COMPARING MICROBIAL COMMUNITY RESULTS FROM DIFFERENT SEQUENCING TECHNOLOGIES

Tyler Bradley*
Jacob R. Price*
Christopher M. Sales*

* Department of Civil, Architectural, and Environmental Engineering, Drexel University

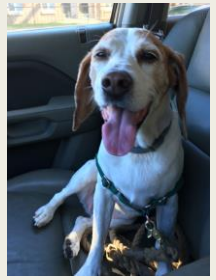
Agenda

- Project Overview
- Sample Collection
- Sequencing Methods and Postprocessing
- Community comparison results

Project Overview

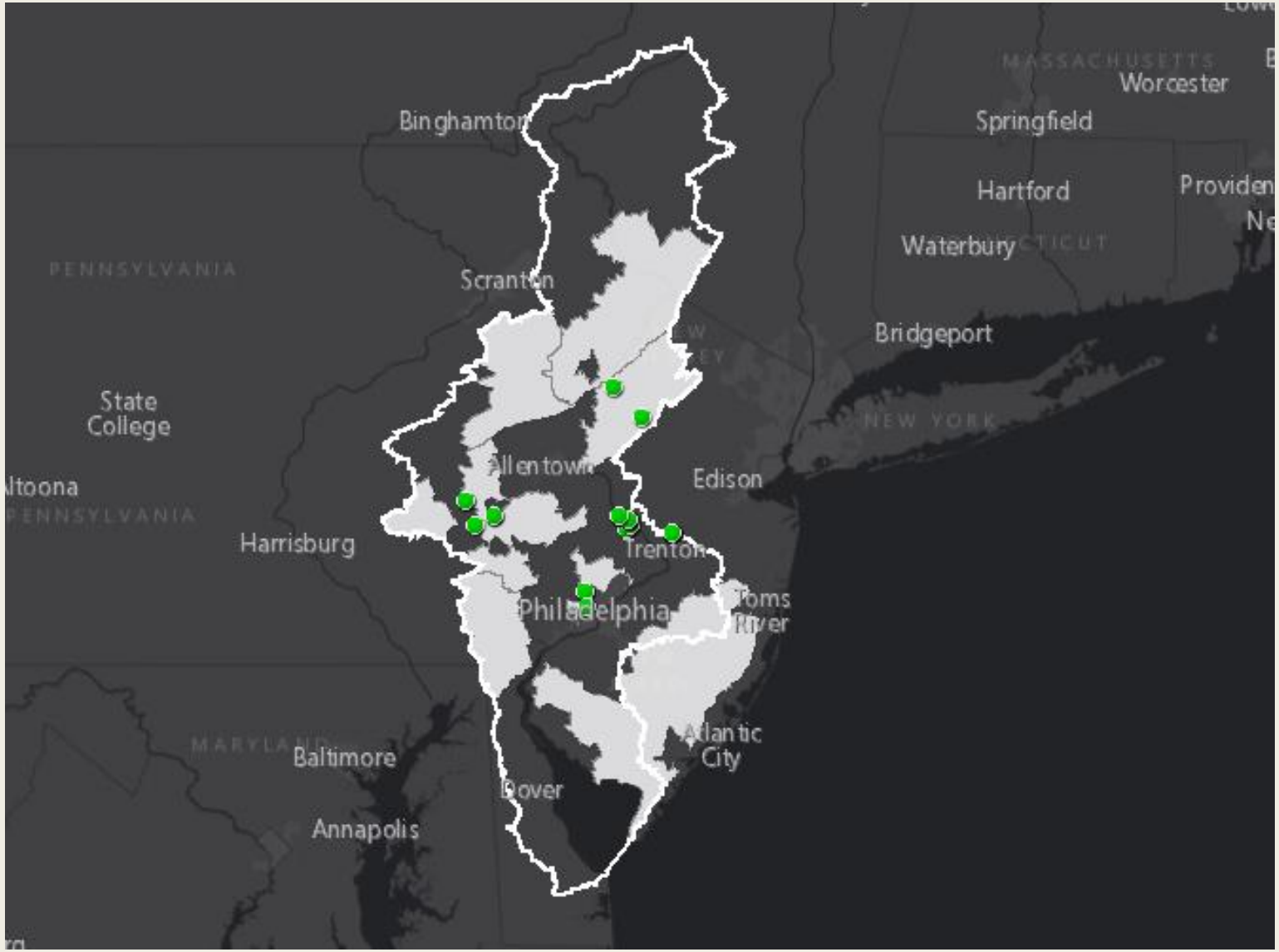
- Microbial Source Tracking (MST) in the Delaware River Watershed
- Objectives:
  1. Generate and analyze high-throughput microbial community (full-length 16S rRNA amplicon) sequencing libraries of different potential fecal sources and water samples collected from a preliminary set of DRWI study sites
  2. Produce high-throughput microbial community (full-length 16S rRNA amplicon) sequencing data of water collected from a preliminary set of DRWI study sites to determine how they correlate with other information being collected at those sites.
  3. Develop and test a preliminary suite of genetic biomarkers based on the sequencing libraries for quantification of microorganisms indicative of specific sources of fecal contamination or presence of particular chemical contaminants.
- Additional Hypothesis: High quality, full length sequencing (16S rRNA gene, ~1.5kbp) via PacBio has improved ability to identify bacteria more precisely

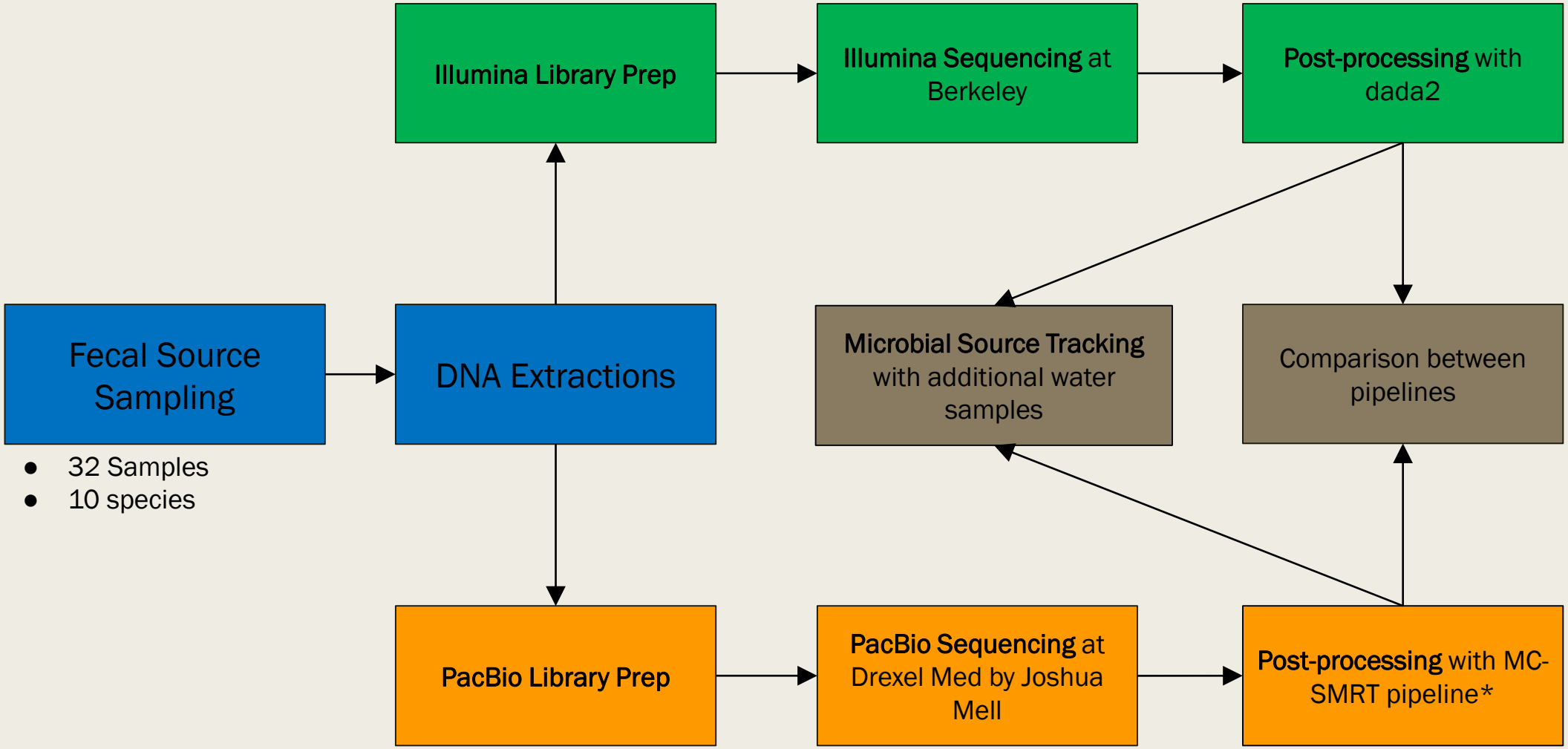
Fecal Source Sample Collection

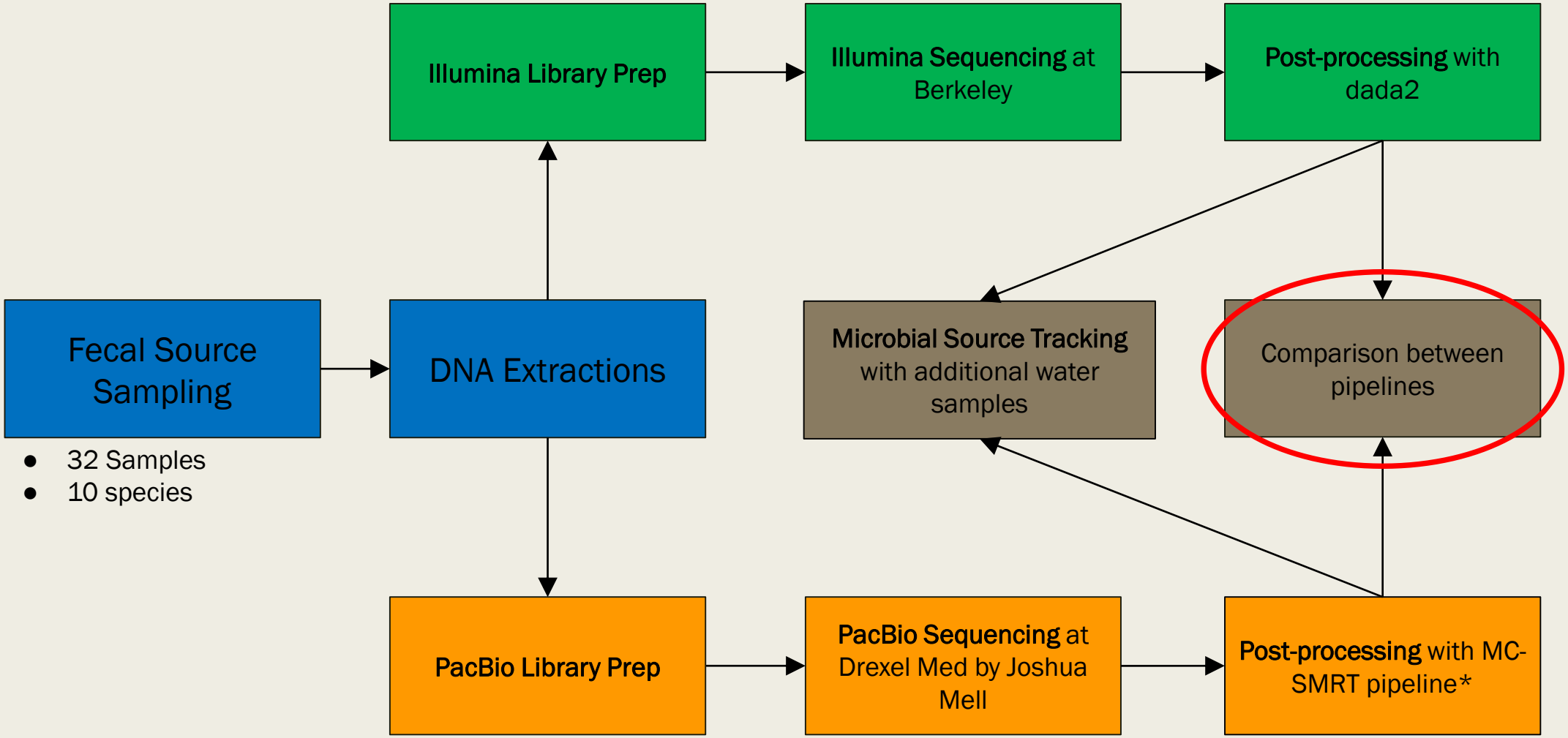










Illumina Library Prep
Illumina Sequencing at Berkeley
Post-processing with dada2
Fecal Source Sampling
DNA Extractions
Microbial Source Tracking with additional water samples
Comparison between pipelines
32 Samples
10 species
PacBio Library Prep
PacBio Sequencing at Drexel Med by Joshua Mell
Post-processing with MC-SMRT pipeline*


Illumina Library Prep
Illumina Sequencing at Berkeley
Post-processing with dada2
Fecal Source Sampling
DNA Extractions
Microbial Source Tracking with additional water samples
Comparison between pipelines
32 Samples
10 species
PacBio Library Prep
PacBio Sequencing at Drexel Med by Joshua Mell
Post-processing with MC-SMRT pipeline*

Comparing Sequencing Technologies

Platform	Illumina MiSeq	PacBio Sequel
Number of Reads	20-180M/lane	500k/SMRT Cell
Yield	Up to 15 to 45 Gb/lane	Up to 1.25 Gb/SMRT cell
Read Length	50 to 150 bp	1,000 to 20,000 bp (avg. 10k-15kbp)
16s analysis cost (this project)	Cost for 96 samples -$3,500 (1 MiSeq lane)	Cost for 32 samples - $12,000 (8 SMRT Cells)

Comparing Sequencing Technologies

Illumina MiSeq

- Targeted specific hypervariable regions of 16S rRNA gene
- Attaches sequences to plate and amplify it to create clusters, clusters are read to identify sequence
- Post-processing: dada2 pipeline
  - Filter for length and quality
  - Dereplication
  - Cluster into ASVs
  - Assign taxonomy via naïve-bayes classifier

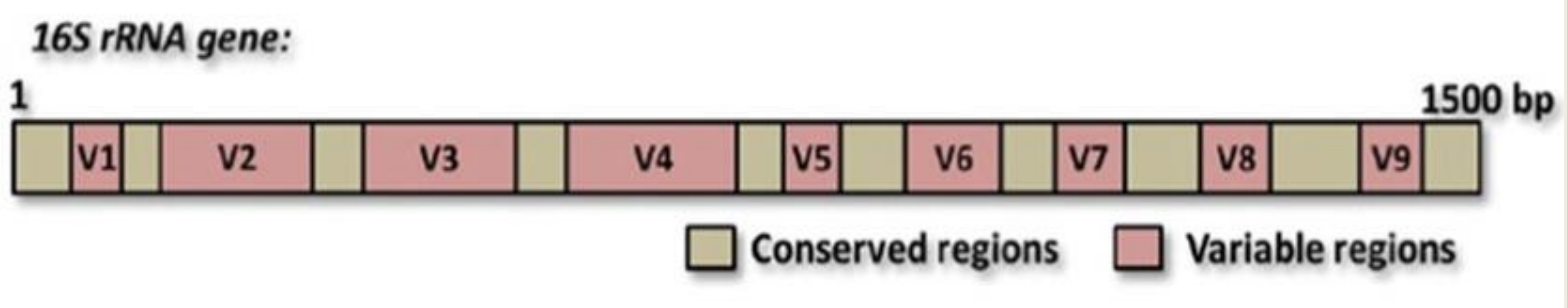
PacBio Sequel

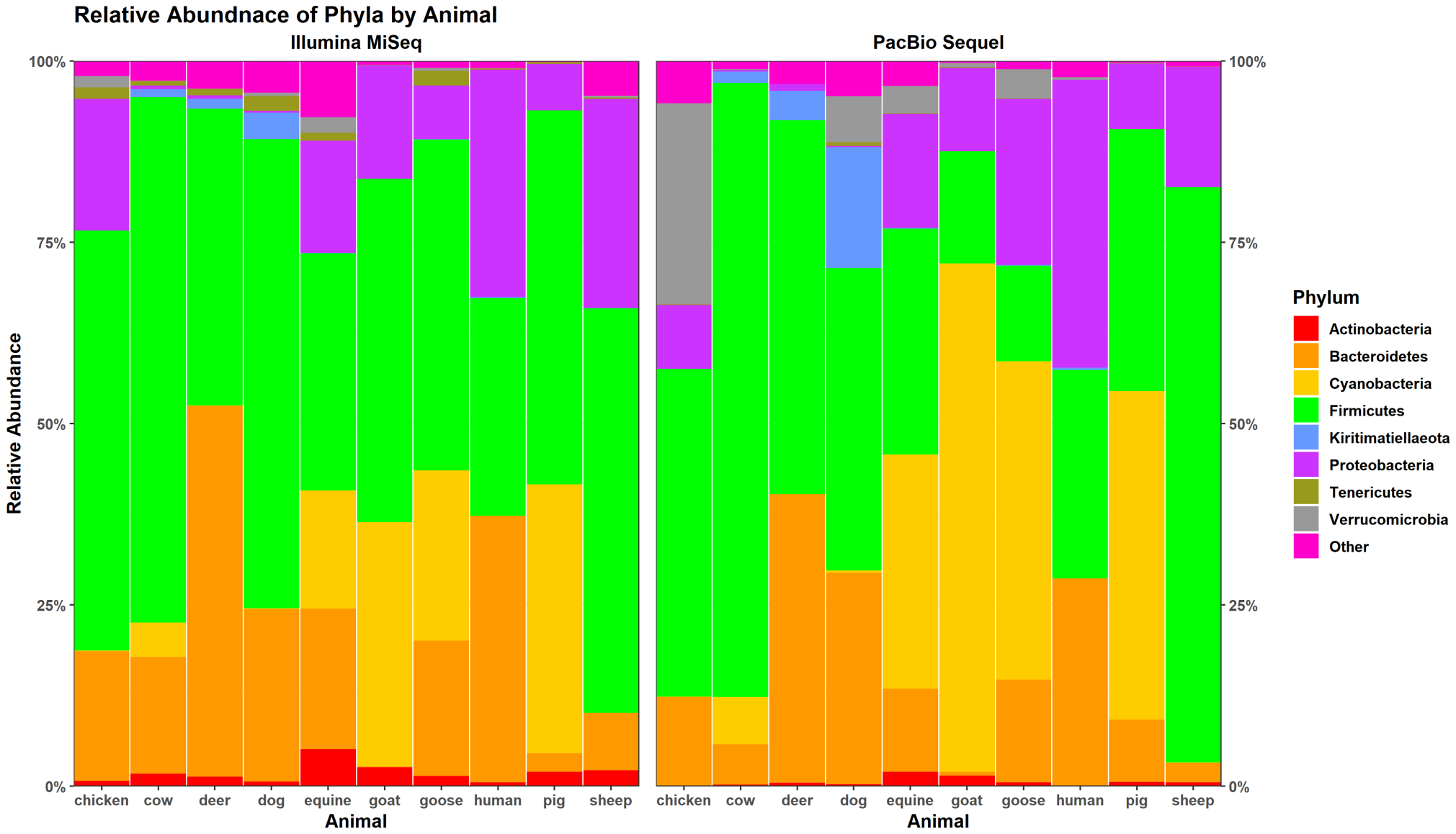
- Targeted full length of 16S rRNA gene
- Single sequence is cycled through single well on plate numerous times to identify sequence
- Post-processing: MC-SMRT pipeline (with slight modification)
  - Demultiplex
  - Filter reads for length and quality
  - Cluster into ASVs
  - Assign taxonomy via naïve-bayes classifier

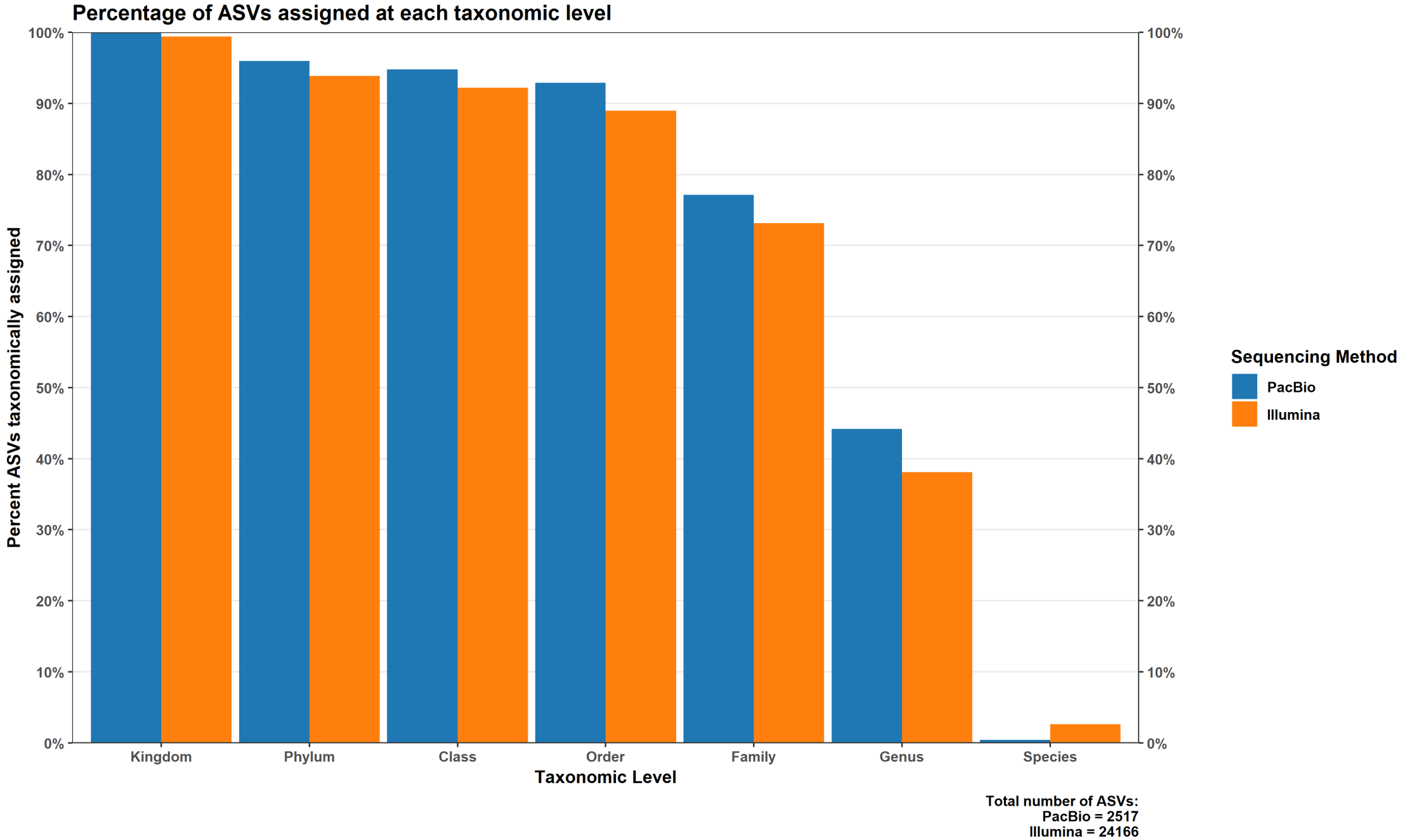
dada2: http://benjjneb.github.io/dada2/index.html
MC-SMRT article: https://doi.org/10.1186/s40168-018-0569-2
MC-SMRT: https://github.com/jpearl01/mcsmrt

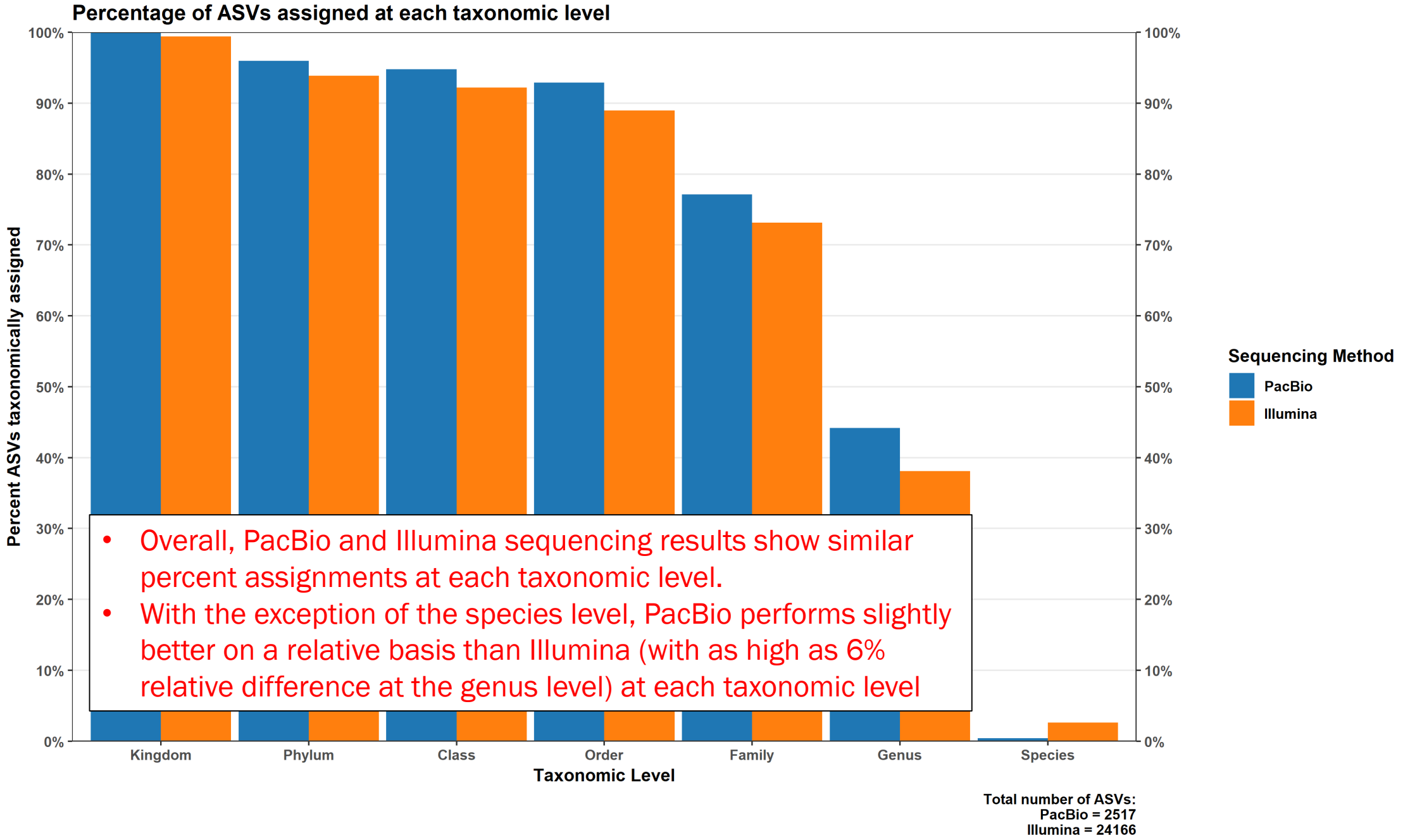
What is 16S?

- Ribosomal RNA (rRNA) gene that is shared by bacteria and archaea
- Ideal candidates for comparing community composition because they are universally distributed, functionally constant, highly conserved, and of adequate length to provide a deep view of evolutionary relationships
- 9 hypervariable regions that allow distinction between different organisms




Relative Abundnace of Phyla by Animal
Illumina MiSeq
PacBio Sequel
Relative Abundance
100%
75%
50%
25%
0%
chicken
cow
deer
dog
equine
goat
goose
human
pig
sheep
Animal
Phylum
Actinobacteria
Bacteroidetes
Cyanobacteria
Firmicutes
Kiritimatiellaeota
Proteobacteria
Tenericutes
Verrucomicrobia
Other


Percentage of ASVs assigned at each taxonomic level
Percent ASVs taxonomically assigned
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Kingdom
Phylum
Class
Order
Family
Genus
Species
Taxonomic Level
Sequencing Method
PacBio
Illumina
Total number of ASVs:
PacBio = 2517
Illumina = 24166



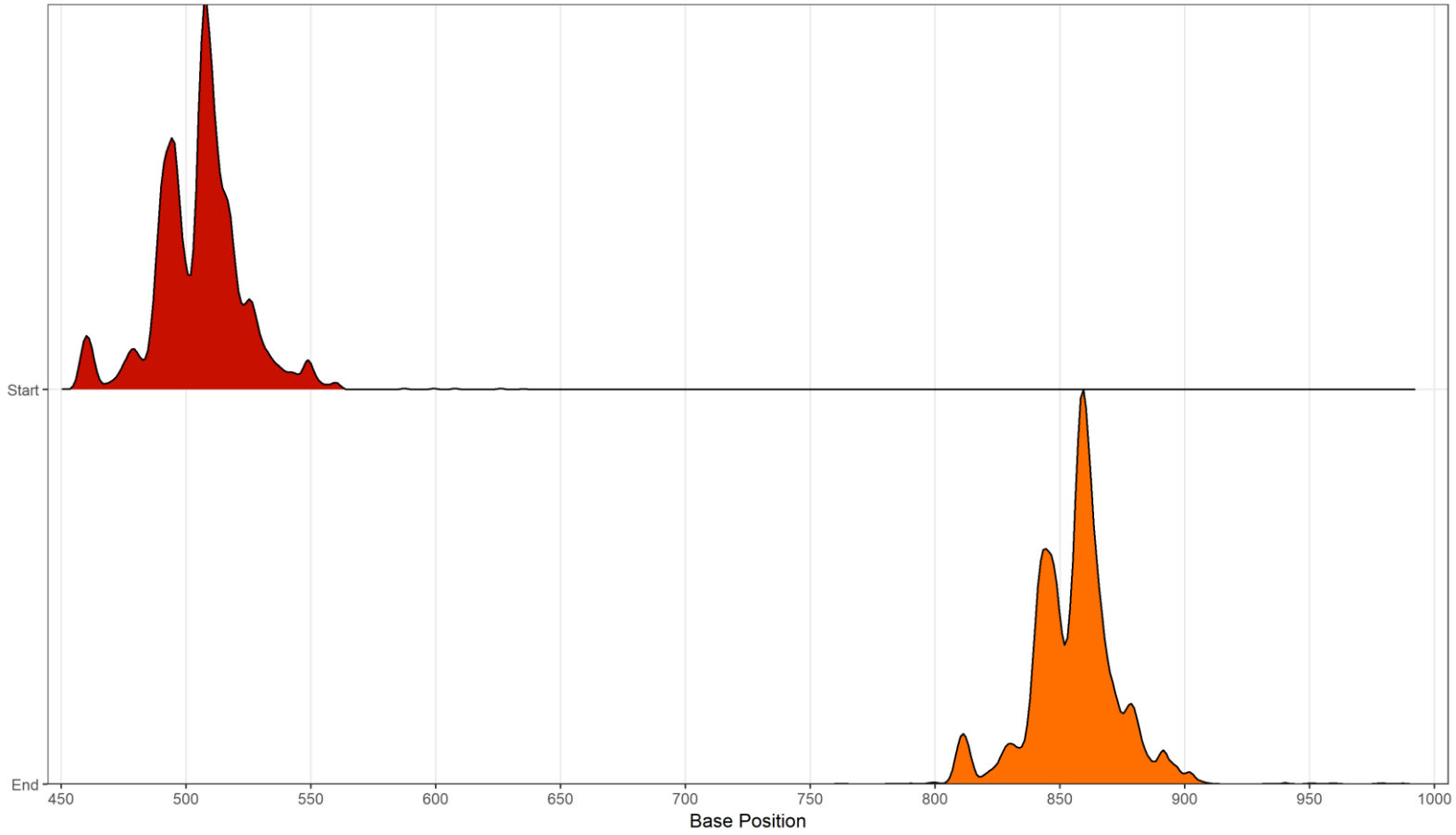
- Overall, PacBio and Illumina sequencing results show similar percent assignments at each taxonomic level.
- With the exception of the species level, PacBio performs slightly better on a relative basis than Illumina (with as high as 6% relative difference at the genus level) at each taxonomic level

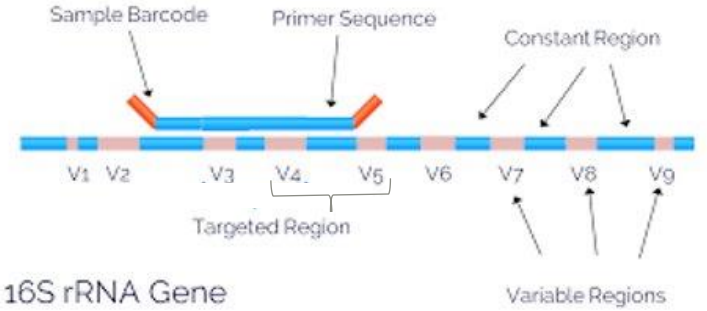
Comparison between community results

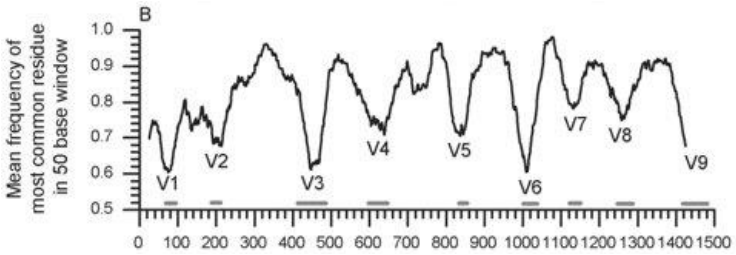
- MiSeq ASV centroid sequences (V4-V5 hypervariable regions of 16S gene) were blasted against Sequel ASV centroid sequence (full-length 16S gene) to compare taxonomic assignment between similar sequences of different lengths
- Best matches were determined by requiring:
  - Alignment length greater than 300 bp
  - Percent identity greater than 97% (less than <11 mismatches)
  - If multiple matches, best taxonomic agreement was selected

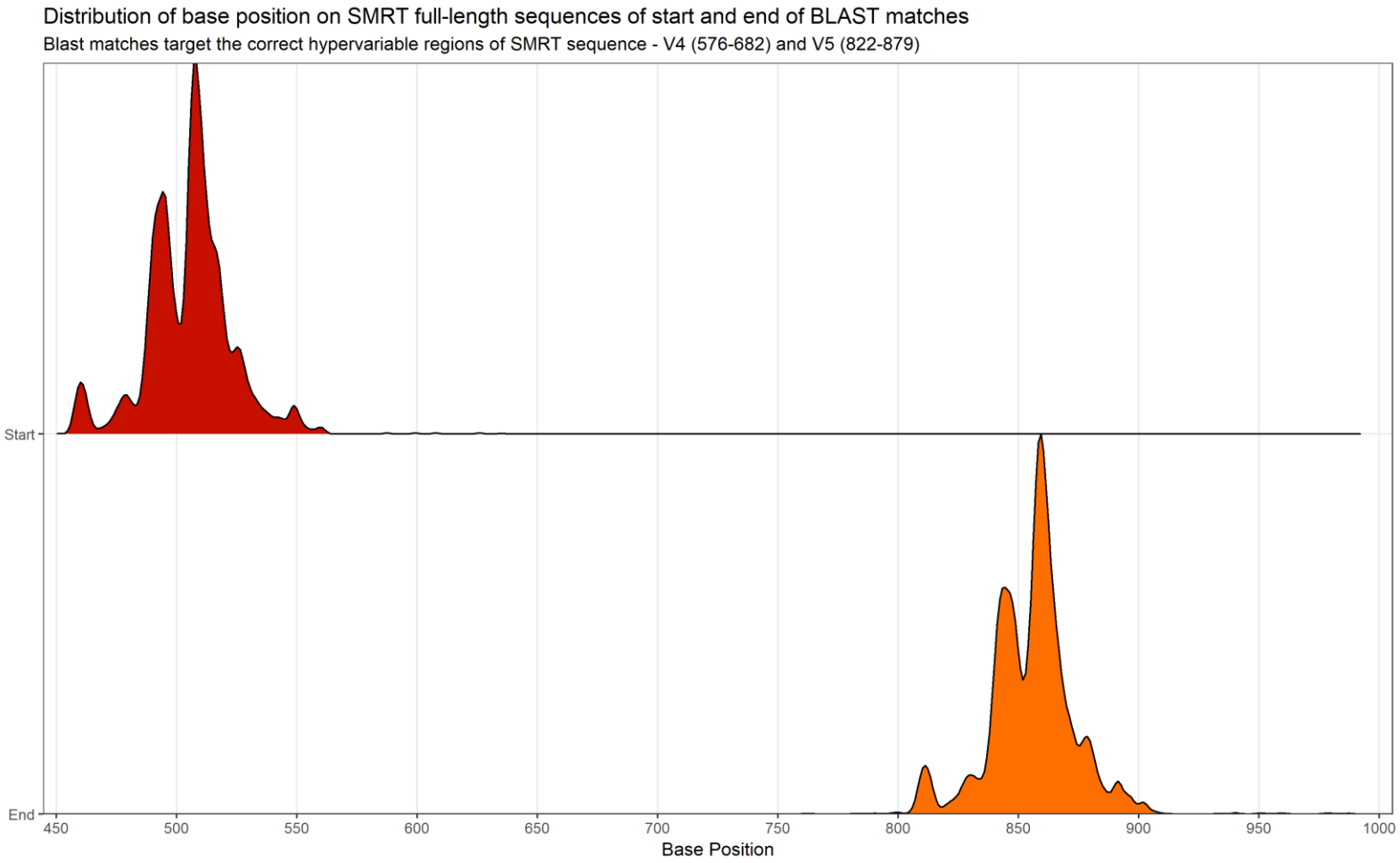
Distribution of base position on SMRT full-length sequences of start and end of BLAST matches

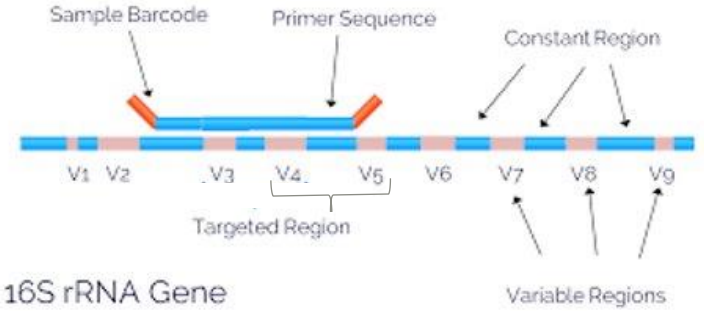
Blast matches target the correct hypervariable regions of SMRT sequence - V4 (576-682) and V5 (822-879)

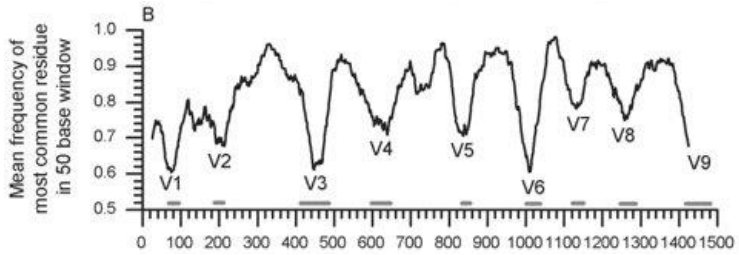




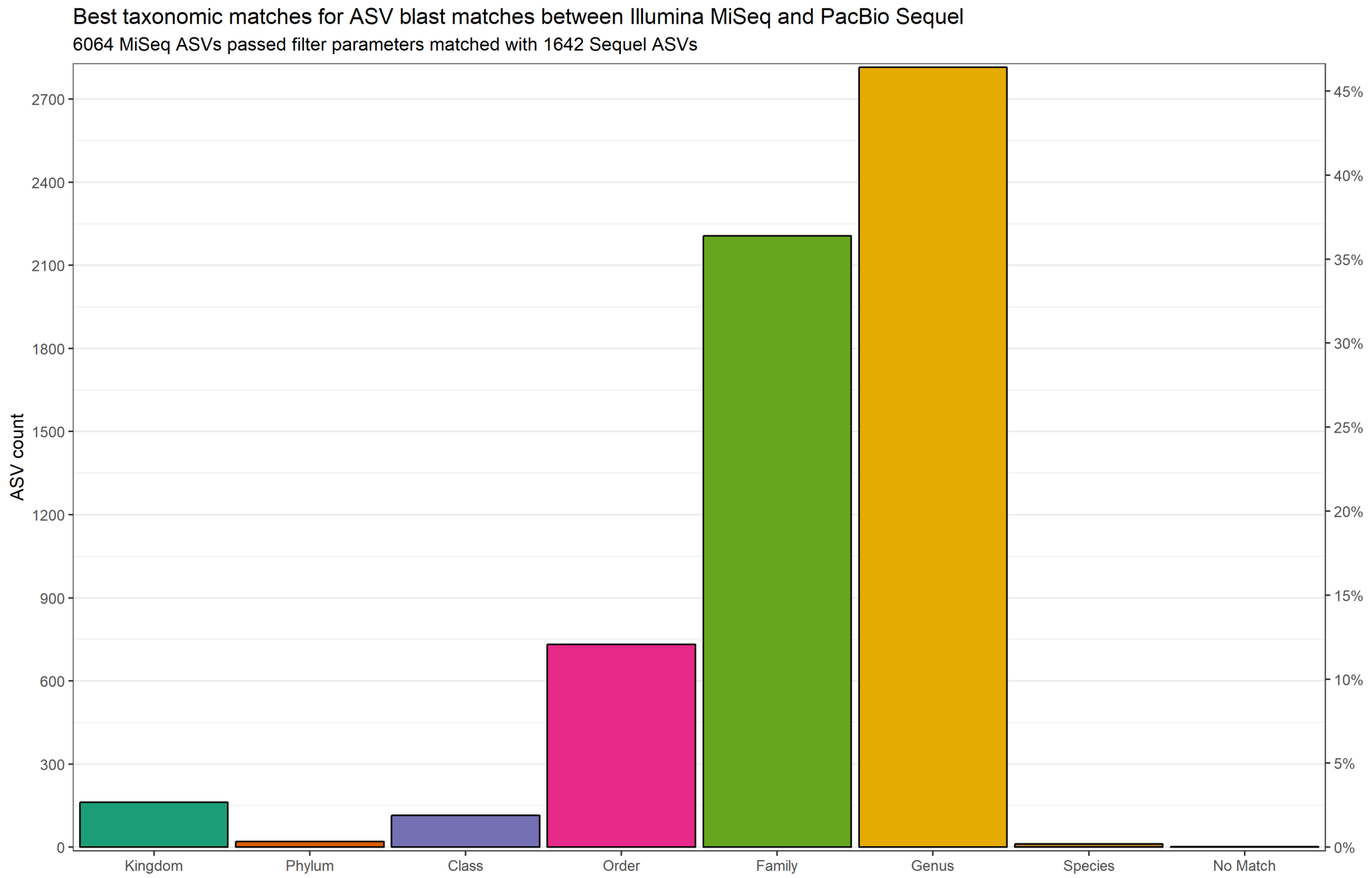


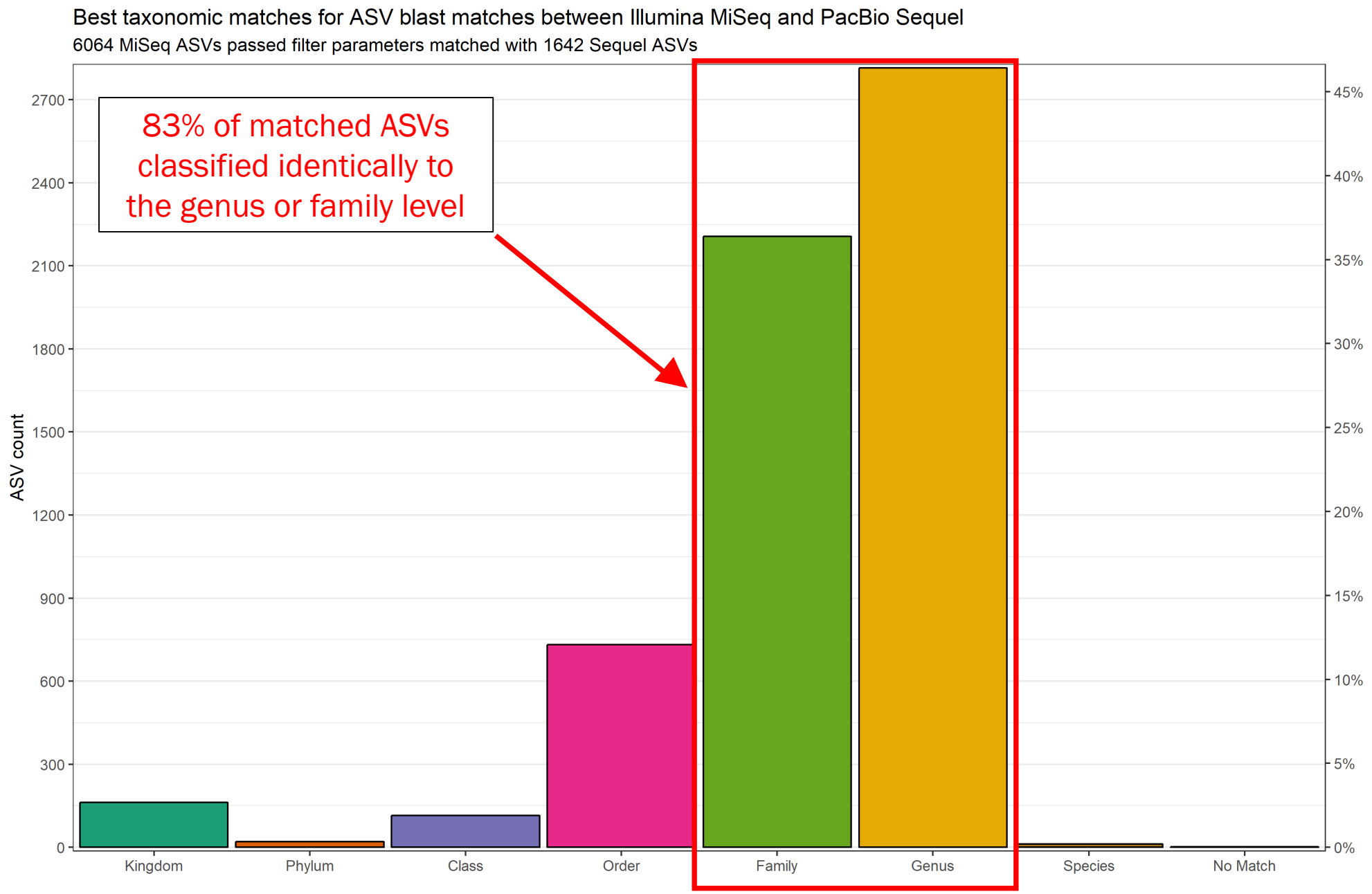






Start and end positions of Illumina blast comparisons match the expected positions of the PacBio full-length 16S rRNA gene


Best taxonomic matches for ASV blast matches between Illumina MiSeq and PacBio Sequel
6064 MiSeq ASVs passed filter parameters matched with 1642 Sequel ASVs
ASV count
2700
2400
2100
1800
1500
1200
900
600
300
0
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Kingdom
Phylum
Class
Order
Family
Genus
Species
No Match


Best taxonomic matches for ASV blast matches between Illumina MiSeq and PacBio Sequel
6064 MiSeq ASVs passed filter parameters matched with 1642 Sequel ASVs
83% of matched ASVs classified identically to the genus or family level
ASV count
2700
2400
2100
1800
1500
1200
900
600
300
0
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Kingdom
Phylum
Class
Order
Family
Genus
Species
No Match

Conclusions from taxonomic assignment comparisons

	Illumina	PacBio
Kingdom	Bacteria	Bacteria
Phylum	Actinobacteria	Actinobacteria
Class	Actinobacteria	Actinobacteria
Order	Corynebacteriales	Corynebacteriales
Family	Mycobacteriaceae	Mycobacteriaceae
Genus	Mycobacterium	Mycobacterium
Species		

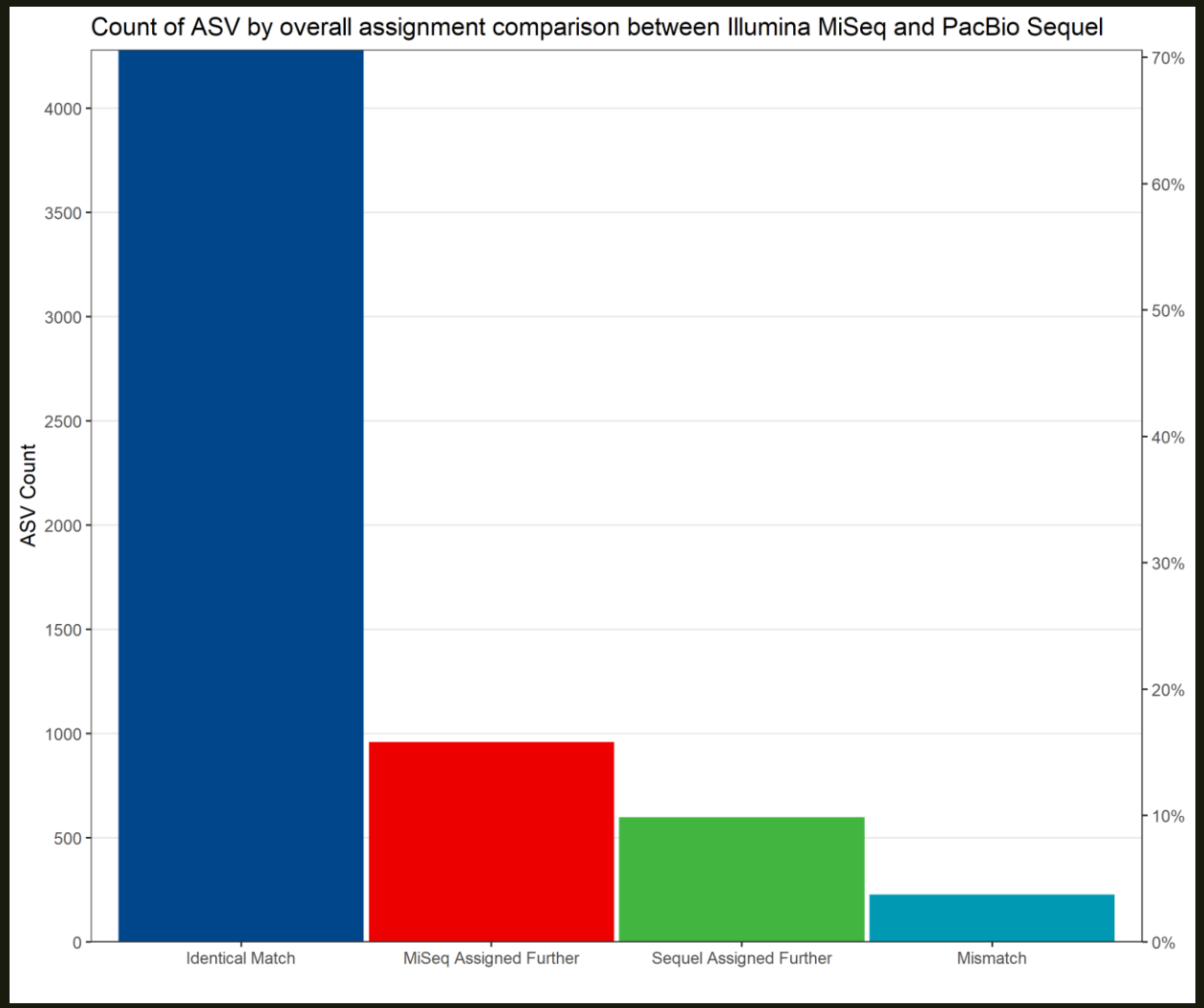
- 46% of matched ASV centroid sequences had identical taxonomic assignment to the genus level

Conclusions from taxonomic assignment comparisons

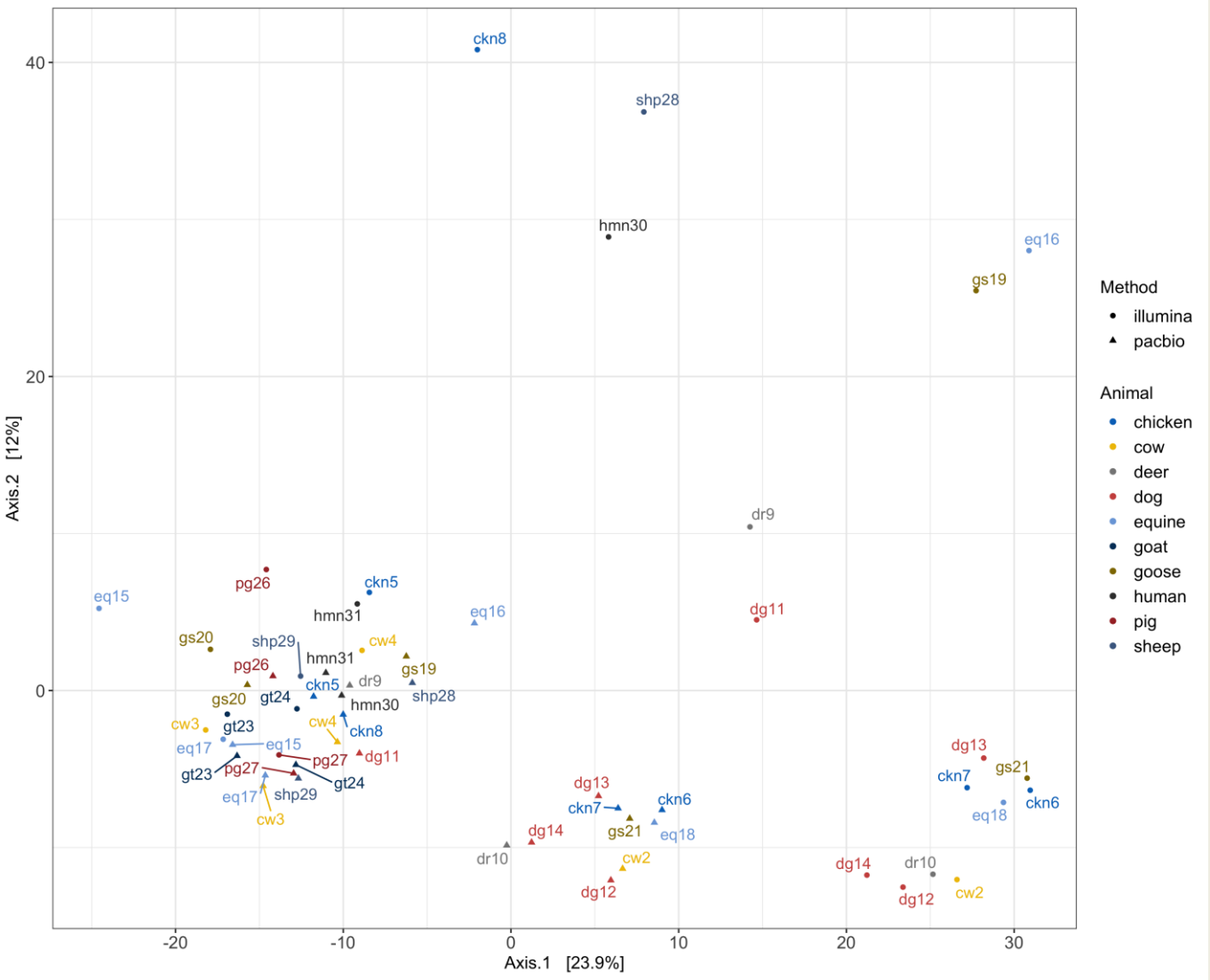
	Illumina	PacBio
Kingdom	Bacteria	Bacteria
Phylum	Proteobacteria	Proteobacteria
Class	Alphaproteobacteria	Alphaproteobacteria
Order	Rhizobiales	Rhizobiales
Family	Xanthobacteraceae	Xanthobacteraceae
Genus	Nitrobacter	Bradyrhizobium
Species	vulgaris	

- Of the remaining matched ASV centroid sequences, 36% had identical taxonomic assignment to the family level
  - *59% were not classified at the genus level in either method*
  - *Only 4.5% were classified differently at the genus level*

Conclusions from taxonomic assignment comparisons



- Overall, 70% of ASVs have identical taxonomic assignment regardless of sequence length when assigned with SILVA v132 with Naïve-Bayes classifier
- Only 3% of matched ASV were assigned for both methods past the com parison's best taxonomic match level



Comparing Sequencing Technologies

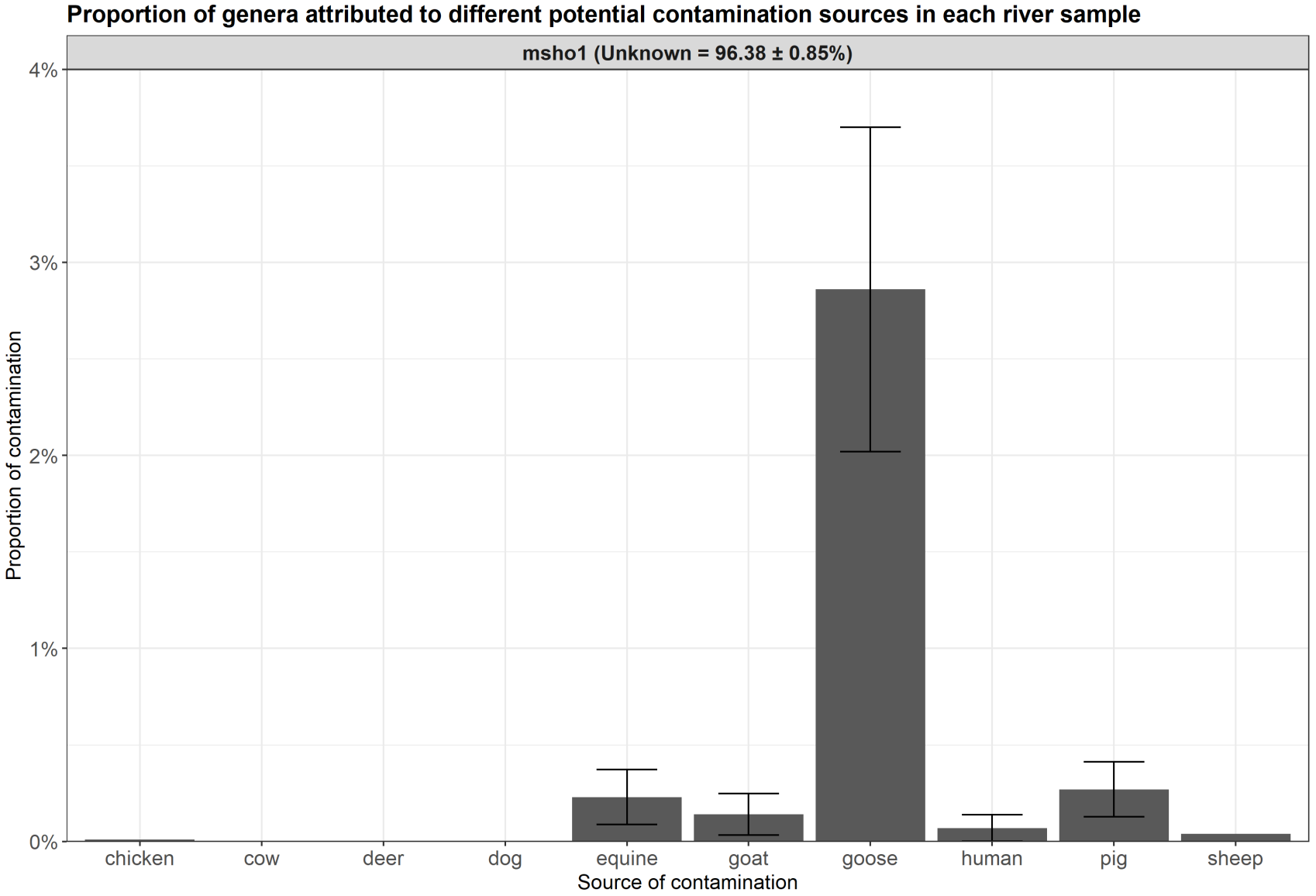
- Now that the taxonomic assignments have been shown to be accurate between the results of the two sequencing technologies, differences between taxa abundances can be more easily assessed
- At the genus level, differential abundance analysis showed that 92.5% (839) of genera shared between the two technologies (888 of 891 total genera) showed no significant difference.
- However, while there is not a large amount of difference between the different genera, there is difference that is best explained by the difference in sequencing method at a sample level.

Conclusions

- Taxonomic assignment via Naïve-Bayes Classifier results in seemingly accurate assignment for both full length and select hypervariable regions of rRNA gene
- Both sequencing methods resulted in roughly similar percentages of OTUs assigned to each of the different taxonomic levels, with PacBio slightly outperforming Illumina
- 92.5% of genera shared between the two sequencing technologies showed no significant differences in abundance between the two technologies
- Overall, the technologies are comparable in their ability to accurately classify the ecological community and in the efficacy of taxonomic assignment. Major differences between the two are seen mostly in cost and overall read abundances

Next Steps

- Identify taxa unique to individual animals within fecal samples
- Determine if these animals are impacting water quality in the waterways downstream of their locations

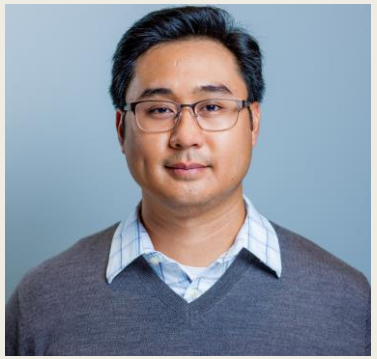

Proportion of genera attributed to different potential contamination sources in each river sample
msho1 (Unknown = 96.38 ± 0.85%)
4%
3%
2%
1%
0%
Proportion of contamination
chicken
cow
deer
dog
equine
goat
goose
human
pig
sheep
Source of contamination

Acknowledgements

The Academy of Natural Sciences of Drexel University

William Penn Foundation

Delaware River Watershed Initiative



Christopher Sales

Drexel University College of Medicine

Genomics Core Facility

Scholarly Research Equipment Award

Drexel University Office of Research

qb3 ucb·ucsc·ucsf

Vincent J. Coates Genomics Sequencing Laboratory

Entomology Group
Microbiology Group

Stroud Water Research Center

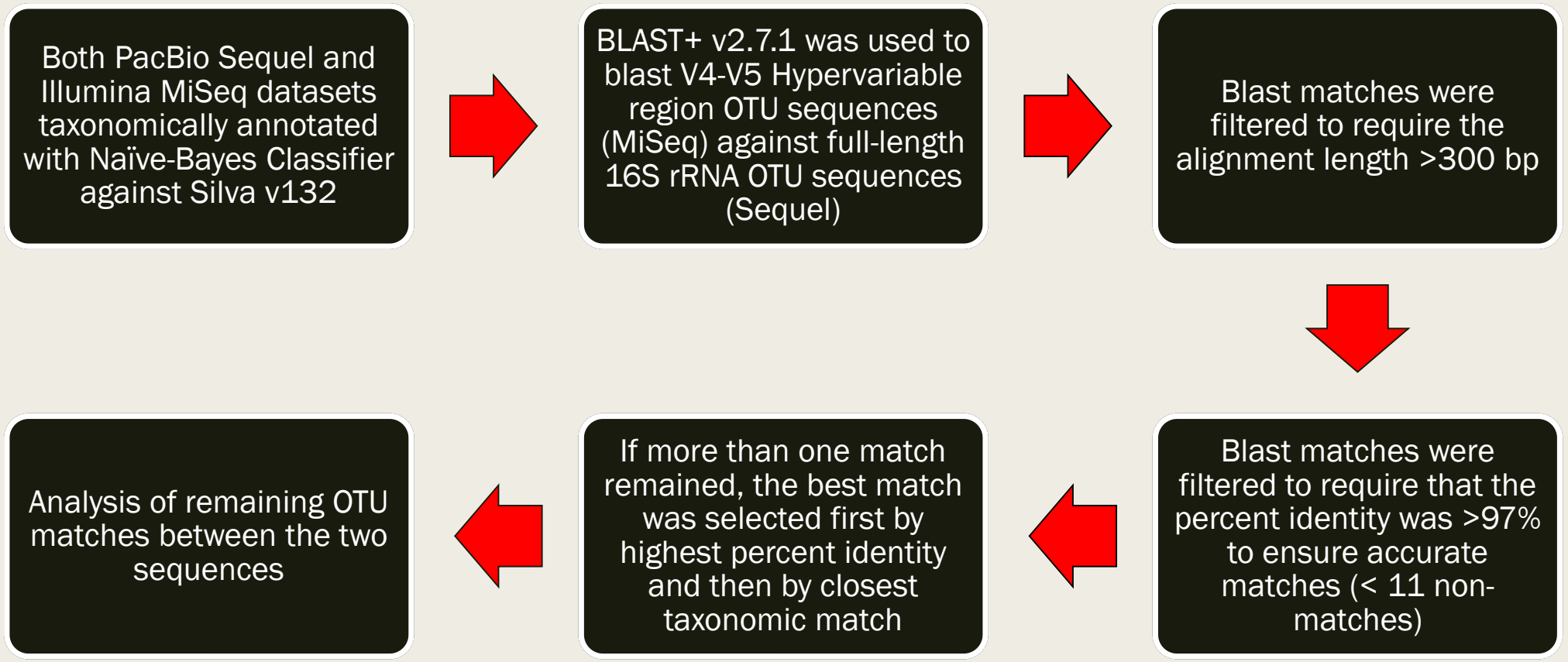
Lin Perez

Jacob Price

Questions?

ADDITIONAL SLIDES

Comparison between community results



MC-SMRT Workflow