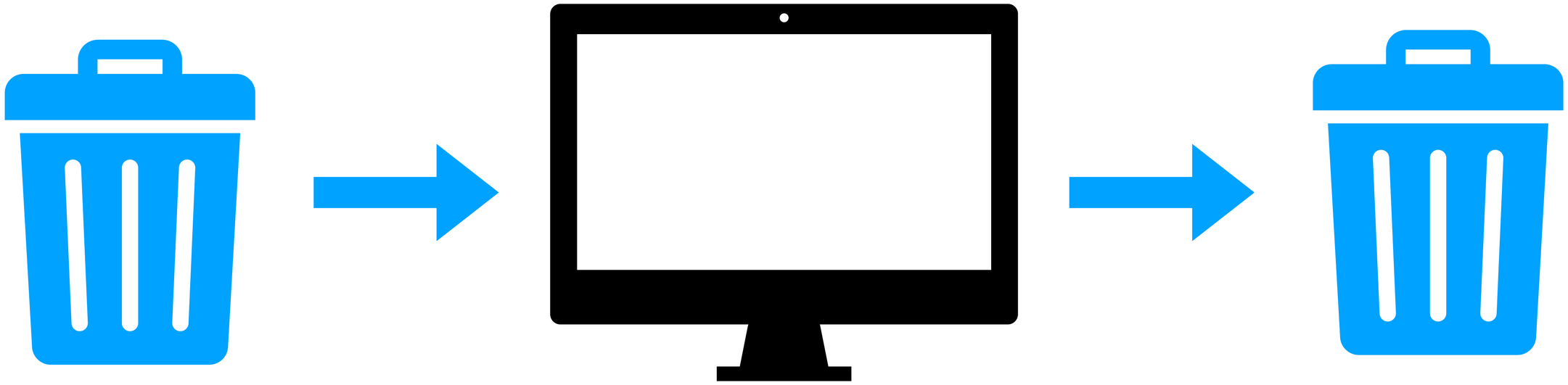
ILLUMINA DATA QC AND CONDA

DARYL DOMMAN, PHD DARRELL DINWIDDIE, PHD

DDOMMAN@GMAIL.COM

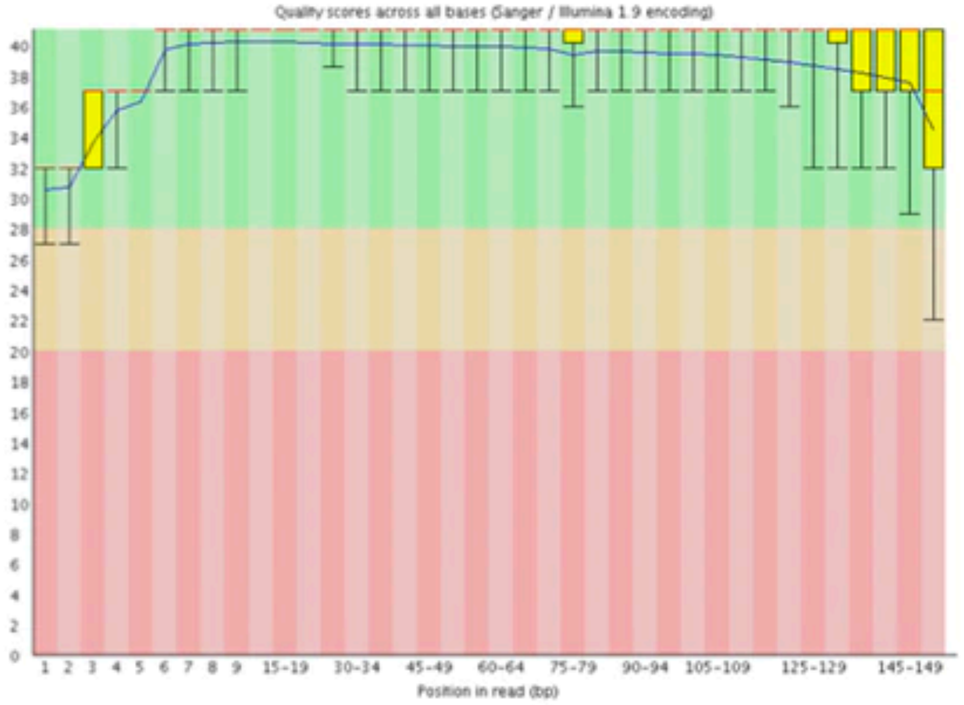
The “Golden” Rule



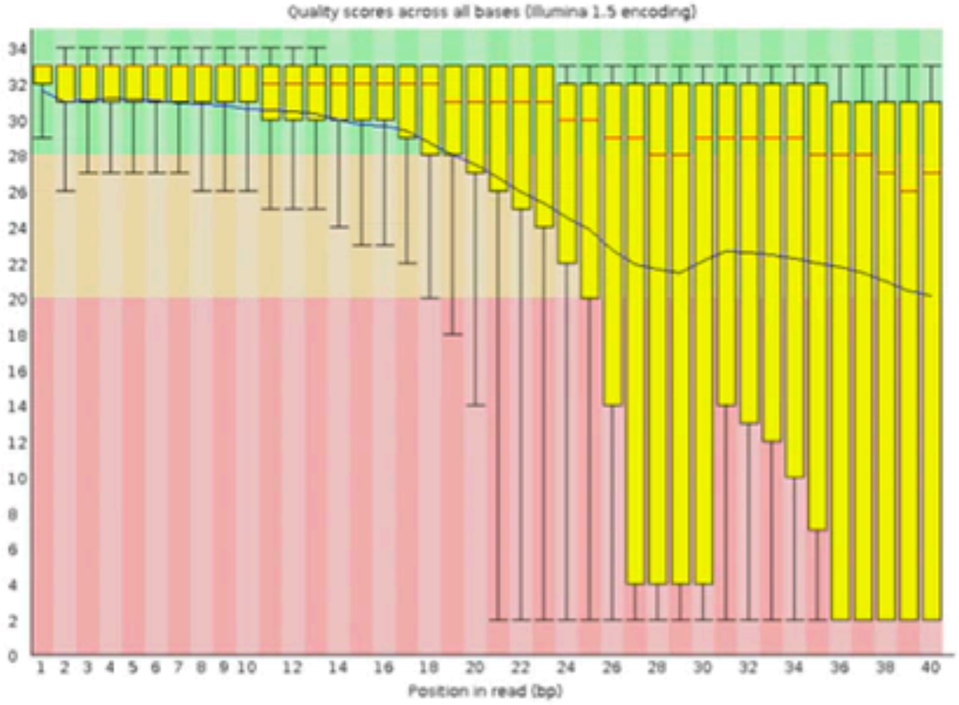
Having good quality fastq data is important!!

FastQC: Per base sequence quality

Good data



Bad data



Quality score interpretation

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
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%

The quality (Q), also called Phred score, is the probability (P) that the corresponding basecall is incorrect.

Many tools for trimming

- Trimmomatic
- sickle
- fastP
- bbduk
- cutadapt
- Trim Galore

Today's Agenda

- Look at two M. tuberculosis datasets
- Run fastqc to look at fastq data quality
- Use conda to install new tools
- Trim poor quality reads with Trim Galore

Questions?