Babelomics

NGS data Preprocessing

MDA – Valencia, March 2011

Javier Santoyo-Lopez
jsantoyo@cipf.es
http://bioinfo.cipf.es
Genomics Department
Centro de Investigacion Principe Felipe (CIPF)
(Valencia, Spain)

INB

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

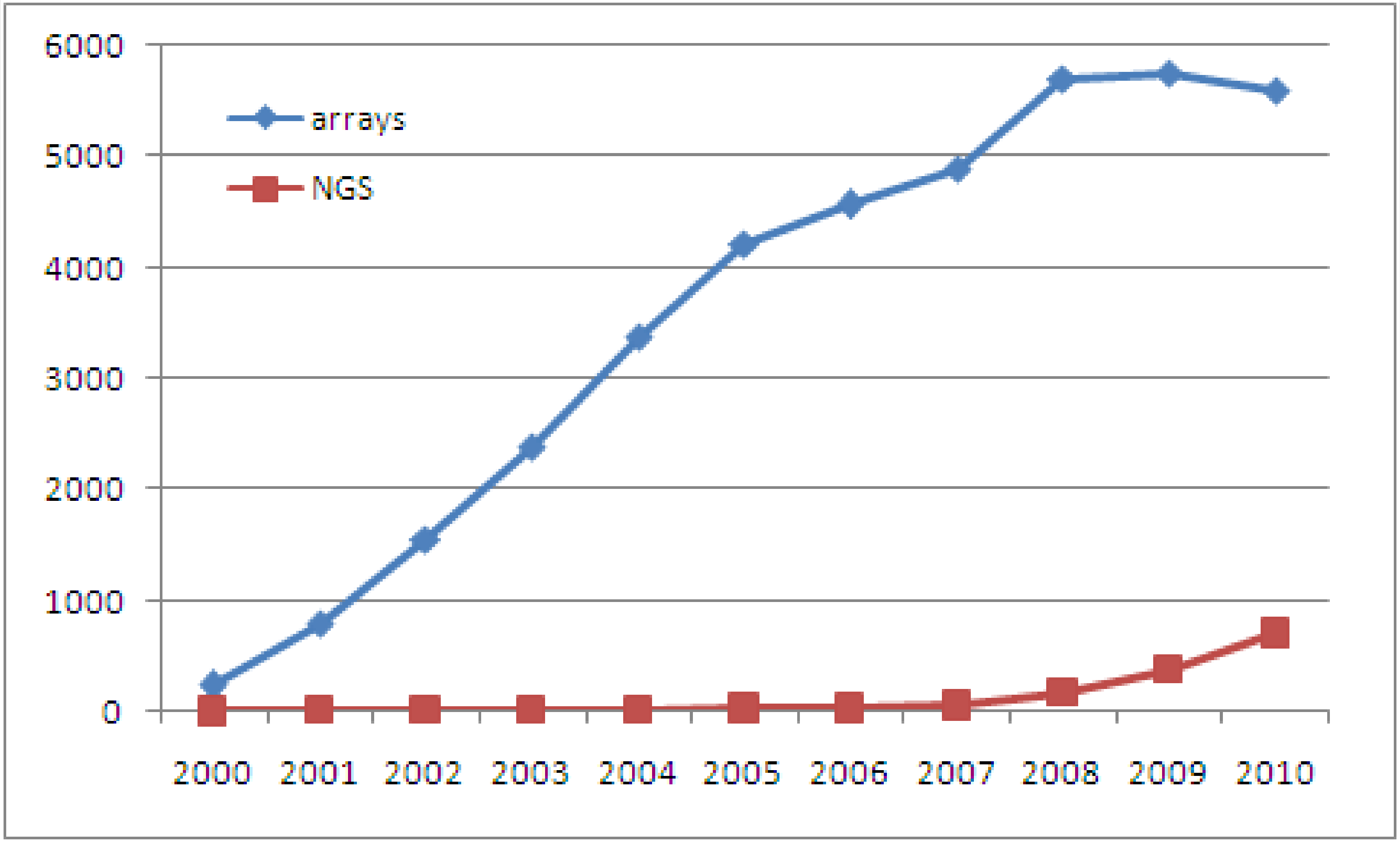
er ciberer
CENTRO DE INVESTIGACION BIOMEDICA EN RED
DE ENFERMEDADES RARAS

History of DNA Sequencing

Adapted from Eric Green, NIH; Adapted from Messing & Llaca, *PNAS* (1998)

Efficiency (bp/person/year)		
	1870	Miescher: Discovers DNA
	1940	Avery: Proposes DNA as ‘Genetic Material’
	1953	Watson & Crick: Double Helix Structure of DNA
1	1965	Holley: Sequences Yeast tRNA^{Ala}
15	1970	Wu: Sequences λ Cohesive End DNA
150		
1,500	1977	Sanger: Dideoxy Chain Termination Gilbert: Chemical Degradation
15,000	1980	Messing: M13 Cloning
25,000	1986	Hood et al.: Partial Automation
50,000	1990	Cycle Sequencing Improved Sequencing Enzymes Improved Fluorescent Detection Schemes
200,000		
50,000,000	2002	
100,000,000,000	2008	Next Generation Sequencing Improved enzymes and chemistry Improved image processing

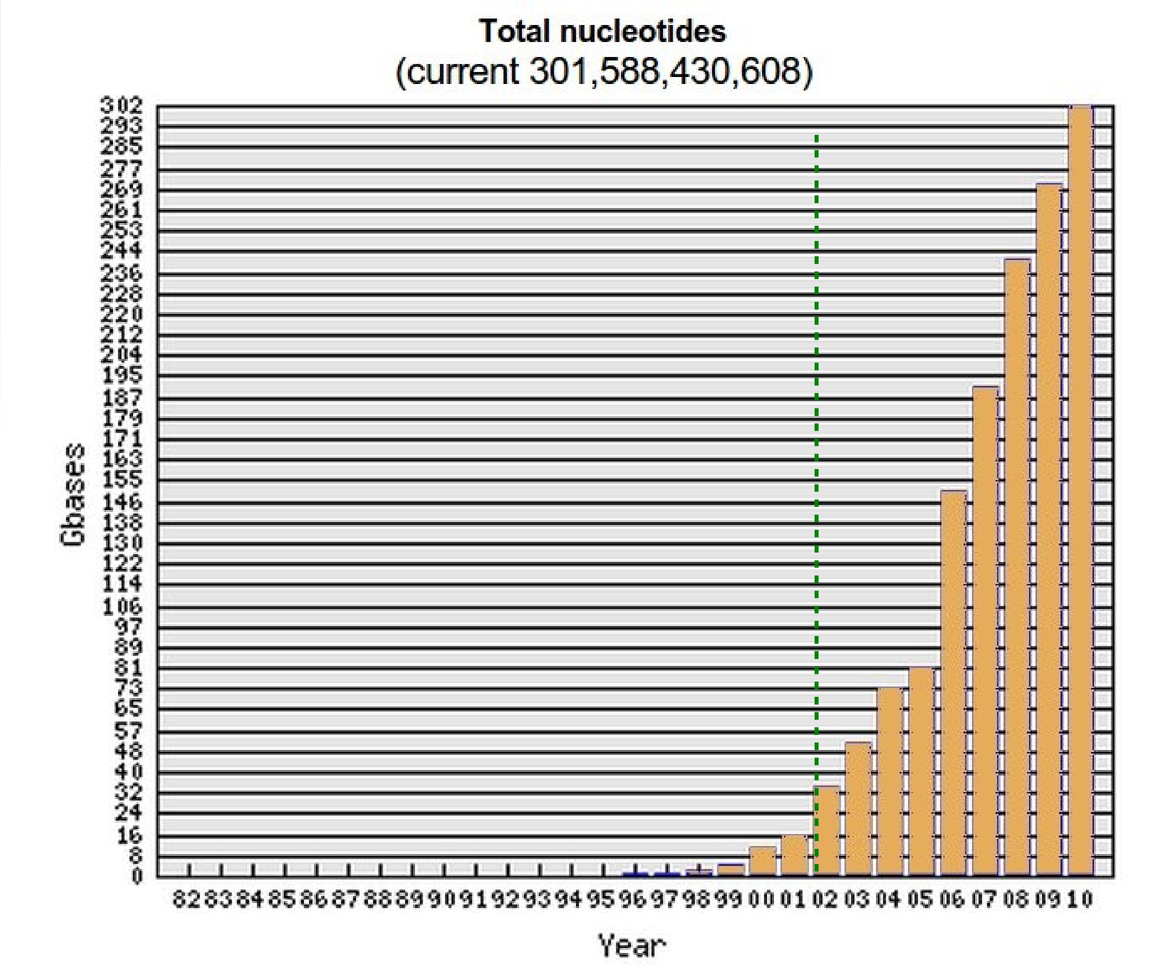
Trends in publications

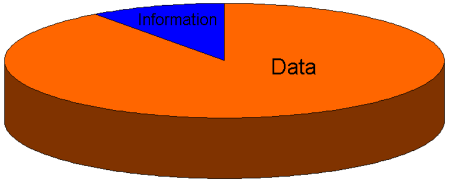


Source Pubmed. Query: "high-throughput sequencing"[Title/Abstract] OR "next generation sequencing"[Title/Abstract] OR "rna seq"[Title/Abstract]) AND year[Publication Date]

Sequence Databases Trend

EMBL database growth (March 2011)





From Michael Metzker, http://view.ncbi.nlm.nih.gov/pubmed/19997069

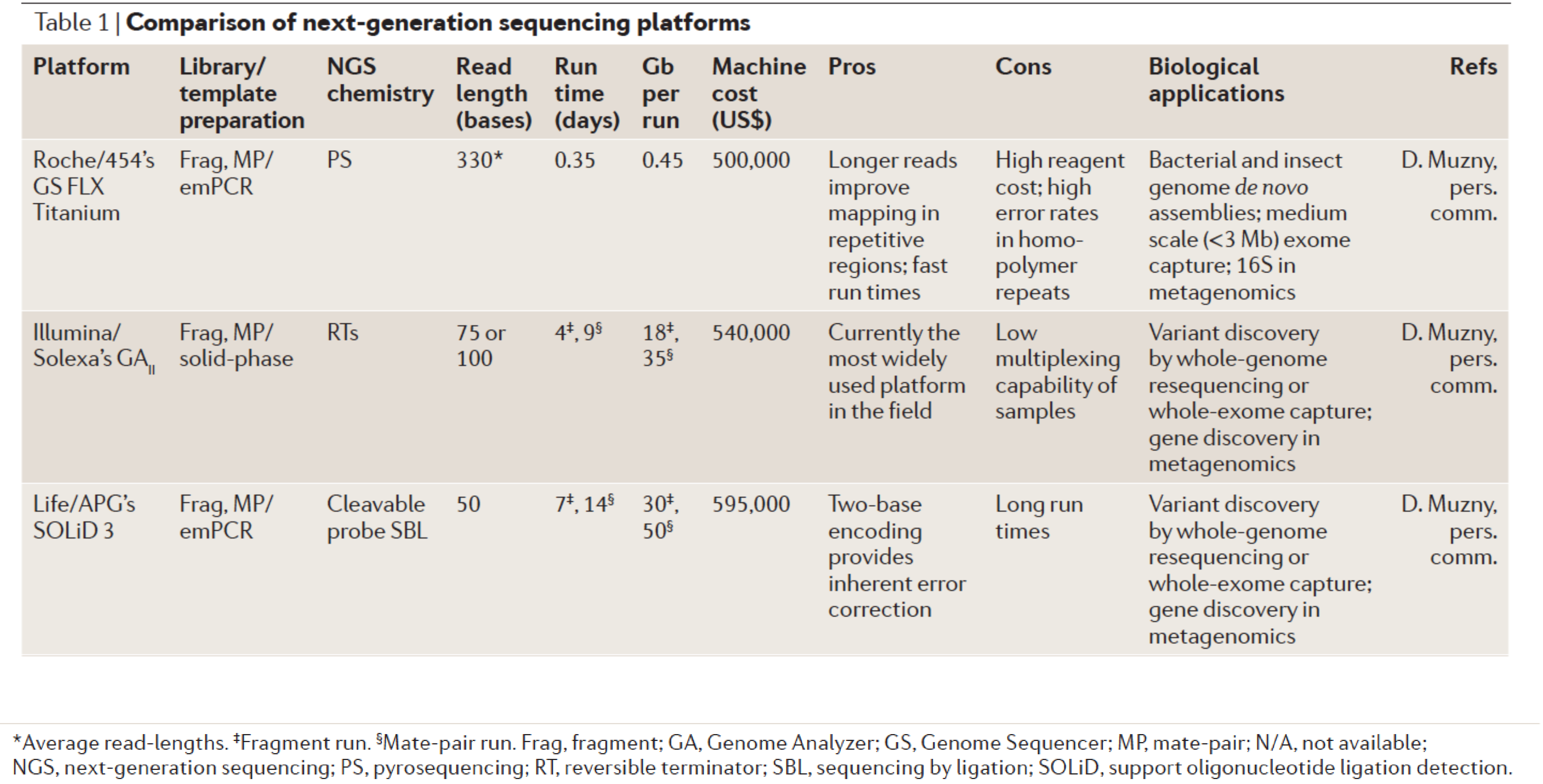
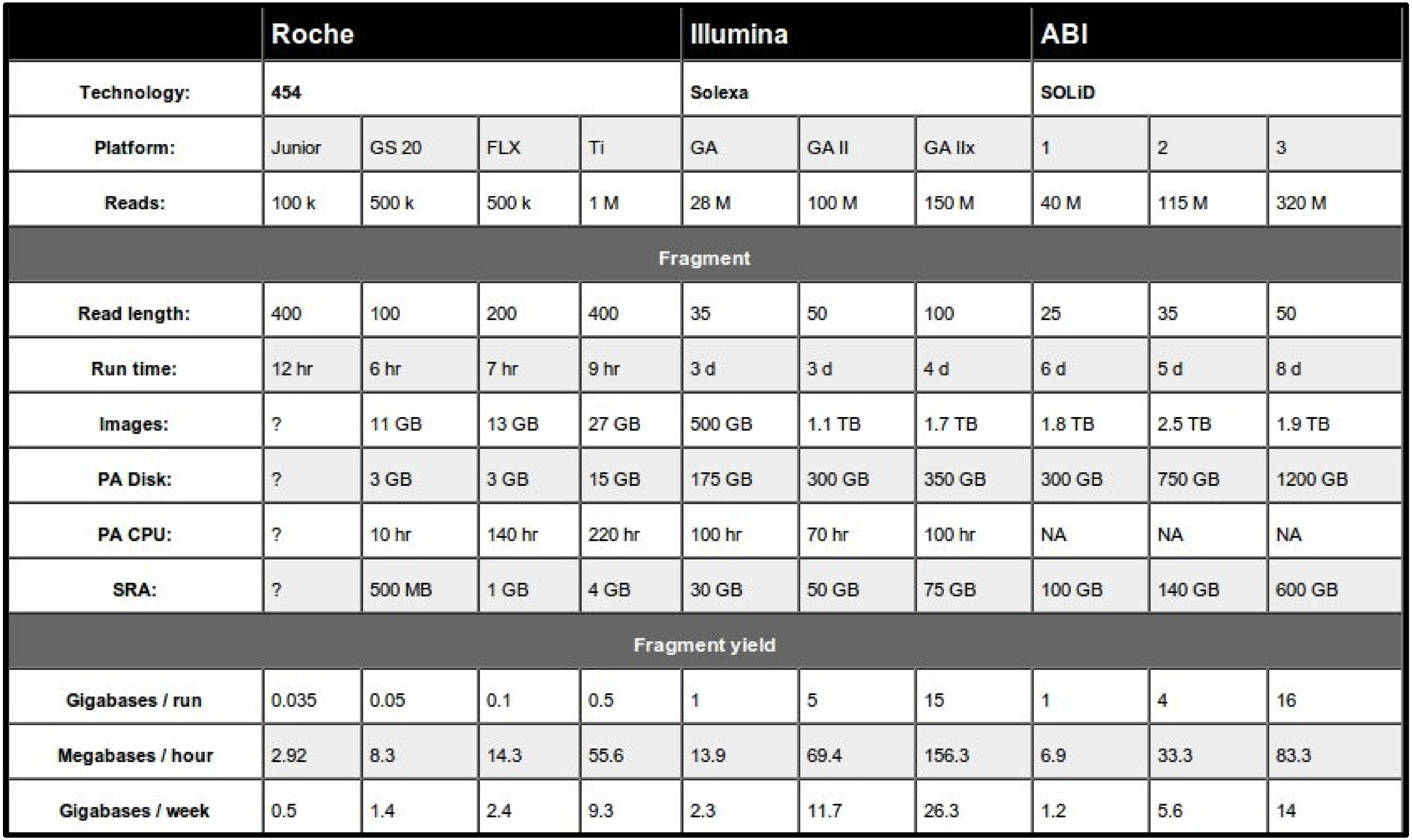


Table 1 | **Comparison of next-generation sequencing platforms**

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA_{II}	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.

*Average read-lengths. ‡Fragment run. §Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

NGS platforms comparison

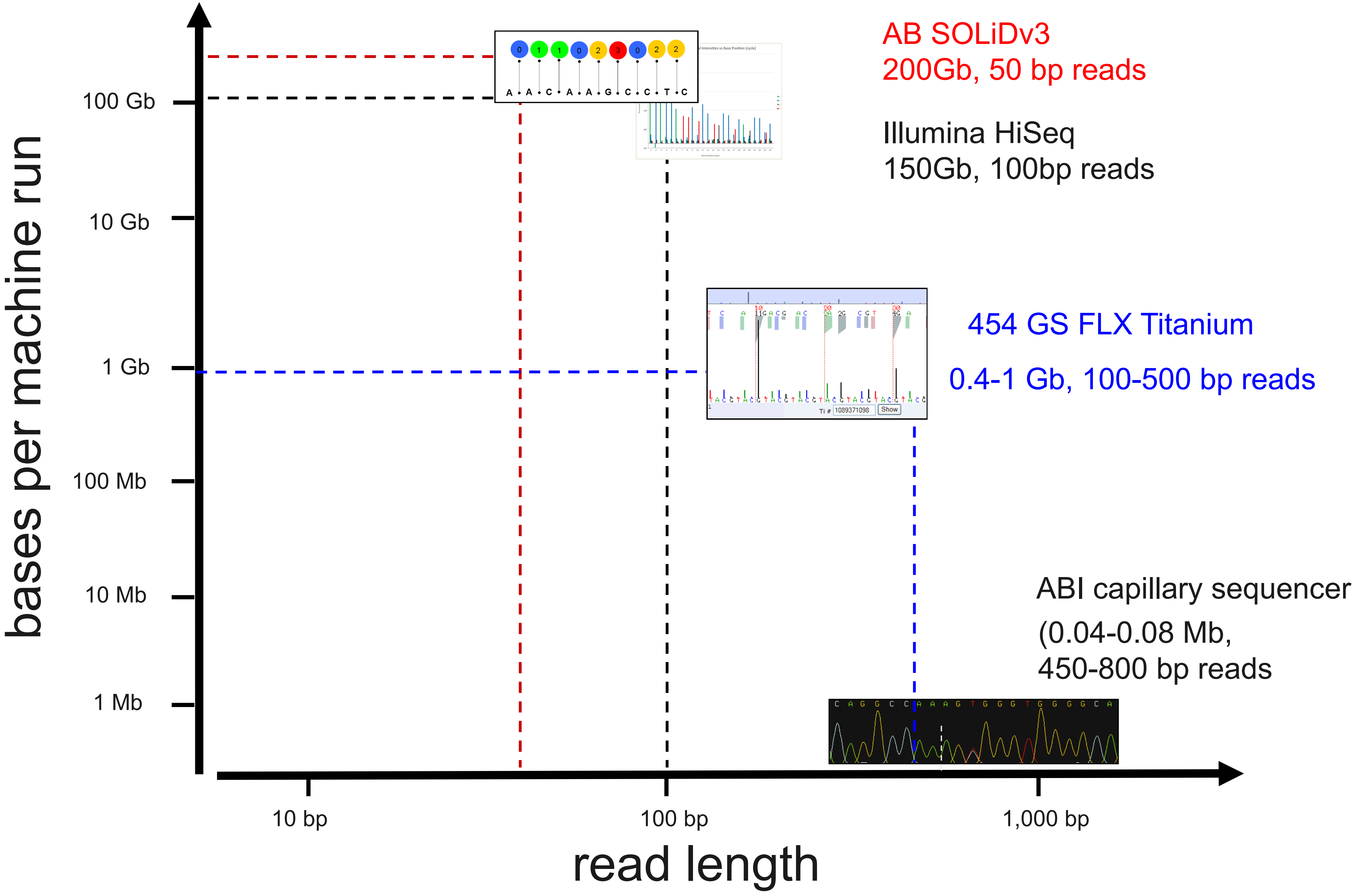


	Roche				Illumina			ABI		
Technology:	**454**				**Solexa**			**SOLiD**		
Platform:	Junior	GS 20	FLX	Ti	GA	GA II	GA IIx	1	2	3
Reads:	100 k	500 k	500 k	1 M	28 M	100 M	150 M	40 M	115 M	320 M
Fragment										
Read length:	400	100	200	400	35	50	100	25	35	50
Run time:	12 hr	6 hr	7 hr	9 hr	3 d	3 d	4 d	6 d	5 d	8 d
Images:	?	11 GB	13 GB	27 GB	500 GB	1.1 TB	1.7 TB	1.8 TB	2.5 TB	1.9 TB
PA Disk:	?	3 GB	3 GB	15 GB	175 GB	300 GB	350 GB	300 GB	750 GB	1200 GB
PA CPU:	?	10 hr	140 hr	220 hr	100 hr	70 hr	100 hr	NA	NA	NA
SRA:	?	500 MB	1 GB	4 GB	30 GB	50 GB	75 GB	100 GB	140 GB	600 GB
Fragment yield										
Gigabases / run	0.035	0.05	0.1	0.5	1	5	15	1	4	16
Megabases / hour	2.92	8.3	14.3	55.6	13.9	69.4	156.3	6.9	33.3	83.3
Gigabases / week	0.5	1.4	2.4	9.3	2.3	11.7	26.3	1.2	5.6	14

Source http://www.clcngs.com/2008/12/ngs-platforms-overview/

Next-gen sequencers

Adapted from John McPherson, OICR



Many Gbs of Sequences and...

- Data management becomes a challenge.
  - Moving data across file systems takes time (several hundred Gbs)

- What structure has the data?
  - Different sequencers output different files, but
  - There are some data formats that are being accepted widely (e.g. FastQ format)

- Raw sequence data formats
  - SFF
  - Fasta, csfasta
  - Qual file
  - Fastq

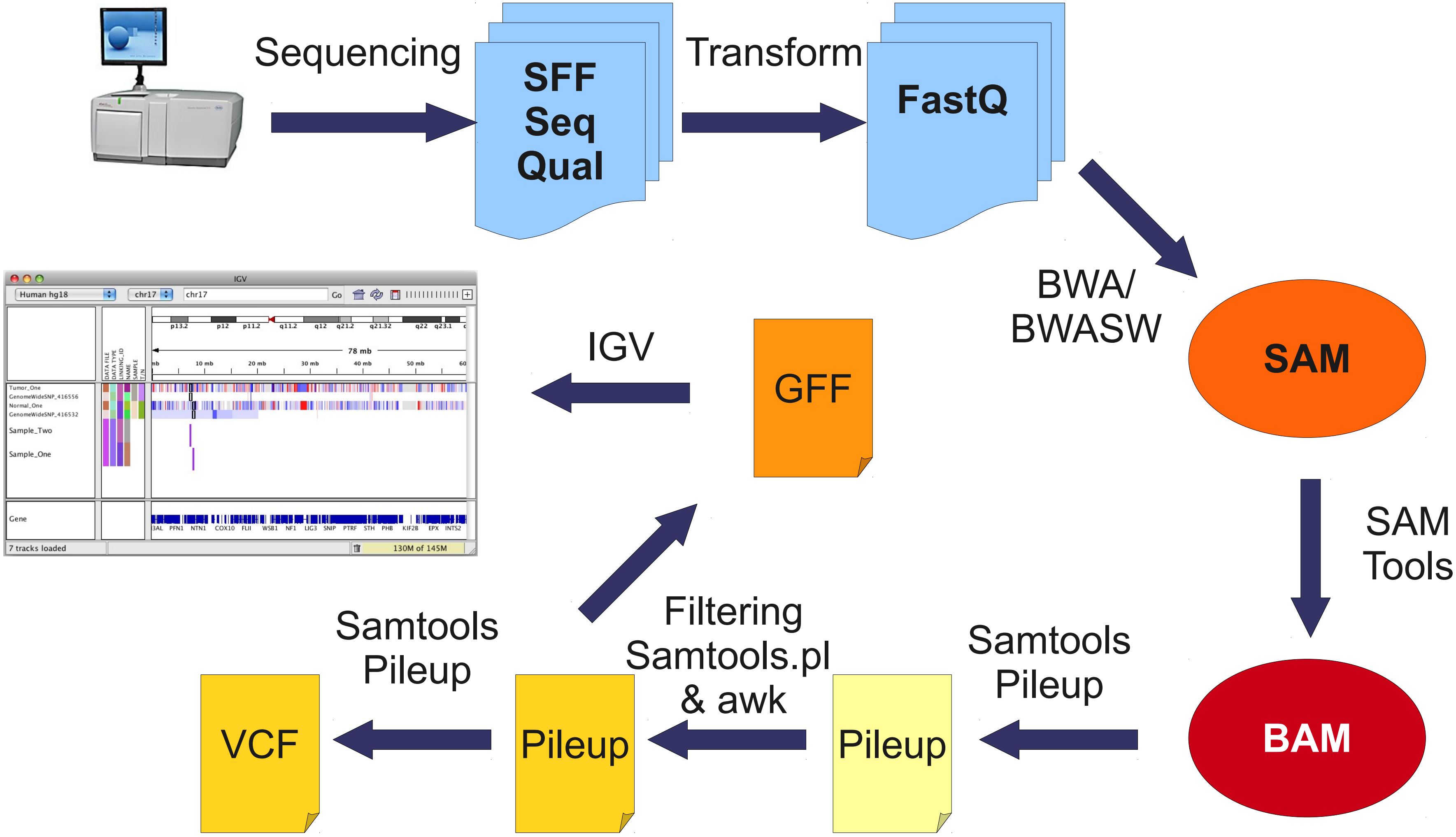
Fasta & Fastq formats

- **FastA** format (everybody knows about it)
  - Header line starts with “>” followed by a sequence ID
  - Sequence (string of nt).

- **FastQ** format
  - First is the sequence (like Fasta but starting with “@”)
  - Then “+” and sequence ID (optional) and in the following line are QVs encoded as single byte ASCII codes
    - Different quality encode variants

- Nearly all downstream analysis take **FastQ** as input sequence

Sequence to Variation Work flow



Sequencing

SFF
Seq
Qual

Transform

FastQ

BWA/
BWASW

SAM

SAM
Tools

BAM

Samtools
Pileup

Pileup

Filtering
Samtools.pl
& awk

Pileup

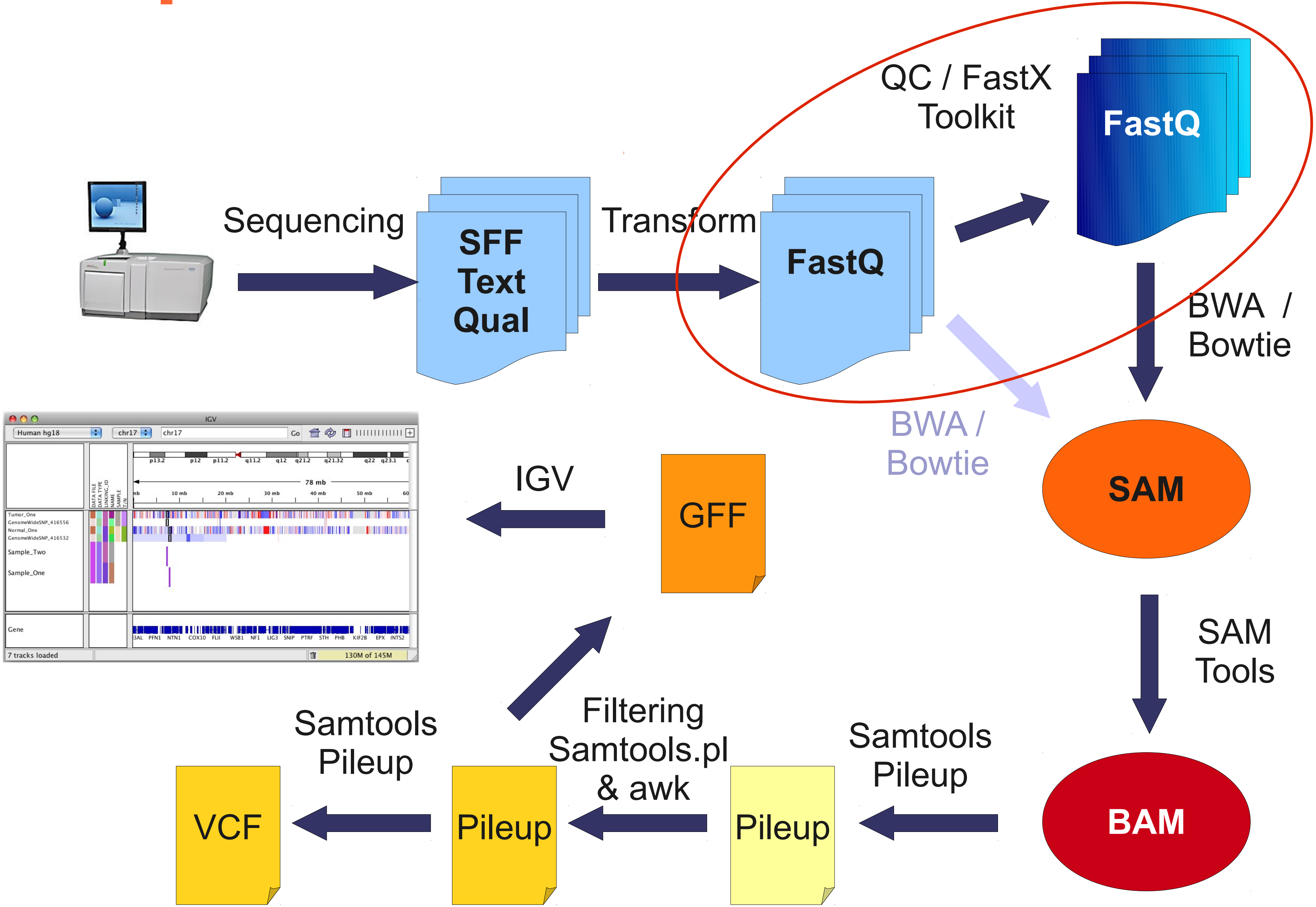
Samtools
Pileup

VCF

GFF

IGV

Sequence to Variation Workflow



Why Quality Control and Preprocessing?

- Sequencer output:
    - Reads + **quality**
    - **Is the quality of my sequenced data OK?**

Why Quality Control and Preprocessing?

- Sequencer output:
  - Reads + **quality**
  - **Is the quality of my sequenced data OK?**

If something is wrong can I fix it?

Why Quality Control and Preprocessing?

- Sequencer output:
  - Reads + **quality**
  - **Is the quality of my sequenced data OK?**

    If something is wrong can I fix it?
- **Problem**:
  - **HUGE** files...

Why Quality Control and Preprocessing?

- Sequencer output:
  - Reads + **quality**
  - **Is the quality of my sequenced data OK?**

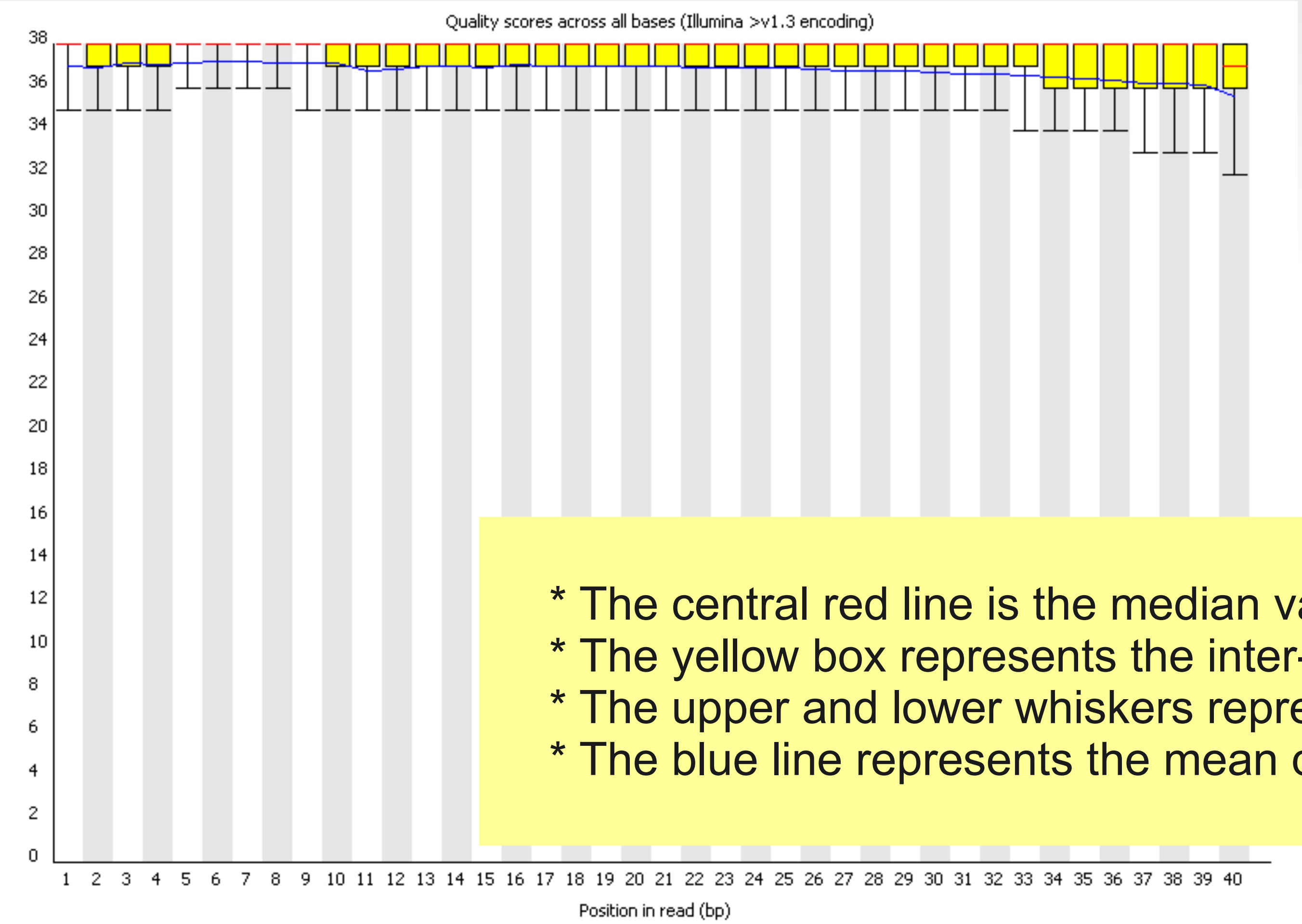
    If something is wrong can I fix it?
- **Problem**:
  - **HUGE** files... How do they look?

```
@HWUSI-EAS460:2:1:368:1089#0/1
TACGTACGTACGTACGTACGTAGATCGGAAGAGCGG
+HWUSI-EAS460:2:1:368:1089#0/1
aa[a_a_a^a^a]VZ]R^P[]YNSUTZBBBBBBBBB

@HWUSI-EAS460:2:1:368:528#0/1
CTATTATAATATGACCGACCAGCTAGATCTACAGTC
+HWUSI-EAS460:2:1:368:528#0/1
abbbbaaaabba^aa`Y``aa`aaa``a`a_\_`[_
```

- Files are flat files and are big... tens of Gbs (please... **don't use MS word** to see or edit them)

Sequence Quality Per base Position

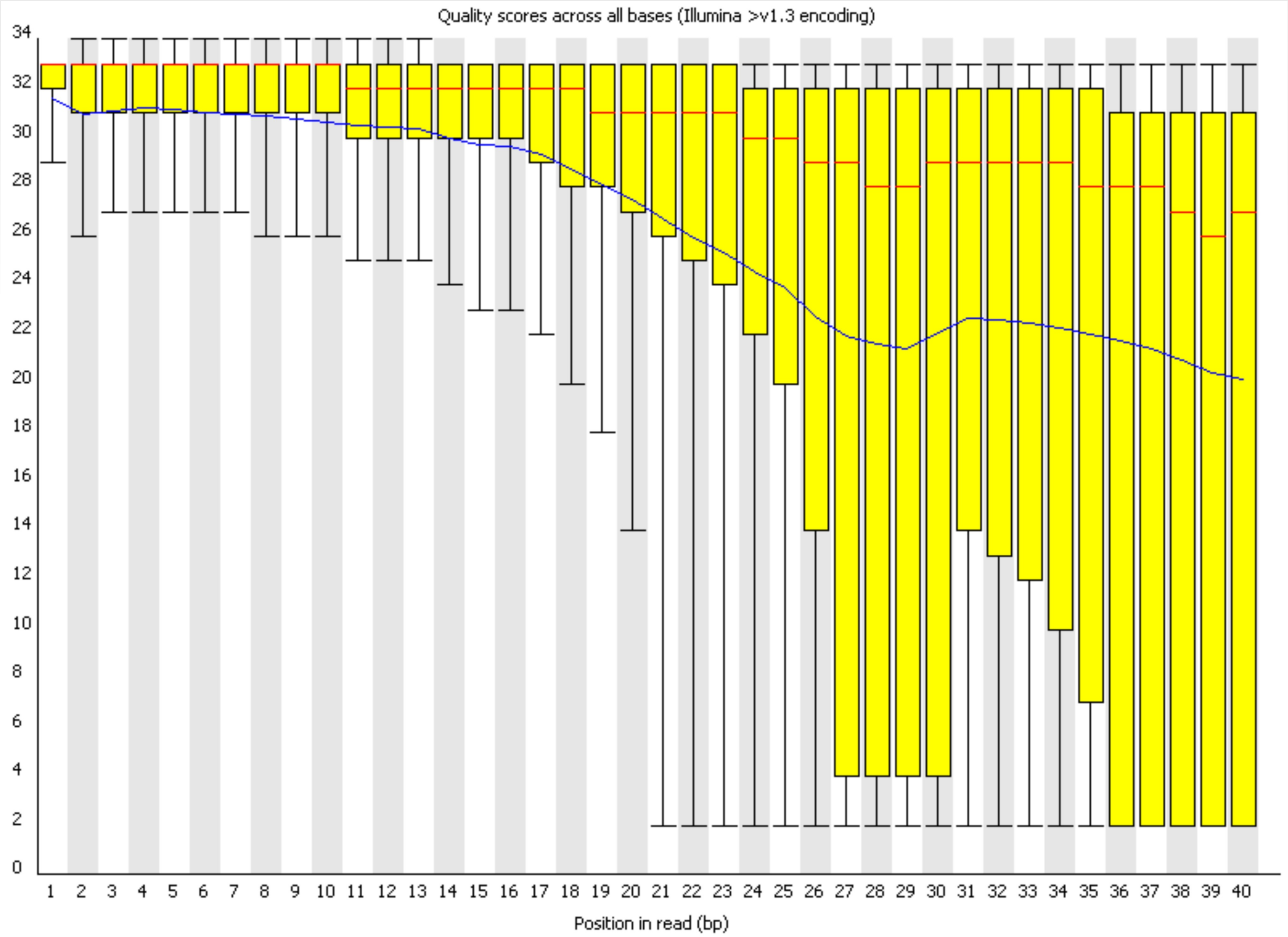


Good data

- Consistent
- High quality along the read

* The central red line is the median value
* The yellow box represents the inter-quartile range (25-75%)
* The upper and lower whiskers represent the 10% and 90% points
* The blue line represents the mean quality

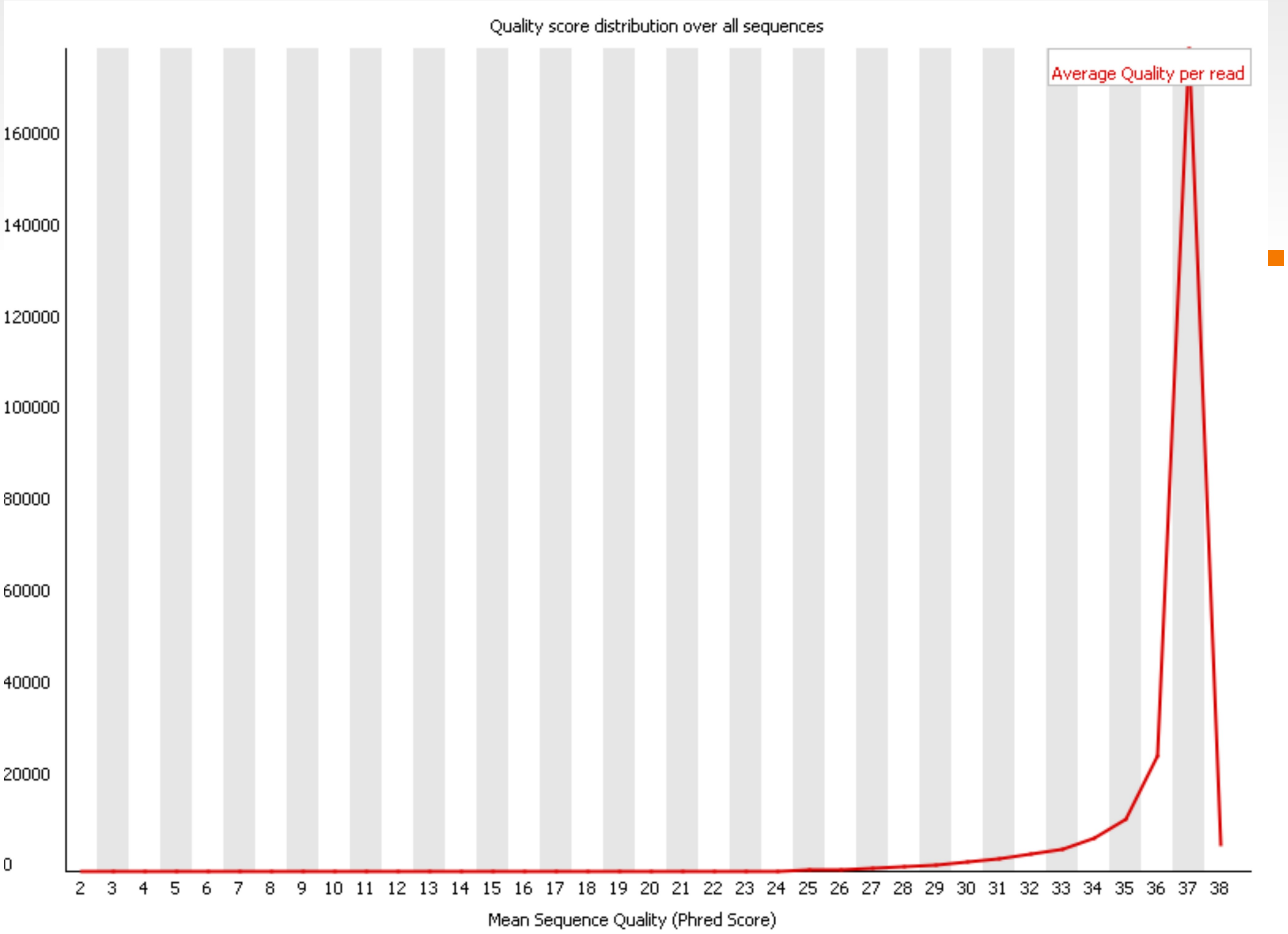
Sequence Quality Per base Position



Bad data

- High variance
- Quality decrease with length

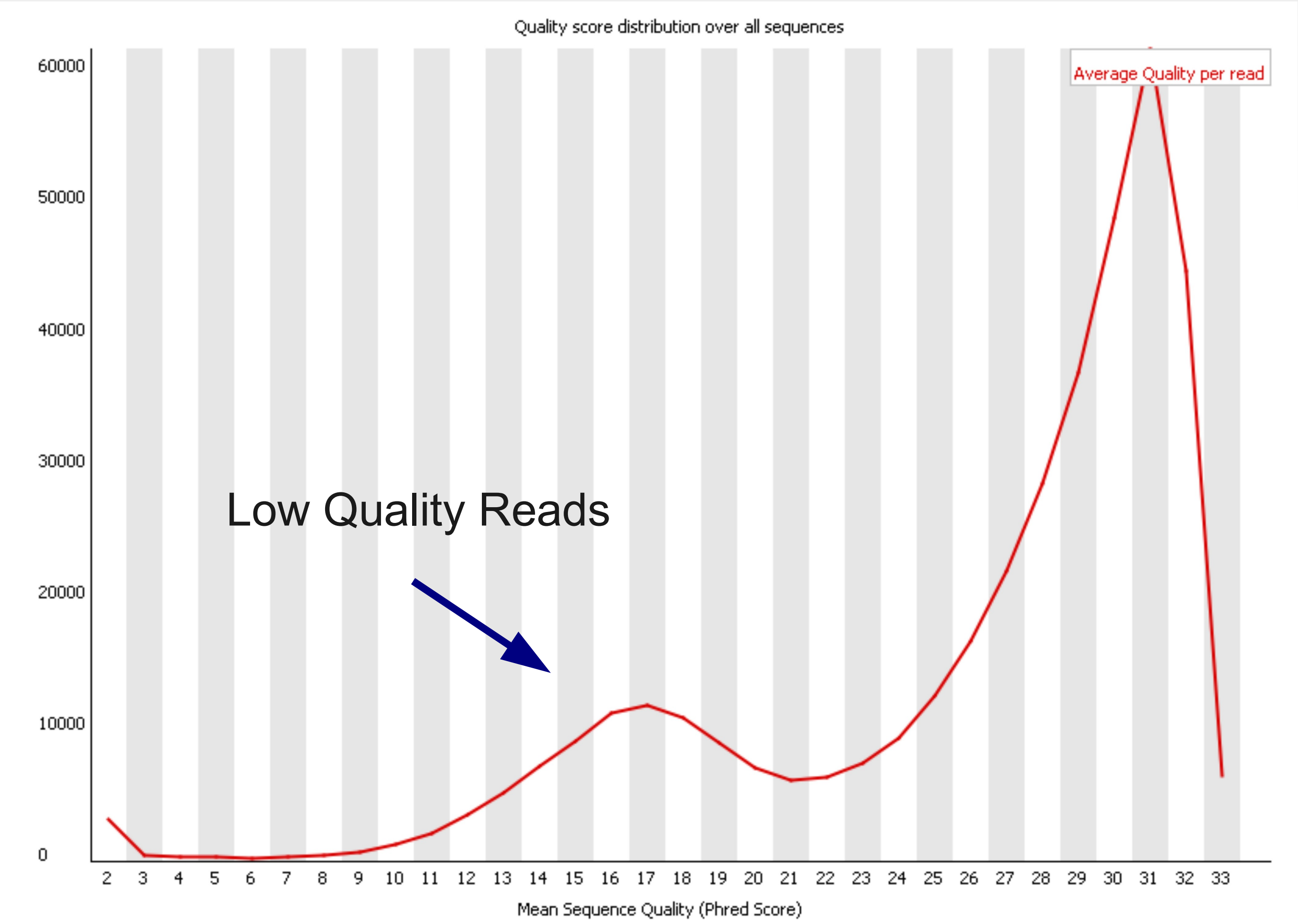
Per Sequence Quality Distribution



Good data

- Most are high-quality sequences

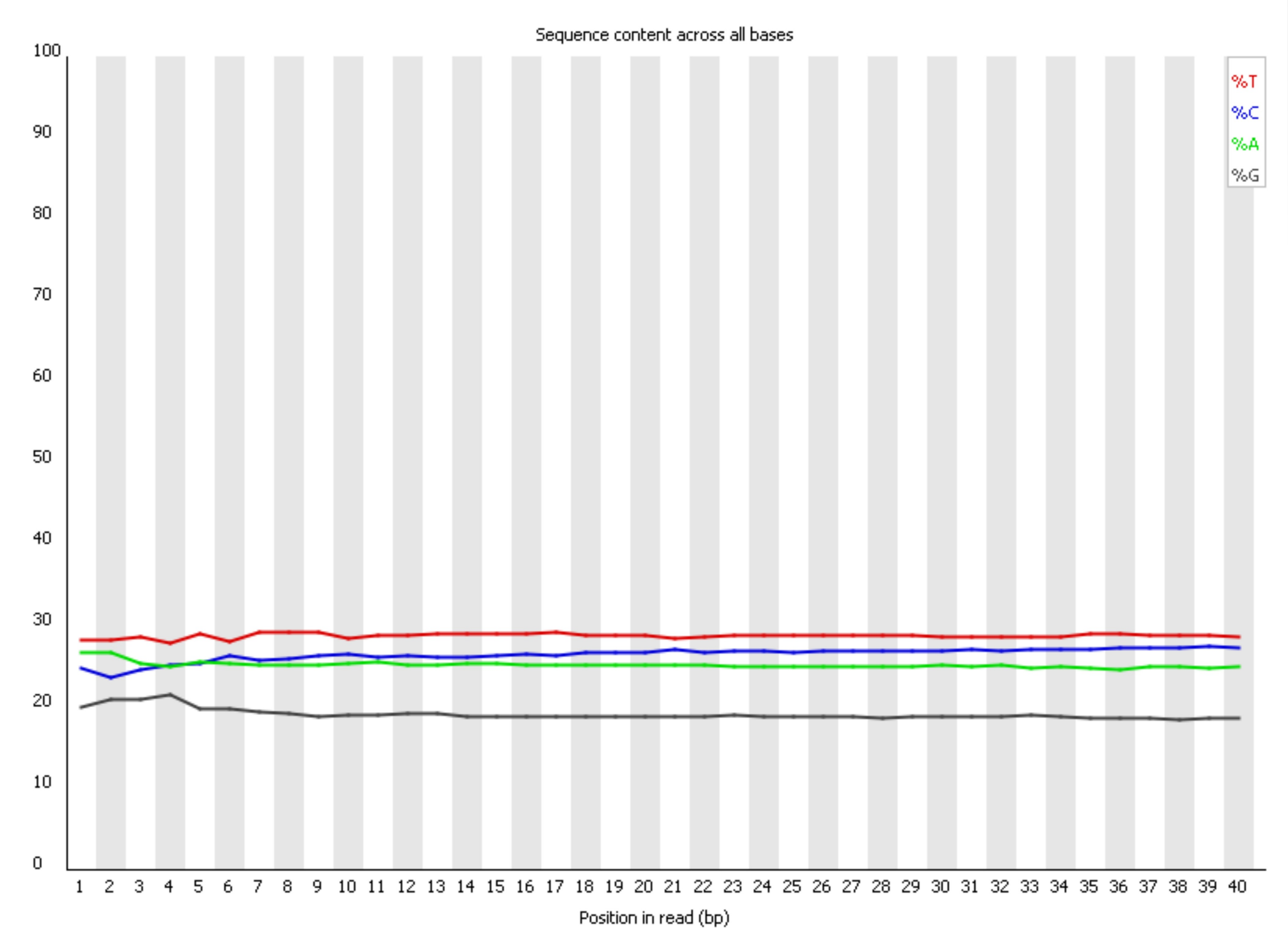
Per Sequence Quality Distribution



Bad data

- Not uniform distribution

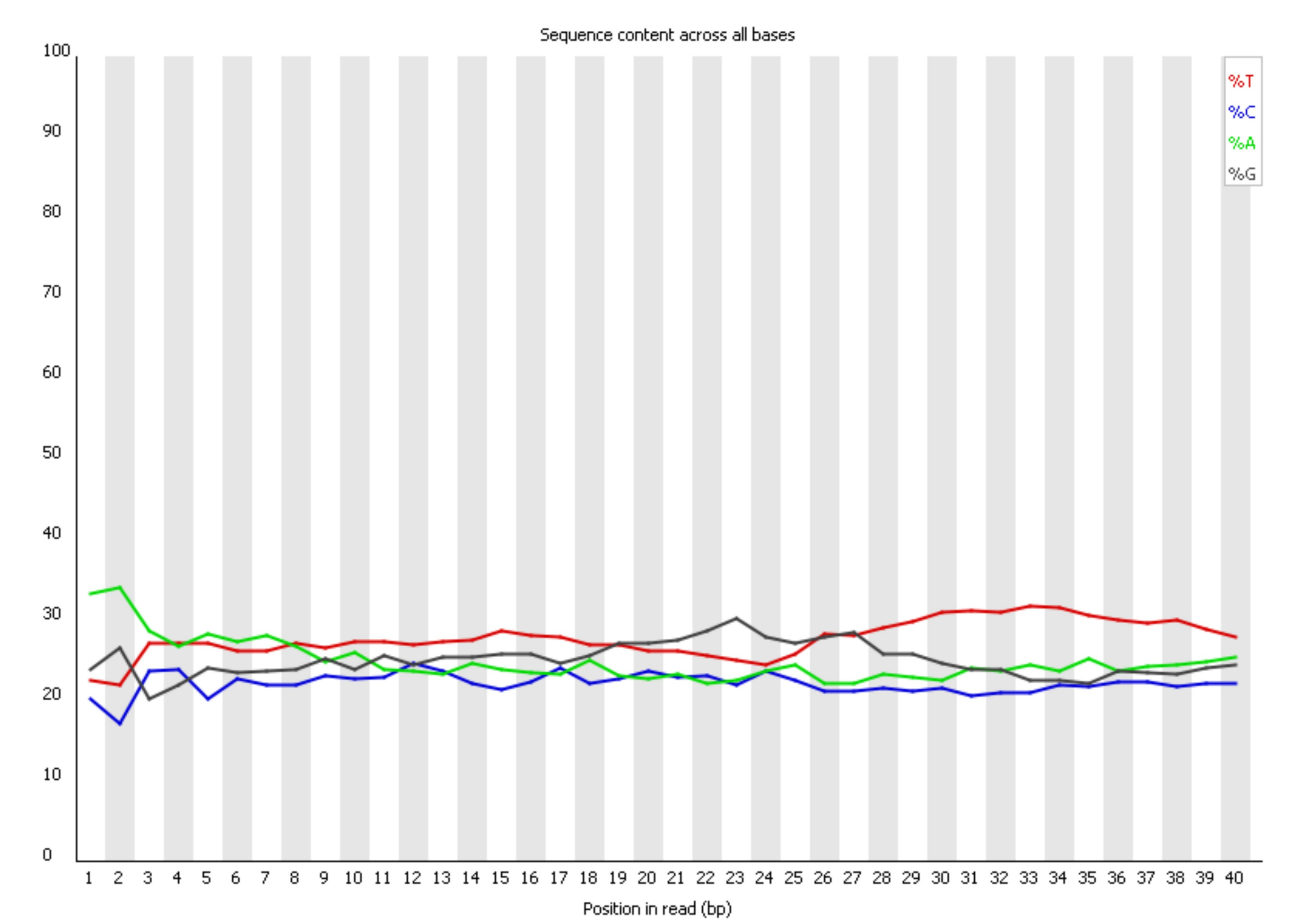
Nucleotide Content per position



Good data

- Smooth over length
- Organism dependent (GC)

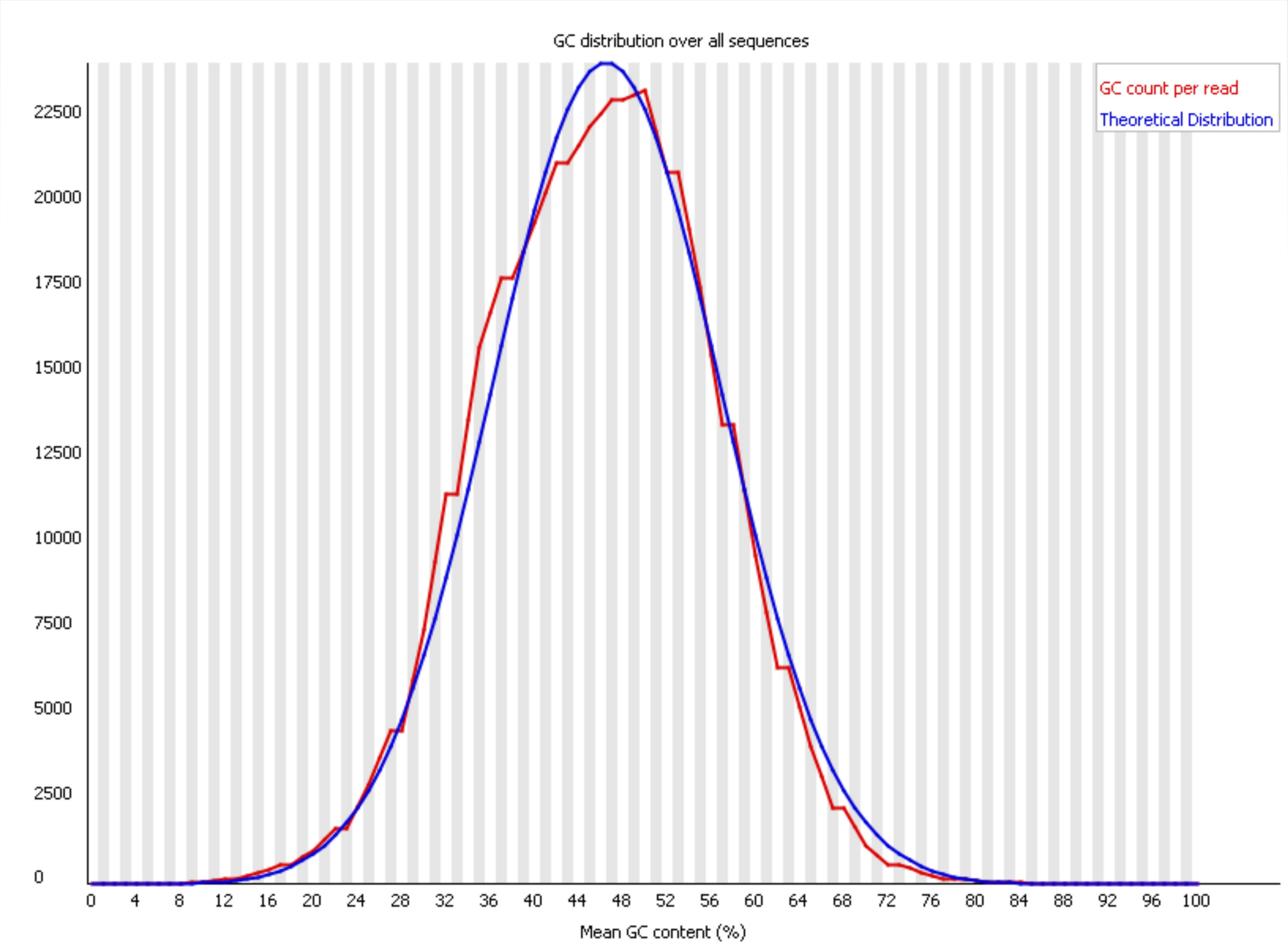
Nucleotide Content per position



Bad data

- Sequence position bias

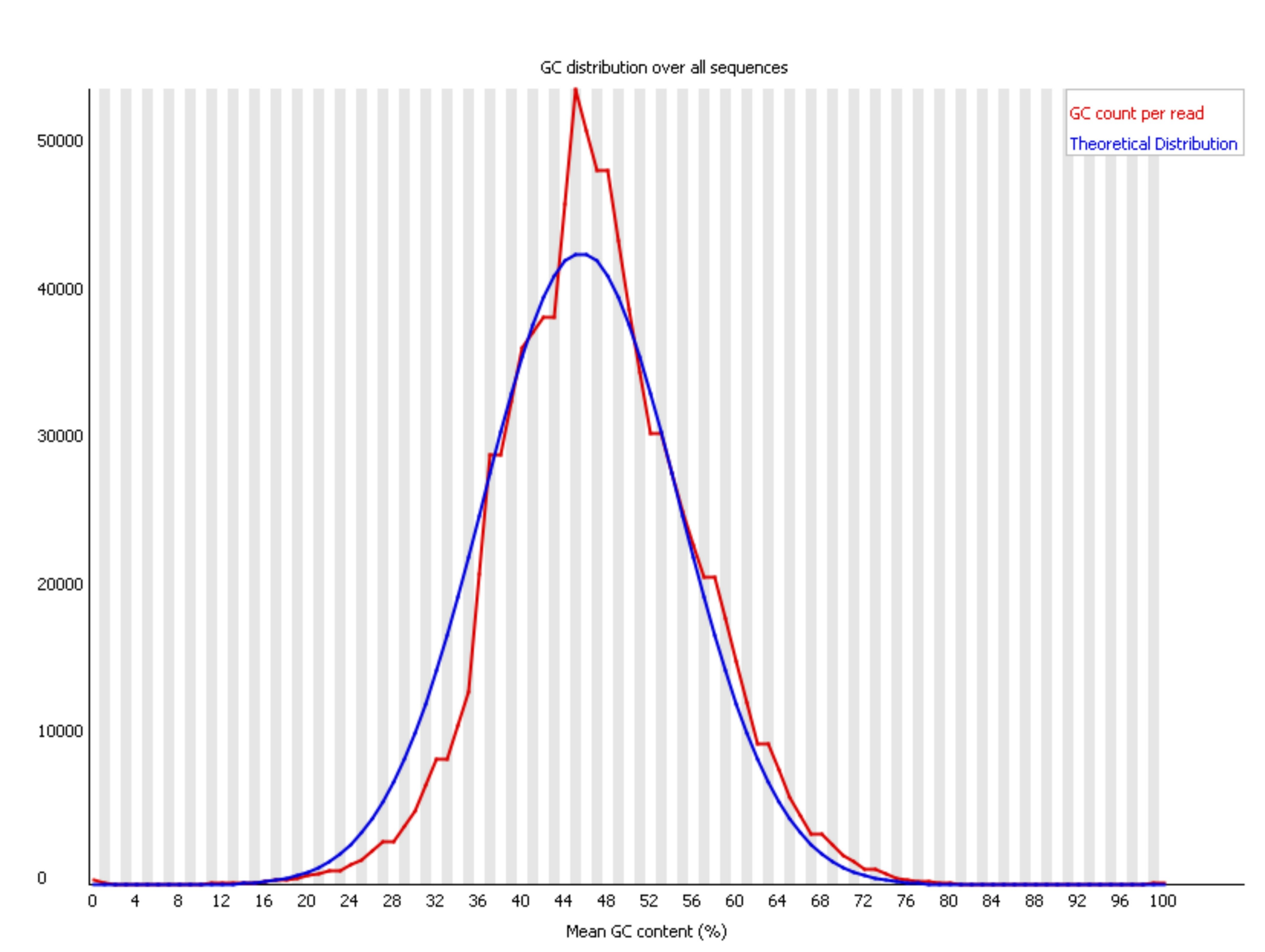
GC Distribution



Good data

- Fits with the expected
- Organism dependent

Per sequence
GC Distribution

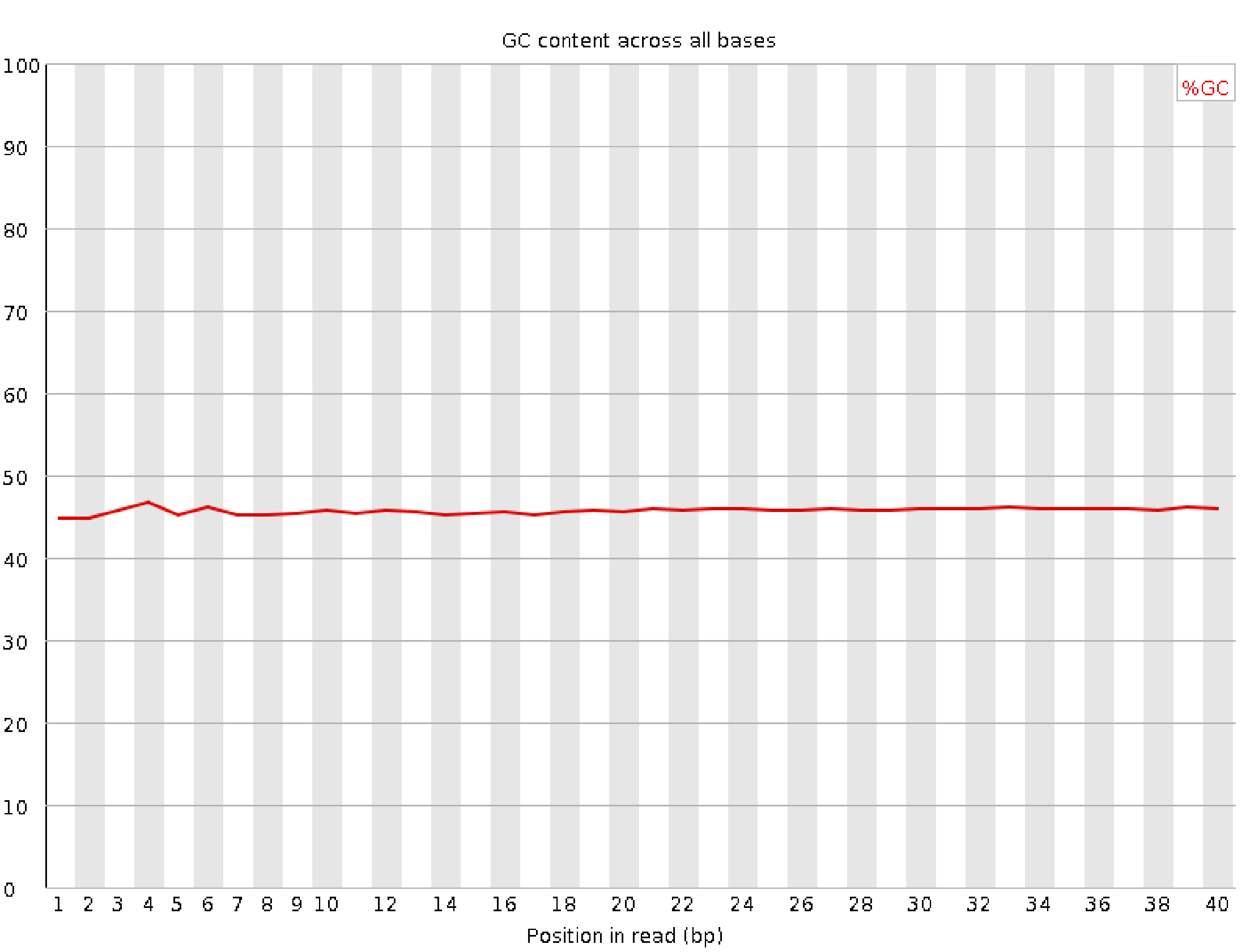


Bad data

- It does not fit with expected
- Organism dependent

Library contamination?

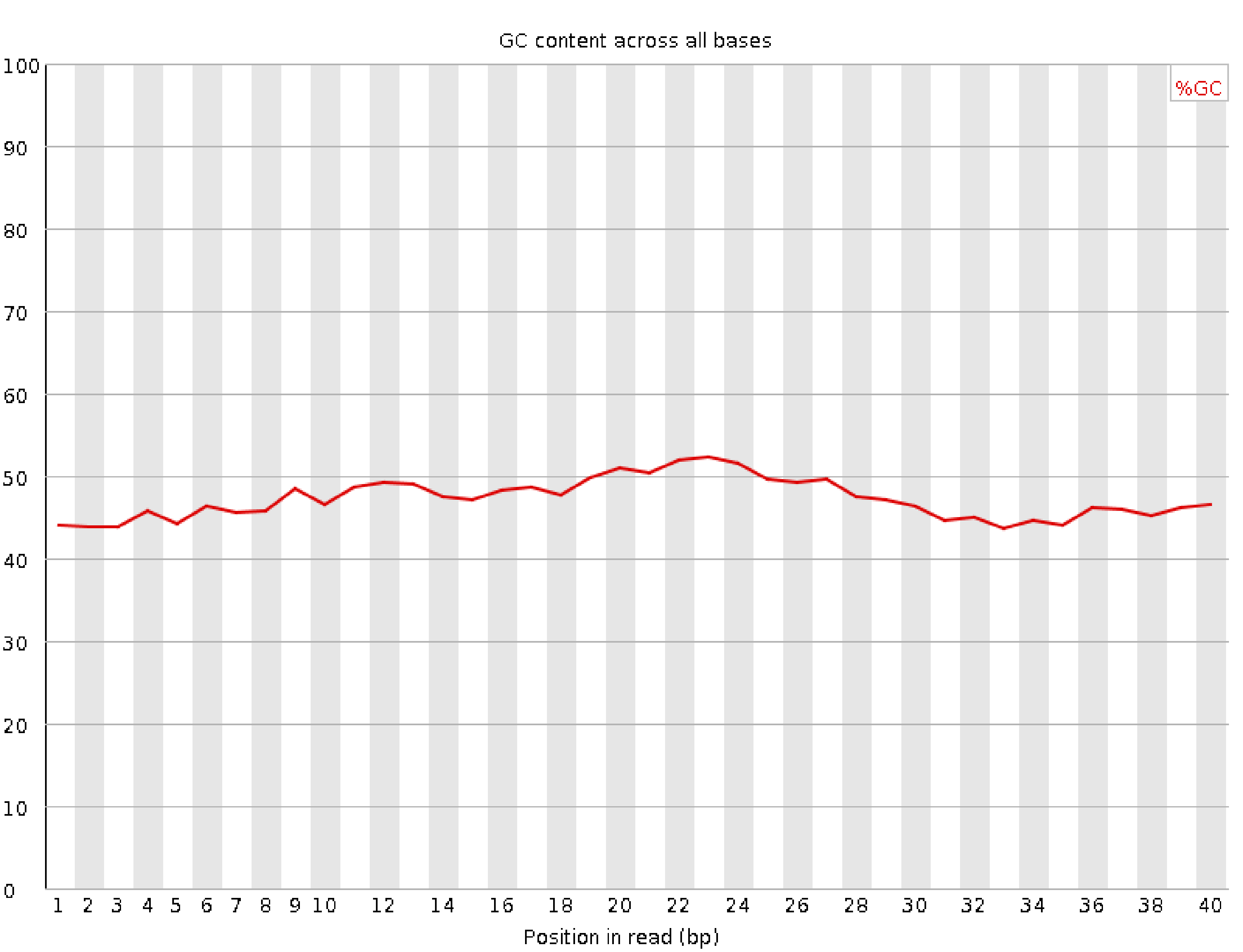
Per base GC Distribution



Good data

- No variation across read sequence

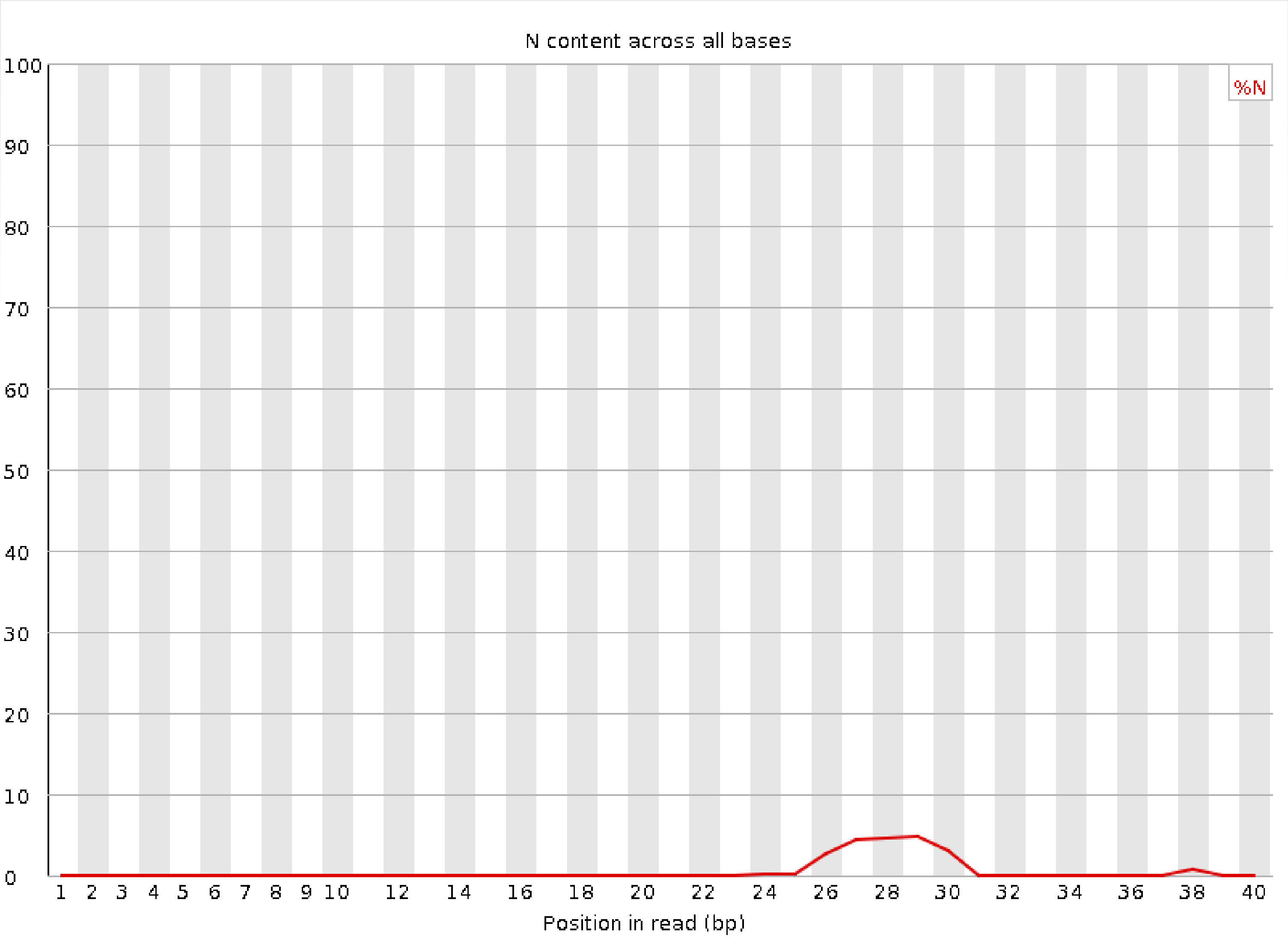
Per base GC Distribution



Bad data

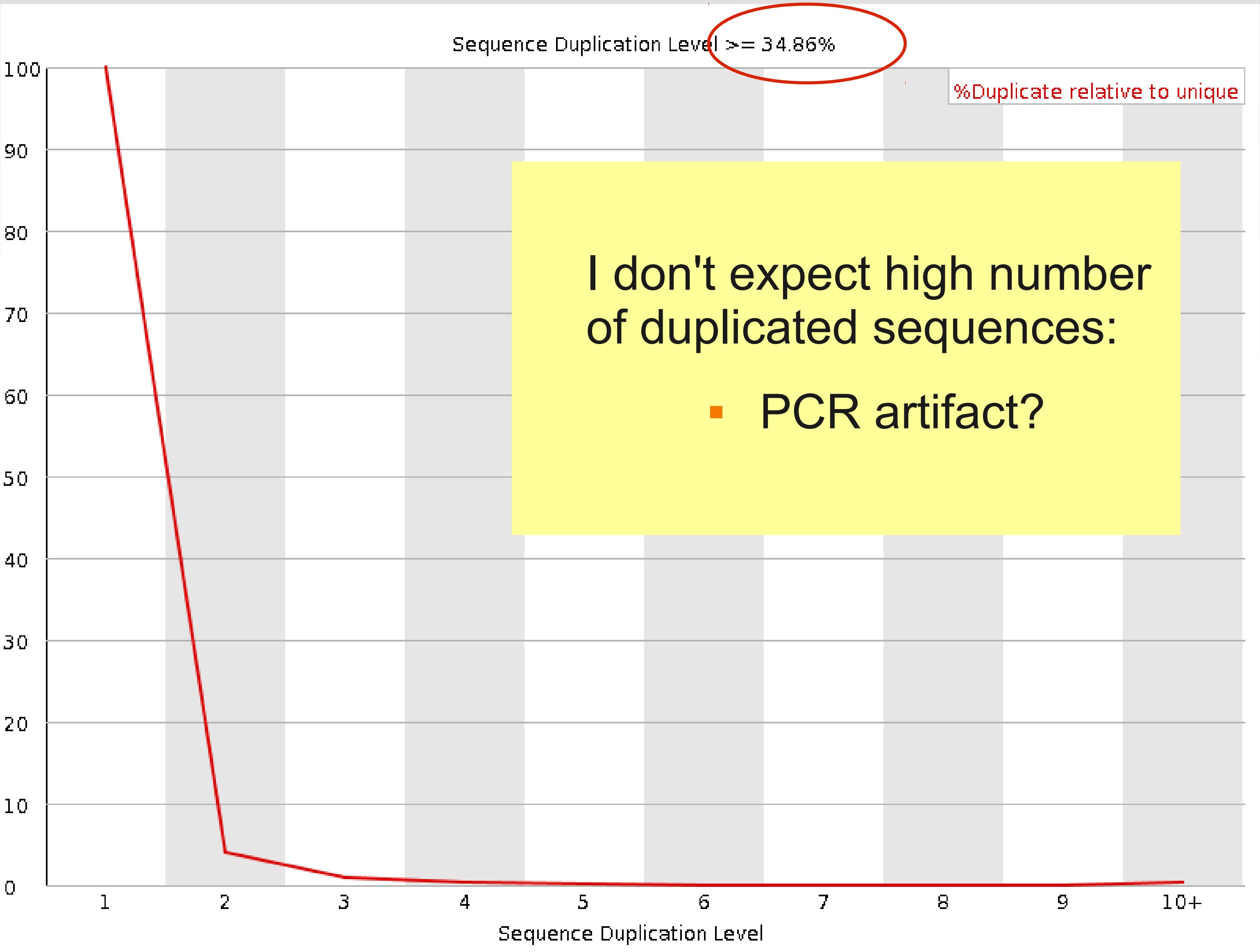
- Variation across read sequence

Per base N content



It's not good if there are N bias per base position

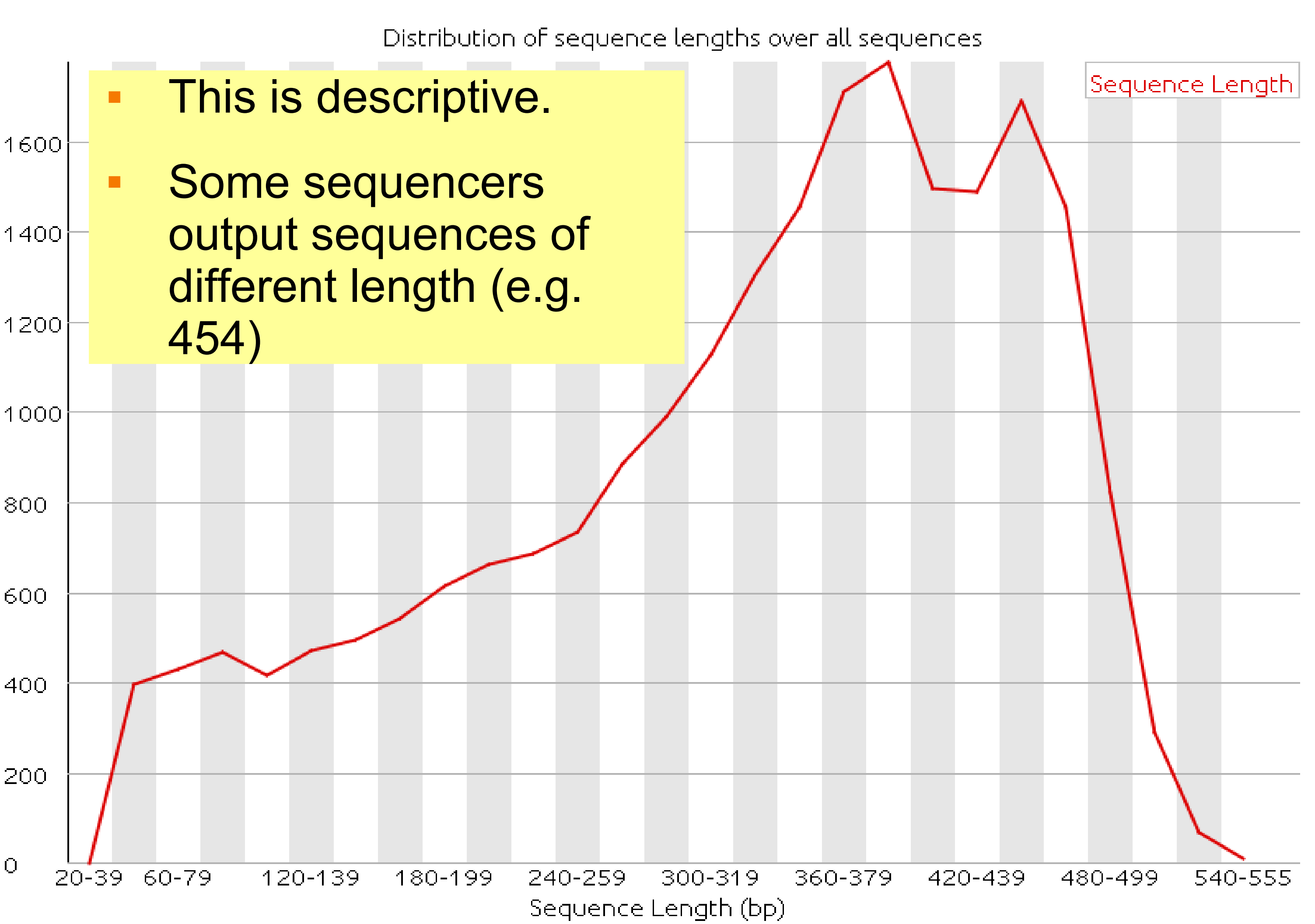
Duplicated Sequences



I don't expect high number of duplicated sequences:

- PCR artifact?

Distribution Length



- This is descriptive.
- Some sequencers output sequences of different length (e.g. 454)

Overrepresented Sequences

Question:

If you obtain the exact same sequence too many times → Do you have a problem?

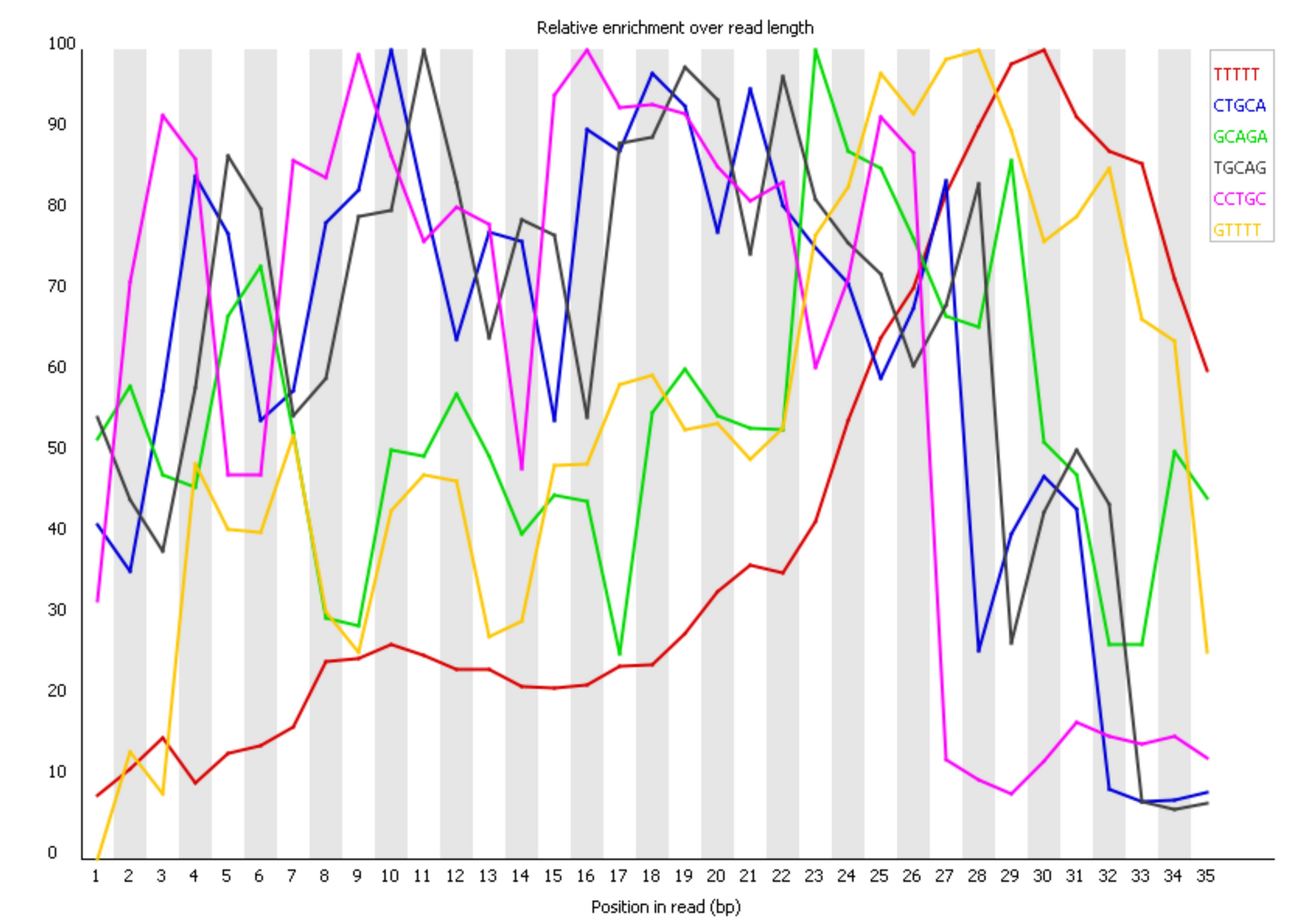
Answer:

Sometimes!

Examples → PCR primers (Illumina)

- GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT
- CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC

K-mer Content



- Helps to detect problems
- Adapters?

Practical:
FastQC and Fastx-toolkit

- Use **FastQC** to see your starting state.
- Use **Fastx-toolkit** to optimize different datasets and then visualize the result with FastQC to prove your success!

Hints: Try trimming, clipping and quality filtering.

Go to the tutorial and try the exercises...