**deBGA: read alignment with de Bruijn Graph-based seed and extension**

**Supplementary information**

Bo Liu[[1]](#footnote-1), Hongzhe Guo\*, Michael Brudno & Yadong Wang[[2]](#footnote-2)

Contents

[Supplementary Figure 1. The genome organization and indexing of deBGA 3](#_Toc451085767)

[Supplementary Figure 2. An illustration of detecting similar seeds with unipaths 4](#_Toc451085768)

[Supplementary Figure 3. An illustration of merging identical extensions with unipaths 6](#_Toc451085769)

[Supplementary Figure 4. Flowchart of read alignment processing 7](#_Toc451085770)

[Supplementary Figure 5. An illustration of Uni-MEM seeding 8](#_Toc451085771)

[Supplementary Figure 6. An illustration of merging seeds within the same unipath 10](#_Toc451085772)

[Supplementary Figure 7. Benchmarking on the datasets from two divergent and two similar *E*. coli strains 11](#_Toc451085773)

[Supplementary Figure 8. The numbers of the real *E. coli* reads aligned to the 62 *E. coli* strains by deBGA (with the more sensitive cl=0.35 setting) 14](#_Toc451085774)

[Supplementary Figure 9. Benchmarking on the datasets from human genome (zoomed-in view) 15](#_Toc451085775)

[Supplementary Figure 10. An example on the difference between deBGA and BWA-MEM for the reads around called duplications of deBGA 16](#_Toc451085776)

[Supplementary Table 1. Datasets for evaluation 17](#_Toc451085777)

[Supplementary Table 2. *E. coli* strains involved in simulation datasets 19](#_Toc451085778)

[Supplementary Table 3. Alignment results on simulated datasets from *E. coli* genomes 21](#_Toc451085779)

[Supplementary Table 4. The repetitiveness of the references 30](#_Toc451085780)

[Supplementary Table 5. Alignment results on simulated datasets from *A. thaliana* genomes 31](#_Toc451085781)

[Supplementary Table 6. Alignment results on real datasets from *E. coli* genomes 34](#_Toc451085782)

[Supplementary Table 7. The numbers of real reads aligned to the 62 *E.coli* strains by deBGA 38](#_Toc451085783)

[Supplementary Table 8. Alignment results on simulated datasets from RefSeq bacteria genomes 40](#_Toc451085784)

[Supplementary Table 9. Alignment results on simulated datasets from Xenograft model 41](#_Toc451085785)

[Supplementary Table 10. Alignment results on pseudo-real metagenomic datasets 42](#_Toc451085786)

[Supplementary Table 11. Alignment results on simulated datasets from human genome 43](#_Toc451085787)

[Supplementary Table 12. The numbers of human simulated reads correctly aligned by the primary and alternative alignments of deBGA 49](#_Toc451085788)

[Supplementary Table 13. Alignment results on the reads spanning retrotransposon regions of human genome 52](#_Toc451085789)

[Supplementary Table 14. Alignment results on real datasets from human genome 53](#_Toc451085790)

[Supplementary Table 15. Statistics on the called STRs of NA12878 57](#_Toc451085791)

[Supplementary Table 16. Statistics on the called duplications and inversions of NA12878 58](#_Toc451085792)

[Supplementary Table 17. Time and memory footprint of RdBG-index construction 59](#_Toc451085793)

[Supplementary Table 18. Memory footprint of the alignment of deBGA 60](#_Toc451085794)

[Supplementary Table 19. The alignment time of deBGA in multiple threads 61](#_Toc451085795)

[Supplementary Notes 62](#_Toc451085796)

[Implementation of the comparisons 62](#_Toc451085797)

[The time cost of the construction of RdBG-index 62](#_Toc451085798)

[The memory footprint of deBGA 63](#_Toc451085799)

[The performance of multiple threads 63](#_Toc451085800)

[Datasets for benchmarking 64](#_Toc451085801)

[Simulated datasets 64](#_Toc451085802)

[Real datasets 64](#_Toc451085803)

[Statistics for the evaluation of alignment 64](#_Toc451085804)

[Command lines of aligners used for comparison 65](#_Toc451085805)

[Benchmarks for deBGA (version 0.1.1) 65](#_Toc451085806)

[Benchmarks for Bowtie2 (version 2.2.4) 65](#_Toc451085807)

[Benchmarks for BWA-MEM (version 0.7.5) 66](#_Toc451085808)

[Benchmarks for Seqalto (version 0.5-r123) 66](#_Toc451085809)

[Benchmarks for BWA (version 0.7.5) 66](#_Toc451085810)

[Benchmarks for GEM (version core\_i3-20121106-022124) 66](#_Toc451085811)

[Benchmarks for STAR (version 2.5.0a) 67](#_Toc451085812)

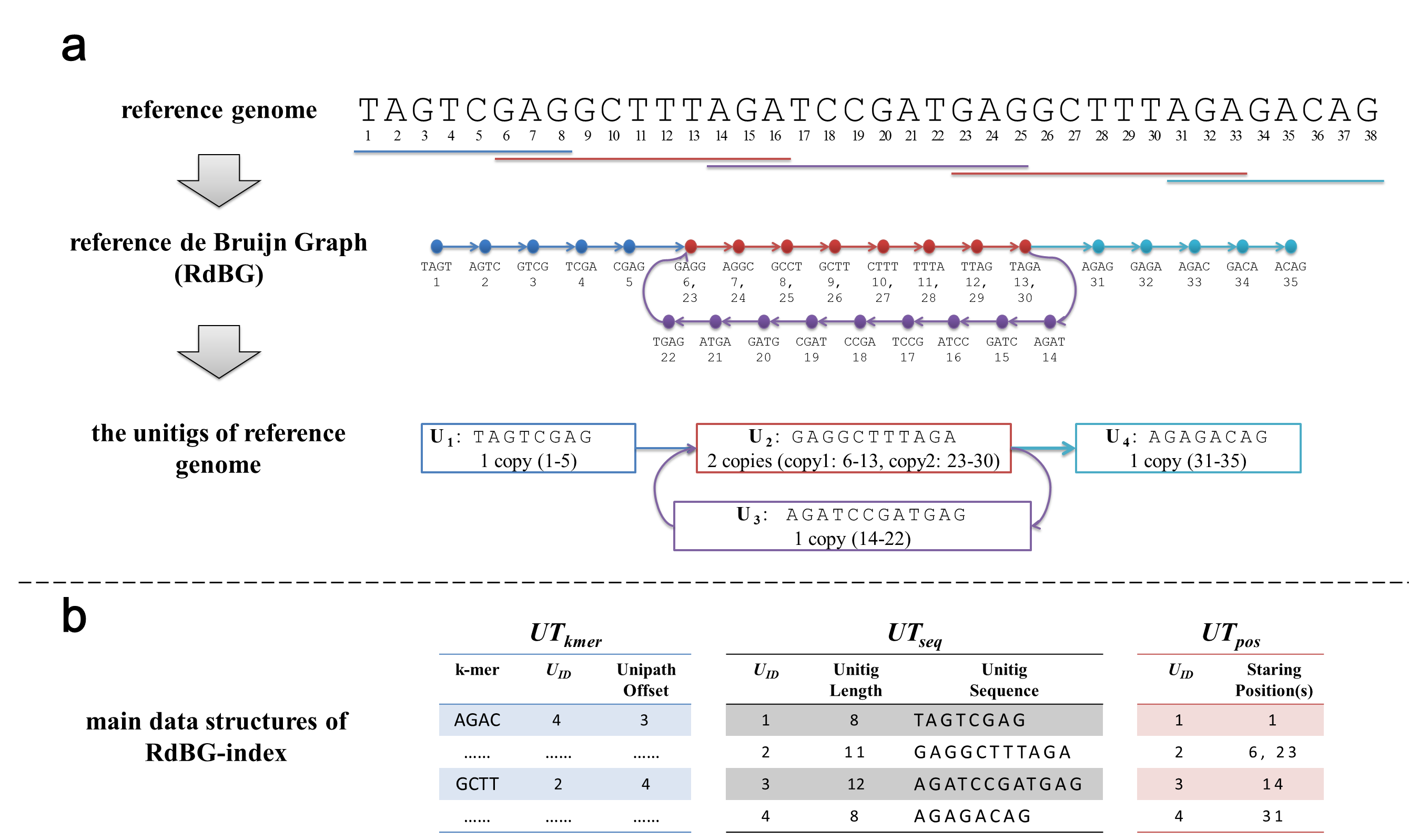
[Benchmarks for GenomeMapper (version 0.4.4) 67](#_Toc451085813)

[Variant calling with GATK HaplotypeCaller (version 3.3.0) 67](#_Toc451085814)

[Variant calling with lobSTR (version 4.0.0) 67](#_Toc451085815)

[Variant calling with Delly (version 0.7.1) 68](#_Toc451085816)

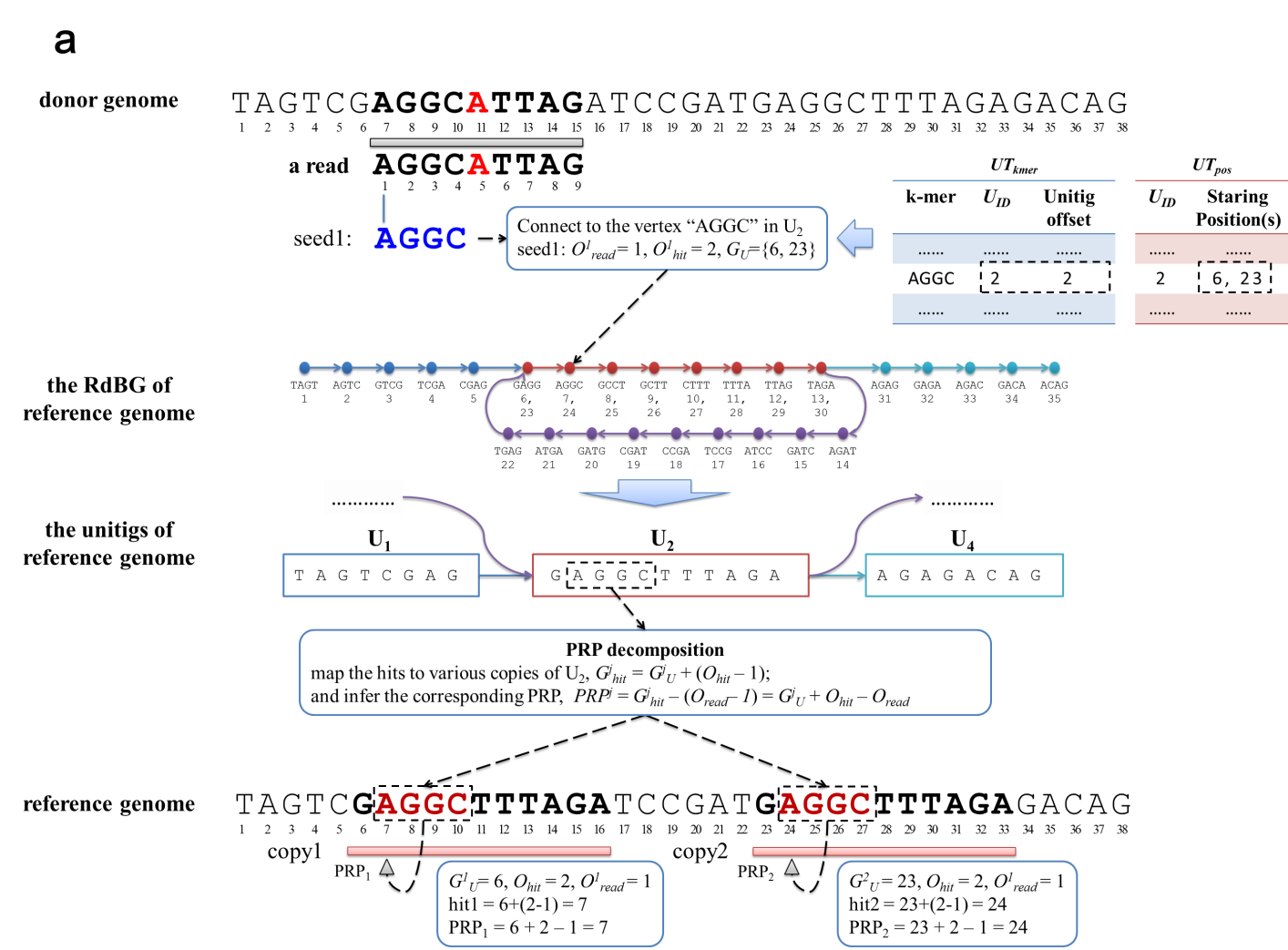
[Options of deBGA 68](#_Toc451085817)

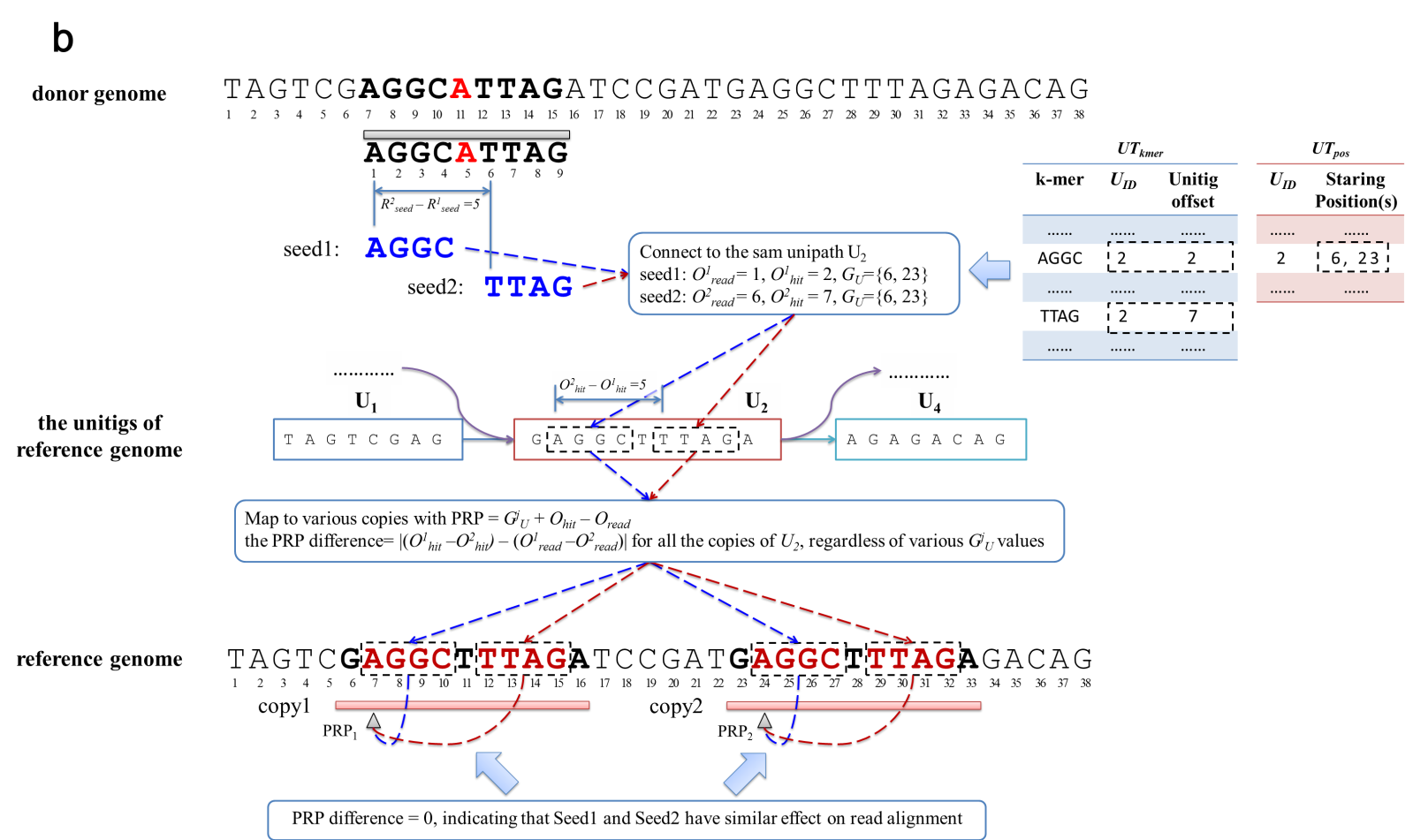


# Supplementary Figure 1. The genome organization and indexing of deBGA

**a) An illustration of organizing the reference genome using a de Bruijn graph.** DeBGA builds the RdBG of the reference genome, and records all its unipaths/unitigs, as well as the starting positions of all the copies of the unitigs. In this case, there are four unipaths, respectively marked by various colors (blue, red, purple and aqua). Moreover, *U2* (“GAGGCTTTAGA”) has two copies respectively starting from the positions 6 and 23. The “toy” reference genome is inspired by the case in the following link which was created by Dr. Daniel Zerbino (<https://plus.maths.org/content/os/issue55/features/sequencing/index>).

**b) The main data structures of RdBG-index.** Mainly, is a hash table recording the tuples of all the *k*-mers and their corresponding unitig coordinates, i.e., the identifiers (s) and offsets of the unitigs. For each unitig, records its length and sequence, and records the starting positions of all its copies.

****

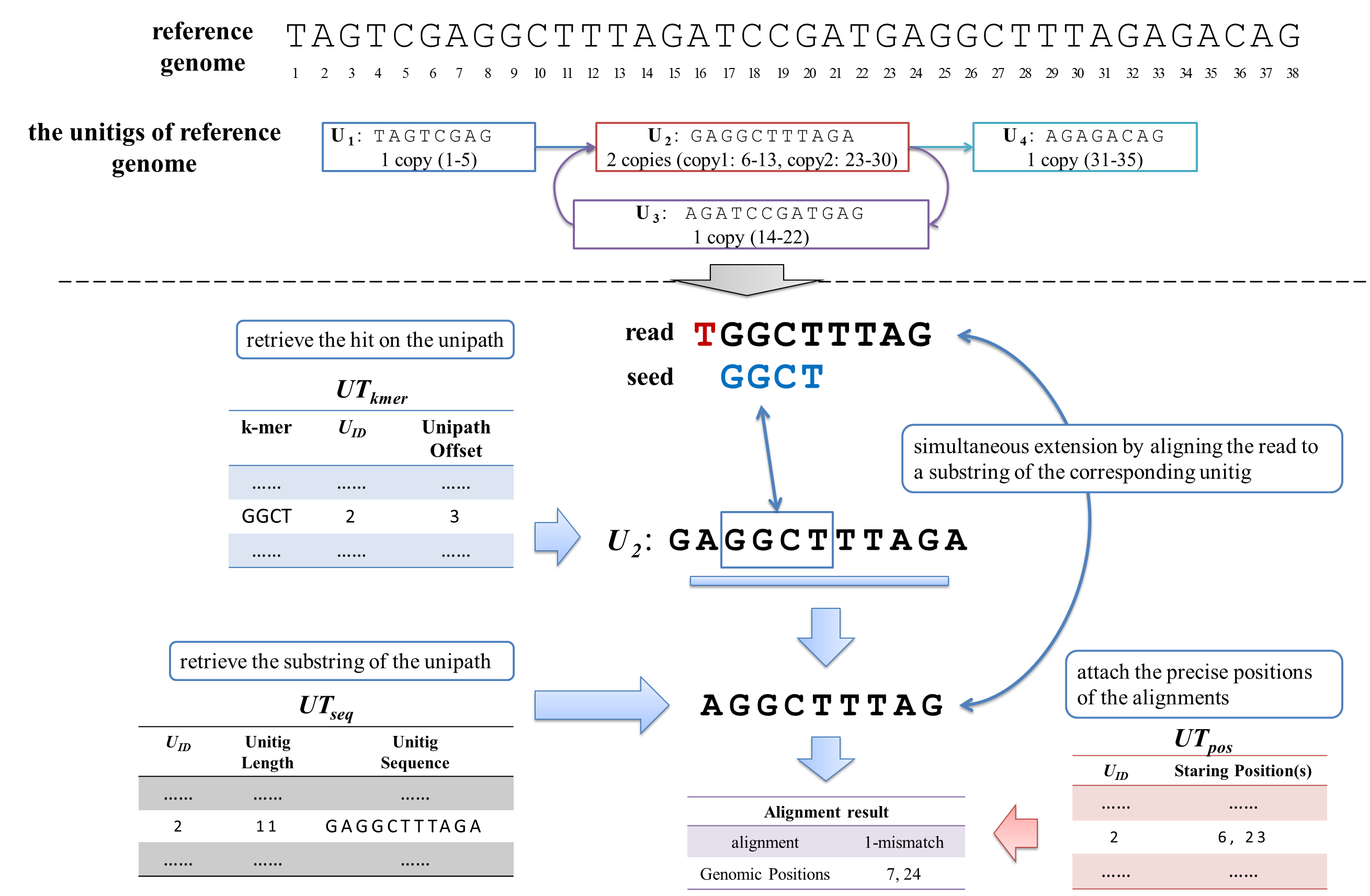
****

# Supplementary Figure 2. An illustration of detecting similar seeds with unipaths

**a) A decomposition of the PRPs of a *k* bp seed with RdBG-index.** For a *k*-mer of the read, all its matches to the reference genome collapse to a unique vertex of RdBG at a specific unitig coordinate. Thus, with the copies of the unitig, each hit of the seed can be formulated as =, where denotes the starting position of the *i*-th copy of the unitig, and denotes the offset on the unitig. This indicates that each of the PRPs inferred by the rule can be decomposed by the , , and values. In this case, a seed, “AGGC” () is matched to a vertex of RdBG (, )by retrieving the *.* Thus, the seed is mapped to the two copies of *U2* whose starting positions are respectively 6 and 23, and the PRPs are respectively inferred as 6 + 2 – 1 = 7 and 23 + 2 – 1 = 24 ( is considered as 0 for simplicity).

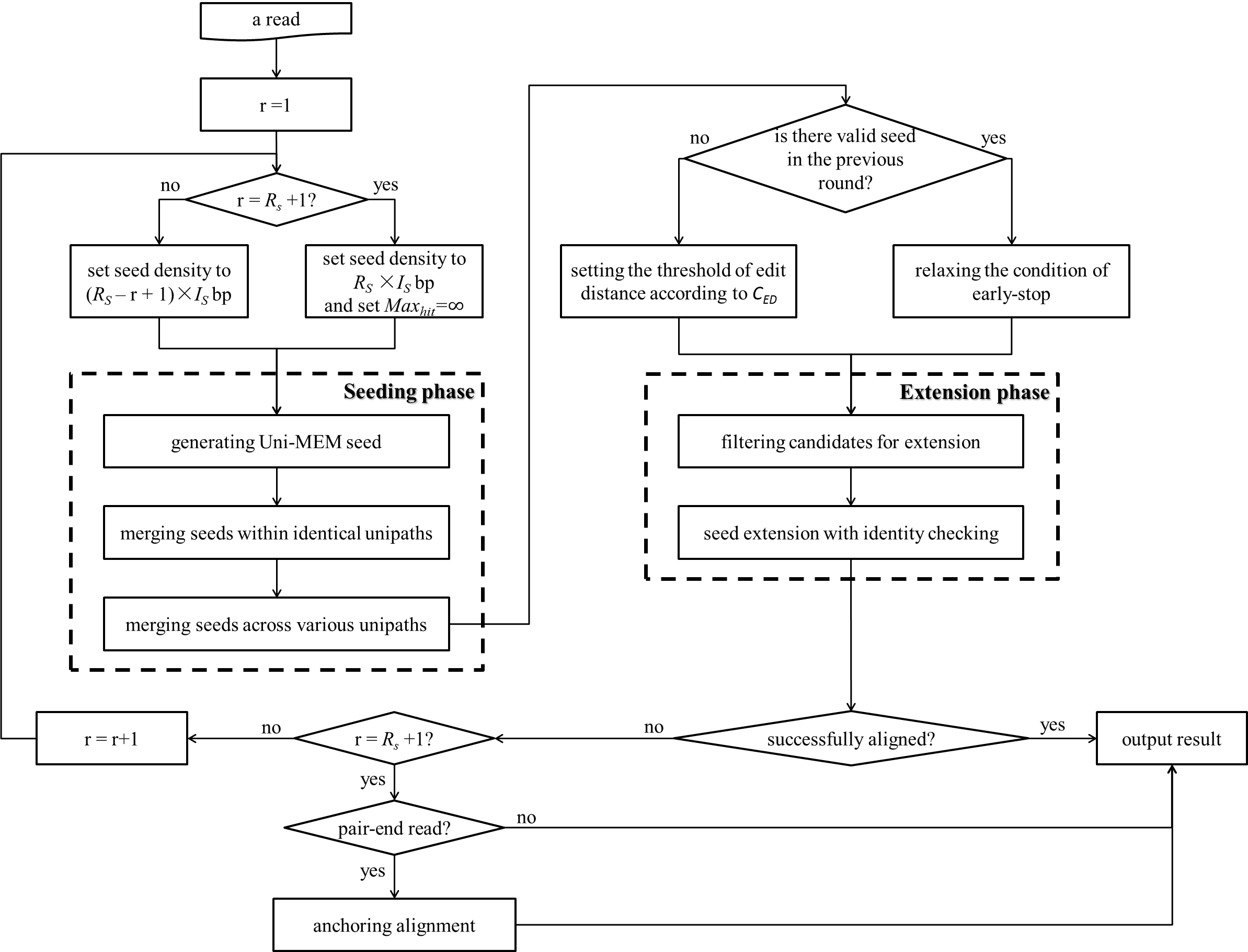
**b) An illustration for detecting two similar seeds.** If two seeds connect to the same unipath, the difference between their PRP sets can be represented by the value . In this case, there are two seeds (“AGGC” and “TTAG”) connecting to the identical unipath (*U2*) and the difference is , indicating PRPs derived from the two seeds are equivalent, thus the two seeds can be considered as having highly similar effect to read alignment.

It is also worthnoting that, in this figure, the red characters in the donor genome and the read denote a variation of the donor genome.

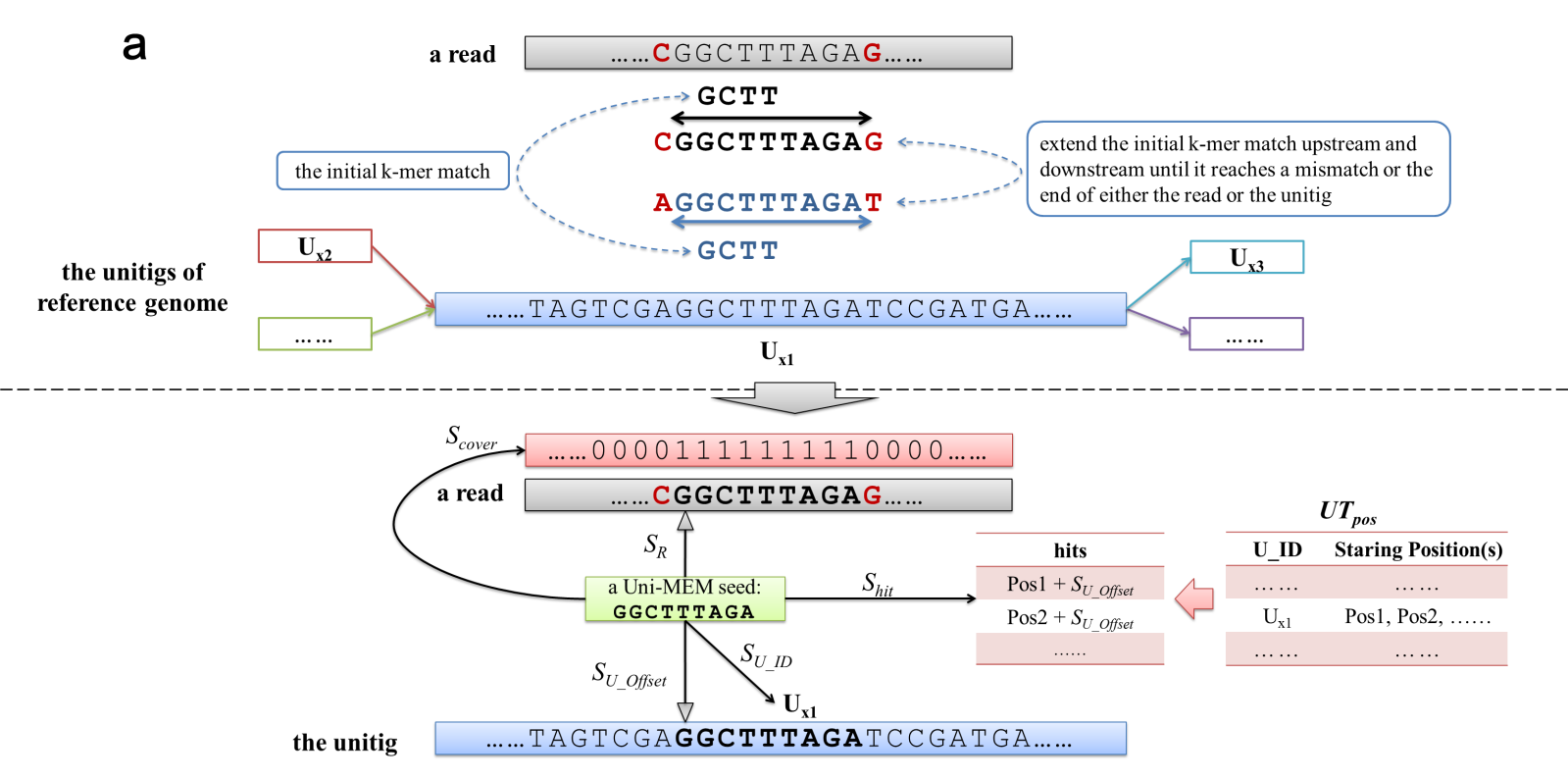
****

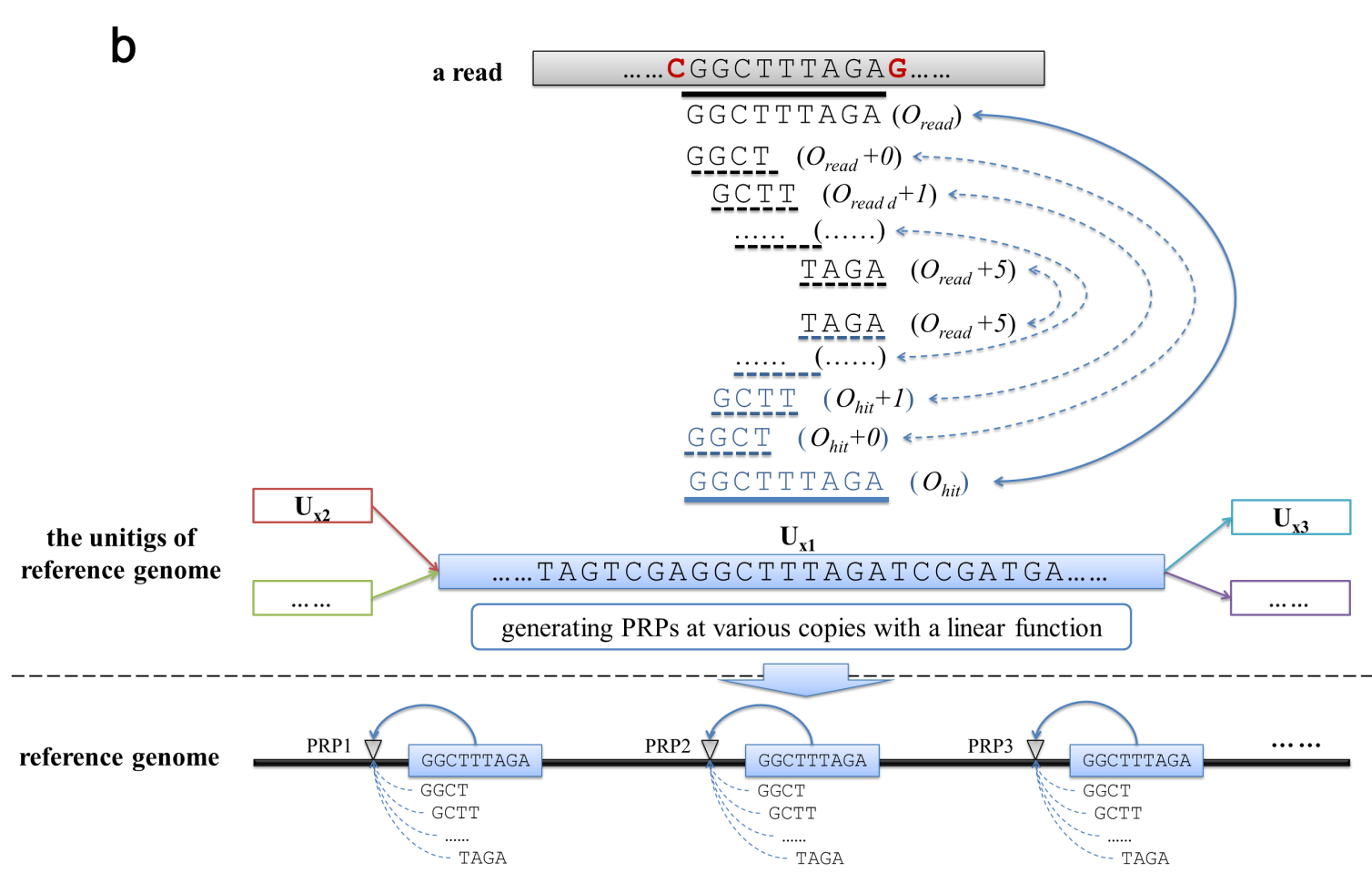
# Supplementary Figure 3. An illustration of merging identical extensions with unipaths

With RdBG-index, the identity of local sequences can be detected in advance. In this case, for a seed hitting on a certain unitig (“AGGC” and *U2* for this case, the red character at the left of the seed denotes a variation of the donor genome), it is easy to check if the unitig can accommodate the whole read to be aligned. If this is the case, all the local sequences are identical to the corresponding substring of the unitig. Thus, a comprehensive extension for all the hits of the seed can be performed by directly aligning the read (“AGGCATTTAG” for this case) with the substring (“AGGCTTTAG” for this case) of the unitig and attaching the positions of the various copies.



# Supplementary Figure 4. Flowchart of read alignment processing

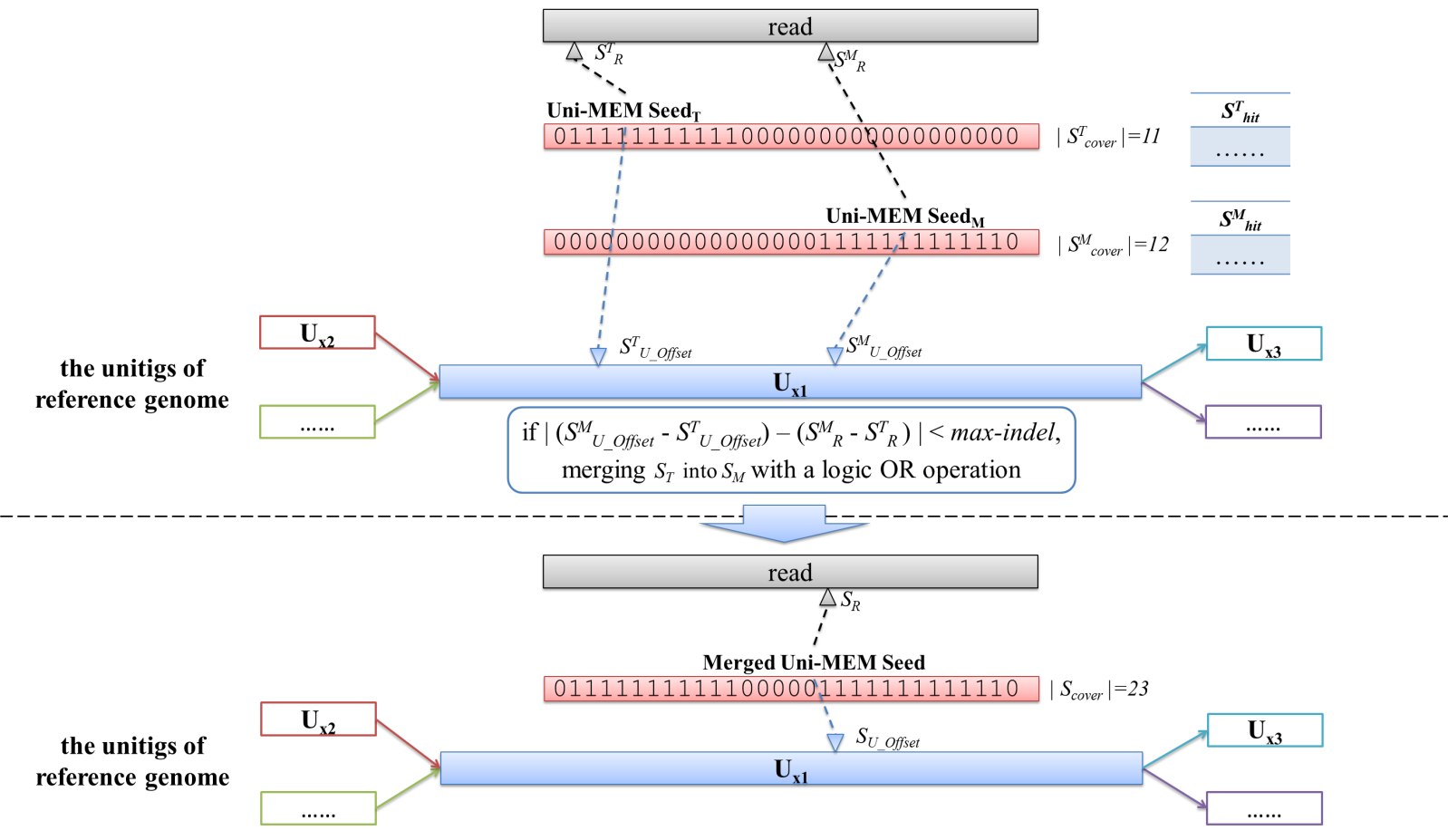




# Supplementary Figure 5. An illustration of Uni-MEM seeding

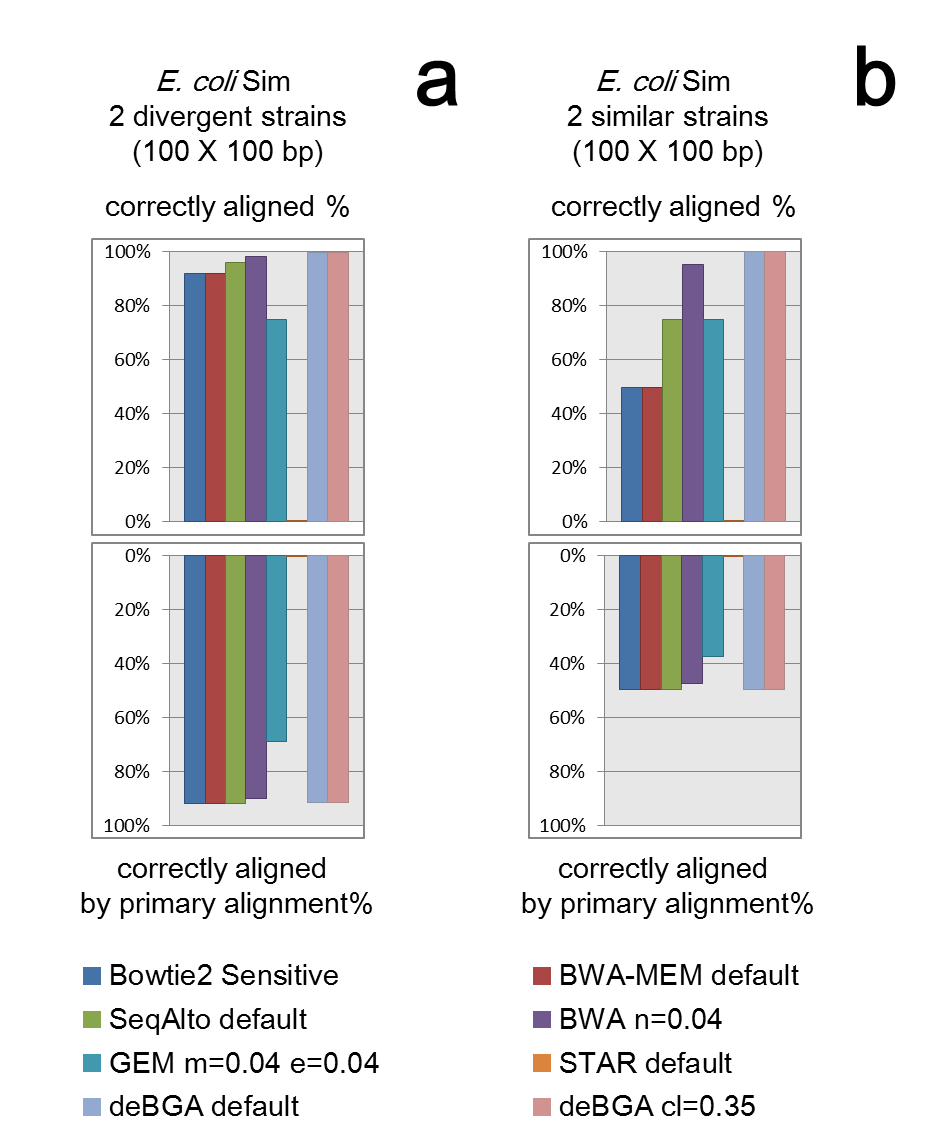
**a) An illustration of the generation of Uni-MEM seed.** A Uni-MEM seed originates from a *k*-mer match to a certain unitig (“GCTT” for this case). The initial match is extended upstream and downstream. In each of the two directions, the extension stops when it meets a mismatch, or reaches the end of either the read or the unitig. In this case, the extension stops due to two mismatches (“C” vs “A” for upstream and “G” vs “T” for downstream, the two mismatches can be seen as the variations of donor genome which are marked as red characters), and the Uni-MEM is “GGCTTTAGA”. The five components of the Uni-MEM seed are illustrated in the lower part of the subfigure. The red bar depicts the bit vector, where the “1”s depict the read positions covered by the seed, and “0”s for other positions. The number of “1”s indicates the coverage length of the seed. Moreover, indicate the set of the hits, which derives from the plus the positions of retrieved from .

**b) The implicit combination of multiple equivalent seeds by Uni-MEM seeding.** A Uni-MEM between a read and a certain unitig implicitly combines a series of successive *k*-mer matches in fixed length, e.g., six 4-mer matches from “GGCT” to “TAGA” for this case. For any two neighboring ones of these matches, both of their displacements on the read and the unitig are exactly 1 bp, so that two identical sets of PRPs would be inferred from any two neighboring seeds. Thus, due to the transitive relation, all these seeds have the same set of PRPs, i.e., all of them have highly similar effect on read alignment, and the Uni-MEM seeding provides an efficient way to implicitly merge them.



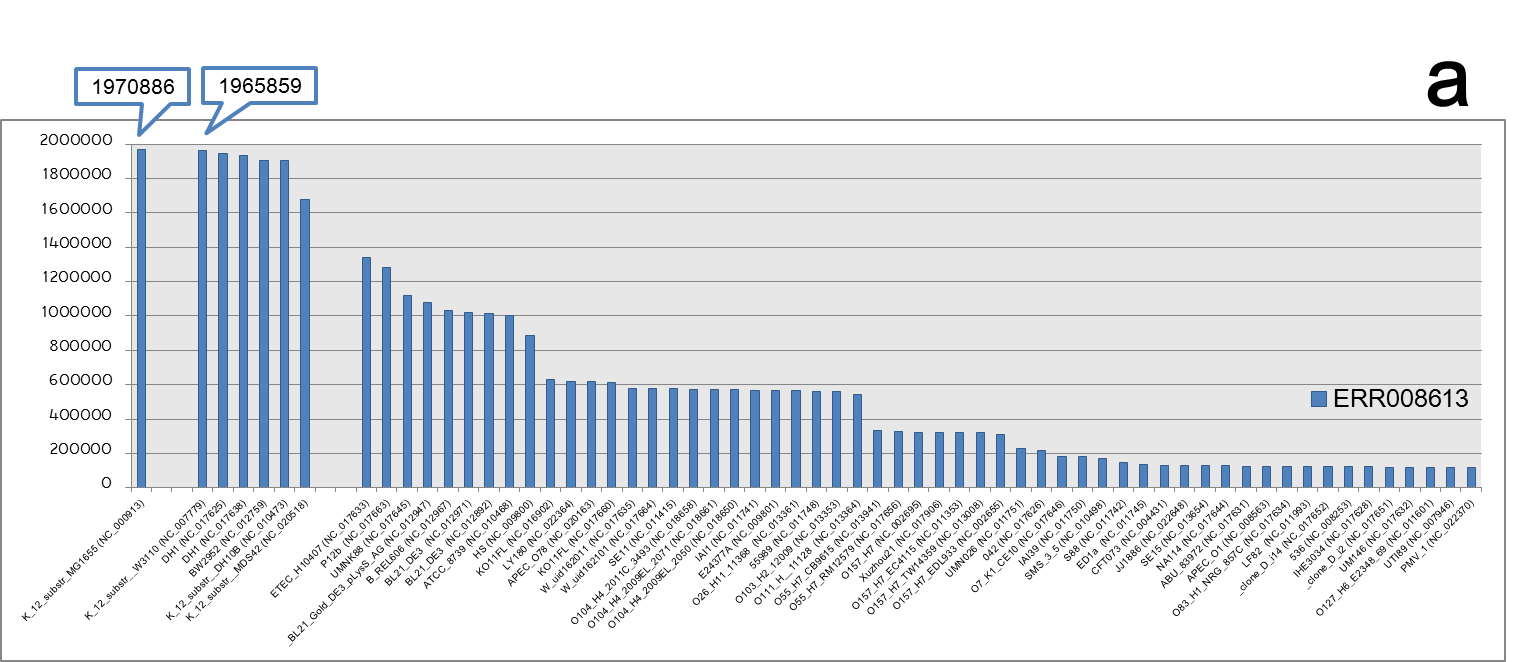
# Supplementary Figure 6. An illustration of merging seeds within the same unipath

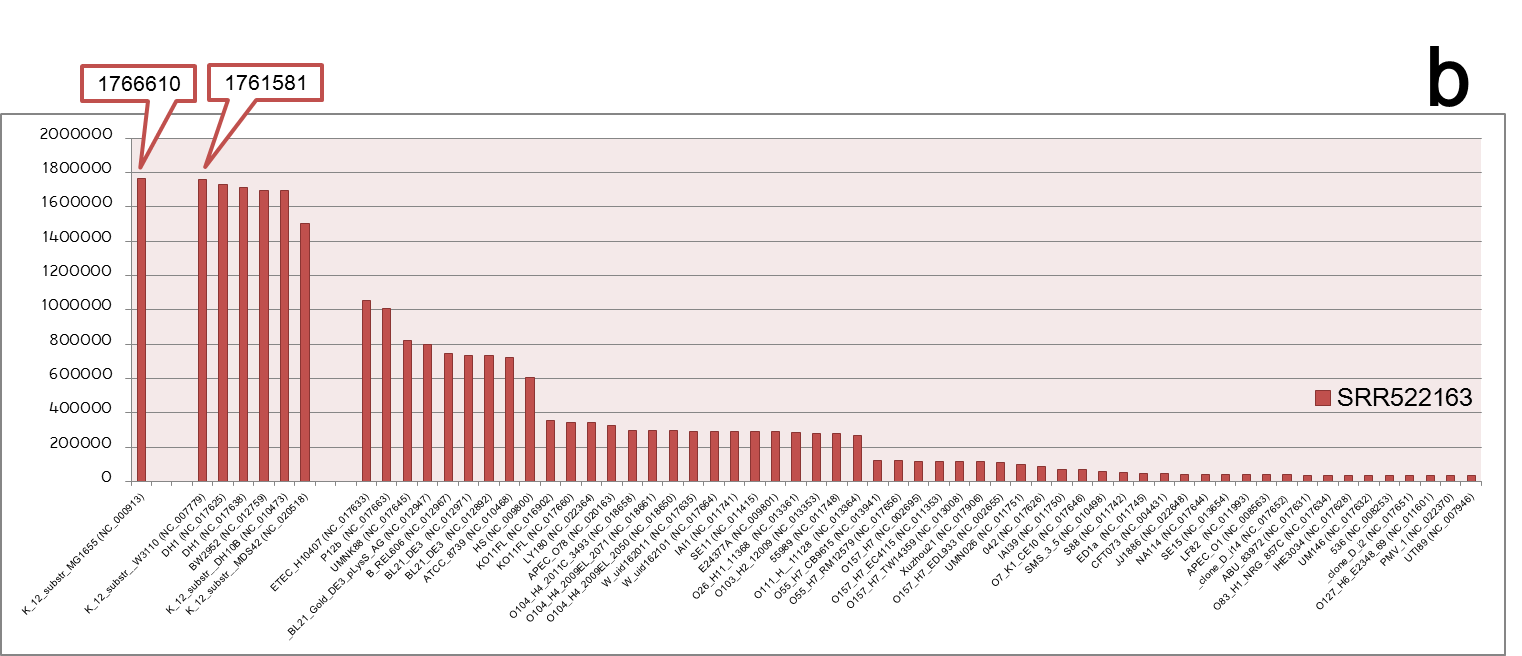
In this case, two Uni-MEM seeds, () and () connects to the same unipath. deBGA clusters them into the same bin, and select the one having larger read coverage length, i.e., , as the merging seed. Assuming that the value is smaller than the maximal allowed in-del size, indicating the two seeds are similar, deBGA merges into . After merging, the new seed inherits the RdBG-connection and the hit set from , and the new bit vector derives from a bit logic OR operation on and *,* which forms a combination of the coverages of the two seeds.

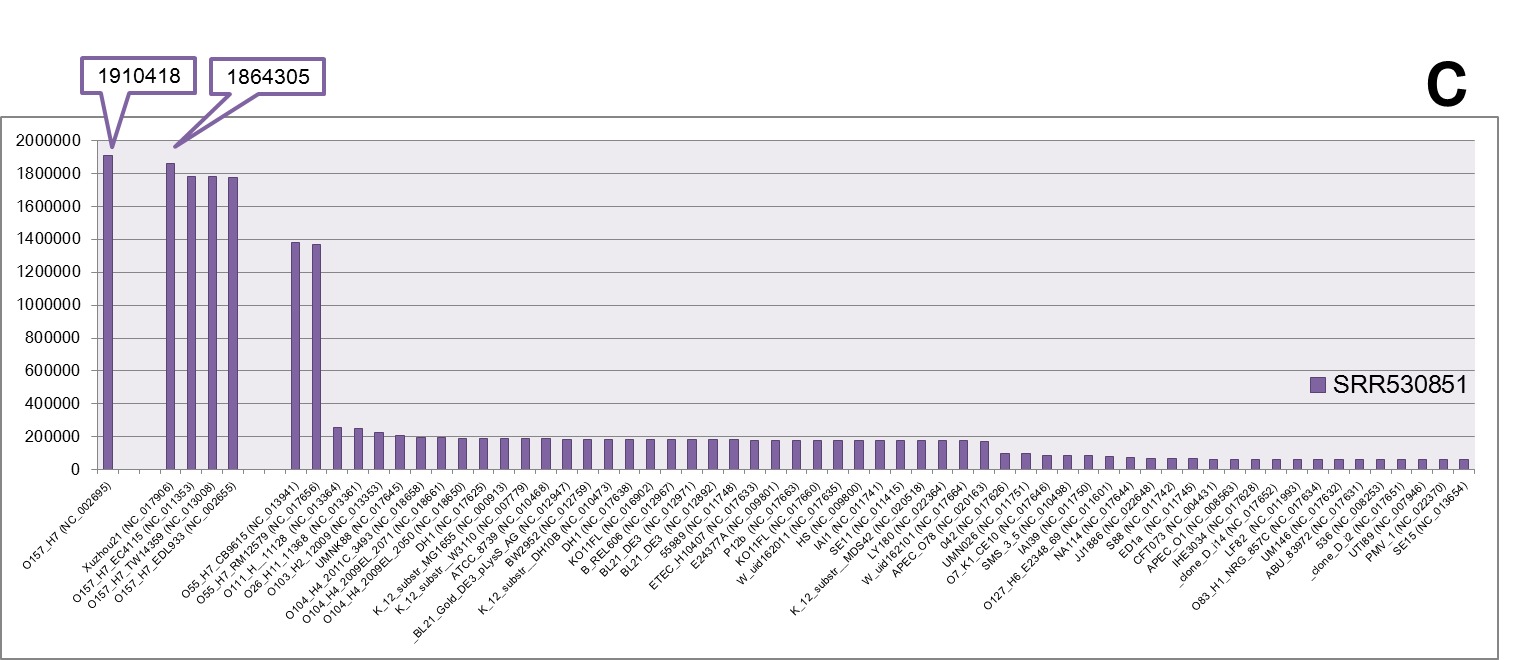


# Supplementary Figure 7. Benchmarking on the datasets from two divergent and two similar *E*. coli strains

(a) Results on the dataset from two divergent *E. coli* strains (K-12 MG1655 substrain (NC\_000913) and O157:H7 strain (NC\_002695)); (b) on the dataset from two highly similar *E. coli* strains (K-12 MG1655 substrain (NC\_000913) and K-12 W3110 substrain (NC\_007779)). “Correct aligned%” indicates the proportion of reads which are aligned to their correct positions by at least one alignment. “Correct aligned by primary alignments%” indicates the proportion of reads which are aligned to their correct positions by their primary alignments. It is worth noting that STAR corrupted when performing alignment for these two datasets.

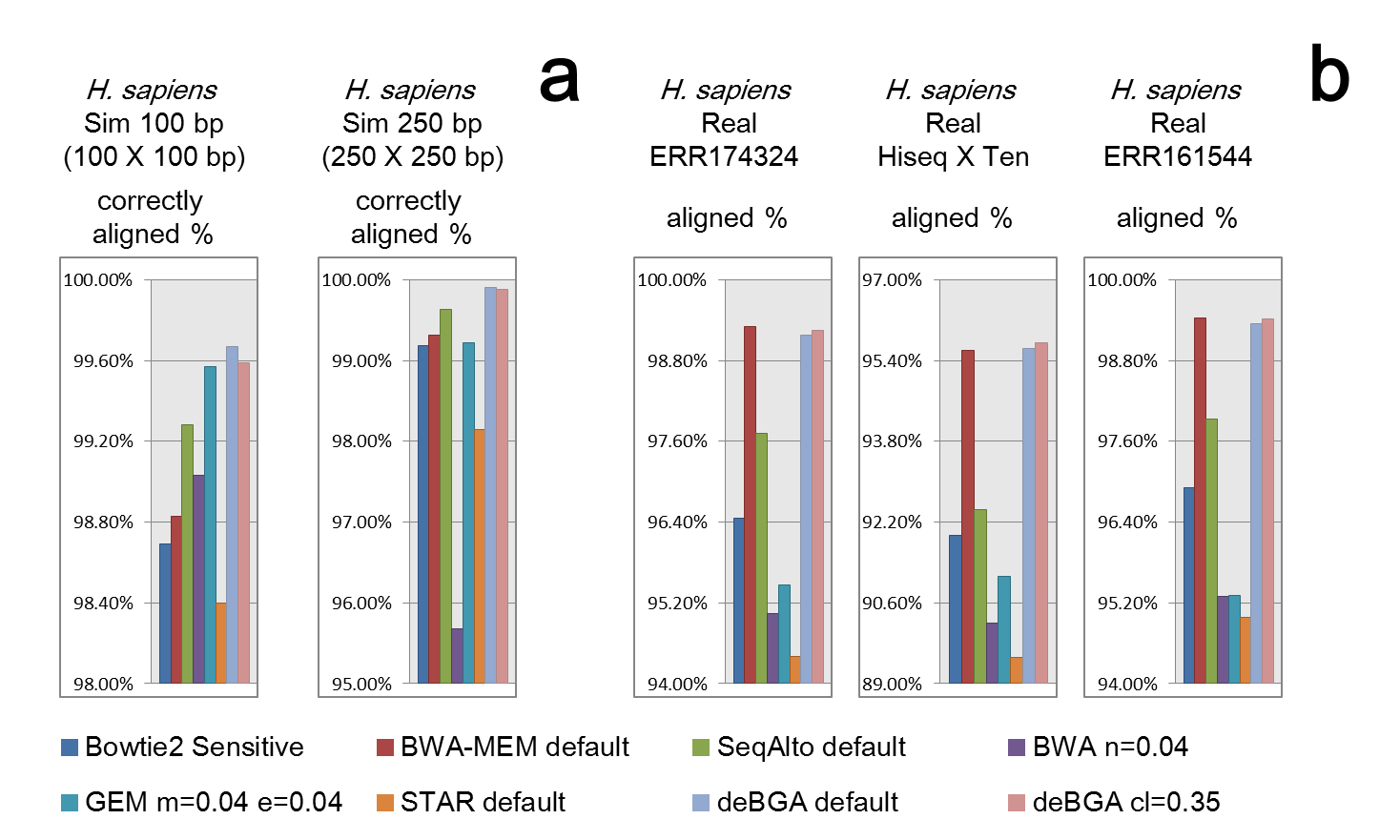






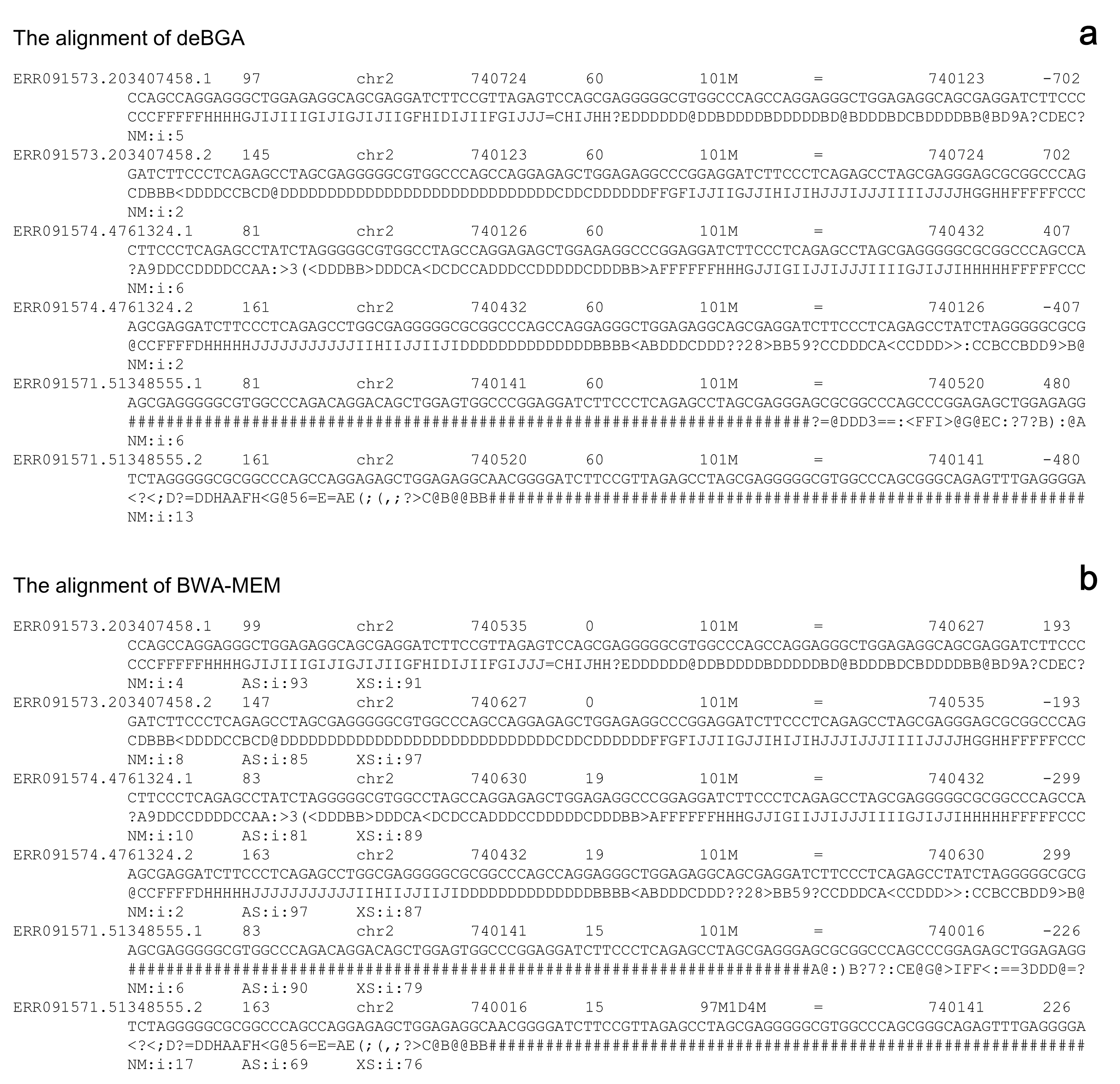
# Supplementary Figure 8. The numbers of the real *E. coli* reads aligned to the 62 *E. coli* strains by deBGA (with the more sensitive cl=0.35 setting)

Results on (a) ERR008613, (b) SRR522163 and (c) SRR530851. The bars respectively indicate the numbers of the reads aligned to the 62 strains (refer to Supplementary Table 7 for detailed numbers). In the subfigures, the *E. coli* strains are organized by the similarity to the correct strains, i.e., K-12 MG1655 substrain (NC\_000913) for ERR008613 and SRR522163, and O157:H7 strain (NC\_002695) for SRR530851. The strains similar to and distant from the correct strain are separately grouped. It is also worthnoting that, for each read, all its alignments output by deBGA are taken into account in this result, i.e., a read may be aligned to multiple genomes, so that the total number of the reads aligned to the genomes is higher than that of the original number of the reads.



# Supplementary Figure 9. Benchmarking on the datasets from human genome (zoomed-in view)

(a) Results on the 100 bp and 250bp simulated human datasets (respectively marked by “*H. sapiens* Sim 100 bp” and “*H. sapiens* Sim 250 bp”); (b) results on the three real human datasets (respectively marked by “ERR174324”, “Hiseq X Ten” and “ERR161544”). The results shown in this figure is same to that of the corresponding parts of Fig. 4 in the main text, but in a “zoomed-in” scale. “Correctly aligned%” indicates the proportion of reads which are aligned to their correct positions by at least one alignment. “Aligned%” indicates the proportion of reads which are aligned by at least one alignment.



# Supplementary Figure 10. An example on the difference between deBGA and BWA-MEM for the reads around called duplications of deBGA

This figure shows the SAM records of three pair-end reads produced by deBGA and BWA-MEM around a duplication called by Delly based on the alignment of deBGA (SV region: chr2: 739940- 740862), as an example on the difference between deBGA and BWA-MEM. DeBGA aligns the three pair-end reads as disconcordant pairs, i.e., the upstream end is aligned to the reverse strand and the downstream end is aligned to the forward strand, which exhibits the feature of tandem duplication event. However, BWA-MEM aligns the reads as normal pairs, but the edit distances (shown in the NM domain) are larger.

# Supplementary Table 1. Datasets for evaluation

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Name/Accession Number** | **Type** | **# of Reads** | **Platform** | **Read Length** | **Species** | **Task** |
| 1 | *E. coli* Sim-1 | simulation | 1,000,000 | Mason Simulator d | 100 bp × 100 bp e | *E. coli* K-12 MG1655f | aligning reads against the genomes of the various *E. coli* strains |
| 2 | *E. coli* Sim-10 | simulation | 1,000,000 | Mason Simulator d | 100 bp × 100 bp e | 10 various *E. coli* strains f |
| 3 | *E. coli* Sim-20 | simulation | 1,000,000 | Mason Simulator d | 100 bp × 100 bp e | 20 various *E. coli* strains f |
| 4 | *E. coli* Sim-40 | simulation | 1,000,000 | Mason Simulator d | 100 bp × 100 bp e | 40 various *E. coli* strains f |
| 5 | *E. coli* Sim-62 | simulation | 1,000,000 | Mason Simulator d | 100 bp × 100 bp e | 62 various *E. coli* strains f |
| 6 | *E. coli* Sim-2d | simulation | 1,000,000 | Mason Simulator d | 100 bp × 100 bp e | 2 divergent *E. coli* strains f |
| 7 | *E. coli* Sim-2s | simulation | 1,000,000 | Mason Simulator d | 100 bp × 100 bp e | 2 similar *E. coli* strains f |
| 8 | ERR008613 a | real | 1,000,000 c | Illumina GA II | 100 bp × 100 bp | *E. coli* K-12 MG1655 f |
| 9 | SRR522163 a | real | 1,000,000 c | Illumina MiSeq | 251 bp × 251 bp | *E. coli* K-12 MG1655 f |
| 10 | SRR530851 a | real | 1,000,000 c | Illumina MiSeq | 151 bp × 151 bp | *E. coli* O157:H7 f |
| 11 | *A. thaliana* Sim-1 | simulation | 1,000,000 | Mason Simulator d | 100 bp × 100 bp e | *A. thaliana* TAIR 10 g | aligning reads against the genomes of the various *A. thaliana* strains |
| 12 | *A. thaliana* Sim-19 | simulation | 1,000,000 | Mason Simulator d | 100 bp × 100 bp e | 19 *A. thaliana* strains g |
| 13 | Meta Sim | simulation | 1,000,000 | Mason Simulator d | 100 bp × 100 bp e | RefSeq bacteria genomes f | aligning reads against all the bacteria genomes recorded in RefSeq database |
| 14 | Xenograft Sim | simulation | 1,000,000 | Mason Simulator d | 100 bp × 100 bp e | *H. sapiens* and  *Mus musculus* | aligning reads against the reference composed by human and mouse genomes |
| 15 | SRR530851 a | real | 250,000 c | Illumina MiSeq | 150 bp × 150 bp | *E. coli* O157:H7 f | aligning reads against all the bacteria genomes recorded in RefSeq database, the dataset is composed by 6 real bacteria genome sequencing datasets to mimic a metagenomics dataset |
| SRR851066 a | real | 250,000 c | Illumina HiSeq 2000 | 100 bp × 102 bp | *K. pneumoniae* MGH 78578 |
| ERR459978 a | real | 250,000 c | Illumina HiSeq 2000 | 100 bp × 100 bp | *M. abscessus* |
| ERR142613 a | real | 250,000 c | Illumina MiSeq | 150 bp × 150 bp | *S. aureus subsp. aureus TW20* |
| SRR592258 a | real | 250,000 c | Illumina HiSeq 2000 | 100 bp × 100 bp | *S. aureus MW2* |
| SRR518692 a | real | 250,000 c | Illumina GA IIx | 150 bp × 150 bp | *A. hydrophila* |
| 16 | *H. sapiens* Sim-100 | simulation | 1,000,000 | Mason Simulator d | 100 bp × 100 bp e | *H. sapiens* | aligning reads against human reference genome |
| 17 | *H. sapiens* Sim-125 | simulation | 1,000,000 | Mason Simulator d | 125 bp × 125 bp e | *H. sapiens* |
| 18 | *H. sapiens* Sim-150 | simulation | 1,000,000 | Mason Simulator d | 150 bp × 150 bp e | *H. sapiens* |
| 19 | *H. sapiens* Sim-200 | simulation | 1,000,000 | Mason Simulator d | 200 bp × 200 bp e | *H. sapiens* |
| 20 | *H. sapiens* Sim-250 | simulation | 1,000,000 | Mason Simulator d | 250 bp × 250 bp e | *H. sapiens* |
| 21 | ERR174324 a | real | 1,000,000 c | Illumina HiSeq 2000 | 100 bp × 102 bp | *H. sapiens* |
| 22 | Hiseq X Ten: NA12878\_L3 b | real | 1,000,000 c | Illumina HiSeq X Ten | 150 bp × 150 bp | *H. sapiens* |
| 23 | ERR161544 | real | 1,000,000 c | Illumina HiSeq 2000 | 100 bp × 100 bp | *H. sapiens* |
| 24 | ERR091571 | real | 211,437,919 | Illumina HiSeq 2000 | 101 bp × 101 bp | *H. sapiens* | aligning reads against human reference genome and call variants |
| ERR091572 | real | 211,710,517 | Illumina HiSeq 2000 | 101 bp × 101 bp | *H. sapiens* |
| ERR091573 | real | 209,646,017 | Illumina HiSeq 2000 | 101 bp × 101 bp | *H. sapiens* |
| ERR091574 | real | 221,290,320 | Illumina HiSeq 2000 | 101 bp × 101 bp | *H. sapiens* |

Simulated and real datasets used for evaluating the performance of deBGA. a) The corresponding datasets were downloaded from NCBI Short Read Archive (SRA) site. b) The HiSeq X Ten dataset was downloaded from Illumina BaseSpace site, the “NA12878\_L3” sample of the “HiSeq X Ten: TruSeq Nano (4 replicates of NA12878)” project. c) For these real datasets, a proportion of reads were randomly extracted from each of them for benchmarking, the number of the reads extracted are given in this column. d) Refer to Supplementary Notes for the command lines of Mason simulator. e) The mean and standard deviation of the insert size of the simulated reads are respectively 500 and 25 bps. f) The bacteria genomes were downloaded from RefSeq. g) The genomes of the *A. thaliana* strains were downloaded from <http://mus.well.ox.ac.uk/19genomes>.

# Supplementary Table 2. *E. coli* strains involved in simulation datasets

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | ***E. coli* Strains**  **Name (Accession Number)a** | ***E. coli* Sim-1 b** | ***E. coli* Sim-10 b** | ***E. coli* Sim-20 b** | ***E. coli* Sim-40 b** | ***E. coli* Sim-62 b** | ***E. coli* Sim-2s b** | ***E. coli* Sim-2d b** |
| 1 | K\_12\_substr\_MG1655 (NC\_000913) |  |  |  |  |  |  |  |
| 2 | O157\_H7\_EDL933 (NC\_002655) |  |  |  |  |  |  |  |
| 3 | O157\_H7 (NC\_002695) |  |  |  |  |  |  |  |
| 4 | CFT073 (NC\_004431) |  |  |  |  |  |  |  |
| 5 | K\_12\_substr\_\_W3110 (NC\_007779) |  |  |  |  |  |  |  |
| 6 | UTI89 (NC\_007946) |  |  |  |  |  |  |  |
| 7 | 536 (NC\_008253) |  |  |  |  |  |  |  |
| 8 | APEC\_O1 (NC\_008563) |  |  |  |  |  |  |  |
| 9 | HS (NC\_009800) |  |  |  |  |  |  |  |
| 10 | E24377A (NC\_009801) |  |  |  |  |  |  |  |
| 11 | ATCC\_8739 (NC\_010468) |  |  |  |  |  |  |  |
| 12 | K\_12\_substr\_\_DH10B (NC\_010473) |  |  |  |  |  |  |  |
| 13 | SMS\_3\_5 (NC\_010498) |  |  |  |  |  |  |  |
| 14 | O157\_H7\_EC4115 (NC\_011353) |  |  |  |  |  |  |  |
| 15 | SE11 (NC\_011415) |  |  |  |  |  |  |  |
| 16 | O127\_H6\_E2348\_69 (NC\_011601) |  |  |  |  |  |  |  |
| 17 | IAI1 (NC\_011741) |  |  |  |  |  |  |  |
| 18 | S88 (NC\_011742) |  |  |  |  |  |  |  |
| 19 | ED1a (NC\_011745) |  |  |  |  |  |  |  |
| 20 | 55989 (NC\_011748) |  |  |  |  |  |  |  |
| 21 | IAI39 (NC\_011750) |  |  |  |  |  |  |  |
| 22 | UMN026 (NC\_011751) |  |  |  |  |  |  |  |
| 23 | LF82 (NC\_011993) |  |  |  |  |  |  |  |
| 24 | BW2952 (NC\_012759) |  |  |  |  |  |  |  |
| 25 | BL21\_DE3 (NC\_012892) |  |  |  |  |  |  |  |
| 26 | \_BL21\_Gold\_DE3\_pLysS\_AG (NC\_012947) |  |  |  |  |  |  |  |
| 27 | B\_REL606 (NC\_012967) |  |  |  |  |  |  |  |
| 28 | BL21\_DE3 (NC\_012971) |  |  |  |  |  |  |  |
| 29 | O157\_H7\_TW14359 (NC\_013008) |  |  |  |  |  |  |  |
| 30 | O103\_H2\_12009 (NC\_013353) |  |  |  |  |  |  |  |
| 31 | O26\_H11\_11368 (NC\_013361) |  |  |  |  |  |  |  |
| 32 | O111\_H\_\_11128 (NC\_013364) |  |  |  |  |  |  |  |
| 33 | SE15 (NC\_013654) |  |  |  |  |  |  |  |
| 34 | O55\_H7\_CB9615 (NC\_013941) |  |  |  |  |  |  |  |
| 35 | KO11FL (NC\_016902) |  |  |  |  |  |  |  |
| 36 | DH1 (NC\_017625) |  |  |  |  |  |  |  |
| 37 | 042 (NC\_017626) |  |  |  |  |  |  |  |
| 38 | IHE3034 (NC\_017628) |  |  |  |  |  |  |  |
| 39 | ABU\_83972 (NC\_017631) |  |  |  |  |  |  |  |
| 40 | UM146 (NC\_017632) |  |  |  |  |  |  |  |
| 41 | ETEC\_H10407 (NC\_017633) |  |  |  |  |  |  |  |
| 42 | O83\_H1\_NRG\_857C (NC\_017634) |  |  |  |  |  |  |  |
| 43 | W\_uid162011 (NC\_017635) |  |  |  |  |  |  |  |
| 44 | DH1 (NC\_017638) |  |  |  |  |  |  |  |
| 45 | NA114 (NC\_017644) |  |  |  |  |  |  |  |
| 46 | UMNK88 (NC\_017645) |  |  |  |  |  |  |  |
| 47 | O7\_K1\_CE10 (NC\_017646) |  |  |  |  |  |  |  |
| 48 | \_clone\_D\_i2 (NC\_017651) |  |  |  |  |  |  |  |
| 49 | \_clone\_D\_i14 (NC\_017652) |  |  |  |  |  |  |  |
| 50 | O55\_H7\_RM12579 (NC\_017656) |  |  |  |  |  |  |  |
| 51 | KO11FL (NC\_017660) |  |  |  |  |  |  |  |
| 52 | P12b (NC\_017663) |  |  |  |  |  |  |  |
| 53 | W\_uid162101 (NC\_017664) |  |  |  |  |  |  |  |
| 54 | Xuzhou21 (NC\_017906) |  |  |  |  |  |  |  |
| 55 | O104\_H4\_2009EL\_2050 (NC\_018650) |  |  |  |  |  |  |  |
| 56 | O104\_H4\_2011C\_3493 (NC\_018658) |  |  |  |  |  |  |  |
| 57 | O104\_H4\_2009EL\_2071 (NC\_018661) |  |  |  |  |  |  |  |
| 58 | APEC\_O78 (NC\_020163) |  |  |  |  |  |  |  |
| 59 | K\_12\_substr\_\_MDS42 (NC\_020518) |  |  |  |  |  |  |  |
| 60 | LY180 (NC\_022364) |  |  |  |  |  |  |  |
| 61 | PMV\_1 (NC\_022370) |  |  |  |  |  |  |  |
| 62 | JJ1886 (NC\_022648) |  |  |  |  |  |  |  |

*E. coli* strains involved in the seven simulation datasets (lines 1-7 of Supplementary Table 1). a) The names and the accession numbers of the reference sequences of *E.coli* strains. The names are obtained the from the original fasta files of the genome sequences. b) For a simulated datasets, the strains involved and not involved in are respectively marked by “●” and “○”.

# Supplementary Table 3. Alignment results on simulated datasets from *E. coli* genomes

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Aligner** | **Parameter** | **Aligned a** | | **Correctly aligned b** | | **Correctly aligned by primary alignment c** | | **Alignment Time d** | | **Index Loading Time e** | |
| # | % | # | % | # | % |
| ***E. coli* Sim-1 (from K-12 MG1655 substrain, 100 bp × 100 bp)** | | | | | | | | | |  | |
| 1 | Bowtie2 | very fast | 1998226 | 99.91% | 1979572 | 98.98% | 1979572 | 98.98% | 286 | | 1 | |
| 2 | Bowtie2 | fast | 1998233 | 99.91% | 1979532 | 98.98% | 1979532 | 98.98% | 287 | | 1 | |
| 3 | Bowtie2 | sensitive | 1999809 | 99.99% | 1981102 | 99.06% | 1981102 | 99.06% | 318 | | 1 | |
| 4 | Bowtie2 | very sensitive | 1999992 | 100.00% | 1981188 | 99.06% | 1981188 | 99.06% | 369 | | 1 | |
| 5 | BWA-MEM | default | 2000000 | 100.00% | 1980979 | 99.05% | 1980979 | 99.05% | 118 | | 1 | |
| 6 | BWA-MEM | -r 5 | 2000000 | 100.00% | 1980957 | 99.05% | 1980957 | 99.05% | 104 | | 1 | |
| 7 | BWA-MEM | -r 10 | 2000000 | 100.00% | 1980947 | 99.05% | 1980947 | 99.05% | 103 | | 1 | |
| 8 | BWA-MEM | -r 100 | 2000000 | 100.00% | 1980947 | 99.05% | 1980947 | 99.05% | 105 | | 1 | |
| 9 | SeqAlto | default | 1999994 | 100.00% | 1990607 | 99.53% | 1981232 | 99.06% | 825 | | 1 | |
| 10 | SeqAlto | fast | 1999994 | 100.00% | 1990614 | 99.53% | 1981264 | 99.06% | 823 | | 1 | |
| 11 | BWA | -n 0.04 | 1993509 | 99.68% | 1978460 | 98.92% | 1974305 | 98.72% | 309 | | 1 | |
| 12 | BWA | -n 0.04 sampe -n 1000 | 1993509 | 99.68% | 1992786 | 99.64% | 1974305 | 98.72% | 311 | | 1 | |
| 13 | BWA | -n 0.06 | 1990362 | 99.52% | 1975399 | 98.77% | 1971268 | 98.56% | 298 | | 1 | |
| 14 | BWA | -n 0.06 sampe -n 1000 | 1990362 | 99.52% | 1989559 | 99.47% | 1971268 | 98.56% | 299 | | 1 | |
| 15 | BWA | -n 0.08 | 1990362 | 99.52% | 1975399 | 98.77% | 1971268 | 98.56% | 299 | | 1 | |
| 16 | BWA | -n 0.08 sampe -n 1000 | 1990362 | 99.52% | 1989559 | 99.47% | 1971268 | 98.56% | 297 | | 1 | |
| 17 | GEM | -m 0.04 -e 0.04 | 1997990 | 99.90% | 1498108 | 74.91% | 1484790 | 74.24% | 116 | | 1 | |
| 18 | GEM | -m 0.06 -e 0.06 | 1999926 | 100.00% | 1499730 | 74.99% | 1486440 | 74.32% | 123 | | 1 | |
| 19 | GEM | -m 0.08 -e 0.08 | 1999996 | 100.00% | 1499870 | 74.99% | 1486569 | 74.33% | 111 | | 1 | |
| 20 | STAR f | default |  |  |  |  |  |  |  | |  | |
| 21 | deBGA | -k 22 -i 5 -n 300 -x 500 -o 300 (default) | 1999934 | 100.00% | 1999016 | 99.95% | 1980311 | 99.02% | 45 | | 1 | |
| 22 | deBGA | -k 22 -i 7 -n 300 -x 500 -o 300 | 1999932 | 100.00% | 1998827 | 99.94% | 1980206 | 99.01% | 45 | | 1 | |
| 23 | deBGA | -k 22 -i 9 -n 300 -x 500 -o 300 | 1999933 | 100.00% | 1998873 | 99.94% | 1980220 | 99.01% | 45 | | 1 | |
| 24 | deBGA | -k 24 -i 5 -n 300 -x 500 -o 300 | 1999950 | 100.00% | 1998982 | 99.95% | 1980342 | 99.02% | 45 | | 1 | |
| 25 | deBGA | -k 26 -i 5 -n 300 -x 500 -o 300 | 1999947 | 100.00% | 1998875 | 99.94% | 1980312 | 99.02% | 45 | | 1 | |
| 26 | deBGA | -k 28 -i 5 -n 300 -x 500 -o 300 | 1999942 | 100.00% | 1998753 | 99.94% | 1980270 | 99.01% | 45 | | 1 | |
| 27 | deBGA | -k 22 -i 5 -n 100 -x 500 -o 300 | 1999934 | 100.00% | 1999016 | 99.95% | 1980311 | 99.02% | 45 | | 1 | |
| 28 | deBGA | -k 22 -i 5 -n 500 -x 500 -o 300 | 1999934 | 100.00% | 1999016 | 99.95% | 1980311 | 99.02% | 47 | | 1 | |
| 29 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 -x 500 -o 300 | 1999974 | 100.00% | 1999691 | 99.98% | 1981025 | 99.05% | 46 | | 1 | |
| 30 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 -x 500 -o 300 | 1999993 | 100.00% | 1999723 | 99.99% | 1981057 | 99.05% | 45 | | 1 | |
| 31 | deBGA | -k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300 | 2000000 | 100.00% | 1999731 | 99.99% | 1981065 | 99.05% | 46 | | 1 | |
| 32 | deBGA | -k 22 -i 5 -n 300 --local -x 500 -o 300 | 1999998 | 100.00% | 1999705 | 99.99% | 1981039 | 99.05% | 45 | | 1 | |
| ***E. coli* Sim-10 (from 10 various *E. Coli* strains, 100 bp × 100 bp)** | | | | | | | | | |  | |
| 33 | Bowtie2 | very fast | 1998320 | 99.92% | 1087924 | 54.40% | 1087924 | 54.40% | 897 | | 1 | |
| 34 | Bowtie2 | fast | 1998367 | 99.92% | 1089024 | 54.45% | 1089024 | 54.45% | 1011 | | 1 | |
| 35 | Bowtie2 | sensitive | 1999760 | 99.99% | 1090274 | 54.51% | 1090274 | 54.51% | 1178 | | 1 | |
| 36 | Bowtie2 | very sensitive | 1999980 | 100.00% | 1092161 | 54.61% | 1092161 | 54.61% | 1304 | | 1 | |
| 37 | BWA-MEM | default | 2000000 | 100.00% | 1090355 | 54.52% | 1090355 | 54.52% | 448 | | 1 | |
| 38 | BWA-MEM | -r 5 | 2000000 | 100.00% | 1090624 | 54.53% | 1090624 | 54.53% | 426 | | 1 | |
| 39 | BWA-MEM | -r 10 | 2000000 | 100.00% | 1090620 | 54.53% | 1090620 | 54.53% | 412 | | 1 | |
| 40 | BWA-MEM | -r 100 | 2000000 | 100.00% | 1090620 | 54.53% | 1090620 | 54.53% | 414 | | 1 | |
| 41 | SeqAlto | default | 2000000 | 100.00% | 1540915 | 77.05% | 1090071 | 54.50% | 4075 | | 3 | |
| 42 | SeqAlto | fast | 2000000 | 100.00% | 1541591 | 77.08% | 1091486 | 54.57% | 4041 | | 3 | |
| 43 | BWA | -n 0.04 | 1919722 | 95.99% | 1284585 | 64.23% | 1046011 | 52.30% | 525 | | 1 | |
| 44 | BWA | -n 0.04 sampe -n 1000 | 1919722 | 95.99% | 1919104 | 95.96% | 1046011 | 52.30% | 672 | | 1 | |
| 45 | BWA | -n 0.06 | 1899931 | 95.00% | 1271761 | 63.59% | 1034489 | 51.72% | 503 | | 1 | |
| 46 | BWA | -n 0.06 sampe -n 1000 | 1899931 | 95.00% | 1899323 | 94.97% | 1034489 | 51.72% | 638 | | 1 | |
| 47 | BWA | -n 0.08 | 1899931 | 95.00% | 1271761 | 63.59% | 1034489 | 51.72% | 499 | | 1 | |
| 48 | BWA | -n 0.08 sampe -n 1000 | 1899931 | 95.00% | 1899323 | 94.97% | 1034489 | 51.72% | 635 | | 1 | |
| 49 | GEM | -m 0.04 -e 0.04 | 1997521 | 99.88% | 1499890 | 74.99% | 814183 | 40.71% | 367 | | 1 | |
| 50 | GEM | -m 0.06 -e 0.06 | 1999802 | 99.99% | 1501594 | 75.08% | 814532 | 40.73% | 371 | | 1 | |
| 51 | GEM | -m 0.08 -e 0.08 | 1999894 | 99.99% | 1501350 | 75.07% | 814406 | 40.72% | 462 | | 1 | |
| 52 | STAR | default | 1985242 | 99.26% | 1973784 | 98.69% | 1084194 | 54.21% | 699 | | 26 | |
| 53 | deBGA | -k 22 -i 5 -n 300 -x 500 -o 300 (default) | 1999971 | 100.00% | 1988941 | 99.45% | 1080553 | 54.03% | 80 | | 1 | |
| 54 | deBGA | -k 22 -i 7 -n 300 -x 500 -o 300 | 1999972 | 100.00% | 1985167 | 99.26% | 1078176 | 53.91% | 78 | | 1 | |
| 55 | deBGA | -k 22 -i 9 -n 300 -x 500 -o 300 | 1999968 | 100.00% | 1986816 | 99.34% | 1078886 | 53.94% | 78 | | 1 | |
| 56 | deBGA | -k 24 -i 5 -n 300 -x 500 -o 300 | 1999970 | 100.00% | 1986826 | 99.34% | 1079795 | 53.99% | 80 | | 1 | |
| 57 | deBGA | -k 26 -i 5 -n 300 -x 500 -o 300 | 1999962 | 100.00% | 1984047 | 99.20% | 1078829 | 53.94% | 78 | | 1 | |
| 58 | deBGA | -k 28 -i 5 -n 300 -x 500 -o 300 | 1999953 | 100.00% | 1980908 | 99.05% | 1077539 | 53.88% | 78 | | 1 | |
| 59 | deBGA | -k 22 -i 5 -n 100 -x 500 -o 300 | 1999971 | 100.00% | 1988936 | 99.45% | 1080546 | 54.03% | 82 | | 1 | |
| 60 | deBGA | -k 22 -i 5 -n 500 -x 500 -o 300 | 1999971 | 100.00% | 1988945 | 99.45% | 1080553 | 54.03% | 80 | | 1 | |
| 61 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 -x 500 -o 300 | 2000000 | 100.00% | 1989571 | 99.48% | 1081535 | 54.08% | 82 | | 1 | |
| 62 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 -x 500 -o 300 | 2000000 | 100.00% | 1989567 | 99.48% | 1081533 | 54.08% | 81 | | 1 | |
| 63 | deBGA | -k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300 | 2000000 | 100.00% | 1989563 | 99.48% | 1081533 | 54.08% | 81 | | 1 | |
| 64 | deBGA | -k 22 -i 5 -n 300 --local -x 500 -o 300 | 2000000 | 100.00% | 1987217 | 99.36% | 1080251 | 54.01% | 82 | | 1 | |
| ***E. coli* Sim-20 (from 20 various *E. Coli* strains, 100 bp × 100 bp)** | | | | | | | | | |  | |
| 65 | Bowtie2 | very fast | 1997957 | 99.90% | 597950 | 29.90% | 597950 | 29.90% | 1138 | | 1 | |
| 66 | Bowtie2 | fast | 1998312 | 99.92% | 607565 | 30.38% | 607565 | 30.38% | 1420 | | 1 | |
| 67 | Bowtie2 | sensitive | 1999816 | 99.99% | 611600 | 30.58% | 611600 | 30.58% | 1765 | | 1 | |
| 68 | Bowtie2 | very sensitive | 1999980 | 100.00% | 610132 | 30.51% | 610132 | 30.51% | 1965 | | 1 | |
| 69 | BWA-MEM | default | 2000000 | 100.00% | 611897 | 30.59% | 611897 | 30.59% | 744 | | 1 | |
| 70 | BWA-MEM | -r 5 | 2000000 | 100.00% | 611960 | 31.00% | 611960 | 31.00% | 682 | | 1 | |
| 71 | BWA-MEM | -r 10 | 2000000 | 100.00% | 611844 | 30.59% | 611844 | 30.59% | 670 | | 1 | |
| 72 | BWA-MEM | -r 100 | 2000000 | 100.00% | 611844 | 30.59% | 611844 | 30.59% | 666 | | 1 | |
| 73 | SeqAlto | default | 1999996 | 100.00% | 1252221 | 62.61% | 609747 | 30.49% | 8621 | | 4 | |
| 74 | SeqAlto | fast | 1999994 | 100.00% | 1252785 | 62.64% | 610677 | 30.53% | 8423 | | 4 | |
| 75 | BWA | -n 0.04 | 1909440 | 95.47% | 657966 | 32.90% | 582248 | 29.11% | 582 | | 1 | |
| 76 | BWA | -n 0.04 sampe -n 1000 | 1909440 | 95.47% | 1908803 | 95.44% | 582248 | 29.11% | 961 | | 1 | |
| 77 | BWA | -n 0.06 | 1887038 | 94.35% | 650151 | 32.51% | 574949 | 28.75% | 553 | | 1 | |
| 78 | BWA | -n 0.06 sampe -n 1000 | 1887038 | 94.35% | 1886456 | 94.32% | 574949 | 28.75% | 914 | | 1 | |
| 79 | BWA | -n 0.08 | 1887038 | 94.35% | 650151 | 32.51% | 574949 | 28.75% | 548 | | 1 | |
| 80 | BWA | -n 0.08 sampe -n 1000 | 1887038 | 94.35% | 1886456 | 94.32% | 574949 | 28.75% | 915 | | 1 | |
| 81 | GEM | -m 0.04 -e 0.04 | 1997158 | 99.86% | 1501991 | 75.10% | 457459 | 22.87% | 576 | | 3 | |
| 82 | GEM | -m 0.06 -e 0.06 | 1999701 | 99.99% | 1503685 | 75.18% | 457682 | 22.88% | 620 | | 3 | |
| 83 | GEM | -m 0.08 -e 0.08 | 1999818 | 99.99% | 1503385 | 75.17% | 457599 | 22.88% | 680 | | 3 | |
| 84 | STAR | default | 1504020 | 75.20% | 1491989 | 74.60% | 567687 | 28.38% | 908 | | 29 | |
| 85 | deBGA | -k 22 -i 5 -n 300 -x 500 -o 300 (default) | 1999960 | 100.00% | 1985772 | 99.29% | 605829 | 30.29% | 112 | | 1 | |
| 86 | deBGA | -k 22 -i 7 -n 300 -x 500 -o 300 | 1999956 | 100.00% | 1979523 | 98.98% | 604380 | 30.22% | 109 | | 1 | |
| 87 | deBGA | -k 22 -i 9 -n 300 -x 500 -o 300 | 1999955 | 100.00% | 1981789 | 99.09% | 604464 | 30.22% | 108 | | 1 | |
| 88 | deBGA | -k 24 -i 5 -n 300 -x 500 -o 300 | 1999957 | 100.00% | 1982683 | 99.13% | 605229 | 30.26% | 108 | | 1 | |
| 89 | deBGA | -k 26 -i 5 -n 300 -x 500 -o 300 | 1999951 | 100.00% | 1978793 | 98.94% | 604974 | 30.25% | 107 | | 1 | |
| 90 | deBGA | -k 28 -i 5 -n 300 -x 500 -o 300 | 1999950 | 100.00% | 1974258 | 98.71% | 604602 | 30.23% | 105 | | 1 | |
| 91 | deBGA | -k 22 -i 5 -n 100 -x 500 -o 300 | 1999946 | 100.00% | 1985077 | 99.25% | 605779 | 30.29% | 117 | | 1 | |
| 92 | deBGA | -k 22 -i 5 -n 500 -x 500 -o 300 | 1999960 | 100.00% | 1985791 | 99.29% | 605831 | 30.29% | 112 | | 1 | |
| 93 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 -x 500 -o 300 | 1999998 | 100.00% | 1986498 | 99.32% | 606347 | 30.32% | 115 | | 1 | |
| 94 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 -x 500 -o 300 | 1999998 | 100.00% | 1986497 | 99.32% | 606348 | 30.32% | 109 | | 1 | |
| 95 | deBGA | -k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300 | 1999998 | 100.00% | 1986499 | 99.32% | 606348 | 30.32% | 111 | | 1 | |
| 96 | deBGA | -k 22 -i 5 -n 300 --local -x 500 -o 300 | 1999997 | 100.00% | 1983792 | 99.19% | 605629 | 30.28% | 117 | | 1 | |
| ***E. coli* Sim-40 (from 40 various *E. Coli* strains, 100 bp × 100 bp)** | | | | | | | | | |  | |
| 97 | Bowtie2 | very fast | 1997979 | 99.90% | 494249 | 24.71% | 494249 | 24.71% | 1335 | | 2 | |
| 98 | Bowtie2 | fast | 1998171 | 99.91% | 502664 | 25.13% | 502664 | 25.13% | 1640 | | 2 | |
| 99 | Bowtie2 | sensitive | 1999776 | 99.99% | 509269 | 25.46% | 509269 | 25.46% | 2130 | | 2 | |
| 100 | Bowtie2 | very sensitive | 1999982 | 100.00% | 510542 | 25.53% | 510542 | 25.53% | 2400 | | 2 | |
| 101 | BWA-MEM | default | 2000000 | 100.00% | 511363 | 25.57% | 511363 | 25.57% | 1283 | | 4 | |
| 102 | BWA-MEM | -r 5 | 2000000 | 100.00% | 511052 | 25.55% | 511052 | 25.55% | 1213 | | 4 | |
| 103 | BWA-MEM | -r 10 | 2000000 | 100.00% | 510922 | 25.55% | 510922 | 25.55% | 1184 | | 4 | |
| 104 | BWA-MEM | -r 100 | 2000000 | 100.00% | 510922 | 25.55% | 510922 | 25.55% | 1175 | | 4 | |
| 105 | SeqAlto | default | 1999998 | 100.00% | 1069731 | 53.49% | 510085 | 25.50% | 15654 | | 7 | |
| 106 | SeqAlto | fast | 1999992 | 100.00% | 1070120 | 53.51% | 510741 | 25.54% | 14528 | | 7 | |
| 107 | BWA | -n 0.04 | 1906212 | 95.31% | 558963 | 27.95% | 482389 | 24.12% | 685 | | 4 | |
| 108 | BWA | -n 0.04 sampe -n 1000 | 1906212 | 95.31% | 1905201 | 95.26% | 482389 | 24.12% | 1388 | | 4 | |
| 109 | BWA | -n 0.06 | 1883099 | 94.15% | 551266 | 27.56% | 475281 | 23.76% | 647 | | 4 | |
| 110 | BWA | -n 0.06 sampe -n 1000 | 1883099 | 94.15% | 1882126 | 94.11% | 475281 | 23.76% | 1325 | | 4 | |
| 111 | BWA | -n 0.08 | 1883099 | 94.15% | 551266 | 27.56% | 475281 | 23.76% | 645 | | 4 | |
| 112 | BWA | -n 0.08 sampe -n 1000 | 1883099 | 94.15% | 1882126 | 94.11% | 475281 | 23.76% | 1314 | | 4 | |
| 113 | GEM | -m 0.04 -e 0.04 | 1979337 | 98.97% | 1605348 | 80.27% | 395989 | 19.80% | 743 | | 5 | |
| 114 | GEM | -m 0.06 -e 0.06 | 1996755 | 99.84% | 1616264 | 80.81% | 398742 | 19.94% | 783 | | 5 | |
| 115 | GEM | -m 0.08 -e 0.08 | 1998999 | 99.95% | 1615337 | 80.77% | 398723 | 19.94% | 822 | | 5 | |
| 116 | STAR | default | 1319918 | 66.00% | 1303448 | 65.17% | 466210 | 23.31% | 720 | | 41 | |
| 117 | deBGA | -k 22 -i 5 -n 300 -x 500 -o 300 (default) | 1999967 | 100.00% | 1979386 | 98.97% | 506122 | 25.31% | 155 | | 1 | |
| 118 | deBGA | -k 22 -i 7 -n 300 -x 500 -o 300 | 1999969 | 100.00% | 1972486 | 98.62% | 504085 | 25.20% | 143 | | 1 | |
| 119 | deBGA | -k 22 -i 9 -n 300 -x 500 -o 300 | 1999967 | 100.00% | 1975074 | 98.75% | 504153 | 25.21% | 142 | | 1 | |
| 120 | deBGA | -k 24 -i 5 -n 300 -x 500 -o 300 | 1999962 | 100.00% | 1976041 | 98.80% | 505427 | 25.27% | 148 | | 1 | |
| 121 | deBGA | -k 26 -i 5 -n 300 -x 500 -o 300 | 1999962 | 100.00% | 1972119 | 98.61% | 505084 | 25.25% | 145 | | 1 | |
| 122 | deBGA | -k 28 -i 5 -n 300 -x 500 -o 300 | 1999956 | 100.00% | 1967262 | 98.36% | 504370 | 25.22% | 143 | | 1 | |
| 123 | deBGA | -k 22 -i 5 -n 100 -x 500 -o 300 | 1999766 | 99.99% | 1974572 | 98.73% | 505710 | 25.29% | 149 | | 1 | |
| 124 | deBGA | -k 22 -i 5 -n 500 -x 500 -o 300 | 1999967 | 100.00% | 1979388 | 98.97% | 506110 | 25.31% | 151 | | 1 | |
| 125 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 -x 500 -o 300 | 2000000 | 100.00% | 1980340 | 99.02% | 506623 | 25.33% | 150 | | 1 | |
| 126 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 -x 500 -o 300 | 2000000 | 100.00% | 1980339 | 99.02% | 506623 | 25.33% | 155 | | 1 | |
| 127 | deBGA | -k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300 | 2000000 | 100.00% | 1980333 | 99.02% | 506623 | 25.33% | 153 | | 1 | |
| 128 | deBGA | -k 22 -i 5 -n 300 --local -x 500 -o 300 | 2000000 | 100.00% | 1977442 | 98.87% | 505997 | 25.30% | 160 | | 1 | |
| ***E. coli* Sim-62 (from 62 various *E. coli* strains, 100 bp × 100 bp)** | | | | | | | | | |  | |
| **129** | Bowtie2 | very fast | 1997910 | 99.90% | 419988 | 21.00% | 419988 | 21.00% | 1505 | | 4 | |
| 130 | Bowtie2 | fast | 1998089 | 99.90% | 426257 | 21.31% | 426257 | 21.31% | 1754 | | 4 | |
| 131 | **Bowtie2** | **sensitive** | **1999766** | **99.99%** | **436978** | **21.85%** | **436978** | **21.85%** | **2366** | | **4** | |
| 132 | Bowtie2 | very sensitive | 1999968 | 100.00% | 439690 | 21.98% | 439690 | 21.98% | 2833 | | 4 | |
| **133** | **BWA-MEM** | **default** | **2000000** | **100.00%** | **441153** | **22.06%** | **441153** | **22.06%** | **1717** | | **5** | |
| 134 | BWA-MEM | -r 5 | 2000000 | 100.00% | 441165 | 22.06% | 441165 | 22.06% | 1650 | | 5 | |
| 135 | BWA-MEM | -r 10 | 2000000 | 100.00% | 441197 | 22.06% | 441197 | 22.06% | 1624 | | 5 | |
| **136** | BWA-MEM | -r 100 | 2000000 | 100.00% | 441197 | 22.06% | 441197 | 22.06% | 1591 | | 5 | |
| 137 | **SeqAlto** | **default** | **1999994** | **100.00%** | **958119** | **47.91%** | **441678** | **22.08%** | **22323** | | 10 | |
| 138 | SeqAlto | fast | 1999914 | 100.00% | 956508 | 47.83% | 440836 | 22.04% | 19218 | | 10 | |
| 139 | BWA | -n 0.04 | 1905980 | 95.30% | 463848 | 23.19% | 417036 | 20.85% | 741 | | 5 | |
| 140 | **BWA** | **-n 0.04 sampe -n 1000** | **1905980** | **95.30%** | **1904405** | **95.22%** | **417036** | **20.85%** | **1700** | | **5** | |
| **141** | BWA | -n 0.06 | 1882914 | 94.14% | 457628 | 22.88% | 411163 | 20.56% | 704 | | 5 | |
| **142** | BWA | -n 0.06 sampe -n 1000 | 1882914 | 94.14% | 1881409 | 94.07% | 411163 | 20.56% | 1621 | | 5 | |
| **143** | BWA | -n 0.08 | 1882914 | 94.14% | 457628 | 22.88% | 411163 | 20.56% | 699 | | 5 | |
| **144** | BWA | -n 0.08 sampe -n 1000 | 1882914 | 94.14% | 1881409 | 94.07% | 411163 | 20.56% | 1617 | | 5 | |
| **145** | **GEM** | **-m 0.04 -e 0.04** | **1971831** | **98.59%** | **1622074** | **81.10%** | **346170** | **17.31%** | **896** | | **8** | |
| 146 | GEM | -m 0.06 -e 0.06 | 1993103 | 99.66% | 1635651 | 81.78% | 349512 | 17.48% | 959 | | 8 | |
| 147 | GEM | -m 0.08 -e 0.08 | 1996095 | 99.80% | 1634656 | 81.73% | 349398 | 17.47% | 1046 | | 8 | |
| **148** | **STAR** | **default** | **1100946** | **55.05%** | **1079133** | **53.96%** | **387323** | **19.37%** | **557** | | **41** | |
| **149** | **deBGA** | **-k 22 -i 5 -n 300 -x 500 -o 300 (default)** | **1999961** | **100.00%** | **1975380** | **98.77%** | **437786** | **21.89%** | **196** | | **1** | |
| 150 | deBGA | -k 22 -i 7 -n 300 -x 500 -o 300 | 1999961 | 100.00% | 1967278 | 98.36% | 436127 | 21.81% | 184 | | 1 | |
| 151 | deBGA | -k 22 -i 9 -n 300 -x 500 -o 300 | 1999964 | 100.00% | 1970433 | 98.52% | 435646 | 21.78% | 181 | | 1 | |
| 152 | deBGA | -k 24 -i 5 -n 300 -x 500 -o 300 | 1999960 | 100.00% | 1971957 | 98.60% | 437499 | 21.87% | 188 | | 1 | |
| 153 | deBGA | -k 26 -i 5 -n 300 -x 500 -o 300 | 1999955 | 100.00% | 1967717 | 98.39% | 437529 | 21.88% | 185 | | 1 | |
| 154 | deBGA | -k 28 -i 5 -n 300 -x 500 -o 300 | 1999943 | 100.00% | 1962661 | 98.13% | 436465 | 21.82% | 183 | | 1 | |
| 155 | deBGA | -k 22 -i 5 -n 100 -x 500 -o 300 | 1999641 | 99.98% | 1971588 | 98.58% | 437197 | 21.86% | 196 | | 1 | |
| 156 | deBGA | -k 22 -i 5 -n 500 -x 500 -o 300 | 1999961 | 100.00% | 1975399 | 98.77% | 437800 | 21.89% | 199 | | 1 | |
| 157 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 -x 500 -o 300 | 2000000 | 100.00% | 1976937 | 98.85% | 438269 | 21.91% | 193 | | 1 | |
| 158 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 -x 500 -o 300 | 2000000 | 100.00% | 1976935 | 98.85% | 438268 | 21.91% | 195 | | 1 | |
| **159** | **deBGA** | **-k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300** | **2000000** | **100.00%** | **1976929** | **98.85%** | **438268** | **21.91%** | **199** | | **1** | |
| 160 | deBGA | -k 22 -i 5 -n 300 --local -x 500 -o 300 | 2000000 | 100.00% | 1973825 | 98.69% | 437657 | 21.88% | 202 | | 1 | |
| 161 | GenomeMapper g | -M 4 -G 2 -E 4 | 943112 | 94.31% | 940316 | 94.03% | 135392 | 13.54% | 32291 | |  | |
| ***E. coli* Sim-2d (from two divergent *E. coli* strains, (K-12 MG1655 substrain (NC\_000913) and O157:H7 strain (NC\_002695), 100 bp × 100 bp)** | | | | | | | | | | | | |
| 162 | Bowtie2 | very fast | 1998407 | 99.92% | 1836338 | 91.82% | 1836338 | 91.82% | 542 | | 4 | |
| 163 | Bowtie2 | fast | 1998421 | 99.92% | 1836370 | 91.82% | 1836370 | 91.82% | 549 | | 4 | |
| **164** | **Bowtie2** | **sensitive** | **1999815** | **99.99%** | **1838834** | **91.94%** | **1838834** | **91.94%** | **678** | | **4** | |
| 165 | Bowtie2 | very sensitive | 1999987 | 100.00% | 1838684 | 91.93% | 1838684 | 91.93% | 655 | | 4 | |
| **166** | **BWA-MEM** | **default** | **2000000** | **100.00%** | **1836549** | **91.83%** | **1836549** | **91.83%** | **197** | | **5** | |
| 167 | BWA-MEM | -r 5 | 2000000 | 100.00% | 1836587 | 91.83% | 1836587 | 91.83% | 163 | | 5 | |
| 168 | BWA-MEM | -r 10 | 2000000 | 100.00% | 1836555 | 91.83% | 1836555 | 91.83% | 172 | | 5 | |
| 169 | BWA-MEM | -r 100 | 2000000 | 100.00% | 1836555 | 91.83% | 1836555 | 91.83% | 214 | | 5 | |
| **170** | **SeqAlto** | **default** | **1999998** | **100.00%** | **1917662** | **95.88%** | **1835778** | **91.79%** | **1230** | | **10** | |
| 171 | SeqAlto | fast | 1999998 | 100.00% | 1917886 | 95.89% | 1836203 | 91.81% | 1228 | | 10 | |
| 172 | BWA | -n 0.04 | 1964289 | 98.21% | 1944192 | 97.21% | 1800221 | 90.01% | 466 | | 5 | |
| **173** | **BWA** | **-n 0.04 sampe -n 1000** | **1964289** | **98.21%** | **1963656** | **98.18%** | **1800221** | **90.01%** | **442** | | **5** | |
| 174 | BWA | -n 0.06 | 1954496 | 97.72% | 1934539 | 96.73% | 1790673 | 89.53% | 391 | | 5 | |
| 175 | BWA | -n 0.06 sampe -n 1000 | 1954596 | 97.73% | 1953823 | 97.69% | 1790673 | 89.53% | 336 | | 5 | |
| 176 | BWA | -n 0.08 | 1954496 | 97.72% | 1934539 | 96.73% | 1790673 | 89.53% | 388 | | 5 | |
| 177 | BWA | -n 0.08 sampe -n 1000 | 1954496 | 97.72% | 1953823 | 97.69% | 1790673 | 89.53% | 422 | | 5 | |
| **178** | **GEM** | **-m 0.04 -e 0.04** | **1998160** | **99.91%** | **1498059** | **74.90%** | **1376008** | **68.80%** | **144** | | **8** | |
| 179 | GEM | -m 0.06 -e 0.06 | 1999951 | 100.00% | 1499297 | 74.96% | 1376735 | 68.84% | 155 | | 8 | |
| 180 | GEM | -m 0.08 -e 0.08 | 1999998 | 100.00% | 1499302 | 74.97% | 1376694 | 68.83% | 170 | | 8 | |
| 181 | STAR f | default |  |  |  |  |  |  |  | |  | |
| **182** | **deBGA** | **-k 22 -i 5 -n 300 -x 500 -o 300 (default)** | **1999972** | **100.00%** | **1997011** | **99.85%** | **1830211** | **91.51%** | **70** | | **1** | |
| 183 | deBGA | -k 22 -i 7 -n 300 -x 500 -o 300 | 1999976 | 100.00% | 1996529 | 99.83% | 1829074 | 91.45% | 67 | | 1 | |
| 184 | deBGA | -k 22 -i 9 -n 300 -x 500 -o 300 | 1999996 | 100.00% | 1996750 | 99.84% | 1829146 | 91.46% | 62 | | 1 | |
| 185 | deBGA | -k 24 -i 5 -n 300 -x 500 -o 300 | 1999970 | 100.00% | 1996742 | 99.84% | 1830079 | 91.50% | 56 | | 1 | |
| 186 | deBGA | -k 26 -i 5 -n 300 -x 500 -o 300 | 1999959 | 100.00% | 1996467 | 99.82% | 1829911 | 91.50% | 58 | | 1 | |
| 187 | deBGA | -k 28 -i 5 -n 300 -x 500 -o 300 | 1999956 | 100.00% | 1996113 | 99.81% | 1829689 | 91.48% | 69 | | 1 | |
| 188 | deBGA | -k 22 -i 5 -n 100 -x 500 -o 300 | 1999972 | 100.00% | 1997011 | 99.85% | 1830211 | 91.51% | 59 | | 1 | |
| 189 | deBGA | -k 22 -i 5 -n 500 -x 500 -o 300 | 1999972 | 100.00% | 1997011 | 99.85% | 1830211 | 91.51% | 57 | | 1 | |
| 190 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 -x 500 -o 300 | 2000000 | 100.00% | 1997554 | 99.88% | 1831026 | 91.55% | 75 | | 1 | |
| 191 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 -x 500 -o 300 | 2000000 | 100.00% | 1997546 | 99.88% | 1831024 | 91.55% | 69 | | 1 | |
| **192** | **deBGA** | **-k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300** | **2000000** | **100.00%** | **1997540** | **99.88%** | **1831022** | **91.55%** | **77** | | **1** | |
| 193 | deBGA | -k 22 -i 5 -n 300 --local -x 500 -o 300 | 1999999 | 100.00% | 1996915 | 99.85% | 1830252 | 91.51% | 80 | | 1 | |
| ***E. coli* Sim-2s (from two similar *E. coli* strains, K-12 MG1655 substrain (NC\_000913) and K-12 W3110 substrain (NC\_007779), 100 bp × 100 bp)** | | | | | | | | | | | | |
| 194 | Bowtie2 | very fast | 1998237 | 99.91% | 993259 | 49.66% | 993259 | 49.66% | 555 | | 4 | |
| **195** | Bowtie2 | fast | 1998242 | 99.91% | 992891 | 49.64% | 992891 | 49.64% | 533 | | 4 | |
| **196** | **Bowtie2** | **sensitive** | **1999790** | **99.99%** | **993946** | **49.70%** | **993946** | **49.70%** | **558** | | **4** | |
| **197** | Bowtie2 | very sensitive | 1999991 | 100.00% | 993606 | 49.68% | 993606 | 49.68% | 613 | | 4 | |
| **198** | **BWA-MEM** | **default** | **2000000** | **100.00%** | **992560** | **49.63%** | **992560** | **49.63%** | **225** | | **5** | |
| 199 | BWA-MEM | -r 5 | 2000000 | 100.00% | 992780 | 49.64% | 992780 | 49.64% | 201 | | 5 | |
| 200 | BWA-MEM | -r 10 | 2000000 | 100.00% | 992782 | 49.64% | 992782 | 49.64% | 175 | | 5 | |
| **201** | BWA-MEM | -r 100 | 2000000 | 100.00% | 992782 | 49.64% | 992782 | 49.64% | 169 | | 5 | |
| **202** | **SeqAlto** | **default** | **1999992** | **100.00%** | **1495249** | **74.76%** | **994653** | **49.73%** | **1742** | | **10** | |
| 203 | SeqAlto | fast | 1999992 | 100.00% | 1494807 | 74.74% | 993748 | 49.69% | 1695 | | 10 | |
| **204** | BWA | -n 0.04 | 1902932 | 95.15% | 1877457 | 93.87% | 944575 | 47.23% | 436 | | 5 | |
| **205** | **BWA** | **-n 0.04 sampe -n 1000** | **1902932** | **95.15%** | **1902926** | **95.15%** | **944575** | **47.23%** | **463** | | **5** | |
| 206 | BWA | -n 0.06 | 1879190 | 93.96% | 1854053 | 92.70% | 932715 | 46.64% | 383 | | 5 | |
| 207 | BWA | -n 0.06 sampe -n 1000 | 1879190 | 93.96% | 1879185 | 93.96% | 932715 | 46.64% | 398 | | 5 | |
| 208 | BWA | -n 0.08 | 1879190 | 93.96% | 1854053 | 92.70% | 932715 | 46.64% | 401 | | 5 | |
| **209** | BWA | -n 0.08 sampe -n 1000 | 1879190 | 93.96% | 1879185 | 93.96% | 932715 | 46.64% | 378 | | 5 | |
| **210** | **GEM** | **-m 0.04 -e 0.04** | **1997969** | **99.90%** | **1499138** | **74.96%** | **743781** | **37.19%** | **170** | | **8** | |
| 211 | GEM | -m 0.06 -e 0.06 | 1999896 | 99.99% | 1500713 | 75.04% | 744541 | 37.23% | 163 | | 8 | |
| 212 | GEM | -m 0.08 -e 0.08 | 1999985 | 100.00% | 1500840 | 75.04% | 744570 | 37.23% | 258 | | 8 | |
| **213** | STAR f | default |  |  |  |  |  |  |  | |  | |
| **214** | **deBGA** | **-k 22 -i 5 -n 300 -x 500 -o 300 (default)** | **1999963** | **100.00%** | **1998365** | **99.92%** | **992219** | **49.61%** | **65** | | **1** | |
| 215 | deBGA | -k 22 -i 7 -n 300 -x 500 -o 300 | 1999962 | 100.00% | 1997006 | 99.85% | 991453 | 49.57% | 55 | | 1 | |
| 216 | deBGA | -k 22 -i 9 -n 300 -x 500 -o 300 | 1999959 | 100.00% | 1997461 | 99.87% | 991953 | 49.60% | 58 | | 1 | |
| 217 | deBGA | -k 24 -i 5 -n 300 -x 500 -o 300 | 1999964 | 100.00% | 1997679 | 99.88% | 992347 | 49.62% | 72 | | 1 | |
| 218 | deBGA | -k 26 -i 5 -n 300 -x 500 -o 300 | 1999960 | 100.00% | 1996685 | 99.83% | 992199 | 49.61% | 74 | | 1 | |
| 219 | deBGA | -k 28 -i 5 -n 300 -x 500 -o 300 | 1999956 | 100.00% | 1995571 | 99.78% | 992399 | 49.62% | 60 | | 1 | |
| 220 | deBGA | -k 22 -i 5 -n 100 -x 500 -o 300 | 1999963 | 100.00% | 1998365 | 99.92% | 992219 | 49.61% | 69 | | 1 | |
| 221 | deBGA | -k 22 -i 5 -n 500 -x 500 -o 300 | 1999963 | 100.00% | 1998365 | 99.92% | 992219 | 49.61% | 99 | | 1 | |
| 222 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 -x 500 -o 300 | 1999998 | 100.00% | 1998997 | 99.95% | 992591 | 49.63% | 76 | | 1 | |
| **223** | deBGA | -k 22 -i 5 -n 300 --cl 0.25 -x 500 -o 300 | 1999998 | 100.00% | 1998997 | 99.95% | 992591 | 49.63% | 81 | | 1 | |
| **224** | **deBGA** | **-k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300** | **2000000** | **100.00%** | **1998999** | **99.95%** | **992591** | **49.63%** | **76** | | **1** | |
| 225 | deBGA | -k 22 -i 5 -n 300 --local -x 500 -o 300 | 1999998 | 100.00% | 1998716 | 99.94% | 992554 | 49.63% | 55 | | 1 | |

Results on simulated *E.coli* datasets (lines 1-7 of Supplementary Table 1). a) - e) Various statistics on the sensitivities, correctness and throughputs of the aligners. Refer to Supplementary Notes (the “Statistics for evaluating alignments” subsection) for definition. The highlighted lines are involved in Fig. 2a and Supplementary Fig. 7. f) On the three datasets, *E. coli* Sim-1, *E. coli* Sim-2d and *E. coli* Sim-2s, the STAR program corrupted, and the results are not available. g) GenomeMapper is only asked to align half of the dataset, due to its slow speed. As the tool could align a read to multiple positions, but does not distinguish the primary and alternative alignments, we considered the first alignment of each read as its primary alignment.

# Supplementary Table 4. The repetitiveness of the references

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | **References** | ***k* = 22** | | ***k* = 24** | | ***k* = 26** | | ***k* = 28** | |
| **#distinct** |  | **#distinct** |  | **#distinct** |  | **#distinct** |  |
| 1 | *E. coli* K-12 MG1655 substrain | 4563650 | 1.02 | 4565593 | 1.02 | 4567279 | 1.02 | 4568778 | 1.02 |
| 2 | 2 divergent *E. coli* strains | 6832544 | 1.48 | 6909030 | 1.47 | 6981205 | 1.45 | 7050645 | 1.44 |
| 3 | 2 similar *E. coli* strains | 5318145 | 1.75 | 5320599 | 1.75 | 5322733 | 1.74 | 5324634 | 1.74 |
| 4 | 10 *E. coli* strains | 16455348 | 3.02 | 16778160 | 2.96 | 17088582 | 2.90 | 17390925 | 2.85 |
| 5 | 20 *E. coli* strains | 18402918 | 5.21 | 18812012 | 5.10 | 19205832 | 4.99 | 19589616 | 4.89 |
| 6 | 40 *E. coli* strains | 25730939 | 7.78 | 26460215 | 7.57 | 27166579 | 7.37 | 27857882 | 7.19 |
| 7 | 62 *E. coli* strains | 31313027 | 9.91 | 32327279 | 9.60 | 33313360 | 9.32 | 34280541 | 9.06 |
| 8 | *A. thaliana* reference genome (TAIR10) | 110973292 | 1.07 | 111601635 | 1.07 | 112114293 | 1.06 | 112556134 | 1.06 |
| 9 | 19 *A. thaliana* strains | 182329469 | 11.49 | 189229970 | 11.07 | 195754679 | 10.69 | 202041197 | 10.35 |
| 10 | RefSeq bacteria genomes | 5639456807 | 1.61 | 5720468920 | 1.58 | 5774723033 | 1.57 | 5820502355 | 1.56 |
| 11 | xenograft model (hg19+MM10) | 4419176798 | 1.25 | 4521884333 | 1.22 | 4601054945 | 1.20 | 4667491155 | 1.18 |
| 12 | human reference genome (hg19) | 2350631887 | 1.22 | 2404675382 | 1.19 | 2448748573 | 1.17 | 2486817647 | 1.15 |

The repetitiveness of the references used in the benchmarking is measured by the copies per distinct *k­*-mers of the references (termed as in the table). We computed the numbers of the distinct *k*-mers of the references at first (shown as “#distinct” in the table), 4 various *k* values (22, 24, 26 and 28) were considered. Then the value can be calculated by , where and are respectively the numbers of all the *k*-mers and the distinct *k*-mers of the reference.

# Supplementary Table 5. Alignment results on simulated datasets from *A. thaliana* genomes

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Aligner** | **Parameter** | **Aligned a** | | **Correctly aligned b** | | **Correctly aligned by primary alignment c** | | **Alignment Time d** | | **Index Loading Time e** | |
| # | % | # | % | # | % |
| ***A. thaliana* Sim-1 (from TAIR10 strain, 100 bp × 100 bp)** | | | | | | | | | |  | |
| 1 | Bowtie2 | very fast | 1998267 | 99.91% | 1976214 | 98.81% | 1976214 | 98.81% | 492 | | 2 | |
| 2 | Bowtie2 | fast | 1998324 | 99.92% | 1976488 | 98.82% | 1976488 | 98.82% | 503 | | 2 | |
| 3 | Bowtie2 | sensitive | 1999793 | 99.99% | 1978080 | 98.90% | 1978080 | 98.90% | 584 | | 2 | |
| 4 | Bowtie2 | very sensitive | 1999970 | 100.00% | 1978328 | 98.92% | 1978328 | 98.92% | 732 | | 2 | |
| 5 | BWA-MEM | default | 2000000 | 100.00% | 1978639 | 98.93% | 1978639 | 98.93% | 281 | | 4 | |
| 6 | BWA-MEM | -r 5 | 2000000 | 100.00% | 1978529 | 98.93% | 1978529 | 98.93% | 232 | | 4 | |
| 7 | BWA-MEM | -r 10 | 2000000 | 100.00% | 1978565 | 98.93% | 1978565 | 98.93% | 225 | | 4 | |
| 8 | BWA-MEM | -r 100 | 2000000 | 100.00% | 1978565 | 98.93% | 1978565 | 98.93% | 226 | | 4 | |
| 9 | SeqAlto | default | 1999994 | 100.00% | 1988874 | 99.44% | 1978217 | 98.91% | 1010 | | 7 | |
| 10 | SeqAlto | fast | 1999992 | 100.00% | 1988797 | 99.44% | 1978072 | 98.90% | 995 | | 7 | |
| 11 | BWA | -n 0.04 | 1992170 | 99.61% | 1983208 | 99.16% | 1970434 | 98.52% | 735 | | 4 | |
| 12 | BWA | -n 0.04 sampe -n 1000 | 1992170 | 99.61% | 1991366 | 99.57% | 1970434 | 98.52% | 786 | | 4 | |
| 13 | BWA | -n 0.06 | 1988925 | 99.45% | 1979916 | 99.00% | 1967239 | 98.36% | 731 | | 4 | |
| 14 | BWA | -n 0.06 sampe -n 1000 | 1988925 | 99.45% | 1988082 | 99.40% | 1967239 | 98.36% | 712 | | 4 | |
| 15 | BWA | -n 0.08 | 1988925 | 99.45% | 1979916 | 99.00% | 1967239 | 98.36% | 709 | | 4 | |
| 16 | BWA | -n 0.08 sampe -n 1000 | 1988925 | 99.45% | 1988082 | 99.40% | 1967239 | 98.36% | 706 | | 4 | |
| 17 | GEM | -m 0.04 -e 0.04 | 1998028 | 99.90% | 1497973 | 74.90% | 1497973 | 74.90% | 182 | | 5 | |
| 18 | GEM | -m 0.06 -e 0.06 | 1999935 | 100.00% | 1499352 | 74.97% | 1499352 | 74.97% | 236 | | 5 | |
| 19 | GEM | -m 0.08 -e 0.08 | 1999998 | 100.00% | 1499505 | 74.98% | 1483487 | 74.17% | 274 | | 5 | |
| 20 | STAR | default | 1899844 | 94.99% | 1980580 | 99.03% | 1980580 | 99.03% | 233 | | 8 | |
| 21 | deBGA | -k 22 -i 5 -n 300 -x 500 -o 300 (default) | 1999966 | 100.00% | 1999133 | 99.96% | 1977399 | 98.87% | 59 | | 1 | |
| 22 | deBGA | -k 22 -i 7 -n 300 -x 500 -o 300 | 1999962 | 100.00% | 1998942 | 99.95% | 1977255 | 98.86% | 61 | | 1 | |
| 23 | deBGA | -k 22 -i 9 -n 300 -x 500 -o 300 | 1999966 | 100.00% | 1998978 | 99.95% | 1977239 | 98.86% | 64 | | 1 | |
| 24 | deBGA | -k 24 -i 5 -n 300 -x 500 -o 300 | 1999969 | 100.00% | 1999007 | 99.95% | 1977298 | 98.86% | 88 | | 1 | |
| 25 | deBGA | -k 26 -i 5 -n 300 -x 500 -o 300 | 1999970 | 100.00% | 1998870 | 99.94% | 1977206 | 98.86% | 54 | | 1 | |
| 26 | deBGA | -k 28 -i 5 -n 300 -x 500 -o 300 | 1999963 | 100.00% | 1998730 | 99.94% | 1977075 | 98.85% | 49 | | 1 | |
| 27 | deBGA | -k 22 -i 5 -n 100 -x 500 -o 300 | 1999965 | 100.00% | 1999112 | 99.96% | 1977384 | 98.87% | 63 | | 1 | |
| 28 | deBGA | -k 22 -i 5 -n 500 -x 500 -o 300 | 1999966 | 100.00% | 1999134 | 99.96% | 1977400 | 98.87% | 66 | | 1 | |
| 29 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 -x 500 -o 300 | 1999999 | 100.00% | 1999786 | 99.99% | 1978087 | 98.90% | 73 | | 1 | |
| 30 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 -x 500 -o 300 | 1999998 | 100.00% | 1999781 | 99.99% | 1978082 | 98.90% | 62 | | 1 | |
| 31 | deBGA | -k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300 | 1999997 | 100.00% | 1999777 | 99.99% | 1978078 | 98.90% | 52 | | 1 | |
| 32 | deBGA | -k 22 -i 5 -n 300 --local -x 500 -o 300 | 1999994 | 100.00% | 1999584 | 99.98% | 1977883 | 98.89% | 70 | | 1 | |
| ***A. thaliana* Sim-19 (from 19 various *A. thaliana* strains, 100 bp × 100 bp)** | | | | | | | | | |  | |
| 33 | Bowtie2 | very fast | 1997919 | 99.90% | 499666 | 24.98% | 499666 | 24.98% | 1487 | | 4 | |
| 34 | Bowtie2 | fast | 1998161 | 99.91% | 507665 | 25.38% | 507665 | 25.38% | 1847 | | 4 | |
| **35** | **Bowtie2** | **sensitive** | **1999744** | **99.99%** | **512541** | **25.63%** | **512541** | **25.63%** | **2429** | | **4** | |
| 36 | Bowtie2 | very sensitive | 1999964 | 100.00% | 514022 | 25.70% | 514022 | 25.70% | 2719 | | 4 | |
| **37** | **BWA-MEM** | **default** | **2000000** | **100.00%** | **518692** | **25.93%** | **518692** | **25.93%** | **1073** | | **5** | |
| 38 | BWA-MEM | -r 5 | 2000000 | 100.00% | 518701 | 25.94% | 518701 | 25.94% | 1450 | | 5 | |
| 39 | BWA-MEM | -r 10 | 2000000 | 100.00% | 519039 | 25.95% | 519039 | 25.95% | 1028 | | 5 | |
| 40 | BWA-MEM | -r 100 | 2000000 | 100.00% | 519039 | 25.95% | 519039 | 25.95% | 975 | | 5 | |
| **41** | **SeqAlto** | **default** | **1999987** | **100.00%** | **1088591** | **54.43%** | **580738** | **29.04%** | **12261** | | **10** | |
| 42 | SeqAlto | fast | 1999251 | 99.96% | 1089028 | 54.45% | 579711 | 28.99% | 11394 | | 10 | |
| 43 | BWA | -n 0.04 | 1909284 | 95.46% | 525528 | 26.28% | 491842 | 24.59% | 1078 | | 5 | |
| **44** | **BWA** | **-n 0.04 sampe -n 1000** | **1909284** | **95.46%** | **1904746** | **95.24%** | **491842** | **24.59%** | **2334** | | **5** | |
| 45 | BWA | -n 0.06 | 1886321 | 94.32% | 518343 | 25.92% | 484782 | 24.24% | 898 | | 5 | |
| 46 | BWA | -n 0.06 sampe -n 1000 | 1886321 | 94.32% | 1882059 | 94.10% | 484782 | 24.24% | 1627 | | 5 | |
| 47 | BWA | -n 0.08 | 1886321 | 94.32% | 518343 | 25.92% | 484782 | 24.24% | 982 | | 5 | |
| 48 | BWA | -n 0.08 sampe -n 1000 | 1886321 | 94.32% | 1882059 | 94.10% | 484782 | 24.24% | 1445 | | 5 | |
| **49** | **GEM** | **-m 0.04 -e 0.04** | **1995845** | **99.79%** | **1989371** | **99.47%** | **511048** | **25.55%** | **1190** | | **8** | |
| 50 | GEM | -m 0.06 -e 0.06 | 1999344 | 99.97% | 1991497 | 99.57% | 510732 | 25.54% | 863 | | 8 | |
| 51 | GEM | -m 0.08 -e 0.08 | 1999694 | 99.98% | 1991524 | 99.58% | 510478 | 25.52% | 2290 | | 8 | |
| **52** | **STAR** | **default** | **970244** | **48.51%** | **701721** | **35.09%** | **441712** | **22.09%** | **542** | | **8** | |
| **53** | **deBGA** | **-k 22 -i 5 -n 300 -x 500 -o 300 (default)** | **1999920** | **100.00%** | **1993278** | **99.66%** | **502975** | **25.15%** | **187** | | **1** | |
| 54 | deBGA | -k 22 -i 7 -n 300 -x 500 -o 300 | 1999939 | 100.00% | 1993244 | 99.66% | 501923 | 25.10% | 174 | | 1 | |
| 55 | deBGA | -k 22 -i 9 -n 300 -x 500 -o 300 | 1999926 | 100.00% | 1993227 | 99.66% | 501800 | 25.09% | 170 | | 1 | |
| 56 | deBGA | -k 24 -i 5 -n 300 -x 500 -o 300 | 1999917 | 100.00% | 1993326 | 99.67% | 503118 | 25.16% | 171 | | 1 | |
| 57 | deBGA | -k 26 -i 5 -n 300 -x 500 -o 300 | 1999895 | 99.99% | 1993092 | 99.65% | 503129 | 25.16% | 187 | | 1 | |
| 58 | deBGA | -k 28 -i 5 -n 300 -x 500 -o 300 | 1999882 | 99.99% | 1992885 | 99.64% | 503114 | 25.16% | 172 | | 1 | |
| 59 | deBGA | -k 22 -i 5 -n 100 -x 500 -o 300 | 1999716 | 99.99% | 1989226 | 99.46% | 502565 | 25.13% | 185 | | 1 | |
| 60 | deBGA | -k 22 -i 5 -n 500 -x 500 -o 300 | 1999943 | 100.00% | 1994360 | 99.72% | 503059 | 25.15% | 195 | | 1 | |
| 61 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 -x 500 -o 300 | 1999944 | 100.00% | 1993749 | 99.69% | 503444 | 25.17% | 166 | | 1 | |
| 62 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 -x 500 -o 300 | 1999946 | 100.00% | 1993606 | 99.68% | 503428 | 25.17% | 167 | | 1 | |
| **63** | **deBGA** | **-k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300** | **1999946** | **100.00%** | **1993560** | **99.68%** | **503419** | **25.17%** | **177** | | **1** | |
| 64 | deBGA | -k 22 -i 5 -n 300 --local -x 500 -o 300 | 1999942 | 100.00% | 1992364 | 99.62% | 502913 | 25.15% | 190 | | 1 | |
| 65 | GenomeMapper f | -M 4 -G 2 -E 4 | 195799 | 97.89% | 194762 | 97.38% | 32878 | 16.44% | 111039 | |  | |

Results on simulated *A.thaliana* datasets (lines 11 and 12 of Supplementary Table 1). a) - e) Various statistics on the sensitivities, correctness and throughputs of the aligners. Refer to Supplementary Notes (the “Statistics for evaluating alignments” subsection) for definition. The highlighted lines are involved in Fig. 2a. f) GenomeMapper is only asked to align one tenth of the dataset, due to its slow speed. As the tool could align a read to multiple positions, but does not distinguish the primary and alternative alignments, we considered the first alignment of each read as its primary alignment.

# Supplementary Table 6. Alignment results on real datasets from *E. coli* genomes

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Aligner** | **Parameter** | **Aligned a** | | **Aligned to the correct strain b** | | **Alignment Time c** | **Index Loading Time d** |
| # | % | # | % |
| **ERR008613 (from K-12 MG1655 substrain, 100 bp × 100 bp)** | | | | | | | | |
| 1 | Bowtie2 | very fast | 1988776 | 99.44% | 130339 | 6.52% | 3630 | 4 |
| 2 | Bowtie2 | fast | 1988844 | 99.44% | 131991 | 6.60% | 4247 | 4 |
| **3** | **Bowtie2** | **sensitive** | **1990186** | **99.51%** | **133062** | **6.65%** | **5153** | **4** |
| 4 | Bowtie2 | very sensitive | 1992370 | 99.62% | 134003 | 6.70% | 5647 | 4 |
| **5** | **BWA-MEM** | **default** | **1989566** | **99.48%** | **1003978** | **50.20%** | **906** | **5** |
| 6 | BWA-MEM | -r 5 | 1989564 | 99.48% | 1003588 | 50.18% | 815 | 5 |
| 7 | BWA-MEM | -r 10 | 1989564 | 99.48% | 1003580 | 50.18% | 804 | 5 |
| 8 | BWA-MEM | -r 100 | 1989564 | 99.48% | 1003580 | 50.18% | 803 | 5 |
| **9** | **SeqAlto** | **default** | **1986260** | **99.31%** | **707818** | **35.39%** | **5500** | **10** |
| 10 | SeqAlto | fast | 1986037 | 99.30% | 706755 | 35.34% | 5036 | 10 |
| 11 | BWA | -n 0.04 | 1829891 | 91.49% | 136945 | 6.85% | 692 | 5 |
| **12** | **BWA** | **-n 0.04 sampe -n 1000** | **1829891** | **91.49%** | **1828551** | **91.43%** | **896** | **5** |
| 13 | BWA | -n 0.06 | 1789104 | 89.46% | 133630 | 6.68% | 546 | 5 |
| 14 | BWA | -n 0.06 sampe -n 1000 | 1789104 | 89.46% | 1787760 | 89.39% | 707 | 5 |
| 15 | BWA | -n 0.08 | 1789104 | 89.46% | 133630 | 6.68% | 546 | 5 |
| 16 | BWA | -n 0.08 sampe -n 1000 | 1789104 | 89.46% | 1787760 | 89.39% | 714 | 5 |
| **17** | **GEM** | **-m 0.04 -e 0.04** | **1960555** | **98.03%** | **1737375** | **86.87%** | **959** | **8** |
| 18 | GEM | -m 0.06 -e 0.06 | 1988968 | 99.45% | 1761356 | 88.07% | 1091 | 8 |
| 19 | GEM | -m 0.08 -e 0.08 | 1997177 | 99.86% | 1766469 | 88.32% | 1229 | 8 |
| **20** | **STAR** | **default** | **564426** | **28.22%** | **521486** | **26.07%** | **804** | **41** |
| **21** | **deBGA** | **-k 22 -i 5 -n 300 -x 500 -o 300 (default)** | **1995720** | **99.79%** | **1964735** | **98.24%** | **250** | **1** |
| 22 | deBGA | -k 22 -i 7 -n 300 -x 500 -o 300 | 1995616 | 99.78% | 1958105 | 97.91% | 217 | 1 |
| 23 | deBGA | -k 22 -i 9 -n 300 -x 500 -o 300 | 1995615 | 99.78% | 1970803 | 98.54% | 210 | 1 |
| 24 | deBGA | -k 24 -i 5 -n 300 -x 500 -o 300 | 1995496 | 99.77% | 1964157 | 98.21% | 211 | 1 |
| 25 | deBGA | -k 26 -i 5 -n 300 -x 500 -o 300 | 1995140 | 99.76% | 1964936 | 98.25% | 214 | 1 |
| 26 | deBGA | -k 28 -i 5 -n 300 -x 500 -o 300 | 1994758 | 99.74% | 1963910 | 98.20% | 212 | 1 |
| 27 | deBGA | -k 22 -i 5 -n 100 -x 500 -o 300 | 1995587 | 99.78% | 1962189 | 98.11% | 225 | 1 |
| 28 | deBGA | -k 22 -i 5 -n 500 -x 500 -o 300 | 1995711 | 99.79% | 1964687 | 98.23% | 223 | 1 |
| 29 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 -x 500 -o 300 | 1997305 | 99.87% | 1971715 | 98.59% | 187 | 1 |
| 30 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 -x 500 -o 300 | 1997401 | 99.87% | 1971206 | 98.56% | 190 | 1 |
| **31** | **deBGA** | **-k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300** | **1997421** | **99.87%** | **1970886** | **98.54%** | **185** | **1** |
| 32 | deBGA | -k 22 -i 5 -n 300 --local -x 500 -o 300 | 1997394 | 99.87% | 1971314 | 98.57% | 181 | 1 |
| **SRR522163 (from K-12 MG1655 substrain, 251 bp × 251 bp)** | | | | | | | | |
| 33 | Bowtie2 | very fast | 1707304 | 85.37% | 45169 | 2.26% | 4369 | 4 |
| 34 | Bowtie2 | fast | 1707826 | 85.39% | 50662 | 2.53% | 6069 | 4 |
| **35** | **Bowtie2** | **sensitive** | **1714352** | **85.72%** | **58108** | **2.91%** | **8722** | **4** |
| 36 | Bowtie2 | very sensitive | 1725592 | 86.28% | 60041 | 3.00% | 10731 | 4 |
| **37** | **BWA-MEM** | **default** | **1825795** | **91.29%** | **912218** | **45.61%** | **5264** | **5** |
| 38 | BWA-MEM | -r 5 | 1825793 | 91.29% | 912124 | 45.61% | 5223 | 5 |
| 39 | BWA-MEM | -r 10 | 1825793 | 91.29% | 911888 | 45.59% | 4855 | 5 |
| 40 | BWA-MEM | -r 100 | 1825793 | 91.29% | 911752 | 45.59% | 4688 | 5 |
| **41** | **SeqAlto** | **default** | **1610365** | **80.52%** | **988713** | **49.44%** | **97725** | **10** |
| 42 | SeqAlto | fast | 1609571 | 80.48% | 987284 | 49.36% | 91859 | 10 |
| 43 | BWA | -n 0.04 | 763422 | 38.17% | 70405 | 3.52% | 3014 | 5 |
| **44** | **BWA** | **-n 0.04 sampe -n 1000** | **763422** | **38.17%** | **761558** | **38.08%** | **3079** | **5** |
| 45 | BWA | -n 0.06 | 763422 | 38.17% | 70405 | 3.52% | 2966 | 5 |
| 46 | BWA | -n 0.06 sampe -n 1000 | 763422 | 38.17% | 761558 | 38.08% | 3042 | 5 |
| 47 | BWA | -n 0.08 | 694758 | 34.74% | 63888 | 3.19% | 2252 | 5 |
| 48 | BWA | -n 0.08 sampe -n 1000 | 694758 | 34.74% | 693009 | 34.65% | 2329 | 5 |
| **49** | **GEM** | **-m 0.04 -e 0.04** | **1643549** | **82.18%** | **1339094** | **66.95%** | **2176** | **8** |
| 50 | GEM | -m 0.06 -e 0.06 | 1725514 | 86.28% | 1400507 | 70.03% | 4342 | 8 |
| 51 | GEM | -m 0.08 -e 0.08 | 1843666 | 92.18% | 1436361 | 71.82% | 17669 | 8 |
| **52** | **STAR** | **default** | **985454** | **49.27%** | **490512** | **24.53%** | **42447** | **40** |
| **53** | **deBGA** | **-k 22 -i 5 -n 300 -x 500 -o 300 (default)** | **1818555** | **90.93%** | **1677640** | **85.53%** | **918** | **1** |
| 54 | deBGA | -k 22 -i 7 -n 300 -x 500 -o 300 | 1815225 | 90.76% | 1687671 | 86.05% | 840 | 1 |
| 55 | deBGA | -k 22 -i 9 -n 300 -x 500 -o 300 | 1817412 | 90.87% | 1694911 | 86.39% | 803 | 1 |
| 56 | deBGA | -k 24 -i 5 -n 300 -x 500 -o 300 | 1814078 | 90.70% | 1683701 | 85.84% | 830 | 1 |
| 57 | deBGA | -k 26 -i 5 -n 300 -x 500 -o 300 | 1810437 | 90.52% | 1685685 | 85.97% | 814 | 1 |
| 58 | deBGA | -k 28 -i 5 -n 300 -x 500 -o 300 | 1809480 | 90.47% | 1694844 | 86.45% | 798 | 1 |
| 59 | deBGA | -k 22 -i 5 -n 100 -x 500 -o 300 | 1818063 | 90.90% | 1674366 | 85.34% | 934 | 1 |
| 60 | deBGA | -k 22 -i 5 -n 500 -x 500 -o 300 | 1818522 | 90.93% | 1677626 | 85.52% | 928 | 1 |
| 61 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 -x 500 -o 300 | 1832266 | 91.61% | 1754076 | 88.34% | 534 | 1 |
| 62 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 -x 500 -o 300 | 1834676 | 91.73% | 1756266 | 88.45% | 490 | 1 |
| **63** | **deBGA** | **-k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300** | **1835078** | **91.75%** | **1766610** | **88.33%** | **460** | **1** |
| 64 | deBGA | -k 22 -i 5 -n 300 --local -x 500 -o 300 | 1835037 | 91.75% | 1726351 | 86.71% | 937 | 1 |
| **SRR530851 (from *E. coli* O157:H7 strain, 151 bp × 151 bp)** | | | | | | | | |
| 65 | Bowtie2 | very fast | 1898341 | 94.92% | 291297 | 14.56% | 3641 | 4 |
| 66 | Bowtie2 | fast | 1898986 | 94.95% | 293981 | 14.70% | 4118 | 4 |
| **67** | **Bowtie2** | **sensitive** | **1900029** | **95.00%** | **296637** | **14.83%** | **4824** | **4** |
| 68 | Bowtie2 | very sensitive | 1900641 | 95.03% | 296134 | 14.81% | 4922 | 4 |
| **69** | **BWA-MEM** | **default** | **1961316** | **98.07%** | **300904** | **15.05%** | **3296** | **5** |
| 70 | BWA-MEM | -r 5 | 1961316 | 98.07% | 300745 | 15.04% | 3042 | 5 |
| 71 | BWA-MEM | -r 10 | 1961315 | 98.07% | 300767 | 15.04% | 3379 | 5 |
| 72 | BWA-MEM | -r 100 | 1961315 | 98.07% | 300767 | 15.04% | 2416 | 5 |
| **73** | **SeqAlto** | **default** | **1930794** | **96.54%** | **1793547** | **89.68%** | **10370** | **10** |
| 74 | SeqAlto | fast | 1930520 | 96.53% | 1792000 | 89.60% | 10244 | 10 |
| 75 | BWA | -n 0.04 | 1793054 | 89.65% | 303579 | 15.18% | 885 | 5 |
| **76** | **BWA** | **-n 0.04 sampe -n 1000** | **1793054** | **89.65%** | **1787101** | **89.36%** | **1078** | **5** |
| 77 | BWA | -n 0.06 | 1793054 | 89.65% | 303579 | 15.18% | 727 | 5 |
| 78 | BWA | -n 0.06 sampe -n 1000 | 1793054 | 89.65% | 1787101 | 89.36% | 828 | 5 |
| 79 | BWA | -n 0.08 | 1793054 | 89.65% | 303579 | 15.18% | 754 | 5 |
| 80 | BWA | -n 0.08 sampe -n 1000 | 1793054 | 89.65% | 1787101 | 89.36% | 898 | 5 |
| **81** | **GEM** | **-m 0.04 -e 0.04** | **1854491** | **92.72%** | **1842169** | **92.11%** | **1104** | **8** |
| 82 | GEM | -m 0.06 -e 0.06 | 1875920 | 93.80% | 1859937 | 93.00% | 1042 | 8 |
| 83 | GEM | -m 0.08 -e 0.08 | 1893057 | 94.65% | 1872163 | 93.61% | 2114 | 8 |
| **84** | **STAR** | **default** | **1510118** | **75.51%** | **1476730** | **73.84%** | **1129** | **40** |
| **85** | **deBGA** | **-k 22 -i 5 -n 300 -x 500 -o 300 (default)** | **1954640** | **97.73%** | **1895888** | **94.79%** | **345** | **1** |
| 86 | deBGA | -k 22 -i 7 -n 300 -x 500 -o 300 | 1954595 | 97.73% | 1895363 | 94.77% | 327 | 1 |
| 87 | deBGA | -k 22 -i 9 -n 300 -x 500 -o 300 | 1954531 | 97.73% | 1893846 | 94.69% | 319 | 1 |
| 88 | deBGA | -k 24 -i 5 -n 300 -x 500 -o 300 | 1954420 | 97.72% | 1896499 | 94.82% | 372 | 1 |
| 89 | deBGA | -k 26 -i 5 -n 300 -x 500 -o 300 | 1954189 | 97.71% | 1896969 | 94.85% | 334 | 1 |
| 90 | deBGA | -k 28 -i 5 -n 300 -x 500 -o 300 | 1954046 | 97.70% | 1898029 | 94.90% | 321 | 1 |
| 91 | deBGA | -k 22 -i 5 -n 100 -x 500 -o 300 | 1954187 | 97.71% | 1894772 | 94.74% | 329 | 1 |
| 92 | deBGA | -k 22 -i 5 -n 500 -x 500 -o 300 | 1954625 | 97.73% | 1895897 | 94.79% | 340 | 1 |
| 93 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 -x 500 -o 300 | 1955745 | 97.79% | 1914479 | 95.72% | 335 | 1 |
| 94 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 -x 500 -o 300 | 1956995 | 97.85% | 1924089 | 96.20% | 335 | 1 |
| **95** | **deBGA** | **-k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300** | **1959854** | **97.99%** | **1910418** | **95.52%** | **337** | **1** |
| 96 | deBGA | -k 22 -i 5 -n 300 --local -x 500 -o 300 | 1960702 | 98.04% | 1887371 | 94.37% | 552 | 1 |

Results on real *E.coli* datasets (lines 8-10 of Supplementary Table 1). a) - d) Various statistics on the sensitivities, correctness and throughputs of the aligners. Refer to Supplementary Notes (the “Statistics for evaluating alignments” subsection) for definition. The highlighted lines are involved in Fig. 2b.

# Supplementary Table 7. The numbers of real reads aligned to the 62 *E.coli* strains by deBGA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | ***E. Coli* Strains**  **Name (Accession Number)** | **ERR008613 a** | **SRR522163 b** | **SRR530851 c** |
| 1 | K\_12\_substr\_MG1655 (NC\_000913) | 1970886 | 1766610 | 188574 |
| 2 | O157\_H7\_EDL933 (NC\_002655) | 310325 | 111471 | 1778381 |
| 3 | O157\_H7 (NC\_002695) | 320663 | 116311 | 1910418 |
| 4 | CFT073 (NC\_004431) | 129747 | 46820 | 64163 |
| 5 | K\_12\_substr\_\_W3110 (NC\_007779) | 1965859 | 1761581 | 188435 |
| 6 | UTI89 (NC\_007946) | 117742 | 34049 | 59745 |
| 7 | 536 (NC\_008253) | 121070 | 36870 | 60302 |
| 8 | APEC\_O1 (NC\_008563) | 123328 | 39123 | 63485 |
| 9 | HS (NC\_009800) | 885775 | 604275 | 175724 |
| 10 | E24377A (NC\_009801) | 567286 | 289606 | 178727 |
| 11 | ATCC\_8739 (NC\_010468) | 1005149 | 721816 | 188263 |
| 12 | K\_12\_substr\_\_DH10B (NC\_010473) | 1904247 | 1697349 | 185286 |
| 13 | SMS\_3\_5 (NC\_010498) | 170857 | 58941 | 87438 |
| 14 | O157\_H7\_EC4115 (NC\_011353) | 318865 | 116182 | 1781127 |
| 15 | SE11 (NC\_011415) | 577874 | 292062 | 174718 |
| 16 | O127\_H6\_E2348\_69 (NC\_011601) | 118558 | 35368 | 76775 |
| 17 | IAI1 (NC\_011741) | 568882 | 292503 | 175707 |
| 18 | S88 (NC\_011742) | 144872 | 54576 | 66762 |
| 19 | ED1a (NC\_011745) | 134952 | 49794 | 65218 |
| 20 | 55989 (NC\_011748) | 561447 | 278467 | 180763 |
| 21 | IAI39 (NC\_011750) | 179750 | 69397 | 85464 |
| 22 | UMN026 (NC\_011751) | 230699 | 101074 | 97436 |
| 23 | LF82 (NC\_011993) | 122471 | 39430 | 61221 |
| 24 | BW2952 (NC\_012759) | 1906036 | 1698064 | 185485 |
| 25 | BL21\_DE3 (NC\_012892) | 1016972 | 731741 | 181091 |
| 26 | \_BL21\_Gold\_DE3\_pLysS\_AG (NC\_012947) | 1080849 | 797633 | 185677 |
| 27 | B\_REL606 (NC\_012967) | 1032203 | 743548 | 181335 |
| 28 | BL21\_DE3 (NC\_012971) | 1017485 | 732630 | 181260 |
| 29 | O157\_H7\_TW14359 (NC\_013008) | 318748 | 115696 | 1780606 |
| 30 | O103\_H2\_12009 (NC\_013353) | 558737 | 279265 | 224549 |
| 31 | O26\_H11\_11368 (NC\_013361) | 566312 | 288774 | 248577 |
| 32 | O111\_H\_\_11128 (NC\_013364) | 545025 | 270275 | 258727 |
| 33 | SE15 (NC\_013654) | 128610 | 43046 | 58115 |
| 34 | O55\_H7\_CB9615 (NC\_013941) | 331037 | 124650 | 1379143 |
| 35 | KO11FL (NC\_016902) | 628026 | 353341 | 182098 |
| 36 | DH1 (NC\_017625) | 1944823 | 1729935 | 188664 |
| 37 | 042 (NC\_017626) | 215235 | 89329 | 100390 |
| 38 | IHE3034 (NC\_017628) | 120964 | 38408 | 61872 |
| 39 | ABU\_83972 (NC\_017631) | 124606 | 38633 | 60477 |
| 40 | UM146 (NC\_017632) | 119248 | 37690 | 60575 |
| 41 | ETEC\_H10407 (NC\_017633) | 1339567 | 1052962 | 178863 |
| 42 | O83\_H1\_NRG\_857C (NC\_017634) | 123071 | 38562 | 60691 |
| 43 | W\_uid162011 (NC\_017635) | 580237 | 294277 | 176324 |
| 44 | DH1 (NC\_017638) | 1933478 | 1711695 | 184379 |
| 45 | NA114 (NC\_017644) | 128446 | 43292 | 71679 |
| 46 | UMNK88 (NC\_017645) | 1119683 | 824555 | 208309 |
| 47 | O7\_K1\_CE10 (NC\_017646) | 182836 | 69353 | 87742 |
| 48 | \_clone\_D\_i2 (NC\_017651) | 119510 | 35796 | 60109 |
| 49 | \_clone\_D\_i14 (NC\_017652) | 121740 | 38799 | 61243 |
| 50 | O55\_H7\_RM12579 (NC\_017656) | 328913 | 123510 | 1370996 |
| 51 | KO11FL (NC\_017660) | 615447 | 344123 | 177347 |
| 52 | P12b (NC\_017663) | 1284626 | 1007660 | 178117 |
| 53 | W\_uid162101 (NC\_017664) | 578904 | 293107 | 174568 |
| 54 | Xuzhou21 (NC\_017906) | 318998 | 114561 | 1864305 |
| 55 | O104\_H4\_2009EL\_2050 (NC\_018650) | 569946 | 296044 | 191649 |
| 56 | O104\_H4\_2011C\_3493 (NC\_018658) | 572433 | 298163 | 193920 |
| 57 | O104\_H4\_2009EL\_2071 (NC\_018661) | 572135 | 296573 | 193852 |
| 58 | APEC\_O78 (NC\_020163) | 617502 | 324739 | 171414 |
| 59 | K\_12\_substr\_\_MDS42 (NC\_020518) | 1678784 | 1503526 | 174670 |
| 60 | LY180 (NC\_022364) | 621232 | 344055 | 174572 |
| 61 | PMV\_1 (NC\_022370) | 117156 | 34317 | 59428 |
| 62 | JJ1886 (NC\_022648) | 129572 | 43409 | 69268 |

a) - c) The number of the reads of ERR008613 (a), SRR522163 (b) and SRR530851 (c) aligned to the 62 *E.coli* strains by deBGA. DeBGA is run in the *--cl* 0.35 configuration (-k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300). It is also worthnoting that, for each read, all its alignments output by deBGA are taken into account in this result, i.e., a read may be aligned to multiple genomes, so that the total number of the reads aligned to the genomes is higher than that of the original number of reads.

# Supplementary Table 8. Alignment results on simulated datasets from RefSeq bacteria genomes

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Aligner** | **Parameter** | **Aligned a** | | **Correctly aligned b** | | **Correctly aligned by primary alignment c** | | **Alignment Time d** | | **Index Loading Time e** | |
| # | % | # | % | # | % |
| ***Meta Sim* (from NCBI RefSeq bacteria genomes, 100 bp × 100 bp)** | | | | | | | | | |  | |
| **1** | **BWA-MEM** | **default** | **2000000** | **100.00%** | **1449547** | **72.48%** | **1449547** | **72.48%** | **713** | | **15** | |
| 2 | BWA-MEM | -r 5 | 2000000 | 100.00% | 1449438 | 72.47% | 1449438 | 72.47% | 1894 | | 15 | |
| 3 | BWA-MEM | -r 10 | 2000000 | 100.00% | 1449381 | 72.47% | 1449381 | 72.47% | 793 | | 15 | |
| 4 | BWA-MEM | -r 100 | 2000000 | 100.00% | 1449381 | 72.47% | 1449381 | 72.47% | 563 | | 15 | |
| **5** | **BWA** | **-n 0.04 sampe -n 1000** | **1955781** | **97.79%** | **1954552** | **97.73%** | **1426936** | **71.35%** | **1773** | | **15** | |
| 6 | BWA | -n 0.06 sampe -n 1000 | 1944090 | 97.20% | 1942869 | 97.14% | 1421158 | 71.06% | 1571 | | 15 | |
| 7 | BWA | -n 0.08 sampe -n 1000 | 1944090 | 97.20% | 1942869 | 97.14% | 1421158 | 71.06% | 1529 | | 15 | |
| **8** | **GEM** | **-m 0.04 -e 0.04** | **1993838** | **99.69%** | **1985391** | **99.27%** | **1443091** | **72.15%** | **1435** | | **15** | |
| 9 | GEM | -m 0.06 -e 0.06 | 1999206 | 99.96% | 1988735 | 99.44% | 1444716 | 72.24% | 1941 | | 15 | |
| 10 | GEM | -m 0.08 -e 0.08 | 1999753 | 99.99% | 1988597 | 99.43% | 1444701 | 72.24% | 7073 | | 15 | |
| **11** | **STAR** | **default** | **1801334** | **90.07%** | **1792692** | **89.63%** | **1422569** | **71.13%** | **480** | | **79** | |
| **12** | **deBGA** | **-k 22 -i 5 -n 300 -x 500 -o 300 (default)** | **1999872** | **99.99%** | **1994005** | **99.70%** | **1443974** | **72.20%** | **130** | | **74** | |
| 13 | deBGA | -k 22 -i 7 -n 300 -x 500 -o 300 | 1999873 | 99.99% | 1991647 | 99.58% | 1442935 | 72.15% | 95 | | 74 | |
| 14 | deBGA | -k 22 -i 9 -n 300 -x 500 -o 300 | 1999871 | 99.99% | 1992221 | 99.61% | 1443031 | 72.15% | 102 | | 74 | |
| 15 | deBGA | -k 22 -i 5 -n 100 -x 500 -o 300 | 1999735 | 99.99% | 1992547 | 99.63% | 1443513 | 72.18% | 101 | | 74 | |
| 16 | deBGA | -k 22 -i 5 -n 500 -x 500 -o 300 | 1999954 | 100.00% | 1994872 | 99.74% | 1444172 | 72.21% | 100 | | 74 | |
| 17 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 -x 500 -o 300 | 1999900 | 100.00% | 1995446 | 99.77% | 1445372 | 72.27% | 106 | | 74 | |
| 18 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 -x 500 -o 300 | 1999901 | 100.00% | 1995438 | 99.77% | 1445368 | 72.27% | 101 | | 74 | |
| **19** | **deBGA** | **-k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300** | **1999901** | **100.00%** | **1995424** | **99.77%** | **1445366** | **72.27%** | **100** | | **74** | |
| 20 | deBGA | -k 22 -i 5 -n 300 --local -x 500 -o 300 | 1999898 | 99.99% | 1994599 | 99.73% | 1444960 | 72.25% | 101 | | 74 | |

Results on simulated metagenomics dataset (line 13 of Supplementary Table 1). a) - e) Various statistics on the sensitivities, correctness and throughputs of the aligners. Refer to Supplementary Notes (the “Statistics for evaluating alignments” subsection) for definition. The highlighted lines are involved in Fig. 3a.

# Supplementary Table 9. Alignment results on simulated datasets from Xenograft model

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Aligner** | **Parameter** | **Aligned a** | | **Correctly aligned b** | | **Correctly aligned by primary alignment c** | | **Alignment Time d** | | **Index Loading Time e** | |
| # | % | # | % | # | % |
| ***Xeon Sim* (from human reference genome hg19/GCRh37 and mouse reference genome mm10, 100 bp × 100 bp)** | | | | | | | | | |  | |
| **1** | **BWA-MEM** | **default** | **2000000** | **100.00%** | **1938905** | **96.95%** | **1938905** | **96.95%** | **2000** | | **16** | |
| 2 | BWA-MEM | -r 5 | 2000000 | 100.00% | 1938645 | 96.93% | 1938645 | 96.93% | 1215 | | 16 | |
| 3 | BWA-MEM | -r 10 | 2000000 | 100.00% | 1938681 | 96.93% | 1938681 | 96.93% | 1199 | | 16 | |
| 4 | BWA-MEM | -r 100 | 2000000 | 100.00% | 1938681 | 96.93% | 1938681 | 96.93% | 1222 | | 16 | |
| **5** | **BWA** | **-n 0.04 sampe -n 1000** | **1990002** | **99.50%** | **1974824** | **98.74%** | **1926671** | **96.33%** | **2485** | | **16** | |
| 6 | BWA | -n 0.06 sampe -n 1000 | 1985915 | 99.30% | 1970979 | 98.55% | 1922959 | 96.15% | 2044 | | 16 | |
| 7 | BWA | -n 0.08 sampe -n 1000 | 1985915 | 99.30% | 1935021 | 96.75% | 1922959 | 96.15% | 1974 | | 16 | |
| **8** | **GEM** | **-m 0.04 -e 0.04** | **1996622** | **99.83%** | **1974179** | **98.71%** | **1923840** | **96.19%** | **492** | | **47** | |
| 9 | GEM | -m 0.06 -e 0.06 | 1999625 | 99.98% | 1976076 | 98.80% | 1925944 | 96.30% | 1434 | | 47 | |
| 10 | GEM | -m 0.08 -e 0.08 | 1999880 | 99.99% | 1975968 | 98.80% | 1925886 | 96.29% | 11942 | | 47 | |
| **11** | **STAR** | **default** | **1973872** | **98.69%** | **1949673** | **97.48%** | **1908506** | **95.43%** | **426** | | **58** | |
| **12** | **deBGA** | **-k 22 -i 5 -n 300 -x 500 -o 300 (default)** | **1999296** | **99.96%** | **1975681** | **98.78%** | **1924334** | **96.22%** | **260** | | **32** | |
| 13 | deBGA | -k 22 -i 7 -n 300 -x 500 -o 300 | 1999171 | 99.96% | 1971463 | 98.57% | 1920521 | 96.03% | 250 | | 32 | |
| 14 | deBGA | -k 22 -i 9 -n 300 -x 500 -o 300 | 1999068 | 99.95% | 1970909 | 98.55% | 1919458 | 95.97% | 255 | | 32 | |
| 15 | deBGA | -k 22 -i 5 -n 100 -x 500 -o 300 | 1999188 | 99.96% | 1969022 | 98.45% | 1920774 | 96.04% | 289 | | 32 | |
| 16 | deBGA | -k 22 -i 5 -n 500 -x 500 -o 300 | 1999344 | 99.97% | 1976345 | 98.82% | 1925237 | 96.26% | 255 | | 32 | |
| 17 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 -x 500 -o 300 | 1999150 | 99.96% | 1975379 | 98.77% | 1924377 | 96.22% | 264 | | 32 | |
| 18 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 -x 500 -o 300 | 1998946 | 99.95% | 1974774 | 98.74% | 1923871 | 96.19% | 256 | | 32 | |
| **19** | **deBGA** | **-k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300** | **1998810** | **99.94%** | **1974383** | **98.72%** | **1923580** | **96.18%** | **253** | | **32** | |
| 20 | deBGA | -k 22 -i 5 -n 300 --local -x 500 -o 300 | 1998727 | 99.94% | 1973895 | 98.69% | 1923169 | 96.16% | 255 | | 32 | |

Results on simulated xenograft model dataset (line 14 of Supplementary Table 1). a) - e) Various statistics on the sensitivities, correctness and throughputs of the aligners. Refer to Supplementary Notes (the “Statistics for evaluating alignments” subsection) for definition. The highlighted lines are involved in Fig. 3b.

# Supplementary Table 10. Alignment results on pseudo-real metagenomic datasets

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Aligner** | **Parameter** | **Aligned a** | | **Aligned to the correct genome b** | | **Alignment Time c** | **Index Loading Time d** |
| # | % | # | % |
| **6 real sequencing datasets from 6 various bacteria genomes** | | | | | | | | |
| **1** | **BWA-MEM** | **default** | **1883121** | **62.77%** | **934191** | **31.14%** | **3160** | **15** |
| 2 | BWA-MEM | -r 5 | 1894452 | 63.15% | 935155 | 31.17% | 2916 | 15 |
| 3 | BWA-MEM | -r 10 | 1894462 | 63.15% | 935458 | 31.18% | 2986 | 15 |
| 4 | BWA-MEM | -r 100 | 1894462 | 63.15% | 935458 | 31.18% | 2787 | 15 |
| **5** | **BWA** | **-n 0.04 sampe -n 1000** | **2116793** | **70.56%** | **2076088** | **69.20%** | **3742** | **15** |
| 6 | BWA | -n 0.06 sampe -n 1000 | 2108709 | 70.29% | 2069240 | 68.97% | 3710 | 15 |
| 7 | BWA | -n 0.08 sampe -n 1000 | 2090079 | 69.67% | 2050665 | 68.36% | 3412 | 15 |
| **8** | **GEM** | **-m 0.04 -e 0.04** | **2580949** | **86.03%** | **2486046** | **82.87%** | **1504** | **15** |
| 9 | GEM | -m 0.06 -e 0.06 | 2623732 | 87.46% | 2520474 | 84.02% | 3176 | 15 |
| 10 | GEM | -m 0.08 -e 0.08 | 2645228 | 88.17% | 2535427 | 84.51% | 50271 | 15 |
| **11** | **STAR** | **default** | **1666100** | **55.54%** | **1471130** | **49.04%** | **680** | **79** |
| **12** | **deBGA** | **-k 22 -i 5 -n 300 -x 500 -o 300 (default)** | **2786041** | **92.87%** | **2588073** | **86.27%** | **582** | **74** |
| 13 | deBGA | -k 22 -i 7 -n 300 -x 500 -o 300 | 2785966 | 92.87% | 2593247 | 86.44% | 539 | 74 |
| 14 | deBGA | -k 22 -i 9 -n 300 -x 500 -o 300 | 2787245 | 92.91% | 2603497 | 86.78% | 510 | 74 |
| 15 | deBGA | -k 22 -i 5 -n 100 -x 500 -o 300 | 2785785 | 92.86% | 2588972 | 86.30% | 584 | 74 |
| 16 | deBGA | -k 22 -i 5 -n 500 -x 500 -o 300 | 2786131 | 92.87% | 2588897 | 86.30% | 586 | 74 |
| 17 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 -x 500 -o 300 | 2798218 | 93.27% | 2666712 | 88.89% | 434 | 74 |
| 18 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 -x 500 -o 300 | 2812655 | 93.76% | 2694544 | 89.82% | 388 | 74 |
| **19** | **deBGA** | **-k 22 -i 5 -n 300 --cl 0.35 -x 500 -o 300** | **2818673** | **93.96%** | **2701294** | **90.04%** | **378** | **74** |
| 20 | deBGA | -k 22 -i 5 -n 300 --local -x 500 -o 300 | 2822201 | 94.07% | 2651921 | 88.40% | 610 | 74 |

Results on “pseudo real” metagenomics dataset (line 15 of Supplementary Table 1). a) - d) Various statistics on the sensitivities, correctness and throughputs of the aligners. Refer to Supplementary Notes (the “Statistics for evaluating alignments” subsection) for definition. The highlighted lines are involved in Fig. 3c.

# Supplementary Table 11. Alignment results on simulated datasets from human genome

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Aligner** | **Parameter** | **Aligned a** | | **Correctly aligned b** | | **Correctly aligned by primary alignment c** | | **Alignment Time d** | | **Index loading time e** | |
| # | % | # | % | # | % |
| ***H. sapiens* Sim-100 (100 bp × 100 bp)** | | | | | | | | | |  | |
| 1 | Bowtie2 | very fast | 1996839 | 99.84% | 1968203 | 98.41% | 1968203 | 98.41% | 612 | | 29 | |
| 2 | Bowtie2 | fast | 1997312 | 99.87% | 1969568 | 98.48% | 1969568 | 98.48% | 668 | | 29 | |
| **3** | **Bowtie2** | **sensitive** | **1999552** | **99.98%** | **1973715** | **98.69%** | **1973715** | **98.69%** | **876** | | **29** | |
| 4 | Bowtie2 | very sensitive | 1999890 | 99.99% | 1975106 | 98.76% | 1975106 | 98.76% | 1698 | | 29 | |
| **5** | **BWA-MEM** | **default** | **2000000** | **100.00%** | **1976623** | **98.83%** | **1976623** | **98.83%** | **722** | | **6** | |
| 6 | BWA-MEM | -r 5 | 2000000 | 100.00% | 1976347 | 98.82% | 1976347 | 98.82% | 445 | | 6 | |
| 7 | BWA-MEM | -r 10 | 2000000 | 100.00% | 1976338 | 98.82% | 1976338 | 98.82% | 435 | | 6 | |
| 8 | BWA-MEM | -r 100 | 2000000 | 100.00% | 1976338 | 98.82% | 1976338 | 98.82% | 450 | | 6 | |
| **9** | **SeqAlto** | **default** | **1999979** | **100.00%** | **1985650** | **99.28%** | **1974919** | **98.75%** | **2567** | | **137** | |
| 10 | SeqAlto | fast | 1999399 | 99.97% | 1984419 | 99.22% | 1973429 | 98.67% | 1358 | | 137 | |
| **11** | **BWA** | **-n 0.04** | **1992278** | **99.61%** | **1980551** | **99.03%** | **1966892** | **98.34%** | **2602** | | **6** | |
| 12 | BWA | -n 0.06 | 1988937 | 99.45% | 1977284 | 98.86% | 1963737 | 98.19% | 1974 | | 6 | |
| 13 | BWA | -n 0.08 | 1988937 | 99.45% | 1977284 | 98.86% | 1963737 | 98.19% | 2111 | | 6 | |
| **14** | **GEM** | **-m 0.04 -e 0.04** | **1997771** | **99.89%** | **1991318** | **99.57%** | **1969103** | **98.46%** | **351** | | **37** | |
| 15 | GEM | -m 0.06 -e 0.06 | 1999909 | 100.00% | 1993323 | 99.67% | 1971177 | 98.56% | 619 | | 37 | |
| 16 | GEM | -m 0.08 -e 0.08 | 1999989 | 100.00% | 1993355 | 99.67% | 1971159 | 98.56% | 1509 | | 37 | |
| **17** | **STAR** | **default** | **1994756** | **99.74%** | **1967997** | **98.40%** | **1951298** | **97.56%** | **297** | | **43** | |
| **18** | **deBGA** | **-k 22 -i 5 -n 300 (default)** | **1999645** | **99.98%** | **1993322** | **99.67%** | **1970877** | **98.54%** | **171** | | **22** | |
| 19 | deBGA | -k 22 -i 7 -n 300 | 1999609 | 99.98% | 1990350 | 99.52% | 1968192 | 98.41% | 172 | | 22 | |
| 20 | deBGA | -k 22 -i 9 -n 300 | 1999529 | 99.98% | 1990619 | 99.53% | 1968035 | 98.40% | 182 | | 22 | |
| 21 | deBGA | -k 24 -i 5 -n 300 | 1999651 | 99.98% | 1993388 | 99.67% | 1971016 | 98.55% | 165 | | 22 | |
| 22 | deBGA | -k 26 -i 5 -n 300 | 1999589 | 99.98% | 1993028 | 99.65% | 1970627 | 98.53% | 140 | | 22 | |
| 23 | deBGA | -k 28 -i 5 -n 300 | 1999537 | 99.98% | 1992679 | 99.63% | 1970409 | 98.52% | 209 | | 22 | |
| 24 | deBGA | -k 22 -i 5 -n 100 | 1999585 | 99.98% | 1993505 | 99.68% | 1970749 | 98.54% | 191 | | 22 | |
| 25 | deBGA | -k 22 -i 5 -n 500 | 1999677 | 99.98% | 1993087 | 99.65% | 1970969 | 98.55% | 168 | | 22 | |
| 26 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 | 1999519 | 99.98% | 1992800 | 99.64% | 1970529 | 98.53% | 175 | | 22 | |
| 27 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 | 1999366 | 99.97% | 1992257 | 99.61% | 1970029 | 98.50% | 171 | | 22 | |
| **28** | **deBGA** | **-k 22 -i 5 -n 300 --cl 0.35** | **1999217** | **99.96%** | **1991881** | **99.59%** | **1969681** | **98.48%** | **165** | | **22** | |
| 29 | deBGA | -k 22 -i 5 -n 300 --local | 1999137 | 99.96% | 1991340 | 99.57% | 1969182 | 98.46% | 170 | | 22 | |
| ***H. sapiens* Sim-125 (125 bp × 125 bp)** | | | | | | | | | | |  | |
| 30 | Bowtie2 | very fast | 1998266 | 99.91% | 1973332 | 98.67% | 1973332 | 98.67% | 750 | | 29 | |
| 31 | Bowtie2 | fast | 1998745 | 99.94% | 1974485 | 98.72% | 1974485 | 98.72% | 819 | | 29 | |
| 32 | Bowtie2 | sensitive | 1999731 | 99.99% | 1977125 | 98.86% | 1977125 | 98.86% | 1039 | | 29 | |
| 33 | Bowtie2 | very sensitive | 1999948 | 100.00% | 1978387 | 98.92% | 1978387 | 98.92% | 2125 | | 29 | |
| 34 | BWA-MEM | default | 2000000 | 100.00% | 1979841 | 98.99% | 1979841 | 98.99% | 781 | | 6 | |
| 35 | BWA-MEM | -r 5 | 2000000 | 100.00% | 1979780 | 98.99% | 1979780 | 98.99% | 533 | | 6 | |
| 36 | BWA-MEM | -r 10 | 2000000 | 100.00% | 1979674 | 98.98% | 1979674 | 98.98% | 471 | | 6 | |
| 37 | BWA-MEM | -r 100 | 2000000 | 100.00% | 1979674 | 98.98% | 1979674 | 98.98% | 493 | | 6 | |
| 38 | SeqAlto | default | 1999983 | 100.00% | 1988035 | 99.40% | 1978484 | 98.92% | 4096 | | 137 | |
| 39 | SeqAlto | fast | 1999750 | 99.99% | 1987207 | 99.36% | 1977507 | 98.88% | 2048 | | 137 | |
| 40 | BWA | -n 0.04 | 1990092 | 99.50% | 1981426 | 99.07% | 1968640 | 98.43% | 4036 | | 6 | |
| 41 | BWA | -n 0.06 | 1987594 | 99.38% | 1978949 | 98.95% | 1966199 | 98.31% | 3365 | | 6 | |
| 42 | BWA | -n 0.08 | 1987594 | 99.38% | 1978949 | 98.95% | 1966199 | 98.31% | 3218 | | 6 | |
| 43 | GEM | -m 0.04 -e 0.04 | 1998819 | 99.94% | 1991583 | 99.58% | 1972285 | 98.61% | 364 | | 37 | |
| 44 | GEM | -m 0.06 -e 0.06 | 1999918 | 100.00% | 1992589 | 99.63% | 1973309 | 98.67% | 627 | | 37 | |
| 45 | GEM | -m 0.08 -e 0.08 | 1999991 | 100.00% | 1992624 | 99.63% | 1973292 | 98.66% | 1902 | | 37 | |
| 46 | STAR | default | 1995990 | 99.80% | 1972255 | 98.61% | 1952822 | 97.64% | 268 | | 43 | |
| 47 | deBGA | -k 22 -i 5 -n 300 (default) | 1999823 | 99.99% | 1995840 | 99.79% | 1976090 | 98.80% | 227 | | 22 | |
| 48 | deBGA | -k 22 -i 7 -n 300 | 1999773 | 99.99% | 1993662 | 99.68% | 1973945 | 98.70% | 211 | | 22 | |
| 49 | deBGA | -k 22 -i 9 -n 300 | 1999778 | 99.99% | 1992169 | 99.61% | 1972677 | 98.63% | 175 | | 22 | |
| 50 | deBGA | -k 24 -i 5 -n 300 | 1999820 | 99.99% | 1996083 | 99.80% | 1976406 | 98.82% | 189 | | 22 | |
| 51 | deBGA | -k 26 -i 5 -n 300 | 1999791 | 99.99% | 1995551 | 99.78% | 1975863 | 98.79% | 159 | | 22 | |
| 52 | deBGA | -k 28 -i 5 -n 300 | 1999804 | 99.99% | 1995460 | 99.77% | 1975787 | 98.79% | 143 | | 22 | |
| 53 | deBGA | -k 22 -i 5 -n 100 | 1999807 | 99.99% | 1996008 | 99.80% | 1975989 | 98.80% | 258 | | 22 | |
| 54 | deBGA | -k 22 -i 5 -n 500 | 1999834 | 99.99% | 1995636 | 99.78% | 1976133 | 98.81% | 212 | | 22 | |
| 55 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 | 1999756 | 99.99% | 1995278 | 99.76% | 1975667 | 98.78% | 206 | | 22 | |
| 56 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 | 1999656 | 99.98% | 1994943 | 99.75% | 1975390 | 98.77% | 213 | | 22 | |
| 57 | deBGA | -k 22 -i 5 -n 300 --cl 0.35 | 1999547 | 99.98% | 1994630 | 99.73% | 1975082 | 98.75% | 211 | | 22 | |
| 58 | deBGA | -k 22 -i 5 -n 300 --local | 1999500 | 99.98% | 1994393 | 99.72% | 1974848 | 98.74% | 221 | | 22 | |
| ***H. sapiens* Sim-150 (150 bp × 150 bp)** | | | | | | | | | |  | |
| 59 | Bowtie2 | very fast | 1997807 | 99.89% | 1974219 | 98.71% | 1974219 | 98.71% | 867 | | 29 | |
| 60 | Bowtie2 | fast | 1998341 | 99.92% | 1975522 | 98.78% | 1975522 | 98.78% | 946 | | 29 | |
| 61 | Bowtie2 | sensitive | 1999810 | 99.99% | 1979379 | 98.97% | 1979379 | 98.97% | 1256 | | 29 | |
| 62 | Bowtie2 | very sensitive | 1999956 | 100.00% | 1980768 | 99.04% | 1980768 | 99.04% | 2435 | | 29 | |
| 63 | BWA-MEM | default | 2000000 | 100.00% | 1981842 | 99.09% | 1981842 | 99.09% | 940 | | 6 | |
| 64 | BWA-MEM | -r 5 | 2000000 | 100.00% | 1981922 | 99.09% | 1981922 | 99.09% | 651 | | 6 | |
| 65 | BWA-MEM | -r 10 | 2000000 | 100.00% | 1981937 | 99.09% | 1981937 | 99.09% | 542 | | 6 | |
| 66 | BWA-MEM | -r 100 | 2000000 | 100.00% | 1981937 | 99.10% | 1981937 | 99.10% | 561 | | 6 | |
| 67 | SeqAlto | default | 1999991 | 100.00% | 1989736 | 99.49% | 1981155 | 99.06% | 6494 | | 137 | |
| 68 | SeqAlto | fast | 1999873 | 99.99% | 1989016 | 99.45% | 1980201 | 99.01% | 3002 | | 137 | |
| 69 | BWA | -n 0.04 | 1984130 | 99.21% | 1977450 | 98.87% | 1965178 | 98.26% | 4823 | | 6 | |
| 70 | BWA | -n 0.06 | 1984130 | 99.21% | 1977450 | 98.87% | 1965178 | 98.26% | 4769 | | 6 | |
| 71 | BWA | -n 0.08 | 1984130 | 99.21% | 1977450 | 98.87% | 1965178 | 98.26% | 4985 | | 6 | |
| 72 | GEM | -m 0.04 -e 0.04 | 1997625 | 99.88% | 1918423 | 95.92% | 1900239 | 95.01% | 420 | | 37 | |
| 73 | GEM | -m 0.06 -e 0.06 | 1999759 | 99.99% | 1920182 | 96.01% | 1901941 | 95.10% | 1271 | | 37 | |
| 74 | GEM | -m 0.08 -e 0.08 | 1999971 | 100.00% | 1920204 | 96.01% | 1901929 | 95.10% | 1721 | | 37 | |
| 75 | STAR | default | 1996692 | 99.83% | 1977627 | 98.88% | 1953561 | 97.68% | 261 | | 434 | |
| 76 | deBGA | -k 22 -i 5 -n 300 (default) | 1999912 | 100.00% | 1996440 | 99.82% | 1978761 | 98.94% | 254 | | 22 | |
| 77 | deBGA | -k 22 -i 7 -n 300 | 1999902 | 100.00% | 1995241 | 99.76% | 1977481 | 98.87% | 234 | | 22 | |
| 78 | deBGA | -k 22 -i 9 -n 300 | 1999885 | 99.99% | 1994204 | 99.71% | 1976585 | 98.83% | 208 | | 22 | |
| 79 | deBGA | -k 24 -i 5 -n 300 | 1999910 | 100.00% | 1996576 | 99.83% | 1978897 | 98.94% | 214 | | 22 | |
| 80 | deBGA | -k 26 -i 5 -n 300 | 1999906 | 100.00% | 1996851 | 99.84% | 1979169 | 98.96% | 207 | | 22 | |
| 81 | deBGA | -k 28 -i 5 -n 300 | 1999896 | 99.99% | 1996906 | 99.85% | 1979196 | 98.96% | 176 | | 22 | |
| 82 | deBGA | -k 22 -i 5 -n 100 | 1999889 | 99.99% | 1996588 | 99.83% | 1978592 | 98.93% | 284 | | 22 | |
| 83 | deBGA | -k 22 -i 5 -n 500 | 1999920 | 100.00% | 1996308 | 99.82% | 1978749 | 98.94% | 244 | | 22 | |
| 84 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 | 1999887 | 99.99% | 1995904 | 99.80% | 1978281 | 98.91% | 240 | | 22 | |
| 85 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 | 1999844 | 99.99% | 1995708 | 99.79% | 1978147 | 98.91% | 254 | | 22 | |
| 86 | deBGA | -k 22 -i 5 -n 300 --cl 0.35 | 1999795 | 99.99% | 1995528 | 99.78% | 1977940 | 98.90% | 247 | | 22 | |
| 87 | deBGA | -k 22 -i 5 -n 300 --local | 1999761 | 99.99% | 1995419 | 99.77% | 1977855 | 98.89% | 256 | | 22 | |
| ***H. sapiens* Sim-200 (200 bp × 200 bp)** | | | | | | | | | |  | |
| 88 | Bowtie2 | very fast | 1998211 | 99.91% | 1976032 | 98.80% | 1976032 | 98.80% | 1210 | | 29 | |
| 89 | Bowtie2 | fast | 1998847 | 99.94% | 1977848 | 98.89% | 1977848 | 98.89% | 1326 | | 29 | |
| 90 | Bowtie2 | sensitive | 1999806 | 99.99% | 1981938 | 99.10% | 1981938 | 99.10% | 1664 | | 29 | |
| 91 | Bowtie2 | very sensitive | 1999968 | 100.00% | 1983624 | 99.18% | 1983624 | 99.18% | 3154 | | 29 | |
| 92 | BWA-MEM | default | 2000000 | 100.00% | 1984616 | 99.23% | 1984616 | 99.23% | 1148 | | 6 | |
| 93 | BWA-MEM | -r 5 | 2000000 | 100.00% | 1984652 | 99.23% | 1984652 | 99.23% | 919 | | 6 | |
| 94 | BWA-MEM | -r 10 | 2000000 | 100.00% | 1984675 | 99.23% | 1984675 | 99.23% | 736 | | 6 | |
| 95 | BWA-MEM | -r 100 | 2000000 | 100.00% | 1984700 | 99.24% | 1984700 | 99.24% | 731 | | 6 | |
| 96 | SeqAlto | default | 1999977 | 100.00% | 1991532 | 99.58% | 1984179 | 99.21% | 12578 | | 137 | |
| 97 | SeqAlto | fast | 1999952 | 100.00% | 1991170 | 99.56% | 1983793 | 99.19% | 5428 | | 137 | |
| 98 | BWA | -n 0.04 | 1970368 | 98.52% | 1965461 | 98.27% | 1954770 | 97.74% | 10251 | | 6 | |
| 99 | BWA | -n 0.06 | 1968455 | 98.42% | 1963619 | 98.18% | 1952984 | 97.65% | 8124 | | 6 | |
| 100 | BWA | -n 0.08 | 1968455 | 98.42% | 1963619 | 98.18% | 1952984 | 97.65% | 8506 | | 6 | |
| 101 | GEM | -m 0.04 -e 0.04 | 1987973 | 99.40% | 1857859 | 92.89% | 1843190 | 92.16% | 840 | | 37 | |
| 102 | GEM | -m 0.06 -e 0.06 | 1999406 | 99.97% | 1867406 | 93.37% | 1852584 | 92.63% | 1732 | | 37 | |
| 103 | GEM | -m 0.08 -e 0.08 | 1999970 | 100.00% | 1866202 | 93.31% | 1851290 | 92.56% | 3140 | | 37 | |
| 104 | STAR | default | 1996926 | 99.85% | 1989630 | 99.48% | 1952250 | 97.61% | 414 | | 43 | |
| 105 | deBGA | -k 22 -i 5 -n 300 (default) | 1999965 | 100.00% | 1997489 | 99.87% | 1982513 | 99.13% | 314 | | 22 | |
| 106 | deBGA | -k 22 -i 7 -n 300 | 1999963 | 100.00% | 1996931 | 99.85% | 1982096 | 99.10% | 299 | | 22 | |
| 107 | deBGA | -k 22 -i 9 -n 300 | 1999955 | 100.00% | 1996301 | 99.82% | 1981339 | 99.07% | 244 | | 22 | |
| 108 | deBGA | -k 24 -i 5 -n 300 | 1999962 | 100.00% | 1997525 | 99.88% | 1982476 | 99.12% | 264 | | 22 | |
| 109 | deBGA | -k 26 -i 5 -n 300 | 1999962 | 100.00% | 1997645 | 99.88% | 1982616 | 99.13% | 258 | | 22 | |
| 110 | deBGA | -k 28 -i 5 -n 300 | 1999963 | 100.00% | 1997634 | 99.88% | 1982600 | 99.13% | 233 | | 22 | |
| 111 | deBGA | -k 22 -i 5 -n 100 | 1999954 | 100.00% | 1997662 | 99.88% | 1982602 | 99.13% | 351 | | 22 | |
| 112 | deBGA | -k 22 -i 5 -n 500 | 1999967 | 100.00% | 1997322 | 99.87% | 1982478 | 99.12% | 299 | | 22 | |
| 113 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 | 1999942 | 100.00% | 1997011 | 99.85% | 1982068 | 99.10% | 284 | | 22 | |
| 114 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 | 1999922 | 100.00% | 1996845 | 99.84% | 1981908 | 99.10% | 295 | | 22 | |
| 115 | deBGA | -k 22 -i 5 -n 300 --cl 0.35 | 1999896 | 99.99% | 1996740 | 99.84% | 1981820 | 99.09% | 280 | | 22 | |
| 116 | deBGA | -k 22 -i 5 -n 300 --local | 1999889 | 99.99% | 1996691 | 99.83% | 1981770 | 99.09% | 287 | | 22 | |
| ***H. sapiens* Sim-250 (250 bp × 250 bp)** | | | | | | | | | |  | |
| 117 | Bowtie2 | very fast | 1998134 | 99.91% | 1976489 | 98.82% | 1976489 | 98.82% | 1497 | | 29 | |
| 118 | Bowtie2 | fast | 1998829 | 99.94% | 1978414 | 98.92% | 1978414 | 98.92% | 1631 | | 29 | |
| **119** | **Bowtie2** | **sensitive** | **1999884** | **99.99%** | **1983544** | **99.18%** | **1983544** | **99.18%** | **2119** | | **29** | |
| 120 | Bowtie2 | very sensitive | 1999969 | 100.00% | 1985541 | 99.28% | 1985541 | 99.28% | 4141 | | 29 | |
| **121** | **BWA-MEM** | **default** | **2000000** | **100.00%** | **1986304** | **99.32%** | **1986304** | **99.32%** | **1463** | | **6** | |
| 122 | BWA-MEM | -r 5 | 2000000 | 100.00% | 1986283 | 99.31% | 1986283 | 99.31% | 1202 | | 6 | |
| 123 | BWA-MEM | -r 10 | 2000000 | 100.00% | 1986212 | 99.31% | 1986212 | 99.31% | 1021 | | 6 | |
| 124 | BWA-MEM | -r 100 | 2000000 | 100.00% | 1986209 | 99.31% | 1986209 | 99.31% | 962 | | 6 | |
| **125** | **SeqAlto** | **default** | **1999970** | **100.00%** | **1992523** | **99.63%** | **1985935** | **99.30%** | **30753** | | **137** | |
| 126 | SeqAlto | fast | 1999965 | 100.00% | 1992394 | 99.62% | 1985775 | 99.29% | 13690 | | 137 | |
| **127** | **BWA** | **-n 0.04** | **1943897** | **97.19%** | **1913602** | **95.68%** | **1903867** | **95.19%** | **16448** | | **6** | |
| 128 | BWA | -n 0.06 | 1943897 | 97.19% | 1913602 | 95.68% | 1903867 | 95.19% | 16294 | | 6 | |
| 129 | BWA | -n 0.08 | 1941541 | 97.08% | 1910969 | 95.55% | 1901313 | 95.07% | 13033 | | 6 | |
| **130** | **GEM** | **-m 0.04 -e 0.04** | **1992035** | **99.60%** | **1984358** | **99.22%** | **1970660** | **98.53%** | **1171** | | **37** | |
| 131 | GEM | -m 0.06 -e 0.06 | 1999804 | 99.99% | 1991104 | 99.56% | 1977303 | 98.87% | 2506 | | 37 | |
| 132 | GEM | -m 0.08 -e 0.08 | 1999997 | 100.00% | 1990140 | 99.51% | 1976302 | 98.82% | 4628 | | 37 | |
| **133** | **STAR** | **default** | **1996920** | **99.85%** | **1963061** | **98.15%** | **1949586** | **97.48%** | **419** | | **43** | |
| **134** | **deBGA** | **-k 22 -i 5 -n 300 (default)** | **1999994** | **100.00%** | **1998012** | **99.90%** | **1984533** | **99.23%** | **365** | | **22** | |
| 135 | deBGA | -k 22 -i 7 -n 300 | 1999998 | 100.00% | 1997534 | 99.88% | 1984152 | 99.21% | 345 | | 22 | |
| 136 | deBGA | -k 22 -i 9 -n 300 | 1999997 | 100.00% | 1996769 | 99.84% | 1983368 | 99.17% | 294 | | 22 | |
| 137 | deBGA | -k 24 -i 5 -n 300 | 1999996 | 100.00% | 1998168 | 99.91% | 1984779 | 99.24% | 338 | | 22 | |
| 138 | deBGA | -k 26 -i 5 -n 300 | 1999993 | 100.00% | 1998176 | 99.91% | 1984563 | 99.23% | 315 | | 22 | |
| 139 | deBGA | -k 28 -i 5 -n 300 | 1999993 | 100.00% | 1998264 | 99.91% | 1984705 | 99.24% | 288 | | 22 | |
| 140 | deBGA | -k 22 -i 5 -n 100 | 1999995 | 100.00% | 1998016 | 99.90% | 1984443 | 99.22% | 414 | | 22 | |
| 141 | deBGA | -k 22 -i 5 -n 500 | 1999995 | 100.00% | 1997992 | 99.90% | 1984551 | 99.23% | 368 | | 22 | |
| 142 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 | 1999992 | 100.00% | 1997720 | 99.89% | 1984258 | 99.21% | 354 | | 22 | |
| 143 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 | 1999988 | 100.00% | 1997633 | 99.88% | 1984176 | 99.21% | 354 | | 22 | |
| **144** | **deBGA** | **-k 22 -i 5 -n 300 --cl 0.35** | **1999981** | **100.00%** | **1997569** | **99.88%** | **1984110** | **99.21%** | **343** | | **22** | |
| 145 | deBGA | -k 22 -i 5 -n 300 --local | 1999979 | 100.00% | 1997523 | 99.88% | 1984064 | 99.20% | 360 | | 22 | |

Results on simulated human datasets (lines 16-20 of Supplementary Table 1). a) - e) Various statistics on the sensitivities, correctness and throughputs of the aligners. Refer to Supplementary Notes (the “Statistics for evaluating alignments” subsection) for definition. The lines in bold font are involved in Fig. 4a.

# Supplementary Table 12. The numbers of human simulated reads correctly aligned by the primary and alternative alignments of deBGA

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Aligner** | **Parameter** | **Correctly aligned by primary alignment a** | | **Correctly aligned by alternative alignment with identical edit distance b** | | **Correctly aligned by alternative alignment with larger edit distance c** | |
| # | % | # | % | # | % |
| ***H. sapiens* Sim-100 (100 bp × 100 bp)** | | | | | | | | |
| 1 | deBGA | -k 22 -i 5 -n 300 (default) | 1970877 | 98.54% | 20311 | 1.02% | 2134 | 0.11% |
| 2 | deBGA | -k 22 -i 7 -n 300 | 1968192 | 98.41% | 20118 | 1.01% | 2040 | 0.10% |
| 3 | deBGA | -k 22 -i 9 -n 300 | 1968035 | 98.40% | 20513 | 1.03% | 2071 | 0.10% |
| 4 | deBGA | -k 24 -i 5 -n 300 | 1971016 | 98.55% | 20228 | 1.01% | 2144 | 0.11% |
| 5 | deBGA | -k 26 -i 5 -n 300 | 1970627 | 98.53% | 20121 | 1.01% | 2280 | 0.11% |
| 6 | deBGA | -k 28 -i 5 -n 300 | 1970409 | 98.52% | 19867 | 0.99% | 2403 | 0.12% |
| 7 | deBGA | -k 22 -i 5 -n 100 | 1970749 | 98.54% | 20545 | 1.03% | 2211 | 0.11% |
| 8 | deBGA | -k 22 -i 5 -n 500 | 1970969 | 98.55% | 20041 | 1.00% | 2077 | 0.10% |
| 9 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 | 1970529 | 98.53% | 20205 | 1.01% | 2066 | 0.10% |
| 10 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 | 1970029 | 98.50% | 20202 | 1.01% | 2026 | 0.10% |
| 11 | deBGA | -k 22 -i 5 -n 300 --cl 0.35 | 1969681 | 98.48% | 20184 | 1.01% | 2016 | 0.10% |
| 12 | deBGA | -k 22 -i 5 -n 300 --local | 1969182 | 98.46% | 20155 | 1.01% | 2003 | 0.10% |
| ***H. sapiens* Sim-125 (125 bp × 125 bp)** | | | | | | | | |
| 13 | deBGA | -k 22 -i 5 -n 300 (default) | 1976090 | 98.80% | 17661 | 0.88% | 2089 | 0.10% |
| 14 | deBGA | -k 22 -i 7 -n 300 | 1973945 | 98.70% | 17663 | 0.88% | 2054 | 0.10% |
| 15 | deBGA | -k 22 -i 9 -n 300 | 1972677 | 98.63% | 17407 | 0.87% | 2085 | 0.10% |
| 16 | deBGA | -k 24 -i 5 -n 300 | 1976406 | 98.82% | 17533 | 0.88% | 2144 | 0.11% |
| 17 | deBGA | -k 26 -i 5 -n 300 | 1975863 | 98.79% | 17502 | 0.88% | 2186 | 0.11% |
| 18 | deBGA | -k 28 -i 5 -n 300 | 1975787 | 98.79% | 17488 | 0.87% | 2185 | 0.11% |
| 19 | deBGA | -k 22 -i 5 -n 100 | 1975989 | 98.80% | 17868 | 0.89% | 2151 | 0.11% |
| 20 | deBGA | -k 22 -i 5 -n 500 | 1976133 | 98.81% | 17471 | 0.87% | 2032 | 0.10% |
| 21 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 | 1975667 | 98.78% | 17619 | 0.88% | 1992 | 0.10% |
| 22 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 | 1975390 | 98.77% | 17590 | 0.88% | 1963 | 0.10% |
| 23 | deBGA | -k 22 -i 5 -n 300 --cl 0.35 | 1975082 | 98.75% | 17585 | 0.88% | 1963 | 0.10% |
| 24 | deBGA | -k 22 -i 5 -n 300 --local | 1974848 | 98.74% | 17580 | 0.88% | 1965 | 0.10% |
| ***H. sapiens* Sim-150 (150 bp × 150 bp)** | | | | | | | | |
| 25 | deBGA | -k 22 -i 5 -n 300 (default) | 1978761 | 98.94% | 15611 | 0.78% | 2068 | 0.10% |
| 26 | deBGA | -k 22 -i 7 -n 300 | 1977481 | 98.87% | 15657 | 0.78% | 2103 | 0.11% |
| 27 | deBGA | -k 22 -i 9 -n 300 | 1976585 | 98.83% | 15419 | 0.77% | 2200 | 0.11% |
| 28 | deBGA | -k 24 -i 5 -n 300 | 1978897 | 98.94% | 15554 | 0.78% | 2125 | 0.11% |
| 29 | deBGA | -k 26 -i 5 -n 300 | 1979169 | 98.96% | 15502 | 0.78% | 2180 | 0.11% |
| 30 | deBGA | -k 28 -i 5 -n 300 | 1979196 | 98.96% | 15504 | 0.77% | 2206 | 0.11% |
| 31 | deBGA | -k 22 -i 5 -n 100 | 1978592 | 98.93% | 15801 | 0.79% | 2195 | 0.11% |
| 32 | deBGA | -k 22 -i 5 -n 500 | 1978749 | 98.94% | 15466 | 0.77% | 2093 | 0.10% |
| 33 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 | 1978281 | 98.91% | 15560 | 0.78% | 2063 | 0.10% |
| 34 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 | 1978147 | 98.91% | 15576 | 0.78% | 1985 | 0.10% |
| 35 | deBGA | -k 22 -i 5 -n 300 --cl 0.35 | 1977940 | 98.90% | 15567 | 0.78% | 2021 | 0.10% |
| 36 | deBGA | -k 22 -i 5 -n 300 --local | 1977855 | 98.89% | 15525 | 0.78% | 2039 | 0.10% |
| ***H. sapiens* Sim-200 (200 bp × 200 bp)** | | | | | | | | |
| 37 | deBGA | -k 22 -i 5 -n 300 (default) | 1982513 | 99.13% | 12514 | 0.63% | 2462 | 0.12% |
| 38 | deBGA | -k 22 -i 7 -n 300 | 1982096 | 99.10% | 12409 | 0.62% | 2426 | 0.12% |
| 39 | deBGA | -k 22 -i 9 -n 300 | 1981339 | 99.07% | 12455 | 0.62% | 2507 | 0.13% |
| 40 | deBGA | -k 24 -i 5 -n 300 | 1982476 | 99.12% | 12560 | 0.63% | 2489 | 0.12% |
| 41 | deBGA | -k 26 -i 5 -n 300 | 1982616 | 99.13% | 12558 | 0.63% | 2471 | 0.12% |
| 42 | deBGA | -k 28 -i 5 -n 300 | 1982600 | 99.13% | 12518 | 0.62% | 2516 | 0.13% |
| 43 | deBGA | -k 22 -i 5 -n 100 | 1982602 | 99.13% | 12612 | 0.63% | 2448 | 0.12% |
| 44 | deBGA | -k 22 -i 5 -n 500 | 1982478 | 99.12% | 12392 | 0.62% | 2452 | 0.12% |
| 45 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 | 1982068 | 99.10% | 12470 | 0.62% | 2473 | 0.12% |
| 46 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 | 1981908 | 99.10% | 12455 | 0.62% | 2482 | 0.12% |
| 47 | deBGA | -k 22 -i 5 -n 300 --cl 0.35 | 1981820 | 99.09% | 12458 | 0.62% | 2462 | 0.12% |
| 48 | deBGA | -k 22 -i 5 -n 300 --local | 1981770 | 99.09% | 12466 | 0.62% | 2455 | 0.12% |
| ***H. sapiens* Sim-250 (250 bp × 250 bp)** | | | | | | | | |
| 49 | deBGA | -k 22 -i 5 -n 300 (default) | 1984533 | 99.23% | 10923 | 0.55% | 2556 | 0.13% |
| 50 | deBGA | -k 22 -i 7 -n 300 | 1984152 | 99.21% | 10787 | 0.54% | 2595 | 0.13% |
| 51 | deBGA | -k 22 -i 9 -n 300 | 1983368 | 99.17% | 10816 | 0.54% | 2585 | 0.13% |
| 52 | deBGA | -k 24 -i 5 -n 300 | 1984779 | 99.24% | 10782 | 0.54% | 2607 | 0.13% |
| 53 | deBGA | -k 26 -i 5 -n 300 | 1984563 | 99.23% | 10994 | 0.55% | 2619 | 0.13% |
| 54 | deBGA | -k 28 -i 5 -n 300 | 1984705 | 99.24% | 10932 | 0.55% | 2627 | 0.13% |
| 55 | deBGA | -k 22 -i 5 -n 100 | 1984443 | 99.22% | 10944 | 0.55% | 2629 | 0.13% |
| 56 | deBGA | -k 22 -i 5 -n 500 | 1984551 | 99.23% | 10876 | 0.54% | 2565 | 0.13% |
| 57 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 | 1984258 | 99.21% | 10902 | 0.55% | 2560 | 0.13% |
| 58 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 | 1984176 | 99.21% | 10897 | 0.54% | 2560 | 0.13% |
| 59 | deBGA | -k 22 -i 5 -n 300 --cl 0.35 | 1984110 | 99.21% | 10896 | 0.54% | 2563 | 0.13% |
| 60 | deBGA | -k 22 -i 5 -n 300 --local | 1984064 | 99.20% | 10888 | 0.54% | 2571 | 0.13% |

The numbers of simulated reads from human genome correctly aligned by the primary and alternative alignments of deBGA. a) indicates the number of reads which are correctly aligned by the primary alignment; b) indicates the number of reads which are correctly aligned by alternative alignments, and the edit distance of the primary alignment is identical to that of the correct alternative alignment; c) indicates the number of reads which are correctly aligned by alternative alignments, and the edit distance of the primary alignment is larger than that of the correct alternative alignment.

# Supplementary Table 13. Alignment results on the reads spanning retrotransposon regions of human genome

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Aligner** | **Parameter** | **Correctly aligned b** | | **Correctly aligned by primary alignment c** | |
| # | % | # | % |
| ***H. sapiens* Sim-100 (100 bp × 100 bp): totally 3100 reads overlapping retrotransposon regions with at least 20 bp** | | | | | | |
| 1 | Bowtie2 | sensitive | 2768 | 89.29% | 2768 | 89.29% |
| 2 | BWA-MEM | default | 2788 | 89.94% | 2788 | 89.94% |
| 3 | SeqAlto | default | 2934 | 94.65% | 2789 | 89.97% |
| 4 | BWA | -n 0.04 | 2847 | 91.84% | 2748 | 88.65% |
| 5 | GEM | -m 0.04 -e 0.04 | 3073 | 99.13% | 2732 | 88.13% |
| 6 | STAR | default | 2894 | 93.35% | 2747 | 88.61% |
| 7 | deBGA | -k 22 -i 5 -n 300 (default) | 3081 | 99.39% | 2720 | 87.74% |
| 8 | deBGA | -k 22 -i 5 -n 300 --cl 0.35 | 3077 | 99.26% | 2718 | 87.68% |
| ***H. sapiens* Sim-250 (250 bp × 250 bp): totally 2962 reads overlapping retrotransposon regions with at least 20 bp** | | | | | | |
| 9 | Bowtie2 | sensitive | 2796 | 94.40% | 2796 | 94.40% |
| 10 | BWA-MEM | default | 2795 | 94.36% | 2795 | 94.36% |
| 11 | SeqAlto | default | 2883 | 97.33% | 2808 | 94.80% |
| 12 | BWA | -n 0.04 | 2748 | 92.78% | 2652 | 89.53% |
| 13 | GEM | -m 0.04 -e 0.04 | 2917 | 98.48% | 2709 | 91.46% |
| 14 | STAR | default | 2869 | 96.86% | 2721 | 91.86% |
| 15 | deBGA | -k 22 -i 5 -n 300 (default) | 2956 | 99.80% | 2768 | 93.45% |
| 16 | deBGA | -k 22 -i 5 -n 300 --cl 0.35 | 2954 | 99.73% | 2766 | 93.38% |

Here, the reads spanning retrotransposon regions indicate the reads have 20 bp overlapping with LTR, LINE and SINE regions. The positions of LTR, LINE and SINE regions are obtained from RepeatMasker (downloaded from UCSC genome browser). The numbers of the reads correctly aligned and correctly aligned by the primary alignments are assessed. DeBGA achieved equally highest number of correctly aligned reads. Only considering the primary alignments, the accuracy of deBGA is also comparable to that of other state-of-the-art aligners. The small decrease of the accuracy of primary alignments is partially caused by that deBGA prioritizes the multiple equally good alignments of the read by their alignment positions, and the alignment having leftmost position is output as primary alignment.

# Supplementary Table 14. Alignment results on real datasets from human genome

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **No** | **Aligner** | **Parameter** | **Aligned a** | | **Alignment Time b** | **Index Loading Time c** |
| # | % |
| **ERR174324 (100 bp × 102 bp)** | | | | | | |
| 1 | Bowtie2 | very fast | 1914430 | 95.72% | 826 | 29 |
| 2 | Bowtie2 | fast | 1918190 | 95.91% | 881 | 29 |
| 3 | **Bowtie2** | **sensitive** | **1929148** | **96.46%** | **1162** | **29** |
| 4 | Bowtie2 | very sensitive | 1935233 | 96.76% | 2087 | 29 |
| 5 | **BWA-MEM** | **default** | **1986275** | **99.31%** | **842** | **6** |
| 6 | BWA-MEM | -r 5 | 1986185 | 99.31% | 593 | 6 |
| 7 | BWA-MEM | -r 10 | 1986183 | 99.31% | 540 | 6 |
| 8 | BWA-MEM | -r 100 | 1986183 | 99.31% | 590 | 6 |
| 9 | **SeqAlto** | **default** | **1954460** | **97.72%** | **2037** | **137** |
| 10 | SeqAlto | fast | 1951171 | 97.56% | 1243 | 137 |
| 11 | **BWA** | **-n 0.04** | **1900876** | **95.04%** | **1945** | **6** |
| 12 | BWA | -n 0.06 | 1891643 | 94.58% | 1419 | 6 |
| 13 | BWA | -n 0.08 | 1891643 | 94.58% | 1401 | 6 |
| 14 | **GEM** | **-m 0.04 -e 0.04** | **1909390** | **95.47%** | **416** | **37** |
| 15 | GEM | -m 0.06 -e 0.06 | 1931560 | 96.58% | 828 | 37 |
| 16 | GEM | -m 0.08 -e 0.08 | 1949248 | 97.46% | 1948 | 37 |
| 17 | **STAR** | **default** | **1888146** | **94.41%** | **330** | **43** |
| **18** | **deBGA** | **-k 22 -i 5 -n 300 (default)** | **1983623** | **99.18%** | **332** | **22** |
| 19 | deBGA | -k 22 -i 7 -n 300 | 1983183 | 99.16% | 320 | 22 |
| 20 | deBGA | -k 22 -i 9 -n 300 | 1982401 | 99.12% | 330 | 22 |
| 21 | deBGA | -k 24 -i 5 -n 300 | 1981538 | 99.08% | 283 | 22 |
| 22 | deBGA | -k 26 -i 5 -n 300 | 1978906 | 98.95% | 260 | 22 |
| 23 | deBGA | -k 28 -i 5 -n 300 | 1975432 | 98.77% | 200 | 22 |
| 24 | deBGA | -k 22 -i 5 -n 100 | 1983534 | 99.18% | 344 | 22 |
| 25 | deBGA | -k 22 -i 5 -n 500 | 1983704 | 99.19% | 346 | 22 |
| 26 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 | 1984713 | 99.24% | 272 | 22 |
| 27 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 | 1984942 | 99.25% | 221 | 22 |
| **28** | **deBGA** | **-k 22 -i 5 -n 300 --cl 0.35** | **1984956** | **99.25%** | **217** | **22** |
| 29 | deBGA | -k 22 -i 5 -n 300 --local | 1984804 | 99.24% | 292 | 22 |
| **Hiseq X Ten: NA12878\_L3 (150 bp × 150 bp)** | | | | | | |
| 30 | Bowtie2 | very fast | 1825215 | 91.26% | 1023 | 29 |
| 31 | Bowtie2 | fast | 1829272 | 91.46% | 1098 | 29 |
| **32** | **Bowtie2** | **sensitive** | **1838705** | **91.93%** | **1432** | **29** |
| 33 | Bowtie2 | very sensitive | 1841771 | 92.09% | 2532 | 29 |
| **34** | **BWA-MEM** | **default** | **1912064** | **95.60%** | **1040** | **6** |
| 35 | BWA-MEM | -r 5 | 1911971 | 95.60% | 789 | 6 |
| 36 | BWA-MEM | -r 10 | 1911971 | 95.60% | 730 | 6 |
| 37 | BWA-MEM | -r 100 | 1911971 | 95.60% | 730 | 6 |
| **38** | **SeqAlto** | **default** | **1848975** | **92.45%** | **3154** | **137** |
| 39 | SeqAlto | fast | 1846677 | 92.33% | 1602 | 137 |
| **40** | **BWA** | **-n 0.04** | **1803964** | **90.20%** | **2360** | **6** |
| 41 | BWA | -n 0.06 | 1803905 | 90.20% | 2339 | 6 |
| 42 | BWA | -n 0.08 | 1803550 | 90.18% | 2350 | 6 |
| **43** | **GEM** | **-m 0.04 -e 0.04** | **1822660** | **91.13%** | **468** | **37** |
| 44 | GEM | -m 0.06 -e 0.06 | 1841606 | 92.08% | 1036 | 37 |
| 45 | GEM | -m 0.08 -e 0.08 | 1858797 | 92.94% | 2734 | 37 |
| **46** | **STAR** | **default** | **1790494** | **89.52%** | **546** | **43** |
| **47** | **deBGA** | **-k 22 -i 5 -n 300 (default)** | **1912847** | **95.64%** | **570** | **22** |
| 48 | deBGA | -k 22 -i 7 -n 300 | 1911884 | 95.59% | 520 | 22 |
| 49 | deBGA | -k 22 -i 9 -n 300 | 1908065 | 95.40% | 453 | 22 |
| 50 | deBGA | -k 24 -i 5 -n 300 | 1908065 | 95.40% | 444 | 22 |
| 51 | deBGA | -k 26 -i 5 -n 300 | 1903797 | 95.19% | 381 | 22 |
| 52 | deBGA | -k 28 -i 5 -n 300 | 1900215 | 95.01% | 340 | 22 |
| 53 | deBGA | -k 22 -i 5 -n 100 | 1912284 | 95.61% | 603 | 22 |
| 54 | deBGA | -k 22 -i 5 -n 500 | 1912826 | 95.64% | 620 | 22 |
| 55 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 | 1914366 | 95.72% | 518 | 22 |
| 56 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 | 1914853 | 95.74% | 349 | 22 |
| **57** | **deBGA** | **-k 22 -i 5 -n 300 --cl 0.35** | **1914949** | **95.75%** | **347** | **22** |
| 58 | deBGA | -k 22 -i 5 -n 300 --local | 1914577 | 95.73% | 476 | 22 |
| **ERR161544 (100 bp × 100 bp)** | | | | | | |
| 59 | Bowtie2 | very fast | 1921894 | 96.09% | 848 | 29 |
| 60 | Bowtie2 | fast | 1926306 | 96.32% | 861 | 29 |
| **61** | **Bowtie2** | **sensitive** | **1938258** | **96.91%** | **1119** | **29** |
| 62 | Bowtie2 | very sensitive | 1945733 | 97.29% | 1720 | 29 |
| **63** | **BWA-MEM** | **default** | **1988711** | **99.44%** | **797** | **6** |
| 64 | BWA-MEM | -r 5 | 1988580 | 99.43% | 514 | 6 |
| 65 | BWA-MEM | -r 10 | 1988580 | 99.43% | 485 | 6 |
| 66 | BWA-MEM | -r 100 | 1988580 | 99.43% | 506 | 6 |
| **67** | **SeqAlto** | **default** | **1958620** | **97.93%** | **1960** | **137** |
| 68 | SeqAlto | fast | 1954627 | 97.73% | 1225 | 137 |
| **69** | **BWA** | **-n 0.04** | **1906004** | **95.30%** | **1391** | **6** |
| 70 | BWA | -n 0.06 | 1896737 | 94.84% | 1001 | 6 |
| 71 | BWA | -n 0.08 | 1896737 | 94.84% | 1029 | 6 |
| **72** | **GEM** | **-m 0.04 -e 0.04** | **1906255** | **95.31%** | **403** | **37** |
| 73 | GEM | -m 0.06 -e 0.06 | 1926126 | 96.31% | 720 | 37 |
| 74 | GEM | -m 0.08 -e 0.08 | 1945637 | 97.28% | 1356 | 37 |
| **75** | **STAR** | **default** | **1899844** | **94.99%** | **320** | **55** |
| **76** | **deBGA** | **-k 22 -i 5 -n 300 (default)** | **1987061** | **99.35%** | **392** | **22** |
| 77 | deBGA | -k 22 -i 7 -n 300 | 1986807 | 99.34% | 410 | 22 |
| 78 | deBGA | -k 22 -i 9 -n 300 | 1986460 | 99.32% | 448 | 22 |
| 79 | deBGA | -k 24 -i 5 -n 300 | 1985731 | 99.29% | 314 | 22 |
| 80 | deBGA | -k 26 -i 5 -n 300 | 1983403 | 99.17% | 260 | 22 |
| 81 | deBGA | -k 28 -i 5 -n 300 | 1980560 | 99.03% | 221 | 22 |
| 82 | deBGA | -k 22 -i 5 -n 100 | 1987027 | 99.35% | 416 | 22 |
| 83 | deBGA | -k 22 -i 5 -n 500 | 1987096 | 99.35% | 405 | 22 |
| 84 | deBGA | -k 22 -i 5 -n 300 --cl 0.15 | 1987946 | 99.40% | 308 | 22 |
| 85 | deBGA | -k 22 -i 5 -n 300 --cl 0.25 | 1988272 | 99.41% | 235 | 22 |
| **86** | **deBGA** | **-k 22 -i 5 -n 300 --cl 0.35** | **1988328** | **99.42%** | **239** | **22** |
| 87 | deBGA | -k 22 -i 5 -n 300 --local | 1988171 | 99.41% | 358 | 22 |

Results on real human datasets (lines 21-23 of Supplementary Table 1). a) - c) Various statistics on the sensitivities and throughputs of the aligners. Refer to Supplementary Notes (the “Statistics for evaluating alignments” subsection) for definition. The lines in bold font are involved in Fig. 4b.

# Supplementary Table 15. Statistics on the called STRs of NA12878

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Aligner** | **#Total a** | **#HighCov b** | **#Matched c** |
| 1 | deBGA | 1558047 | 1454880 | 603017 |
| 2 | BWA-MEM | 1592775 | 1506495 | 616251 |

a) The total number of the STRs called by lobSTR based on the alignments of the corresponding aligner. b) The number of the called STRs having >5x coverage, this number is automatically calculated by lobSTR. c) The number of the called STRs which can be matched by at least one record of the STR callset of 1000 Genomes Phase1 data, which is produced by the author of lobSTR with the tool. Here, a called STR of deBGA or BWA-MEM is considered as matched only if there is at least one 1000 Genomes STR within 10 bp of the called STR. The callset of 1000 Genomes Phase1 data is available at <http://melissagymrek.com/lobstr-code/download.html>.

# Supplementary Table 16. Statistics on the called duplications and inversions of NA12878

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Aligner** | **Duplication** | | | **Inversion** | |  |
| **#Total a** | **#Matched b** | **#Recovered c** | **#Total a** | **#Matched b** | **#Recovered c** |
| 1 | deBGA | 4682 | 10 | 9 | 8140 | 49 | 28 |
| 2 | BWA-MEM | 2295 | 10 | 10 | 7955 | 50 | 28 |

a) The total number of the variants called by Delly based on the alignments of the corresponding aligner. b) The number of the called variants which can be matched by 1000 Genomes Project Release3 SV callset of NA12878. Here, a called duplication (inversion) is considered as matched, only if there is a variant marked as ‘‘SVTYPE=DUP’ or ‘SVTYPE=CNV’ (‘SVTYPE=INV’ for inversion) in the NA12878 SV callset meeting the reciprocal 50% criteria with the called duplication (inversion). The 1000 Genomes Project Release3 SV callset of NA12878 is available at: <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/ALL.wgs.integrated_sv_map_v2.20130502.svs.genotypes.vcf.gz>. c) The number of the recovered variants of the 1000 Genomes Project NA12878 callset. Here, a variant of the 1000 Genomes Project NA12878 callset is considered as recovered only if there is a called variant by deBGA (BWA-MEM) meeting the reciprocal 50% criteria for this variant.

# Supplementary Table 17. Time and memory footprint of RdBG-index construction

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Reference** | **Time a** | **Memory b** |
| 1 | *E. coli* K-12 MG1655 substrain | 21 | 2.02 |
| 2 | 10 *E. coli* strains | 116 | 3.02 |
| 3 | 20 *E. coli* strains | 200 | 3.02 |
| 4 | 40 *E. coli* strains | 280 | 4.03 |
| 5 | 62 *E. coli* strains | 729 | 5.04 |
| 6 | *A. thaliana* reference genome (TAIR10) | 972 | 7.05 |
| 7 | 19 *A. thaliana* strains | 4993 | 11.09 |
| 8 | RefSeq bacteria genomes | 69912 | 153.30 |
| 9 | xenograft model (hg19+MM10) | 42851 | 130.03 |
| 10 | human reference genome (hg19) | 15786 | 69.55 |

a)-b) The wall time and memory footprint of the index construction for the references (the *k*-mer size of the index is 22). The time is in seconds, and the memory footprints are in Gigabytes.

# Supplementary Table 18. Memory footprint of the alignment of deBGA

|  |  |  |
| --- | --- | --- |
| **No.** | **Dataset** | **Memory a** |
| 1 | *E. coli* Sim-62 | 9.07 |
| 2 | ERR008613 | 9.07 |
| 3 | SRR522163 | 9.07 |
| 4 | SRR530851 | 9.07 |
| 5 | *A. thaliana* Sim-19 | 12.1 |
| 6 | Meta Sim | 80.64 |
| 7 | Xenograft Sim | 60.48 |
| 8 | Meta Pseudo-real | 80.64 |
| 9 | *H. sapiens* Sim-100 | 40.32 |
| 10 | *H. sapiens* Sim-250 | 40.32 |
| 11 | ERR174324 | 40.32 |
| 12 | Hiseq X Ten: NA12878\_L3 | 40.32 |
| 13 | ERR161544 | 40.32 |

a) The memory footprint of deBGA when aligning the reads from the listed datasets (the alignment is conducted with the default setting of deBGA). The memory footprint are in Gigabytes.

# Supplementary Table 19. The alignment time of deBGA in multiple threads

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Dataset** | **1 thread a** | | **2 threads a** | | **4 threads a** | | **8 threads a** | |
| **default b** | ***--cl* 0.35 b** | **default b** | ***--cl* 0.35 b** | **default b** | ***--cl* 0.35 b** | **default b** | ***--cl* 0.35 b** |
| 1 | *E. coli* Sim-62 | 196 | 199 | 146 | 143 | 99 | 96 | 67 | 67 |
| 2 | *A. thaliana* Sim-19 | 250 | 185 | 152 | 141 | 105 | 88 | 73 | 68 |
| 3 | ERR008613 | 918 | 460 | 523 | 289 | 295 | 170 | 177 | 106 |
| 4 | SRR522163 | 345 | 337 | 223 | 210 | 138 | 125 | 95 | 83 |
| 5 | SRR530851 | 187 | 177 | 120 | 117 | 74 | 73 | 55 | 56 |
| 6 | Meta Sim | 130 | 100 | 68 | 68 | 45 | 44 | 33 | 32 |
| 7 | Xenograft Sim | 260 | 253 | 166 | 164 | 96 | 97 | 60 | 59 |
| 8 | Meta Pseudo-real | 582 | 378 | 350 | 236 | 204 | 146 | 127 | 97 |
| 9 | *H. sapiens* Sim-100 | 171 | 165 | 103 | 101 | 61 | 58 | 35 | 35 |
| 10 | *H. sapiens* Sim-250 | 365 | 343 | 206 | 198 | 112 | 109 | 67 | 63 |
| 11 | ERR174324 | 332 | 217 | 200 | 133 | 109 | 75 | 65 | 48 |
| 12 | Hiseq X Ten: NA12878\_L3 | 570 | 347 | 329 | 201 | 173 | 109 | 97 | 61 |
| 13 | ERR161544 | 392 | 239 | 249 | 158 | 136 | 88 | 76 | 51 |

a) DeBGA was run with various numbers of (1, 2, 4, and 8) threads to benchmark the alignment time. b) Both of the default and the more sensitive --*cl=*0.35 settings of deBGA were assessed. The alignment time is recorded in seconds.

# Supplementary Notes

## Implementation of the comparisons

Bowtie2 (version 2.2.4), BWA (version 0.7.5), BWA-MEM (version 0.7.5), SeqAlto (version 0.5-r123), GEM (version core\_i3-20121106-022124) and STAR (version 2.5.0a) were employed for comparison. Bowtie2, BWA and BWA-MEM were built from source files with default arguments. For SeqAlto and GEM, the executable files were obtained respectively from their official websites. For these aligners, we tuned the parameters most related to their sensitivities and throughputs referring to previous studies. We also tuned a variety of parameters for deBGA to investigate their effects, mainly on the *k*-mer size (*-k*), the minimal interval of seeding (*-i*), the max allowed number of hits per seed (*-n*) and the early-stop heuristic (*--cl*, *--local*). See Supplementary Tables 3, 4, 6 and 8 for detailed parameters used for the comparison.

All the benchmarks were conducted on a server with an Intel Xeon E4820 CPU at 2.00 GHz and 1 Terabytes RAM, running Linux Ubuntu 14.04. All the aligners were run with single thread if not specifically mentioned. Considering the compatibility to downstream analysis, the throughput of each aligner was assessed by the wall clock time from the beginning of read alignment to the completion of SAM output. The time of loading index is separately considered as it is constant. BWA and GEM need to call multiple modules during the alignment. For each of these modules, the execution time was separately recorded.

## The time cost of the construction of RdBG-index

The time costs of the construction of RdBG-index for various references are in Supplementary Table 17. The index construction is slower than that of some hash table-based aligners, e.g., SeqAlto, but on the same order to that of some suffix trie-based aligners like STAR and GEM. Furthermore, considering that the index usually needs to be only built once for the same reference, the time cost is still affordable to many tasks.

The index construction of deBGA is implemented a little straightforwardly, which is one of the main causes to the time cost. That is, deBGA collects all the tuples of (*k*+1-mer, position), and sorts them, to recognize the unipaths of RdBG and build the , and tables simultaneously. The sorting process costs quite a lot of time (time complexity: , where *n* is the numbers of the tuples). But it is possible to further improve the overall speed of index construction. Potential solutions include separately building the reference de Bruijn graph, recognizing the unipaths, and indexing them in multiple steps with more advanced algorithms. Moreover, as deBGA may be employed in many tasks to align reads against multiple genomes, it could be also very useful if new genomes can be added to existing index to efficiently build the index of the new composite reference. However, this task is also non-trivial as it needs to efficiently mix the de Bruijn graphs of two sets of genomes and re-build the index, which is still an open problem, especially for large amount of genome sequences. This is an important future work for us, to develop new algorithms for the fast construction and merging of RdBG-index.

## The memory footprint of deBGA

The memory footprints of deBGA for processing various datasets are in Supplementary Table 18. Generally, deBGA uses more RAM space than state-of-the-art aligners, especially those aligners using FM-index to index the references, e.g., BWA, Bowtie2 and GEM. The major RAM space cost stems from the two tables of RdBG-index, and . In theory, the sizes of these two tables respectively depend on the numbers of the *k*-mers and the copies of the unipaths of the reference. In practice, to support reference > 4 Gbp, we used 64-bit variables to record the IDs and genomic positions of unipaths. Thus, each line of requires at least 20 bytes, and each line of requires at least 8 bytes.

Although RdBG-index costs more RAM space, it is helpful to accelerate the alignment. That is, *k*-mers can be efficiently matched to the unipaths through , and the obtained unipath IDs can be used to fast merge the seeds. After merging, the positions of the merged seeds can also be directly retrieved through with very few operations. We think the tradeoff on the RAM space is worthwhile, considering the fast alignment speed. Meanwhile, this memory requirement can still be fulfilled by modern servers and workstations. Moreover, we also realized that it is critical to reduce the space cost of deBGA to make it suited to PCs or other computers with less RAM. This is also an important future work. Potential solutions include using compressed data structures to represent the hash table (). It is also feasible to reduce the redundancy of the data structures, e.g., the 64-bit variables may be unnecessary as 40-bit is enough to support the references having upto 1 Tbp.

The memory footprints of the construction of RdBG-index are also investigated (Supplementary Table 17). The construction of the index also has large memory footprints, mainly due to that deBGA keeps all the tuples of (*k-*mer, position) to build the unipaths as well as their positions. This memory footprint could be greatly reduced by using more advanced de Bruijn graph construction algorithms. Meanwhile, as it only needs to construct the index once for the same reference, the RAM cost problem could also be mitigated by performing index construction with the computers having large RAM, and using the index for read alignment on other computers.

## The performance of multiple threads

Since parallelization is also very important to read alignment tasks, we also benchmarked the speed of deBGA in multiple threads. The results on a series of datasets are in Supplementary Table 19. DeBGA were run with 1, 2, 4 and 8 threads.

It is observed from the results that, on all the tested datasets, deBGA gradually speedups with the increase of threads. It is also worthnoting that the speedup ratio decreased when many threads (e.g., 8 threads) are used. We further investigated the running of deBGA, and found that this is mainly caused by the I/O operations. That is, although the read alignment processing is parallelized, the input of reads and the output of alignment are still executed serially. Especially, there are quite a lot of serial operations to convert the alignment results recorded in specifically designed data structures to text-based SAM format, and output to files. This module costs a proportion of the running time, which is a major overhead of the program. The cost is especially obvious for the tasks of read alignment against multiple genomes, due to that the reads usually have multiple equally best alignments.

This I/O module of deBGA should be further optimized. Potential solutions include re-implement the module to let it run in parallel with other read processing threads, and further improve the executions of the module. This is also an important future work for us.

## Datasets for benchmarking

### Simulated datasets

The simulated datasets were produced by using Mason simulator (version 0.1.2) with the following command lines.

For *E.coli* strains datasets:

mason illumina -N 1000000 -n *read\_length* -i -snN -mp -sq -ll 500 -le 25 -rn 2 -hn 1 -hnN -nN *E.coli*\_*reference\_library.fasta* -o *output.fastq*

For human datasets:

mason illumina -N 1000000 -n *read\_length* -i -snN -mp -sq -ll 500 -le 25 -rn 2 -hn 2 -hnN -nN *human*\_*reference\_genome.fasta* -o *output.fastq*

### Real datasets

Fourteen datasets with SRA accession numbers ERR008613, SRR522163, SRR530851, SRR851066, ERR459978, ERR142613, SRR592258, SRR518692, ERR174324, ERR161544, ERR091571, ERR091572, ERR091573 and ERR091574, were downloaded from NCBI Short Read Archive (SRA) site, and the HiSeq X Ten dataset was downloaded from Illumina BaseSpace site, the “NA12878\_L3” sample of the “HiSeq X Ten: TruSeq Nano (4 replicates of NA12878)” project.

All the reads of ERR091571-74 were used for benchmarking. For other datasets, one million pair-end reads were randomly extracted with an in-house Python script for benchmarking.

## Statistics for the evaluation of alignment

The following statistics were used for evaluating the alignments produced by various aligners.

**The number of aligned reads** (the “Aligned” columns of Supplementary Tables 3, 5, 6, 8-11 and 14): the number of the reads which are aligned by at least one alignment of the corresponding aligner.

**The number of correctly aligned simulated reads** (the “Correctly aligned” columns of Supplementary Tables 3, 5, 8, 9, 11 and 13): the number of the reads which have a correct alignment generated by the corresponding aligner. Here, a correct alignment indicates an alignment which is within 15 bp around the grand truth genome coordinate given by Mason simulator, and the alignment is in proper direction.

**The number of simulated reads correctly aligned by the primary alignment** (the “Correctly aligned by primary alignment” columns of Supplementary Tables 3, 5, 8, 9, 11 and 13): the number of the reads whose primary alignment is a correct alignment, i.e., it is within 15 bp around the grand truth genome coordinate given by Mason simulator, and in proper direction.

**The number of aligned real reads aligned to the correct strain** (the “Aligned to the correct strain” column of Supplementary Table 6 and the “Aligned to the correct genome” column of Supplementary Table 10): the number of the reads which are aligned to the correct strain (*E. coli* K-12 MG1655) in proper direction by at least one alignment of the corresponding aligner.

**The time of read alignment** (the “Alignment Time” columns of Supplementary Tables3, 5, 6, 8-11 and 14): the wall clock time from start of read alignment to the completion of SAM output. For BWA, the execution time includes the execution time of finding the suffix array (SA) coordinates for the two ends of the reads, and the execution time of generating alignments in SAM format. For GEM, the execution time includes the execution time of the main alignment module, and the execution time of generating alignments in SAM format. The time is recorded in seconds.

**The time of loading genome index** (the “Index Loading Time” columns of Supplementary Tables 3, 5, 6, 8-11 and 14): the wall clock time of loading genome index into the RAM. For aligners with multiple components, i.e., BWA and GEM, and time of loading index in various components are summed up. The time is recorded in seconds.

## Command lines of aligners used for comparison

### Benchmarks for deBGA (version 0.1.1)

Genome indexing:

deBGA index -k *kmer\_size reference.fasta index\_route*

Read alignment:

deBGA aln -k *kmer\_size* -s *num\_of\_iterations* -i *seed\_interval* -o *num\_of\_output* -x *num\_of\_anchor\_alignment* *index\_route* *read1.fastq read2.fastq output.sam*

Read alignment with the *--cl* option:

deBGA aln -k *kmer\_size* -s *num\_of\_iterations* -i *seed\_interval* --cl *adjusted\_edit\_distance\_threshold* -o *num\_of\_output* -x *num\_of\_anchor\_alignment* *index\_route* *read1.fastq read2.fastq output.sam*

Read alignment with the *--local* option:

deBGA aln -k *kmer\_size* -s *num\_of\_iterationss* -i *seed\_interval* --local -o *num\_of\_output* -x *num\_of\_anchor\_alignment index\_route* *read1.fastq read2.fastq output.sam*

### Benchmarks for Bowtie2 (version 2.2.4)

Read alignment with the *very fast* configuration:

Bowtie2 –very-fast –x *index* -1 *read1.fastq read2.fastq* –S *output.sam*

Read alignment with the *fast* configuration:

Bowtie2 -fast –x *index* -1 *read1.fastq read2.fastq* –S *output.sam*

Read alignment with the *sensitive* configuration:

Bowtie2 -sensitive –x *index* -1 *read1.fastq read2.fastq* –S *output.sam*

Read alignment with the *very sensitive* configuration:

Bowtie2 –very-sensitive –x *index* -1 *read1.fastq read2.fastq* –S *output.sam*

### Benchmarks for BWA-MEM (version 0.7.5)

Read alignment with the default configuration:

bwa mem *index* –f *output.sam read1.fastq read2.fastq*

Read alignment with various -r configurations:

bwa mem –r *ratio* *index* –f *output.sam read1.fastq read2.fastq*

### Benchmarks for Seqalto (version 0.5-r123)

Read alignment with the *default* configuration:

Seqalto\_basic align *index* -1 *read1.fastq* -2 *read2.fastq* –f > *output.sam*

Read alignment with the *fast* configuration:

Seqalto\_basic align *index* -1 *read1.fastq* -2 *read2.fastq* > *output.sam*

### Benchmarks for BWA (version 0.7.5)

Read alignment with default output:

bwa aln –n *error* –f *aln1.sai index read1.fastq*

bwa aln –n *error* –f *aln2.sai index read2.fastq*

bwa sampe –f *output.sam index aln1.sai aln2.sai read1.fastq read2.fastq*

Read alignment with upto 1000 outputs:

bwa aln –n *error* –f *aln1.sai index read1.fastq*

bwa aln –n *error* –f *aln2.sai index read2.fastq*

bwa sampe –n 1000 –f *output.sam index aln1.sai aln2.sai read1.fastq read2.fastq*

### Benchmarks for GEM (version core\_i3-20121106-022124)

Read alignment with the *e =4%, m=4%* configuration:

gem-mapper –I *index* -1 *read1.fastq* -2 *read2.fastq* –o *output* –q *offset-33* –m 0.04 –e 0.04 –s 0 –p –E 0.3

gem-2-sam –I *index* –i *output* –o *output.sam* –c

Read alignment with the *e =6%, m=6%* configuration:

gem-mapper –I *index* -1 *read1.fastq* -2 *read2.fastq* –o *output* –q *offset-33* –m 0.06 –e 0.06 –s 0 –p –E 0.3

gem-2-sam –I *index* –i *output* –o *output.sam* –c

Read alignment with the *e =8%, m=8%* configuration:

gem-mapper –I *index* -1 *read1.fastq* -2 *read2.fastq* –o *output* –q *offset-33* –m 0.08 –e 0.08 –s 0 –p –E 0.3

gem-2-sam –I *index* –i *output* –o *output.sam* –c

### Benchmarks for STAR (version 2.5.0a)

Read alignment with default configuration:

STAR --runThreadN *1* --genomeDir *Index* --outFileNamePrefix *output* --readFilesIn *read1.fastq**read2.fastq*

### Benchmarks for GenomeMapper (version 0.4.4)

Read alignment:

genomemapper –I *reference.fa* -q *read.fastq* -M *4* -G *2* -E *4* >*output*

### Variant calling with GATK HaplotypeCaller (version 3.3.0)

Preparation of BAM files with Samtools (version 1.2):

samtools view -b –o *output.bam input.sam*

samtools sort *input.bam output*

samtools index –b *input.bam*

Read processing with Picard (version 1.119):

java -jar AddOrReplaceReadGroups.jar I=input.bam O=output.bam LB=Wessim PL=illumina PU=Wessim1 SM=NA12878

Variant calling with GATK HaplotypeCaller:

java -jar GenomeAnalysisTK.jar -R *reference.fa* -T HaplotypeCaller -I *input.bam* -nct *8* -o *output.vcf*

Filtering called variants with VQSR filter:

java -jar GenomeAnalysisTK.jar -T VariantRecalibrator -R reference.fa -input input.vcf -resource:hapmap, known=false, training=true, truth=true, prior=15.0 hapmap\_3.3.b37.sites.new.vcf -resource:omni, known=false,training=true, truth=true,prior=12.0 1000G\_omni2.5.b37.sites.new.vcf -resource:1000G, known=false, training=true, truth=true, prior=10.0 1000G\_phase1.snps.high\_confidence.hg19.vcf -resource:dbsnp, known=true, training=false, truth=false, prior=2.0 Data/All\_20151104.new.vcf -an DP -an QD -an FS -an MQ -an MQRankSum -an ReadPosRankSum -mode SNP -tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 -recalFilerecalibrate.recal -tranchesFilerecalibrate.tranches -rscriptFilerecalibrate\_plots.R

java -jar GenomeAnalysisTK.jar -T ApplyRecalibration -R reference.fa -input input.vcf -mode BOTH --ts\_filter\_level 99.0 -recalFile recalibrate.recal -tranchesFile recalibrate.tranches -o output.vcf

### Variant calling with lobSTR (version 4.0.0)

allelotype --command classify --bam *input.bam* --noise\_model *lobstr\_route/share/lobSTR/models/illumina\_v3.pcrfree* --out *output* --strinfo l*obstr\_route /hg19\_v3.0.2/lobstr\_v3.0.2\_hg19\_strinfo.tab* --index-prefix *lobstr\_route /hg19\_v3.0.2/lobstr\_v3.0.2\_hg19\_ref/lobSTR\_*

### Variant calling with Delly (version 0.7.1)

delly -t *INV* -o *OUTPUT.INV* -g *reference.fa input.bam*

delly -t *DUP* -o *OUTPUT.DUP* -g *reference.fa input.bam*

## Options of deBGA

We document the command line options of deBGA as following.

**1. Installation of deBGA**

The source code package of deBGA can be obtained from the Supplementary Software or the following link: <https://github.com/hitbc/deBGA>.

deBGA is implemented in C language, and mainly designed for running in Linux system. To install deBGA, one needs to use the makefile with the “make” command to build the executable file. The makefile is included in the source code package.

**2. Genome indexing**

*Synopsis:*

deBGA index [options] reference.fasta <index\_route>

The input reference genome (reference.fasta) should be in fasta format. For reference having multiple chromosomes or genomes, the corresponding sequences should be concatenated in advance.

*Options:*

**-k (integer, default: 22)**, the *k*-mer length of the vertices of RdBG. This is a basic parameter for building the RdBG-index. For the current version of deBGA, the range of -k parameter is restricted to 21-28 bp, considering both of the effectiveness of the seeds and memory footprint.

**3. Read alignment**

*Synopsis:*

deBGA aln [options] <index\_route> <single\_end\_read.fastq [pair\_end\_read1.fastq pair\_end\_read2.fastq]> <result\_file.sam>

The input read file(s), single\_end\_read.fastq or (pair\_end\_read1.fastq and pair\_end\_read2.fastq), should be in fastq or fasta format.

*Options:*

**-k (integer, default: 22)**, the minimum length of a valid Uni-MEM seed. For current version of deBGA, this setting should be equal to the *k*-mer length of the RdBG-index.

**-s (integer, default: 4),** the number of iterations of re-seeding. deBGA iteratively aligns a read in at most (-s + 1) iterations with various set of seeds. This parameter works combining with the minimum interval of seeding (the -i option) and the maximum allowed number of hits per seed (the -n option). That is, in the r-th iteration (r = 1 ,…, -s), deBGA tries to generate seeds at every ((-s) – r +1)\*(-i) bp along the read. If the read still cannot be successfully aligned after -s iterations, deBGA would ignore -n option to handle very repetitive reads in the (-s+1)-th iteration.

**-i (integer, default: 5),** the minimum interval of seeding. This parameter determines the density of seeds, which is related to the sensitivity and efficiency of alignment. Configuring this parameter with lower value will make deBGA generate seed more densely, which could improve the sensitivity, but at the expense of throughput.

**-n (integer, default: 300),** the maximum allowed number of hits per seed. In the first -s iterations of the alignment process, the seeds with more than -n hits would be discarded for achieving faster speed. DeBGA ignores this restriction to introduce repetitive seeds if the read still cannot be successfully aligned after -s iterations.

**-c (number, default: 0.05),** the threshold on the edit distance for early stop. In each iteration, deBGA checks the edit distance of the obtained best alignment. If the ratio ED\_best/RL < (-c), where ED\_best and RL are respectively the edit distance of the best alignment and the read length, deBGA considers that the read is confidently aligned and early-stops the alignment.

**--cl (number, default: following the setting of -c),** the adjusted threshold on the edit distance for early stop. When --cl option is set, in any given iteration, if there is at least one Uni-MEM seed available for extension, but no successful alignment is obtained, the threshold on the edit distance (-c) can be dynamically adjusted to the value of --cl in next iterations. This is a heuristic may accelerate the alignment of divergent reads, e.g., reads having many low quality bases. If --cl is not set, there will be no change on the -c option during the process.

**--local (default: not set),** the local alignment option for confident alignment. When --local option is set, in any given iteration, if there is at least one Uni-MEM seed available for extension, but no successful alignment is obtained, deBGA perform local alignment instead of end-to-end alignment in following iterations. The best obtained local alignment will be output as the result. It is also worthnoting that the --cl option and --local option should not be simultaneously set.

**--local-match (integer, default: 1),** the score for a matched base in the local alignment. This option will take effect only if --local option is set.

**--local-mismatch (integer, default: 4),** the penalty for a mismatched base in the local alignment. This option will take effect only if --local option is set.

**--local-gap-open (integer, default: 6),** the penalty for a gap open in the local alignment. This option will take effect only if --local option is set.

**--local-gap-extension (integer, default: 1),** the penalty for gap extension in the local alignment. This option will take effect only if --local option is set.

**-u (integer, default: 1000),** the upper limit of insert size. For a pair-end read, deBGA pairs the alignments of the two ends according to the upper (-u option) and lower (-f option) limits of the insert size. deBGA will consider it as a suitable pair-end alignment only if the inferred insert size is within the range [-f, -u].

**-f (integer, default: 50),** the lower limit of insert size. For a pair-end read, deBGA pairs the alignments of the two ends according to the upper (-u option) and lower (-f option) limits of the insert size. deBGA will consider it as a suitable pair-end alignment only if the inferred insert size is within the range [-f, -u].

**-o (integer, default: 20),** the maximum number of alignment output. deBGA outputs at most -o alignments for the read. This is except for the pair-end reads which are handled with the anchoring alignment strategy. For those reads, the number of outputs is determined by the -x option.

**-x (integer, default: 150),** the maximum number of alignment output for anchoring alignment. For the pair-end reads aligned with the anchoring alignment strategy, deBGA will output at most -x alignments.

**-l (integer, default: 512),** the maximum allowed read length. For the current version of deBGA, reads shorter than -l bp will be normally processed, and for reads longer than -l bp, only the first -l bp will be aligned, and the other parts will be trimmed. Set -l option with a larger number may slightly increase the memory footprint. For most nowadays next generation sequencing reads, e.g., reads from Illumina platforms, the default setting is long enough to work without the trimming. Moreover, the current version of deBGA can support reads upto 4096 bp (setting -l to 4096).

**-e (integer, default: 100),** the budget for single-end alignment. In single-end read alignment, deBGA sets a budget on the computation resource in advance for balancing the efficiency and the sensitivity. More precisely, in the extension phase, deBGA subsequently extend the candidate seeds in order of their coverage lengths, until more than -e extension operations have been totally executed after handling some of the seeds, or all the seeds are extended.

**-p (integer, default: 1),** the number of threads. The current version of deBGA supports upto 32 threads in read alignment.

**--stdout (default: not set),** output alignments by stdout. This option will let deBGA directly output alignments by stdout instead of user defined file.

1. These authors contributed equally to this work. [↑](#footnote-ref-1)
2. To whom correspondence should be addressed. [↑](#footnote-ref-2)