High Throughput Sequencing the Multi-Tool of Life Sciences

Lutz Froenicke

DNA Technologies and Expression Analysis Cores

UCD Genome Center

DNA Technologies & Expression Analysis Cores

- HT Sequencing (Illumina & PacBio)
- Illumina microarray (for genotyping – Illumina has discontinued expression analysis)

- consultations
- introducing new technologies to campus
- shared equipment (accessible after training)
- teaching (workshops)

Complementary Approaches

Illumina	PacBio
Still-imaging of clusters (~1000 clonal molecules)	Movie recordings fluorescence of single molecules
Short reads - 2x300 bp Miseq	Up to 60 kb, N50 23 kb
Repeats are mostly not analyzable	spans retro elements
High output - up to 100 Gb per lane	up to 1,3 Gb and 5 Gb per SMRT-cell
High accuracy (< 0.5 %)	Error rate 15 %
Considerable base composition bias	No base composition bias
Very affordable	Costs 5 to 10 times higher
De novo assemblies of thousands of scaffolds	"Near perfect" genome assemblies

Genome Resequencing

Rearrangements

SNPs, Indels CNVs

De novo genome Sequencing

Metagenomics

Exome Sequencing

RNA-seq Gene Expression

High Throughput Short Read Sequencing: Illumina

DNA Methylation

Splice Isoform Abundance

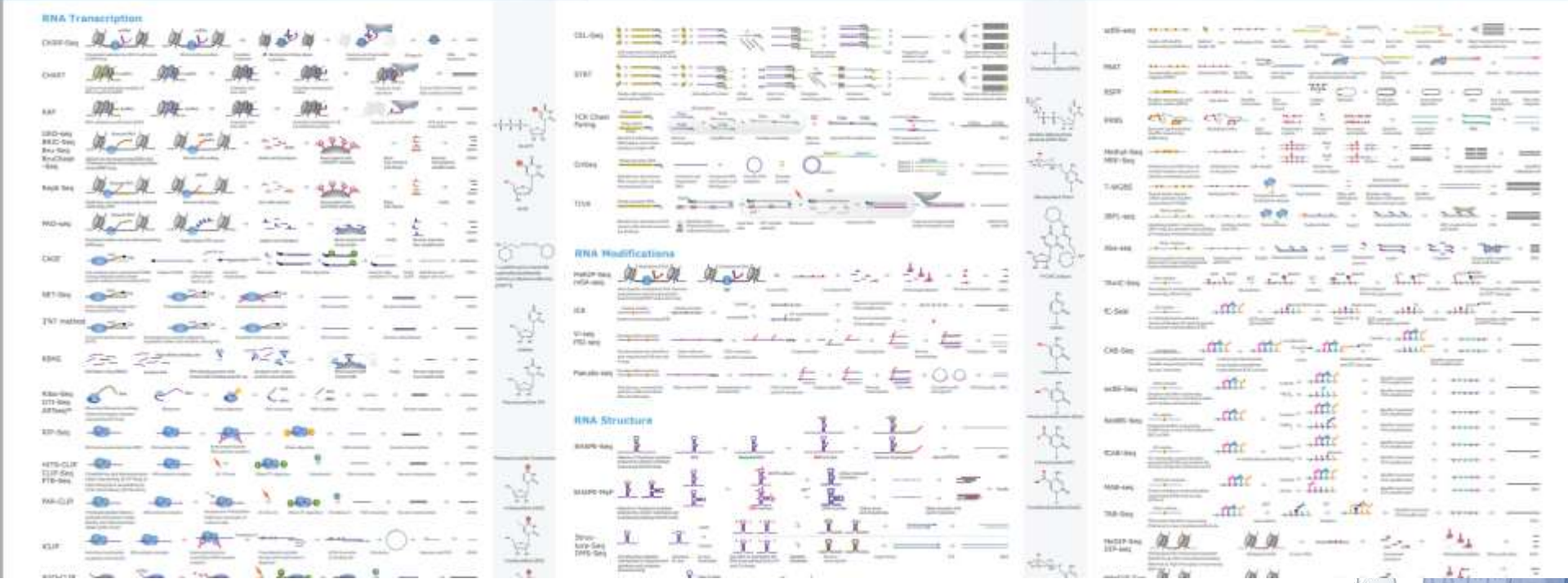
ChIP-SEQ

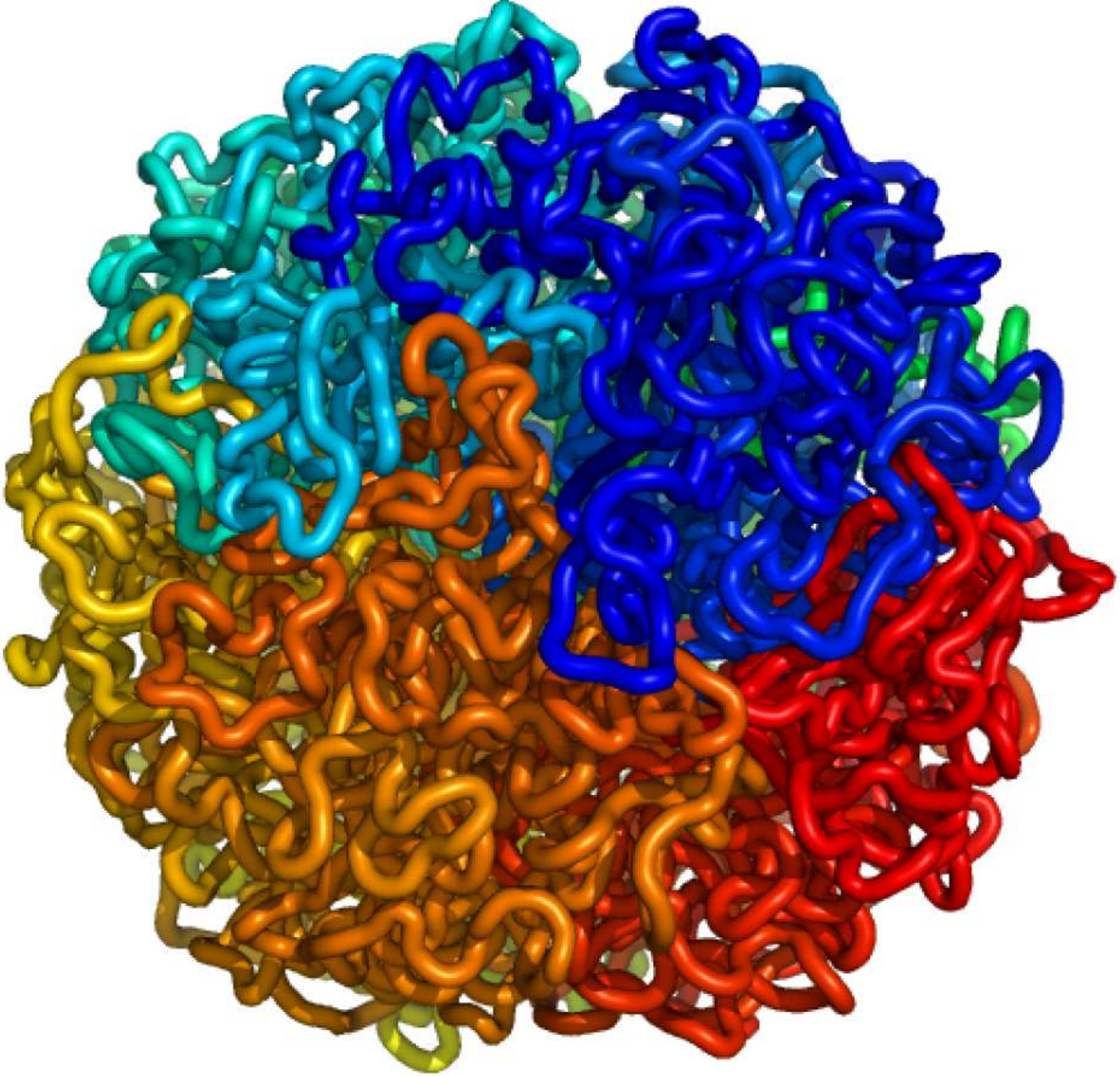
3D Organization

Genotyping

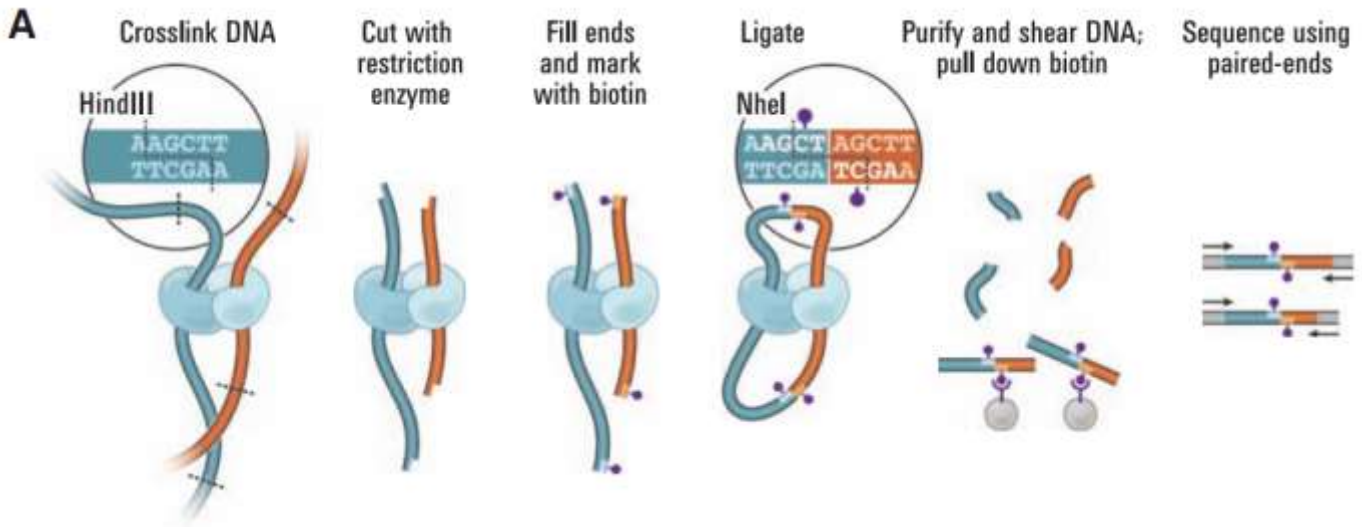
Small RNA

For all you seq...



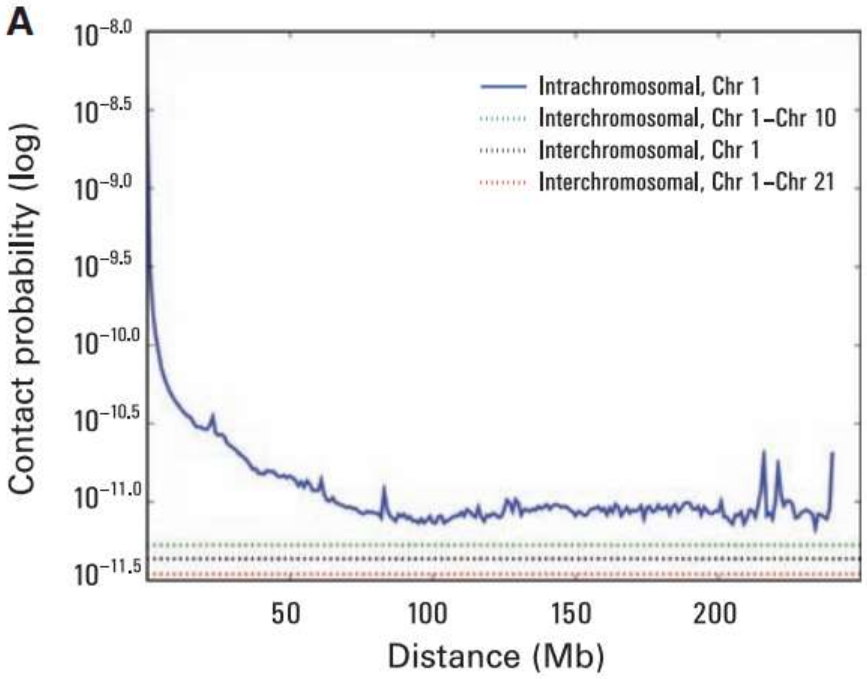


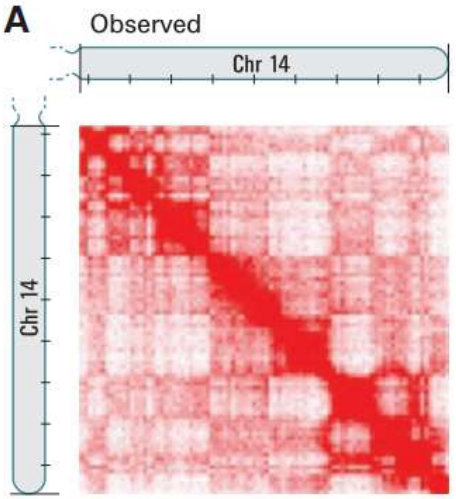
Hi-C



Lieberman-Aiden 2010

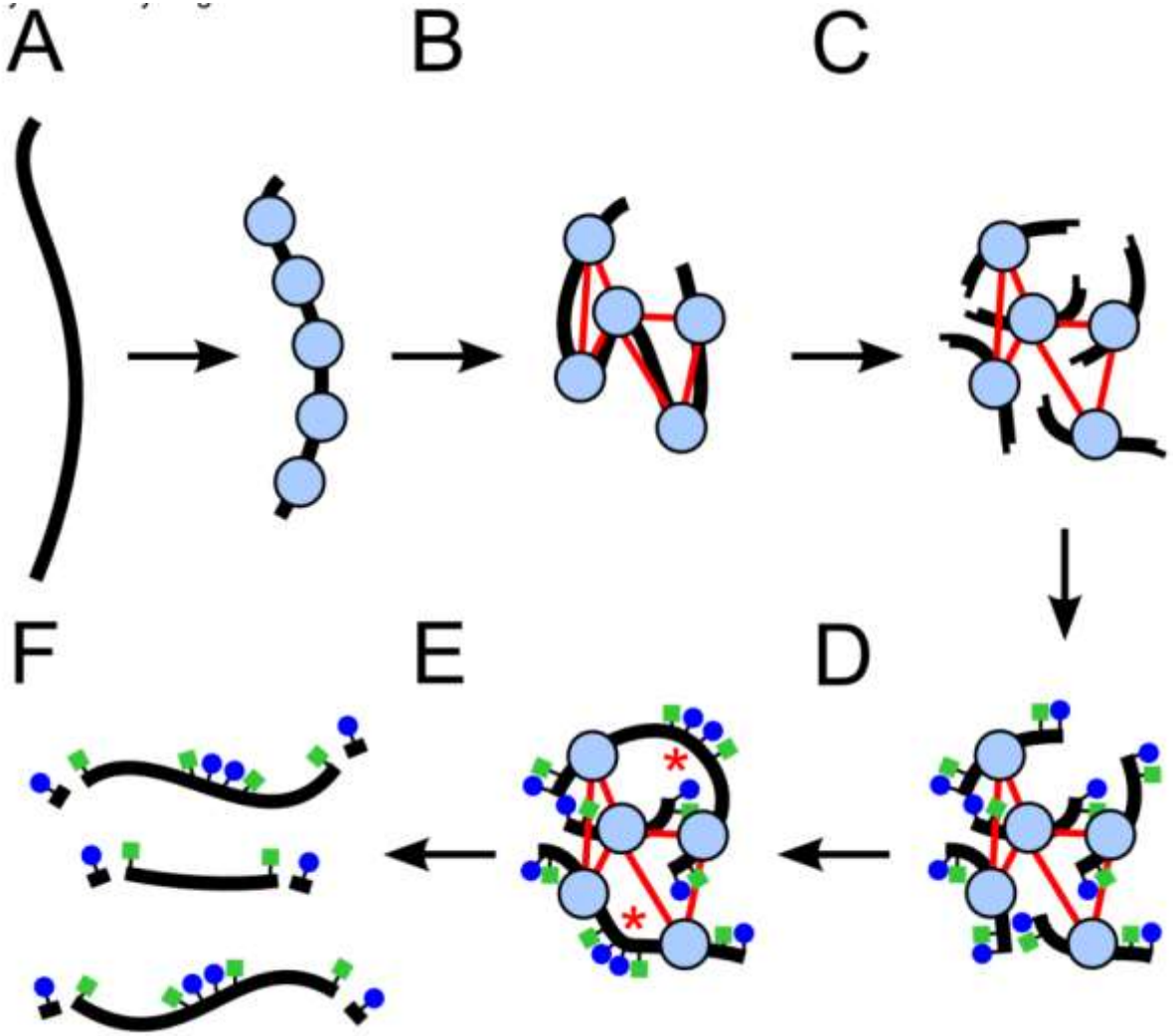
Hi-C



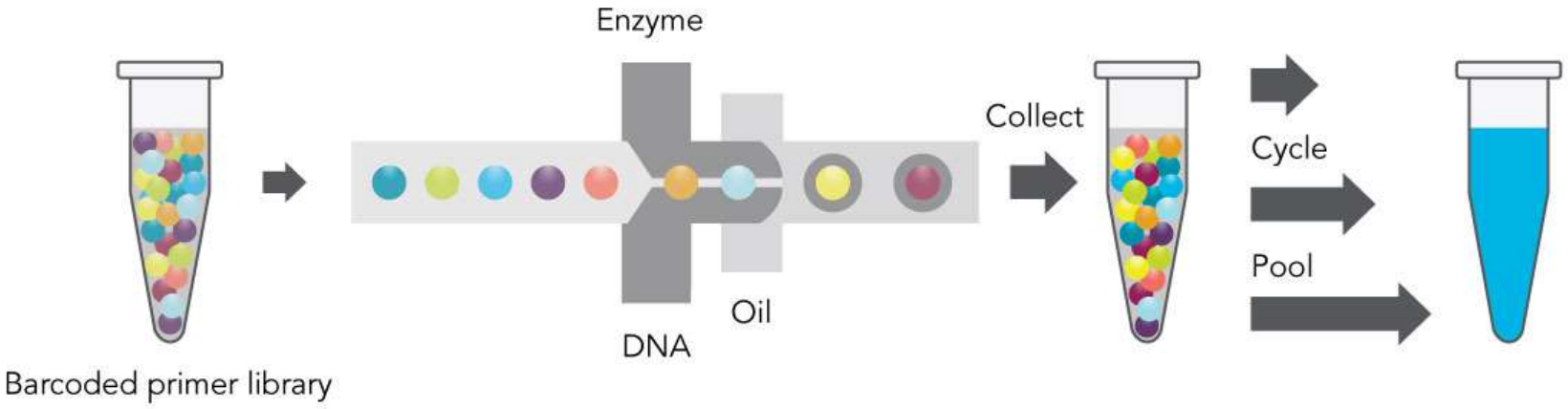


Lieberman-Aiden 2010

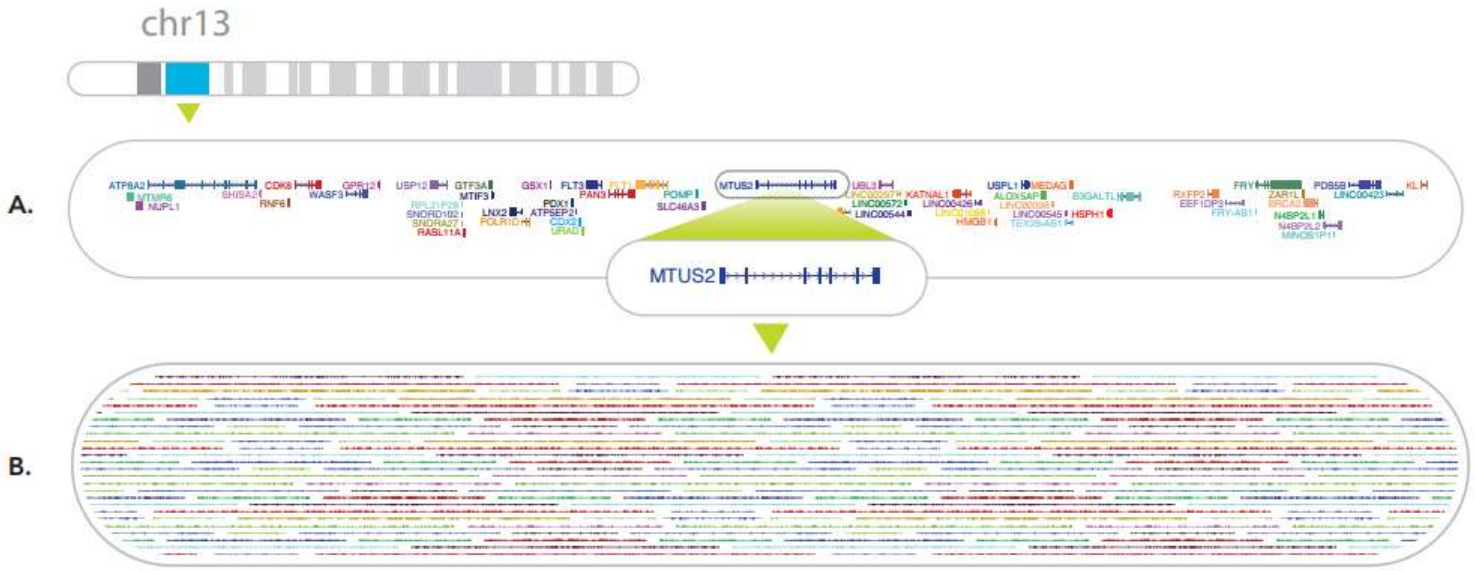
Dovetail Sequencing (Putnam 2015)



10X Genomics

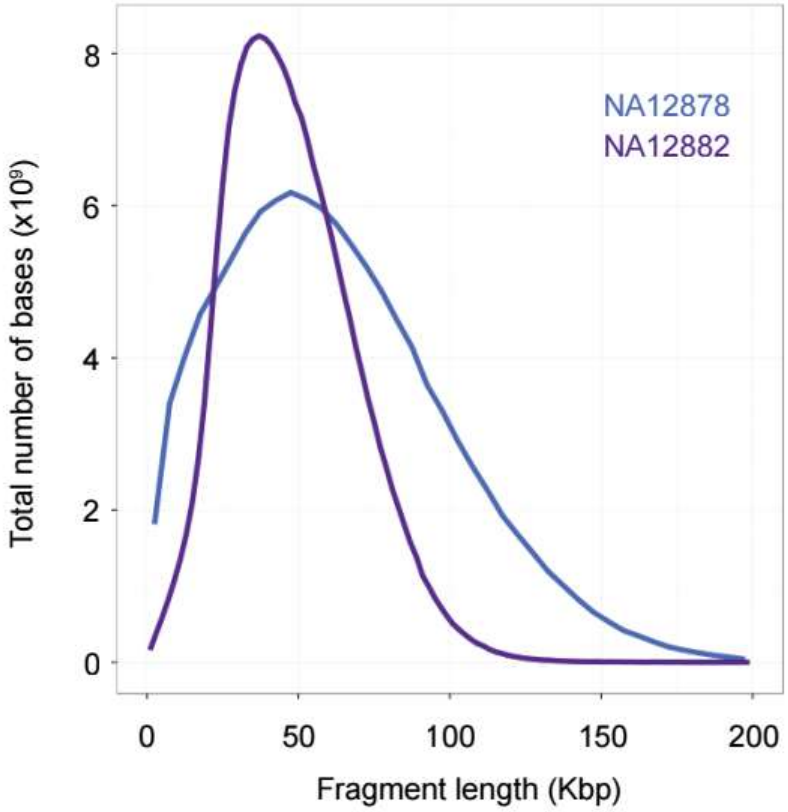

Enzyme
Collect
Cycle
Pool
Oil
DNA
Barcoded primer library

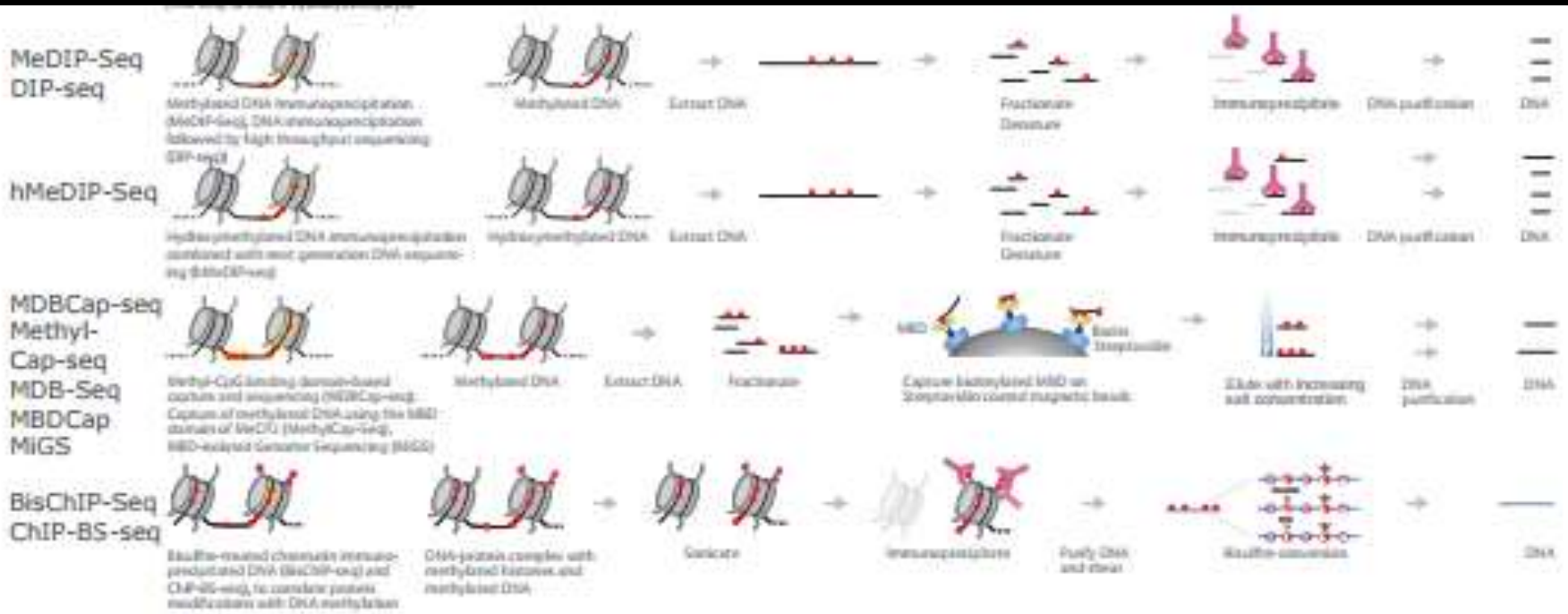
25x linked –read coverage



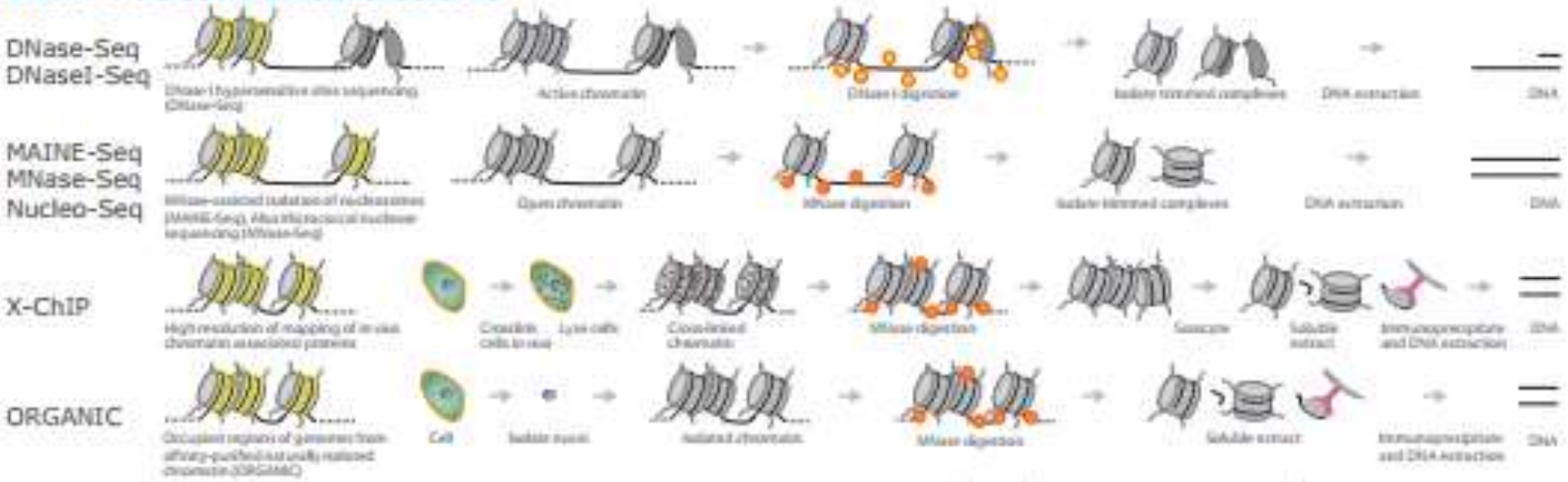
- 60 kb deletion


10X GENOMICS
Loupe · Sample 5728
File
SUMMARY
PHASING
STRUCTURAL VARIANTS
chr2+29989813-chr2+30039813,chr2+42493841-chr2+42543841
chr2
29,995,000
30,000,000
30,005,000
30,010,000
30,015,000
30,020,000
30,025,000
30,030,000
30,035,000
HAPLOTYPE 1
UNPHASED
HAPLOTYPE 2
GENES
EXONS
COVERAGE
BREAKPOINTS
ALK
SV #61
chr2
42,495,000
42,500,000
42,505,000
42,510,000
42,515,000
42,520,000
42,525,000
42,530,000
42,535,000
42,540,000
EML4
SV #61
SV #71
SV #279


NA12878
NA12882
Total number of bases (x10⁹)
0
2
4
6
8
0
50
100
150
200
Fragment length (Kbp)



DNA-Protein Interactions



Genome Resequencing

Rearrangements

SNPs, Indels CNVs

De novo genome Sequencing

Metageno mics

Exome Sequencing

RNA-seq Gene Expression

Long Read Sequencing PacBio , (Nanopore)

DNA Methylation

Splice Isoform Abundance

ChIP-SEQ

3D Organization

Genotyping

Small RNA

Illumina sequencing workflow

- Library Construction
- Cluster Formation
- Sequencing
- Data Analysis

Fragmentation

- Mechanical shearing:
  - BioRuptor
  - Covaris

  DNA, RNA
- Enzymatic:
  - Fragmentase, RNAse3

  DNA, RNA
- Chemical: Mg2+, Zn2+ → **RNA**

DNA library construction

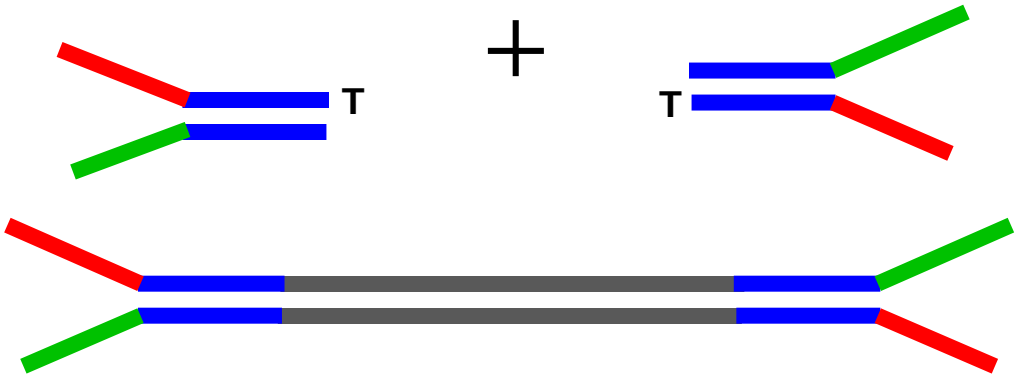
Fragmented DNA

End Repair

Blunt End Fragments

"A" Tailing

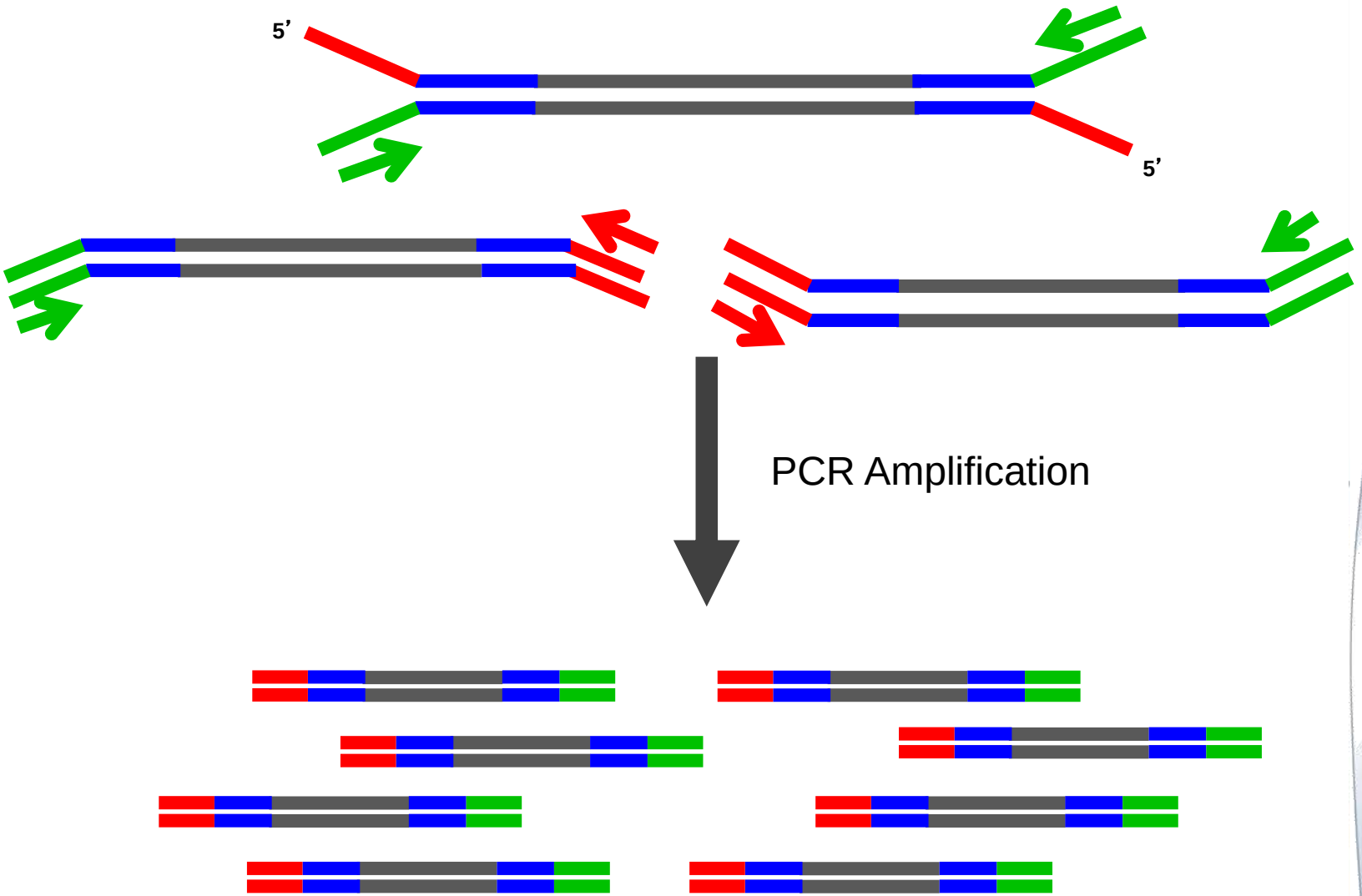
Single Overhang Fragments



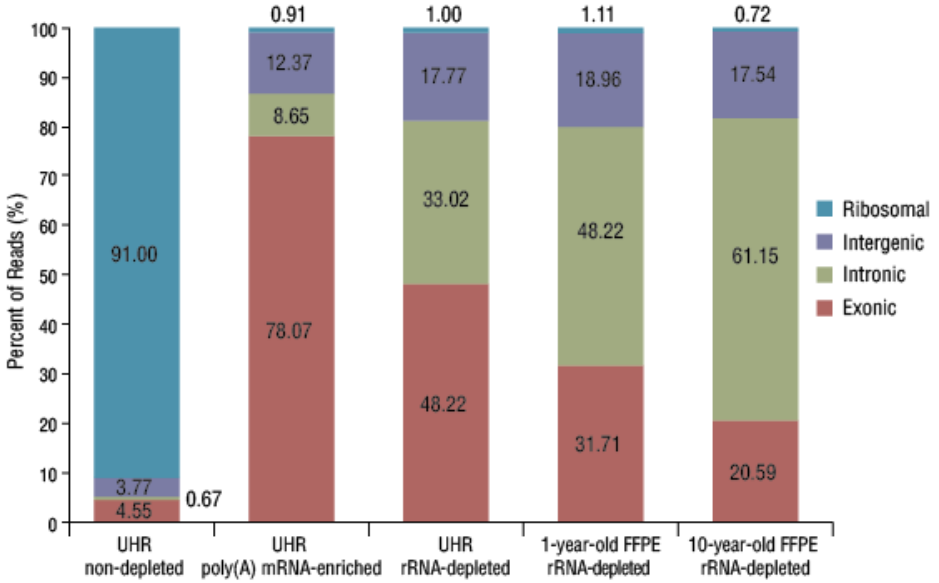
Adapter Ligation

DNA Fragments with Adapter Ends

Enrichment of library fragments



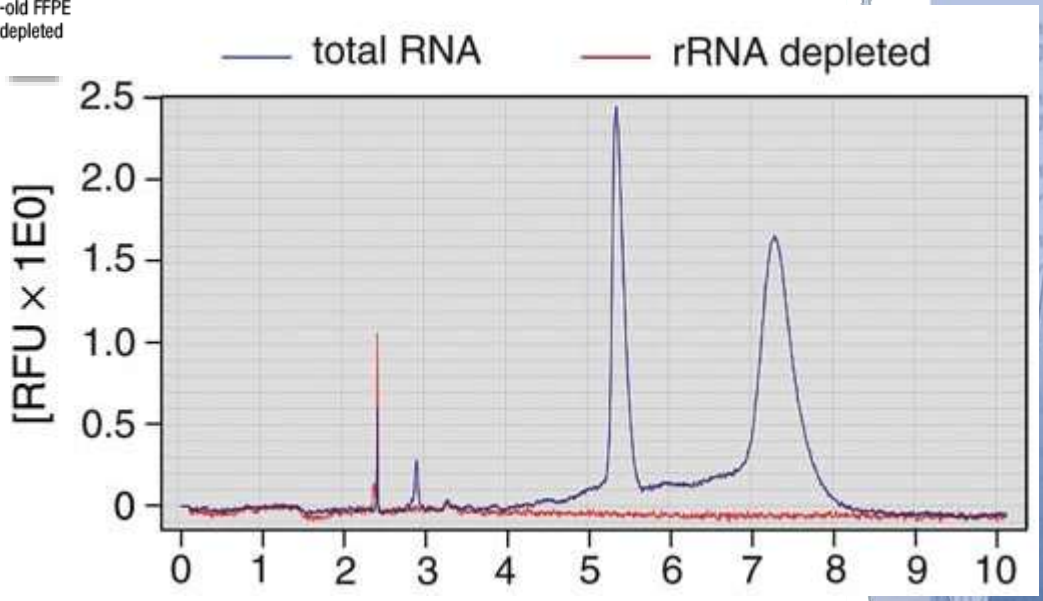
mRNA makes up only about 2% of a total RNA sample



- more than 90% rRNA content

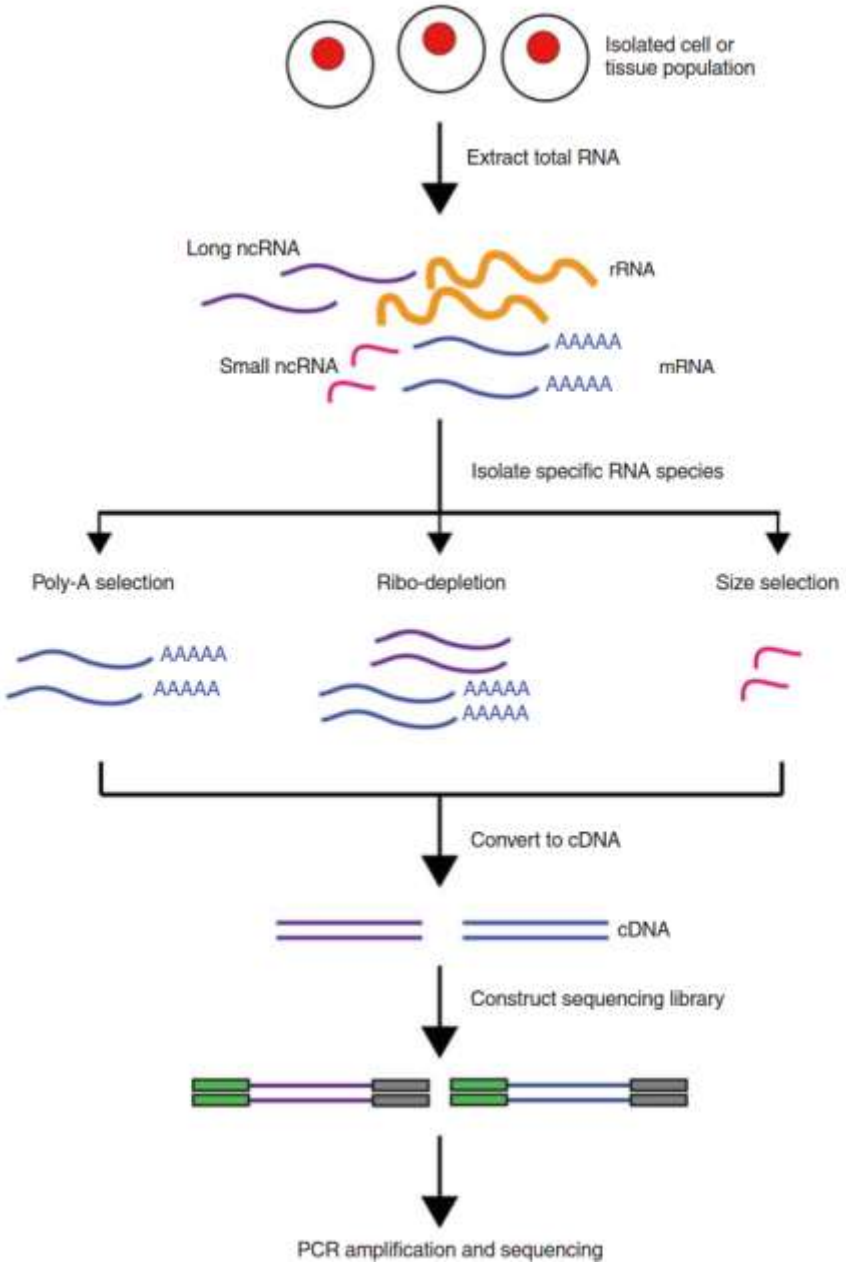
- multiple other non-coding RNA species

Bioanalyzer trace before and after ribo-depletion



RNA-Seq library prep procedure

1. RNA-sample QC, quantification, and normalization
2. Removal of ribosomal RNA sequences:
   via positive or negative selection: Poly-A enrichment or ribo-depletion
3. Fragment RNA:
   heating in Mg++ containing buffer – chemical fragmentation has little bias
4. First-strand synthesis:
   random hexamer primed reverse transcription
5. RNAse-H digestion:
   - creates nicks in RNA strand; the nicks prime 2nd-strand synthesis
   - dUTP incorporated into 2nd strand only
6. A-tailing and adapter ligation exactly as for DNA-Seq libraries
7. PCR amplification of only the first strand to achieve strand-specific libraries - archeal polymerases will not use dUTP containing DNA as template

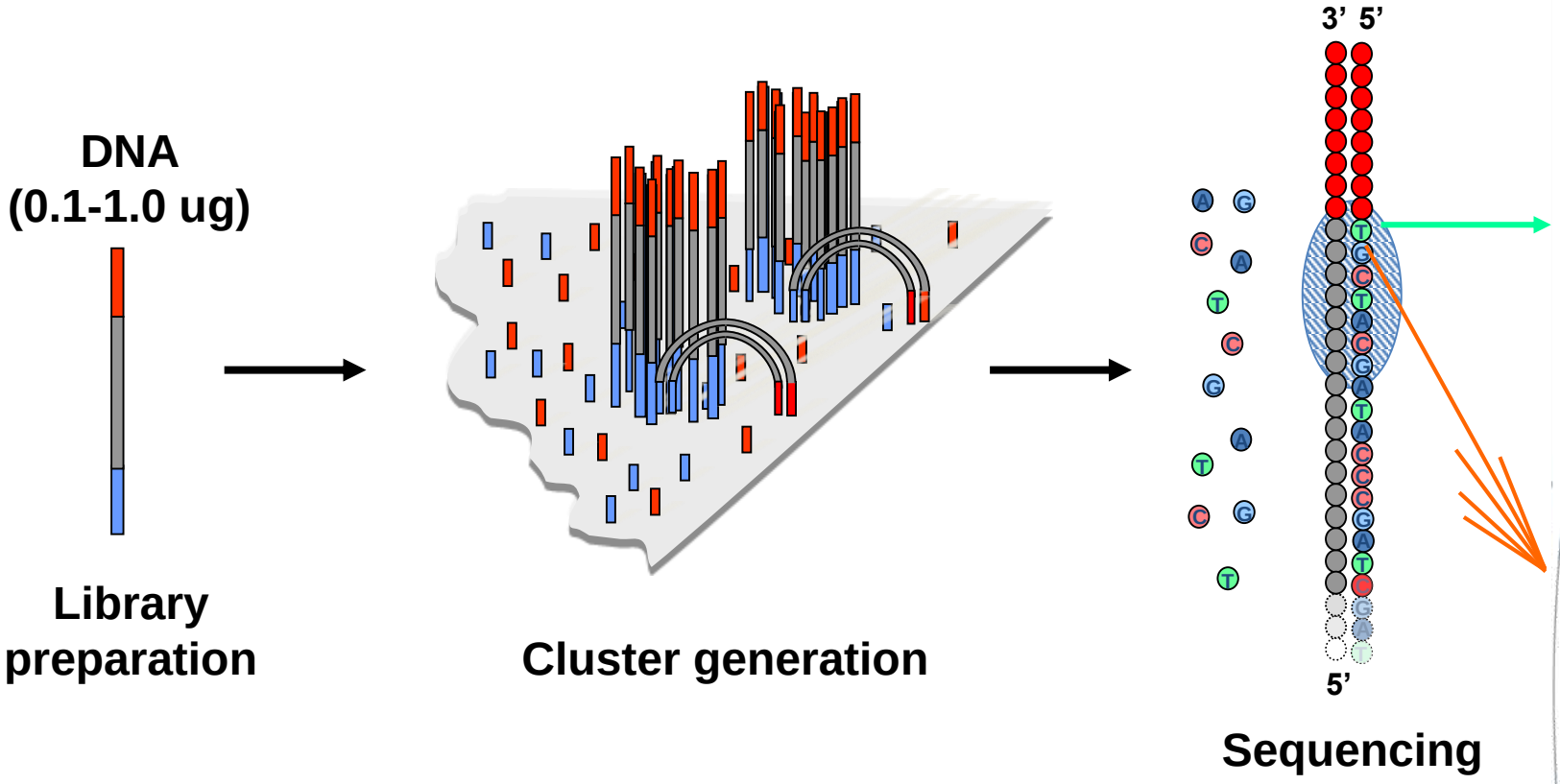


RNA-seq?

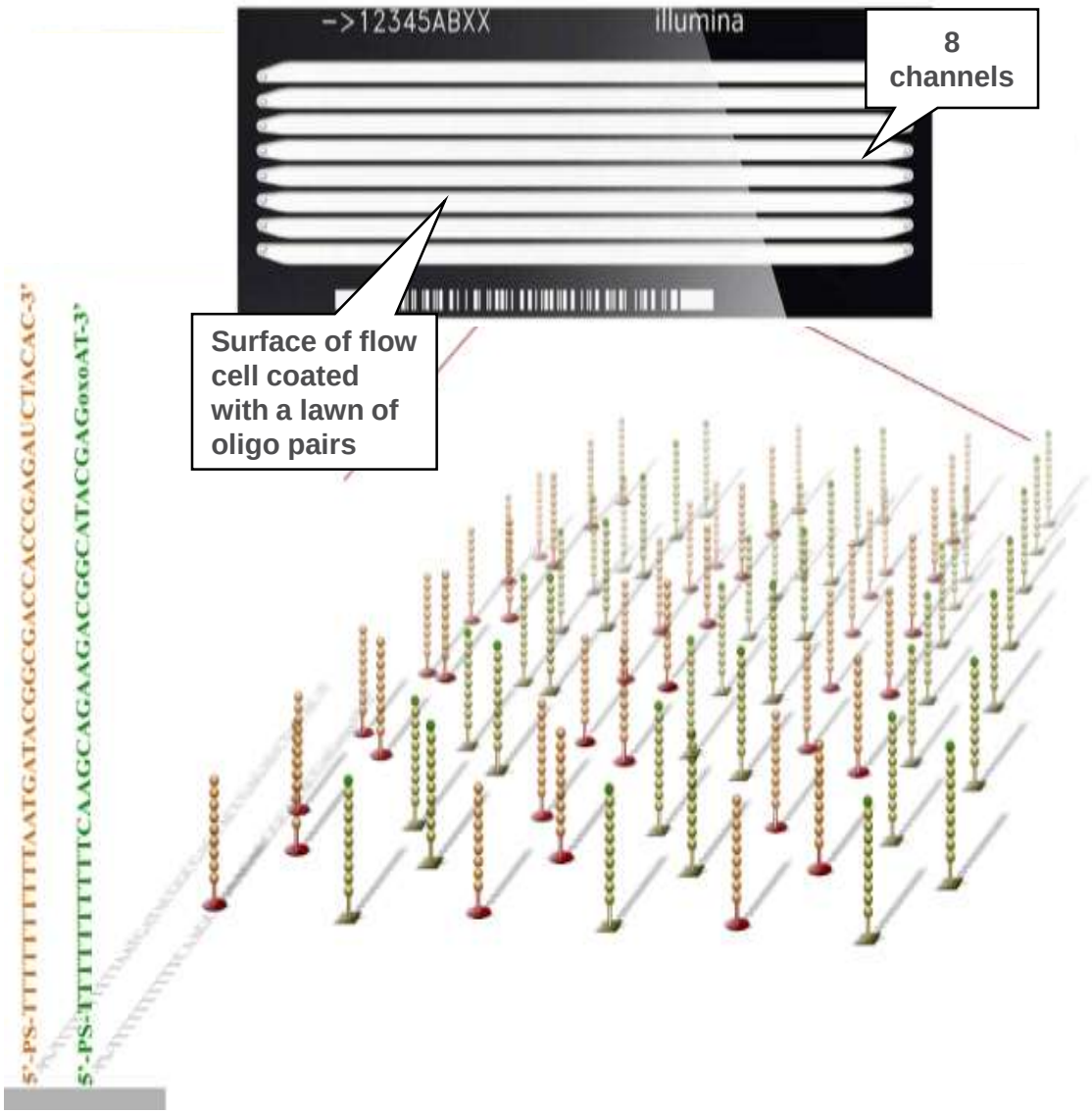
Sorry - we are only sequencing DNA.

Illumina Sequencing Technology

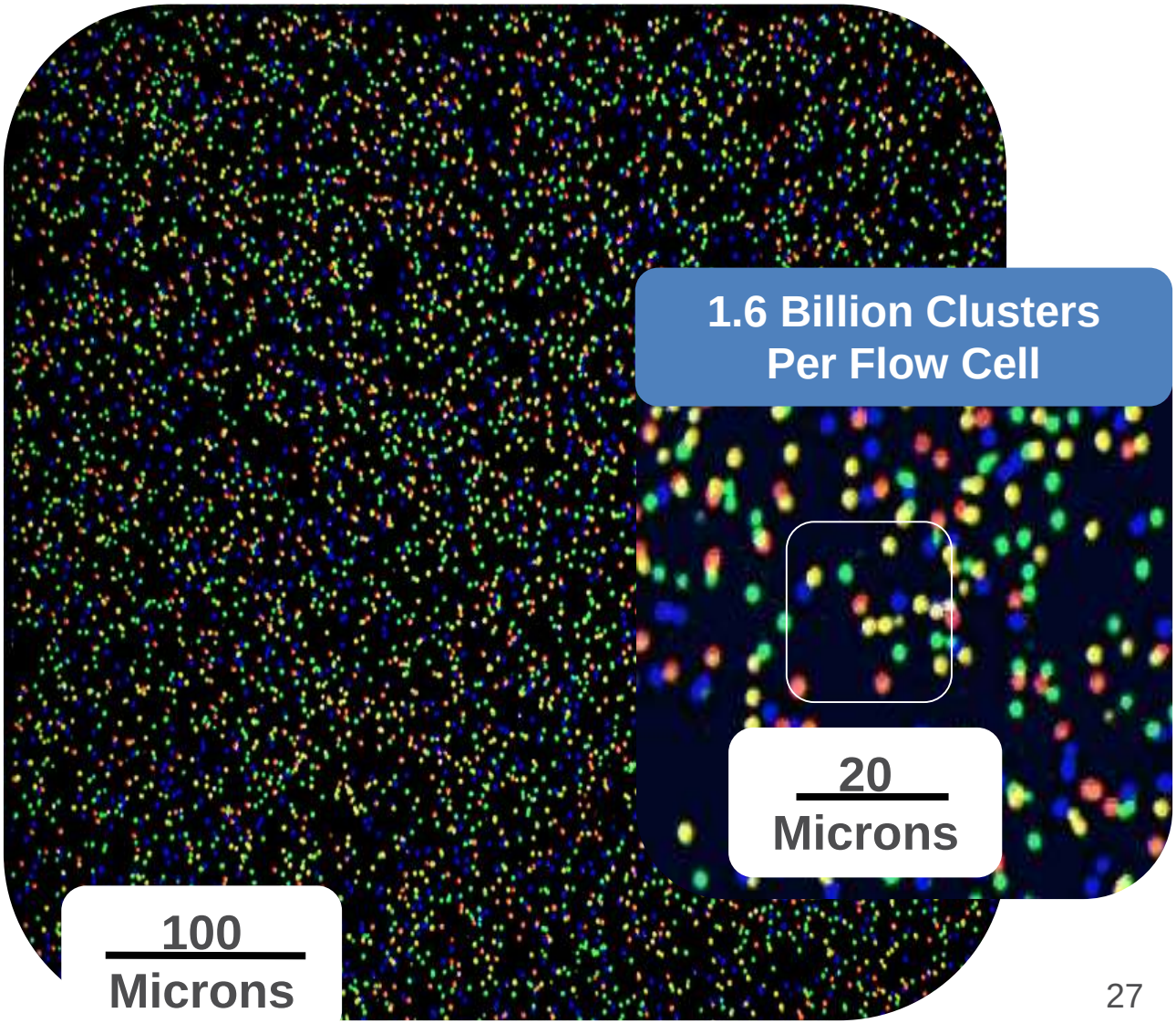
Sequencing By Synthesis (SBS) Technology



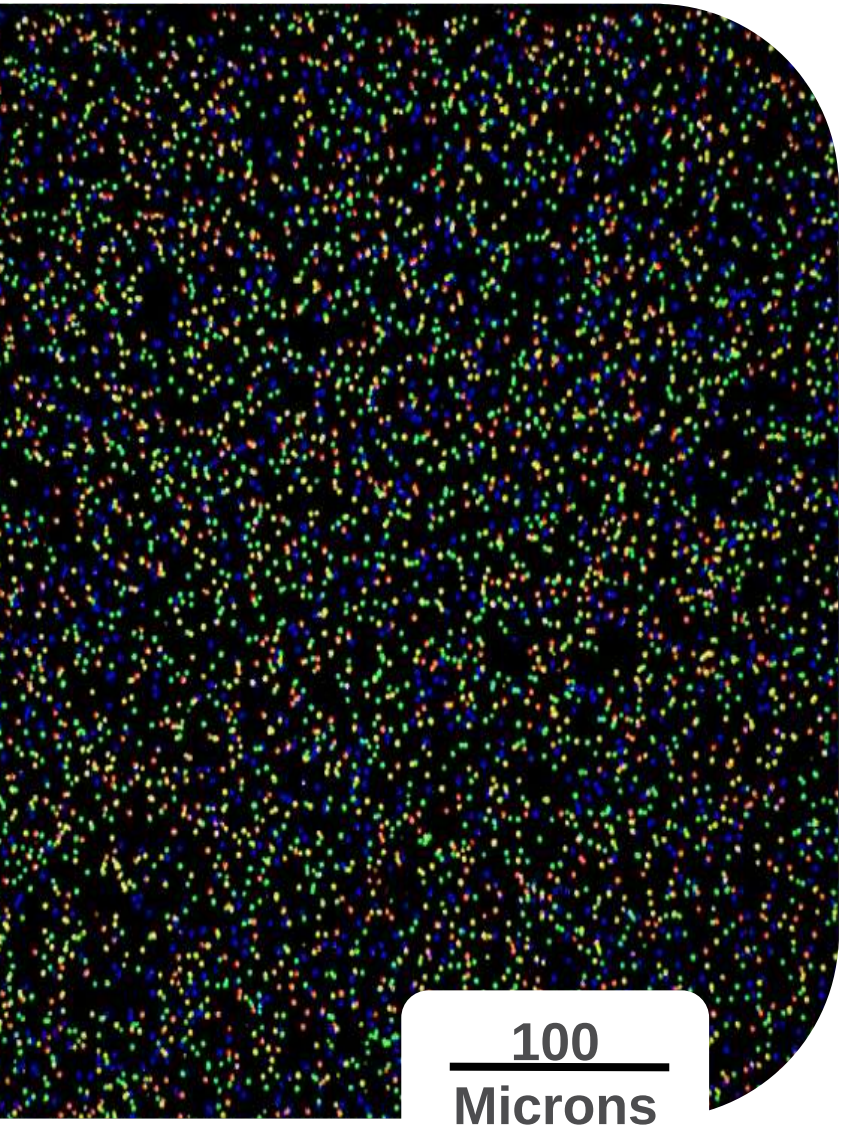
TruSeq Chemistry: Flow Cell

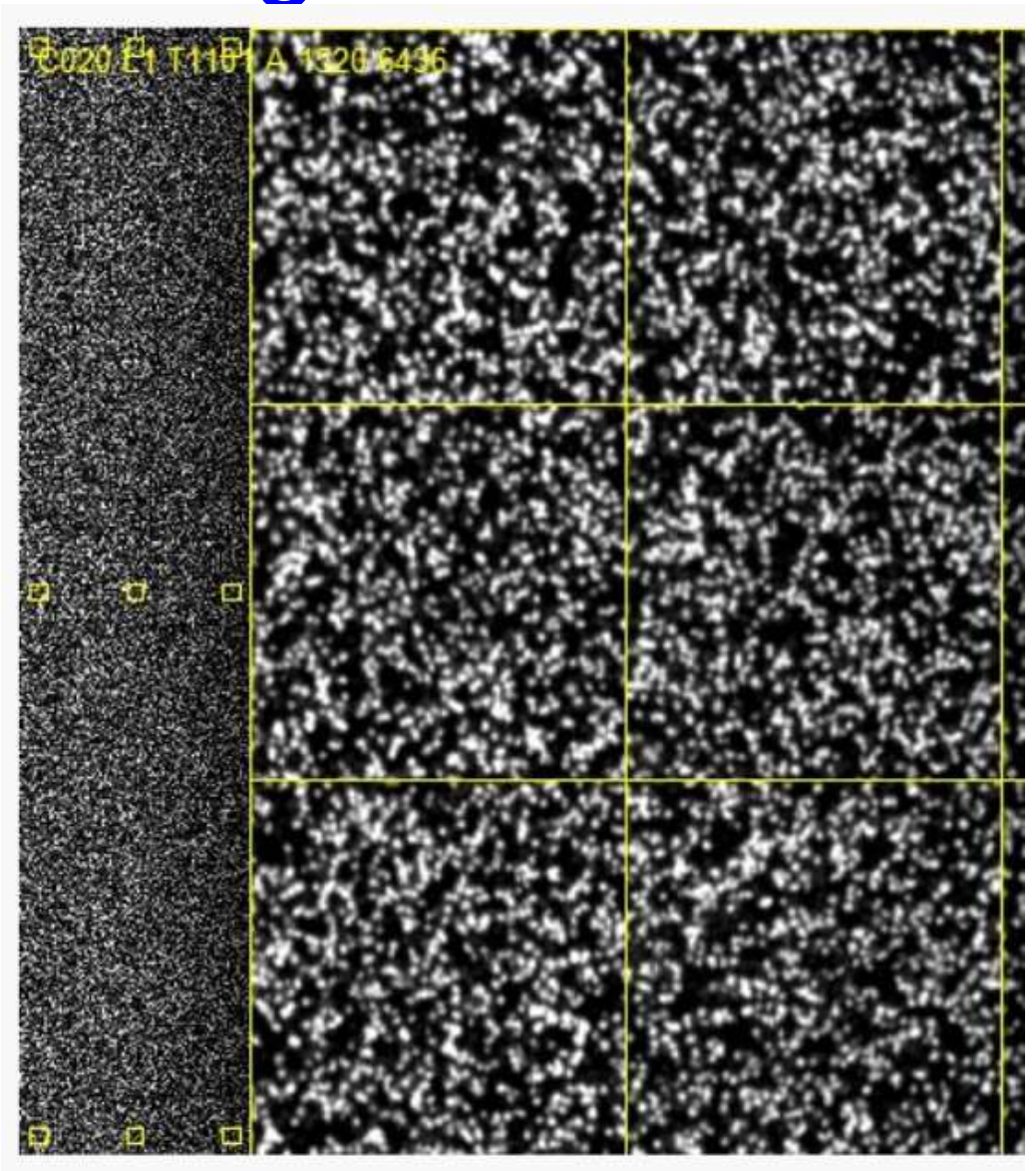


Sequencing

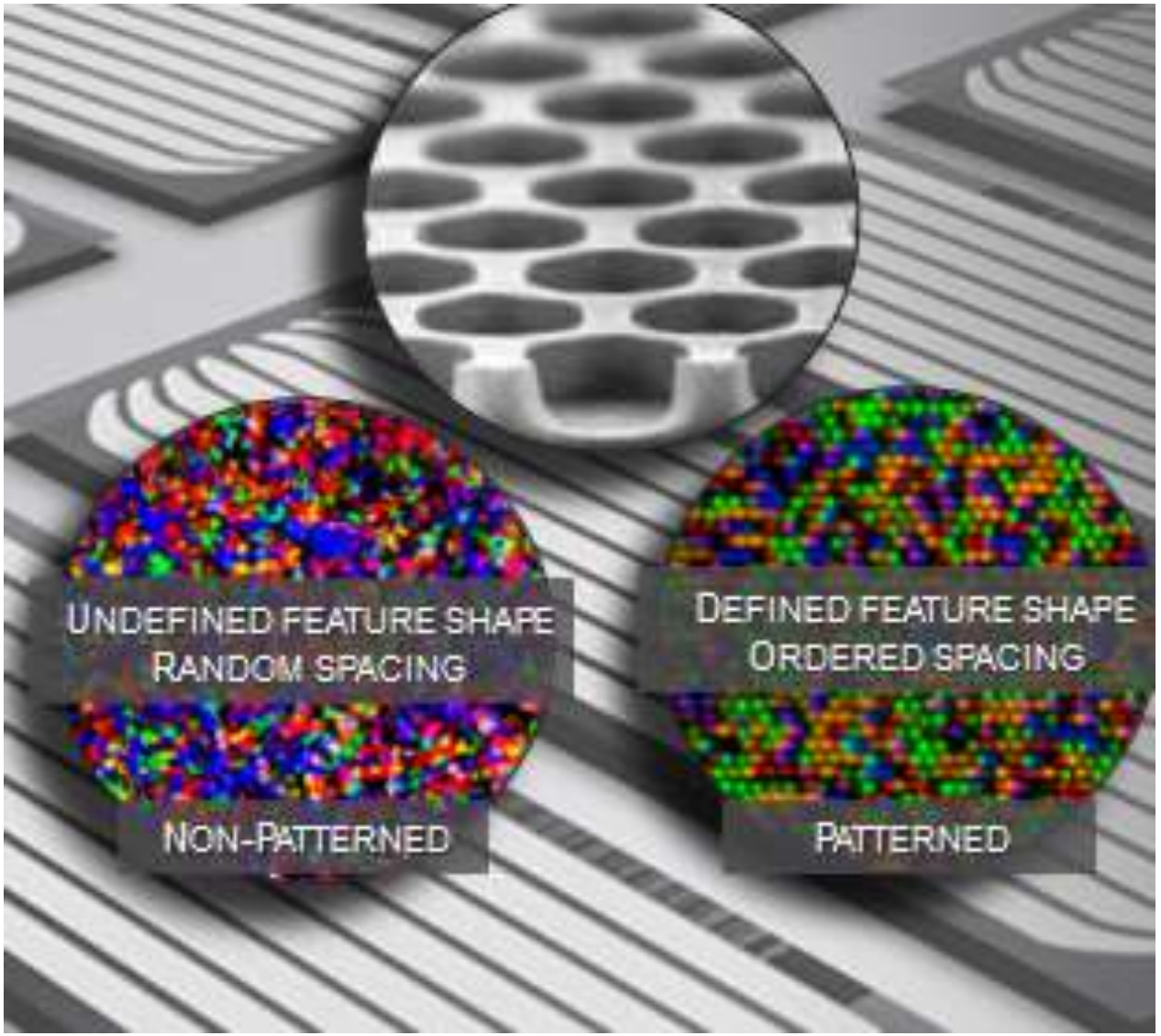


Sequencing

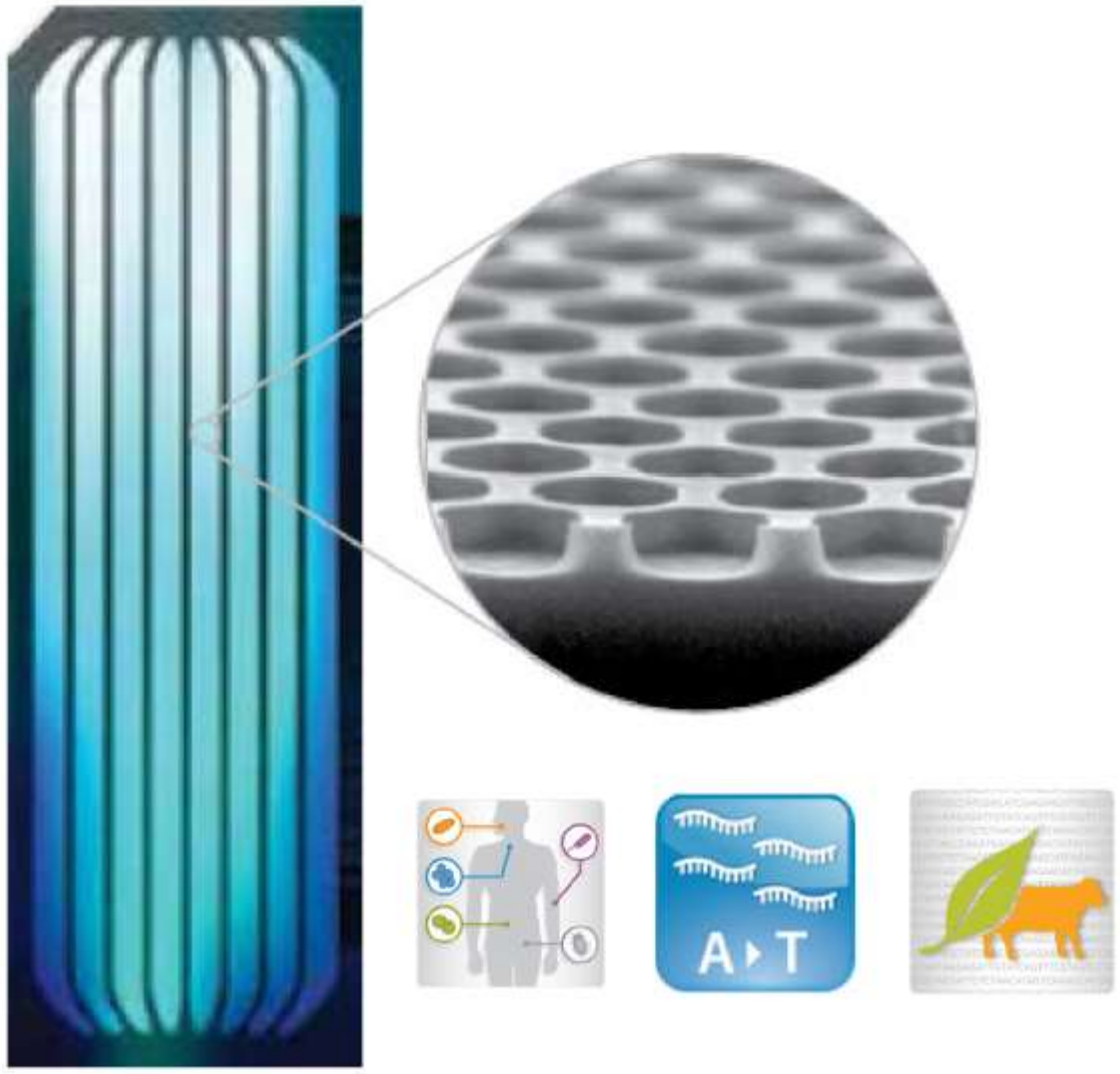




Patterned Flowcell

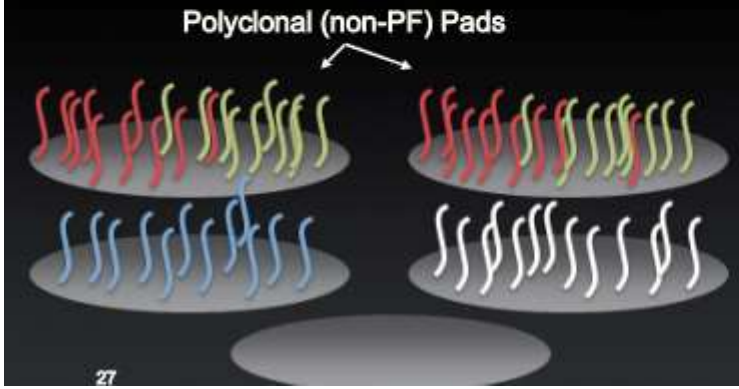


Hiseq 3000: 478 million nanowells per lane

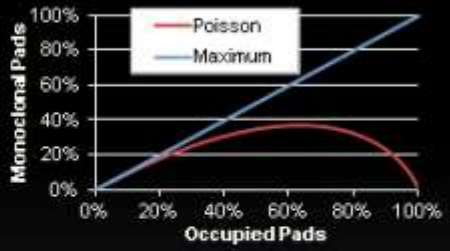


CONCEPTUAL CHALLENGE— BEATING POISSON

Amplification Phase



Maximizing Well Occupancy and Monoclonality

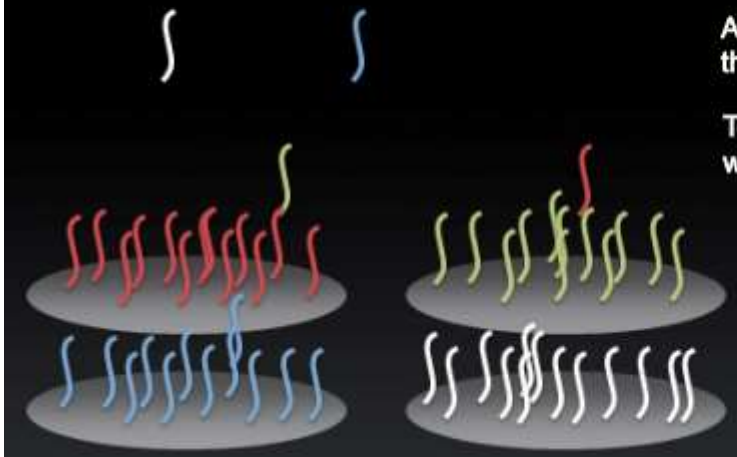


Poisson statistics limit max monoclonal occupancy < 40%

Polyclonality rises as occupancy increases

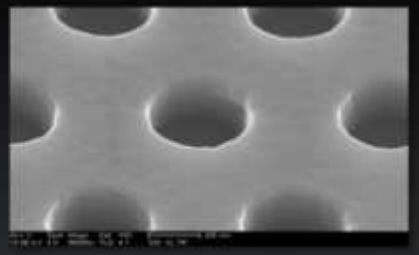
SIMULTANEOUS SEEDING AND AMPLIFICATION

Maximizing Well Occupancy and Monoclonality



Amplification occurs at rate >> faster than seeding rate

Templates excluded from occupied wells



What will go wrong ?

- cluster identification
- bubbles
- synthesis errors:

ClusterCluster
ClustsrCluster
ClusterCluster
ClusterCluster
CllsterCluster

What will go wrong ?

- synthesis errors:

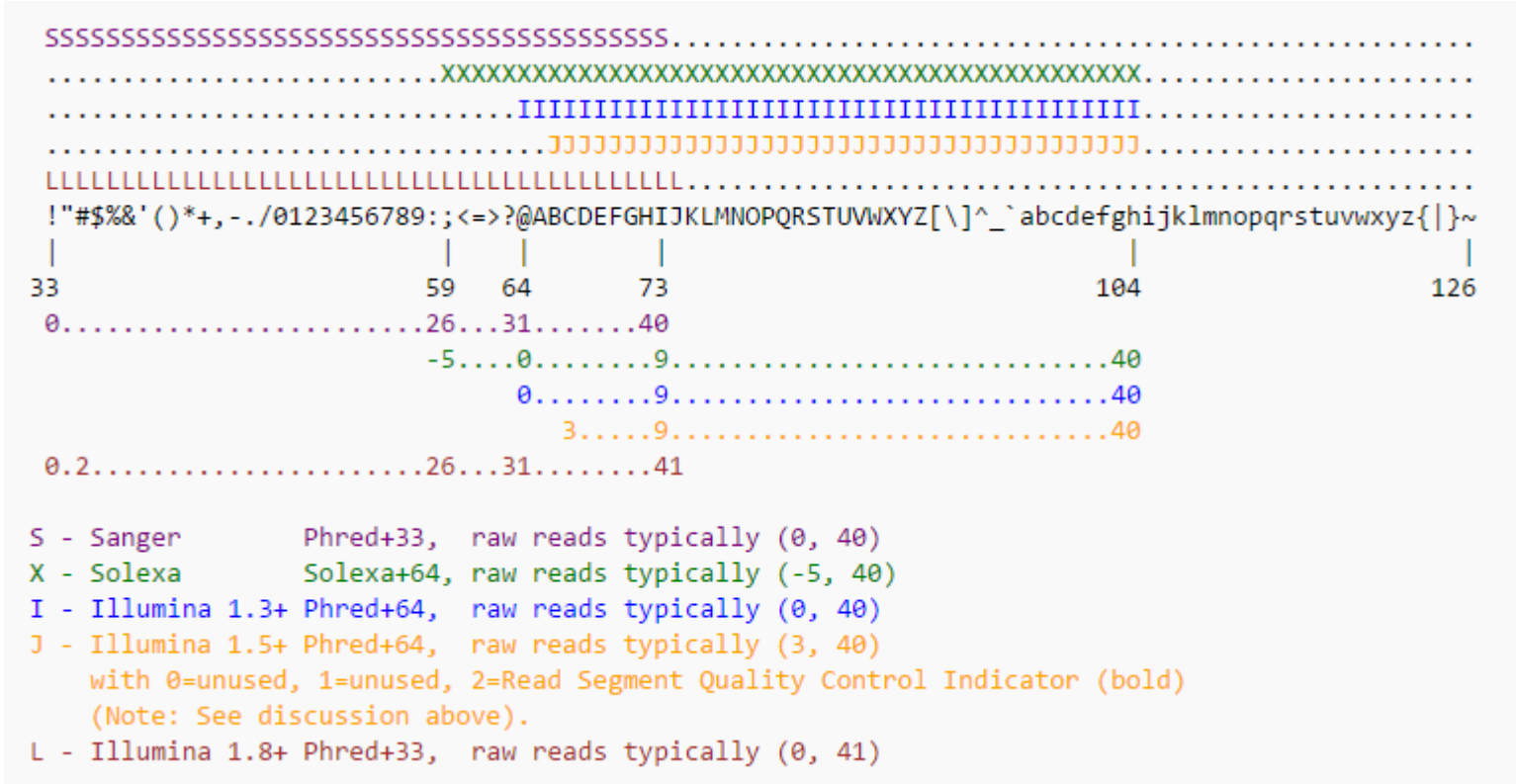
ClusterCluster	ClsterClusterC
ClustsrCluster	ClusterCluster
ClusterCluster	ClusterCluster
ClusterCluster	ClusterCluste
ClsterCluster	ClusterCluster

Phasing & Pre-Phasing problems

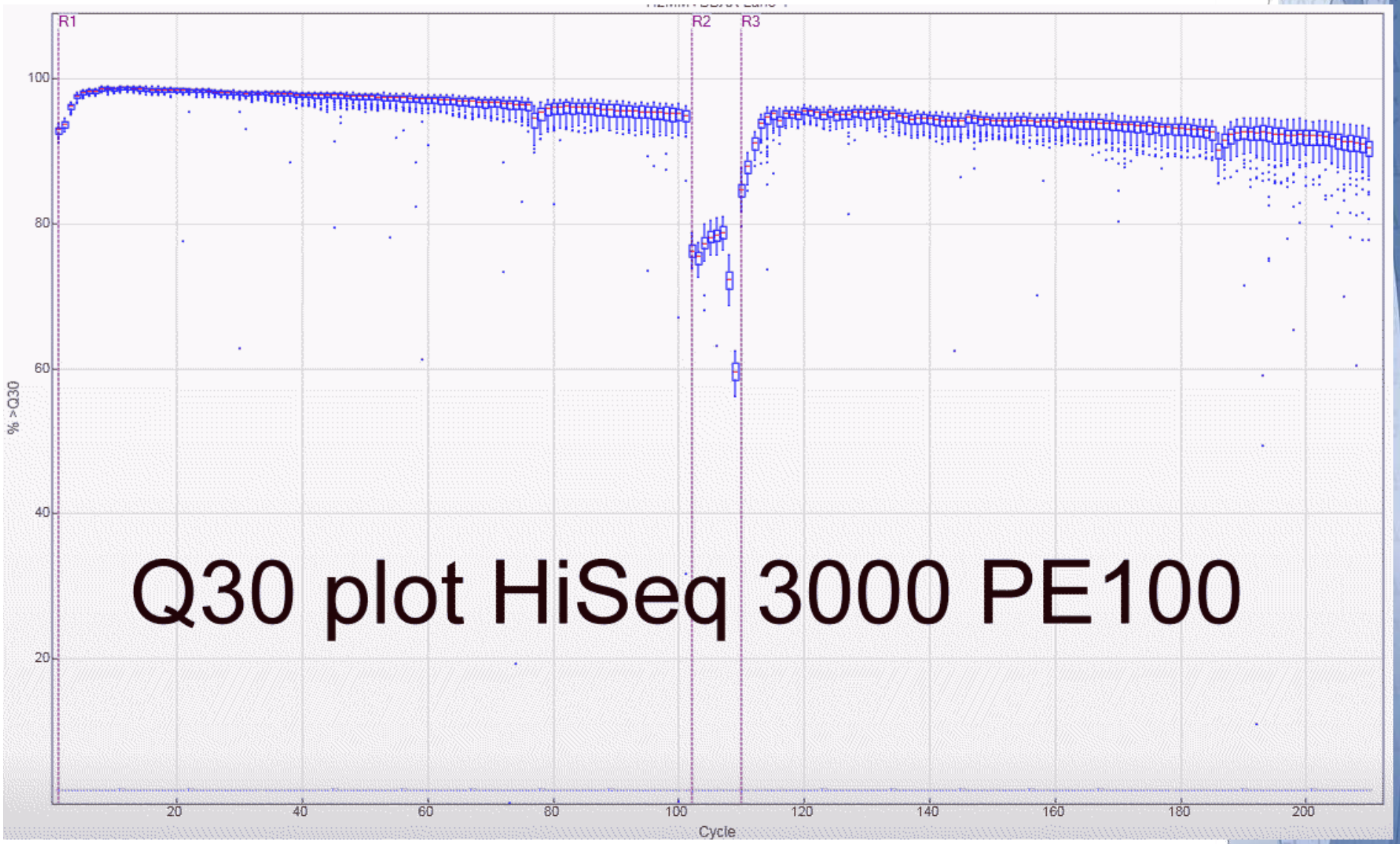
The first lines of your data

```
@700593F:586:HTWJJBCXX:1:1107:2237:10031 2:N:0:GGAGAACA
NNNNNNNNNNNNNNNNNNNNNAGGCCAGCCATAGAACGCTCCCGGCTTCACGGACGT
CATATAGTCAGGCACGAGGTCGGCGCCGAGTTCGTCACGCTCGTTGACGACCGCCCAT
ACCGCTTGATTTGCGGGGTTGATCGCTAGCGCGGTCGGATTGCGAATGCCCGAGGCAT
ACGTCCGATGGGCCCCGCTGGCGCGGTCCACTTCCCATACAACCGCGCGTTCCTCTTC
CAGCGCCATGCCGCGTT
+
@@@@@@@@@@@@@@@@@@@@@@@IHHFIIIIIHIIIIIIIIEHIIIIIHHFHIIIIII
IIGHHHHIHHIIIIEHIIHIHIHGIIHHDHIIIIIHHHHEEHIHIHCG/EH@GHHII
GIHHIIHHIIIICHGHHIHIIHHIIIIIHIHIIHIGIIIGIEH?H?HHHIGHIIIGHI
GHIDHH@AHHIHIH=HI=GHGHHGD
```

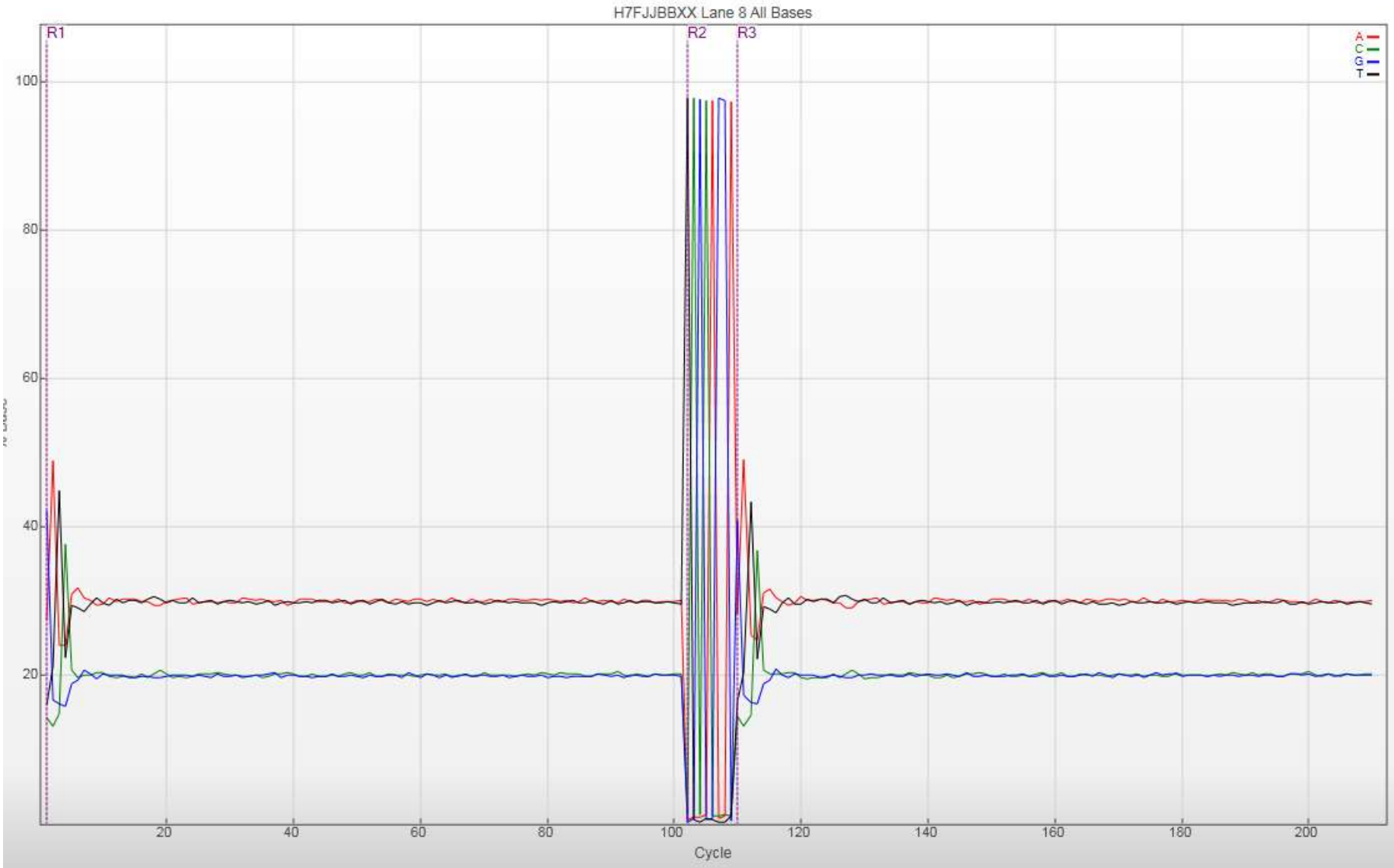
```
@M02034:265:000000000-AN3L2:1:2102:8707:16197 2:N:0:85
GATGAACATAATAAGCAATGACGGCAGCAATAAACTCAACAGGAGCAGGA
+
AAAAAFFFFFFFGFFGFFBE5GEAAAEDCFDFAEG5CFGHFGGFEGHHHG
```



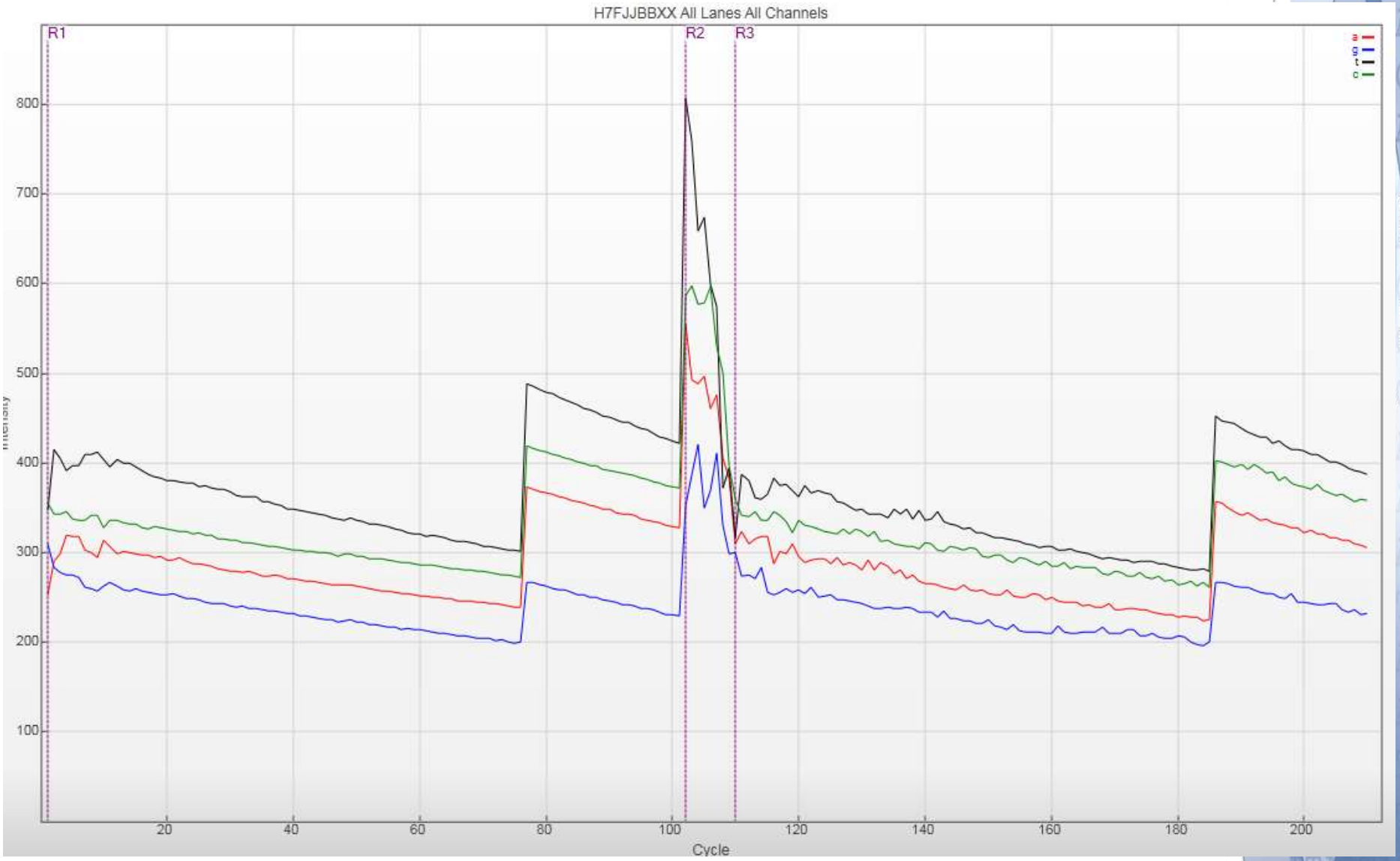
Illumina SAV viewer



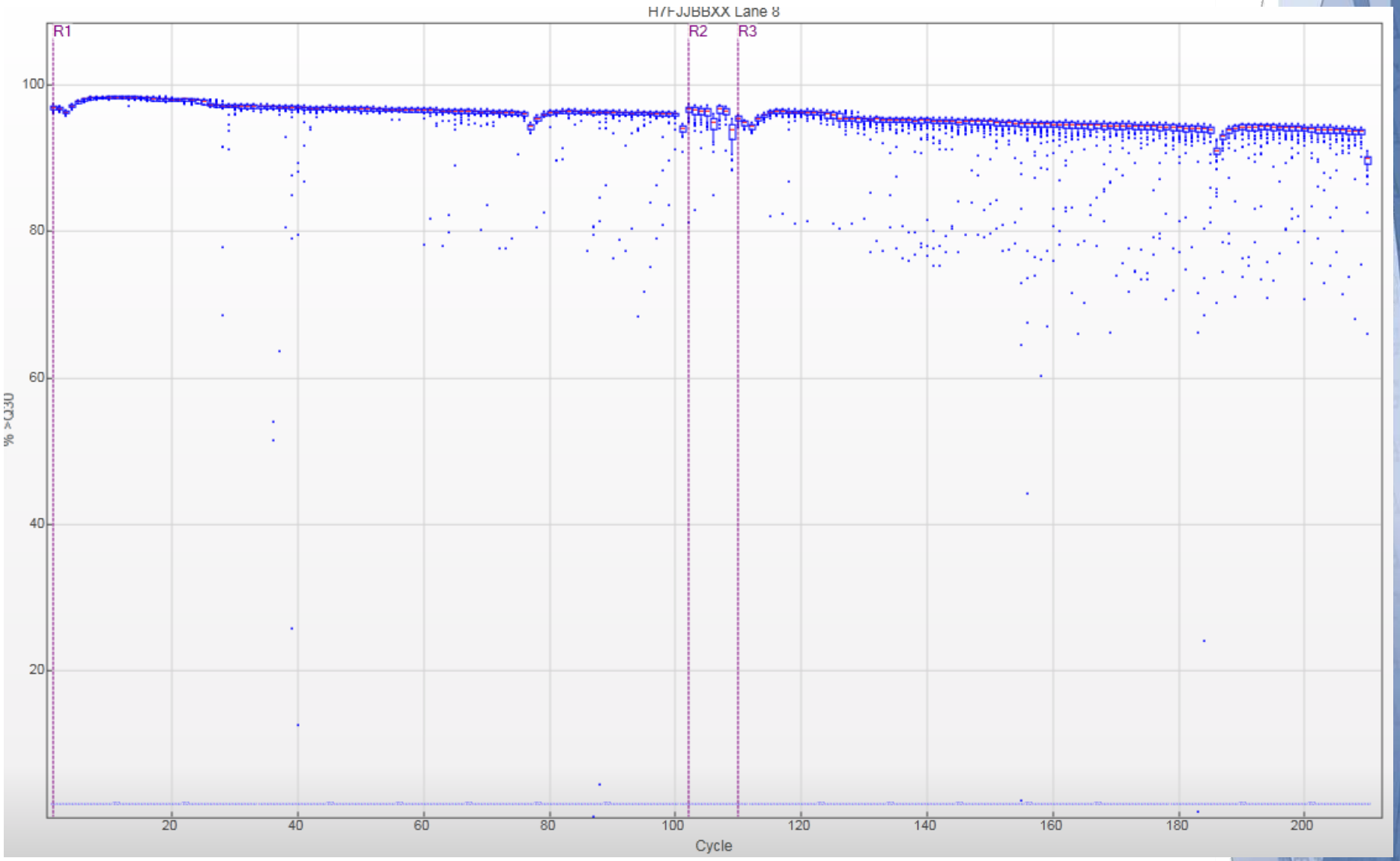
base composition



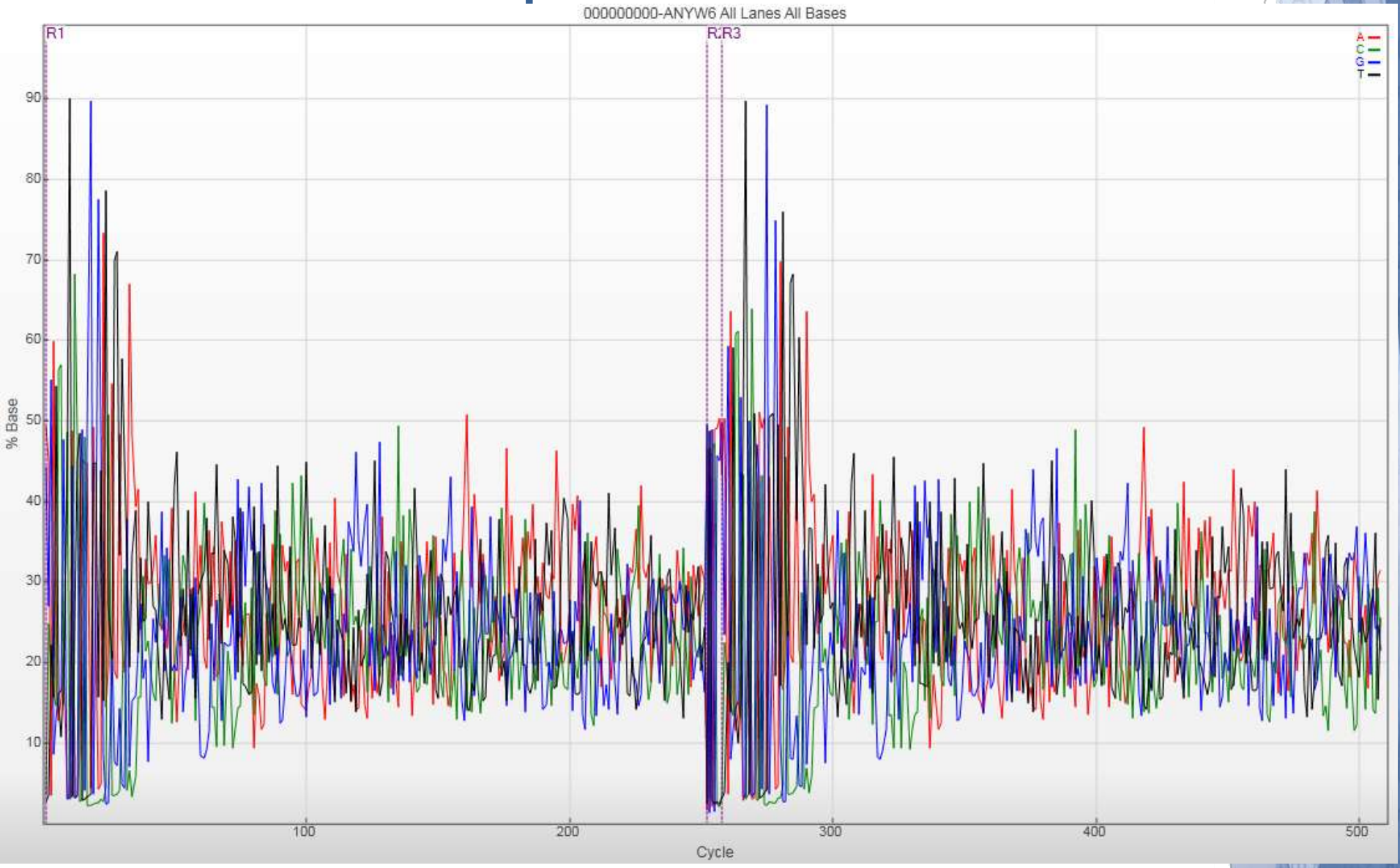
fluorescence intensity



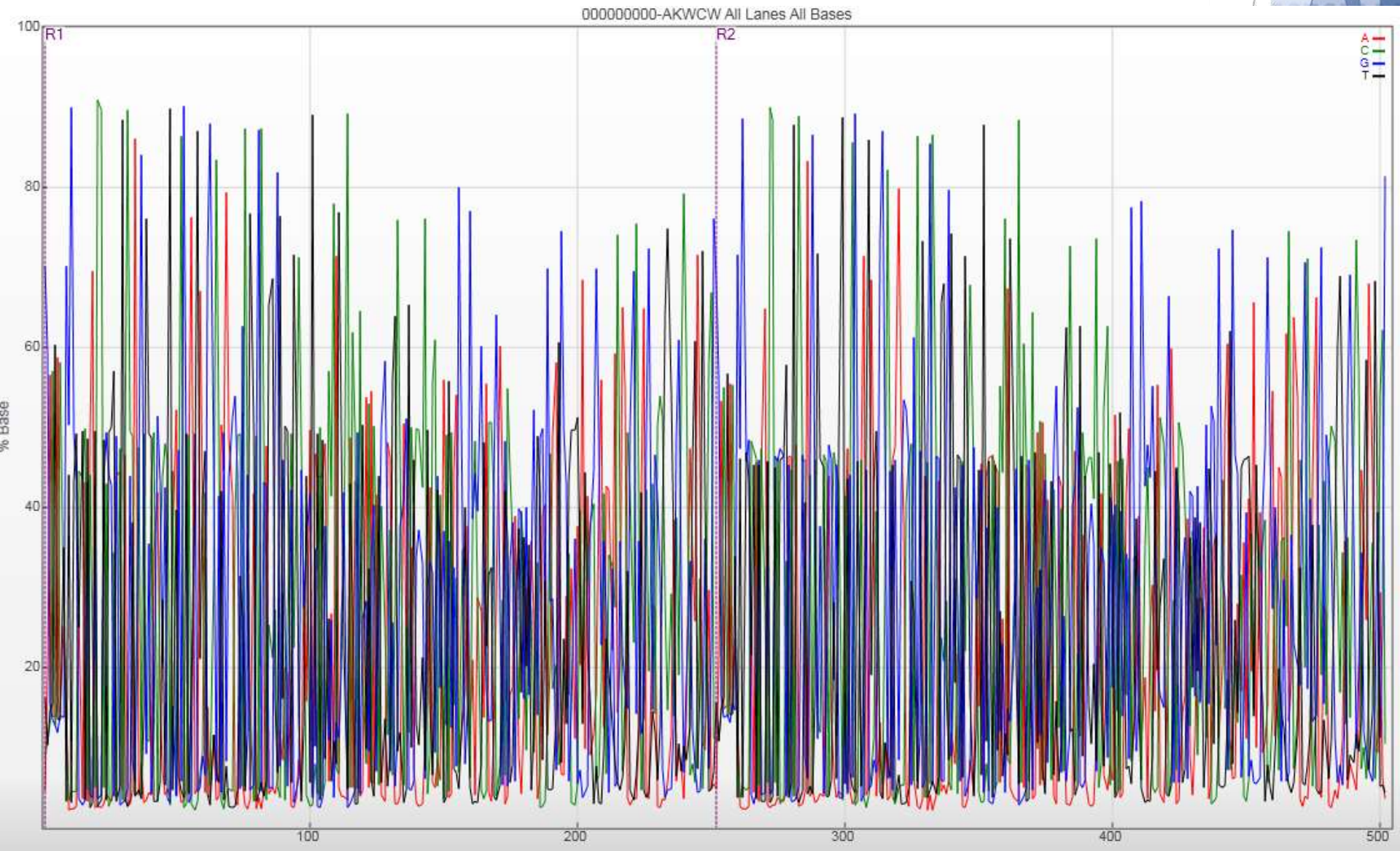
fluorescence intensity



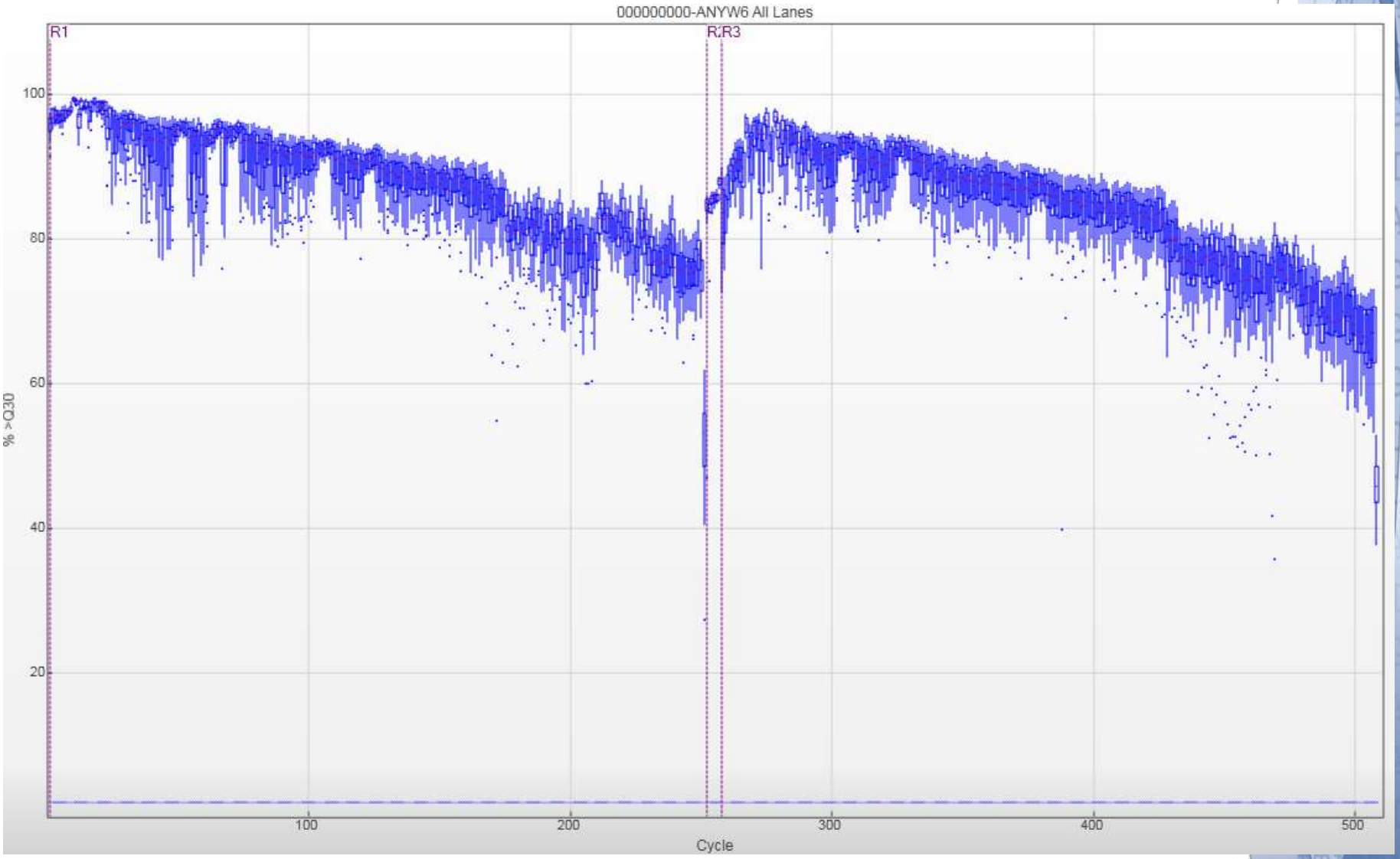
amplicon mix



amplicon


000000000-AKWCW All Lanes All Bases
R1
R2
A
C
G
T
% Base
100
80
60
40
20
100
200
300
400
500

amplicon mix Q30



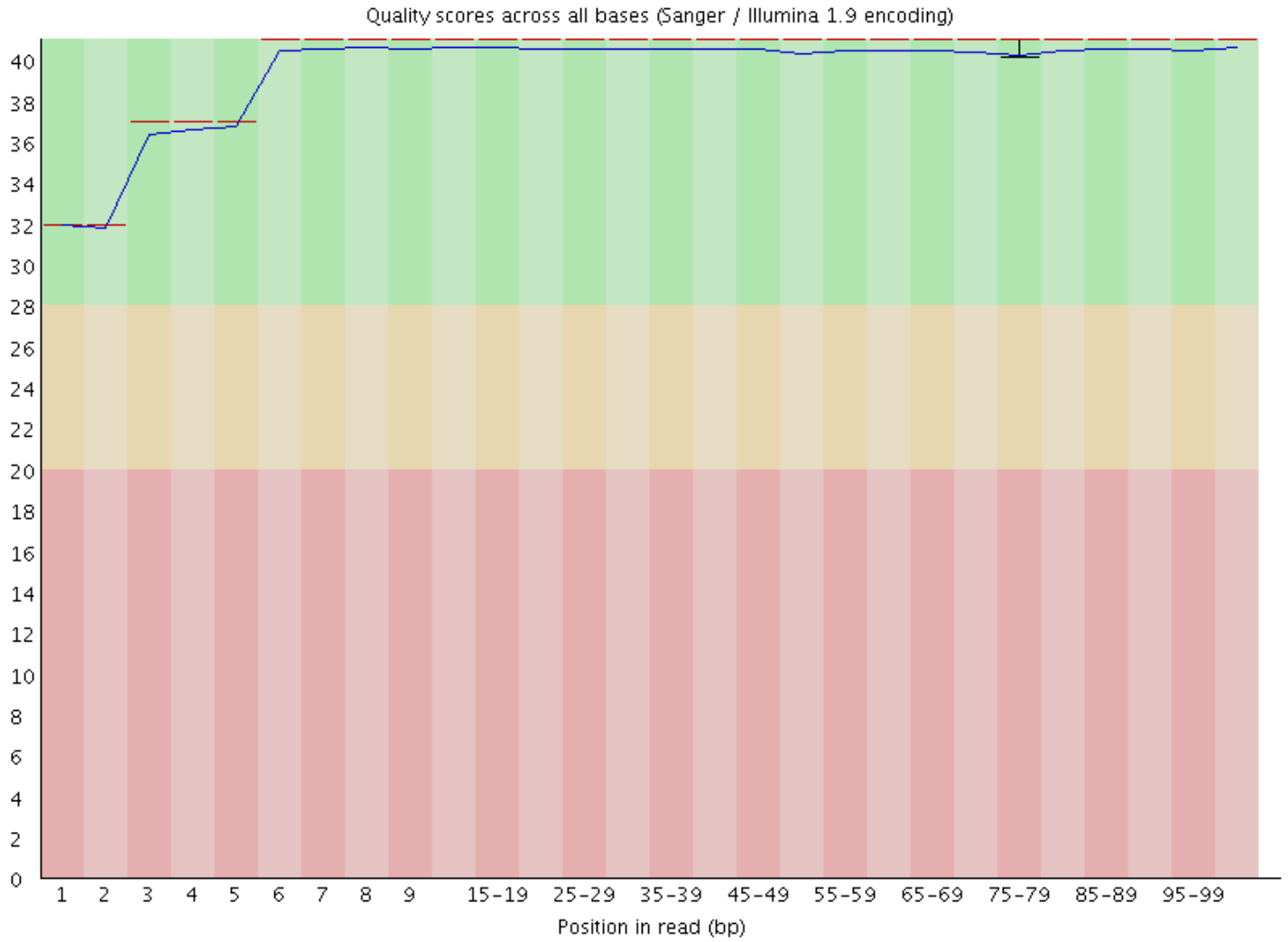
FASTQC

Basic Statistics

Measure	Value
Filename	3_S16_L008_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	16574908
Sequences flagged as poor quality	0
Sequence length	150
%GC	40

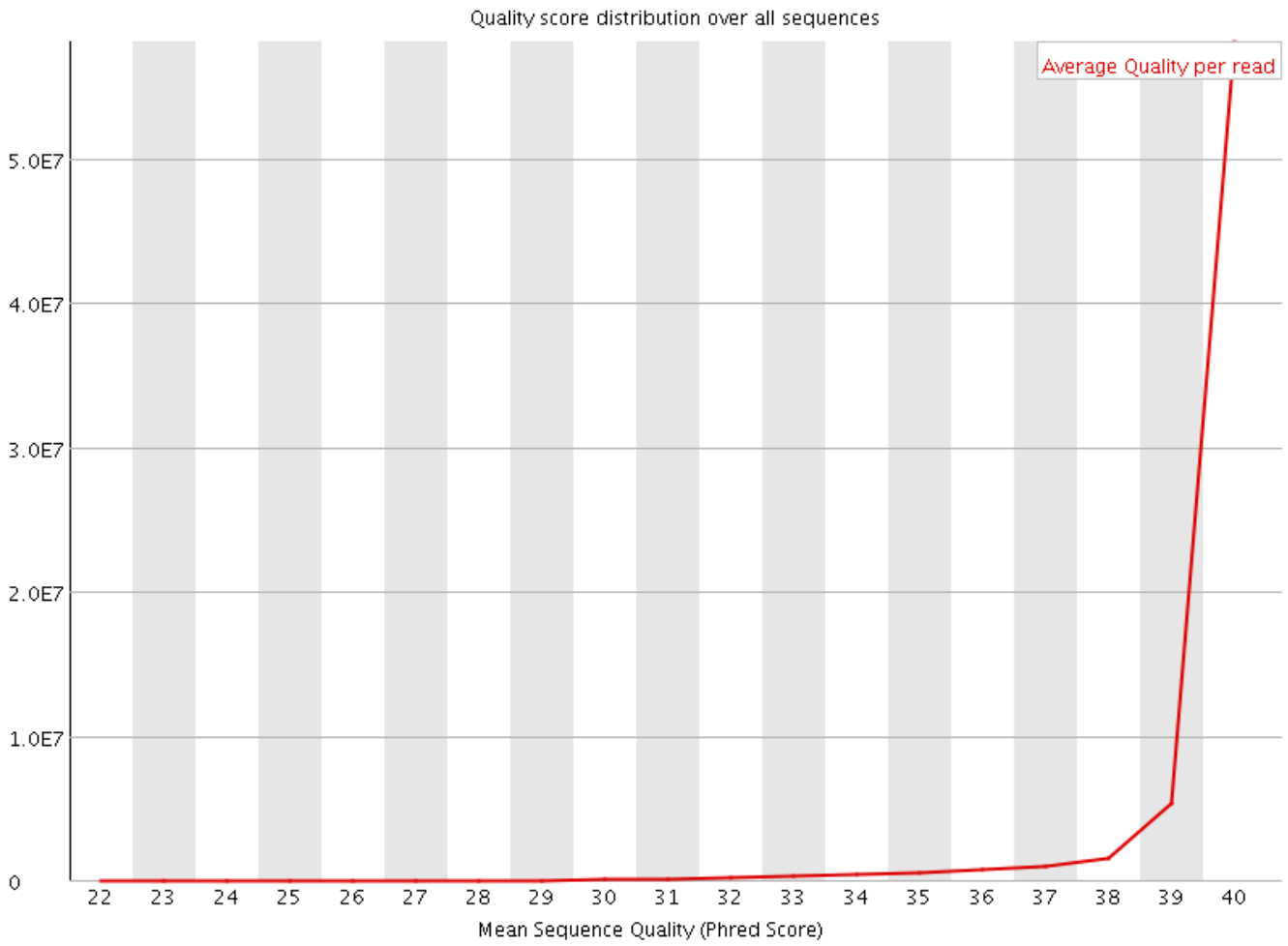
FASTQC

Per base sequence quality



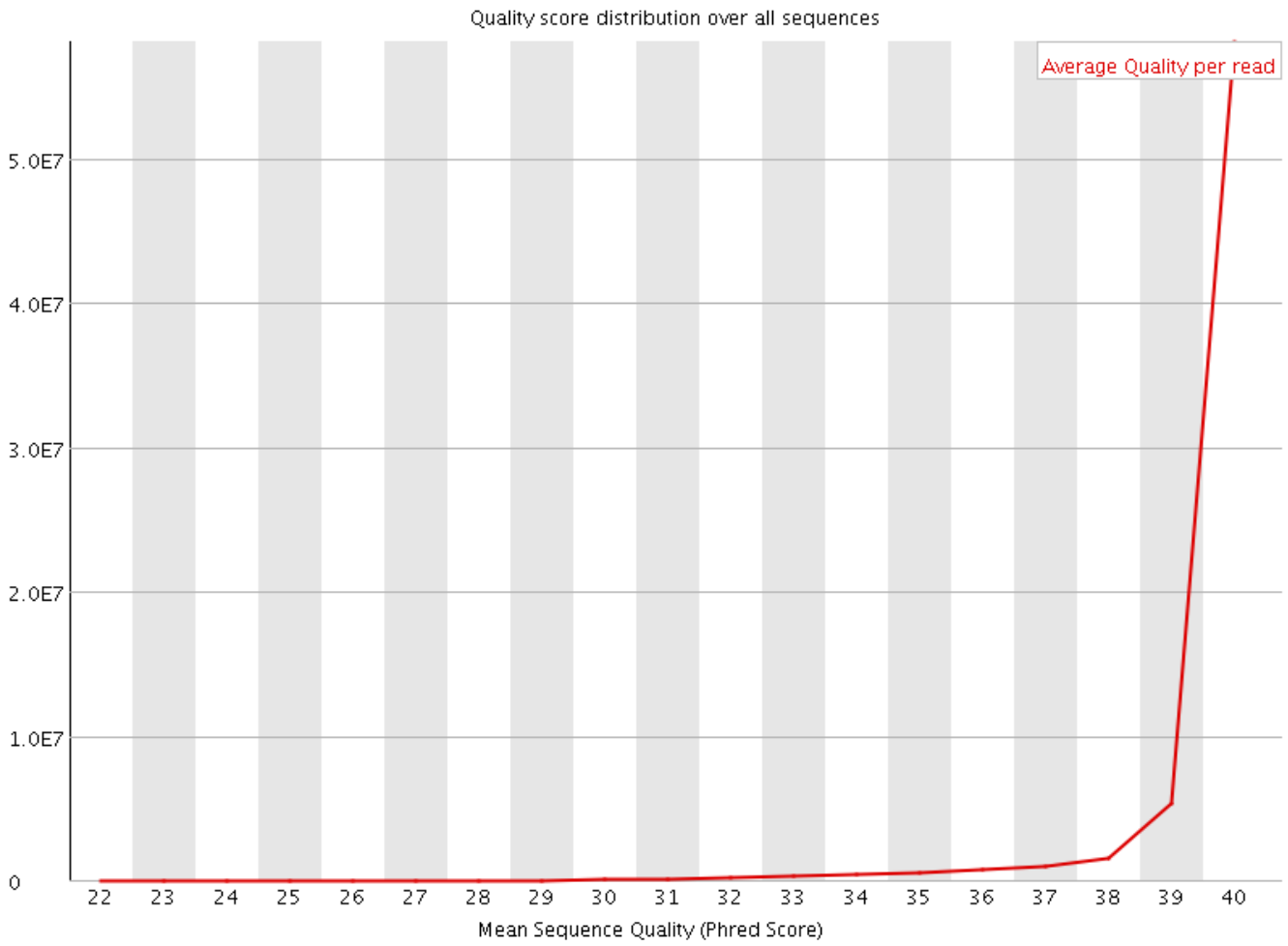
FASTQC

Per sequence quality scores



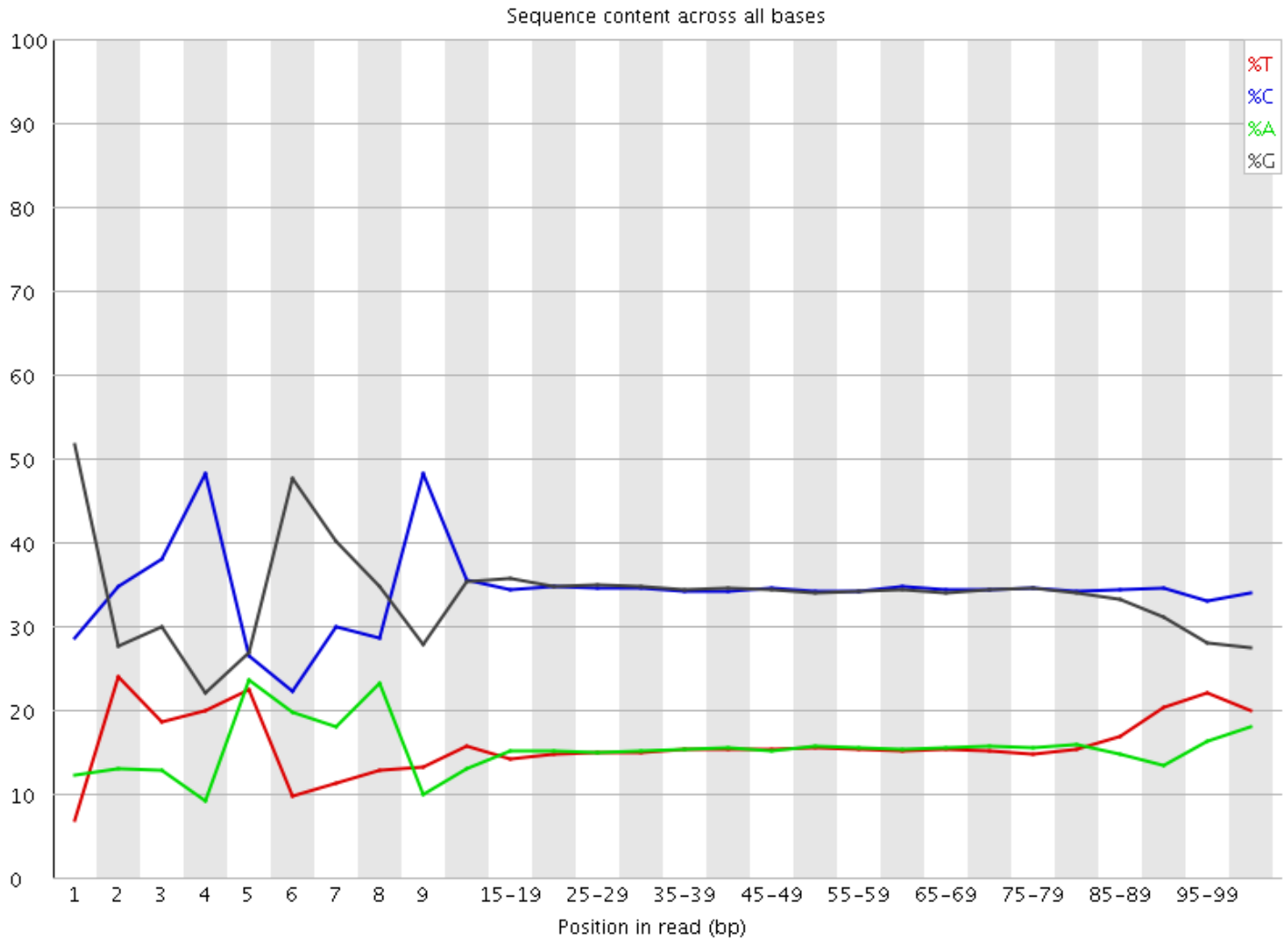
FASTQC

Per sequence quality scores



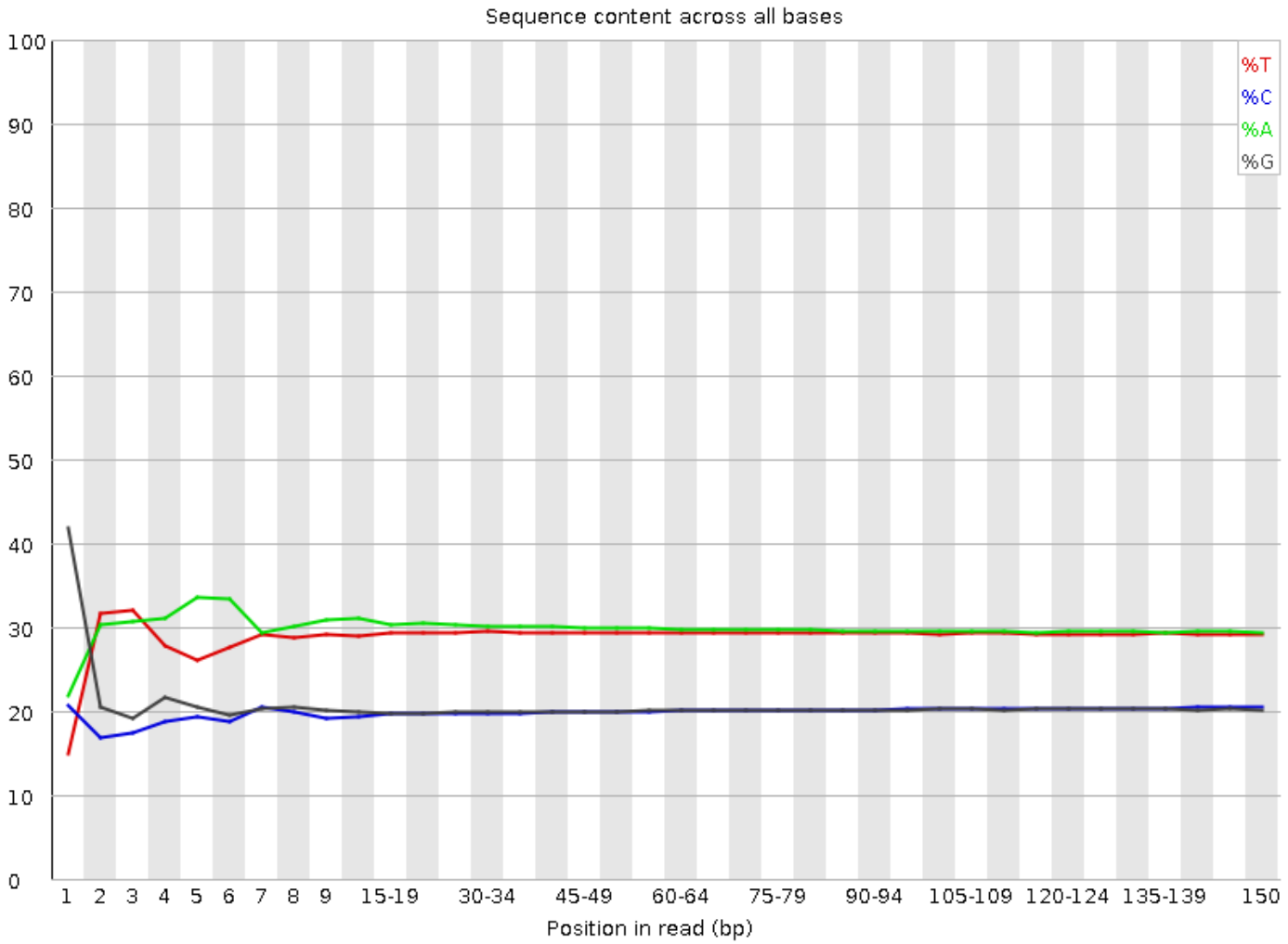
FASTQC - Nextera

Per base sequence content



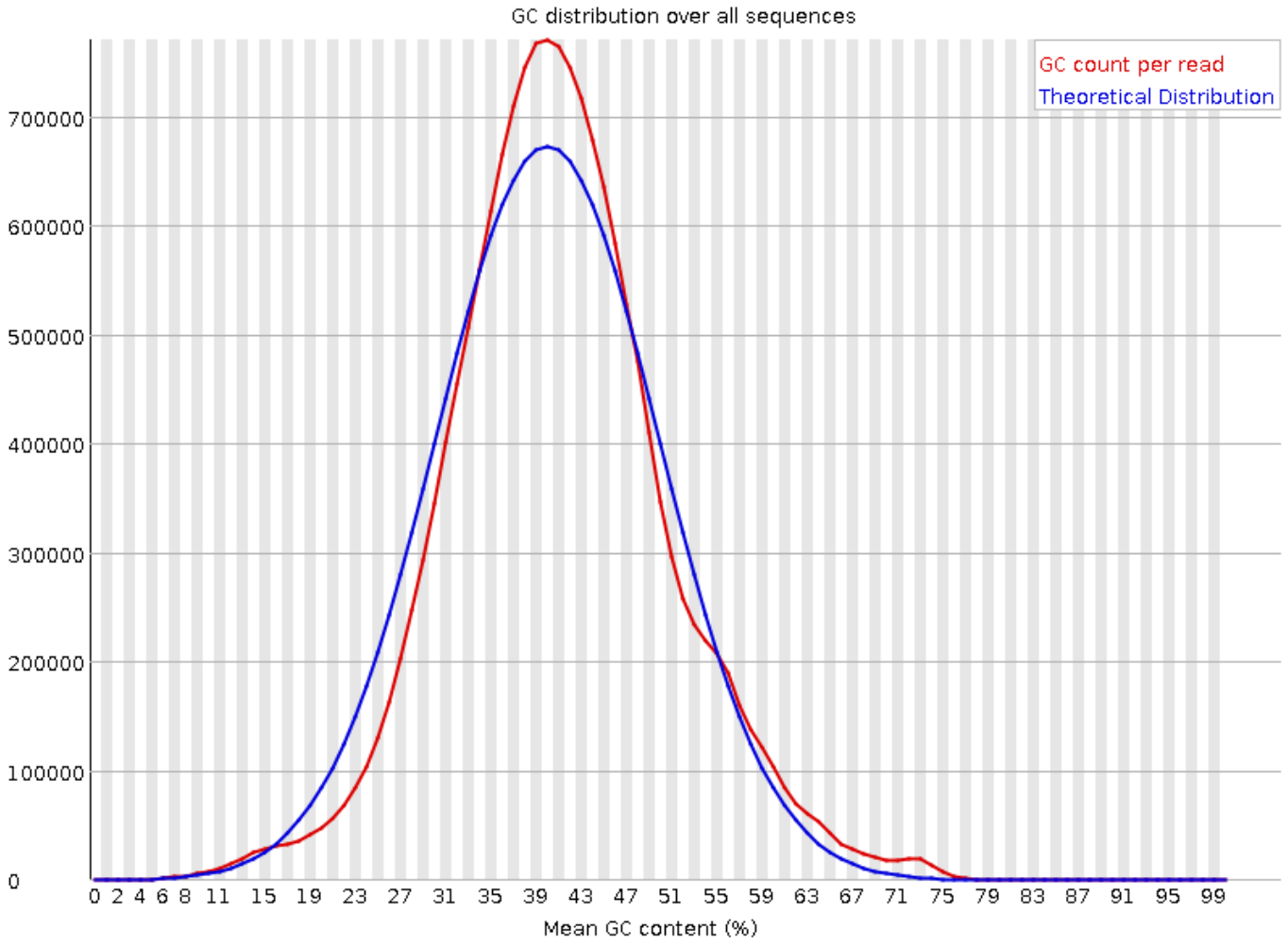
FASTQC

Per base sequence content



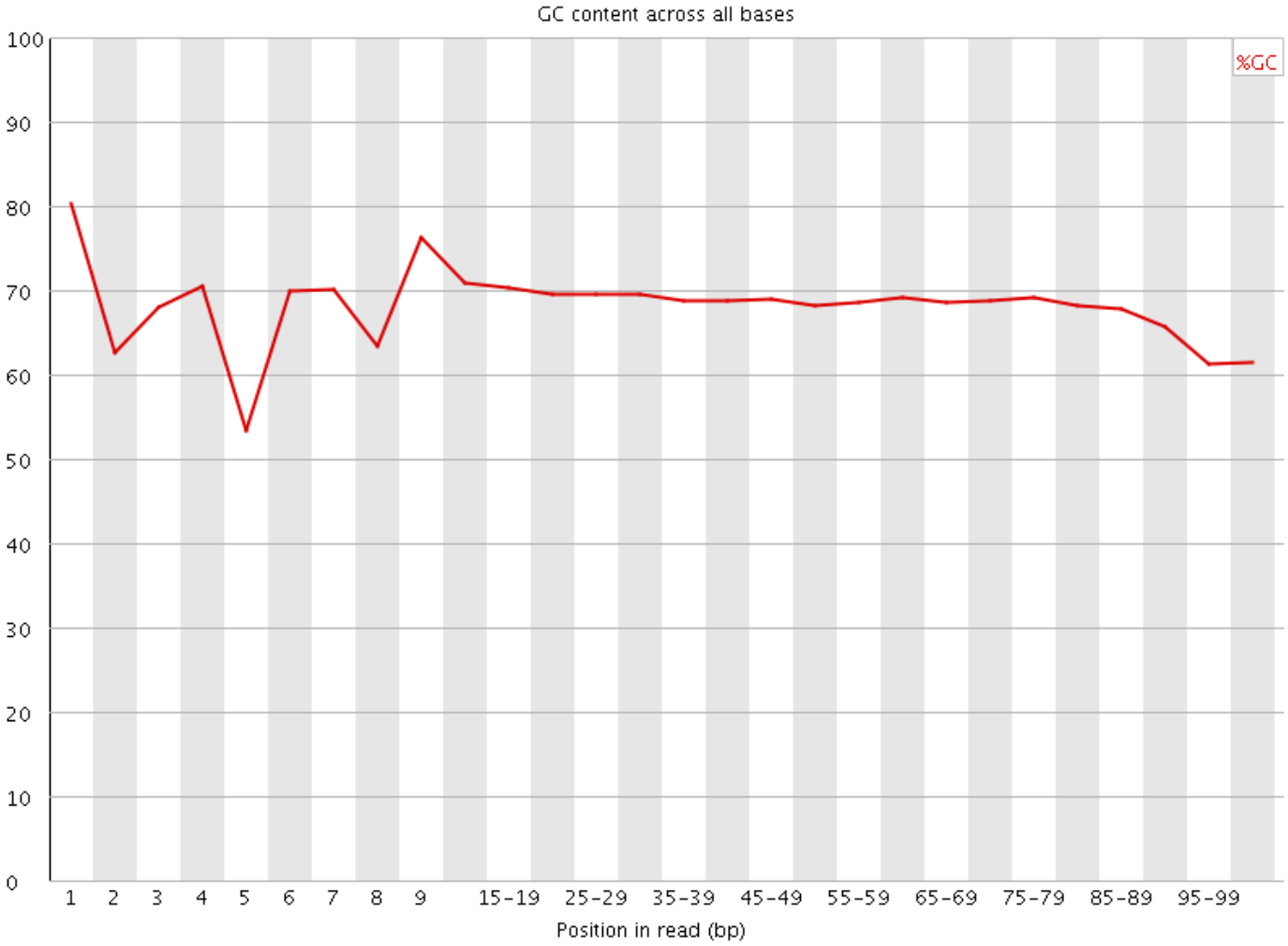
FASTQC

Per sequence GC content



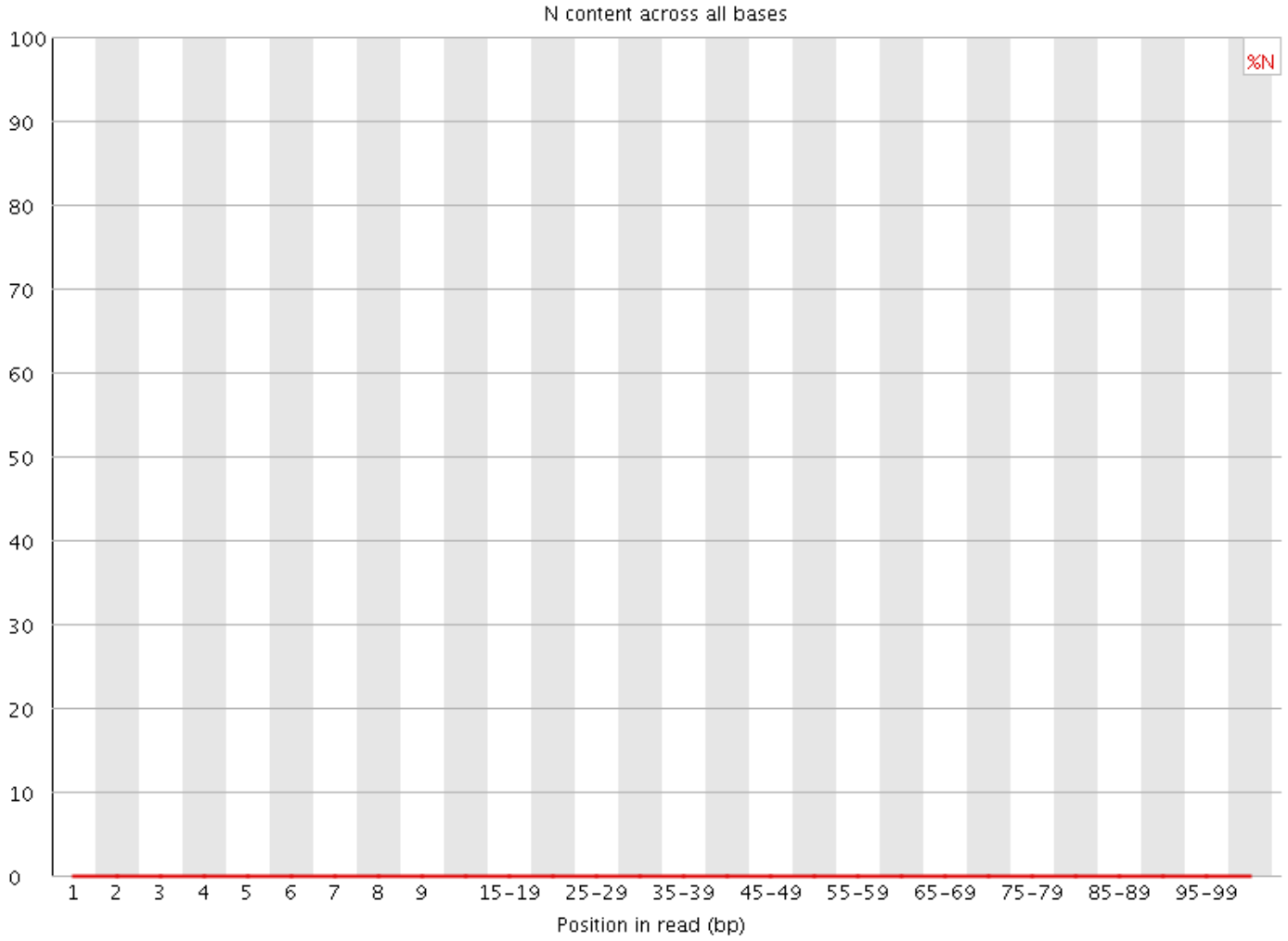
FASTQC

Per base GC content



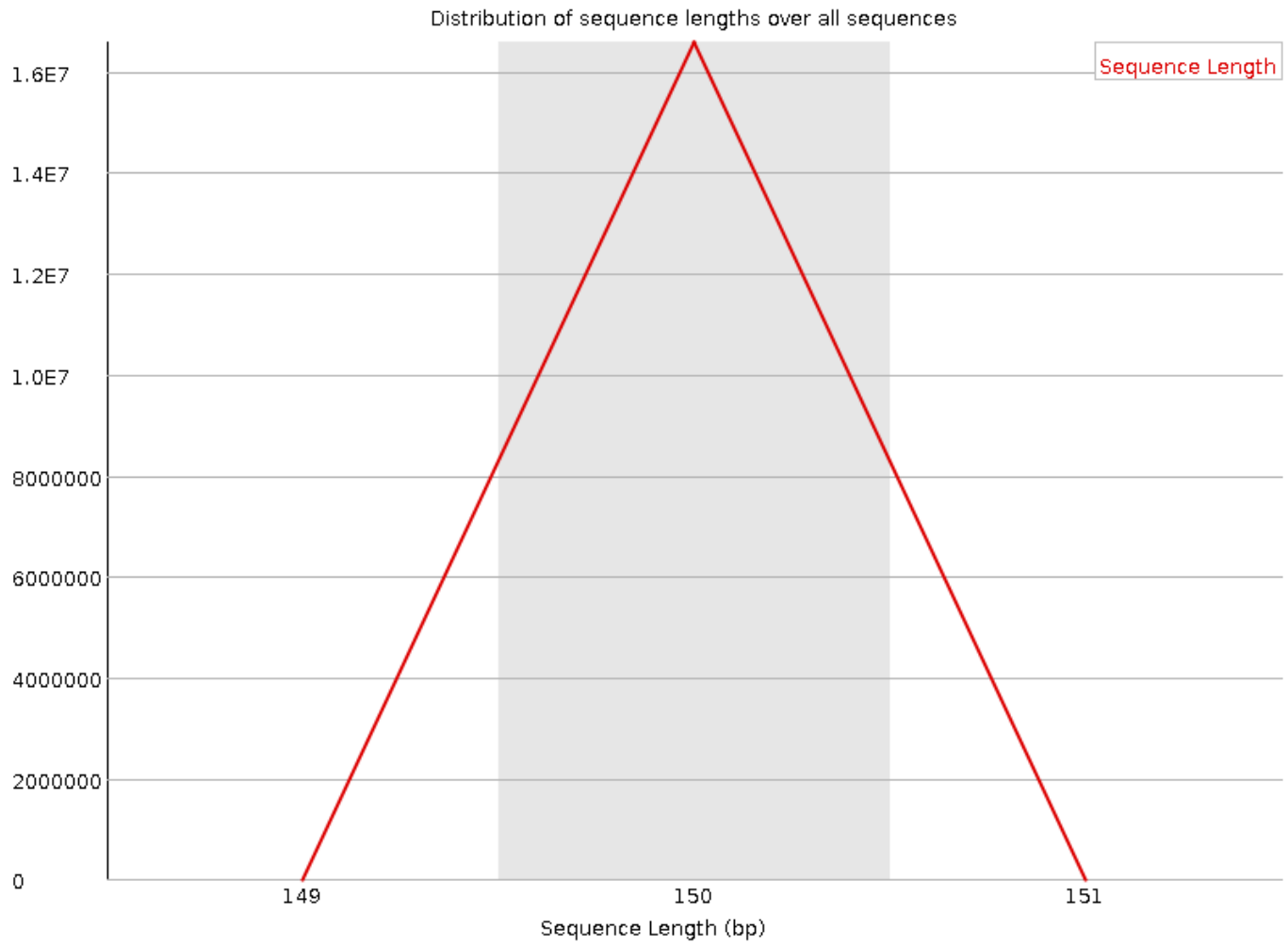
FASTQC

Per base N content



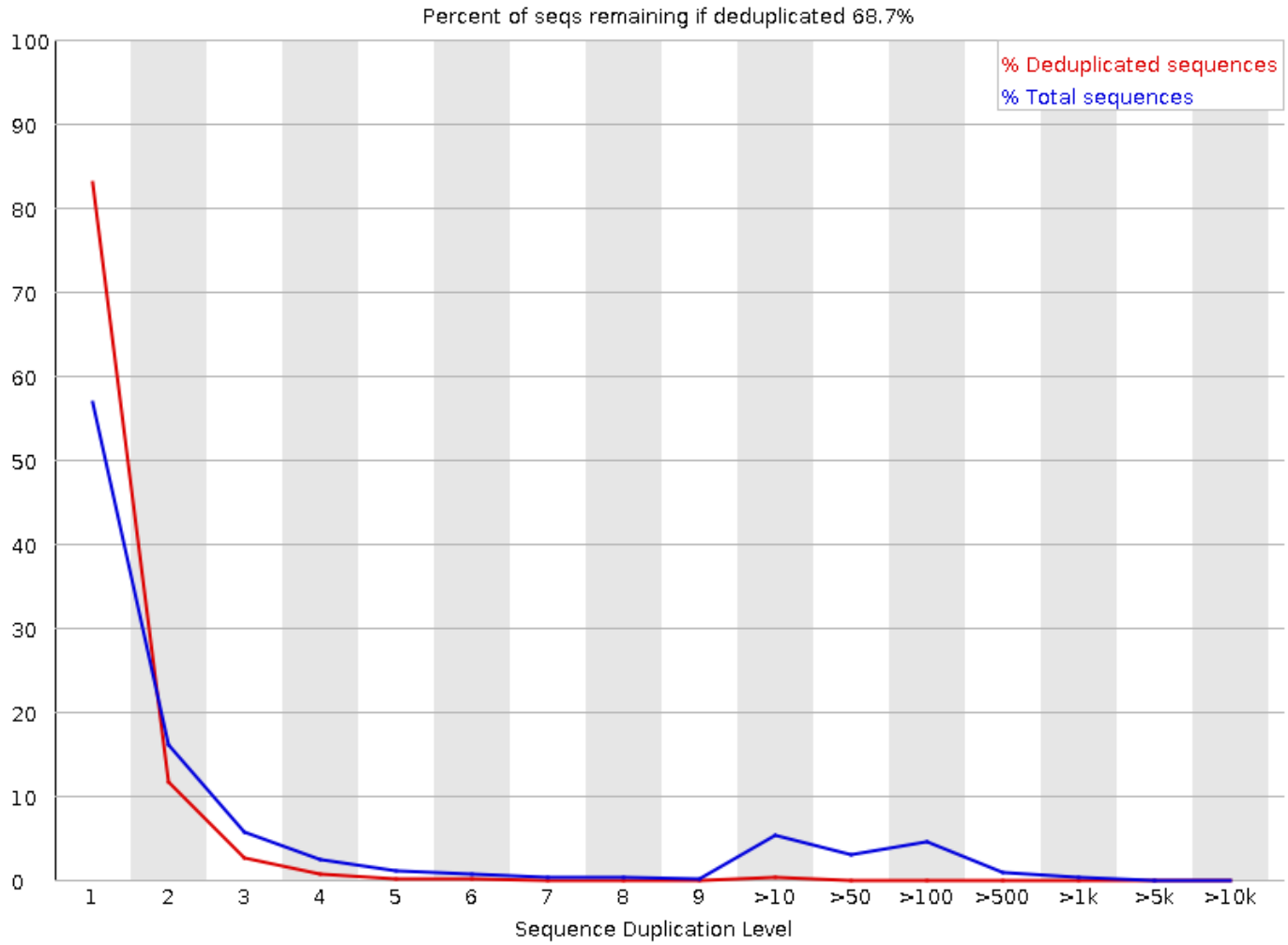
FASTQC

Sequence Length Distribution



FASTQC

Sequence Duplication Levels

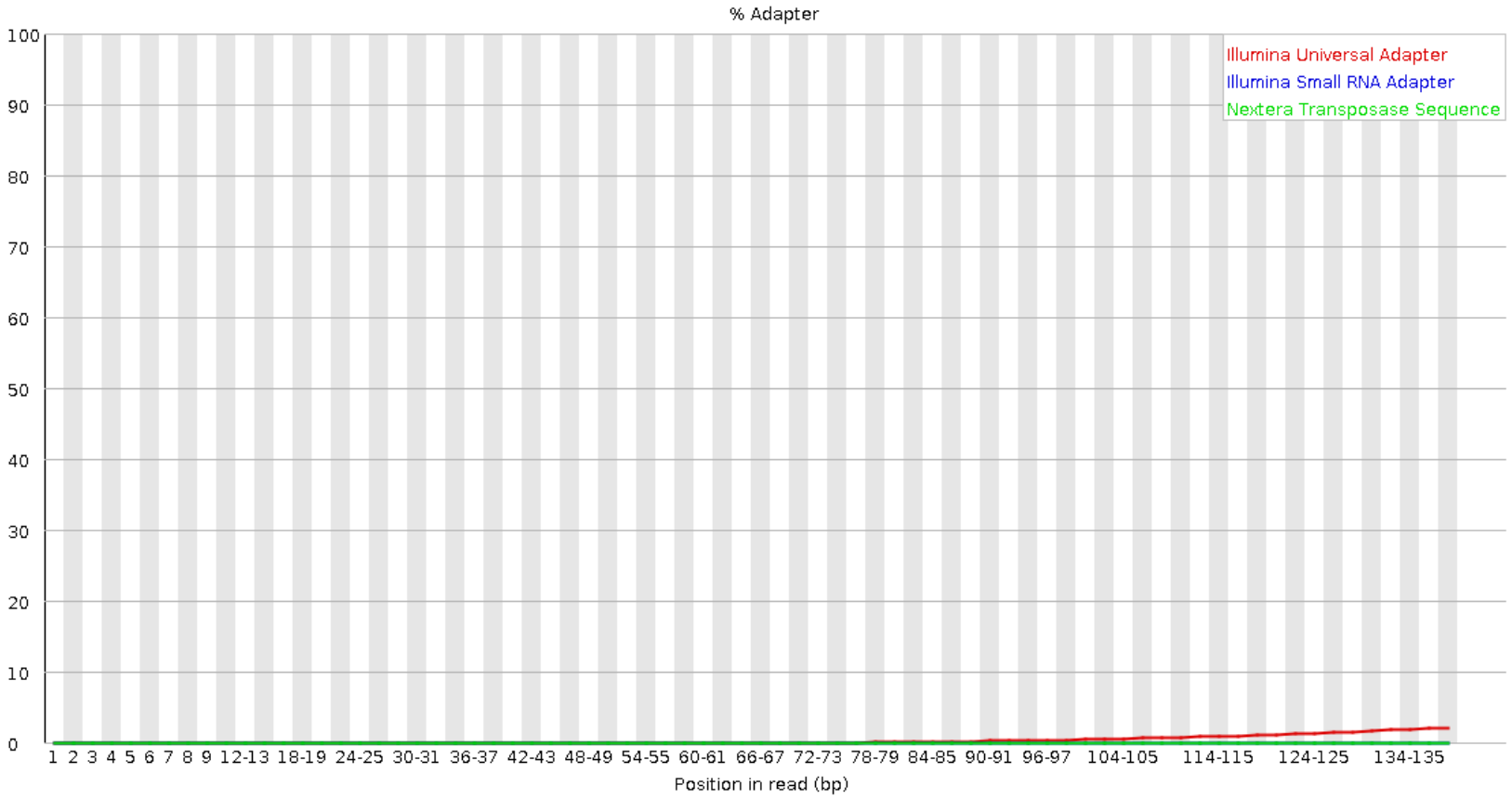


FASTQC

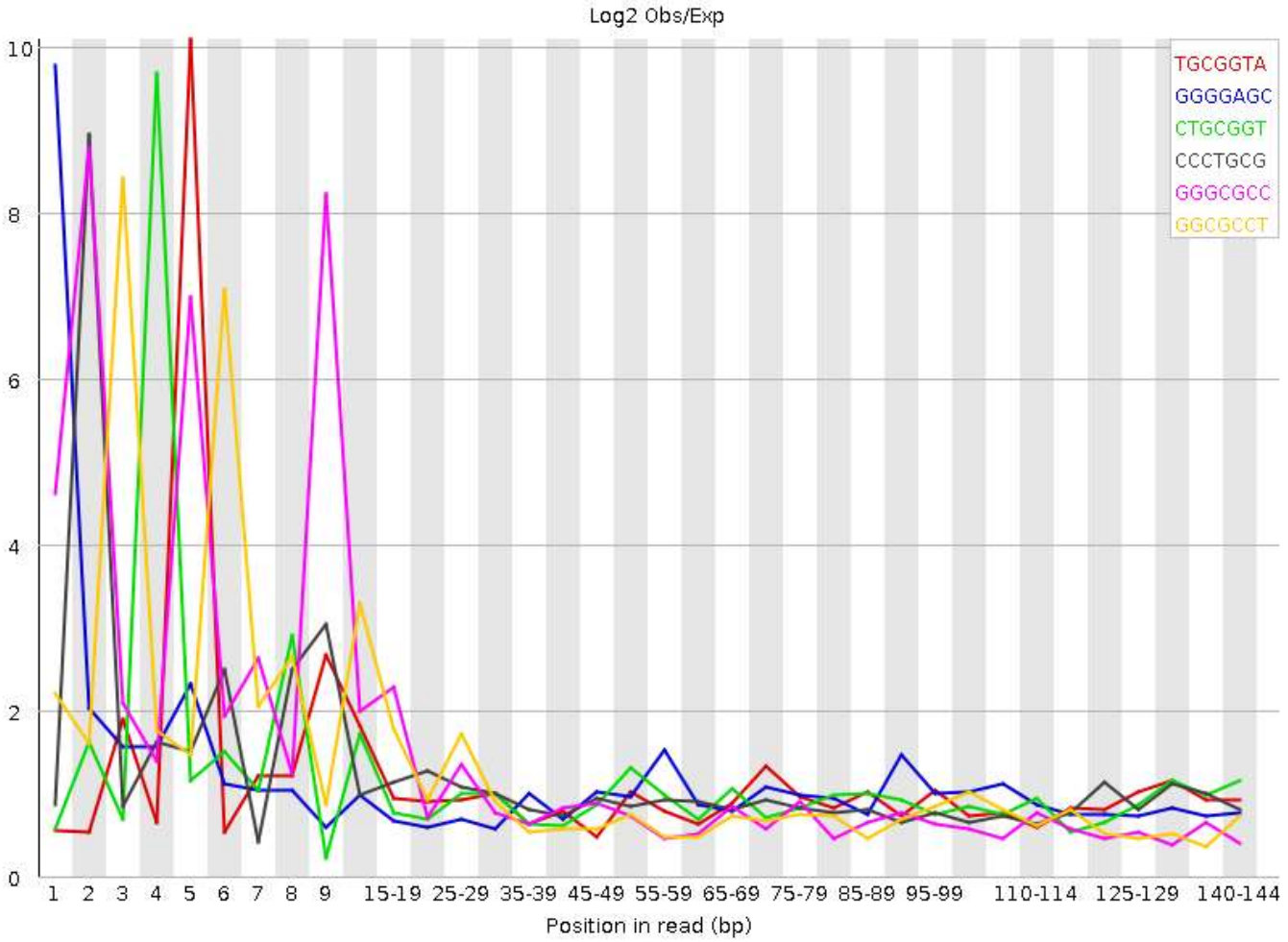
Overrepresented sequences

No overrepresented sequences

Adapter Content

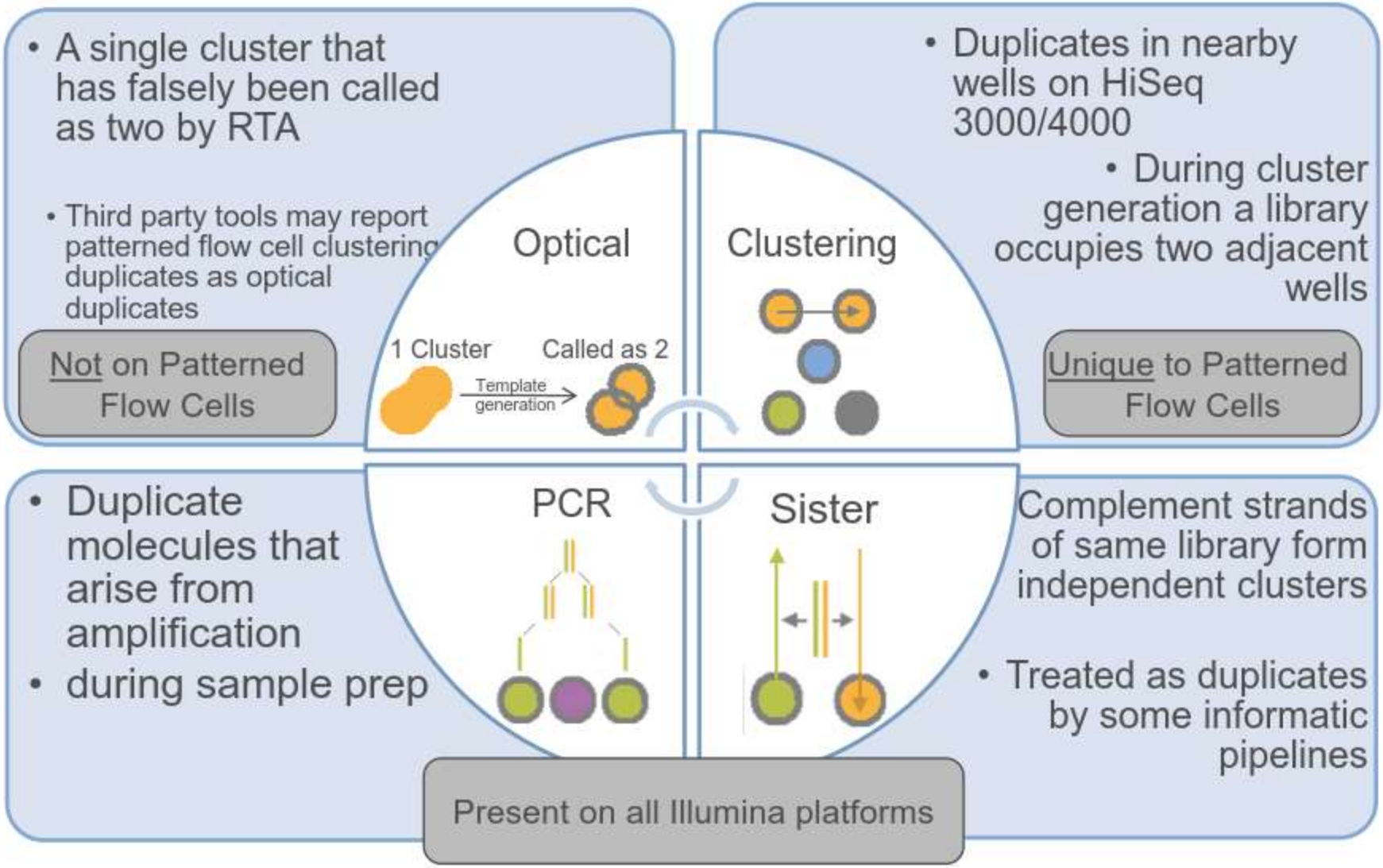


Kmer Content



Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
TGCGGTA	6425	0.0	10.080686	5
GGGGAGC	9540	0.0	9.778594	1
CTGCGGT	6170	0.0	9.680999	4
CCCTGCG	6605	0.0	8.939233	2
GGGCGCC	5155	0.0	8.799765	2

A Review of Sequencing Duplicate Types

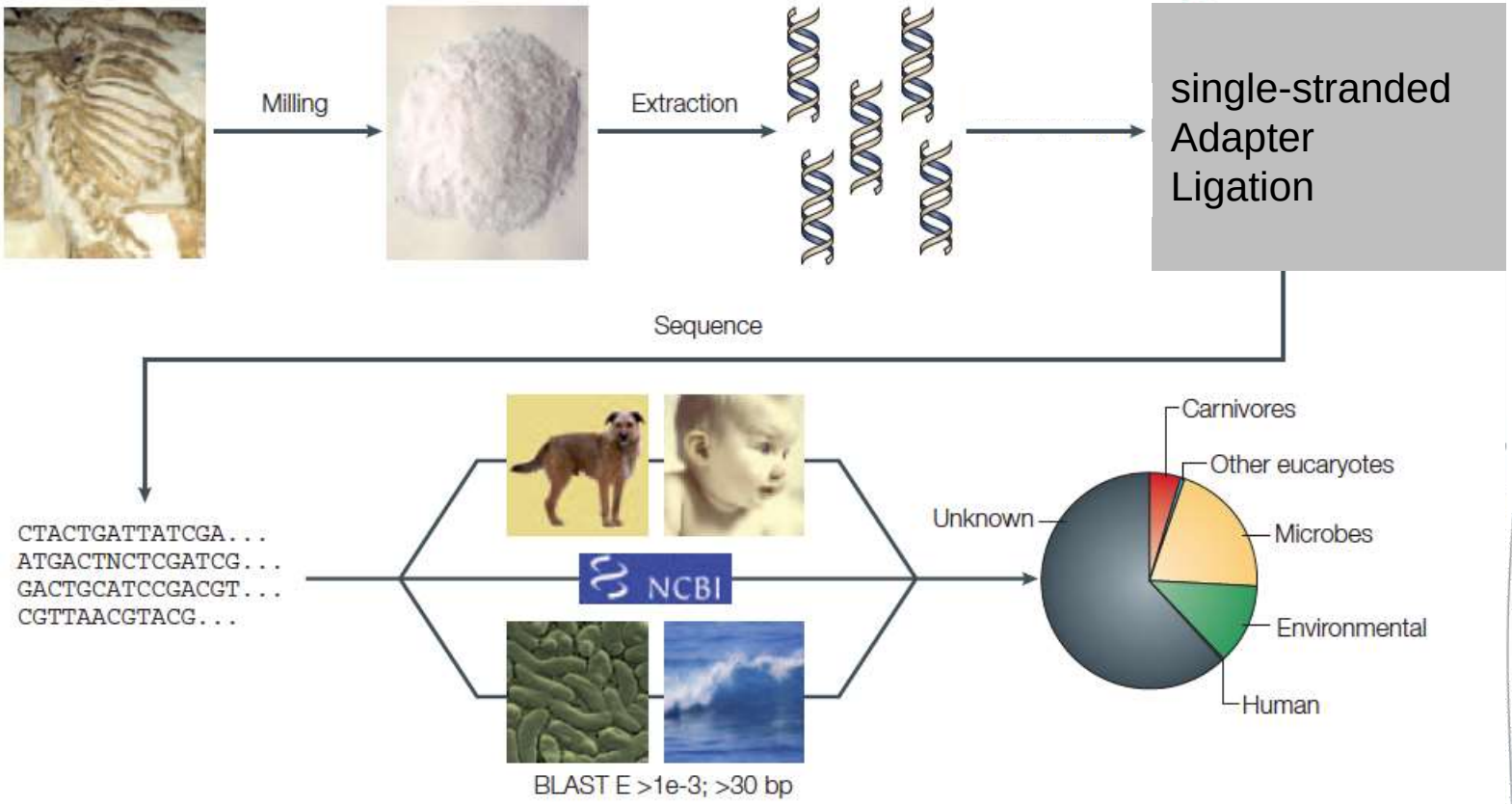


illumina

“If you can put adapters on it,

we can sequence it!”

Know your sample



No need to be scared of HTS

UC Davis Center for Plant Diversity/Herbarium

- The Herbarium archives contain over 300,000 dried specimens.

- Search for Grapevine Red Blotch-Associated Virus
- Virus traces found by PCR

Maher Al Rwahnih
UCD Plant Foundation
Plant Services

Studying historic Bean varieties from herbarium samples

- GBS (Genotyping-By-Sequencing)
- 60 year old herbarium samples

Sarah Dohle,
Gepts Lab

Quantitation & QC methods

- Intercalating dye methods (PicoGreen, Qubit, etc.):

Specific to dsDNA, accurate at low levels of DNA
Great for pooling of indexed libraries to be sequenced in one lane
Requires standard curve generation, many accurate pipetting steps

- Bioanalyzer:

Quantitation is good for rough estimate
Invaluable for library QC
High-sensitivity DNA chip allows quantitation of low DNA levels

- qPCR

Most accurate quantitation method
More labor-intensive
Must be compared to a control

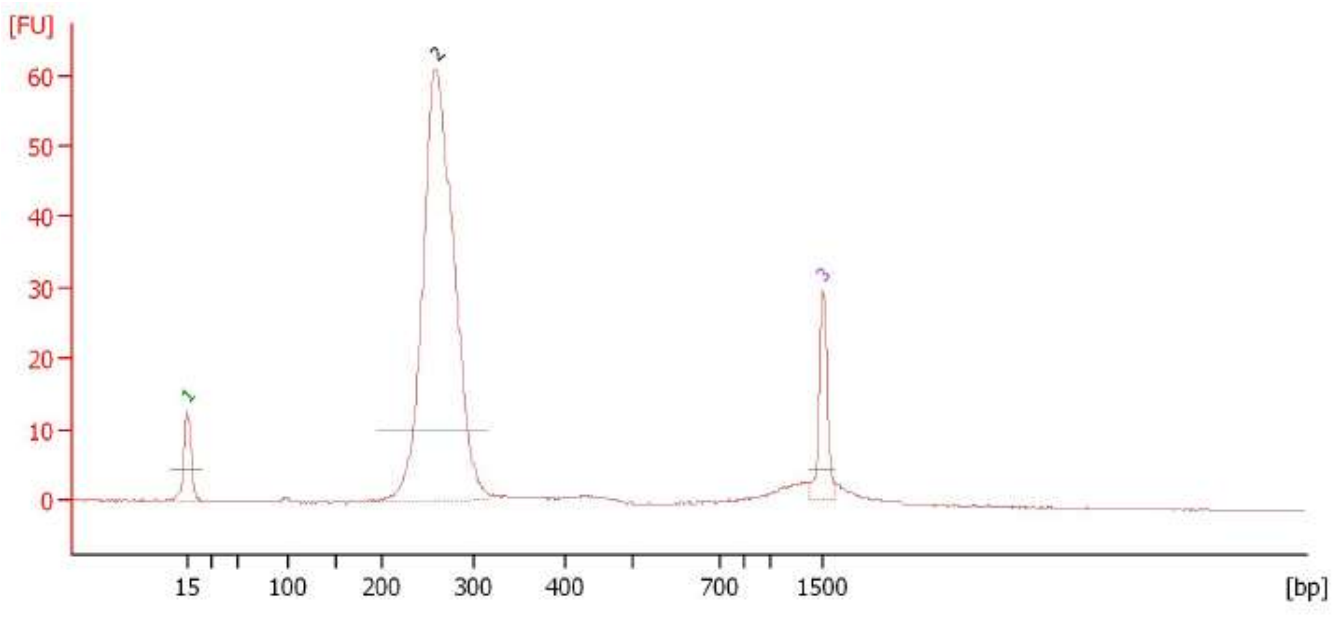
Optional: PCR-free libraries

- PCR-free library:
  - if concentration allows
  - Reduction of PCR bias against e.g. GC rich or AT rich regions, especially for metagenomic samples

OR

- Library enrichment by PCR:
  - Ideal combination: high input and low cycle number; low-bias polymerase

Library QC by Bioanalyzer

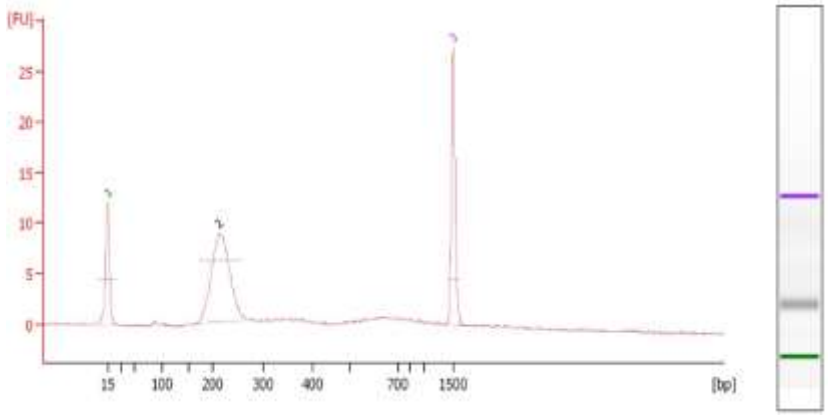


Predominant species of appropriate MW

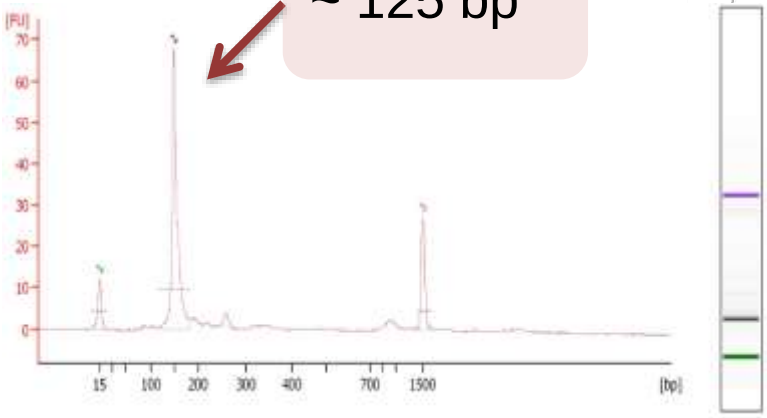
Minimal primer dimer or adapter dimers

Minimal higher MW material

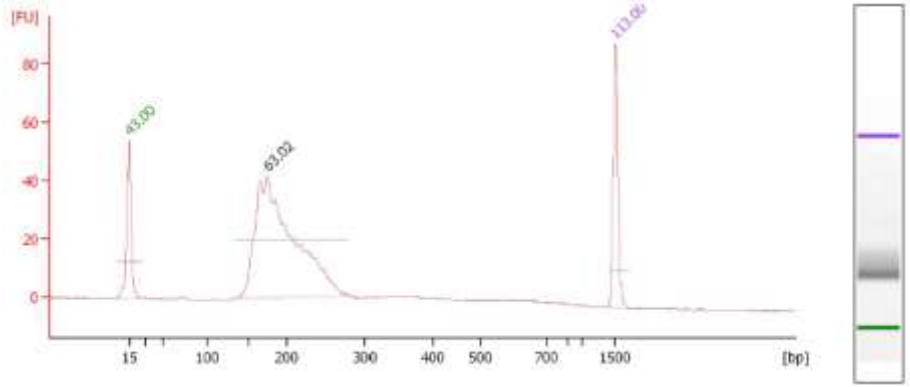
Library QC by Bioanalyzer



Beautiful

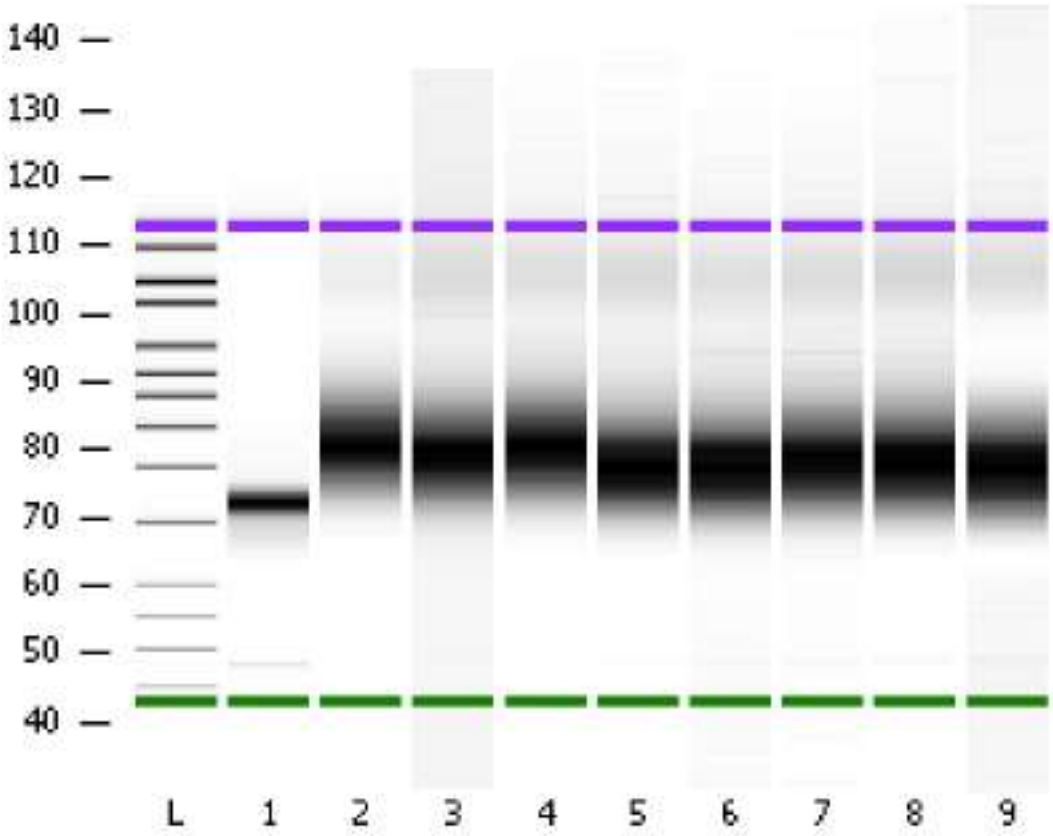


100% Adapters

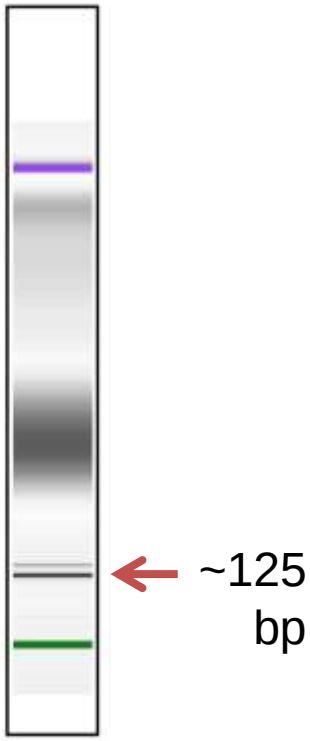


Beautiful

Library QC

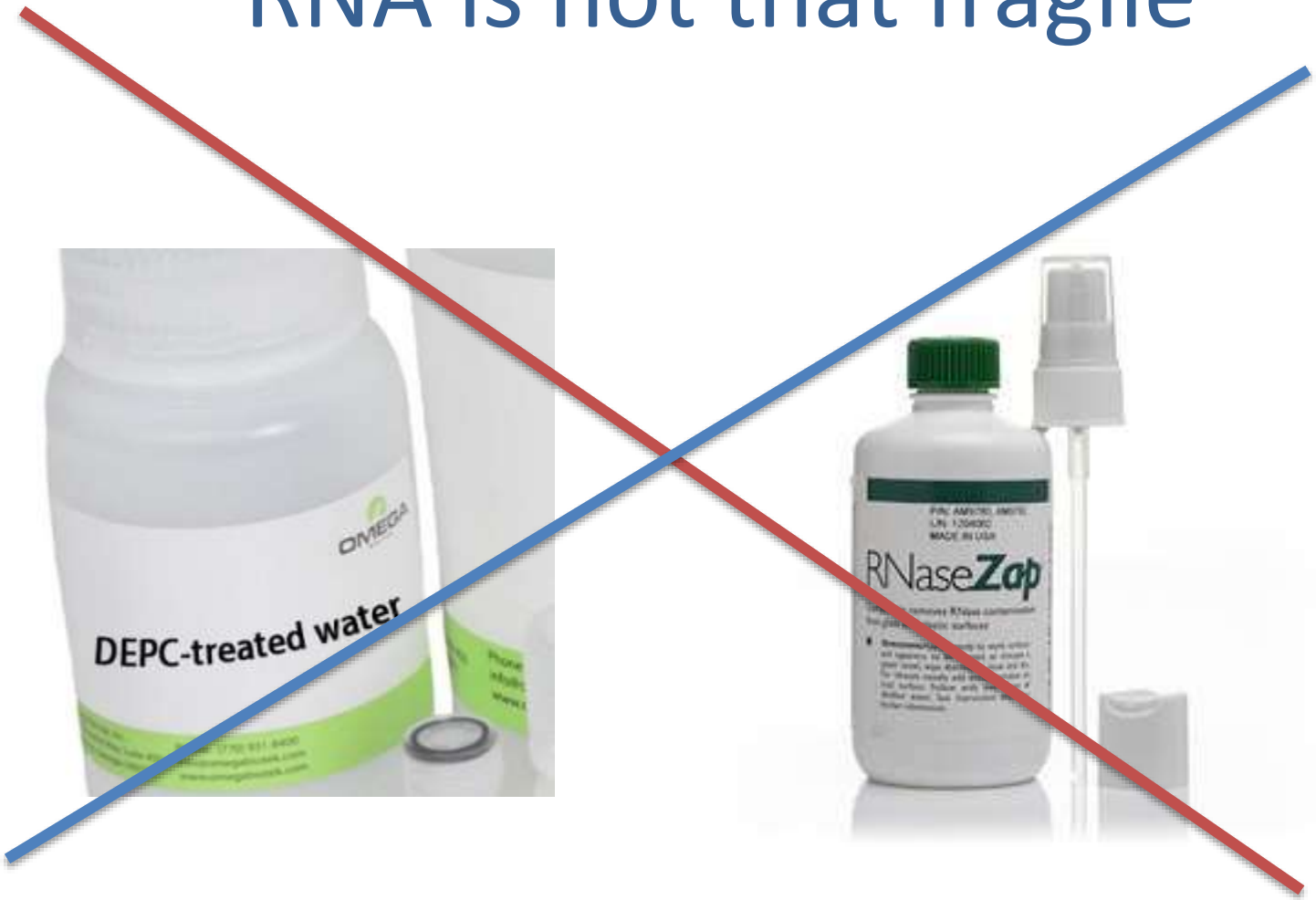


Examples for successful libraries



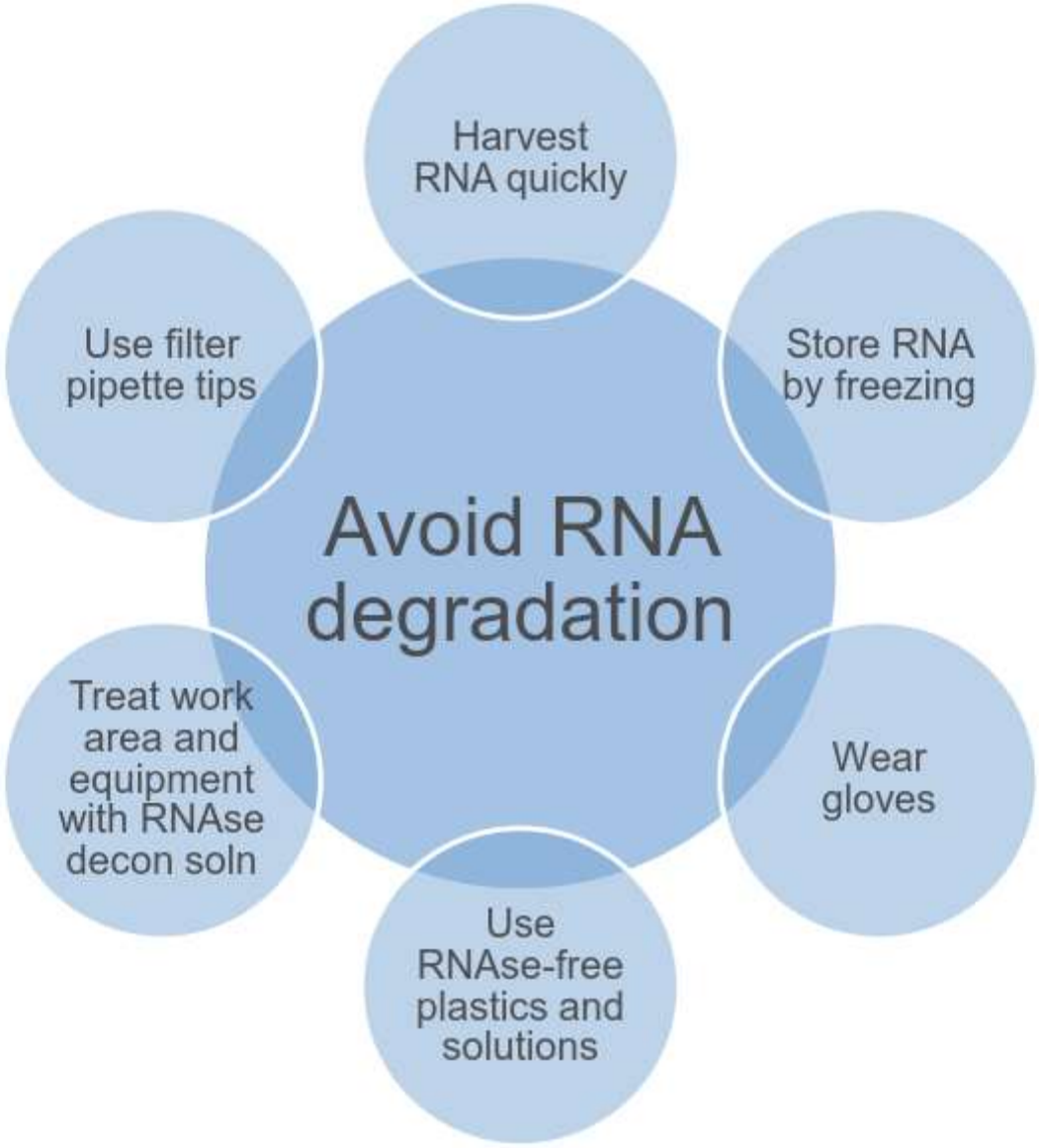
Adapter contamination at ~125 bp

RNA is not that fragile



Actually: Avoid DEPC-treated reagents -- remnants can inhibit enzymes

RNA Handling Best Practices



Recommended RNA input

Library prep kit	Starting material
mRNA (TruSeq)	100 ng – 4 µg total RNA
Directional mRNA (TruSeq)	1 – 5 µg total RNA or 50 ng mRNA
Apollo324 library robot (strand specific)	100 ng mRNA
Small RNA (TruSeq)	100 ng -1 µg total RNA
Ribo depletion (Epicentre)	500 ng – 5 µg total RNA
SMARTer™ Ultra Low RNA (Clontech)	100 pg – 10 ng
Ovation RNA seq V2, Single Cell RNA seq (NuGen)	10 ng – 100 ng

Standard RNA-Seq library protocol

- QC of total RNA to assess integrity
- Removal of rRNA (most common)
  - mRNA isolation
  - rRNA depletion
- Fragmentation of RNA
- Reverse transcription and second-strand cDNA synthesis
- Ligation of adapters
- PCR Amplification
- Purify, QC and Quantify

- 18S (2500b) , 28S (4000b)

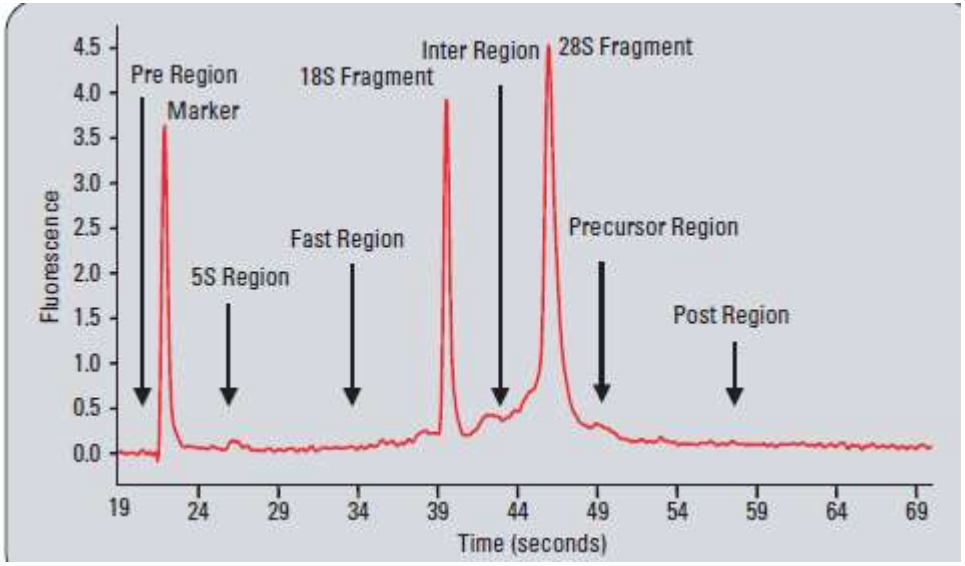
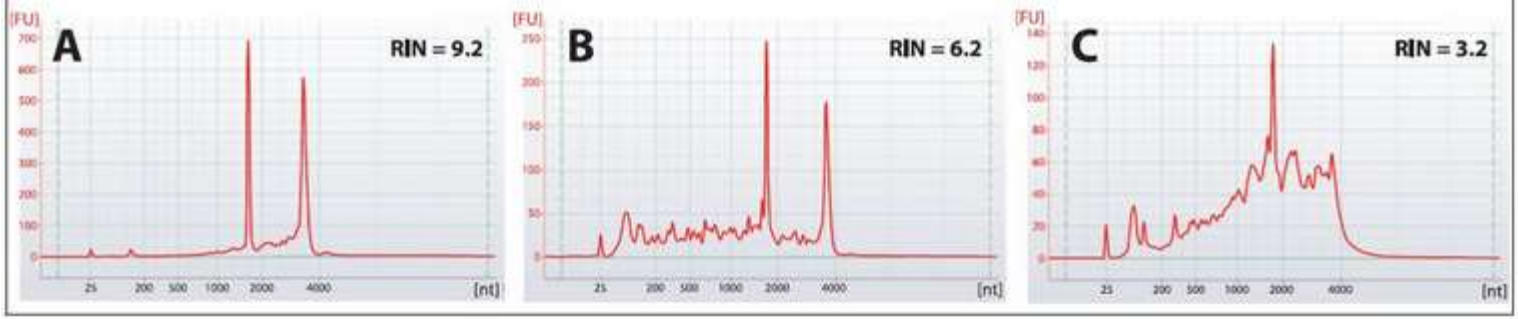
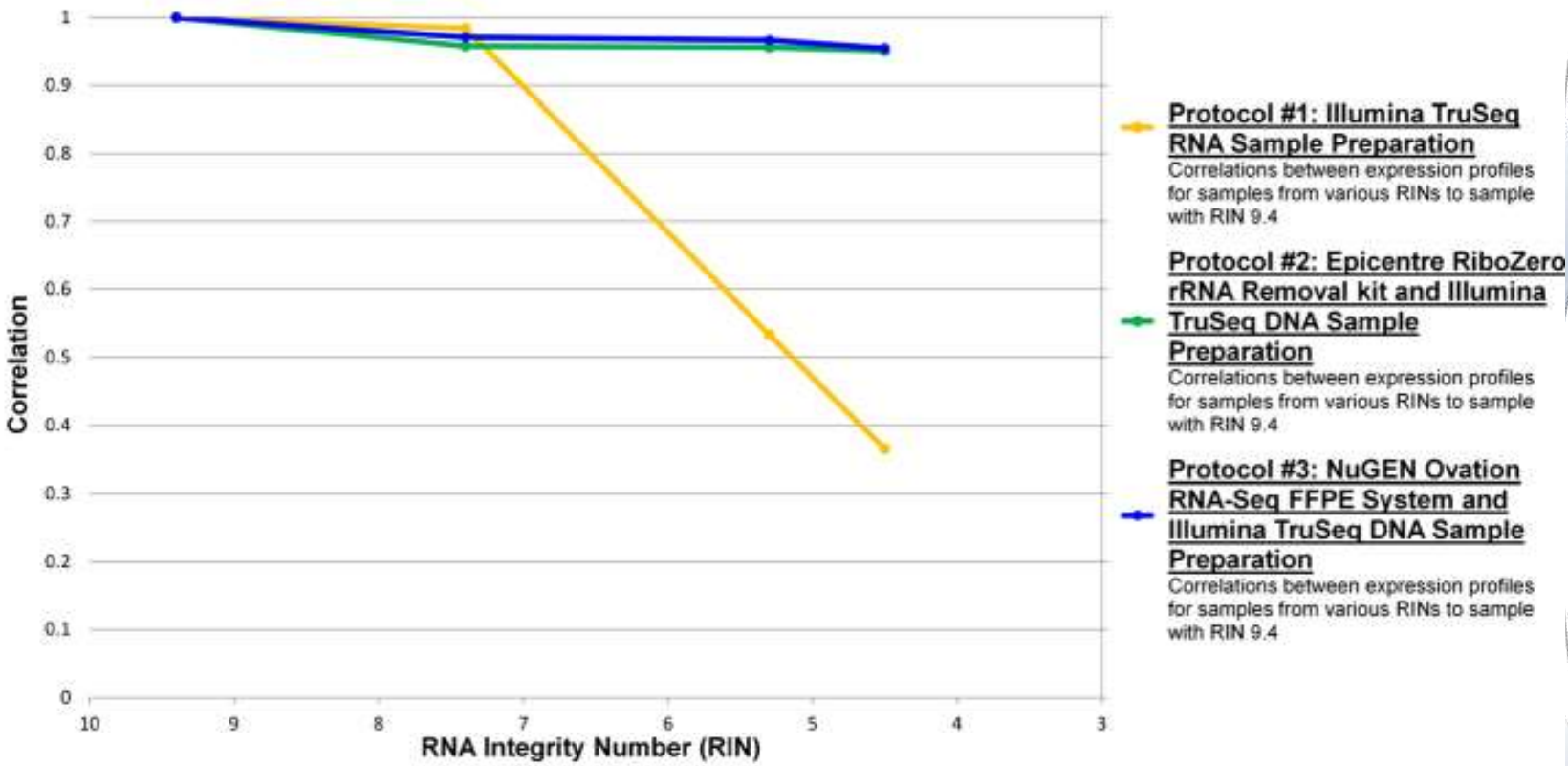


Figure 2.1 Example Agilent Bioanalyzer Electropherograms from three different total RNAs of varying integrity. Panel [A] represents a highly intact total RNA (RIN = 9.2), panel [B] represents a moderately intact total RNA (RIN = 6.2), and panel [C] represents a degraded total RNA sample (RIN = 3.2).



RNA integrity <> reproducibility



Chen et al. 2014

Considerations in choosing an RNA-Seq method

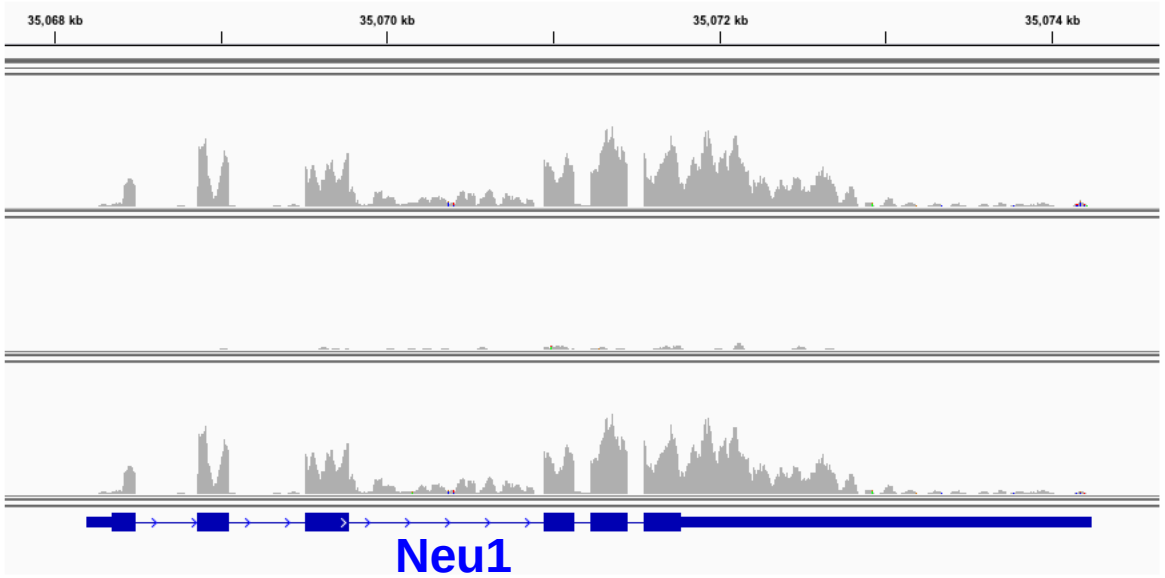
- Transcript type:
  - mRNA, extent of degradation
  - small/micro RNA
- Strandedness:
  - un-directional ds cDNA library
  - directional library
- Input RNA amount:
  - 0.1-4ug original total RNA
  - linear amplification from 0.5-10ng RNA
- Complexity:
  - original abundance
  - cDNA normalization for uniformity
- Boundary of transcripts:
  - identify 5’ and/or 3’ ends
  - poly-adenylation sites
  - Degradation, cleavage sites

Is strand-specific information important?

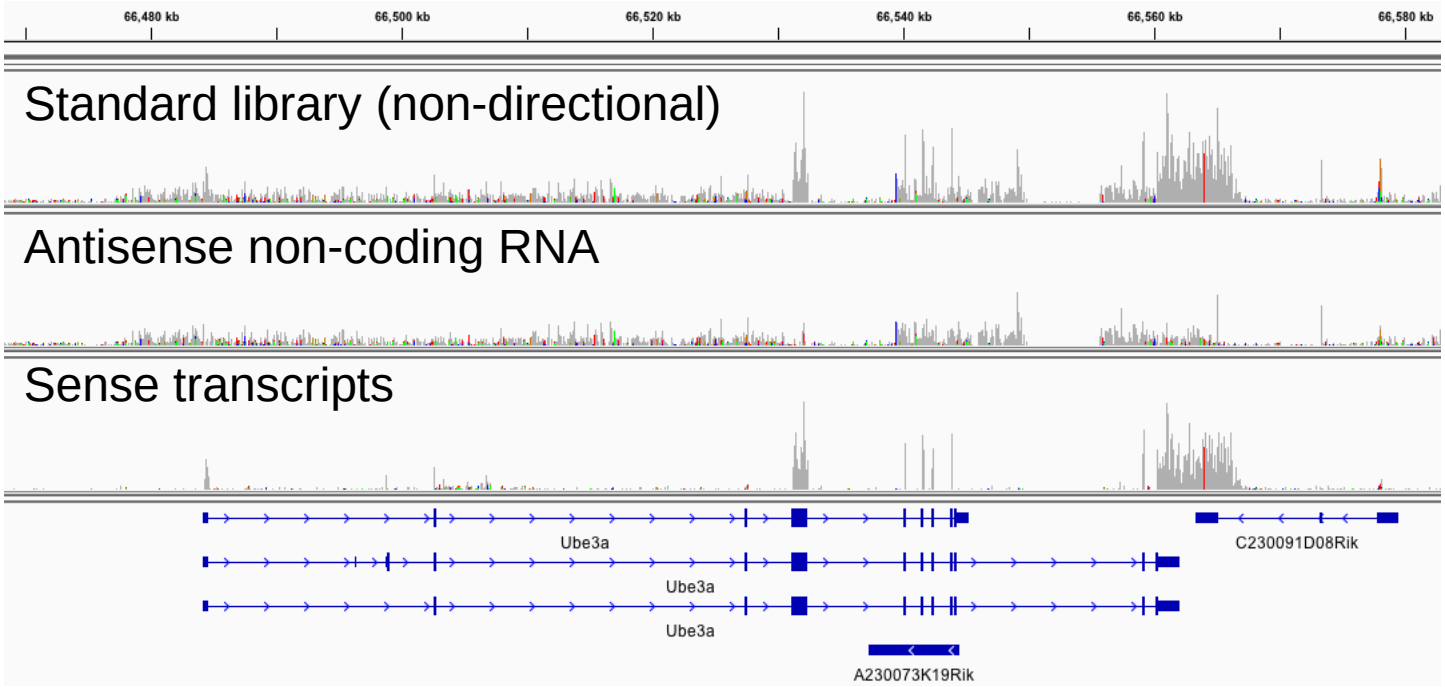
Standard library (non-directional)

antisense

sense



Strand-specific RNA-seq



- Informative for non-coding RNAs and antisense transcripts
- Essential when NOT using polyA selection (mRNA)
- No disadvantage to preserving strand specificity

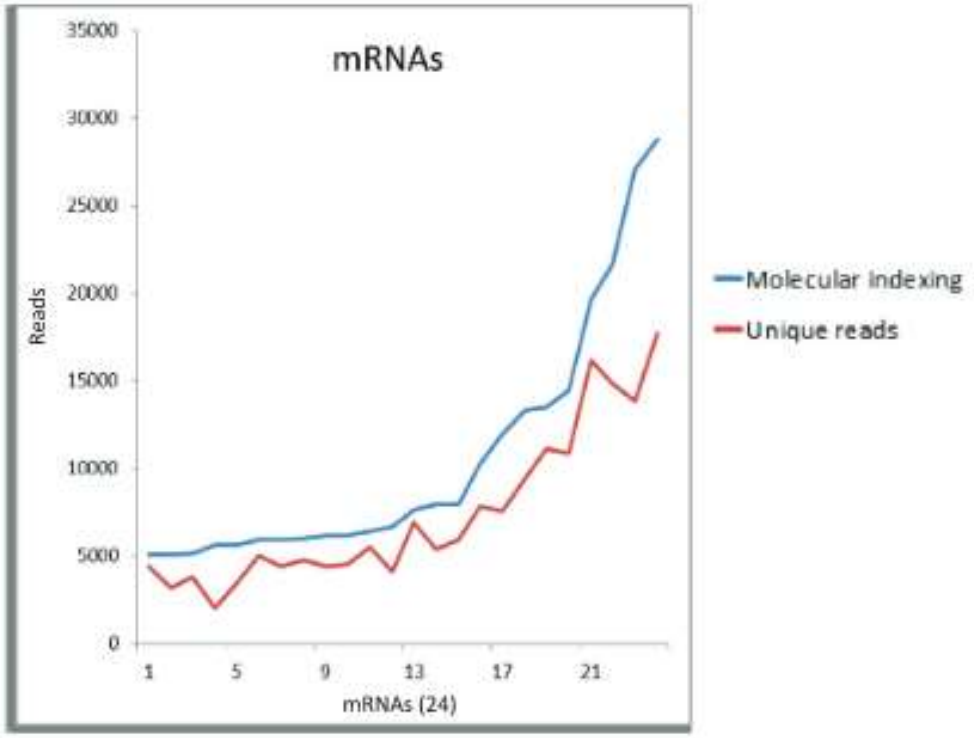
RNA-seq for DGE

- Differential Gene Expression (DGE)
  - 50 bp single end reads
  - 30 million reads per sample (eukaryotes)
    - 10 mill. reads > 80% of annotated genes
    - 30 mill. . reads > 90% of annotated genes
  - 10 million reads per sample (bacteria)

Molecular indexing – for precision counts

Molecular indexing – for precision counts

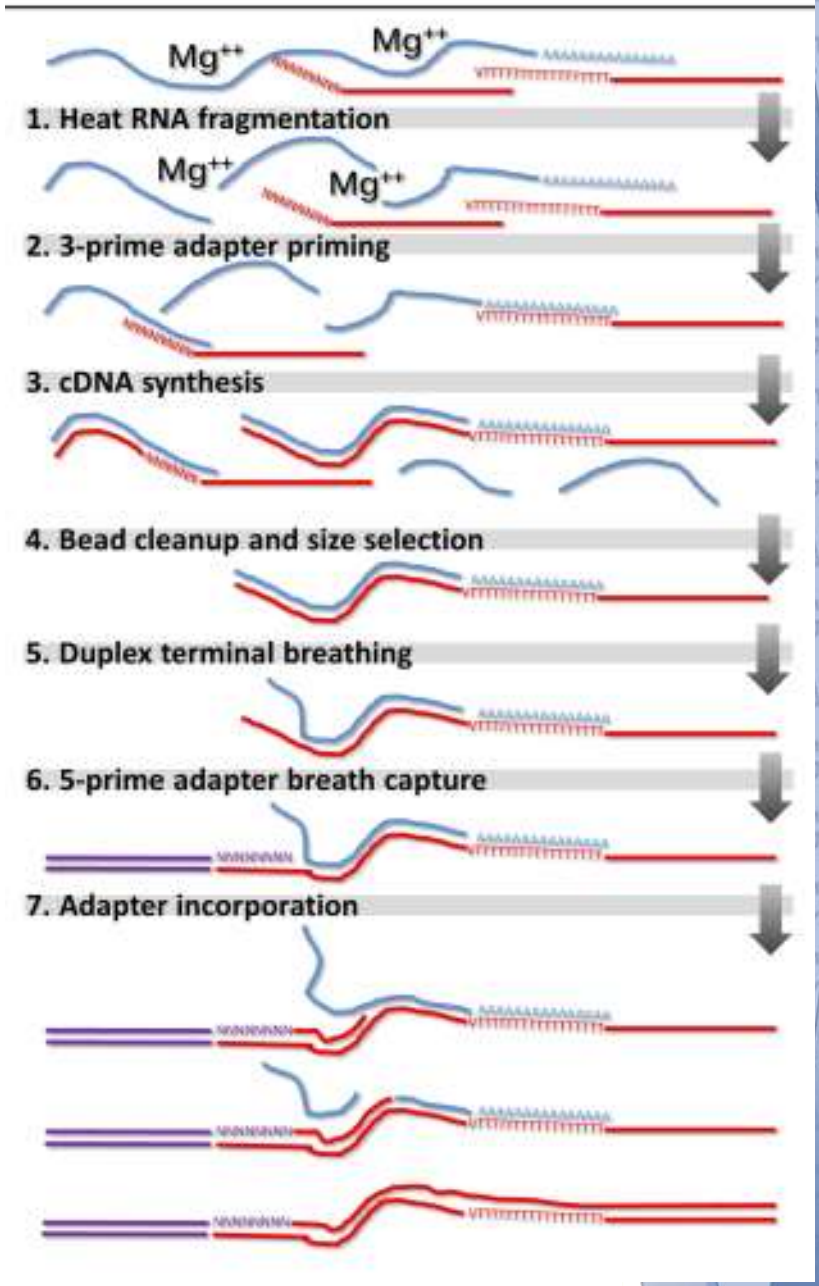
B



RNA-seq: cheap and dirty

- 3’ Tag-sequencing
- Micro-array-like data
- Quant-Seq
- Brad-Seq (Townsley 2015)

AMARYLLIS NUCLEICS
BETTER SEQUENCING



3’-Tag-Seq

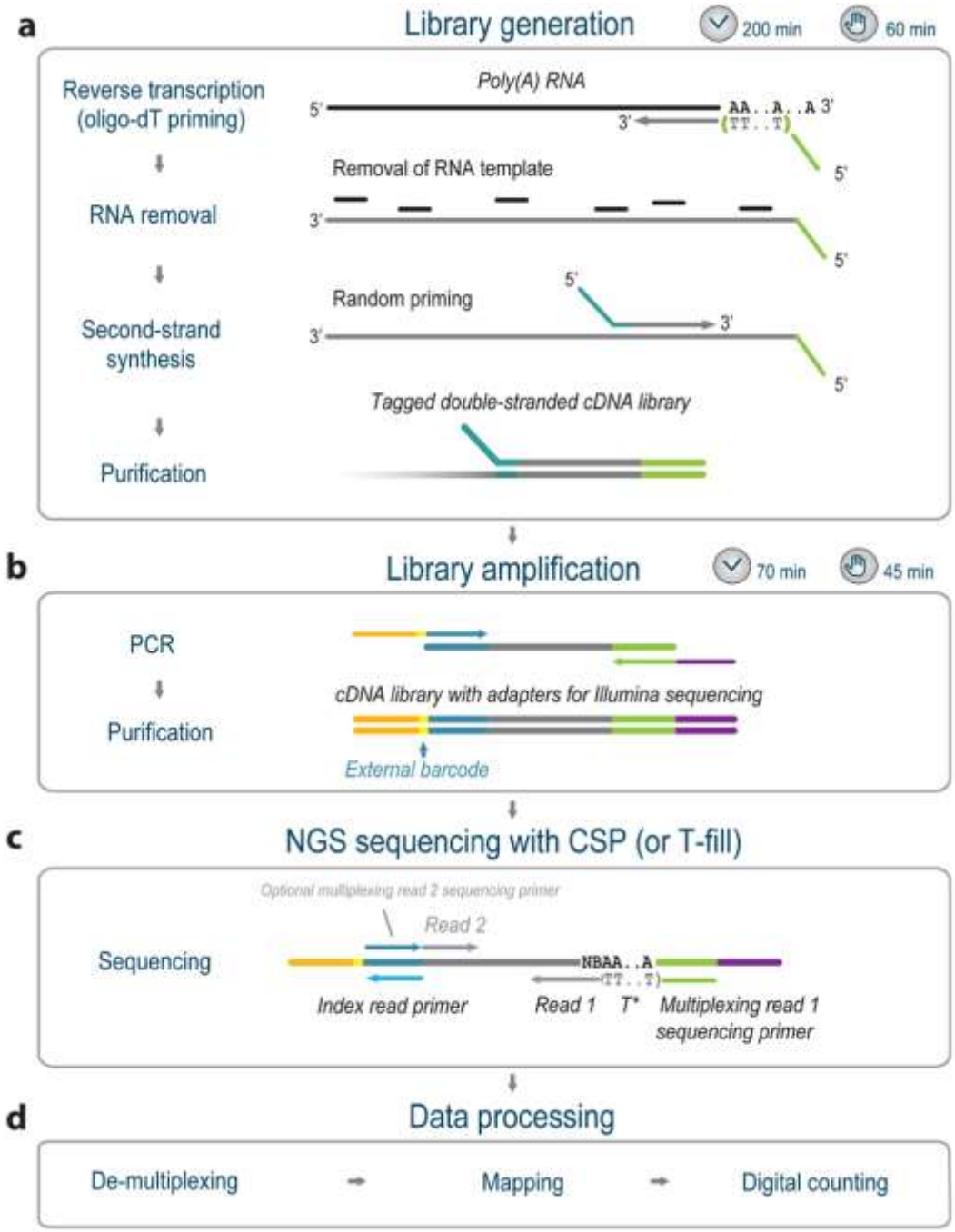
- In contrast to full length RNA-seq
- Sequencing 1/10 for the average transcript
- Less dependent on RNA integrity
- Microarray-like data

- Options:

- **BRAD-Seq : 3’ Digital Gene Expression**

- **Lexogen Quant-Seq**

Lexogen Quant-Seq



Other RNA-seq

- Transcriptome assembly:
  - 300 bp paired end plus
  - 100 bp paired end
- Long non coding RNA studies:
  - 100 bp paired end
  - 60-100 million reads
- Splice variant studies:
  - 100 bp paired end
  - 60-100 million reads

RNA-seq targeted sequencing:

- Capture-seq (Mercer et al. 2014)
- Nimblegen and Illumina

- Low quality DNA (FFPE)
- Lower read numbers 10 million reads
- Targeting lowly expressed genes.

RNA-seq reproducibility

- Two big studies multi-center studies (2014)
- High reproducibility of data given:

  - same library prep kits, same protocols
  - same RNA–samples
  - RNA isolation protocols have to be identical
  - robotic library preps?

PACIFIC BIOSCIENCES™

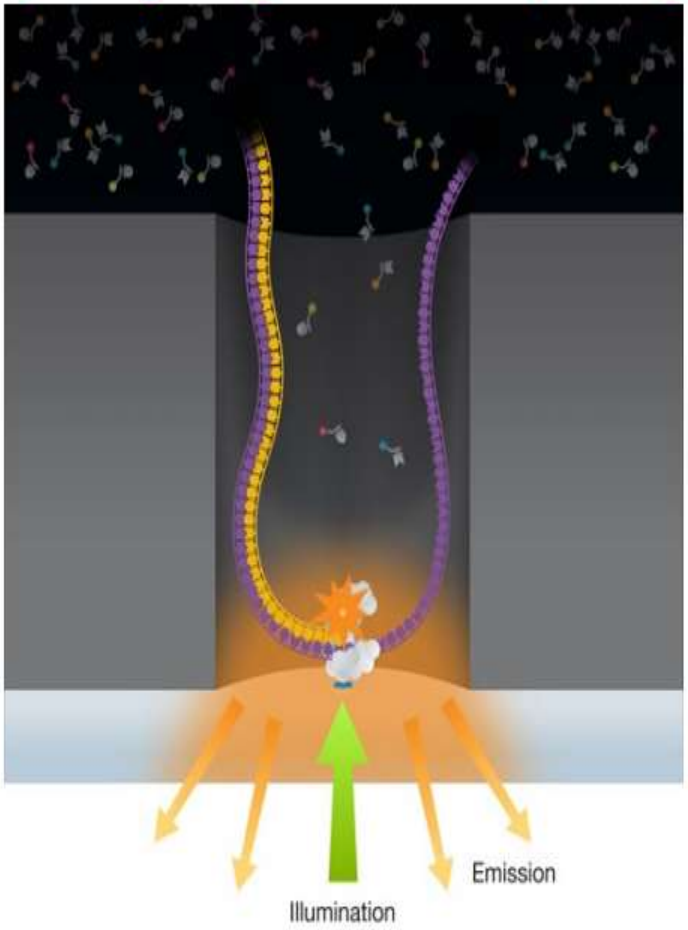
http://pacificbiosciences.com

THIRD GENERATION DNA SEQUENCING

Single Molecule Real Time (SMRT™) sequencing
Sequencing of single DNA molecule by single polymerase
Very long reads: average reads over 8 kb, up to 30 kb
High error rate (~13%).
Complementary to short accurate reads of Illumina

Third Generation Sequencing : Single Molecule Sequencing

Pacific Biosciences

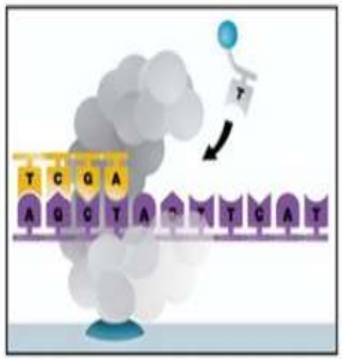


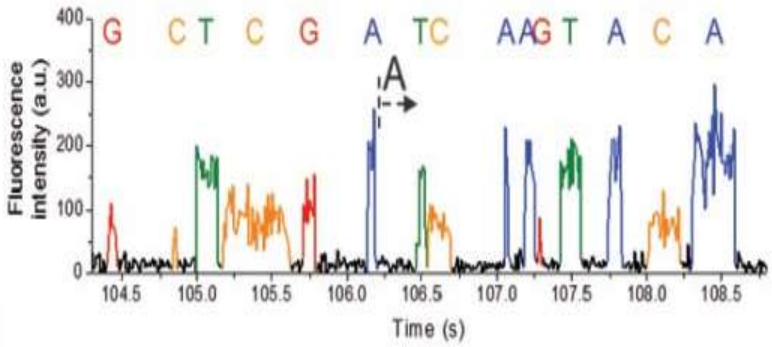
4 nucleotides with different fluorescent dye simultaneous present

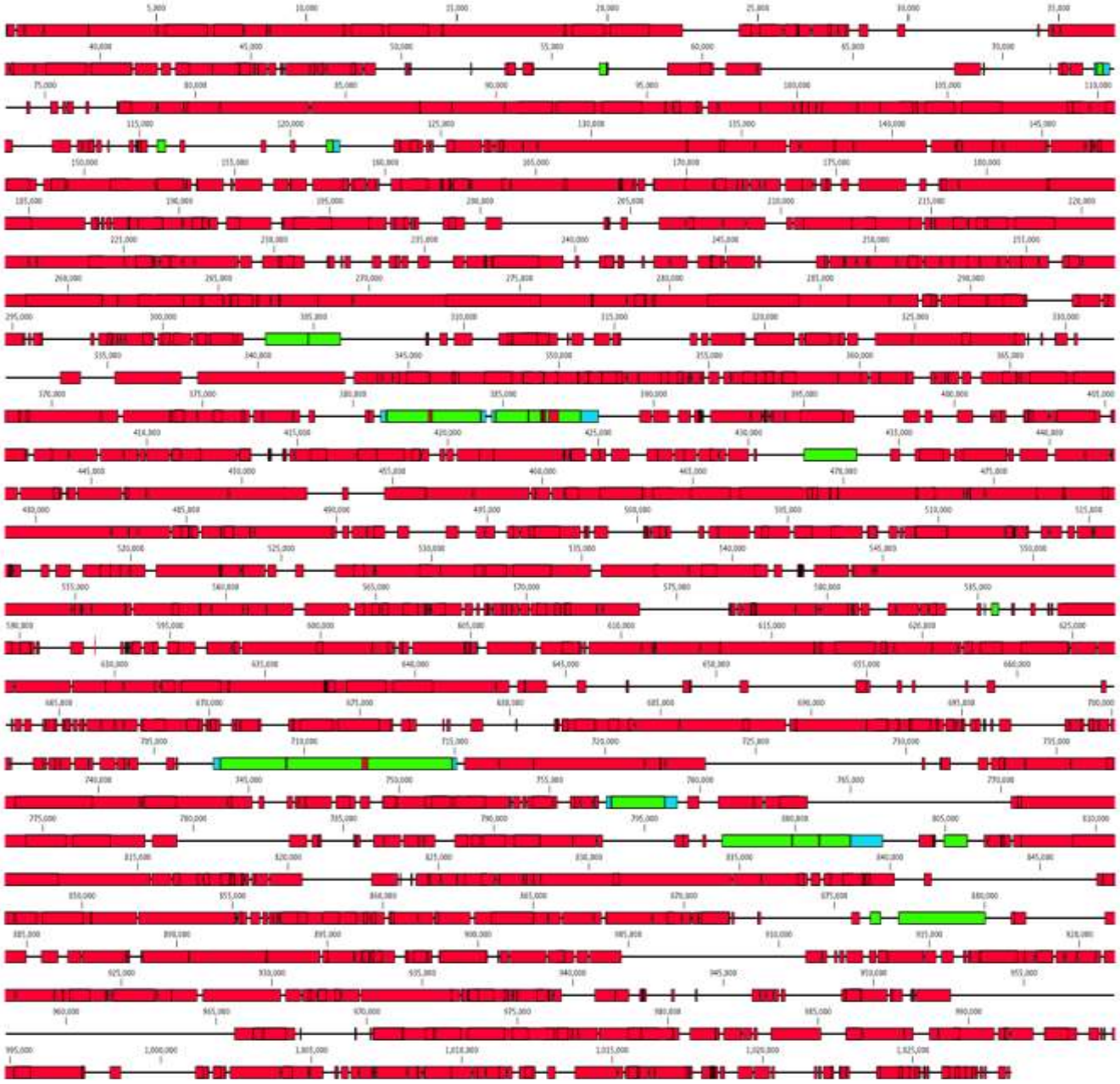
2-3 nucleotides/sec
2-3 Kb (up to 50) read length
6 TB data in 30 minutes

laser damages polymerase

70 nm aperture
“Zero Mode Waveguide”





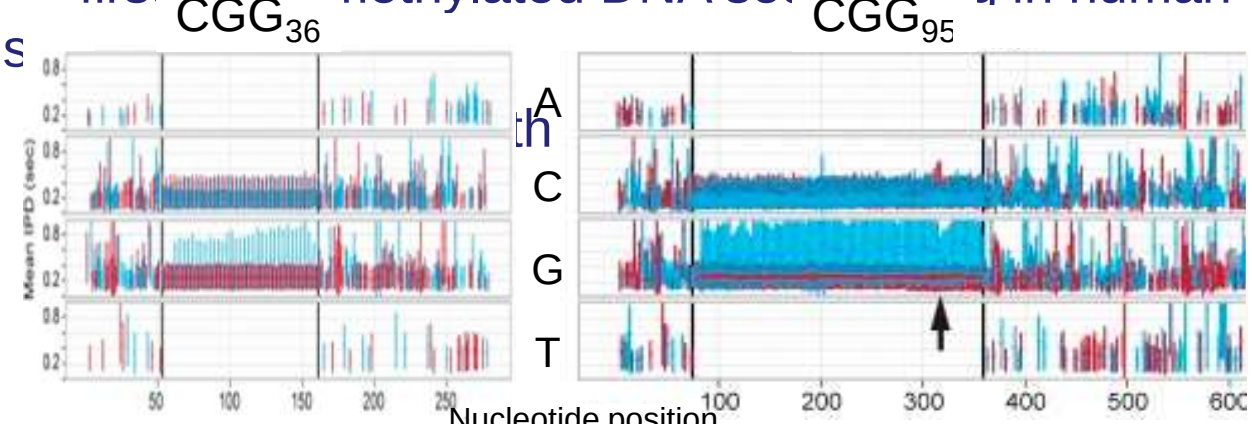


Damien Pelt

First Sequencing of CGG-repeat Alleles in Human Fragile X Syndrome using PacBio RS Sequencer

Paul Hagerman, Biochemistry and Molecular Medicine, SOM.

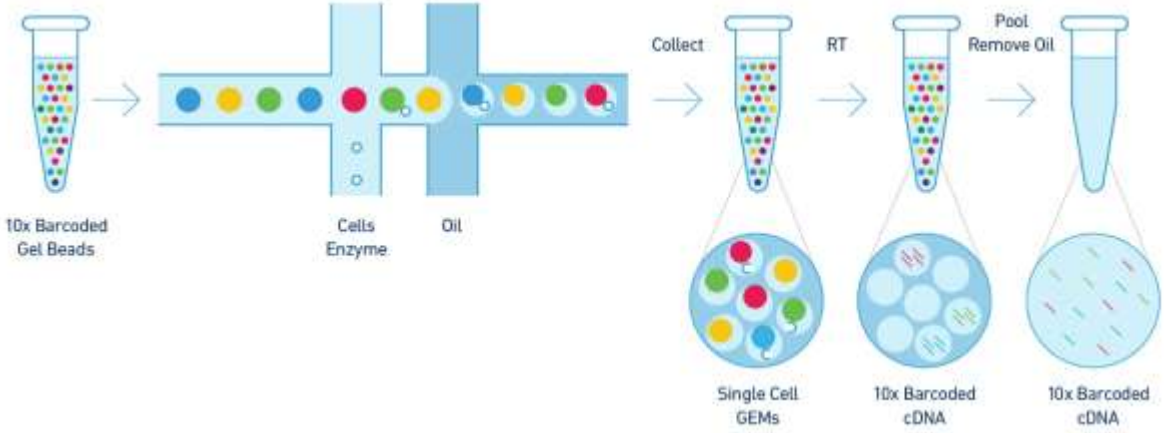
- Single-molecule sequencing of pure CGG array,
  - first for disease-relevant allele. Loomis *et al*. (2012) *Genome Research*.
  - applicable to many other tandem repeat disorders.
- Direct genomic DNA sequencing of methyl groups,
  - direct epigenetic sequencing (paper under review).
- Discovered 100% bias toward methylation of 20 CGG-repeat allele in female,
  – first [illegible] methylated DNA sequencing in human dis
- [illegible]th



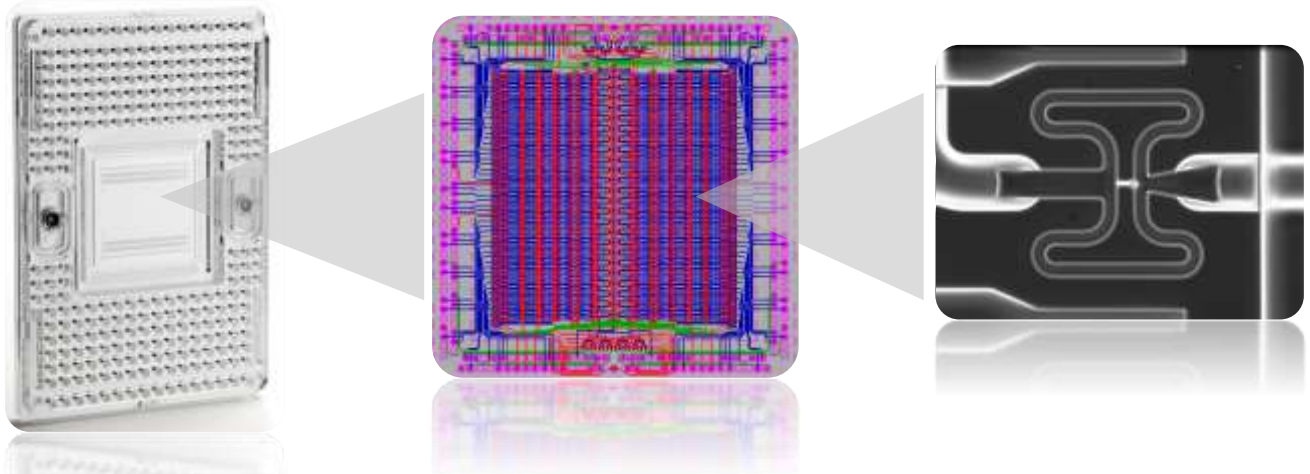
Iso-Seq Pacbio

- Sequence full length transcripts
  → no assembly
- High accuracy (except very long transcripts)
- More than 95% of genes show alternate splicing
- On average more than 5 isoforms/gene
- Precise delineation of transcript isoforms (PCR artifacts? chimeras?)

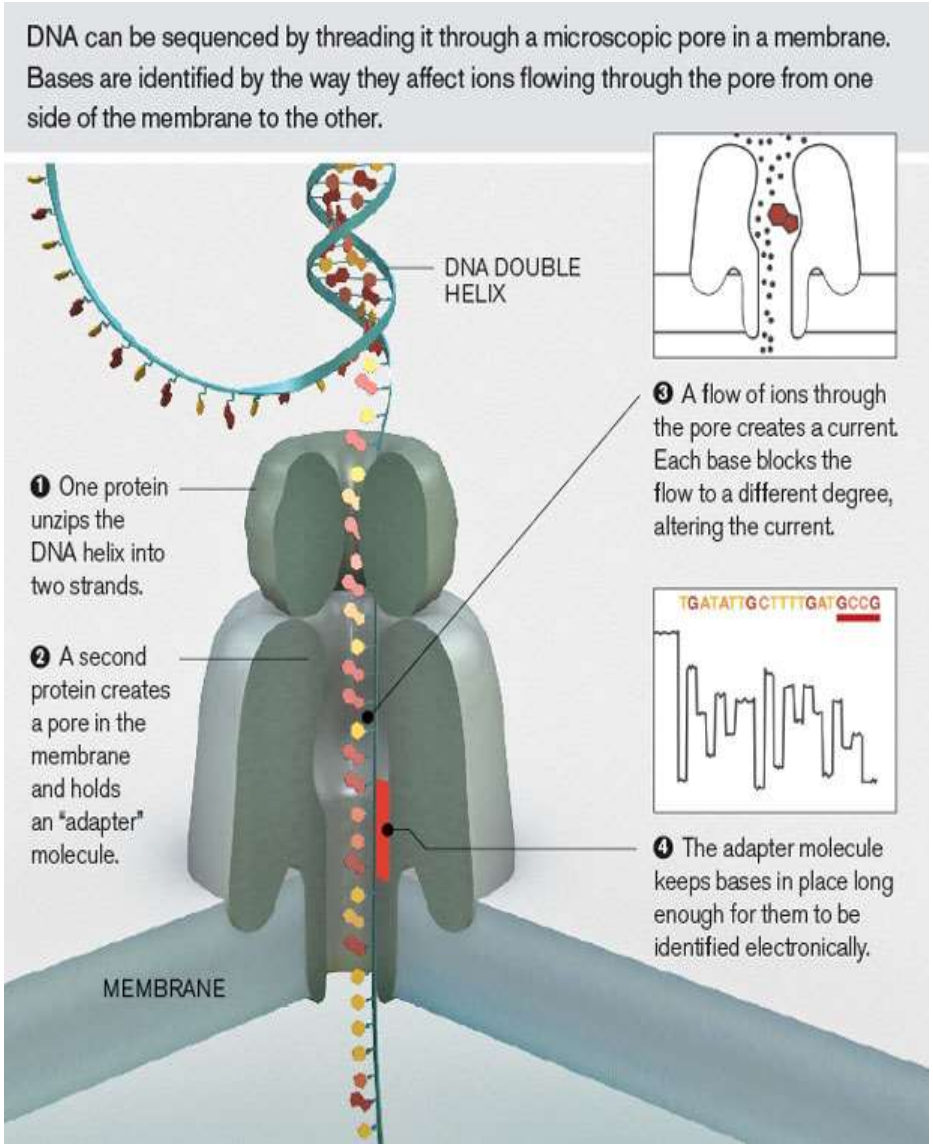
10X Genomics single-cell Drop-Seq



C_1 Single cell capture



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.



Future's so bright

Thank you!