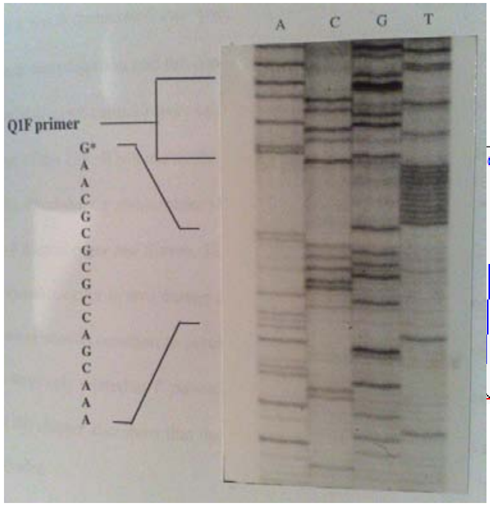
Usegalaxy.org

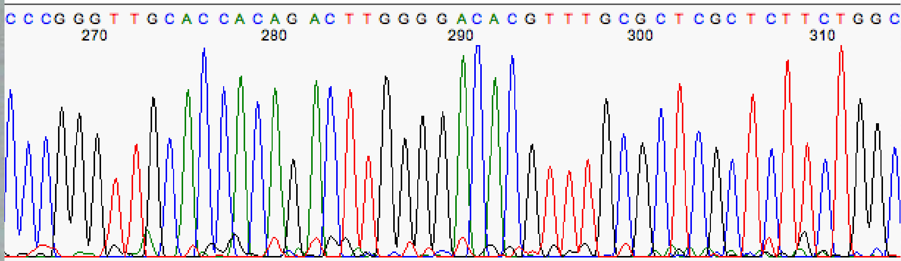
1. Upload your files to galaxy
2. FASTQ-groomer on files (your files are Illumina 1.3-1.7)

What you know: Sanger Sequencing



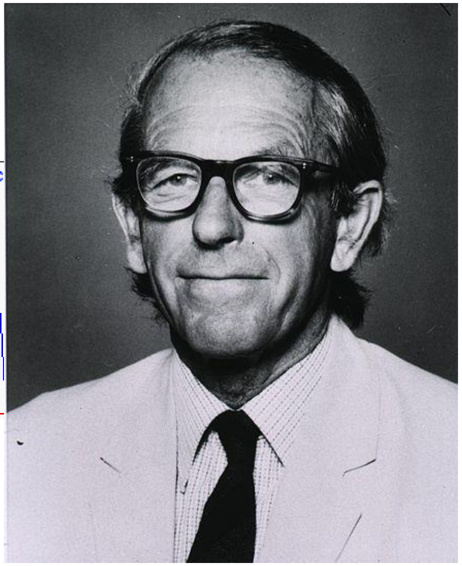
Traditional sanger / chain termination methods
Radioactive slab gels.
(70s,80s,90s)
4 separate lanes, 1 per base

Fluorescent dye-terminator sequencing
(90s – today)



How does this work?
What is the data like? How long?
How much money (per base)?

Length: 500-1000bp
Data is bad at beginning and end of read
~cents per base



Fred Sanger
Nobel laureate 1958,1980

“Next Gen Sequencing”

- All the sequences technologies since Sanger sequencing.
- Many sequencing technologies, but one is hugely dominant.

https://www.illumina.com/

http://allseq.com/knowledgebank
For modern-ish review of other sequencing platforms

*we used Illumina sequencing for our RNA-Seq experiment

How does Illumina sequencing work?

- Massively parallel sequencing of short reads – 40bp-300bp

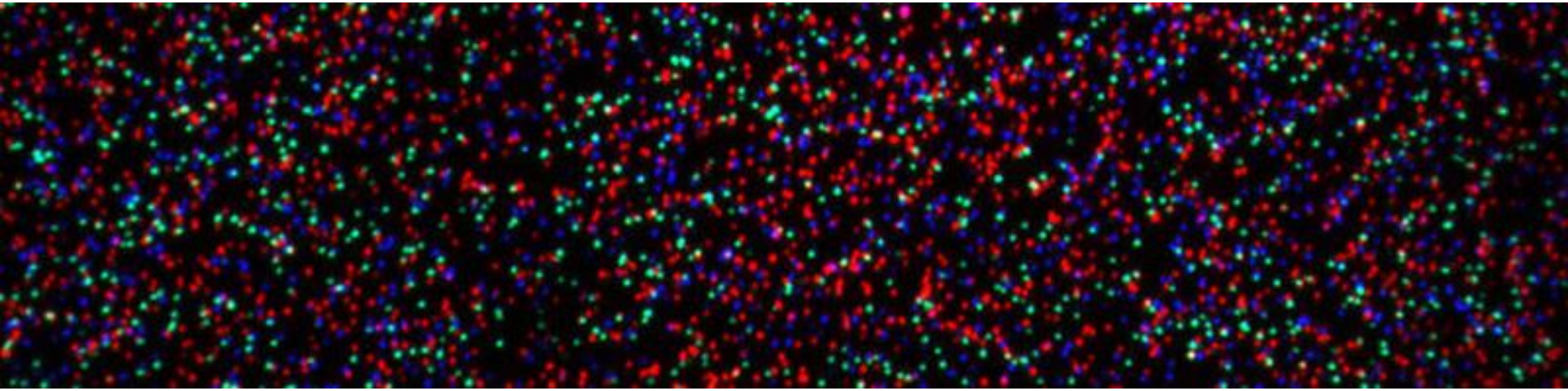


Image of Illumina HiSeq flowcell
Every spot (cluster) on the flowcell is a unique sequencing reaction.
Each spot is 1um or less.

A bit outdated, but detailed resource on Illumina sequencing:
https://www.broadinstitute.org/scientific-community/science/platforms/genome-sequencing/broadillumina-genome-analyzer-boot-camp

Illumina HiSeq2500 flowcell photograph removed due to copyright restrictions.

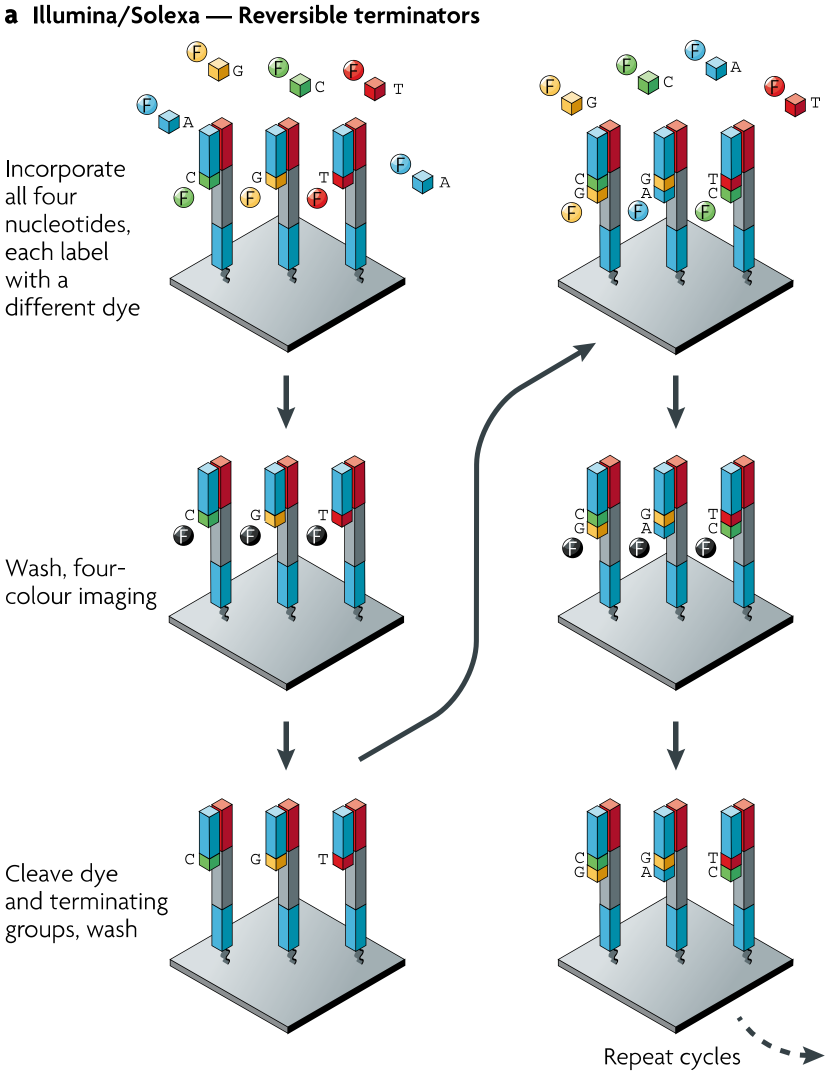
Please see: http://www.cisd.ethz.ch/software/openBIS/Deep_Sequencing;
Illumina Flow Cell v3: http://www.cisd.ethz.ch/software/openBIS/Deep_Sequencing/Illumina_Flow_Cell_v3.jpg?hires

Illumina HiSeq2500 flowcell

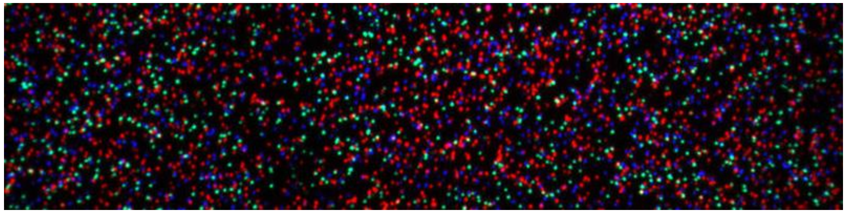
Sequencing happens on a flowcell, you buy sequencing capacity by lane (the flowcell above has 8)

Each lane gives you 200M + reads, and costs upwards of $1000

Illumina / Solexa Sequencing



a Illumina/Solexa — Reversible terminators
F
G
C
T
A
Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles



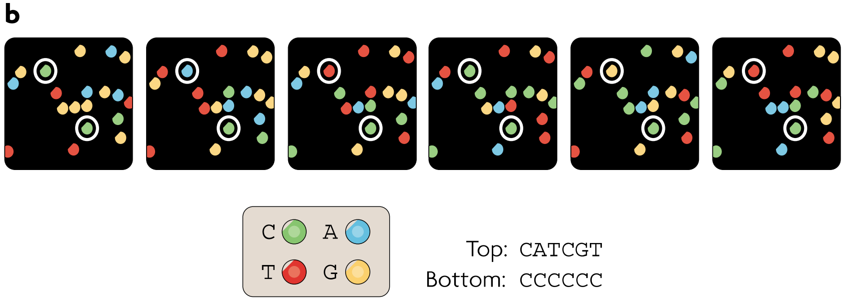
1 base per cycle.

~1 hour per cycle (20mins chemistry, 40mins imaging)

~1000 molecules per cluster

<1um per cluster

*varies somewhat depending on Illumina instrument



Reprinted by permission from Macmillan Publishers Ltd: *Nature Review Genetics*.
Source: Metzker, M. L. "Sequencing Technologies — the Next Generation." *Nature Reviews Genetics* 11 (2010): 31–46. © 2010.

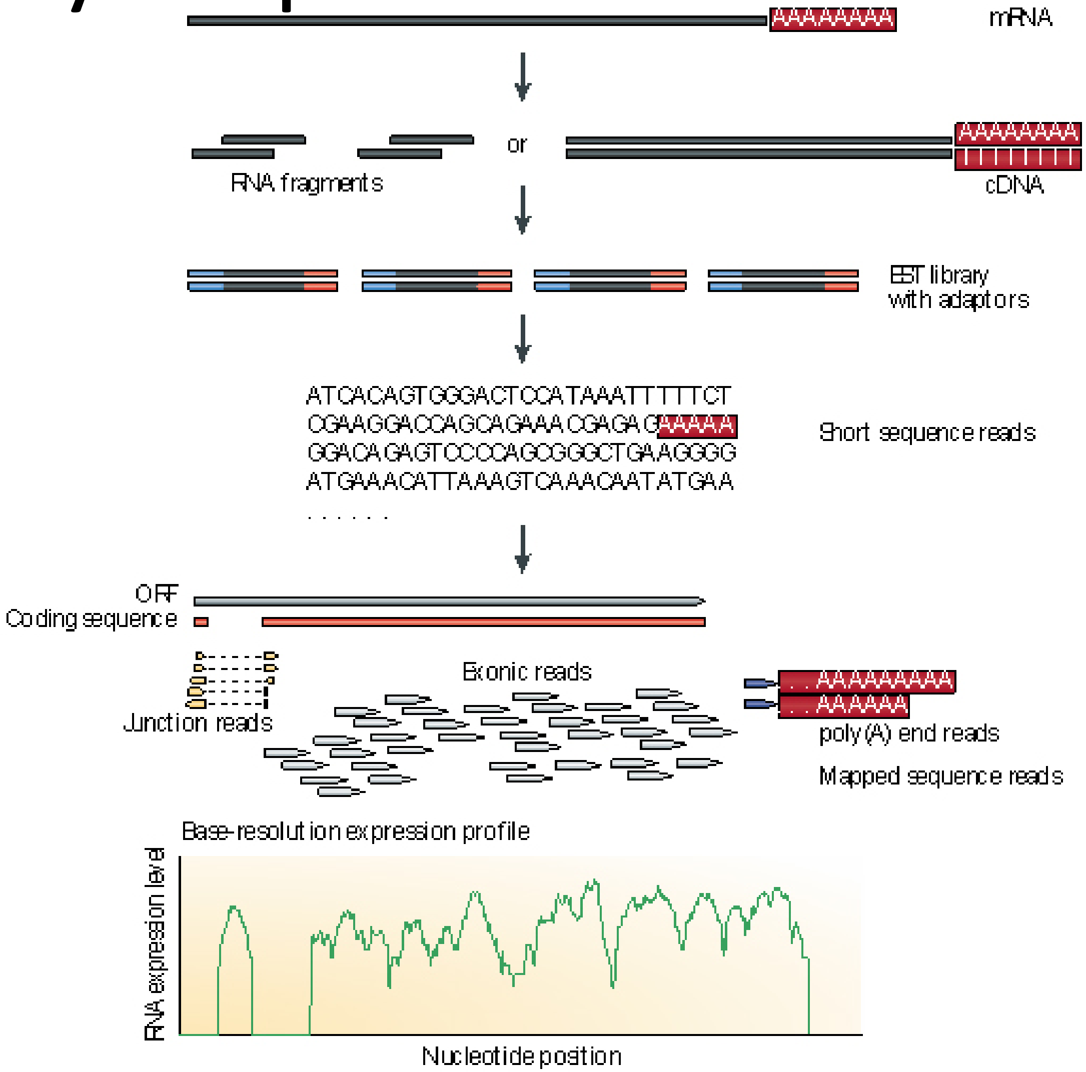
Movie on Illumina sequencing

https://www.youtube.com/watch?v=womKfikWlxM

How to get DNA suitable for sequencing? Library Prep

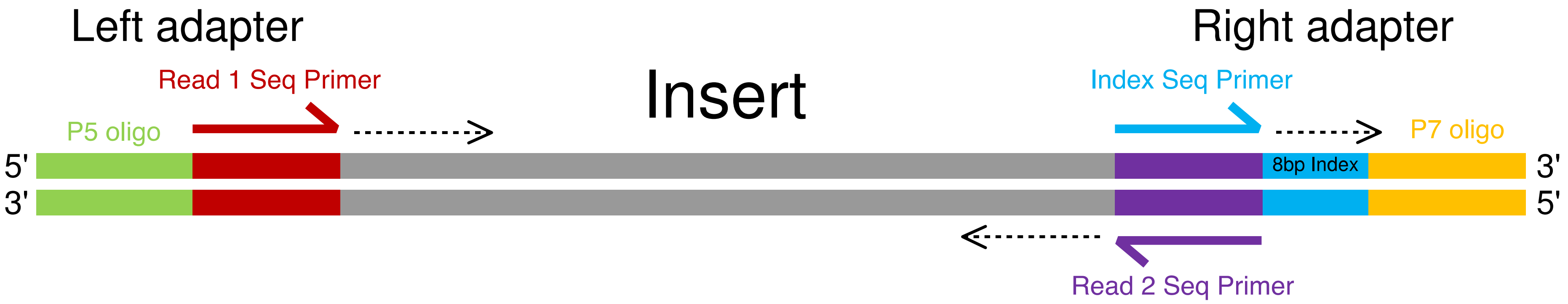
Depending on what type of library prep you do, can have totally different types of experiments (DNA RNA, methylation, ribosome profiling)

But ultimately, everything looks the same when it is converted to the final flowcell ready fragment: a dsDNA fragment with asymmetric adaptors



Reprinted by permission from Macmillan Publishers Ltd: *Nature Review Genetics*. Source: Wang, Z., Gerstein, M., et al. "RNA-Seq: A Rvolutionary Tool for Transcriptomics." *Nature Reviews Genetics* 10 (2009): 57–63. © 2009.

Anatomy of an Illumina sequencing fragment



- P5 & P7 oligos bind fragment to flowcell

Can have single ended sequencing. Only Read 1 (plus index read if multiplexing/barcoding)

- Paired end sequencing. Read 1&2 (plus index read)
- Index read gives you the multiplexing / barcoding that lets you put multiple samples onto the same “lane”

Whats the data like - FASTQ

```
@WIGTC-HISEQ:4:1107:1232:1988#TTAGGC/1;0
TGAAACTATTTTCACCCAGACAGATGCCATATTTGAATTC
+WIGTC-HISEQ:4:1107:1232:1988#TTAGGC/1;0
]\Z``RS\_baaS^__bPR_J^V\\[VbR[\[_aSI^V^B
@WIGTC-HISEQ:4:1107:1117:1992#TTAGGC/1;1
GTGGGGATGTTCGACTGGATTCATGGCAACTCCTCTGACA
+WIGTC-HISEQ:4:1107:1117:1992#TTAGGC/1;1
___eeeecgbeeefghffhiffiiiifhhhbghhhhfhfb
@WIGTC-HISEQ:4:1107:1647:1958#TTAGGC/1;1
CTGTAATTGGCTTCCGACGACTTGGAATGATAGCATCGAA
+WIGTC-HISEQ:4:1107:1647:1958#TTAGGC/1;1
\__S`cdeffeggfghfihhihiiifghbffhihfhhfgh
@WIGTC-HISEQ:4:1107:1629:1991#TTAGGC/1;1
GGCAACAGCGGTCTTGGAGACGGCAGCAGCGGTACCTCCT
+WIGTC-HISEQ:4:1107:1629:1991#TTAGGC/1;1
__bJ`cdeffceghhhihiffdghgghhihfdUedgibg]
@WIGTC-HISEQ:4:1107:1516:1994#TTAGGC/1;1
GTCCATCGAGCCATGGGGTCTTGACTGTGGTGATGAAGAA
+WIGTC-HISEQ:4:1107:1516:1994#TTAGGC/1;1
_abeeeeeggfggiiiiicfhihihiihhiegbgffhhhi
@WIGTC-HISEQ:4:1107:2130:1974#TTAGGC/1;1
GTCCGTCGTTTCCTGGTGCTCCTGGTTGTCCATCAGCTCC
+WIGTC-HISEQ:4:1107:2130:1974#TTAGGC/1;1
bb_ceeeegfggghiiffgihhhfighhihfhfiihhiii
@WIGTC-HISEQ:4:1107:2078:1977#TTAGGC/1;1
ATGGAGTTGTCTCAAACGTCTGCACGATCTCCTTCACGAT
+WIGTC-HISEQ:4:1107:2078:1977#TTAGGC/1;1
bbbeeedeggggghiiiifgiiiiiihiiiiiiiiiihhh
```

Orange -> Sequence data

Line 1: Read identifying metadata

@WIGTC-HISEQ -> Instrument name

4 -> Flowcell lane #4

1107:1117:1992 -> X,Y and tile #

#TTAGGC -> Barcode

/1 -> Forward read (/2 is reverse read)

Line 2:

ATCG… The actual nucleotide sequence data

Blue -> Quality data

Line 3: Read identifying metadata (same as line 1)

Line 4: Quality data. 1 character per base.

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

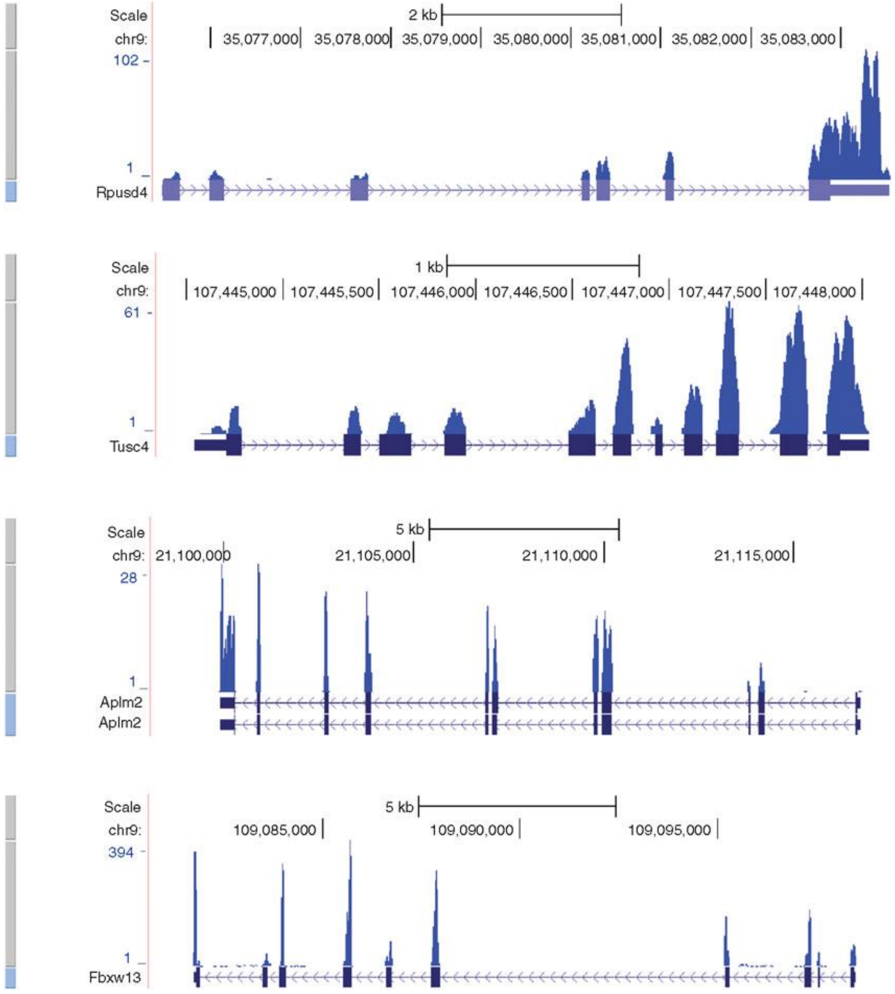
Quality data can be in different encodings!

Check encoding of a given FASTQ file with the FastQC program (on galaxy and standalone)

FASTQ groomer on Galaxy can convert quality score encodings

RNA-seq in a nutshell

1. Start with raw reads: FASTQ
2. Some quality filtering, dropping bad reads, removing or trimming reads include sequence from adaptors
3. "Map" to a reference genome using a splice-junction-aware mapper/sequence aligner. TopHat is commonly used. Produces .bam file.
4. Count how many reads map to a given gene. That count is proportional to the abundance of that transcript in the original sample. Cuffdiff of the Cufflinks suite can do this. Produces a spreadsheet.



Tools on Galaxy you should know about

- *Reminder that Galaxy is mainly a wrapper around existing command-line programs.
- NGS: QC and manipulation -> FastQC
  - Give it a FASTQ, bam, or sam file, and it computes summary statistics to help evaluate if your sequencing run worked, or just how the reads along the process look globally.
- NGS: QC and manipulation -> FASTQ Groomer
  - Converts FASTQ quality encoding to something Galaxy likes ("Sanger" format). FastQC can tell you what the quality encoding is for a given FASTQ file.

More Galaxy tools

- Bowtie & Bowtie2 -> Genome mapper for fragments
- TopHat -> Splice junction aware mapper. Use for mapping RNA reads to a genome
- Cufflinks suite -> Suite of tools for doing expression analysis (and other things) on RNA-Seq data (Trapnell et al. 2012) - http://cole-trapnell-lab.github.io/cufflinks/

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protocols 7, 562–78.