SEQUENCING CULTURED REPRESENTATIVES IN THE HUMUN GUT MICROBIOME

Jian Xu[1], Makedonka Mitreva[1], Lucinda Fulton[1], Robert Fulton[1], Kym Hallsworth-Pepin[1], Sandra W. Clifton[1], Priya Sudarsanam[2], Ruth Ley[2], Michael Mahowald[2], Janaki Guruge[2], Jeffrey I. Gordon[2], Elaine R. Mardis[1], Richard K. Wilson[1]

[1]Genome Sequencing Center and [2]Center for Genome Sciences, Washington University School of Medicine, St. Louis, MO 63108

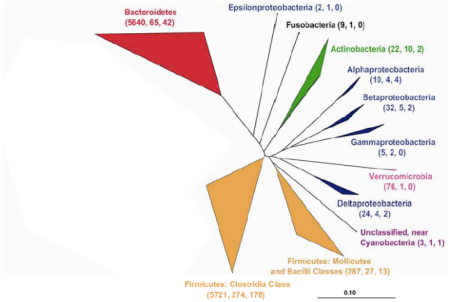
Introduction

The human intestine is a complex and dynamic ecosystem where hundreds of bacterial species co-exist, co-adapt and co-evolve with their host. Interactions between this community and the host are highly complex and poorly understood. Our Human Gut Microbiome Initiative (HGMI) aims to produce reference genomes for ~100 representatives from the human gut microbiota. Through this and subsequent metagenomic analysis we will explore the biodiversity in human gut and identify specific correlations among microbial organismal and gene lineages and host biology.

At least 15X 454 coverage (454 Life Science, GS20) and at least ~4X 3730 coverage (ABI 3730xl) are generated for each cultured representative and the data compiled into a 'hybrid' assembly using PCAP. One round of automated targeted sequence improvement is performed, followed by manual inspection and improvement. A vigorous experimental approach for identity assessment and a sensitive and highly automated computational QC system are combined to detect contamination at each step. Here we will present our sequencing approach and quality assessment of the assemblies.

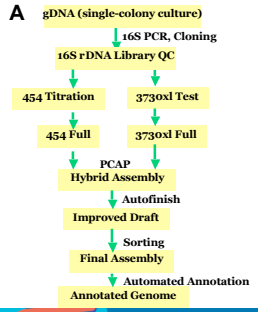
1 Composition of human intestinal microbiota.

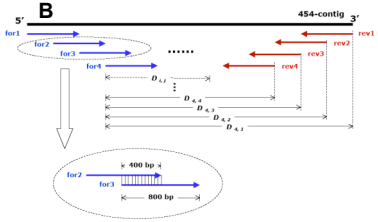
16S rRNA gene-based surveys (e.g. Eckburg, et al, *Science*, 2005; Ley et al *Nature*, 2006) indicate that two bacterial divisions, the Bacteroidetes and Firmicutes, contain more than 90% of the phylotypes in the distal gut microbiota of healthy humans. Our sequencing efforts mainly target cultured members from these two divisions.



2 Sequencing and assembly strategy.

Genomic DNA is prepared from an anaerobic culture started with a single colony of the targeted organism. At least 15X 454 and ~4X ABI 3730 xl coverage is generated (**A**). 454-reads are assembled into 454-contigs: these contigs are converted to 3730xl-like paired-end reads (**B**), which are then assembled together with 3730xl plasmid reads into a 'hybrid' assembly.





3 Assessment of identity and detection of contaminations.

Purification of gDNA from double streaked single-colony-derived culture

↓ 16S PCR using 27F - 1391R primer pair (Bacteria-specific); sequencing the amplicon.

Confirmation of gDNA identity.

↓ 16S rRNA gene PCR of gDNA using two primer pairs respectively: 27F - 1391R and 515F-1391R (universal); cloning and sequencing of 768 amplicons

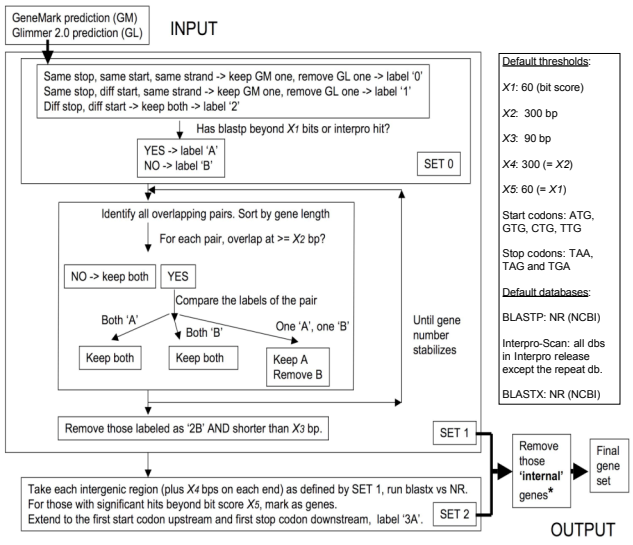
Detection of prokaryotic and eukaryotic contamination

↓
- GC Plot
- Searches against DB of phylogenetic markers
- Searches against NR (NCBI)
- Searches against other projects in our pipeline
- Abnormality in assembly, such as very low contig contiguity or bias in sequence coverage

Detection of contamination possibly introduced in the sequencing pipeline

4 Identification of protein-coding genes.

Predictions of GeneMark and Glimmer 2.0 are combined; an evidence-based approach is used to prioritize genes for inclusion into final gene set. Genes missed by the two *ab initio* gene predictors are identified using BlastX.



Note: *An **'internal'** gene is one that overlaps over its full length with another gene on the same frame of larger or similar size.

5 Impact of 454 and 3730xl read coverage on the utility of the hybrid assemblies for annotating coding sequences.

To determine the optimal mixture of 454 and 3730xl coverage, simulations were performed using the 1.6Mb finished genome of *H. pylori* HPAG1 (Oh et al *PNAS*, 2006): annotated nucleotide sequences of all coding sequences on the *H. pylori* HPAG1 chromosome were searched against the following simulated assemblies using BLASTN with an E value of 0.1. The reference for calculating cost ratios is 6.2X 3730xl plasmid reads. A draft assembly of 24X 454 coverage plus 6X plasmid reads yields 98% protein-coding genes in complete and accurate form.

	Finished sequence	6.2X plasmid reads	6.2X plasmid plus 0.8X fosmid	24X 454 reads	24X 454 reads plus 2.7X 3730 plasmid reads	24X 454 reads plus 4.5X 3730 plasmid reads	24X 454 reads plus 6.2X 3730 plasmid reads
ORFs fully recovered	1537 (100%)	1174 (76%)	1256 (82%)	1291 (84%)	1447 (94%)	1486 (97%)	1500 (98%)
ORFs not found	0	12	9	0	0	0	0
ORFs partially recovered	0	351	272	246	90	51	37
truncation errors only	-	205	147	53	25	14	11
substitution/indel errors only	-	31	27	180	59	30	19
truncation plus substitution/indel errors	-	115	98	13	6	7	7
tRNA fully recovered	36 (100%)	28 (78%)	30 (83%)	36 (100%)	36 (100%)	36 (100%)	36 (100%)
rRNA fully recovered	3 (100%)	3 (100%)	3 (100%)	3 (100%)	3 (100%)	3 (100%)	3 (100%)
Cost ratio	-	1.0	1.1	1.8	2.0	2.1	2.3

Note: 'not found': those that did not return any hits; 'fully recovered': those that generated alignments that cover 100% query length with 100% identity; 'truncation errors': those with alignments that have 100% identity yet cover < 100% of query length; 'substitution errors': those with alignments that cover 100% query length yet have < 100% identity; 'both truncation and substitution errors': the remaining coding sequences.

6 HGMI genomes status.

Only those genomes for which manual improvement has been completed are listed. For a complete list, see Gordon, *et al.* http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/HGMISeq.pdf

Genus	Species	Strain ID	Genome size (Mb)	GC%	Number of major contigs	Contig N50 size (Kb)
Actinomyces	*odontolyticus*	DSM 4331	2.44	65.2%	17	315
Bacteroides	*caccae*	ATCC 43185T	4.96	42.0%	20	432
Bacteroides	*capillosus*	ATCC 29799	4.25	59.1%	92	110
Bacteroides	*merdae*	ATCC 43184T	4.52	45.3%	31	302
Bacteroides	*ovatus*	ATCC 8483T	6.50	41.9%	39	426
Bacteroides	*uniformis*	ATCC 8492	4.68	46.8%	63	122
Bifidobacterium	*adolescentis*	L2-32	2.62	59.8%	16	289
Collinsella	*aerofaciens*	JCM 10188	2.47	60.5%	22	87
Dorea	*longicatena*	CCUG 45247	2.94	41.5%	34	117
Eubacterium	*ventriosum*	ATCC 27560	2.89	35.0%	28	227
Ruminococcus	*gnavus*	ATCC 29149	3.50	42.9%	51	101
Ruminococcus	*obeum*	ATCC 29174	2.61	41.6%	43	112
Ruminococcus	*torques*	ATCC 27756	2.80	42.0%	39	124

Summary

- We have tested and optimized a cost-effective, high-throughput sequencing, quality control and annotation pipeline for cultured bacterial isolates.
  - Our hybrid sequencing strategy typically includes at least 15X 454 GS-20 reads and 4X plasmid 3730xl reads.
  - Extensive and vigorous QC measures have been implemented to detect and remove contaminations.
- The first phase of HGMI aims to sequence ~100 cultured human gut symbiont genomes, mainly of Firmicutes and Bacteroidetes.
  - To date, a total of 13 species has undergone automated improvement and two more dozen are in various stages in the pipeline;
  - Project status, genome sequences and automated annotation are available at: http://genome.wustl.edu/

Acknowledgements: NHGRI Grant HG003079