A Catalog of Reference Genomes from the Human Microbiome

The Human Microbiome Jumpstart Reference Strains Consortium

***The list of authors is presented at the end of the manuscript**

Supporting Online Material

Materials and Methods

Selection of strains for DNA sequencing. The HMP is targeting five body sites to keep efforts focused: the gastrointestinal tract, the oral cavity, the vagina, the skin, and the nasal passage.

To accomplish this goal, the Jumpstart Centers formed a Strains working group (WG) and sub-WGs for each body site, composed of experts on the particular body site and members of relevant NIH institutes. These sub-WGs made recommendations for organisms of the human microbiome. Among the criteria used to select strains were 1) unknown species, that occupy a novel phylogenetic position; 2) strains implicated in a disease process; 3) previously known species, not yet sequenced; 4) known species, with previously sequenced genomes, but showing high sequence diversity and/or significant variation in phenotypic profiles.

Prior to the start of the HMP, the GCs had acquired strains from collaborators. An early effort was the Human Gut Microbiome Initiative at the Washington University supported by the National Human Genome Research Institute (NHGRI) (7), a project to sequence 100 strains from the gastrointestinal tract, exemplifying collaborations between GCs and the research community.

A subsequent NHGRI project expanded this effort to include two additional GCs (Baylor College of Medicine-Human Genome Sequencing Center and Broad Institute) and targeted another 200 strains, principally from the gastrointestinal tract and vagina (8).

The reference genomes that have been sequenced during the initial efforts by the GCs have been combined with the current list of targeted strains produced by the NIH "body-site" WGs to generate a combined HMP Project Catalog (1) hosted and maintained by the HMP DACC (9).This Project Catalog includes significant details on all genomes targeted for sequencing. This Project Catalog also includes metadata for all reference genomes. This metadata is fully compliant with the specification developed by the Genome Standards Consortium (10) and includes information such as habitat, gram-staining, isolate source and sequencing status. The Project Catalog is dynamic, and is revised on a regular basis as new organisms are nominated for inclusion and move through the sequencing process. The DACC web site also supports community input, and community-based recommendations are encouraged. Information on how to nominate an organism is at the HMP DACC web site(2).

All strains selected for sequencing were either already available in a public repository, or were made available through the Biodefense and Emerging Infections Research Resources Repository (BEI (11)). The Jumpstart Centers do not initiate sequencing an organism until it is in a public repository and available to the community. The strains that currently make up the Project Catalog are all culturable, since current methodology for sequencing uncultured organisms is not mature. However, four of the HMP Roadmap RFAs (RFA-RM-08-010, -011, -026, and -027) provide support for developing methods to sequence uncultivated organisms, and it is expected that genomes from uncultivated organisms will be sequenced in the future, and the Strains WG has formed a sub-group that is focused on this topic.

Provisional standards for draft genome sequences. The majority of the 900 genomes to be sequenced for the NIH HMP will be completed to draft genome quality, with ~15% being further improved. The two traditional categories of genome sequence quality, draft and finished, are not universally defined for bacterial genomes and one of the first tasks of the Jumpstart Centers was to agree on such definitions. This was accomplished by experimental comparison of genomes that had been sequenced to varying levels of quality.

The two principal deliverables for each genome are a genome sequence and a gene list. The sequence is to be used to assign sequences from metagenomics experiments while the gene list is to be used to model the metabolic capabilities and other phenotypes of the microbial community. Completeness and accuracy are important qualities for the utility of both of these products.

Producing 900 sequences in a two-year period is feasible by the high throughput and low cost of new sequencing technologies. The metrics and standards that define the high-quality draft genome for the HMP had to be platform independent since different methods are used and technology is expected to change during the project. Manual intervention, the previous norm for microbial sequencing, is not practical given the rate of genome production in this initiative. Adherence to strict standards is important for use of automated assembly and annotation in this project, with limited opportunities for downstream correction of sequences.

To set initial quality standards, three genomes with Gold Standard finished sequences and annotations were selected as test cases. *Staphylococcus aureus* USA300 (Sau, accession number CP000730) an AT-rich genome containing two plasmids, *Rhodobacter sphaeroides* 2.4.1 (Rsp, accession numbers CP000143 and CP000144) a GC-rich genome containing two chromosomes (one linear, one circular) as well as 5 plasmids, and *Escherichia coli* MG1655 (Eco, accession number U00096) a genome of balanced base composition and reasonable complexity, were selected for study.

Each genome was sequenced at the GCs using both Roche-454 and Illumina-Solexa methods to assess reproducibility. No significant differences in data were found between instruments, GCs, or protocols. The 454 datasets were assembled at 1x to 16x coverage with the Newbler assembler. Statistics from each assembly for each organism were computed to determine contiguity of contigs and scaffolds, genome coverage, and accuracy of assembly and base sequence was judged by comparison to the finished Gold Standards. Results for different assembly quality are shown in Figures S1 and S2.

Assemblies were also tested for efficacy in producing gene predictions, which were evaluated for completeness and accuracy. In this way a provisional set of metrics was determined that could be used by the HMP Consortium to ensure each genome released at high-quality draft status would meet a rigorously defined and consistent quality level. The standards are considered provisional and will be revisited after the first round of analyses on the completed genomes (see below).

Additional details of all metrics are available on the DACC website (12).

The provisional HMP draft genome standards, with their rationales, are as follows:

(1) >90% of the genome included in contigs (≥ 500bp) so as to ensure completeness for identification of source species for metagenomic sequences. Genome size is estimated as the sum of contigs for fragment assemblies without read pairs, or sum of scaffold spans for assemblies with read pairs. Only contigs submitted to NCBI (≥ 500 bp and containing at least two reads) were used in genome size estimation;

(2) >90% of bases at greater than 5x read coverage to give assurance of high base quality in the consensus sequence (this allows base quality to be assessed independent of sequencing platform as use of quality values may be inconsistent between platforms);

(3) >5 kb contig N50 length to ensure long enough contiguous sequences so the most genes are intact;

(4) >20 kb scaffold N50 length to ensure long enough scaffolds to capture large operons;

(5) Average contig length > 5kb to provide uniformity throughout the assembly, i.e., assembly is not a few large contigs and many small ones.

(6) >90% of "core genes" present in the gene list, to ensure completeness. The core genes comprise single copy genes conserved among all sequenced genomes in the super kingdom Bacteria. A similar set of core genes for Archaea was derived (4);

The provisional high-quality draft standards have operational definitions so they can be reliably measured in novel genomes that have no reference for comparison. The standards define when a high-quality draft genome sequence is completed and released into the public domain. The genomes that are produced by the HMP Consortium exceed these standards with the exception of some genomes produced before standards were in place (Table 1). More stringent metrics (N75 and N90 for contig and scaffold continuity) are presented, and nearly all genomes satisfy these higher standards. It is not known how many "difficult" genomes may be encountered in the future, and whether some of the standards will need modification, especially as the genomes of uncultured organisms are sequenced.

Standards for upgraded genome sequences. The Jumpstart Centers arrived at definitions for a series of genome sequence grades that capture different levels of improvement of high-quality draft genomes. HMP standards are aligned with those recently generated by a multi-center group (13), but include a higher level of detail in support of HMP scientific goals. The grades, Improved High Quality Draft, Annotation-directed finishing, Non-contiguous Finished, and Finished (see below), are provisional as they are based on a limited set of improved genomes. Like high-quality draft standards, the definitions are independent of sequencing platform and assembly software.

Improved High-Quality Draft. A sequence grade characterized by automated or manual work involving manipulation of existing shotgun data or addition of automated directed reads. With minimal work per genome this standard may be applied to a wide subset of the HMP reference genome collection. Unclosed areas require no annotation. HMP genomes with this designation will exhibit a minimum 50 kb contig N50 and are free of N base calls.

Annotation-Directed Improvement. Finishing work is targeted to clearly defined areas identified by an automated annotation pipeline. A coordinate key is included with the submission describing boundaries of finished vs. draft sequence. Such annotation includes information regarding improved areas not meeting finished standards. Assemblies subjected to Annotation-grade will exhibit a 50 kb minimum contig N50 and will carry a representational full-length or attempted full-length 16S rRNA copy. These genomes will be subject to a second automated annotation after improvement is complete to confirm improvement in quality of gene content.

Noncontiguous Finished. This intermediate finished sequence level reflects the comparative grade finished sequence previously applied to BACs (14). Non-contiguous finished sequence will be subject to manual closure approaches for all sequence problems. Minimal effort, however, will be expended on areas of low complexity. Full annotation of any areas not meeting finished standard is required. HMP non-contiguous finished assemblies are limited to a maximum of 3 scaffolds/Mb, must cover 97% of the captured genome, require identification and processing of bacterial plasmids and contain one finished 16S rRNA gene. Base quality is expected at finished quality unless otherwise noted, including removal of low confidence data at contig ends, and resolution of ambiguous bases and potentially misassembled regions. It is expected that most finished HMP genomes will fall into this category.

Finished. Finished reflects the traditional understanding of finished sequence, where the genome is completely deciphered along with extra-chromosomal elements such as plasmids. Consensus quality is upgraded to 10^{-5} maximum error rate. The assembly is expected to be free of misassembly and has been subjected to a QC review after completion. Any exceptions to completely finished sequence are noted with the submission.

Selection of genomes for upgrading. As mentioned above, the GCs are upgrading ~15% of the draft genome sequences for the HMP. The biological justifications for upgrading a strain's genome sequence as agreed upon by the Jumpstart Centers include: (1) species representing a novel phylogenetic lineage; (2) species with established clinical significance as causative or involved in disease; (3) species that may have a closely related previously sequenced genome, but showing significant intraspecies diversity and/or significant variation in phenotypic profiles; (4) genomes not attaining quality definition status with automated methods; (5) species that have documented abundance (dominance) in a body site; (6) duplicate species that are found in different body sites (e.g. isolates that are found both on skin and in the vaginal tract); (7) species that present an opportunity to explore pan-genomes.

Annotation pipeline and standards. The GCs' gene annotation process includes generating both *ab initio* and evidence-based (BLAST) predictions using one or more gene finding algorithms (Glimmer (15), GeneMark (16), and/or Metagene (17)). Loci are then defined by clustering predictions with the same reading frame. The best prediction at each locus is selected by evaluating all predictions against the best evidence (non-redundant, NR and Pfam) and resolving overlaps between adjacent coding genes as well as non-coding features such as tRNAs and rRNAs. The same evidence is used to determine and agree upon minimal gene lengths for genes with (60 bp) and without (120 bp) evidence, and is a guide in allowing shorter gene predictions. The centers established a hierarchical set of criteria, using homologies to the Pfam (18) and NR/GenBank entries, to resolve complex loci with overlapping predictions, on both strands or in the same strand, resulting in the deletion of spurious predictions and selection of the best gene model. Details on the individual gene prediction pipelines are provided by the Annotation WG as a set of Standard Operating Procedures (SOPs) and are available on the DACC (6).

Evaluation of the GCs' automated annotation pipelines. We evaluated the performance of the automated annotation pipelines using three finished and annotated genomes from GenBank as the gold standard: *E. coli* MG1655; *S. aureus* USA300; and *R. sphaeroides* 2.4.1. Each reference genome provides unique annotation challenges due to varied GC content, genome size, and prevalence of closely related sequences in public databases. The annotations generated by each of the GCs were compared with the reference GenBank annotations to identify false positives (FPs, the annotated gene models without corresponding GenBank annotations), and false negatives (FN, GenBank annotations with no corresponding GC-annotated genes). To assess the accuracy of draft genome annotation, the reference GenBank annotations from the finished genomes were mapped to the draft genomes, and the contiguous regions of the genome spanning complete reference genes were used for evaluation.

The automated gene identification accuracy for the three complete genomes ranged from

1.4-3.7% FN and 4.6-12.5% FP across centers, indicating that few known genes were missed by the automated annotation pipelines. The FP rate may be artificially high since several such predicted genes have matches to Pfam domains and may correspond to real genes missing from the original reference GenBank annotations. The draft genomes of varied sequence coverage and assembly quality differ in their approximation of the corresponding complete genome sequences. The higher the draft genome coverage and assembly quality (e.g., contig N50 value), the better the approximation to the complete genome as evidenced by the number of reference genes that map to the draft assemblies and by their representation as complete vs. fragmented genes. The accuracy of the draft genome annotation is also affected by the quality of the assembly. Although the FN rates remain low, the number of FP tends to increase with lower assembly quality. Hence, efforts to improve upon the draft genome assemblies should result in more accurate gene identifications

The GCs evaluated gene name assignments for the three reference genomes from each center's automated pipelines. There were similarities and differences in gene nomenclature due to variations in methodologies and type of underlying evidence used by each pipeline. Additionally, there were differences in annotation data types assigned to proteins by the centers. To achieve consistency and completeness in the functional annotation of reference genomes, the GCs and the DACC are working together to develop a standard set of SOPs. Upon implementation, the methods will become integral parts of each center's pipeline and will assign uniform data types to include descriptive protein name, gene symbol, Enzyme Commission (EC) number, functional role categories (a set of terms pertaining to general activities in the cell) and Gene Ontology (GO) terms to each predicted protein. The GCs will continue to evaluate and integrate new methods and trusted evidence sources into their pipelines as they become available. The pipeline modification will be recorded in updated SOPs.

Calculation of draft genome metrics. The draft genome metrics are a set of predefined quality measures that were used to assess the sequencing, assembly, and annotation of each reference strain's genome. The draft genome metrics were computed using a fully automated process that relies on the public sequence and annotation data that are available for each reference strain in GenBank. The requisite sequence and annotation data were downloaded from GenBank via NCBI's Entrez Utilities ("EUtils") web API and processed at the DACC, to ensure that only data guaranteed to be available to the wider research community were used in the evaluation and that subsequent reporting and the possibility of sequencing center-specific bias was minimized.

The draft metrics were computed and reported by a pair of custom Perl programs: computeProjectStats.pl, which downloads the contigs, scaffolds, and annotation for a single reference strain genome sequencing project and computes the draft metrics; and summarizeResults.pl, which reads the output of one or more runs of the former program, and generates a set of tabular plain text and HTML summary reports. The cells in the HTML tables that correspond to the tested metrics are color-coded green for those metrics that meet or exceed the defined standard, and red for those that do not, allowing a relatively large number of reference strains to be scanned rapidly.

The following statistics/metrics were computed and reported by the programs. The tested metrics that were used as quality controls are those with the "passing" threshold listed in square brackets after the metric. Note that not every metric could be computed for every reference strain:

1. Number of contigs

2. Number of contigs of length < 5kb

3. Average contig length [pass if >5kb]

4. Percentage of the total genome sequence in contigs [pass if >= 90%] (*)

5. Contig N50, N75, and N90 [pass if > 5kb] (*)

6. Number of scaffolds

7. Average scaffold length

8. Scaffold N50, N75, and N90 [pass if >20kb] (*)

9. Number of sequencing gaps (**)

10. Number of sequencing gaps per 5kb [pass if < 1] (***)

11. Number and percentage of "core genes" matched

12. Number and percentage of "core gene groups" matched [pass if >= 90%]

(*) Note that the total genome sequence size is approximated by simply summing the lengths of the available sequence scaffolds or contigs.

(**) For strains with scaffolds only.

(***) Since the gaps are represented in the scaffold sequences as stretches of Ns, the denominator in this fraction includes the gaps themselves.

Phylogenetic analyses. Phylogenetic trees were created using PAUP (19). In order to reduce the dataset of 16S rRNA gene sequences from ~7000 sequences to ~1500, only one representative per genus was chosen. The alignments of these sequences were extracted from the SILVA database (20), the 16S rRNA sequences from our sequenced genomes were aligned to these sequences using ARB (21) and a bootstrapped neighbor-joining tree was created. As bootstrapped PAUP trees lack meaningful branch lengths, these were estimated by loading the resulting PAUP tree into TREE-PUZZLE (22) as a user defined tree.

Selection of sets of Bacterial and Archaeal core genes. A set of the bacterial core genes was identified as follows: 12,087 proteins originating from four bacterial species (*E. coli* str. K-12 substr. MG1655, *R. sphaeroides* 2.4.1, *T. pallidum* subsp. *pallidum* str. Nichols and *S. aureus* subsp. *aureus*) were clustered using OrthoMCL with default parameters (23). From the 290 groups of orthologous genes in all 4 organisms, 235 had a single copy gene per species. These genes were interrogated with a published (24) set of 111 single copy core gene groups, yielding 99 genes that were present in both data sets. Evaluation based on their presence in the 621 finished genomes in GenBank (May 2008) resulted in a final set of 66 core genes. The SOP is available (25).

Similarly, to build a set of Archaeal core genes, 112,992 proteins from 52 finished Archaeal genomes (16 Crenarchaeota, 34 Euryarchaeota, 1 Nanoarchaeon and 1 Korarchaeon; database built 05/13/2008) were clustered using OrthoMCL. OrthoMCL identified 11,410 orthologous groups from these sequences with default parameter settings. From these groups, 119 groups with 6,445 genes from all of the 52 species resulted in 84 single-copy orthologs (i.e. having only one gene from each species). This dataset was compared to a previous core set (26), which constructed 166 Archaeal Clusters of Orthologous Genes (arCOGs) from 41 genomes (13 Crenarchaeota, 27 Euryarchaeota and 1 Nanoarchaeon). Thirty-nine of these 41 genomes were included in our analysis. The refined core gene set resulted in 104 core groups. For more details on both analyses, refer to the DACC (27).

Pan-genome analysis.

Genome sequences. The annotated protein and DNA sequences for all publicly available genomes of *B. longum*, *E. faecalis*, *L. reuteri* and *S. aureus* were obtained from the NCBI Web site on May 23, 2009.

Pan-genome calculations. All versus all comparisons within each species set were performed using BLASTP and tBLASTN. Core genome and pangenome extrapolation was performed as described (3).

Average nucleotide identity vs. base composition analysis.

Genome sequences. The annotated protein and DNA sequences for all publicly available genomes of *Bacteroides*, *Bifidobacterium*, *Lactobacillus*, and *Clostridium* were obtained from the NCBI Web site on May 23, 2009.

Genetic versus ecological diversity between genomes. For each pair of genomes within a genus, the sequence-based evolutionary distance and the percentage shared gene content are calculated. Both values are based on the set of orthologous genes between two genomes, determined by a reciprocal best BLAST hit approach, as previously described (28). Only reciprocal hits between two proteins that share at least 30% amino acid sequence identity within 70% or more of the whole gene are considered as orthologous. The number of genes for which an ortholog can be found divided by the total number of genes within a genome, is the percentage of shared gene content. The sequence-based evolutionary distance, also referred to as the average nucleotide identity (ANI), is calculated by averaging all nucleotide sequence similarity values over all orthologous gene pairs between two genomes.

Identification of novel genes. To identify novel polypeptides, a total of 547,968 predicted polypeptides were extracted from the 178 reference strains, corresponding to the entire annotated gene complement of these strains. These were searched against the bacterial and viral divisions of NCBI's nonredundant protein database (nr) using WU-BLASTP. Polypeptides from 165 strains were downloaded using NCBI's EUtils service and those from the remaining 13 strains were downloaded manually through the Entrez web interface. The May 18, 2009 release of the NCBI non-redundant protein database, (NR), was downloaded and all of the sequences from the BCT, PHG, SYN, UNA, VRL, and ENV divisions were extracted and formatted into a BLAST-searchable database. This subset of nr contained 5,562,464 of the 8,865,153 nr sequences. Each reference polypeptide was searched against the NR subset using WU-BLASTP 2.0 and the following command-line options: -E 1e-5 -matrix BLOSUM62 -wordmask seg -B 150 -V 150 -gspmax 5 -shortqueryok -novalidctxok -cpus 1. The BLAST output was subsequently parsed to determine which hits were to "reference strain" nr entries (i.e., those NR entries entirely composed of sequences drawn from the 178 reference strains themselves). Each polypeptide was also run against a merged HMM database of TIGRFAM and Pfam HMMs using version 2a of the HMMER3 package, with the --cut_ga option. A set of candidate "novel" polypeptides was defined by selecting those that met both of the following criteria: 1) Unmasked sequence length > 100 amino acids and 2) no BLASTP match to any NON-reference entry in the nr subset with E-value < 1e-10. A second dataset of 747,522 polypeptides was extracted from 178 randomly selected bacterial and archaeal genomes, which were not in the set of HMP reference strains and contained at least 1500 annotated gene predictions each. This dataset was searched using the same methods for comparison.

Analysis of metagenomic shotgun data. A one percent subset of the reads from each metagenomic dataset was aligned to all the publically available complete and draft microbial genomes. After adjusting the depth coverage to account for sub-sampling a genome was used for recruitment if any individual metagenomic dataset or if all the metagenomic datasets combined would be projected to have more than one-fold coverage. Recruitment was then carried out as described (5) with the additional criterion that 90% of a 454 read had to align to a reference genome to be considered recruited.

Figure S1. ***E. coli*** **metrics compared to assembly coverage.** The quality metrics are shown for assemblies of the *E. coli* genome at sequence coverage from 1x to 16x. Sequence divergence from the finished reference is shown in (A). Blue indicates substitutions, red indicates deletions and yellow indicates insertions in the draft assembly, green is the sum of all sequence errors. Very few differences are seen by 10x coverage. The Contig N50 is shown in (B). The amount of the finished genome covered by the draft assembly is shown in (C). The number of gaps remaining in the draft assembly is shown in (D).

Figure S2. Coding Sequences (CDS) in the ***E. coli*** **draft assemblies.** The quality of the *E. coli* draft genome assemblies was assessed by counting the number of CDS that were mapped or missed. Assemblies were investigated at 5x, 6x, 9x, 10x, and 16x sequence coverage and assembled using different assembly algorithms at three Genome Centers. (A) shows the percentage of CDSs that were not mapped to the different assemblies. The breakdown of theses CDSs by class is shown in (B), where large CDSs are shown in green, known gene CDSs in gold, and the core gene set in light blue. The reasons for the missed CDSs are shown in (C) where small proteins are shown in lavender, frameshift induced changes to the CDS shown in red, indels non-modulo 3 in yellow, and gaps in light blue.

Figure S3. ***Lactobacillus reuteri*** **pangenome analysis.** Each circle represents the number of new genes (A) or shared genes (B) identified as new genomes are sequentially added to the analysis. Colored diamonds indicate medians, and means are colored triangles. The exponential fit was plotted on the median points on a linear scale. (A) Strain-specific gene analysis. The exponential fit is shown by a blue line ($y = 3.57 + 20495.28e^{(-x/0.45)}$). The extrapolated

average number of strain-specific genes is shown by the dotted line. (B) Core genome analysis. The exponential fit is shown by the green line (y = 1602.42 + 442.98e^(−x /2.48)). The extrapolated core genome size is shown by the dotted line. The inset graph represents the *L. reuteri* pangenome analysis, with the red curve showing the calculated pangenome size (y = 3381.56 + −1263.57e^(−x/9.55)). The dotted line shows the extrapolated number of genes at which point the pangenome would approach a closed pangenome model.

Figure S4. *Bifidobacterium longum* pangenome analysis. Each circle represents the number of new genes (A) or shared genes (B) identified as new genomes are sequentially added to the analysis. Colored diamonds indicate medians, and means are colored triangles. The exponential fit was plotted on the median points on a linear scale. (A) Strain-specific gene analysis. The exponential fit is shown by a blue line (y = 110.77 + 1739.43e^(−x/0.93)). The extrapolated average number of strain-specific genes is shown by the dotted line. (B) Core genome analysis. The exponential fit is shown by the green line (y = 1433.42 + 906.26e^(−x /1.36)). The extrapolated core genome size is shown by the dotted line. The inset graph represents the *B. longum* pangenome analysis, with the red curve showing the calculated pangenome size (y = 3372 + −3089.62e^(−x /1.79)). The dotted line shows the extrapolated number of genes at which point the pangenome would approach a closed pangenome model.

Figure S5. *Enterococcus faecalis* pangenome analysis. Each circle represents the number of new genes (A) or shared genes (B) identified as new genomes are sequentially added to the analysis. Colored diamonds indicate medians, and means are colored triangles. The exponential fit was plotted on the median points on a linear scale. (A) Strain-specific gene analysis. The

exponential fit is shown by a blue line (y = 99.01 + 574.74e^(−x/1.72)). The extrapolated average number of strain-specific genes is shown by the dotted line. (B) Cure genome analysis. The exponential fit is shown by the green line (y = 1794.37 + 1121.88e^(−x /13.98)). The extrapolated core genome size is shown by the dotted line. The inset graph represents the *E. faecalis* pangenome analysis, with the red curve showing the calculated pangenome size (y = 4079.66 + −1242.99e^(−x /2.65)). The dotted line shows the extrapolated number of genes at which point the pangenome would approach a closed pangenome model.

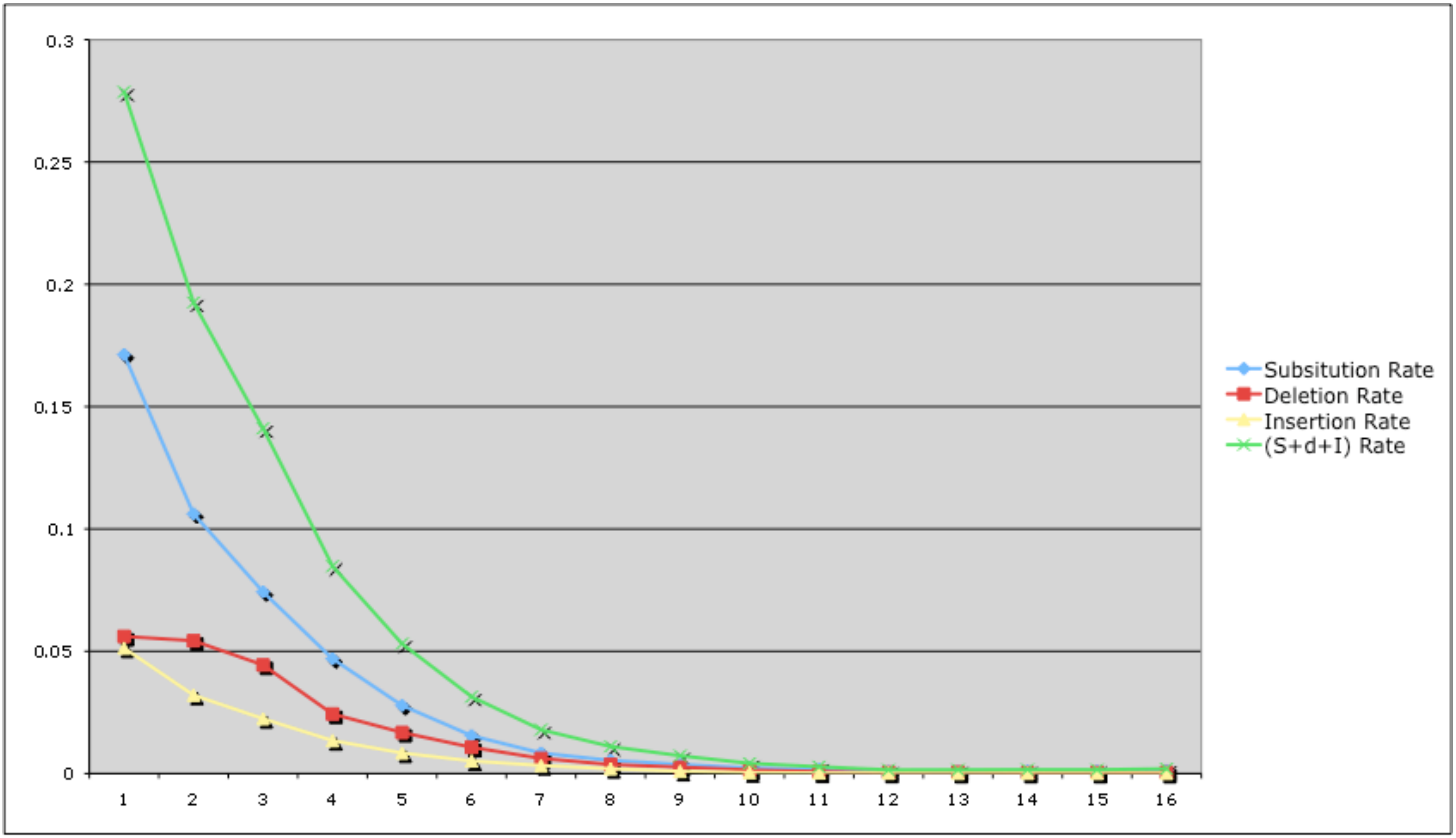
Figure S6. *Staphylococcus aureus* pan-genome analysis. Each circle represents the number of new genes (A) or shared genes (B) identified as new genomes are sequentially added to the analysis. Colored diamonds indicate medians, and means are colored triangles. The exponential fit was plotted on the median points on a linear scale. (A) Strain-specific gene analysis. The exponential fit is shown by a blue line (y = 7.92 + 385.47e^(−x/ 1.96)). The extrapolated average number of strain-specific genes is shown by the dotted line. (B) Core genome analysis. The exponential fit is shown by the green line (y = 2295.16 + 334.99e^(−x/ 4.65)). The extrapolated core genome size is shown by the dotted line. The inset graph represents the *S. aureus* pangenome analysis, with the red curve showing the calculated pan-genome size (y = 3210.05 + −527.08e^(−x /5.6)), and the dotted line demonstrating that this dataset has nearly reached the extrapolated pan-genome size.

Figure S7. Inter-strain diversity among *Bifidobacterium* genomes. Each point represents a whole-genome comparison between two *Bifidobacterium* genomes and shows the percentage average nucleotide identity (ANI) on the x-axis as a measure of evolutionary distance, plotted

against the percentage of gene content similarity on the y-axis. Only comparisons with ANI values above 85% are shown. The horizontal line at 95% corresponds to a recommended cut-off of 70% DNA–DNA reassociation for species delineation. Different intra- and inter-species comparisons are color-coded, with full or open circles respectively, and labeled with given taxonomical name in corresponding color. Colored ovals assist in identifying related data points belonging to a single named species.

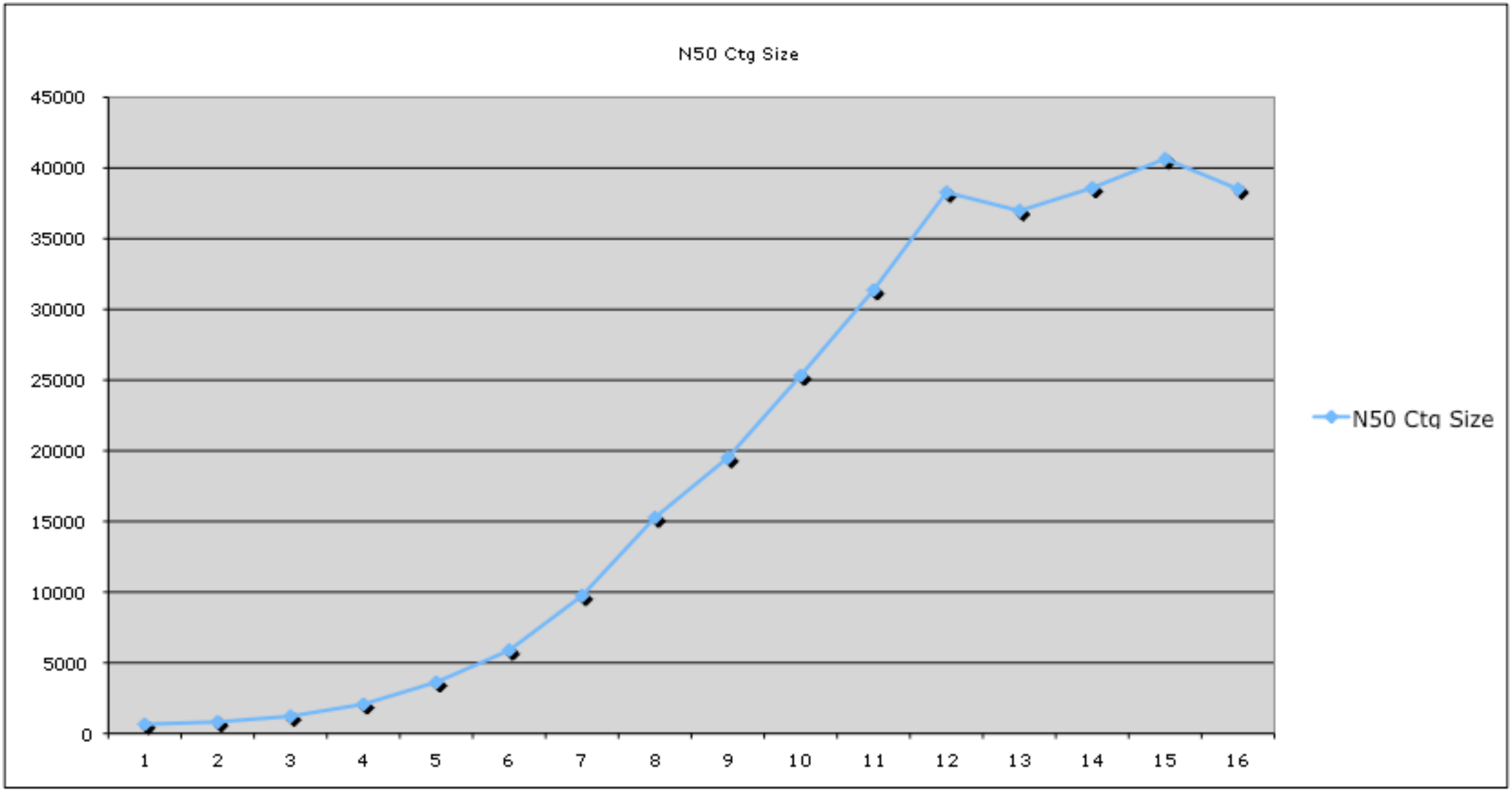
Figure S8. Inter-strain diversity among *Bacterioides* genomes. Each point represents a whole-genome comparison between two *Bacteroides* genomes and shows the percentage average nucleotide identity (ANI) on the x-axis as a measure of evolutionary distance, plotted against the percentage of gene content similarity on the y-axis. Only comparisons with ANI values above 85% are shown. The horizontal line at 95% corresponds to a recommended cut-off of 70% DNA–DNA reassociation for species delineation. Different intra- and inter-species comparisons are color-coded, with full or open circles respectively, and labeled with given taxonomical name in corresponding color. Colored ovals assist in identifying related data points belonging to a single named species.

Figure S1A. E. coli Error Rates vs Assembly



Depth of coverage used for assembly

Figure S1B. E. coli N50 Contig Size vs Assembly



Depth of coverage used for assembly

Figure S1C. Amount of E. coli genome covered vs Assembly

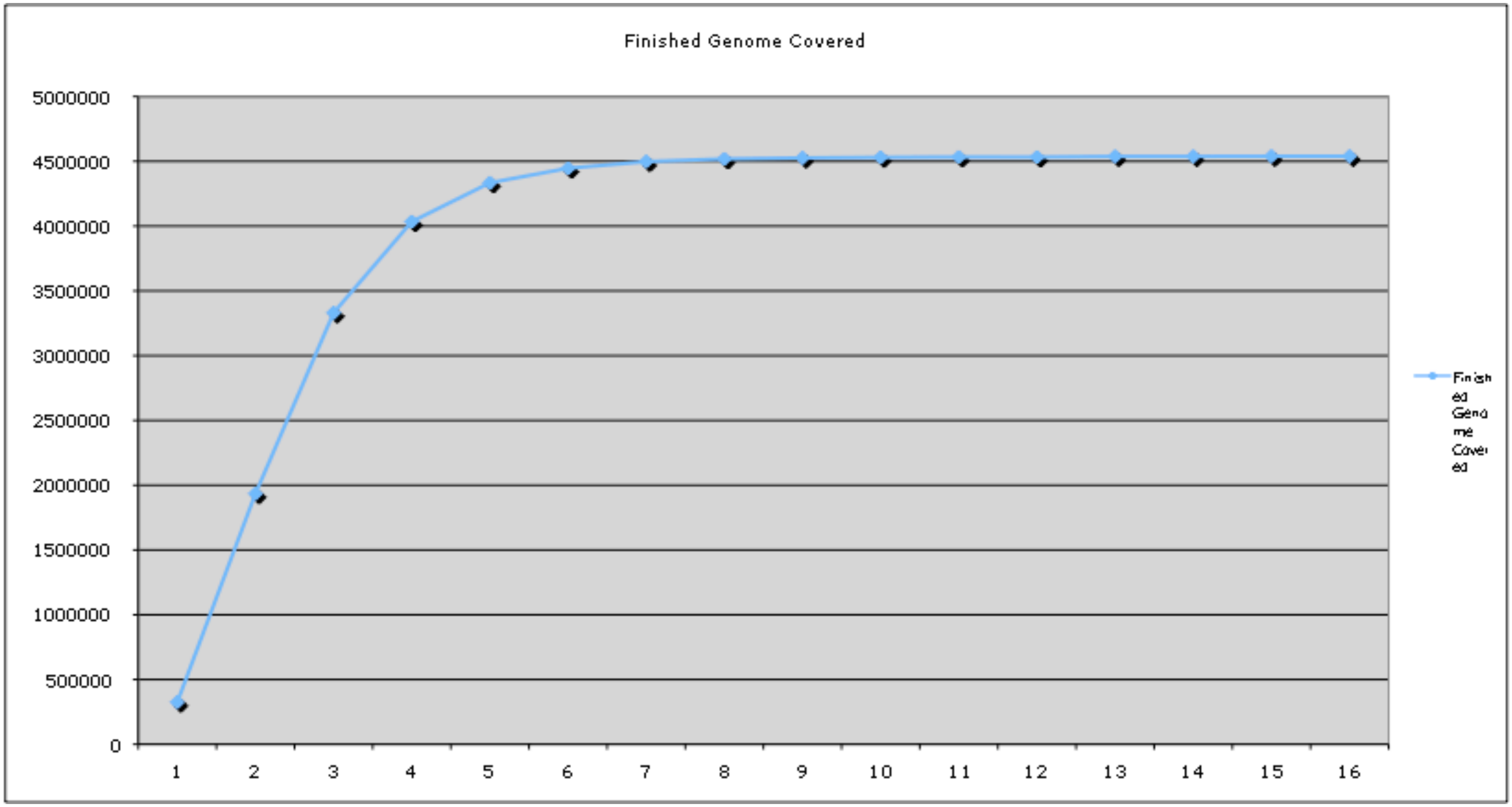


Figure S1D. Number of gaps in E. coli assembly vs Assembly

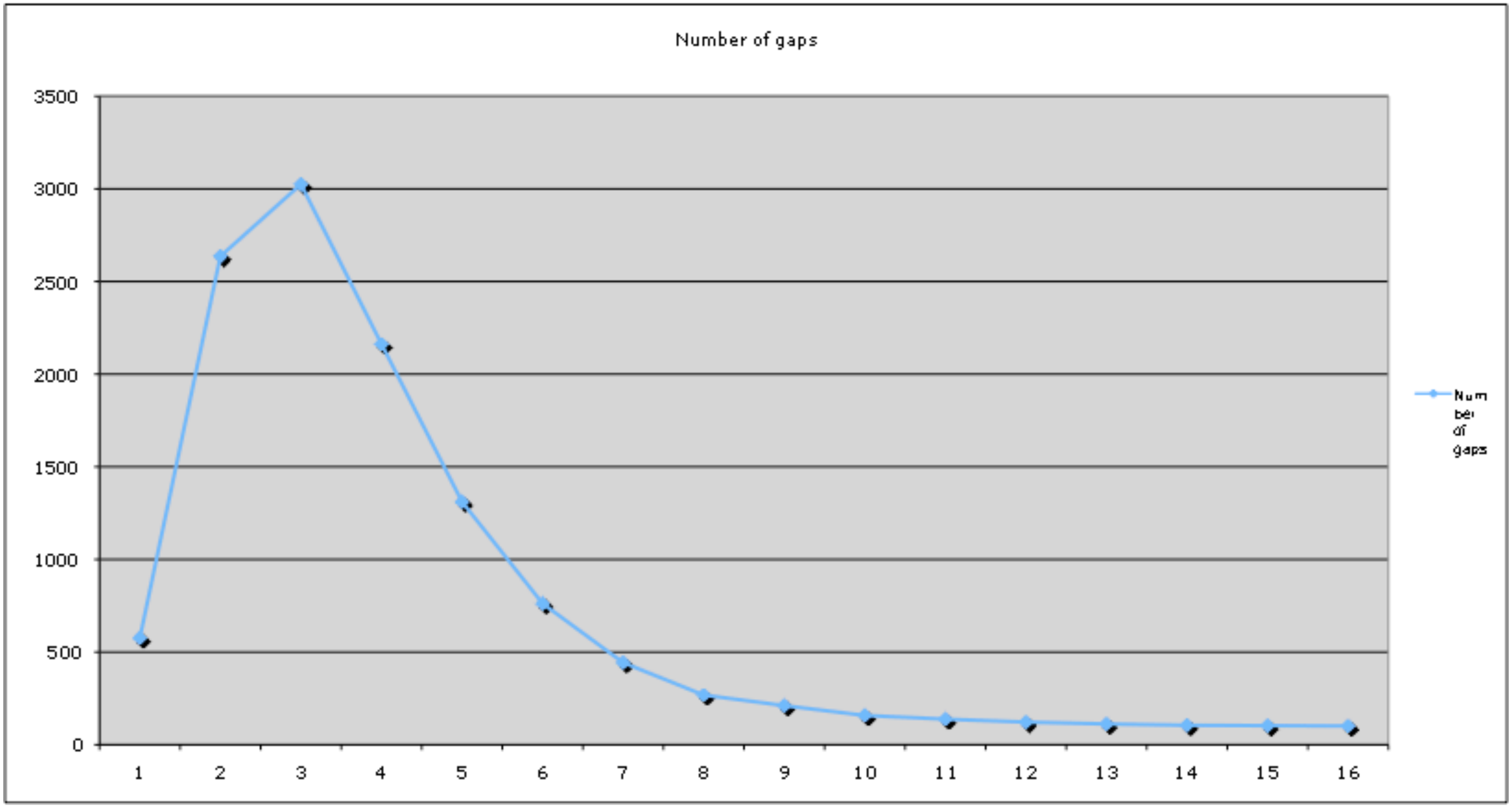


Figure S2A. Percent ORFs missing from different draft E.coli assemblies

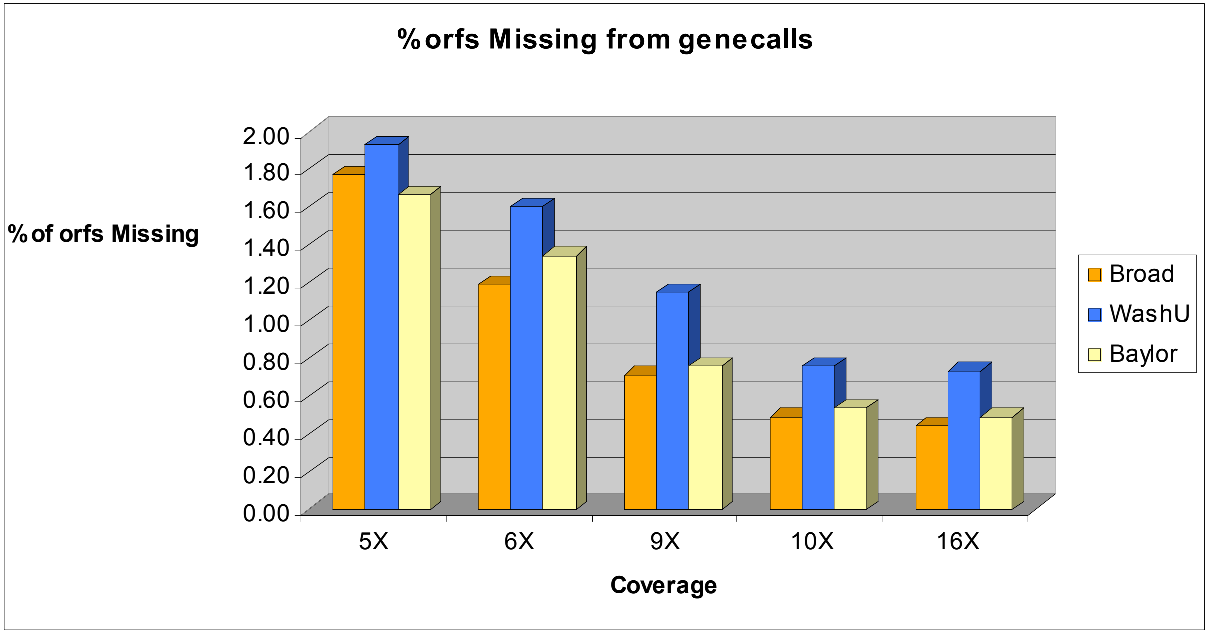


Figure S2B. Coverage of Mapped GenBank ORFs in Different Assemblies.

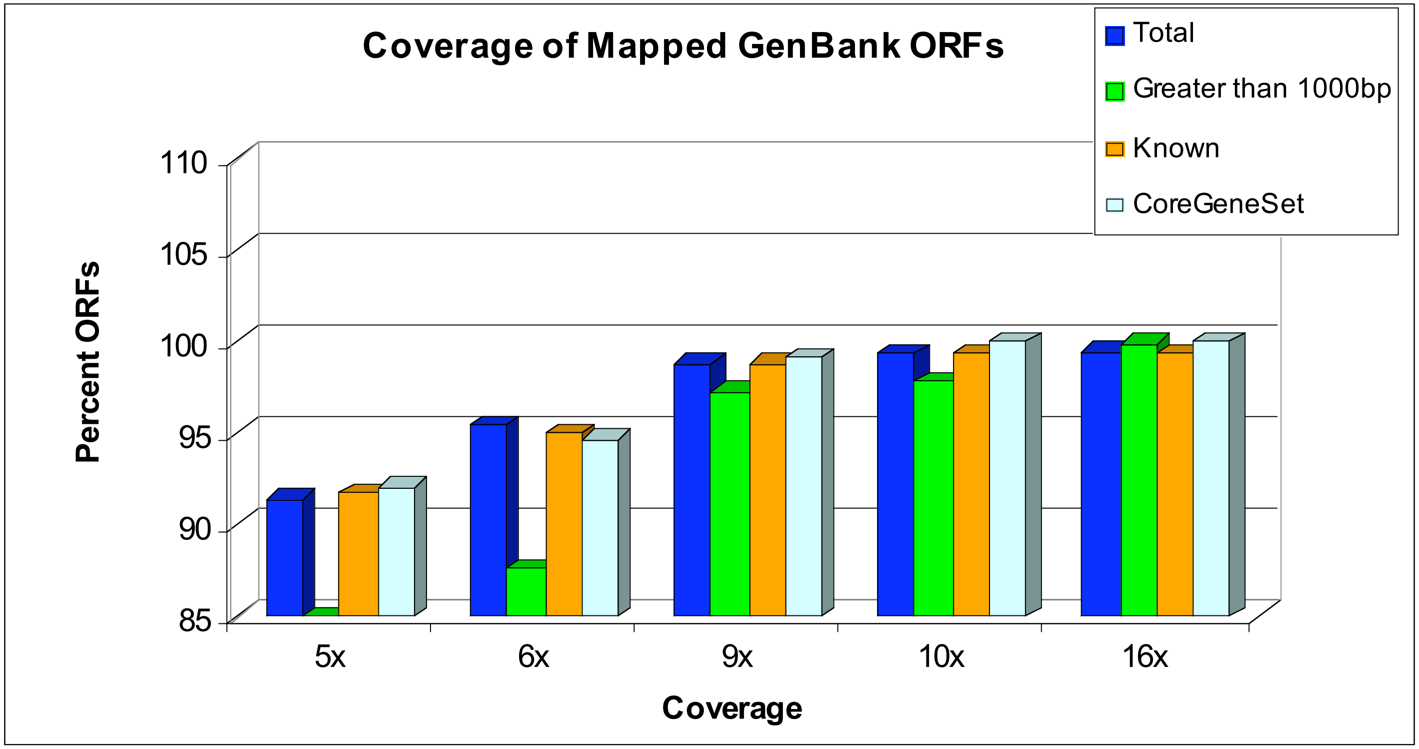
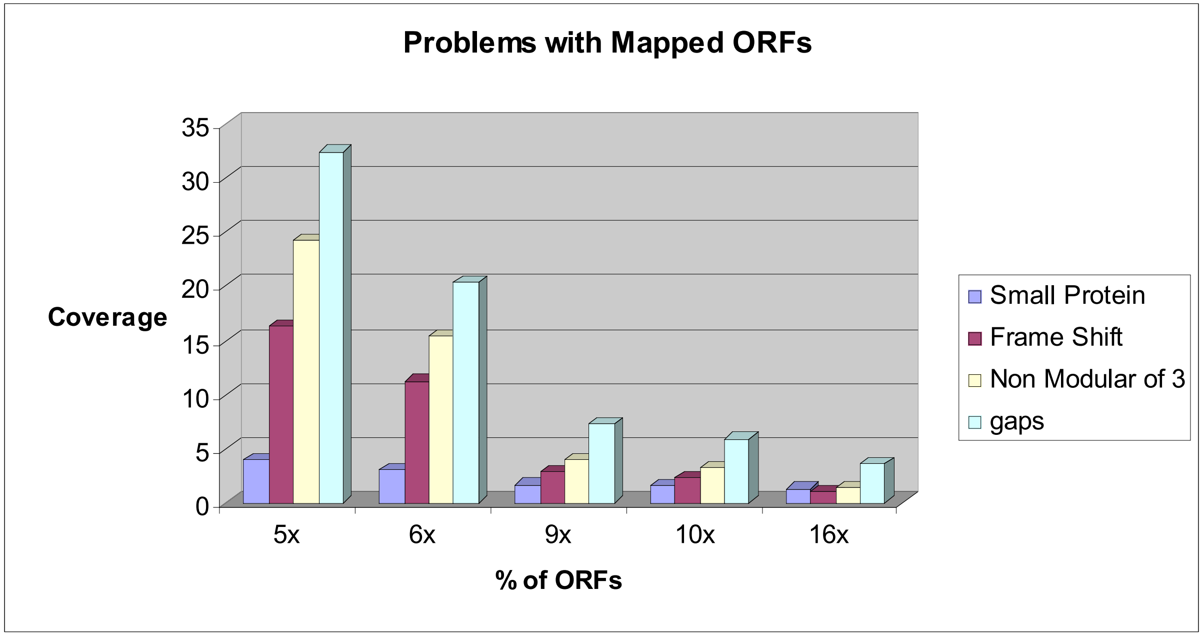
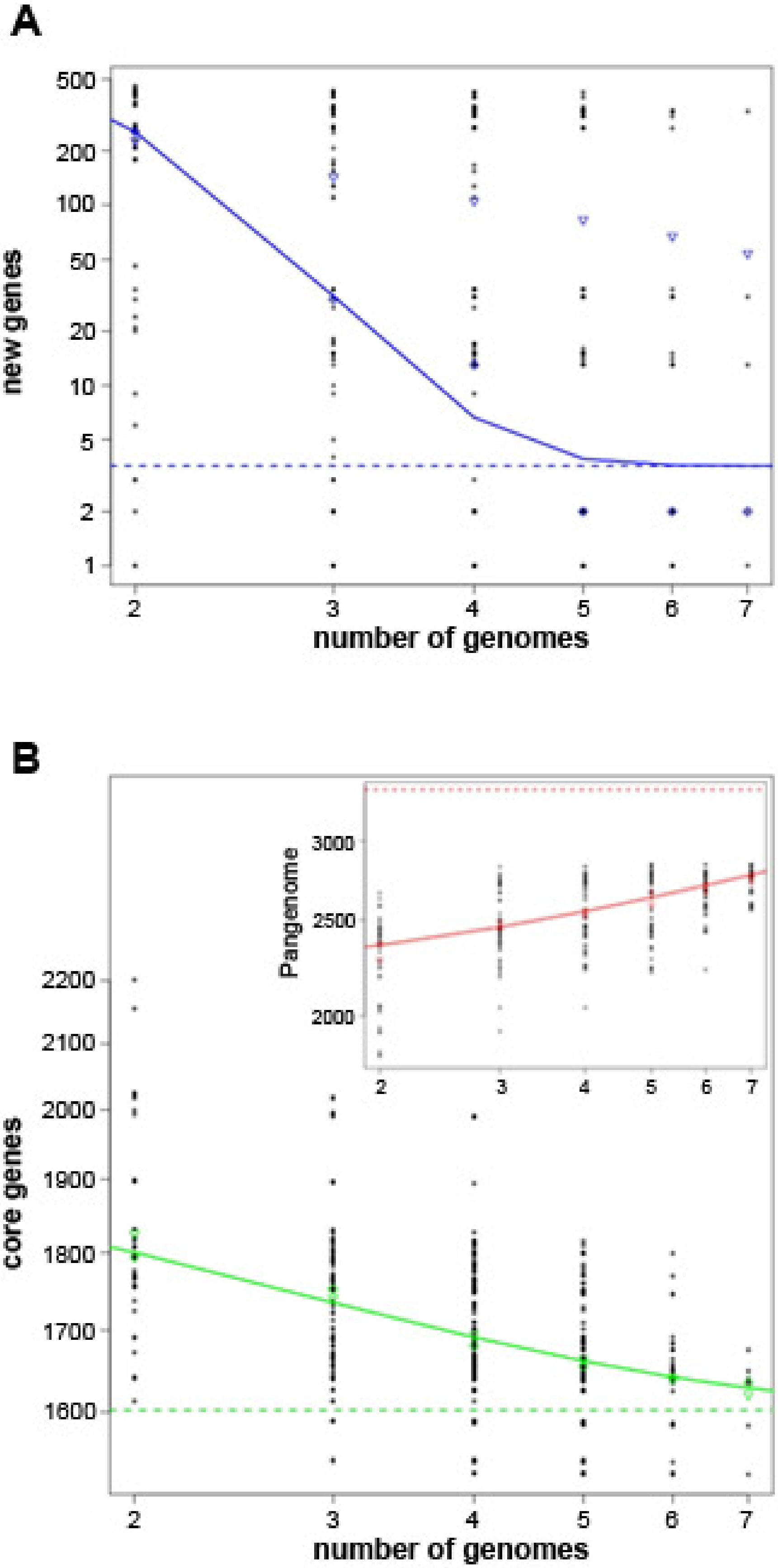
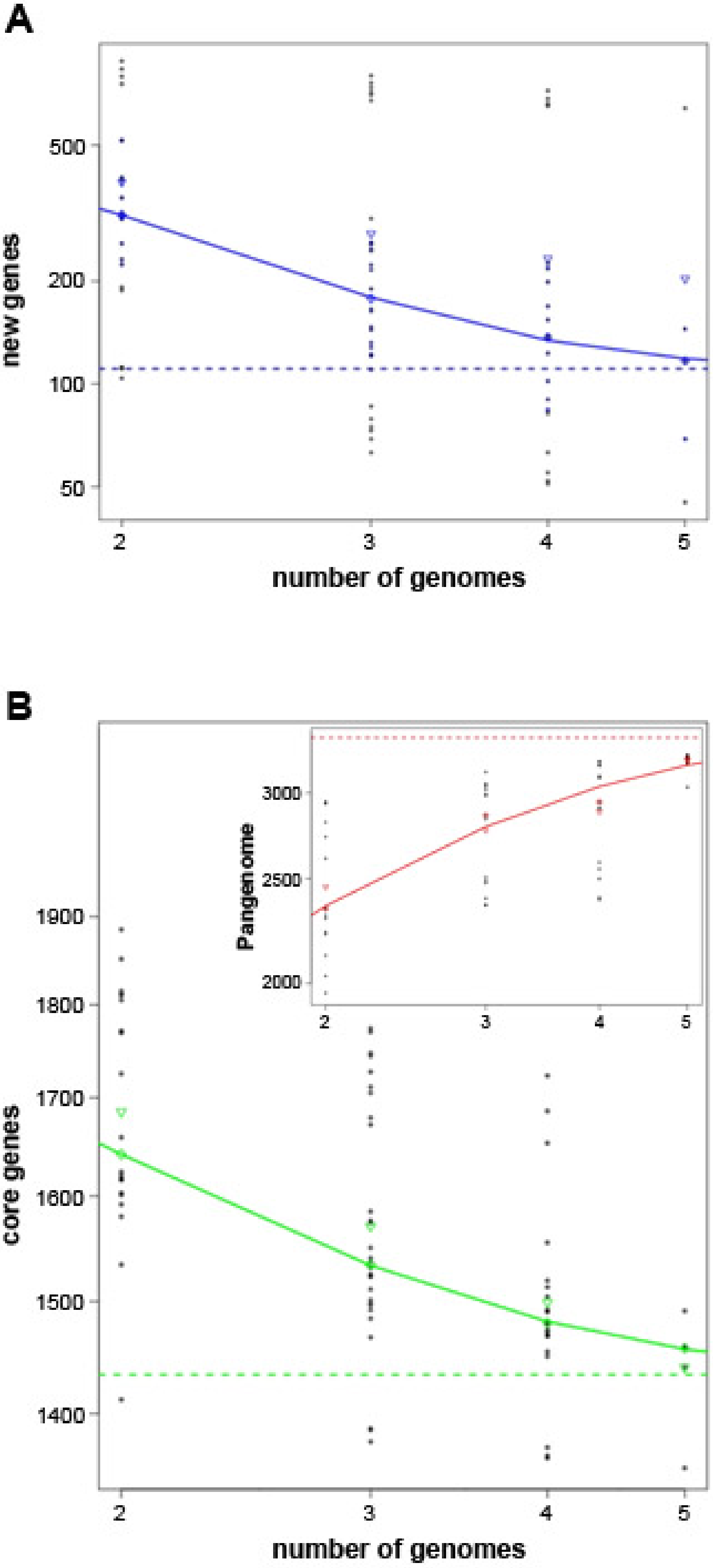
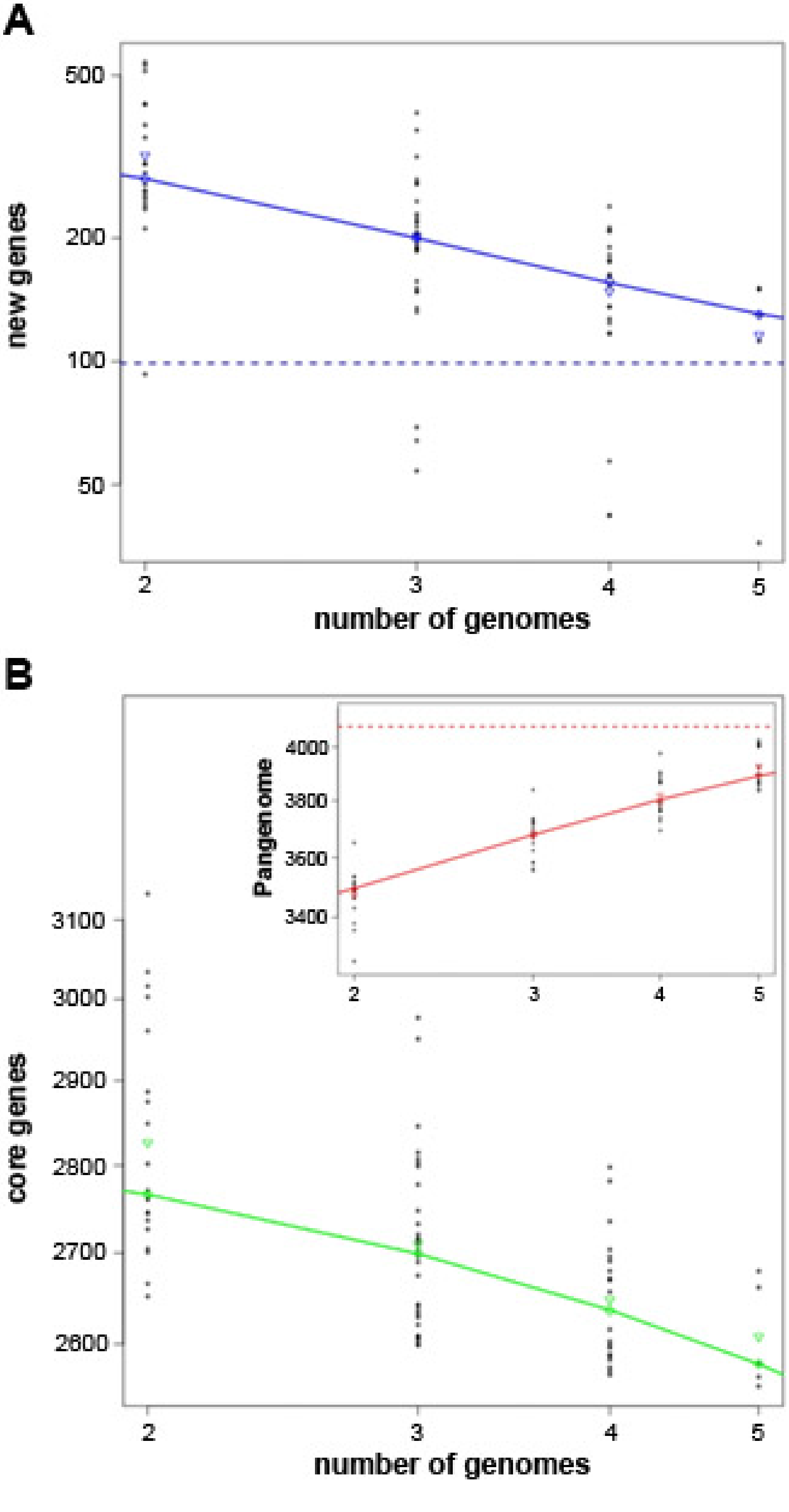


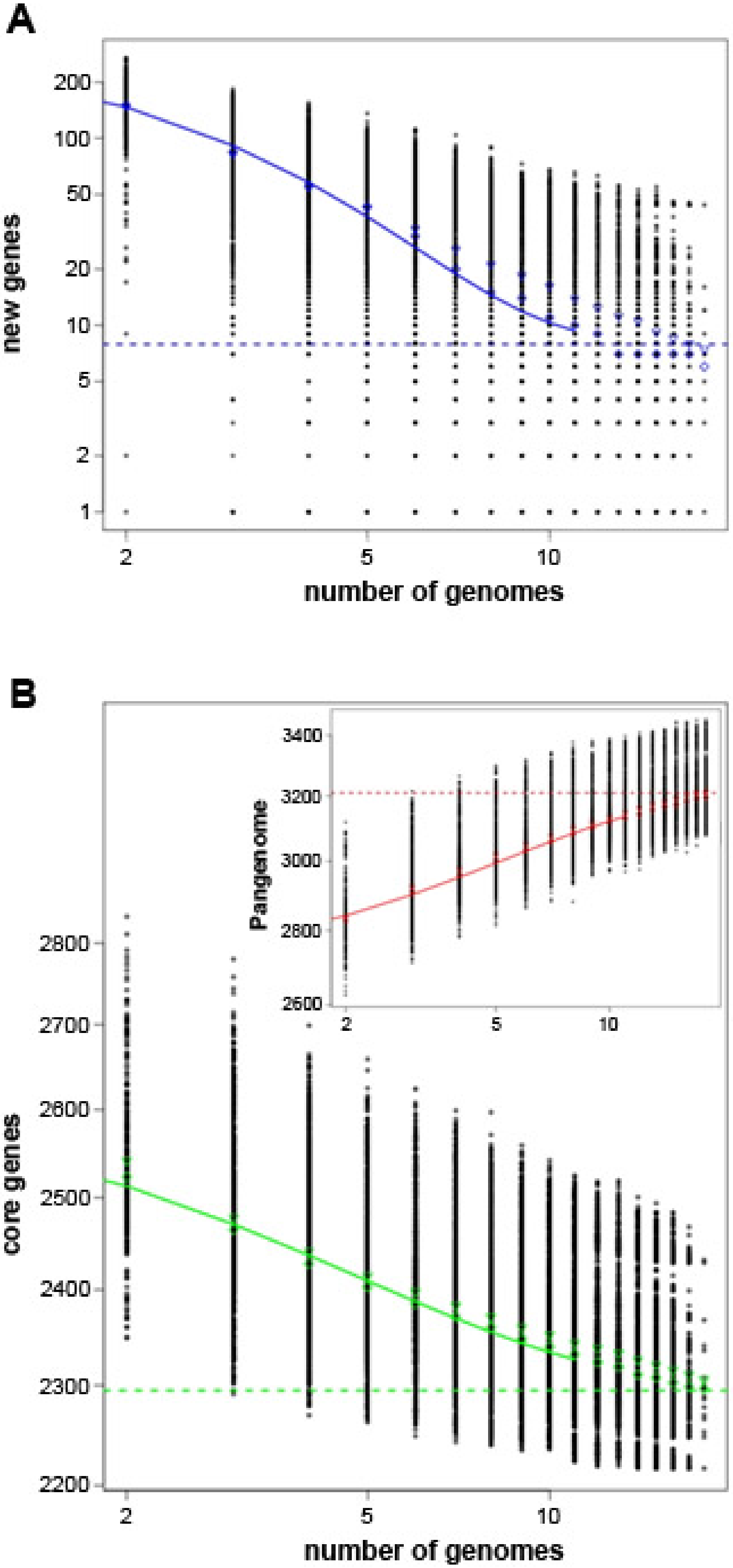
Figure S2C. Problems with Mapped Genbank ORFs on different coverage Assemblies

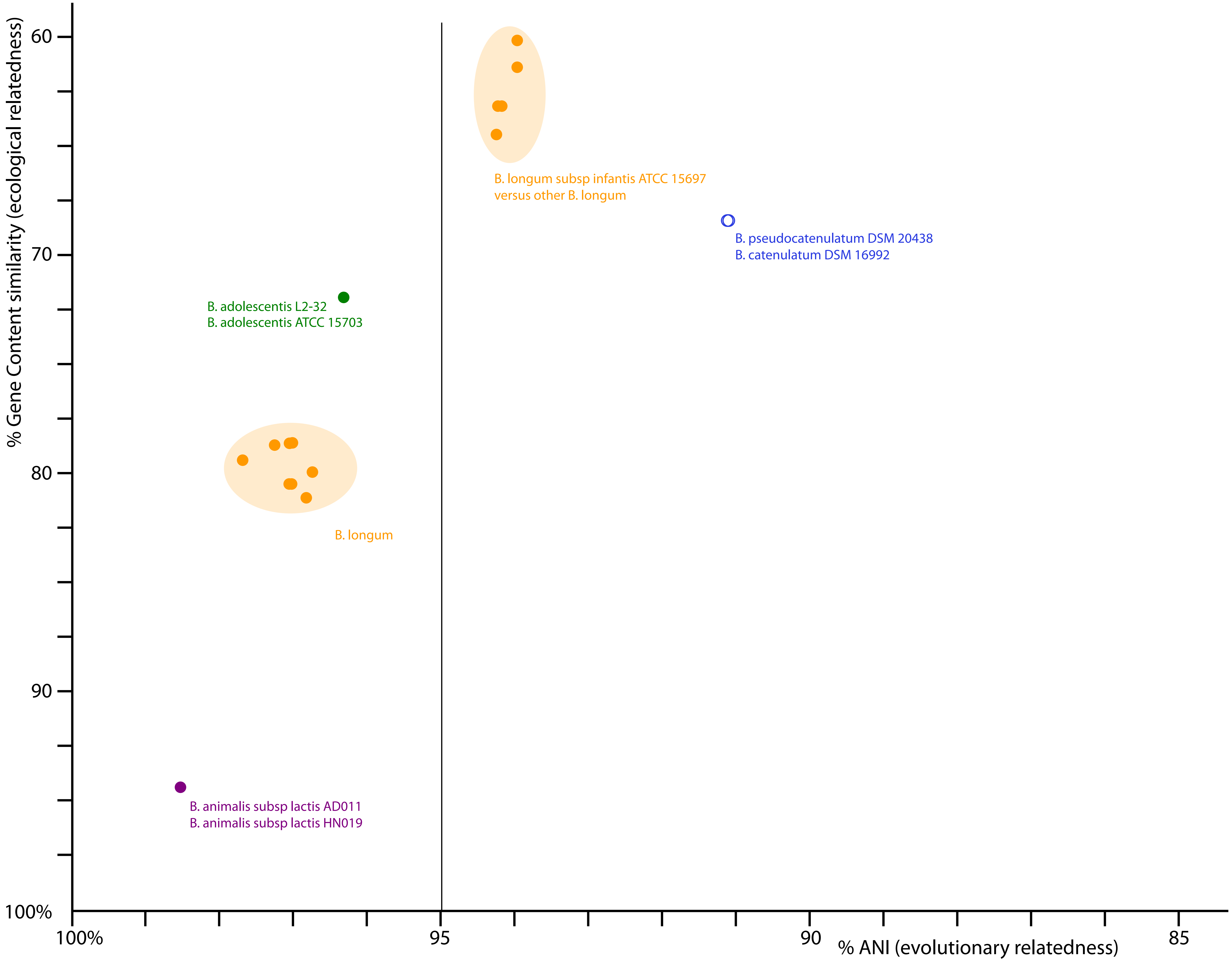


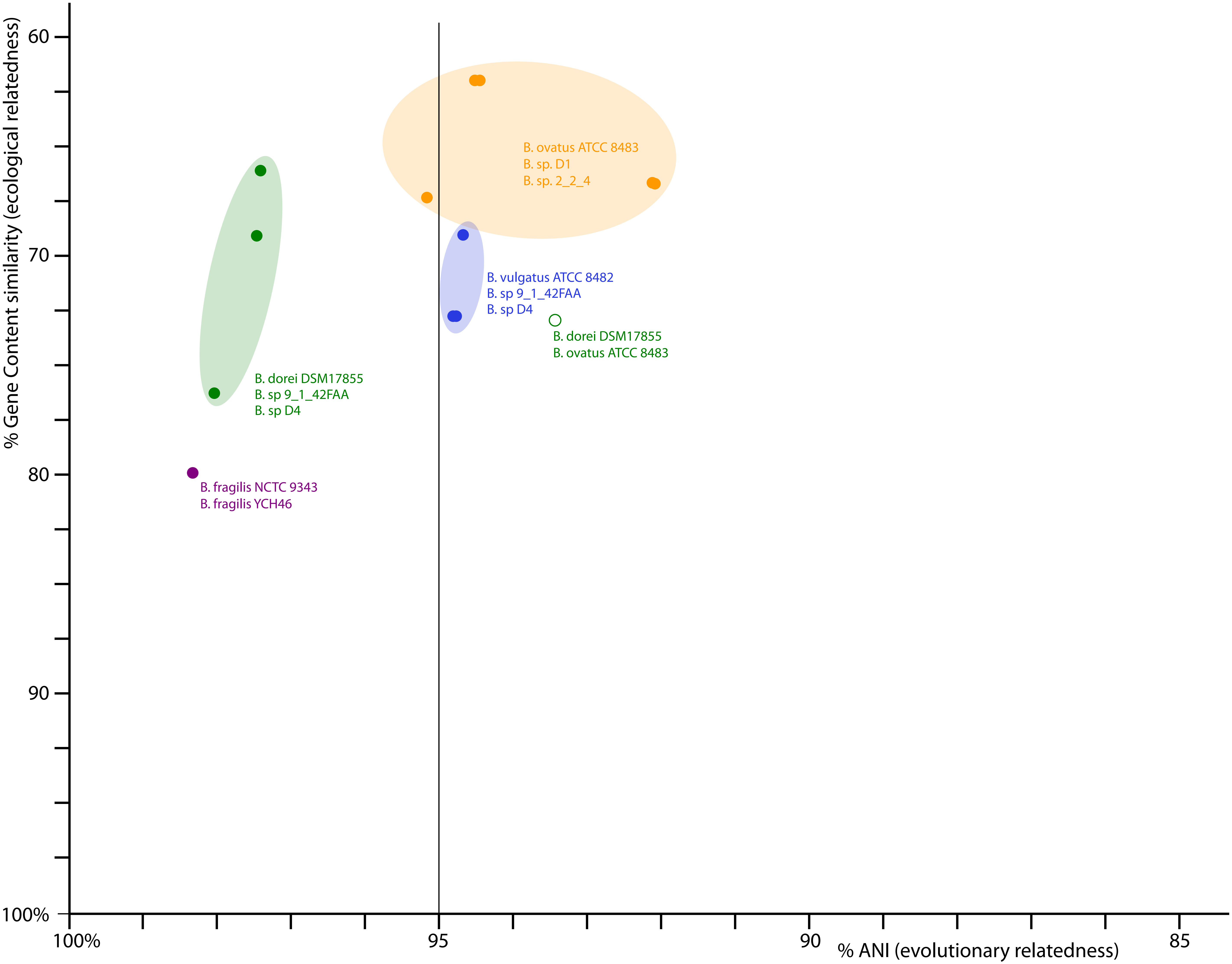

A
500
200
100
50
20
10
5
2
1
new genes
2
3
4
5
6
7
number of genomes
B
2200
2100
2000
1900
1800
1700
1600
core genes
Pangenome
3000
2500
2000
2
3
4
5
6
7
2
3
4
5
6
7
number of genomes


A
500
200
100
50
new genes
2
3
4
5
number of genomes
B
1900
1800
1700
1600
1500
1400
core genes
Pangenome
3000
2500
2000
2
3
4
5
2
3
4
5
number of genomes


A
500
200
100
50
new genes
2
3
4
5
number of genomes
B
4000
3800
3600
3400
Pangenome
2
3
4
5
3100
3000
2900
2800
2700
2600
core genes
2
3
4
5
number of genomes


A
new genes
200
100
50
20
10
5
2
1
2
5
10
number of genomes
B
core genes
2800
2700
2600
2500
2400
2300
2200
2
5
10
number of genomes
Pangenome
3400
3200
3000
2800
2600
2
5
10


% Gene Content similarity (ecological relatedness)
60
70
80
90
100%
100%
95
90
85
% ANI (evolutionary relatedness)
B. longum subsp infantis ATCC 15697
versus other B. longum
B. pseudocatenulatum DSM 20438
B. catenulatum DSM 16992
B. adolescentis L2-32
B. adolescentis ATCC 15703
B. longum
B. animalis subsp lactis AD011
B. animalis subsp lactis HN019


% Gene Content similarity (ecological relatedness)
60
70
80
90
100%
100%
95
90
85
% ANI (evolutionary relatedness)
B. ovatus ATCC 8483
B. sp. D1
B. sp. 2_2_4
B. vulgatus ATCC 8482
B. sp 9_1_42FAA
B. sp D4
B. dorei DSM17855
B. ovatus ATCC 8483
B. dorei DSM17855
B. sp 9_1_42FAA
B. sp D4
B. fragilis NCTC 9343
B. fragilis YCH46

References and Notes

1. "HMP Project Catalog," Human Microbiome Project Data Analysis Coordinating Center, (http://www.hmpdacc.org/project_catalog.html).
2. "Reference Genomes of the Human Microbiome Project," Human Microbiome Project Data Analysis Coordinating Center, (http://hmpdacc.org/reference_genomes.php).
3. H. Tettelin *et al.*, *Proc Natl Acad Sci U S A* **102**, 13950 (Sep 27, 2005).
4. Materials and methods are available as supporting material on Science Online.
5. D. B. Rusch *et al.*, *PLoS Biol* **5**, e77 (Mar, 2007).
6. "Documentation and SOPs," Human Microbiome Project Data Analysis Coordinating Center, (http://www.hmpdacc.org/sops.php).
7. Jeffrey I. Gordon, Richard Wilson, Elaine Mardis, Jian Xu, Claire M. Fraser, David A. Relman, "Extending Our View of Self: the Human Gut Microbiome Initiative (HGMI)," (http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/HGMISeq.pdf).
8. George Weinstock, Richard Wilson, Jeffrey Gordon, Sandra Clifton, Bruce Birren, Chad Nusbaum, "Pilot Project to Expand the Number of Sequences of Culturable Microbes from the Human Body," (www.genome.gov/Pages/Research/Sequencing/SeqProposals/HMPP_Proposal.pdf).
9. "Human Microbiome Data Analysis Coordiating Center Main Page," Human Microbiome Project Data Analysis Coordinating Center, (http://www.hmpdacc.org/).
10. D. Field *et al.*, *Nat Biotechnol* **26**, 541 (May, 2008).
11. "BEI resources Biodefense and Emerging Infections Research Resources Repository," (http://www.beiresources.org).
12. "Provisional Assembly Metrics," Human Microbiome Project Data Analysis Coordinating Center, (http://hmpdacc.org/doc/ProvisionalAssemblyMetrics.pdf).
13. F. Chain, C. Cote-Beaulieu, F. Belzile, J. G. Menzies, R. R. Belanger, *Mol Plant Microbe Interact* **22**, 1323 (Nov, 2009).
14. R. W. Blakesley *et al.*, *Genome Res* **14**, 2235 (Nov, 2004).
15. A. L. Delcher, D. Harmon, S. Kasif, O. White, S. L. Salzberg, *Nucleic Acids Res* **27**, 4636 (Dec 1, 1999).
16. M. Borodovsky, R. Mills, J. Besemer, A. Lomsadze, *Curr Protoc Bioinformatics* **Chapter 4**, Unit4 5 (May, 2003).
17. H. Noguchi, J. Park, T. Takagi, *Nucleic Acids Res* **34**, 5623 (2006).
18. R. D. Finn *et al.*, *Nucl. Acids Res.* **36**, D281 (January 11, 2008, 2008).
19. J. C. Wilgenbusch, D. Swofford, *Curr Protoc Bioinformatics* **Chapter 6**, Unit 6 4 (Feb, 2003).
20. Microbial Genomics Group at the Max Planck Institute for Marine Microbiology with the Department of Microbiology at the Technical University Munich and Ribocon, "Silva comprehensive ribosomal RNA database," (http://www.arb-silva.de/).
21. W. Ludwig *et al.*, *Nucleic Acids Res* **32**, 1363 (2004).
22. H. A. Schmidt, K. Strimmer, M. Vingron, A. von Haeseler, *Bioinformatics* **18**, 502 (Mar, 2002).

23. F. Chen, A. J. Mackey, C. J. Stoeckert, Jr., D. S. Roos, *Nucleic Acids Res* **34**, D363 (Jan 1, 2006).
24. S. J. Callister *et al.*, *PLoS One* **3**, e1542 (2008).
25. M. Mitreva, "Bacterial Core Gene Set," (http://hmpdacc.org/doc/sops/reference_genomes/metrics/Bacterial_CoreGenes_SOP.pdf).
26. K. Makarova, A. Sorokin, P. Novichkov, Y. Wolf, E. Koonin, *Biology Direct* **2**, 33 (2007).
27. M. Mitreva. "Archeal Core Gene Set," (http://hmpdacc.org/doc/sops/reference_genomes/metrics/Archaeal_CoreGenes_SOP.pdf)
28. K. T. Konstantinidis, J. M. Tiedje, *Proc Natl Acad Sci U S A* **102**, 2567 (Feb 15, 2005).

The Human Microbiome Jumpstart Reference Strains Consortium

Manuscript Preparation
Karen E. Nelson [1†]
George M. Weinstock [2]
Sarah K. Highlander [3,4]
Kim C. Worley [3,5]
Heather Huot Creasy [6]
Jennifer Russo Wortman [7,6]
Douglas B. Rusch [8]
Makedonka Mitreva [9]
Erica Sodergren [2]
Asif T. Chinwalla [2]
Michael Feldgarden [9]
Dirk Gevers [9]
Brian J. Haas [9]
Ramana Madupu [8]
Doyle V. Ward [9]

Principal Investigator
Bruce W. Birren [9]
Richard A. Gibbs [3,5]
Sarah K. Highlander [3,4]
Barbara Methe [1]
Karen E. Nelson [1]
Joseph F. Petrosino [3,4]
Robert L. Strausberg [1]
Granger G. Sutton [8]
George M. Weinstock [2]
Owen R. White [10,6]
Richard K. Wilson [2]

Annotation
Asif T. Chinwalla [2]
Heather Huot Creasy [6]
Scott Durkin [8]
Michelle Gwinn Giglio [6]
Sharvari Gujja [9]
Brian J. Haas [9]
Sarah K. Highlander [3,4]
Clint Howarth [9]
Chinnappa D. Kodira [11]
Nikos Kyrpides [12]
Ramana Madupu [8]
Teena Mehta [9]
Makedonka Mitreva [9]
Donna M. Muzny [3,5]
Matthew Pearson [9]
Kymberlie Pepin [2]
Amrita Pati [12]
Xiang Qin [3,5]
Kim C. Worley [3,5]
Jennifer Russo Wortman [7,6]
Chandri Yandava [9]
Qiandong Zeng [9]

Lan Zhang [3,5]

Assembly
Aaron M. Berlin [9]
Lei Chen [2]
Theresa A. Hepburn [9]
Justin Johnson [8]
Jamison McCorrison [8]
Jason Miller [8]
Pat Minx [2]
Donna M. Muzny [3,5]
Chad Nusbaum [9]
Xiang Qin [3,5]
Carsten Russ [9]
Granger G. Sutton [8]
Sean M. Sykes [9]
Chad M. Tomlinson [2]
Sarah Young [9]
Wesley C. Warren [2]
Kim C. Worley [3,5]

Data Analysis
Jonathan Badger [13]
Jonathan Crabtree [6]
Heather Huot Creasy [6]
Michael Feldgarden [9]
Dirk Gevers [9]
Sarah K. Highlander [3,4]
Ramana Madupu [8]
Victor M. Markowitz [14]
Makedonka Mitreva [2]
Donna M. Muzny [3,5]
Joshua Orvis [6]
Joseph F. Petrosino [3,4]
Douglas B. Rusch [8]
Granger G. Sutton [8]
Doyle V. Ward [9]
Kim C. Worley [3,5]
Jennifer Russo Wortman [7,6]

DNA Sequence Production
Andrew Cree [3,5]
Steve Ferriera [15]
Lucinda L. Fulton [2]
Robert S. Fulton [2]
Marcus Gillis [1]
Lisa D. Hemphill [3,5]
Vandita Joshi [3,5]
Christie Kovar [3,5]
Donna M. Muzny [3,5]
Manolito Torralba [1]
Xiang Qin [3,5]

Funding Agency Management
Kris A. Wetterstrand [16]

Genome Improvement
Amr Abouellleil [9]
Aye M. Wollam [2]
Christian J. Buhay [3,5]
Yan Ding [3,5]
Shannon Dugan [3,5]
Michael G. FitzGerald [9]
Lucinda L. Fulton [2]
Robert S. Fulton [2]
Mike Holder [3,5]
Jessica Hostetler [1]
Ramana Madupu [8]
Donna M. Muzny [3,5]
Xiang Qin [3,5]
Granger G. Sutton [8]

Project Leadership
Bruce W. Birren [9]
Sandra W. Clifton [2]
Sarah K. Highlander [3,4]
Karen E. Nelson [1]
Joseph F. Petrosino [3,4]
Erica Sodergren [2]
Robert L. Strausberg [1]
Granger G. Sutton [8]
George M. Weinstock [2]
Owen R. White [10,6]

Strain Management
Emma Allen-Vercoe [17]
Jonathan Badger [13]
Sandra W. Clifton [2]
Heather Huot Creasy [6]
Ashlee M. Earl [9]
Candace N. Farmer [2]
Michelle Gwinn Giglio [6]
Marcus Gillis [1]
Sarah K. Highlander [3,4]
Konstantinos Liolios [12]
Karen E. Nelson [1]
Erica Sodergren [2]
Michael G. Surette [18]
Granger G. Sutton [8]
Manolito Torralba [1]
Doyle V. Ward [9]
George M. Weinstock [2]
Jennifer Russo Wortman [7,6]
Qiang Xu [19]

Submissions
Asif T. Chinwalla [2]
Craig Pohl [2]
Scott Durkin [8]
Granger G. Sutton [8]
Katarzyna Wilczek-Boney [3,5]
Dianhui Zhu [3,5]

†Corresponding author.

1. Human Genomic Medicine, J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, Maryland, 20850, USA

2. The Genome Center, Washington University School of Medicine, 4444 Forest Park Ave, St. Louis, Missouri, 63108, USA

3. Human Genome Sequencing Center, Baylor College of Medicine, BCM226, One Baylor Plaza, Houston, Texas, 77030, USA

4. Department of Molecular Virology and Microbiology, BCM280, Baylor College of Medicine, One Baylor Plaza, Houston, Texas, 77030, USA

5. Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas, 77030, USA

6. Institute for Genome Sciences, University of Maryland School of Medicine, 801 W. Baltimore St. Baltimore, Maryland, 21201, USA

7. Department of Medicine, University of Maryland School of Medicine, Department of Genetics, 801 W. Baltimore St. Baltimore, Maryland, 21201, USA

8. Bioinformatics, J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, Maryland, 20850, USA

9. Genome Sequencing and Analysis Program, Broad Institute, 7 Cambridge Center, Cambridge, Massachusetts, 02142, USA

10. Department of Epidemiology and Preventive Medicine, University of Maryland School of Medicine, 801 W. Baltimore St., Baltimore, Maryland, 21201, USA

11. Genome Sequencing and Analysis Program, 454 Sequencing, 15 Commercial Street, Branford, Connecticut, 06405, USA

12. DOE-Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California, 94598, USA

13. Microbial and environmental genomics, J. Craig Venter Institute, 10355 Science Center Drive, La Jolla, California, 92121, USA

14. Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, California, 94720, USA

15. Sequencing, J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, Maryland, 20850, USA

16. NHGRI, 5635 Fishers Lane, Bethesda, Maryland, 20892, USA

17. Molecular and Cellular Biology, University of Guelph, 50 Stone Road, Guelph, Ontario, N1G 2W1, Canada

18. Microbiology & Infectious Diseases, University of Calgary, 3330 Hospital Drive, Calgary T2N4N1 Canada, Alberta, T2N4N1, Canada

19. Osel Inc., 4008 Burton Drive, Santa Clara, California, 95054, USA

Acknowledgements.

The authors gratefully acknowledge James Warren, Jingkun Zhang, R Gerald Fowler, Peter Pham, Dan Haft, Jeremy Selengut, Tanja Davidsen, Phil Goetz, Derek Harkins, Susmita Shrivastava, Sergey Koren, Brian Walenz, Les Foster, Indresh Singh, Jian Xu, Shiaw-Pyng Yang and Seth Schobel for bioinformatics support; Yu-hui Rogers, the Broad Genome Sequencing Platform, Yi Han, Viktoriya Korchina, Mark Scheel, Rebecca Thornton and the BCM-HGSC production team, Laura Courntey, Catrina Fronick, Otis Hall, Michelle O'Laughlin, Mark Cunningham, David O'Brien, Brenda Theising and the GCWU production team for sequencing; Jeff Gordon, Floyd Dewhirst, Brenda Wilson, Bryan White, Robert Mandrell, Martin Blaser, Roy H. Stevens, Sharon Hillier, Yang Liu, Zeli Shen, David Schauer, James Fox, Milton Allison, Chris D. Sibley, Delphine M. Saulnier, and Glenn R. Gibson for providing strains; and Maria Y. Giovanni, Carl L. Baker, Vivien Bonazzi, Carolyn D. Deal, Susan Garges, Robert W. Karp, R. Wayne Lunsford, Jane Peterson, Michael Wright, Tsegahiwot T. Belachew, and Christopher R. Wellington for funding agency management. We would like to acknowledge the National Institutes of Health for funding this project to The J. Craig Venter Institute (N01 AI 30071; U54-AI084844), Washington University (U54-HG003079; U54-HG004968), Baylor College of Medicine (U54-HG003273; U54-HG004973) and Broad Institute (HHSN272200900017C; U54-HG004969).