# A Workflow for the Analysis of Contigs from the Metagenomic Shotgun Assembly of SMRT® Sequencing Data

Richard J Hall[1], Meredith Ashby[1], Brett Bowman[1], Cheryl Heiner[1], Sudeep Mehrotra[2], Nikoleta Juretic[2], Jessica Wasserscheid[2] and Ken Dewar[2]
[1] Pacific Biosciences, Menlo Park, CA 94025
[2] McGill University, Montreal, Quebec, Canada

## Abstract

The throughput of SMRT® Sequencing and long reads allows microbial communities to be analyzed using a shotgun sequencing approach. Key to leveraging this data is the ability to cluster sequences belonging to the same member of a community. Long reads of up to 40 kb provide a unique capability in identifying those relationships, and pave the way towards finished assemblies of community members. Long reads are highly valuable when samples are more complex and containing lower intra-species variation, such as a larger number of closely related species, or high intra-species variation.

Here, we present a collection of tools tailored for the analysis of PacBio® metagenomic assemblies. These tools allow for improvements in the assembly results, and greater insight into the complexity of the study communities.

Supervised classification is applied to a large set of sequence characteristics (e.g. GC content, raw read coverage, k-mer frequency, and gene prediction information) and to cluster contigs from single or highly related species. Assembly in isolation of the raw data associated with these contigs is shown to improve assembly statistics. A unique feature of SMRT Sequencing is the availability to leverage simultaneously collected base modification / methylation data to aid the clustering of contigs expected to comprise a single or very closely related species. We demonstrate the added value of base modification information to distinguish and study variation within metagenomic samples based on differences in the methylated DNA motifs involved in the restriction modification system.

Application of these techniques is demonstrated on a mock community and monkey intestinal microbiome sample.

## Monkey Intestinal Microbiome - Clustering

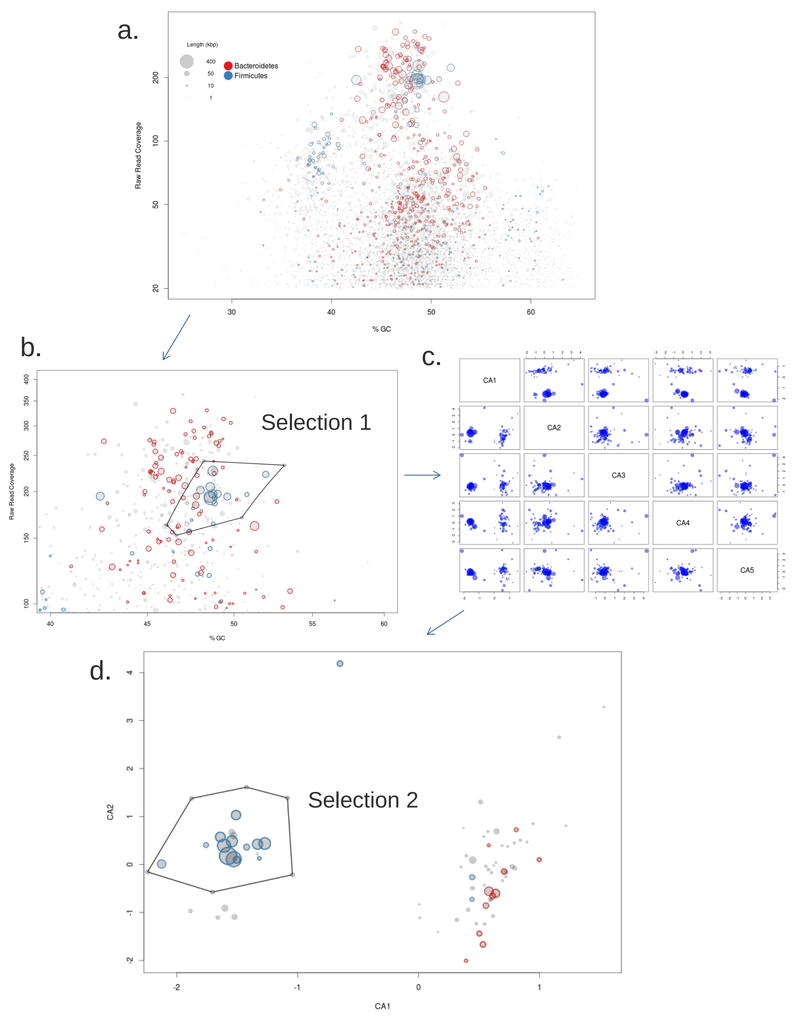


**Figure 1.** Workflow for clustering HGAP-generated contigs by sequence context, method adapted from Albertsen *et. al.* The first plot (a) shows contigs plotted by raw read coverage vs. GC content, size indicates contig size, and color taxonomic prediction. The second plot (b) shows an enlarged region of the first plot, with a manual attempt at selecting contigs belonging to one OTU. Selection 1 obviously includes contigs belonging to at least two OTUs, as indicated by the color, the bottom two panels show a further representation of the data in order to improve classification. Plot c shows the results of a correspondence analysis of the 4-mer sequence context from the contigs in Selection 1. The final plot (d) shows the contigs from Selection 1 re-plotted using the first two components from the correspondence analysis. It can be seen from the final plot that the 4-mer sequence context allows a finer classification of the contigs than plotting by coverage and GC content. Again manual selection is used to select for contigs belonging to one OTU (Selection 2).

## Monkey Intestinal Microbiome - Clustering Results

| Attribute | Selection 1 | Selection 2 |
|---|---|---|
| #Bases | 4140269 | 2685615 |
| #Contigs | 73 | 21 |
| Mean Length | 56716 | 127886.4 |
| Max Length | 467651 | 467651 |
| % GC | 48.5 | 48.7 |
| Coverage | 195.3 | 196.5 |
| Essential Genes | 105 | 85 |
| Unique Essential Genes | 86 | 80 |

**Table 1.** Results from the selection of data from the above manual clustering of HGAP-generated contigs. Selection 1 is the initial set of contigs by manual selection from a plot of coverage vs. GC content. Selection 2 is a subset of Selection 1 based on a manual selection of contigs from a correspondence analysis of 4-mer sequence context. The number of contigs is significantly reduced in Selection 2 (73 to 21), while the number of unique essential genes (Albertsen *et. al*.) is maintained (86 to 80) and the mean contig length is increased.

| Motif | Modification Type | % Motifs Detected Selection 1 | % Motifs Detected Selection 2 |
|---|---|---|---|
| CAT**A**TG | m6A | 63.14 | 72.42 |
| GNGYA**A**G | m6A | 28.21 | Not Detected |
| CTGC**A**GD | m6A | 25.07 | 79.47 |

**Table 2.** Base Modification detection in the two sets of selected contigs. Two out of three motifs detected in Selection 1 with >50 modification QV are found in Selection 2, but with much higher % detection. Within a single OTU 100% of a given motif would be expected to be modified.

## Strain Variation – Mock Dataset

| Sample | Size | Coverage |
|---|---|---|
| *E. coli* | 4.5 Mb | 335x |
| Streptomyces A | 8 Mb | 180x |
| Streptomyces B | 7.8 Mb | 60x |
| *C. difficile* A | 4 Mb | 265x |
| *C. difficile* B | 4 Mb | 160x |

**Table 3.** A model dataset was assembled from five independent sequencing experiments. The table above shows the makeup of the dataset. A distinct individual (*E. coli*) was included as a control, two Streptomyces at ~80% sequence identity in similar regions, and two *C. difficile* at ~97% identity across the genome.

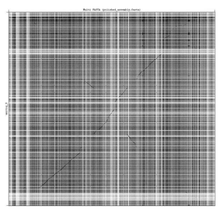

**Figure 2.** Dot plot showing the similarity between Streptomyces A & B

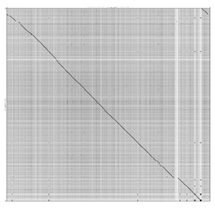

**Figure 3.** Dot plot showing the similarity between *C. difficile* A & B

## Strain Variation – Assembly Workflow

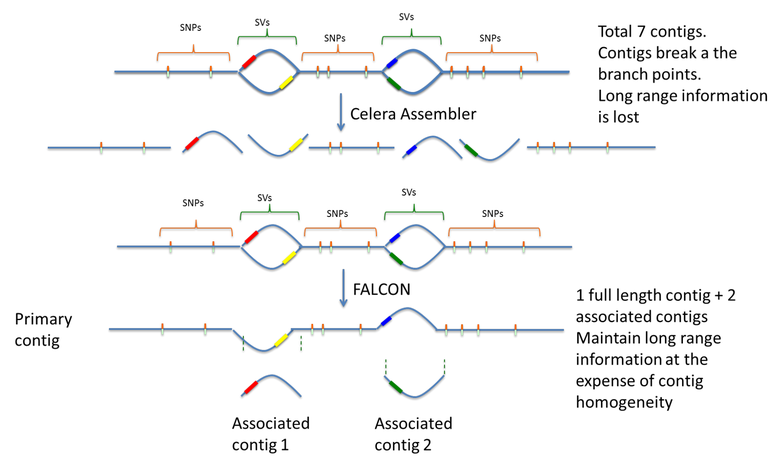


**Figure 4.** Workflow for assembling data from a mixed population with structural differences. The standard OLC (Overlap-Layout-Consensus) approach implemented in Celera Assembler correctly breaks the above graph structure at the branch points generating 7 contigs. The FALCON string graph assembler (https://github.com/PacificBiosciences/falcon) maintains long-range information by forming a primary contig from the longest path, and associated contigs that contain alternative sequence representing the structural variation. Note the primary contig is not phased and will be chimeric for the two strains.

## Strain Variation – Assembly Results

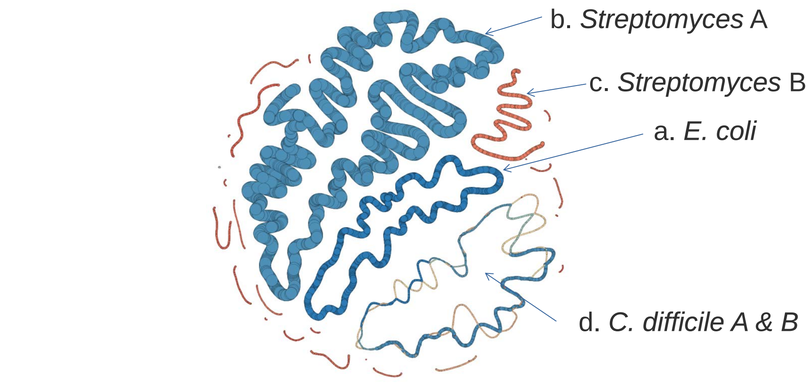


**Figure 5.** Overlap graph generated from the output of a Celera assembly of the model dataset. Color indicates the read coverage of the generated contigs, red low, blue high. Node size corresponds to total contig size. (**a**) indicates the circular graph of the *E. coli* control, (**b**) and (**c**) indicate the complete high-coverage *Streptomyces* A, and the fragmented graph from the low-coverage *Streptomyces* B. (**d**) indicates the circular graph for the assembly of the two strains of *C. difficile,* note the bubbles in the graph caused by the structural differences between the two strains. The contigs generated from d are fragmented when compared to the genome of either strain.

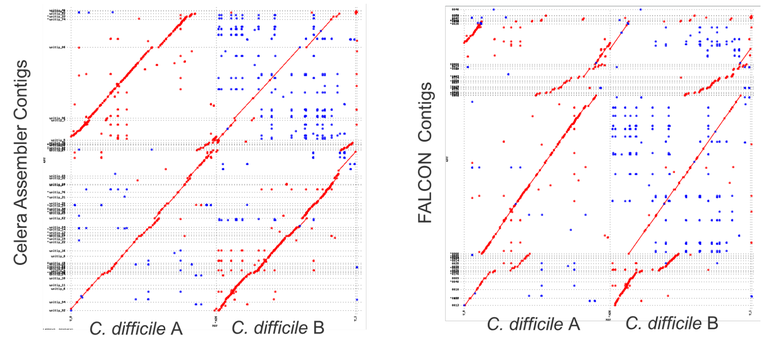


**Figure 6.** Results from the assembly of the same error-corrected reads using Celera Assembler and FALCON. Strain variation in the mixed sample results in a fragmented assembly when using Celera Assembler, the resulting contigs are largely homogeneous for a single strain, but long range information is lost. The FALCON assembly maintains long-range information for the loss of contig homogeneity. The largest contig in the FALCON assembly ~3.1 Mb compares to 1.8 Mb for Celera Assembler.

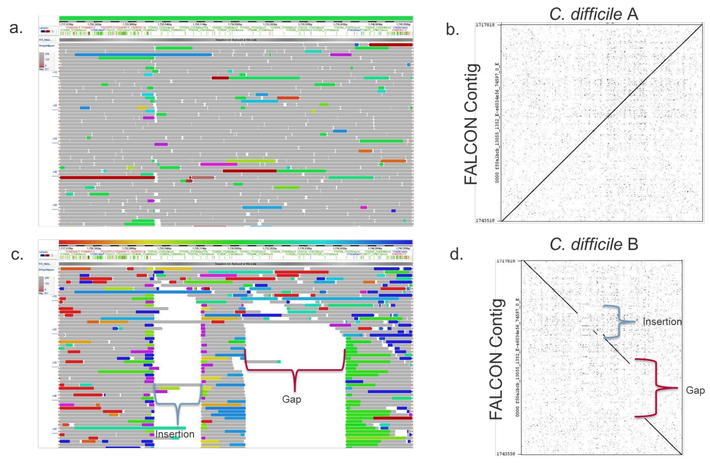


**Figure 7.** The Bridgemapper protocol in the SMRT Analysis system can be used to investigate structural variations by aligning the read data to the large FALCON contig. (**a**) shows the alignment of all the data, (**c**) shows the same alignment, but selecting for reads that have a bridged mapping. The insertion shows that reads, while primarily mapping here also map to a secondary contig indicated by the purple and blue. Also shown are reads mapping across a gap, green and light blue. The dot plots (**b**) and (**d**) show the reference genomes compared to the region shown in the alignment, from this it can be seen that the reads in **c** belong to *C. difficile* B which has structural variation with respect the to FALCON contig.

## Conclusion

Unique characteristics of SMRT Sequencing data, in particular the long reads and base modification information, can be leveraged in metagenomic analysis. We present a number of assembly procedures that maximize the information that can be gained from complex metagenomic samples to deliver strain-level information from a whole-genome shotgun experiment.

## References

Chin, *et al*. “Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data” Nat Methods, 10, 563-569 (2013)

Albertsen, *et al*. “Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes.” *Nat Biotechnol*, 31(6) (2013).