

1 **Title:** Bacterial genome reduction as a result of short read sequence assembly

2
3 **Authors:** Charles H.D. Williamson, Andrew Sanchez, Adam Vazquez, Joshua
4 Gutman, Jason W. Sahl

5
6 **Abstract**

7
8 High-throughput comparative genomics has changed our view of bacterial
9 evolution and relatedness. Many genomic comparisons, especially those
10 regarding the accessory genome that is variably conserved across strains in a
11 species, are performed using assembled genomes. For completed genomes, an
12 assumption is made that the entire genome was incorporated into the genome
13 assembly, while for draft assemblies, often constructed from short sequence
14 reads, an assumption is made that genome assembly is an approximation of the
15 entire genome. To understand the potential effects of short read assemblies on
16 the estimation of the complete genome, we downloaded all completed bacterial
17 genomes from GenBank, simulated short reads, assembled the simulated short
18 reads and compared the resulting assembly to the completed assembly.
19 Although most simulated assemblies demonstrated little reduction, others were
20 reduced by as much as 25%, which was correlated with the repeat structure of
21 the genome. A comparative analysis of lost coding region sequences
22 demonstrated that up to 48 CDSs or up to ~112,000 bases of coding region
23 sequence, were missing from some draft assemblies compared to their finished
24 counterparts. Although this effect was observed to some extent in 32% of
25 genomes, only minimal effects were observed on pan-genome statistics when
26 using simulated draft genome assemblies. The benefits and limitations of using
27 draft genome assemblies should be fully realized before interpreting data from
28 assembly-based comparative analyses.

29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

1 **Introduction**

2
3 Advances in DNA sequencing technologies have allowed for large-scale
4 whole genome sequencing of bacterial genomes. Short read technologies, such
5 as those employed on the Illumina sequencing platforms, have facilitated high-
6 throughput analyses of organisms for the purposes of comparative genomics (1),
7 phylogeography (2), and association of genomic attributes with antimicrobial
8 resistance (3). While reference-guided methods, including the identification of
9 single nucleotide polymorphisms (SNPs), are important for understanding
10 population genetics (4), many analyses are typically performed with assembled
11 genomes making genome assembly an important and standard method in the
12 analysis of bacterial organisms.

13 Studies that rely on assembled genomes include analyzing the conservation
14 of genomic features within a set of isolates and estimating core and pan-
15 genomes. Core and pan-genome analyses, introduced by Tettelin and
16 colleagues (5), have been applied to many bacterial species (6), and a number of
17 tools have been developed to calculate and analyze the pan-genome (7-13). All
18 of these tools rely on assembled genomes (or protein/nucleotide sequences from
19 assemblies) as input. Most of the assembled genomes currently available in
20 public databases are draft assemblies. Of the approximately 80,000 bacterial
21 genomes available from NCBI on November 1, 2016, less than 6000 are
22 complete.

23 As assemblies generated from short read sequencing data have become an
24 integral part of many research projects, potential limitations of this type of data
25 must be considered. For instance, contaminating reads can be incorporated into
26 assemblies (14-16) requiring post-assembly screening and quality control.
27 Additionally, genome assemblies generated from short read technologies are
28 typically fragmented due to the inability of short reads (and insert regions) to
29 span large repeat regions of a genome (17), which often breaks assemblies into
30 multiple contigs. This fragmentation can drop genomic regions from an assembly,
31 which look like missing regions in comparative analyses. In this study, we
32 evaluated how well assemblies generated from short read data estimate
33 complete bacterial genomes.

34 **Methods and Materials**

35
36
37 **Complete Genomes Used.** We downloaded (September 16, 2016) all bacterial
38 genomes from GenBank, then filtered the genomes to only include completed
39 assemblies (n=5676). We then filtered out genomes that contained >10 non-
40 nucleotide characters (non A,T,G,C), which could indicate problems with genome
41 assembly (n=203). A complete list of genomes (n=5473) used in this study is
42 shown in Table S1.

43
44 **Read simulation.** Paired end illumina reads were simulated for each complete
45 genome using ART (18) vMountRainier with the following parameters: -ss MSv3 -
46 l 250 -f 75 -m 300 -s 30. Genomes were then assembled with SPAdes v3.7.1

1 (19) using the following parameters: -t 4 -k 21,33,55,77,99,127 -cov-cutoff auto -
2 careful -1 pair1 -2 pair2. Following assembly, genomes were polished with Pilon
3 v.1.7, using the following parameters: --threads 4 --fix all,amb. Contigs shorter
4 than 200bp were filtered from the assembly to stay consistent with GenBank
5 standards. The genome assembly was automated with the UGAP assembly
6 pipeline (<https://github.com/jasonsahl/UGAP>), which was run using the Slurm
7 management system on a high-performance computing (HPC) cluster at
8 Northern Arizona University.

9 In order to identify how well the simulated reads represented the completed
10 genomes, we mapped the reads to the completed genome with BWA-MEM (20).
11 The per base coverage was calculated with the GenomeCoverageBed method in
12 BEDTOOLS (21). The number of bases with a minimum coverage of 1 was then
13 divided by the total number of bases in the completed genome to calculate the
14 percent coverage of simulated reads across each genome.

15
16 **Genome validation.** In addition to simulated reads, we also analyzed a set of 49
17 complete, or near complete, genomes that have been assembled separately with
18 both Illumina and PacBio sequencing platforms (Table S2). To test the ability of
19 ART to simulate representative short sequencing reads, we ran the Illumina
20 reads through SPAdes using the same parameters as with the simulated reads.

21
22 **Genome size calculation.** For each genome, we summed the entire sequence
23 length across all sequences with a Python script
24 (<https://gist.github.com/jasonsahl/64d88d2858a915ee730b5f86e305e5d4>). We
25 divided the size of the simulated assembly by the size of the completed assembly
26 to determine the amount of the genome retained.

27
28 **Repeat characterization.** To identify the percentage of the genome associated
29 with repeated regions, we aligned each genome against itself with NUCmer (22).
30 We then divided the number of bases in repeated regions by the total length of
31 the genome to characterize the repeat percentage. The identification of repeat
32 regions was facilitated by methods implemented in the NASP pipeline (4). Using
33 default parameters, NUCmer is unable to detect repeats shorter than 21
34 nucleotides.

35
36 **Multi-locus Sequence Typing comparisons.** The sequence type of *E. coli* and
37 *S. aureus* assemblies was identified using the PubMLST system and a custom
38 script (https://github.com/jasonsahl/mlst_blast.git). Each allele was assigned if an
39 exact match to the database was observed.

40
41 **Comparative genomics.** To identify the impact of regions collapsed or lost
42 during the genome assembly using simulated reads, a large-scale Blast Score
43 Ratio (LS-BSR) (12, 23) analysis was performed. Coding regions were predicted
44 from the completed genome and the simulated genome with Prodigal (24). All
45 coding regions were clustered with USEARCH (25) at an ID of 0.9 and aligned
46 against both genomes with BLAT (26). The BSR values were then compared

1 between the simulated and the completed genome to identify the number and
2 combined length of regions that had a BSR value > 0.8 (~80% peptide identity
3 over 100% of the peptide length) in the completed genome and a BSR value <
4 0.4 in the simulated genome. These regions represent those that were lost from
5 the assembly and could confound comparative analysis using genome
6 assemblies of short read sequence data.

7
8 **Publicly available genomes.** To characterize the quality of all genomes from a
9 single species in public databases, all *Escherichia coli* genome assemblies
10 (n=4842) were downloaded on September 16, 2016. Genomes were assessed
11 for contig number and assembly size.

12 Results

13
14
15 **Extent of genome reduction using simulated, short read assemblies.** In
16 order to understand the effects of short read assembly on the retention of
17 sequence from bacterial genomes, we downloaded all completed genomes from
18 GenBank with fewer than 10 ambiguities (n=5473) (Table S1) and simulated
19 paired-end Illumina MiSeq reads with ART (18) at an average coverage of 75x.
20 We assembled all genomes with SPAdes as it performs well compared to other
21 assemblers (27), it recovers larger portions of reference genomes than other
22 short read assembly algorithms (28), and we wanted to keep the assembly
23 algorithm constant. The sizes of the complete and the simulated genomes were
24 compared to understand the extent of reduction due to assembly problems.
25 While the vast majority of the genome was recovered in most cases, some
26 genomes showed significant reduction due to short read assembly (Figure 1,
27 Table S3). The maximum percentage of observed genome reduction was
28 approximately 25% in *Orientia tsutsugamushi*, which has been described as
29 having one of the most duplicated genomes (29). In some cases, the simulated
30 genome assembly was slightly larger than the complete genome (maximum of
31 ~0.76% larger), which may be due to the presence of contigs in the simulated
32 genome that should have been merged during assembly.

33 We then calculated the breadth of coverage of the completed genome, at a
34 minimum depth of 1x, with simulated reads (Table S3). The breadth of coverage
35 was meant to estimate how well the simulated reads represented the complete
36 genome. Breadth of coverage values range from approximately 73% to 100%. A
37 correlation of breadth of coverage and genome reduction (correlation
38 coefficient=0.76) demonstrates that different methods (genome assembly and
39 short read mapping) return a similar result (Figure 2, Table S3).

40
41 **Genome reduction using actual sequence data.** To confirm that genome
42 reduction wasn't solely due to the short read simulation, a set of 49 complete or
43 near complete *Burkholderia* genomes (30) was compared to the same isolates
44 where the genomes were also sequenced on the Illumina MiSeq platform. When
45 the genome reduction percentages were compared between real and simulated
46 reads, similar results were observed (correlation coefficient=0.50) (Table 1). In

1 some cases, the Illumina assembly was larger than the completed genome,
2 which may be due to bleed over between multiplexed samples on the same
3 sequencing run (31) or assembly error. This analysis demonstrates that the
4 simulated short reads should be generally representative of the extent of genome
5 reduction across other species.

6
7 **Repeat structure of all genomes.** In order to understand the repeat structure of
8 each genome, NUCmer self-alignments were performed on all genomes and the
9 summed repeat regions were divided by the entire genome length. The results
10 demonstrate that several of the genomes with a high level of reduction were also
11 highly repetitive (Figure 3). In general, genomes with a low level of repeats also
12 had a low level of reduction. The inability to span repeats largely explains the
13 reduction in genome size following genome assembly. As mentioned above,
14 genome reduction is correlated to breadth of coverage (short read mapping),
15 which highlights the limitations of short reads in resolving repeats using
16 independent approaches.

17
18 **Draft genome assembly effects on comparative genomics.** The potential
19 effects of a reduced genome on comparative genomics was investigated using
20 LS-BSR. The number and length of regions that were missing from the simulated
21 genome was calculated (Table S3). In 3729 of 5473 queried genomes, there
22 were no coding regions (CDSs) that were missing from the simulated genome
23 compared to the completed genome, despite seeing simulated genome assembly
24 sizes that were up to 16% smaller than the completed genome. Of all simulated
25 genomes, 780 were missing more than one CDS identified in the complete
26 genome. The maximum number of CDSs missing from a simulated genome
27 compared to a completed genome was 48, while the maximum length of coding
28 region sequence lost in any genome was approximately 112,000 nucleotides.

29 Reads were aligned to CDSs identified in complete genome assemblies but
30 missing from simulated genome assemblies to determine if short read alignment
31 could be used to verify the presence or absence of CDSs in a genome. The
32 breadth of coverage was determined at a minimum depth of 1X as described
33 above. In two test cases (GCA_000017805.1, missing 48 CDSs in the simulated
34 genome; GCA_000147815.3, missing 8 CDSs corresponding to ~112,000
35 nucleotides), all missing CDSs were at least partially covered by simulated reads
36 (minimum of ~46% coverage breadth). Mapped reads provided 100% breadth of
37 coverage for 50 of the 56 CDSs evaluated for both genomes, which suggests
38 that read mapping is a valuable method for confirming the presence/absence of
39 potentially missing genomic features.

40
41 **Draft genome assembly effects on pan-genome calculations.** The effect of
42 genome reduction on core and pan genome calculations was identified in an
43 analysis of *Escherichia coli*, *Staphylococcus aureus*, and *Salmonella enterica*,
44 species for which numerous (>100) complete genomes are available. In each
45 case, the core genome was calculated with LS-BSR for coding region sequences
46 with a BSR value of > 0.8 across all genomes tested; in each case, the average

1 core genome was calculated across 10 replicates at each level of sampling. The
2 core genome results demonstrate the simulated and completed genomes
3 generally return a consistent core genome size (Figure 4). Additionally, the pan-
4 genome size was slightly larger using simulated reads, which is likely a result of
5 fragmented coding regions that appear to be separate sequences during the
6 clustering step in LS-BSR. The same general trends were observed across each
7 species.

9 **MLST comparisons between complete and simulated genome assemblies.**

10 The relationships between bacterial isolates has typically been performed with
11 multi-locus sequence type (MLST) approaches (32). To test the quality of
12 assembled genomes, we extracted the 7 genes from the *E. coli* and *S. aureus*
13 MLST schemes (33) and compared the sequence type (ST) calls between
14 finished and simulated genome assemblies. In both species, all called sequence
15 types matched between complete and simulated genome assemblies. This
16 demonstrates that high quality draft genome assemblies can often provide
17 important sequence type information for comparison to previous or future studies.

18
19 **Comparison of real and simulated data.** Although simulated draft genome
20 assemblies provide comparable MLST and core genome information, they don't
21 represent real data, which can be of variable quality. A comparison of contig
22 numbers between *E. coli* genomes downloaded from GenBank and simulated
23 assemblies generated in this study demonstrates this variability (Figure 5, panel
24 A). The genome size is also highly variable in the real data (Figure 5, panel B),
25 which could be due to either insufficient coverage or contamination with other
26 genomes. If strict filtering on real genome sequence data is implemented, then
27 much of this variation can and should be eliminated prior to comparative
28 analyses.

30 **Discussion**

31
32 Short read sequencing technologies have been key in understanding the
33 movement (34) and population structure (35) of bacterial species. Recent
34 advances in DNA sequencing now allow for the push button assembly of
35 bacterial genomes using long read sequencing approaches (36), which holds the
36 promise of automated and complete genome assembly even for highly duplicated
37 and repetitive genomes. However, due to cost limitations, many laboratories still
38 rely on short read technologies for high-throughput SNP identification and
39 genome assembly ensuring that short read applications will continue to be used
40 for large-scale comparative genomics. While benefits to these approaches exist
41 in the consensus calling of variants, limitations also exist due to the short nature
42 of the read composition, which depending on the length of the read and length of
43 the repeat, cannot span many repeat regions resulting in fragmented genome
44 assemblies. Previous work has demonstrated the effects of different genome
45 assembly algorithms on the recovery of a reference genome using short read
46 technologies (27, 37, 38), and the GAGE-B study (27) evaluated assemblies of

1 eight different bacterial species genomes with a number of different assemblers.
2 However, less is known about how the composition of genomes from diverse
3 species affects the ability to resolve the full genome with short read sequencing
4 technology by keeping the assembly algorithm constant. In this study, we
5 performed a comprehensive analysis of this issue through the assembly of
6 greater than 5000 finished bacterial genomes. The results demonstrate that
7 simulated short read assemblies recovered high percentages of most genomes;
8 however, significant genome reduction was observed in some highly repetitive
9 genomes, which has the ability to affect downstream comparative analyses.

10 Comparative genomics studies include the identification of genomic features
11 that are differentially conserved between genomes from isolates in the same or
12 closely-related species. These comparisons are important for identifying gene
13 differences that may be associated with diagnostics, virulence, or differential
14 phenotypes (39-44). Artifacts generated from the assembly of short read data
15 could potentially impact these sorts of comparisons. Our results indicate that
16 coding region sequences identified in simulated draft genome assemblies were
17 representative of the coding regions identified in complete genomes in most
18 cases. Thus, draft assemblies can provide important information on genomic
19 feature variation between strains, core and pan-genome comparisons, and
20 isolate relationships based upon MLST genes extracted from draft assemblies.

21 The results of this study also demonstrate that draft genome quality in public
22 repositories is variable and that quality control and filtering should be applied
23 prior to comparative genomics studies. The results also indicate that genome
24 reduction due to short read assembly can be a problem in downstream analyses
25 for some genomes, although the impacts are variable, and perhaps predictable
26 based on the repeat structure of a given genome. For large-scale comparative
27 analyses, results must be interpreted with these limitations in mind. If missing
28 genes are observed between groups of genomes, raw read mapping can be
29 used to verify the gene presence or absence, although short read mapping may
30 also suffer from some of the same limitations as short read genome assembly.
31 Additionally, complete genomes representing species or clades of interest can
32 provide a reference point for evaluating draft genome assemblies (e.g. provide
33 information about repeat structure). This study indicates that draft genome
34 assemblies generated from short read data often provide an acceptable
35 representation of a bacterial genome for many comparative genomics
36 applications.

37 38 **Acknowledgments**

39 This work was facilitated by the Monsoon High Performance Cluster (HPC)
40 resource at Northern Arizona University.

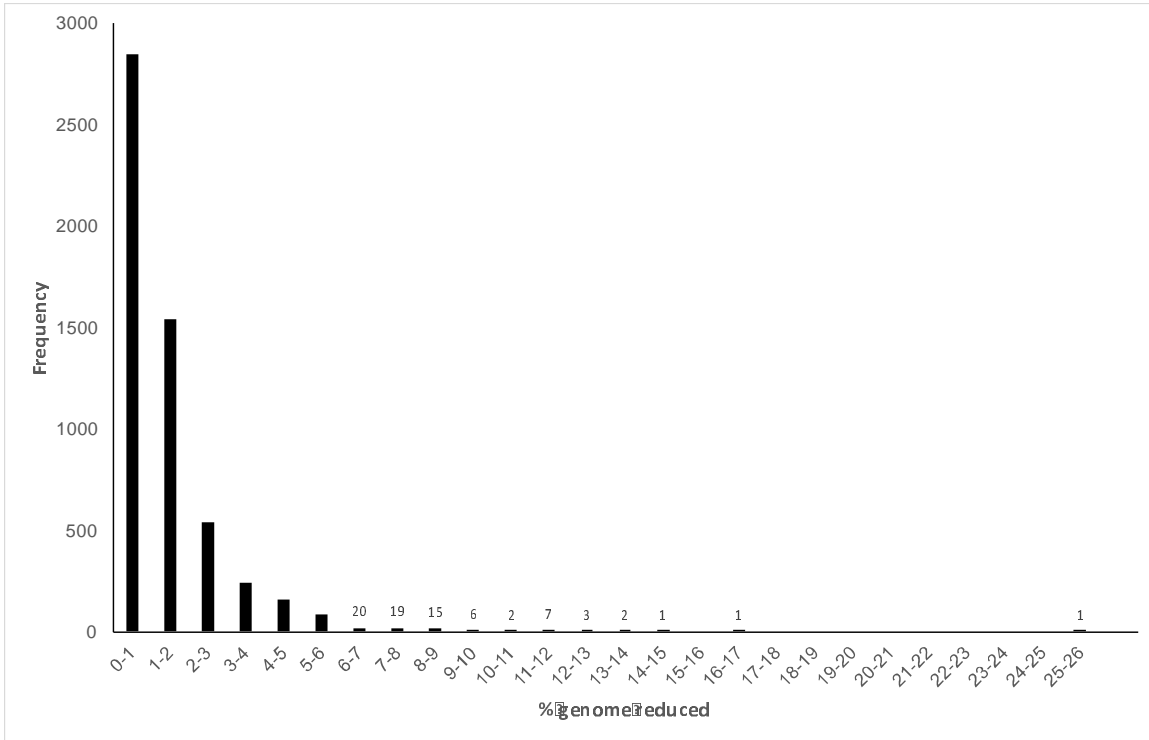
41

| Strain | PacBio Assembly Size | Illumina Assembly Size | Simulated Assembly Size | %reduction (Illumina) | %reduction (simulated) |
|--------|-------------------------|---------------------------|----------------------------|--------------------------|---------------------------|
| Bp7003 | 6590904 | 6498086 | 6511321 | 1.41 | 1.21 |
| Bp7004 | 7357522 | 7251364 | 7263334 | 1.44 | 1.28 |
| Bp7005 | 7159534 | 7122062 | 7120222 | 0.52 | 0.55 |
| Bp7006 | 6858431 | 6782684 | 6786320 | 1.10 | 1.05 |
| Bp7046 | 6708057 | 6592936 | 6610712 | 1.72 | 1.45 |
| Bp7047 | 6739513 | 6624496 | 6646837 | 1.71 | 1.38 |
| Bp7049 | 6785295 | 6673578 | 6683615 | 1.65 | 1.50 |
| Bp7097 | 6899618 | 6819885 | 6815275 | 1.16 | 1.22 |
| Bp7270 | 6787216 | 6701472 | 6651437 | 1.26 | 2.00 |
| Bp0071 | 7319559 | 7263748 | 7209107 | 0.76 | 1.51 |
| Bp0072 | 7135010 | 7072535 | 7077251 | 0.88 | 0.81 |
| Bp0073 | 6729272 | 6633203 | 6642243 | 1.43 | 1.29 |
| Bp0150 | 6857828 | 6807568 | 6812424 | 0.73 | 0.66 |
| Bp0379 | 6902331 | 6857306 | 6861916 | 0.65 | 0.59 |
| Bp0422 | 7961561 | 7902985 | 7909489 | 0.74 | 0.65 |
| Bp0713 | 7079942 | 7036491 | 7043175 | 0.61 | 0.52 |
| Bp2202 | 7337349 | 7298664 | 7307467 | 0.53 | 0.41 |
| Bp3330 | 7956735 | 7871833 | 7878640 | 1.07 | 0.98 |
| Bp6994 | 7273038 | 7209301 | 7216862 | 0.88 | 0.77 |
| Bp6997 | 7129510 | 7024776 | 7036295 | 1.47 | 1.31 |
| Bp7014 | 8170759 | 8126665 | 8127581 | 0.54 | 0.53 |
| Bp7021 | 6815241 | 6740025 | 6746707 | 1.10 | 1.01 |
| Bp7030 | 6330734 | 6503504 | 6304216 | -2.73 | 0.42 |
| Bp7031 | 7648890 | 7608554 | 7612789 | 0.53 | 0.47 |
| Bp7035 | 7424194 | 7378871 | 7381392 | 0.61 | 0.58 |
| Bp7043 | 8354771 | 8296564 | 8294334 | 0.70 | 0.72 |
| Bp7052 | 6767931 | 6704757 | 6702085 | 0.93 | 0.97 |
| Bp7053 | 6970128 | 6778809 | 6751479 | 2.74 | 3.14 |
| Bp7055 | 6950257 | 6771434 | 6750755 | 2.57 | 2.87 |
| Bp7062 | 7235050 | 7149035 | 7147490 | 1.19 | 1.21 |
| Bp7080 | 7250560 | 6938686 | 7161963 | 4.30 | 1.22 |
| Bp7344 | 6620927 | 6532626 | 6523523 | 1.33 | 1.47 |
| Bp7345 | 6835061 | 6735080 | 6747730 | 1.46 | 1.28 |
| Bp7347 | 6614472 | 6526402 | 6536593 | 1.33 | 1.18 |
| Bp7353 | 6764528 | 6692224 | 6699728 | 1.07 | 0.96 |
| Bp7422 | 8000745 | 7624755 | 7686199 | 4.70 | 3.93 |
| Bp7432 | 8028759 | 8096860 | 7956373 | -0.85 | 0.90 |
| Bp7434 | 8028759 | 7906481 | 7956220 | 1.52 | 0.90 |
| Bp7583 | 6883100 | 6810027 | 6817056 | 1.06 | 0.96 |
| Bp7621 | 7780595 | 7705087 | 7719455 | 0.97 | 0.79 |
| Bp7630 | 6891707 | 6797470 | 6822127 | 1.37 | 1.01 |
| Bp7634 | 6913771 | 6822861 | 6839604 | 1.31 | 1.07 |
| Bp7657 | 7583794 | 7518484 | 7536053 | 0.86 | 0.63 |
| Bp7702 | 7463455 | 7418001 | 7428963 | 0.61 | 0.46 |
| Bp7709 | 6845356 | 6764904 | 6787311 | 1.18 | 0.85 |
| Bp7064 | 7427587 | 7302761 | 7286670 | 1.68 | 1.90 |
| Bp7071 | 6784408 | 7273481 | 6721097 | -7.21 | 0.93 |

1
2
3

Table 1: Correlations between simulated and true assemblies

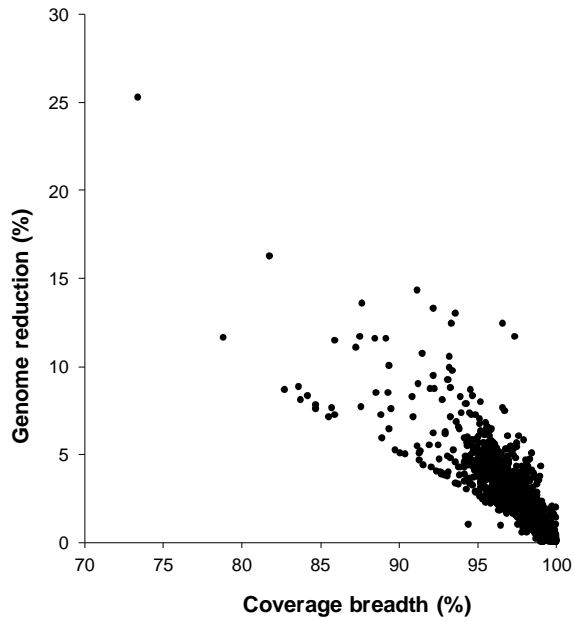
1
2
3



4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27

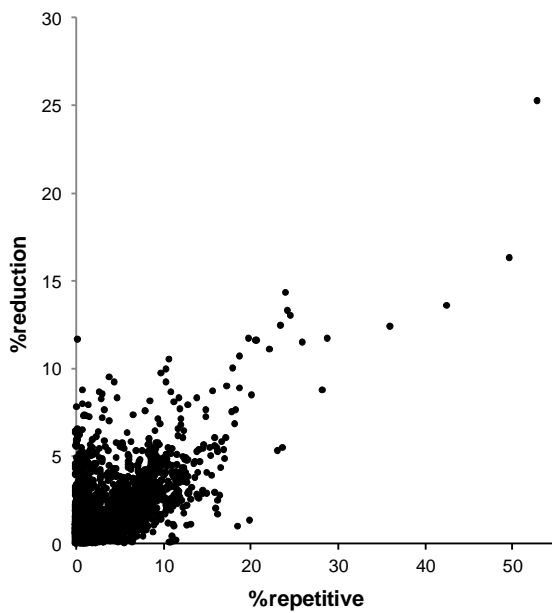
Figure 1: Frequency plot of the number of genomes against the extent of genome reduction.

1
2
3



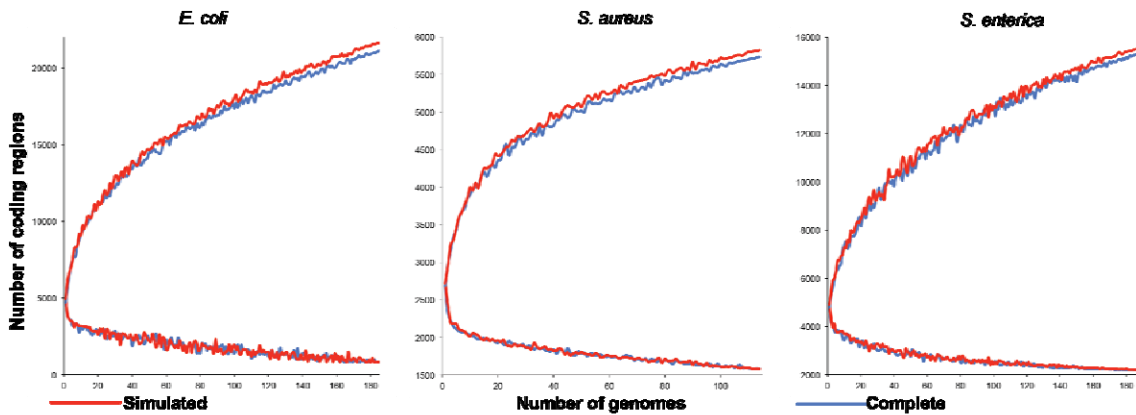
4
5
6
7
8
9

Figure 2: Breadth of coverage of simulated sequencing reads across complete genomes compared to the genome reduction of simulated genome assemblies compared to completed genomes.



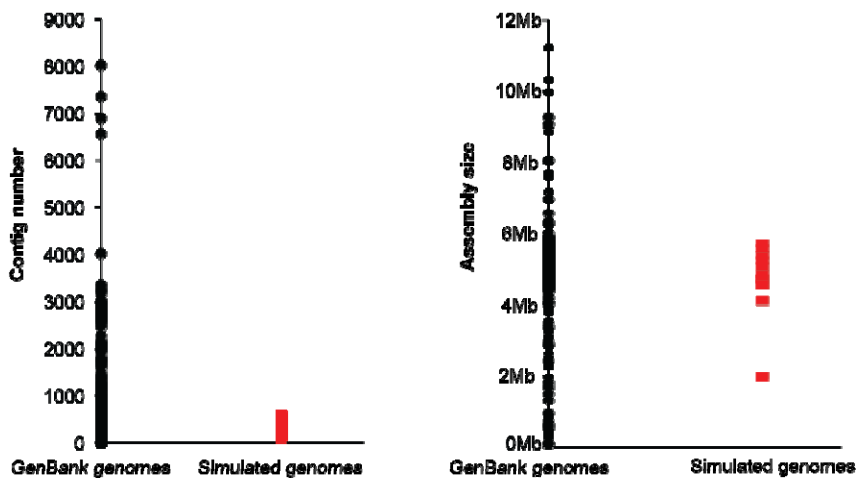
10
11
12
13
14
15

Figure 3: A plot of the % of the genome that is repetitive against the % of the genome that is reduced



1
2
3
4
5
6
7
8

Figure 4: Comparative pan-genome plots for 3 species with a large number of complete genomes. The plots either demonstrate the accumulation of coding regions in the pan-genome (upper lines) or reduction of coding regions in the core genome (lower lines).



9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

Figure 5: A comparison between all *Escherichia coli* genomes in Genbank (black) and simulated short read assemblies (red) in terms of (A) the number of total contigs, and (B) the summed genome assembly size.

References

1. **Sahl JW, Gillece JD, Schupp JM, Waddell VG, Driebe EM, Engelthaler DM, Keim P.** 2013. Evolution of a pathogen: a comparative genomics analysis identifies a genetic pathway to pathogenesis in *Acinetobacter*. *PLoS ONE* **8**:e54287.
2. **Pearson T, Giffard P, Beckstrom-Sternberg S, Auerbach R, Hornstra H, Tuanyok A, Price EP, Glass MB, Leadem B, Beckstrom-Sternberg JS, Allan GJ, Foster JT, Wagner DM, Okinaka RT, Sim SH, Pearson O, Wu Z, Chang J, Kaul R, Hoffmaster AR, Brettin TS, Robison RA, Mayo M, Gee JE, Tan P, Currie BJ, Keim P.** 2009. Phylogeographic reconstruction of a bacterial species with high levels of lateral gene transfer. *BMC Biol* **7**:78.
3. **Chen PE, Shapiro BJ.** 2015. The advent of genome-wide association studies for bacteria. *Curr Opin Microbiol* **25**:17-24.
4. **Sahl JW, Lemmer D, Travis J, Schupp JM, Gillece JD, Aziz M, Driebe EM, Drees KP, Hicks ND, Williamson CHD, Hepp CM, Smith DE, Roe C, Engelthaler DM, Wagner DM, Keim P.** 2016. NASP: an accurate, rapid method for the identification of SNPs in WGS datasets that supports flexible input and output formats. *Microbial Genomics* DOI: **10.1099/mgen.0.000074**.
5. **Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Ros IMY, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou LW, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM.** 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America* **102**:13950-13955.
6. **Vernikos G, Medini D, Riley DR, Tettelin H.** 2015. Ten years of pan-genome analyses. *Current Opinion in Microbiology* **23**:148-154.
7. **Benedict MN, Henriksen JR, Metcalf WW, Whitaker RJ, Price ND.** 2014. ITEP: An integrated toolkit for exploration of microbial pan-genomes. *Bmc Genomics* **15**.
8. **Chaudhari NM, Gupta VK, Dutta C.** 2016. BPGA- an ultra-fast pan-genome analysis pipeline. *Scientific Reports* **6**.
9. **Contreras-Moreira B, Vinuesa P.** 2013. GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. *Applied and Environmental Microbiology* **79**:7696-7701.
10. **Laing C, Buchanan C, Taboada EN, Zhang YX, Kropinski A, Villegas A, Thomas JE, Gannon VPJ.** 2010. Pan-genome sequence analysis

- 1 using Panseq: an online tool for the rapid analysis of core and accessory
2 genomic regions. *Bmc Bioinformatics* **11**.
- 3 11. **Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG,**
4 **Fookes M, Falush D, Keane JA, Parkhill J.** 2015. Roary: rapid large-
5 scale prokaryote pan genome analysis. *Bioinformatics* **31**:3691-3693.
- 6 12. **Sahl JW, Caporaso JG, Rasko DA, Keim P.** 2014. The large-scale blast
7 score ratio (LS-BSR) pipeline: a method to rapidly compare genetic
8 content between bacterial genomes. *PeerJ* **2**:e332.
- 9 13. **Zhao YB, Wu JY, Yang JH, Sun SX, Xiao JF, Yu J.** 2012. PGAP: pan-
10 genomes analysis pipeline. *Bioinformatics* **28**:416-418.
- 11 14. **Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A.** 2015.
12 Large-scale contamination of microbial isolate genomes by Illumina PhiX
13 control. *Standards in Genomic Sciences* **10**.
- 14 15. **Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW.** 2015.
15 CheckM: assessing the quality of microbial genomes recovered from
16 isolates, single cells, and metagenomes. *Genome Research* **25**:1043-
17 1055.
- 18 16. **Tennessen K, Andersen E, Clingenpeel S, Rinke C, Lundberg DS,**
19 **Han J, Dangi JL, Ivanova N, Woyke T, Kyrpides N, Pati A.** 2016.
20 ProDeGe: a computational protocol for fully automated decontamination of
21 genomes. *Isme Journal* **10**:269-272.
- 22 17. **Treangen TJ, Salzberg SL.** 2011. Repetitive DNA and next-generation
23 sequencing: computational challenges and solutions. *Nat Rev Genet*
24 **13**:36-46.
- 25 18. **Huang W, Li L, Myers JR, Marth GT.** 2012. ART: a next-generation
26 sequencing read simulator. *Bioinformatics* **28**:593-594.
- 27 19. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov**
28 **AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV,**
29 **Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA.** 2012.
30 SPAdes: a new genome assembly algorithm and its applications to single-
31 cell sequencing. *J Comput Biol* **19**:455-477.
- 32 20. **Li H.** 2013. Aligning sequence reads, clone sequences and assembly
33 contigs with BWA-MEM. *arXiv.org*.
- 34 21. **Quinlan AR, Hall IM.** 2010. BEDTools: a flexible suite of utilities for
35 comparing genomic features. *Bioinformatics* **26**:841-842.
- 36 22. **Delcher AL, Salzberg SL, Phillippy AM.** 2003. Using MUMmer to identify
37 similar regions in large sequence sets. *Curr Protoc Bioinformatics*
38 **Chapter 10**:Unit 10 13.
- 39 23. **Rasko DA, Myers GS, Ravel J.** 2005. Visualization of comparative
40 genomic analyses by BLAST score ratio. *BMC Bioinformatics* **6**:2.
- 41 24. **Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ.**
42 2010. Prodigal: prokaryotic gene recognition and translation initiation site
43 identification. *BMC bioinformatics* **11**:119.
- 44 25. **Edgar RC.** 2010. Search and clustering orders of magnitude faster than
45 BLAST. *Bioinformatics* **26**:2460-2461.

- 1 26. **Kent WJ.** 2002. BLAT--the BLAST-like alignment tool. *Genome Res* 12:656-664.
- 2
- 3 27. **Magoc T, Pabinger S, Canzar S, Liu XY, Su Q, Puiu D, Tallon LJ, Salzberg SL.** 2013. GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics* 29:1718-1725.
- 4
- 5
- 6 28. **Gurevich A, Saveliev V, Vyahhi N, Tesler G.** 2013. QUASt: quality assessment tool for genome assemblies. *Bioinformatics* 29:1072-1075.
- 7
- 8 29. **Akayama KN, Amashita AY, Urokawa KK, Orimoto TM, Gawa MO, Ukuhara MF, Rakami HU, Hnishi MO, Chiyama IU, Gura YO, Oka TO, Shima KO, Amura AT, Attori MH, Ayashi TH.** 2008. The Whole-genome Sequencing of the Obligate Intracellular Bacterium *Orientia tsutsugamushi* Revealed Massive Gene Amplification During Reductive Genome Evolution. *DNA Research* 15:185-199.
- 9
- 10
- 11
- 12
- 13
- 14 30. **Sahl JW, Vazquez AJ, Hall CM, Busch JD, Tuanyok A, Mayo M, Schupp JM, Lummis M, Pearson T, Shippy K, Colman RE, Allender CJ, Theobald V, Sarovich DS, Price EP, Hutcheson A, Korfach J, LiPuma JJ, Ladner J, Lovett S, Koroleva G, Palacios G, Limmathurotsakul D, Wuthiekanun V, Wongsuwan G, Currie BJ, Keim P, Wagner DM.** 2016. The effects of signal erosion and core genome reduction on the identification of diagnostic markers. *MBio* 7:e00846-00816.
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22 31. **Jeong H, Pan J-G, Park S-H.** 2016. Contaminatin as a major factor in poor Illumina assembly of microbial isolate genomes. *bioRxiv* doi:<http://dx.doi.org/10.1101/081885>.
- 23
- 24
- 25 32. **Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America* 95:3140-3145.
- 26
- 27
- 28
- 29
- 30
- 31 33. **Jolley KA, Maiden MC.** 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC bioinformatics* 11:595.
- 32
- 33 34. **Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, Engelthaler DM, Bortolaia V, Pearson T, Waters AE, Upadhyay BP, Shrestha SD, Adhikari S, Shakya G, Keim PS, Aarestrup FM.** 2011. Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. *mBio* 2:e00157-00111.
- 34
- 35
- 36
- 37
- 38 35. **Van Ert MN, Easterday WR, Huynh LY, Okinaka RT, Hugh-Jones ME, Ravel J, Zanecki SR, Pearson T, Simonson TS, U'Ren JM, Kachur SM, Leadem-Dougherty RR, Rhoton SD, Zinser G, Farlow J, Coker PR, Smith KL, Wang B, Kenefic LJ, Fraser-Liggett CM, Wagner DM, Keim P.** 2007. Global genetic population structure of *Bacillus anthracis*. *PloS one* 2:e461.
- 39
- 40
- 41
- 42
- 43
- 44 36. **Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korfach**
- 45

- 1 J. 2013. Nonhybrid, finished microbial genome assemblies from long-read
2 SMRT sequencing data. *Nat Methods* **10**:563-569.
- 3 37. **Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO,**
4 **Buffalo V, Zerbino DR, Diekhans M, Nguyen N, Ariyaratne PN, Sung**
5 **WK, Ning Z, Haimel M, Simpson JT, Fonseca NA, Birol I, Docking TR,**
6 **Ho IY, Rokhsar DS, Chikhi R, Lavenier D, Chapuis G, Naquin D,**
7 **Maillet N, Schatz MC, Kelley DR, Phillippy AM, Koren S, Yang SP, Wu**
8 **W, Chou WC, Srivastava A, Shaw TI, Ruby JG, Skewes-Cox P,**
9 **Betegon M, Dimon MT, Solovyev V, Seledtsov I, Kosarev P, Vorobyev**
10 **D, Ramirez-Gonzalez R, Leggett R, MacLean D, Xia F, Luo R, Li Z, Xie**
11 **Y, et al.** 2011. Assemblathon 1: a competitive assessment of de novo
12 short read assembly methods. *Genome Res* **21**:2224-2241.
- 13 38. **Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S,**
14 **Treangen TJ, Schatz MC, Delcher AL, Roberts M, Marcais G, Pop M,**
15 **Yorke JA.** 2012. GAGE: A critical evaluation of genome assemblies and
16 assembly algorithms (vol 22, pg 557, 2012). *Genome Research* **22**:1196-
17 1196.
- 18 39. **Sahl JW, Allender CJ, Colman RE, Califf KJ, Schupp JM, Currie BJ,**
19 **Van Zandt KE, Gelhaus HC, Keim P, Tuanyok A.** 2015. Genomic
20 Characterization of Burkholderia pseudomallei Isolates Selected for
21 Medical Countermeasures Testing: Comparative Genomics Associated
22 with Differential Virulence. *Plos One* **10**.
- 23 40. **Sahl JW, Del Franco M, Pournaras S, Colman RE, Karah N,**
24 **Dijkshoorn L, Zarrilli R.** 2015. Phylogenetic and genomic diversity in
25 isolates from the globally distributed Acinetobacter baumannii ST25
26 lineage. *Scientific Reports* **5**.
- 27 41. **Sahl JW, Morris CR, Emberger J, Fraser CM, Ochieng JB, Juma J,**
28 **Fields B, Breiman RF, Gilmour M, Nataro JP, Rasko DA.** 2015.
29 Defining the Phylogenomics of Shigella Species: a Pathway to
30 Diagnostics. *Journal of Clinical Microbiology* **53**:951-960.
- 31 42. **Sahl JW, Sistrunk JR, Fraser CM, Hine E, Baby N, Begum Y, Luo Q,**
32 **Sheikh A, Qadri F, Fleckenstein JM, Rasko DA.** 2015. Examination of
33 the Enterotoxigenic Escherichia coli Population Structure during Human
34 Infection. *MBio* **6**:e00501.
- 35 43. **Hazen TH, Donnenberg MS, Panchalingam S, Antonio M, Hossain A,**
36 **Mandomando I, Ochieng JB, Ramamurthy T, Tamboura B, Qureshi S,**
37 **Quadri F, Zaidi A, Kotloff KL, Levine MM, Barry EM, Kaper JB, Rasko**
38 **DA, Nataro JP.** 2016. Genomic diversity of EPEC associated with clinical
39 presentations of differing severity. *Nat Microbiol* **1**:15014.
- 40 44. **Baig A, McNally A, Dunn S, Paszkiewicz KH, Corander J, Manning G.**
41 2015. Genetic import and phenotype specific alleles associated with
42 hyper-invasion in Campylobacter jejuni. *BMC Genomics* **16**:852.
- 43