RNAseq analysis

Overview

- Data QC
- File formats
- Aligning sequence reads
- Viewing data on a genome
- Calculate expression levels

Pipeline RNAseq

Tool	Function(s)
FastQC	QC
Bowtie/TopHat	Splice-aware alignment
Samtools	Manipulate alignment (SAM/BAM) files
IGV/UCSC	Data viewing
htseq-count	Calculate gene expression

Fastq - sequence format

```
@HWI-ST586:57:C04AYACXX:1:1101:4625:16394/1
TTGAAAAACTCCTGTAGTGAAGAAAGAAAATAACTAATATACAGTGACAATATAAATATTATAAGTATTTATTGTATTATCGCCCTGAAATTTGATAATA
+
CCCFFFFFHHHGHJIIJHIJEIJIJIJJJJGIIJJJJIFGIJJJGIIJIGIGIJJIIIIJJJIJJGIJJJJIJJJHIJIJHGJHHHFFFFFDEEEEEEDC
@HWI-ST586:57:C04AYACXX:1:1104:15832:86791/1
CAGCTTTGCAGCTAAGTTTGCATTCCCAGCCCTTCTCAATAGCCATTTTGAAATATATCAAGGAAATATATTTAGGGGTAAAATATATTAGTTTCCTTCA
+
CCCFFFFFGHHHHJIIHIIJIGHJJJIIJJJIIJJJIIIJJIJIJHIJJIJJJJJJGGIIIIJJEIJJJIIIHGIGIIBBHHGDDFFFFDECCEDEEEDD
@HWI-ST586:57:C04AYACXX:1:1202:17226:150911/1
GAGTAAGATCTTCTCTTGTTCCTTTGGACCATGGACCAGTGTTCCTCACTGGGAACGTGGGCTCCTGACACTTCAAAACTGCCAATTTGCTGGGGCAAAG
+
BBCDFFFFHHHHHJJJJJHJJJJJJJJJJJJJJJJJJJIIIIIJJJJJJJJJJJJJJJIJJJJJJJJJJJHHHHHFDBDFFEEEECDDDDDDDDDDDDDDD
@HWI-ST586:57:C04AYACXX:1:1306:1790:105947/1
CTGGGTTTAGAAAAAGGGTTTTGGAGCCAAGGTTGCATCATGCAGATGAAGCCTCCAGGTAGCAGGCTTCAGAGAGAATAGATTGTAATTGTTTCTTAGT
+
CCCFFEFFHHHHHJJJJJGHIJJJGIJJJJJJIIJJGIJJJJJJJIJJIIJJJJJJJJJHIIJJHHHHFFFFDECA=CCDDDEEDDEEFEDDDDDDDDDD
@HWI-ST586:57:C04AYACXX:1:1106:2186:124922/1
CAGTTTCCATGTAGTTGAGCAGTTTTGAGTGAGTTTCTTAATCCTGAGTTCTAGTTTGATTGCACTGTGGTCTGAGAGACAGTTTGTTATAATTTCTGTT
+
?@@D;DD>F?FHDHIIBFCGEHIHIIC=EHIII9EF>FHIIHI@ECAH:BBFDE0?DFBFHIIIIIIIGEFGIIICFGHGHHHCHFHF@BD>BEC@CCEE
```

Fastq format – fasta with qualities

- p = the probability that the corresponding base call is wrong
- Qualities

$$Q_{\text{sanger}} = -10 \log_{10} p$$

  - p = 0.1 Q = 10
  - p = 0.01 Q = 20
  - P = 0.001 Q = 30
- Encoding: Sanger/Phred format can encode a quality score from 0 to 93
- using ASCII 33 to 126: Q + 33 - ASCII code

Source: http://en.wikipedia.org/wiki/FASTQ_format

Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
32	20	040		Space	64	40	100	@	@
33	21	041	!	!	65	41	101	A	A
34	22	042	"	"	66	42	102	B	B
35	23	043	#	#	67	43	103	C	C
36	24	044	$	$	68	44	104	D	D
37	25	045	%	%	69	45	105	E	E
38	26	046	&	&	70	46	106	F	F
39	27	047	'	'	71	47	107	G	G
40	28	050	(	(	72	48	110	H	H
41	29	051	)	)	73	49	111	I	I
42	2A	052	*	*	74	4A	112	J	J
43	2B	053	+	+	75	4B	113	K	K
44	2C	054	,	,	76	4C	114	L	L
45	2D	055	-	-	77	4D	115	M	M
46	2E	056	.	.	78	4E	116	N	N
47	2F	057	/	/	79	4F	117	O	O
48	30	060	0	0	80	50	120	P	P
49	31	061	1	1	81	51	121	Q	Q
50	32	062	2	2	82	52	122	R	R
51	33	063	3	3	83	53	123	S	S
52	34	064	4	4	84	54	124	T	T
53	35	065	5	5	85	55	125	U	U
54	36	066	6	6	86	56	126	V	V
55	37	067	7	7	87	57	127	W	W
56	38	070	8	8	88	58	130	X	X
57	39	071	9	9	89	59	131	Y	Y
58	3A	072	:	:	90	5A	132	Z	Z
59	3B	073	;	;	91	5B	133	[	[
60	3C	074	<	<	92	5C	134	\	\
61	3D	075	=	=	93	5D	135	]	]
62	3E	076	>	>	94	5E	136	^	^
63	3F	077	?	?	95	5F	137	_	_

Illumina sequence identifiers

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

Run/read QC

- First thing when receive a FASTQ file to check:
  - data amount
  - data quality
  - get first idea of whether sequenced what were aiming to sequencing

- Influences downstream e.g. checking that there are no non-genomic sequences in reads that would prevent mapping to reference

- Extremely important for feedback upstream to wet-lab

Potential QC issues

- Numbers of reads (no adaptors?, too little/much DNA loaded?, imaging problem?)

- Qualities (reagent problems?, imaging?)
- by cycle: fall with increasing cycle (Illumina)
- by read: usually bad bases distributed across reads

- Base distribution
- strong skewness at particular cycles (indicative of adaptors present)
- skewness across all cycles (may be indicative of not sequencing targeted region)

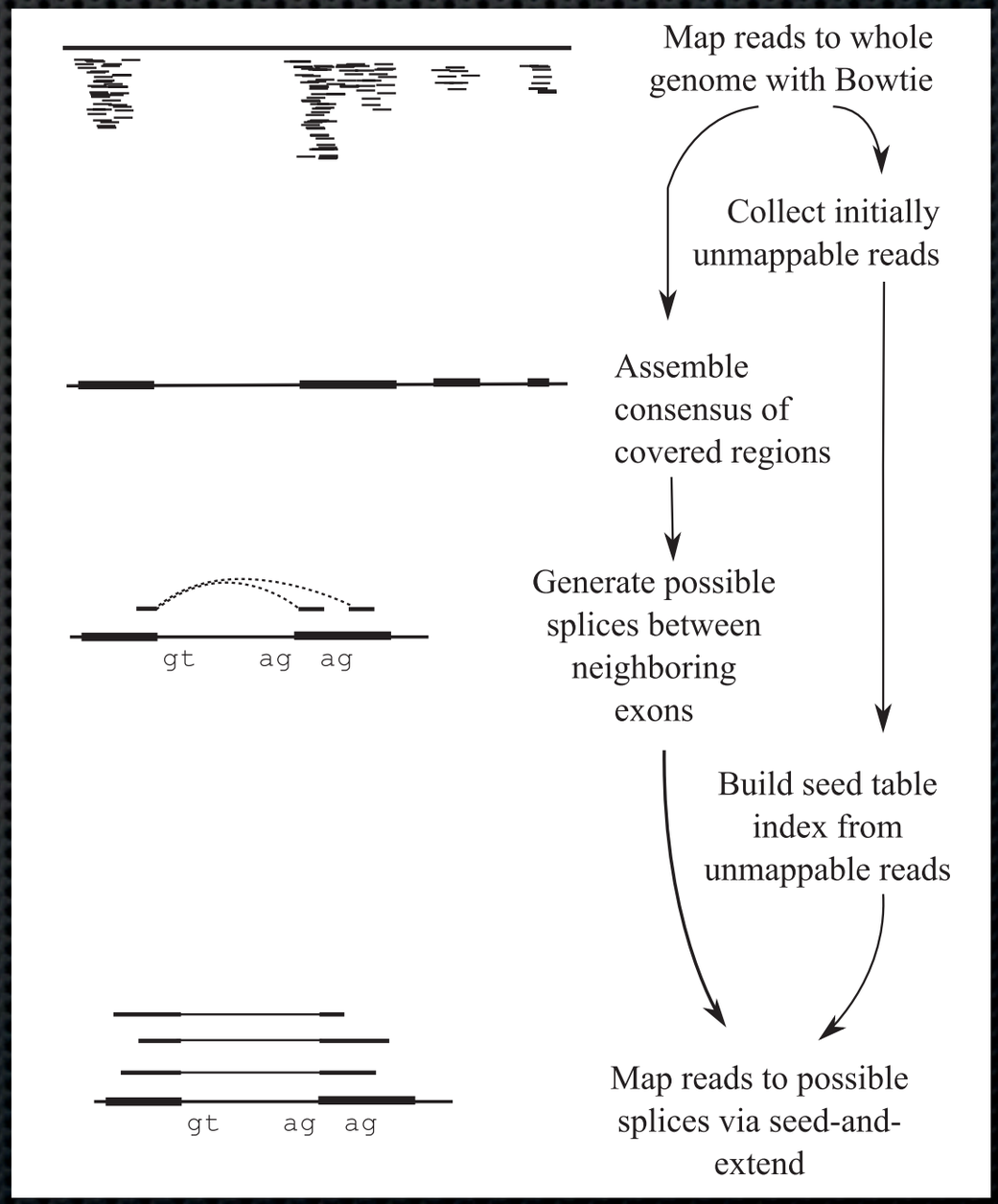
- Levels of duplication

FastQC

- http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/
- Don’t take all warnings at face value
- Has both a GUI and command line interface.
- Used by NSC to generate our standard QC report
- Example...

Aligning RNAseq reads to genome

- Bowtie - fast short read aligner
- TopHat
  - Split reads
  - Align
  - Predict splicing
  - Realign



Commands I

```
# Make directories for course

mkdir ref
mkdir reads
mkdir results

# Files
reads/sample_1_hsa21_R1.fastq, reads/sample_1_hsa21_R2.fastq
ref/chr21.fa
ref/Homo_sapiens_chr21.GRCh37.64.gtf
```

```
# Prepare bowtie index

bowtie-build chr21.fa chr21



# Run tophat to do alignment

tophat -r 50 -G ref/Homo_sapiens_chr21.GRCh37.64.gtf --no-novel-juncs \
--library-type=fr-unstranded -p 1 -o results \
ref/chr21 reads/sample_1_hsa21_R1.fastq reads/sample_1_hsa21_R2.fastq
```

Commands II

```
# View coverage

genomeCoverageBed -split -bg -ibam accepted_hits_np.bam > accepted_hits_coverage.bed

# Sort hits, and make sam file

samtools sort -n accepted_hits.bam accepted_hits_n
samtools view -h accepted_hits_n.bam > accepted_hits_n.sam

# View results: UCSC, IGV
```

http://genome.ucsc.edu/
http://www.broadinstitute.org/igv/

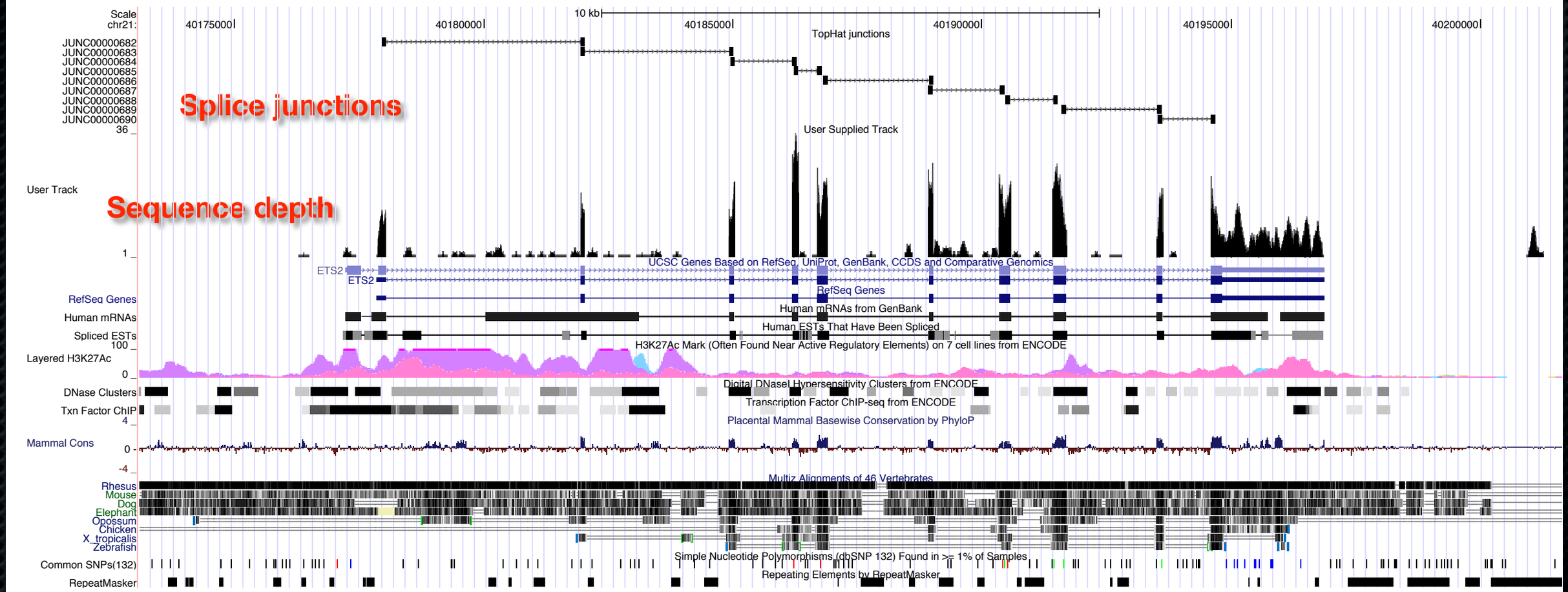
```
# Count expression levels

htseq-count -m intersection-strict --stranded=no -t exon -i gene_id \
results/accepted_hits_n.sam ref/Homo_sapiens_chr21.GRCh37.64.gtf \
> results/sample_1_hsa21.count
```

Viewing data

- UCSC
- IGV

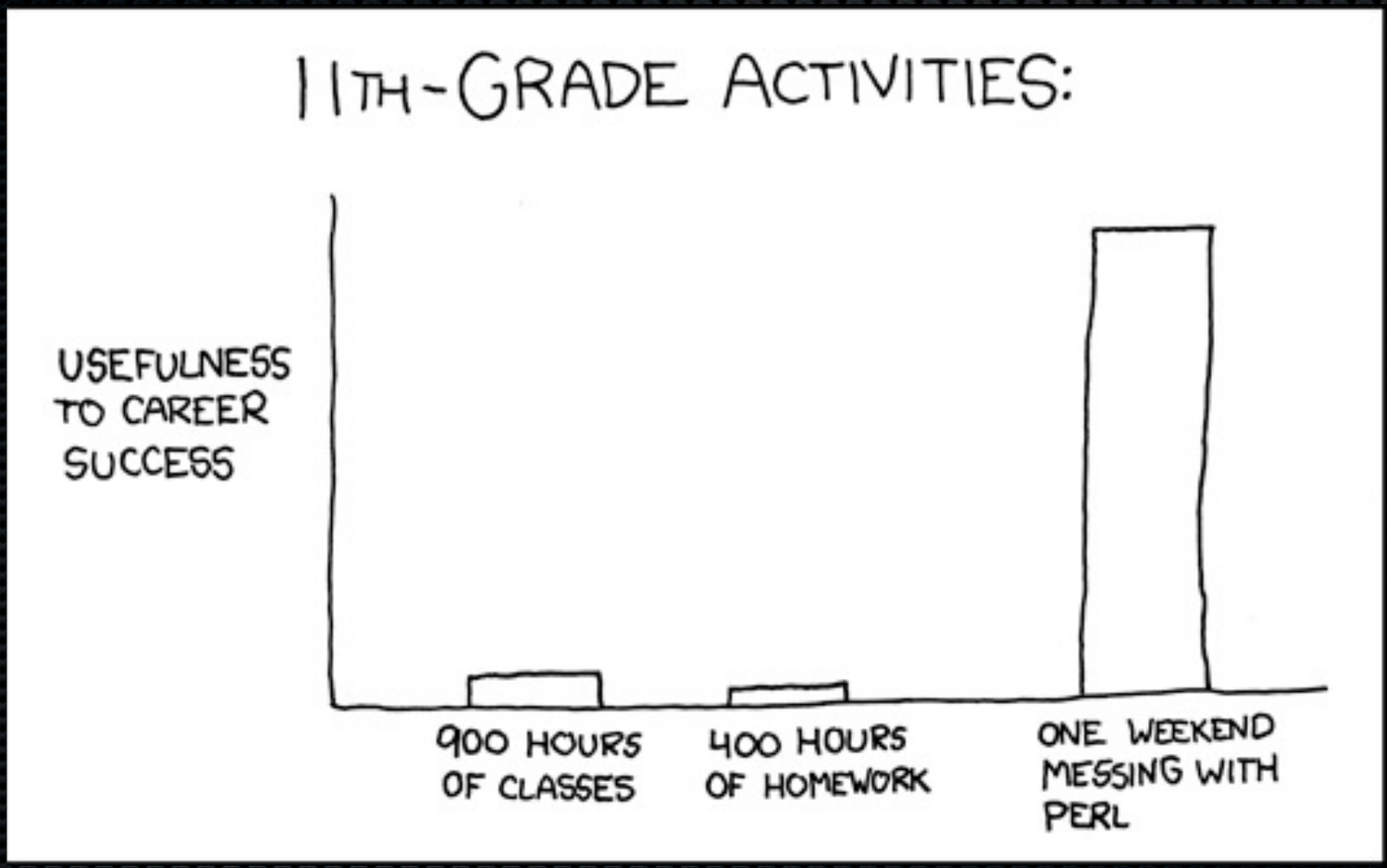
Viewing RNAseq data



Expression levels

- Count expression levles from alignment
  - HTSeq package
- Differential expression
  - R
    - DEseq (and others)
  - Cufflinks

Questions?



xkcd.com