Comparison of mapping softwares for next generation sequencing data

Julien Fayolle, Jean-François Gibrat, Valentin Loux, Sophie S. Schbath

Comparison of mapping softwares for next generation sequencing data

Julien FAYOLLE, Jean-François GIBRAT, Valentin LOUX, Sophie SCHBATH
contact : prenom.nom@jouy.inra.fr

Introduction

Recent DNA sequencers, usually called "next generation", produce reads that are shorter and in much larger number than previous sequencers. New alignment programs have been developed for these new type of reads. Our study evaluates the efficiency, strong points and weaknesses of these tools.

We have identified about 40 softwares that are currently used to map onto known genomes the reads produced by next generation sequencers (NGS). Our study focuses on short reads (produced by Illumina sequencers).

We focus on 9 of the most used softwares (bwa, Novoalign, Bowtie, MOM, ProbeMatch, SOAP2, BFAST, SHRiMP, and maq) to align the simulated reads on the genome.

Methodology

We simulate two sets of reads of length 40 bp, that are drawn uniformly in a dataset. To reflect the diversity of genomic data, we use 2 kinds of datasets: the human genome (2.7G bp) and a concatenation of 900 bacterial genomes (1.7G bp). The sets contain 10M reads, close to the actual amount produced by NGS tools.

In the first set reads are simulated without errors, in the second, three mismatches are added at random positions. We monitor several indicators of the performance of each software: CPU time used, whether the read matches at its "original" position, number of match positions found for a given read, number of uniquely mapped reads.

Tools

	Output	Algorithm	Input	Multithreaded	Gap	Version
Bwa	SAM	**Burrows-Wheeler**	NT-space	yes	yes	0.5.6
Novoalign	SAM	Indexing the reference genome (proprietary source)	NT-space and colorspace	yes	no	2.06.09
Bowtie	SAM	**Burrows-Wheeler**	NT and colorspace	yes	no	0.12.5
MOM	Own	**Hash**-table on reference genome or read sequences	NT-space	yes	no	0.4
ProbeMatch	Own	**Hash**-table on reference genome	NT-space	no	no	
SOAP2	SAM-like	**Burrows-Wheeler**	NT-space	yes	yes	2.20
BFAST	SAM	**Hash**-table on reference genome	NT-space and colorspace	yes	yes	0.6.4d
SHRiMP	SAM	**Hash**-table on read sequences	NT-space and colorspace	yes	yes	2.0.1
Maq	SAM	**Hash**-table on read sequences	NT-space	no	no	0.7.1

Results

We separate reads mapped once and mapped several times. Reads mapped too often provide little information to the end user and are actually rarely considered.
A **unique read** is a read mapped at only one location.
Reads originate from a random position (drawn uniformly) in the genome. We look whether this **original position is retrieved** by the software.

Softwares' performance with 0 mismatch

				Unique reads		Non unique reads					
								Orig pos retrieved		Orig pos not retrieved	
Software	Indexing time	Mapping time	Nb mapped reads	Nb	Orig pos retrieved	Nb	Mean hits [sd]	Nb [%]	Mean rank [sd]	200 hits	Less than 200 hits
bwa	1h 28mn	48mn	9999998 100%	8739090	8330833 95.32%	1260908	30.50 44.92	1260908 100	6.71 3.47	0	0
Novoalign	23mn	10h 50mn	9999320 99.99%	8875324	8874639 99.99%	1124676	59.56 81.07	932754 82.9	15.83 33.00	191912	9
Bowtie	3h 32mn	21mn	9999950 99.99%	8874680	8874631 99.99%	1125320	59.64 81.12	932623 82.8	15.83 33.00	192696	0
SOAP2	1h 34mn	56mn									
BFAST	14mn+ 10× 1d 10h	13h 39mn	9294641 92.9%								
maq											

A **hit** (for a read) is a position in the genome where the read is mapped.
A read with a single hit is a unique read. Some reads have up to 30k hits.

Each hit of a read is written to an output file. The first hit in the output file has **rank** one, the second has rank two, etc.

CPU time is split in two : **indexing** and **mapping** times.
Indexing is done once for a given reference genome.
Mapping is done for each set of reads.

The maximum **number of hits** asked to the software is **limited to 200**.

Some softwares were not able to align 10 millions reads on the reference genomes (mostly because the memory requirements were too demanding).
These softwares do not appear in the tables.

Softwares were run on a single CPU core (even though most softwares allow multithreading).

We used a 2.3 GHz CPU (64 bits) with 16 GB of memory.

Softwares' performance with 3 mismatches

				Unique reads		Non unique reads					
								Orig pos retrieved		Orig pos not retrieved	
Software	Indexing time	Mapping time	Nb mapped reads	Nb	Orig pos retrieved	Nb	Mean hits [sd]	Nb [%]	Mean rank [sd]	200 hits	less than 200 hits
bwa	1h 28mn	3h 16mn	5781876	4790181	4566774	991695	32.37 44.15	576532 58.13	12.64 24.27	319	414528
Novoalign	23mn	4d 8h	9999949	8695303	8471634	1304697	14.46 26.9	839463 64.34	5.97 12.95	4186	461047
Bowtie	3h 32mn	3h 31mn	9999950	8495019	8494971	1504981	72.14 85.44	1189287 79.02	20.42 38.01	315692	1
SOAP2											
BFAST	14mn + 10 × 1d 10h		9999950								
maq											

Future developments

- On SOLiD reads (colorspace)
- On paired-ends and mate-paired reads
- On larger datasets (100M, 1G reads)
- On reads produced by sequencers (instead of simulated ones)

- Incorporate model on the errors produced by sequencers
- Other types of variations for the reads (gaps, mutations)
- Automate the evaluation process

References

- **[bwa]** Li H and Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.
- **[ProbeMatch]** Kim YJ, Teletia N, *et alii* (2009). ProbeMatch: rapid alignment of oligonucleotides to genome allowing both gaps and mismatches. Bioinformatics. Jun 1;25(11):1424-5.
- **[SHRiMP]** Rumble SM, Lacroute P, *et alii* (2009). SHRiMP: Accurate Mapping of Short Color-space Reads. PLoS Comput Biol 5(5).
- **[Novoalign]** Novocraft.com
- **[Bowtie]** Langmead B, Trapnell C, Pop M, Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.
- **[SOAP2]** Li R *et alii* (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, doi:10.1093/bioinformatics/btp336
- **[BFAST]** Homer N, Merriman B, Nelson SF (2009). BFAST: An alignment tool for large scale genome resequencing. PLoS ONE 4(11): e7767 doi:10.1371/journal.pone.0007767
- **[maq]** Li H, Ruan J, Durbin R (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18(11):1851-8.
- **[MOM]** Eaves H and Gao Y (2009). MOM: maximum oligonucleotide mapping. Bioinformatics 25(7):969-970; doi:10.1093/bioinformatics/btp092

MIG – Mathématique, informatique et génome
Domaine de Vilvert,
78352 Jouy-en-Josas
This work is funded as part of the ANR project CBME

Mathématique Informatique & Génome

INRA