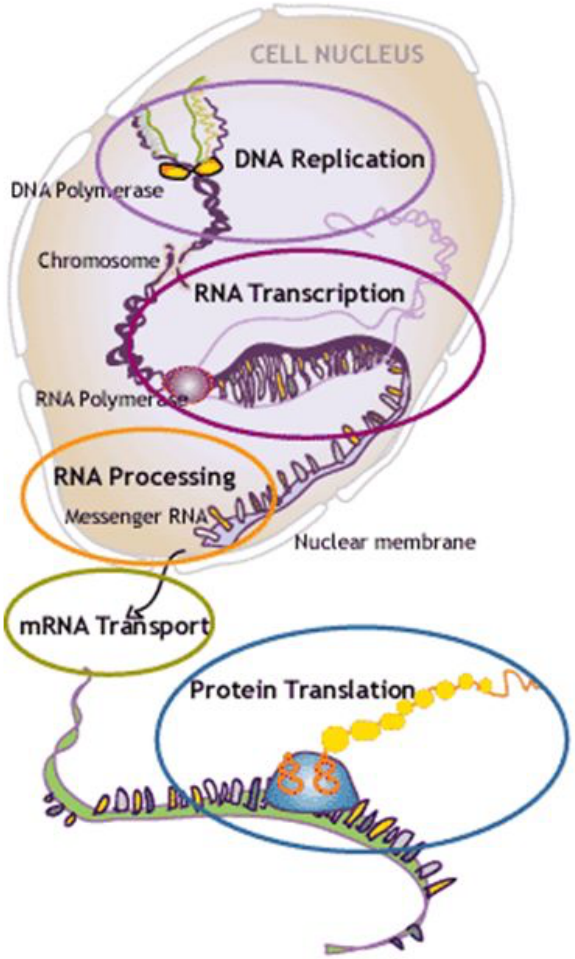
RNA-SEQ WEEKS

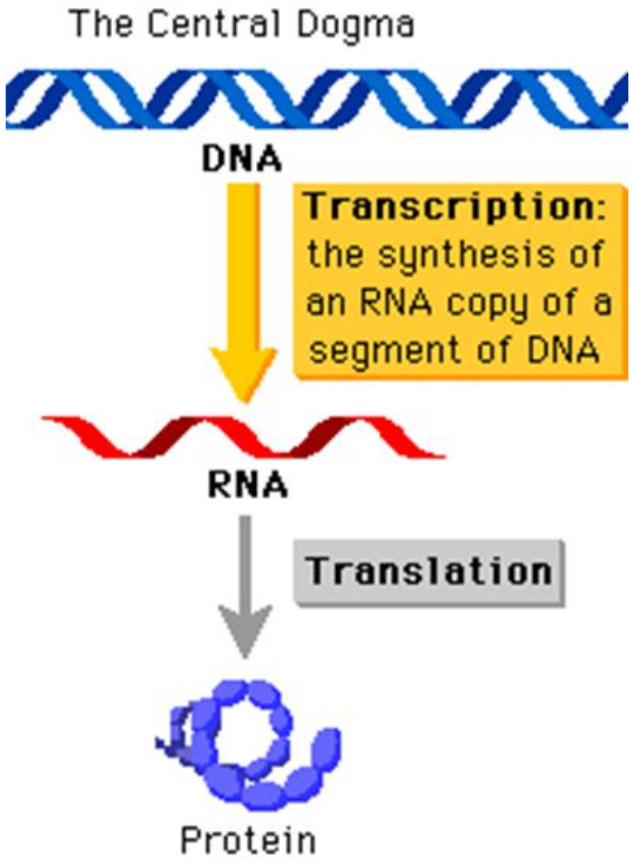
- Registration, Introduction to RNA-Seq Analysis, Part 1
  (Today - now)
- Registration, Introduction to RNA-Seq Analysis, Part 2
  Tuesday, Nov 10 @ 1 PM
- RNA-Seq Analysis on the DNAnexus platform
  Thursday, Dec 3 @ 1 PM
- CCBR-Pipeliner for analysis of RNA-Seq data
  Thursday, Dec 10 @ 1 PM
- Bulk RNA-Seq Analysis on the NIDAP platform
  Thursday, Dec 17 @ 1 PM

BTEP - Calendar **https://btep.ccr.cancer.gov**

RNA-SEQ WEEKS **https://btep.ccr.cancer.gov/rna-seq-weeks-coming-in-october/**

DNA → RNA → Protein





What is RNASEQ ?

RNA-Seq (**RNA sequencing**), uses next-generation sequencing (NGS) to reveal the presence and quantity of **RNA** in a biological sample at a given moment. (*Wikipedia*)

- Strictly speaking this could be any type of RNA (mRNA, rRNA, tRNA, snoRNA, miRNA) from any type of biological sample.
- For the purpose of this talk we will be limiting ourselves to **mRNA**.
- Technically, with a few exceptions, we are not actually sequencing **mRNA** but rather **cDNA**.

(RNASeq is only valid within the context of Differential Expression)

RNASEQ - WorkFlow

- **Experimental Design**
  - What question am I asking
  - How should I do it (*does it need to be done*)
- **Sample Preparation**
  - Sample Prep
  - Library Prep
  - Quality Assurance
- **Sequencing**
  - Technology/Platform
  - Detail Choices
- **Data Analysis (Computation)**

Generating the Data

Experimental Design
Sample Preparation
Sequencing

Only Sequence the RNA of interest

- Remember ~90% of RNA is ribosomal RNA
- Therefore enrich your total RNA sample by:
- polyA selection (oligodT affinity) of mRNA (eukaryote)
- rRNA depletion - RiboZero is typically used (costs extra)

Remember

- RNASEQ looks at steady state mRNA levels which is the sum of transcription and degradation
- Protein levels are assumed to be driven by mRNA levels
- RNASEQ can measure relative abundance not absolute abundance
- RNASEQ is really all about sequencing cDNA

What question(s) are you asking?

- Which gene are expressed?
- Which genes are differentially expressed?
- Are different splicing isoforms expressed?
- Are there novel genes or isoforms expressed?
- Are you interested in structural variants or SNPs, indels
- Are you interested in non-coding RNAs
- Does your interest lie in micro RNAs
- If this a standalone experiment, a pilot, or a “fishing trip”

Data Analysis Questions

- Where will the primary data be stored (fastq)?
- Where will the processed data be stored (bam)?
- Who will do the primary analysis?
- Who will do the secondary analysis?
- Where will the published data be deposited and by who? (what metadata will they require)
- Are you doing reproducible science?

***Talk** to the people who will be analyzing your data*
***BEFORE** doing the experiment*

Decissions, decisions, decisions!

- MiSeq
- NextSeq
- HiSeq
- NovaSeq
- PacBio
- OxfordNanopore

- Short Reads
- Long Reads
- Very Long Reads
- Very Very Long Reads

- Single End
- Paired End
- Stranded
- Unstranded

- mRNA
- rRNA
- miRNA

- Coding RNA
- non-Coding RNA
- Novel Genes
- Splice Variants
- Gene Fusions
- SNPs
- Structural Variants

Read Choices

- **Read Depth**
  - More depth needed for lowly expressed genes
  - Detecting low fold differences need more depth
- **Read Length**
  - The longer the length the more likely to map uniquely
  - Paired read help in mapping and junctions
- **Replicates**
  - Detecting subtle differences in expression needs more replicates
  - Detecting novel genes or alternate iso-forms need more replicates

Increasing depth, length, and/or replicates increase costs

Replicates

- **Technical Replicates**
  - It's generally accepted that they are not necessary because of the low technical variation in RNASeq experiments
- **Biological Replicates** (Always useful)
  - Not strictly needed for the identification of novel transcripts and transcriptome assembly.
  - Essential for differential expression analysis - must have 3+ for statistical analysis
  - Minimum number of replicates needed is variable and difficult to determine:

    3+ for cell lines

    5+ for inbred samples

    20+ for human samples (rarely possible)
  - More is always better

Best Practice Guidelines from Bioinformatic Core (CCBR):

1. Factor in at least 3 replicates (absolute minimum), but 4 if possible (optimum minimum). Biological replicates are recommended rather than technical replicates.

2. Always process your RNA extractions at the same time. Extractions done at different times lead to unwanted batch effects.

3. There are 2 major considerations for RNA-Seq libraries:

- If you are interested in coding mRNA, you can select to use the mRNA library prep. The recommended sequencing depth is between 10-20M paired-end (PE) reads. Your RNA has to be high quality (RIN > 8).
- If you are interested in long noncoding RNA as well, you can select the total RNA method, with sequencing depth ~25-60M PE reads. This is also an option if your RNA is degraded.

4. Ideally to avoid lane batch effects, all samples would need to be multiplexed together and run on the same lane. This may require an initial MiSeq run for library balancing. Additional lanes can be run if more sequencing depth is needed.

5. If you are unable to process all your RNA samples together and need to process them in batches, make sure that replicates for each condition are in each batch so that the batch effects can be measured and removed bioinformatically.

6. For sequence depth and machine requirements, visit Illumina Sequencing Coverage website

For cost estimates, visit Sequencing Facility pricing for NGS

For further assistance in planning your RNA-Seq experiment or to discuss specifics of your project, please contact us by email: ***CCBR@mail.nih.gov*** *OR visit us during office hours on Fridays 10am to noon (Bldg37/Room3041). For cost and specific information about setting up an RNA-Seq experiment, please visit the Sequencing Facility website or contact Bao Tran*

Sample Preparation

Costs (mRNA total)

CCR Sequencing Facility (subsidized pricing**)**

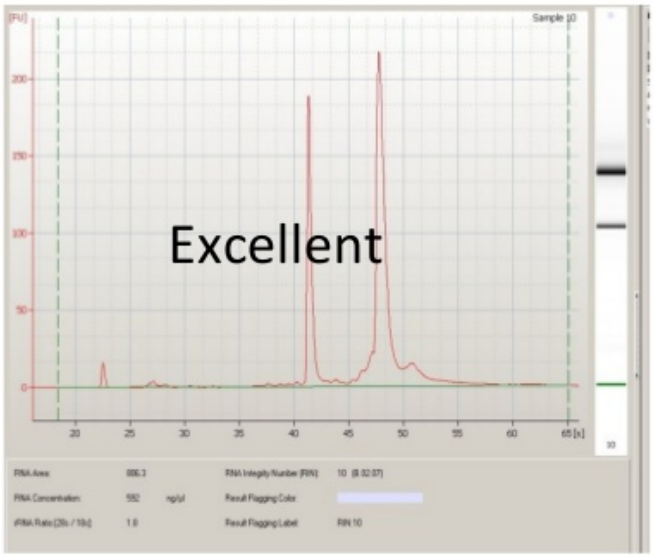
Library Construction	$87	
Illumina HiSeq 4000	$1007/lane	PE 2 x 75 (all 8 lanes)
Illumina NovaSeq	$4382/lane	1 x 100 bp
Illumina NextSeq High Output	$1956	2 x 75 bp (V2)
Illumina MiSeq	$623	PE 2 x 75 bp (V3)

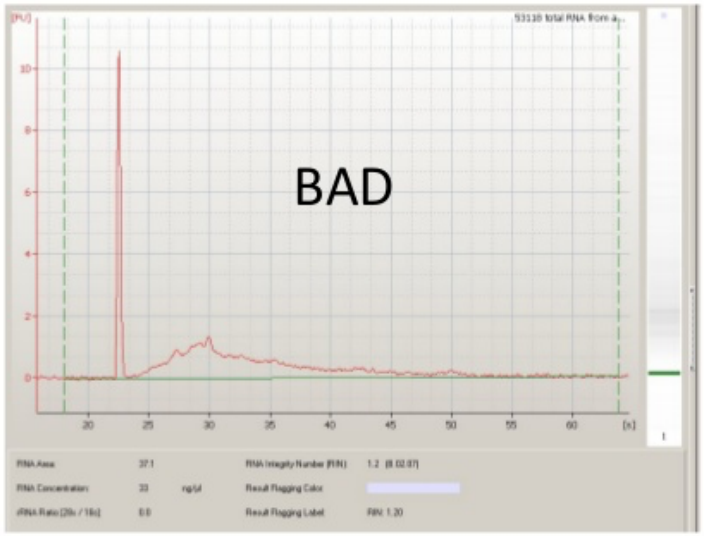
https://ostr.ccr.cancer.gov/resources/sequencing-facility/

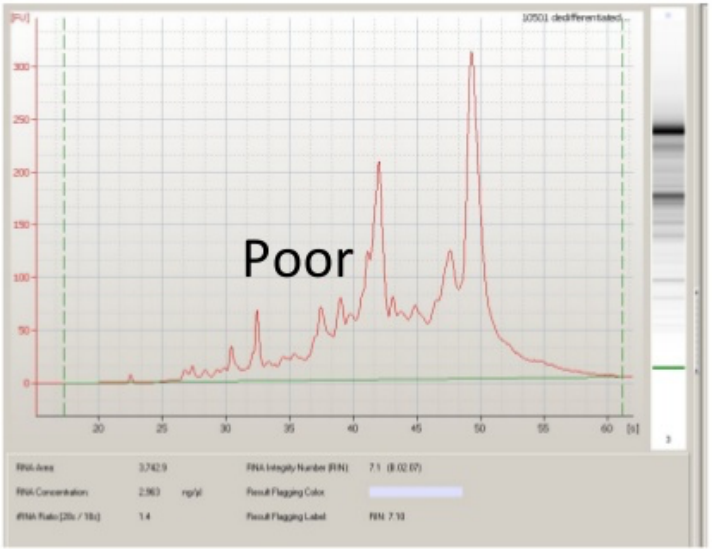
General Rules for Sample Preparation

- Prepare all samples at the same time or as close as possible. The same person should prepare all samples
- Do not prepare “experiment” and “control” samples on different days or by different people. (Batch effects).
- Use high quality means to determine sample quality (**R**NA **I**ntegrity **N**umber) (RIN >0.8) and quantity, and size (Tapestation, Qibit, Bioanalyzer)
- Don’t assume everything will work the first time (do pilot experiments) or every time (prepare extra samples)

Determining Library size distribution







CENTER FOR CANCER COMPUTATIONAL BIOLOGY

Sample Amounts

Type of Library	Minimum DNA/RNA Requirement for Library Construction	Recommended DNA/RNA for Optimal Library Construction	Maximum Sample Volume Requirement for Library Construction	Additional Requirements
mRNA Sequencing	100 ng	1 µg	50 µL	RIN should be at least 8.0, DNase treated
mRNA ultralow Clonetech	100 pg	10 ng	10 µL	RIN should be at least 8.0, DNase treated
microRNA Sequencing	100 ng	1 µg	6 µL	
Total RNA Sequencing	100 ng	1 µg	10 µL	DNase treated, FFPE and degraded RNA can be used
Total RNA ultralow	10 ng	1 µg	10 µL	DNase treated, FFPE and degraded RNA can be used

RNA-Seq Sample Recommendations (CCBR)

QC Metric Guidelines	mRNA	total RNA
RNA Type(s)	Coding	Coding + non-coding
RIN	8 [low RIN = 3' bias]	> 8
Single-end vs Paired-end	Paired-end	Paired-end
Recommended Sequencing Depth	10-20M PE reads	25-60M PE reads
FastQC	Q30 > 70%	Q30 > 70%
Percent Aligned to Reference	70%	> 65%
Million Reads Aligned Reference	7M PE reads (or > 14M reads)	16.5M PE reads (or > 33M reads)
Percent Aligned to rRNA	< 5%	< 15%
Picard RNAseqMetrics	Coding > 50%	Coding > 35%
Picard RNAseqMetrics	Intronic + Intergenic < 25%	Intronic + Intergenic < 40%

Sequencing

Illumina Sequencing Platforms

Illumina
Sequencing by Synthesis (SbS)
/NovaSeq/HiSeq/NextSeq/MiSeq
Short read length (50 to 300 bp)

Selection driven by cost, precision, speed, number of samples and number of reads required

Consult with the Sequencing Core

Illumina
NovaSeq

Illumina
NextSeq

Illumina
MiSeq

Long Read Sequencing Platforms

PacBio
120,000 bases per molecule, with maximum read lengths > 200,000 bases. Good for repetitive regions and isomers, modified bases.

Oxford Nanopore
Direct DNA or RNA sequencing (Max length 2 Mb) Good for modified bases, repetitive regions, isomers, small genomes.

Consult with the Sequencing Cores

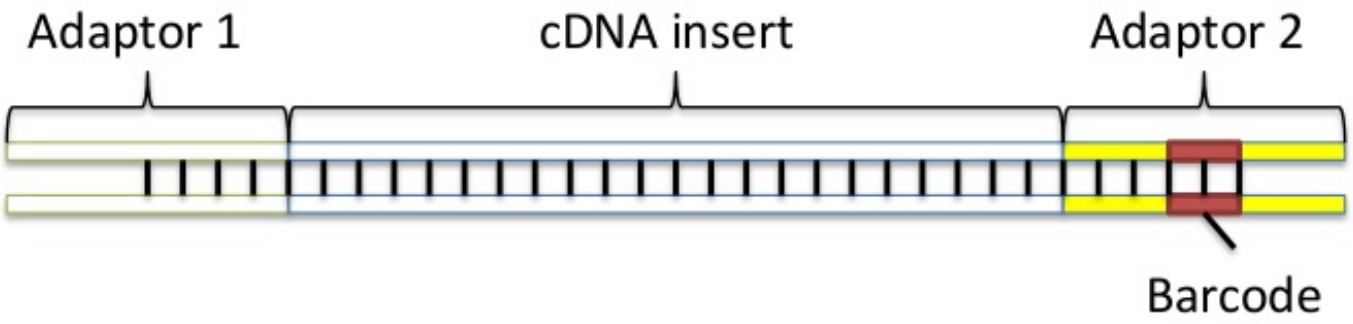
PacBio Sequel II

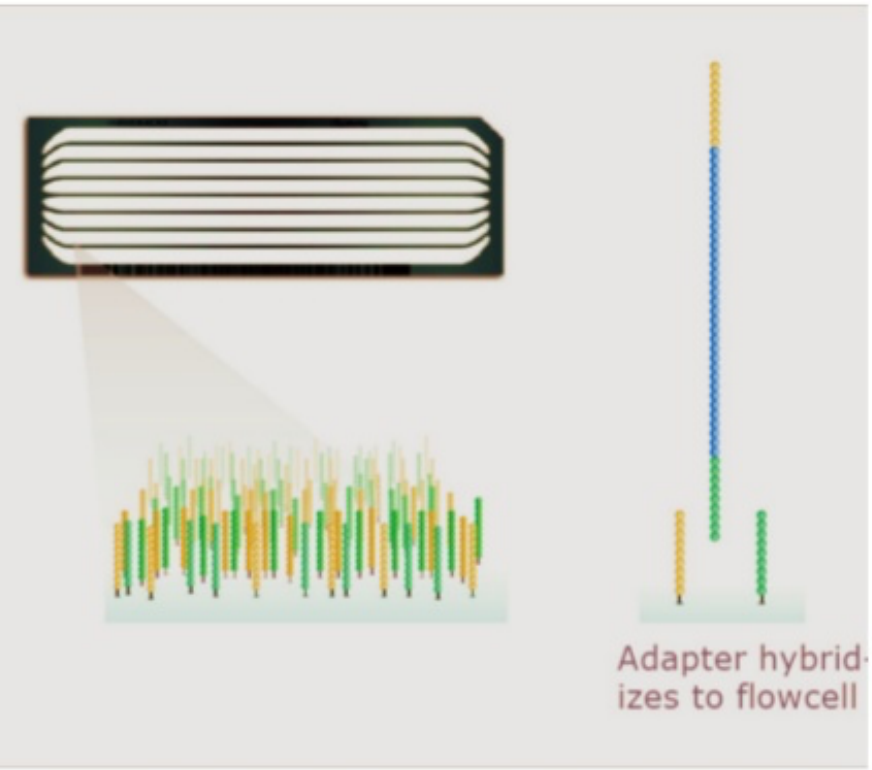
MinION

GridION

Oxford Nanopore

Sequencing Library Structure



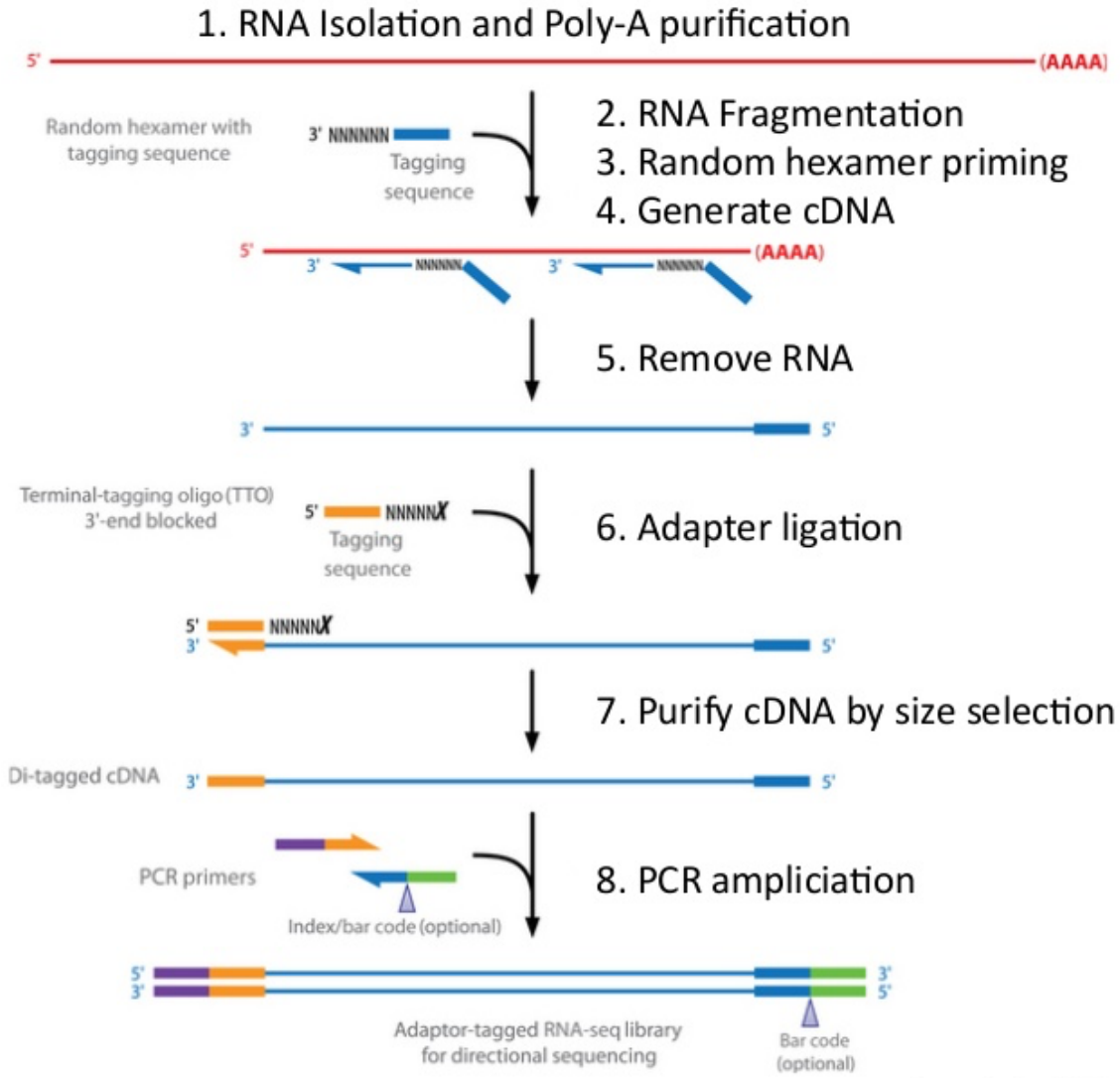


Adaptor – 58 bp nucleotide sequence to fix sequence library onto flow cell

Barcode – optional index sequence that is typically 6 nucleotide bases long for associating sequence with a particular sample (can be present on both adaptor)

cDNA insert – fragmented cDNA sequence generated from mRNA of interest. The insert typically range between 300-500bp for mRNA

Illumina SBS RNASeq

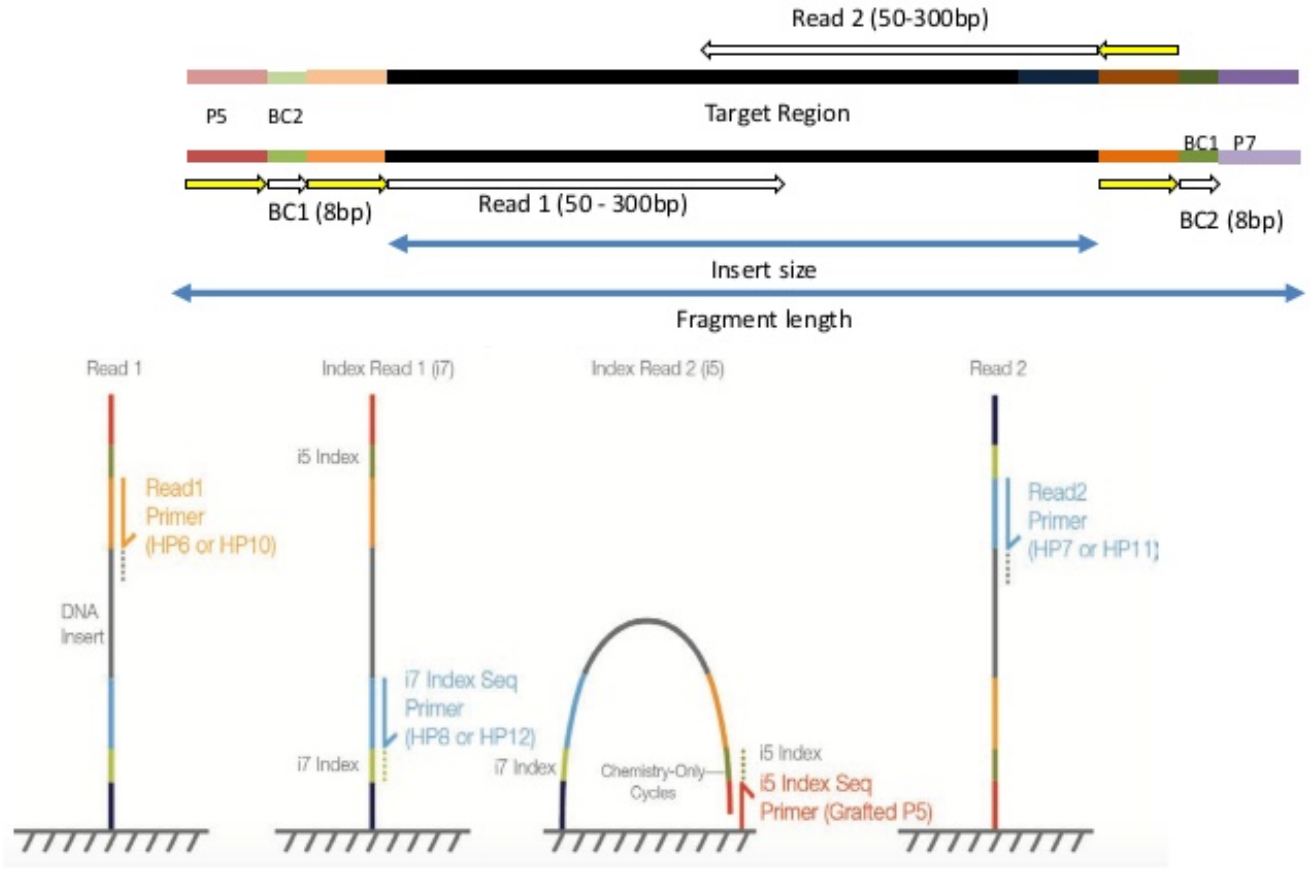


Pease, Nature Methods 9 (2012)

UCDAVIS Bioinformatics Core

Illumina sequencing
sequencing by synthesis

[Illumina SBS]



Data Analysis

RNASEQ - Data Analysis WorkFlow

- **Quality Control**
  - **Sample Cleanup**
  - **Trimming**
- **Alignment/Mapping**
  - Reference Target (Sequence and annotation)
  - Alignment Program
  - Alignment Parameters
  - Post-Alignment Quality Assurance
- **Quantification**
  - Counting Method
  - Counting Parameters
- **Visualization**

Computational Considerations
THE GOOD NEWS

For the most part the computational aspects have been taken care of for you.
(no need to develop new algorithms or code).

There are pre-built workflows that can automate many of the processes involved, and facilitate reproducability

Computational Considerations
THE BAD NEWS

Like most of NGS data analysis, the complexity of RNA-Seq data analysis revolves around data and information management and the dealing with "expected issues.

Consider the simplest experiment

(*Two conditions three replicates*)

6-12 fastq starting files

6-12 quality control files

6-12 fastq files post trimming of adaptors

6 bam file, and 6 bam index files

6 gene count files

36-48 files minimum (big files)

Computational Considerations
The Challenges

There is no single **best method** for RNA-Seq data analysis - it depends on your definition of best, and even then it varies over time and with the particular goals and specifics of a given experiment

It's for this reason that you should learn enough about the process to make "sensible choices" and to know when the results are reasonable and correct.

Treating an RNA-Seq (or any NGS) analysis as a black box is a "recipe for disaster" (*or at least bad science*). That's not to say that you need to know the particulars of every algorithm involved in a workflow, but you should know the steps involved and what assumptions and/or limitations are build into the whole workflow

Computational Prerequisites

- High performance Linux computer (multi core, high memory, and plenty of storage)
- Familiarity with the “command line” and at least one programming/scripting language.
- Basic knowledge of how to install software
- Basic knowledge of R and/or statistical programming
- Basic knowledge of Statistics and model building

Data Analysis

Pre-alignment QC & cleanup

Alignment

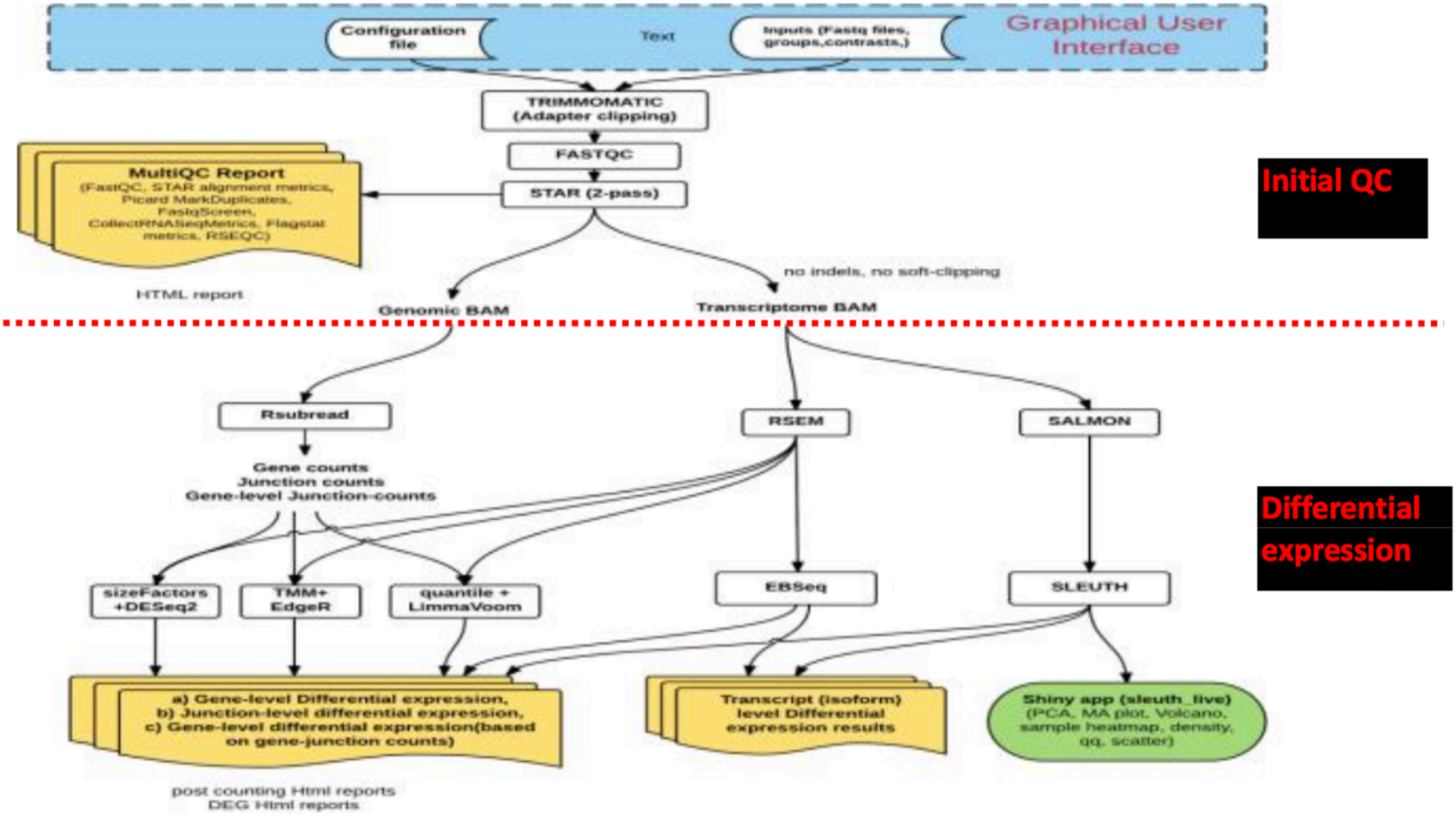
Post-alignment QC & filtering

Quantification

Differential Expression

RNASEQ Pipeline

https://github.com/CCBR/Pipeliner/blob/master/RNASeqDocumentation.pdf



Quality Control/Assesment (Pre-Alignment)

Data Quality Assessment

- **Evaluate the read quality to determine**

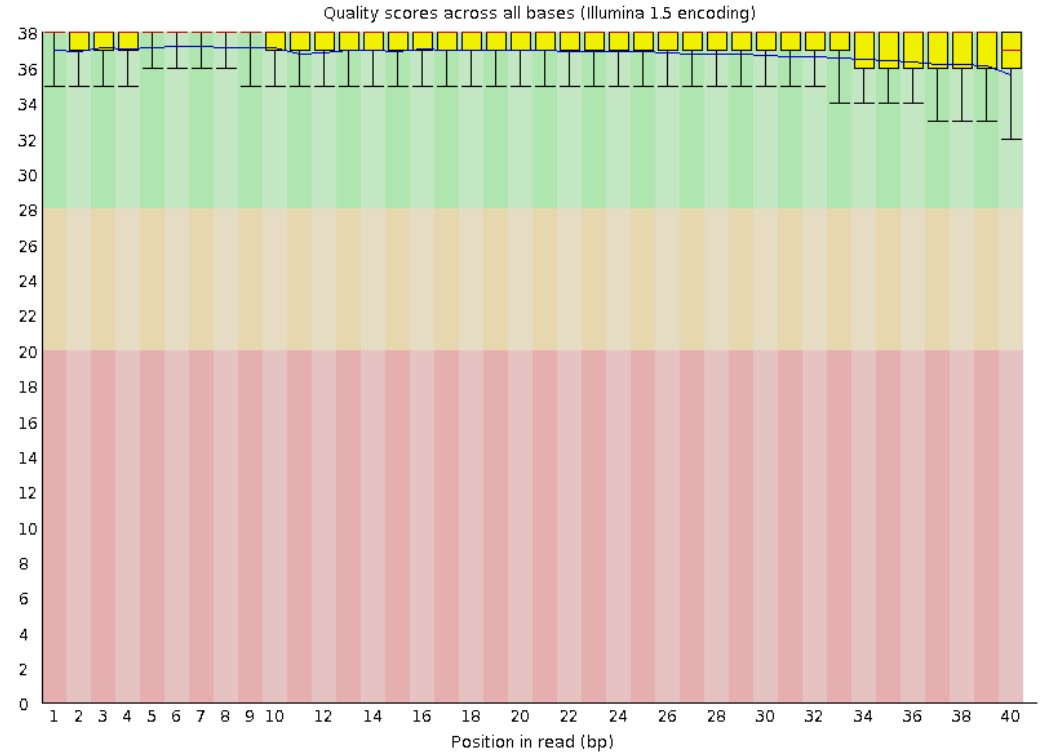
  (Tells us nothing about whether the experiment worked)

  - Is the data of sufficiently high quality to be analyzed?
  - Are there technical artifacts?
  - Are there poor quality samples?
- **Evaluate the following features**
  - Overall sequencing quality scores and distributions
  - GC content distribution
  - Presence of adapter or contamination
  - Sequence duplication levels
- **Data should be filtered, trimmed, or rejected as appropriate**

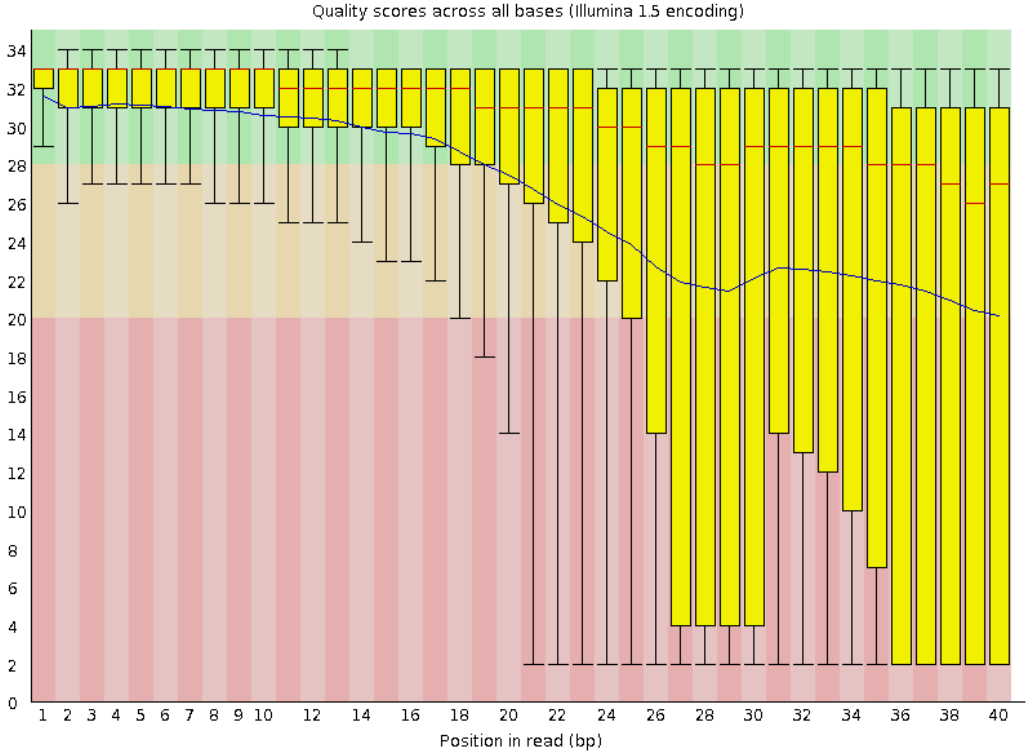
Sequencing cores generally provide some/all of this analysis

FastQC

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
good_sequence_short_fastqc.html



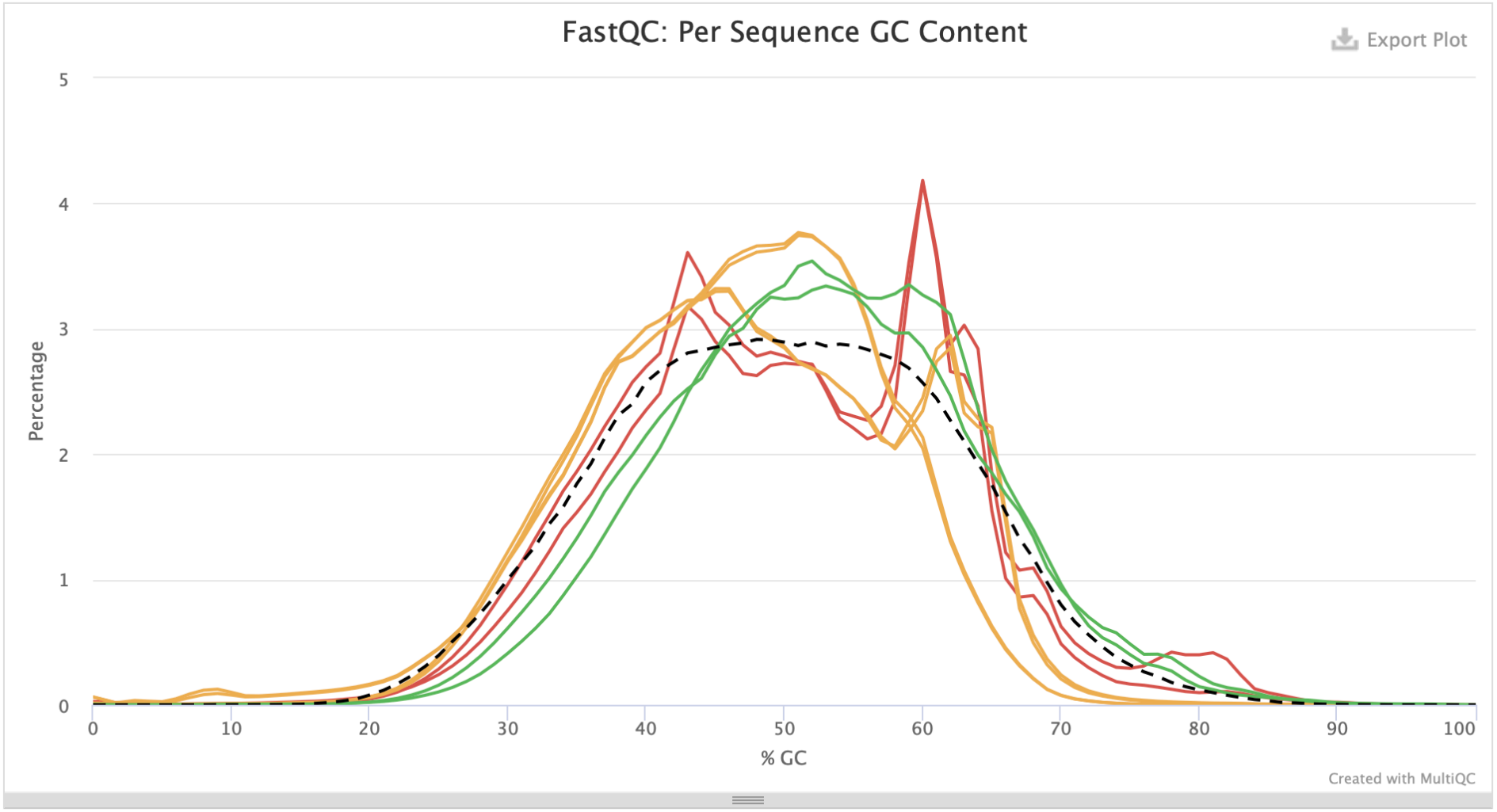
GOOD



BAD

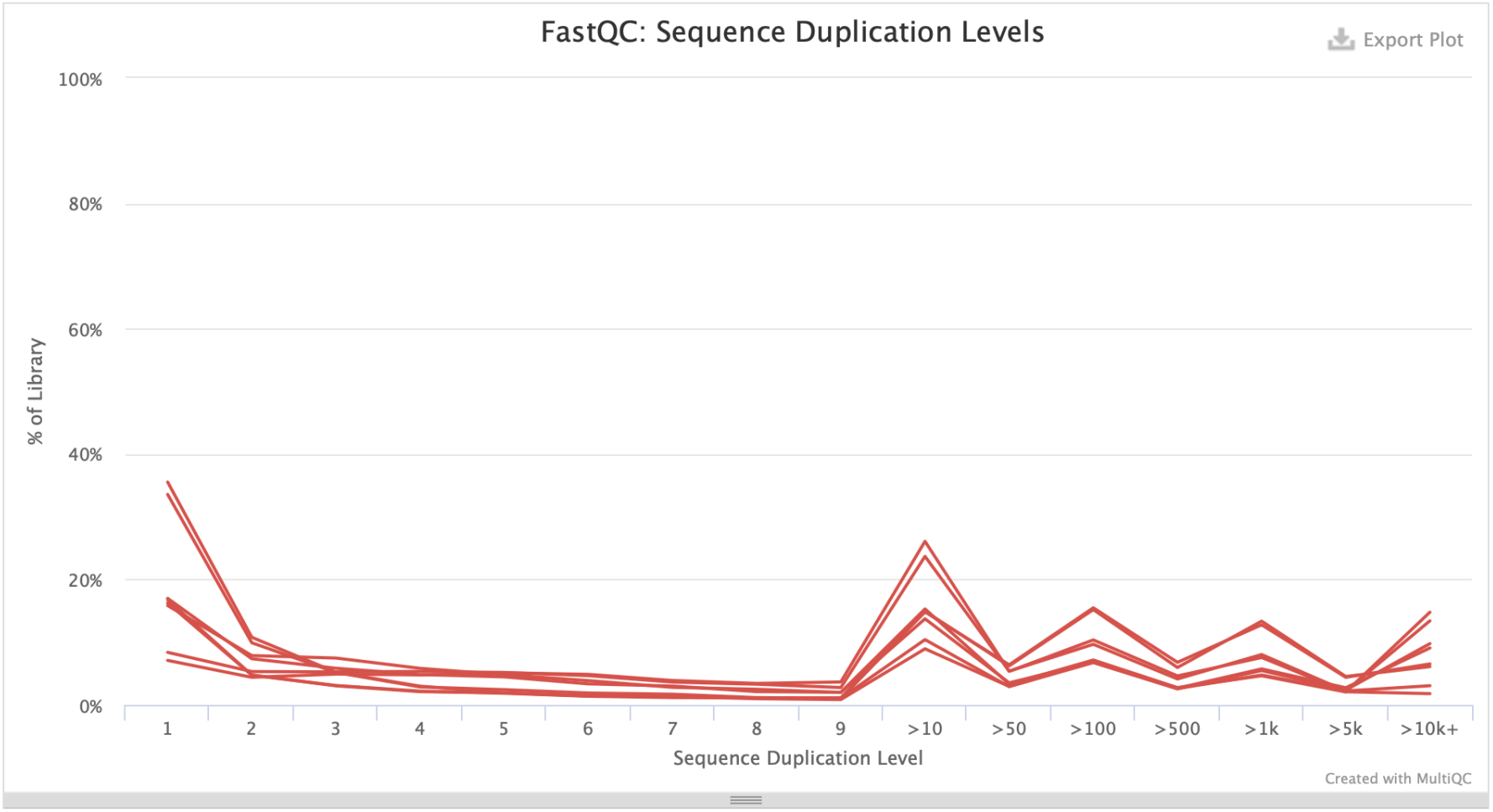
MultiQC

https://multiqc.info/examples/rna-seq/multiqc_report.html



MultiQC

https://multiqc.info/examples/rna-seq/multiqc_report.html



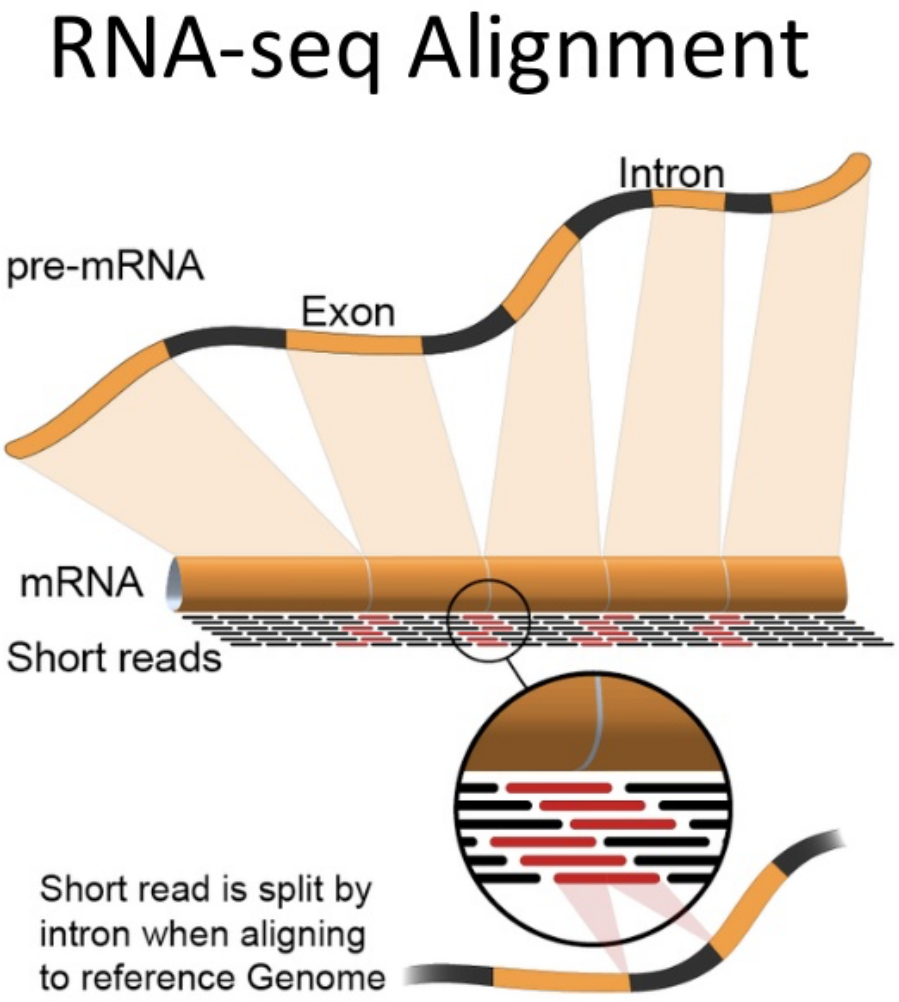
Raw Sequence Cleanup

Trim and/or filter sequence to remove sequencing primers/adaptor and poor quality reads. Example programs:

- **FASTX-Toolkit** is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.
- **SeqKit** is an ultrafast comprehensive toolkit for FASTA/Q processing.
- **Trimmomatic** is a fast, multithreaded command line tool that can be used to trim and crop Illumina (FASTQ) data as well as to remove adapters.
- **TrimGalore** is a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries.
- **Cutadapt** finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads.

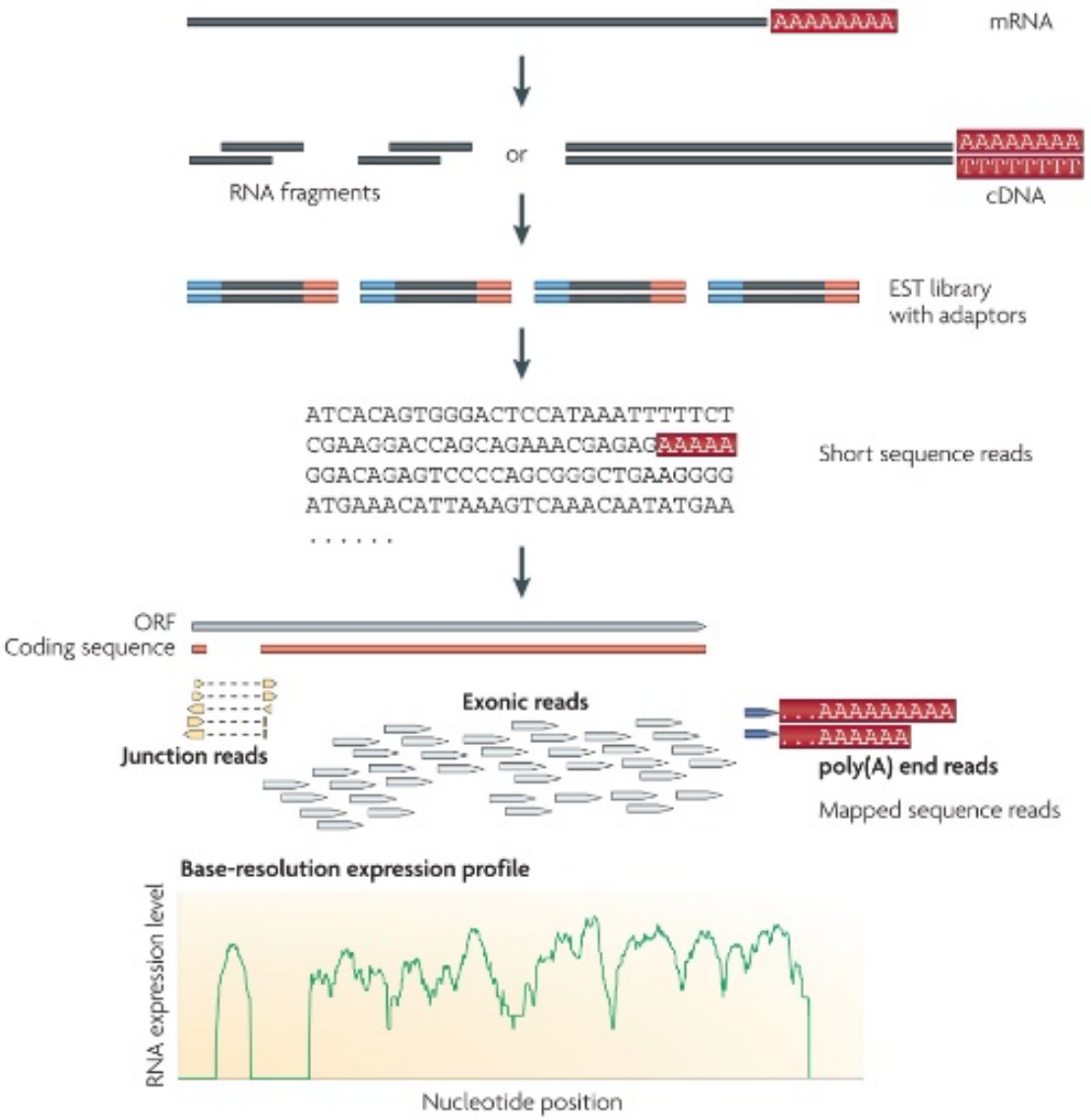
Alignment

RNASeq Mapping Challenges



The majority of mRNA derived from eukaryotes is the result of splicing together discontinuous exons.

RNA-seq protocol schematic

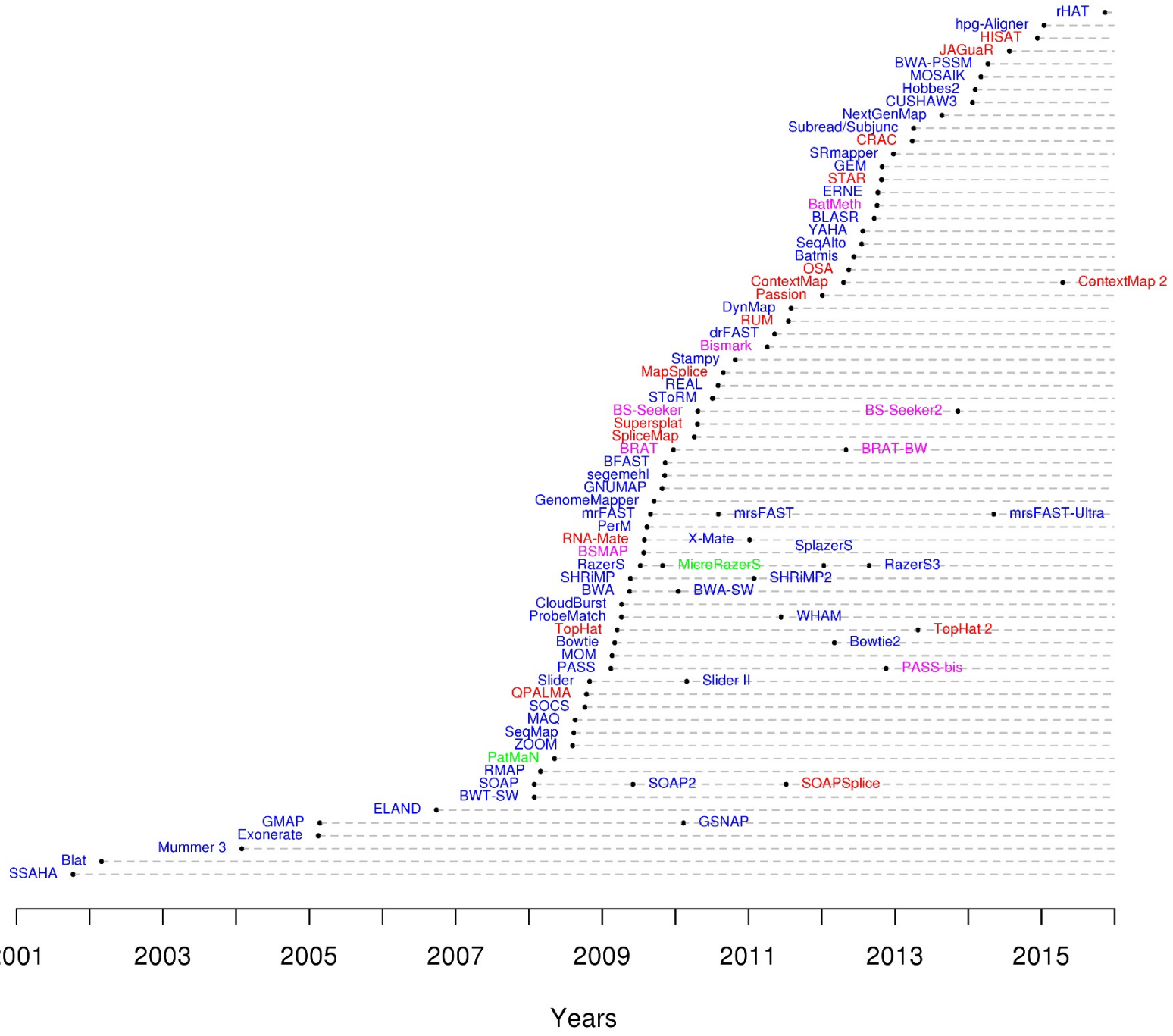


Mapping Challenges

- Reads not perfect
- Duplicate molecules (PCR artifacts skew quantitation)
- Multimapped reads - Some regions of the genome are thus classified as unmappable
- Aligners try **very** hard to align **all** reads, therefore fewest artifacts occur when all possible genomic locations are provides (genome over transcriptome)

RNASeq Mapping Solutions

- **Align against the transcriptome**
  - Many/All transcriptomes are incomplete
  - Can only measure known *genes*
  - Won't detect non-coding RNAs
  - Can't look at splicing variants
  - Can't detect fusion genes or structure variants
- **De novo assembly of RNASeq reads**
  - Largely used for uncharacterized genomes
- **Align against the genome using a splice-aware aligner**
  - Most versatile solution
- **Pseudo-Aligner - quasi mappers (Salmon and Kalisto)**
  - New class of programs - blazingly fast
  - Map to transcriptome (not genome) and does quantitation
  - Surprisingly accurate except for very low abundance signals
  - With bootstrapping can give confidence values


rHAT
hpg-Aligner
HISAT
JAGuaR
BWA-PSSM
MOSAIK
Hobbes2
CUSHAW3
NextGenMap
Subread/Subjunc
CRAC
SRmapper
GEM
STAR
ERNE
BatMeth
BLASR
YAHA
SeqAlto
Batmis
OSA
ContextMap
ContextMap 2
Passion
DynMap
RUM
drFAST
Bismark
Stampy
MapSplice
REAL
SToRM
BS-Seeker
BS-Seeker2
Supersplat
SpliceMap
BRAT
BRAT-BW
BFAST
segemehl
GNUMAP
GenomeMapper
mrFAST
mrsFAST
mrsFAST-Ultra
PerM
RNA-Mate
X-Mate
SplazerS
BSMAP
RazerS
MicroRazerS
RazerS3
SHRiMP
SHRiMP2
BWA
BWA-SW
CloudBurst
ProbeMatch
WHAM
TopHat
TopHat 2
Bowtie
Bowtie2
MOM
PASS
PASS-bis
Slider
Slider II
QPALMA
SOCS
MAQ
SeqMap
ZOOM
PatMaN
RMAP
SOAP
SOAP2
SOAPSplice
BWT-SW
ELAND
GSNAP
GMAP
Exonerate
Mummer 3
Blat
SSAHA
2001
2003
2005
2007
2009
2011
2013
2015
Years

Common Aligners

Most alignment algorithms rely on the construction of auxiliary data structures, called indices, which are made for the sequence reads, the reference genome sequence, or both. Mapping algorithms can largely be grouped into two categories based on properties of their indices: algorithms based on hash tables, and algorithms based on the Burrows-Wheeler transform

- Bowtie2
- BWA/BWA-mem
- **STAR**
- HISAT
- HISAT2
- TopHat
- TopHat2

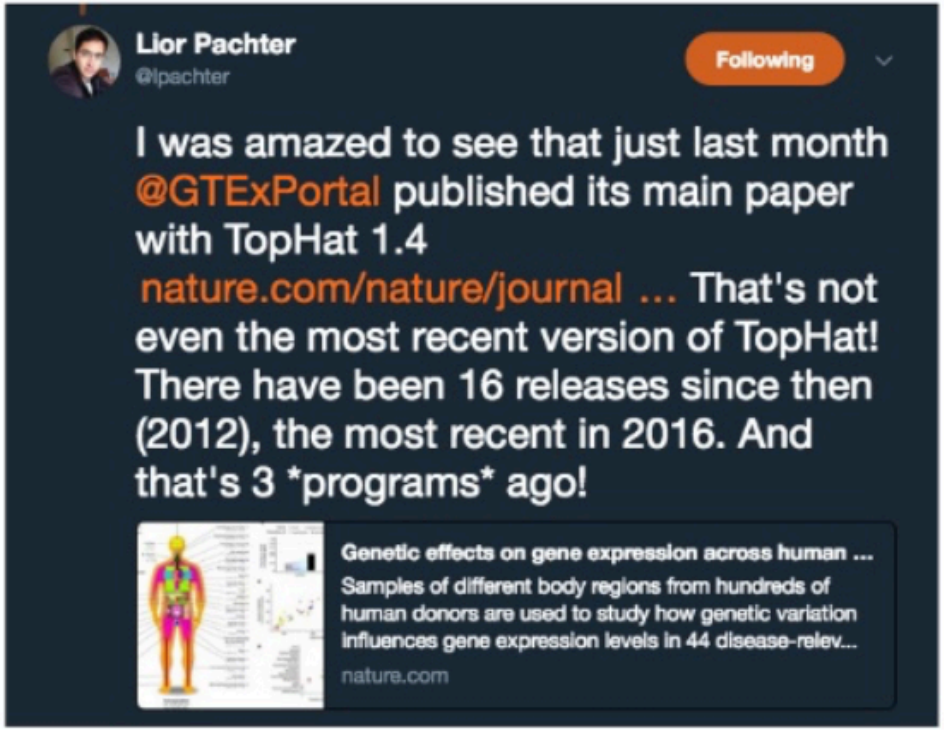
Tools for mapping high-throughput sequencing data

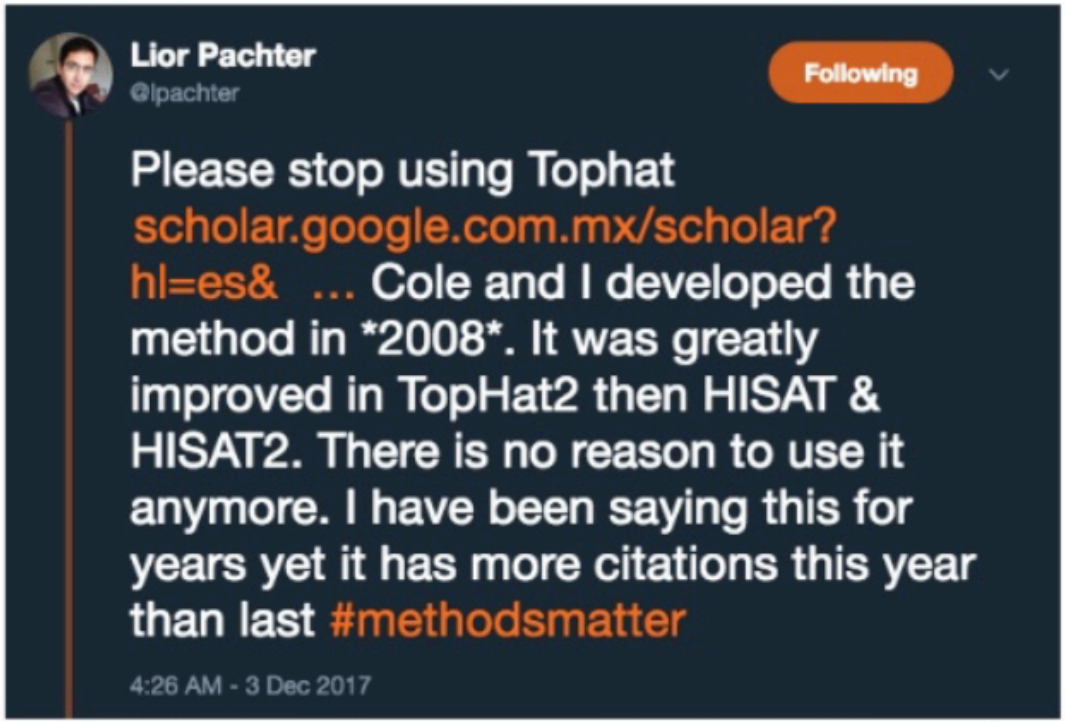
Nuno A. Fonseca Johan Rung Alvis Brazma John C. Marioni Author Notes

Bioinformatics, Volume 28, Issue 24, 1 December 2012, Pages 3169–3177, https://doi.org/10.1093/bioinformatics/bts605

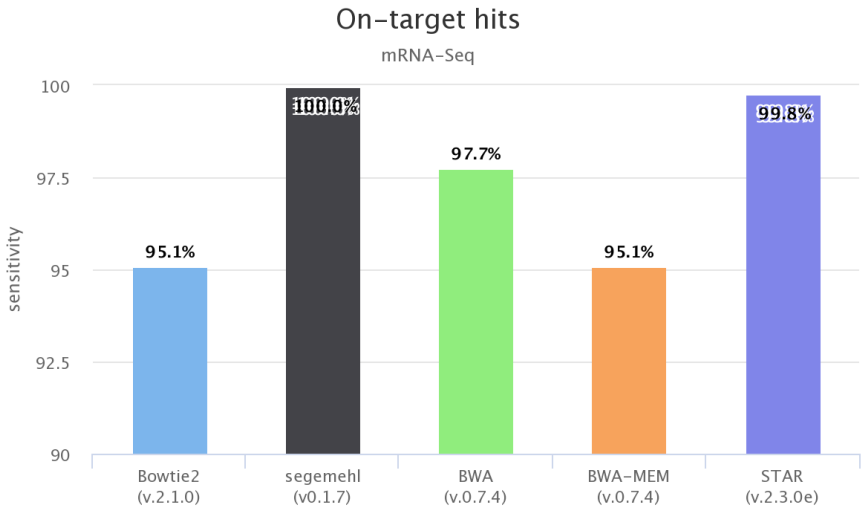
The Times they are a Changin !!

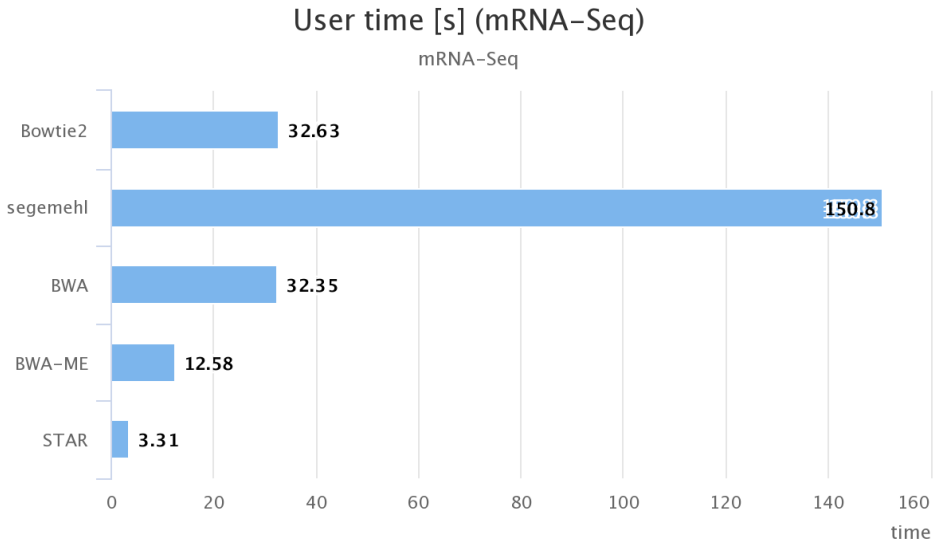
Check or new versions… try new software

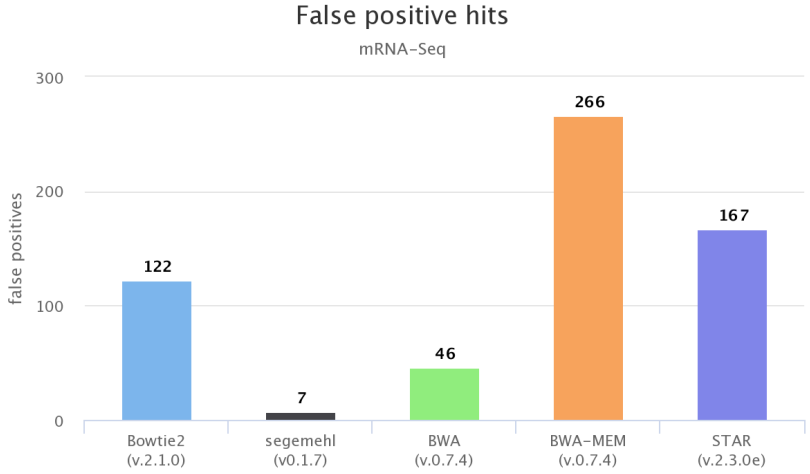


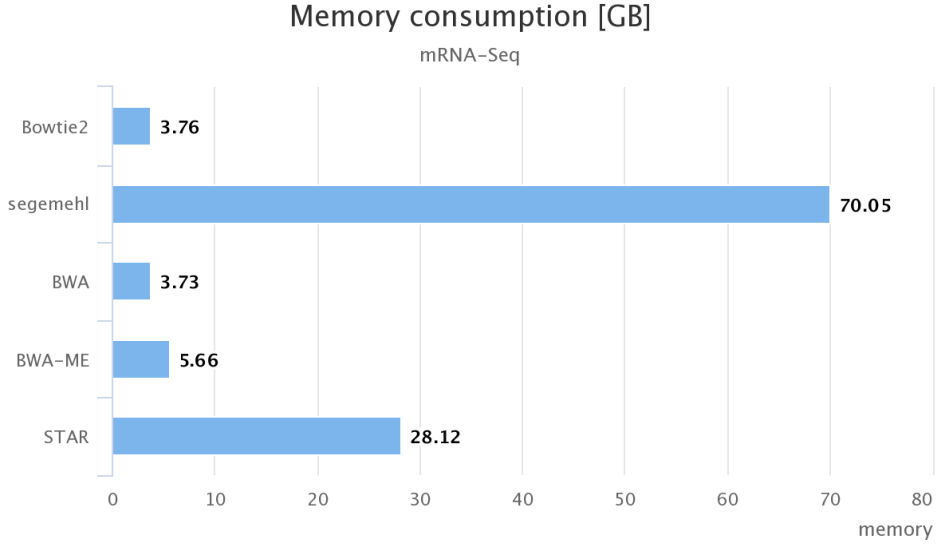


Source: Twitter









ECSEQ BIOINFORMATICS

Pseudo-Aligners

kallisto is a program for quantifying abundances of transcripts from RNA-Seq data, or more generally of target sequences using high-throughput sequencing reads. It is based on the novel idea of pseudoalignment for rapidly determining the compatibility of reads with targets, without the need for alignment. (*https://doi.org/10.1038/nbt.3519*)

Salmon uses new algorithms (specifically, coupling the concept of quasi-mapping with a two-phase inference procedure) to provide accurate expression estimates very quickly (i.e. wicked-fast) and while using little memory. Salmon performs its inference using an expressive and realistic model of RNA-seq data that takes into account experimental attributes and biases commonly observed in real RNA-seq data. (*https://doi.org/10.1038/nmeth.4197*)

To Align or not to Align

Aligners typically align against the entire genome and provide a output where the results can be **visibly inspected** (bam file via IGV). The must be used for detecting novel genes/transcripts. Quantitation of aligned reads to specific genes is typically done by separate program

PseudoAligners assign reads to the most appropriate transcript… can't find novel genes/transcripts or other anomalies. Generally much faster than aligner and are likely more accurate (Recent improvements in salmon have increased its accuracy, at the expense of being somewhat slower than the original)

Typical Questions about alignment

- What is the best aligner to use?
- What Genome version should I use?
- What Genome annotation should I use?

Answers

- STAR - (**Salmon** or Kallisto) - *subjective*
- Depends ! most recent or best annotated
- GeneCode with caveats - know what is being annotated and what is not and how it effects your results

Questions not asked

- What parameters should I use?

Answers

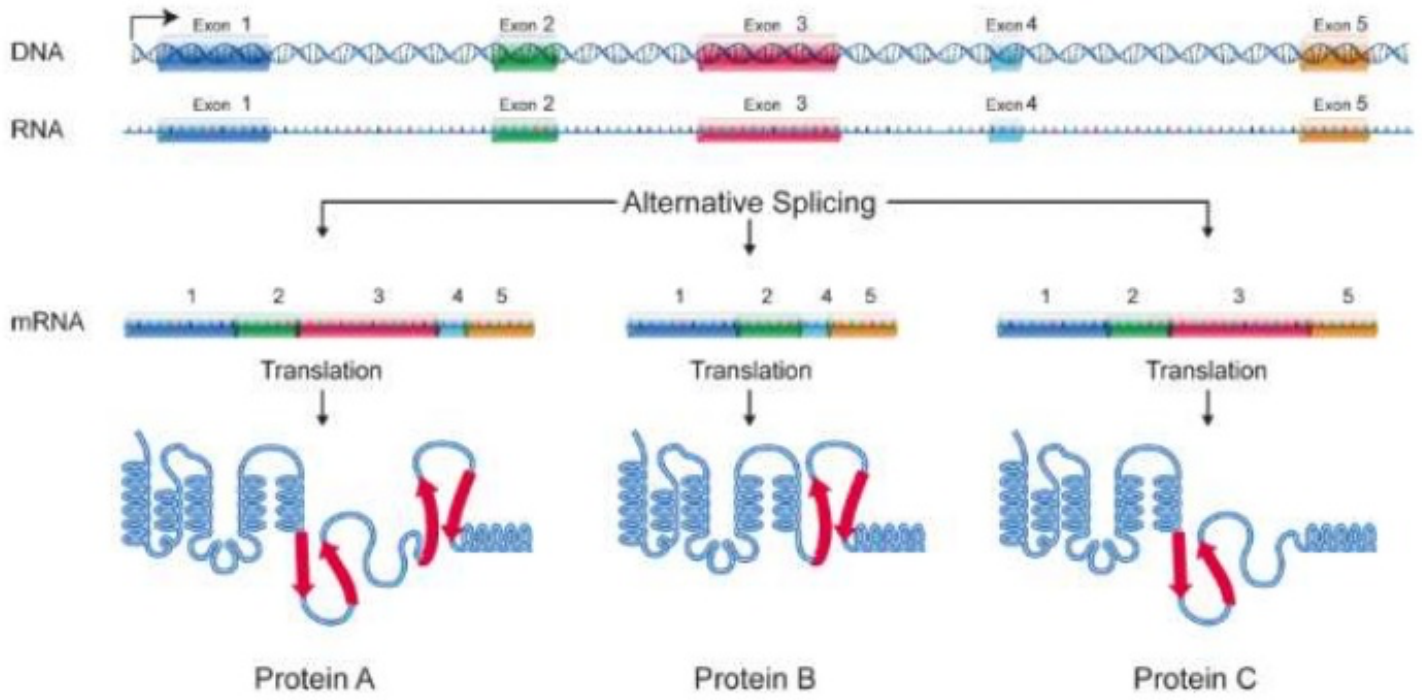
- Most programs have lots of optional parameters that can tweak the results, but most are set to defaults that should work in most common situations.
  (*Don't touch what you don't understand -* ***especially*** *if it gets you, your favorite answer*)

Additional Parameters For Star In CCBR Pipeliner Program

```
STAR
--runThreadN 32
--genomeDir /fdb/STAR_2.4.2a/GENCODE/Gencode_mouse/
release_M4/genes-125
--readFilesIn R1_all.fastq.gz   R2_all.fastq.gz
--readFilesCommand zcat
--limitSjdbInsertNsj 2000000
--outFileNamePrefix Ker_RNA.Rep01.p2.
--outSAMtype BAM   SortedByCoordinate
--outSAMstrandField None
--outSAMunmapped Within
--outWigType None
--outWigStrand Stranded
--outFilterType BySJout
--outFilterMultimapNmax 10
--outFilterMismatchNmax 10
--outFilterMismatchNoverLmax 0.3
--outFilterIntronMotifs RemoveNoncanonicalUnannotated
--clip3pAdapterSeq -
--alignIntronMin 21
--alignIntronMax 0
--alignMatesGapMax 0
--alignSJoverhangMin 5
--alignSJDBoverhangMin 3
--sjdbFileChrStartEnd Ker_RNA.Rep01.SJ.out.tab
--sjdbGTFfile /fdb/GENCODE/Gencode_mouse/release_M4/
gencode.vM4.annotation.gtf
--quantMode Transcriptome
```

RNA-Seq: Special Mapping Concerns

Alternate Splicing



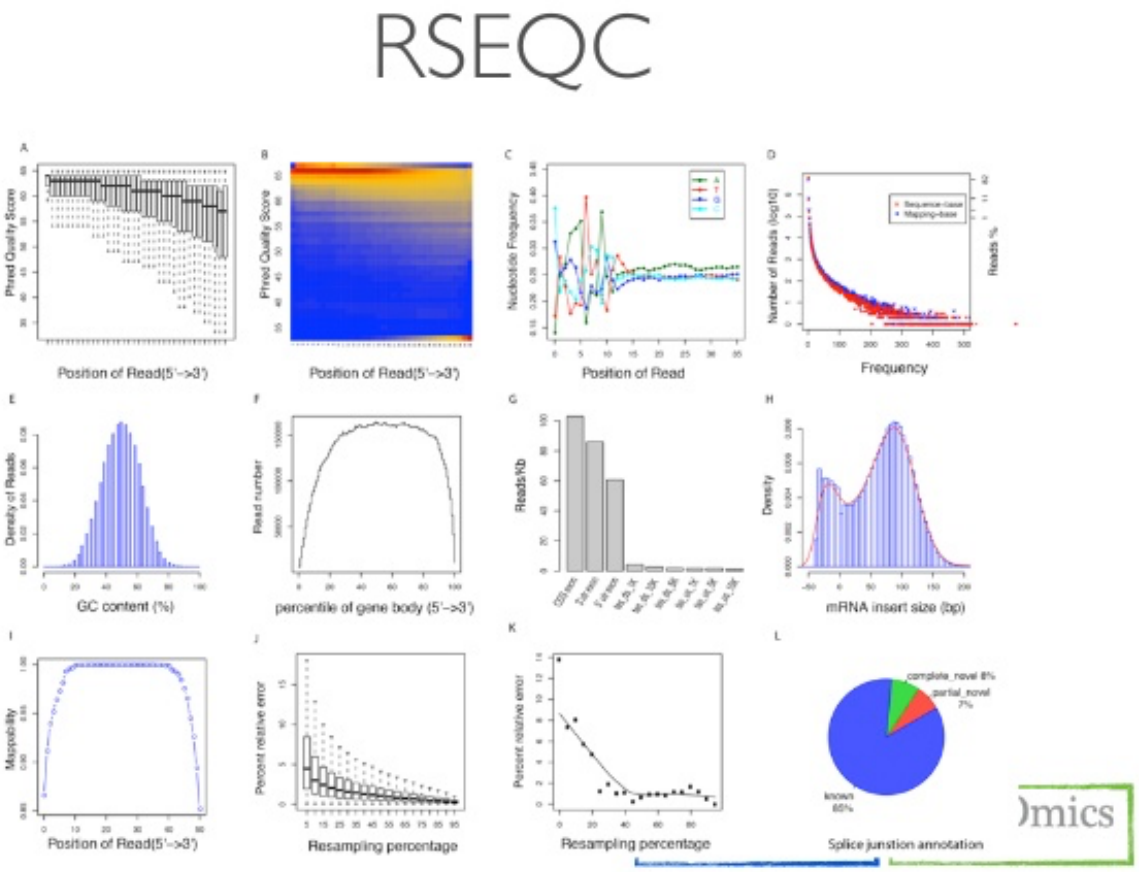
Post Alignment QC

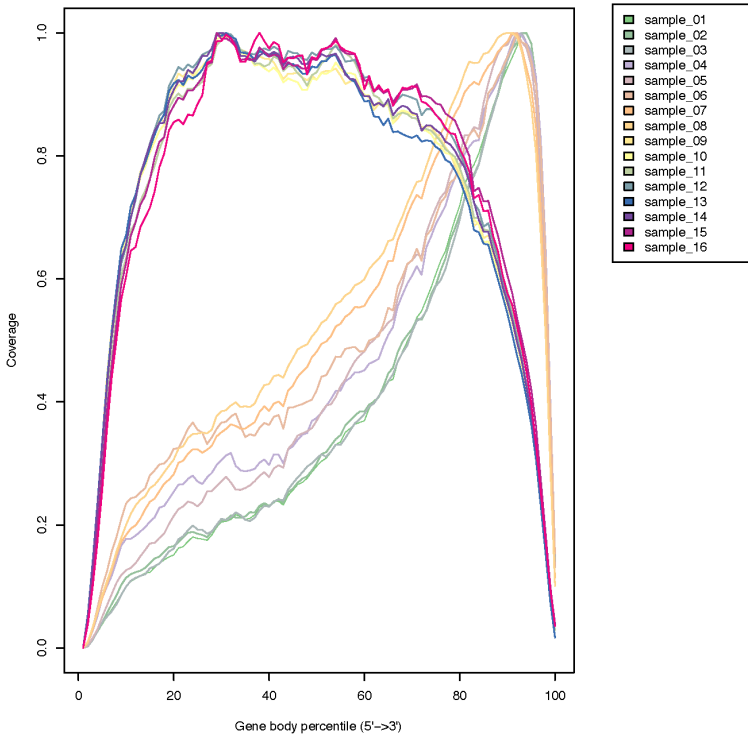
RSeQC package provides a number of useful modules that can comprehensively evaluate high throughput sequence data especially RNA-seq data. “Basic modules” quickly inspect sequence quality, nucleotide composition bias, PCR bias and GC bias, while “RNA-seq specific modules” investigate sequencing saturation status of both splicing junction detection and expression estimation, mapped reads clipping profile, mapped reads distribution, coverage uniformity over gene body, reproducibility, strand specificity and splice junction annotation.

MultiQC is a modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Picard Tools - RNAseqMetrics is a module that produces metrics about the alignment of RNA-seq reads within a SAM file to genes

RSeQC example of plot types





Post Alignment Cleanup

Picard is a set of command line tools for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. (mark pcr duplicates)

Samtools provide various utilities for manipulating alignments in the SAM/BAM format, including sorting, merging, indexing and generating alignments in a per-position format.

BamTools is a command-line toolkit for reading, writing, and manipulating BAM (genome alignment) files.