SUPPLEMENTARY MATERIAL

Fast and sensitive read mapping with approximate seeds and multiple backtracking

Enrico Siragusa, David Weese and Knut Reinert (2012)

S1 Multiple backtracking

Figure S1: Multiple backtracking. The suffix trie representing the text GGTAACGTT-GCGGGC. Only the dotted subtree is depicted in Figure 2.
S2 Banded edit distance verification

We propose a bit-parallel banded alignment algorithm that can be used to compute the edit distance between a pattern and a text within a parallelogram of the DP matrix. For a given text \( t \) of length \( n \) and a given pattern \( p \) of length \( m \) we consider the DP matrix which has \( m + 1 \) rows and \( n + 1 \) columns. Let a band of \( w \) consecutive diagonals be given where the left-most diagonal is the main diagonal shifted by \( c \) diagonals to the left, see Figure S2a. The algorithm diagonally slides a column vector \( D \) of \( w + 1 \) cells over the band. \( D \) is encoded by delta bit-vectors \( V_P \) and \( V_N \) of size \( w \) and a variable \( \text{score} \) that tracks the cell values of the lower band boundary (dark cells in Figure S2a). Each sliding step consist of a horizontal and a vertical step. In the horizontal step the delta vectors \( D_0, H_P \) and \( H_N \) are computed as in Myers’ algorithm, see Figure S2c and lines 12–15 in Algorithm 1. Then \( V_P \) and \( V_N \) of the next column are deduced from these delta vectors and shifted by 1 bit to the right in the vertical step, see Figure S2c and lines 16–18. In the beginning, \( D \) covers the intersection of the first column and all band diagonals plus the diagonal left of the band as shown Figure S2c. As \( D \) initially represents cells beyond the DP matrix, they have to be initialized such that they have no unintended influence on the cells within the DP matrix and such that the first DP row contains increasing values for the edit distance computation. Setting the pattern bitmasks to zero for cells beyond the DP matrix, \( V_N = 0^w \) and \( V_P = 1^{c+1}0^w-c-1 \) results in the desired initialization pattern depicted in Figure S2b.

Before the pattern bitmasks can be used to compute the next \( D \) vector, they need to be shifted by 1 bit to the right and for \( \text{pos} + c < m \) in one bitmask bit \( w \) must be set to represent the next pattern character \( p[\text{pos} + c] \). This is done in line 11. Eventually, \( \text{score} \) must be tracked properly. As long as the second last cell of \( D \) is within the DP matrix the last bit of \( D_0 \) is used to track \( \text{score} \) down the left-most band diagonal in line 20. Otherwise, the horizontal deltas of the last matrix row are used in lines 23 and 24 to update \( \text{score} \). A simple heuristic allows to stop the verification earlier and improve the average running time. Cell values along a DP diagonal are monotonically increasing from top to bottom and \( \text{score} \) can only decrease along the last matrix row. The last row contains \( n + c - m \) band cells and thus the search can be stopped if \( \text{score} > k + n + c - m \).

To extend an \( e \)-error seed to the right, Masai searches the remaining part of a read with at most \( x = k - e \) errors in the text. To this end, the above described algorithm is parametrized with \( c = x \) and \( w = 2x + 1 \). The extension of a seed to the left is done analogously using the reversed read and text.

Figure S2: (a) Band parameters. For two sequences of length \( m \) and \( n \) the band is uniquely defined by the number of consecutive diagonals \( w \) and the row \( c \) that intersects the left-most diagonal and the first column. (b) DP initialization for the edit distance computation. (c) The initial state and the two substeps of one DP recursion step are shown on the right.
Algorithm 1: bandedMyers\((t, p, k, w, c)\)

**input** : text \(t\), pattern \(p\), errors \(k\), and parameters \(w, c\)

**output** : text end positions of matches with up to \(k\) errors

1. \textbf{foreach} \(x \in \Sigma\) \textbf{do}
   // initialize pattern bitmasks
   2. \(B[x] \leftarrow 0^w\)
   3. \textbf{for} \(j \leftarrow 0\) \textbf{to} \(c - 1\) \textbf{do}
      \(B[p[j]] \leftarrow B[p[j]] \;|\; 0^{c-j-1}10^{w-c+j}\)
   4. \(\text{VP} \leftarrow 1^w, \; \text{VN} \leftarrow 1^{c+1}0^{w-c-1}\)
   // initialize bit-vectors
   5. \(\text{score} \leftarrow c\)

7. \textbf{for} \(\text{pos} \leftarrow 0\) \textbf{to} \(n - 1\) \textbf{do}
   // shift pattern bitmasks
   8. \textbf{foreach} \(x \in \Sigma\) \textbf{do}
      \(B[p[\text{pos} + c]] \leftarrow B[p[\text{pos} + c]] \;|\; 10^{w-1}\)
   9. \(X \leftarrow B[t[\text{pos}]] \;|\; \text{VN}\)
   // compute horizontal bit-vectors
   10. \(D0 \leftarrow ((\text{VP} + (X \& \text{VP})) \land \text{VP}) \;|\; X\)
   11. \(HN \leftarrow \text{VP} \& D0\)
   12. \(HP \leftarrow \text{VN} \;|\; \sim((\text{VP} \;|\; D0)\)
   13. \(X \leftarrow D0 \gg 1\)
   // compute and shift vertical bit-vectors
   14. \(\text{VN} \leftarrow X \& HP\)
   15. \(\text{VP} \leftarrow HN \;|\; \sim(X \;|\; HP)\)
   16. \textbf{if} \(\text{pos} \leq m - c\) \textbf{then}
      // scoring and output
      17. \(\text{score} \leftarrow \text{score} + 1 - ((D0 \gg (w-1)) \& 1)\)
      18. \(s = (w - 2) - (\text{pos} - (m - c + 1))\)
      19. \(\text{score} \leftarrow \text{score} + ((HP \gg s) \& 1)\)
   20. \(\text{score} \leftarrow \text{score} - ((HN \gg s) \& 1)\)
   21. \textbf{else}
   22. \(\text{report occurrence ending at pos}\)
   23. \(\text{if} \; \text{pos} \geq m - c \; \text{and} \; \text{score} \leq k \; \text{then}\)
   24. \text{report occurrence ending at pos}
S3 Read mapper parametrization

In the following we describe how we configured all read mappers considered in the manuscript.

**Masai.** Version 0.5 was used. In order to use Masai as an all-mapper we passed the argument `--all`, otherwise the argument `--any-best` is used by default. We set the maximal edit distance using the parameter `-e`. We configured the seed length with the parameter `--seed-length`; on E. coli, D. melanogaster and C. elegans we chose a seed length of 16, while on H. sapiens we chose a seed length of 33. We selected the SAM output format with `-os` and enabled CIGAR output with `-oc`.

**Bowtie 2.** Version 2.0.0-beta6 was used. We used the parameter `--end-to-end` to enforce semi-global read alignments. For the Rabema benchmark we used the parameter `-k 100`.

**BWA.** Version 0.6.1-r104 was used. For the Rabema experiment we passed the parameter `-N` to `aln` and `-n 100` to `samse`.

**Soap 2.** Version 2.1 was used.

**RazerS 3.** Version 3.1 was used. We mapped with indels using the pigeonhole filter (default) and set the error rate through the parameter `-i`, e.g. `-i 95` to map within an error rate of 5%. We selected the native or SAM output format with `-of 0` or `-of 4`.

**Hobbes.** Version 1.3 was used. We built the index using the recommended q-gram length 11. Since we focus on edit distance, we used the 16 bit bit-vector version. We enabled indels with `--indels` and set maximal edit distance using the parameter `-v`. For resource measurement we used the output without CIGAR, for analyzing the results we enabled CIGAR output using `--cigar`.

**mrFAST.** Version 2.1.0.6 was used. We set maximal edit distance using the parameter `-e`.

**SHRiMP 2.** Version 2.2.2 was used.
S4 Simulation details

We used Mason with the Illumina model to simulate reads. We chose a read length of 100 bp both for the Rabema benchmark and the variant detection experiment.

We set the haplotype SNP rate to 0.6% and the indel rate to 0.1% along with random indel sizes between 1 and 32. This was the variation rate we observed from the 1000 genomes individual NA19240.

Sequencing errors were simulated using an indel probability of 0.03% at each position and replacements were simulated with the standard model using a factor of 3. The positional replacement, insert and deletion probabilities can be seen in Figure S3. The probabilities were determined after realigning the reads using the Needleman-Wunsch algorithm with a score that slightly prefers mismatches over gaps. This explains the artifacts of the curves at the ends and in the center.

![Figure S3: Positional replacement, insert and deletion probabilities for the simulated reads used in the Rabema benchmark.](image-url)
S5 Additional results

Table S1: Runtime results for E. coli and D. melanogaster. Remarks. Hobbes was not able to map completely the E. coli dataset.

| dataset         | E. coli          | D. melanogaster |
|-----------------|------------------|-----------------|
| method          | time [min:s]     | memory [Mb]     | Rabema any-best mapped reads [%] | time [min:s]     | memory [Mb]     | Rabema any-best mapped reads [%] |
| Masai           | 0:44             | 2389            | 100.00           | 98.03                  | 99.99           | 99.99           |
| Bowtie2         | 12:11            | 28              | 99.71            | 99.32                  | 99.97           | 99.76           |
| BWA             | 10:13            | 160             | 99.73            | 97.98                  | 99.72           | 97.98           |
| Soap2           | 2:19             | 609             | 97.50            | 95.68                  | 99.39           | 95.11           |
| SHRIMP 2        | 1:13             | 2274            | 100.00           | 98.03                  | 100.00          | 98.03           |
| RazerS 3        | 1:46             | 5706            | 100.00           | 98.03                  | 100.00          | 98.03           |
| Hobbes          | 9:14             | 692             | 95.09            | 93.22                  | 99.99           | 96.62           |
| mrfAST          | 4:34             | 690             | 100.00           | 98.03                  | 100.00          | 98.03           |
| SHRIMP 2        | 41:40            | 969             | 99.82            | 99.28                  | 99.71           | 99.71           |

Table S2: Runtime results for huge H. sapiens datasets. The ERR012100 dataset consists of 60M × 100bp reads, the ERR161544 dataset of 150M × 100bp reads. Remarks. A memory footprint of 20GB can be attained by processing these datasets in blocks of 10M reads.

| dataset         | ERR012100        | ERR161544        |
|-----------------|------------------|------------------|
| method          | time [min:s]     | memory [Mb]      | Rabema any-best mapped reads [%] | time [min:s]     | memory [Mb]      | Rabema any-best mapped reads [%] |
| Masai           | 118:32           | 31217            | 99.99            | 93.84                  | 99.99           | 93.16           |
| Bowtie2         | 343:58           | 3181             | 99.44            | 96.76                  | 99.99           | 96.12           |
| BWA             | 502:29           | 4475             | 99.53            | 93.61                  | 99.99           | 92.90           |
| Soap2           | 65:45            | 5357             | 95.77            | 89.91                  | 95.02           | 88.58           |
### Table S3: Rabema benchmark results for E. coli.

| method       | all         | all-best     | any-best     | precision  | recall  |
|--------------|-------------|--------------|--------------|------------|---------|
| Masai        | 98.20       | 98.59        | 98.59        | 98.59      | 98.59   |
| Bowtie 2     | 98.02       | 98.41        | 98.41        | 98.38      | 98.23   |
| BWA          | 97.10       | 97.49        | 97.49        | 98.33      | 97.25   |
| Soap 2       | 70.23       | 70.56        | 70.56        | 98.58      | 68.66   |

### Table S4: Rabema benchmark results for D. melanogaster.

| method       | all         | all-best     | any-best     | precision  | recall  |
|--------------|-------------|--------------|--------------|------------|---------|
| Masai        | 95.01       | 95.95        | 95.95        | 95.97      | 95.97   |
| Bowtie 2     | 94.84       | 95.77        | 95.77        | 95.77      | 95.63   |
| BWA          | 93.90       | 94.83        | 94.83        | 95.73      | 94.63   |
| Soap 2       | 67.79       | 68.54        | 68.54        | 95.91      | 68.66   |

### Table S5: Rabema benchmark results for C. elegans.

| method       | all         | all-best     | any-best     | precision  | recall  |
|--------------|-------------|--------------|--------------|------------|---------|
| Masai        | 94.49       | 97.38        | 97.38        | 97.33      | 97.33   |
| Bowtie 2     | 94.23       | 97.07        | 97.07        | 97.11      | 96.92   |
| BWA          | 93.43       | 96.30        | 96.30        | 97.03      | 96.92   |
| Soap 2       | 66.90       | 69.24        | 69.24        | 97.34      | 69.34   |
### S6 Filtration results

We assessed the contribution of approximate seeds and multiple backtracking on runtime results. To this intent we performed all-mapping with Masai on each previously considered dataset, this time using either exact or approximate seeds in combination with either single or multiple backtracking. The optimal combination of seeding and backtracking that was used for runtime results is shown in bold.

Table S5 shows the results. Filtration time consists of the time spent to index the seeds (in case of multiple backtracking) and to perform backtracking. Candidates reports the number of candidate locations reported by the filter for which seed extension is subsequently performed.

Since we concentrate on filtration, we did not consider the time spent performing seed extensions and I/O, i.e. loading the reference genome and its index, loading the reads, writing the results. Such time is independent of any combination of seeding or backtracking and can be extrapolated by subtracting bold filtration times of Table S5 from respective Masai all-mappers times of Table 3 and Table S4.

On E. coli, D. melanogaster and C. elegans approximate seeds reduce the number of candidates respectively by 2.1 times, 9.9 times, and 4.3 times. Nevertheless we still prefer exact seeds since filtration dominates the total runtime. Multiple backtracking on exact seeds compared to single backtracking speeds up filtration by 2.9 times on E. coli, and 3.8 times on D. melanogaster and C. elegans. Without the contribution of multiple backtracking Masai would not be faster than RazerS3, the second fastest all-mapper.

Approximate seeds become effective on H. sapiens, where they reduce the number of candidates by 10.8 times. On H. sapiens seed extensions largely dominate the total runtime, therefore we prefer approximate seeds. Multiple backtracking on approximate seeds provides a speed-up of 3.2 times over single backtracking. The combination of the two methods makes Masai an order of magnitude faster than any other all-mapper.

| organism     | dataset     | seeding   | backtracking | filtration time | candidates  |
|--------------|-------------|-----------|--------------|-----------------|-------------|
| E. coli      | ERR022075   | exact     | single       | 3:55            | 69.17 M     |
| E. coli      | ERR022075   | exact     | multiple     | 1:20            | 69.17 M     |
| E. coli      | ERR022075   | approximate | single     | 38:42           | 33.08 M     |
| E. coli      | ERR022075   | approximate | multiple     | 9:00            | 33.08 M     |
| D. melanogaster | SRR497711 | exact     | single       | 8:15            | 1020.28 M   |
| D. melanogaster | SRR497711 | exact     | multiple     | 2:11            | 1020.28 M   |
| D. melanogaster | SRR497711 | approximate | single     | 100:18          | 102.78 M    |
| D. melanogaster | SRR497711 | approximate | multiple     | 20:48           | 102.78 M    |
| C. elegans   | SRR065390   | exact     | single       | 8:25            | 1065.70 M   |
| C. elegans   | SRR065390   | exact     | multiple     | 2:11            | 1065.70 M   |
| C. elegans   | SRR065390   | approximate | single     | 102:02          | 246.65 M    |
| C. elegans   | SRR065390   | approximate | multiple     | 21:33           | 246.65 M    |
| H. sapiens   | ERR012100   | exact     | single       | 55:54           | 294943.86 M |
| H. sapiens   | ERR012100   | exact     | multiple     | 41:52           | 294943.86 M |
| H. sapiens   | ERR012100   | approximate | single     | 165:45          | 27396.01 M  |
| H. sapiens   | ERR012100   | approximate | multiple     | 52:15           | 27396.01 M  |
S7 Performance on real data using different indexing methods

Table S7: Performance on real data using different indexing methods.

| dataset   | ERR022075 E. coli | SRR497711 D. melanogaster | SRR065390 C. elegans | ERR012100 H. sapiens |
|-----------|-------------------|---------------------------|----------------------|----------------------|
| **method** | time [min:s]      | memory [Mb]               | time [min:s]         | memory [Mb]          | time [min:s]         | memory [Mb] |
| **best-mode** |                  |                           |                      |                     |                      |             |
| SA        | 0:44              | 2389                      | 4:52                 | 3051                | 3:10                 | 2936          | 22:35       | 19711          |
| Esa       | 0:46              | 2424                      | 5:01                 | 3969                | 3:15                 | 3701          | 26:24       | 42780          |
| FM-index  | 0:52              | 2362                      | 10:17                | 2611                | 5:29                 | 2569          | 42:26       | 8772           |
| **all-mode** |                  |                           |                      |                     |                      |             |             |                |
| SA        | 1:33              | 2274                      | 7:34                 | 2936                | 10:49                | 2821          | 307:16      | 20130          |
| Esa       | 1:42              | 2310                      | 8:02                 | 3855                | 11:13                | 3586          | 297:13      | 43191          |
| FM-index  | 2:38              | 2248                      | 17:15                | 2497                | 21:12                | 2455          | 480:23      | 9156           |