The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes

Background. As whole genome sequence data from bacterial isolates becomes cheaper to generate, computational methods are needed to correlate sequence data with biological observations. Here we present the large-scale BLAST score ratio (LS-BSR) pipeline, which rapidly compares the genetic content of hundreds to thousands of bacterial genomes, and returns a matrix that describes the relatedness of all coding sequences (CDSs) in all genomes surveyed. This matrix can be easily parsed in order to identify genetic relationships between bacterial genomes. Although pipelines have been published that group peptides by sequence similarity, no other software performs the rapid, large-scale, full-genome comparative analyses carried out by LS-BSR. **Results.** To demonstrate the utility of the method, the LS-BSR pipeline was tested on 96 *Escherichia coli* and *Shigella* genomes; the pipeline ran in 163 minutes using 16 processors, which is a greater than 7-fold speedup compared to using a single processor. The BSR values for each CDS, which indicate a relative level of relatedness, were then mapped to each genome on an independent core genome single nucleotide polymorphism (SNP) based phylogeny. Comparisons were then used to identify clade specific CDS markers and validate the LS-BSR pipeline based on molecular markers that delineate between classical *E. coli* pathogenic variant (pathovar) designations. Scalability tests demonstrated that the LS-BSR pipeline can process 1,000 *E. coli* genomes in 27-57h, depending upon the alignment method, using 16 processors. **Conclusions.** LS-BSR is an open-source, parallel implementation of the BSR algorithm, enabling rapid comparison of the genetic content of large numbers of genomes. The results of the pipeline can be used to identify specific markers between user-defined phylogenetic groups, and to identify the loss and/or acquisition of genetic information between bacterial isolates. Taxa-specific genetic markers can then be translated into clinical diagnostics, or can be used to identify broadly conserved putative therapeutic candidates.

1 **Authors:**

2 Jason W. Sahl*
3 Division of Pathogen Genomics
4 Translational Genomics Research Institute
5 3051 W. Shamrell Blvd, Suite 106
6 Flagstaff, AZ, 86001, USA
7 Email: jsahl@tgen.org

8 J. Gregory Caporaso
9 Department of Biological Sciences
10 Northern Arizona University
11 1298 S. Knoles Drive, PO Box 4073
12 Flagstaff, AZ, 86011, USA
13 Email: gregcaporaso@gmail.com

14 David A. Rasko
15 Department of Microbiology and Immunology
16 University of Maryland School of Medicine
17 Baltimore, MD, 21202, USA
18 Email: drasko@som.umaryland.edu

19 Paul Keim
20 Center for Microbial Genetics and Genomics
21 Northern Arizona University
22 1298 S. Knoles Drive, PO Box 4073
23 Flagstaff, AZ, 86011, USA
24 Email: paul.keim@nau.edu

25 *To whom correspondence should be addressed

26 **INTRODUCTION**

27   Whole genome sequence (WGS) data has changed our view of bacterial relatedness
28   and evolution. Computational analyses available for WGS data include, but are not
29   limited to, single nucleotide polymorphism (SNP) discovery (DePristo et al. 2011), core
30   genome phylogenetics (Sahl et al. 2011), and gene based comparative methods (Hazen
31   et al. 2013; Sahl et al. 2013).   In 2005, a BLAST score ratio (BSR) method was
32   introduced in order to compare peptide identity from a limited number of bacterial
33   genomes (Rasko et al. 2005). However, the "all vs. all" implementation of this method
34   scales poorly with a larger number of sequenced genomes.

35   Here we present the Large Scale BSR method (LS-BSR) that can rapidly compare
36   gene content of a large number of bacterial genomes. Comparable methods have been
37   published in order to group genes into gene families, including OrthoMCL (Li et al.
38   2003), TribeMCL (Enright et al. 2002), and GETHOGs (Altenhoff et al. 2013). Although
39   grouping peptides into gene families is not the primary focus of LS-BSR, the output can
40   be parsed to identify the pan-genome (Tettelin et al. 2008) structure of a species; scripts
41   are included with LS-BSR that classify coding sequences (CDSs) into pan-genome
42   categories based on user-defined identity thresholds.

43   Pipelines have also been established to perform comprehensive pan-genome
44   analyses, including the pan-genome analysis pipeline (PGAP) (Zhao et al. 2012), which
45   requires specific gene annotation from GenBank and complicates the analysis of large
46   numbers of novel genomes. PGAP also doesn't allow for the screen of specific genes of
47   interest against query genomes in order to identify patterns of distribution.
48   GET_HOMOLOGUES (Contreras-Moreira & Vinuesa 2013) is a recently published tool
49   that can be used for pan-genome analyses, including the generation of dendrograms
50   based on the presence/absence of homologous genes; by only using presence/absence
51   based on gene homology, more distantly related gene relatedness cannot be fully
52   investigated. The integrated toolkit for the exploration of microbial pan-genomes (ITEP)
53   toolkit (Benedict et al. 2014) was recently published and performs similar functions to
54   LS-BSR, including the identification of gene gain/loss at nodes of a phylogeny.   ITEP
55   relies on multiple dependencies and workflows, which are available as a pre-packaged
56   virtual machine. The authors of ITEP report that an analysis of 200 diverse genomes
57   would take ~ 6 days on a server with 12 processors and scales quadratically with
58   additional genomes.

MATERIALS AND METHODS

The LS-BSR method can either use a defined set of genes, or can use Prodigal (Hyatt et al. 2010) to predict CDSs from a set of query genomes. When using Prodigal, all CDSs are concatenated and then de-replicated using USEARCH (Edgar 2010) at a pairwise identity of 0.9 (identity threshold can be modified by the user). Each unique CDS is then translated with BioPython (www.biopython.org) and aligned against its nucleotide sequence with TBLASTN (Altschul et al. 1997) to calculate the reference bit score; if BLASTN or BLAT (Kent 2002) is invoked, the nucleotide sequences are aligned. Each query sequence is then aligned against each genome with BLAT, BLASTN, or TBLASTN and the query bit score is tabulated. The BSR value is calculated by dividing the query bit score by the reference bit score, resulting in a BSR value between 0.0 and 1.0 (values slightly higher than 1.0 have been observed due to variable bit score values obtained by TBLASTN). The results of the LS-BSR pipeline include a matrix that contains each unique CDS name and the BSR value in each genome surveyed. CDSs that have more than one significant BSR value in at least one genome are also identified in the output. A separate file is generated for CDSs where one duplicate is significantly different than the other in at least one genome; these regions could represent paralogs and may require further detailed investigation. Once the LS-BSR matrix is generated, the results can easily be visualized as a heatmap or cluster with the Multiple Experiment Viewer (MeV) (Saeed et al. 2006) or R (R Core Team 2013); the heatmap represents a visual depiction of the relatedness of all peptides in the pan-genome across all genomes. The Interactive Tree Of Life project (Bork et al. 2008) can also be used to generate heatmaps from LS-BSR output and correlate heatmap data with a provided phylogeny. A script is included with LS-BSR (compare_BSR.py) to rapidly compare CDSs between user-defined sub-groups, using a range of BSR thresholds set for CDS presence/absence. Annotation of identified CDSs can then be applied using tools including RAST (Aziz et al. 2008) and prokka (http://www.vicbioinformatics.com/software.prokka.shtml). LS-BSR source code, unit tests, and test data can be freely obtained at https://github.com/jasonsahl/LS-BSR under a GNU GPL v3 license.

RESULTS AND DISCUSSION

LS-BSR algorithm speed and scalability. To determine the scalability of the LS-BSR method, 1,000 *Escherichia coli* and *Shigella* genomes were downloaded from Genbank (Benson et al. 2012); *E. coli* was used as a test case due to the large number of genomes deposited in Genbank. Genomes were sub-sampled at different depths (100 through 1000, sampling every 100) with a python script (https://gist.github.com/jasonsahl/115d22bfa35ac932d452) and processed with LS-BSR using 16 processors. A plot of wall time and the number of genomes processed demonstrates the scalability of the method (Figure 1A) using three different alignment methods. To demonstrate the parallel nature of the algorithm, 100 *E. coli* genomes were processed with different numbers of processors. The results demonstrate decreased runtime of LS-BSR with an increase in the number of processors used (Figure 1B).

Improvements on a previous BSR implementation. The LS-BSR method is an improvement on a previous BSR implementation (http://bsr.igs.umaryland.edu/) in terms of speed and ease of use. The former BSR algorithm (Rasko, et al., 2005) requires peptide sequences and genomic coordinates of CDSs to run. LS-BSR only requires genome assemblies in FASTA format, which is the standard output of most genome assemblers. To test the speed differences between methods, 10 *E. coli* genomes (Supplemental Table 1) were processed with both methods. Using the same number of processors (n=2) on the same server, the original BSR method took ~14 hours (wall time) to complete, while the LS-BSR method, using TBLASTN, took ~25 minutes to complete (wall time). Because the original BSR method is an "all vs. all" comparison and the LS-BSR method is a "one vs. all" comparison, this difference is expected to be more pronounced as the number of genomes analyzed increases.

Test case: analysis of 96 *E. coli* and *Shigella* genomes. To demonstrate the utility of the LS-BSR pipeline, a set of 96 *E. coli* and *Shigella* genomes were processed (Supplemental Table 1); these genomes are in various stages of assembly completeness and have been generated with various sequencing technologies from Sanger to Illumina. The BSR matrix was generated with TBLASTN in 2h34m from a set of ~20,000 unique CDSs using 16 processors. In addition to the LS-BSR analysis, a core genome single nucleotide polymorphism (SNP) phylogeny was inferred on 96 genomes using methods published previously (Sahl et al. 2011); the SNP phylogeny

121 with labels is shown in Supplemental Figure 1. Briefly, all genomes were aligned with
122 Mugsy (Angiuoli & Salzberg 2010) and the core genome was extracted from the whole
123 genome alignment; the alignment file was then converted into a multiple sequence
124 alignment in FASTA format. Gaps in the alignment were removed with Mothur (Schloss
125 et al. 2009) and a phylogeny was inferred on the reduced alignment with FastTree2
126 (Price et al. 2010).

127   The compare_BSR.py script included with LS-BSR was used to identity CDS
128 markers that are unique to specific phylogenetic clades (Figure 2). Identified CDSs had
129 a BSR value ≥0.8 in targeted genomes and a BSR value <0.4 in non-targeted genomes;
130 the gene annotation of all marker CDSs is detailed in Supplemental Table 2. The
131 conservation and distribution of all clade-specific markers was visualized by correlating
132 the phylogeny with a heatmap of BSR values (Figure 2). This presentation provides an
133 easy way for the user to highlight features conserved in one or more phylogenomic
134 clades.

135   *E. coli* and *Shigella* pathogenic variants (pathovars) are delineated by the presence
136 of genetic markers primarily present on mobile genetic elements (Rasko et al. 2008).
137 The conservation of these markers was used as a validation of the LS-BSR method. A
138 representative sequence from each pathovar-specific marker (Supplemental Table 2)
139 was screened against the 96-genome test set and the BSR values (Supplemental Table
140 3) were visualized as a heatmap (Figure 2). The BSR matrix demonstrates that
141 pathovar-specific genes were accurately identified in each targeted genome
142 (Supplemental Table 3, Figure 2). For example, the *ipaH3* marker was positively
143 identified in all *Shigella* genomes and the Shiga toxin gene (*stx2a*) was conserved in the
144 clade including O157:H7 *E. coli* (Figure 2). A sub-set of these 96 *E. coli* genomes is
145 included with LS-BSR as test data to characterize the conservation and distribution of
146 pathovar specific genes.

147   Finally, the BSR values were used to cluster all 96 genomes with an average linkage
148 algorithm implemented in MeV and the structure of the resulting dendrogram was
149 compared to the core SNP phylogeny. The BSR based clustering method incorporates
150 both the core and accessory genome, while the SNP phylogeny relies on core genomic
151 regions alone. A comparison of the tree structures demonstrates that while *Shigella*
152 genomes share a diverse evolutionary history (Figure 3A), they all cluster together
153 based on gene presence and conservation (Figure 3B). This result was also observed

154 using a k-mer frequency method (Sims & Kim 2011), which uses all possible k-mer
155 values to infer a phylogeny and validates the findings of the LS-BSR pipeline. The
156 dendrogram also differed from the core SNP phylogeny in other genomes, which could
157 represent either assembly problems, or more likely the acquisition of accessory genomic
158 regions that are not a product of direct descent.

159 The functionality of LS-BSR was compared to recently released pan-genome
160 software packages including GET_HOMOLOGUES (Contreras-Moreira & Vinuesa
161 2013), ITEP (Benedict et al. 2014), and PGAP (Zhao et al. 2012). A set of 11
162 *Streptococcus pyogenes* genomes was chosen for the comparative analysis, as it was
163 also used as a test set in the PGAP publication; the comparative analysis and results
164 are shown in Table 1. Overall, the size of the core genome was comparable between
165 methods, with LS-BSR (BLASTN) and GET-HOMOLOGUS calculating differing core
166 genome numbers compared to the other methods. However, small differences were
167 expected due to differing thresholds and clustering algorithms. Based on these results,
168 LS-BSR represent a significant improvement in terms of speed and ease of use
169 compared to comparable methods, while having comparable utility.

170 **Pan-genome analyses**. One application in comparative genomics is the analysis of the
171 pan-genome, or the combined genome, of isolates within a species. Post matrix-
172 building scripts are available to visualize the pan-genome of a given dataset. One script
173 (BSR_to_PANGP.py) creates a matrix compatible with PanGP (Zhao et al. 2014), for
174 visualization of pan-genome statistics. The pan_genome_stats.py script provides data
175 that can be used to visualize the conservation of CDSs at different genome depths
176 (Figure 4A). An additional script randomly subsamples the CDS distribution at all depths
177 and produces data that can be plotted to visualize core genome convergence (Figure
178 4B), accumulation of CDSs (Figure 4C), and the number of unique CDSs for each
179 genome analyzed (Figure 4D). All analyses were conducted on a set of 100 *E. coli*
180 genomes, with 100 iterations.

181 **CONCLUSIONS**
182 The LS-BSR method can rapidly compare the gene content of a relatively large
183 number of bacterial genomes in either draft or complete form, though with more
184 fragmented assemblies LS-BSR is likely to perform sub-optimally. As sequence read

185 lengths improve, assembly fragmentation should become less problematic due to more
186 contiguous assemblies. LS-BSR can also be used to rapidly screen a collection of
187 genomes for the conservation of known virulence factors or genetic features. By using a
188 range of peptide relatedness, instead of a defined threshold, homologs and paralogs
189 can also be identified for further characterization.

190   LS-BSR is written in python, with many steps conducted in parallel. This allows the
191 script to scale well from hundreds to thousands of genomes. The LS-BSR method is a
192 major improvement on a previous BSR implementation in terms of speed, ease of use,
193 and utility. As more WGS data from bacterial genomes become available, methods will
194 be required to quickly compare their genetic content and perform pan-genome analyses.
195 LS-BSR is an open-source software package to rapidly perform these comparative
196 genomic workflows.

197 **ACKNOWLEDGEMENTS**

198 Thanks to Darrin Lemmer for his critical review of the LS-BSR code.

199 **REFERENCES**

200 Altenhoff AM, Gil M, Gonnet GH, and Dessimoz C. 2013. Inferring hierarchical
201   orthologous groups from orthologous gene pairs. *PLoS ONE* 8:e53786.
202 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ.
203   1997. Gapped BLAST and PSI-BLAST: a new generation of protein database
204   search programs. *Nucleic Acids Res* 25:3389-3402.
205 Angiuoli SV, and Salzberg SL. 2010. Mugsy: Fast multiple alignment of closely related
206   whole genomes. *Bioinformatics*.
207 Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S,
208   Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA,
209   McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R,
210   Vassieva O, Vonstein V, Wilke A, and Zagnitko O. 2008. The RAST Server: rapid
211   annotations using subsystems technology. *BMC genomics* 9:75.
212 Benedict MN, Henriksen JR, Metcalf WW, Whitaker RJ, and Price ND. 2014. ITEP: An
213   integrated toolkit for exploration of microbial pan-genomes. *BMC genomics* 15:8.
214 Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, and Sayers EW. 2012.
215   GenBank. *Nucleic Acids Res* 40:D48-53.
216 Bork P, Hugenholtz P, Kunin V, Raes J, Harris JK, Spear JR, Walker JJ, Ivanova N, von
217   Mering C, Bebout BM, and Pace NR. 2008. Millimeter-scale genetic gradients
218   and community-level molecular convergence in a hypersaline microbial mat. *Mol*
219   *Sys Biol* 4:198-198.

220 Contreras-Moreira B, and Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software
221   package for scalable and robust microbial pangenome analysis. *Appl Environ*
222   *Microbiol* 79:7696-7701.
223 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del
224   Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM,
225   Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, and Daly MJ. 2011. A
226   framework for variation discovery and genotyping using next-generation DNA
227   sequencing data. *Nature genetics* 43:491-498.
228 Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST.
229   *Bioinformatics* 26:2460-2461.
230 Enright AJ, Van Dongen S, and Ouzounis CA. 2002. An efficient algorithm for large-scale
231   detection of protein families. *Nucleic acids research* 30:1575-1584.
232 Hazen TH, Sahl JW, Fraser CM, Donnenberg MS, Scheutz F, and Rasko DA. 2013.
233   Refining the pathovar paradigm via phylogenomics of the attaching and effacing
234   *Escherichia coli*. *Proc Natl Acad Sci U S A* 110:12810-12815.
235 Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, and Hauser LJ. 2010. Prodigal:
236   prokaryotic gene recognition and translation initiation site identification. *BMC*
237   *Bioinformatics* 11:119.
238 Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* 12:656-664.
239 Li L, Stoeckert CJ, Jr., and Roos DS. 2003. OrthoMCL: identification of ortholog groups
240   for eukaryotic genomes. *Genome Res* 13:2178-2189.
241 Price MN, Dehal PS, and Arkin AP. 2010. FastTree 2--approximately maximum-
242   likelihood trees for large alignments. *PLoS ONE* 5:e9490.
243 R Core Team RCT. 2013. R: A language and environment for statistical computing.
244   *Available at* http://www.R-project.org.
245 Rasko DA, Myers GS, and Ravel J. 2005. Visualization of comparative genomic
246   analyses by BLAST score ratio. *BMC Bioinformatics* 6:2.
247 Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J,
248   Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, and Ravel
249   J. 2008. The pangenome structure of *Escherichia coli:* comparative genomic
250   analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* 190:6881-
251   6893.
252 Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan
253   M, White JA, and Quackenbush J. 2006. TM4 microarray software suite. *Methods*
254   *Enzymol* 411:134-193.
255 Sahl JW, Gillece JD, Schupp JM, Waddell VG, Driebe EM, Engelthaler DM, and Keim P.
256   2013. Evolution of a pathogen: a comparative genomics analysis identifies a
257   genetic pathway to pathogenesis in *Acinetobacter*. *PLoS ONE* 8:e54287.
258 Sahl JW, Steinsland H, Redman JC, Angiuoli SV, Nataro JP, Sommerfelt H, and Rasko
259   DA. 2011. A comparative genomic analysis of diverse clonal types of
260   enterotoxigenic *Escherichia coli* reveals pathovar-specific conservation. *Infect*
261   *Immun* 79:950-960.
262 Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA,
263   Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn
264   DJ, and Weber CF. 2009. Introducing mothur: Open-Source, Platform-
265   Independent, Community-Supported Software for Describing and Comparing
266   Microbial Communities. *Appl Environ Microbiol* 75:7537-7541.

267 Sims GE, and Kim SH. 2011. Whole-genome phylogeny of *Escherichia coli/Shigella*
268   group by feature frequency profiles (FFPs). *Proc Natl Acad Sci U S A* 108:8329-
269   8334.
270 Tettelin H, Riley D, Cattuto C, and Medini D. 2008. Comparative genomics: the bacterial
271   pan-genome. *Curr Opin Microbiol* 11:472-477.
272 Zhao Y, Jia X, Yang J, Ling Y, Zhang Z, Yu J, Wu J, and Xiao J. 2014. PanGP: A tool for
273   quickly analyzing bacterial pan-genome profile. *Bioinformatics*.
274 Zhao Y, Wu J, Yang J, Sun S, Xiao J, and Yu J. 2012. PGAP: pan-genomes analysis
275   pipeline. *Bioinformatics* 28:416-418.

276 **Figure Legends:**

277 **Figure 1.** Time performance of the LS-BSR pipeline. Panel A) 1000 *Escherichia coli*
278 and *Shigella* genomes were randomly sub-sampled and analyzed using default LS-BSR
279 parameters and 16 processors. Wall time was plotted against the number of genomes
280 analyzed. The results demonstrate that the LS-BSR pipeline scales well with increasing
281 numbers of genomes. Panel B) The same set of 100 *E. coli* genomes was processed
282 with different numbers of processors and the wall time was plotted.   The results
283 demonstrate that using additional processors decreases the overall run time of LS-BSR.

284 **Figure 2**. The distribution of virulence factors and phylogenomic markers associated
285 with a core single nucleotide polymorphism (SNP) phylogeny. The core SNP phylogeny
286 was inferred from a whole genome alignment produced by Mugsy (Angiuoli & Salzberg
287 2010).   Known virulence genes (Supplemental Table 2) were screened against 96
288 *Escherichia coli* and *Shigella* genomes using BLASTN within LS-BSR. Clade specific
289 markers were identified at defined nodes in the phylogeny (A through Q).   Gene
290 annotations for these markers are detailed in Supplemental Table 2.

291 **Figure 3**. A comparison of 96 *Escherichia coli / Shigella* genomes between (Panel A) a
292 core single nucleotide polymorphism (SNP) phylogeny or (Panel B) a cluster generated
293 with the Multiple Experiment Viewer (Saeed et al. 2006) from BLAST Score Ratio (BSR)
294 values that include the entire pan-genome. Colors applied to each classical *E. coli*
295 phylogroup were applied to the SNP phylogeny and transferred to the BSR cladogram.
296 *Shigella* genomes are marked with a red circle.

297   **Figure 4**. Pan-genome plots generated from LS-BSR output. Analyses were conducted

298   on a set of 100 *Escherichia coli* genomes. The distribution of coding region sequences

299   (CDSs) across the set of genomes surveyed is shown in Panel A. A supplemental script

300   can be used to better understand the convergence of the core genome (Panel B), the

301   accumulation of CDSs (Panel C), and the number of unique CDSs for each genome

302   analyzed (Panel D); each analysis was conducted with 100 random sum-samplings and

303   means are depicted with red diamonds.

304   **Supplemental Figure 1**. A core genome SNP phylogeny of 96 *Escherichia coli* and

305   *Shigella* genomes. The core genome was extracted from the output of Mugsy (Angiuoli

306   & Salzberg 2010) and the phylogeny was inferred with FastTree2 (Price et al. 2010).

307   This phylogeny contains labels that can be used to identify specific genomes in Figures

308   2 and 3.

Comparison of four pan-genome methods on a test set of 11 *Streptococcus pyogenes* genomes

Table 1. Comparison of four pan-genome methods on a test set of 11 *Streptococcus pyogenes* genomes

	LS-BSR	GET_HOMOLOGUES	PGAP	ITEP
clusters orthologs?	Yes	Yes	Yes	Yes
open source?	Yes	Yes	Yes	Yes
pan-genome calculation?	Yes	Yes	Yes	Yes
lineage specific gene identification?	Yes	Yes	Yes	Yes
functional annotation?	No	Yes	Yes	Yes
analyzes user-defined genes?	Yes	No	No	Yes
Input files	".fasta"	GenBank or ".faa"	".faa",".fna",".ppt"	GenBank
supported platforms	linux,OSX	linux/OSX	linux	linux/OSX
core genome size	1318,1350,1426[a]	1232,1234[b]	1332,1366[c,d]	1342
time (2 cores), only runtime	5m59s,1m53s,1m17s[a]	25m14s	29m59s,199m58s[c,d]	24m22s

[a]TBLASTN,BLASTN,BLAT
[b]COG,MCL
[c]MP,GF
[d]Taken from publication

Figure 1

Time performance of the LS-BSR pipeline.

Panel A) 1000 *Escherichia coli* and *Shigella* genomes were randomly sub-sampled and analyzed using default LS-BSR parameters and 16 processors. Wall time was plotted against the number of genomes analyzed. The results demonstrate that the LS-BSR pipeline scales well with increasing numbers of genomes. Panel B) The same set of 100 *E. coli* genomes was processed with different numbers of processors and the wall time was plotted. The results demonstrate that using additional processors decreases the overall run time of LS-BSR.

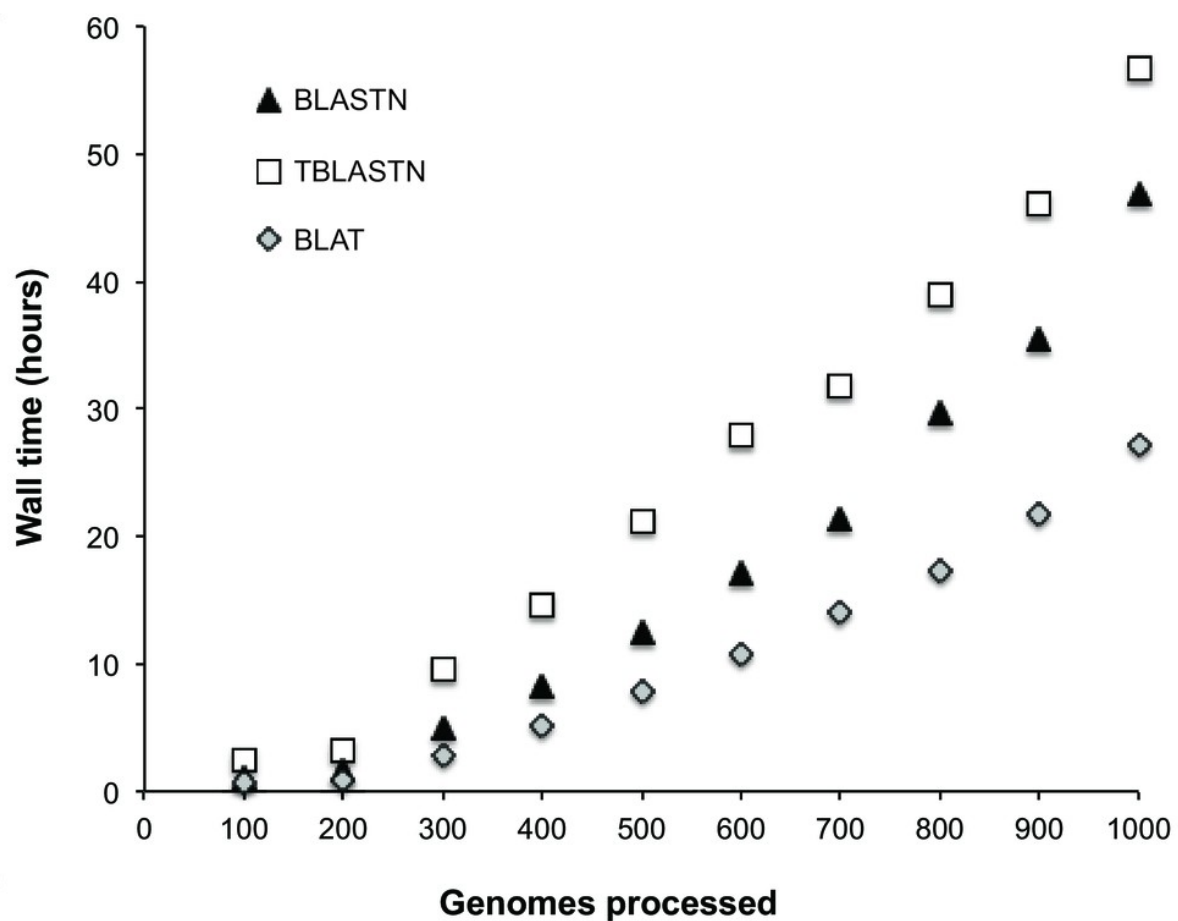
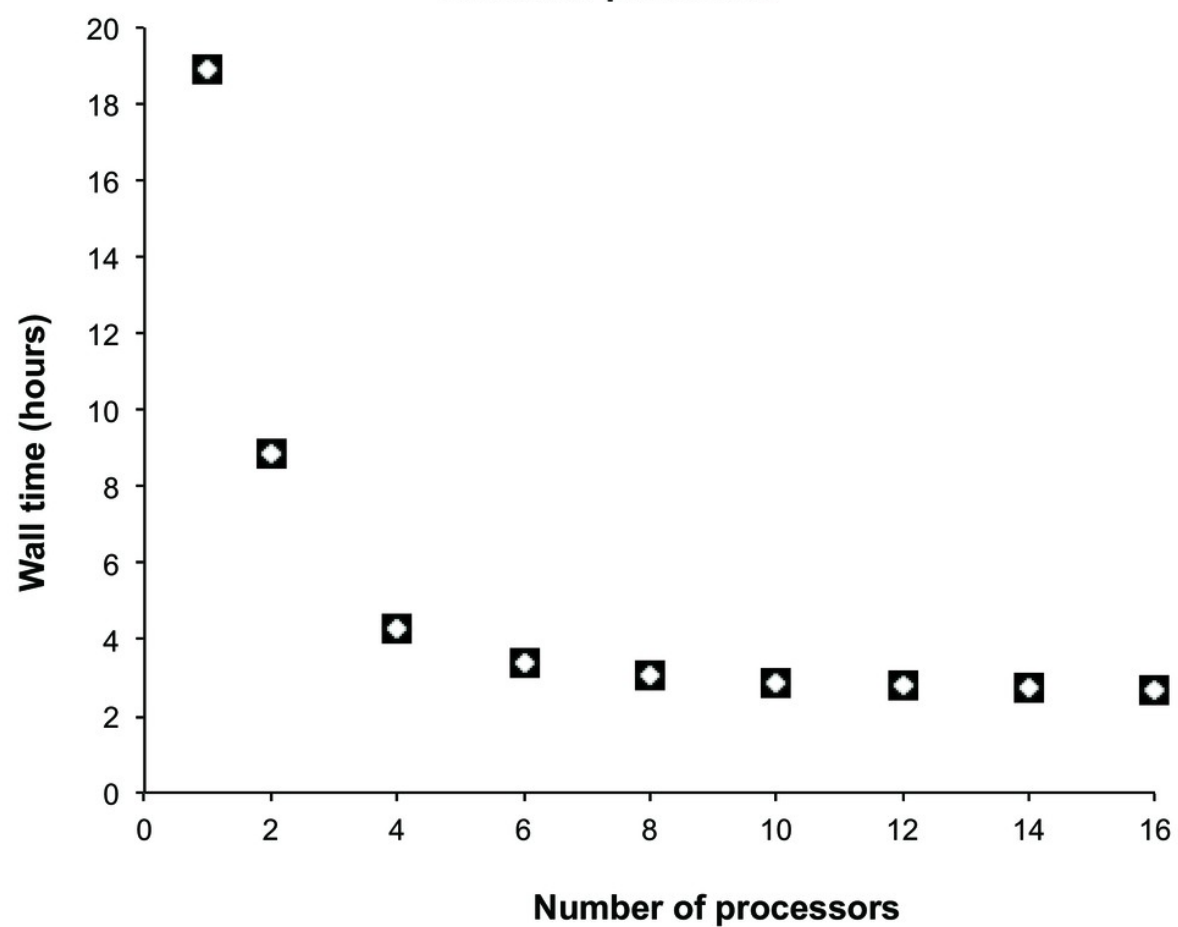



Figure 2

The distribution of virulence factors and phylogenomic markers associated with a core single nucleotide polymorphism (SNP) phylogeny

The core SNP phylogeny was inferred from a whole genome alignment produced by Mugsy (Angiuoli & Salzberg 2010) . Known virulence genes (Supplemental Table 2) were screened against 96 *Escherichia coli* and *Shigella* genomes using BLASTN within LS-BSR. Clade specific markers were identified at defined nodes in the phylogeny (A through Q). Gene annotations for these markers are detailed in Supplemental Table 2.

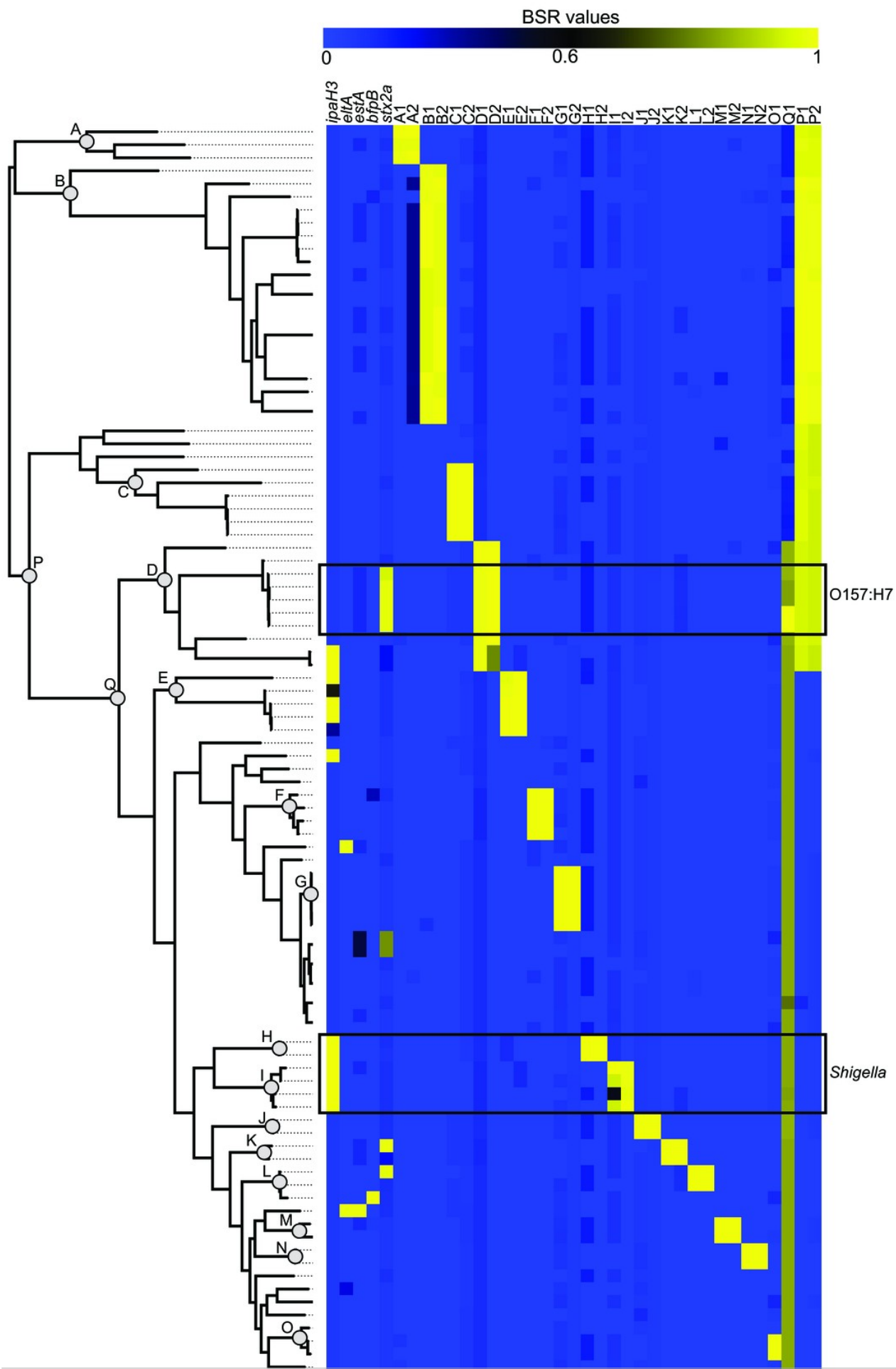



Figure 3

Comparison of LS-BSR cluster with core genome SNP phylogeny

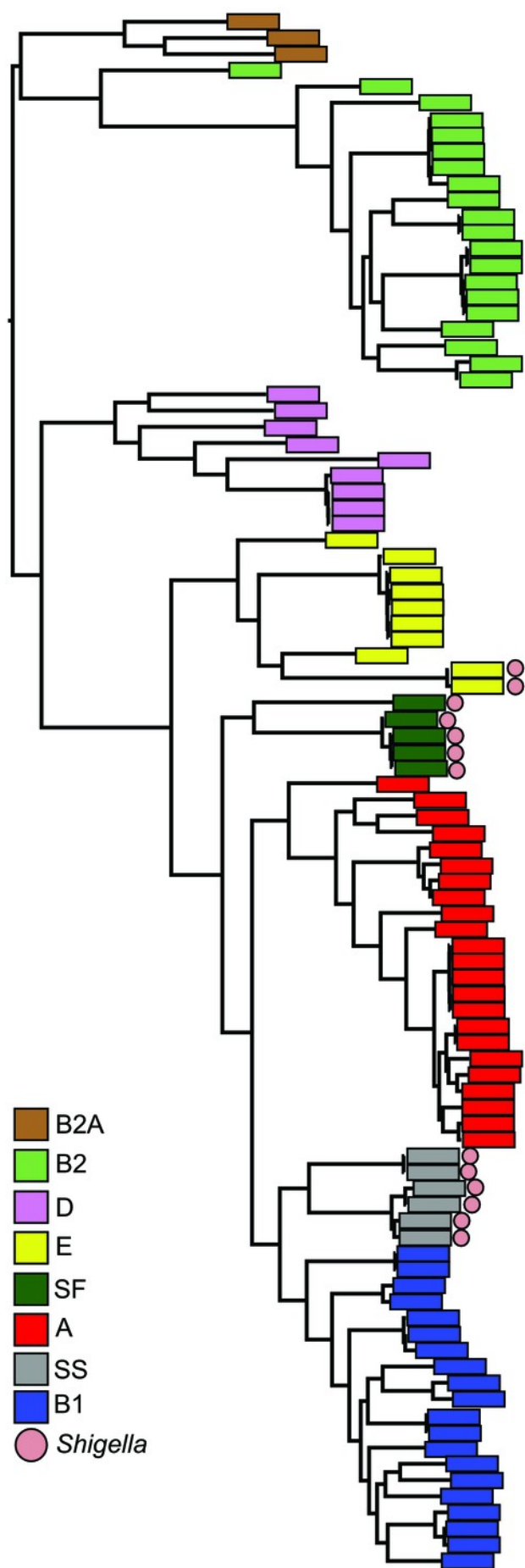
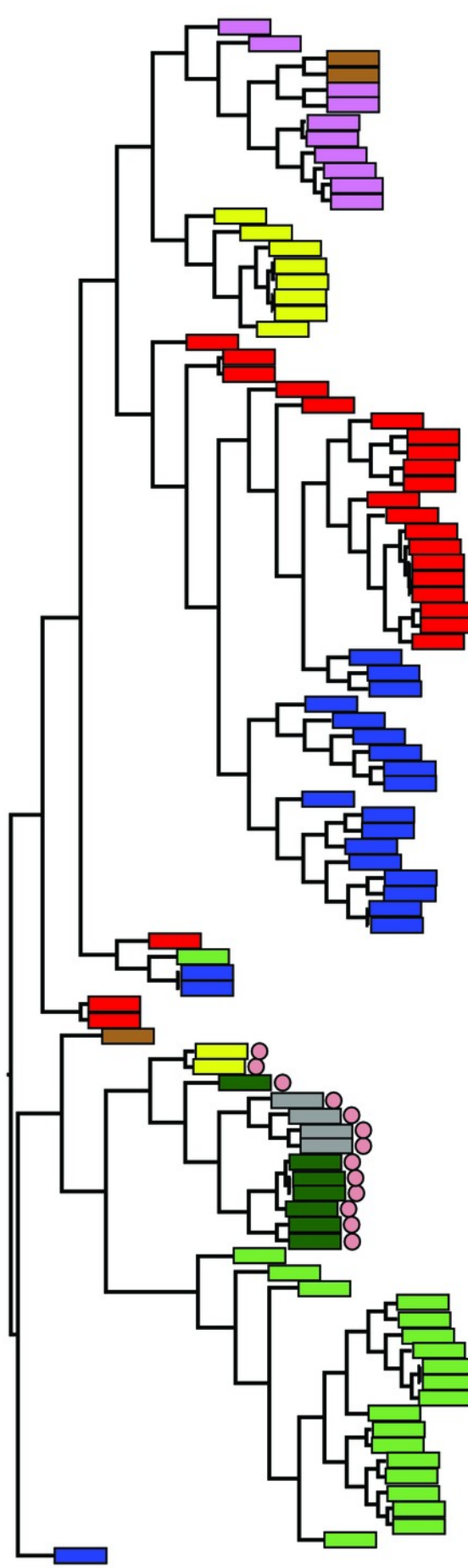
A comparison of 96 *Escherichia coli* / *Shigella* genomes between (Panel A) a core single nucleotide polymorphism (SNP) phylogeny or (Panel B) a cluster generated with the Multiple Experiment Viewer (Saeed et al. 2006) from BLAST Score Ratio (BSR) values that include the entire pan-genome. Colors applied to each classical *E. coli* phylogroup were applied to the SNP phylogeny and transferred to the BSR cladogram. *Shigella* genomes are marked with a red circle.

B2A	
B2	
D	
E	
SF	
A	
SS	
B1	
Shigella	

Figure 4

Pan-genome plots generated from LS-BSR output

Analyses were conducted on a set of 100 *Escherichia coli* genomes. The distribution of coding region sequences (CDSs) across the set of genomes surveyed is shown in Panel A. A supplemental script can be used to better understand the convergence of the core genome (Panel B), the accumulation of CDSs (Panel C), and the number of unique CDSs for each genome analyzed (Panel D); each analysis was conducted with 100 random sum-samplings and means are depicted with red diamonds.