Welcome to the RNASeq Workshop/Course

Instructors

ENTOMOLOGY &
PLANT PATHOLOGY
UTIA INSTITUTE OF AGRICULTURE
THE UNIVERSITY OF TENNESSEE

- Meg Staton
- Abdullah Almsaeed, mobile and web developer
- Matt Huff, researcher
- Fang Liu, PhD student in EPP
- Jiali Yu, PhD student in GST

A UT/ORNL PARTNERSHIP
NATIONAL INSTITUTE FOR COMPUTATIONAL SCIENCES
NICS

- Nick Csercsevits, Joint Institute for Computational Sciences (JICS)

Learning Objectives

- Students will be able to apply basic bioinformatic theory and tools to analyze transcriptome datasets
- Students will be able to effectively communicate and critically assess the application of bioinformatic tools to transcriptome data
- Students will have basic competence in the UNIX shell and usage of bioinformatic tools from the command line

Syllabus Review

Course website:

https://github.com/statonlab/rnaseq_workshop/wiki

- Office hours 1:30 - 2:30
- Additional Linux command line practice today, if anyone requests it

Other Stuff

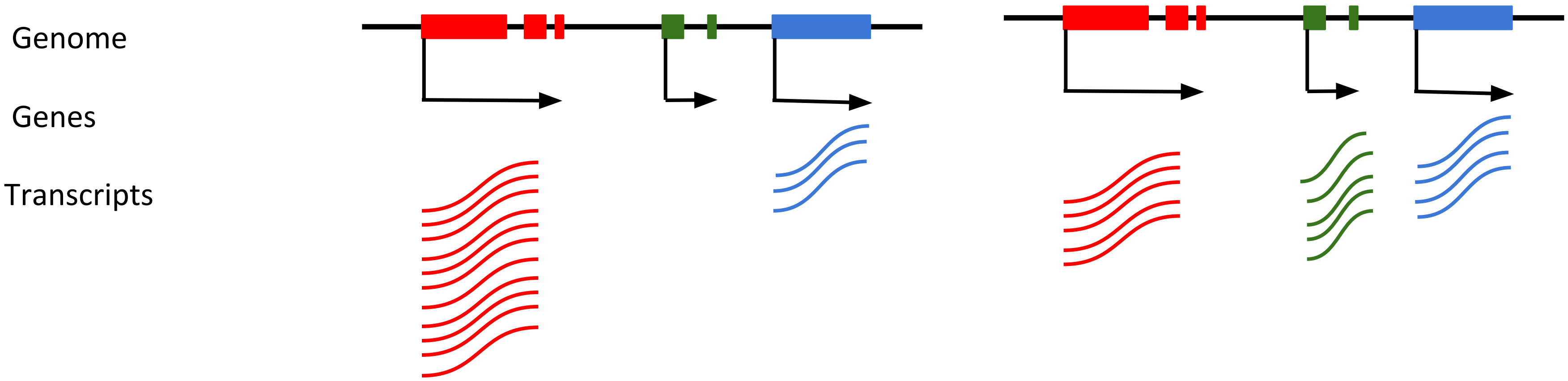
- Help is available!
- Ask questions, have discussion with your neighbors
- Use your stickies

Basics of RNASeq

Outline

- The Big Picture
- Illumina platform
- Fasta format
- Fastq format
- RNASeq
- Data set for the class

The Big Picture

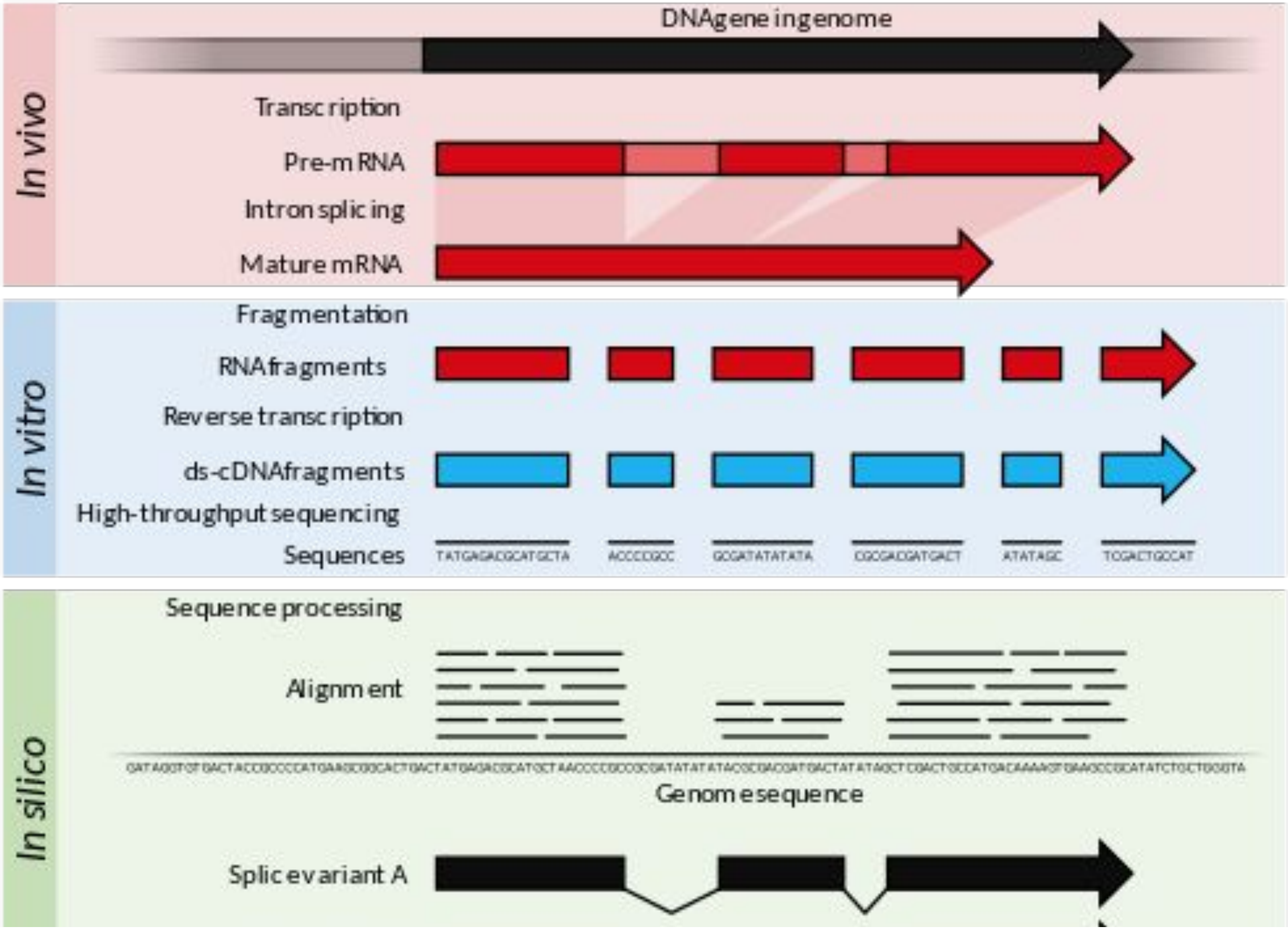


Genome

Genes

Transcripts

The Big Picture



illumina

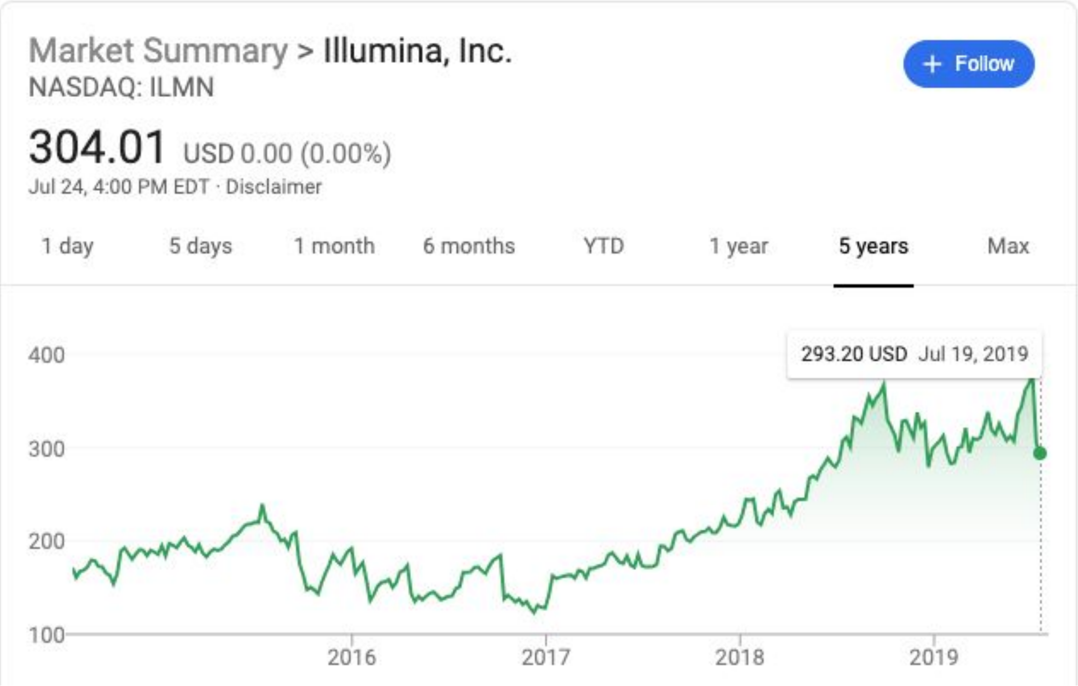
Illumina Sequencing Technology

$3.3 billion in revenue in 2018

Market share (estimated):
90% in 2016
75% in 2018

Why are they so popular?

- Low price
- High throughput
- High base calling fidelity
- Paired end sequencing

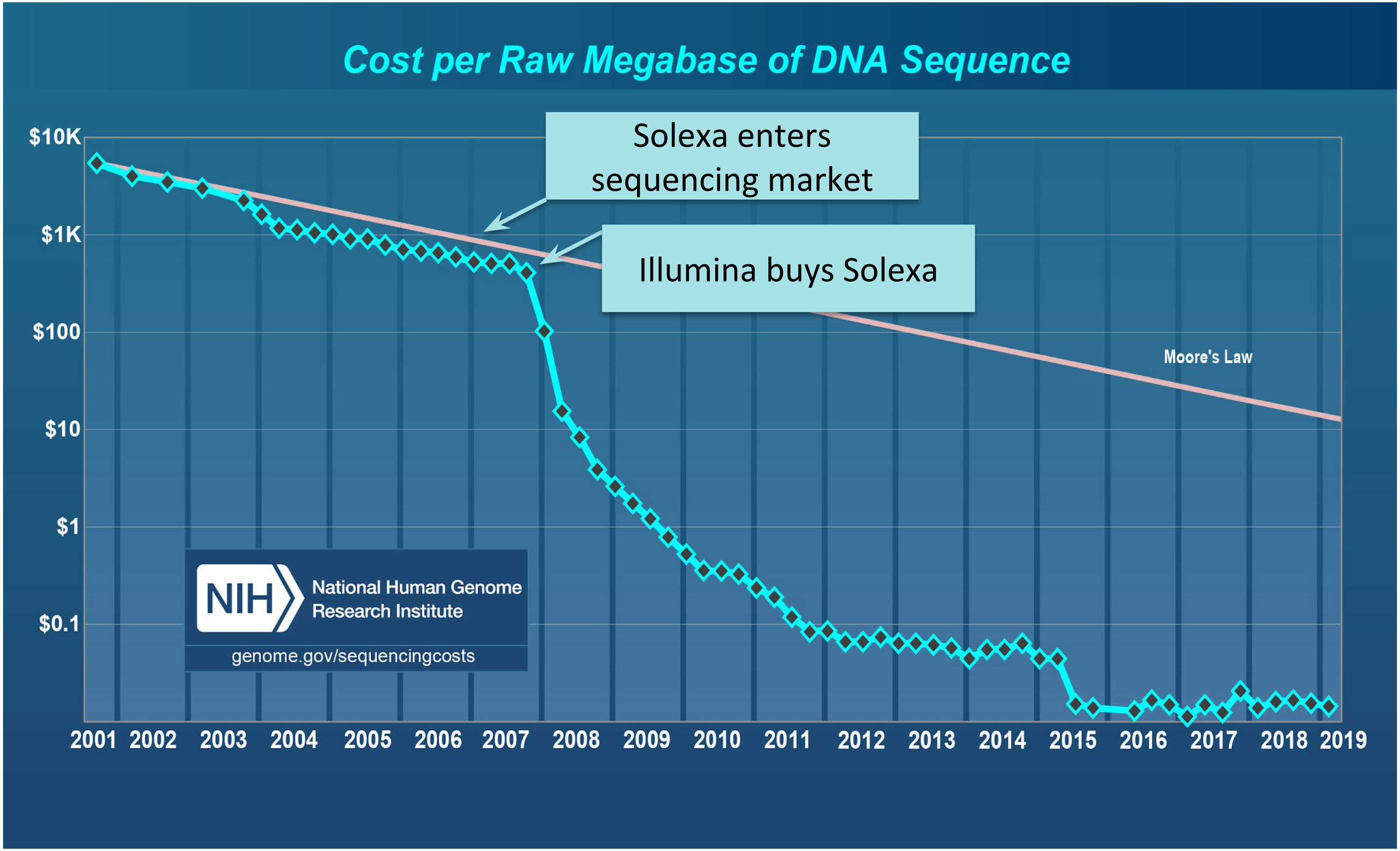


Market Summary > Illumina, Inc.
NASDAQ: ILMN

304.01 USD 0.00 (0.00%)
Jul 24, 4:00 PM EDT · Disclaimer

Announced acquisition of PacBio in late 2018 - may or may not go through.

More info: https://www.fool.com/investing/2019/07/22/what-happens-if-pacific-biosciences-isnt-acquired.aspx



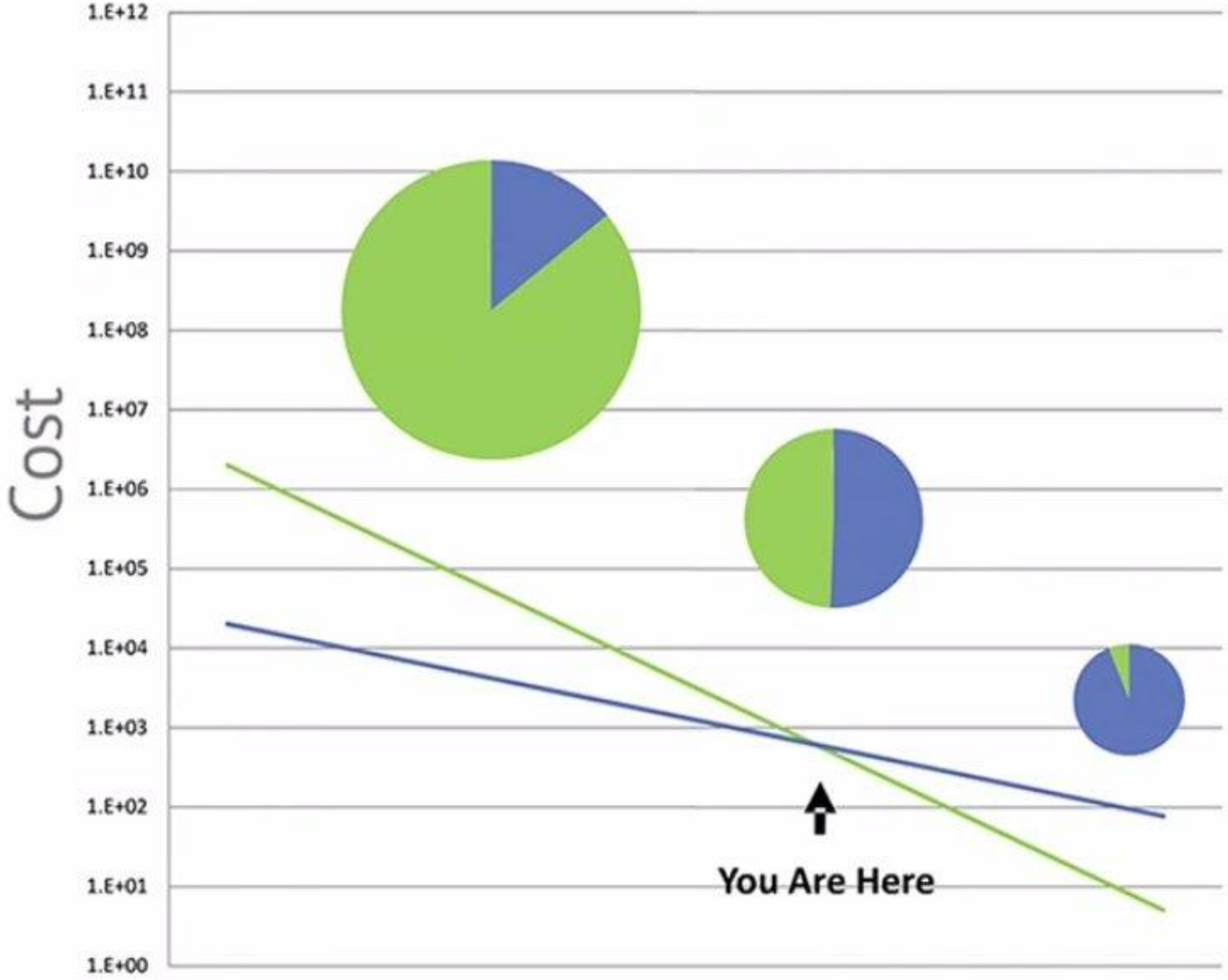
Cost per Raw Megabase of DNA Sequence

Solexa enters sequencing market

Illumina buys Solexa

https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

DNA Sequencing Economics



Sequencing cost dropping ~5x per year

Informatics cost dropping ~2x per year

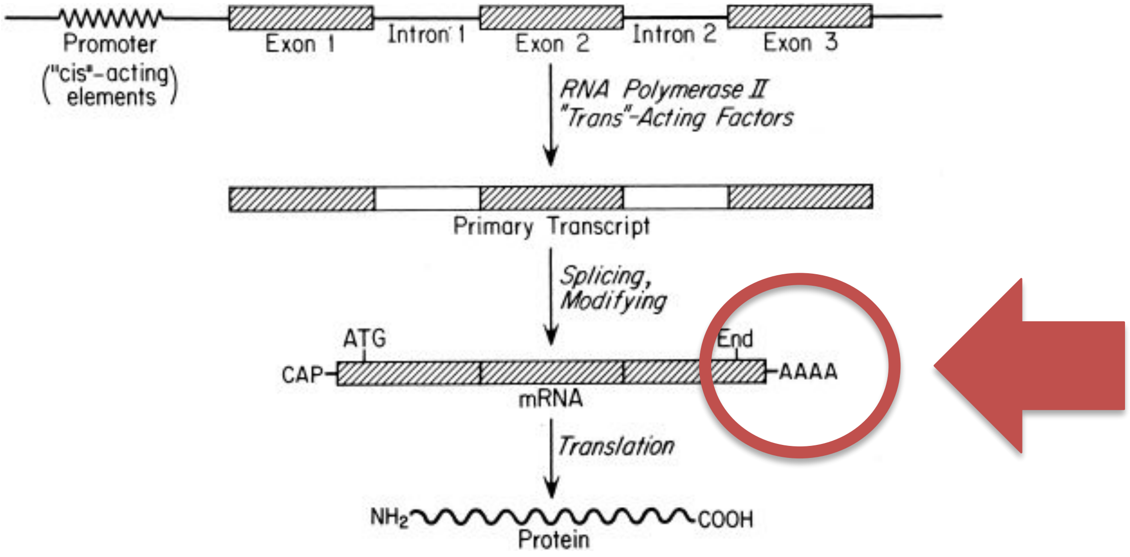
Informatics is now the bottleneck

Raymond McCauley 2014 https://www.youtube.com/watch?v=j88APStUcp4&feature=youtu.be

How does it work?

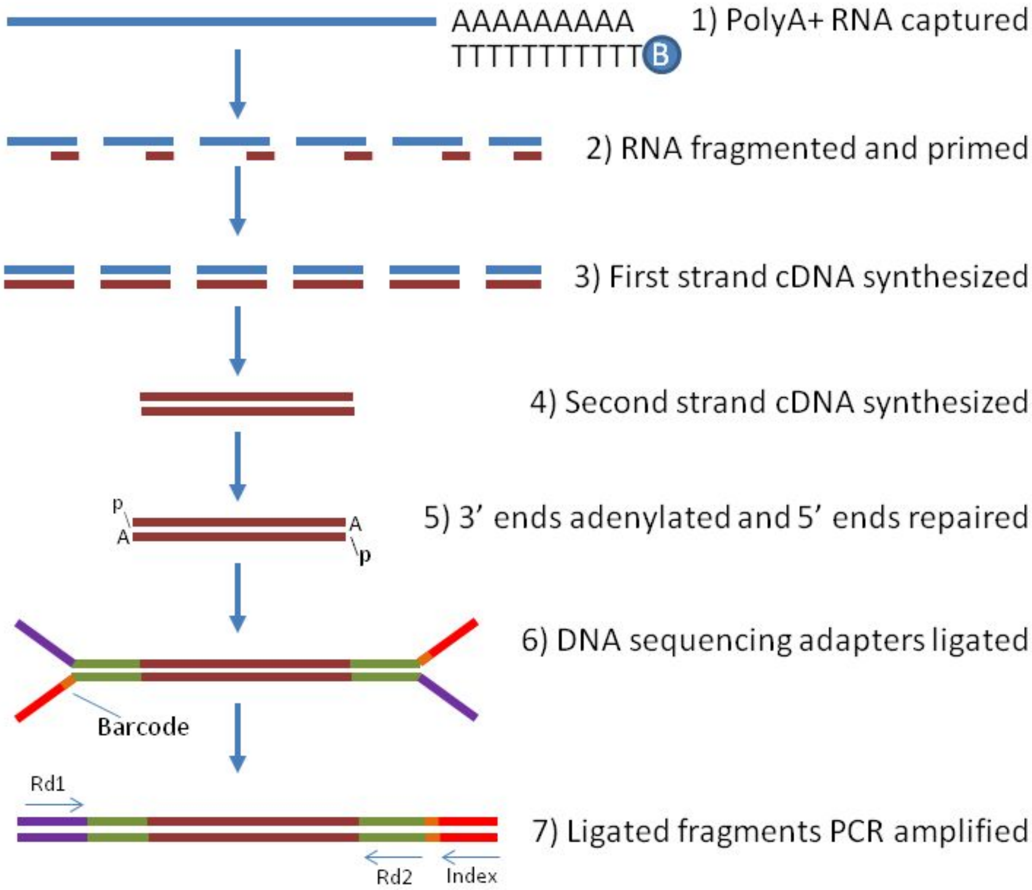
Targeting coding RNA for sequencing

1. Poly-A enrichment - purify the poly-A containing mRNA molecules using poly-T oligo attached magnetic beads (eukaryotes)



2. Riboremoval - rRNA depletion

Basics of Library Prep



1) PolyA+ RNA captured

2) RNA fragmented and primed

3) First strand cDNA synthesized

4) Second strand cDNA synthesized

5) 3’ ends adenylated and 5’ ends repaired

6) DNA sequencing adapters ligated

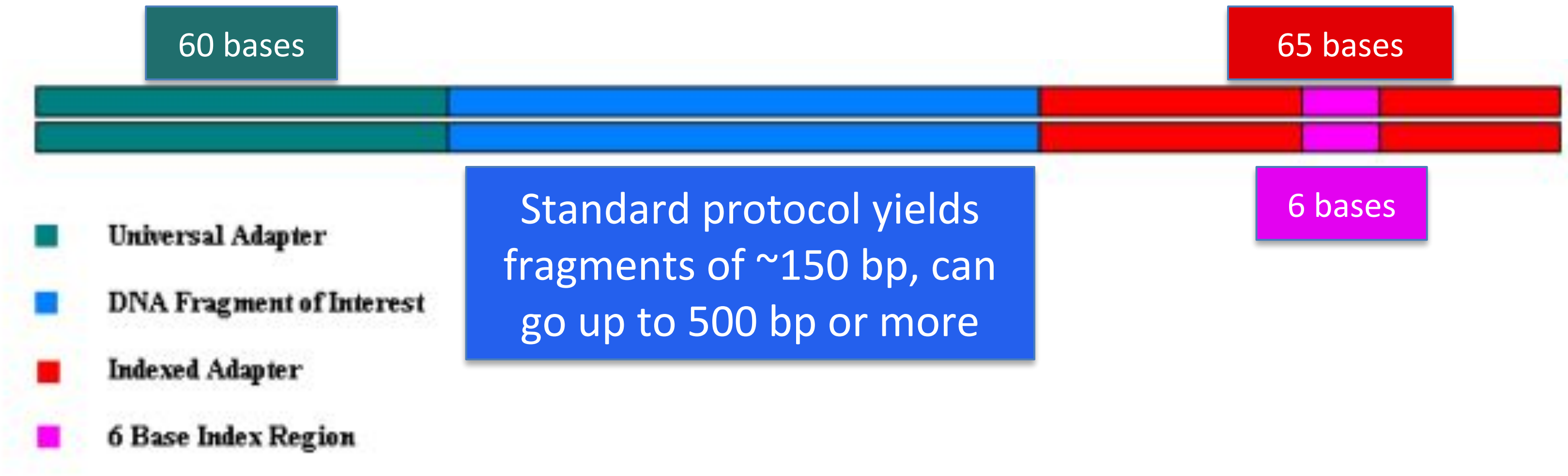
7) Ligated fragments PCR amplified

https://www.labome.com/method/RNA-seq.html

Final library fragments that are sequenced.

Library construction can vary by kit

TruSeq Example:



60 bases

65 bases

6 bases

Standard protocol yields fragments of ~150 bp, can go up to 500 bp or more

You will need the adapter sequences and a good understanding of adapter locations to later trim them out of your data

Illumina TruSeq DNA Adapters De-Mystified
https://www.med.unc.edu/pharm/calabreselab/files/tufts-sequencing-primer

Videos

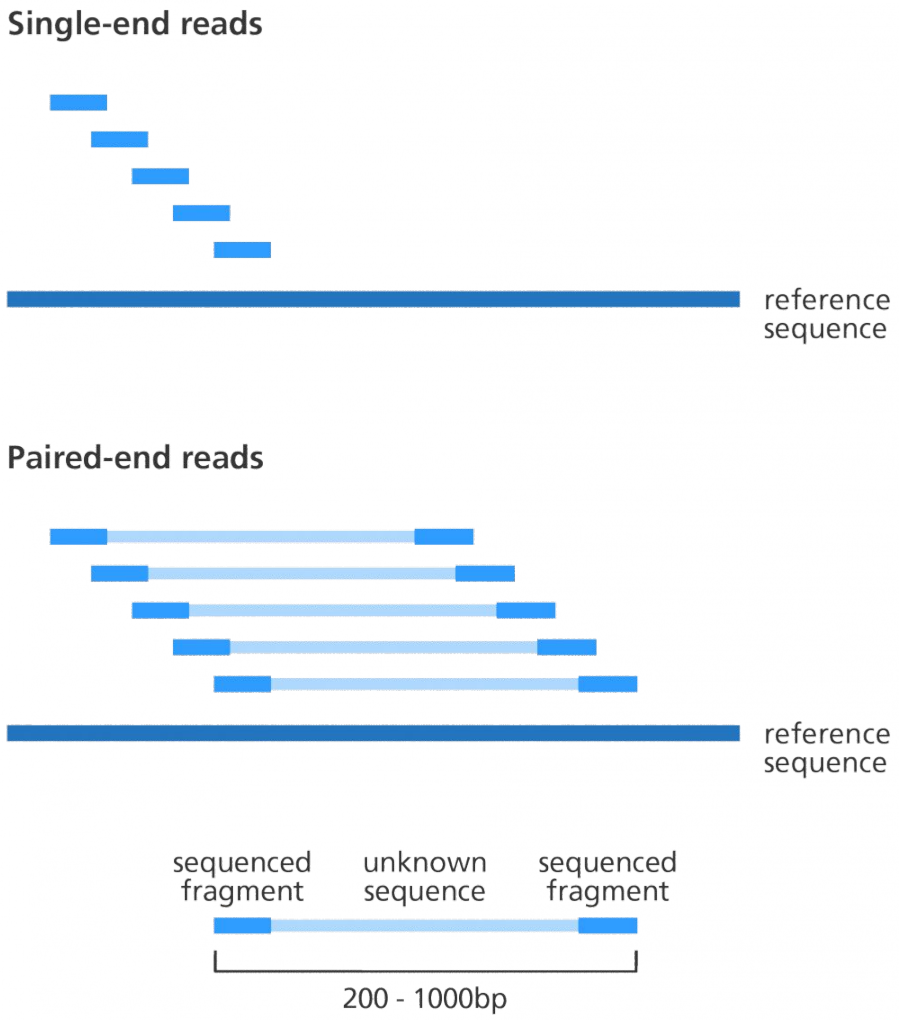
Illumina Sequencing by Synthesis by Illumina
https://www.youtube.com/watch?v=fCd6B5HRaZ8

Illumina Sequencing Technology by Illumina
https://www.youtube.com/watch?v=womKfikWlxM

Library Prep by ThermoFisher
https://www.youtube.com/watch?v=_yC0Bzw3WbQ

Also many platform specific videos

Paired End Sequencing

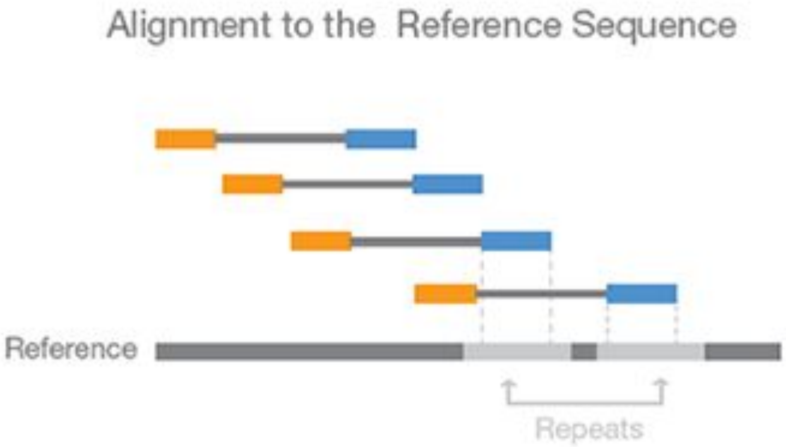


Devon Ryan https://www.biostars.org/p/267167/

Paired End Sequencing

Why?

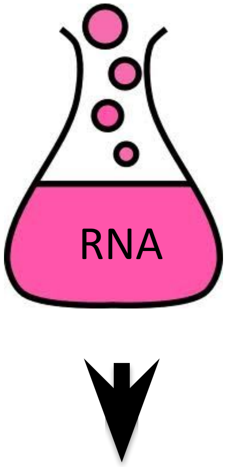
- Overcome lack of length.
- Map accurately to repetitive regions.
- Identify insertion/deletion mutations
- Better assembly.



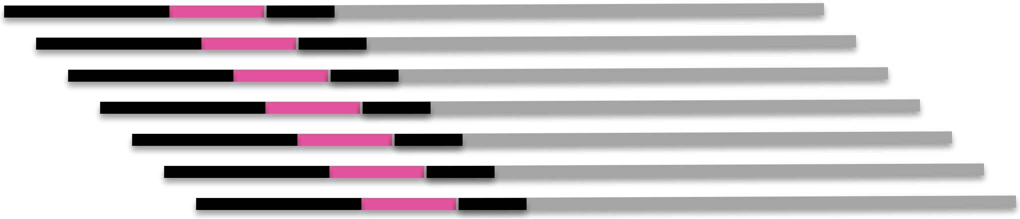
https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html

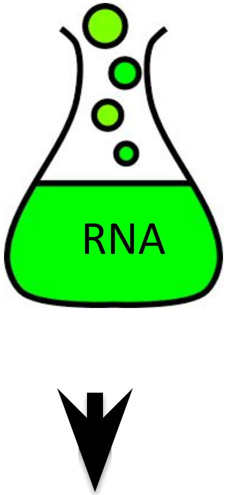
Multiplexing

Loading many samples into one lane.

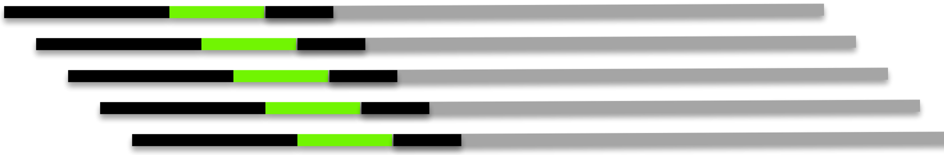


Pink Sample With CGATGT

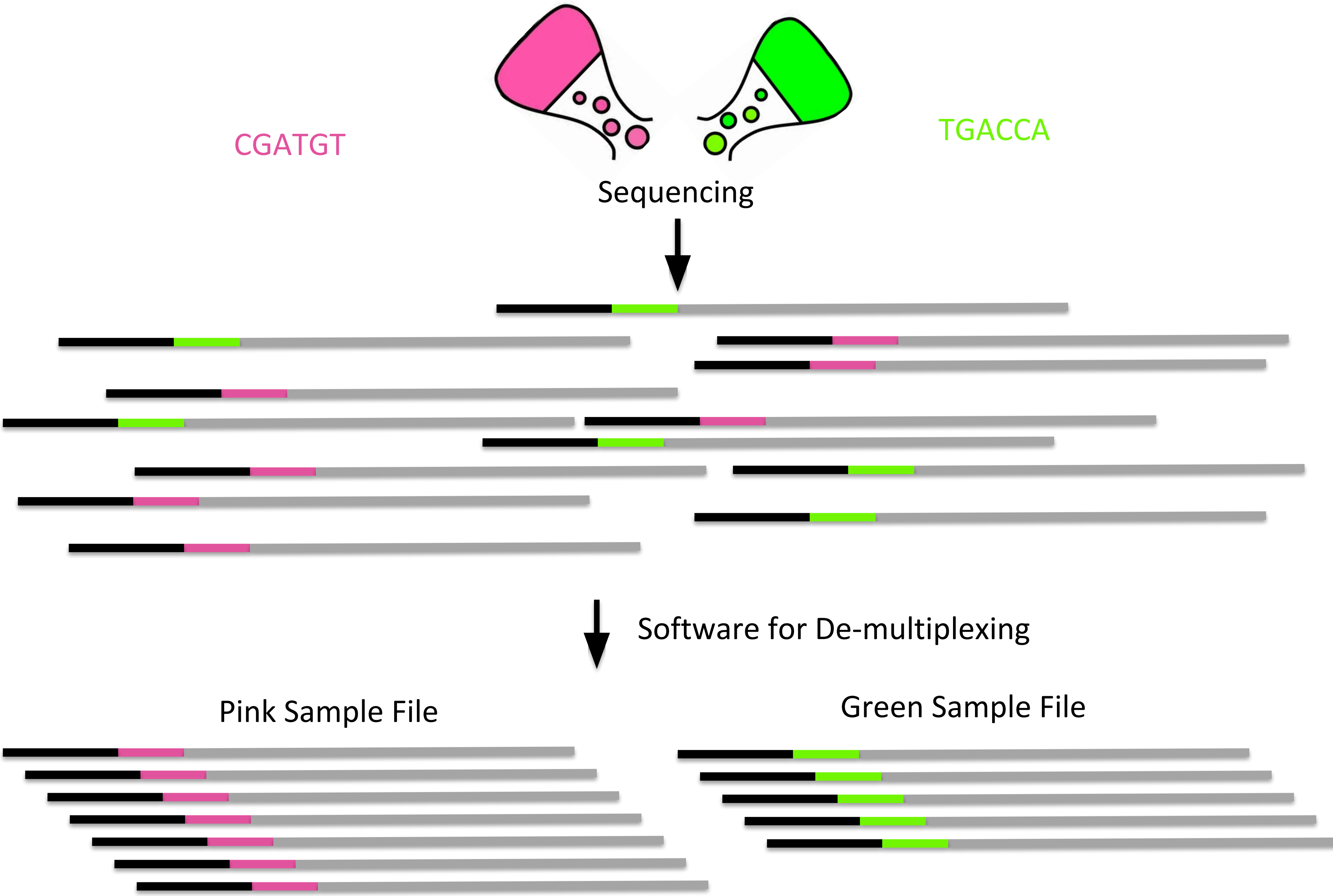




Green Sample with TGACCA



http://www.illumina.com/technology/next-generation-sequencing/multiplexing-sequencing-assay.html



CGATGT

TGACCA

Sequencing

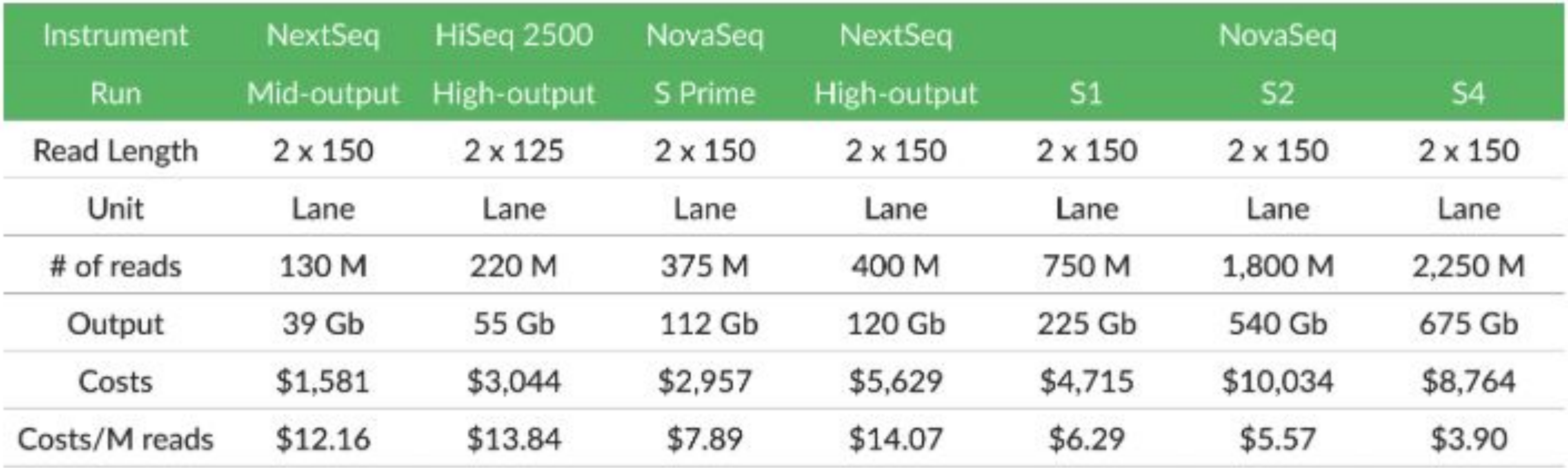
Software for De-multiplexing

Pink Sample File

Green Sample File

http://www.illumina.com/technology/next-generation-sequencing/multiplexing-sequencing-assay.html

Price and Throughput



Instrument	NextSeq	HiSeq 2500	NovaSeq	NextSeq	NovaSeq		
Run	Mid-output	High-output	S Prime	High-output	S1	S2	S4
Read Length	2 x 150	2 x 125	2 x 150	2 x 150	2 x 150	2 x 150	2 x 150
Unit	Lane	Lane	Lane	Lane	Lane	Lane	Lane
# of reads	130 M	220 M	375 M	400 M	750 M	1,800 M	2,250 M
Output	39 Gb	55 Gb	112 Gb	120 Gb	225 Gb	540 Gb	675 Gb
Costs	$1,581	$3,044	$2,957	$5,629	$4,715	$10,034	$8,764
Costs/M reads	$12.16	$13.84	$7.89	$14.07	$6.29	$5.57	$3.90

University of Minnesota: http://genomics.umn.edu/nextgen-novaseq.php

UNIVERSITY OF MINNESOTA

Price and Throughput

Companies offer regular deals:

RNASeq library prep + sequencing of 20 million reads per library
$189

GENEWIZ®
Solid science. Superior service.

Novogene

Experimental Goals for mRNA Seq

- Catalog of genes
- Gene expression levels
- Differential gene expression levels
- All of the above for alleles and splice variants
- Annotating the genes in a reference genome
- Variant (Genetic marker) discovery
- Post-transcriptional modifications, RNA-editing

Limitations

- Reverse transcription, PCR and fragmentation steps can introduce biases
  - Depletion of reads at both 5’ and 3’ ends
    - Difficult to identify the true start and end of novel transcripts
    - May underestimate expression level of short genes
    - PCR-free preps are available
  - GC bias, length bias
- Some genes are difficult to accurately quantify - too short, overlapping, gene family members
- Experimental design!

Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 2010 Jul;38(12):e131.
Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol.2011;12(3):R22.

Long Read Technologies

PacBio (IsoSeq for RNA) and Nanopore

- full length transcripts
- no fragmentation, no amplification
- more expensive
- great if you don’t have a reference genome
- great for discovering and profiling alternative splicing variants

File Formats

Fasta Format

```
>gi|31563518|ref|NP_852610.1|
microtubule-associated proteins 1A/1B
light chain 3A isoform b [Homo sapiens]
MKMRFFSSPCGKAAVDPADRCKEVQQIRD
QHPSKIPVIIERYKGEKQLPVLDKTKFLVPDHV
NMSELVKIIRRRLQLNPTQAFFLLVNQHSMV
SVSTPIADIYEQEKDEDGFLYMVYASQETFGF
>FN640832
CCTGGTAGCTATGGCTTGCCTTTACTAAGA
CCCATCTCAAACAGGCTCAATTATTTTTGGT
TCCAAGGGCCTGAAACATTCTTAAAGAAGC
GAATAGAGAAACACAGGAGCACAGTTTTT
CGCACCAATATCCCTCCAACTTTCCCTTTCT
TCTCCAATGTTAATCCCAGCGTTGTTGCTGT
CCTTGACACCAAGTCTTTTGCACACCTC
```

A sequence must start with a header line

- Begins with a >
- First “word” is the sequence id
- Rest of line may contain more sequence descriptors

Fasta Format

```
>gi|31563518|ref|NP_852610.1|
microtubule-associated proteins 1A/1B
light chain 3A isoform b [Homo sapiens]
MKMRFFSSPCGKAAVDPADRCKEVQQIRD
QHPSKIPVIIERYKGEKQLPVLDKTKFLVPDHV
NMSELVKIIRRRLQLNPTQAFFLLVNQHSMV
SVSTPIADIYEQEKDEDGFLYMVYASQETFGF
>FN640832
CCTGGTAGCTATGGCTTGCCTTTACTAAGA
CCCATCTCAAACAGGCTCAATTATTTTTGGT
TCCAAGGGCCTGAAACATTCTTAAAGAAGC
GAATAGAGAAACACAGGAGCACAGTTTTT
CGCACCAATATCCCTCCAACTTTCCCTTTCT
TCTCCAATGTTAATCCCAGCGTTGTTGCTGT
CCTTGACACCAAGTCTTTTGCACACCTC
```

The header is followed by the sequence

- May be amino acid or nucleotide
- May be a single line or multiple lines
- Should be consistent within a file

No empty line between sequence entries

Fastq Format

```
@SRR070570.1 HWUSI-EAS455:3:1:1:1388 length=41
CAGCACTAATGCACCGGATCCCATCAGAACTCCGCAGTTAA
+SRR070570.1 HWUSI-EAS455:3:1:1:1388 length=41
BACBC9BCC@.>C>96;CB@:?BB7@5>BA=:4.:B9>BB@
@SRR070570.2 HWUSI-EAS455:3:1:1:1785 length=41
CCAGAACACAAAGCTCATGACACGTTCACCTCCTGGAAGTT
+SRR070570.2 HWUSI-EAS455:3:1:1:1785 length=41
>AB@ACBB<BCA:>B;AA;@<B=;-=;<?@?<?=1-?B<8A
@SRR070570.3 HWUSI-EAS455:3:1:1:1679 length=41
ATCGATGAAGAACGTAGCGAAATGCGATACTTGGTGTGAAT
+SRR070570.3 HWUSI-EAS455:3:1:1:1679 length=41
BA=:==4?:8>A:8:>6:4:;2<07,<:@582+22'-';@>
```

Fastq Format

Sequence Identifier

Optional Description

```
@SRR070570.1 HWUSI-EAS455:3:1:1:1388 length=41
CAGCACTAATGCACCGGATCCCATCAGAACTCCGCAGTTAA
+SRR070570.1 HWUSI-EAS455:3:1:1:1388 length=41
BACBC9BCC@.>C>96;CB@:?BB7@5>BA=:4.:B9>BB@
```

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

Fastq Format

The Sequence

```
@SRR070570.1 HWUSI-EAS455:3:1:1:1388 length=41
CAGCACTAATGCACCGGATCCCATCAGAACTCCGCAGTTAA
+SRR070570.1 HWUSI-EAS455:3:1:1:1388 length=41
BACBC9BCC@.>C>96;CB@:?BB7@5>BA=:4.:B9>BB@
```

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

Fastq Format

Totally useless line that begins with a + but does not need anything else; id and description are sometimes repeated.

```
@SRR070570.1 HWUSI-EAS455:3:1:1:1388 length=41
CAGCACTAATGCACCGGATCCCATCAGAACTCCGCAGTTAA
+SRR070570.1 HWUSI-EAS455:3:1:1:1388 length=41
BACBC9BCC@.>C>96;CB@:?BB7@5>BA=:4.:B9>BB@
```

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

Fastq Format

Quality values for each base.

```
@SRR070570.1 HWUSI-EAS455:3:1:1:1388 length=41
CAGCACTAATGCACCGGATCCCATCAGAACTCCGCAGTTAA
+SRR070570.1 HWUSI-EAS455:3:1:1:1388 length=41
BACBC9BCC@.>C>96;CB@:?BB7@5>BA=:4.:B9>BB@
```

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

FASTQ Quality Scores

Scores are encoded as a single character. From lowest score to highest score:

!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI

0…   …41

Can calculate the likelihood of a base being wrong with a logarithmic formula.

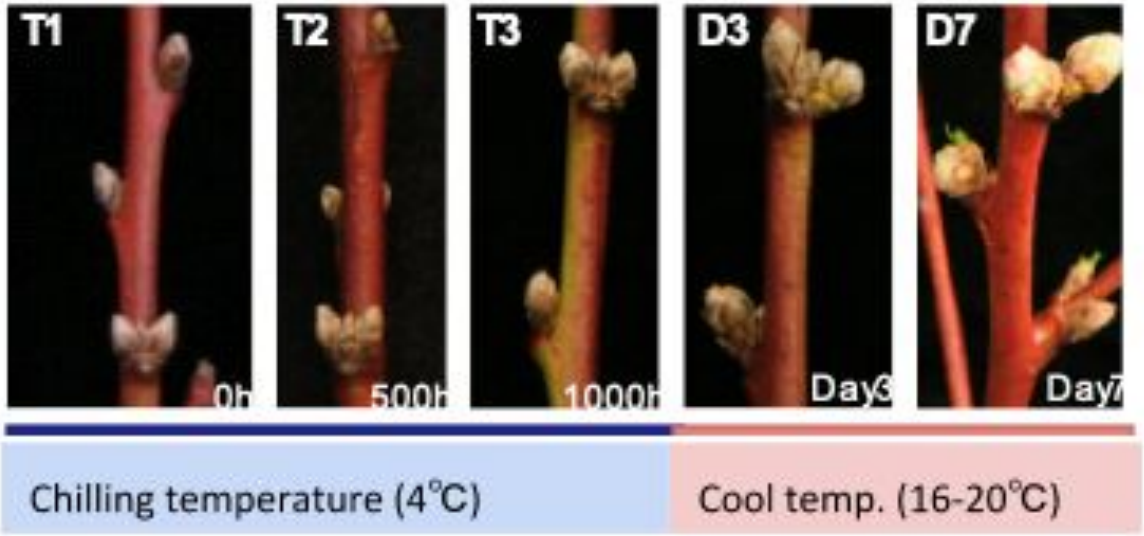
An I is 99.99% likely be correct.

A * is only 90% likely to be correct.

https://en.wikipedia.org/wiki/Phred_quality_score
Ewing et al, 1998

Data

- USDA grant “Abiotic Stress Response And Adaptive Phenology In Fruit Trees”
- Dormancy in apricots (*Prunus armeniaca*) and peaches (*Prunus persica*)
- Late blooming (high chill) variety – adapted to northern climates
- Early blooming (low chill) variety – adapted to southern climates
-



Questions before we begin?