Current and Future Outlook of Genomic Technologies

Jeffrey Fisher and Mostafa Ronaghi

Overview:

Genomics has emerged as a scientific field since the invention of DNA sequencing technique back in 1977 by Fredrick Sanger.[16, 17] Back then Sanger introduced a chemical method reading about 100 nucleotides which would take around six months of preparation. Eventually, Sanger technique did evolve thanks to a large community of scientists to become the technology of choice for draft sequencing of the first human genome, which took around 13 years to complete in a project funded for three billion dollars. Pyrosequencing [14, 15] was the second alternative technology that was introduced and could be parallelized enabling 100 folds higher throughput. Pyrosequencing was used to sequence thousands of microbial and larger genomes. Reversible dye-terminator sequencing-by-synthesis[3] was introduced to the market by Illumina in 2006. This technology has increased the throughput by ~10,000 folds in the last four years reducing the cost below $10,000 for full human genome sequencing. The most recent system based on this chemistry allows sequencing of several human genomes at 30X coverage in a single run. Here, this technology will be discussed in details.

Background

At its most fundamental level, sequencing the genome consists of just a handful of basic biochemical steps. What presents the challenge is the enormous scale of molecularly-encoded information that must processed through those steps: two almost identical strands each consisting of 3.2 billion base pairs (Gbp) of information for the human genome. Furthermore, a typical genome is read to 30x coverage, which means that each base pair is read on average 30 times (on separate strands of DNA), giving a total throughput per genome of 100 billion base pairs. The processing and readout of such immense amounts of information has been made possible by the adoption of engineering-based approached to achieve massive parallelization of the sequencing reactions. All current-generation sequencing platforms coordinate Chemical, Engineering, and Computation subsystems on an unprecedented scale (measured in information throughput).

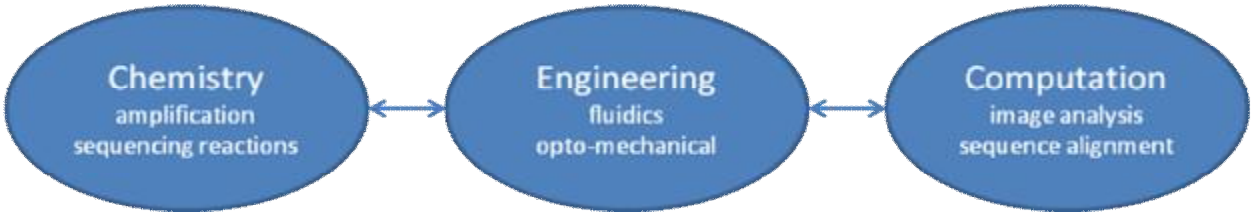


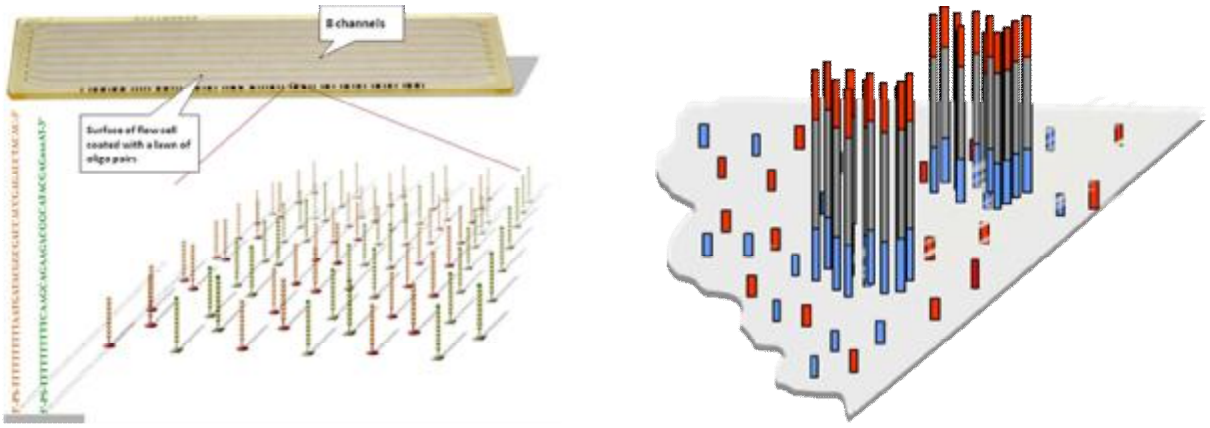
Figure 1. Modern sequencing requires highly coordinated subsystems to handle the massive throughput of information.

DNA Sequencing

Sequencing, which determines the arrangement of the four genetic bases (A, T, C, and G) in a given stretch of DNA, relies on the four steps below. [12, 13, 19]

1. Fragment: break the genome into manageable segments, usually a few hundred base-long.
2. Isolate: the segments will be captured separated from one another to present distinct signals.
3. Amplify: although single-molecule techniques can theoretically proceed without this step, most systems apply some form of clonal amplification to increase the signal and accuracy.
4. Readout: the genetic information is transformed base-by-base into a machine-readable form, typically an optical (fluorescent) signal.

While the field has evolved in recent years to include a diverse array of systems, including ones that do not necessarily follow this exact pattern, the majority of commercial platforms utilize all four steps in one form or another.

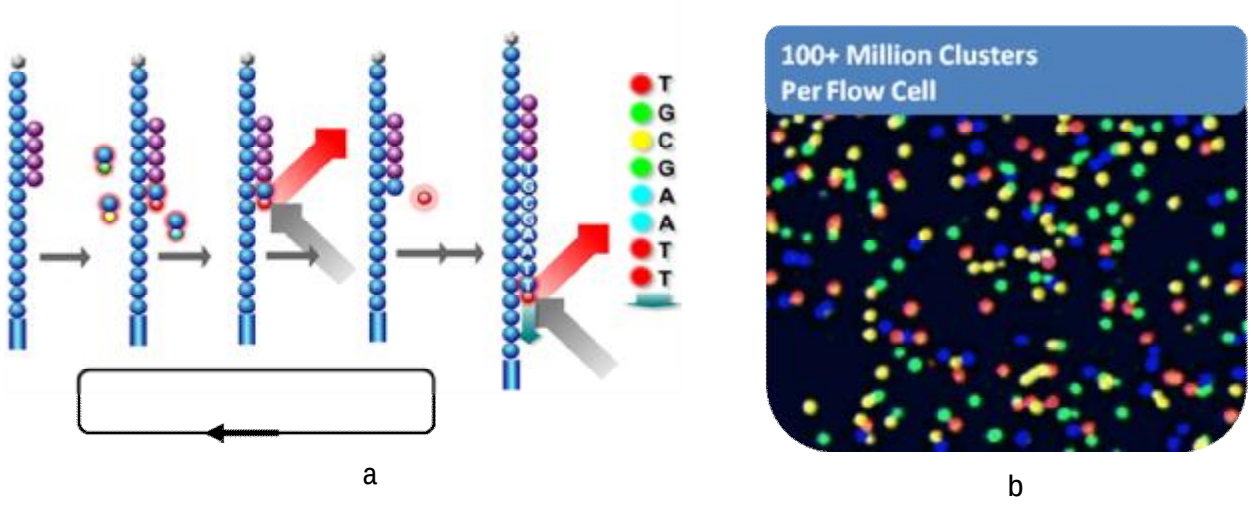


a b

Figure 2. Adapter-ligated segments of DNA are loaded into a flow cell which is coated with a lawn of oligos to capture and bind the DNA segments (a). Once attached, the DNA is PCR amplified in-place to form clonal clusters (b).

The Illumina Genome Analyzer and Hi-Seq systems are examples of the massively parallel nature of this type of biochemical workflow.[2, 3] First a sample of DNA is fragmented into segments ~400 bp long, and oligonucleotides of known sequence are ligated to the ends. These ligated adapters function as "handles" for the segment, allowing it to be manipulated in downstream reactions. For example, they provide a means to trap the DNA segment in the flow cell and later release it, and areas for primers to bind for the sequencing reaction. Next, the sample is injected into a flow cell containing a lawn of oligonucleotides which will bind to the adapters on the DNA segments (Figure 2a). The concentration is carefully controlled so that only one strand is present in a given area of the chip—representing the

signal isolation step. Then, the segment is amplified in place by means of a substrate-bound PCR process (Bridge PCR), until each single segment has grown into a cluster of thousands of identical copies of the sequence (Figure 2b). A single flow cell will contain approximately 100 million individual clusters. Although they have grown larger than the initial single strand, the clusters remain immobile and physically separated from one another, such that they can be visually distinguished during the readout step. The genetic sequence is then transformed into a visual signal by synthesizing a complementary strand one base at a time using nucleotides with four separate color tags (Figure 3a). For each cycle (during which a single base will be read per cluster), DNA polymerase will incorporate a single nucleotide matching the next base on the template sequence. Only one nucleotide is incorporated because in addition to fluorescent tags, each nucleotide possesses a terminator group which blocks subsequent nucleotides from being added. The entire flow cell is then imaged, and the color of each cluster indicates which base was added for that sequence (Figure 3b). Finally, the terminator group and fluorophore are cleaved and the cycle is repeated.



a b

Figure 3. Sequencing-by-Synthesis. During each round fluorescently-labeled nucleotides (one color for each of A, T, C and G) are added to the flow cell (a). The nucleotide matching the next open position is incorporated. The entire flow cell is then imaged, producing a map (b) showing exactly which nucleotide was incorporated at each cluster. The label and terminator are then cleaved to allow the reaction to start again. This process is repeated hundreds of times to read out a single contiguous sequence.

This process is repeated until each cluster has been read 100-150 times; then the segment can be "flipped over" and another 100-150 bases of sequence information can be read from the other end. Therefore, the total amount of information that can be garnered from a single flow cell is directly proportional to the number of clusters and the read length per cluster, both of which represent targets for improvement as we continually increase system throughput.

Moore's Law and Genomics

The often quoted Moore's law posits that the number of transistors on an integrated circuit will double every 18-24 months, consequently reducing the cost per transistor (Figure 4). Sequencing costs have demonstrated a similar exponential decrease over time, accelerating at even a faster rate.

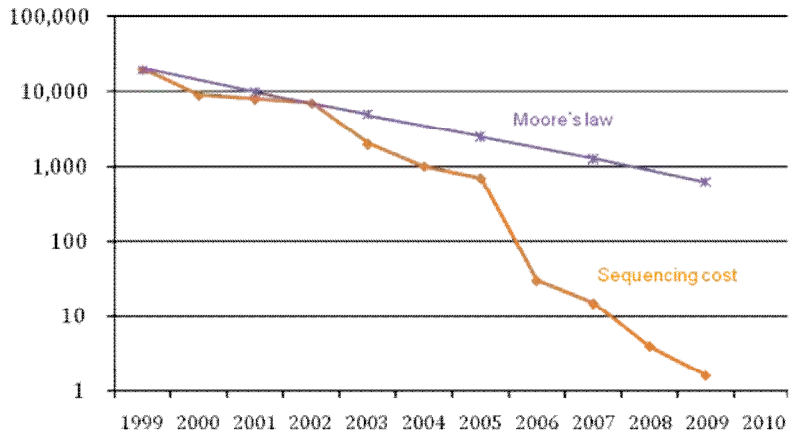


Figure 4. As the throughput of sequencing systems increases exponentially, the costs drop accordingly, out pacing the rate of Moore's law.

One factor that has made this possible is that while transistor density is limited to increases in the 2-D area of the chip, sequencing has a "third dimension" which is the read length. Therefore, each of the subsystems in Figure 1 can be improved in parallel to increase the total throughput of the system. Improvements to the chemistry have resulted in longer read lengths and shorter cycle times. The total cluster number has grown both by increasing the flow cell area and the cluster density. The engineering subsystem doubled the throughput by using both the top and bottom of the flow cell for cluster growth. Cluster density has been increased by improving the optics and the algorithms that detect the clusters. Total run time is regularly decreased by using faster chemistries, faster optical scanning, and faster algorithms. On the one hand, improvements in each subsystem independently contribute to increases in throughput, but on the other hand, improvements in one system are often the leading driver to advances in the others.

Frontiers in Genomics

The way forward is exciting because it lies in improving the technology such that it can be adapted to an ever broader range of applications. Two ways to achieve this are 1) increasing the sensitivity of the system such that it can more robustly handle lower signal-to-noise ratios, and 2) increasing the throughput to drive down costs.

Using the methods described above, one can sequence an entire human genome starting with as little as 1 ug of DNA. However, there are types of samples for which even this relatively modest amount is difficult to come by. For example, many researchers are beginning to look at the genomics of single

cells—and not just one single cell, but processing small populations individually to evaluate the heterogeneity across the group.[7, 21, 23] However, since a cell contains only about 6 pg of genomic DNA and 10 pg of RNA, the corresponding signal is many orders of magnitude less than normal. Another example of sample types that would benefit from improved assay sensitivity are formalin-fixed, paraffin embedded (FFPE) samples. FFPEs are histological tissue samples that have been chemically fixed to reduce degeneration so that they can be stained and examined under a microscope. They are important because there are huge archives of the samples for which detailed patient outcomes are already known. However, the fixing, staining, and storage conditions break down the genetic material making sequencing or genotyping much more difficult. The ability to go back to these samples and perform genomic analysis represents an invaluable resource for tracking down the genetic contributions to disease and wellness.[4, 9, 18, 25]

There are also cases where the signal itself is not the problem; rather, the background possesses a much higher degree of noise. For example, there is a lot of interest recently in studying the microbiome of different environments, whether that is soil or seawater, or the human gut.[5, 22, 24] In cases such as these, the genetic diversity of the sample can make it difficult to separate the components of different organisms. Likewise, genomics plays a vital role in the study of cancer,[1, 6, 20] which is defined by its genetic instability and pathology.[8, 10, 11] However, this means that cells taken even from the same tumor can possess extreme genetic heterogeneity making increased sensitivity and detection key to distinguishing the often subtle differences leading to one outcome versus another. This requires a much deeper read: 7200x coverage rather than the typical 30x coverage needed for a homogenous sample.

Increases in throughput will not only have an impact in terms of the quantity of genetic information available, but the resultant decrease in cost will open up completely new markets, representing a qualitative shift in the ways in which genomics impacts our daily lives. Once the cost of sequencing an entire genome reaches what it now costs to analyze a single gene, a watershed moment will occur in the market, releasing a flood of applications for sequencing. Diagnosis, prognosis, pharmacogenomics, drug development, agriculture—all will be changed in a fundamental way.

Challenges

Of course, achieving the improvements described above will require overcoming technical obstacles, but the most significant challenges facing genomics are non-technical in nature. Like many information-based fields, the pace of innovation is out-stripping the rate at which legislation and regulation can keep up. Laws designed prior to the genomic revolution are being shoe-horned to fit technologies and situations for which there is no clear precedent. There is a definite need to more clearly define the regulatory landscape, so that companies can move forward with confidence leveraging these innovations to improve people's lives. Simultaneously, we must work to raise the public's awareness of the science and the technology, dispelling myths and fostering an understanding of its importance to the health of both as individuals and society as a whole.

Summary

Genomics has emerged as an important tool for the studies of biological systems today. Significant cost reduction has accelerated adaptation of this technology in new fields; Including, research, diagnostics, consumer, agrigenomics as well as forensics. We predict that the cost will continue to drop for the next few years. Increased throughput per day is the most important factors in reducing cost, which will be achieved through increased density, increased read-length and shorter cycle time.

References

[1] A. Balmain, J. Gray, and B. Ponder. The genetics and genomics of cancer. NATURE GENETICS, 33:238–244, Mar 2003.

[2] D. R. Bentley. Whole-genome re-sequencing. Curr Opin Genet Dev, 16(6):545–52, Dec 2006.

[3] D. R. Bentley et. al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 456(7218):53–9, Nov 2008.

[4] M. Bibikova, D. Talantov, E. Chudin, J. M. Yeakley, J. Chen, D. Doucet, E. Wickham, D. Atkins, D. Barker, M. Chee, Y. Wang, and J.-B. Fan. Quantitative gene expression profiling in formalin-fixed, paraffin-embedded tissues using universal bead arrays. Am J Pathol, 165(5):1799–807, Nov 2004.

[5] S. R. Gill, M. Pop, R. T. DeBoy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson. Metagenomic analysis of the human distal gut microbiome. SCIENCE, 312(5778):1355–1359, Jun 2006.

[6] P. A. Jones and S. B. Baylin. The fundamental role of epigenetic events in cancer. NATURE REVIEWS GENETICS, 3(6):415–428, Jun 2002.

[7] K. Kurimoto and M. Saitou. Single-cell cdna microarray profiling of complex biological processes of differentiation. Curr Opin Genet Dev, Jul 2010.

[8] C. Lengauer, K. W. Kinzler, and B. Vogelstein. Genetic instabilities in human cancers. Nature, 396(6712):643–9, Dec 1998.

[9] F. Lewis, N. J. Maughan, V. Smith, K. Hillan, and P. Quirke. Unlocking the archive–gene expression in paraffin-embedded tissue. J Pathol, 195(1):66–71, Sep 2001.

[10] L. A. Loeb. Mutator phenotype may be required for multistage carcinogenesis. Cancer Res, 51(12):3075–9, Jun 1991.

[11] L. A. Loeb. A mutator phenotype in cancer. Cancer Res, 61(8):3230–9, Apr 2001.

[12] M. L. Metzker. Sequencing technologies - the next generation. NATURE REVIEWS GENETICS, 11(1):31–46, Jan 2010.

[13] E. Pettersson, J. Lundeberg, and A. Ahmadian. Generations of sequencing technologies. Genomics, 93(2):105–11, Feb 2009.

[14] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén. Real-time dna sequencing using detection of pyrophosphate release. Anal Biochem, 242(1):84–9, Nov 1996.

[15] M. Ronaghi, M. Uhlén, and P. Nyrén. A sequencing method based on real-time pyrophosphate. Science, 281(5375):363, 365, Jul 1998.

[16] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phi x174 dna. Nature, 265(5596):687–95, Feb 1977.

[17] F. Sanger, S. Nicklen, and A. R. Coulson. Dna sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A, 74(12):5463–7, Dec 1977.

[18] M. R. Schweiger, M. Kerick, B. Timmermann, M. W. Albrecht, T. Borodina, D. Parkhomchuk, K. Zatloukal, and H. Lehrach. Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (ffpe) tumor tissues for copy-number- and mutation-analysis. PLOS ONE, 4(5):e5548, May 2009.

[19] J. Shendure and H. Ji. Next-generation dna sequencing. Nat Biotechnol, 26(10):1135–45, Oct 2008.

[20] M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. NATURE, 458(7239):719–724, Apr 2009.

[21] K. Taniguchi, T. Kajiyama, and H. Kambara. Quantitative analysis of gene expression in a single cell by qpcr. Nat Methods, 6(7):503–6, Jul 2009.

[22] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. The human microbiome project. NATURE, 449(7164):804–810, Oct 2007.

[23] A. Walker and J. Parkhill. Single-cell genomics. Nat Rev Microbiol, 6(3):176–7, Mar 2008.

[24] T. Woyke, G. Xie, A. Copeland, J. M. González, C. Han, H. Kiss, J. H. Saw, P. Senin, C. Yang, S. Chatterji, J.-F. Cheng, J. A. Eisen, M. E. Sieracki, and R. Stepanauskas. Assembling the marine metagenome, one cell at a time. PLoS One, 4(4):e5299, 2009.

[25] J. M. Yeakley, M. Bibikova, E. Chudin, E. Wickham, J. B. Fan, T. Downs, J. Modder, M. Kostelec, A. Arsanjani, and J. Wang-Rodriguez. Gene expression profiling in formalin-fixed, paraffin-embedded (ffpe) benign and cancerous prostate tissues using universal bead arrays. JOURNAL OF CLINICAL ONCOLOGY, 23(16):843S–843S, Jun 2005.