Comparative Report of the Oxford Nanopore DNA Sequencing Technology and Current Sequencers

Molly LoSchiavo, Corianne Galloway, Kevin Boehme

Comparative Report of the Oxford Nanopore DNA Sequencing Technology and Current Sequencers

Submitted to

Sister Hadden

for

English 316

Brigham Young University

Provo, Utah

December 8, 2014

by

Molly LoSchiavo, Corianne Galloway, Kevin Boehme

Letter of Transmittal

December 8, 2014

English 316
Sister Hadden
3004 JKB
Brigham Young University
Provo, UT 84602
(801) 422-4704

Dear Sister Hadden,

As a group, we are prepared to submit a copy of the Technical Report assignment for English 316 to you and our peers.

The purpose of the report is to provide a comparison of current DNA sequencing technologies with the new Oxford Nanopore technology. We focused on three of the "next-generation" sequencing technologies to use for comparison: Illumina, Pacific Biosciences, and Ion Torrent. To compare these technologies, we concentrated on three criteria to determine the best recommendation for the most efficient technology based on cost per megabase, time per run, and output per run.

Please take time to look over our report. We are anxious for feedback, so we can make the necessary improvements. If you have questions or comments please contact us at coric8@gmail.com, mollyloschiavo@gmail.com, or kevinlboehme@gmail.com.

Sincerely,

Corianne Galloway, Molly LoSchiavo, Kevin Boehme

Table of Contents

List of Figures

List of Tables

Abstract

This report is a comparative analysis of desktop DNA sequencers: Oxford Nanopore (MinION), Illumina (MiSeq), Ion Torrent (PGM), and Pacific Bioscience (RS II). Oxford Nanopore is a new technology that is garnering a lot of hype in the media. Our goal is to compare the results of Oxford Nanopore's desktop DNA sequencer (MinION) with the three most popular DNA sequencers in the market. We aim to determine 1) if the MinION measures up to the hype it's received and 2) which sequencing technology is the most efficient in terms of money, time, and output.

We collected data from studies that compared the older sequencing technologies as well as recent data from experiments which used the Oxford Nanopore sequencer. With this data, we made comparisons between the DNA sequencers based on cost per megabase, time per run, and output per run. Our results indicate that Oxford Nanopore costs about 3 times more than the next most expensive sequencer (PacBio) and 10 times more than the cheapest sequencer (Illumina). Oxford Nanopore also has the longest time per run and the second-lowest output per run of all the technologies. In our analysis, we combined the time per run and output per run measurements to formulate a new standard, throughput, or output per hour. We found that the MinION also has the lowest throughput of all the technologies.

Our conclusion is that the Oxford Nanopore technology performs far worse in terms of time, money, output, and throughput than the established sequencing technologies. We further determine that Illumina's MiSeq is the most efficient desktop sequencer, capable of sequencing DNA cheaper and faster than any other desktop machine included in the study.

Comparative Report of the Oxford Nanopore DNA Sequencing Technology and Current Sequencers

I Introduction

The subject of our project is comprised of the up-and-coming Oxford Nanopore DNA sequencing technology and the most popular *next-generation DNA sequencers* (see Glossary). As for the purpose of our study, we aim to provide a digestible and practical comparison of current sequencing technologies with that of the new Oxford Nanopore technology to determine if Oxford Nanopore deserves all of the hype it's been receiving and to determine which technology's desktop unit is the most efficient in terms of money, time, and output.

The scope of this study involves the comparison of Oxford Nanopore's desktop unit, the MinION, to the three most popular desktop next-generation DNA sequencing machines: Illumina's MiSeq, Ion Torrent's PGM, and Pacific Biosciences' (PacBio) RS II. To identify which of these technologies is most efficient, we focused on three main criteria: cost per *megabase*, time per run, and output per run.

The development of our project consisted of three steps: research, analysis, and solution. We first collected data and information regarding the three criteria mentioned above as well as any other research previously done on this subject. This information helped us formulate our background material as well as lay a foundation for analysis. We gathered this information from prominent next-generation sequencing technology websites such as www.allseq.com, www.biomedcentral.com, and www.illumina.com. In addition, we found studies from reputable biotechnology journals such as Nature Biotechnology, Science, and Genome Research. We also found information regarding recent results from Oxford Nanopore's MinION unit in a study performed by Quick, et al., as well as recent installments on its development and implementation[1]. With this data, we created visual comparisons between the technologies based on the criteria mentioned in the scope of our project. In addition, we combined the

time per run and output per run measurements to formulate a new standard, throughput, or output per hour. We felt that this new measurement was more indicative of the productivity of each machine. After we completed our analysis, we formulated a recommendation for the overall most efficient sequencing technology.

We believe that our work will prove useful to a variety of groups. For example, the field of bioinformatics deals with DNA sequence processing, and any potential "game-changing" technology (i.e. Oxford Nanopore) will be of intense interest to everyone in this field. Health professionals and pharmaceutical companies will be interested in this study if the results indicate a potential for rapid clinical applications by significantly reducing costs or time for sequencing. In addition, many individuals in the biotechnology arena follow sequencing technologies closely, and this study may provide tractable information to them.

II Background

Introduction to Sequencing

DNA is a molecule that is commonly referred to as "the blueprint of life". That is because the DNA found in just one of your trillions of cells contains all of the information needed to make, grow, and operate your body! This information is stored in the form of DNA sequence. DNA is composed of four kinds of molecules called *nucleotides*: adenine, guanine, cytosine, and thymine. For simplicity, these nucleotides are referred to as A, G, C, and T, respectively. DNA sequence is the order in which these nucleotides occur. Within the 3 billion nucleotides that make up the human *genome* lie many of the answers to our questions about ancestry, life processes, and diseases such as cancer, Alzheimer's, and autism.

The first efforts to unearth this valuable sequence began in the 1970s. At this time, the prevailing sequencing technology (Sanger sequencing) was primitive and too expensive to sequence anything larger than a couple hundred thousand nucleotides long. However, as

new technologies emerged, and public interest increased over the following decades, the world of DNA sequencing experienced an event that changed it forever: The Human Genome Project.

The mapping of the human genome was one of the largest collaborative projects in human history. With the presentation of this invaluable data to the public came a rush of new DNA sequencing technology. Many universities, research facilities, and private companies wanted to hitch a ride on the coattails of this new, exciting opportunity to perhaps get their own chance at changing the world. Through this, many brand-name technologies emerged such as Illumina, Ion Torrent, and PacBio. These technologies came to be known as "next-generation sequencing technologies" because of their technological advancement as compared to earlier sequencing technologies like Sanger sequencing[2]. Although much progress has been made in the field of genomics, researchers are still eager for cheaper, faster, and better sequencing technologies. Specific goals that the field has set include reducing the cost of sequencing an individual's genome to less than $1,000 and advancing medical and genomic research to someday realign the medical diagnostic framework which will allow patients to receive treatments specific to their genomic makeup[3]. While current technologies have gotten us close to achieving these goals, there still remains room for improvements.

Oxford Nanopore is an up-and-coming sequencing technology that claims to be leading the world to a $1,000 genome sequence with longer read lengths and shorter run time due to the fundamental design of the technology[4]. The advancement in technology that the new sequencing method uses (called strand sequencing) is what put Oxford Nanopore in the "third-generation sequencing technologies" category. We will now introduce each technology individually.

Illumina

Nucleotides perform base-pairing between strands of DNA. In *base-pairing*, A and T only pair with each other, and C and G only pair with each other. Using this, Illumina employs "sequencing by synthesis" to sequence DNA. Sequencing by synthesis works by first isolating a single strand of DNA. Next, all four nucleotides are introduced to the strand. Because of base-pairing, only the nucleotide that matches the sequence will attach to the strand. On this newly attached nucleotide is a *fluorophore*. Fluorophores fluoresce a specific color when activated by a laser. Each nucleotide is given a separate color. In Figure 1, T is red, G is green, C is yellow, and A is blue. After the nucleotide base-pairs to the DNA strand, the Illumina machine flashes a laser, and a camera records what color fluoresces. This tells the machine which nucleotide was attached to the strand. This process is then repeated until the entire DNA strand has been sequenced and recorded.

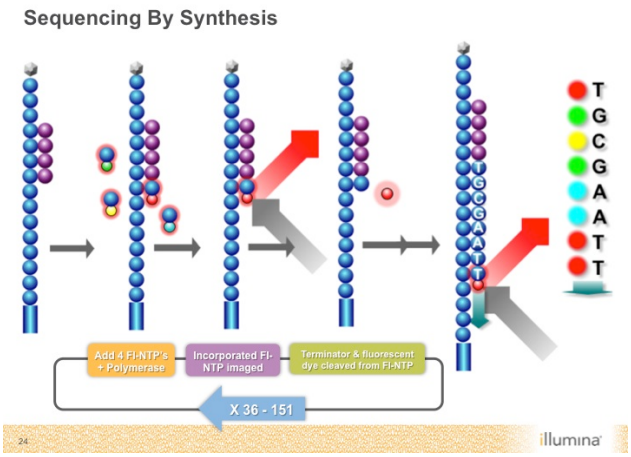

Figure 1: Illumina[5]

An advantage to this technology is its reliability. Because bases are incorporated into the sequence one at a time, the chance of the machine making an error when it reads and records the sequence is reduced. In addition, this reaction is particularly easy to mass-produce, which provides opportunities for greater efficiency. Another way to think about it is to consider Illumina as the "Costco" of DNA sequencers. If you sequence in bulk, you can reduce your costs.

Ion Torrent

Ion Torrent's process for sequencing DNA, called "semiconductor sequencing", is similar to Illumina, but with some key differences. Both begin by isolating a single DNA strand. Rather than introducing all four nucleotides at the same time, Ion Torrent only introduces one base at a time. When a nucleotide base-pairs to the DNA strand, it releases a hydrogen ion. In the Ion Torrent machine, there is a hypersensitive pH meter. It can detect the single addition of a hydrogen ion in a solution. If, when the nucleotide is introduced to the DNA strand, the machine does not detect a change in pH, that nucleotide is removed, and the next kind is brought in. The process is repeated until the machine detects a change in pH. When that happens, the computer records which base was incorporated and repeats the cycle until the sequencing is complete.

Because Ion Torrent doesn't need highly specialized reagents such as nucleotides with color-coded fluorophores, it is generally cheaper than most technologies. However, it does not benefit from an "economies of scale" as does Illumina. Additionally, Ion Torrent's machines are designed so that, as technology advances, the only upgrades ever needed are new semiconductor sequencing chips, rather than new machines entirely. This attribute makes Ion Torrent machines more appealing as long-run investments.

PacBio

This technology is referred to as "single-molecule, real-time DNA sequencing", or "SMRT sequencing." Its concept is also similar to that of Illumina and Ion Torrent, with its own important distinctions. SMRT sequencing begins by anchoring a single *DNA polymerase* molecule to the bottom of a tiny well. Next, a single DNA strand is introduced to the DNA polymerase. In addition, all four kinds of nucleotides are added to the solution. These nucleotides have specifically colored fluorophores attached, as in Illumina. When DNA polymerase grabs the nucleotide that matches the DNA's sequence, it cleaves the fluorophore from

the nucleotide, activating it in the process and emitting light (See Figure 2). A specialized camera recognizes the color of the light and records the respective base.

The major advantage to this technology is DNA polymerase works extremely fast (around 1,000 base pairs per second)[7]. This attribute allows PacBio to sequence DNA very quickly. However, it is not easily mass-produced because there can only be one DNA strand sequenced per well.

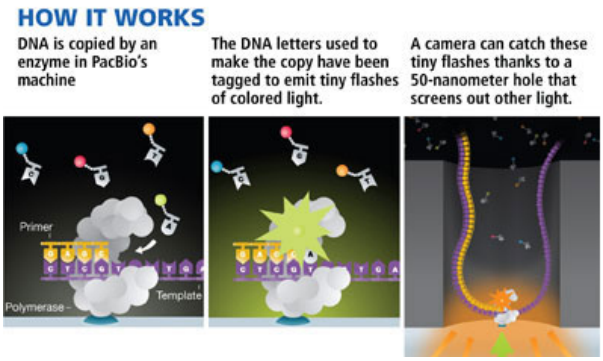

Figure 2: PacBio[6]

Oxford Nanopore

Oxford Nanopore utilizes "strand sequencing." The key to the workings of strand sequencing is a synthetically produced protein designed specifically for this task. This protein has a tube in the middle that is nanometers in diameter, making it small enough so only one strand of DNA can pass through at a time. This protein is also designed to fit into a tiny pore on a synthetic membrane. An electric current is then run through the membrane. Whenever anything passes through the small tube in the protein, it disrupts the current. Each molecule that passes through the protein creates its own distinct disruption.

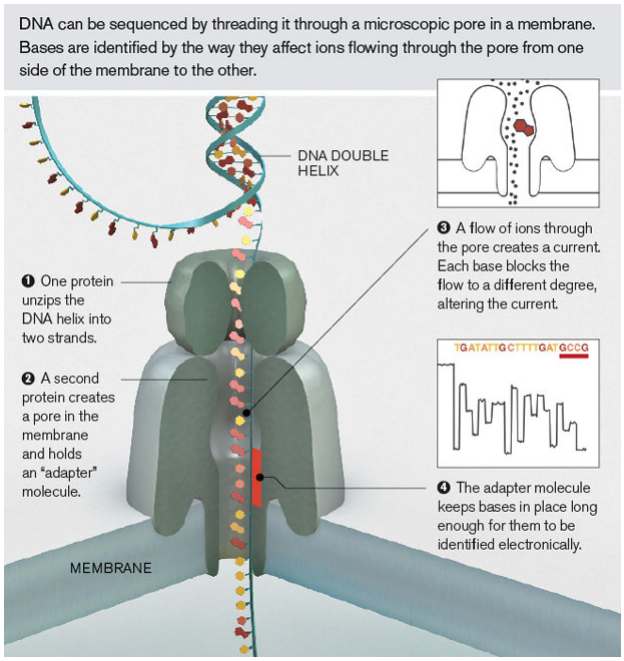

Figure 3: Oxford Nanopore[8]

When a DNA strand, for example, is run through the protein, the disruptions made can be categorized into each of the four bases. In other words, the disruptions can be translated into the DNA's sequence (See Figure 3)[9]. Ideally, because strand sequencing does not require any DNA synthesis and very few reagents, as do the other three technologies, it should be faster, more accurate, and cheaper[10]. Therefore, Oxford Nanopore is expected to be the most efficient of the sequencing technologies. To determine if this is true, we selected three parameters (cost per megabase, time per run, and output per run) to measure and compare each technology.

III Methodology

In order to accurately compare sequencers, we elected to compare each technology based on cost per megabase, time per run, and output per run. These metrics are a good measure of the efficacy of the machine and provide a solid foundation to allow us to make comparisons. We will now explain what each measurement is, why it was chosen, and how we obtained our data. In addition, we provide a "caveats" section describing some limitations of this comparison.

Cost per Megabase

Considering the fact that the Human Genome Project cost around $2.7 billion dollars, the cost of sequencing is a major limiting factor when considering projects[3]. Utilizing technologies that sequence at lower costs enable researchers to take on bigger projects and obtain more data[2]. Therefore, the cost it takes to obtain one megabase (1 million bases) of data is heavily considered by researchers when choosing what kind of sequencing technology to use. In fact, it is used by the National Human Genome Research Institute in their widely known and important benchmark graphs illustrating the decreasing costs associated with DNA sequencing. This measurement captures all of the direct "production" costs of producing the

raw sequencing data. These production costs include[11]:

1. Labor, administration, management, utilities, reagents, and consumables
2. Sequencing instruments and other large equipment (amortized over three years)
3. Informatics activities directly related to sequence production (e.g., laboratory information management systems and initial data processing)
4. Shotgun library construction (required for preparing DNA to be sequenced)
5. Submission of data to a public database
6. Indirect Costs as they relate to the above items

These are all important costs associated with producing DNA sequence data and should be captured in a cost per megabase measurement. We collected cost per megabase data for the established sequencers from van Dijk, Loman, and Mikheyev[12,13,14] and data for the Oxford Nanopore from Quick[1].

The price of purchasing the actual sequencing machine is a critically important factor for those institutions looking to be able to sequence in-house. However, for others looking to outsource their DNA sequencing needs, this cost is not as important as the cost per megabase. For the scope of this report, we will refrain from incorporating the cost of the actual machines into our comparison.

Time per Run

Time per run is an indicator of how quickly a machine can turn DNA template into sequence data. Researchers hold this quality of high importance because, as of now, genomic sequencing takes a considerable amount of time to complete (the Human Genome Project, for example, took about 13 years)[15]. Therefore, the shorter the run-time for sequencing, the quicker each project can be completed. This is a crucial consideration not only for re-

searchers, but also for investors who are anxious for good returns in a timely manner. This information was found by combining data from Quick and van Dijk[1,12].

Output per Run

Output per run is the number of base pairs sequenced in a single run on a machine. It is important to consider output per run alongside time per run because a machine with a short run-time is inefficient if it also produces a low output. In addition, a higher output per run increases efficiency in terms of cost. Each run of a machine costs a considerable amount of money, so the more output generated by each run, the more cost-efficient the machine. We used data from van Dijk, Loman, and Mikheyev for our analysis[12,13,14].

Caveats

Validity of Machine Comparison

Each technology we have discussed comes in many forms. For example, Illumina sells four different kinds of machines that use the same technology, but are built for different purposes[16]. We found that all but one of our chosen technologies offers a "desktop" machine, made to be efficient both in terms of money and time. However, these machines were not intended for use on large, data-intensive projects. Because the purpose of our report is to compare these technologies in terms of efficiency, and in an effort to be as unbiased as possible, we chose to use data collected from the desktop version of each technology. For Illumina, we used the MiSeq sequencer. For Ion Torrent, we used PGM, and for Oxford Nanopore, we used MinION.

One problem with this decision is that PacBio does not offer a desktop sequencer. Their only machine, the PacBio RS II, is noted for its high accuracy and long reads, but not necessarily efficiency[17]. While it would've been ideal to use data from a PacBio desktop unit, that was not possible, so we collected data for the RS II.

These circumstances reduce the validity of our comparison because the RS II was not built for the same purposes as the other three desktop sequencers. Despite this, we feel that given the situation, we have produced the best comparison possible to determine the most cost and time efficient technology.

Sources of MinIon data

Since the Oxford Nanopore technology is so new, there is significantly less data on its capabilities compared to the more established sequencers. The public's main source of unbiased data comes from a handful of published studies done by "early access" customers. The early access program allowed a small number of selected labs, institutions, and individuals to purchase an Oxford Nanopore sequencer in early spring 2014. Some of these groups went on to do small sequencing projects using the MinION, usually sequencing small bacterial genomes, and publishing the results. Besides giving a brief but useful explanation of the MinION sequencer itself, these studies provide unbiased information on the data this machine produces and represent our main source of information regarding the price, speed, and output of this technology. The disadvantage of this source is that it uses a small sample size. Due to the MinION's novelty, not many programs have tested it to provide results. In the future, there are sure to be more reliable sources with MinION data, but for our time frame, this was the best option.

Variability of MinION results

The results of a sequencing run can vary by a lot of factors. Because MinION is such a new technology, many of its protocols and reagents are in flux. This is especially true of the critical flow cells that form the basic consumable reagent of the MinION. Early access users were given a generation of flow cells called R6. Within a short time, improvements were applied, and versions R7 (released in July 2014) and R7.3 (released in September 2014) were made available.

The advanced chemistry of these flow cells, while outside the scope of this report, plays a huge role in the quality of the results. This is demonstrated by certain studies, which used the older R6 technology, and concluded that the results of Oxford Nanopore sequencing are so error prone that the technology is practically useless, yet others using the R7 technology were able to produce vastly improved results[14]. We provide a table (Table 1) showing the average results of a study which implemented the most recent iterations of the flow cell technology: The R7 and R7.3[1].

	R7	R7.3
Run Time	72 hr	48 hr
template reads	43,656 (272Mb)	39,819 (163 Mb)
complement reads	23,338 (125 Mb)	18,889 (84 Mb)
2D reads	20,087 (131 Mb)	11,823 (64.53 Mb)
Read Length (2D)	6,543	5,458

Table 1: MinIon results of older R7 flow cell vs. newer R7.3.

IV Data Summary

Cost per Megabase

Figure 4 displays our results for cost per megabase of sequenced data. The Oxford Nanopore cost was calculated using the reported flow cell price of $1,000 and dividing it by the average output per run, which is 150 Mb[1]. As evidenced from Figure 4, Oxford Nanopore produces the most expensive reads by a factor of at least 3 when compared to the other machines. Illumina had the cheapest cost per megabase at around 50 cents per million bases. This means that a human genome (3,000 Mb) could be sequenced for around

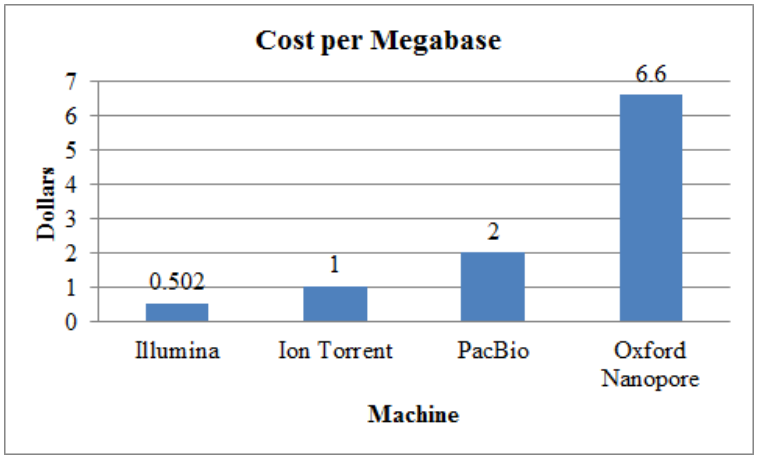

Figure 4: Cost per Megabase[1]

$1,500 using Illumina, whereas Oxford Nanopore would cost $19,800.

Time per Run

The data we compiled for time per run are shown in Figure 5, where time is measured in hours. We see that Oxford Nanopore has the longest time per run by 78% (when compared to Illumina). Oxford Nanopore also takes 15 times longer than Ion Torrent and PacBio. While this measurement does serve as an indicator for efficiency in terms of time, it must also be considered in conjunction with output per run to determine true efficiency. We will discuss output per run next.

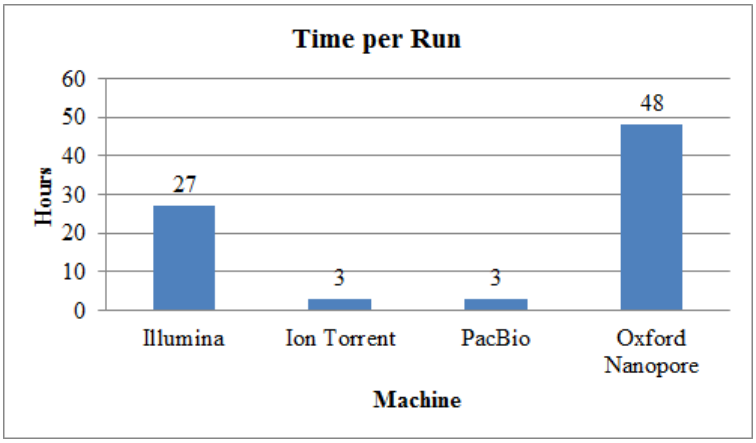

Figure 5: Time per Run[1,12]

Output per Run

Output per run is measured in megabases (Mb) of data. The results of our research can be seen in Figure 6. We found that Illumina has the highest output per run, being more than three times larger than PacBio, more than ten times larger than Oxford Nanopore, and about twenty times larger than Ion Torrent. In our discussion, we will consider time per run and output per run together in order to develop a more holistic analysis of efficiency.

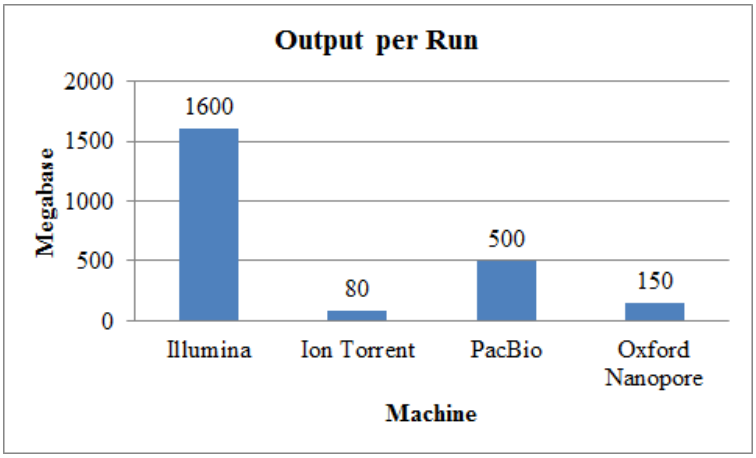

Figure 6: Output per Run[1,12,13,14]

V Discussion

To interpret our results, we combined our data found for time per run and output per run to create a new measurement, throughput, or output per hour, measured in megabases. We did this by multiplying the inverse of our time per run statistics by output per run as shown in Figure 7 below:

$$\frac{output}{run} \times \frac{run}{hour} = \frac{output}{hour}(throughput)$$

Figure 7: Throughput Equation

The results from these calculations are shown in Figure 8. This interpretation suggests that PacBio is by far the most efficient in terms of time and productivity, delivering 166.67 megabases of sequencing data per hour. Oxford Nanopore definitely underperforms in this category, producing a mere 3.13 megabases of sequencing data per hour.

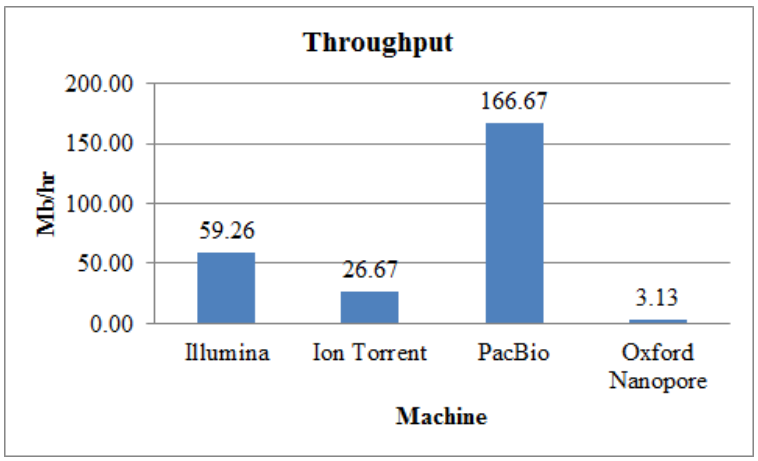

Figure 8: Throughput (Output per hour)

Having determined the most productive and expeditive machine, we will now discuss the machine that is the most efficient in terms of cost. Referring back to Figure 4, we see that Illumina is the most cost efficient, costing only .502 dollars to sequence one million bases, with Oxford Nanopore, once again, as the least efficient, costing $6.60 to sequence one million bases. As mentioned before, this seemingly small price difference is very significant when

sequencing billions of bases.

From this analysis, we have determined that Illumina's MiSeq is more efficient in terms of cost, and PacBio's RS II is more efficient in terms of time. However, the aim of this report is to determine the most overall efficient machine. To accomplish this, we will compare the degrees to which each is the most efficient. From our data, we calculated ratios of each measurement and determined that Illumina is 3.98 times more cost-efficient than PacBio, whereas PacBio's throughput is only 2.81 times larger than Illumina. Next, we will take these interpretations and formulate our conclusion.

VI Conclusion

In this report, we aim to determine if Oxford Nanopore deserves the hype it's receiving by the media and if it's desktop MinION unit is more efficient in terms of money, time, and output than the leading desktop DNA sequencing machines. If it is not, then we aim to determine which machine is the most efficient. To determine this, we gathered research on four technologies (Illumina, Ion Torrent, PacBio, and Oxford Nanopore). We compiled statistics of cost per megabase, time per run, and output per run. From these data, we combined time per run and output per run to generate a new measurement, throughput, or output per hour. Our analysis of throughput indicated to us which machine was the most efficient in terms of time and productivity, and cost per megabase indicated to us which machine was the most efficient in terms of money.

From our analysis, we conclude that Oxford Nanopore is far from being more efficient than the other leading technologies. In fact, it is, by a large margin, the least efficient of those that we compared. Because of this result, we have also determined which of the current technologies is the leader in efficiency.

We conclude that Illumina provides the most overall efficient desktop machine (MiSeq). Even though PacBio has a higher throughput, its sequencing cost is four times more than

Illumina. In practice, this means that while PacBio can sequence a human genome in 18 hours (50 hours for Illumina), it would cost $6,000 to do so compared to about $1,500 for Illumina.

As the field of genomics grows, many researchers and institutions around the world are considering investing in a DNA sequencing machine. This research can be of great value to those who are considering purchasing Oxford Nanopore's MinION unit because of all the public hype the machine is receiving. However, we find that they will save a lot more time and money if they instead invest in purchasing Illumina's MiSeq unit.

Although our results indicate that the Oxford Nanopore technology is not up to par with the standard sequencing technologies yet and may have prematurely received praises from the media, we still believe it has unrealized potential. The fundamental design utilizing a synthetic protein that allows for a straight read of DNA rather than reading the sequence by synthesis is miles ahead of the technology currently being used. We believe that the advantages from this technology will become more apparent as the process is developed and refined. In fact, all of the established technologies that we compared Oxford Nanopore to went through a development period of their own before they became the reliable machines they are today. Still, these development periods typically take years or even decades. While we would not recommend purchasing the MinION unit now or even in a few years, we also advise our readers to keep a close eye on its development, for it may one day result in the most powerful and efficient DNA sequencer the world has seen yet.

Thank you for reading our report. If there are any questions, comments, or clarifications, we can be contacted by email at any of these addresses: mollyloschiavo@gmail.com, kevinlboehme@gmail.com, coric8@gmail.com.

Bibliography

1. Quick, J., Quinlan, A. R., & Loman, N. J. A reference bacterial genome dataset generated on the minion portable single-molecule nanopore sequencer. *Gigascience* **3**, 22 (2014).

2. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).

3. Salto-Tellez, M. & Castro, D. G. Next-generation sequencing: a change of paradigm in molecular diagnostic validation. *J. Pathol.* **234**, 5–10 (2014).

4. Maitra, R. D., Kim, J., & Dunbar, W. B. Recent advances in nanopore sequencing. *Electrophoresis* **33**, 3418–3428 (2012).

5. Illumina sequencing by synthesis. http://www.yerkes.emory.edu/nhp_genomics_core/Services/Sequencing.html (2014).

6. Herper, M. The new, fast gene machine. http://www.forbes.com/forbes/2009/1005/revolutionaries-science-genomics-gene-machine.html (2009).

7. Kelman, Z. & O'Donnell, M. DNA polymerase III holoenzyme: structure and function of a chromosomal replicating machine. *Annu. Rev. Biochem.* **64**, 171–200 (1995).

8. Schaffer, A. Nanopore sequencing. http://www2.technologyreview.com/article/427677/nanopore-sequencing/ (2012).

9. Oxford nanopore technologies. https://www.nanoporetech.com/ (2014).

10. Xuan, J., Yu, Y., Qing, T., Guo, L., & Shi, L. Next-generation sequencing in the clinic: Promises and challenges. *Cancer Lett.* **340**, 284–295 (2013).

11. National Human Genome Research Institute (NHGRI). DNA sequencing costs. http://www.genome.gov/sequencingcosts/ (2014).

12. van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. Ten years of next-generation sequencing technology. *Trends in genetics* **30**, 418–426 (2014).

13. Loman, N., Misra, R., Dallman, T., Constantinidou, C., Gharbia, S., Wain, J., & Pallen, M. Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30**, 562–562 (2012).

14. Mikheyev, A. S. & Tin, M. M. A first look at the oxford nanopore minion sequencer. *Mol. Ecol. Resour.* **14**, 1097–1102 (2014).

15. The human genome project completion: Frequently asked questions. http://www.genome.gov/11006943 (2014).

16. Sequencing systems. http://systems.illumina.com/systems/sequencing.html (2014).

17. PACBIO RS II. http://www.pacificbiosciences.com/products/ (2014).

Appendix

Glossary

Base-pair: A chemical bond between two nucleotides. A and T bond exclusively, and C and G bond exclusively. DNA is naturally found in a state where all of it's nucleotides are base-paired to the nucleotides of another DNA strand. This term is often interchangeable with "nucleotide" and "base".

DNA: Deoxyribonucleic acid. A molecule whose sequence of nucleotides (see below) contains the information a cell needs to produce all proteins necessary for survival. Each cell in the human body contains a complete set of these "instructions".

DNA polymerase: Enzyme (protein) that dramatically increases the speed of the nucleotide base-pairing process by matching each base in a DNA strand with it's complementary base and creating the bond that holds the two bases together.

Flow cells: A consumable reagent for the Oxford Nanopore MinION that contains all components necessary for one "sequencing run". Houses the chemistry and nanopores needed to sequence and produce the DNA data.

Fluorophore: Molecule that fluoresces a particular color when activated. Can be activated by a laser (as in Illumina) or by DNA polymerase (as in PacBio). In sequencing, particular colors of fluorophores are attached to particular bases. For example, all C's may be attached to red fluorophores, all T's may be attached to green fluorophores, all G's may be attached to blue fluorophores, and all A's may be attached to yellow fluorophores.

Genome: The entire DNA sequence found in an organism's cell. Contains all the information for creating, growing, and sustaining that organism. The human genome is around 3 billion base pairs long.

Megabase: One million base-pairs. Used as a unit of measurement for how many bases were sequenced.

Next-generation DNA sequencing technologies: Series of DNA sequencing technologies that emerged in the 1990's and new millennium in response to the Human Genome Project. These technologies vastly outperform those that had been used before this time, which is why they were considered to be a whole new generation.

Nucleotide: A building block of DNA. Also known as a "base" or "base pair". There are four kinds: A, G, C, and T. The order of nucleotides determines the DNA's sequence.

Throughput: Output per hour given in Megabases/hour produced by a DNA sequencer. Helpful metric to compare sequencing technologies, indicating productivity.

nature biotechnology

Next-generation DNA sequencing

Jay Shendure[1] & Hanlee Ji[2]

DNA sequence represents a single format onto which a broad range of biological phenomena can be projected for high-throughput data collection. Over the past three years, massively parallel DNA sequencing platforms have become widely available, reducing the cost of DNA sequencing by over two orders of magnitude, and democratizing the field by putting the sequencing capacity of a major genome center in the hands of individual investigators. These new technologies are rapidly evolving, and near-term challenges include the development of robust protocols for generating sequencing libraries, building effective new approaches to data-analysis, and often a rethinking of experimental design. Next-generation DNA sequencing has the potential to dramatically accelerate biological and biomedical research, by enabling the comprehensive analysis of genomes, transcriptomes and interactomes to become inexpensive, routine and widespread, rather than requiring significant production-scale efforts.

The field of DNA sequencing technology development has a rich and diverse history[1,2]. However, the overwhelming majority of DNA sequence production to date has relied on some version of the Sanger biochemistry[3]. Over the past five years, the incentive for developing entirely new strategies for DNA sequencing has emerged on at least four levels, undeniably reinvigorating this field (for a review, see ref. 4). First, in the wake of the Human Genome Project, there are few remaining avenues of optimization through which significant reductions in the cost of conventional DNA sequencing can be achieved. Second, the potential utility of short-read sequencing has been tremendously strengthened by the availability of whole genome assemblies for *Homo sapiens* and all major model organisms, as these effectively provide a reference against which short reads can be mapped. Third, a growing variety of molecular methods have been developed, whereby a broad range of biological phenomena can be assessed by high-throughput DNA sequencing (e.g., genetic variation, RNA expression, protein-DNA interactions and chromosome conformation). And fourth, general progress in technology across disparate fields, including microscopy, surface chemistry, nucleotide biochemistry, polymerase engineering, computation, data storage and others, have made alternative strategies for DNA sequencing increasingly practical to realize.

Here, we review the current crop of next-generation DNA sequencing platforms: how they work, their relative strengths and limitations, and current and emerging applications. We briefly discuss related developments in this field, such as new software tools and front-end methods for isolating arbitrary genomic subsets. We emphasize that the DNA sequencing technology field has become a quickly moving target, and we can at best provide a snapshot of this particular moment.

[1]Department of Genome Sciences, University of Washington, Foege Building S-250, Box 355065, 1705 NE Pacific St., Seattle, Washington 98195-5065, USA. [2]Stanford Genome Technology Center and Division of Oncology, Dept. of Medicine, Stanford University School of Medicine, CCSR 3215, 269 Campus Drive, Stanford, California 94305, USA. Correspondence should be addressed to J.S. (shendure@u.washington.edu) or H.J. (genomics_ji@stanford.edu).

Published online 9 October 2008; doi:10.1038/nbt1486

Sanger sequencing

Since the early 1990s, DNA sequence production has almost exclusively been carried out with capillary-based, semi-automated implementations of the Sanger biochemistry[3,5,6] (**Fig. 1a**). In high-throughput production pipelines, DNA to be sequenced is prepared by one of two approaches: first, for shotgun *de novo* sequencing, randomly fragmented DNA is cloned into a high-copy-number plasmid, which is then used to transform *Escherichia coli;* or second, for targeted resequencing, PCR amplification is carried out with primers that flank the target. The output of both approaches is an amplified template, either as many 'clonal' copies of a single plasmid insert present within a spatially isolated bacterial colony that can be picked, or as many PCR amplicons present within a single reaction volume. The sequencing biochemistry takes place in a 'cycle sequencing' reaction, in which cycles of template denaturation, primer annealing and primer extension are performed. The primer is complementary to known sequence immediately flanking the region of interest. Each round of primer extension is stochastically terminated by the incorporation of fluorescently labeled dideoxynucleotides (ddNTPs). In the resulting mixture of end-labeled extension products, the label on the terminating ddNTP of any given fragment corresponds to the nucleotide identity of its terminal position. Sequence is determined by high-resolution electrophoretic separation of the single-stranded, end-labeled extension products in a capillary-based polymer gel. Laser excitation of fluorescent labels as fragments of discreet lengths exit the capillary, coupled to four-color detection of emission spectra, provides the readout that is represented in a Sanger sequencing 'trace'. Software translates these traces into DNA sequence, while also generating error probabilities for each base-call[7,8]. The approach that is taken for subsequent analysis—for example, genome assembly or variant identification—depends on precisely what is being sequenced and why. Simultaneous electrophoresis in 96 or 384 independent capillaries provides a limited level of parallelization.

After three decades of gradual improvement, the Sanger biochemistry can be applied to achieve read-lengths of up to ~1,000 bp, and per-base 'raw' accuracies as high as 99.999%. In the context of high-throughput shotgun genomic sequencing, Sanger sequencing costs on the order of $0.50 per kilobase.

Second-generation DNA sequencing

Alternative strategies for DNA sequencing can be grouped into several categories (as discussed previously in ref. 4). These include (i) microelectrophoretic methods[9] (**Box 1**), (ii) sequencing by hybridization[10] (**Box 2**), (iii) real-time observation of single molecules[11,12] (**Box 3**) and (iv) cyclic-array sequencing (J.S. *et al.*[13] and ref. 14). Here, we use 'second-generation' in reference to the various implementations of cyclic-array sequencing that have recently been realized in a commercial product (e.g., 454 sequencing (used in the 454 Genome Sequencers, Roche Applied Science; Basel), Solexa technology (used in the Illumina (San Diego) Genome Analyzer), the SOLiD platform (Applied Biosystems; Foster City, CA, USA), the Polonator (Dover/Harvard) and the HeliScope Single Molecule Sequencer technology (Helicos; Cambridge, MA, USA). The concept of cyclic-array sequencing can be summarized as the sequencing of a dense array of DNA features by iterative cycles of enzymatic manipulation and imaging-based data collection[15] (Shendure and colleagues[16]). Two reports in 2005 described the first integrated implementations of cyclic-array strategies that were both practical and cost-competitive with conventional sequencing (J.S. *et al.*[13] and ref. 14), and other groups have quickly followed[17,18].

Although these platforms are quite diverse in sequencing biochemistry as well as in how the array is generated, their work flows are conceptually similar (**Fig. 1b**). Library preparation is accomplished by random fragmentation of DNA, followed by *in vitro* ligation of common adaptor sequences. Alternative protocols can be used to generate jumping libraries of mate-paired tags with controllable distance distributions[13,19]. The generation of clonally clustered amplicons to serve as sequencing features can be achieved by several approaches, including *in situ* polonies[15], emulsion PCR[20] or bridge PCR[21,22] (**Fig. 2**). What is common to these methods is that PCR amplicons derived from any given single library molecule end up spatially clustered, either to a single location on a planar substrate (*in situ* polonies, bridge PCR), or to the surface of micron-scale beads, which can be recovered and arrayed (emulsion PCR). The sequencing process itself consists of alternating cycles of enzyme-driven biochemistry and imaging-based data acquisition (**Fig. 3**). The platforms that are discussed here all rely on sequencing by synthesis, that is, serial extension of primed templates, but the enzyme driving the synthesis can be either a polymerase[16,23] or a ligase[13,24]. Data are acquired by imaging of the full array at each cycle (e.g., of fluorescently labeled nucleotides incorporated by a polymerase).

Global advantages of second-generation or cyclic-array strategies, relative to Sanger sequencing, include the following: (i) *in vitro* construction of a sequencing library, followed by *in vitro* clonal amplification to generate sequencing features, circumvents several bottlenecks that restrict the parallelism of conventional sequencing (that is, transformation of *E. coli* and colony picking). (ii) Array-based sequencing enables a much higher degree of parallelism than conventional capillary-based sequencing. As the effective size of sequencing features can be on the order of 1 μm, hundreds of millions of sequencing reads can potentially be obtained in parallel by rastered imaging of a reasonably sized surface area. (iii) Because array features are immobilized to a planar surface, they can be enzymatically manipulated by a single reagent volume. Although microliter-scale reagent volumes are used in practice, these are essentially amortized over the full set of sequencing features on the array, dropping the effective reagent volume per feature to the

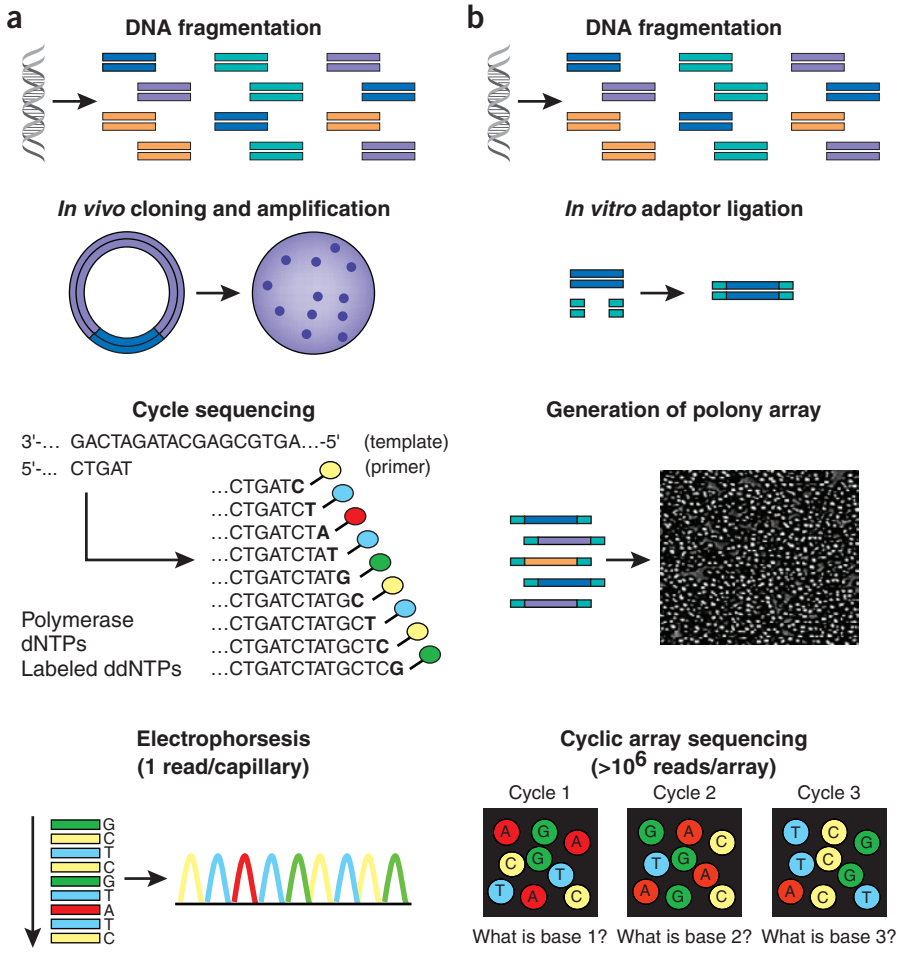

Figure 1 Work flow of conventional versus second-generation sequencing. (**a**) With high-throughput shotgun Sanger sequencing, genomic DNA is fragmented, then cloned to a plasmid vector and used to transform *E. coli*. For each sequencing reaction, a single bacterial colony is picked and plasmid DNA isolated. Each cycle sequencing reaction takes place within a microliter-scale volume, generating a ladder of ddNTP-terminated, dye-labeled products, which are subjected to high-resolution electrophoretic separation within one of 96 or 384 capillaries in one run of a sequencing instrument. As fluorescently labeled fragments of discrete sizes pass a detector, the four-channel emission spectrum is used to generate a sequencing trace. (**b**) In shotgun sequencing with cyclic-array methods, common adaptors are ligated to fragmented genomic DNA, which is then subjected to one of several protocols that results in an array of millions of spatially immobilized PCR colonies or 'polonies'[15]. Each polony consists of many copies of a single shotgun library fragment. As all polonies are tethered to a planar array, a single microliter-scale reagent volume (e.g., for primer hybridization and then for enzymatic extension reactions) can be applied to manipulate all array features in parallel. Similarly, imaging-based detection of fluorescent labels incorporated with each extension can be used to acquire sequencing data on all features in parallel. Successive iterations of enzymatic interrogation and imaging are used to build up a contiguous sequencing read for each array feature.

scale of picoliters or femtoliters. Collectively, these differences translate into dramatically lower costs for DNA sequence production.

The advantages of second-generation DNA sequencing are currently offset by several disadvantages. The most prominent of these include read-length (for all of the new platforms, read-lengths are currently much shorter than conventional sequencing) and raw accuracy (on average, base-calls generated by the new platforms are at least tenfold less accurate than base-calls generated by Sanger sequencing). Although these limitations create important algorithmic challenges for the immediate future, we should bear in mind that these technologies will continue to improve with respect to these parameters, much as conventional sequencing progressed gradually over three decades to reach its current level of technical performance.

454 pyrosequencing. The 454 system was the first next-generation sequencing platform available as a commercial product[14]. In this approach, libraries may be constructed by any method that gives rise to a mixture of short, adaptor-flanked fragments. Clonal sequencing features are generated by emulsion PCR[20], with amplicons captured to the surface of 28-μm beads (**Fig. 2a**). After breaking the emulsion, beads are treated with denaturant to remove untethered strands, and then subjected to a hybridization-based enrichment for amplicon-bearing beads (that is, those that were present in an emulsion compartment supporting a productive PCR reaction). A sequencing primer is hybridized to the universal adaptor at the appropriate position and orientation, that is, immediately adjacent to the start of unknown sequence.

Sequencing is performed by the pyrosequencing method[25] (**Fig. 3a**). In brief, the amplicon-bearing beads are preincubated with *Bacillus stearothermophilus* (*Bst*) polymerase and single-stranded binding protein, and then deposited on to a microfabricated array of picoliter-scale wells (with dimensions such that only one bead will fit per well) to render this biochemistry compatible with array-based sequencing. Smaller beads are also added, bearing immobilized enzymes also required for pyrosequencing (ATP sulfurylase and luciferase). During the sequencing, one side of the semi-ordered array functions as a flow cell for introducing and removing sequencing reagents, whereas the other side is bonded to a fiber-optic bundle for CCD (charge-coupled device)-based signal detection. At each of several hundred cycles, a single species of unlabeled nucleotide is introduced. On templates where this results in an incorporation event, pyrophosphate is released. Via ATP sulfurylase and luciferase, incorporation events immediately drive the generation of a burst of light, which is detected

Box 1 Microchip-based electrophoretic sequencing

Significant progress has been made toward developing methods whereby conventional electrophoretic sequencing can be carried out on a microfabricated device[78,79]. The primary advantages of this approach include faster processing times and substantial reductions in reagent consumption. An ideal device for this purpose would integrate all aspects of sample processing, with microfluidic transport of the reaction volume between steps, for example, clonal amplification by nanoliter-scale PCR from a single cell or a single template molecule; template purification; cycle sequencing reaction; isolation and concentration of extension fragments; injection into a microchannel for electrophoretic separation (potentially parallelized; e.g., with 384 or more channels concentrically arranged around a rotating fluorescence scanner[80]). Many of the key challenges have already been overcome in proof-of-concept experiments[9,81–83]. Although it is unclear in the immediate moment whether these efforts will be able to keep pace with cyclic-array sequencing and other strategies, it is worth bearing in mind that the Sanger biochemistry coupled to electrophoretic separation remains by far the best option for DNA sequencing in terms of read-length and accuracy; we simply lack methods to parallelize it to the extent possible with cyclic-array strategies. One could imagine that 'lab-on-a-chip' nucleic acid analysis could supplant conventional DNA sequencing for low-scale applications and may also prove useful in the context of point-of-care diagnostics.

Box 2 Sequencing by hybridization

The basic concept of sequencing by hybridization is that the differential hybridization of labeled nucleic acid fragments to an array of oligonucleotide probes can be used to precisely identify variant positions. Usually, the oligos tethered to the array are designed as a tiling representation of the reference sequence corresponding to the genome of interest. With the approach taken by Affymetrix (Santa Clara, CA, USA) and Perlegen (Mountain View, CA, USA) (in performing extensive SNP discovery in human[84], mouse[85] and yeast[86], for example), each possible single-base substitution is represented on the array by an independent feature. Roche NimbleGen (Madison, WI, USA), in performing sequencing by hybridization of microbial genomes, takes a two-tier approach, with an initial array directed at performing approximate localization, and a second custom array directed at pinpointing and confirmation of variant positions[87]. Although microarrays are clearly useful and cost effective for genomic resequencing as well as a range of other genome-scale applications[10], it is unclear what will happen as next-generation sequencing technologies begin to compete for many of the same applications (e.g., resequencing, but also expression analysis, structural variation analysis, DNA-protein binding).

In terms of sequencing, limitations of microarrays include the following: (i) sequences that are repetitive or subject to cross-hybridization cannot easily be interrogated; (ii) it remains unclear how *de novo* sequencing can be achieved with hybridization-based strategies; and (iii) without very careful data analysis, false positives pose an important problem, and it is not clear how to obtain the equivalent of redundant coverage that is possible with conventional and cyclic-array sequencing. Thus far, sequencing by hybridization has likely had its greatest impact in the context of genome-wide association studies, which rely on array-based interrogation (that is, genotyping by hybridization) of a highly defined set of discontiguous genomic coordinates.

A different (and earlier) take on the idea of 'sequencing by hybridization' involves serial or parallel interrogation with comprehensive sets of short oligonucleotides (e.g., 4,096 × 6-mers or 8,192 × 7-mers) followed by sequence reconstruction[88,89]. Recently, this basic strategy was used in the context of an array of rolling circle-amplified sequencing features[90] to perform resequencing of an *E. coli* genome[91]. This successful proof-of-concept is perhaps better classified as a cyclic-array method, where serial hybridization rather than polymerase-driven synthesis was used for the actual sequencing.

Box 3 Sequencing in real time

Several academic groups and companies are working on technologies for ultra-fast DNA sequencing that are substantially different from the current crop of available next-generation platforms. One approach is nanopore sequencing, in which nucleic acids are driven through a nanopore (either a biological membrane protein such as alpha-hemolysin or a synthetic pore)[92]. Fluctuations in DNA conductance through the pore, or, potentially, the detection of interactions of individual bases with the pore, are used to infer the nucleotide sequence. Although progress has been made in achieving early proof-of-concept demonstrations with such methods[11,12,93,94], major technical challenges remain along the path to a truly practical nanopore-based sequencing platform. Another approach involves the real-time monitoring of DNA polymerase activity. Nucleotide incorporations can potentially be detected through FRET (fluorescence resonance energy transfer) interactions between a fluorophore-bearing polymerase and gamma phosphate-labeled nucleotides (Visigen; Houston), or with zero-mode waveguides (Pacific Biosciences; Menlo Park, CA, USA), with which illumination can be restricted to a zeptoliter-scale volume around a surface-tethered polymerase such that incorporation of nucleotides (with fluorescent labels on phosphate groups) can be observed with low background[95]. Pacific Biosciences recently demonstrated substantial progress toward a working technology, including the potential for longer reads than Sanger sequencing, in several presentations and publications[96,97]. Although technical hurdles remain and the bar has been raised by cyclic-array methods, we are also unlikely to run out of nucleotides to sequence anytime soon.

by the CCD as corresponding to the array coordinates of specific wells. In contrast with other platforms, therefore, the sequencing by synthesis must be monitored 'live' (that is, the camera does not move relative to the array). Across multiple cycles (e.g., A-G-C-T-A-G-C-T…), the pattern of detected incorporation events reveals the sequence of templates represented by individual beads. Like the HeliScope (discussed below), the sequencing is 'asynchronous' in that some features may get ahead or behind other features depending on their sequence relative to the order of base addition.

A major limitation of the 454 technology relates to homopolymers (that is, consecutive instances of the same base, such as AAA or GGG). Because there is no terminating moiety preventing multiple consecutive incorporations at a given cycle, the length of all homopolymers must be inferred from the signal intensity. This is prone to a greater error rate than the discrimination of incorporation versus nonincorporation. As a consequence, the dominant error type for the 454 platform is insertion-deletion, rather than substitution. Relative to other next-generation platforms, the key advantage of the 454 platform is read-length. For example, the 454 FLX instrument generates ~400,000 reads per instrument-run at lengths of 200 to 300 bp. Currently, the per-base cost of sequencing with the 454 platform is much greater than that of other platforms (e.g., SOLiD and Solexa) but it may be the method of choice for certain applications where long read-lengths are critical (e.g., *de novo* assembly and metagenomics).

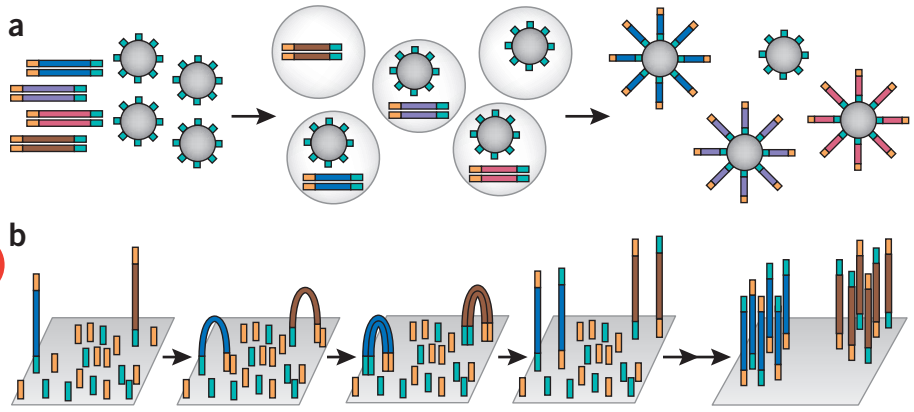

Figure 2 Clonal amplification of sequencing features. (**a**) The 454, the Polonator and SOLiD platforms rely on emulsion PCR[20] to amplify clonal sequencing features. In brief, an *in vitro*–constructed adaptor-flanked shotgun library (shown as gold and turquoise adaptors flanking unique inserts) is PCR amplified (that is, multi-template PCR, not multiplex PCR, as only a single primer pair is used, corresponding to the gold and turquoise adaptors) in the context of a water-in-oil emulsion. One of the PCR primers is tethered to the surface (5′-attached) of micron-scale beads that are also included in the reaction. A low template concentration results in most bead-containing compartments having either zero or one template molecule present. In productive emulsion compartments (where both a bead and template molecule is present), PCR amplicons are captured to the surface of the bead. After breaking the emulsion, beads bearing amplification products can be selectively enriched. Each clonally amplified bead will bear on its surface PCR products corresponding to amplification of a single molecule from the template library. (**b**) The Solexa technology relies on bridge PCR[21,22] (aka 'cluster PCR') to amplify clonal sequencing features. In brief, an *in vitro*–constructed adaptor-flanked shotgun library is PCR amplified, but both primers densely coat the surface of a solid substrate, attached at their 5′ ends by a flexible linker. As a consequence, amplification products originating from any given member of the template library remain locally tethered near the point of origin. At the conclusion of the PCR, each clonal cluster contains ~1,000 copies of a single member of the template library. Accurate measurement of the concentration of the template library is critical to maximize the cluster density while simultaneously avoiding overcrowding.

Illumina Genome Analyzer. Commonly referred to as 'the Solexa', this platform has its origins in work by Turcatti and colleages[22,23] and the merger of four companies—Solexa (Essex, UK), Lynx Therapeutics (Hayward, CA, USA), Manteia Predictive Medicine (Coinsins, Switzerland) and Illumina. Libraries can be constructed by any method that gives rise to a mixture of adaptor-flanked fragments up to several hundred base-pairs (bp) in length. Amplified sequencing features are generated by bridge PCR[21,22] (**Fig. 2b**). In this approach, both forward and reverse PCR primers are tethered to a solid substrate by a flexible linker, such that all amplicons arising from any single template molecule during the amplification remain immobilized and clustered to a single physical location on an array. On the Illumina platform, the bridge PCR is somewhat unconventional in relying on alternating cycles of extension with *Bst* polymerase and denaturation with formamide. The resulting 'clusters' each consist of ~1,000 clonal amplicons. Several million clusters can be amplified to distinguishable locations within each of eight independent 'lanes' that

are on a single flow-cell (such that eight independent libraries can be sequenced in parallel during the same instrument run). After cluster generation, the amplicons are single stranded (linearization) and a sequencing primer is hybridized to a universal sequence flanking the region of interest. Each cycle of sequence interrogation consists of single-base extension with a modified DNA polymerase and a mixture of four nucleotides (**Fig. 3b**). These nucleotides are modified in two ways. They are 'reversible terminators', in that a chemically cleavable moiety at the 3′ hydroxyl position allows only a single-base incorporation to occur in each cycle; and one of four fluorescent labels, also chemically cleavable, corresponds to the identity of each nucleotide[23]. After single-base extension and acquisition of images in four channels, chemical cleavage of both groups sets up for the next cycle. Read-lengths up to 36 bp are currently routine; longer reads are possible but may incur a higher error rate.

Read-lengths are limited by multiple factors that cause signal decay and dephasing, such as incomplete cleavage of fluorescent labels or terminating moieties. The dominant error type is substitution, rather than insertions or deletions (and homopolymers are certainly less of an issue than with other platforms such as 454). Average raw error-rates are on the order of 1–1.5%, but higher accuracy bases with error rates of 0.1% or less can be identified through quality metrics associated with each base-call. As with other systems, modifications have recently enabled mate-paired reads; for example, each sequencing

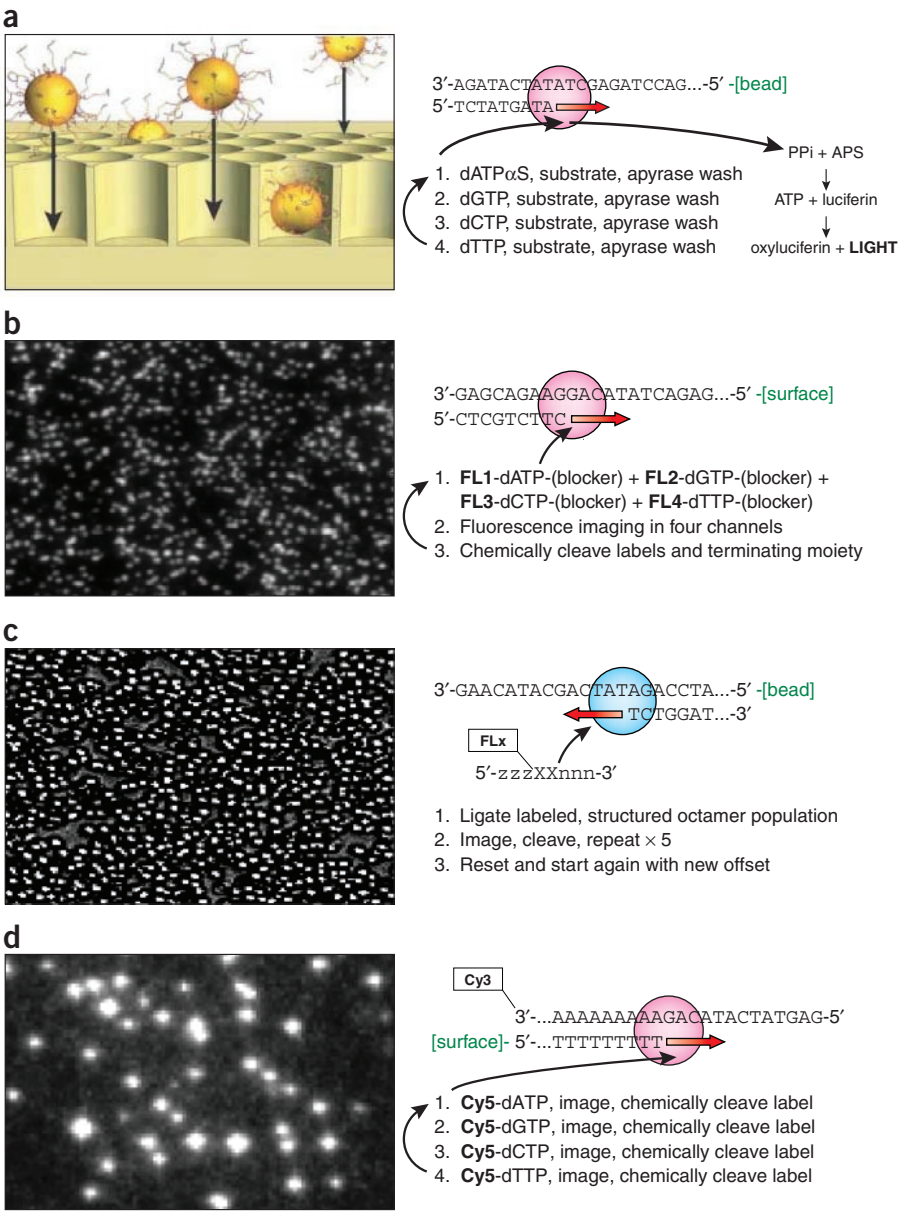

Figure 3 Strategies for cyclic array sequencing. (**a**) With the 454 platform, clonally amplified 28-μm beads generated by emulsion PCR serve as sequencing features and are randomly deposited to a microfabricated array of picoliter-scale wells. With pyrosequencing, each cycle consists of the introduction of a single nucleotide species, followed by addition of substrate (luciferin, adenosine 5′-phosphosulphate) to drive light production at wells where polymerase-driven incorporation of that nucleotide took place. This is followed by an apyrase wash to remove unincorporated nucleotide. Image from Margulies *et al.* (2005)[14]. (**b**) With the Solexa technology, a dense array of clonally amplified sequencing features is generated directly on a surface by bridge PCR (aka cluster PCR). Each sequencing cycle includes the simultaneous addition of a mixture of four modified deoxynucleotide species, each bearing one of four fluorescent labels and a reversibly terminating moiety at the 3′ hydroxyl position. A modified DNA polymerase drives synchronous extension of primed sequencing features. This is followed by imaging in four channels and then cleavage of both the fluorescent labels and the terminating moiety. (**c**) With the SOLiD and the Polonator platforms, clonally amplified 1-μm beads are used to generate a disordered, dense array of sequencing features[13]. Sequencing is performed with a ligase, rather than a polymerase[13,24,26–28]. With SOLiD, each sequencing cycle introduces a partially degenerate population of fluorescently labeled octamers. The population is structured such that the label correlates with the identity of the central 2 bp in the octamer (the correlation with 2 bp, rather than 1 bp, is the basis of two-base encoding)[26]. After ligation and imaging in four channels, the labeled portion of the octamer (that is, 'zzz') is cleaved via a modified linkage between bases 5 and 6, leaving a free end for another cycle of ligation. Several such cycles will iteratively interrogate an evenly spaced, discontiguous set of bases. The system is then reset (by denaturation of the extended primer), and the process is repeated with a different offset (e.g., a primer set back from the original position by one or several bases) such that a different set of discontiguous bases is interrogated on the next round of serial ligations. (**d**) With the HeliScope platform, single nucleic acid molecules are sequenced directly, that is, there is no clonal amplification step required. Poly-A–tailed template molecules are captured by hybridization to surface-tethered poly-T oligomers to yield a disordered array of primed single-molecule sequencing templates. Templates are labeled with Cy3, such that imaging can identify the subset of array coordinates where a sequencing read is expected. Each cycle consists of the polymerase-driven incorporation of a single species of fluorescently labeled nucleotide at a subset of templates, followed by fluorescence imaging of the full array and chemical cleavage of the label. Image from Braslavsky *et al.* (2003)[30].

Table 1 Second-generation DNA sequencing technologies

	Feature generation	Sequencing by synthesis	Cost per megabase	Cost per instrument	Paired ends?	1° error modality	Read-length	References
454	Emulsion PCR	Polymerase (pyrosequencing)	~$60	$500,000	Yes	Indel	250 bp	14,20
Solexa	Bridge PCR	Polymerase (reversible terminators)	~$2	$430,000	Yes	Subst.	36 bp	17,22
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)	~$2	$591,000	Yes	Subst.	35 bp	13,26
Polonator	Emulsion PCR	Ligase (nonamers)	~$1	$155,000	Yes	Subst.	13 bp	13,20
HeliScope	Single molecule	Polymerase (asynchronous extensions)	~$1	$1,350,000	Yes	Del	30 bp	18,30

The pace with which the field is moving makes it likely that estimates for costs and read-lengths will be quickly outdated. Vendors including Roche Applied Science, Illumina, and Applied Biosystems have major upgrade releases currently in progress. Estimated costs-per-megabase are approximate and inclusive only of reagents. Read-lengths are for single tags. Subst., substitutions; indel, insertions or deletions; del, deletions.

feature yielding 2 × 36 bp independent reads derived from each end of a given library molecule several hundred bases in length.

AB SOLiD. This platform has its origins in the system described by J.S. and colleagues[13] in 2005 and in work by McKernan and colleagues[26] at Agencourt Personal Genomics (Beverly, MA, USA) (acquired by Applied Biosystems (Foster City, CA, USA) in 2006). Libraries may be constructed by any method that gives rise to a mixture of short, adaptor-flanked fragments, though much effort with this system has been put into protocols for mate-paired tag libraries with controllable and highly flexible distance distributions[13,19]. Clonal sequencing features are generated by emulsion PCR, with amplicons captured to the surface of 1-μM paramagnetic beads[20] (**Fig. 2a**). After breaking the emulsion, beads bearing amplification products are selectively recovered, and then immobilized to a solid planar substrate to generate a dense, disordered array. Sequencing by synthesis is driven by a DNA ligase[13,24,26–28], rather than a polymerase. A universal primer complementary to adaptor sequence is hybridized to the array of amplicon-bearing beads. Each cycle of sequencing involves the ligation of a degenerate population of fluorescently labeled octamers (**Fig. 3c**). The octamer mixture is structured, in that the identity of specific position(s) within the octamer (e.g., base 5) correlate with the identity of the fluorescent label. After ligation, images are acquired in four channels, effectively collecting data for the same base positions across all template-bearing beads. Then, the octamer is chemically cleaved between positions 5 and 6, removing the fluorescent label. Progressive rounds of octamer ligation enable sequencing of every 5th base (e.g., bases 5, 10, 15, 20). Upon completing several such cycles, the extended primer is denatured to reset the system. Subsequent iterations of this process can be directed at a different set of positions (e.g., bases 4, 9, 14, 19) either by using a primer that is set back one or more bases from the adaptor-insert junction, or by using different mixtures of octamers where a different position (e.g., base 2) is correlated with the label. An additional feature of this platform involves the use of two-base encoding, which is an error-correction scheme in which two adjacent bases, rather than a single base, are correlated with the label[26]. Each base position is then queried twice (once as the first base, and once as the second base, in a set of 2 bp interrogated on a given cycle) such that miscalls can be more readily identified.

A related system to the SOLiD is the Polonator, also based in part on the system developed by J.S. and the Church group[13] at Harvard. This platform also uses sequencing features generated by emulsion PCR and sequencing by ligation. The cost of the instrument, however, is substantially lower than that of other second-generation sequencing instruments. Additionally, the instrument is open source and programmable, potentially enabling user innovation (e.g., the use of alternative biochemistries). The current read-lengths, however, may be significantly limiting.

An additional disadvantage, common to 454, SOLiD and the Polonator, is that emulsion PCR can be cumbersome and technically challenging. On the other hand, it is possible that sequencing on a high-density array of very small (1 μm) beads (with sequencing by ligation, polymerase extension, or another biochemistry) may represent the most straightforward opportunity to achieve extremely high data densities, simply because 1-μm beads physically exclude one another at a spacing that is on the order of the diffraction limit. Furthermore, high-resolution ordering of 1-μm bead arrays, as recently described[29], may enable the limit of one pixel per sequencing feature to be closely approached.

HeliScope. The Helicos sequencer[18], based on work by Quake's group[30], also relies on cyclic interrogation of a dense array of sequencing features. However, a unique aspect of this platform is that no clonal amplification is required. Instead, a highly sensitive fluorescence detection system is used to directly interrogate single DNA molecules via sequencing by synthesis. Template libraries, prepared by random fragmentation and poly-A tailing (that is, no PCR amplification), are captured by hybridization to surface-tethered poly-T oligomers to yield a disordered array of primed single-molecule sequencing templates. At each cycle, DNA polymerase and a single species of fluorescently labeled nucleotide are added, resulting in template-dependent extension of the surface-immobilized primer-template duplexes (**Fig. 3d**). After acquisition of images tiling the full array, chemical cleavage and release of the fluorescent label permits the subsequent cycle of extension and imaging. As described in a recent report[18], several hundred cycles of single-base extension (that is, A, G, C, T, A, G, C, T…) yield average read-lengths of 25 bp or greater. Notable aspects of this system include the following. First, like the 454 platform, the sequencing is asynchronous, as some strands will fall ahead or behind others in a sequence-dependent manner. Chance also plays a role, as some templates may simply fail to incorporate on a given cycle despite having the appropriate base at the next position. However, because these are single molecules, dephasing is not an issue, and such events do not in and of themselves lead to errors.

Second, no terminating moiety is present on the labeled nucleotides. As with the 454 system, therefore, homopolymer runs are an important issue. However, because single molecules are being sequenced, the problem can be mitigated by limiting the rate of incorporation events. Additionally, Harris *et al.*[18] noted that consecutive incorporations of labeled nucleotide at homopolymers produced a quenching interaction that enabled the authors to infer the discreet number of incorporations (e.g., A versus AA versus AAA).

Third, the raw sequencing accuracy can be substantially improved by a two-pass strategy in which the array of single-molecule templates (here with adaptors at both ends) is sequenced as described above, and then fully copied. As the newly synthesized strand is surface-teth-

ered, the original template can be removed by denaturing. Sequencing primed from the distal adaptor then yields a second sequence for the same template, obtained in the opposite orientation. Positions that are concordant between the two reads have *phred*-like quality scores approaching 30 (refs. 8,18).

And finally, largely secondary to the incorporation of contaminating, unlabeled or nonemitting bases, the dominant error type is deletion (2–7% error rate with one pass; 0.2–1% with two passes). However, substitution error rates are substantially lower (0.01–1% with one pass). With two passes, the per-base raw substitution error rate (approaching 0.001%) may currently be the lowest of all the second-generation platforms.

Advantages and disadvantages of different approaches

In terms of costs, limitations and practical aspects of implementation, clear differences between conventional sequencing and the second-generation platforms determine which general strategy represents the best option for any given project. The applications of conventional sequencing (that is, Sanger) have grown diverse, and for small-scale projects in the kilobase-to-megabase range, this will likely remain the technology of choice for the immediate future. This is a consequence of its greater 'granularity' (that is, the ability to efficiently operate at either small or large production scales) relative to the new technologies. Even so, it is clear that despite limitations relative to Sanger sequencing (e.g., in terms of read-length and accuracy), large-scale projects will quickly come to depend entirely on next-generation sequencing. As an example of the advantages of the new platforms, consider that large-scale resequencing studies for identifying germline variation or somatic mutations have relied on Sanger-based resequencing approaches, that in turn are reliant on one-at-a-time PCR amplification of each targeted region[31,32]. In this context, the requirements of a Sanger sequencing approach include major costs beyond just reagents. These include robotic support of reagents, processing of multiple samples in 96- or 384-well formats, maintenance of capillary-based sequencers, extensive bioinformatics infrastructure to handle the flow of data and dedicated support staff to maintain complicated equipment. In a recent informal survey we conducted of the overall cost to conventionally sequence 100 genes from 100 samples, assuming each gene has an average of 10 exons, quoted estimates from noncommercial genome centers and commercial sequence service providers ranged from $300,000 to over $1,000,000. Clearly, this cost is beyond the range of most individual laboratories. In addition to reducing the per-base cost of sequencing by several orders of magnitude, second-generation instruments have fewer infrastructure requirements; instead, the principle challenge is downstream data management.

There are important differences among the second-generation platforms themselves (**Table 1**) that may result in advantages with respect to specific applications (**Table 2**). Some applications (e.g., resequencing) may be more tolerant of short read-lengths than others (e.g., *de novo* assembly). For applications relying on tag counting (e.g., quantification of protein-DNA interactions), one would actually prefer a given amount of sequencing to be split into as many reads as possible (above some minimum length that allows placement to a reference). The overall accuracy as well as the specific error distributions of individual technologies (e.g., the rate of insertion-deletion versus substitution errors; the propensity for systematic consensus errors) may also be highly relevant. Mate-paired reads, useful in *de novo* assembly and for mapping structural variants, for example, are now available with all of the second-generation platforms, but the extent to which the distance distribution with which the read pairs are separated can be controlled or varied may be an important factor. Finally, of course, the cost of sequencing varies greatly between the second-generation platforms, and as consumers, we hope for more competition between vendors than was the case with conventional sequencing in the past decade. Comparisons of 'per-base' costs can be helpful but occasionally misleading, as, for example, more accurate bases may be worth more than less accurate bases.

Software and standards for next-generation sequencing data

The diversity and rapid evolution of next-generation sequencing technology is posing challenges for bioinformatics in areas including sequence quality scoring, alignment, assembly and data release. These are discussed in more detail below.

Sequence quality. The topic of sequence quality scoring has become an area of intense interest, given the relatively low quality of raw data from the new sequencing platforms, and the various context-dependent error distributions associated with different sequencing by synthesis biochemistries. As second-generation sequencing matures and the range of biological and clinical problems to which it is applied expands, it will be critical to have clear metrics in place for data quality, reliability, reproducibility and biological relevance. Given the nascent state of the field, there is an opportunity to establish an early consensus of standardized benchmarks for comparing current and newly introduced platforms, and, it is hoped, avoid the dilemmas of data comparison that have occurred in the past with multi-vendor genomic technologies, such as microarrays. Multiple applications will benefit from standardized quality metrics. This might include metrics to quantify the general quality of *de novo* sequence assemblies; metrics for the confidence associated with individual read alignments to a reference; metrics for confidence in raw and consensus base-calls for improved polymorphism and mutation discovery; and general quality control and assurance metrics for large-scale sequencing projects.

Examples of areas that should be systematically evaluated in assess-

Table 2 Applications of next-generation sequencing

Category	Examples of applications	Refs
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes	44
Reduced representation sequencing	Large-scale polymorphism discovery	45
Targeted genomic resequencing	Targeted polymorphism and mutation discovery	46–52
Paired end sequencing	Discovery of inherited and acquired structural variation	53,54
Metagenomic sequencing	Discovery of infectious and commensal flora	55
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations	56–63
Small RNA sequencing	microRNA profiling	64
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA	60,65,66
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions	67–70
Nuclease fragmentation and sequencing	Nucleosome positioning	69
Molecular barcoding	Multiplex sequencing of samples from multiple individuals	61,71

ing data quality with new sequencing platforms include: (i) technical reproducibility; (ii) distribution of estimated accuracies for raw base-calls (e.g., *phred*-like scores); (iii) systematic error patterns in raw or consensus sequence data; (iv) bias and skewing of true ratios in tag-counting applications. Quality data should be included by default along with sequences, ideally in a simple, standardized format. All of the current vendors are adopting quality metrics in their data reporting, and consumers of these technologies need to be able to make cross-comparisons among data sets, as well as to potentially integrate data from multiple platforms while exploiting their individual advantages (e.g., mixed-technology genome assemblies). Current resequencing approaches for polymorphism and mutation discovery from short-read data sets sometimes rely on majority voting schemes, which are easy to implement but prone to error and false discovery. More appropriate schemes for single-nucleotide polymorphism (SNP) detection, such as those developed by Brockman *et al.*[33] and Quinlan *et al.*[34], estimate the error probabilities associated with individual base-calls (based on data quality and context) and use this information in making consensus calls. We anticipate that much as *phred* and related tools did for Sanger sequencing, the development of 'third party' algorithms will substantially improve base-calling and polymorphism detection with all of the new platforms.

Software and bioinformatics tools for data analysis. Even at this early stage of commercial availability, a variety of software tools are available for analyzing next-generation sequencing data (**Table 3**). Their functions fit into several general categories, including: (i) alignment of sequence reads to a reference; (ii) base-calling and/or polymorphism detection; (iii) *de novo* assembly, from paired or unpaired reads; and (iv) genome browsing and annotation.

Alignment and assembly represent particularly interesting problems. Whereas alignment solutions like BLAST or BLAT are largely adequate for long reads such as those generated by conventional sequencing, these are unlikely to be the best algorithms for handling short-read sequence data. An increasing number of alignment tools have been developed specifically for rapid alignment of large sets of short reads, while allowing for mismatches and/or gaps. Some of these tools take advantage of well-established alignment algorithms, such as Smith-Waterman, but there also has been significant innovation in developing new algorithms specifically tailored for short reads. For example, SOAP, a software package for efficient gapped or ungapped alignment, uses a memory-intensive seed and look-up table algorithm to accelerate alignment, while allowing iterative trimming of the 3′ end of reads (usually associated with a higher error rate)[35]. Other approaches used to accelerate processing include 'bit encoding' to compress sequence data into a computationally more manageable and efficient format[35,36]. Alignment software is increasingly taking into account the estimated quality of the underlying data in generating read-placements, as is the case with MAQ[37], an alignment and variation discovery tool that works with data from either Solexa or SOLiD data, and SHRiMP (http://compbio.cs.toronto.edu/shrimp/), which includes a novel "color-space to letter-space" Smith-Waterman algorithm compatible with two base–encoded sequence data from the SOLiD platform.

As with alignment algorithms, the short read-lengths, relatively lower accuracies and quantity of data associated with the new technologies make *de novo* assembly into a challenging problem once again. Several assembly tools have recently been adapted or independently developed for generating assemblies from short, unpaired sequencing reads[38–40]. Mate-paired reads, now possible with all of the major platforms, are anticipated to have a major impact on the overall success of *de novo* assembly with short reads, and several algorithms have been already developed that take advantage of these[38–41].

Table 3 Bioinformatics tools for short-read sequencing

Program	Categories	Author(s)	Reference	URL
Cross_match	Alignment	Phil Green, Brent Ewing and David Gordon		http://www.phrap.org/phredphrapconsed.html
ELAND	Alignment	Anthony J. Cox		http://www.illumina.com/
Exonerate	Alignment	Guy S. Slater and Ewan Birney	72	http://www.ebi.ac.uk/~guy/exonerate
MAQ	Alignment and variant detection	Heng Li	37	http://maq.sourceforge.net
Mosaik	Alignment	Michael Strömberg and Gabor Marth		http://bioinformatics.bc.edu/marthlab/Mosaik
RMAP	Alignment	Andrew Smith, Zhenyu Xuan and Michael Zhang	73	http://rulai.cshl.edu/rmap
SHRiMP	Alignment	Michael Brudno and Stephen Rumble		http://compbio.cs.toronto.edu/shrimp
SOAP	Alignment	Ruiqiang Li *et al.*	35	http://soap.genomics.org.cn
SSAHA2	Alignment	Zemin Ning *et al.*	36	http://www.sanger.ac.uk/Software/analysis/SSAHA2
SXOligoSearch	Alignment	Synamatix		http://synasite.mgrc.com.my:8080/sxog/NewSXOligoSearch.php
ALLPATHS	Assembly	Jonathan Butler *et al.*	38	
Edena	Assembly	David Hernandez *et al.*	74	http://www.genomic.ch/edena
Euler-SR	Assembly	Mark Chaisson and Pavel Pevzner	75	
SHARCGS	Assembly	Juliane Dohm *et al.*	76	http://sharcgs.molgen.mpg.de
SHRAP	Assembly	Andreas Sundquist *et al.*	39	
SSAKE	Assembly	René Warren *et al.*	40	http://www.bcgsc.ca/platform/bioinfo/software/ssake
VCAKE	Assembly	William Jeck	77	http://sourceforge.net/projects/vcake
Velvet	Assembly	Daniel Zerbino and Ewan Birney	41	http://www.ebi.ac.uk/%7Ezerbino/velvet
PyroBayes	Base caller	Aaron Quinlan *et al.*	34	http://bioinformatics.bc.edu/marthlab/PyroBayes
PbShort	Variant detection	Gabor Marth		http://bioinformatics.bc.edu/marthlab/PbShort
ssahaSNP	Variant detection	Zemin Ning *et al.*		http://www.sanger.ac.uk/Software/analysis/ssahaSNP

Incomplete list compiled from sources, including http://seqanswers.com/forums/showthread.php?t=43 and http://www.sanger.ac.uk/Users/lh3/seq-nt.html.

Guidelines for publication, release and archiving. There are active ongoing discussions regarding the establishment of consensus guidelines for reporting and archiving short-read sequence data, similar to the MIAME (Minimum Information About a Microarray Experiment) guidelines for microarray experiments[42] (http://uhts.lbl.gov/). These early efforts have proposed that metadata include annotation of the biological or clinical nature of the data, key experimental details such as sample properties and treatments, specific format styles for sequence reads among multiple samples, policies for release and publication of sequence data and data processing methods. Public archiving of next-generation sequence data is another area that will require substantial effort given the sheer quantity of data being generated. The National Center for Biotechnology Information at the National Institutes of Health (Bethesda, MD) has recently established a Short Read Archive (SRA), and is defining submission formats for data sets corresponding to major next-generation sequencing platforms[43]. The SRA is designed to integrate with the Entrez system and to include metadata, such as experimental annotation and parameters from actual instrument runs. Current efforts include the development of online tools for efficiently searching the archive and for data visualization; these are expected to appear in the coming year[43].

Box 4 Targeted capture of genomic subsets

For genomic resequencing (that is, sequencing for somatic or germline variation discovery in individual(s) of a species for which a reference genome is available), it is frequently the case that investigators would prefer to use finite resources to sequence a specific subset of the genome across more individuals, rather than the whole genome of fewer individuals. Examples of genomic subsets that may be highly relevant include: (i) a specific megabase-scale region of the genome to which a disease phenotype has been mapped; (ii) exons of specific candidate genes belonging to a disease-related pathway; (iii) the full complement of protein-coding DNA sequences. These subsets generally total to megabases, raising the question of how they can be efficiently isolated barring hundreds or thousands of individual PCR reactions. In other words, analogous to how PCR served as an effective 'front-end' for resequencing of kilobase-sized targets with capillary electrophoresis, there is a strong need for flexible targeting methods that are matched to the megabase-scale granularity at which the next-generation sequencing platforms operate. Fortunately, a variety of such methods have shown convincing proof-of-concept demonstrations in the past several years. These include methods that, like PCR, rely on a combination of oligonucleotide hybridization and enzymatic activity (e.g., polymerase or ligase) to confer specificity, but unlike PCR, are more compatible with high degrees of multiplexing. For example, Ji and colleagues[52] described the multiplex capture of 177 exons by selective circularization of restriction fragments; Fredriksson, *et al.*[53] (Ji is a coauthor) described a modified version of multiplex PCR that allowed single-reaction amplification of 170 exons; and Shendure and colleagues[54] demonstrated that a complex mixture of molecular inversion probes obtained by parallel synthesis on and release from the surface of a programmable microarray could be used to capture ~10,000 exons in a single aqueous-phase reaction (>50,000 with further optimizations; J.S., unpublished observations). Another approach is capture by hybridization. Bashiardes *et al.*[55] demonstrated 10,000-fold hybridization-based enrichment of sequences derived from BAC (bacterial artificial chromosome)-sized genomic regions. More recently, using NimbleGen microarrays, three groups[56–58] carried out capture by hybridization of megabase-scale genomic regions or tens-of-thousands of exons per array. It is likely that targeted capture methods such as these will play an increasingly important role in driving resequencing applications of next-generation platforms.

Applications of next-generation sequencing

The past several years have seen an accelerating flurry of publications in which next-generation sequencing is applied for a variety of goals (**Table 2**). Important applications include: (i) full-genome resequencing or more targeted discovery of mutations or polymorphisms (**Box 4**); (ii) mapping of structural rearrangements, which may include copy number variation, balanced translocation breakpoints and chromosomal inversions; (iii) 'RNA-Seq', analogous to expressed sequence tags (EST) or serial analysis of gene expression (SAGE), where shotgun libraries derived from mRNA or small RNAs are deeply sequenced; the counts corresponding to individual species can be used for quantification over a broad dynamic range, and the sequences themselves can be used for annotation (e.g., splice junctions and transcript boundaries); (iv) large-scale analysis of DNA methylation, by deep sequencing of bisulfite-treated DNA; (v) 'ChIP-Seq', or genome-wide mapping of DNA-protein interactions, by deep sequencing of DNA fragments pulled down by chromatin immunoprecipitation. Over the next few years, the list of applications will undoubtedly grow, as will the sophistication with which existing applications are carried out.

Conclusions

Over the past several years, next-generation DNA sequencing technologies have catapulted to prominence, with increasingly widespread adoption of several platforms that individually implement different flavors of massively parallel cyclic-array sequencing. Common characteristics extend beyond the technologies themselves, to the quantity and quality of data that are generated, such that all raise a similar set of new challenges for experimental design, data analysis and interpretation. The reduction in the costs of DNA sequencing by several orders of magnitude is democratizing the extent to which individual investigators can pursue projects at a scale previously accessible only to major genome centers. The dramatic increase in interest in this area is also evident in the number of groups that are now working on sequencing methods to supplant even the new technologies discussed here. Given the current state of flux, it is difficult to peer even a few years into the future, but we anticipate that next-generation sequencing technologies will become as widespread, commoditized and routine as microarray technology has become over the past decade. Also analogous to microarrays, we also expect that the challenges will quickly shift from mastering of the technologies themselves to the question of how best to go about extracting biologically meaningful or clinically useful insights from a very large amount of data.

ACKNOWLEDGMENTS

We thank G. Church and G. Porreca for helpful comments on the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Hutchison, C.A., III. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res.* **35**, 6227–6237 (2007).

2. Sanger, F. Sequences, sequences, and sequences. *Annu. Rev. Biochem.* **57**, 1–28 (1988).
3. Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687–695 (1977).
4. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* **5**, 335–344 (2004).
5. Swerdlow, H., Wu, S.L., Harke, H. & Dovichi, N.J. Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *J. Chromatogr.* **516**, 61–67 (1990).
6. Hunkapiller, T., Kaiser, R.J., Koop, B.F. & Hood, L. Large-scale and automated DNA sequence determination. *Science* **254**, 59–67 (1991).
7. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
8. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
9. Blazej, R.G., Kumaresan, P. & Mathies, R.A. Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing. *Proc. Natl. Acad. Sci. USA* **103**, 7240–7245 (2006).
10. Gresham, D., Dunham, M.J. & Botstein, D. Comparing whole genomes using DNA microarrays. *Nat. Rev. Genet.* **9**, 291–302 (2008).
11. Soni, G.V. & Meller, A. Progress toward ultrafast DNA sequencing using solid-state nanopores. *Clin. Chem.* **53**, 1996–2001 (2007).
12. Healy, K. Nanopore-based single-molecule DNA analysis. *Nanomed.* **2**, 459–481 (2007).
13. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
14. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
15. Mitra, R.D. & Church, G.M. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res.* **27**, e34 (1999).
16. Mitra, R.D., Shendure, J., Olejnik, J., Edyta Krzymanska, O. & Church, G.M. Fluorescent in situ sequencing on polymerase colonies. *Anal. Biochem.* **320**, 55–65 (2003).
17. Bentley, D.R. Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* **16**, 545–552 (2006).
18. Harris, T.D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106–109 (2008).
19. Ng, P. *et al.* Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods* **2**, 105–111 (2005).
20. Dressman, D., Yan, H., Traverso, G., Kinzler, K.W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. USA* **100**, 8817–8822 (2003).
21. Adessi, C. *et al.* Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.* **28**, e87 (2000).
22. Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.* **34**, e22 (2006).
23. Turcatti, G., Romieu, A., Fedurco, M. & Tairi, A.P. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res.* **36**, e25 (2008).
24. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000).
25. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. & Nyren, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**, 84–89 (1996).
26. McKernan, K., Blanchard, A., Kotler, L. & Costa, G. Reagents, methods, and libraries for bead-based sequencing. US patent application 20080003571 (2006).
27. Housby, J.N. & Southern, E.M. Fidelity of DNA ligation: a novel experimental approach based on the polymerisation of libraries of oligonucleotides. *Nucleic Acids Res.* **26**, 4259–4266 (1998).
28. Macevicz, S.C. DNA sequencing by parallel oligonucleotide extensions. US patent 5750341 (1998).
29. Barbee, K.D. & Huang, X. Magnetic assembly of high-density DNA arrays for genomic analyses. *Anal. Chem.* **80**, 2149–2154 (2008).
30. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S.R. Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. USA* **100**, 3960–3964 (2003).
31. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
32. Wood, L.D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113 (2007).
33. Brockman, W. *et al.* Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.* **18**, 763–770 (2008).
34. Quinlan, A.R., Stewart, D.A., Stromberg, M.P. & Marth, G.T. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat. Methods* **5**, 179–181 (2008).
35. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).
36. Ning, Z., Cox, A.J. & Mullikin, J.C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
37. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* published online, doi:10.1101/gr.078212.108 (19 August 2008).
38. Butler, J. *et al.* ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res.* **18**, 810–820 (2008).
39. Sundquist, A., Ronaghi, M., Tang, H., Pevzner, P. & Batzoglou, S. Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS ONE* **2**, e484 (2007).
40. Warren, R.L., Sutton, G.G., Jones, S.J. & Holt, R.A. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23**, 500–501 (2007).
41. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
42. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).
43. Wheeler, D.L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **36**, D13–D21 (2008).
44. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
45. Van Tassell, C.P. *et al.* SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* **5**, 247–252 (2008).
46. Dahl, F. *et al.* Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc. Natl. Acad. Sci. USA* **104**, 9387–9392 (2007).
47. Fredriksson, S. *et al.* Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. *Nucleic Acids Res.* **35**, e47(2007).
48. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
49. Bashiardes, S. *et al.* Direct genomic selection. *Nat. Methods* **2**, 63–69 (2005).
50. Albert, T.J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4**, 903–905 (2007).
51. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* **39**, 1522–1527 (2007).
52. Okou, D.T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* **4**, 907–909 (2007).
53. Campbell, P.J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–729 (2008).
54. Chen, W. *et al.* Mapping translocation breakpoints by next-generation sequencing. *Genome Res.* **18**, 1143–1149 (2008).
55. Cox-Foster, D.L. *et al.* A metagenomic survey of microbes in honey bee colony collapse disorder. *Science* **318**, 283–287 (2007).
56. Wilhelm, B.T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239–1243 (2008).
57. Sugarbaker, D.J. *et al.* Transcriptome sequencing of malignant pleural mesothelioma tumors. *Proc. Natl. Acad. Sci. USA* **105**, 3521–3526 (2008).
58. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
59. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
60. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
61. Kim, J.B. *et al.* Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1481–1484 (2007).
62. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).
63. Bainbridge, M.N. *et al.* Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7**, 246 (2006).
64. Morin, R.D. *et al.* Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.* **18**, 610–621 (2008).
65. Korshunova, Y. *et al.* Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. *Genome Res.* **18**, 19–29 (2008).
66. Ordway, J.M. *et al.* Identification of novel high-frequency DNA methylation changes in breast cancer. *PLoS ONE* **2**, e1314 (2007).
67. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
68. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
69. Schones, D.E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–898 (2008).
70. Wold, B. & Myers, R.M. Sequence census methods for functional genomics. *Nat. Methods* **5**, 19–21 (2008).
71. Meyer, M., Stenzel, U. & Hofreiter, M. Parallel tagged sequencing on the 454 platform. *Nat. Protocols* **3**, 267–278 (2008).
72. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
73. Smith, A.D., Xuan, Z. & Zhang, M.Q. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* **9**, 128 (2008).
74. Hernandez, D., Francois, P., Farinelli, L., Osteras, M. & Schrenzel, J. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Res.* **18**, 802–809 (2008).
75. Chaisson, M.J. & Pevzner, P.A. Short read fragment assembly of bacterial genomes. *Genome Res.* **18**, 324–330 (2008).
76. Dohm, J.C., Lottaz, C., Borodina, T. & Himmelbauer, H. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res.* **17**, 1697–1706 (2007).
77. Jeck, W.R. *et al.* Links extending assembly of short DNA sequences to handle error. *Bioinformatics* **23**, 2942–2944 (2007).

78. Paegel, B.M., Blazej, R.G. & Mathies, R.A. Microfluidic devices for DNA sequencing: sample preparation and electrophoretic analysis. *Curr. Opin. Biotechnol.* **14**, 42–50 (2003).
79. Hong, J.W. & Quake, S.R. Integrated nanoliter systems. *Nat. Biotechnol.* **21**, 1179–1183 (2003).
80. Emrich, C.A., Tian, H., Medintz, I.L. & Mathies, R.A. Microfabricated 384-lane capillary array electrophoresis bioanalyzer for ultrahigh-throughput genetic analysis. *Anal. Chem.* **74**, 5076–5083 (2002).
81. Toriello, N.M., Liu, C.N., Blazej, R.G., Thaitrong, N. & Mathies, R.A. Integrated affinity capture, purification, and capillary electrophoresis microdevice for quantitative double-stranded DNA analysis. *Anal. Chem.* **79**, 8549–8556 (2007).
82. Blazej, R.G., Kumaresan, P., Cronier, S.A. & Mathies, R.A. Inline injection microdevice for attomole-scale sanger DNA sequencing. *Anal. Chem.* **79**, 4499–4506 (2007).
83. Hong, J.W., Studer, V., Hang, G., Anderson, W.F. & Quake, S.R. A nanoliter-scale nucleic acid processor with parallel architecture. *Nat. Biotechnol.* **22**, 435–439 (2004).
84. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
85. Frazer, K.A. *et al.* A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448**, 1050–1053 (2007).
86. Gresham, D. *et al.* Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. *Science* **311**, 1932–1936 (2006).
87. Albert, T.J. *et al.* Mutation discovery in bacterial genomes: metronidazole resistance in *Helicobacter pylori*. *Nat. Methods* **2**, 951–953 (2005).
88. Drmanac, S. *et al.* Accurate sequencing by hybridization for DNA diagnostics and individual genomics. *Nat. Biotechnol.* **16**, 54–58 (1998).
89. Drmanac, R., Labat, I., Brukner, I. & Crkvenjakov, R. Sequencing of megabase plus DNA by hybridization: theory of the method. *Genomics* **4**, 114–128 (1989).
90. Lizardi, P.M. *et al.* Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat. Genet.* **19**, 225–232 (1998).
91. Pihlak, A. *et al.* Rapid genome sequencing with short universal tiling probes. *Nat. Biotechnol.* **26**, 676–684 (2008).
92. Deamer, D.W. & Akeson, M. Nanopores and nucleic acids: prospects for ultrarapid sequencing. *Trends Biotechnol.* **18**, 147–151 (2000).
93. Meller, A., Nivon, L., Brandin, E., Golovchenko, J. & Branton, D. Rapid nanopore discrimination between single polynucleotide molecules. *Proc. Natl. Acad. Sci. USA* **97**, 1079–1084 (2000).
94. Cockroft, S.L., Chu, J., Amorin, M. & Ghadiri, M.R. A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution. *J. Am. Chem. Soc.* **130**, 818–820 (2008).
95. Levene, M.J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682–686 (2003).
96. Lundquist, P.M. *et al.* Parallel confocal detection of single molecules in real time. *Opt. Lett.* **33**, 1026–1028 (2008).
97. Korlach, J. *et al.* Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc. Natl. Acad. Sci. USA* **105**, 1176–1181 (2008).