# BBMap: A Fast, Accurate, Splice-Aware Aligner

**Brian Bushnell**[1*]

[1] LBNL Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, USA

**To whom correspondence should be addressed*: Email: bbushnell@lbl.gov

March 21, 2014

**ACKNOWLEDGMENTS:**

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Special thanks for testing, suggestions and feedback to: Rob Egan, Alex Copeland, Brian Foster, Alicia Clum, Hui Sun, Kurt LaButti, Vasanth Singan, Andrew Tritt, Alexander Spunde, Joel Martin, James Han, Mat Nolan and Matt Scholz.

# BBMap: A Fast, Accurate, Splice-Aware Aligner


Brian Bushnell

Department of Energy, Joint Genome Institute, 2800 Mitchel Drive, Walnut Creek, California 94598, USA


## Introduction

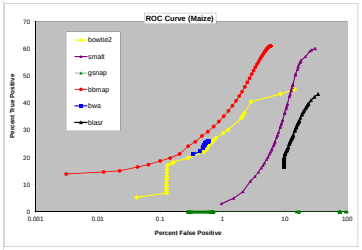
Alignment of reads is one of the primary computational tasks in bioinformatics. Of paramount importance to resequencing, alignment is also crucial to other areas - quality control, scaffolding, string-graph assembly, homology detection, assembly evaluation, error-correction, expression quantification, and even as a tool to evaluate other tools. An optimal aligner would greatly improve virtually any sequencing process, but optimal alignment is prohibitively expensive for gigabases of data. Here, we will present BBMap [1], a fast splice-aware aligner for short and long reads. We will demonstrate that BBMap has superior speed, sensitivity, and specificity to alternative high-throughput aligners bowtie2 [2], bwa [3], smalt, [4] GSNAP [5], and BLASR [6].

## Problem Description

Mapping perfect reads is easy, but real reads have errors and mutations; any reference substring can be transformed into any read by applying a series of insertions, deletions, substitutions, and no-calls. The alignment game is played by assigning scores to these operations, then finding the location(s) in a reference maximizing that function. A function correctly reflecting probabilities of errors and mutations will yield its highest score at a read's most likely origin, allowing correct alignments to be made. A good aligner will be able to map reads rapidly and accurately in the presence of mutations.

## Materials & Methods

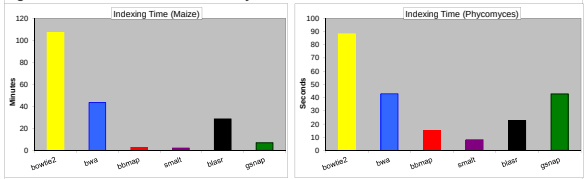
To evaluate the aligners, synthetic reads were generated, transmuted, and tagged with their genomic start and stop. Aligners then mapped reads to the reference, and the resulting sam file was graded. A mapping was considered 'strictly correct' if the start and stop matched the genomic origin. To reflect the diversity of JGI research, multiple genomes were used: 1 fungus, 1 bacteria, 1 plant, and 1 soil metagenome. Prior to mapping, the genomes were fully indexed, and the times recorded.




**Strict True Positive**: Both endpoints of read map to their correct origins. Essential for calling variations.

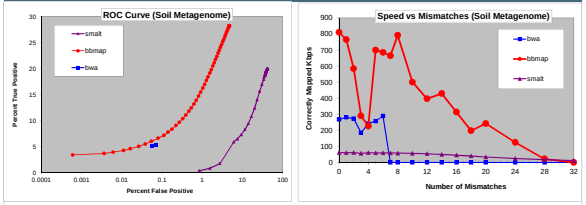
**Strict False Positive**: Any read not a strict true positive. These may be mismapped, or just misaligned

The tests were run on an exclusive NERSC Mendel node with 128GB RAM and 16 Ivy Bridge E cores in two sockets. All programs used their default settings and 32 threads, as hyperthreading was enabled. All reads were single-ended 150bp reads except synthetic PacBio, which was 400bp reads. ROC curves are generated from reads with a variety of SNPs, indels, and nocalls.

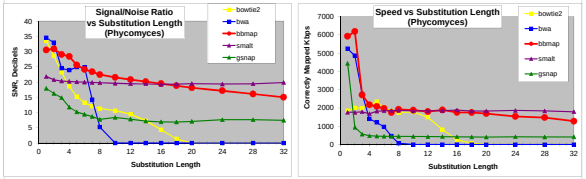



Smalt and BBMap have much faster indexing than FM-transform aligners.

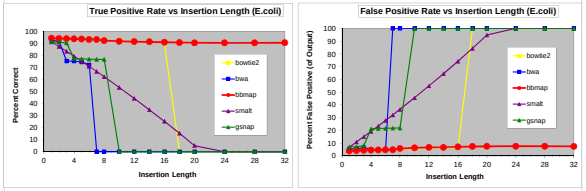
## Results



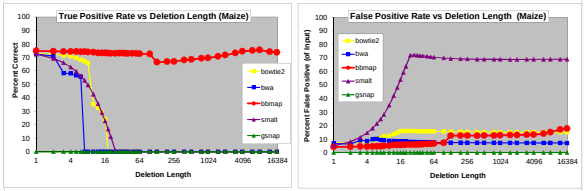

Only 3 aligners were capable of indexing the metagenome, though it was not particularly large, at 5Gbp and 22M scaffolds.



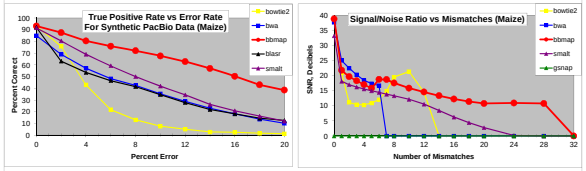

BBMap is good with contiguous substitutions that trouble FM-indexed aligners.




Bowtie2 and BBMap both handle insertions well, but BBMap handles longer ones.




BBMap is unrivalled at processing long deletions. Bowtie2 is second, while smalt has an exceptionally high error rate.




BBMap is the most accurate with PacBio's error profile, even moreso than PacBio's own BLASR. Real PacBio data is around 15% error.

BBMap is always first or second in this graph, and is capable of accurately mapping reads with more errors than other aligners.

## Conclusions

- **BBMap** is shown to be a fast and accurate aligner, capable of correctly handling an overall wider variety of references, reads, and mutations than others. It has particularly outstanding performance with deletions, especially long ones, that other aligners cannot handle at all. While less common than SNPs, such large-scale features indicating (for example) the complete absence of a gene or promoter will not even be detected by other aligners.
- **GSNAP**'s performance was unexpectedly bad. It was incapable of indexing soil, and yielded 100% incorrect mappings against Maize, for unknown reasons. It had generally inferior performance on the two genomes it seemed able to process, Phycomyces and E.coli. And despite being billed as a de novo splice aligner, GSNAP was incapable of mapping reads across any long deletions.
- **Bwa** was fairly fast and showed fairly good results as long as the edit distance was less than 7, but was incapable of handling low-identity reads and performed poorly with indels. Though these tests were run with default settings, more sensitive settings were also explored, causing an exponential increase in runtime with marginal improvement in results.
- **Bowtie2** did a fairly good job in all cases. Though slower than bwa with perfect reads, it was much better able to handle lower-identity reads and indels. Overall, it seemed like a better tool than bwa for general use; but as it failed to index the soil metagenome, it can't be recommended for metagenomics.
- **Smalt** was the only aligner to maintain a consistent speed with decreasing read identity, and was able to map more highly mutated reads than anything else, even BBMap. However, it does so at the cost of extremely high false positive rates, particularly for indels.
- **BLASR** had acceptable speed but poor accuracy on synthetic PacBio data, mapping fewer reads and generating more false positives than alternatives. This may in part be because BLASR is optimized for its native format, rather than fastq. Regardless, it does not seem like a good choice for mapping short Illumina reads to PacBio for error correction.

## References


[1] sourceforge.net/projects/bbmap/

[2] bowtie-bio.sourceforge.net/bowtie2/

[3] bio-bwa.sourceforge.net/

[4] research-pub.gene.com/gmap/

[5] www.sanger.ac.uk/resources/software/smalt/

[6] Chaisson MJ, Tesler G (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13:238 doi:10.1186/1471-2105-13-238


## Acknowledgements


Special thanks for testing, suggestions, and feedback go to:

Rob Egan, Alex Copeland, Brian Foster, Alicia Clum, Hui Sun, Kurt LaButti, Vasanth Singan, Andrew Tritt, Alexander Spunde, Joel Martin, James Han, Matt Nolan, and Matt Scholz..



For any questions or comments, please contact the author at bbushnell@lbl.gov



The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.