Comparative analysis of algorithms for next-generation sequencing read alignment

Matthew Ruffalo [1,*], Mehmet Koyutürk [1,3] and Thomas LaFramboise [2,3]

[1]Department of Electrical Engineering & Computer Science
[2]Department of Genetics
[3]Center for Proteomics and Bioinformatics
Case Western Reserve University

[*]To whom correspondence should be addressed: matthew.ruffalo@case.edu

March 7, 2011

1 Abstract

The advent of next-generation sequencing (NGS) techniques presents many novel opportunities for genomic applications. The vast number of short reads produced by these techniques, however, pose significant computational challenges. In many applications, the first step in any kind of analysis is the mapping of short reads to a reference genome, and many groups have developed algorithms and software packages to perform this function. As the developers of these packages optimize their algorithms with respect to various considerations, the relative merits of different software packages remain unclear. However, for scientists who generate and use NGS data for their specific research projects, an important question is to choose the software that is most suitable for their application.

With a view to comparing existing short read alignment software, we develop a simulation and evaluation suite, Seal, which simulates NGS runs for different configurations of various factors, including sequencing error, indels, and coverage. We also develop performance criteria to compare the performance of software with disparate output structure (*e.g.*, some packages return a single alignment while some return multiple possible alignments). Using these criteria, we comprehensively evaluate the performances of Bowtie, BWA, mr- and mrsFAST, Novoalign, SHRiMP and SOAPv2, with regard to sensitivity, specificity, and runtime on resequencing the human genome, as well as carefully designed synthetic genomes.

We expect that the results presented here will be useful to investigators in choosing the alignment software that is most suitable for their specific research aims. Our results also provide insights into the factors that should be considered to use alignment results effectively. Seal can also be used to evaluate the performance of algorithms that use deep sequencing data for various purposes (*e.g.*, identification of genomic variants).