Announcements

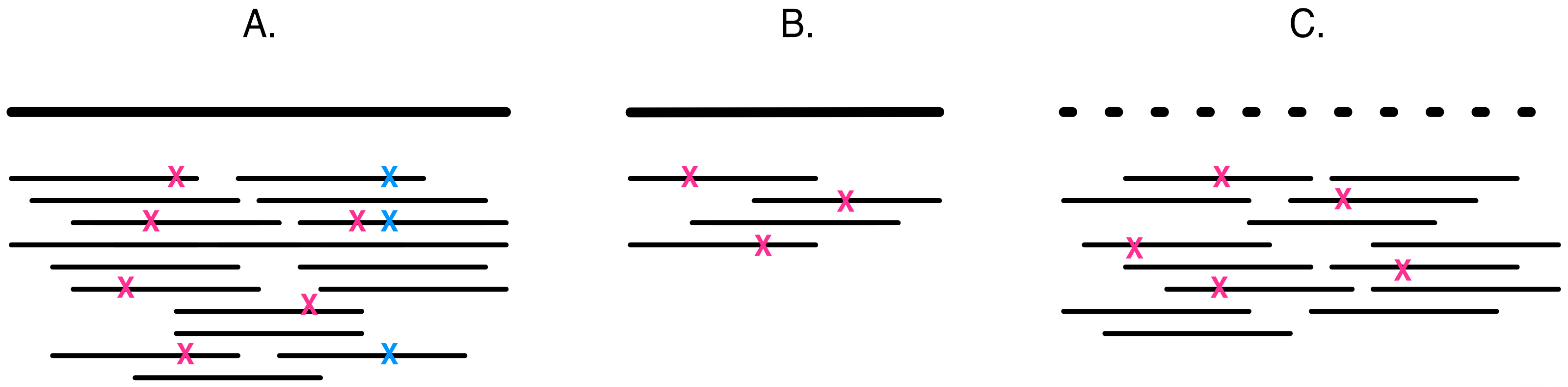
- Coffee!
- Evaluation.

Dr. Yoshiki Sasai, R.I.P.

Sequencing considerations

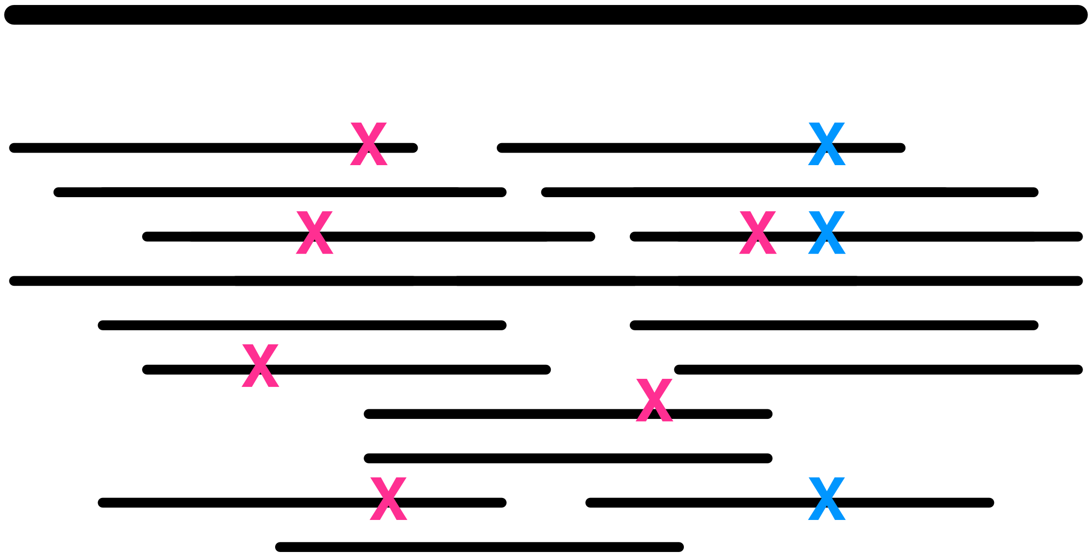
Three basic problems

Resequencing, counting, and assembly.



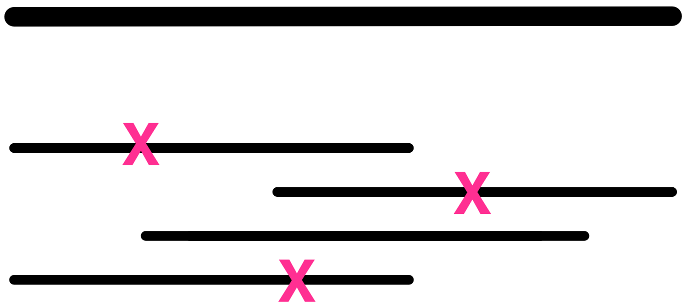
1. Resequencing analysis

We know a reference genome, and want to find *variants* (blue) in a background of errors (red)



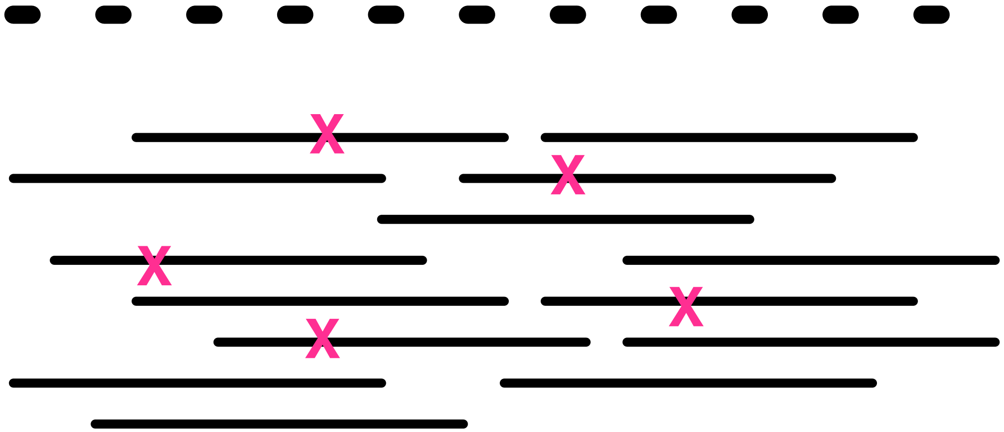
2. Counting

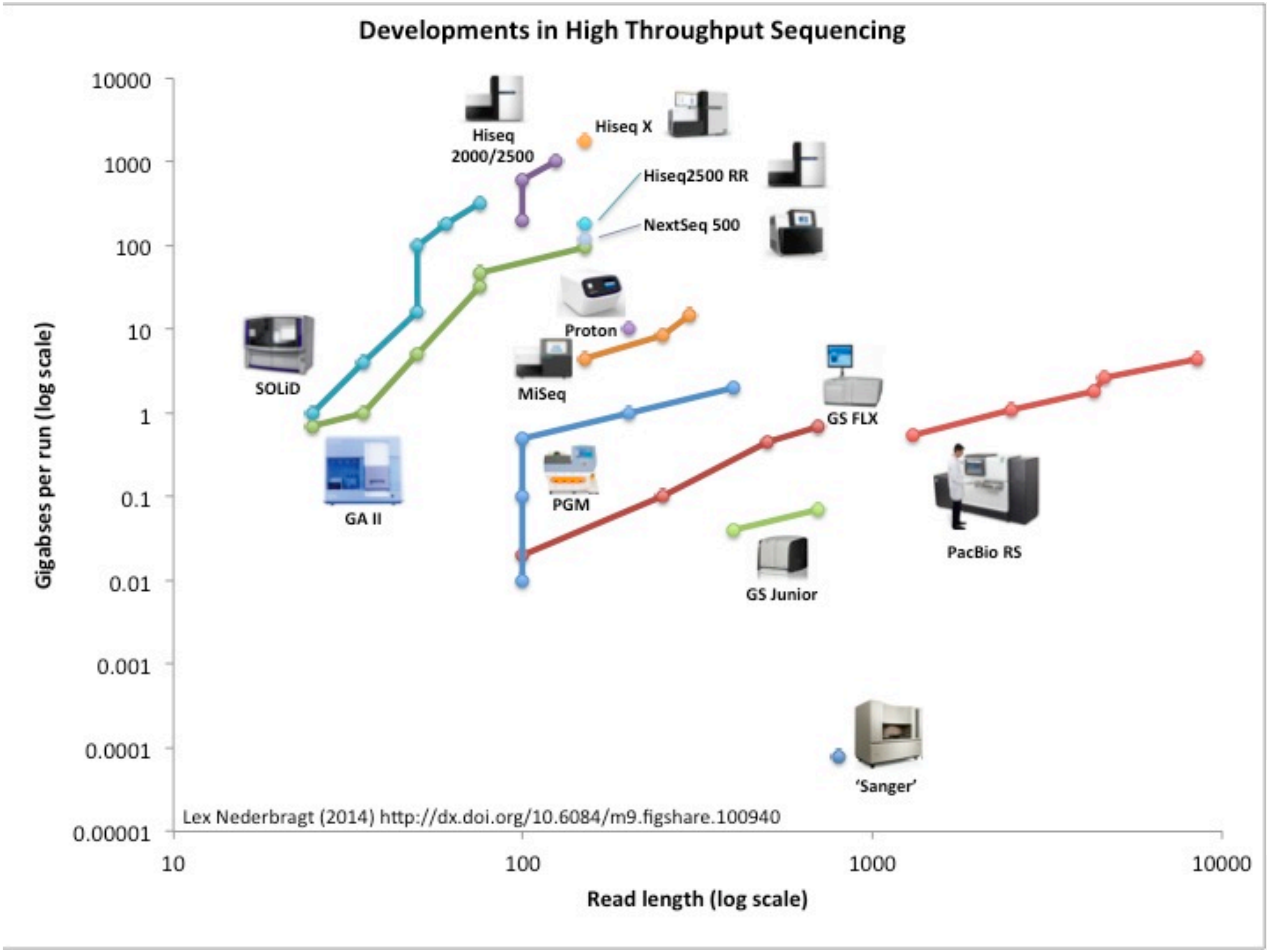
We have a reference genome (or gene set) and want to know how *much* we have. Think gene expression/ microarrays.



3. Assembly

We don't have a genome or any reference, and we want to construct one.
(This is how all new genomes are sequenced.)





Developments in High Throughput Sequencing

Lex Nederbragt (2014) http://dx.doi.org/10.6084/m9.figshare.100940

Outline

- Shotgun sequencing
- The magic of polonies, and how Illumina sequencing works
- Sequencing depth, read length, and coverage
- Paired-end sequencing and insert sizes
- Coverage bias
- Long reads: PacBio and Nanopore sequencing

Shotgun sequencing

It was the best of times, it was the worst of times, it was
the age of wisdom, it was the age of foolishness

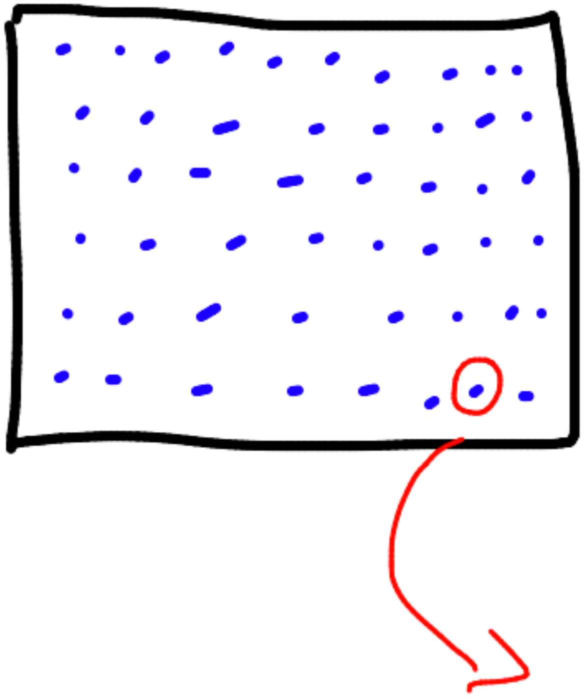
It was the best of times, it was the wor

, it was the worst of times, it was the

isdom, it was the age of foolishness

mes, it was the age of wisdom, it was th

Ion Torrent



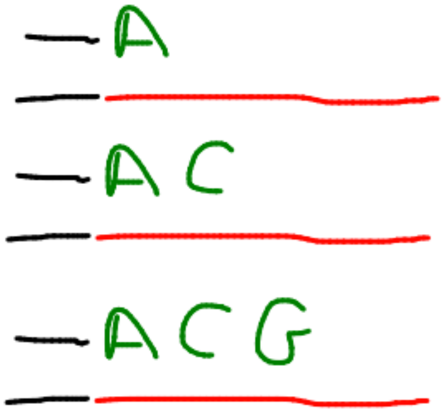
1e8 wells

Each one is a mini
pH meter

Add A

Did H get released
for this well?

Yes? Then next
base was A.



~ 6 hrs for sample
prep plus run => data

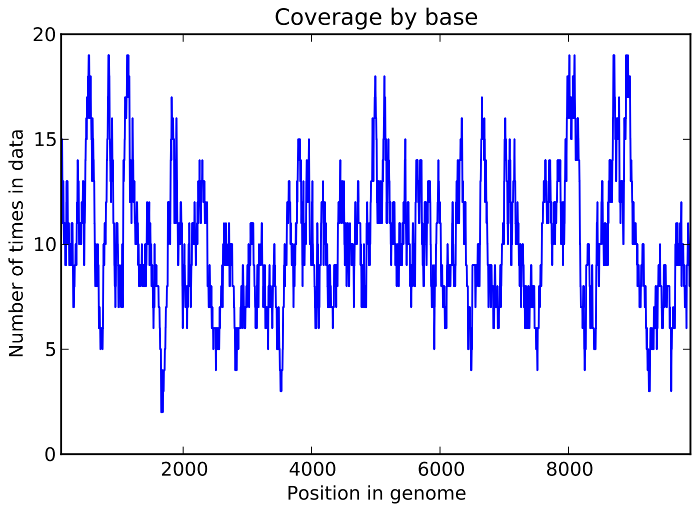
$500 or so.

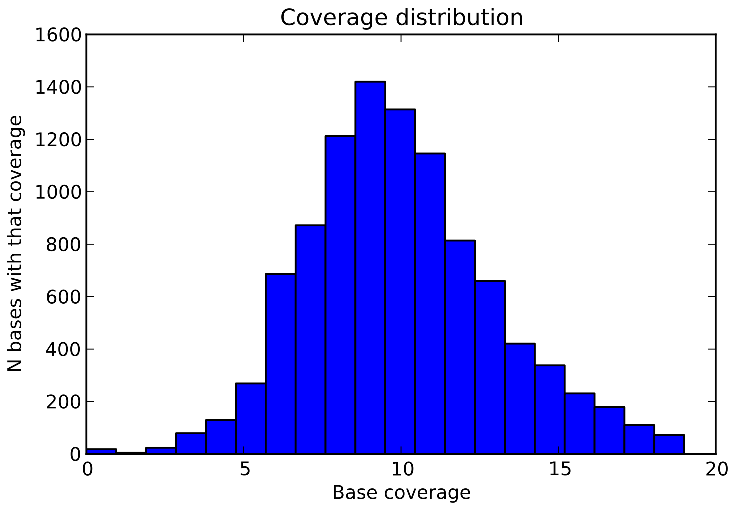
Two specific concepts:

- First, sequencing everything **at random** is very much easier than sequencing a specific gene region. (For example, it will soon be easier and cheaper to shotgun-sequence all of *E. coli* then it is to get a single good plasmid sequence.)

- Second, if you are sequencing on a 2-D substrate (wells, or surfaces, or whatnot) then any increase in **density** (smaller wells, or better imaging) leads to a **squared** increase in the number of sequences yielded.

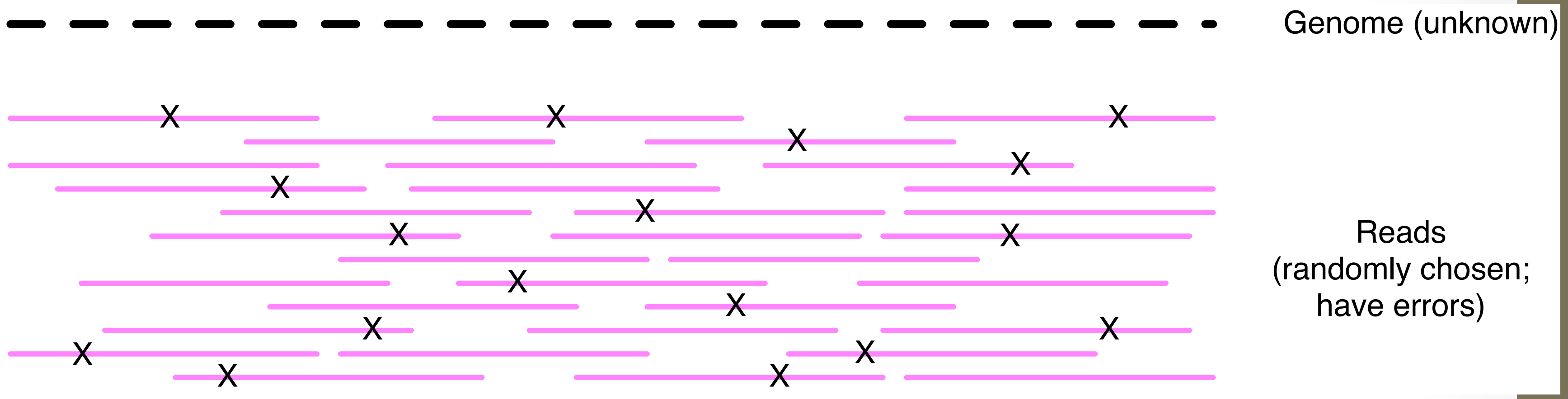
Random sampling => deep sampling needed





Typically 10-100x needed for robust recovery (300 Gbp for human)

“Coverage”



“Coverage” is simply the average number of reads that overlap each true base in genome.

Here, the coverage is ~10 – just draw a line straight down from the top through all of the reads.

Illumina yields the *deepest* sequencing available

- MiSeq
  - 30 million reads per run
  - 300 base paired-end reads

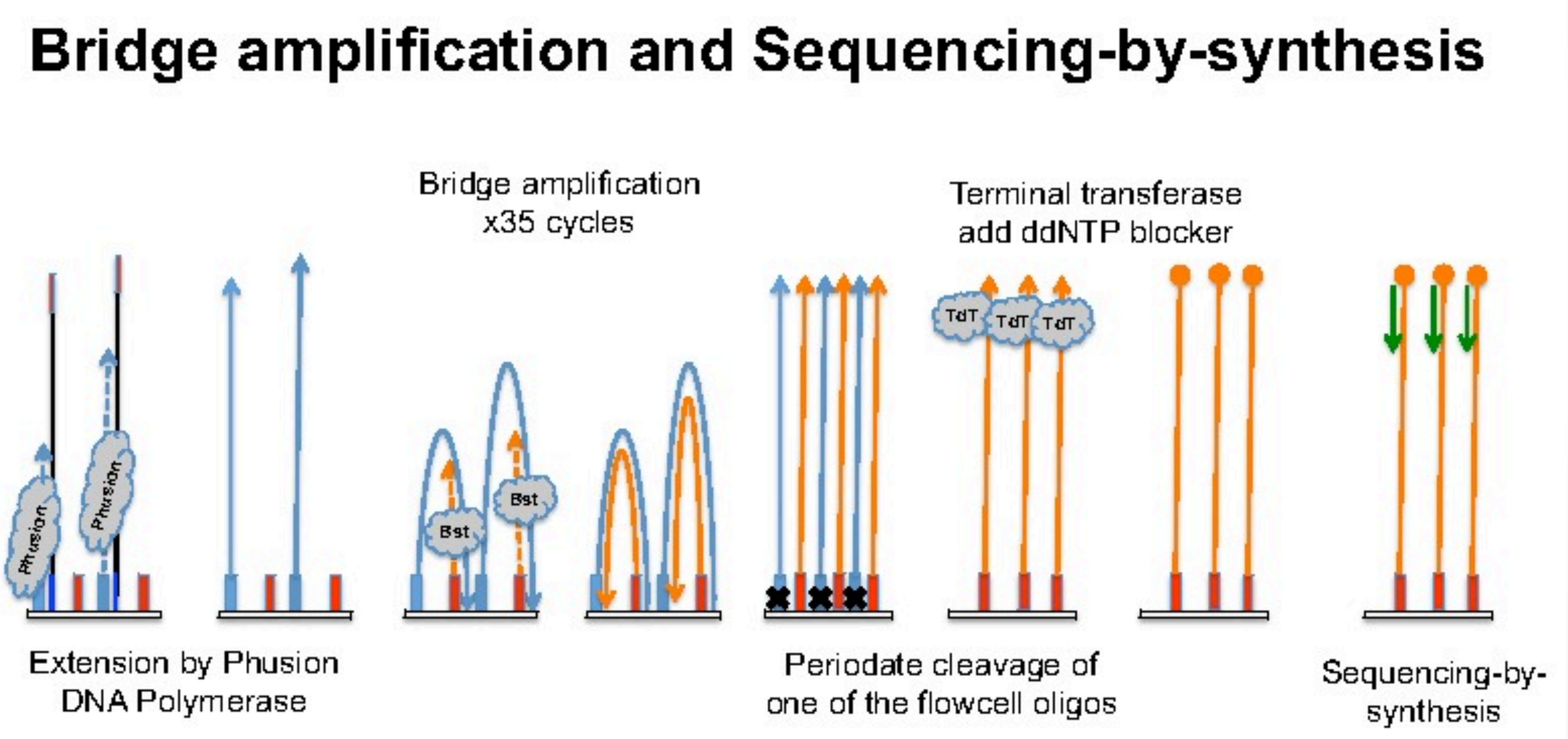
- HiSeq 2500 RR/X 10
  - 6 billion reads per run
  - 150 base paired-end reads

- PacBio
  - 44,000 reads per run
  - 8500 bp in length

http://flxlexblog.wordpress.com/2014/06/11/developments-in-next-generation-sequencing-june-2014-edition/

Illumina basics

(See http://seqanswers.com/forums/showthread.php?t=21 for details)



Bridge amplification and Sequencing-by-synthesis

http://ted.bti.cornell.edu/cgi-bin/epigenome/method-1.cgi

A movie of Illumina sequencing:

https://www.youtube.com/watch?v=tuD-ST5B3QA#t=61

FASTQ

- @895:1:1:1246:14654/1
- CAGGCGCCCACCACCGTGCCCTCCAACCTGATGGT
- +
-][aaX__aa[`ZUZ[NONNFNNNNNO_____^RQ_
- @895:1:1:1246:14654/2
- ACTGGGCGTAGACGGTGTCCTCATCGGCACCAGC
- +
- \UJUWSSV[JQQWNP]]SZ]ZWU^]ZX][^TXR`
- @895:1:1:1252:19493/1
- CCGGCGTGGTTGGTGAGGTCACTGAGCTTCATGTC
- +
- OOOKONNNNN__`R]O[TGTRSY[IUZ]]]__X__

Read length and reconstructability

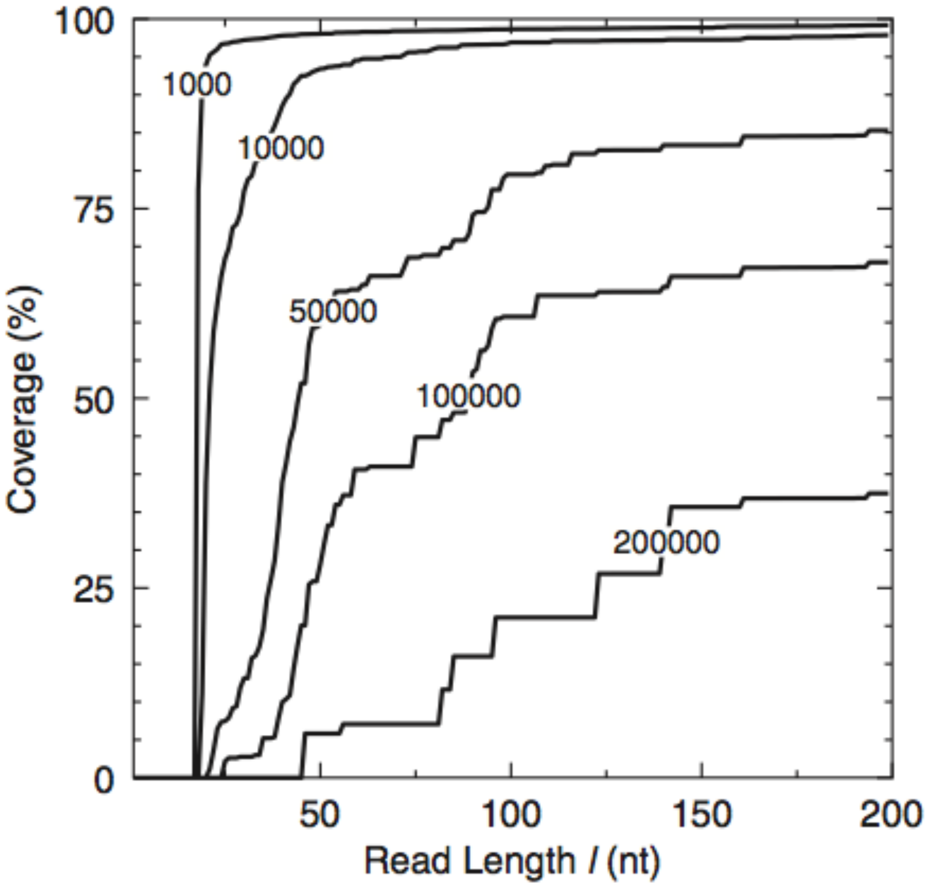


Figure 3. Percentage of the *E.coli* genome covered by contigs greater than a threshold length as a function of read length.

Whiteford et al., Nuc. Acid Res, 2005

“Reconstructability”

- Assembling new genomes or transcriptomes…

- *Haplotyping* - think human genetics & viruses, both.

Real problem? Our data can't uniquely specify solution!

Here, there is no direct way to know if last exon is connected to first exon.

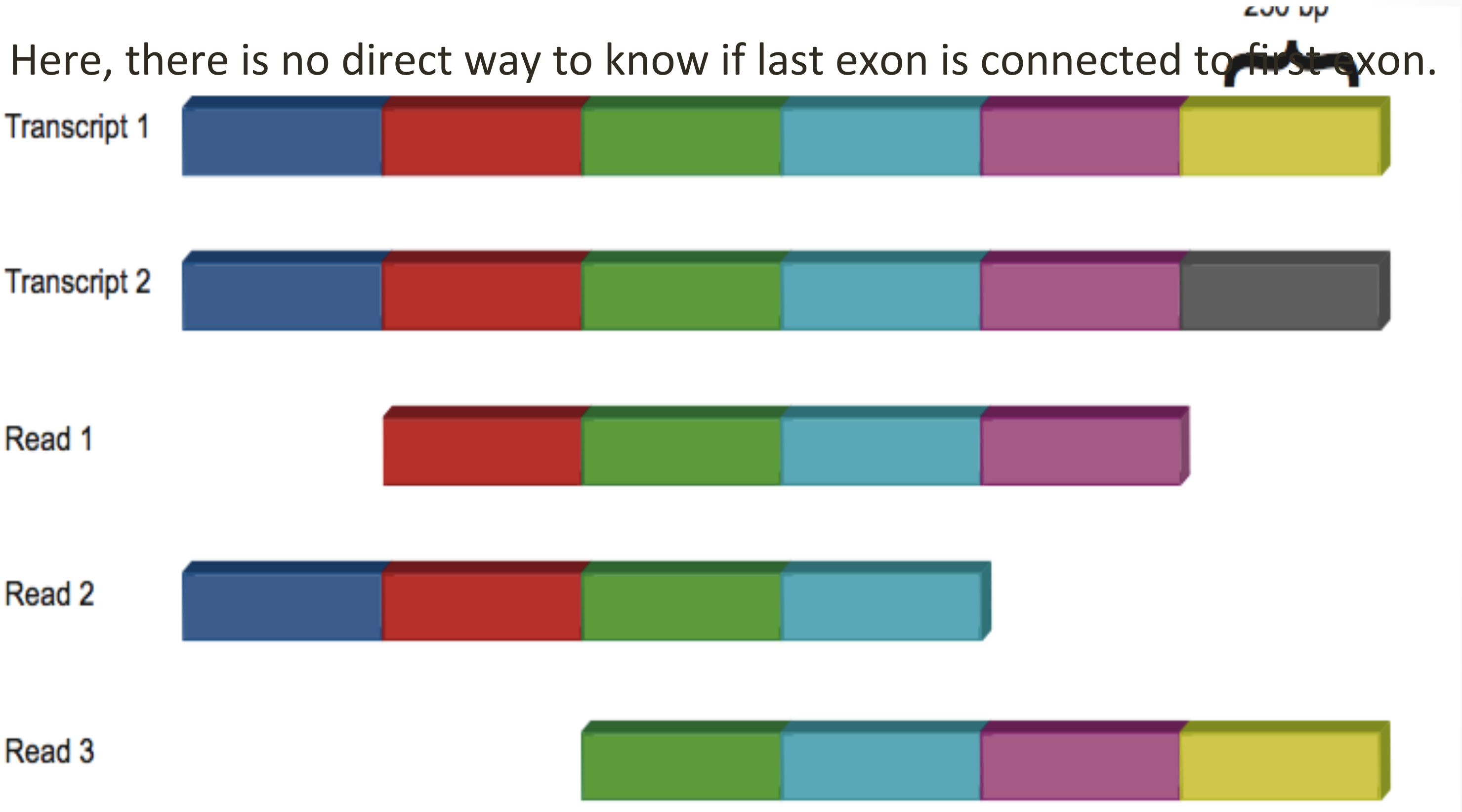
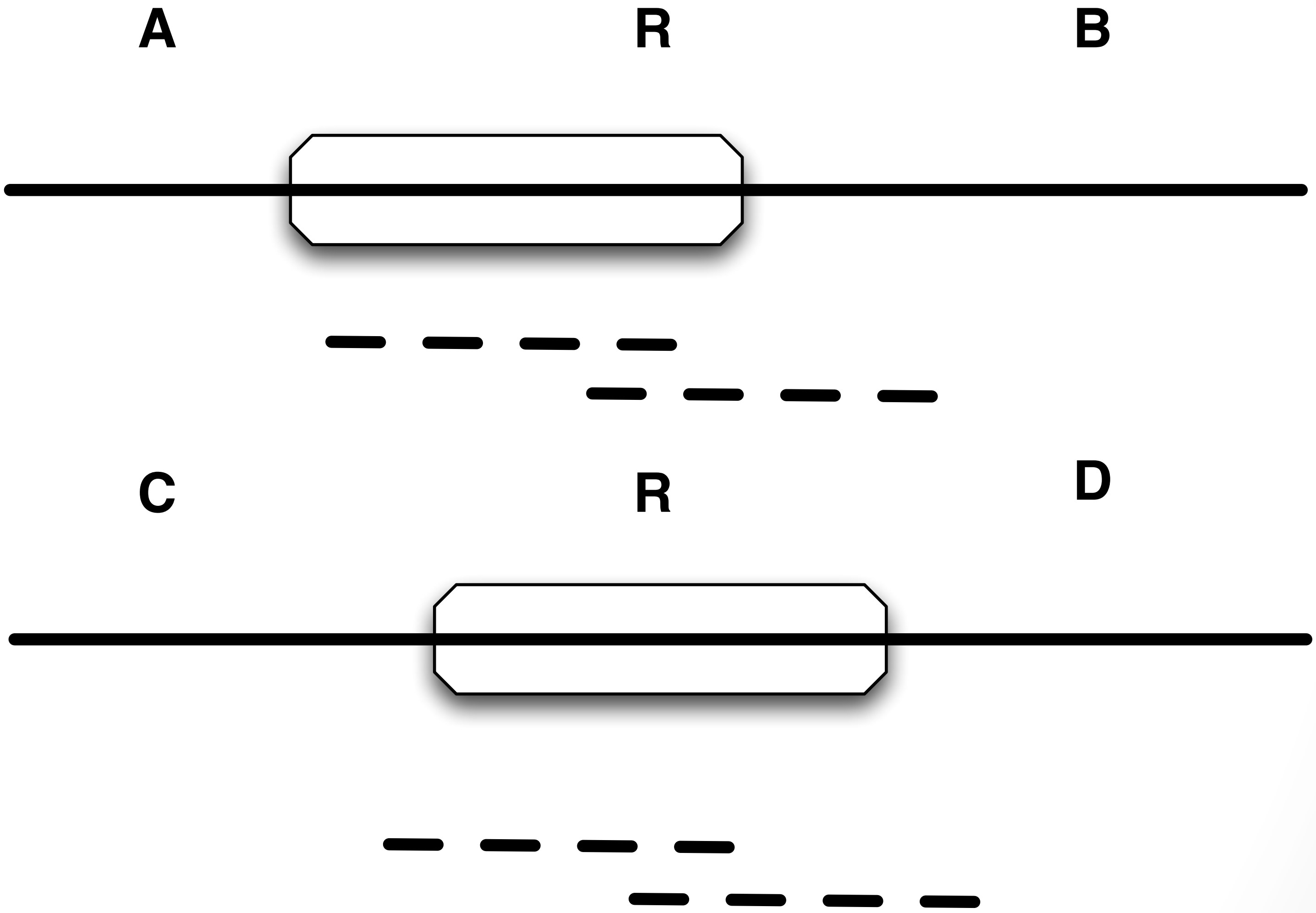


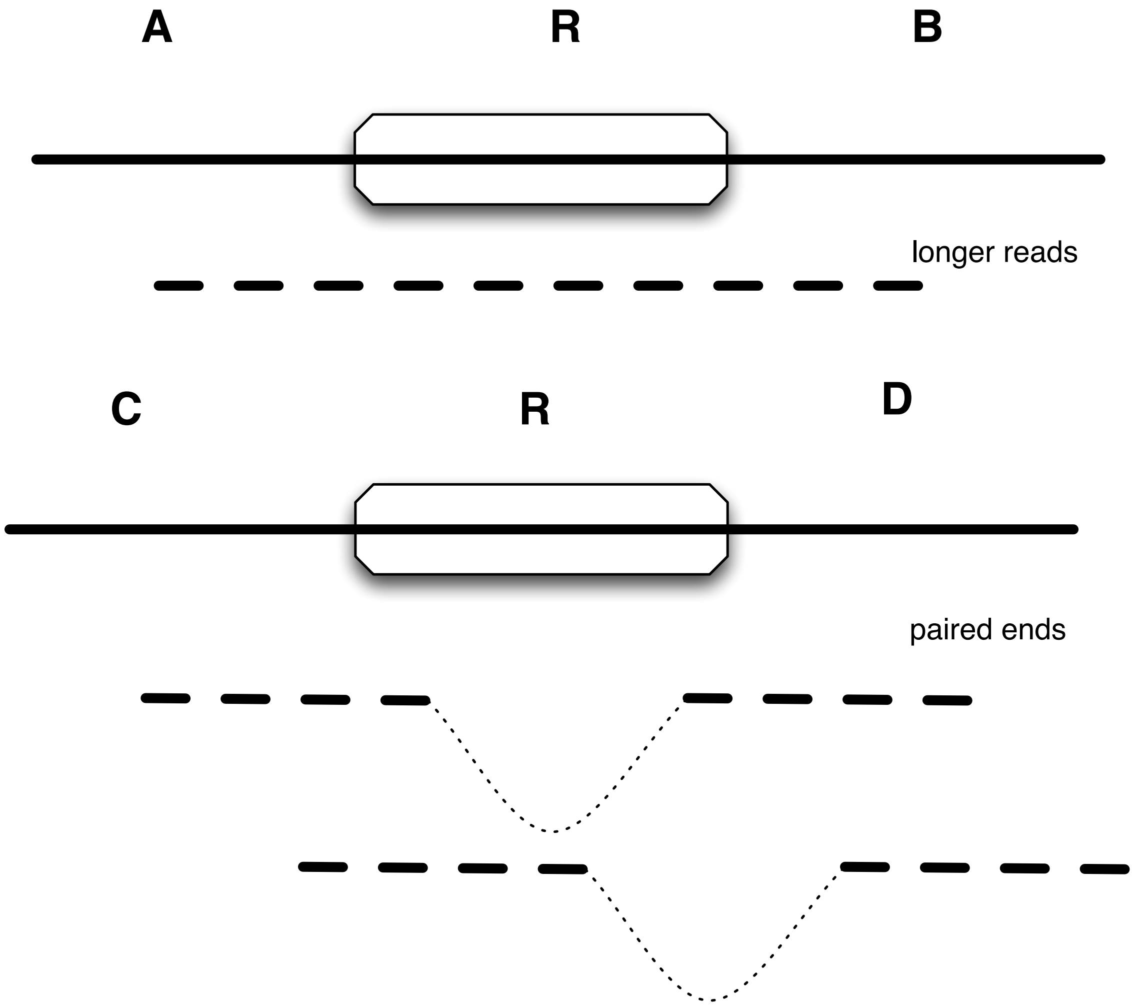
Figure 6. Hypothetical example of 1 kb multimap reads. Only Read 3 can be uniquely

Pyrkosz et al., unpub.

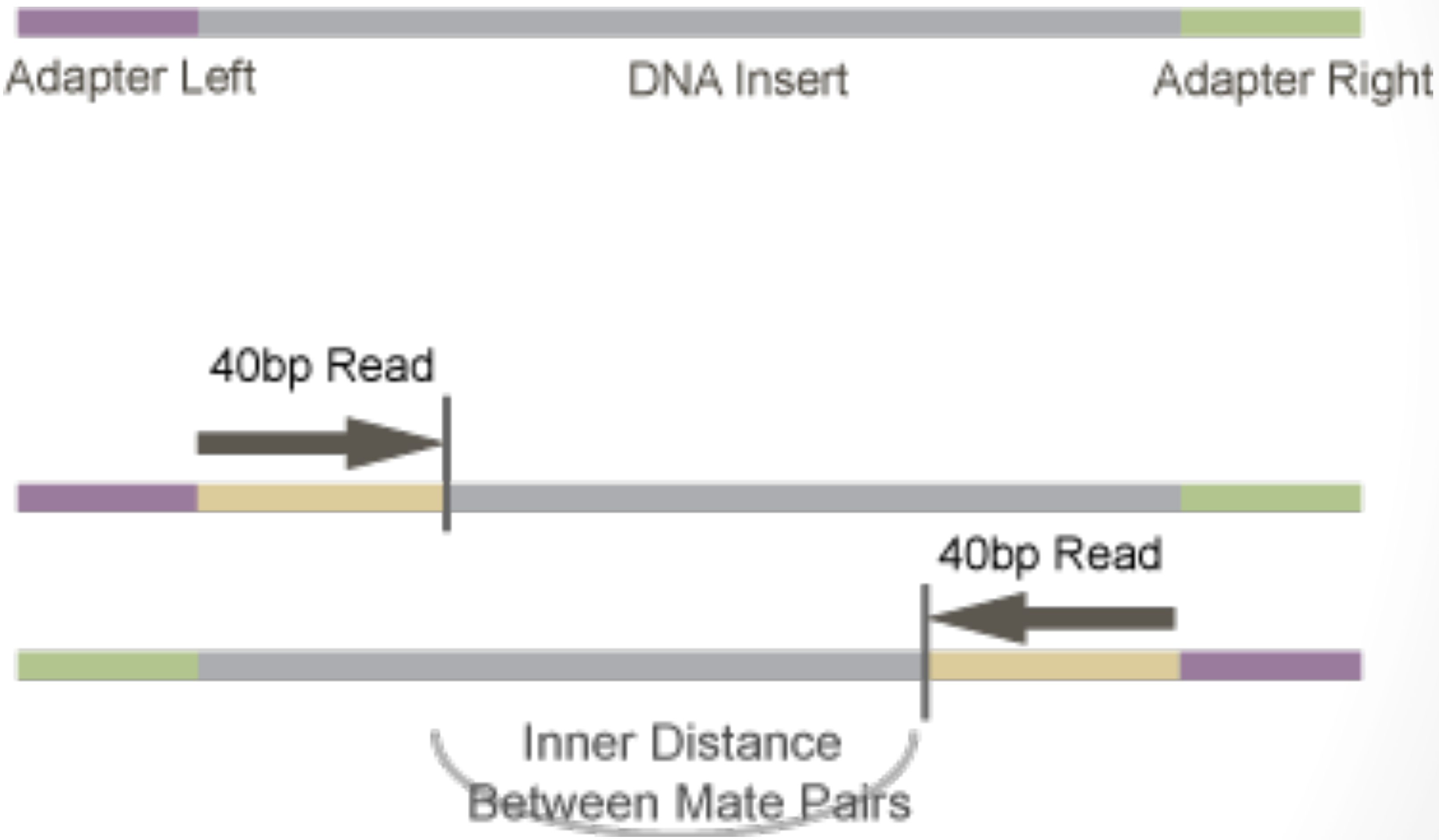
Repeats! (and shared exons)



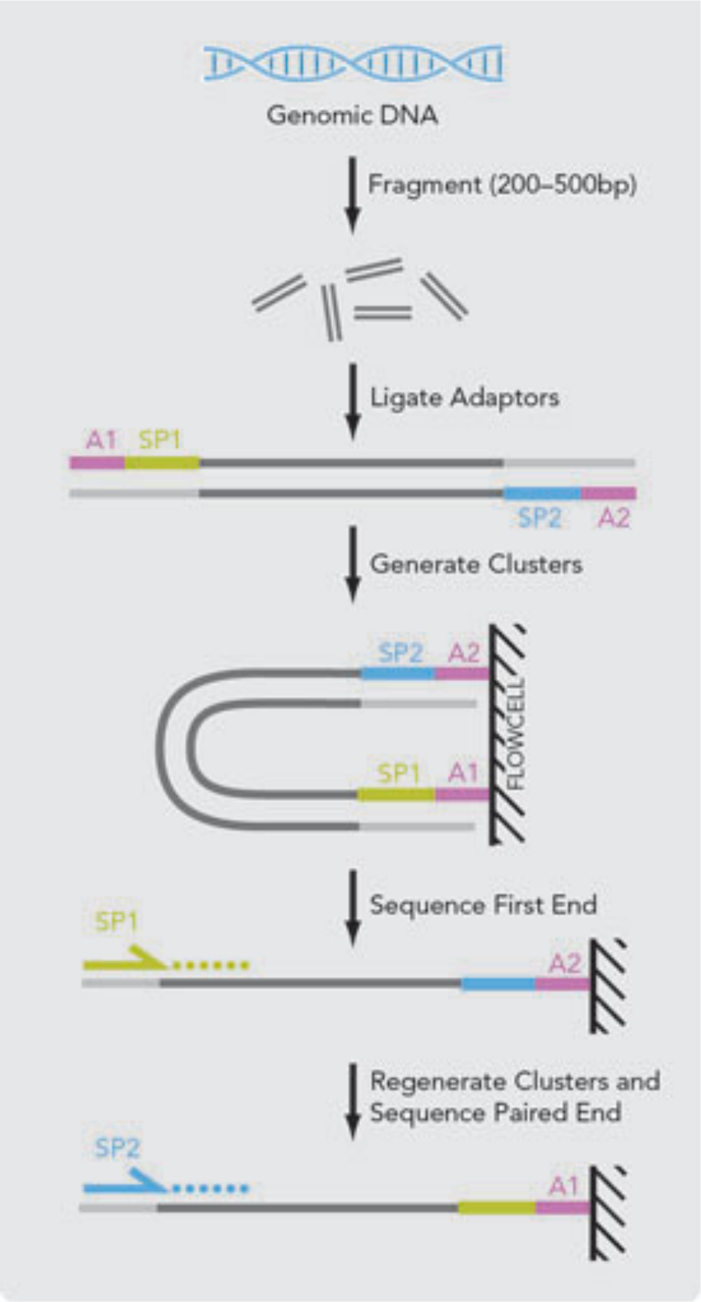
Longer reads ... OR ...
Paired-end/mate pair sequencing

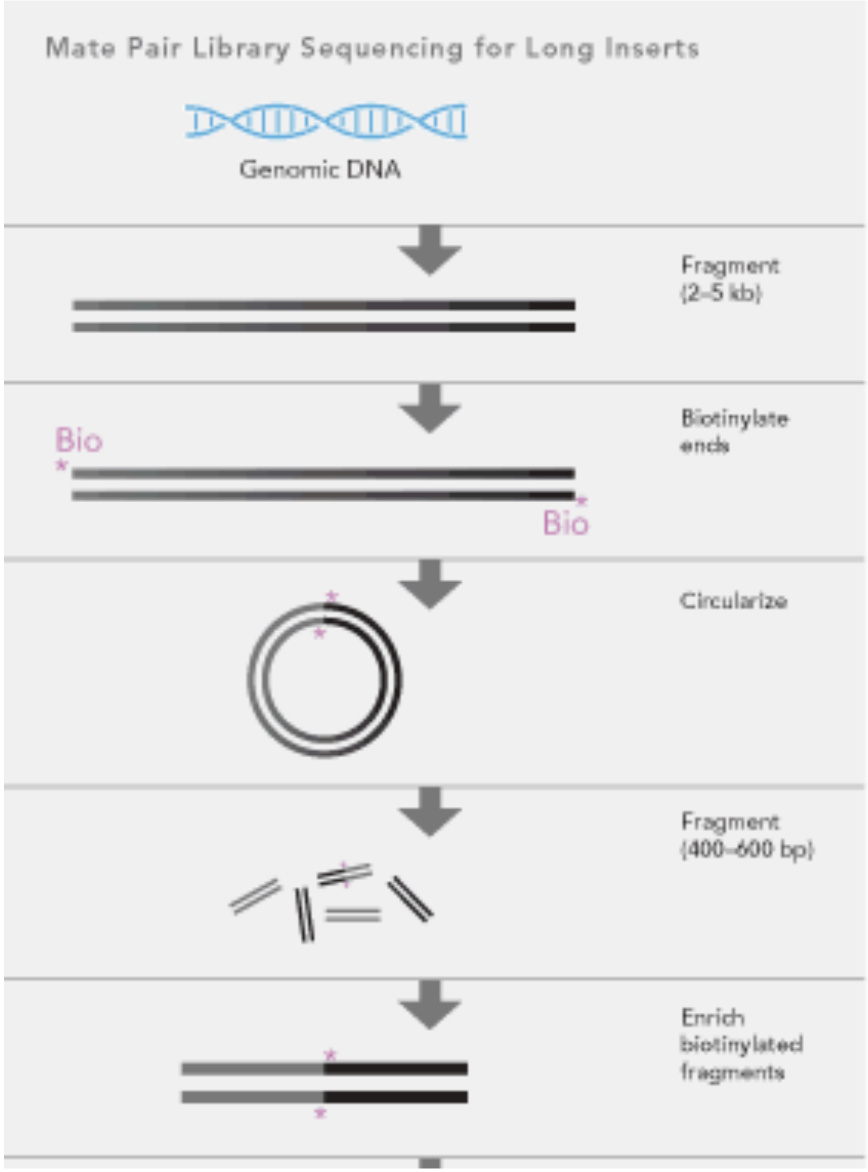


Paired-end sequencing



http://vallandingham.me/RNA_seq_differential_expression.html


Genomic DNA
Fragment (200–500bp)
Ligate Adaptors
A1
SP1
SP2
A2
Generate Clusters
SP2
A2
SP1
A1
FLOWCELL
Sequence First End
SP1
A2
Regenerate Clusters and
Sequence Paired End
SP2
A1

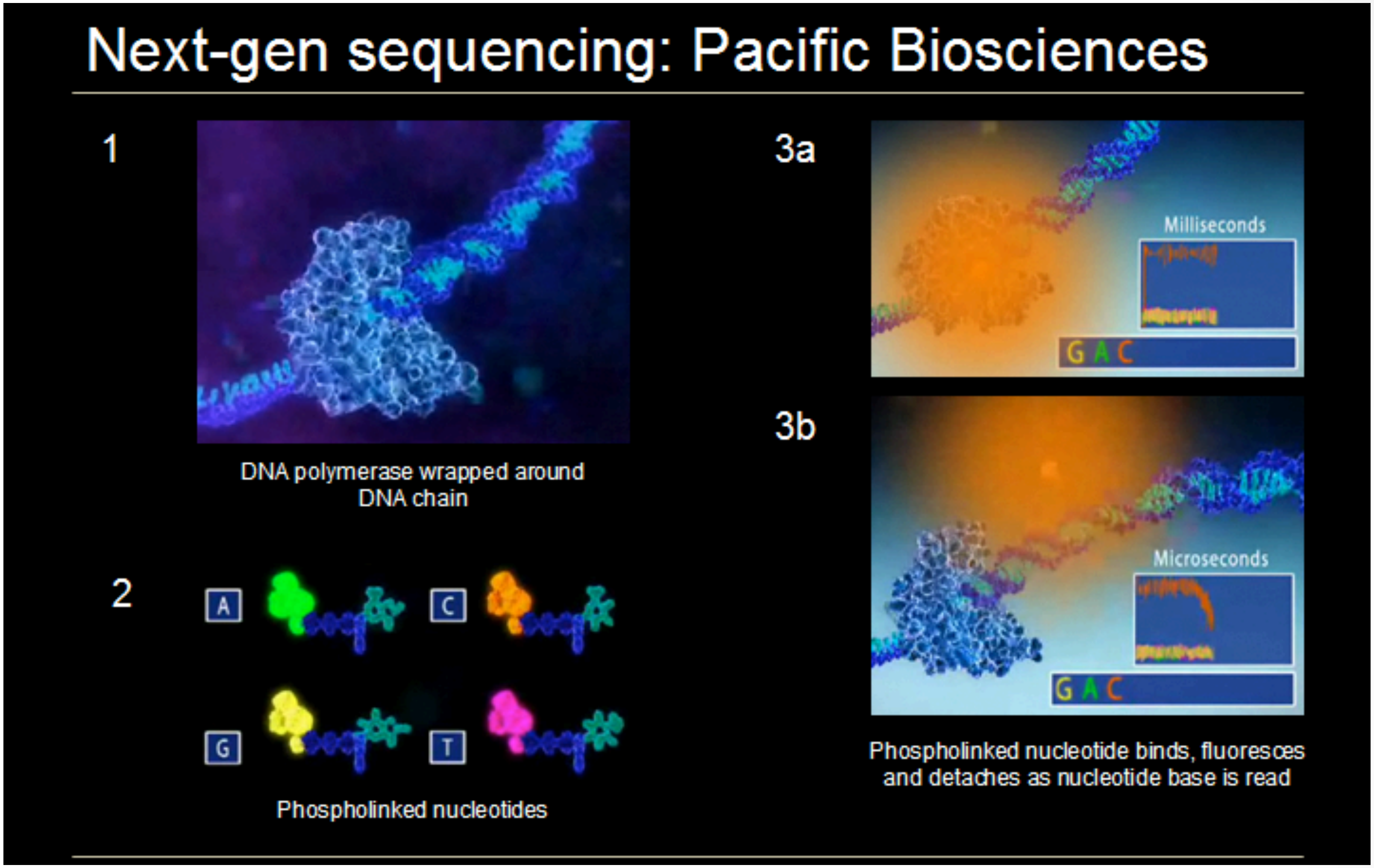


Mate Pair library preparation is designed to generate short fragments that consist of two segments that originally had a separation of several kilobases in the genome. Fragments of sample genomic DNA are end-biotinylated to tag the eventual mate pair segments.
Self-circularization and refragmentation of these large fragments generates a population of small fragments, some of which contain both mate pair segments with no intervening sequence. These Mate Pair fragments are enriched using their biotin tag. Mate Pairs are sequenced using a similar two-adapter strategy as described for paired-end sequencing.

Mate-pair sequencing (long insert)

Longer reads

- PacBio
- Moleculo
- Nanopore



Next-gen sequencing: Pacific Biosciences

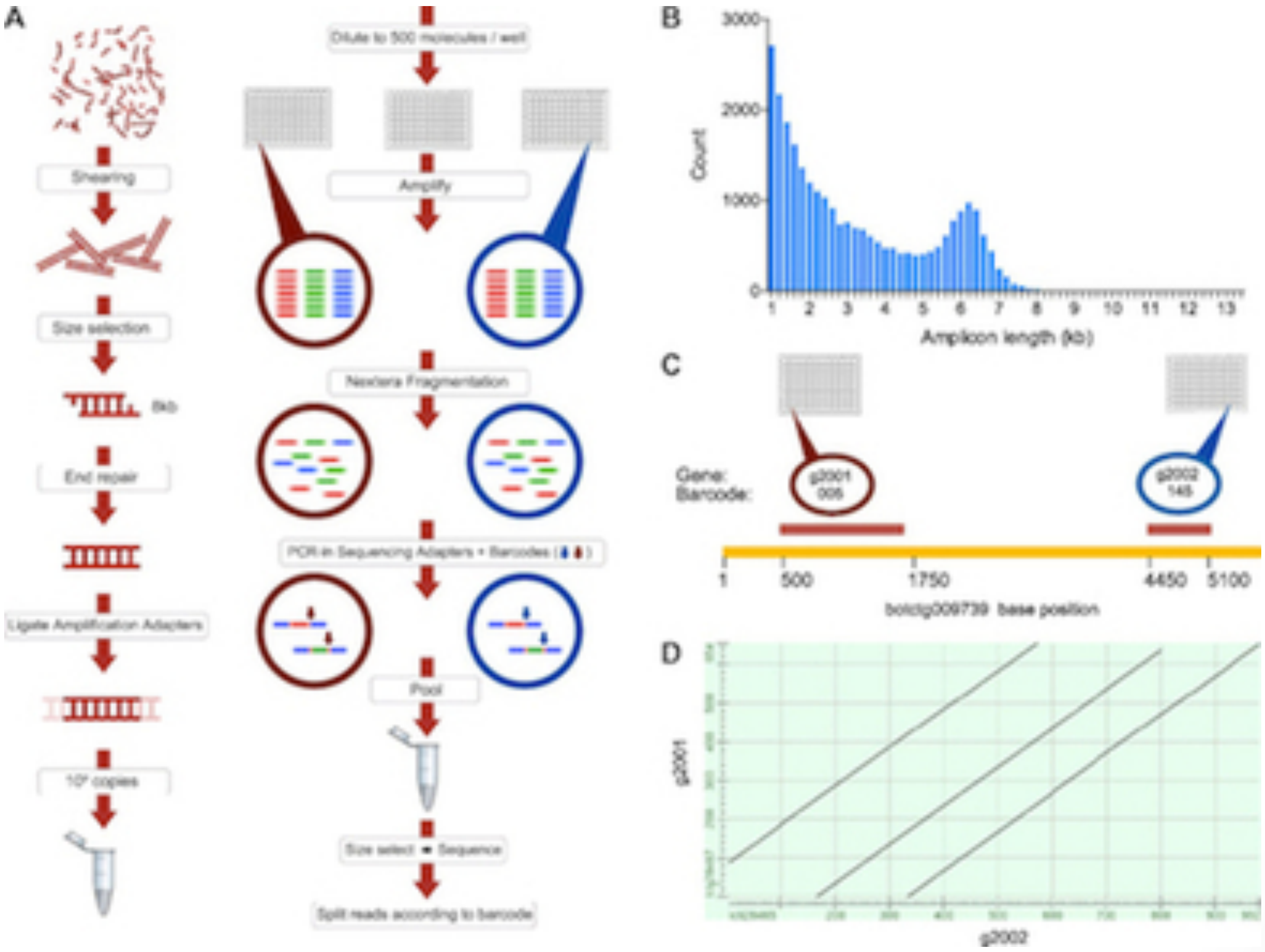
DNA polymerase wrapped around DNA chain

Phospholinked nucleotides

Phospholinked nucleotide binds, fluoresces and detaches as nucleotide base is read

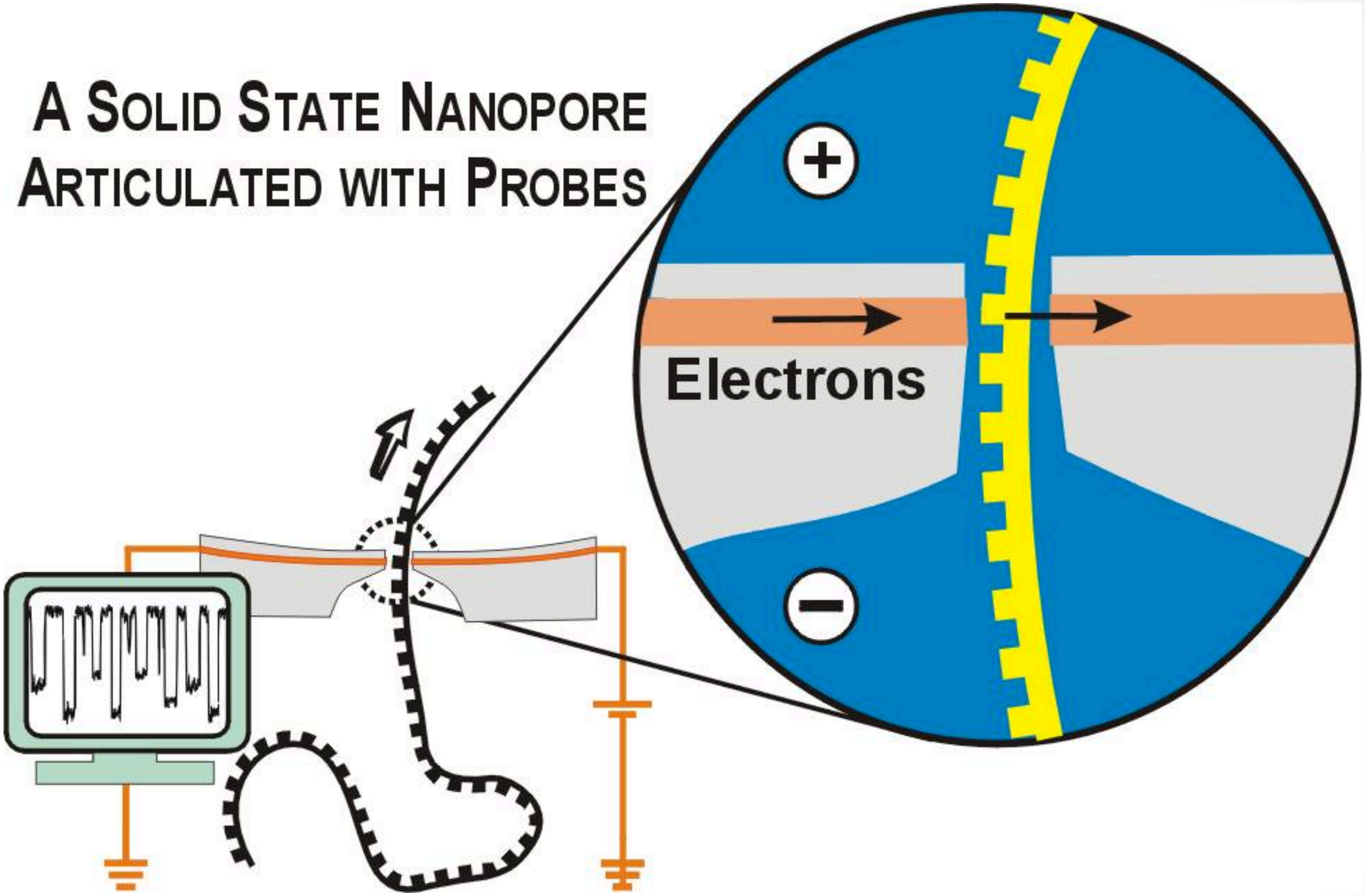
http://www.melanieswan.com/FOLS.html

Moleculo (Illumina)



http://nextgenseek.com/2013/07/illumina-announces-moleculo-long-read-technology-and-phasing-as-service/

A Solid State Nanopore Articulated with Probes



http://labs.mcb.harvard.edu/branton/projects-NanoporeSequencing.htm

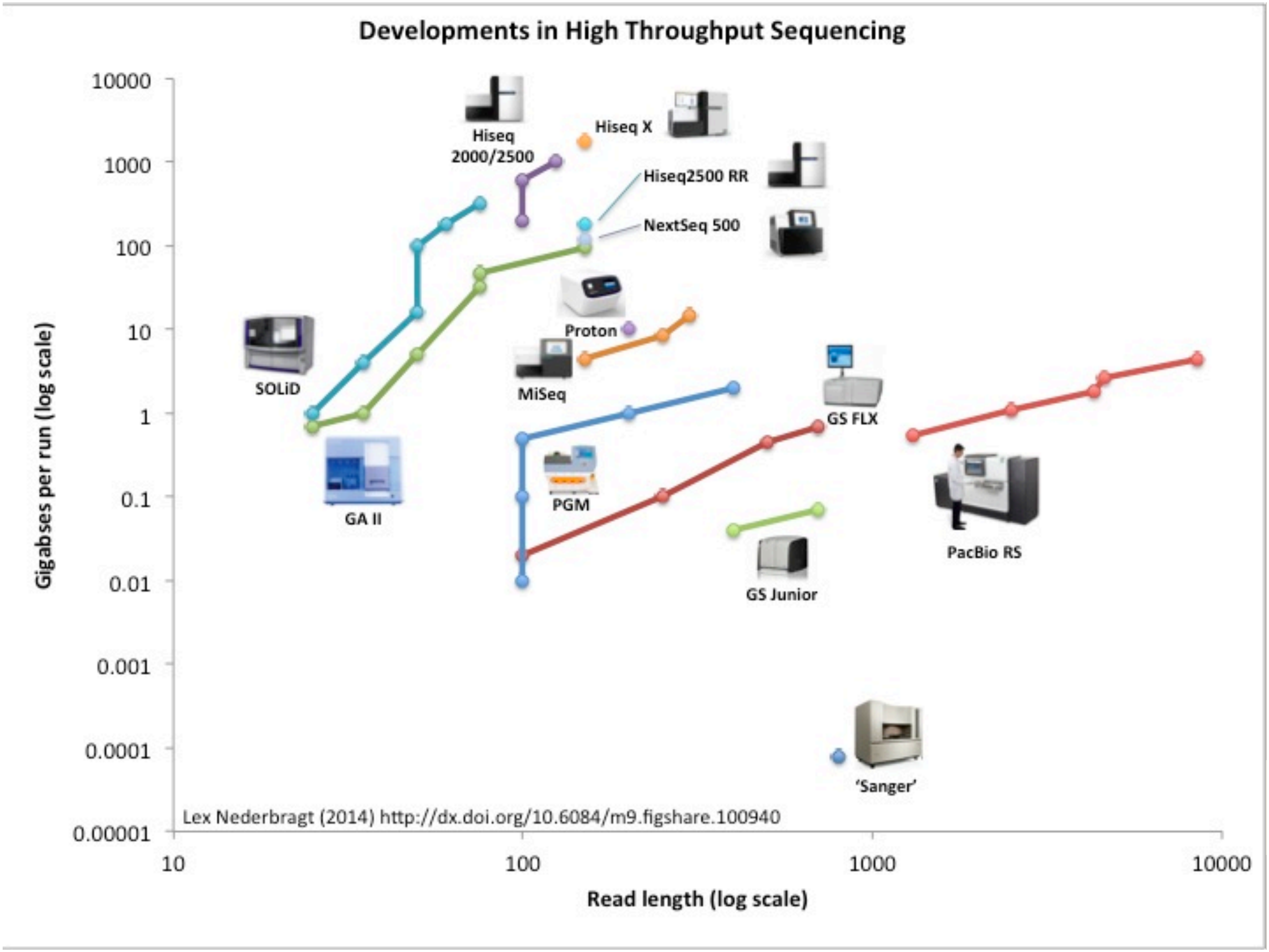
Actual yields

- MiSeq
  - 30 million reads per run
  - 300 base paired-end reads

- HiSeq 2500 RR/X 10
  - 6 billion reads per run
  - 150 base paired-end reads

- PacBio
  - 44,000 reads per run
  - 8500 bp in length

http://flxlexblog.wordpress.com/2014/06/11/developments-in-next-generation-sequencing-june-2014-edition/



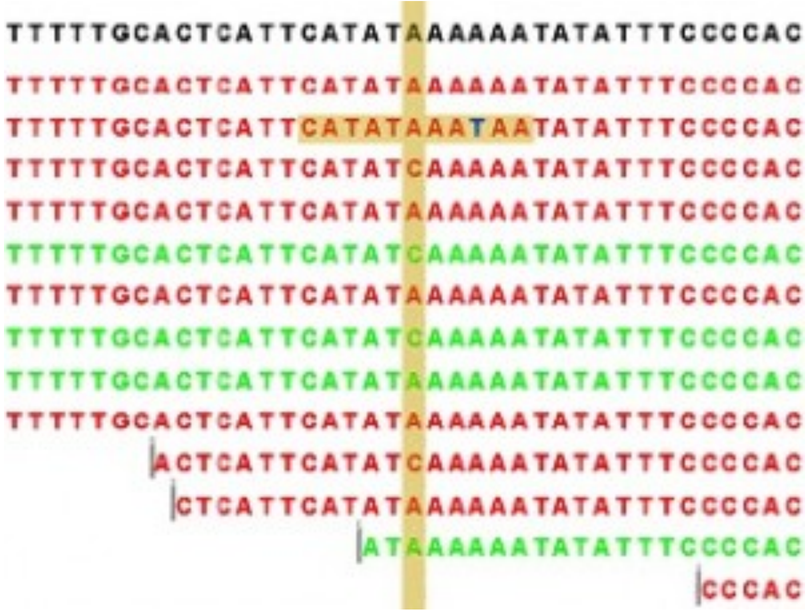
Developments in High Throughput Sequencing

Lex Nederbragt (2014) http://dx.doi.org/10.6084/m9.figshare.100940

Your basic data (FASTQ)

- @895:1:1:1246:14654/1
- CAGGCGCCCACCACCGTGCCCTCCAACCTGATGGT
- +
-][aaX__aa[`ZUZ[NONNFNNNNNO_____^RQ_
- @895:1:1:1246:14654/2
- ACTGGGCGTAGACGGTGTCCTCATCGGCACCAGC
- +
- \UJUWSSV[JQQWNP]]SZ]ZWU^]ZX][^TXR`
- @895:1:1:1252:19493/1
- CCGGCGTGGTTGGTGAGGTCACTGAGCTTCATGTC
- +
- OOOKONNNNN__`R]O[TGTRSY[IUZ]]]__X__

Mapping



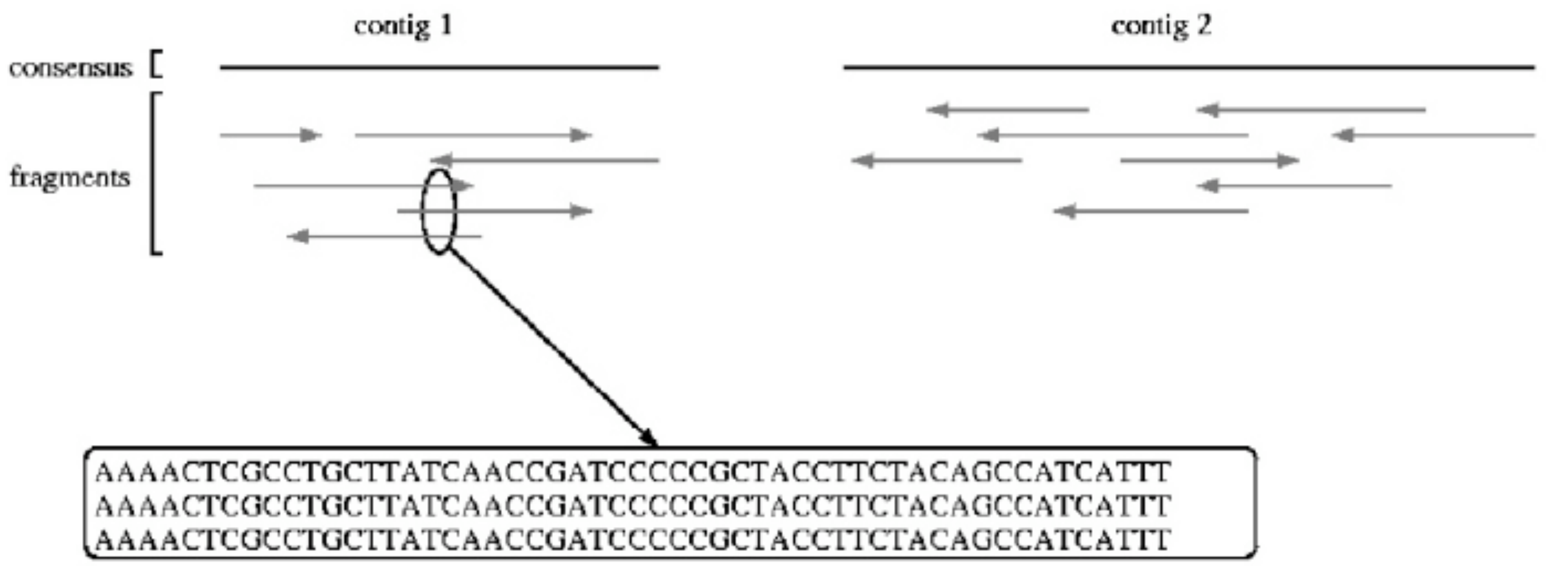
```
TTTTTGCACTCATTCATATAAAAAATATATTTCCCCAC
TTTTTGCACTCATTCATATAAAAAATATATTTCCCCAC
TTTTTGCACTCATTCATATAAATAATATATTTCCCCAC
TTTTTGCACTCATTCATATCAAAAATATATTTCCCCAC
TTTTTGCACTCATTCATATAAAAAATATATTTCCCCAC
TTTTTGCACTCATTCATATCAAAAATATATTTCCCCAC
TTTTTGCACTCATTCATATAAAAAATATATTTCCCCAC
TTTTTGCACTCATTCATATCAAAAATATATTTCCCCAC
TTTTTGCACTCATTCATATAAAAAATATATTTCCCCAC
TTTTTGCACTCATTCATATAAAAAATATATTTCCCCAC
       ACTCATTCATATCAAAAATATATTTCCCCAC
        CTCATTCATATAAAAAATATATTTCCCCAC
                 ATAAAAAATATATTTCCCCAC
                                 CCCAC
```

U. Colorado
http://genomics-course.jasondk.org/?p=395

- Many fast & efficient computational solutions exist.
- You have to figure out how to choose parameters to maximize sensitivity/specificity, and when to validate.

Assembly

Reassemble random fragments computationally.



```
AAAACTCGCCTGCTTATCAACCGATCCCCCGCTACCTTCTACAGCCATCATTT
AAAACTCGCCTGCTTATCAACCGATCCCCCGCTACCTTCTACAGCCATCATTT
AAAACTCGCCTGCTTATCAACCGATCCCCCGCTACCTTCTACAGCCATCATTT
```

UMD assembly primer (cbcb.umd.edu)

Shotgun sequencing

It was the best of times, it was the wor

, it was the worst of times, it was the

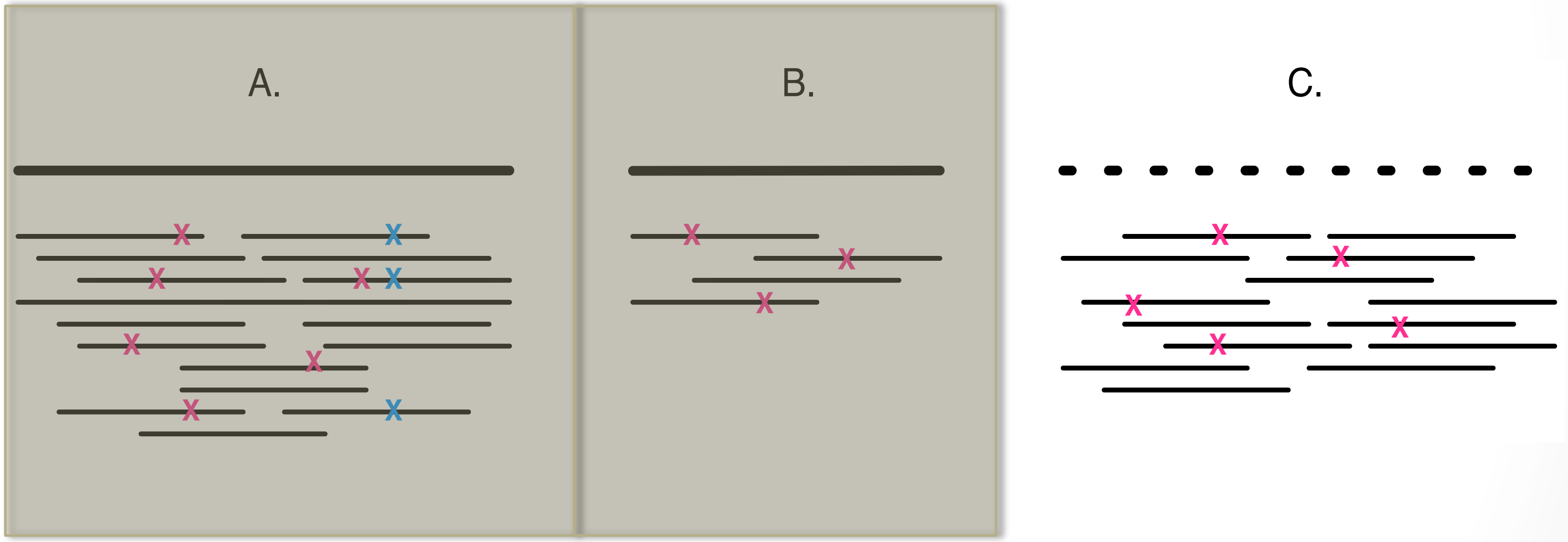
isdom, it was the age of foolishness

mes, it was the age of wisdom, it was th

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness

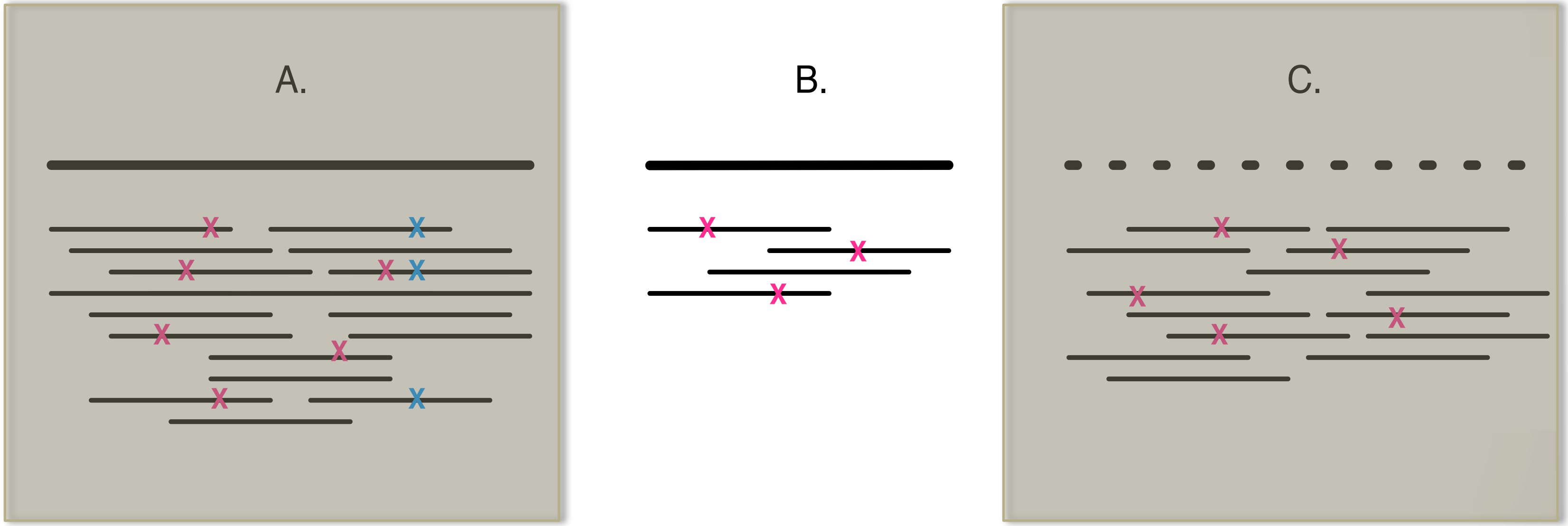
Where does # of reads count?

Resequencing, counting, and assembly.



Where does reconstructability matter?

Resequencing, counting, and assembly.



Summary

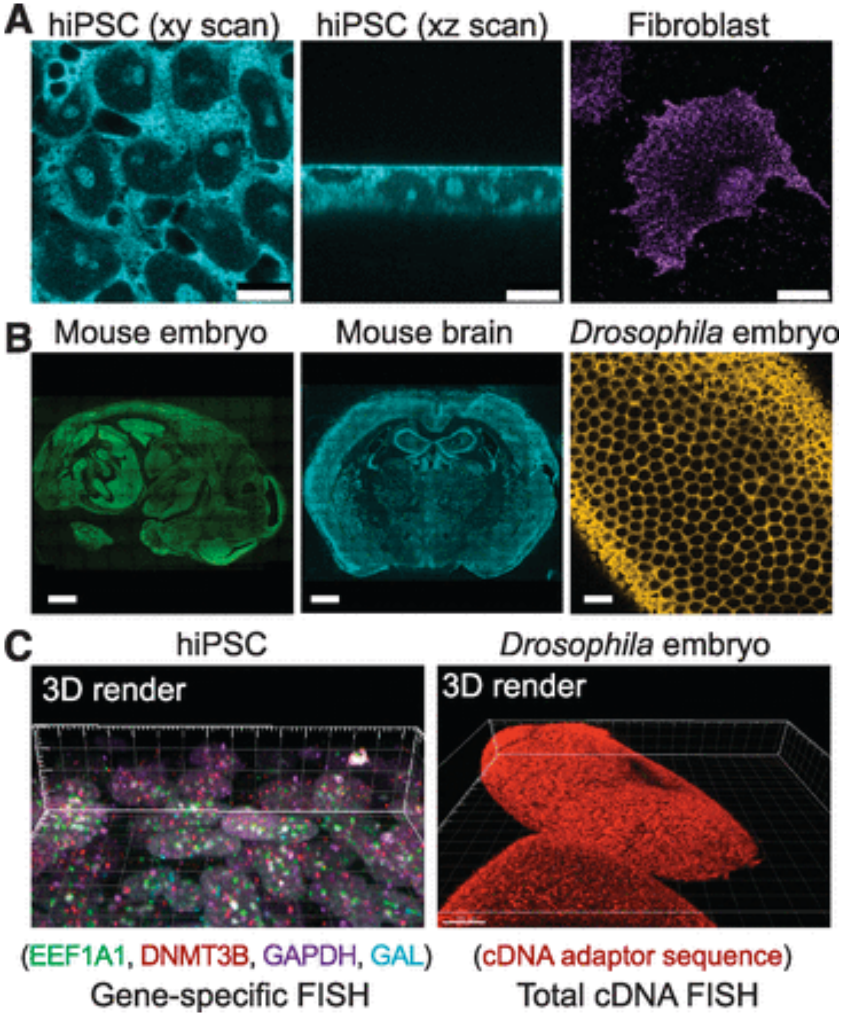
- Coverage matters for SNP calls and assembly;

- # of reads matters for counting;

- Length of reads matters for reconstructability (assembly & haplotyping);

- Illumina is still "best" for high coverage;
- PacBio and Moleculo => genome assembly;
- Nanopore??

Sequencing in situ!?



Lee et al., http://www.sciencemag.org/content/343/6177/1360.full

Today

- More command line stuff
- Working with actual data!!
- Evaluating the quality of your data…