Comparaison d'outils de mapping pour les données NGS

Sophie S. Schbath

Comparaison d'outils de mapping pour les données NGS

Sophie Schbath

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

IRISA, Rennes, 19 mai 2011

Outils testés

Outils	Format	Algorithme	Input	Threads	Gaps
bwa	SAM	Burrows-Wheeler	nucléotides	oui	oui
Bowtie	SAM	Burrows-Wheeler	NT & color	oui	oui
Novoalign	SAM	indexe la référence	NT & color	-	-
MOM	perso	hash sur réf/reads	NT	oui	non
ProbeMatch	perso	hash sur référence	NT	non	oui
SOAP2	perso	Burrows-Wheeler	NT	oui	non
BFAST	SAM	hash sur référence	NT & color	oui	oui
SHRiMP	SAM	hash sur reads	NT & color	oui	oui
maq	SAM	hash sur reads	NT	non	non
SSAHA2	SAM	hash sur référence	NT	non	non
MPscan	perso	exact	NT	non	non

Outils testés

Outils	Format	Algorithme	Input	Threads	Gaps
bwa	SAM	Burrows-Wheeler	nucléotides	oui	oui
Bowtie	SAM	Burrows-Wheeler	NT & color	oui	oui
Novoalign	SAM	indexe la référence	NT & color	-	-
MOM	perso	hash sur réf/reads	NT	oui	non
ProbeMatch	perso	hash sur référence	NT	non	oui
SOAP2	perso	Burrows-Wheeler	NT	oui	non
BFAST	SAM	hash sur référence	NT & color	oui	oui
SHRiMP	SAM	hash sur reads	NT & color	oui	oui
maq	SAM	hash sur reads	NT	non	non
SSAHA2	SAM	hash sur référence	NT	non	non
MPscan	perso	exact	NT	non	non

Génération des reads $\mathcal{H}_0$

- Genome de référence = genome humain
- Longueur de reads = 40 bps
- Nombre de reads = 10 millions
- Tirage uniforme sur les 2 brins

$\longrightarrow$ 49 reads avec des ’N’ (1 à 10 N)
$\longrightarrow$ environ 1 123 000 reads non uniques
$\longrightarrow$ nombre maximal d’occurrences = 53162

Mapping exact de $\mathcal{H}_0$ sur génome humain

Software	Indexing time	Mapping time	No map. reads	Mapped reads	Original position	
					retrieved	not ret.
bwa	1h 28mn	48mn	49	9 999 951	9 999 951	0
Novoalign	23mn	10h 50mn	632	9 999 368	9 999 368	0
Bowtie	3h 32mn	21 mn	49	9 999 951	9 999 951	0
SOAP2	1h 34mn	56mn	49	9 999 951	9 996 385	3 566
BFAST	14mn +	13h 39mn	703 910	9 296 090	9 253 677	42 413
maq	6 mn	8 h 46 mn	42	9 999 958		
SSAHA2	29 mn	1d 12 h	35 875	9 964 125	9 770 914	193 211
MPscan	-	2×40 mn	26	9 999 974	9 999 974	0

TAB.: Global characteristics of the run of each software on the exact human dataset ($\mathcal{H}_0$) : computation times, number of reads which have not been mapped among the 10 millions reads, number of mapped reads and number of mapped reads whose original position has been retrieved in the complete list of hits.

Mapping exact de $\mathcal{H}_0$ (suite)

		Unique reads		Non unique reads			
Software	No map. reads	Nb	Ori. pos. not retr.	Nb	Nb hits mean [sd]	Ori. pos. not retr.	Mean rank [sd]
bwa	49	*8 877 061*	*0*	*1 122 890*	*722.81* [*2424.86*]	0	*362.56* [*1414.67*]
Novoalign	632	8 877 101	0	1 122 267	*698.62* [*2171.48*]	0	*350.27* [*1269.76*]
Bowtie	49	8 877 061	0	1 122 890	722.81 [2424.86]	0	361.56 [1415.16]
SOAP2	49	8 877 061	0	1 122 890	*653.26* [*1804.9*]	*3566*	*312.36* [*1016.2*]
BFAST	703 910	8 851 822	20 676	444 268	2.95 [1.47]	21 737	1.98 [1.16]
maq	42	7 503 738	496 642	2 496 217	101.66 [91.88]	1 574 273	22.6 [40.45]
SSAHA2	35 875	8 886 204	9 847	1 077 921	*79.52* [*151.7*]	*183 364*	*20.96* [*59.33*]
MPscan	26	8 877 081	0	1 122 893	*722.81* [*2424.86*]	0	*361.94* [*1407.16*]

Génération des reads $\mathcal{H}_3$

- A partir de $\mathcal{H}_0$
- 3 mutations réparties uniformément sur chaque read
- mutations équiprobables

Mapping jusqu'à 3 mismatchs de $\mathcal{H}_3$

Software	Indexing time	Mapping time	No map. reads	Mapped reads	Original position	
					retrieved	not ret.
bwa	1h 28mn	3h 16mn	4241618	5758382	5073020	685362
Novoalign	23mn	4d 8h	27	9999973	9334497	665476
Bowtie	3h 32mn	3h 31mn	49	9999951	9998521	1430
SOAP2	1h 34mn	48 mn	9334543	665457	0	665457
BFAST	14mn +	10h 30 mn	78651	9921349	8946942	974407
maq	6 mn	1 d 3h	43	9999957		
SSAHA2	29 mn	5 d 22 h	213	9999787	5537652	4462135
MPscan	2×25 mn	-	9990633	9367	0	9367

TAB.: Global characteristics of the run of each software on the 3 mismatches human dataset ($H3$) : computation times, number of reads which have not been mapped among the 10 millions reads, number of mapped reads and number of mapped reads whose original position has been retrieved in the complete list of hits.

Mapping jusqu'à 3 mismatchs de $\mathcal{H}_3$ (suite)

		Unique reads		Non unique reads			
Software	No map. reads	Nb	Ori. pos. not retr.	Nb	Nb hits mean [sd]	Ori. pos. not retr.	Mean rank [sd]
bwa	*4241618*	*4630867*	*213999*	*1127515*	*124.92* [*403.28*]	*471363*	*66.44* [266.01]
Novoalign	27	8699126	202456	1300847	14.48 [26.93]	*463020*	*6.34* [19.74]
Bowtie	49	*8494619*	0	*1505332*	*1103* [*3715.08*]	*1430*	*567* [*2171*]
SOAP2	9334543	202437	202437	463020	*21.36* [40.67]	463020	-
BFAST	78651	8759952	367478	1161397	3.95 [2.67]	606929	2.23 [1.63]
SSAHA2	213	8286416	3085913	1713371	6.8 [14.87]	*1376222*	*3.6* [*9.13*]
MPscan	9990633	5750	5750	3617	*14.97* [*101.18*]	3617	-

Remerciements

Collaborateurs :

- Julien Fayolle
- Véronique Martin
- Valentin Loux
- Jean-François Gibrat

Soutien : ANR CBME