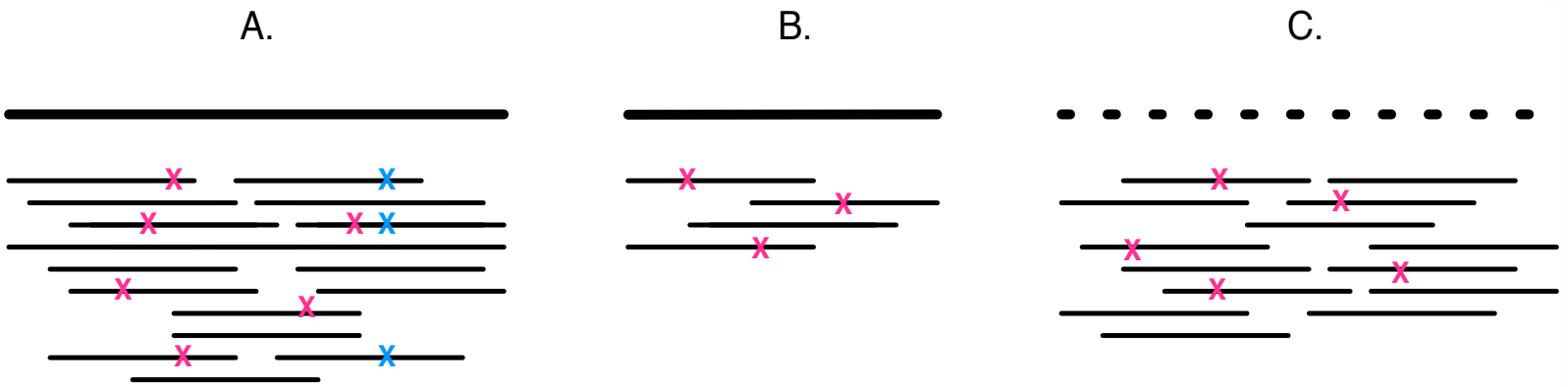
(Shotgun) sequencing

Titus Brown

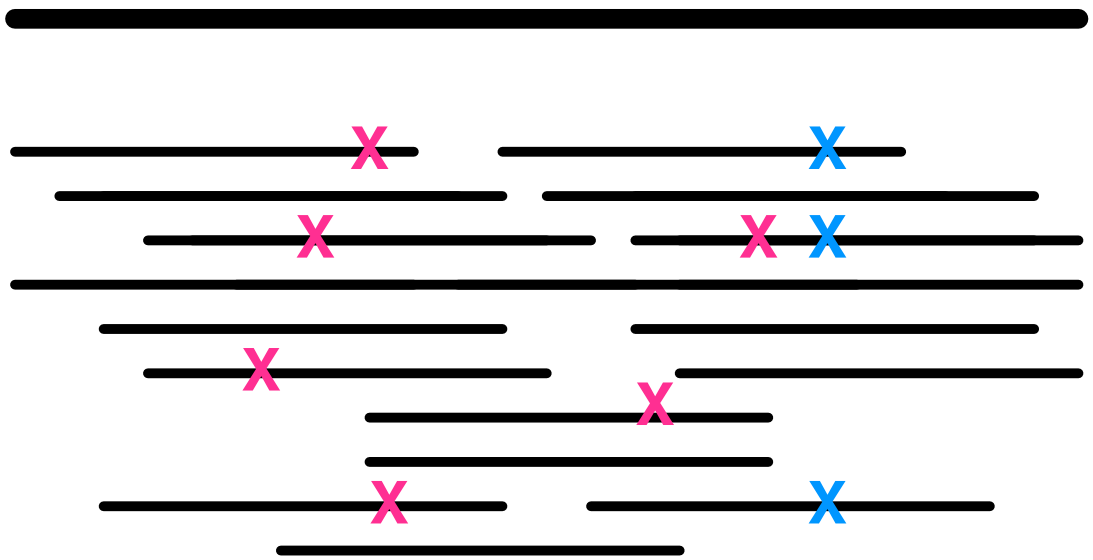
Three basic problems

Resequencing, counting, and assembly.



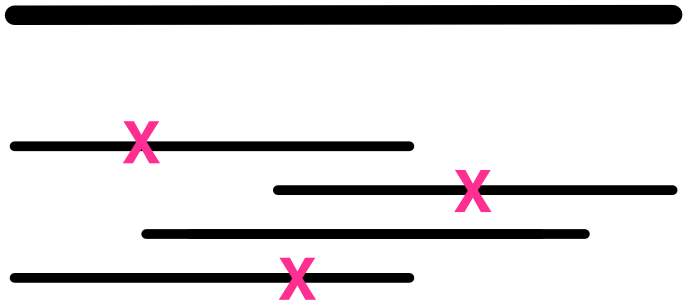
1. Resequencing analysis

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



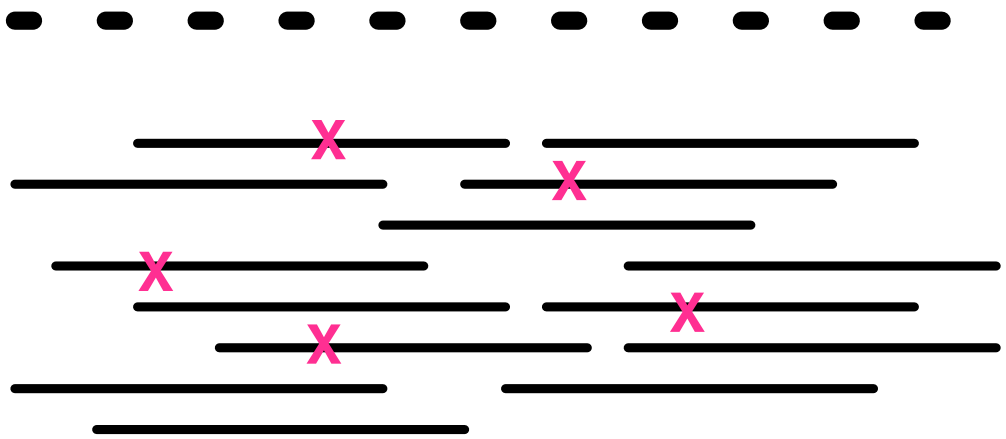
2. Counting

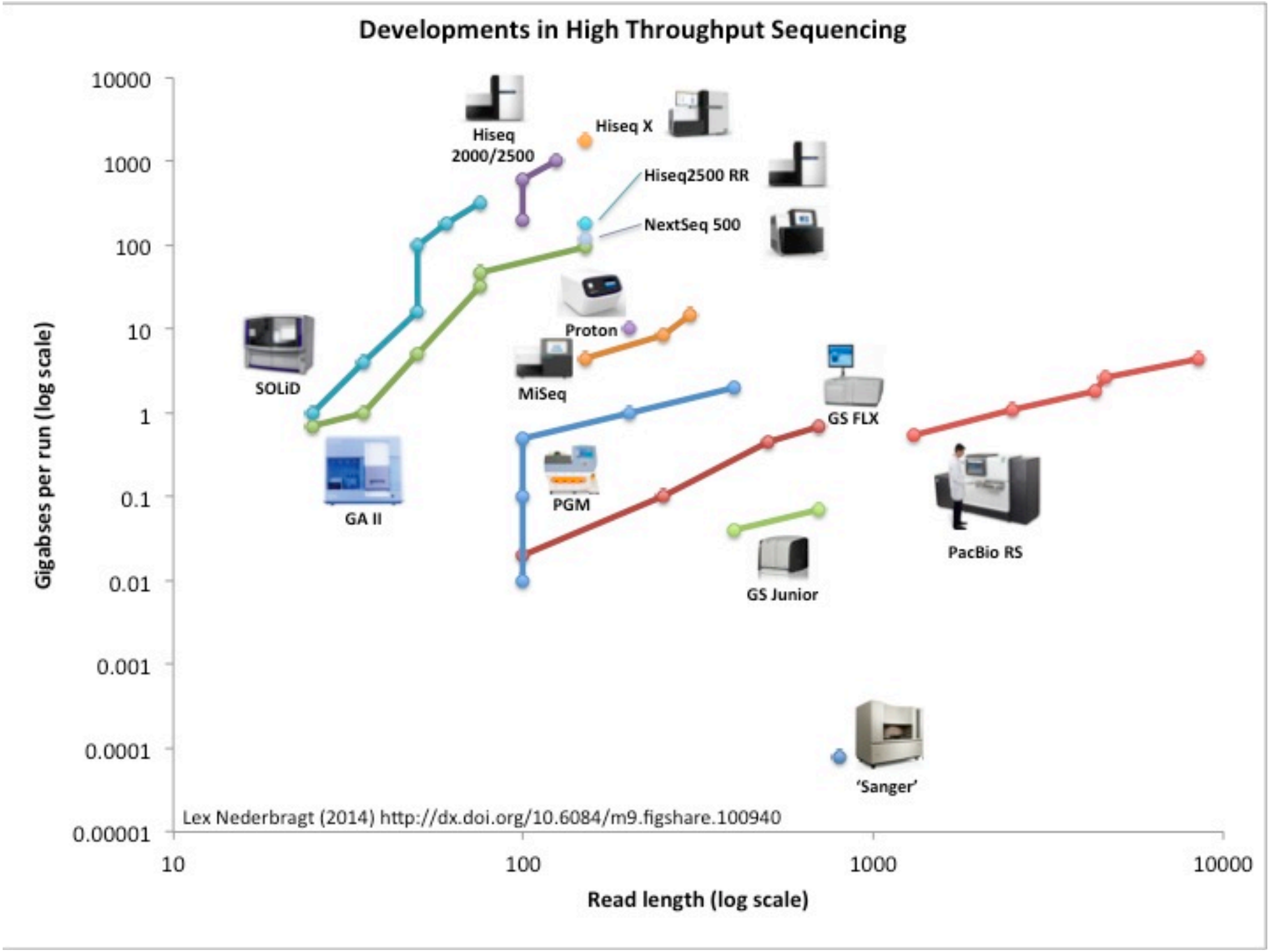
We have a reference genome (or gene set) and want to know how *much* we have. Think gene expression/ microarrays.



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
Gigabases per run (log scale)
10000
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1000
10000
Read length (log scale)
Hiseq 2000/2500
Hiseq X
Hiseq2500 RR
NextSeq 500
Proton
SOLiD
MiSeq
GS FLX
GA II
PGM
PacBio RS
GS Junior
'Sanger'
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

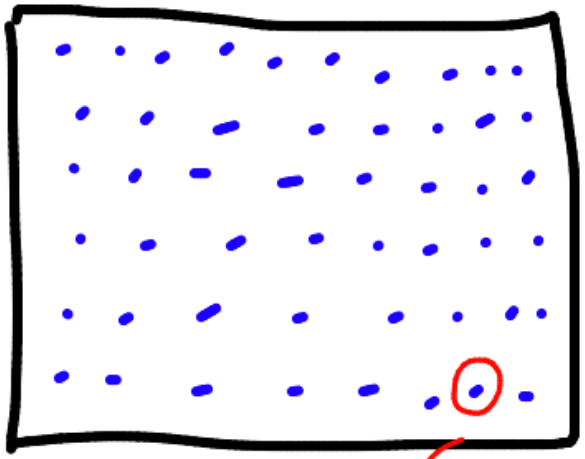
It was the best of times, it was the wor

, 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

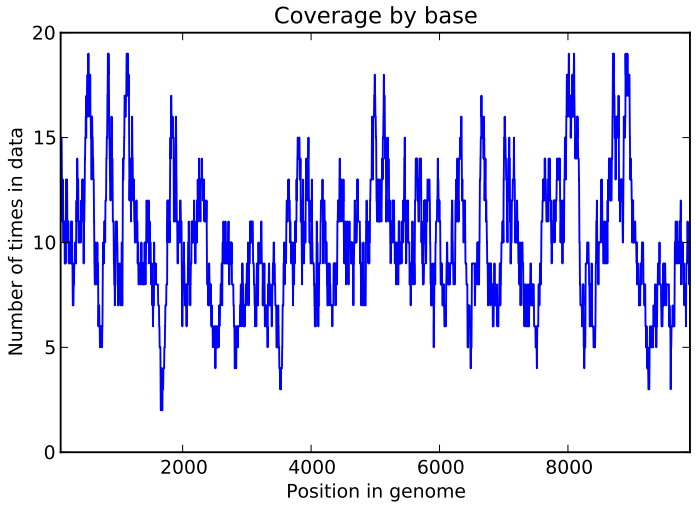
$500 or so.

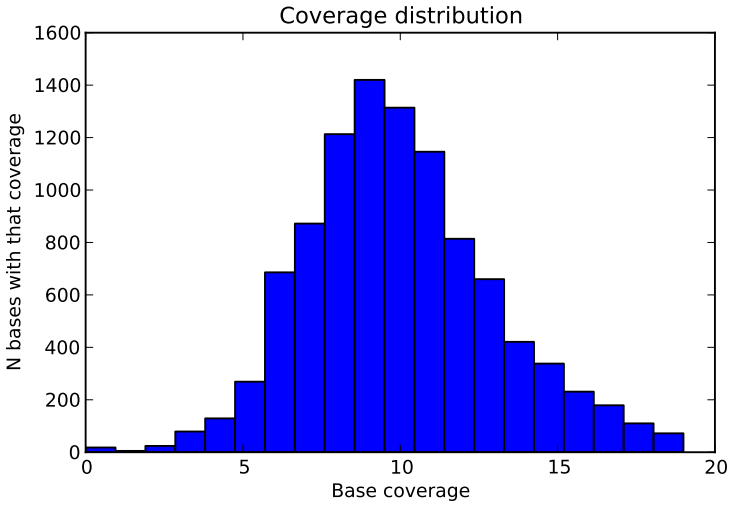
Two specific concepts:

- 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.

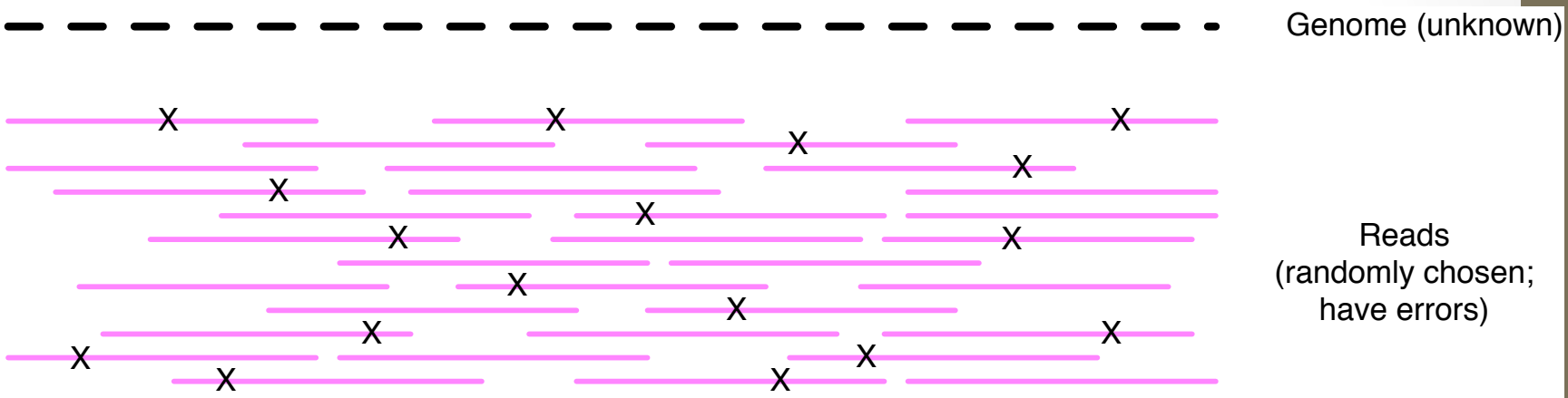
Random sampling => deep sampling needed





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.

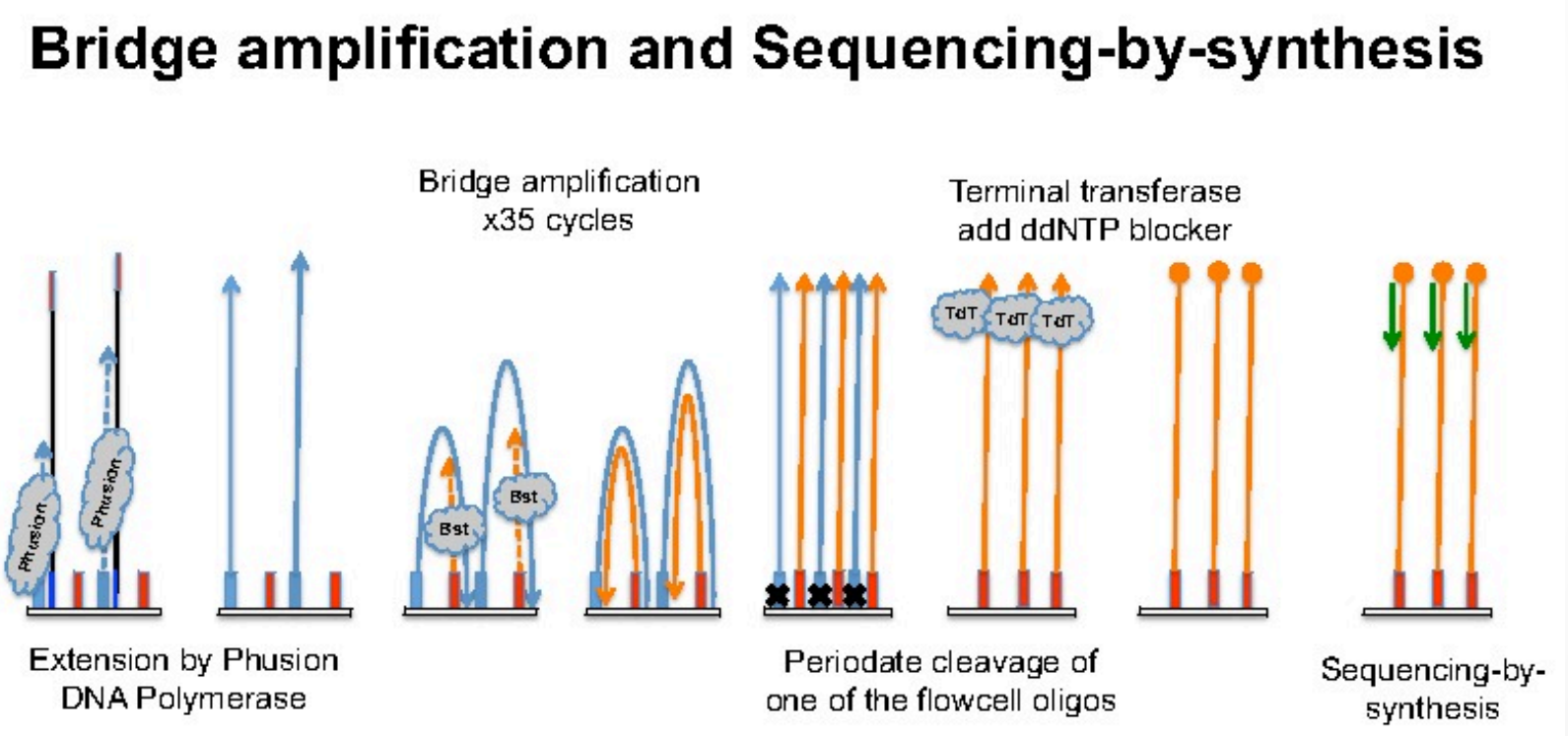
Illumina yields the *deepest* sequencing available

- 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)



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

What goes wrong with basic assumptions?

- Not all sequence is as easily sequenced as other, depending on your sequencing technology (e.g. GC/AT bias);
- Some RNA not be as accessible as others (secondary structure);

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

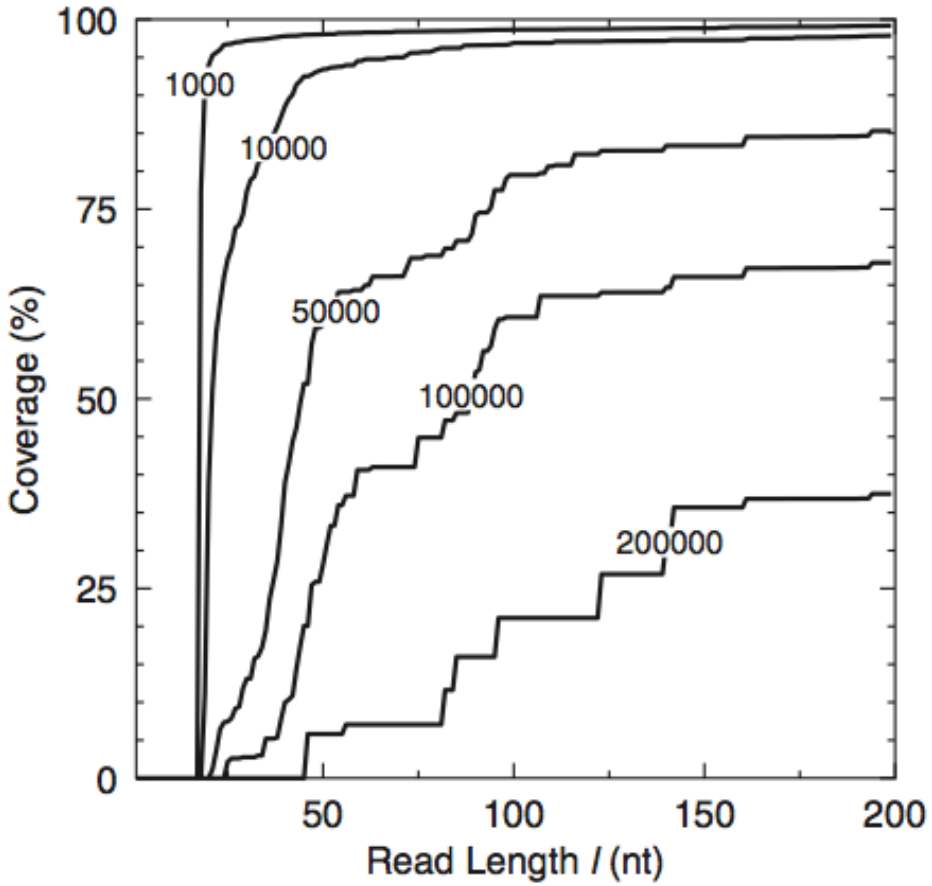


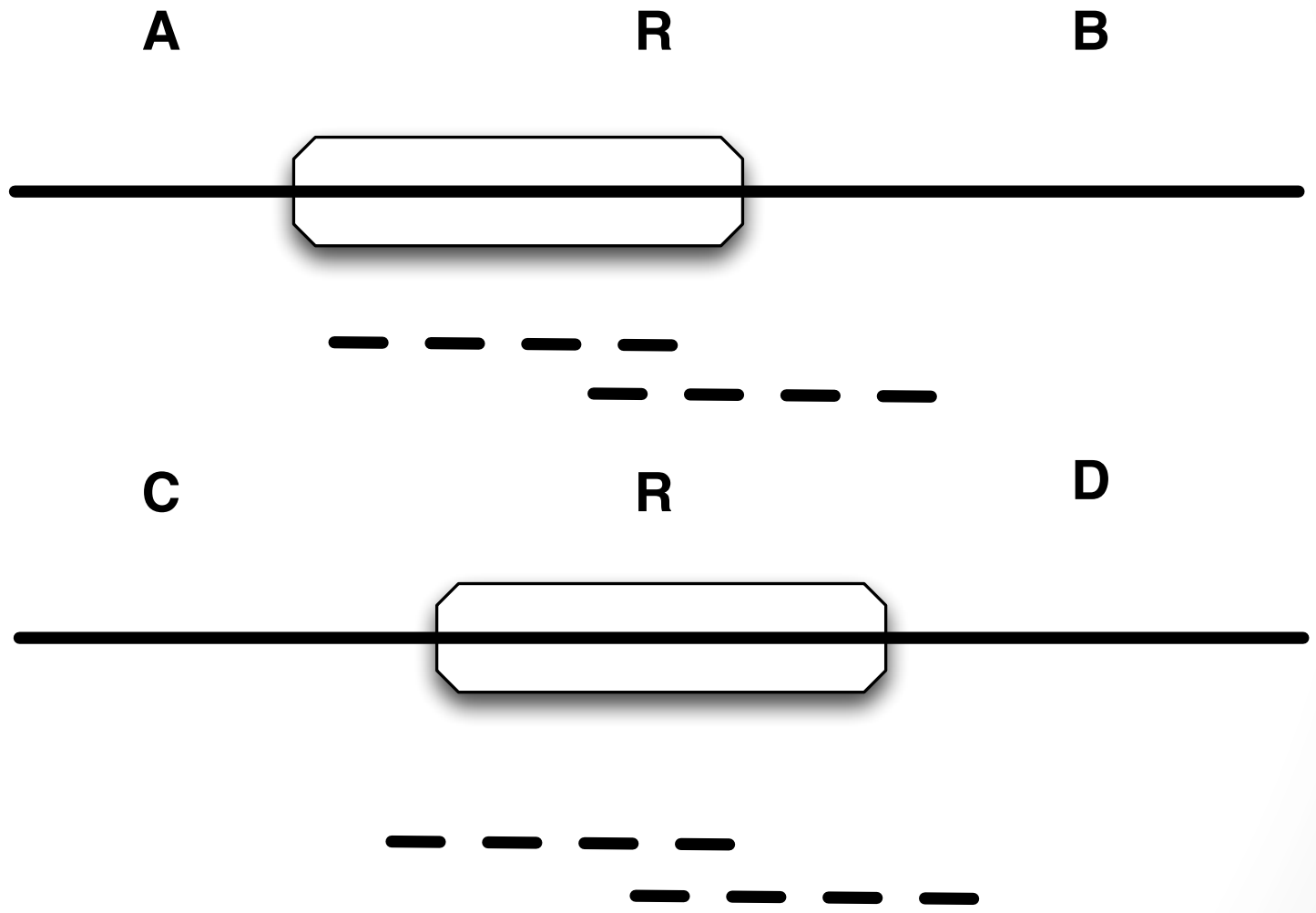
Figure 3. Percentage of the *E.coli* genome covered by contigs greater than a threshold length as a function of read length.

Whiteford et al., Nuc. Acid Res, 2005

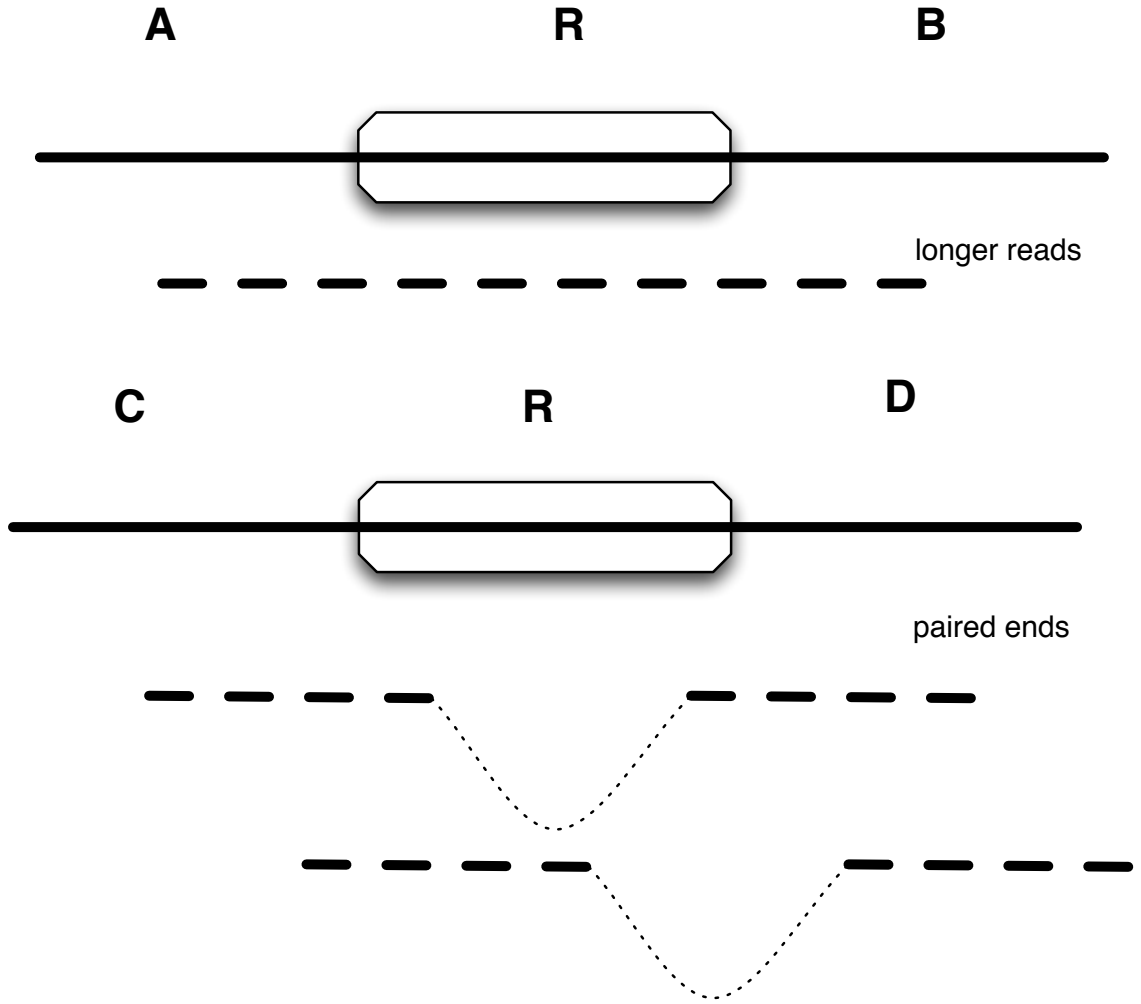
“Reconstructability”

- Assembling new genomes or transcriptomes…
- *Haplotyping* - think human genetics & viruses, both.

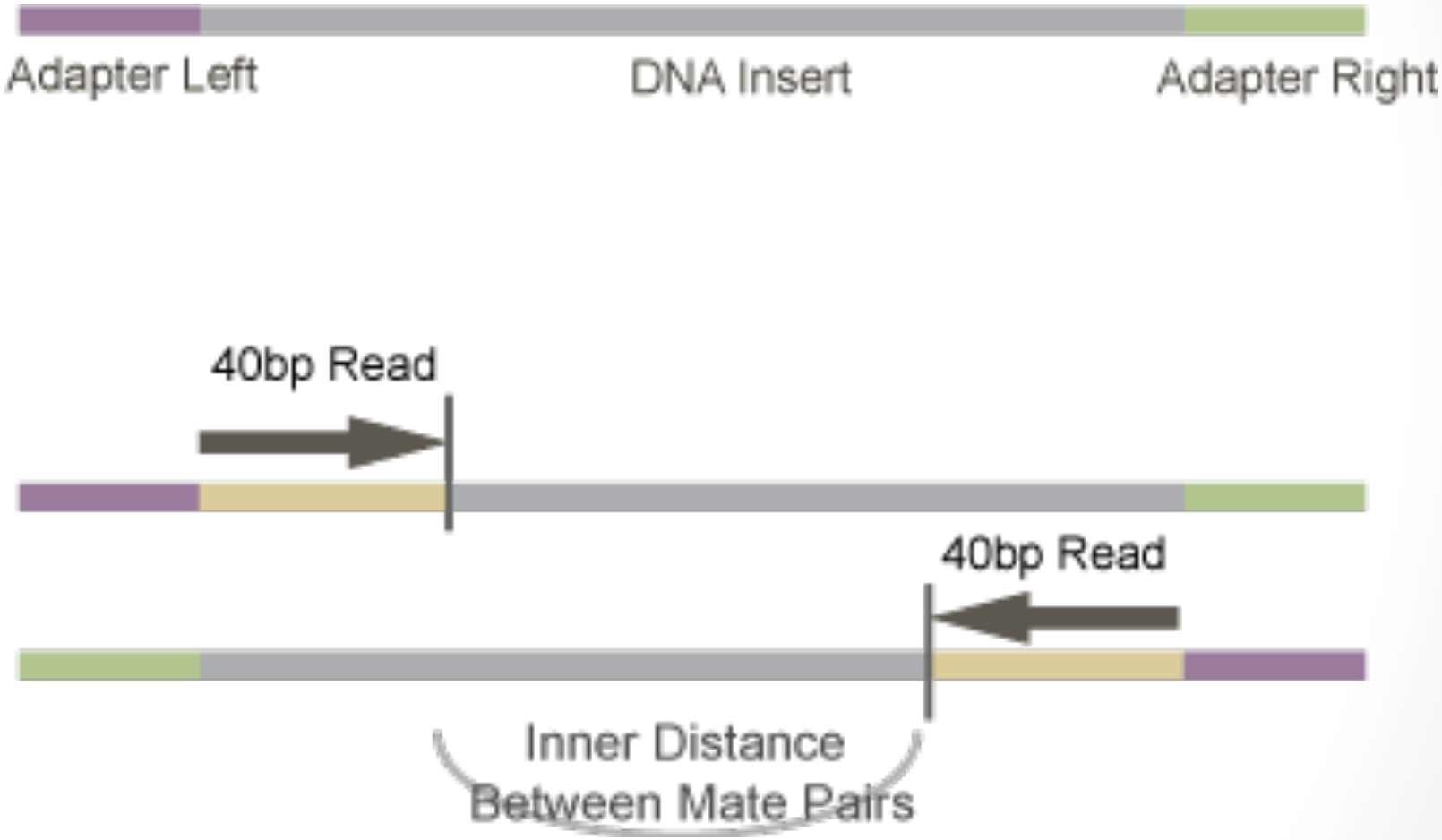
Repeats! (and shared exons)



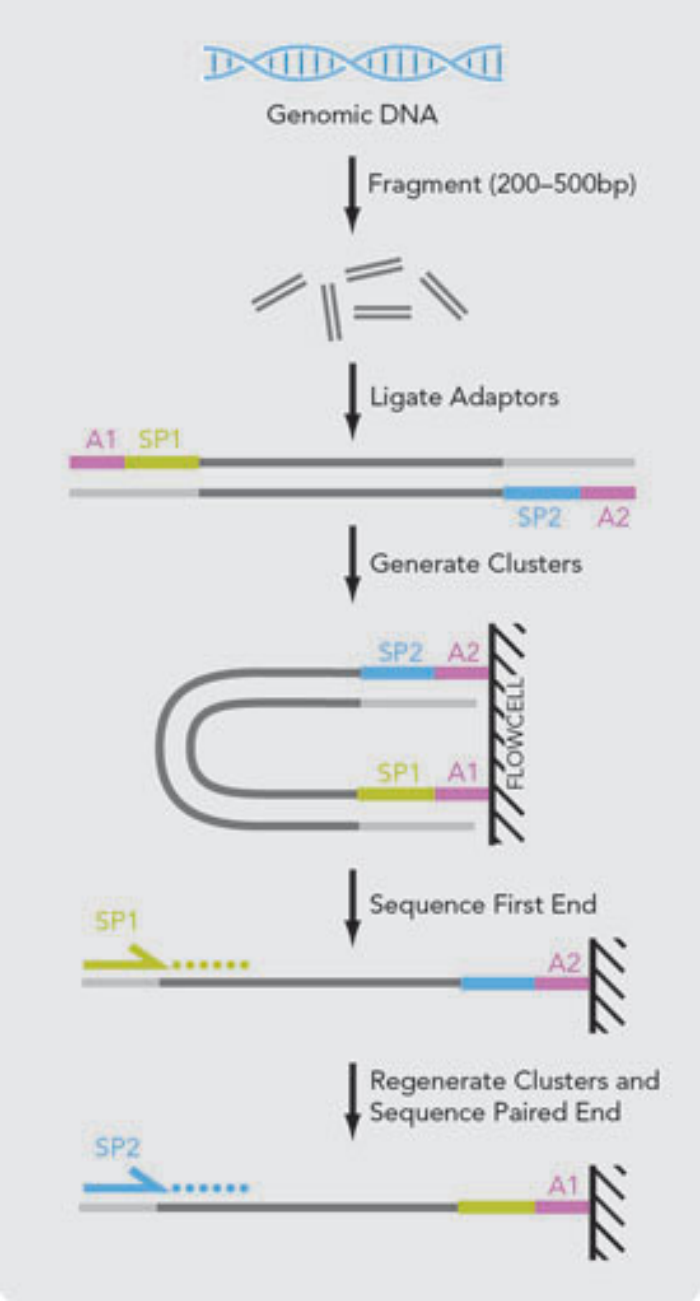
Longer reads … OR …
Paired-end/mate pair sequencing

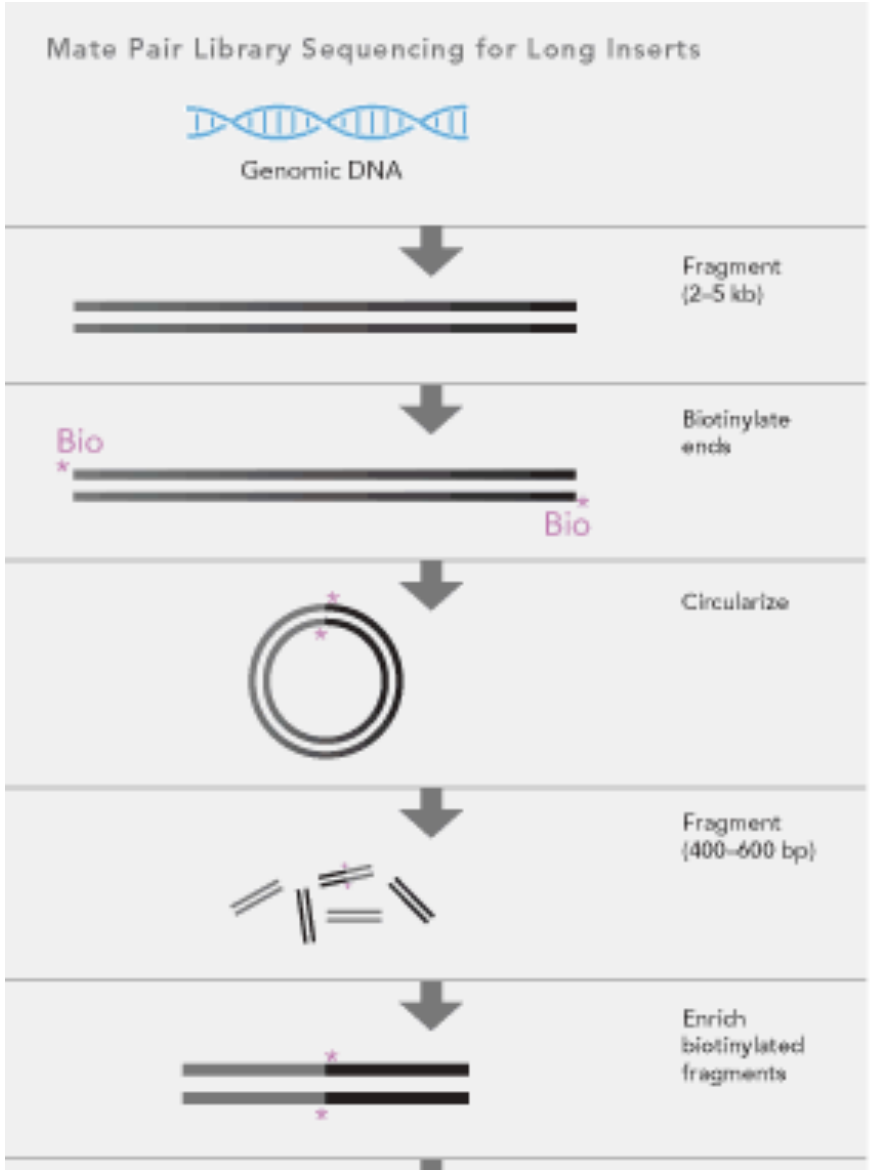


Paired-end sequencing



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.

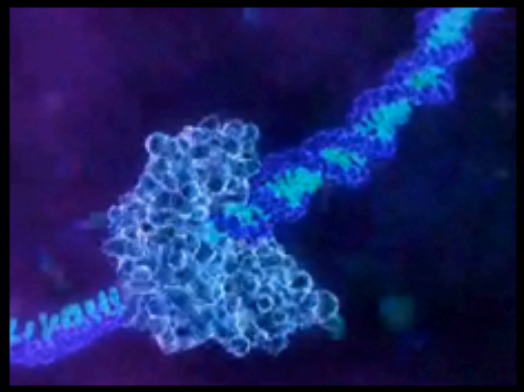
Mate-pair sequencing (long insert)

Longer reads

- PacBio
- Moleculo
- Nanopore

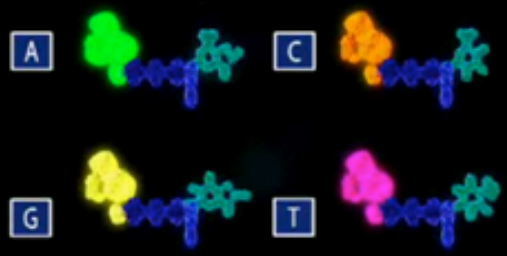
Next-gen sequencing: Pacific Biosciences

1



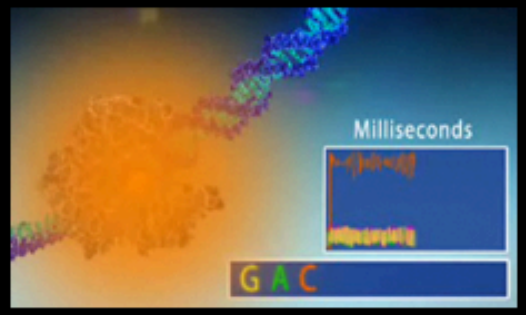
DNA polymerase wrapped around DNA chain

2

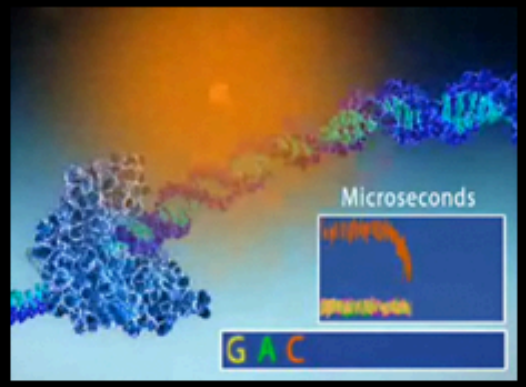


Phospholinked nucleotides

3a



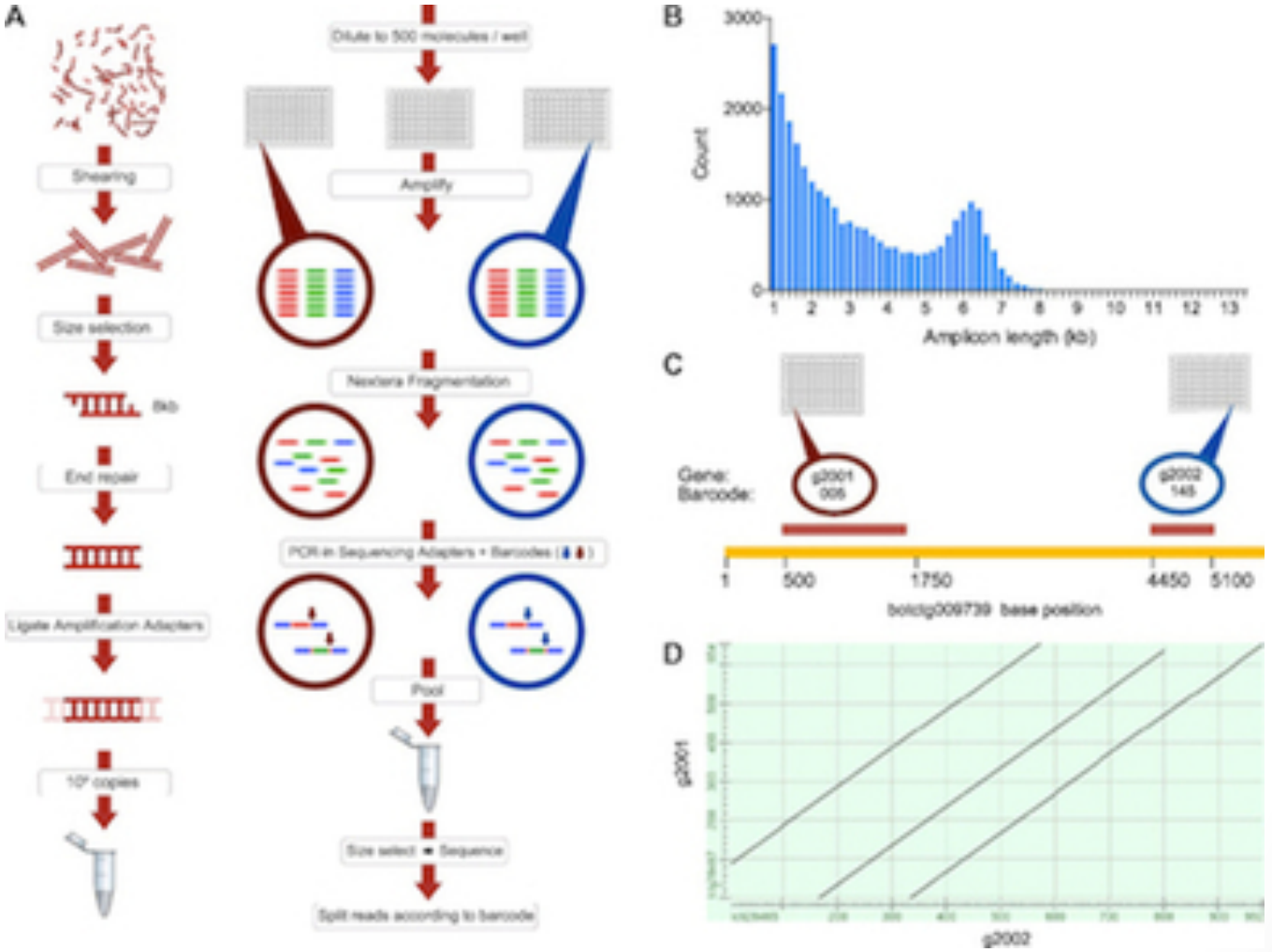
3b



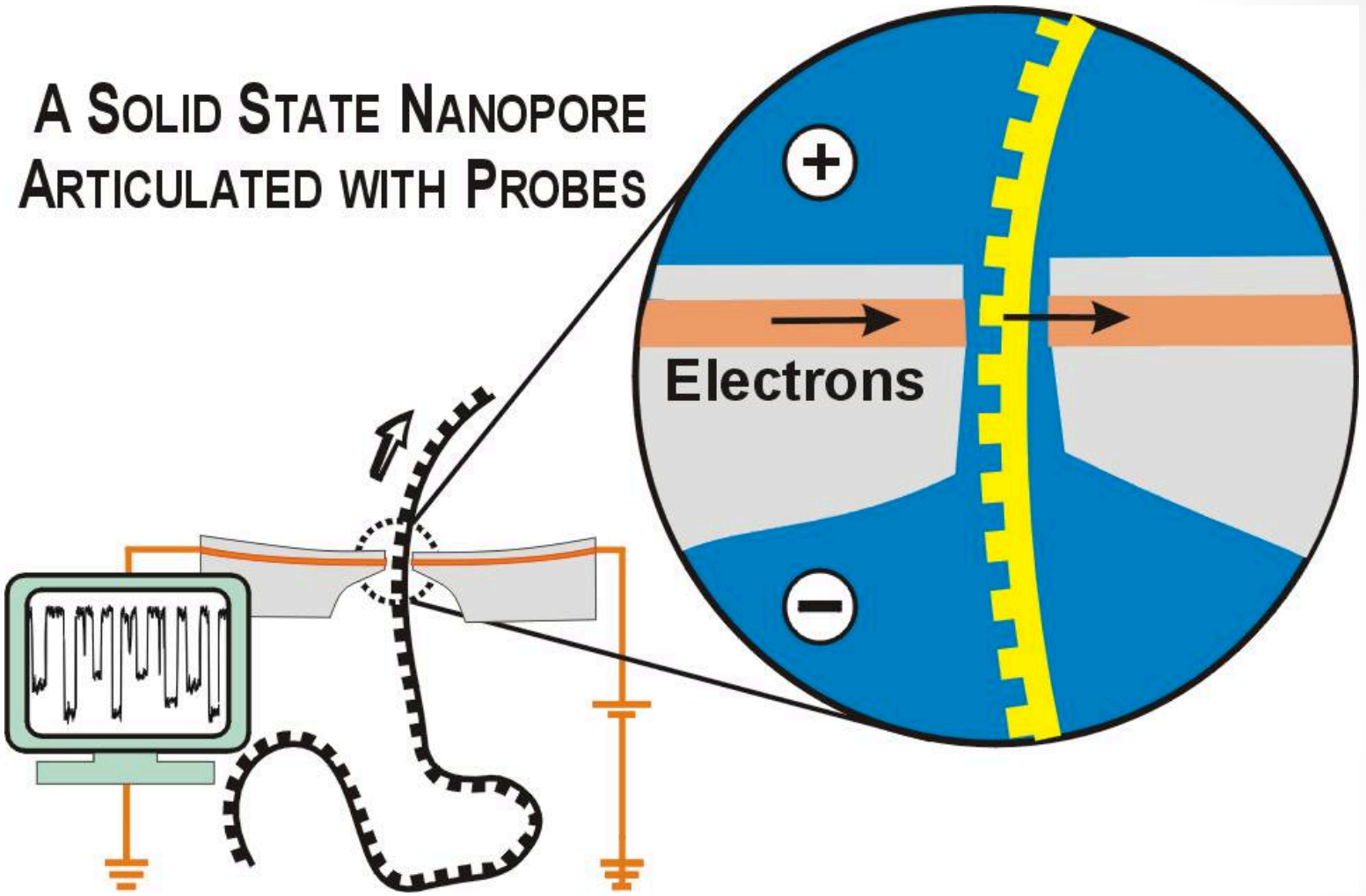
Phospholinked nucleotide binds, fluoresces and detaches as nucleotide base is read

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

Moleculo (Illumina)



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

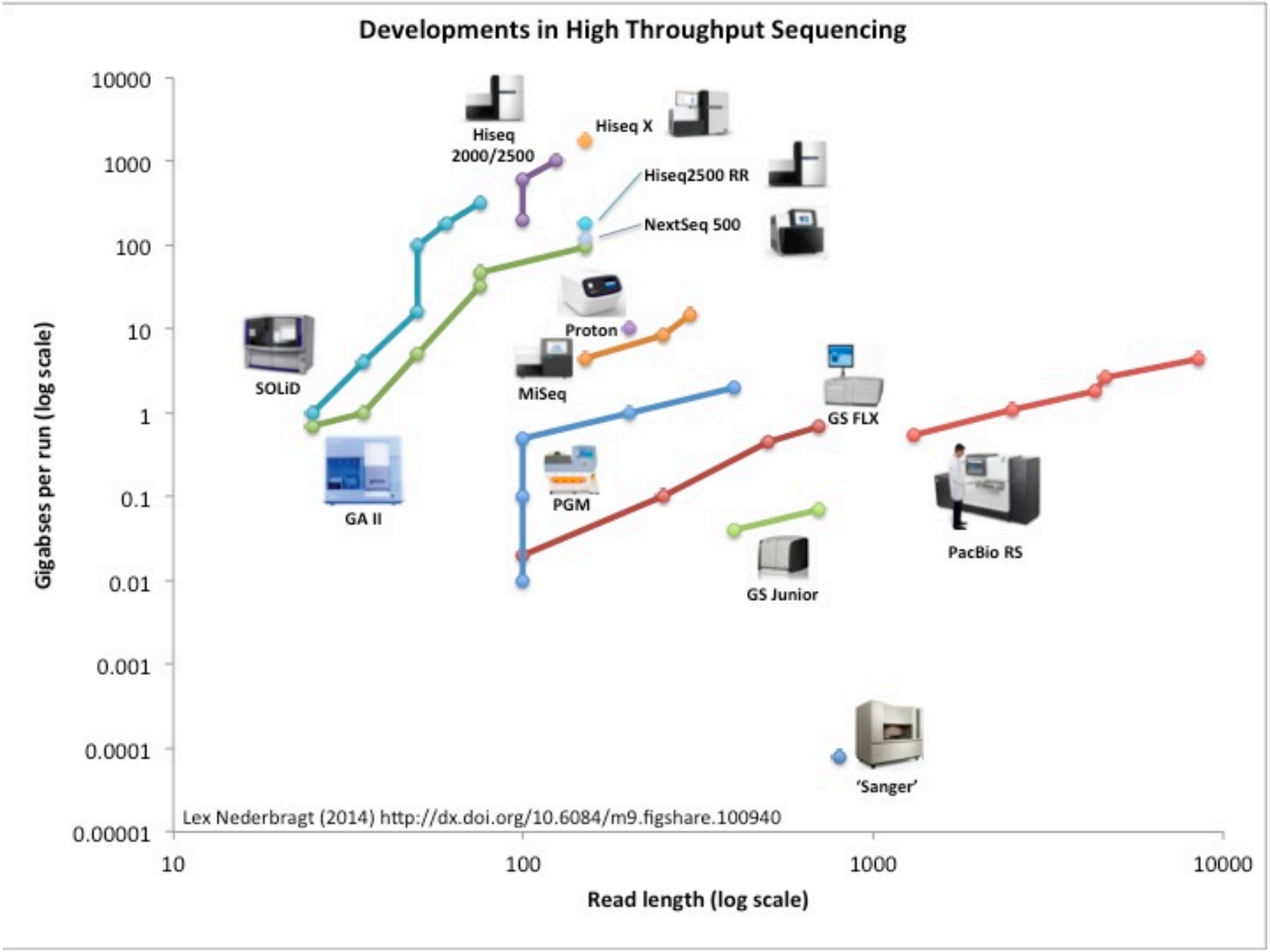


http://labs.mcb.harvard.edu/branton/projects-NanoporeSequencing.htm

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
Gigabases per run (log scale)
10000
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1000
10000
Read length (log scale)
Hiseq 2000/2500
Hiseq X
Hiseq2500 RR
NextSeq 500
Proton
SOLiD
MiSeq
GS FLX
GA II
PGM
PacBio RS
GS Junior
'Sanger'
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__

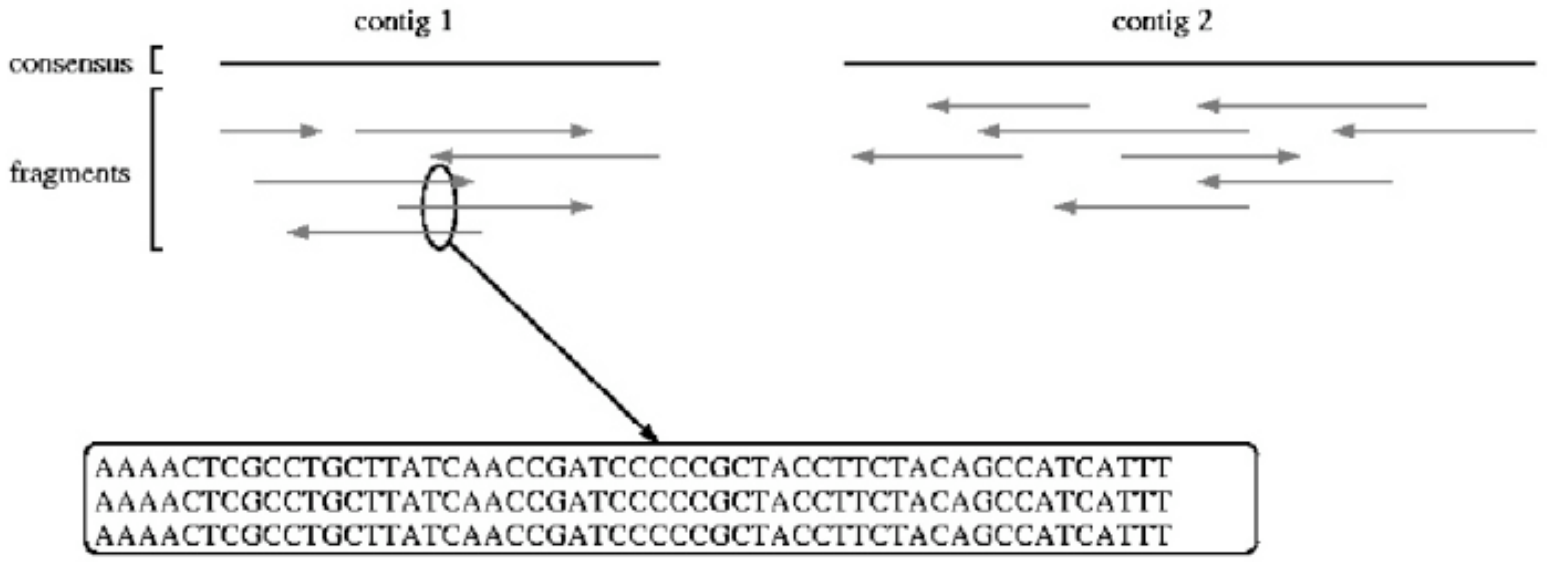
Mapping

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.



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

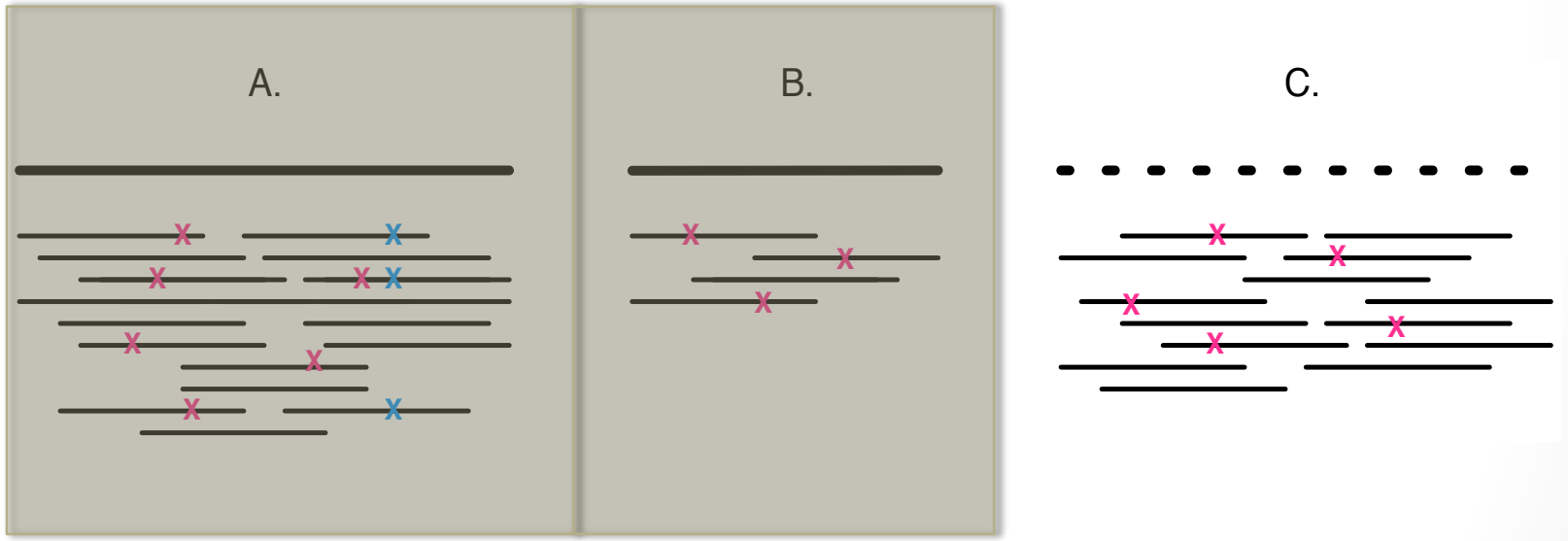
isdom, it was the age of foolishness

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

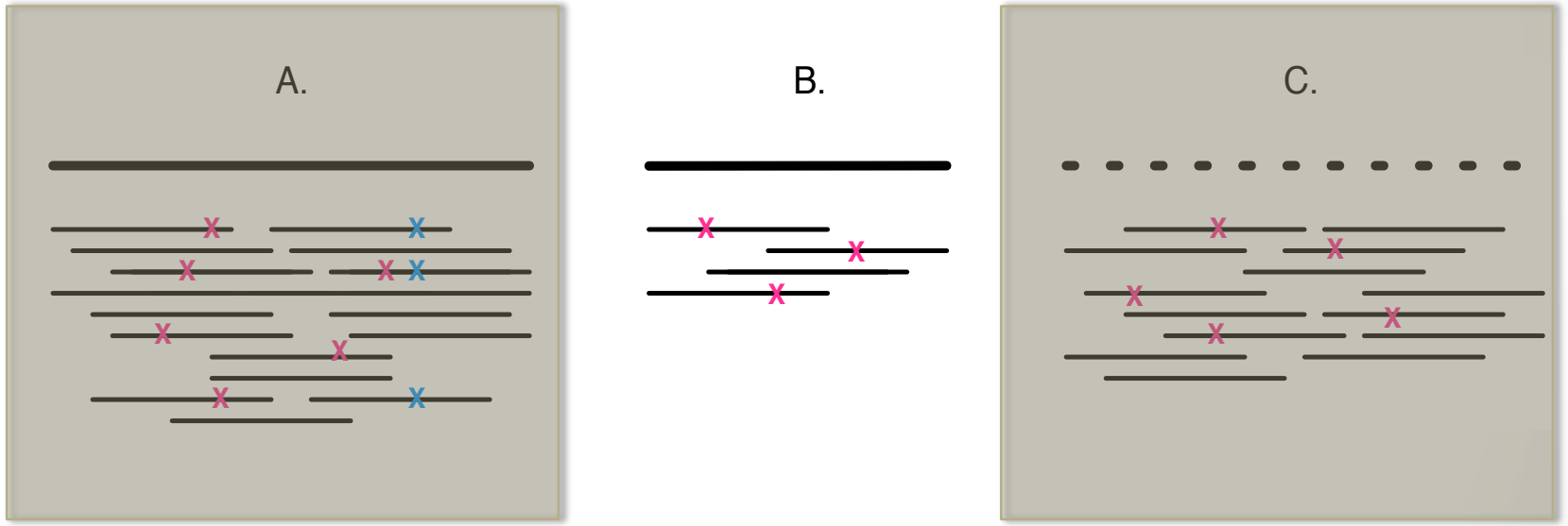
Where does # of reads count?

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: still tricky but lots of progress being made.

Bad data

I asked:
https://twitter.com/ctitusbrown/status/624721875252420608

I received:

- http://www.bioinfo-core.org/index.php/Interesting_NGS_failures
- http://bioinfo-core.org/index.php/9th_Discussion-28_October_2010
- https://biomickwatson.wordpress.com/2013/01/21/ten-things-to-consider-when-choosing-an-ngs-supplier/

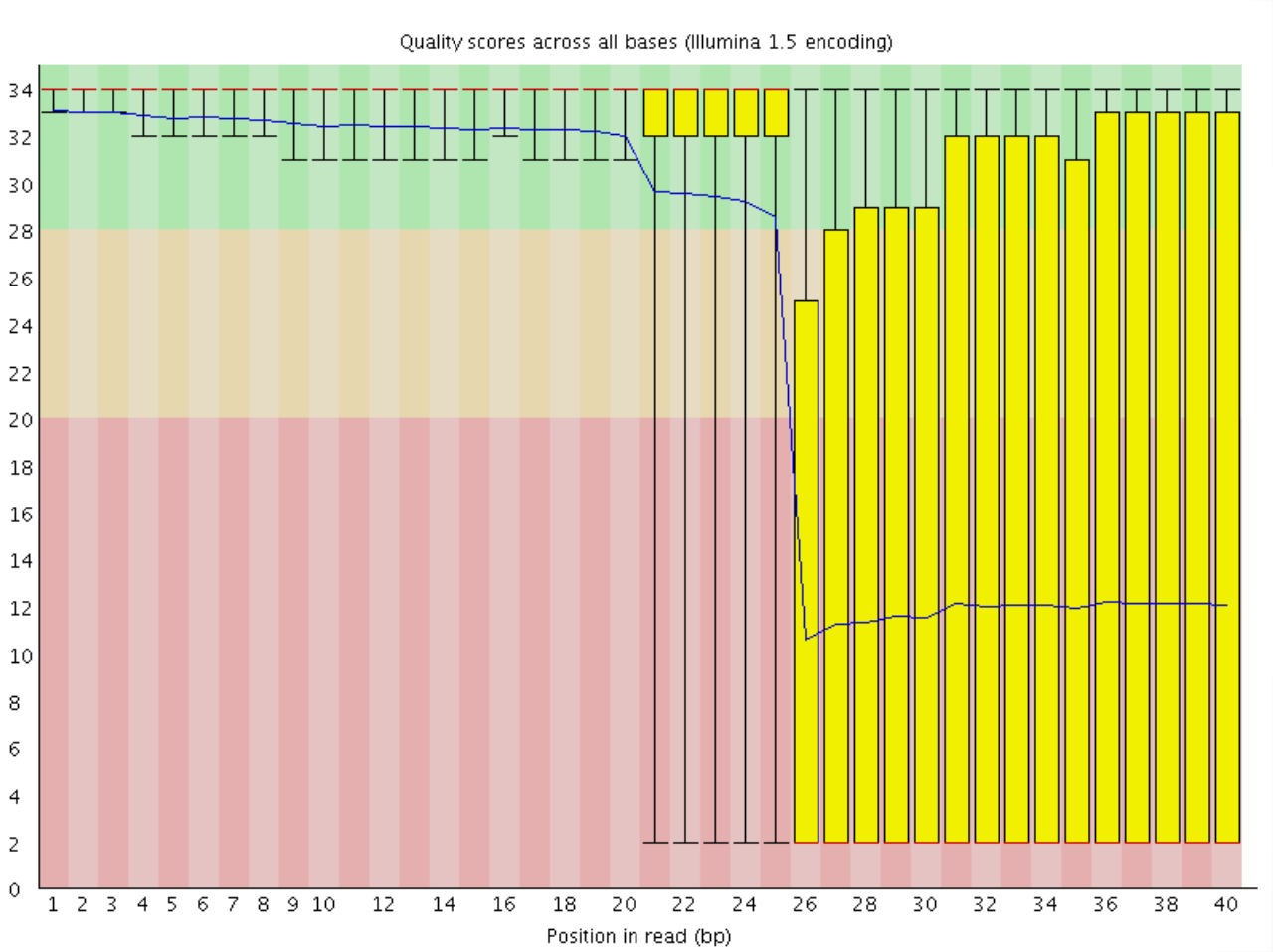
Sequencing Bloopers

Simon Andrews

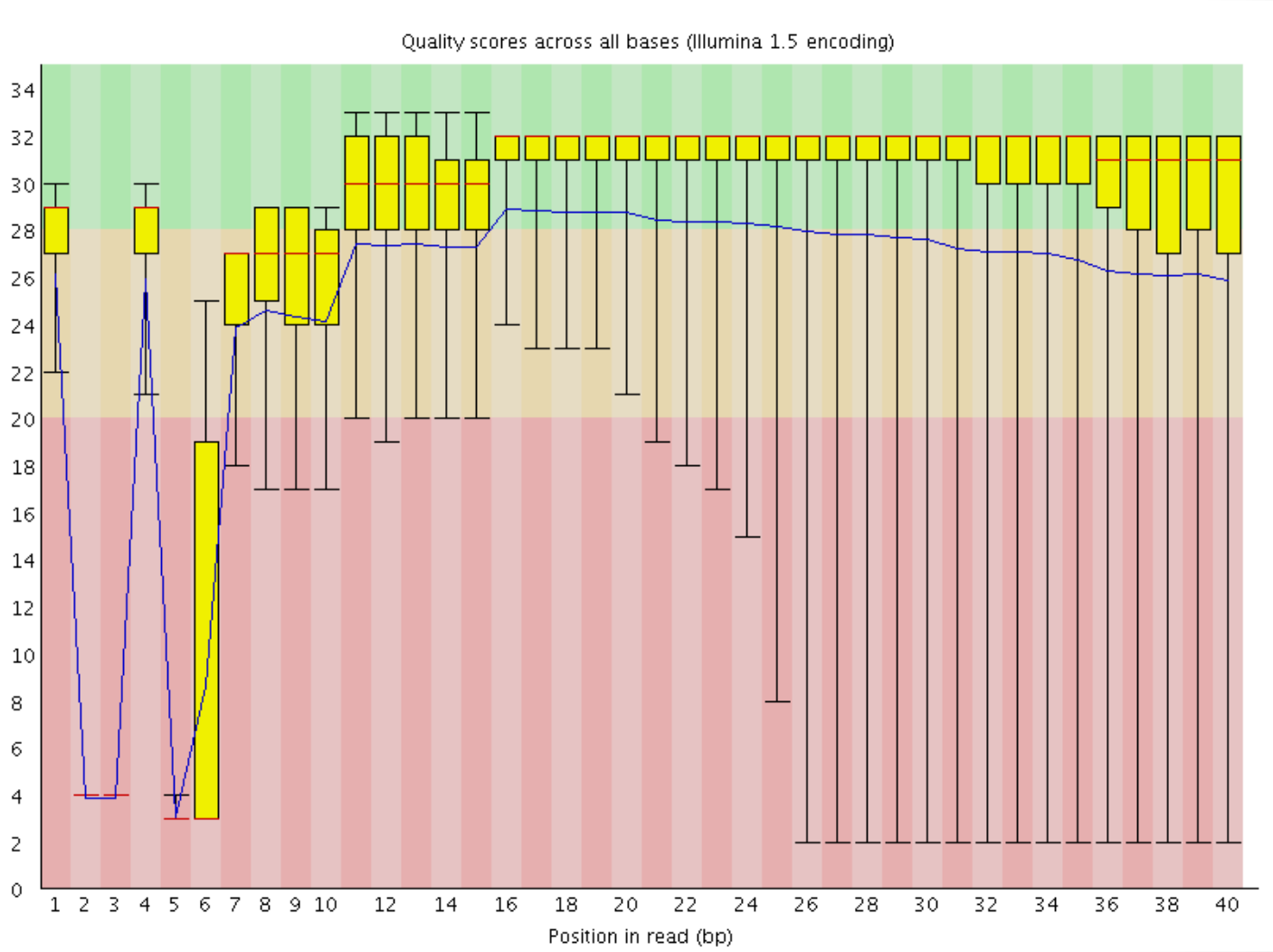
Tim Stevens

Technical sequencer problems

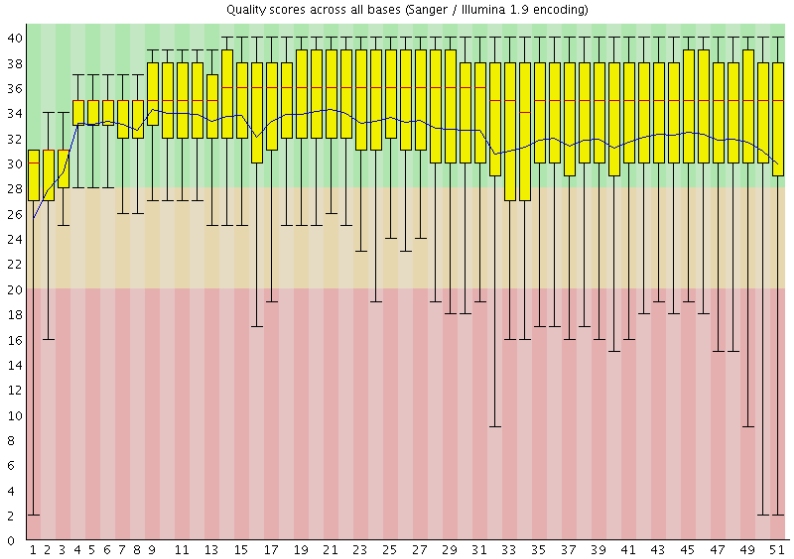
Manifold burst in cycle 26

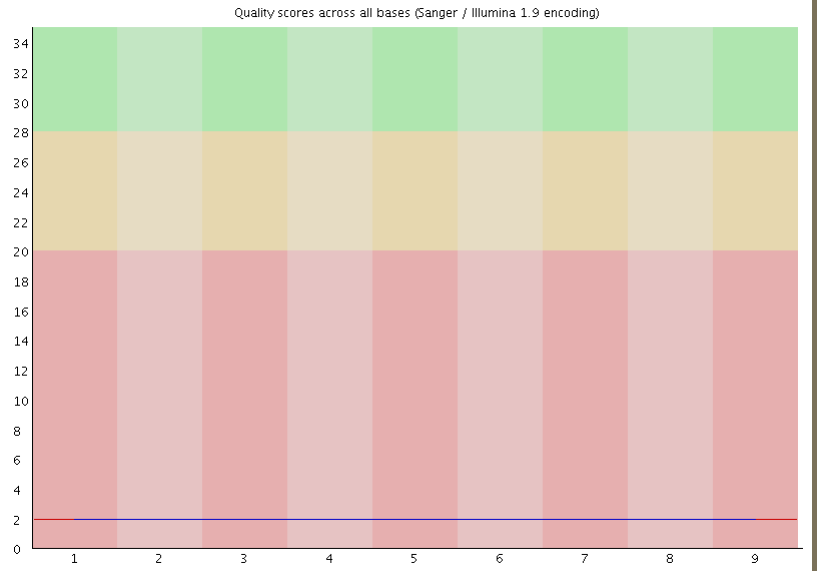


Specific cycles lost

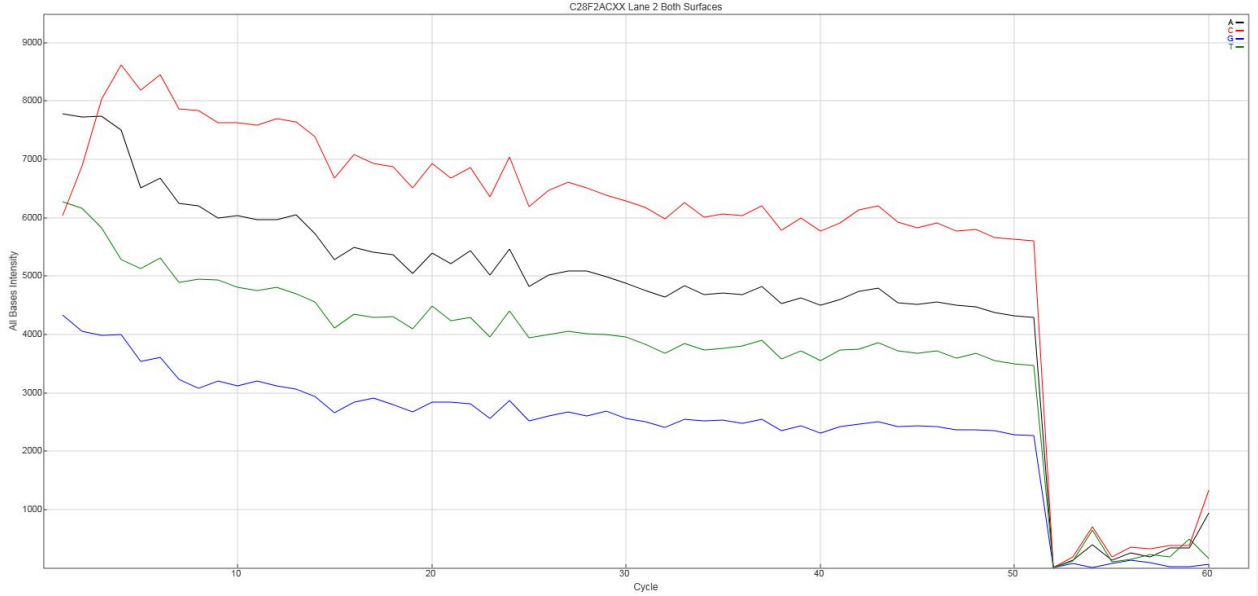


No priming /signal
(Wrong adapters used)



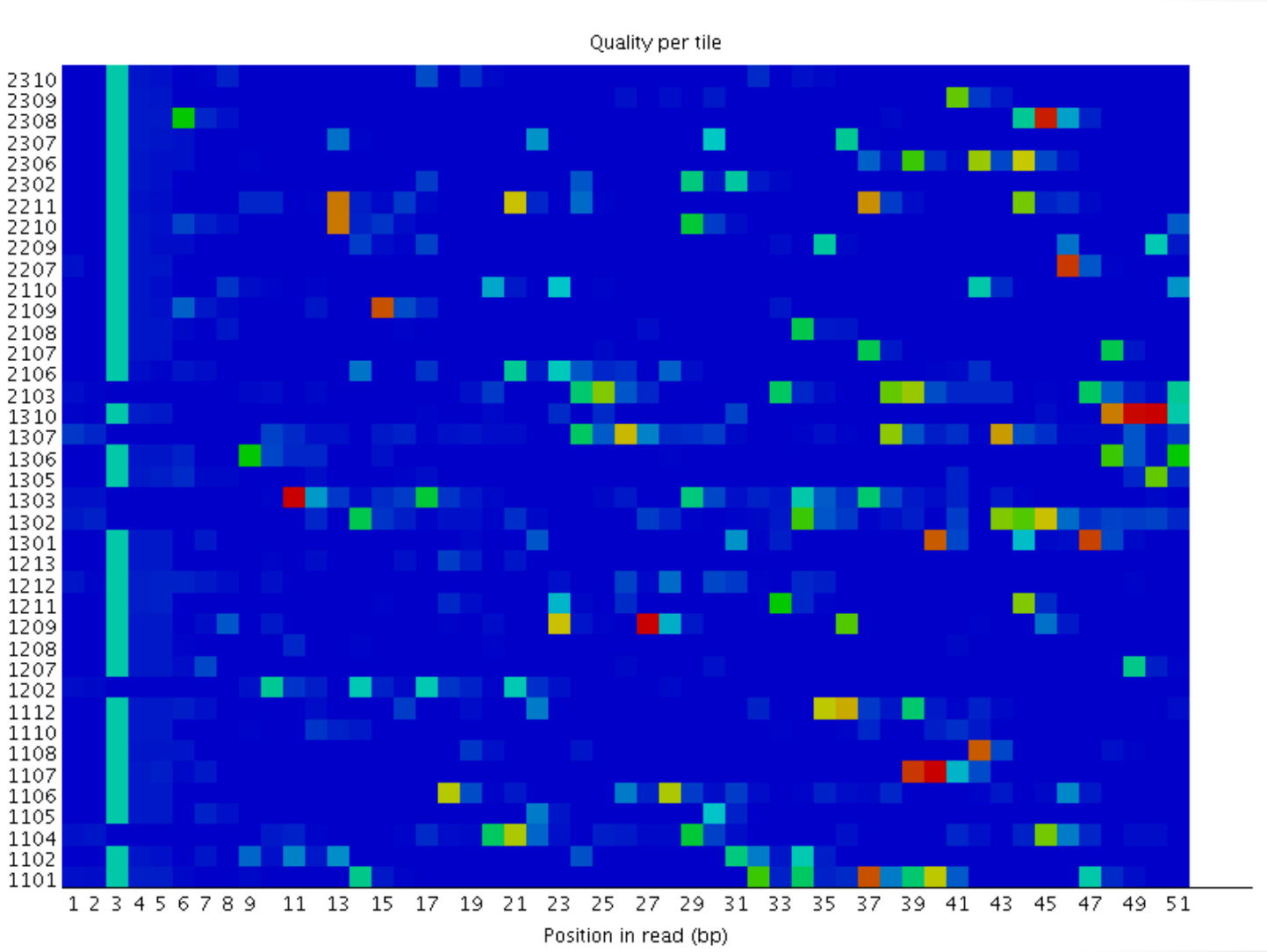


Read 1

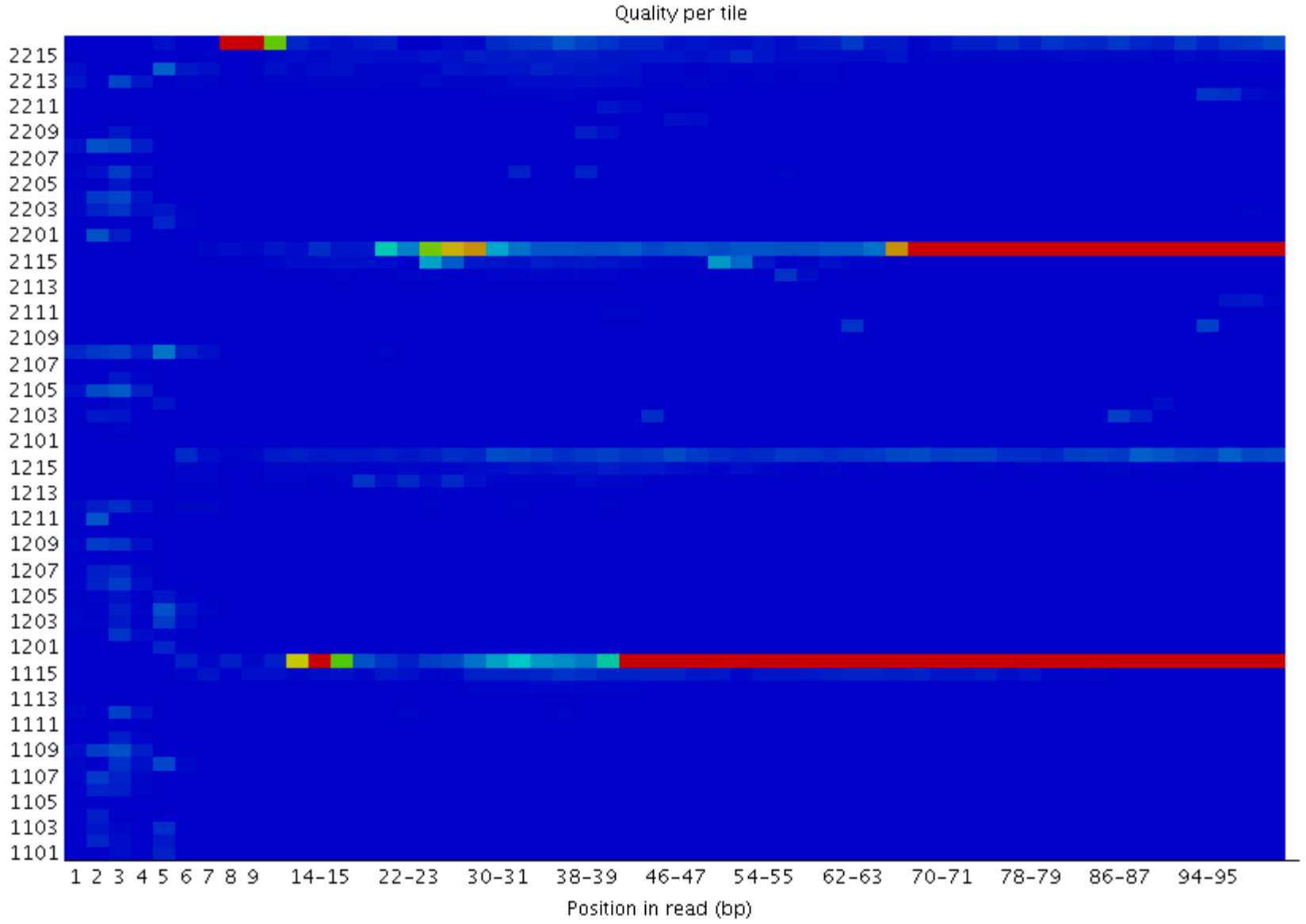


Read 2 (barcode)

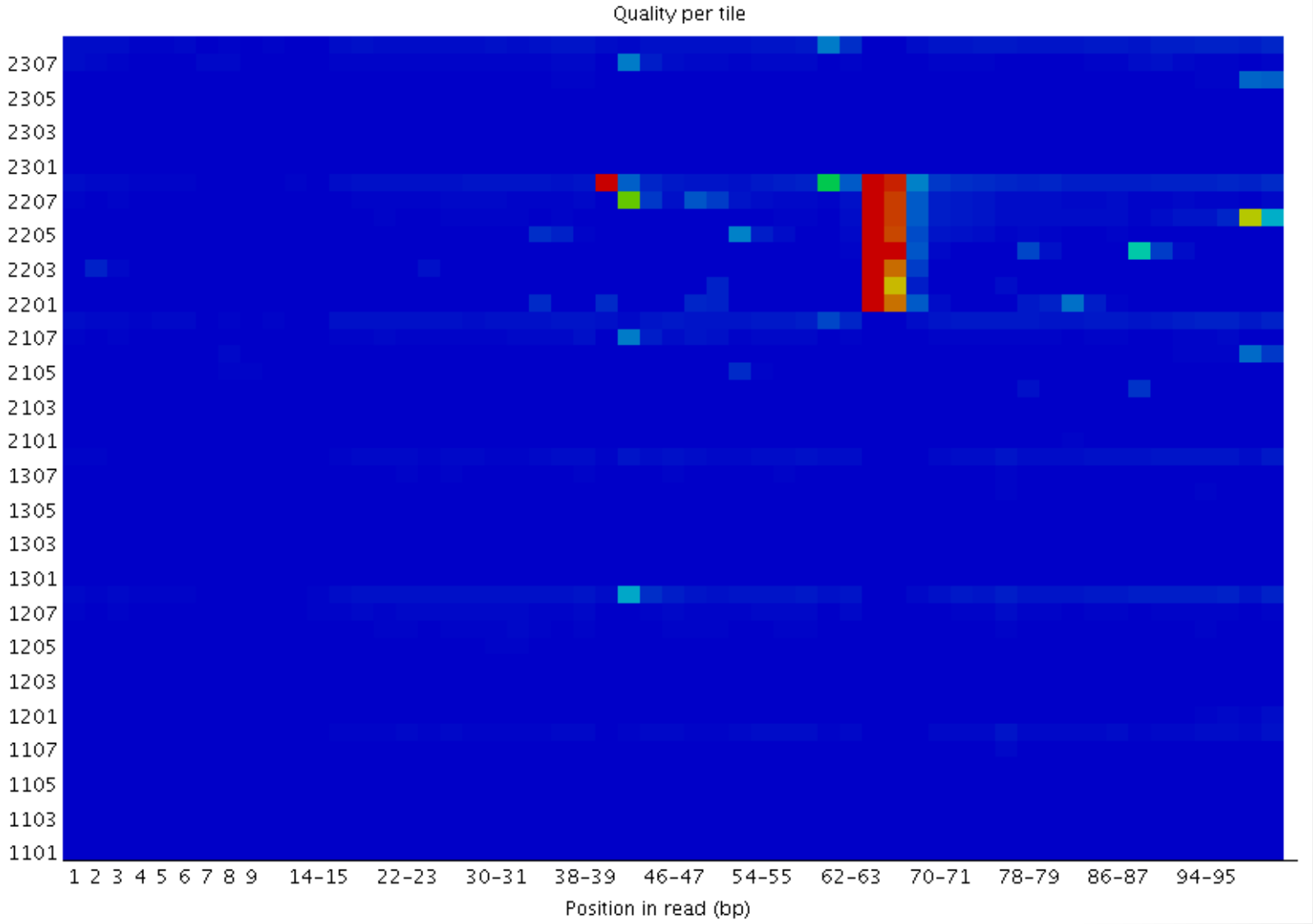
Tile Problems - Overclustering



Tile Problems – Consistent tile fail



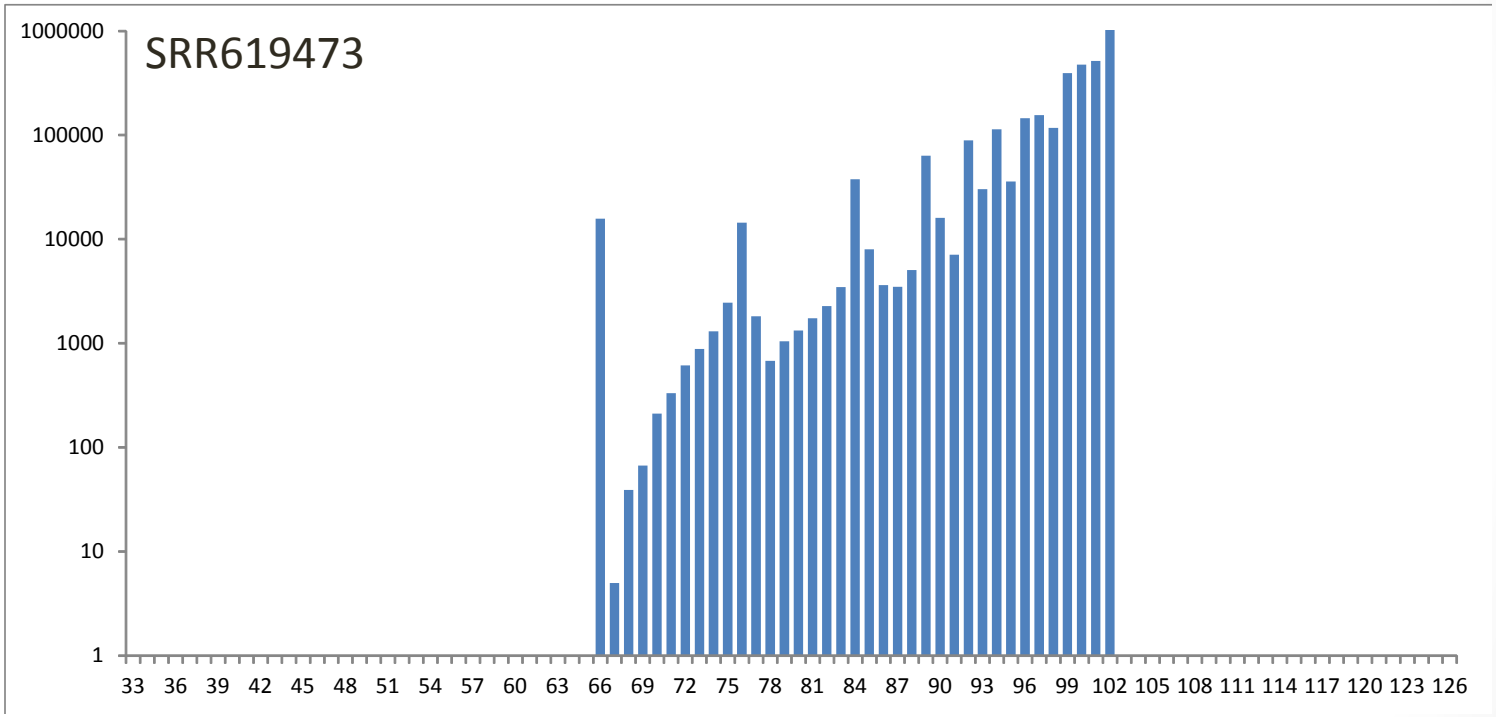
Tile problems – transient tile fail



Incorrect Phred Scores

"the NCBI SRA makes all its data available as standard Sanger FASTQ files (even if originally from a Solexa/Illumina machine)"

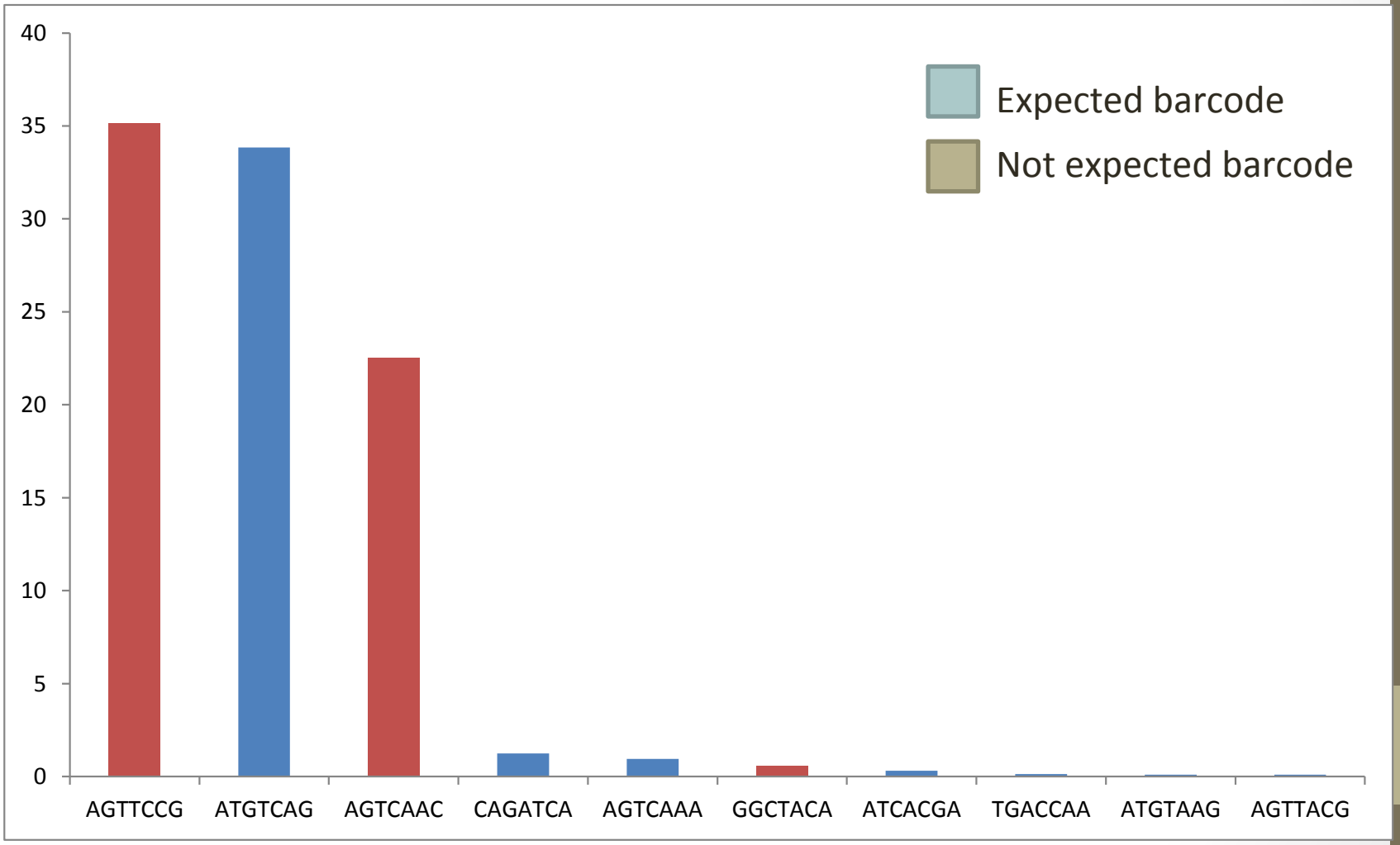
Nucleic Acids Res. 2010 Apr; 38(6): 1767–1771.



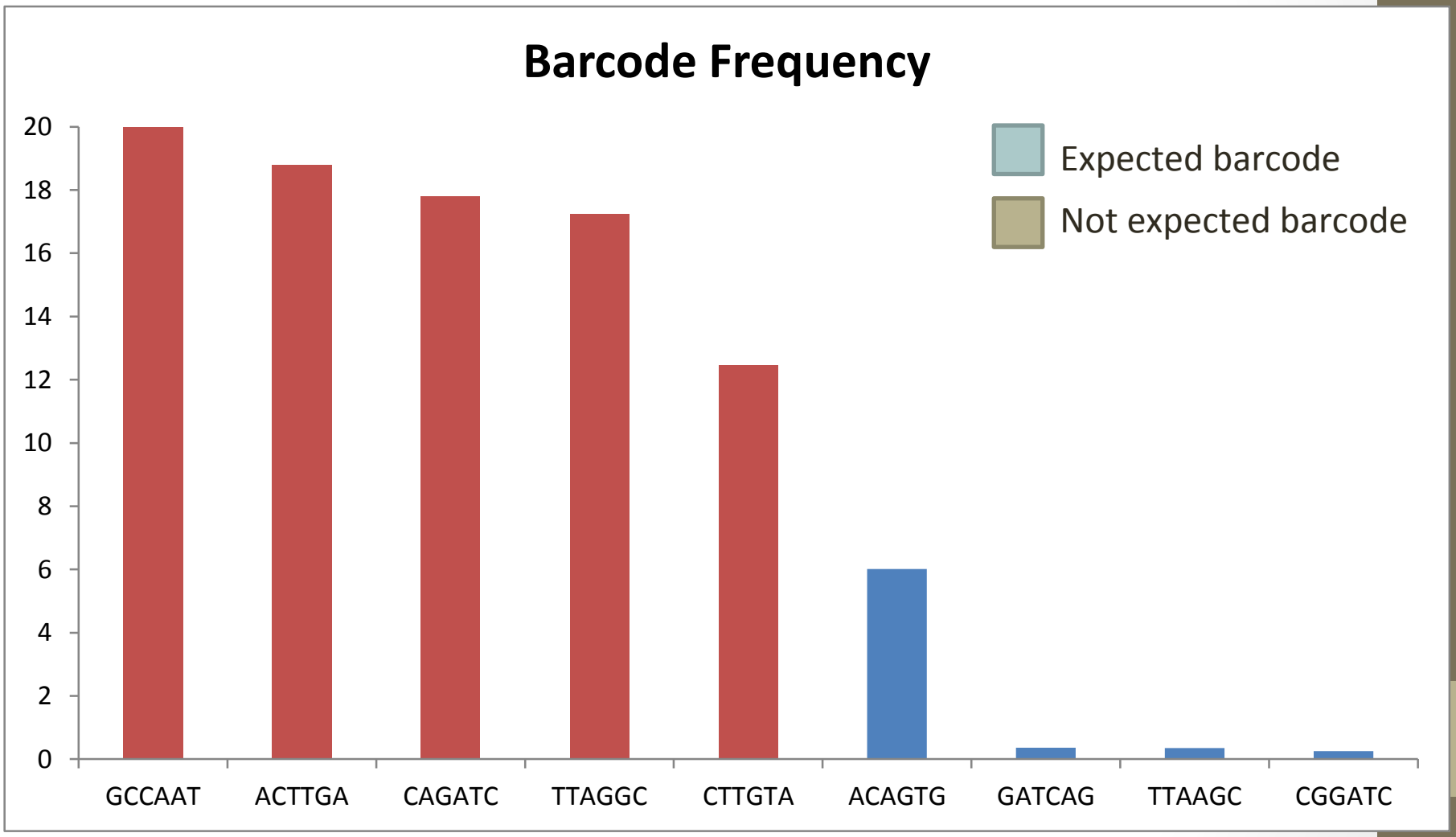
Found LOTS of examples of this in the SRA

Data Extraction

Wrong barcode annotation

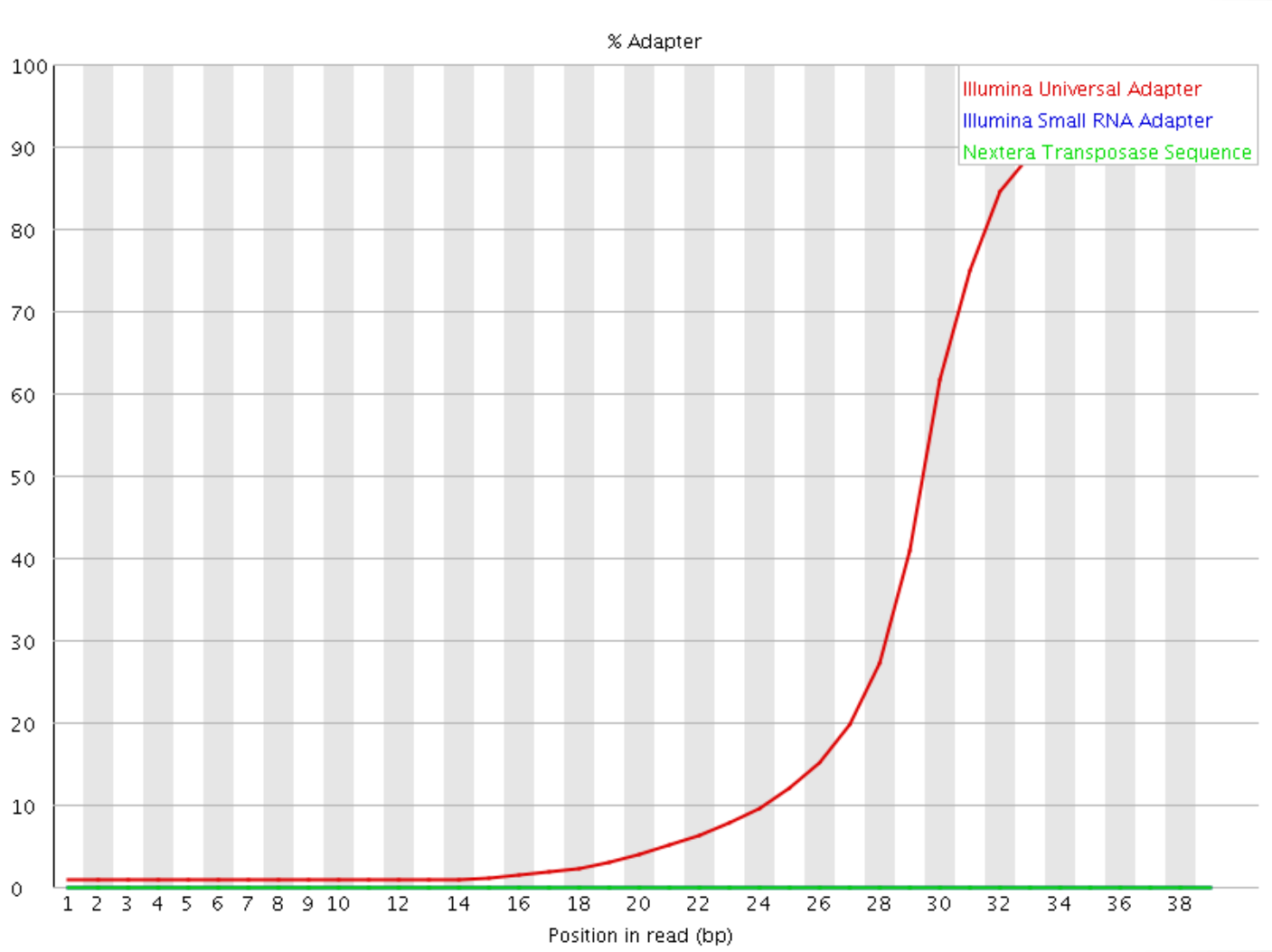


Contaminated Barcode Stocks



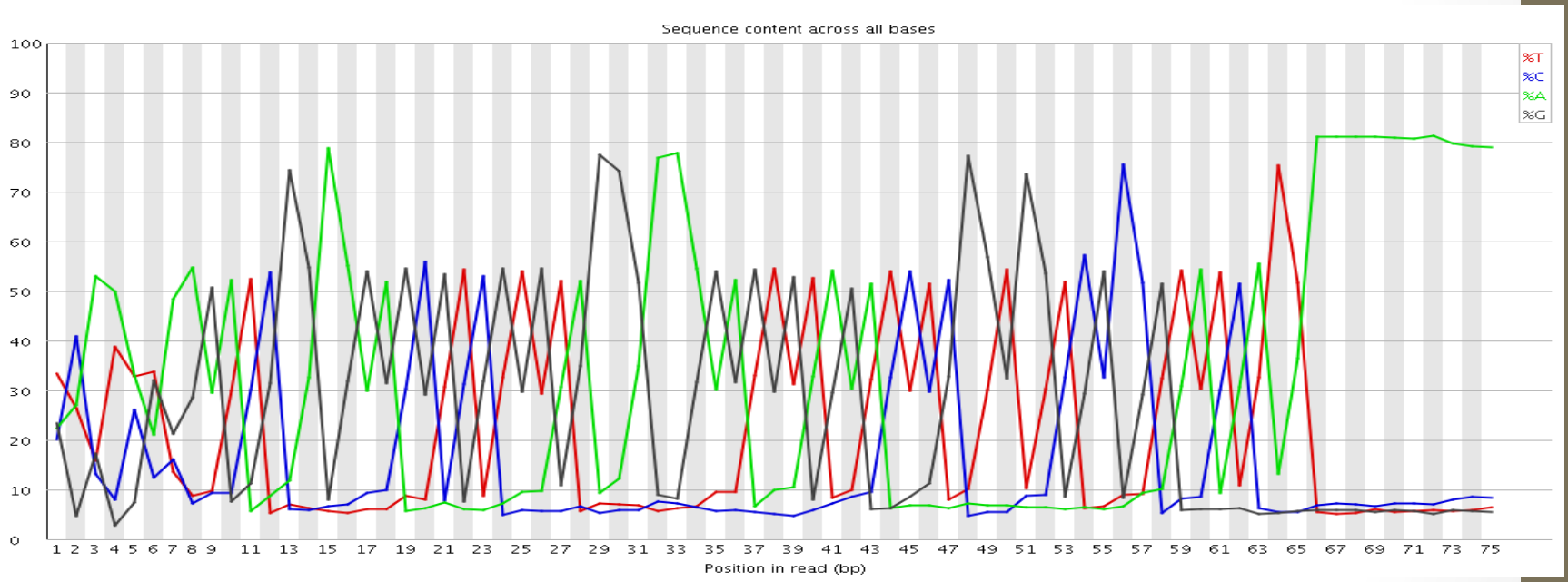
Odd sequence composition

Read through adapter



Adapter dimer overload

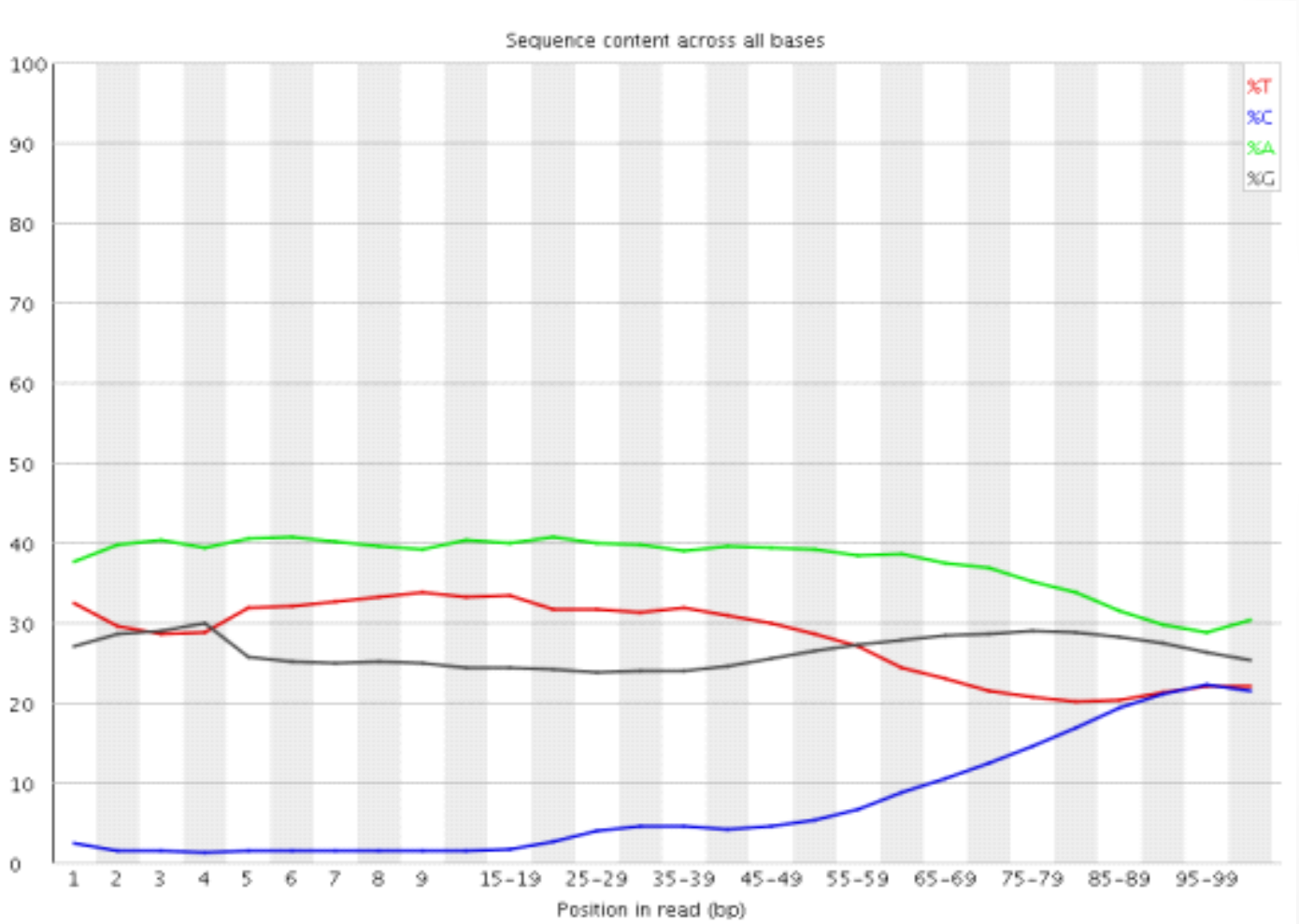
Sequence	%	Possible Source
CCTAAGGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAAAAAAA	9.42	Illumina Single End PCR Primer 1
TCAATGAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAAAAAAA	7.30	Illumina Single End PCR Primer 1
GAGACTCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAAAAAAA	5.65	Illumina Single End PCR Primer 1



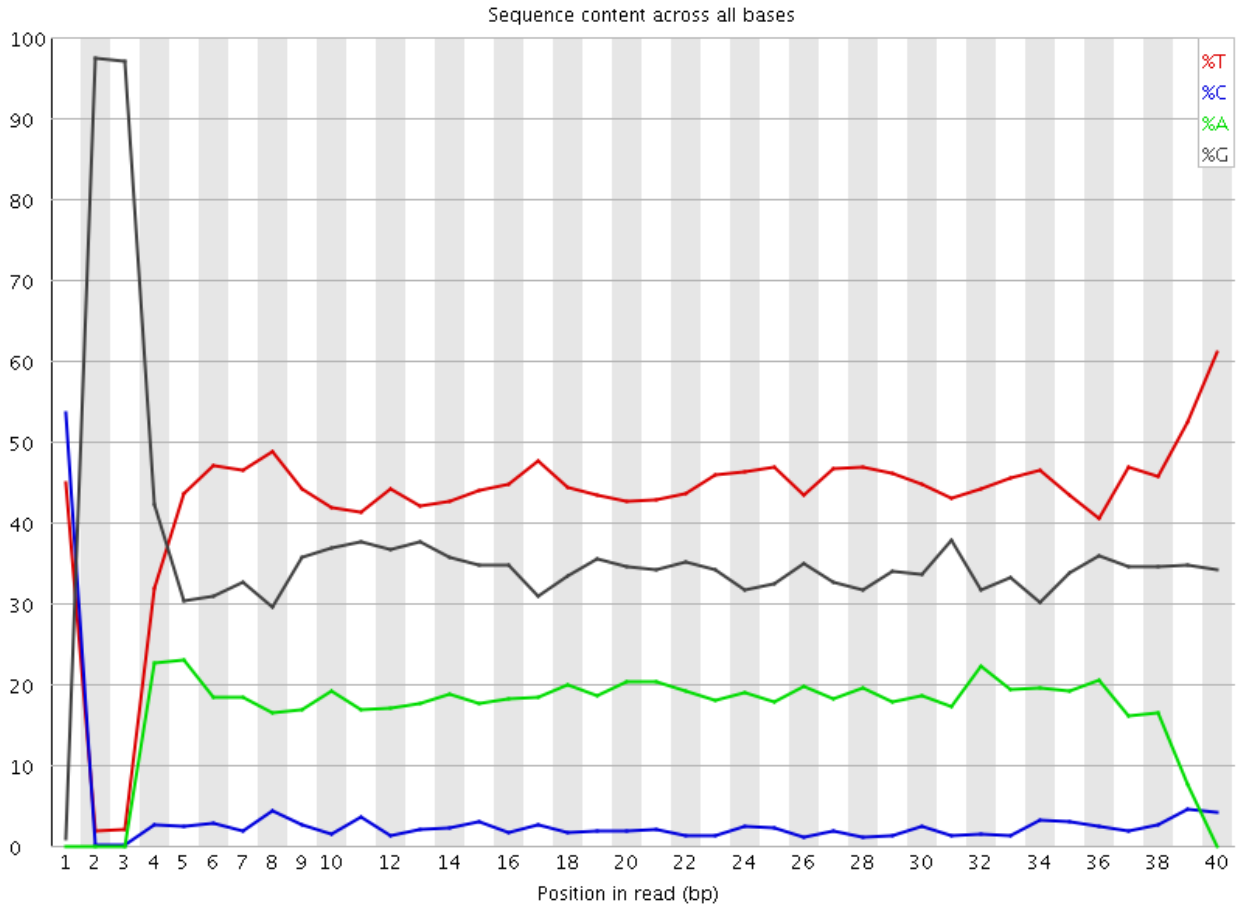

```
gi|372098977|ref|NT_039624.8| Mus musculus chr16 GRCm38
CTGGAAGGGAGAAAAGTCCAAACATTCTGGCTCTAACTTCT
|||||||||||||||||||||||||||||||| || ||||
CTGGAAGGGAGAAAAGTCCAAACATTCTGGCTCCAAGTTCT
gi|372098992|ref|NT_039500.8| Mus musculus chr10 GRCm38
CTTTCTCTATCTGAATTATAAACAAAAGCACACAGGCCCGCTTACATTTACATGATAAAATGTGCACTTTG
|||||||||| || |||||||||||||||||||||||||||||||   ||||||||||||||| | ||||
CTTTCTCTATATGCATTATAAACAAAAGCACACAGGCCCGCTTACAGGGACATGATAAAATGTGAAATTTG
```

(Single-cell Hi-C)

Positional Sequence Bias Application Specific – BS-Seq

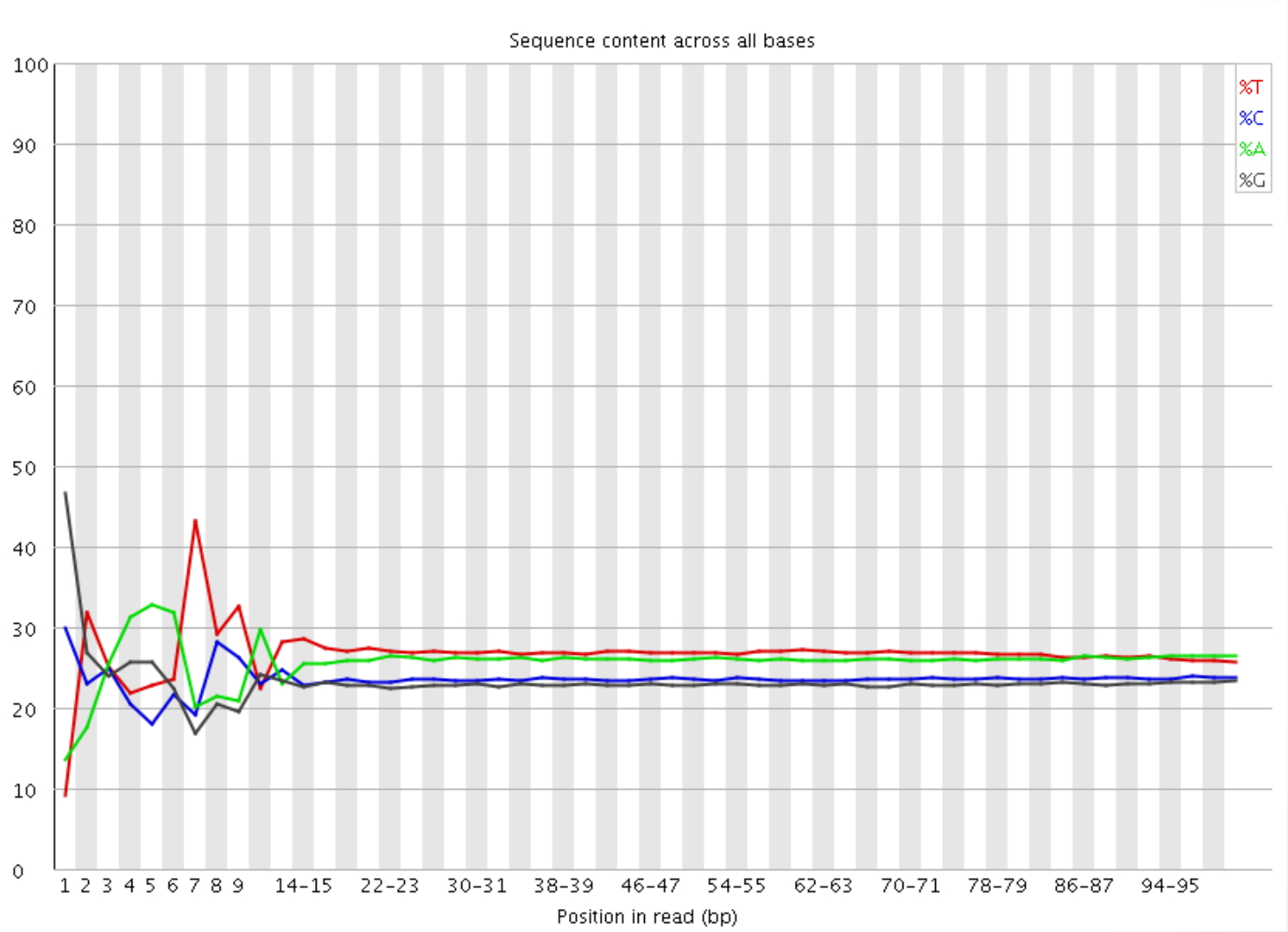


Positional Sequence Biases Expected - RRBS

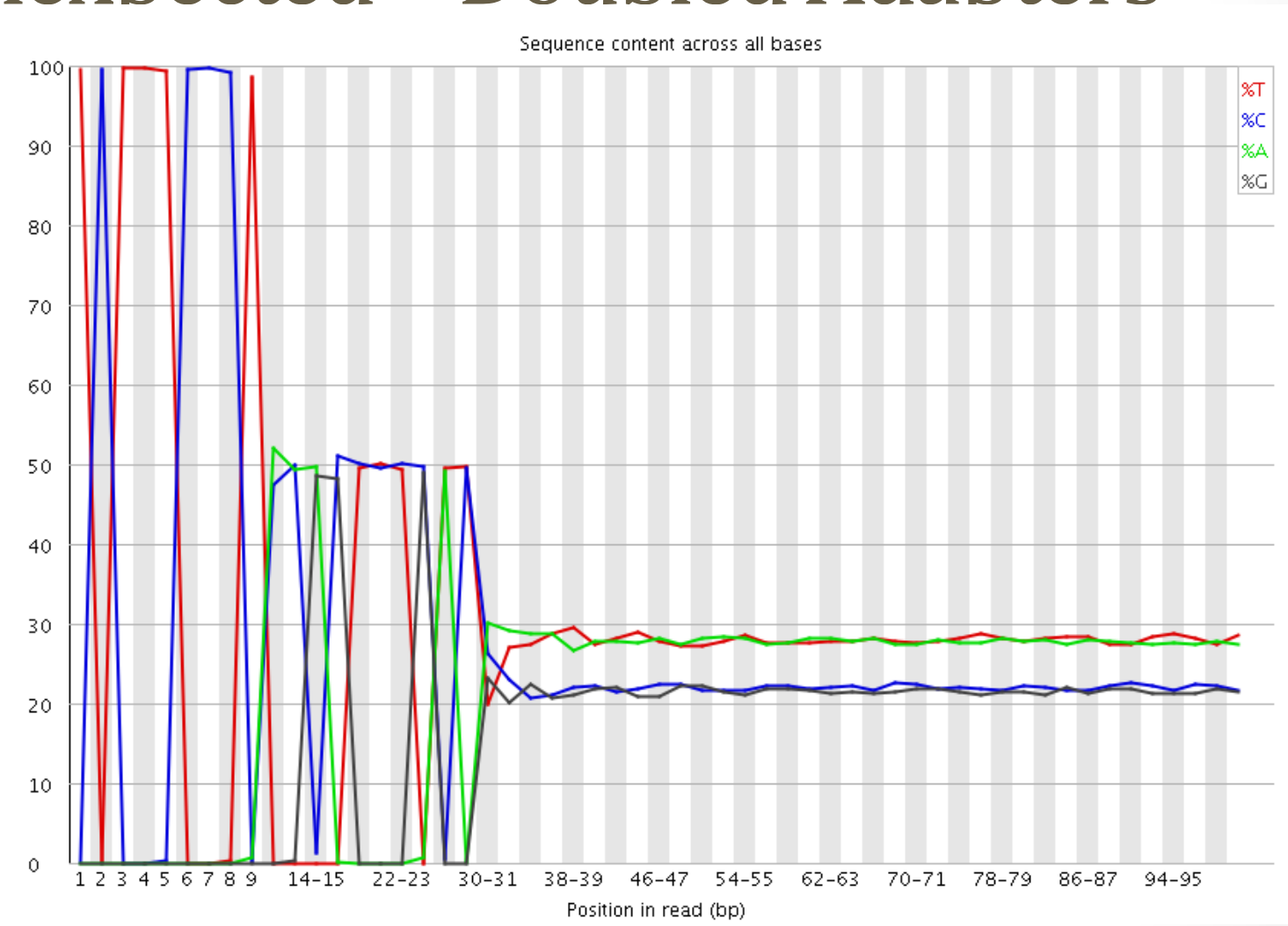


Also reports of a ‘Chinese CRO’ whose RRBS libraries have the MspI sites missing due to their proprietary and unexplained pre-processing

Positional Sequence Biases Unavoidable – RNA-Seq



Positional Sequence Biases Unexpected – Doubled Adapters

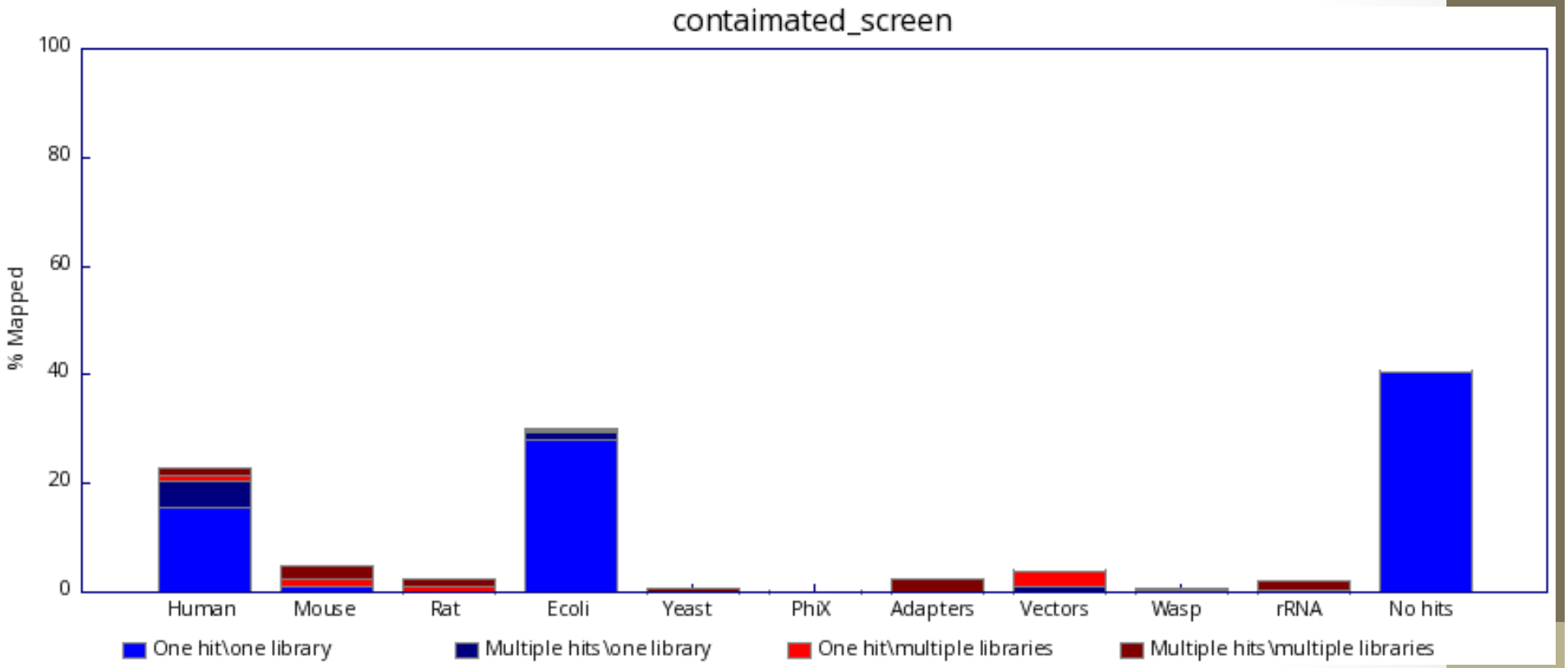


Overrepresented Individual Sequences

- Adapter dimers
- rRNA
- Satellite sequences

My data doesn’t map well…

Contaminated with guessable sequence



www.bioinformatics.babraham.ac.uk/projects/fastq_screen

Contaminated with guessable sequence

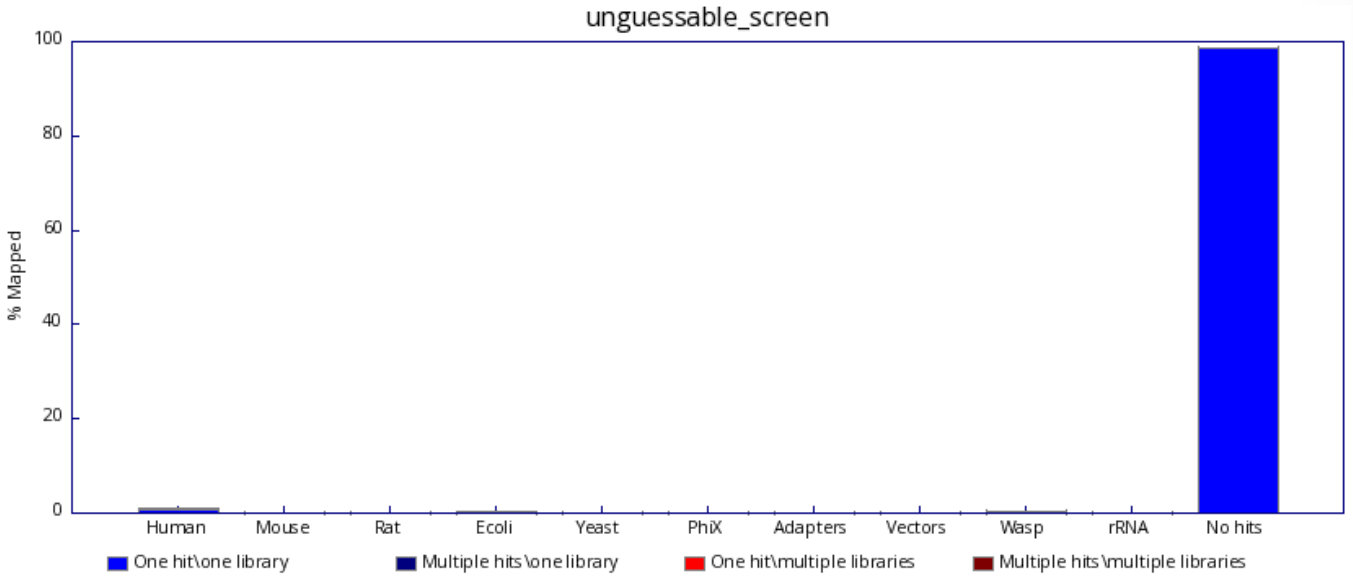
4

Yield (Gbases): 8.11
Sequences: 162,116,141
Sampled: 100,000

Reference ID	Species/Reference Genome	Aligned	Aligned %	Error rate	Assigned	Assigned %
Hsa.GRCh37	Homo sapiens (human)	91337	91.3%	0.28%	91337	91.3%
phix174	Phi X 174	4483	4.5%	0.15%	4483	4.5%
Ptr.CHIMP2	Pan troglodytes (chimpanzee)	86152	86.2%	1.21%	641	0.6%
Ggo.gorGor3	Gorilla gorilla	83354	83.4%	1.47%	104	0.1%
fungi.RefSeq	Fungi	3092	3.1%	4.17%	33	0.0%
Nle.Nleu1	Nomascus leucogenys (northern white cheeked gibbon)	67793	67.8%	2.40%	27	0.0%
Cja.calJac1	Callithrix jacchus (marmoset)	29717	29.7%	3.08%	7	0.0%
Mml.MMUL1	Macaca mulatta (macaque)	52073	52.1%	2.90%	6	0.0%
Hsa.NCBI36	Hsa.NCBI36	91106	91.1%	0.28%	5	0.0%
Cfa.BROADD2	Canis familiaris (dog)	3093	3.1%	3.95%	1	0.0%
Xtr.JGI4_1	Xenopus tropicalis (Western clawed frog)	4796	4.8%	3.81%		
Hsa.GRCh37.assembled	Hsa.GRCh37.assembled	91065	91.1%	0.29%		
Hsa.NCBI36.assembled	Hsa.NCBI36.assembled	90978	91.0%	0.29%		
Other	20 others				126	0.1%
Unmapped		3230	3.2%			
Adapter		0	0.0%			

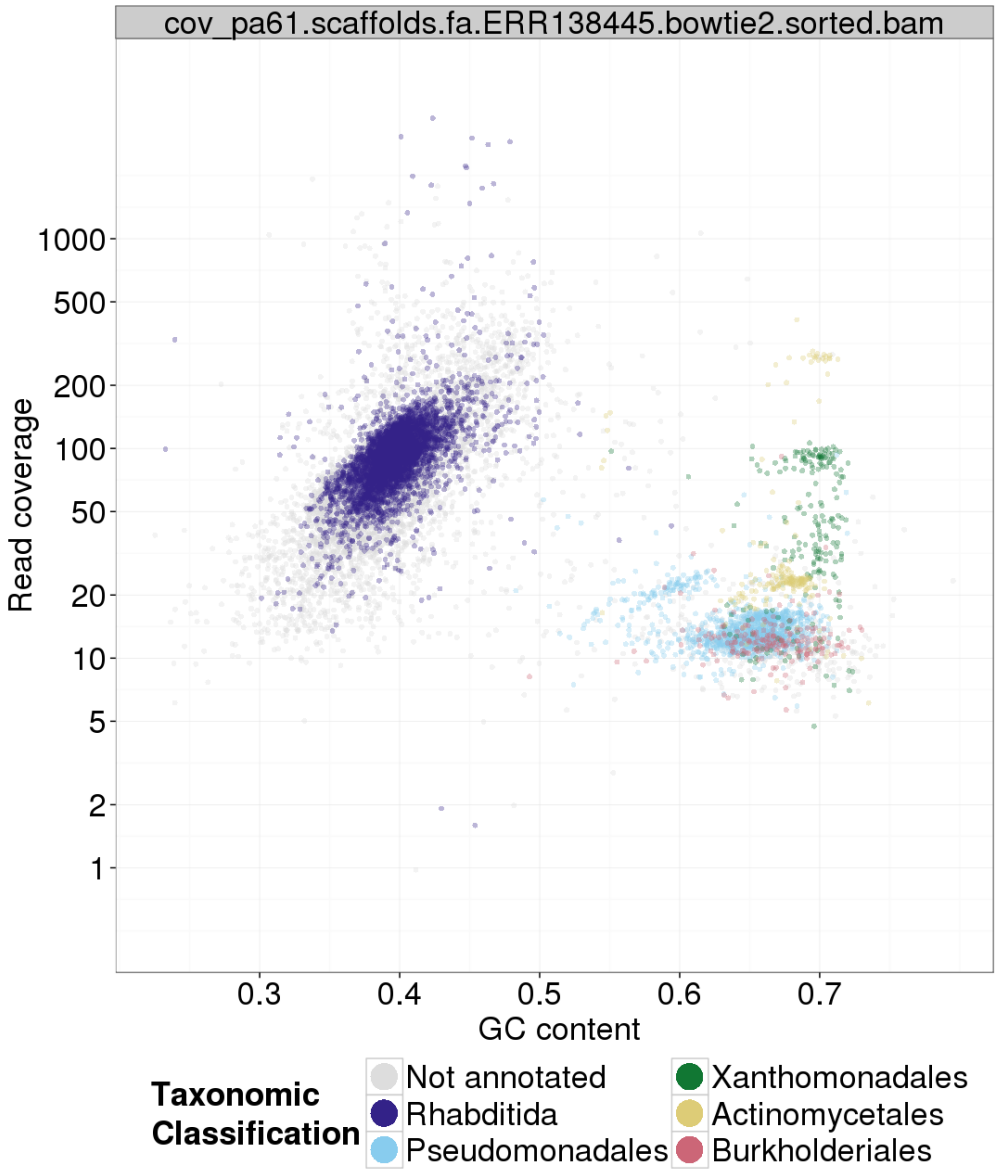
CRUK Multi-genome alignment system (MGA)

Contamination with unguessable sequence




```
>AF431889 AF431889.1 Acinetobacter lwoffii type IIs modification

Query: 1    cggtgagcaggcattagaaattgatttttagaaggtgtgttgaagaaactgggccgctt 60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||| |||||
Sbjct: 4661 cggtgagcaggcattagaaattgatttttagaaggtgtgttgaagaaactgggtcgctt 4720

>GQ352402 GQ352402.1 Acinetobacter baumannii strain AbSK-17 plasmid

Query: 1    ggtgagcagtggtttacatggttaattgaacaagacatcaacttctgcattcgtg 55
             |||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8213 ggtgagcagtggtttacatggttaattgaacaagacatcaacttctgcattcgtg 8159

>AF431889 AF431889.1 Acinetobacter lwoffii type IIs modification

Query: 1    acttgctgcgattaaagcagaaaaaacacttgctgaattgagtgct 46
             |||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 4484 acttgctgcgattaaagcagaaaaaacacttgctgaattgagtgct 4529
```

TAGC Plots



Assemble

↓

Filter contigs

↓

Plot %GC vs Coverage

↓

Sample and blast

https://github.com/blaxterlab/blobology

Reagent contamination

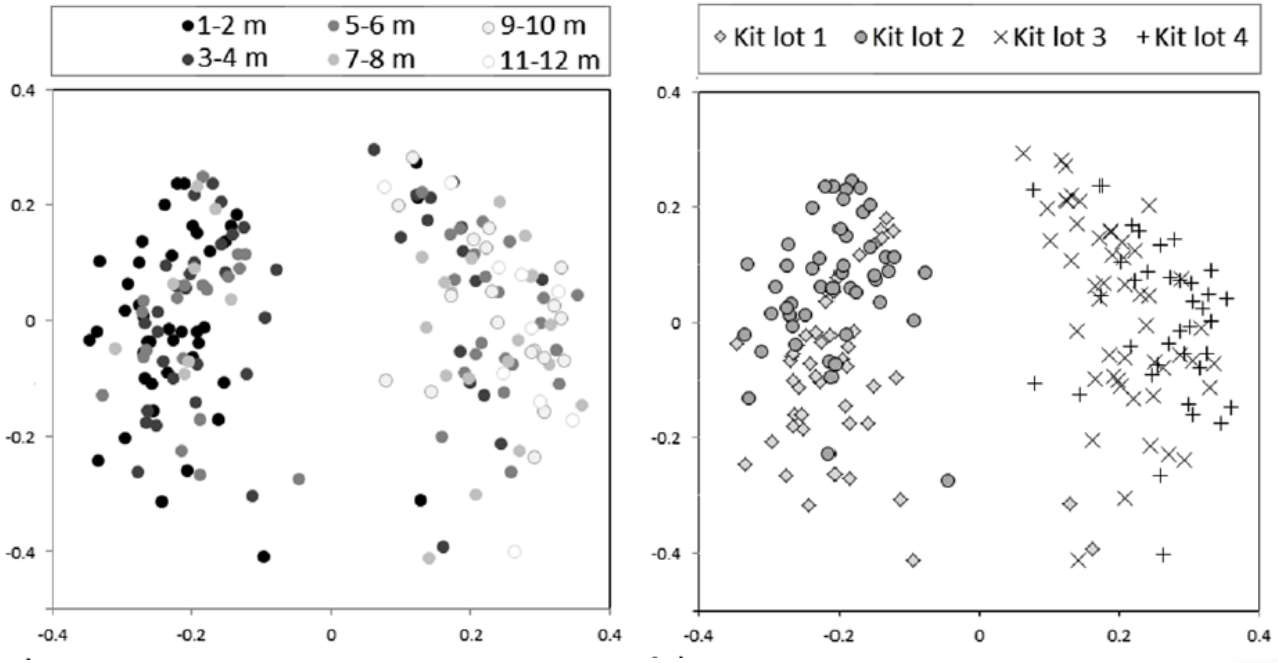
Salter *et al. BMC Biology* 2014, **12**:87
http://www.biomedcentral.com/1741-7007/12/87

BMC Biology

RESEARCH ARTICLE Open Access

Reagent and laboratory contamination can critically impact sequence-based microbiome analyses

Susannah J Salter[1*], Michael J Cox[2], Elena M Turek[2], Szymon T Calus[3], William O Cookson[2], Miriam F Moffatt[2], Paul Turner[4,5], Julian Parkhill[1], Nicholas J Loman[3] and Alan W Walker[1,6*]



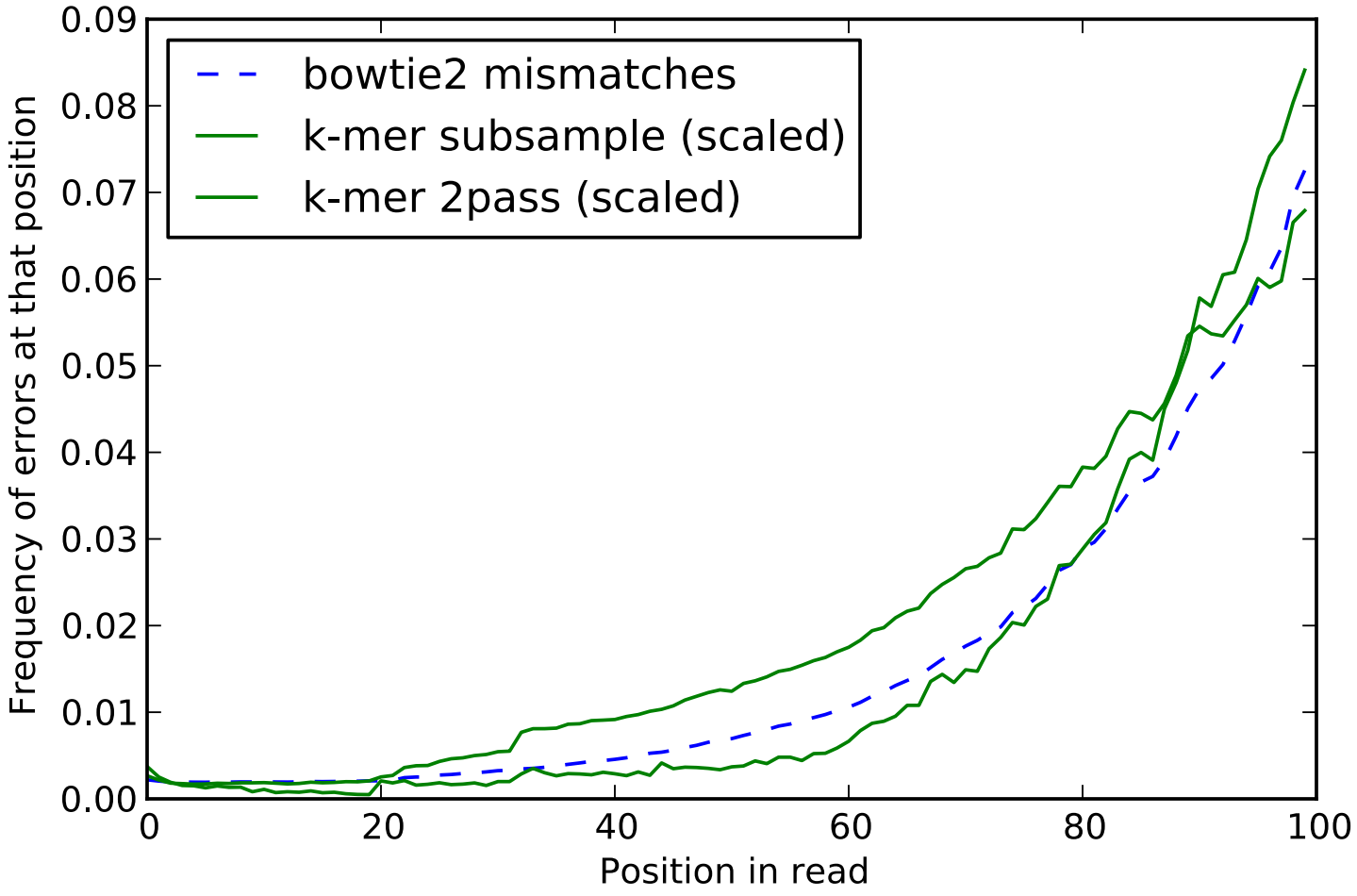
Molbio grade water is not the same as DNA free water – heat treated but DNA survives

Later this week --

Many different approaches to evaluating quality/mismatches:

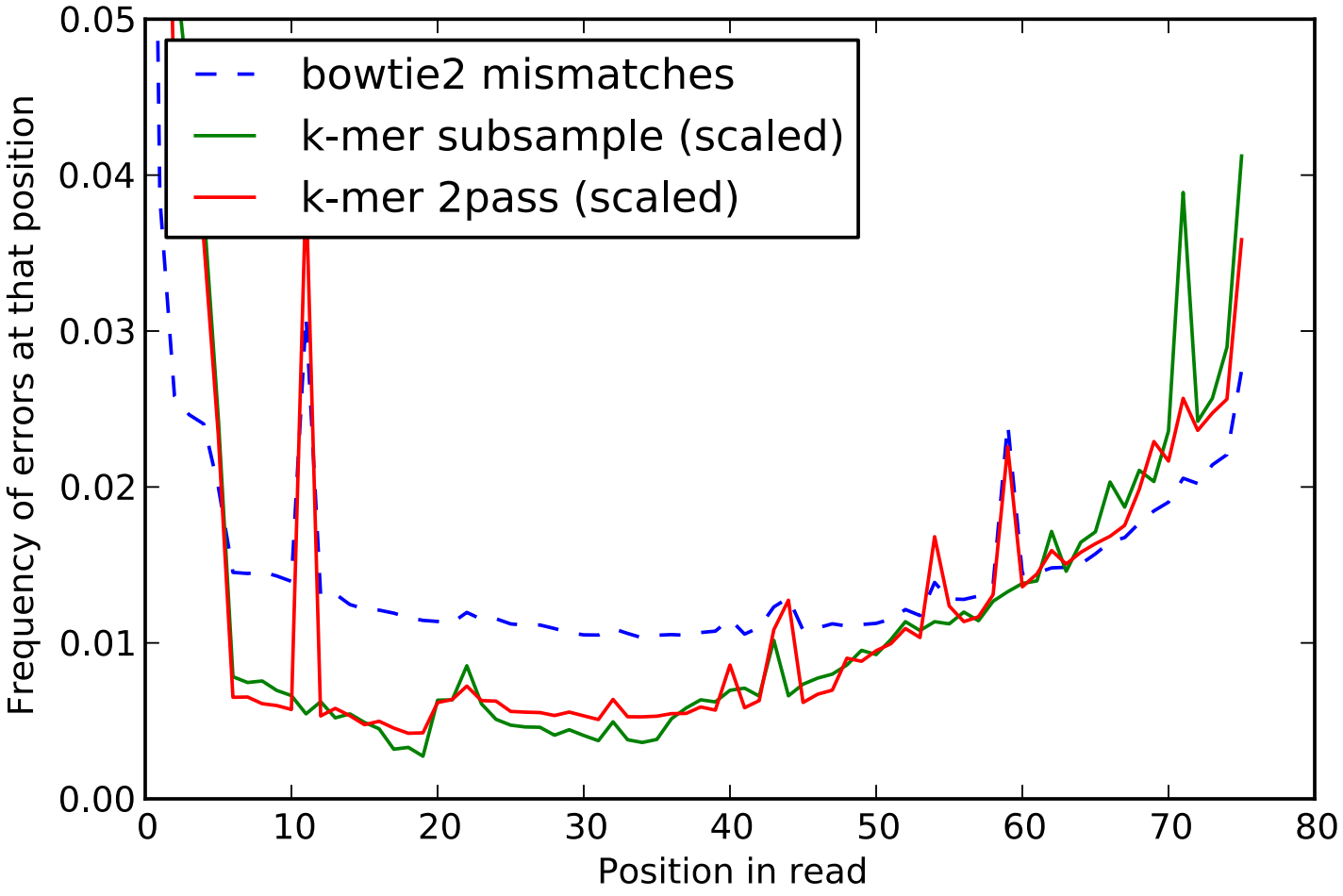
1. Quality-score based (FastQC etc)
2. Composition based (FastQC etc)
3. Reference based ("I know what the answer should look like")
4. Assembly-graph / k-mer based

Reference & quality-score independent approaches (k-mers)



Zhang et al., https://peerj.com/preprints/890/

...from a well known data set...



Zhang et al., https://peerj.com/preprints/890/