



HAL
open science

Mapping reads on a genomic sequence: a practical comparative analysis

Sophie S. Schbath, Veronique V. Martin, Matthias Zytnicki, Julien Fayolle, Valentin Loux, Jean-François Gibrat

► **To cite this version:**

Sophie S. Schbath, Veronique V. Martin, Matthias Zytnicki, Julien Fayolle, Valentin Loux, et al.. Mapping reads on a genomic sequence: a practical comparative analysis. Kick-of meeting of PF7 RADIANT project, Jan 2013, Manchester, United Kingdom. 19 diapos. hal-02804577

HAL Id: hal-02804577

<https://hal.inrae.fr/hal-02804577>

Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mapping reads on a genomic sequence: a practical comparative analysis

Sophie Schbath, Véronique Martin, Matthias Zytnicki, Julien
Fayolle, Valentin Loux et Jean-François Gibrat

Unité Mathématique, Informatique et Génome
INRA, Jouy-en-Josas



Kick-off RADIANT meeting, Manchester, January 14, 2013

Next Generation Sequencing reads mapping

- Short read mapping is the initial step of many NGS analyses (SNPs calling, RNA-Seq, CHIP-Seq, ...)
- A lot of tools have been released between 2007 and 2012 (76 tools in the survey of Fonseca et al. (2012))
- Few complete, controlled and fully understandable benchmarks

Aim of our benchmark

Controlled benchmark, simple questions :

- Are the tools capable to systematically map a read occurring exactly (with no mismatch) in the reference genome ?
- Can they always do it for a read having as many errors as the maximum number of mismatches allowed in the alignments ?
- For reads occurring at several positions, do/can they retrieve all the occurrences or only a subset ?
- Do the reads reported as unique really occur only once along the genome ?

Evaluated mappers

Mapper	Format	Algorithm	Input	Threads	Gaps
bwa	SAM	Burrows-Wheeler	nt	yes	yes
Bowtie	SAM	Burrows-Wheeler	nt & color	yes	no
SOAP2	dedicated	Burrows-Wheeler	nt	yes	no
Novoalign	SAM	hash on ref.	nt & color	yes	yes
BFAST	SAM	hash on ref.	nt & color	yes	yes
SSAHA2	SAM	hash on ref.	nt	no	no
GASSST	SAM	hash on ref.	nt	yes	yes
PerM	SAM	hash on ref.	nt & color	no	no
MPscan	dedicated	suffix tree	nt	no	no

SAM : Simple Alignment Map

nt : Nucleotide space

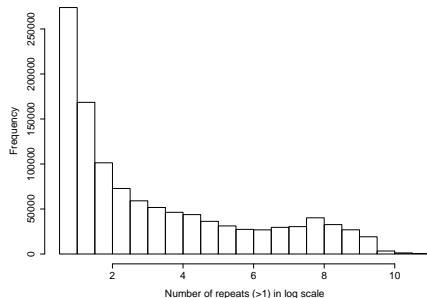
Read dataset generation : \mathcal{H}_0

Dataset

- Human genome (2.7 Gbp) as reference
- 10 millions of 40 bp reads
- Uniformly drawn from both strands

Characteristics

- 49 reads with 'N'
- 1 122 893 reads non unique
- Most frequent read : 53162 occurrences



Other read datasets

Human reference genome

- \mathcal{H}_1 , \mathcal{H}_2 and \mathcal{H}_3 : exactly 1, 2 and 3 mismatches uniformly generated within each read from \mathcal{H}_0
- $\mathcal{H}l_0$ and $\mathcal{H}l_3$: longer reads of 100 bps

Bacterial reference genome

- \mathcal{B}_0 : 10 millions of 40bps reads drawn from 904 bacterial genomes
- \mathcal{B}_3 : 3 mismatches uniformly generated within each read from \mathcal{B}_0

Results for exact mapping of \mathcal{H}_0 on human genome

Exact mapping of \mathcal{H}_0 : memory usage

Software	Memory usage (Gb)	Indexing time	Mapping time	Unmapped reads	Orig. pos. not ret.
BWA	2.18	1h 36mn	1h 13mn	49	0
Novoalign	8.12	8mn	13h 24mn	632	0
Bowtie	7.36	3h 25mn	2h 42mn	49	0
SOAP2	51.87	1h 56mn ^(‡)	56mn ^(‡)	49	3 566
BFAST	9.68	18h 01mn ^(*)	15h 02mn	726 332	20 026
SSAHA2	9.60	24mn	1d 1h	35 875	193 211
MPscan	2.67	1h 20mn		26	0
GASSST	57.93	8h 45 ^(††)		49	54
PerM	13.77	13h 05mn		115 871	4

Exact mapping of \mathcal{H}_0 : computation time

Software	Memory usage (Gb)	Indexing time	Mapping time	Unmapped reads	Orig. pos. not ret.
BWA	2.18	1h 36mn	1h 13mn	49	0
Novoalign	8.12	8mn	13h 24mn	632	0
Bowtie	7.36	3h 25mn	2h 42mn	49	0
SOAP2	51.87	1h 56mn ^(‡)	56mn ^(‡)	49	3 566
BFAST	9.68	18h 01mn ^(*)	15h 02mn	726 332	20 026
SSAHA2	9.60	24mn	1d 1h	35 875	193 211
MPscan	2.67	1h 20mn		26	0
GASSST	57.93	8h 45 ^(†‡)		49	54
PerM	13.77	13h 05mn		115 871	4

Test performed on a Intel Quad Core 2.33 GHz 16 Gb RAM, except ^(‡) four Intel Six Core 2.40 GHz 132 Gb RAM

(*) Average indexing time per spaced seed computed on 10 seeds.

(†) This time does not include the running time of the `gassst_to_sam` command.

(‡) This time is slightly over-estimated.

Exact mapping of \mathcal{H}_0 : failures

Software	Memory usage (Gb)	Indexing time	Mapping time	Unmapped reads	Orig. pos. not ret.
BWA	2.18	1h 36mn	1h 13mn	49	0
Novoalign	8.12	8mn	13h 24mn	632	0
Bowtie	7.36	3h 25mn	2h 42mn	49	0
SOAP2	51.87	1h 56mn ^(‡)	56mn ^(‡)	49	3 566
BFAST	9.68	18h 01mn ^(*)	15h 02mn	726 332	20 026
SSAHA2	9.60	24mn	1d 1h	35 875	193 211
MPscan	2.67	1h 20mn		26	0
GASSST	57.93	8h 45 ^(††)		49	54
PerM	13.77	13h 05mn		115 871	4

Exact mapping of \mathcal{H}_0 : unique matches

Software	Unmapped reads	Reads uniquely retrieved		Reads with multiple hits		
		Nb	Orig. pos. not retr.	Nb	Nb hits mean	Orig. pos. not retr.
BWA	49	8 877 061	0	1 122 890	722.81	0
Novoalign	632	8 877 107	0	1 122 261	698.63	0
Bowtie	49	8 877 061	0	1 122 890	722.81	0
SOAP2	49	8 877 061	0	1 122 890	653.26	3566
BFAST	726 332	8 840 305	9 193	433 363	2.96	10 833
SSAHA2	35 875	8 886 204	9 847	1 077 921	79.52	183 364
MPscan	26	8 877 081	0	1 122 893	722.81	0
GASSST	49	8 877 061	0	1 122 890	722.47	54
PerM	115 871	8 877 068	3	1 007 061	126.42	1
Reference		8 877 107		1 122 893	722.81	

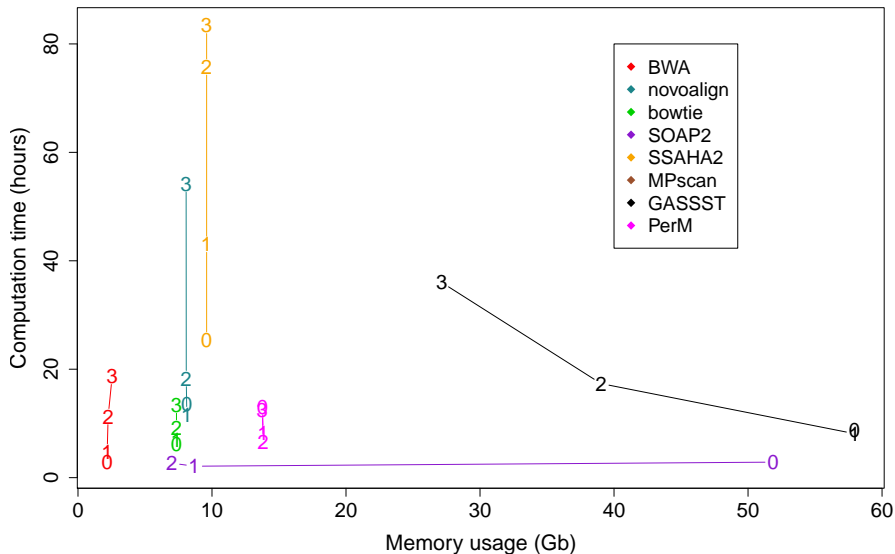
Exact mapping of \mathcal{H}_0 : multiple matches

Software	Unmapped reads	Reads uniquely retrieved		Reads with multiple hits		
		Nb	Orig. pos. not retr.	Nb	Nb hits mean	Orig. pos. not retr.
BWA	49	8 877 061	0	1 122 890	722.81	0
Novoalign	632	8 877 107	0	1 122 261	698.63	0
Bowtie	49	8 877 061	0	1 122 890	722.81	0
SOAP2	49	8 877 061	0	1 122 890	653.26	3566
BFAST	726 332	8 840 305	9 193	433 363	2.96	10 833
SSAHA2	35 875	8 886 204	9 847	1 077 921	79.52	183 364
MPscan	26	8 877 081	0	1 122 893	722.81	0
GASSST	49	8 877 061	0	1 122 890	722.47	54
PerM	115 871	8 877 068	3	1 007 061	126.42	1
Reference		8 877 107		1 122 893	722.81	

Trends as increasing the number of mismatches

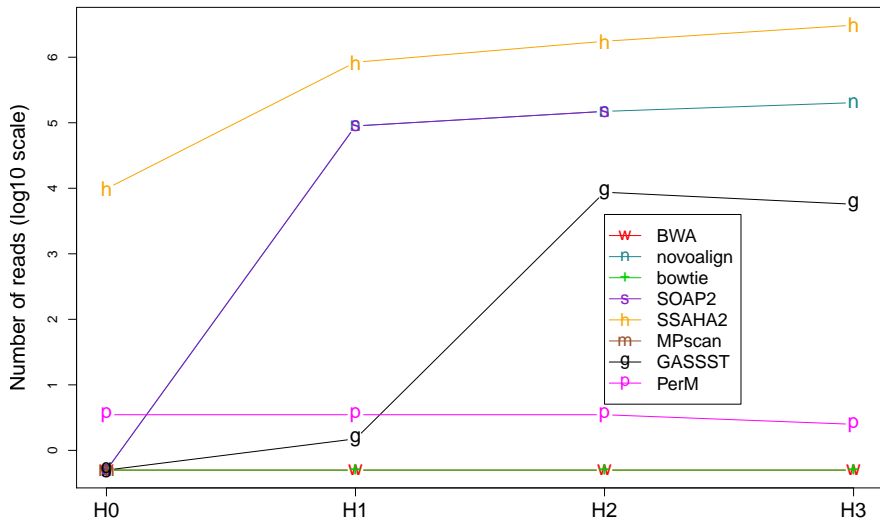
- exact mapping of \mathcal{H}_0 on human genome
- mapping with up to 1 mismatch of \mathcal{H}_1 on human genome
- mapping with up to 2 mismatches of \mathcal{H}_2 on human genome
- mapping with up to 3 mismatches of \mathcal{H}_3 on human genome

Memory usage versus Computation time



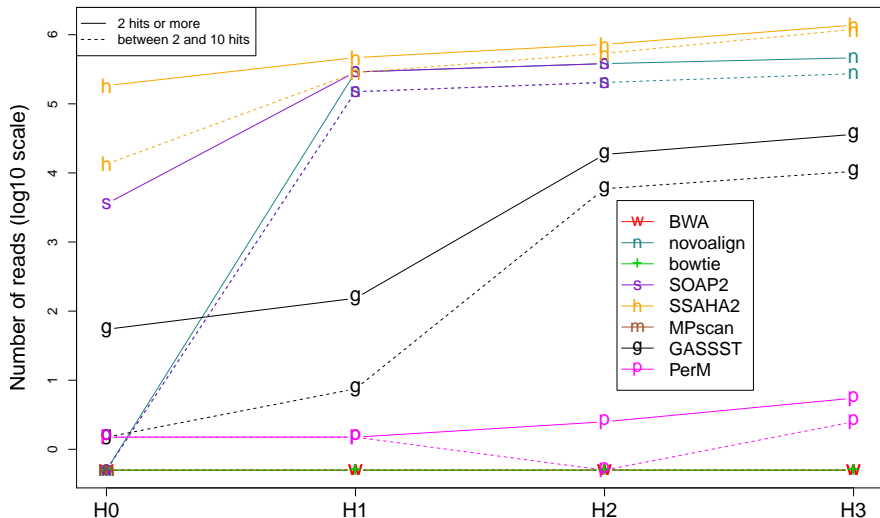
Failures for unique matches

Reads uniquely retrieved but not at their original position



Failures for multiple matches

Reads with multiple hits but not at their original position



Conclusions and future work

Benchmark

- Each software has specific heuristics that leads to different results, even at \mathcal{H}_0
- Dealing with multiple hits is a strong difference point
- Choice of software depend of the biological question
- Tuning of parameters is important
- Similar trends for the bacterial datasets.
- Two software, BWA and Bowtie seem a little ahead

Conclusions and future work

Benchmark

- Each software has specific heuristics that leads to different results, even at H_0
- Dealing with multiple hits is a strong difference point
- Choice of software depend of the biological question
- Tuning of parameters is important
- Similar trends for the bacterial datasets.
- Two software, BWA and Bowtie seem a little ahead

Ongoing work

- Running latest release of some software (Bowtie2, Soap3, . . .)
- paired-reads, indels

Acknowledgments

People involved in this work

- Véronique Martin
- Matthias Zytnicki
- Julien Fayolle
- Valentin Loux
- Jean-François Gibrat
- Pierre Nicolas
- Jérôme Compain

Computational resources



Funding

ANR CBME : Computational
Biology for Metagenomics
Experiments

Mapping Reads on a Genomic Sequence : An Algorithmic Overview and a Practical Comparative Analysis

Journal of Computational Biology 2012 vol. 19, pp. 796-813.

<http://genome.jouy.inra.fr/ngs/mapping>