Intro to Bootcamp

What we will cover

- TSCC (Triton Shared Computing Cluster) and coding in UNIX
- RNA-seq pipeline using open-source data
- Visualize and interpreting Data in Python

Big-Picture Objectives

Learn how to code in UNIX and Python

Read documentation for software packages

Use Python to analyze and present data

Answer your own research questions with computational techniques

Week Schedule - check website

Monday 9-1pm

Tuesday 9-11am

Wednesday 9-1pm

Thursday 11am - 1pm

Wednesday (Oct 2) 9-1pm

Presentations this day

Group Project

Groups of 5

Find your own published dataset starting with fastq files

Run data and create plots

Conclude what your results tell us about biology

Office hours

All office hours will be held on slack!

This way everyone can see and help with questions.

Request to join here:

https://join.slack.com/t/bmsbioinforma-kn18317/shared_invite/enQtNzQ2ODcxOTAyNTE2LTY2OTMzMjZhMmU0ZGY3NjkzMTAyMTc4ZTFhOTQ2ZWM2ZGU3YjU5NjQ5ZDliOTk1MDkzZDEzODQ0MTk1MDBhNWQ

Class materials on Github

Bioinformatics Bootcamp 2019

This is the page for the BMS BIOM200 Bootcamp 2019

View the Project on GitHub
macatbu/biom200_bootcamp_2019

Class Schedule

If notebooks won't load view them in nbviewer

Pre-Class

- Set up your TSCC account
  - Intructions Mac Windows
  - Practice UNIX here
  - Download Anaconda using these instructions

Sep 23 (Mon) 9am – 1pm

Intro to Sequencing, Unix, Data Download, Fastqc

- Notebooks
  - 1 TSCC login and Downloads
  - 2 Intro to bash commands
  - 3 Data downloads and quality checks

Website view:
https://macatbu.github.io/biom200_bootcamp_2019/

Github view:
https://github.com/macatbu/biom200_bootcamp_2019

What is bioinformatics, and why do you need it?

- Intersection of computer science and biology
- Analyze large omics datasets
- Useful for many jobs/research fields

Bootcamp Tips

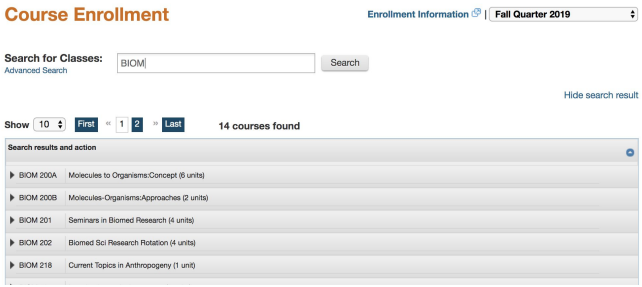
- Learning a new language is hard, be patient with yourself.

- We are focusing on learning concepts, not finishing the activity fast, so don't stress if you fall a little behind.

- This is just a tiny taste of bioinformatics, and if you are interested to learn more we can point you to more resources.

What if i want to learn more?

- Classes at UCSD



- Online resources like CodeAcademy and Datacamp

- Best way to learn is to practice/ have a relevant project!

Today’s agenda

- Login to TSCC
- Practice in UNIX
- Lecture on RNA-seq

BREAK

- Download Data
- Fastqc

Today’s objectives

- Login to TSCC
- Be able to move around in UNIX
- Be able to download data from ENCODE
- Be able to check data quality

TSCC

2019

TSCC

- **Triton Shared Computing Cluster (TSCC)**
  - Housed within the San Diego Supercomputer Center (SDSC)
  - Other clusters include *Comet* and *Gordon*
  - More information about TSCC:
    - http://www.sdsc.edu/support/user_guides/tscc-quick-start.html

TSCC Structure: Branches from Root

- **Scratch** – Faster file processing, runs on a parallel file system
  - Each user is given their own directory in scratch
  - Essentially unlimited space, but untouched files get purged after 3 months
  - https://en.wikipedia.org/wiki/Lustre_(file_system)

- **Home –** Permanent storage for each user, relatively small space, runs on network file system
  - Each user has their own home folder
  - Very minimal space
  - https://en.wikipedia.org/wiki/Network_File_System

- **Projects –** Labs purchase storage space for permanent files that are shared among members of the lab
  - Each lab has their own projects folder this is space for more permanent storage (e.g. ps-yeolab)

What is UNIX?

- Operating system with command line interface

- Language we will use to run most of our RNA-seq analysis

Intro to sequencing

BMS Bootcamp 2019

Outline

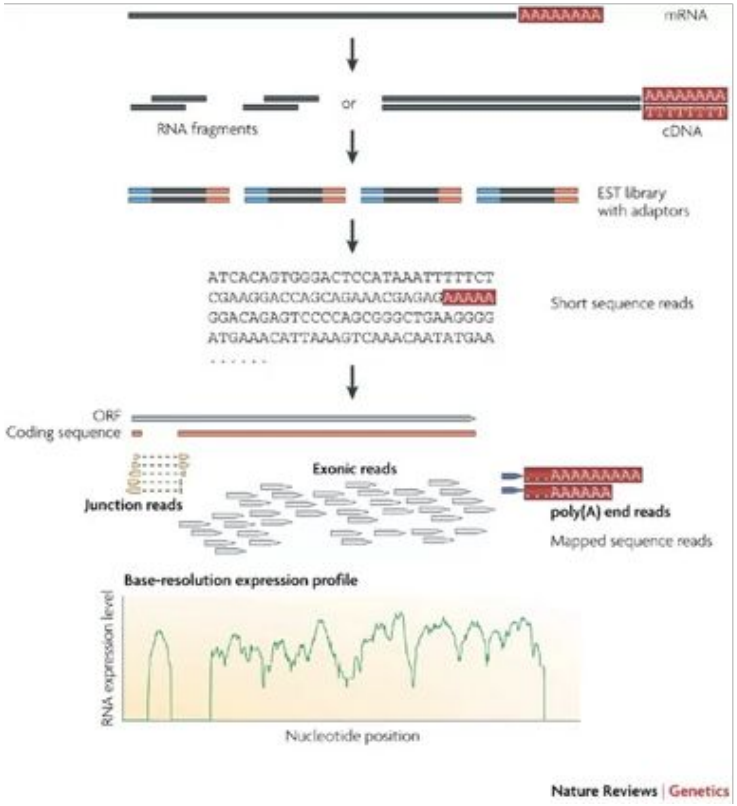
- Intro to RNA-seq
- Processing pipeline and workflow

Differential Gene Expression (DGE) by RNA-Seq

- How does a certain condition affect which genes are transcribed into RNA and translated into protein?

- DGE by RNA-Seq: count up the mRNA molecules in each condition, look at genes where abundance of mRNA copies changes

RNA-seq Pipeline



RNA Isolation and Fragmentation

Library Preparation

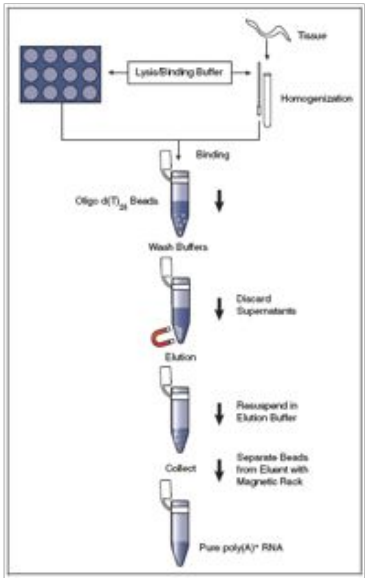
Sequencing

Genomic Alignment of Sequencing Reads

Step 1: What RNA do you want to profile?

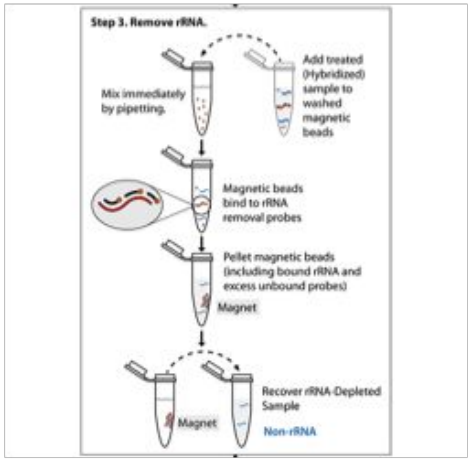
mRNA only -> PolyA selection

(All mRNAs are polyadenylated at the 3' end – can use $d(T)_{25}$ beads to select)



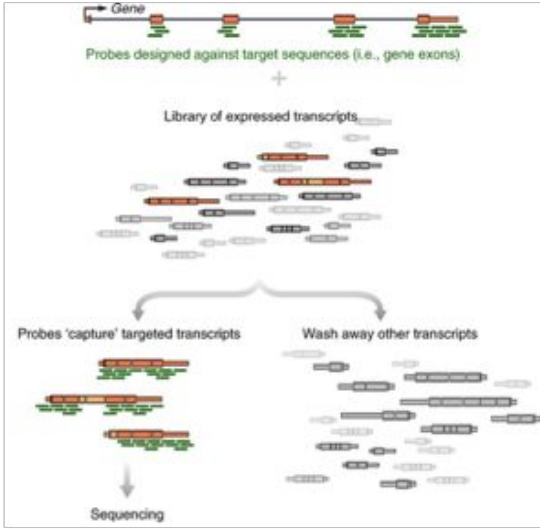
rRNA-depleted RNAs -> rRNA negative selection

(Probes matching rRNA are used to pull-out rRNA molecules)



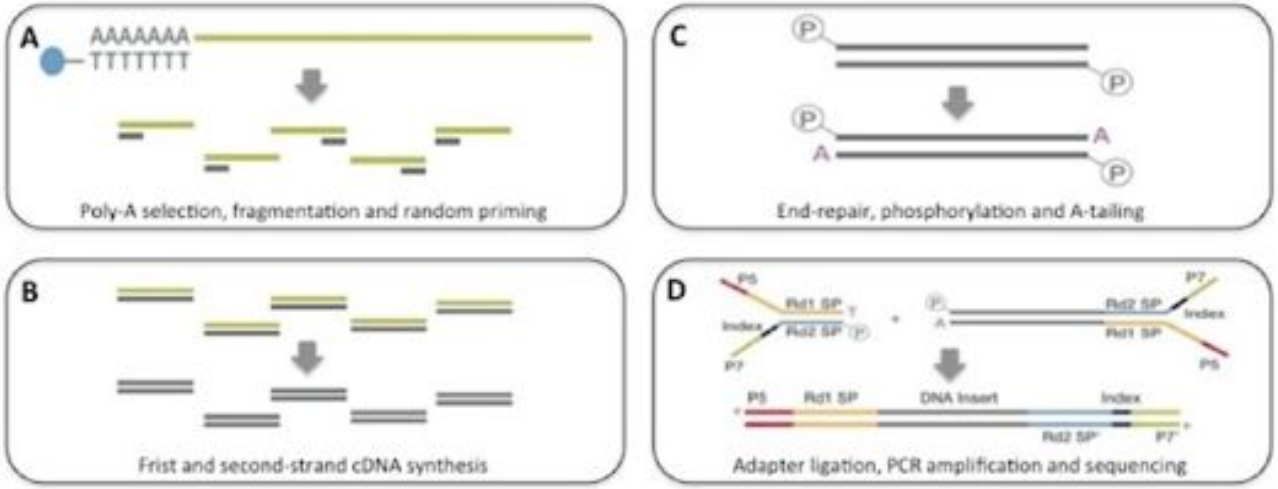
(Other methods – hybridize targeted DNA oligos + RNAseH treat)

Specific RNAs -> targeted enrichment

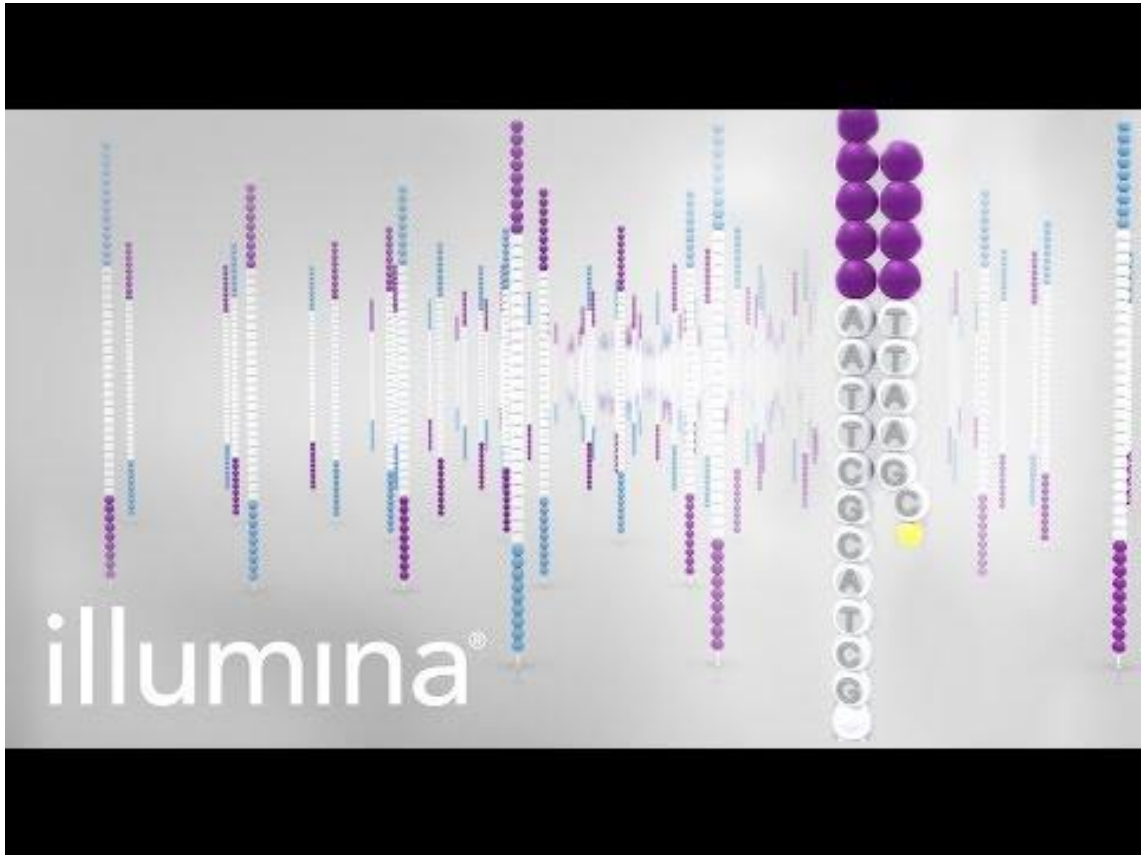


General Library Preparation

Illumina Tru-Seq RNA-seq protocol

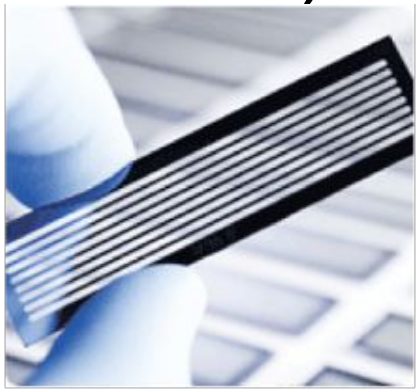


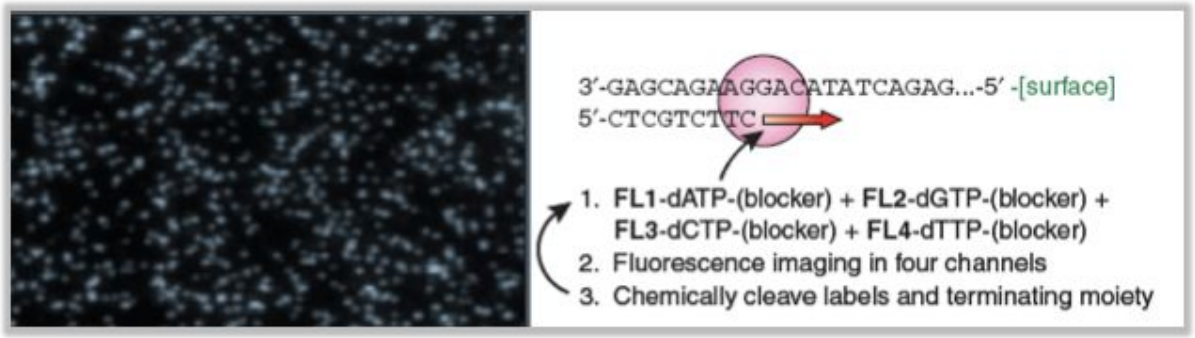
Library prep begins from 100ng-1ug of Total RNA which is poly-A selected (A) with magnetic beads. Double-stranded cDNA (B) is phosphorylated and A-tailed (C) ready for adapter ligation. The library is PCR amplified (D) ready for clustering and sequencing.



https://www.youtube.com/watch?v=fCd6B5HRaZ8

Sequencing by synthesis: HiSeq 2500 (Illumina)



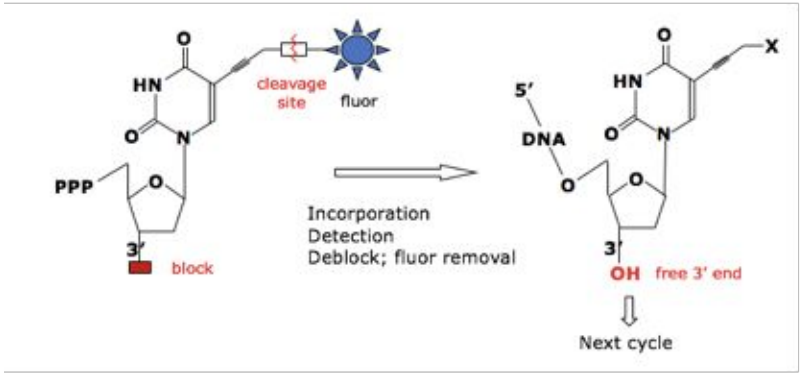


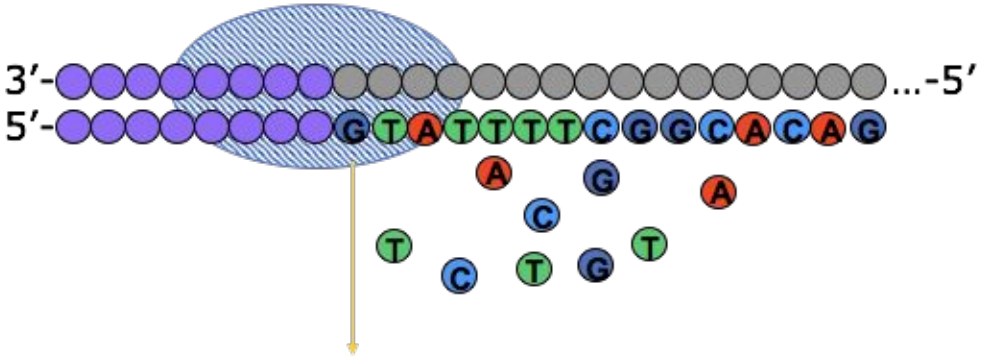
Shendure & Lee, Nat. Biotech. 2008

- **Can do 50bp to 250bp single-end or paired-end reads**
- **~300 million reads per lane x (2 or 8) lanes**
- **4-8 days run time**
- **200 billion bp output each run**

Sequencing by Synthesis

Reversible Terminator Chemistry





Cycle 1: Add sequencing reagents

First base incorporated

Remove unincorporated bases

Detect signal

Cycle 2-n: Add sequencing reagents and repeat

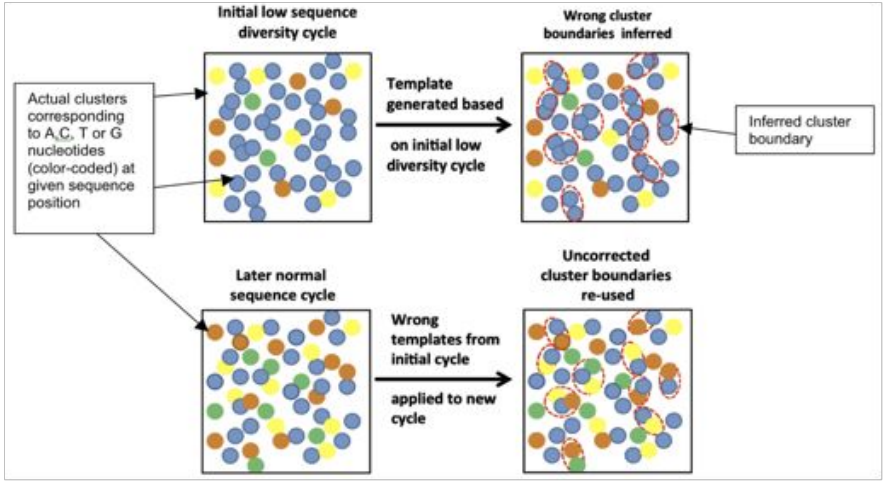
Key considerations

"Cluster Density" = how many clusters are there per mm^2

- If too high, hard to properly draw cluster boundaries

"Library Complexity" = how diverse are the sequences?

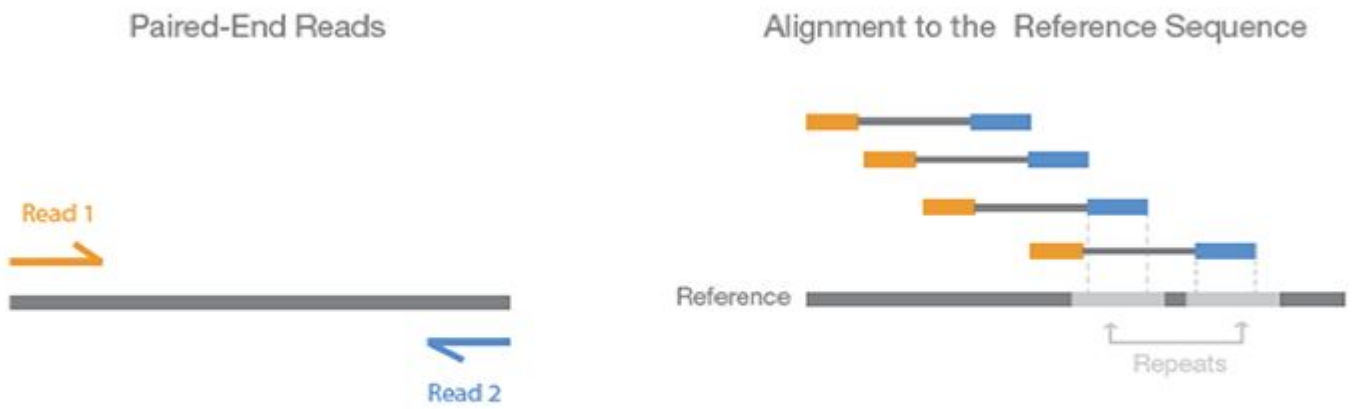
- Illumina identifies clusters in the first 5 cycles – if those 5 cycles are identical for nearby clusters, the software doesn't know to split them into two



Mitra, A. *et al.* Plos ONE (2015)

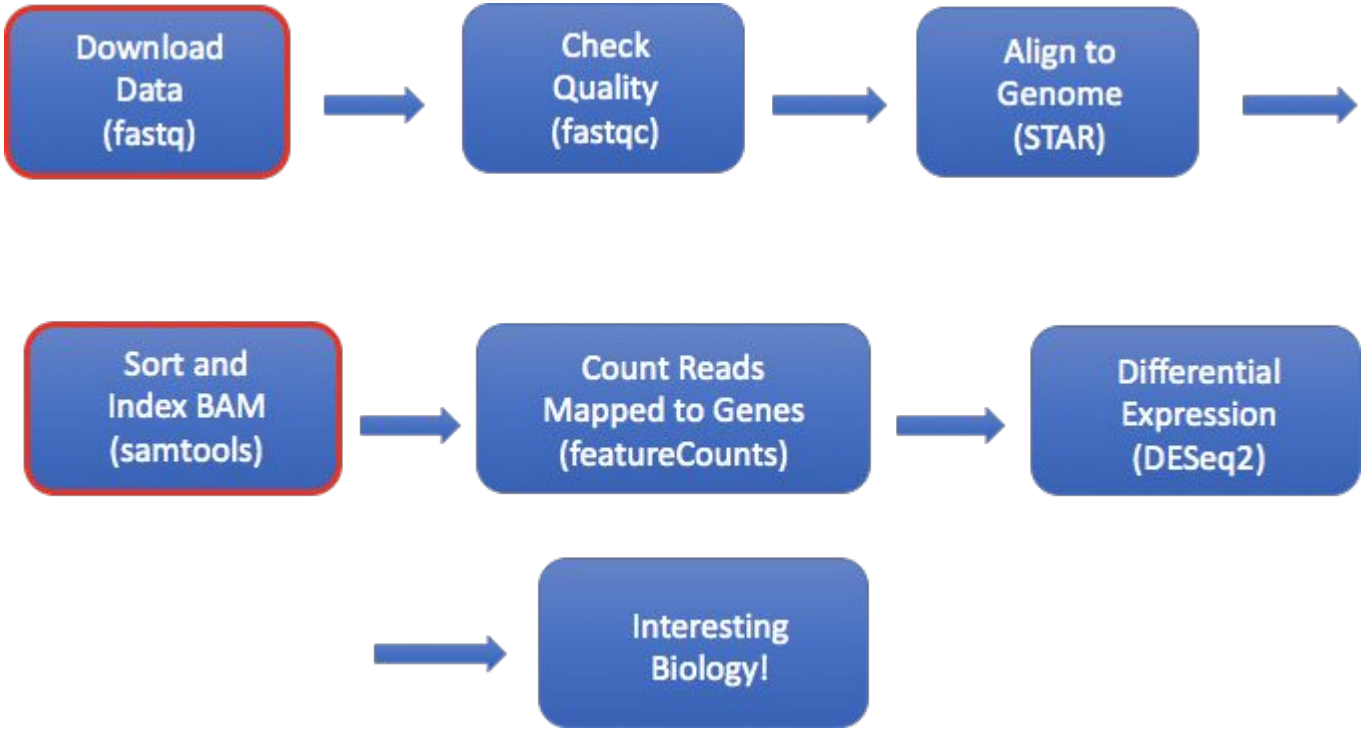
Paired-end sequencing results in more accurate sequence alignment

Figure 4. Paired-End Sequencing and Alignment



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

RNA-Seq Analysis Overview:



Fastq File

- First line is the information about the location of the read and specific sequencing machine used:

```
@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos>
<read>:<is filtered>:<control number>:<index sequence>
```

- Second line is the nucleotide sequence called
- Third line is "+" and can optionally be followed by a repeat of the filename in line 1
- Fourth line contains the quality score as determined by the sequencer

```
@NS500672:54:HL775BGXX:1:11101:22716:1042 1:N:0:CCCCGG
CCGCCNATGCCCATGCCACAGTTGTTGAGCTTGAGTTCCTGCAGGGTGAAGCAGGCTGAGCTCTTGAG
CAGGGCCTCGAA
+
AAAAA#EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEE
EEEE
```

Fastq File – Phred Quality Score

- Quality scores report the probability that the base call is incorrect

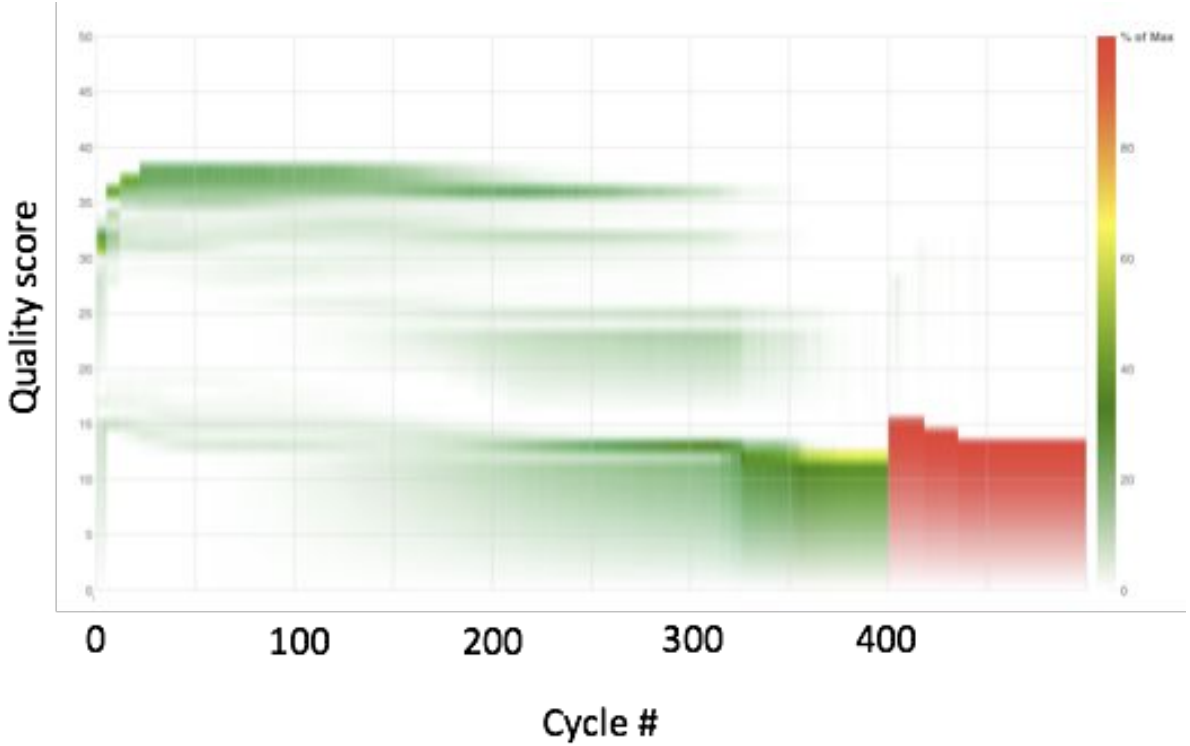
$$Q = -10 \log_{10} P$$

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

- Field standard is to accept bases with quality >20

Illumina sequencing – great for read #, not great for read length



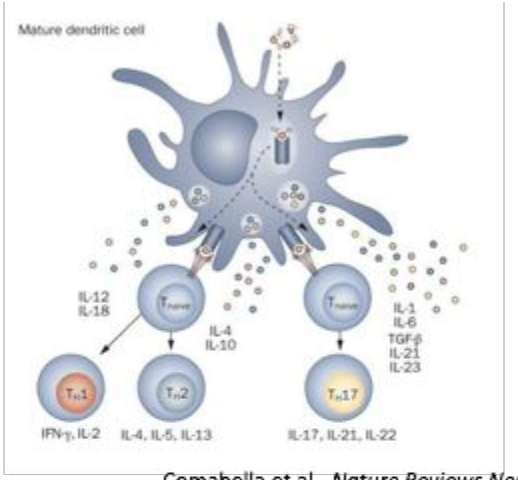
Dataset: Primary dendritic cells following exotoxin activation

- Dendritic cells: antigen-presenting cells
  - Phagocytose pieces of pathogens in surrounding environment
  - Immature: sample pathogens and degrade proteins into presentable peptides
  - Mature: localize to lymph nodes and activate T cells through antigen presentation

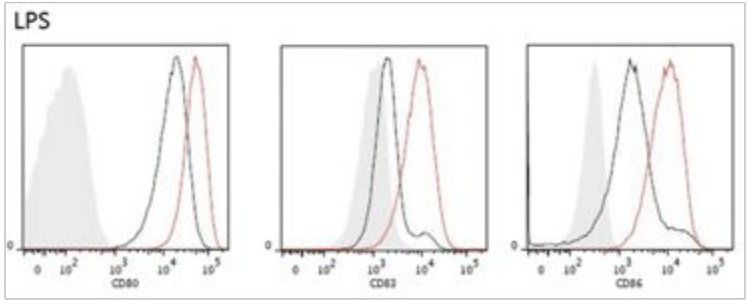
- Exposure to bacterial, viral, or fungal pathogen evokes specific transcriptional program through interaction will cell receptors (e.g. TLRs)

- Serves as link between innate and adaptive immunity

- Goal: look at immune response gene patterns following exposure to pathogen component lipopolysaccharide (LPS) (bacterial stimulation) throughout maturation



Comabella et al., *Nature Reviews Neurology* 2010



Street et al., *Journal of Orthopaedic Surgery and Research* 2015

Dataset: Primary dendritic cells following exotoxin activation

- Dendritic cells: antigen-presenting cells
  - Phagocytose pieces of pathogens in surrounding environment
  - Immature: sample pathogens and degrade proteins into presentable peptides
  - Mature: localize to lymph nodes and activate T cells through antigen presentation

- Exposure to bacterial, viral, or fungal pathogen evokes specific transcriptional program through interaction will cell receptors (e.g. TLRs)

- Serves as link between innate and adaptive immunity

- Goal: look at immune response gene patterns following exposure to pathogen component lipopolysaccharide (LPS) (bacterial stimulation) throughout maturation

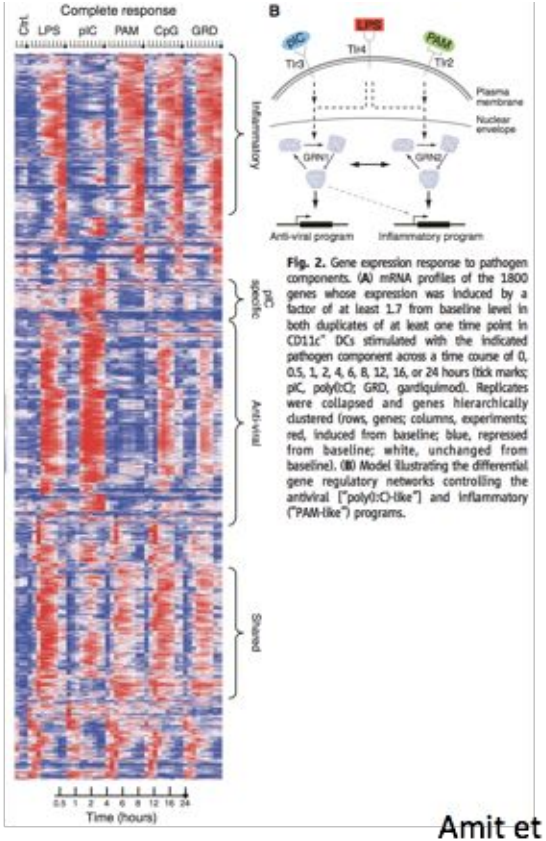


Fig. 2. Gene expression response to pathogen components. (**A**) mRNA profiles of the 1800 genes whose expression was induced by a factor of at least 1.7 from baseline level in both duplicates of at least one time point in CD11c$^+$ DCs stimulated with the indicated pathogen component across a time course of 0, 0.5, 1, 2, 4, 6, 8, 12, 16, or 24 hours (tick marks; pIC, poly(I:C); GRD, gardiquimod). Replicates were collapsed and genes hierarchically clustered (rows, genes; columns, experiments; red, induced from baseline; blue, repressed from baseline; white, unchanged from baseline). (**B**) Model illustrating the differential gene regulatory networks controlling the antiviral ("poly(I:C)-like") and inflammatory ("PAM-like") programs.

Amit et al., *Science* 2009

Dataset: Primary dendritic cells following exotoxin activation

- Samples: bone marrow-derived dendritic cells (differentiated in vitro from mouse bone marrow)
- Exposed to 100 ng/ml LPS for 0 to hours
- 4-hour time point: cells characterized to have pronounced migratory ability and poor antigen uptake ability
- Compare initial and long-term/terminal gene signatures of maturing DCs

https://www.encodeproject.org/rna-seq/long-rnas/

- Each replicate should have 30 million aligned reads, although older projects aimed for 20 million reads.
- Replicate concordance: the gene level quantification should have a Spearman correlation of >0.9 between isogenic replicates and >0.8 between anisogenic replicates (i.e. replicates from different donors).

RNA-Seq Analysis Overview: