Minireview

The need for speed

Paul Flicek

Address: European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Email: flicek@ebi.ac.uk

Published: 27 March 2009

*Genome **Biology*** 2009, **10:**212 (doi:10.1186/gb-2009-10-3-212)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/3/212

Abstract

DNA sequence data are being produced at an ever-increasing rate. The Bowtie sequence-alignment algorithm uses advanced data structures to help data analysis keep pace with data generation.

In this month's *Genome Biology*, Langmead and colleagues [1] present the Bowtie algorithm. Bowtie is designed to align large numbers of relatively short DNA sequencing reads to an entire reference genome. It does so by first taking the reference genome assembly and changing the order of the sequence using something called the Burrows-Wheeler Transform. Why is this useful? Speed is the best answer: Bowtie is more than 30 times faster than other published tools designed to do the same task. Let's step back and see why the need for speed in our analysis algorithms is greater now than at any time in the genomic age.

Over the past three years massively high-throughput sequencing, often called 'next-generation' sequencing, has developed from a few beta devices in key genome centers to a large installed base in research labs around the world. The success of sequencing machines such as Illumina/Solexa, ABI SOLiD and 454 FLX has facilitated the development of sequencing as a general-purpose experimental tool for many biological applications. The range of possible uses is rapidly establishing DNA sequencing as the microscope of modern biology.

The scale of data generation is amazing; for example, in the course of its pilot phase the 1000 Genomes Project [2] has already generated almost 2,000-fold total coverage of the human genome from 180 individual samples, an amount orders of magnitude larger than the original Human Genome Project. There is a very real chance that before 2012 the amount of data generated by worldwide DNA sequencing will exceed the expected 15 petabytes of data per year produced by CERN's Large Hadron Collider.

In the light of these spectacular developments in data-generation capacity, it should come as no surprise that the computational requirements for supporting large-scale genome sequencing are growing dramatically. A key question is whether bioinformaticians are up to the task. Fortunately, the sheer number of new algorithms - some, like Bowtie, are based on data structures and methods either newly introduced to biology or rediscovered in the light of challenges posed by next-generation sequence data - suggest that bioinformatics, if not yet entering a new golden age [3], is responding to the waves of data by building better surfboards rather than running for higher ground.

Alignment is one of the first and most fundamental problems for any sequencing-based project in which a reference genome assembly already exists for the species concerned. Today's resequencing and functional studies (Box 1) directly leverage the effort required to create high-quality finished and draft genome assemblies such as those available for the human and mouse genomes. For next-generation sequencing studies the collected DNA sequencing reads are almost completely meaningless until they are aligned. Even the knowledge of whether the experiment succeeded is unknown until the sequencing reads are aligned to the reference genome.

How do we address this essential step in the analysis and get as quickly as possible to the point where we can start to make sense of the biology? Programs such as Bowtie dramatically accelerate the alignment step by storing the reference genome in a highly ordered manner that facilitates very rapid searching of sequence. The key technology in Bowtie is called the Burrows-Wheeler Transform (BWT), which was originally developed for data compression. It works by reordering the original genome sequence such that certain patterns within the sequence are made explicit and therefore simplifies compression of the sequence. Importantly, the

Box 1. Resequencing and functional studies.

A small sampling of recent work leveraging the developments in DNA sequencing technology.

Resequencing projects

Individual genomes [12-14]

1000 Genomes project [2]

Large-scale resequencing of individual genomes originally done with short read sequencing as a proof of principle. The 1000 Genomes project is being done comprehensively using relatively low sequencing coverage over a large number of individuals to create a deep catalogue of human genetic variation.

Cancer genome sequencing [15]

Sequencing cancer genomes requires the sequencing of both the tumour genome and a matched normal sample from the same individual. Finding the potentially small number of differences between theses two samples currently requires that both genomes be sequenced to high coverage to ensure accurate mutation discovery.

Functional studies

Any experimental technique able to isolate a fraction of the genome involved in a specific biological function is a potential candidate for DNA sequence analysis.

ChIP-seq [16,17]

ChIP isolates regions of protein-DNA interaction, including transcription factor binding and locations of modified histones.

Nucleosome mapping [18]

By directly isolating nucleosomes and sequencing the DNA sequence that is wound around each one it is possible to directly assess chromatin state. For example, regions with consistently placed nucleosomes and apparently stable chromatin architecture are distinguishable from more dynamic regions.

DNase Seq [19]

Directly measuring DNase I hypersensitive regions is conceptually complementary to techniques for nucleosome mapping and is an effective genome-wide technique to identify many regulatory regions.

DNA methylation [20-22]

The methylated fraction of the genome can be assessed using a wide variety of methods amenable to DNA sequencing including MeDIP (methylated DNA immunoprecipitation) and techniques involving bisulphite conversion of methylated cytosines before sequencing.

Transcriptomics [23,24]

Transcriptome mapping has nearly limitless applications in normal and disease states. Unlike array-based methods, mapping transcription with direct DNA sequencing makes analysis of alternative splicing and discovery of novel transcripts relatively easy.

BWT reordering is reversible, so we are always able to reconstruct the original sequence. In fact, those readers who have ever downloaded compressed files from the Internet have probably already benefited from the BWT, which is at the heart of the bzip2 data compression algorithm [4].

Once the BWT has been constructed for the given genome assembly it is indexed for optimal searching by creating an FM index, which is, roughly speaking, a compressed suffix array of the genome sequence. These existing techniques and novel modifications by Langmead *et al*. [1] to existing sequencing matching algorithms allow Bowtie to use the FM index to rapidly align both exactly matching DNA sequencing reads and those with mismatches caused by sequencing error or sequence polymorphism, all while maintaining a memory footprint low enough to run on many standard laptop computers.

The BWT and the FM index are not complete strangers to bioinformatics. Several groups have adopted the data structure to solve specific problems mostly related to comparing many short segments of the genome to the genome as a whole. Before massive resequencing datasets existed, a common application of this problem was microarray probe design [5,6]. In this case, one effective way to estimate cross-hybridization potential for a given array design is to do a brute-force comparison of all short DNA segments (that is, possible array probes) to the genome as a whole.

Even when there are hundreds of billions of short sequencing reads the problem of alignment remains relatively easy compared with the problem of *de novo* genome assembly from short sequencing reads (especially for mammalian-sized genomes). A key difference comes from how easy it is to distribute the required computational work over the nodes of the compute clusters that are commonly used for bioinformatics analysis.

For example, alignment is considered 'embarrassingly parallel', so named because of how easy it is to achieve parallelization. For the case of read alignment to the reference genome, the most common way to distribute the task across a compute cluster is to store the complete reference genome on each of the nodes of the cluster and then distribute the collection of reads equally across the nodes. The read alignments can be merged at the end of the process. *De novo* assembly requires that essentially all the information needed to solve the problem (that is, how sequencing reads are related to each other) is available to the assembly program. For short-read datasets and mammalian-sized genomes, this generally leads to extremely large memory requirements that grow with the genome size and number of sequencing reads or to software implementations based on complex message passing between compute nodes.

To achieve large-scale alignment parallelization one only needs to be able to store the entire reference genome in memory available at each compute node. Without the BWT and the data compression it provides, storing a search-optimized data structure such as a suffix array for the entire genome is not feasible on each of the compute nodes found in today's clusters (see [5] for a more detailed discussion of the memory requirements of a mammalian genome suffix array both before and after a BWT).

Bowtie is not the only alignment program designed for next-generation sequence data using an index based on the BWT, but it does appear to be the first reported in the literature. The creators of SOAP [7] have recently introduced SOAP2 [8] and the creators of MAQ [9] have produced BWA [10], both of which provide a significant improvement in speed over the hash-table-based implementations of SOAP and MAQ.

For applications such as ChIP-seq and for rapid confirmation that the sequencing experiment performed as expected, Bowtie is likely to be the most effective solution. For some other applications, including whole-genome, paired-end resequencing projects, it may not yet be the right choice. Although much faster, Bowtie is not as accurate as MAQ in the case of a real dataset aligned with Bowtie's default parameters [1]. Parameter choices can increase Bowtie's accuracy, but at the cost of speed. Bowtie is also currently missing some critical functionality (for example, the ability to align paired reads). This functionality will certainly be added soon - either by the Bowtie developers, who have already implemented preliminary support for pair-end alignment in the most up-to-date version available on the Bowtie website [11], or by someone else enabled by Bowtie's open-source license.

Bowtie is yet another example of a common story in bioinformatics. Whereas default alignment programs are provided by the instrument manufacturers, the wider scientific community has developed the programs now used by many, if not most, researchers. This is a testament to the software-development skills within the research community and the desire within that community to create tools that are easy to deploy and use within existing analysis pipelines. There can be no doubt that open data formats and the ability to tap into the widest segment of the community in the search for solutions is the best way forward for DNA sequence analysis.

For now, sequence-alignment algorithms based on the BWT allow us to keep pace with the sequencing machines for at least another year. In today's fast-moving world of sequence generation, this is indeed a dramatic development.

References

1. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biology* 2009, **10:**R25.
2. **1000 Genomes** [http://www.1000genomes.org/page.php]

3. Stein LD: **Bioinformatics: alive and kicking.** *Genome Biol* 2008, **9:**114.
4. **bzip2** [http://bzip.org]
5. Gräf S, Nielsen FG, Kurtz S, Huynen MA, Birney E, Stunnenberg H, Flicek P: **Optimized design and assessment of whole genome tiling arrays.** *Bioinformatics* 2007, **23:**i195-i204.
6. Healy J, Thomas EE, Schwartz JT, Wigler M: **Annotating large genomes with exact word matches.** *Genome Res* 2003, **13:**2306-2315.
7. Li R, Li Y, Kristiansen K, Wang J: **SOAP: short oligonucleotide alignment program.** *Bioinformatics* 2008, **24:**713-714.
8. **SOAP: short oligonucleotide analysis package** [http://soap.genomics.org.cn]
9. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18:**1851-1858.
10. **MAQ** [http://maq.sourceforge.net]
11. **Bowtie: an ultrafast memory-efficient short read aligner** [http://bowtie-bio.sourceforge.net/index.shtml]
12. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, *et al.*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456:**53-59.
13. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, *et al.*: **The diploid genome sequence of an Asian individual.** *Nature* 2008, **456:**60-65.
14. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452**:872-876.
15. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, *et al.*: **DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.** *Nature* 2008, **456:**66-72.
16. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129:**823-837.
17. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.** *Nature* 2007, **448**:553-560.
18. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, Sidow A, Fire A, Johnson SM: **A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning.** *Genome Res* 2008, **18**:1051-1063.
19. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE: **High-resolution mapping and characterization of open chromatin across the genome.** *Cell* 2008, **132:**311-322.
20. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Bäckdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJ, Durbin R, Tavaré S, Beck S: **A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis.** *Nat Biotechnol* 2008, **26:**779-785.
21. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: **Highly integrated single-base resolution maps of the epigenome in Arabidopsis.** *Cell* 2008, **133**:523-536.
22. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES: **Genome-scale DNA methylation maps of pluripotent and differentiated cells.** *Nature* 2008, **454**:766-770.
23. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM: **Transcriptome sequencing to detect gene fusions in cancer.** *Nature* 2009, **458**:97-101.
24. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**:621-628.