GENOME ASSEMBLY
FINAL PIPELINE AND RESULTS

Faction 1

Yanxi Chen | Carl Dyson | Sean Lucking | Chris Monaco
Shashwat Deepali Nagar | Jessica Rowell | Ankit Srivastava
Camila Medrano Trochez | Venna Wang | Seyed Alireza Zamani

CREATING THE NEXT®

OBJECTIVES

- Research industry standard tools for cleaning reads, assembling genomes, and evaluating the assemblies
  - Identify and evaluate new tools
- Devise a pipeline for read quality assessment, cleaning, assembly, evaluation, and polishing
- Assemble reads using two standard methods
- Compare results and choose the best assembly, considering the downstream analyses planned

DATA AND BACKGROUND

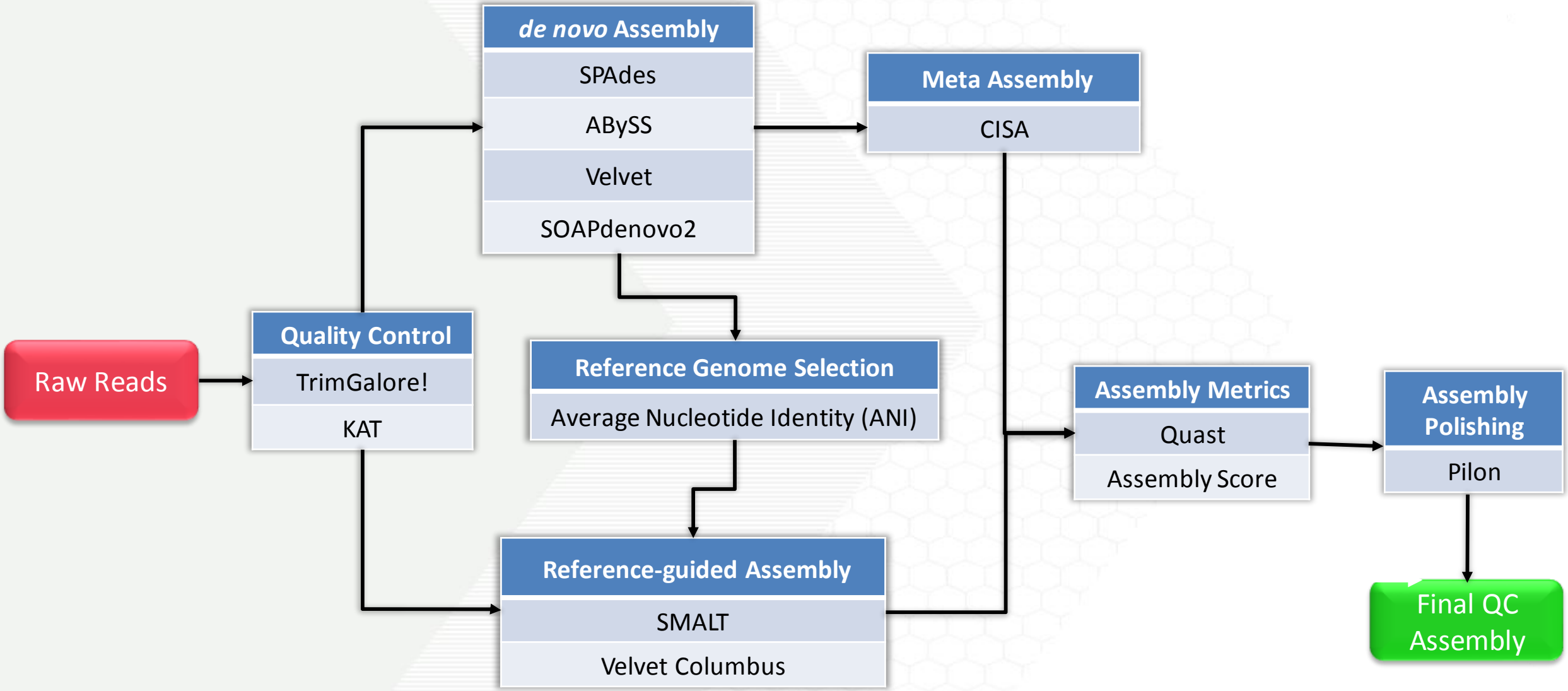
What do we have?

- Reads for 24 isolates of *Salmonella enterica* subsp. *enterica* serovar Heidelberg from an outbreak in 2013
- Short paired-end reads sequenced using Illumina MiSeq (2nd Generation)

Species characteristics

- GC Content: 52.8% | 1 chromosome + plasmids | Genome size 4.7 - 5.1 Mb

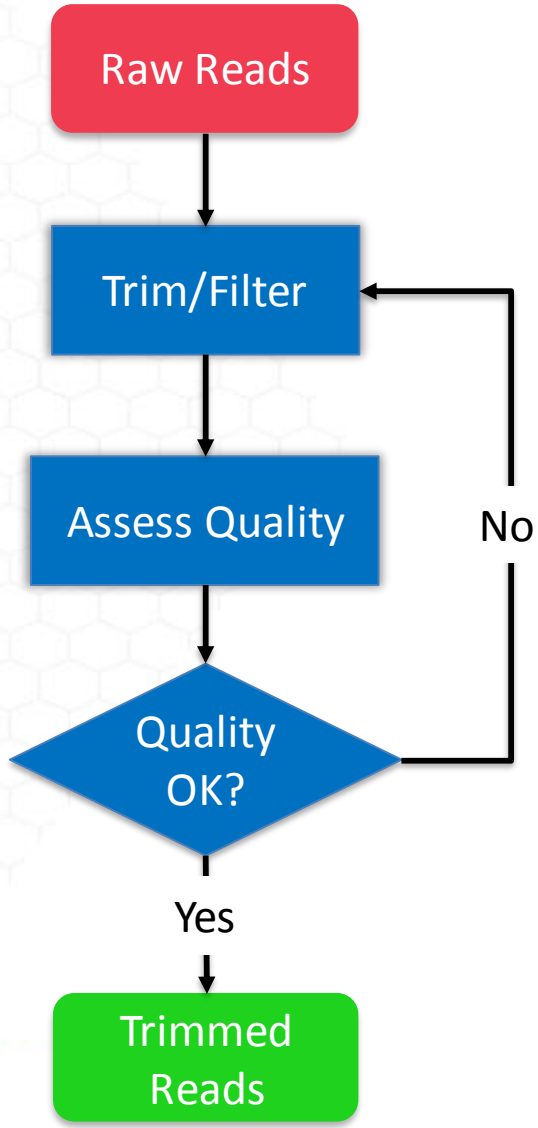
FINAL PIPELINE



QUALITY CONTROL

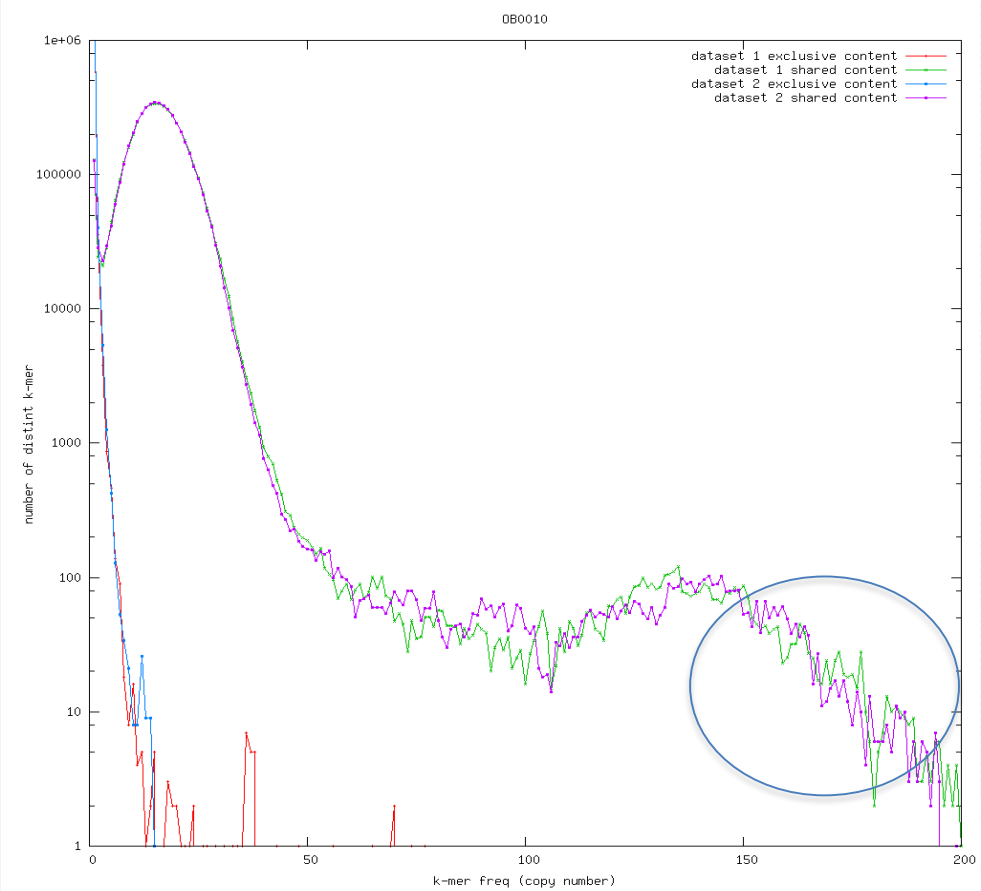
- Did iterative blanket trimming and filtering of reads using TrimGalore! and assessed the quality using FastQC

- There were a few problematic samples, which were handled separately (OB0012, OB0015, OB0022)

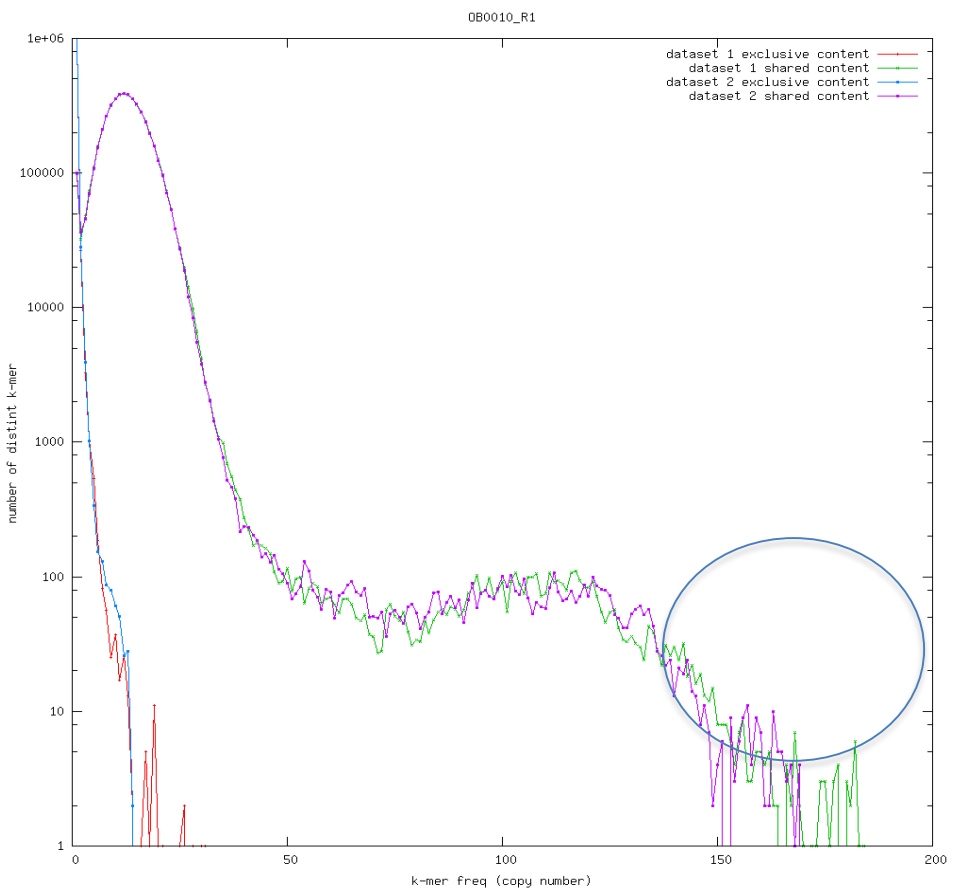


QUALITY CONTROL - KAT

Raw data (left) has strong divergence at low kmer frequency and the slight shift to lower kmer frequencies compared to the corrected (right) graph



Before trimming



After trimming

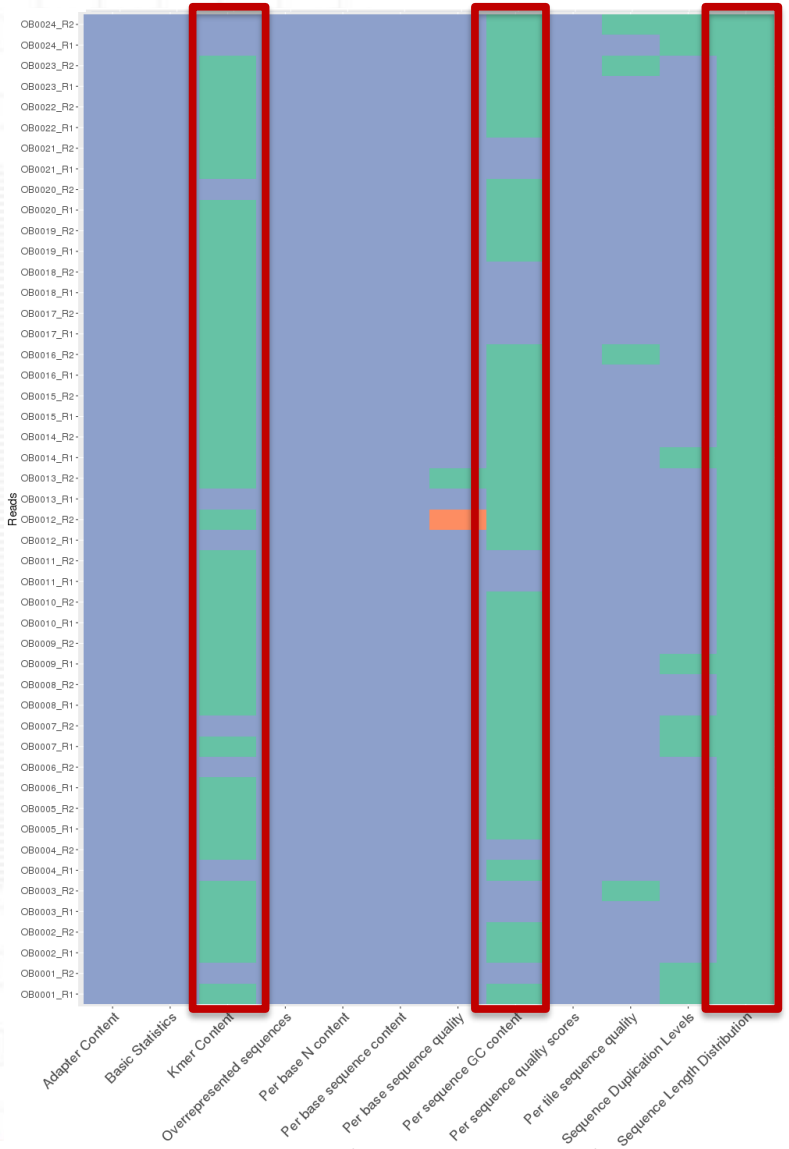
QUALITY CONTROL - FASTQC

Raw reads

Kmer Content

GC Content

Sequence Length Distribution



OK

Warning

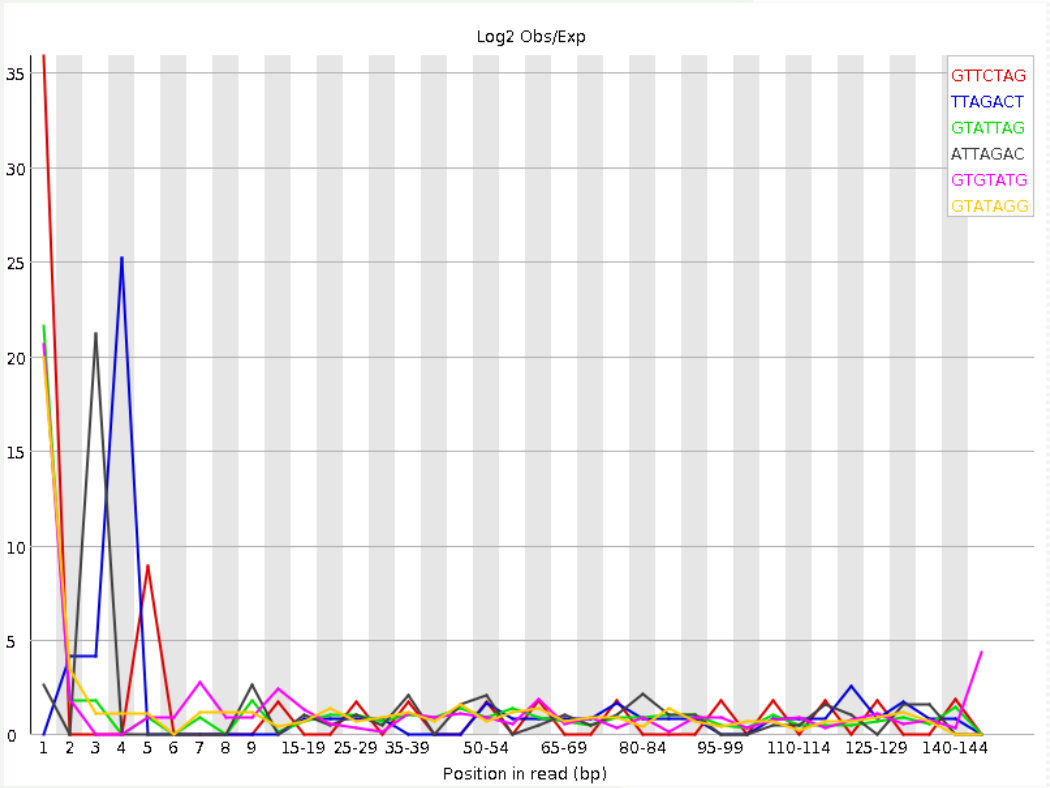
Fail

Clip: 10 on 5' end, 5 on 3' end

Length cutoff: 150

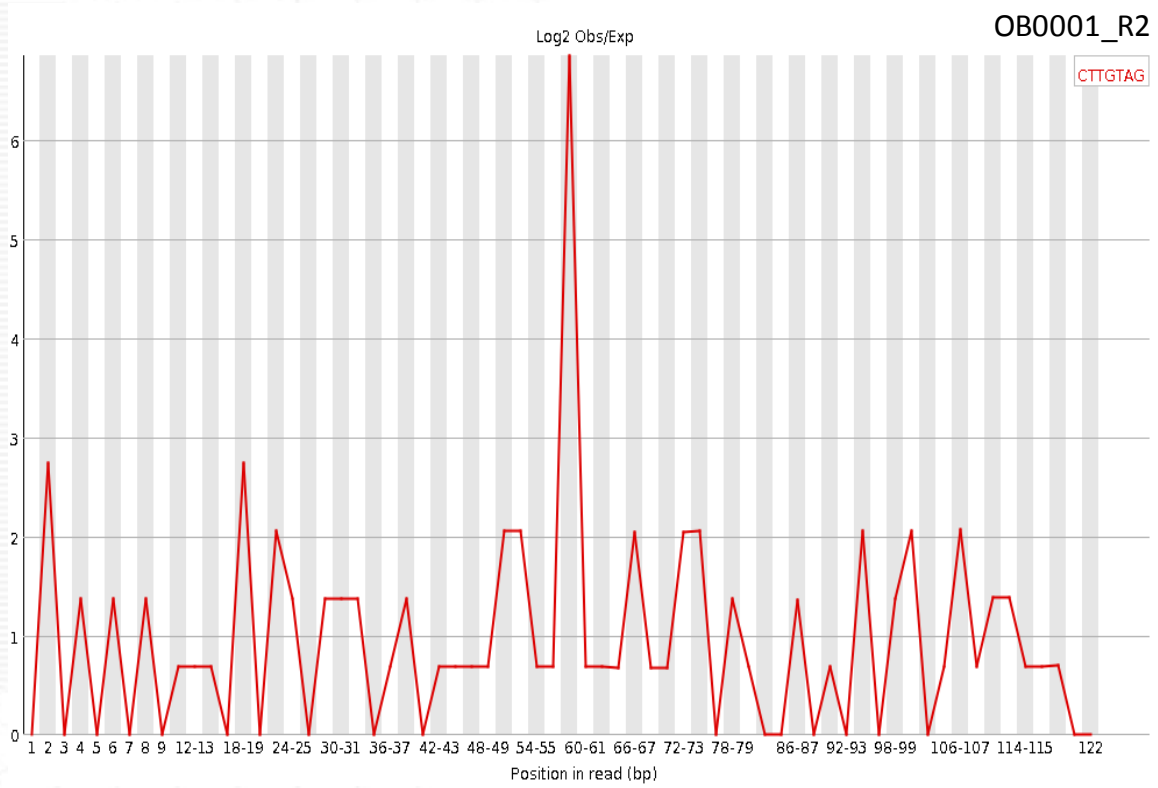
KMER CONTENT: A CLOSER LOOK

- Warning occurs if any kmer is imbalanced with a binomial $10^{-5} < \text{p-value} < 10^{-2}$



Before Trimming

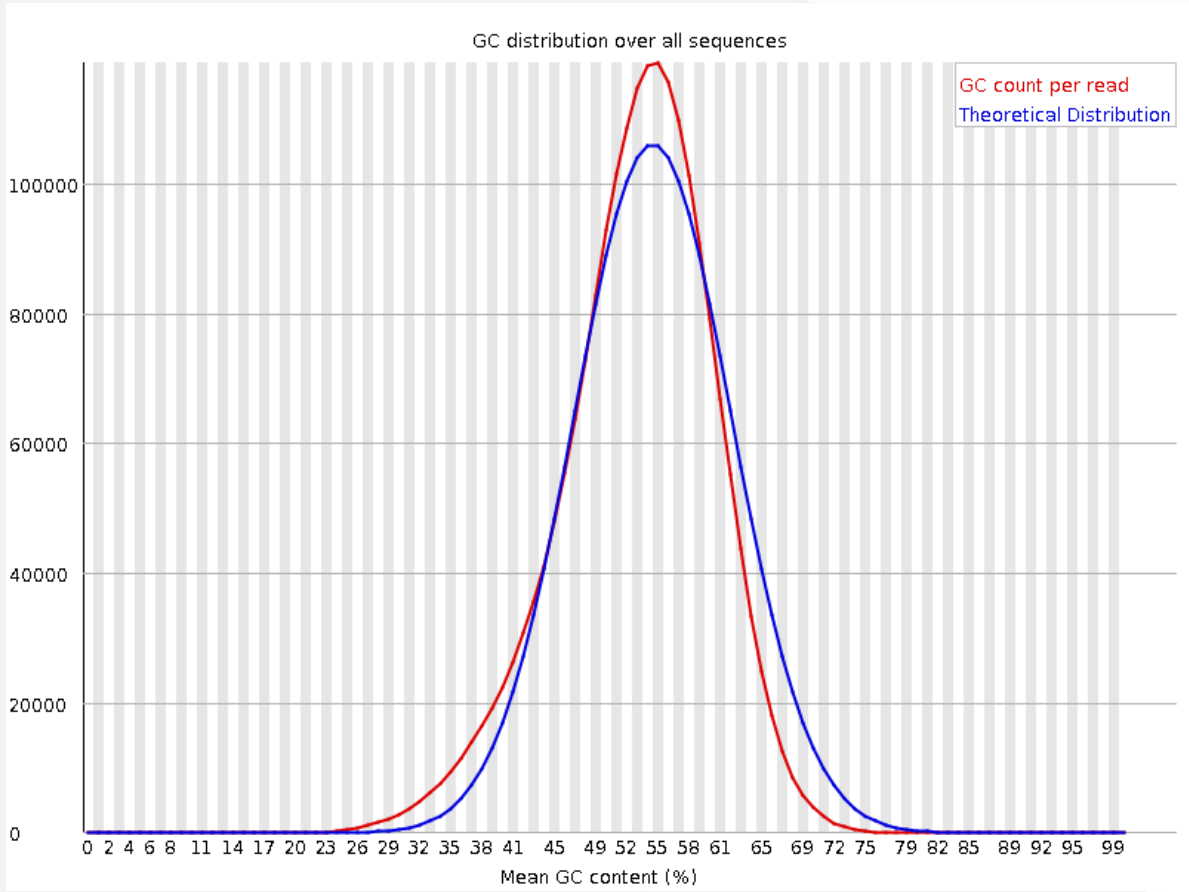
OB0001_R2

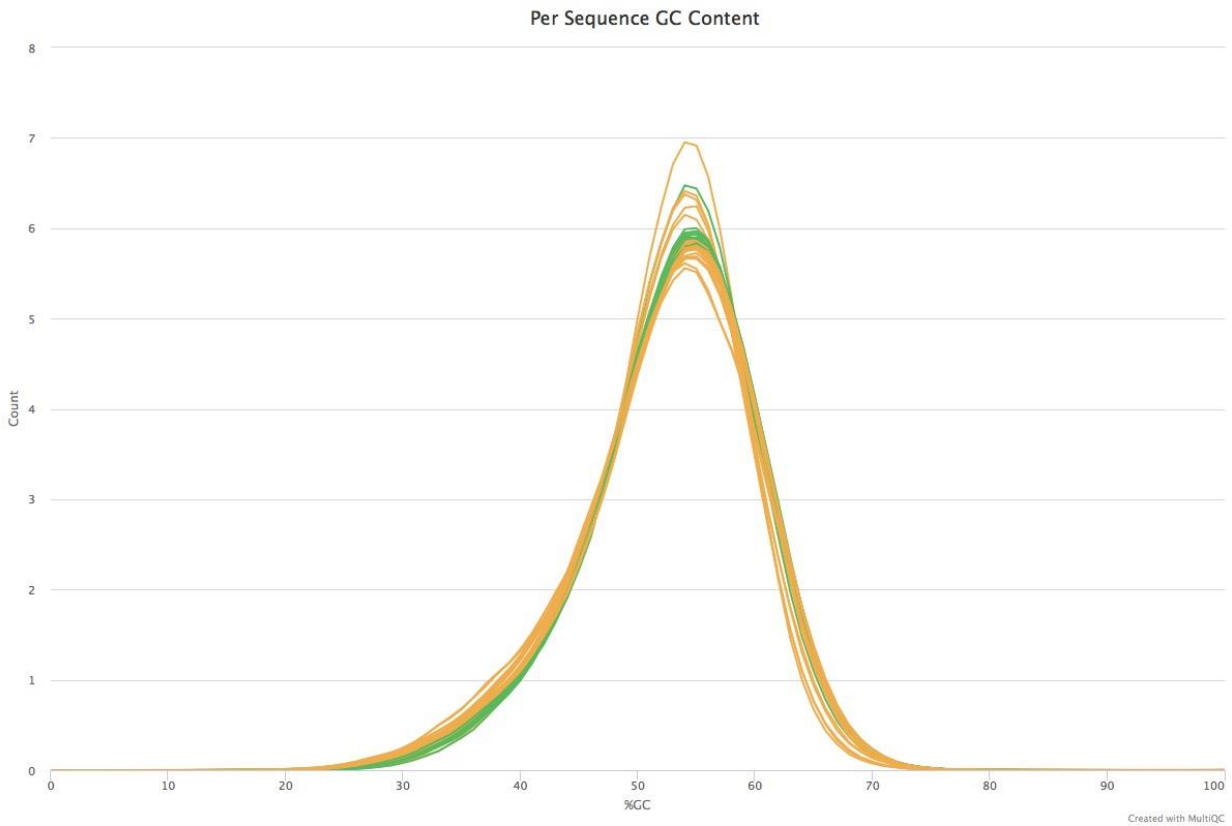


After Trimming

GC CONTENT: A CLOSER LOOK

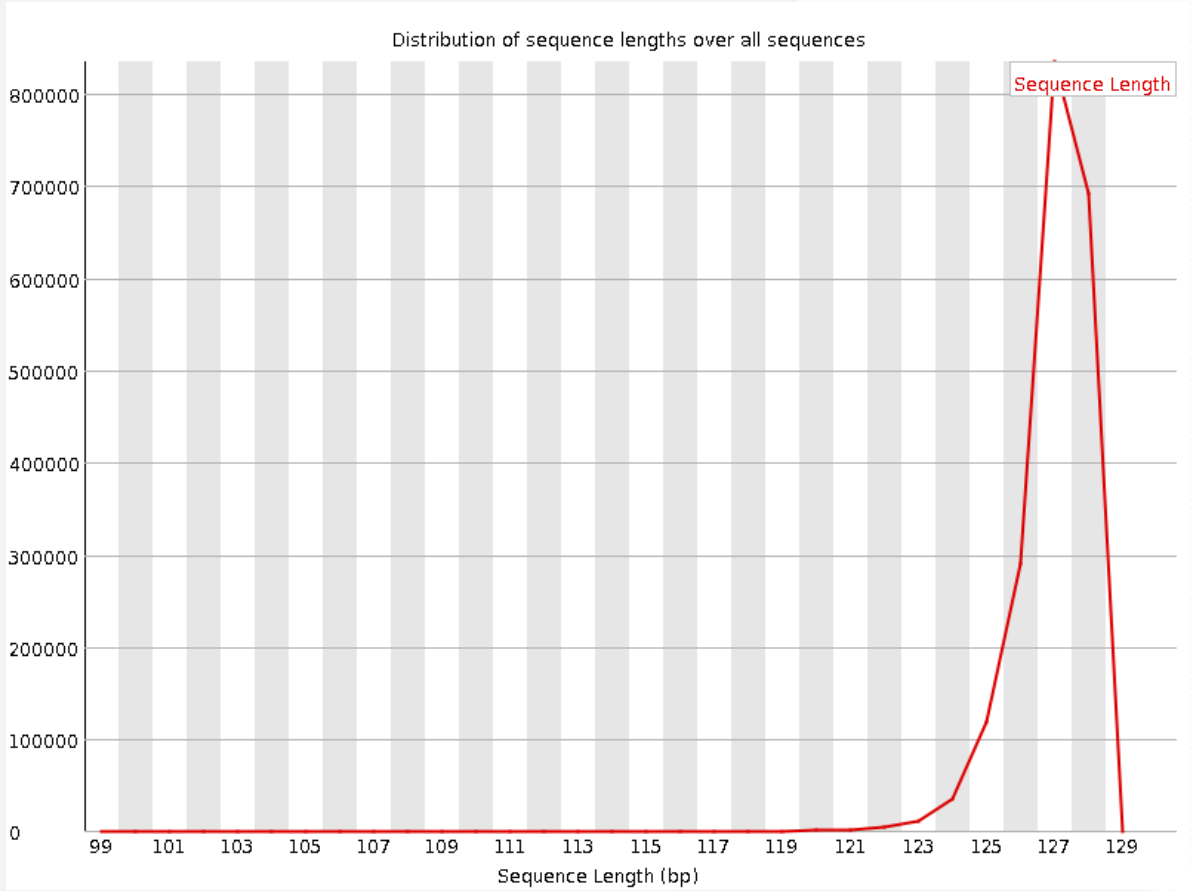
- Warning occurs if sum of deviations from theoretical (Gaussian) distribution differs by between 15% and 30%

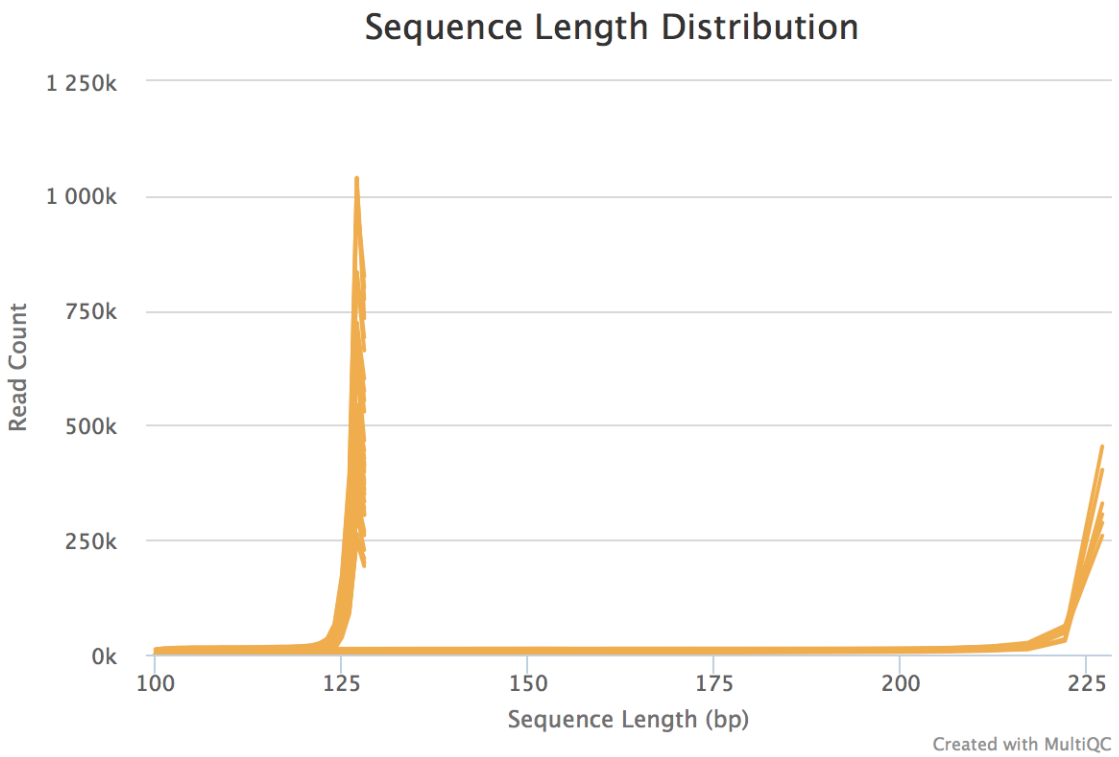




SEQUENCE LENGTH DISTRIBUTION: A CLOSER LOOK

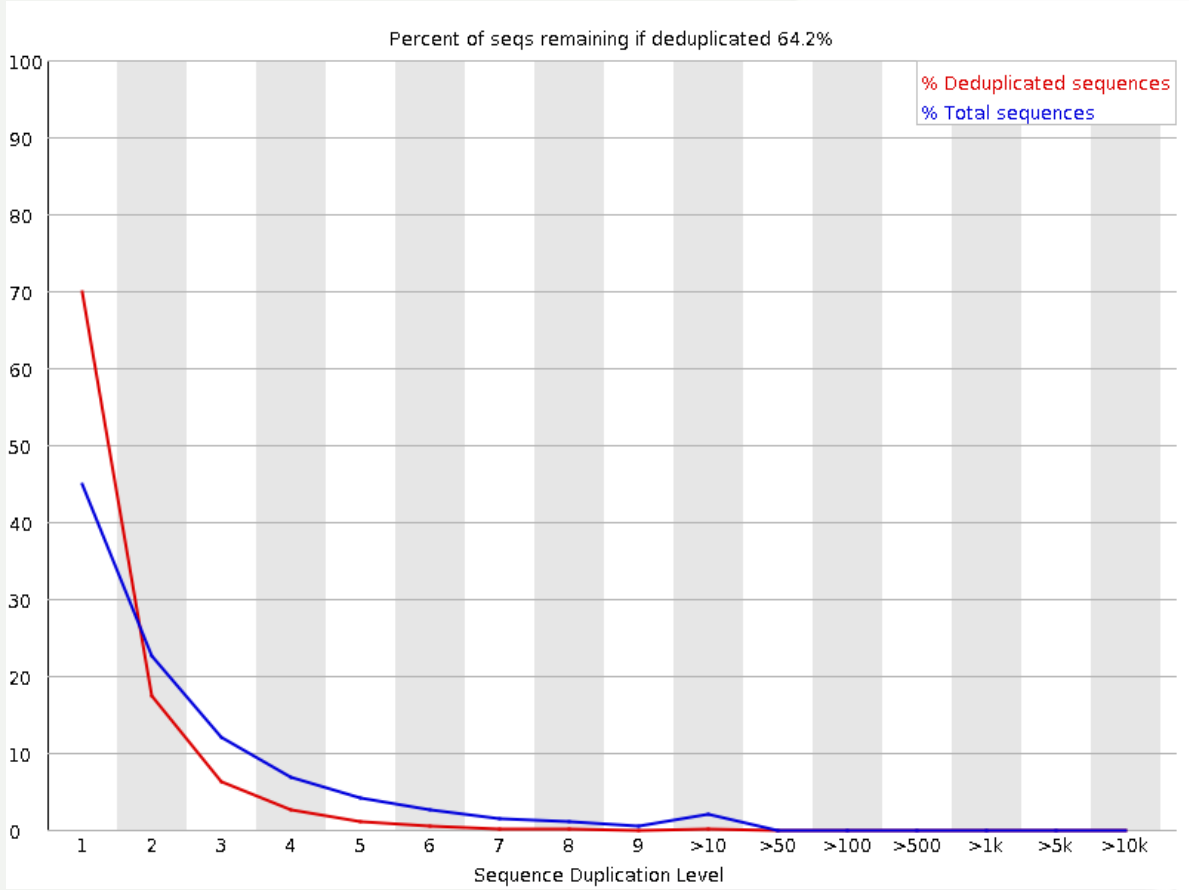
- Warning occurs if all the sequences are not of the same length

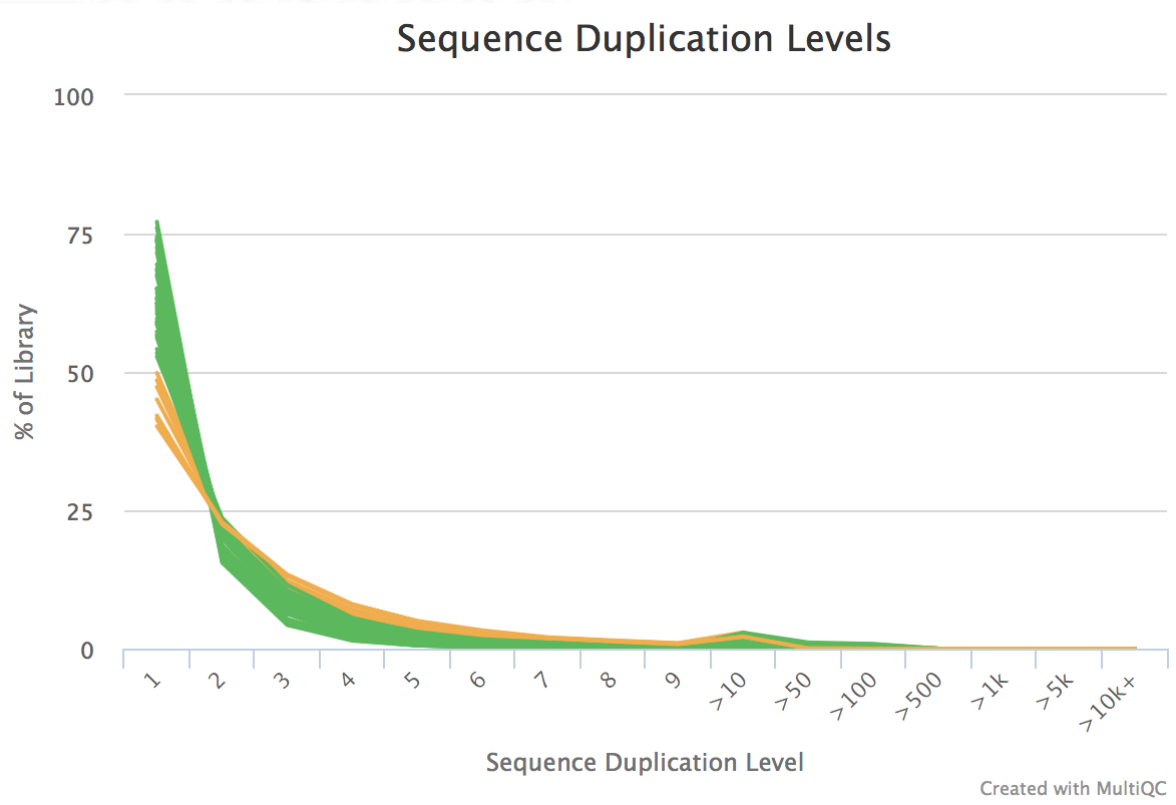




SEQUENCE DUPLICATION LEVELS: A CLOSER LOOK

- Warning occurs if non-unique sequences are between 20% and 50% of total





DE NOVO ASSEMBLERS

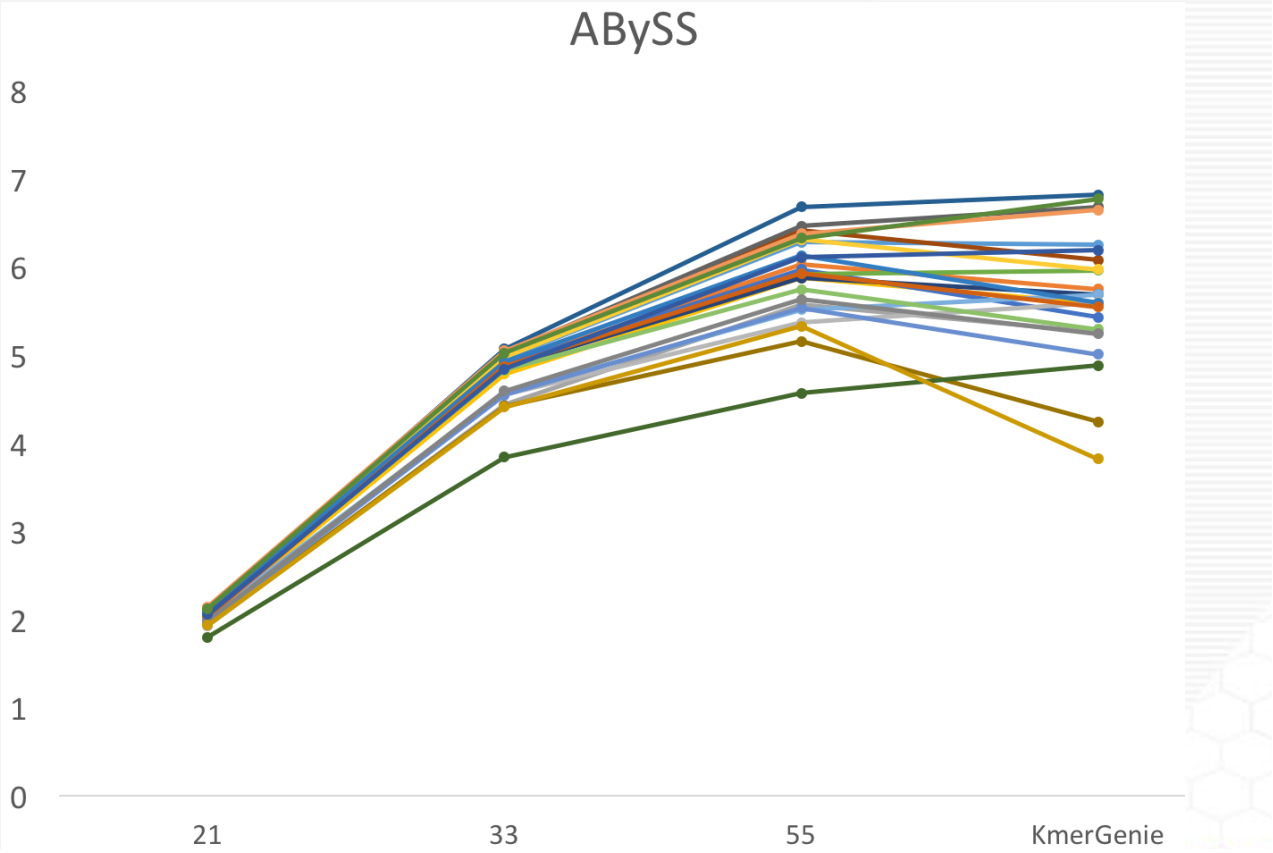
- Used four De Bruijn graph based *de novo* assemblers: ABySS, SOAPdenovo2, SPAdes, and Velvet

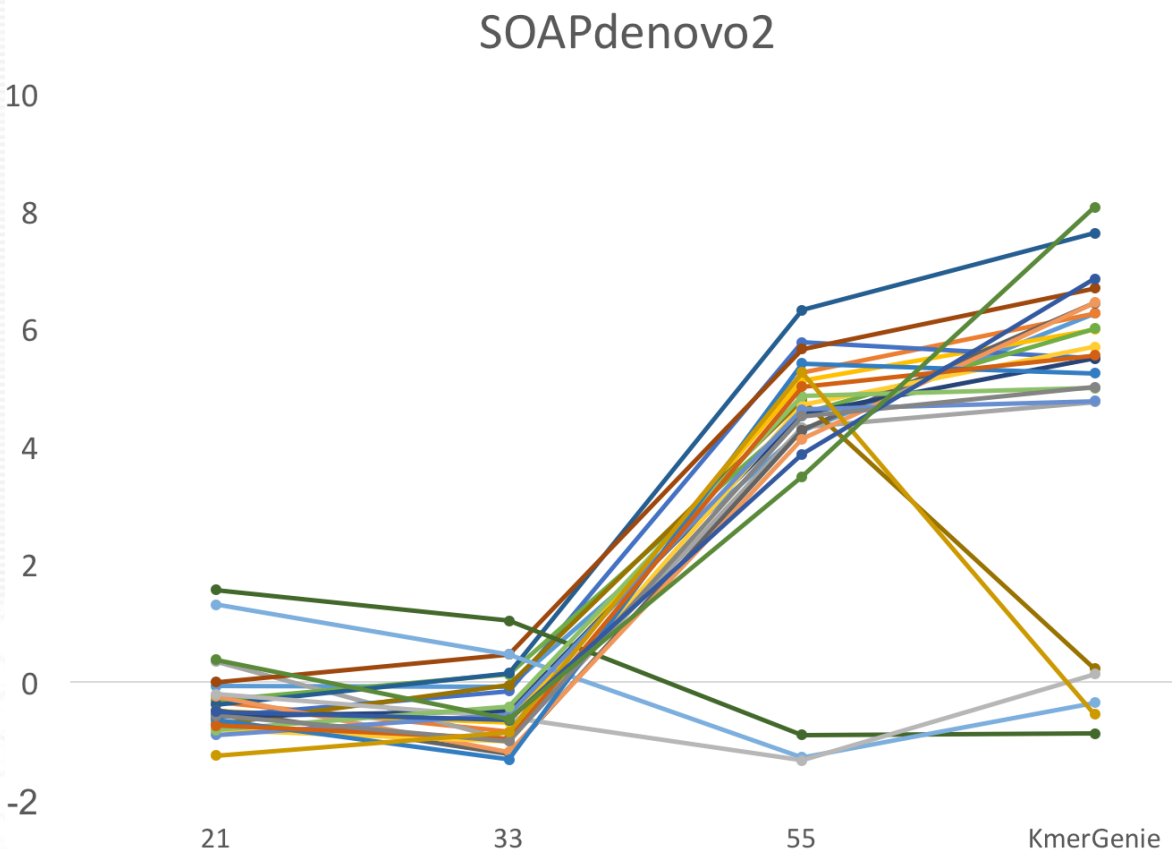
- ABySS and SOAPdenovo2 require kmer size as input
  - Used KmerGenie for getting the optimal kmer size

KMERGENIE: OPTIMAL KMER SELECTION

- KmerGenie gave **k=71** as the optimal kmer size for all the samples, except for two
- For validating the KmerGenie prediction, we ran the assemblers for different kmer sizes

$$\text{score} = \log\left(\frac{N50 \times \%\, cov}{\#\ of\ contigs}\right)$$





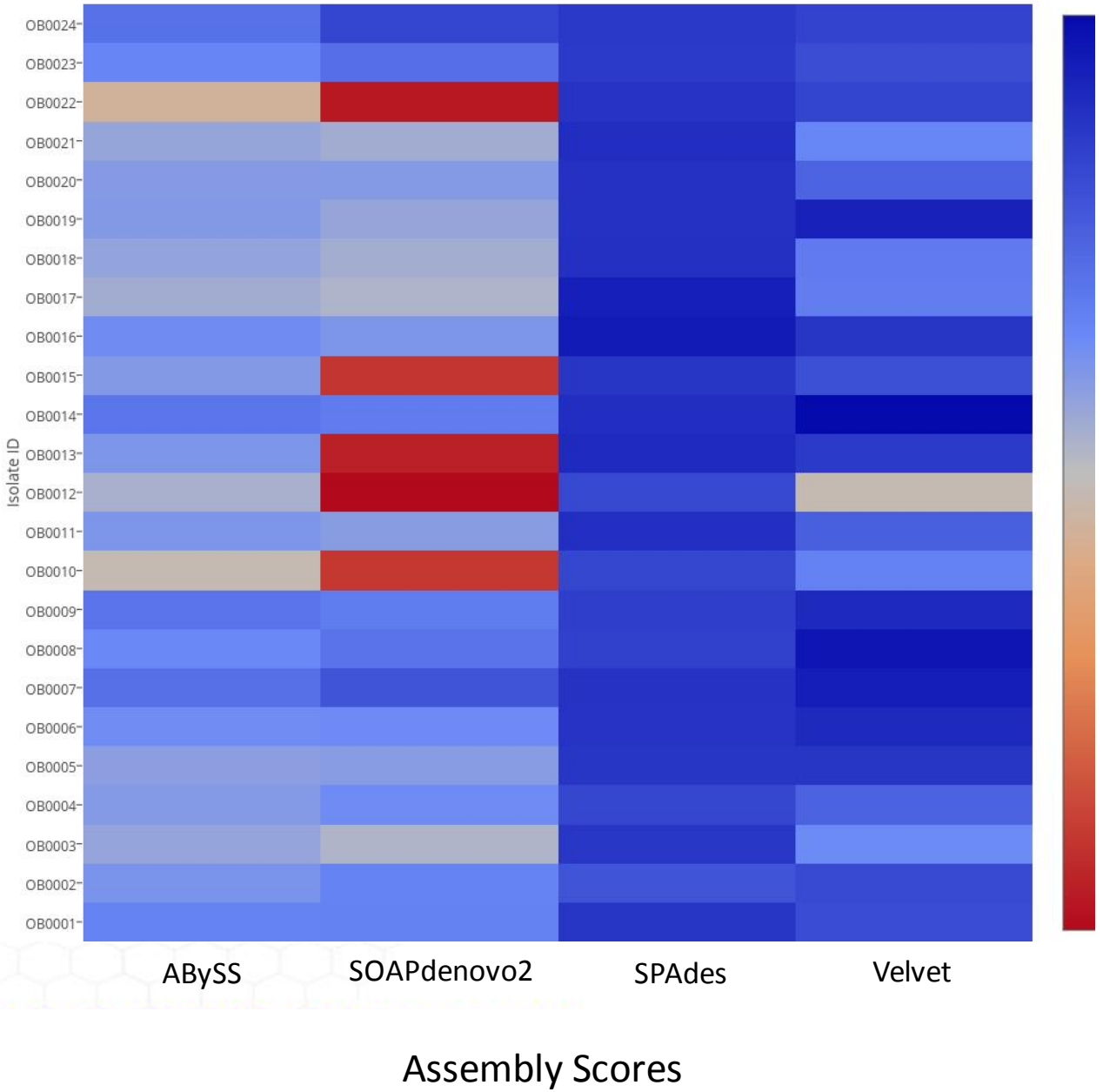
Assembly scores versus kmer sizes

RESULTS: *DE NOVO* ASSEMBLY

- SPAdes gave the best assembly scores among the *de novo* assemblers

$$\text{score} = \log\left(\frac{N50 \times \%\, cov}{\#\, of\ contigs}\right)$$

- Some samples scored consistently low across all the assemblers (OB0012, OB0015, OB0022)

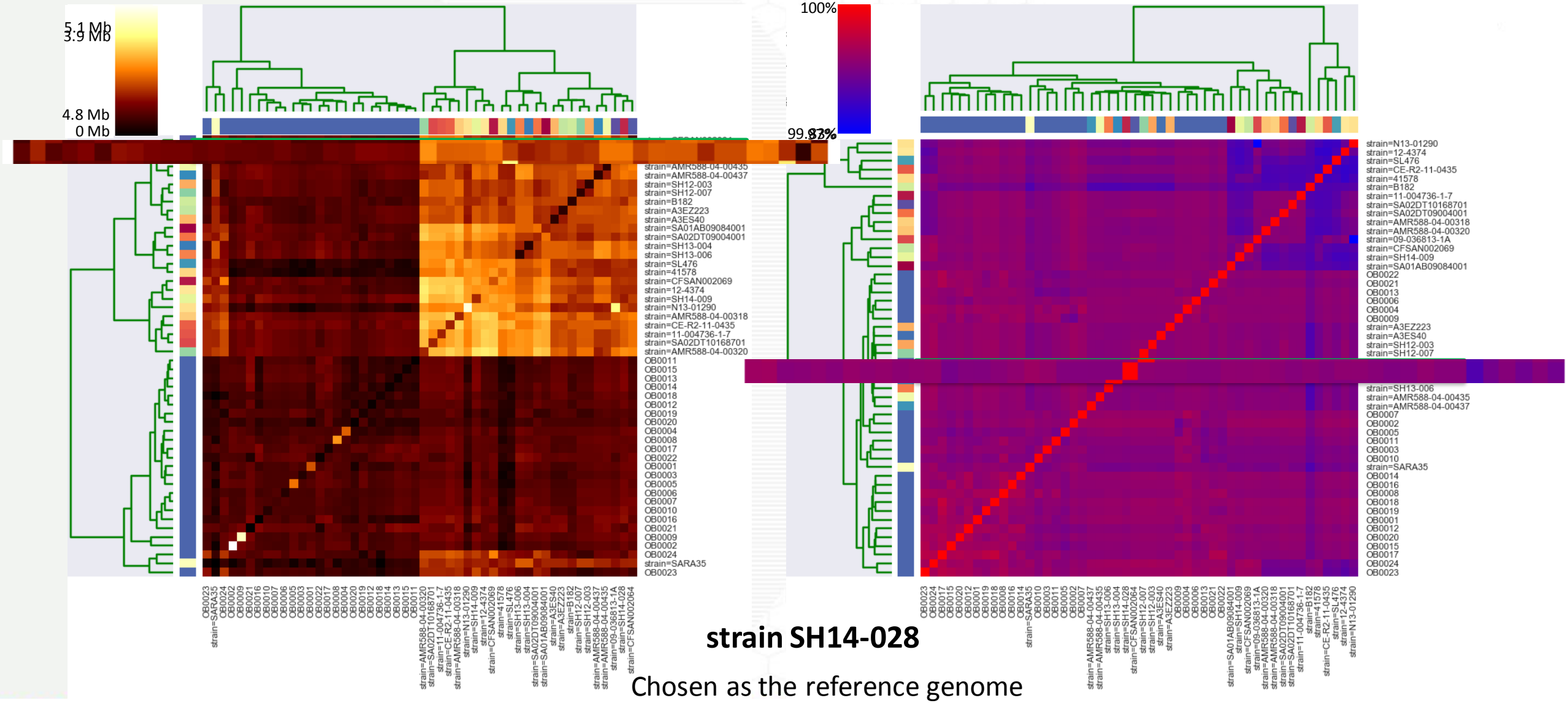


CHOOSING A REFERENCE GENOME

- For choosing a reference genome, we compared 26 different assemblies of *S. enterica* ser. Heidelberg with our *de novo* assembled samples

- Used a Python library called *pyani* for aligning whole genomes against each other (using MUMmer) and calculating average nucleotide identity (ANI)
- Relied on two parameters for making the choice:
  - Alignment length
  - Percentage Identity

CHOOSING A REFERENCE GENOME: ANI



strain SH14-028

Alignment length

Chosen as the reference genome for reference-based assembly

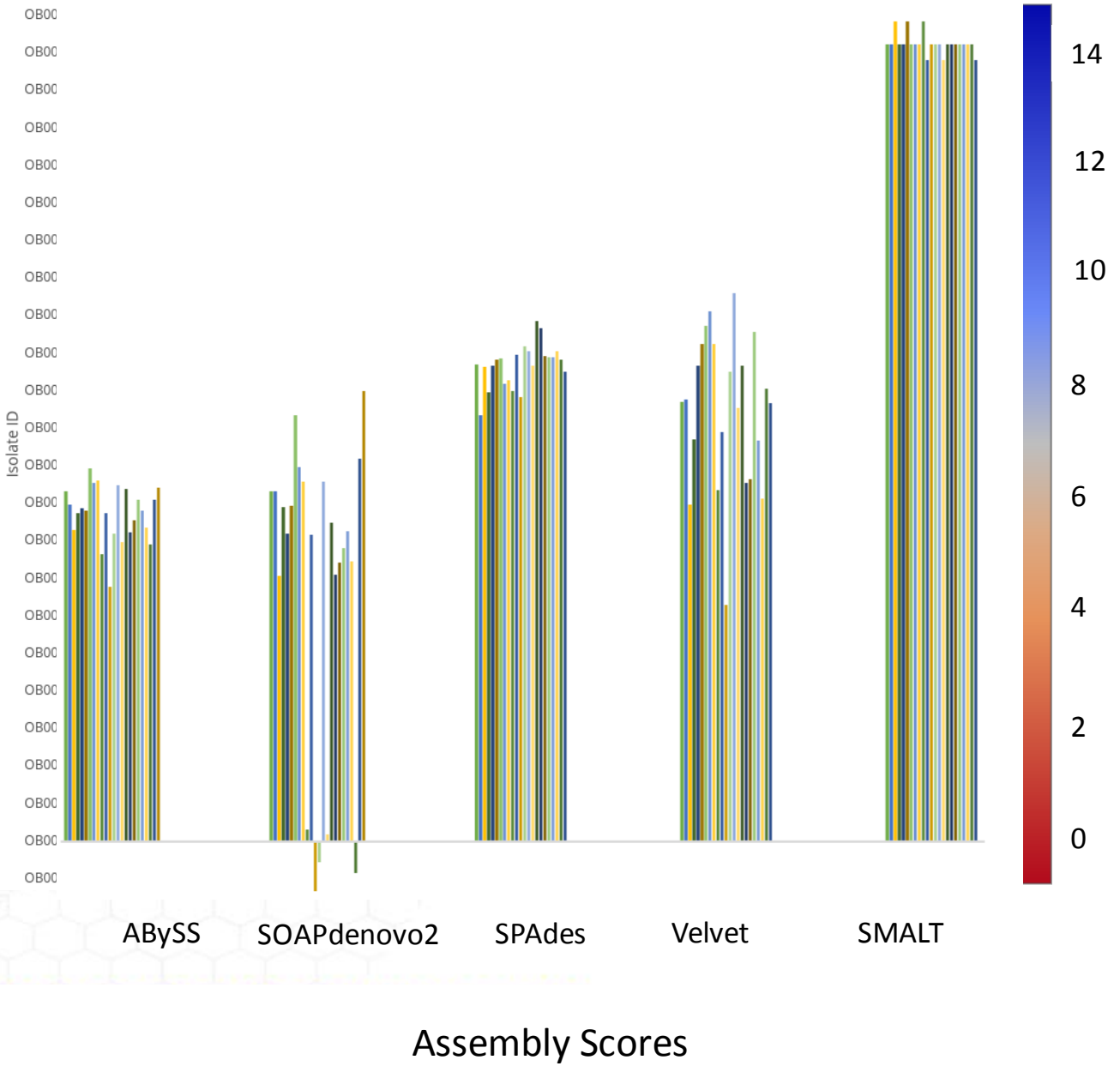
Percentage identity

RESULTS: SMALT VS. *DE NOVO* ASSEMBLERS

- SMALT assembly scores were consistent across all the samples, and better than those for *de novo* assemblies

$$\text{score} = \log\left(\frac{N50 \times \%\, cov}{\#\ of\ contigs}\right)$$

- SPAdes and SMALT scores were consistent across all the samples, whereas SOAPdenovo2 and Velvet scores showed a lot of variation

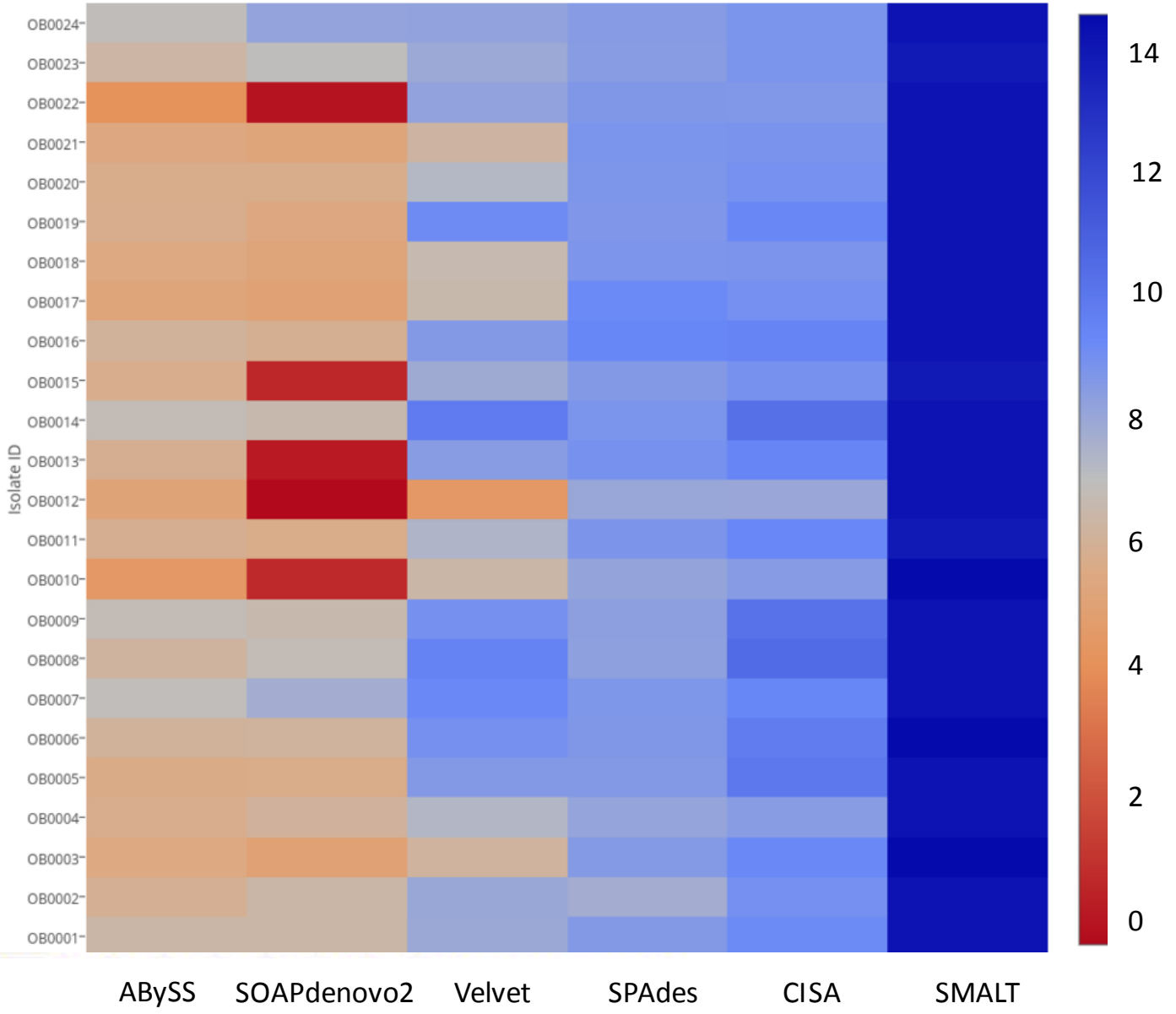


META ASSEMBLY

- Every assembly tool implements its own features for selecting the assembly
- We used a meta-assembly tool to integrate contigs from all the *de novo* assemblers
- We chose CISA (Contig Integrator for Sequence Assembly)
  - Specific for paired end reads
  - Implements 4 phases to integrate contigs:
    1. Identification of representative contigs and possible extensions
    2. Removal and splitting of misassembled contigs
    3. Iterative merging of contigs with a minimum 30% overlap
    4. Merging of contigs based on size of repetitive regions

META ASSEMBLY: CISA

- CISA-generated meta-assembly scores were higher than scores for *de novo* assemblies
- SMALT scores were higher than CISA scores
- Meta-assemblies produces longer length assemblies



FINAL ASSEMBLIES: SUMMARY TABLE

Sample	Genome Length	N50	L50	GC (%)	# of Contigs
OB0001	5091201	368767	5	52.06	39
OB0002	4870895	411917	4	52.03	54
OB0003	4787899	380392	5	52.08	37
OB0004	5011647	204220	7	52.1	46
OB0005	4807094	439006	4	52.11	21
OB0006	4797153	412049	4	52.05	23
OB0007	5132398	395938	5	52.06	35
OB0008	4468280	548552	4	52.18	14
OB0009	5040559	440299	5	52.04	16
OB0010	4785987	221278	7	52.08	47
OB0011	4763307	397502	5	52.08	36
OB0012	4814729	156944	9	52.09	56
OB0013	5052251	426056	4	52.04	35
OB0014	5022308	539810	4	52.03	18
OB0015	5020248	270079	5	52.07	38
OB0016	5073863	412034	4	52.09	33
OB0017	4818198	298803	5	52.11	39
OB0018	4826057	286074	5	52.08	46
OB0019	4981537	381484	4	52.03	35
OB0020	4832658	298639	5	52.1	41
OB0021	4869918	270023	6	52.12	41
OB0022	4774807	235849	5	52.08	43
OB0023	4894310	276120	6	52.03	43
OB0024	4894310	276120	6	52.03	43

ASSEMBLY SELECTION

- Reference-guided assembly might perform more poorly than de novo assembly on bacterial genomes that contain mobile elements like plasmid DNA and/or transposons
  - Plasmids are extra-chromosomal segments of DNA
    - A mechanism for antibiotic resistance – so of interest to preserve them
  - Transposons are sequences of DNA that can move within the genome
    - Can be found in plasmid or chromosomal DNA
  - Horizontal gene transfer may also result in the reference-guided assembly performing poorly

CONCLUSION

- The reference-guided assembly performed better than the best *de novo* assembly and the best meta-assembly
  - ANI plots showed the 26 references to be over 99% similar to our 24 samples
  - Since the goal of this exercise is to explore the pathogenicity of the samples, it would be wise to use a *de novo* assembly-based methodology to help identify novel* genes involved in pathogenesis.
- Although the reference-guided assembly performed better, it may have dropped some indels, genes, and transposable elements that didn't align to the reference
  - The assembly may be better, but we are interested in capturing these elements for downstream analysis
  - The best *de novo* meta assembly is the best compromise to maintain specificity and sensitivity for mobile DNA elements

MORE INFORMATION

Our github page:

https://github.com/GenomeAssembly2017

Our Wiki:

http://compgenomics2017.biology.gatech.edu/index.php/Faction_I_Genome_Assembly_Group

Our final assemblies are in:
/data/home/jrowell32/data/final_assemblies

THANK YOU!