UNIVERSITY OF WATERLOO
Faculty of Engineering

Sequencing the Human Genome: A Look at the Human Genome Project and Celera Genomics

For

Professor Stashuk

SYDE 444

Prepared by

Zhan Huan Zhou

April 4, 2001

Table of Contents

List of Figures

List of Tables

1.0 Introduction

1.1 What is DNA?

Deoxyribonucleic acid, commonly known as DNA, contains the genetic information for higher life forms. James Watson and Francis Crick first correctly described its famous double-helix structure in 1953 [1]. The double-helix structure is similar to a ladder twisted upon itself. The "rungs" of the ladder are composed of one of two base pairs (bp): adenine-thymine (A-T) or, cytosine-guanine (C-G). The exact sequence of A, T, C, G directs the construction of proteins and ultimately such physical attributes as eye or hair colour. It can also determine susceptibility to genetic diseases such as cystic fibrosis. The human genome is composed of approximately 3 billion bases.

1.2 History of the Human Genome Project

When a plan to sequence the entire human genome was first proposed in 1985, it was met with much criticism. There was no doubt that the entire sequence was useful, what was in question was the cost. The entire project was estimated to consume roughly thirty thousand person years over fifteen years, costing about $3 billion [2]. After much debate, the publicly funded Human Genome Project (HGP) was jointly launched by the US Department of Energy and National Institutes of Health in 1990. The goal of the HGP was to complete a detailed map of the human genome by 2005. The map was to aid in finding genes that could be used to treat genetic diseases. As the project continued, other international centres joined the monumental task set out by the HGP.

By 1998, the HGP was over-budget and well behind its projected schedule [3]. On 9 May 1998, a private company, Celera Genomics, was formed with the goal of sequencing the entire human genome in just three years at a tenth the cost of the HGP using a radical new approach. This helped fuel a so-called race between the public and private sectors. This race accelerated the sequencing process on both sides. By June 2000, both teams jointly announced the completion of a rough draft of the human genome years ahead of schedule.

This paper describes the basics of DNA sequencing and how technology played an integral role in achieving such a milestone ahead of schedule. A brief description of the methods used by the HGP is followed by a detailed look at the methods employed at Celera Genomics to generate the draft genome. The following discussion assumes the reader has a basic understanding of genetics and biochemistry.

2.0 Sequencing Techniques

Sequencing of DNA requires the following key steps: the preparation and replication of short segments of DNA; the creation of partial copies of the segments each one base longer than the next; identification of the last base of each copy; and ordering of the bases [4].

Short segments of DNA are created by fracturing a source strand with sound (sonification) or passing it through a nozzle under pressure (nebulation) [5]. These short DNA segments are inserted in a vector, typically a bacterial virus (phage). The virus then infects a bacterium with the DNA segment, also known as an insert. The insert is now part of the Bacteria Artificial Chromosome (BAC). Typical insert sizes are 50 000 to 300 000 base pairs (bp) [4].

To prepare for sequencing, the human DNA insert must be extracted from the BAC and then amplified. This is accomplished with a routine process known as polymerase chain reaction (PCR). For each sample to be sequenced, copies are made with each one varying in length by one base using restriction enzymes. The fragments are then labelled with one of four fluorescent dyes, corresponding to the last base.

Sequencers employ one of two strategies: slab-gel or the more advanced capillary electrophoresis. The Applied Biosystems Inc. (ABI) 377 is a slab-gel device while the ABI PRISM 3700 is capillary electrophoresis-based device. The theory of operation for both types of devices is essentially the same, but the exact mechanics differ.

In slab-gel devices, an electric field is applied across the gel matrix. Since DNA is a negatively charged molecule, it migrates across the electric field through the gel. However, smaller fragments move through the gel faster than larger fragments. The time of migration indicates the size of the fragments. As the fragments emerge, a laser causes the dye to fluoresce and the colour is detected optically by a charge-coupled device (CCD). The sequence of the DNA fragment can then constructed from the series of colours seen [4].

The main differences between slab-gel and capillary-based sequencers are the sequencing time and reliability. The ABI 377 requires 5 to 6 hours to sequence 500 bp while the ABI 3700 can accomplish the same task in 2 to 3 hours. The maximum read length of each device is approximately 600 bp. Reads of up to 1000 bp are possible, but it takes longer and the error rate increases substantially. The main disadvantage of the ABI 377 is that the slabs of gel have to be manually prepared when needed. Quality of the gels was difficult to maintain, varying from batch to batch and even with the time of year. Furthermore, the slabs contained 96 lanes allowing 96 samples to be loaded and sequenced. Since the lanes were close together, the optical software could get confused and take reads from adjacent lanes if the DNA fragment travelled outside its lane. By design, the capillary-based ABI 3700 does not face such lane-tracking difficulties [4].

The emergence of the ABI 3700 not only solved lane-tracking errors, but also ushered in sequencing automation. It can hold two 384-well plates, about a day's worth of analysis. Once loaded, a robotic arm takes samples two at a time and loads them until 96 samples are loaded. A voltage is applied to draw the samples along the capillaries. When the samples emerge from the 50 cm capillaries, they flow into a stream of polymer where a laser detects the 96 separate outputs. There are no lanes so there are no such tracking problems as in slab-gel devices [4].

Automated sequencing has greatly accelerated the sequencing phase. This is illustrated in Figure 1. Note the sharp increase in data in 1997 when the ABI 3700 was introduced. The facility at Celera employs ABI 3700s and has been running uninterrupted since May 1999 and can produce 175,000 reads per day [6]. At HGP facilities, the equivalent of one-fold coverage was produced every six weeks [7].

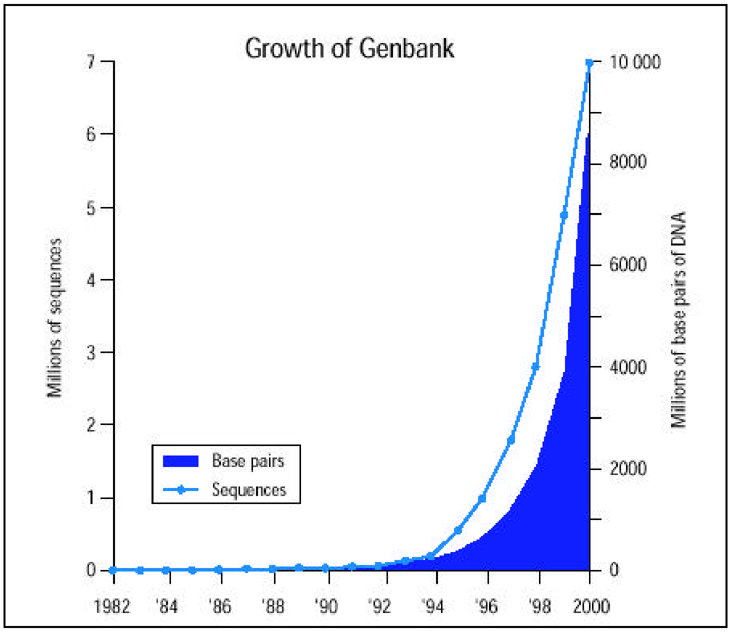

Figure 1: Growth of GenBank

Source: National Center for Biotechnology Information

3.0 Basic Shotgun Sequencing

To determine longer sequences of DNA, the shotgun sequencing strategy was introduced soon after the invention of DNA sequencing methods. First, fragments of the source sequence are randomly selected. The first 300 to 900 bases of one end of a fragment are then sequenced. If enough fragments are sequenced and the sampling is sufficiently random, it should be possible to reconstruct the source sequence from the overlapping fragments [5]. A simplified procedure is illustrated in Table 1 with read lengths of 4 bases. The first read overlaps with the known start sequence, highlighted in grey. The second read overlaps with the first read. This overlapping procedure continues until the entire source sequence is constructed from the set of random reads.

Table 1: Sample Shotgun Sequencing

Sequence	ATGCGATCAT…AGACAGTAAAGA
Read 1	ATGC
Read 2	GCGA
…	…
Read N-1	AAAG
Read N	AAGA

3.1 Challenges to Shotgun Sequencing

There are two key sources of failure for shotgun sequencing. The first is when the sampling is not entire randomly and the second is caused by repeat sequences in the genome.

Non-random sampling is quite common due to clone biasing. The most common form of clone biasing occurs when an insert/vector combination is unstable, possibly causing a toxic environment for the host/vector environment. This potential problem can be overcome by picking host/vector combinations where the insert DNA will produce a relatively inert reaction [5].

The second challenge to shotgun sequencing is the presence of repeats in the genome. This was not a problem with simpler prokaryotic organisms. However, higher-order eukaryotic organisms, such as humans, have a repeat-rich genome introducing a computational challenge. To solve this challenge, the nature of repeats in the genome must first be understood.

3.1.1 Repeat Sequences

Repeats occur on three levels in the human genome. First, there are large-scale repeats. For example, there is a five-fold repeat of a trypsinogen gene that is 4 kbp long and varies 3 to 5% between copies [5]. Three of the repeats are close enough that they appear in a single shotgun-sequenced insert [8]. This poses a problem because reads with unique portions outside of the repeat cannot span it. This makes it impossible to determine the correct sequence upon exiting the repeat as shown in Table 2. The highlighted string "GATTACA" is repeated in the sequence twice. It can be seen that read 1 can be uniquely placed. However it is impossible to determine which of read 2 or read 3 is correct since the correct sequence is not known *a priori*.

Table 2: Shotgun Sequencing With Repeats

Sequence	ATCG**GATTACA**AAAGGG**GATTACA**GGGAAA
Read 1	TCG**GATT**
Read 2	**TTACA**GGG (incorrect)
Read 3	**TTACA**AAA (correct)

Second, smaller repeat elements of about 300 bp exist. Even though the repeat sequence can be spanned, they are still problematic because they cluster and can represent up to 60% of the source sequence, with copies varying from 5 to 15% [9][10]. Finally, there are microsatellite repeats of the form x^n near the centromeres and telomeres [9]. The repeated "satellite" x is three to six bases long, n is very large, and has a variation of 1 to 2%. It is estimated that the human genome contains roughly 10% repeated Alu elements, 5% LINE (long interspersed nucleotide elements), and about 25% repeat of genes [5]. The impact of repeats must be carefully accounted for in shotgun assembly algorithms.

3.2 Double-barrelled Shotgun Sequencing

A variation of the shotgun approach involves sequencing an insert from both ends, producing a pair of reads, known as mates. These mate-pairs are in opposite orientation and separated by a known distance. The information contained in these mate pairs can help with the assembly of large uninterrupted stretches of DNA sequence (contigs). For instance, if a read in one contig has a mate in another contig, the relative spacing and orientation of the contigs can be determined. These set of arranged contigs form a scaffold. Unlike contigs, scaffolds are not contiguous, but have gaps. Using mating information, however, the sizes of these gaps are known. With 7.5-fold coverage of the genome, simulations show contigs to have an average length of 66 kbp and gaps of 66 bp [5]. The gaps can then be filled using PCR. The use of double-barrelled shotgun sequencing to produce contigs and scaffolds is shown below in Figure 2.

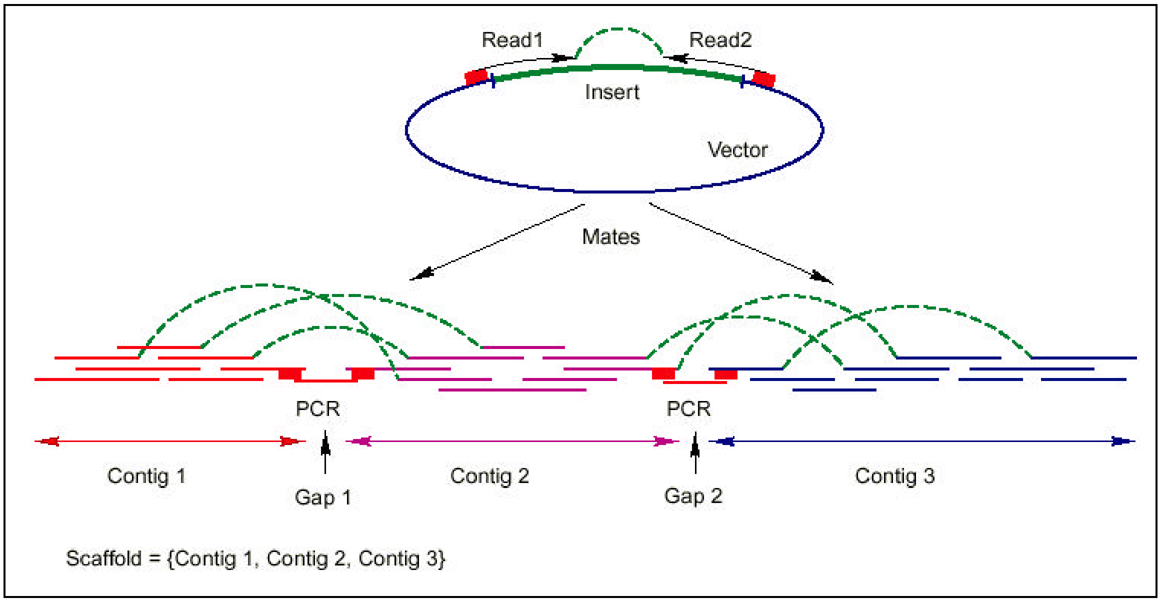

Figure 2: Double-Barrelled Shotgun Sequencing

Source [5].

In addition to building scaffolds, mating information can be used to resolve repeats. For the example shown in Table 2, it was impossible to determine which of read 2 or read 3 was correct. However, if a mate-pair spans the repeat, it is possible to determine which read is correct as shown below in Table 3. Using the mate-pair information in read 4 allows us to determine that read 3 is correct and should be used for sequencing.

Table 3: Resolving Repeats

Sequence	ATCG**GATTACA**AAAGGG**GATTACA**GGGAAA
Read 1	TCG**GATT**
Read 2	**TTACA**GGG (incorrect)
Read 3	**TTACA**AAA (correct)
Read 4	CG-------AA (mate-pair)

Until recently, using mate-pair information for repeat resolution was rare because of the high false-positive rate. About 10% of mate-pairs are mistakenly assigned, that is, 10% of mate-pair information is actually unrelated [5]. The primary source of this error is due to lane tracking errors in the slab-gel sequencing machines. The material does not migrate in a straight line so the optical software misnumbers the 32 to 96 lanes, causing these errors [5]. Mate-pair information can be used to resolve repeats if the error rate is sufficiently low.

3.3 Mathematical Analysis

Intuitively, as the number of reads increases, so should the quality of the final assembled sequence. Before going into more detailed analysis, we must first define a set of terms shown in Table 4.

Table 4: Definition of Terms

Symbol	Description
G	Length of target sequence
$\overline{L}$	Average length of sequence read
R	Number of sequence reads in shotgun data set
N	$R\overline{L}$, total number of base pairs sequenced
$\overline{I}$	Average length of a clone inset
$\overline{c}$	N/G, average sequence coverage
$\overline{m}$	$R\overline{I}/2G$, average clone or map coverage

Source [5].

A typical BAC of length G = 100 kbp is sequenced R = 1500 times with a length $\overline{L}$ = 500 bp. In total this is $N = R\overline{L}$ = 750 kbp of raw data for an average coverage of $\overline{c} = N/G$ = 7.5-fold. In practice, we want to sequence to a known level of coverage so we sequence until we get $N = G\overline{c}$ base pairs of data. Assuming perfectly uniform random sampling, the following results follow:

1. The probability that a base is not sequenced is $e^{-\overline{c}}$
2. Average contig lengths of $\left(\overline{L}/\overline{c}\right)e^{\overline{c}}$
3. Gaps of average length $\overline{L}/\overline{c}$

Source [5].

This information can now guide in the selection of $\bar{I}$, the average size of the insert. For double-barrelled shotgun sequencing, it follows that $\overline{m} = \bar{c}\left(\bar{I}/2\bar{L}\right)$ is greater than $\bar{c}$, so there is a factor of $e^{\bar{c}}/e^{\overline{m}} = e^{-\bar{I}/2\bar{L}}$ fewer gaps in the source inserts than gaps in the assembly of the clone. For an insert length of 5 kbp, there is a factor of e^{-5} (or 148) fewer clone gaps than sequence gaps. From another point of view, scaffolds are 148 times larger than contigs. With 7.5-fold coverage using a 200 kbp source, it is expected that all contigs could be ordered with mate information [5].

3.4 Clone-by-Clone Approach

The clone-by-clone approach is a hierarchal two-tiered method of sequencing. First, the entire human DNA sequence is fractured into 50- to 300-kbp fragments and then inserted into BACs to create a library. The first step is to produce a low-resolution physical map from the inserts and create a minimal tiling set of inserts that span the entire genome. These inserts, or clones, are then shotgun sequenced to reveal the entire genome. The physical map is created by using fingerprint information about each BAC insert. The most common type of fingerprinting is the STS (Sequence Tagged Site) probe which is the presence or absence of a pair of 18-length substrings between 200 and 1,000 bases apart in the insert [5]. Figure 3 below illustrates the clone-by-clone procedure.

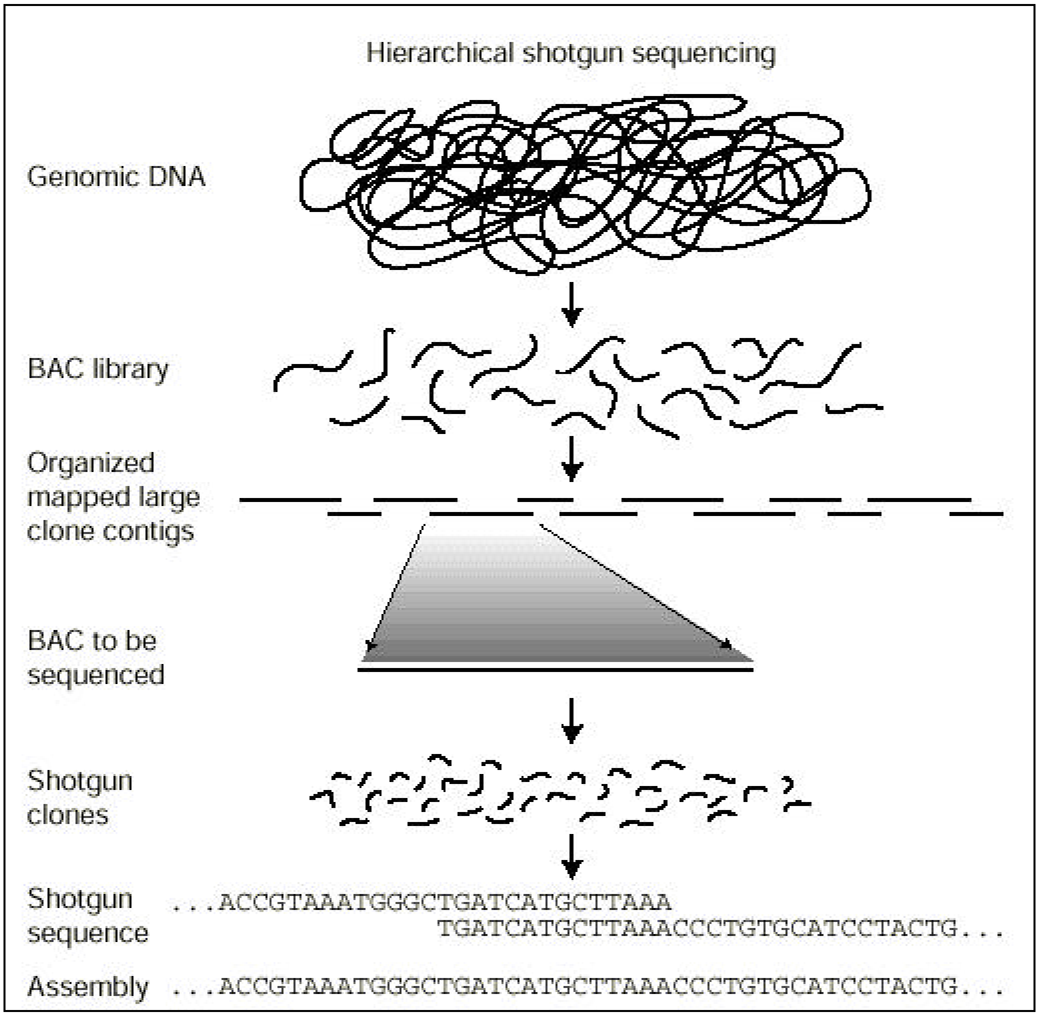
Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

Figure 3: Clone-by-Clone Sequencing

Source [7].

3.5 Whole-Genome Shotgun Assembly

In the early 1980s, typical source sequence sizes ranged from 5 to 10 kbp. By 1990, it was routine to shotgun sequences of about 40 kbp, comparable to the length of cosmid-sized clones. By 1995, the entire genome of the 1.8 Mbp bacteria *Haemophilus influenzae* was successfully shotgun sequenced [11]. It was then proposed that larger eukaryotic organisms could be sequenced with a whole-genome shotgun approach. By March 2000, the fruit fly *Drosophila melanogaster* had been successfully sequenced with this method. It was validated that large complex organism could be sequenced with the whole-genome approach. The main advantage of the whole-genome method is that it does not require a physical map to be built first. This step is both costly and time consuming.

4.0 The Human Genome Project

There has been much debate over which method of sequencing the Human Genome Project should employ. Simulations by proponents of the whole-genome shotgun approach suggested it was feasible and that it would be more efficient [12]. This stance was challenged by arguing that the likely risks outweighed the benefits and that the clone-by-clone approach should be used [13]. In the end, the HGP decided to use the clone-by-clone approach mainly because it was safer [7].

4.1 Selection of Clones

The clones were chosen from eight large-insert libraries containing BAC or P1-derived artificial chromosome (PAC). Partial digestion of genomic DNA with restriction enzymes was used to create the libraries. In total, the library represents approximately 65-fold coverage. It must be noted that libraries based on other vectors, such as cosmids, were also used earlier in the project. This may potentially introduce clone bias [7].

In the large-scale sequencing phase, a genome-wide physical map of overlapping clones was first constructed by systematic analysis of BAC clones for 20-fold coverage. DNA from each BAC was fingerprinted with a restriction enzyme. The fingerprint pattern was then positioned with STS markers from existing genetic maps. This allowed for BACs to be easily retrieved for later analysis. Where possible, clones were selected to form a minimum tiling set. However, since construction of the physical map was concurrent with sequencing, it was not possible to select such a set. The clones used for the HGP are therefore not a minimally overlapping set, but was justified because it allowed sequencing to start earlier. Furthermore, it allowed many SNPs to be discovered [7].

4.2 Sequencing and Assembly

Each of the twenty centres involved with the project had different sequencing equipment and standards. Average length of insert size varied, as well as the use of double-barrelled or single-ended sequencing. They also differed in the fluorescent labels and the degree to which dye-primers or dye-terminators were used. Both slab gel- and capillary-based sequencers were used. However, the resulting data could still be compared directly because the raw sequences were all processed with the Phred, Phrap, and Consed software packages [7].

The Phred (Phil's read editor) program assesses the quality of the fluorescent signals. Using Fourier methods, it determines the likelihood that a given base has been correct identified by the sequencer. The score is logarithmic, so a quality of 15 indicates a 1-in-$10^{15/10}$ chance that the base is incorrectly assigned [4]. A distribution scores for the draft sequence is shown below in Table 5. Phrap (phragment assembly program, or Phil's revised assembly program) uses these scores to assemble the shotgun data. The assembled data is visually output to Consed, a consensus visualizing and editing program. A human user can examine the assembly created by Phrap and correct any mistakes and identify possible gaps to fill [4]. The contigs generated by Phrap were then assembled into scaffolds with GigAssembler using mRNA, mate-pairs, and other information [7].

Table 5: Distribution of Phred Scores

Phred Score	Percentage of bases in the draft genome sequence
0 – 9	0.6
10 – 19	1.3
20 – 29	2.2
30 –39	4.8
40 – 49	8.1
50 – 59	8.7
60 – 69	9.0
70 – 79	12.1
80 – 89	17.3
>90	35.9

Source [7].

As of 7 October 2000, the draft contained 1,246 fingerprint clone contigs. A total of 4.26 Gbp had been sequenced from 29,298 overlapping BACs. This results in 23 Gbp of raw shotgun data for 7.5-fold coverage. However, since some clones have not been 'finished,' the overall draft genome has an average of 4.5-fold coverage [7].

The quality of the draft produced by the HGP was measured against a statistic called the 'N50 length', defined as the largest length *L* such that 50% of all nucleotides are contained in contigs of at least *L* [7]. The N50 length for intial sequence contigs is 21.7 kbp, 82 kbp for sequence contigs, 274 kbp for a sequence-contig-scaffold, 826 kbp for a sequence-cloned contig, and 8.4 Mbp for a fingerprint clone contig [7]. An illustration of these different contigs and scaffolds is shown below in Figure 4.

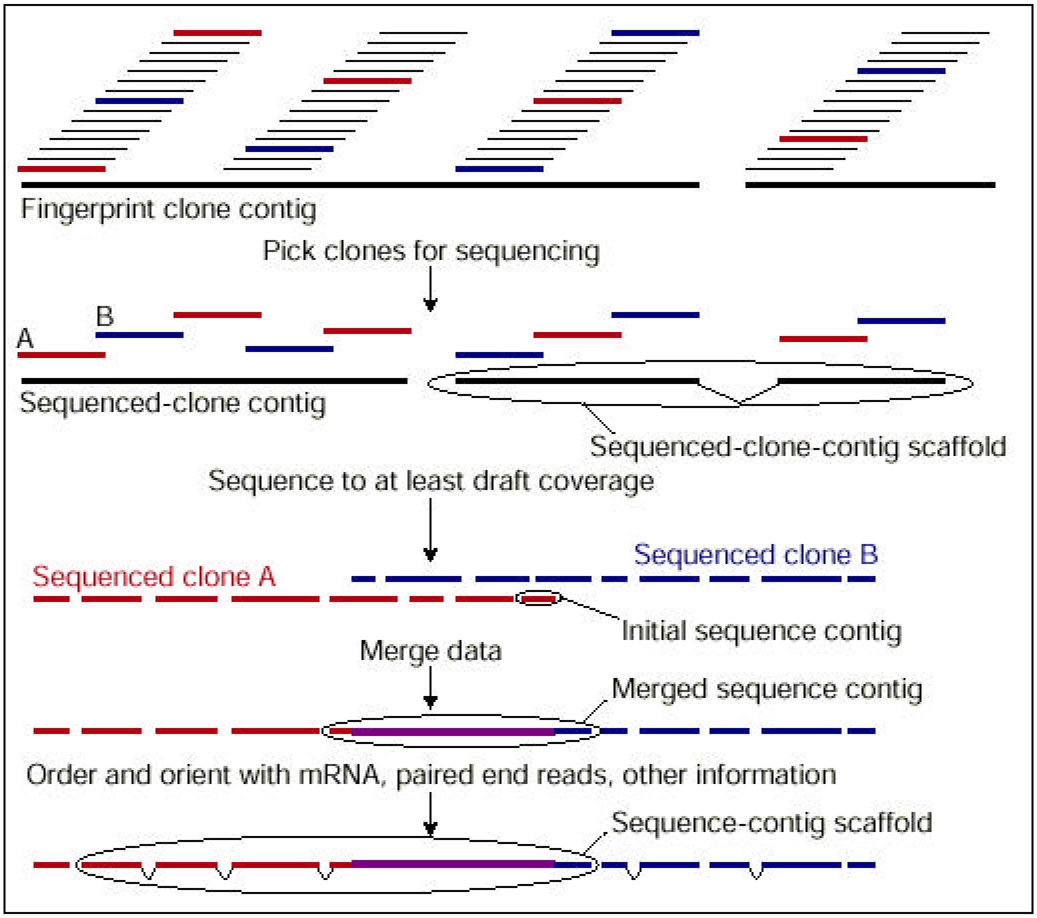

Figure 4: Illustration of Contigs used for HGP

Source [7].

5.0 Celera Genomics

In contrast to the HGP which used the conservative clone-by-clone approach, Celera employed a largely untested sequencing method. With the success of the *Drosophila* genome, Celera was ready to tackle the larger, more repeat-intensive human genome with the whole-genome assembly method. This approach to generate a draft human genome is discussed below.

5.1 Selection of Clones

The whole-genome shotgun strategy revolves around high-quality libraries consisting of varying insert sizes with mate-pairing information. The libraries must have an equal representation across the genome, a small number of clones without inserts, and no contamination from mitochondrial or *E. coli* DNA. Libraries consisted of inserts of three sizes: 2 kbp, 10 kbp, and 50 kbp [6].

5.2 Sequencing

The process for DNA sequencing at Celera was modular and automated. The four modules at the sequencing facility were: (i) library transformation, plating, and colony picking; (ii) DNA template preparation; (iii) dideoxy sequencing reaction set-up and purification; and (iv) sequence determination with the ABI PRISM 3700 [6]. It is important to recall that the ABI 3700 is a capillary-based sequencer capable of creating mate-pair information with high accuracy. This mate-pair information can then be reliably used to resolve repeats and assemble contigs and scaffolds

After quality and vector trimming, the average sequence length was 543 bp, and the sequencing accuracy was exponentially distributed with a mean of 99.5% and with less than 1 in 100 reads being less than 98% accurate [14]. Each sequence was then screened against vector alone, *E. coli* genomic DNA and human mitochondrial DNA. A total of 713 reads match *E. coli* genomic DNA, and 2114 matched mitochondrial DNA [6]. These samples were discarded and not used for assembly. Other sequences not used for assembly were reads from highly repetitive regions, data from other organisms introduced through various routes as found in many genome projects, and data of poor quality or untrimmed vector [6].

Quality control was maintained because all sequencing was performed in a single facility. The successful assembly of the *Drosophila* genome confirmed the validity of sequence data and quality control standards [14].

5.3 Assembly

Celera used two independent data sets for their assemblies. The first was a random shotgun data set of 27.27 million reads with average length of 543 bp produced by Celera. Combining the 2 kbp, 10 kbp, and 50 kbp libraries and mate-pair information, this sequence gave 5.1-fold coverage of the genome, and clone coverage of 3.42-fold, 16.40-fold, and 18.84-fold for the 2-, 10-, and 50-kbp libraries, respectively, for a total of 38.7-fold clone coverage. The second data set was derived from the HGP, and downloaded from GenBank on 1 September 2000 for a total of 4443.3 Mbp of sequence at various levels of completion [6].

To prepare for whole-genome assembly, the HGP data was first disassembled, or "shredded" into a synthetic shotgun data set of 550 bp reads that form a perfect 2-fold coverage of the bactigs. This resulted in 16.05 million "faux" reads for 2.96-fold genome coverage. The combined data set of 43.32 million reads (8-fold coverage) was then subject to the whole-genome assembly algorithm [6].

The whole-genome assembly (WGA) routines from the *Drosophila* project were extended for the 25-times larger human genome. The WGA assembler consists of a pipeline of five stages: Screener, Overlapper, Unitigger, Scaffolder, and Repeat Resolver, respectively. The Screener finds and marks all microsatellite repeats with less than a 6 bp element, and screens out all known interspersed elements, including Alu, LINE and ribosomal DNA. The marked regions get searched for overlaps, but screened regions do not. The Overlapper compares every read against every other read in search of complete end-to-end overlaps of at least 40 bp and with no more than 6% variation. Statistically, every overlap is a 1-in-10^{17} event, making it unlikely to be a coincidental event [6].

Overlaps may be incorrectly assigned due to large-scale repeat in the genome not screened for earlier in the sequence. This is known as a repeat-induced overlap. The Unitigger resolves repeat-induced overlaps. First, all assemblies of reads that appear to be uncontested with respect to all other reads are found. These subassemblies are known as unitigs (uniquely assembled contigs). Even if some of these subassemblies are indeed correct, some are actually collections of reads from several copies of a repetitive element that have been overcollapsed into a single subassembly. Fortunately, this is very easy to identify. The depth of coverage for the overcollapsed assemblies will be inconsistent with overall average coverage. A simple statistical discriminator was used to determine if the unitig was composed of unique DNA or of a repeat consisting of two or more copies. With the correct discriminator threshold, a subset of unitigs that are certain can be identified. Using a less stringent threshold, a subset of unitigs can be created that are almost certainly correct because they will consistently be assembled. The collection of these two sets is dubbed U-unitigs [6].

The result of the Unitigger was a set of correctly assembly contigs estimated to cover 73.6% of the human genome. Using mate-pair information, the Scaffolder ordered the contigs into scaffolds. Assuming mate-pairs are false less than 2% of the time, the information can link a given pair of U-unitigs with a certain orientation and distance, with a 1-in-10^{10} probability of being wrong. The U-unitigs can then be assembled with confidence using the 2- or 10-kbp mate pairs. These intermediate sizes scaffolds can then be recursively linked with 50 kbp mates and BAC end sequences. The scaffolds were typically of megabase size with gaps between between their contigs that generally correspond to repetitive elements and occasionally to small sequencing gaps. The resulting scaffolds reconstructed the majority of the unique sequence of a genome [6].

The next step was resolving repeats in the genome. This was done using a progressively more aggressive strategy. Using the “rocks” strategy, all unitigs with a good, but not definitive score was placed in a scaffold gap. This was only done on the condition that two or more mate pairs with one of their reads placed it unambiguously within the gap of the scaffold. The chance of an incorrect insertion is estimated to be less than 1-in-10^7 [6].

Using the “stones” approach, gaps are filled with mate pairs. A read may be placed in the gap because its mate pair is in the contig of the scaffold. In other words, a read may be inferred to be in the gap because of distance and location information in its mate pair. All such inferred mate pair information is collected and used to fill these gaps. External gap “walking” attempts to fill the remaining gaps. The gaps are filled with assembled BAC data that cover these gaps [6].

The final step in assembling the genome was to order and orient the scaffolds along the chromosomes. The scaffolds were aligned against two maps: a fingerprint map of BAC clones and GeneMap99, a high-density STS map [6]. The final assembly of scaffolds was produced by aligning them with both maps.

5.4 Computing Power Required

Assembly of the human genome is a very computationally intensive task. This section outlines the computing power required for selected processes, computing limitations, and modifications from the *Drosophila* algorithms.

A straightforward application of the *Drosophila* software would have required 600 GB of RAM, so the routines had to be modified for the assembly of the human genome. The Overlapper and Unitigger were made incremental so that the maximum instantaneous usage of memory was only 28 GB. The computing power required by the Overlapper to find all overlaps was roughly 10,000 CPU hours with a suite of four-processor Alpha SMPs with 4 GB of RAM. With 40 such machines operating in parallel, this process require 4 to 5 days. Sequence construction routines were run in parallel with the Overlapper [6].

Since the first three stages were now incremental, new data could be added at any time. The Scaffolder and Repeat Resolution could then be completed in 7 days with the new data. Assembly operations used 10 four-processor SMPs with 4 GB of memory per cluster (Compaq's ES40, Regatta) and a 16 processor NUMA machine with 64 GB of memory (Compaq's GS160, Wildfire). Assembly required approximately 20,000 CPU hours [6].

5.5 Quality of Assembled Data

When all data had been assembled, the scaffolds spanned 2.848 Gbp and contigs consisting of 2.586 Gbp of sequence data. More than 84% of the genome was covered by scaffolds greater than 100 kbp, averaging 91% sequence and 9% gaps for a total of 2.297 Gbp of sequence. In total, there were 93,857 gaps among these 1637 scaffolds. The average scaffold length was 1.5 Mbp, the average contig size was 24.06 kbp, and the average gap size was 2.43 kbp with exponential distribution. More than 50% of all gaps were less than 500 bp long, and more than 62% were less than 1 kbp, with no gaps greater than 100 kbp. Moreover, 65% of the sequence is in contigs greater than 30 kbp, more than 31% in contigs > 100 kbp, with the largest contig at 1.22 Mbp long [6].

Completeness is defined as “the percentage of eucrhomatic sequence represented in the assembly” [6]. However, since the entire euchromatin sequence has not been completed, the completeness of the sequence can only be estimated. Using a comparison with GeneMap99, it is estimated that the Celera assembled sequence contains 93.4% of the human genome and 5.5% in unassembled data for a total of 98.9% coverage [6].

Correctness is defined as “the structural and sequence accuracy of the assembly” [6]. Based on a statistical analysis of the quality readings of the underlying data, it is estimated that the Celera sequence is 99.96% correct. Another method using the clone coverage of 39-times estimates that at least 99% of the assembly is correct [6].

6.0 Comparison of Drafts

The two draft sequences produced by the HGP and Celera each have their strengths and weaknesses. The HGP draft has 0.65% unidentified bases while the Celera draft has 8.7%. After these unidentified bases have been removed from the sequence, the HGP sequence has 2.84 Gb of nucleotide sequence while the Celera sequence has 2.66 Gb [15]. Using 'N' to represent an unidentified base, the HGP sequence has 181,079 strings of 100 Ns, with strings up to 2,500 Ns. Meanwhile, the Celera sequence contains 21,684 strings of 50 Ns, but contains strings of up to 168,735 Ns. This is shown in Figure 5. However, since the annotation of the two drafts differ, this does not imply that gaps in the HGP draft are smaller than that of Celera's. As both these drafts move towards their completed form, it is expected that the differences between the two will diminish [15].

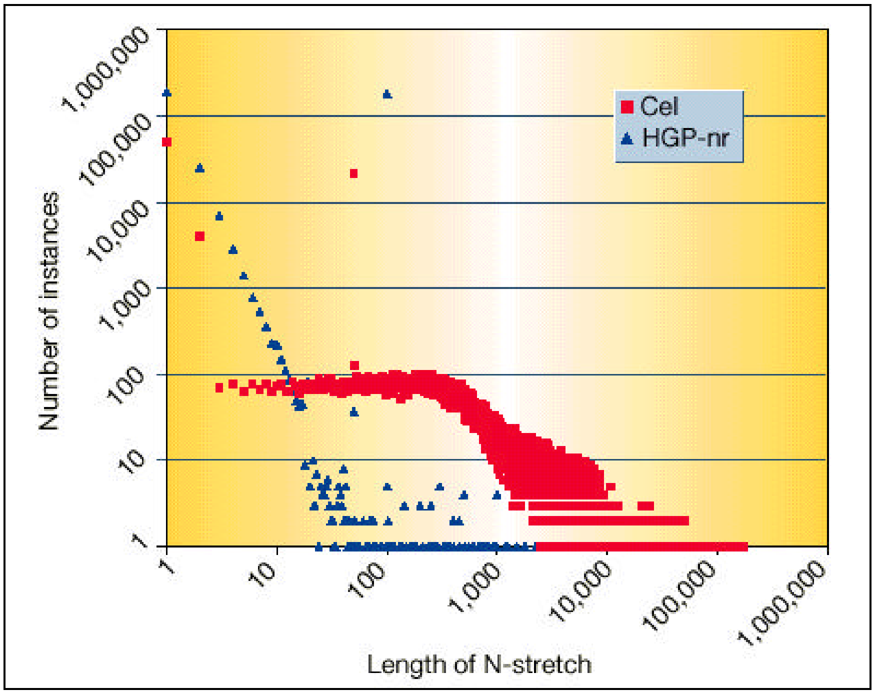

Figure 5: Comparison of HGP and Celera Data

Source [15].

7.0 Critical Analysis

The whole-genome shotgun assembly of the human genome was successful. It delivered high-quality reconstruction in unique regions of the genome and less in the repetitive regions [6]. Combined with the success of the *Drosophila* genome, there is not doubt about the utility and validity of the whole-genome approach.

The expensive cost and inefficiency of the clone-by-clone approach compared to whole-genome assembly makes it difficult to justify for subsequent large-scale projects. Still, other options such as BAC walking [5] and hybrid methods [6] may be explored for efficiency and cost. Celera will continue to sequence animals and plants at their facility using the whole-genome assembly approach.

Furthermore, sequencing rates appear to follow “Moore’s Law.” The amount of sequence reads doubles approximately every 18 months while the cost to do so is also cut in half [4]. With the conventional progression of Moore’s Law, computing time required for genome assembly will also decrease rapidly. This will allow future genomes to be sequenced even faster for even cheaper.

8.0 The Future

The draft sequence of human genome marks the beginning, not the end, of a revolution in biology. With the book of man in hand, researchers have a plethora of new challenges ahead. Three such important fields are in the analysis of single nucleotide polymorphisms (SNPs), comparative genomics, and proteomics.

As its name suggests, a SNP is a variation of a single nucleotide among individuals. These variations are important in determining differences between people. It may also cause certain individuals to be more susceptible to certain disease than others. So far, over 1.4 million SNPs have been identified [7]. Understanding of SNPs can lead to personalized medicine. Companies, such as Affymetrix, have developed “GeneChips” to analyze fragments of DNA for SNPs.

Since some biological tests cannot be ethically conducted on humans, an understanding of common laboratory animals is essential. By comparing genomes across species, researchers can gain better insight into the functions of specific genes. Gene manipulation can be performed on lab animals before being tested on humans.

Since the DNA ultimately codes proteins, the human genome can be used for proteomics. Proteomics involves analyzing the complex folding of the protein from the given genome sequence. The proteins function comes from this three-dimensional folding pattern. Understanding of protein behaviour can help towards the development of targeted drugs, radically reducing development time.

9.0 References

[1] James D. Watson and Francis Crick, "Molecular Structure of Nucleic Acids: A Structure of Deoxyribose Nucleic Acid," *Nature*, vol. 171, pp. 737 – 738, 1953.

[2] Robert H. Tamarin, Principles of Genetics (4th Ed.), Wm. C. Brown Publishers, USA, 1993.

[3] Leslie Roberts, *Science*, vol. 291, pp. 1182 – 1188, 2001.

[4] John Hodgson, "Gene sequencing's Industrial Revolution," *IEEE Spectrum*, November 2000, pp 36 – 42.

[5] Gene Myers, "Whole-Genome DNA Sequencing," *Computing in Science & Engineering*, May – June 1999, pp. 33 – 43.

[6] J. Craig Venter *et al*, "The Sequence of the Human Genome," *Science*, vol. 291, pp. 1304 – 1351, 2001

[7] International Human Genome Sequencing Consortium, "Initial Sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860 – 921, 2001.

[8] L. Rowen, B.F. Koop, and L. Hood, "The Complete 685-Kilobase DNA Sequence of the Human Beta T cell Receptor Locus," *Science*, vol. 272, pp. 1775 – 1762, 1996.

[9] G.I. Bell, "Roles of Repetitive Sequences," *Computers Chemistry*, vol. 16, pp. 135 – 143, 1994.

[10] F.J.M. Iris, "Optimized Methods for Large-Scale Sequencing in Alu-Rich Genomic Regions," *Automated DNA Sequencing and Analysis*, M.D. Adams, C. Fields, and J.C. Venter, eds., Academics Press, Long, pp. 199 – 210, 1994.

[11] J. Craig Venter *et al.*, "Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd," *Science*, vol. 269, pp. 496 – 512, 1995.

[12] James L. Weber and Eugene W. Myers, "Human Whole-Genome Shotgun Sequencing," *Genome Research*, **7**, pp. 401 – 409, 1997.

[13] Philip Green, "Against a Whole-Genome Shotgun," *Genome Research*, **7**, pp. 410 – 417, 1997.

[14] M.D. Adams et al, *Science* vol. **287**, pp. 2185, 2000.

[15] John Aach *et al*, "Computational comparison of two draft sequences of the human genome," *Nature*, vol. 409, pp. 856 – 859, 2001.