

33 purpose. Our endeavor is to establish a benchmark method for short read
34 aligning tools.

35

36 De novo assembly of long sequence reads from Sanger-based sequencing
37 process produces reliable genomic sequences [27]. Sanger sequence reads are
38 typically 650 to 850 bases long while the NGS methods produce much shorter
39 reads that are 50-450 bases long. The reads are assembled to chromosomes
40 using well established algorithms, such as Celera Assembler[23], Arachne [2],
41 Atlas, CAP3 [15], Euler [26], PCAP[16], Phrap [11, 12], RePS [34], Phusion
42 [22]. Most of the assemblers follow the “overlap-layout-consensus” algorithmic
43 strategy [25] or are based on a de Bruijn graph[7]. Usually, the “overlap”
44 portion of the assembly process is the most computationally intensive. Us-
45 ing NGS reads for assembling whole genomes significantly reduces the costs
46 of genome sequencing.

47

48 However, most of the existing sequence assembly programs are not suffi-
49 cient enough for short sequence reads generated by NGS methods [24]. This
50 is partly because, the information contained in a short read is not sufficient
51 to find a position of a read in a genome [36]. Moreover, the number of
52 NGS reads is several orders of magnitude larger than the Sanger sequencing
53 reads. For a novel or little-explored genomes, this can prove very difficult.
54 Therefore, different algorithmic strategies more suitable for the short reads
55 assembly have been developed. Usage of two sets of restriction enzymes cre-
56 ates overlapping libraries and reduces errors. It is also possible to use long
57 and short reads together to take advantage of the low cost of NGS sequenc-
58 ing and computational unambiguity of long reads [35, 32, 8]. Finally, there
59 is an “alignment-layout-consensus” approach that uses a reference genome.
60 One of the implementations of this strategy is AMOS Comparative Assem-
61 bler [27].

62

63 When a reference genome is used to guide a sequence assembly, the
64 quality of the resulting assembly depends on the specific algorithm used,
65 on frequency of repeats in the pair of genomes, and evolutionary distance
66 between them. In addition, insertions in the target genome cannot be assem-
67 bled using the “alignment-layout-consensus” approach and presence of re-
68 arrangements will negatively affect the quality of assembled contigs [27]. It
69 has been demonstrated [27] that the “alignment-layout-consensus” approach
70 works well for a pair of strains of the same bacteria (92-94% coverage of the
71 target genome), but fails for more distinct sequences (11.4% coverage of the
72 target genome using more diverse organisms such as *Streptococcus agalactiae*

73 vs. *Streptococcus pyogenes*). *S. pyogenes* is a human pathogen, exclusively
74 adapted to the human host, and *S. agalactiae* is one of the principal causes
75 of bovine *Streptococcal mastitis* [21]. An array of computationally efficient
76 tools for mapping of short reads onto reference genomes, such as SOAP,
77 Bowtie, SHRiMP and BWA, has been developed. Well-established sequence
78 alignment tools like BLAST [1] can also handle short reads alignment.

79
80 It is important to determine the limits of applicability of the reference-
81 based alignment method depending on the divergence between the reference
82 and target species. In this paper we chose a simulation approach using global
83 whole genome alignments as gold standards. Simulation enables us to gen-
84 erate “NGS reads” of arbitrary length without investing in sequencing, map
85 them to a reference genome and assess the correctness of a mapped position.
86 To estimate error rate of these programs we propose a benchmark, which
87 uses the large-scale alignment between syntenic regions of genome sequences
88 as the true alignment. The aligned fragments of the whole genome alignment
89 were cut into short sequence ‘reads and the ability of different programs to
90 reproduce the true alignment using these reads was tested. This proposed
91 benchmark is a convenient way to select programs that are most suitable
92 for the reference-based genome assembly. It gives clear, realistic and robust
93 estimates of the accuracy of the alignment programs. The benchmark also
94 defines the limits of sequence similarities for selecting a reference genome.

95
96 In this paper, we compare performance of the five popular freely avail-
97 able alignment programs using whole-genome alignment between between
98 several strains of model bacteria *Escherichia coli* and between *E. coli* and
99 several species of *Salmonella*. We focused our analysis on bacterial species
100 for a number of reasons. They have manageable size genomes, variety of
101 nucleotide composition, and alignment of bacterial reads to genomes is es-
102 sential for environmental and clinical applications, annotation of variants,
103 determination of toxicity, drug resistance and pathogenicity of the analyzed
104 strain [3],[33].

105

106 Proposed Methods

107 We propose the following procedures to institute the benchmark method.
108 To evaluate the effectiveness of an alignment tool, we propose to compare
109 the alignments done by the tool with a gold standard alignment from other
110 independent sources. Researchers at various laboratories have invested ef-
111 fort in obtaining a consensus global alignment among several model species.
112 We intend to make use of these alignments to achieve a single benchmark
113 score for a given tool.

114
115 Our procedure starts by extracting the reference genome and query
116 genome from a peer reviewed global alignment. We call this the “Gold
117 Standard Alignment (GSA)”. By removing all the gaps from aligned se-
118 quences, we form the complete genome for reference and query genome. We
119 then split the query sequence into short reads of variant base pair lengths.
120 The philosophy is that, a “perfect” tool *per se*, will be able to align these
121 small sequence fragments to their accurate alignment positions within the
122 reference genome, replicating the results of GSA. The precision rate close to
123 1.0 will present a “near perfect” tool [28].

124
125 Different alignment tool produces alignment results in different format.
126 Our procedure do not discriminate the tools based on the tool’s own claim
127 of accuracy. For example, the E-value reported by the BLAST tool is not
128 carried towards the result of our scoring. We collect an information set, \mathbb{R}
129 from the alignment results containing: (i) read id ($r(n)$), (ii) reference se-
130 quence identifier (ref), (iii) start position of the read (stp). This information
131 is then compared with their counterparts from the GSA.

132
133 To evaluate the quality of mapping of reads to the reference genome, we
134 used a scoring method formed of True positives (TP), False Positive (FP),
135 False Negative (FN). When a short read or fragment is mapped exactly to
136 the same position on the reference genome as defined by GSA, we award one
137 point towards TP. If a fragment is mapped to a different position than the
138 one defined by the global alignment, a penalty is awarded to FP. However, if
139 the candidate tool failed to align a fragment to the correct location as GSA,
140 then a penalty point is awarded to FN.

141
142 To conclude benchmark of a candidate tool, we use Rijsbergen’s $F1$ score
143 as a measure of test accuracy [28].

144

145 We used true positive rate (r) and positive predictive rate (p), to com-
146 pute $F1$ score. Sensitivity or true positive rate, alternatively called as Recall.
147 A recall measures the probability of actually mapped reads. The true posi-
148 tive rate is computed by dividing the total number of correct results by the
149 number of alignments that were expected: $r = TP/(TP + FN)$.

150
151 Positive predictive value or alternatively called as Precision. Precision
152 measures the probability of positively mapped reads by dividing the total
153 number of correct alignments by total number of alignments detected by the
154 tool: $p = TP/(TP + FP)$.

155
156 The $F1$ score can be interpreted as a weighted average of the precision
157 and recall, where an $F1$ score reaches its best value at 1 and worst score at
158 0:

$$159 \quad F1 = 2 \times ((p \times r)/(p + r)).$$

160

Algorithm 1: Benchmarking of a list of NGS Short Reads Aligner

Data: **GSA:** Gold Standard Alignment between two sequences

Model, M: Reference genome of GSA

Query, Q: Query genome of GSA

Tools, T: List of the candidate tools

Initialization;

Data Preparation: Simulate short reads, $q(n) \subset Q$ of variant
bp lengths $n \in \{50, 100, 150, 400\}$;

foreach $t \in T$ **do**

foreach $q(n)$ **do**

 Align the reads to the model genome;

 From new alignment results generate $\mathbb{R} \leftarrow \{q(n), ref, stp\}$:

 Compare \mathbb{R} with GSA and produce a set $S \leftarrow \{TP, FP, FN\}$
 where

 True Positive Rate, $r \leftarrow \frac{TP}{TP+FN}$;

 Positive Predictive Value, $\rho \leftarrow \frac{TP}{TP+FP}$;

 Rijsbergen's accuracy measurement score, $\mathbb{F}_1 = 2 * \frac{\rho * r}{\rho + r}$

end

end

Result: Benchmark Score, \mathbb{F}_1

Table 1: List of paired strains and their whole genome alignment statistics

Genome	Accession	Identity	Al. Length
<i>S. enterica</i> Typhi Ty2	NC_004631.1	56.58	29480
<i>S. enterica</i> Typhi CT18	NC_003198.1	54.43	29159
<i>S. typhimurium</i> LT2	NC_003197.1	58.02	29025
<i>S. enterica</i> Paratyphi-A SARB42	NC_006511.1	52.91	32221
<i>E. coli</i> O157:H7 EDL933	NC_002655.2	76.38	34335
<i>E. coli</i> K12	NC_000913.2	79.48	38457
<i>E. coli</i> Sakai O157:H7	NC_002695.1	77.46	34316

161 Implementation

162 We designed an experiment using model species *Escherichia sp.* and
163 *Salmonella sp.* For our test cases we used pre-computed global alignments
164 of the following pairs of bacterial strains done by the VISTA consortium of
165 Lawrence Berkeley National Laboratory and Joint Genome Institute [13, 10].

166

167 For this experiment, we used seven pairs of alignments between *Es-*
168 *cherichia coli* O6 CFT073 and seven other strains from *Escherichia* and
169 *Salmonella*. Table 1 contains a list of paired strains together with whole
170 genome alignment statistics. Average percent identity is calculated as the
171 number of identical nucleotides divided by the alignment length. Aver-
172 age alignment length computed as from all fragments in the corresponding
173 whole-genome alignment.

174 GSA Selection Justification

175 We chose VISTA global alignments as GSA as the used technique gener-
176 ates long continuous DNA fragments of Orthologous genomics intervals.
177 VISTA uses a combination of global and local alignment methods consisting
178 of three steps; (a) obtaining a map of large blocks of conserved synteny be-
179 tween the two species by applying Shuffle-LAGAN global chaining algorithm
180 [5] to local alignments by translated BLAT [17]; (b) using Supermap [9], the
181 fully symmetric whole-genome extension to the Shuffle-LAGAN [4], and (c)
182 applying Shuffle-LAGAN the second time on each syntenic block to obtain
183 a more fine-grained map of small-scale rearrangements.

184 Short Reads Simulation

185 As proposed in the method, to maintain consistency we used *Escherichia*
186 *coli* O6 CFT073 genome as a reference genome. We used the second genome

Table 2: List of alignment tools used

Tool Name	Version Used
BLAST+: NCBI Basic Local Alignment Search Tool	2.2.26
Bowtie 2: Bowtie Short Read Aligner	2.1.0
SHRiMP: SHort Read Mapping Package	2.2.3
SOAP2: Short Oligonucleotide Analysis Package	2.2.1
BWA: Burrows-Wheeler Aligner	0.7.0

187 from each pairings as the query genome. Using a simple R program, we
188 simulated short reads of lengths of n bp where $n=50, 100, 150, 400$ from the
189 reads. Each nucleotide was used as a start point of a new read as long as
190 they ended with a read of expected length (n bp).

191 *Selection of Candidate Tools*

192 There is a large number of alignment tools available in the public domain.
193 We intend to use most (if not all) of the tools to produce a comprehensive
194 benchmark database. However, for this case study we used the most popular
195 alignment tool, BLAST from NCBI and four other relatively new alignment
196 tools. Table 2 presents a list of the tools and their versions that was used.
197 To maintain consistency, we did not use the latest versions of all the tools
198 and rather dependent on the stable releases of the tools from a contemporary
199 release time.

200 All of the tools were used as-is and without modification. Default pa-
201 rameters were used and the user guides were consulted only to install and
202 run examples as recommended by developers.

203 Results and Discussion

204 The aim of the experiment was to examine how the tools perform with
205 reads of varying lengths. Very short reads of 50 base pairs and relatively
206 longer reads 400 base pairs were of special interest. We used the evolutionary
207 tree as a biological reference to observe accuracy of the benchmark. We
208 expect that, if the genomes are identical, all five candidate programs should
209 provide close alignments with high precision, yielding in F1 scores close to
210 1. Likewise, the tools are expected to yield a lower F1 scores for alignments
211 performed between more distant organisms.

212
213 Our experiment demonstrated limits of sequence similarity for differ-
214 ent programs. As expected, for alignments between various strains of same
215 species (*E. coli*), all programs performed reasonably well, with the excep-
216 tion of SOAP2. For shorter reads of 50bp and 100bp, all five candidate
217 tools demonstrated good F1 scores. However, as the reads lengths started
218 to increase, at 150bp and 400bp, SOAP and BWA did not stay in-par with
219 BLAST+ and SHRiMP.

220
221 For closely related genomes, BLAST+'s performance matched its repu-
222 tation, however, for distant genomes, the performance was rather poor. For
223 alignments between *Salmonella ep.* and *E. coli*, for reasonably shorter reads
224 (50-100bp), BLAST+ was outperformed by SHRiMP. As the read lengths
225 increased, BLAST+ showed a recovering trend.

226
227 In our experiment, Bowtie started with a below-par accuracy score for
228 short reads, and with the increase of reads lengths, the accuracy continued
229 to decrease.

230
231 In almost all cases SOAP2 ability was behind BLAST+ and SHRiMP,
232 which can be explained by the fact that, mapping of the reads in SOAP2
233 is mismatch-dependent. In an earlier study [36] it was observed that the
234 suboptimal hits reduce from 21% to 1%, when mismatch rate was changed
235 from 2 to 6 mismatches invoking the different behavior of the tool, which is
236 partly dependent on the mismatch. More recently, it has been demonstrated
237 that SOAP2 has a lower read mapping accuracy in meta-genome experiments
238 and it shows 256 significant differences in the coverage depth[], which agrees
239 with our findings.

240 For more distant species, SHRiMP performs significantly better. In al-
241 most all cases, SOAP showed the worst performance. Poor performance of

242 SOAP can be explained by the fact that mapping of the reads in SOAP
243 is mismatch-dependent. In an earlier study [14] it was observed that the
244 suboptimal hits reduce from 21% to 1%, when mismatch rate was changed
245 from 2 to 6 mismatches invoking the different behavior of the tool, which
246 partially depends on the mismatch. More recently, it was demonstrated that
247 SOAP has a lower read mapping accuracy in meta-genome experiments and
248 it shows significant differences in the coverage depth [20], which agrees with
249 our demonstrated results.

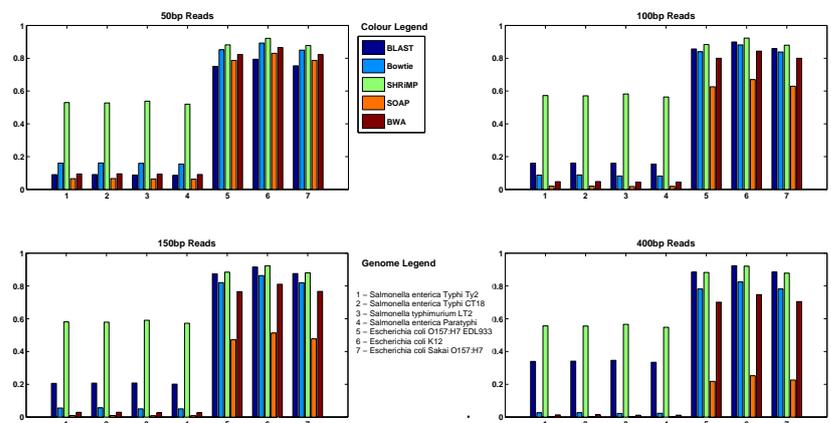


Figure 1: F1 Score for different reads lengths.

250 CONCLUSIONS

251 We developed and experimented a benchmark strategy to assess the cor-
252 rectness of alignments produced by different tools. We tested our method
253 on five tools and on a set of case study data. Our tested method proves
254 our hypothesis about closely related genomes. If the genomes are identical,
255 the tools perform well. If the genomes are distantly related by evolution
256 such as E.coli and Salmonella, the tools perform differently. In our case,
257 SHRiMP over-performs rest of the tools and SOAP performed reasonably
258 bad. BLAST and Bowtie performed well after SHRiMP. BLAST showed
259 consistent result as per our hypothesis. We conducted this experiment on a
260 set of data to establish the benchmark method. We aim to extend our study
261 for different species (i.e *Homo sapiens* vs *Pan troglodytes*) and adding a
262 range of different tools for comparative analysis.

263 **Availability of Supporting Data**

264 The gold standard global alignments were collected from VISTA website
265 available at :
266 <http://pipeline.lbl.gov/data/ecoli2/>.

267
268 Simulated reads and outputs of BLAST, Bowtie, SHRiMP and SOAP are
269 accessible via <http://cbio.uk/benchNGS/>. UNIX executable of a program
270 created using this algorithm is also available at the same link.

271 **ACKNOWLEDGEMENTS**

272 The authors are grateful to Alexandre Poliakov for his work on whole-
273 genome alignments.

274 **Funding**

275 FR was supported by HPC-Wales and Fujitsu Lab Europe. TT was
276 supported by grants from The National Institute for General Medical Studies
277 (GM068968), and the Eunice Kennedy Shriver National Institute of Child
278 Health and Human Development (HD070996).

279 **Authors Contributions**

280 TT and AN conceptualized the benchmark and proposed initial frame-
281 work. ID, FR, AK, MH designed the experiment, performed the case study,
282 generated results and prepared manuscript. FR, TT, ID and NA interpreted
283 the results and wrote the manuscript.

284 **References**

- 285 [1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J.,
286 et al., 1990. Basic local alignment search tool. *Journal of molecular*
287 *biology* 215 (3), 403–410.
- 288 [2] Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli,
289 E., Berger, B., Mesirov, J. P., Lander, E. S., Jan. 2002. Arachne: a
290 whole-genome shotgun assembler. *Genome Res* 12 (1), 177–89.
- 291 [3] Bolshoy, A., Salih, B., Cohen, I., Tatarinova, T., 2014. Ranking of
292 prokaryotic genomes based on maximization of sortedness of gene
293 lengths. *J Data Mining Genomics Proteomics* 5 (151).

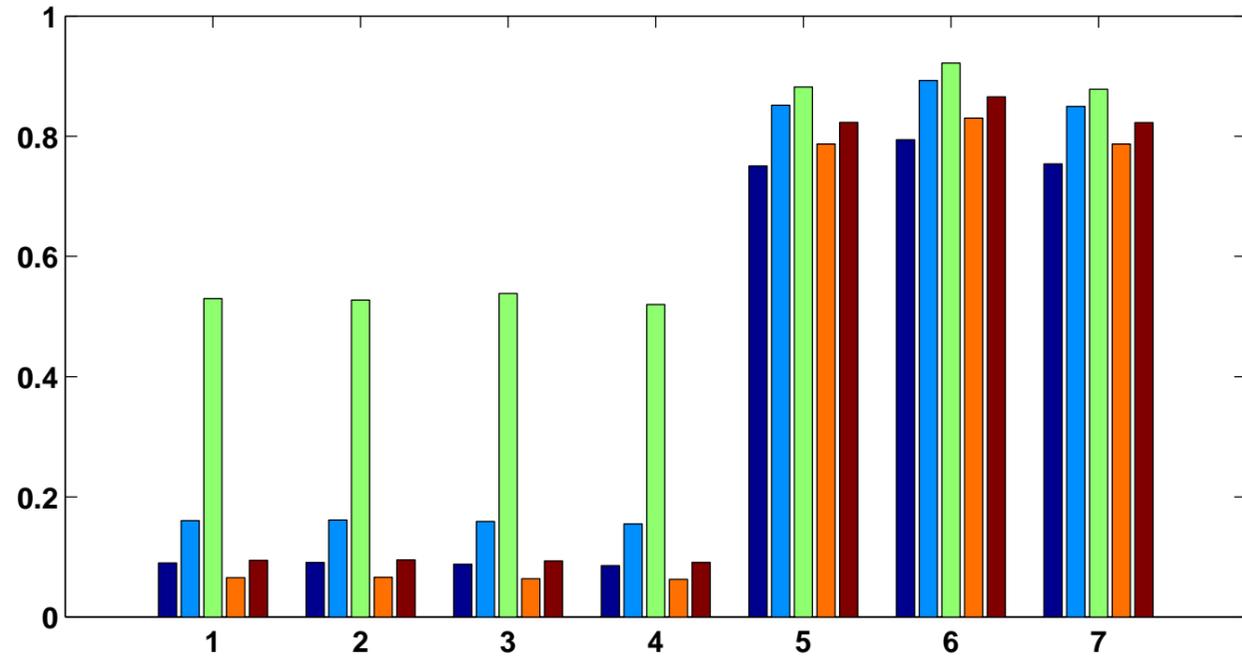
- 294 [4] Brudno, M., 2007. An introduction to the lagan alignment toolkit.
295 *Methods Mol Biol* 395 (4), 205–220.
- 296 [5] Brudno, M., Malde, S., Poliakov, A., Do, C. B., Couronne, O.,
297 Dubchak, I., Batzoglou, S., 2003. Glocal alignment: finding rearrange-
298 ments during alignment. *Bioinformatics* 19 (suppl 1), i54–i62.
- 299 [6] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J.,
300 Bealer, K., Madden, T., 2009. Blast+: architecture and applications.
301 *BMC Bioinformatics* 10 (1), 421.
302 URL <http://www.biomedcentral.com/1471-2105/10/421>
- 303 [7] de Bruijn, N. G., Jun. 1946. A Combinatorial Problem. *Koninklijke*
304 *Nederlandsche Akademie Van Wetenschappen* 49 (6), 758–764.
- 305 [8] DiGiustini, S., Liao, N., Platt, D., Robertson, G., Seidel, M., Chan,
306 S., Docking, T. R., Birol, I., Holt, R., Hirst, M., Mardis, E., Marra,
307 M., Hamelin, R., Bohlmann, J., Breuil, C., Jones, S., 2009. De novo
308 genome sequence assembly of a filamentous fungus using Sanger, 454
309 and Illumina sequence data. *Genome Biology* 10 (9), R94+.
310 URL <http://dx.doi.org/10.1186/gb-2009-10-9-r94>
- 311 [9] Dubchak, I., Poliakov, A., Kislyuk, A., Brudno, M., 2009. Multiple
312 whole-genome alignments without a reference organism. *Genome Re-*
313 *search* 19 (4), 682–689.
- 314 [10] Dubchak, I., Poliakov, A., Kislyuk, A., Brudno, M., 2009-04-01
315 00:00:00.0. Multiple whole-genome alignments without a reference or-
316 ganism. *Genome Research* 19 (5), 682–9.
- 317 [11] Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces
318 using phred. ii. error probabilities. *Genome Research* 8 (3), 186–194.
319 URL <http://genome.cshlp.org/content/8/3/186.abstract>
- 320 [12] Ewing, B., Hillier, L., Wendl, M. C., Green, P., 1998. Base-calling of
321 automated sequencer traces using phred.i. accuracy assessment. *Genome*
322 *Research* 8 (3), 175–185.
323 URL <http://genome.cshlp.org/content/8/3/175.abstract>
- 324 [13] Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., Dubchak, I.,
325 2004. VISTA: computational tools for comparative genomics. *Nucleic*
326 *Acids Research* 32 (Web-Server-Issue), 273–279.
327 URL <http://dx.doi.org/10.1093/nar/gkh458>

- 328 [14] Hatem, A., Bozdog, D., Toland, A., Catalyurek, U., 2013. Benchmark-
329 ing short sequence mapping tools. BMC Bioinformatics 14 (184).
- 330 [15] Huang, X., Madan, A., 1999. Cap3: A dna sequence assembly program.
331 Genome Research 9 (9), 868–877.
332 URL <http://genome.cshlp.org/content/9/9/868.abstract>
- 333 [16] Huang, X., Wang, J., Aluru, S., Yang, S.-P., Hillier, L., 2003. Pcap: A
334 whole-genome assembly program. Genome Research 13 (9), 2164–2170.
335 URL <http://genome.cshlp.org/content/13/9/2164.abstract>
- 336 [17] Kent, W. J., 4 2002. BLAT – The BLAST-Like Alignment Tool.
337 Genome Research 12 (4), 656–664.
- 338 [18] Langmead, B., Trapnell, C., Pop, M., Salzberg, S., 2009. Ultrafast
339 and memory-efficient alignment of short dna sequences to the human
340 genome. Genome Biology 10 (3), R25.
- 341 [19] Li, R., Li, Y., Kristiansen, K., 0004, J. W., 2008. Soap: short oligonu-
342 cleotide alignment program. Bioinformatics 24 (5), 713–714.
- 343 [20] Martin, J., Sykes, S., Young, S., Kota, K., Sanka, R., et al., 2012.
344 Optimizing read mapping to reference genomes to determine composi-
345 tion and species prevalence in microbial communities. PLoS ONE 7 (6),
346 e36427.
- 347 [21] Mickelson, M., 1966. Effert of lactoperoxidase and thiocyanate on the
348 growth of streptococcus pyogenes and streptococcus agalactiae in a
349 chemically defined culture medium. Journal of general microbiology
350 43 (1), 31–43.
- 351 [22] Mullikin, J. C., Ning, Z., 2003. The phusion assembler. Genome Re-
352 search 13 (1), 81–90.
353 URL <http://genome.cshlp.org/content/13/1/81.abstract>
- 354 [23] Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P.,
355 Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H. J.,
356 Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H.-H., Jordan,
357 C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C.,
358 Chen, L., Dunn, P. J., Lai, Z., Liang, Y., Nusskern, D. R., Zhan, M.,
359 Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D., Venter, J. C., 2000.
360 A whole-genome assembly of drosophila. Science 287 (5461), 2196–2204.
361 URL <http://www.sciencemag.org/content/287/5461/2196.abstract>

- 362 [24] Paszkiewicz, K., Studholme, D., 2010. De novo assembly of short se-
363 quence reads. *Briefings in bioinformatics* 11 (5), 457–472.
- 364 [25] Peltola, H., Soderlund, H., Ukkonen, E., 1984. Seqaid: a dna sequence
365 assembling program based on a mathematical model. *Nucleic Acids*
366 *Research* 12 (1), 307–321.
- 367 [26] Pevzner, P. A., Tang, H., Waterman, M. S., 2001. A new approach
368 to fragment assembly in dna sequencing. In: *Proceedings of the Fifth*
369 *Annual International Conference on Computational Biology. RECOMB*
370 *'01*. ACM, New York, NY, USA, pp. 256–267.
371 URL <http://doi.acm.org/10.1145/369133.369230>
- 372 [27] Pop, M., Phillippy, A., Delcher, A. L., Salzberg, S. L., 2004. Compara-
373 tive genome assembly. *Briefings in Bioinformatics* 5 (3), 237–248.
374 URL <http://bib.oxfordjournals.org/content/5/3/237.abstract>
- 375 [28] Rijsbergen, C. J. V., 1979. *Information Retrieval*, 2nd Edition.
376 Butterworth-Heinemann, Newton, MA, USA.
- 377 [29] Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A.,
378 Brudno, M., 2009. Shrimp: Accurate mapping of short color-space
379 reads. *PLoS Computational Biology* 5 (5).
- 380 [30] Schuster, S. C., 2007. Next-generation sequencing transforms today's
381 biology. *Nature Methods* 5 (1).
- 382 [31] Solovyev, V., Tatarinova, T., 2011. Towards the integration of ge-
383 nomics, epidemiological and clinical data. *Genome Medicine* 3 (7), 48.
384 URL <http://genomemedicine.com/content/3/7/48>
- 385 [32] Swaminathan, K., Alabady, M., Varala, K., De Paoli, E., Ho, I.,
386 Rokhsar, D., Arumuganathan, A., Ming, R., Green, P., Meyers, B.,
387 Moose, S., Hudson, M., 2010. Genomic and small rna sequencing of
388 miscanthus x giganteus shows the utility of sorghum as a reference
389 genome sequence for andropogoneae grasses. *Genome Biology* 11 (2),
390 R12.
391 URL <http://genomebiology.com/2010/11/2/R12>
- 392 [33] Tatarinova, T., Salih, B., Dien Bard, J., Cohen, I., Bolshoy, A., 2014.
393 Lengths of orthologous prokaryotic proteins are affected by evolutionary
394 factors. *BioMed Research International*.

- 395 [34] Wang, J., Wong, G. K.-S., Ni, P., Han, Y., Huang, X., Zhang, J., Ye, C.,
396 Zhang, Y., Hu, J., Zhang, K., Xu, X., Cong, L., Lu, H., Ren, X., Ren,
397 X., He, J., Tao, L., Passey, D. A., Wang, J., Yang, H., Yu, J., Li, S.,
398 2002. Repeats: A sequence assembler that masks exact repeats identified
399 from the shotgun data. *Genome Research* 12 (5), 824–831.
- 400 [35] Wicker, T., Schlagenhauf, E., Graner, A., Close, T., Keller, B., Stein,
401 N., 2006. 454 sequencing put to the test using the complex genome of
402 barley. *BMC Genomics* 7 (1), 275.
403 URL <http://www.biomedcentral.com/1471-2164/7/275>
- 404 [36] Young, A. L., Abaan, H. O., Zerbino, D., Mullikin, J. C., Birney, E.,
405 Margulies, E. H., Feb. 2010. A new strategy for genome assembly us-
406 ing short sequence reads and reduced representation libraries. *Genome*
407 *Research* 20 (2), 249–256.
408 URL <http://dx.doi.org/10.1101/gr.097956.109>

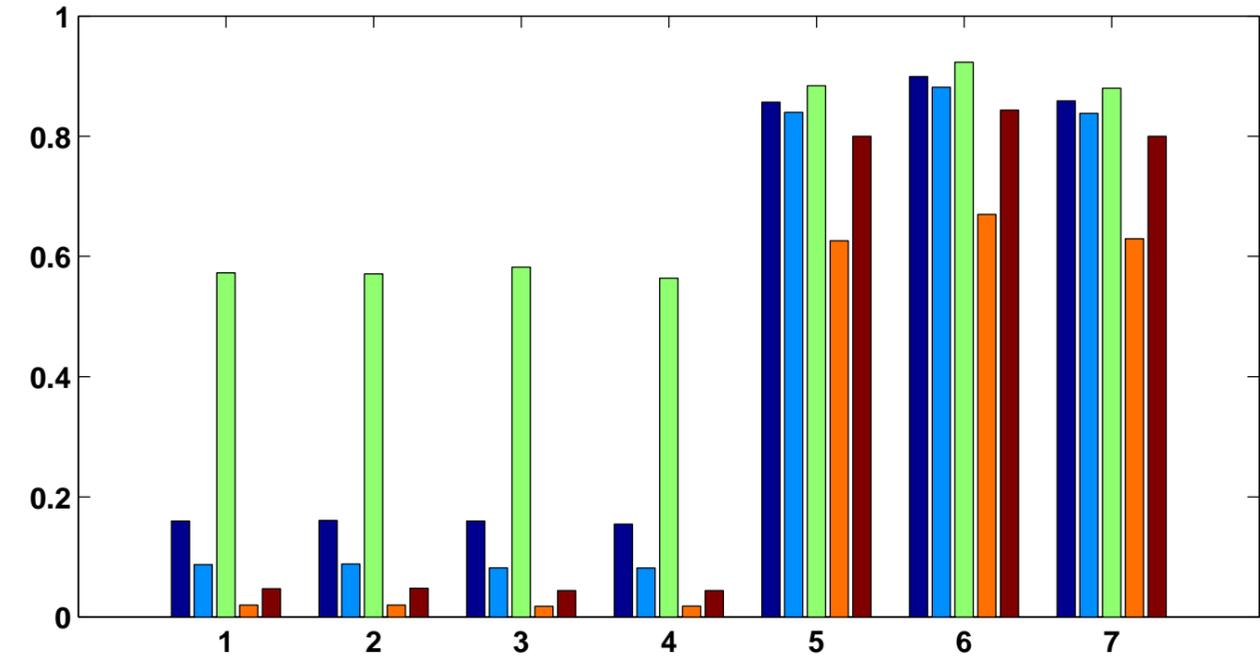
50bp Reads



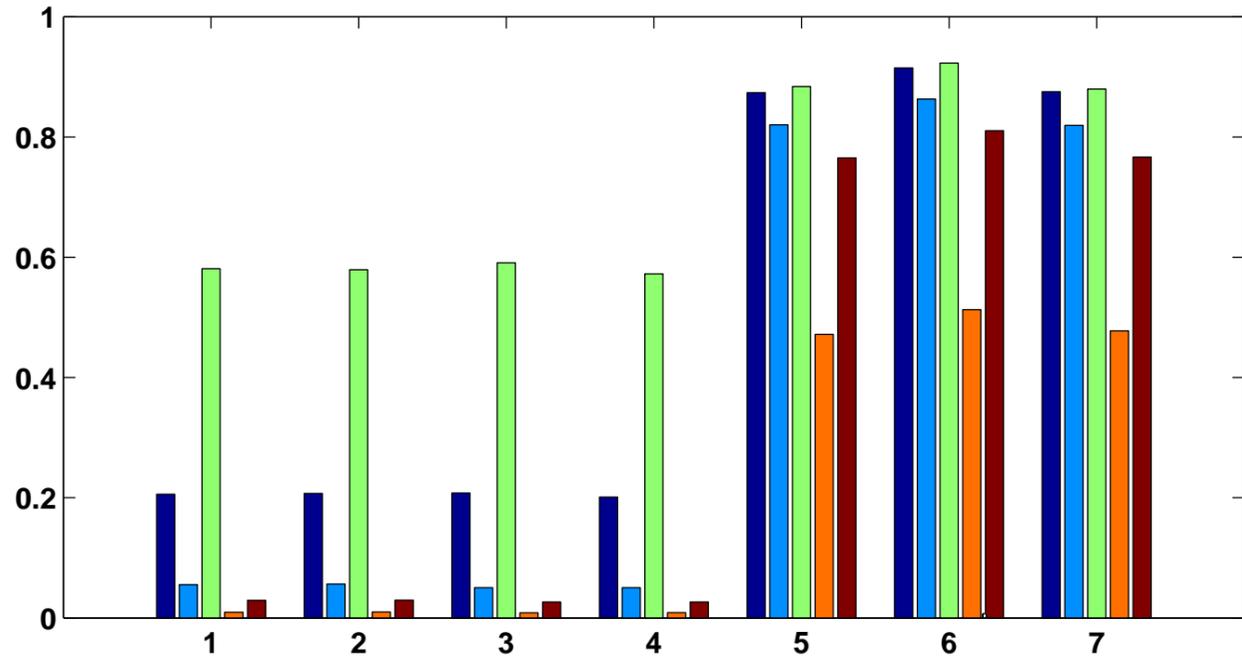
Colour Legend



100bp Reads



150bp Reads



Genome Legend

- 1 – *Salmonella enterica* Typhi Ty2
- 2 – *Salmonella enterica* Typhi CT18
- 3 – *Salmonella typhimurium* LT2
- 4 – *Salmonella enterica* Paratyphi
- 5 – *Escherichia coli* O157:H7 EDL933
- 6 – *Escherichia coli* K12
- 7 – *Escherichia coli* Sakai O157:H7

400bp Reads

