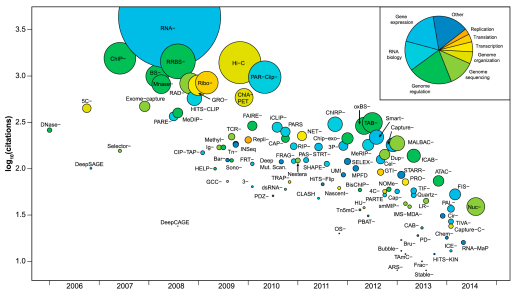
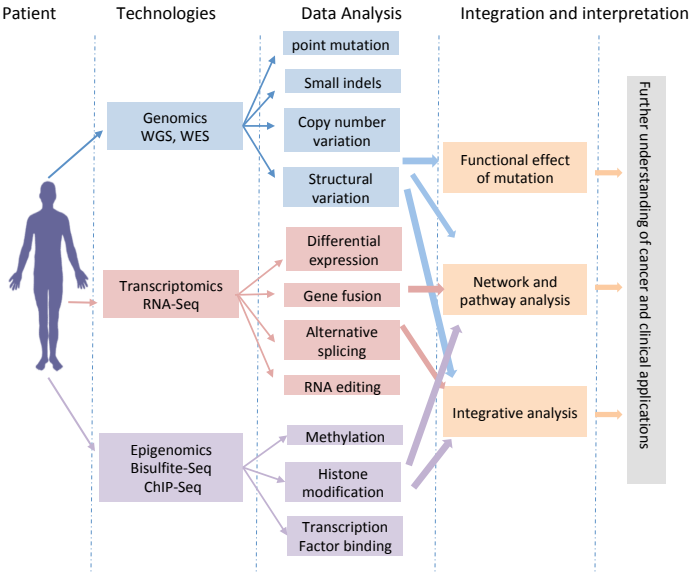
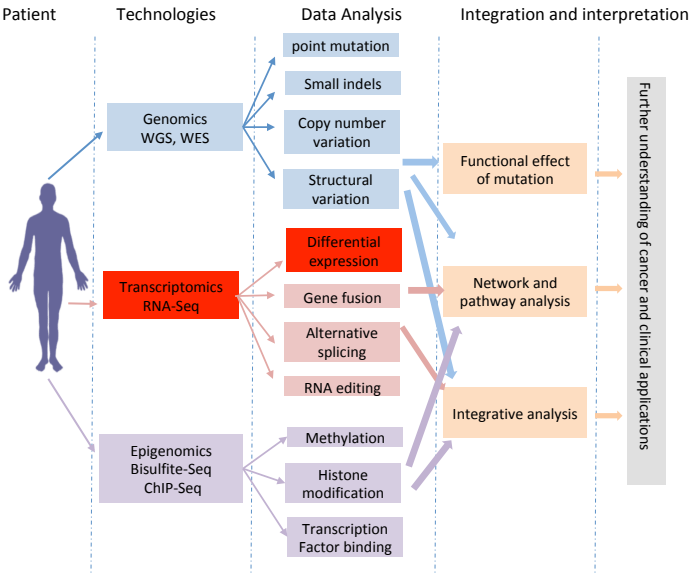
Case Study: RNA-seq analysis

What protocol answers my question of interest?

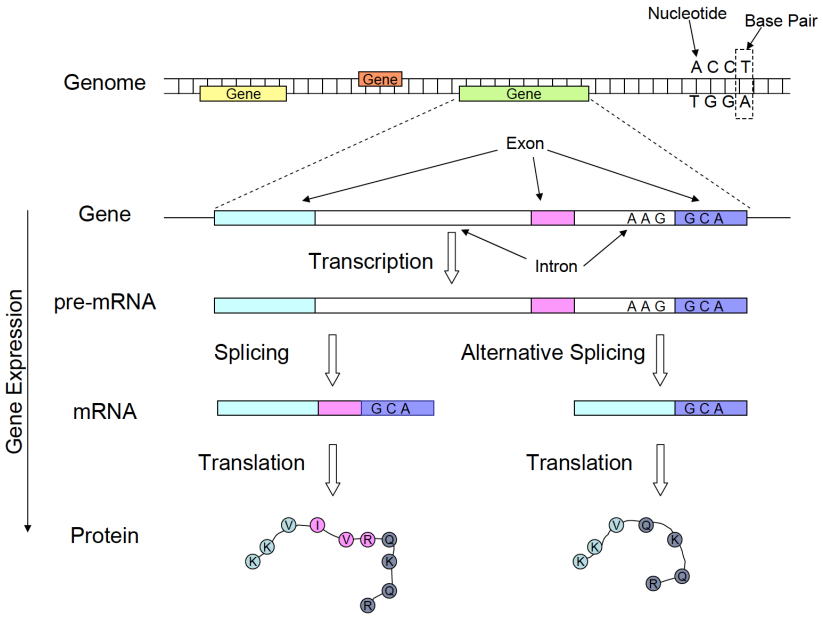




Shyr D, Liu Q. Biol Proced Online. (2013)15,4



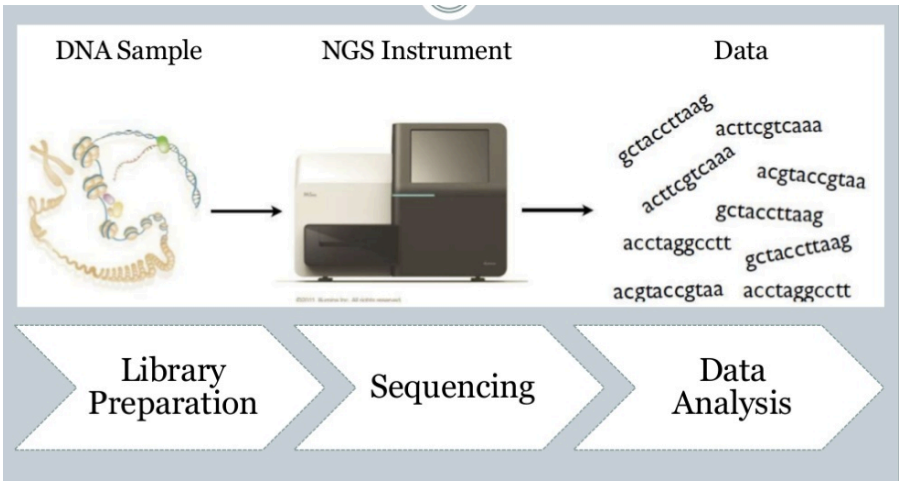
Shyr D, Liu Q. Biol Proced Online. (2013)15,4



92–94% of human genes undergo alternative splicing, 86% with a minor isoform frequency of 15% or more

E.T. Wang, et al, Nature 456, 470-476 (2008)

NGS sequencing pipeline



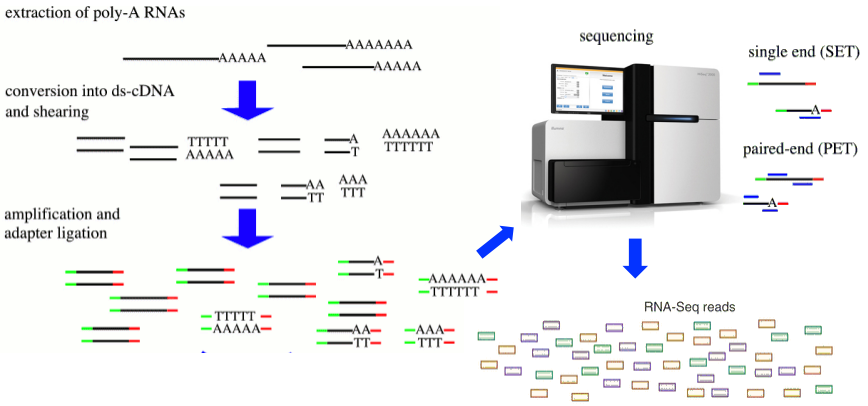
http://www.slideshare.net/mkim8/a-comparison-of-ngs-platforms

RNA-seq library options

- How do I select the RNAs of interest?
  - Poly-A selection?
  - Ribosomal subtraction (Total RNA)?
  - Size selection? (more rare)
- What sequencing strategy was utilized?
  - Single end sequencing (inexpensive, least informative)
  - Paired end sequencing (expensive but highly informative)

Overview of RNA-Seq

Transcriptome profiling using NGS



Depth of Sequencing

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3227109/

- How much depth is necessary?
  - Different total transcription levels
  - Different numbers of transcribed genes
  - Different levels of transcriptome complexity
  - Different distributions of expression levels
- What analysis is intended?
  - Gene level summaries needs less depth (~30M)
  - Isoform inference requires much higher depth (100-200M)
  - Detect everything? (> 800M reads)

Replicates are ALWAYS more important than depth.

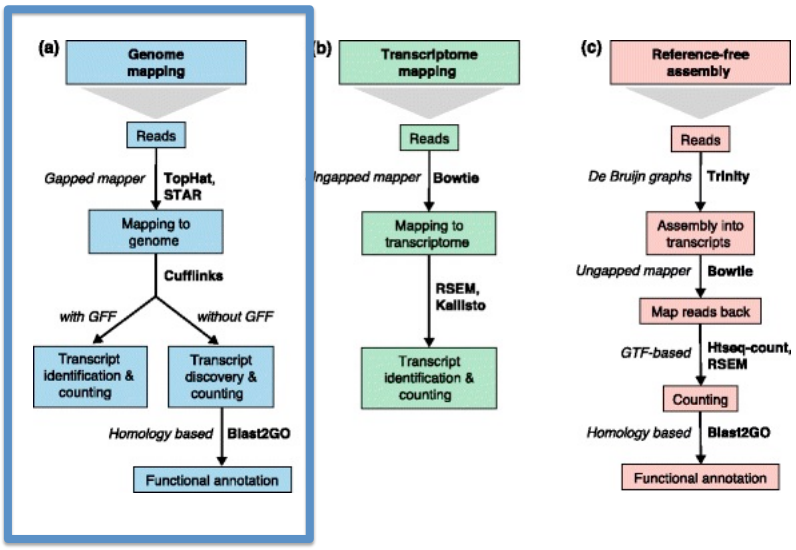
- Only via BIOLOGICAL replicates can variability in growth conditions, facilities, handling and individuals be managed.
- Statistical power is predominantly via REPLICATION.

https://www.nature.com/articles/nbt.1910

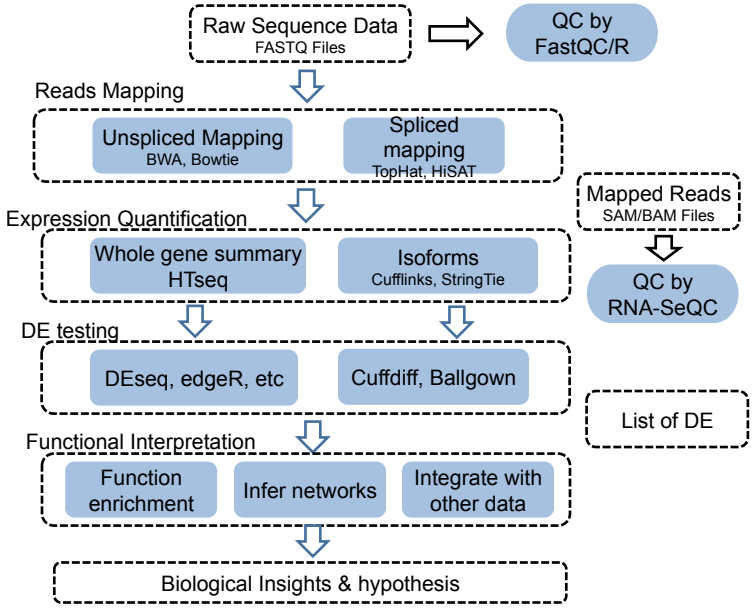
Probability of detecting differential expression in a single test

	Replicates per group		
	3	**5**	**10**
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

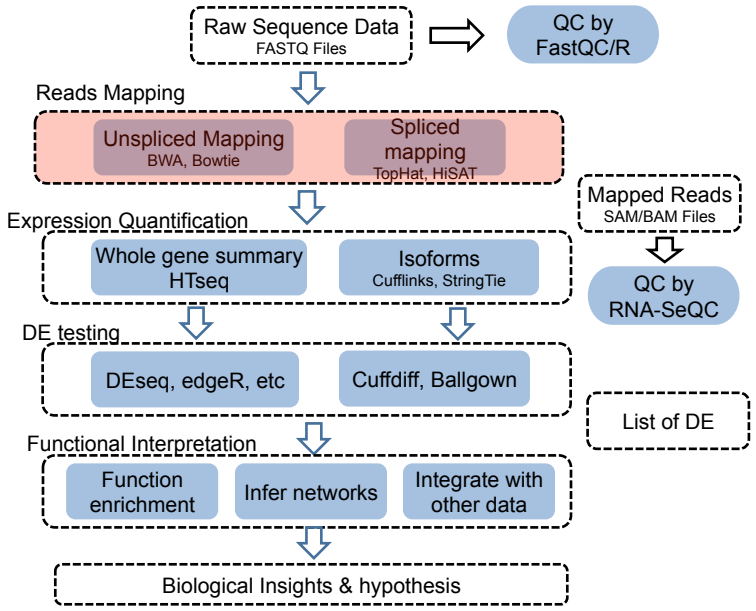
How do I figure out an analysis pipeline?



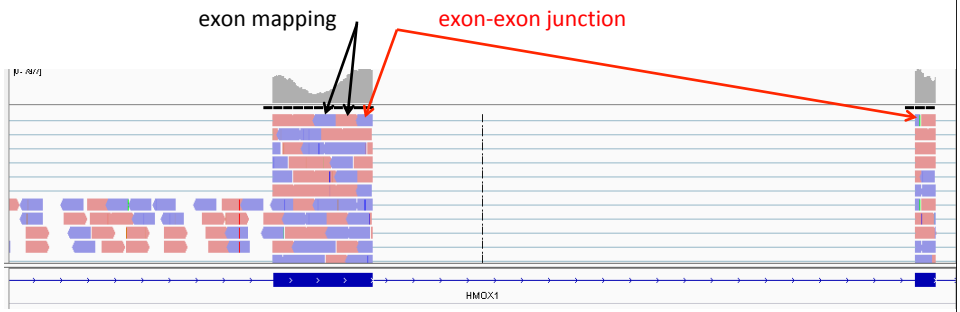
From reads to differential expression



From reads to differential expression

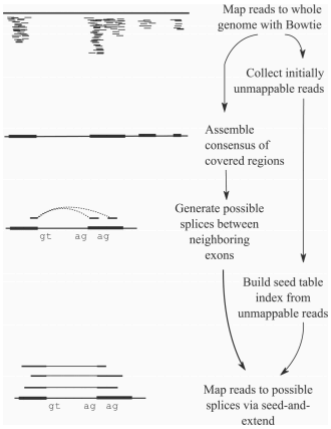


Read mapping



Unlike DNA-Seq, when mapping RNA-Seq reads back to reference genome, we need to pay attention to **exon-exon junction reads**

Reference Mapping - TOPHAT



INPUT
FASTQ (processed)
Output (4 files)
Insertions (.bed)
Deletions (.bed)
Junctions (.bed)
Accepted Hits (.bam)

TOPHAT provides both identifying and quantifying information

.bed files can be downloaded to excel
-sam (Sequence Aligment/Map) or bam (binary compressed version of sam) – can be used to visualize reads using UCSC Genome Browser or Integrative Genomics Viewer

Link to File type descriptions
https://genome.ucsc.edu/FAQ/FAQformat.html#format1

How do I run TopHat or HiSAT?

https://ccb.jhu.edu/software/tophat/manual.shtml

http://www.ccb.jhu.edu/software/hisat/manual.shtml

HiSAT is newer (published 2015), uses a different indexing scheme, requires less memory, and is a bit faster than TopHat. Results are comparable.

Running TopHat

```
### Assign path variables
TOPHAT="/scratch/Users/USERNAME/tophat_out"
SCRATCH='/scratch/Users/USERNAME'

### Make tophat output directory in scratch
mkdir /scratch/Users/USERNAME/tophat_out/

### Get annotation file
rsync /Users/dowellde/4521/Hg18/Hg18.chr10.refseq.gtf $SCRATCH/

### Load the tophat2 module and its dependencies
module load bowtie/2.2.9
module load samtools/1.3.1
module load tophat/2.1.1

### Run my commands, or whatever else IN SCRATCH

### map reads with tophat
### tophat [options] <bowtie_index> <reads1[,reads2,...]>
tophat2 --b2-fast -p 12 -r 325 --mate-std-dev 150 --microexon-search \
--library-type fr-firststrand --rg-id USERNAME --rg-sample Human \
--no-novel-juncs -o $TOPHAT -G $SCRATCH/Hg18.chr10.refseq.gtf \
$SCRATCH/Bowtie2Indexes/Hg18 $SCRATCH/RNA_Eli_repA_R1.fastq \
$SCRATCH/RNA_Eli_repA_R2.fastq
```

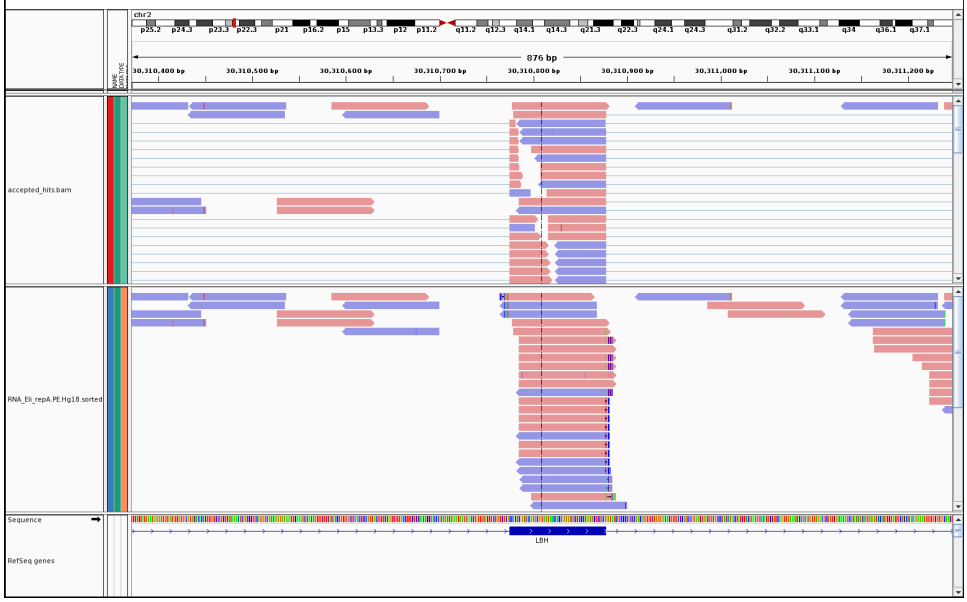
```
### rename tophat file (optional, but helps to document!)
mv $TOPHAT/accepted_hits.bam $TOPHAT/RNA_Eli_repA.tophat.accepted_hits.bam

### get alignment stats
samtools flagstat $TOPHAT/RNA_Eli_repA.tophat.accepted_hits.bam > $TOPHAT/
RNA_Eli_repA.tophat.accepted_hits.alignment_stats.txt

### create an index for accepted_hits.bam
samtools index $TOPHAT/RNA_Eli_repA.tophat.accepted_hits.bam

### MOVE MY OUTPUTS BACK TO HOME STORAGE
rsync $TOPHAT/RNA_Eli_repA.chr10.tophat.accepted_hits.bam $HOME/
rsync $TOPHAT/RNA_Eli_repA.chr10.tophat.accepted_hits.bam.bai $HOME/
rsync $TOPHAT/RNA_Eli_repA.chr10.tophat.accepted_hits.alignment_stats.txt $HOME/
```

Running HiSAT

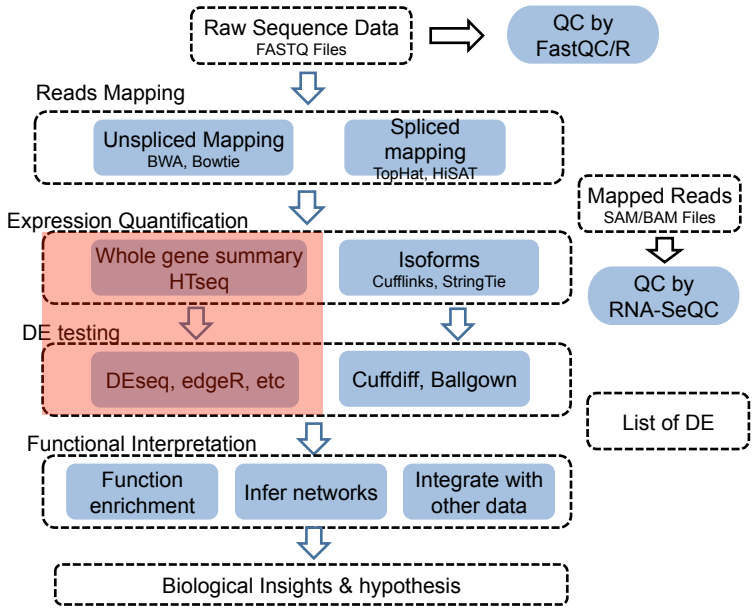
```
### Define input, output, and stderr path along with name for input files (the basename: up to
'_R*' part of filename).
INPATH='<PATH_TO_TRIMMED_FASTQ_DIRECTORY>'
OUTPATH='<PATH_TO_DIRECTORY_WHERE_SAM_AND_BAM_FILES_ARE_TO_BE_WRITTEN>'
ERRPATH=${OUTPATH}'stderr/'

INPUTFILE='<BASENAME_OF_INPUT_FILES>'

# Make directories you will be writing to
mkdir ${OUTPATH}
mkdir ${ERRPATH}


### Define genome index directory (include index file prefix, up to the .#.ht2 suffix)
GENOMEIDX='/scratch/Users/USERNAME/HISAT2_indexes/rn6/genome_rn6'
printf "\nYou are using genome index: ${GENOMEIDX}\n"


# Load modules
MODULES=('samtools/1.3.1' 'hisat2/2.1.0')
```

```
# Define read1, read2, sam, and err file names for paired reads
PAIR_R1=${INPATH}${INPUTFILE}'_R1.paired.trim.fq.gz'
PAIR_R2=${INPATH}${INPUTFILE}'_R2.paired.trim.fq.gz'
PAIR_SAM=${OUTPATH}${INPUTFILE}'.paired.trim.sam'
PAIR_SAMERR=${ERRPATH}${INPUTFILE}'.paired.trim.sam.stderr'
### RUN HISAT ON PAIRED READ FASTQ
#########################################################
hisat2 -p 32 -x ${GENOMEIDX} -1 ${PAIR_R1} -2 ${PAIR_R2} -S ${PAIR_SAM} 2> ${PAIR_SAMERR}



# Define read1, read2, sam, and err file names for unpaired reads
UNPAIR_R1=${INPATH}${INPUTFILE}'_R1.unpaired.trim.fq.gz'
UNPAIR_R2=${INPATH}${INPUTFILE}'_R2.unpaired.trim.fq.gz'
UNPAIR_SAM=${OUTPATH}${INPUTFILE}'.unpaired.trim.sam'
UNPAIR_SAMERR=${ERRPATH}${INPUTFILE}'.unpaired.trim.sam.stderr'
### RUN HISAT2 ON UNPAIRED READ FASTQ
#########################################################
hisat2 -p 32 -x ${GENOMEIDX} -U ${UNPAIR_R1},${UNPAIR_R2} -S ${UNPAIR_SAM} 2> $
{UNPAIR_SAMERR}
```

Tophat vs Bowtie

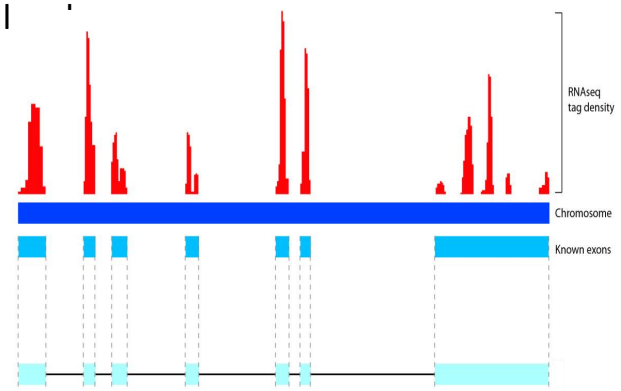


From reads to differential expression



Expression quantification

- Count data
  - Summarized mapped reads to CDS, gene or exon



Expression quantification

The number of reads is roughly proportional to

- the length of the gene
- the total number of reads in the library

Question:

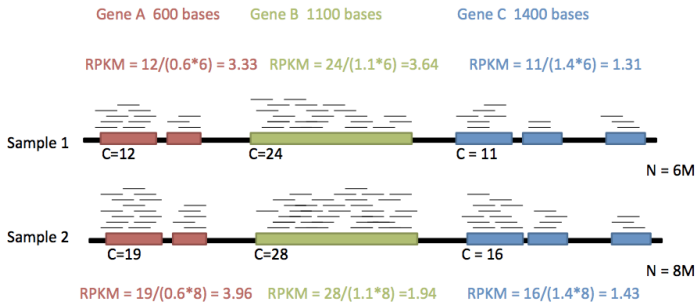
Gene A: 200

Gene B: 300

Expression of Gene A < Expression of Gene B?

How do I quantify expression from RNA-seq?

RPKM: Reads per Kb million (Mortazavi et al. Nature Methods 2008)
FPKM: Fragments per Kb million

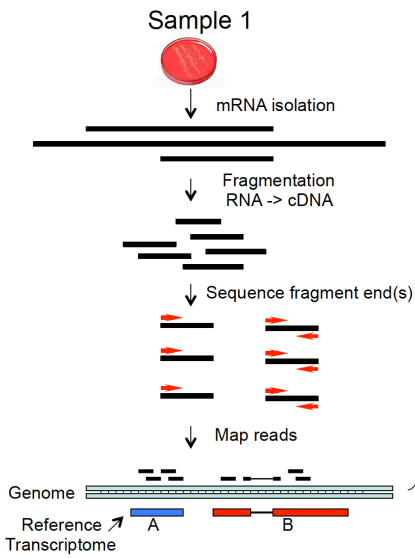


Longer and more highly expressed transcripts are more likely be represented among RNA-seq reads

RPKM normalizes by transcript length and the total number of reads captured and mapped in the experiment

Sequencing depth can alter RPKM values

$$FPKM = \frac{Counts\ of\ mapped\ fragments}{Total\ mapped\ fragments\ (million) \times Exon\ length\ of\ transcript\ (KB)}$$



Calculate transcript abundance

	Gene A	Gene B
Sample 1	4	4

# of Reads

	Gene A	Gene B
Sample 1	4	2

Reads per kilobase of exon

	Gene A	Gene B	Total
Sample 1	4	2	6
Sample 2	7	5	12

Reads per kilobase of exon

	Gene A	Gene B	Total
Sample 1	.7	.3	6
Sample 2	.6	.3	12

Reads per kilobase of exon per million mapped reads

RPKM

bedtools/2.25.0

How do I calculate RPKM/FPKM?

BedTools: http://bedtools.readthedocs.io/en/latest/index.html

coverageBed

http://bedtools.readthedocs.io/en/latest/content/tools/coverage.html

BedTools has a large number of useful programs for genomic arithmetic. ***VERY useful.***

But for quantifying coverage, it is limited by needing "regions" (i.e. it isn't aware of gene structure − YOU have to describe it).

HTseq

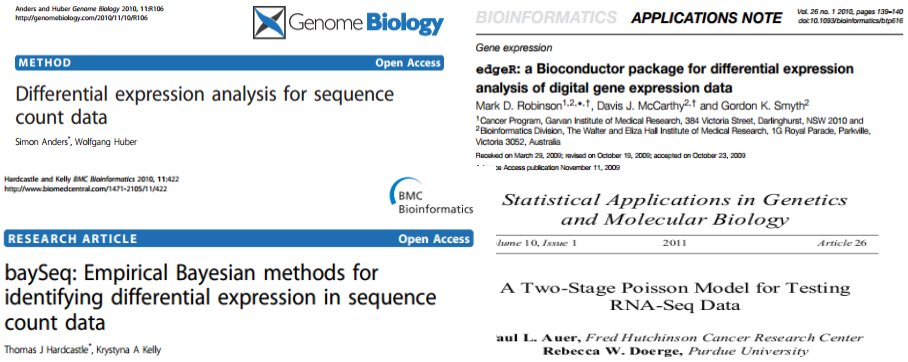
http://htseq.readthedocs.io/en/master/count.html

Specialized tool for specifically counting coverage over annotated genes. Can handle exon/intron structure (provided in a gff or gtf file) and function "correctly".

Lots of options to specify what "correctly" means in YOUR case.

Count-based methods (R packages)

1. **DESeq** -- based on negative binomial distribution
2. **edgeR** -- use an overdispersed Poisson model
3. **baySeq** -- use an empirical Bayes approach
4. **TSPM** -- use a two-stage poisson model



Anders and Huber *Genome Biology* 2010, 11:R106
http://genomebiology.com/2010/11/10/R106

Genome Biology

METHOD Open Access

Differential expression analysis for sequence count data

Simon Anders*, Wolfgang Huber

Hardcastle and Kelly *BMC Bioinformatics* 2010, 11:422
http://www.biomedcentral.com/1471-2105/11/422

BMC Bioinformatics

RESEARCH ARTICLE Open Access

baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data

Thomas J Hardcastle*, Krystyna A Kelly

BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 1 2010, pages 139–140
doi:10.1093/bioinformatics/btp616

Gene expression

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

Mark D. Robinson[1,2,*,†], Davis J. McCarthy[2,†] and Gordon K. Smyth[2]

[1]Cancer Program, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, NSW 2010 and [2]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia

Received on March 29, 2009; revised on October 19, 2009; accepted on October 23, 2009

Advance Access publication November 11, 2009

Statistical Applications in Genetics and Molecular Biology

Volume 10, Issue 1 2011 Article 26

A Two-Stage Poisson Model for Testing RNA-Seq Data

Paul L. Auer, *Fred Hutchinson Cancer Research Center*
Rebecca W. Doerge, *Purdue University*

R/3.3.0

How do I run DEseq?

http://bioconductor.org/packages/release/bioc/html/DESeq.html

Like many bioinformatics programs, it's in R.

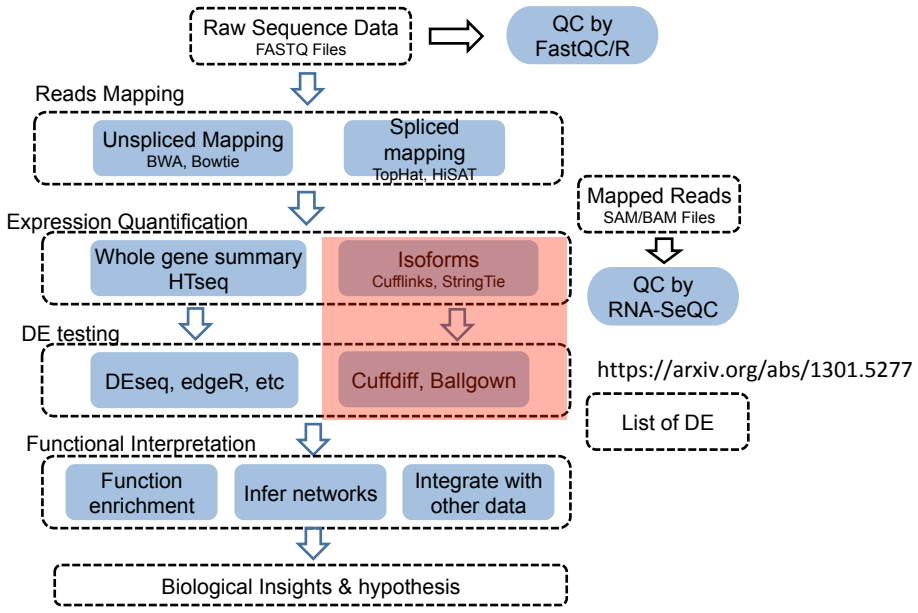
However, you can run R programs on the cluster by using "scripts" of R code …

```
R CMD BATCH --no-save --no-restore
Hah2011bidirectional_hits_intervals_092015.bed.count.bed.vehicleE2_40m.R
```

So now I have to learn R too???!

- No … and Yes.
  - For this class, you are fine to just do Cuffdiff and forgo learning R. That said, DEseq is a far better statistical model, particularly in the case of replicates.
  - In next few weeks, we'll play with R ***a little*** on the cluster and see some examples of running DEseq.

From reads to differential expression



Isoform inference

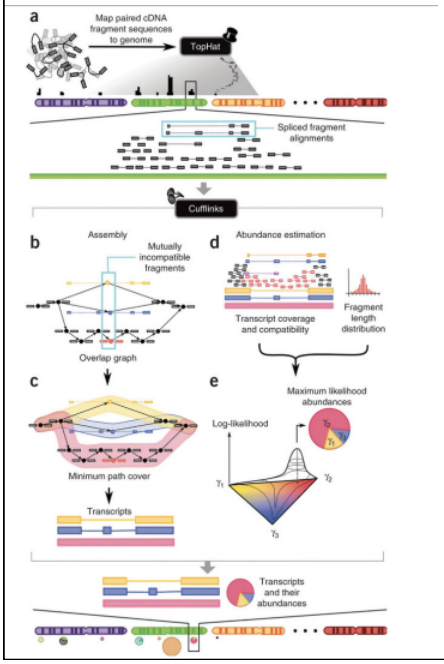
- Cufflinks expects TopHat output

  http://cole-trapnell-lab.github.io/cufflinks/manual/

- StringTie expects HiSAT output

  http://ccb.jhu.edu/software/stringtie/index.shtml?t=manual

Estimating Transcript Abundance - Cufflinks



<u>INPUT</u>
.bam file (Accepted Hits, e.g. from TopHat)

<u>Reference (.gtf)</u>
Refseq, Ensembl, etc

<u>Output</u> (tabular form, excel)
FPKM quantifiable

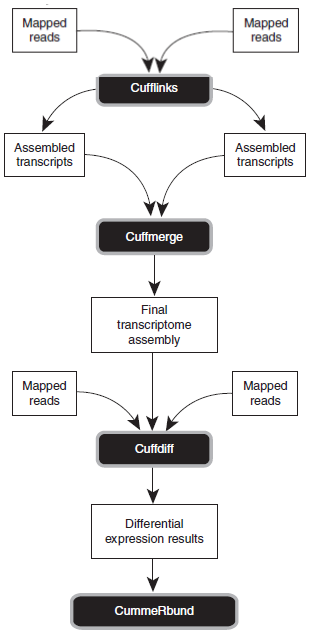
Cufflinks & Cuffdiff

Nature Protocols 7, 562-578 (2012)

PROTOCOL

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell[1,2], Adam Roberts[3], Loyal Goff[1,2,4], Geo Pertea[3,6], Daehwan Kim[3,7], David R Kelley[1,2], Harold Pimentel[3], Steven L Salzberg[3,8], John L Rinn[1,2] & Lior Pachter[3,8,9]



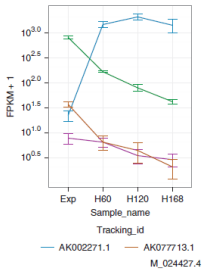
How do I run Cufflinks?

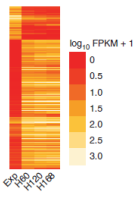
```
### Assign path variables
SCRATCH='/scratch/Users/USERNAME'
TOPHAT='/scratch/Users/USERNAME/tophat_out'

### Load modules
module load cufflinks/2.2.1

### Run my commands, or whatever else IN SCRATCH
### cufflinks [options] <aligned_reads.(sam/bam)> where BAM/SAM is sorted by position
cufflinks -p 12 -b $SCRATCH/Hg18.fa -u --library-type fr-firststrand \
-m 325 -s 150 -G $SCRATCH/Hg18.refseq.gtf --no-faux-reads \
--no-update-check -o $TOPHAT/ $TOPHAT/accepted_hits.bam


### MOVE MY OUTPUTS BACK TO HOME STORAGE
rsync $TOPHAT/transcripts.gtf $HOME/RNA_Eli_repA.cufflinks.transcripts.gtf
rsync $TOPHAT/isoforms.fpkm_tracking $HOME/
RNA_Eli_repA.cufflinks.isoforms.fpkm_tracking
rsync $TOPHAT/genes.fpkm_tracking $HOME/RNA_Eli_repA.cufflinks.genes.fpkm_tracking
```

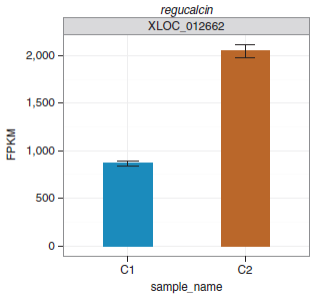
CummeRbund

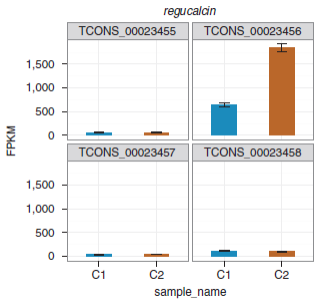
http://compbio.mit.edu/cummeRbund/manual_2_0.html

Much like DEseq, CummeRbund is an R package.









Differential Gene Expression Analysis

CuffDiff: If you have two samples, cuffdiff tests, for each transcript whether there is evidence that the concentration of this transcript is not the same in the two samples

DESeq/EdgeR: If you have two different experimental conditions, with replicates for each condition, DESeq tests whether, for a given gene, the change in the expression strength between the two conditions is large as compared to the variation within each group.

You will get different answers with different tests

Differential Gene Expression Analysis

RPKM
- -Can calculate Fold change
- -Input sequence reads must be similar
- -replicates not needed
- -provides NO statistical test for differential gene expression
- -useful for Cluster based classification of genes and other metrics/graphing

CuffDiff (part of Cufflinks package)
- -Input .bam file
- -Can set statistical threshold (p<0.05 or whatever)
- -replicates encouraged but not needed
- -Input sequence reads can be somewhat dissimilar
- -can provide differential splicing and promoter usage

DESeq (Technically an R program)
- -Input .bam file
- -Can set statistical threshold
- -Input sequence reads can be somewhat dissimilar
- -**Must have replicates**

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4728800/