Schwartz DC, Waterman MS. New generations: Sequencing machines and their computational challenges. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 25(1): 3–9 Jan. 2010

New Generations: Sequencing Machines and Their Computational Challenges

David C. Schwartz[1] and Michael S. Waterman[2,3]

[1] *Laboratory for Molecular and Computational Genomics, Department of Chemistry and Laboratory of Genetics University of Wisconsin-Madison, WI 53706, U.S.A.*

[2] *Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089-2910, U.S.A.*

[3] *Department of Automation, Tsinghua University, Beijing 100084, China*

E-mail: dcschwartz@facstaff.wisc.edu; msw@usc.edu

Received September 5, 2009; revised November 24, 2009.

Abstract New generation sequencing systems are changing how molecular biology is practiced. The widely promoted $1000 genome will be a reality with attendant changes for healthcare, including personalized medicine. More broadly the genomes of many new organisms with large samplings from populations will be commonplace. What is less appreciated is the explosive demands on computation, both for CPU cycles and storage as well as the need for new computational methods. In this article we will survey some of these developments and demands.

Keywords genome sequencing, new generation sequencing, read mapping, optical mapping, sequence assembly, Eulerian graphs

1 Introduction

It may be somewhat futile to attempt to track perfectly an explosion. But here we hope to give some hints about the technological and computational challenges that will surely be addressed along the path to the commoditization of sequence information. As the cost of sequence information drops, its utility will grow as sequencing directly alters medical care, the type and safety of our food supply, and of course, now unfathomable applications: who would have predicted 50 years ago that lasers would find broad application as "pointers"? Accordingly, we expect that the experimental and computational challenges will become progressively intermingled in ways that may foster development of completely new disciplines for tackling the even greater challenges that are now unthinkable. In this regard, we present here a brief overview of the current state of DNA sequencing, and our best guesses for how technology and computation may interact for creating this future.

2 Current Technology

Although commercial next generation platforms differ from each other in how sequence is actually obtained, they share the common advantage of not requiring bacterial clone libraries. In many ways, the obviation of clone library construction and handling is a major reason why genome sequencing costs have plummeted, while platform throughput is dramatically increasing. Templates for large scale DNA sequencing are made from a library spread across massive culture plates and individual clones are isolated by "picking robots" for downstream sequencing reactions. Such operations, for large genomes such as human, require factory floor settings bristling with robots and technicians before any sequencing data is acquired. In contrast, next generation platforms construct "clone" libraries directly from individual genomic DNA molecules, which are amplified by emulsion or bridge PCR (polymerase chain reaction). Entire genome libraries consist of small vesicles, or surfaces laden with amplicons, but there is one company[1] whose libraries comprise unamplified genomic templates that are bound to surfaces.

2.1 Next-Generation Sequencing

Today, an investigator can choose between four commercially available systems, each offering a panoply of technical strengths and weaknesses that need to be considered against overall cost and application: 1) Illumina's Genome Analyzer, 2) Life Technologies' SOLiD

Survey

This work is supported by NIH under Grant No. R01 HG000225 (DCS) and NSF of USA under Grant No. DBI-0501818 (DCS).

©2010 Springer Science + Business Media, LLC & Science Press, China

System, 3) Roche's 454 GS FLX, and 4) Helicos' Heliscope Sequencer. See [2] for an extensive review. This list of platforms should be considered as merely an initial offering, because another half-dozen systems will be announced through 2010. These commercial sequencing systems are now regarded as "next generation" platforms and use variations of a venerable approach — SBS, or sequencing by synthesis, for the acquisition of sequence reads using cycles of polymerase action incorporating labeled nucleotides, which are tracked by microscopy. A quick summary of next generation sequencing systems follows (company websites offering informative media presentations).

2.1.1 *Illumina's Genome Analyzer* (*http://www.illumina.com/*)

Genomic templates, after attachment of complementary adaptors, are captured and tethered to a glass slide for amplification by bridge PCR[3]. Reads lengths are about ∼75 bp and paired-end sequencing is possible using 200 bp∼5 kbp inserts.

2.1.2 *Life Technologies' SOLiD System* (*https://products.appliedbiosystems.com*)

Emulsion PCR amplicons using genomic DNA molecules are captured on beads and enriched for extended products. Beads are then deposited onto a glass slide for sequencing cycles. In place of polymerase-mediated synthesis, the SOLiD system is distinguished by its cycles of ligation that sequentially incorporate strings of labeled di-base (2 nt) probes. Each cycle of ligation is followed by cleavage with a nicking restriction enzyme and then the entire extension product is removed. A new primer, offset by one base-pair, is annealed and a new round of ligation mediated extension proceeds. These operations are repeated 5 times. SOLiD achieves read length of 50 bp and accommodates inserts for paired end sequencing spanning 600 bp∼10 kbp.

2.1.3 *Roche's 454 GS FLX* (*http://www.454.com/*)

Like the Illumina platform, the Roche system uses emulsion PCR and amplicon attachment to beads to set up templates for sequencing through cycles of polymerase-mediated addition of unlabeled nucleotides. Templates and immobilized enzymatic reagents are deposited into individual microwells for cycles of sequencing. Unlike all other platforms, the Roche system does not use a fluorescence detection scheme, but instead leverages chemiluminescent signals transduced from pyrophosphate[4-5] liberated during each cycle of polymerase-mediated primer extension. Luciferase action is triggered by ATP through transduction of available pyrophosphate into ATP via sulfurylase. Luciferase plus ATP converts luciferin into oxyluciferin accompanied by photon emission, which is optically detected. This transduction step and subsequent photon emission are quantitative meaning that the count of nucleotides added per cycle is precisely measured. However, some issues do arise when long homopolymer tracts (long strings of the same nucleotide) are sequenced[6]. Because chemiluminescence requires no external excitation, the 454 system presents low background (excitation) for any optical detection scheme. Lastly, the long 400 bp∼600 bp read lengths and the capability for paired end sequencing, up to 20 kbp, enable this platform to deliver data supporting de novo assembly of bacterial genomes.

2.1.4 *Heliscope Sequencer* (*http://www.helicosbio.com/*)

The Heliscope handles genomic DNA samples in much the same way as the Illumina Genome Analyzer, but dispenses with any amplification steps once individual template strands are captured onto a glass surface for sequencing. Read lengths are ∼25 bp and the Heliscope will accept up to 5 kbp inserts for paired end sequencing.

2.2 Next-Next Generation (Gen-3) Sequencing

Even more advanced sequencing platforms, called Gen-3, or next-next generation, will likely appear during 2010. Gen-3 systems will accrue even greater throughput (lower costs) by real-time, direct detection of polymerase action during synthesis, using labeled nucleotides and a single molecule template. Essentially, this is sequencing by synthesis or SBS, but without serial biochemical steps for acquiring reads. Since this type of SBS is "free running," nucleotides have 5′ fluors that are cleaved off when polymerase incorporates them during elongation. Detection of these events requires measurement of transient emissions for each added nucleotide, since the fluor is not detected long after cleavage from the nucleotide. The principal advantages here are the obviation of slow, expensive sequencing cycles manifesting next generation platforms and the need for template amplification steps. In fact, it would be quite difficult to coordinate polymerase action on multiple, amplified templates, so that these platforms intrinsically require single molecule samples. There are two major platforms under development for real-time, cycle-free sequencing:

2.2.1 *Pacific BioSciences* (*PacBio*)

The PacBio platform uses a very unique detection

scheme based on the properties of zero mode waveguides (ZMWs) to control the radiative "reach" of the excitation entering it. The ZMWs are an array of aluminum nanowells, just 100 nm wide, fabricated over a glass slide the slide providing a tiny window for each ZMW. The action of a single polymerase bound to a template, sitting at the bottom of each nanowell, is actively imaged during incorporation of labeled nucleotides, which is accompanied by concerted release of their fluorochrome labels. Because ZMWs limit excitation close to their glass windows, ∼10 nm detection is limited to only those fluorochrome labeled nucleotides engaging polymerase during strand elongation. This feat ensures normal polymerase action by allowing optimized nucleotide concentrations, while eliminating the fluorescence background caused by spectator nucleotides bearing fluorochromes[7]. At this stage of development it is too early to definitively state reads lengths; however, PacBio uses a highly processive DNA polymerase, which they have shown capable of strand displacement synthesis stretching multiple kbp.

2.2.2 Life Technologies

The details regarding Life Technologies' Gen-3 system are still somewhat obscure, but sufficiently compelling to discuss here. This Gen-3 is uses similar chemistries and attenuation of detection volume as the PacBio system, but ingeniously places the detection burden onto an individual DNA polymerase. Although reagent nucleotides are fluorochrome labeled, they are only visible by FRET (fluorescence resonance energy transfer) excitation, when incorporated within a growing strand, mediated by a labeled polymerase. Not detected are labeled nucleotides that are freely diffusing, and like the PacBio system, this advantage allows optimized concentrations of nucleotides during sequencing. Such specificity is engendered by FRET excitation, which requires a fluorochrome donor — acceptor pair, and here the labeled polymerase provides the donor, while the acceptor is attached to the incoming nucleotide.

2.3 Next-Next-Next Generation: Nanopore Sequencing

Characterizing the previous two sequencing systems as Gen-3 means that nanopore sequencing should be called "Gen-4." Conceptually, it may be the simplest way to sequence DNA, but possibly the most difficult to implement. Briefly, a single strand of DNA is electro-kinetically threaded through a tiny pore of comparable width (∼1 nm ∼ 2 nm), and the measured electrical signature of each base, acquired as it moves through the pore, reports how much current it blocks. Purines (`A, G`), being more bulky than pyrimidines (`C, T`), block more current, as electrically measured across a pore[8]. For this purpose, nature has provided an almost ideal pore protein — hemolysin — when suitably engineered[9] creates new routes for very inexpensive sequencing[10]. Nanofabrication approaches are also creating synthetic pores for DNA sequencing offering new opportunities to place novel detection schemes within the same device[11]. The main advantages of nanopore sequencing are that DNA samples can be immediately analyzed without any labeling, or involved preparation. The non-optical electronic detection scheme lays the basis for the development of supremely miniaturized instrumentation, which will dramatically reduce costs. Recent developments in nanopore sequencing are reinvigorating an early single molecule sequencing approach championed by Keller and colleagues over 20 years ago[12]. Here, both old and new approaches use exonuclease for serially clipping bases (Keller used fluor labels) from a DNA strand for separate downstream detection, but the modern approach identifies liberated bases by nanopore detection[13], obviating enzymatically troublesome labeled nucleotides and optical detection.

2.4 The Central Problem: Cost

Even when we reach the goal of a $1000 genome, this price is still far too costly for supporting the commoditization of sequence information. In the final analysis, there is an almost an unquenchable need for more sequence information within dramatically expanded databases and from citizens, who will expect to perform numerous assays as a normal part of their daily lives.

3 Computational New Generation Sequence Assembly

In this section we will focus on computational aspects of the assembly of genomic DNA sequence from reads. There are a multiplicity of issues resulting from determining the reads from the machines discussed in the previous section, issues which are specific to each technology. With Sanger sequencing the quality scores, determined by estimating the probability that the read was actually the reported letter, became very popular and useful although these q-values are not always from an accurate estimate of the probability. Here we neglect such matters in order to provide a more general discussion.

While our discussion is organized around what is called *de novo* sequence assembly, a host of other challenges result from using a sequenced genome to "map" the reads onto, cataloging the RNA gene transcripts,

studying epigenomics data such as that which results from bisulfide sequencing. See the paper by Zhang and Smith in this issue for epigenomics issues[14], Morozova *et al.* for functional genomics applications[15] and Chen *et al.* for an in depth discussion of mapping reads onto genomes[16].

3.1 Classical Sequence Assembly

The chain termination method of DNA sequencing was developed by Fred Sanger in the 1970s and became increasingly popular in the last 25 years of the 20th century. Dideoxynucleotide triphosphates (ddNTPs) are used as DNA chain terminators and initially there was a gel electrophoresis lane for each of four reactions. Gel resolution of one base pair allowed the calling of the sequence of bases for a read. Later dye terminator sequencing allowed all four reactions to be run in a single channel. Technological advances in Sanger sequencing created the workhorse of the Human Genome Project and allowed early completion of that project.

In the same way that increasingly sophisticated improvements of Sanger sequencing allowed production of the data for the Human Genome Project, there was a parallel development of computational programs to assemble DNA sequence from random reads. The reads were randomly located on the target DNA as well as randomly either strand of the double helix. (We ignore issues of diploid targets for now.) Roger Staden was called on by Fred Sanger to write code that assembled DNA sequence. Initially he employed a greedy approach. As the DNA was produced, he assembled overlaps into longer pieces (super-reads) by taking the longest overlaps first[17]. At the same time Gingeras *et al.* employed a similar approach[18]. Later that strategy was used with a collection of reads. A simplification of the greedy method was presented to the theoretical computer science community: given a collection of reads, what is the length of the shortest superstring which contains each read (exactly)? See [19]. The known results require some sophisticated computer science. See [20] for a proof that the sequence provided by the greedy algorithm is at most 4 times optimal. It is known that the answer is at most 2.5 times optimal and there is a conjecture that 2 times optimal is the best possible bound.

However greedy is a weak algorithm and the wide occurrence of repeats in many genomes including the human genome makes the answers provided by greedy unsatisfactory. The first systematic study of DNA assembly was in [21]. Many valuable papers by Gene Myers and collaborators brought sophisticated improvements to assembly, and Myers led the Celera assembly team to that company's resulting Human Genome Sequence[22-25]. Just as celebrated was the Santa Cruz assembly of the data from the public project, see [26]. [27] is a comparison of the results of these justly celebrated projects.

The sequence assembly algorithms we have been describing can be outlined in three steps: Overlap, Layout and Consensus. Overlap requires considering for each pair of reads whether they overlap in the presented orientation or in another orientation. This means two comparisons, conceptually. The most straightforward comparison is to use overlap dynamic programming, which takes time proportional to the product of the read lengths. The Celera project had approximately 26.5 million reads of length approximately 550, so this direct approach was not feasible, but it indicated the magnitude of the problem. The next Layout step is even more challenging, requiring the determination of sets of reads with mutually consistent overlaps, and determining approximately how they might be arranged on the genome. Here the difficult problem of determining which orientation each overlapping pair needs to be solved. With 10^6 reads there are $2^{10^6} = 10^{301000}$ possible orientations. Finally for Consensus, given an approximate alignment of a set of consistent overlapping reads, the multiple alignment problem must be solved to produce a consensus sequence. This too is a very challenging computational task.

3.2 Euler Sequence Assembly

With the New Generation Sequencing methods described in Section 2, we see many critical differences from the classical Sanger data. First of all the reads are often much shorter, and also there are many more reads. An approach using Overlap-Layout-Consensus is doomed to failure and a new method must be found. Fortunately there was a paper in 1995 by Idury and Waterman which, although virtually unnoticed at the time, can be exploited for these uses[28]. Essentially it trades huge computing costs for a substantial (and sometimes huge) storage cost. It was a quite new approach for DNA assembly.

The idea for the new 1995 method is based on Eulerian graphs. There is a natural way to take a sequence and encode it into a graph. One chooses a value of k and extracts all of the $n - k + 1$ k-words from the sequence. Then the vertices of the sequence graph are the various $k - 1$ words from the k-words and the edges are the corresponding k-words. Note that we have gone from a sequence to a graph. If the data were the k-words from an unknown sequence, then an identical graph could be produced and the job would be to infer that unknown sequence from the data. Note that this job is simple when there is a unique answer from our

graph. Of course the answer need not be unique but one sequence consistent with the data can be produced in linear time, or it can be decided in linear time that no such sequence exists, if indeed there is none. These ideas and procedures are part of the theory of Eulerian graphs which for sequences goes back over 100 years. Contrast these easy operations with the difficulties for another encoding where the k-word data are identified with graph vertices. In that case we have a Hamiltonian path problem which is NP-hard.

The extension of sequence graphs to sequence assembly begins with a specification of k and every fragment is decomposed into its overlapping k-words. The total set of these k-words for all fragments is merged, where attached to a k-word is a list of all fragments and fragment positions where the word occurs. Then a sequence graph is constructed from this set, carrying along all the annotation, which is of course associated with edges. The choice of k is based on the shortest length where most k-words in the genome will be unique. The target coverage needs to be deep enough so that the correct k words will appear in the fragments fairly often; in this way sequencing errors will be out voted. Errors, of a few letters either substitutions of indels, will then cause bubbles to diverge in the graph from heavier weighted edges by which the erroneous k-words appear on the graph. The 1995 paper[28] recommended trimming such edges. In the programs that have appeared in the last few years, so-called error correction is instead recommended. Pevzner has introduced what he calls the Eulerian superpath problem where the goal is to extract Eulerian paths from the sequence graph which breaks as few original fragment paths in the graph[29-30]. That is just what all the heuristic methods attempt to do. Pevzner also has extended the technique to handle paired-end sequencing. Myers has shown this problem is NP-hard, but the heuristic algorithms still run linear time and work quite well[31]. In some sense the final sequence graphs after processing data from a genome project are simply elegant data structures containing what is known about the target genome. Here is a schematic of the method:

Algorithm Assembly. Set N, k: input: $f_1, f_2, \ldots, f_N$.

1) Take all fragments and their reverse complements.

2) Construct the sequence graph on $(k-1)$-tuples for the k-tuples from step 1).

3) Perform an Eulerian tour(s) and infer the sequence(s).

4) Align fragments to sequence(s) produced by step 3).

See [32-34].

As mentioned earlier, this discussion leaves out any mention of the error properties of the particular sequencing technologies. Not all reads are equally reliable and this should be taken into account. Errors can arise from runs of the same base in the material to be sequenced and other issues such as bias in the reads which are recorded. We do not go into details of the various platforms here but they are essential in constructing effective assembly and read mapping packages.

Storage is a serious drawback to Eulerian path methods, and it is difficult to imagine that several more orders of magnitude improvement to sequencing technologies can proceed without a breakthrough in the basic computational methods (including storage).

4 The Future

Overview: five years ago, one company dominated the market for the platforms and reagents used for sequencing. Today, this is no longer true and the current situation is characterized by vibrant activity in both the academic and commercial settings that are creating a torrent of new ways for sequencing DNA. The problems, challenges, and promises to be encountered, as sequencing costs drop, will certainly have evolutionary components, but will likely also present radical shifts in how we approach scientific problems and view ourselves within the continuity of life, or society.

4.1 Absentee Genomics

Given the just described characteristics of Gen-2 and Gen-3 sequencing systems, there is an enormous gap developing between the quantity and quality of sequence data. Aside from the great variety of new error processes associated with new sequencing platforms, short, but plentiful sequence reads are presenting new opportunities for reliable sequencing. However, there looms the issue of completeness if a sequenced genome harbors any gaps, or misassemblies, it is incomplete, but can accommodate most gene hunting activities. Modest read length and the absence of physical and genetic maps aggravates this problem. It is somewhat sobering to realize that there are only handful of "finished" genomes despite current advances in sequencing technology and assembly. Remarkably, the "finished" human genome reference map does not include much sequence information regarding centromeric regions — critical chromosomal regions that anchor cell division. Unfortunately, an imperfectly and incompletely sequenced genome is a compromised reference for comparative studies. Any study aimed at discovering genomic polymorphisms and mutations is weakened by questions of reference map accuracy, whose validations often are commensurate in scope with the efforts required for cataloging the genomic alterations.

In response to the problem outlined above, we project that over the next 2 years, reference genomes[35-36] will be constructed using new

algorithms[36] combining long-range physical maps with voluminous Gen-2/3 datasets[37]. In this regard, the Optical Mapping System constructs genome-wide ordered restriction maps from individual (∼500 kbp) genomic DNA molecules[38-39]. Like the Gen-3 sequencing platforms, Optical Mapping is a single molecule system that uses microscopy for imaging dense arrays of individual DNA molecules[40]. Although sequence information is not acquired, a dense restriction map is constructed for each 500 kbp molecule, enabling maps to bridge across complex genomic regions. Optical Mapping data sets comprise hundreds of thousands of maps that are assembled into genome-wide physical maps using *de novo*[36,41] and anchoring approaches utilizing provisional sequence scaffolds[35,42]. A successor to Optical Mapping — "Nanocoding"[43] — promises dramatic boosts to throughput by leveraging novel DNA nanoconfinement effects (long DNA molecules are electrically squeezed in tight tubes). Other advancements are aimed at new systems that directly combine sequencing and long-range mapping within a single approach. Here, a string of short sequence reads is acquired along a large molecule that is mapped[44-46]; called "Optical Sequencing." In this way, many of the issues surrounding modest read lengths are addressed and such developments support that assembly of truly finished genomes.

4.2 Every Genome Matters

Although new sequencing approaches will drive the compilation of a Noah's ark catalog of sequenced genomes, somatic genomes present an almost infinite collection of genotypes for discovery by even more advanced sequencing approaches. Analysis of tissue sections will yield subtle genomic alterations when ultra-deep sequencing becomes economically feasible. These rare genomic alterations sometime arise from inherent genome instability[47], and may evolve further into pervasive cancerous states. Consider that gathering 1 000 000 X coverage is about 100 000 times deeper than the publicly available human reference genome. Another way of thinking about this is 1 000 000 X genome coverage is equivalent to sequencing 100 000 individuals. The computational challenges arising from this are far beyond our current capacities; even today's Eulerian methods are bound from the excessive memory requirements. As always the solutions must be closely linked to the technologies.

Extensive resequencing will reveal outlier populations of cellular genomes (the genome of an individual cell) when sequencing is done on a per cell basis. If cellular DNAs are pooled then sequenced, a more practical approach, a spectrum of rare genotypes spanning the entire genome will be apparent for a given tissue sample. It is likely that most of these altered genotypes are mutations, but more interestingly, we will have the chance to discover new developmental programs that may guide differentiation and ensure pluripotency.

Acknowledgement We are grateful to Frederick Sanger for his discoveries which laid the basis for knowing ourselves at the molecular level.

References

[1] Harris T D, Buzby P R, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch J W, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake S R, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z. Single-molecule DNA sequencing of a viral genome. *Science*, 2008, 320(5872): 106-109.

[2] Fuller C W, Middendorf L R, Benner S A, Church G M, Harris T, Huang X, Jovanovich S B, Nelson J R, Schloss J A, Schwartz D C, Vezenov D V. The challenges of sequencing by synthesis. *Nat. Biotechnol.*, 2009, 27(11): 1013-1023.

[3] Pemov A, Modi H, Chandler D P, Bavykin S. DNA analysis with multiplex microarray-enhanced PCR. *Nucleic Acids Res.*, 2005, 33(2): e11.

[4] Ronaghi M, Uhlen M, Nyren P. A sequencing method based on real-time pyrophosphate. *Science*, 1998, 281(5375): 363-365.

[5] Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.*, 1996, 242(1): 84-89.

[6] Bentley D R. Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.*, 2006, 16(6): 545-552.

[7] Eid J, Fehr A, Gray J *et al.* Real-time DNA sequencing from single polymerase molecules. *Science*, 2009, 323(5910): 133-138.

[8] Kasianowicz J J, Brandin E, Branton D, Deamer D W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. USA*, 1996, 93(24): 13770-13773.

[9] Astier Y, Braha O, Bayley H. Toward single molecule DNA sequencing: Direct identification of ribonucleoside and deoxyribonucleoside 5′-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. *J. Am. Chem. Soc.*, 2006, 128(5): 1705-1710.

[10] Branton D, Deamer D W, Marziali A *et al.* The potential and challenges of nanopore sequencing. *Nat. Biotechnol.*, 2008, 26(10): 1146-1153.

[11] Sigalov G, Comer J, Timp G, Aksimentiev A. Detection of DNA sequences using an alternating electric field in a nanopore capacitor. *Nano. Lett.*, 2008, 8(1): 56-63.

[12] Jett J H, Keller R A, Martin J C, Marrone B L, Moyzis R K, Ratliff R L, Seitzinger N K, Shera E B, Stewart C C. High-speed DNA sequencing: An approach based upon fluorescence detection of single molecules. *J. Biomol. Struct. Dyn.*, 1989, 7(2): 301-309.

[13] Clarke J, Wu H C, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.*, 2009, 4(4): 265-270.

[14] Zhang M Q, Smith A D. Challenges in understanding genome-wide DNA methylation. *J. Comput. Sci. & Technol.*, 2010, 25(1): 26-34.

[15] Morozova O, Marra M. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 2008, 92(5): 255-264.

[16] Chen Y, Souaiaia T, Chen T. PerM: Efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. *Bioinformatics*, 2009, 25 (19): 2514-2521.
[17] Staden R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.*, June 11, 1979, 6(7): 2601-2610.
[18] Gingeras T R, Milazzo J P, Sciaky D, Roberts R J. Computer programs for the assembly of DNA sequences. *Nucleic Acids Res.*, September 25, 1979, 7(9): 529-545.
[19] Gallant J, Maier D, Storer J. On finding minimal length superstrings. *J. Computer System Sci.*, 1980, 20(1): 50-58.
[20] Waterman M S. Introduction to Computational Biology. Chapman & Hall, 1995.
[21] Kececioglu J D, Myers E W. Combinatiorial algorithms for DNA sequence assembly. *Algorithmica*, 1995, 13(1/2): 7-51.
[22] Kececioglu J D. Exact and approximation algorithms for DNA sequence reconstruction [Ph.D. Dissertation]. University of Arizona, Tucson, USA, 1992.
[23] Myers E W. Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology*, 1995, 2(2): 275-290.
[24] Myers E S. The fragment assembly string graph. *Bioinformatics*, 2005, 21(Suppl. 2): ii79-ii85.
[25] Venter J C, Adams M D, Myers E W *et al.* The sequence of the human genome. *Science*, 2001, 291: 1304-1351.
[26] Lander E S, Linton L M, Birren B *et al.* Initial sequencing and analysis of the human genome. *Nature*, 2001, 409(6822): 860-921.
[27] Istrail S, Sutton G, Florea L *et al.* Whole genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl. Acad. Sci. USA*, 2003, 101(7): 1916-1921.
[28] Idury R, Waterman M S. A new algorithm for DNA sequence. *J. Comput. Biol.*, 1995, 2(2): 291-306.
[29] Chaisson M J, Pevzner P A. Short read fragment assembly of bacterial genomes. *Genome Res.*, 2008, 18(2): 324-330.
[30] Chaisson M J, Tang H, Pevzner P A. Fragment assembly with short reads. *Bioinformatics*, 2004, 20(13): 2067-2074.
[31] Myers E W. The fragment assembly string graph. *Bioinformatics*, 2005, 21(Suppl. 2): ii79-ii85, doi:10.1093/ bioinformatics/bti7114.
[32] Pevzner P A, Tang H. Fragment assembly with double-barreled data. *Bioinformatics*, 2001, 17(Suppl. 1): S225-S233.
[33] Pevzner P A, Tang H, Waterman M S. A Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. USA*, 2001, 98(17): 9748-9753.
[34] Zerbino D R, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, 2008, 18(5): 821-829.
[35] Church D M, Goodstadt L, Hillier L W *et al.* Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.*, 2009, 7(5): e1000112.
[36] Valouev A, Schwartz D C, Zhou S, Waterman M S. An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proc. Natl. Acad. Sci. USA*, 2006, 103(43): 15770-15775.
[37] Nagarajan N, Read T D, Pop M. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics*, 2008, 24(10): 1229-1235.
[38] Schwartz D C, Li X, Hernandez L, Ramnarain S P, Huff E J, Wang Y K. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science*, 1993, 262(5130): 110-114.
[39] Zhou S, Herschleb J, Schwartz D C. A Single Molecule System for Whole Genome Analysis. New Methods for DNA Sequencing, Mitchelson K R (ed)., Amsterdam: Elsevier, 2007.
[40] Dimalanta E T, Lim A, Runnheim R *et al.* A microfluidic system for large DNA molecule arrays. *Anal. Chem.*, 2004, 76(18): 5293-5301.
[41] Valouev A, Zhang Y, Schwartz D C, Waterman M S. Refinement of optical map assemblies (original paper). *Bioinformatics*, 2006, 22(10): 1217-1224.
[42] Valouev A, Li L, Liu Y C, Schwartz D C, Yang Y, Zhang Y, Waterman M S. Alignment of optical maps. *J. Comput. Biol.*, 2006, 13(2): 442-462.
[43] Jo K, Dhingra D M, Odijk T *et al.* A single-molecule barcoding system using nanoslits for DNA analysis. *Proc. Natl. Acad. Sci. USA*, 2007, 104(8): 2673-2678.
[44] Ramanathan A, Pape L, Schwartz D C. High-density polymerase-mediated incorporation of fluorochrome-labeled nucleotides. *Analytical Biochemistry*, 2005, 337(1): 1-11.
[45] Ramanathan A, Huff E J, Lamers C C, Potamousis K D. Forrest D K, Schwartz D C. An integrative approach for the optical sequencing of single DNA molecules. *Analytical Biochemistry*, 2004, 330(2): 227-241.
[46] Zhou S, Pape L, Schwartz D C. Optical Sequencing: Acquisition from Mapped Single Molecule Templates. Next Generation Genome Sequencing: Towards Personalized Medicine, Janitz M (ed.), 2008, Weinheim: Wiley-VCH Verlag & Co., pp.133-149.
[47] Aguilera A, Gomez-Gonzalez B. Genome instability: A mechanistic view of its causes and consequences. *Nat. Rev. Genet.*, 2008, 9(3): 204-217.

David C. Schwartz has been a professor of chemistry, genetics, and biotechnology at the University of Wisconsin-Madison since 1999. Previous faculty appointments were at New York University and The Carnegie Institution of Washington-Baltimore. He works at the interface between nanotechnology and genomic science through his invention of single molecule platforms for genome analysis. As a graduate student, he invented Pulsed Field Gel Electrophoresis and continues to train students in the science of how to conceive and develop new tools for biological investigation as the director of a training program.

Michael S. Waterman is university professor at the University of Southern California where he has been a faculty member since 1982. Prior to that he held positions at Los Alamos National Laboratory and Idaho State University. He also currently holds a chair professor team at Tsinghua University. Professor Waterman works in the area of computational biology, particularly on the analysis DNA, RNA and protein sequence data. He is the co-developer of the Smith-Waterman algorithm for sequence comparison and of the Lander-Waterman formula for physical mapping. He is a member of the U.S. National Academy of Sciences and of the French Académie des Sciences.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.