Introduction to RNA sequencing and analysis

Monday, April 3, 2017

Mikhail Dozmorov, Ph.D.
mikhail.dozmorov@vcuhealth.org

https://github.com/mdozmorov/presentations

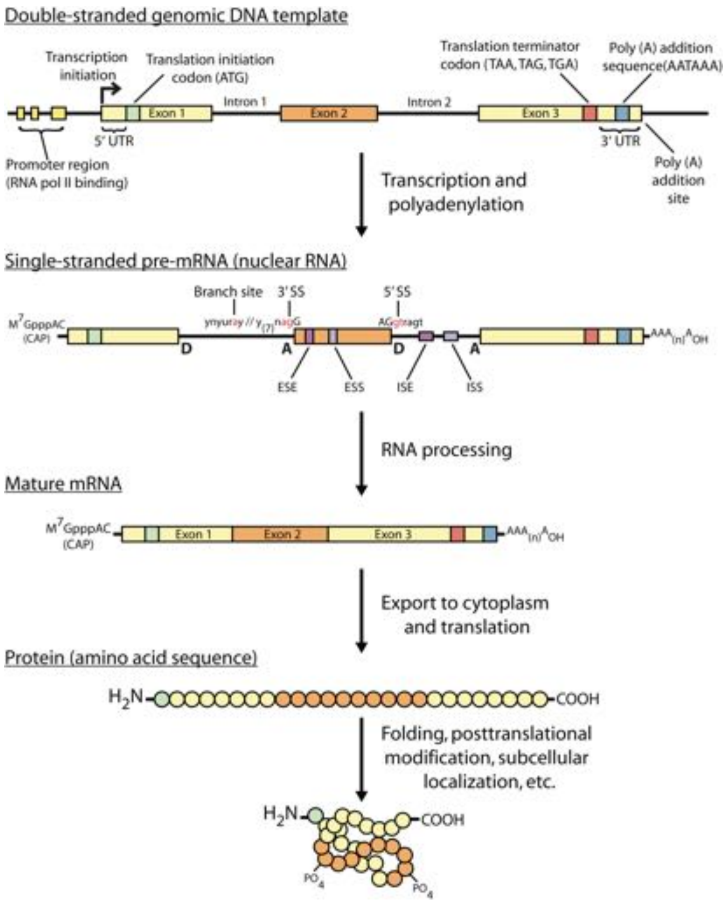
Objectives

Introduction to the theory and practice of RNA sequencing (RNA-seq) analysis

- Goals for RNA sequencing
- RNA-seq technology
- Experimental design
- RNA-seq analysis workflow
- Gene expression analysis
- Functional interpretation analysis
- Alternative splicing

Goals for RNA sequencing

Gene expression


Double-stranded genomic DNA template
Transcription initiation
Translation initiation codon (ATG)
Translation terminator codon (TAA, TAG, TGA)
Poly (A) addition sequence(AATAAA)
Intron 1
Exon 1
Exon 2
Intron 2
Exon 3
5' UTR
3' UTR
Promoter region (RNA pol II binding)
Poly (A) addition site
Transcription and polyadenylation
Single-stranded pre-mRNA (nuclear RNA)
Branch site
3'SS
5'SS
ynyuray // y(7)nagG
AGgtragt
M^7GpppAC (CAP)
D
A
D
A
AAA(n)A_{OH}
ESE
ESS
ISE
ISS
RNA processing
Mature mRNA
M^7GpppAC (CAP)
Exon 1
Exon 2
Exon 3
AAA(n)A_{OH}
Export to cytoplasm and translation
Protein (amino acid sequence)
H_2N
COOH
Folding, posttranslational modification, subcellular localization, etc.
H_2N
COOH
PO_4
PO_4

Source: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

Why RNA-sequencing?

- Gene Expression snapshot
- Allele-specific expression
- Transcriptome assembly
- Fusion detection
- Alternative splicing
- Detection of genomic variants

And many more, http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s003

Advantage of RNA-Seq over Microarray

- Much richer information beyond quantitation
  - Boundary of gene transcripts: both 5' and 3' end, to nucleotide level
  - Alternative exon usage, novel splicing junction detection
  - SNP/indel discovery in transcripts: both coding and UTRs
  - Allele specific expression: critical in imprinting, cancer

- Not relying on gene annotation by mapping to the whole genome
  - No longer biased by probe design
  - Novel gene and exon discovery enabled

Advantage of RNA-Seq over Microarray

- Better performance at quantitation
    - Unlimited dynamic range: by increasing depth as needed
    - Higher specificity and accuracy: digital counts of transcript copies, very low background noise
    - Higher sensitivity: more transcripts and more differential genes detected
- Re-analysis easily done by computation, as gene annotation keeps evolving
- *De novo* assembly possible, not relying on reference genome sequence
- Comparable cost, continuing to drop

RNA-seq technology

Sequencing technologies

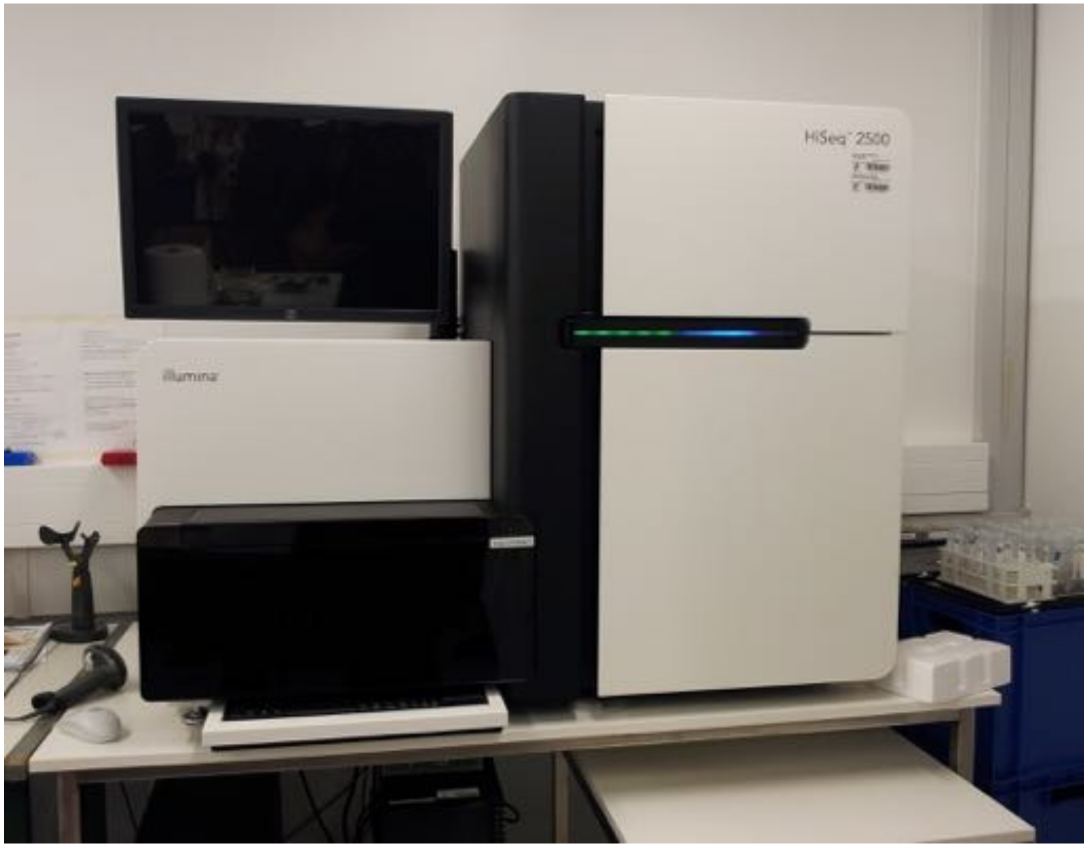
Commercially available

- Illumina/Solexa - short reads, sequencing-by-synthesis
- Life Technologies Ion Torrent/Proton - short reads, Ion Semiconductor sequencing
- Pacific Biosciences - long reads, Single Molecule Real Time sequencing

Experimental

- Nanopore sequencing - continuous sequencing (very long reads), fluctuations of the ionic current from nucleotides passing through the nanopore

Sequencing technologies



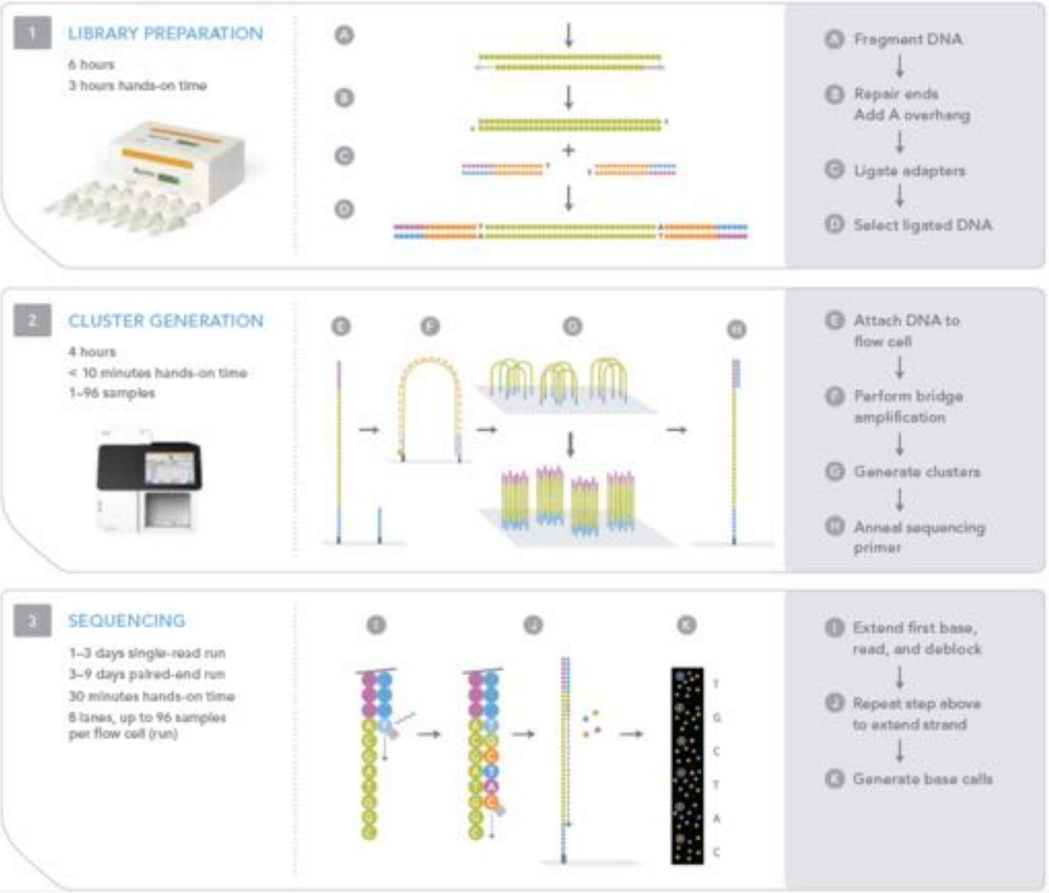
- Specifications depend on the library preparation kits, Single or Dual flow cells, High-Output or Rapid-Run modes

Sequencing technologies

- Example:
    - 100 bp paired-end reads
    - Dual Flow cell run-time
    - High-Output Run Mode
        - Up to 4 billion reads
        - 5 days run time
    - Rapid-Run Mode
        - Up to 600 million reads
        - 27 hours

Source: https://www.illumina.com/documents/products/datasheets/datasheet_hiseq2500.pdf

Illumina sequencing workflow


1 LIBRARY PREPARATION

6 hours
3 hours hands-on time

A Fragment DNA
↓
B Repair ends
Add A overhang
↓
C Ligate adapters
↓
D Select ligated DNA

2 CLUSTER GENERATION

4 hours
< 10 minutes hands-on time
1–96 samples

E Attach DNA to flow cell
↓
F Perform bridge amplification
↓
G Generate clusters
↓
H Anneal sequencing primer

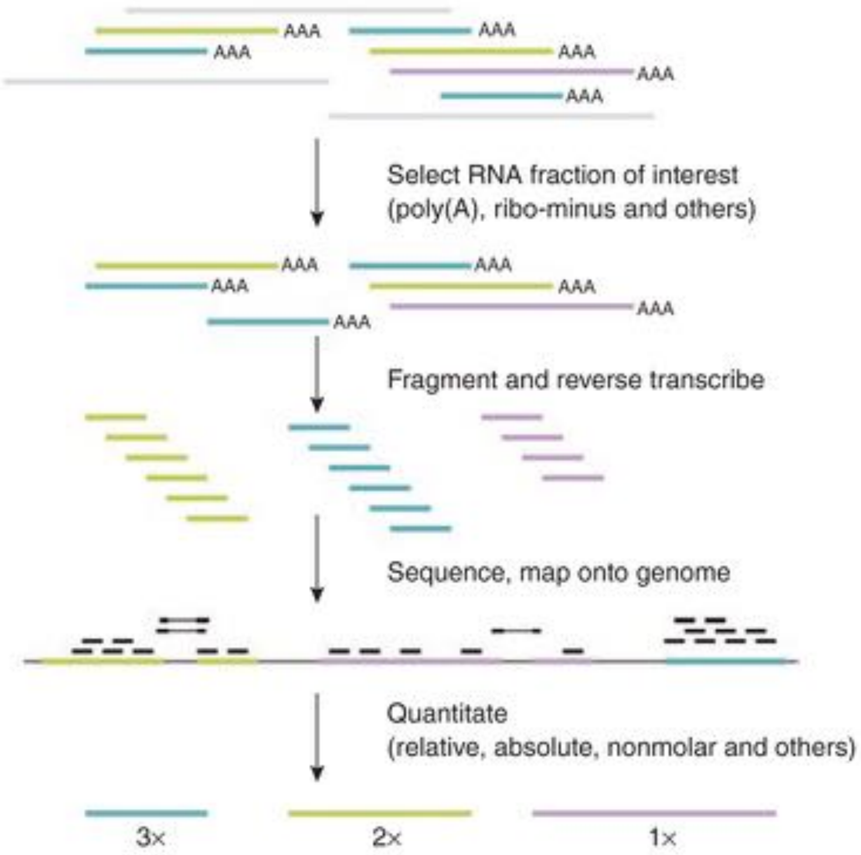
3 SEQUENCING

1–3 days single-read run
3–9 days paired-end run
30 minutes hands-on time
8 lanes, up to 96 samples per flow cell (run)

I Extend first base, read, and deblock
↓
J Repeat step above to extend strand
↓
K Generate base calls

Video: https://www.youtube.com/watch?v=77r5p8IBwJk

Overview of RNA sequencing technology


AAA
Select RNA fraction of interest
(poly(A), ribo-minus and others)
Fragment and reverse transcribe
Sequence, map onto genome
Quantitate
(relative, absolute, nonmolar and others)
3×
2×
1×

Source: http://www.nature.com/nmeth/journal/v6/n11s/fig_tab/nmeth.1371_F5.html

RNA-Seq Limitations

- Quantitation influenced by many confounding factors
  - "Sequenceability" - varying across genomic regions, local GC content and structure related
  - Varying length of gene transcripts and exons
  - Bias in read ends due to reverse transcription, subtle but consistent
  - Varying extent of PCR amplification artifacts
  - Effect of RNA degradation in the real world
  - Computational bias in aligning reads to genome due to aligners

RNA-Seq Limitations

- SNP discovery in RNA-seq is more challenging than in DNA
    - Varying levels of coverage depth
    - False discovery around splicing junctions due to incorrect mapping
- *De novo* assembly of transcripts without genome sequence: computationally intensive but possible, technical improvements will help
    - longer read length
    - lower error rate
    - more uniform nucleotide coverage of transcripts - more equalized transcript abundance

Library preparation

- **RNA isolation**
  - $0.1 - 1\mu g$ original total RNA
- **Ribosomal RNA (rRNA) depletion**
  - rRNAs constitute over 90 % of total RNA in the cell, leaving the 1–2 % comprising messenger RNA (mRNA) that we are normally interested in.
  - Enriches for mRNA + long noncoding RNA.
  - Hybridization to bead-bound rRNA probes

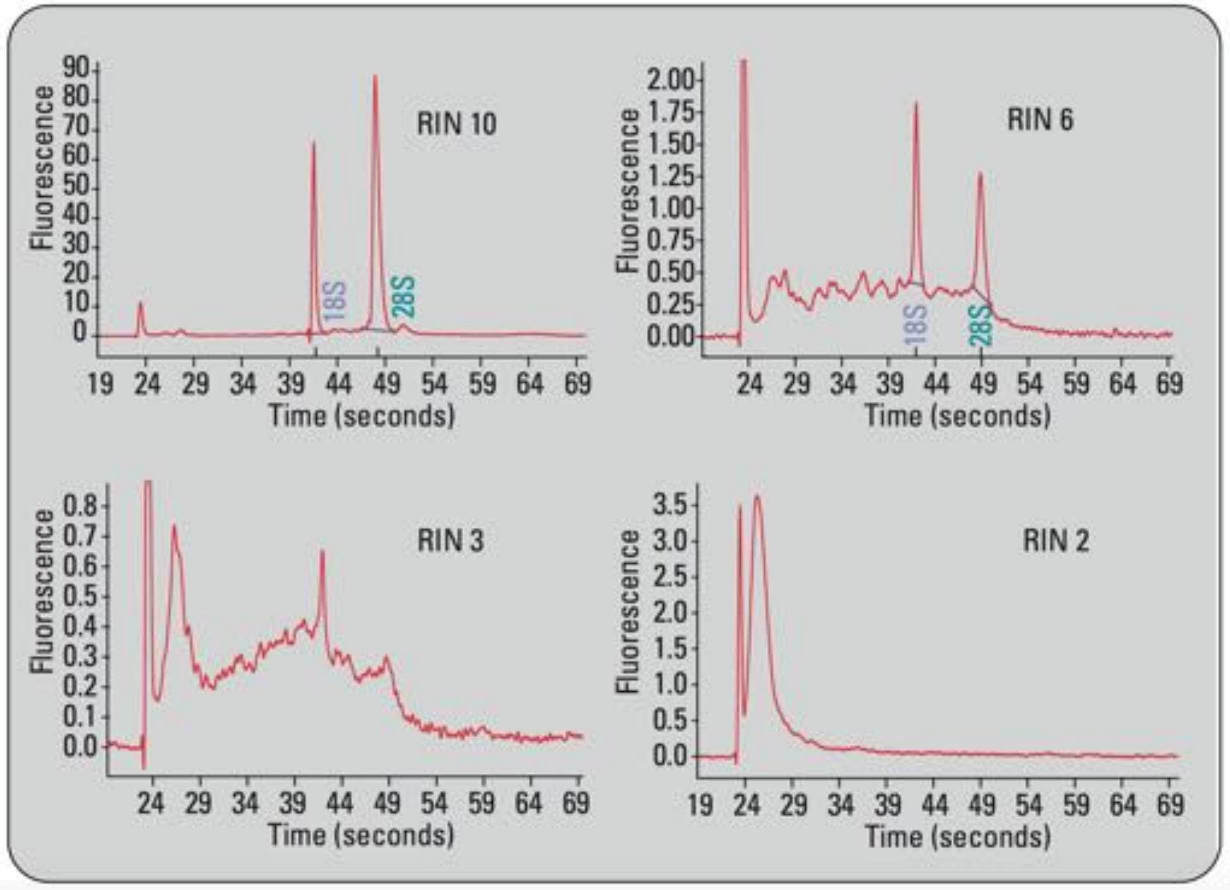
Library preparation

- **Poly(A) selection (for eukaryotes only)**
  - Enrich for mRNA.
  - Hybridization to oligo-dT beads
- **Small RNA extraction**
  - Specific kits required to retain small RNAs
  - Optionally, size-selection by gel

More at http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s006

RNA quality

Agilent 2100 bioanalyzer. RIN - RNA integrity number (should be >7)



18/90

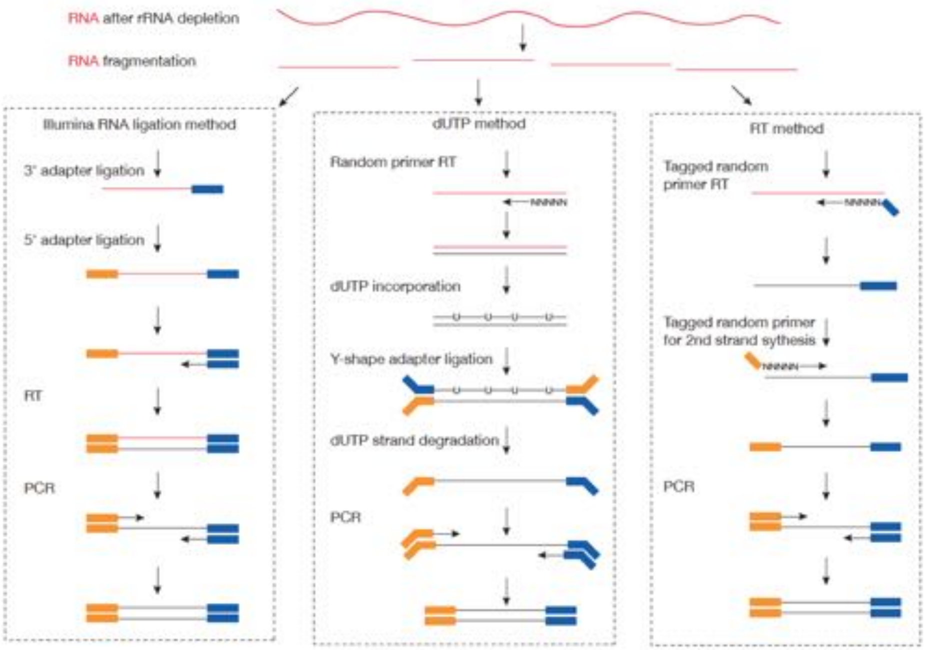
Library preparation steps

- **Fragmentation**, to recover short reads across full length of long genes
- **Size selection**, suitable for RNA sequencing. 300-500bp - mRNA, 20-150bp - small/miRNA
- **Amplification**, typically by PCR. Up to $0.5 - 10ng$ of RNA
- **Library normalization/Exome capture**

Unstranded vs. Strand-specific library

Unstranded: Random hexamer priming to reverse-transcribe mRNA

Stranded: dUTP method - incorporating UTP nucleotides during the second cDNA synthesis, followed by digestion of the strand containing dUTP



More at http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s007

Library preparation steps

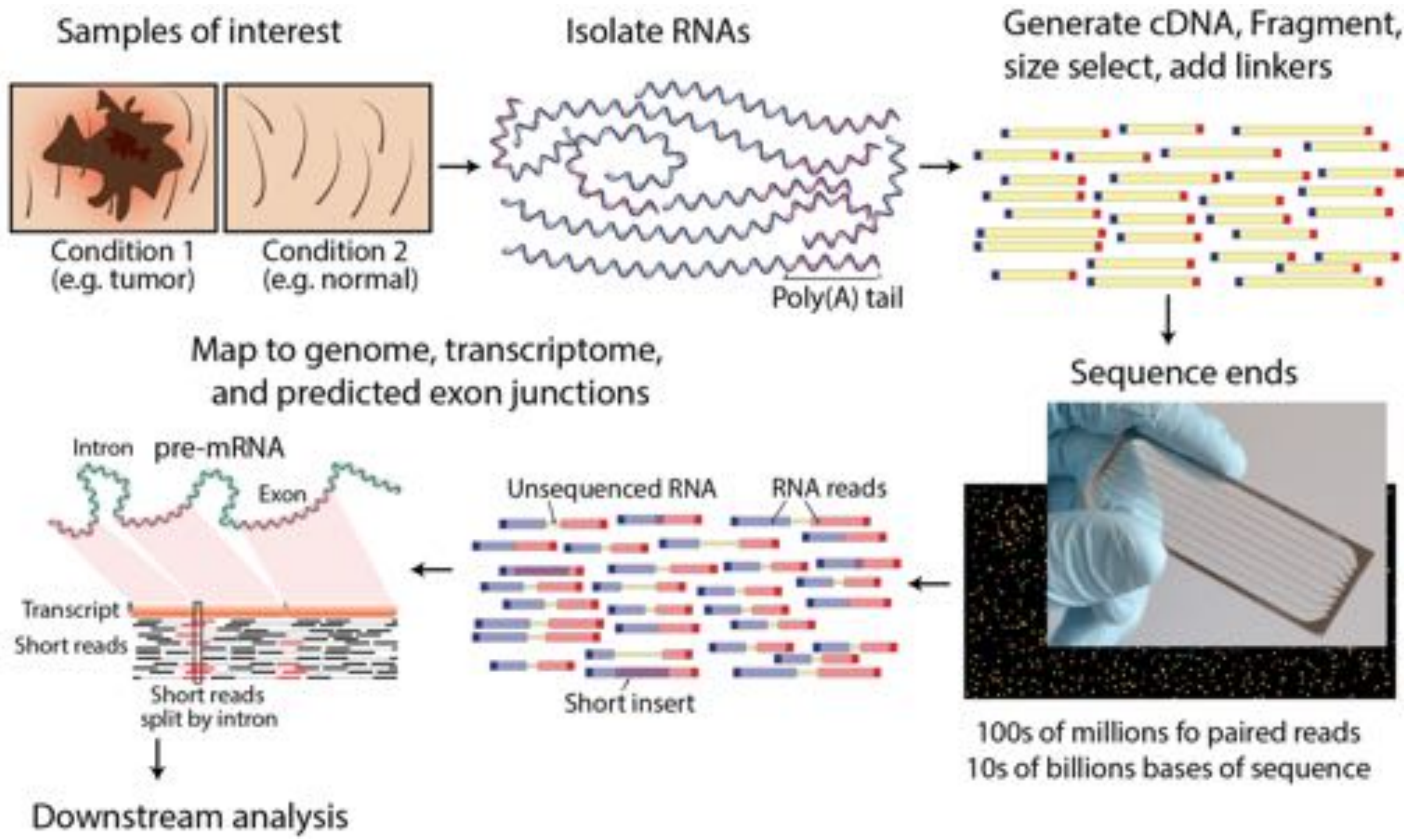
- **Barcoding** and **multiplexing**
- Optionally, add **spike-in controls**
- **Single** or **paired end** sequencing. The latter is preferrable for *de novo* transcript discovery or isoform expression analysis

More at http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s005

Sequencing length/depth

- Longer reads improve mappability and transcript quantification
- More transcripts will be detected and their quantification will be more precise as the sample is sequenced to a deeper level
- Up to 100 million reads is needed to precisely quantify low expressed transcripts. In reality, 20-30 million reads is OK for human genome.

Overview of RNA sequencing technology


Samples of interest
Condition 1
(e.g. tumor)
Condition 2
(e.g. normal)
Isolate RNAs
Poly(A) tail
Generate cDNA, Fragment, size select, add linkers
Sequence ends
100s of millions fo paired reads
10s of billions bases of sequence
Unsequenced RNA
RNA reads
Short insert
Map to genome, transcriptome, and predicted exon junctions
Intron
pre-mRNA
Exon
Transcript
Short reads
Short reads split by intron
Downstream analysis

Source: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

Experimental design

Experimental design

- **Replication**. It allows the experimenter to obtain an estimate of the experimental error
- **Randomization**. It requires the experimenter to use a random choice of every factor that is not of interest but might influence the outcome of the experiment. Such factors are called nuisance factors
- **Blocking**. Creating homogeneous blocks of data in which a nuisance factor is kept constant while the factor of interest is allowed to vary. Used to increase the accuracy with which the influence of the various factors is assessed in a given experiment
- **Block what you can, randomize what you cannot**

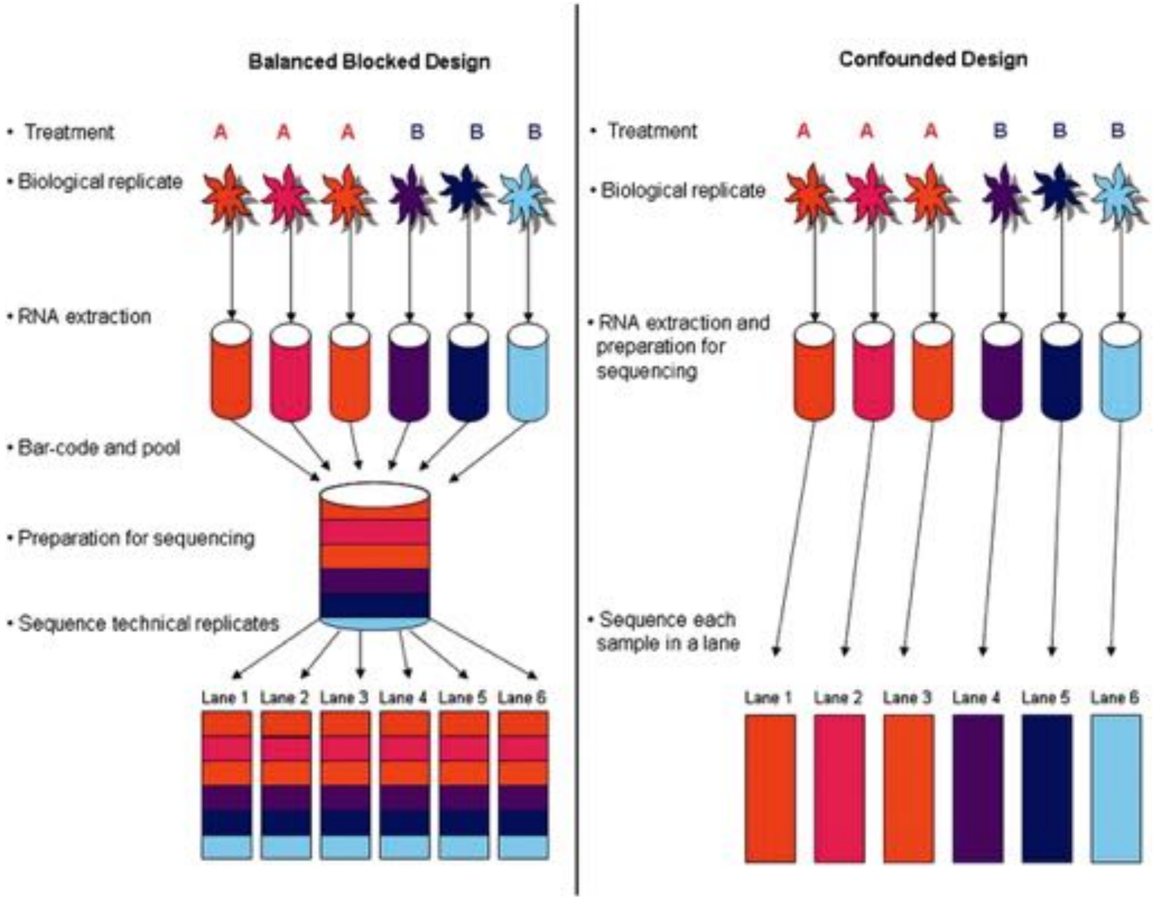
Experimental design

In RNA-seq, we have multiple levels of randomness:

- Biological variability in samples
- Stochasticity of RNA content
- Randomness of fragments being sequenced
- Technical variability

Auer, P.,RW Doerge. "Statistical Design and Analysis of RNA Sequencing Data." Genetics, 2010 http://www.genetics.org/content/185/2/405.long

Experimental design: Multiplexing balances technical variability


Balanced Blocked Design
• Treatment
A A A B B B
• Biological replicate
• RNA extraction
• Bar-code and pool
• Preparation for sequencing
• Sequence technical replicates
Lane 1 Lane 2 Lane 3 Lane 4 Lane 5 Lane 6
Confounded Design
• Treatment
A A A B B B
• Biological replicate
• RNA extraction and preparation for sequencing
• Sequence each sample in a lane
Lane 1 Lane 2 Lane 3 Lane 4 Lane 5 Lane 6

Number of replicates

Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

Replicates per group	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %

Source: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Number of replicates

Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

Replicates per group	3	5	10
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

Source: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Power calculations

- **Scotty**
  - Power Analysis for RNA Seq Experiments, http://scotty.genetics.utah.edu/
- **powerSampleSizeCalculator**
  - R scripts for power analysis and sample size estimation for RNA-Seq differential expression, http://www2.hawaii.edu/~lgarmire/RNASeqPowerCalculator.htm
- **RnaSeqSampleSize**
  - R package and a Shiny app for RNA sequencing data sample size estimation, https://cqs.mc.vanderbilt.edu/shiny/RNAseqPS/

Busby MA, Stewart C, Miller CA, Grzeda KR, Marth GT. "Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression". *Bioinformatics* 2013 https://www.ncbi.nlm.nih.gov/pubmed/23314327

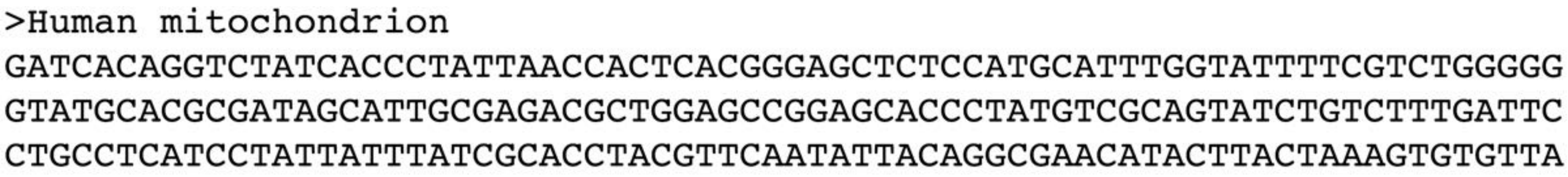
Travers C. et.al. "Power analysis and sample size estimation for RNA-Seq differential expression" *RNA* 2014 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4201821/

Guo et.al. "RNAseqPS: A Web Tool for Estimating Sample Size and Power for RNAseq Experiment" *Cancer Informatics* 2014 http://insights.sagepub.com/rnaseqps-a-web-tool-for-estimating-sample-size-and-power-for-rnaseq-ex-article-a4433

RNA-seq analysis workflow

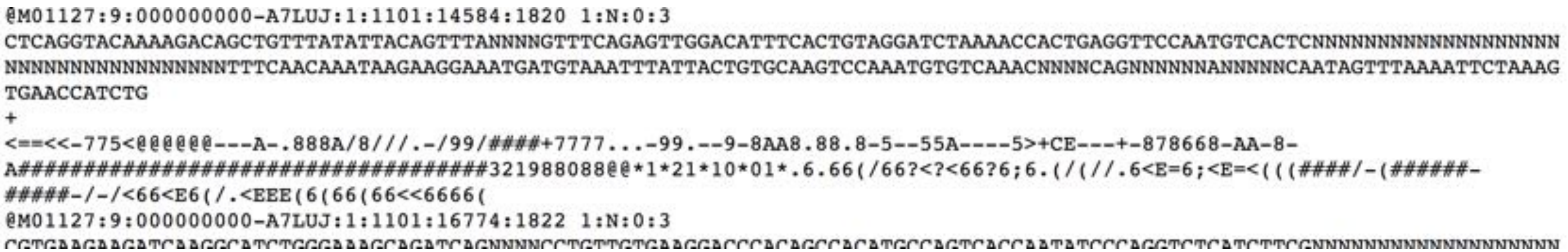
FASTA/FASTQ format

FASTA: text-based representation of nucleotide sequence.
http://zhanglab.ccmb.med.umich.edu/FASTA/



```
>Human mitochondrion
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGG
GTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTC
CTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTA
```

FASTQ: sequence and quality info



```
@M01127:9:000000000-A7LUJ:1:1101:14584:1820 1:N:0:3
CTCAGGTACAAAAGACAGCTGTTTATATTACAGTTTANNNNGTTTCAGAGTTGGACATTTCACTGTAGGATCTAAAACCACTGAGGTTCCAATGTCACTCNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNTTTCAACAAATAAGAAGGAAATGATGTAAATTTATTACTGTGCAAGTCCAAATGTGTCAAACNNNNCAGNNNNNNANNNNNCAATAGTTTAAAATTCTAAAG
TGAACCATCTG
+
<==<<-775<@@@@@@---A-.888A/8///.-/99/####+7777...-99.--9-8AA8.88.8-5--55A----5>+CE---+-878668-AA-8-
A####################################3219880888@@*1*21*10*01*.6.66(/66?<?<66?6;6.(/(//.6<E=6;<E=<(((####/-(######-
#####-/-/<66<E6(/.<EEE(6(66(66<<6666(
@M01127:9:000000000-A7LUJ:1:1101:16774:1822 1:N:0:3
CGTGAAGAAGATCAAGGCATCTGGGAAAGCAGATCAGNNNNNCCTGTTGTGAAGGACCCACAGCCACATGCCAGTCACCAATATCCCAGGTCTCATCTTCGNNNNNNNNNNNNNNNNNNNN
```

Quality control

- **FASTQC**
  - Quality of raw sequencing data, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
  - Video tutorial how to interpret, https://www.youtube.com/watch?v=bz93ReOv87Y

- **RNASeQC**
  - quality of mapped (aligned) data, http://rseqc.sourceforge.net/

Quality of base calling

- **Phred quality score** is widely used to characterize the quality of base calling
- Phred quality score = $-10 * log_{10}(P)$, where 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

Adapter trimming

- **FASTX-Toolkit**: set of tools for low-level sequence trimming/cutting, http://hannonlab.cshl.edu/fastx_toolkit/
- **Trimmomatic**: well-documented and easy-to-use adapter trimmer using multiple algorithms. Handles single- and paired-end reads. http://www.usadellab.org/cms/?page=trimmomatic
- **Flexbar**: similar to Trimmomatic by functionality. https://github.com/seqan/flexbar/wiki/Manual

Alignment

- RNA-seq aligners face an additional problem, not encountered in DNA-only alignment: many RNA-seq reads will span introns
- The average human intron length is >6,000 bp (some are >1 Mbp in length)
- In a typical human RNA-seq experiment using 100-bp reads, >35% of the reads will span multiple exons - align over splice junctions
- Aligners must be splice-aware, especially when aligning longer (>50bp) reads

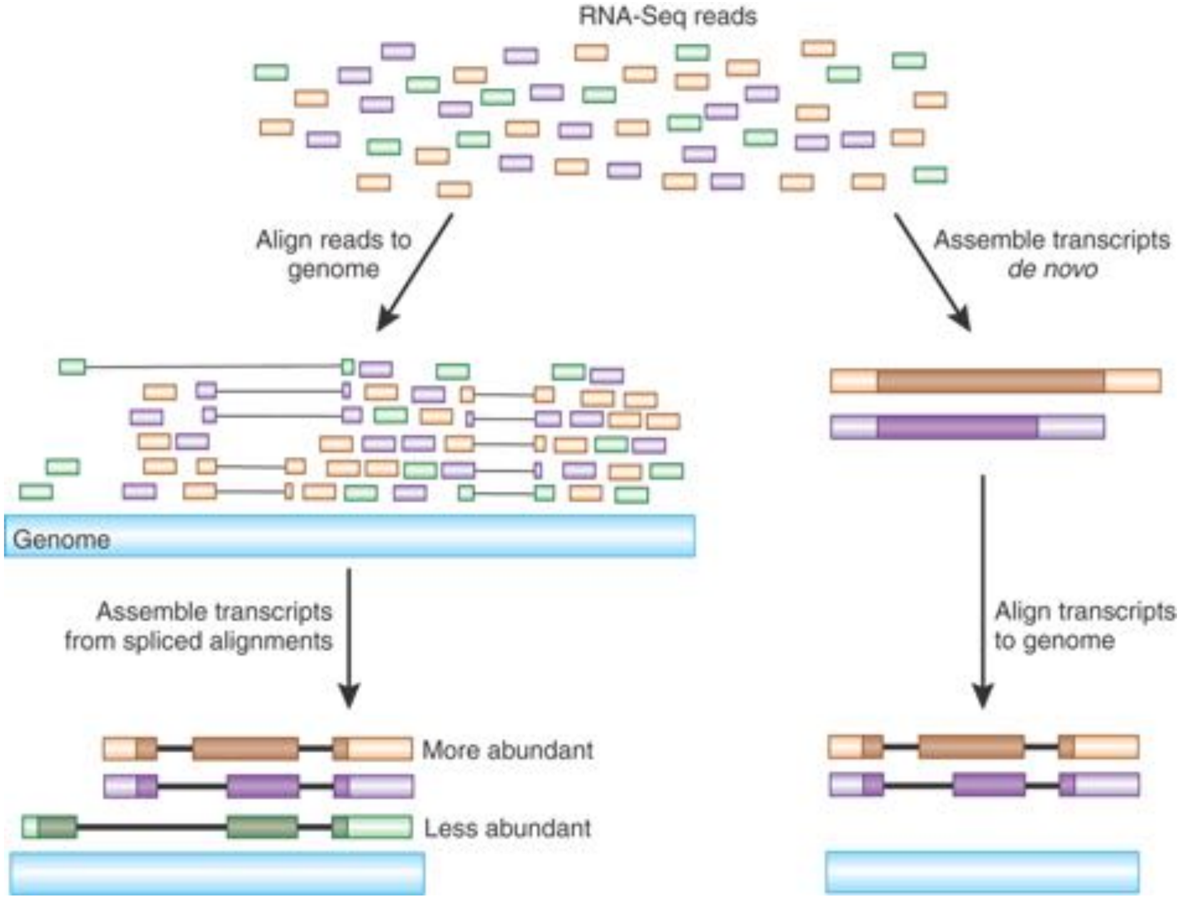
Duplicates removal

- Duplicates may correspond to biased PCR amplification of particular fragments
- For highly expressed, short genes, duplicates are expected even if there is no amplification bias
- Removing them may reduce the dynamic range of expression estimates

Generally, do not remove duplicates from RNA-seq data

- If you ultimately want to remove duplicates, use Picard tools' `MarkDuplicates` command, https://broadinstitute.github.io/picard/command-line-overview.html#MarkDuplicates

Alignment strategies


RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Align to the reference genome is the most common for transcript quantification

Alignment - Mapping RNA-seq reads to the genome

- **BWA**: general purpose algorithms tuned for different tasks, http://bio-bwa.sourceforge.net/
- **STAR**: fast and accurate aligner, https://github.com/alexdobin/STAR
- **HISAT**: (hierarchical indexing for spliced alignment of transcripts) uses two types of indexes for alignment: a global, whole-genome index and tens of thousands of small local indexes. Can detect novel splice sites, transcription initiation and termination sites. A part of the new "Tuxedo suite", including StringTie and Ballgown, http://ccb.jhu.edu/software/hisat2/index.shtml.
- **subread**: a fast and accurate aligner, R and command line. The whole package includes `subjunc` for junction detection, and `featureCounts` for extracting read counts per gene from aligned SAM/BAM files, http://subread.sourceforge.net/

Timeline and extensive comparison of aligners: https://www.ebi.ac.uk/~nf/hts_mappers/

SAM format of aligned data

– SAM stands for Sequence Alignment/Map format. The SAM format consists of two sections:

Header section

- Used to describe source of data, reference sequence, method of alignment, etc.

Alignment section

- Used to describe the read, quality of the read, and nature alignment of the read to a region of the genome

SAM format specification https://samtools.github.io/hts-specs/SAMv1.pdf

BAM file format of aligned data

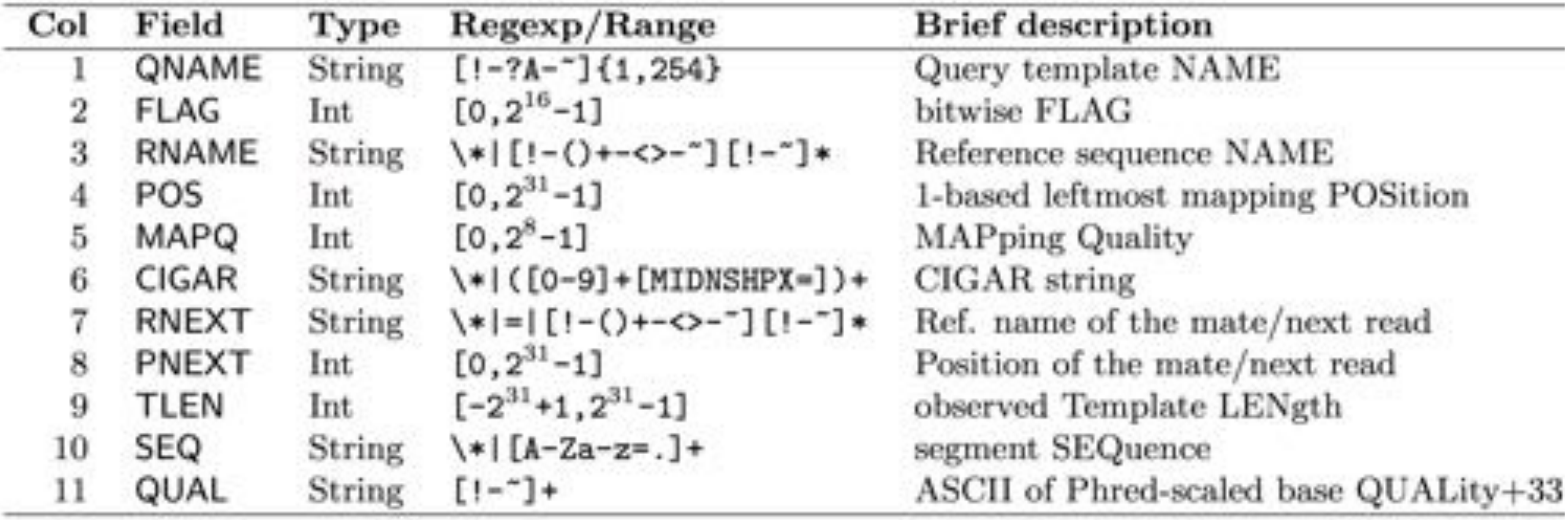
- BAM is the binary version of a SAM file. Smaller, but not easily readable.
- Compressed using lossless BGZF format
- Other BAM compression strategies are a subject of research. See ‘CRAM’ format for example, http://www.internationalgenome.org/faq/what-are-cram-files/
- BAM files are usually indexed. An index is stored alongside the BAM file with a ".bai" extension
- Indexing aims to achieve fast retrieval of alignments overlapping a specified region without going through the whole alignments.
- BAM must be sorted before indexing. Depending on the downstream tools, sort by
  - Name
  - Coordinate

SAM/BAM header section

- Used to describe source of data, reference sequence, method of alignment, etc.
- Each section begins with character ‘@’ followed by a two-letter record type code. These are followed by two-letter tags and values

```
@HD The header line
VN: format version
SO: Sorting order of alignments
@SQ Reference sequence dictionary
SN: reference sequence name
LN: reference sequence length
SP: species
@RG Read group
ID: read group identifier
CN: name of sequencing center
SM: sample name
@PG Program
PN: program name
VN: program version
```

SAM/BAM alignment section

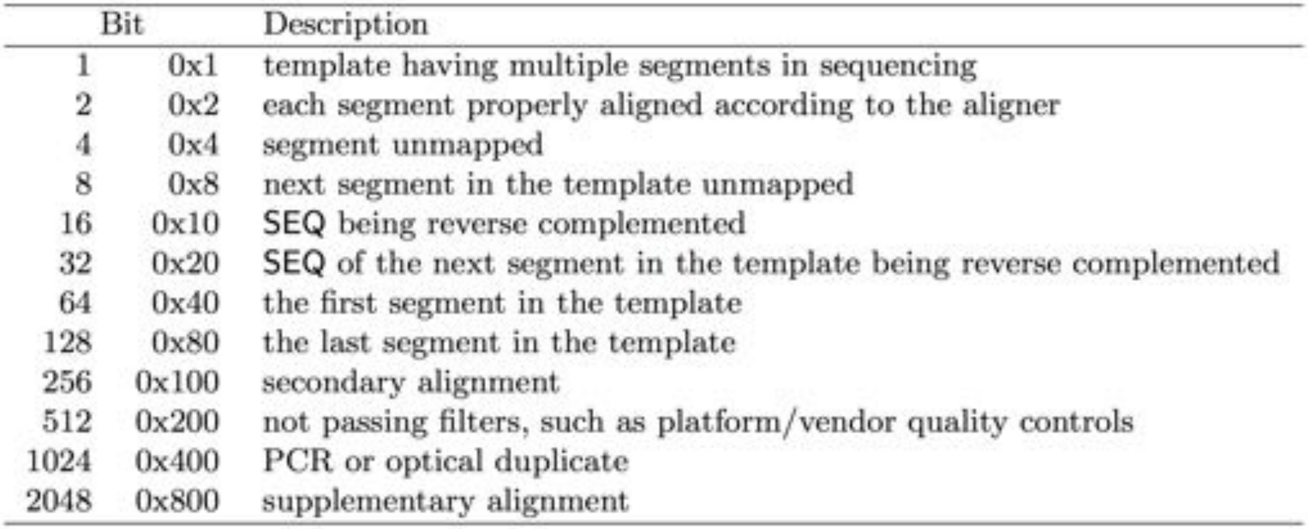


Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{31}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,2^{31}-1]	Position of the mate/next read
9	TLEN	Int	[-2^{31}+1,2^{31}-1]	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Source: https://samtools.github.io/hts-specs/SAMv1.pdf

Using SAM flags to filter subsets of reads

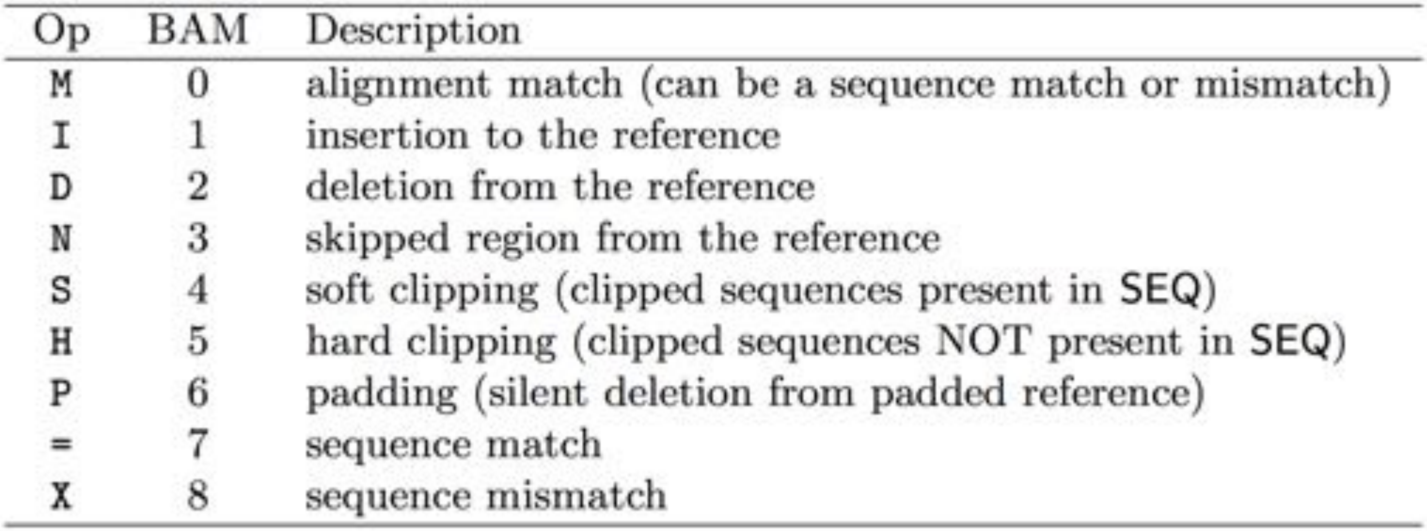
- 12 bitwise flags describing the alignment
- These flags are stored as a binary string of length 11
- Value of ‘1’ indicates the flag is set. e.g. 00100000000
- All combinations can be represented as a number from 1 to 2048 (i.e. $2^{11} - 1$). This number is used in the BAM/SAM file. You can specify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



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

Source: https://samtools.github.io/hts-specs/SAMv1.pdf

CIGAR string



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

- The CIGAR string is a sequence of base lengths and associated "operations" that are used to indicate which bases align to the reference (either a match or mismatch), are deleted, are inserted, represent introns, etc.
  - e.g. 81M859N19M
  - Read as: A 100 bp read consists of: 81 bases of alignment to reference, 859 bases skipped (an intron), 19 bases of alignment

Source: https://samtools.github.io/hts-specs/SAMv1.pdf

Browser Extensible Data (BED) format

- When working with BAM files, it is very common to want to examine reads aligned to a focused subset of the reference genome, e.g. the exons of a gene
- Focus on location - genomic coordinates
- Basic BED format (plain text, tab separated):
- Chromosome name, start position, end position
- Coordinates in BED format are 0 based

https://genome.ucsc.edu/FAQ/FAQformat#format1

GFF/GTF file format

- Generic feature format for storing genomic annotation data.
- Tab delimited text file (with optional header lines beginning ##):
    - contig (chromosome)
    - source
    - type
    - start
    - end
    - score
    - strand
    - phase
    - attributes

```
CR940346  Genes  CDS   1    1516  .  +  .  locus_tag "Tap370b08.q2ca38.01";product
"cytochrome C oxidase subunit I (COX1 homologue), putative"
CR940346  Exons  EXON  1    1516  .  +  .  gene_id "Tap370b08.q2ca38.01";transcript_id
"Tap370b08.q2ca38.01"
```

GFF specifications: https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

More about file formats, http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s008

Tools to work with SAM/BAM/BED files

SAM/BAM files

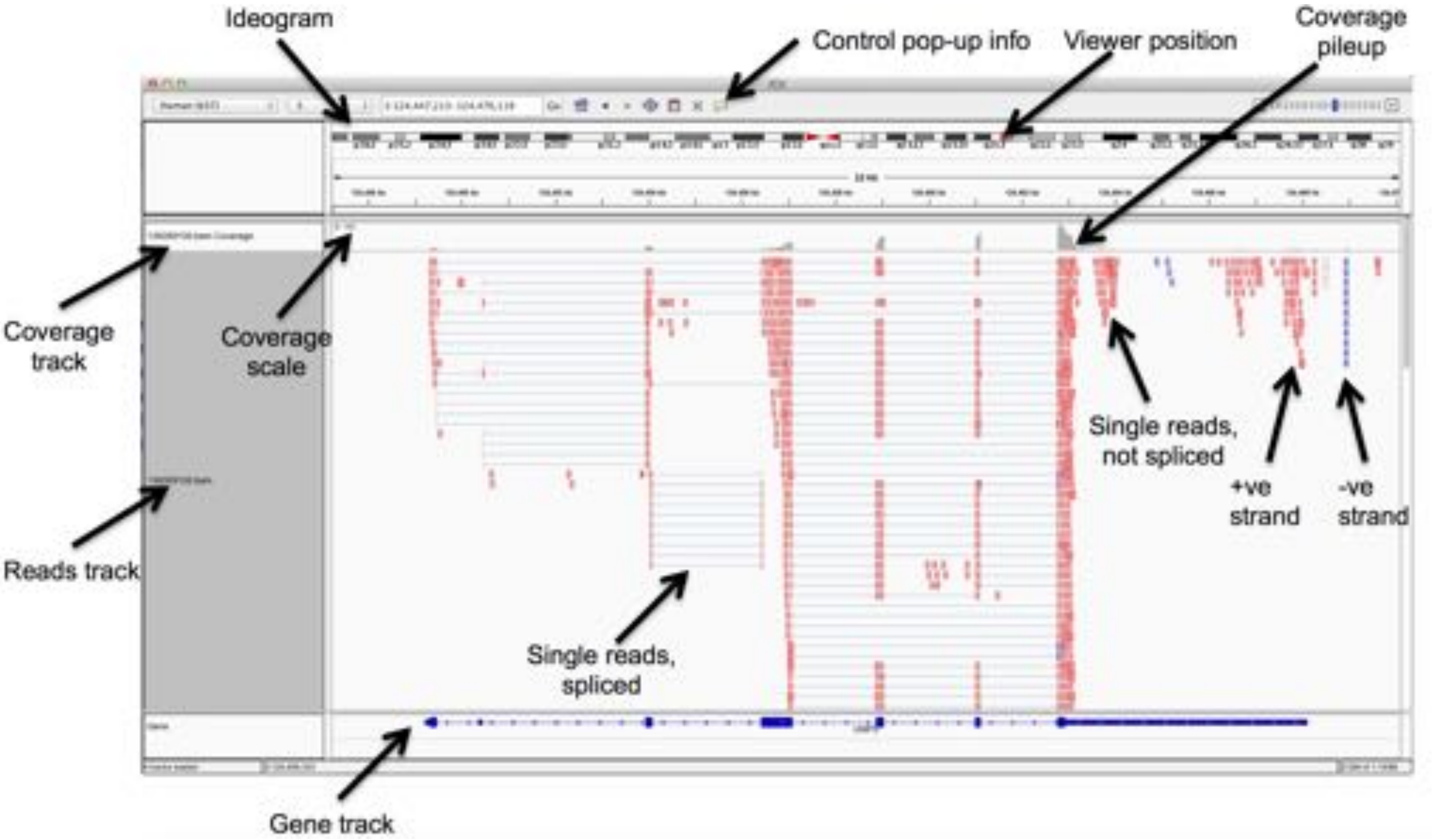
- `samtools`, view, sort, index, QC, stats on SAM/BAM files, and more, https://github.com/samtools/samtools
- `sambamba`, view, sort, index, merge, stats, mark duplicates. fast laternative to `samtools`, https://lomereiter.github.io/sambamba/index.html
- `picard`, QC, validation, duplicates removal and many more utility tools, https://broadinstitute.github.io/picard/

BED files

- `bedtools`, universal tools for manipulating genomic regions, https://bedtools.readthedocs.io/en/latest/
- `bedops`, complementary to `bedtools`, providing additional functionality and speedup, https://bedtools.readthedocs.io/en/latest/

Visualization

Integrative Genomics Viewer (IGV), http://software.broadinstitute.org/software/igv/



IGV

Features

- Explore large genomic datasets with an intuitive, easy-to-use interface.
- Integrate multiple data types with clinical and other sample information.
- View data from multiple sources:
  - local, remote, and "cloud-based".
  - Intelligent remote file handling - no need to download the whole dataset
- Automation of specific tasks using command-line interface
- Tutorial: https://github.com/griffithlab/rnaseq_tutorial/wiki/IGV-Tutorial

Gene expression analysis

Expression estimation for known genes and transcripts

HTSeq (htseq-count), http://www-huber.embl.de/HTSeq/doc/count.html

-
  ```
  htseq-count --mode intersec=on-strict --stranded no --minaqual 1
  --type exon --ida_r transcript_id accepted_hits.sam chr22.gff >
  transcript_read_counts_table.tsv
  ```
- Issues with `htseq-count`: http://seqanswers.com/forums/showthread.php?t=18068

featureCounts, http://bioinf.wehi.edu.au/featureCounts/

- Summarize multiple datasets at the same time:
-
  ```
  featureCounts -t exon -g gene_id -a annotation.gtf -o counts.txt
  library1.bam library2.bam library3.bam
  ```

Expression estimation for known genes and transcripts

- **Counts of reads**: The relative expression of a transcript is proportional to the number of cDNA fragmets that originate from it ~ number of aligned reads. Disadvantages: longer gene produce more reads, library depth (total counts) influence counts of individual transcripts
- **Counts per million**: counts scaled by the library depth in million units. $CPM = C * 10^6/N$
- **RPKM**: Reads Per Kilobase of transcript per Million mapped reads.
- **FPKM**: Fragments Per Kilobase of transcript per Million mapped reads.

Expression estimation for known genes and transcripts

- **FPKM** (or **RPKM**) attempt to normalize for gene size and library depth

$$RPKM\ (or\ FPKM) = (10^9 * C)/(N * L)$$

- C - number of mappable reads/fragments for a gene/transcript/exon/etc.
- N - total number of mappable reads/fragments in the library
- L - number of base pairs in the gene/transcript/exon/etc.

- **RSEM**: RNA-Seq by Expectation-Maximization, https://www.ncbi.nlm.nih.gov/pubmed/21816040

https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/

TPM: Transcript per Kilobase Million

FPKM is calculated as

1. Sum sample/library fragments per million
2. Divide gene/transcript fragment counts by #1 – fragments per million, FPM
3. Divide FPM by length of gene in kilobases (FPKM)

TPM reverses the order - length first, library size second

1. Divide fragment count by length of transcript – fragments per kilobase, FPK
2. Sum all FPK for sample/library per million
3. Divide #1 by #2 (TPM)

- http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/
- https://www.ncbi.nlm.nih.gov/pubmed/22872506

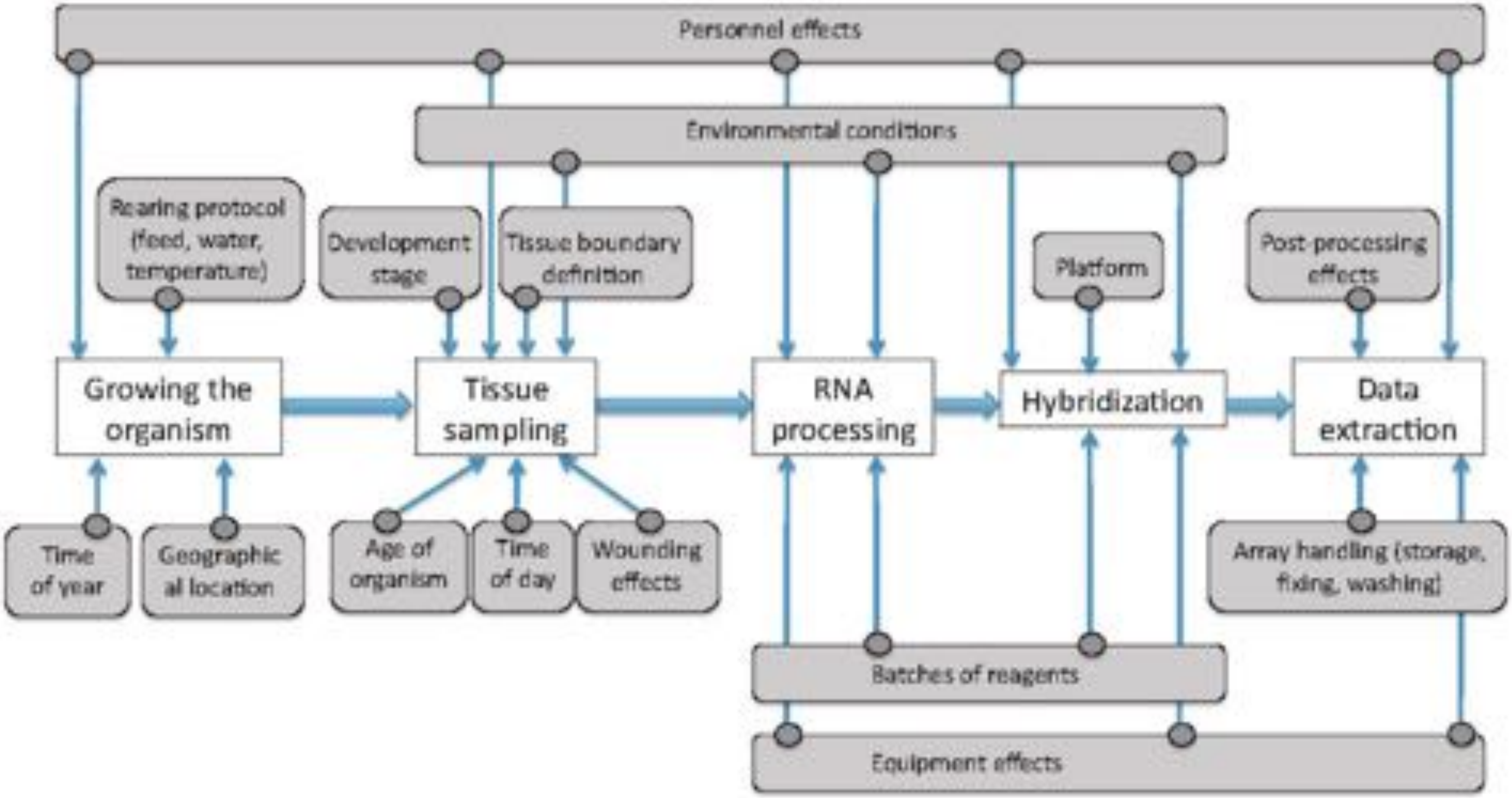
Scripts for RNA-seq data analysis

- https://github.com/mdozmorov/dcaf/tree/master/ngs.rna-seq - an example of a pipeline
- https://github.com/chapmanb/bcbio-nextgen - Validated, scalable, community developed variant calling, RNA-seq and small RNA analysis https://bcbio-nextgen.readthedocs.org

Batch effects

- Batch effects are widespread in high-throughput biology. They are artifacts not related to the biological variation of scientific interests.
- For instance, two experiments on the same technical replicates processed on two different days might present different results due to factors such as room temperature or the two technicians who did the two experiments.
- Batch effects can substantially confound the downstream analysis, especially meta-analysis across studies.

Batch sources



58/90

ComBat

ComBat - Location-scale method

The core idea of ComBat was that the observed measurement Y_{ijg} for the expression value of gene g for sample j from batch i can be expressed as

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg}$$

where X consists of covariates of scientific interests, while γ_{ig} and δ_{ig} characterize the additive and multiplicative batch effects of batch i for gene g.

https://www.bu.edu/jlab/wp-assets/ComBat/Abstract.html

ComBat

After obtaining the estimators from the above linear regression, the raw data Y_{ijg} can be adjusted to Y_{ijg}^*:

$$Y_{ijg}^* = \frac{Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g - \hat{\gamma}_{ig}}{\hat{\delta}_{ig}} + \hat{\alpha}_g + X\hat{\beta}_g$$

For real application, an empirical Bayes method was applied for parameter estimation.

https://www.bu.edu/jlab/wp-assets/ComBat/Abstract.html

SVA

When batches were unknown, the surrogate variable analysis (SVA) was developed.

The main idea was to separate the effects caused by covariates of our primary interests from the artifacts not modeled.

Now the raw expression value Y_{jg} of gene g in sample j can be formulated as:

$$Y_{jg} = \alpha_g + X\beta_g + \sum_{k=1}^{K} \lambda_{kg}\eta_{kj} + \epsilon_{jg}$$

where η_{kj}s represent the unmodeled factors and are called as “surrogate variables”.

SVA

Once again, the basic idea was to estimate η_{kj}s and adjust them accordingly.

An iterative algorithm based on singular value decomposition (SVD) was derived to iterate between estimating the main effects $\hat{\alpha}_g + X\hat{\beta}_g$ given the estimation of surrogate variables and estimating surrogate variables from the residuals $r_{jg} = Y_{jg} - \hat{\alpha}_g - X\hat{\beta}_g$

`sva` package in Bioconductor

- Contains `ComBat` function for removing effects of known batches.
- Assume we have:
    - `edata`: a matrix for raw expression values
    - `batch`: a vector named for batch numbers.

```
modcombat = model.matrix(~1, data=as.factor(batch))

combat_edata = ComBat(dat=edata, batch=batch, mod=modcombat, par.prior=TRUE, prior.plot=FALSE)
```

https://bioconductor.org/packages/release/bioc/html/sva.html

SVASEQ

For sequencing data, `svaseq`, the generalized version of SVA, suggested applying a moderated log transformation to the count data or fragments per kilobase of exon per million fragments mapped (FPKM) first to account for the nature of discrete distributions

Instead of a direct transformation on the raw counts or FPKM, remove unwanted variation (RUV) adopted a generalized linear model for Y_{jg}

BatchQC - Batch Effects Quality Control

A Bioconductor package with a GUI (shiny app).

https://github.com/mani2012/BatchQC

Differential expression analysis

- Many tools for differential expression analysis design their statistics around raw read counts
- Poisson distribution (single-parameter model, mean = variance)?
- Genes with larger average expression (counts) have on average larger variance across samples.
- Negative Binomial is a good approximation, https://www.ncbi.nlm.nih.gov/pubmed/23975260
- Variability is modeled by the dispersion parameter

Differential expression analysis

- `DESeq2` - https://bioconductor.org/packages/release/bioc/html/DESeq.html, https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8
- `edgeR` - https://bioconductor.org/packages/release/bioc/html/edgeR.html, https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp616

- Both use Negative Binomial distribution
- Differ in estimation of the dispersion parameter

Differential expression analysis

- `limma` - Linear Models for Microarray Data, https://bioconductor.org/packages/release/bioc/html/limma.html
- `voom` - variance modeling at the observational level transformation. Uses the variance of genes to create weights for use in linear models. https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-2-r29

- After `voom` transformation, the RNA-seq data can be analyzed using `limma`.
- https://gist.github.com/mdozmorov/fb7a1f40eb18699298442c3e77a0de02 - Differential expression analysis in RNA-seq, short
- https://gist.github.com/stephenturner/e34e32b3d054bb850ae2 - Differential expression analysis in RNA-seq, long

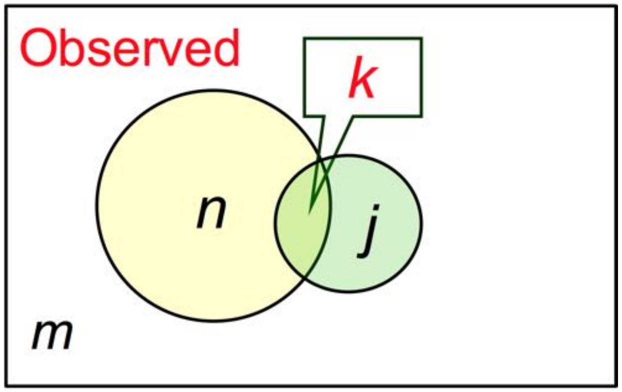
Functional interpretation analysis

Interpretation

- **Enrichment analysis** - high-level understanding of the biology behind hundreds differentially expressed genes
- **Gene annotations** - sets of genes with shared functions, or structured _a priori_knowledge about genes. Gene Ontology (http://geneontology.org/), MSigDb (http://software.broadinstitute.org/gsea/msigdb/) and MSigDf (https://github.com/stephenturner/msigdf), KEGG pathways (http://www.genome.jp/kegg/) and many more.

Enrichment analysis, Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category



Enrichment analysis, Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category

The expected value of k would be $k_e = (n/m) * j$.

If $k > k_e$, functional category is said to be enriched, with a ratio of enrichment $r = k/k_e$

Enrichment analysis, Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category

	Diff. exp. genes	Not Diff. exp. genes	Total
In gene set	k	j-k	j
Not in gene set	n-k	m-n-j+k	m-j
Total	n	m-n	m

Enrichment analysis, Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category

What is the probability of having k or more genes from the category in the selected n genes?

$$P = \sum_{i=k}^{n} \frac{\binom{m-j}{n-i}\binom{j}{i}}{\binom{m}{n}}$$

Enrichment analysis, Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category

$k < (n/m) * j$ - underrepresentation. Probability of k or less genes from the category in the selected n genes?

$$P = \sum_{i=0}^{k} \frac{\binom{m-j}{n-i}\binom{j}{i}}{\binom{m}{n}}$$

Enrichment analysis, Hypergeometric test

1. Find a set of differentially expressed genes (DEGs)
2. Are *DEGs in a set* more common than *DEGs not in a set*?

- Fisher test `stats::fisher.test()`
- Conditional hypergeometric test, to account for directed hierachy of GO `GOstats::hyperGTest()`

Example: https://github.com/mdozmorov/MDmisc/blob/master/R/gene_enrichment.R

Problems with Fisher's exact test

- The outcome of the overrepresentation test depends on the significance threshold used to declare genes differentially expressed.

- Functional categories in which many genes exhibit small changes may go undetected.

- Genes are not independent, so a key assumption of the Fisher’s exact tests is violated.

Many GO enrichment tools

- **GOStat**, http://gostat.wehi.edu.au/
- **GOrilla**, Gene Ontology enRIchment anaLysis and visuaLizAtion tool http://cbl-gorilla.cs.technion.ac.il/
- **g:Profiler**, http://biit.cs.ut.ee/gprofiler/
- **Metascape**, http://metascape.org/
- **ToppGene**, https://toppgene.cchmc.org/
- **WebGestalt** - WEB-based GEne SeT AnaLysis Toolkit, http://www.webgestalt.org/
- R packages, **clusterProfiler**, https://www.bioconductor.org/packages/devel/bioc/html/clusterProfiler.htn

Enrichment analysis, Functional Class Scoring (FCS)

- **Gene set analysis (GSA)**. Mootha et al., 2003; modified by Subramanian, et al. "**Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.**" PNAS 2005 http://www.pnas.org/content/102/43/15545.abstract
- Main rationale – functionally related genes often display a coordinated expression to accomplish their roles in the cells
- Aims to identify gene sets with "subtle but coordinated" expression changes that would be missed by DEGs threshold selection

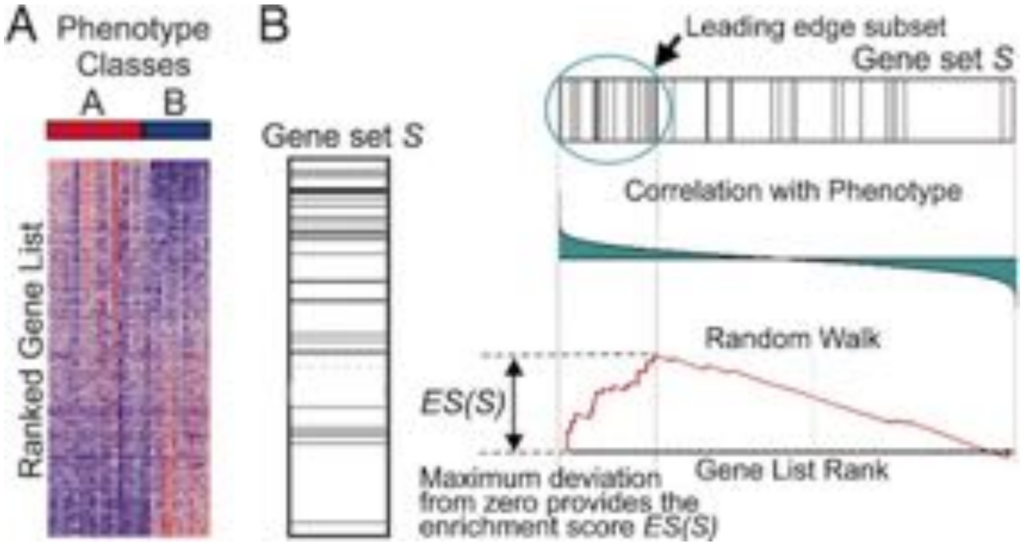
GSEA: Gene set enrichment analysis

- The null hypothesis is that the **rank ordering** of the genes in a given comparison is **random** with regard to the case-control assignment.

- The alternative hypothesis is that the **rank ordering** of genes sharing functional/pathway membership is **associated** with the case-control assignment.

GSEA: Gene set enrichment analysis

1. Sort genes by log fold change
2. Calculate running sum - increment when gene in a set, decrement when not
3. Maximum of the runnig sum is the enrichment score - larger means genes in a set are toward top of the sorted list
4. Permute subject labels to calculate significance p-value



Other approaches

Linear model-based

- **CAMERA** (Wu and Smyth 2012)
- **C**orrelation-**A**djusted **ME**an **RA**nk gene set test
- Estimating the variance inflation factor associated with inter-gene correlation, and incorporating this into parametric or rank-based test procedures

Other approaches

Linear model-based

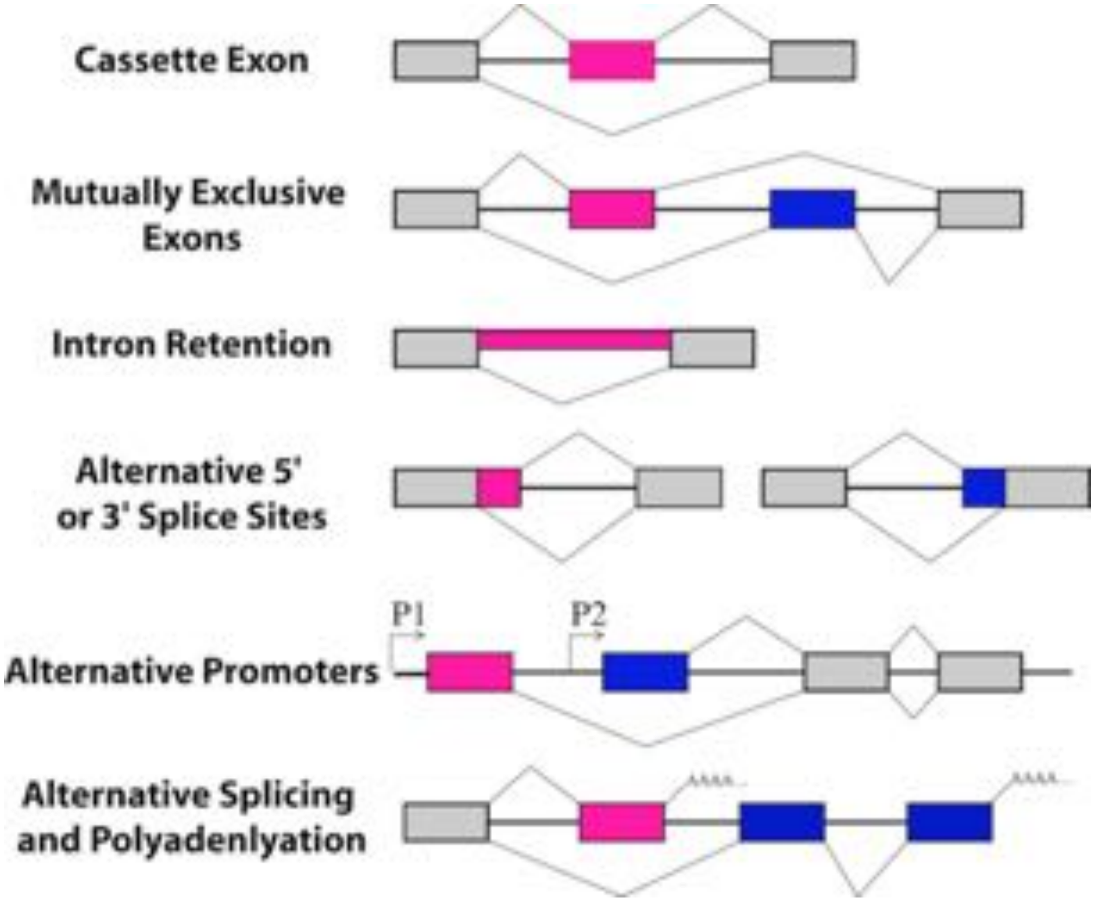
- **ROAST** (Wu et.al. 2010)
- Under the null hypothesis (and assuming a linear model) the residuals are independent and identically distributed $N(0, \sigma_g^2)$.
- We can *rotate* the residual vector for each gene in a gene set, such that gene-gene expression correlations are preserved.

Genome analysis platform: Galaxy

- Open Source academic project, https://usegalaxy.org/
- A web-based user-friendly interface that allows you to run existing workflows or create custom analyses by combining tools in the Galaxy 'toolshed'
- Example of RNA-seq workflow: https://usegalaxy.org/workflow/display_by_username_and_slug?username=mwolfien&slug=rnaseq-wolfien-pipeline

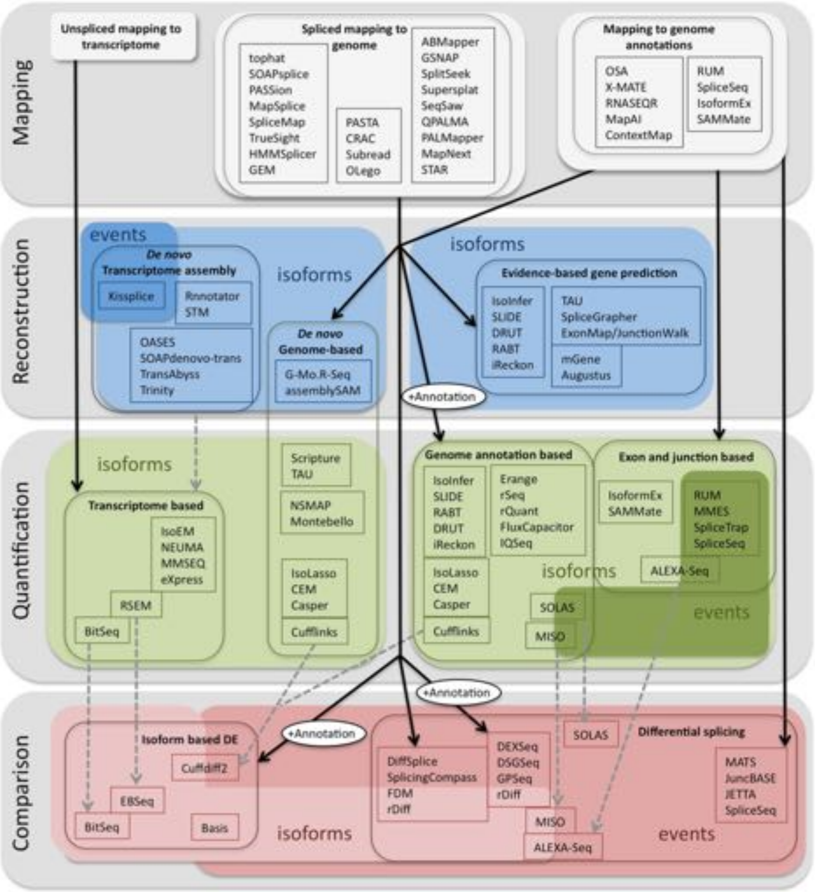
List of other genome analysis plaforms - https://docs.google.com/spreadsheets/d/1o8iYwYUy0V7IECmu21Und3XALw0itk/pubhtml

Alternative splicing



85/90

Alternative splicing



Alamancos, G. et.al. "Methods to Study Splicing from High-Throughput RNA Sequencing Data." *Spliceosomal Pre-mRNA Splicing: Methods and Protocols*, 2014 https://www.ncbi.nlm.nih.gov/pubmed/24549677

Alternative splicing

Best approach to predict novel and alternative splicing events from RNA-seq data

- https://www.biostars.org/p/68966/
- https://www.biostars.org/p/62728/

Alternative splicing detection

- https://www.biostars.org/p/65617/
- https://www.biostars.org/p/11695/

Identifying genes that express different isoforms in cancer vs normal RNA-seq data

- https://www.biostars.org/p/50365/

Visualization of alternative splicing events using RNA-seq data

- https://www.biostars.org/p/8979/

References

- Griffith, M. et.al. "Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud." *PLoS Computational Biology* 2015 - RNA-seq overview and extensive supplementary material http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393. The complete practical RNA-seq tutorial https://github.com/griffithlab/rnaseq_tutorial

- Conesa, A. et.al. "A Survey of Best Practices for RNA-Seq Data Analysis."" *Genome Biology* 2016. http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

- Law, C. et.al. "RNA-Seq Analysis Is Easy as 1-2-3 with Limma, Glimma and edgeR." *F1000Research* 2016. - Latest Rsubread-limma plus pipeline https://f1000research.com/articles/5-1408/v2. The complete R code for RNA-seq analysis tutorial https://www.bioconductor.org/help/workflows/RNAseq123/

References

- Pertea, M. et.al. “Transcript-Level Expression Analysis of RNA-Seq Experiments with HISAT, StringTie and Ballgown.” *Nature Protocols* 2016. http://www.nature.com/nprot/journal/v11/n9/full/nprot.2016.095.html

- Williams, A. et.al. "RNA-Seq Data: Challenges in and Recommendations for Experimental Design and Analysis: RNA-Seq Data: Experimental Design and Analysis." 2014. http://onlinelibrary.wiley.com/doi/10.1002/0471142905.hg1113s83/abstract

- Tools for RNA-seq data analysis, http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s004

Thank you

Questions?

This presentation on GitHub:

https://github.com/mdozmorov/presentations

Mikhail Dozmorov, Ph.D.

Assistant professor, Department of Biostatistics, VCU

mikhail.dozmorov@vcuhealth.org