Part One
Sanger DNA Sequencing

Next-Generation Genome Sequencing: Towards Personalized Medicine. Edited by Michal Janitz

ISBN: 978-3-527-32090-5

1
Sanger DNA Sequencing

Artem E. Men, Peter Wilson, Kirby Siemering, and Susan Forrest

1.1
The Basics of Sanger Sequencing

From the first genomic landmark of deciphering the phiX174 bacteriophage genome achieved by F. Sanger's group in 1977 (just over a 5000 bases of contiguous DNA) to sequencing several bacterial megabase-sized genomes in the early 1990s by The Institute for Genomic Research (TIGR) team, from publishing by the European Consortium the first eukaryotic genome of budding yeast *Saccharomyces cerevisiae* in 1996 to producing several nearly finished gigabase-sized mammal genomes including our own, Sanger sequencing definitely has come a long and productive way in the past three decades. Sequencing technology has dramatically changed the face of modern biology, providing precise tools for the characterization of biological systems. The field has rapidly moved forward now with the ability to combine phenotypic data with computed DNA sequence and therefore unambiguously link even tiny DNA changes (e.g., single-nucleotide polymorphisms (SNPs)) to biological phenotypes. This allows the development of practical ways for monitoring fundamental life processes driven by nucleic acids in objects that vary from single cells to the most sophisticated multicellular organisms.

"Classical" Sanger sequencing, published in 1977 [1], relies on base-specific chain terminations in four separate reactions ("A", "G", "C", and "T") corresponding to the four different nucleotides in the DNA makeup (Figure 1.1a). In the presence of all four 2′- deoxynucleotide triphosphates (dNTPs), a specific 2′,3′-dideoxynucleotide triphosphate (ddNTP) is added to every reaction; for example, ddATP to the "A" reaction and so on. The use of ddNTPs in a sequencing reaction was a very novel approach at the time and gave far superior results compared to the 1975 prototype technique called "plus and minus" method developed by the same team. The extension of a newly synthesized DNA strand terminates every time the corresponding ddNTP is incorporated. As the ddNTP is present in minute amounts, the termination happens rarely and stochastically, resulting in a cocktail of extension

Next-Generation Genome Sequencing: Towards Personalized Medicine. Edited by Michal Janitz

ISBN: 978-3-527-32090-5

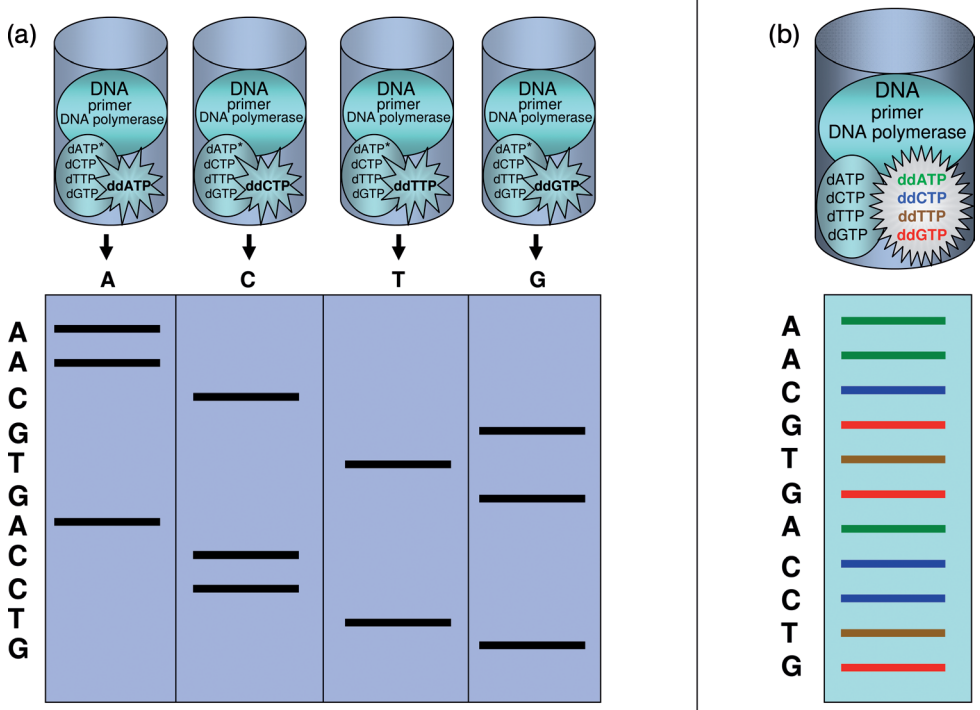

Figure 1.1 Schematic principle of the Sanger sequencing method. (a) Four separate DNA extension reactions are performed, each containing a single-stranded DNA template, primer, DNA polymerase, and all four dNTPs to synthesize new DNA strands. Each reaction is spiked with a corresponding dideoxynucleoside triphosphate (ddATP, ddCTP, ddTTP, or ddGTP). In the presence of dNTPs, one of which is radioactively labeled (in this case, dATP), the newly synthesized DNA strand would extend until the available ddNTP is incorporated, terminating further extension. Radioactive products are then separated through four lanes of a polyacrylamide gel and scored according to their molecular masses. Deduced DNA sequence is shown on the left. (b) In this case, instead of adding radioactive dATP, all four ddNTPs are labeled with different fluorescent dyes. The extension products are then electrophoretically separated in a single glass capillary filled with a polymer. Similar to the previous example, DNA bands move inside the capillary according to their masses. Fluorophores are excited by the laser at the end of the capillary. The DNA sequence can be interpreted by the color that corresponds to a particular nucleotide.

products where every position of an "N" base would result in a matching product terminated by incorporation of ddNTP at the 3′ end.

The second novel aspect of the method was the use of radioactive phosphorus or sulfur isotopes incorporated into the newly synthesized DNA strand through a labeled precursor (dNTP or the sequencing primer), therefore, making every product detectable by radiography. Finally, as each extension reaction results in a very complex

mixture of large radioactive DNA products, probably the most crucial achievement was the development of ways to individually separate and detect these molecules. The innovative use of a polyacrylamide gel (PAG) allowed very precise sizing of termination products by electrophoresis followed by *in situ* autoradiography. Later, the autoradiography was partially replaced by less hazardous techniques such as silver staining of DNA in PAGs.

As innovative as they were 30 years ago, slab PAGs were very slow and laborious and could not be readily applied to interrogating large genomes. The next two major technological breakthroughs took place in (i) 1986 when a Caltech team (led by Leroy Hood) and ABI developed an automated platform using fluorescent detection of termination products [2] separating four-color-labeled termination reactions in a single PAG tube and in (ii) 1990 when the fluorescent detection was combined with electrophoresis through a miniaturized version of PAGs, namely, capillaries [3] (Figure 1.1b). Capillary electrophoresis (CE), by taking advantage of a physically compact DNA separation device coupled with laser-based fragment detection, eventually became compatible with 96- and 384-well DNA plate format making highly parallel automation a feasible reality. Finally, the combination of dideoxy-based termination chemistry, fluorescent labeling, capillary separation, and computer-driven laser detection of DNA fragments has established the four elegant "cornerstones" on which modern building of high-throughput Sanger sequencing stands today.

Nowadays, the CE coupled with the development of appropriate liquid-handling platforms allows Sanger sequencing to achieve a highly automatable stage whereby a stand-alone 96-capillary machine can produce about half a million nucleotides (0.5 Mb) of DNA sequence per day. During the late 1980s, a concept of "highly parallel sequencing" was proposed by the TIGR team led by C. Venter and later successfully applied in human and other large genome projects. Hundreds of capillary machines were placed in especially designed labs fed with plasmid DNA clones around the clock to produce draft Sanger reads (Figure 1.2). The need for large volumes of sequence data resulted in the design of "sequencing factories" that had large arrays of automated machines running in parallel together with automated sample preparation pipelines and producing several million reads a month (Figure 1.3). This enabled larger and larger genome projects to be undertaken, culminating with the human and other billion base-sized genome projects.

Along the way, numerous methods were developed that effectively supported template production for feeding high-throughput sequencing pipelines, such as the whole genome shotgun (WGS) approach of TIGR and Celera, or strategies of subgenome sample pooling of YAC, BAC, and cosmid clones based on physical maps of individual loci and entire chromosomes (this strategy was mainly used by the International Human Genome Project team). Not only did the latter methods help to perform sequencing cheaper and faster but also facilitated immensely the genome assembly stage, where the daunting task of putting together hundreds of thousands of short DNA pieces needed to be performed. Some sophisticated algorithms based on paired end sequencing or using large-mapped DNA constructs, such as fingerprinted BACs from physical maps, were developed. Less than 20 years ago,

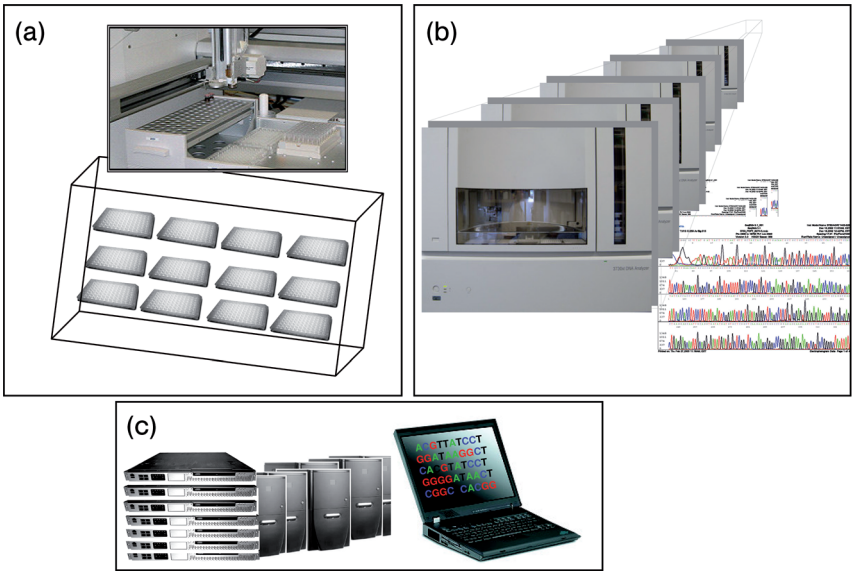

Figure 1.2 Sanger sequencing pipeline. (a) DNA clone preparation usually starts with the isolation of total DNA (e.g., whole genomic DNA from an organism or already fragmented DNA, cDNA, etc.), followed by further fragmentation and cloning into a vector for DNA amplification in bacterial cells. As a result, millions of individual bacterial colonies are produced and individually picked into multiwell plates by liquid-handling robots for isolation of amplified DNA clones. This DNA then goes through a sequencing reaction described in Figure 1.1. (b) Processed sequenced DNA undergoes capillary electrophoresis where labeled nucleotides (bases) are collected and scanned by the laser producing raw sequencing traces. (c) Raw sequencing information is converted into computer files showing the final sequence and quality of every scanned base. The resultant information is stored on dedicated servers and also is usually submitted into free public databases, such as the GeneBank and Trace Archive.

assembling a 1.8 Mb genome of *Haemophilus influenzae* sequenced by the WGS approach [4] was viewed as a computational nightmare, as it required putting together about 25 000 DNA pieces. Today, a typical next-generation sequencing machine (a plethora of which will be described in the following chapters of this book) can produce 100 Mb in just a few hours with data being swiftly analyzed (at least to a draft stage) by a stand-alone computer.

1.2 Into the Human Genome Project (HGP) and Beyond

The HGP, which commenced in 1990, is a true landmark of the capability of Sanger sequencing. This multinational task that produced a draft sequence published in 2001 [5] was arguably the largest biological project ever undertaken. Now, 7 years later, to fully capitalize on and leverage the data from the Human Genome Project, sequencing technologies need to be taken to much higher levels of output to study

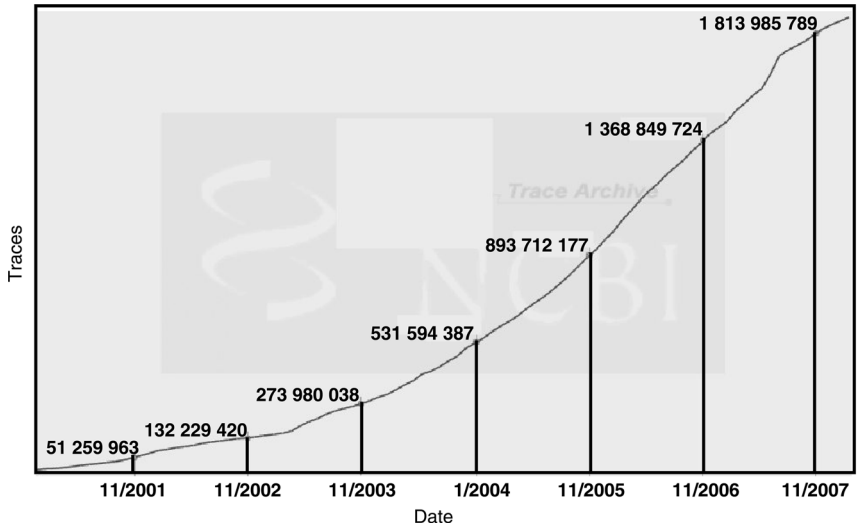

Figure 1.3 Growth of the sequencing information. Number of sequencing traces (reads) submitted to the Trace Archive grew more than 30 times between November 2001 and November 2007. Graph has been modified from reports available at http://www.ncbi.nlm.nih.gov/Traces/trace.cgi. Some more statistics of interest: (i) a major genome center produces about 1000 nucleotides per second; (ii) between November 2007 and February 2008, the Trace Archive received about 200 million trace submissions; (iii) in a single week in February 2008, just the top 10 submissions to the Trace Archive constituted 6 209 892 600 nucleotides; (iv) in 1997, there were 15 finished and published genomes of various sizes and by the end of 2007, there were 710; and (v) there are currently 442 eukaryotic and 965 microbial genome sequencing projects in progress.

multiple genomes cost effectively. Based on the capabilities already available to the medical and other research communities, numerous goals can be envisaged, such as deciphering entire genomes of many individuals, resequencing exons in large cohorts to discover new gene variants, and ultradeep analysis of cellular transcription activities and epigenetic changes that underlie multiple biological phenomena. Opportunities for discovery are virtually endless, from complex diseases to paleogenomics and "museomics" (analysis of ancient DNA), from searching for new organisms in the deep ocean and volcanoes to manipulating valuable traits in livestock and molecular plant breeding. This is where the challenges as well as major opportunities lie in the future.

1.3 Limitations and Future Opportunities

Despite the fact that the Sanger method is still considered by the research community as the "gold standard" for sequencing, it has several limitations. The first is the "biological bias" as the methodology is based on cloning foreign DNA in vectors that

have to be "bacteria friendly" and compatible with the replication machinery of the *E. coli* cells. It has been shown that some parts of chromosomes, such as centromeres and heterochromatic knobs, are practically unclonable. This limitation, in some cases, can be overcome by generating and directly sequencing PCR products but practically it is a very low-throughput and tricky approach. The second challenge is the very restricted ability of Sanger sequencing to handle and analyze allele frequencies. Often, even finding a heterozygous SNP in a PCR product is cumbersome, let alone any bases that are not represented at 1 : 1 ratios.

The third and the most significant burden of the Sanger methodology is the cost. At about $1 per kilobase, it would cost $10 000 000 to sequence a 1 Gb genome to 10× coverage! It means that average research laboratories cannot even contemplate sequencing projects that go beyond a megabase scale, thus often totally relying on the large genome centers to get the job done when it comes to sequencing your favorite genome.

Another limitation of Sanger sequencing lies at the genome assembly stage. Although Sanger reads are still the longest on the market, *de novo* assembly of single reads containing repeats is practically impossible without high-resolution physical maps of those regions if a high-quality genome draft is the goal. In regard to the length of a single read, with the current setup of CE separation of dye-tagged extension products, it probably will not reach far beyond 1 kb, despite the development of new fluorophores, with better physical characteristics, and new recombinant polymerases. Nevertheless, further miniaturization of the CE setup or replacing capillaries with chip-based systems with nanochannels that would allow analysis of molecules in the picomolar concentration range, combined with amplification of and signal detection from single template molecules [6], potentially looks like something that will keep Sanger sequencing in the game. In addition, options of combining Sanger outputs with the next-generation reads are quite promising. There of course will be still plenty of low-throughput projects that require only a few reads to be performed for a particular task, for which Sanger sequencing undoubtedly is an excellent and mature technology and will remain the gold standard for quite some time.

1.4 Bioinformatics Holds the Key

In the past 5 years, about a dozen genomes larger than a billion nucleotides in size were sequenced and assembled to various finished stages. There are 905 eukaryotic genomes currently in production as of February 2008 (http://www.genomesonline.org/gold.cgi). Most importantly, every new bit of data is being immediately made available to the general research community through databases such as the Trace Archive (http://0-www.ncbi.nlm.nih.gov.catalog.llu.edu/Traces/trace.cgi and http://www.tracearchive.ntu.ac.uk) (Figure 1.3). This enormous terabyte-sized data flow generates previously unseen possibilities for computer-based analysis and boosts fields such as comparative and population genomics to new levels of biological

discoveries via *in silico* data manipulation. The importance of the role of the bioinformatician as a major player in modern biology cannot be understated, and it will only grow with the advent of next-generation sequencers and sequencing pipelines. The larger genome projects already undertaken with Sanger sequencing have required the development of many analytical algorithms and quality assessment tools. With the significant growth in the output of DNA sequence information from the Sanger method to the next-generation DNA sequencers comes a concomitant rise in the amount of sequence information to be checked, assembled, and interpreted.

1.5 Where to Next?

A few questions obviously stand out. How can we move from just the ability to sequence a genome of a chosen individual to practical solutions that can be used in population studies or personalized medicine? How to use DNA-based information in routine medical checkups and for personalized drug prescription based on prediction of potential diseases? How one can access and explore genetic diversity in a given population, whether it is a study of diversity in birds or a search for new drought-resistant traits in agricultural crops?

The impact of genome sequencing on everyday life is getting more and more obvious. Making it affordable is the next big challenge. Many methods aimed at decreasing the cost of individual sequences are being developed very rapidly, such as genome partitioning through filtering or hybridization processes that reduce the complexity of the DNA sample to its most informative fraction, say a set of particular exons. From early experiments that involved filtration for nonrepetitive DNA via DNA reassociation followed by the sequencing of the nonrepetitive fraction to recently published array-based capturing of a large number of exons [7–9], the "genome reduction" concept seems to hold one of the keys to cheaper sequencing, as it strips down the complexity of a given genome to its gene-coding essence (almost two orders of magnitude) making it more readily accessible to sequencing analysis. The hype about developing new, cheaper ways for deciphering individual genomes was certainly boosted by several prizes offered for developing new platforms, with the paramount goal being the "$1000-genome" mark [10]. The book you hold in your hands is dedicated to the most recent ideas of how this goal might be achieved. It presents a number of totally new, exciting approaches taken forward in just a last few years that have already contributed immensely to the field of sequencing production in general and cost-efficiency in particular. Although not in the scope of this introductory chapter, it is truly worth mentioning that up to a 500 megabases a day is the average productive capacity of a current next-generation sequencing platform that is at least a thousand times more efficient than the standard 96-capillary machines used for the HGP.

The growing power of genome sequencing from the ability to sequence a bacteriophage in the late 1970s to a bacterial genome and then finally to human genome suggests that the sequencing capacity increases by about three orders of

magnitude every decade. It is hard to predict which method(s) will dominate the sequencing market in the next decade, just as it was hard to predict 30 years ago whether Sanger's or Maxam–Gilbert's method would become a major player. Both methods were highly praised in their initial phase and secured Nobel prizes for both team leaders. At that time, probably nobody would have been able to predict that the Sanger sequencing would take preference over the Maxam–Gilbert technology as a method of choice, largely thanks to subsequent development and application of shotgun cloning, PCR, and automation. In any case, only time will tell whether the next champion in DNA sequencing production will be a highly parallel data acquisition from hundreds of millions of short DNA fragments captured in oil PCR "nanoreactors," or attached to a solid surface (see following chapters) or individual analysis of unlabeled and unamplified single nucleic acids with data collection in real time, based on their physical changes detected by Raman or other spectral methods [11].

Rough extrapolation suggests that, with the current progress of technology, in 2020, we will be able to completely sequence a million individuals or produce a hundred million "exome" data sets (assuming one set being 20 000 gene exons of 1.5 kb each). Quite an impressive number, but regardless of cost it still is only about 1% of the planet's population. Nevertheless, the recent online announcement from U.S. and Chinese scientists of a very ambitious plan of sequencing up to 2000 human genomes in the coming 3 years (http://www.insequence.com/issues/2_4/features/144575-1.html) has set the bar much higher for every aspect of the technology, from the accumulation of reads to the analysis and storage of terabytes of data. It is an exciting time in genome biology, and the combination of existing sequencing methods such as Sanger and the numerous next-generation sequencing tools will result in a wealth of data ready for mining by intrepid bioinformaticians and then given back to scientists, doctors, criminologists, and farmers.

References

1 Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463–5467.

2 Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S. and Hood, L. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature*, **321**, 674–679.

3 Swerdlow, H. and Gesteland, R. (1990) Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research*, **18**, 1415–1419.

4 Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J., McKenney, K., Sutton, G., FitzHugh, W., Fields, C. and Venter, J. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **268**, 496–498.

5 International Human Genome Sequencing Consortium . (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

6 Xiong, Q. and Cheng, J. (2007) Chip capillary electrophoresis and total genetic analysis systems, in *New High Throughput*

Technologies for DNA Sequencing and Genomics (ed. K.R. Mitchelson), Elsevier, Amsterdam.

7 Albert, T., Molla, M., Muzny, D., Nazareth, L., Wheeler, D., Song, X., Richmond, T., Middle, C., Rodesch, M., Packard, C., Weinstock, G. and Gibbs, R. (2007) Direct selection of human genomic loci by microarray hybridization. *Nature Methods*, **4**, 903–905.

8 Okou, D., Stinberg, K., Middle, C., Cutler, D., Albert, T. and Zwick, M. (2007) Microarray-based genomic selection for high-throughput resequencing. *Nature Methods*, **4**, 907–909.

9 Porreca, G., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S., LeProust, E., Peck, B., Emig, C., Dahl, F., Gao, Y., Church, G. and Shendure, J. (2007) Multiplex amplification of large sets of human exons. *Nature Methods*, **4**, 931–936.

10 Bennett, S., Barnes, C., Cox, A., Davies, L. and Brown, C. (2005) Toward the $1000 human genome. *Pharmacogenomics*, **6**, 373–382.

11 Bailo, E. and Deckert, V. (2008) Tip-enhanced Raman spectroscopy of single RNA strands: towards a novel direct-sequencing method. *Angewandte Chemie*, **4**, 1658–1661.