Topic Introduction

Mapping Billions of Short Reads to a Reference Genome

Jui-Hung Hung and Zhiping Weng

Rapid development and commercialization of instruments that can accurately, rapidly, and cheaply sequence billions of DNA bases is revolutionizing molecular biology and medicine. Because a reference genome is usually available, the first bioinformatics challenge presented by the new generation of high-throughput sequencers is the genome mapping problem, where each read is mapped to a reference genome to reveal its location(s). An introduction to mapping algorithms, as well as factors that influence their results, is provided here.

INTRODUCTION

Myriad technologies have been developed to take advantage of high-throughput sequencers to sequence individual human genomes (Altshuler et al. 2010), an entire microbial community (microbiome) (Blow 2008), cancer genomes (Mardis and Wilson 2009), specific regions of a genome bearing epigenetic marks or bound by transcriptional factors (ChIP-seq) (Robertson et al. 2007), mRNAs (RNA-seq) (Marioni et al. 2008), and small silencing RNAs (Trapnell and Salzberg 2009).

Compared with a Sanger-style sequencer, which produces thousands of long sequences (also called reads) per run, each ~1000-nt long, the new sequencers manufactured by Illumina (HiSeq 2000) and Applied Biosystems (ABI SOLiD 4) can deliver hundreds of millions of short reads per run (paired-end reads up to 2×100 nt). Roche/454 sequencers that are already in production deliver >400-nt reads. Pacific Biosciences sequencers generate >1000-nt reads in early testing, although this technology is currently of medium throughput.

Because traditional sequence alignment algorithms such as BLAST and BLAT are too slow for such a task, many new algorithms have been developed in the past 3 yr for mapping short-read sequences to a reference genome (for review, see Trapnell and Salzberg 2009). Currently, the most widely used algorithms are all based on the Burrows–Wheeler (BW) transform of the reference genome. These include Bowtie (<http://bowtie.cbcb.umd.edu/>; Langmead et al. 2009), BWA (<http://maq.sourceforge.net/bwa-man.shtml>; Li and Durbin 2009), and SOAP2 (<http://soap.genomics.org.cn/>; Li et al. 2009b). These three algorithms have similarly good performance, although they differ in strategies to deal with mismatches and gaps.

THE BURROWS–WHEELER (BW) TRANSFORM

The BW transform is a technique originally developed for compressing large files and has been adapted to compress genome sequences. BW transform couples very well with a type of data structure called suffix/prefix trie that is designed for exact string-matching algorithms. Trie (the term comes from

From the Molecular Cloning collection, edited by Michael R. Green and Joseph Sambrook.

© 2017 Cold Spring Harbor Laboratory Press

Cite this introduction as *Cold Spring Harb Protoc*; doi:10.1101/pdb.top093153

retrieval) is a data structure that stores all of the suffixes/prefixes of a string, in our case, a reference genome sequence, to facilitate fast searching. Searching for a query sequence (a sequencing read) can be achieved by searching for the suffixes that have their prefixes exactly matched with the query. Because all suffixes are sorted, the search can be done very rapidly. To be exact, the time complexity of determining whether a query sequence has an exact match in a trie is linearly proportional to the length of the query and independent of the length of the reference genome. However, the memory required to store all suffixes of a genome is not feasible for a desktop computer. BW transform can compress the trie to occupy small memory footage without substantially increasing processing time.

The human genome is a text file with 3 billion As, Gs, Cs, and Ts, and its BW transform only requires ~2.2 Gb of memory (RAM), which is available in most modern desktop or laptop computers. BW transform-based algorithms align a query sequence one base at a time with the BW transform of a reference genome, with each base growth winnowing down the list of possible matches in the genome. Building on an exact string-matching algorithm, one of the main innovations in Bowtie, BWA, etc., is dealing with query sequences that do not map to the genome entirely and perfectly (due to sequencing errors or polymorphisms). The algorithms search for mismatches in the query sequence at positions of high likelihood of sequencing error so that a match in the genome can be located.

Although extremely fast, these algorithms can take hours for a typical data set (millions of reads against a mammalian genome) and need to be run on the user's computer, as opposed to a web-accessible program. When the data set is large, a computer cluster can be used so that the data set is split into jobs that can be distributed across multiple compute nodes. However, the reference genome should not be split (e.g., one chromosome per job); otherwise, a computationally extensive merging procedure will be required, because a read that maps to multiple locations of the genome may appear to map to a unique location after the split.

FACTORS THAT INFLUENCE MAPPABILITY

The percentage of reads that map uniquely to the genome is often taken as a major result of an alignment algorithm. In the best-case scenario, 85%–95% of the reads map uniquely. However, this percentage is affected by several factors independent of the algorithm.

1. *Extent of polymorphism.* If the underlying genome of the sample is evolutionarily distant from the reference genome, then even if both are from the same species, the number of polymorphisms can be high, especially in regions of the genome that are under low evolutionary selection pressure (e.g., intergenic and heterochromatin regions). To account for polymorphism, one can allow for a small number of mismatches (e.g., one mismatch per 36 bases). More accurately, if the genome has been genotyped (e.g., by the HapMap and 1000 Genomes Projects, in the case of the human genome), one can explicitly include the polymorphisms in the reference genome. For example, when only considering the two major alleles in polymorphic positions, one can concatenate two copies of the human genome, each with one allele at each position, as the input for the BW transform and indexing.
2. *Quality of sequence data.* Low data quality is often the culprit of low mappability. This can be due to poor library preparation or issues with sequencing. For example, a human DNA sample may be so heavily contaminated that many of the reads are from the genomes of other organisms. Low sequencing quality can be revealed with a box plot of sequencing-quality score at each position of the reads—a high-quality data set corresponds to consistently high scores at all positions, without much deterioration toward the 3' end.
3. *Lengths of DNA molecules being sequenced compared with size of reference genome.* For some experiments, the DNA molecules being sequenced may be shorter than the read length of the sequencer. For example, you may wish to investigate the binding sites of a DNA-binding protein in

high resolution by digesting away the flanking bases of the bound DNA. For a genome of N bases of random composition (i.e., no higher-order patterns such as repeats), reads shorter than $\log_4(N)$ bases are expected to map to the genome at least once purely by chance. By the same logic, all reads shorter than 16 bases are expected to map to the human genome by chance. Thus, if the binding sites are shorter than 16 nt, it will be difficult to pin down their genomic locations. In this case, one might consider sequencing longer DNA molecules and then improving the resolution by computationally identifying the embedded binding sites.

4. *Degree of sequence repetition in regions of the genome to which the reads map.* By definition, reads that map to repetitive regions of the genome map to many locations. Some alignment algorithms may not report all of the locations for the reads that map to too many locations in the genome; for example, because there are $\sim 10^6$ Alu elements in the human genome, reads that map to Alu are typically discarded. This problem can only be alleviated by increasing the read length or by using paired-end reads in the hopes that one of the pairs maps uniquely.

Existing sequencing technologies constantly strive for longer reads, and new technologies that can yield longer reads continue to be developed. Allowing a small number of mismatches is no longer the best strategy for mapping long reads, because long reads with several mismatches and indels (insertions and deletions) can still be mapped uniquely to the reference genome. Algorithms such as BWA-SW (<http://maq.sourceforge.net/bwa-man.shtml>; Li and Durbin 2010) are slower than short-read aligners (e.g., Bowtie, BWA, and SOAP2) and yet are much faster than the traditional alignment algorithms such as BLAST and BLAT. In the foreseeable future, there will be a continuum of algorithms from which to choose, depending on the sequencing technology.

The output of a mapping algorithm becomes the input data for subsequent analysis. An example of downstream analysis is given in Protocol: **Identifying Regions Enriched in a ChIP-seq Data Set (Peak Finding)** (Hung and Weng 2017a). The mapping results can also be visualized with tools such as the UCSC Genome Browser (Fujita et al. 2010) (see Protocol: **Visualizing Genomic Annotations with the UCSC Genome Browser** [Hung and Weng 2016a]). Most alignment algorithms can produce outputs in the SAM format (or its binary equivalent, the BAM format) (Li et al. 2009a), which can be directly uploaded to the UCSC Genome Browser. With the help of SAMtools (Li et al. 2009a) and BEDTools (Quinlan and Hall 2010) (see Box 1), users can call variants (e.g., single-nucleotide polymorphisms [SNPs]) and pile reads and transform SAM/BAM to BED format for inputting to other software or websites (see Protocol: **Mapping Short Sequence Reads to a Reference Genome** [Hung and Weng 2017b] for a pipeline based on these two tools).

Because of the rapidly evolving nature of sequencing technologies and mapping algorithms, the traditional means of information dissemination via journal and book publications can be too slow paced. SeqAnswers (<http://www.seqanswers.com>) is an online forum followed by users at all levels of experience in the next-generation sequencing community. There are many useful threads on existing software for mapping short reads and subsequent data analysis and visualization. Novice users can find fixes to most errors encountered while running these programs, experienced users can get advice on adjusting parameters to best suit their biological problems, and algorithm developers can provide their code for testing before publication and solicit feedback.

BOX 1. BEDTools AND SAMtools

The BEDTools package consists of utilities to use primarily for manipulating BED files, including functionality for tasks such as finding overlaps and computing coverage. It also has some operations that take SAM/BAM files as input. SAMtools is a set of command-line utilities that manipulate SAM and BAM files. SAMtools can perform sorting, merging, and indexing, and generate alignments in a per-position format, which is an intuitive way of revealing possible single-nucleotide polymorphisms (SNPs).

Familiarity with both SAMtools and BEDTools is strongly recommended when dealing with high-throughput sequencing data (see Introduction: **Data Formats in Bioinformatics** [Hung and Weng 2016b]).

REFERENCES

- Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, De La Vega FM, Donnelly P, Egholm M, et al. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.
- Blow N. 2008. Metagenomics: Exploring unseen communities. *Nature* **453**: 687–690.
- Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, et al. 2010. The UCSC Genome Browser database: Update 2011. *Nucleic Acids Res* **39** (Database issue): D876–D882.
- Hung J-H, Weng Z. 2016a. Visualizing genomic annotations with the UCSC Genome Browser. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot093062.
- Hung J-H, Weng Z. 2016b. Data formats in bioinformatics. *Cold Spring Harb Protoc* doi: 10.1101/pdb.top093211.
- Hung J-H, Weng Z. 2017a. Identifying regions enriched in a ChIP-seq data set (peak finding). *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot093187.
- Hung J-H, Weng Z. 2017b. Mapping short sequence reads to a reference genome. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot093161.
- Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.
- Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589–595.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
- Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. 2009b. SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **25**: 1966–1967.
- Mardis ER, Wilson RK. 2009. Cancer genome sequencing: A review. *Hum Mol Genet* **18**: R163–R168.
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**: 1509–1517.
- Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.
- Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**: 651–657.
- Trapnell C, Salzberg SL. 2009. How to map billions of short reads onto genomes. *Nat Biotechnol* **27**: 455–457.

Cold Spring Harbor Protocols

Mapping Billions of Short Reads to a Reference Genome

Jui-Hung Hung and Zhiping Weng

Cold Spring Harb Protoc; doi: 10.1101/pdb.top093153 originally published online August 29, 2016

Email Alerting Service

Receive free email alerts when new articles cite this article - [click here](#).

Subject Categories

Browse articles on similar topics from *Cold Spring Harbor Protocols*.

- [Alignment of Sequences](#) (46 articles)
- [Alignment of Sequences, general](#) (24 articles)
- [Bioinformatics/Genomics, general](#) (192 articles)
- [Computational Biology](#) (100 articles)
- [DNA Sequencing](#) (97 articles)
- [Genetic Variation](#) (86 articles)
- [Genetics, general](#) (375 articles)
- [Genome Analysis](#) (191 articles)
- [High-Throughput Analysis, general](#) (155 articles)
- [Multiple Sequence Alignment](#) (23 articles)

To subscribe to *Cold Spring Harbor Protocols* go to:
<http://cshprotocols.cshlp.org/subscriptions>