Mapdecode : inventory and benchmark of read mapping tools

Jérome Compain, Jullien Renaud, Sivagangari Nandy, Olivier Collin, Jean-François Gibrat, Valentin Loux, Veronique V. Martin, Sophie S. Schbath

Poster_ECCB2014 Submission 115

Update information
Update authors
Withdraw

If you want to **change any information** about your paper or withdraw it, use links in the upper right corner.

For all questions related to processing your submission you should contact the conference organizers. Click here to see information about this conference.

All **reviews sent to you** can be found at the bottom of this page.

Paper 115 (abstract only)

Title:	Mapdecode : inventory and benchmark of read mapping tools
Track:	A: Sequencing and sequence analysis for genomics
Author keywords:	benchmark NGS mapping
Abstract:	New sequencing technologies are able to produce enormous amounts of data, up to a billion reads per run. The first step of many analyses of these data, for instance the study of gene differential expression (RNA-seq), the study of gene regulation (ChIP-seq, Methyl-seq), the search for genomic variants (SNPs, chromosomal rearrangements) to name but a few, starts with mapping the reads on the reference genome. Over the last seven years, many mapping methods have been proposed to efficiently cope with the avalanche of data produced by the new generation sequencing (NGS) technologies. There exist exact algorithms for mapping reads on reference genomes, with or without indels but these algorithms are far too slow to be used with NGS data and large genomes. Therefore, most mapping methods implement heuristics that provide a trade-off between speed and accuracy. This trade-off often leads to the development of complex software with many ad-hoc options whose effects on the mapping results are usually difficult to predict beforehand. To help users choosing a particular mapping program that best suits their needs a number of benchmarks have been recently published that differs in the mapping tools they consider and in their methodology for carrying out the benchmarks (criteria for evaluating the mapping results, data used, etc.). In this work, we extend the work done in Schbath et al., 2012 by focusing, more specifically, on paired reads and extending the test cases (longer reads, consideration of indels, etc.) For the purpose of this study, we first compiled an inventory of 93 published mapping tools. However, only 25 of those tools seem actively maintained and have been updated since March 2013. From this list, we evaluated the performance of seven mapping tools on simulated datasets. We generated 3 different datasets from the human genome. The first one contains 10 millions 40 bp reads, the second one, contains 10 millions 100 bp reads. The last one contains 5 millions 2x100 bp paired reads. From these 3 datasets we then created further datasets with 1, 2 or 3 mismatches per reads. For the long and paired datasets, we also created a dataset with an insertion of three consecutive random nucleotides and a dataset with a deletion of three consecutive nucleotides. Finally, we generated a more realistic dataset by using a software that can simulate reads according to errors profiles observed in real datasets (we used bacterial data sequenced with an Illumina Hiseq2000). All the mapping tools were evaluated for correctness of mapping against these datasets. The inventory of the mapping tools and the benchmark results for the one tested are available on the Mapdecode website (http://mapdecode.genouest.org). This work is supported by the "France Génomique" project (ANR-10-INBS-0009)
Time:	May 30, 12:24 GMT

Time:	May 30, 12:34 GMT

Authors

first name	last name	email	country	organization	Web site	corresponding?
Jérôme	Compain	jerome.compain@jouy.inra.fr	France	Mathématique, Informatique et Génome - INRA		
Renaud	Jullien	renaud.jullien@irisa.fr	France	Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA)		
Sivasangari	Nandy	sivasangari.nandy@irisa.fr	France	Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA)		
Olivier	Collin	olivier.collin@irisa.fr	France	Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA)		
Jean-Francois	Gibrat	Jean-Francois.Gibrat@jouy.inra.fr	France	Mathématique, Informatique et Génome - INRA	http://mig.jouy.inra.fr	
Valentin	Loux	valentin.loux@jouy.inra.fr	France	Mathématique, Informatique et Génome - INRA		✔
Véronique	Martin	veronique.martin@jouy.inra.fr	France	Mathématique, Informatique et Génome - INRA		
Sophie	Schbath	sophie.schbath@jouy.inra.fr	France	Mathématique, Informatique et Génome - INRA		

Reviews

Review 1

Reviewer's confidence:	**5**: (expert)
Review:	The paper deals with the comparison of NGS mapping tools It perfectly fit with ECCB program