UCDAVIS Bioinformatics Core

Sequence Preprocessing: A perspective

Dr. Matthew L. Settles

Genome Center
University of California, Davis
settles@ucdavis.edu

Why Preprocess reads

- We have found that aggressively “cleaning” and processing reads can make a large difference to the **speed** and **quality** of assembly and mapping results. Cleaning your reads means, removing reads/bases that are:
  - other unwanted sequence (polyA tails in RNA-seq data)
  - artificially added onto sequence of primary interest (vectors, adapters, primers)
  - join short overlapping paired-end reads
  - low quality bases
  - originate from PCR duplication
  - not of primary interest (contamination)
- Preprocessing also produces a number of statistics that are technical in nature that should be used to evaluate “experimental consistancy”

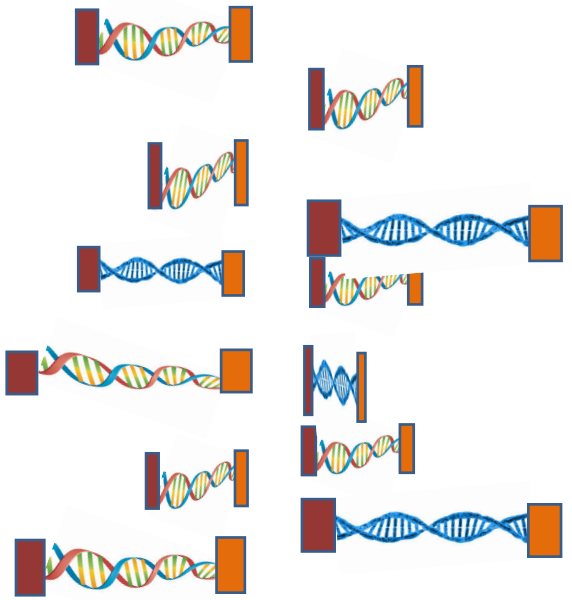
Read Preprocessing strategies, many over time

- Identity and remove contaminant and vector reads
  - Reads which appear to fully come from extraneous sequence should be removed.
- Quality trim/cut
  - “end” trim a read until the average quality > Q (Lucy)
  - remove any read with average quality < Q
- eliminate singletons/duplicates
  - If you have excess depth of coverage, and particularly if you have at least x-fold coverage where x is the read length, then eliminating singletons is a nice way of dramatically reducing the number of error-prone reads.
  - Read which appear the same (particularly paired-end) are often more likely PCR duplicates and therefor redundant reads.
- eliminate all reads (pairs) containing an “N” character
  - If you can afford the loss of coverage, you might throw away all reads containing Ns.
- Identity and trim off adapter and barcodes if present
  - Believe it or not, the software provided by Illumina, either does not look for, or does a mediocre job of, identifying adapters and removing them.

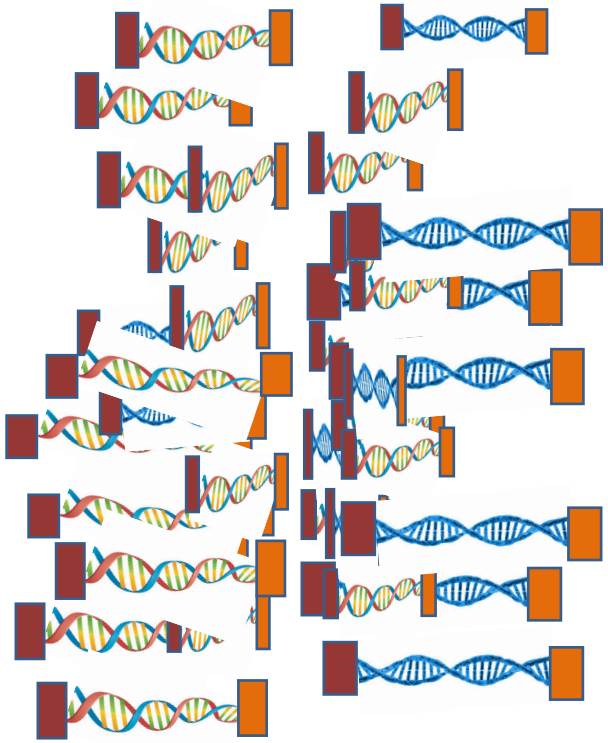
UCDAVIS Bioinformatics Core

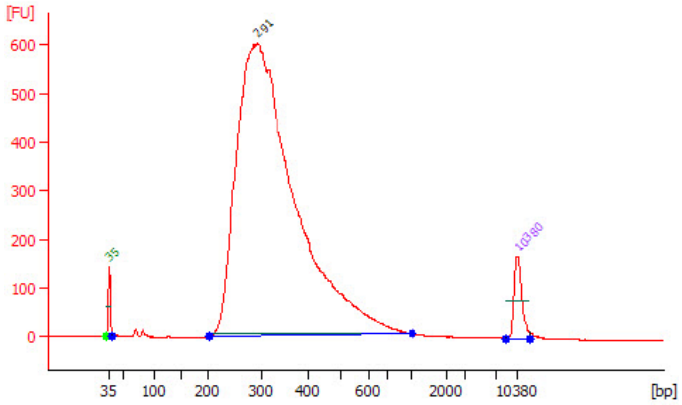
DNA/RNA, could contain ‘contamination’

Library prep, fragmentation, adapter addition

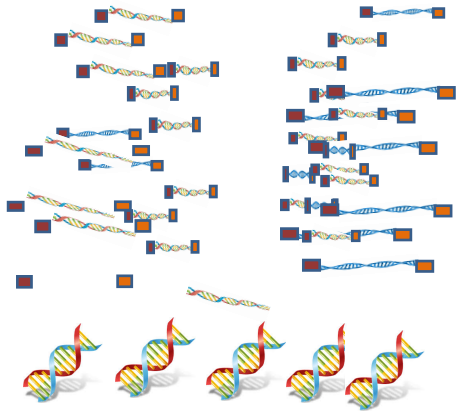


PCR enrichment

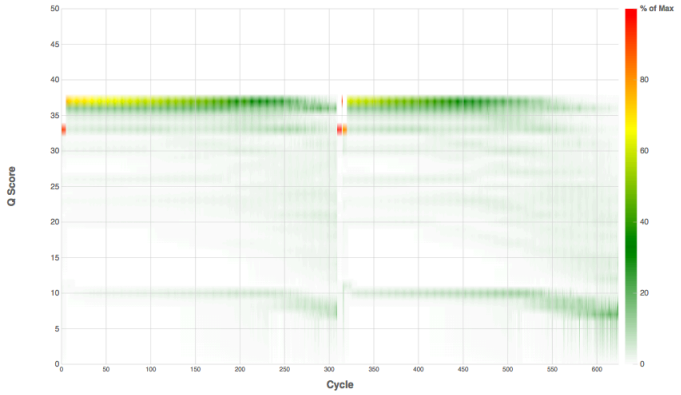




Final Library, size distribution



Possible addition of phiX



Sequencing Characteristics/ Quality

Preprocessing

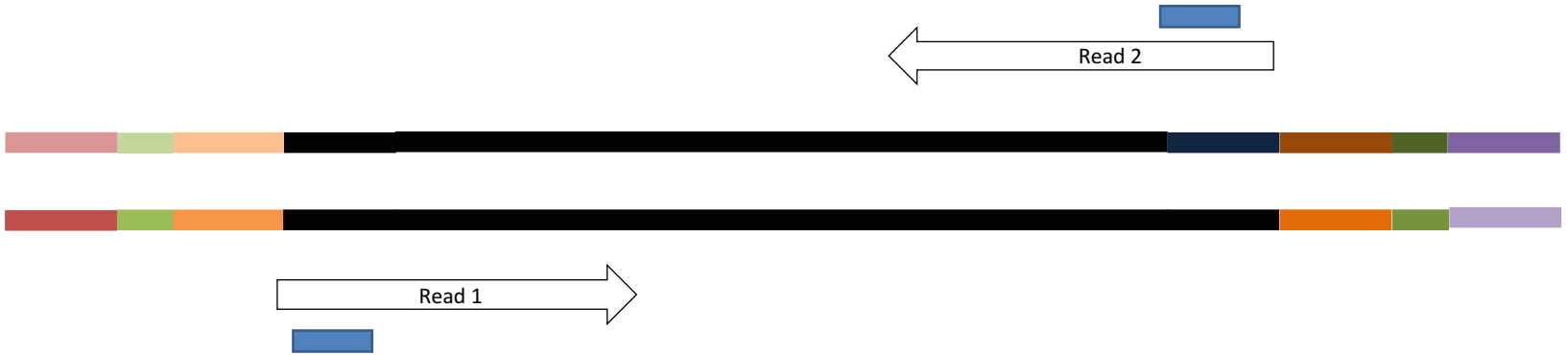
- Map reads to contaminants/PhiX and extract unmapped reads [bowtie2 --local
  - Remove contaminants (at least PhiX), uses bowtie2 then extracts all reads (pairs) that are marked as unmapped.
- Super-Deduper [PE reads only]
  - Remove PCR duplicates (we use bases 10-35 of each paired read)
- FLASH2 [PE reads only]
  - Join and extend, overlapping paired end reads
  - If reads completely overlap they will contain adapter, remove adapters
  - Identify and remove any adapter dimers present
- Scythe [SE Reads only]
  - Identify and remove adapter sequence
- Sickle
  - Trim sequences (5’ and 3’) by quality score (I like Q20)
- cleanup
  - Run a polyA/T trimmer
  - Remove any reads that are less then the minimum length parameter
  - Produce preprocessing statistics

Why Screen for PhiX

- PhiX is a common control in Illumina runs, facilities rarely tell you if/when PhiX has been spiked in
  - Does not have a barcode, so in theory should not be in your data
- **However**
  - When I know PhiX has been spiked in, I find sequence every time
  - When I know PhiX has not been spiked in, I **do not** find sequence
- Better safe than sorry and screen for it.

UCDAVIS Bioinformatics Core

Super Deduper



Data	Alignment Algorithm	MarkDuplicates	Rmdup	Super Deduper	FastUniq	Fulcrum	Total # of Reads
PhiX	BWA MEM	1,048,278 (0.25%)	1,011,145 (1.05%)	1,156,700 (13.7%)	4,202,526	3,092,155	4,750,299
	Bowtie 2 Local	1,054,725 (6.62%)	948,784 (10.2%)	1,166,936 (14.0%)	4,236,647	3,103,872	4,790,972
	Bowtie 2 Global	799,524 (0%)	800,868 (0.12%)	896,487 (9.92%)	3,768,641	2,704,114	4,293,787
Acropora digitifera	BWA MEM	5,132,111 (2.26%)	6,906,634 (44.5%)	5,133,339 (10.2%)	12,968,469	2,103,567	54,108,240
	Bowtie 2 Local	4,688,809 (4.03%)	5,931,862 (38.9%)	3,971,743 (9.32%)	9,893,903	4,259,619	41,728,154
	Bowtie 2 Global	1,457,865 (3.62%)	1,512,966 (24.2%)	1,185,838 (11.4%)	3,014,498	1,286,031	11,600,847

https://github.com/dstreett/Super-Deduper

Super Deduper

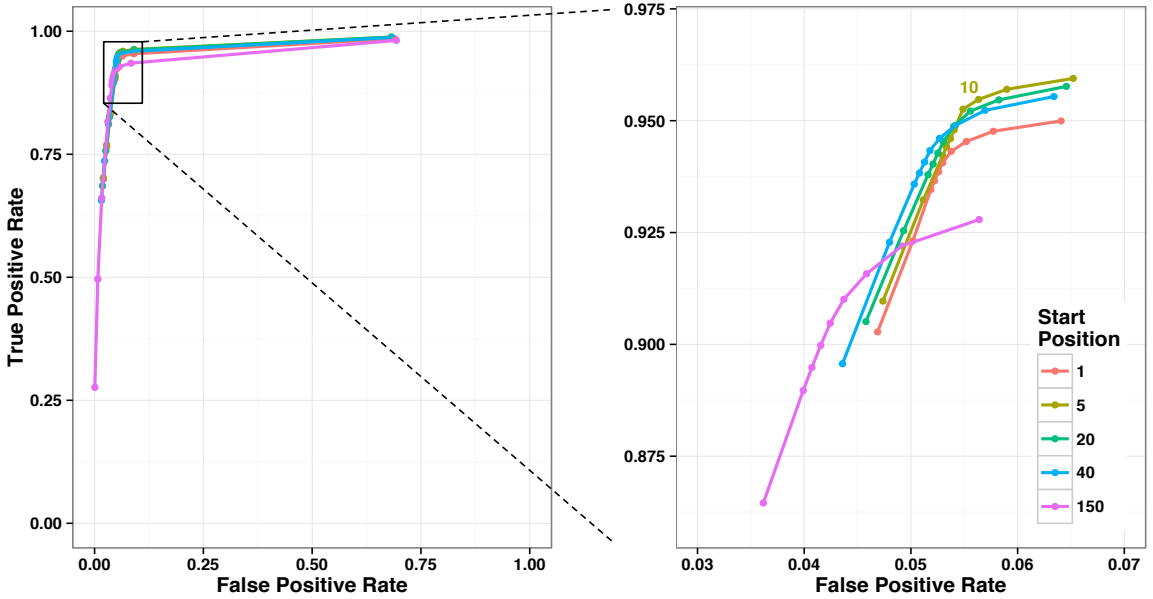
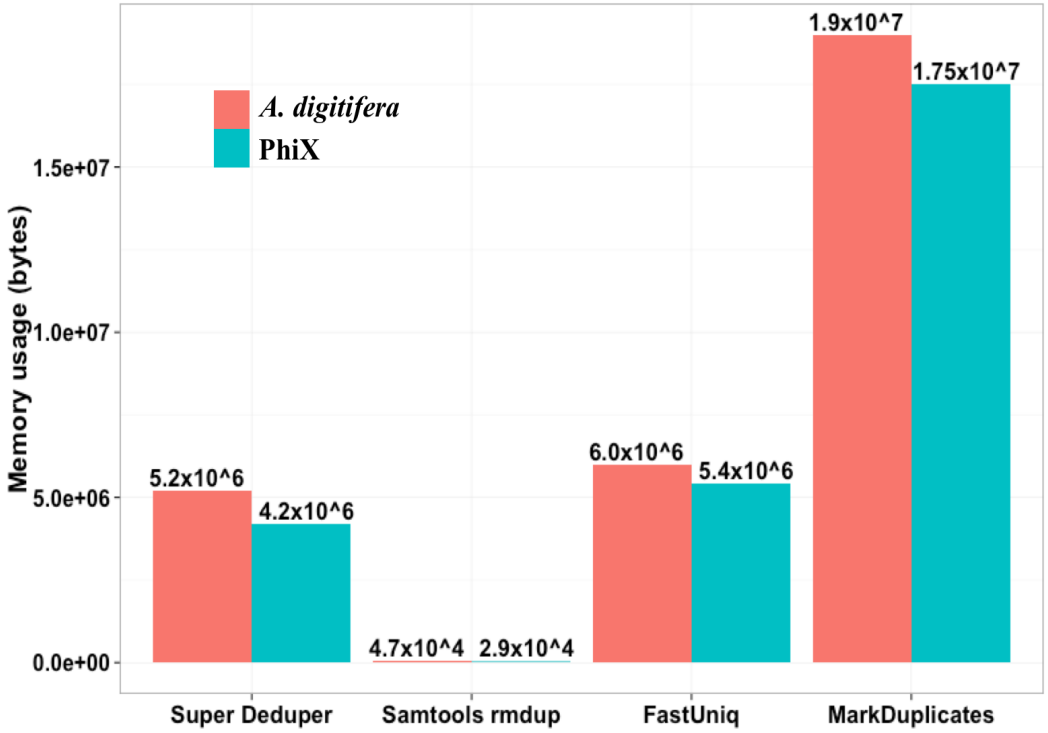


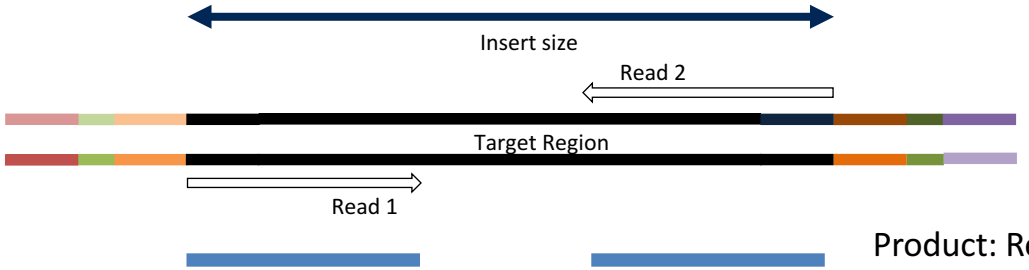
Figure 1: ROC curves. Only a representative subset of the different start positions is shown. The image on the left shows the full ROC curves and the image on the left is a zoomed in view of corner of the curves. Each curve represents a start position and each point represents a length. The labeled point in the image on the right is the default start and length for Super Deduper.



We calculated the Youden Index for every combination tested and the point that acquired the highest index value (as compared to Picard MarkDuplicates) occurred at a start position of 5bp and a length of 10bps (20bp total over both reads)

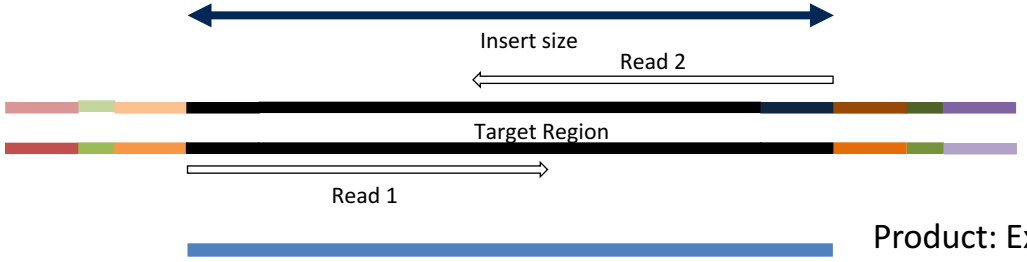
Flash2 – overlapping of reads and adapter removal in paired end reads

Insert size > length of the number of cycles



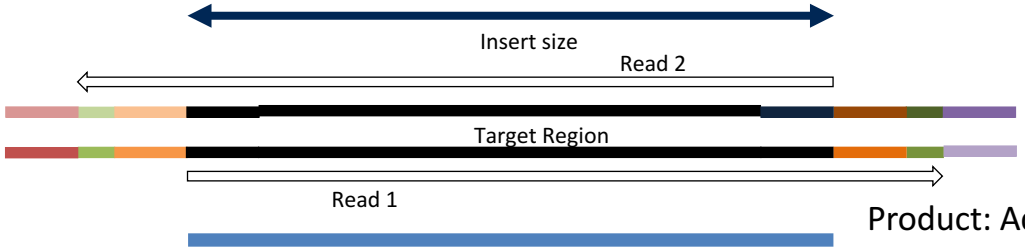
Product: Read Pair

Insert size < length of the number of cycles (10bp min)



Product: Extended, Single

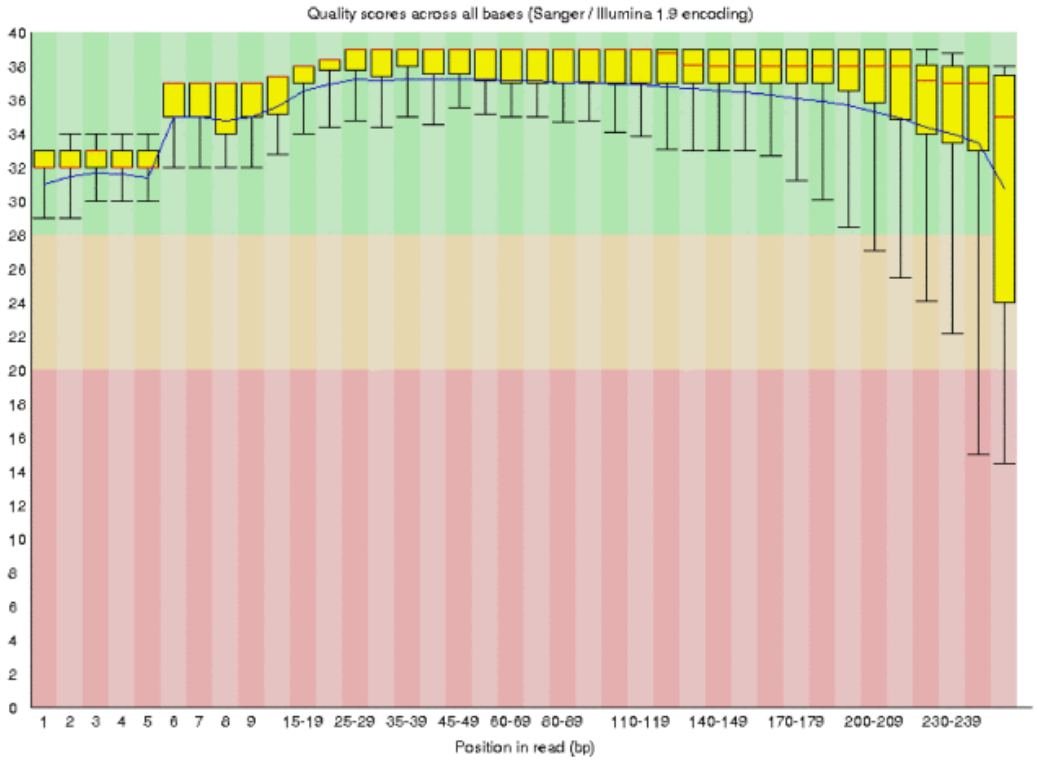
Insert size < length of the read length



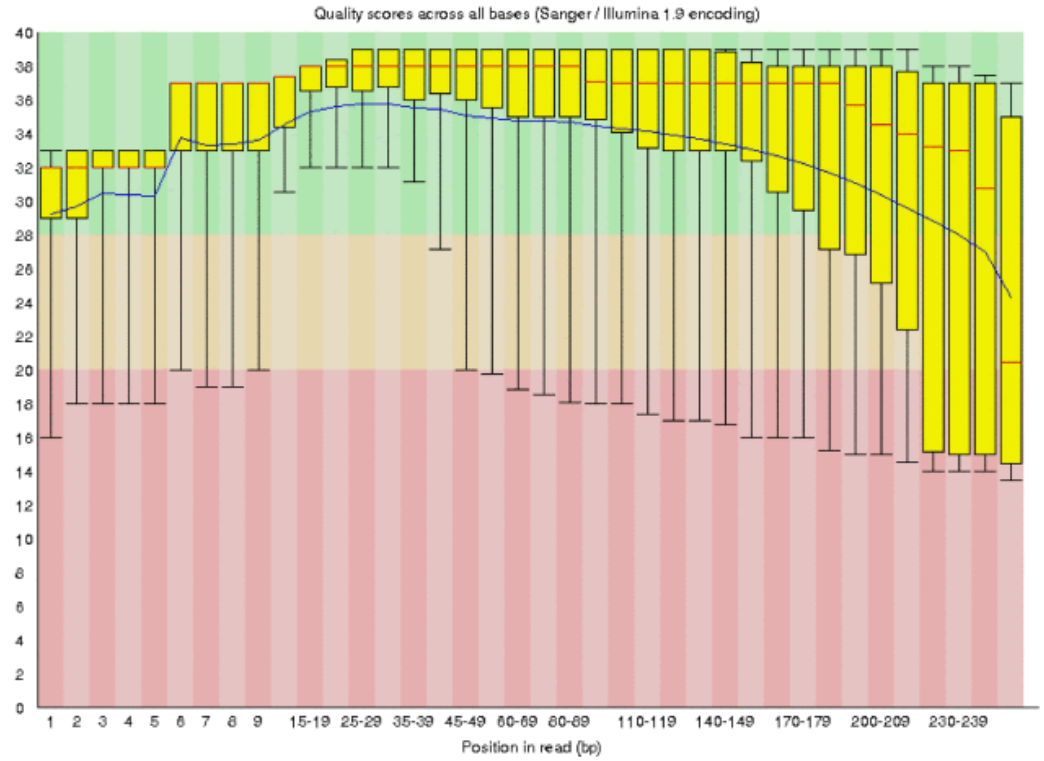
Product: Adapter Trimmed, Single

https://github.com/dstreett/FLASH2

Quality Trimming - Sickle



Forward reads



Reverse reads

Remove “poor” quality sequence from both the 5’ and 3’ ends

Ribosomal RNA

- Ribosomal RNA makes up 90% or more of a typical total RNA sample.
- Library prep methods reduce the rRNA representation in a sample
  - oligoDt only binds to polyA tails to enrich a sample for mRNA
  - Ribo-depletion binds rRNA sequences

Neither technique is 100% efficient

Can screen (map reads to rRNA sequences) to determine rRNA efficiency and potentially remove those reads.

Effects in Variant Analysis

- Reads resulting from contaminants may align to the genome
- PCR duplicate reads are clearly redundant data
- Overlapping reads also produce redundant and possibly error prone basepairs
  - Freebayes does not pseudo overlap reads
  - Samtools
    - Sets one of the base pair qualities to 0 and the other to the sum of both qualities
    - In conflict, sets both qualities to 0
  - GATK Unified Genotype
    - Tries to model the reads
  - GATK Haplotype Caller
    - Sets one of the base pair qualities to 0 and the other to the sum of both qualities
    - In conflict, sets both qualities to 0

QA/QC

- Beyond generating 'better' data for downstream analysis, cleaning statistics also give you an idea as to the quality of the sample, library generation, and sequencing quality used to generate the data.
- This can help inform you of what you might do in the future.
- I've found it best to perform QA/QC on both the run as a whole (poor samples can affect other samples) and on the samples themselves as they compare to other samples (REMEMBER, BE CONSISTANT).
  - Reports such as Basespace for Illumina, are great ways to evaluate the runs as a whole.
  - PCA/MDS plots of the preprocessing summary are a great way to look for technical bias across your experiment