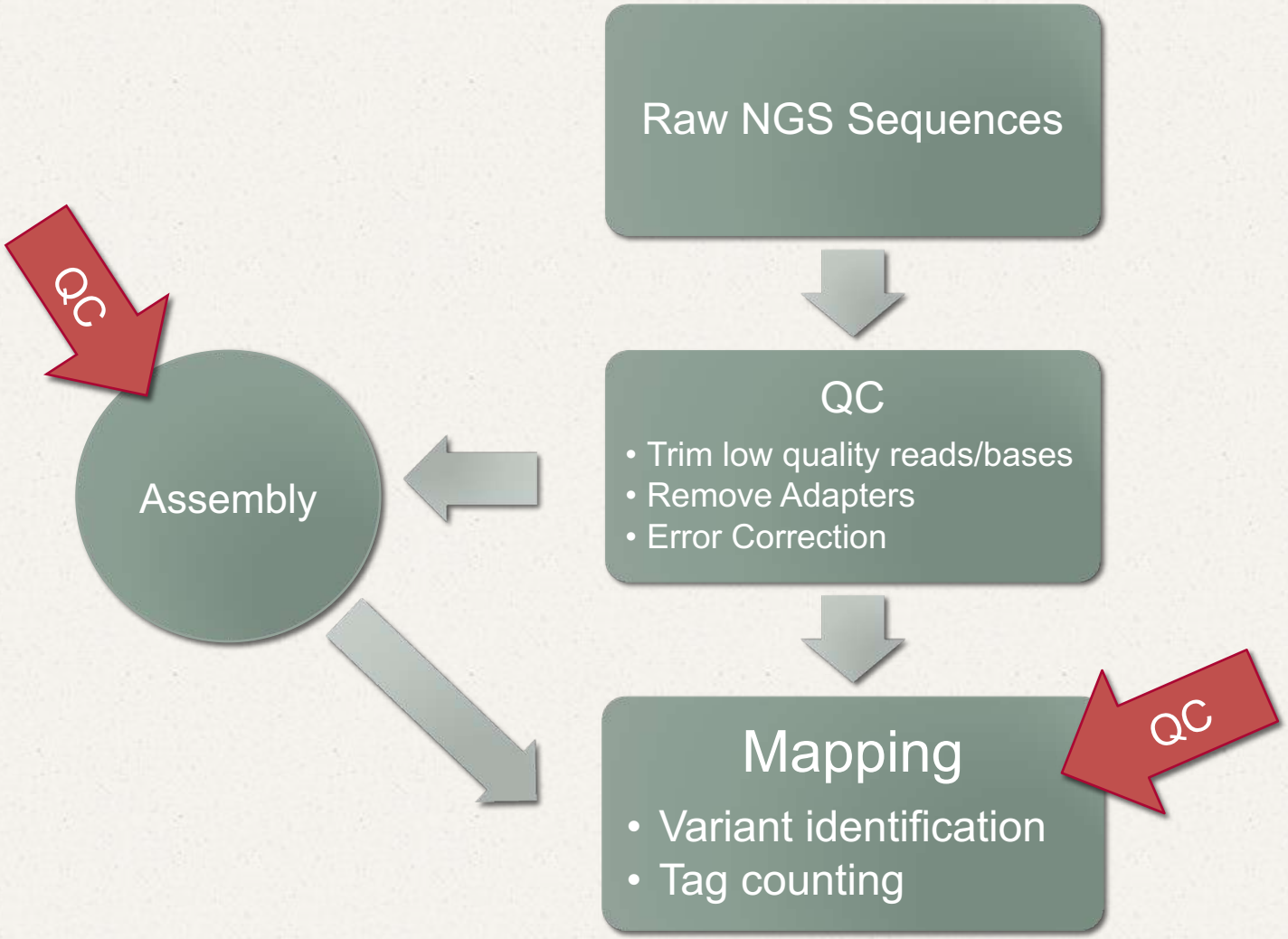
NGS QC

An introduction

Goals:

- Raw NGS Data Formats
- Evaluating Raw Data
- Cleaning Raw Data
- K-mer Counting

The Big Picture




PHASE TWO: INTERPRETATION
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

The FASTA format

```
>sequence 1
CATCGATCGCATGCTACTGACTG
CATGCTCGCGCCCCCCCCGATG
>sequence 2
ACTGACTCGCGCGCGCGGGGG
GAGCTGATGTG
>sequence 3
CATCGATCGCATGCTACTGACTG
CATGCTCGCGCCCCCCCCGATG
ACTGACTCGCGCGCGCGGGGG
GAGCTGATGTG
```

FASTQ format

@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
hhhhhhhhhhhghhhhhhhhehhhedhhhhfhhhhhh

FASTQ format

Sequence ID

@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
hhhhhhhhhhhghhhhhhhhehhhedhhhhfhhhhhh

FASTQ format

Sequence

@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
hhhhhhhhhhhghhhhhhhhehhhedhhhhfhhhhhh

FASTQ format

+ description (or empty)

```
@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
hhhhhhhhhhhghhhhhhhhehhhedhhhhfhhhhhh
```

FASTQ format

+ description (or empty)

@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG
+
hhhhhhhhhhhghhhhhhhhehhhedhhhhfhhhhhh

FASTQ format

Quality score of each base

@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
hhhhhhhhhhghhhhhhhehhhedhhhhfhhhhhh

Illumina Sequence ID Lines: A Decoder

@M01137:30:000000000-AA299:1:1101:10929:1966	
M01137	the unique instrument name
30	the run id
000000000-AA29	the flowcell id
1	flowcell lane
1101	tile number within the flowcell lane
10929	'x'-coordinate of the cluster within the tile
1966	'y'-coordinate of the cluster within the tile
1 or 2 (not shown, optional)	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
ATCACG (not shown, optional)	index sequence

Quality Scores

- Phred Score
- $Q = -10*\log_{10}P$ P = probability the base call is incorrect
- ASCII (character) - 33

Phred Quality Score	Probability of incorrect base call	Base call accuracy
0	1	0 %
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %
93	1 in 2000000000	99.9999995 %

Why is trimming important?

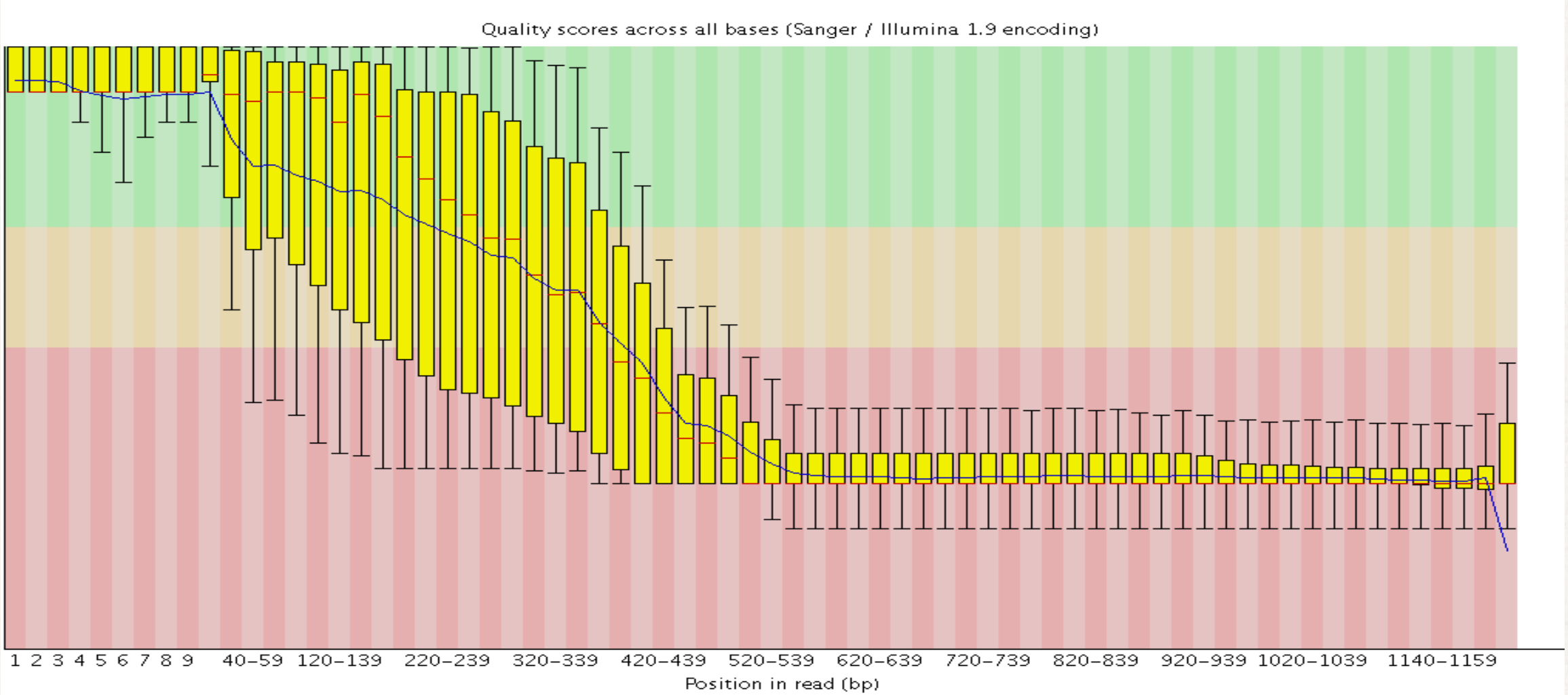
OPEN ACCESS Freely available online PLOS ONE

An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis

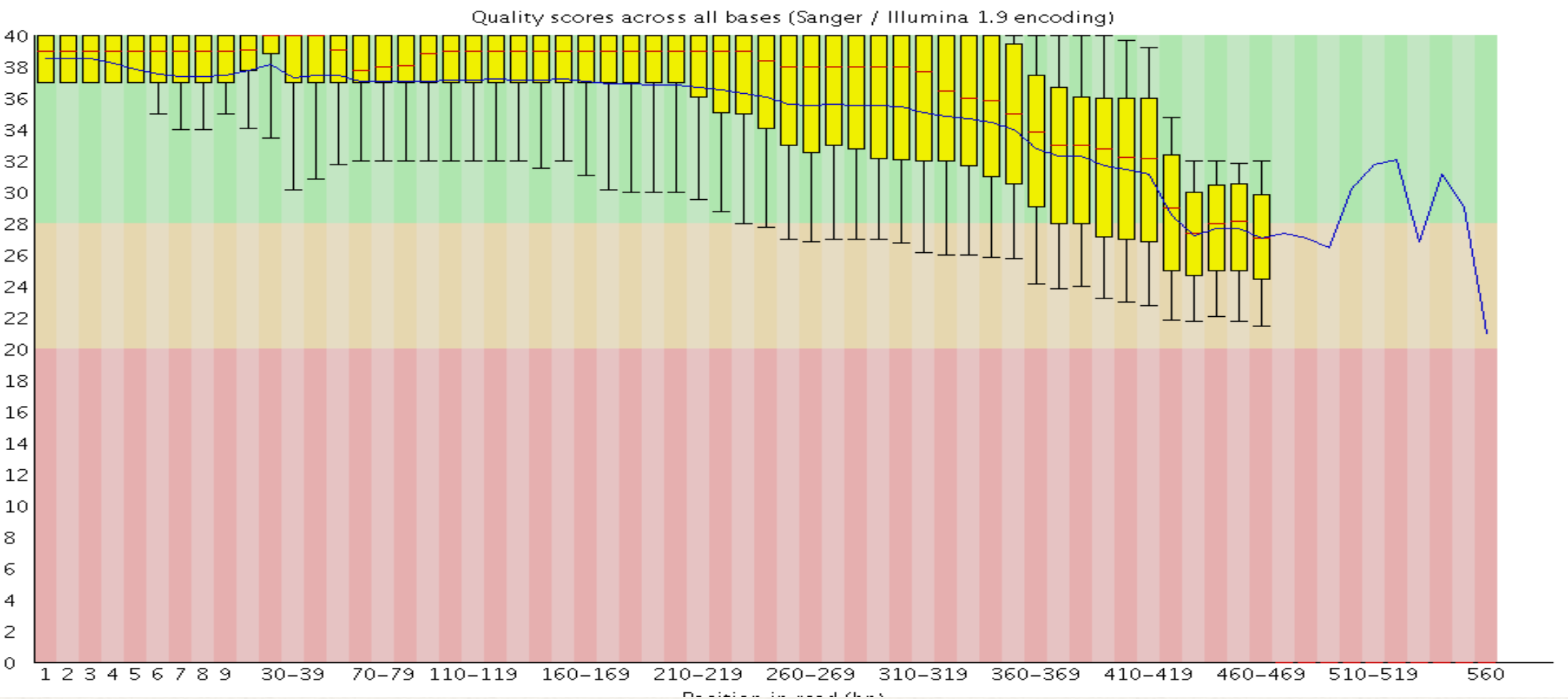
Cristian Del Fabbro[1☯], Simone Scalabrin[2☯], Michele Morgante[1], Federico M. Giorgi[1,3*]

"Trimming is shown to increase the quality and reliability of the analysis, with concurrent gains in terms of execution time and computational resources needed"

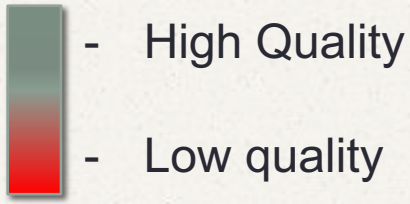
Low Quality Sequences Before Trimming (Puma 454 sequences)



Same Sequences After Trimming (Puma 454 sequences)

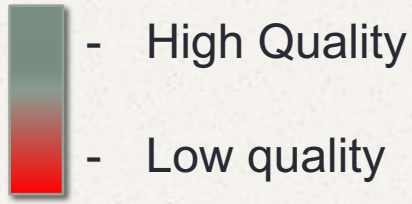


Types of Trimming

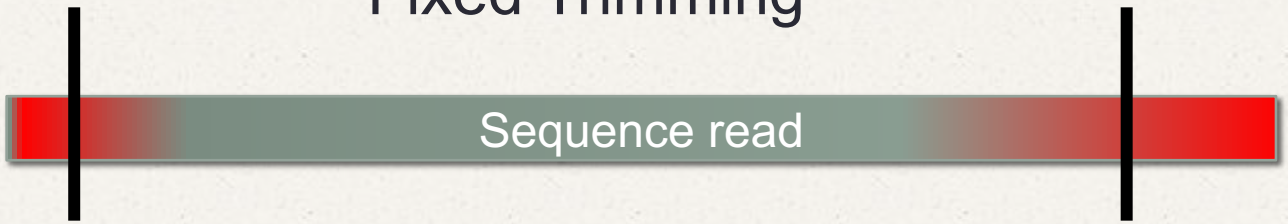


- High Quality
- Low quality

Types of Trimming



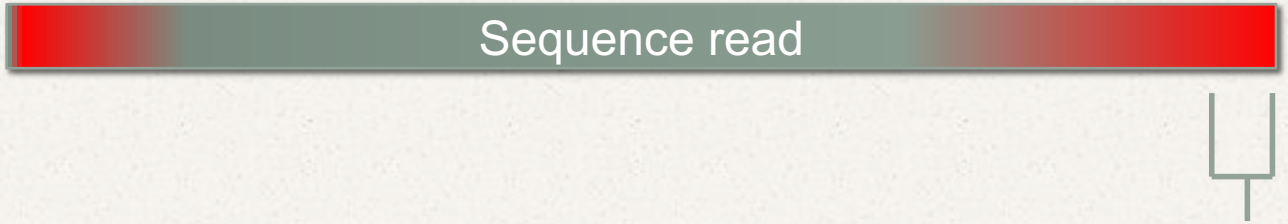
Fixed Trimming



Types of Trimming

- High Quality
- Low quality

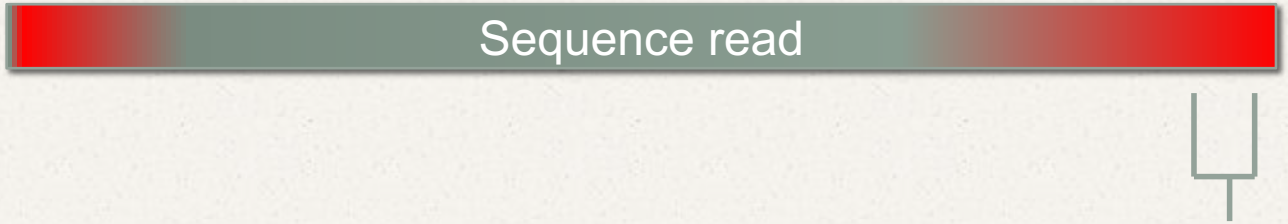
Sliding Window Trimming



Types of Trimming

- High Quality
- Low quality

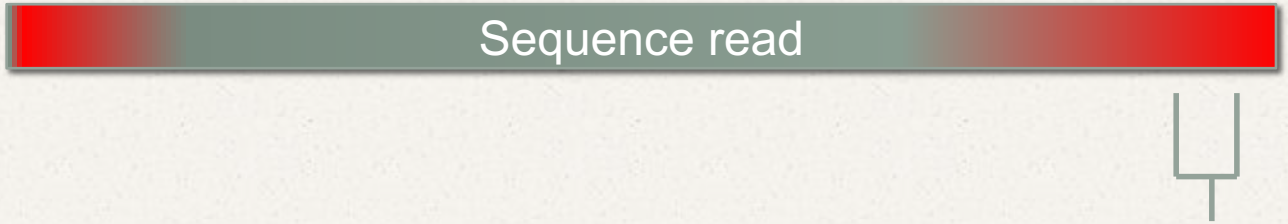
Sliding Window Trimming



Types of Trimming

- High Quality
- Low quality

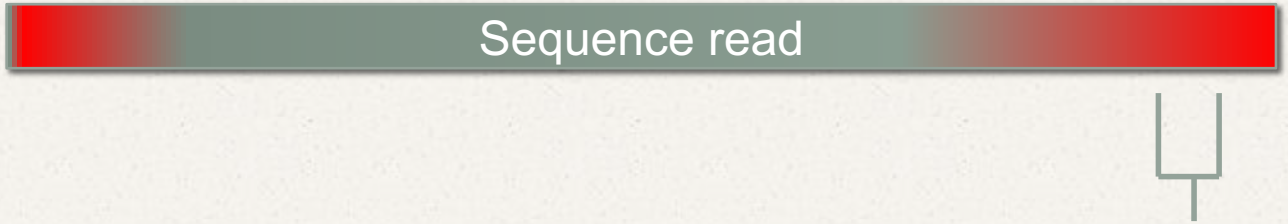
Sliding Window Trimming



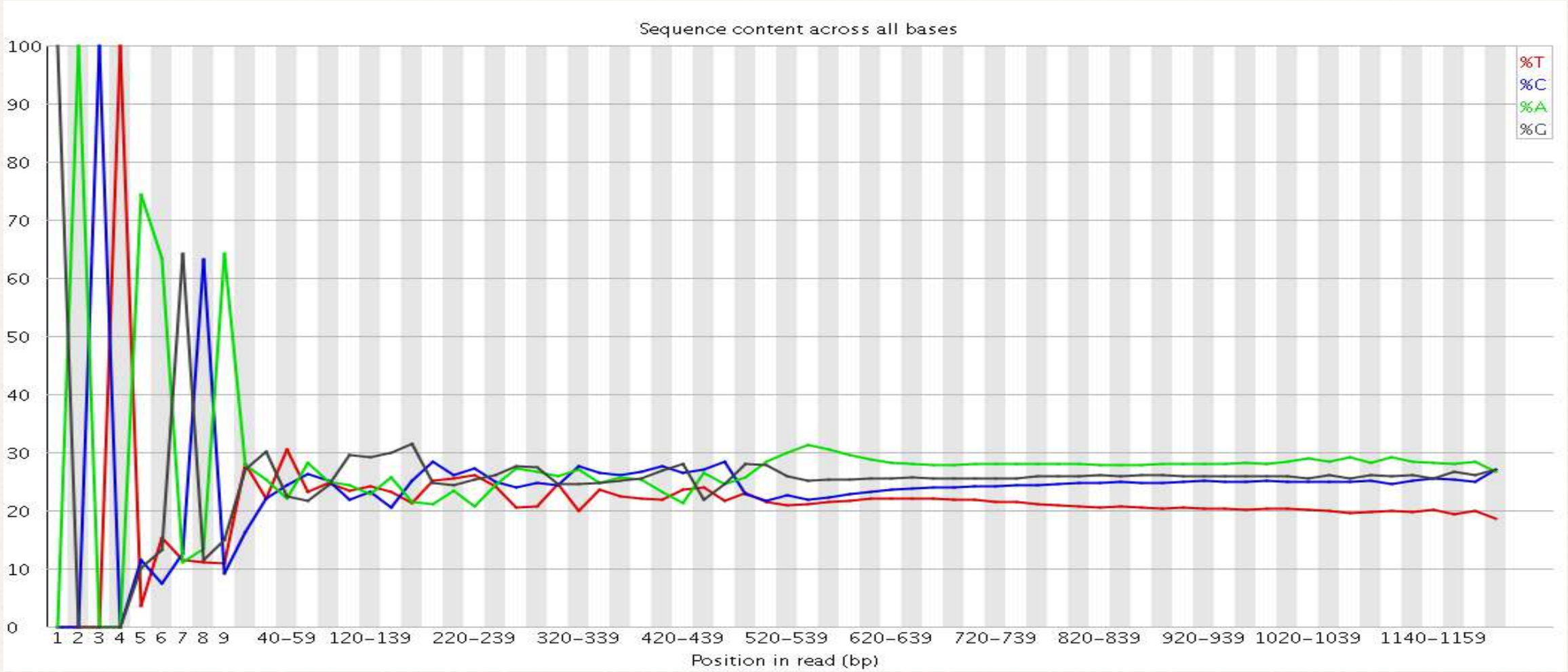
Types of Trimming

- High Quality
- Low quality

Sliding Window Trimming



Adapter Contamination



Adapter Contamination

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GACTAAGCAGTGGTATCAACGCAGAGTACATGGGGACACTTGTTTCTGAC	19391	5.415739186535921	No Hit
GACTAAGCAGTGGTATCAACGCAGAGTACATGGGGACACTTGCTTCTGAC	11325	3.162974900083508	No Hit
GACTAAGCAGTGGTATCAACGCAGAGTACATGGGACACTTGTTTCTGACA	9229	2.5775801636088915	No Hit

Download Graphics

gnl|uv|NGB00593.1:1-30 Evrogen Mint PlugOligo-1 adapter

Sequence ID: **Length:** 30 **Number of Matches:** 1

Range 1: 1 to 25 Graphics Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
50.6 bits(25)	0.001	25/25(100%)	0/25(0%)	Plus/Plus

```
Query  5   AAGCAGTGGTATCAACGCAGAGTAC  29
           |||||||||||||||||||||||||
Sbjct  1   AAGCAGTGGTATCAACGCAGAGTAC  25
```

Error Correction

GAGE: A critical evaluation of genome assemblies and assembly algorithms

Steven L. Salzberg,[1,7] Adam M. Phillippy,[2] Aleksey Zimin,[3] Daniela Puiu,[1] Tanja Magoc,[1] Sergey Koren,[2,4] Todd J. Treangen,[1] Michael C. Schatz,[5] Arthur L. Delcher,[6] Michael Roberts,[3] Guillaume Marçais,[3] Mihai Pop,[4] and James A. Yorke[3]

"For all four genomes and for all eight assemblers used in GAGE, the best assemblies were created from reads that had been processed through extensive error correction routines"

Illumina Sequencing Errors: ~1%, Substitution errors

Error-correction: K-mer counting

k = 4

AGCTGTGG

Error-correction: K-mer counting

k = 4

AGCTGTGG

AGCT

Error-correction: K-mer counting

k = 4

AGCTGTGG

AGCT

GCTG

Error-correction: K-mer counting

k = 4

AGCTGTGG

AGCT

GCTG

CTGT

Error-correction: K-mer counting

k = 4

AGCTGTGG

AGCT

GCTG

CTGT

TGTG

Error-correction: K-mer counting

k = 4

AGCTGTGG

AGCT

GCTG

CTGT

TGTG

GTGG

Error-correction: K-mer counting

k = 6

AGCTGTGG

Error-correction: K-mer counting

k = 6

AGCTGTGG

AGCTGT

Error-correction: K-mer counting

k = 6

AGCTGTGG

AGCTGT

GCTGTG

Error-correction: K-mer counting

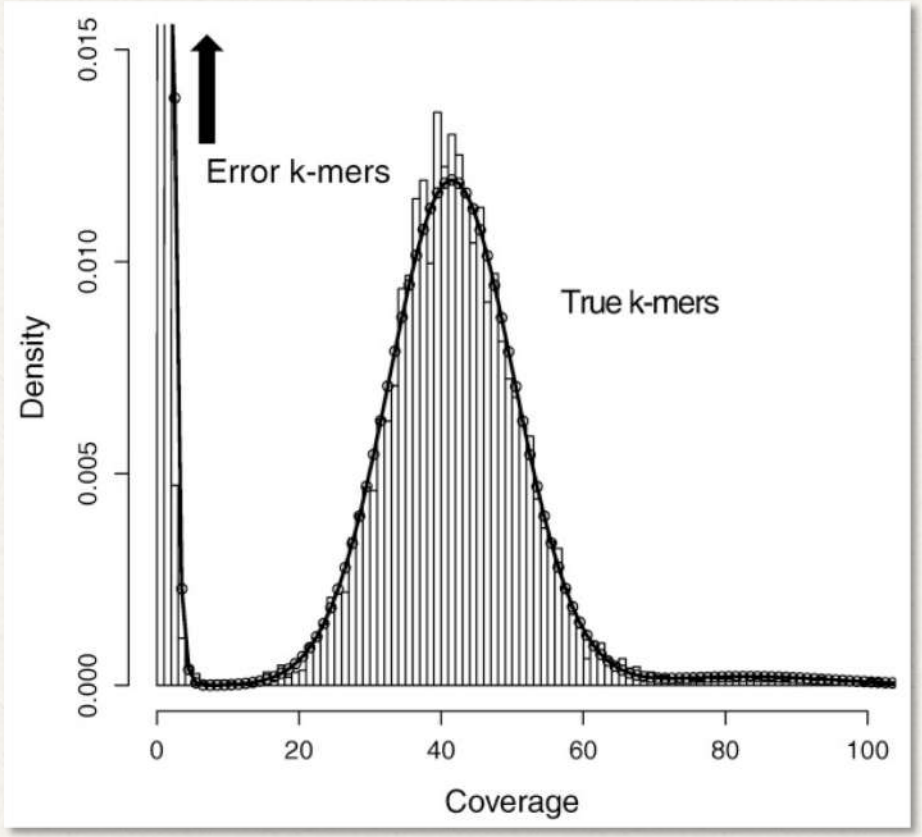
k = 6

AGCTGTGG

AGCTGT
GCTGTG
CTGTGG

K-mers and Error-correction

- Expected Distribution of k-mer frequency

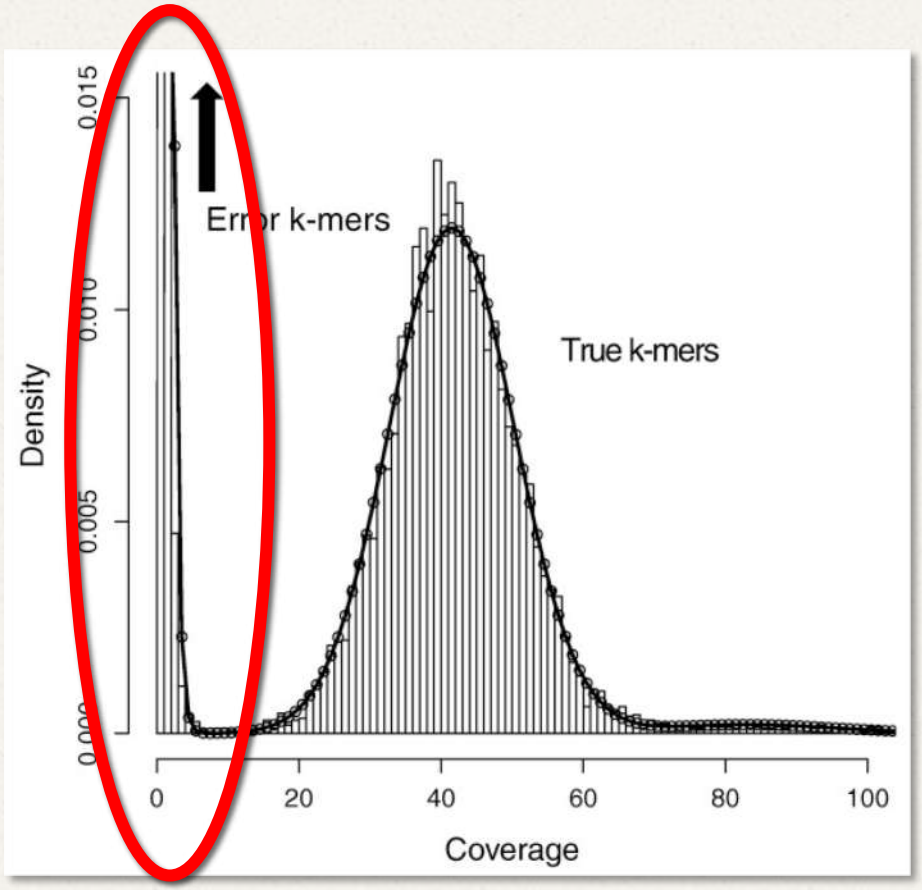


DSK; Rizk et al. 2013

K-mers and Error-correction

- Expected Distribution of k-mer frequency

Corrected

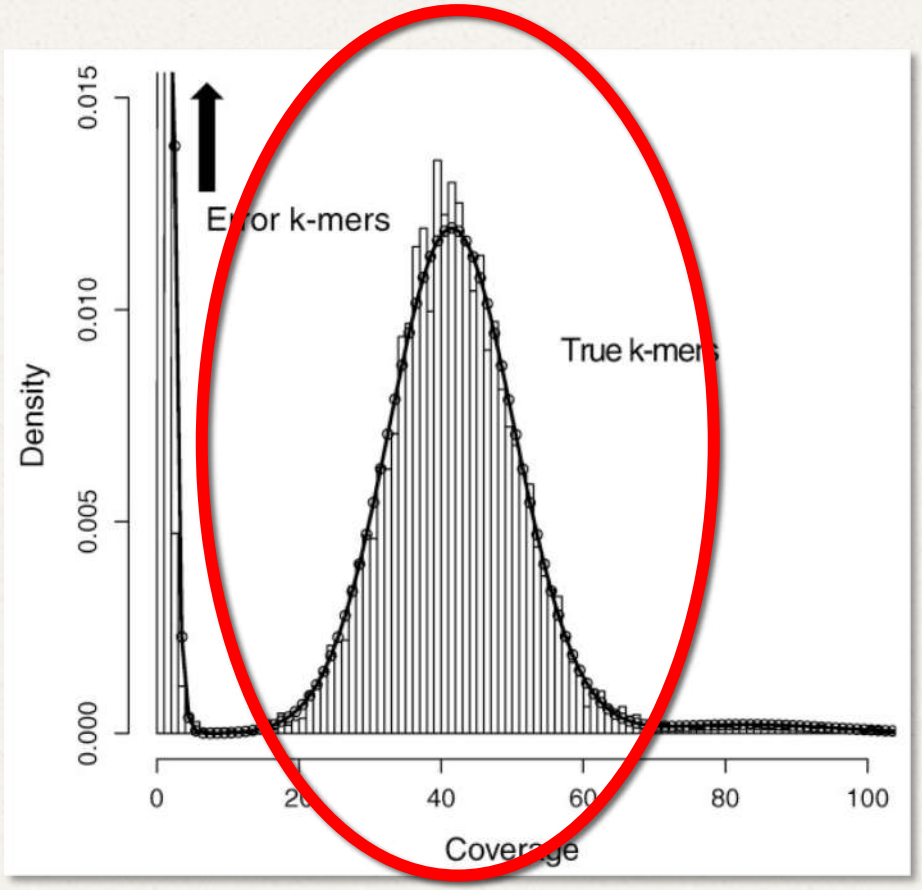


DSK; Rizk et al. 2013

K-mers and Error-correction

- Expected Distribution of k-mer frequency

Estimate genome size

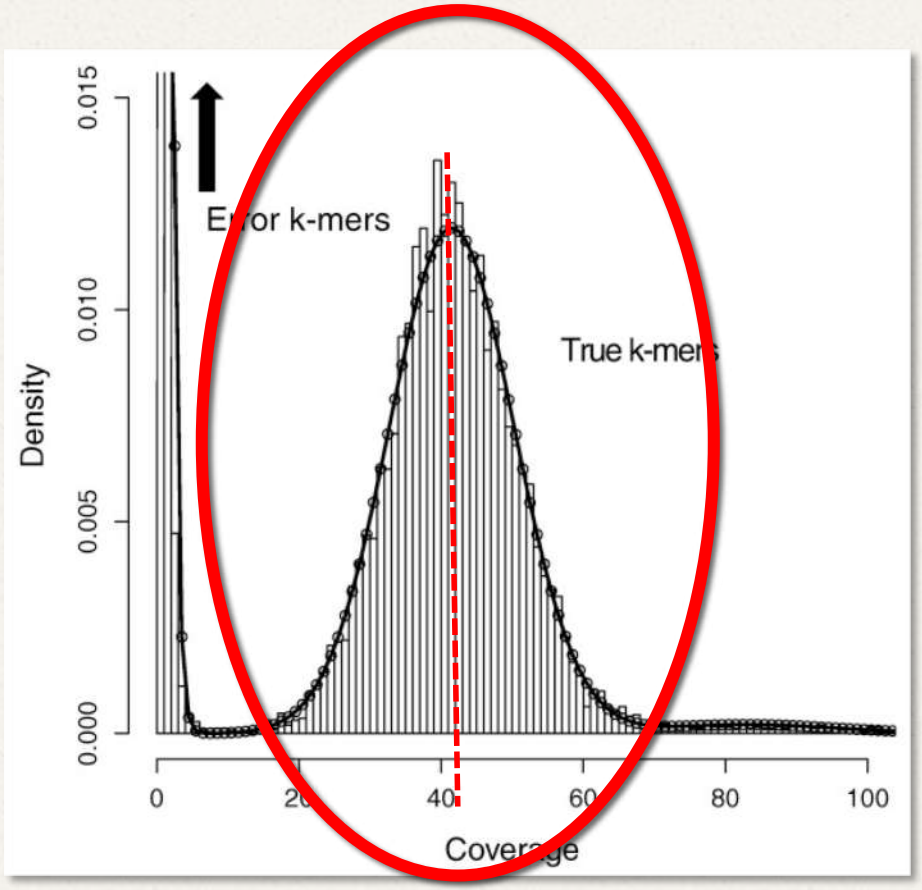


DSK; Rizk et al. 2013

K-mers and Error-correction

- Expected Distribution of k-mer frequency

Estimate genome size



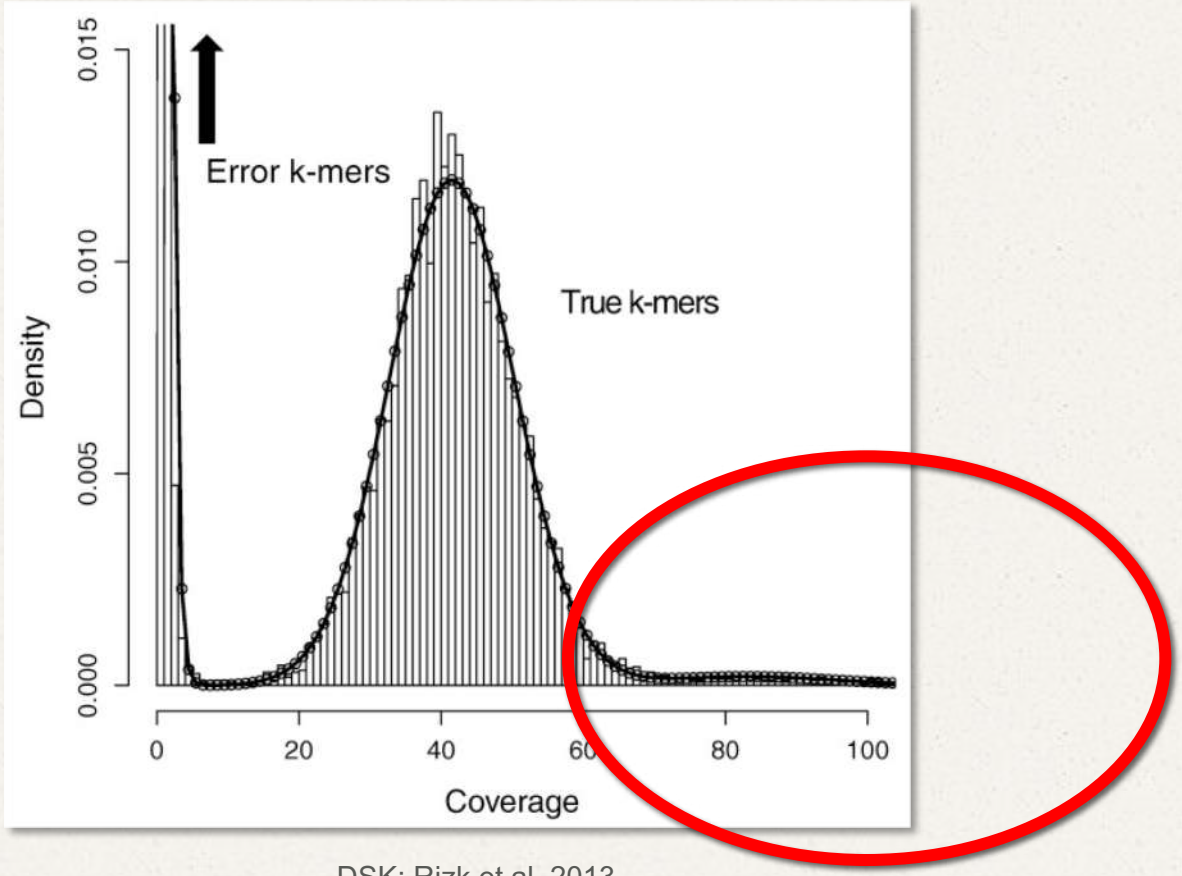
DSK; Rizk et al. 2013

G = C / P
G = genome size
C = total count of true k-mers
P = peak coverage

K-mers and Error-correction

- Expected Distribution of k-mer frequency

Estimate repetitive content



DSK; Rizk et al. 2013

Recap: NGS QC

- Remove low quality bases and reads
- Identify and remove adapter contamination
- Optional: Correct substitution sequencing errors
- Optional: De-duplication

To your MacBookPro Terminal!