THE AUTHOR FILE

# Ben Langmead

## Building a better sequence alignment program

Ben Langmead hopes to be a special kind of translator, one who lets powerful instruments speak to scientists. The throughput of modern sequencing machines and other instruments is astonishing, he says. "But all that throughput has to go through computers before it goes back to scientists. As a computational biologist, I want to bridge the gap."

Ben Langmead

This goal was sufficiently strong to pull Langmead out of a comfortable job at a New York consulting company and into graduate school. As a consultant, Langmead had specialized in fast pattern-matching algorithms that network processors used to fend off spam and computer viruses. He started looking for ways to apply these skills in biology just as next-generation sequencing machines were hitting the marketplace, forcing scientists to think in terms of millions of short reads. After talking with researchers using graphical processing units to align reads to a reference genome, he wanted to try the task using network processers. "I thought, 'Oh, wow, this sounds perfect,'" Langmead recalls. In 2007, he joined Steven Salzberg's group, then at the University of Maryland, and set to work.

It took him about two months to realize that the route he had chosen was a blind alley. He knew how to get network processers to filter text for words like "Nigerian" followed by "prince" with all sorts of text in between, but handling sequencing data involved aligning strings with close mismatches, and that meant more possibilities than a network processor's memory was equipped to deal with efficiently.

Langmead began scouring the literature for something that could handle data in a more compact way and identified a recently developed method called full-text minute indexing. "It was a heck of a lot faster than scanning through the whole text," he says. "It could be a way of taking the genome and jumping to good candidates [for alignment]. Instead of scanning and asking 'Is this a candidate? Is this a candidate?', you can jump right to the right one."

This resulted in an alignment program based on an approach called the Burrows-Wheeler Transform, which Langmead dubbed ebwt. Salzberg worried no one would be able to pronounce the name, and proposed "Bowtie," which still contains the key letters. Bowtie was published in 2009 and quickly became a standard tool for aligning sequencing data.

Then sequencing technology changed. As reads became longer, sequencing machines began making different kinds of errors. Instead of just swapping one nucleotide for another (putting, say, an A or C or G where a T belonged), machines would also insert or delete nucleotides. The mathematics varies from read to read, but assuming an interrogated sequence is just 10 nucleotides long, the sequence space that must be searched to align mismatches is $4^{10}$. However, searching for mismatches as well as single-nucleotide insertions and deletions would require a search of $9^{10}$ possibilities. Searches of this magnitude are too much for full-text minute indexing.

Langmead thought that dividing the labor between two types of algorithms might allow such searches to proceed, so he combined full-text minute indexing with a technique called hardware-accelerated dynamic programming, which allows an arithmetic operation to be performed on an array of numbers simultaneously instead of on each number individually. The result was a new program, Bowtie 2. It first uses full-text minute indexing to identify a set of positions in a reference genome where portions of reads align, and then uses dynamic programming to identify which of these can be extended to give alignment of the full read.

"We're positioned well for a longer, gappier future."

This approach speeds up the very processes that previously slowed Bowtie down. An added bonus: the algorithm is now better at finding insertions and deletions in an analyzed genome relative to the reference genome. But the fix didn't come instantly, says Langmead. "Working with software is tricky, because something can seem like a good idea, and it takes months to implement it well enough to run on real data." And only real data can reveal whether an idea is feasible, he says. "It goes from 'I have a good idea' to 'That was a waste of time.'"

With perseverance, good ideas can show their worth. For example, when Langmead fed Bowtie 2 a couple of million HiSeq Illumina sequencing reads, its default settings could align those reads in 12 minutes. The next-fastest alignment tool took 32 minutes. Bowtie 2 also performs well on data from Ion Torrent and 454 machines, and Langmead predicts that it is likely to be the choice for reads longer than 40 or 50 nucleotides. "We're positioned well for a longer, gappier future," he says.

**Monya Baker**

Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).