BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–3

Supplementary Material for Comparative Assembly Hubs: Web Accessible Browsers for Comparative Genomics

Ngan Nguyen[1+], Glenn Hickey[1+], Brian J. Raney[1+], Joel Armstrong[1], Hiram Clawson[1], Ann Zweig[1], Jim Kent[1], David Haussler[1], Benedict Paten[1*]

[1]Center for Biomolecular Sciences and Engineering, CBSE/ITI, UC Santa Cruz, 1156 High St, Santa Cruz, CA 95064, USA.

+ These authors contributed equally to this work.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

SUPPLEMENTARY METHODS

E. coli/Shigella Comparative Assembly Hubs

In this work, we generated three different *E. coli/Shigella* comparative assembly hubs. One with duplications allowed (`http://compbio.soe.ucsc.edu/reconstruction/ecoliComparativeHubs/ecoliWithDups/hub/hub.txt`), one with duplications disallowed (`http://compbio.soe.ucsc.edu/reconstruction/ecoliComparativeHubs/ecoliNoDups/hub/hub.txt`), and one that disallowed duplications and required all genomes to be present in every block (`http://compbio.soe.ucsc.edu/reconstruction/ecoliComparativeHubs/ecoliCore/hub/hub.txt`). The first two hubs can also be accessed via UCSC public hubs webpage (`http://tinyurl.com/UCSC-public-hubs`, hubs named "EcoliCompHubWtDups" and "EcoliCompHub"). For more information on viewing assembly hubs, see `http://hgwdev.cse.ucsc.edu/goldenPath/help/hgTrackHubHelp.html#View`.

The browser sessions of Figures 1a and 2 are available at `http://tinyurl.com/ecoliInversion` and `http://tinyurl.com/tandemDups`, respectively.

Each of the hubs was generated by the two following commands:

1. *runProgressiveCactus.sh –legacy –configFile config.xml –maxThreads 24 –ktType snapshot seqFile.txt outdir outdir/alignment.hal*
2. *hal2assemblyHub.py alignment.hal outHubDir –maxThreads 24 –lod –bedDirs Genes,RNA,GI,PI,PathogenicGenes,ARGB –rmskDir rmskTracks –gcContent –alignability –conservation conservationRegions.bed –conservationGenomeName reference –conservationTree tree.nw –tree tree.nw –rename shortnames.txt –hub ecoliCompHub –shortLabel EcoliCompHub –longLabel "Escherichia coli Comparative Assembly Hub"*

All related files can be found at `http://compbio.soe.ucsc.edu/reconstruction/ecoliComparativeHubs`, under directories "ecoliWithDups", "ecoliNoDups" and "ecoliCore', respectively. For more details of the options, please see the $hal2assemblyHub$ documentation at `https://github.com/glennhickey/hal`.

Nucleotide sequences of 57 *E. coli* and 9 *Shigella* spp. complete genomes were downloaded from the NCBI ftp site (`ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz`, January 2013). The sequences were repeat-masked using RepeatMasker (1) with the '*-xsmall*' option and otherwise default settings. The repeat-masked sequences were used as inputs to construct the MSA. Other outputs of RepeatMasker were converted into bigBed format to build the "Repetitive Elements" track for each genome (`http://genomewiki.ucsc.edu/index.php/RepeatMasker`). For the 9 genomes ATCC 873, DH1 161951, KO11FL 162099, KO11FL 52593, O104 H4 2009EL 2050, O104 H4 2009EL 2071, O104 H4 2011C 3493, UM146, BL21 Gold DE3 pLysS AG, we used the reverse complement of their assemblies as the majority portion of those assemblies aligned to the reverse strand of other (57) genomes.

Gene, protein and non-coding RNA annotations for each genome were also obtained from NCBI (`ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.gff.tar.gz`, `ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.faa.tar.gz` and `ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.rnt.tar.gz`, respectively).

Assessing The Genome Alignment

In a comparative assembly hub all genome comparison and lifted track displays are driven, consistently, by a single underlying genome alignment (and summaries of it). This provides great consistency and is likely to lead to less confusion in interpretation. For example, when looking at high-level views, which can always be drilled down on to reach the original base-level alignment, and when looking at lifted annotations, because any lift-over can be easily interrogated via a snake track that shows the actual alignment used to do the lift-over. However, this does mean that the accuracy of the genome alignment is important. Alignments for assembly hubs can be created by any aligner that can export a MAF file (a simple flat-file format), however, the currently most general and tested solution - and demonstrated here - is the Cactus alignment program.

*to whom correspondence should be addressed

Gene and Operon Alignment Assessment

Category	Total	Shared	Conserved	%
Gene Families	4751	3374	3333	98.80
Operons	535	452	452	100.00

Table 1. Proportion of orthologous gene families aligned in the multiple sequence alignment and proportion of K12 MG1655 operons present in other genomes with the gene order and orientation conserved. 'Total': average number of gene families each genome has or the total number of K12 MG1655 operons analyzed. 'Shared': average number of gene families each genome shares with another genome (pairwise comparisons) or average number of operons with all constituent genes conserved in another genome (pairwise comparisons). 'Conserved': average number of shared gene families that are aligned by the MSA or average number of 'shared' operons with the gene order and orientation conserved. '%': percentage of 'Shared' that are 'Conserved'.

Cactus has been developed primarily with the aim of aligning large eukaryotic genomes. In the recent Alignathon (`http://compbio.soe.ucsc.edu/alignathon/`) competition it proved highly accurate for this purpose (personal communication). To assess the alignment methodology for the alignment of bacteria, we assessed the *E. coli*/*Shigella* alignment to see how well orthologous genes of input genomes were aligned to each other. Gene annotations of each input genome obtained from NCBI and BLAT (2) pairwise alignments were used to group genes into orthologous groups (see Methods). For each pair of genomes we computed the number of orthologous coding gene families that were aligned in the Cactus alignment. On average each genome contains 4751 gene families and shares 3374 gene families with another genome (Supp. Table). Across all possible pairs of genomes, the vast majority (99%, 3333/3374) of each pair's orthologous groups were aligned to each other in the multiple alignment.

Rearrangements and gene gain and loss are commonly observed in *E. coli* and subsequently result in the gain and loss of operons (3). However, if an operon of one genome has all its constituent genes (individually) conserved in another genome, the order and orientation of these genes are often conserved as well (4). As another assessment of the alignment, we analyzed the conservation of *E. coli* K12 MG1655 operons when these operons are mapped by the alignment to other genomes (target genomes, 65 comparisons total). A total of 535 K12 MG1655 operons, each comprised of two or more genes, were included in the analysis (see Methods). On average, 452 of these operons had all their constituent genes (individually) conserved in the target genome. 'Conserved' was defined as being mapped by the alignment to the target genome with at least 90% coverage. Of the 452 operons, we found only two cases of operons in which the gene orders and orientations were broken, both due to rearrangements in the target genomes and not alignment errors. One was operon *envY-ompT* broken in five O157 genomes (EC4115, EDL933, Sakai, TW14359 and Xuzhou21) as a result of recombination. The other was operon *fumAC* broken in *Shigella sonnei* 53G due to an inversion.

E. coli/Shigella Core Genome and Pan-Genome

The core genome (for 66 strains) computed by the CAH pipeline is 2.7 Mbp in size. It is expected that the core genome size decreases as the number of genomes increases, until enough genomes are added, at which point the core genome size becomes stabilized (3; 5; 6). This observation is recapitulated here, as shown in Supp. Figure 1.

The pan-genome is ∼11 Mbp (Supp. Figure 3, `http://tinyurl.com/ecoli-pangenome`).

Gene and Operon Analyses

Paralogous and orthologous annotated coding genes were identified by BLAT amino acid sequence pairwise alignments. For each genome, genes were grouped into a single gene family if they shared at least 90% amino acid identity over at least 90% of the length of the longest gene.

To identify orthologous gene families shared among the genomes, we used the divide and conquer approach. Briefly, we started by breaking the input set of genomes into pairs. For each pair of genomes, we computed their union list of gene families by grouping orthologous gene families together. The resulted union gene family lists of all pairs were recursively treated as a new set of genomes and the process of finding union lists was repeated until orthologous gene families of all genomes were grouped together and one union gene family list was obtained. Two gene families of two genomes was identified as orthologous if at least one gene of one family had a reciprocal match with at least one gene of the other family. A match was defined as having at least 90% amino acid identity and 90% coverage.

To assess the multiple sequence alignment, for each pair of genomes, we computed the number of orthologous gene families that were aligned in the MSA and reported the average statistics of all pairs. Two orthologous gene families were considered as aligned in the MSA if at least one gene of one family was aligned to one gene of the other family by the MSA with a minimum coverage of 90% of the longer gene.

As another assessment of the MSA, we analyzed the gene order and orientation conservation of the well-annotated *E. coli* K12 MG1655's operons that were also present in other genomes. Operons (or more accurately, transcription units) of K12 MG1655 were downloaded from RegulonDB. As the orders and orientations of the genes were of interest, only operons with two or more genes (and no pseudogene) were included in the analysis. In addition, we filtered out annotations without strong evidences, which we defined as operons with no other evidence than one of the following: "Inferred by computational analysis" (ICA), "Inferred computationally without human oversight" (ICWHO), "Non-traceable author statement" (NTAS) and "Polar mutation" (PM). After filtering, there were 535 operons total. For each genome other than K12 MG1655 (called target genome, 65 genomes total), we calculated the number of K12 MG1655 operons that had all their constituent genes present (having an ortholog) in the target genome and the percentage of these operons with the gene order and orientation conserved by the MSA. We reported the average statistics in Sup. Table .

REFERENCES

[1]Smit, A.F.A., Hubley, R., Green, P.: RepeatMasker. open-3.0 (1996-2013)

[2]Kent, W.J.W.: BLAT–the BLAST-like alignment tool. Genes & Development **12**(4) (April 2002) 656–664

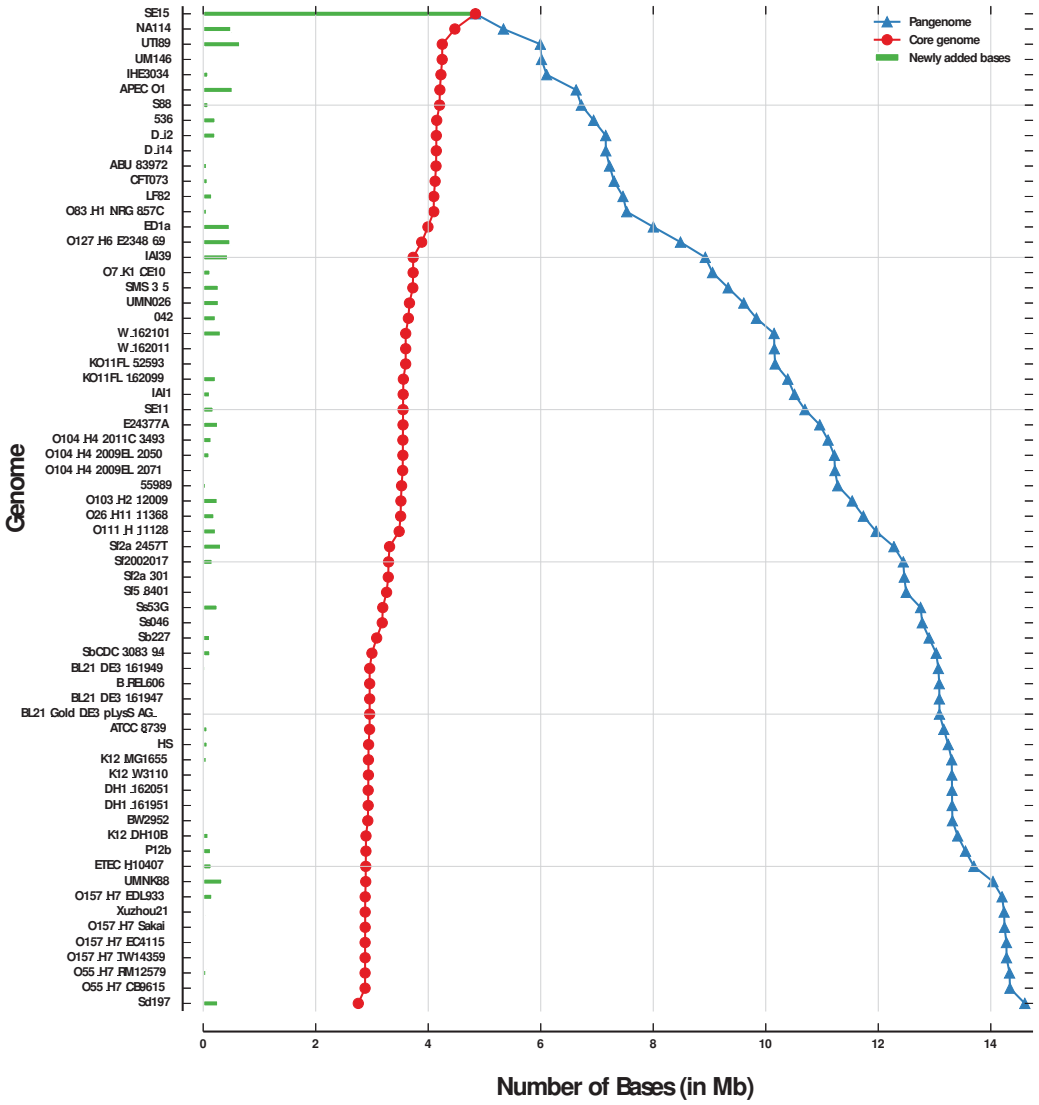

Fig. 1. Pan-genome and core genome sizes, an adaptation of Figure 4 in Lukjancenko *et al.* (5), supporting the open pan-genome and the stable core genome models in *E. coli*. The x-axis shows the number of bases (in Mb) in the pan-genome (blue triangle), core genome (red circle) and the number of the new bases added to the pan-genome (horizontal green bar) as more genomes are added into the analysis. The y-axis shows the genome that is added each step. Genomes were added in the order guided by the phylogenetic tree in Supp. Figure 2.

[3]Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller, S., Danchin, A., Diard, M., Dossat, C., Karoui, M.E., Frapy, E., Garry, L., Ghigo, J.M., Gilles, A.M., Johnson, J., Le Bouguénec, C., Lescat, M., Mangenot, S., Martinez-Jéhanne, V., Matic, I., Nassif, X., Oztas, S., Petit, M.A., Pichon, C., Rouy, Z., Ruf, C.S., Schneider, D., Tourret, J., Vacherie, B., Vallenet, D., Médigue, C., Rocha, E.P.C., Denamur, E.: Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genetics **5**(1) (January 2009) e1000344

[4]Rocha, E.P.C.: The Organization of the Bacterial Genome. Annual Review of Genetics **42**(1) (December 2008) 211–233

[5]Lukjancenko, O., Wassenaar, T.M., Ussery, D.W.: Comparison of 61 sequenced Escherichia coli genomes. Microbial ecology **60**(4) (November 2010) 708–720

[6]Leimbach, A., Hacker, J., Dobrindt, U.: E. coli as an all-rounder: the thin line between commensalism and pathogenicity. Current topics in microbiology and immunology **358** (2013) 3–32

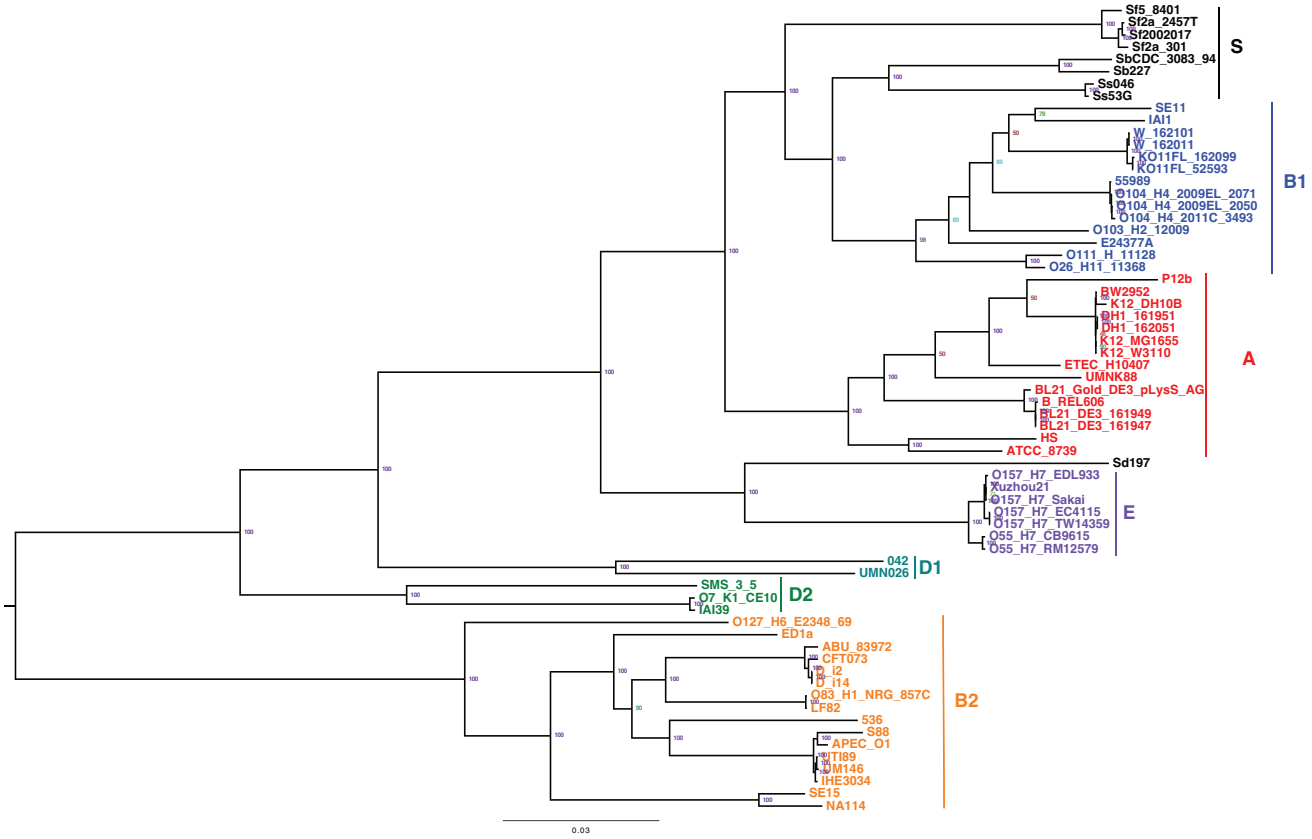

Fig. 2. Maximum-likelihood based phylogenetic tree of 66 *E. coli* and *Shigella* spp. genomes, constructed from their core genome alignment using RAxML and FigTree `http://tree.bio.ed.ac.uk/software/figtree/`. The genomes are colored by their annotated phylogroups: orange: B2, green: D2, teal: D1, purple: E, red: A, blue: B1 and black: *Shigella*.

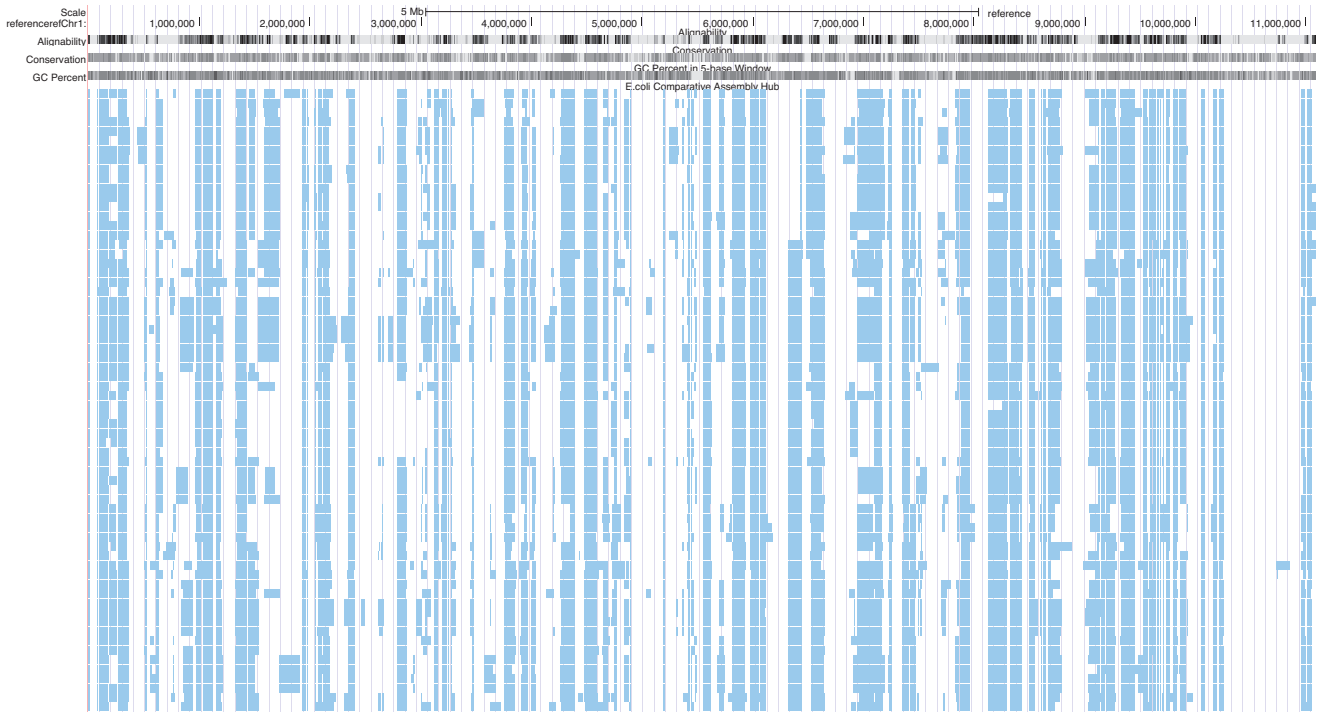

Fig. 3. Pan-genome browser of 66 *E. coli*/*Shigella* genomes. The snake tracks (in blue) are in dense mode.