Comparison of short read aligners with Galaxy

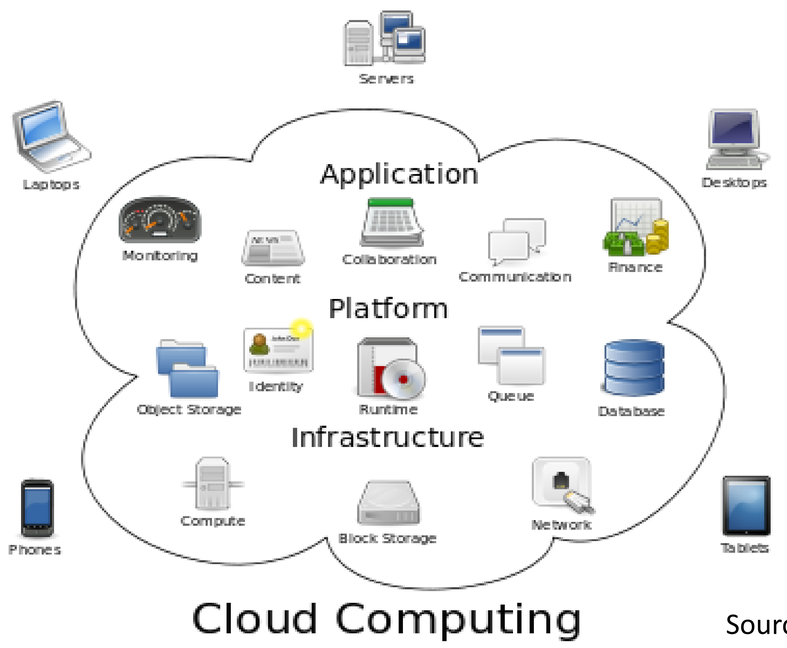
Subazini Thankaswamy Kosalai, Jens Nielsen, Intawat Nookaew*
Systems and Synthetic Biology,
Department of Chemical and Biological Engineering,
Chalmers University of Technology, Gothenburg, Sweden 41296.

sys²bio
www.sysbio.se

Abstract

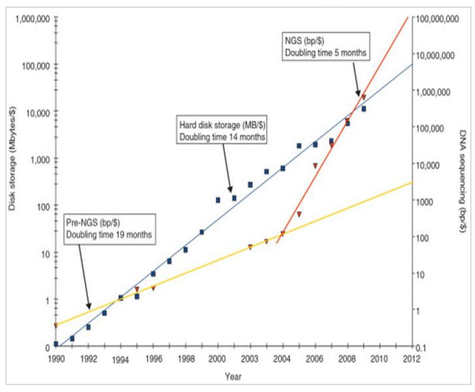
The emergence of Next generation sequencing (NGS) technology ensued production of large-scale data in fast pace demanding increased storage resource and computational power. The essential step in NGS analysis is read alignment or mapping with reference genome to determine the desired DNA sequence. The 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 command-line. 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



Source: Wikipedia

NGS analysis



Stein *Genome Biology* 2010 **11**:207

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

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

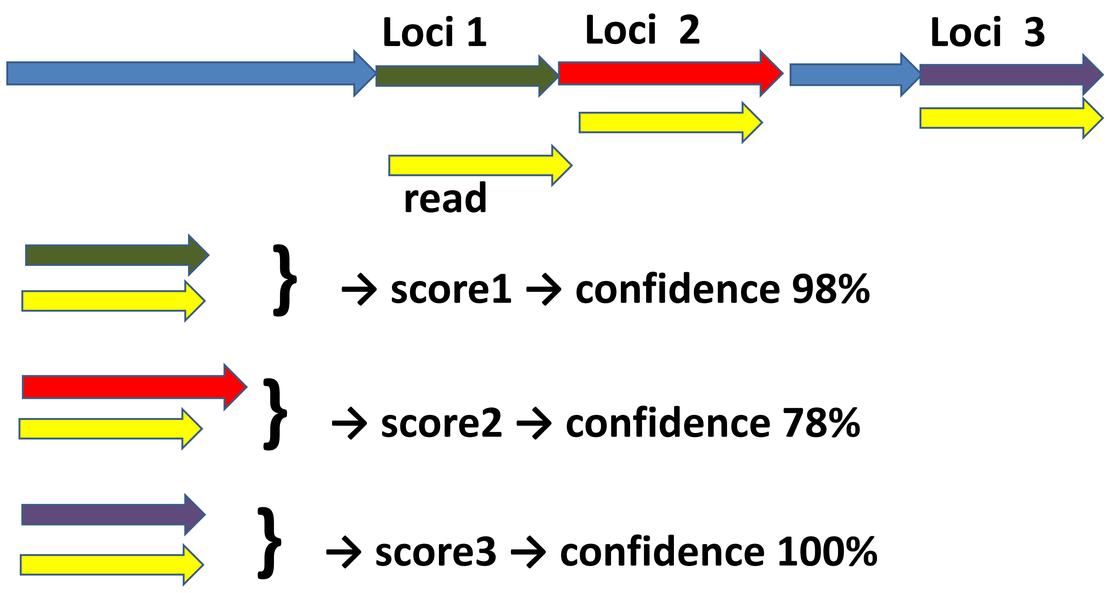
Advantages

- Provides computational and storage resources
- 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

Parameters selection

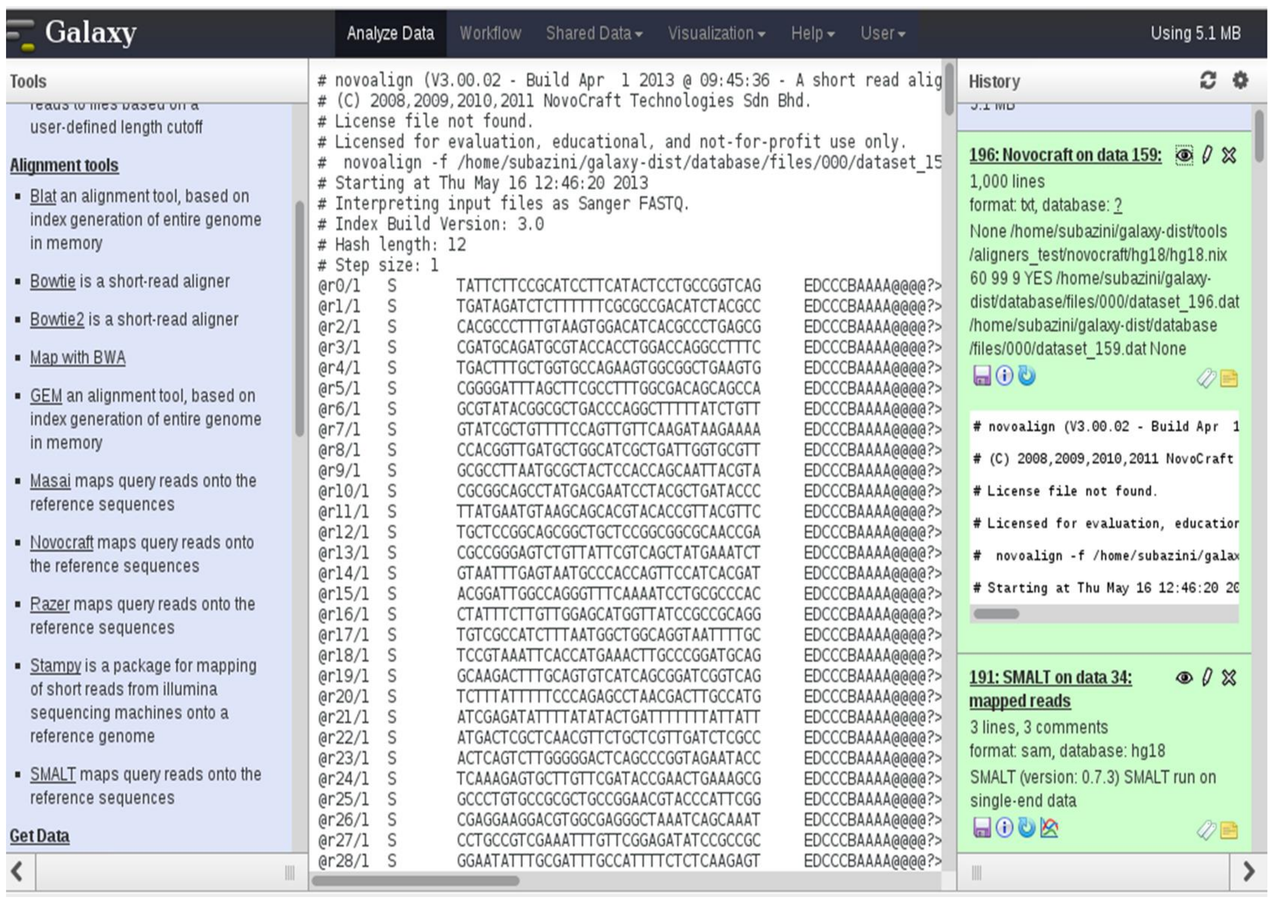
	A	B	C	D	E	F	G	H	I	J
1		Blat	Bowtie2	BWA	GEM	Novocraft	RAZER	smalt	stampy	soap
2	GAP	maxGap	-i <func>	-o INT						-g <int>
3		The size of max gap between tiles in a clump. Usually between 0 and 3	Sets a function governing the interval between seed substrings to use during multiseed alignment.	Maximum number of gap opens [1]						maximum gap size allowed on a read, default=0bp
4	Extension	extendthroughN			--max-matches-per-					
5		Allows extension of alignment through large blocks of N's			Selects how many extensions per match should be attempted.					
6	Max insert		-X/--maxins <int>		[--max-insert-size number]			[-i]	[--insertsize=n]	-x <int>
7			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
8	Min insert		-I/--minins <int>					[-j]		-m <int>
9			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	Score	minScore	--score-min <func>							
11		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	Gap open penalty		--rdg <int1>,<int2>	-O INT		-g 99		[-s]	[--gapopen=n]	-g <int>
13			Sets the read gap open (<int1>) and extend (<int2>) penalties.	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 on a read, default=0bp
14	Gap Extend penalty		--rfg <int1>,<int2>	-E INT		-x 99		[-s]	[--gapextend=n]	
15			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.	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	Mismatch penalty	oneOff	--mp MX,MN	-M INT	-m			[-s]		-v <int>
17		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

Read mapping



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

Galaxy CloudMan



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.

CHALMERS