# SUPPLEMENT MATERIALS

This document provides the implementation details of CompMap, the experimental results on test data sets, the guidelines for parameter setting, and the effect of BWA read mapping report.

## 1. Implementation Details of CompMap

The implementation details of CompMap are provided as follows using pseudo codes. The main procedure of the program is outlined in Algorithm 1. The corresponding sub procedures of *k*mer indexing, reference-based sequence compression, and recovery of short read mapped positions in the original database are provided in Algorithms 2-5.

**Algorithm 1**: The main procedure of CompMap

**Input:** a NGS short read file ***F***, a sequence database ***D***, a set of *k*mer prefixes ***P***, the length of *k*mers *k*, the mismatch tolerance rate *e*, and the valid repeat length *L*.

**Output:** the mapping results of ***F*** on ***D*** in SAM format.

```
BEGIN
    Select one or multiple sequences from $D$ based on heuristics, e.g., length or similarity, to form a reference $R$;
    Build an index table INDEX of the $k$mers in $R$ with predefined prefixes given in $P$ (see Algorithm 2);
    Concatenate the non-reference sequences $D$-$R$ to form $M$;
    Locally align $M$ to $R$ around the $k$mers present in INDEX, and then remove the repeats from $M$ (see Algorithm 3);
    Concatenate $R$ and the remaining segments in $M$ to form a representation sequence $R'$ of $D$;
    Map $F$ to $R'$ using some standard read mapping tool, e.g., BWA, Bowtie 2, or Novoalign, and then recover the mapped positions of the short reads in $D$ (see Algorithm 5).
END
```

**Algorithm 2:** The procedure of indexing the *k*mers in the reference

**Input:** the reference ***R***, the set of *k*mer prefixes ***P***, and the length of *k*mers *k*;

**Output:** an index table **INDEX** storing the positions of *k*mers in ***R*** with predefined prefixes.

```
BEGIN
For $i$=1 to $|\boldsymbol{R}|$ do
    If $R_iR_{i+1} \in \boldsymbol{P}$ then
        key= $Hashfunc\ (R_iR_{i+1},\ldots, R_{i+k})$;
        INDEX<key> = INDEX<key> $\cup$ $i$;
    End If
End For
END
```

Here, the prefix set ***P*** could contain any dimers. **INDEX** is a hash table of numeric keys calculated by a hash function. The value associated with each key, i.e., **INDEX**<key>, is a set of occurrence positions of a *k*mer in ***R***. The hash function $Hashfunc(R_iR_{i+1},\ldots, R_{i+k})$ converts a *k*mer '$R_iR_{i+1},\ldots, R_{i+k}$' to a binary number with 'A'=00, 'C'=01, 'G'=10, and 'T'=11 (The other rare symbols in DNA sequence like 'W', 'M', 'N', etc. could be converted randomly with little effect). For example, *Hashfunc*('CGATTTAA') = 0110001111110000 or 25,584 in decimal.

**Algorithm 3**: Reference-based sequence compression

**Input:** the reference ***R***, the concatenation of non-reference sequences ***M***, the set of *k*mer prefixes ***P***, the length of *k*mers *k*, the index table **INDEX**, and the valid repeat length *L*.

**Output:** a non-redundant representation sequence ***R'*** of ***D*** and a log file ***S***.

```
BEGIN
For $i$=1 to $|M|$ do
    If $M_iM_{i+1} \in P$ and '$M_iM_{i+1}\ldots M_{i+k}$' is not located in a junction of $M$ then
        key= $Hashfunc$ ($M_iM_{i+1}\ldots M_{i+k}$);
        For each $t$ in INDEX<key> do
            Locally align $M$ to $R$ at positions $i$ and $t$, respectively (see Algorithm 4);
            If the aligned length $l$ >=$L$ then
                Write a three-tuple {$i$,$t$,$l$ } to a log file $S$;
                $i$=$i$+$l$;
            Else
                $i$=$i$+$L$;
            End If
        End For
    End If
End For
Remove all aligned repeats recorded in $S$ from $M$;
Sequentially concatenate $R$ and all remaining segments in $M$ to form $R'$;
Record the positions of all remaining segments in $M$ to $S$ using three-tuples like that of
 aligned repeats;
END
```

**Algorithm 4**: Local alignment between ***M*** and ***R***

**Input:** the reference ***R***, the concatenation of non-reference sequences ***M***, the starting position *t* at ***R***, the starting position *i* at ***M***, the size of prospecting window *N*, and the mismatch tolerance rate *e*;

**Output:** the aligned length *l*.

```
BEGIN
$l$=0;
$E$=0; //total number of mismatches
While $i+l<=|M|$ do    //match forward within the length of $M$
    If $R_{t+l} == M_{i+l}$ then
        $l=l+1$;
    Else
        Count the number of mismatches $N_e$ in the following $N$ bases of $R_{t+l}$ and $M_{i+l}$.
        If    $N_e >= N/2$ or $E+N_e > l*e$ then
               Break;
        End If
        $E=E+N_e$;
        $l=l+N$ ;
    End If
End While
$l$=0;
$E$=0;
While $i-l >0$ do //match backward
    If $R_{t-l} == M_{i-l}$ then
        $l=l+1$;
    Else
        Count the number of mismatches $N_e$ in the previous $N$ bases of $R_{t-l}$ and $M_{i-l}$.
        If    $N_e >= N/2$ or $E+N_e > l*e$ then
               Break;
        End If
        $E=E+N_e$ ;
        $l=l+N$ ;
    End If
End While
END
```

The mismatches include 1-bp substitutions, insertions and deletions.

**Algorithm 5:** Compressive short read mapping

**Input:** the NGS short read file ***F***, the non-redundant representation sequence ***R’***, and the log file ***S***;

**Output:** the mapped positions of the short reads in ***D***.

```
BEGIN
Map F to R’ using some standard read mapping tool like BWA, Bowtie 2, or Novoalign;
For each mapped read in F do
    Obtain the mapped positions $A=\{a_1,a_2,\dots\}$ of the short read in R’;
    $A_D=\{\}$; //the set of mapped positions in D
    For $i=1$ to $|A|$ do
        $A_D= A_D \cup a_i$;
        For $j=1$ to $|S|$ do //$|S|$ denotes the number of three-tuple recorded in $S$;
            Retrieve the $j$-th three-tuple $\{p_1,p_2,l\}$ recorded in $S$;
            If $p_2 \le a_i \le p_2+l$ then
                $A_D= A_D \cup ( p_1+a_i-p_2)$
            End If
        End For
    End For
    Output $A_D$ to the target file in SAM format.
End For
END
```

## 2. Experimental Results on Heterologous Data

We conducted two experiments using different data sets on a cluster running 64-bit Red Hat 4.4.4-13 with 32-core 3.1GHz Intel(R) Xeon(R) CPU E31220. CompMap was run using parameter setting of *k*=10, *k*mers prefixes={'CG','AT'}, *e*=0.05, *N*=10 and *L*=1000. BWA was used as the read mapping tool with commands: 'bwa index ref.fa' and 'bwa mem –a ref.fa reads.fq > aln-se.sam', where the option '-a' enables BWA to generate multi-mapped results. We also run BWA on the raw input as a comparison.

Three heterologous NGS data sets namely SRR031601, SRR032505, and SRR032501 derived from three bacteria agents namely *Y. kristensenii*, *Y. ruckeri*, and *Y. rohdei*, respectively, were mapped to a database (589 MB) containing 170 complete chromosomes of eight bacterial agents of bioterrorism identified by the CDC, i.e., *B. anthracis*, *B. mallei*, *B. pseudomallei*, *Brucella sp.*, *C. botulinum*, *E. coli O157:H7*, *F. tularensis*, and *Y. pestis*. The data sets and database were collected from (Francis et al. 2013). The mapping results of BWA and CompMap are shown in Table S1. The database was compressed to 304 MB by CompMap in 86 seconds using 620 MB memory. Seventeen sequences selected to form the reference are listed in Table S2.

**Table S1. The mapping results of BWA and CompMap on the heterologous sequences**

| | **SRR031601** | **SRR032505** | **SRR032501** |
|---|---:|---:|---:|
| **Num of reads** | 1,374,452 | 299,889 | 199,435 |
| **Read length** | 65~310 bp | 50~195 bp | 53~198 bp |
| **File size** | 404 MB | 88 MB | 58.3 MB |
| **Platform** | Roche 454 GS 20 | Roche 454 GS 20 | Roche 454 GS 20 |
| **Source genome** | Y. kristensenii ATCC 33638 | Y. ruckeri ATCC 29473 | Y. rohdei ATCC 43380 |
| | **BWA** | | |
| **Num of un-mapped reads** | 642,474 | 235,921 | 117,887 |
| **Mapped percentage** | 53.26 % | 21.33% | 40.89% |
| **Mapped positions (A)** | 4,392,698 | 1,043,447 | 701,707 |
| **Running time** | 1763s | 1278s | 1217s |
| **Max memory used** | 1,326 MB | 1,326 MB | 1,326 MB |
| | **CompMap** | | |
| **Num of un-mapped reads** | 643,412 | 237,014 | 118,137 |
| **Mapped percentage** | 53.19% | 20.97% | 40.76% |
| **Num of reads mapped on junctions** | 5,210 | 681 | 372 |
| **Mapped positions (B)** | 4,681,827 | 1,129,698 | 735,579 |
| **A∩B*** | 4,190,167 | 986,332 | 663,982 |
| **(A∩B)/A** | 95.39% | 94.53% | 94.62% |
| **Running time** | 1,084s | 698s | 656s |
| **Max memory used** | 673 MB | 673 MB | 673 MB |

* A∩B indicates the number of consistent mapping positions found by BWA and CompMap. Two mapping positions of a read are considered consistent if their distance < 8bp.

All data used in this experiment are available at http://csse.szu.edu.cn/staff/zhuzx/CompMap/

**Table S2. The 17 selected sequences to form the reference**

|  | **Sequence Information** |
|---|---|
| 1 | gi\|209395693\|ref\|NC_011353.1\|Escherichia coli O157:H7 str. EC4115, complete genome |
| 2 | gi\|218901206\|ref\|NC_011773.1\| Bacillus cereus AH820, complete genome |
| 3 | gi\|51594359\|ref\|NC_006155.1\| Yersinia pseudotuberculosis IP 32953, complete genome |
| 4 | gi\|226947222\|ref\|NC_012563.1\| Clostridium botulinum A2 str. Kyoto, complete genome |
| 5 | gi\|76808520\|ref\|NC_007434.1\| Burkholderia pseudomallei 1710b chromosome I, complete sequence |
| 6 | gi\|225851546\|ref\|NC_012441.1\| Brucella melitensis ATCC 23457 chromosome I, complete genome |
| 7 | gi\|163844199\|ref\|NC_010167.1\| Brucella suis ATCC 23445 chromosome II, complete genome |
| 8 | gi\|76817237\|ref\|NC_007435.1\| Burkholderia pseudomallei 1710b chromosome II, complete sequence |
| 9 | gi\|17988344\|ref\|NC_003318.1\| Brucella melitensis 16M chromosome II, complete sequence |
| 10 | gi\|153946813\|ref\|NC_009708.1\|Yersinia pseudotuberculosis IP 31758 chromosome, complete genome |
| 11 | gi\|22123922\|ref\|NC_004088.1\|Yersinia pestis KIM10+ chromosome, complete genome |
| 12 | gi\|170731356\|ref\|NC_010508.1\|Burkholderia cenocepacia MC0-3 chromosome 1, complete sequence |
| 13 | gi\|110798562\|ref\|NC_008261.1\|Clostridium perfringens ATCC 13124 chromosome, complete genome |
| 14 | gi\|126445587\|ref\|NC_009079.1\|Burkholderia mallei NCTC 10247 chromosome II, complete sequence |
| 15 | gi\|77358719\|ref\|NC_006349.2\|Burkholderia mallei ATCC 23344 chromosome 2, complete sequence |
| 16 | gi\|25010075\|ref\|NC_004368.1\|Streptococcus agalactiae NEM316, complete genome |
| 17 | gi\|89255449\|ref\|NC_007880.1\|Francisella tularensis subsp. holarctica LVS chromosome, complete genome |

The top nine sequences represent the majority of the database for each of them shares 50%+ similarity with 8~20 non-reference sequences, while the last eight reference sequences represent the minority of the database for each of them shares 50%+ similarity with only two non-reference sequences.

## 3. Experimental Results on Homogeneous Data

In this experiment, CompMap and BWA use the same parameter setting as the previous experiment on heterologous data sets. Three homogeneous *E. coli* NGS files namely SRR1063349, ERR385912, and ERR231645 in different size scales were mapped to a sequence database (437 MB) consisting of up to 5,338 genomes or plasmids of different E.coli strains. One randomly selected sequence (‘gi|110640213|ref|NC_008253.1| Escherichia coli 536, complete genome’) was used as the reference to compress the database. The database was compressed to 152 MB by CompMap in 40 seconds using 720 MB memory. The mapping results of BWA and CompMap on these homogeneous data sets are shown in Table S3.

**Table S3. The mapping results of BWA and CompMap on the homogeneous sequences**

| | SRR1063349 | ERR385912 | ERR231645 |
|---|---|---|---|
| **Num of reads** | 66,629 | 2,728,935 | 6,344,039 |
| **Read length** | 202 bp | 51 bp | 51 bp |
| **File size** | 30.2 MB | 640.7 MB | 1,447.3 MB |
| **Platform** | Illumina HiSeq 2000 | Illumina HiSeq 2000 | Illumina HiSeq 2000 |
| **Source genome** | E. coli K02 | E. coli K-12 MG1655 | E. coli BW2952 |
| **BWA** | | | |
| **Num of un-mapped reads** | 23,147 | 17,077 | 65,793 |
| **Mapped percentage** | 65.26 % | 99.37 % | 98.96% |
| **Mapped positions (A)** | 3,424,991 | 170,399,130 | 331,042,307 |
| **Running time** | 1,210s | 5,438s | 12,185s |
| **Max memory used** | 1,393 MB | 2,816 MB | 2,432 MB |
| **CompMap** | | | |
| **Num of un-mapped reads** | 23,148 | 18,791 | 76,467 |
| **Mapped percentage** | 65.25% | 99.31% | 98.79% |
| **Num of reads mapped on junctions** | 9,282 | 121,875 | 207,878 |
| **Mapped positions (B)** | 6,034,470 | 186,318,744 | 378,342,097 |
| **A∩B*** | 3,152,773 | 154,102,317 | 287,903,520 |
| **(A∩B)/A** | 92.05% | 90.44% | 86.97% |
| **Running time** | 354s | 1,845s | 3,603s |
| **Max memory used** | 720 MB | 858 MB | 858 MB |

* A∩B indicates the number of consistent mapping positions found by BWA and CompMap. Two mapping positions of a read are considered consistent if their distance < 8bp.

All data used in this experiment are available at http://csse.szu.edu.cn/staff/zhuzx/CompMap/

## 4. The Effects and Setting of the Parameters

In this part, we discuss the effects of the following parameters and try to provide guidelines for setting these parameters.

- the mismatch tolerance rate: *e*
- the size of prospecting window in local alignment: *N*
- the minimum length of a valid repeat: *L*
- the *k*mer prefixes
- the *k*mer length: *k*

### 4.1 The effects and setting of *e*

The mismatch tolerance rate *e* used in local alignment controls the balance between the compression rate and mapping accuracy. Larger *e* leads to higher compression rate and more time and space saving in short read mapping, yet the mapping accuracy is lower. We conducted experiments with *e*=0.01, 0.03, and 0.05 on two representative data sets, i.e., SRR032501 from heterologous data and SRR1063349 from homogeneous data. The other parameters were set consistently with the previous experiments in Sections 2 and 3. The results are reported in Tables S4 and S5. It is shown that as *e* increases from 0.01 to 0.05, the compression rate of the sequences increases from 21.3% to 65.2%. CompMap attains less time and/or space saving if the compression rate is lower. It is unsurprising that the mapping consistency between BWA and CompMap reduces as *e* increases, because more lossy compression is performed. Setting *e* in 0.03~0.05 would be a safe choice to obtain compromise of mapping accuracy and time/space saving. If lower mapping accuracy is tolerable, more time and space saving is achievable by increasing *e*.

**Table S4. The mapping results of BWA and CompMap with different *e* on the heterologous data SRR032501 (Y. rohdei ATCC 43380, 58.3 MB, read length=53~198 bp)**

| | No compression | *e*=0.01 | *e*=0.03 | *e*=0.05 |
|---|---:|---:|---:|---:|
| | (BWA) | (CompMap) | (CompMap) | (CompMap) |
| Reference size | 589 MB | 420 MB | 329 MB | 304 MB |
| Compression rate | 0% | 28.7% | 44.1% | 48.4% |
| Num of un-mapped reads | 117,887 | 117,937 | 118,055 | 118,137 |
| Mapped percentage | 40.89% | 40.86% | 40.81% | 40.76% |
| Num of reads mapped on junctions | 0 | 402 | 399 | 372 |
| Mapped positions | 701,704 (A) | 714,959 (B) | 726,947 (B) | 735,579 (B) |
| A∩B* | - | 680,157 | 665,967 | 663,982 |
| (A∩B)/A | - | 96.93% | 94.91% | 94.62% |
| Running time | 1217s | 840s | 682s | 656s |
| Max memory used | 1326 MB | 673 MB | 673 MB | 673 MB |

* A∩B indicates the number of consistent mapping positions found by BWA and CompMap. Two mapping positions of a read are considered consistent if their distance < 8bp.

**Table S5. The mapping results of BWA and CompMap with different *e* on the homogeneous data SRR1063349 (E. coli K02, 30.2 MB, read length=202 bp)**

| | No compression<br>(BWA) | *e*=0.01<br>(CompMap) | *e*=0.03<br>(CompMap) | *e*=0.05<br>(CompMap) |
|---|---|---|---|---|
| Reference size | 437 MB | 344 MB | 190 MB | 152 MB |
| Compression rate | 0% | 21.3% | 56.5% | 65.2% |
| Num of un-mapped reads | 23,116 | 23,149 | 23,154 | 23,154 |
| Mapped percentage | 65.31% | 65.26% | 65.25% | 65.25% |
| Num of reads mapped on junctions | 0 | 9,131 | 9,773 | 9,282 |
| Mapped positions | 3,424,965(A) | 4,276,653(B) | 5,599,187(B) | 6,034,470(B) |
| A∩B* | - | 3,246,643 | 3,142,155 | 3,152,773 |
| (A∩B)/A | - | 94.79% | 91.74% | 92.05% |
| Running time | 1,210s | 759s | 425s | 354s |
| Max memory used | 1,393 MB | 720 MB | 720 MB | 720 MB |

* A∩B indicates the number of consistent mapping positions found by BWA and CompMap. Two mapping positions of a read are considered consistent if their distance < 8bp.

## 4.2 The effects and setting of *N*

The size of prospecting window *N* used in local alignment decides how many bases the alignment should search ahead to estimate the number of mismatches. Experiments were conducted on two representative data sets, i.e., SRR032501 and SRR1063349, with *N*=5, 10, and 20 to show the effects of *N*. The other parameters were set consistently with the previous experiments in Sections 2 and 3. The results reported in Tables S6 and S7 show that the magnitude of *N* has little influence on the alignment accuracy. CompMap obtains similar results with *N*=5, 10, and 20. According to our empirical studies, setting *N* in10~20 should be a good choice for most of the data.

**Table S6. The mapping results of BWA and CompMap with different *N* on the heterologous data SRR032501 (Y. rohdei ATCC 43380, 58.3 MB, read length=53~198 bp)**

| | No compression<br>(BWA) | *N*=5<br>(CompMap) | *N*=10<br>(CompMap) | *N*=20<br>(CompMap) |
|---|---|---|---|---|
| Reference size | 589 MB | 320 MB | 304 MB | 299 MB |
| Compression rate | 0% | 45.7% | 48.4% | 49.2% |
| Num of un-mapped reads | 117,887 | 118,082 | 118,137 | 118,166 |
| Mapped percentage | 40.89% | 40.79% | 40.76% | 40.75% |
| Num of reads mapped on junctions | 0 | 496 | 372 | 355 |
| Mapped positions | 701,704 (A) | 732,494(B) | 735,579 (B) | 744,406(B) |
| A∩B* | - | 665,438 | 663,982 | 664,648 |
| (A∩B)/A | - | 94.83% | 94.62% | 94.72% |
| Running time | 1217s | 684s | 656s | 633s |
| Max memory used | 1326 MB | 673 MB | 673 MB | 673 MB |

* A∩B indicates the number of consistent mapping positions found by BWA and CompMap. Two mapping positions of a read are considered consistent if their distance < 8bp.

**Table S7. The mapping results of BWA and CompMap with different *N* on the homogeneous data SRR1063349 (E. coli K02, 30.2 MB, read length=202 bp)**

|  | No compression (BWA) | *N*=5 (CompMap) | *N*=10 (CompMap) | *N*=20 (CompMap) |
|---|---|---|---|---|
| Reference size | 437 MB | 174 MB | 152 MB | 142 MB |
| Compression rate | 0% | 60.2% | 65.2% | 67.5% |
| Num of un-mapped reads | 23,116 | 23,153 | 23,154 | 23,151 |
| Mapped percentage | 65.31% | 65.25% | 65.25% | 65.25% |
| Num of reads mapped on junctions | 0 | 12,918 | 9,282 | 8,724 |
| Mapped positions | 3,424,965(A) | 5,710,986 (B) | 6,034,470(B) | 6,132,821(B) |
| A∩B* | - | 3,135,913 | 3,152,773 | 3,159,955 |
| (A∩B)/A | - | 91.56% | 92.05% | 92.26% |
| Running time | 1,210s | 382s | 354s | 362s |
| Max memory used | 1,393 MB | 720 MB | 720 MB | 720 MB |

* A∩B indicates the number of consistent mapping positions found by BWA and CompMap. Two mapping positions of a read are considered consistent if their distance < 8bp.

### 4.3 The effects and setting of *L*

The minimum length of a valid repeat *L* affects the number of repeats identified in local alignment. Theoretically, the smaller *L* is, the more repeats could be identified and eliminated from the sequences, resulting in more compression rate. We conducted experiments with *L*=200, 600, and 1000 on two representative data sets, i.e., SRR032501 and SRR1063349. The other parameters were set consistently with the previous experiments conducted in Sections 2 and 3. The results are reported in Tables S8 and S9. It is observed that the setting of *L* does affect the compression rate to some extent but it imposes limited effects on the mapping precision of CompMap. The mapping consistencies between BWA and CompMap are not significantly different with *L*=200, 600, and 1000. The reason for this observation could be the greedy local alignment applied in CompMap, which aims to find the longest alignment despite *L*. Hence, most identified alignments are much longer than *L* and the effect of *L* to the mapping results is weakened. Nevertheless, it is better to set *L* larger than the short read length, such that a short read can be mapped within one valid repeat and the number of mappings on junctions could be reduced. We suggest setting *L* to at least three times of the short read length.

### 4.4 The effects and setting of *k*mer prefixes and *k*

The *k*mer prefixes determinate how many *k*mers are used in the local alignment. Since ‘CG’ and ‘AT’ are the most frequently identified dimers in DNA sequences and each valid repeat normally contains hundreds of bases, it is almost sure that each valid repeat contains a *k*mer prefixed with ‘CG’ or ‘AT’. We conducted experiments on homogeneous sequences with only one *k*mer prefix ‘CG’ and a shorter *L* say 200 to test the effect of *k*mer prefixes. According to the experimental results shown in Table S10, CompMap using only one *k*mer prefix ‘CG’ still obtains satisfactory performance. Yet, we suggest using both ‘CG’ and‘AT’ for the sake of stability.

The parameter *k* decides the length of *k*mers. Longer *k*mers lead to more precisely local alignment, but less mismatches tolerance. The user can assign more prefixes and larger *k* to allow higher alignment sensitivity at the cost of more memory consumption. Following the setting of *k*mer in many short reads assembling applications, *k* is set to ~10 by default in CompMap.

**Table S8. The mapping results of BWA and CompMap with different *L* on the heterologous data SRR032501 (Y. rohdei ATCC 43380, 58.3 MB, read length=53~198 bp)**

|  | No compression<br>(BWA) | *L*=200<br>(CompMap) | *L*=600<br>(CompMap) | *L*=1,000<br>(CompMap) |
|---|---|---|---|---|
| Reference size | 589 MB | 282 MB | 295 MB | 304 MB |
| Compression rate | 0% | 52.1% | 49.9% | 48.4% |
| Num of un-mapped reads | 117,887 | 118,297 | 118,197 | 118,137 |
| Mapped percentage | 40.89% | 40.68% | 40.73% | 40.76% |
| Num of reads mapped on junctions | 0 | 446 | 462 | 372 |
| Mapped positions | 701,704 (A) | 771,842(B) | 753,059(B) | 735,579 (B) |
| A∩B* | - | 661,076 | 662,059 | 663,982 |
| (A∩B)/A | - | 94.21% | 94.35% | 94.62% |
| Running time | 1217s | 632s | 645s | 656s |
| Max memory used | 1326 MB | 673 MB | 673 MB | 673 MB |

* A∩B indicates the number of consistent mapping positions found by BWA and CompMap. Two mapping positions of a read are considered consistent if their distance < 8bp.

**Table S9. The mapping results of BWA and CompMap with different *L* on the homogeneous data SRR1063349 (E. coli K02, 30.2 MB, read length=202 bp)**

|  | No compression<br>(BWA) | *L*=200<br>(CompMap) | *L*=600<br>(CompMap) | *L*=1,000<br>(CompMap) |
|---|---|---|---|---|
| Reference size | 437 MB | 128 MB | 140 MB | 152 MB |
| Compression rate | 0% | 70.7% | 68.0% | 65.2% |
| Num of un-mapped reads | 23,116 | 23,152 | 23,122 | 23,154 |
| Mapped percentage | 65.31% | 65.25% | 65.30% | 65.25% |
| Num of reads mapped on junctions | 0 | 10,362 | 10,074 | 9,282 |
| Mapped positions | 3,424,965(A) | 6,246,030(B) | 6,129,952(B) | 6,034,470(B) |
| A∩B* | - | 3,164,716 | 3,156,169 | 3,152,773 |
| (A∩B)/A | - | 92.40% | 92.15% | 92.05% |
| Running time | 1,210s | 335s | 330s | 354s |
| Max memory used | 1,393 MB | 720 MB | 720 MB | 720 MB |

* A∩B indicates the number of consistent mapping positions found by BWA and CompMap. Two mapping positions of a read are considered consistent if their distance < 8bp.

**Table S10. The mapping results of BWA and CompMap on the homogeneous data with *k*mer prefix={'CG' }, *k*=10, *e*=0.05, *N*=10 and *L*=200**

| | SRR1063349 | ERR385912 | ERR231645 |
|---|---|---|---|
| **Num of reads** | 66,629 | 2,728,935 | 6,344,039 |
| **Read length** | 202 bp | 51 bp | 51 bp |
| **File size** | 30.2 MB | 640.7 MB | 1,447.3 MB |
| **Platform** | Illumina HiSeq 2000 | Illumina HiSeq 2000 | Illumina HiSeq 2000 |
| **Source genome** | E. coli K02 | E. coli K-12 MG1655 | E. coli BW2952 |
| **BWA** | | | |
| **Num of un-mapped reads** | 23,147 | 17,077 | 65,793 |
| **Mapped percentage** | 65.26 % | 99.37 % | 98.96% |
| **Mapped positions (A)** | 3,424,991 | 170,399,130 | 331,042,307 |
| **Running time** | 1,210s | 5,438s | 12,185s |
| **Max memory used** | 1,393 MB | 2,816 MB | 2,432 MB |
| **CompMap** | | | |
| **Num of un-mapped reads** | 23,117 | 20,922 | 83,019 |
| **Mapped percentage** | 65.30% | 99.23% | 98.69% |
| **Num of reads mapped on junctions** | 10,843 | 193,463 | 327,976 |
| **Mapped positions (B)** | 6,155,141 | 198,064,331 | 399,777,896 |
| **A∩B** | 3,143,134 | 156,091,007 | 290,760,501 |
| **(A∩B)/A** | 91.77% | 91.60% | 87.83% |
| **Running time** | 358s | 1,852s | 3,633s |
| **Max memory used** | 720 MB | 858 MB | 858 MB |

## 5. The Effect of BWA Read Mapping Report

It has been observed that CompMap has 5-10% inconsistent mapped positions when compared to BWA, partially because of lossy compression (see section 4.1). In addition, the inconstancy is also partially attribute to that BWA may not exhaustively identify read mapping locations. We demonstrate this effect by comparing BWA applied to an input reference database to when it is applied separately to the partitions of the same database. For example, for the heterologous data, we split the reference database of 170 genomes into two parts, consisting of 93 and 77 genomes, respectively. Then, SRR032501 was mapped to the original database and the two separate parts. The mapping results are summarized in Table S11, where BWA identified 11.4% more mapped positions on the two separate parts than on the original database. Similar results in Table S12 are obtained on homogenous data where BWA identifies 14.0% more mapped positions on the partition reference genomes than that on the original database.

**Table S11. The mapping results of BWA on the heterologous data SRR032501 using all 170 genomes as whole and the partition i.e., 93+77 genomes.**

| | 170 Genomes | 93 Genomes | 77 Genomes |
|---|---:|---:|---:|
| Reference size | 589 MB | 305 MB | 284 MB |
| NGS file | SRR032501 | SRR032501 | SRR032501 |
| Num of reads | 199,435 | 199,435 | 199,435 |
| Mapped positions | 701,707 (A) | 482,235 | 299,214 |
| | | 781,449 (B) | |
| (B-A)/A | 11.4% | | |

**Table S12. The mapping results of BWA on the homogeneous data SRR1063349 using all 5338 genomes as whole and the partition i.e., 1000+4338 genomes.**

| | 5338 Genomes | 1000 Genomes | 4338 Genomes |
|---|---:|---:|---:|
| Reference size | 437 MB | 148 MB | 289 MB |
| NGS file | SRR1063349 | SRR1063349 | SRR1063349 |
| Num of reads | 66,629 | 66,629 | 66,629 |
| Mapped positions | 3,424,991 (A) | 1,506,418 | 2,397,599 |
| | | 3,904,017 (B) | |
| (B-A)/A | 14.0% | | |