FINAL REVIEW

Equations

$$Q = -10 \log_{10} P$$ **Phred Score**

P = probability that the base is incorrect
Q = Phred quality score

Chastity

$$\frac{\text{Brightest Intensity}}{\text{Brightest Intensity} + \text{Second Brightest Intensity}} \geq 0.6$$

MACS2

$$\frac{\#\text{ of reads that align in input}}{\text{length of bin in nt}} = \lambda$$

STAR

$$\text{sequence read length} - 1 = \text{--sjdbOverhang}$$

$$\frac{\log_2(\text{Genome Length})}{2} - 1 = \text{--genomeSAindexNbases}$$

Clones

$$n = \frac{\ln(1-P_o)}{\ln(1-f)}$$

n = number of clones in a gene library
P_o = desired probability of gene in library
f = fraction of genome in one insert

Read Coverage

$$R_N = \frac{CI}{rL(P_f)}$$

R_N = # of reads needed to complete target sequence
C = depth of genome coverage
I = length of target DNA sequence in bases
rL = average length of a read (i.e. $Q > 20$)
P_f = pass rate, fraction of reads above Q threshold

Nucleotide Recovery

$$P_o = e^{-c} = e^{-\frac{LN}{G}}$$ L = read length

P_o = probability that a base is NOT sequenced
C = fold sequence coverage $= \frac{LN}{G}$

N = number of reads sequenced
G = length of target DNA sequence in bases

$$Gm = \sum_{i=1}^{\ell} n_i G_i$$

Metagenomes

Gm = metagenome size in bases
ℓ = number of genomes in sample
n_i = number of copies of genome G_i
G_i = size of any given genome in sample of ℓ genomes

Distributions

Poisson Distribution

- emulsion clonal amplification
- MACS2 (read count distribution)
- Lander-Waterman model (nucleotide recovery)

Normal Distribution

- insert size
- microarrays

Negative Binomial Distribution

- RNA seq count data

Algorithms

BWA

index the reference into 32mer table

↓

extract first (5') 32 bases of read sequence (seed) and matched to index table (2 mismatches allowed in the seed region)

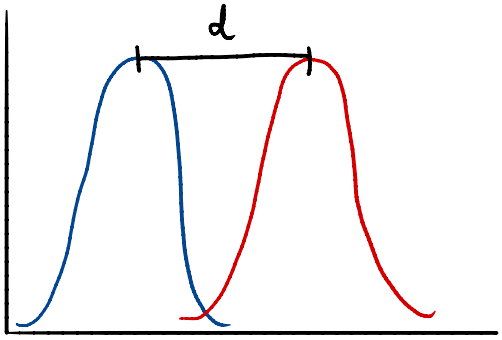
↓

extension from seed (3' only) (mismatches allowed in extension set by parameters)

↓

read is assigned a mapping quality

MACS2



find significantly enriched bins (2 x average fragment size) with counts 'mfold' higher than random genome average

↓

for 1000 randomly chosen enriched bins, calculate the difference between max of distribution of (+) and (−) read starts

↓

shift all (+) read starts by +d/2 and (−) reads by by −d/2

↓

scale control experiment to the same number of reads as IP

↓

scan genome with 2d bins and calculate the mean number of sequence reads from IP that align within it $\frac{\#\text{ of reads}}{2d}$

↓

calculate the mean number of mean number of sequence reads from INPUT for the next 1kb, 5kb, 10kb, and genome-wide. (max = λ)

↓

Using Poisson distribution for λ, calculate p-value using IP mean

↓

Estimate an empirical FDR for each bin (Benjamini-Hochberg)

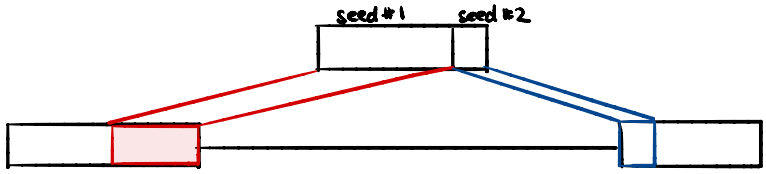
STAR

SEED SEARCHING

for each read, find longest sequence that matches exactly one or more locations on reference (MMPs)

↓

different parts of the reads that are mapped separately are "seeds"



↓

search ONLY for unmapped portion of read to find the next longest sequence that matches the reference

↓

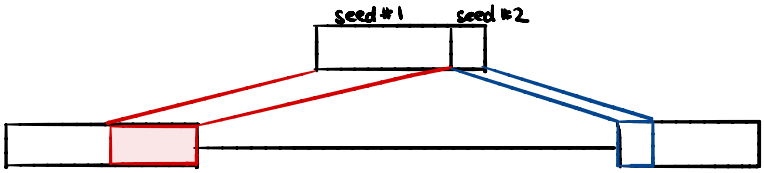
extend previous MMP/seeds to accomodate mismatches

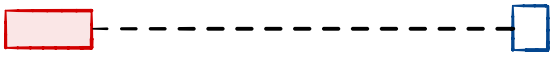
↓

Separate seeds are stitched together to create a complete read

↓

cluster seeds together based on how close they are to a set of "anchor" seeds (uniquely mapped seeds)



stitched read 

DESeq2

Take the natural log of all the values

↓

Average each row (=gene)

↓

Filter out genes with infinity

↓

Subtract the average ln value from ln(counts

↓

Calculate the median of the ratios for each sample

↓

Convert the medians to get the final scaling factors for each sample

↓

Divide the orignal read counts by the scaling factor

remove repeats, mask low complexity regions based on a priori knowledge and sequence quality scores

↓

identify overlaps between reads at user defined length and identity thresholds

↓

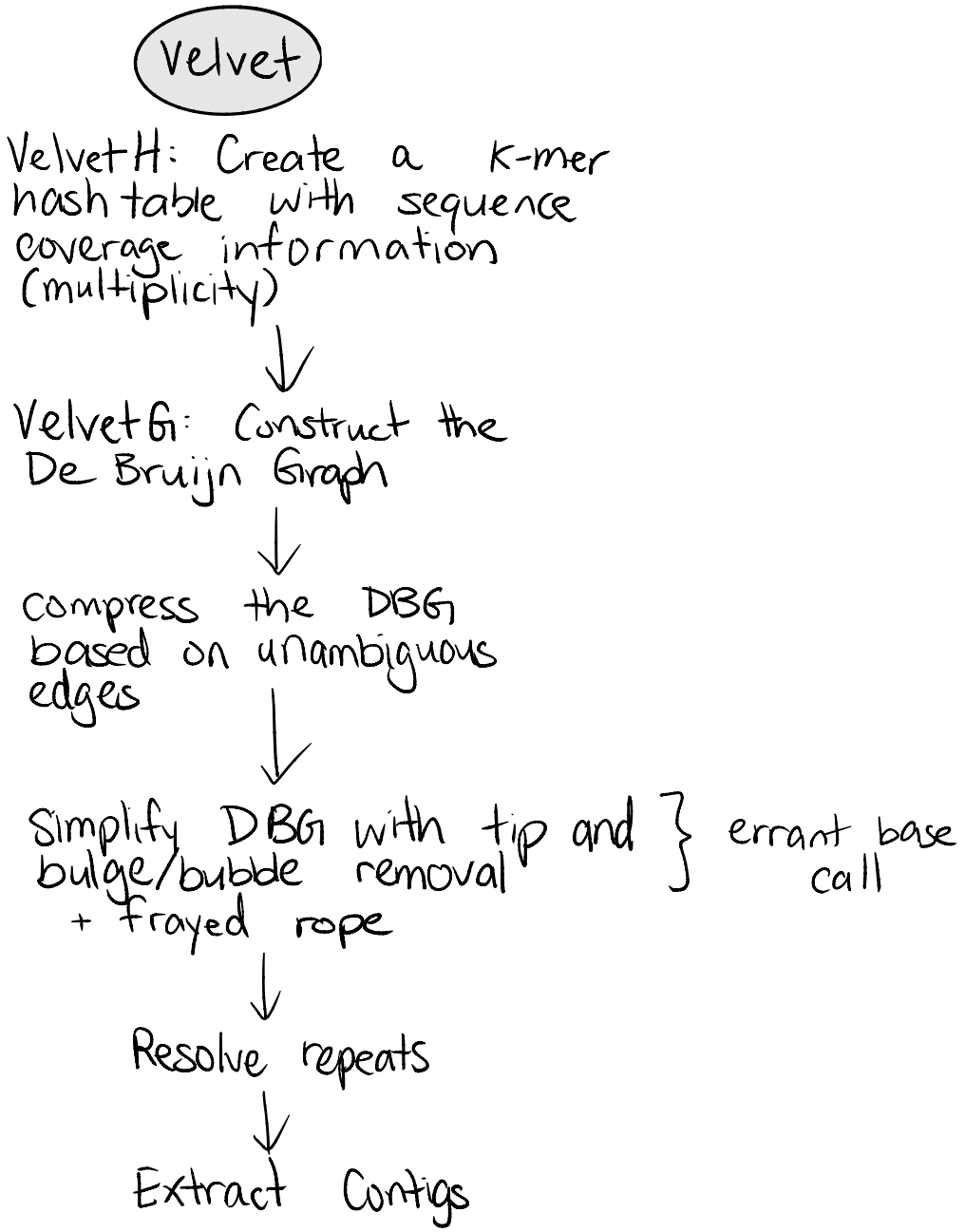
assemble high confidence sequences from overlapping reads

↓

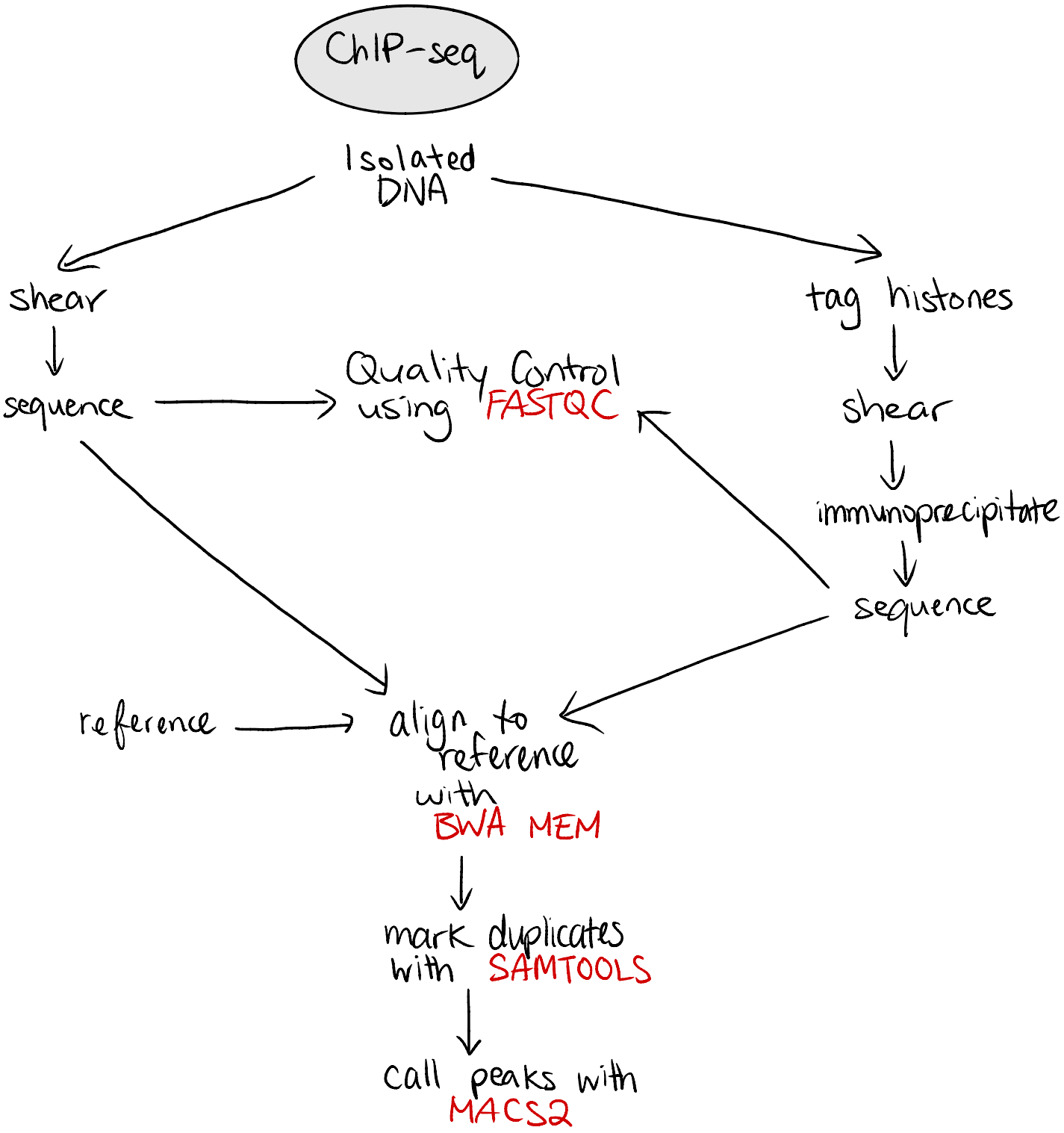
order and orient contigs using mate pair information (fill gaps with Ns)

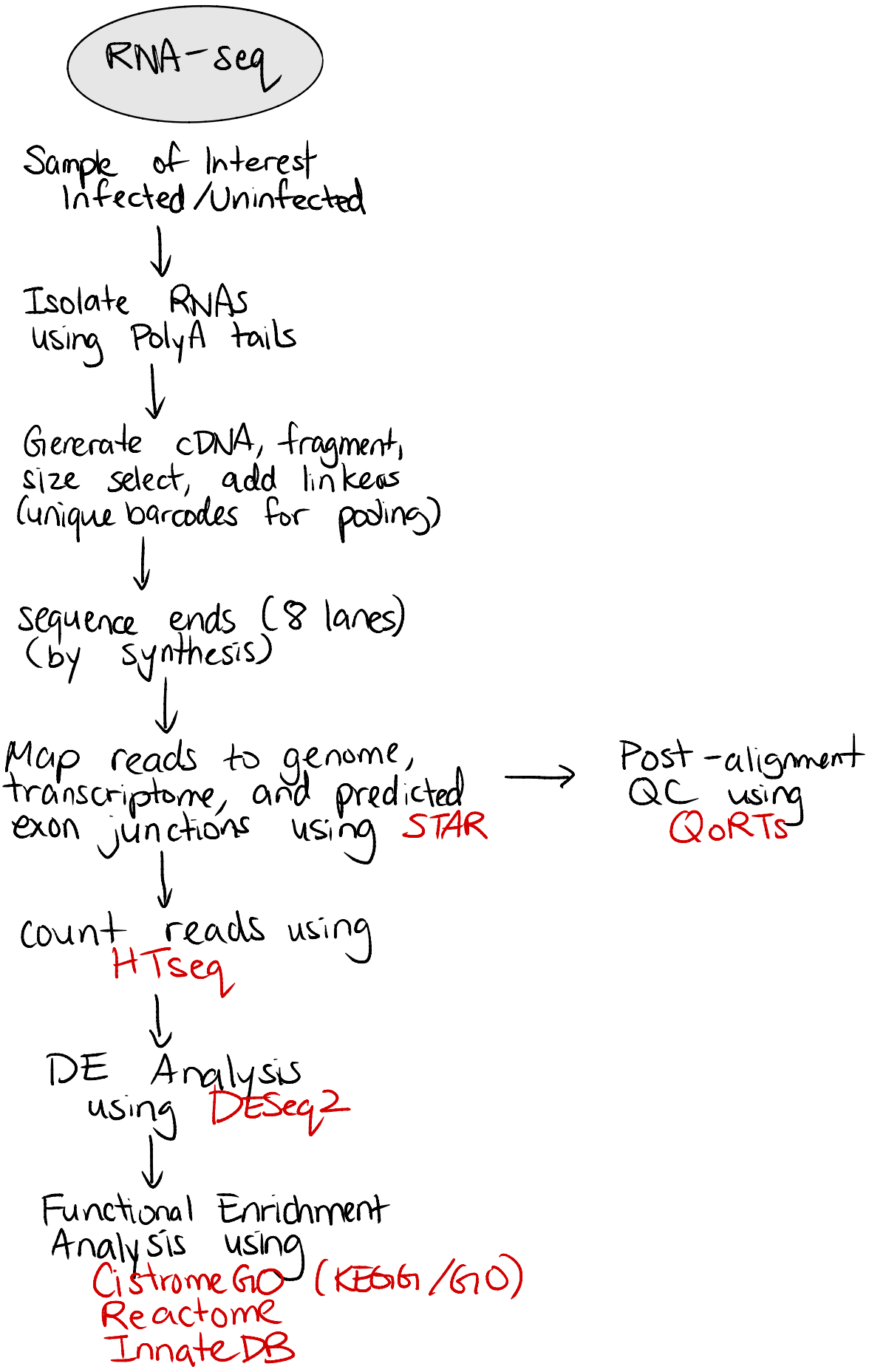
↓

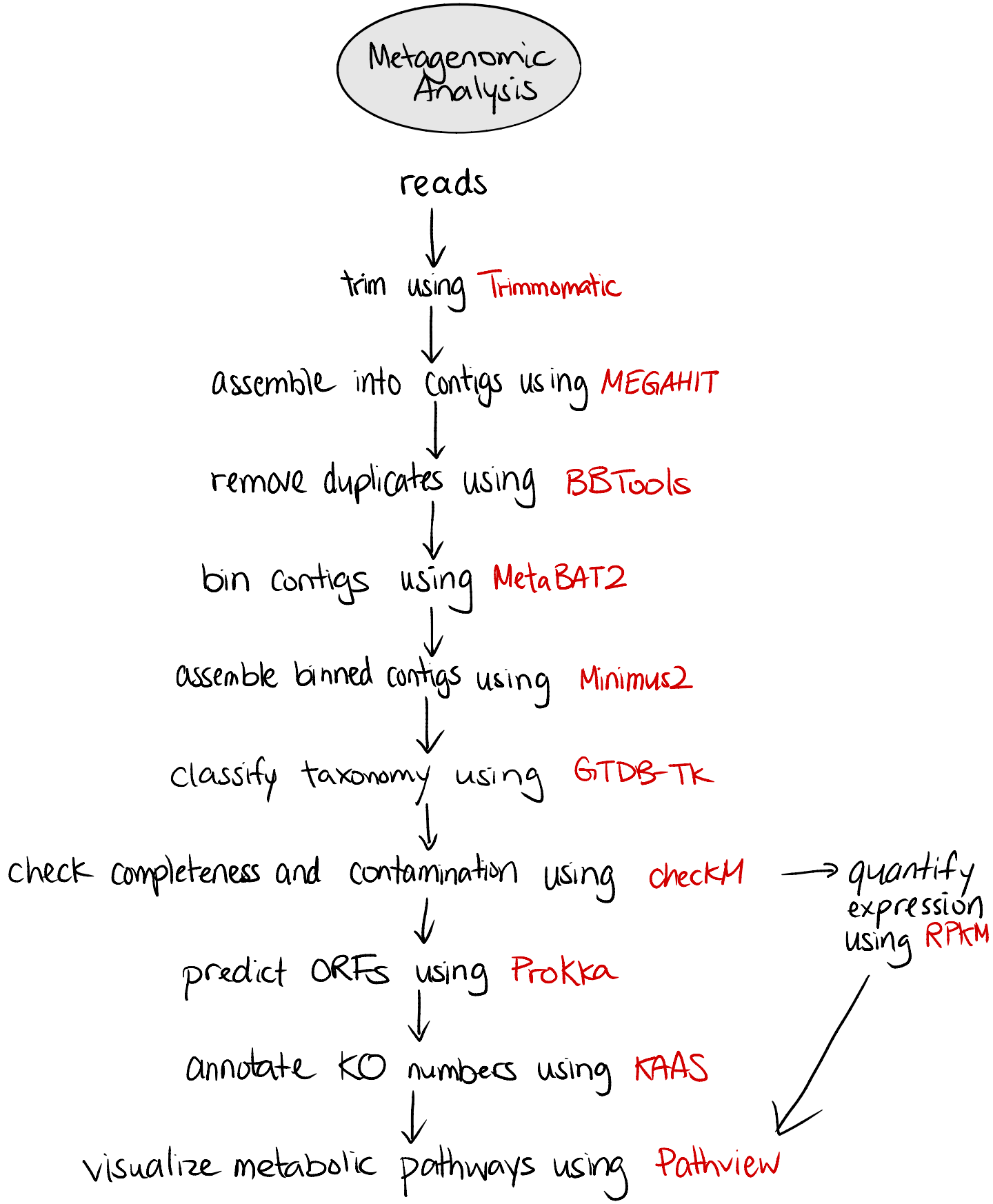
attempt to resolve sequencing errors


Velvet
VelvetH: Create a k-mer hash table with sequence coverage information (multiplicity)
VelvetG: Construct the De Bruijn Graph
compress the DBG based on unambiguous edges
Simplify DBG with tip and bulge/bubble removal + frayed rope
} errant base call
Resolve repeats
Extract Contigs

Workflows


ChIP-seq
Isolated DNA
shear
sequence
tag histones
shear
immunoprecipitate
sequence
Quality Control using FASTQC
reference
align to reference with BWA MEM
mark duplicates with SAMTOOLS
call peaks with MACS2


RNA-seq
Sample of Interest
Infected/Uninfected
Isolate RNAs
using PolyA tails
Gererate cDNA, fragment,
size select, add linkers
(unique barcodes for pooling)
sequence ends (8 lanes)
(by synthesis)
Map reads to genome,
transcriptome, and predicted
exon junctions using STAR
Post-alignment
QC using
QoRTs
count reads using
HTseq
DE Analysis
using DESeq2
Functional Enrichment
Analysis using
Cistrome GO (KEGG/GO)
Reactome
InnateDB


Metagenomic Analysis
reads
trim using Trimmomatic
assemble into contigs using MEGAHIT
remove duplicates using BBTools
bin contigs using MetaBAT2
assemble binned contigs using Minimus2
classify taxonomy using GTDB-Tk
check completeness and contamination using checkM
quantify expression using RPKM
predict ORFs using Prokka
annotate KO numbers using KAAS
visualize metabolic pathways using Pathview

DESeq2 Normalization

Gene	Sample 1	Sample 2	Sample 3
A	0	10	4
B	2	6	12
C	33	55	200

Gene	ln(Sample 1)	ln(Sample 2)	ln(Sample 3)
A	NA	2.3	1.4
B	0.7	1.8	2.5
C	3.5	4.0	5.3

Gene	Average of ln
A	NA
B	1.7
C	4.3

Gene	Average of ln
B	1.7
C	4.3

Gene	ln(Sample 1)	ln(Sample 2)	ln(Sample 3)
B	-1	0.1	0.8
C	-0.8	-0.3	1

Gene	ln(Sample 1)	ln(Sample 2)	ln(Sample 3)
B	-1	0.1	0.8
C	-0.8	-0.3	1
Median	-0.9	-0.1	0.9

Gene	ln(Sample 1)	ln(Sample 2)	ln(Sample 3)
Scaling Factor	0.4	0.9	2.5

Gene	Sample 1	Sample 2	Sample 3
A	0	11.1	1.6
B	5	6.7	4.8
C	82.5	61.1	80