

HORIZON SCANNING

A solid 'scan of the horizon' can provide the background to develop strategies for anticipating future developments and thereby gain lead time. It can also be a way to assess trends to feed into a scenario development process.



Horizon Scanning

“Horizon scanning is a technique for detecting early signs of potentially important developments through a systematic examination of potential threats and opportunities, with emphasis on new technology and its effects on the issue at hand. The method calls for determining what is constant, what changes, and what constantly changes. It explores novel and unexpected issues as well as persistent problems and trends, including matters at the margins of current thinking that challenge past assumptions.”

<https://www.oecd.org/site/schoolingfortomorrowknowledgebase/futuresthinking/overviewofmethodologies.htm>



Horizon Scanning

Trends in Ecology & Evolution

Volume 31, Issue 1, January 2016, Pages 44-53



Feature Review

A Horizon Scan of Global Conservation Issues for 2016

THE ECOLOGICAL SOCIETY OF AMERICA

Frontiers in Ecology and the Environment

[Explore this journal >](#)

Review

Emerging threats in urban ecosystems: a horizon scanning exercise

HYPOTHESIS & THEORY ARTICLE

Front. Genet., 21 October 2015 | <https://doi.org/10.3389/fgene>.

Prospects and challenges for the conservation of farm animal genomic resources, 2015-2025



Goals

- By Friday:
 - Contribute 5 anonymous questions of highest priority for research, infrastructure, and policy development as it relates to Genomics of Disease in Wildlife
- The speakers will organize or cluster these questions in 4-6 key concept areas
- On Saturday:
 - We will assign a moderator to each concept area and participants will be divided among them
 - The groups will meet and discuss their area, highlighting the prospects and challenges of their area.
- Sometime in the future:
 - Assemble the results (hopefully there are good note takers) into a symposium article describing our findings.

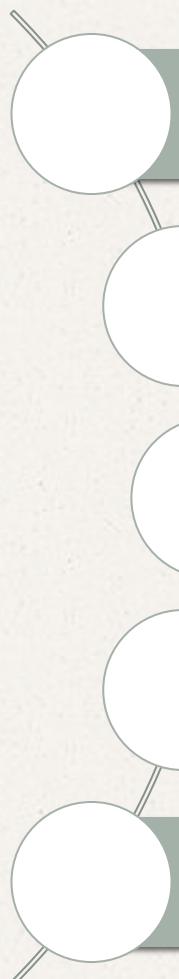


NGS QC

An introduction



Goals for the Morning:



Raw NGS Data Formats

Evaluating Raw Data

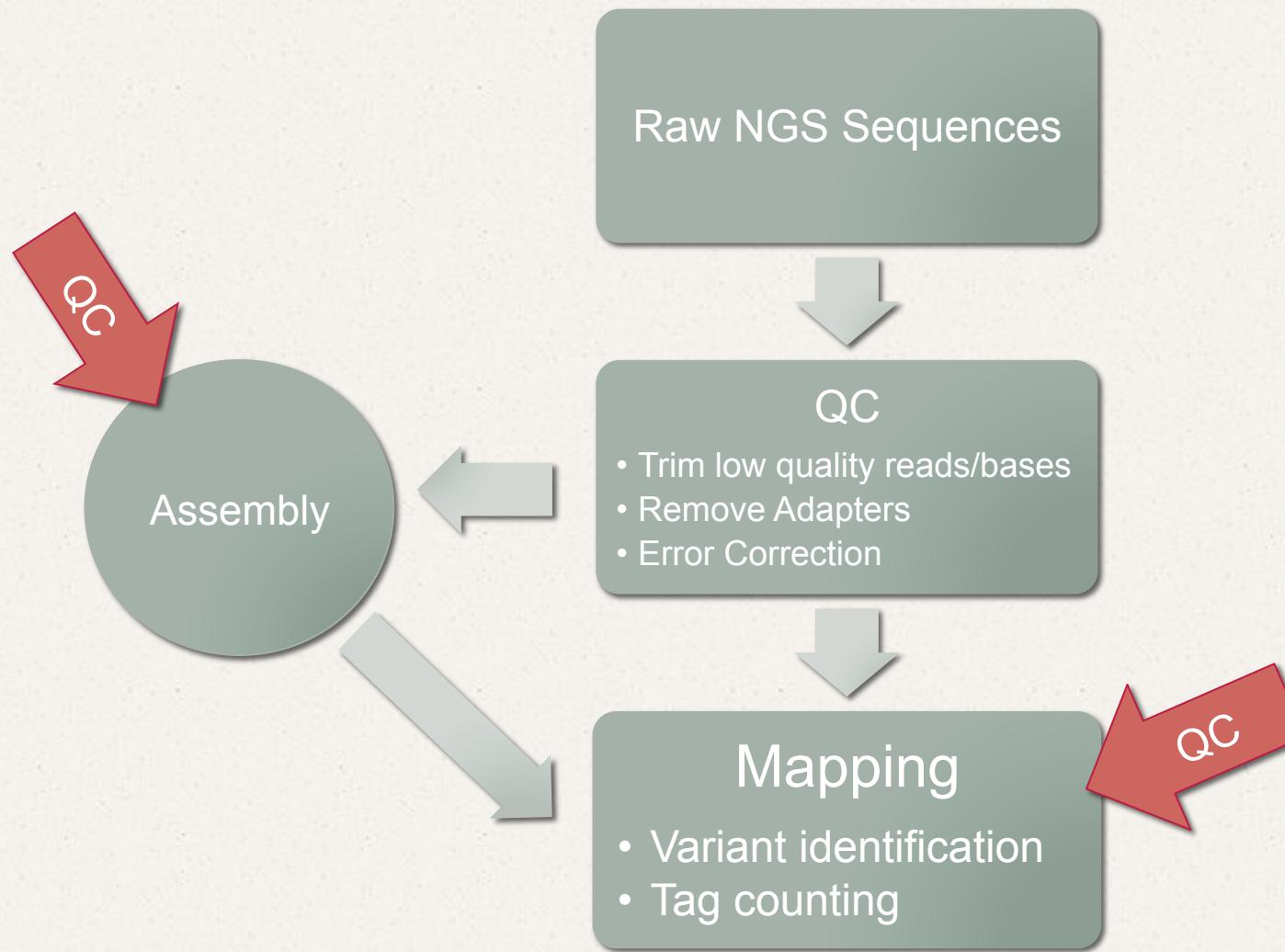
Cleaning Raw Data

K-mer counting

Command line tutorial



The Big Picture





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
hhhhhhhhhhghhhhhhehhhedhhhhfhhhhh
```



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
hhhhhhhhhhghhhhhhehhhedhhhhfhhhhh
```



FASTQ format

Sequence

```
@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
GTAGAACTGGTACGGACAAAGGGGAATCTGACTGTAG
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
hhhhhhhhhhghhhhhhhhehhhedhhhhfhhhhh
```



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
hhhhhhhhhhghhhhhhehhhedhhhfhhhhh
```



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
hhhhhhhhhhghhhhhhehhhedhhhfhhhhh
```



Illumina Sequence ID Lines: A Decoder

@M01137:30:00000000-AA299:1:1101:10929:1966

M01137	the unique instrument name
30	the run id
00000000-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 (<i>paired-end or mate-pair reads only</i>)
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



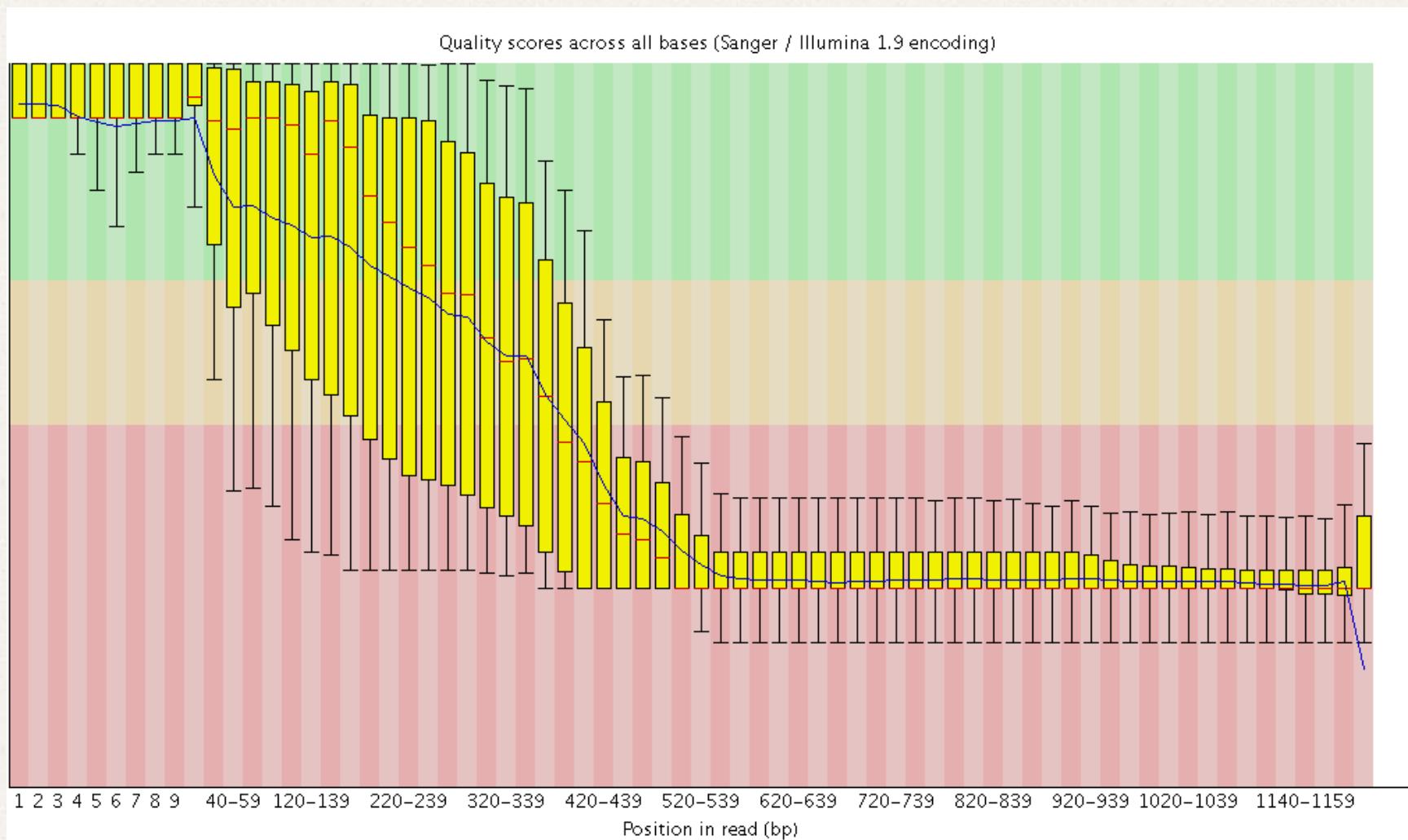
An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis

Cristian Del Fabbro¹✉, Simone Scalabrin²✉, Michele Morgante¹, Federico M. Giorgi^{1,3*}

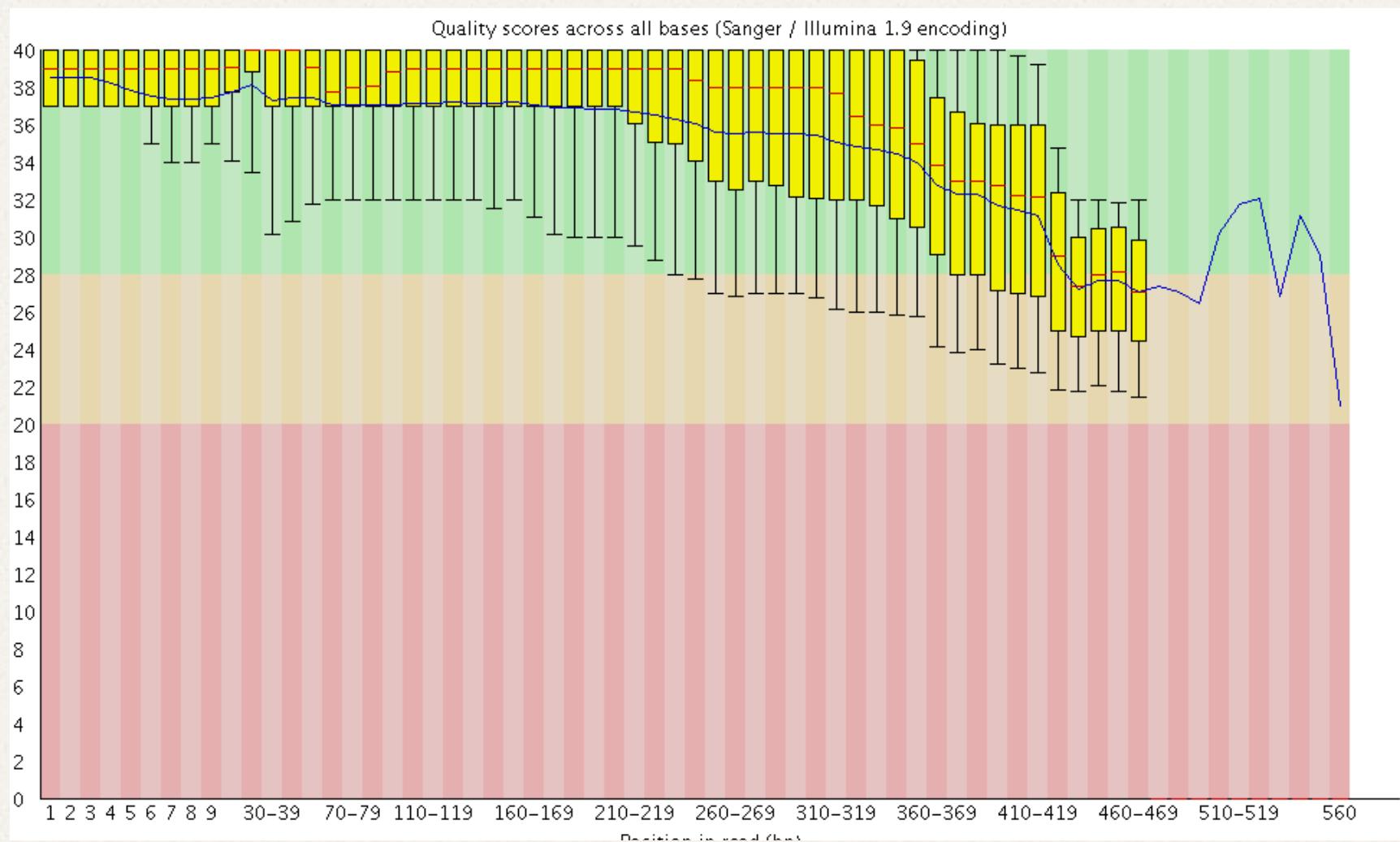
“Trimming is shown to increase the quality and reliability of the analysis, with concurrent gains in terms of execution time and computational resources needed“



Low Quality Sequences Before Trimming (Puma 454 sequences)



Same Sequences After Trimming (Puma 454 sequences)



Types of Trimming

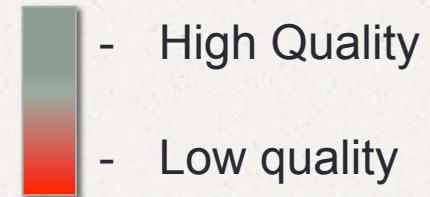


- High Quality
- Low quality

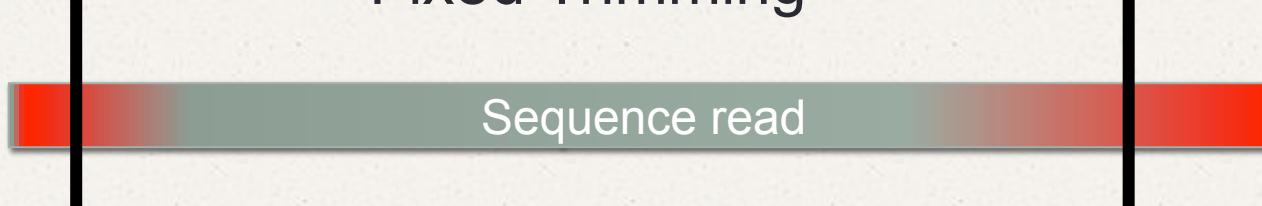
Sequence read



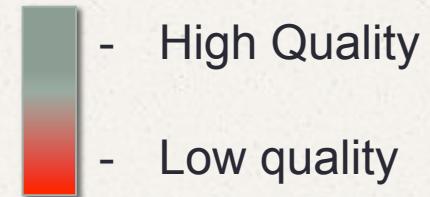
Types of Trimming



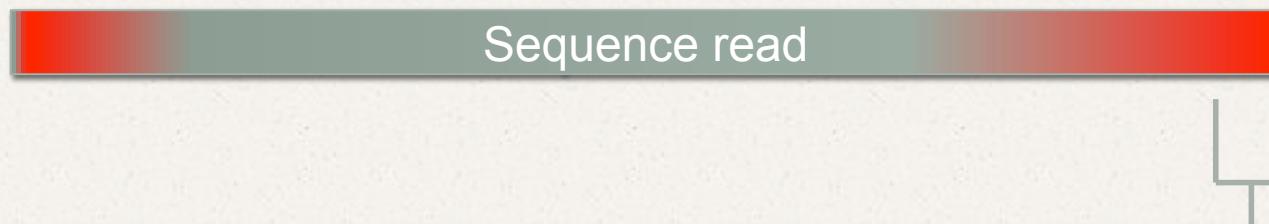
Fixed Trimming



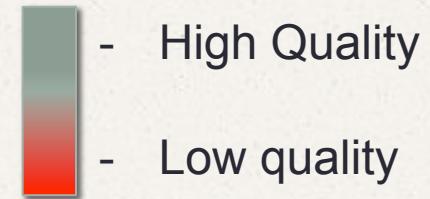
Types of Trimming



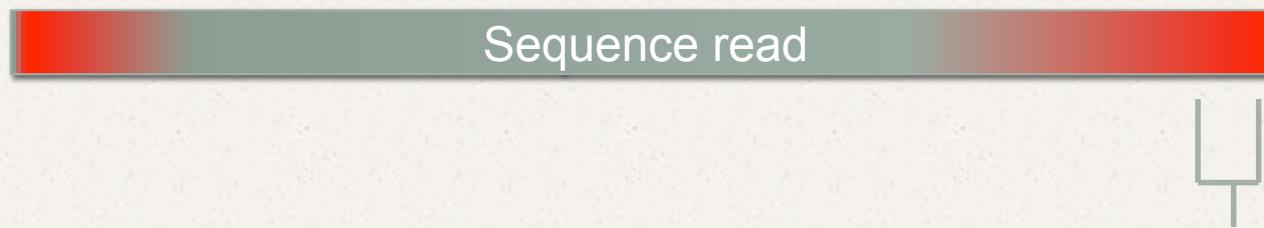
Sliding Window Trimming



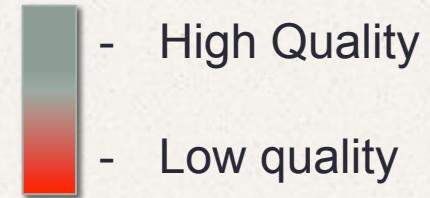
Types of Trimming



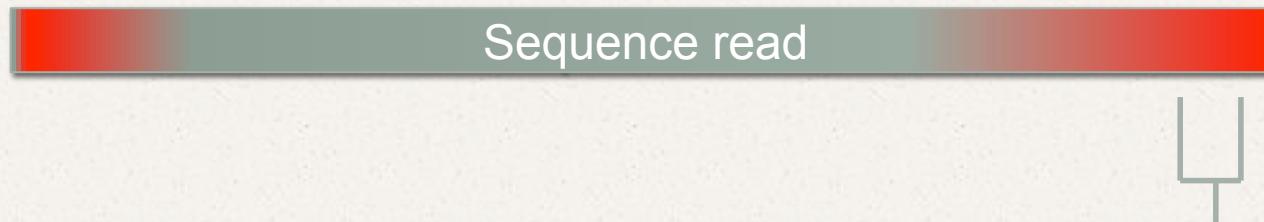
Sliding Window Trimming



Types of Trimming



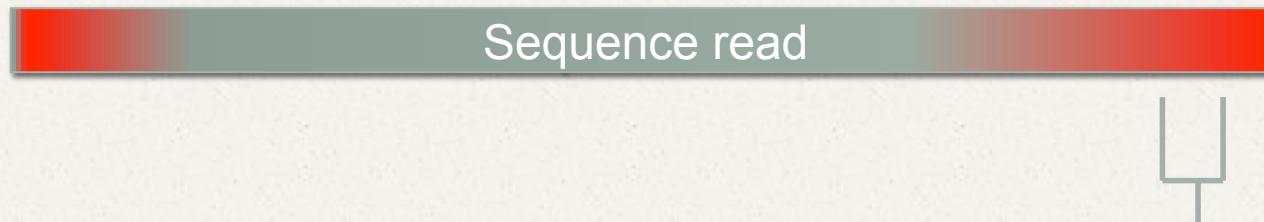
Sliding Window Trimming



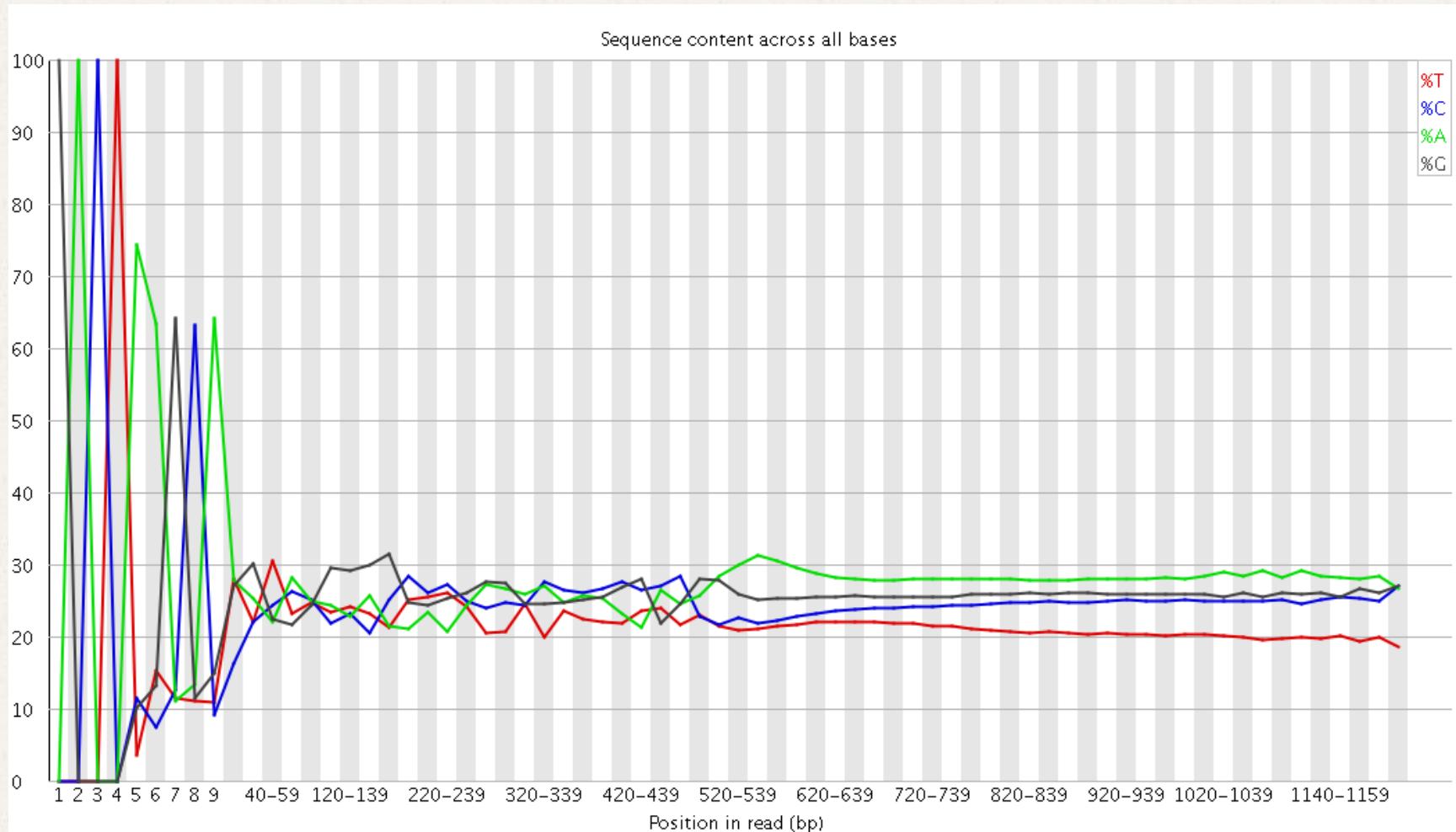
Types of Trimming

- High Quality
- Low quality

Sliding Window Trimming



Adapter Contamination



Adapter Contamination



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GACTAACCGAGTGGTATCAACGCCAGACTACATGGGACACTTGTTCTGAC	19391	5.415739186535921	No Hit
GACTAACCGAGTGGTATCAACGCCAGACTACATGGGACACTTGCTTCTGAC	11325	3.162974900083508	No Hit
GACTAACCGAGTGGTATCAACGCCAGACTACATGGGACACTTGTTCTGACA	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 AAGCAGTGGTATCAACGCCAGACTAC 29

Sbjct 1 AAGCAGTGGTATCAACGCCAGACTAC 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. Yorke³

“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

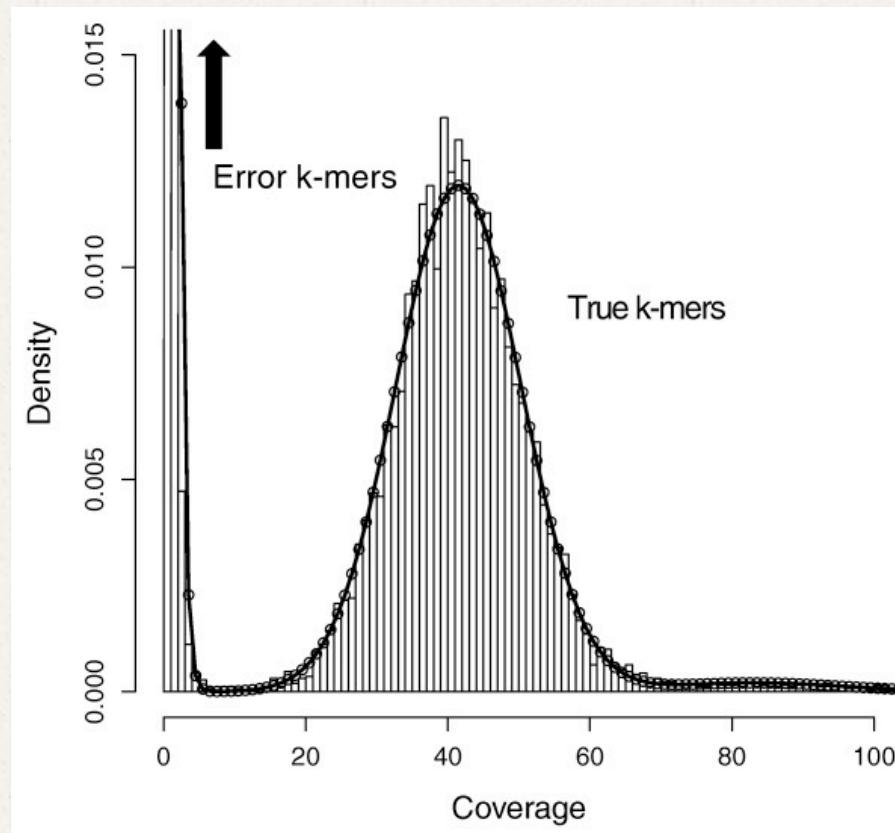


AGCTGT
GCTGTG
CTGTGG



K-mers and Error-correction

- Expected Distribution of k-mer frequency



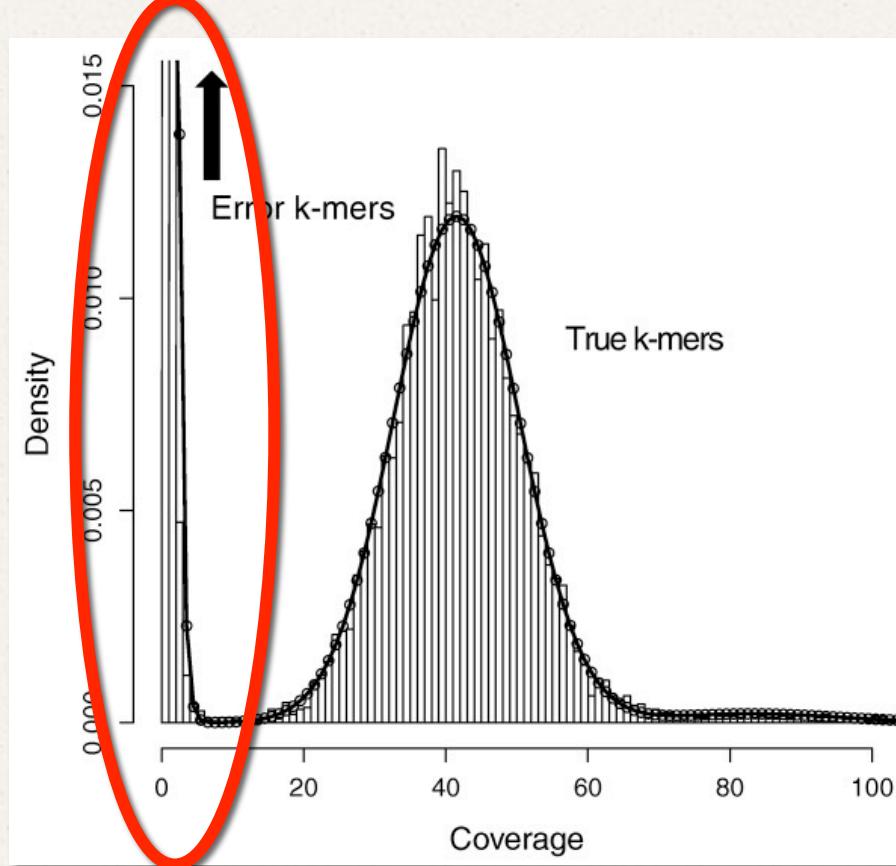
DSK; Rizk et al. 2013



K-mers and Error-correction

- Expected Distribution of k-mer frequency

Corrected



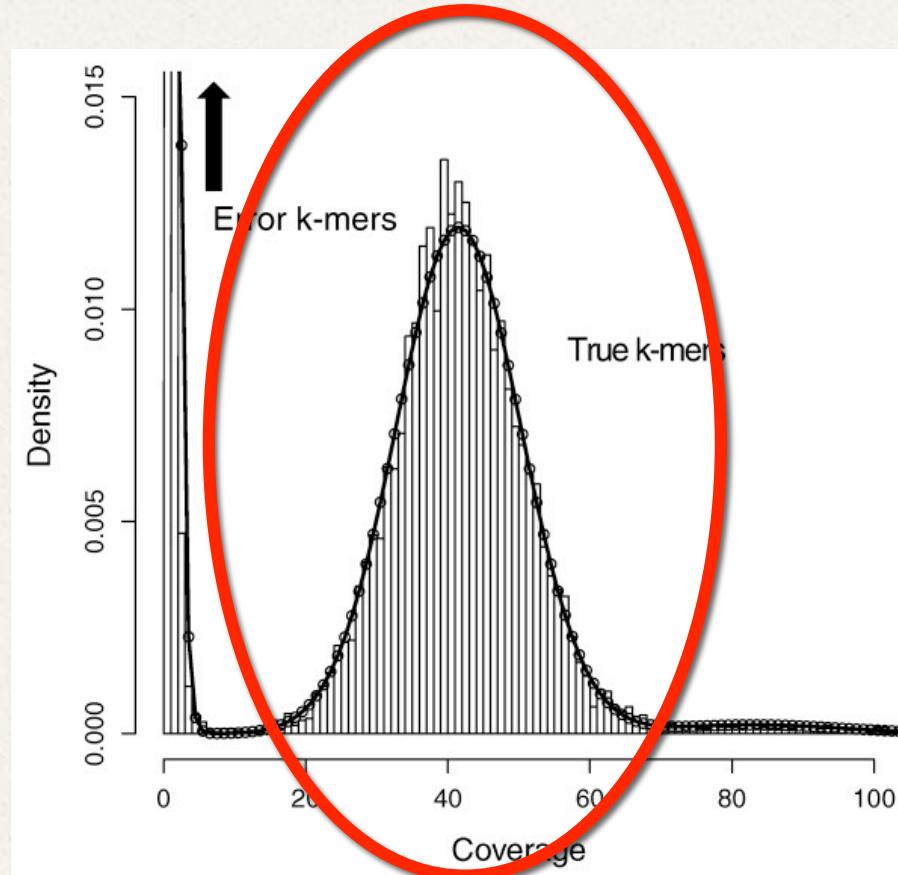
DSK; Rizk et al. 2013



K-mers and Error-correction

- Expected Distribution of k-mer frequency

Estimate genome size



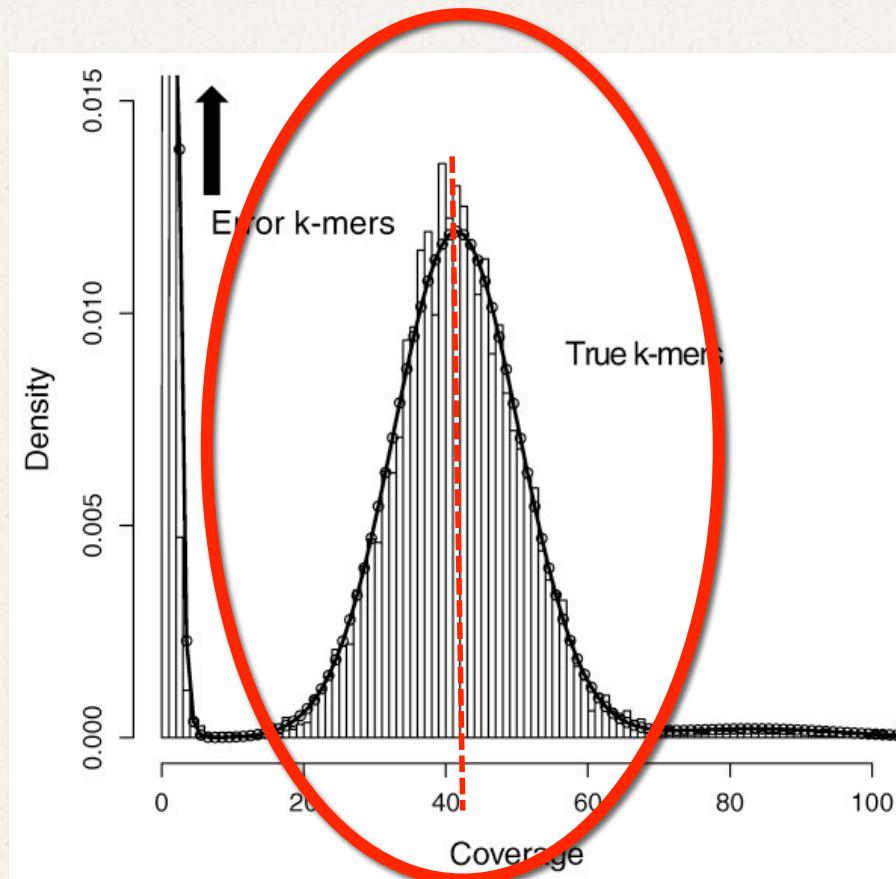
DSK; Rizk et al. 2013



K-mers and Error-correction

- Expected Distribution of k-mer frequency

Estimate genome size



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

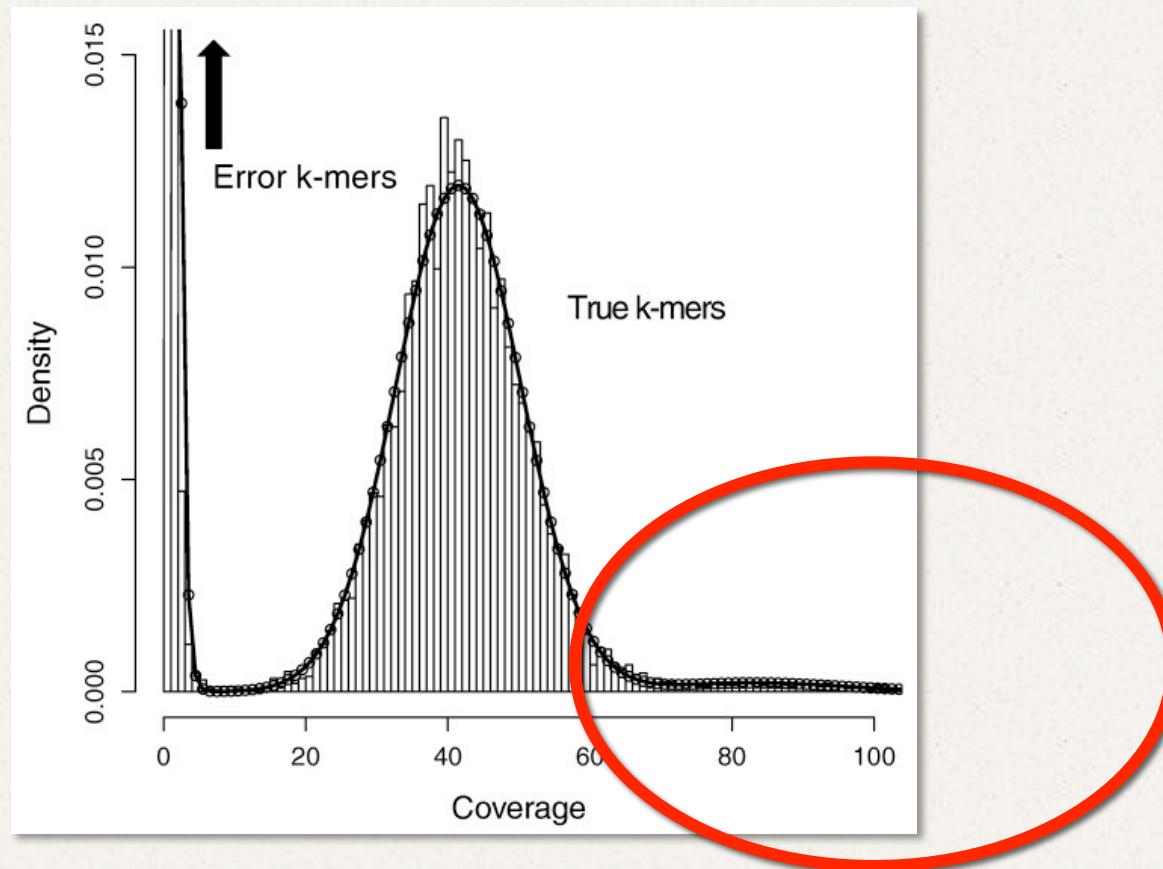
DSK; Rizk et al. 2013



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!

