MrsFAST: A cache-oblivious algorithm for short-read mapping

ARTICLE *in* NATURE METHODS · AUGUST 2010

Impact Factor: 32.07 · DOI: 10.1038/nmeth0810-576 · Source: PubMed

7 AUTHORS, INCLUDING:

Faraz Hach
Simon Fraser University
23 PUBLICATIONS **502** CITATIONS
SEE PROFILE

Fereydoun Hormozdiari
University of California, Davis
100 PUBLICATIONS **8,040** CITATIONS
SEE PROFILE

Can Alkan
Bilkent University
177 PUBLICATIONS **14,579** CITATIONS
SEE PROFILE

Cenk Sahinalp
Simon Fraser University
183 PUBLICATIONS **7,557** CITATIONS
SEE PROFILE

Published in final edited form as:
Nat Methods. 2010 August ; 7(8): 576–577. doi:10.1038/nmeth0810-576.

mrsFast: a cache-oblivious algorithm for short-read mapping

Faraz Hach[1], **Fereydoun Hormozdiari**[1], **Can Alkan**[2,3], **Farhad Hormozdiari**[1], **Inanc Birol**[1,4], **Evan E Eichler**[2,3], and **S Cenk Sahinalp**[1,2]

S Cenk Sahinalp: cenk@cs.sfu.ca

[1] School of Computing Science, Simon Fraser University, Burnaby, Canada

[2] Department of Genomics, University of Washington School of Medicine, Seattle, Washington, USA

[3] Howard Hughes Medical Institute, Seattle, Washington, USA

[4] Genome Sciences Center, British Columbia Cancer Agency, Vancouver, Canada

To the Editor

In addition to single-nucleotide variations and small insertions-deletions (indels), larger-sized structural variations (for example, insertions, deletions, inversions, segmental duplications and copy-number polymorphisms) contribute to human genetic diversity. In almost all recent structural variation discovery (SVD) studies, short reads from a donor genome have been mapped to a reference genome as a first step. The accuracy of such an SVD study is directly correlated to the accuracy of this mapping step, which also provides the main computational bottleneck of the SVD study.

Next-generation sequencing technologies provide increasingly longer reads (currently ~400 base pairs (bp) for the Roche 454 platform and 2 × 100 bp for the Illumina platform). However, even with the increased read lengths, ambiguity in read mapping remains a problem. A human genome resequencing study[1] using 36-bp reads has reported, on average, 1,628 mapping locations per read within two mismatches and indels. In our study, on a set of one million 36-bp reads from a Yoruban individual (NA18507), we observed an average of 1,486 mapping locations within two mismatches and indels and 1,411 mapping locations when we allowed only two mismatches but no indels. The mapping multiplicity only reduced to 615 locations for 50-bp reads within three mismatches, 185 locations for 75-bp reads within four mismatches and 140 locations for 100-bp reads within six mismatches (Table 1).

As structural variants are typically observed in repeat regions, it is critical to consider all possible mapping locations for each read. To address this need, recently developed mapping tools such as Maq[2], Bowtie[3], RazerS[4] and Burrows-Wheeler alignment (BWA)[5] have options to report read multiplicities, but they do not capture all possible mapping locations. Bowtie and BWA use the Ferragina-Manzini index[6] (FMI), which is designed (and works effectively) for finding exact matches. Extending the FMI to handle mismatches or indels is only achieved by heuristic generalizations. As the read length and the corresponding number

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Note: Supplementary information is available on the Nature Methods website.

of mismatches and indels to be tolerated increase, these methods deteriorate exponentially in terms of speed and/or accuracy.

We developed 'micro-read (substitutions only) fast alignment and search tool' (mrsFAST), a cache-oblivious short read mapping algorithm that rapidly finds all mapping locations of a collection of short reads from a donor genome in the reference genome within a user-specified number of mismatches through indexing both the reference genome and the short reads, and executing a simple cache-oblivious, all-to-all list comparison algorithm (Supplementary Note). We also developed mrFAST-CO, a version of mrsFAST that can handle indels and substitutions (equivalently, mrFAST-CO is a doubly indexed, cache-oblivious version of previously developed mrFAST[1], a simple 'seed and extend'–type mapping algorithm).

Like mrFAST, mrsFAST and mrFAST-CO are seed-and-extend algorithms. Such algorithms work by first placing a *k*-mer (seed) from a read by interrogating the index (in the form of a hash table for all *k*-mers and their respective loci) of the reference genome and then extending them by allowing at most a user-specified number of mismatches or indels. During the execution of the algorithms, the operating system copies the information related to the seed locations from the main memory to the much faster levels of cache memory, and the extension step is performed using the information stored in the cache. In a naive execution (in comparison to a cache-oblivious execution) of such a seed-and-extend algorithm, the seed mapping locations to be compared to the read would be streamed through the cache. As cache capacity is very limited, before such read locations can be used for another read, they will be overwritten by new mapping locations.

mrsFAST and mrFAST-CO, in contrast, establish for each possible seed sequence S, the list $L1$ of reads that include S; they also establish the list $L2$ of locations in the reference genome in which S is observed. Instead of using two nested loops to compare each element of $L1$ with every element of $L2$ requiring a total of $O(|L1| \times |L2|)$ comparisons and cache misses, our cache-oblivious algorithms partition the two lists recursively until the subproblems can fit in the cache hierarchy and compare the sublists with each other. Although the number of comparisons stays the same, the order in which they are performed mathematically guarantees that the number of cache misses are minimized asymptotically without any specific knowledge of the existing cache sizes or structure. Because all available short read mapping tools spend a substantial amount of execution time handling cache misses, the cache-obliviousness paradigm provides means to improve their performance drastically (Supplementary Note).

Given a user-specified number of mismatches or indels, mrs-FAST and mrFAST-CO mathematically guarantee to return the coordinates of all mapping locations of each read—or optionally up to a user-specified maximum multiplicity. Reporting coordinates of multiple mapping locations does not have a substantial impact on run time or memory load. Our algorithms coupled with state-of-the-art structure variation detection algorithms such as Variation Hunter[7] captured longer than insert size deletions that could not be detected by single mapping based approaches (Supplementary Note).

In comparison to mapping tools using the FMI, *q*-gram filtering (filtering based on distribution of length *q* substrings such as RazerS) and other seed-and-extend techniques, mrsFAST and mrFAST-CO are substantially faster and more accurate (Table 1). mrsFAST and mrFAST-CO source codes are available for public use through sourceforge (http://mrsfast.sourceforge.net/).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the following funding agencies for providing support: Genome British Columbia Science Opportunities Fund (to S.C.S. and I.B.), Mathematics of Information Technology and Complex Systems Accelerate Program (to F. Hach and Fe. Hormozdiari), National Sciences and Engineering Research Council Discovery Grant Program (to S.C.S.), Simon Fraser University Community Endowment Trust Fund (to S.C.S.), Howard Hughes Medical Institute (to E.E.E.), National Institutes of Health (HG004120 to E.E.E.). We thank A. Ghane and V. Kazempour for their help with the use of the OProfile tool.

References

1. Alkan C, et al. Nat Genet. 2009; 41:1061–1067. [PubMed: 19718026]
2. Li H, Ruan J, Durbin R. Genome Res. 2008; 18:1851–1858. [PubMed: 18714091]
3. Langmead B, Trapnell C, Pop M, Salzberg S. Genome Biol. 2009; 10:R25. [PubMed: 19261174]
4. Weese D, Emde AK, Rausch T, Doring A, Reinert K. Genome Res. 2009; 19:1646–1654. [PubMed: 19592482]
5. Li H, Durbin R. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
6. Ferragina P, Manzini G. Proc IEEE FOCS. 2000:390–398.
7. Hormozdiari F, Alkan C, Eichler E, Sahinalp SC. Genome Res. 2009; 19:1270–1278. [PubMed: 19447966]

Table 1

Speed and accuracy of mapping

Read length (errors):	36 bp (2 errors)			50 bp (3 errors)			75 bp (4 errors)			100 bp (6 errors)		
Algorithm	**Time (h:min)[a]**	**Reads mapped (%)[b]**	**Locations reported (millions)[c]**	**Time (h:min)[a]**	**Reads mapped (%)[b]**	**Locations reported (millions)[c]**	**Time (h:min)[a]**	**Reads mapped (%)[b]**	**Locations reported (millions)[c]**	**Time (h:min)[a]**	**Reads mapped (%)[b]**	**Locations reported (millions)[c]**
Bowtie	5:14	91.65	1,404	3:13	92.73	610	NA	NA	NA	NA	NA	NA
BWA	3:10	92.05	1,581	10:23	93.38	729	59:35	90.16	212	67:38	87.91	42
Maq	6:45	90.91	1,609	10:05	89.25	458	NA	NA	NA	NA	NA	NA
mrFAST-CO	6:12	92.18	1,486	9:21	93.39	663	11:32	90.22	193	17:54	88.55	155
mrsFAST	2:00	91.79	1,411	1:55	92.91	613	2:00	89.35	177	2:49	87.27	138
RazerS[d]	10:17	91.79	<100	12:17	92.91	<100	12:00	89.35	<100	25:10	87.27	<100
BWA[e]	0:10	92.05	<1	0:15	93.38	<1	0:25	90.16	<1	7:04	87.91	<1

We mapped one million reads of indicated read lengths and within the given number of errors, to the human reference genome HG18 build 36 by indicated algorithms. All rows (except the last two) denote the time needed to report all mapping locations. Because of its high memory requirement, we could not run RazerS for read multiplicities >100. Note that in some columns the total number of mapping locations is higher for Maq or BWA than for mrsFAST or mrFAST-CO because Maq often returns mapping locations with an error rate higher than the user-specified rate and BWA returns certain mapping locations multiple times. NA, not applicable.

[a] Time required for mapping (on a single personal computer).

[b] Percentage of the reads mapped.

[c] Total map locations reported (in millions).

[d] Maximum multiplicity, 100.

[e] Single location.