Comparison of short read aligners with Galaxy

Subazini Thankaswamy Kosalai, Jens Nielsen, Intawat Nookaew* Systems and Synthetic Biology, Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden 41296.

Abstract

The emergence of Next generation sequencing (NGS) technology ensued production of large-scale data in fast pace demanding increased storage resource and computational power. The essential step in NGS analysis is read alignment or mapping with reference genome to determine the desired DNA sequence. The genetic difference between strains attained on mapping can also be used in variant detection and annotation. It is difficult to determine the position of short reads by mapping, mostly in the case of repetitive regions. Many tools developed for short read sequence alignment are available public and mostly commandline. On the other hand end-users find it more convenient when the tools are with user-interface. Galaxy is an integrated frame, which can be used in resolving computational issues, by allowing the tools to be deployed in cloud called Galaxy CloudMan. It also allows user to create a well-defined user-interface for command-line tools in XML. In this work, we have deployed different mappers or aligners based on different algorithms in Galaxy CloudMan and compared them for sensitivity and speed with allowed mismatch. XML Wrapper files are generated to create user-defined interface for the command-line mappers and deployed in galaxy so that it can be utilized for constructing workflows. The challenge is to select a mapping tool with fundamental priorities of speed, sensitivity and minimal memory usage. We made criteria for setting different parameters suitable for researchers' project and evaluating the aligners using mapping speed, RAM occupancy, sensitivity and accuracy using short read simulators and some real data.

Service over network



Parameters selection

1	Α	В	С	D	E	F	G	Н	I	J
1		Blat	Bowtie2	BWA	GEM	Novocraft	RAZER	smalt	stampy	soap
2		maxGap	-i <func></func>	-o INT						-g <int></int>
3	GAP	The size of max gap between tiles in a clump. Usually between 0 and 3	Sets a function governing the interval between seed substrings to use during multiseed alignment.	Maximum number of gap opens [1]						maximum gap size allowed on a read, default=0bp
4		extendthroughN			max-matches-per-					
5	Extension	Allows extension of alignment through large blocks of N's			Selects how many extensions per match should be attempted.					
6			-X/maxins <int></int>		[max-insert-size number]			[-i]	[insertsize=n]	-x <int></int>
7	Max insert		The maximum fragment length for valid paired-end alignments.		Specifies the maximum acceptable insert size for the pair.			insertmax Maximum insert size for paired-end reads. insertmax is a positive integer (default 500).	(Initial) mean insert size for paired-end reads [250]	maximal insert size allowed, default=600
9	Min insert		The minimum fragment length for valid paired-end alignments.					insertmin Minimum insert size for paired-end reads insertmax is a positive integer (default 0).		minimal insert size allowed, default=400
10		minScore	score-min <func></func>							
11	Score	match minus mismatch minus gap penalty. Default 30	Sets a function governing the minimum alignment score needed for an alignment to be considered "valid"							
12		50		-O INT		-n 00		[_c]	[ganonen=n]	-a cinth
13	Gap open penalty		Sets the read gap open (<int1>) and extend (<int2>) penalties.</int2></int1>	Gap open penalty [11]		Sets the gap opening penal ty. Default 40		scorspec Specify alignment penalty scores for a match or mismatch (substitution), or for opening or extending a gap.	Gap open penalty (phred score) [40]	maximum gap size allowed
14			rfg <int1> <int2></int2></int1>	-F INT		-y 99		[-s]	[gapertend=n]	
14	Gap Extend penalty		Sets the reference gap open (<int1>) and extend (<int2>) penalties. A reference gap of length N gets a penalty of <int1> + N * <int2>. Default: 5, 3.</int2></int1></int2></int1>	Gap extension penalty [4]		Sets the gap extend penalty . Default 6 Bisulphite Alignment Optio ns (Novoalign Only):		scorspec Specify alignment penalty scores for a match or mismatch (substitution), or for opening or extending a gap. scorspec is a comma speparated list of integer	Gap extension penalty (phred score) [3]	
16		oneOff	mp MX,MN	-M INT	-m			[-s]		-v <int></int>
17	Mismatch penalty	Allows one mismatch in tile	Sets the maximum (MX) and minimum (MN) mismatch penalties, both integers.	Mismatch penalty. BWA will not search for suboptimal hits with a score lower than (bestScore-misMsc). [3]	The maximum number of nucleotide substitutions allowed while mapping each read.			penalty scores for a match or mismatch (substitution), or for opening or extending a gap.		maximum number of mismatches allowed on a read, <=5. default=2bp

NGS analysis



Stein Genome Biology 2010 11:207

Galaxy

Galaxy provides a web based interface for deploying tools in cloud and also for analyzing and manipulating NGS data

www.sysbio.se

Advantages

 Provides computational and storage resources

Challenges

- Big data size of NGS sequences
- Extensive memory required for data storage
- Servicing of certain softwares used in NGS analyses imposes additional cost
- Moving data is non trivial
- Requires extensive computing power for data management and analysis

Galaxy CloudMan

- Independent queries on genomic data from different sources (UCSC, Yeast mine ...)
- Can share history with other Galaxy users.
- Workflow, can be developed in combination, refinement, calculation, extraction and visualization of queries
- Multiple analysis by Query intersection, subtractions and proximity searches

Read mapping									
	Loci 1	Loci 2	Loci 3						
	read Score1 -	> confidence 98%							

- Galaxy	Analyze Data	Workflow Shared Data -	Visualization - H	lelp - User -	Using 5.1 MB
Tools #	novoalign (V3 (C) 2008,2009,	.00.02 - Build Apr 1 20 ,2010,2011 NovoCraft Tec	13 @ 09:45:36 - A hnologies Sdn Bhd	short read alig	History C ¢
user-defined length cutoff #	License file r Licensed for e	not found. evaluation, educational,	and not-for-prof	it use only.	106: Novocraft on data 150: (A) (2)
Alignment tools #	novoalign -f	/home/subazini/galaxy-d	ist/database/file	es/000/dataset_15	1.000 lines
 Blat an alignment tool, based on 	Interpreting at In	innut files as Sander FA	format bt database: 2		
index generation of entire genome	Index Build Ve	ersion: 3.0	510.		Normal M, ualabase. <u>r</u>
in memory #	Hash length: 1	12			None /nome/subazini/galaxy-dist/tools
#	Step size: 1				/aligners_test/novocraft/hg18/hg18.nix
 <u>Bowtie</u> is a short-read aligner 	r0/1 S	TATTCTTCCGCATCCTTCATACT	CCTGCCGGTCAG	EDCCCBAAAA@@@@?>	60 99 9 YES /home/subazini/galaxy-
(e)	rl/l S	TGATAGATCTCTTTTTCGCGCC	GACATCTACGCC	EDCCCBAAAA@@@@@?>	dist/database/files/000/dataset_196.dat
 <u>Bowtie2</u> is a short-read aligner 	r2/1 S	CACGCCCTTTGTAAGTGGACATC	ACGCCCTGAGCG	EDCCCBAAAA@@@@?>	/home/subazini/galaxy-dist/database
Man with RWA @	r3/1 S	CGATGCAGATGCGTACCACCTGG	ACCAGGCCTTTC	EDCCCBAAAA@@@@@?>	/files/000/dataset_159.dat None
- <u>map with DWA</u> (0)	r4/1 S	TGACTITGCTGGTGCCAGAAGTG	GCGGCTGAAGTG	EDCCCBAAAA@@@@?>	
 GEM an alignment tool, based on 	r5/1 S	CGGGGATITAGCTTCGCCTITGG	CGACAGCAGCCA	EDCCCBAAAA@@@@?>	
index generation of entire genome	r6/1 S	GCGTATACGGCGCTGACCCAGGC	TTTTTATCTGTT	EDCCCBAAAA@@@@@?>	# several days (VO 00 00 Dudld Ass. 1
in memory	r7/1 S	GTATCGCTGTTTTCCAGTTGTTC	AAGATAAGAAAA	EDCCCBAAAA@@@@@?>	# novoalign (V3.00.02 - Build Apr 1
e	r8/1 S	CCACGGIIGAIGCIGGCAICGCI	GATTGGTGCGTT	EDCCCBAAAA@@@@?>	# (C) 2008,2009,2010,2011 NovoCraft
 Masai maps query reads onto the 	r9/1 S	GCGCCTTATCACCTATCC	AGCAATTACGTA	EDCCCBAAAA@@@@@?>	Aliense file ask found
reference sequences	$r_{10/1}$ S	TTATCAATCTAACCACCACCTAC	ACCOLLATACCC	EDCCCBAAAAddddd:>	# License file not found.
(d)	$r_{12/1}$ S	TGCTCCGGCAGCGCCTGCTCCGG	CGGCGCAACCGA	EDCCCBAAAA@@@@?>	# Licensed for evaluation, educatior
 <u>Novocraft</u> maps query reads onto 	$r_{12/1} S$		GCTATGAAACCOA	EDCCCBAAAA@@@@@:>	#
the reference sequences	r14/1 S	GTAATTTGAGTAATGCCCACCAG	TTCCATCACGAT	EDCCCBAAAAaaaaa?>	# novoalign -r /nome/subazini/galax
	$r_{15/1}$ S	ACGGATTGGCCAGGGTTTCAAAA	TCCTGCGCCCAC	EDCCCBAAAAaaaaa?>	# Starting at Thu May 16 12:46:20 20
 <u>Razer</u> maps query reads onto the 	r16/1 S	CTATTTCTTGTTGGAGCATGGTT	ATCCGCCGCAGG	EDCCCBAAAA@@@@?>	
reference sequences	r17/1 S	TGTCGCCATCTTTAATGGCTGGC	AGGTAATTTTGC	EDCCCBAAAA@@@@?>	
- Stompujo o poskogo for monning	r18/1 S	TCCGTAAATTCACCATGAAACTT	GCCCGGATGCAG	EDCCCBAAAA@@@@?>	
 <u>Stampy</u> is a package for mapping (a) 	r19/1 S	GCAAGACTTTGCAGTGTCATCAG	CGGATCGGTCAG	EDCCCBAAAA@@@@?>	191: SMALT on data 34:
of short reads from Illumina	r20/l S	TCTTTATTTTTCCCAGAGCCTAA	CGACTTGCCATG	EDCCCBAAAA@@@@?>	mapped reads
sequencing machines onto a	r21/1 S	ATCGAGATATTTTATATACTGAT	TTATTATT	EDCCCBAAAA@@@@@?>	3 lines 3 comments
reference genome @	r22/1 S	ATGACTCGCTCAACGTTCTGCTC	GTTGATCTCGCC	EDCCCBAAAA@@@@?>	format: com databaco: ba10
e SMALT mono quenuro de ente the	r23/1 S	ACTCAGTCTTGGGGGGACTCAGCC	CGGTAGAATACC	EDCCCBAAAA@@@@?>	outure (and a constant of the
<u>SMALT</u> maps query reads onto the @	r24/1 S	TCAAAGAGTGCTTGTTCGATACC	GAACTGAAAGCG	EDCCCBAAAA@@@@@?>	SMALT (Version: 0.7.3) SMALT run on
reference sequences @	r25/1 S	GCCCTGTGCCGCGCGCTGCCGGAAC	GTACCCATTCGG	EDCCCBAAAA@@@@?>	single-end data
Get Data	r26/1 S	CGAGGAAGGACGTGGCGAGGGGCT	AAATCAGCAAAT	EDCCCBAAAA@@@@?>	📊 🛈 🖄 🖄
<u>ocroata</u> (el	r2//1 S	CCTGCCGTCGAAATTTGTTCGGA	TCTCTCAACACT	EDCCCBAAAA@@@@?>	
<	128/1 5	GGAATATTTGCGATTTGCCATTT	TCTCTCAAGAGI	EDCCCBAAAA@@@@(>	

- $\Rightarrow score2 \rightarrow confidence 78\%$
- $\} \rightarrow \text{score3} \rightarrow \text{confidence 100\%}$

Alignment with best score determines the quality of mapping. It depends on similarity of reads and number of alignments reported.

References:

- Thakur RS, Bandopadhyay R, Chaudhary B, Chatterjee S. Now and next-generation sequencing techniques: future of sequence analysis using cloud computing.
 Front Genet. 2012;3:280
- Afgan E, Baker D, Coraor N, Chapman B, Nekrutenko A, Taylor J. Galaxy CloudMan: delivering cloud compute clusters. BMC Bioinformatics. 2010 Dec 21;11 Suppl 12:S4.

