Basics of RNASeq

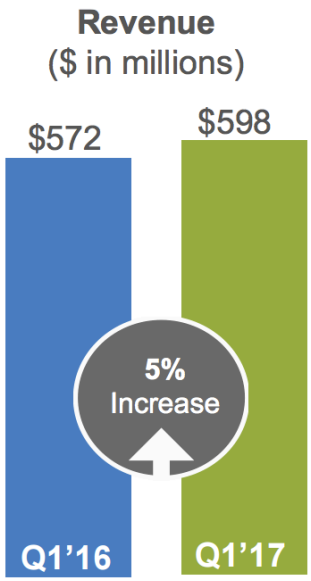
Outline

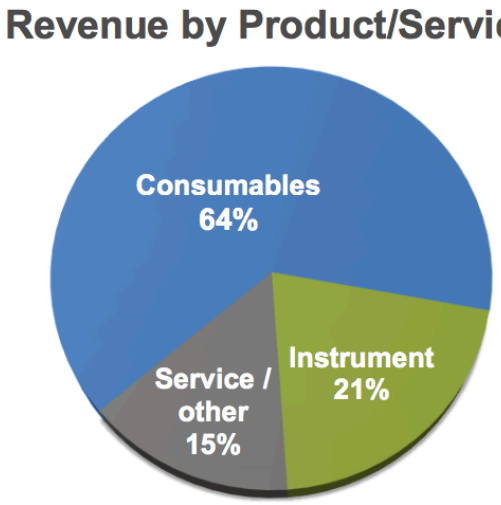
- Illumina platform
- RNA Review
- RNASeq
- First Steps in Data Analysis
- Data set for the class

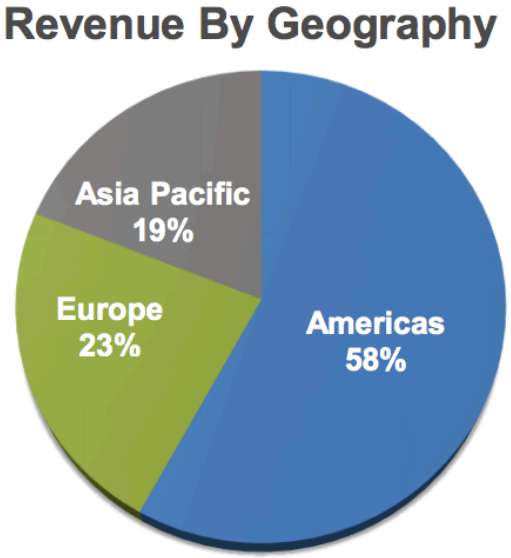

illumina

Illumina Sequencing Technology

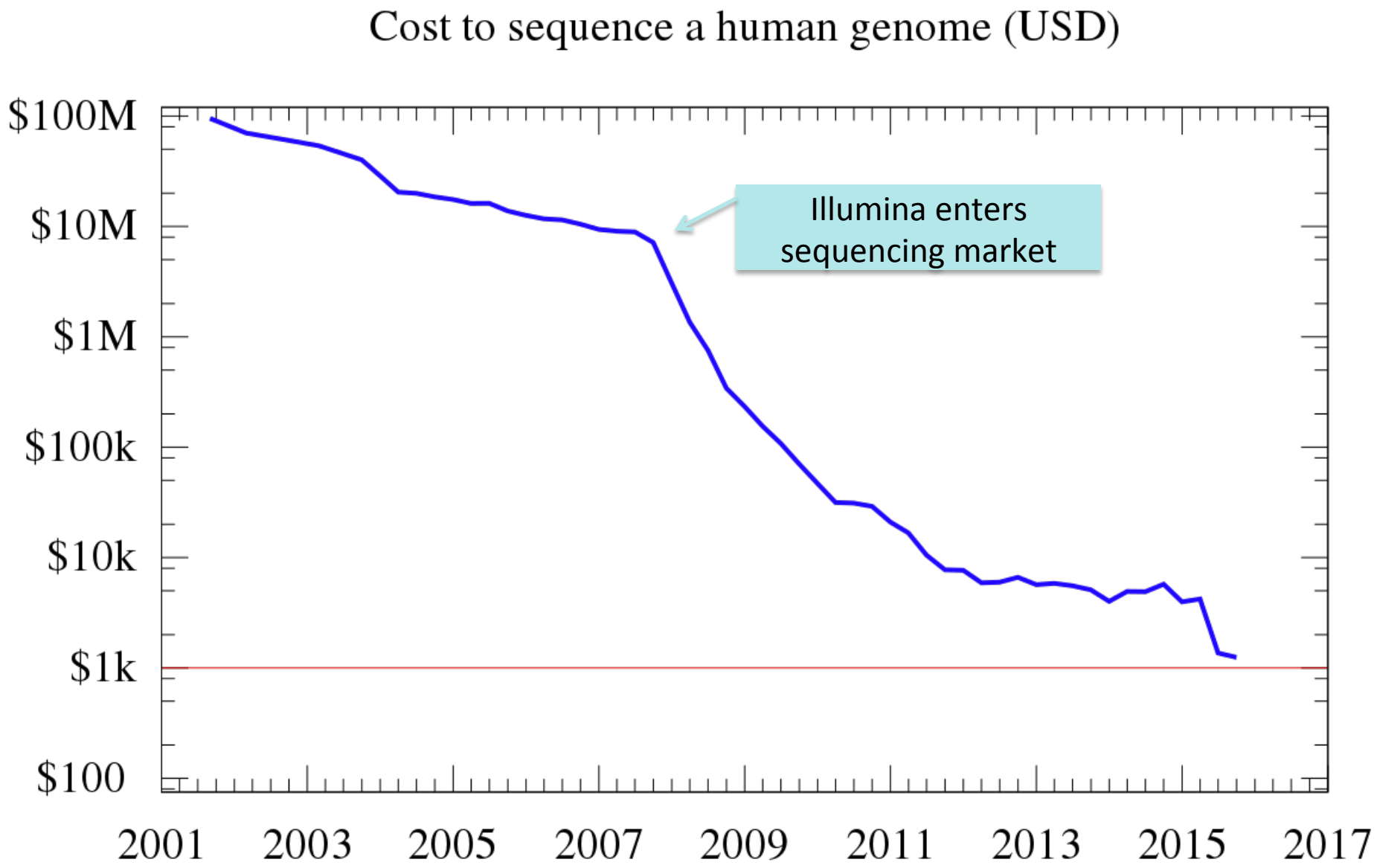
- $2.4 billion in revenue in 2016
- 90% sequencing market share (estimated) in 2016
- Why are they so popular?
  - Low price
  - High throughput
  - High base calling fidelity
  - Paired end sequencing





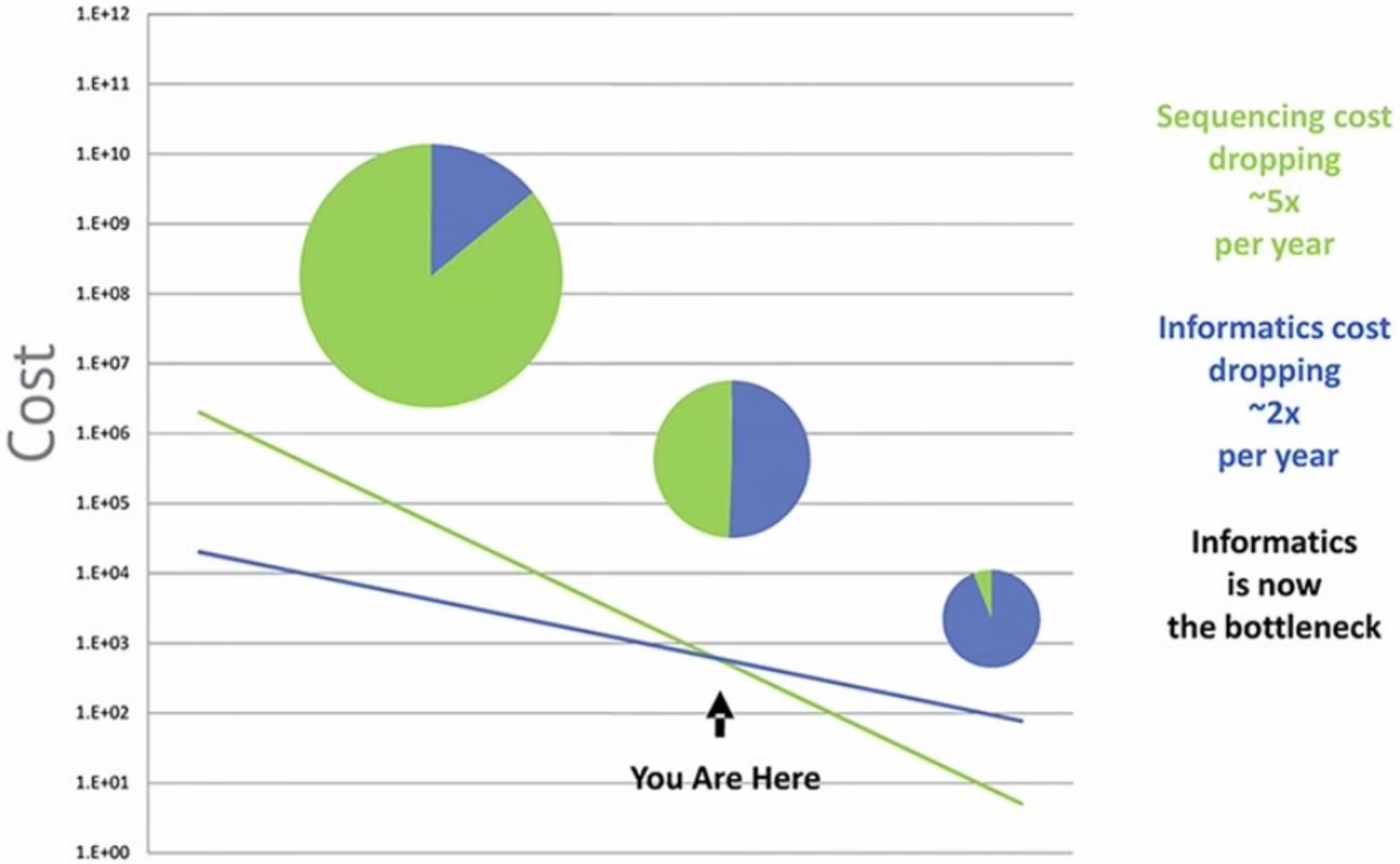


https://www.illumina.com/content/dam/illumina-marketing/documents/company/investor-relations/investor_presentations/illumina_investor_presentation.p



https://en.wikipedia.org/wiki/Whole_genome_sequencing

DNA Sequencing Economics



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

Price and Throughput

	read type	price	Average read pair yield
MiSeq v3	Paired End (2x300)	$1,775	22 million
HiSeq 4000	Paired End (2x150)	$3,202	240 million
NextSeq 500	Paired End (2x150)	$6,636	330 million

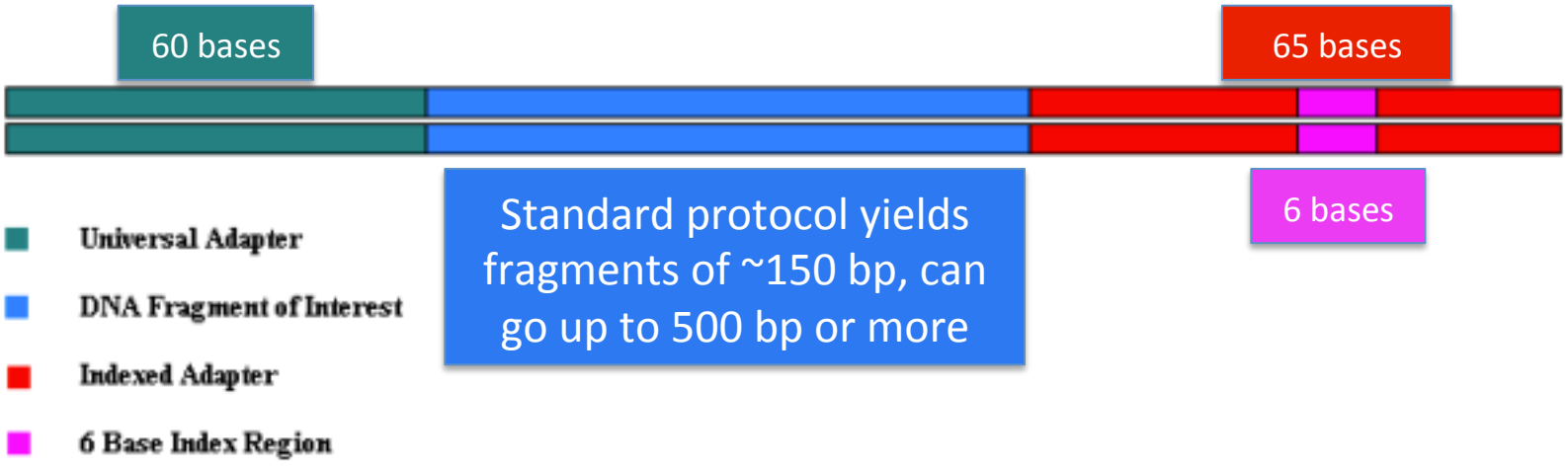
Now also have NovaSeq!

UT Austin Genome Sequencing and Analysis Center https://wikis.utexas.edu/display/GSAF/Pricing

How does it work?

Library construction can vary by kit

TruSeq Example:



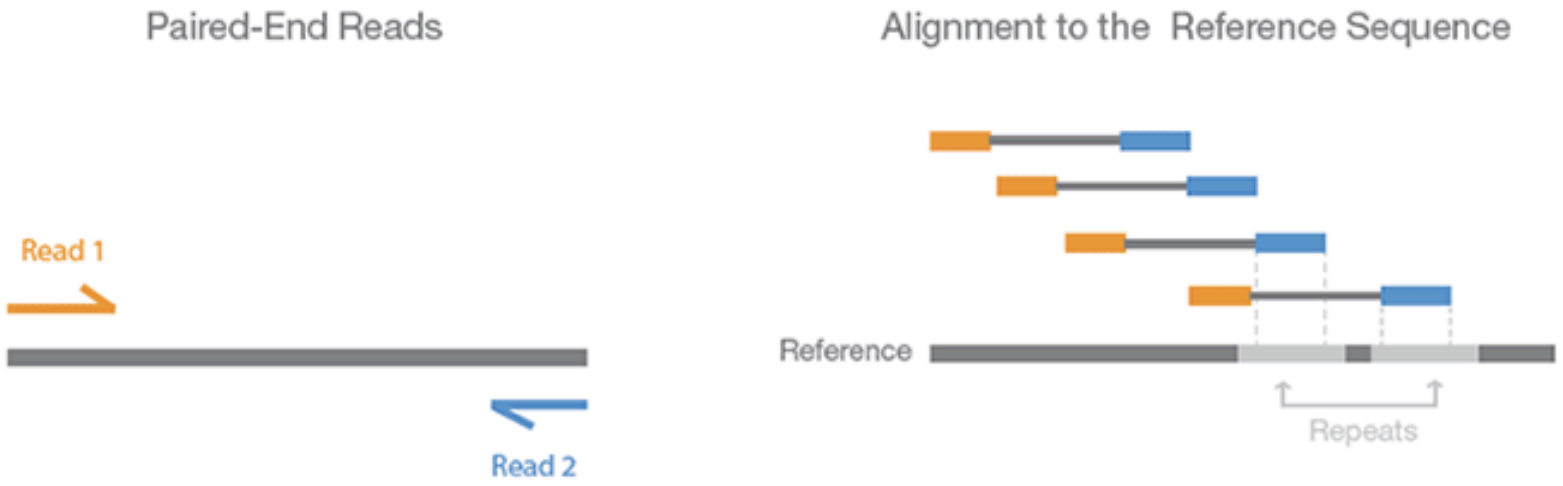
You will need the adapter sequences and a good understanding of adapter locations to later trim them out of your data

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

Paired End Sequencing

Overcome lack of length.

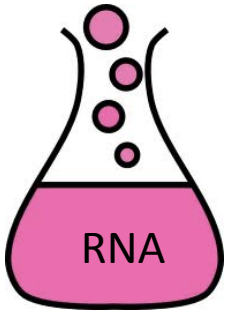
Figure 4. Paired-End Sequencing and Alignment



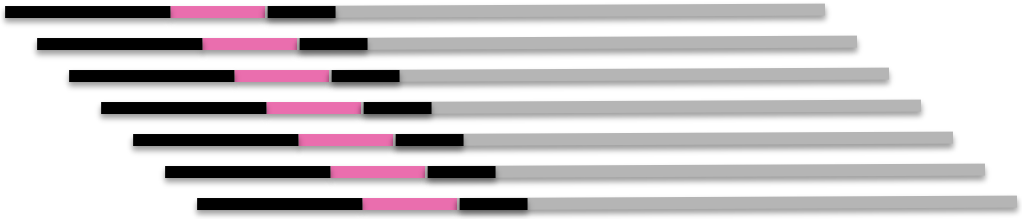
Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across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

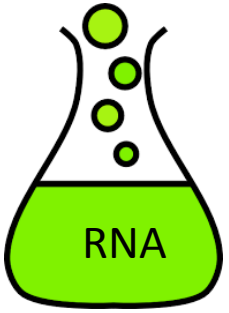
Multiplexing

Loading many samples into one lane.



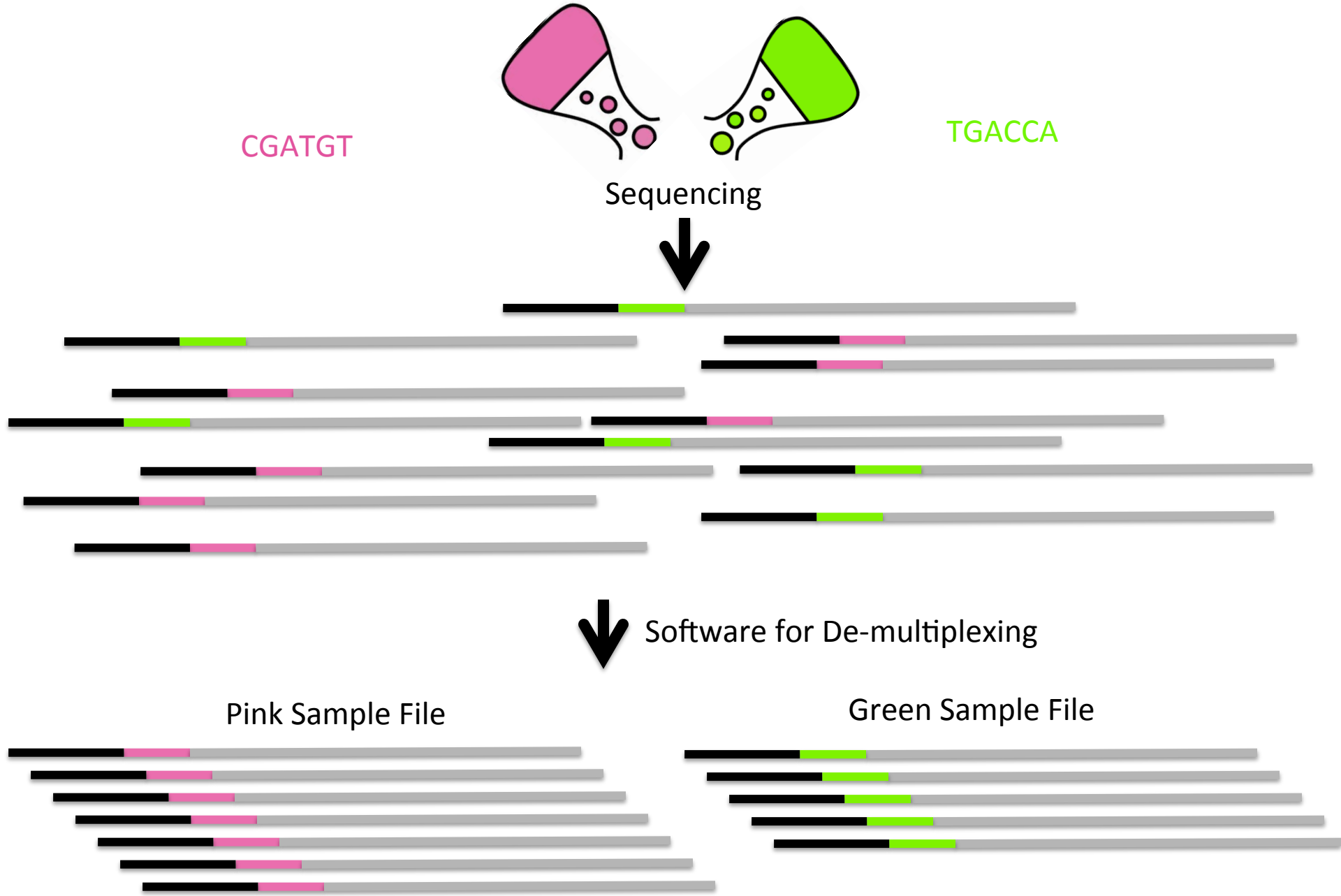
Pink Sample With CGATGT





Green Sample with TGACCA

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

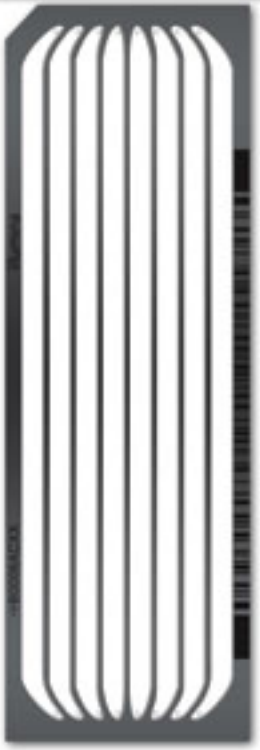


http://www.illumina.com/technology/next-generation-sequencing/multiplexing-sequencing-assay.html

Run vs. Lane

- Used interchangeably or as something different?

MiSeq
1 run, 1 input sample



HiSeq
Flow Cell
8 lanes, 8 samples

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

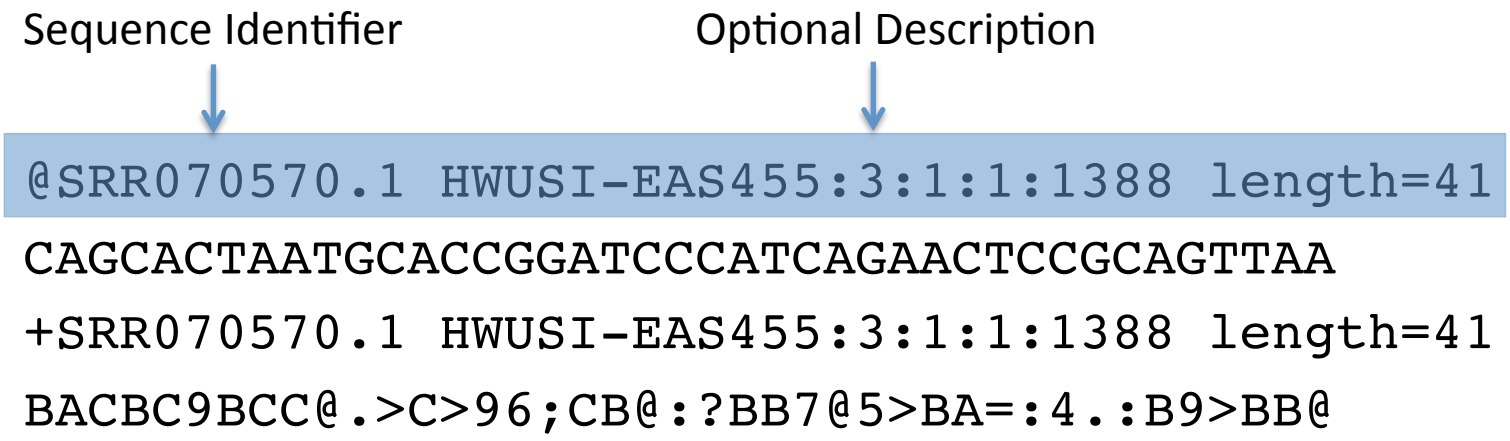
- May be amino acid or nucleotide
- May be a single line or multiple lines
- Should be consistent within a file

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



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.

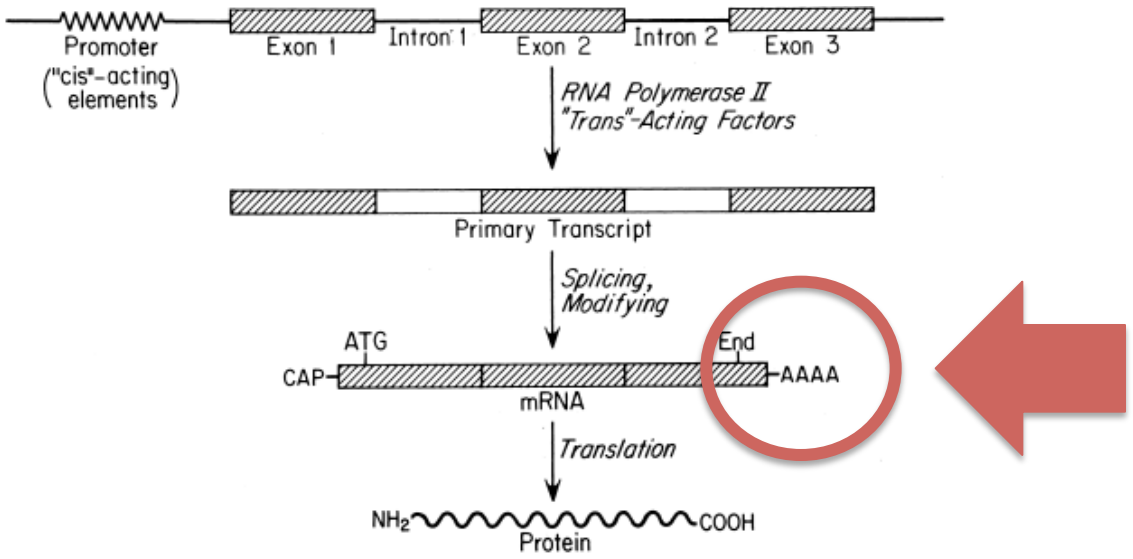
A * is only 90% likely to be correct.

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

RNA Sequencing

Targeting mRNA for sequencing

- To target mRNA
  - Poly-A enrichment - purify the poly-A containing mRNA molecules using poly-T oligo attached magnetic beads
  - Only works for eukaryotes

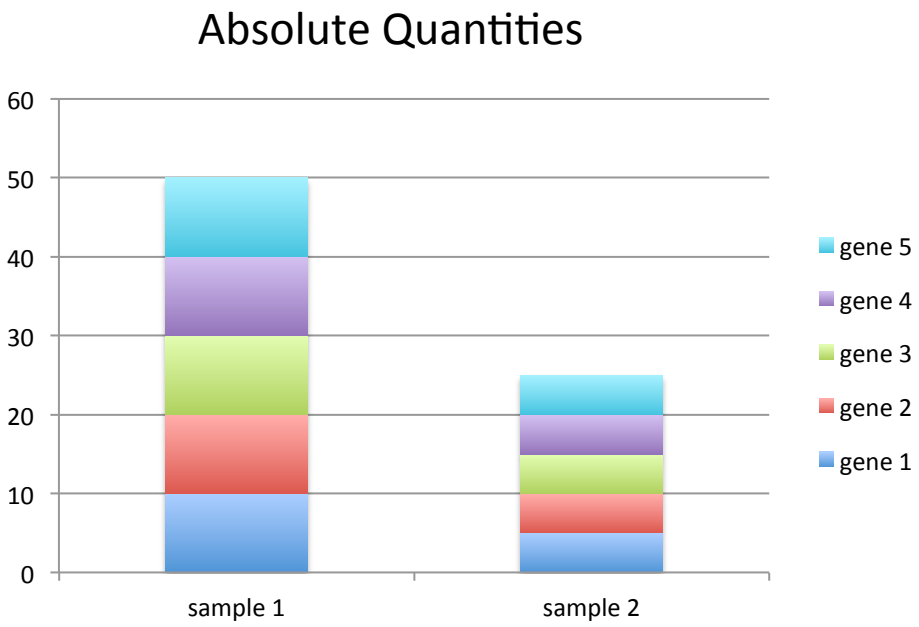


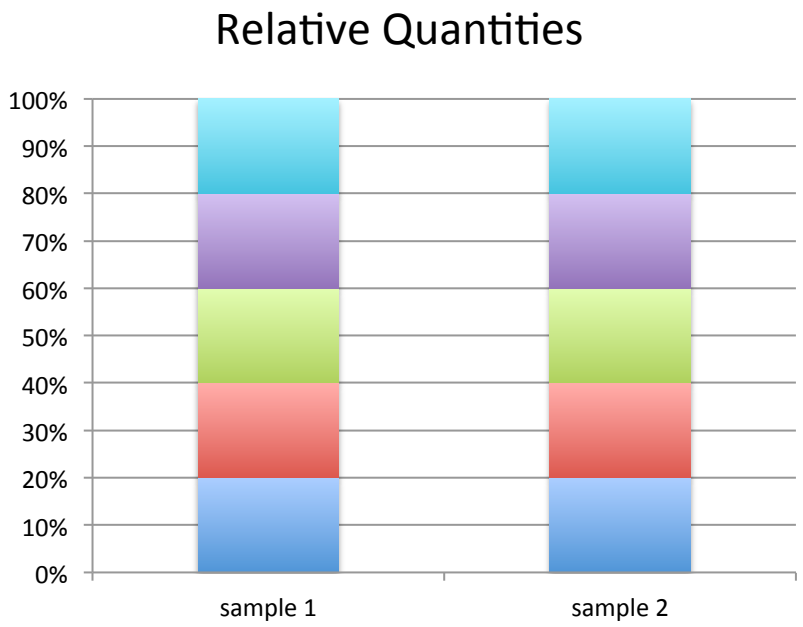
Experimental Goals for mRNA Seq

- 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

RNASeq gives you relative abundance only





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
  - GC bias, length bias
- PCR-free preps are available

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.

Data

- USDA grant “Abiotic Stress Response And Adaptive Phenology In Fruit Trees”
- Dormancy in Apricots (*Prunus armeniaca*)
- Late blooming (high chill) variety – adapted to northern climates
- Early blooming (low chill) variety – adapted to southern climates
- At 800 chill hours, how is gene expression different inside the bud?

Bert Abbott
Forest Health Research Center
University of Kentucky

Questions before we begin?