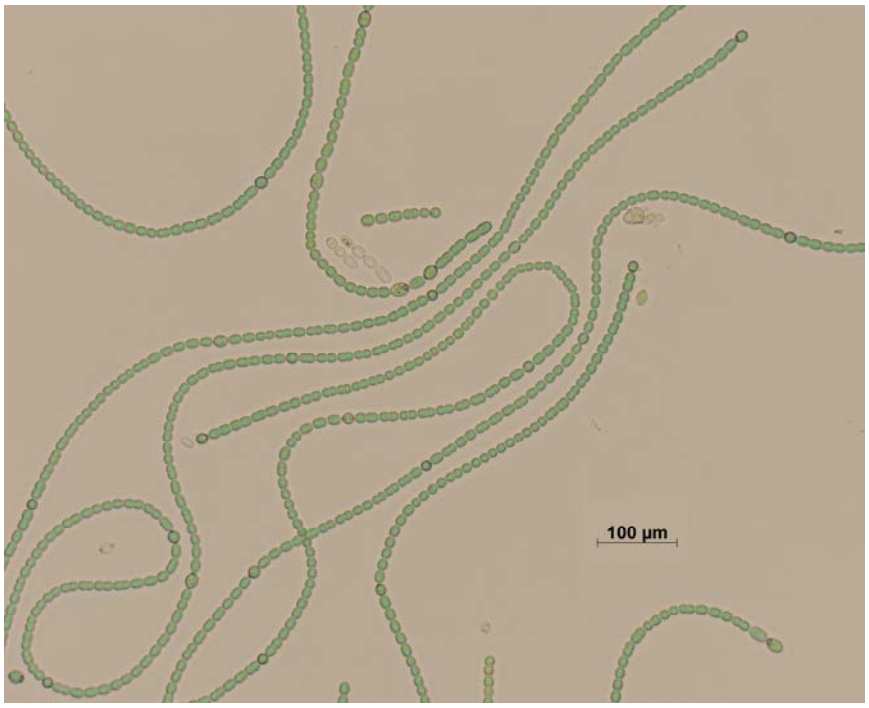
Whole genome assembly and gap closure of the toxic bloom-forming cyanobacterium *Anabaena* sp. strain 90



Hao Wang

Master Thesis

Department of Applied Chemistry and Microbiology

Division of Microbiology

University of Helsinki

2008

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Faculty of Agriculture and Forestry	Department of Applied Chemistry and Microbiology

Tekijä — Författare — Author

Hao Wang

Työn nimi — Arbetets titel — Title

Whole genome assembly and gap closure of the toxic bloom-forming cyanobacterium *Anabaena* sp. strain 90

Oppiaine —Läroämne — Subject

Biotechnology

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro Gradu	December 2008	80 + 16 appendix pages

Tiivistelmä — Referat — Abstract

Anabaena is a common member of the phytoplankton in lakes, reservoirs and ponds throughout the world. It is a filamentous, nitrogen-fixing cyanobacterial genus and is frequently present in the lakes of Finland. *Anabaena* sp. strain 90 was isolated from Lake Vesijärvi. It produces microcystins, anabaenopeptilides and anabaenopeptins. A whole genome shotgun sequencing project was undertaken to obtain the complete genome of this organism in order to better understand the physiology and environmental impact of toxic cyanobacteria. This work describes the genome assembly and finishing, the genome structure, and the results of intensive computational analysis of the *Anabaena* sp. strain 90 genome. Altogether 119,316 sequence reads were generated from 3 genomic libraries with 2, 6 and 40 kb inserts from high throughput Sanger sequencing. The software package `Phred/Phrap/Consed` was used for whole genome assembly and finishing. A combinatorial PCR method was used to establish relationships between remaining contigs after thorough scaffolding and gap-filling. The final assembly results show that there is a single 4.3 Mb circular chromosome and 4 circular plasmids with sizes of 820, 80, 56 and 20 kb respectively. Together, these 4 plasmids comprise nearly one-fifth of the total genome. Genomic variations in the form of 79 single nucleotide polymorphisms and 3 sequence indels were identified from the assembly results. Sequence analysis revealed that 7.5 percent of the *Anabaena* sp. strain 90 genome consists of repetitive DNA elements. The genome sequence of *Anabaena* sp. strain 90 provides a more solid basis for further studies of bioactive compound production, photosynthesis, nitrogen fixation and akinete formation in cyanobacteria.

Avainsanat — Nyckelord — Keywords

Cyanobacteria, Anabaena, genome, assembly, scaffold, gap closure, finishing, repeat

Säilytyspaikka — Förvaringsställe — Where deposited

Library of Microbiology Division

Muita tietoja — Övriga uppgifter — Further information

This Pro gradu work was supervised by Academy Professor Kaarina Sivonen and PhD David Fewer, and was funded by Academy of Finland (53305, 118637) and University of Helsinki.

Table of Contents

List of tables

Lists of figures

Glossary

Shotgun library: A collection of clones that over-sample the target genomic DNA and consists of inserts with similar size.

Reads: DNA strings that derived from large scale sequencing of shotgun library clone ends.

Pair-ends (or mated reads): two sequencing reads that derive from both ends of the same clone.

Contig: a stretch of DNA sequence produced by joining a collection of overlapping sequence reads.

Scaffold: a series of ordered non-overlapping contigs that are linked by pair-end information.

Genome coverage: a parameter for sequencing amount measurement, which can be calculated by dividing the total length of all sequenced reads by the genome size.

Sequencing gaps: non-sequenced regions between contigs belonging to the same scaffold, which had been cloned into shotgun libraries and could be filled by walking over the corresponding clones.

Physical gaps: DNA sequences that are not present in any clone of the libraries constructed for genome sequencing.

Copy number: the number of times of a particular sequence is present in the target DNA.

N50 Contig: the contig such that the contigs larger than it constitute 50% of the bases of the whole assembly, the bigger the N50 Contig size, the more complete the assembly.

Physical map: an information source which provides an ordered set of DNA fragments, there fragments are from restriction enzyme digestion of chromosomal DNA or large insert clones.

Abbreviations

BAC:	Bacterial Artificial Chromosome
BLAST:	Basic Local Alignment and Search Tool
CCD:	Charge-Coupled Device
DMSO:	Dimethyl sulfoxide
FISH:	Fluoresence In *Situ* Hybridization
HGP:	Human Genome Project
HS:	Hierarchical Shotgun
IPTG:	Isopropyl-β-D-1-thiogalactopyranoside
IS:	Insertion Sequences
LB:	Luria-Bertani
NCBI:	National Center for Biotechnology Information
NRPS:	Non-Ribosomal Peptide Synthetase
PFGE:	Pulsed Field Gel Electrophoresis
RFLP:	Restriction Fragment Length Polymorphism
VNTR:	Variable Number Tandem Repeat
WGA:	whole genome assembly
WGS:	whole genome shotgun
X-Gal :	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
Mb:	mega bases
Kb:	kilo bases

Abstract

Anabaena is a common member of the phytoplankton in lakes, reservoirs and ponds throughout the world. It is a filamentous, nitrogen-fixing cyanobacterial genus and is frequently present in the lakes of Finland. *Anabaena* sp. strain 90 was isolated from Lake Vesijärvi. It produces microcystins, anabaenopeptilides and anabaenopeptins. A whole genome shotgun sequencing project was undertaken to obtain the complete genome of this organism in order to better understand the physiology and environmental impact of toxic cyanobacteria. This work describes the genome assembly and finishing, the genome structure, and the results of intensive computational analysis of the *Anabaena* sp. strain 90 genome. Altogether 119,316 sequence reads were generated from 3 genomic libraries with 2, 6 and 40 kb inserts from high throughput Sanger sequencing. The software package `Phred/Phrap/Consed` was used for whole genome assembly and finishing. A combinatorial PCR method was used to establish relationships between remaining contigs after thorough scaffolding and gap-filling. The final assembly results show that there is a single 4.3 Mb circular chromosome and 4 circular plasmids with sizes of 820, 80, 56 and 20 kb respectively. Together, these 4 plasmids comprise nearly one-fifth of the total genome. Genomic variations in the form of 79 single nucleotide polymorphisms and 3 sequence indels were identified from the assembly results. Sequence analysis revealed that 7.5 percent of the *Anabaena* sp. strain 90 genome consists of repetitive DNA elements. The genome sequence of *Anabaena* sp. strain 90 provides a more solid basis for further studies of bioactive compound production, photosynthesis, nitrogen fixation and akinete formation in cyanobacteria.

1. Introduction

1.1 Cyanobacteria

The cyanobacteria (or blue-green algae) are important photosynthetic prokaryotes, and thought to be progenitors of chloroplasts (Douglas 1994, Giovannoni *et al.,* 1988). Evidence from the morphological fossil record also suggests that cyanobacteria were present 3.0-3.5 billion years ago (Ga) (Des Marais 2000). The photosynthetic activity of ancient cyanobacteria is thought to have resulted in increased oxygen levels in the biosphere (Schopf 2000), which affected nearly all life on earth. Cyanobacteria also play an important role in the Earth's nitrogen cycle, especially in the ocean (Howarth *et al.,* 1988). The important physiological properties of cyanobacteria have made them the subjects of research in diverse biological studies, and 32 genomes of bacteria belonging to this phylum have been completed (**Table 1**).

Cyanobacteria frequently form water blooms in worldwide aquatic environments, and these mass occurrences of cyanobacteria can be toxic (Sivonen and Jones 1999). The well-known cyanotoxins are the hepatotoxic peptides, which are comprised of microcystins and nodularins (Sivonen and Jones 1999). These peptides are toxic to eukaryotic cells after crossing the cell membrane because of their inhibition of phosphatases 1 and 2A (MacKintosh *et al.,* 1990, Yoshizawa *et al.,* 1990). They are produced from the non-ribosomal peptide synthetase (NRPS) gene clusters, which also produce a number of other bioactive secondary metabolites in cyanobacteria (Welker and Döhren 2006). Over 65 structural variants of microcystins have been reported (Sivonen and Jones 1999). Recent study indicates that the gene clusters encoding the enzyme complexes which direct the biosynthesis of microcystins and nodularin have an ancient origin (Rantala *et al.,* 2004). However, the biological function of these toxic and other non-ribosomal peptides is still unclear. To clarify the physiological mechanisms of toxicity and environmental impact of the toxic peptides, a genome sequencing project was initiated to fully characterize the genomic structure of *Anabaena* sp. strain 90, an organism for bloom-forming, filamentous, toxic cyanobacteria. This strain produces toxic

Table 1. Completed cyanobacterial genomes. (collected from NCBI at 3.6.2008)

Organism	Size (Mb)	GC (%)	# chromosome	# plasmid	Accession	Release date	Sequencing Center
Acaryochloris marina MBIC11017	8.36	47.0	1	9	NC_009925.1	16/10/07	Washington University (WashU)
Anabaena variabilis ATCC 29413	7.07	41.4	1	3	NC_007413.1	15/09/05	DOE Joint Genome Institute
Cyanothece sp. ATCC 51142	5.43	37.9	2	4	NC_010546.1	01/04/08	Washington University
Gloeobacter violaceus PCC 7421	4.66	62	1		NC_005125.1	19/11/03	Kazusa
Microcystis aeruginosa NIES-843	5.8	41.6	1		NC_010296.1	24/01/08	Kazusa
Nostoc sp. PCC 7120	7.21	41.3	1	6	NC_003272.1	05/01/02	Kazusa
Prochlorococcus marinus str. AS9601	1.7	31.3	1		NC_008816.1	19/01/07	J. Craig Venter Institute
Prochlorococcus marinus str. MIT 9211	1.7	39.7	1		NC_009976.1	13/11/07	J. Craig Venter Institute
Prochlorococcus marinus str. MIT 9215	1.7	30.8	1		NC_009840.1	20/09/07	DOE Joint Genome Institute
Prochlorococcus marinus str. MIT 9301	1.6	31.3	1		NC_009091.1	05/03/07	Gordon and Betty Moore Foundation Marine Microbiology Initiative
Prochlorococcus marinus str. MIT 9303	2.7	50	1		NC_008820.1	22/01/07	J. Craig Venter Institute
Prochlorococcus marinus str. MIT 9312	1.71	31.2	1		NC_007577.1	02/11/05	DOE Joint Genome Institute
Prochlorococcus marinus str. MIT 9313	2.41	50.7	1		NC_005071.1	15/08/03	DOE Joint Genome Institute
Prochlorococcus marinus str. MIT 9515	1.7	30.8	1		NC_008817.1	19/01/07	J. Craig Venter Institute
Prochlorococcus marinus str. NATL1A	1.9	35	1		NC_008819.1	22/01/07	J. Craig Venter Institute
Prochlorococcus marinus str. NATL2A	1.8	35.1	1		NC_007335.2	09/08/05	DOE Joint Genome Institute
Prochlorococcus marinus subsp. marinus str. CCMP1375	1.75	36.4	1		NC_005042.1	15/08/03	CNRS
Prochlorococcus marinus subsp. pastoris str. CCMP1986	1.7	30.8	1		NC_005072.1	15/08/03	DOE Joint Genome Institute
Synechococcus elongatus PCC 6301	2.7	55.5	1		NC_006576.1	21/12/04	Nagoya University, Japan
Synechococcus elongatus PCC 7942	2.75	55.4	1	1	NC_007604.1	10/03/04	DOE Joint Genome Institute
Synechococcus sp. CC9311	2.61	52.4	1		NC_008319.1	01/09/06	TIGR
Synechococcus sp. CC9605	2.51	59.2	1		NC_007516.1	27/10/05	DOE Joint Genome Institute
Synechococcus sp. CC9902	2.23	54.2	1		NC_007513.1	26/10/05	DOE Joint Genome Institute
Synechococcus sp. JA-2-3B'a(2-13)	3.05	58.5	1		NC_007776.1	06/02/06	TIGR
Synechococcus sp. JA-3-3Ab	2.93	60.2	1		NC_007775.1	06/02/06	TIGR
Synechococcus sp. PCC 7002	3.4	49.2	1	6	NC_010475.1	14/03/08	Penn State Univ.
Synechococcus sp. RCC307	2.2	60.8	1		NC_009482.1	19/05/07	Genoscope
Synechococcus sp. WH 7803	2.4	60.2	1		NC_009481.1	19/05/07	Genoscope
Synechococcus sp. WH 8102	2.43	59.4	1		NC_005070.1	15/08/03	DOE Joint Genome Institute
Synechocystis sp. PCC 6803	3.95	47.4	1	4	NC_000911.1	31/08/95	Kazusa
Thermosynechococcus elongatus BP-1	2.59	53.9	1		NC_004113.1	21/09/02	Kazusa
Trichodesmium erythraeum IMS101	7.8	34.1	1		NC_008312.1	06/07/06	DOE Joint Genome Institute

peptide microcystins (Fujii *et al*., 1996, Rouhiainen *et al.,* 2004), anabaenopeptilides (Fujii *et al*., 1996, Rouhiainen *et al.,* 2000) and anabaenopeptins (Fujii *et al*., 1996), the biosynthesis of these peptides has been characterized. It was expected that characterization of the genome context and concurrent gene sets would help to expand our understanding of this toxic cyanobacterial species.

1.2. Genome sequencing

Rapid progress has been made in all aspects of large scale sequencing, with the success of human genome projects both from the public (Lander *et al.,* 2001) and private (Venter *et al.,* 2001) sectors. This has led to the exponential growth of DNA and protein sequences in public databases. Two different strategies were used in the process of sequencing the human genome. The international Human Genome Project (HGP) used a hierarchical shotgun (HS) approach, whereas Celera Genomics adopted a whole-genome shotgun (WGS) approach (Waterston *et al.,* 2002). In the first method, a mapping step was implemented (McPherson *et al.,* 2001) to produce a tiling path of large insert clones along the chromosomes **(Figure 1)**. A genome sequence could be obtained by merging sequences of adjacent clones, which are individually subjected to shotgun sequencing. The WGS approach gets rid of the complicated mapping step, but results in a heavy burden on the assembly and finishing steps for the sequencing data generated at the whole genome level. Although there have been debates in the literature as to the effectiveness of the WGS approach in obtaining the human genome sequence (Green 2002, Myers *et al.,* 2002, Waterston *et al.,* 2003, Adams *et al.,* 2003), it is generally accepted that the former method can guarantee an accurately finished genome with low sequence coverage, but entails a time-consuming mapping step; while the latter approach could yield most of the genomic DNA in a short period, but with the risk of an unfinished genome.

The cost of genome sequencing has been continuously reduced with the implementation of technical innovations in the fields of DNA sequencing (Smith *et al.,* 1986) and

genomic library construction (Shizuya *et al.,* 1992). Frequently upgraded computer systems and bioinformatics tools for genome assembly and finishing have provided a platform for processing an ever increasing amount of sequencing data. These advances have further increased the efficiency of the WGS approach, and led to its wider usage in genome projects of microbial organisms due to their lower complexity and smaller size. So far, more than six hundred completed microbial genomes have been deposited in the National Center for Biotechnology Information (NCBI) database since 1995, when the first bacterial genome, that of *Haemophilus influenza*, was published (Fleischmann *et al.,* 1995). At present there are 32 complete cyanobacterial genomes that have been released (**Table 1**), nearly all of which were obtained by the WGS approach.

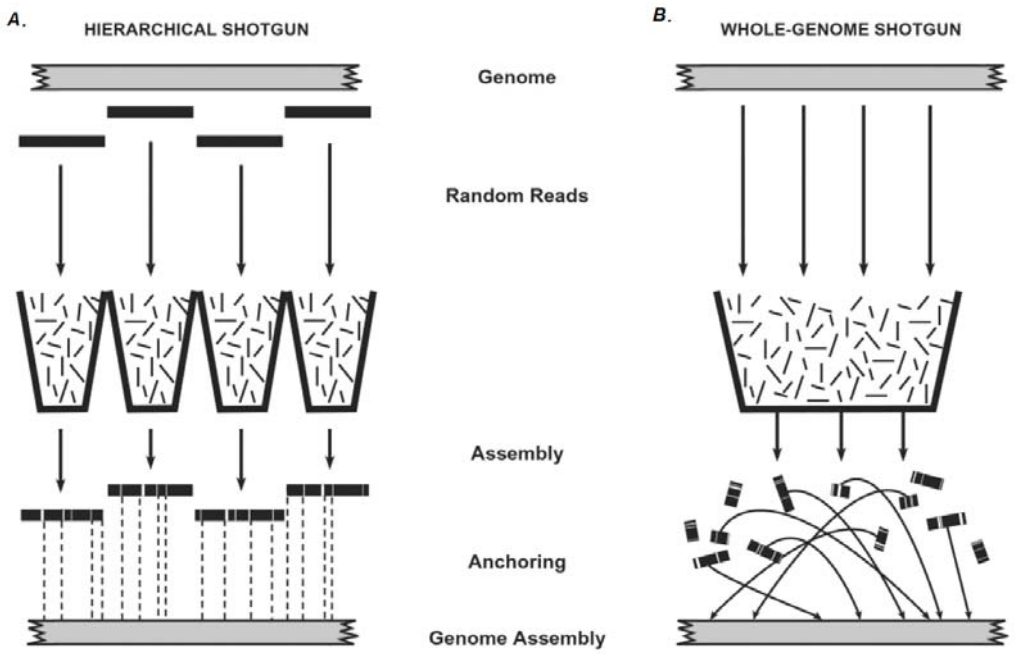


Figure 1. Comparison of current sequencing strategies. Adapted from Waterston *et al.,* (2002). (*A*) In the hierarchical shotgun (HS) strategy, large insert clones are ordered in an overlapping manner. Each mapped clone is shotgun sequenced and assembled individually. Then the complete genome sequence is obtained by merging the sequences of adjacent clones. This method has the advantage that assembly complexity is restricted to a smaller scale. (*B*) The whole genome shotgun (WGS) strategy attempts to assemble all shotgun sequencing data simultaneously. The complexity of this approach multiplies as genome size increases.

The WGS approach (**Figure 2**) normally begins with libraries of different insert sizes (Fleischmann *et al.,* 1995). Two libraries, one with small inserts and the other with large inserts, are shown in **Figure 2**, but even more libraries are recommended when working with larger and more complex genomes (Venter *et al.,* 1998). In each constructed library, the product of the average clone size and the number of clones used in sequencing should be above a certain multiple of the genome size (Lander and Waterman 1988), usually 15-20 times. This is to ensure that full coverage of the target genome can be achieved by end-sequencing (Venter *et al.,* 1998). In the high throughput sequencing phase, clones from all these libraries are sequenced from both ends with universal primers that are located at the edges of the vector sequences. The sequencing reads at this stage typically amount to 8X to 10X coverage according to the Lander-Waterman model (Lander and Waterman 1988). All reads generated from the large scale sequencing are then used to construct an initial whole genome assembly (WGA), in which each individual read would be aligned to others so that longer continuous sequence pieces (contigs) could be constructed through extending overlapped reads. Sequence reads obtained from small insert library clones always comprise a large portion of the genome sequencing, because it is easier to make libraries from short inserts than from large ones. Contigs are then further organized with respect to paired ends where each pair was derived from the same clone, and clustered to form a scaffold (or super-contig) **(Figure 2)**. The reads from large insert clones (cosmid, fosmid, or BAC) provide information which can be used to link and order contigs over a longer range. Although bacterial genomes are thought to be small and have a lower complexity, genome features like repetitive regions and unclonable sections, and a lack of physical data can result in an unfinished genome. The addition of the paired ends from large insert libraries can resolve these difficulties by anchoring separate contigs together and establishing a physical relationship (**Figure 2**). At the finishing stage, gaps within scaffolds are typically closed by primer walks over PCR products or clones that span the gap regions (International Human Genome Sequencing Consortium 2004). Each nucleotide base in a finished genome needs to be sequenced to a sufficiently high quality level and sequenced on both strands, in order that low quality sites and single stranded regions in the assembly will be properly modified

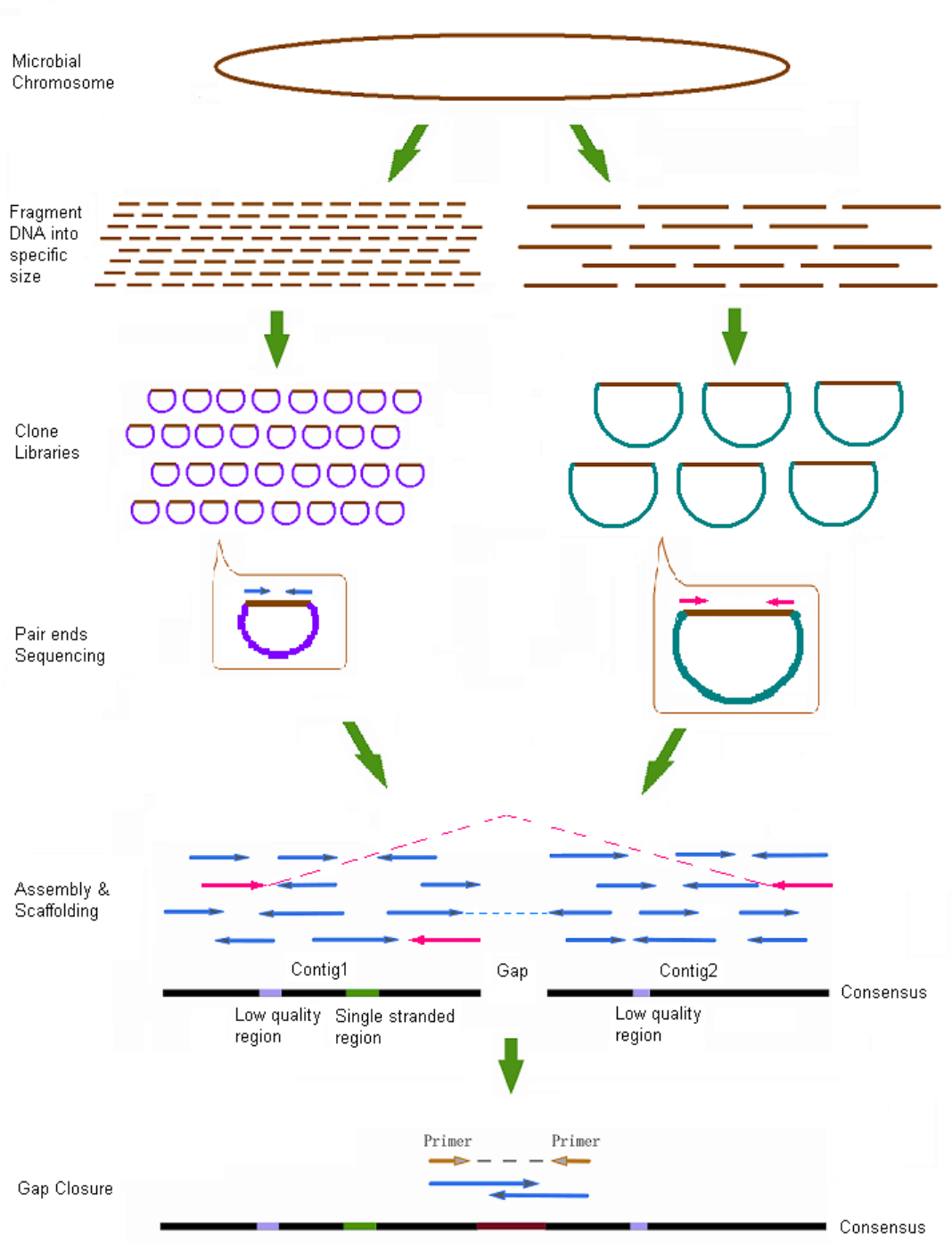


Figure 2. Basic strategy of the Whole Genome Shotgun (WGS) sequencing method.

(International Human Genome Sequencing Consortium 2004).

1.3 Genomic library construction

A typical WGS sequencing project may use a collection of libraries with both small inserts (2-10 kb) and large inserts (cosmid and fosmid: ~40 kb, BAC: 100-300 kb) to facilitate the scaffolding process during the finishing stage (Frangeul *et al.,* 1999). Inserts fragmented by sonication are preferred since their distribution is more random than those generated by site-specific nuclease cleavage (Fleischmann *et al.,* 1995). Homogeneity of insert sizes is another important requirement in building each genomic library, thus insert size selection is usually an indispensable step. This is to ensure that mated reads from the same clone are always apart from each other with certain distance (Frangeul *et al.,* 1999), as well as a few hundred bases coverage. For an unknown genome, one or more large insert libraries are vital, because they are the basis of the HS approach, and their pair-end sequences are very useful in finding the relationships and orders from the separated contigs in the WGS approach (Osoegawa *et al.,* 2001).

1.4 Physical map

A physical map normally provides an ordered set of DNA fragments from restriction enzyme digestions of genomic DNA or library clones (largely from BAC libraries). For complex eukaryotic genomes, they have been constructed through building clusters of BAC clones by intensive Restriction Fragment Length Polymorphism (RFLP) comparison, then mapping the physical locations of these clusters with Fluoresence *In Situ* Hybridization (FISH). Until now, enormous efforts have been made to obtain a few physical maps of mammalian genomes, such as those for human (McPherson *et al.,* 2001), mouse (Gregory *et al.,* 2002) and bovine (Snelling *et al.,* 2007). A physical map of a bacterial genome is usually generated by digestion of genomic DNA with rare-cutting restriction enzymes, followed with separation by Pulse Field Gel Eletrophoresis (PFGE). Comparison of RFLP results of overlapping PFGE fragments obtained with different

enzymes, and hybridization of genomic clones or known genes with bands in the PFGE gel, make it possible to order the PFGE segments and determine their locations along the bacterial chromosome, building up a physical map. In the hierarchical shotgun (HS) approach, shotgun-sequencing of a selected set of clones or DNA fragments would be sufficient to finish the target genome. Thus the amount of sequencing and the assembly complexity are minimized.

A number of physical maps have been constructed for the cyanobacteria *Nostoc* PCC 7120 (Bancroft *et al.,* 1989), *Synechococcus* sp. strain PCC 7002 (Chen and Widger 1993), *Synechocystis* sp. strain PCC 6803 (Churin *et al.,* 1995), and *Synechococcus* sp. strain PCC 6301 (Kaneko *et al.,* 1996a). These maps have assisted the sequencing of these cyanobacterial genomes using the HS approach (Kaneko *et al.,* 1996b), and have acted as supplementary scaffolding information for the WGS approach (Kaneko *et al.,* 2001, Nakamura *et al.,* 2002, Nakamura *et al.,* 2003).

1.5 High throughput sequencing methods

1.5.1 Sanger sequencing

Sanger sequencing (Sanger and Coulson 1975) became the dominant sequencing method soon after its development in the 1970s. The introduction of fluorescent tagged dideoxynucleotides (Smith *et al.,* 1986) and capillary electrophoresis separation (Karger *et al.,* 1990, Smith *et al.,* 1990, Dovichi *et al.,* 1990) promoted the automation and improved the accuracy of this method. Industrialized instruments later further multiplied the sequencing throughput by generating dozens or hundreds of sequences in parallel.

1.5.2 454 sequencing

The 454 sequencing method was developed from an existed pyrosequencing approach (Ronaghi *et al.,* 1998), which is based on the principle of "sequencing by synthesis". The

throughput volume in this method dramatically increases the number of sequencing reads in a single batch. A fibre-optic slide with highly condensed wells was invented to perform emulsion PCR on a very large scale (Margulies *et al.,* 2005). While sequencing, only one type of nucleotide is added at a time for the synthesis reactions in all wells, and the sequencing result of the single step is obtained by detecting synthesis products with a high resolution Charge-Coupled Device (CCD) camera. Since there are about 1.6 million wells on the slide, a huge volume of sequencing data can be produced in a short period even with a shorter sequence length (100-200 bp). The ability of this sequencing method to deal with novel and large genomes still needs to be validated because the data generated is too fragmented and lacks pair-end information. However, it should be noted that cloning biases might, to some degree, be circumvented by this library-free method.

1.6 The Lander-Waterman model

The assembly complexity is greatly simplified if there are some reference genomic sequences or marker genes, or physical map information available. However, for a novel organism, it is necessary to accumulate enough sequences to give sufficient genome coverage (>8X) before assembly (Frangeul *et al.,* 1999). This number is based on an important study undertaken by Lander and Waterman in the early days of the Human Genome Project (Lander and Waterman 1988). The formulas from their work were originally developed as a guide for mapping BAC clones over genomic DNA. But the conclusion of the Lander-Waterman model can be directly used for predicting sequencing completeness in WGS sequencing. Based on this model, the probability of any base that was sequenced **y** times could be deduced by the following Poisson formula:

$$P(\mathbf{y}) = (\lambda^{y} * e^{-\lambda}) / \mathbf{y}! \qquad \textbf{(1)}$$

In the above equation, λ is the fold of genome coverage. If we assign **y** as 0, this formula is then modified to calculate the probability of non-sequenced region existing in the genome when sequencing coverage is given as λ.

$$P(y=0) = (\lambda^0 * e^{-\lambda}) / 0! = (1 * e^{-\lambda}) / 1 = e^{-\lambda} \quad \textbf{(2)}$$

Base upon equation **(2)**, we can calculate the estimated assembly integrity under varying genome coverage **(Table 2)**.

Table 2. Estimated genome integrities under different genome coverages according to equation **(2)**.

Coverage(λ)	P(0)= $e^{-\lambda}$	% not sequenced	% sequenced 1-P(0)
1	0.37	37 %	63 %
2	0.135	13.50 %	87.50 %
3	0.05	5 %	95 %
4	0.018	1.80 %	98.20 %
5	0.0067	0.60 %	99.40 %
6	0.0025	0.25 %	99.75 %
7	0.0009	0.09 %	99.91 %
8	0.0003	0.03 %	99.97 %
9	0.0001	0.01 %	99.99 %
10	0.000045	0.005 %	99.995 %

Genome completeness through the shotgun approach is predicted to reach 99.9% when the coverage is 7X or above (**Table 2**). This is the theoretical basis of the WGS sequencing approach which normally use 8-10X as a minimum coverage for an unknown organism. Accordingly, awareness of genome size is important for planning when initializing a WGS genome project.

1.7 Base calling

Base calling is the step of extracting nucleotide bases from the original raw data (i.e. the chromatograph files generated by automatic sequencers) and converting them into

appropriate formats for assembly programs. There are two major publicly available programs: `phred` (Ewing and Green 1998, Ewing *et al.,* 1998) and the `pregap4` module in the `Staden` package (Staden *et al.,* 1998), which can be used to call bases and for quality assessment. They function in a similar style and are compatible with nearly all assembly programs. Of the two, `phred` has achieved a wider application in genome sequencing.

`Phred` can read trace data from SCF format, ABI model 3700 and 3730 DNA sequencer chromatograms, and MegaBACE ESD files, and automatically detect the file format whether the chromatogram files were compressed by gzip or compress (UNIX programs). `Phred` first calls the bases then writes the sequences to files in either PHD format (**Appendix 9.1**) or FASTA format (**Appendix 9.2**). Quality values for the bases are written to PHD files or FASTA format files, which can be used by assembly programs in order to increase the accuracy of the assembled sequences.

1.8 Genome assembly programs

A large number of 300-1000 bp sequencing reads are required to piece together a long chromosome. This step needs intensive computational resources to build up continuous genomic pieces (contigs) by assembly of overlapping reads. Assembly is normally performed by specific computer programs, such as `Phrap` (Green 1994), `CAP3` (Huang and Madan 1999), `Celera Assembler` (Myers *et al.,* 2000), and the newly developed eukaryotic genome assembly programs `ARACHNE` (Batzoglou *et al.,* 2002, Jaffe *et al.,* 2003) and `AMOScmp` (Pop *et al.,* 2004).

Algorithms for assembly programs function under similar principles because of the common aim for which they were designed. The `phrap` assembly algorithm involves two sequence comparison steps (Green 1994). In the initial comparison, each sequence is compared to every other in a pairwise fashion to identify overlaps. An LLR score is assigned to each pair of an alignment based on quantities of matching and mismatch

bases. By default, one matching residue receives a score of +1, one base mismatch gets a penalty of -2, a gap opening residue gets -4, and a gap extension residue gets -3. With these parameters, the highest score (i.e. LLR score) would be determined when scoring a local nucleotide alignment between a pair of reads. Normally, two sequences that are about 70% or more identical will tend to have a positive alignment LLR score; otherwise there tends to be a negative LLR score. Phrap then sorts all matches in order of decreasing LLR score to build layouts of contigs, and consensus sequences are constructed as a mosaic of the highest quality parts of overlapping reads by a second round comparison. The quality values for the resulting sequence are mostly inherited from the read segments of which it is comprised, and are combined with the information on strand coverage and chemistry types.

Phrap (Green 1994) is the most widely used assembly program, and was developed by Phil Green and his coworker Ewing Brent. It was initially designed for shotgun assembly of mapped BAC clones in the Human Genome Project (HGP), and later improved for whole genome assembly. Phrap allows the use of the entire read (not just a trimmed high-quality part) and computes data quality information (generated by phred) to improve the accuracy of the assembly. It outputs extensive information about assembly (including quality values for contig sequences) to assist trouble-shooting. Note that phrap does not provide editing or viewing capabilities; these functions are available with consed (Gordon *et al.,* 1998). Due to its popularity, phrap had been incorporated, as a standalone assembly engine, into a number of genome assembly systems, such as the Staden package (Staden *et al.,* 1998), RePS (Wang *et al.,* 2002), Phusion (Mullikin and Ning 2003) and Atlas (Havlak *et al.,* 2004).

The CAP3 (Huang and Madan 1999) assembly program had similar attributes to phrap. However the CAP3 assembler takes into account forward-reverse constraints. Comparison of these two programs showed that CAP3 often generates a more accurate but fragmented result than phrap (Huang and Madan 1999). PCAP (Huang *et al.,* 2003) is the upgraded version of CAP3. Its capacity was extended to the whole genome level,

and the efficiency of this new program has been tested with the mouse and the human genome data (Huang *et al.,* 2003).

The `Celera Assembler` (Myers *et al.*, 2000) is the classic assembly tool developed by The Institute for Genomic Research (TIGR) and the Celera Genomics between 1998 and 2002. Its performance is comparable to that of `phrap` and other greedy algorithm based assemblers. It is the first assembler that was modified to perform whole genome level assembly of eukaryotic organisms (Myers *et al.,* 2000).

The `ARACHNE` (Batzoglou *et al.,* 2002, Jaffe *et al.,* 2003) assembler was developed in the year 2000 at the Whitehead institute, MIT. It was designed for whole genome assembly of complex mammalian genomes (a few hundred Mb or more). Its effective performance has been successfully verified by WGS assembly of the mouse genome (Mouse Genome Sequencing Consortium 2002).

`AMOScmp` (Pop *et al.,* 2004) is an assembly program developed as part of the open-source `AMOS` project (see chapter **1.12**). The novel character of this program is the incorporation of genome sequences of other closely related organisms when assembling a new genome.

1.9 Causes of gaps

Theoretically, a genome should be finished with a 10X coverage assembly (see chapter **1.6**). However, empirical data shows that this is often not the case (Fleischmann *et al.,* 1995). The Lander-Waterman model should be used only as a guide because of the two prerequisites for this model: 1) clones (or reads) will be randomly distributed along the genome; and 2) the genome is made up of unique sequences and contains no repeats. These predefined assumptions of the Lander-Waterman model deviate from natural conditions. Normally, primary assembly of one genome project with 8-10X sequencing coverage produces hundreds of contigs, even for a small bacterial genome (Fleischmann

et al., 1995). This indicates that the genomic sequence is not arranged as a linear string with four randomly distributed letters (A, G, C and T). A gapped assembly is the typical result of genomic complexities: i) repetitive sequences (transposon, gene duplication, ribosomal RNA operons); ii) library bias resulting from an uneven distribution of clones; and iii) unclonable regions.

Repetitive sequences (or repeats) are a noteworthy characteristic of genomic DNA (Brown 2002). The size of repeats range from several bases to mega bases (Mb), and causes problems in sequencing and assembly. It has been shown that most gaps in the finishing of the human genome were attributed to repetitive DNA elements (Eichler *et al.,* 2004). A classic case of misassembly caused by repeats is a 'sequence jump' (**Figure 3**). At present this situation cannot be resolved by any assembly program (Salzberg and Yorke 2005). As the genome size increases, the size of the 'orphan' contigs (the yellow section in **Figure 3**) increases, whereas the possibility of recognizing such a situation decreases. Heterochromatic blocks are the extreme conditions of long repetitive sequences (in Mbs) that are frequently present in tandem around the telomeres and centromeres of large eukaryotic chromosomes. These regions cannot be resolved with current techniques since they cannot be cloned or sequenced with general protocols. Hence, repeat-rich eukaryotic genome projects only attempt to obtain the sequences of euchromatic regions (Adams *et al.,* 2000, Lander *et al.,* 2001, Venter *et al.,* 2001).

Unclonable regions might be lethal to bacterial host cells during the cloning process (Hayes 1995), and a biased library is a non-random representation of target DNA. Both of these would result in the assembly of a partial genome from genomic libraries. Because both HS and WGS are library-based sequencing approaches, these missing genomic regions in libraries would result in physical gaps in the assembly.

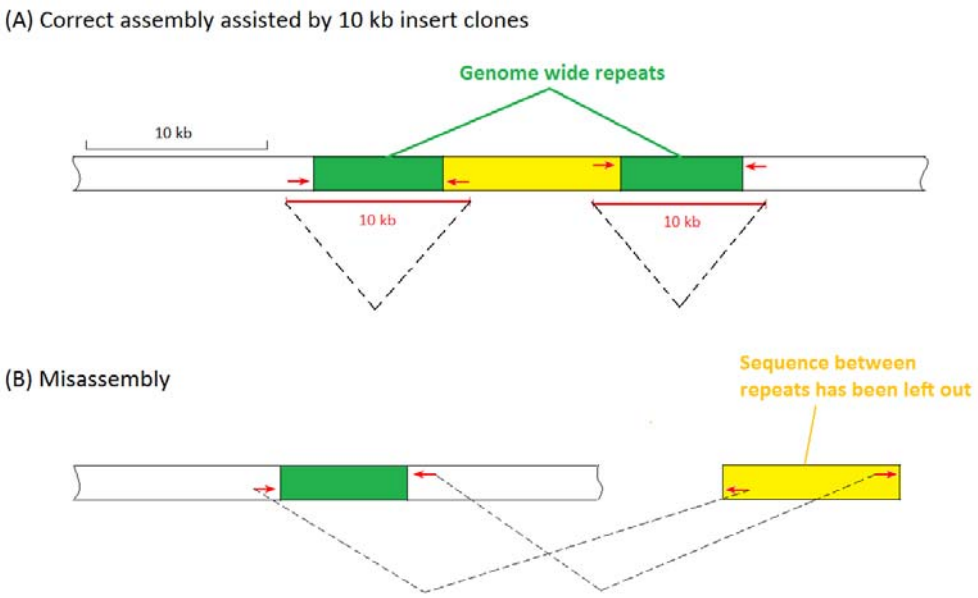


Figure 3. A classic case of 'sequence jump' misassembly. Modified from Salzberg and Yorke (2005). (A) Two genome-wide repeats (green sections) are distributed along the genomic DNA. Assembly of both end-sequences from 10 kb insert clones can anchor the consensus in the correct order. (B) A classic misassembly derived from the repeats. Two repeats had been joined together, and the middle part (yellow section) has been excluded. This could be identified by long range pair-ends spanning the entire repeat region.

1.10 Finishing and Scaffolding

The process of converting a gapped and unordered assembly into a complete genome is referred to as ‘finishing’, which mainly involves two associated steps: scaffolding and gap closure.

Scaffolding is the linking and ordering of contigs into clusters (scaffolds) with regard to pair-end reads located in different contigs (**Figure 2**). For the hierarchical shotgun (HS) approach, long range scaffolding is done in the initial mapping of large insert clones, and the assembly complexity has been restricted by the scale of the large insert clone size (40-300 kb). Thus the scaffolding process is greatly simplified. In contrast, scaffolding complexity in the WGS approach is immense, and the difficulty level multiplies as the genome size increases. The task of scaffolding a 5 Mb bacterial genome could be

challenging in some cases (Latreille *et al.,* 2007). Moreover, almost all current eukaryotic genomes sequenced by the WGS approach are still a collection of scaffolds with multi-megabase continuity, such as *Homo sapiens* (Levy *et al.,* 2007) and *Oriza sativa* (Yu *et al.,* 2005).

Misassembly checking is another part of the finishing process. In the WGS method, each read pair is always apart a particular distance from each other because the lengths of clone inserts are within a specific range (Venter *et al.,* 1998). Such positioning information could be applied to resolve the ambiguities caused by repeats (**Figure 3**).

Large insert libraries are essential in finishing. In WGS sequencing, long range pair-ends relationships could not only correlate individual contigs into an associated manner but also help to find misassemblies. In HS sequencing, mapping of large insert clones is the prerequisite for sequencing, and could also be viewed as a pre-finishing step. Meanwhile, the application of multiple cloning systems is also encouraged so as to reduce potential problems of library bias and unclonable regions (Venter *et al.,* 1998).

Nowadays, computer programs, like `consed` (Gordon *et al.,* 1998), and `hawkeye` (Schatz *et al.,* 2007), are widely used in scaffold construction and assembly verification, and they are largely operated in graphic interfaces. But these program-generated scaffolds and potential misassemblies need to be manually checked, because there is always a certain amount of false pairing information encountered in any shotgun sequencing project (Adams *et al.,* 2000, Venter *et al.,* 2001).

The closure process involves iterative cycles of computational analysis and laboratory work. Sequencing walking is a basic strategy for closing intra-scaffold gaps (closure of inter-scaffold gaps will be discussed in the next section). The walking templates normally are the clones that span the gap region and provide information for building scaffolds (**Figure 2**). PCR products amplified using primers designed from contig ends could also be used as templates if clones covering gaps are unavailable. Such sequencing walks

generally are employed in a bidirectional manner. Alternative methods for closing recalcitrant gaps have previously been introduced (International Human Genome Sequencing Consortium 2004).

Another step, which is normally carried out in parallel to gap closure, is resolving low-quality and single-stranded regions in assembly. An error event of less than 1 per 10,000 base pairs is a commonly accepted quality standard for finished sequences (International Human Genome Sequencing Consortium 2004). Additional sequences from resequencing existing clones and primer walks can be used for accomplishing this goal.

1.11 Closure of physical gaps

Inter-scaffold gaps, or physical gaps, are largely attributed to the cloning bias of large insert libraries both in WGS sequencing and HS projects (Lander *et al.,* 2001). To identify these gaps, a hybridizing method (Kirkness *et al.,* 1991) has been commonly used in which additional large insert clones are screened with probe sequences designed from scaffold ends. Shared positive clones from different scaffold ends are used as templates for the corresponding physical gaps. Two other methods were proposed for identifying physical gaps in bacterial genomes (Fleischmann *et al.,* 1995) which can be successful due to the lower complexities of bacterial genomes. The first method, peptide link, is based on BLASTX searching of protein databases. Whenever sequences of two scaffold ends are similar to the same protein, it might indicate a potential joint. The second method is based on PFGE results of the chromosome DNA. Probes designed from scaffold ends would be hybridized to the PFGE bands. If two probes hybridized with same band, their original scaffolds might be connected. Adjacent scaffolds identified with the above methods should be targeted for specific PCR (or long PCR) confirmation. Besides, it has been widely reported that PCR reactions with randomly paring primers from scaffold ends were utilized in the final finishing of many WGS projects. The multiplex PCR (combinatorial PCR) method (Sorokin *et al.,* 1996) has also been used to increase the efficiency of physical gap characterization. Recently, an optical mapping

method was introduced to finish bacterial genomes with highly repetitive sequences (Latreille *et al.,* 2007). Optical mapping is a system for creating whole-genome ordered restriction endonuclease maps from high molecular weight DNA (Reslewic *et al.,* 2005). Pattern comparison of optical maps and *in silico* restriction maps of assembled contigs turned out to be an effective way of guiding the finishing and validating misassemblies (Latreille *et al.,* 2007).

1.12 Software packages for genome projects

To cope with different aspects of a genome project, it is always necessary to develop diversified programs. These programs are normally organized as packages.

The package `Phred/Phrap/Consed` was developed and is curated by Phil Green's group at the University of Washington, and has been widely used for large sequencing projects at genome centers. This package is mainly comprised of `phred` (see chapter **1.7**), `phrap` (see chapter **1.8**) and `consed` (see chapter **1.10**). These programs work together to reach an optimal performance with large datasets from genome or EST sequencing.

The `Staden` Package (Staden *et al.,* 1998) is also a well-known computational system for genome assembly, which was developed at the MRC Laboratory of Molecular Biology, Cambridge, UK. It consists of a series of tools for DNA base calling and quality assessment (`pregap4`), assembly and editing (`GAP4`), and DNA/protein sequence analysis (`spin`).

`AMOS` (A Modular, Open-Source assembler) is a recently-initiated consortium which is committed to developing outstanding open-source assembly tools for next-generation sequencing projects through bringing together the efforts of leading genome assembly software developers. A list of an increasing number of modules is provided for different aspects of genome assembly. Currently available modules are: `Figaro` - statistical

vector trimmer; `minimus` - basic genome assembler for small datasets; `AMOScmp` - comparative assembler; `Bambus` (Pop *et al.,* 2004) - scaffolder; `amosvalidate` - assembly validation; `Hawkeye` - assembly viewer; and file conversion utilities for converting input and output files to and from different formats. The complete release of the entire `AMOS` package is anticipated.

The website references for all packages and programs mentioned in this thesis are collected in **Appendix 9.6**.

1.13 Cyanobacterial genomes

To date, 32 complete genomes of cyanobacteria have been released (**Table 1**). They were sequenced solely by the WGS approach with the exception of a few strains for which there was physical data (see chapter **1.4**). The taxonomic distribution of these organisms is biased in this phylum; three-quarters of these genomes are centralized in only two photosynthetic marine genera: *Prochlorococcus* (12) and *Synechococcus* (11). Genome studies demonstrated their compact gene sets (Dufresne *et al.,* 2003) and strong adaptation abilities (Palenik *et al.,* 2003, Palenik *et al.,* 2006). Additionally, a few more phylogenetically diverse cyanobacteria strains have been sequenced (Kaneko *et al.,* 2001, Kaneko *et al.,* 2007, Nakamura *et al.,* 2002, Nakamura *et al.,* 2003, Swingley *et al.,* 2008), mainly by the Kazusa DNA Research Institute, Japan.

2. Background and aims of this study

The general aim of this project was to sequence the complete genome of *Anabaena* sp. strain 90. This sequencing project was initiated by Academy Professor Kaarina Sivonen and was carried out as a scientific collaboration between the University of Helsinki and the Beijing Genomics Institute, Chinese Academy of Sciences. In this collaboration, strain culturing and DNA extraction were done in the Cyanobacteria group at the University of Helsinki. The high throughput sequencing was undertaken by the genome center of the Beijing Genomics Institute. Prior to starting this thesis in autumn 2006, the large scale sequencing phase was completed and preliminary assemblies were made.

The aim of this thesis was to finish the genome of *Anabaena* sp. strain 90. In order to obtain a gap-free genome, I started a new version of whole genome assembly at the University of Helsinki as illustrated in **Figure 4**, which included the cascading work of scaffolding, misassembly verification and primer design. For finishing, all PCRs and sequencing experiments were performed at the BGI Life Tech Co., Beijing, China.

This thesis is based on the bioinformatics studies conducted by Hao Wang at the University of Helsinki and the specific aims are:

- To set up a computational system and build up the whole genome assembly of *Anabaena* sp. strain 90 from the large amount of sequences generated by the WGS sequencing.

- To develop effective methods for scaffolding assembled contigs by using information derived from bioinformatics and experimental analysis.

- To close gaps and fix low quality regions in the assembly in order to obtain the genome with a sufficiently high quality and error-free sequences.

3. Materials and methods

3.1 Strain culture and DNA extraction of *Anabaena* sp. strain 90

The strain of *Anabaena* sp. strain 90 was isolated from the Vesijärvi lake, Finland, in 1986 (Sivonen *et al.,* 1992). This strain has been maintained in continuous liquid culture in the Cyanobacteria group at the University of Helsinki. *Anabaena* sp. strain 90 was grown in Z8 medium without nitrogen at 23-25℃ with continuous illumination of 20-25 μmol/m^2s. DNA extractions for genomic library constructions were made by Leo Rouhiainen in 1992, 2003 and 2004. The DNA was isolated according to the method of Golden *et al.,* (1988).

3.2 Construction of DNA libraries

Three sizes of genomic libraries were used for end sequencing. The large insert library was a cosmid library with an insert size of approximately 40 kb, and was constructed as previously described (Rouhiainen *et al.,* 2000). The two shotgun libraries were constructed from 2 kb and 6 kb inserts. They were made by shearing with sonication of the DNA, size selection of the fragments, ligation into pUC18 plasmid vector, and electroporation into the DH5 alpha competent *E. coli* cells (Invitrogen, Carlsbad, CA, USA).

3.3 Large scale sequencing

Transformants from all three libraries were grown on Luria-Bertani (LB) broth plates. Ampicilli / Isopropyl-β-D-1-thiogalactopyranoside (IPTG) / 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal) agar was used in selection of small insert clones, which were then picked and inoculated into 96-well plates. Each clone was assigned a unique name reflecting library type, cloning attempt, plate ID and location in the 96-well plate. DNA inserts within 96-well plates were stored at -20℃ after purification. Sequencing reactions were performed for each clone, using an ET MegaBACE Dye Terminator Kit (GE Healthcare UK Ltd, Chalfont St. Giles, UK), with appropriate vector-specific

universal primers for all libraries (Lorist6: T7 & Sp6; pUC18: M13F & M13R). Finally, the dideoxy Sanger method (Sanger and Coulson 1975), and the sequencers Megabase 1000 (GE Healthcare UK Ltd, Chalfont St. Giles, UK) and ABI 3730 (Applied Biosystems, Foster City, CA, USA) were used for high throughput end-sequencing. The criterion for qualified sequence reads had been set as 500 bases of Q-15 quality level or higher (Ewing and Green 1998). Since the genome size of *Anabaena* sp. strain 90 was unknown, a 10X shotgun sequencing coverage was calculated by assuming a genome size of approximately 6 Mb. The genome size estimation was based on the sequenced heterocyst forming cyanobacterium *Nostoc* sp. PCC 7120. This part of the work was undertaken by the sequencing center of Beijing Genomics Institute, Chinese Academy of Sciences in 2004.

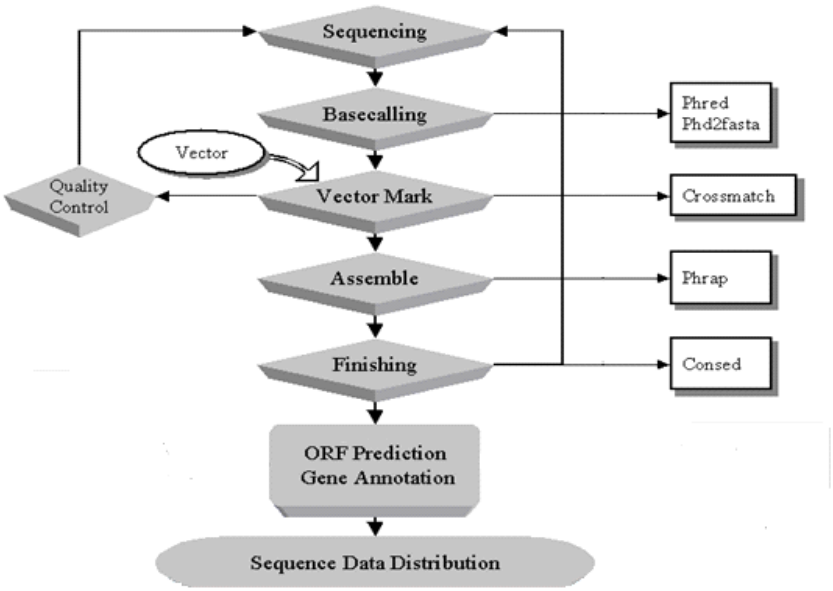


Figure 4. Bioinformatics procedures in genome sequencing. There are four steps in *in silico* analysis in this genome assembly: Base calling, Vector Marking, Assembly and Finishing. And the following programs were applied respectively: `Phred` & `Phd2fasta`, `Crossmatch`, `Phrap` and `Consed`.

3.4 Genome assembly

The *Anabaena* sp. strain 90 genome was assembled with the package `Phred/Phrap/Consed` (**Figure 4)**. The versions of the programs used in this study were 0.020425.c (**`Phred`**), 0.990622.f (**`phd2fasta`**), 0.990329 (**`cross_match`**), 0.990329 (**`Phrap`**), and 16.0 (**`Consed`**).

3.4.1 Base calling

Chromatogram trace files for the *Anabaena* sp. strain 90 genome were accumulated to give ~10X genome coverage for base calling, which was processed by `phred` (Ewing *et al.,* 1998). Nucleotide bases were extracted out with default settings (Ewing *et al.,* 1998). Base qualities, which reflect the log-odds score of each called base being correct (Ewing and Green 1998), were assigned to discern error probabilities. The PHD output format, which contains both bases and their quality values, was chosen for the assembly. The command line used in this study for running `phred` is presented below:

Command line for running `Phred`:

> phred –id chromat_dir –pd phd_dir

The command line given above forces `phred` to read the trace files from the "chromat_dir" folder and write the generated PHD files to the directory "phd_dir". In each PHD file, comments on the file conversion process were followed by the base values and their corresponding qualities (**Appendix 9.1**). The program `phd2fasta` was then used to create FASTA files from PHD files with the following command:

Command line for running `phd2fasta`:

> phd2fasta –id phd_dir –os An90.fasta –oq An90.fasta.qual

This operation generated two files. The file '***An90.fasta***' records the sequences of all the reads in FASTA format (**Appendix 9.2**) and the file '***An90.fasta.qual***' records the quality

values of each base in all the sequences (**Appendix 9.3**).

3.4.2 Vector screening

Following the creation of the two FASTA files with the raw read sequences and their quality scores, the next task was to screen out vector sequences before assembly. This is to avoid vector sequences causing reads to be identified as "chimeras" (recombinant inserts). The Smith-Waterman algorithm-based (Smith and Waterman 1981) alignment program, "`cross_match`" (Green 1994), was used for vector masking because it is more sensitive than BLAST (Altschul *et al.,* 1990). Before carrying out the screening, a FASTA-format file (vector.fasta) containing the sequences of the pUC18 and Lorist6 vectors was placed in specified location defined in `phredPhrap` script (see chapter **3.4.3**). In our case, the called reads file is "***An90.fasta***", and the `cross_match` running command is:

Command line for running `cross_match`:

> cross_match An90.fasta vector.fasta -minmatch 12 -minscore 20 -screen > screen.out

The '***-minmatch***' argument defines the minimum length of matching nucleotides in each comparison, and '***-minscore***' defines the minimum alignment score for a matching sequence pair. The '***-screen***' option creates a file named "***An90.fasta.screen***" containing vector-masked versions of the original sequences: i.e. any region that matches any part of a vector sequence is replaced by 'X's (**Appendix 9.4**).

3.4.3 Assembly

In this project, ancillary information on sequence orientation and mating was arranged into read names according to the CodonCode naming convention (Green 1994). A default directory structure of relevant files for assembly was constructed while using the `Phred/Phrap/Consed` package as recommended (Green 1994). In our case, a project

folder "***An90***" was created, and the sub-directory structure was deployed in following manner:

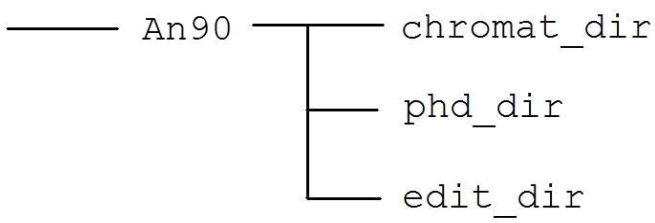


Figure 5. The directory structure of `Phred/Phrap/Consed` assembly. A folder named "*An90*" was created for genome assembly and three sub-folders were created for proper operation of `Phred/Phrap/Consed`. The sub-folder "chromat_dir" was used for storing all the chromatography files; "phd_dir" was the directory for `phred` output base calling files in PHD format; the sequence file and the quality file in FASTA format for `phrap` assembly and other cascading output files were located within the "edit_dir" folder.

To take full advantage of the potential of `phrap`, the file containing quality values generated by `phred` was included with the vector screened full-length sequences ("***An90.fasta.screen***") as input. Its name was manually modified from "***An90.fasa.qual***" to "***An90.fasa.screen.qual***" in the "edit_dir" folder. The format of this quality file was similar to the FASTA sequence file, both having identical header lines starting with a '>' character. In the quality file, this header line was followed by one or more lines giving the qualities of each base, which were separated by spaces and ranged from 0 to 99 (**Appendix 9.3**). The total number of quality values of each read matches the number of bases in the sequence file. Base quality information was used by `phrap` both in pairwise alignment and in determining the quality values of consensus contig sequences.

After the previous steps were set up properly, whole genome assembly (WGA) was implemented with the `phrap` program with the following command:

Command line for running `phrap`:

> phrap An90.fasta.screen -minmatch 14 –minscore 30 -new_ace > phrap.out

The quality file was not specified in the command line, but instead was arranged as above. Thresholds of '***-minmatch***' set the minimum matching length of nucleotides in each comparison, and '***-minscore***' is the minimum alignment LLR score. Both of them are basic parameters of `phrap` and `cross_match`. When the ***-old_ace*** or ***-new_ace*** option was specified, `phrap` would output the assembly result file in ACE format (a detailed description of the ACE file format can be found on the `consed` website: http://bozeman.mbt.washington.edu/consed/distributions/README.16.0.txt). Finally, the standard output was redirected to a file "***phrap.out***" for trouble shooting.

In addition to the standard output, `phrap` generated a series of other output files. These files were named by appending the input read file name with specific suffixes, for instance ".ace". The following is a list of the `phrap` output files used in this study:

(i) ***.contigs*** file. This is a FASTA file containing all contig sequences. Bases in this file are in upper case if and only if the quality is >= qual_show (default is 15). It also includes singleton contigs which consist of single read and match with some other contigs but which could not be merged consistently together.
(ii) ***.contigs.qual*** file. This has the phrap-generated qualities for the contig bases.
(iii) ***.singlets*** file. A FASTA file containing the singlet reads (i.e. the reads which have no match to any other read).
(iv) Standard error file (which is normally printed to the screen, but may be redirected to a file). This contains information indicating the point in the run that `phrap` has reached, summary results for some of the steps, and various warnings and error messages.
(v) ***.log*** file. It includes various diagnostic information and a summary of aspects of the assembly; it is useful in the trouble-shooting of assembly.
(vi) ***.ace*** file (generated when the option ***-new_ace*** or ***-old_ace*** is used). This is the most important output file of `phrap`; it can be used in visualizing assembly results by `consed` (see chapter **3.5.1**).

Within the Phred/Phrap/Consed package, there is a Perl script phredPhrap, which was applied to carry out the aforementioned data processing steps (i.e. base calling, vector screening, assembly, etc.) in batch mode. To guarantee smooth operation, several modifications were made manually beforehand: a) set up a proper directory structure (**Figure 5**) with all chromatogram files deposited into the "chromat_dir" folder; b) adhere to the proper read naming convention for reads, and arrange the information of template/chemistry/read and orientation in the right format; and c) ensure that corresponding clone vector sequences were included in a FASTA file and located in an appropriate location for screening with cross_match (see chapter **3.4.2**).

3.5 Finishing

3.5.1 Assembly viewing and editing

Consed (Gordon *et al.,* 1998) was used for viewing and editing the phrap assembly of the *Anabaena* sp. strain 90 genome. The parameter settings of consed used in this study (**Appendix 9.5**) were modified according to the feature of this genome project in advance.

Consed is graphical interface software (**Figure 6**). In practice, the Contig Window (**Figure 7**) and Assembly View (**Figure 8**) were used for scaffolding and misassembly checking. Firstly, correlated contig clusters (scaffolds) were viewed in Assembly View window, in which the pair-ends that define contig relations were labeled with blue or purple lines (**Figure 8**). These scaffolds defining reads were thoroughly confirmed by checking the alignment from individual Contig windows (**Figure 7**). When pair-ends were found in repeat regions, the corresponding scaffold structure was subjected to misassembly verification. In the case where only one end of a clone was a repetitive sequence, the placement of this end was decided by considering the physical location of the mated read. In the case of both ends of a clone being repeat sequences, they were discarded together. Insert sizes were also taken into consideration when building scaffolds. Inconsistent pair-ends, which were too far away or too close to each other and

were labeled with red lines, were subjected to intensive navigation in the Contig Window. In summary, the following criteria were used in scaffolding and assembly verification: i) any two contigs would be linked when there were at least two pair-ends; ii) single out-of-range pair-ends, which were illustrated as red lines in the Assembly View window, would be neglected when no more evidence was presented; iii) two or more out-of-range pair-ends present in the same region suggested potential misassemblies and were subjected to further checks. The ambiguities in which there was more than one pair of conflicting mated reads were resolved by PCR or long PCR amplifications on the genomic DNA. It was found that, in some cases, scaffold structures were not detected due to erratic patterns of trace files that were supposed to provide pair-end information for building scaffolds. Such problems were identified and corrected by checking the chromatography of the corresponding reads in the Trace Window (**Figure 9**), which was also used in the manual confirmation of single nucleotide polymorphism (SNP).

In finishing, the process of improving base quality was partially aided by the Autofinish module (Gordon *et al.,* 2001) of `consed` in candidate clone template selection and primer picking.

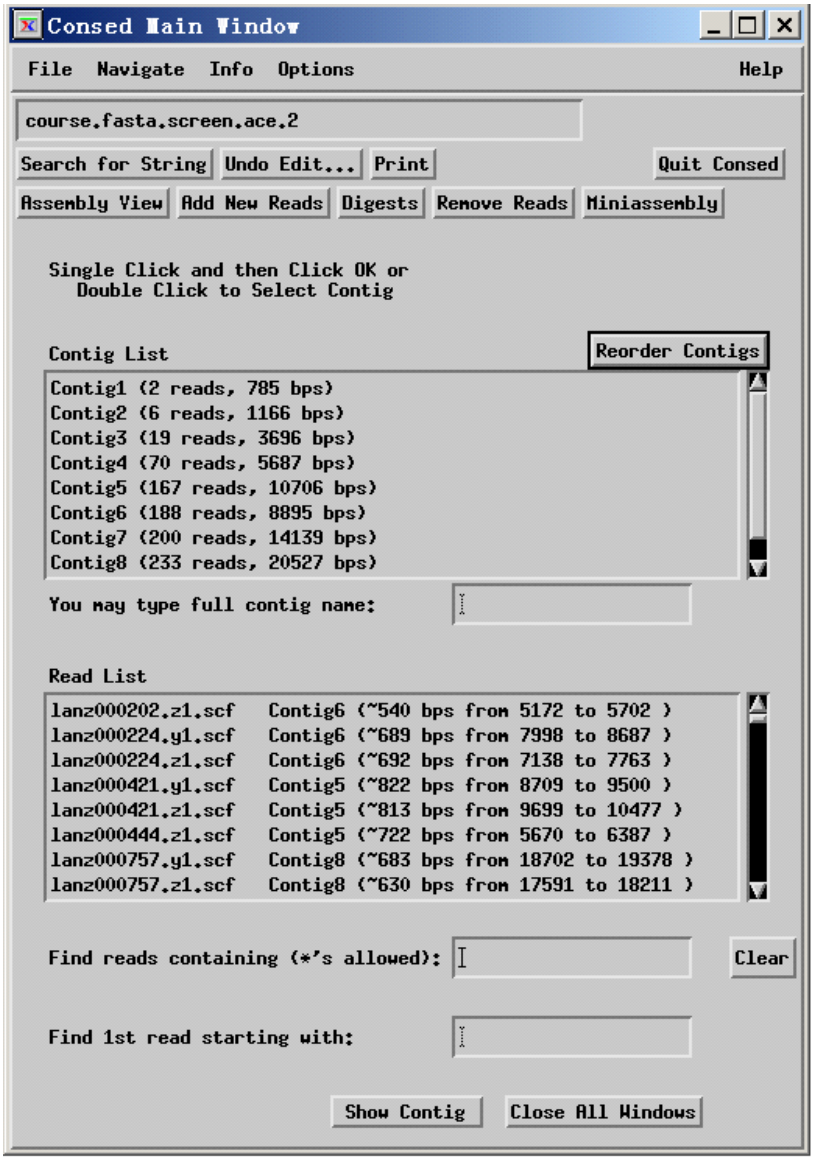


Figure 6. `Consed` main window. This is the central panel of `consed`. Lists of assembled contigs and reads are present in the upper and lower part of this window for browsing. Search boxes supporting wildcards are also provided for finding particular contigs or reads. The buttons provided here, "Add New Reads" and "Remove Reads", are used for adding and deleting reads from the assembly respectively; "Miniassembly" is for reassembling reads after pulling them out of their original contigs (this function is normally used in correcting misassembly); and "Assembly View" is for opening the Assembly View window. A detailed description of these buttons can be found in the manual (http://www.phrap.org/consed/consed.html).

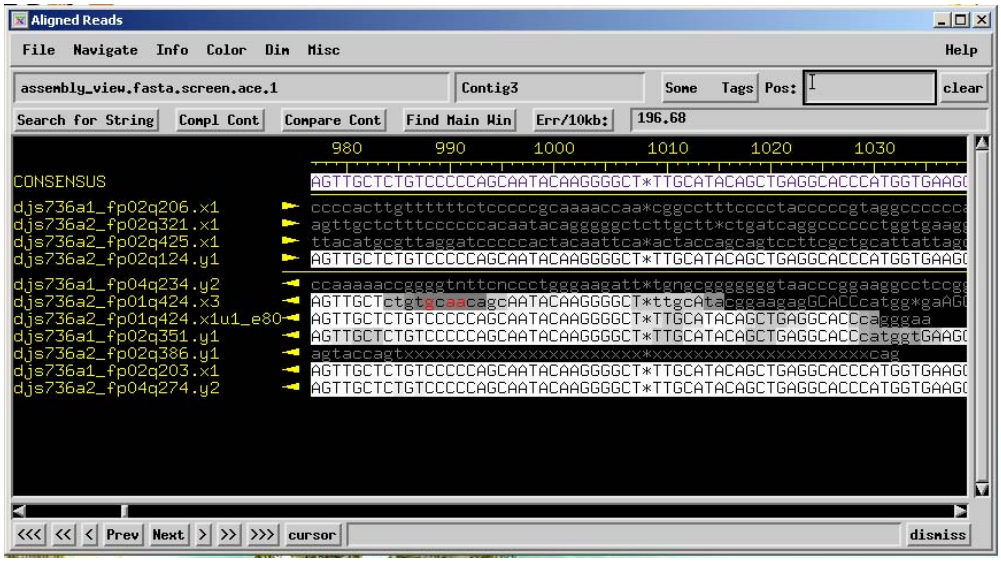


Figure 7. Contig Window of `Consed`. "Contig Window" will pop up when double-clicking any listed contig or read. Alignments of all overlapping sequences in a contig can be navigated from this window. The consensus sequence is on the top line with coordinates while assembled reads are aligned below according to their matching positions. The darker region of a read represents low-quality bases (Q value < 15); high quality bases are displayed in upper case with a white background color; discrepant bases are marked in red.

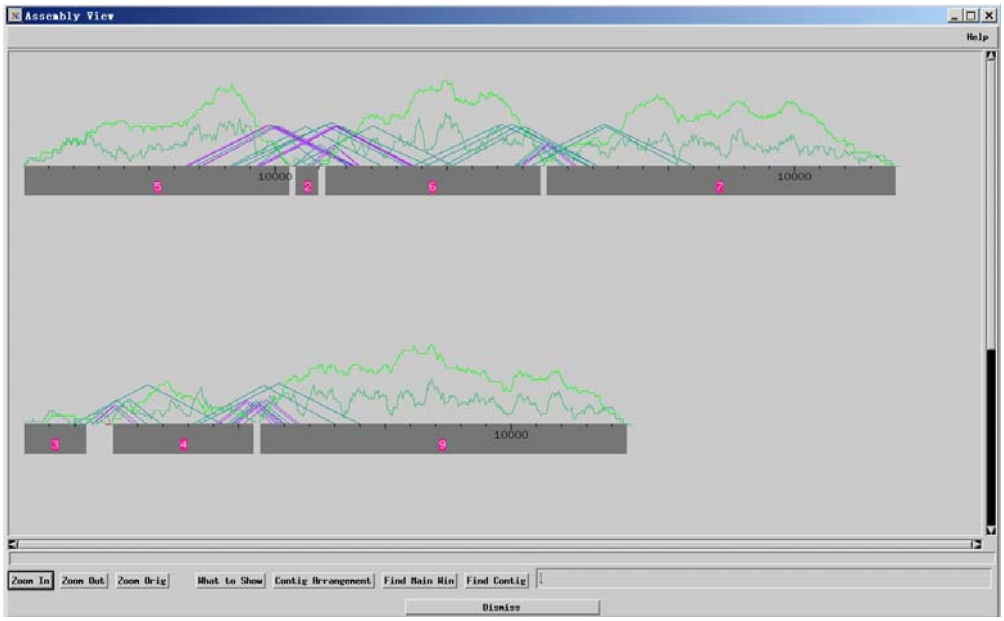


Figure 8. Assembly View window of a sample project. Two scaffolds (contig 5-2-6-7 and contig 3-4-9) are shown in this window. These contigs are clustered according to pair-end information from two different libraries (2kb and 6kb).

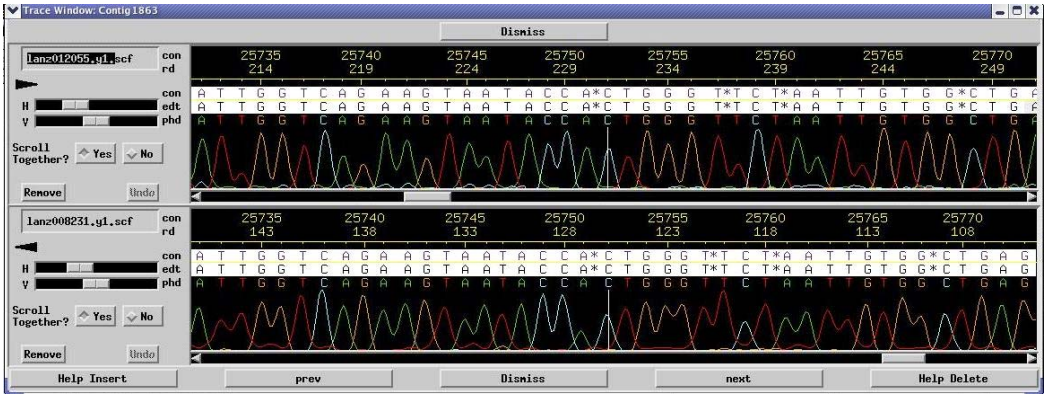


Figure 9. Trace window showing two aligned sequences. Chromatograph files can be zoomed into in both horizontal and vertical dimensions, and erroneous bases can be manually modified by clicking the middle button of the mouse over the bases and typing in the correct bases.

3.5.2 Supplementary scaffolding methods

In addition to evidence on mated reads, the peptide link method (Fleischmann *et al*., 1995) was extensively applied for inferring relationships between the remaining contigs. Both Blastn searching with completed genomes of closely related strains (*Nostoc* sp. PCC 7120: NC_003272.1 and *Anabaena variabilis* ATCC 29413: NC_007413.1) and Blastx searching against the Non-redundant (http://www.ncbi.nlm.nih.gov/blast/) protein database were utilized. The scaffolding search was further extended by screening for the ribosomal RNA operon and known cyanobacterial operons, which were collected from the Operon DataBase (ODB) (Okuda *et al*., 2006), to find links between remaining contigs. These operons were compared using the BLAST program against contig sequences in the assembly. The search results were manually screened for hits to any operons which were located at two different contig ends; this could be evidence of a potential link between the contigs. PCR reactions with primers designed from contig end sequences were then used to confirm this predicted relationship.

3.5.3 Combinatorial PCR

To further advance the finishing process, a method of combinatorial PCR was used to

join the remaining merged contigs through multiplex PCR experiments, in which a pool of primers (4-12) were added in each amplification and the genomic DNA was used as template. This method can be summarized in three steps: first, outward-directed primers are carefully designed from the ends of all remaining contigs; and then the primers were categorized into different groups; next, PCR experiments were performed in pairs of primer pools in a combinatorial manner. To illustrate the details of this strategy, the third round Combinatorial PCR test is presented below as an example. In this round, 39 primers were grouped into 7 pools, from L1 to L7 (**Table 3**).

In order to exhaustively obtain all possible PCR results from each primer pool, a combinatorial way for paring these primer pools was applied. PCRs were amplified in the following order: **L1- L2**, **L1 - L3**, **L1 - L4**, **L1 - L5**, **L1 - L6**, **L1 - L7**, **L2 - L3**, **L2 - L4**, **L2 - L5**, **L2 - L6** …… **L5 - L6**, **L5 - L7**, **L6 - L7**. So the total number (N) of reactions was:

$$\mathbf{N = {}_7C_2 = 7!/[2!(7-2)!] = 7\times6\div2 = 21}$$

Then a series of PCR results were visualized after electrophoresis separation (**Figure 10)**. Whenever there was a product, it was used as templates, and the corresponding two pools were used in priming subsequent sequencing over this product. If the sequencing was successful, reassembly with the newly generated reads was used to find out which two contigs were joined by the PCR products. For instance of the product L1—L5 (**Figure 10**), the L1 and L5 pools were used, rather than a single primer, in priming the sequencing reactions over this product. Reassembly with the two new sequences showed that the two reads were located at the left end of Contig166 and left end of Contig18, respectively, so the relationship between this pair of contig ends was validated. Following this strategy, eight pairs of contig ends were successfully correlated from this round of combinatorial tests, and they are highlighted with different colors in **Table 3**. It was necessary to walk many steps in both directions to fill in the gap when the product was very long, for instance the product L3—L5 (13.5 kb).

Table 3. Primer pools of the 3rd round of Combinatorial PCR tests. In this round, 39 primers were grouped into 7 pools for amplification. The product of each combinatorial PCR was used as a template for walk sequencing with the two corresponding primer pools. Re-assembly of the sequences generated provided the following scaffolding information, for instance, the band from **L1** and **L5** was generated with primers designed from Contig166's left end and Contig18's left end (labeled in the same green color), which showed that the two contigs were physically adjacent to each other.

L1	L2	L3	L4	L5	L6	L7
53LR	4LR	1RF	66LR	18LR	47RF	162RF
127RF	69RF	78LR	42LR	59LR	158RF	97RF
166LR	158LR	76RF	38RF	159LR	189RF	66RF
137LR	22LR	162LR	35RF	15RF	78RF	13RF
23LR	11LR	97LR	126RF	30RF	76LR	38LR
			6RF	101LR	114RF	35LR

It was possible that a PCR product could be generated within a primer pool (intra-pool product) simply because primers were grouped in a random manner. In this case, a band with the same size would appear in all lanes containing this pool, such as the **L2** group in **Figure 10**. To determine the origin of this intra-pool band, a method of sub-pooling PCRs was used in this study. PCR reactions were done with L2 sub-pools by removing one primer from the pool each time (**Figure 11)**. It clearly illustrated that this band was the product of primers 158LR and 11LR, because the band disappeared when either of them was missing. One further amplification between primers 158LR and 11LR confirmed this relationship.

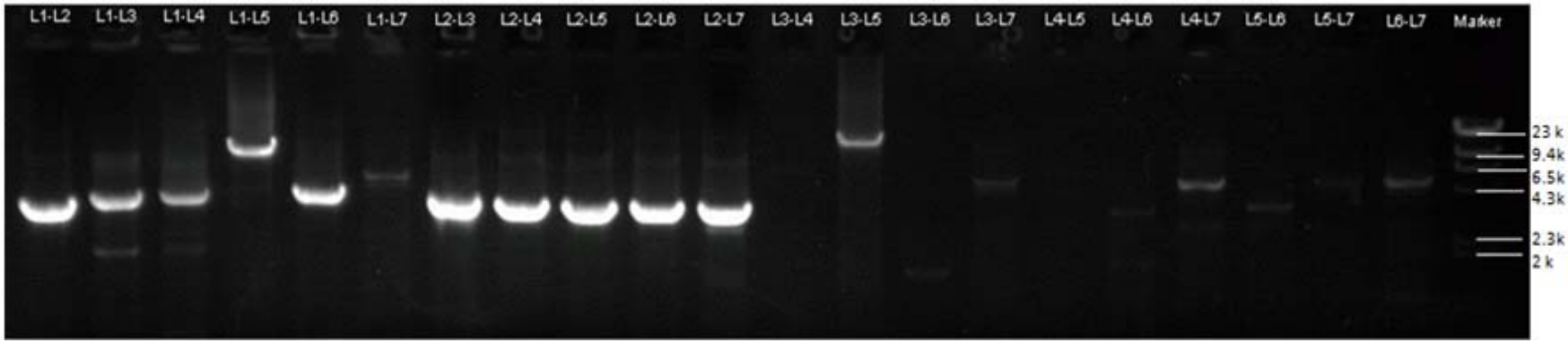


Figure 10. Gel image of the 3rd round Combinatorial long PCR results. In this round, 39 primers were grouped into 7 pools as **Table 3** shown. These 7 primer pools were paired in a combinatorial manner for long PCR amplifications. About 15 lanes showed clear bands from this gel image. The purified products were used as templates for walk sequencing with the corresponding two primer pools.

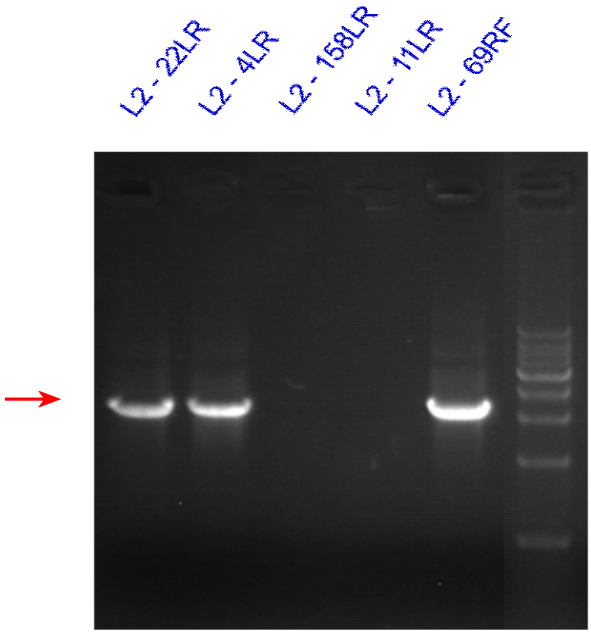


Figure 11. Gel image of five sub-pool multiplex PCR reactions in **L2** group, which has five primers originally. The product of the L2 intra-pool band (marked by red arrow) disappeared when primer 158LR or 11LR was absent.

In these combinatorial amplifications, PCR reaction conditions were 2 min of denaturation at 94°C, and 35 cycles of 20 s of melting at 94°C, 50 s of annealing at 55°C and 2 min of polymerization at 72°C, then 10 min of final extension at 72°C. The conditions for long PCR reactions were 2 min denaturation at 94°C, 35 cycles of 20 s of melting at 94°C, 50 s of annealing at 59°C, 10 min of polymerization at 68°C, and 10 min of final extension at 68°C. In each amplification, 0.2 μM of each primer and 200 μM dNTP were mixed with 1 unit of Takara LA Taq enzyme (Takara Bio Inc., Seta 3-4-1, Otsu, Shiga 520-2193, Japan) and 10X LA PCR buffer, 50 ng of genomic DNA were used as template, and RNase-free water was added to reach a final volume of 25 μl. All reactions for Combinatorial PCR tests were performed in an Eppendorf ™ Mastercycler gradient PCR machine (Eppendorf, Barkhausenweg, Hamburg, Germany). The PCR reactions were carried out at BGI Life Tech. Co., according to the experimental design of Hao Wang and were conducted at the University of Helsinki.

3.6 Primer selection

In the finishing phase, primer walks were commonly used for filling in gaps and fixing low quality regions. The built-in "Primer picking" function in consed was intensively utilized for the high throughput design of primers. It provided candidate primers in defined regions. Of the two types of primers, for the PCR product amplification and for the sequencing walks over clones or PCR products, different sets of parameters were used (**Table 4)** in screening the two groups of candidates.

Table 4. Threshold values of primer picking parameters in consed.

Parameters in primer selection	Threshold values	
	PCR primer	Walking primer
Window size in looking (bp)	2000	450
Min. Length (bp)	18	15
Max. Length (bp)	30	25
Min. Melting Temperature (^{o}C)	55	55
Max. Melting Temperature (^{o}C)	58	60
Max. Match Elsewhere LLR Score*	21	17
Max. Length of Mononucleotide Repeat	4	4
Max. Self-Match LLR Score	-	6
Max. Primer Dimer LLR Score	14	-
Max. Melting Temperature Difference between primer pair (^{o}C)	3	-

*LLR score is a parameter for illustrating the similarity of two sequences (see chapter **1.8**)

Sometimes PCR amplification primers were also used for sequencing walking, and vice versa. In most cases, they were interchangeable and worked quite well. Repeat regions were always carefully avoided in the primer picking process.

3.7 Methods of Gap closure

Based on the scaffolding results (e.g. as shown in **Figure 8**), the intra-scaffold gaps were normally closed by primer walking over spanning clones or PCR products. The detail methods for gap closure in this study could be largely categorized as following (**Figure 12**): 1. Resequencing reads that were located at contig edges by using optimized

conditions, for instance, using Big Dye terminator chemistry in didoxide reactions and/or a longer capillary for separation. This would lead to longer sequencing reads, allowing closure of smaller gaps. 2. Resequencing reads that would fill the gap region. 3. Primer walking over the clones spanning gap regions. 4. Amplifying the gap region from total DNA, and then walking over the PCR product for closure; this method was used when clones covering gap regions were no longer usable.

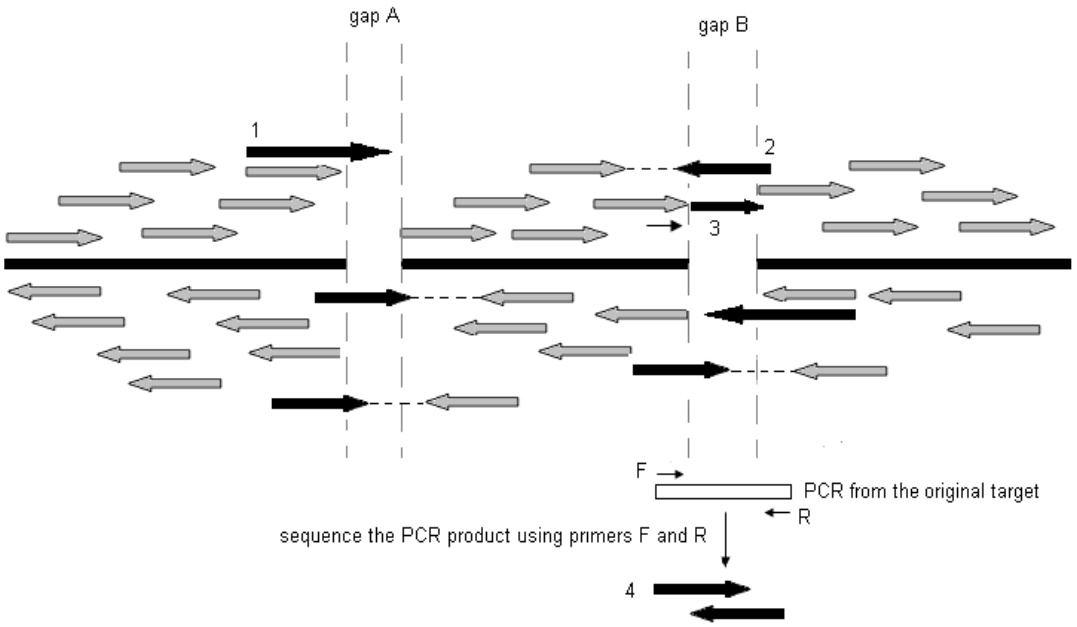


Figure 12. Illustration of gap closure methods. In this study, four methods were applied in filling the sequencing gaps between contigs: 1. resequencing reads located at contig edges by using optimized conditions (Big Dye terminator chemistry, longer capillary separation) in order to generate longer sequencing reads, allowing closure of smaller gaps. 2. resequencing clone ends that would fill the gap region. 3. primer walking over the clones spanning the gap region. 4. amplifying the gap region from total DNA, and walking over the PCR product for closure.

Long repetitive sequences lead to "sequence jump" (**Figure 3**) or other misassemblies because of the inherited limitation of assembly programs (Salzberg and Yorke 2005). The following strategy was utilized to ensure the correct assembly of these long repeats in this study. Here, the case of rRNA operons is used as an example. At first, the orientations and orders of the adjacent contigs spanning rRNA operons were defined from long range pair-

ends (see chapter **3.5.1**) or long PCRs. To guarantee the correct internal assembly of each rRNA region, the five operon regions were firstly manually "cut" out from their original contigs arbitrarily and five additional small "rRNA conitgs" were thus generated. Only the reads, whose mated reads located just outside the operon and orientated toward the operon, were pulled out from the "rRNA conitgs" and assembled back to fill in the corresponding rRNA gaps. Secondly, such shrunken gaps were then filled in by walking over long PCR products or large insert clones that spanned the rRNA regions. The reads remaining on these small "rRNA contigs" either lacked paring information or fell inside the repeat region together with their mating pair. These contigs were excluded from the assembly finally.

3.8 Quality criteria for the *Anabaena* sp. strain 90 genome

Genome sequencing requires more stringent criteria than other projects, like metagenomes or EST sequencing. In this study, we set an even higher standard than normal, i.e. our criteria were that each base should have: 1) three independent, high-quality reads as minimal coverage, 2) sequence coverage on both strands, and 3) a `phred` quality value >= Q40. The human genome sequence quality standards set the error rate of the finished sequence (estimated by `phred` quality score) to be less than one per 10 kb (International Human Genome Sequencing Consortium 2004). Here the accuracy standard was increased by 100 fold (an error rate of less than one base per 1 Mb). To reach such standards, the low quality regions were improved by resequencing clone ends and primer walking over clones or PCR products.

For quality improvement of the Variable Number Tandem Repeat (VNTR) sequences, new plasmid clones were constructed from the PCR products (300-700 bp) of VNTR regions. For each product, 3-5 colonies were picked for DNA extraction. By virtue of the instability of clones when their inserts carry tandem repeats (see chapter **1.9**), sequencing reactions based on modified protocols (see chapter **3.9**) were used to obtain sequences of both ends of these clones.

3.9 Sequencing protocols for finishing at BGI Life Tech Co.,

To achieve better sequencing qualities, ABI Sequencing kits (Applied Biosystems, Foster City, CA, USA) with either BigDye Terminator or dGTP BigDye Terminator were used for all clone end resequencing and primer walking in the finishing phase. Long capillary separation and addition of Dimethyl sulfoxide (DMSO) were also used for sequencing recalcitrant gap regions, like tandem repeats or regions with possible secondary structures.

3.10 Sequence analysis

In this study, sequence alignments were performed by BLAST programs (Altschul *et al.,* 1997, Schäffer *et al.,* 2001). The platform-specific BLAST package was downloaded from the NCBI webpage, and installed locally as instructed. Genomic sequences extracted from assembled contigs were compared with known genes (blastn) and proteins (blastx) for position inference. Multi-sequence alignment was undertaken using the ClustalW program (Thompson *et al.,* 1994).

The Non-redundant protein database was downloaded from the NCBI ftp server (ftp://ftp.ncbi.nlm.nih.gov/blast/db) and formatted locally for 'peptide link' analysis (see chapter **3.5.2**). The reference 16S gene sequence used in this study was AJ133156.

3.11 Repeat analysis

An all-versus-all comparison of the entire *Anabaena* sp. strain 90 genome was made through BLAST search. After removing self-hits, the search results were modified for *de novo* identification and classification of dispersed repeat families by `RECON` (Bao and Eddy 2002).

3.12 Computer system

In this study, all *in silico* analysis of assembly and finishing was performed on a Red Hat Linux system, which was installed on a Dell™ OptiPlex™ GX620 workstation (Intel Pentium D 2.66G CPU, 220G HD, 4G RAM) at the Cyanobacteria group of the University of Helsinki. The total size of raw sequences and edits amounted to over 100 Gigabytes.

4. Results

4.1 High throughput sequencing

In this study, three levels of genomic libraries were prepared for clone end sequencing; the details are listed in **Table 5** below:

Table 5. Summary of high throughput sequencing results.

Library Type	Vector	Clone Size	Clone Number	Clone coverage	Sequencing Primers	Number of Reads
Plasmid	pUC18	2 kb	55,060	18.4 X	M13F, M13R	97,772
Plasmid	pUC18	6 kb	4,383	4.4 X	M13F, M13R	8,067
Cosmid	Lorist6	~40 kb	952	6.3 X	T7, SP6	1,487
Total			60,295			107,326

60,295 clones from the three libraries were picked during the high throughput sequencing phase. The total length of 107,326 derived sequencing reads amounts to 59,294,418 bp, which represented 9.9X coverage according to the estimated genome size (6 Mb).

4.2 Assembly, Scaffolding and Finishing

Based on the `phred` base calling check and primary assembly, an additional 8,673 reads from clone end resequencing and 451 reads from primer walking were added to compensate for unsuccessful reads and allow filling of gaps within scaffolds. Afterwards, 226 sequence contigs were generated. At this point, the genome assembly was fragmented into many separate contigs, which ranged from 46 bp to 284 kb in length. The sum of all contigs was 5,639,373 bp, and the size of the N50 contig (the contig such that the contigs larger than it constitute 50% of the bases of the whole assembly) was 148,038 bp.

With the aid of `consed`, all mated reads in the assembly were screened to build scaffolds and identify misassembly. 736 new cosmid ends were sequenced to speed up this process. Then other sources of information, from BLAST searches and operon checking (see

chapter **3.5.2**), were also applied to aid the finishing process. However, there was no improvement obtained using the 'peptide link' (Fleischmann *et al.,* 1995) scaffolding method (see chapter **3.5.2**). In the results from the operon check, parts of the *nif* (nitrogen fixing) operon and ribosomal RNA operons were found in a few contig ends, suggesting that they could be linked to form scaffolds, and further PCRs also confirmed these predicted relations. All together, four pairs of contig ends (three pairs from rRNA operons and one pair from the *nif* operon) were joined together. Six misassemblies were also corrected with the aid of the cosmid pair-end information. A combination of resequencing of clone ends and primer walks was used to close the internal gaps of these scaffolds (see chapter **3.7**). Finally, 280 newly designed primers and 825 new reads were used for gap closure. At the end of this step, the major contig number decreased to 35 and the N50 contig size increased to 257,679 bp.

The remaining physical gaps in the assembly were resolved using the combinatorial PCR method (see chapter **3.5.3**). Starting with 70 primers in the beginning, six rounds of combinatorial PCR tests were applied all together. The first round was normal PCR and the following five rounds were long PCR reactions. Finally, physical relationships between all contigs were established and the contigs were assembled into five circles. In the last three rounds, eight primers were optimized for increasing specificities and adjusting product sizes. Since primers were purposely selected to avoid the regions containing repetitive sequences and potential secondary structure motifs at contig edges, products of combinatorial multiplex PCRs were always longer than the actual gap sizes. Except for one 13.5-kb long PCR product, the rest of the bands from the combinatorial PCR tests were under 10 kb, and most of them ranged from 3 to 6 kb in size. Thus it was always necessary to do a few walking reactions to fill in the gaps. In all, 210 primers were designed for PCRs and walks, and 240 reads (not including unsuccessful ones) were sequenced in this step.

In order to reach the desired high standard for the quality criteria, resequencing clone ends and primer walks over clones or PCR products were utilized (see chapter **3.8**) again

for resolving remaining low quality and single stranded regions. 118 primers were further designed for PCR reactions and walk sequencing. Together with 87 reads from the newly-made plasmid clones of VNTR inserts, 1,108 sequence reads were generated at this step.

In summary, 60,882 clones from all three size libraries were sequenced, 1,009 primers were designed in the finishing stage, and finally 119,316 reads were collected, which amounted to a final genome coverage of 12.5X.

After checking the finished genome, it was found that the sum of all unclonable regions was only approximately 12 kb in total, which is less than 0.2 % of the total length of the genomic DNA. But these unclonable regions were scattered around the chromosome and plasmids, which led to a laborious finishing process. More than half of the physical gaps were found to be flanked with repetitive sequences.

4.3 Genome structure

The final assembly contained five major contigs as circular structure. These contigs represent one large chromosome and four plasmids (**Table 6**). The total genomic length was 5,306,097 bp, and the overall GC content was 38 %.

Table 6: *Anabaena* sp. strain 90 genome structure.

	Length (bp)	Errors/Mb	GC (%)
Chromosome	4,329,353	0.02	38.1
Plasmid A	819,965	0.03	38.2
Plasmid B	80,384	0.03	37.2
Plasmid C	56,037	0.01	37.3
Plasmid D	20,025	0	37.4
Total	5,306,097		38

Note that the present quality level of the assembly is very high (less than 1 error per 10 Mb). The error rate of each contig (estimated from summing the error probabilities of

each consensus base) was far below the normal standard, which has been defined as less than 100 errors per 1 Mb (International Human Genome Sequencing Consortium 2004). Moreover, there was no evident GC or AT skew variation found in the chromosome and Plasmid A of *Anabaena* sp. strain 90 (**Figure 13**).

4.4 Single nucleotide polymorphisms (SNPs)

Initially, 2,760 putative single nucleotide polymorphism sites were found in the chromosome, 502 in Plasmid A, 48 in Plasmid B, 1,062 in Plasmid C and 21 in Plasmid D by detecting high quality sequence discrepancies in the assembly alignments. With the application of filtration (one SNP site requires confirmation of at least two overlapping reads) and manual checking, only 79 putative SNPs (~2%) were confirmed. 16 were identified on the chromosome, and 63 from Plasmid C, on which the SNPs could be largely categorized into two haplogroups according to the assembly. No SNPs were identified in Plasmids A, B and D. It was surprising that Plasmid C had such a high rate of SNPs given its smaller size. The average density of SNPs over the *Anabaena* sp. strain 90 genome is 15 SNPs per Mb.

4.5 Genome rearrangements

In this project, small insert libraries were constructed with genomic DNAs that were extracted at three different times. Consequently, three sequence indels were identified in small contigs which have partial similarity with major contigs. By designing primers at the edges of the indel boundaries, all three regions were re-amplified from total genomic DNAs extracted at different times, and the rearrangement phenomena were repeated in varied degrees. We further sequenced the PCR products from both directions with the PCR primers. Assembly of the sequencing results confirmed the existence of those sequence divergences. **Figure 14** demonstrates one case of a sequence indel found during assembly. Among the three DNA extractions, a sequence insertion was found in the third DNA sample at a specific location (**Figure 14**). A negative control was amplified in parallel, and sequencing results confirmed the presence of this 529 bp insert.

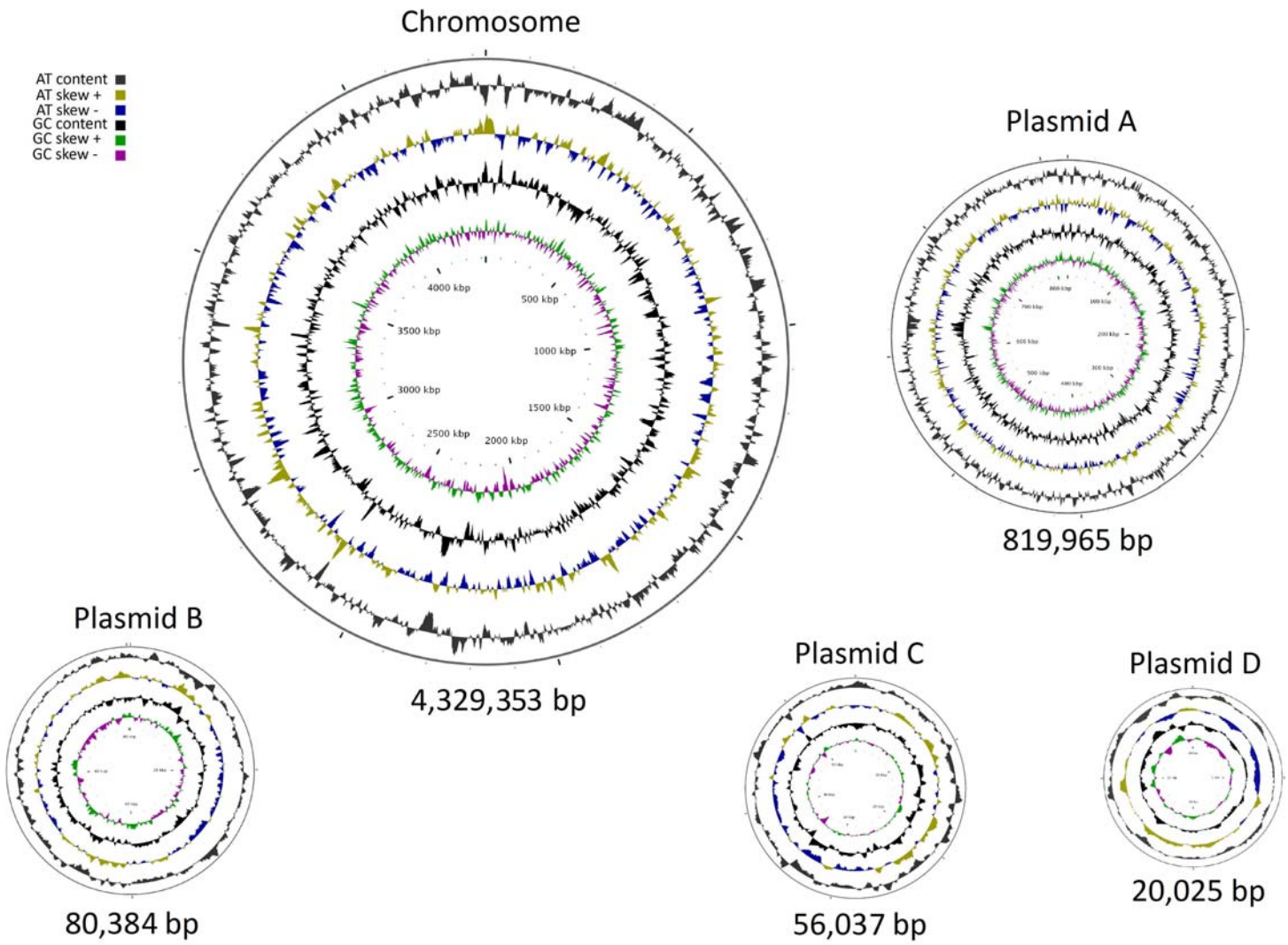


Figure 13. Circular demonstration of *Anabaena* sp. strain 90 genome. AT content (dark grey), AT skew (+: yellow, -: blue), GC content (black) and GC skew (+: green, -: purple) are drawn according to the genomic coordinates. The perimeters of the five circular figures are not proportional to the actual genomic lengths.

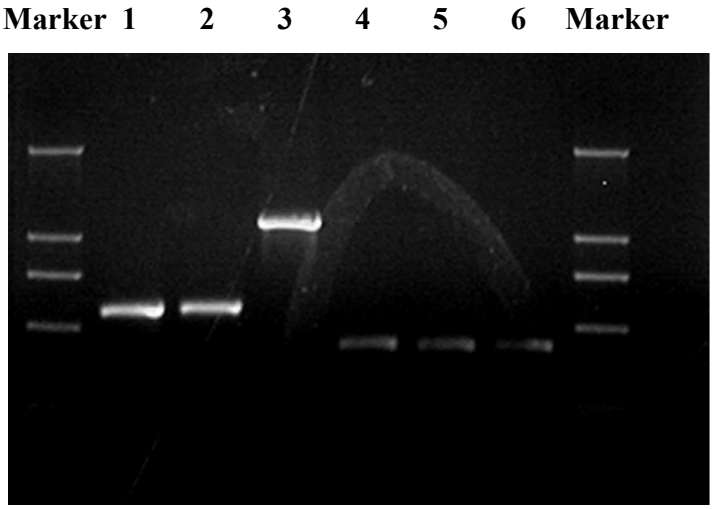


Figure 14. Image of PCR results for three different DNA extractions. Lanes 1, 2 and 3 used the same primer pair for amplifying the insert sequence; Lanes 4, 5, and 6 used the same primer pair for amplifying a control region; Lanes 1 and 4 used the DNA template from the first extraction; Lanes 2 and 5 used the DNA from the second extraction; and Lanes 3 and 6 from the third extraction.

4.6 Repeat sequence families

With pair-wise alignment of genomic sequences (see chapter **3.11**), 613 highly identical repeat families were identified in the *Anabaena* sp. strain 90 genome. These repeat families comprise 3,350 repeat elements in total, and on average there are five elements in each repeat family. The average lengths of most repeat families range from 30 bp to 200 bp (**Figure 15**). The biggest repeat family has two highly identical 9,121-bp elements. The total size of all identified elements is 421,103 bp, which amounts to 7.5% of the total genome.

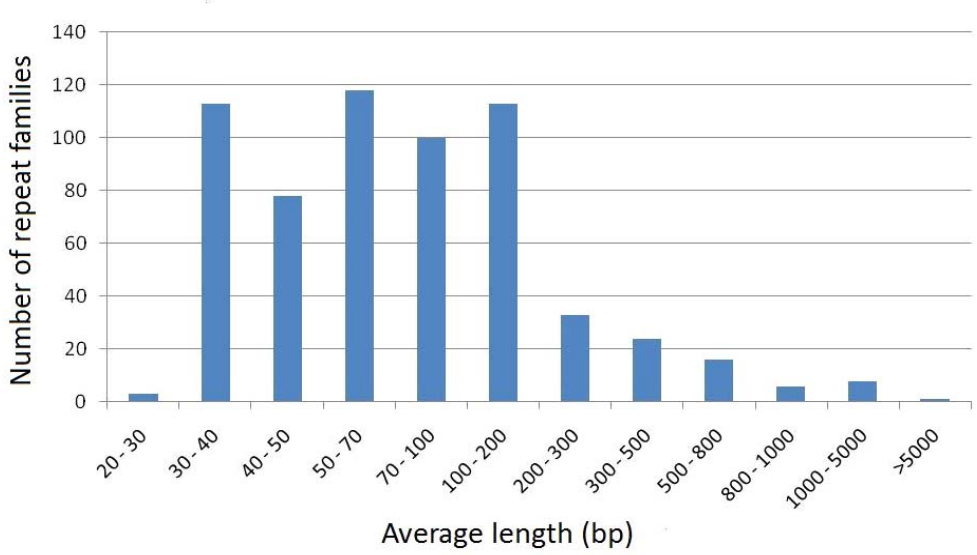


Figure 15. Distribution of average lengths of repeat sequence families in the *Anabaena* sp. strain 90 genome.

4.7 Variable Number Tandem Repeats (VNTRs)

Tandem repeats usually are absent from the list of repeat families identified by local alignment of genomic sequences (Bao and Eddy 2002). In this study, 14 putative VNTRs were recognized in the process of resolving remaining quality defects (see chapter **3.8**). They were all distributed over the chromosome. PCR result (**Figure 16**) revealed the existence of variable length products from these loci. It seems that the length variation of each locus derived from the genomic DNAs with different duplication number of the repeat unit, because differences in the repeat number of those elements were found in the assembled clone end sequences. These clones contained the corresponding VNTR regions and might come from the genomic DNA of different cells or different chromosomes of same cell. The lengths of these tandem repetitive elements were around 8-10 nucleotides (**Table 7**), so that they should be classified as short tandem repeats (STRs). Each VNTR cluster was smaller than 400 bp in size. However, high throughput sequencing methods normally fail to sequence through these areas. An additional cloning step (see chapter **3.8**) and modified sequencing protocols (see chapter **3.9**) were used to characterize these GC-rich tandem repeats. There was a common poly-G/C motif found in nearly all the elements (**Table 7**), and two VNTRs (#8 and #9) were found physically adjacent to each

other.

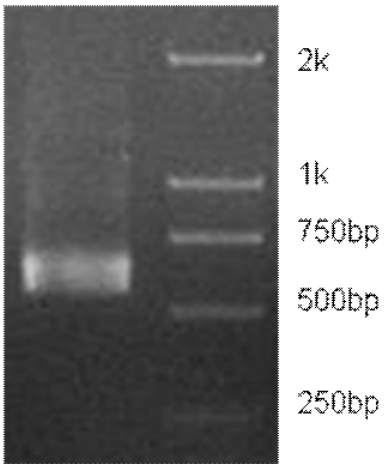


Figure 16. PCR result for one VNTR region. The smearing of the band indicates a variable length of the PCR products.

Table 7. Putative VNTRs found in the *Anabaena* sp. strain 90 strain genome.

Locus ID	Start position (bp)	Repeat element	Repeat number*
1	442,879	TAGGGGAG	3—17
2	696,750	GTAGGGGAG	3—9
3	700,341	GTAGGGGAG	10—12
4	1,411,424	TCCCCAACC	4—19
5	1,516,174	TACTCCCC(C)	12—20
6	1,629,834	AGGGGGG	1—3
7	1,674,734	AAGCAGGGG	3—15
8	1,723,613	AGTTGGGG(G)	12—16
9		TACCTCCCC	2—18
10	1,826,094	(T)ACTTCCCC(C)	17—24
11	2,227,698	TAGGGGA(A)G	2—9
12	2,875,611	AGGTGGGGG	2—10
13	3,910,332	WGGGGWGGT	2—11
14	3,981,006	(G)GGGGAG(G)T	2—14

* The range in the repeat number of the element in each VNTR region was inferred from the end sequences of the clones which contained the corresponding VNTR regions and might come from the genomic DNA of different cells or different chromosomes of same cell.

4.8 Ribosomal RNA operons

Five ribosomal RNA operons were found in the *Anabaena* sp. strain 90 chromosome. Multi-sequence alignment results (**Figure 17**) revealed that three SNPs (537T->C, 549A->G, 940A->G) divide the five 16S RNA genes into two sub-groups. Group one contains three identical 16S RNA genes: *16SGamma*, *16SDelta* and *16SBeta*. Group two is comprised of *16SAlpha* and *16SEpsilon*, which diverge in 5 nucleotides at loci 203, 209-210 and 213-214 (the coordinates refer to those of the submitted sequence AJ133156).

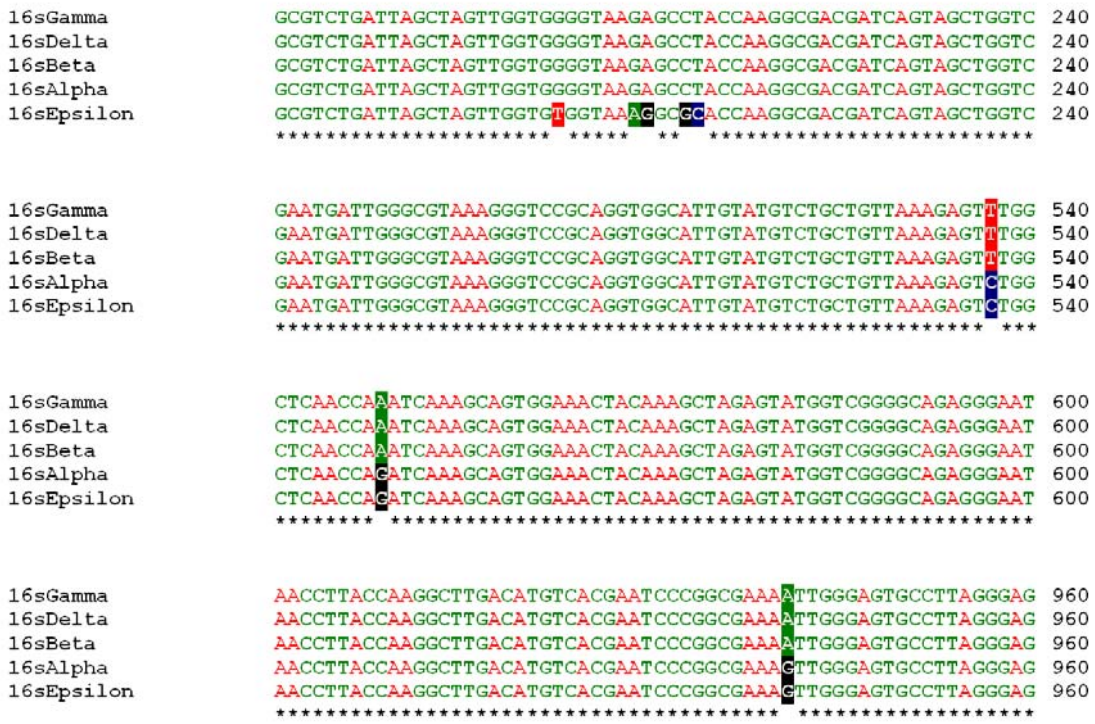


Figure 17. Multi-sequence alignment of the 16S genes in the *Anabaena* sp. strain 90 genome. Sequence coordinates are referenced to the submitted 16S sequence of *Anabaena* sp. strain 90 (AJ133156).

4.9 Contamination

Contamination is common in biological experiments. For the genome project of *Anabaena* sp. strain 90, there were 72 small contigs identified as contaminating

sequences through BLAST searching, and the statistical information for these contigs was tabulated in **Table 8**. Besides *E. coli* contamination, about one-third of the contaminated contigs have homology with Human herpes virus 4 or chicken sequences, which probably originated from errors in the large scale sequencing. Based on **Table 8**, these contaminating contigs only accounted for 0.14 % of reads (171 from 119,316), and the sum of all contaminating contigs (43 kb) was no more than one percent of the complete genome size.

Table 8. Statistics for contaminating contigs.

Contig name	Contig size (bp)	# reads in Contig	Homology
Contig78	359	1	*E. coli*
Contig79	806	1	*E. coli*
Contig86	641	2	Chicken
Contig89	693	2	Chicken
Contig90	680	2	Chicken
Contig93	107	2	Human herpes virus 4
Contig98	107	2	Human herpes virus 4
Contig100	130	2	Human herpes virus 4
Contig102	150	2	Human herpes virus 4
Contig103	121	2	Human char 8
Contig104	94	2	Human herpes virus 4
Contig108	76	2	Human herpes virus 4
Contig110	314	2	Human chr 19
Contig117	123	2	Human herpes virus 4
Contig125	81	2	Human herpes virus 4
Contig129	84	2	Human herpes virus 4
Contig130	146	2	Human herpes virus 4
Contig131	102	2	Human herpes virus 4
Contig132	130	2	Human chr 16
Contig133	152	2	Human herpes virus 4
Contig134	185	2	Human herpes virus 4
Contig135	139	2	Human herpes virus 4
Contig138	94	2	Human herpes virus 4
Contig140	159	2	Human herpes virus 4
Contig143	116	2	Human chr 15
Contig172	1379	2	*E. coli*
Contig173	1160	2	*E. coli*
Contig174	1439	2	*E. coli*
Contig175	98	2	*E. coli*
Contig176	1323	2	*E. coli*

Contig177	1209	2	*E. coli*
Contig178	1311	2	*E. coli*
Contig179	1221	2	*E. coli*
Contig180	590	2	*E. coli*
Contig181	682	2	*E. coli*
Contig182	736	2	*E. coli*
Contig183	1264	2	*E. coli*
Contig184	1267	2	*E. coli*
Contig185	1168	2	*E. coli*
Contig186	855	2	*E. coli*
Contig187	260	2	*E. coli*
Contig188	1356	2	*E. coli*
Contig189	1271	2	*E. coli*
Contig190	75	2	*E. coli*
Contig191	1215	2	*E. coli*
Contig192	1251	2	*E. coli*
Contig193	1309	2	*E. coli*
Contig194	1108	2	*E. coli*
Contig195	1268	2	*E. coli*
Contig196	1181	2	*E. coli*
Contig197	832	2	*E. coli*
Contig198	850	2	*E. coli*
Contig199	1250	2	*E. coli*
Contig201	1184	2	*E. coli*
Contig202	541	2	*E. coli*
Contig208	156	3	Human herpes virus 4
Contig210	167	3	Human chr 9
Contig212	68	3	Human herpes virus 4
Contig214	71	3	Human herpes virus 4
Contig215	93	3	Human herpes virus 4
Contig217	131	3	Human herpes virus 4
Contig218	459	3	Vector
Contig223	1359	2	*E. coli*
Contig224	1305	3	*E. coli*
Contig225	423	3	*E. coli*
Contig229	159	4	Human herpes virus 4
Contig230	115	4	Human herpes virus 4
Contig232	1308	4	*E. coli*
Contig234	142	5	Human herpes virus 4
Contig235	571	5	Vecotr and Mouse cDNA
Contig236	164	6	Human herpes virus 4
Contig239	168	6	Human herpes virus 4
Total	43,301	171	

5. Discussion

5.1 Genome structure

The final assembly contained five major contigs, representing the chromosome and four plasmids respectively. All these replicons were found to have a circular structure according to three pieces of evidence: i) reliable mated reads relationships existed in the head and tail of the contigs; ii) PCR products could be amplified from genomic DNA template with outward-directed primers designed from the head and tail of the contigs; iii) identical sequences were found at the head and tail of the contigs. The circular structure of Plasmids B, C and D were supported by all three standards, while the mega plasmid and the chromosome were supported by ii) and iii). Of the 32 released complete cyanobacterial genomes, six were found to contain plasmids (**Table 1**). The number of plasmids in these organisms ranges from 1 to 9 (Swingley *et al.,* 2008), and the lengths of these plasmids are rather divergent, from a minimum of 2.1 kb (Swingley *et al.,* 2008) to a maximum of 408 kb (Kaneko *et al.,* 2001). Comparatively, the number and size of the plasmids in the *Anabaena* sp. strain 90 genome are intermediate within this range, except that the newly identified mega plasmid (820 kb) was the largest of all known cyanobacterial plasmids sequenced so far. The four plasmids together comprised nearly one-fifth of the total genome length. It should be noted again that the identification of these four plasmids was solely based on the *in silico* analysis of sequencing data. The presence of these plasmids has not been confirmed experimentally.

In general, the copy number of plasmids is higher than that of the chromosome within a bacterial cell. From the assembly of the *Anabaena* sp. strain 90 genome, it could be predicted that Plasmids B and C have relatively higher copy numbers than the chromosome and the megaplasmid in each cell because of the higher ratios of percentage of cosmid clone number versus length percentage in Plasmid B (7.56% versus 1.5%) and Plasmid C (5.99% versus 1.06%).

GC content is a basic parameter describing DNA sequences. Currently, the two more

richly sequenced genera in released cyanobacterial genomes, *Prochlorococcus* and *Synechococcus*, largely defined the lower (30.8%) and the higher (60.8%) bounds of GC content in this phylum (**Table 1**). The GC content of the *Anabaena* sp. strain 90 genome (38%) is relatively low. The chromosome and four plasmids also have similar GC contents to each other, which could be regarded as an evidence of a common origin for these DNA molecules.

5.2 Genome assembly and finishing

Nowadays, the commonly used sequencing approaches (HS and WGS) are both library-based strategies. The finishing process thus mainly aims to fill in the gaps resulting from unclonable regions and genomic areas missing from the library clones, as well as to resolve the misassemblies derived from repetitive sequences. The results of scaffolding and finishing also function as a test of whether the previous work has been adequate or not; sample contamination, clone bias, and other sequencing problems can be identified at this stage. For *Anabaena* sp. strain 90 genome sequencing, three sizes of genomic libraries were utilized. The large insert library (cosmid library) was supposed to link and scaffold contigs on a longer range. However, finishing results revealed that this library had a serious bias in clone distribution. Only 53.3% of the chromosome was covered by cosmid clones (**Figure 18**) while the statistical coverage of cosmid clones was over 6 fold (**Table 5**). It was found that pair-ends of cosmid clones (blue and purple lines) were accumulated in some parts of the chromosome, and left many blank regions in between (**Figure 18**). In addition, a significant fraction of the *Anabaena* sp. strain 90 genome (7.5 percent) was repetitive sequences, the size of repeat elements range from dozens of bases to several kb (**Figure 15**). 68 dispersed repeat families (including 324 repeat elements) have a size over 300 bp, which necessitated intensive efforts of manual misassembly verification. A similar situation has been reported in another newly sequenced toxic cyanobacteria genome *Microcystis aeruginosa* NIES-843 (Kaneko *et al.,* 2007). Nevertheless, more evidence is required to reach the conclusion that a high content of repeats is a common trait of toxic cyanobacterial genomes.

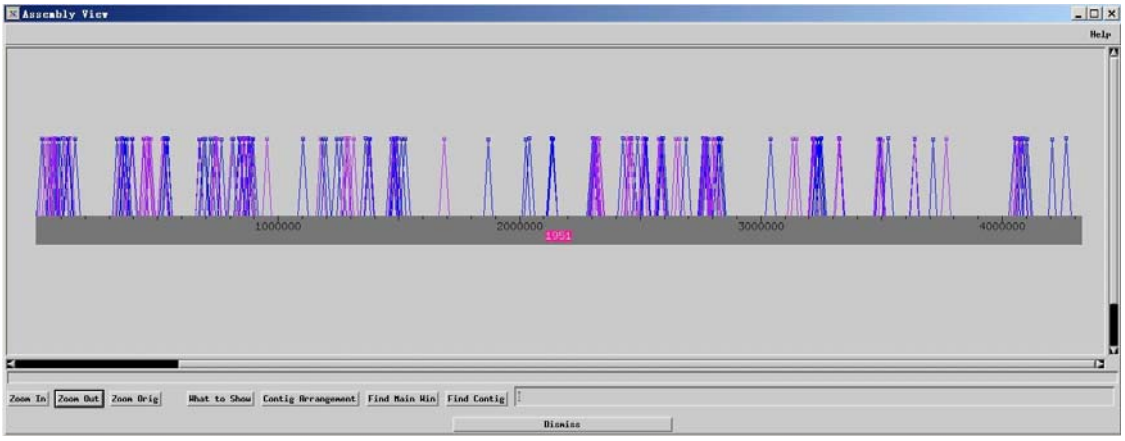


Figure 18. Distribution of cosmid clones along the chromosome. Each blue line represents a pair of cosmid clones, and the purple lines represent the existence of overlapping paired ends.

Since there was no physical data available for this project, a combinatorial multiplex PCR method was tentatively applied to link the remaining contigs in the assembly. In contrast to the random PCR method, combinatorial PCR greatly increases efficiency by using a mixture of primers (up to 16) that are located in the vicinity of contig ends (Sorokin *et al.,* 1996). In the beginning, it had been considered that only gaps under a certain size could possibly be determined because of the ultimate limitation of PCR. Fortunately, all remaining physical gaps in this study were resolvable by PCR or long PCR experiments. Among the products of all combinatorial PCRs, eighty percent of them were less than 6 kb and the largest one was 13.5 kb. In the process, false negatives, which have been reported before (Sorokin *et al.,* 1996), were seen in these combinatorial PCR tests. This might be caused by competition between the multiple primers in one test tube. It was assumed that only the most efficient pair could generate clear enough signals in such multiplex PCR results. This could also explain another phenomenon, which was that there was almost no reaction yielding two or more bands in the PCR results. Even ineffective primers might partly contribute to this outcome; supporting evidence is that six PCR products appeared just in the latter round of tests, but not in the earlier round, where corresponding primers had been pooled together. These primers should be the uncompetitive ones compared to others. Accordingly, the combinatorial PCR test was repeated six times to exhaustively position contig relationships. In general, too many

primers in one pool would cause indecipherable results; while inversely, too few primers would be less efficiency. Based on the combinatorial PCR tests in this study, a better result would be achieved if the number of primers in each pool was in the range of 4 to 6, which would also be helpful in the follow-up data analysis and sequencing. We also found that more than half of the physical gaps were flanked with repetitive sequences. This was consistent with the recurrent scenario of repetitive sequences being unstable in library clones and thus usually absent from genomic libraries. For instance in this genome assembly, it has been shown that one 3-kb physical gap was derived from a tandem repeat region, which was identified as the coding sequence of *GvpA* protein by BLAST search. This protein is known to be lethal to the *E. coli* hosts (Hayes 1995).

In spite of the higher content of repeats in this genome and clone bias of the cosmid library, the completeness and correctness of assembly could be warranted by the middle level (~6 kb) shotgun library. Over 95% of the genome was covered by clones of this library. Firstly, inserts for this library were sheared by sonication; and this fragmentation method has proved to be effective in avoiding library bias. Secondly, the clone inserts were long enough for discernment of most assembly problems caused by repeats in this genome. Results from repeat analysis proved that the repeat sequence families in this genome were smaller than 5 kb, except for one 9-kb repeat family (see chapter **4.6**), which was finally resolved by the combinatorial long PCR tests. This evenly distributed middle level library functioned in verifying the assembly of the chromosome and plasmids. Since reliable and unbiased large insert libraries are always hard to construct, the use of middle level libraries is usually preferred, even in some eukaryotic genome projects (Myers *et al.,* 2000, Venter *et al.,* 2001).

With this completed genome, it could be calculated that the physical gaps identified by the combinatorial PCR were derived from a total of approximately 12 kb of unclonable regions. The average size of physical gaps was only 342 bp, and the maximum one was less than 10 kb. The PCR method is quite suitable for identifying gaps of such a range; this was confirmed by the fluent processing of the combinatorial PCRs during finishing.

However, it is obvious that the PCR method would not help when working with gapped genomes which have long segments (> 40 kb) of physical gaps and/or repeat elements.

A sequencing coverage of 12.5X in the final assembly indicated a higher complexity level of this genome. Large amount of sequencing work had to be employed to resolve the recalcitrant regions for gap closure and quality improvement. In particular, PCR walks always failed in sequencing through the VNTR regions because of their special structure; usually uninterpretable traces were generated from identical units of homopolymers or oligomer sequences when variable length templates existed. In high throughput sequencing, most clones containing tandem repeat sequences could hardly be sequenced through probably due to their instability. The method used in this study, sequencing of freshly made clones containing VNTRs with the modified protocols (see chapter **3.9**), turned out to be quite successful in obtaining high quality sequences of VNTR regions. Here it should be noted that the variation ranges of those repeat elements (**Table** 7) were putatively estimated from reads that covered the corresponding loci. VNTR is characterized as a subset of tandem repeats, which have been discovered previously in cyanobacteria through Southern analysis (Mazel *et al.,* 1990, Rouhiainen *et al.,* 1995) and genome sequencing (Meeks *et al.,* 2001). However, the polymorphism of tandem repeats has not been reported within any strain of the cyanobacteria whose genomes have been sequenced (**Table 1**). It is also known that variation in the unit number of adjacent tandem repeats usually happens between genera or species in bacteria, which can be identified by comparative genomic analysis (Chang *et al.,* 2006). Therefore, VNTRs have been used as genetic markers for typing different bacterial strains (Yazdankhah and Lindstedt 2007). In the *Anabaena* sp. strain 90 genome, interestingly, such length variations of tandem repeats were detected in the same strain, probably between individual cells or different copies of the chromosome within one cell. The cause of this novel phenomenon has yet to be clarified.

5.3 Genomic variations

Anabaena sp. strain 90 has been maintained in continuous pure culture since it was isolated in 1986 (Sivonen *et al.,* 1992). The goal of genome sequencing is to get a snapshot of the genomic content of an organism at a given time point in its long evolutionary history. Magnitudes of genomic variation, ranging from a single base (SNP), a few bases (VNTR), to hundreds of bases (indels), were discovered within the genome of *Anabaena* sp. strain 90. This may be because genomic libraries used in this study were constructed in different years (the cosmid library was constructed in 1992 and plasmid libraries in 2003 and 2004). Therefore, to obtain a coherent assembly and avoid problems with genetic rearrangements, it is preferred that in the beginning a large amount of genomic DNA is extracted all at once for shotgun library construction, PCR templates, and other procedures.

In this study, evidence of genome rearrangements and transpositions were also found. Two ends of some cosmid clones were found located in different DNA molecules of the *Anabeana* sp. strain 90 genome. For instance, one end of some clones was assembled into the chromosome but the other end was found in a plasmid; in other cases, the two ends were found in two different plasmids. This phenomenon is unlikely to be caused by misassembly or sequencing errors because the incidences were found in clones which resided at different genomic sites. Based on the fact that this cosmid library had been constructed twelve years earlier than the small insert libraries, one possible explanation is that DNA molecules within the genome have evolved actively in the intervening years. Mobile genomic elements, like insertion sequences (IS), have been proposed as the cause of genomic rearrangements (Kaneko *et al.,* 2007) due to their frequent presence within cyanobacterial genomes (Kaneko *et al.,* 2001, Kaneko *et al.,* 2007, Nakamura *et al.,* 2002, Nakamura *et al.,* 2003). Cyanophage infection was also suggested as a cause of rearrangements of cyanobacterial genomes (Rocap *et al.,* 2003), and it could also be the origin of the plasmids in the *Anabaena* sp. strain 90 genome. Therefore, the identification of IS and phage proteins will be a natural task in the follow-up genome annotation for the elucidation of the cause of the genome rearrangements.

5.4 Contamination

Based on **Table 8**, it can be concluded that the level of contamination in this project was extremely low with respect to the total volume of data. This indicates that this strain has been purely cultured, and contamination was properly controlled in this sequencing project.

5.5 Origin of replication

Sequence bias (i.e. GC and AT skew) has been frequently used for identification of the origin of replication in bacterial chromosomes (Salzberg *et al.,* 1998, Frank and Lobry 2000, Zhang and Zhang 2002, Worning *et al.,* 2006). The basis of these approaches is the putative cause-and-effect relationship between replication mechanism and asymmetry of nucleotide composition in prokaryotic genomes (Lobry 1996). This connection has been confirmed in most sequenced bacterial strains (Necsulea and Lobry 2007). However, some cyanobacteria belong to the small fraction bacterial species in which base composition skews are absent (Necsulea and Lobry 2007). For the *Anabeana* sp. strain 90 chromosome, there was no evidence for either GC or AT skews (**Figure 13**). A plausible explanation for the total lack of strand bias in some cyanobacterial chromosomes is the existence of multiple functional origins of replication (Nikolaou and Almirantis 2005). As some archaeal chromosomes also share this feature (Necsulea and Lobry 2007), it probably suggests a closer divergent time between cyanobacteria and archaea during their phylogenetic history. With the addition of the *Anabeana* sp. strain 90 genome, a more solid base is being created for more advanced studies on the identification of replication origins of cyanobacteria, both experimentally and computationally.

6. Future perspectives

For a completed genome, annotation is always an important part of subsequent analysis. Its purpose is to pinpoint all biological elements (genes, RNAs and control sequences) from the genomic DNA, and this task is under way now for *Anabaena* sp. strain 90. In the future, the complete gene sets of the *Anabaena* sp. strain 90 genome will open a door to in-depth study of the biology of this toxic cyanobacterium. Functional genomics analysis of toxic and non-toxic peptide production and other aspects of *Anabaena* sp. strain 90 will be deployed in a broad context. Systematic research at the transcriptome and proteome levels, as well as comparative genomic methods, will be utilized to this organism for a comprehensive understanding of physiological features of cyanobacteria such as photosynthesis, nitrogen fixation, and akinetes formation.

7. Acknowledgements

I want to express my sincere thanks to Professor Kaarina Sivonen and Dr. Bin Liu, who provided me the opportunity to be involved in this exciting project. It's their trusting support that ensured the smooth progress of this study. I also owe great thanks to Mr. Chi Zhenfen for his excellent work in the Combinatorial PCRs experiments, and Mr. Yang Shudong for his continuous insistence on tedious sequencing work during the finishing process, as well as to Mr. Li Zhijie for his coordination of the project. I really want to give my special thanks to Dr. David Fewer and Dr. Leo Rouhiainen who had valuable discussions with me and gave me useful comments. My grateful thanks go to all members of the Cyanobacteria group in the Department of Applied Chemistry and Microbiology, University of Helsinki.

Above all, I would like to present my special thanks to my dear wife, Yanhua, for her unlimited support. Also it is the coming little baby who has been constantly pushing me to finish the whole manuscript in the past few months. I will never forget the struggling time in the finishing period. Finally, I would attribute the success of this project to a little bit of lucky and my resolution, which comes from the support of all of the people mentioned above.

This work was financially supported by Research Center of Excellence grants from the Academy of Finland and from the University of Helsinki to Kaarina Sivonen (projects Microbial Resources 53305, Photobiomics 118637).

8. References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., *et al.* 2000. The genome sequence of *Drosophila melanogaster*. Science. 287: 2185-2195.

Adams, M. D., Sutton, G. G., Smith, H. O., Myers, E. W., and Venter, J. C. 2003. The independence of our genome assemblies. Proc. Natl. Acad. Sci. U.S.A. 100: 3025-3026.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389-3402.

Bancroft, I., Wolk, C. P., and Oren, E. V. 1989. Physical and genetic maps of the genome of the heterocyst-forming cyanobacterium *Anabaena* sp. strain PCC 7120. J. Bacteriol. 171: 5940-5948.

Bao, Z. and Eddy, S. R. 2002. Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res. 12: 1269-1276.

Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. P., and Lander, E. S. 2002. ARACHNE: a whole-genome shotgun assembler. Genome Res. 12: 177-189.

Brown, T. A. 2002. The repetitive DNA content of genomes. In: Carlson S. (Ed). Genomes 2. p. 59-66. BIOS Scientific Publishers Ltd, Oxford, UK.

Chang, C. H., Chang, Y. C., Underwood, A., Chiou, C. S., and Kao, C. Y. 2007. VNTRDB: a bacterial variable number tandem repeat locus database. Nucleic Acids Res.

35: D416-21.

Chen, X. and Widger, W. R. 1993. Physical genome map of the unicellular cyanobacterium *Synechococcus* sp. strain PCC 7002. J. Bacteriol. 175: 5106-5116.

Churin, Y. N., Shalak, I. N., Börner, T., and Shestakov, S. V. 1995. Physical and genetic map of the chromosome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC 6803. J. Bacteriol. 177: 3337-3343.

De Marais, D. J. 2000. When did photosynthesis emerge on Earth? Science. 289: 1703-1705.

Douglas, S. E. 1994. Chloroplast origins and evolution. In: Bryant, D. B. (Ed). The Molecular Biology of Cyanobacteria. p. 91–118. Kluwer Academic Publishers, Dordrecht, Boston.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V., Le Gall, F., Makarova, K. S., Ostrowski, M., Oztas, S., Robert, C., Rogozin, I. B., Scanlan, D. J., Tandeau de Marsac, N., Weissenbach, J., Wincker, P., Wolf, Y. I., *et al.* 2003. Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc. Natl. Acad. Sci. U.S.A. 100: 10020-10025.

Eichler, E. E., Clark, R. A., and She, X. 2004. An assessment of the sequence gaps: unfinished business in a finished human genome. Nat. Rev. Genet. 5: 345-354.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8: 186-194.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. 1998. Base-calling of automated

sequencer traces using phred. I. Accuracy assessment. Genome Res. 8: 175-185.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., and Merrick, J. M. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 269: 496-512.

Frangeul, L., Nelson, K. E., Buchrieser, C., Danchin, A., Glaser, P., and Kunst, F. 1999. Cloning and assembly strategies in microbial genome projects. Microbiology. 145: 2625-2634.

Frank, A. C. and Lobry, J. R. 2000. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. Bioinformatics. 16: 560-561.

Fujii, K., Harada, K.-I., Suzuki, M., Kondo, F., Ikai, Y., Oka, H., Carmichael, W.W. and Sivonen, K. 1996. Occurrence of novel cyclic peptides together with microcystins from toxic cyanobacteria, *Anabaena* species. In: T. Yasumoto, Y. Oshima & Y. Fukuyo (Eds). Harmful and Toxic Algal Blooms, p. 559-562. Intergovernmental Oceanographic Comission of UNESCO, Paris.

Gibson, T. J., Rosenthal, A., and Waterston, R. H. 1987. Lorist6, a cosmid vector with BamHI, NotI, ScaI and HindIII cloning sites and altered neomycin phosphotransferase gene expression. Gene. 53: 283-286.

Giovannoni, S. J., Turner, S., Olsen, G. J., Barns, S., Lane, D. J., and Pace, N. R. 1988. Evolutionary relationships among cyanobacteria and green chloroplasts. J. Bacteriol. 170: 3584-3592.

Golden, J. W., Carrasco, C. D., Mulligan, M. E., Schneider, G. J., and Haselkorn, R. 1988. Deletion of a 55-kilobase-pair DNA element from the chromosome during

heterocyst differentiation of *Anabaena* sp. strain PCC 7120. J. Bacteriol. 170: 5034-5041.

Gordon, D., Abajian, C., and Green, P. 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8: 195-202.

Gordon, D., Desmarais, C., and Green, P. 2001. Automated finishing with autofinish. Genome Res. 11: 614-625.

Green, P. 1994. "PHRAP documentation: ALGORITHMS". (URL: http://www.phrap.org/phredphrap/phrap.html).

Green, P. 2002. Whole-genome disassembly. Proc. Natl. Acad. Sci. U.S.A. 99: 4143-4144.

Gregory, S. G., Sekhon, M., Schein, J., Zhao, S., Osoegawa, K., Scott, C. E., Evans, R. S., Burridge, P. W., Cox, T. V., Fox, C. A., Hutton, R. D., Mullenger, I. R., Phillips, K. J., Smith, J., Stalker, J., Threadgold, G. J., Birney, E., Wylie, K., Chinwalla, A., Wallis, J., *et al*. 2002. A physical map of the mouse genome. Nature. 418: 743-750.

Guttman, A., Cohen, A., Heiger, D., and Karger, B. L. 1990. Analytical and micropreparative ultrahigh resolution of oligonucleotides by polyacrylamide gel high-performance capillary electrophoresis. Anal. Chem. 62: 137-141.

Havlak, P., Chen, R., Durbin, K. J., Egan, A., Ren, Y., Song, X. Z., Weinstock, G. M., and Gibbs, R. A. 2004. The Atlas genome assembly system. Genome Res. 14: 721-732.

Hayes, P.K. **1995.** The *gvp* operon of *Anabaena flos-aquae* has multiple copies of a GVPa encoding gene; Expression of the genes in Escherichia coli is lethal. GenBank Accession No. M32060.

Howarth, R. W., Marino, R., Lane, J., and Cole, J. J. 1988. Nitrogen fixation in freshwater, estuarine, and marine ecosystems. 1. Rates and importance. Limnol.Oceanogr. 33: 669-687.

Huang, X. and Madan, A. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9: 868-877.

Huang, X., Wang, J., Aluru, S., Yang, S. P., and Hillier, L. D. 2003. PCAP: a whole-genome assembly program. Genome Res. 13: 2164-2170.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature. 431: 931-945.

Jaffe, D. B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J. P., Zody, M. C., and Lander, E. S. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res. 13: 91-96.

Kaneko, T., Matsubayashi, T., Sugita, M., and Sugiura, M. 1996a. Physical and gene maps of the unicellular cyanobacterium *Synechococcus* sp. strain PCC6301 genome. Plant Mol. Biol. 31: 193-201.

Kaneko, T., Nakamura, Y., Wolk, C. P., Kuritz, T., Sasamoto, S., Watanabe, A., Iriguchi, M., Ishikawa, A., Kawashima, K., Kimura, T., Kishida, Y., Kohara, M., Matsumoto, M., Matsuno, A., Muraki, A., Nakazaki, N., Shimpo, S., Sugimoto, M., Takazawa, M., Yamada, M., *et al.* 2001. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. DNA Res. 8: 205-213; 227-253.

Kaneko, T., Nakajima, N., Okamoto, S., Suzuki, I., Tanabe, Y., Tamaoki, M., Nakamura, Y., Kasai, F., Watanabe, A., Kawashima, K., Kishida, Y., Ono, A.,

Shimizu, Y., Takahashi, C., Minami, C., Fujishiro, T., Kohara, M., Katoh, M., Nakazaki, N., Nakayama, S., *et al.* 2007. Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. DNA Res. 14: 247-256.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., *et al.* 1996b. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3: 109-136.

Kirkness, E. F., Kusiak, J. W., Fleming, J. T., Menninger, J., Gocayne, J. D., Ward, D. C., and Venter, J. C. 1991. Isolation, characterization, and localization of human genomic DNA encoding the beta 1 subunit of the GABAA receptor (GABRB1). Genomics. 10: 985-995.

Kuritz, T., Ernst, A., Black, T. A., and Wolk, C. P. 1993. High-resolution mapping of genetic loci of *Anabaena* PCC 7120 required for photosynthesis and nitrogen fixation. Mol.Microbiol. 8: 101-110.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., and FitzHugh, W. 2001. Initial sequencing and analysis of the human genome. Nature. 409: 860-921.

Lander, E. S. and Waterman, M. S. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics. 2: 231-239.

Latreille, P., Norton, S., Goldman, B. S., Henkhaus, J., Miller, N., Barbazuk, B., Bode, H. B., Darby, C., Du, Z., Forst, S., Gaudriault, S., Goodner, B., Goodrich-Blair, H., and Slater, S. 2007. Optical mapping as a routine tool for bacterial genome

sequence finishing. BMC Genomics. 8: 321.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., Busam, D. A., *et al.* 2007. The diploid genome sequence of an individual human. PLoS Biol. 5: e254.

Lobry, J. R. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13: 660-665.

Luckey, J. A., Drossman, H., Kostichka, A. J., Mead, D. A., D'Cunha, J., Norris, T. B., and Smith, L. M. 1990. High speed DNA sequencing by capillary electrophoresis. Nucleic Acids Res. 18: 4417-4421.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., and Chen, Z. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 437: 376-380.

Mazel, D., Houmard, J., Castets, A. M., and Tandeau de Marsac, N. 1990. Highly repetitive DNA sequences in cyanobacterial genomes. J.Bacteriol. 172: 2755-2761.

McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E. R., Wilson, R. K., Fulton, R., Kucaba, T. A., Wagner-McPherson, C., Barbazuk, W. B., Gregory, S. G., Humphray, S. J., French, L., Evans, R. S., Bethel, G., Whittaker, A., *et al.* 2001. A physical map of the human genome. Nature. 409: 934-941.

Meeks, J. C., Elhai, J., Thiel, T., Potts, M., Larimer, F., Lamerdin, J., Predki, P., and Atlas, R. 2001. An overview of the genome of Nostoc punctiforme, a multicellular,

symbiotic cyanobacterium. Photosynth Res. 70: 85-106.

Mullikin, J. C. and Ning, Z. 2003. The Phusion assembler. Genome Res. 13: 81-90.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H. H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., *et al.* 2000. A whole-genome assembly of *Drosophila*. Science. 287: 2196-2204.

Myers, E. W., Sutton, G. G., Smith, H. O., Adams, M. D., and Venter, J. C. 2002. On the sequencing and assembly of the human genome. Proc. Natl. Acad. Sci. U.S.A. 99: 4145-4146.

Nakamura, Y., Kaneko, T., Sato, S., Ikeuchi, M., Katoh, H., Sasamoto, S., Watanabe, A., Iriguchi, M., Kawashima, K., Kimura, T., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Nakazaki, N., Shimpo, S., Sugimoto, M., Takeuchi, C., Yamada, M., *et al.* 2002. Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. DNA Res. 9: 123-130.

Nakamura, Y., Kaneko, T., Sato, S., Mimuro, M., Miyashita, H., Tsuchiya, T., Sasamoto, S., Watanabe, A., Kawashima, K., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Nakazaki, N., Shimpo, S., Takeuchi, C., Yamada, M., and Tabata, S. 2003. Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. DNA Res. 10: 137-145.

Necsulea, A. and Lobry, J. R. 2007. A new method for assessing the effect of replication on DNA base composition asymmetry. Mol. Biol. Evol. 24: 2169-2179.

Nikolaou, C. and Almirantis, Y. 2005. A study on the correlation of nucleotide skews

and the positioning of the origin of replication: different modes of replication in bacterial species. Nucleic Acids Res. 33: 6816-6822.

Okuda, S., Katayama, T., Kawashima, S., Goto, S., and Kanehisa, M. 2006. ODB: a database of operons accumulating known operons across multiple genomes. Nucleic Acids Res. 34: D358-62.

Osoegawa, K., Mammoser, A. G., Wu, C., Frengen, E., Zeng, C., Catanese, J. J., and de Jong, P. J. 2001. A bacterial artificial chromosome library for sequencing the complete human genome. Genome Res. 11: 483-496.

Palenik, B., Brahamsha, B., Larimer, F. W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E. E., McCarren, J., Paulsen, I., Dufresne, A., Partensky, F., Webb, E. A., and Waterbury, J. 2003. The genome of a motile marine *Synechococcus*. Nature. 424: 1037-1042.

Palenik, B., Ren, Q., Dupont, C. L., Myers, G. S., Heidelberg, J. F., Badger, J. H., Madupu, R., Nelson, W. C., Brinkac, L. M., Dodson, R. J., Durkin, A. S., Daugherty, S. C., Sullivan, S. A., Khouri, H., Mohamoud, Y., Halpin, R., and Paulsen, I. T. 2006. Genome sequence of *Synechococcus* CC9311: Insights into adaptation to a coastal environment. Proc. Natl. Acad. Sci. U.S.A. 103: 13555-13559.

Pop, M., Phillippy, A., Delcher, A. L., and Salzberg, S. L. 2004. Comparative genome assembly. Briefings in Bioinformatics. 5: 237-248.

Rantala, A., Fewer, D. P., Hisbergues, M., Rouhiainen, L., Vaitomaa, J., Borner, T., and Sivonen, K. 2004. Phylogenetic evidence for the early evolution of microcystin synthesis. Proc. Natl. Acad. Sci. U.S.A. 101: 568-573.

Reslewic, S., Zhou, S., Place, M., Zhang, Y., Briska, A., Goldstein, S., Churas, C.,

Runnheim, R., Forrest, D., Lim, A., Lapidus, A., Han, C. S., Roberts, G. P., and Schwartz, D. C. 2005. Whole-genome shotgun optical mapping of *Rhodospirillum rubrum*. Appl.Environ.Microbiol. 71: 5511-5522.

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., Johnson, Z. I., Land, M., Lindell, D., Post, A. F., Regala, W., Shah, M., Shaw, S. L., Steglich, C., Sullivan, M. B., Ting, C. S., *et al.* 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature. 424: 1042-1047.

Ronaghi, M., Uhlen, M., and Nyren, P. 1998. A sequencing method based on real-time pyrophosphate. Science. 281: 363, 365.

Rouhiainen, L., Paulin, L., Suomalainen, S., Hyytiainen, H., Buikema, W., Haselkorn, R., and Sivonen, K. 2000. Genes encoding synthetases of cyclic depsipeptides, anabaenopeptilides, in *Anabaena* strain. 90: 156–167.

Rouhiainen, L., Sivonen, K., Buikema, W. J., and Haselkorn, R. 1995. Characterization of toxin-producing cyanobacteria by using an oligonucleotide probe containing a tandemly repeated heptamer. J.Bacteriol. 177: 6021-6026.

Rouhiainen, L., Vakkilainen, T., Siemer, B. L., Buikema, W., Haselkorn, R., and Sivonen, K. 2004. Genes coding for hepatotoxic heptapeptides (microcystins) in the cyanobacterium *Anabaena* strain 90. Appl.Environ.Microbiol. 70: 686-692.

Salzberg, S. L., Salzberg, A. J., Kerlavage, A. R., and Tomb, J. F. 1998. Skewed oligomers and origins of replication. Gene. 217: 57-67.

Salzberg, S. L. and Yorke, J. A. 2005. Beware of mis-assembled genomes. Bioinformatics. 21: 4320-4321.

Sanger, F. and Coulson, A. R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 94: 441-448.

Schopf, J. W. 2000. The fossil record: tracing the roots of the cyanobacterial lineage. In: Whitton, B. A. & Potts, M. (Eds). The Ecology of Cyanobacteria: Their Diversity in Time and Space. p. 13-35. Kluwer Academic Publisher, Dordrecht, The Netherlands.

Schatz, M. C., Phillippy, A. M., Shneiderman, B., and Salzberg, S. L. 2007. Hawkeye: an interactive visual analytics tool for genome assemblies. Genome Biol. 8: R34.

Schäffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V., and Altschul, S. F. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 29: 2994-3005.

Shizuya, H., Birren, B., Kim, U. J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M. 1992. 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. U.S.A. 89: 8794-8797.

Sivonen, K., and G. Jones. 1999. Cyanobacterial toxins. In: Chorus, I. & Bartram, J. (Eds). Toxic Cyanobacteria in Water: a Guide to Public Health Significance, Monitoring and Management. p. 41-111. Für WHO durch E & FN Spon /Chapman & Hall, London.

Sivonen, K., Namikoshi, M., Evans, W. R., Carmichael, W. W., Sun, F., Rouhiainen, L., Luukkainen, R., and Rinehart, K. L. 1992. Isolation and characterization of a variety of microcystins from seven strains of the cyanobacterial genus *Anabaena*. Appl. Environ. Microbiol. 58: 2495-2500.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B., and Hood, L. E. 1986. Fluorescence detection in automated DNA sequence analysis. Nature. 321: 674-679.

Smith, T. F. and Waterman, M. S. 1981. Identification of common molecular subsequences. J. Mol. Biol. 147: 195-197.

Snelling, W. M., Chiu, R., Schein, J. E., Hobbs, M., Abbey, C. A., Adelson, D. L., Aerts, J., Bennett, G. L., Bosdet, I. E., Boussaha, M., Brauning, R., Caetano, A. R., Costa, M. M., Crawford, A. M., Dalrymple, B. P., Eggen, A., Everts-van der Wind, A., Floriot, S., Gautier, M., Gill, C. A., *et al.* 2007. A physical map of the bovine genome. Genome Biol. 8: R165.

Sorokin, A., Lapidus, A., Capuano, V., Galleron, N., Pujic, P., and Ehrlich, S. D. 1996. A new approach using multiplex long accurate PCR and yeast artificial chromosomes for bacterial chromosome mapping and sequencing. Genome Res. 6: 448-453.

Staden, R., Beal, K. F., and Bonfield, J. K. 1998. The Staden package, Computer Methods in Molecular Biology. 132: 115-130.

Swerdlow, H., Wu, S. L., Harke, H., and Dovichi, N. J. 1990. Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. J. Chromatogr. 516: 61-67.

Swingley, W. D., Chen, M., Cheung, P. C., Conrad, A. L., Dejesa, L. C., Hao, J., Honchak, B. M., Karbach, L. E., Kurdoglu, A., Lahiri, S., Mastrian, S. D., Miyashita, H., Page, L., Ramakrishna, P., Satoh, S., Sattley, W. M., Shimada, Y., Taylor, H. L., Tomo, T., Tsuchiya, T., *et al.* 2008. Niche adaptation and genome expansion in the chlorophyll d-producing cyanobacterium *Acaryochloris marina*. Proc.

Natl. Acad. Sci. U.S.A. 105: 2005-2010.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673-4680.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., *et al.* 2001. The sequence of the human genome. Science. 291: 1304-1351.

Venter, J. C., Adams, M. D., Sutton, G. G., Kerlavage, A. R., Smith, H. O., and Hunkapiller, M. 1998. Shotgun sequencing of the human genome. Science. 280: 1540-1542.

Wang, J., Wong, G. K. S., Ni, P., Han, Y., Huang, X., Zhang, J., Ye, C., Zhang, Y., Hu, J., and Zhang, K. 2002. RePS: A sequence assembler that masks exact repeats identified from the shotgun data. Genome Res. 12: 824-831.

Waterston, R. H., Lander, E. S., and Sulston, J. E. 2002a. On the sequencing of the human genome. Proc. Natl. Acad. Sci. U.S.A. 99: 3712-3716.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., and An, P. 2002b. Initial sequencing and comparative analysis of the mouse genome. Nature. 420: 520-562.

Waterston, R. H., Lander, E. S., and Sulston, J. E. 2003. More on the sequencing of the human genome. Proc. Natl. Acad. Sci. U.S.A. 100: 3022-4; author reply 3025-6.

Welker, M. and von Döhren, H. 2006. Cyanobacterial peptides - nature's own combinatorial biosynthesis. FEMS Microbiol. Rev. 30: 530-563.

Worning, P., Jensen, L. J., Hallin, P. F., Staerfeldt, H. H., and Ussery, D. W. 2006. Origin of replication in circular prokaryotic chromosomes. Environ. Microbiol. 8: 353-361.

Yazdankhah, S. P. and Lindstedt, B. A. 2007. Variable number tandem repeat typing of bacteria. Methods Mol. Biol. 396: 395-405.

Yoshizawa, S., Matsushima, R., Watanabe, M. F., Harada, K., Ichihara, A., Carmichael, W. W., and Fujiki, H. 1990. Inhibition of protein phosphatases by microcystis and nodularin associated with hepatotoxicity. J. Cancer Res. Clin. Oncol. 116: 609-614.

Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C., Zhang, J., Zhang, Y., Li, R., Xu, Z., Li, S., Li, X., Zheng, H., Cong, L., Lin, L., Yin, J., *et al.* 2005. The genomes of *Oryza sativa*: a history of duplications. PLoS Biol. 3: e38.

Zhang, R. and Zhang, C. T. 2002. Single replication origin of the archaeon *Methanosarcina mazei* revealed by the Z curve method. Biochem.Biophys.Res.Commun. 297: 396-400.

9. Appendix

9.1 PHD file example

The following is an example of a PHD file. In such a file, there are two basic parts: the 'comment' section and the 'DNA' section. The 'comment' section is the first part of a PHD file and resides between 'BEGIN_COMMENT' and 'END_COMMENT'; it contains basic information about the sequence generated during sequencing and base calling. The 'DNA' section is after the 'comment' section, and resides between 'BEGIN_DNA' and 'END_DNA'. This part includes three columns: the first column is the base value which has been called; the second column is the corresponding base quality; the numbers in the third column represent the physical coordinates of base peaks in the trace file. The third part of a PHD file is the ancillary information about this sequence; it contains library type, clone ID and primer orientation (forward or reverse). Generation of the ancillary information is optional.

```
BEGIN_SEQUENCE djs74-690.x1

BEGIN_COMMENT
CHROMAT_FILE: djs74-690.x1
ABI_THUMBPRINT: 062036221324000057246033324011
PHRED_VERSION: 0.990722.g
CALL_METHOD: phred
QUALITY_LEVELS: 99
TIME: Thu Jul 27 15:33:49 2000
TRACE_ARRAY_MIN_INDEX: 0
TRACE_ARRAY_MAX_INDEX: 7802
TRIM: 42 559 0.0500
CHEM: term
DYE: big
END_COMMENT

BEGIN_DNA
a 4 11
t 4 21
t 10 31
g 9 45
a 4 55
n 0 66
n 0 78
n 0 91
c 4 101
c 9 107
c 4 113
n 0 140
t 4 148
n 0 164
n 0 176
n 0 187
a 4 197
a 4 216
```

```
c 6 220
t 6 224
g 6 243
a 6 249
a 6 259
c 10 267
t 10 276
g 7 294
c 7 308
t 6 324
g 8 332
t 6 345
c 6 346
c 8 359
t 14 367
g 20 378
c 11 390
a 4 398
n 0 410
g 4 417
t 17 431
c 4 444
c 8 458
a 9 465
c 16 476
t 20 489
c 29 501
t 29 512
a 29 520
/...
 [Part of this file has been deleted for brevity]
/...
a 26 6876
a 29 6889
g 24 6900
c 24 6914
a 16 6924
g 10 6942
g 12 6956
g 8 6964
a 8 6972
a 8 6986
a 17 7002
t 20 7014
a 24 7023
t 29 7036
g 29 7046
a 27 7061
c 25 7074
a 14 7086
c 10 7098
t 8 7109
t 7 7114


END_DNA

END_SEQUENCE

WR{
template determineReadTypes 000727:153352
name: djs74-690
}

WR{
primer determineReadTypes 000727:153352
type: univ fwd
}
```

9.2 sequence file in FASTA format

```
>djs74-690.x1 PHD_FILE: djs74-690.x1.phd.1
attgannncccntnnnaactgaactgctgtcctgcangtccactctagag
gatcccatggcatatccactgcctaggggcccgatgacatgccaaactgc
ccagcccacctctgtcactgatggcatccaaacaagctgactggaggccc
cccaagaattggcccacctggaccatctcacactggtgccagcgtatgcc
acactggggcctaaggacaggcatactcagcctaccactgccatcgctgg
gacctaaagattggcccacctgacacccctgtccccagcaaaacttcacc
acagcctccactaaaaaccacaccctaagacactgagaaaatcaaacact
actgacactgttcctgacaaaattcatacagagactacactatggtatac
atacagaatcaaagccaaagtggcctacccgaacaccaccatagacccat
ctttaggaaaaagtgcccccactacaaaagcaaatgcaaaaaactgaaga
aggactgttataccacatgcacagatatcaacataaggatagaagaaaca
tgaaaaagcagggaaatatgacactttcaagggacgtaattctcagcatg
gatttcagtgaaaaaaaaattgtgaaa
```

9.3 Quality file in FASTA format

The following is an example of a quality file in FASTA format. The quality file always has identical header lines to that of the corresponding sequence file. Each header line starts with a '>' character, and is followed by one or more lines giving the qualities of each base, which are separated by spaces and range from 0 to 99 (from worst to best). The total number of quality values of each read always matches the number of bases in the corresponding sequence file.

```
>djs74-690.x1 PHD_FILE: djs74-690.x1.phd.1
4 4 10 9 4 0 0 0 4 9 4 0 4 0 0 0 4 4 6 6 6 6 6 10 10
7 7 6 8 6 6 8 14 20 11 4 0 4 17 4 8 9 16 20 29 29 29
29 29 30 35 35 44 34 34 34 34 34 34 33 35 37 37 40
35 35 33 29 29 29 34 30 33 40 36 42 42 40 56 56 56
40 40 42 42 42 42 48 56 42 38 38 45 33 24 24 20 27
19 22 33 39 35 35 35 37 26 26 19 10 10 10 17 16 19
33 42 42 42 42 42 44 44 44 47 44 43 43 42 42 42 56
56 44 44 44 44 42 42 42 42 42 42 42 44 42 42 42 56
56 56 56 56 42 42 42 42 42 50 42 39 37 50 46 40 40
40 40 40 35 35 37 40 35 35 35 36 42 43 42 42 42 50
50 42 34 42 42 42 42 42 42 42 50 50 50 44 50 50 56
56 43 42 42 42 42 41 50 50 42 42 43 42 42 43 42 42
42 42 42 44 42 50 41 41 41 41 42 42 50 42 42 41 41
41 41 41 40 36 35 35 35 50 50 50 42 42 35 42 42 41
41 40 40 40 38 41 41 44 44 50 42 41 35 41 41 45 41
41 41 41 41 50 43 43 43 56 56 50 44 44 44 44 42 46
46 50 44 42 50 50 50 50 42 42 42 42 42 42 42 56 50
42 42 42 42 37 35 35 40 40 38 38 38 44 56 43 43 43
43 43 46 56 46 43 42 42 42 42 43 43 46 42 50 44 44
50 50 56 56 41 41 41 51 46 46 41 40 40 40 40 40 42
42 42 42 42 42 42 30 30 37 50 50 42 42 42 40 40 40
56 56 45 40 40 40 40 40 42 42 44 50 50 40 40 40 42
42 42 50 47 42 42 35 35 35 35 36 39 42 43 42 42 42
42 44 44 40 40 40 42 44 44 36 36 35 35 35 44 44 33
33 23 23 23 33 33 42 42 44 42 42 42 40 40 40 40 38
38 37 35 35 38 38 37 40 40 40 40 37 40 29 32 29 26
26 19 19 19 26 25 29 24 24 22 30 30 44 46 42 42 38
35 35 35 42 42 44 42 42 42 42 42 46 42 37 30 29 29
37 29 29 29 29 19 12 12 9 7 7 7 7 8 24 21 19 10 10
10 19 16 20 23 24 24 29 40 37 35 30 30 25 29 29 32
27 29 19 19 19 29 22 26 27 25 19 14 10 14 10 11 12
24 25 32 29 29 29 29 25 26 26 25 25 26 26 29 24 24
16 10 12 8 8 8 17 20 24 29 29 27 25 14 10 8 7 7 6 7
7 15 8 8 11 11 13 14 13 14 12 8 10 11 13 11 10 10 10
10 7 7 13 7 7 7 9 9 7 8 8 9 7 6 6 6 9 9 9 7 10 9 9
8 9 10 10
```

9.4 Vector screened file in FASTA format

The following is an example of a screened file in FASTA format. It was generated by the vector screening program '`cross_match`', and had an identical format to the corresponding sequence file. Vector sequences found in the reads by accurate alignment would be marked by 'X', which would be neglected in assembly.

```
>djs74-690.x1 PHD_FILE: djs74-690.x1.phd.1
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXATGGCATATCCACTGCCTAGGGGCCCGATGACATGCCAAACTGC
CCAGCCCACCTCTGTCACTGATGGCATCCAAACAAGCTGACTGGAGGCCC
CCCAAGAATTGGCCCACCTGGACCATCTCACACTGGTGCCAGCGTATGCC
ACACTGGGGCCTAAGGACAGGCATACTCAGCCTACCACTGCCATCGCTGG
GACCTAAAGATTGGCCCACCTGACACCCCTGTCCCCAGCAAAACTTCACC
ACAGCCTCCACTAAAAACCACACCCTAAGACACTGAGAAAATCAAACACT
ACTGACACTGTTCCTGACAAAATTCATACAGAGACTACACTATGGTATAC
ATACAGAATCAAAGCCAAAGTGGCCTACCCGAACACCACCATAGACCCAT
CTTTAGGAAAAAGTGCCCCCACTACAAAAGCAAATGCAAAAAACTGAAGA
AGGACTGTTATACCACATGCACAGATATCAACATAAGGATAGAAGAAACA
TGAAAAAGCAGGGAAATATGACACTTTCAAGGGACGTAATTCTCAGCATG
GATTTCAGTGAAAAAAAAATTGTGAAA
```

9.5 List of `consed` parameters used in this study

The following list is the complete set of `consed` parameters used in this study. They were modified to meet the specific conditions and requirements in the finishing of *Anabaena* sp. strain 90 genome.

```
consed.showProteinTranslation: false
consed.colorHighlightBackground: Yellow
consed.colorHighlightForegroundX: Black
consed.colorMeansMatch_same: CadetBlue1
consed.colorMeansMatch_different: Orange
consed.colorMeansMatch_clippedEnds: Grey50
consed.colorMeansMatch_consensus: Yellow
consed.colorMeansMatch_lowQualConsensus: Gold3
consed.colorConsensusLabel: Yellow
consed.colorConsensusLabelBackground: Black
consed.colorMeansQualityAgree: Black
consed.colorMeansQualityDisagree: red2
consed.colorMeansQualityConsensusForeground: Purple4
consed.colorMeansQuality0_4: grey39
consed.colorMeansQuality5_9: grey47
consed.colorMeansQuality10_14: grey54
consed.colorMeansQuality15_19: grey60
consed.colorMeansQuality20_24: grey66
consed.colorMeansQuality25_29: grey77
consed.colorMeansQuality30_34: grey86
consed.colorMeansQuality35_39: grey91
consed.colorMeansQuality40_97: white
consed.colorMeansQuality98: grey39
consed.colorMeansQuality99: white
consed.colorMeansEditedAgree: Dark Goldenrod
consed.colorMeansEditedDisagree: red2
consed.colorMeansEditedUnedited: Black
consed.colorMeansEdited98: grey66
consed.colorMeansEdited99: white
consed.colorVerticalScrollbarScrolledDown: green
consed.colorMode_editCursorForeground: White
consed.colorMode_editCursorBackground: Red
consed.colorTracesA: chartreuse3
consed.colorTracesC: CadetBlue1
consed.colorTracesG: Orange
consed.colorTracesT: Red
consed.colorTracesN: Purple
consed.colorTracesPad: Magenta
consed.colorScale: Yellow
consed.colorScaleBackground: Black
consed.colorHighlightedReadNames: Magenta
consed.colorSequencingDirectionArrow: Yellow
consed.colorSequencingDirectionArrowTracesUp: Magenta
consed.colorProteinTranslation: Yellow
consed.colorProteinTranslationLabels: Yellow
consed.colorProteinStartCodon: deep pink
consed.colorProteinStopCodon: deep pink
consed.colorUnmatchedRestrictionFragment: Red
consed.colorRestrictionFragmentPairTooFarApart: Red
consed.colorRestrictionFragmentInContig: black
consed.colorRestrictionFragmentPartlyOffContig: yellow
consed.colorRestrictionFragmentEntirelyOffContig: blue
```

```
consed.colorActualGelRestrictionFragment: black
consed.colorRestrictionDigestScale: black
consed.colorRestrictionDigestCursorIndicator: black
consed.colorRestrictionFragmentsOnTopOfEachOther: purple
consed.colorAssemblyViewScale: black
consed.colorAssemblyViewScaleNumbers: black
consed.colorAssemblyViewContigs: grey45
consed.colorAssemblyViewContigNamesForeground: floral white
consed.colorAssemblyViewContigNamesBackground: Deep Pink
consed.colorAssemblyViewConsistentFwdRevPairs: blue1
consed.colorAssemblyViewConsistentFwdRevPairDepth: green
consed.colorAssemblyViewReadDepth: medium sea green
consed.colorAssemblyViewTooFewConsistentFwdRevPairs: red2
consed.colorAssemblyViewInconsistentFwdRevPair: red2
consed.colorAssemblyViewConsistentGapSpanningFwdRevPair: cyan4
consed.colorAssemblyViewHighlight: yellow
consed.colorAssemblyViewMultipleItemsOnTopOfEachOther: purple
consed.colorAssemblyViewDirectSequenceMatches: darkorange
consed.colorAssemblyViewInvertedSequenceMatches: black
consed.colorAssemblyViewConsistentRestrictionDigestFragment1: cyan1
consed.colorAssemblyViewConsistentRestrictionDigestFragment2: forest green
consed.colorAssemblyViewConsistentRestrictionDigestFragment3: pale green
consed.colorAssemblyViewConsistentRestrictionDigestFragment4: plum
consed.colorAssemblyViewConsistentRestrictionDigestFragment5: burlywood
consed.colorAssemblyViewConsistentRestrictionDigestFragment6: medium sea green
consed.colorAssemblyViewConsistentRestrictionDigestFragment7: peachpuff3
consed.colorAssemblyViewConsistentRestrictionDigestFragment8: plum
consed.colorAssemblyViewConsistentRestrictionDigestFragment9: khaki
consed.colorAssemblyViewConsistentRestrictionDigestFragment10: tan
consed.colorAssemblyViewInconsistentRestrictionDigestFragment: red
consed.colorAssemblyViewCloneEnd: light coral
consed.colorReadPrefixDefault: blue
consed.tagColorEdit: DarkOliveGreen1
consed.tagColorBecomeConsensus: NavajoWhite3
consed.tagColorIgnoreMismatches: SeaGreen1
consed.tagColorIgnoreMatches: orange
consed.tagColorSignificantDiscrepancy: Plum1
consed.tagColorCompression: brown3
consed.tagColorDataNeeded: gold
consed.tagColorComment: SkyBlue1
consed.tagColorTagsOverlap: magenta4
consed.tagColorSequencingVector: LightPink1
consed.tagColorCloningVector: Cyan
consed.tagColorVector: LightSalmon
consed.tagColorOligo: Yellow
consed.tagColorOligo3PrimeEnd: Red
consed.tagColorChimera: DarkSeaGreen3
consed.tagColorContigName: Aquamarine
consed.tagColorPolymorphism: SlateBlue2
consed.tagColorHomozygoteAA: SlateBlue2
consed.tagColorHomozygoteCC: SlateBlue2
consed.tagColorHomozygoteGG: SlateBlue2
consed.tagColorHomozygoteTT: SlateBlue2
consed.tagColorHeterozygoteAC: Pink
consed.tagColorHeterozygoteAG: Pink
consed.tagColorHeterozygoteAT: Pink
consed.tagColorHeterozygoteCG: Pink
consed.tagColorHeterozygoteCT: Pink
consed.tagColorHeterozygoteGT: Pink
consed.tagColorRepeat: DodgerBlue
consed.tagColorPolyPhredRank1: Red
consed.tagColorPolyPhredRank2: Orange
consed.tagColorPolyPhredRank3: MediumSeaGreen
```

```
consed.tagColorPolyPhredRank4: Blue
consed.tagColorPolyPhredRank5: orchid1
consed.tagColorPolyPhredRank6: purple
consed.tagColorIndelSite: DarkCyan
consed.tagColorHeterozygoteIndel: DarkOrange
consed.tagColorHomozygoteIndel: SlateBlue2
consed.tagColorPolymorphismConfirmed: orange
consed.tagColorMatchElsewhereHighQual: green
consed.tagColorMatchElsewhereLowQual: aquamarine
consed.tagColorG_dropout: Plum1
consed.tagColorMarkedHighQuality: orange
consed.tagColorMarkedLowQuality: NavajoWhite3
consed.tagColorAutoFinishExp: turquoise4
consed.tagColorDoNotFinish: saddle brown
consed.tagColorDoNotDoPCR: yellow green
consed.tagColorConsedFixedGoldenPath: tomato4
consed.tagColorCloneEnd: aquamarine
consed.tagColorEditable: aquamarine
consed.tagColorContigEndPair: RosyBrown1
consed.tagColorChangedGenotype: cyan3
consed.tagColorStartNumberingConsensus: DarkSeaGreen3
consed.customTag1:
consed.tagColorCustomTag1:
consed.customTag2:
consed.tagColorCustomTag2:
consed.customTag3:
consed.tagColorCustomTag3:
consed.customTag4:
consed.tagColorCustomTag4:
consed.customTag5:
consed.tagColorCustomTag5:
consed.customTag6:
consed.tagColorCustomTag6:
consed.customTag7:
consed.tagColorCustomTag7:
consed.customTag8:
consed.tagColorCustomTag8:
consed.customTag9:
consed.tagColorCustomTag9:
consed.customTag10:
consed.tagColorCustomTag10:
consed.customTag11:
consed.tagColorCustomTag11:
consed.customTag12:
consed.tagColorCustomTag12:
consed.customTag13:
consed.tagColorCustomTag13:
consed.customTag14:
consed.tagColorCustomTag14:
consed.customTag15:
consed.tagColorCustomTag15:
consed.customConsensusTag1:
consed.tagColorCustomConsensusTag1:
consed.customConsensusTag2:
consed.tagColorCustomConsensusTag2:
consed.customConsensusTag3:
consed.tagColorCustomConsensusTag3:
consed.customConsensusTag4:
consed.tagColorCustomConsensusTag4:
consed.customConsensusTag5:
consed.tagColorCustomConsensusTag5:
consed.customConsensusTag6:
consed.tagColorCustomConsensusTag6:
```

```
consed.customConsensusTag7:
consed.tagColorCustomConsensusTag7:
consed.customConsensusTag8:
consed.tagColorCustomConsensusTag8:
consed.customConsensusTag9:
consed.tagColorCustomConsensusTag9:
consed.customConsensusTag10:
consed.tagColorCustomConsensusTag10:
consed.customConsensusTag11:
consed.tagColorCustomConsensusTag11:
consed.customConsensusTag12:
consed.tagColorCustomConsensusTag12:
consed.customConsensusTag13:
consed.tagColorCustomConsensusTag13:
consed.customConsensusTag14:
consed.tagColorCustomConsensusTag14:
consed.customConsensusTag15:
consed.tagColorCustomConsensusTag15:
consed.defaultTagType: polymorphism
consed.defaultTagOnConsensusNotReads: true
consed.autoFinishMinNumberOfErrorsFixedByAnExp: 0.01
consed.autoFinishRedundancy:   2
consed.autoFinishAverageInsertSize: 1500
consed.primersMaxInsertSizeOfASubclone: 50000
consed.primersMaxMeltingTemp: 60
consed.primersMaxMeltingTempForPCR: 58
consed.primersPickTemplatesForPrimers: false
consed.primersSubcloneFullPathnameOfFileOfSequencesForScreening:
/usr/local/genome/lib/screenLibs/primerSubcloneScreen.seq
consed.primersCloneFullPathnameOfFileOfSequencesForScreening:
/usr/local/genome/lib/screenLibs/primerCloneScreen.seq
consed.primersMinMeltingTemp: 55
consed.primersMinMeltingTempForPCR: 55
consed.searchFunctionsUseUnalignedEndsOfReads: false
consed.searchFunctionsUseLowQualityEndsOfReads: true
consed.inexactSearchForStringMaxPerCentMismatch: 5
consed.onlyAllowOneReadWriteConsedAtATime: false
consed.autoFinishAllowHighQualityDiscrepanciesInTemplateIfConsistentForwardReversePair:
true
consed.printWindowCommand: /usr/bin/X11/xwd | /usr/bin/X11/xpr | /bin/lp -dlevulose
consed.fileOfTagTypes:
consed.assemblyViewShowConsistentFwdRevPairs: false
consed.assemblyViewShowConsistentFwdRevPairDepth: true
consed.assemblyViewShowConsistentFwdRevPairsBetweenDifferentScaffolds: true
consed.assemblyViewShowLegsOnSquaresForConsistentFwdRevPairs: false
consed.assemblyViewShowGapSpanningFwdRevPairs: true
consed.assemblyViewShowWhichInconsistentFwdRevPairs: filtered
consed.assemblyViewShowReadDepth: true
consed.assemblyViewShowRestrictionDigestCutSites: true
consed.assemblyViewFilterSequenceMatchesBySize: false
consed.assemblyViewSequenceMatchesMinSize: 100
consed.assemblyViewSequenceMatchesMaxSize: 10000
consed.assemblyViewAutomaticallyStartWithConsed: false
consed.assemblyViewDisplayTheseTagTypesOnTheseLines:  edit  0  matchElsewhereHighQual  1
matchElsewhereLowQual 2
consed.assemblyViewShowTags: true
consed.autoEditRecalculateHighQualitySegmentsOfReads: true
consed.autoEditConvertCloneEndBasesToXs: true
consed.autoEditTellPhrapNotToOverlapMultiplyDiscrepantReads: true
consed.autoEditTagEditableLowConsensusQualityRegions: true
consed.autoEditMakeFakeRead: false
consed.autoEditMakeFakeReadFromRead1: read1
consed.autoEditMakeFakeReadFromRead2: read2
```

```
consed.autoEditMakeFakeReadName: mama
consed.autoEditMakeFakeReadFastaFilename: mama.fa
consed.autoEditMergeAssembly: false
consed.autoEditSecondaryAceFile: mama.ace
consed.showAllTracesJustShowGoodTraces: true
consed.addAlignedSequenceQualityOfBases: 40
consed.makeLightBackgroundInAlignedReadsWindowAndTracesWindow: false
consed.autoReportPrintScaffolds: false
consed.numberUnpaddedConsensusAtUserDefined: true
consed.autoReportPrintHighQualityDiscrepancies: false
consed.autoReportHighQualityDiscrepanciesExcludeCompressionOrG_dropoutTags: true
consed.autoReportHighQualityDiscrepanciesExcludeMostPads: true
consed.autoReportPrintLowConsensusQualityRegions: false
consed.autoReportPrintSingleSubcloneRegions: false
consed.autoReportPrintSingleStrandedRegions: false
consed.autoReportPrintLinkingForwardReversePairs: false
consed.autoReportPrintFilteredInconsistentForwardReversePairs: false
consed.showAllTracesDoNotShowTraceIfTheseTagsPresent: dataNeeded
consed.nameOfFakeJoiningReadsIncludesAceFileName: false
consed.whenUserScrollsOffWindowMillisecondsBetweenScrolling: 250
consed.whenUserScrollsOffWindowBasesToScrollEachTime: 15
consed.compareContigsUseBandedRatherThanFullSmithWaterman: true
consed.compareContigsBandSize: 50
consed.assemblyViewShowFwdRevPairDepthsInRedIfOnlyThisMany: 1
consed.assemblyViewShowSequenceMatches: true
consed.assemblyViewOKToShowSequenceMatchesBetweenContigs: true
consed.assemblyViewOKToShowSequenceMatchesWithinContigs: true
consed.assemblyViewOKToShowDirectSequenceMatches: true
consed.assemblyViewOKToShowInvertedSequenceMatches: true
consed.assemblyViewOnlyShowSequenceMatchesToAParticularRegion: false
consed.assemblyViewOnlyShowSequenceMatchesToThisContig:
consed.assemblyViewOnlyShowSequenceMatchesToThisRegionLeft: 0
consed.assemblyViewOnlyShowSequenceMatchesToThisRegionRight: 0
consed.assemblyViewOnlyShowSequenceMatchesToEndsOfContigs: false
consed.assemblyViewOnlyShowSequenceMatchesToEndsOfContigsThisFar: 1000
consed.defaultReadPrefix: *
consed.readPrefixesFile: readPrefixes.txt
consed.maxCharsDisplayedForReadPrefix: 1
consed.autoFinishDoNotDoPCRIfThisManyAvailableGapSpanningTemplates: 2
consed.autoFinishDoNotDoUnorientedPCRIfThisManyOrMoreUnorientedPCRReactions: 6
consed.autoFinishDoNotDoOrientedPCRIfGapSizeLargerThanThis: 10000
consed.autoFinishDoNotDoPCRIfEndIsExtendedByReads: true
consed.autoFinishMaxAcceptableErrorsPerMegabase: 0
consed.autoFinishIfNotEnoughFwdRevPairsUseThisPerCentOfInsertSize: 90
consed.primersNumberOfBasesToBackUpToStartLooking: 50
consed.primersMakePCRPrimersThisManyBasesBackFromEndOfHighQualitySegment: 100
consed.primersOKToChoosePrimersInSingleSubcloneRegion: true
consed.primersOKToChoosePrimersWhereHighQualityDiscrepancies: false
consed.primersOKToChoosePrimersWhereUnalignedHighQualityRegion: false
consed.autoFinishCallReversesToFlankGaps: true
consed.autoFinishAllowWholeCloneReads: true
consed.autoFinishAllowCustomPrimerSubcloneReads: true
consed.autoFinishAllowResequencingReads: true
consed.autoFinishAllowResequencingReadsOnlyForRunsAndStops: false
consed.autoFinishAllowDeNovoUniversalPrimerSubcloneReads: true
consed.autoFinishAllowMinilibraries: true
consed.autoFinishAllowPCR: true
consed.autoFinishAllowUnorientedPCRReactions: true
consed.autoFinishAllowResequencingAUniversalPrimerAutofinishRead: true
consed.autoFinishAlwaysCloseGapsUsingMinilibraries: true
consed.autoFinishMaximumFinishingReadLength: 2000
consed.autoFinishSuggestMinilibraryIfGapThisManyBasesOrLarger: 800
consed.autoFinishSuggestSpecialChemistryForRunsAndStops: true
```

```
consed.autoFinishSuggestThisManyMinilibrariesPerGap: 2
consed.primersWindowSizeInLooking: 450
consed.primersAssumeTemplatesAreDoubleStrandedUnlessSpecified: false
consed.alignedReadsWindowInitialCharsWide: 60
consed.alignedReadsWindowInitialCharsHigh: 20
consed.alignedReadsWindowMaxCharsForReadNames: 20
consed.alignedReadsWindowAutomaticallyExpandRoomForReadNames: true
consed.autoFinishAllowResequencingReadsToExtendContigs: true
consed.autoFinishCallHowManyReversesToFlankGaps: 2
consed.autoFinishCloseGaps: true
consed.autoFinishContinueEvenThoughReadInfoDoesNotMakeSense: false
consed.autoFinishCostOfResequencingUniversalPrimerSubcloneReaction:  20
consed.autoFinishCostOfCustomPrimerSubcloneReaction:  60
consed.autoFinishCostOfCustomPrimerCloneReaction:  80
consed.autoFinishCostOfDeNovoUniversalPrimerSubcloneReaction:  60
consed.autoFinishCostOfMinilibrary: 500
consed.autoFinishCoverSingleSubcloneRegions: true
consed.autoFinishCoverLowConsensusQualityRegions: true
consed.autoFinishDebugUniversalPrimerReadsFile: gordon_debug.txt
consed.autoFinishDebugCustomPrimerReadsFile: debug_custom.txt
consed.autoFinishDoNotAllowSubcloneCustomPrimerReadsCloserThanThisManyBases: 200
consed.autoFinishDoNotAllowWholeCloneCustomPrimerReadsCloserThanThisManyBases: 300
consed.autoFinishDoNotFinishWhereTheseTagsAre: doNotFinish editable
consed.autoFinishDoNotExtendContigsWhereTheseTagsAre: doNotFinish
consed.autoFinishDoNotExtendContigsIfTagsAreThisCloseToContigEnd: 50
consed.dumpContigOrderAndOrientationInfoToThisFile:
consed.autoFinishDumpTemplates: false
consed.autoFinishExcludeContigIfOnlyThisManyReadsOrLess: 10
consed.autoFinishExcludeContigIfDepthOfCoverageGreaterThanThis:  50
consed.autoFinishExcludeContigIfThisManyBasesOrLess: 8000
consed.autoFinishHowManyTemplatesYouIntendToUseForCustomPrimerSubcloneReactions: 3
consed.primersMinNumberOfTemplatesForPrimers: 1
consed.autoFinishMinBaseOverlapBetweenAReadAndHighQualitySegmentOfConsensus: 70
consed.autoFinishNumberOfVectorBasesAtBeginningOfAUniveralPrimerRead: 40
consed.autoFinishCDNANotGenomic: false
consed.autoFinishConfidenceThatReadWillCoverSingleSubcloneRegion: 90
consed.autoFinishPrintForwardOrReverseStrandWhenPrintingSubcloneTemplatesForCustomPrime
rReads: true
consed.autoFinishPrintMinilibrariesSummaryFile: false
consed.autoFinishNearGapsSuggestEachMissingReadOfReadPairs: true
consed.autoFinishDoNotIgnoreLCQIfThisManyBasesFromEndOfContigForLCQTagger: 300
consed.checkIfTooManyWalks: true
consed.numberOfColumnsBeforeReadNameInAlignedReadsWindow: 1
consed.compareContigsAlignsThisManyBasesMax: 2000
consed.compressedChromatExtension: .gz
consed.dimLowQualityEndsOfReads: true
consed.dimUnalignedEndsOfReads: false
consed.fakeReadsSpecifiedByFilenameExtension: true
consed.fullPathnameOfAddReads2ConsedScript: /usr/local/genome/bin/addReads2Consed.perl
consed.fullPathnameOfCrossMatch: /usr/local/genome/bin/cross_match
consed.fullPathnameOfPhred: /usr/local/genome/bin/phred
consed.fullPathnameOfMiniassemblyScript: /usr/local/genome/bin/phredPhrap
consed.gunzipFullPath: /usr/local/bin/gunzip
consed.hideSomeTagTypesAtStartup: false
consed.maximumNumberOfTracesShown: 4
consed.navigateAutomaticTracePopup: false
consed.navigateAutomaticAllTracesPopup: false
consed.primersMinimumLengthOfAPrimer: 15
consed.primersMaximumLengthOfAPrimer: 25
consed.primersMinimumLengthOfAPrimerForPCR: 18
consed.primersMaximumLengthOfAPrimerForPCR: 30
consed.primersMaxMeltingTempDifferenceForPCR:   3
consed.primersMaxPCRPrimerPairsToDisplay: 100000
```

```
consed.primersCheckJustSomePCRPrimerPairsRatherThanAll: true
consed.primersNumberOfTemplatesToDisplayInFront: 2
consed.primersMaxLengthOfMononucleotideRepeat: 4
consed.primersBadLibrariesFile: badLibraries.txt
consed.primersLibrariesInfoFile: librariesInfo.txt
consed.primersBadTemplatesFile: badTemplates.txt
consed.primersChooseTemplatesByPositionInsteadOfQuality: true
consed.primersWhenChoosingATemplateMinPotentialReadLength: 350
consed.primersWindowSizeInLookingForPCR: 2000
consed.qualityThresholdForFindingHighQualityDiscrepancies: 40
consed.defaultVectorPathnameForRestrictionFragments:
/usr/local/genome/lib/screenLibs/singleVectorForRestrictionDigest.fasta
consed.fileOfAdditionalRestrictionEnzymes:
consed.commonRestrictionEnzymes: BglII EcoRV NsiI HindIII BamHI XhoI PstI
consed.defaultSelectedRestrictionEnzymes: EcoRV HindIII
consed.restrictionEnzymesActualFragmentsFile: fragSizes.txt
consed.restrictionDigestInitialWindowSizeInTextRows: 45
consed.restrictionDigestDoNoShowAreaOfFragmentsOverThisSize: 50000
consed.showReadsAlphabetically: false
consed.showReadsInAlignedReadsWindowOrderedByFile: false
consed.showReadsInAlignedReadsWindowOrderedByThisFile: readOrder.txt
consed.showABIBasesInTraceWindow: false
consed.tracesWindowInitialPixelHeight: 50
consed.assemblyViewWindowInitialPixelHeight: 500
consed.assemblyViewFileOfTemplatesToNotShow: doNotShowInAssemblyView.fof
consed.assemblyViewCrossMatchMinmatch: 30
consed.assemblyViewCrossMatchMinscore: 60
consed.assemblyViewFindSequenceMatchesForConsedScript:
/usr/local/genome/bin/findSequenceMatchesForConsed.perl
consed.assemblyViewCrossmatchMinmatch: 50
consed.assemblyViewCrossmatchMinscore: 50
consed.assemblyViewSequenceMatchesMinimumSimilarity: 90
consed.tracesWindowInitialPixelWidth: 800
consed.assemblyViewWindowInitialPixelWidth: 800
consed.automaticallyScaleTraces: true
consed.automaticallyScaleTracesSamplePeakHeightFractionOfWindowHeight: 0.99
consed.automaticallyScaleTracesSamplePeakPercentile: 100
consed.verticalTraceMagnification: 30
consed.userDefinedKeys: 14 15
consed.programsForUserDefinedKeys: /bin/echo /bin/echo
consed.argumentsToPassToUserDefinedPrograms:                    argument_for_first_key
argument_for_second_key
consed.tagsToApplyWithUserDefinedKeys: none polymorphismConfirmed
consed.listOfTagTypesToHide: matchElsewhereHighQual matchElsewhereLowQual
consed.listOfOptionalWordsToSaveInListOfReadNames:    forward    reverse    ET    BigDye
customOligo SeqEx FS dyePrimer dyeTerminator doubleStranded singleStranded
consed.extendConsensusWithHighQuality: false
consed.fastStartup: true
consed.fastStartupFile: phd.ball
consed.alwaysRunProgramToGetChromats: false
consed.programToRunToGetChromats: /usr/local/bin/myFavoriteProgram
consed.autoFinishUseLongModelReadRatherThanShort: false
consed.askAgainIfWantToQuitConsedIfThisManyReads: 5000
consed.printWindowInstructions:  Make  sure  that  the  window  you  want  to  print  is
unobscured.  Then click "Yes" to dismiss this box.  Then click on the window you want
to print.  You will hear a beep immediately, then another beep a little later.  Then
the  copy  of  the  window  should  come  off  the  printer  specified  by  your  environment
variable LPDEST.
consed.allowMultipleSearchForStringWindows: false
consed.autoPCRAmplifyFalseProductsOKIfLargerThanThis: 3000
consed.autoPCRAmplifyMakePrimerOutOfFirstRegion: false
consed.autoPCRAmplifyMaybeRejectPrimerIfThisCloseToDesiredProduct: 5000
consed.addNewReadsRecalculateConsensusQuality: true
```

```
consed.addNewReadsPutReadIntoItsOwnContig: ifUnaligned
consed.assemblyViewNumberOfRowsOfTags: 4
consed.warnUserWhenTryingToEditAllReads: true
consed.autoFinishEmulate9_66Behavior: false
consed.primersPCRPrimersGroupedIntoWindowOfThisManyBases: 200
consed.primersLookForThisManyPCRPrimerPairsPerPairOfGroups: 2
consed.autoFinishStandardDeviationsFromMeanFromGapToLookForTemplatesForSuggestingEachMi
ssingReadOfReadPairs:  -1
consed.autoFinishCheckThatReadsFromTheSameTemplateAreConsistent: true
consed.autoFinishDoNotAllowSubcloneCustomPrimerReadsCloseTogether: true
consed.autoFinishDoNotAllowWholeCloneCustomPrimerReadsCloseTogether: true
consed.autoFinishMinilibrariesPreferTemplateIfSizeThisManyStdDevsFromMean:   2
consed.autoFinishMinNumberOfForwardReversePairsInLibraryToCalculateAverageInsertSize: 5
consed.autoFinishIfEnoughFwdRevPairsUseThisManyStdDevBelowMeanForInsertSize: 0.2
consed.autoFinishNewCustomPrimerReadThisFarFromOldCustomPrimerRead: 50
consed.autoFinishMinNumberOfSingleSubcloneBasesFixedByAnExp: 1
consed.autoFinishNumberOfBasesBetweenContigsAssumed: 1000
consed.autoFinishPotentialHighQualityPartOfReadStart: 80
consed.autoFinishPotentialHighQualityPartOfReadEnd: 300
consed.autoFinishPrintCustomNavigationFileForChosenReads: true
consed.autoFinishReversesForFlankingGapsTemplateMustProtrudeFromContigThisMuch: 100
consed.autoFinishTagOligosWhenDoExperiments: true
consed.countPads: false
consed.debugging: 0
consed.ignoreHighQualityDiscrepanciesThisManyBasesFromEndOfAlignedRegion: 5
consed.ignoreUnalignedHighQualitySegmentsShorterThanThis: 20
consed.primersLookThisFarForForwardVectorInsertJunction: 125
consed.primersDNAConcentrationNanomolar:  50
consed.primersMaxMatchElsewhereScore: 17
consed.primersMaxMatchElsewhereScoreForPCR: 21
consed.primersMaxSelfMatchScore: 6
consed.primersMaxPrimerDimerScoreForPCR: 14
consed.primersMinQuality: 30
consed.primersPrintInfoOnRejectedTemplates: true
consed.primersSaltConcentrationMillimolar:  50
consed.primersScreenForVector: true
consed.primersToleranceForDifferentBeginningLocationOfUniversalPrimerReads: 100
consed.primersTooManyVectorBasesInWalkingRead: 10
consed.qualityThresholdForLowConsensusQuality: 25
consed.tagColorPerCentOfBase: 50
consed.uncompressedChromatDirectory: /tmp
consed.whenMakingFakeReadToJoinContigsAddThisManyBasesOnEitherSideOfAlignedRegion: 200
consed.writeThisAceFormat: 2
consed.dumpCoreIfBoundsError: false
consed.autoFinishMinSmithWatermanScoreOfARun: 20
consed.autoFinishDoNotComparePCRPrimersMoreThanThisManyTimes: 1e+09
consed.restrictionDigestMaximumBasesToCompareToVector: 200
consed.restrictionDigestZoomFactor:   2
consed.restrictionDigestZoomFactorForNavigate:  10
consed.restrictionDigestToleranceInPositionUnits: 20
consed.autoPCRAmplifyTooManySeriousFalseMatches: 100
consed.assemblyViewZoomFactor: 1.5
consed.assemblyViewFilterInconsistentFwdRevPairsIfThisClose: 2000
consed.assemblyViewGridCellWidthInPixels:   4
consed.assemblyViewCursorSensitivityInPixels: 4
consed.assemblyViewReadDepthQuality: 20
consed.showAllTracesMaxNumberOfTracesToShowAtOnce: 100
consed.allowFwdRevPairScaffoldsToBeMergedIfThisManyBasesIntersectionOrLess: 1000
consed.justForPrimateProject: false
consed.autoReportAllNeededSpeciesCode: 1
consed.autoReportUseCommasInBigNumbers: true
consed.autoReportPrintToCompareToReich: false
consed.autoReportOnlyAllowSitesThatAreBetweenAcceptableSites: false
```

```
consed.autoReportDeaminationMutationsDeterminedByMoreAccurateMethod: true
consed.autoReportChooseTreesUsingBadData: true
consed.autoReportChooseTreesByCountingDeaminationMutations: true
consed.autoReportChooseTreesUsingKimura: true
consed.autoReportPrintCrudeChimpHumanMutations: false
consed.autoReportPrintPositionsForGraham: false
consed.autoReportPrintAncestralCpGs: false
consed.autoReportPrintCpGMutations: false
consed.autoReportPrintMutationsWithContext: false
consed.autoReportCountAllMutationsML: false
consed.autoReportCountAllMutations: false
consed.autoReportIgnoreMultipleTrees: false
consed.autoReportCountAcceptableColumnsWithNoneOnLeft: false
consed.autoReportPrintFlankedColumns4: false
consed.autoReportUseAnnotationFormat: false
consed.autoReportPrintFlankedColumns3: false
consed.autoReportPrintFlankedColumns2: false
consed.autoReportPrintFlankedColumns: false
consed.autoReportHighQualitySegmentData: false
consed.autoReportGoodReadsBug: false
consed.autoReportDiscrepancyRateInFlankedRegions: false
consed.autoReportDiscrepancyRateInFlankedRegions2: false
consed.autoReportDiscrepancyRateInFlankedRegions4: false
consed.autoReportDiscrepancyRateInFlankedRegions5: false
consed.autoReportSingleSignalOrQuality: false
consed.autoReportLowQualityBasesInHQS: false
consed.autoReportCompareHQSWithLQS: false
consed.autoReportCountColumnsForGroupsOfSpecies: false
consed.autoReportSingleSignalInfo: false
consed.autoReportSingleSignalInfo2: false
consed.autoReportCompareTopAndBottomStrands: false
consed.autoReportCompareTopAndBottomStrandsNoHuman: false
consed.autoReportCompareTopAndBottomStrands2: false
consed.autoReportCompareTopAndBottomStrands3: false
consed.autoReportCompareTopAndBottomStrands4: false
consed.autoReportTopStrandPinnedPosition: 0
consed.autoReportBottomStrandPinnedPosition: 0
consed.autoReportCompareTopAndBottomStrandsWithHuman: false
consed.autoReportPrintLengthsOfAlignedSegmentsOfReads: false
consed.autoReportPrintLengthsOfUnalignedHighQualitySegmentsOfReads: false
consed.autoReportPrintIfReadsAreCorrectlyAligned: false
consed.autoReportCalculateErrorProbabilitiesByComparingPTroPPan: false
consed.autoReportPrintAgreeDisagreeBetweenPairsOfSpecies: false
consed.autoReportPrintAgreeDisagreeBetweenPairsOfSpecies2: false
consed.autoReportFilterSingleSignal: true
consed.autoReportGoodHitReads:
/me2/gordon/primates/checkHumanGenomeBothYears/goodReadsBothYears.txt
consed.autoReportQualityWindowLow: 10
consed.autoReportQualityWindowHigh: 15
consed.autoReportPrintNumberOfIsolatedPadsForEachSpecies: false
consed.autoReportPrintNumberOfIsolatedPads: false
consed.autoReportIsolatedPadsOfReadsWithThisPattern: PTro
consed.autoReportMinNumberOfPerfectlyAlignedBasesBeforeDiscrepancy: 5
consed.autoReportMaxSizeOfDiscrepantRegion: 3
consed.autoReportSizeOfDiscrepantRegion: 1
consed.autoReportPrintMinimumQualityHistogram: false
consed.autoReportPrintDiscrepantRegions: false
consed.autoReportPrintBasesInDiscrepantRegions: false
consed.autoReportPrintDiscrepantRegionsButIgnoreReadsContainingThis:
consed.autoReportBackboneReadHasThisStringInIt: HSap
consed.autoReportPrintDiscrepantRegionsButOnlyIfAboveQualityThreshold: false
consed.autoReportPrintSpeciesAlignment: false
consed.autoReportPrintReadAlignment: false
```

```
consed.autoReportPrintTheseReads: readsToPrint.txt
consed.autoReportPrintReadPositions: false
consed.autoReportPrintChosenReadName: false
consed.autoReportNumbersOfCharactersOfChosenReadNameToBePrinted: 1
consed.autoReportPrefix: 1
consed.autoReportUseOldCriteriaForDeletingColumnsOfPads: false
consed.autoReportDeleteColumnsOfPadsBeforeAdjustingReadQualityValues: true
consed.autoReportFlankingBasesMustBeSingleSignal: false
consed.autoReportMinimumQualityOfFlankingBases: 0
consed.autoReportFlankingBasesMustBeInHighQualitySegment: false
consed.autoReportSpecies: PPan PTro GGor PPyg MMul
```

9.6 Web site references

Phred/Phrap/Consed
http://www.phrap.org/phredphrapconsed.html

The staden package
http://staden.sourceforge.net/

CAP3 and PCAP
http://seq.cs.iastate.edu/

Celera Assembler
http://wgs-assembler.sourceforge.net/

ARACHNE
http://www.broad.mit.edu/wga/

AMOS
http://amos.sourceforge.net/

Phusion Assembler
http://www.sanger.ac.uk/Software/production/phusion/

Atlas Whole Genome Assembly Suite
http://www.hgsc.bcm.tmc.edu/downloads/software/atlas/