SciLifeLab

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

K-mer and contaminant analysis

elixir

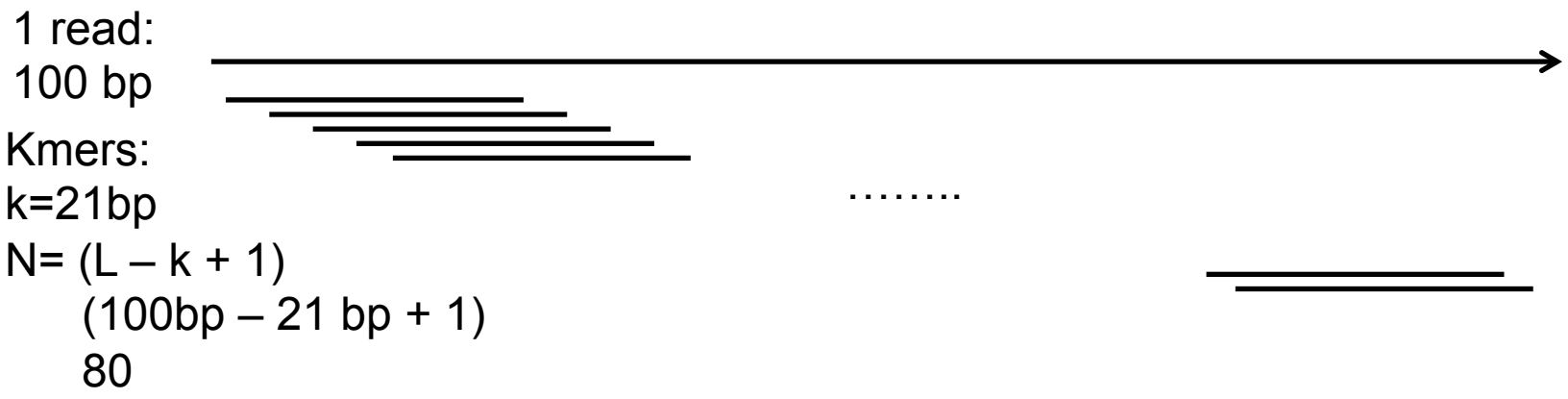
excelerate

K-mer Analysis

SciLifeLab

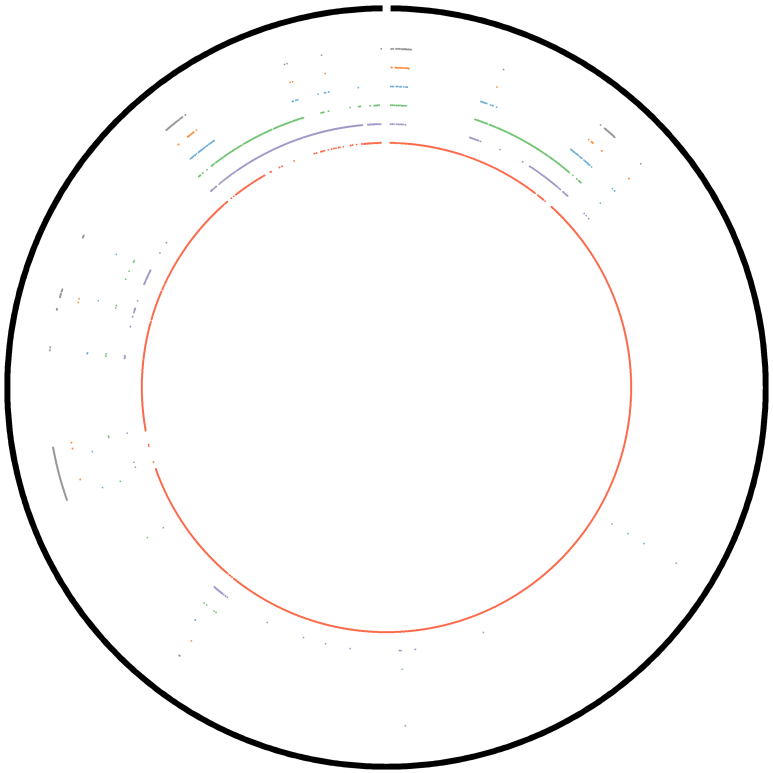
- What is a k-mer?
  - A k-mer is a sequence of nucleotides of length k.
  - Examples of a 6-mer
    - ACGTCT
    - TGACTA
    - GATCCC

- A read of length L has L-k+1 k-mers.



K-mer Analysis

How many k-mers are distinct in your genome?

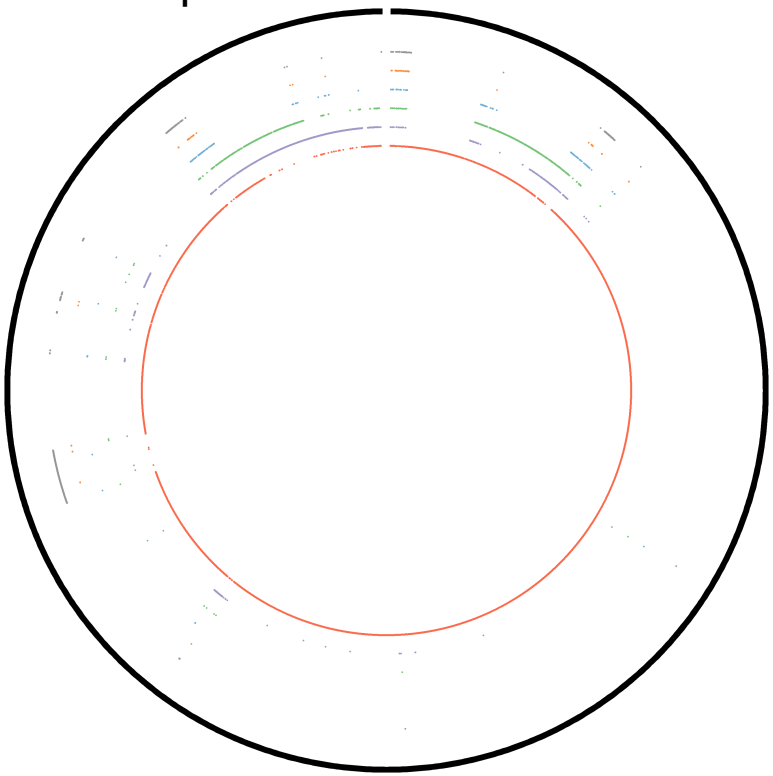


- Chr 1 of yeast strain s288c
  - Haploid
  - Linear, but plotted as circle
  - Black: position along chromosome
  - K-mer frequency in genome
    - K=27
    - Red: 1
    - Purple: 2
    - Green: 3
    - Blue: 4
    - Orange: 5
    - Grey: 6+
  - Majority of k-mers are distinct.

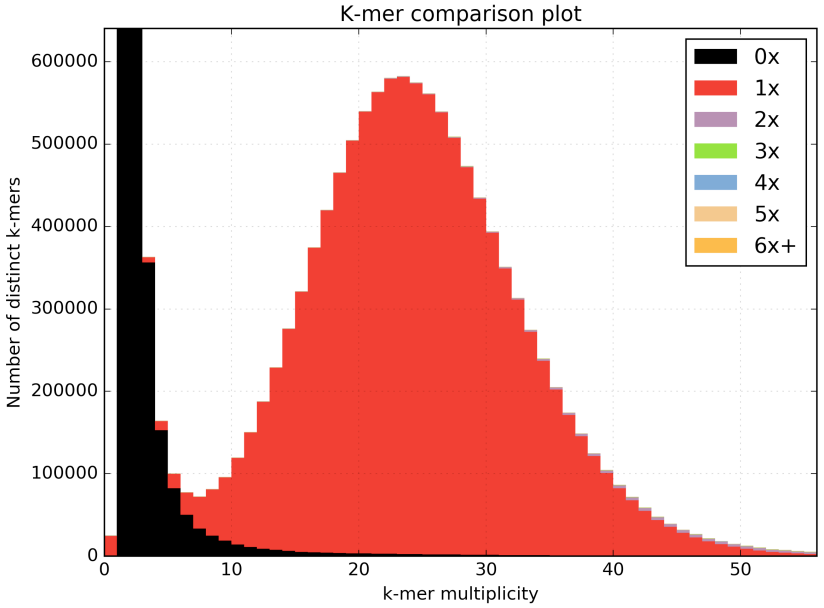
K-mer Analysis

SciLifeLab

Frequency in the genome
Haploid



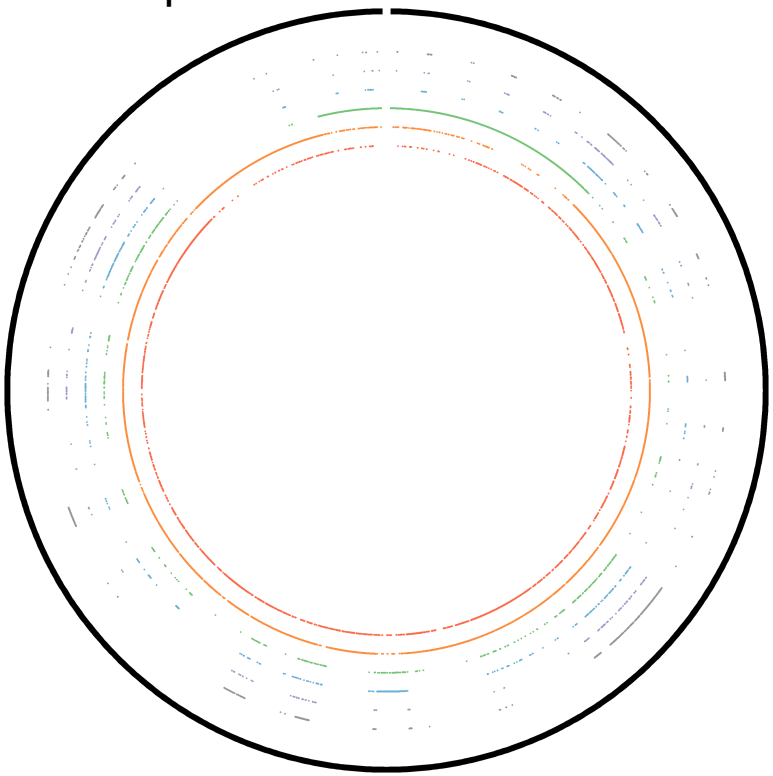
Frequency of those k-mers in the sequence data set



x-axis: Frequency of the distinct k-mer in the data
y-axis: Count of distinct k-mers with frequency x

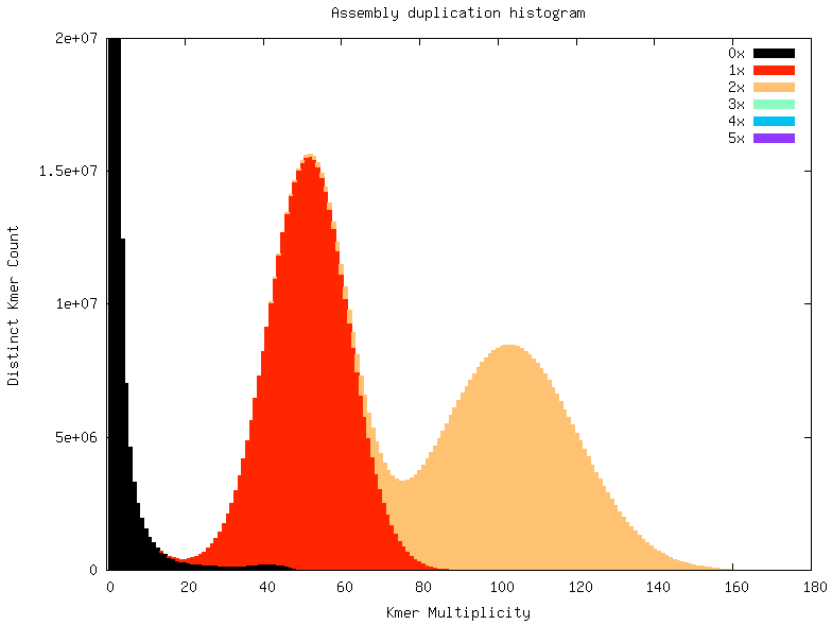
K-mer Analysis

Frequency in the genome
Diploid



Contig from A.thaliana phased genome

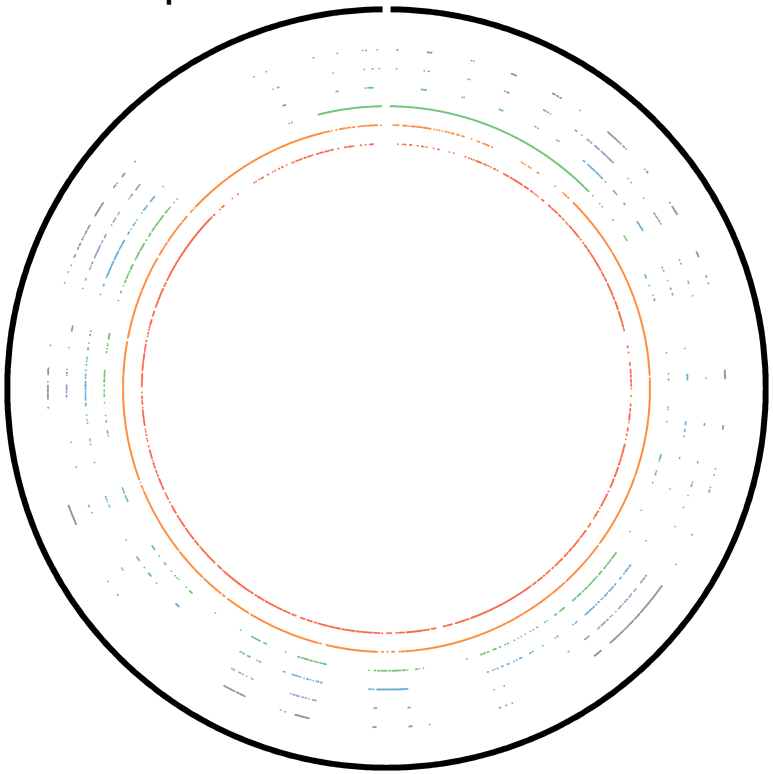
Frequency of those k-mers in the sequence data set



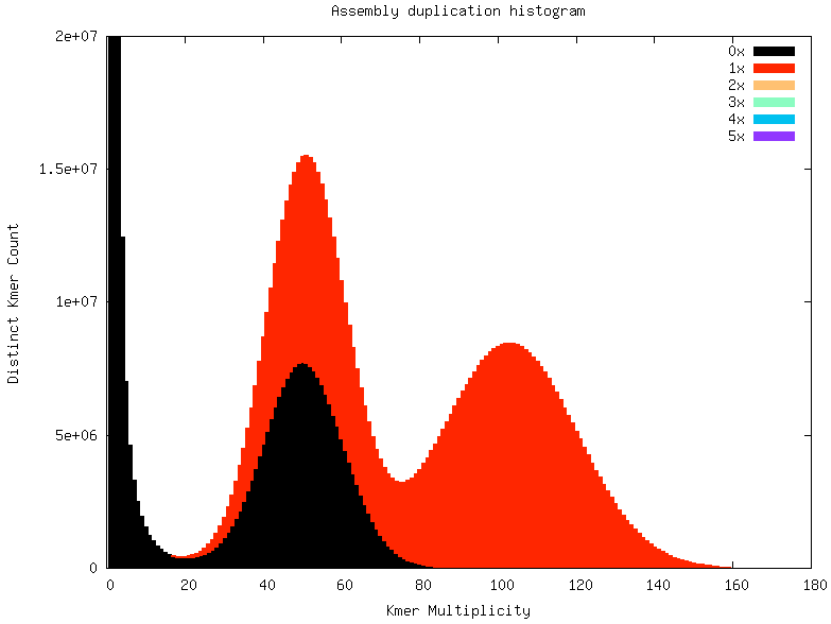
x-axis: Frequency of the distinct k-mer in the data
y-axis: Count of distinct k-mers with frequency x

K-mer Analysis

Frequency in the genome
Diploid



Frequency of those k-mers in the sequence data set

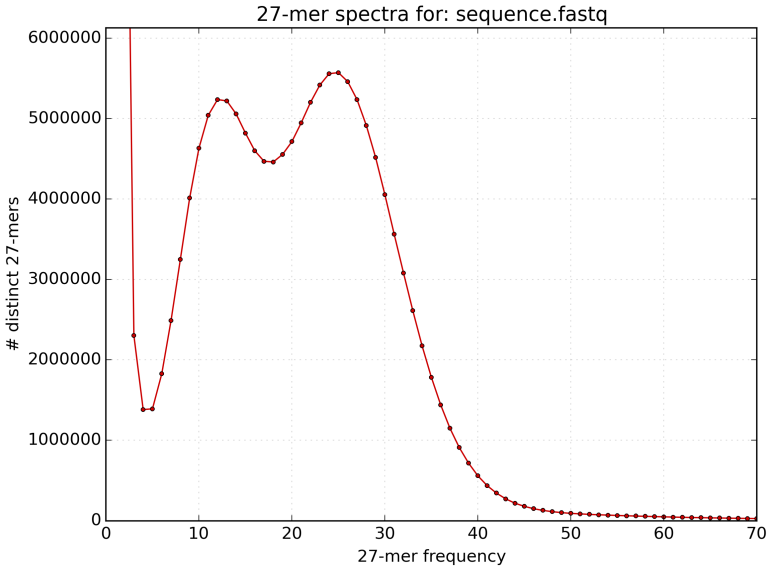


x-axis: Frequency of the distinct k-mer in the data
y-axis: Count of distinct k-mers with frequency x

Estimating Genome Size

SciLifeLab

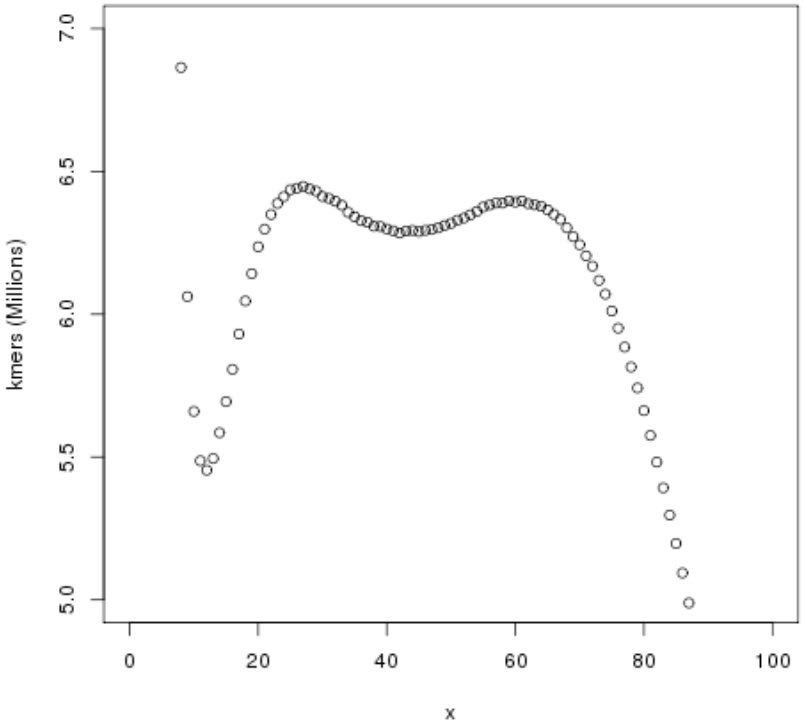
- We can use coverage to estimate genome size
  - The peak with the largest k-mer multiplicity is the mean k-mer coverage across the genome.
  - N = M * L / (L - K + 1)
    - N is Depth of Read Coverage
    - M is mean k-mer coverage
    - L is read length
    - K is k-mer size
  - G = T / N
    - G is the genome size
    - T is the total number of bases

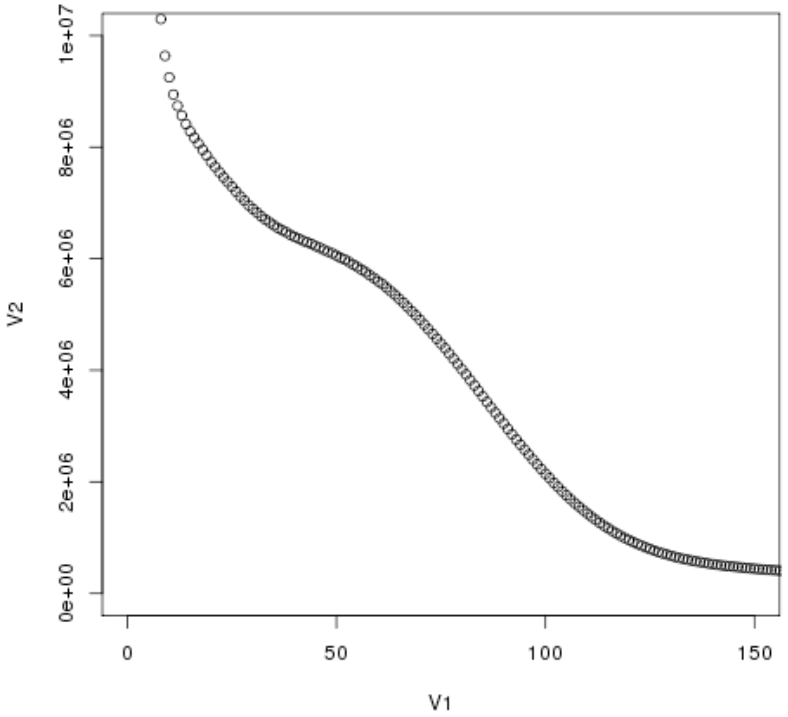


Estimating Genome Size

SciLifeLab

- Not so easy: estimating complexity





Estimating Genome Size

SciLifeLab

- Distribution decomposition analysis
  - **`kat_distanalysis.py --plot kat.hist`**

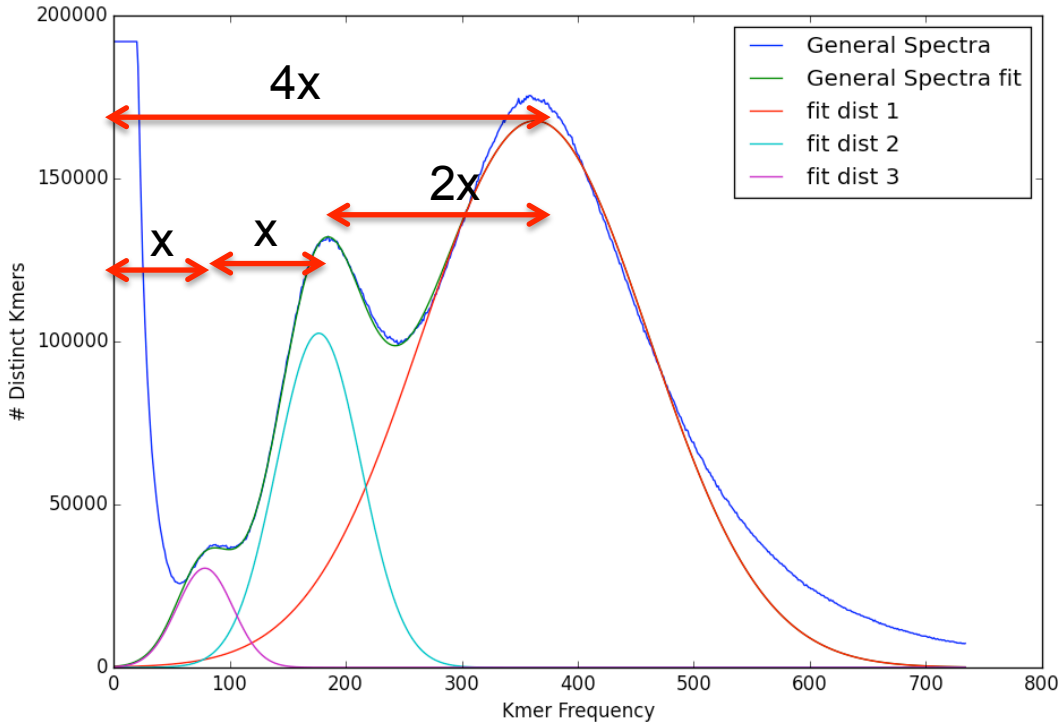
```
Main spectra statistics
-----------------------
K-value used: 27
Peaks in analysis: 3
Index   Freq    Max     Volume
1       79      35596   1844622
2       177     132229  9354614
3       364     175454  41202694
Mean k-mer frequency: 320x
Homozygous peak index: 3
Estimated genome size: 48.05 Mbp
Estimated heterozygous rate: 0.86 %
Estimated assembly completeness: 95.73%

Breakdown of copy number composition for each peak
--------------------------------------------------

---- Report for f=364.209 (total elements 36483749)----
 0x: No significant content
 1x: 100.00% (36483749 elements at f=367.62)
 2x: No significant content
 3x: No significant content

---- Report for f=177.838 (total elements 9214569)----
 0x: 47.33% (4361617 elements at f=183.02)
 1x: 52.67% (4852951 elements at f=176.43)
 2x: No significant content
 3x: No significant content

---- Report for f=79.618 (total elements 858966)----
 0x: 100.00% (858966 elements at f=77.18)
 1x: No significant content
 2x: No significant content
 3x: No significant content
```

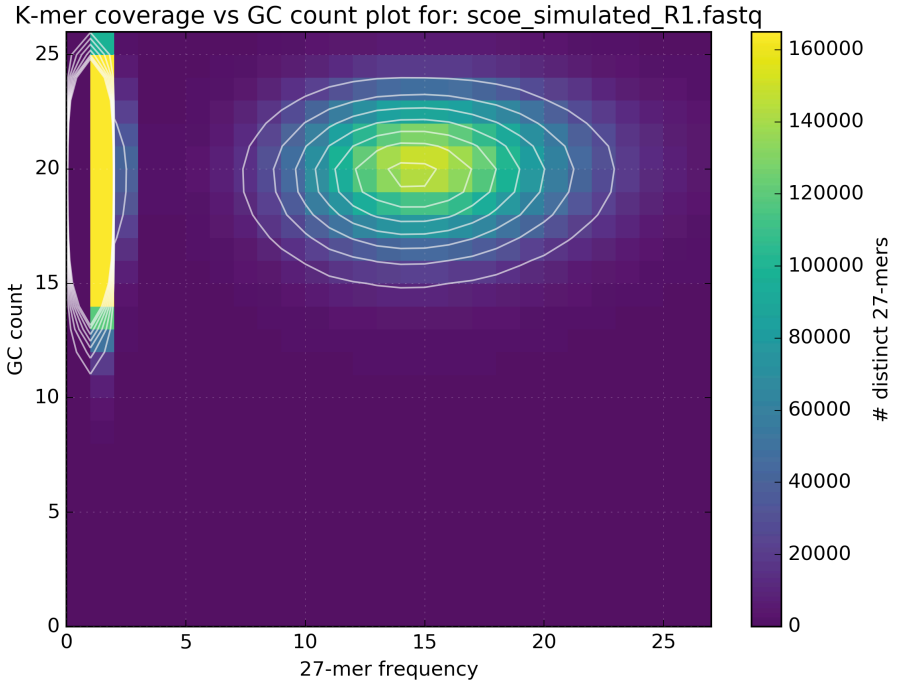


Tetraploid genome copy number spectra

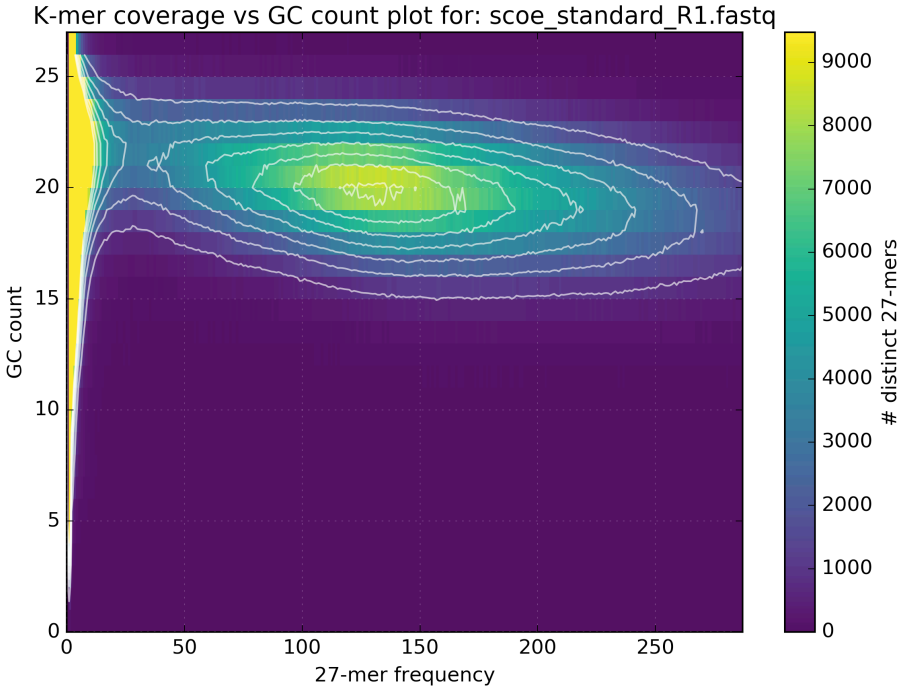
GC Content in depth

- Monitor GC Bias

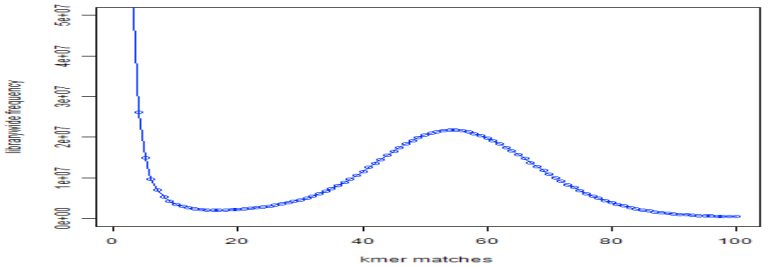
Simulated data



GC bias in Standard Illumina protocol



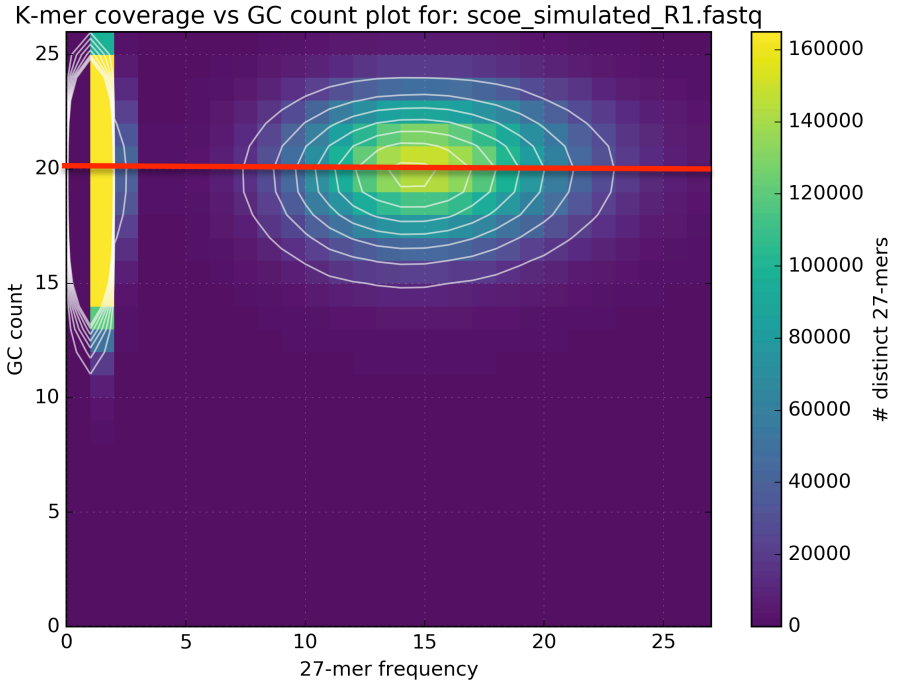

```
ACTCAGGATTA   GC=4 Count=1
AATAGCCGGGG   GC=7 Count=2
```



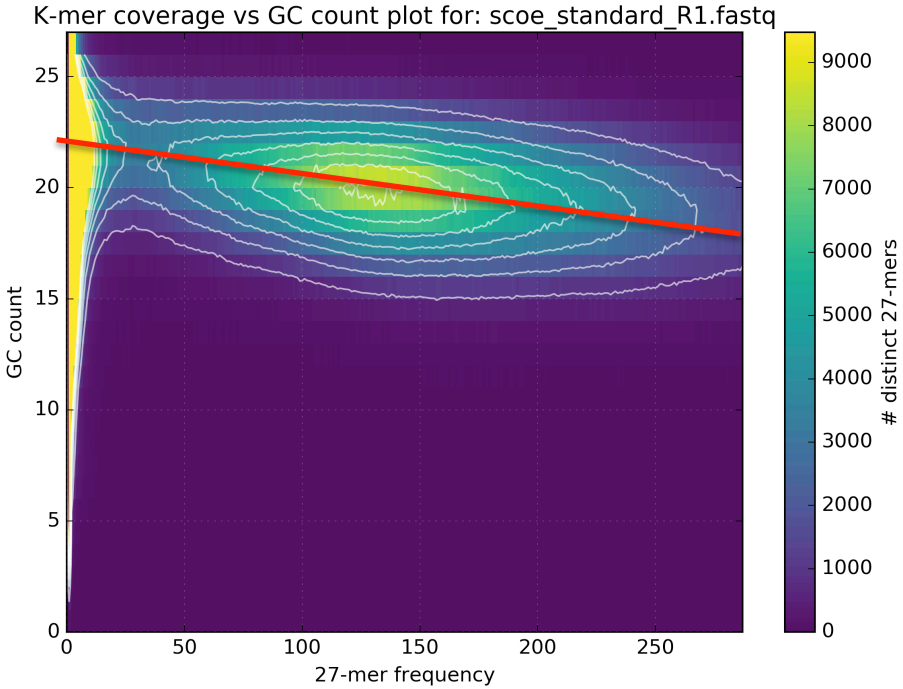
GC Content in depth

- Monitor GC Bias

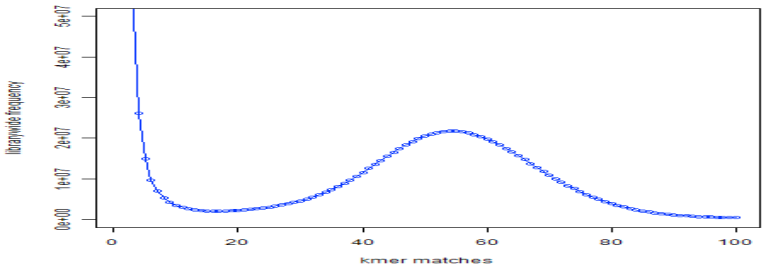
Simulated data



GC bias in Standard Illumina protocol



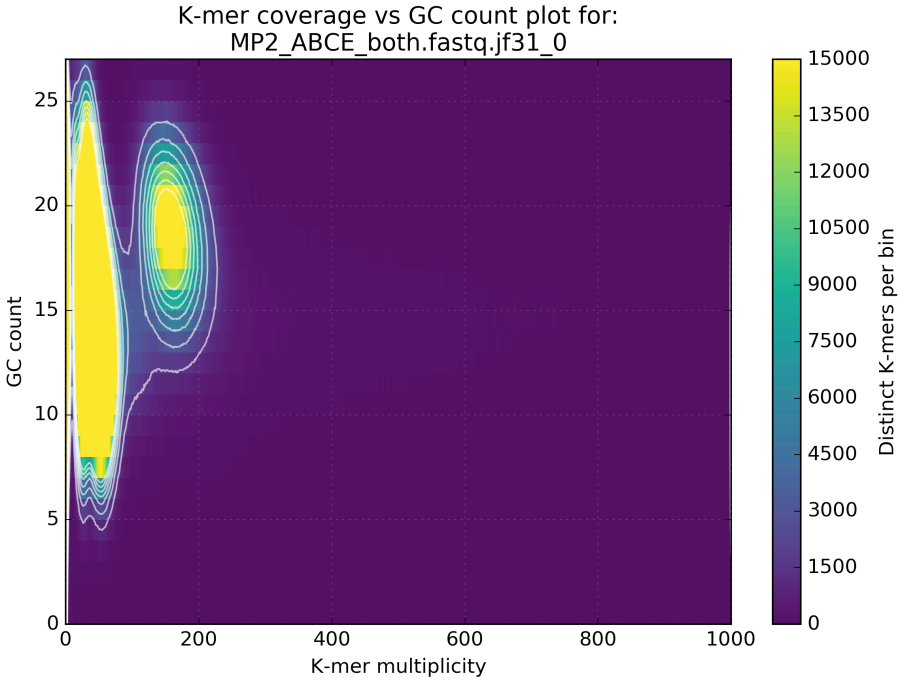

```
ACTCAGGATTA   GC=4 Count=1
AATAGCCGGGG   GC=7 Count=2
```

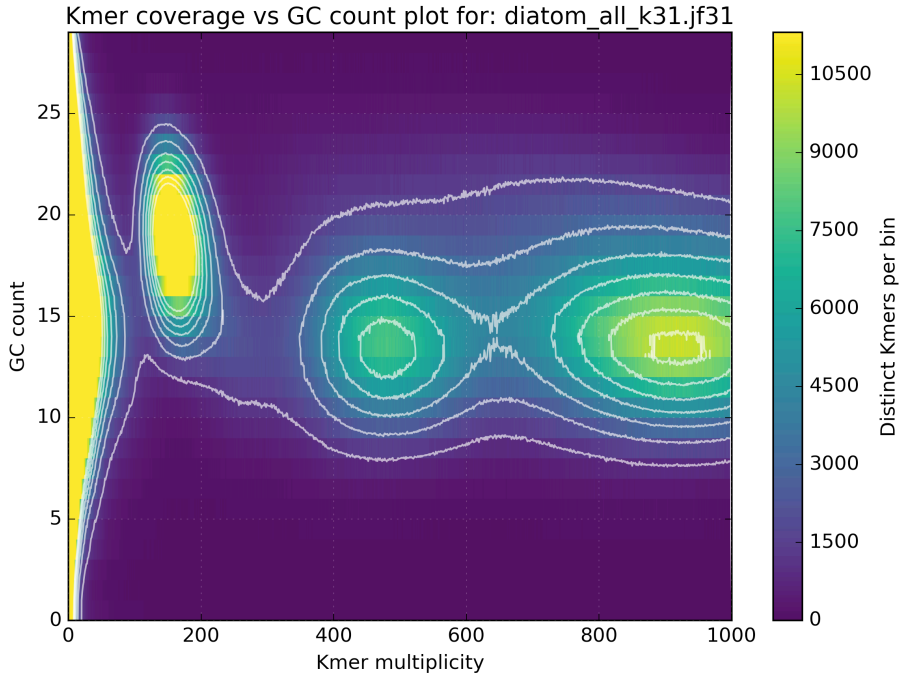


GC Content in depth

SciLifeLab

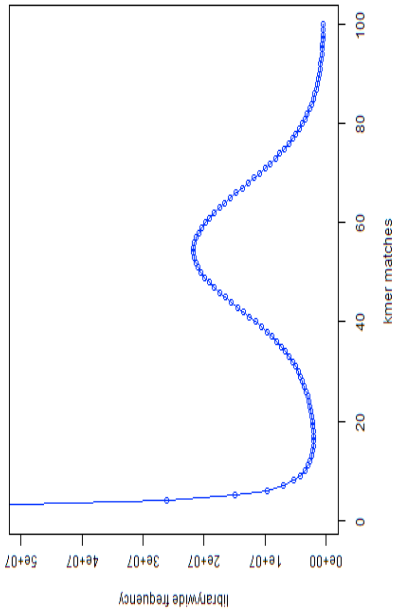
- Uncover contamination
  - Separate bacteria from eukaryote
  - Separate organelle from nuclear genome

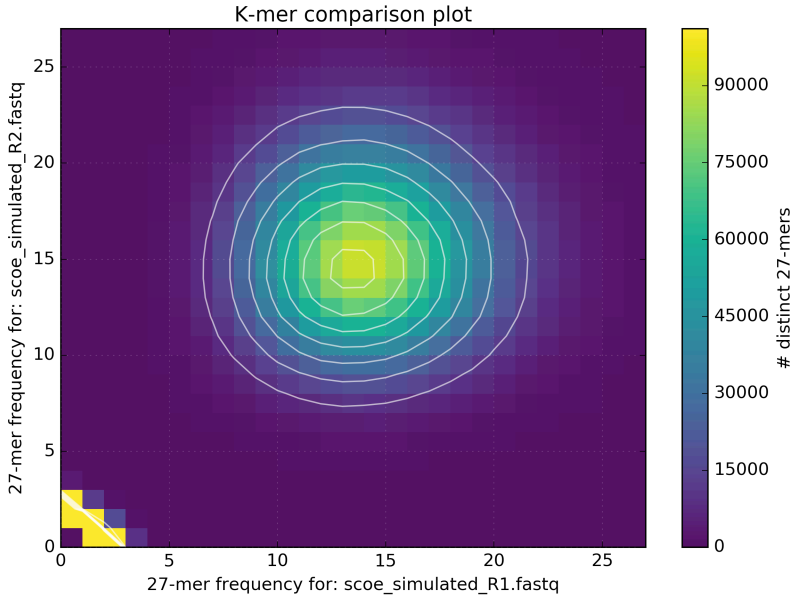


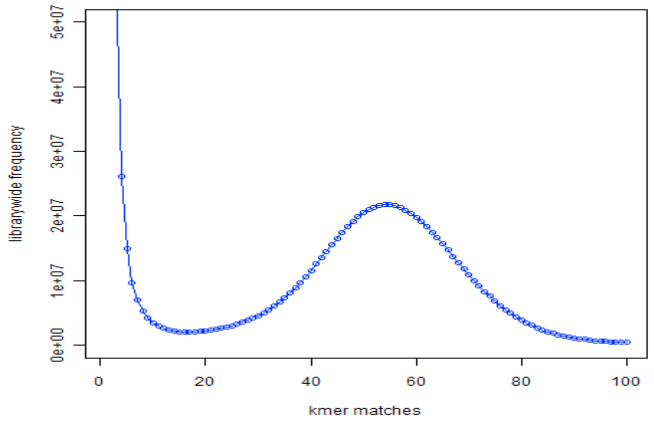


What can k-mers tell us?

- Comparing k-mer counts between data reveals biases
  - R1 vs R2
  - Lib1 vs Lib2

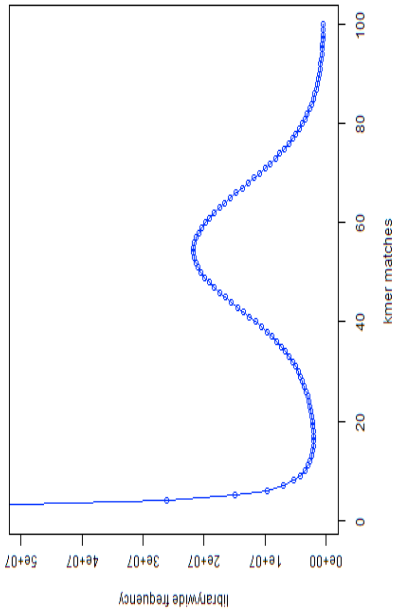


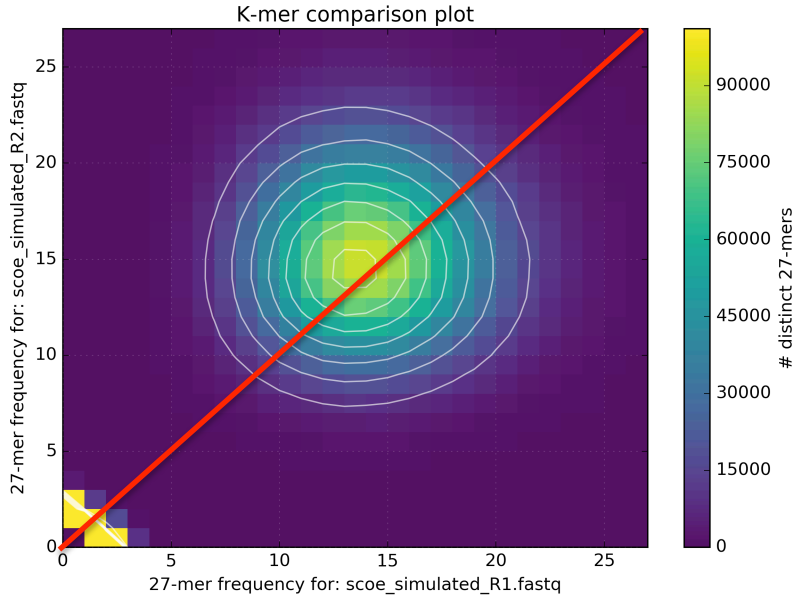


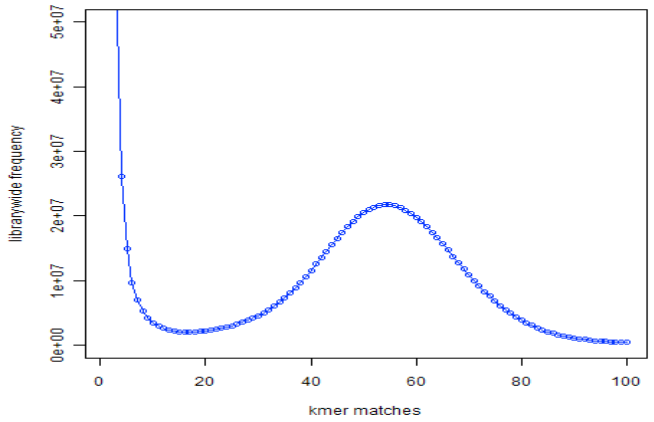


What can k-mers tell us?

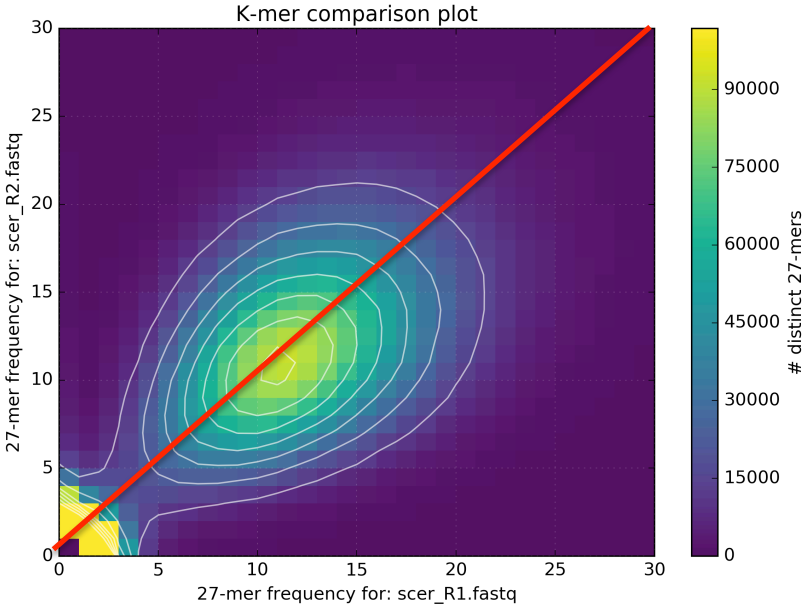
- Comparing k-mer counts between data reveals biases
  - R1 vs R2
  - Lib1 vs Lib2



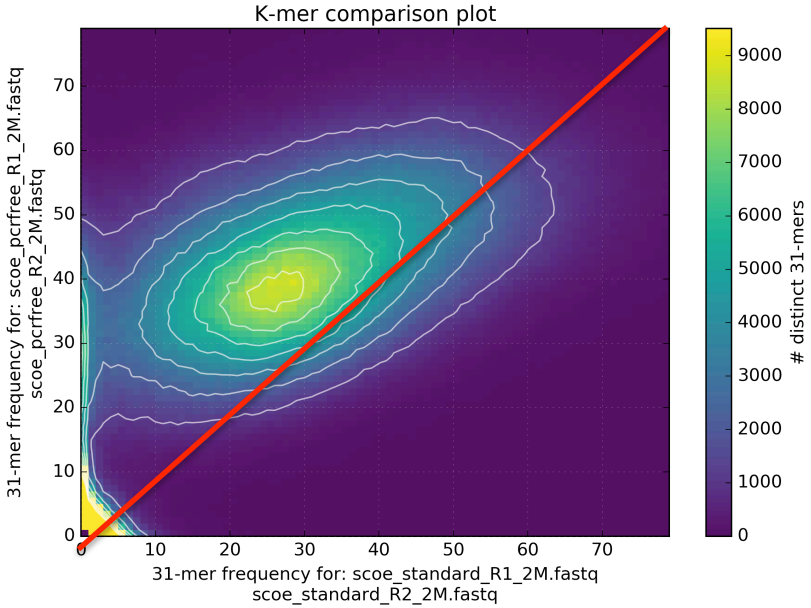




Data comparison



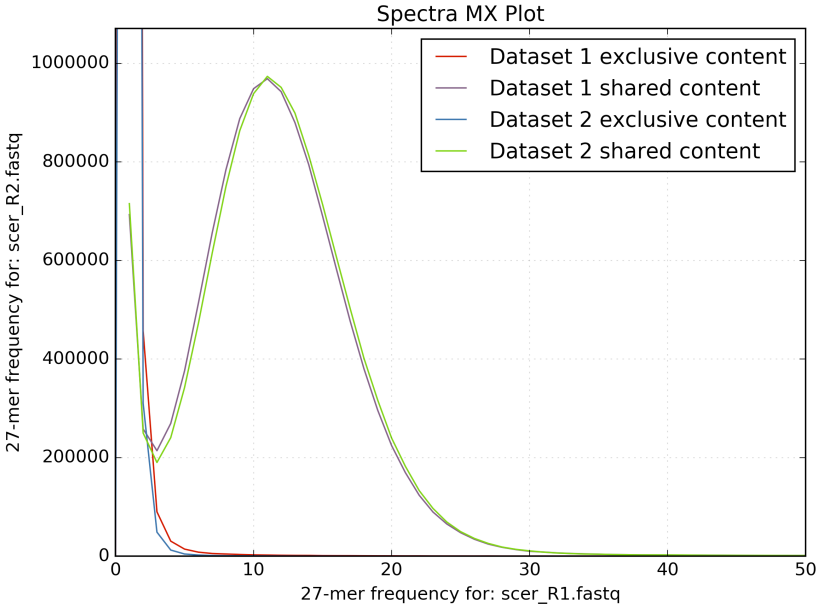
R1 vs R2

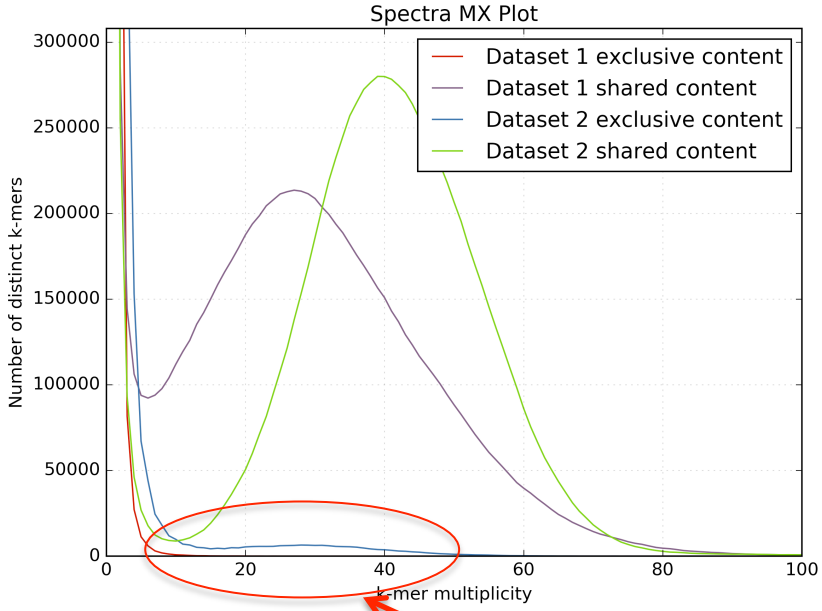


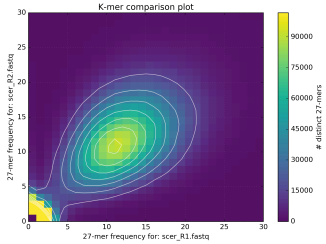
Standard vs PCR free

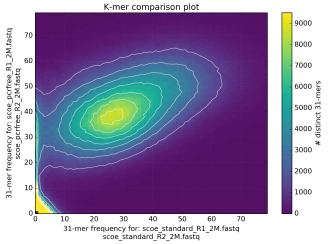
- PCR free captures data missing in standard protocol

Data comparison







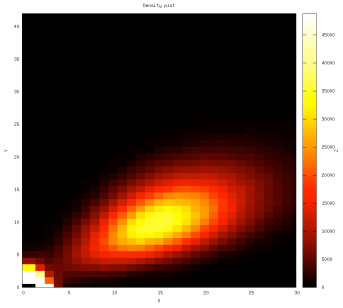


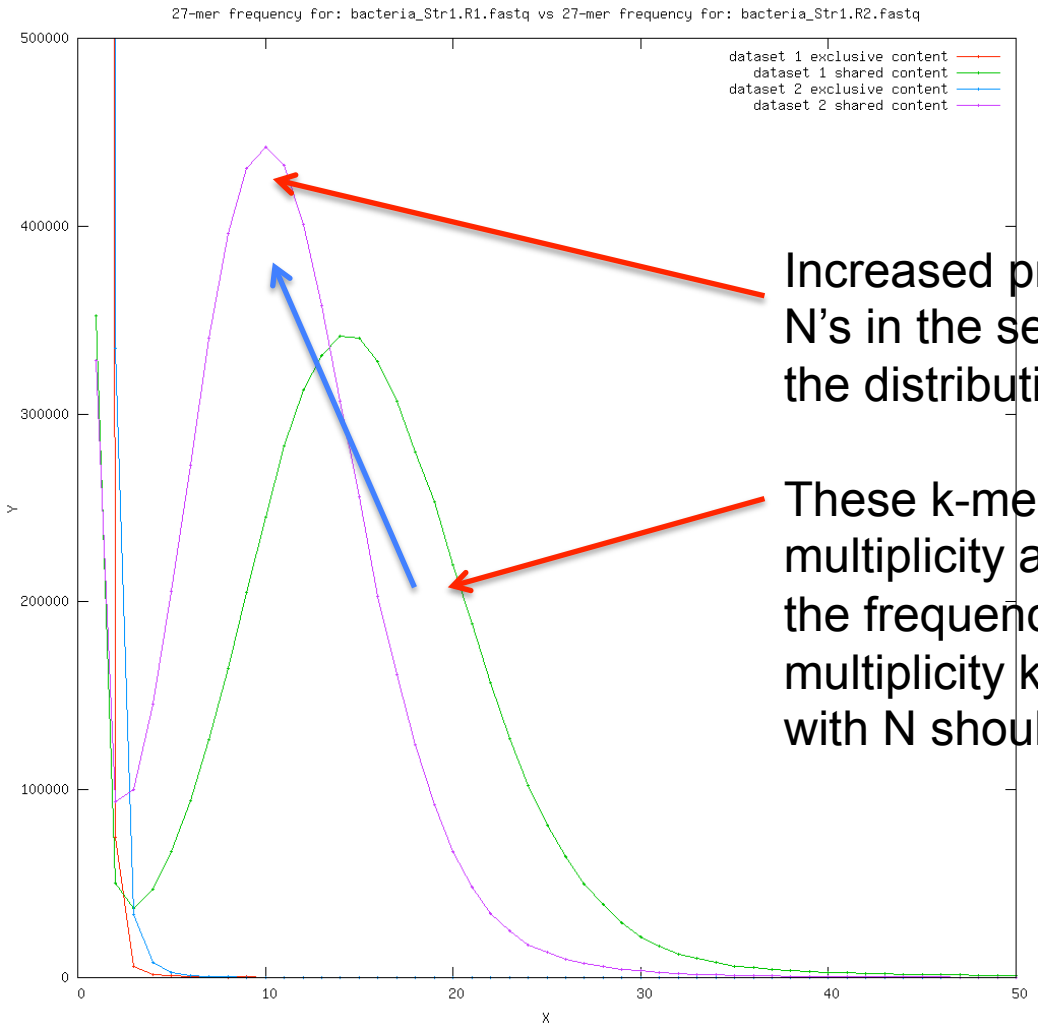
Low proportion of k-mers present only in dataset 2

Data comparison

SciLifeLab

R1 v R2





Lower quality dataset = More N's in the sequence

Increased prevalence of N's in the sequence shift the distribution.

These k-mers have lower multiplicity and so increase the frequency of lower multiplicity k-mers. K-mers with N should not be counted

K-mer Filtering

SciLifeLab

- K-mer based analyses provides a lot of information about the information within your data.

- K-mer filtering can be used to assist exploratory analyses.
  - Digital normalization can help to process uneven coverage.
  - Error correction can be used to remove low frequency errors.
  - Frequency based filtering can separate organelles or contamination for assembly.

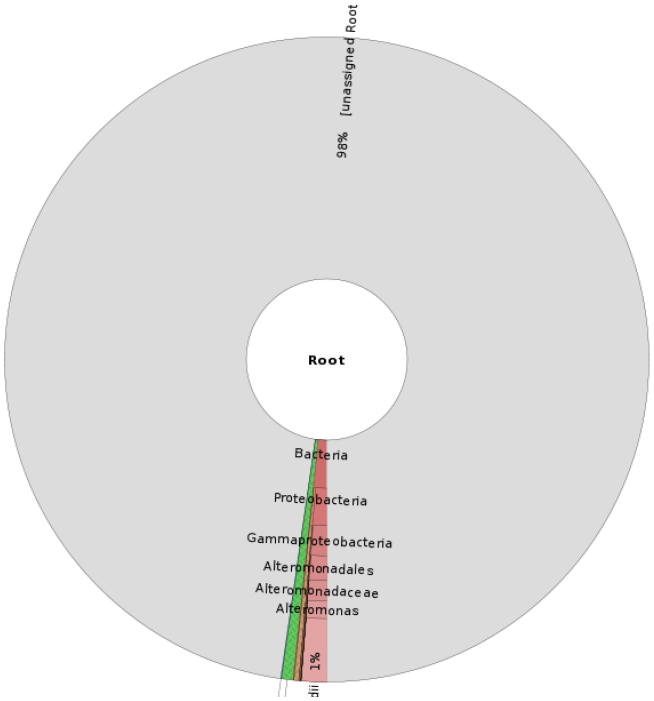
PacBio and Nanopore Quality

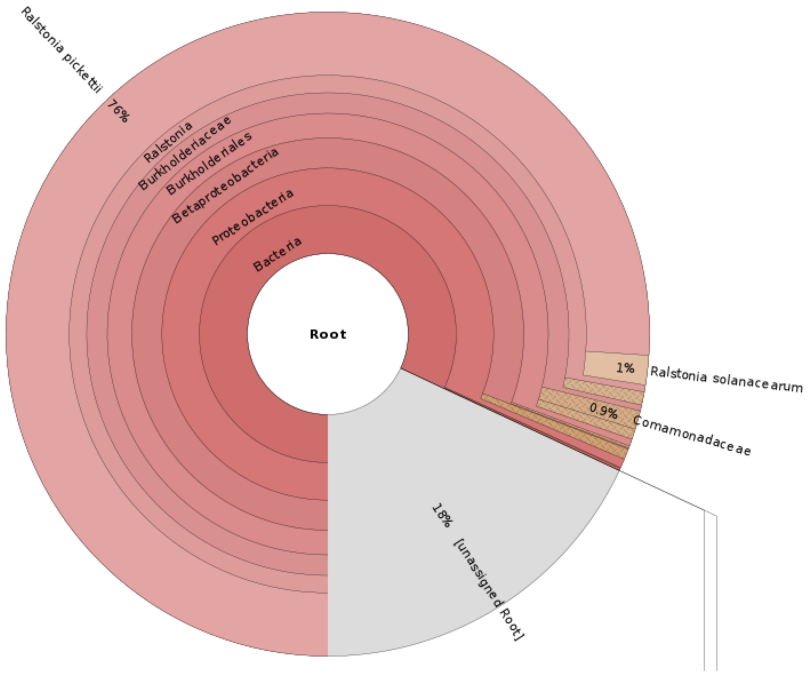
SciLifeLab

- PacBio and Nanopore have greater error rates per base. This precludes many analyses that one can do with Illumina data.
  - K-mer analyses are not feasible because there are too many unique k-mers in the reads.
    - Reads of Insert / CCS reads require at least 30x coverage in order to increase the accuracy to the necessary threshold.
  - Biases and data differences are more difficult to detect.
  - K-mer based contamination analyses have lower accuracy.

Contamination Analyses

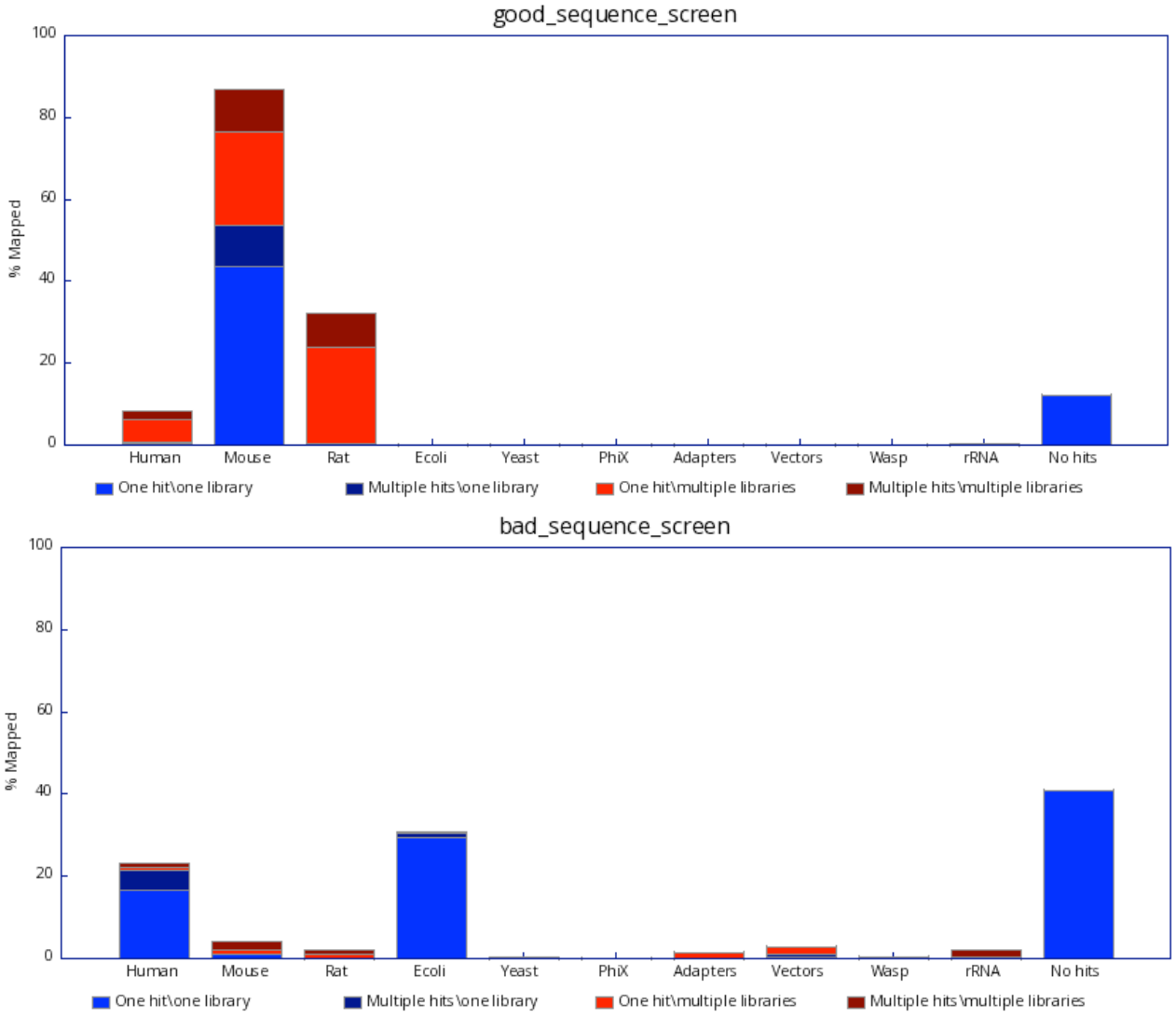
- Read based contamination analyses are tricky
  - Entirely dependent on your reference database
  - Short k-mer matching increases alignment to multiple targets
  - Unrelated organisms can contain similar strings of nucleotides





FastQ Screen

SciLifeLab



Summary

SciLifeLab

- K-mer analyses are limited by:
  - sufficient depth of coverage
  - sequence error rates

- K-mer analyses can:
  - Help estimate genome size and infer ploidy
  - Detect library biases within and between data sets
  - Help find contaminants

- K-mer based filtering can make data easier to work with

- Contamination assessment:
  - Entirely dependent on your subject database
  - Loss of accuracy with shorter strings