benchNGS : An approach to benchmark short reads alignment tools

Farzana Rahman[a], Mehedi Hassan[a], Alona Kryshchenko[b], Inna Dubchak[d], Nickolai Alexandrov[c], Tatiana V. Tatarinova[b]

[a] *University of South Wales, UK*
[b] *University of Southern California, USA*
[c] *International Rice Research Institute, Philippines*
[d] *Joint Genome Institute, USA*

Abstract

In the last decade a number of algorithms and associated software were developed to align next generation sequencing (NGS) reads to relevant reference genomes. The results of these programs may vary significantly, especially when the NGS reads are contain mutations not found in the reference genome. Yet there is no standard way to compare these programs and assess their biological relevance.

We propose a benchmark to assess accuracy of the short reads mapping based on the pre-computed global alignment of closely related genome sequences. In this paper we outline the method and also present a short report of an experiment performed on five popular alignment tools.

Keywords: Benchmark, NGS, alignment, short reads, BLAST, SOAP, Bowtie, bwa, SHRiMP

Introduction

Next Generation Sequencing (NGS) technologies provide fast and cost-effective alternatives to the established Sanger sequencing, and powers impressive scientific achievements and development of novel biological applications in medicine, ecology, forensics, epidemiology and other fields of science [30, 31]. High throughput NGS technology comes with challenges in managing large datasets and the "big data" questions in biology. Open access publications and public domain data liberation, made way for development of a plethora of tools for analysis of these datasets. With hourly paid cloud-based computing services being increasingly available, researchers are now in need of a benchmark method to select the perfect tool, that is fit for

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

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

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

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

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

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

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

Proposed Methods

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

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

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

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

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

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

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

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

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

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

Data: **GSA**: **G**old **S**tandard **A**lignment between two sequences
Model, $\mathbb{M}$: Reference genome of GSA
Query, $\mathbb{Q}$: Query genome of GSA
Tools, $\mathbb{T}$: List of the candidate tools

Initialization;
Data Preparation: Simulate short reads, $q(n) \subset \mathbb{Q}$ of variant bp lengths $n \subset \{50, 100, 150, 400\}$;
foreach $t \subset \mathbb{T}$ **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
S. enterica Typhi Ty2	NC_004631.1	56.58	29480
S. enterica Typhi CT18	NC_003198.1	54.43	29159
S. typhimurium LT2	NC_003197.1	58.02	29025
S. enterica Paratyphi-A SARB42	NC_006511.1	52.91	32221
E. coli O157:H7 EDL933	NC_002655.2	76.38	34335
E. coli K12	NC_000913.2	79.48	38457
E. coli Sakai O157:H7	NC_002695.1	77.46	34316

Implementation

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

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

GSA Selection Justification

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

Short Reads Simulation

As proposed in the method, to maintain consistency we used *Escherichia 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

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

Selection of Candidate Tools

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

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

Results and Discussion

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

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

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

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

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

For more distant species, SHRiMP performs significantly better. In almost all cases, SOAP showed the worst performance. Poor performance of

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

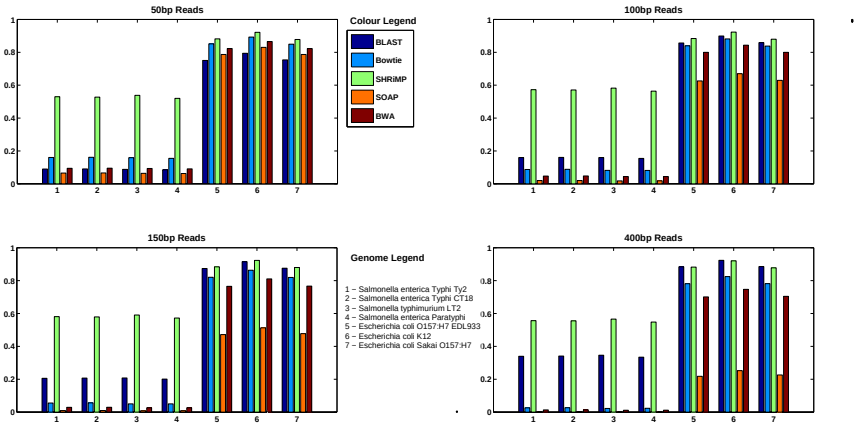


Figure 1: F1 Score for different reads lengths.

CONCLUSIONS

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

Availability of Supporting Data

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

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

ACKNOWLEDGEMENTS

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

Funding

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

Authors Contributions

TT and AN conceptualized the benchmark and proposed initial framework. ID, FR, AK, MH designed the experiment, performed the case study, generated results and prepared manuscript. FR, TT, ID and NA interpreted the results and wrote the manuscript.

References

[1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., et al., 1990. Basic local alignment search tool. Journal of molecular biology 215 (3), 403–410.

[2] Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. P., Lander, E. S., Jan. 2002. Arachne: a whole-genome shotgun assembler. Genome Res 12 (1), 177–89.

[3] Bolshoy, A., Salih, B., Cohen, I., Tatarinova, T., 2014. Ranking of prokaryotic genomes based on maximization of sortedness of gene lengths. J Data Mining Genomics Proteomics 5 (151).

[4] Brudno, M., 2007. An introduction to the lagan alignment toolkit. Methods Mol Biol 395 (4), 205–220.

[5] Brudno, M., Malde, S., Poliakov, A., Do, C. B., Couronne, O., Dubchak, I., Batzoglou, S., 2003. Glocal alignment: finding rearrangements during alignment. Bioinformatics 19 (suppl 1), i54–i62.

[6] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T., 2009. Blast+: architecture and applications. BMC Bioinformatics 10 (1), 421.
URL http://www.biomedcentral.com/1471-2105/10/421

[7] de Bruijn, N. G., Jun. 1946. A Combinatorial Problem. Koninklijke Nederlandsche Akademie Van Wetenschappen 49 (6), 758–764.

[8] DiGuistini, S., Liao, N., Platt, D., Robertson, G., Seidel, M., Chan, S., Docking, T. R., Birol, I., Holt, R., Hirst, M., Mardis, E., Marra, M., Hamelin, R., Bohlmann, J., Breuil, C., Jones, S., 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biology 10 (9), R94+.
URL http://dx.doi.org/10.1186/gb-2009-10-9-r94

[9] Dubchak, I., Poliakov, A., Kislyuk, A., Brudno, M., 2009. Multiple whole-genome alignments without a reference organism. Genome Research 19 (4), 682–689.

[10] Dubchak, I., Poliakov, A., Kislyuk, A., Brudno, M., 2009-04-01 00:00:00.0. Multiple whole-genome alignments without a reference organism. Genome Research 19 (5), 682–9.

[11] Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred. ii. error probabilities. Genome Research 8 (3), 186–194.
URL http://genome.cshlp.org/content/8/3/186.abstract

[12] Ewing, B., Hillier, L., Wendl, M. C., Green, P., 1998. Base-calling of automated sequencer traces usingphred.i. accuracyassessment. Genome Research 8 (3), 175–185.
URL http://genome.cshlp.org/content/8/3/175.abstract

[13] Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., Dubchak, I., 2004. VISTA: computational tools for comparative genomics. Nucleic Acids Research 32 (Web-Server-Issue), 273–279.
URL http://dx.doi.org/10.1093/nar/gkh458

[14] Hatem, A., Bozdag, D., Toland, A., Catalyurek, U., 2013. Benchmarking short sequence mapping tools. BMC Bioinformatics 14 (184).

[15] Huang, X., Madan, A., 1999. Cap3: A dna sequence assembly program. Genome Research 9 (9), 868–877.
URL http://genome.cshlp.org/content/9/9/868.abstract

[16] Huang, X., Wang, J., Aluru, S., Yang, S.-P., Hillier, L., 2003. Pcap: A whole-genome assembly program. Genome Research 13 (9), 2164–2170.
URL http://genome.cshlp.org/content/13/9/2164.abstract

[17] Kent, W. J., 4 2002. BLAT – The BLAST-Like Alignment Tool. Genome Research 12 (4), 656–664.

[18] Langmead, B., Trapnell, C., Pop, M., Salzberg, S., 2009. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. Genome Biology 10 (3), R25.

[19] Li, R., Li, Y., Kristiansen, K., 0004, J. W., 2008. Soap: short oligonucleotide alignment program. Bioinformatics 24 (5), 713–714.

[20] Martin, J., Sykes, S., Young, S., Kota, K., Sanka, R., et al., 2012. Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. PLoS ONE 7 (6), e36427.

[21] Mickelson, M., 1966. Effert of lactoperoxidase and thiocyanate on the growth of streptococcus pyogenes and streptococcus agalactiae in a chemically defined culture medium. Journal of general microbiology 43 (1), 31–43.

[22] Mullikin, J. C., Ning, Z., 2003. The phusion assembler. Genome Research 13 (1), 81–90.
URL http://genome.cshlp.org/content/13/1/81.abstract

[23] Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H. J., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H.-H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., Lai, Z., Liang, Y., Nusskern, D. R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D., Venter, J. C., 2000. A whole-genome assembly of drosophila. Science 287 (5461), 2196–2204.
URL http://www.sciencemag.org/content/287/5461/2196.abstract

[24] Paszkiewicz, K., Studholme, D., 2010. De novo assembly of short sequence reads. Briefings in bioinformatics 11 (5), 457–472.

[25] Peltola, H., Soderlund, H., Ukkonen, E., 1984. Seqaid: a dna sequence assembling program based on a mathematical model. Nucleic Acids Research 12 (1), 307–321.

[26] Pevzner, P. A., Tang, H., Waterman, M. S., 2001. A new approach to fragment assembly in dna sequencing. In: Proceedings of the Fifth Annual International Conference on Computational Biology. RECOMB '01. ACM, New York, NY, USA, pp. 256–267.
URL http://doi.acm.org/10.1145/369133.369230

[27] Pop, M., Phillippy, A., Delcher, A. L., Salzberg, S. L., 2004. Comparative genome assembly. Briefings in Bioinformatics 5 (3), 237–248.
URL http://bib.oxfordjournals.org/content/5/3/237.abstract

[28] Rijsbergen, C. J. V., 1979. Information Retrieval, 2nd Edition. Butterworth-Heinemann, Newton, MA, USA.

[29] Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A., Brudno, M., 2009. Shrimp: Accurate mapping of short color-space reads. PLoS Computational Biology 5 (5).

[30] Schuster, S. C., 2007. Next-generation sequencing transforms today's biology. Nature Methods 5 (1).

[31] Solovyev, V., Tatarinova, T., 2011. Towards the integration of genomics, epidemiological and clinical data. Genome Medicine 3 (7), 48.
URL http://genomemedicine.com/content/3/7/48

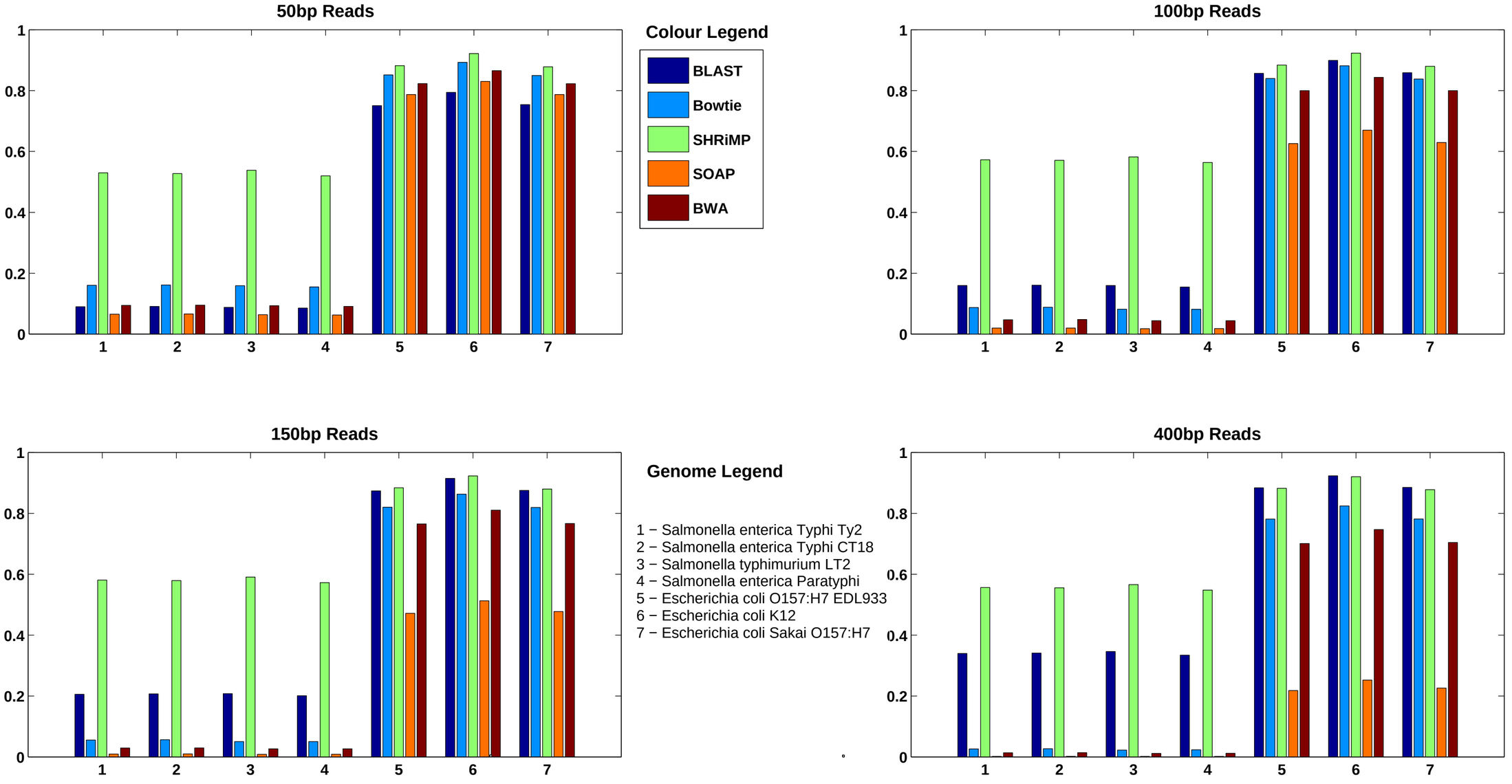
[32] Swaminathan, K., Alabady, M., Varala, K., De Paoli, E., Ho, I., Rokhsar, D., Arumuganathan, A., Ming, R., Green, P., Meyers, B., Moose, S., Hudson, M., 2010. Genomic and small rna sequencing of miscanthus x giganteus shows the utility of sorghum as a reference genome sequence for andropogoneae grasses. Genome Biology 11 (2), R12.
URL http://genomebiology.com/2010/11/2/R12

[33] Tatarinova, T., Salih, B., Dien Bard, J., Cohen, I., Bolshoy, A., 2014. Lengths of orthologous prokaryotic proteins are affected by evolutionary factors. BioMed Research International.

[34] Wang, J., Wong, G. K.-S., Ni, P., Han, Y., Huang, X., Zhang, J., Ye, C., Zhang, Y., Hu, J., Zhang, K., Xu, X., Cong, L., Lu, H., Ren, X., Ren, X., He, J., Tao, L., Passey, D. A., Wang, J., Yang, H., Yu, J., Li, S., 2002. Reps: A sequence assembler that masks exact repeats identified from the shotgun data. Genome Research 12 (5), 824–831.

[35] Wicker, T., Schlagenhauf, E., Graner, A., Close, T., Keller, B., Stein, N., 2006. 454 sequencing put to the test using the complex genome of barley. BMC Genomics 7 (1), 275.
URL `http://www.biomedcentral.com/1471-2164/7/275`

[36] Young, A. L., Abaan, H. O., Zerbino, D., Mullikin, J. C., Birney, E., Margulies, E. H., Feb. 2010. A new strategy for genome assembly using short sequence reads and reduced representation libraries. Genome Research 20 (2), 249–256.
URL `http://dx.doi.org/10.1101/gr.097956.109`


50bp Reads
100bp Reads
150bp Reads
400bp Reads
Colour Legend
BLAST
Bowtie
SHRiMP
SOAP
BWA
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