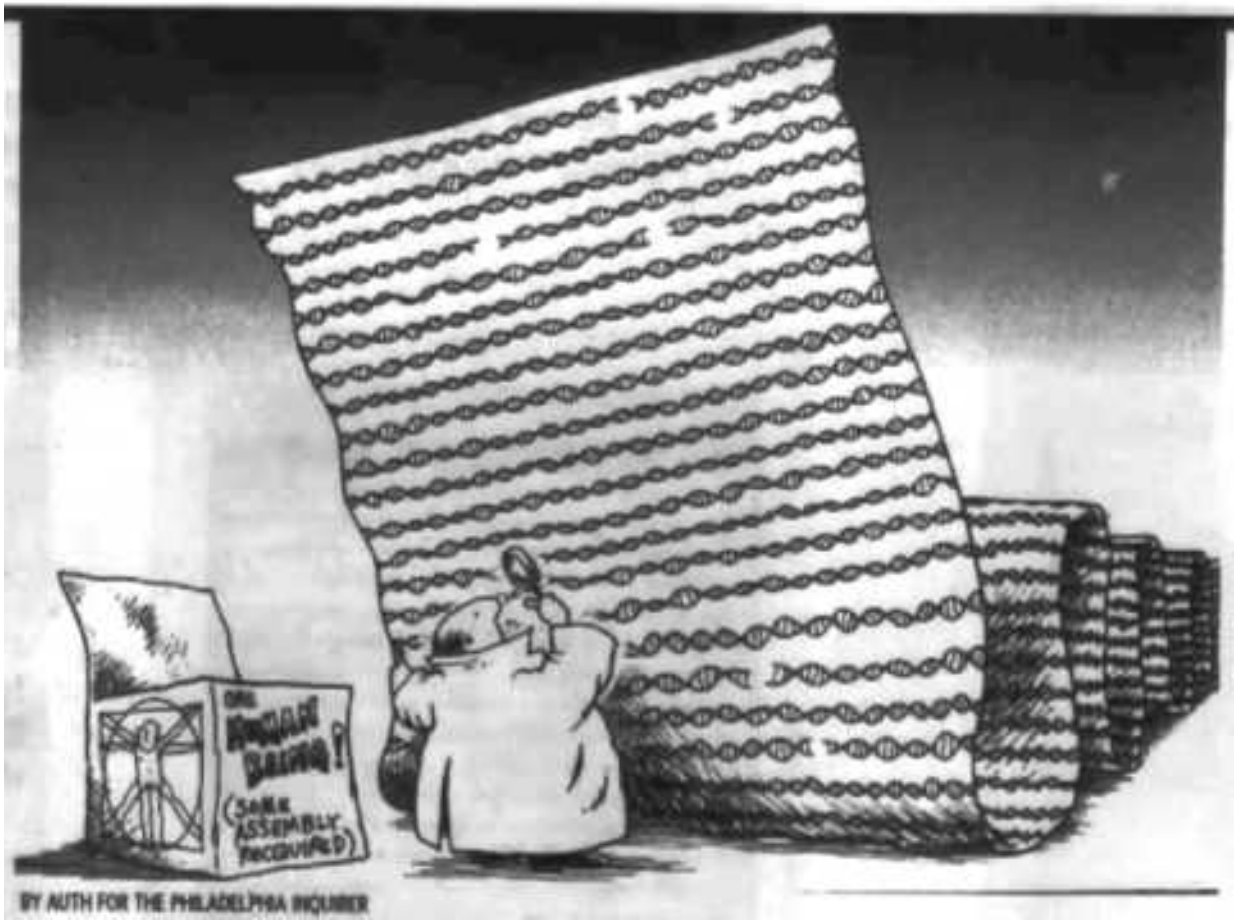
Overview of High-throughput Sequencing

Outline

- Omics
- History of NGS
- NGS platform chemistry
- Illumina platform
- File formats

NGS =
Next generation sequencing

HTS =
High throughput sequencing



Omics

Birth of the word Genome

Translated from German:

"I propose the expression Genom for the haploid chromosome set, which, together with the pertinent protoplasm, specifies the material foundations of the species."

-Hans Winkler, German botanist,

from "Spread and cause of pathogenesis in plant and animal kingdoms," 1920

Genome becomes Omics

- Biome established as a word in 1916, genome in 1920
- In molecular biology, "ome" has come to mean all constituents considered collectively
- Surprisingly, unrelated to the word chromosome

- All from 1990s:
  - Transcriptomics
  - Metabolomics
  - Proteomics

"the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms."

Eisen's Bad Omics

- The –ome usage is a bit out of control. This is a series of great examples of terrible usage.
- It is often meant to convey large amounts of data or a new global approach
- Do these new words really bring anything new or meaningful to the discussion?

- Nutrimetabonomics
- sexome
- circomics
- nascentome
- negatome
- diseasome, receptorome, uniqueome, drugome, adversomics, bibliome, N-terminome, transactome, nutriome, miRNAome, tRNomics, variome, speechome, vaccinomics, pharmacomicrobiomics, and museomics.

Eisen, "Badomics words ands and the power and peril of the ome-meme", GigaScience, 2012

Next generation sequencing

Advent of next generation sequencing

2005

Roche released the 454 platform (same amount of data as 50 capillary sequencers at 1/6th the cost)

Chemistry: Pyrosequencing

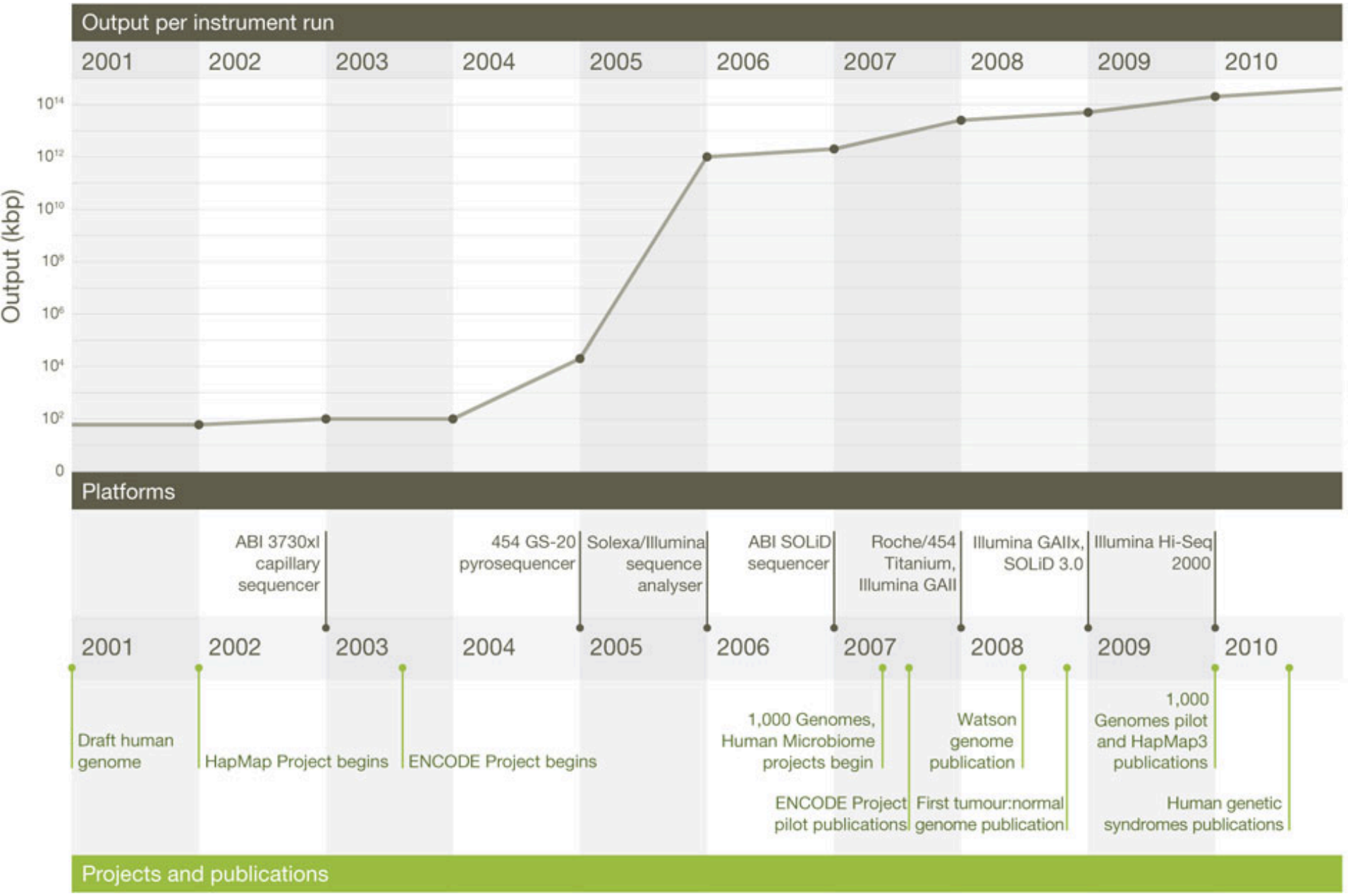
illumina

2006

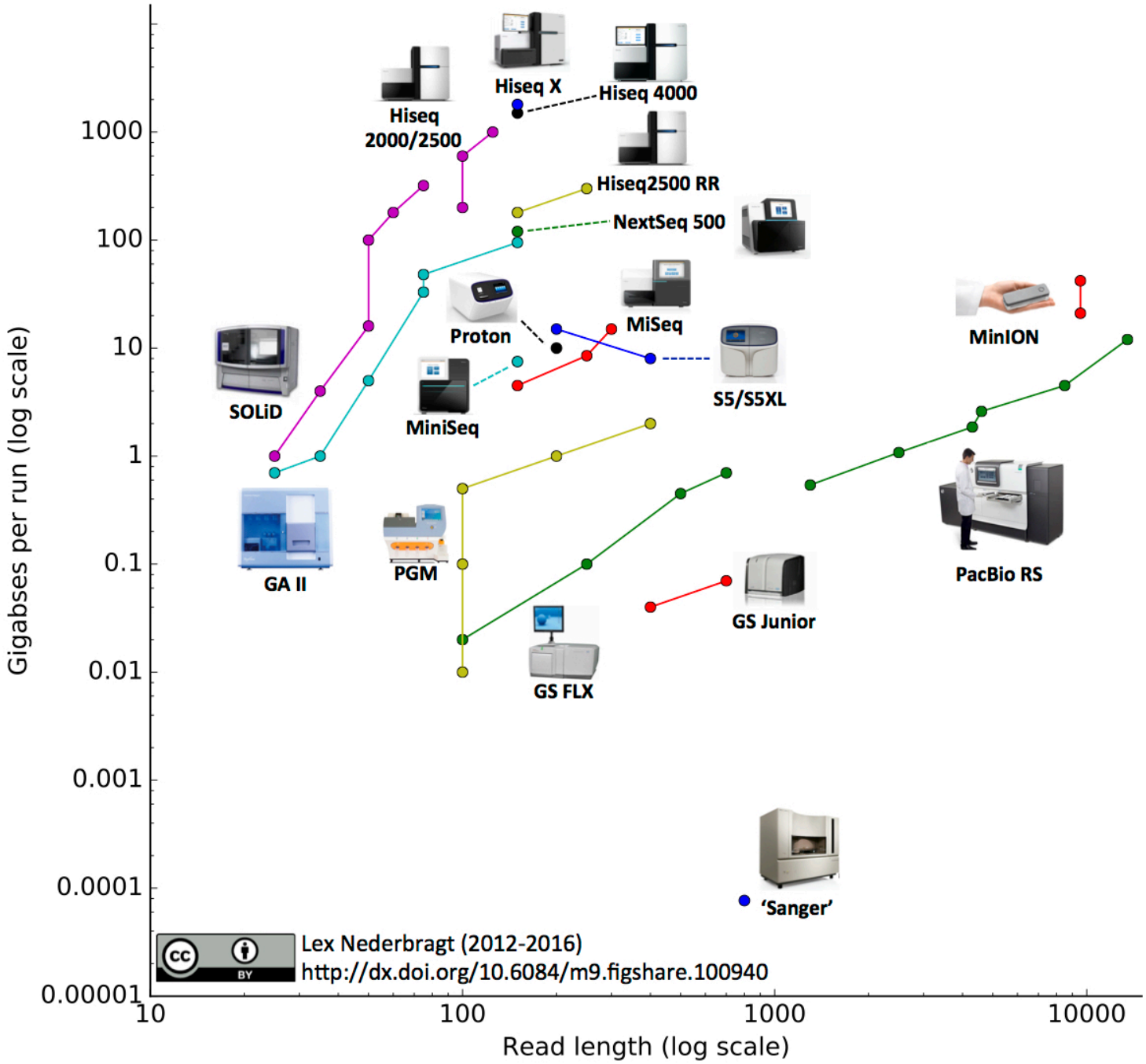
Illumina releases Solexa Genome Analyzer

Chemistry:

Sequening by Synthesis

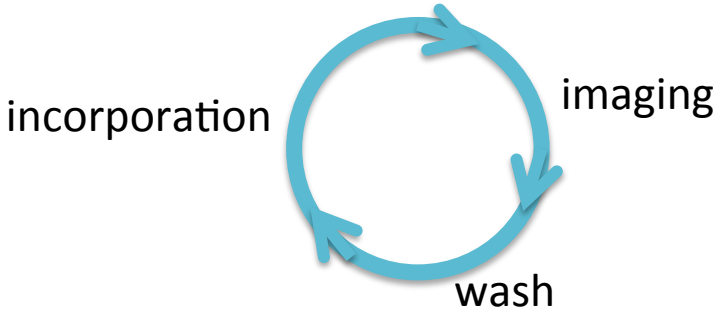


Mardis, A decade's perspective on DNA sequencing technology, Nature, 2011


Hiseq X
Hiseq 4000
Hiseq 2000/2500
Hiseq2500 RR
NextSeq 500
Proton
MiSeq
MinION
SOLiD
MiniSeq
S5/S5XL
GA II
PGM
PacBio RS
GS Junior
GS FLX
'Sanger'
Gigabses per run (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1000
10000
Read length (log scale)
Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940

Shared by most NGS

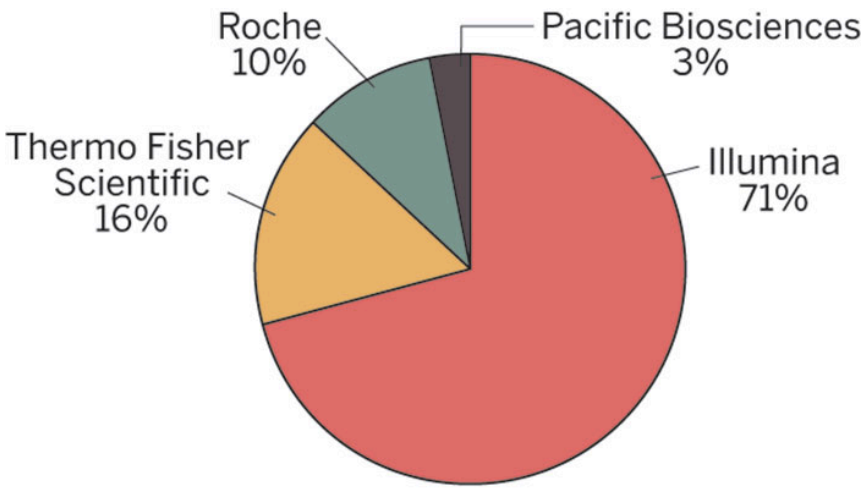
- DNA sample preparation
  - addition of defined sequences, known as “adapters,” to the ends of randomly fragmented DNA
  - Creates a “library”
- Immobilization
  - Adapters are used to anchor the individual DNA fragments to a solid surface like a glass slide
  - amplification is required to form spatially distinct and detectable sequence features (except PacBio)
- Sequencing
  - DNA polymerase synthesis with fluorescent nucleotides




illumina

Illumina Sequencing Technology

- Company was worth $28 billion in 2015
- 80% market share in 2014 *
- Predicted to continue at 75%+ market share until 2020
- Why are they so popular?
  - Low price
  - High throughput
  - Paired end sequencing

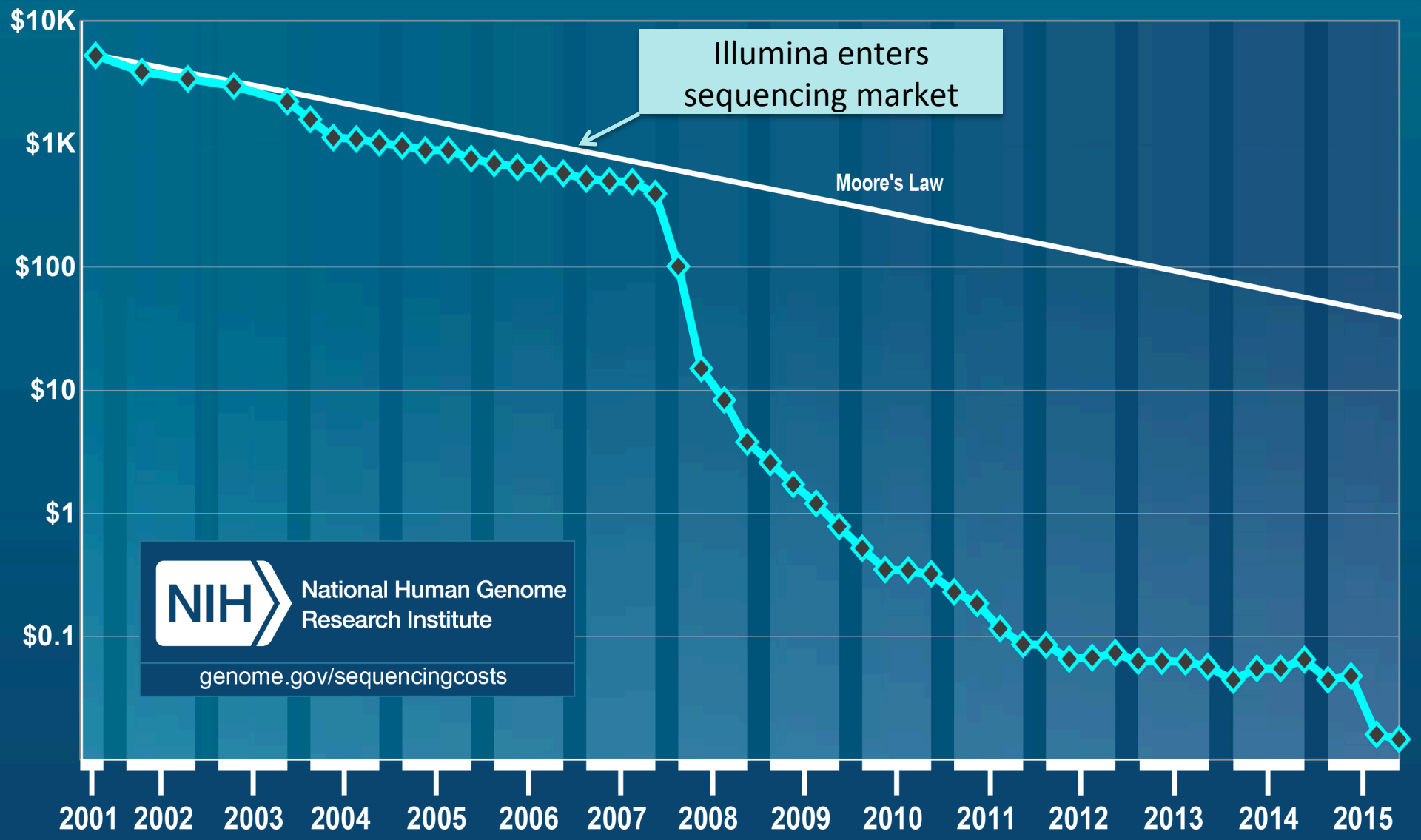


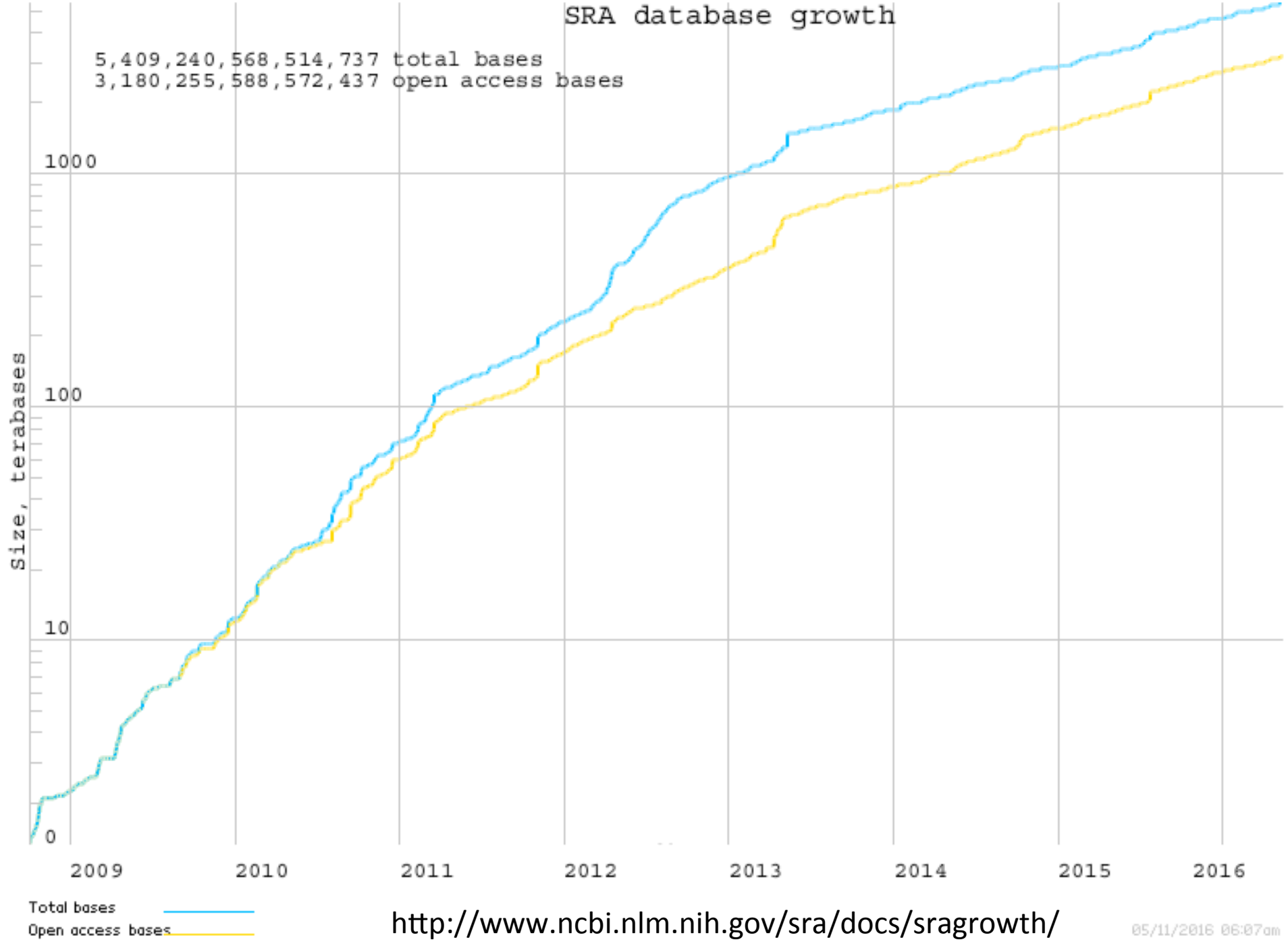
World market in 2013 = $1.3 billion

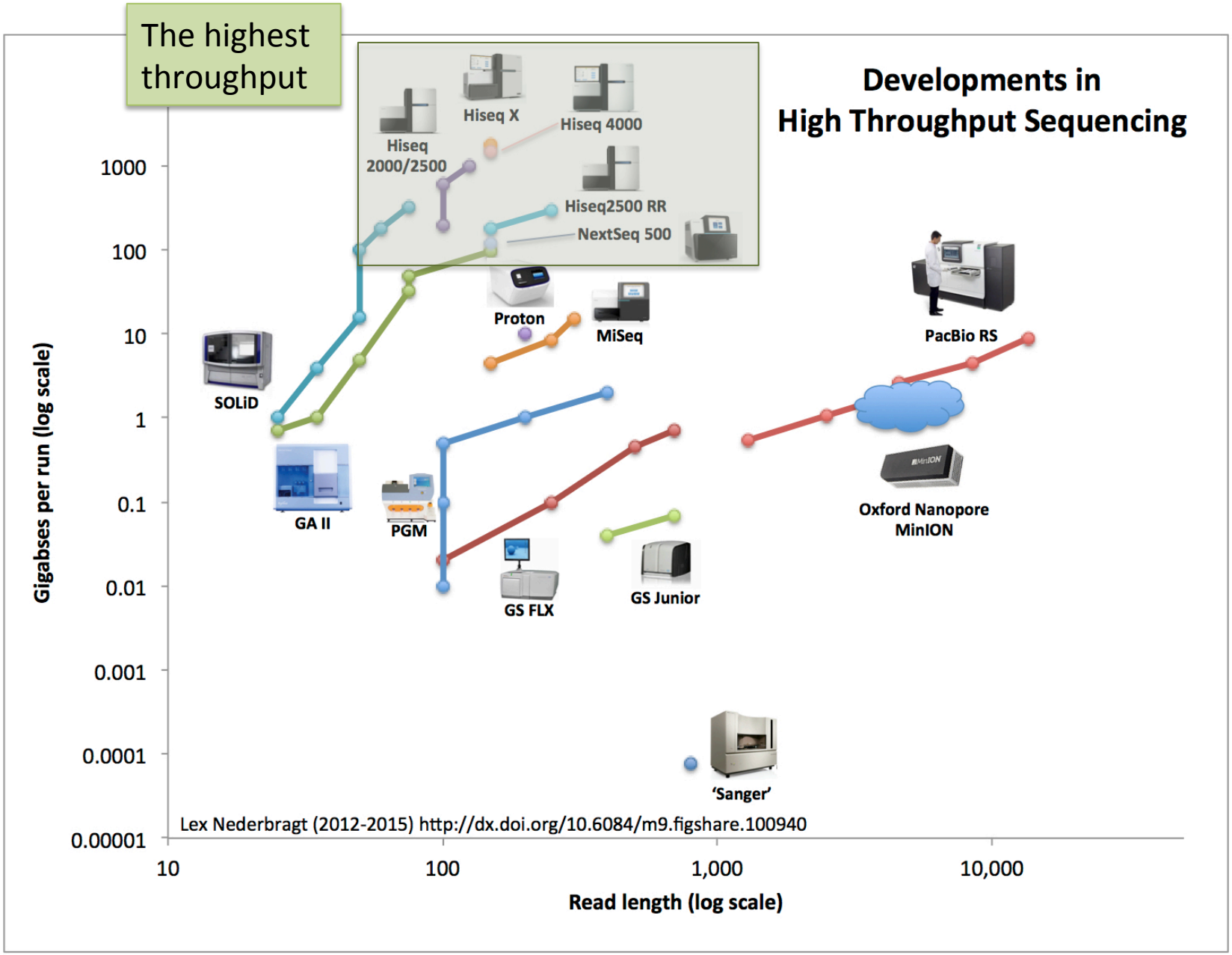
World market predicted to increase to $20 billion in 2016

- Timmerman L "DNA Sequencing Market will Exceed $20 billion, says Illumina CEO Jay Flatley" Forbes, Apr 29, 2015
- * Herper, M. "Flatley's Law: The Company Speeding A Genetic Revolution" Forbes Aug 20th, 2014

Cost per Raw Megabase of DNA Sequence




SRA database growth
5,409,240,568,514,737 total bases
3,180,255,588,572,437 open access bases
Size, terabases
1000
100
10
0
2009
2010
2011
2012
2013
2014
2015
2016
Total bases
Open access bases
http://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/
05/11/2016 06:07am



October 2013 Lex Nederbragt https://flxlexblog.wordpress.com/2013/10/01/developments-in-next-generation-sequencing-october-2013-edition/

Price and Throughput

	read type	price	Average read pair yield
MiSeq v3	Paired End (2x300)	$2,312	22 million
HiSeq 4000	Paired End (2x150)	$3,267	240 million
NextSeq 500	Paired End (2x150)	$6,486	330 million

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

Illumina Sequencing at UTK

- MiSeq in UT Genomics Core

  http://mbrf.utk.edu/

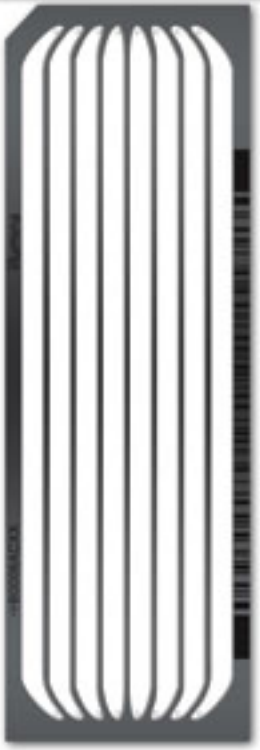
- MiSeq in Center for Environmental Biotechnology

  http://www.ceb.utk.edu/dnasequence.html

Run vs. Lane

- Used interchangeably or as something different?

MiSeq
1 run, 1 input sample



HiSeq
Flow Cell
8 lanes, 8 samples

Illumina Limitations

- Short read length
  - MiSeq (smaller throughput instrument)
    - 2x300
  - HiSeq/NextSeq
    - 2x150
- Bias against sequencing through GC-rich regions or AT-rich regions
- Errors are likely to be SNPs and are likely to cluster at the ends of sequences

Specifics for Illumina

Illumina Sequencing Technology

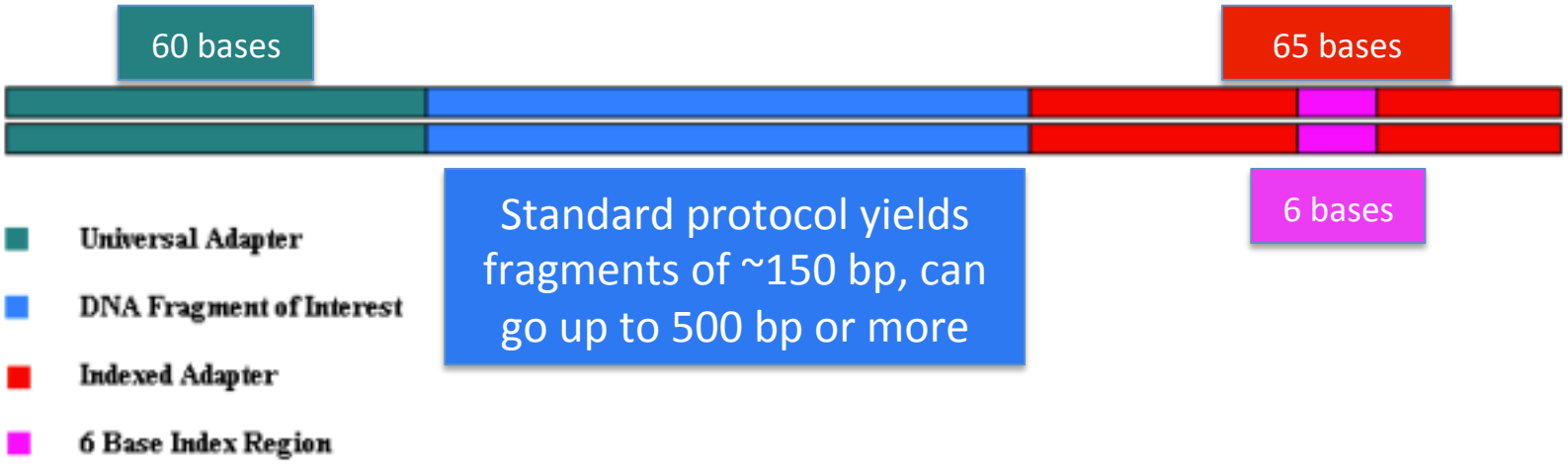
- Video of Illumina Sequencing Technique w Nextera:
  - https://www.youtube.com/watch?v=womKfikWlxM

- Newest sequencing technologies (MiniSeq and NextSeq) use only 2 dyes instead of 4
  - http://www.illumina.com/technology/next-generation-sequencing/sequencing-technology/2-channel-sbs.html

How does it work?

Library construction can vary by kit

TruSeq Example:



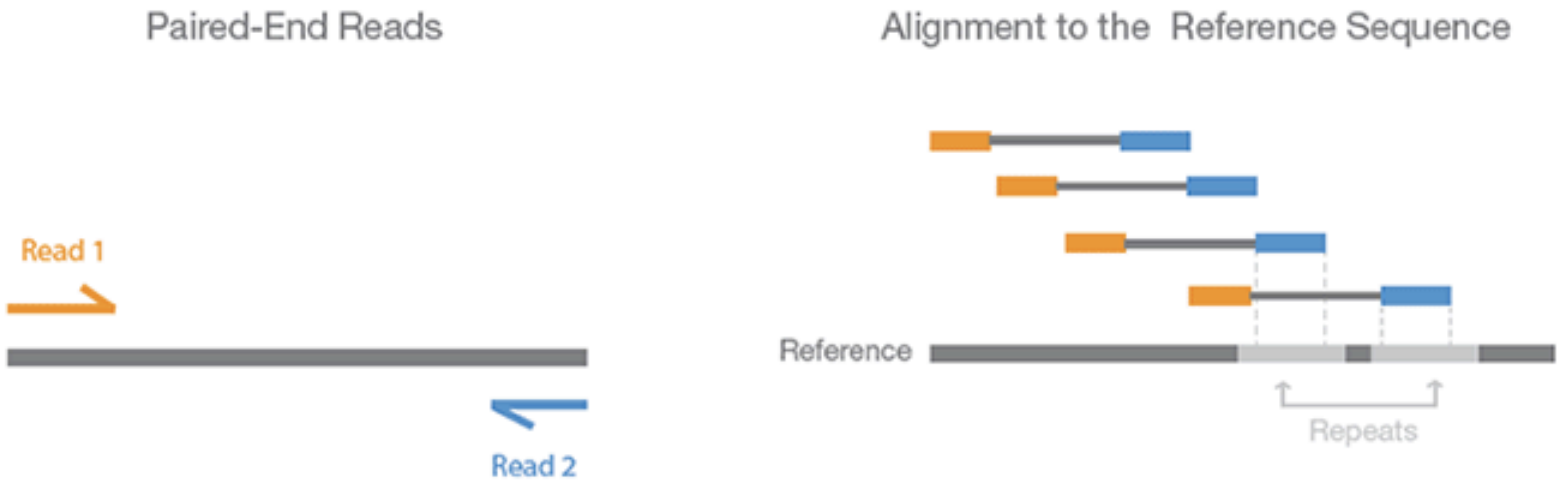
You will need the adapter sequences and a good understanding of adapter locations to later trim them out of your data

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

Paired End Sequencing

Overcome lack of length.

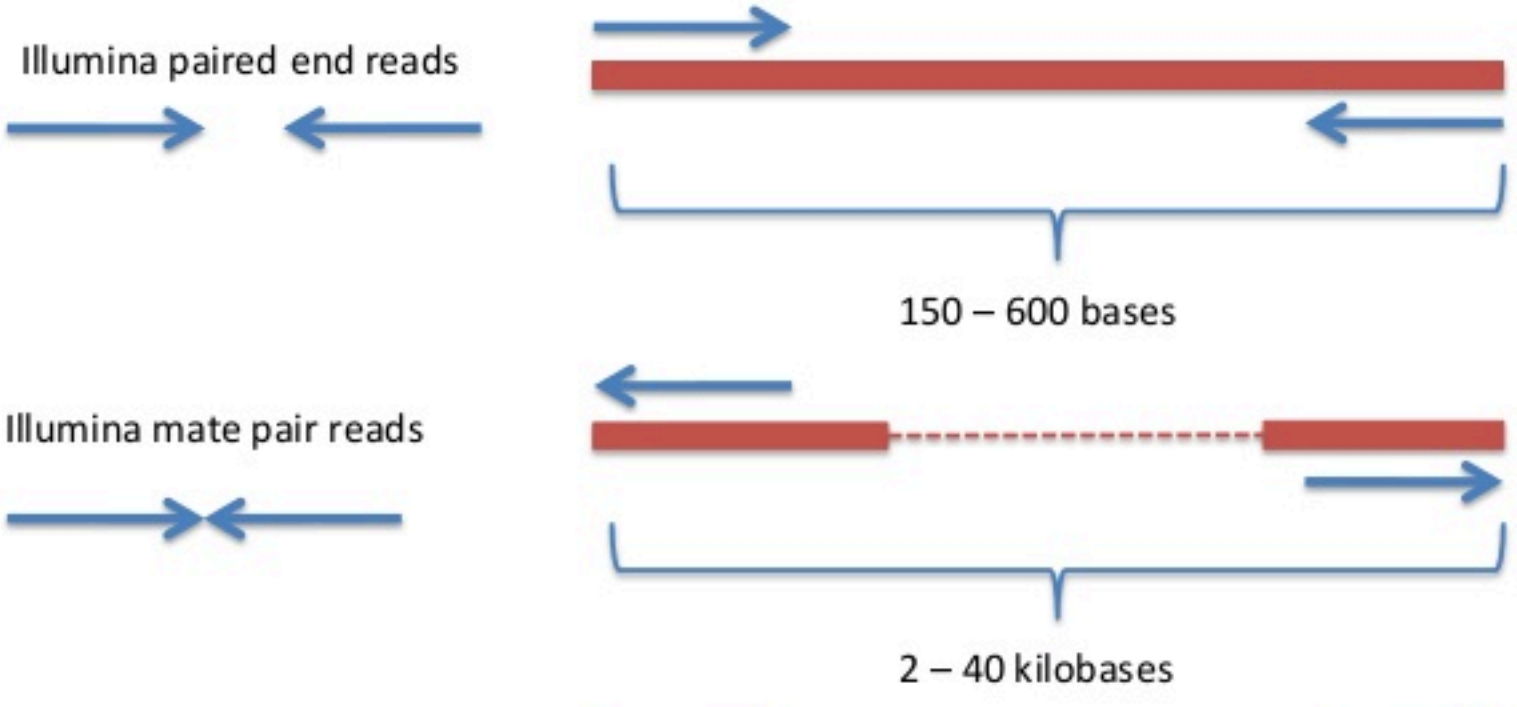
Figure 4. Paired-End Sequencing and Alignment



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.

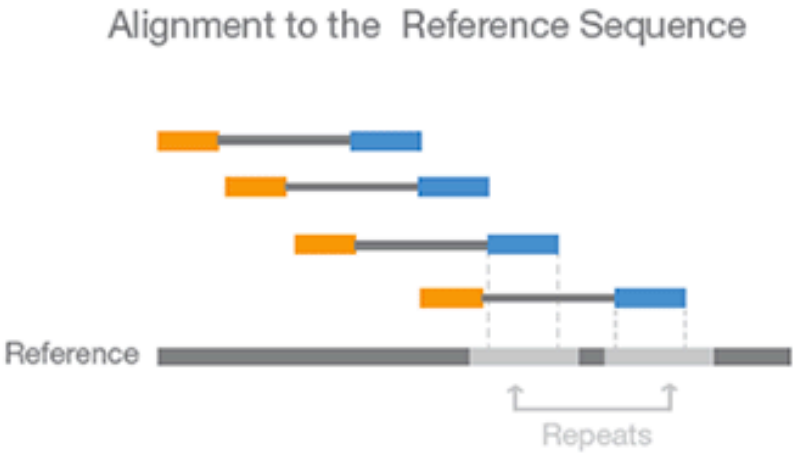
Mate Pair Sequencing

- Even longer jump distances



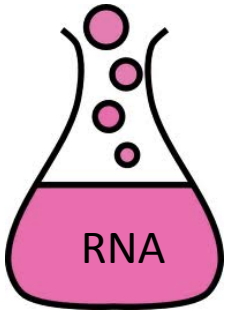
Utility of PE and MP for Repeats

- Can be used to sort out the number of repeats and the sequence of repeats
- Optimally, one end is in a unique region, the other end is in the repetitive region (or both are repetitive)

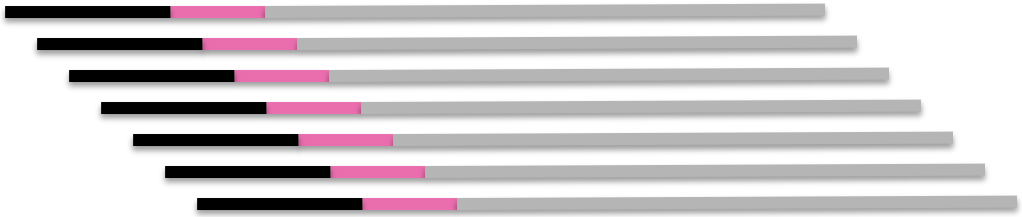


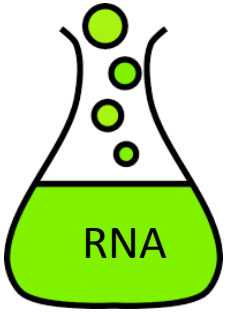
Multiplexing

Loading many samples into one lane.



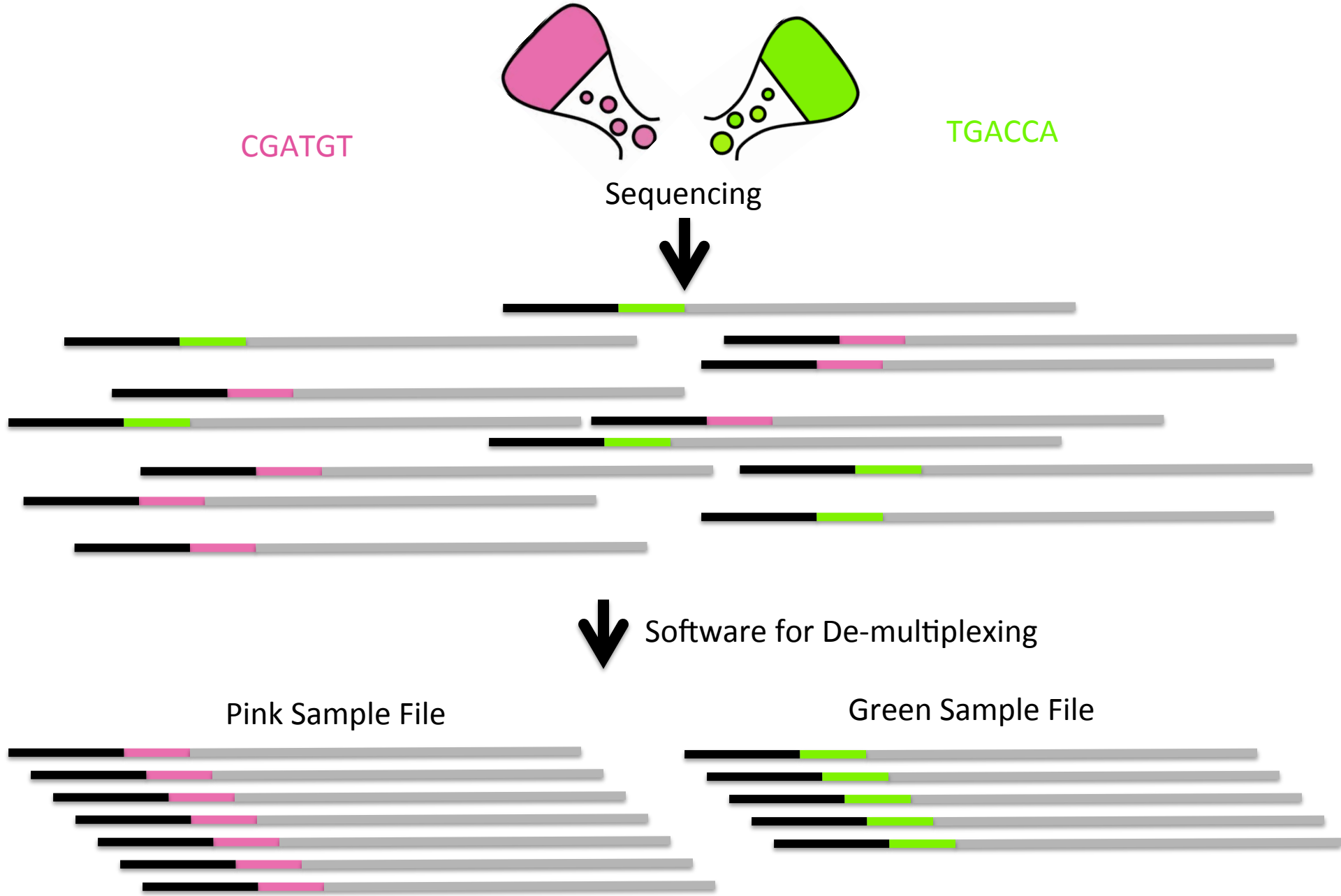
Pink Sample With CGATGT





Green Sample with TGACCA

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



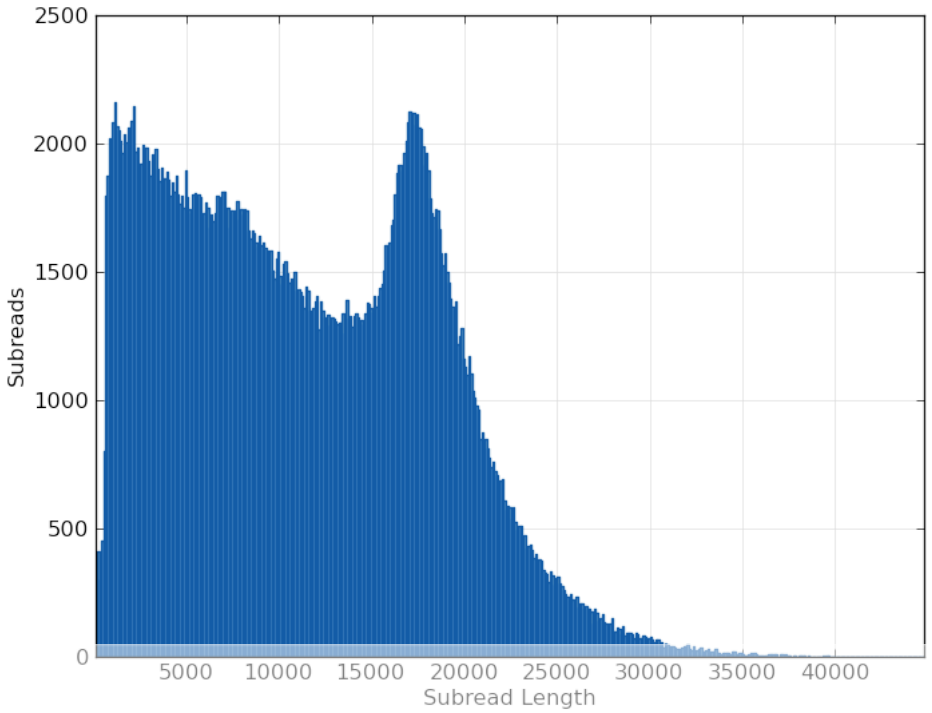
http://www.illumina.com/technology/next-generation-sequencing/multiplexing-sequencing-assay.html

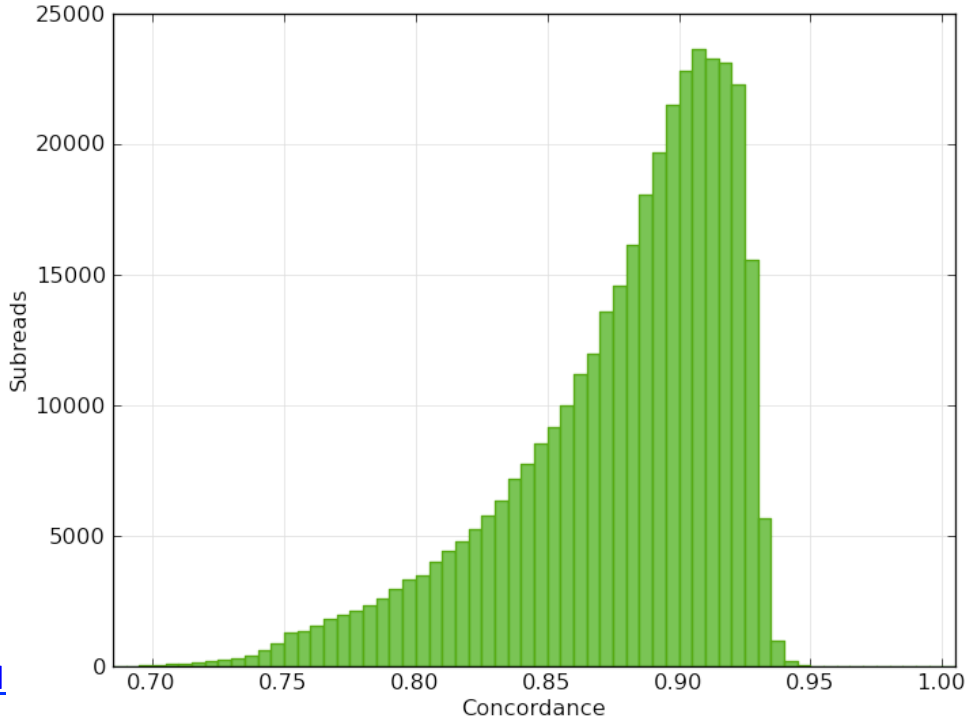
Multiplexing

- Up to 384 barcodes available in some commercial kits
- Can be beneficial to select your barcodes wisely
  - Want an even distribution of nucleotides per cycle
  - If barcodes differ from each other at more than one position, you may be able to assign reads to a sample even with a sequencing error

PACIFIC BIOSCIENCES®

- "PacBio"
- Commercially released 2011
- SMRT = single molecule real time
- No amplification needed
- More expensive
- Very long reads
- 15-20% error
  - Truly random error
- No GC bias
- Indels instead of SNP errors
- Can detect methylation of nucleotides without alterations to the DNA
- Requires a large amount of high quality DNA





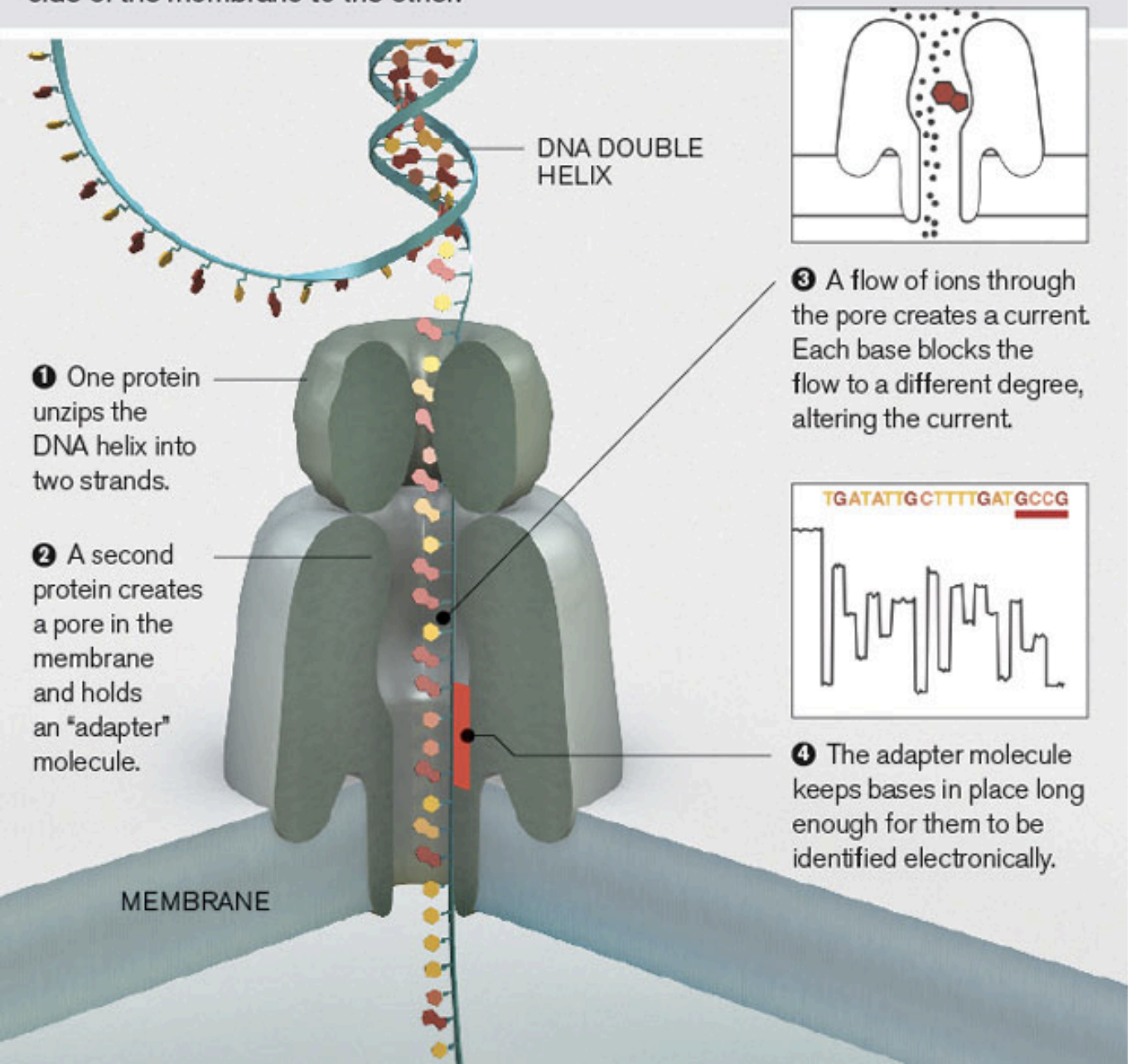
Video: https://www.youtube.com/watch?v=v8p4ph2MAvI

Oxford NANOPORE Technologies®

- MinION
- Extremely different mode for sequencing
  - Tiny (handheld) instruments and disposable chips
  - No fluorescence and no polymerase
  - uses ion current disruptions while “unzipping” the DNA
  - No fixed run time
  - Long reads, can read each strand twice

Does still require library prep, but field based applications are very exciting!

DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.



Video: https://www.youtube.com/watch?v=3UHw22hBpAk

First Step to Data Analysis

Preprocessing

- Turning the raw signal of the instrument into base calls and quality scores
- Images are converted to text and numbers
- Supplied by vendor-provided software “on rig” (i.e. with the computer supplied with the equipment)
- Image files are usually not kept because of their size
- You usually will not have to deal with this, but best to be aware that it is happening

Additional Steps in Data Analysis

- Why would we want to learn about the sequencing chemistry to do bioinformatics?
  - Understand biases
  - Spot nonsensical results
  - Plan for robust statistical analysis
  - Good experimental design
- Always get the sequences for your adapters and barcodes for later analysis!
  - Illumina has an open letter with most of their adapters:
  - https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-customer-sequence-letter.pdf
  - Kits should come with additional documentation

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

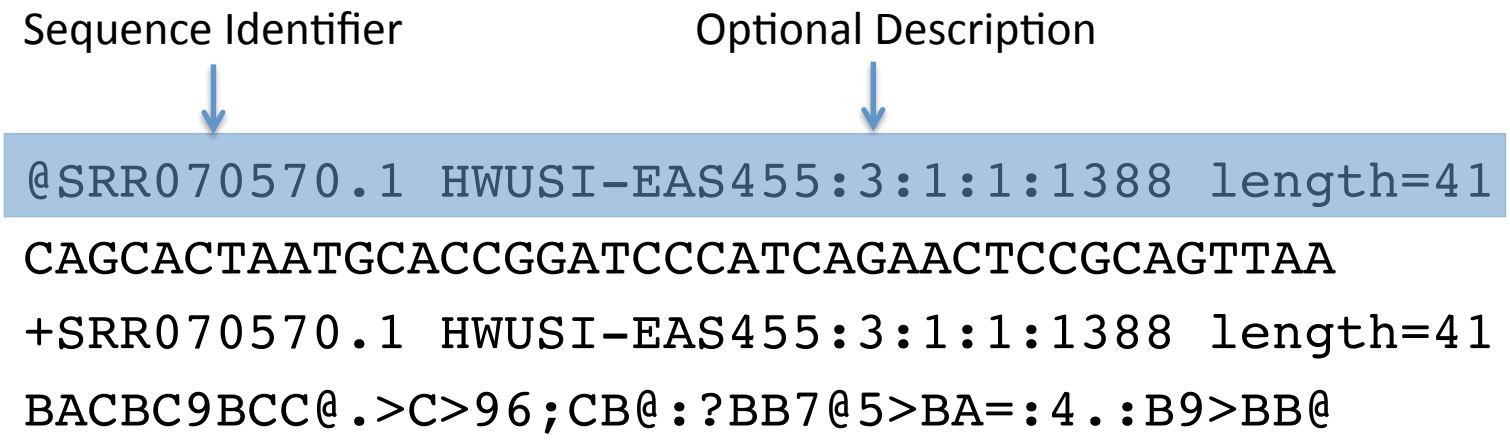
- May be amino acid or nucleotide
- May be a single line or multiple lines

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 Values

- Based on phred quality scoring system (developed in the 1990s)

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%

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

FASTQ Quality Values

- Storing it based on the numbers 0-60 takes up too much space
- Sequence: `ACTGATC`
- Quality: `10 15 25 15 17 32 35`

- Instead, assign individual letters, numbers and symbols to represent numeric quality values
- No need for space
- New Quality: `JOXOLb`

The Quality Value Debacle

Or, an excellent example of how NOT to create a standard format (and make your bioinformaticians weep)

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                          |    |        |                              |                     |
33                         59   64       73                            104                   126
 0........................26...31.......40
                           -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

See wiki page for converters. http://en.wikipedia.org/wiki/FASTQ_format

Lesson

- Python: Lists and loops