Illumina Reads QC

Marco van Zwetselaar

Kilimanjaro Clinical Research Institute

Sources for Illumina Reads QC

- Basic summaries in:
  - MiSeq Control Software (MCS) primary analysis: run-time statistics and predictions
  - MiSeq Reporter after secondary analysis: summaries of clusters, phasing, errors across the run

- Sequencing Analysis Viewer (SAV): very detailed (per lane, tile, cycle) in real time and post-run
  - Ideal for troubleshooting

- FastQC: open source tool, analyses reads files

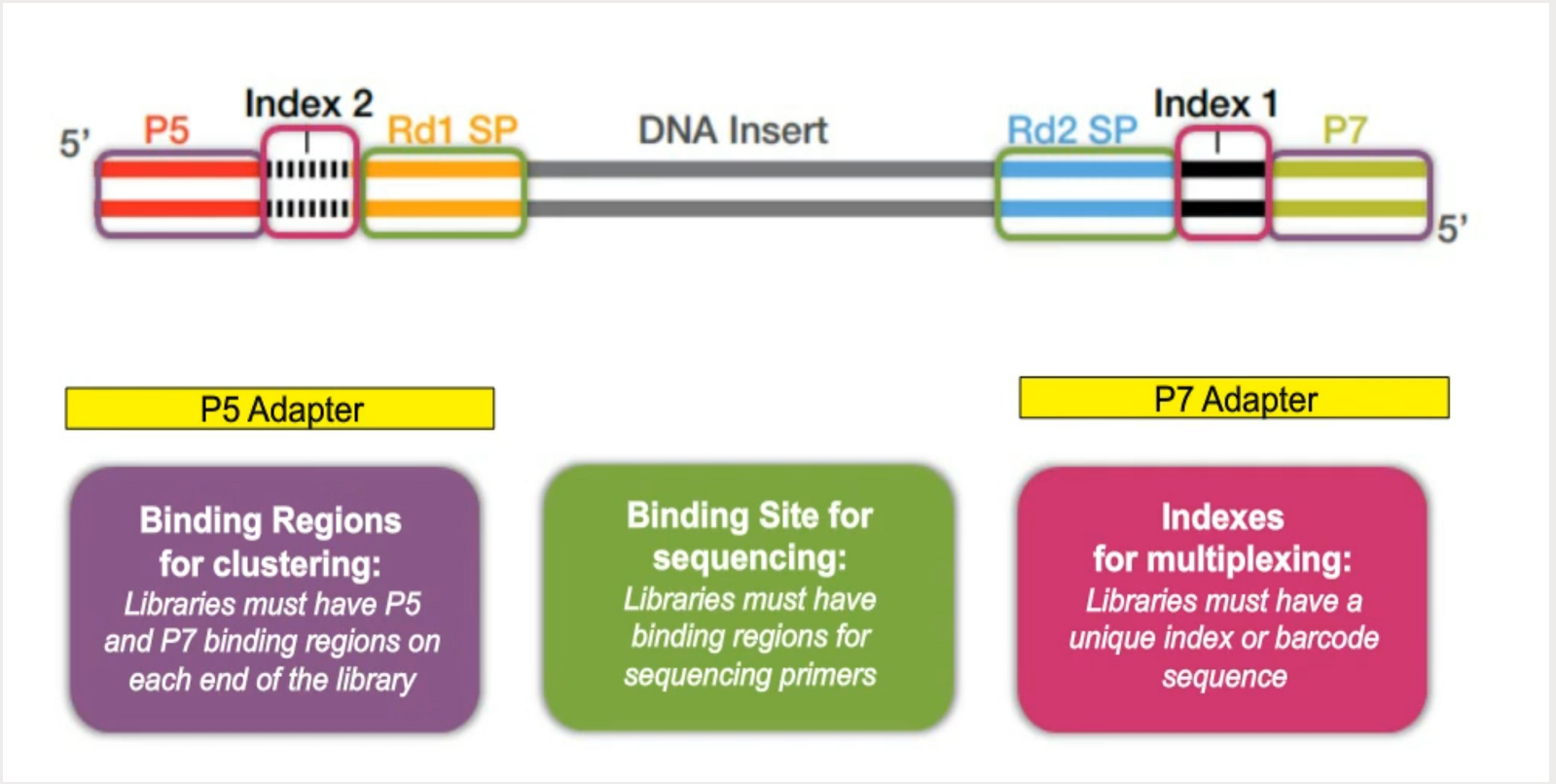
Phred Scale Q score

- $Q = -10 \cdot \log_{10} P_{incorrect}$ $\Rightarrow$ $P_{incorrect} = 1 / 10^{Q/10}$
- “10 Q more is 10 times better” (as in: 10 times fewer errors)
- Q-score divided by 10 is “number of nines”

Phred Score (Q)	Probability of incorrect call	Probability of incorrect call	Base call accuracy	Characters (Phred+33)
0	$1/10^0$	1 (100%)	-	!"#$%&'()*
10	$1/10^1$	0.1 (10%)	90%	+,-./01234
13	$1/10^{1.3}$ (~1/20)	0.05 (5%)	95%	.
17	$1/10^{1.7}$ (~1/50)	0.02 (2%)	98%	2
20	$1/10^2$	0.01 (1%)	99%	56789:;<=>
30	$1/10^3$	0.001	99.9%	?@ABCDEFGH
40	$1/10^4$	0.0001	99.99%	IJKLMNOPQR
50	$1/10^5$	0.00001	99.999%	STUVWXYZ[\

The Fleming Fund
Regional Grants

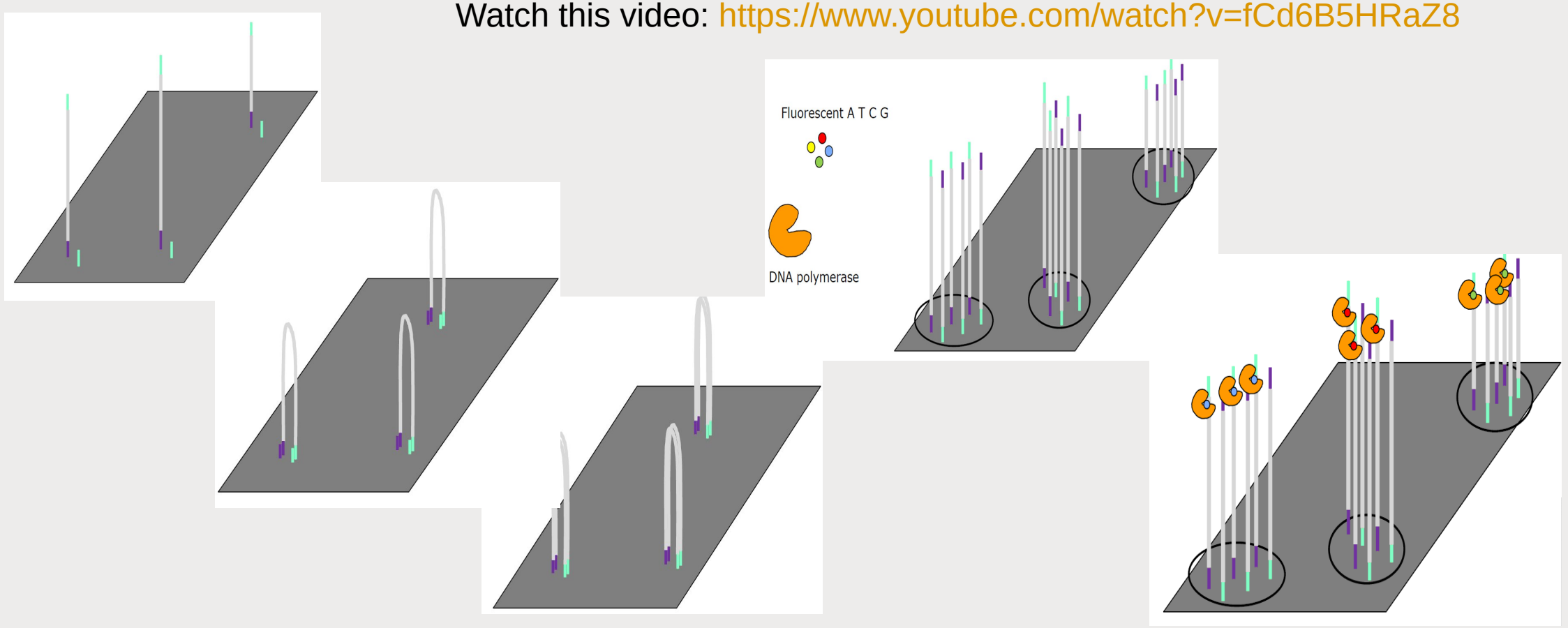
What are Adapters?



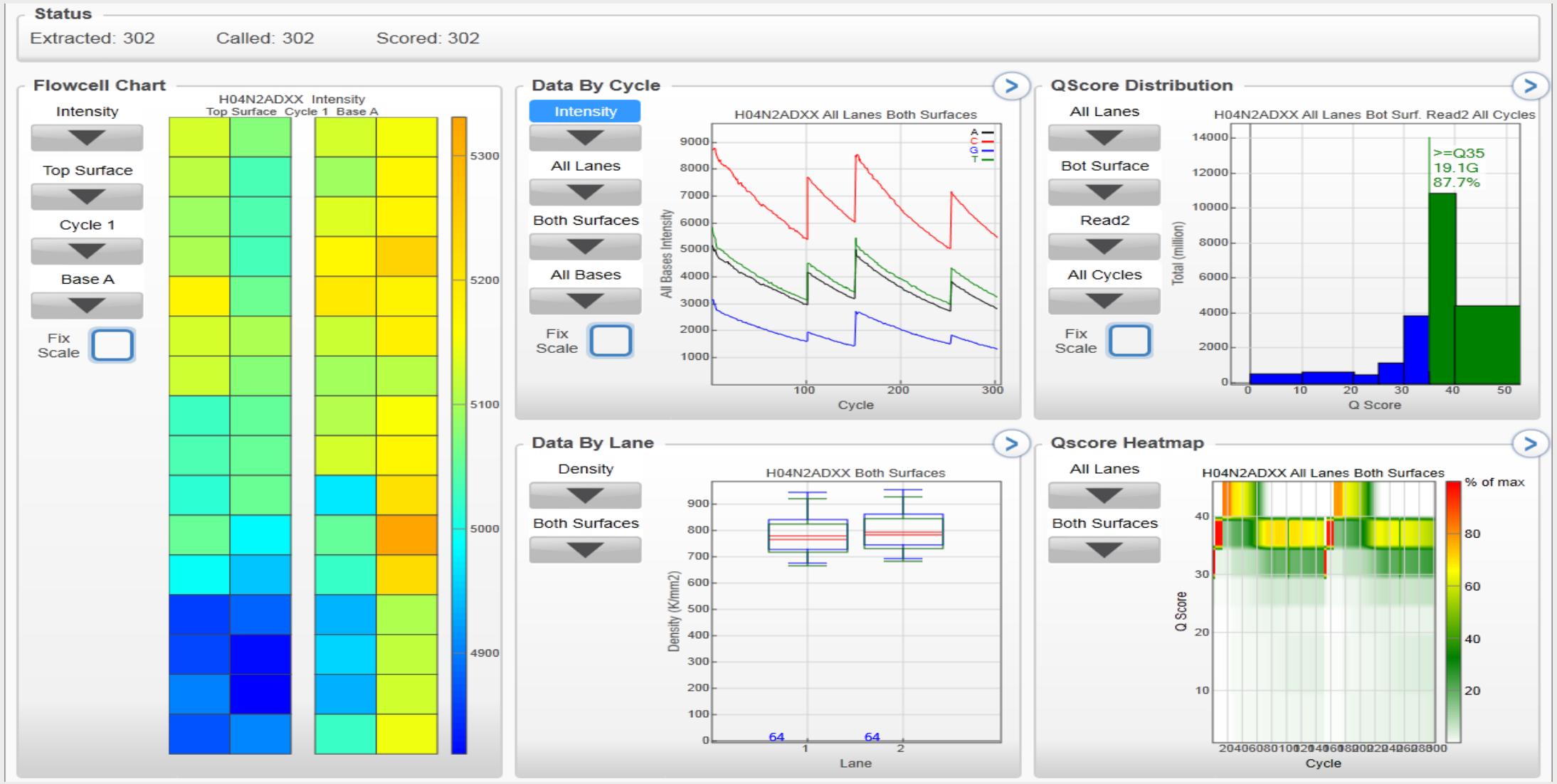
What are Clusters and Cycles?

(and lanes, tiles, and swathes?)

Watch this video: https://www.youtube.com/watch?v=fCd6B5HRaZ8



Sequencing Analysis Viewer



Sequencing Analysis Viewer

Sequencing Analysis Viewer

Run Folder: Y:\101029_P22_0759_BFC805GRAB

Browse | Refresh

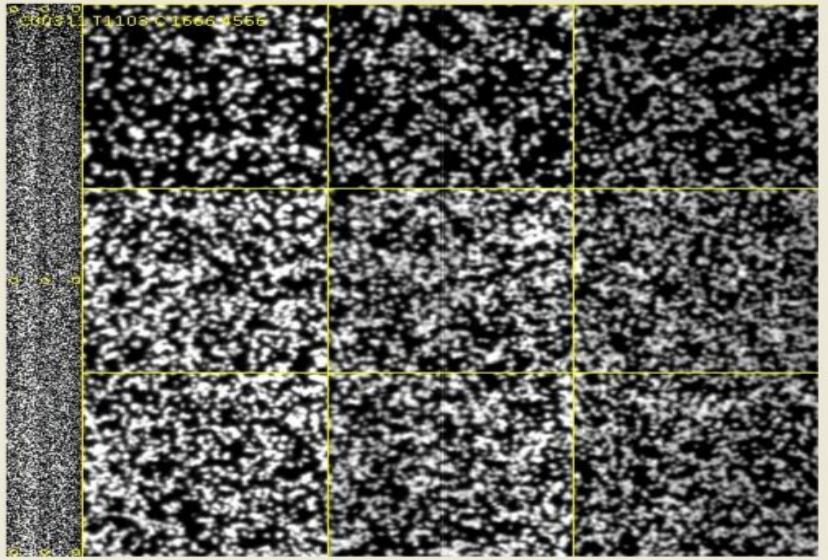
Analysis | Imaging | Summary | Tile Status | Controls

Cycle 1 | Lane 1 | Surface Top | Swath All | Section 3

A C G T

Index	Lane	Tile	Section	Cycle	Surface	Swath	Time	P90 A	P90 C	P90 G	P9
1297	1	1103	3	1	Top	1	10/29/201...	3326	3944	1729	41
1298	1	1103	3	2	Top	1	10/29/201...	3306	3906	1741	4(
1299	1	1103	3	3	Top	1	10/29/201...	3259	3853	1699	4(
1300	1	1103	3	4	Top	1	10/29/201...	3211	3806	1720	4(
1301	1	1103	3	5	Top	1	10/30/201...	3087	3703	1691	37
1302	1	1103	3	6	Top	1	10/30/201...	3066	3677	1610	35
1303	1	1103	3	7	Top	1	10/30/201...	2959	3554	1577	35
1304	1	1103	3	8	Top	1	10/30/201...	2945	3547	1558	35
1305	1	1103	3	9	Top	1	10/30/201...	2918	3518	1548	35
1306	1	1103	3	10	Top	1	10/30/201...	2906	3491	1536	35
1307	1	1103	3	11	Top	1	10/30/201...	2881	3461	1512	34
1308	1	1103	3	12	Top	1	10/30/201...	2869	3452	1498	34
1309	1	1103	3	13	Top	1	10/30/201...	2835	3423	1510	35
1310	1	1103	3	14	Top	1	10/30/201...	2854	3443	1493	34
1311	1	1103	3	15	Top	1	10/30/201...	2857	3445	1482	34
1312	1	1103	3	16	Top	1	10/30/201...	2811	3392	1452	33
1313	1	1103	3	17	Top	1	10/30/201...	2779	3356	1471	34
1314	1	1103	3	18	Top	1	10/30/201...	2756	3338	1435	33
1315	1	1103	3	19	Top	1	10/30/201...	2778	3364	1419	32
1316	1	1103	3	20	Top	1	10/30/201...	2749	3335	1435	33
1317	1	1103	3	21	Top	1	10/30/201...	2740	3324	1434	33
1318	1	1103	3	22	Top	1	10/30/201...	2692	3271	1441	33
1319	1	1103	3	23	Top	1	10/30/201...	2721	3308	1423	33
1320	1	1103	3	24	Top	1	10/30/201...	2687	3270	1420	33
1321	1	1103	3	25	Top	1	10/30/201...	2656	3229	1398	32

Rows=20736 | Disp=162 | Sel=1 | Filter



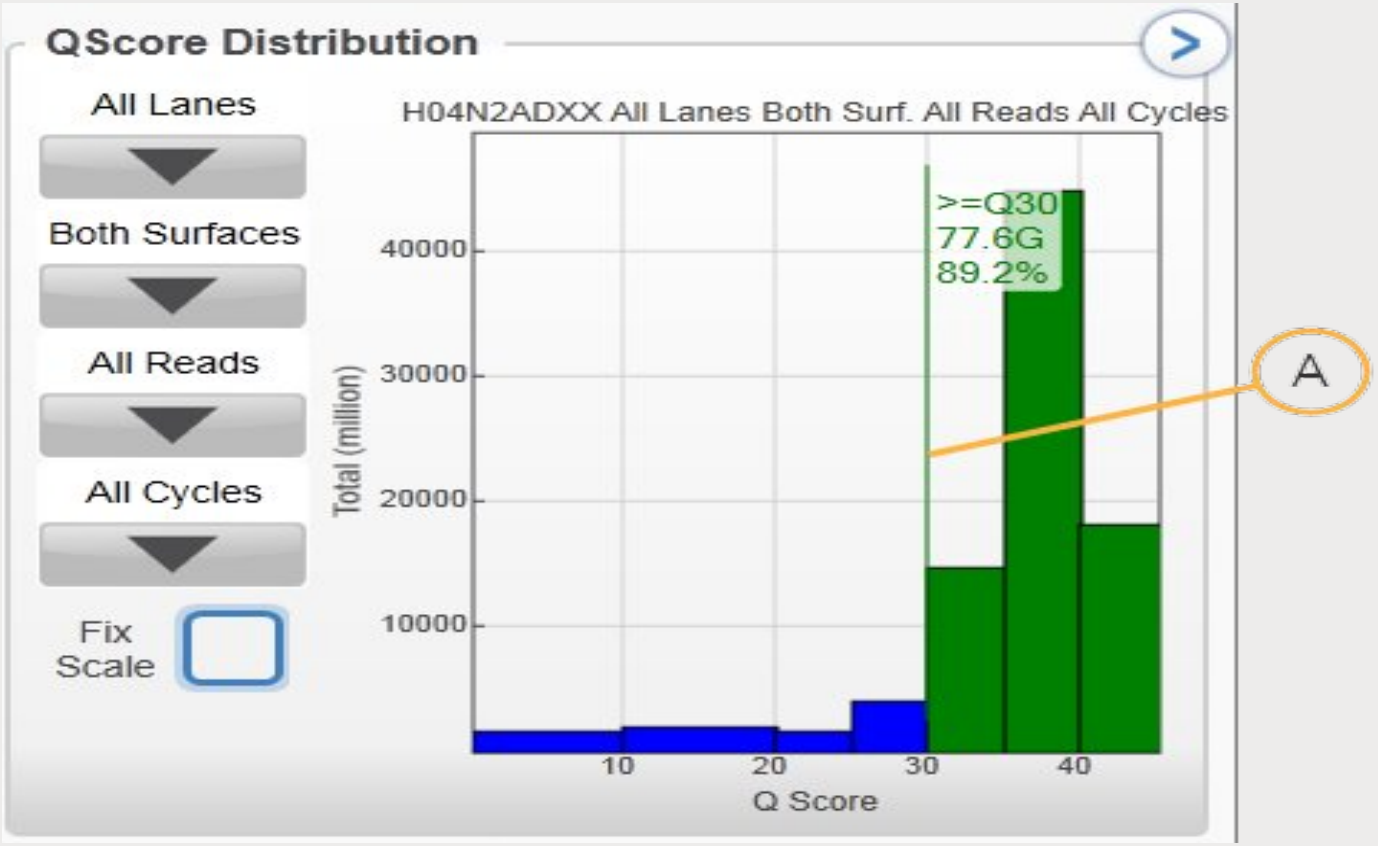
Main Metrics

- Yield (# bases): total number of bases read
- Error rate (%): percentage called incorrectly
- Q30+ (%): percentage bases with Q30+ score
- Density (K/mm^2): thousands of clusters per mm^2
- Clusters PF (%): percentage passing chastity filter
- (Pre-)phasing (%): ‘jumping’ or lagging bases

Metric: Error Rate

- Percentage of bases called incorrectly
- How does it know? PhiX
- Also:
  - % Aligned to PhiX (should match spiked %)
  - % Perfect reads (relative to PhiX)

Metric: Percentage Q30+



Metric: (Pre-)Phasing

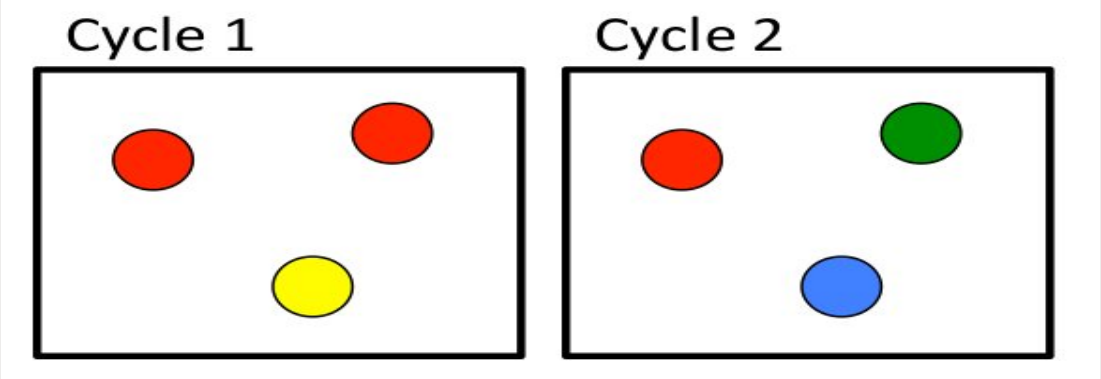
- Percentage of bases in cluster lagging or ‘jumping ahead’
  - Phasing: no base attaches, sequencing falls behind
  - Pre-phasing: skipping a base, sequencing jumps ahead
- Sequencer corrects for this (to some extent)
- Rule of thumb: below 0.5%, preferably <0.1%
- Causes: unbalanced bases, reagents or flow cell quality, temperature

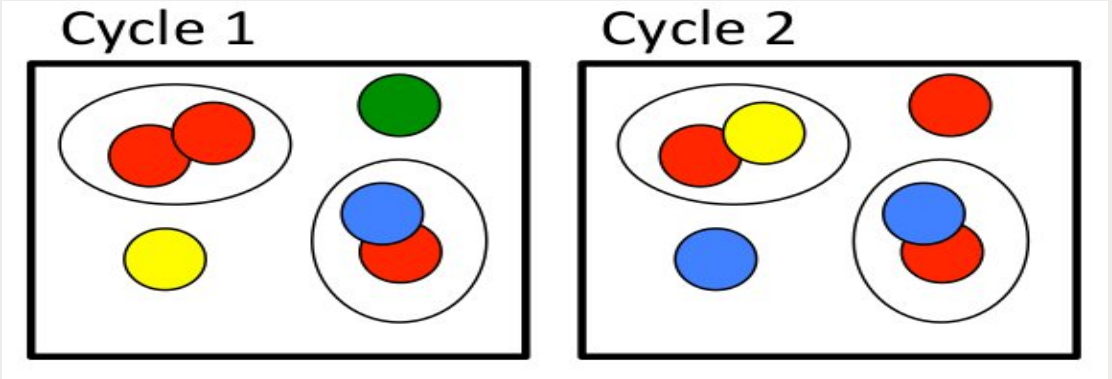
Metric: Clusters PF (passing filter)

- Chastity value:
  - Intensity of called base divided by sum of called base and second brightest
  - Threshold value is 0.6 (so must be “50% better than” second brightest base)
- Passing rule:
  - In first 25 cycles at most 1 base may have chastity under 0.6
- Subsequent SAV statistics are for Clusters PF
- Rule of thumb: 80% or higher

Metric: Cluster Density

- Bases hard to call
  - Q values down, %Error up
  - %PF and yield decrease
  - Index read failures
- Reverse read even worse
- Low diversity exacerbates
  - Early cycles determine cluster locations
  - Spike in more PhiX
- Underclustering: intensity and focus issues





FastQC Outputs

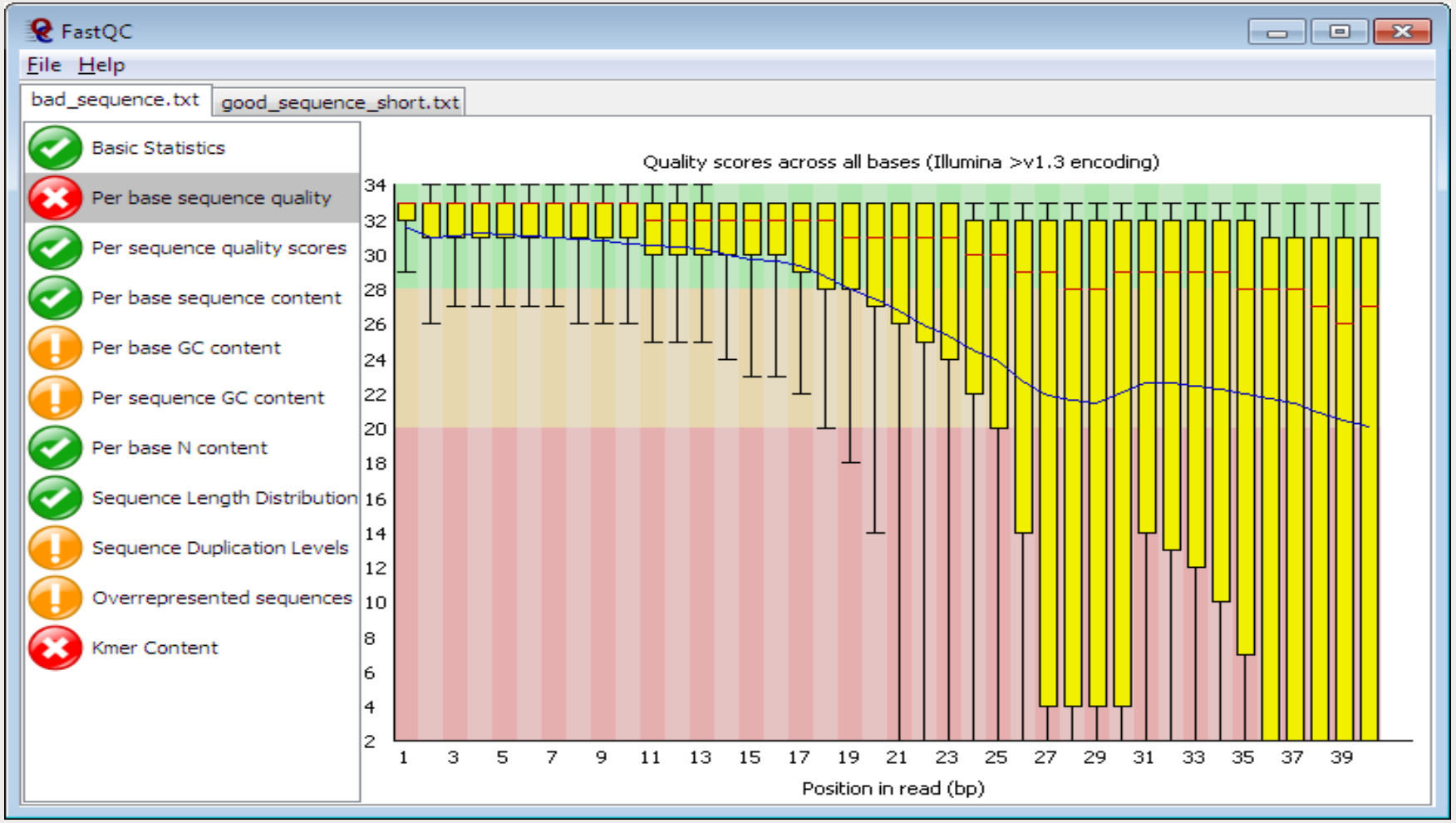
- http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- https://sequencing.qcfail.com/

FastQC: Basic Stats

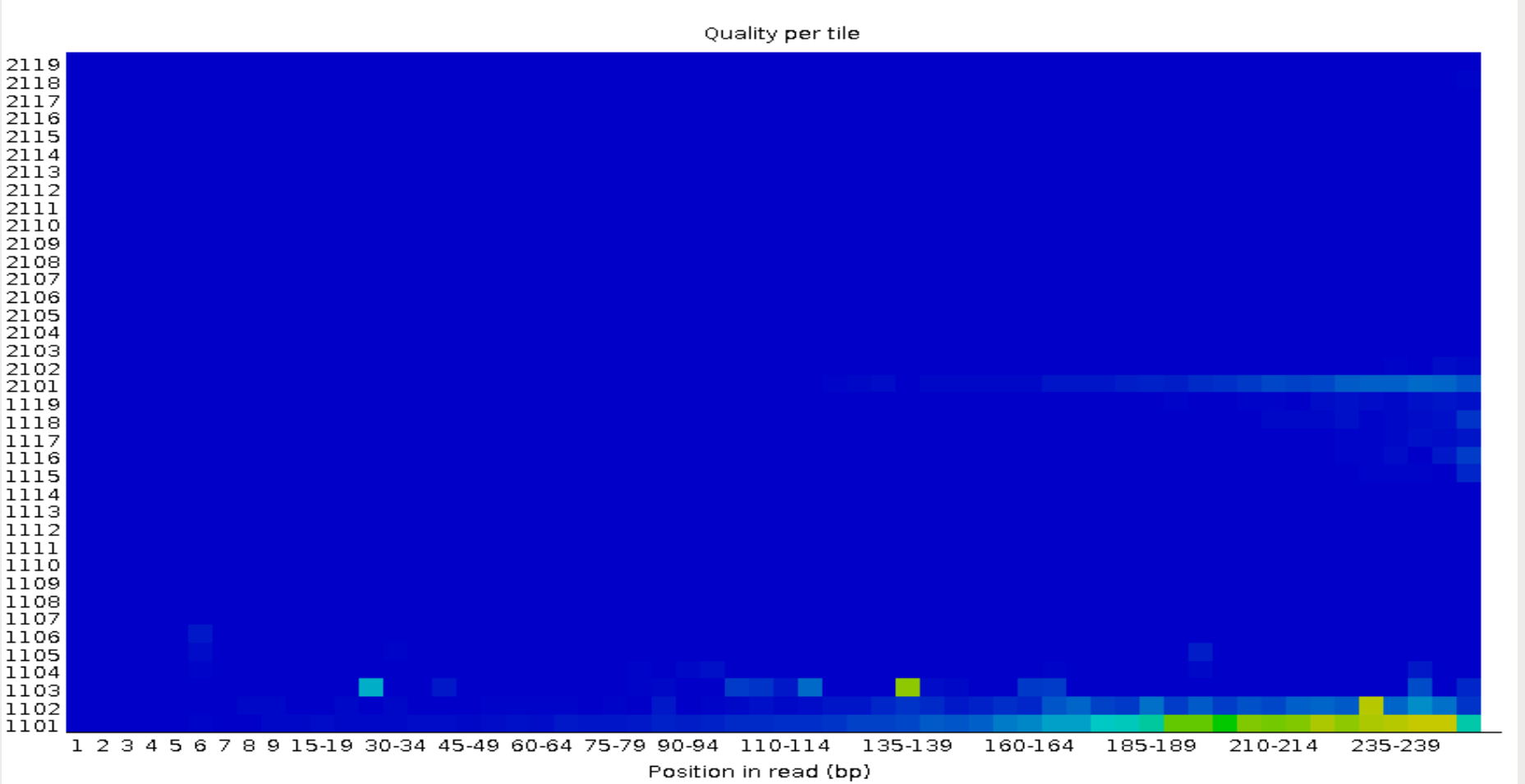
Basic Statistics

Measure	Value
Filename	110LF_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	743046
Sequences flagged as poor quality	0
Sequence length	35-251
%GC	56

FastQC: per base sequence quality

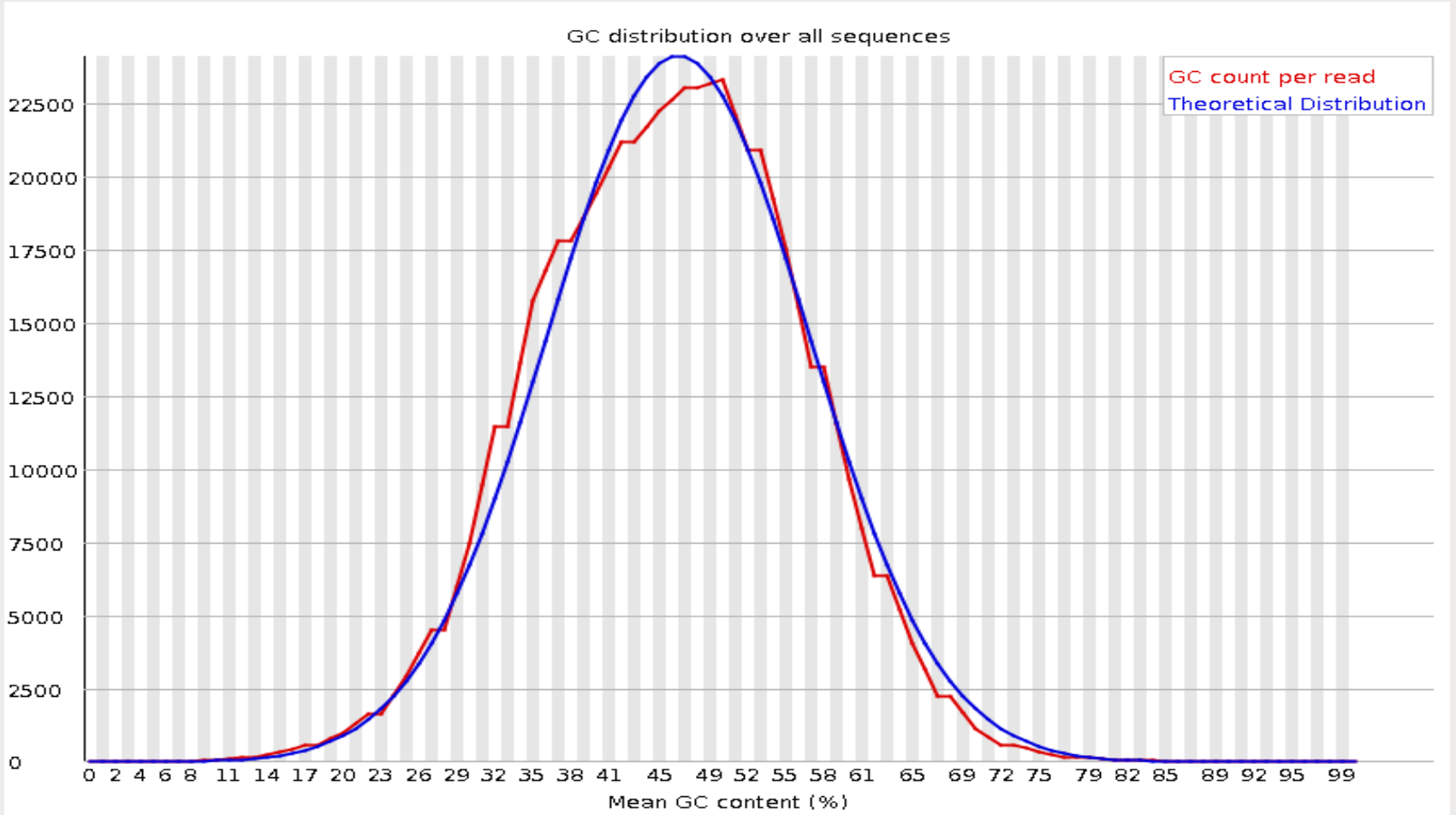


FastQC: Per tile (and cycle) quality

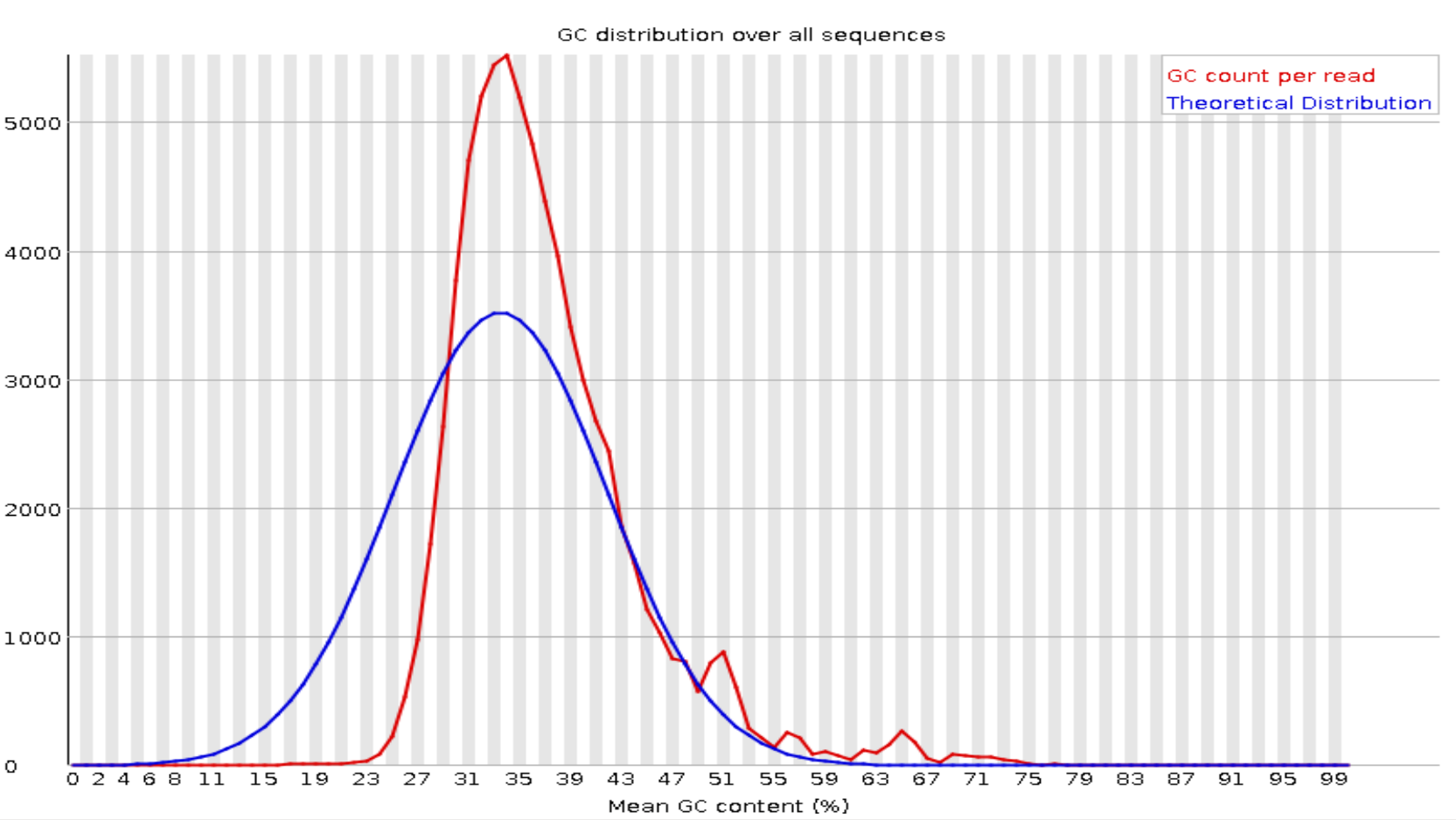


The Fleming Fund
Regional Grants

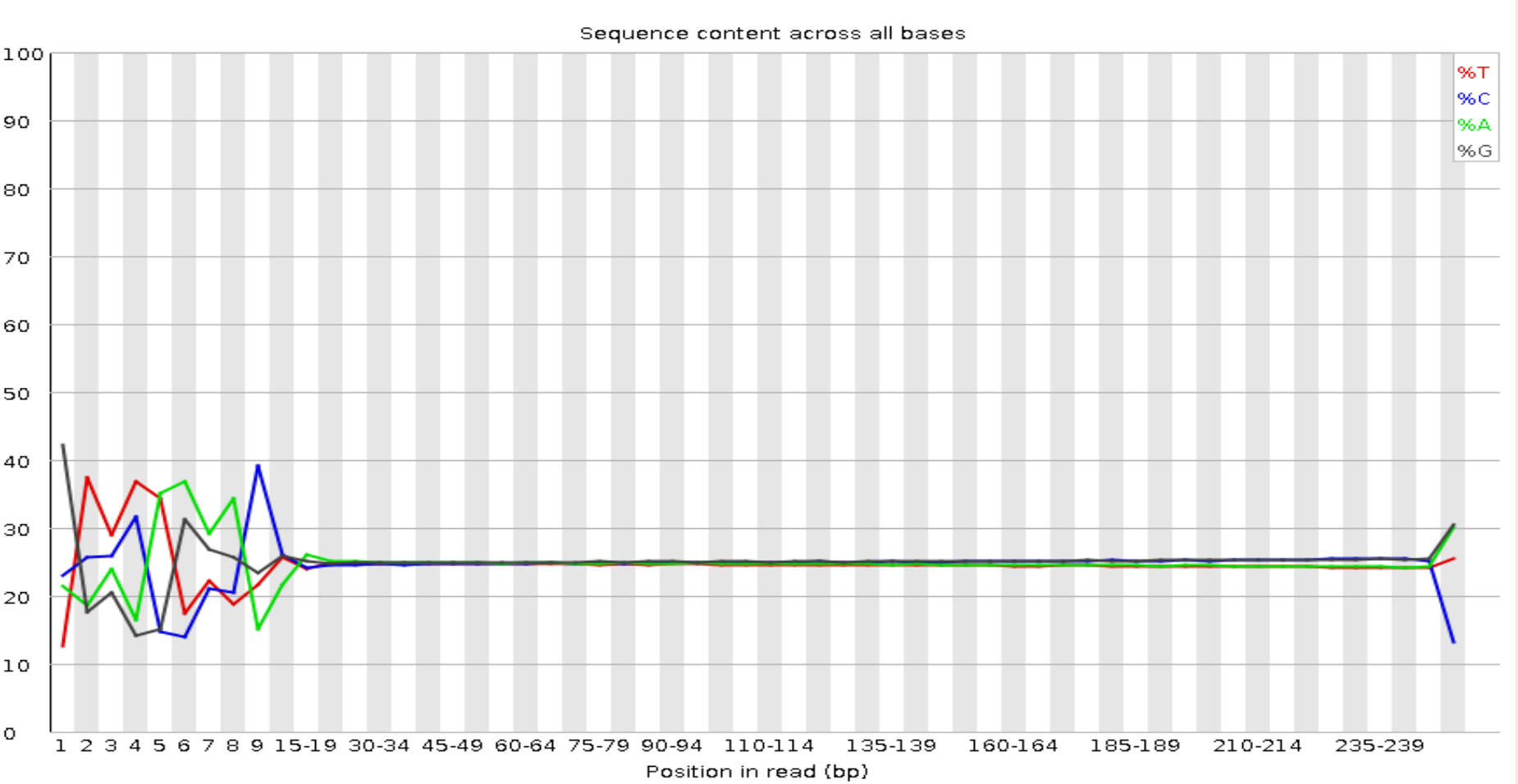
FastQC: GC distribution



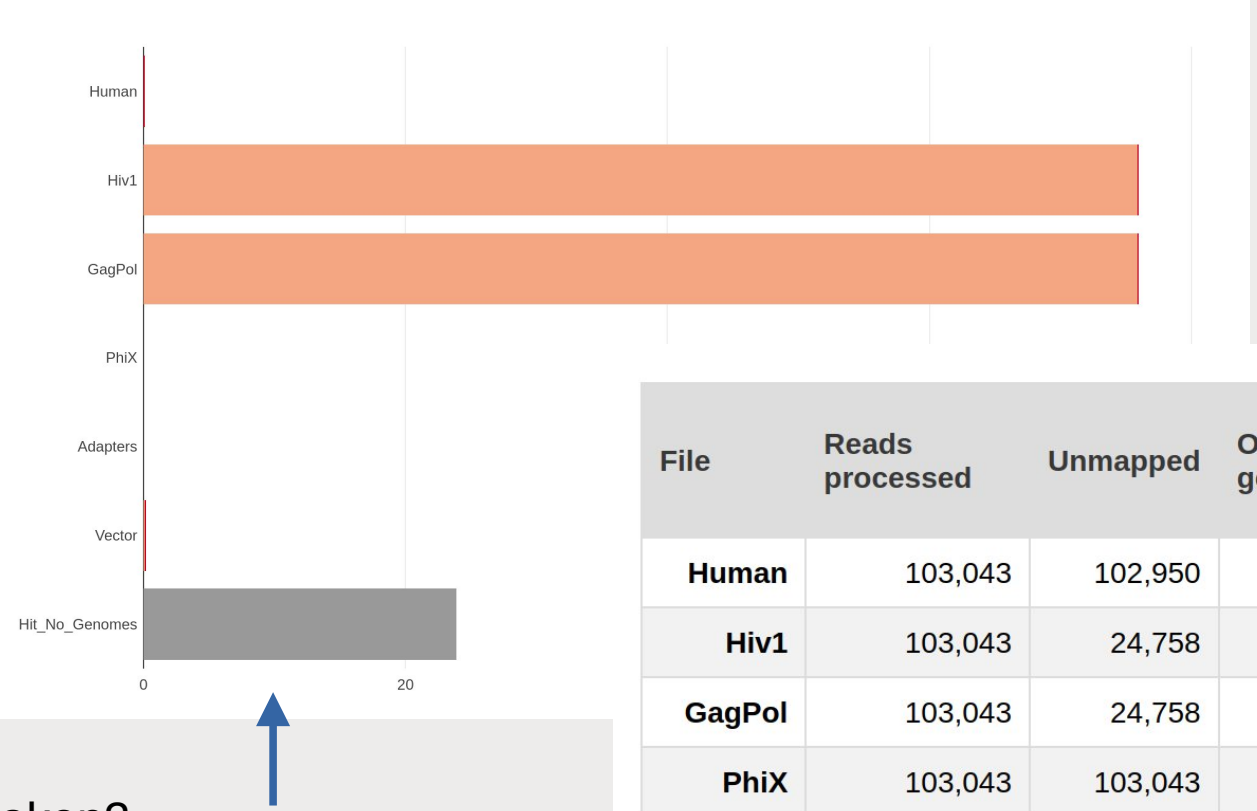
FastQC: unlikely GC distribution



FastQC: proportions of bases



Bonus: FastqScreen



File	Reads processed	Unmapped	One hit / one genome	Multiple hits / one genome	One hit / multiple genomes	Multiple hits / multiple genomes
Human	103,043	102,950	52	23	5	13
Hiv1	103,043	24,758	0	0	78,262	23
GagPol	103,043	24,758	0	0	78,262	23
PhiX	103,043	103,043	0	0	0	0
Adapters	103,043	103,043	0	0	0	0
Vector	103,043	102,903	50	4	72	14

Kraken2
(metagenomic binner)
may actually be more useful!

The Fleming Fund
Regional Grants

Thank you

NOGUCHI MEMORIAL INSTITUTE FOR MEDICAL RESEARCH
UNIVERSITY OF GHANA , LEGON

UNIVERSITY OF IBADAN
RECTE SAPERE FONS

NATIONAL INSTITUTE FOR COMMUNICABLE DISEASES
Division of the National Health Laboratory Service

KCRI Kilimanjaro Clinical Research Institute
Academic Centre for Evidence Based Health Interventions
An Institution of the Good Samaritan Foundation

UKaid
from the British people

MOTT MACDONALD

SEQAFRICA

DTU

This programme is being funded by the UK Department of Health and Social Care.
The views expressed do not necessarily reflect the UK Government's official policies.