An interactive workflow for the analysis of contigs from the metagenomic shotgun assembly of SMRT® Sequencing data

Richard J Hall[1], Chen-Shan Chin[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

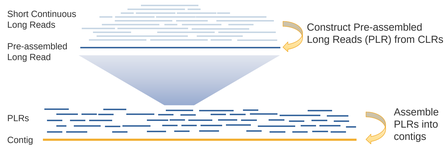
Abstract

The data throughput of next-generation sequencing allows whole microbial communities to be analyzed using a shotgun sequencing approach. Because a key task in taking advantage of these data is the ability to cluster reads that belong to the same member in a community, single-molecule long reads of up to 30 kb from SMRT Sequencing provide a unique capability in identifying those relationships and pave the way towards finished assemblies of community members. Long reads become even more valuable as samples get more complex with lower intra-species variation, a larger number of closely related species, or high intra-species variation. Here we present a collection of tools tailored for PacBio® data for the analysis of these fragmented metagenomic assembles, allowing improvements in the assembly results, and greater insight into the communities themselves.

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, allowing the clustering of contigs from single or highly related species. A unique feature of SMRT Sequencing data is the availability of base modification / methylation information, which can be used to further analyze clustered contigs expected to be comprised of single or very closely related species. Here we show base modification information can be used to further study variation, based on differences in the methylated DNA motifs involved in the restriction modification system.

Application of these techniques is demonstrated on a monkey intestinal microbiome sample and an in silico mix of real sequencing data from distinct bacterial samples.

HGAP Assembly



HGAP (Hierarchical Genome Assembly Process) (Chin *et. al.*). The HGAP algorithm, implemented in the SMRT Analysis package (http://www.pacificbiosciences.com/devnet/) is used for assembly in all stages of the workflow.

Clustering – Monkey Intestinal Microbiome

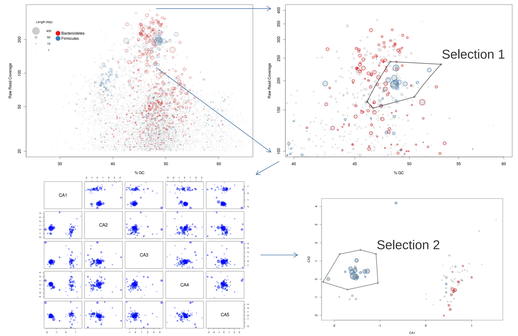


Figure showing the workflow for clustering HGAP generated contigs by sequence context, method adapted from Albertsen *et. al.* The first plot shows contigs plotted by raw read coverage vs. GC content, size indicates contig size, and color taxonomic prediction. The second plot shows an enlarged region of the first plot, with a manual attempt at selecting contigs belonging to one OTU. The initial selection 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. The first plot shows the results of a correspondence analysis of the 4-mer sequence context from the contigs in Selection 1. The final plot 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.

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

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 conitgs 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.

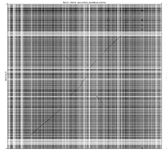
Motif	Modification Type	% Motifs Detected Selection 1	% Motifs Detected Selection 2
CATATG	m6A	63.14	72.42
GNGYAAG	m6A	28.21	Not Detected
CTGCAGD	m6A	25.07	79.47

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.

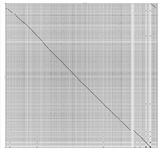
Strain Variation – Dataset

A model dataset was assembled from 5 independent sequencing experiments. The table below 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.

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



Dot plot showing the similarity between *Streptomyces* A & B



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

Strain Variation - Assembly

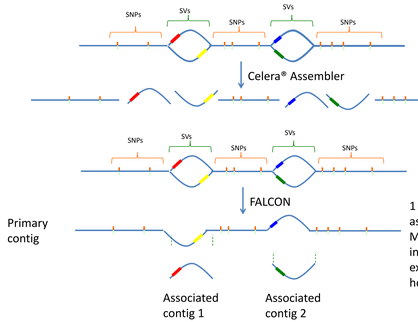
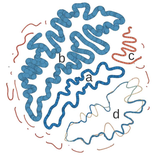
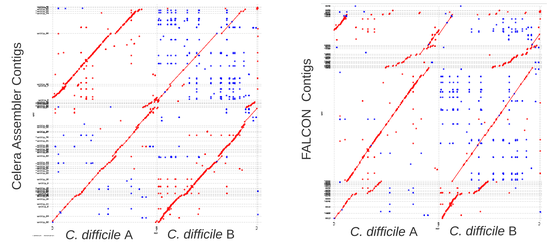


Figure showing the results of assembling data from a mixed population with structural differences. The standard OLC (Overlap-Layout-Consensus) approach implemented in Celera Assembler correctly breaks the above graph structure at the branch points generating 7 contigs. The FALCON string graph assembler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

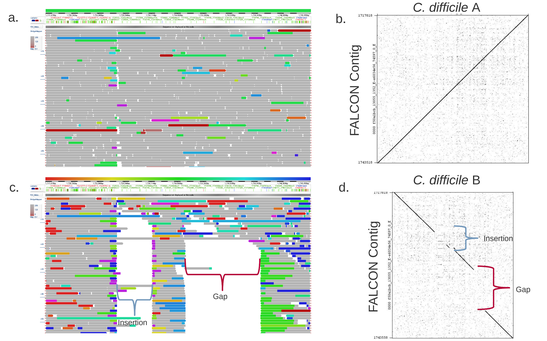


Overlap graph generated from the output of a Celera assembly of the model dataset. Color indicates the read coverage of the generated contigs, red low, blue high. Node size corresponds to total contig size. a indicates the circular graph of the *E. coli* control; b & c indicate the complete high coverage *Streptomyces* A, and the fragmented graph from the low coverage *Streptomyces* B, respectively. 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.

Strain Variation - Results



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 compared to 1.8 Mb for Celera Assembler.



The Bridgemapper protocol in the SMRT Analysis system can be used to investigate structural variations be 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 to the FALCON contig.

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)