STRATEGIES FOR THE SYSTEMATIC SEQUENCING OF COMPLEX GENOMES

Eric D. Green

Recent spectacular advances in the technologies and strategies for DNA sequencing have profoundly accelerated the detailed analysis of genomes from myriad organisms. The past few years alone have seen the publication of near-complete or draft versions of the genome sequence of several well-studied, multicellular organisms — most notably, the human. As well as providing data of fundamental biological significance, these landmark accomplishments have yielded important strategic insights that are guiding current and future genome-sequencing projects.

Biology and medicine are in the midst of a revolution, the full extent of which will probably not be realized for many years to come. The catalyst for this revolution is the Human Genome Project[1] and related activities that aim to develop improved technologies for analysing DNA, to generate detailed information about the genomes of numerous organisms, and to establish powerful experimental and computational approaches for studying genome structure and function. The past few years have seen a remarkable crescendo in accomplishments related to DNA sequencing, with genome sequences being generated for several key experimental organisms, including a yeast (*Saccharomyces cerevisiae*), a nematode (*Caenorhabditis elegans*), a fly (*Drosophila melanogaster*), a plant (*Arabidopsis thaliana*) and the human (*Homo sapiens*). Collectively, the generation of these sequence data and others is launching the 'sequence-based era' of biomedical research.

Associated with the above accomplishments has been the refinement of existing strategies for genome sequencing, as well as the development of new ones. Among these are approaches that make extensive use of large-insert clones and associated physical maps, some that take a whole-genome approach without using clone-based physical maps, and others that use a hybrid strategy that involves elements of the other two. Each of these general strategies for genome sequencing is described in this review.

There are many potential uses of genome-sequence data. In some cases, a detailed and accurate sequence-based 'blueprint' of a genome is required (for example, to establish a comprehensive gene catalogue and/or to gain insight into long-range genome organization), whereas in other cases, an incomplete survey will suffice (for example, to acquire information about the repetitive sequences in a genome and/or to carry out simple, non-comprehensive comparisons to sequences from other organisms). Importantly, the intended use(s) of genome-sequence data must be carefully considered when choosing a specific sequencing strategy and defining the end point of a particular project. These issues, as well as the plans for future sequencing initiatives by the Human Genome Project, are also discussed.

Contemporary sequencing methods

Shortly after the Human Genome Project began in 1990, pilot projects were initiated that aimed to sequence the smaller genomes of several key model organisms (for example, *Escherichia coli*, *S. cerevisiae*, *C. elegans* and *D. melanogaster*) using available technologies. At the time, the general idea was that the eventual sequencing of the human and other vertebrate genomes could not begin in earnest without the development of a new, revolutionary sequencing technique(s). In reality, such methods were not forthcoming. However, numerous incremental improvements, each evolutionary in nature, were made

Genome Technology Branch and NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.
e-mail: egreen@nhgri.nih.gov

during the early sequencing efforts of the Human Genome Project; together, these created a viable path for the sequencing of vertebrate-sized genomes. The crucial enhancements were geared towards increasing the accuracy and efficiency of the dideoxy chain termination sequencing method, originally developed by Fred Sanger and colleagues in the 1970s[2]. This simple but elegant technique, which involves the electrophoretic separation and detection of *in vitro* synthesized, single-stranded DNA molecules terminated with dideoxy nucleotides, has been central to virtually all past and current genome-sequencing projects of any significant scale.

The evolutionary improvements to the Sanger sequencing method that have been devised in recent years are diverse in nature. First, significant advances have been made in the laser-based instrumentation that allows the automated detection of fluorescently labelled DNA molecules[3,4], including the development of capillary-based sequencing instruments in the late 1990s capable of analysing 96 samples in parallel[5,6]. Second, numerous enhancements in the biochemical components required for the Sanger method sequencing reactions were developed, such as improved thermostable polymerases[7], fluorescent-dye-labelled dideoxy terminators[8] and more robust fluorescent dye systems[9–12]. Of note, the current state-of-the-art methods for fluorescence-based DNA sequencing using capillary-based instruments provide ~500–800 bases of quality sequence per reaction (that is, per 'sequence read'). Third, various robotic systems have been designed to automate specific steps in the sequencing process[13,14]. These include systems that facilitate subclone library construction, that do the picking and arraying of bacterial subclones, that purify template DNA from individual subclones, that prepare sequencing reactions, and that load samples on slab gels or capillaries before electrophoresis. The net effect of these various improvements has been a profound increase in the quality and overall throughput of DNA sequencing, with a corresponding decrease in the associated costs — estimated to have been reduced >100-fold over the past decade[15]. It is worth noting that this cost reduction has, in part, been realized by the creation of very large sequencing centres that greatly benefit from the economies-of-scale and efficiencies associated with large production facilities.

In addition to the biochemical and mechanical improvements, crucially important software systems have been developed for analysing primary sequence data and for carrying out sequence assembly. For example, new and more powerful programs are now available that analyse the data associated with each sequence read; this includes both calling the nucleotide base at each position and assigning a corresponding quality score to reflect the statistical likelihood that the indicated base call is correct. The resulting quality scores provide both the ability to monitor the quality of the raw data and to detect authentic overlaps among sequence reads in a more accurate fashion. Additional programs have been developed that use the assigned quality scores for assembling sequences in a statistically robust way, yielding sequence assemblies with calculated accuracy rates that provide an objective criterion for subsequent sequence finishing and analysis. Yet other programs provide user-friendly viewers for inspection and editing of the resulting sequence assemblies. A particularly popular suite of programs for these various steps is Phred, Phrap and Consed[16–18], which are designed for base calling, sequence assembly and the viewing of sequence assemblies, respectively. Another commonly used sequence assembly program is GAP[19]. In addition to improving the overall quality of the generated sequence by using sophisticated, statistically driven approaches to base calling and sequence assembly, such programs facilitate the establishment of more automated routines for sequence finishing[20] (see below) and for assessing the accuracy of FINISHED SEQUENCE[21,22]. In a related fashion, numerous software systems have also been developed for monitoring and assimilating the vast quantities of primary sequence data generated at high-throughput sequencing facilities[23–26].

The availability of improved methods and technologies for DNA sequencing prompted the testing of various general strategies for genome sequencing early in the Human Genome Project. These included approaches that used transposons to create random insertions (and thus sequence priming sites) in cloned DNA[27], as well as those that used multiplex sequencing strategies coupled with various detection schemes[28–30]. As the cost of individual sequence reads generated by the Sanger method decreased, these other approaches were ultimately found to be less efficient for sequencing large genomes than the general strategy known as shotgun sequencing. The basic theory of shotgun sequencing, which was first described in the early 1980s[31–35], is that a large piece of DNA can be sequenced by first fragmenting it into smaller pieces, then generating redundant amounts of sequence data from random fragments and finally piecing the individual sequence reads back together (that is, assembling them) in a fashion that accurately reveals the sequence of the starting DNA. Laboratory manuals that detail the various experimental steps involved in shotgun sequencing are available[36–39]. Below, the principal approaches for shotgun sequencing, as implemented for the systematic sequencing of complex genomes, are described.

Clone-by-clone shotgun sequencing

So far, the most commonly used approach for establishing the sequence of large, complex genomes involves the shotgun sequencing of individual mapped clones. This strategy is best exemplified by the efforts of the Human Genome Project to sequence the yeast[40], nematode[41] and human[15] genomes (TABLE 1). Also referred to as hierarchical shotgun sequencing or map-based shotgun sequencing[15], this strategy follows a 'map first, sequence second' progression: the target DNA (either an entire genome or a smaller genomic region) is first analysed by clone-based physical mapping methods, and then individual mapped clones that together span the region of interest are selected and subjected to shotgun sequencing (FIG. 1a). The process of clone-by-clone shotgun sequencing can be conceptually divided into a series of discrete and sequential steps[42], each of which is described below.

FINISHED SEQUENCE
Complete sequence of a clone or genome, with a defined level of accuracy and contiguity.

Table 1 | **The sequencing of genomes from multicellular organisms***

Organism	Size‡	General strategy	Status	Relevant URLs
Nematode[41]	~100	Clone-by-clone	Virtually complete	genome.wustl.edu/gsc/Projects/C.elegans www.sanger.ac.uk/Projects/C_elegans www.wormbase.org
Arabidopsis thaliana[68]	~125	Clone-by-clone	Virtually complete	www.arabidopsis.org
Drosophila[86,87]	~120§	Hybrid‖	Draft complete, finishing in progress	www.fruitfly.org flybase.bio.indiana.edu www.celera.com www.hgsc.bcm.tmc.edu/drosophila
Human[15]	~3,200	Clone-by-clone	Draft complete, finishing in progress	www.nhgri.nih.gov/genome_hub.html
Human[92]	~3,200	Whole-genome shotgun	Draft complete	www.celera.com
Mouse	~3,200	Hybrid	In progress	www.informatics.jax.org www.nih.gov/science/models/mouse www.ncbi.nlm.nih.gov/genome/seq/MmHome.html
Mouse	~3,200	Whole-genome shotgun	In progress	www.celera.com
Rat	~3,200	Hybrid	In progress	www.hgsc.bcm.tmc.edu/rat www.celera.com www.nih.gov/science/models/rat www.rgd.mcw.edu
Zebrafish	~1,700	Hybrid	In progress	www.sanger.ac.uk/Projects/D_rerio zfin.org/ZFIN
Pufferfish (*Fugu rubripes*)	~400	Whole-genome shotgun¶	In progress	www.jgi.doe.gov/tempweb/programs/fugu.htm
Pufferfish (*Tetraodon nigroviridis*)	~400	Whole-genome shotgun¶	In progress	www.genoscope.cns.fr/externe/tetraodon
Rice[102]	~400	Clone-by-clone	In progress	rgp.dna.affrc.go.jp/Seqcollab.html

*This table is not intended to be comprehensive; listed here are the highlights of the main genome-sequencing projects involving multicellular organisms. ‡In megabase pairs (Mb); in most cases, the indicated genome size is an estimate. §Euchromatic portion (total genome size estimated at ~180 Mb). ‖Involved a component for which an assembly was carried out using sequence reads generated by whole-genome shotgun sequencing; however, the ultimate finishing of the sequence to high accuracy is being done with sequence reads generated in a clone-by-clone fashion; so, this represents a hybrid shotgun-sequencing approach. ¶So far, plans only include the generation of sequence data by whole-genome shotgun sequencing.

Map construction. In clone-based physical mapping, pieces of genomic DNA are cloned using a suitable host–vector system (based in bacteria or yeast). Individual clones are then analysed for the presence of unique DNA landmarks (for example, SEQUENCE-TAGGED SITES (STSs)[43], restriction sites and/or other sequence-based elements[44]), with data about the sharing of common landmarks among clones then used to assemble overlapping clone maps, called CONTIGS (FIG. 2). Each clone contig contains the DNA from a contiguous segment of the source genome.

Two main cloning systems have been used for constructing physical maps of large genomic regions: yeast artificial chromosomes (YACs)[45,46] and bacterial artificial chromosomes (BACs)[47] (or the closely related P1-derived artificial chromosomes[48] (PACs)). YACs contain cloned inserts upwards of a megabase pair in size and, for example, were used to construct the first-generation physical maps of the human[49–52] and mouse[53] genomes. However, for various scientific and logistical reasons, YACs are not a good starting template for shotgun sequencing (although they are used for this purpose in some rare instances). By contrast, BACs, which carry 100–200-kb inserts in a reasonably stable fashion, have proven extremely useful for constructing second-generation physical maps of large genomes. In addition, these clones are well suited as starting templates for shotgun sequencing.

Efficient approaches for constructing BAC-based physical maps have been developed in recent years[54,55]. These methods essentially represent contemporary adaptations of the classic restriction-mapping techniques used to generate clone-based physical maps of the *E. coli*[56], yeast[57,58] and nematode[59] genomes. In short, restriction enzyme digest-based fingerprints are derived for each BAC clone. Pairwise comparisons of the fingerprints are made and the resulting information is analysed to infer clone overlaps, which in turn are used to assemble BAC contigs[54,55] (FIG. 2). The BAC-fingerprinting process has been made relatively high-throughput in nature, allowing high-resolution BAC contig maps to be constructed for the plant *A. thaliana*[60,61], human[62] and mouse (M. Marra, unpublished data) genomes. Similar BAC-based maps of the rat, zebrafish and bovine genomes are now being constructed. The resulting BAC contig maps often provide continuity that extends for several megabase pairs. In addition, supplementary mapping data is typically generated (for example, information about the presence of STSs, genetic markers and gene sequences in the assembled

SEQUENCE-TAGGED SITE (STS). Short (for example, <1,000 bp), unique sequence associated with a PCR assay that can be used to detect that site in the genome.

CONTIG
Overlapping series of clones or sequence reads (for a clone contig or sequence contig, respectively) that corresponds to a contiguous segment of the source genome.

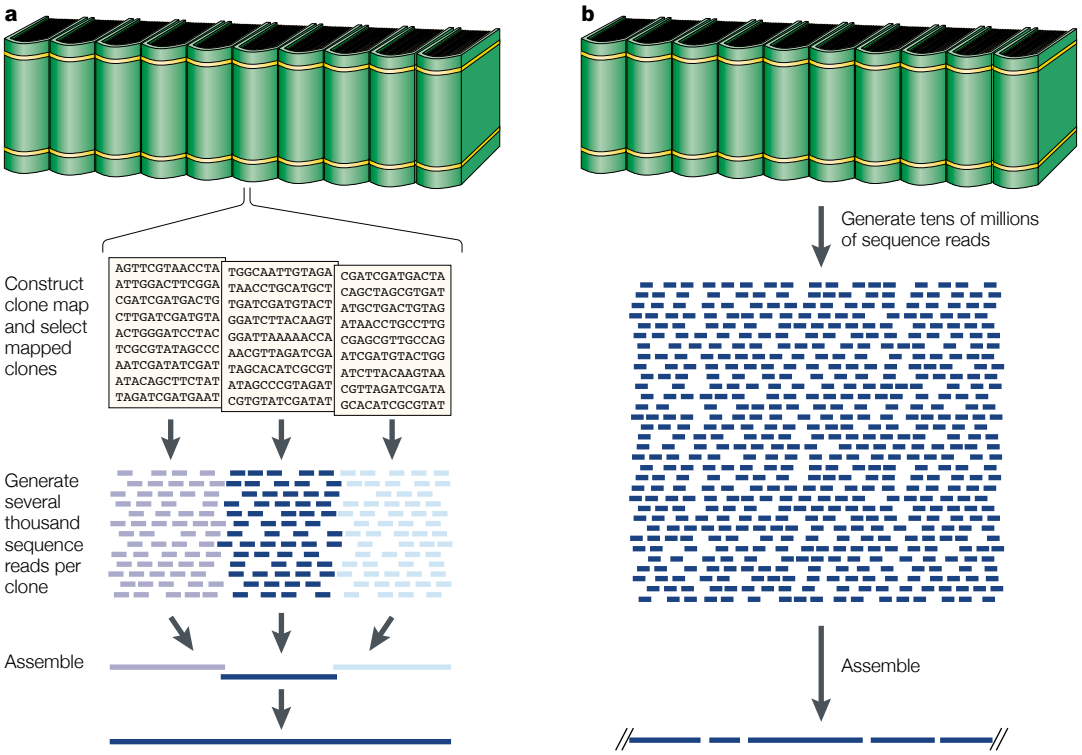

Figure 1 | **Two main shotgun-sequencing strategies. a** | Schematic overview of clone-by-clone shotgun sequencing. A representation of a genome is made by analogy to an encyclopaedia set, with each volume corresponding to an individual chromosome. The construction of clone-based physical maps produces overlapping series of clones (that is, contigs), each of which spans a large, contiguous region of the source genome. Each clone (for example, a bacterial artificial chromosome (BAC)) can be thought of as containing the DNA represented by one page of a volume. For shotgun sequencing, individual mapped clones are subcloned into smaller-insert libraries, from which sequence reads are randomly derived. In the case of BACs, this typically requires the generation of several thousand sequence reads per clone. The resulting sequence data set is then used to assemble the complete sequence of that clone (see FIGS 3,4). **b** | Schematic overview of whole-genome shotgun sequencing. In this case, the mapping phase is skipped and shotgun sequencing proceeds using subclone libraries prepared from the entire genome. Typically, tens of millions of sequence reads are generated and these in turn are subjected to computer-based assembly to generate contiguous sequences of various sizes.

BAC contigs), thereby providing landmark-content data[63] that can be used for the relative positioning of the BAC contigs in the genome. Note that the above mapping efforts only correspond to ~1% of the total cost of a clone-based shotgun-sequencing project[15].

Clone selection. Once provided with an assembled BAC contig map, minimally overlapping clones (that together represent a MINIMAL TILING PATH across a genomic region) are then selected for shotgun sequencing (FIG. 2). The most important criterion for selecting a BAC from such a map is evidence that the clone is authentic. Typically, the restriction enzyme digest-based fingerprint of a candidate BAC is examined and compared with that of all overlapping clones; a BAC is only selected for sequencing if each of the restriction fragments in its fingerprint is also present in at least one overlapping clone. Common anomalies, such as an internal deletion or a chimeric insert, would typically result in the presence of a restriction fragment(s) not present in an overlapping clone(s); the detection of such an anomaly would make that BAC inappropriate for sequencing. Note that properly constructed BAC-based physical maps with highly redundant clone COVERAGE (for example, tenfold or greater) can be readily used for selecting suitable sets of clones for sequencing across most genomic regions[62] (hence, the name SEQUENCE-READY MAPS[64] (FIG. 2)).

Subclone library construction. For each selected BAC, the cloned DNA is purified and subjected to random fragmentation (FIG. 3), most often by physical shearing methods. After enzymatic repair of the broken ends and size fractionation, the DNA fragments in a defined size range (for example, 2–5 kb) are recovered and subcloned into a plasmid- or M13-based vector. The chief advantages of plasmid subclones are that the resulting double-stranded templates can be used for deriving sequence reads from both ends of the subcloned fragment (at the cost of purifying only one template) and that the pair of sequence reads from each subclone ('read pairs' or 'mate pairs') can be used to facilitate and/or assess the subsequent sequence assembly[65] (see below). The principal advantages of M13 subclones are that the template DNA is simpler to prepare and that the sequence data generated with the resulting single-stranded template is generally of

MINIMAL TILING PATH
A minimal set of overlapping clones that together provides complete coverage across a genomic region.

COVERAGE
The average number of times a genomic segment is represented in a collection of clones or sequence reads (synonymous with redundancy).

SEQUENCE-READY MAP
Typically considered an overlapping bacterial clone map (for example, a BAC contig map) with sufficiently redundant clone coverage to allow for the rational selection of clones for sequencing.

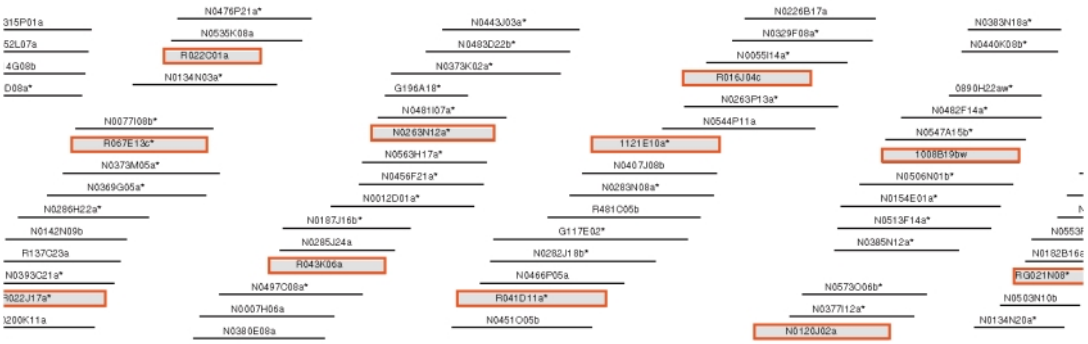

Figure 2 | **Sequence-ready BAC contig map.** A collection of overlapping bacterial artificial chromosome (BAC) clones that contain human DNA was subjected to restriction enzyme digest-based fingerprint analysis[54]. The resulting data was analysed using the program FPC[100,101], which constructed the depicted BAC contig map that spans >1 Mb. The 11 clones outlined in red, which provide a minimal tiling path across the corresponding genomic region, were selected for sequencing.

higher quality. Plasmid and M13 vectors have slightly different cloning biases, with certain genomic regions being more readily recovered in one type of subclone compared with the other[66]; as a result, a combination of plasmid and M13 subclones is sometimes used in a shotgun-sequencing project to minimize the inherent cloning biases associated with any one system. The main disadvantage of using both types of subclones is that it introduces additional complexity to the sequencing process, with the need to construct two subclone libraries for each BAC and to purify two types of template DNA for sequencing.

Random shotgun phase. In the initial sequencing phase (FIG. 3), subclones are picked at random and sequence reads are derived from the UNIVERSAL PRIMING SITE(S) located on one or both sides of the cloned insert (in the case of M13 and plasmid subclones, respectively). After the generation of a sufficient amount of redundant sequence data (relative to the starting BAC insert), the sequence reads are computationally assembled on the basis of detected sequence overlaps (FIG. 4). The resulting assembly typically yields a series of sequence contigs, each of which consists of a collection of overlapping reads and a deduced consensus sequence.

For producing highly accurate sequence (for example, >99.99% accurate, the agreed-on minimal standard for finished human sequence[21,22]), data that provide ~8–10-fold sequence coverage (that is, redundancy) are typically generated; this is often called FULL-SHOTGUN SEQUENCE. For example, in the case of an ~150-kb BAC, ~3,000 usable sequence reads (excluding those that failed to produce any data, those derived from the BAC vector or contaminating *E. coli* genomic DNA introduced during subclone library construction, and those from non-recombinant subclones) that average ~500 quality bases in length are required to provide roughly tenfold coverage. Once such a level of sequence coverage is attained, the clone is ready for the next step in the shotgun-sequencing process.

Assemblies generated with lower levels of coverage can give tremendous insight into the sequence of a clone. For example, assemblies generated with even ~3–5-fold sequence coverage can be used for a wide array of important analyses[67]. Such a product is often referred to as WORKING DRAFT SEQUENCE, which represents both an intermediate that is en route to the generation of highly accurate and complete sequence, as well as an analytical tool of immense value for preliminary analyses. Both of these features are nicely illustrated by the recently completed working draft sequence of the human genome produced by the Human Genome Project[15].

Directed finishing phase. The initial assembly of the sequence reads generated from a BAC (assuming ~8–10-fold coverage) typically yields a handful of sequence contigs that together reflect virtually all of the starting clone. The remaining problems usually include discontinuities between sequence contigs (that is, gaps), areas of low sequence quality, ambiguous bases in the consensus sequence and contig misassemblies. In the final sequencing phase (FIG. 3), the preliminary sequence assembly (or PREFINISHED SEQUENCE) is refined in a directed fashion to produce a final, highly accurate sequence; this process is referred to as sequence finishing. In contrast to the routine, high-throughput-orientated nature of the random shotgun phase, the sequence finishing phase is highly customized and generally associated with a lower throughput. Often, each BAC presents its own unique set of problems. Resolving these involves additional sequencing of existing subclones, as well as deriving sequence data directly from purified BAC DNA or from amplified PCR products; it often also requires the use of alternate sequencing chemistries and modified experimental protocols.

The process of sequence finishing is now greatly facilitated by the availability of recently developed computer software systems. Base-calling programs that assign a quality score to each base of a sequence read (for example, Phred[16,17]) provide the statistical foundation for more robust sequence assembly programs (for example, Phrap), which in turn provide the means to monitor the accuracy of the resulting sequence assemblies. This allows the end point of the sequence finishing phase to be clearly defined, including the attainment of established accuracy rates for the final sequence. In addition, software tools for facilitating sequence finishing (for example, the program Autofinish[20]) can be used to partially automate the finishing process. Specifically,

UNIVERSAL PRIMING SITE
A short sequence (for example, 16–24 bases) in a cloning vector, immediately adjacent to the vector–insert junction to which a common (that is, universal) sequencing primer can anneal.

FULL-SHOTGUN SEQUENCE
A type of prefinished sequence, in this case with sufficient coverage to make it ready for sequence finishing (typically on the order of 8–10-fold coverage).

WORKING DRAFT SEQUENCE
A type of prefinished sequence, often meant to correspond to sequence with coverage that puts it at roughly the halfway point towards full-shotgun sequence.

PREFINISHED SEQUENCE
Sequence derived from a preliminary assembly during a shotgun-sequencing project (at this stage, the sequence is often not contiguous nor highly accurate).

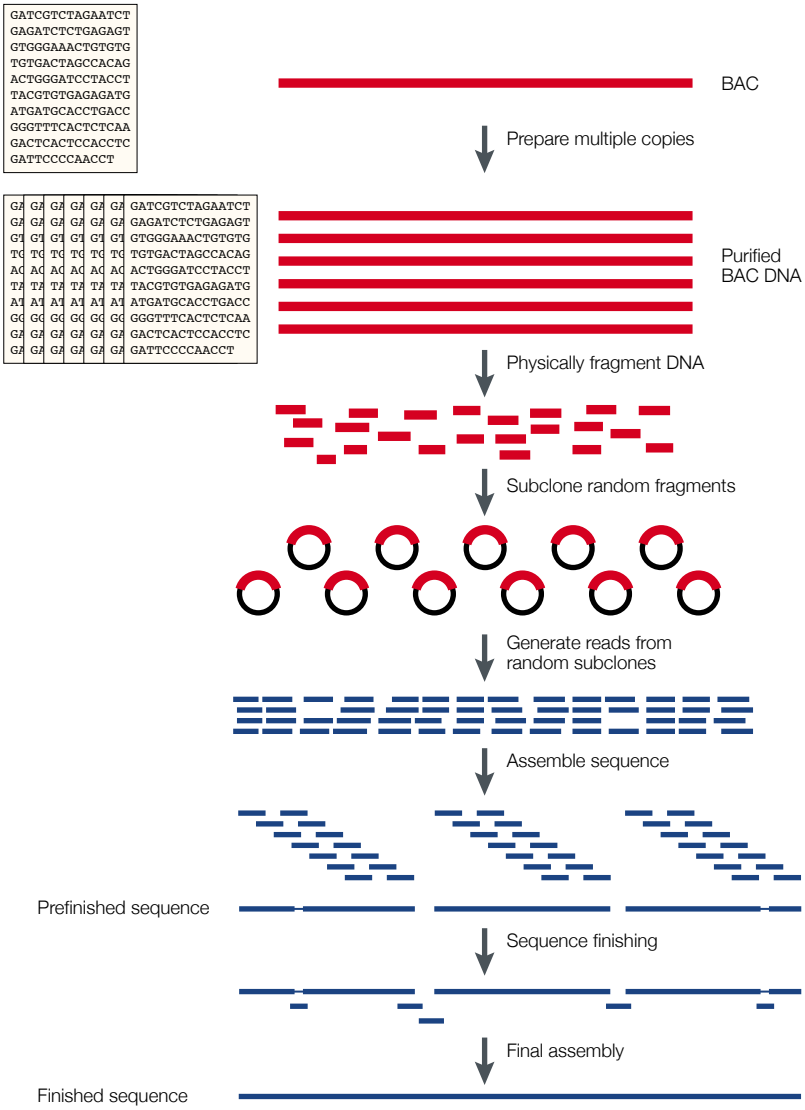

Figure 3 | **Main steps in clone-by-clone shotgun sequencing.** An individual clone, such as a bacterial artificial chromosome (BAC), is selected, and a large amount of BAC DNA is purified (analogous to making several photocopies of a specific page from a given volume of an encyclopaedia set). The purified DNA is then fragmented by physical shearing methods (analogous to passing the copied pages through a paper shredder). The random DNA fragments (typically 2–5 kb in size) are then subcloned. Sequence reads are then generated from one or both ends of randomly selected subclones (for example, several thousand sequence reads are typically generated from each ~100–150-kb BAC). The random reads are then assembled on the basis of sequence overlaps (see FIG. 4), yielding preliminary sequence assemblies (often referred to as prefinished sequence). Such sequence is imperfect, being associated with both gaps (indicated by breaks between the horizontal lines) and areas of poor sequence quality (indicated by the thinner horizontal lines). Often, the order and orientation of some of the sequence contigs is also not known. Subsequent customized sequence finishing, involving the generation of additional sequence data for closing gaps and bolstering areas of poor sequence quality, yields finished, highly accurate sequence across the entire clone.

RADIATION HYBRID MAP
Physical map of markers (typically STSs) positioned on the basis of the frequency with which they are separated by radiation-induced breaks (map construction involves the PCR analysis of rodent cell lines, each containing different fragments of the source genome).

such programs analyse the assemblies after the random shotgun phase and recommend specific additional sequencing reactions, thereby eliminating some of the labour-intensive (mostly computational) work involved in the early stages of sequence finishing. The net effect can be a significant increase in the efficiency and a decrease in the overall cost associated with this phase of the shotgun-sequencing process.

Sequence authentication. After final assembly, the finished sequence is typically analysed for several important features, such as the presence and correct order of known sequence-based markers (for example, STSs, genetic markers and genes) and its concordance with established restriction enzyme digest-based fingerprint(s). The latter involves comparing an *in silico* restriction digest(s) of the assembled sequence with the previously established fingerprint of the starting clone, with any discrepancy indicating the possible presence of a sequence misassembly or a rearrangement of the cloned insert before sequencing. Such authentication checks are crucial for ensuring that the final sequence produced in a shotgun-sequencing project is highly accurate.

Clone-by-clone shotgun-sequencing projects. The general process of clone-by-clone shotgun sequencing, as detailed above, has proven to be highly robust in several of the large genome-sequencing projects (TABLE 1). The sequencing of the yeast *S. cerevisiae*[40], the nematode *C. elegans*[41], and the plant *A. thaliana*[68] genomes was done in a clone-by-clone fashion, using previously constructed clone-based physical maps[57–61,69]. The most notable use of this strategy has been in the sequencing of the human genome by the Human Genome Project[70]. In this case, a carefully crafted plan for clone-by-clone shotgun sequencing was formulated based on the early experiences gained in a series of pilot projects[71,72]. The execution of this plan relied extensively on a newly constructed BAC-based physical map[62] and a mechanism for coordinating the activities of several sequencing groups around the world[73]. So far, a working draft sequence has been generated for virtually all of the readily clonable human genome[15], with finished sequence now available for roughly half of the genome (including some entire chromosomes[74,75]). This sequence is extensively integrated with various other maps of the human genome, including cytogenetic[76], genetic[77], STS[49,51,52], gene/transcript[78] and RADIATION HYBRID MAPS[79]. Importantly, the high standards for accuracy established for the human genome sequence[21] are seemingly being met[22].

In addition to the traditional implementation of clone-by-clone shotgun sequencing (TABLE 1), alternative approaches for the selection of clones for sequencing have been proposed. For example, as a substitute or a supplement to the construction of a BAC contig map, large collections of BAC insert-end sequences can be generated. In principle, such 'BAC-end sequences' can be used to identify clones that minimally overlap with an already sequenced clone (specifically, by matching unique BAC-end sequences near the insert end of the sequenced clone[80]). When used in conjunction with available BAC-fingerprint data, such a strategy provides a potential route for the rational selection of clones for sequencing[81]. Another variant strategy involves starting a genome-sequencing project by initially sequencing numerous randomly selected BACs[82] and then eventually switching to a map-based approach for clone selection. Such a strategy has been advocated for cases in which no clone-based physical map is available at the time that sequencing is ready to begin.

Figure 4 | **Shotgun-sequence assembly.** A key component of shotgun sequencing involves the computational assembly of redundant collections of sequence reads, from which an accurate consensus sequence can be deduced. Shown here is a representative display from the program Consed[18], which is the viewing tool typically used with the associated base-calling program Phred[16,17] and sequence-assembly program Phrap. In this ~60-bp view from a shotgun-sequencing project, the indicated sequence reads have been assembled and used to deduce the consensus sequence (highlighted in white along the top). The primary data for one of the sequence reads (the sequence trace) is shown along the bottom. See REF. 18 for additional details about the various features of Consed shown here.

Whole-genome shotgun sequencing

An alternative strategy for genome sequencing, called whole-genome shotgun sequencing, involves the assembly of sequence reads generated in a random, genome-wide fashion (FIG. 1b), theoretically bypassing the need for a clone-based physical map. Specifically, the entire genome of an organism is fragmented into pieces of defined sizes, which in turn are subcloned into suitable plasmid vectors. Sequence reads are generated from both insert ends of a very large number of subclones, so as to produce highly redundant sequence coverage across the genome. Computational methods are then used to assemble the sequence reads and to deduce a corresponding consensus sequence. A key aspect of this strategy, which is especially important for dealing with the problems presented by repetitive sequences, is the generation of sequence reads from both ends of most subclones[65]. The expected physical distances separating these juxtaposed read pairs are an important factor in the process of deriving an accurate sequence assembly.

The most common implementation of whole-genome shotgun sequencing has been for elucidating the sequence of smaller bacterial genomes[83], as first shown for the landmark sequencing of the *Haemophilus influenzae* genome[84]. Indeed, this strategy is now routinely used for sequencing repeat-poor microbial genomes, with numerous examples reported in recent years[85].

The application of whole-genome shotgun sequencing to eukaryotic genomes is more difficult owing to their larger size and higher repeat content. In particular, repetitive sequences pose significant challenges to the process of sequence assembly, especially when carried out using data derived in a genome-wide fashion. The sequencing of the *Drosophila* genome represented the first large-scale use of whole-genome shotgun sequencing for the comprehensive analysis of a eukaryotic genome[86,87]. This effort provided valuable insight about the use of this strategy for sequencing eukaryotic genomes, in this case resulting in the assembly of virtually all of the euchromatic portion of the *Drosophila* genome (TABLE 1). It is important to note, though, that the finishing of the *Drosophila* genome is proceeding in a BAC-by-BAC fashion using an established clone-based physical map[88]; so, the *Drosophila* genome will ultimately be sequenced by a hybrid strategy (see below).

The most prominently featured application of whole-genome shotgun sequencing has been for analysing the human genome. Lively discussions about the potential use of this strategy for sequencing large, repeat-rich mammalian genomes (for example, the human genome) began several years ago[89,90], even before its use for sequencing the *Drosophila* genome. Using a modified version of an earlier plan[91], a whole-genome shotgun-sequencing strategy was used by Celera Genomics to sequence the human genome[92]. Specifically, this involved: first, generating an approximately fivefold coverage of the human genome in sequence reads derived from three types of subclone libraries (containing ~2-, ~10- and ~50-kb inserts); and second, augmenting this data with an approximately threefold coverage in sequence reads artificially created by shredding available human-sequence assemblies produced by the Human Genome Project[15] (which was generated in a BAC-by-BAC fashion using mapped clones). The combined data sets were used to carry out two types of sequence assembly, in each case involving a sophisticated, multistep algorithm[92]. The resulting sequence contigs were then organized into scaffolds (FIG. 5), each consisting of a group of sequence contigs held together by read pairs (in each case, with one read assembled into one contig and the other read assembled into a nearby contig). The scaffolds were then aligned relative to the genome by the use of long-range mapping information, most often by the identification of mapped STSs, genetic markers and genes in the assembled sequence contigs.

These initial experiences in applying a whole-genome shotgun-sequencing strategy to the analysis of eukaryotic genomes have yielded important lessons. First, it is now appreciated that the use of several size classes of subclones is crucial. Each class has a slightly different role in the assembly process, such as spanning

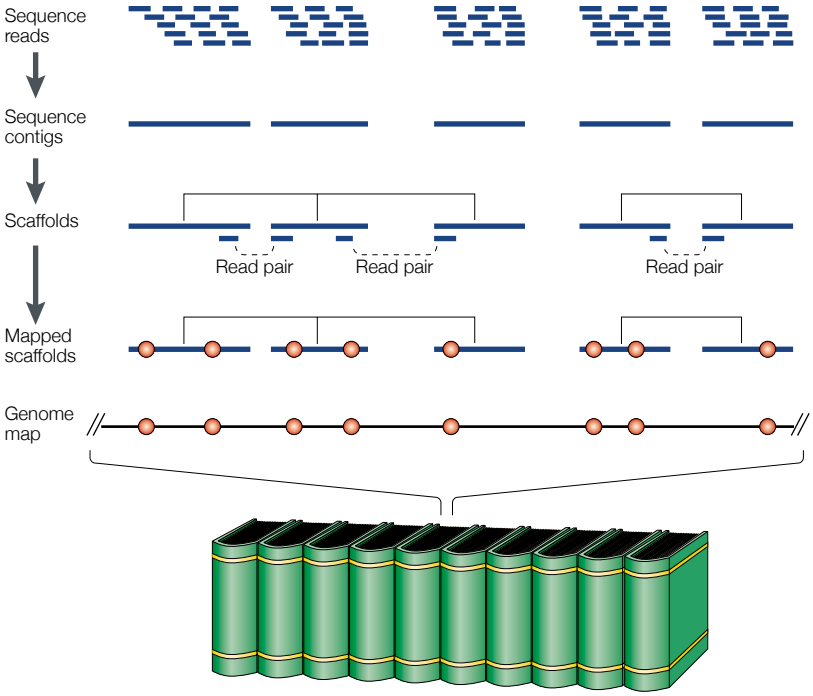

Figure 5 | **Long-range sequence assembly in whole-genome shotgun sequencing.** Individual sequence reads generated in a whole-genome shotgun-sequencing project are initially assembled into sequence contigs. Groups of sequence contigs are then organized into scaffolds on the basis of linking information provided by read pairs (in each case, with one sequence read from a pair assembling into one contig and the other read into another contig). In turn, the scaffolds can be aligned relative to the source genome (represented by an encyclopaedia set) by the identification of already mapped, sequence-based landmarks (for example, STSs, genetic markers and genes; depicted as red circles) in the sequence contigs, thereby associating them with a known location on the genome map. Adapted from REF. 92.

certain blocks of repetitive sequence or providing a long-range organizational framework for the sequence contigs and scaffolds. Typically, the largest numbers of sequence reads are generated with the subclones that contain the smallest inserts. Second, the availability of supplementary long-range mapping data is crucially important. Indeed, assimilating the large collection of assembled sequence contigs into an organized, useful resource requires their accurate cross-referencing with available mapping information (FIG. 5). Finally, perhaps the most essential element of any whole-genome shotgun-sequencing strategy is the availability of a robust assembly program that can accommodate the inevitably large collection of sequence reads[87,92]. Such software systems must assemble sequence reads derived from several gigabase pairs of genomic DNA, in contrast to the more traditional clone-based sequence assembly programs that only encounter data sets derived from 100–200-kb clones. The former include algorithms that account for the anticipated spatial relationship of read pairs emanating from individual subclones, which help to avoid misassemblies due to repetitive sequences. The development of more powerful whole-genome sequence assembly programs, including those that use data produced in both whole-genome and clone-by-clone shotgun-sequencing projects, represents an active area of genome research.

The above experiences with whole-genome shotgun sequencing have also revealed important advantages and disadvantages of the approach. The former include the ability to initiate the sequencing of a genome without an existing clone-based physical map and to generate large amounts of sequence data from an entire genome relatively quickly for use in identifying conserved sequences[93] and polymorphisms[92,94]. The main problems with the strategy, especially when used in the absence of supplementary clone-derived sequence data, include the sequence gaps and misassemblies that are caused by repetitive sequences (particularly those associated with low-copy duplicated segments) and the uncertainty about how best to proceed from data generated exclusively by whole-genome shotgun sequencing to a highly accurate, finished sequence of an entire genome.

Hybrid strategies for shotgun sequencing

The scientific and popular press has frequently highlighted the distinct approaches taken by the Human Genome Project and Celera Genomics in their respective efforts to sequence the human genome (that is, clone-by-clone versus whole-genome shotgun sequencing, respectively). In reality, the two strategies are not mutually exclusive and, in fact, are quite complementary. Indeed, there has been remarkable convergence in the use of these sequencing approaches, resulting in the advent of hybrid (or mixed) strategies that incorporate elements of both.

In a hybrid shotgun-sequencing strategy, sequence reads are generated in both a clone-by-clone and a whole-genome fashion (FIG. 6a). The sequence reads from individual BACs are then used to identify additional matching reads generated by whole-genome shotgun sequencing. The combined collection of reads, which essentially reflect sequencing data for a BAC-sized bin of the genome, is then used for sequence assembly and finishing. Note that this is precisely the strategy being used for finishing the *Drosophila* genome sequence[86–88].

A hybrid shotgun-sequencing strategy can, in principle, capture the advantageous elements of both clone-by-clone and whole-genome approaches. For example, the whole-genome shotgun component provides rapid insight about the sequence of the entire genome. Such data can be used to learn about the repertoire of repetitive sequences in a genome and to find matching conserved sequences in common with other organisms[93], as is now being done with mouse whole-genome shotgun-sequence data for the purpose of annotating the human genome sequence. At the same time, the whole-genome shotgun sequence reads are also useful when pooled with the corresponding data generated in a clone-by-clone fashion, thereby creating data sets with enhanced sequence coverage. Meanwhile, the clone-by-clone component simplifies the process of sequence assembly to individual clone-sized genomic segments, thereby minimizing the likelihood of serious misassemblies. Importantly, this provides a well-established path for sequence finishing.

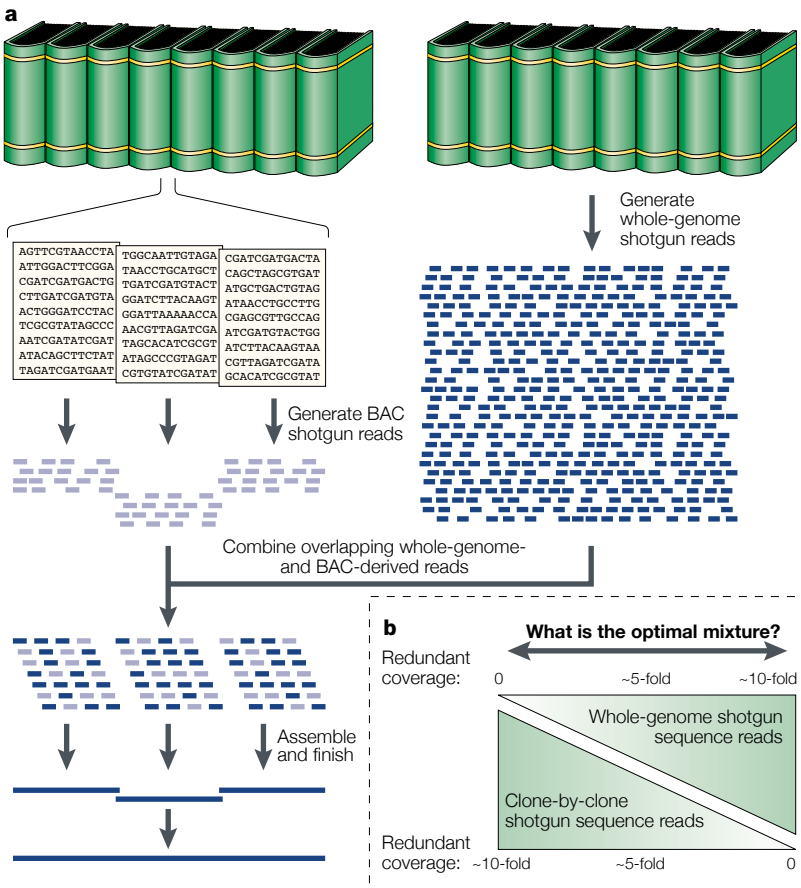

Figure 6 | **Hybrid shotgun-sequencing approach. a** | In a hybrid approach, elements of both clone-by-clone and whole-genome shotgun sequencing are amalgamated. A subclone library from a whole genome (represented by an encyclopaedia set) is prepared, and numerous sequence reads (depicted in dark blue) are generated in a genome-wide fashion. Meanwhile, individual mapped BACs are also subjected to shotgun sequencing. The BAC-derived sequence reads (depicted in light blue) can then be used to identify overlapping sequences in the larger collection of whole-genome-derived sequence reads, in essence reducing the complexity of the whole-genome shotgun data set to a series of individual BAC-sized bins. The combined set of sequence reads for each BAC can then be individually assembled and subjected to sequence finishing. **b** | Optimal balance of generating clone-by-clone versus whole-genome shotgun sequence reads in a hybrid sequencing strategy. In the hybrid shotgun-sequencing approach illustrated in **a**, sequence reads derived from individual clones and those generated in a genome-wide fashion are assembled together. Although an overall sequence redundancy of ~8–10-fold is typically desired, the optimal mixture of clone-by-clone and whole-genome shotgun sequence reads is at present not well established. Current projects that are using a hybrid approach for sequencing the mouse, rat and zebrafish genomes (see TABLE 1) should provide valuable insight into this issue.

A current research focus of several sequencing groups involves establishing the optimal balance between generating sequence reads in a clone-by-clone versus whole-genome fashion when implementing a hybrid shotgun-sequencing strategy. Although there is general consensus that ~8–10-fold redundant coverage is required for a project that aims to deliver high-quality, finished sequence, the relative amount of coverage contributed by the reads derived from a clone-by-clone versus whole-genome component is actively being investigated (FIG. 6b). In particular, the current efforts to sequence the mouse, rat and zebrafish genomes (TABLE 1), all of which are using a hybrid sequencing strategy, should provide crucial insight into this issue.

Future genome sequencing efforts

Buoyed by the information gleaned from the genome sequences now available for the human and several other model organisms, there is ever-increasing interest in acquiring sequence data from myriad other genomes. In contemplating additional large-scale sequencing projects, significant consideration is rightfully being given to the type(s) of information desired in each case. The latter can include specific biological insight into the genetic blueprint of an organism, data for use as a comparative tool to analyse an already available genome sequence (for example, that of the human) and a general cataloguing of genes, repetitive sequences, conserved elements or polymorphisms in a genome. Another important consideration relates to the degree of completeness, accuracy and continuity desired and/or required for the intended use of the resulting sequencing data. Although each of the current and future genome-sequencing projects has its own characteristic features, some generalizations about sequencing strategies can be made. For example, a whole-genome shotgun-sequencing strategy is particularly appropriate for studying a genome in a global and survey-orientated fashion, especially when a finished genome sequence is not an immediate goal. By contrast, clone-by-clone and hybrid shotgun-sequencing strategies are typically used when a more detailed, highly accurate (most likely finished) genome sequence is desired.

The mouse, rat and zebrafish genomes are expected to be sequenced over the next few years (TABLE 1). Smaller efforts, at present mostly involving whole-genome shotgun sequencing, aim to sequence the compact genomes of two pufferfish species: *Tetraodon nigroviridis*[95] and *Fugu rubripes*[96]. Various efforts are also sequencing the genomes of several plants, such as rice (TABLE 1). Meanwhile, the selection and prioritization of other genomes for systematic sequencing is an active process, in many cases associated with vigorous debate (such as in contemplating the sequencing of the chimpanzee genome[97–99]). The generation of additional genome sequences should be accompanied by improved insight into the level of completeness and accuracy needed, especially for analyses that will be able to capitalize on already available, high-quality genome sequences.

Although considerably lower than a decade ago, the current costs associated with large-scale sequencing remain a central limiting factor for genome-sequencing efforts. For example, establishing approximately onefold sequence coverage of a mammalian genome requires the generation of roughly 6 million sequence reads and at present costs about US $10–20 million (and so, even a working draft sequence with approximately fivefold coverage would probably cost in excess of US $50 million). As a result, only a handful of different genomes can realistically be sequenced in a comprehensive fashion at any point in time, at least by the available methods.

An alternative approach for comparative genome exploration involves the sequencing of well-defined, delimited genomic regions. Such targeted sequencing,

which typically involves the shotgun sequencing of mapped BACs, provides the ability to directly compare orthologous sequences generated from numerous different organisms. So, in addition to comprehensive genome sequencing for a select set of organisms (TABLE 1), the near future should also see the generation of sequence from the same targeted genomic regions of an even larger set of organisms that will be used for detailed comparative sequence analysis.

Conclusions

The experience so far in analysing the various, recently generated genome sequences has led to one definitive conclusion — that the availability of genome-sequence data has a profound and positive impact on the infrastructure of experimental biology. As such, despite the impressive sequencing triumphs of late, the thirst for genome-sequence data is far from quenched. The Human Genome Project, and related efforts in genome analysis, have brought tremendous advances in the methods and technologies available for sequencing complex genomes. Having recruited creative and ambitious investigators with diverse backgrounds, the field of large-scale sequencing now represents an exciting and ever-challenging area of genome science that continues to attract significant interest and attention. The future should bring a steady stream of additional sequencing efforts and with these will come a continued maturation of the portfolio of strategic alternatives for sequence-based genome exploration.

Links

FURTHER INFORMATION Human Genome Project | *Saccharomyces cerevisiae* | *Caenorhabditis elegans* | *Drosophila melanogaster* | *Arabidopsis thaliana* | *Homo sapiens* | *Escherichia coli* | Phred | Phrap | Consed | GAP | mouse | BAC fingerprint map of the mouse genome | rat | zebrafish | TIGR comprehensive microbial resource | Celera Genomics | *Tetraodon nigroviridis* | *Fugu rubripes* | rice

1. Green, E. D. in *The Metabolic and Molecular Bases of Inherited Disease* (eds Scriver, C. R. *et al.*) 259–298 (McGraw–Hill, New York, 2001).
2. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74**, 5463–5467 (1977).
Reports the Nobel prize-winning method developed by Fred Sanger and colleagues for sequencing DNA — called dideoxy chain termination sequencing. Roughly 25 years later, this continues to be the state-of-the-art technique for large-scale DNA sequencing.
3. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986).
First notable description of fluorescence-based DNA sequencing and the use of automated instrumentation for detecting the sequencing reaction products.
4. Hunkapiller, T., Kaiser, R. J., Koop, B. F. & Hood, L. Large-scale and automated DNA sequence determination. *Science* **254**, 59–67 (1991).
5. Mullikin, J. C. & McMurray, A. A. Sequencing the genome, fast. *Science* **283**, 1867–1868 (1999).
6. Meldrum, D. R. Sequencing genomes and beyond. *Science* **292**, 515–516 (2001).
7. Tabor, S. & Richardson, C. C. A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc. Natl Acad. Sci. USA* **92**, 6339–6343 (1995).
8. Prober, J. M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**, 336–341 (1987).
9. Ju, J., Ruan, C., Fuller, C. W., Glazer, A. N. & Mathies, R. A. Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis. *Proc. Natl Acad. Sci. USA* **92**, 4347–4351 (1995).
10. Rosenblum, B. B. *et al.* New dye-labeled terminators for improved DNA sequencing patterns. *Nucleic Acids Res.* **25**, 4500–4504 (1997).
11. Metzker, M. L., Lu, J. & Gibbs, R. A. Electrophoretically uniform fluorescent dyes for automated DNA sequencing. *Science* **271**, 1420–1422 (1996).
12. Lee, L. G. *et al.* New energy transfer dyes for DNA sequencing. *Nucleic Acids Res.* **25**, 2816–2822 (1997).
13. Meldrum, D. Automation for genomics. I. Preparation for sequencing. *Genome Res.* **10**, 1081–1092 (2000).
14. Meldrum, D. Automation for genomics. II. Sequencers, microarrays, and future trends. *Genome Res.* **10**, 1288–1303 (2000).
15. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
Landmark paper about the initial sequence of the human genome generated by the public Human Genome Project using a clone-by-clone shotgun-sequencing strategy.
16. 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).
17. Ewing, B. & Green, P. Base-calling of automated sequencer traces using *Phred.* II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
18. Gordon, D., Abajian, C. & Green, P. *Consed*: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).
The most commonly used suite of computer programs for carrying out base calling, sequence assembly and viewing of sequence assemblies are Phred (references 16 and 17), Phrap and Consed (reference 18), respectively. Reference 20 describes an important extension of Consed (a program called Autofinish) that automates some of the key steps in sequence finishing.
19. Bonfield, J. K., Smith, K. F. & Staden, R. A new DNA sequence assembly program. *Nucleic Acids Res.* **23**, 4992–4999 (1995).
20. Gordon, D., Desmarais, C. & Green, P. Automated finishing with Autofinish. *Genome Res.* **11**, 614–625 (2001).
21. Olson, M. & Green, P. A 'quality-first' credo for the Human Genome Project. *Genome Res.* **8**, 414–415 (1998).
22. Felsenfeld, A., Peterson, J., Schloss, J. & Guyer, M. Assessing the quality of the DNA sequence from The Human Genome Project. *Genome Res.* **9**, 1–4 (1999).
23. Huang, G. M. High-throughput DNA sequencing: a genomic data manufacturing process. *DNA Seq.* **10**, 149–153 (1999).
24. Wendl, M. C., Dear, S., Hodgson, D. & Hillier, L. Automated sequence preprocessing in a large-scale sequencing environment. *Genome Res.* **8**, 975–984 (1998).
25. Dedhia, N. N. & McCombie, W. R. Kaleidaseq: a web-based tool to monitor data flow in a high throughput sequencing facility. *Genome Res.* **8**, 313–318 (1998).
26. Lawrence, C. B. *et al.* The Genome Reconstruction Manager: a software environment for supporting high-throughput DNA sequencing. *Genomics* **23**, 192–201 (1994).
27. Kimmel, B. E., Palazzolo, M. J., Martin, C. H., Boeke, J. D. & Devine, S. E. in *Genome Analysis: A Laboratory Manual. 1. Analyzing DNA* (eds Birren, B. *et al.*) 455–532 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1997).
28. Church, G. M. & Kieffer-Higgins, S. Multiplex DNA sequencing. *Science* **240**, 185–188 (1988).
29. Cherry, J. L. *et al.* Enzyme-linked fluorescent detection for automated multiplex DNA sequencing. *Genomics* **20**, 68–74 (1994).
30. Smith, D. R. *et al.* Multiplex sequencing of 1.5 Mb of the *Mycobacterium leprae* genome. *Genome Res.* **7**, 802–819 (1997).
31. Gardner, R. C. *et al.* The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. *Nucleic Acids Res.* **9**, 2871–2888 (1981).
32. Anderson, S. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.* **10**, 3015–3027 (1981).
33. Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide sequence of the bacteriophage lambda DNA. *J. Mol. Biol.* **162**, 729–773 (1982).
References 31–33 represent some of the earliest papers that reported the use of shotgun sequencing as a strategy for establishing the sequence of large pieces of DNA.
34. Deininger, P. L. Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. *Anal. Biochem.* **129**, 216–223 (1983).
35. Messing, J. The universal primers and the shotgun DNA sequencing method. *Methods Mol. Biol.* **167**, 13–31 (2001).
36. Ansorge, W., Voss, H. & Zimmermann, J. (eds) *DNA Sequencing Strategies* (Wiley & Sons, Inc., New York, 1997).
37. Adams, M. D., Fields, C. & Venter, J. C. (eds) *Automated DNA Sequencing and Analysis* (Academic, Inc., San Diego, 1994).
38. Spurr, N. K., Young, B. D. & Bryant, S. P. (eds) *ICRF Handbook of Genome Analysis* (Blackwell Science Ltd, Oxford, 1998).
39. Green, E. D., Birren, B., Klapholz, S., Myers, R. M. & Hieter, P. (eds) *Genome Analysis: A Laboratory Manual* Vols 1–4 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1997).
40. Goffeau, A. *et al.* The yeast genome directory. *Nature* **387**, S1–S105 (1997).
Describes the genome sequence of the first eukaryotic organism, the yeast *Saccharomyces cerevisiae*, by a collection of numerous sequencing groups (large and small) around the world.
41. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
Reports the genome sequence of the first multicellular organism, the nematode worm *Caenorhabditis elegans*, by the sequencing groups at Washington University and the Sanger Centre.
42. Wilson, R. K. & Mardis, E. R. in *Genome Analysis: A Laboratory Manual. 1. Analyzing DNA* (eds Birren, B. *et al.*) 397–454 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1997).

43. Olson, M., Hood, L., Cantor, C. & Botstein, D. A common language for physical mapping of the human genome. *Science* **245**, 1434–1435 (1989).
44. Vollrath, D. in *Genome Analysis: A Laboratory Manual. 4. Mapping Genomes* (eds Birren, B. *et al.*) 187–215 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1999).
45. Burke, D. T., Carle, G. F. & Olson, M. V. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236**, 806–812 (1987).
46. Green, E. D., Hieter, P. & Spencer, F. A. in *Genome Analysis: A Laboratory Manual. 3. Cloning Systems* (eds Birren, B. *et al.*) 297–565 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1998).
47. Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl Acad. Sci. USA* **89**, 8794–8797 (1992).
48. Ioannou, P. A. *et al.* A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nature Genet.* **6**, 84–89 (1994).
49. Hudson, T. J. *et al.* An STS-based map of the human genome. *Science* **270**, 1945–1954 (1995).
50. Chumakov, I. M. *et al.* A YAC contig map of the human genome. *Nature* **377**, 175–297 (1995).
51. Bouffard, G. G. *et al.* A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. *Genome Res.* **7**, 673–692 (1997).
52. Nagaraja, R. *et al.* X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. *Genome Res.* **7**, 210–222 (1997).
53. Nusbaum, C. *et al.* A YAC-based physical map of the mouse genome. *Nature Genet.* **22**, 388–393 (1999).
54. Marra, M. A. *et al.* High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7**, 1072–1084 (1997).
Approach for constructing sequence-ready BAC contig maps by restriction enzyme digest-based fingerprint analysis. This general method, which essentially represents an extension of earlier mapping techniques (for example, see references 57–59), has been used to generate BAC contig maps of the human, mouse, *Arabidopsis thaliana* and other genomes.
55. Gregory, S. G., Howell, G. R. & Bentley, D. R. Genome mapping by fluorescent fingerprinting. *Genome Res.* **7**, 1162–1168 (1997).
56. Kohara, Y., Akiyama, K. & Isono, K. The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50**, 495–508 (1987).
57. Olson, M. V. *et al.* Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl Acad. Sci. USA* **83**, 7826–7830 (1986).
58. Riles, L. *et al.* Physical maps of the six smallest chromosomes of *Saccharomyces cerevisiae* at a resolution of 2.6 kilobase pairs. *Genetics* **134**, 81–150 (1993).
59. Coulson, A., Sulston, J., Brenner, S. & Karn, J. Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **83**, 7821–7825 (1986).
References 57–59 represent classic descriptions of restriction enzyme digest-based fingerprint analysis, as used to construct physical maps of the *Saccharomyces cerevisiae* and *Caenorhabditis elegans* genomes. In both cases, the resulting maps paved the way towards the sequencing of these genomes, as well as provided key insight into the strategies required for mapping and sequencing the human genome.
60. Marra, M. *et al.* A map for sequence analysis of the *Arabidopsis thaliana* genome. *Nature Genet.* **22**, 265–270 (1999).
61. Mozo, T. *et al.* A complete BAC-based physical map of the *Arabidopsis thaliana* genome. *Nature Genet.* **22**, 271–275 (1999).
62. The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409**, 934–941 (2001).
Paper reporting the BAC-based physical map of the human genome constructed by the Human Genome Project.
63. Green, E. D. & Olson, M. V. Chromosomal region of the cystic fibrosis gene in yeast artificial chromosomes: a model for human genome mapping. *Science* **250**, 94–98 (1990).
64. McPherson, J. D. Sequence ready — or not? *Genome Res.* **7**, 1111–1113 (1997).
65. Edwards, A. *et al.* Automated DNA sequencing of the human HPRT locus. *Genomics* **6**, 593–608 (1990).
66. Chissoe, S. L. *et al.* Representation of cloned genomic sequences in two sequencing vectors: correlation of DNA sequence and subclone distribution. *Nucleic Acids Res.* **25**, 2960–2966 (1997).
67. Bouck, J., Miller, W., Gorrell, J. H., Muzny, D. & Gibbs, R. A. Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Res.* **8**, 1074–1084 (1998).
68. The *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
Describes the genome sequence of the first plant, *Arabidopsis thaliana*, by an international consortium of sequencing groups.
69. Coulson, A., Waterston, R., Kiff, J., Sulston, J. & Kohara, Y. Genome linking with yeast artificial chromosomes. *Nature* **335**, 184–186 (1988).
70. Bentley, D. R. Decoding the human genome sequence. *Hum. Mol. Genet.* **9**, 2353–2358 (2000).
71. Waterston, R. & Sulston, J. E. The Human Genome Project: reaching the finish line. *Science* **282**, 53–54 (1998).
72. The Sanger Centre & The Washington University Genome Sequencing Center. Toward a complete human genome sequence. *Genome Res.* **8**, 1097–1108 (1998).
73. Bentley, D. R., Pruitt, K. D., Deloukas, P., Schuler, G. D. & Ostell, J. Coordination of human genome sequencing via a consensus framework map. *Trends Genet.* **14**, 381–384 (1998).
74. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
75. The Chromosome 21 Mapping and Sequencing Consortium. The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
References 74 and 75 announce the completion of finished sequence for the first two human chromosomes — 22 and 21, respectively.
76. The BAC Resource Consortium. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409**, 953–958 (2001).
77. Yu, A. *et al.* Comparison of human genetic and sequence-based physical maps. *Nature* **409**, 951–953 (2001).
78. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
79. Olivier, M. *et al.* A high-resolution radiation hybrid map of the human genome draft sequence. *Science* **291**, 1298–1302 (2001).
80. Venter, J. C., Smith, H. O. & Hood, L. A new strategy for genome sequencing. *Nature* **381**, 364–366 (1996).
81. Mahairas, G. G. *et al.* Sequence-tagged connectors: a sequence approach to mapping and scanning the human genome. *Proc. Natl Acad. Sci. USA* **96**, 9739–9744 (1999).
82. Wendl, M. C. *et al.* Theories and applications for sequencing randomly selected clones. *Genome Res.* **11**, 274–280 (2001).
83. Fraser, C. M. & Fleischmann, R. D. Strategies for whole microbial genome sequencing and analysis. *Electrophoresis* **18**, 1207–1216 (1997).
84. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
Publication reporting the genome sequence of the first prokaryotic organism, the bacterium *Haemophilus influenzae*. This effort was the first report of using a whole-genome shotgun-sequencing strategy to sequence the genome of a free-living organism.
85. Fraser, C. M., Eisen, J. A. & Salzberg, S. L. Microbial genome sequencing. *Nature* **406**, 799–803 (2000).
86. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
87. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000).
References 86 and 87 report the initial genome sequence of *Drosophila melanogaster* generated by a hybrid strategy that involved both whole-genome shotgun sequencing and clone-by-clone shotgun sequencing. This project reflected a collaboration between the public Human Genome Project and Celera Genomics.
88. Hoskins, R. A. *et al.* A BAC-based physical map of the major autosomes of *Drosophila melanogaster*. *Science* **287**, 2271–2274 (2000).
89. Weber, J. L. & Myers, E. W. Human whole-genome shotgun sequencing. *Genome Res.* **7**, 401–409 (1997).
90. Green, P. Against a whole-genome shotgun. *Genome Res.* **7**, 410–417 (1997).
References 89 and 90 provide point/counter-point perspectives that detail the opposing views on the use of a whole-genome shotgun-sequencing strategy for sequencing the human genome.
91. Venter, J. C. *et al.* Shotgun sequencing of the human genome. *Science* **280**, 1540–1542 (1998).
92. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
Landmark paper reporting the initial sequence of the human genome generated by Celera Genomics using a whole-genome shotgun-sequencing strategy in conjunction with available clone-by-clone data provided by the Human Genome Project.
93. Bouck, J. B., Metzker, M. L. & Gibbs, R. A. Shotgun sample sequence comparisons between mouse and human genomes. *Nature Genet.* **25**, 31–33 (2000).
94. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
95. Crollius, H. R. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25**, 235–238 (2000).
96. Brenner, S. *et al.* Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* **366**, 265–268 (1993).
97. McConkey, E. H. & Varki, A. A primate genome project deserves high priority. *Science* **289**, 1295–1296 (2000).
98. Varki, A. A chimpanzee genome project is a biomedical imperative. *Genome Res.* **10**, 1065–1070 (2000).
99. VandeBerg, J. L., Williams-Blangero, S., Dyke, B. & Rogers, J. Examining priorities for a primate genome project. *Science* **290**, 1504–1505 (2000).
100. Soderlund, C., Longden, I. & Mott, R. FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.* **13**, 523–535 (1997).
101. Soderlund, C., Humphray, S., Dunham, A. & French, L. Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res.* **10**, 1772–1787 (2000).
102. Sasaki, T. & Burr, B. International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. *Curr. Opin. Plant Biol.* **3**, 138–141 (2000).

Acknowledgements

I thank F. Collins, J. Touchman and R. Wilson for critical reading of this manuscript.