Comparison of sequencing approaches applied to complex soil metagenomes to resolve proteins of interest

Jain S[1], Shilova IN[1], Johnson AJ[1], **Heiner C**[2], Hall R[2], Chang C[2], Wong J[2], Loriaux P[1], DeSantis TZ[1]

[1]Second Genome, Inc, 341 Allerton Ave, South San Francisco, CA USA, [2]Pacific Biosciences of California, Inc, Menlo Park, CA USA

FRIDAY - AES-1171

ABSTRACT

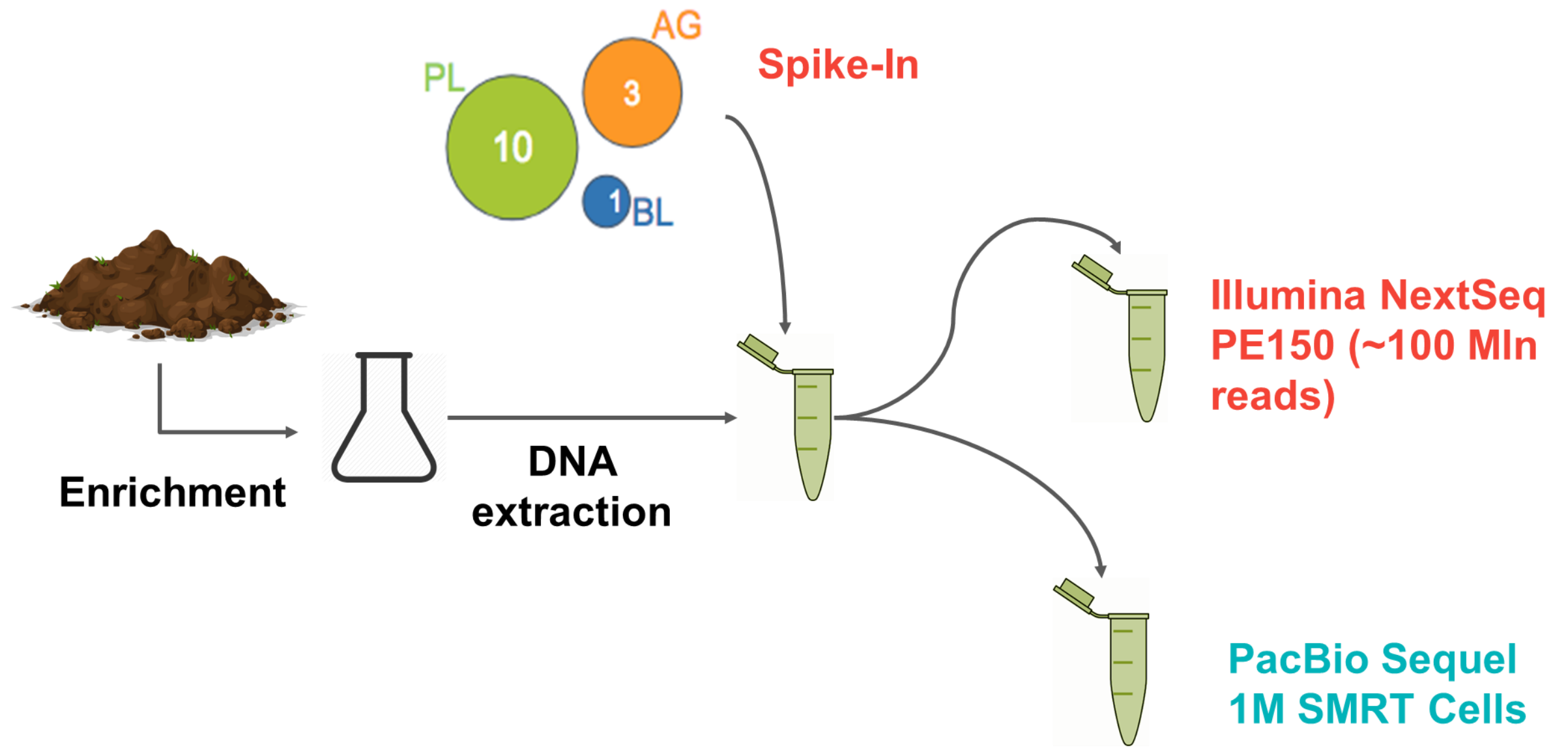
Background: Long-read sequencing presents several potential advantages for providing more complete gene profiling of metagenomic samples. Long reads can capture multiple genes in a single read, and longer reads typically result in assemblies with better contiguity, especially for higher abundance organisms. However, a major challenge with using long reads has been the higher cost per base, which may lead to insufficient coverage of low-abundance species. Additionally, lower single-pass accuracy can make gene discovery for low-abundance organisms difficult.

Methods: To evaluate the pros and cons of long reads for metagenomics, we directly compared PacBio and Illumina sequencing on a soil-derived sample, which included spike-in controls of known concentrations of pure referenced samples. For PacBio sequencing, a 10 kb library was sequenced on the Sequel System with 3.0 chemistry. Highly accurate long reads (HiFi reads) with Q20 and higher were generated for downstream analyses using PacBio Circular Consensus Sequencing (CCS) mode. Results were assessed according to the following criteria: DNA extraction capacity, bioinformatics pipeline status, % of proteins with ambiguous AA's, total unique error-free genes/$1000, total proteins observed in spike-ins/$1000, proteins of interest/$1000, median length of contigs with proteins, and assembly requirements.

Results: Both methods had areas of superior performance. DNA extraction capacity was higher for Illumina, the bioinformatics pipeline is well-tested, and there was a lower proportion of proteins with ambiguous AA's. On the other hand, with PacBio, twice as many unique error-free genes, twice as many total proteins from spike-ins, and ~6 times more proteins of interest were found per $1000 cost. PacBio data produced on average 5 times longer contigs capturing proteins of interest. Additionally, assembly was not required for gene or protein finding, as was the case with Illumina data.

Conclusions: In this comparison of PacBio Sequel System with Illumina NextSeq on a complex microbiome, we conclude that the sequencing system of choice may vary, depending on the goals and resources for the project. PacBio sequencing requires a longer DNA extraction method, and the bioinformatics pipeline may require development. On the other hand, the Sequel System generates hundreds of thousands of long HiFi reads per SMRT Cell, producing more genes, more proteins, and longer contigs, thereby offering more information about the metagenomic samples for a lower cost.

I. APPROACH

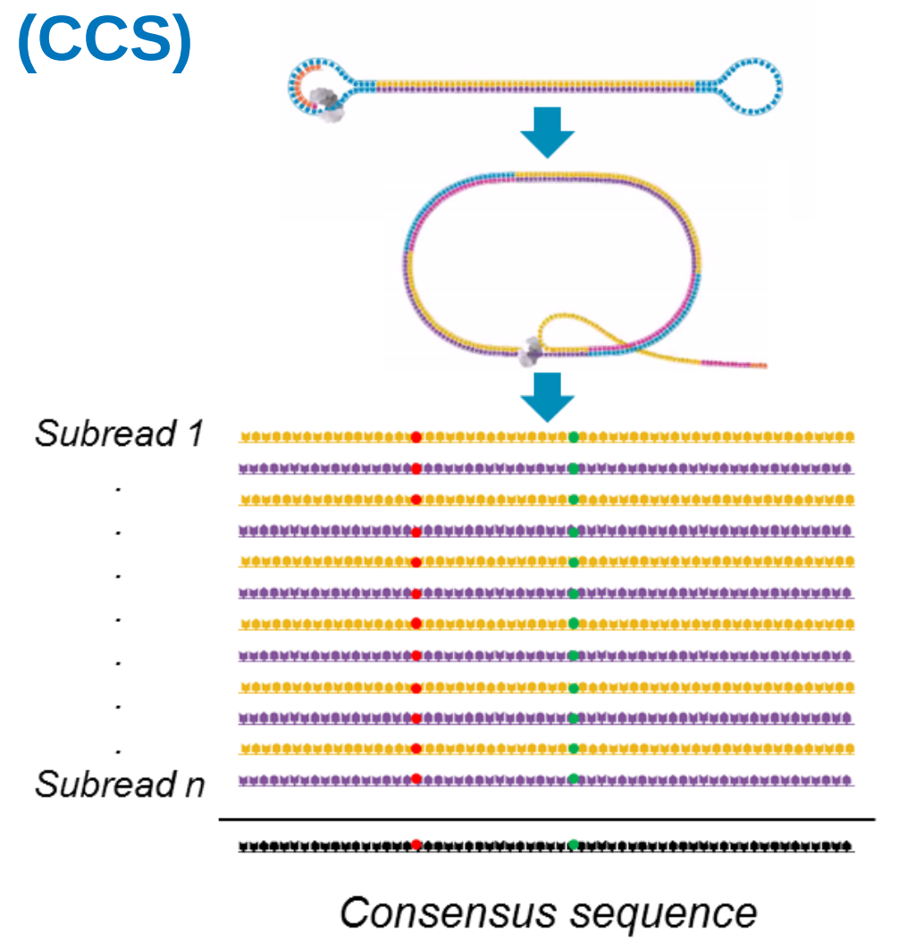

PL
10
AG
3
1
BL
Spike-In
Enrichment
DNA extraction
Illumina NextSeq PE150 (~100 Mln reads)
PacBio Sequel 1M SMRT Cells

II. YIELD OF LONG READS ON THE SEQUEL SYSTEM

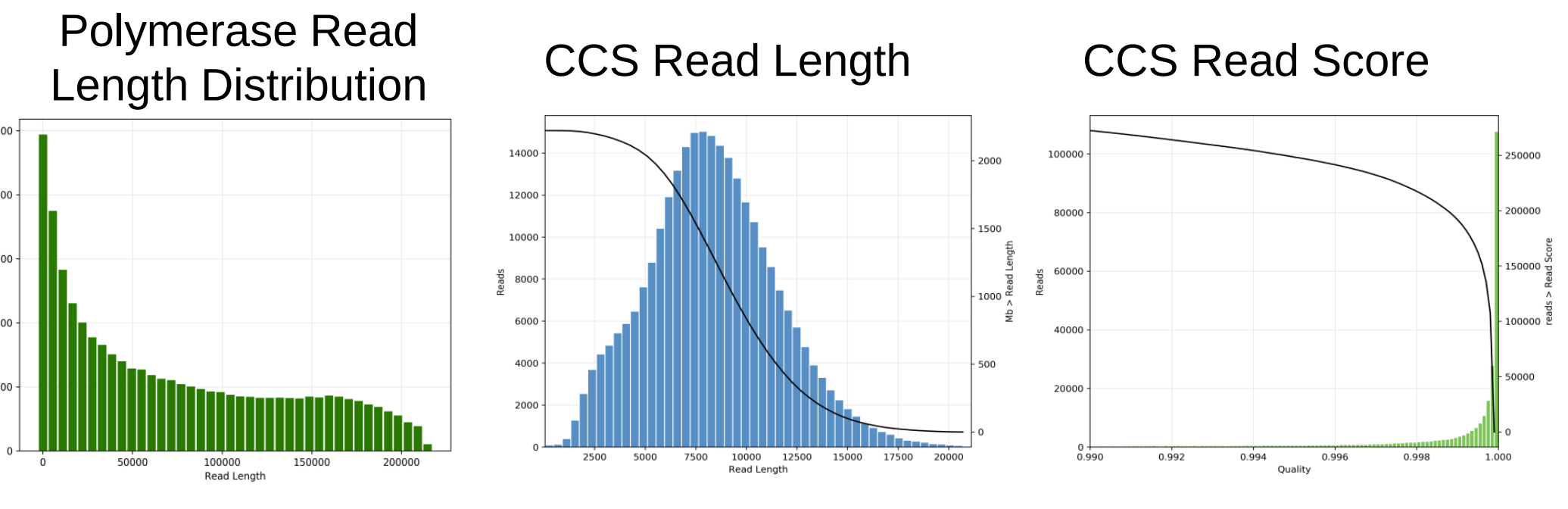
Sample	# of Primary Bases	Mean Polymerase Read Length	# CCS Reads	# CCS Bases	Mean CCS Read Length	Mean CCS Predicted Accuracy
SMRT Cell 1	38 Gb	77,846 bases	273,261	2.3 Gbp	8,345	0.999
SMRT Cell 2	37 Gb	76,318 bases	271,184	2.1 Gbp	7,705	0.999

CCS filtering @ ≥99% predicted accuracy

Circular Consensus Sequencing (CCS)



Read Length and Predicted Accuracy – SMRT Cell 1


Polymerase Read Length Distribution

CCS Read Length

CCS Read Score

III. RECOVERY OF UNIQUE GENES

➢ PacBio recovers over 2x more unique genes compared to Illumina short-read assemblies per $1000

Cost-normalized protein counts (per $1000)

	Illumina		PacBio	
	Prodigal	FragGeneScan	Prodigal	FragGeneScan
# Proteins	29,325	32,009	127,750	117,145
After error penalties	17,573	26,074	unavailable	84,656
Dereplicated at 100%	24,195	21,143	80,705	49,658
After error penalties	14,499	17,222	unavailable	35,886

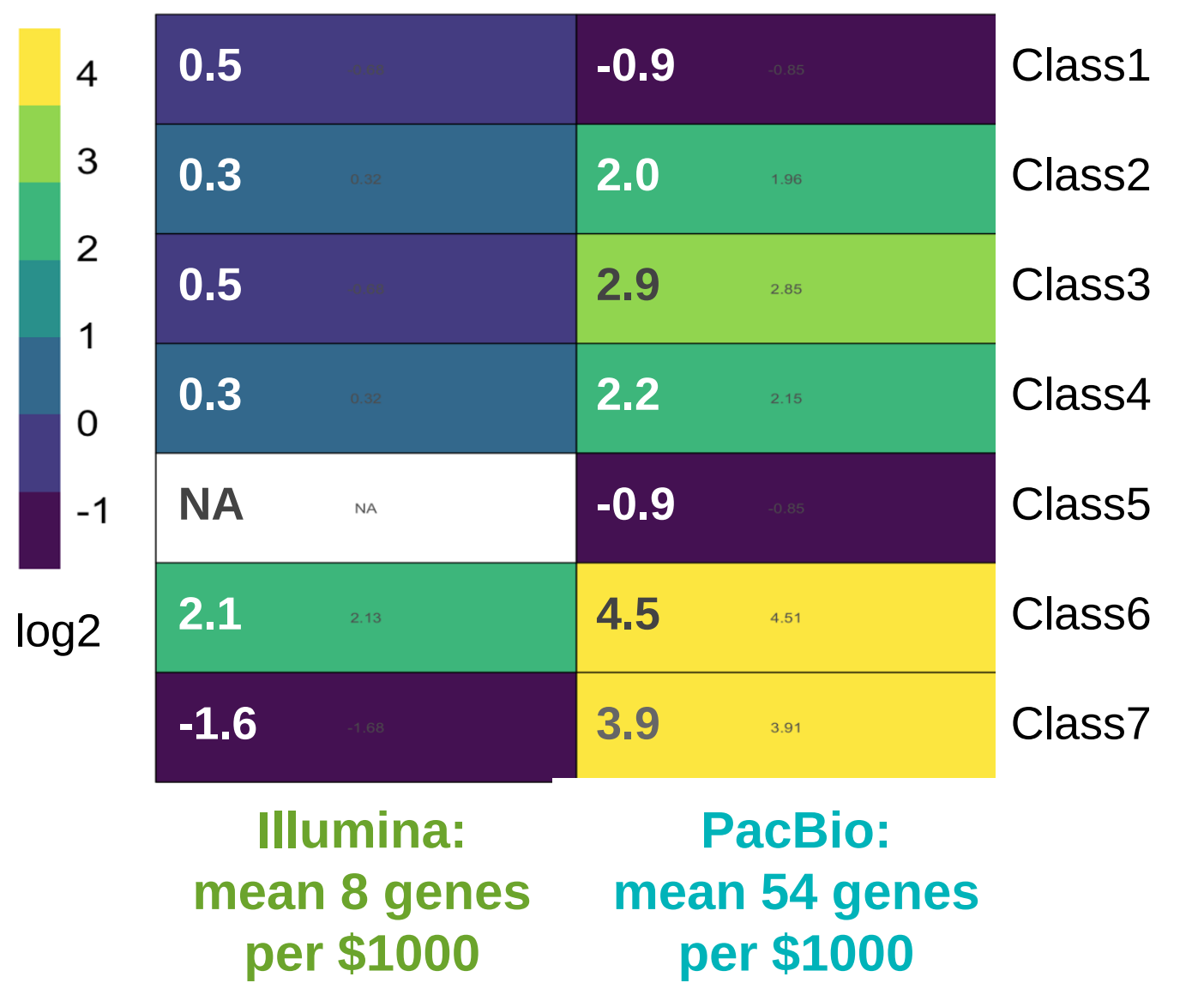
Cost-normalized protein counts per spike-in genome (per $1000)

ID 100 + COV 100		Illumina		PacBio	
Organism	Source SeqTech	Prodigal	FragGeneScan	Prodigal	FragGeneScan
AG	Illumina (ALLPATHS-LG)	1,321	1,182	2,180	2,038
BL	Sanger (MIRA)	1,191	986	2,028	1,744
PL	PacBio (HGAP)	1,197	990	2,097	1,814

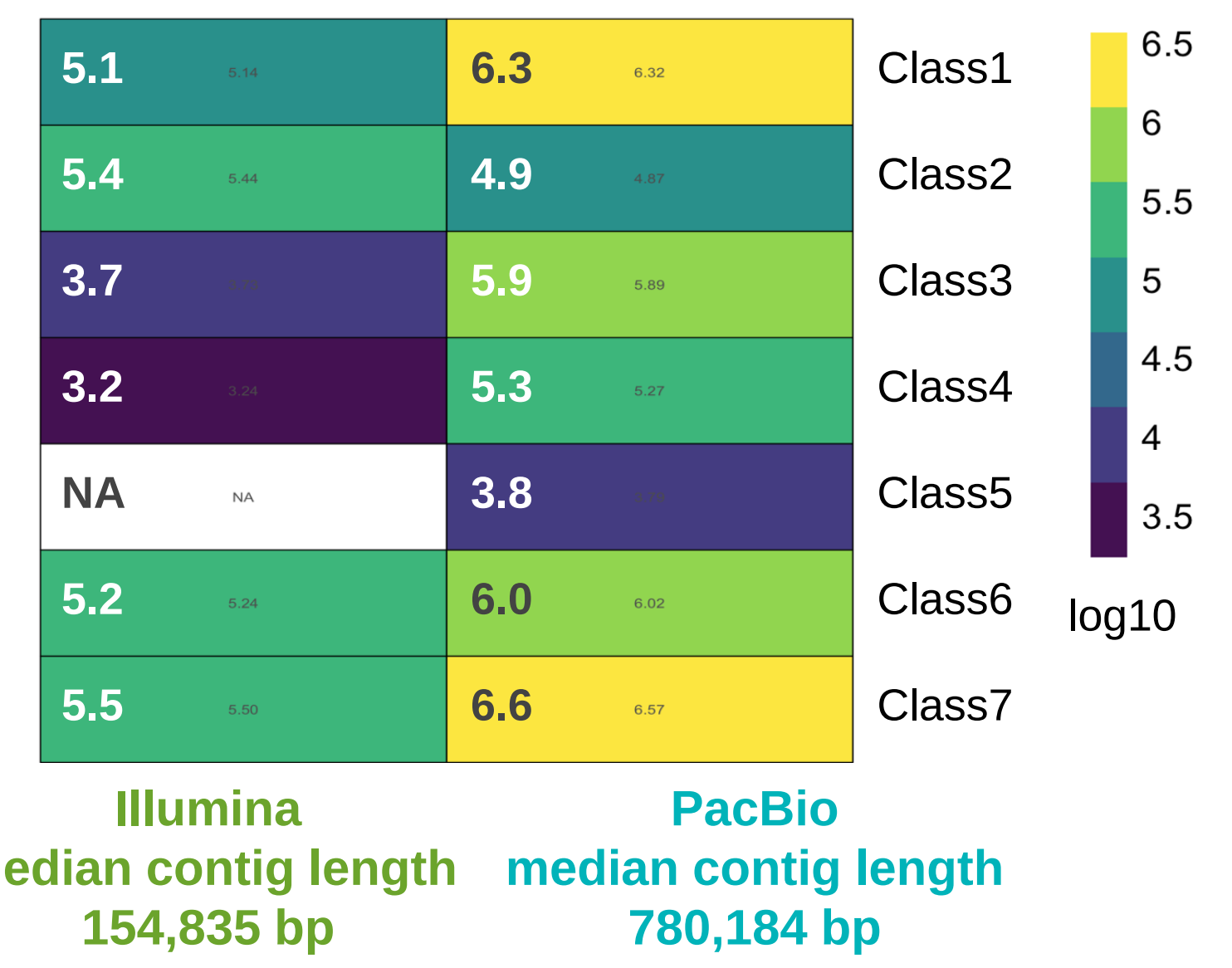
- Normalized to costs, PacBio with Prodigal recalls the highest number of correct genes in spike-in genomes. However, FragGeneScan PacBio predicted proteins were at the ~expected lengths for complete genes, longer than Prodigal PacBio genes.

IV. YIELD AND CONTIG LENGTH FOR GENES OF INTEREST

➢ Grouped by protein class, PacBio assembles higher number of potential genes of interest, on longer contigs



Occurrences



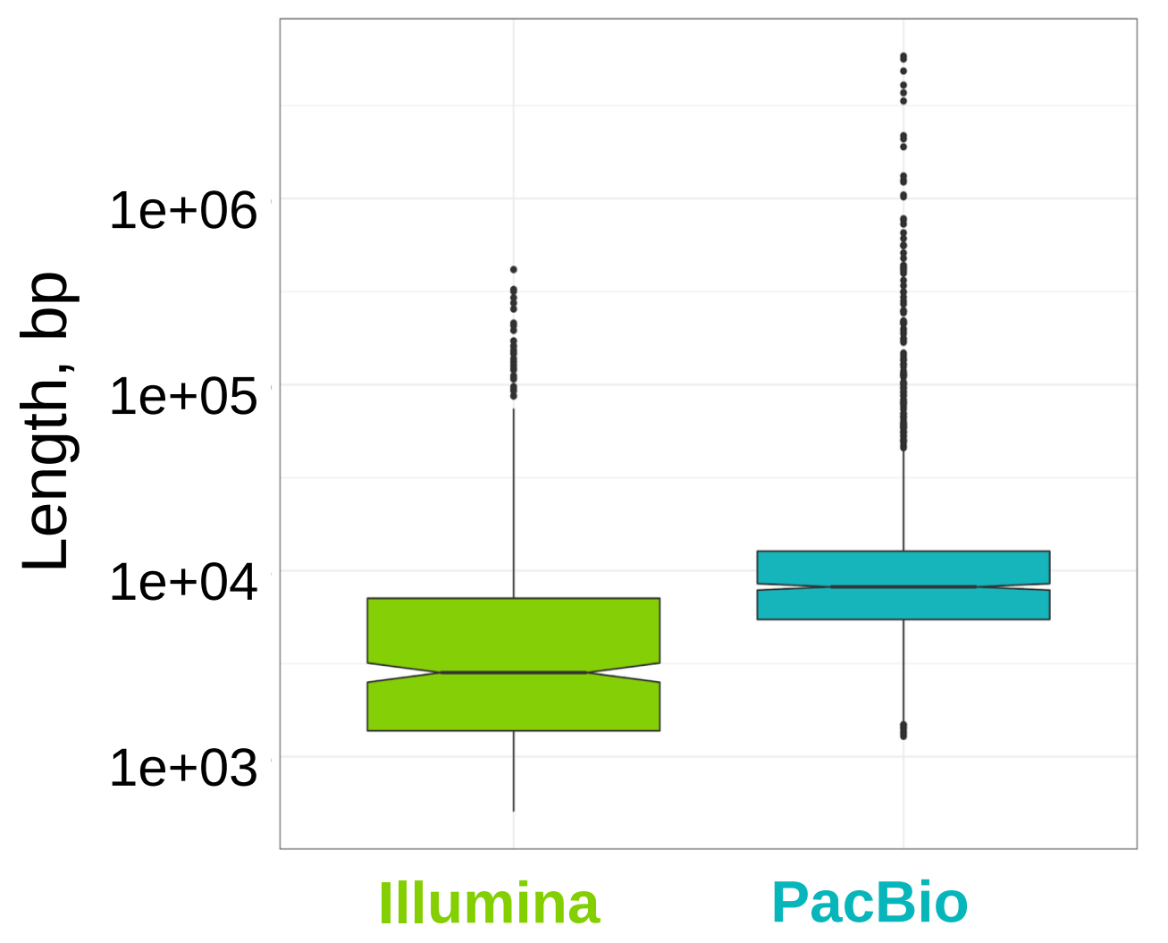
Length of contigs with genes of interest

Method: blastp against proteins of interest
Sequence data:

- PacBio: Assemblies (Canu + Bandage) + unmapped CCS reads over 99.9% accuracy from Sequel System data
- Illumina: Assemblies only

V. CONTIG LENGTH COMPARISON

➢ PacBio assemblies produce longer contigs, including multi-megabase contigs and complete genomes



PacBio Assembly Stats	
Number of contigs	1,081
Total size of contigs	100,739,059 bp
Longest contig	5,861,669 bp
Shortest contig	1,221 bp
Mean contig size	93,191 bp
Median contig size	24,014 bp
N50 contig length	507,735 bp

- PacBio assembly produced 3 complete genomes, 2 in single contigs and 1 in two contigs.

VI. SUMMARY OF RESULTS

Evaluation Criteria	PacBio	Illumina	Preference
DNA Extraction Capacity (1 FTE)	8h hands-on and 24h passive time for 8 samples	7 h per 90 samples (using KF)	Illumina
Bioinformatics Pipelines	Needs Development	Well Tested	Illumina
Portion of all observed proteins with ambiguous AA	28%	18%	Illumina
Total unique error-free genes observed	35,886 per $1000	17,222 per $1000	PacBio
Total proteins observed from spike-ins	6,305 per $1000	3,709 per $1000	PacBio
Spike-In proteins of interest recovered	7/13	7/13	Wash
Proteins of Interest per $1000	54.5	8.4	PacBio
Length of Contigs w/ Proteins of interests (median)	780,184	154,836	PacBio
Assembly Required	No	Yes	PacBio

➢ CONCLUSION: Short and long read technologies complement each other

Summary	PacBio	Illumina
Similar	• Resolve relatively abundant (at least 1%) genomes from complex metagenomes • Assemblies with multiple genes of interest per contig	
Advantages	• More unique error-free genes and proteins of interest / $1000 • Proteins of interest located in longer contigs • Whole genome assembly • Longer NG50 • One CCS read can contain full gene(s) - no assembly necessary	• Streamlined defined sample prep • Bioinformatics pipeline is well-tested • Lower error rate
Disadvantages	• Development required for DNA extraction and bioinformatics pipelines	• Lower number of genes per contig; shorter contigs

ACKNOWLEDGEMENTS

We acknowledge T. Ogawa for providing the soil sample, MAD Lab at SG for Illumina sequencing.

Contact Information:
cheiner@pacificbiosciences.com
amyjo@secondgenome.com
irina@secondgenome.com