A Basic Overview of RNASeq Data Analysis

Peter FitzGerald, PhD

Head Genome Analysis Unit
Director of BTEP
CCR, NCI

RNA Workshop

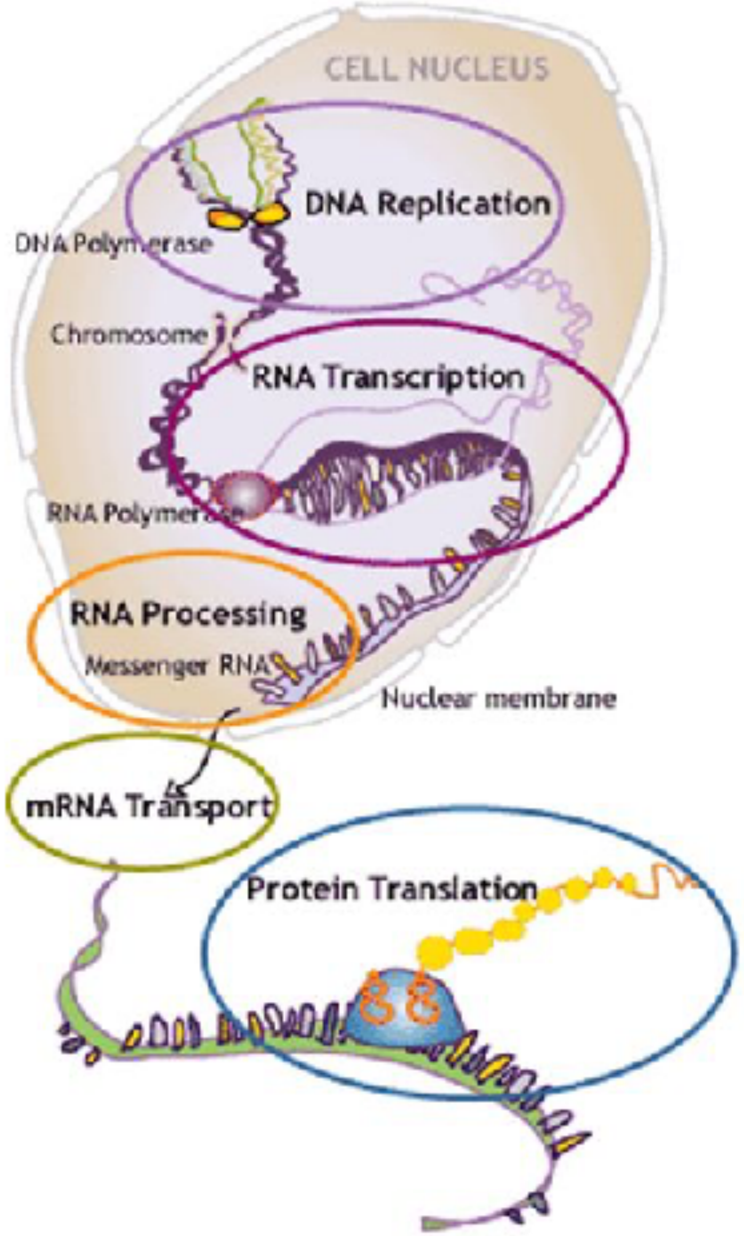
- Introduction To RNA-Seq Technology, Overview And Analyses

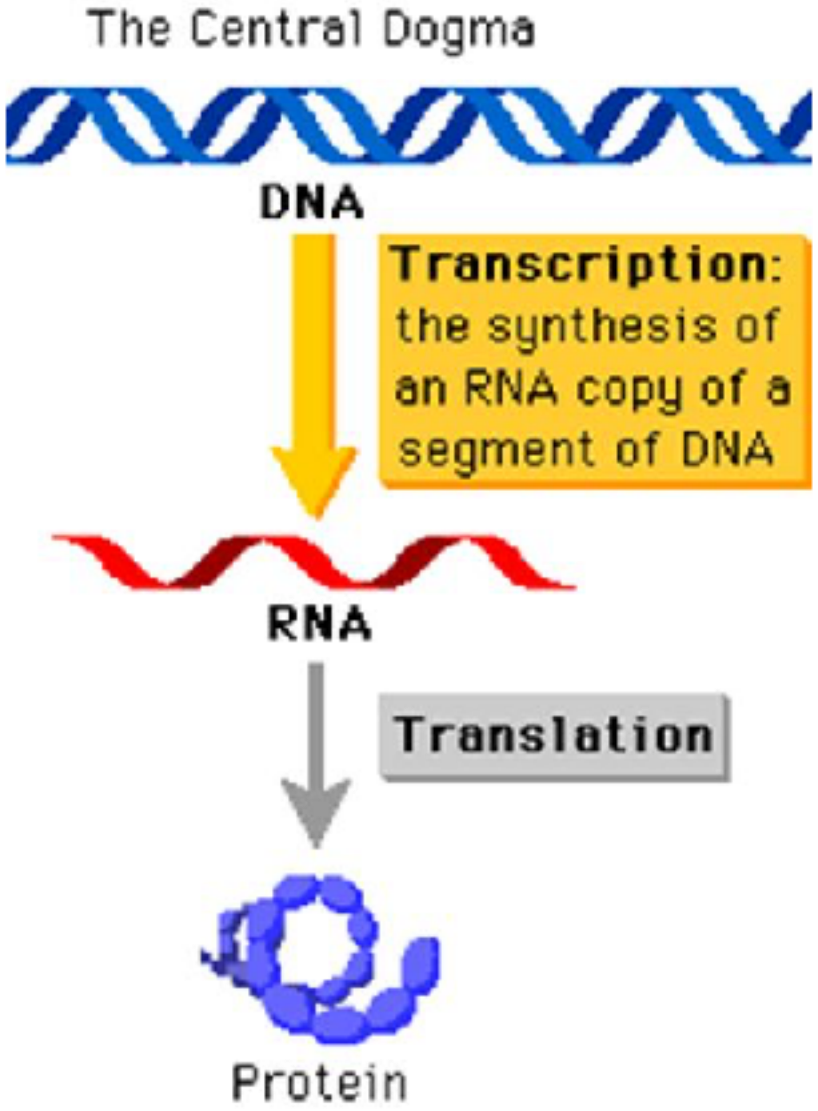
  (Now)
- Introduction To IPA (Ingenuity Pathway Analysis) And The Core Analysis

  (This Afternoon)
- RNA-Seq Data Analysis In Partek Flow

  (Tomorrow Morning)

DNA → RNA → Protein




The Central Dogma
DNA
Transcription: the synthesis of an RNA copy of a segment of DNA
RNA
Translation
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 - WorkFlow

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

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**

RNASEQ - Differential Expression WorkFlow

- **Sample Consistency**
  - Check for sample outliers
- **Differential Expression Program**
  - Filtering
  - Normalization
  - Fit to Statistical Model
  - Generate Comparison Ratios
  - Adjust for Multiple Testing
  - Check results for confidence
  - Annotation

RNASEQ - WorkFlow

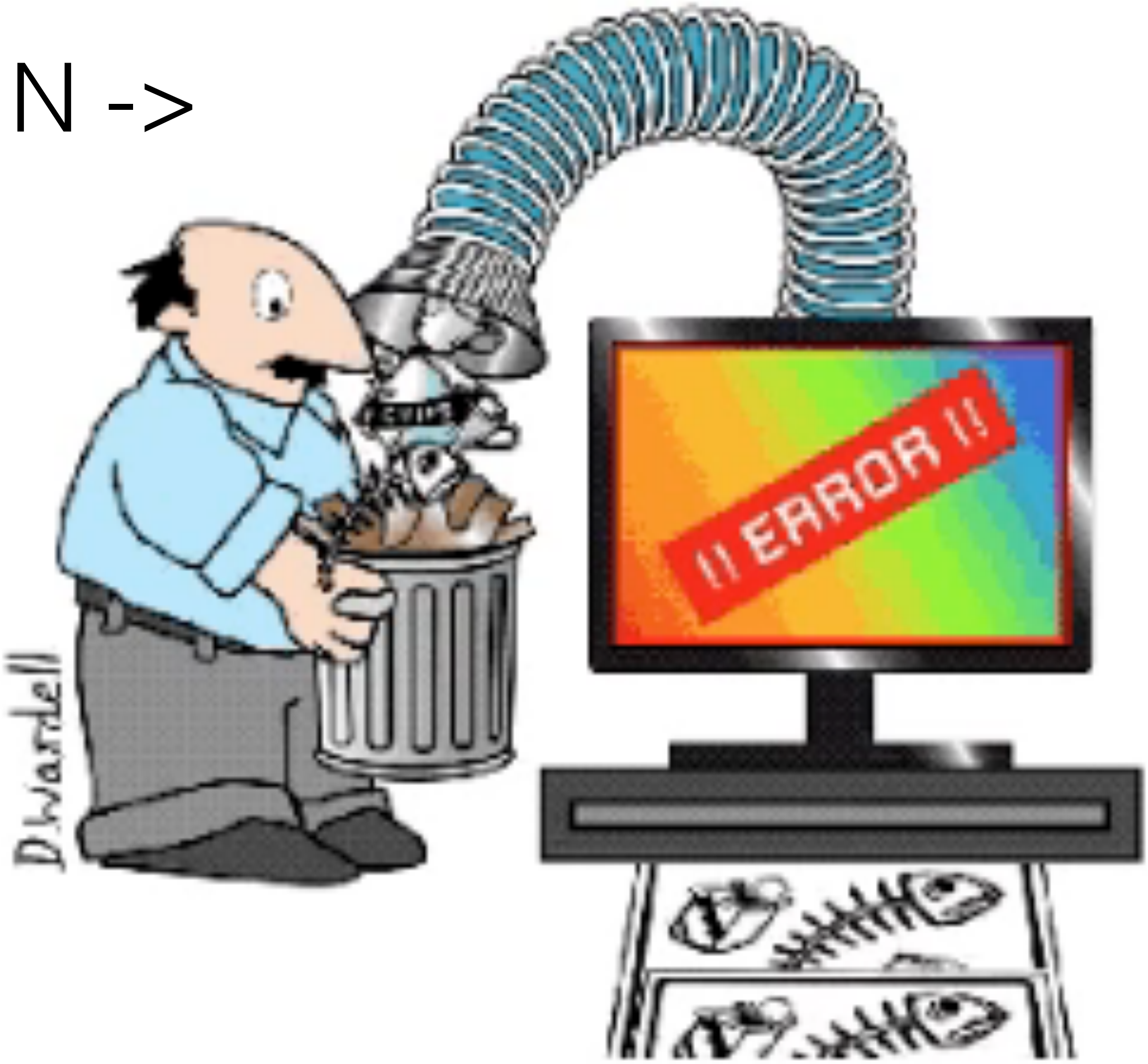
- **Differential Expressed Gene List**
  - Log fold change
  - pvalue
  - FDR
- **Visualization**
  - IGV - genomic context and raw data
  - Clustering
  - Scatter plots
- **Biological Interpretation**
  - GSEA
  - Pathway Analysis

Generating the Data

Experimental Design
Sample Preparation
Sequencing

Experimental Design

GARBAGE IN ->



GARBAGE OUT ->

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

Next Generation Sequencing Platforms

Illumina
Sequencing by Synthesis (SbS)
/NovaSeq/HiSeq/NextSeq/MiSeq
Short read length (30 to 300 bp)
High throughput
“Industry Standard”

PacBio
Sequencing by Synthesis
single-molecule, real-time (SMRT) technology
Long Reads ~10,000 bp
No PCR bias and artifacts

Minion (Oxford Nanopore)
Long Reads ~100,000 bp
No PCR bias and artifacts
RNA and DNA

Need size range and technology

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.

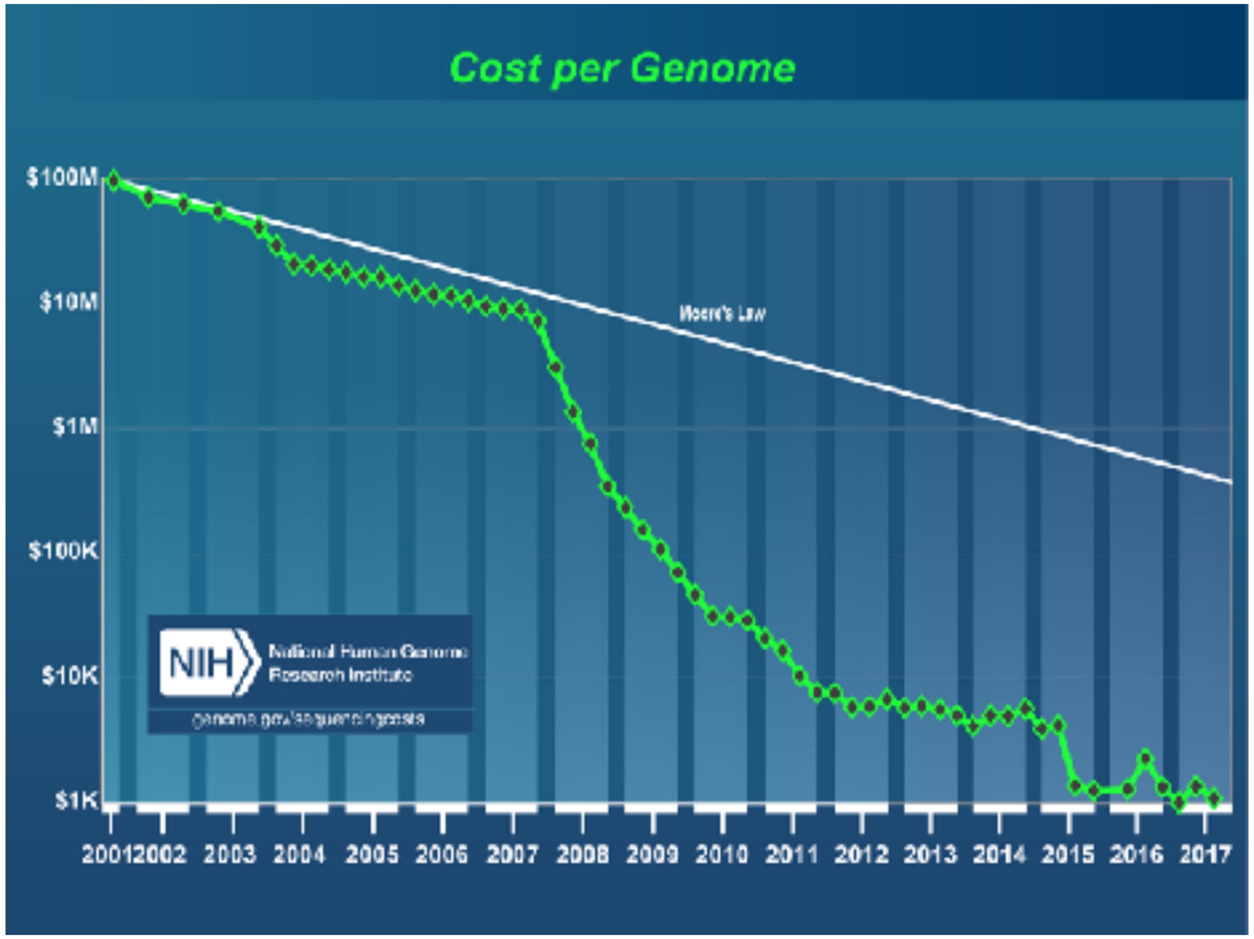
5. If you are unable to process all your RNA samples together and need to process them in batches, make sure that replicates for each condition are in each batch so that the batch effects can be measured and removed bioinformatically.

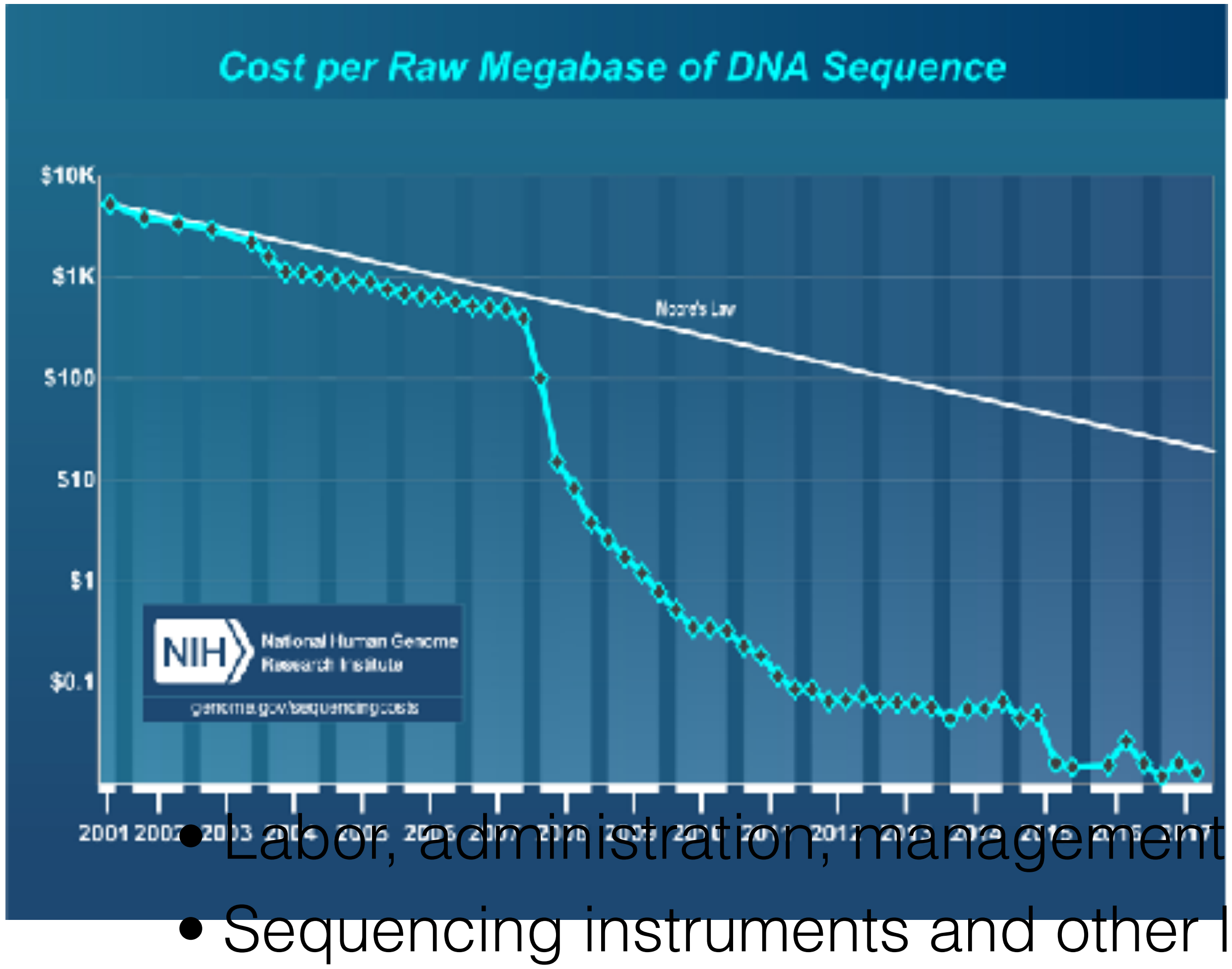
6. For sequence depth and machine requirements, visit Illumina Sequencing Coverage website

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*

http://genome.gov/sequencingcosts/





- Labor, administration, management, utilities, reagents, and consumables
- Sequencing instruments and other large equipment (amortized over three years)
- Informatics activities directly related to sequence production (e.g., laboratory information management systems and initial data processing)
- Submission of data to a public database
- Indirect Costs as they relate to the above items

Costs (mRNA)

CCR Sequencing Facility (subsidized pricing**)**

Library Construction	$61	
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)

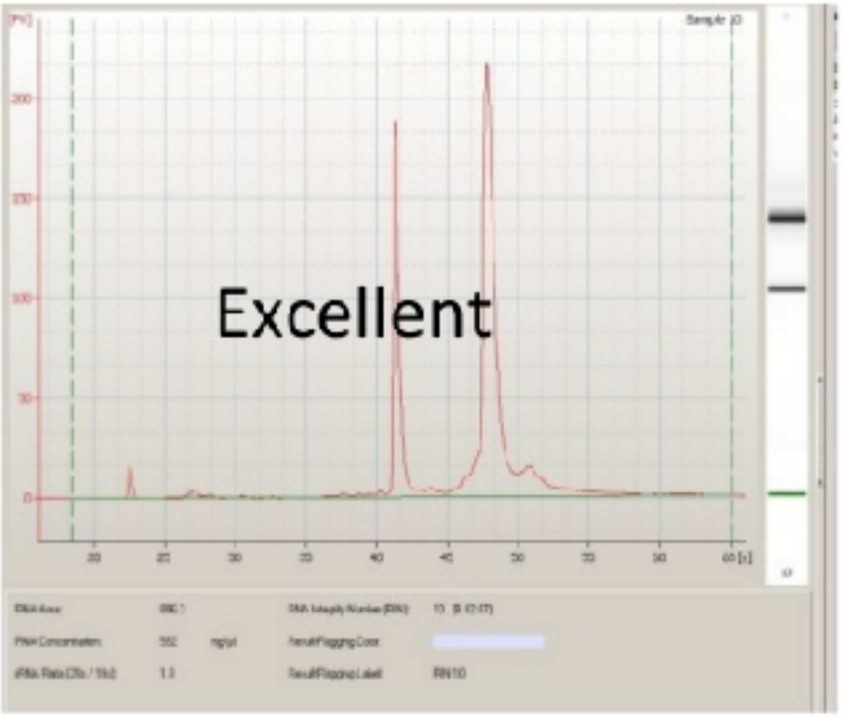
https://ostr.ccr.cancer.gov/resources/sequencing-facility/

Sample Preparation

General Rules for Sample Preparation

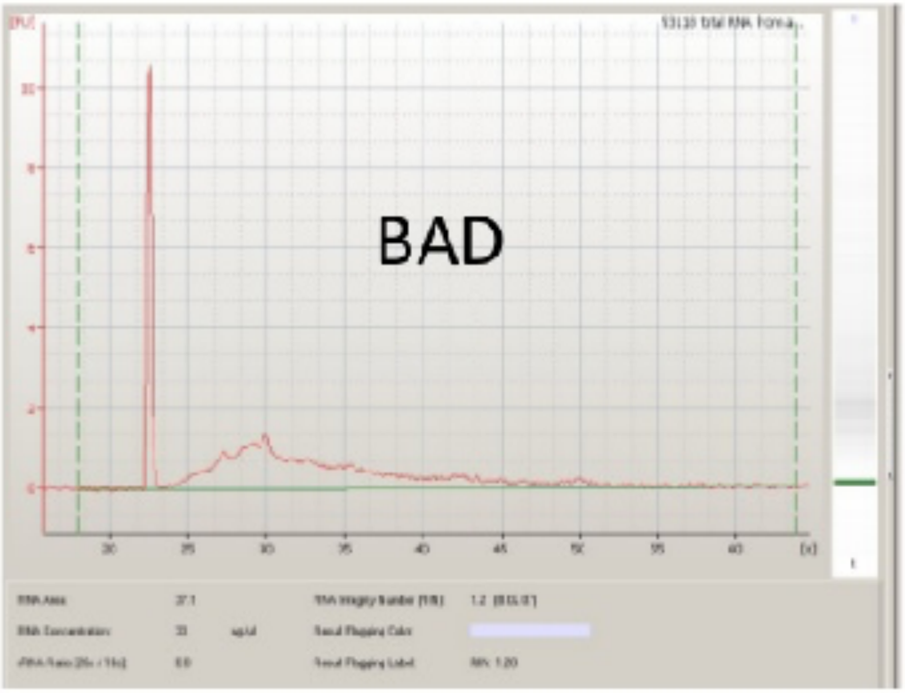
- Prepare all samples at the same time or as close as possible. The same person should prepare all samples
- Do not prepare “experiment” and “control” samples on different days or by different people. (Batch effects).
- Use high quality means to determine sample quality (**R**NA **I**ntegrity **N**umber) (RIN >0.7) 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



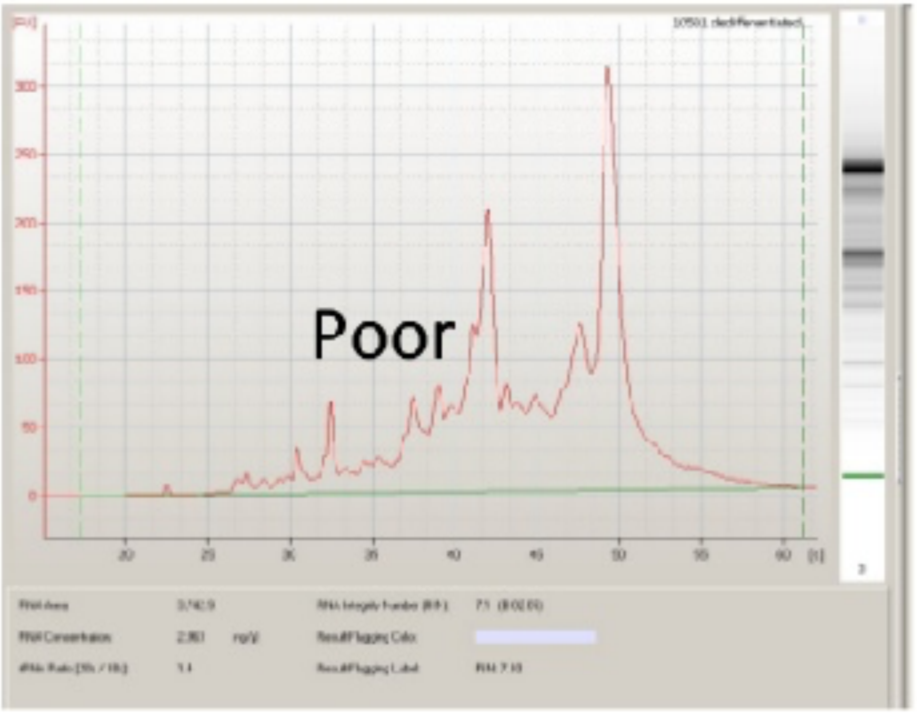
Excellent

RNA Area 880.1
RNA Concentration 552 ng/µl
rRNA Ratio [28s / 18s] 1.3
RNA Integrity Number (RIN) 10 (B.02.07)
Result Flagging Color
Result Flagging Label RIN 10



BAD

RNA Area 27.1
RNA Concentration 33 ng/µl
rRNA Ratio [28s / 18s] 0.0
RNA Integrity Number (RIN) 1.2 (B.02.07)
Result Flagging Color
Result Flagging Label RIN: 1.20



Poor

RNA Area 3,742.9
RNA Concentration 2,361 ng/µl
rRNA Ratio [28s / 18s] 1.1
RNA Integrity Number (RIN) 7.1 (B.02.05)
Result Flagging Color
Result Flagging Label RIN: 7.10

CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

Sequencing

Illumina Sequencing Platforms

Illumina
Sequencing by Synthesis (SbS)
/NovaSeq/HiSeq/NextSeq/MiSeq
Short read length (30 to 300 bp)

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

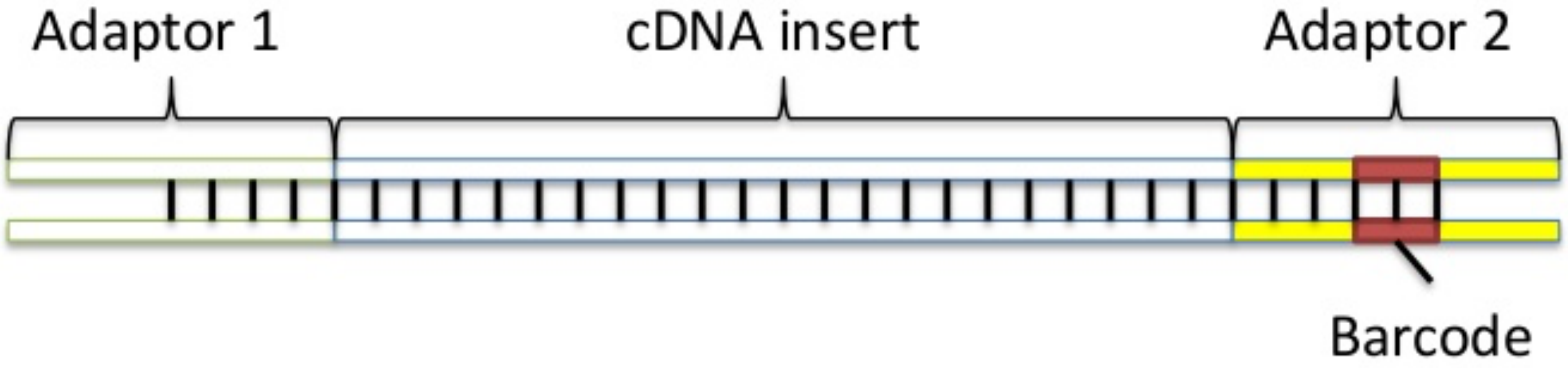
Consult with the Sequencing Core

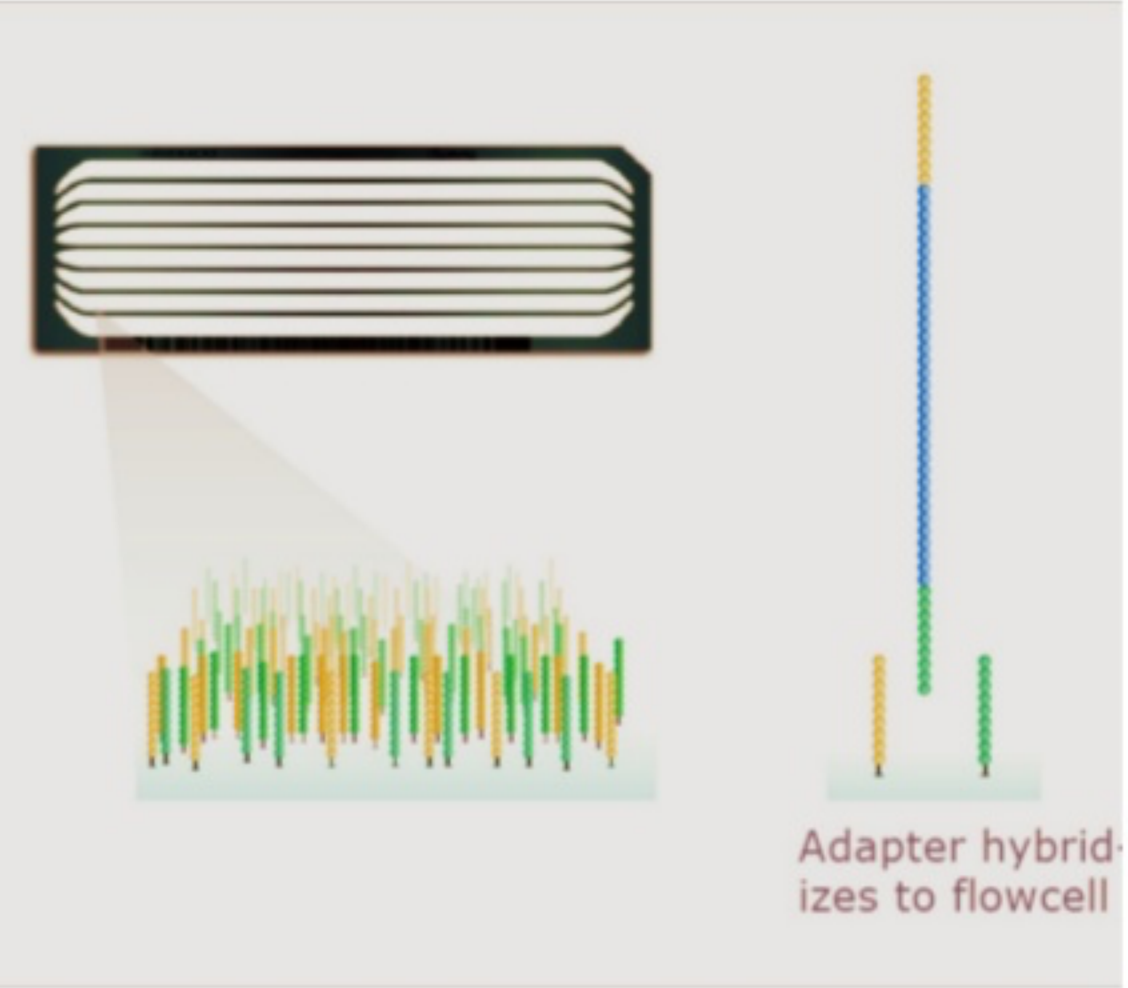
Illumina
NovaSeq

Illumina
NextSeq

Illumina
MiSeq

Sequencing Library Structure



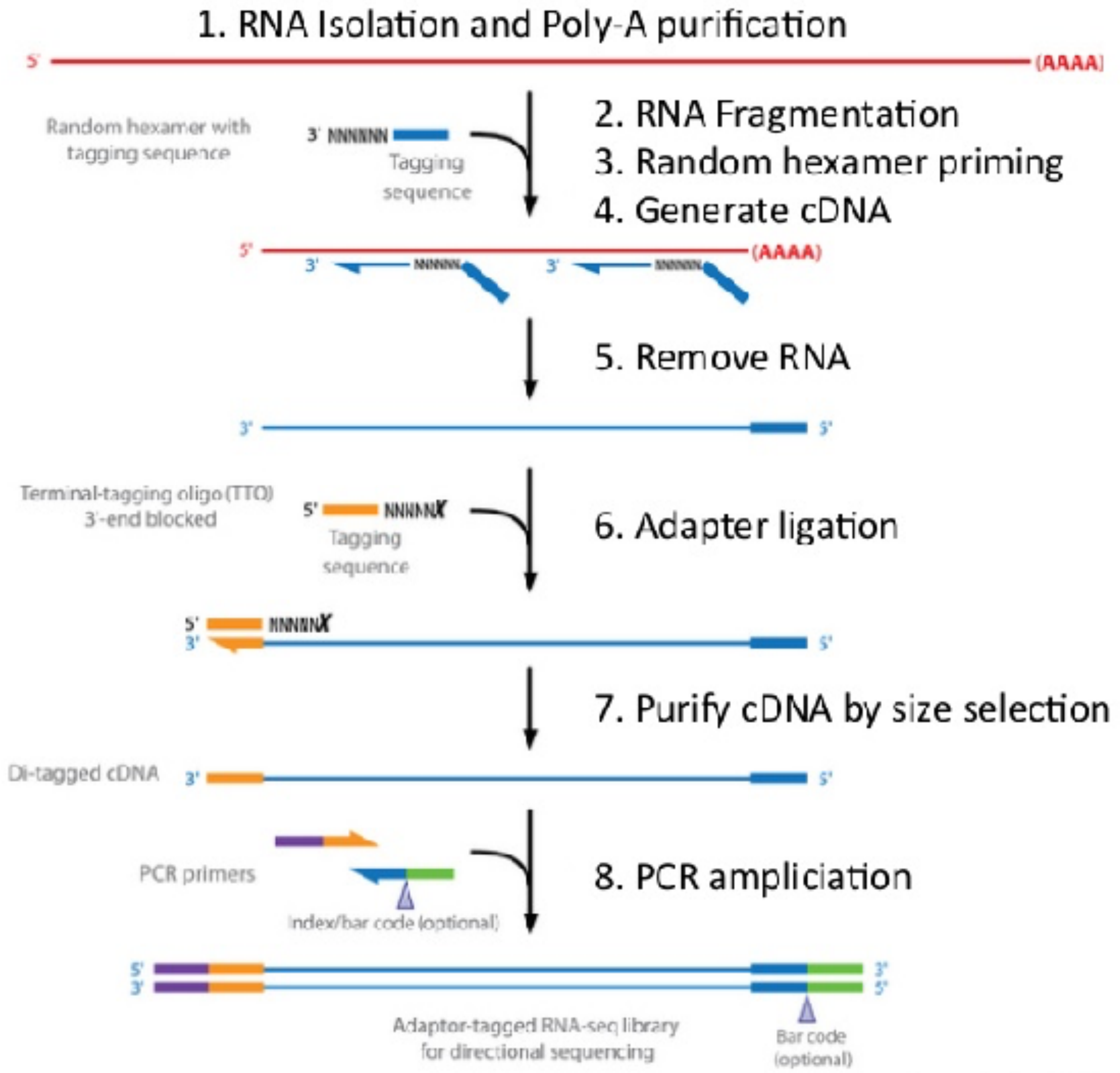


Adaptor – 58 bp nucleotide sequence to fix sequence library onto flow cell

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



1. RNA Isolation and Poly-A purification

2. RNA Fragmentation

3. Random hexamer priming

4. Generate cDNA

5. Remove RNA

6. Adapter ligation

7. Purify cDNA by size selection

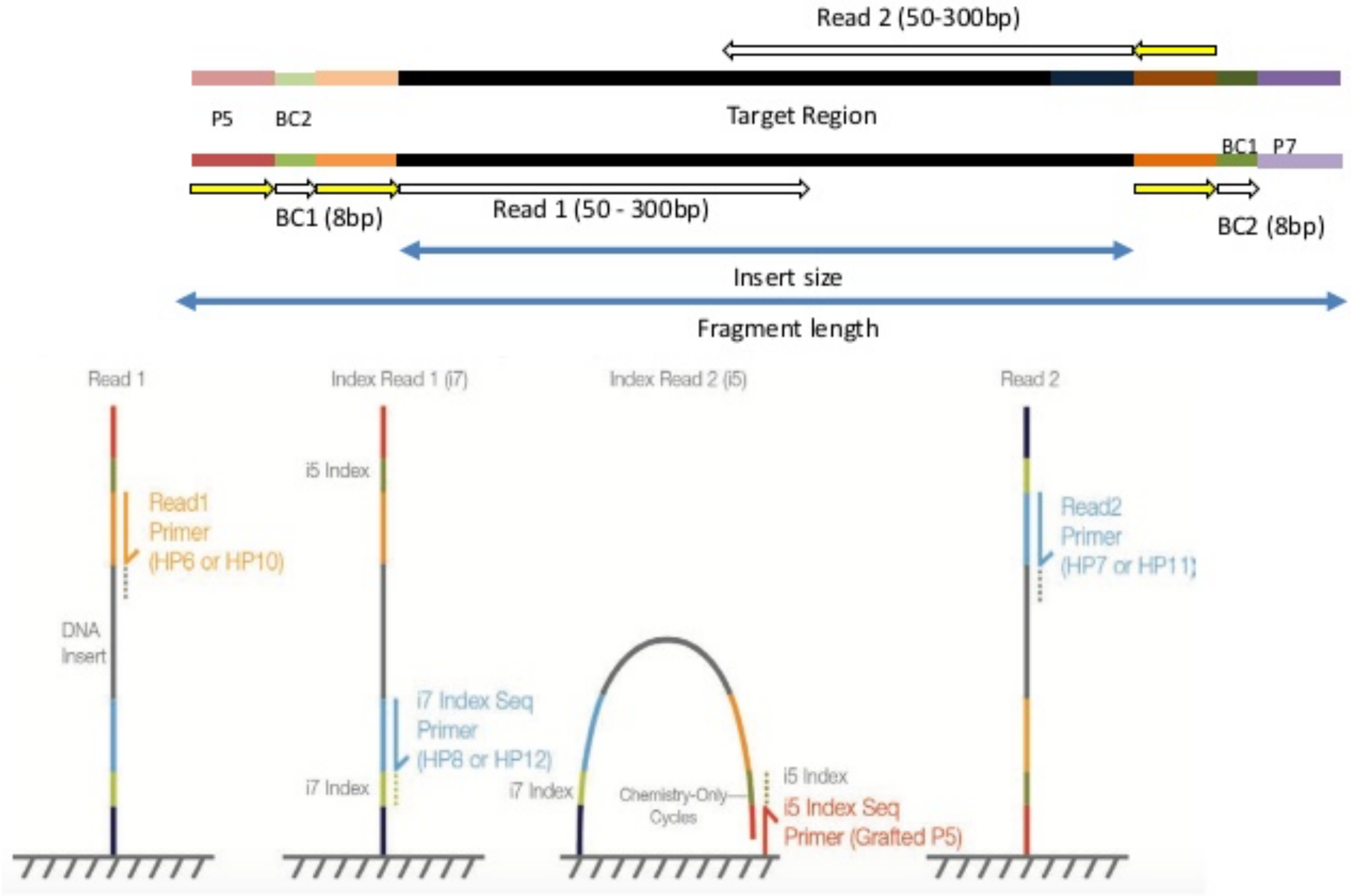
8. PCR ampliciation

5'
(AAAA)
Random hexamer with tagging sequence
3' NNNNNN
Tagging sequence
5'
(AAAA)
3'
NNNNN
3'
NNNNN
3'
5'
Terminal-tagging oligo (TTO) 3'-end blocked
5'
NNNNNX
Tagging sequence
5'
3'
NNNNNX
5'
Di-tagged cDNA
3'
5'
PCR primers
Index/bar code (optional)
5'
3'
3'
5'
Adaptor-tagged RNA-seq library for directional sequencing
Bar code (optional)

Pease, Nature Methods 9 (2012)

Illumina sequencing
sequencing by synthesis

Illumina SBS



SEQUENCE FILE FORMATS

FASTA FORMAT

FASTA

Single sequence example:

```
>HWI-ST398_0092:1:1:5372:2486#0/1
TTTTTCGTTCTTTTCATGTACCGCTTTTTGTTCGGTTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAT
ACGTAGCAGCAGCATCAGTACGACTACGACGACTAGCACATGCGACGATCGATGCTAGCTGACTATCGATG
```

Multiple sequence example:

```
>Sequence Name 1
TTTTTCGTTCTTTTCATGTACCGCTTTTTGTTCGGTTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAT
ACGTAGCAGCAGCATCAGTACGACTACGACGACTAGCACATGCGACGATCGATGCTAGCTGACTATCGATG
>Sequence Name 2
ACGTAGACACGACTAGCATCAGCTACGCATCGATCAGCATCGACTAGCATCACACATCGATCAGCATCACGACTAGCAT
AGCATCGACTACACTACGACTACGATCCACGTACGACTAGCATGCTAGCGCTAGCTAGCTAGCTAGTCGATCGATGAGT
AGCTAGCTAGCTAGC
>Sequence Name 3
ACTCAGCATGCATCAGCATCGACTACGACTACGACATCGACTAGCATCAGCAT
```

SEQUENCE FILE FORMATS

FASTQ FORMAT

FASTQ

Text based format for storing sequence data and corresponding quality scores for each base.
To enable a one-one correspondence between the base sequence and the quality score the score is stored as a single one letter/number code using an offset of the standard ASCII code.
Quality scores range from 0 to 40 and represent a log^{10} score for the probability of being wrong.
E.g. score of 30 => 1:1000 chance of error

SEQUENCE FILE FORMATS

FASTQ FORMAT

FASTQ

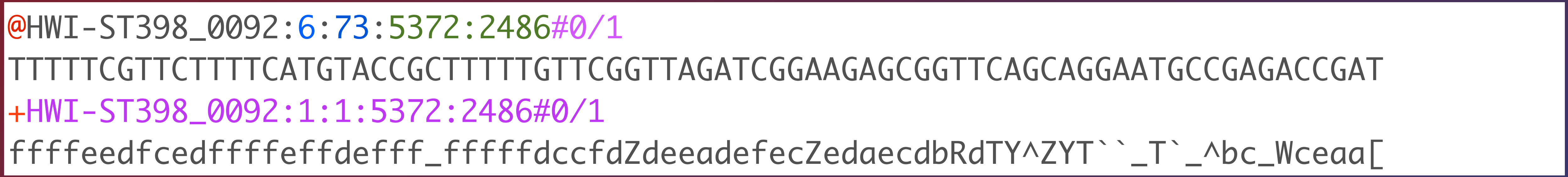
Each fastq file contain multiple entries and each entry consists of 4 lines:

1. header line beginning with “@“ and sequence name
2. sequence line
3. header line beginning with “+” which can have the name but rarely does
4. quality score line

SEQUENCE FILE FORMATS

FASTQ

FASTQ FORMAT



```
@HWI-ST398_0092:6:73:5372:2486#0/1
TTTTTCGTTCTTTTCATGTACCGCTTTTTGTTCGGTTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAT
+HWI-ST398_0092:1:1:5372:2486#0/1
ffffeedfcedffffeffdefff_fffffdccfdZdeeadefecZedaecdbRdTY^ZYT``_T`_^bc_Wceaa[
```

6 - Flowcell lane

73 - Tile number

5372:2486 - 'x','y'-coordinates of the cluster within the tile
#0 - index number for a multiplexed sample (0 for no indexing)

/1 - the member of a pair, /1 or /2 (paired-end or mate-pair reads only)

For paired end reads fastq files come in pairs, typically labelled R1 and R2 (reads are in same order in both files…header often does not distinguish between read1 and read2

SEQUENCE FILE FORMATS

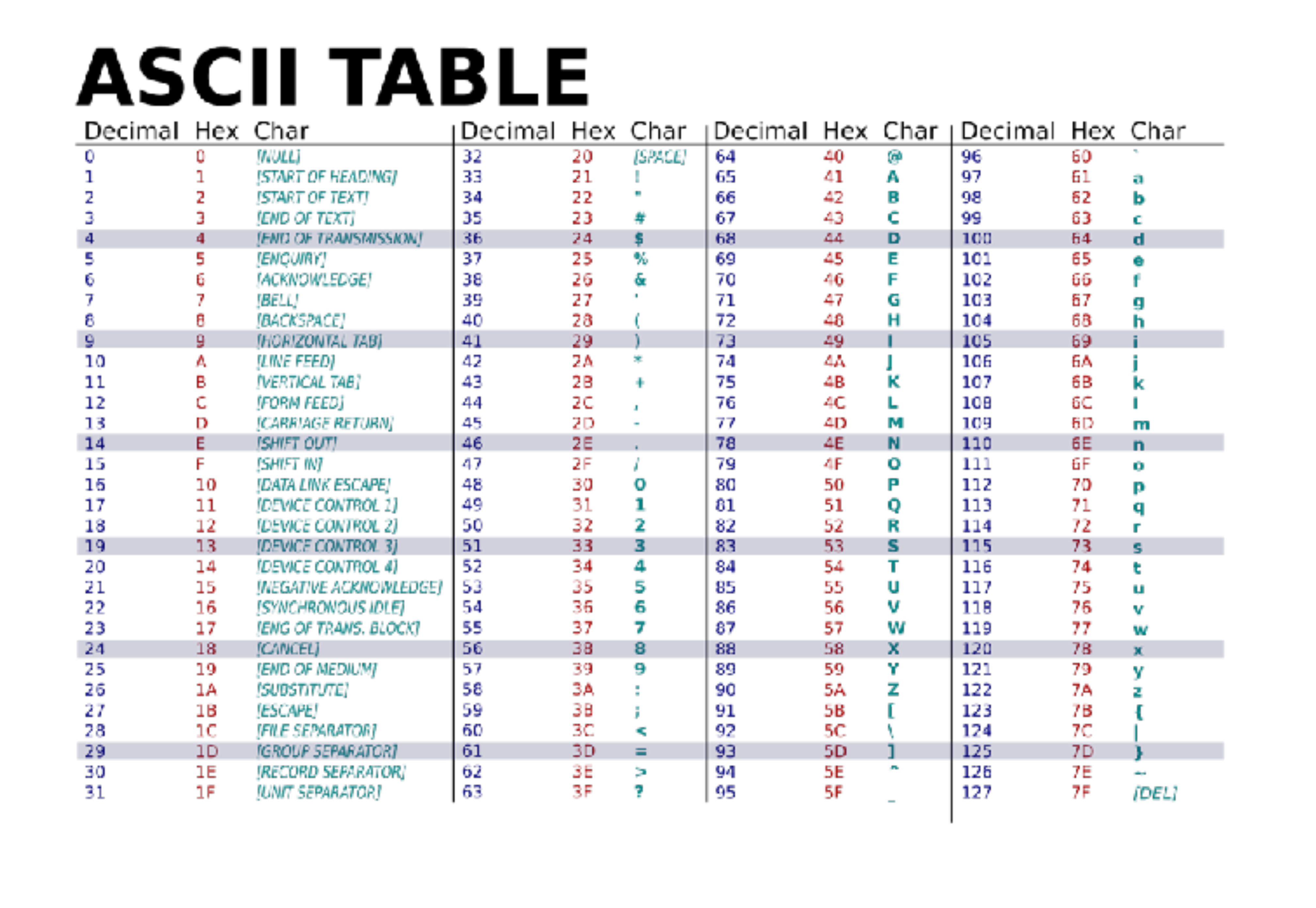
QUALITY SCORES

$$\text{Quality (Q)} = -10\mathit{log}_{10}\text{P}$$

Quality Score	Probabiliy that the base has been called incorrectly
10	1 in 10
20	1 in 100
30	1 in 1,000
40	1 in 10,000

SEQUENCE FILE FORMATS

QUALITY SCORES

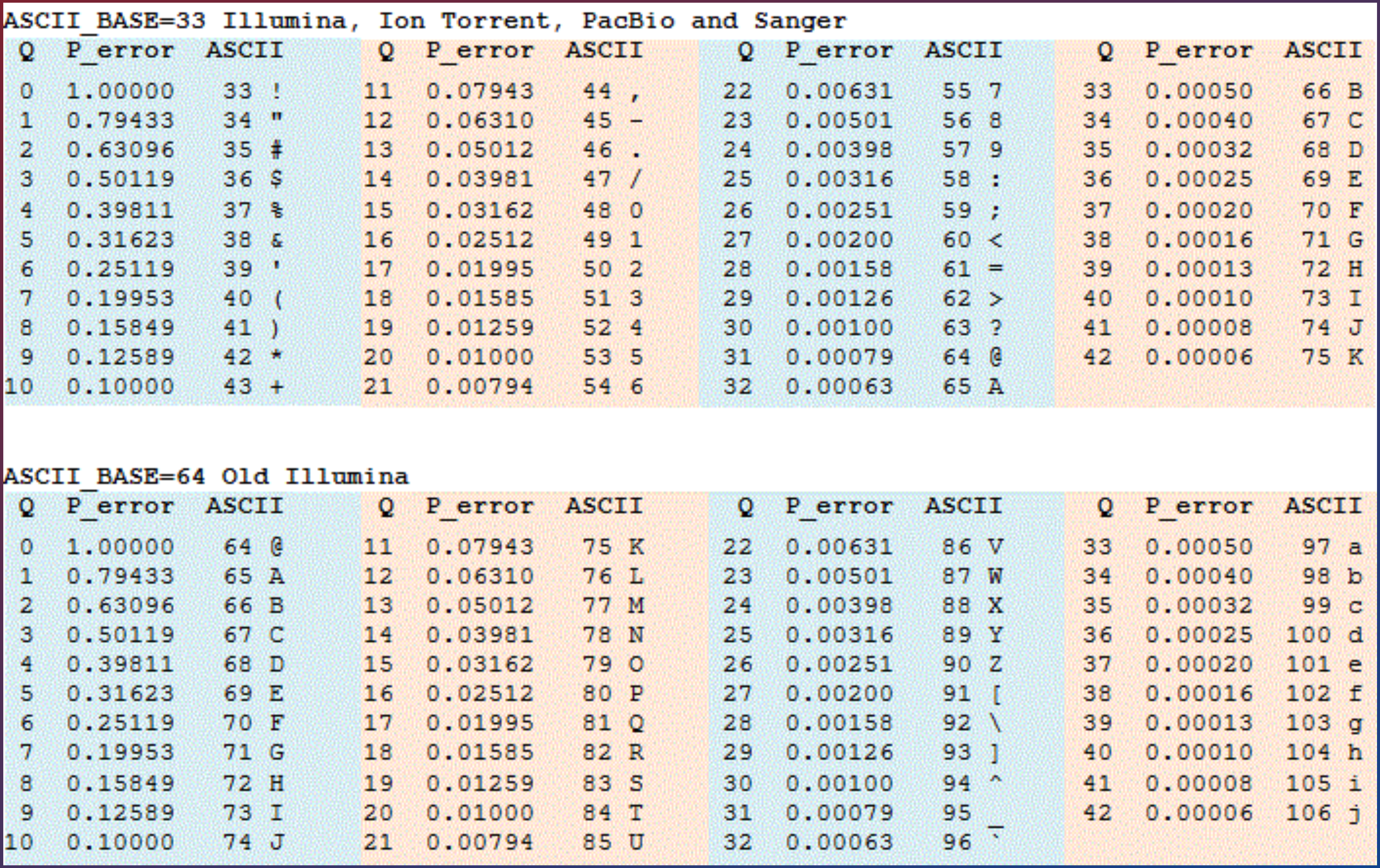


ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

SEQUENCE FILE FORMATS

QUALITY SCORES



ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

Data Analysis

Computational Prerequisites

- High performance Linux computer (multi core, high memory, and plenty of storage)
- Familiarity with the “command line” and at least one programming 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

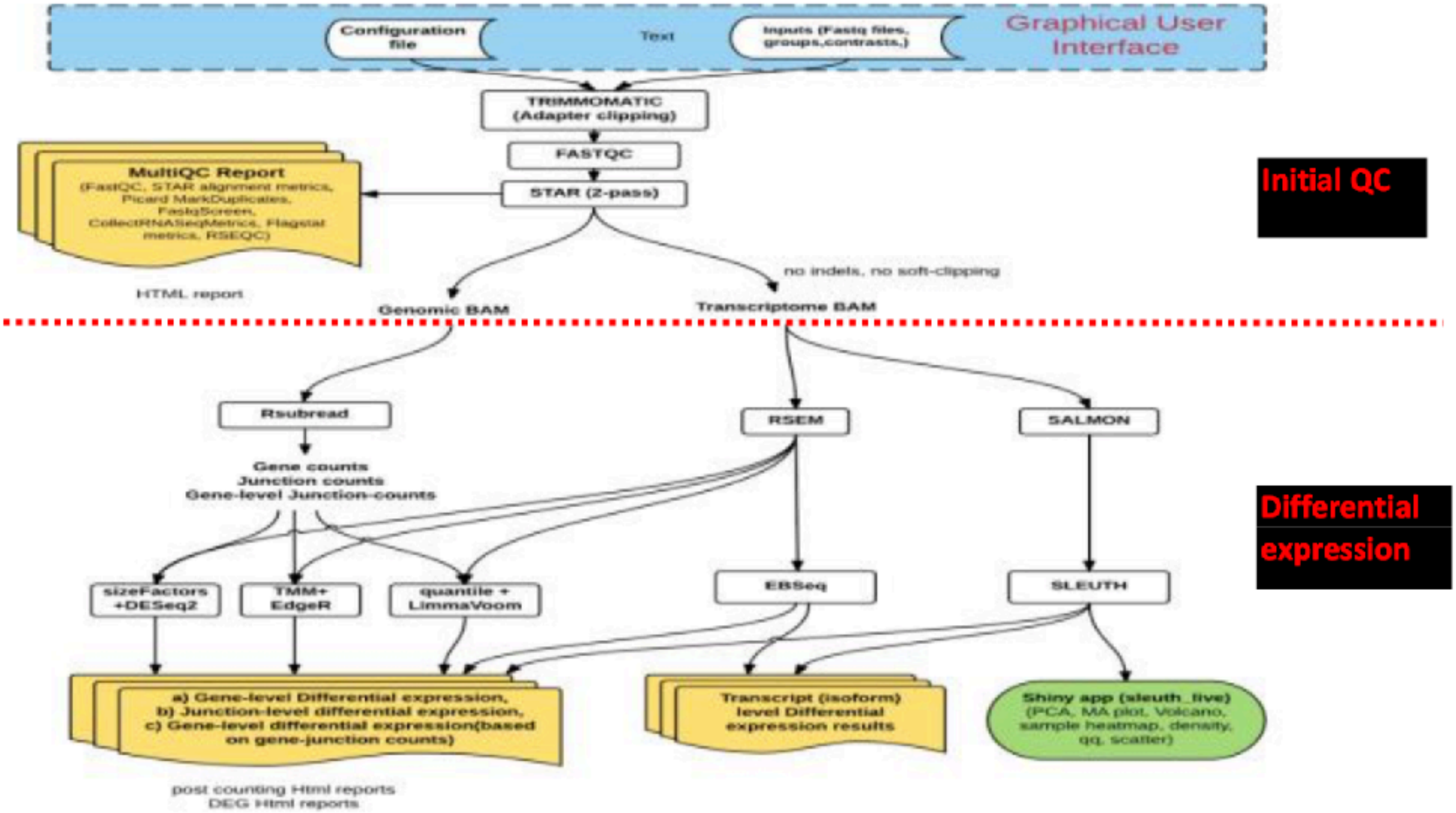
Post-alignment QC & filtering

Quantification

Differential Expression

RNASEQ Pipeline

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


Configuration file
Text
Inputs (Fastq files, groups,contrasts,)
Graphical User Interface
TRIMMOMATIC (Adapter clipping)
FASTQC
STAR (2-pass)
MultiQC Report (FastQC, STAR alignment metrics, Picard MarkDuplicates, FastqScreen, CollectRNASeqMetrics, Flagstat metrics, RSEQC)
HTML report
Initial QC
no indels, no soft-clipping
Genomic BAM
Transcriptome BAM
Rsubread
Gene counts
Junction counts
Gene-level Junction-counts
RSEM
SALMON
Differential expression
sizeFactors +DESeq2
TMM+ EdgeR
quantile + LimmaVoom
EBSeq
SLEUTH
a) Gene-level Differential expression, b) Junction-level differential expression, c) Gene-level differential expression(based on gene-junction counts)
Transcript (isoform) level Differential expression results
Shiny app (sleuth_live) (PCA, MA plot, Volcano, sample heatmap, density, qq, scatter)
post counting Html reports
DEG Html reports

Quality Control/Assesment (Pre-Alignment)

Data Quality Assessment

- **Evaluate the read quality to determine**

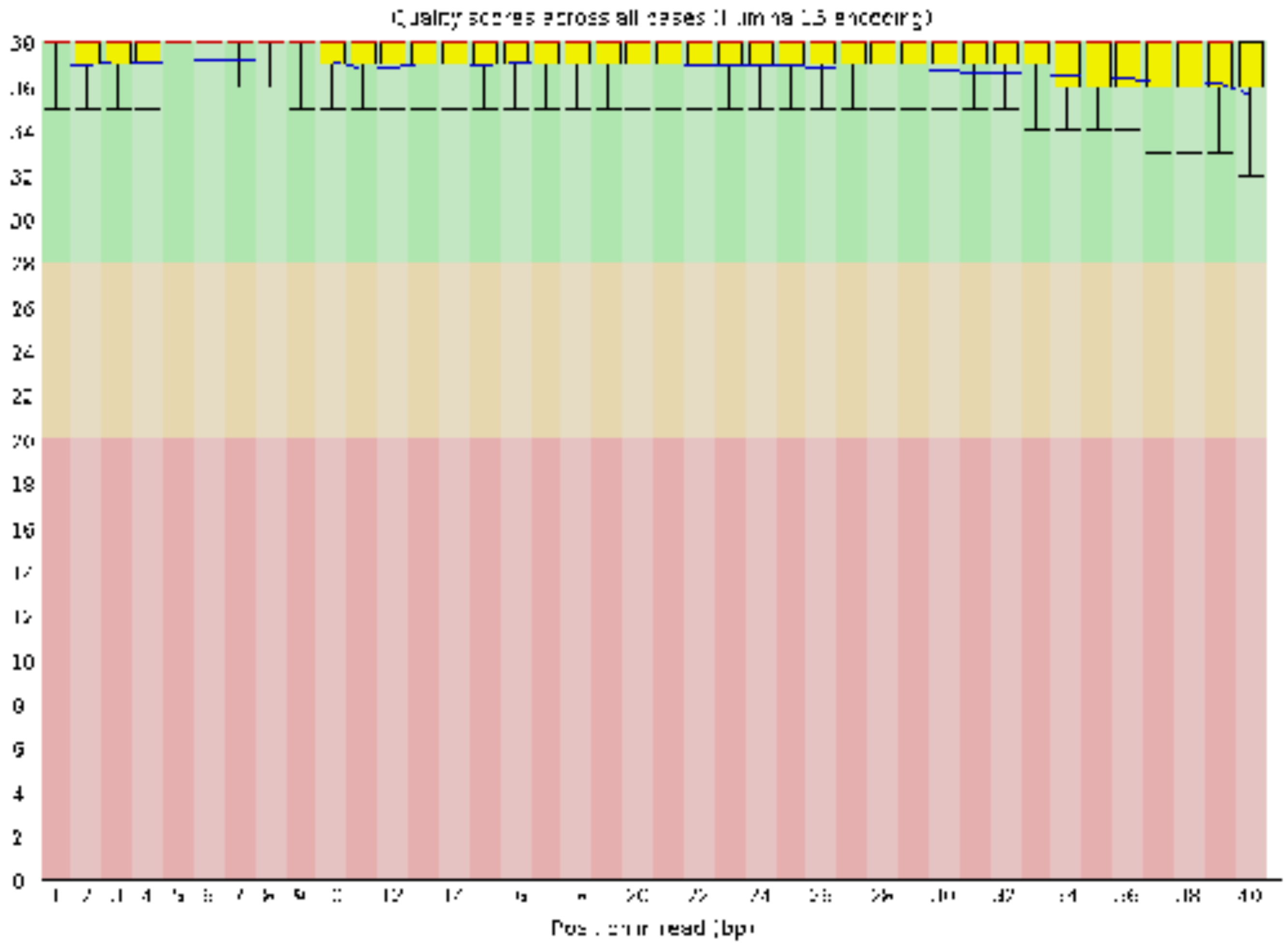
  (Tells us nothing about whether the experiment worked)

  - 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**

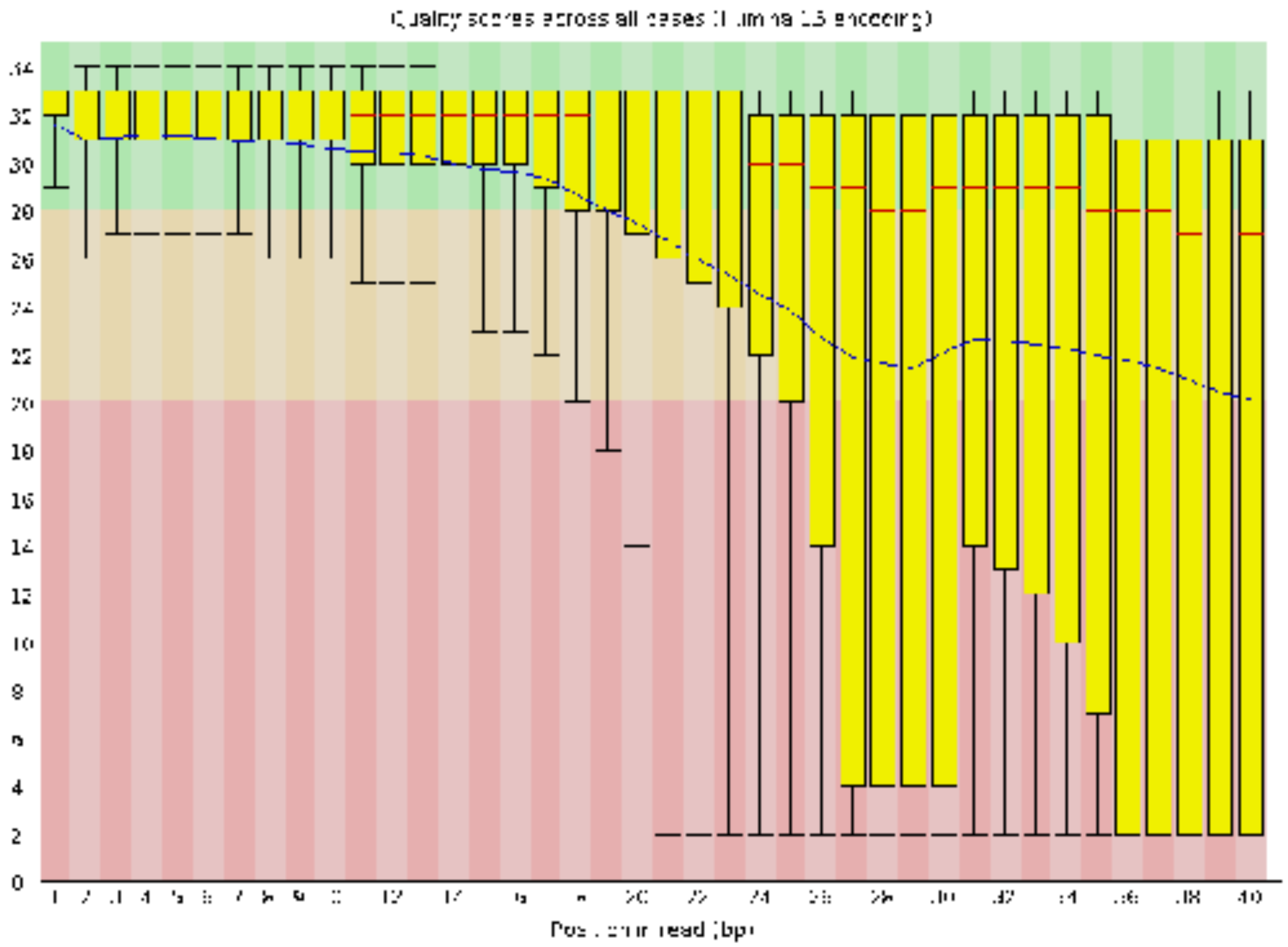
Sequencing cores generally provide some/all of this analysis

FastQC

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



GOOD



BAD

Raw Sequence Cleanup

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

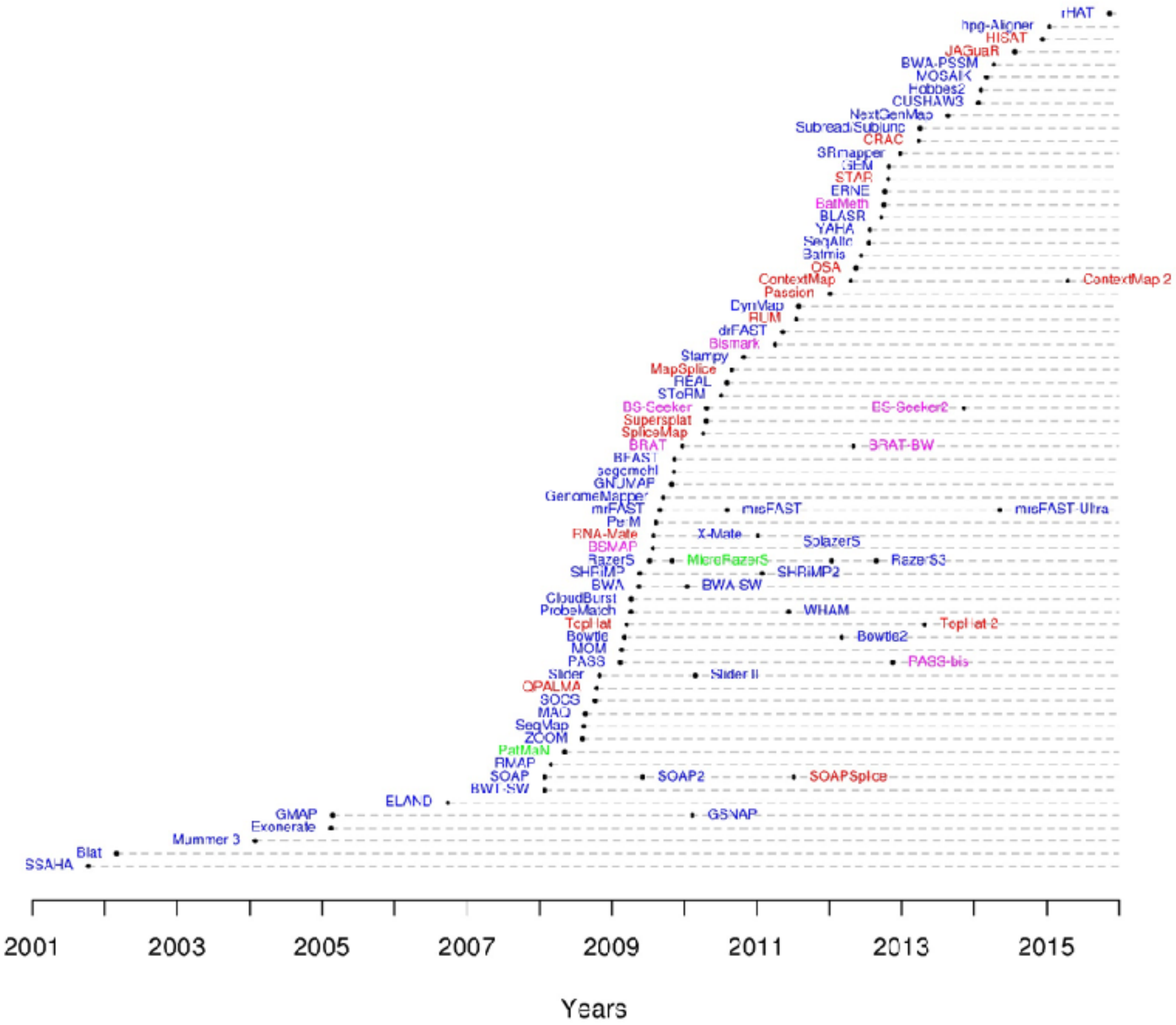
- **Trimmomatic** is a fast, multithreaded command line tool that can be used to trim and crop Illumina (FASTQ) data as well as to remove adapters.

- **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.

- **FASTX-Toolkit** is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.

Alignment


rHAT
hpg-Aligner
HISAT
JAGuaR
BWA-PSSM
MOSAIK
Hobbes2
CUSHAW3
NextGenMap
Subread/Subjunc
CRAC
3Rmapper
GEM
STAR
ERNE
BatMeth
BLASR
YAHA
SeqAlto
Batmis
OSA
ContextMap
ContextMap 2
Passion
GymMap
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
GMAP
GSNAP
Exonerate
Mummer 3
Blat
SSAHA
2001
2003
2005
2007
2009
2011
2013
2015
Years

Common Aligners

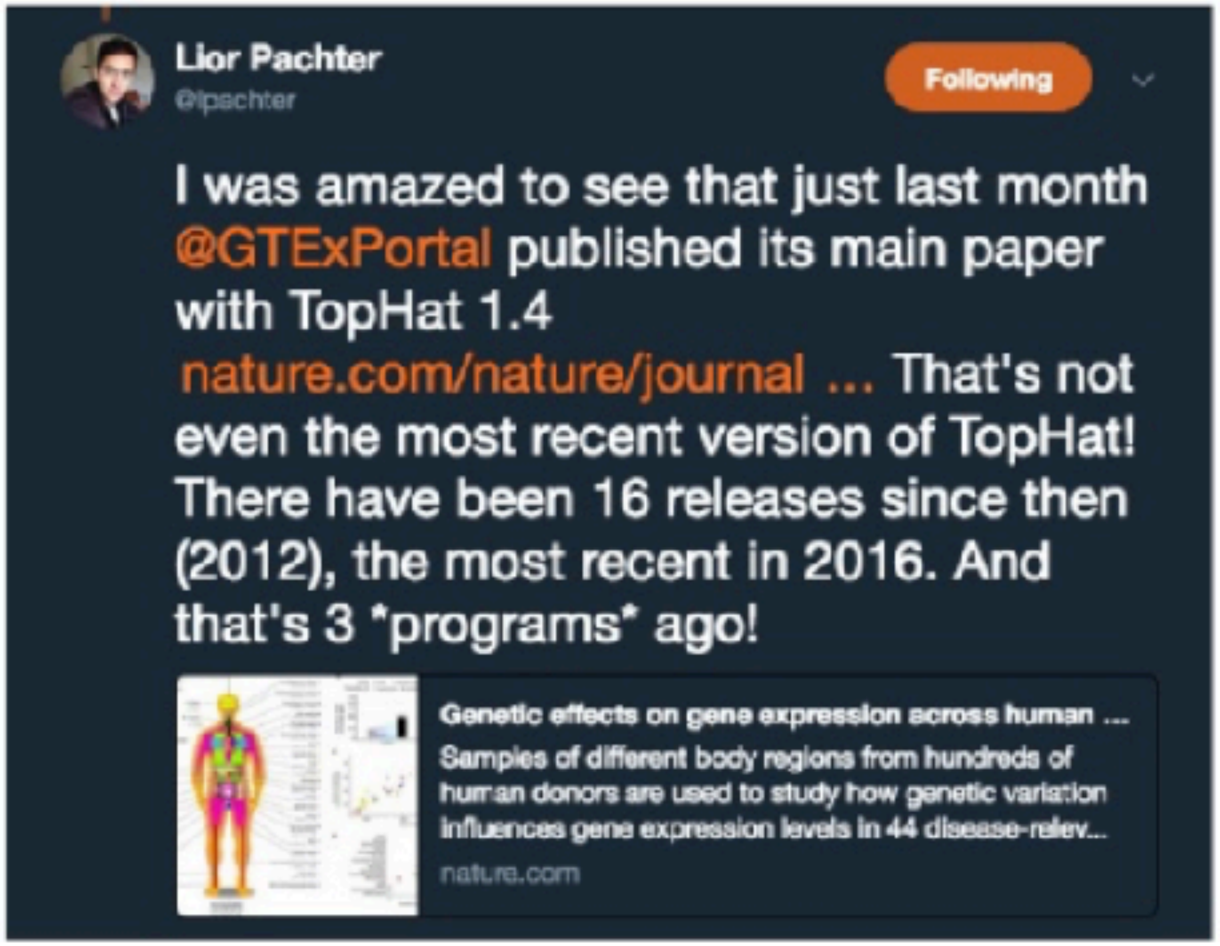
Most alignment algorithms rely on the construction of auxiliary data structures, called indices, which are made for the sequence reads, the reference genome sequence, or both. Mapping algorithms can largely be grouped into two categor on properties of their indices: algorithms based on hash tables, and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

- 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 !!

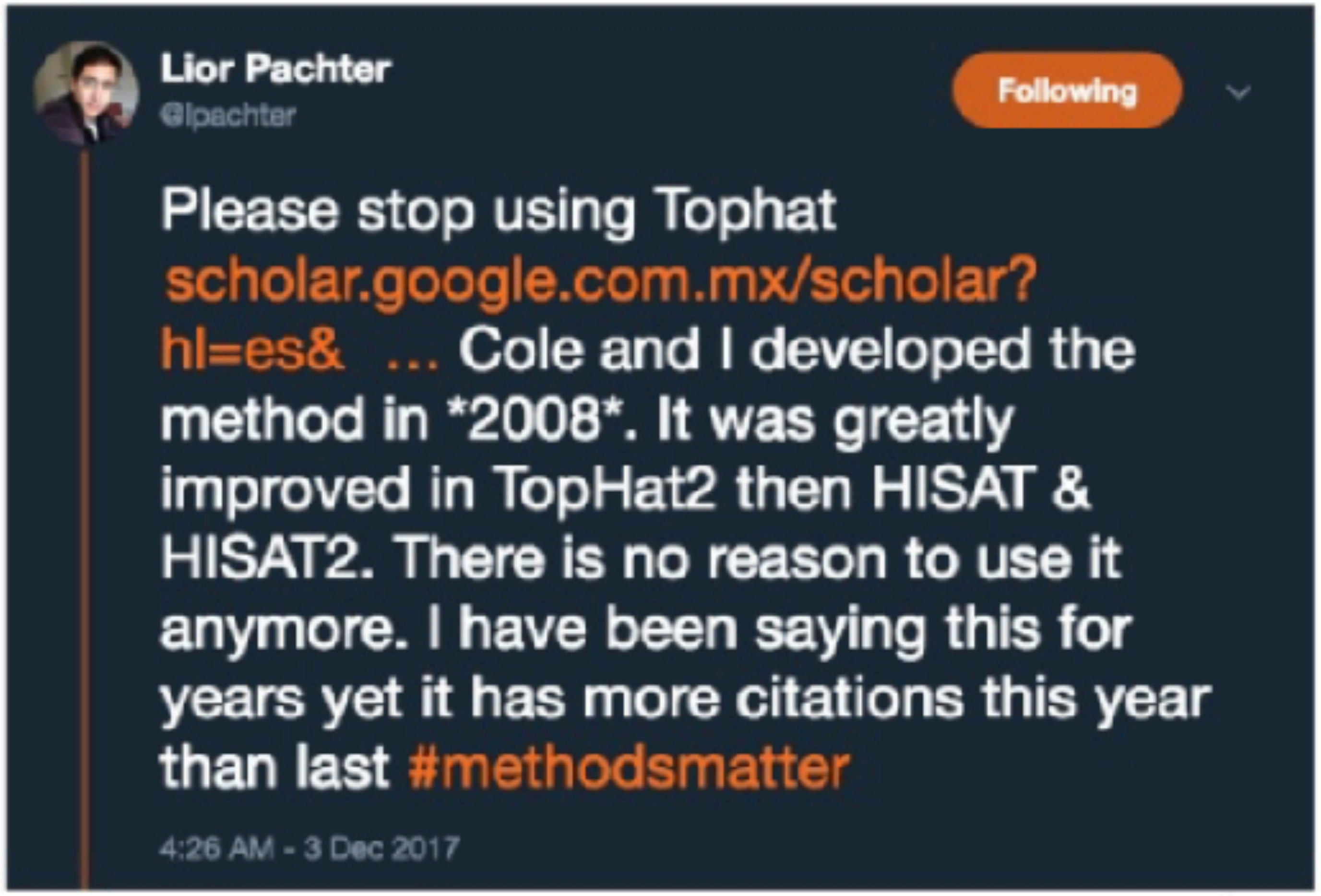
Check or new versions… try new software



Lior Pachter
@lpachter

I was amazed to see that just last month @GTExPortal published its main paper with TopHat 1.4 nature.com/nature/journal … That's not even the most recent version of TopHat! There have been 16 releases since then (2012), the most recent in 2016. And that's 3 *programs* ago!

Genetic effects on gene expression across human …
Samples of different body regions from hundreds of human donors are used to study how genetic variation influences gene expression levels in 44 disease-relev...
nature.com

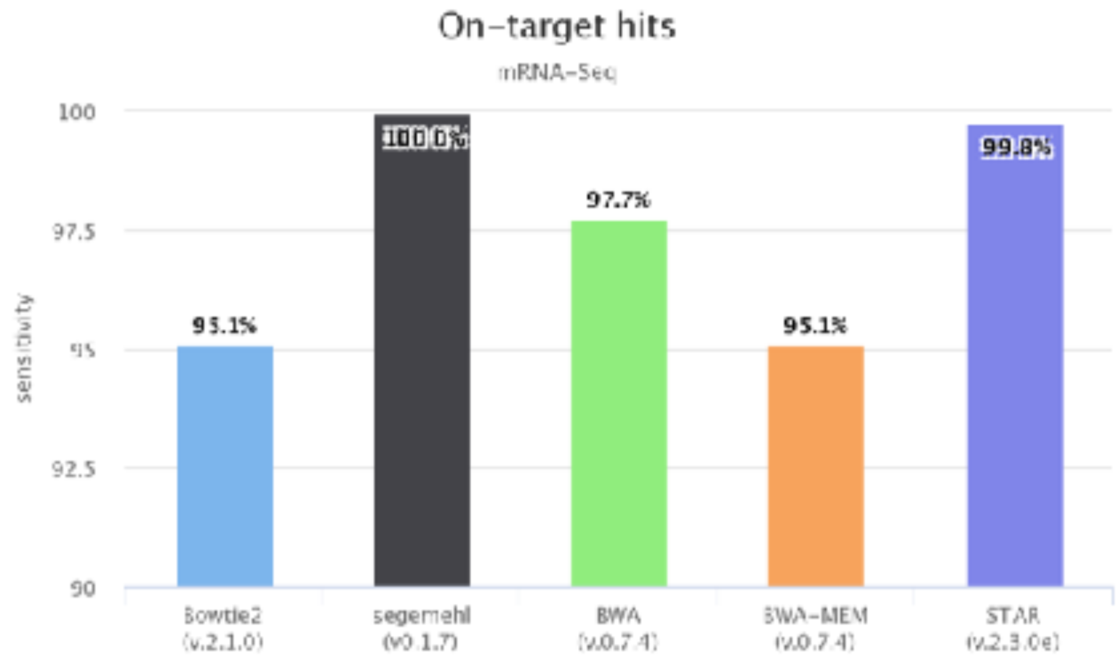


Lior Pachter
@lpachter

Please stop using Tophat scholar.google.com.mx/scholar?hl=es& … Cole and I developed the method in *2008*. It was greatly improved in TopHat2 then HISAT & HISAT2. There is no reason to use it anymore. I have been saying this for years yet it has more citations this year than last #methodsmatter

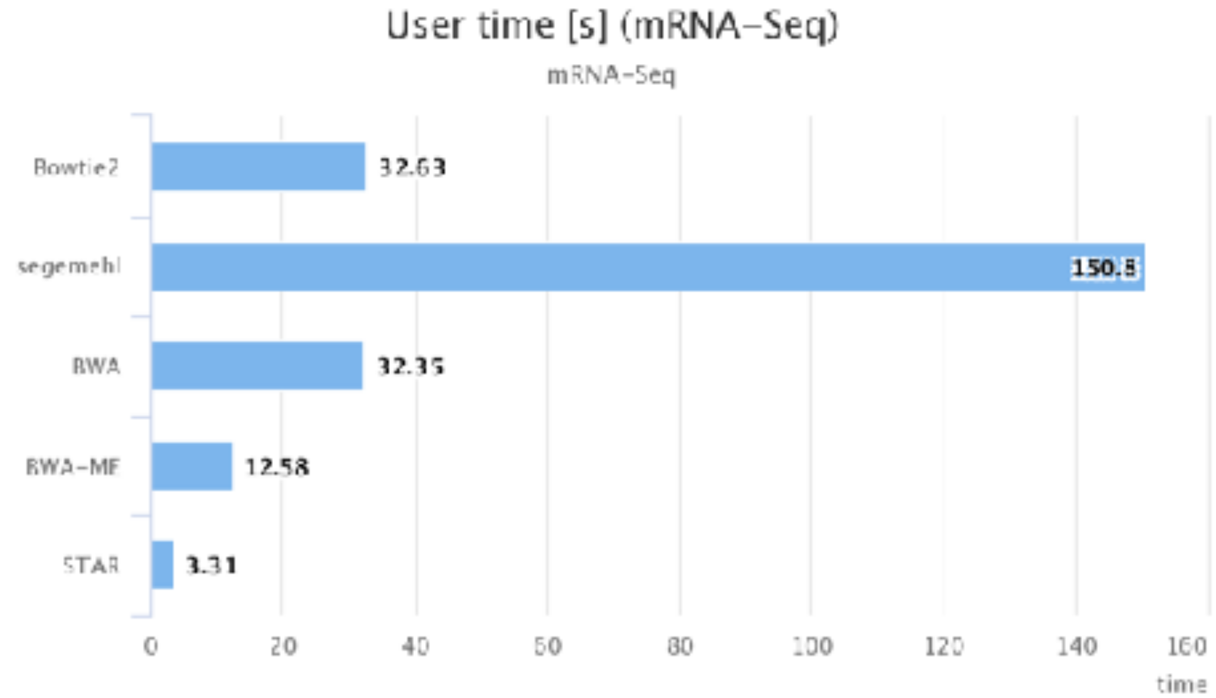
4:26 AM - 3 Dec 2017

Source: Twitter



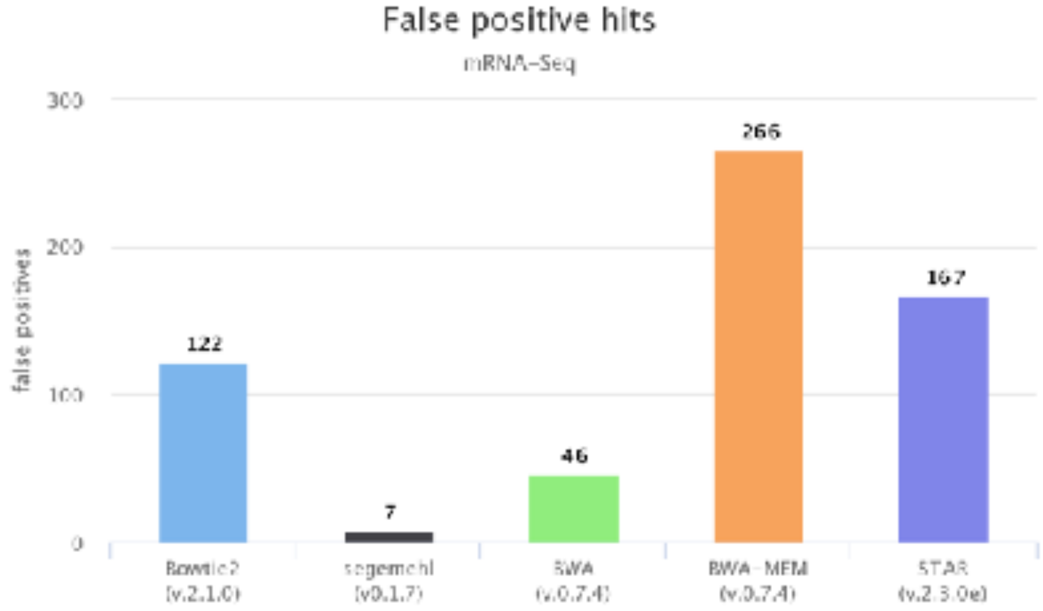
On-target hits

mRNA-Seq



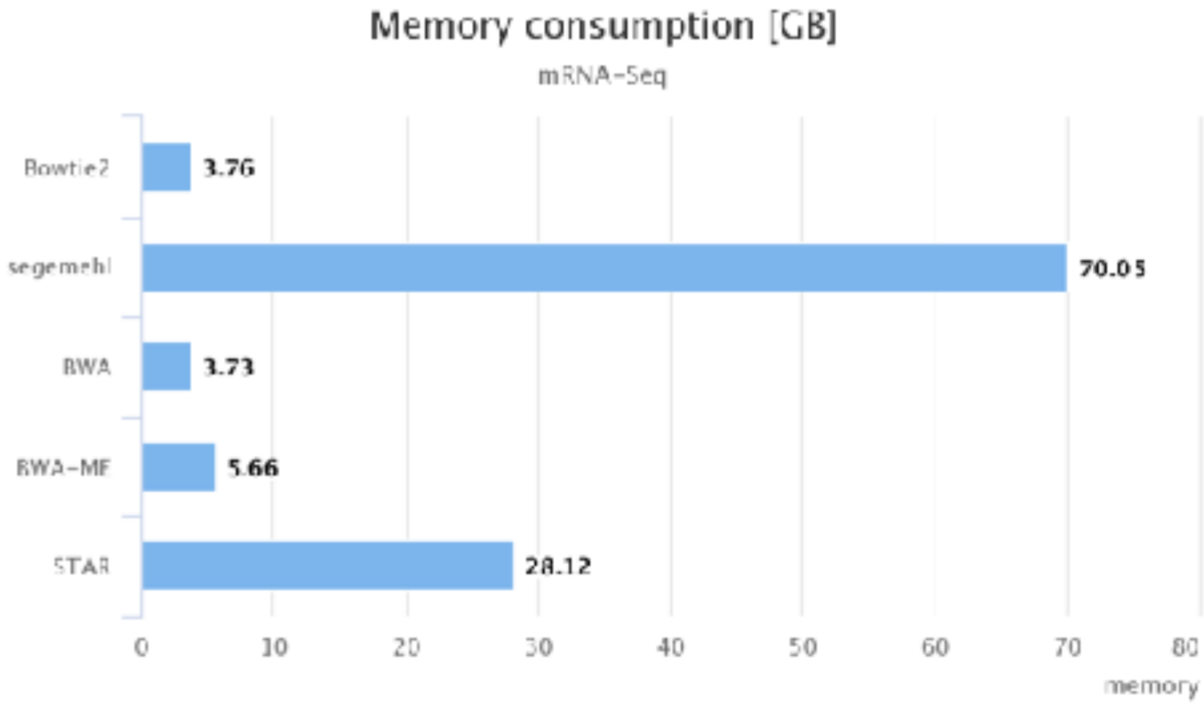
User time [s] (mRNA-Seq)

mRNA-Seq



False positive hits

mRNA-Seq



Memory consumption [GB]

mRNA-Seq

Typical Questions about alignment

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

Answers

- STAR - (also Kalisto or Salmon) - *subjective*
- Depends !
- GeneCode with caveats

Questions not asked

- What parameters should I use?
- What about non-aligners?

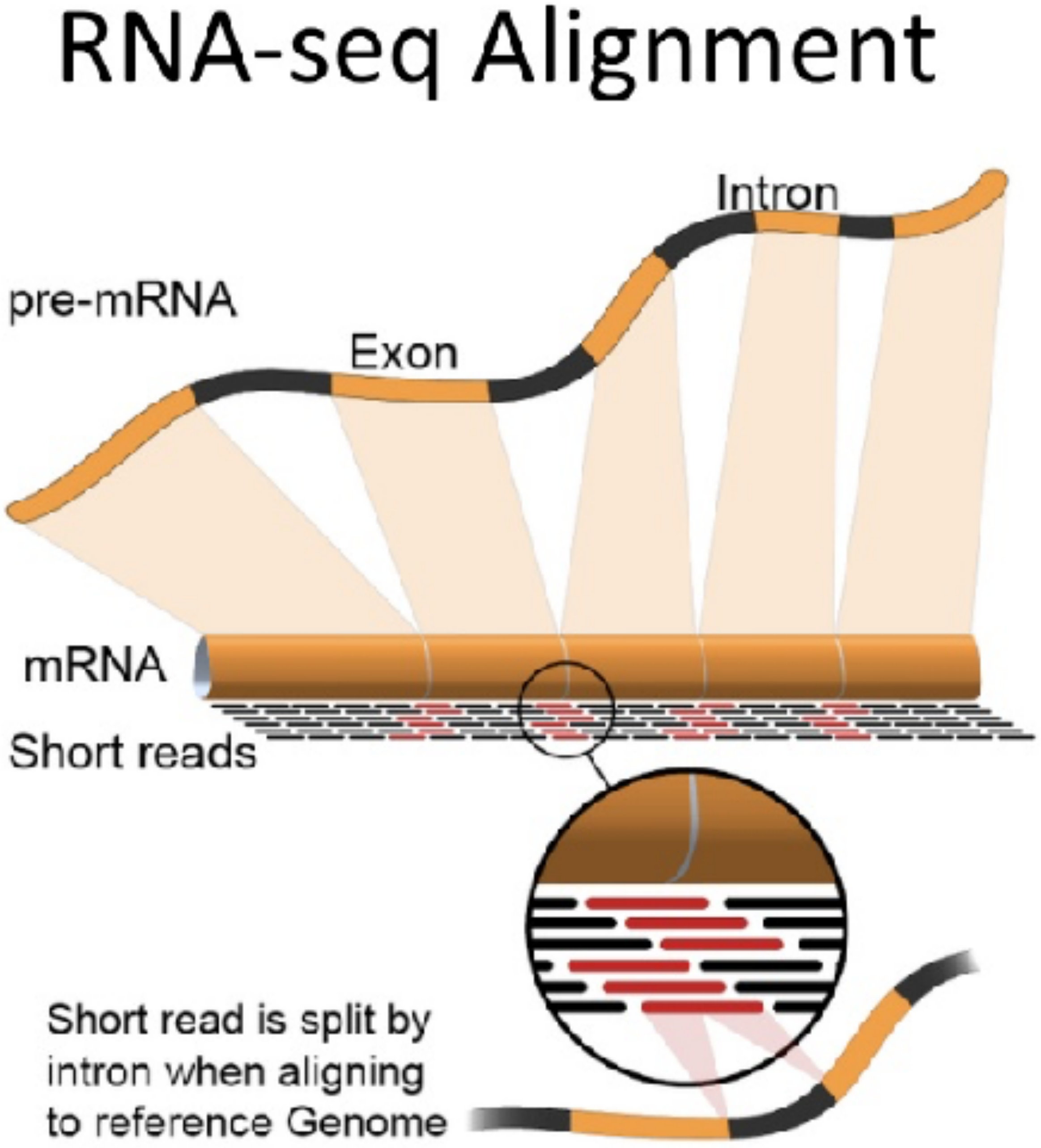
Answers

- Lots ??
- Depends

Additional Parameters For Star In CCBR Pipeliner Program

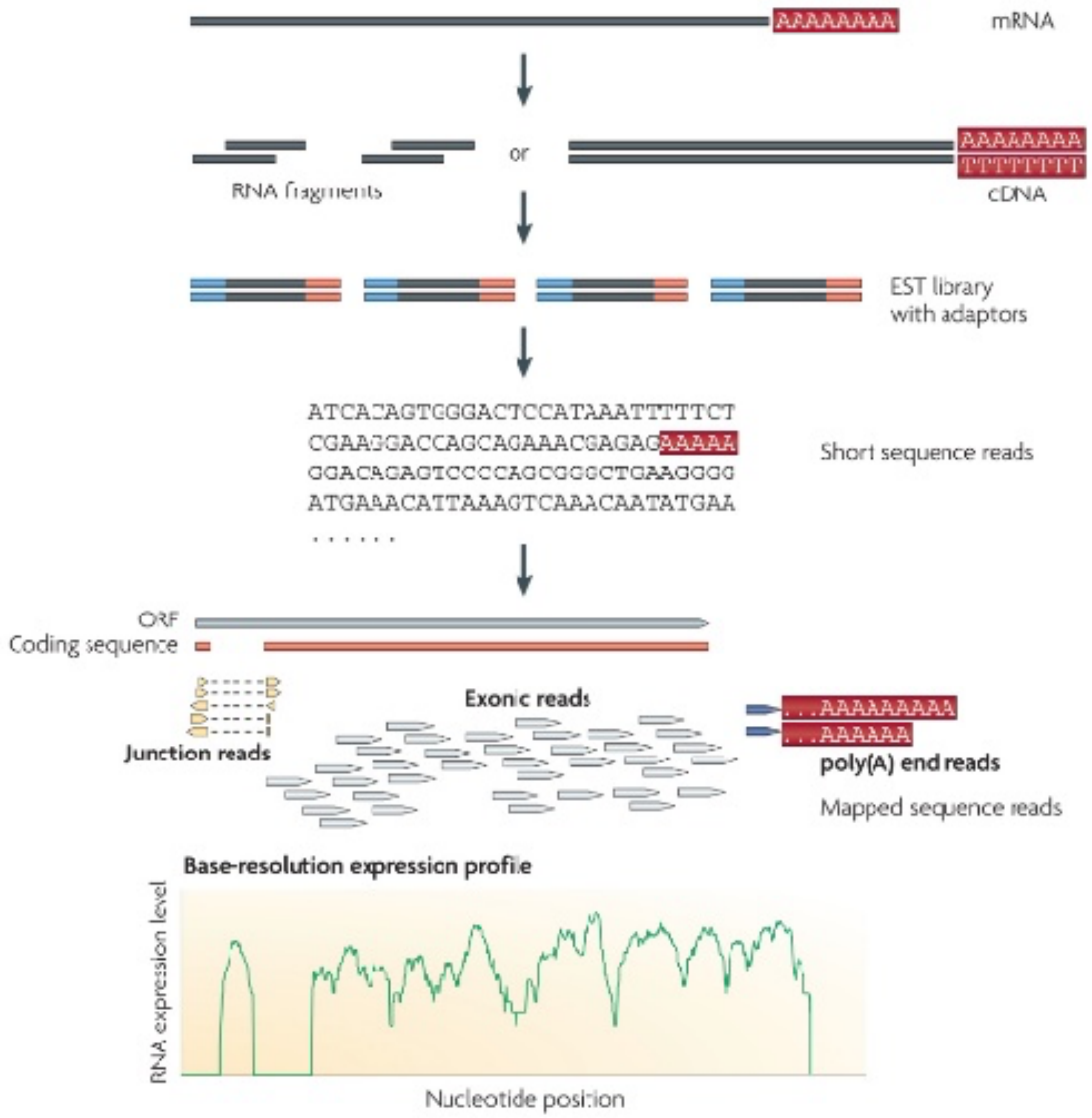
```
STAR
--runThreadN 32
--genomeDir /fdb/STAR_2.4.2a/GENCODE/Gencode_mouse/
release_M4/genes-125
--readFilesIn R1_all.fastq.gz   R2_all.fastq.gz
--readFilesCommand zcat
--limitSjdbInsertNsj 2000000
--outFileNamePrefix Ker_RNA.Rep01.p2.
--outSAMtype BAM   SortedByCoordinate
--outSAMstrandField None
--outSAMunmapped 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
```

RNASeq Mapping Challenges


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

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

RNA-seq protocol schematic

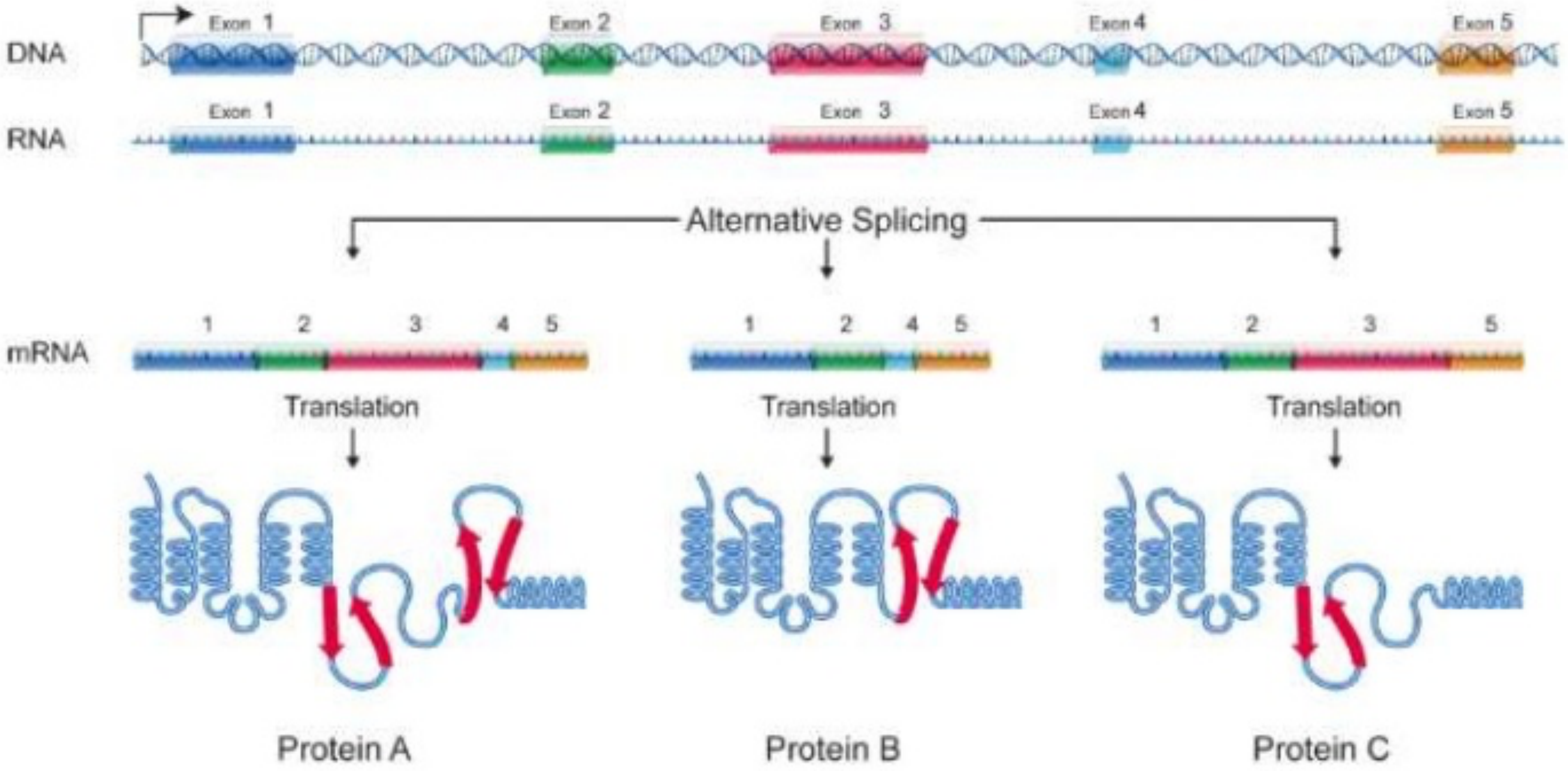


Mapping Challenges

- 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)

RNA-Seq: Special Mapping Concerns

Alternate Splicing



genome.gov

RNASeq Mapping Solutions

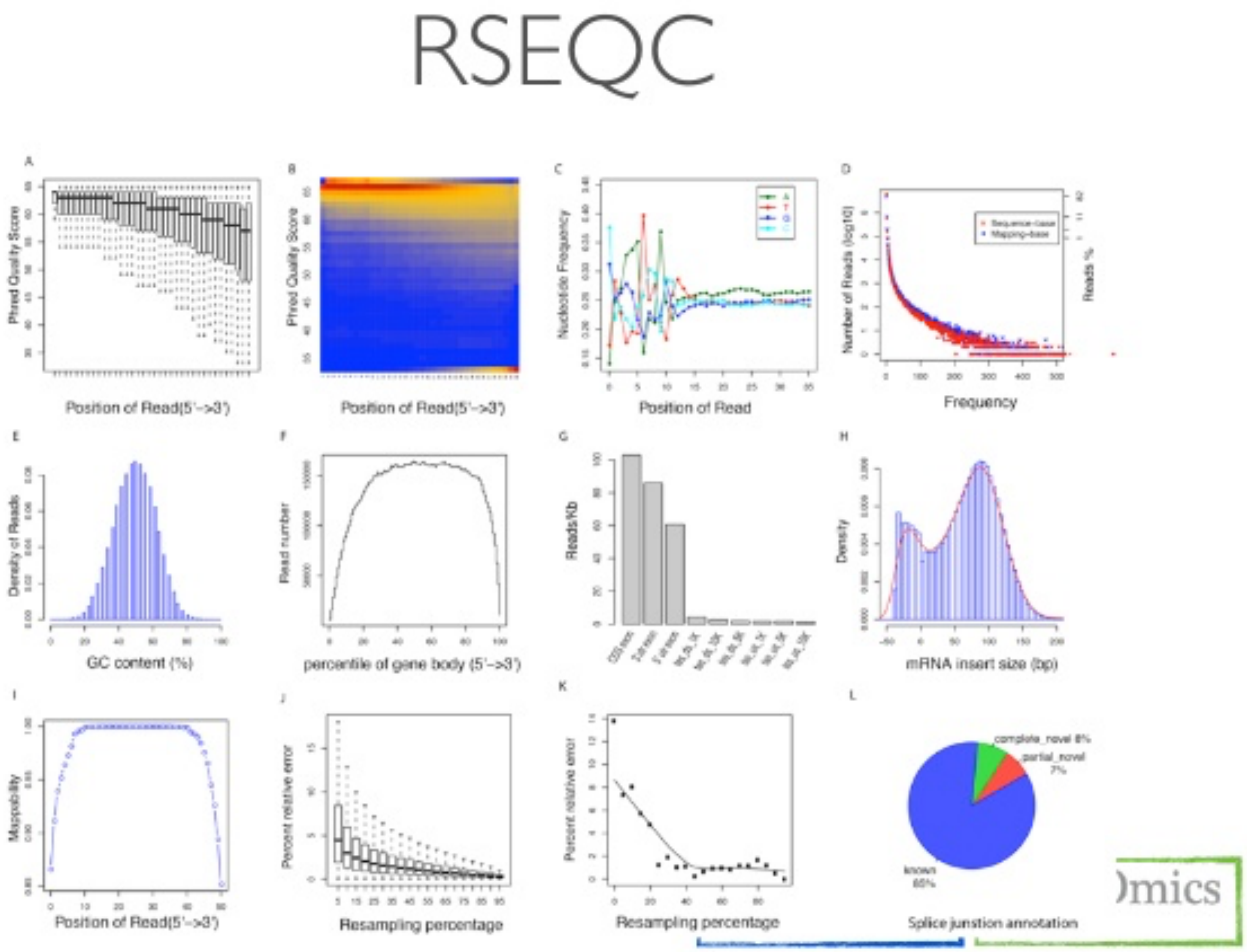
- **Align against the transcriptome**
  - Many/All transcriptomes are incomplete
  - Can only measure known genes
  - Won’t detect non-coding RNAs
  - Can’t look at splicing variants
  - Can’t detect fusion genes or structure variants
- **De novo assembly of RNASeq reads**
  - Largely used for uncharacterized 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

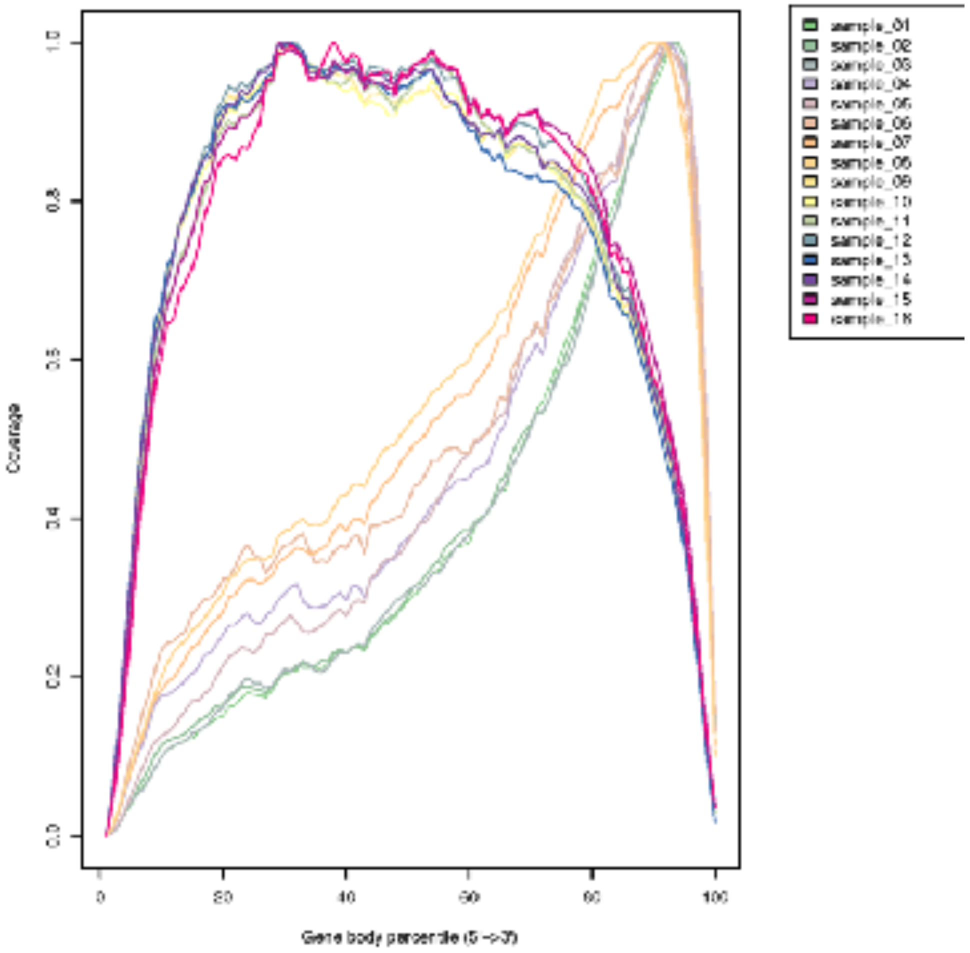
Post Alignment QC

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.

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.

ALIGNMENT FILE FORMATS

SAM FORMAT

```
8_100_10000_12419      163     chrVII  271183  255     40M     =       271294  151
TGGTGTATTATACGCTACCGTGCGGTGCCGGGGCAACCG
bbbabbbbbbbbbbbbbbbcbbbbcbbbbbbbbbbbbb       XA:i:0  MD:Z:40 NM:i:0
```

The **SAM Format** (**S**equence **A**lignment/**M**ap) is a text format for storing sequence alignment data in a series of tab delimited ASCII columns.

The file has two parts:

1. **Header** - Each line starts with a "@".
   @HD, @SQ, @RG, @PG
2. **Alignments** - One line for each entry.

ALIGNMENT FILE FORMATS

SAM FORMAT

Example of SAM Header

```
@HD  VN:1.0      SO:unsorted
@SQ  SN:chr1     LN:195471971
@SQ  SN:chr2     LN:182113224
@SQ  SN:chr3     LN:160039680
@SQ  SN:chr4     LN:156508116
@SQ  SN:chr5     LN:151834684
@SQ  SN:chr6     LN:149736546
@SQ  SN:chr7     LN:145441459
@SQ  SN:chr8     LN:129401213
@SQ  SN:chr9     LN:124595110
@SQ  SN:chr10    LN:130694993
@SQ  SN:chr11    LN:122082543
@SQ  SN:chr12    LN:120129022
@SQ  SN:chr13    LN:120421639
@SQ  SN:chr14    LN:124902244
@SQ  SN:chr15    LN:104043685
@SQ  SN:chr16    LN:98207768
@SQ  SN:chr17    LN:94987271
@SQ  SN:chr18    LN:90702639
@SQ  SN:chr19    LN:61431566
@SQ  SN:chrX     LN:171031299
@SQ  SN:chrY     LN:91744698
@SQ  SN:chrM     LN:16299
@PG  ID:bowtie2  PN:bowtie2  VN:2.2.9    CL:"/usr/local/apps/bowtie/2-2.2.9/bowtie2-align-s --wrapper basic-0 -x /fdb/bowtie2.DELETE/mm10 -q jun_minus_dex_rep1a -S jun_minus_dex_rep1a_mm10.sam -p8"
```

ALIGNMENT FILE FORMATS

SAM FORMAT

```
8_100_10000_12419     163     chrVII  271183  255     40M     =      271294  151
TGGTGTATTATACGCTACCGTGCGGTGCCGGGGCAACCG
bbbabbbbbbbbbbbbbbbcbbbcbbbbbbbbbbbbb      XA:i:0  MD:Z:40 NM:i:0
```

8_100_10000_12	163	chr7	271183	255	40M	=	271294	151	TGGTGTA TTATACG	bbbabbbb bbbbbbbb	XA:i:0 MD:Z:40
QNAME	**FLAG**	**RNAME**	**POS**	**MAPQ**	**CIGAR**	**MRNM**	**MPOS**	**TLEN**	**SEQ**	**QUAL**	**OPT**

Col	Field	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe (‘=’ if same as RNAME)
8	MPOS	1-based Mate POSistion
9	TLEN	inferred Template LENgth (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

ALIGNMENT FILE FORMATS

SAM FORMAT

```
8_100_10000_12419      163     chrVII  271183  255     40M     =      271294  151
TGGTGTATTATACGCTACCGTGCGGTGCCGGGGGCAACCG
bbbabbbbbbbbbbbbbbbbcbbbbcbbbbbbbbbbbbbb      XA:i:0  MD:Z:40 NM:i:0
```

Understanding Flag codes

http://broadinstitute.github.io/picard/explain-flags.html

1	read paired
2	read mapped in proper pair
4	read unmapped
8	mate unmapped
16	read reverse strand
32	mate reverse strand
64	first in pair
128	second in pair
256	not primary alignment
512	read fails platform/vendor quality checks
1024	read is PCR or optical duplicate
2048	supplementary alignment

ALIGNMENT FILE FORMATS

BAM/CRAM FORMAT

BAM (*.bam) is the compressed binary version of the Sequence Alignment/Map (SAM) format, a compact and index-able representation of nucleotide sequence alignments. **BAM** is compressed in the **BGZF** format that supports random access through the BAM file index (*.bam.bai).

HINT: Filename.bam and filename.bai always go together

CRAM (*.cram) - newer implementation of BAM like binary data.

1. Significantly better lossless compression than BAM
2. Full compatibility with BAM
3. Effortless transition to CRAM from using BAM files
4. Support for controlled loss of BAM data

ANNOTATION FILE FORMATS

BED FORMAT

1. **chrom** - name of the chromosome
2. **chromStart** - Start of feature (0-based)
3. **chromEnd** - End of the feature (not included in display)

+ 9 optional columns - most common are:

4. **name** - a label for the feature
5. **score** - a score (0-1000)
6. **strand** - which strand the feature on (+/-)

chr1	15000	20000	gene1	50	+
chr2	106000	108000	gene2	400	-

ANNOTATION FILE FORMATS

BED FORMAT

7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays). When there is no thick part, thickStart and thickEnd are usually set to the chromStart position.
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line itemRgb attribute is set to "On", this RBG value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. **blockCount** - The number of blocks (exons) in the BED line.
11. **blockSizes** - A comma-separated list of the block sizes. The number of items in this list should correspond to blockCount.
12. **blockStarts** - A comma-separated list of block starts. All of the blockStart positions should be calculated relative to chromStart. The number of items in this list should correspond to blockCount.

ANNOTATION FILE FORMATS

GFF FORMAT

GFF (General Feature Format) GFF lines have nine required fields that *must* be tab-separated [GFF2 - UCSC & GFF3 - EMBL]

1. **squid** - The name of the chromosome or scaffold.
2. **source** - The program that generated this feature.
3. **feature** - The name of this type of feature. Some examples of standard feature types are "CDS" "start_codon" "stop_codon" and "exon"li>
4. **start** - The starting position of the feature in the sequence. The first base is numbered 1.
5. **end** - The ending position of the feature (inclusive).
6. **score** - floating point value
7. **strand** - Valid entries include "+", "-", or "." (for don't know/don't care).
8. **phase** - If the feature is a coding exon, frame should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be ".".
9. **attributes**- A list of feature attributes in the format tag=value pairs separated by **";"**

GFF2 http://genome.ucsc.edu/FAQ/FAQformat.html#format3
GFF3 https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md
http://useast.ensembl.org/info/website/upload/gff3.html

ANNOTATION FILE FORMATS

GFF FORMAT

GFF example

```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS             5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS             7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

ANNOTATION FILE FORMATS

GTF FORMAT

GTF (Gene Transfer Format) is a refined form of the GFF with group attributes - essentially the same as GFF2

1. **seqname** - The name of the sequence. Must be a chromosome or scaffold. (chr1 or 1)
2. **source** - The program that generated this feature.
3. **feature** - The name of this type of feature. Some examples of standard feature types are "CDS" "start_codon" "stop_codon" and "exon"li>
4. **start** - The starting position of the feature in the sequence. The first base is numbered 1.
5. **end** - The ending position of the feature (inclusive).
6. **score** - A score between 0 and 1000 (UCSC) **OR** floating point value
7. **strand** - Valid entries include "+", "-", or "." (for don't know/don't care).
8. **frame** - If the feature is a coding exon, frame should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be ".".
9. **attributes/group** - A list of feature attributes in the format tag=value pairs separated by “**;**”

GTF/GFF2 http://useast.ensembl.org/info/website/upload/gff.html

GRAPHING FILE FORMATS

WIG (BIGWIG) FORMAT

1) FixedStep

fixedStep	chrom=chr1 start=3001 step=1
24	
56	
100	

2) VariableStep

variableStep	chrom=chr1
3001	24
3002	56
3003	100

variableStep	chrom=chr1
3001	24
3003	56
3010	100

GRAPHING FILE FORMATS

BEDGRAPH FORMAT

1. **chrom** - name of the chromosome
2. **chromStart** - Start of feature (0-based)
3. **chromEnd** - End of the feature (not included in display)
4. **score** - a score (integer or real positive / negative number)

chr1	15000	20000	1
chr2	106000	108000	0.75

Format Conversion Utilities

- Galaxy (http://galaxy.psu.edu/ - http://galaxy.cit.nih.gov/)
  - Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.
- Samtools (http://samtools.sourceforge.net)
  - SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format. Also, note TABIX for indexing generic tab delimited files.
- Picard (http://picard.sourceforge.net/)
  - Picard comprises Java-based command-line utilities that manipulate SAM files, and a Java API (SAM-JDK) for creating new programs that read and write SAM files. Both SAM text format and SAM binary (BAM) format are supported.
- UCSC Utilities (http://hgdownload.cse.ucsc.edu/admin/exe/)

Format Conversion Utilities

- Bamtools -(https://github.com/pezmaster31/bamtools)
  - BamTools provides both a programmer's API and an end-user's toolkit for handling BAM files.
- Bedtools (http://bedtools.readthedocs.io/en/latest/)
  - Collectively, the bedtools utilities are a swiss-army knife of tools for a wide-range of genomics analysis tasks. The most widely-used tools enable genome arithmetic: that is, set theory on the genome. For example, bedtools allows one to intersect, merge, count, complement, and shuffle genomic intervals from multiple files in widely-used genomic file formats such as BAM, BED, GFF/GTF, VCF. While each individual tool is designed to do a relatively simple task (e.g., intersect two interval files), quite sophisticated analyses can be conducted by combining multiple bedtools operations on the UNIX command line.
- FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit//)
  - The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.
- SRA ToolKit (https://github.com/ncbi/sra-tools)
  - The SRA Toolkit and SDK from NCBI is a collection of tools and libraries for using data in the INSDC Sequence Read Archives.

Binary Formats & Indices

Indexed binary file formats are much more efficient.

Only the portions of the files needed for the region currently being processed or visualized are transferred and loaded as needed. Thus for large data sets they are considerably faster than regular files.
(e.g. bigBED, bigWIG, BAMindexed)

Quantitation

Counting as a measure of Expression

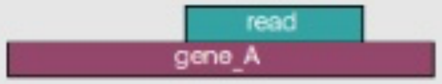
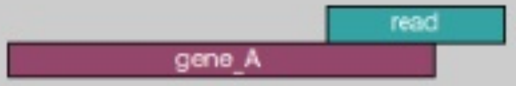
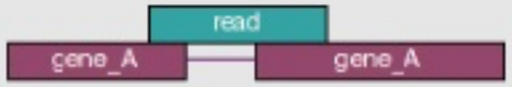
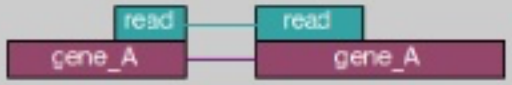
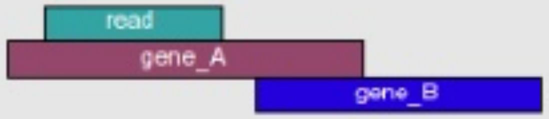
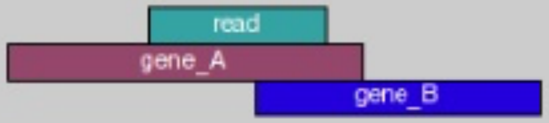
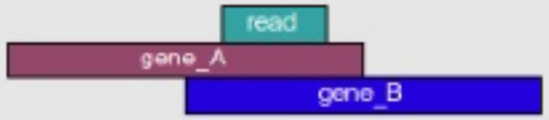
- Most RNASEQ techniques deal with count data.
  The reads are mapped to a reference and the number of reads mapped to each gene/transcript is counted
- Read counts are roughly proportional to gene-length and abundance
- The more reads the better

- Artifacts occur because of:
  - Sequencing Bias
  - Positional bias along the length of the gene
  - Gene annotations (overlapping genes)
  - Alternate splicing
  - Non-unique genes
  - Mapping errors

Counting as a measure of Expression

- Count mapped **reads**
- Count each read once (deduplicate)
- Discard reads that:
  - The alignment has a poor quality score
  - Are not uniquely mapped
  - Alignment overlaps several genes
  - Pair reads do not map together
  - Document what was done

Read Counting

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Counting as a measure of Expression

- Subread (featureCount)
- STAR (quantmode)
- HTseq (counts)
- RSEM (RNA-Seq by Expectation Maximization

Differential Expression

Differential Expression

Differential expression involves the comparison of **normalized** expression counts of different samples and the application of **statistical measures** to identify quantitative changes in gene expression between the different samples.

The two most important step are:

- The normalization of the data to ensure all samples are comparable (variable gene length, read depth)
- The statistical test that determines whether an observed difference is statistically significant (i.e. the likelihood of the observation is greater than that expected from random biological variability).

Differential Expression

Biological replicates are essential to derive a meaningful result. Don’t mistake the high precision of the technique for the need for biological replicates.

Final output generally a rank order list of differentially expressed (DE) genes with expression values with associated p-values.

If technical or biological variability exceeds that of the experimental perturbation you will get zero DEs.

Remember not all DE may be directly due to the experimental perturbation, but could be do to cascading effects of other genes.

Inferring Differential Expression (DE)

Method	Normalization	Needs replicas	Input	Statistics for DE	Availability
edgeR	Library size	Yes	Raw counts	Empirical Bayesian estimation based on Negative binomial distribution	R/Bioconductor
DESeq	Library size	No	Raw counts	Negative binomial distribution	R/Bioconductor
baySeq	Library size	Yes	Raw counts	Empirical Bayesian estimation based on Negative binomial distribution	R/Bioconductor
LIMMA	Library size	Yes	Raw counts	Empirical Bayesian estimation	R/Bioconductor
CuffDiff	RPKM	No	RPKM	Log ratio	Standalone

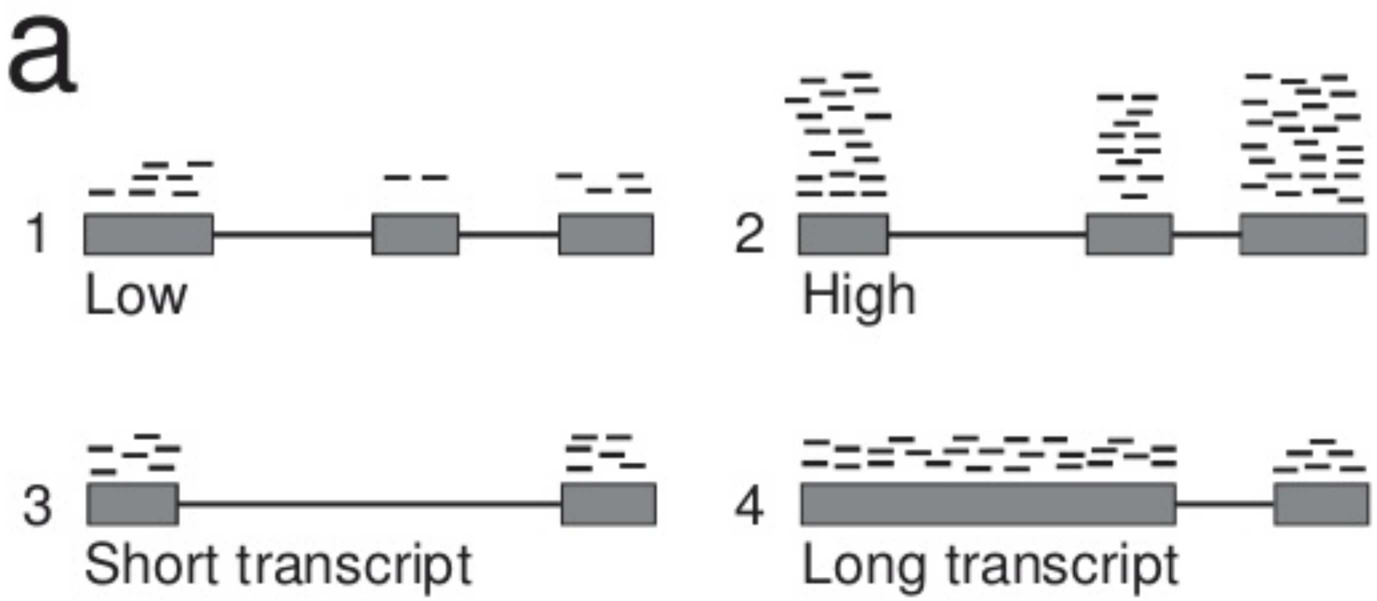
Models for RNA-seq

- Count-based models
- Multi-reads (isoform resolution)
- Paired-end reads (include length resolution step)
- Positional bias along transcript length
- Sequence bias

Count Normalization

- Number of reads aligned to a gene gives a measure of its level of expression

- Normalization of the count data
  - Sequencing depth
  - Length bias



Nature Methods 8, 469–477 (2011), Doi:10.1038/nmeth.1613

Normalization

There are three metrics commonly used to attempt to normalize for sequencing depth and gene length.

- **RPKM = Reads Per Kilobase Million**

  Total Reads/1,000,000 = PM
  Gene read-count/PM = RPKM
  RPM/gene-length (kb) = RPKM

- **FPKM = Fragments Per Kilobase Million**

FPKM is very similar to RPKM. RPKM was made for single-end RNASEQ, where every read corresponded to a single fragment that was sequenced. FPKM was made for paired-end RNA-seq.

- **TPM = Transcripts Per Million (***Sum of all TPM in samples is the same***)**

TPM is very similar to RPKM and FPKM. The only difference is the order of operations

Gene read-count/gene-length (kb) = RPK
(Sum all RPKs)/1,000,000 = PM
Gene RPK/PM = TPM

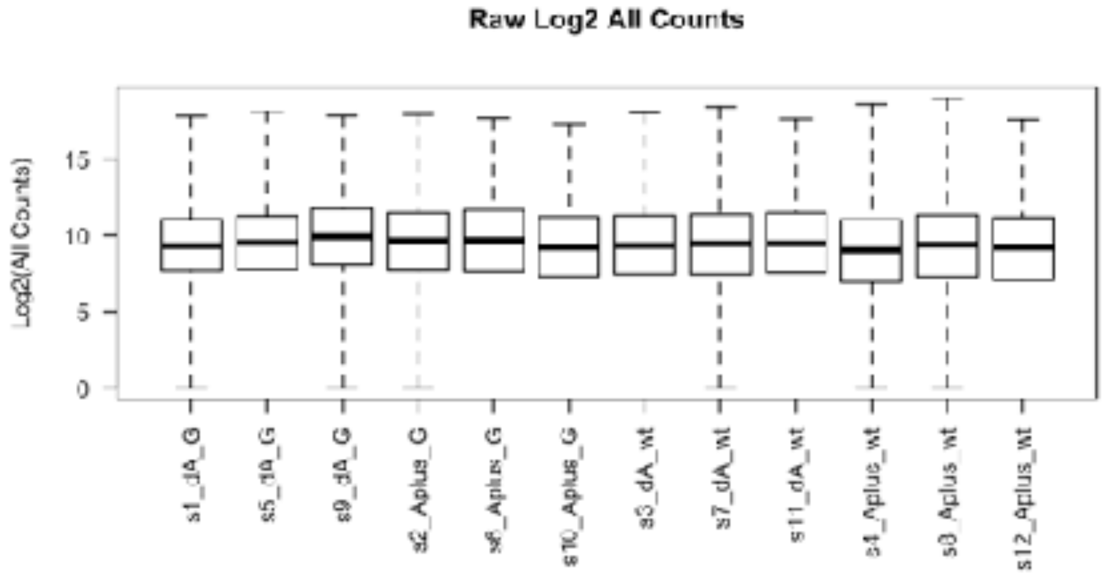
https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

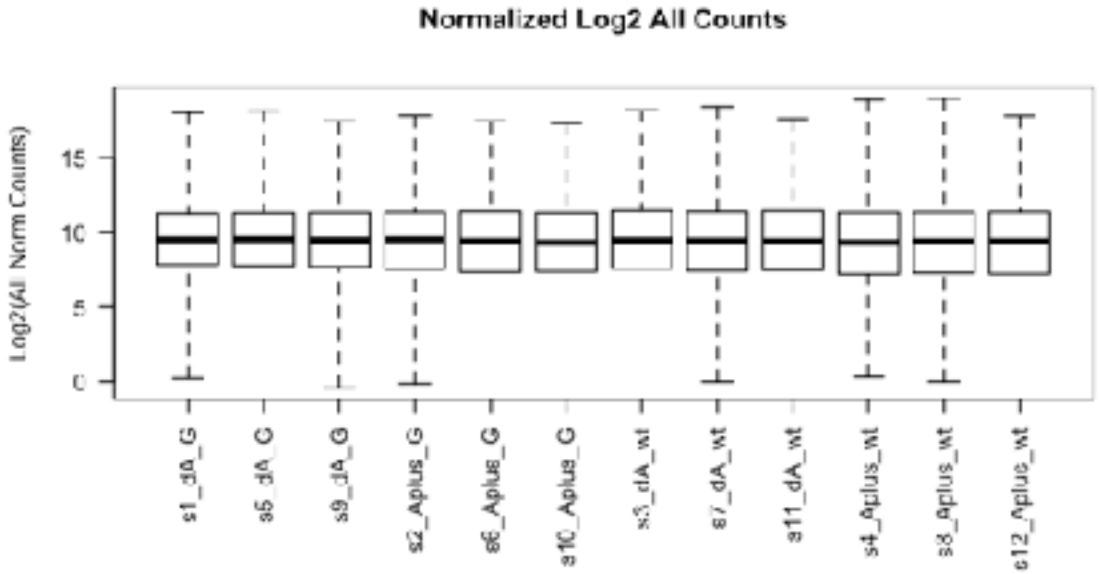
Multiple Testing Correction

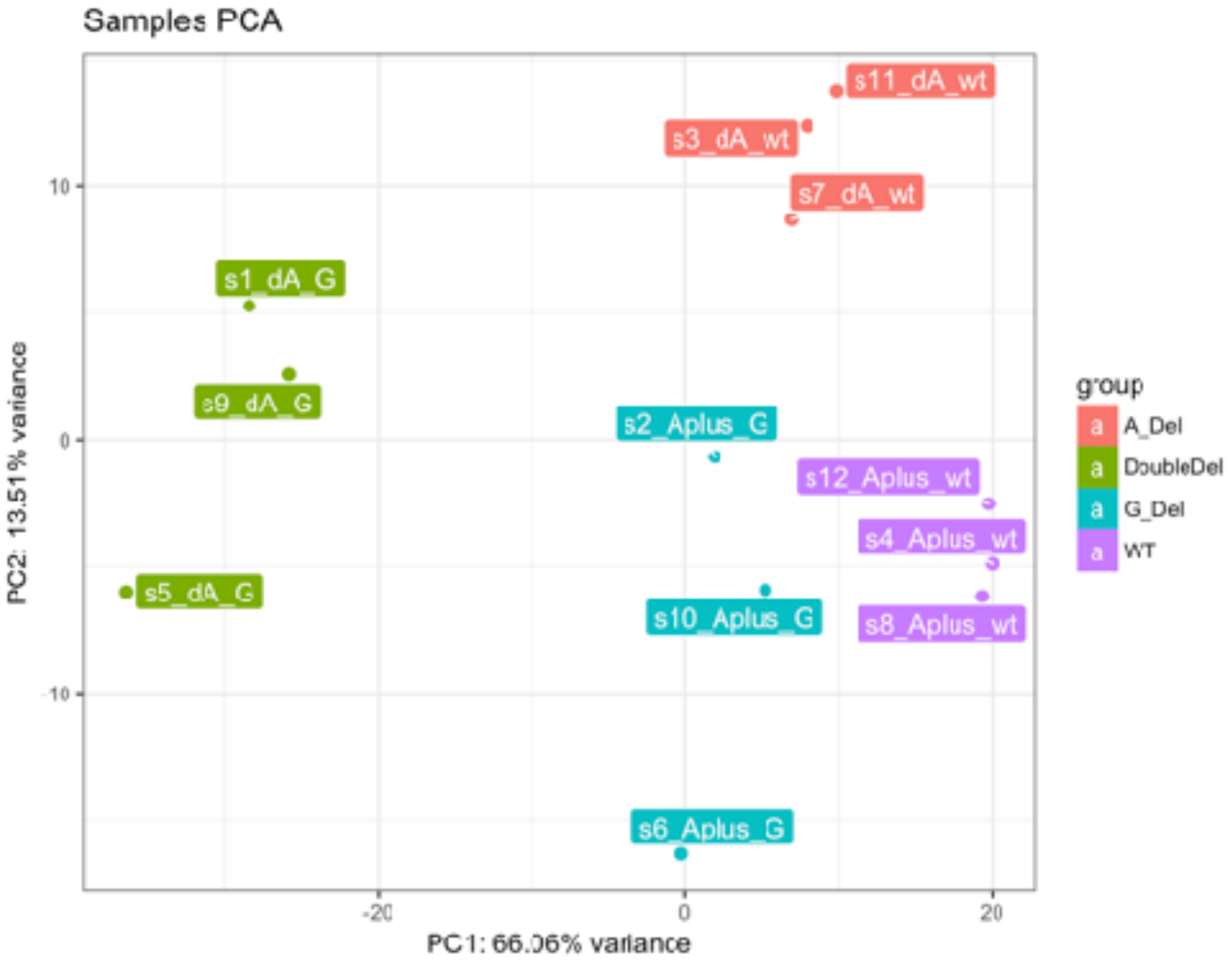
Differential Expression data **must** be corrected for multiple testing. Two common methods are the “Bonferroni procedure” and “Benjamini–Hochberg procedure”. These forms or statistical correction will result in a “corrected pvalue”, or a qvalue or FDR.

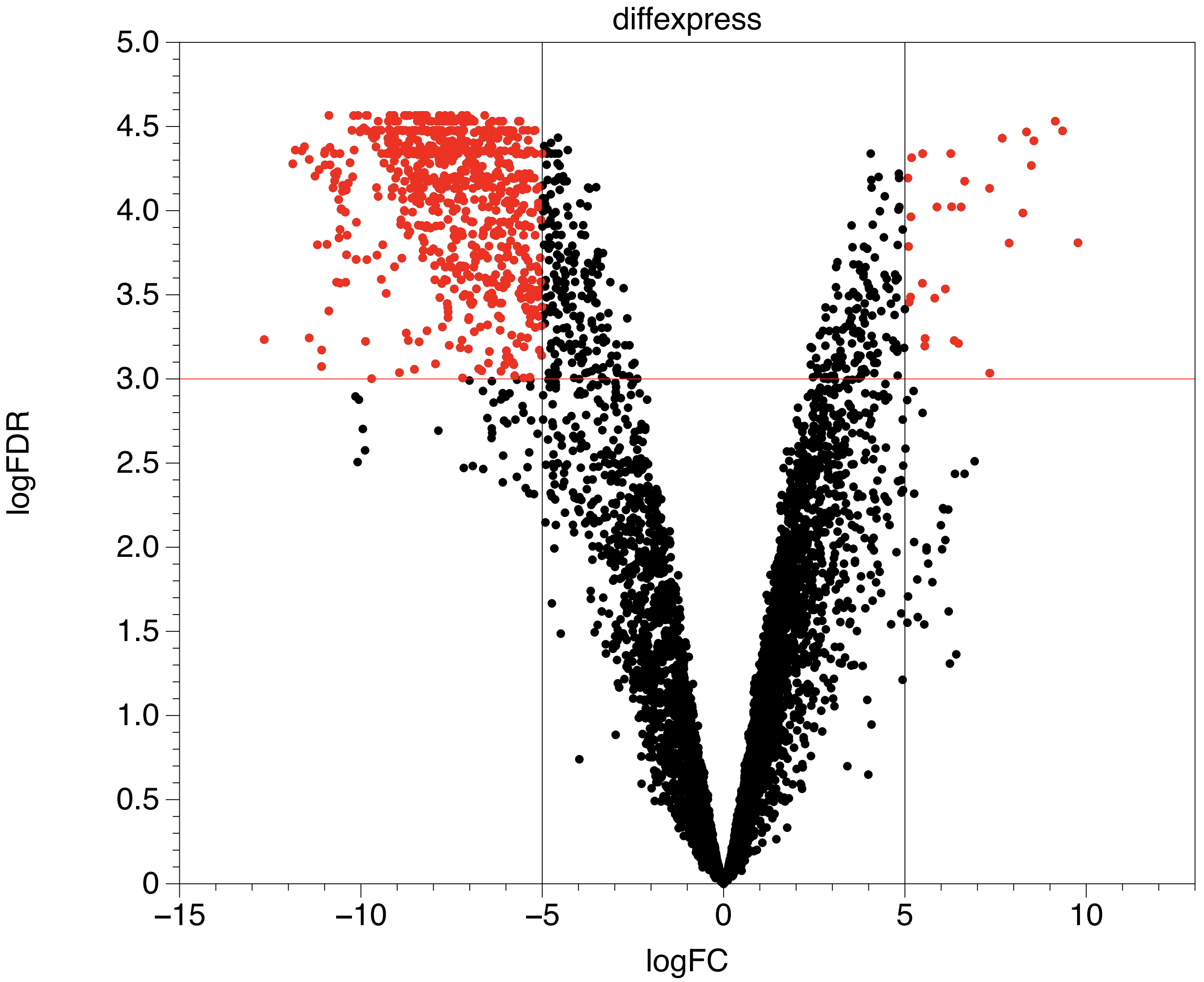
Note pvalues refer to the each gene, whereas an FDR (or qvalue) is a statement about a list. So using FDR cuff of 0.05 indicates that you can expect 5% false positives in the list of genes with an FDR of 0.05 or less.

Plotting the Data

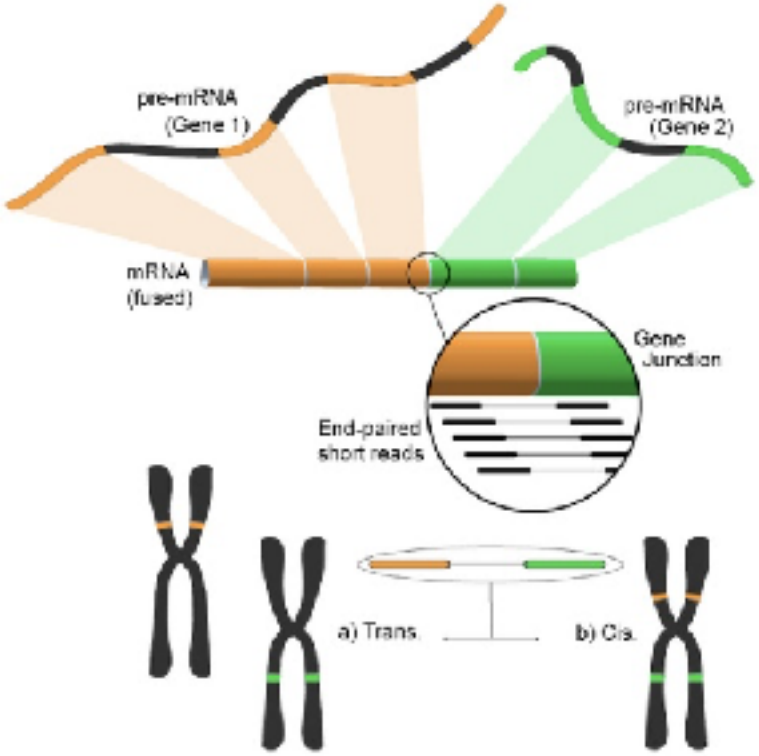


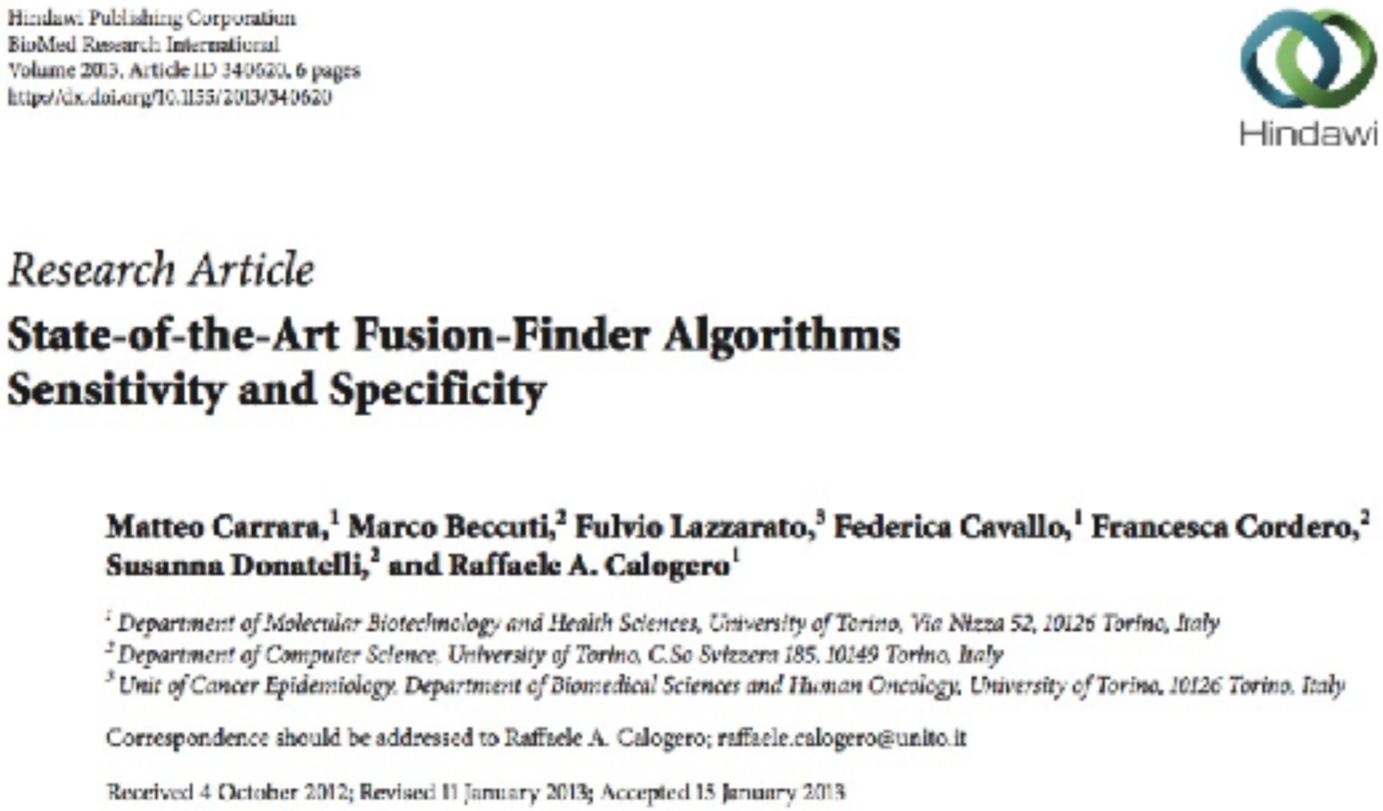






Fusion gene schematic




Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 340620, 6 pages
http://dx.doi.org/10.1155/2013/340620

Research Article

State-of-the-Art Fusion-Finder Algorithms Sensitivity and Specificity

Matteo Carrara,[1] Marco Beccuti,[2] Fulvio Lazzarato,[3] Federica Cavallo,[1] Francesca Cordero,[2] Susanna Donatelli,[2] and Raffaele A. Calogero[1]

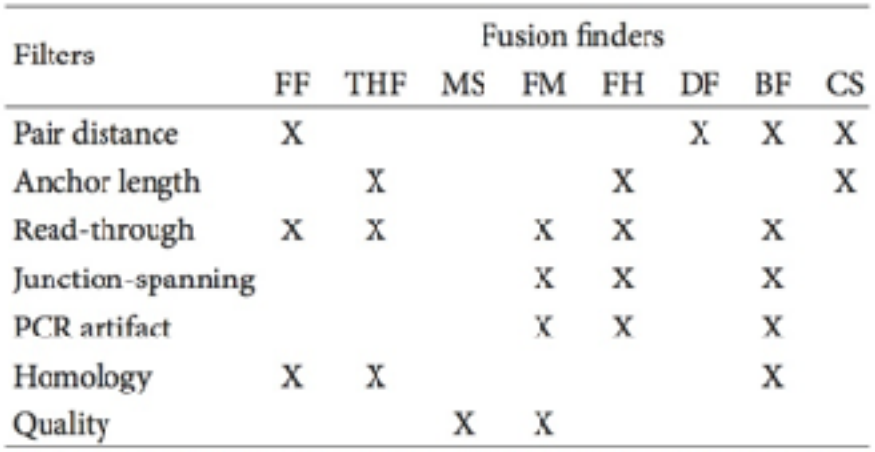
[1] *Department of Molecular Biotechnology and Health Sciences, University of Torino, Via Nizza 52, 10126 Torino, Italy*
[2] *Department of Computer Science, University of Torino, C.So Svizzera 185, 10149 Torino, Italy*
[3] *Unit of Cancer Epidemiology, Department of Biomedical Sciences and Human Oncology, University of Torino, 10126 Torino, Italy*

Correspondence should be addressed to Raffaele A. Calogero; raffaele.calogero@unito.it

Received 4 October 2012; Revised 11 January 2013; Accepted 15 January 2013

Fusion Detection

Table 1: Filtering steps embedded in the algorithms.



Filters	Fusion finders							
	FF	THF	MS	FM	FH	DF	BF	CS
Pair distance	X					X	X	X
Anchor length		X			X			X
Read-through	X	X		X	X		X	
Junction-spanning				X	X		X	
PCR artifact				X	X		X	
Homology	X	X					X	
Quality			X	X				

FF: FusionFinder; THF: TopHat-fusion; MS: MapSplice; FM: FusionMap; FH: FusionHunter; DF: deFuse; BF: Bellerophontes; CS: ChimeraScan.

False Positive Fusion Detection

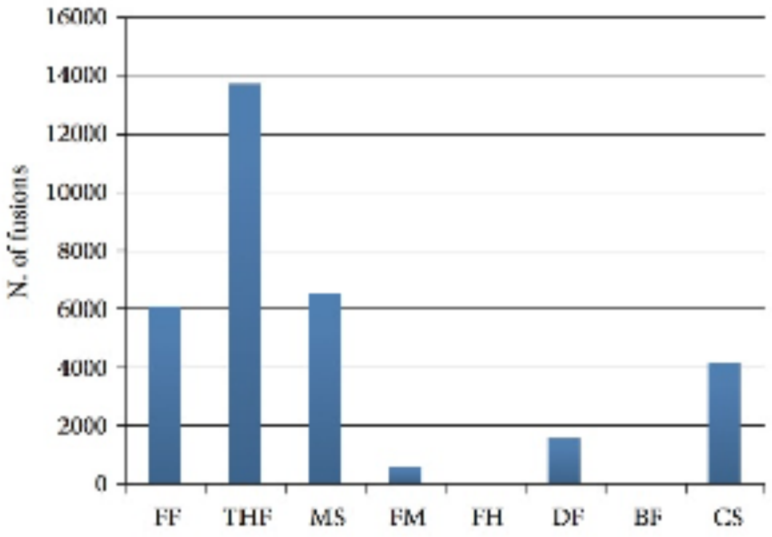
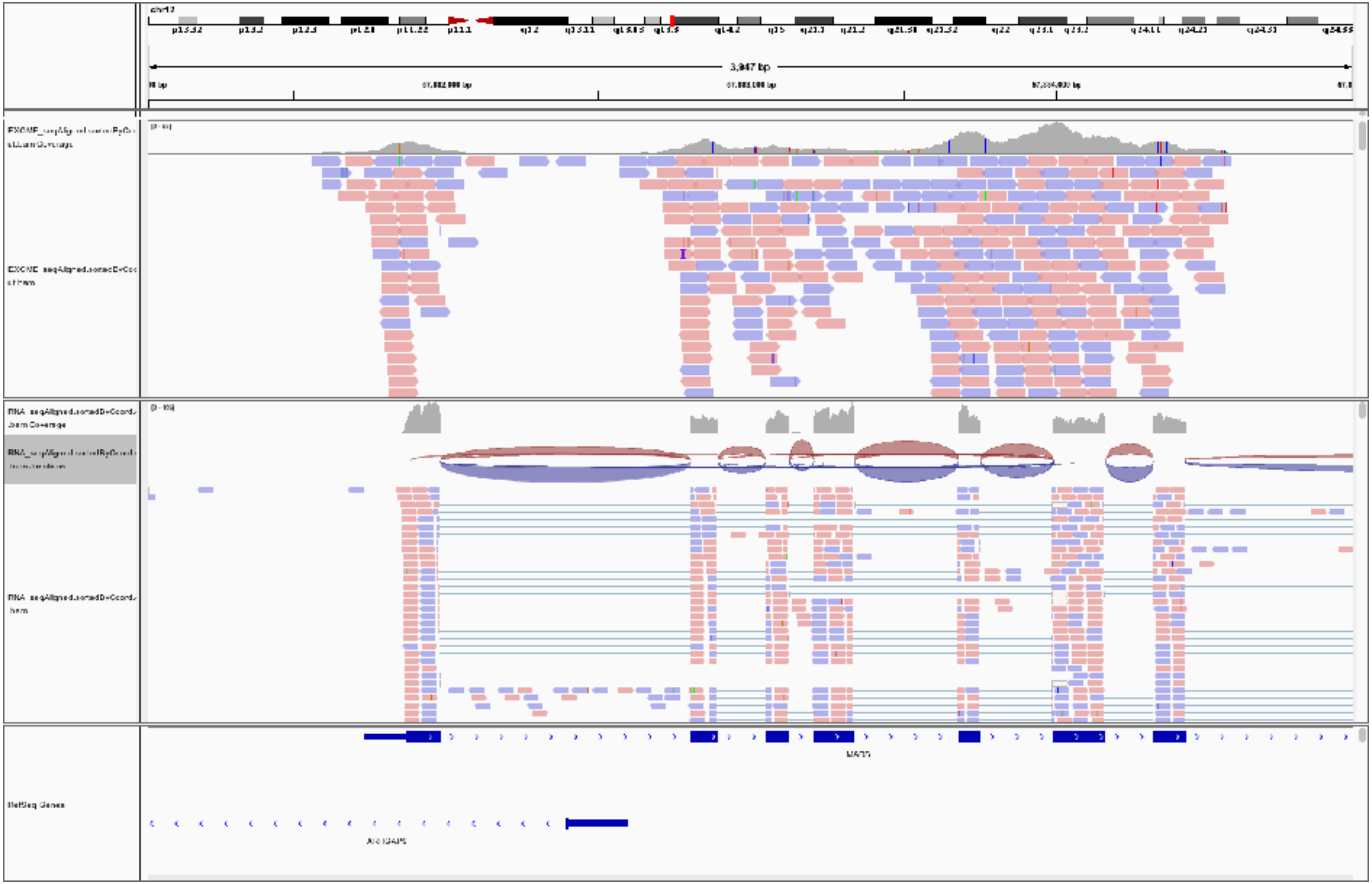
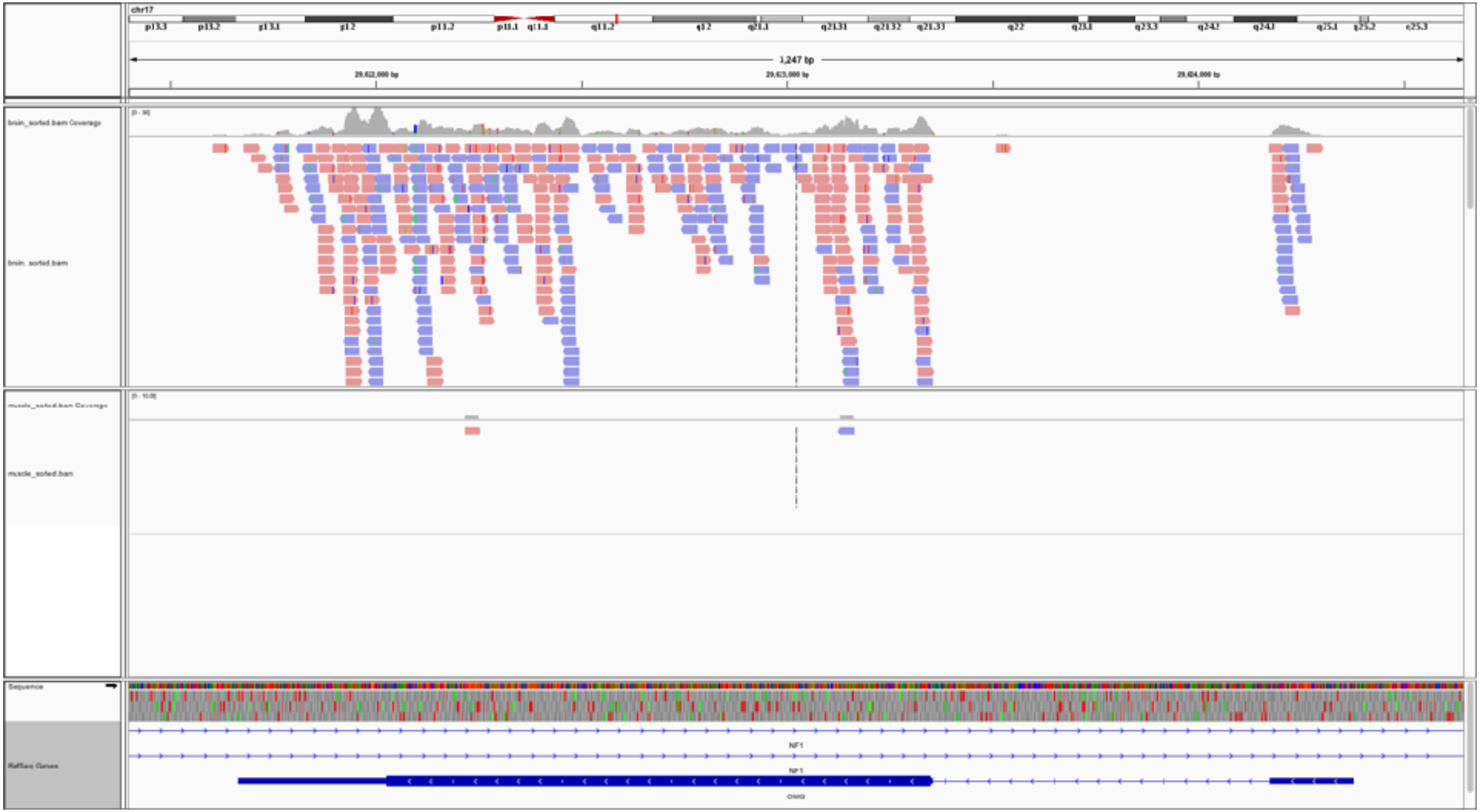
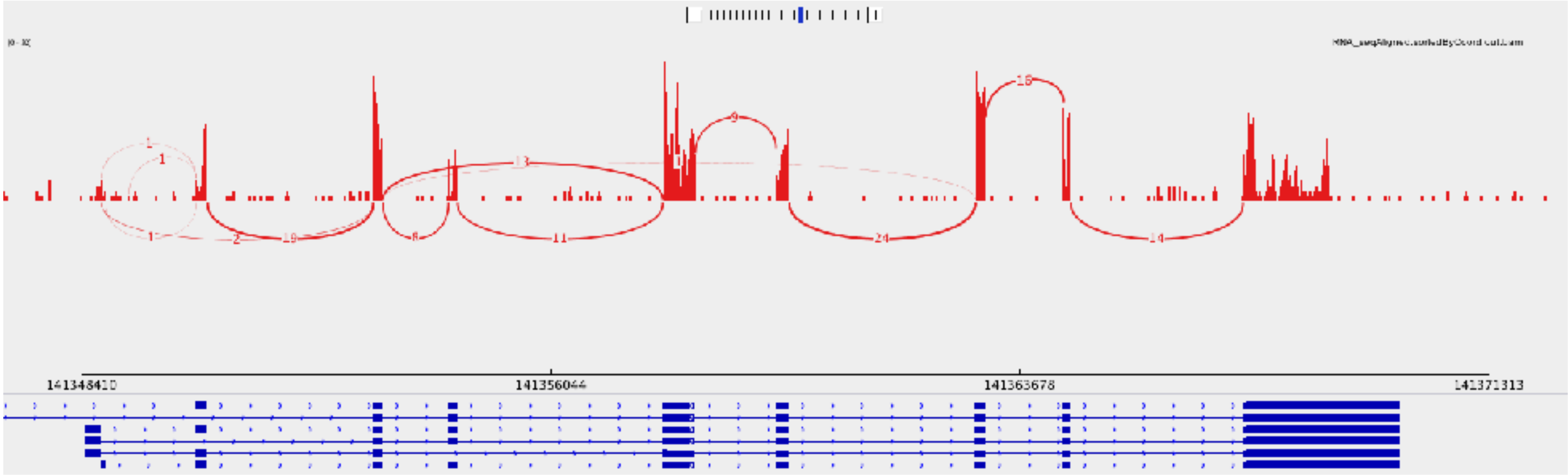


Figure 4: *False positive fusion detected using a synthetic dataset without chimeras.* FF: FusionFinder; THF: TopHat-fusion; MS: MapSplice; FM: FusionMap; FH: FusionHunter; DF: defuse; BF: Bellerophontes; CS: ChimeraScan.

Visualization

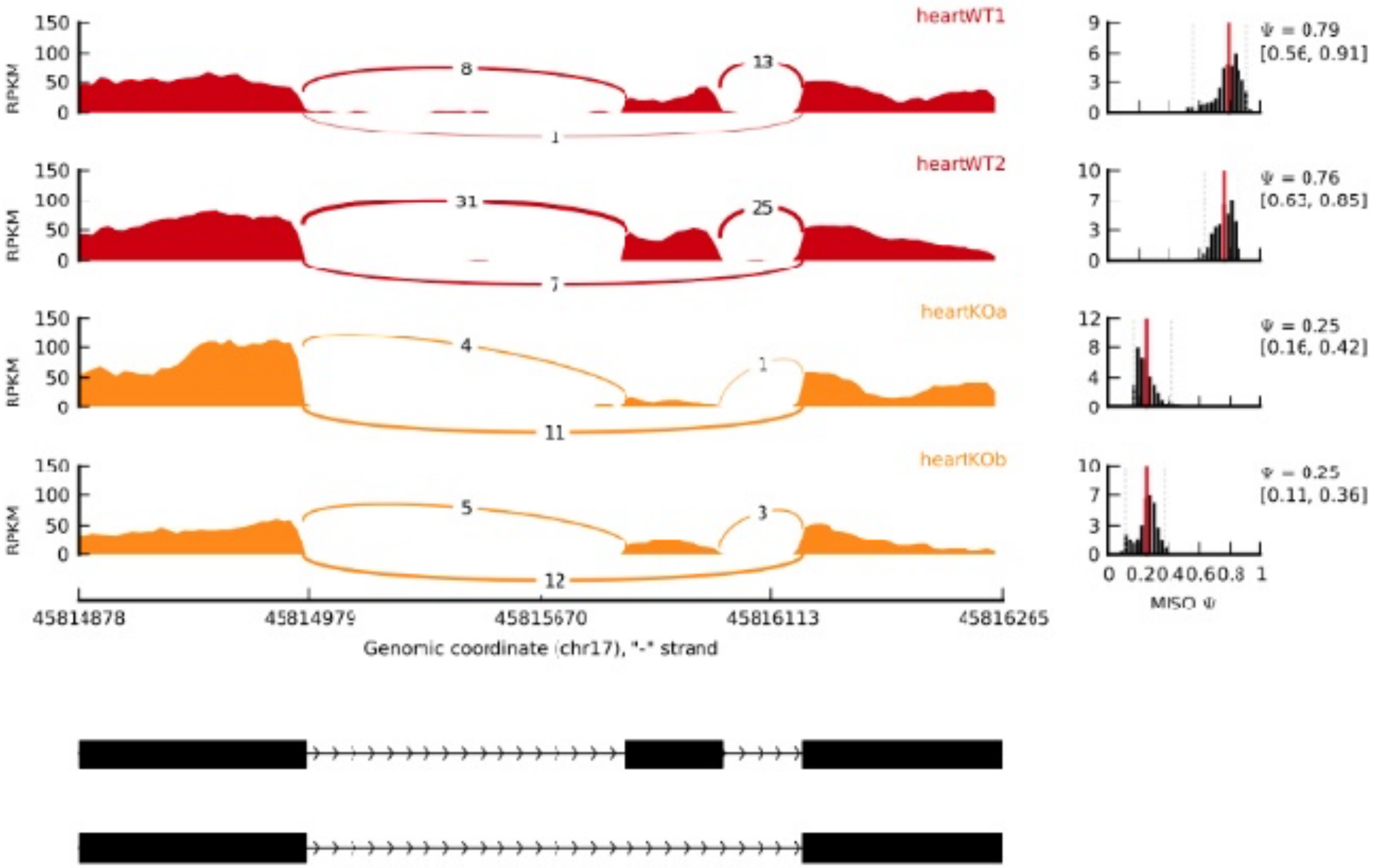

chr12
p13.32
p13.2
p12.3
p12.1
p11.22
p11.1
q12
q13.11
q13.13
q13.3
q14.2
q15
q21.1
q21.2
q21.31
q21.32
q22
q23.1
q23.3
q24.11
q24.23
q24.31
q24.33
3,947 bp
bp
57,882,000 bp
57,883,000 bp
57,884,000 bp
EXOME_seqAligned.sortedByCoord.out.bam Coverage
[0 - 10]
EXOME_seqAligned.sortedByCoord.out.bam
RNA_seqAligned.sortedByCoord.out.bam Coverage
[0 - 169]
RNA_seqAligned.sortedByCoord.out.bam Junctions
RNA_seqAligned.sortedByCoord.out.bam
MARS
RefSeq Genes
ARHGAP9


chr17
p13.3
p13.2
p13.1
p12
p11.2
p11.1
q11.1
q11.2
q12
q21.1
q21.31
q21.32
q21.33
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.2
q25.3
1,247 bp
29,622,000 bp
29,623,000 bp
29,624,000 bp
brain_sorted.bam Coverage
[0 - 30]
brain_sorted.bam
muscle_sorted.bam Coverage
[0 - 10.00]
muscle_sorted.bam
Sequence
RefSeq Genes
NF1
NF1
OMG


[0 - 32]
RNA_seqAligned.sortedByCoord.out.bam
1
1
1
2
19
8
13
11
9
16
24
14
141348410
141356044
141363678
141371313

Visualizing Splicing

chr17:45816186:45816265:-@chr17:45815912:45815950:-@chr17:45814875:45814965:-



Visualization and Next step tools

Visualization

1. Integrated Genome Viewer (https://www.broadinstitute.org/igv/)

Further Annotation of Genes

1. DAVID (http://david.abcc.ncifcrf.gov/tools.jsp)
2. ConsensusPathdb (http://cpdb.molgen.mpg.de/)
3. NetGestalt (http://www.netgestalt.org/)
4. Molecular Signatures Database (http://www.netgestalt.org/)
5. PANTHER (http://www.pantherdb.org/)
6. Cognoscente (http://vanburenlab.medicine.tamhsc.edu/cognoscente.shtml)
7. Pathway Commons (http://www.pathwaycommons.org/)
8. Readctome (http://www.reactome.org/)
9. PathVisio (http://www.pathvisio.org/)
10. Moksiskaan (http://csbi.ltdk.helsinki.fi/moksiskaan/)
11. Weighed Gene Co-Expression Network Analysis (WGCNA)s
12. More tools in R Bioconductor

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*)

Software Solutions

CCR staff have access to a number of resources

- Biowulf (Helix) - CIT maintained large cluster with a huge software library (Unix command line)
- CCBR Pipeliner (Biowulf)
- Partek Flow (Local Web Service)
- DNAnexus (Cloud Solution)
- CLCBio Genomic Workbench (Small genomes)

Public sources of RNA-Seq data

- **Gene Expression Omnibus** (GEO) (http://www.ncbi.nlm.nih.gov/geo/)
  - Both microarray and sequencing data
- **Sequence Read Archive (SRA)** (http://www.ncbi.nlm.nih.gov/sra)
  - All sequencing data (not necessarily RNA-Seq)
- **ArrayExpress** (https://www.ebi.ac.uk/arrayexpress/)
  - European version of GEO
- **Homogenized data**: MetaSRA, Toil, recount2, ARCHS^4

File Transfer

- Globus (https://hpc.nih.gov/storage/globus.html)
- Active Echo (https://activecho.cit.nih.gov/signin)
- (s)FTP
- Network Drives
- Flash Drives

Questions ?

Contacts:

Peter Fitzgerald	fitzgepe@nih.gov
Amy Stonelake	amy.stonelake@nih.gov
BTEP	ncibtep@nih.gov