HHS Public Access

Author manuscript

Proc IEEE Inst Electr Electron Eng. Author manuscript; available in PMC 2018 March 01.

Published in final edited form as:

Proc IEEE Inst Electr Electron Eng. 2017 March ; 105(3): 436–458. doi:10.1109/JPROC.2015.2455551.

Short Read Mapping: An Algorithmic Tour

Stefan Canzar and **Steven L. Salzberg**

Abstract

Ultra-high-throughput next-generation sequencing (NGS) technology allows us to determine the sequence of nucleotides of many millions of DNA molecules in parallel. Accompanied by a dramatic reduction in cost since its introduction in 2004, NGS technology has provided a new way of addressing a wide range of biological and biomedical questions, from the study of human genetic disease to the analysis of gene expression, protein-DNA interactions, and patterns of DNA methylation. The data generated by NGS instruments comprise huge numbers of very short DNA sequences, or 'reads', that carry little information by themselves. These reads therefore have to be pieced together by well-engineered algorithms to reconstruct biologically meaningful measurments, such as the level of expression of a gene. To solve this complex, high-dimensional puzzle, reads must be mapped back to a reference genome to determine their origin Due to sequencing errors and to genuine differences between the reference genome and the individual being sequenced, this mapping process must be tolerant of mismatches, insertions, and deletions. Although optimal alignment algorithms to solve this problem have long been available, the practical requirements of aligning hundreds of millions of short reads to the 3 billion base pair long human genome have stimulated the development of new, more efficient methods, which today are used routinely throughout the world for the analysis of NGS data.

Index Terms

sequence alignment; string matching; suffix trees; Burrows-Wheeler transform; DNA sequencing

I. Introduction

SEQUENCE alignment has a long history in molecular biology and genetics, dating back to the first decoding of protein sequences in the 1960s. A pairwise alignment of two proteins identifies subsequences of amino acids that are position-wise similar, respecting the order in which they occur in the two proteins. The original goal of alignment was to uncover the evolutionary relationship between two proteins: positions that match represent the sequence of a common ancestor, and positions that differ represent mutations that occurred in one or both proteins since the divergence of their most recent common ancestor.

The first elegant algorithm to compute the optimal alignment of two sequences, in some sense the archetype for many alignment algorithms, was proposed by Needleman and Wunsch in 1970 [92]. It is based on the *dynamic programming* paradigm, which relates the

optimal alignment of two sequences to the optimal alignment of two shorter substrings and orders the resulting sub-problems in a way that avoids the repeated computation of identical nodes in the recursion tree While such a *global alignment* is meaningful if two sequences are similar in their entirety, it might fail to discover local similarities. In 1981, Smith and Waterman [109] proposed an adaption of the Needleman-Wunsch algorithm that is able to find such *local alignments*, with a computational cost that is $O(nm)$, where n and m are the lengths of the two input sequences.

The biological and medical relevance of comparative sequence analysis emerged quickly as biological sequence data became more widely available. This was exemplified by R. Doolittle's discovery [26] that the *sis* oncogene, a gene from a simian sarcoma virus, was highly similiar to a previously known gene, platelet-derived growth factor (PDGF), which is involved in controlling cell growth. This very early alignment result, in 1983, led to the realization that oncogenes might be abnormally expressed versions of normal growth factors.

Sequence databases grew rapidly in the 1980s (GenBank[1] was started in 1982), and as a result, performing a full dynamic programming comparison of a query sequence to every known sequence soon became computationally very costly. The need to efficiently align a query sequence against a database motivated a heuristic algorithm proposed by Wilbur and Lipman [117] and refined and implemented in the FASTA program suite [94]. The basic principle underlying this algorithm is to *exclude* large parts of the database from the expensive dynamic programming comparison by rapidly identifying candidate sequences that share short stretches (k-tuples) of highly similar sequence with the query. FASTA was followed by the BLAST program [7], which achieved additional speed advantages and added a very useful feature: an estimate of the statistical likelihood that each matching sequence had been found by chance. BLAST soon became the most widely used search program for biological sequence databases, and the two primary publications (Altschul *et al.* 1990 [7] and 1997 [8] now have over 90,000 citations. BLAST's combination of speed and sensitivity has made it the workhorse of biological sequence searches for over twenty years.

As whole-genome sequencing accelerated in the late 1990s and early 2000s, other alignment challenges emerged. For example, microbial genomicists began sequencing multiple strains of the same bacteria, motivating the need to align two complete genomes to one another. This requirement, which later included far larger insect, plant, and animal genomes, presented a requirement that BLAST did not address. The first solution to this problem, the MUMmer program [24], used the suffix tree data structure, which had not previously been used for DNA sequence alignment. Suffix trees have the advantage that they only occupy linear space (in terms of the length of the input sequences) and they can be searched for exact matches in linear time.

Other methods for fast indexing also emerged at this time, such as the BLAT [50] and SSAHA [93] programs, both of which (similarly to FASTA) used a hash index based on k-tuples. These methods focused on hashing into vertebrate genomes (especially human) using

[1] http://www.ncbi.nlm.nih.gov/genbank/

an index that could fit into RAM and therefore allowed much faster searching than BLAST, although neither provided probability estimates in their output.

The advent of ultra-high-throughput next-generation sequencing (NGS) technology in 2007 presented major new challenges for alignment. The first sequencers from Solexa (later Illumina) were able to generate, in a single run, tens of millions of very short DNA sequences, or "reads." Initially, read lengths were only 25 base pairs (bp), which although short is nonetheless long enough to map approximately 79% of the human genome [69]; i.e., 79% of the genome is composed of 25-bp sequences that occur only once. Because 25-bp reads can be aligned uniquely to the genome, this technology gained immediate use as a new method for studying human genetic variation and disease. Read lengths subsequently increased, and have now reached 150 bp in the highest throughput sequencing machines.

The rapid improvement in sequencing capacity has dramatically outpaced Moore's Law, according to which computing speed doubles every two years. Today, an Illumina HiSeq 2500 in rapid mode can produce 120 gigabases, the equivalent of 40 human genomes, in just 27 hours[2], representing a speedup of approximately 500,000-fold since 2001. This ultrahigh-throughput technology and the accompanied dramatic reduction in cost have opened the door to a remarkable variety of biological and biomedical applications. Besides determining the complete genomes of numerous species [100], it allows the resequencing of large cohorts of individuals and has spurred efforts to sequence thousands of individual humans [1]. Illumina' s recent HiSeq X Ten platform can sequence more than 18,000 human genomes per year at a marginal cost of just $1000 per genome.[3] Additional new techniques have led to the widespread use of sequencing to study gene expression (RNA-seq [88], [73], [21]), protein-DNA interactions (ChIP-seq [46]), and DNA methylation (MeDIP-seq [27]).

These advances in sequencing throughput have presented a major challenge for alignment algorithms. Because CPU speeds have been largely constant, the time required merely to align reads to the genome has increased enormously. This in turn has led to new efforts to improve alignment algorithms. In this perspective, we describe these efforts and the current state of the art in NGS alignment algorithms.

Since 2007, computational biologists have developed more than 70 read mapping tools [36] that try to tackle these challenges. This survey elucidates the algorithmic underpinnings that paved the way to extremely efficient read mapping software, which today is used throughout the world for next-generation sequence alignment. We do not attempt to evaluate the mapping tools themselves, but rather to describe their main algorithmic strategies and their strategies for making speed versus accuracy trade-offs from an *algorithm engineering* point of view.

A. The Short Read Mapping Problem

1) The Data—Each next-generation sequencing platform implements a different technology, but the workflow can be illustrated by a brief summary of the Illumina method

[2]http://www.illumina.com/systems/hiseq_2500_1500/performance_specifications.ilmn
[3]http://www.illumina.com/systems/hiseq-x-sequencing-system/system.ilmn

(see Figure 1). First, genomic DNA is sheared into *fragments* that can be size-selected to obtain a template library of uniformly short DNA fragments. Platform-specific adapter sequences are then attached to both ends of the fragments, which allow them to be attached to a solid surface. Through a local PCR amplification step, many copies of each template molecule are produced in a tightly clustered location on the surface. These molecules emit a signal during the sequencing reaction that is strong enough to be detected by the optical system of the instrument. During sequencing, every molecule is cyclically extended by a single base, and all clusters are read in parallel before the cycle is repeated. A *base-calling* algorithm determines individual bases from raw signal data, to which quality scores are assigned. Since the precise meaning and range of quality scores is platform dependent, here we assume that a higher quality score indicates a lower likelihood of the base call being incorrect. In *paired-end* sequencing, both ends of the original DNA fragment are sequenced, and the instrument keeps track of this pairing information.

2) Solving the Inverse Problem—The first step in the analysis of NGS data is to determine where in the genome each of the short reads originated (Figure 1, step D). The obvious criteria for mapping such a read to a genomic location is sequence similarity. Due to sequencing errors and genuine differences between the reference genome and the sequenced organism, a read might not match its corresponding location in the reference genome exactly. We therefore need an alignment method that permits some number of mismatches, insertions, and deletions. Although older alignment software such as BLAST [7] or BLAT [50] can be used to map short reads to a reference genome, those methods are simply too slow. The sheer amount of data, sometimes comprising billions of short reads that have to be aligned to a large (e.g. mammalian) genome, has required innovative new algorithms that run orders of magnitude faster, at least for the specialized problem of short-read alignment to a fixed reference genome. In order to be most useful, these algorithms should also exploit additional information associated with the experimental protocol, such as read pairing information, and take into account technology-specific error profiles.

In the case of paired-end reads, we expect the two *mates* to map at a distance and orientation that is consistent with the library fragment length and the relative sequencing direction, respectively. In the formalization of the read mapping problem below, however, we ignore pairing information on reads and treat the mates as independent single-end reads. In Section VII-A we describe strategies to incorporate the additional constraints imposed by paired-end information.

Because the DNA strand from which a read originates is unknown, either the read sequence itself or its reverse complement may be mapped to the reference genome. In the problem formulation below and in our description of the algorithms, we will not make a distinction between a read and its reverse complement and simply refer to them both as *reads.*

B. Read Mapping as String Matching

Given a reference genome G and a set of reads $\mathscr{R}$, in the *read mapping* problem we want to determine for every read $R \in \mathscr{R}$ its origin in G. The reference G and the reads in $\mathscr{R}$ can be modelled as *strings*, (mathematical) sequences of symbols from alphabet $\Sigma = \{A, C, G, T\}$.

The symbols in Σ represent the four nucleotides *adenine, cytosine, guanine*, and *thymine.* To conform to biological usage, we use the terms *string* and *sequence* synonymously. In contrast to a *subsequence* of a string *S*, however, a *substring* denotes a *contiguous* subsequence of *S*.

1) Exact String Matching—The origin of a read *R* might be detected through *sequence identity* of *R* to a substring in *G*. Because multiple occurrences of *R* in *G* cannot be distinguished without incorporating additional information, we are interested in finding *all* occurrences of *R* in *G*.

Definition 1: Let Σ be a finite alphabet of size $|\Sigma| = \sigma$. Given a *text* $T \in \Sigma^*$ of length $|T| = n$ and a *pattern* $P \in \Sigma^*$ of length $|P| = m$, the *exact string matching* problem is to find all occurrences of *P* in *T*.

Here, the pattern *P* corresponds to a single read in ℛ, the reference genome is represented by *T*, and $\sigma = 4$. Since ℛ typically comprises tens to hundreds of millions of reads and $|T|$ is huge compared to the read length, classical matching algorithms like Boyer-Moore [13] and Knuth-Morris-Pratt [53] are impractical for the read mapping problem. For every pattern *P*, these algorithms require time proportional to the $|T|$ to find occurrences of *P* in *T*.

Indexed string matching builds an auxiliary data structure (an index) for *T* such that occurrences of pattern *P* can be found efficiently without scanning the complete text *T*. Such an index structure is indispensible in the context of mapping short reads to a genome, because it speeds up the mapping of many tens of millions of reads and thereby marginalizes the cost of constructing the index. In addition, the reference genome (*T*) is static and thus the cost of maintaining the index is low.

Suffix trees [9] and *suffix arrays* [84] are classical *full-text* index structures, the latter being, as an array of integers specifying the lexicographical order of the suffixes of *T*, a space efficient alternative to suffix trees. They permit (almost) optimal $\mathcal{O}(m+occ)$ and $\mathcal{O}(m+\log n+occ)$ search time, respectively, where *occ* is the number of occurrences of *P* in *T*. Because search time is a function of the length of the pattern *P* rather than the genome, these represent a much faster way to map short reads. The initial barrier to their widespread use in computational biology was their space requirement. Both data structures require $\Theta(n \log n)$ bits and are thus asymptotically larger then the text itself, which needs $n[\log \sigma]$ bits. With $\sigma = 4$ and $n > 3 \cdot 10^9$ for the human genome, the resulting demand in memory resources can be prohibitively large, in particular when the index is to be stored in main memory to allow fast access.

Although several attempts were made to reduce the space requirement of index structures while preserving fast search functionality (e.g. [49]), it was not until Ferragina and Manzini developed their full-text index [33], based on the Burrows-Wheeler transform, that anyone had an index that was comparable in size to the text itself. They introduced the first *self index*, which takes space proportional to the *compressed* text, provides fast searching, and contains sufficient information to reproduce the text and thus can replace the text. In essence, the proposed *FM-index*) requires space close to the desired entropy bound for

constant-sized alphabets and allows text search with almost optimal time complexity. Theoretical concepts underlying this exciting invention and its further developments are surveyed in [91], their engineering aspects are addressed in [32].

2) Approximate String Matching—Alignment methods that only allow exact matches of reads to the genome are, for all practical purposes, useless. Real data has two significant sources of mismatches: sequencing errors and true variation. Sequencing errors vary with the technology of sequencing; the current error rate of high-throughput Illumina sequencers is less than 0.5%. True variation refers to differences between any human subject (if we are aligning human sequence) and the reference genome. In the human population, the overall variation has been estimated at approximately 0.1% (1 base per 1000), but this varies greatly depending on the background: the reference human genome is a European male, and the genomes of individuals with different ancestry have much higher rates of variation. To be useful, aligners must be able to tolerate at least as much variation as is found in the population.

More formally, any practically useful algorithm must be able to to find all approximate occurrences of a read; i.e., substrings of T whose similarity to P exceeds a specific threshold, or whose distance is less than a similar threshold. The two most commonly used distance functions are *Hamming distance* [99] and *Levenshtein* or *edit distance* [62]. While Hamming distance simply counts the number of substitutions implied by mismatched symbols of two strings at the same position, edit distance additionally accounts for inserted and deleted symbols (*indels*) in one string with respect to the other.

Definition 2: Given text T and pattern P, the *k-mismatch problem* asks for all $|P|$-length substrings of T within Hamming distance k, i.e. that match at least $|P| - k$ characters in P.

In contrast, a similarity measure is based on a scoring scheme that assigns a higher score to matched symbols than to mismatched symbols, insertions and deletions, and in the context of short read mapping ideally incorporates base quality scores. A set of insertions and deletions implying matches and mismatches between two sequences is represented by an *alignment* (Figure 2). A (global) alignment of two DNA sequences inserts a dash symbol "–" $\notin \Sigma$ into either string to obtain two strings of the same length. This alignment establishes a positionwise correspondence between nucleotides in the augmented strings. Nucleotides aligned to dash symbols represent either an insertion in one string or a deletion in the other. The *score* or *value* of an alignment is the sum of its position-wise scores.

The (weighted) edit distance or *similarity* (score) between two strings is defined by an optimal alignment that minimizes the number (or total weight) of edits (substitutions, insertions, deletions) or that has a maximum score, respectively. Assigning a score of 1 to matching nucleotides, −3 to mismatches, and −2 to indels, the alignment in Figure 2 is optimal and implies a similarity score of 1. The (unit cost) edit distance of the two sequences is 5.

Definition 3: Given text T and pattern P, the *k-errors problem* asks for all substrings of T within edit distance k.

Algorithms that compute an optimal alignment for two sequences of length n and m, based on the original Smith-Waterman [109] paradigm, run in time $\mathcal{O}(nm)$ for typical scoring schemes. Variants that simply bound the number of errors k (Defintions 2 and 3) allow for a faster $\mathcal{O}(kn)$ solution.

The deletion or insertion of several consecutive nucleotides by a single mutational event can by captured by the concept of a *gap*. A gap is a maximal sequence of dash symbols in one string, that is penalized by a weight that is a function of the length of the gap. In the context of short read mapping, the most commonly used model gap penalties are *affine* functions of the form $\alpha + q\beta$, where q denotes the length of the gap. In the case of affine gap weights, the optimal alignment can be found within the same time bound of $\mathcal{O}(nm)$ as for scoring schemes without gaps.

Finding *all* approximate occurrences of a short read in a reference genome requires a modification of the base condition in the dynamic programming algorithm to allow the read to start at any position in the genome. Although this will solve the alignment problem, it is clearly infeasible to spend $\mathcal{O}(nm)$ time, or even $\mathcal{O}(kn)$ in case of an error bound k, to match each of hundreds of millions of reads approximately to a large reference genome.

C. Overview

The static nature of most genomes (human, mouse, and other reference genome assemblies) allows the use of a precomputed index to speed the search process. Indexing a text to support approximate matching queries is a challenging problem, primarily because of their space requirements. Classical full-text indexes such as suffix trees and suffix arrays require, in the most optimized implementations, approximately 3 bytes per base or 36 GB of memory for a mammalian genome.

In Section II we present a class of methods that attempt to quickly eliminate large parts of the genome where the read does *not* match. *Filtering algorithms* typically rely on unaltered parts of the read that do not contain any substitutions or indels and reduce the search for inexact-matching regions to the exact matching problem, which can be solved efficiently through a lookup in a hash table. Only regions that match a sufficiently large part of the read exactly must be verified by a more expensive approximate matching algorithm.

In Section III we describe how state of the art software tools apply the FM-index to map short reads to a reference genome. The FM-index has originally been designed for the exact string matching problem and as such supports filtration based approaches. Furthermore, it allows to navigate through occurrences of related patterns and thus facilitates an inexact search of the read.

Section V discusses adaptations of algorithms and data structures that are necessary to support the massive parallelism provided by a graphics processing unit (GPU). Figure 16 organizes methods described in Sections II–V in a tree of algorithmic design decisons.

Section VI summarizes efforts made by read mappers to identify the true origin of a read among its approximate occurrences and to provide an estimate on the probability that the true origin was found.

In Section VII we address aspects of the read mapping problem that are not captured by an approximate string matching model. We conclude in Section VIII.

II. Hashing-Based Methods

The alignment challenges posed by NGS sequencing technology were initially addressed by methods that used the seed- and-extend strategy (e.g. [8], [82]). This approach assumes that in the true alignment of a read, parts of the read, the *seeds*, match the reference without errors. If these seeds are long enough to match in only a small number of places, the aligner can collect all the seed matches and then extend each of them. seeding quickly discards large parts of the genome where the pattern does *not* match according to a given error model. For example, if neither ACA nor AGT occur in a given text, then ACAAGT with at most one mismatch cannot exist either. Only regions around the seeds need to be evaluated with a more costly approximate string matching algorithm.

Assuming that seeds have length k, we can locate k-mers shared between the reads and the genome through a hash table index. Aligners either build a hash table of subsequences found in the genome and scan the reads for matching k-mers, or the other way around. The hash table is accessed through a k-mer and stores the positions where the corresponding subsequence occurs in the genome or the read. Tools that index the reads include Eland (AJ Cox, Illumina, unpublished), mrFAST/mrsFAST [6]/[40], MAQ [65], RMAP [107], ZOOM [71], SeqMap [45], and SHRiMP [98]. Tools that build a hash index of the genome included SOAP [67], Novoalign[4], Mosaik [61], SRmapper [37], Stampy [80], BFAST [43], SHRiMP2 [22], Hobbes [3], and FastHASH [119]. For a more comprehensive list of sequence aligners, see the “Short-Read Sequence Alignment” section at http://en.wikipedia.org/wiki/List_of_sequence_alignment_software and http://wwwdev.ebi.ac.uk/fg/hts_mappers/.

Figure 3 depicts a typical workflow for a hash-based mapper. First, it extracts k-mers from a query read by applying the same seed template as was used to index the reference genome. In the example shown in Figure 3, the 5-mers are obtained from (contiguous) substrings of length 5. Some methods, including BLAST and Subread [70], only consider an informative subset of the resulting k-mers in subsequent steps. Second, the hash table is queried to locate all occurrences of the query k-mers in the reference genome. Third, genomic regions containing single or multiple seed matches are aligned using an approximate alignment method that attempts to extend the seeds to a full read alignment, using a scoring function that is particular to each mapper.

Next we discuss different seeding strategies to select k-mers that have to be matched between read and reference genome. We give examples of additional filters that can further

[4] http://www.novocraft.com

reduce the number of candidate regions to be verified in Section II-B. Algorithms employed during the final verification step depend on the error model and are reviewed in Section II-C.

A. Seed Template Design

The "guessing" of an exact matching seed is usually guided by a filtering criterion that relies on parts of the read that must remain unaffected by the edit operations allowed by a specific error model. Seed templates specify these potentially unaffected parts (subsequences) of the reads and are important for the efficiency of filtering-based methods. A low *selectivity* gives rise to a large number of seeds that have to be extended by computationally intensive dynamic programming algorithms (see Section II-C). A low *sensitivity* will result in many approximate matches that are missed by the filter, which might include the true origin of a read. In contrast, a *lossless* filter is guaranteed to not discard any true occurrence.

1) Lossy Filtration—In the simplest case, as typified by the BLAST program [7], the algorithm creates seeds from consecutive sequences of characters that must occur in both the read and the genome, similar to the six matching seeds in Figure 3. Requiring a long sequence to match exactly might cause the correct alignment to be missed, while a short sequence may have an excessive number of false hits that will not extend to full alignments. The highly repetitive structure of many genomes may greatly increase the cost of overly short seeds, which can match in literally millions of locations on large plant and animal genomes.

In PatternHunter [82], Ma *et al.* addressed this dilemma by introducing *spaced seeds* that require a *non*-consecutive sequence of characters to match. A spaced seed *template* can be represented by fixed-length words over the alphabet {0,1}, where 1's indicate positions that are required to match and 0-positions are ignored. Figure 4 depicts a seed match implied by template 1101011. In this model, consecutive templates as used by BLAST correspond to templates of the form 1^k.

If appropriately designed, a spaced seed has a higher chance of hitting a true approximate match on the genome while producing fewer random hits. Besides spanning a larger interval on the read with the same number of positions required to match, the main property responsible for this advantageous behavior is the increased independence of a spaced seed template and its shifted copies. Simply speaking, a spaced seed template and its shifted instance share fewer positions that are required to match and thus together capture a wider range of mismatch scenarios. Notice that spaced seeds seek for parts of the read that match with small Hamming distance, but do not account for insertions and deletions.

Again, increasing the selectivity of a spaced seed by requiring a higher number of positions to match usually comes at the cost of a lower sensitivity. An alternative scheme was proposed in [110] and implemented in PatternHunter II [66]. It uses *multiple* (spaced) seeds in *disjunctive* form, that is, a seed must match under at least one template. Intuitively, the loss in sensitivity resulting from an increased number of positions required to match is compensated by alternative templates. On the downside, multiple seeds multiply the computational effort necessary to detect matching seeds. BFAST [43], for example, aims at

determining a small set of seed templates by greedily adding a seed template that maximizes the gain in sensitivity.

SEME [20], on the other hand, is based on a single consecutive seed template (similar to BLAST), which is however not simply shifted along the read but skips a certain number of bases. Intuitively, the idea is similar to the one underlying spaced seeds: two consecutive seeds with less overlap capture a wider range of error scenarios. The sequential seed mapping strategy of SEME, which stops as soon as the first hit is encountered, leverages this effect. SEME does not use a hash index to map the seeds but finds (exact) occurrences on the genome through a tailored binary search in a sorted list of integers representing 32-mers from the genome.

The methods discussed so far identify candidate mapping locations based on individual seed matches between read and genome, that is, seeds are treated *disjunctively.* In contrast, the *seed-and-vote* scheme proposed in [70] relies on multiple seeds that *jointly* define a candidate region. The idea is to use the *number* of exactly matching (consecutive) seeds as a proxy for the edit distance. Using the expectation that a small edit distance should yield more contiguous stretches of exact matches in the corresponding alignment, the Subread program [70] allows all the seeds of a read to vote for the best mapping location. The region receiving the highest number of votes is identified as the (one) final mapping location, which is not guaranteed to lie within a certain error threshold.

Similarly, NextGenMap [104] and SHRiMP [98] consider candidate regions that are supported by a sufficiently large number of seed matches. The corresponding threshold is inferred in NextGenMap from the distribution of the number of mutually consistent seed matches. Simply speaking, reads from repetitive regions necessitate the verification of a higher number of candidates compared to reads from unique regions.

2) Lossless Filtration—Lossless filtration algorithms attempt to remove potential regions from consideration without losing any true matches. ZOOM [71], for instance, designs minimum cardinality sets of spaced seeds needed to achieve full sensitivity for a given read length, allowed number of mismatches, and number of positions required to match. Most of the methods in this section, however, either apply the pigeonhole principle or rely on a lower bound on the number of required seed matches. While the former naturally applies to the k-mismatch problem, the lower bound criterion is used to identify candidate regions for k-error matches.

a) Seed Design by the Pigeonhole Principle: Methods that are guaranteed to detect all k-mismatch occurrences of a read often employ a combinatorial argument related to the pigeonhole principle to design (multiple) spaced seed templates. For example, Eland, MAQ, SeqMap, and SOAP subdivide the read into four parts of roughly the same length, such that two mismatches are guaranteed to leave two parts unaffected. MAQ applies this strategy only to the first, most reliable, 28 bases of the read. Six (spaced) seed templates cover all possible combinations of two error-free parts. For 8 bases, for example, the templates are

11110000 00111100
00001111 11001100
11000011 00110011

Alternatively, RMAP originally used the Baeza-Yates and Perleberg filtration method [10] and partitioned the read into $k + 1$ contiguous templates, such that an approximate match with at most k mismatches is guaranteed to contain one error-free part. A similar partitioning strategy is applied in one of the two filters implemented in RazerS 3 [116]. On the negative side, a higher number k of allowed mismatches shrinks the minimum length $\left\lfloor \frac{m}{k+1} \right\rfloor$ of the contigous templates required to achieve 100% sensitivity, yielding a higher number of false positive hits that have to be evaluated subsequently. A more selective (and relatively large) set of spaced seed templates is employed by an updated version of RMAP, which obtains full sensitivity to 2 mismatches in the first 32 bases of the read.

If the read is partitioned into $s < k + 1$ non-overlapping seeds, their increased length comes at the cost of losing the guarantee that an error-free seed exists. Nevertheless, by the (generalized) pigeonhole principle each approximate occurrence of the read with at most k errors contains a seed with at most $\left\lfloor \frac{k}{s} \right\rfloor$ errors. Such an *approximate occurrence* can be found by *backtracking* search in the Masai program [106] (see Section III-D2).

In contrast, Hobbes2 [51] selects $k + 2$ non-overlapping seeds and infers candidate regions from two or more seed matches. A bounded number of indels are accounted for by allowing gaps between two matched seeds and by expanding the candidate region.

Because the repetitiveness of equal-length seeds can vary considerably for a given read, the GEM mapper [85] and Hobbes(2) [3], [51] determine the partitioning into seeds adaptively, keeping the number of matches of each seed or the total number of matches small, respectively. From the resulting number of seeds, GEM infers the maximal number of errors that has to be allowed to guarantee full-sensitivity (see Section III-D2).

b) Lower Bound on Seed Matches: Following a similar intuition as the seed-and-vote paradigm (Section II-A1) more rigorously, Rasmussen *et al.* [96] proposed a filter that is based on multiple seeds that *conjunctively* indicate a candidate mapping location. It is based on the observation that a pattern P and an approximate occurrence of P in the text (genome) with (Levenshtein or Hamming) distance k have at least $|P| + 1 - q(k + 1)$ substrings of length q (q-grams) in common (q-gram lemma) [47]. This bound is optimal. Intuitively, each of the k errors affects at most q of the $|P| - q + 1$ many q-grams of P. For example, the single mismatch between strings in Figure 5 affects the three 3-grams marked by the red lines and leaves $8 + 1 - 3.2 = 3$ 3-grams indicated by blue lines unchanged.

The q-gram lemma was first used in the database searching algorithm QUASAR [15] and was later generalized to gapped q-grams [16] and families of gapped q-grams [55].

In contrast to the spaced seeds discussed above, a filter criterion that builds upon this lower bound on the number of q-grams takes into account insertions and deletions by which the read may differ from its approximate match. Relying on multiple seeds that flag a candidate region of approximate occurrence is a technique commonly applied in long-read alignment; e.g., see BLAT [50], AGILE [87] and Section III-F.

Rasmussen *et al.* [96] show that the search for candidate regions sharing a sufficient number of q-grams (the q-gram lemma) can be restricted to specific parallelograms in the dot plot of the sequences (Figure 6). After building an index containing the locations of q-grams, these parallelograms can be found using a sliding window technique. If the number of diagonals in the edit matrix that are summarized into bins is chosen appropriately, the update of the number of common q-grams falling into certain bins as the window proceeds can be performed using fast bit-operations. Such a scheme is implemented, for example, in RazerS [115] and its successor RazerS 3 [116].

FastHASH [119], on the other hand, uses a lower bound on the number of consecutive seeds that are required to satisfy an adjacency condition. Seeds that are adjacent in the read are required to match to adjacent positions on the reference, where the adjacency condition is slightly relaxed if indels are allowed. The minimum number of adjacent seed matches that define a candidate region is determined by the maximal number of errors allowed and the number of seeds into which the read is divided (using the pigeonhole principle). Compared to the set of candidate regions implied by independent seed matches, the adjacency requirement eliminates false positive hits only and thus achieves full sensitivity with respect to a given edit distance threshold. To reduce in turn the comutational cost of the adjacency filter, which involves a binary search on the list of seed locations on the reference, FastHASH chooses “cheap” seeds in increasing order of the number of matches on the reference genome. The authors of FastHASH argue that this simple sorting scheme exhibits a better cost-efficiency trade-off than the dynamic program applied by Hobbes (Section II-A2a) to determine an optimal division of reads. The latter involves a significant number of hash-table accesses which potentially outweigh the reduction in k-mer query costs.

B. Additional Filters

Since the verification of a large number of candidate regions by a dynamic programming scheme (Section II-C) can be computationally expensive, GASSST [97] and Hobbes [3] apply additional filters to further eliminate false positive seed hits prior to their extension. Two types of filters in GASSST provide lower bounds on the edit distance between a read and a potential mapping location on the reference. If the lower bounds exceed a given error threshold the seed hit is discarded. The *frequency vector filter*, introduced in [113], eliminates seed hits that, together with flanking regions, exhibit too much deviation in nucleotide frequency from the genomic region. A similar filter is applied by Hobbes [3]. Second, GASSST utilizes, similarly to the PASS aligner [18], precomputed alignment scores of all possible words of a given (short) length to derive a lower bound on the alignment score of a seed and its flanking regions. To support this filtering mechanism, the hash table additionally stores flanking regions of k-mers. Notice however that these filters are based on a simple, unweighted edit distance. Hobbes additionally allows one to compute a lower

bound on the number of mismatching characters (Hamming distance) through bitwise operations.

C. Seed Extension

Candidate regions returned by, for example, a pigeonhole filter or a *q*-gram based filter (Section II-A2) and which have optionally passed additional filters (Section II-B) have to be verified to contain an approximate occurrence of the read according to a specified error model. More specific distance measures usually allow faster algorithms to be applied.

1) Hamming-Based Extension—If an approximate match is defined with respect to a maximal number of mismatches (*k*-mismatch problem) a genomic region can be verified simply through a linear scan counting the number of mismatches, as is done e.g. in RazerS and RazerS 3. At essentially no additional cost MAQ minimizes the sum of quality scores of mismatched bases to obtain “more reliable” alignments. Alternatively, bitwise operations (mainly XORs) between an appropriate binary representation of read and genomic region can be employed to obtain a bit vector that indicates mismatches. The number of 1 bits in this vector and thus the number of mismatches can be determined through a look-up table (e.g. SOAP) or using a technique proposed by Warren [114] (e.g. ZOOM, RMAP). This bit-parallel scheme does not directly account for quality scores. RMAP incorporates quality values at a lower resolution by treating low quality bases, defined in terms of a threshold, as wild cards always inducing a match.

2) Alignment-Based Extension—A very restricted edit distance is considered in SEME [20]. Relying on a low indel error rate of the Illumina sequencing platform and a small number of indel polymorphisms between reference and sequenced subject, it allows only one insertion or deletion. The proposed extension algorithm then runs in time linear in the length of the read.

Generally, if the approximate matching criteria is based on a similarity score, the Smith-Waterman algorithm is the method of choice to compute a (gapped) local alignment (see e.g. BFAST). The gaps between mapped seeds in the seed-and-vote scheme [70] (Section II-A1) are filled by a Smith-Waterman dynamic program that takes into account the indel length that is inferred from distance discrepencies of seeds between read and genome.

Several schemes have been proposed that build on Single-Instruction Multiple-Date (SIMD) instructions available on modern CPUs to compute several cells of the dynamic programming matrix in parallel [30]. Stampy, Novoalign, and SHRiMP are example methods that employ SIMD-vectorized Smith-Waterman implementations. SHRiMP devises a new scheme that increases the degree of parallelism at the cost of supporting only uniform match and mismatch scores.

Even faster bit-parallel algorithms exist if quality scores are neglected and the mere number of edit operations (*k*-errors problem) is used as a measure of relatedness. RazerS, RazerS 3, Masai [106], Hobbes [3], and GEM [85] implement (a banded variant [44] of) Myers bit-vector algorithm [89].

III. BWT-Based Methods

Among the most widely used software tools used for mapping short DNA reads to a reference genome are Bowtie [60] and its successor Bowtie 2 [58], as well as BWA [63] and BWA-SW [64]. At the core of all of these programs lies the FM-index [34], which allows them to use the *Burrows-Wheeler transform* (BWT) [17] as a suffix array-like index structure. Simply speaking, the FM index augments the space-efficient BWT with additional data that permits very fast exact string matching. In Sections III-A and III-B, we briefly summarize key properties of the BWT and the main algorithmic ingredients that make the BWT searchable at the cost of a small amount of additional memory.

Prior to their use in NGS alignment, BWT-based indexes such as the FM index and the *compressed suffix array* [39] had been used for several other problems in computational biology, including repeat finding [42], whole-genome comparison [72], the design of whole-genome tiling arrays [38], [95], the local alignment of a long pattern (e.g., a gene) to a reference genome [57], and more recently, it has also been employed in genome assembly [105]. Although the problem studied in [57] is mathematically very similar to the short read alignment problem, the exact method the authors proposed does not provide a practical solution to the alignment of hundreds of millions of short reads. In Sections III-C and III-D we show how state-of-the-art mapping tools modify and extend the BWT and the FM index so they can handle inexact matching as well as many specific characteristics of next-generation sequencing reads.

A. Burrows-Wheeler Transform

The BWT T^{bwt} is a reversible permutation of a string T that facilitates compression through its repetitive structure, and that has proved useful in lossless data compression techniques such as bzip2. It is constructed by (i) appending character $\$ \notin \Sigma$ to T, (ii) creating a conceptual matrix $\mathscr{M}_T$ (the Burrows-Wheeler matrix) whose rows are the cyclic rotations of $T\$$ in lexicographical order, and finally (iii) extracting the last column of $\mathscr{M}_T$. Figure 7 illustrates steps (i)–(iii) for an example DNA sequence. Sorting the rows of $\mathscr{M}_T$ is equivalent to sorting the suffixes of T, which shows the close relationship between matrix $\mathscr{M}_T$ and the suffix array built on T.

A key feature of the Burrows-Wheeler Matrix $\mathscr{M}_T$ is the *Last-to-First column mapping LF* [17]: the ith occurrence of a character c in the last column corresponds to the ith occurrence of c in the first column (Figure 7). More formally, let $C(c)$ be the total number of characters in T that are alphabetically smaller than c, and $Occ(c, i)$ the number of occurrences of character c in prefix $T^{bwt}[1,i]$. Then character $T^{bwt}[i]$ is located in the first column at position

$$LF(i)=C(T^{bwt}[i])+Occ(T^{bwt}[i],i) \quad (1)$$

B. Backward Search

The FM-index allows one to determine the range of lexicographically ordered suffixes of T that have pattern P as prefix without storing the corresponding suffix array. Algorithm BACKWARDSEARCH shown below computes in iteration i the maximal range of rows in $\mathscr{M}_T$ prefixed by $P[i,m]$, or equivalently the maximal range of entries in the corresponding suffix array pointing to suffixes prefixed by $P[i, \mathrm{m}]$.

Algorithm 1

BACKWARDSEARCH(P, m, n, C, Occ)

```
1 sp ← 1, ep ← n
2 for i ← m to 1 do
3 |   sp ← C(P[i]) + Occ(P[i], sp − 1) + 1
4 |   ep ← C(P[i]) + Occ(P[i], ep)
5 |   if sp > ep then return ∅
6 end
7 return [sp, ep]
```

The maximal range of rows in $\mathscr{M}_T$ prefixed by $P[i-1,m]$, possibly empty (line 5), can be obtained from the maximal range of rows in $\mathscr{M}_T$ prefixed by $P[i, m]$, by determining the immediate predecessors of the first characters in rows sp and ep in the original text T, which are the given by $T^{bwt}[sp]$ and $T^{bwt}[ep]$, respectively, and by computing their corresponding occurrences in the first column through the LF mapping, see lines 3 and 4. The correctness of this relation was demonstrated by Ferragina and Manzini [34]. Figure 8 illustrates the steps of Algorithm BACKWARDSEARCH for an example pattern.

Implementing function C as an array requires just $\sigma \log n$ bits and allows retrieval of any entry in lines 3 and 4 in constant time. Using auxiliary data structures and the *four Russians trick* [14], Ferragina and Manzini [34] showed how to find $Occ(c, i)$ in constant time too, using $\mathcal{O}(nH_k)$ bits of space. H_k denotes the kth order entropy, a lower bound on the number of bits any kth order compressor requires to encode a symbol. Algorithm BACKWARDSEARCH thus runs in optimal $\mathcal{O}(m)$ time.

After determining the relevant range of rows in $\mathscr{M}_T$ in $\mathcal{O}(m)$ time, the task remains to locate the corresponding starting positions in the reference genome, which would be directly given by the entries of a suffix array. Ferragina and Manzini [34] store for a suitable subset of (marked) rows of M_T their text offset. They implicitly move backwards in T by jumping between the corresponding rows in M_T using the LF mapping until a marked row is encountered. The text position returned is then equal to the text offset assigned to the final row plus the number of "jumps". A jump can be performed in constant time, and the data structure containing the text offsets of marked rows supports queries in constant time too. If every ηth row is marked, $\eta = \lceil \log^{1+\varepsilon} n \rceil$, occ occurrences of P in T can be located in $\mathcal{O}(occ \log^{1+\varepsilon} n)$ time. Bowtie [60] tries to balance the space-time tradeoff by setting $\eta = 32$,

which corresponds to $\varepsilon < 0.01$. Similarly, BWA[63] samples the suffix array at intervals of size 32 and traverses the text in forward direction using the inverse function of the LF mapping [39].

Subsequent work on the FM index aimed at reducing its dependence on the alphabet size σ [35], [91]. The results in [34] assume that the size of the alphabet is constant, and indeed the space requirement depends exponentially on σ. Because DNA only uses four nucleotides ($\sigma = 4$), this dependence is not an issue. In the original Bowtie implementation, the BWT-based FM index for the human genome requires only about 1.3 gigabytes (GB) of memory [60], which allows it to fit easily in the main memory (RAM) of a standard laptop or desktop computer. This also facilitates distribution of pre-computed indexes for most major model organisms; for example, the Bowtie site[5] provides pre-built indexes for the human, mouse, dog, rat, fruit fly, and other genomes.

C. FM-Index-Based k-Mismatch Search

The indexes discussed above are designed for the exact string matching problem. Exact string matching is an idealized version of the real read mapping problem. Sequencing errors and differences between the reference genome and the sequenced organism can cause the read to deviate substantially from the the reference genome (see Section I-B2).

In the simplest case, a BWT index replaces the hash index in a scheme as discussed in Section II-A2a. When two mismatches are allowed, SOAP2 [68], for instance, splits the read into three parts such that one is guaranteed to match perfectly, and the exact match is located through the BWT of the reference genome.

Below we describe how the high-throughput read mappers Bowtie [60] and BWA [63] implement their own backtracking methods that allow for (a small number of) mismatches and that are engineered for the efficient processing of millions of short sequence reads.

1) Backtracking—The general principle of backtracking during backward search can be illustrated by considering its application to the k-mismatch problem. The pseudocode of a backtracking aware backward search that finds all k-mismatch occurrences of P in T is given in Algorithm kMISMATCHSEARCH [83].

[5] http://bowtie-bio.sourceforge.net/index.shtml

Algorithm 2

*k*MISMATCHSEARCH(*P, k, j, sp, ep*)

```
1  if sp > ep then return ∅
2  if j = 0 then return [sp, ep]
3  foreach c ∈ {A, C, G, T} do
4  |  sp' ← C(c) + Occ(c, sp − 1) + 1
5  |  ep' ← C(c) + Occ(c, ep)
6  |  if P[j] ≠ c then k' ← k − 1 else k' ← k
7  |  if k' ≥ 0 then
8  |  |  kMismatchSearch(P, k', j − 1, sp', ep')
9  |  end
10 end
```

The main difference to algorithm BACKWARDSEARCH (without backtracking) lies in line 3, where in addition to current nucleotide $P[j]$ (**else** block in line 6) all possible substitutions of $P[j]$ are considered (**then** block in line 6). The total number of substitutions is not allowed to exceed k (**if** condition in line 7). Each node of the recursion tree maintains the range of suffixes [*sp, ep*] in the suffix array that are prefixed by the current modification of $P[j,m]$. Figure 9 shows the recursion tree for an example search with one mismatch, i.e. $k = 1$.

BWA [63] generalizes algorithm *k*MISMATCHSEARCH in a straightforward way to find *k-errors* occurrences, allowing for both mismatches and insertions and deletions. Note that the practicability of Algorithm III-C1 strongly depends on the alphabet size, since all possible substitutions are considered in line 3. In other words, current read mapping methods that apply the FM index through a backtracking procedure to the inexact matching case exploit a small alphabet size of $\sigma = 4$.

2) Engineering Backtracking—The worst case running time of *k*MISMATCHSEARCH is $\mathcal{O}(|\sum|^k m^{k+1})$ [83]. To further improve their practical performance, Bowtie [60] and BWA [63] invest some additional memory to not only build the FM index for the genome T, called the *forward* index, but also for the reverse (not complemented) genomic sequence $T^r = t_n t_{n-1} \cdots t_1$, called the *reverse* or *mirror* index. However, the two methods exploit the reverse FM index in different ways, as we will illustrate in the following sections. In Section III-C2d we briefly summarize algorithm design decisions that are influenced by specific properties of the data such as base quality values and systematic errors.

a) Pruning by Bounding: BWA uses the reverse index to precompute for each prefix $P[1,i]$ of a given read P the minimum number of parts $D[i]$ into which $P[1,i]$ has to be broken into such that each part occurs at least once in genome T (exactly) [83]. Array D can be computed in time linear in the read length through a backward search of the reverse read in the reverse FM index, increasing the current entry of D by one every time range $[sp,ep]$ becomes empty. $D[i]$ provides a (not necessarily tight) lower bound on the number of errors

an approximate occurrence of $P[1,i]$ in T exhibits and can thus be used to prune the space of potential matches explored by backward backtracking (see Section III-C1) as follows. Let e_v be the total number of errors introduced in line 3 of Algorithm III-C1 up to current node v of the recursion tree reached after backward searching suffix $P[j, m]$. Then, if $e_v + D[j-1] > k$ proceeding the search from node v cannot yield a match with at most k errors and therefore the current branch can be pruned.

b) Pruning by Case Sensitivity: The pruning strategy of Bowtie can be seen as an extension of the pigeonhole principle underlying the seed-and-extend paradigm described in Section II. The *case sensitive backtracking* scheme considers all possibilities to distribute k mismatches to 2 segments of the read and disallows backtracking in a segment that is assumed to be error-free in the current mismatch scenario. To prune the search space, backtracking on higher levels (i.e. lower depth) of the search tree is avoided by searching read P from right or left using forward or mirror index, depending on where the error-free segment of the read is assumed to lie. This idea can be illustrated for the case where only one mismatch is allowed in the alignment ($k = 1$) and the read is split into a left and a right half (Figure 10). The mismatch can lie either in the left half or in the right half of the read. In the first case (Figure 10a) the right half must be error-free and therefore in line 3 of Algorithm III-C1 no substitution is necessary, that is, c must be equal to $P[j]$ if j lies in the right half of the read. Searching P from right to left using the forward index restricts branching to lower levels (i.e. higher depth) of the search tree. The second case (Figure 10b) is a mirror of the first case and thus P is searched from left to right using the mirror index, disallowing substitutions in the left half of the read and thus again avoiding branching on higher levels of the search tree.

For two or more mismatches ($k \geq 2$), not every assignment of mismatches to read segments implies an error-free segment. In fact, for every $k \geq 2$ there is a worst case scenario where both read segments have at least $\left\lfloor \frac{k}{2} \right\rfloor$ erros. Bowtie switches in different *phases* between forward and mirror indexes, depending on where the segment assigned a smaller number of errors by the current scenario lies. Again, such a segment is preferably treated on higher levels of the search tree.

c) Pruning versus Filtering: In [83] the pruning heuristics introduced in Sections III-C2a and III-C2b were experimentally evaluated on ChIP-seq data taken from [111] comprising 10,000 36 bp long reads. For a varying number of mismatches k these heuristics were compared against the *suffix filter* [48], a state of the art *filtering* algorithm (see Section II) developed for the approximate string matching problem. The case sensitive backtracking strategy implemented by Bowtie slightly outperformed the suffix filter in terms of search time per read, when either the best, or all occurrences were located. A novel pruning heuristic proposed in [83] that combines suffix filter and case sensitive backtracking was superior in speed, at the cost of a significantly higher space consumption.

d) Application Engineering: Compared to the straightforward recursive scheme applied by Algorithm III-C1, Bowtie and BWA exploit the quality of a partial alignment, i.e. its score, to guide the (implicit) traversal of the suffix tree. BWA uses a heap structure to implement a

best-first strategy, Bowtie traverses the tree in a slightly modified *depth-first* manner, selecting greedily substitution positions with minimal quality value.

Both methods employ a *seeding* strategy (see Section II) that assumes a small number of sequencing errors in a high-quality part of the read, which usually comprises a few tens of bases at the $5'$ end of the read. A smaller number of errors reduces the need to backtrack while backward searching the seed and allows to find promising candidate regions on the genome efficiently. For example, the effectiveness of Bowtie's pruning strategy (see III-C2b) depends on the existence of a read segment with a small number of errors (ideally error-free). Therefore, the division into halves is with respect to the seed instead of the full read.

Several restrictions are imposed to further balance the efficiency-sensitivity trade-off. For example, the total number of errors and backtracks allowed is limited by BWA and Bowtie, respectively. A relatively low error rate exhibited by the read sequences and a typically low divergence between the sequenced individual (from which reads P are derived) and the reference genome T render this trade-off suitable for most applications.

D. FM-Index-Based Similarity Search

Lam *et al.* [57] describe an extension of BACKWARD-SEARCH to the search for most similar occurrences of P in T that may exhibit mismatches, insertions and deletions. Gaps are penalized by an affine function in their length. Insertions and deletions occur more often in longer reads produced by current sequencing technology. In particular, indels are the predominant type of sequencing error in data generated by single-molecule sequencing technologies [54].

Their tool BWT-SW, which is designed for the alignment of relatively long queries (e.g., 3000 bp) against the human genome, resorts to local alignments computed by dynamic programming. The main difference to the Smith-Waterman algorithm lies in the (implicit) suffix trie representation of T against which pattern P is aligned to. Roughly speaking, when the alignment of P and a path in the suffix tree reaches a node v, P has implicitly been aligned to all occurrences of the string represented by v in T, due to the common prefix structure of the trie. Representing T by a BWT, BWT-SW generates all words in T by traversing the suffix trie in a pre-order fashion (backtracking), determining edges on the fly using backward search.

The search can be restricted to *meaningful* alignments [57], which allows to prune the search space without sacrificing optimality. Overall, in their experiments Lam *et al.* observed a 1000 fold speed-up compared to the Smith-Waterman algorithm. For longer patterns however, BLAST ran several times faster than BWT-SW and missed only few significant alignments. In general, the time required to compute an alignment increases to $\mathcal{O}(m^2)$, compared to $\mathcal{O}(m)$ if only mismatches are considered, and renders an exhaustive backtracking-based approach infeasible for the read mapping problem.

To find approximate occurrences of a read in a more general error model including (affine-weight) gaps efficiently, Bowtie 2 [58] and BWA-SW [64] resort to the seed and extend paradigm that was already applied successfully by earlier methods discussed in Section II.

The crucial difference lies in the seeding step, where candidate mapping locations on the genome are determined efficiently using the FM index. While BWA-SW permits gaps in the seed alignments, Bowtie 2 tries to “guess” a short substring of the read that aligns with few mismatches only. This anchoring idea is taken one step further by Chaisson and Tesler [19] who developed BLASR to address the computational challenge posed by aligning multi-kilobase long reads generated by single-molecule sequencing technologies with a high error rate. We summarize the main algorithmic concepts of BLASR and BWA-MEM, which complements the BWA software package [63], [64] for the scenario of long-read alignments, in Section III-F.

1) Gapped Seeds—Li and Durbin [64] combine and extend the dynamic programming scheme on suffix tries from [57] and the heuristic seed and extend paradigm of BLAST [7]. BWA-SW determines seeds by a dynamic program between a prefix trie and a prefix directed acyclic word graph (DAWG) [12], allowing mismatches and indels. A prefix DAWG is obtained from a prefix trie by collapsing nodes that correspond to the same suffix array interval and is implemented, as the prefix trie, by an FM index. In that way, the evaluation of a single pair of nodes essentially compares a set of substrings in T with a set of substrings in P, exploiting repetitiveness both in T and in P. See Figure 11 for an illustration. As soon as these sets become too small, the overhead involved in traversing prefix trie and prefix DAWG exceeds the benefits and these seeds are extended by the Smith-Waterman algorithm. Two heuristics are applied to accelerate the computation. The DP is pruned at low scoring nodes, and only seeds that are likely to lead to distinct alignments are retained.

2) Ungapped Seeds—Bowtie 2 [58] relies on seed strings, i.e. possibly overlapping (consecutive) substrings of P, obtained from evenly spaced intervals along the read. See Figure 12 for an example. The seed strings are chosen to be short enough (e.g. 20 nt) that their alignment can ignore indels and has to account only for a small number of mismatches (≤ 1). Such an alignment can be computed efficiently by an FM index supported backward search with very limited backtracking (see Algorithm III-C1). In this step, Langmead and Salzberg apply a scheme similar to the one developed in Bowtie [60], with a bi-directional BWT [56] supporting an efficient switch between alignment directions. Seed alignments are selected in an order that favors those with a smaller number of alternative occurrences in T and are subsequently extended by a hardware-accelerated dynamic program. Single-Instruction Multiple-Date (SIMD) instructions, supported by most of the modern CPUs, are used to parallelize the Smith-Waterman algorithm. It organizes the data for the vector instructions similarly to [31], with slight adaptions to the specifics of the read alignment problem.

E. FM-Index-Based k-Error Search

As mentioned in Section II-A2a, the seeding strategies of Masai [106] and the GEM mapper [85] aim at full sensitivity by employing the pigeonhole principle, according to which there always exists a seed with at most $\left\lfloor \frac{k}{s} \right\rfloor$ errors in a partitioning of a k-error occurrence into s seeds.

The GEM mapper determines the partitioning adaptively based on the number of matches of each seed in the genome. During backward search in the FM-index a new seed is started as soon as the number of matches ($ep - sp + 1$, see Section III-B) of the current seed falls below a certain threshold, keeping the number of candidate regions to be verified small. If the resulting number of seeds is large enough ($s \geq k + 1$), full sensitivity can be guaranteed by searching for exact seed matches, otherwise approximate matches of seeds have to be taken into account.

Masai searches for a seed with at most $\left\lfloor \frac{k}{s} \right\rfloor$ errors through backtracking [112] in a (conceptual) suffix tree of the reference genome, implemented as a suffix array or FM-index. A second index organizes the non-overlapping seeds of the reads in a radix tree and allows to search all seeds in the suffix tree simultaneously. Alternatively, Masai permits to partition the read into $s = k + 1$ seeds per read and searches for exact matches using a simpler and more efficient algorithm.

Notice that both the filtration and in particular the extension based on Myers bit-vector algorithm [89] (see Section II-C2) as employed by Masai and GEM rely on the edit distance measure and do not take into account quality scores.

F FM-Index-Based Long-Read Alignment

Chaisson and Tesler [19] tailor the idea of anchoring the alignment at multiple short (exact) seed matches to the problem of mapping longer and error-prone reads produced by single-molecule sequencing methods. Aligning reads several kilobases long containing a higher number of errors, most of which are insertions and deletions, poses a computational challenge that shifts the NGS alignment problem slightly towards the problem of aligning whole genomes (WGA) [103], [24]. Their tool BLASR therefore combines a successive refinement strategy typically employed in WGA with space- and time-efficient data structures that can yield a boost in alignment speed. It applies a distance measure that does not simply count the number of edit operations, but assigns alphabet-dependent weights.

BLASR first finds candidate intervals on the reference genome that are subsequently evaluated by a pairwise alignment routine. Candidate intervals resemble dense clusters of short exact matches between read and genome that are consistent in terms of coordinate range, order, and orientation. Exact matches are obtained by scanning the read from left to right and searching the longest common prefix of the current read suffix and the genome using either a suffix array or BWT-FM index. The resulting anchors are clustered by chaining [2] a maximal set of non-overlapping anchors in increasing order (by coordinate in genome and read) without exceeding the read length.

The anchor chains define the intervals to which the read is aligned, extending the boundaries slightly to take into account insertion and deletion errors that blur the start or end position of the read (Figure 13A). The pairwise alignment of the read and the intervals is further subdivided into a more coarse alignment using sparse dynamic programming (SDP) [28] (Figure 13B) and a more comprehensive banded alignment guided by the coarse alignment (Figure 13C). SDP again determines a consistent set of small exact matches (anchors)

between the read and the genomic interval. The final dynamic programming considers only entries in the DP matrix that lie in a band around the anchors returned by the previous SDP step. The scoring of the alignment is based on the quality values and the alternative base calls provided by the PacBioRS platform.

The feasibility of this approach clearly depends on (i) a sufficiently high number of anchors marking the true mapping location, and (ii) an overall low number of anchors lying outside this true genomic interval. Chaisson and Tesler propose a method to computed the probability of event (i) for a given read length and sequencing accuracy on the one hand, and study the repetitive structure of the human genome to assess (ii) on the other. They conclude that for the human genome, both conditions are satisfied, rendering their successive refinement strategy feasible in this context.

BWA-MEM, available as part of the BWA software package [63], [64], also relies on maximal exact matches to seed alignments of longer and error-prone reads generated by, e.g., the PacBioRS platform. Compared to the longest common prefix that BLASR computes for every read suffix, BWA-MEM restricts the search space to supermaximal exact matches (SMEMs) between reference and read, i.e., exact matches that cannot be extended at either end and that are not contained in any other exact match on the read. The unique variants of SMEMs were previously used by Delcher *et al.* [24] as maximal unique matches (MUMs) to anchor the alignment of two whole genome sequences.

BWA-MEM's seed extension uses a banded Smith-Waterman algorithm with heuristic modifications that increase its robustness to sequencing errors: A suboptimal end-to-end alignment can be chosen heuristically over a local alignment that achievs a higher score, and extensions through low-scoring regions are avoided without penalizing long gaps in either of the sequences.

IV. The Agony of Choice

Despite several algorithmic advances (e.g. spaced seeds) that benefitted hashing-based methods, this class of methods were largely replaced in practice by tools relying on the FM index, after Bowtie and BWA first introduced this space and time efficient data structure in the context of short read mapping. In the experiments performed in Langmead *et al.* [60], Bowtie mapped 35-bp reads 38–350 times faster than state-of-the-art hashing-based methods, with almost no loss in sensitivity, making such methods the choice of short read aligner for large-scale mammalian re-sequencing studies.

The low error rate of high-throughput sequencers together with a typically low divergence between reference and sequenced organism allowed aligners like Bowtie and BWA to traverse the space of mutations for very short reads efficiently. With read lengths increasing to 100–150 bp, however, the alignment throughput of purely index-based methods decreases. Bowtie 2 and BWA-SW therefore explore the larger search space of longer, gapped alignments by resorting to the seed-and-extend paradigm.

Much longer and more error-prone reads generated by recent single-molecule sequencing technologies, with lengths exceeding 10,000 bp and error rates of 15% or higher, pose new

challenges to read mapping tools. With insertions being the predominant type of sequencing error of this technology [54], error models that neglect or restrict gaps are inadequate. Algorithmic recipes successful in this context combine and generalize concepts developed in the short read mapping and the whole-genome comparison literature. To avoid incurring quadratic cost for aligning a long read, algorithms can anchor the alignment at multiple seeds which ideally correspond to maximal stretches (e.g. longest common prefix [19], MUMs [24], or SMEMs as used in BWA-MEM) of error-free nucleotides. These seeds can be matched efficiently using a structure like the FM-index and guide a coarse alignment method [19] developed for the alignment of whole genomes.

Although these observations suggest a certain class of alignment method to be applied in a particular scenario, they typically leave the user with a large number of mapping tools to choose from. As previous benchmarks showed (see [41] and references therein), alignment sensitivity and throughput also varies with quantitative data parameters such as read length, genome size, and genome repetitiveness, and there is no single best aligner. Moreover, the choice is not only among available tools but also among the many parameters that can be provided to adjust an algorithm to the characteristics of the data, available hardware, and the requirements imposed by the downstream analysis. For example, BWA-MEM offers options to adjust the scoring scheme, seed length and filtering strategy when run on reads generated by particular sequencing technology. The aim of this review is not to suggest the use of specific tools, but rather to enable and guide the reader to select or design mapping software and parameters that best suit a particular need.

V. Parallelization

The enormous data sets generated by modern sequencing machines require greater speed than any of the methods described above can deliver on a single processor. To address this challenge, many methods implement at least one form of parallel computing. The widely-used aligners Bowtie1/2 [58] and BWA [63] include built-in methods to distributing work (e.g. reads) across multiple threads to make use of multicore CPUs or multi-CPU nodes. Taking parallelization one step further, CloudBurst [101] and Crossbow [59] are two prominent examples of systems that utilize *cloud computing* for parallel read alignment implementations. Using Hadoop[6], an open source implementation of Google's Map-Reduce programming model [23], these two systems parallelize the execution of RMAP and Bowtie, respectively, using large clusters of commodity hardware.

A cost-effective alternative to cope with the plateau in CPU clock speeds is to leverage the massive parallelism provided by a graphics processing unit (GPU), whose high-level architecture is shown in Figure 14. Designed primarily to perform computer graphics calculations, GPUs are optmized for throughput rather than latency. on a GPU, hundreds of high performance streaming processors execute small tasks in parallel. This processing power can also be harnessed by data-parallel algorithms developed for diverse general-purpose applications (known as General Purpose computing on GPUs, or GPGPU), including the alignment of short NGS reads to a reference genome.

[6] http://hadoop.apache.org

GPU-based aligners rely on the same general principles discussed in Sections II and III, such as the seed-and-extend paradigm, searching an FM-index for approximate occurrences through backtracking, or the verification of candidate genomic regions through a Smith-Waterman alignment. However, exploiting the computing power of GPUs in the highly parallel alignment of short reads involves more than simply assigning to each GPU thread a single read and launching thousands of threads. The massive parallelism of GPUs comes at the cost of a more restrictive programming model compared to CPUs. From a software development point of view, the following properties of the GPU architecture determine the design of data structures and algorithms.

1. Similar to classical SIMD pocessors, GPU cores execute sequential threads on different data in a *single instruction, multiple thread (SIMT)* fashion. The hardware groups several threads together that execute the same instruction at the same time, on different operands.
2. Repeated accesses to the global memory and spacehungry data structures can cause memory contention in the highly parallel environment of a GPU.
3. Global memory access is considerably slower than performing arithmetic operations.
4. Many consecutive locations in global memory can be accessed in parallel.

Furthermore, the engineering of GPU-based software requires one to take into account features of the underlying hardware such as shared memory and cache (hierarchy) configurations. Although these aspects can have a large impact on the overall performance of the software, we focus on the algorithm design decisions made by current GPU aligners to adhere to the data-parallelism properties listed in 1–4 above. The hashing-based methods in the following section divide the work between CPU and GPU in a straightforward way, while BWT-based methods in Section V-B make more explicit use of the GPU properties 1–4.

A. Hashing-Based GPU Aligners

Blom *et al.* [11] developed SARUMAN, the first short read aligner that makes use of GPU computing power. In their algorithm, the host CPU pre-computes a hash index from the reference genome and performs filtration based on the pigeonhole principle, see Section II-A2a. The resulting candidate mapping locations are verified in parallel through GPU threads that compute a global alignment by a modified Needleman-Wunsch algorithm.

Similarly, GPU-RMAP [5] builds on RMAP (see Section II-A2a) and indexes the reads on the host. In contrast to SARUMAN, however, the genome is divided into independent segments and GPU threads scan these segments in parallel to find and score (number of mismatches) mapping locations of all reads. In a second stage, reads are distributed to GPU threads that select the best scoring location. Compared to the sequential implementation of RMAP [108], GPU-RMAP replaces the hash table containing the reads by a binary search tree, placing frequently accessed levels of the tree into faster cache memory.

B. BWT-Based GPU Aligners

BarraCUDA [52], SOAP3 [74], SOAP3-dp [81], CUSHAW [78], and CUSHAW2-GPU [76] are all based on the BWT and variants of the FM-index (see Section III), and each employs different strategies to address GPU properties 1–4:

1) Avoiding Divergence of Execution—A consequence of property 1 is that divergent execution paths (branching) of threads in the same group drastically harms performance. The exploration of potential substitutions (see backtracking in Section III-C1) will lead to many branches for some reads that will cause other cores with few branches to be idle and wait for the other threads to reach the same point of execution. To avoid this loss in parallel efficiency, SOAP3 and SOAP3-dp align reads in groups of similar branching complexity, which is estimated at runtime from the number of suffix array ranges. The most complicated reads are handled by the CPU. For the same reason, BarraCUDA divides reads into segments and explores potential (inexact) matchings by consecutive depth-first searches, each restricted to a segment.

2) Reducing Memory Consumption—One obvious way to cope with the limited device memory is to organize reads or genomic candidate regions in batches that are scheduled one-by-one (CUSHAW, CUSHAW2-GPU).

To reduce the memory consumption per thread (see property 2), BarraCUDA and CUSHAW traverse the space of possible base substitutions and indels (BarraCUDA only) in a depth-first search manner. During DFS, BarraCUDA trades efficiency for reduced memory consumption by storing only the currently best branch, evaluating certain nodes multiple times. CUSHAW further prunes the search space by imposing additional constraints on the overall quality score.

Compared to the jumping strategy proposed by Ferragina and Manzini [34] to locate a read occurrence on the genome, CUSHAW and CUSHAW2-GPU avoid auxiliary data structures by resorting to a reduced suffix array that stores only every ηth entry. In contrast to [34] however, a search cannot be guaranteed to be successful after η jumps. Similarly, CUSHAW stores Occ (see Section III-A) only at every 128th position and interpolates the remaining entries through T^{bwt} as needed.

3) Reducing and Coalescing Memory Access—SOAP3 and SOAP3-dp reduce random memory access (properties 2 and 3) required by the 2way-BWT index [56] that facilitates the switch in search direction during the approximate matching of a read. In particular, an auxiliary array supporting $Occ(c, i)$ queries (see Section III-B) is built following a simple one-level sampling strategy. The sampling ratio is chosen such that (i) a single access to T^{bwt} exploits the full memory bus width, and (ii) the organization of the auxiliary array into groups of Occ values for the same position allows to determine $Occ(c, i)$ for all four nucleotides c through a single memory access (see property 4). The latter operation is of key importance for the efficient search in the 2way-BWT index.

To accelerate memory access (property 3 and 4), SOAP3 and SOAP3-dp *coalesce* simultaneous access of different threads from the same SIMT group to global memory. The

idea is to arrange reads in the memory in a way such that memory requests of threads in the same SIMT group can be satisfied in a single memory transaction. For that, reads are partitioned into groups of the same size as there are threads in a SIMT group and the jth words of the reads in the same group occupy a consecutive region in memory (see Figure 15). The access to the BWT index in global memory is more random and thus can be coalesced only to a small degree.

The strategies discussed so far mostly involved adapted data structures and memory orgranization. From an algorithmic point of view, few memory accesses are desired (properties 2 and 3) when verifying a candidate region using a Smith-Waterman algorithm. SOAP3-dp applies a GPU-tailored variant of the Smith-Waterman algorithm [75] that reduces the number of memory operations per iteration. Similarly, CUSHAW2-GPU computes alignment scores using a slightly modified variant of a GPU-accelerated SW algorithm [77] and reconstructs the best scoring alignments by backtracking in 4×4 tiles. Since 2 bits suffice to encode the three possible edit operations in each cell, an alignment restricted to a tile can be represented by a single 32-bit integer and thus proceeding in units of a tile is supported by single read/write operations to the matrix storing the corresponding edit operations.

C. GPU versus CPU aligners

A general statement about the relative performance and practicability of GPU-based and CPU-based aligners is difficult to make, since it would be based on a comparison of algorithms that run on hardware implying a different cost and availability. Nevertheless, the gain in alignment throughput of GPU-based tools over methods that run on multiple cores on the CPU has been rather modest. While BarraCUDA and CUSHAW achieved a throughput that was comparable to that of BWA run on 6 cores [52] and Bowtie using 1–4 threads [78], respectively, they were outperformed [81] by CPU-based aligners like Bowtie 2 [58] and GEM [85]. The successor CUSHAW2-GPU achieved a 1.5-fold speedup over Bowtie 2 when employing one GPU in addition to the 12 CPU threads Bowtie 2 used. SOAP3-dp seems to utilize the computing power of GPUs most efficiently. In their experiemts [81], SOAP3-dp ran 3.5 times faster than GEM and 7 times faster than Bowtie 2 when using one GPU device in addition to 4 CPU threads.

Wilton *et al.* [118] identify two main obstacles that hinder GPU-based alignment software to profit more directly from the massive parallelism provided by a GPU. First, the Smith-Waterman dynamic program formulates dependencies that favor the independent computation of entire alignments in parallel compared to a data-parallel implementation of the algorithm suitable for GPUs. Second, few seeds extracted from reads match a very large number of subsequences in the reference, implying an enormous computational burden for the alignment software. In [118] the authors propose “GPU-friendly” heuristics based on list-manipulation operations for pruning this search space. By assigning those components of the sequence alignment pipeline to the GPU that are amenable to a data-parallel implementation, Arioc [118] achieves a 10-fold speedup over state-of-the-art CPU-based aligners Bowtie 2 and BWA-MEM.

We believe that GPU-based aligners will play an important role in the analysis of datasets generated by largescale sequencing studies involving thousands of individuals. GPU-based aligners scale well with an increasing number of GPU devices used [118], [52] and thus provide a means to support GPU-tailored algorithmic advances with additional hardware [118].

VI. Making an Educated Guess

The distance or similarity threshold k is generally chosen such that the true origin of a read is likely to be among the approximate occurrences implied by k. Read mappers like RazerS (3), Zoom, Hobbes, GEM, Masai, and BWT-SW simply enumerate and report *all* occurrences, leaving it to the downstream analysis to pick the right location. They sometimes allow for an adjustable sensitivity or limit the number of reported occurrences of highly repetitive sequences.

A. Picking the Best Mappings

Alternatively, several approaches have been proposed that make an attempt to further narrow down the set of potential mapping locations, or even guess the true read origin as the best alignment. In the simplest case, the alignments are ranked by the distance or similarity score. For instance, SOAP (2) by default returns the alignment imposing the minimal number of mismatches or the smallest gap. Similarly, BFAST, BWA and Bowtie 2 (default mode) prioritize alignments achieving the highest score when guessing the true read origin, while Shrimp outputs the best n scoring alignments. BLASR only considers up to a certain number of the least repetitive candidate genomic regions.

Methods that are based on a Hamming distance or edit distance rather than a more general scoring scheme often take into account base quality scores to guide the subsequent prediction of the read origin. MAQ, for example, picks the location where the sum of the qualities at mismatched bases is minimized. Accordingly, Bowtie greedily seeks alignments with low qualities at mismatched positions. BWA-SW, on the other hand, aims at selecting distinct ones among the highest scoring alignments.

B. Estimating the Qualtiy of a Mapping

Some real aligners provide, with each alignment, an estimate of the probability that the alignment is the true genomic origin of the read. This is not a measurement of the alignment quality itself: it does not depend on the number of mismatches, consistently mapped read pairs, or the quality of the sequenced DNA. Rather, it captures the uncertainty that arises when a read's true source is a repetitive region of the genome. If a read can be mapped equally well to two different genomic locations, then in statistical terms we can only have a 50% probability of assigning it to the correct location. Downstream analysis protocols, especially variant calling in human DNA sequencing, typically rely on these mapping qualities to decide which variants can be called with confidence. Li *et al.* [65] introduced the mapping quality score Q (or MAPQ) and scaled it [29] as

$$Q = -10 \log Pr[\text{mapping is incorrect}];$$

i.e., a quality value of 20 corresponds to a probability of 0.01 that the mapping is incorrect. Higher quality values indicate higher confidence in the mapping.

Assuming that genomic positions are chosen by the sequencer with equal probability (uniform prior distribution), the posterior probability of a read sequence R originating from a position m in the reference sequence G is

$$Pr\left[m|R,G\right] = \frac{Pr\left[R|G,m\right]}{\sum_{i=1}^{|G|} Pr\left[R|G,i\right]} \quad (2)$$

The numerator gives the probability of obtaining a particular sequence R when sequencing begins at position m in the reference. It depends on the divergence between reference and sequenced individual and the error rate of the sequencing process. Typically the former is ignored and reads are treated as being obtained from the reference. The latter is ideally captured by the scoring scheme that the alignment algorithm optimizes for. While the likelihood of a mismatch to be caused by a sequencing error can be estimated from the base quality scores (see Section I-A1), the scoring of indels usually is more empirical. MAQ [65], for instance, minimizes the sum of quality scores of mismatched bases. Bowtie 2 [58] additionally applies an affine gap penalty. The banded dynamic program employed by BLASR [19] explicitly takes into account quality values provided by PacBioRS for substition, insertion, and deletion events. SHRiMP [98], on the other hand, estimates the rate of mutations (substitutions, indels), and sequencing errors via bootrapping. Note, however, that these estimates play a role only in the assessement of mapping quality; the alignment algorithm optimizes a different scoring scheme.

Algorithms computing one (arbitrary) best alignment in terms of alignment score only capture the similarity in the sense of the numerator in equation (2). To provide a mapping quality for an alignment, its uniqueness is measured by the denominator of (2). It captures the probability that read R was obtained from anywhere in the genome. Since summing Pr[R|G, i] over all genomic positions i is computationally infeasible, existing methods only consider the sums from genomic regions exhibiting a high similarity to R. For instance, MAQ estimates the uniqueness character of an alignment from the best alignment and all second best alignments. Bowtie 2, BLASR, BWA [63], and SOAP3-dp [81] apply a similar uniqueness criteria in assigning mapping qualities as MAQ. In contrast, SHRiMP counts the number of substrings of a random genome of the same length to which a read can align with the same number of mismatches to guess the uniqueness of its alignment on the genome.

Note that computing a mapping quality imposes a significant cost on alignment. For example, if a user has specified that the program should only return the single best alignment (which is a very common usage), it can halt immediately when it finds a perfect match to a query. However, if the program must return a mapping quality, it needs to search – in every case – for at least one more alignment, in order to estimate the likelihood that the read should map to a different location. Therefore, mappers that do not provide alignment qualities or that use a random genome as proxy, like SHRiMP, have an inherent speed

advantage, although they do not provide a crucial output that many downstream analysis systems require.

VII. Beyond String Matching

A. Paired-End Reads

One important aspect in which the task of mapping short reads differs from the (approximate) string matching problem as modelled in Section I-B, is the availability of read pair information. As mentioned in Section I-A1, current instruments typically sequence both ends of the original DNA fragment, providing valuable information concerning the relative orientation and distance of the mates.

Some read mappers align both mates independently and then employ the pairing constraints to guide the selection of the true read origin among a set of candidate alignments. More specifically, Bowtie 2, BWA, SOAP2, MAQ, GEM, ZOOM, and CUSHAW2 constrain a pair of mates to align in a consistent orientation and separated by a distance that is concordant with an empirically estimated fragment length distribution. Pairs of alignments that violate these constraints are either marked as discordant in the output, or removed entirely. The orientation and distance constraints often help to resolve a repetitive read if its mate can be aligned unambiguously. BWA-MEM goes one step further and selects one best pair based on a score that considers orientation and distance of the mates, as well as the individual alignment scores.

Besides improving the accuracy of the mapping, read pair information can also be used to reduce the computational cost of alignment. In one strategy, candidate alignment locations are verified only if they satisfy the read pairing contraints (e.g., RazerS, Hobbes). In another strategy, a mapped read defines an anchor that restricts the search for aligning its mate to a limited genomic region (e.g., Bowtie 2, CUSHAW). The latter approach also allows the aligner to spend more resources aligning reads by applying the computationally expensive Smith-Waterman alignment algorithm to the genomic region defined by a successfully mapped read and the likely distance to its mate (e.g., BWA-SW, BWA-MEM, GEM, SOAP3-dp, CUSHAW2, CUSHAW2-GPU).

B. Horizontal and Vertical Integration of Alignments

1) From Pairwise to Multiple Alignments—Another simplification that is implied by the (approximate) string matching abstraction is the *independent* alignment of individual reads or read pairs. A single base in the genome is usually contained in more than one read. The read alignment problem can therefore be seen as a *multiple sequence alignment* problem. Since a global multiple alignment of all reads is computational infeasible, and since most of the reads do not overlap, the simplification to pairwise alignments employed by all current read alignment tools is well justified.

Nevertheless, the context of multiple alignment can be used in *local re-alignments* to improve the accuracy of the mapping. For instance, in contrast to sequencing errors, a true variant in the sequenced individual compared to the reference is expected to cause read alignments spanning this variant to *consistently* indicate the variant. While an independent

pairwise aligment of reads cannot exploit such a signal, a local re-alignment (see GATK [25]) performing a *multiple alignment* of sequences in the vicinity of the variant in question (particularly indels) can partially distinguish between true variants and sequencing errors.

Similarly, Scalpel [90] combines the information provided by multiple reads through a microassembly of reads initially mapped to regions of interest (e.g. exons). Based on a gapped alignment of the assembled sequences, insertions and deletions can be called with higher accuracy.

2) Vertical Integration—TotalReCaller [86] employs the reference genome sequence during base calling (see Section I-A1) by simultaneously aligning the partially generated read sequence that grows base-by-base. This integration of base-calling and alignment decreases the error rate of the former which in turn allows to align more reads back to the reference. The resulting bias towards the reference genome, however, might lower the sensitivity in detecting true single nucleotide polymorphisms.

VIII. Conclusion

At least 70 read mapping tools have been published since the advent of NGS technology. Although many of them make valuable theoretical and methodological contributions, only very few software tools are routinely used in the analysis of the enormous amounts of NGS data that are being generated at an ever increasing speed throughout the world. The reasons are manifold, but the overarching challenge is to develop sophisticated computational algorithms that also address the practical constraints imposed by DNA sequencing technology and the vagaries of biological experimentation.

The first and most obvious criterion for the practical utility of a read mapper is its *efficiency*, both in time and space. For these very large data sets, even constant factors matter, and an algorithm that runs twice as fast can save literally years of CPU time. Space matters as well, in part because the most common “text”, the human genome, is 3 billion bases long. An index that fails to represent the text efficiently may require tens of gigabytes of real memory, an amount that exceeds the standard memory available on most computing grids. This explains why the BWT-based FM-index, first implemented in the Bowtie and BWA read mappers, was adopted almost immediately by large numbers of users.

The second prerequisite for the widespread use of mapping software is *accuracy.* On the one hand, the true origin of a read is expected to lie in the set of candidate positions returned by the software. On the other hand, a large number of false positive alignments will essentially hide the true origin from downstream analysis. To map reads with high accuracy, the underlying (error) model must take into account the specifics of the sequencing technology as well as additional data provided by the instrument, such as read pairing information and base quality values. The confidence that the mapper has found the true origin of a read should be reflected by the mapping quality, a valuable piece of information that some downstream analyses rely upon.

Third, the *usability* of the software plays a significant role in its adoption. Typically, scientists who develop and write alignment software and those who run the software have

very different backgrounds and skills. Thus, software must be easy to use and well documented, ideally providing some form of user support. For the most popular software packages, user networks have emerged to provide support to one another.

Fourth, *maintenance* is absolutely critical for a package to maintain its usefulness. Although less glamorous than the original development work, maintenance is required to keep up with changes in sequencing technology, changes in underlying operating systems, and the ever-changing ways in which sequence data is used. For example, the leading aligners in 2009 were optimized for read lengths of 35 bp, which was the standard length at the time. This quickly increased to 75 and then 100 bp, and more recently jumped to 300 bp with the new (but lower throughput) MiSeq instrument. Aligners had to modify their seed lengths and other internal methods in order to adjust to these longer reads. At the same time, many users continue to use older, shorter read technology, requiring alignment developers to maintain older versions of their systems simultaneously with the new versions.

As human sequencing work increases, the limitations of using a single reference genome have become apparent. In the near future, we are likely to see additional reference human genomes representing sub-populations or ethnic groups, which in turn may facilitate the analysis of personal genomes [4]. As these become available, it may be useful (see e.g. [102]) to align reads to multiple genomes simultaneously. This in turn could eliminate the bias towards a single reference and thus improve the accuracy of read mapping. The growing gap between sequencing capacity and computing power, however, needs be filled by clever (mapping) algorithms, that scale sublinearly with genomic data size. The redundancy inherent in collections of genomes can be exploited by storing similarities and variations in a compressed format on which alignment algorithms might operate directly [79].

In the near future, alignment algorithms will not only have to cope with exponentially increasing data volumes, but also with changing data attributes that might require the development of novel techniques. Most notably, the trend towards longer reads as provided by single-molecule sequencing technologies shifts the DNA read alignment problem slightly towards the problem of aligning whole genomes, which has been studied before. If, on the other hand, longer reads are obtained from RNA transcripts, the mapping algorithms must account for a large number of introns, which might decisively change the nature of the RNA read mapping problem. Furthermore, the higher information content that a longer read carries by itself might allow one to intertwine the read mapping problem and methods that are currently distinct, such as the assembly of full-length transcripts from RNA-seq data.

Acknowledgments

This work was supported in part by the U.S. National Institutes of Health under grant R01 HG006102 to SLS.

Biographies

Steven L. Salzberg received his B.A. degree in English and M.S. and M.Phil. degrees in Computer Science from Yale University, and his Ph.D. in Computer Science from Harvard University in 1989.

From 1989–1998 he was a faculty member in Computer Science at Johns Hopkins University, and from 1997–2005 he was Senior Director of Bioinformatics at The Institute for Genomic Research (TIGR) in Rockville, Maryland, one of the world's leading DNA sequencing centers at the time. From 2005–2011, he was the Director of the Center for Bioinformatics and Computational Biology (CBCB) and the Horvitz Professor of Computer Science at the University of Maryland, College Park. He is currently the Bloomberg Distinguished Professor of Biomedical Engineering, Computer Science, and Biostatistics and the Director of the Center for Computational Biology in the McKusick-Nathans Institute of Genetic Medicine at Johns Hopkins University.

Dr. Salzberg has authored or co-authored over 250 publications in leading scientific journals, and his h-index is 113. He is a Fellow of the American Association for the Advancement of Science, a Fellow of the International Society for Computational Biology, and a former member of the Board of Scientific Counselors of the National Center for Biotechnology Information at NIH. He was the 2013 winner of the Benjamin Franklin Award for Open Access in the Life Sciences, and the 2013 winner of the Robert G. Balles Prize in Critical Thinking for his Forbes science column. In 2001 and again in 2014 he was listed as a Highly Cited Researcher by Thomson Reuters, a compilation of the 1% most-cited researchers in the world.

Stefan Canzar received his Diploma in Computer Science from the Technische Universität München, Munich, Germany, in 2004, and his bi-national doctoral degree in Computer Science from Universität des Saarlandes, Saarbrücken, Germany, and Université Henri Poincaré, Nancy, France, in 2008.

From 2009 to 2011 and from 2012 to 2014,he was a postdoctoral researcher at the Centrum Wiskunde & Informatica, Amsterdam, The Netherlands, and the Johns Hopkins Institute of Genetic Medicine, Baltimore, respectively. Since 2014, he has been a Research Assistant Professor with the Toyota Technological Insitute at Chicago. His research interests include the development of algorithmic solutions to problems arising in the analysis of high-throughput sequencing data. The goal of his research is to develop advanced computational approaches that help to transform the generated data into information and ultimately knowledge in research and medicine.

References

1. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. Oct.2010 467(7319):1061–1073. [PubMed: 20981092]
2. Abouelhoda, MI., Ohlebusch, E. A local chaining algorithm and its applications in comparative genomics. In: Benson, G., Page, RDM., editors. WABI, volume 2812 of Lecture Notes in Computer Science. Springer; 2003. p. 1-16.
3. Ahmadi A, Behm A, Honnalli N, Li C, Weng L, Xie X. Hobbes: optimized gram-based methods for efficient read alignment. Nucleic Acids Research. 2011
4. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim DS, Kim BC, Kim SY, Kim WY, Kim C, Park D, Lee YS, Kim S, Reja R, Jho S, Kim CG, Cha JY, Kim KH, Lee B, Bhak J, Kim SJ. The first korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. Genome Research. 2009; 19(9):1622–1629. [PubMed: 19470904]

5. Aji A, Zhang L, chun Feng W. GPU-RMAP: Accelerating Short-Read Mapping on Graphics Processors. Computational Science and Engineering (CSE), 2010 IEEE 13th International Conference on. 2010:168–175.
6. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, Sahinalp SC, Gibbs RA, Eichler EE. Personalized copy number and segmental duplication maps using next-generation sequencing. Nature genetics. Oct.2009 41(10):1061–1067. [PubMed: 19718026]
7. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. Oct.1990 215(3):403–410. [PubMed: 2231712]
8. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. Sep; 1997 25(17):3389–3402. [PubMed: 9254694]
9. Apostolico, A. volume 12 of NATO Advance Science Institute Series. Berlin: Springer Verlag; 1985. The Myriad Virtues of Suffix Trees; p. 85-95.Series F: Computer and Systems Sciences
10. Baeza-Yates RA, Perleberg CH. Fast and practical approximate string matching. Inf Process Lett. 1996; 59(1):21–27.
11. Blom J, Jakobi T, Doppmeier D, Jaenicke S, Kalinowski J, Stoye J, Goesmann A. Exact and complete short-read alignment to microbial genomes using graphics processing unit programming. Bioinformatics. 2011; 27(10):1351–1358. [PubMed: 21450712]
12. Blumer A, Blumer J, Haussler D, Ehrenfeucht A, Chen MT, Seiferas JI. The smallest automaton recognizing the subwords of a text. Theor Comput Sci. 1985; 40:31–55.
13. Boyer RS, Moore JS. A fast string searching algorithm. Commun ACM. Oct.1977 20(10):762–772.
14. Brodnik A, Munro JI. Membership in constant time and almostminimum space. SIAM J Comput. 1999; 28(5):1627–1640.
15. Burkhardt S, Crauser A, Ferragina P, Lenhof HP, Rivals E, Vingron M. *q*-gram based database searching using a suffix array (QUASAR). RECOMB. 1999:77–83.
16. Burkhardt, S., Kärkkäinen, J. Better filtering with gapped q-grams. In: Amir, A., Landau, GM., editors. CPM, volume 2089 of Lecture Notes in Computer Science. Springer; 2001. p. 73-85.
17. Burrows M, Wheeler DJ. A block-sorting lossless data compression algorithm. Technical Report 124, Digital SRC Research Report. 1994
18. Campagna D, Albiero A, Bilardi A, Caniato E, Forcato C, Manavski S, Vitulo N, Valle G. Pass: a program to align short sequences. Bioinformatics. 2009; 25(7):967–968. [PubMed: 19218350]
19. Chaisson M, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (blasr): Theory and application. BMC Bioinformatics. 2012; 13:238. [PubMed: 22988817]
20. Chen S, Wang A, Li LM. SEME: A Fast Mapper of Illumina Sequencing Reads with Statistical Evaluation. Journal of Computational Biology. Nov.2013 20(11):847–860. [PubMed: 24195707]
21. Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Meth. Jul; 2008 5(7):613–619.
22. David M, Dzamba M, Lister D, Ilie L, Brudno M. Shrimp2: Sensitive yet practical short read mapping. Bioinformatics. 2011; 27(7):1011–1012. [PubMed: 21278192]
23. Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters. Commun ACM. Jan.2008 51(1):107–113.
24. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. Alignment of whole genomes. Nucleic Acids Res. 1999; 27(11):2369–76. [PubMed: 10325427]
25. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation dna sequencing data. Nat Genet. May; 2011 43(5):491–498. [PubMed: 21478889]
26. Doolittle R, Hunkapiller M, Hood L, Devare S, Robbins K, Aaronson S, Antoniades H. Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. Science. 1983; 221(4607):275–277. [PubMed: 6304883]

27. 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, Birne y E, Hubbard TJ, Durbin R, Tavaré S, Beck S. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nature biotechnology. Jul; 2008 26(7):779–785.
28. Eppstein D, Galil Z, Giancarlo R, Italiano GF. Sparse dynamic programming i: Linear cost functions. J ACM. Jul; 1992 39(3):519–545.
29. Ewing B, Hillier L, Wendl M, Green P. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Research. Mar.1998 8(3):175–185. [PubMed: 9521921]
30. Farrar M. Striped smith-waterman speeds database searches six times over other simd implementations. Bioinformatics. Jan.2007 23(2):156–161. [PubMed: 17110365]
31. Farrar M. Striped smith-waterman speeds database searches six times over other simd implementations. Bioinformatics. 2007; 23(2):156–161. [PubMed: 17110365]
32. Ferragina P, González R, Navarro G, Venturini R. Compressed text indexes : From theory to practice. J Exp Algorithmics. Feb.2009 13:12:1.12–12:1.31.
33. Ferragina, P., Manzini, G. Proceedings of the 41st Annual Symposium on Foundations of Computer Science, FOCS '00. Washington, DC, USA: IEEE Computer Society; 2000. Opportunistic data structures with applications; p. 390
34. Ferragina P, Manzini G. Indexing compressed text. J ACM. 2005; 52(4):552–581.
35. Ferragina P, Manzini G, Mäkinen V, Navarro G. Compressed representations of sequences and full-text indexes. ACM Trans Algorithms. May.2007 3(2)
36. Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. Bioinformatics. 2012; 28(24):3169–3177. [PubMed: 23060614]
37. Gontarz PM, Berger J, Wong CF. Srmapper: a fast and sensitive genome-hashing alignment tool. Bioinformatics. 2013; 29(3):316–321. [PubMed: 23267171]
38. Gräf S, Nielsen FGG, Kurtz S, Huynen MA, Birney E, Stunnenberg H, Flicek P. Optimized design and assessment of whole genome tiling arrays. Bioinformatics. 2007; 23(13):i195–i204. [PubMed: 17646297]
39. Grossi, R., Vitter, JS. Compressed suffix arrays and suffix trees with applications to text indexing and string matching (extended abstract). In: Yao, FF., Luks, EM., editors. STOC. ACM; 2000. p. 397-406.
40. Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC. mrsFAST: a cache-oblivious algorithm for short-read mapping. Nature methods. Aug.2010 7(8):576–577. [PubMed: 20676076]
41. Hatem A, Bozdag D, Toland A, Catalyurek U. Benchmarking short sequence mapping tools. BMC Bioinformatics. 2013; 14(1):184. [PubMed: 23758764]
42. Healy J, Thomas EE, Schwartz JT, Wigler M. Annotating large genomes with exact word matches. Genome research. Oct.2003 13(10)
43. Homer N, Merriman B, Nelson SF. BFAST: An Alignment Tool for Large Scale Genome Resequencing. PLoS ONE. Nov.2009 4(11):e7767+. [PubMed: 19907642]
44. Hyyrö H. A bit-vector algorithm for computing levenshtein and damerau edit distances. Nordic J of Computing. Mar.2003 10(1):29–39.
45. Jiang H, Wong WH. Seqmap: mapping massive amount of oligonucleotides to the genome. Bioinformatics. 2008; 24(20):2395–2396. [PubMed: 18697769]
46. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-dna interactions. Science. 2007; 316(5830):1497–1502. [PubMed: 17540862]
47. Jokinen P, Ukkonen E. Two algorithms for approximate string matching in static texts. MFCS. 1991:240–248.
48. Kärkkäinen J, Na JC. Faster filters for approximate string matching. ALENEX SIAM. 2007
49. Kärkkäinen, J., Ukkonen, E. Proc 3rd South American Workshop on String Processing (WSP'96. Carleton University Press; 1996. Lempel-ziv parsing and sublinear-size index structures for string matching (extended abstract); p. 141-155.

50. Kent WJ. Blatthe blast-like alignment tool. Genome Research. 2002; 12(4):656–664. [PubMed: 11932250]
51. Kim J, Li C, Xie X. Improving read mapping using additional prefix grams. BMC Bioinformatics. 2014; 15(1):42. [PubMed: 24499321]
52. Klus P, Lam S, Lyberg D, Cheung M, Pullan G, McFarlane I, Yeo G, Lam B. BarraCUDA - a fast short read sequence aligner using graphics processing units. BMC Research Notes. 2012; 5(1):27. [PubMed: 22244497]
53. Knuth DE, Morris JH, Pratt VR. Fast Pattern Matching in Strings. SIAM Journal on Computing. Mar; 1977 6(2):323–350.
54. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Phillippy AM. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotech. Jul; 2012 30(7):693–700.
55. Kucherov, G., No, L., Roytberg, M. Multi-seed lossless filtration. In: Sahinalp, S.Muthukrishnan, S., Dogrusoz, U., editors. Combinatorial Pattern Matching, volume 3109 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2004. p. 297-310.
56. Lam T, Li R, Tam A, Wong S, Wu E, Yiu S. High throughput short read alignment via bi-directional BWT. Bioinformatics and Biomedicine, 2009 BIBM '09 IEEE International Conference on. 2009:31–36.
57. Lam TW, Sung WK, Tam SL, Wong CK, Yiu SM. Compressed indexing and local alignment of DNA. Bioinformatics. Mar.2008 24(6):791–797. [PubMed: 18227115]
58. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. Apr; 2012 9(4):357–359. [PubMed: 22388286]
59. Langmead B, Schatz M, Lin J, Pop M, Salzberg S. Searching for snps with cloud computing. Genome Biol. 2009; 10(11):R134. [PubMed: 19930550]
60. Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009; 10(3):R25. [PubMed: 19261174]
61. Lee WP, Stromberg M, Ward A, Stewart C, Garrison E, Marth GT. MOSAIK: A hash-based algorithm for accurate next-generation sequencing read mapping. ArXiv e-prints. Sep.2013
62. Levenshtein V. Binary codes capable of correcting spurious insertions and deletions of ones. Problems of Information Transmission. 1965; 1:8–17.
63. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. Jul; 2009 25(14):1754–1760. [PubMed: 19451168]
64. Li H, Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. Bioinformatics. 2010; 26(5):589–595. [PubMed: 20080505]
65. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008; 18:1851. [PubMed: 18714091]
66. Li M, Ma B, Kisman D, Tromp J. Patternhunter ii: Highly sensitive and fast homology search. J Bioinformatics and Computational Biology. 2004; 2(3):417–440. [PubMed: 15359419]
67. Li R, Li Y, Kristiansen K. SOAP: short oligonucleotide alignment program. Bioinformatics. 2008; 24(5):713–714. J. W. 0004. [PubMed: 18227114]
68. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. Soap2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009; 25(15):1966–1967. [PubMed: 19497933]
69. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang Jian an d. De novo assembly of human genomes with massively parallel short read sequencing. Genome research. Feb.2010 20(2)
70. Liao Y, Smyth GK, Shi W. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research. 2013; 41(10):e108. [PubMed: 23558742]
71. Lin H, Zhang Z, Zhang MQ, Ma B, Li M. Zoom! zillions of oligos mapped. Bioinformatics. 2008; 24(21):2431–2437. [PubMed: 18684737]
72. Lippert RA. Space-efficient whole genome comparisons with Burrows-Wheeler transforms. Journal of Computational Biology. May; 2005 12(4):407–415. [PubMed: 15882139]

73. 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. May; 2008 133(3): 523–536. [PubMed: 18423832]
74. Liu CM, Wong T, Wu E, Luo R, Yiu SM, Li Y, Wang B, Yu C, Chu X, Zhao K, Li R, Lam TW. SOAP3 : ultra-fast GPU-based parallel alignment tool for short reads. Bioinformatics. 2012; 28(6): 878–879. [PubMed: 22285832]
75. Liu Y, Maskell D, Schmidt B. CUDASW++: optimizing Smith-Waterman sequence database searches for CUDA-enabled graphics processing units. BMC Research Notes. 2009; 2(1):73. [PubMed: 19416548]
76. Liu Y, Schmidt B. CUSHAW2-GPU: empowering faster gapped short-read alignment using GPU computing. IEEE Design & Test of Computers. 2014; 31(1):31–39.
77. Liu Y, Schmidt B, Maskell D. CUDASW++2.0: enhanced Smith-Waterman protein database search on CUDA-enabled GPUs based on SIMT and virtualized SIMD abstractions. BMC Research Notes. 2010; 3(1):93. [PubMed: 20370891]
78. Liu Y, Schmidt B, Maskell DL. CUSHAW: a CUDA compatible short read aligner to large genomes based on the Burrows-Wheeler transform. Bioinformatics. 2012
79. Loh PR, Baym M, Berger B. Compressive genomics. Nature Biotechnology. Jul; 2012 30(7):627–630.
80. Lunter G, Goodson M. Stampy: A statistical algorithm for sensitive and fast mapping of illumina sequence reads. Genome Research. 2010
81. Luo R, Wong T, Zhu J, Liu CM, Zhu X, Wu E, Lee LK, Lin H, Zhu W, Cheung DW, Ting H-F, Yiu SM, Peng S, Yu C, Li Y, Li R, Lam TW. SOAP3-dp: Fast, Accurate and Sensitive GPU-Based Short Read Aligner. PLoS ONE. 2013; 8(5):e65632. [PubMed: 23741504]
82. Ma B, Tromp J, Li M. Patternhunter: faster and more sensitive homology search. Bioinformatics. 2002; 18(3):440–445. [PubMed: 11934743]
83. Mäkinen, V., Välimäki, N., Laaksonen, A., Katainen, R. Unified View of Backward Backtracking in Short Read Mapping. In: Elomaa, T.Mannila, H., Orponen, P., editors. Algorithms and Applications, volume 6060 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg; Berlin, Heidelberg: 2010. p. 182-195.chapter 13
84. Manber U, Myers EW. Suffix arrays: A new method for on-line string searches. SIAM J Comput. 1993; 22(5):935–948.
85. Marco-Sola S, Sammeth M, Guigo R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Meth. Dec.2012 9(12):1185–1188.
86. Menges F, Narzisi G, Mishra B. TotalReCaller: improved accuracy and performance via integrated alignment and base-calling. Bioinformatics. 2011; 27(17):2330–2337. [PubMed: 21724593]
87. Misra S, Agrawal A, Liao W-k, Choudhary A. Anatomy of a hash-based long read sequence mapping algorithm for next generation dna sequencing. Bioinformatics. 2011; 27(2):189–195. [PubMed: 21088030]
88. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by rna-seq. Nat Meth. Jul; 2008 5(7):621–628.
89. Myers G. A fast bit-vector algorithm for approximate string matching based on dynamic programming. J ACM. 1999; 46(3):395–415.
90. Narzisi G, O'Rawe JA, Iossifov I, Fang H, Lee Y-H, Wang Z, Wu Y, Lyon GJ, Wigler M, Schatz MC. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. Nature Methods. Oct.2014 11(10):1033–1036. [PubMed: 25128977]
91. Navarro G, Mäkinen V. Compressed full-text indexes. ACM Comput Surv. Apr.2007 39(1)
92. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology. 1970; 48(3):443–453. [PubMed: 5420325]
93. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large dna databases. Genome Res. Oct.2001 11(10):1725–9. [PubMed: 11591649]
94. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences of the United States of America. Apr.1988 85(8):2444–2448. [PubMed: 3162770]

95. Phillippy AM, Deng X, Zhang W, Salzberg SL. Efficient oligonucleotide probe selection for pan-genomic tiling arrays. BMC Bioinformatics. 2009; 10:293. [PubMed: 19758451]
96. Rasmussen KR, Stoye J, Myers EW. Efficient q-gram filters for finding all epsilon-matches over a given length. Journal of Computational Biology. 2006; 13(2):296–308. [PubMed: 16597241]
97. Rizk G, Lavenier D. Gassst: global alignment short sequence search tool. Bioinformatics. 2010; 26(20):2534–2540. [PubMed: 20739310]
98. Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. SHRiMP: Accurate Mapping of Short Color-space Reads. PLoS Comput Biol. May.2009 5(5):e1000386+. [PubMed: 19461883]
99. Sankoff, D., Kruskal, JB., editors. Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. Addison-Wesley; Reading, MA: 1983.
100. Schatz M, Delcher A, Salzberg S. Assembly of large genomes using second-generation sequencing. Genome Research. 2010
101. Schatz MC. CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics. 2009; 25(11):1363–1369. [PubMed: 19357099]
102. Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, Kohlbacher O, Weigel D. Simultaneous alignment of short reads against multiple genomes. Genome Biology. 2009; 10(9):R98. [PubMed: 19761611]
103. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. Human-Mouse Alignments with BLASTZ. Genome Research. 2003; 13(1):103–107. [PubMed: 12529312]
104. Sedlazeck FJ, Rescheneder P, von Haeseler A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics. 2013; 29(21):2790–2791. [PubMed: 23975764]
105. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome research. Mar.2012 22(3):549–556. [PubMed: 22156294]
106. Siragusa E, Weese D, Reinert K. Fast and accurate read mapping with approximate seeds and multiple backtracking. Nucleic Acids Research. 2013; 41(7):e78. [PubMed: 23358824]
107. Smith A, Xuan Z, Zhang M. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics. Feb.2008 9(1):128+. [PubMed: 18307793]
108. Smith AD, Chung WY, Hodges E, Kendall J, Hannon G, Hicks J, Xuan Z, Zhang MQ. Updates to the RMAP short-read mapping software. Bioinformatics. 2009; 25(21):2841–2842. [PubMed: 19736251]
109. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of molecular biology. Mar.1981 147(1):195–197. [PubMed: 7265238]
110. Sun Y, Buhler J. Designing multiple simultaneous seeds for dna similarity search. Journal of Computational Biology. 2005; 12(6):847–861. [PubMed: 16108721]
111. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Bjorklund M, Wei G, Yan J, Niittymaki I, Mecklin JP, Jarvinen H, Ristimaki A, Di-Bernardo M, East P, Carvajal-Carmona L, Houlston RS, Tomlinson I, Palin K, Ukkonen E, Karhu A, Taipale J, Aaltonen LA. The common colorectal cancer predisposition snp rs6983267 at chromosome 8q24 confers potential to enhanced wnt signaling. Nat Genet. Aug; 2009 41(8):885–890. [PubMed: 19561604]
112. Ukkonen, E. Approximate string-matching over suffix trees. In: Apostolico, A.Crochemore, M.Galil, Z., Manber, U., editors. CPM, volume 684 of Lecture Notes in Computer Science. Springer; 1993. p. 228-242.
113. Wang W, Zhang P, Liu X. Short read dna fragment anchoring algorithm. BMC Bioinformatics. 2009; 10(S-1)
114. Warren, HS. Hacker’s Delight. Addison-Wesley Longman Publishing Co., Inc.; Boston, MA, USA: 2002.
115. Weese D, Emde AK, Rausch T, Dring A, Reinert K. Razersfast read mapping with sensitivity control. Genome Research. 2009; 19(9):1646–1654. [PubMed: 19592482]
116. Weese D, Holtgrewe M, Reinert K. Razers 3: Faster, fully sensitive read mapping. Bioinformatics. 2012
117. Wilbur WJ, Lipman DJ. Rapid similarity searches of nucleic acid and protein data banks. PNAS. Feb; 1983 80(3):726–730. [PubMed: 6572363]

118. Wilton R, Budavari T, Langmead B, Wheelan S, Salzberg SL, Szalay AS. Arioc: high-throughput read alignment with GPU-accelerated exploration of the seed-and-extend search space. PeerJ. 2015

119. Xin H, Lee D, Hormozdiari F, Yedkar S, Mutlu O, Alkan C. Accelerating read mapping with fasthash. BMC Genomics. 2013; 14(S-1):S13.

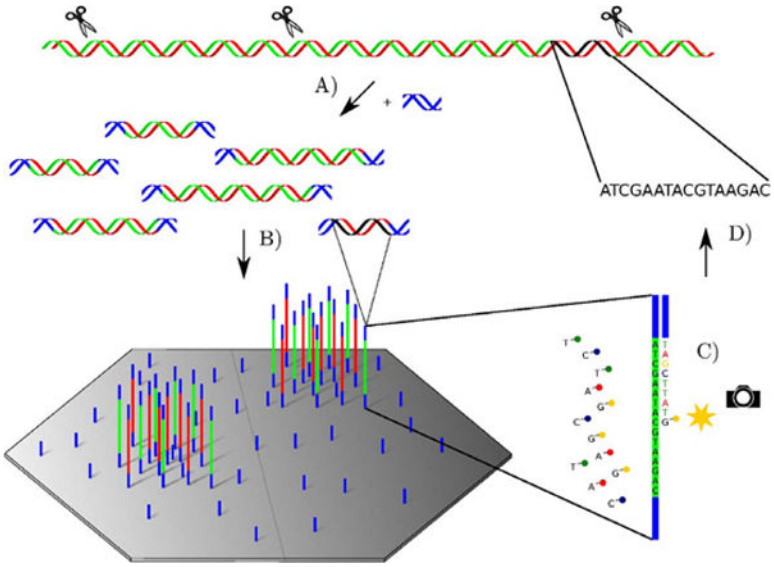

Fig. 1.

Workflow of next-generation sequencing. (A) Genomic DNA is sheared into fragments, and platform-specific adapter sequences (blue) are then attached. After amplification on a solid surface (B), the sequence of nucleotides is "read out" (C) from signals emitted when a base is added to the complement of a template strand. A read mapping algorithm then must find (D) for the genomic origin of the resulting reads.

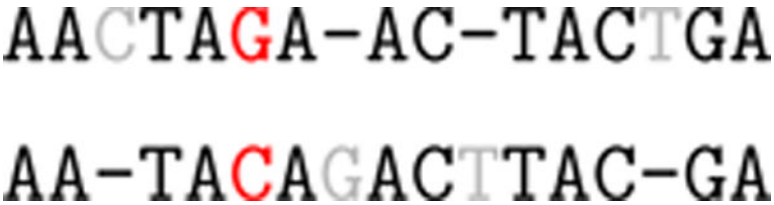

Fig. 2.
An alignment of two strings, implying one substitution (red) and four insertions/deletions (gray).

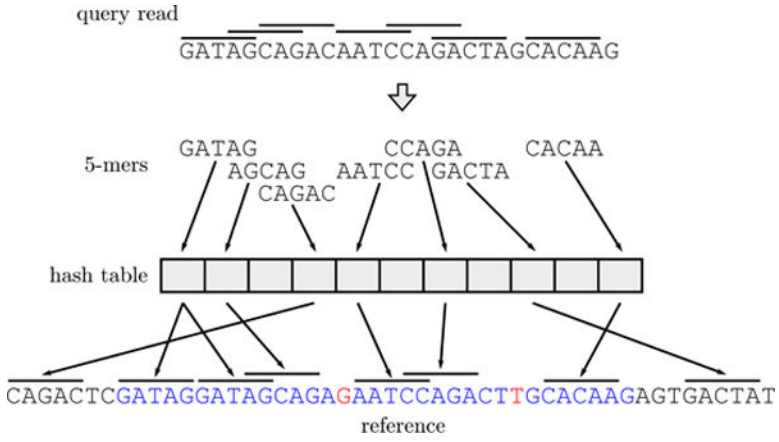

Fig. 3.

Workflow of hashing-based methods. First, *k*-mers are extracted from the query read applying seed templates. Here, *k*-mers correspond to substrings of the read, indicated by black lines. Typically the seed template is applied at all positions of the read. A hash table returns the position of *k*-mer occurrences in the genome. In the example shown, an approximate match (blue) of the query read with 2 mismatches (red) is hit by six seeds.

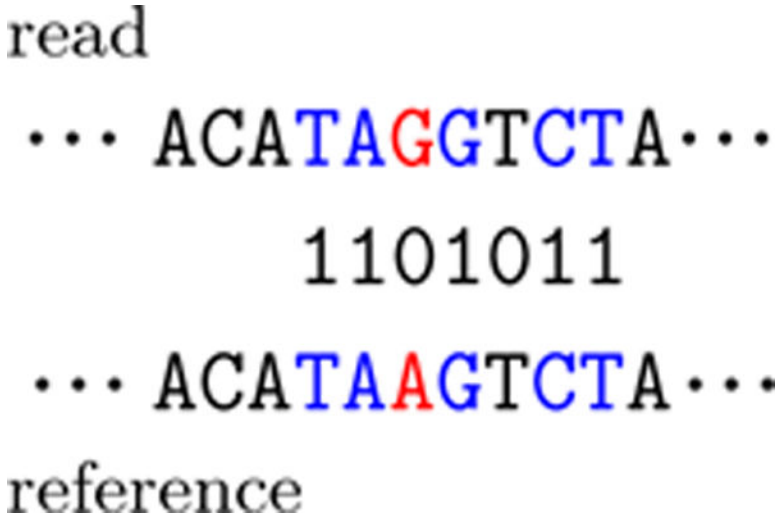

Fig. 4.
The 5-mer TAGCT matched by the seed template 1101011 is effectively TANGNCT, where N's correspond to "don't care" positions. Mismatches (red) at these positions are ignored.

Fig. 5.
The single mismatch does not affect 3 *q*-grams (blue lines) but invalidates 3 seed matches (red lines).

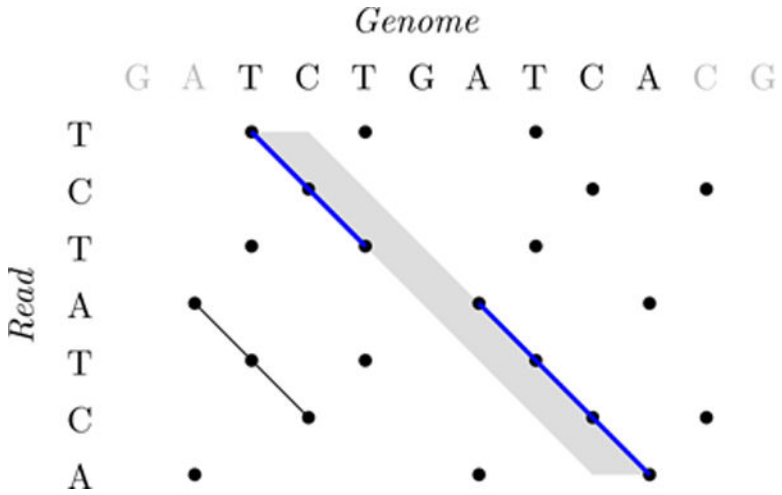

Fig. 6.

Dot plot between read TCTATCA and its approximate occurrence of Levenshtein distance 1 in the genome. Dots mark identical nucleotides in the two sequences written along the vertical and horizontal axes. The parallelogram (grey) spans 8 columns and 2 diagonals and contains three 3-hits (blue).

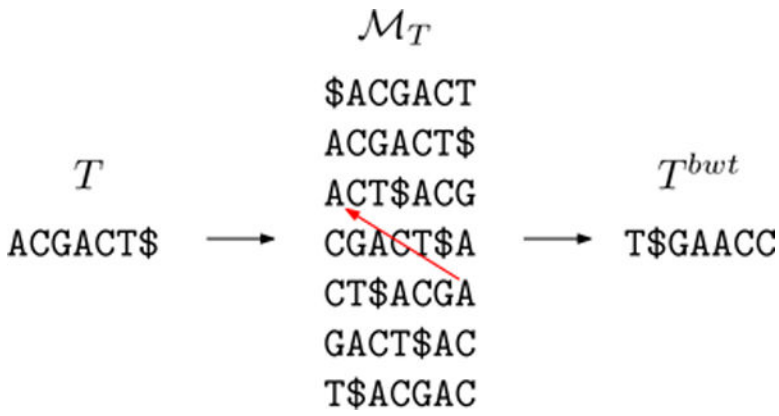

Fig. 7.

Constructing the BWT of ACGACT. *LF* mapping (red arrow): The second occurrence of A in the last column corresponds to the second occurrence of A in the first column.

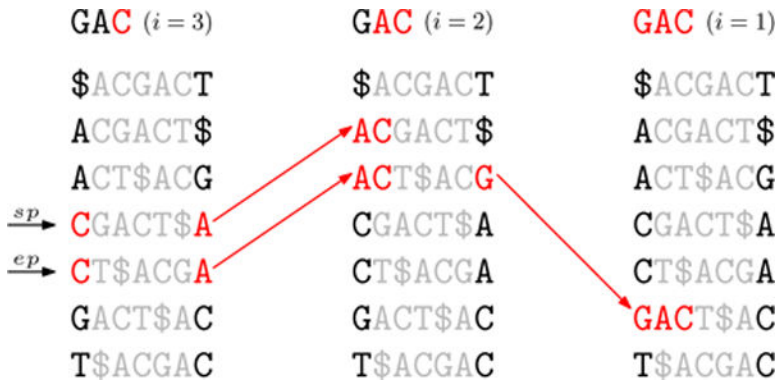

Fig. 8.

Backward search of P = GAC in BWT of T = ACGACT. The red arrows denote the update of *sp* and *ep* in lines 3 and 4 of Algorithm III-B.

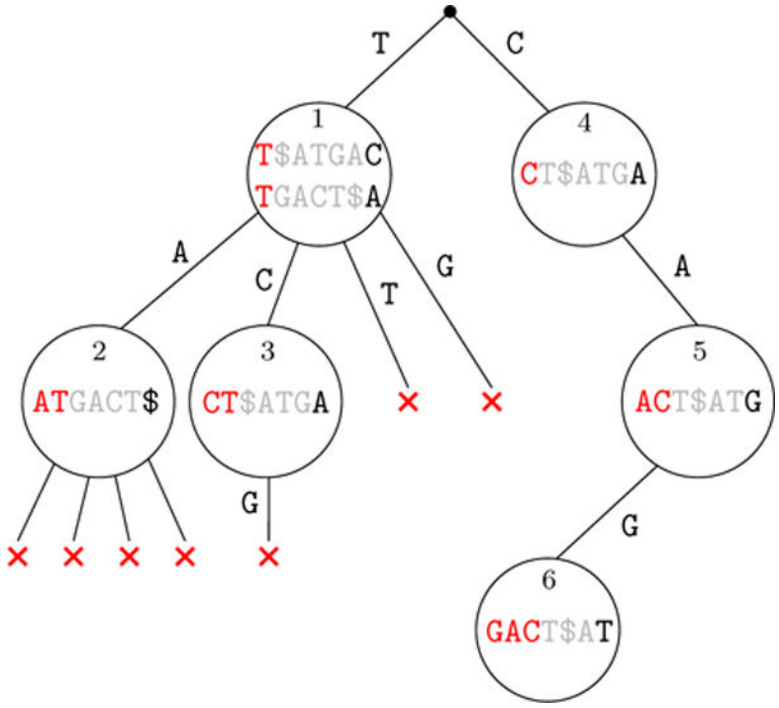

Fig. 9.

Recursion tree as traversed by Algorithm *k*MISMATCHSEARCH for P = GAT, T = ATGACT, and k = 1. Nodes contain strings falling in current range [*sp, ep*] and are expanded in depth-first manner as indicated by their numbering. Red crosses denote pruning in line 1 of the algorithm. Notice that for $c \neq$ G the condition in line 1 evaluates to false in node 3. When reaching node 6, an occurrence with 1 mismatch has been found.

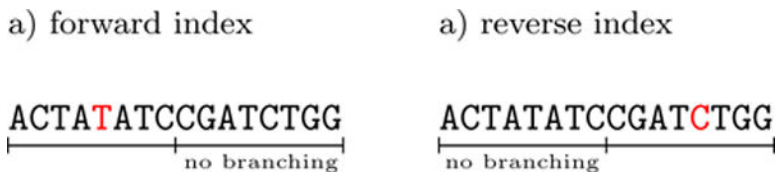

Fig. 10.

To minimize backtracking, the search direction in Bowtie depends on the location of the mismatch (red).

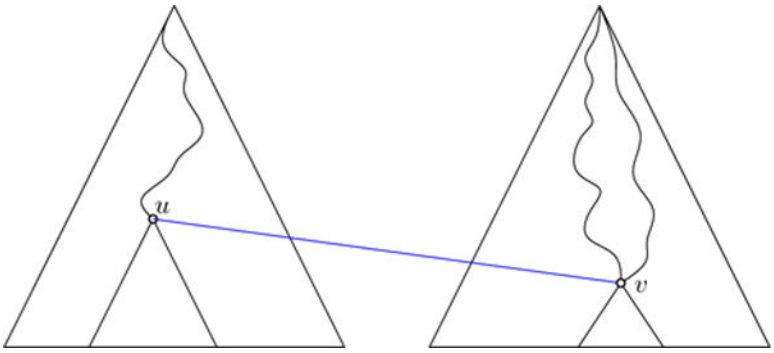

Fig. 11.

Alignment of a prefix trie representation T of the genome (left) and a DAWG G capturing the read sequence (right) by BWA-SW [64]. Starting at the roots of the graphs, the DP recursion relates the optimal alignment of substrings represented by a node u in T and a node v in G to the optimal alignment of substrings expressed by their parent nodes. Node u in T and v in G represent all occurrences of the string spelled by the (unique) u-to-root path in the genome and all occurrences of the strings spelled by the v-to-root paths in the read, respectively. The v-to-root paths in G by construction spell substrings of the read such that one is a prefix of the other.

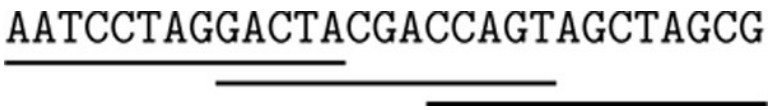

Fig. 12.
Seed Strings in Bowtie 2: In this example, 13 nt substrings spanned by the lines are extracted every 8 nt.

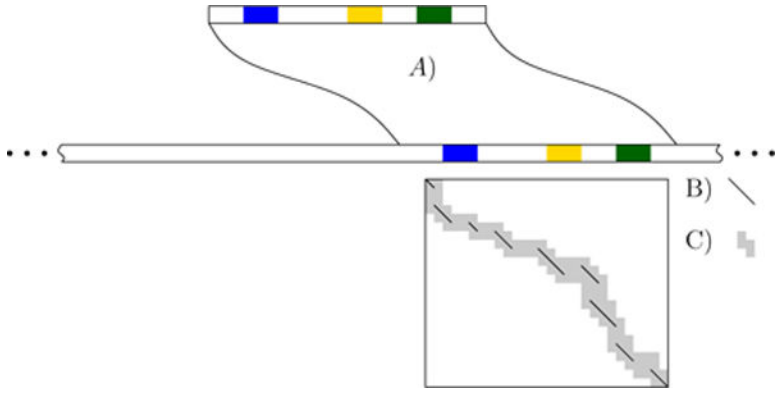

Fig. 13.

The three main steps performed by BLASR [19]: A) Candidate genomic regions are identified by clusters of consistent exact matchtes. B) A sparse dynamic program pre-aligns the read to the highest scoring candidate regions. C) Guided by the anchors found through the sparse alignment, a final base-resolution alignment is computed.

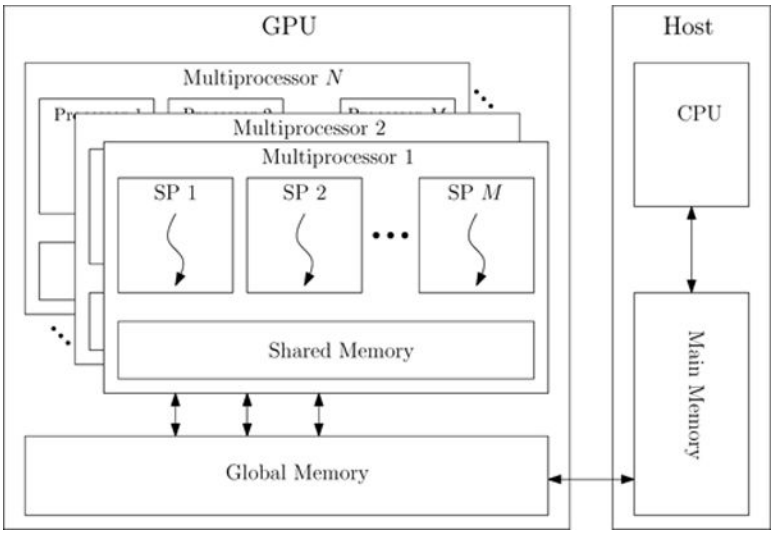

Fig. 14.

The architecture of a graphics processing unit (GPU). GPUs comprise a number N of multiprocessors that can simultaneously access the *global memory.* Multiprocessors consist of several *stream processors* (SP), or *cores*, that can all access the *shared memory* located on the same multiprocessor. The cores execute the sequential threads (curved arrows) in single instruction, multiple thread (SIMT) mode (see text).

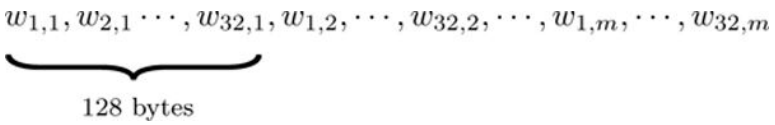

Fig. 15.

Arrangement of reads in memory by SOAP3 and SOAP3-dp. For $1 \leq j \leq m$, the jth 4-byte word $w_{i,j}$ of all reads i in a group of 32 reads occupy a contiguous 128-byte region, which can be retrieved by a single memory access operation.

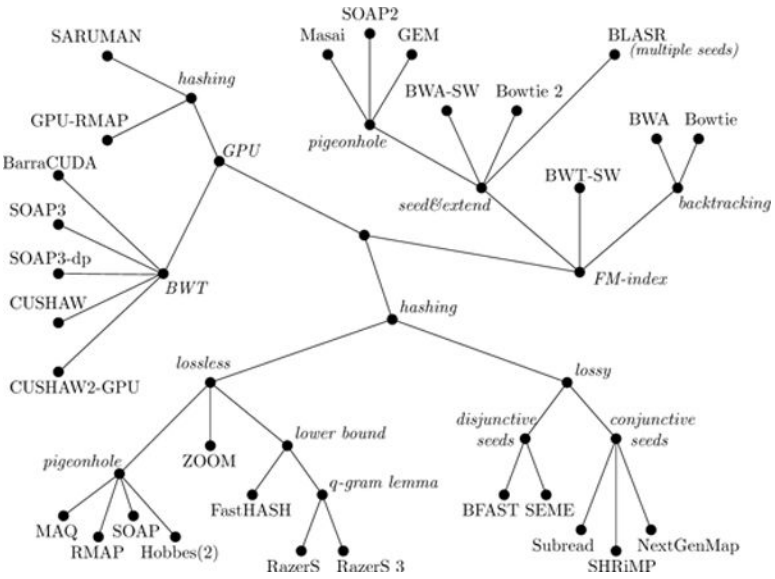

Fig. 16.

Methods at leaves of a tree that branches along algorithmic design decisions. The hierarchical structure roughly reflects the organization of Sections II–V.