Metagenomic Analysis of Type II Diabetes Gut Microbiota Using PacBio HiFi Reads Reveals Taxonomic and Functional Differences

Daniel Portik[1], Joan Wong[1], Khi Pin Chua[1], Baoli Zhu[2], Na Lyu[2], Fei Liu[2], Xiaofeng Song[3], Yong Xue[4], Weigang Zhao[5], Shixuan Liu[5], Tao Yuan[5], Yong Fu[5], Gregory Young[1], Richard Hall[1]

[1]Pacific Biosciences, Menlo Park, CA, [2]Chinese Academy of Sci., Beijing, China, [3]Chinese Ctr. for Disease Control and Prevention, Beijing, China, [4]China Agricultural Univ., Beijing, China, [5]Peking Union Med. Coll. & Chinese Academy of Med. Sci., Beijing, China

Introduction

In the past decade, the human microbiome has been increasingly shown to play a major role in health. For example, imbalances in gut microbiota appear to be associated with Type II diabetes mellitus (T2DM) and cardiovascular disease. Coronary artery disease (CAD) is a major determinant of the long-term prognosis among T2DM patients, with a 2- to 4-fold increased mortality risk when present. However, the exact microbial strains or functions implicated in disease need further investigation.

From a large study with 523 participants (185 healthy controls, 186 T2DM patients without CAD, and 106 T2DM patients with CAD), 3 samples from each patient group were selected for long read sequencing. Each sample was prepared and sequenced on one Sequel II System SMRT Cell, to assess whether long accurate PacBio HiFi reads could yield additional insights to those made using short reads.

Each of the 9 samples was subject to metagenomic assembly and binning, taxonomic classification and functional profiling. Results from metagenomic assembly and binning show that it is possible to generate a significant number of complete MAGs (Metagenome Assembled Genomes) from each sample, with over half of the high-quality MAGs being represented by a single circular contig. We show that differences found in taxonomic and functional profiles of healthy versus diabetic patients in the small 9-sample study align with the results of the larger study, as well as with results reported in literature. For example, the abundances of beneficial short-chain fatty acid (SCFA) producers such as *Phascolarctobacterium faecium* and *Faecalibacterium prausnitzii* were decreased in T2DM gut microbiota in both studies, while the abundances of quinol and quinone biosynthesis pathways were increased as compared to healthy controls.

In conclusion, metagenomic analysis of long accurate HiFi reads revealed important taxonomic and functional differences in T2DM versus healthy gut microbiota. Furthermore, metagenome assembly of long HiFi reads led to the recovery of many complete MAGs and a significant number of complete circular bacterial chromosome sequences.

Data

Sampling Design and HiFi Data Yields

Three samples in each group:

- Healthy controls
- Type II diabetes + cardiovascular disease (CAD)
- Type II diabetes only

→ Each sample sequenced on one Sequel II System SMRT Cell 8M

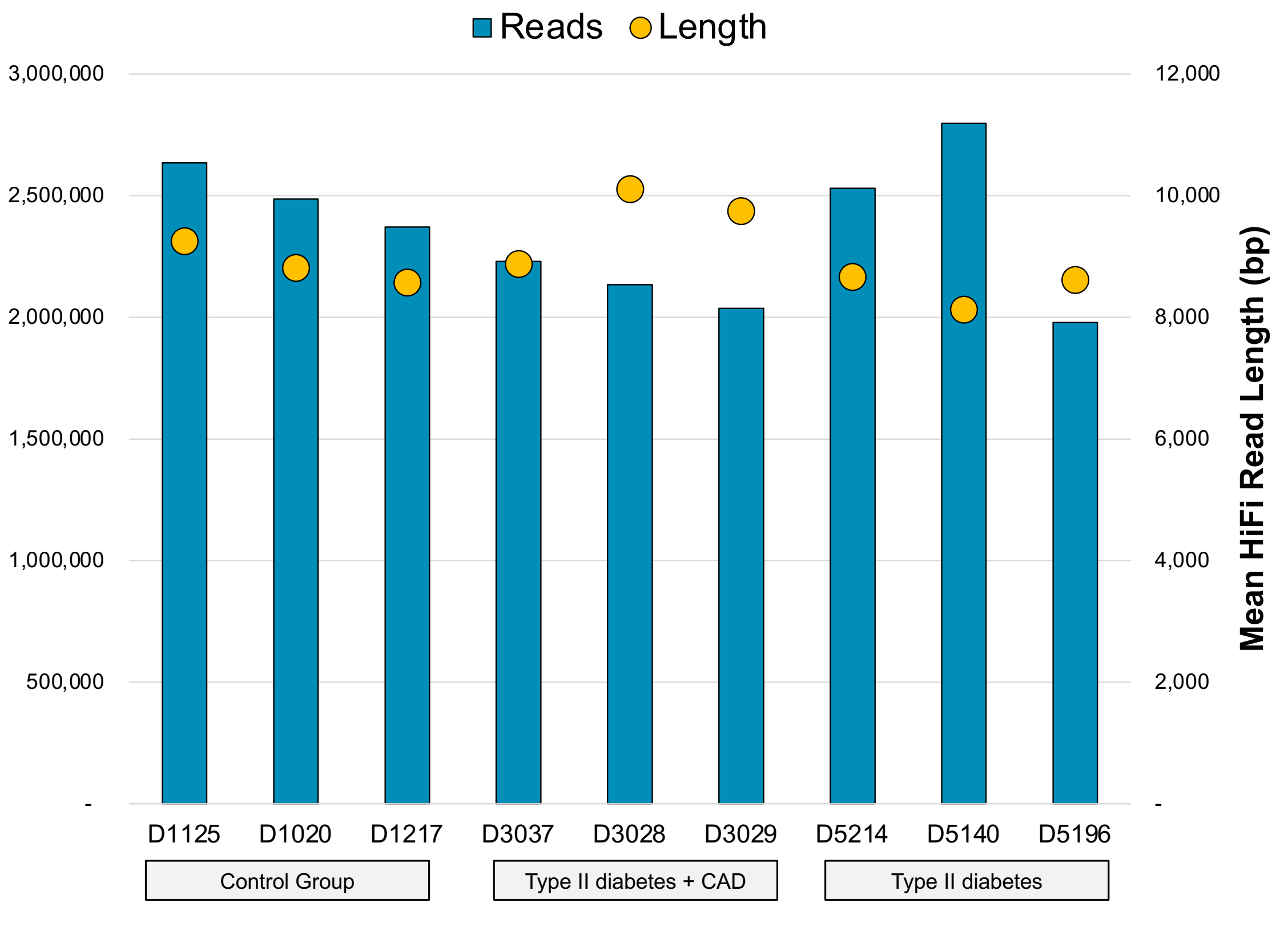


Figure 1. The number of HiFi reads and the average HiFi read length for all 9 samples. Typical yields were 2.0–2.5M reads, with mean read lengths 8–10 kb.

Assembly

HiFi reads for each sample were assembled using Canu[1] (v1.8) with settings specific to PacBio HiFi data.

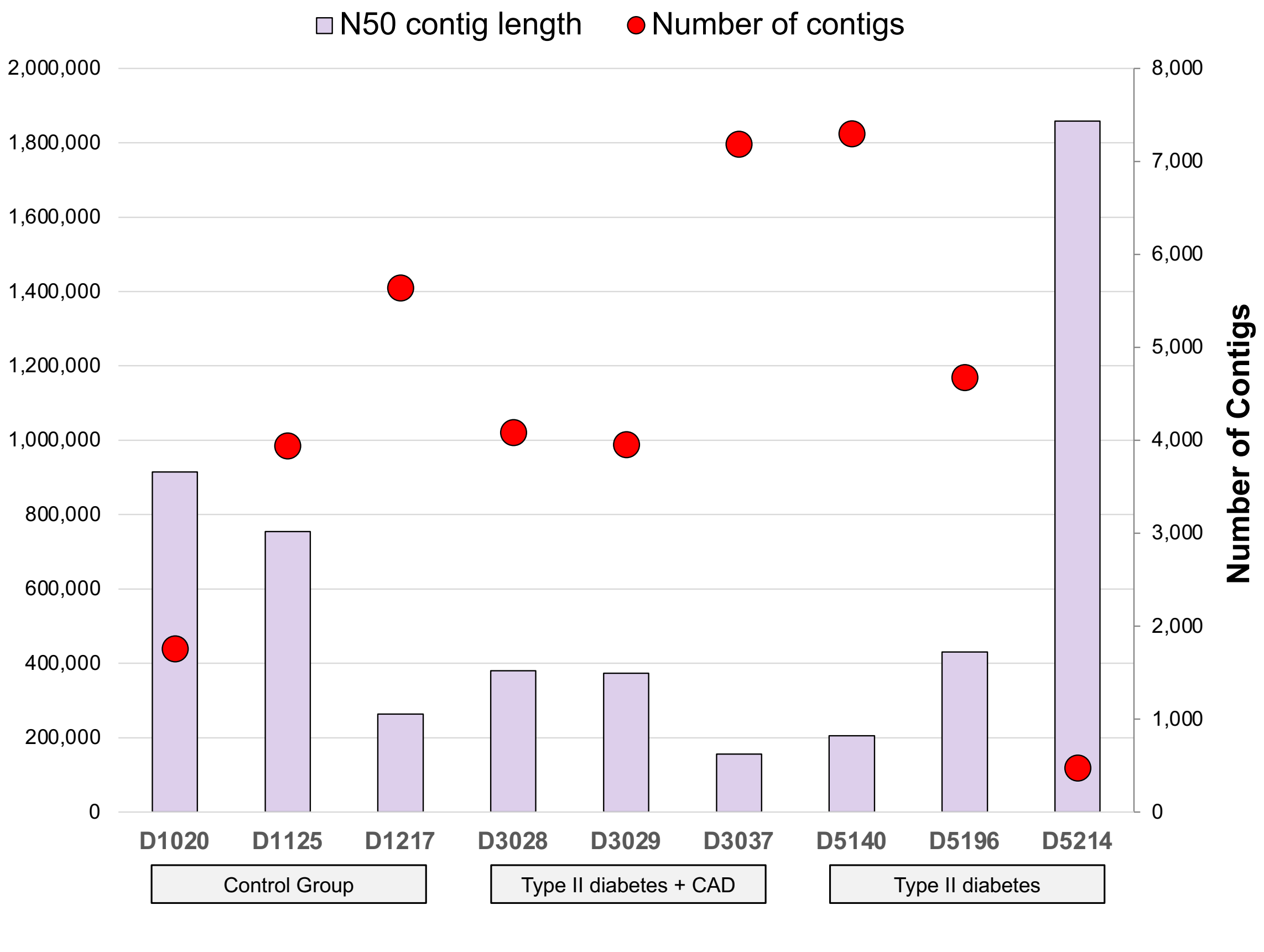


Figure 2. Summary of the N50 length and number of assembled contigs per sample. The N50 length and the number of assembled contigs vary considerably across samples. However, most samples had 40–70 contigs that were greater than 1Mbp in length.

Metagenomic Binning

Genome Binning Workflow

1. Map HiFi reads to Canu contigs with Minimap2[2]
2. Generate bins using MetaBAT2[3]
3. Assess bin quality with CheckM[4]
4. Select highest quality bins:
   - >70% completeness
   - <10% contamination
   - <10 contigs
5. Identify taxonomy of high-quality bins using GTDB-Tk[5]

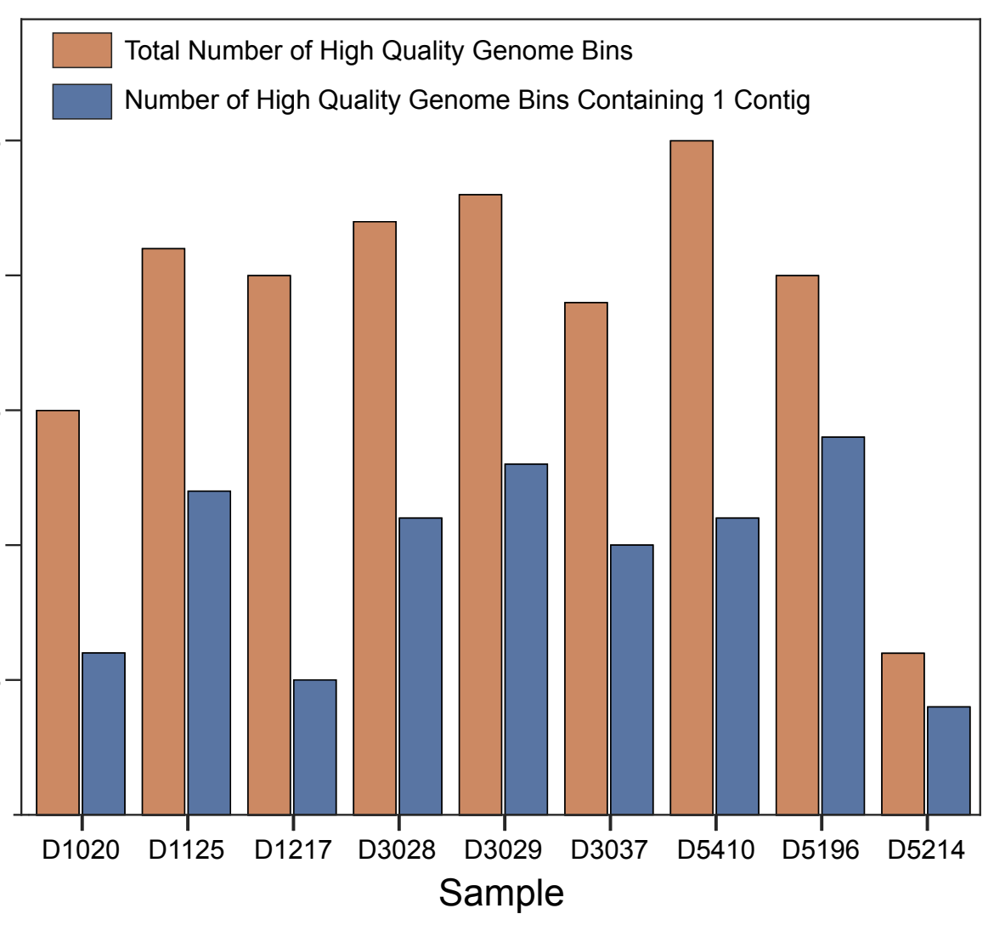
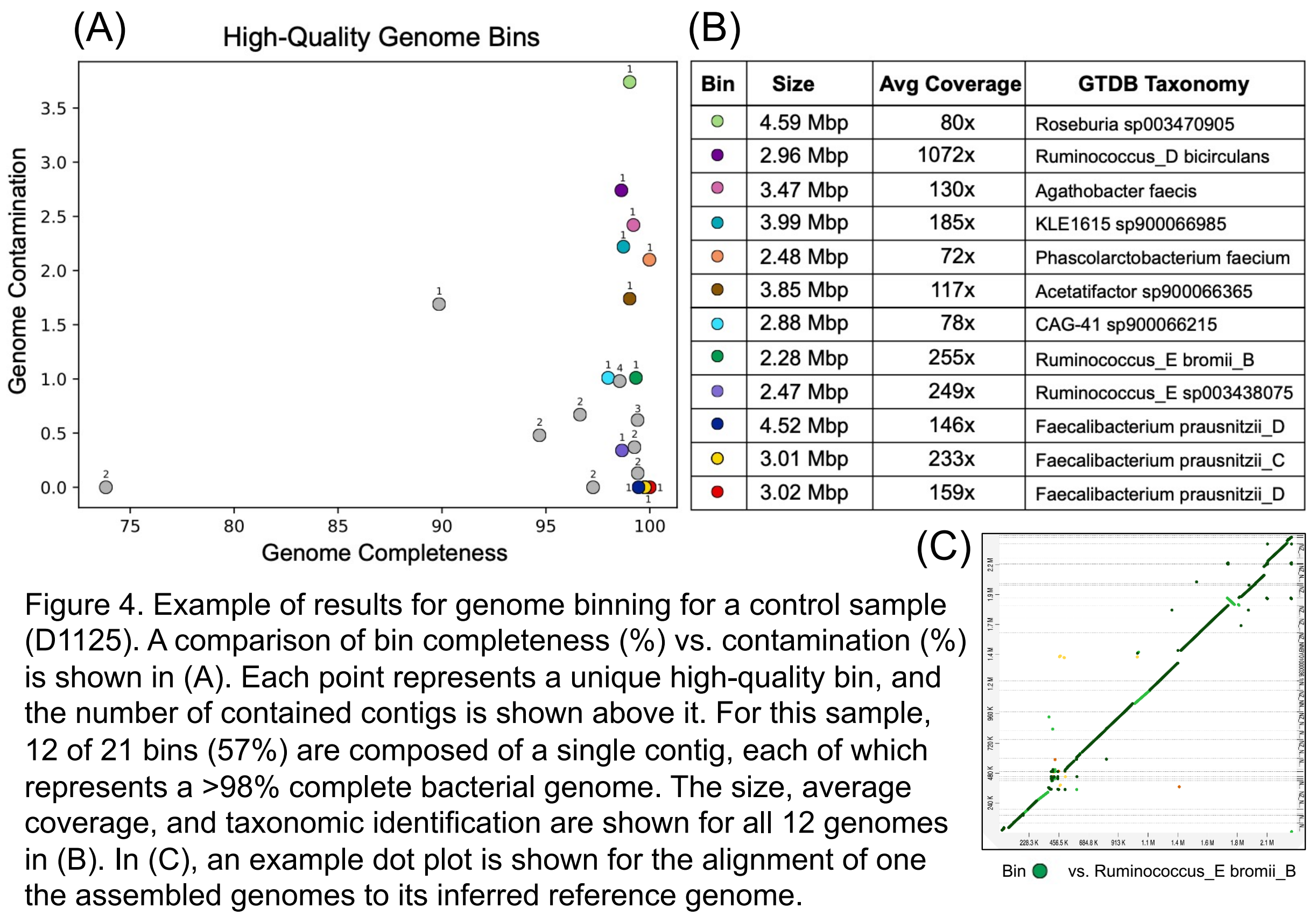


Figure 3. The number of high-quality genome bins recovered per sample (orange), which typically included 15–25 bins. For most samples, more than half the bins contained a single contig (blue). These represent high-quality assembled bacterial genomes.



Bin	Size	Avg Coverage	GTDB Taxonomy
●	4.59 Mbp	80x	Roseburia sp003470905
●	2.96 Mbp	1072x	Ruminococcus_D bicirculans
●	3.47 Mbp	130x	Agathobacter faecis
●	3.99 Mbp	185x	KLE1615 sp900066985
●	2.48 Mbp	72x	Phascolarctobacterium faecium
●	3.85 Mbp	117x	Acetatifactor sp900066365
●	2.88 Mbp	78x	CAG-41 sp900066215
●	2.28 Mbp	255x	Ruminococcus_E bromii_B
●	2.47 Mbp	249x	Ruminococcus_E sp003438075
●	4.52 Mbp	146x	Faecalibacterium prausnitzii_D
●	3.01 Mbp	233x	Faecalibacterium prausnitzii_C
●	3.02 Mbp	159x	Faecalibacterium prausnitzii_D

Figure 4. Example of results for genome binning for a control sample (D1125). A comparison of bin completeness (%) vs. contamination (%) is shown in (A). Each point represents a unique high-quality bin, and the number of contained contigs is shown above it. For this sample, 12 of 21 bins (57%) are composed of a single contig, each of which represents a >98% complete bacterial genome. The size, average coverage, and taxonomic identification are shown for all 12 genomes in (B). In (C), an example dot plot is shown for the alignment of one the assembled genomes to its inferred reference genome.

Functional Profiling

MEGAN6 Taxonomic and Functional Profiling workflow

1. Align HiFi reads, to a protein database such as NCBI nr or RefSeq using DIAMOND[6]
2. Convert DIAMOND output file to MEGAN6[7,8] long read input format
3. Perform taxonomic classification and functional profiling in MEGAN6

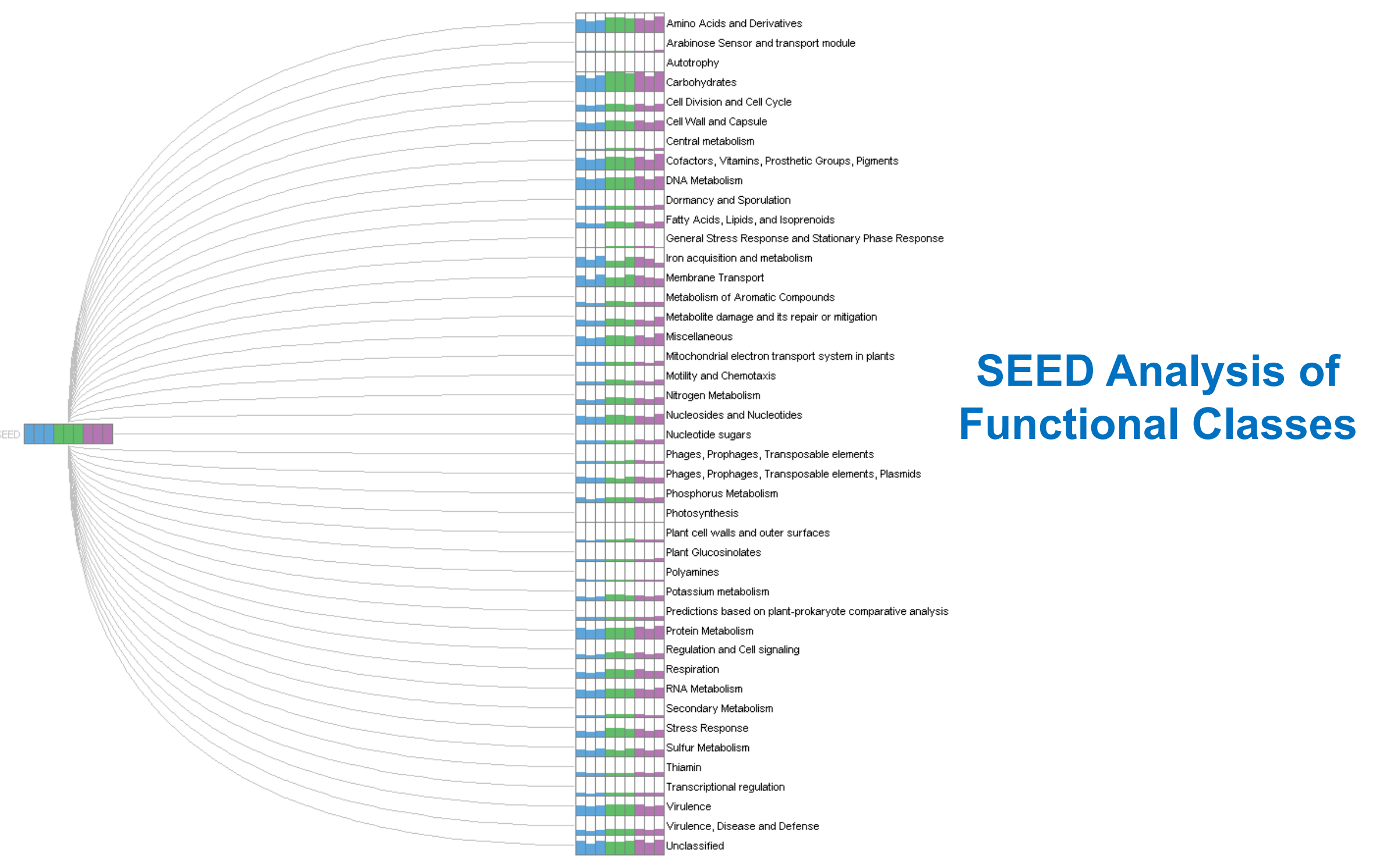


Figure 5. The SEED viewer in MEGAN6 showing a hierarchical tree of major functional classes. The representation of the different classes is shown for each sample in the bar charts, including samples in the control group (blue), Type II diabetes + CAD (green), and Type II diabetes (purple). Functional classes can be expanded to examine representation in various subcategories.

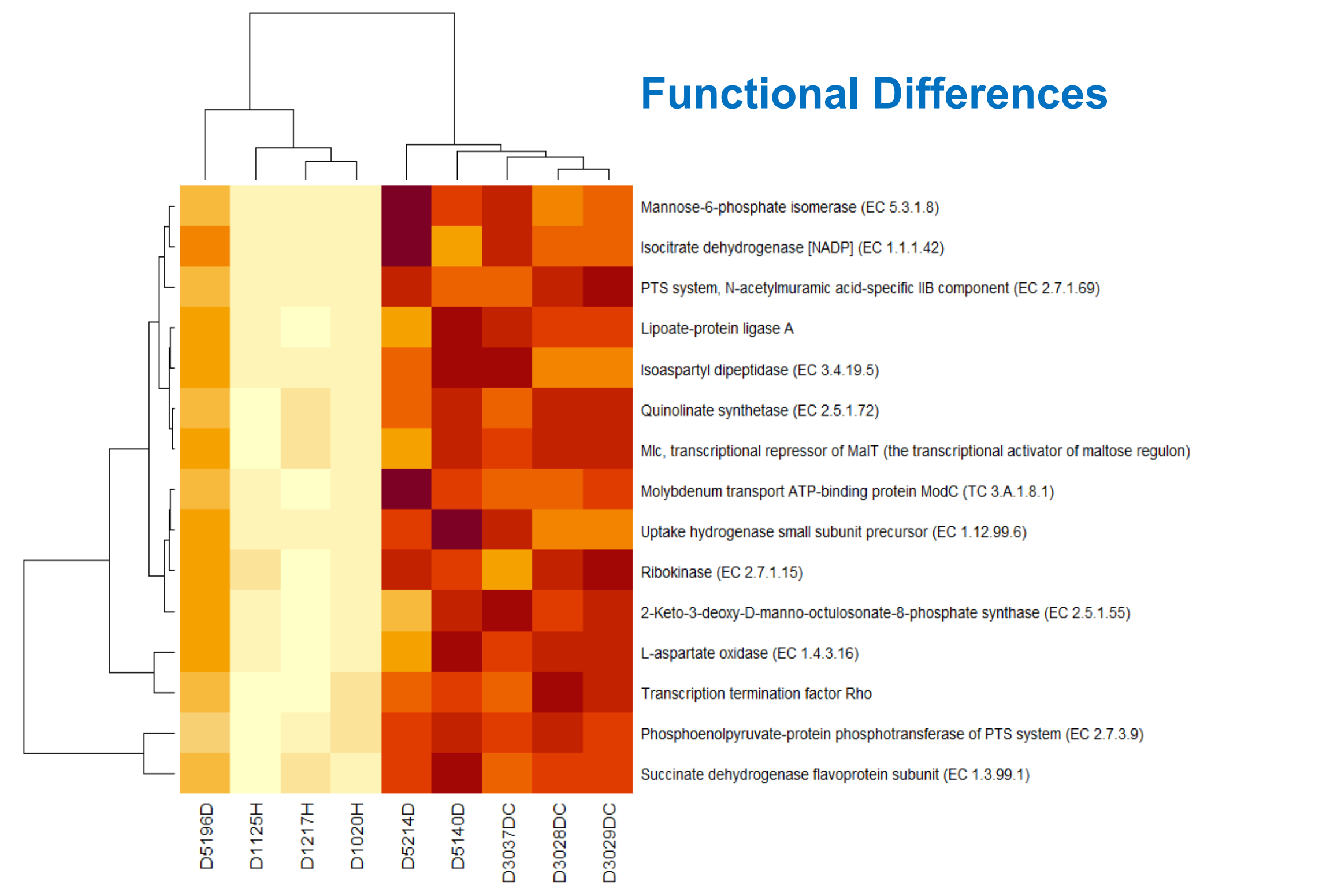
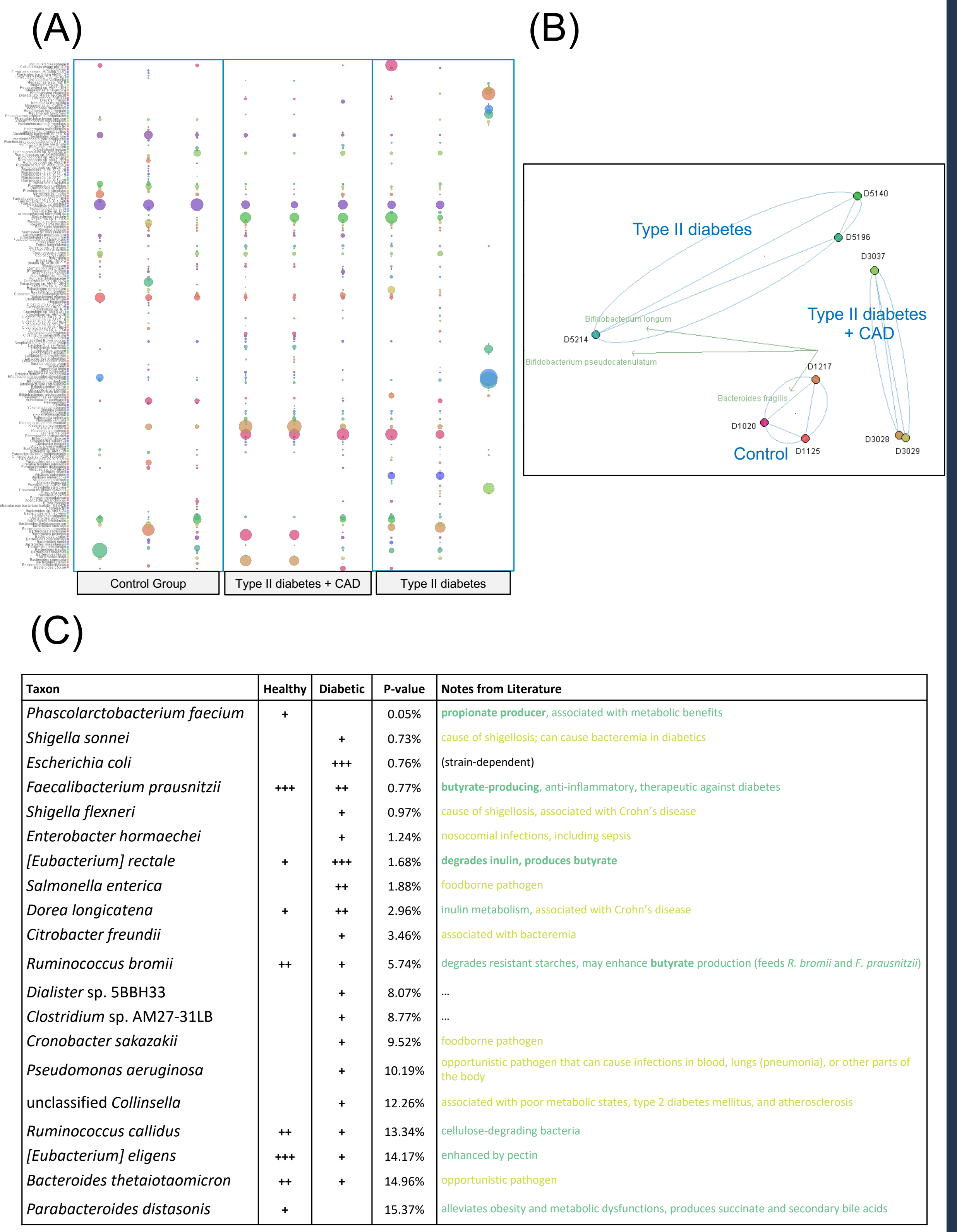


Figure 6. Heat map of the top 20 functional differences found across sample groups (H: Control; D: Type II diabetes; DC: Type II diabetes + CAD). Functional classification counts were exported from MEGAN6. The control group is distinguishable from other samples.

Taxonomic Classification

Taxonomic classification was performed using MEGAN6.



Taxon	Healthy	Diabetic	P-value	Notes from Literature
Phascolarctobacterium faecium	+		0.05%	**propionate producer**, associated with metabolic benefits
Shigella sonnei		+	0.73%	cause of shigellosis; can cause bacteremia in diabetics
Escherichia coli		+++	0.76%	(strain-dependent)
Faecalibacterium prausnitzii	+++	++	0.77%	**butyrate-producing**, anti-inflammatory, therapeutic against diabetes
Shigella flexneri		+	0.97%	cause of shigellosis, associated with Crohn's disease
Enterobacter hormaechei		+	1.24%	nosocomial infections, including sepsis
[Eubacterium] rectale	+	+++	1.68%	**degrades inulin, produces butyrate**
Salmonella enterica		++	1.88%	foodborne pathogen
Dorea longicatena	+	++	2.96%	inulin metabolism, associated with Crohn's disease
Citrobacter freundii		+	3.46%	associated with bacteremia
Ruminococcus bromii	++	+	5.74%	degrades resistant starches, may enhance **butyrate** production (feeds *R. bromii* and *F. prausnitzii*)
Dialister sp. 5BBH33		+	8.07%	…
Clostridium sp. AM27-31LB		+	8.77%	…
Cronobacter sakazakii		+	9.52%	foodborne pathogen
Pseudomonas aeruginosa		+	10.19%	opportunistic pathogen that can cause infections in blood, lungs (pneumonia), or other parts of the body
unclassified *Collinsella*		+	12.26%	associated with poor metabolic states, type 2 diabetes mellitus, and atherosclerosis
Ruminococcus callidus	++	+	13.34%	cellulose-degrading bacteria
[Eubacterium] eligens	+++	+	14.17%	enhanced by pectin
Bacteroides thetaiotaomicron	++	+	14.96%	opportunistic pathogen
Parabacteroides distasonis	+		15.37%	alleviates obesity and metabolic dysfunctions, produces succinate and secondary bile acids

Figure 7. A visualization of the differences in the abundance of bacterial species/strains across samples is show in (A). A principle component analysis of samples based on their taxonomic profiles is shown in (B). The bacterial taxa contributing the most to the variation are shown as vectors indicating the direction of the steepest increase. Table (C) details the 20 species with the most significant change in abundance between the healthy and diabetic patient groups. The included 'Notes from Literature' detail some of the important functions of the specific species (particularly those related to short-chain fatty acids).

Conclusion

Metagenomics analysis tools for HiFi data

We present several tools and analysis pipelines suitable for working with shotgun metagenomic HiFi data.

Metagenome assembly:

- Canu assembler generates high quality metagenomic assemblies
- New binning workflow can generate high quality MAGs and identify single contigs that represent complete bacterial chromosomes
- Pipeline is available on github: PacificBiosciences/pb-metagenomics-tools

Taxonomic classification and functional binning:

- DIAMOND can be used to align HiFi reads to a protein database
- Results can be visualized in the MEGAN6 software
- The SEED database can be used to investigate function
- Pipeline is available on github: PacificBiosciences/pb-metagenomics-tools

Functional profiling and taxonomic classification

While only a small number of samples were sequenced, the results of the functional profiling and taxonomic classification match what was seen in a larger short read dataset and are supported by the literature. The taxonomic classification shows fewer beneficial strains of bacteria in the guts of affected individuals. Given the small number of samples sequenced the results are largely observational, and a much larger cohort would need to be sequenced to show strong statistical significance.

The ability to generate high quality MAGs together with taxonomic and functional data demonstrates the high value of HiFi datasets in metagenomic analysis.

References

1. Koren, S., Walenz, B.P., Berlin, K., *et al.* (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* 27: 722-736
2. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34: 3094-3100.
3. Kang, D.D., Li, F., Kirton, E., *et al.* (2019). MetaBAT2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7: e7359.
4. Parks, D.H., Imelfort, M., Skennerton, C.T., *et al.* (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* 25: 1043-1055.
5. Chaumeil, P.-A., Mussing, A.J., Hugenholtz, P., *et al.* (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 36(6): 1925-1927.
6. Buchfink, B., Xie, C., & Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12: 59-60.
7. Huson, D.H., Beier, S., Flade, I., *et al.* (2016). MEGAN Community Edition – interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Computational Biology* 12(6):e1004957.
8. Huson, D.H., Albrecht, B., Bagci, C., *et al.* (2018). MEGAN-LR: New algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology Direct* 13(1): 6.