# MINING COMPLEX METAGENOMES FOR PROTEIN DISCOVERY WITH LONG-READ SEQUENCING

**SEQUENCE WITH CONFIDENCE**

PACBIO®

The bacteria living on and within us can impact health, disease, and even our behavior, but there is still much to learn about the breadth of their effects. The torrent of new discoveries unleashed by high-throughput sequencing has captured the imagination of scientists and the public alike.

Scientists at Second Genome are hoping to apply these insights to improve human health, leveraging their bioinformatics expertise to mine bacterial communities for potential therapeutics. Recently they teamed up with scientists at PacBio to explore how long-read sequencing might supplement their short-read-based pipeline for gene discovery, using an environmental sample as a test case. They were especially interested in identifying unique, complete, and error-free gene clusters in metagenomic assemblies.

Second Genome is a biopharmaceutical company with a mission to redefine disease in the context of microbiome medicine and create therapeutics that can address unmet medical needs.

SECOND GENOME

## A Collaborative Approach to Gene Discovery

The team began by spiking the sample with internal controls and generating a 10 kb insert library. The library was sequenced on two Single Molecule, Real-Time (SMRT®) Cells with 20-hour movies on the Sequel® System. The resulting data was analyzed in two ways. First, the circular consensus sequencing (CCS) algorithm was applied to generate ~270,000 HiFi reads per SMRT Cell 1M with ≥99% accuracy. Since these reads are on average 10-times longer than the typical bacterial gene, full-length genes can be discovered even without assembly. Bioinformaticians at Second Genome then applied two different gene prediction programs to evaluate the usability of HiFi reads for gene discovery. Next, the raw data was assembled with Canu, and the discovered genes were mapped back onto the contigs.

## Results Reveal Novel Insights

Two SMRT Cells of data revealed an impressive number of genes and a highly contiguous assembly with a mean contig size of 93 kb (Table 1). In addition, among the contigs were two closed bacterial genomes.

**Table 1. Metagenome assembly statistics**

| Metric | Base Pairs |
|---|---|
| **Number of contigs** | 1,081 |
| **Total size of contigs** | 100,739,059 |
| **Longest contig** | 5,861,669 |
| **Mean contig size** | 93,191 |
| **Median contig size** | 24,041 |
| **N50 contig size** | 507,735 |

Comparing the performance of the two gene discovery algorithms, scientists at Second Genome found Prodigal predicted a large number of genes that were concordant with expected genes regardless of the sequencing technology, assembler, or annotation tool used in the source genome (Table 2). However, more analysis is needed to make a definitive comparison of the two protein prediction tools.

**Table 2. Comparison of gene discovery programs for all identified genes**

| Gene Discovery | FragGeneScan | Prodigal |
|---|---|---|
| **Number of genes** | 2,691,726 | 3,271,709 |
| **Mean size (aa)** | 317 | 251 |
| **100% dereplicated (aa)** | 695,181 | 773,648 |
| **99% dereplicated (aa)** | 451,483 | 624,207 |

## Long-Read Sequencing – A Cost-Effective Method for Protein Prediction

Second Genome then took the analysis one step further to evaluate whether long-read sequencing was a good value for their business. Rather than resort to the commonly used 'cost per base' metric, they developed their own more pertinent way of measuring success: what was the cost per error-free, unique predicted protein?

**Table 3. Recovery of proteins from spike-in genomes, per $1,000 of sequencing data**

| Organism | Reference Information | | | Illumina Short Reads | | PacBio Long Reads | |
|---|---|---|---|---|---|---|---|
| | Assembly Method | Annotation Method | Predicted Genes | Prodigal | FragGeneScan | Prodigal | FragGeneScan |
| **AG** | Illumina (ALLPATHS-LG) | User Submitted | 4,470 | 1,321 | 1,182 | 2,180 | 2,038 |
| **BL** | Sanger (MIRA) | GeneMarkS+ | 4,319 | 1,191 | 986 | 2,028 | 1,744 |
| **PL** | PacBio (HGAP) | GeneMarkS+ | 4,310 | 1,197 | 990 | 2,097 | 1,814 |

By normalizing their data to calculate unique predicted proteins per $1,000, they found that PacBio sequencing was twice as cost-effective as short-read technology at discovering complete genes from the same DNA sample (Table 3). Whereas short-read technology predicted ~17,000 full-length proteins per $1,000 of data, PacBio data yielded ~36,000 predicted proteins. Similarly, PacBio sequencing recovered approximately twice as many spike-in sequences per $1,000 invested.

## Complete Characterization of Microbial Communities is within Reach

Reviewing the results, Todd DeSantis, Second Genome Co-founder and VP of Informatics, said "PacBio's long, accurate single molecule reads allowed us to more accurately discover novel and complete proteins that were encoded in our valuable and complex microbiome specimens. The prospect of bypassing assembly may increase our rate of discovery." Furthermore, given the median contig size of 24 kb, most of the PacBio genes are collocated with numerous other genes on the same contig.

The results demonstrate that long-read sequencing technology can be successfully applied to metagenomes from complex communities as a complement to short-read technology. PacBio sequencing was more effective at discovering complete genes than short-read sequencing in this study. Even more exciting is the prospect of replicating this study on the forthcoming Sequel II System which, with ~8-times more data, will make the cost comparison even more favorable. Dr. Irina Shilova, a Metagenomic Scientist at Second Genome, said "We anticipate the Sequel II System may enable scientists to more completely characterize the genetic potential of microbial communities. This could ultimately yield significant discoveries in the microbiome and metagenomic space at a reduced cost to what we are seeing today with Sequel System and short reads."

**Todd DeSantis, Co-founder and VP Informatics, Second Genome**

"PacBio's long, accurate single molecule reads allowed us to more accurately discover novel and complete proteins that were encoded in our valuable and complex microbiome specimens. The prospect of bypassing assembly may increase our rate of discovery."

**– Todd DeSantis**
Co-founder and
VP Informatics, Second Genome

Learn more about the Sequel II System at **pacb.com/sequel**

MICROBIOLOGY AND INFECTIOUS DISEASE



PN: CS123-052919