Galaxy

Sequence Data Quality Control

www.galaxyproject.org

@jxtx / #usegalaxy

Quality control analysis: Get the Data

Import

Shared Data → Data Libraries →

Illumina iDEA Datasets (sub-sampled) →

BT20 paired-end RNA-seq subsampled (end 1)

NGS Data Quality Control

- FASTQ format
- Examine quality in an Chip-Seq dataset
- Trim/filter as we see fit, hopefully without breaking anything.

What is FASTQ?

- Specifies sequence (FASTA) and quality scores (PHRED)
- Text format, 4 lines per entry

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- Several variants to how FASTQ scores are encoded

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
  ............................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126

 S - Sanger       Phred+33,  93 values  (0, 93) (0 to 60 expected in raw reads)
 I - Illumina 1.3 Phred+64,  62 values  (0, 62) (0 to 40 expected in raw reads)
 X - Solexa       Solexa+64, 67 values (-5, 62) (-5 to 40 expected in raw reads)
```

http://en.wikipedia.org/wiki/FASTQ_format

NGS Data Quality: Assessment tools

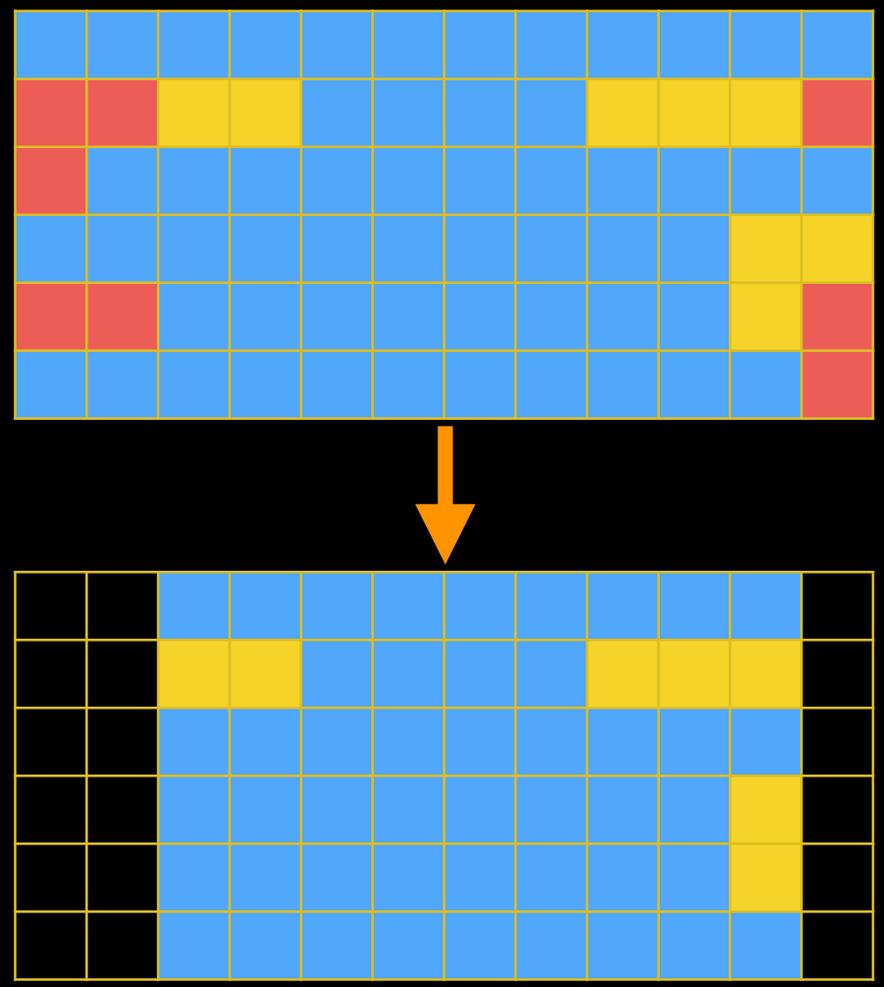
NGS QC and Manipulation → FastQC

Gives you a lot of information but little control over how it is calculated or presented.

http://bit.ly/FastQCBoxPlot

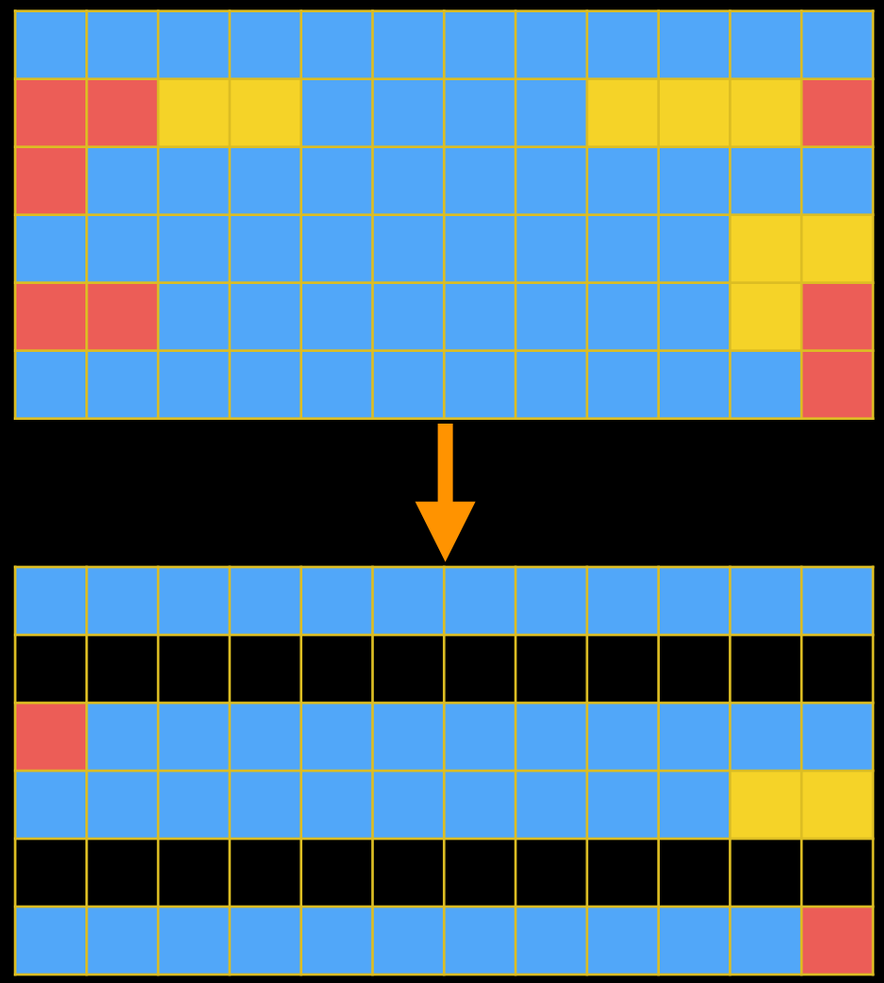
NGS Data Quality: Trim as we see fit

- Trim as we see fit: Option 1
  - **NGS QC and Manipulation → FASTQ Trimmer by column**
  - Trim same number of columns from every record
- Can specify different trim for 5′ and 3′ ends



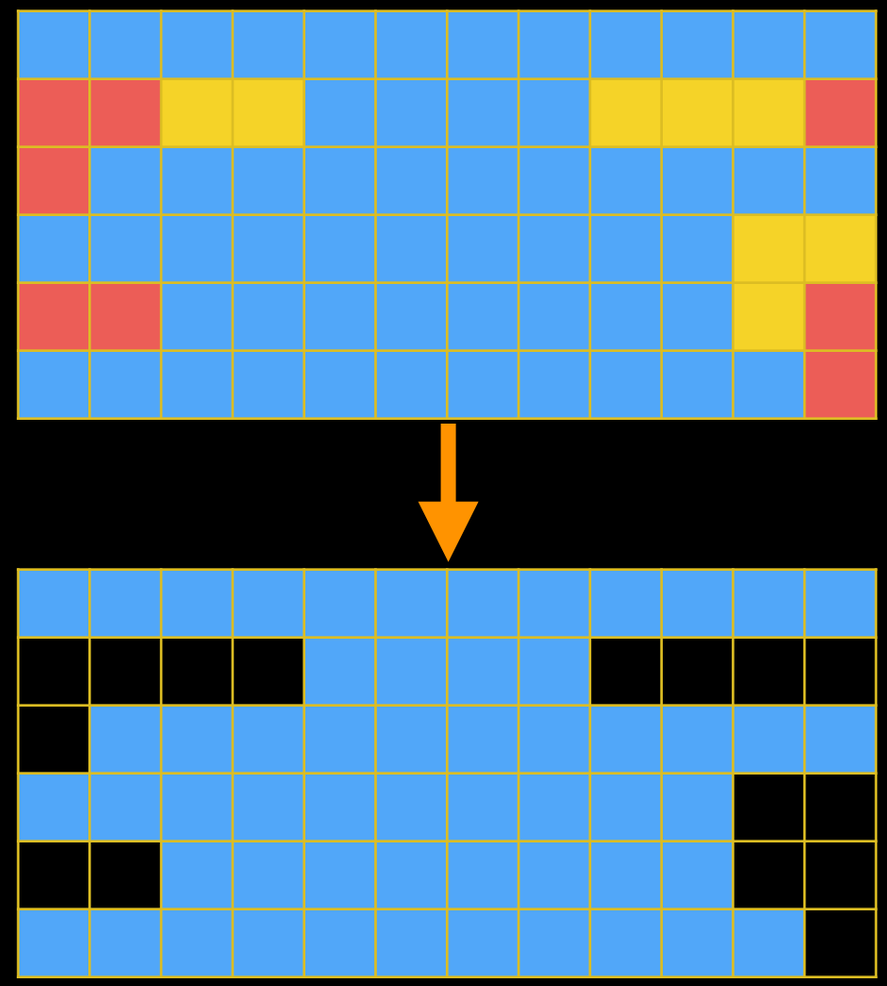
NGS Data Quality: Base Quality Trimming

- ~~Trim~~ Filter as we see fit: Option 2
  - NGS QC and Manipulation → **Filter FASTQ reads by quality score and length**
  - Keep or discard whole reads
  - Can have different thresholds for different regions of the reads.
  - Keeps original read length.

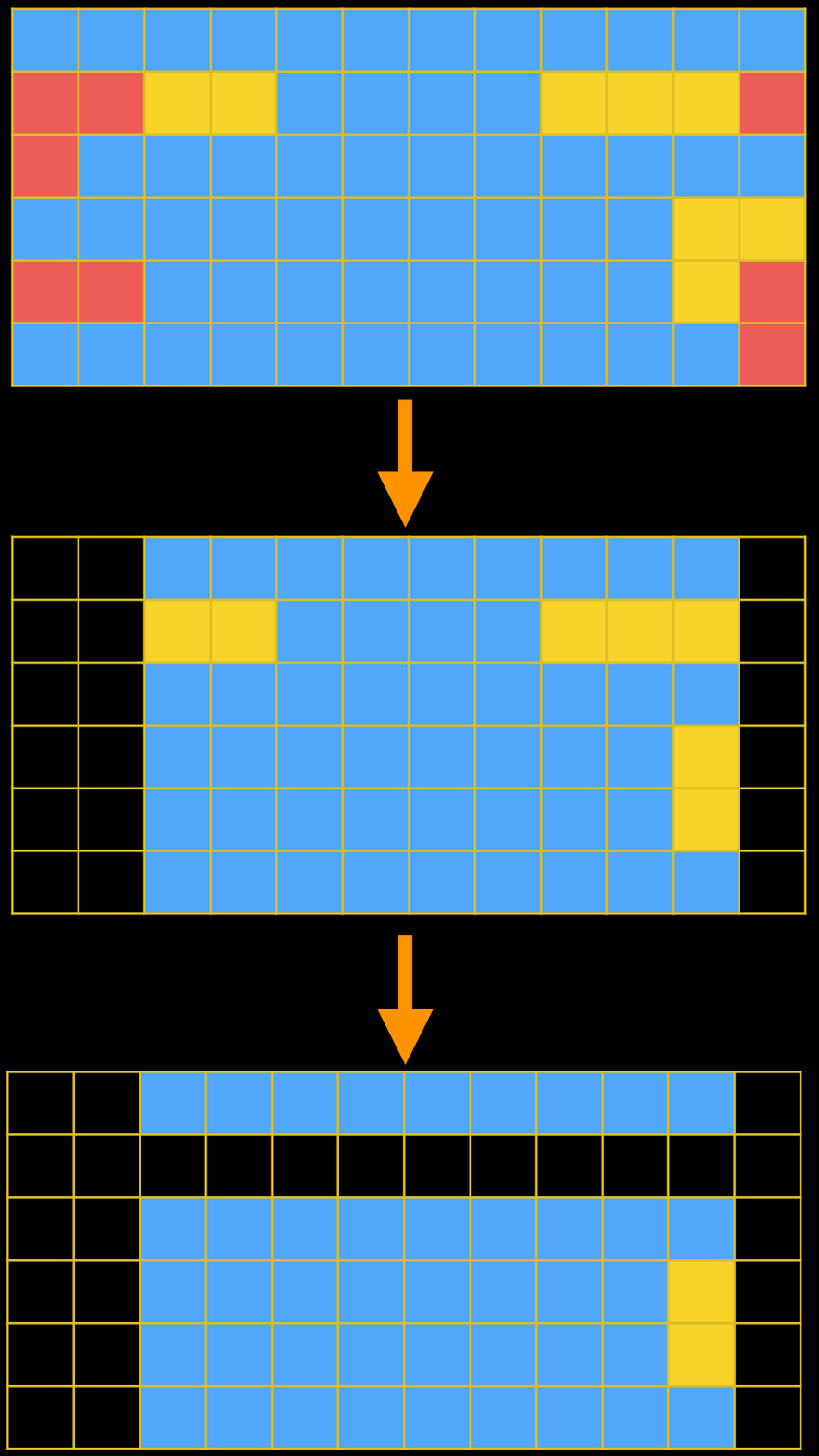


NGS Data Quality: Base Quality Trimming

- Trim as we see fit: Option 3
  - NGS QC and Manipulation → **FASTQ Quality Trimmer by sliding window**
  - Trim from both ends, using sliding windows, until you hit a high-quality section.
  - Produces variable length reads



Options are not mutually exclusive



Option 1
(by column)

+

Option 2
(by entire row)

Trim? *As we see fit?*

- Introduced 3 options
  - One preserves original read length, two don't
  - One preserves number of reads, two don't
  - Two keep/make every read the same length, one does not

Trim? *As we see fit?*

- **Choice depends on downstream tools**
- **Find out assumptions & requirements for downstream tools and make appropriate choice(s) now.**
- **How to do that?**
  - **Read the tool documentation**
  - **http://biostars.org/**
  - **http://seqanswers.com/**
  - **http://galaxyproject.org/search**

Summary

Many factors can effect the quality of DNA sequencing data

FASTQC is one tool that allows evaluating a number of quality metrics for FASTQ datasets from many sequencing platforms

The Galaxy FASTQ manipulation tools can help to salvage quality sequences from lower quality data

When filtering or trimming reads, be sure to take into account the requirements of downstream analysis