SciLifeLab

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

K-mer Analyses, Contamination Detection, And Mapping Based Analyses

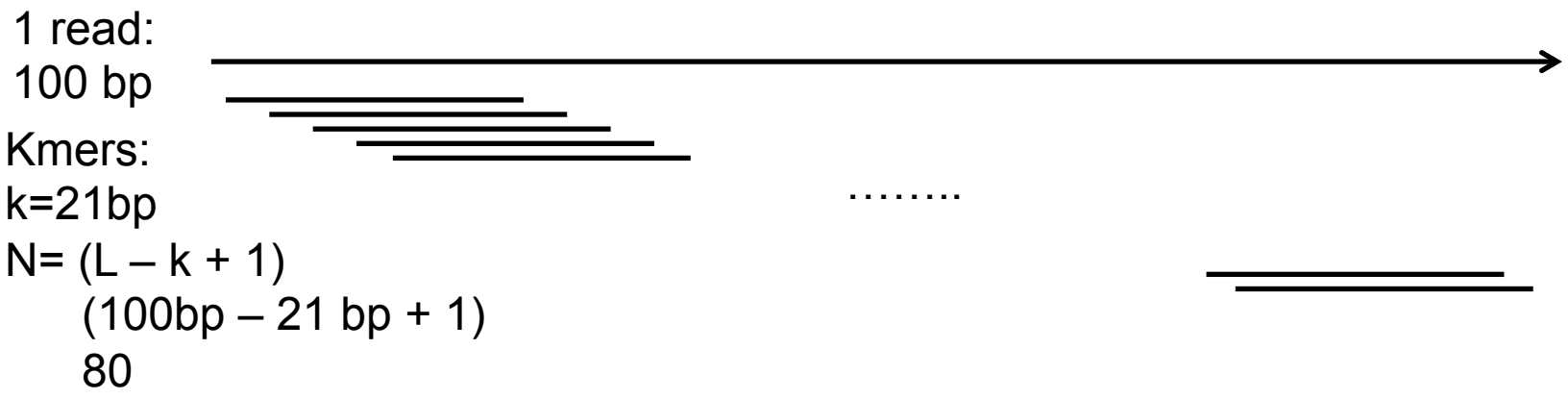
elixir excelerate

K-mer Analysis

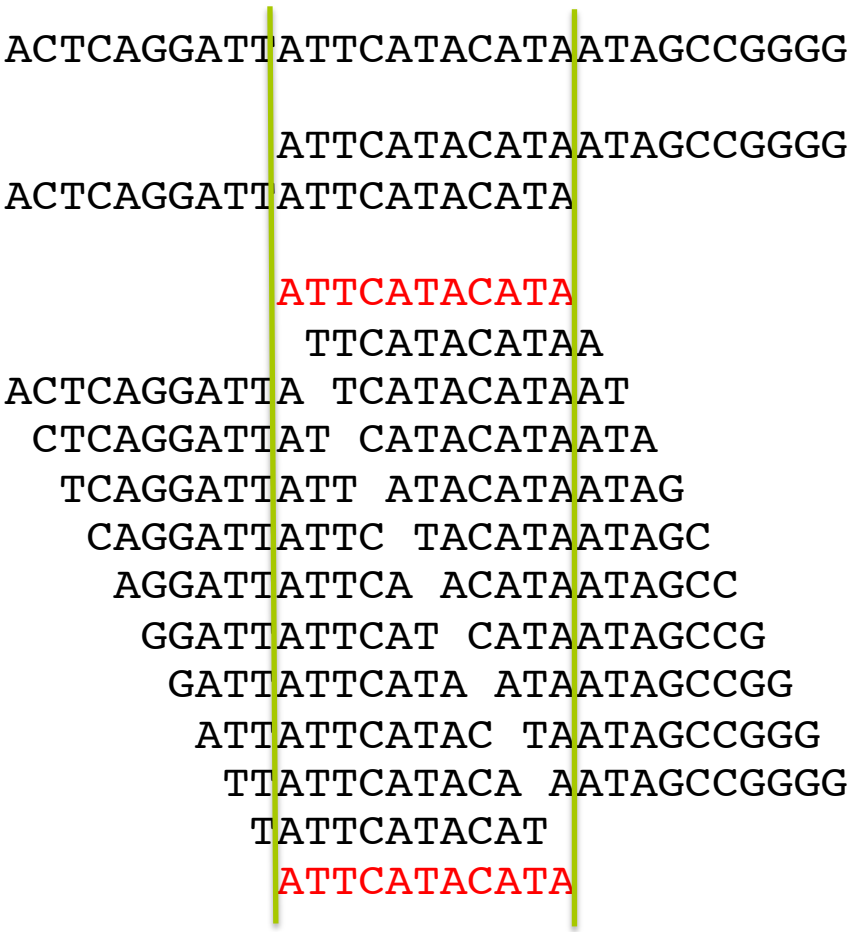
SciLifeLab

- 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



- The sequence has 21 distinct 11-mers.
- The reads have 21 distinct 11-mers even though 22 11-mers are generated.
- The frequency of a distinct k-mer increases where reads overlap

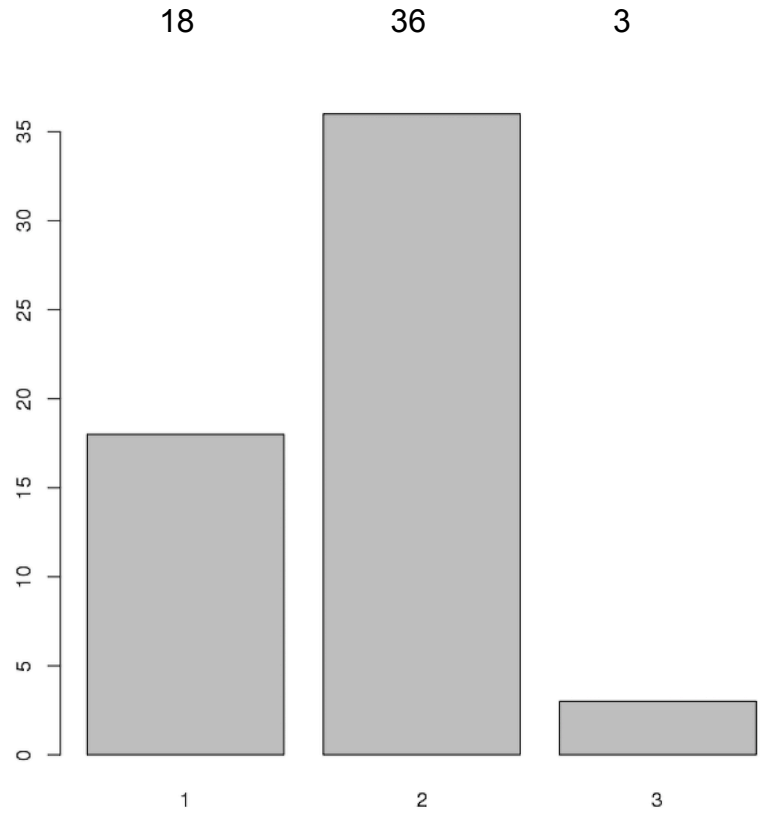
K-mer Analysis

- The sequence has 21 distinct 11-mers.
- The reads have 21 distinct 11-mers even though 33 11-mers are generated.
- The frequency of a distinct k-mer increases where reads overlap

K-mer Analysis

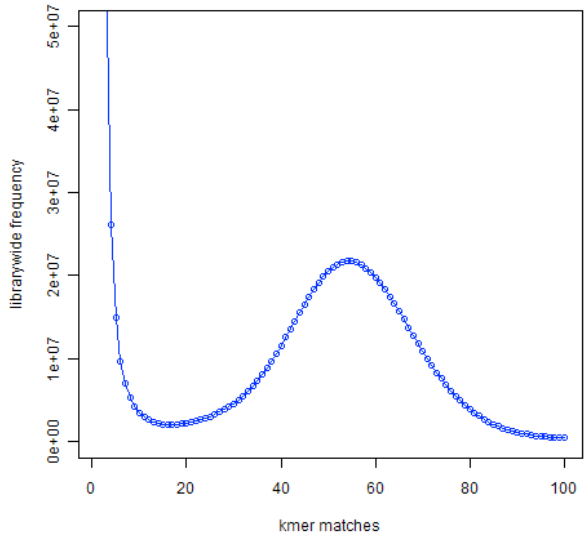
SciLifeLab

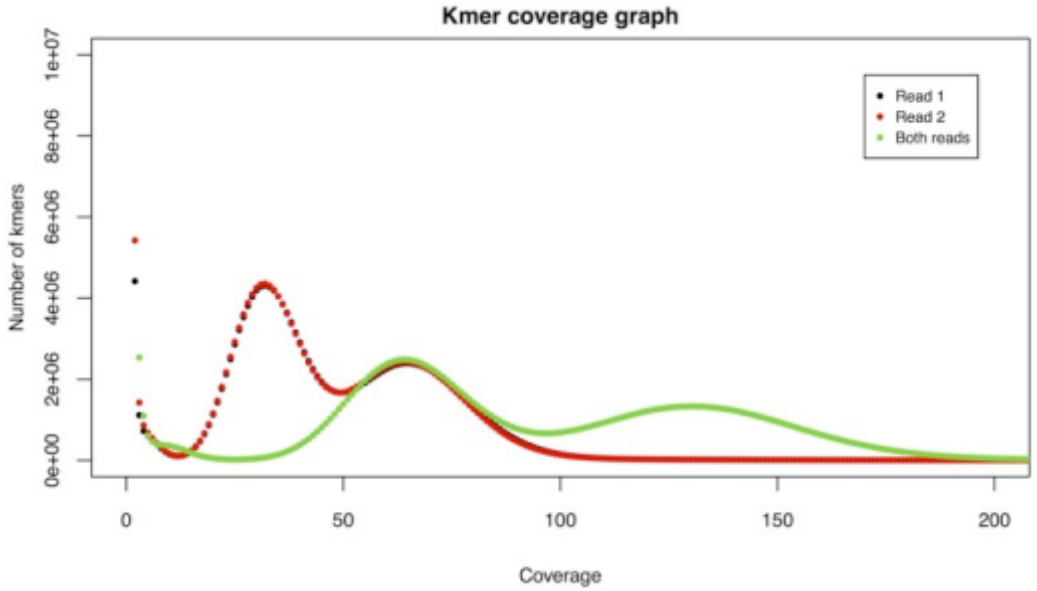
```
ACTCAGGATTATTCATACATAATAGCCGGGGTACATTTCATTAAGCGGCGATTACGATTACATACGT

          ATTCATACATAATAGCCGGGG  CATTTCATTAAGCGGCGATTA
ACTCAGGATTATTCATACATA   GCCGGGGTACATTTCATTAAG GGCGATTACGATTACATACGT
     GGATTATTCATACATAATAGC GGGGTACATTTCATTAAGCGG
               ACATAATAGCCGGGGTACATT      AGCGGCGATTACGATTACATA

          ATTCATACATA
           TTCATACATAA
ACTCAGGATTA TCATACATAAT
 CTCAGGATTAT CATACATAATA   GGGGTACATTT        GGCGATTACGA
  TCAGGATTATT ATACATAATAG   GGGTACATTTC        GCGATTACGAT
   CAGGATTATTC TACATAATAGC   GGTACATTTCA        CGATTACGATT
    AGGATTATTCA ACATAATAGCC   GTACATTTCAT        GATTACGATTA
     GGATTATTCAT CATAATAGCCG   TACATTTCATT        ATTACGATTAC
      GATTATTCATA ATAATAGCCGG   ACATTTCATTA        TTACGATTACA
       ATTATTCATAC TAATAGCCGGG   CATTTCATTAA        TACGATTACAT
        TTATTCATACA AATAGCCGGGG   ATTTCATTAAG        ACGATTACATA
         TATTCATACAT               TTTCATTAAGC        CGATTACATAC
          ATTCATACATA   GCCGGGGTACA TTCATTAAGCG        GATTACATACG
     GGATTATTCAT          CCGGGGTACAT TCATTAAGCGG        ATTACATACGT
      GATTATTCATA          CGGGGTACATT
       ATTATTCATAC          GGGGTACATTT     AGCGGCGATTA
        TTATTCATACA          GGGTACATTTC     GCGGCGATTAC
         TATTCATACAT          GGTACATTTCA     CGGCGATTACG
          ATTCATACATA          GTACATTTCAT     GGCGATTACGA
           TTCATACATAA          TACATTTCATT     GCGATTACGAT
            TCATACATAAT          ACATTTCATTA     CGATTACGATT
             CATACATAATA          CATTTCATTAA     GATTACGATTA
              ATACATAATAG          ATTTCATTAAG     ATTACGATTAC
               TACATAATAGC                          TTACGATTACA
                                                     TACGATTACAT
                ACATAATAGCC       CATTTCATTAA        ACGATTACATA
                 CATAATAGCCG       ATTTCATTAAG
                  ATAATAGCCGG       TTTCATTAAGC
                   TAATAGCCGGG       TTCATTAAGCG
                    AATAGCCGGGG       TCATTAAGCGG
                     ATAGCCGGGGT       CATTAAGCGGC
                      TAGCCGGGGTA       ATTAAGCGGCG
                       AGCCGGGGTAC       TTAAGCGGCGA
                        GCCGGGGTACA       TAAGCGGCGAT
                         CCGGGGTACAT       AAGCGGCGATT
                          CGGGGTACATT       AGCGGCGATTA
```



What can k-mers tell us?

- Plotting the frequencies of k-mers tells us how complicated our genome is.
  - There are a proportion of distinct k-mers that appear only once in the genome.
  - Distinct k-mers with count f (x-axis) vs Frequency of distinct k-mers with count f (y-axis).



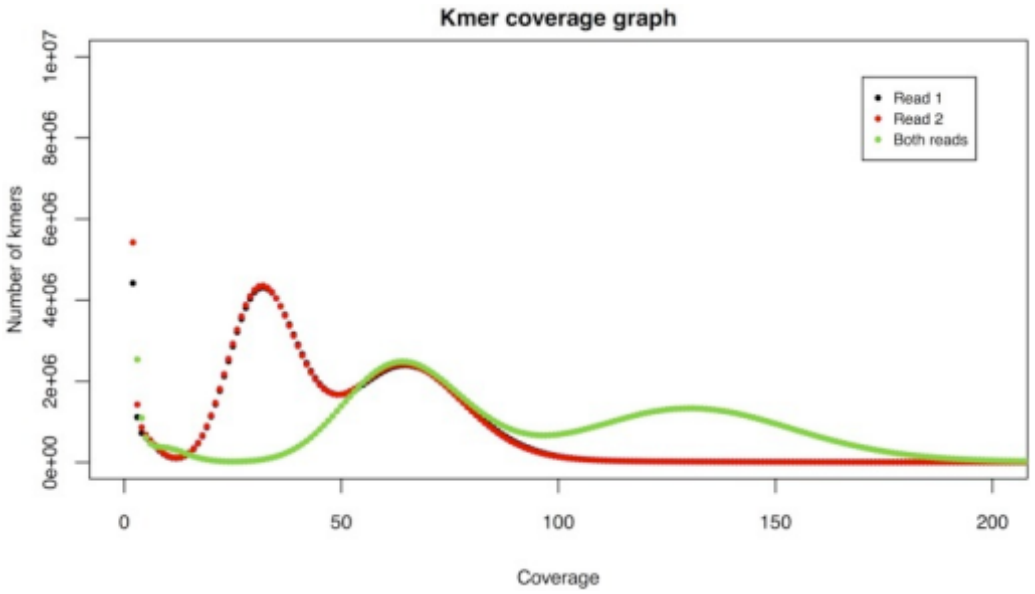


K-mer Analysis

SciLifeLab

```
ACTCAGGATTATTCATACATAATAGCCGGGG
ACTCCGGATTATTCATACATAATAGCCGGGG

ACTCAGGATTATTCATACATA
ACTCCGGATTATTCATACATA
           ATTCATACATAATAGCCGGGG
           ATTCATACATAATAGCCGGGG

           ATTCATACATA
            TTCATACATAA
ACTCAGGATTA TCATACATAAT
 CTCAGGATTAT CATACATAATA
  TCAGGATTATT ATACATAATAG
   CAGGATTATTC TACATAATAGC
    AGGATTATTCA ACATAATAGCC
     GGATTATTCAT CATAATAGCCG
      GATTATTCATA ATAATAGCCGG
       ATTATTCATAC TAATAGCCGGG
        TTATTCATACA AATAGCCGGGG
         TATTCATACAT
           ATTCATACATA
           ATTCATACATA
            TTCATACATAA
ACTCCGGATTA TCATACATAAT
 CTCCGGATTAT CATACATAATA
  TCCGGATTATT ATACATAATAG
   CCGGATTATTC TACATAATAGC
    CGGATTATTCA ACATAATAGCC
     GGATTATTCAT CATAATAGCCG
      GATTATTCATA ATAATAGCCGG
       ATTATTCATAC TAATAGCCGGG
        TTATTCATACA AATAGCCGGGG
         TATTCATACAT
           ATTCATACATA
```

- Variation generates additional distinct k-mers
- Variants are at half the frequency of the rest of the distinct genome



Estimating Genome size

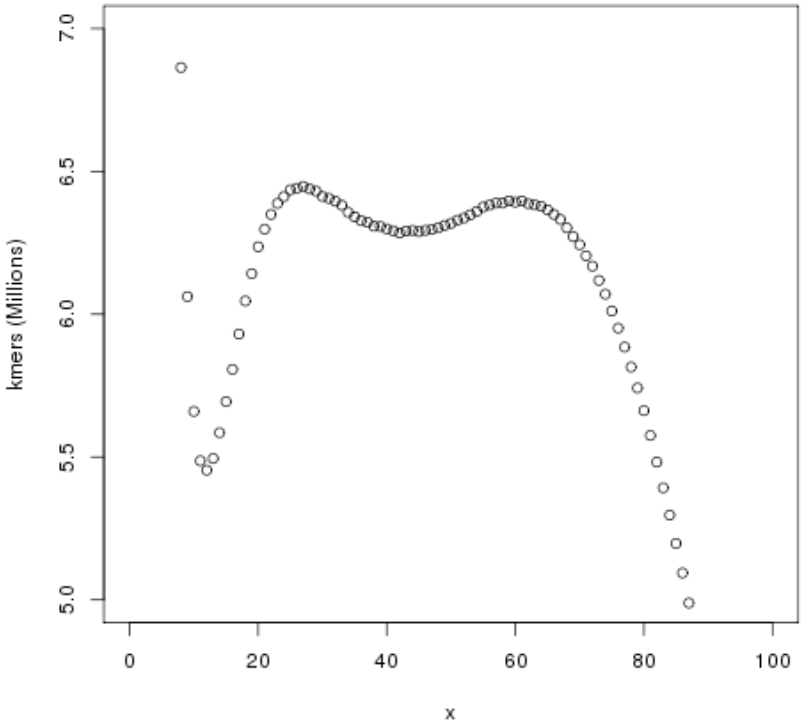
SciLifeLab

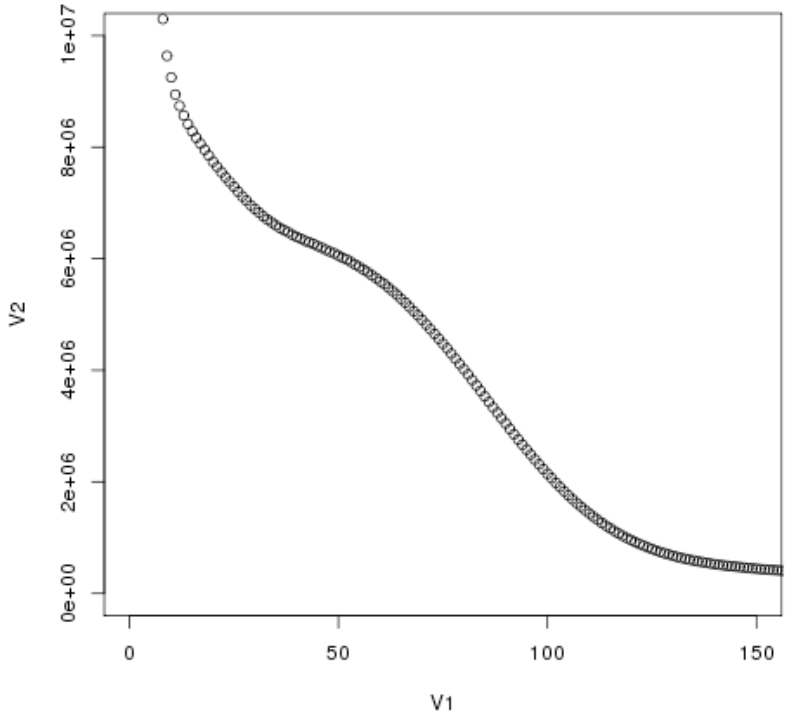
- We can use coverage to estimate genome size
  - N = M * L / (L - K + 1)
    - N is Depth of Read Coverage
    - M is mean k-mer coverage
    - L is read length
    - K is k-mer size
  - G = T / N
    - G is the genome size
    - T is the total number of bases

Estimating Genome size

SciLifeLab

- Not so easy: estimating complexity

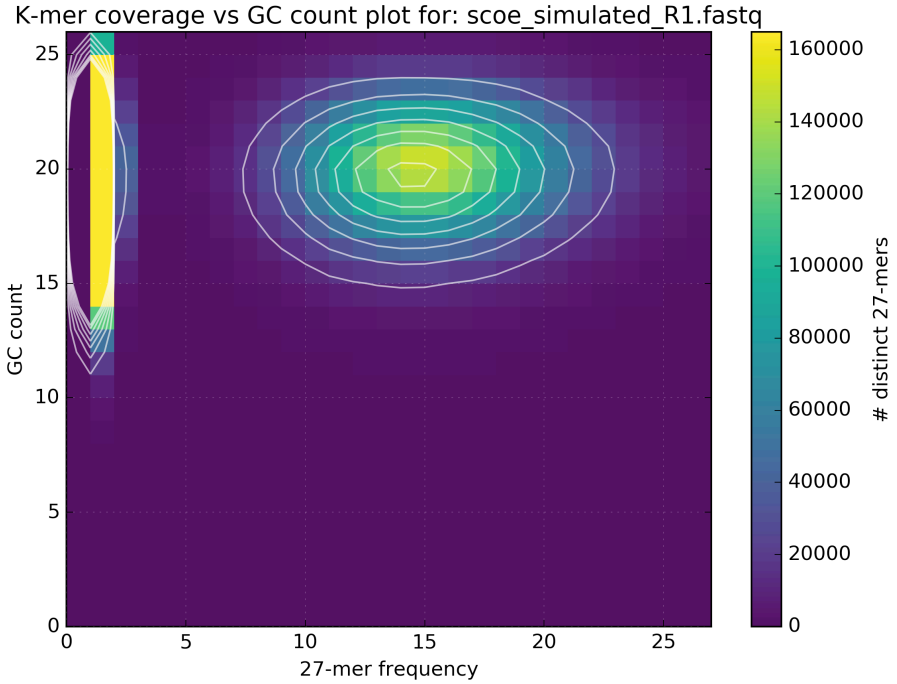




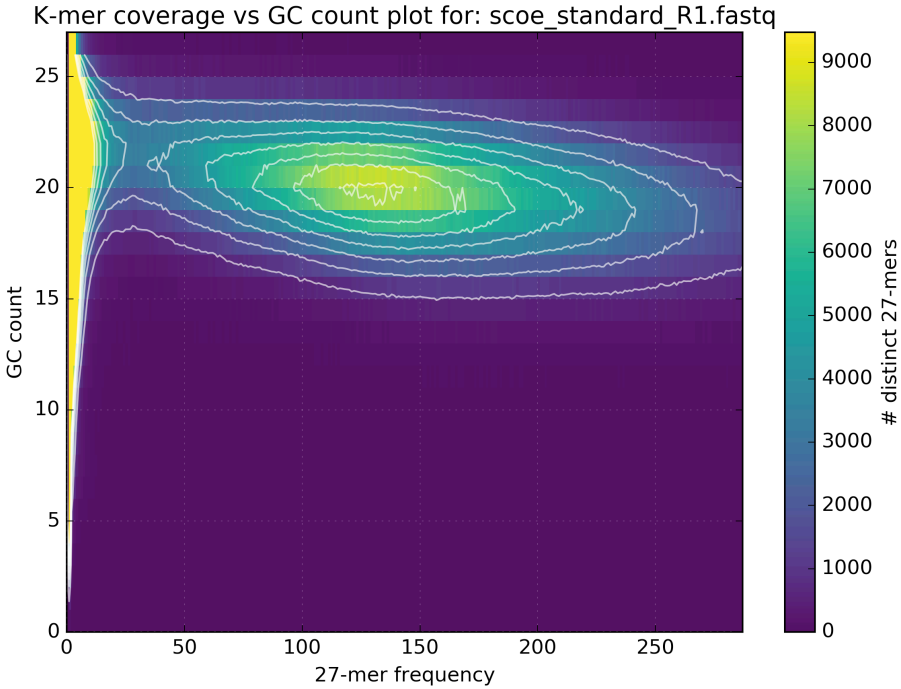
GC Content in depth

- Monitor GC Bias

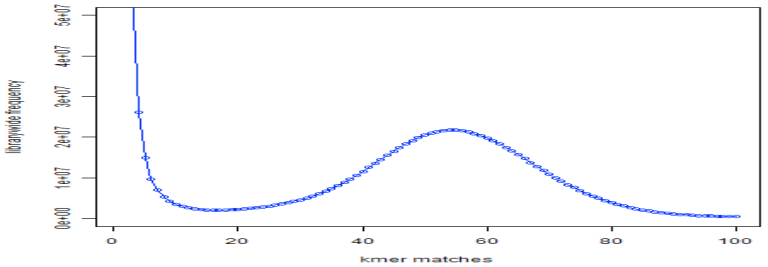
Simulated data



GC bias in Standard Illumina protocol



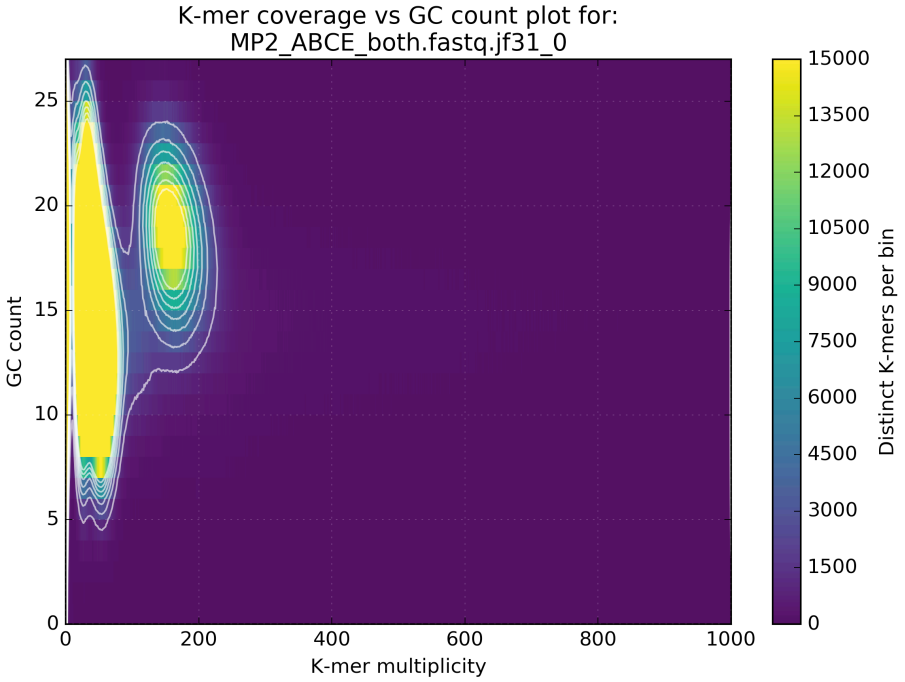

```
ACTCAGGATTA   GC=4 Count=1
AATAGCCGGGG   GC=7 Count=2
```

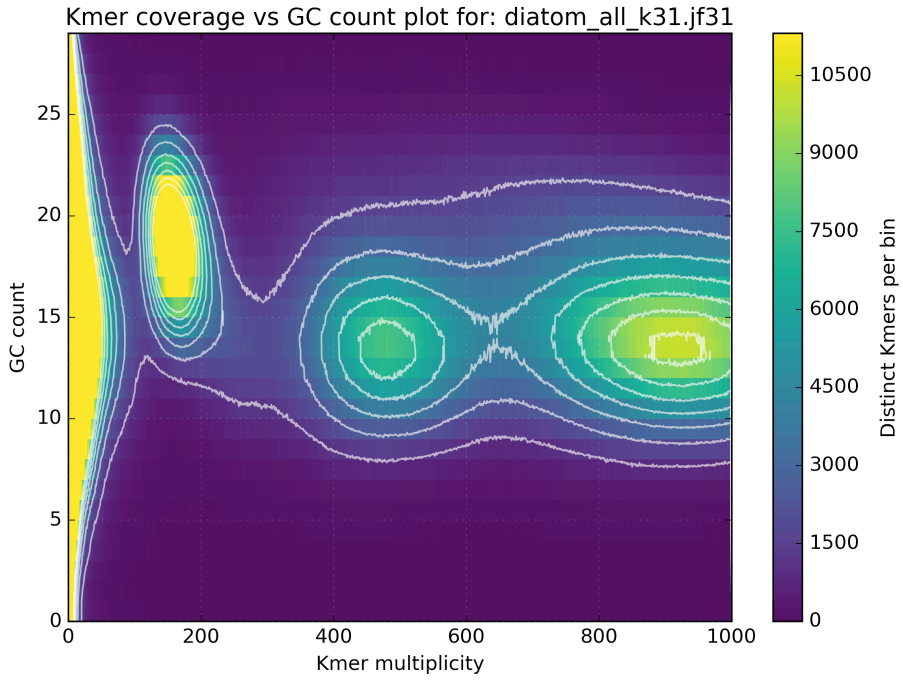


GC Content in depth

SciLifeLab

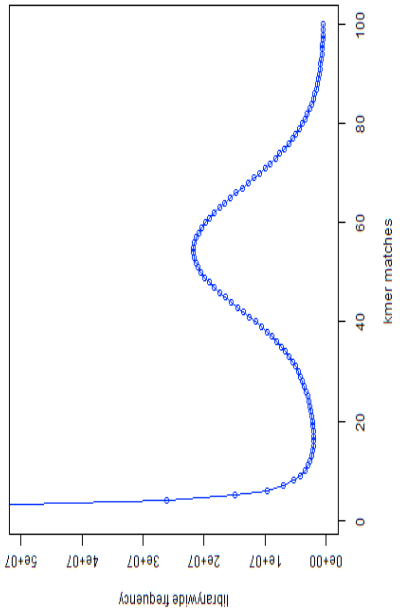
- Uncover contamination
  - Separate bacteria from eukaryote
  - Separate organelle from nuclear genome

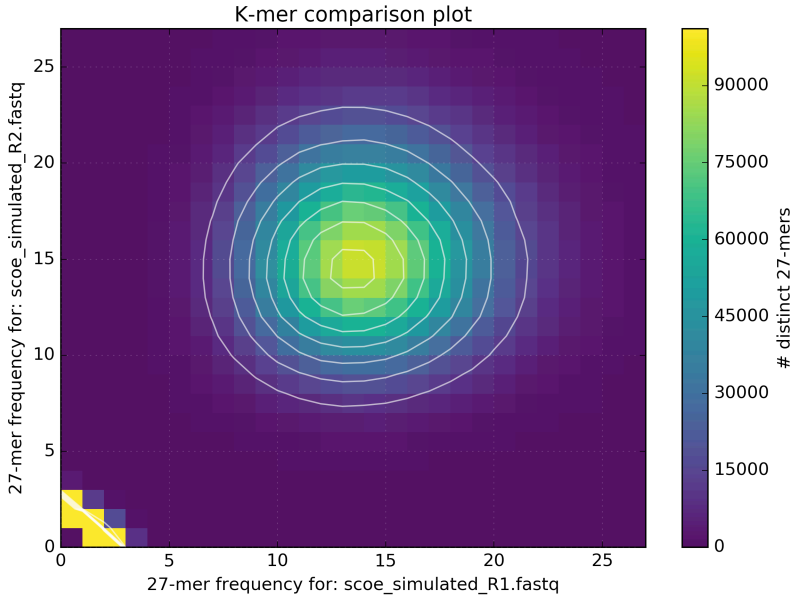


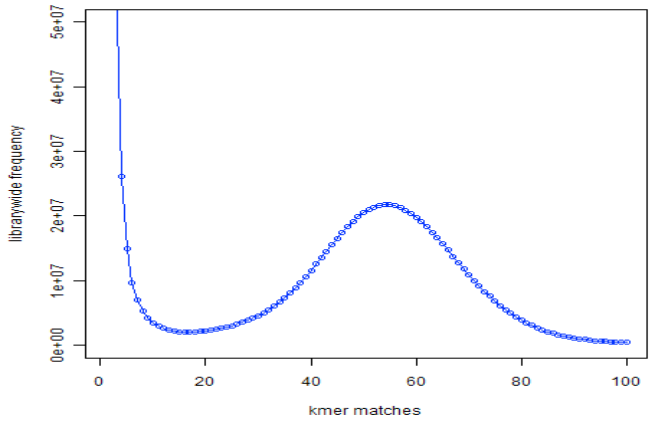


What can k-mers tell us?

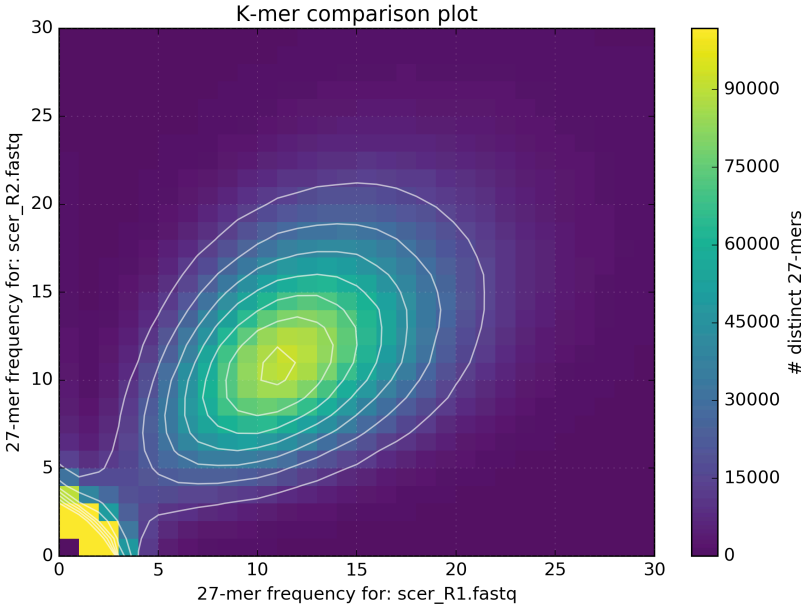
- Comparing k-mer counts between data reveals biases
  - R1 vs R2
  - Lib1 vs Lib2



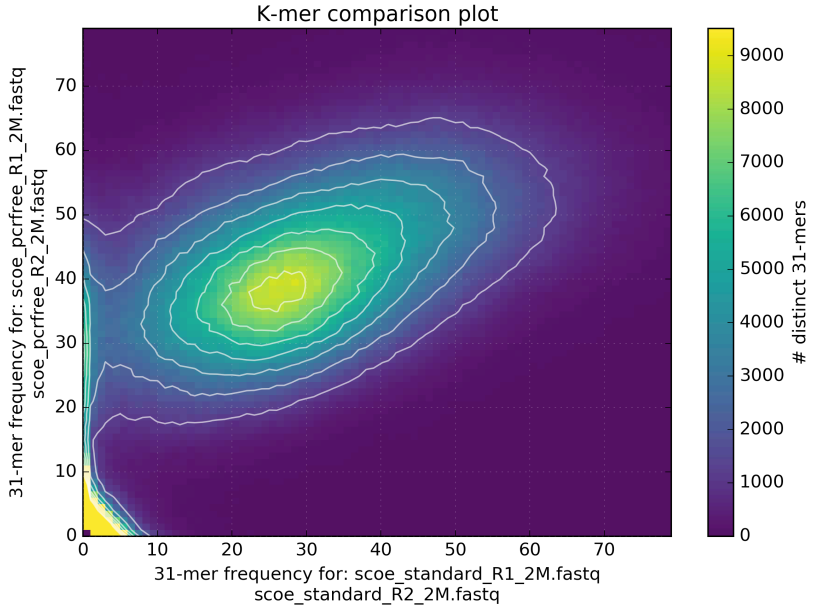




Data comparison



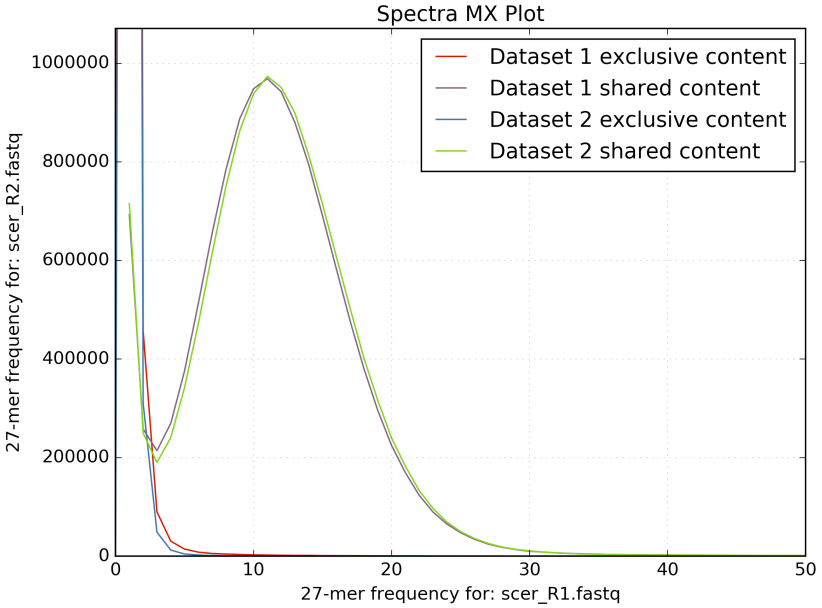
R1 vs R2

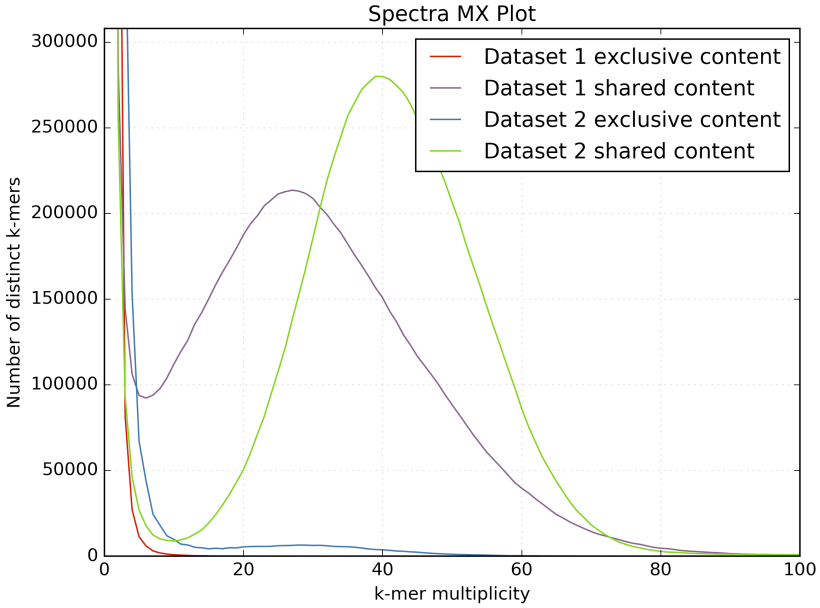


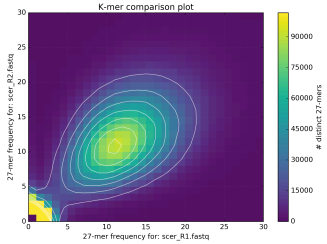
Standard vs PCR free

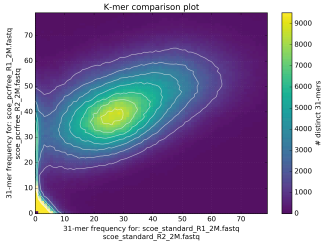
- PCR free captures data missing in standard protocol

Data comparison









K-mer Filtering

SciLifeLab

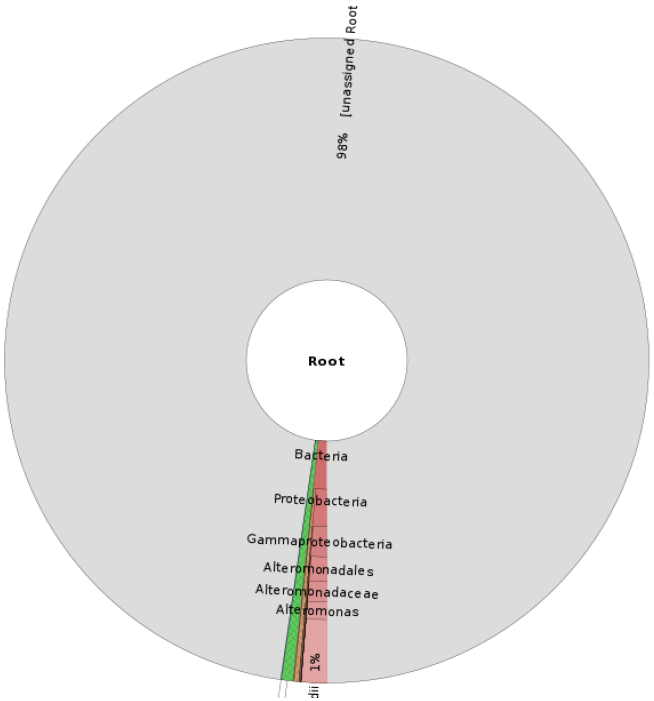
- Methods:
  - Digital Normalisation
  - Error Correction
  - Cut-off based
- The result is difficult to predict
  - Removes low coverage data
  - Overcorrects repeats
  - Removal of similar alleles
  - Large computation time for potentially little gain

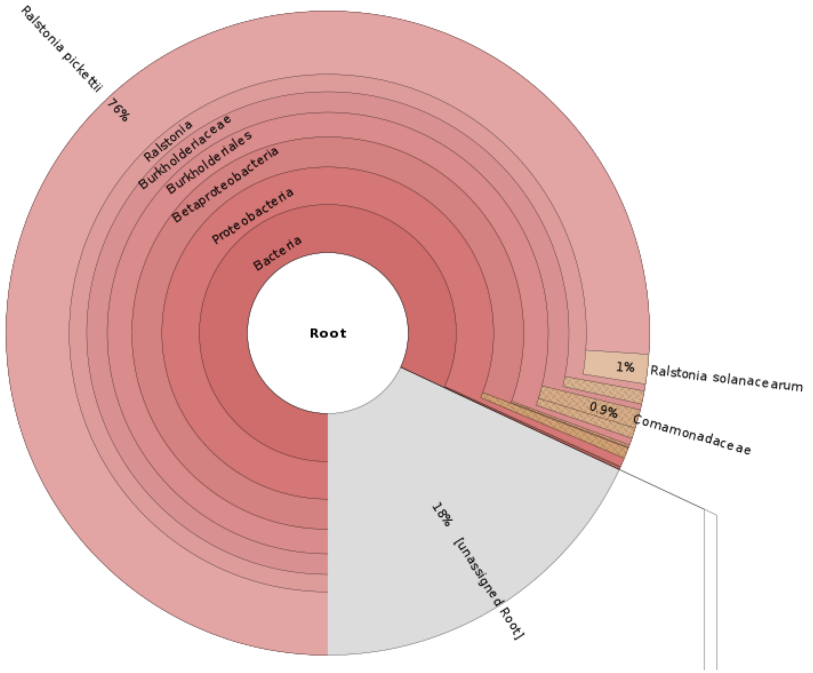
Contamination Analyses

SciLifeLab

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

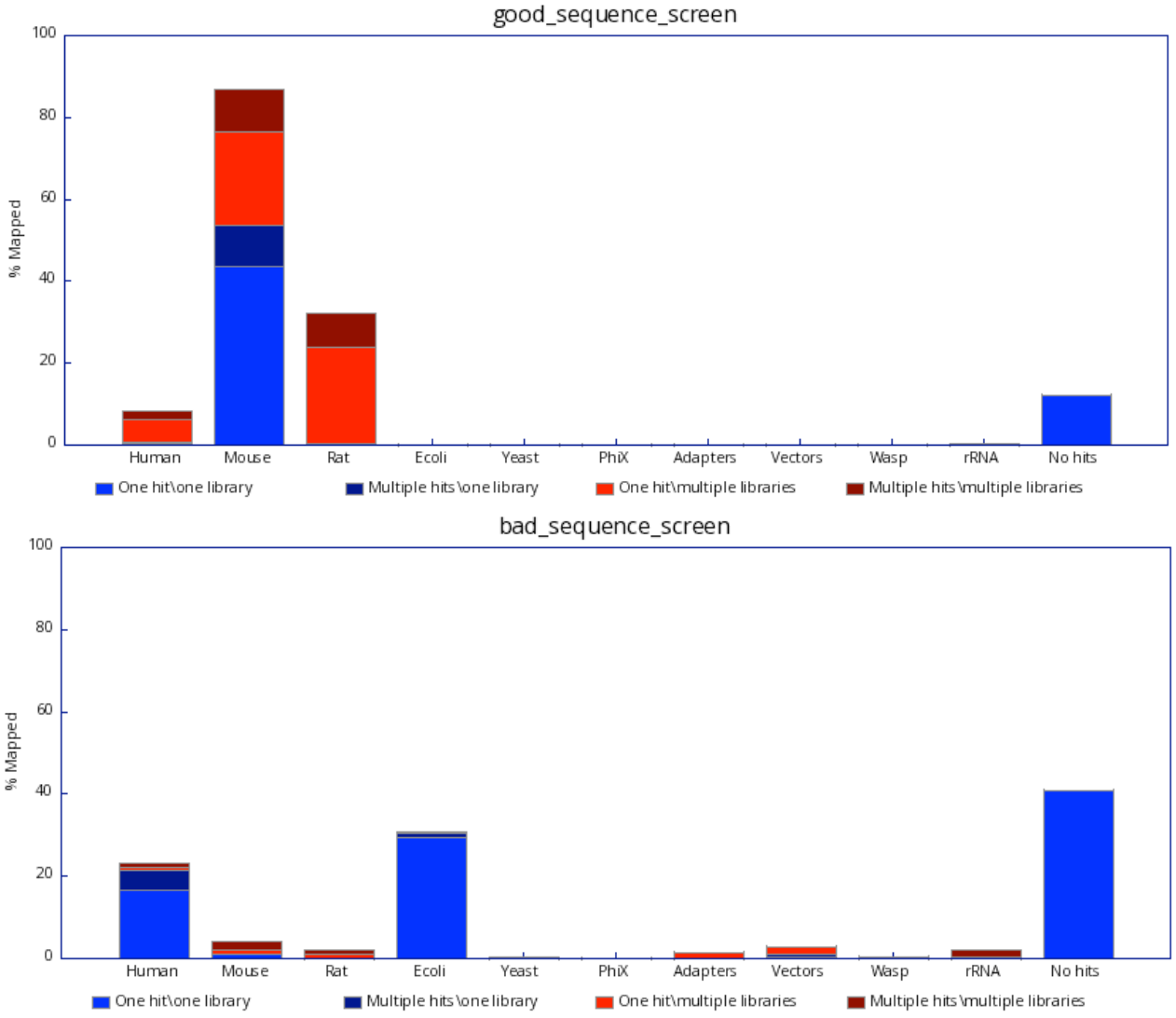
Kraken taxonomic classification





FastQ Screen

SciLifeLab



Reference Alignment

SciLifeLab

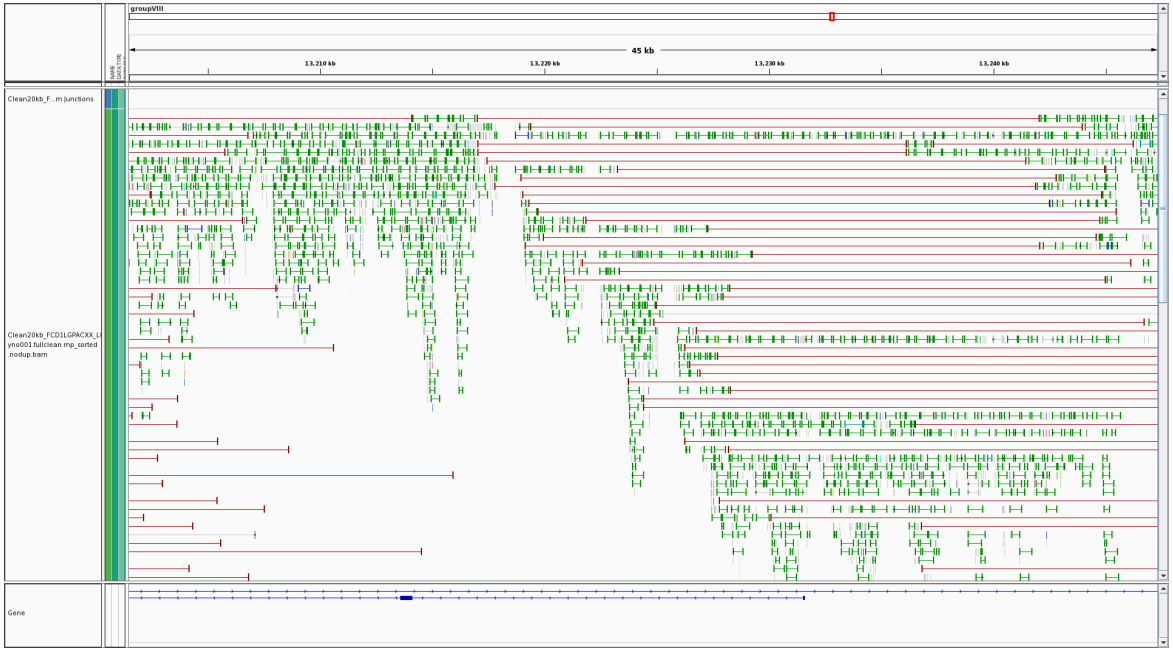
- Alignment to a known close relative.
  - Check % alignment.

```
5351663 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
19003 + 0 supplementary
0 + 0 duplicates
543253 + 0 mapped (10.15% : N/A)
5332660 + 0 paired in sequencing
2666330 + 0 read1
2666330 + 0 read2
0 + 0 properly paired (0.00% : N/A)
524250 + 0 with itself and mate mapped
0 + 0 singletons (0.00% : N/A)
12048 + 0 with mate mapped to a different chr
160 + 0 with mate mapped to a different chr (mapQ>=5)
```

Paranoia

SciLifeLab

- Other QC issues
  - Incorrect fragment size
    - Alignment -> Picard Metrics / Genome Browser



  - Incorrect Demultiplexing
    - Check data proportions

Summary

SciLifeLab

- Check data integrity
- Check basic statistics
- Check for contaminants, adapter, duplication
- Check data sets share the same information content
- Check platform specific diagnostics
- Trash in -> trash out.

- If something looks wrong, talk to your sequencing provider
  - (but get a second opinion on the answer they give you)