CHAPTER 1

DNA Sequencing Strategies

1.1 THE EVOLUTION OF DNA SEQUENCING TECHNOLOGIES

Methods for DNA sequencing were invented multiple times, first by chemical means (Maxam and Gilbert, 1977) and later using biochemical approaches (Sanger and Coulson, 1975). Initially, radioactive compounds were used to make the DNA bands, which represent base pairs (bp), detectable by autoradiography after an electrophoretic separation. A major step forward was the switch to fluorescent labels in the 1980s, which could be detected during the electrophoretic separation and automatically recorded by computers, instead of humans reading autoradiograms, which were then manually converted to computer files. Naturally, the manual editing necessary for the creation of the early DNA sequence files led to a high error rate, which should be kept in mind when comparing DNA sequences downloaded from public repositories, such as GenBank.

The early automated DNA sequencing machines allowed for the production of a few kilobase pairs of DNA sequence per day and machine. Today, this has increased tremendously through the invention of the so-called next-generation DNA sequencing technologies and the integration of robotic workstations into the DNA sequencing workflow. It is now possible to produce millions of reads per day and device. Still the read length obtained in 2012, with few exceptions, does not yet reach that of the first-generation automated DNA sequencing machines, which peaked at about 1200 bp. The Roche 454-type systems (www.454.com), which are currently the mainstream long-read machines, generate an average read length of around 450 bp at the most. Finally, this limit is about to

be broken by yet another generation of DNA sequencing devices, which are expected to produce read lengths of at least several kilobase pairs per read. The new system from Pacific Biosciences, called PacBio RS (www.pacificbiosciences.com), can already reach this level of longer read lengths but currently with major trade-offs in sequence accuracy.

Regardless of the DNA sequencing technology used for automated DNA sequencing, all approaches ultimately lead to the creation of a so-called trace file, which captures essentially a sequence of base pairs, which never represents more than a fraction of a genome. The trace files are essentially graphical representations of the DNA sequencing progress, either represented as a succession of DNA fragments throughout a separation by size or the representation of a DNA synthesis on a chip or in a flow cell over time. The DNA sequence can be extracted from the trace file as a character file. Typically, the character file is formatted as a "FASTA"-format file, which essentially contains the DNA sequence (usually 60 characters per line) and a single description line, which is preceded by a ">" character on the first line. Multiple sequence files can be combined into a single file with multiple pairs of a description line and a sequence, each corresponding to one sequence and preceded by a ">" character, which leads to a so-called multiple FASTA file.

1.2 DNA SEQUENCE ASSEMBLY STRATEGIES

Until now, there is no DNA sequencing technology that directly leads to a complete genome with a single sequencing run. In all cases, the complete contiguous sequences (i.e., contigs) has to be constructed from small individual sequence reads, which range from 30 to a few hundred base pairs. The necessary sequence assembly can be done using three fundamental genome assembly strategies: primer walking, shotgun assembly, or a mixed strategy.

Initially, much of the DNA sequence production was done by so-called primer walking (Voss et al., 1993). This essentially meant that after each sequence run, once the new DNA sequences were integrated into the assembly project, a set of oligonucleotides was calculated (to obtain the next DNA sequencing primers), which could extend the existing contigs. This approach was quite slow, as the new primers needed to be synthesized before every new sequencing run and could fail for a number of reasons, and it was also expensive, as dedicated primers needed to be synthesized for each individual sequencing run. First and foremost, problems arose from repetitive elements within the clone that was being sequenced, which were not yet known to the researcher and therefore not taken into account during the primer calculation, and resulted in DNA sequencing errors.

This was especially true for single-stranded DNA sequences, where the usual sequencing errors led to the calculation of primers that were not priming correctly due to built-in mistakes. An advantage of the primer-walking strategy was the low coverage necessary, even when both DNA strands were sequenced completely. In many cases, the final coverage was between three and four times the genome equivalent. Figure 1.1 shows an overview of the primer-walking sequencing strategy.

Craig Venter, then at The Institute for Genomic Research (TIGR), and his team can be credited for the invention of the "shotgun sequencing" approach (Venter et al., 1996). In the early version of this DNA sequencing strategy, which was still based on cloned DNA fragments, only universal primer pairs were used to sequence into the cloned DNA inserts with much higher redundancy (6- to 10-fold genome coverage) than in the primer-walking approach. Once enough end sequences were obtained, they were assembled into contigs. Figure 1.2 shows an overview of the shotgun sequencing strategy.

If all regions of a genome would be equally clonable, this strategy could have rapidly yielded complete genome sequences as only standard components, such as universal primers, were required during sequence production, but this was rarely ever the case. Therefore, applying this strategy in its pure form led to almost completely sequenced genomes, but in the case of microbial genomes as an example, many gaps remained, often numbering several hundred gaps per 2 to 3 mega bp of genomic sequence. In addition, in most cases many regions within the genome were only

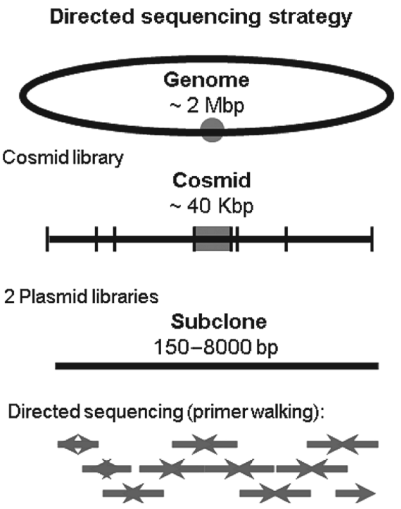

FIGURE 1.1 Overview of the primer-walking DNA sequencing strategy.

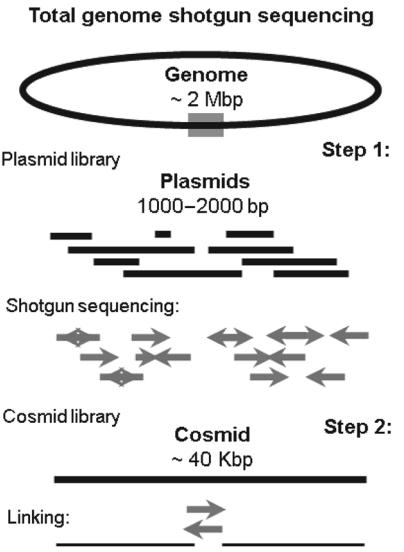

FIGURE 1.2 Overview of the shotgun DNA sequencing strategy.

sequenced on a single DNA strand (most likely also due to cloning artifacts). This increased the error rate in the finished genome, as multiple instances of a single base occurring in sequence are generally best resolved by the sequencing of both DNA strands. The redundancy of a shotgun-sequenced genome is typically much higher than that of one sequenced by primer walking, with 8- to 10-fold, but sometimes more than 20-fold being reported for the finished product.

Ultimately, almost all groups attempting to sequence complete genomes settled on a mixed strategy, where the bulk of the genome sequencing was accomplished through shotgun sequencing and the gaps were closed through primer walking on a set of large-insert clones. Figure 1.3 shows an overview of the mixed sequencing strategy.

This approach yields complete genomic sequences at a much lower cost than the genomes finished entirely by primer walking, as only very few dedicated DNA sequencing primers are required with almost the same ultimate DNA sequence accuracy. Typically, in the case of a microbial genome with a size of 2 to 3 mega bp, between 200 and up to 500 gaps need to be closed by primer walking, most of these being fairly small. In support of gap closure, primer-walking technology using large-insert libraries, such as cosmids, fosmids, or lambda clones were developed, which can be used to efficiently complement the shotgun sequencing results. Today,

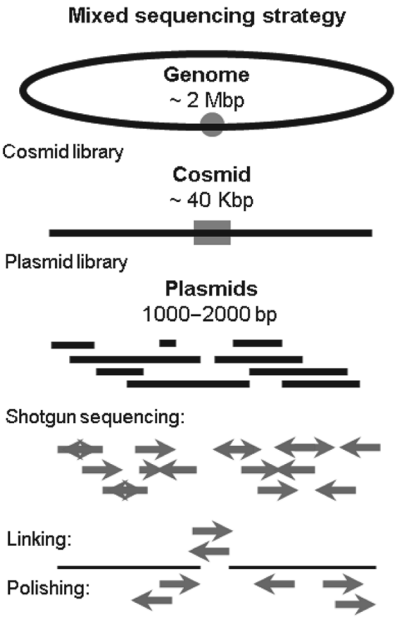

FIGURE 1.3 Overview of the mixed DNA sequencing strategy.

most of this is already history, as next-generation DNA sequencing is replacing most of the above. Still the aforementioned techniques have led to the generation of more than a thousand genomes, many of them of very good quality (1 error per 10,000 bp or less). These early complete genomes provide the templates for today's sequencing experiments.

1.3 NEXT-GENERATION SEQUENCING

The latest DNA sequencing technologies certainly influence the current DNA sequencing strategy and lead to new approaches in biochemistry and biotechnology. Two major types of next-generation DNA sequencing technology are being developed: short-read systems, with read lengths up to approximately 100 bp per DNA fragment, and long-read systems, with average read lengths of at least 300 bp and potential read lengths of several kilobase pairs per fragment. There are many different technologies that are currently being explored for next-generation sequencing, with the Illumina, ABI SOLiD, and Roche 454 sequencing approaches being in the lead at present. The latest sequencing technologies do not require cloned DNA as a prerequisite, thus sequencing can be accelerated tremendously. A single machine run can yield several million reads, which is sufficient for the shotgun assembly of a microbial genome.

As several hundred genomes, including the human genome, have been completely sequenced to date and thousands more have been completed with various grades of quality, it is now often of little interest to complete genomes entirely. As the cost for DNA sequence production has been lowered considerably, when compared to the older capillary DNA sequencing technology (a human genome equivalent can be generated for less than $10,000 in 2012 and might be possible to be generated for less than $1,000 shortly), sequence redundancy can be increased considerably to factors of 30- to 40-fold or even higher coverages. This often leaves only 20 to 50 gaps in the genomic sequence of a microbial genome, which might or might not be closed through traditional primer-walking strategies. Often, the new genomic sequence is close enough in similarity to an existing complete genome that the fully completed genome can be used as a "scaffold" to align the newly generated sequence in a meaningful way. The low number of gaps means that in a microbial genome less than 1% of the genes are not fully characterized. This is sufficient for almost all genome research needs.

New analysis strategies are being developed, which supersede many older molecular biological techniques, including PCR experiments, DNA mapping technologies, and even gene chips. Complete sequencing of the "total genome" of an organism is now possible. In the human case, this means not only the generation of the DNA sequence of the human chromosomes, but also the sequencing of all DNA-containing organisms and viruses in the body fluids, such as blood and also those in the gut. The total DNA content of a human being is at least a thousand times larger than the DNA content of the chromosomes, posing major challenges for the annotation of this entire conglomerate of DNA molecules.

1.4 SEQUENCING BIAS AND ERROR RATES

In our opinion, to date, there is really no "complete and error free" genome sequence that was ever characterized by humans. Many sources of error exist, leading to certain error rates in all DNA sequences, which have been submitted to the public repositories. Typically, the generation of a new sequence involves a large number of individual steps, from the isolation of the original DNA, to the preparation for DNA sequencing using various kits, to the actual operation of the DNA sequencing equipment, and finally to the assembly and annotation. In most steps, direct human involvement is still necessary. Errors are manifold in nature, from single-base pair differences (Meacham et al., 2011), which can lead to frameshifts in the sequence, to different assemblies of the same sequence which lead to a false

representation of the genomic arrangement and gene order (for example, two assemblies of the bovine genome currently exist, which are based on the same data, but differ considerably) (Florea et al., 2011), to "missing" genes or regions based on the erroneous assembly of repetitive regions (for example, in the human genome) (Semple et al., 2002) in a genome. Thus already the input into the genome annotation process is not without flaws. The goal should be to use the best possible input sequence. This can be best achieved by using at least two different sequencing techniques simultaneously while generating the final sequence. This also helps to iron out the bias, which is introduced by every sequencing technology. Some machines cannot deal with high-GC (guanine–cytosine) content DNA fragments the same way they deal with low-GC content fragments, some machines balk at homopolymer stretches in the DNA sequence, and some machines lose accuracy during the sequencing process to a large degree, calling the final bases in the sequencing run with no more than 70% accuracy.

In the future, the combination of short reads (around 100 bp), with reads generated at both ends of the DNA fragment (paired-end sequencing) and very long reads (over 1000 bp) will probably prevail for *de novo* sequencing, while short reads alone will be used if a more or less complete genome template already exists (for example, in the case of the human genome).

REFERENCES

Florea, L., Souvorov, A., Kalbfleisch, T.S., Salzberg, S.L. 2011. Genome assembly has a major impact on gene content: A comparison of annotation in two *Bos taurus* assemblies. *PLoS One* 6:e21400.

Maxam, A.M., Gilbert, W. 1977. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA* 74(2):560–564.

Meacham, F., Boffelli, D., Dhahbi, J., Martin, D.I., Singer, M., Pachter, L. 2011. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* 21:451.

Sanger, F., Coulson, A.R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* 94(3):441–448.

Semple, C.A., Morris, S.W., Porteous, D.J., Evans, K.L. 2002. Computational comparison of human genomic sequence assemblies for a region of chromosome 4. *Genome Res.* 12:424–429.

Venter, J.C., Smith, H.O., Hood, L. 1996. A new strategy for genome sequencing. *Nature* 381(6581):364–366.

Voss, H., Wiemann, S., Grothues, D., et al. 1993. Automated low-redundancy large-scale DNA sequencing by primer walking. *Biotechniques* 15:714–721.