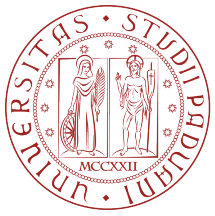


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli studi di Padova

Dipartimento di Biologia

SCUOLA DI DOTTORATO DI RICERCA IN BIOSCIENZE E BIOTECNOLOGIE
INDIRIZZO BIOTECNOLOGIE
CICLO: XXV

Development and Application of a Novel Method for Genome Mapping using Next Generation Sequencing

Direttore della scuola: Ch.mo Prof. Giuseppe Zanotti

Coordinatore d'indirizzo: Ch.mo Prof. Giorgio Valle

Supervisore: Ch.mo Prof. Giorgio Valle

Dottorando: Fabio De Pascale

Sommario

In questa Tesi viene presentato un lavoro svolto nell'ambito del sequenziamento dei genomi. In particolare viene affrontato il problema legato alla creazione di mappe fisiche dei genomi. Le mappe fisiche sono formate da un insieme di informazioni genetiche la cui posizione sul genoma è nota. Queste informazioni genetiche, dette marcatori, possono essere ad esempio geni legati alla manifestazione di caratteri fenotipici. In ultima analisi, comunque, qualsiasi sequenza di DNA può essere considerata un marcatore genetico. Le mappe genomiche sono utlili nel sequenziamento di genomi di nuovi organismi in quanto forniscono dei punti di riferimento per la ricostruzione della sequenza completa.

Questo lavoro presenta un nuovo metodo per produrre le mappe genomiche sfruttando le grandi potenzialità offerte dai sequenziatori di nuova generazione. L'idea centrale del metodo è quella di produrre dei profili di presenza e assenza dei marcatori genetici. Questi profili vengono ottenuti sequenziando porzioni di genoma che ne rappresentino al massimo il 40-50%. Per ottenere queste porzioni di genoma viene utilizzata una libreria di cloni BAC. Riunendo un certo numero di cloni selezionati da questa libreria è possibile produrre dei pool di BAC che rappresentano la porzione desiderata di genoma. Tramite il sequenziamento è quindi possibile identificare i marcatori genetici presenti in ciascun pool di BAC producendo così i profili di presenza e assenza.

Le differenze tra questi profili sono indicative della distanza fisica tra i diversi marcatori. Una volta prodotti questi profili è quindi possibile compararli tra loro in modo da identificare i profili più simili. Profili simili staranno ad indicare che due marcatori sono vicini sul genoma consentendo quindi di posizionarli vicini in una mappa. Alla fine del processo si otterrà quindi una mappa del genoma. Utilizzando i sequenziatori di nuova generazione è possibile utilizzare qualsiasi sequenza si desideri come marcatore.

Questo progetto è stato sviluppato all'intero di un più ampio progetto di sequenziamento del genoma dell'alga unicellulare *Nannochloropsis gaditana*. Il genoma di questo organismo è stato infatti scelto come prova sul campo per questo nuovo metodo.

Abstract

In this Thesis it is presented a new method to produce genome maps. Genome maps are formed by a set of genetic markers whose sequences and positions on the genome are known and defined. Genetic markers could be any kind of DNA sequence, from genes to even smaller sequences. The entire ordered set of genetic markers of a genome constitute its maps. The availability of a such a map in a genome sequencing project could be very useful. In fact, it provides landmarks along the entire target genome that could be used to produce the final and complete sequence.

The aim of the new method proposed in this work is to produce physical maps taking advantage of the next generation sequencing technology. With the high throughput of sequencing that could be reached with these machines any DNA sequence could be a genetic marker. The rationale of this method is to produce profiles of presence and absence of the desired genetic markers. These profiles are produced by sequencing several fractions of the genome, each representing at least its 40-50%. Once these fractions are sequenced it is possible to see, in each of them, which genetic markers are present obtaining the profiles of presence and absence for all genetic marker.

The differences in these profiles give information about the distances on the genome of the genetic markers. By comparing all the profiles one another it is possible to see if two markers are close in the genome. In fact, if two profiles are identical it will means that the two markers are physically close. These information could be used to ordinate the markers on the genome producing its complete map.

In this work this method is developed and applied. The organism chosen as a test filed is the unicellular algae *Nannochloropsis gaditana*. Its genome size (around 30 Mbp) was believed to have the right size to be suitable as a test for this genome sequencing project. Moreover, the presence of a parallel project of sequencing its genome offers the chance to compare such a new method with a sequence produced in a classical way.

Contents

Chapter 1

Introduction

Genomics is a branch of biology whose target is the study of genomes. A genome is the complete set of the genetic information stored within an organism in chromosomes and in any other DNA molecule. Uncover the final and complete sequence of the genetic material of an organism is not an easy task. The genome sequencing projects require several month or even many years to be completed. The final product, the complete sequence of the genetic material, gives access to an enormous amount of information. These information could be useful to many branches of life sciences ranging from genetic engineering to bio-remediation, from studies on human pathogens to studies on hereditary diseases. The final and complete genome sequence of an organism is thus the first step to open all these possibilities to life scientists. Many researchers, all around the world, spend time and money in order to produce genome sequences. Some of them spend their time and money in searching new possibilities and new methods, aimed to produce high quality genome sequences. This Thesis covers the story of some time and money spent to participate at this effort.

1.1 A brief History of Genomics

Genome sequencing projects date back to 1970's. In 1976 and 1977 two bacteriophage genomes were sequeced: the first was the RNA virus MS2 sequenced by Fiers and collegues [1] and the second was the DNA sequence of the phage ϕX174 published by Sanger *et al.* [2]. Few months later, another pubblication by Frederick Sanger actually opened the way to DNA sequencing projects: it was the DNA sequencing with chain terminators method [3]. However, it took many years to uncover the complete sequence of genomes larger than those of virus. In the middle of 1990's three organisms, one for each domain of life, were sequenced: the bacterium *Haemophilus influenzae* [4], the arachaeon *Methanococcus jannaschii* [5] and the eukaryote *Saccharomyces cerevisiae* [6].

Within the following six years, many other genomes throughout the tree of life have been published: the first animals were the model organisms *Caenorhabditis elegans* and *Drosophila melanogaster* [7, 8]; many more eubacteria and some archea, most of them with biomedical interest such as *Eschericia coli* and *Vibrio cholerae* [9, 10]; and the plant model *Arabidopsis thaliana* [11]. However, the most important goal for genomics was the sequencing of the Human genome. Two independent groups published the draft sequence in 2001: the public one, the Interantional Human Genome Sequencing Consortium [12] and the private one headed by J. C. Venter at Celera Genomics [13].

Sequencing strategies

Two strategies were initally proposed for sequencing genomes: one is the original shotgun sequencing and was applied to *H. influenzae* [4] ; and a second one was designed to study larger genomes such those of *S. cerevisiae* and *C. elegans* [6, 7] and is known as hierarchical shotgun.

The **shutgun approach** imply the random shearing of the target genome in smaller picies of given size. These genomic DNA fragments are cloned in recombinant plasmid vector and the resulting colonies are randomly selected and sequenced. The production of the final complete sequence relies on the possibility to find overlaps between the random sequences. Because of this any given base should be sequenced many times. *H. influenzae* has a genome size of 1.8 Mb and to complete the entire genome a 20$\times$ coverage of raw data was produced [4]. The overlapping analysis of such a high number of random sequences requires a lot of computational power. The presence of repeated sequences could complicate this overlapping analysis. For this reason and for

the infrastuctured required by a similar analysis on a large genome, it was thought that this strategy would be more suitable for small genomes with low presence of repeats, such as those of bacteria.

The **hierarchical approach** was thus designed for the study of eukarya genomes that present many more repeats than bacteria genomes. This method relies on the production of a great number of large insert clones, typically BAC [1] clones, that span the entire genome and that are positioned along the chormosomes. Each of these large fragments is then subcloned and shotgun sequenced and assembled. The final complete sequence could be obtained positioning this large sequences back to the map. By means of this process the problem of analyzing random senquences is restricted to each large clone instead than to the entire genome. Because of the procedure, this approach is also known as map-based or clone by clone strategy.

Despite the strategy used for sequencing, the core of the assembly procedure is performed by assembly softwares. These programs are based on algorithms that identified shared portion by overlapping the produced sequences. In this way they produce a consensus for the overlapping sequences. These contiguos stretch of bases are called contigs.

To improve the assembly methods in TODO it was proposed the paired-end method. This technique relies on the possibility to sequence both ends of the insert of a proper recombinant vector. Once both extremities are sequenced, the length of the insert gives the information about the physical distance between the two reads. This method is useful for the assembly procedure because both provides a physical constrain with the insert size and permits a to "jump" between separate contigs to join them in scaffolds[2].

It is possible to identify three main phases within a genome sequencing project: the production of sequences, the assembly of sequences in contigs, and the finishing of the genome. The production phase, despite the strategy chosen for sequencing, largely depends on the throughput of the sequencing facilities. The assembly step, as explained above, relies on the compute power available to perform the overlapping algorithms and in part on the presence of dedicated sequencing strategies such for example the paired-end reads.

The finishing phase can represent the hardest part of the work. This is due to the need of obtaining a high quality complete sequence of the genome

[1] Bacterial Artificial Chromosomes, cloning vectors that could contain from 50 to 150 kb or more of insert.

[2] A scaffold is an ordered sets of contigs and gaps between them. Contigs are joined by means of evidences such as paired reads from both ends of a plasmid insert or mate-pair from next generation sequencing.

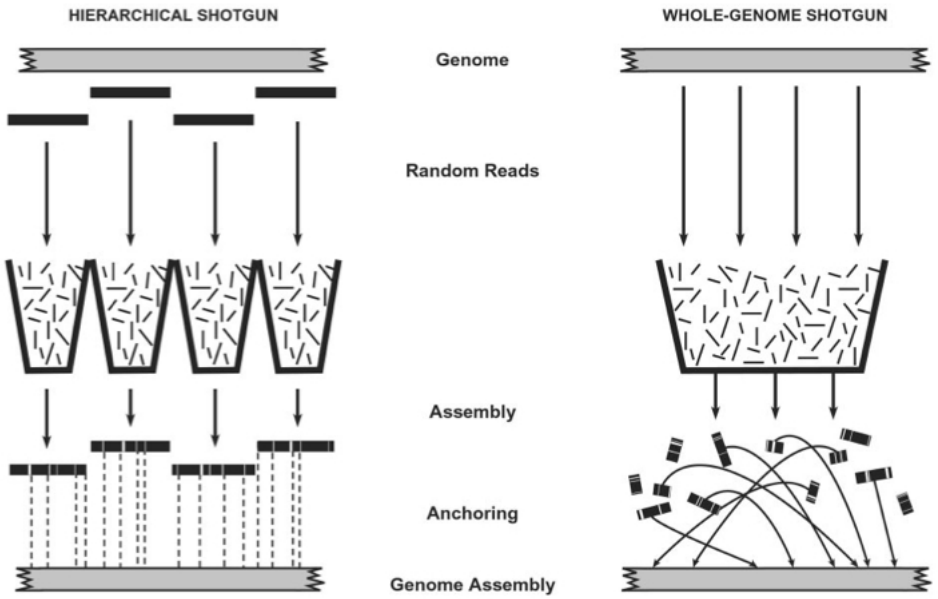


Figure 1.1: Sequencing strategies. (*Left*) The hierarchical shotgun strategy involves the production of a tiling path of overlapping BAC clones covering the entire genome. Each BAC is shotgun sequenced and reassembled, and then the sequences of adjacent clones are merged. The advantage is that all contigs and scaffolds produced from a shotgun sequencing a BAC belong to a single region that is already positioned on genome. (*Right*) Whole-genome shotgun strategy involves the shotgun sequencing on the entire genome and the subsequent reassembly of the produced reads. With this method, each contig and scaffold is an independent component that must be anchored to the genome. To do this, many scaffolds could need directed efforts. Source: Waterston *et al.* PNAS 2002.

of interest. This step usually consist in join together contigs and scaffolds in chromosomes by means of different evidences such as paired-end sequencing of very large insert clones or the presence of maps of the genome. These maps consist in a set of genetic evidences (such as genes for phenotypic traits) along the genome whose position and sequence is known. For these characteristics these information are called **genetic markers**. The availability of such a map of the genome in study facilitate the finishing phase making more easy to place contigs and scaffolds along the genome.

After the successful sequencing of *H. influenzae*, *M. jannaschii* and *M. genitalium* with the shotgun strategy [4, 5, 14] Weber and Myers advanced the hypothesis of a human whole-genome shotgun project [15]. They proposed that with the creation of libraries of different insert sizes sequenced at both extremities, and a proper computational power, the WGS approach could be extended also to very complex genomes. At that time the Human Genome

Project was already started with a hierarchical strategy and genetic and physical maps was under production. Moreover, the hierarchical approach was considered more reliable for the creation of a high quality final sequence of the human genome [16].

However, the challenge lunched by Weber and Myers was taken by J.C. Venter header of the three teams that published *H. influenzae*, *M. jannaschii* and *M. genitalium* and also founder of Celera Genomics. In 2000 Celera published the proof of concept of sequencing a large eukaryotic genome with the whole-shotgun sequencing of the euchromatic portion of *Drosophila melanogaster* [8]. The following year it reached the objective with the publication of the human genome [13].

1.1.1 Human genome sequencing projects

International Human Genome Sequencing Consortium

The Human genome sequencing project performed by the International Consortium (HGP) took a decade to be accomplished and involved twenty centers from six different countries all around the world.

The background of the project is very complex and covers many fields of genomics. Many different studies such as genetic and physical maps, published independently from the HGP paper, contributed in different ways to the final result. However, the proper sequencing project started with production of a large number of large insert clones by digesting the human genome with different restriction enzymes to produce a final library coverage of 65× of the human genome. A genome-wide scale physical map was created by BAC DNA fingerprinting. BAC DNA were digested with a restriction enzyme to create BAC fingerprints to produce fingerprint clone contigs[3] in which BACs are ordered and overlapped. This fingerprint contigs were then positionated on the chromosomes using Sequence Tagged Sites (STS) from existing genetic and physical maps. This mapping procedure was performed using probe hybridization and, later on in the project, sequencing itself.

Selected fingerprint clones were then shotgun sequenced. The sequencing strategies adopted by the different centers varied in terms of library insert sizes, single-strand or double-strand sequencing and in production of paired end sequences or one end sequences. Each center processed, assembled and deposited data according to defined parameters. They produced a total of

[3]In genome assembly, contigs are contiguous blocks of sequence. Here the HGP authors refer to contigs as a contiguos block of fingerprinted BAC.

23 Gb of raw data starting from 29,298 total clones and resulting in a 7.5× average genome coverage.

The assemblies deposited from single large clone sequencing were then assembled in the final draft of the human genome. This process was performed assigning each sequence to its proper fingerprint clone contigs. Then, the fingerprint contigs were mapped on the genome using STS maps, human radiation hybrid maps and genetic maps.

The final draft accounted for 942 fingerprint clone contigs with a N50[4] of 8,398 kb. The published data included both finished and draft sequences.

The actual strength of the project was the worldwide shearing of information: all the genomic sequence data were released without restriction within 24 hours of assembly. Thanks to this organization it was possible to proceed in many different aspect of the project at the same time.

The Celera Human Genome Sequencing Project

The Human genome sequencing project performed at Celera genomics took three years to be accomplished mainly performed in a single big sequencing center at Celera producing in total 175,000 reads per day.

They constructed three libraries with different insert size: 2, 10 and 50 kbp. Both ends of each insert were sequenced resulting in a total of 27,27 million of reads of average length of 543 bp for a total genome coverage of 5.1×.

To realize their assembly they also used data produced by the HGP such as the assembled sequences from BAC clone sequencing, and physical maps information. The assembled data were virtually fragmented in a "synthetic shotgun" data set for a total of 16.05 millions of "faux" reads 550-bp long for a final 2.96× genome coverage.

They performed two different assembly strategies: a whole-genome assembly (WGA) and a compartmentalized shotgun assembly (CSA). In the former the entire set of reads, WGS and faux-WGS, were shotgun assembled without any mapping information. In the latter the WGS data set was divided in subsets by matching with the HGP assembled BAC sequences. After this process the BAC sequences were reduced to "faux" reads and each subset were shotgun assembled. The assemblies resulted in 2218 WGA scaffolds and 1717 CSA scaffolds, for a total of 2.087 and 2.474 Gb.

The resulting scaffolds were then mapped to the genome using STSs physical maps and BAC fingerprinting information produced by the HGP.

[4]The N50 length represents the length L at which the 50% of all assembled nucleotides are contained in contig or scaffold of length L.

Both papers concluded that the proposed sequences are a draft version of the euchromatic portion of the human genome. The complete euchromatic portion of the human genome was published by the HGp in 2004 with the release of the Build-35 version [17].

Comparison between the two human genome sequencing projects

In ref. [18, 19] there is a detailed analysis about the human whole-genome shotgun assemblies performed in the Celera paper. In particular they focused on the intimate effects of using the sequences produced by the HGP as synthetic set of whole-genome shotgun reads. They pointed out that the tiling approach used to reproduce shotgun sequences from the HGP data [13] intimately conserved its own original assembly information [18]. They concluded that the Celera paper did not produced any evidence in supporting the possibility a WGS approach to sequence complex genomes They rather believed that WGS is a very good method to obtain good draft assemblies.

However, despite the discussed success of the whole-genome shotgun strategy to sequence large genome, many more sequencing project started to use WGS or a hybrid approach, like the one used in mouse [20], to sequence even large eukaryotic genomes.

1.1.2 Sequencing technology improvement

The technology improvement to Sanger based sequencers driven by both Human genome sequencing projects made possible the sequencing of many more organisms. Many model organisms and higher eukaryots were sequenced and the number of genomes presents in the public databases rapidly increase. But in 2005, 2006 and again in 2007, sequencing technology undergone an amazing development that causes both a drop in cost and an impressive increase in number of bases that could be produced with a single run [21].

The United States National Human Genome Research Institute (NHGRI) periodically performs an analysis of the costs fo DNA sequencing. To exemplify the revolution on sequencing technologies they perform a comparison between sequencing cost and the Moore's law [22], see figure 1.2. The Moore's law describes the trend of computer hardware development and its associated costs, and predicts a doubling of compute power every two years. As it can be seen in the graph, the sequencing development beats any possible prediction. The impressive drop between 2007 and 2008 marks the transition of

the big sequencing centers from Sanger based sequencers to the next generation sequencing machines.

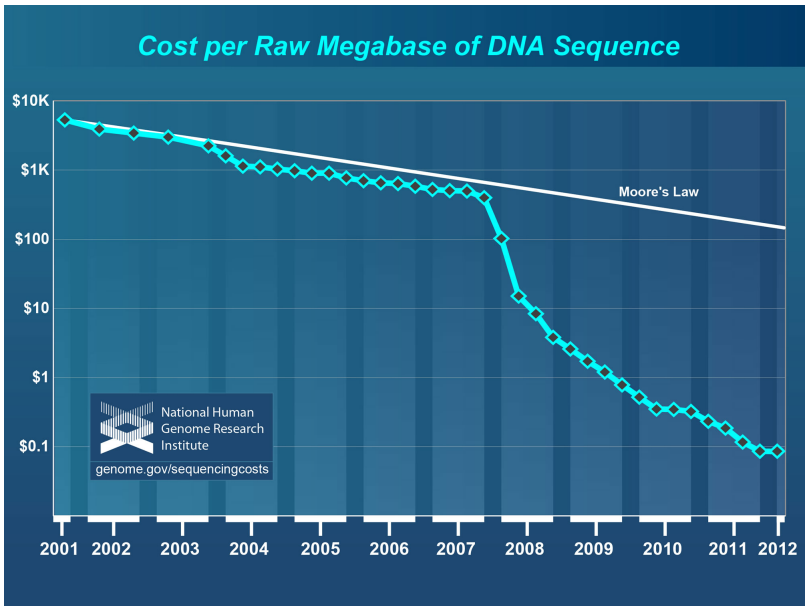


Figure 1.2: Comparison between sequencing costs (cyan line) and Moore's law (withe line). Note the logarithm scale on the y-axis.

During those years, in fact, a number of new sequencers became avaiable to the scientific community. In 2005 454™ Corporation lunches the 454 pyrosequencer, in 2006 Illumina annouches its Genome Analyzer instrument and in 2007-2008 Applied Biosystem commercializes the SOLiD™ system. These new sequencers, and few others with minor commercial success, are globally known as next-generation sequencing (NGS) platforms. The innovative aspects of these machine covers many technology and molecular biology fileds that can be summarized in the following three focal points: (i) each of them adopts a different sequencing chemistry that escapes from the traditional Sanger sequencing, (ii) each one cuts out any *in vivo* cloning steps, (iii) they are capable of producing a huge amount of sequences with a single run. The drawback is that the produced reads are shorter than the classic Sanger reads ranging from 75 with SOLiD™ system to 500 bp with 454™ . As an example of the high throughput reached by these sequencers, SOLiD™ system could produce up to 3 Gb in 10 days.

The sequencing improvement affects not only the actual chemistry and the related technology but also the approaches to sequencing projects. Many new genomics applications have been developed thanks to the presence of next generation sequencing. For example, metagenomics emerged as the tool to

investigate microbial community at the genome level and many aspects once reserved to microarray technology gradually moved into sequencing such as transcriptome or epigenetic studies. NGS also opened the possibility to perform re-sequencing projects to identify genomic variation between individuals or related species such as single nucleotide polymorphisms (SNP) or genomic rearrangements. For purpose of genome assemblies projects it was improved the paired-end sequencing approach.

In the meanwhile, the whole-genome shotgun sequencing approach gradually became an effective method to sequence even very large genomes. This change was mainly due to the huge amount of work required to map clones in the hierarchical approach. Furthermore, with the progress in computing power and the advent of NGS platforms the whole genome shotgun became much more feasible and tempting.

1.1.3 *De novo* assembly with NGS

The sequencing projects are now carried out mainly with next generation sequencing technologies. The re-sequencing projects find their optimal tools in NGS thanks to the high throughput of bases they could reach. In presence of a reference genome the data analysis do not represent a big issue. On the other hand, for *de novo* sequencing projects the analysis is much more complicated. Despite those projects that take advantage of long 454™ reads, the short reads could indeed represent a very big issue. Classic overlapping algorithms in fact, could not computationally manage such a high number of short reads.

Many new genome assembler software, more suitable for assembly short reads data, are based on de Bruijn graph. These programs work essentially in a way that was firstly described by Pevzner *et al.* [23]. Reads are decomposed in seed words named k-mer of given length k. Each k-mer is a node of the graph and nodes are connected if their k-mer are present consecutively on one or more reads. In ref. [24] there is a detailed overview about the different de Bruijn graph based assemblers.

In 2010 Li *et al.* published a *de novo* assembly of two human genomes, one from an asian individual and one from an african one, using short read sequencing [25]. The sequencing was performed entirely with NGS technology. The system chosen by the author was the Illumina Genome Analyzer. The impressive depth of sequencing, however reported only for the asian genome, was a total of 200 Gb divided in 72 Gb for a single-end library and 128 Gb for paired-end libraries of many different size. In the paper they also presented

SOAPdenovo the de Bruijn assembler software they actually used to assemble the produced reads. The assemblies of the two genomes resulted in a N50 for contig of 1050 and 886 and for scaffolds of 446,283 and 61,880 respectively for the Asian and for the African genomes. It has to be pointed out that for the african genome sequencing the paired-end libraries were smaller in respect to those for the asian genome. To evaluate the goodness of the two assemblies they performed a comparison with the reference human genome. They reported a genome coverage of 87.4% for the asian assembly and of 85.4% for the african one, and a gene coverage of 95.5% and 89.2% respectively.

This work represent the ultimate development of the original WGS strategy: in fact, next-generation sequencing and assembly require no maps to be created and rely essentially on a shotgun library preparation. Considering the obtained results, this new method gave assemblies even if not entirely equal but at least comparable to classically produced reference genomes.

Despite this success, in ref. [26] Alkan and co-workers deeply revised the assemblies by Li *et al.* focusing on a comparison of the asian assembly with the reference human genome. In particular they highlighted the shortness of the asian assembly mainly due to mis-assembled sequences. For instance, they identified 420.2 Mbp missing common repeat sequences such as LINE1 and Alu and moreover, they evaluated that only the 56.3% of the genes in the assembly had more than the 95% of their sequence. They concluded that it is critical for comparative genomic studies that the published genomes must be high-quality sequences. Moreover, they suggested that new hybrid approaches that couple many different sequencing technology should be developed to fullfil this target.

1.1.4 Genome Mapping in NGS era

The same suggestion expressed by Alkan *et al.* about *de novo* sequencing and assembly with NGS was already delineated by Lewin and O'Brien *et al.* in 2009 [27]. Their concern regarded the possibility to perform valid studies of comparative genomics in vertebrate: they argued that assemblies of mammalian genomes performed without any physical maps information are poorly useful for comparative genomics. This is because current short-reads sequencing technologies and short-reads assemblers are not able to solve long repetitive regions and chromosome rearrangements, that indeed occurred within vertebrate genomes. They conclude that some efforts should be focused on the production of new methods to produce high-quality physical maps in a rapid and cost-effective way.

In 2011 van Oeveren and collegues published a method, patent by Keygene N.V., to perform physical maps by whole genome profiling [28]. This method coupled the physical consistency offered by BAC libraries with the high-throughput of next-generation sequencing. Whole genome profiling relies on the production and subsequent positioning of "tags" generated by endonuclease digestion of BAC clones. The BAC clones are pooled in a 2 dimensional fashion, row and column of each plate, resulting in a big number of super-samples. These pools are digested and the produced sites recovered with *ad hoc* adapters and then sequenced, each pool independently. A deconvolution step is performed on raw sequences in order to obtain restriction sites originated from a single BAC clone. In the 2-d pooling, in fact, each clone is present in two pools, a row pool and a column pool. This indeed permits the identification of those unique reads that derive from the digestion of a single BAC. The resulting sequences or "tags" are used to construct a fingerprint map ordering restriction sites on the genome according to fingerprint profiles of the BAC clones.

Some plant genome projects adopted this strategy to construct a physical map, for instance the tomato genome consortium [29] and the wheat genome [30]. This method has its major drawback in the 2-d pooling strategy because it implies the production of a high number of pools that must be sequenced independently.

1.2 My Method

Hereafter I will briefly overview the works and principles underlying the method that I propose. Firstly I will present the linkage mapping technique that is the classical approach to perform maps of genetic markers on the genome. Then it follows a discussion about the methods that actually inspired my work: the Radiation Hybrid and the Happy Mapping techniques [31, 32].

1.2.1 Linkage mapping

The principle underlying classical genetic linkage maps is that the probability of recombination between alleles during meiosis could give an estimate about genetic distance between given loci. Assuming an equal frequency of crossing over along the entire chromosome and given for example two loci or genes that lie at the extremities of a chromosome, then it will occur a great number of recombination events between them. This means that a great number of recombinants for those genes will be observed. With the Haldane function it is

possible to convert the observed frequency of recombinants into an estimation of the genetic distance between loci. For review see [33].

Linkage maps produce very good and robust results only in presence of single-gene variable traits [33] and, moreover, they essentially give an information about the order of the genes in the chromosomes and their associated genetic distance. This kind of information could be very useful but in this way only genes with clear phenotypic effects could be mapped and this genes could be assent in wide genomic regions. Moreover, these phenotypic information are usually available only for model organisms and not for "new" species. However, one of the major drawbacks is the amount of time needed to perform cross test between a big number of individuals with different genotypes.

1.2.2 Radiation Hybrid

The Radiation hybrid method allows the analysis of a single chromosome at a time [31]. In brief, a single copy of the target chromosome is contained in a rodent cell; the cell is then irradiated with an high dose of x-rays that cause the breakage of all the chromosomes within it. Treated cell tend to die and so they are rescued by fusion with non irradiated rodent cells. Some of the resulting hybrids will contain fragments of the target chromosome. With southern hybridization it is then possible to verify the presence of given markers in the resulting hybrids. If two markers are present together this means that they are localized on the same fragment of the target chromosome and so that they were close enough to be not separated during x-rays irradiation. However, this is true only in those cases in which only one fragment of the target chromosome is incorporated in the resulting hybrid, otherwise the analysis is much more difficult or even impossible because of the presence of a big number of markers at a time.

Despite this, that was discussed in the original paper, radiation hybrid presents two other major drawbacks. The first is the possibility to analyze just one chromosome at a time with little possibility to implement for high throughput studies. The second is that with x-rays the resulting fragments are about 500 kb long and to obtain a higher density of markers in map, the method needs to be coupled with pulsed field gel electrophoresis [31].

1.2.3 Happy Mapping

The happy mapping method is simpler than the radiation hybrid [32]. Briefly, the genomic DNA is fragmented with γ-rays for long range mapping or by

shearing for short range high resolution mapping. At this point aliquots are taken from the pool of fragments in order to represent one haploid equivalent of the genome. The exact amount of DNA is on the order of pico grams and depends on the estimated size of the genome of interest. Aliquots undergo to PCR to amplify the desired genetic markers and the results are analyzed with gel electrophoresis. As for the radiation hybrid, if two markers lie close on the genome, they will be always on the same random fragments and so they will be detected in the same aliquots.

One of the major drawbacks to face in the happy mapping method is the requirement of haploid equivalents of the genome. This in fact, implies not only the selection of pico gram quantities of DNA but also the need of an amplification step to detect the desired genetic markers.

An important feature that Radiation Hybrid and Happy Mapping share in commom is the need of a priori information about the genomic sequence of the organism in study. In fact, where in the former there is an hybridization procedure to detect markers, in the latter there is a PCR step to amplify the desired markers.

Next generation sequencers could offer the possibility to overcome this limitation. The very high coverage achievable by means of short reads sequencing could permit the development of mapping methods useful for *de novo* sequencing projects in which there is the absence of any previous genomic or genetic information. Moreover, with the present cost per base offered by Illumina Genome Analyzer or Life Technology SOLiD™ system, there is actually the possibility to realize a cost effective and rapid method to perform genome maps.

1.2.4 Profiling of genetic markers

As explained above, classical linkage mapping takes advantage of meiotic events to produce genome maps. Crossing over is the intimate tool to measure distance, while segregation, that separates recombinant chromosomes in different gamets giving rise to new genotypes, gives the possibility to detect recombination events within chromosomes.

Radiation Hybrid and Happy Mapping methods developed different *in vitro* analogues of these natural events. The crossing over is simulated through a random mechanical fragmentation of the genome of interest, respectively through irradiation with x-rays or γ-rays. *In vitro* segregation is accomplished

with two different approaches. In radiation hybrid, in which a single chromosome is analyzed at a time, the formation of the fusion cell separates chromosome fragments. In happy mapping, segregation is obtained by selecting aliquots from the DNA solution that represent one haploid equivalent of the genome.

Fragmentation and segregation are, thus, the core of these genome mapping methods. These two concepts constitute, with slight modification, the basis of the method proposed in this work.

The **fragmentation** of the genome forces genetic markers to co-segregate within genomic fragments. A random breakage assures an even representation of all the markers across the whole fragmented genome. Moreover, if many molecule of the target genome are fragmented, any given marker will be present in more than one random fragment.

Segregation can be easly obtained by producing aliquots of the fragmented genome, as it happens with the Happy Mapping method. Only in presence of 0.5 equivalents or lower of the genome it is possibile to see if two genetic markers are present together because they are actually on the same fragments. In fact, anlyzing the entire genome at a time, even if it is fragmented, the entire set of genetic markers will be present, giving no more information than just that they are on the same genome.

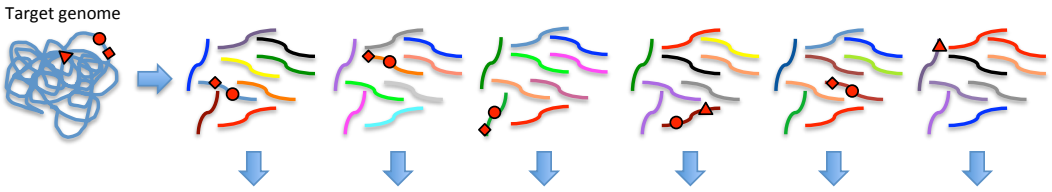
It should be pointed out that, for the purpose of this method, segregation is accomplished in two steps:

- The *co-segregation* of genetic markers during the random fragmentation;
- The *segregation* of the produced fragments in the different aliquots of the genome.

The new and focal aspect of this method is the procedure with which genetic and physical distance are estimated. In presence of a random fragmentation of the genome, genetic markers *co-segregate* depending on their physical distance and on the size of the fragments. At this point, *segregation* it is crucial in order to analyze several different fractions of the fragmented genome: by searching a given marker in all these fractions it is possible to produce its peculiar profile of presence and absence. If this process is performed for any desired markers, each of them will have its proper profile of presence and absence across the entire set of fractions.

If two genetic markers are close in the genome they will be present on the same random fragments. This means that they will segregate in the same fractions and thus, they will be always detected together. In light of this, their

profiles of presence and absence will be exactly the same. This concept can be extended to the analysis of all the markers of the genome: by comparing their profiles it is thus possible to evaluate their distance on the genome. In fact, differences in presence and absence will account for differences in *co-segregation* and so in physical distance, see figure 1.3.



Marker	Fraction 1	Fraction 2	Fraction 3	Fraction 4	Fraction 5	Fraction 6
◆	Yes	Yes	Yes	No	Yes	No
●	Yes	Yes	Yes	Yes	Yes	No
▲	No	No	No	Yes	No	Yes

Figure 1.3: This figure exemplifies the profiling of markers in fractions of the genome. On the left of the cartoon, in ligth blue, there is the original genome. The three geometric figures in red represent three different markers located in the same genomic region but at different physical distances. On the rigth of the cartoon the six group of colored lines represent six different fractions while the lines are different genomic fragments. The table below indicates the results of hypothetic sequencing of the different fractions to identify the presence of the three genetic markers. As can be seen the diamond and the circle have much more evidence in common than with the triangle. In fact, looking back to the genome cartoon we can see that diamond and circle are much more close than each of them with the triangle. For this reason there are few fractions with the triangle and any other markers.

Next generation BAC clone sequencing

To fulfill fragmentation and segregation it was decided to use a BAC library. BAC libraries are produced through partial restriction digestion of the genome of interest. This procedure ensures an acceptable random fragmentation of genomic DNA. Moreover, the procedure of library preparation is customizable in terms of fragment size and library coverage. The size of the DNA fragments can be controlled by tuning the reaction conditions, whereas the high coverage is obtained by producing an high number of clones. This latter aspect is important in order to have many different fragments that come from different portion of the same genomic region. This means that a given marker will be present in many different BAC clones. At the end of the procedure, the actual

average insert size could be estimated with a pulsed field gel electrophoresis on a sub-sample of BAC clones.

Using the information about the average insert size and the estimatimated genome size, it is possible to obtain the appropriate number of clones that represent the desired fraction of the genome. For example, if the disired portion is 12 Mbp and the average insert size of the library is 120 kbp, a hundered of BACs should be selected to obtain this portion. At this point, a pool of BACs could be created by randomly selecting the clones from the library. Iterating this process, several pools can be produced in order to have many different aliquots of genome. Each pool is then sequenced in order to detect any kind of genetic marker. With a next generation sequencer, that produces high coverage with short reads, any sequence could be a genetic marker without any previous knowledge about the DNA sequences of the target markers.

1.2.5 Two sequencing approaches

A genome mapping method shuold be suitable by any genome sequencing project. It is poorly useful for the genomic community a method that is *ad hoc* designed just for one organism or its closely related species. With this vision in mind two complementary methods were developed. These two methods are developed starting from the same theoretical assumptions exposed above about profiling of genetic markers. Both methods thus rely on the production of several BAC pools to produce the final genome map.

The size of the target genome is the first point to be considered in designing the proper approach for that the sequencing project should pursue. These two approaches were designed to be suitable on sequencing projects of organisms with different genome sizes. In presence of a relatively small genome the proposed approach is based on the shotgun sequencing of the BAC pool DNA. On the other hand, in the presence of large genome the proposed method analyzes only the endonuclease restriction sites of the BAC pool DNA.

Shotgun sequencing and mapping

The first approach implies a fragment library preparation protocol for each pool. The idea is to sequence all the inserts of the BAC clones within each pool. The produced reads can give information about presence and absence of any sequence chosen as genetic marker. For instance these markers could be any sequences or contigs produced by an independent sequencing or assembly.

Endonuclease restriction site mapping

In the second method the genetic markers used are the sequence of restriction site and its flanking bases. After the production of the BAC pools the DNA is digested with a single restriction enzyme that recognizes a site of four or six bases depending on the predicted genome size. The digested sites are recovered with a biotinilated custom adapter and then sequenced. This procedure allows sequencing of both side of the restriction site.

Sequencing restriction sites could be useful in mapping large genomes because they represent a fraction of all the possible genetic markers. This in fact, could reduce the complexity during the presence and absence profiling step because of the lower number of profiles to be analyzed.

The custom adapters were designed in order to be used in the SOLiD™ system and contain the sequencing primer and a small barcode: this is a tag sequence, four or eight bases long, useful to identify the produced reads that belong to a given sample. The commercially avaiable kit for multiplex sample preparation by Applied Biosystems™ has the barcodes in the amplification primer and requires a dedicated sequencing reaction. A barcode within the sequencing primer makes possible to mix together many different samples in one *super-sample* starting from the beginning of the protocol. Moreover, the barcodes will be sequenced within the main reaction. These advantages considerably reduce both time and costs during sequencing and library preparation. The drawback is that the bases of the barcodes are stolen from the proper template but the four or eigth bases of the barcode do not affect a proper alignment with a reference.

Figure 1.4: This cartoon illustrates the structure of the resulting construct after ligation of both adapters to the DNA fragment. The barcode next to the amplification primer is the commercial one while the one between sequencing primer and the DNA fragment is the custom one. The blue portion in the DNA fragment marks the bases that flank the endonuclease site in the genome. The red spot marks the position of the biotin.

This approach was designed also to be used in genotyping projects based on the original work published by Miller and collegues [34]. They proposed a method, called RAD, to identify polymorphisms that are associated to endonuclease restriction sites. The protocol implyed the digestion of the genome with a single enzyme, the recovery of the digested sites and their hybridization on a microarray. The hybridization made possible to identify

SNPs and thus to simply genotype the organisms of interest.

A number of different works implemented RAD, or a similar method, with next generation sequencing technologies [35, 36, 37, 38], but few of them take advantage of SOLiD™ platform. In the presence of a reference genome, as in the case of RAD-like genotyping project, the usage of short reads will not affect the identification of polimorphisms. On the other hand, the very high coverage reached by SOLiD™ system will rather improve the number of sites that could be investigated.

Both this methods are covered in this Thesis. For both of them, a number of BAC pools were processed and sequenced. However, a test should be performed to validate the theoretical bases of the overall strategy. The unicellular algae *Nannochloropsis gaditana* was chosen as a test on the field for this genome mapping method. The genome size of this organism was predicted to be around 30 Mbp so the approach used in order to bild its map was the shotgun sequencing of BAC pools.

The Materials and Methods chapter will describe the laboratory procedures to fulfill both the sequencing approaches. The Result and Discussion chapter will describe the preliminary analysis of the sequences produced with both methods but for the largest part discusses the efforts to develop the physical maps method focusing on data obtained with the shotgun sequencing approach.

Chapter 2

Materials and Methods

2.1 The chosen organism

The organism chosen to develop and test this method was *Nannochloropsis gaditana*. This is a unicellular algae that is very promising for the production of biofuels. The sequencing project was included in a wider study focused on the analysis of the conditions that could permit higher production of biofuels. Moreover the availability of the genome sequence was required to permit genetic engineering of target metabolic pathways. The genome sequencing project started more than three years ago with the aim of producing a high quality final sequence. Thus, the strategy chosen was a hybrid approach involving different NGS platforms designed to obtain a high quality final sequence.

A whole-genome shotgun library were constructed and sequenced with Roche 454™ FLX. Two SOLiD™ system mate-paired libraries were sequenced with insert sizes: one 1.5-2 kbp and the other one 2-3 kbp. The 454 sequencing produced an estimated average coverage of 20×. The reads were assembled with Newbler 2.6 resulting in 5910 contigs for a total of 27.96 Mbp with a N50 length of 40.85 kbp. The mate-pair libraries were used to join contigs in order to bild scaffolds using a scaffolding custom program. The final assembly resulted in a total 26.3 Mbp distributed in 58 chromosome scale scaffolds (N50 of 1,052 kbp) for the nuclear genome and a complete assembly for chloroplast

and mitochondrion genomes. Results are in publication [39].

Beside this whole-genome shotgun approach, a BAC library was obtained in order to have a physical reference to confirm possible controversial chromosomes and scaffolds.

This organism and its genome sequencing project represented a great opportunity to develop the genome mapping method proposed in this Thesis.

2.2 BAC Library

The BAC library was purchased from Bio S&T in Montreal, Canada and consists of 11,520 clones in 384 wells microtiter plates. This library was constructed starting from agarose plugs of intact N. gaditana cells. According to our purpose the high molecular weight DNA was partially digested with HindIII endonuclease and cloned into pCC1BAC™ vector from Epicentre®. The average insert size was attested by the supplier at 120,000 bp (data not shown). The predicted library coverage was 45×.

The principal features of pCC1BAC™ vector are the *oriV* high copy origin of replication, and the chloramphenicol resistance. Replication starting from oriV requires the *trfA* gene product. The E. coli strain used for transformation through electroporation was the Epicentre® TransforMax™ Epi300™. The presence of the *trfA* gene in this strain, regulated by an inducible promoter, allows the controlled high copy replication of pCC1BAC™. The inducible solution (arabinose 2%) is added to the growth medium prior or shortly after the bacteria inoculum to a final concentration of 0.01%.

The BAC library was stored at -80°C upon arrival.

2.3 Bacterial cell growth and DNA Extraction

The first problem to face in a project that implies the use of a BAC library, is the DNA extraction from many samples. The second problem is the production of high quality DNA from each clone. Next generation sequencing ensure the high throughput sequencing of sample, but even a small contamination in the preparation will be sequenced with a high coverage. It is mandatory to obtain samples as pure as possible in order to maximize the results.

This project of genome mapping implies the sequencing of a high number of pools of BAC clones. The total number of BAC clones to be processed is thus very high. For the genome of *N. gadiatna* the number of desired pools

is 64 each one containing 96 different BAC clones. The pools will be used to create proper profiles of presence and absence for the genetic markers. These profiles should permit to discriminate between different markers. The number of pools was decided in order to maximize this power of discrimination. If two genetic markers have long profiles and these are equal, these two markers will be certainly together on the genome. The number of BAC clones per pools were decided on the basis of genome size and average insert size of the library see below.

This method thus needs the producing a very high number of single BAC DNA samples. Moreover, the DNA quantity for each sample should be sufficiently high to ensure both trials and sequencing. To achieve high throughput BAC DNA preparation and high quality and quantity of the DNA there was the needs to develop an *ad hoc* method to extract DNA. This procedure was develop by modifying a method previously published by Klein *et al.* [40].

To achieve the high throughput sample preparation the method was designed to be performed on a robotic platform the liquid handling workstation MICROLAB® STAR One from Hamilton Robotics. This robotic platform ensures fast, robust, flexible, parallel and automated procedures on a high number of both 384-well and 96-well microtiter plates as well as on single vials. This robotic platform was *ad hoc* programmed to account to all the needs and steps of this protocol.

2.3.1 Bacterial cell growth

Here follows the list of materials and solutions used to grow bacterial cells.

- Luria-Bertani medium, LB (1 L): 10 g of tryptone, 5 g of yeast extract, 10 g NaCl;
- Terrific Broth medium, TB (1 L): 12 g of tryptone, 24 g of yeast extract, 4 mL of glycerol;
- 384 wells Corning Costar plate: square bottom wells with a maximum capacity of 110 μL;
- 96 deep-wells plate: round bottom wells with a maximum capacity of 1200 μL;
- 384 pin replicator tool V & P Scientific Inc;
- Inducible solution: arabinose 2%;

- Chloramphenicol: 37.5 mg/mL, added to medium to allow selective growth of *E. coli* cells carrying the BAC.

Sixteen 384 wells microtiter plates were selected from the BAC library to be processed. These account for a total of 6,144 BAC clones, the total number of clones required for the method. Plates were grown with the following procedure. All liquid handling steps were performed with MICROLAB® STAR One; volumes refer to quantities for each well.

1. 384 wells microtiter plates containing 80 μL of LB medium in each well (plus chloramphenicol 18.7 μg/mL) were inoculated with frozen BAC library cells using a replicator 384 pin tool;

2. Plates were covered with plastic lids and grown for 21 hours at 37°C in owen with orbital shaking at 900 rpm (Heidolph Titramax 1000 coupled with Heidolph Inkubator 1000);

3. After growth, 384 wells microtiter plates were split in 96 deep-well plates, one quadrant per 96-well plates. 6 μL of grown culture were inoculated in 300 μL of TB medium (plus chloramphenicol 18.7 μg/mL and arabinose 0.01%);

4. 96-well plates were covered with plastic lids and grown for 21 hours at 37°C in owen with orbital shaking at 900 rpm.

2.3.2 DNA extraction

The actual DNA extraction procedure is a proper alkaline lysis method. The lysis and recovery steps are designed to minimize DNA molecule breakage during liquid handling steps and plates manipulations. All liquid handling steps were performed with MICROLAB® STAR One; volumes refer to quantities for each well.

1. After bacterial growth, deep-well plates were centrifuged at 2397 g-force at 4°C for 27 minutes in Eppendorf Centrifuge 5810R;

2. Plates were gently inverted to discard medium and gently tapped on paper towels;

3. Plates were centrifuged at 2397 g-force at 4°C for 5 minutes to remove any residual medium;

4. Plates were vigorously inverted to remove medium and gently tapped on paper towels; 125 μL of cold Solution 1 (50 mM EDTA, 50 mM Tris-HCl, at 4 C) was added to each well;
5. Pellets were resuspended using vortex for short time period;
6. 300 μL of Solution 2 (0.2 NaOH, 1% SDS) were added to each well;
7. Plates were gently shaken in circle and incubated at room temperature for 4 minutes;
8. 225 μL of ice cold Solution 3 (3 M Potassium, 5M Acetate) were added to each well;
9. Plates were then incubated on ice for 20 minutes;
10. Plates were centrifuged at 3202 g-force at room temperature for 30 minutes in Eppendorf Centrifuge 5810;
11. 560 μL were recovered from each well and transferred into a new 96 wells plate;
12. New plates were centrifuged at 3202 g-force at room temperature for 30 minutes;
13. 460 μL were recovered from each well and transferred into a new 96 wells plate;
14. 460 μL of isopropanol were added to each well;
15. DNA was precipitated for at least 22 hours at -20°C;
16. DNA was pelletted at 2300 g-force at 4°C for 40 minutes in Thermo Scientific GR4-auto centrifuge;
17. Supernatant was discarded by vigorously inverting plates;
18. Pellets were washed with 600 μL of ethanol 80%;
19. Plates were centrifuged at 3202 g-force at room temperature for 40 minutes; supernatant discarded by vigorously inverting plates;
20. Pellets were washed with 400 μL of ethanol 80%;
21. Plates were centrifuged at 3202 g-force at room temperature for 40 minutes; supernatant discarded by vigorously inverting plates;

22. Pellets were air dried on bench;
23. Pellets were resuspended in 50 μL of pure H_2O from SIGMA®;
24. Plates were sealed with aluminum foil using ABGene ALPS-300;
25. Plates were incubated at room temperature in orbital shaking at speed 6 in Heidolph Titramax 101 overnight;
26. DNA Plates were then stored at -20°C until use.

The high volume of the wells and the soft manipulation of plates during lysis should ensure the precipitation of high molecular weight DNA removing most of the *E. coli* genomic DNA form the preparation leaving intact the BAC DNA in solution.

2.4 Pooling strategy

The focal point of the project is the pool: each pool should represent the 40-50% of the entire genome of interest or a lower fraction (see section 1.2.4). The number of BAC to be mixed in each pool was estimated considering the predicted size of the genome of *N. gaditana* (30Mb) and the average insert size of the BAC library (see section 2.2). The proper number of BAC per pool was estimated to be 96, in order to have on average 11,520,000 bases, that represent indeed the 38% of the genome.

The pooling was performed with MICROLAB® STAR One collecting 10 μL of DNA solution form each well of a 96-well plate into a single vial. This process was performed for 64 96-well plates resulting in a total of 64 pools.

2.5 DNA processing

Each pool was treated to remove as much as possible both E. coli genome and RNA contaminants. E. coli genome was removed using Plasmid-Safe™ DNase from Epicentre®. RNA was digested using RNase A form Sigma-Aldrich®.

2.5.1 Plasmid-Safe™ reaction

Plasmid-Safe™ DNase ensures the digestion of linear double stranded DNA: genomic DNA in a BAC DNA preparation is fragmented as so it is a substrate for this DNase. Genomic DNA is present in all BAC preparations. In a scenario

in which the quantities of E. coli contaminants in any single DNA preparation could be very high, in the resulting pool the quantities of any single BAC will be very small compered to that of E. coli genome. For this reason the presence of a very low quantity of E. coli DNA is mandatory for our method: otherwise we will waste too many sequences during sequencing.

To evaluate the extent of E. coli genomic DNA contamination in the preparations and its effective removal with Plasmid-Safe digestion, a comparison was performed between treated and non treated samples. 16 out of 64 total pools were not treated with Plasmid-Safe™ and the results were compared at the sequencing level, see section 3.1.1. The reaction for the 48 remaining pools is set up in: 33 mM Tris-acetate, 66 mM Potassium-acetate, 10 mM Magnesium-acetate, 0.5 mM DTT, 1 mM ATP and 12 Units of Plasmid-Safe™ DNase. Reactions incubated at 37℃ for 30 minutes in water bath and then incubated at 70℃ for 30 minutes in water bath for Plasmid-Safe™ inactivation.

2.5.2 RNase A reaction

RNA digestion was performed with RNase A form Sigma-Aldrich® to a final concentration of 12.5 μg/mL and incubated at 70℃ for 30 minutes in water bath during the inactivation step of Plasmid-Safe™.

2.6 Shotgun sequencing library preparation

As explained in the introduction, in section 1.2.5 two sequencing approaches were designed in order to achieve the production of a genome map. One method impies the shotgun sequencing of the DNA BAC pools while the other implies the sequencing of endonulcease digested sites from these DNA BAC pools. Both these method were developed and carried out to the sequencing phase. Of the two, the one chosen to develop the mapping procedure was the first one, the shotgun project. The other one was developed as trial to test the whole procedure of sequencing endonuclease sites.

This section describes the steps performed to prepare the shotgun libraries for 64 pools. These steps follow with slightly modification the 5500xl SOLiD™ protocol for a fragment library preparation.

2.6.1 DNA fragmentation

After RNA digestion each pool was fragmented with Covaris™ System with a target size between 150 and 350 bp. DNA was not purified prior to fragmentation for money, time and material saving. 1 μg of DNA was fragmented in microtubes with AFA (Adaptive Focused Acoustics™) technology, low TE buffer (10 mM Tris-HCl, 0.1 mM EDTA) from Applied Biosystem™ by Life Technologies™ was added to samples up to a final volume of 130 μL. The instrument parameters were set as follows: water bath temperature, 7°C; duty, 20%; intensity, 10; cycle/burst, 1000; 10 cycle of 60 seconds each. DNA was then purified with XP beads and eluted in 130 μL of low TE buffer, see below for DNA purification protocol.

2.6.2 DNA purification

Samples were purified adding 1.5 volumes of Agencourt AMPure XP beads by Beckman Coulter. After completely mixing samples and beads, binding was carried out for a maximum of 7 minutes at room temperature. Reactions were placed in magnetic rack for up to 5 minutes to separate beads from solution, supernatant discarded. Off the magnet, at least 100 μL (or larger as the initial volume of DNA sample) of ethanol 70% was added to wash beads, thoroughly mixed. Samples were placed in magnetic rack until clearing of solution, ethanol discarded. Wash was repeated once and beads were let to air dry on the magnetic rack. DNA was eluted from beads by adding the desired volume of low TE buffer, the solution was thoroughly mixed and incubated at room temperature, off magnet, for up to 10 minutes. Reactions were placed in magnetic rack until clearing of solution and supernatant recovered into new tubes.

Hereafter I will refer to this protocol of DNA purification as "purification with XP beads" indicating both beads quantity and volume of low TE used for elution.

2.6.3 DNA quantification

DNA was quantified using Qubit® 1.0 Fluorometer from Invitrogen™ by Life Technologies™ with the Qubit® High Sensitivity assay kit that ensures detection of a DNA range from 0.2 ng to 100 ng. Hereafter I will refer to this protocol of DNA quantification as Qubit quantification.

2.6.4 DNA ends repair reaction

DNA was end repaired to allow subsequent blunt end ligation of the adapters. This reaction was performed using the commercial "DNA end repair mix" kit from Invitrogen™ by Life Technologies™. This contains a combination of two enzymes T4 polynucleotide kinase and T4 DNA polymerase (respective concentrations are not disclosed by the company) that ensure clonability of mechanically broken DNA fragments, such as those produced by fragmentation with Covaris™ System. The reaction was performed for 30 ng of XP beads purified DNA in a final volume of 20 μL: 0.5 μL of Enzyme Mix, 50 mM Tris-HCl (pH 7.5), 10 mM MgCl2, 10 mM DTT, 1 mM ATP, 0.4 mM dATP, 0.4 mM dCTP, 0.4 mM dGTP, 0.4 mM dTTP. Reaction was incubated at 37℃ for 30 min in thermal cycler and purified with 1.5 volumes of XP beads, DNA eluted in 20 μL of low TE buffer.

2.6.5 Adapters ligation reaction

SOLiD™ specific sequencing adapters (called P1 and P2) were ligated to the entire 20 μL repaired and purified DNA. Reaction was performed in a final volume of 40 μL. Adapters were added in 80 fold excess compared to the estimated number of extremities of DNA. Reaction was set up as follows: 50 mM Tris-HCl (pH 7.6), 10 mM MgCl2, 1 mM ATP, 1 mM DTT, 5% polyethylene-glycol 8000, 35 U (Weiss) of T4 ligase, 50 μmol of P1, 50 μmol of P2. Reaction was incubated at room temperature for 15 minutes and purified with 1.5 volumes of XP beads, eluted in 25 μL of low TE.

For our application we used multiplex P2 adapters commercially available from Applied Biosystem™ that contain bar code sequences to allow multiplexing of different libraries. We used 16 of such P2 adapters.

2.6.6 DNA amplification

DNA samples were amplified with 7 cycles of PCR with Platinum® PCR master mix from Applied Biosystem™ by Life Technologies™ prior to emulsion PCR. Within this amplification cycle there is a nick repair step to ensure the covalent binding of both strands of the adapters to the DNA ends. These few cycles, that slightly increment the template amount, are also required to remove the single stranded nick resulting from adapters ligation. The number of cycles was experimentally determined (data not shown) and kept as lower as possible.

PCR reaction: 10 μL of DNA sample, 1.5 μL of P1 (50 μM) and P2 (50 μM) PCR primers and 62 μL of Platinum® mix.

PCR cycle: nick translation step 72℃ for 20 minutes; initial denaturation at 95℃ for 5 minutes; 7 cycles of 95℃ for 15 seconds, 62℃ for 15 seconds, 70℃ for 1 minute; and final extension at 70℃ for 5 minutes.

DNA was purified with 1.6 volumes of XP beads, eluted in 70 μL of low TE and quantified with Qubit HS.

2.6.7 Sequencing library preparation

Four different sequencing libraries were prepared each containing 16 different libraries (each library is a single BAC pools) in equal amount. Each one of these 4 super-libraries was processed for sequencing according to 5500 SOLiD™ library E80 protocol. The workflow can be summarized as follows: emulsion PCR reaction set up; emulsion PCR amplification; positive beads enrichment. The emulsion is prepared by properly mixing aqueous phase (PCR reagents, beads, DNA library) and oil phase to create a highly homogenous emulsion. The emulsion is then transferred in a special pouch and amplified in a modified thermal cycler. During this step the template DNA will covers the sequencing beads. At the end, amplified emulsion is broken within a peculiar machine that selects and purifies positive beads. Positive beads are those beads that succesfully go into amplification of a template DNA molecule; these beads can be identified thanks to the presence of the P2 adapter. At this point beads are almost ready to be sequenced: DNA that covers the beads need to be modify at 3'-end in order to bind the glass surface of the flowchip. After a couple of washes the beads are ready to be loaded in the flowchip and then in the sequencer.

These four libraries were loaded each one in a single lane of a 5500xl SOLiD™ flowchip.

2.7 Endonuclease digested library preparation

In this section is described the protocol performed to produce the samples to be sequenced with the endonuclease digestion method. The starting samples for this method are the same BAC pools prepared for the shotgun sequencing (see sections from 2.3 to 2.5).

The focal point of this method is the recovery of the digested sites. After the digestion of the template with the endonuclease the DNA is still bigger

than the desired size for sequencing. For instance an enzyme with a recognition site four bases long produces fragments with an average size of TODO bp. On the other hand the desired size of template DNA for NGS library preparation is around 150-300 bp. For this reason the digested DNA should be processed as well as not digested DNA. During this process however it could be difficult to preserve the digested sites.

In order to achieve this task the sites should be recovered as soon as possible to ensure their actual sequencing. This issue was resolved in this method with the aim of biotinilated adapters to bind to the digested sites. Biotinilated DNA molecules could be recovered with proper magnetic beads. Moreover, to reduce material loss the steps following the recovery were designed to be performed directly on the beads.

Of the 64 BAC pools sixteen were randomly chosen in order to be processed with this protocol.

2.7.1 Custom adapters design

Eight different custom adapters were designed. These adapters contain the sequence of the P1 primer, that is the sequencing primer, and a four bases long barcode located at the 3' of the P1. These adapters present no overhang at the extremities. At the 5' end of the P1 primer sequence a biotin was attached to allow recovery of ligated sites. Both extremities of the adapter were dephosphorilated. The adapters were obtained by hybridization of two complementary oligo purchased from Invitrogen™. The hybridization was performed in pure water in thermal cycler for thirty minutes: from 95℃ to 1℃ with -1℃ steps every 30 seconds.

The barcode sequences were carefully designed on the basis of their corresponding color-space sequences. SOLiD™ system produces sequences in color-space and thus the barcodes were designed in order to have a color-space sequence as different as possible one each other. In table 2.1 are reported base-space and color-space sequences for each barcodes.

The reliability of identification of these P1 barcoded custom adapters was tested by pairing each of them with a different multiplex P2 barcode adapter from Applied Biosystem™.

2.7.2 Enzymatic digestion

1.5 μg of DNA treated with both Plasmid-Safe™ and RNase A were purified with XP beads and digested with Sau3AI endonuclease from New England Biolabs®.

Base Space	Color Space
P1(T)-CGGT	T-2301
P1(T)-AGTT	T-3210
P1(T)-TGAT	T-0123
P1(T)-GGCT	T-1032
P1(T)-GAAT	T-1203
P1(T)-TACT	T-0312
P1(T)-AAGT	T-3021
P1(T)-CATT	T-2130

Table 2.1: Barcodes seqeuences. In the left column P1 indicate the sequence and the position of the P1 primer its sequence is omitted because patented. The T in parenthesis is the last base of the P1 primer and is fondamental for conversion from base space to color space. In the rigth are reported the color-space sequences of each barcode, the T is the same as in left column.

Reaction conditions: 6 units of Sau3AI, 10 mM Bis-Tris-Propane-HCl, 10 mM MgCl2, 1 mM DTT (NEB Buffer 1), BSA 1 ng/μL, incubated at 37℃ in water bath for 2 hours. Enzyme was inactivated at 65℃ for 20 minutes. DNA was purified with 1.5 volumes of XP beads and eluted in 60 μL of low TE buffer.

The 5' protruding extremities of digested sites could affect sequencing because of the peculiar chemistry of SOLiD™ System. The 5' protruding extremities were digested with Mung Bean nuclease from Takara Bio creating blunt ends in double-stranded DNA fragments. Reaction condition: 45 units of enzyme, 30 mM sodium acetate (pH 5.0), 100 mM NaCl, 1 mM zinc acetate, 5% glycerol, incubated at 37℃ for 20 minutes. DNA was purified with 1.5 volumes of XP beads and eluted in 30 μL of low TE buffer.

2.7.3 Ligation of P1 custom barcode adapters

The P1 custom barcoded adapters are conjugated at one 5' end with a biotin. This biotin is necessary at the end of the protocol to recover digested over the mechanically fragmented sites. These latter sites will be more abundant in respect to the digested sites. P1 adapters were added in 80 fold excess with respect to the estimated number of 5' ends. Reaction condition: 50 mM Tris-HCl (pH 7.6), 10 mM MgCl2, 1 mM ATP, 1 mM DTT, 5% polyethylene-glycol 8000, 25 units (Weiss) of T4 DNA ligase and 1.52 μL of P1 50 μM, in a final volume of 50 μL. Incubated at room temperature for 20 minutes. DNA purified with 1.6 volumes of XP beads and eluted in 25 μL of low TE buffer.

2.7.4 Mechanical Fragmentation

Samples were then fragmented with Covaris™ System in order to obtain fragments ranging form 150 to 350 bp. Samples fragmented in microtubes with AFA (Adaptive Focused Acoustics™) technology, low TE buffer (10 mM Tris-HCl, 0.1 mM EDTA) from Applied Biosystem™ was added to samples up to a final volume of 130 μL. DNA purified with 1.6 volumes of XP beads and eluted in 20 μL of low TE buffer.

2.7.5 DNA ends repair reaction

DNA was end repaired to allow ligation of P2 adapters. The reaction was performed using the commercial "DNA end repair mix" kit from Invitrogen™ by Life Technologies™. Reaction was performed as reported above (see paragraph 2.6.4) and DNA was purified with 1.6 volumes of XP beads and eluted in 15 μL of low TE buffer.

2.7.6 Ligation of multiplex P2 adapters

Ligation of P2 multiplex adapters was performed with an 80 fold excess of adapter compared to the entire set of extremities, both P1 ends and fragmented ends. Reaction condition was the same as those of P1 ligation (see paragraph 2.7.3) with a final volume of 30 μL and incubated at room temperature for 20 minutes. DNA was purified with 1.6 volumes of XP beads and eluted in 20 μL of low TE.

2.7.7 Biotin capturing

After the ligation of the P2 adapters, it is likely that the most abundant fragments are those carrying the P2 at both ends (P2-P2 fragments). But the fragments needed for sequencing are those carrying the P1 at one end and the P2 at the opposite. These are indeed the digested sites. To overcame this problem the strategy is to take advantage of the biotin on the P1 custom adapters to selectively capture the correct costruct (P1-P2) by using the streptavidin coated magnetic beads, Dynabeads® MyOne™ Streptoavidin C1 from Invitrogen™ by Life Technologies™. Binding and washing (B&W) buffer composition 2x: 10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 2 M NaCl.

10 μL of beads were prepared for biotin binding as follows:

- Magnetize until clearing of solution, supernatant removed;

- Resuspend beads in 50 μL of 2x B&W buffer, magnetize until clearing and discard;
- Repeat wash;
- Resuspend beads in 50 μL of 10 ng/μL BSA, magnetize until clearing and discard;
- Resuspend beads in 50 μL of 2x B&W buffer, change tube and magnetize until clearing and discard;
- Resuspend beads in 20 μL of 2x B&W buffer.

Beads are now ready to bind DNA:

- Add 20 μL of DNA sample to binding beads;
- Incubate at room temperatures on rotor for 30 minutes;
- Magnetize until clearing and remove supernatant;
- Wash with 50 μL of 1x B&W buffer;
- Wash with 50 μL of E1 buffer (10 mM Tris-HCl, pH 8.5) from Invitrogen™ ;
- Resuspend beads in 15 μL of E1 buffer.

Beads can now be used as template for subsequent application.

2.7.8 DNA amplification

To recover DNA, beads were amplified with Platinum® PCR master mix from Applied Biosystem™ by Life Technologies™. Also in this case during amplification there is a step of nick repair to ligate both strand of the adapters to the DNA ends.

PCR reaction: 15 μL of beads, 1.2 μL of P1 (50 μM) and P2 (50 μM) PCR primers and 45.6 μL of Platinum® mix.

PCR cycle: nick translation step 72℃ for 20 minutes; denaturation at 95℃ for 5 minutes; 9 cycles of 95℃ for 15 seconds, 62℃ for 15 seconds, 70℃ for 1 minute; and a final extension at 70℃ for 5 minutes.

PCRs were then magnetized to remove MyOne™ C1 beads prior to DNA purification with 1.6 volumes of XP beads, DNA was eluted in 40 μL of low TE and quantified with Qubit® HS.

2.7.9 Library preparation

Two different libraries were prepared, each containing 8 digested BAC pools in equal amount. Each one of these 2 libraries or super-pools was processed for sequencing according to 5500 SOLiD™ library E80 protocol, see paragraph 2.6.7.

2.8 Reads alignment

The sequences of the 64 pools shotgun sequenced were aligned with PASS [41], a program to align short reads on a reference. A unique file containing all the reads from the sixty-four pools was created, reads from each single pool were marked. All analysis were performed considering only reads uniquely mapped with no gap opening allowed.

One alignment was performed against a database containing the draft assembly of the genome produced on 454™ reads, the reference genome of *E. coli* and the sequence of the BAC vector. This alignment was performed in order to filter contaminant reads coming from *E. coli* and BAC vector and to produce preliminary statistics.

A second alignment was performed to construct the distance scoring matrix (see section 3.2.3): the database in this case contains the virtually fragmented 454™ contigs, the *E. coli* reference sequence and the BAC vector. To create these fragments two different average sizes were used: 5000 and 2500. The virtual fragmentation was performed in order to obtain consecutive fragments from all the contigs. In this process no information about connections among contigs were considered nor preserved. The actual length of the resulting fragments depends also on the length of the fragmented contig. Despite the average length the fragments from a given contig have minimal differences in length. These virtual smaller contigs are colled *smaltigs*. This alignment was used for subsequent mapping programs.

A third alignment was performed against contigs genereted starting from mate-pair reads (see section 2.9). This alignment was used for subsequent mapping programs.

2.9 Short reads assembly

A short-reads assembly was perform on reads obtained from one of the mate pair libraries produced for the genome sequencing project of *N. gaditana*. The

insert size was from 1.5 to 2 kb. The library was sequenced with SOLiD™ 3plus version. The produced sequences are 50 base long. A first assembly was performed without considering the information of insert size. This means that the two mate reads of a single DNA fragment were considered and used independently, as it happend for reads from fragment sequencing. Reads were assembled with Velvet [42], a short reads assembly program based on de Brujin graphs [23]. The last 5 bases at the 3' end of each sequence were removed because of low quality score of the final bases. The k-mer size used to build the graph was 21 bp. These paramenters of trimming and k-mer size were decided on the basis of a trial assembly on a sub-set of reads, see section 5.2. The assembly resulted in 55556 contigs with a N50 of 727 bases.

2.10 Custom programs and scripts

Several scripts were produced to manage and analyze data. These scripts were self written in Python language (www.python.org).

The DOT language was used to construct user friendly visualization for map-scaffolds results see section **??** and 3.3.

The main algorithms to count reads aligning in contigs, produce matrixes and analyze distances between profiles were developed in intimate collaboration with professor Giorgio Valle. These programs because of computing power and the complexity of the algorithm they were written in C language by prof. Giorgio Valle.

Chapter 3

Results and Discussion

3.1 Preliminary analysis

In this Thesis are presented the results for both the strategies of mapping genomes explained in section 1.2.5. The two projects rely on the sequencing of several pools of BAC clones. One project is based on the shotgun sequencing of the DNA of each pool. The second project is based on the sequencing of the endonuclease restriction sites obtained from the DNA of each pool. The number of pools processed with the two methods is different: for the former 64 DNA BAC pools were sequenced, while for the latter 16 digested DNA BAC pools were sequenced.

A total of six sequencing reactions were performed each one within a single lane of SOLiD™ system 5500 xl flow-chip. Each run ensured the independent sequencing of several pools by means of commercial barcodes sequences (see section 2.6.5). Each run was thus performed on a single multiplex library. Runs 1 and 2 were performed on pools processed with the "endonuclease protocol" while runs from 3 to 6 were performed on pools shotgun sequenced. Table 3.1 summarize the reads produced by each run.

The lower numbers of pools in sequencing reaction 1 and 2 is due to the number of custom barcodes. In fact, because the barcodes are four bases long only eight different barcodes were designed among all the possible 4 bases

Run	Pools	Sequences
1	8	112,198,801
2	8	140,865,340
3	16	99,552,241
4	16	113,666,328
5	16	117,264,515
6	16	108,660,500

Table 3.1: Runs summary. Column "Pool" indicates the number of single BAC pools sequenced within each run by means of multiplex sequencing. Column "Sequences" indicates the total number of sequences produced within each run.

sequences, see section 2.7.1. Runs 3 to 6 were performed on a higher number of pools by means of the commercial barcoding kit. This offers the possibility to sequence up to 96 different samples in a single library. Sixteen pools were sequenced within each run to ensure a sufficient throughput of sequencing for each pool. The reads were 75 bases long, accounting for a total of 51.9 Gbp.

The "shotgun sequenced" pools were actually used for the development of the genome mapping method presented in this Thesis. For this purpose an high number of pools was sequenced. The "endonuclease digested sequenced" pools were preliminary analyzed in order to verify the goodness of the custom barcodes.

3.1.1 Alignment results for shotgun sequencing project

For the genome mapping project sixty-four DNA BAC pools were sequenced in four different sequencing reactions. A total of 439.1 millions of reads were produced. The reads were aligned against a database containing *N. gaditana* draft assembly, *E. coli* reference genome and pCC1-BAC vector sequence. During the alignment, the program filters low quality reads: on the total reads more than 47 millions were removed. Of the remaining reads the 74.12% presented an unique alignment. The results of this alignment are summarized in table 3.2, detailed results for all pools are reported in table 5.1.

The remarkably high number of reads aligning on the BAC vector sequence was expected. A single pool is composed by 96 BAC clones. Each BAC sequence is formed by the vector and the insert. The insert accounts for the larger part of the BAC (on average 120 kbp) while the vector accounts fo a minor part of this sequence (around 8 kbp). In a single pool the insert is different for each BAC but the vector is always the same. This implies that the sequence coverage

Run	Produced seq.	Aligned seq.	*E. coli* seq.	BAC seq.	*N. gaditana* seq.
3	6,222,015	3,969,535 (64.54)	39,933 (1.01)	439,179 (11.35)	3,490,423 (87.64)
	± 969,373	± 802,902 (12.82)	± 20,887 (0.51)	± 90,263 (2.47)	± 748,409 (2.71)
4	7,104,146	4,758,573 (67.17)	27,038 (0.55)	497,330 (10.55)	4,234,205 (88.89)
	± 2,330,236	± 1,623,181 (5.98)	± 18,461 (0.26)	± 155,992 (0.90)	± 1,458,606 (0.89)
5	7,329,032	4,953,694 (67.57)	42,466 (0.84)	414,033 (8.42)	4,497,195 (90.74)
	± 1,187,326	± 833,740 (3.73)	± 38,968 (0.70)	± 57,944 (0.61)	± 762,630 (0.81)
6	6,791,281	4,470,426 (65.82)	41,675 (0.93)	387,679 (8.69)	4,041,071 (90.39)
	± 394,003	± 299,819 (1.99)	± 31,658 (0.68)	± 44,138 (0.96)	± 285,219 (1.36)

Table 3.2: In this table are summarized the results of sequencing reactions and reads alignments. Seq. = sequences. The sequencing reactions are indicated on the first column. Each run was performed on sixteen pools thus each raw indicates the average values for a single pool in each reaction. Values marked with ± indicate the standard deviation from the relative mean. Columns *E. coli* seq., BAC seq. and *N. gaditana* seq. indicate the average values of aligned reads on the corresponding reference. Values in parenthesis indicate the percentages. In column "Aligned seq." the percentages refer to the total number of produced sequences, while in columns *E. coli* seq., BAC seq. and *N. gaditana* seq. the percentages refer to the corresponding number of aligned sequences.

of the vector will be very high. A possible strategy to eliminate the vector sequence from a BAC DNA preparation could be the endonuclease digestion with a rare cutter enzyme, i.e. NotI whose recognition sites are placed at the two ends of the poly-cloning site. In this way the insert will be "released" from the vector. With a size selection in agarose gel it is then possible to select the insert. However, this strategy was not viable for this project because of the high number of samples to be processed. Anyway, within a single SOLiD™ system lane it is possible to produce a very high number of reads. Given this very high throughput, the reads that will be lost in sequencing the vector will not compromise the production of a high coverage for the desired reads covering the insert sequences.

As it happens for the vector, also the *E. coli* genome is present in all BAC DNA preparation and so its relative amount increases during BAC pooling. In contrast with the vector sequence, the *E. coli* genome is pretty much bigger. Thus it is mandatory to eliminate as much as possible this genome from the DNA preparations. In order to achieve this task the BAC DNA was extracted taking into account peculiar strategies in order to reduce *E. coli* genome abundance in the purified DNA. Moreover a treatment with Plasmid-Safe™ exonuclease was performed on purified DNA to further remove *E. coli* genome, see 2.5.1.

The success of the DNA extraction procedure by itself in removing the *E. coli* genome was evaluated with an experiment performed on libraries of run 3 and 4. These two libraries were sequenced prior to the others. The sixteen pools of run 3 were not treated with Plasmid-Safe™ enzyme while the sixteen pools of library 4 were treated with this enzyme. These pools in fact were processed for sequencing with the same procedure extraction procedure except that for the Plasmid-Safe™ reaction.

The percentages of reads aligning on *E. coli* genome in the sixteen treated pools were compared with those of non treated pools. The results are summarized in the box-plot in figure 3.1. The treatment with Plasmid-Safe™ enzyme significantly reduces the amount of genomic DNA. However, the low fraction of genomic DNA in non treated pools indicates that the BAC DNA preparation method developed in this work is useful by itself to reduce genomic contamination. The remaining 32 pools were treated with Plasmid-Safe™ to ensure the highest removal of *E. coli* genome.

The low numbers of reads aligning on the *E. coli* reference genome in all the pools (the total percentage of reads aligning on E. coli genome is 0.83%) confirms the overall success of the DNA preparation strategy developed.

3.1.2 Analysis of custom P1 barcodes

The endonuclease restriction mapping project were performed on sixteen pools that were sequenced in two SOLiD™ system sequencing lanes. As reported in section 2.7.1 eight custom barcodes four bases long were designed; of all the possible combinations were chosen those that imply the highest number of sequencing errors to became one of the other barcodes.

In order to test the reliability of these custom barcodes they were coupled with the commercial kit from Applied Biosystem™ to produce multiplex library sequencing. This system assigns sequences to each library reading the barcode sequences prior to start the sequencing run. At the end, it produces different

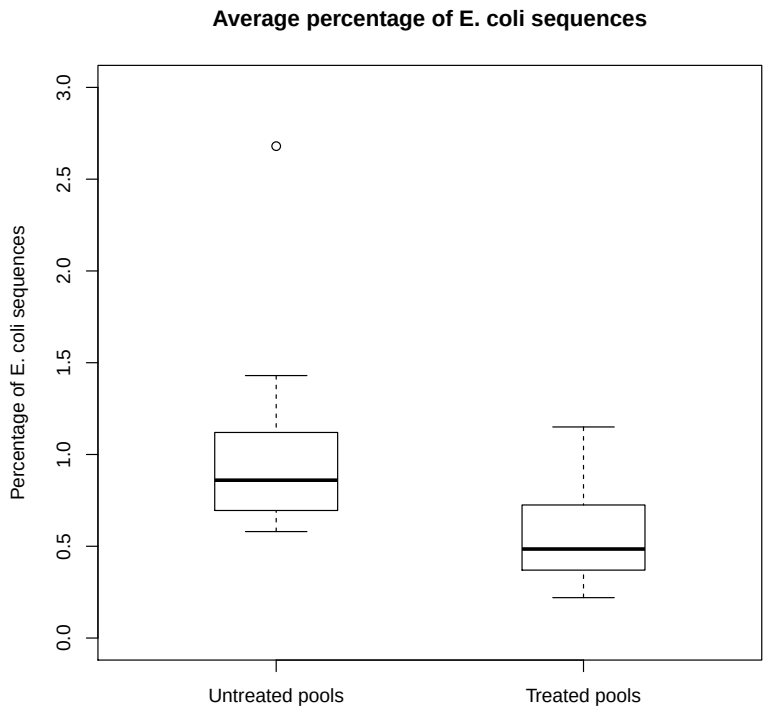


Figure 3.1: These box plots indicate the distribution of the percentage of *E. coli* sequences in pools of run 3, containing untreated pools, and run 4, containing treated pools. The percentages refer to the total number of aligning reads.

files containing the reads belonging to each library. The initial assumption in this test is that the commercial system is assumed to be free or errors. This means that it will not assign a wrong barcode to a given sequence and so a sequence to the wrong library. It has to be pointed out that to our knowledge the sequencer does not provide any information about the rejected sequences, those sequences that do not match with any of the assigned barcodes.

At the end of the sequencing reactions the reads are divided according to their proper commercial barcode. Is that possible to see if the custom barcodes could work as well as the commercial ones? Given that each custom barcode was coupled with a commercial one there should be no discrepancy between the two. To evaluate if the custom barcodes are useful to discriminate between different libraries a simple strategy is to look at the first four bases within each sequences of the sixteen different pools. If there are no errors in the custom barcodes, within each pool there will be only the elected barcode at the beginning of each read.

By looking at the reads produced from all the 16 pools, the majority of these

sequences presents the proper barcode in the starting position but unfortunately a considerably high amount have a different sequence. These sequences present one to four errors in respect to the proper barcode, see table 3.3. These differences could be due to errors occurred during both sequencing or synthesis of the barcodes. Unfortunately, many of these sequences with errors actually have the sequence of another barcodes.

Total sequences	**251,270,552**
No errors	224,018,078
One error	17,506,856
Two errors	5,473,007
Three errors	2,432,757
Four errors	1,839,854
Total sequences with errors	**27,252,474**
Cross called	2,269,682

Table 3.3: Number of sequences with errors within barcode sequence. Here is summarized the number of sequences for the sixteen digested pools that have zero or one to four errors in the first four bases. The number of total sequences refers to the sum of useful sequences produced within each pool (useful are those sequences that do not present one or more gap, the lack of a base, within the first four positions). The number of errors refers to the number of wrong bases in a given sequence in respect to its proper barcodes. "Total sequences" with errors refers to sequences that present at least one error in the barcode. "Cross called" refers to those wrong sequences that specify for another barcode, independently from the number of errors.

DNA sequencers give a quality score for each base they call within a read. These scorese give informations about the fidelity with which the sequencer assign a base (or a color in case of SOLiD) to a given position. The higher the score the lowest is the probability that a base call is an error. In light of this, sequencing errors should have low quality scores whereas synthesis errors should have high quality scores. This is because errors occurred prior to sequencing, for example during oligo synthesis or during PCR template amplification, could not be identified by the sequencer that indeed reads "what is written". The four bases of the wrong barcodes could then be analyzed by looking at their quality scores, in order to see if they are sequencing errors o synthesis errors. Figure 3.2 shows the distribution of qualities scores for the four bases of the barcodes. The sequences were divided in five different classes according to their errors in barcode sequence: sequences with no errors; sequences with one error

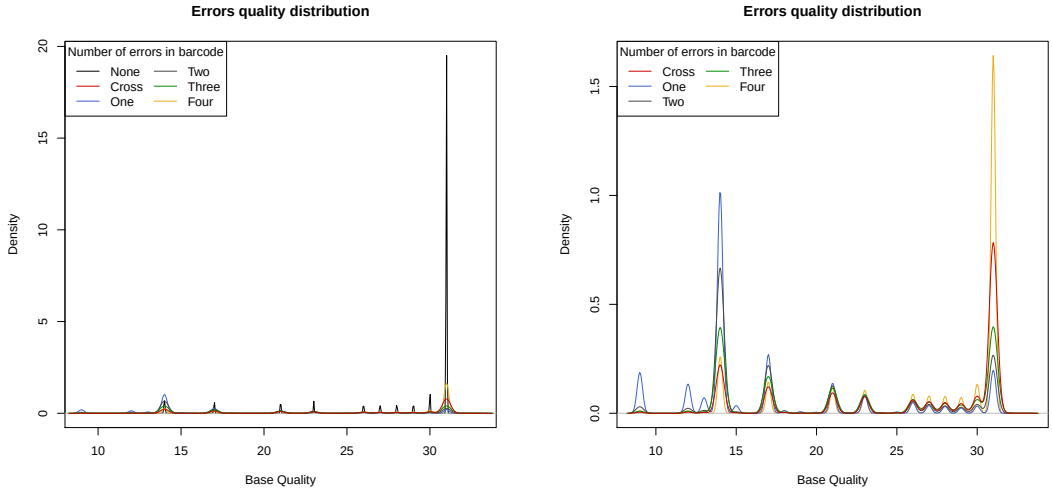


Figure 3.2: Base quality distributions for the four bases that compose the barcodes in the custom adapters. Base quality values range from 0 to 31 where 31 expresses a very high quality of a given base. In the graphs the curves indicate sequences that present none, one, two, three and four errors according to legend. In the graph on the left are shown also the distribution for those sequences that does not present any error. In the graph on the right the no-errors sequences are removed in order to make more clear visible the distributions of wrong sequences.

(class 1); sequences with two errors (class 2); sequences with three errors (class 3); and sequences with four errors (class 4).

As it can be seen in the graphs the vast majority of the bases that present no errors has the maximum quality score, 31. The bases of the class 1 have low quality scores, indicating that the presence of one error in the barcode is mainly due to sequencing errors. The same is true also for class 2 although the density of the curve is higher in high values indicating that many error occurred during barcode synthesis.

The cross called sequences those that actually can be confused with a right barcode, result only from class 3 and 4. As it can be seen in the graph the quality scores for these two class of errors are mainly distributed on high values. This is especially true for class 4 rather than for class 3 where in fact the quality values are almost equally distributed between 14 and 31.

These results indicate that a barcode four bases long could not be enough to ensure a proper discrimination of sequences in a multiplex library. It has to be pointed out that the commercial barcodes have a much more longer sequence that could permits a proper call of the barcodes. In order to eliminate the fraction of cross-called sequences the number of bases within these custom barcodes should be increased. With a longer barcode it could be also possible

to correctly assign also sequences that present one error within the barcode allowing the recovery of much more sequences.

3.2 Genome mapping project development

The genome mapping project was developed analyzing the reads produced with the shotgun sequencing of the 64 BAC pools, see section 3.1.1. Thus all the following analysis refer to data obtained from sequencing runs 3, 4, 5 and 6.

3.2.1 Genome fraction in pools

Considering reads length and the number of reads aligning on the *N. gaditana* draft assembly see table 3.2, each pool accounts for an average of 304 million bases sequenced. Each pool represents a fraction of the genome, so this total number of bases should represent only a portion of the draft assembly, in theory the 38% (see section 2.4). However, the real amount of target DNA within each pool could be different from this prediction. The genome portion present in each pool can be estimated by the coverage per base on the draft assembly. With self written Python scripts the amount of bases covered at least once was calculated. This threshold was decided because one reads is sufficient to give information about the presence of a sequence.

On average the 25% of the genome is covered at least once in each pool. However some pools covers even lower fractions of genome. For instance pool 35 covers the 16% of the assembly and pools 12 and 9 each covers only the 17%. On the other hand, some pools account for higher fractions such as the 32% of pool 12 or even the 30% (pools 44 and 48).

These percentages are lower than the 38% a priori decided during pooling process. A first explanation to this is that the draft assembly accounts for 27.9 million bases, lower than the predicted size of 30 Mb. In fact, the number of BAC per pools was calculated according to the predicted size of the genome not to the total bases assembled. This could indicates that some genomic regions are present in the BAC library but are absent in the draft assembly. Even assuming a reliable genome size prediction, this could be true only for a small number of these "lost fractions". In fact, the whole-genome shotgun strategy whit next generation sequencing adopted for the production of the draft assembly does not suffer of any cloning biases. On the contrary, some genomic regions could be toxic for *E. coli* and get lost during cloning.

There are two other possible explanations for this lower representation of

the genome in the BAC pools. (i) Some systematic errors could occur during pooling process. Some BAC clones could be lost during automated DNA extraction or they could be present in very low quantity in BAC DNA pools. (ii) Another possible explanation is that the average insert size of the BAC library could be lower than expected. The latter should account for the largest part of these "lost fractions". This hypothesis is suggested by the fraction of reads aligning on the draft genome in respect to the produced reads. For instance the worst pool, the 35, has a total number of reads of 5,710,940 and 2,920,695 of these align on the assembly accounting for a 16% of the genome. On the other hand pools such as 6, 22 and 31 have a lower number of reads aligned but these represent the 25% of the genome. Moreover the systematic errors should be distributed in all the preparations whereas there are only a number of pools that has a very low fraction.

Another possibility is that the genome could be larger than predicted but published draft genome of *Nannochloropsis* species have comparable genome sizes [43, 44].

3.2.2 Creation of profiles of presence and absence

Despite the problems during pooling, each pool actually contains a fraction of the target genome. In light of this, is that possible to obtain, by looking at the pools, an information about presence and absence of a desired genetic markers? Is it possible to produce profiles of presence and absence of these genetic markers? And moreover, what kind of markers could be used in this strategy?

In a shotgun sequencing approach any produced sequences could be used as genetic marker. The sole restriction is that this sequence should be unique on the genome. A repeated sequence will give little information about genomic position given that it accounts for multiple regions that may not be physically connected. Thus, genetic markers could be a set of unique sequences of a given length, for example 21 bases. These sequences could be searched in the reads produced from each pool to see how many times and in which pool they are present. At the end of the process same sequences will be discarded because they are present in all the pools or, on the other hand, they are never present. But some reads will present a profile in which it is possible to see when that sequence is present. All the sequences with similar profiles are hypothetically close on the genome. Although this strategy is viable, it was not used during this work due to the informatics resources and knowledge required to perform

such a kind of analysis.

Another strategy is to use contigs from an independent assembly, if present, and to profile these contig along the whole set of pools. This strategy is simpler than the previous one because requires only the alignment of the reads of each pool on the assembly. The results can give information about presence and absence of each contig. This was the strategy actually used to develop the genome mapping method.

The draft assembly of the *N. gaditana* genome produced in our laboratory contains very large contigs that indeed account for large genomic regions. These large genomic regions could be covered by many different BACs that could be randomly sorted in different pools. This can compromise the presence and absence analysis because larger contigs will have spurious profiles. Trying to eliminate this problem the pool-reads alignments on contigs were counted in windows of given size. But this approach gave some problems in managing those reads that align at the edge of two windows. To reduce this problem the reads were aligned against fragments of the assembled contigs, called *smaltigs*. In this way only the best unique match is considered as useful information for the presence of that *smaltigs*.

Is this alignment information a viable method to produce profiles of presence and absence of the *smaltigs*?

To answer this question let's take a look at figure 3.3. In this matrix each raw corresponds to a single *smaltigs* whereas the columns correspond to pools from 1 to 39. The numbers within the matrix indicate the number of reads for a given pool aligning on a given *smaltigs*. The *smaltigs* in the picture are 5000 bases long and are created from contig00001. They are ordered from the beginning to the end of the contig, homogeneously covering its entire sequence.

As it can be seen in the picture, *smaltigs* that lie close on the genome (i.e. that are next to each other on the contig), share similar number of reads counts in the same pools. This is due to the presence of single BAC clones that cover a portion of the genome. Once sequenced these clones give information about the physical connection of two close sequences.

With these evidences four initial assumptions can be confirmed:

Figure 3.3: Figure on next page. Matrix of reads counts for *smaltigs* belonging to contig00001. In raws there are the *smaltigs* from 1 to 82 of the contig00001. Contig00001 is more than 500,000 bases long. In columns there are the 39 BAC pools, only 39 pools are displayed because of space. The numbers represent the total reads that for the pool in column aligning on the *smaltigs* in raw.

Snail_ID	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25	P26	P27	P28	P29	P30	P31	P32	P33	P34	P35	P36	P37	P38	P39
00001.001	0	2	0	1	2420	6	3696	3	331	3494	6	4	5	6	180	0	3	4	694	1	128	8	244	54	2836	3	543	461	14	4247	21	22	23	7851	4	0	11	629
00001.002	7	2	5	9	2919	3	4131	9	405	3879	2	4	0	4	102	1	1	1	661	3	98	2	223	16	2461	0	577	400	10	3577	10	19	19	8968	2	11	5	790	12
00001.003	2	6	12	3	4374	0	5291	7	534	2690	5	13	0	7	6	0	7	6	1263	5	197	5	438	94	4306	11	936	725	30	6880	10	44	30	4560	7	4	3	499	14
00001.004	7	6	13	7	3305	5	4644	5	419	2305	5	2	0	4	6	0	3	6	908	4	210	2	358	66	3418	0	779	66	31	5191	8	26	30	6228	3	3	4	7	10
00001.005	17	2	3	5	3285	26	5376	3	559	2725	9	9	0	6	6	0	5	4	889	2	126	2	390	88	3947	0	918	0	16	2599	8	14	30	10530	2	3	11	14	19
00001.006	2	4	5	18	3597	6	6465	5	536	2595	9	4	5	7	2	7	7	2	998	2	315	2	286	76	4294	3	841	4	0	182	2	2	51	9452	9	0	4	1	10
00001.007	2	13	7	16	4565	6	9524	1	599	3299	9	6	3	2	2	11	4	6	1366	2	225	5	3	184	12245	5	1306	7	4	222	5	2	19	6773	4	6	8	4	16
00001.008	2	8	11	5	3798	6	9598	5	655	3218	64	11	2	6	15	3	8	6	1470	3	274	8	3	180	16218	15	3095	18	8	273	32	5	6058	23	6375	9	6	7	8	16
00001.009	5	8	14	10	4241	9	10090	3	674	3513	32	4	2	1	10	1	1	1	1579	2	304	19	3	52	18157	0	3417	16	16	307	8	2	8898	34	6776	7	9	14	11	8
00001.010	2	15	17	12	3543	3	6237	1	435	2073	5	8	5	4	5	4647	5	10	570	2	93	0	3	5	5869	3	1290	2	0	121	2	4689	15	5839	2	4	2	4	8
00001.011	2	0	7	9	3715	5	2796	5	405	2633	13	4	3	22	4	5874	5	8	1673	4	133	13	5	82	10836	1	2432	2	6	146	2	5	5063	16	6261	7	0	4	1	8
00001.012	0	4	7	9	2889	3	3401	3	534	2029	4	4	2	6	8	7520	8	4	1193	4	89	35	1	114	9151	1	2436	2	4	175	2	5	5418	10	6577	9	1	13	4	14
00001.013	0	4	1	5	3147	2	3102	4	547	2938	9	6	3	6	6	6723	5	6	1427	2	30	5	0	120	9817	0	2786	2	25	248	0	0	7353	17	6573	9	3	11	3	4
00001.014	0	2	3	1	965	6	2961	7	611	5	9523	4	6076	34	18	17107	5	4	1340	4	6	10	3	3	8085	0	2605	7	4	217	16	6	6492	17	6401	4	1	2	12	6
00001.015	10	8	4	3	1731	1	3041	1	496	8	9373	4	5690	59	16	17107	5	4	1204	3	12	5	3	24	8128	3	2425	11	0	175	8	1	6196	38	6604	4	2	1	4	5
00001.016	2	3	9	11	4546	6	3623	3	132	16	8728	18	3870	4	16	16117	6	0	1112	2	4	1	1	1	4814	1	2255	11	8	21	13	2	5564	19	6240	2	3	2	1	5
00001.017	5	6	9	9	4589	6	3493	5	195	11	11009	9	6796	13	22	19997	13	6	1654	6	8	16	13	38	11129	17	13434	9	8	40	13	5	9042	10	7123	9	3	2	0	8
00001.018	5	6	3	3	3476	6	4131	9	674	3513	32	6	3870	4	22	20451	17	15	1533	9	6	12	30	40	10917	11	13440	14	4	40	13	2	8578	17	7376	1	2	4	0	5
00001.019	5	6	3	9	4241	3	6237	1	435	2073	5	8	5	4	5	4647	5	10	570	2	93	0	3	5	5869	3	1290	2	0	121	2	4689	15	5839	2	4	2	4	8
00001.020	7	8	3	12	4421	11	4733	16	3699	2826	7311	167	4	6	2	8750	144	10	1427	4	6	2	3	24	8128	3	2425	11	0	175	8	1	6196	38	6604	4	2	1	4	5
00001.021	12	8	3	33	4509	23	3493	5	195	11	11009	9	6796	13	22	19997	13	6	1654	6	8	16	13	38	11129	17	13434	9	8	40	13	5	9042	10	7123	9	3	2	0	8
00001.022	7	0	12	4	4421	13	3476	6600	198	16	11130	32	6022	5015	25	20451	17	15	1533	9	4	16	12	30	40	10917	11	13440	14	4	40	13	2	8578	17	7376	1	2	4	0	5
00001.023	5	6	7	9	4733	16	3699	9044	244	8	10996	13	6470	4907	26	19574	19	2	1214	10	8	3	3	40	9614	9	14167	4	8	31	13	4086	4	8358	4	0	1	0	1
00001.024	0	17	5	5	5416	26	3282	2826	7311	167	8749	27	4865	3838	22	15681	7	4	625	3	24	1320	32	8904	23	10662	9	7106	2	97	48	8389	6728	8317	0	3	1	0	1
00001.025	5	4	1	16	5109	26	3282	2826	7311	167	8749	27	4865	3838	22	15681	7	4	625	3	24	1320	32	8904	23	10662	9	7106	2	97	48	8389	6728	8317	0	3	1	0	1
00001.026	25	5	14	18	5403	9	3706	13	174	8	9032	19	11	3802	20	15566	2	4	974	9	38	1	3	14	9113	0	5375	30	7	1233	9	5	5421	14	8245	4	0	4	0	3
00001.027	2	11	5	16	4565	6	9524	1	599	3299	9	6	3	2	2	11	4	6	1366	2	225	5	3	184	12245	5	1306	7	4	222	5	2	19	6773	4	6	8	4	16
00001.028	0	6	2	10	4589	6	3493	5	195	11	11009	9	6796	13	22	19997	13	6	1654	6	8	16	13	38	11129	17	13434	9	8	40	13	5	9042	10	7123	9	3	2	0	8
00001.029	5	12	6	20	4375	16	6150	8024	25	8	9804	32	3605	4	15	16168	11	4	1410	22	16	11	3	24	11665	10	9768	37	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.030	2	6	2	5	4375	16	6150	8024	25	8	9804	32	3605	4	15	17560	16	4	1410	22	16	11	3	24	11665	10	9768	37	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.031	5	2	20	5	5414	9	6460	9582	22	18	11130	32	6900	32	7	11210	16	16	4679	29	10	32	18	71	10718	17	7242	25	17	17831	14	10	7336	5	3915	5	3	7	19	3
00001.032	2	14	11	18	4196	19	4752	6076	25	1	6900	32	7	3126	8	5662	1	15	633	2	2	8	33	15	8288	18	48	24	17	16240	7	6	6918	9	4898	1395	6	4	2	12	4
00001.033	3	14	18	26	6163	23	7070	10469	50	13	12412	23	1	6204	10	10718	9	71	2241	11	68	33	13667	92	20977	7	27102	16	317	31	5970	18	32	1702	31	8590	10	1	10	
00001.034	3184	13	3	14	4096	9	5733	7002	13	11	8404	0	7	3934	6	7259	8	17	1106	6	43	13	5375	30	10939	5	14671	11	179	7	6078	8	380	1484	21	7182	4	3	8	
00001.035	2838	0	5	11	4651	16	5119	6413	18	13	7481	13	3	3383	31	6392	6	26	491	2	43	13	5375	30	9113	5	11832	2	133	7	6840	10	402	1370	14	6771	4	3	3	
00001.036	1662	6	3	9	2699	6	3264	3664	4	13	4555	18	3	2917	8	3776	1	4	3	4	12	41	11	27	2319	0	3972	0	29	7	6404	0	231	1538	24	4994	1	8	3	
00001.037	3139	10	7	20	4704	13	5548	7002	36	6	8150	18	3	6307	12	6721	12	4	1673	2	41	11	5274	12	8374	3	11169	4	137	20	5595	8	485	1538	24	7469	1	8	5	
00001.038	2870	4	892	20	3966	3	3307	7095	11	1	9498	9	3	6531	11	6248	14	13	1193	3	30	3	3158	44	9089	7	6526	2	120	7	5949	14	339	1432	16	7620	7	8	54	
00001.039	3733	4	8511	18	4949	16	5204	6534	2	4	4	70	13	4040	12	4726	14	26	1427	6	21	2	3	15	3984	5	13267	7	146	14	6179	14	402	1400	12	2940	5	16	28	
00001.040	2363	2	5574	13	3468	13	3373	4177	6	3	4	4	1	1675	6	1	9	6	7	1	34	21	11782	2645	5675	2	14671	11	179	7	6078	8	380	1484	21	7182	4	3	8	
00001.041	3279	10	6979	7	4471	16	4948	6078	4	5	2	13	5	2161	6	5	16	8	1340	4	58	14	21712	4765	20	7269	3	10599	9	5544	3	1	9491	3	3	11	10	2	0	4
00001.042	3926	2	9026	16	4986	23	6104	6907	0	16	4	18	5	2587	16	7	6	26	25	9	51	68	32432	6526	42	10599	9	5681	7	10137	7	126	9	140	21	5125	8	428	1358	24	7469	1	8	5	
00001.043	2777	14	6927	14	3839	29	4213	5351	11	3	0	27	3	1940	20	6	0	30	9	3	43	13	17613	3450	14	5544	3	6733	9	13900	9	2	8898	34	6776	7	9	14	11	8
00001.044	3708	4	8131	5	3403	19	5314	7026	0	11	4	0	9	2592	6	0	11869	34	17	13	51	68	26192	5064	12	5681	7	10137	7	126	9	140	21	5125	8	428	1358	24	7469	1	8	5	
00001.045	3945	6	8267	0	194	13	2713	6725	4	6	0	0	7	2466	8	9783	17	13	5198	41	23	4	25551	5406	2	385	7	9336	2	140	21	5125	8	428	1358	24	7469	1	8	5	
00001.046	3856	8	9222	3	208	0	2773	6854	6	0	18	0	5	2573	10	9988	21	28	10853	54	54	32	28377	5495	32	2341	11	10484	9	194	21	5714	8	411	1622	2	1	5	3	7
00001.047	2667	1694	6895	1	171	26	3047	6908	6	6	11	0	5	2043	16	7175	5	15	6379	1	23	23	15886	3164	4	1311	11	6220	4	95	21	3215	8	307	1288	2	3	4	10	3
00001.048	4497	2053	3861	1	2150	15	3217	6908	2	10	4	4	3	2552	20	6921	24	24	1552	11	32	0	4779	6349	38	3440	9	6220	4	95	21	3215	8	307	1288	2	3	4	10	3
00001.049	4315	2598	3933	7	197	9	3180	5	4	13	6	6	2	2098	695	13028	28	19	16791	9	28	30	5952	15	9163	9	6855	9	10	9	5	10	34	468	807	4	0	1	0	1
00001.050	4421	2614	15984	5	253	6	3356	3	11	3	2	13	6	1940	20	6	0	30	9	3	43	13	5375	30	9113	5	11832	2	133	7	6840	10	402	1370	14	6771	4	3	3	
00001.051	4823	2426	14073	14	265	9	4176	3	2	8	9	46	772	4086	1107	10051	42	43	31908	29	126	36	11665	10	9768	37	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.052	4477	2684	13542	3	202	0	3643	1	2	13	13	13	668	3230	975	13704	235	85767	29271	12	23	46	12547	86	5664	4	16	19	8	1418	1976	21	18	4728	7	6	2	4	5
00001.053	4585	4285	15176	20	249	56	3456	25	36	33	47	75	668	3230	975	13704	235	85767	22222	160	190	225	9022	153	5126	233	212	250	160	5	18	8	182	4429	2	4	5	24	25
00001.054	3780	2402	13334	11	167	3	2967	0	6	5	13	46	580	2856	767	13032	21	17	17078	18	4	28	5905	20	7291	1	3881	4	5	14	4	12	435	4366	4	3	7	19	3548
00001.055	4556	2616	14732	7	204	9	3727	3	6	15	4	37	743	3689	942	16643	36	19	23303	4	15	48	8288	18	6139	25	3704	14	18	8	12	4	518	5040	16	1	10	17	4156
00001.056	2764	2368	11979	7	215	3	2126	0	0	6	0	9	642	2907	822	13427	24	17	17831	14	10	10	7336	5	3915	5	1	2	2	12	4	466	4697	7	4	7	19	3627
00001.057	35	2697	4736	7	167	6	1	0	0	6	0	11	617	2785	651	12942	25	17	16240	7	6	6	6918	9	4898	1395	6	4	2	4	428	4661	7	4	0	14	6	3557	
00001.058	22	2558	4324	5	200	3	1	0	0	6	0	5	593	2680	765	8084	25	28	16513	7	5	6	6893	25	3442	5	6	2	6	16	6	16	441	4284	4	3	5	9	3583
00001.059	12	2682	4264	1	179	0	1	0	0	6	0	5	498	2655	456	7414	35	41	15278	6	8	2	6627	5	3442	5	6	2	6	16	16	16	1922	3318	7	5	19	11	2958
00001.060	7	1861	3208	1	159	0	1	0	0	6	0	4	470	1954	454	1441	33	4	8240	6	8	2	6627	5	3442	5	6	2	6	16	16	16	1922	3318	7	5	19	11	2958
00001.061	2	2499	4193	1	214	6	0	0	0	6	0	4	545	2135	601	1837	40	41	21749	38	10	10	5712	10	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.062	10	2312	4065	1	214	6	0	0	0	6	0	4	530	11	603	1837	29	54	18762	22	17	24	4950	12	2628	2	5766	9	11	4428	1216	2709	7	16	2	2	1	4	17	5
00001.063	10	2543	4687	1	200	0	0	0	0	6	0	4	545	2135	601	1837	40	41	21749	38	10	10	5712	10	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.064	12	2838	5062	1	251	0	0	0	0	6	0	4	530	11	603	1837	29	54	18762	22	17	24	4950	12	2628	2	5766	9	11	4428	1216	2709	7	16	2	2	1	4	17	5
00001.065	7	3450	6832	1	247	0	0	0	0	6	0	4	545	2135	601	1837	40	41	21749	38	10	10	5712	10	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.066	15	3048	5408	1	186	0	0	0	0	6	0	4	530	11	603	1837	29	54	18762	22	17	24	4950	12	2628	2	5766	9	11	4428	1216	2709	7	16	2	2	1	4	17	5
00001.067	10	3304	5526	14	249	0	0	0	0	6	0	4	545	2135	601	1837	40	41	21749	38	10	10	5712	10	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.068	17	3012	5546	11	164	0	0	0	0	6	0	4	530	11	603	1837	29	54	18762	22	17	24	4950	12	2628	2	5766	9	11	4428	1216	2709	7	16	2	2	1	4	17	5
00001.069	17	2552	4545	1	164	3	0	0	0	6	0	4	545	2135	601	1837	40	41	21749	38	10	10	5712	10	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.070	7	2838	5062	1	251	0	0	0	0	6	0	4	530	11	603	1837	29	54	18762	22	17	24	4950	12	2628	2	5766	9	11	4428	1216	2709	7	16	2	2	1	4	17	5
00001.071	17	3012	5546	11	164	0	0	0	0	6	0	4	545	2135	601	1837	40	41	21749	38	10	10	5712	10	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.072	17	2552	4545	1	164	3	0	0	0	6	0	4	530	11	603	1837	29	54	18762	22	17	24	4950	12	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.073	7	2552	3846	5	150	0	0	0	0	6	0	4	545	2135	601	1837	40	41	21749	38	10	10	5712	10	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.074	7	2552	4545	1	164	3	0	0	0	6	0	4	530	11	603	1837	29	54	18762	22	17	24	4950	12	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.075	17	3012	5546	11	164	0	0	0	0	6	0	4	545	2135	601	1837	40	41	21749	38	10	10	5712	10	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.076	17	2552	4545	1	164	3	0	0	0	6	0	4	530	11	603	1837	29	54	18762	22	17	24	4950	12	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.077	7	2552	3846	5	150	0	0	0	0	6	0	4	545	2135	601	1837	40	41	21749	38	10	10	5712	10	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.078	0	2552	4545	1	164	3	0	0	0	6	0	4	530	11	603	1837	29	54	18762	22	17	24	4950	12	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.079	0	2552	3846	5	150	0	0	0	0	6	0	4	545	2135	601	1837	40	41	21749	38	10	10	5712	10	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.080	2	2684	3	7	133	0	0	0	0	6	0	4	530	11	603	1837	29	54	18762	22	17	24	4950	12	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.081	10	2173	9	0	133	3228	8	5140	2	4878	11	3	545	2135	601	1837	40	41	21749	38	10	10	5712	10	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3
00001.082	5	2387	3	1	188	3306	3761	10	4315	9	6	1071	13	530	11	603	1837	29	54	18762	22	17	24	4950	12	2628	2	6001	9	5	1233	9	5	5421	14	8245	4	0	4	0	3

Figure 3.3: Caption for this image is on previous page.

- BAC pool sequencing gives the possibility to produce fractions of a target genome;
- Reads produced from BAC pools sequencing and aligned on a given sequence actually gives information about its presence in a given pool (i.e. in that pool there is a BAC clone covering the relative genomic region);
- A profile of presence and absence for a given sequence can be created by looking at its number of aligned reads in each pool;
- Sequences that lie close on the genome have comparable profiles of reads count across the whole set of pools.

Estimate profile distances

The situation presented in figure 3.3 derives form the profile analysis of a large region of the genome already assembled. In such a picture it is quite easy to see profiles that look similar because they are already close one to the other. But is it possible to do this backward? In other words, is it possible to place *smaltigs* one next to the other by looking only at their profiles? Is it possible to develop a method able to estimate distances between profiles?

To answer this question all the contigs of the draft assembly were fragmented to *smaltigs* of an average decided length. Reads from all the pools were aligned on these database of *smaltigs* in order to produce profiles of each of them along the whole set of pools. The reads counts of each *smaltig* in each pool went through a double normalization step.

A first normalization take into account the length of each *smaltigs*. The length of the *smaltigs* depends also on the length of the native contig. The virtual fragmentation step was tuned in order to reduce differences in length between *smaltigs* coming from the same contig. However, contigs that are shorter than the chosen *smaltig* length will have smaller size. The number of reads aligning on a sequence is directly correlated with the length of that sequence – i.e. a longer sequence will have an higher number of sequences on it. To eliminate this bias the reads counts for each *smaltig* were normalized on its length.

A second normalization take into account the total number of reads obtained in sequencing each pool. In fact, the sequencing reaction of one pool could performed better than the one of another pool, resulting in a higher total number of reads (see table 5.1). This could compromise the comparison between profiles. In order to prevent a possible bias caused by the highly variable number of

total reads produced per pool (see standard deviations in table 3.2) reads count per pool were normalized on the total number of reads produced by that pool.

On the other hand for an analysis of presence and absence one may wonder if it is really useful to know the actual number of reads aligning on a given sequence instead of just knowing that same align on it and therefore manage only a sort of "Yes" or "Not" information. The advantage of analyzing the actual number of reads aligning on a fragment come from the fact that BAC clones from the same pool are present in different quantities. This can be seen in figure 3.3 where some regions of some pools, for example pool 13, are uniformly covered but with different "intensity". This indicate that two BACs have a different representation in the pool, probably due to its different DNA concentrations within the pool.

Is it possible to estimate differences between profiles taking into account also differences on reads counts due to BAC concentration? BACs with different concentrations in a pool will produce different amounts of reads. The method to perform the profiles comparisons should be able to manage this kind of data. In this way the it will be possible to see the *smaltigs* that are present actually on the same BAC clone and not just those that are in the same pool.

The initial strategy was to perform a clustering analysis of the profiles using Pearson Correlation index. This strategy is largely used to analyze gene expression data in microarray experiments. It permits to cluster together genes that have similar pattern of up-regulation and down-regulation despite their absolute values of expression. This kind of data are similar to those of reads counts for *smaltig* except for the absence of negative data. The results of these analysis (data not shown) actually permitted to cluster together *smaltigs* belonging to a portion of their native contigs but does not go further than this. In other word, it does not permit to produce anything more than small clusters of very similar profiles, almost identical. Moreover, the Pearson index does not give a proper estimation of differences among profiles but an indication of correlation among profiles.

3.2.3 Matrix development

Pearson correlation index clusters together counts profiles that are very similar. But only *smaltigs* that are close on the genome will have identical profiles. On the contrary, *smaltigs* that are distant on the genome would share some reads counts in common in different pools. The profiles of these distant *smaltigs* will be similar not identical and their similarity will decrease with increasing

distance. The most interesting information in a genome mapping view are actually those about genomic regions that are physically distant on the genome. If the method to compare profiles is not able to manage these differences most of the information will be lost. Is it possible to develop a method to compare these profiles that could be able to give information about their distances more than just about their similarity?

00001.016	2	1
00001.017	10	7
00001.018	5	1
00001.019	5	3
00001.020	7	3
00001.021	12	5
00001.022	7	6609
00001.023	5	9044
00001.024	0	8667
00001.025	5	7311
00001.026	25	8350
00001.027	2	8158
00001.028	0	7231
00001.029	5	7095
00001.030	12	8024
00001.031	3088	9502
00001.032	2388	6076
00001.033	4565	10469
00001.034	3184	7002
00001.035	2838	6413
00001.036	1662	3664
00001.037	3139	7002
00001.038	2270	4795
00001.039	2807	6831
00001.040	3733	6534
00001.041	2363	4177
00001.042	3279	6078
00001.043	3926	6907
00001.044	2777	5351
00001.045	3708	7026
00001.046	4342	7638
00001.047	3945	6725
00001.048	3856	6854
00001.049	2667	4878
00001.050	4009	6199
00001.051	4477	6908
00001.052	4315	5
00001.053	4421	3
00001.054	4823	3
00001.055	4477	1
00001.056	4585	25
00001.057	3780	0
00001.058	4556	3
00001.059	2764	0
00001.060	35	0
00001.061	22	1
00001.062	12	1
00001.063	7	0
00001.064	7	1
00001.065	10	5
00001.066	10	1
00001.067	12	0

Figure 3.4: In this figure are reported the read counts for some *smaltigs* of the contig00001 in the pools P01 on the left and P08 on the right. For convenience *smaltigs* are indicated with the number of the native contig *dot* a progressive number according to its original position on the native *contig*.

In any column of figure 3.3 it can be seen that some regions are uniformly and continuously covered – look for example at pools P01 and P08 highlighted

in figure 3.4. Recall that *smaltigs* are ordered along the native contig, so they represent the continuous sequence of contig 1. Thus, those regions homogeneously covered indicate the presence of a single BAC clone for that genomic region. A BAC clone is a physical indication of the proximity of two sequences. This concept could be extended to all the pool and to all the other contigs. So is it possible to use the physical information provided by BAC clones that cover large contig, to infer something about the physical distance?

A single BAC clone in a given pool gives a comparable number of reads counts to all the *smaltigs* that it covers. This means that those *smaltigs* could be identified as close on the genome by looking at their read counts. But in presence of many BAC pools each with many BAC clones, looking at single numbers of counts will be difficult and probably meaningless. A more useful tool could be the estimation of the probability that two sequences that share the same number of reads in a given pool, do it because they are actually close on the genome and not just for chance.

The read count values in a pool for *smaltigs* belonging to the same contig change when a BAC begins or finishes whereas in the middle they are more or less constant. Looking at these increases and drops of counts it is possible to see the number of times that a BAC clone starts or ends within a contig. On the other hand, if the count values do not change too much, it could means that the two *smaltigs* are on the same BAC, thus, that they are close on the genome. In order to reduce the small differences between counts, that can be observed within BAC clones, a viable strategy is to use classes of counts, like for example those in table 3.4.

By looking at the whole draft assembly in terms of *smaltigs* belonging to the same contig, it is possible to estimate the number of times that a variation in counts values occurs in any pools. In this way it is possible to produce a matrix of observed occurrences of each possible transition from one class to another one like the one shown in table 3.5.

With these observed occurrences it is possible to estimate their frequency in respect to the total of occurrences. This total occurrences refers to the total number of changes from one class to another one. The frequency values could be useful to calculate the expected occurrences for each class transitions. Expected occurrences are thus calculated by multiplying the total occurrences for each class per each frequency of transition. In table 3.6 are reported the expected values calculated from matrix in table 3.5.

At this point it is possible to produce a scoring system that represent the probability to change from a given number of counts to another one for

Class	Lower limit	Upper limit
0	0	1
1	2	3
2	4	7
3	8	15
4	16	31
5	32	63
6	64	127
7	128	255
8	256	511
9	512	1023
10	1024	2047
11	2048	∞

Table 3.4: Classes of read counts.

fragments that are close on the genome. The scores for each class transitions are calculated with the logarithm of the ratio between observed and expected multiplied for a constant k. Note that values for a transition and for the opposite one are different both in observed and expected matrixes. To eliminate this bias a symmetric scoring matrix is obtained by calculating the media between scores for the same class transitions, see table 3.7.

Some of the classes used for the construction of the scoring matrix covers a wide range of values. To create smooth intervals between classes the final scoring matrix is interpolated in order to create many more classes with their relative scores, mainly for high numbers of reads count. The interpolated matrix is not shown here for problems of space.

Scoring matrix validation

The scores indicated in this matrix represent the probabilities to observe a given transition in reads count between two fragments that are close on the genome. With this scoring matrix it could be possibile to compare the profiles of two *smaltigs*. The profiles are compared by looking at one pool at a time: for each pool the two values, each belonging to one *smaltig*, are compared and a score is assigned according to the matrix. At the end, the comparison of the two profiles will have as many scores as the total number of pools, in this case sixty-four. A global score for the comparison is obtained from the sum of any

Class	0	1	2	3	4	5	6	7	8	9	10	11
0	12475	5497	5403	2430	410	58	36	28	50	49	78	125
1	5551	3975	4986	2930	562	84	30	24	29	40	52	103
2	5263	5088	8456	7080	1950	201	47	51	39	85	96	169
3	2387	2909	6995	10808	5492	683	81	58	53	81	109	181
4	447	583	1986	5372	7534	2354	169	61	40	40	66	140
5	53	58	213	695	2351	3514	779	62	35	25	41	91
6	30	34	52	74	165	802	1576	402	53	24	39	47
7	34	18	44	57	61	50	404	1585	568	99	51	62
8	39	34	71	49	61	23	61	550	2738	896	145	110
9	43	43	65	92	60	24	29	81	908	4654	1645	358
10	66	53	87	99	67	36	31	34	166	1645	9223	3092
11	128	84	179	179	145	67	34	62	128	381	3037	46399

Table 3.5: Matrix of observed occurrences.

Class	0	1	2	3	4	5	6	7	8	9	10	11
0	3308	2281	3543	3706	2334	983	409	376	593	993	1813	6312
1	2281	1572	2442	2555	1609	677	282	259	409	685	1250	4352
2	3543	2442	3793	3968	2499	1053	438	403	635	1064	1941	6759
3	3706	2555	3968	4151	2614	1101	458	421	664	1113	2031	7070
4	2334	1609	2499	2614	1646	693	288	265	418	701	1279	4453
5	983	677	1053	1101	693	292	121	111	176	295	538	1876
6	409	282	438	458	288	121	50	46	73	123	224	781
7	376	259	403	421	265	111	46	42	67	113	206	718
8	593	409	635	664	418	176	73	67	106	178	325	1132
9	993	685	1064	1113	701	295	123	113	178	298	544	1896
10	1813	1250	1941	2031	1279	538	224	206	325	544	993	3459
11	6312	4352	6759	7070	4453	1876	781	718	1132	1896	3459	12043

Table 3.6: Matrix of expected occurrences.

single score. This global score should represent the physical distance between the two *smaltigs*.

Given this scoring system, the global scores for profiles of *smaltigs* that are very close on the genome will have high positive values. On the other hand, *smaltigs* that are very distant on the genome or even unrelated (i.e. in different chromosomes) will have a negative global score. In the middle all the positive scores could indicate a physical relation between the relative *smaltigs*. Does this method of scoring system works as predicted? Is it true that by comparing profiles with this method, *smaltigs* that are distant on the genome have negative scores? On the other hand *smaltigs* that are physically close have actually high values?

In figure 3.5 are displayed the distribution of the scores for the comparison of the real profiles of the *smaltigs* and for a set of random profiles for the same *smaltigs* (the scoring matrix is constructed using the real profiles). The random profiles are created starting from the real profiles, by randomly mixing all the

Class	0	1	2	3	4	5	6	7	8	9	10	11
0	53	35	16	-17	-67	-115	-100	-99	-104	-122	-129	-156
1	35	37	28	5	-41	-90	-87	-100	-103	-112	-127	-153
2	16	28	32	22	-9	-65	-87	-85	-97	-106	-122	-146
3	-17	5	22	38	29	-18	-71	-79	-102	-102	-118	-146
4	-67	-41	-9	29	60	48	-21	-58	-84	-105	-118	-137
5	-115	-90	-65	-18	48	99	75	-27	-72	-100	-106	-126
6	-100	-87	-87	-71	-21	75	138	86	-9	-62	-74	-118
7	-99	-100	-85	-79	-58	-27	86	145	84	-9	-63	-97
8	-104	-103	-97	-102	-84	-72	-9	84	130	64	-29	-90
9	-122	-112	-106	-102	-105	-100	-62	-9	64	109	44	-65
10	-129	-127	-122	-118	-118	-106	-74	-63	-29	44	89	-4
11	-156	-153	-146	-146	-137	-126	-118	-97	-90	-65	-4	53

Table 3.7: Matrix of symmetric scores for transition from a reads count class to another one.

reads count of all the *smaltigs*. These profiles lose all the physical information because *smaltigs* that are originally close will not share any reads count in common.

As it can be seen in the graphs the scores for the random profiles have all negative values. On the contrary, the scores for the real profiles display a wide distribution of values with many negative values but with a considerably high fraction of positive values in respect to that of random profiles.

These graphs indicate that there are meaningful differences between distances for random profiles and for real profiles. Could this indicate that the real profiles actually carry information about physical distance? Are the positive scores a useful indication for the proximity of *smaltigs*? Each *smaltig* can be compared against all the others in order to obtain a huge list of scores, both positives and negatives. Within this list, will the neighboring *smaltigs* have the higher positive scores? In other words, does this scoring method works properly in joining together *smaltigs* that are actually close on the genome?

Each *smaltig* created from the draft assembly can present a profiles of read counts as a result of the alignment of the 64 pools. These profiles can be compared with the scoring system presented above and the result is a set of lists, one for each *smaltig*. Each of these lists indicates a set of candidate neighbor *smaltigs* with their relative scores. Look for example at table 3.8 in which are reported the two lists of neighbors for the two first *smaltigs* of contig00001.

In these two lists are indicated only the *smaltigs* with positive scores, those that are indeed candidate neighbors. The highest score in each list is the result of the comparison of the query profile against itself (here, *query* is intended as the *smaltig* that has been searched for neighbors). It can be seen that many

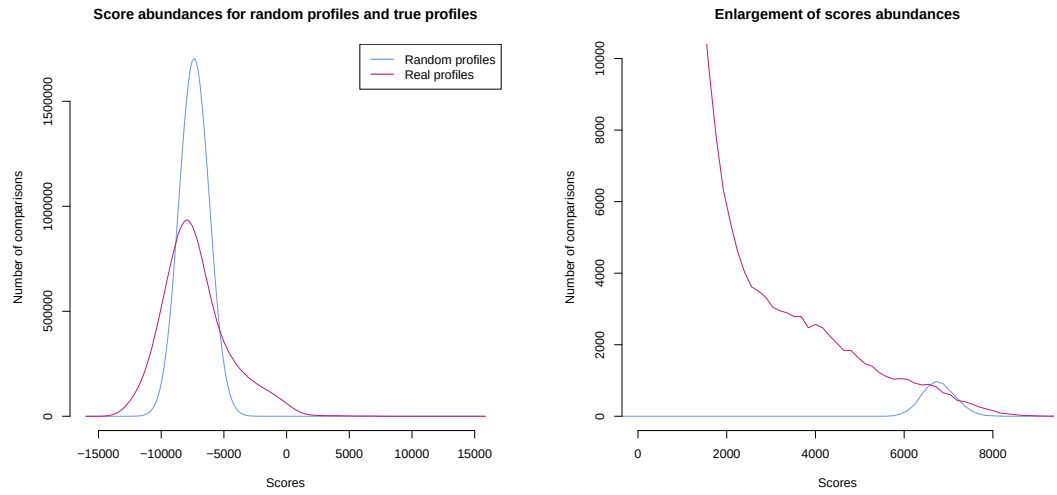


Figure 3.5: Distribution of scores for random profiles and real profiles for *smaltigs* of 5000 bp. The graph on the left shows the complete distribution of the two population of scores, the graph on the right shows only positive scores. The small peak of positive values in the random distribution indicates the scores of each query profile against itself.

of the candidates are themselves part of the contig00001. Moreover, many of these *smaltigs* belonging to contig00001 are those that actually lie next to the two query *smaltigs* (recall that the three numbers after the dot in the *smaltig* code indicate the position on the native contig, see figure 3.4).

However, these two lists present some *smaltigs* belonging to contigs different from the contig of the query *smaltigs*. This could be embarrassing because it could mean that the scoring system does not work properly. If this would be the case, the presence of some *smaltigs* from the same contig of the query *smaltig* could be given just by chance. But what about *smaltigs* that lie in the middle of a large contig? If the scoring system actually works their candidate neighbors would be only *smaltigs* from one side or the other one of the contig. Otherwise, if some *foreign smaltigs* are present the scoring system should be revised.

However, in this case another possibility could be considered. In fact, contigs are generated by whole-genome shotgun assembly programs, which are not free of errors. For this reason a contig could be misassembled, producing a chimera that indeed could cause the calling by our system of foreign contigs even from its middle. But the assembly has been produced by Newbler staring from long 454 reads: the creation of chimera with this software is quite unusual. Moreover, there are other independent evidences that confirm that the assembly is a high quality draft assembly (i.e.: SOLiD mate-pair mapping into the contigs).

Neighbors of *smaltig* 00001.001	Scores
00001.001	7146
00421.001	5964
00084.011	5904
00084.010	5711
00001.002	5672
00421.002	5424
00759.001	5256
00084.009	4896
00084.008	4325
00001.003	3502
00001.005	3081
00001.006	2832
00001.007	2809
00084.007	2538
00084.006	2339
00001.004	2174
00084.005	1889
00001.008	1764
00001.013	714
00001.012	669
00001.009	600
00001.010	472
00001.014	91

Neighbors of *smaltig* 00001.002	Scores
00001.002	7000
00001.001	5672
00084.011	5554
00421.001	5466
00759.001	5387
00421.002	5384
00084.010	5218
00084.009	4328
00084.008	3655
00001.003	3592
00001.006	3072
00001.005	2914
00001.007	2883
00001.004	2466
00084.007	2254
00001.008	1977
00084.006	1884
00084.005	1801
00001.012	1223
00001.009	823
00001.013	711
00001.010	170
00001.011	92

Table 3.8: Candidate neighbors for *smaltigs* 00001.001 and 00001.002. These two *smaltigs* lie at one extremity of the contig00001. In light blue are indicated the *smaltigs* that belong to contigs different from contig00001.

In table 3.9 are listed the neighbors for the two *smaltigs* that are at the middle of contig00001.

As it can be seen in these two lists, the *smaltigs* that are identified as possible neighbors in the middle of contig00001 that is more than 500 kb long, are only *smaltigs* of the same native contig. Moreover, the candidate neighbor *smaltigs* are not far away from the middle of the contig.

A similar situation should happen also in other large contigs. The contig length threshold for this analysis depends on the number of *smaltigs* per contig that indeed depends on the length of the *smaltigs*, see section 2.8. In fact, if the contigs are large enough the lists of neighbors of a smaltig placed in the middle will present only smaltigs of the same contig, whereas, in smaller contigs the "middle smaltig" will link also to foreign contigs. Given that each list has on average 20 neighbor candidates this analysis can be carried out in contigs that are larger than 120 kb. In fact with an average *smaltigs* size of 5000 bases a contig of 120 kb has 24 different *smaltigs*.

In the draft assembly there are 29 contigs larger than 120 kb. By looking at the three middle *smaltigs* of each of these large contigs only in four out of 87 lists there are some foreign contigs. In the 95.4% of the middle *smaltigs*

Neighbors of *smaltig* 00001.050	Scores
00001.050	7436
00001.051	5358
00001.047	4504
00001.048	4432
00001.046	4398
00001.049	4079
00001.043	3271
00001.044	3107
00001.040	3107
00001.042	3096
00001.058	2585
00001.045	2558
00001.052	2505
00001.054	2422
00001.057	2171
00001.039	2022
00001.053	1986
00001.041	1918
00001.055	1711
00001.059	1583
00001.056	934
00001.033	592
00001.061	536
00001.035	510
00001.037	389
00001.060	292
00001.062	184
00001.034	44

Neighbors of *smaltig* 00001.051	Scores
00001.051	7344
00001.050	5358
00001.048	4803
00001.046	4543
00001.049	4266
00001.047	4128
00001.052	3335
00001.042	3049
00001.043	2937
00001.040	2888
00001.058	2809
00001.045	2676
00001.059	2619
00001.044	2581
00001.055	2502
00001.053	2446
00001.054	2272
00001.057	2186
00001.041	1781
00001.039	1548
00001.060	1274
00001.056	1038
00001.061	959
00001.062	300

Table 3.9: Candidate neighbors for two *smaltigs* at the middle of contig00001.

the candidate neighbors are *smaltigs* from the same contig. The *smaltigs* in which there are some foreign contigs are the 00023.016, 00028.011, 00028.012 and 00029.014. However it has to be pointed out that the lengths of their native contigs, c00023 c00028 and c00029, are close to the threshold value and moreover, the foreign *smaltigs* have very low scores.

For reasons of space in these examples are shown results obtained with *smaltigs* of an average length of 5,000 bases. Also analyzing *smaltigs* of smaller size, 2,500 bases on average, the results completely agree with those presented here. Even in this case the 95.4% of the middle *smaltigs* present as candidate neighbors only *smaltigs* of the same native contig. Those that call foreign contigs are the central *smaltigs* of contigs c00023 c00028 and c00029.

With these evidences the specificity of the scoring system seems to be confirmed. So, what is going on at the extremities of the contig? Given that the scoring system seems to work properly, the foreign contigs called at the beginning of contig00001 should be close to it on the genome. And what about the opposite extremity? Is this event present also there?

Table 3.10 shows the lists for the two *smaltigs* that lie at the other extremity of contig00001.

Neighbors of *smaltig* 00001.099	Scores
00001.099	7248
00001.100	5737
00018.002	5446
00018.003	5223
00018.001	5069
00001.097	5013
00018.004	4936
00001.096	4908
00001.098	4893
00001.094	4101
00001.095	3914
00001.093	3781
00018.006	3036
00018.005	2877
00001.092	2673
00018.007	1985
00001.090	999
00018.008	960
00001.091	878
00001.086	271
00001.087	85
00001.088	61
00018.009	57
00001.089	7

Neighbors of *smaltig* 00001.100	Scores
00001.100	6990
00001.099	5737
00018.002	5510
00018.003	5424
00018.001	5310
00018.004	5140
00001.098	5035
00001.097	4840
00001.096	4802
00001.095	4032
00001.094	3949
00001.093	3946
00018.006	3227
00018.005	3141
00001.092	2767
00018.007	2420
00018.008	1276
00001.090	1184
00001.091	831
00001.086	333
00001.087	153
00001.088	53

Table 3.10: Candidate neighbors for the two final *smaltigs* of contig00001. In light blue are indicated the *smaltigs* that belong to contigs different from contig00001.

Even at this extremity some foreign *smaltigs* are identified as candidate neighbors together with *smaltigs* that are known to be close to the query ones. If these contigs called from the extremities are true positives, they should indicate that the scoring system is able to join not only *smaltigs* that are already close on the assembly but also *smaltigs* that are on different contigs. This may mean that the system developed to estimate distances could actually join together different contigs on a physical bases.

The team that worked on the *N. gaditana* genome sequencing project moved on from the draft assembly of 454 reads to a final draft of the whole genome. The project was carried out independently from the work presented in this Thesis. Beside the 454 shotgun sequencing efforts, the production of the final draft took advantage of two mate-pair libraries, a number of BAC-ends sequences and several transcriptome experimental data [39].

In this Thesis the inferences about connection between contigs are based only on the distances between profiles of presence and absence obtained with the scoring system presented above. The data used as genetic markers are

only the *smaltigs* obtained from all the contigs of the assembly: there are no information about connections among contigs. Given this, a comparison between the final draft and the connection inferred with the profile distances could be made without any bias. Moreover, the two assemblies, the one in publication and the one proposed in this Thesis, could confirm each other.

Thus, this final draft could be useful to confirm that the foreign *smaltigs* called at the extremities of contig00001 are true positive but also to confirm that the method works also in joining different contigs. To do this the final draft could be simply searched for the desired contigs. If these contigs are on the same scaffold or even better on the same chromosome, this will means that there two independent evidences that confirm the same genome structure.

Table 3.11 shows the position on the final draft of the four contigs identified by searching similar profiles to some *smaltigs* of contig00001. In particular, *smaltigs* of contig00018 are called at one extremity of contig00001 by *smaltigs* 00001.099 and 00001.100. At the opposite extremity *smaltigs* 00001.001 and 00001.002 call *smaltigs* that belong to contigs c00421, c00759 and c00084.

Chromosome	Start	End	Contig	Contig length
NG-chr08	197887	376143	contig00018	178257
NG-chr08	376244	377096	contig02446	853
NG-chr08	377197	378099	contig02362	903
NG-chr08	378200	880455	contig00001	502256
NG-chr08	880556	892541	contig00421	11986
NG-chr08	892642	898689	contig00759	6048
NG-chr08	898790	955919	contig00084	57130

Table 3.11: Genomic region for the contigs identified as neighbors of the contig00001 as in the final draft of the *N. gaditana* genome. NG-chr stays for *Nannochloropsis gaditana* chromosome. Start and End columns indicate the position of the contigs within the chromosome.

In the table it can be seen that in chromosome 8 contig00018 is placed before contig00001 that at the other end it is next to contigs c00421, c00759 and c00084. Given that the foreign contigs called at the extremities of contig00001 are confirmed, even the foreign contigs called from the middle *smaltigs* 00023.016, 00028.011, 00028.012 and 00029.014 could be real positive neighbors. In particular, *smaltig* 00023.016 calls contigs c00528 and c00126, the two *smaltigs* of contig00028 both call contig00042 and *smaltig* 00029.014 calls contig00152.

In table 3.12 are reported the regions of the final draft for these contigs: it can be seen that also the contigs that are called from *smaltigs* placed in the

Chromosome	Start	End	Contig	Contig length
NG-chr03	147155	243751	contig00042	96597
NG-chr03	243852	371429	contig00028	127578
NG-chr03	476462	602473	contig00029	126012
NG-chr03	440347	476361	contig00152	36015
NG-scf01	356457	513530	contig00023	157074
NG-scf01	513631	522525	contig00528	8895
NG-scf01	522626	565816	contig00126	43191

Table 3.12: Genomic regions for the contigs identified as neighbors of the contigs c00023, c00028 and 00029 as n the final draft of the *N. gaditana* genome. NG-chr stays for *Nannochloropsis gaditana* chromosome whereas scf stays for scaffold. Start and End columns indicate the position of the contig within the chromosome or scaffold.

middle of these contigs are confirmed as neighbors of the query contigs.

All these evidences suggest that the method developed to estimate distances between profiles of read counts works properly. In fact, with this method it is possible to identify real differences between profiles. The scores assigned to each comparison actually give information about the physical distance between profiles. Moreover, these scores are useful not only in reconstructing contigs starting from their constituent *smaltigs*, but also in joining together different contigs by looking at their *smaltigs*. Finally these candidate neighbor contigs are confirmed to be together by independent evidences.

3.2.4 Building map-scaffolds

The lists of neighbors for each *smaltig* actually indicate physical proximity. In light of this, is that possible to perform a map of the genome? Is that possible to place *smaltigs* one next to the other on a long range scale by looking at their scores?

Two strategies can be pursuit to reach this target. The first one, denoted as mini-scaffolds strategy, it is designed to produce several scaffolds, one for each *smaltig*. The second one called global scaffold strategy aims to produce the largest possible scaffolds.

The mini-scaffolds strategy focus on a single *smaltig* at a time. It consists in ordinating the different candidate neighbors of the list on the right side or left side in respect to the query *smaltig*. The idea is that the candidates with the lower scores are likely to be far away from the query *smaltig*. So if two of these lower-score candidates are selected, it is possible that one will be on a

side of the query and one on the opposite side. These lower-score candidates that fall on opposite sides are called attractors. Each one of the attractors has its own list of candidate neighbors. If the two attractors actually fall on opposite sides, they will share only some *smaltigs* or, in an ideal situation, only the original query *smaltig* because they are very distant one from the other. If this analysis is performed for a number of lower-score candidates it is possible to place the candidate neighbors of the original query *smaltig* on one side or on the other according to their presence in the neighbor lists of the attractors.

With this strategy each *smaltig* has its own mini-scaffold that represents an ordinated boundary around the query *smaltig*. However, it could be difficult to join each of them together in a bigger scaffold because in some cases mini-scaffolds could disagree on the order of the *smaltigs*. A possible strategy would be the creation of a consensus for overlapping mini-scaffolds. The data for this strategy are very preliminary and are not shown in this Thesis. At the moment of writing it is possible to produce only the mini-scaffolds because of the difficulties to develop a overlapping-like algorithm.

The global scaffold strategy aims to directly produce large scaffolds of the genome. This method focuses on identifying connections between different *smaltigs*. It starts from an arbitrary *smaltig* and looks at its neighbors list. Within this list the system selects the neighbor with the highest score – the score of the query *smaltig* against itself is not considered. The selected *smaltig* it is likely to be the closest to the query one and thus a connection between the two is made. The system now moves to the *smaltig* just called and repeats the procedure. If a neighbor list presents *smaltigs* that have been already called they are not considered in the selection of the highest score. In this way the scaffold could be extended until there are *smaltigs* that can be positioned. Once a scaffold could not be extended any more, an uncalled *smaltig* is selected as new starting point for a new scaffold.

The advantage of this method compared to the previous one is that this looks at the possible connection between different *smaltigs*, whereas the mini-scaffolds strategy produces many ordinated regions but that remain unconnected, at least at the moment. A drawback of the global scaffold it is that it could not produce a consensus for the order of the *smaltigs* along the scaffolds.

Testing global scaffold approach

A test of the global scaffold approach can be perfomed on *smaltigs* of 2,500 bp. This size is more suitable than 5,000 bp because it could allow a better

resolution.

Once the reads of each pool are aligned on these *smaltigs* the profiles of read counts can be created. Then each profile is compared against all the others. A score is assigned to all the comparisons and the positive ones are selected to create neighbors lists. Within these lists are searched the possible connections with the global scaffold method described above. Each set of connections between *smaltigs* proposed with the global scaffold approach is called a map-scaffold. A total of the 77 map-scaffolds are produced, 53 of these have more than 13 *smaltigs*. The remaining ones are formed by isolated couples or little groups of *smaltigs*.

These map-scaffolds can be compared with the final assembly of the genome of *N. gaditana*. This comparison can be useful to identify if the proposed map-scaffolds of *smaltigs* are consistent with the final assembly in terms of chromosomes or scaffolds or contigs. To allow a faster representation, each *smaltigs* name is coupled with its relative chromosome or scaffold. An easy representation of the connections identified by this method is the graph shown in figure 3.6 obtained with Graphviz software.

In the figure it can be seen that different contigs are joined together, for instance *smaltigs* of contig00016 are connected to *smaltigs* of contigs c01263, c00490, c00488 and some others. These connection are confirmed by the fact that each of these contigs are actually part of the same chromosome, the NG-chr12 as indicated in the figure. Despite the ramifications, that will be discussed below, this example shows that this global scaffold approach could permit a visual representation of the connections between contigs.

Another example is the contig00001 discussed above. This contig is contained in two separate large map-scaffolds. The map-scaffold 1 contains almost every *smaltigs* of c00001, one portion of map-scaffold 1 is shown in figure 3.7. The portion shown here includes the extremity of contig00001 connected with contig c00421, c00759 and c00084 as previous seen in table 3.11. By looking at the figure it seems that the map-scaffold starts form the first *smaltig* of c00001 and procedes in two separate ways (the blue path). This is actually an artifact generated by the algorithm of global scaffolding.

The procedure, in fact, chooses the first available uncalled *smaltig* as starting point for a new map-scaffold. From here, it then looks for possible neighbors and starts building the map-scaffold. For this reason the start *smaltig* can be anywhere in respect to the resulting map-scaffold. In figure 3.7 in fact, the connections highlighted in blue show the longest path across map-scaffold 1 indicating that the contig00001 should belong to the central portion of a

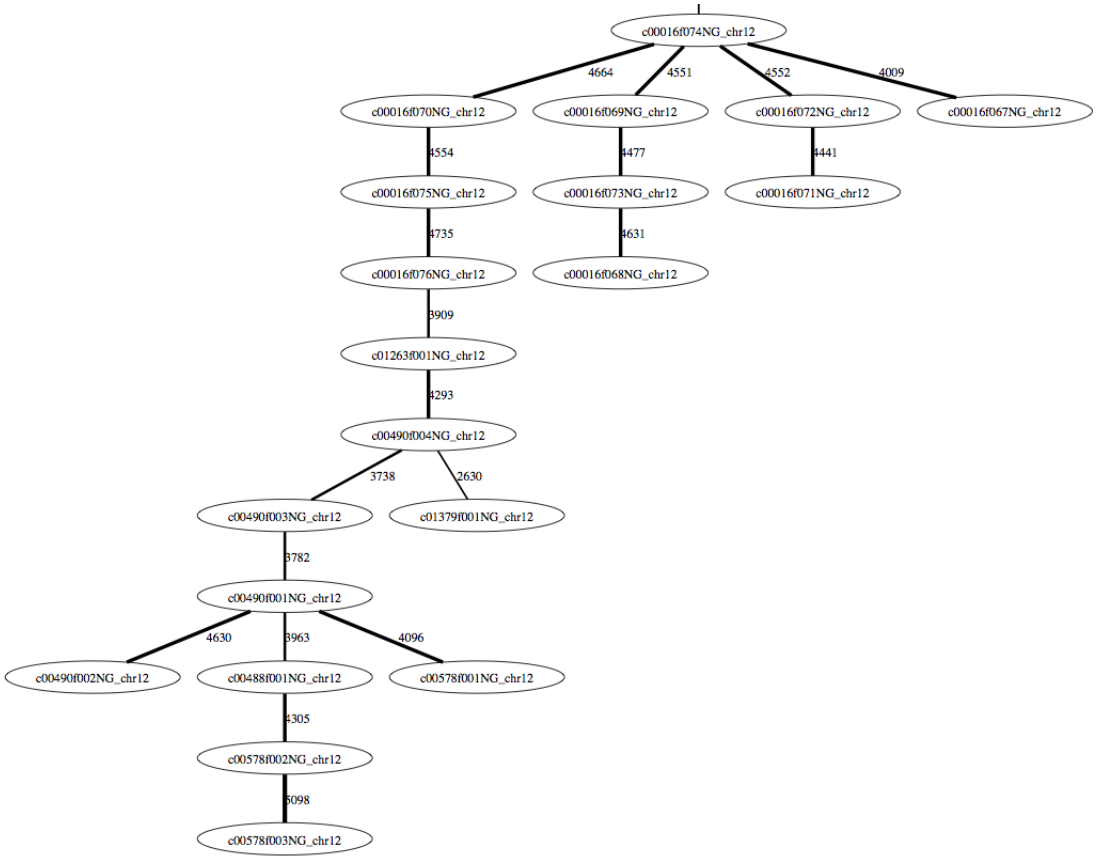


Figure 3.6: In this figure is shown the end portion of map-scaffold 17. Each oval is a single *smaltig*. Their names are coded with the following criteria: for instance c00016 indicates the native contig, f074 indicates the fragment within the native contig and NG-chr12 represents the relative chromosome in the final assembly. The numbers at the connections represent the scores of that comparison.

chromosome. This is coherent with table 3.11.

Most of contig00001 belong to map-scaffold 1, while the remaining portion of c00001 belongs to map-scaffold 2. As shown above, in table 3.11, this extremity should be connected at least to contig00018. Figure 3.8 shows the upper part of this map-scaffold. Curiously, only few *smaltigs* of c00001, those that are connected with c00018 are contained in this scaffold.

In some map-scaffolds the situation looks more clear. In fact, many regions present few ramifications and the path looks more linear. Figure 3.9 shows an example for these situations in chromosome one around the contigs c00096 and c00115.

However, in some of these cases in which the path is linear, the order of some *smaltigs* that are close on the genome is not strictly respected. For

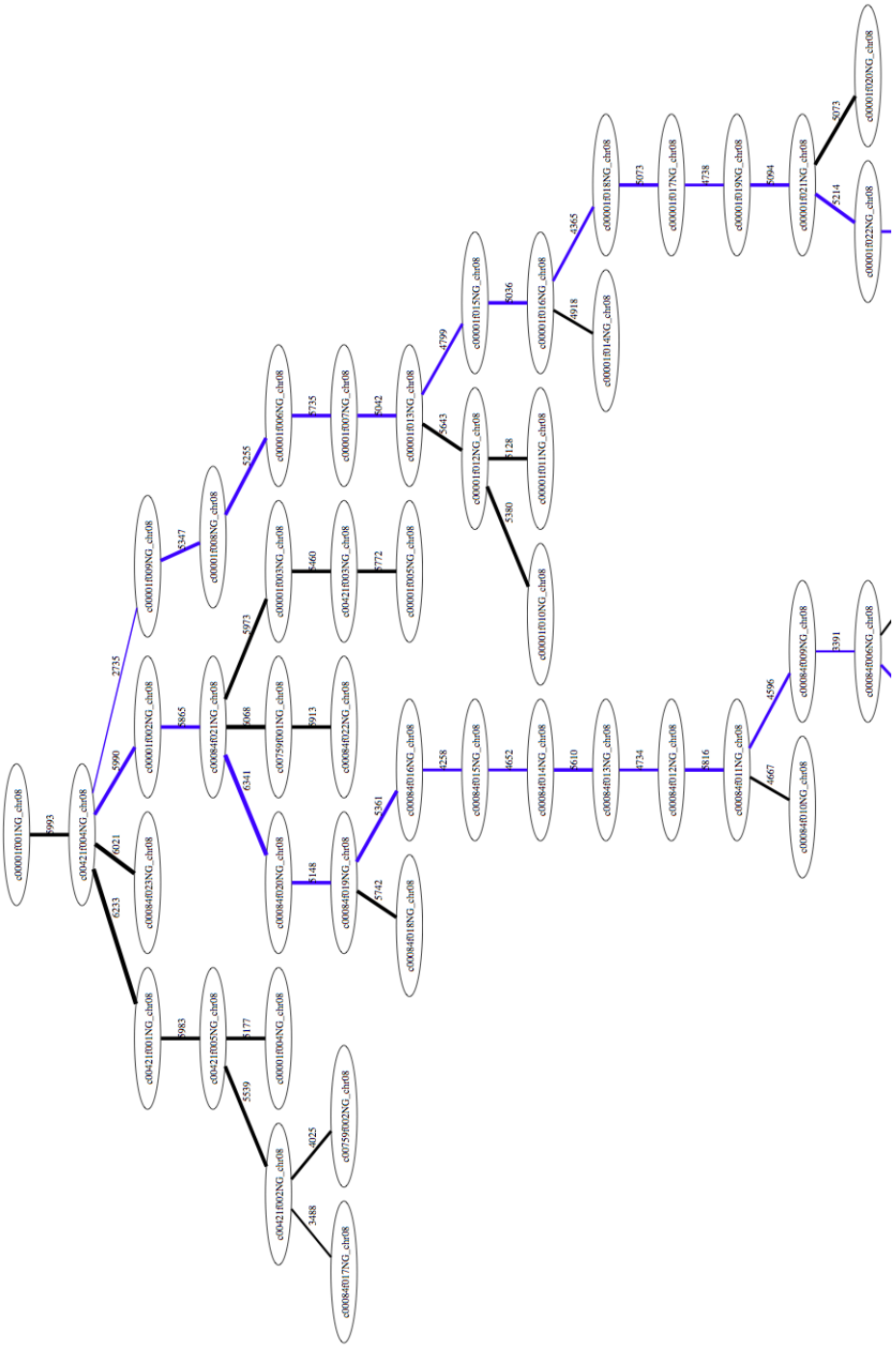


Figure 3.7: Initial portion of Scaffold 1. The blue connections indicate the longest path across the scaffold.

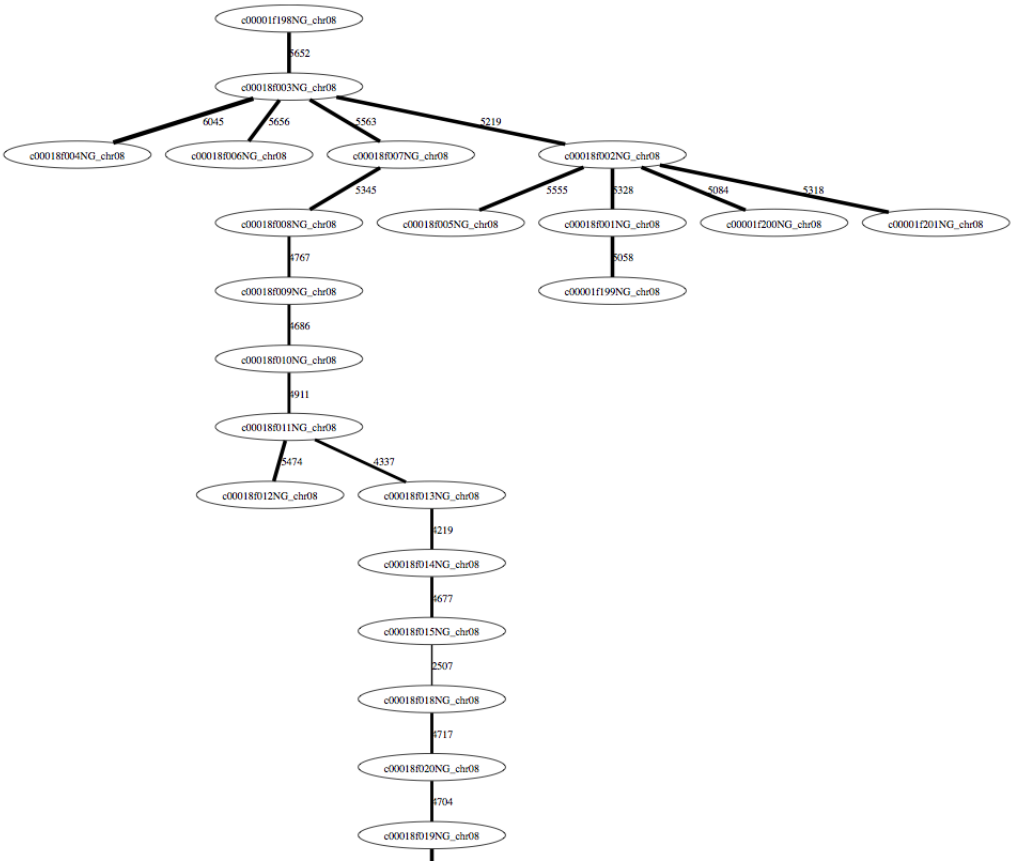


Figure 3.8: Initial portion of Scaffold 2.

example, *smaltigs* that lie one next to the other on the native contig (those that have consecutive numbers as for example c00096f014 and c00096f013) may be erroneously placed on the map-scaffold. In the figure, for instance, *smaltig* c00096f013 is placed after *smaltigs* c00096f014 and c00096f015. The same thing could happen also for different contigs that indeed lie one next to the other on the genome. In the figure, in fact, some *smaltigs* of contig c00115 are placed within *smaltigs* of contig c00096. This could indicate that the two native contigs are very close one to the other. In fact, by looking at the assembly of chromosome 1 the two contigs are actually close, see table 3.13.

Chromosome	Start	End	Contig	Contig length
NG-chr01	1236070	1290247	contig00096	54178
NG-chr01	1290348	1337214	contig00115	46867

Table 3.13: Genomic region for chromosome 1 in the final assembly. The 100 bases of difference between the end of contig00096 and the beginning of contig00115 for convenience are N so there are no other contigs between them in the assembly.

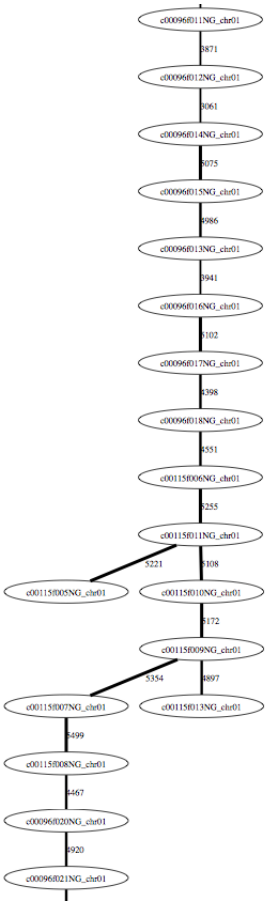


Figure 3.9: Central portion of Scaffold 4. Here are shown connections between *smaltigs* belonging to contig00096 and contig000115.

These examples show that the proposed connection obtained with the global scaffold approach are confirmed in the final assembly of the genome. Moreover, in some other map-scaffolds there are new connections that are not present in the final assembly. This happend especially for isolated small contigs but even for same scaffolds. For example, 101 out of 625 isolated contigs in the final assembly are placed within map-scaffolds obtained with this approach, together with larger scaffolds or even chromosomes.

On the other hand, several chromosomes or scaffolds are divided in different map-scaffolds. Chromosome 1, for example, is split in map-scaffold 4 that accounts for the 77% of the contigs of the chromosome, and scaffold 65 that contains a minor fraction of the entire chromosome. Another example is chromosome 8 that, as seen above, is split in two very large map-scaffolds at contig00001.

Considerations

Some consideration should be done about these results of the global scaffold approach.

The fact that the contig00001 is split in two different map-scaffolds, one for each extremity, is probably due to the fact that the global scaffold method considers only the highest value within a list and marks that one as a connection. Moreover, when a *smaltig* is called it can not be called again. Further investigation should be done on this kind of splitting because the few *smaltigs* of contig00001 that are present in map-scaffold 2 are actually called by some *smaltigs* of the map-scaffold 1 as neighbors, but any connections is done between them.

The ramifications that can be seen in the scaffolds are created during a secondary analysis. Within this step the system traces back all the neighbors lists of the *smaltigs* assigned to a scaffold to find not called neighbors. When one not called neighbor is found, it is attached in that position. In this way it could happen that some *smaltigs* are not perfectly positioned. However, the ramifications within these scaffolds involve regions that are close in the genome, especially *smaltigs* that are one next to the other in the native contig or in final assembly, like for example c00096 and c00115 shown above.

The system presents some problems in assigning the correct order to *smaltigs* that are very close on the native contig or on the genome. This problem it is probably due to the resolution of the method strictly connected to the usage of a BAC library. It is not possible that every single base of the genome becomes the starting point of BAC inserts: inserts will be different for at least some contiguos nucleotides. In this way, these portions of the genome will never belong to different BAC inserts and so they will always be together. This implies that the genetic markers of these undivided regions will have identical profiles. Thus, in comparing the profiles of *smaltigs* from these regions with other *smaltigs* the former will obtain the same scores making impossible to place them in any order.

In has to be pointed out that, even if the entire assembly is not reconstructed with this method, there is no map-scaffold that proposes a connection between different chromosomes of the assembly.

3.3 Test on mate-pair assembly

Given these results in analyzing *smaltigs* from the assembly of 454 reads, is that possible to produce map-scaffolds starting from a different assembly?

As described in Materials and Methods the genome sequencing project of *N. gaditana* implies also the sequencing of mate-pair libraries with SOLiD™ system. In sections 2.9 and 5.2 are described the efforts in producing an assembly from these reads using a short reads assembler. The assembly results in more than 55,000 contigs with an N50 of 727 bases. Is that possible to analyze the contigs obtained from this assembly, hereafter called *veltigs*, with the mapping method proposed in this Thesis?

To perform this preliminary test the reads form the 64 pools can be aligned on the *veltigs* to create profiles of read counts, as well as with the *smaltigs*. However, with the *veltigs* it is not possible to produce the scoring matrix. This is because the matrix is constructed using the informations about variation of reads count classes between *smaltigs* belonging to the same contig. The *veltigs* are single contigs so it is not possible to know which of them are close on the genome and thus looking at variations in read counts. The solution is to use a matrix constructed on *smaltigs*. The matrix used in this test is the one constructed with *smaltigs* 2,500 bases long, the same used for the map-scaffolds reported above.

The number of *veltigs* is very high and could complicate the profiling analysis. Moreover, the majority of them presents very few reads in very few pools or even no reads in any pools. These are thus filtered resulting in 19,690 usable *veltigs*. The profiles of these *veltigs* can be compared according to the scoring matrix in order to identify possible neighbors. Given that these scores are based on a matrix builded on a different set of counts, there is the need to evaluate if the scores could be informative about physical distances. A useful indication could come from the comparison to scores obtained from random profiles. The two distributions are shown in figure 3.10.

The graphs show that there are a considerably large fraction of comparisons with positive scores suggesting that also for this assembly the scoring system could work in identifying physically related *veltigs*.

These scores can thus be analyzed with the global scaffold approach in order to build map-scaffolds of *veltigs*. The analysis produces a total of 73 map-scaffolds but only 46 of them has more than 19 *veltigs*. This threshold is chosen because the map-scaffolds with less than 19 *veltigs* could represent regions very small in the genome. Moreover, of these excluded map-scaffolds

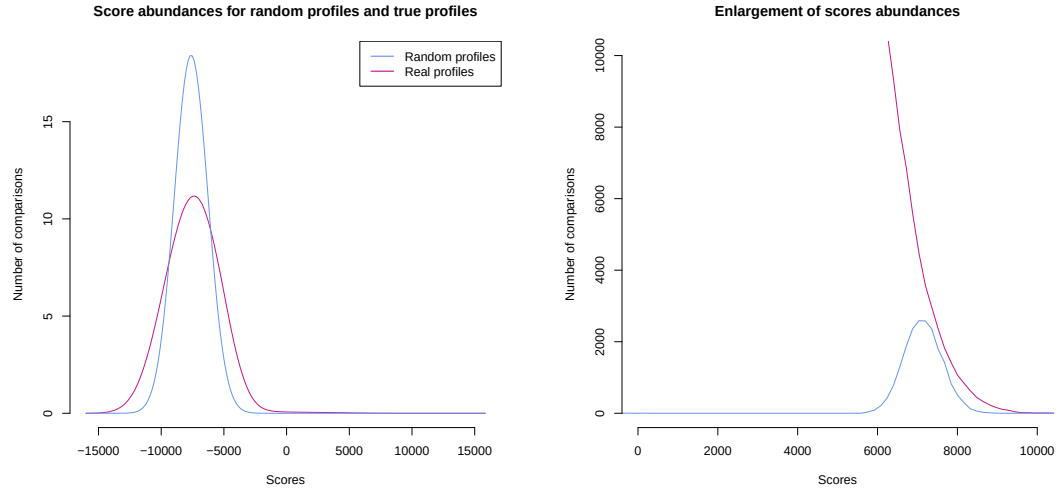


Figure 3.10: Distribution of scores for random profiles and real profiles for *veltigs* comparisons. The graph on the left shows the complete distribution of the two population of scores, y-axis values indicate millions of comparisons. The graph on the right shows only positive scores. The small peak at positive values in the random distribution indicates the scores of each profile against itself.

only five have more than 4 *veltigs*.

The map-scaffolds obtained on these *veltigs* are a little more branched than those obtained on the *smaltigs*. However, there are some regions in which the path is linear as for example those shown in figure 3.11.

Each one of the *veltigs* has its proper position inside a larger 454 contig. These position on the draft assembly can be assigned with a BLAST alignment (see section 5.2) and could be useful to evaluate if the scoring system worked properly also for this assembly. In fact, the path of the *veltigs* in the map-scaffolds should at least respect their positions inside the larger contigs. Table 3.14 shows the relative positions of the *veltigs* in figure 3.11 inside the draft assembly.

The order of the *veltigs* in the two tables is the same that they have on figure 3.11. In the table of scaffold 2 all the *veltigs* belong to the same contig but their positions in the map-scaffold do not respect their actual positions in the relative contig. This is actually the same problem faced with *smaltigs* belonging to the same native contig: the scores can not discriminate between regions that are relatively close on the genome.

The left side table shows the positions of *veltigs* of scaffold 1 in the draft assembly. As it can be seen, some of them belongs to different contigs. These connections indicate that these 454 contigs could be close in the genome. To confirm this suggestion it can be looked at the final assembly of the genome.

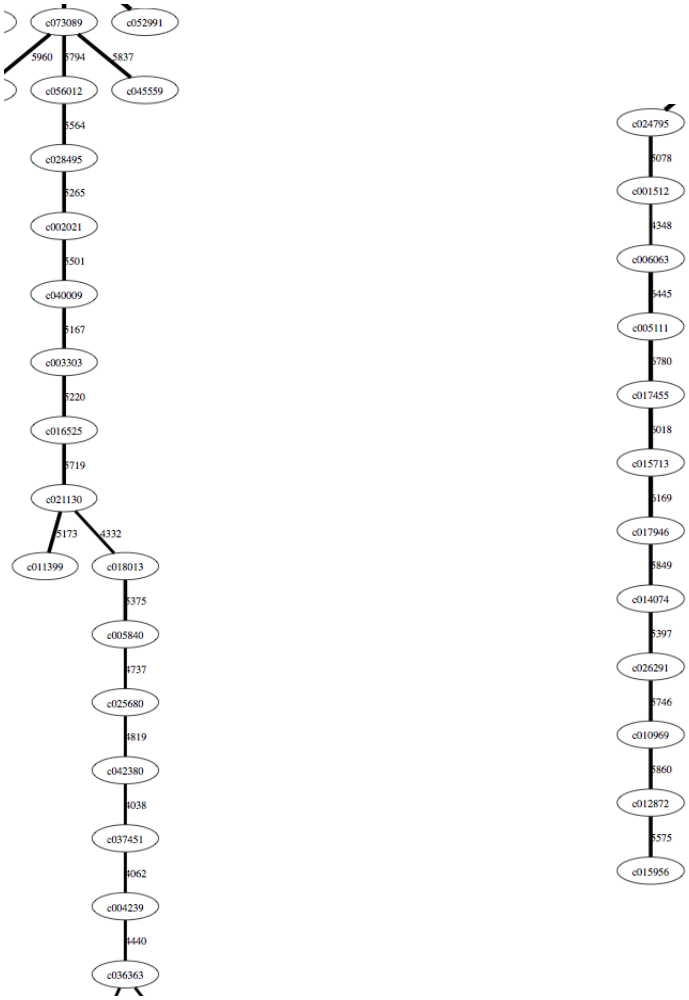


Figure 3.11: Here are shown two part of two different map-scaffolds builded on *veltigs*. The path on the left is a part of the scaffold 1 while the path on the right is the final portion of scaffold 2. The codes within the ovals indicate the number of the *veltig*.

These contigs are inserted in one large chromosome, the chromosome 3. Table 3.15 shows the region of this chromosome containing these contigs.

This table indicates that the positions of the *veltigs* suggested in the map-scaffold reflect the actual position of the contigs in the chromosome. In fact, despite the already discussed problem in positioning very close portions of the genome, the order of the identified contigs is the same. The absence of the contigs in the middle could be due to a bad assembly of those contigs.

Veltig	Contig	Start	End
c073089	00443	2341	2669
c056012	00443	1910	2344
c028495	00443	4896	5373
c002021	00443	1921	1371
c040009	00443	6102	5385
c003303	00575	5338	4637
c016525	00575	5377	6209
c021130	00575	764	1921
c011399	00443	37	854
c018013	00851	986	1860
c005840	00585	4947	4032
c025680	00585	6135	4960
c042380	00851	1869	3173
c037451	00373	8449	7115
c004239	00373	2590	1760
c036363	00373	6648	5962

Veltig	Contig	Start	End
c024795	00004	106374	108075
c001512	00004	97922	101016
c006063	00004	54593	50700
c005111	00004	60541	58311
c017455	00004	73192	71422
c015713	00004	47419	48862
c017946	00004	55527	56884
c014074	00004	75759	74583
c026291	00004	66735	67582
c010969	00004	54789	55538
c012872	00004	68541	67594
c015956	00004	50241	49126

Table 3.14: These two tables illustrate the position on 454 contigs of the *veltigs* shown in figure 3.11. On the left there are the *veltigs* of scaffold 1 and on the right the *veltigs* of scaffold 2. Columns "start" and "end" indicate the relative starting and ending positions of the *veltigs* inside the 454 contigs.

The results about the mapping of these contigs generated by short-reads assembly are very preliminary. A better assembly focused on reducing the small contigs and the possible chimera would result in more useful map-scaffolds.

However, the regions like those shown here, in which the path through the *veltigs* reflects the real situation on the genome, suggest that the method to estimate distances works also on different assemblies.

This pilot test on short reads assembly suggests that also the reads from the pools could be used as "markers" to build the map-scaffolds. The reads from the pools could be assembled as well as the mate-pair reads – these reads were in fact assembled without using the insert size information. Assembling independently the reads from the pool it is possible to produce several pools of contigs. These contigs will represent a fraction of the genome but some of them will be redundant because a given region will be present in many different pools. Because a pool contains only a fraction of the genome it could be possible to take the chance to solve possible repeated regions. In fact, in a given pool a repeated region could be present just once offering the opportunity to solve it. In these cases the resulting contigs could represent unique portion of the genome. The subsequent analysis of the profile distances on these contigs could

Chromosome	Start	End	Contig	Contig length
NG-chr03	1191960	1203172	contig00443	11213
NG-chr03	1203273	1209055	contig00780	5783
NG-chr03	1209156	1209451	contig04544	296
NG-chr03	1209552	1210060	contig03336	509
NG-chr03	1210161	1211002	contig02472	842
NG-chr03	1211103	1219215	contig00575	8113
NG-chr03	1219316	1220149	contig02484	834
NG-chr03	1220250	1225316	contig00851	5067
NG-chr03	1225417	1225739	contig04319	323
NG-chr03	1225840	1233777	contig00585	7938
NG-chr03	1233878	1235713	contig01720	1836
NG-chr03	1235814	1236548	contig02680	735
NG-chr03	1236649	1238166	contig01852	1518
NG-chr03	1238267	1251699	contig00373	13433

Table 3.15: Genomic region for chromosome 3 in the final assembly. The 100 bases of difference between the end of a contig and the begin of the next one for convenience are N so there is no other contigs between them in the assembly.

make much more easy to map them on the genome even if they have repeat.

Chapter 4

Conclusions

With the advent of next generation sequencing technologies, physical maps were somewhat neglected in favor of faster and cheaper whole-genome shotgun projects. Somehow, the improvements in sequencing technology does not stimulate improvements in physical mapping methods. However, they still remain an extremely useful method to produce high quality and complete genome sequences. The work presented in this Thesis is proposed as a new method aimed to produce physical maps of genomes taking advantage of next generation sequencing technology.

The rationale of the project is the creation of profiles of presence and absence for a set of genetic markers. To produce these profiles the method relies on the sequencing of several genome fractions. These fractions of the genome are created by pooling together a given number of BAC clones in order that the sum of their average insert size represents the desired fraction of the genome. These BACs are chosen randomly from a BAC library that should presents same peculiar characteristics: (i) it has to be produced by a random fragmentation of the genome, (ii) the average insert size should be around 100 kbp and (iii) its genome coverage should be higher enough to ensure the presence of a given portion of the genome in many different BAC clones, for instance a 30× library.

Once the sequencing of the BAC pools is completed it is possible to produce

the profiles of presence and absence for the desired genetic markers. These profiles are obtained by aligning the reads coming from the sequencing of the pools on the genetic markers. The working hypothesis of this method is that by looking at these profiles it is possible to estimate the distances of the genetic markers or at least their positions on the genome.

Two approaches for sequencing the BAC pools were developed: one based on shotgun sequencing and a second one based on sequencing endonuclease digested sites. Both the methods were confirmed to be viable and are proposed as complementary strategies for genome mapping respectively in small and large genome.

Nannochloropsis gaditana was chosen as test for this mapping method. The genome of this unicellular algae was believed to be the proper size to permit both the development and the application of this mapping method. Thus, sixty-four BAC DNA pools were shotgun sequenced in order to bild the genome map. The parallel and independent project for sequencing the genome of *Nannochloropsis gaditana* represented a good opportunity to perform the comparison of this mapping method with a standard sequencing approach.

The results shown in this Thesis suggest that the proposed method could be a viable strategy to produce genome maps with next generation sequencing. The initial assumptions, at the basis of the method were confirmed. (I) With the BAC pooling procedure many fractions of the genome could be created. (II) By sequencing these BAC DNA pools it is possible to obtain profiles of presence and absence of desired genetic markers. (III) These profiles are expressed in terms of reads aligning on given target sequences. (IV) The test on *smaltigs*, the little virtual fragments obtained from the contigs of the draft assembly, clearly showed that the profiles of presence and absence are similar for regions that lie one next to the other on the genome.

During this work it was also developed a scoring system aimed to compare these profiles of presence and absence. The developed scoring matrix is based on the observed profiles and expresses the probability to see a given difference between read counts in two near DNA fragments. The profiles of read counts of all the *smaltigs* was compared according to this scoring matrix and the positive scores were analyzed. The positive scores actually gave indication about the physical proximity of the compared *smaltigs*.

Using these scores it was developed a preliminary mapping procedure to place *smaltigs* on scaffold-like maps. The map-scaffolds obtained with this method were confirmed by the independent assembly produced during the *N. gaditana* genome sequencing project. With the same scoring system and the

same mapping procedure it was possible to place contigs from an independent assembly (the one obtained with short reads assembly on mate-pair sequences) in a comparable order. However some regions remain unsolved in both these "maps". But it is here demonstrated the effectiveness of the scoring system and the mapping procedure.

The whole results shown in this Thesis are very promising and suggest that the method could actually produce a good genome map. However, some aspects should be improved in order to achieve a better system. The scoring system will move to data simulation instead of considering observed versus expected scores. The genetic markers that are profiled will move from *smaltigs* to unique sequences obtained directly from the sequencing of the BAC pools. The mapping procedure, now based on the global scaffold approach will likely move to a mixed approach between global scaffold and mini-scaffold approach, in order to propose more strong connections.

Some efforts will be spent in order to move away form the need of a BAC library to produce fraction of the genome. This aspect is very important because the BAC library requires time and money to be produced and processed. Some new ideas will be pursuit to overcome this limitation. A possible strategy will be the production of gel slices from pulsed filed gel electrophoresis of the entire genome. Another strategy could be the implementation of strategies similar to exome capturing to select portions of the genome of interest.

However it has to be pointed out that the method proposed here is intended to be a genome mapping method. It is not proposed as an alternative approach to sequence genomes but as a complementary strategy to classical sequencing project in the aim of obtaining high quality final genome sequences.

Chapter 5

Supplementary Informations

5.1 Reads alignment

Table 5.1 reports the results of the alignment of the pools against the draft assembly of the algae genome, the E. coli reference genome and the vector sequence. The pools were sequenced in four different sequencing reactions. Each pool was produced pooling together DNA BAC clones from a single 96-well plate.

5.2 Trial assembly of mate pair reads

The mate pair reads were assembled with Velvet [42]. SOLiD™ mate-pair sequencing has a peculiar chemistry: the two reads are sequenced from the same DNA strand and so they have the same orientation. On contrary, Velvet

Table 5.1: In this table are shown the results of the alignment of each pool on the complete draft assembly of *N. gaditana*. Seq. = sequences. In column "Aligned seq." in parenthesis are indicated the percentages of aligned reads in respect to the total number of produced reads. In columns *E. coli* seq., BAC seq. and *N. gaditana* seq., values in parenthesis indicate percenteges of aligned reads in respect to total aligned reads for that pool.

Pool	Produced seq.	Aligned seq.	*E. coli* seq.	BAC seq.	*N. gaditana* seq.
1	5,153,404	3,765,057 (73.06)	100,808 (2.68)	498,228 (13.23)	3,166,021 (84.09)
2	6,060,838	4,247,473 (70.08)	60,926 (1.43)	519,566 (12.23)	3,666,981 (86.33)
3	6,396,583	4,574,972 (71.52)	62,937 (1.38)	637,353 (13.93)	3,874,682 (84.69)
4	5,866,065	4,440,241 (75.69)	36,193 (0.82)	553,182 (12.46)	3,850,866 (86.73)
5	5,686,692	4,219,313 (74.2)	28,893 (0.68)	414,790 (9.83)	3,775,630 (89.48)
6	4,053,779	2,525,155 (62.29)	18,275 (0.72)	313,156 (12.4)	2,193,724 (86.87)
7	7,022,680	4,807,037 (68.45)	31,714 (0.66)	490,327 (10.2)	4,284,996 (89.14)
8	5,589,364	4,093,616 (73.24)	28,987 (0.71)	384,650 (9.4)	3,679,979 (89.9)
9	8,348,319	3,587,527 (42.97)	42,221 (1.18)	376,545 (10.5)	3,168,761 (88.33)
10	6,788,846	4,622,658 (68.09)	48,843 (1.06)	398,787 (8.63)	4,175,028 (90.32)
11	6,685,681	3,786,205 (56.63)	34,062 (0.9)	513,195 (13.55)	3,238,948 (85.55)
12	6,249,073	1,901,463 (30.43)	19,229 (1.01)	339,245 (17.84)	1,542,989 (81.15)
13	6,368,872	4,476,860 (70.29)	40,419 (0.9)	427,944 (9.56)	4,008,497 (89.54)
14	5,551,909	3,669,192 (66.09)	25,024 (0.68)	344,782 (9.4)	3,299,386 (89.92)
15	7,303,934	3,898,118 (53.37)	22,574 (0.58)	353,381 (9.07)	3,522,163 (90.36)
16	6,426,202	4,897,680 (76.21)	37,827 (0.77)	461,738 (9.43)	4,398,115 (89.8)
17	12,273,325	8,445,791 (68.81)	63,153 (0.75)	801,498 (9.49)	7,581,140 (89.76)
18	7,128,011	3,928,494 (55.11)	16,418 (0.42)	427,057 (10.87)	3,485,019 (88.71)
19	9,053,379	4,750,915 (52.48)	13,098 (0.28)	526,357 (11.08)	4,211,460 (88.65)
20	10,551,048	7,667,649 (72.67)	72,137 (0.94)	774,723 (10.1)	6,820,789 (88.96)
21	6,250,758	4,089,226 (65.42)	18,865 (0.46)	394,769 (9.65)	3,675,592 (89.88)
22	4,940,590	3,108,390 (62.92)	13,215 (0.43)	383,582 (12.34)	2,711,593 (87.23)
23	6,789,006	4,754,966 (70.04)	27,479 (0.58)	456,574 (9.6)	4,270,913 (89.82)
24	5,643,770	3,884,112 (68.82)	17,155 (0.44)	337,088 (8.68)	3,529,869 (90.88)
25	6,052,821	4,496,820 (74.29)	10,044 (0.22)	468,235 (10.41)	4,018,541 (89.36)
26	5,895,847	4,227,774 (71.71)	11,464 (0.27)	466,389 (11.03)	3,749,921 (88.7)
27	11,066,515	7,535,374 (68.09)	24,247 (0.32)	816,888 (10.84)	6,694,239 (88.84)
28	5,427,068	3,654,671 (67.34)	18,572 (0.51)	425,858 (11.65)	3,210,241 (87.84)
29	5,891,135	4,110,006 (69.77)	47,234 (1.15)	429,900 (10.46)	3,632,872 (88.39)
30	5,879,919	3,911,625 (66.53)	22,245 (0.57)	418,374 (10.7)	3,471,006 (88.74)
31	4,509,038	3,272,334 (72.57)	22,960 (0.7)	362,117 (11.07)	2,887,257 (88.23)
32	6,314,098	4,299,016 (68.09)	34,316 (0.8)	467,878 (10.88)	3,796,822 (88.32)
33	5,982,633	3,812,537 (63.73)	29,373 (0.77)	337,440 (8.85)	3,445,724 (90.38)
34	6,437,061	4,588,472 (71.28)	25,561 (0.56)	365,903 (7.97)	4,197,008 (91.47)
35	5,710,940	3,304,680 (57.87)	75,295 (2.28)	308,690 (9.34)	2,920,695 (88.38)
36	6,736,499	4,729,285 (70.2)	18,580 (0.39)	398,075 (8.42)	4,312,630 (91.19)
37	8,181,057	5,522,853 (67.51)	29,088 (0.53)	464,638 (8.41)	5,029,127 (91.06)
38	7,109,679	5,033,218 (70.79)	29,846 (0.59)	423,295 (8.41)	4,580,077 (91)
39	6,457,548	4,492,918 (69.58)	22,929 (0.51)	418,454 (9.31)	4,051,535 (90.18)
40	7,693,833	5,105,335 (66.36)	19,174 (0.38)	440,753 (8.63)	4,645,408 (90.99)
41	6,000,501	4,191,534 (69.85)	10,746 (0.26)	380,815 (9.09)	3,799,973 (90.66)
42	6,827,101	4,679,166 (68.54)	11,154 (0.24)	403,677 (8.63)	4,264,335 (91.13)
43	7,357,093	5,081,943 (69.08)	14,215 (0.28)	408,877 (8.05)	4,658,851 (91.67)
44	7,520,440	5,414,103 (71.99)	7,282 (0.13)	458,991 (8.48)	4,947,830 (91.39)
45	10,005,254	6,776,957 (67.73)	142,671 (2.11)	551,610 (8.14)	6,082,676 (89.76)
46	7,547,618	4,983,573 (66.03)	77,819 (1.56)	367,372 (7.37)	4,538,382 (91.07)
47	8,750,735	6,005,060 (68.62)	103,259 (1.72)	428,204 (7.13)	5,473,597 (91.15)
48	8,946,523	5,537,477 (61.9)	62,465 (1.13)	467,735 (8.45)	5,007,277 (90.43)
49	7,275,643	4,807,421 (66.08)	29,161 (0.61)	397,046 (8.26)	4,381,214 (91.13)
50	7,514,601	5,109,649 (68)	31,469 (0.62)	382,631 (7.49)	4,695,549 (91.9)
51	6,778,551	4,588,525 (67.69)	32,726 (0.71)	362,615 (7.9)	4,193,184 (91.38)
52	6,564,339	4,510,053 (68.71)	26,092 (0.58)	418,437 (9.28)	4,065,524 (90.14)
53	7,311,455	4,625,215 (63.26)	66,051 (1.43)	448,565 (9.7)	4,110,599 (88.87)
54	6,808,873	4,666,225 (68.53)	32,829 (0.7)	367,480 (7.88)	4,265,916 (91.42)
55	6,713,417	4,367,494 (65.06)	45,011 (1.03)	481,321 (11.02)	3,841,162 (87.95)
56	5,991,530	3,873,174 (64.64)	24,719 (0.64)	336,566 (8.69)	3,511,889 (90.67)
57	6,365,897	4,074,486 (64)	17,850 (0.44)	313,972 (7.71)	3,742,664 (91.86)
58	6,993,851	4,606,441 (65.86)	15,735 (0.34)	399,999 (8.68)	4,190,707 (90.97)
59	6,588,937	4,399,183 (66.77)	15,444 (0.35)	377,500 (8.58)	4,006,239 (91.07)
60	6,455,427	4,146,757 (64.24)	13,784 (0.33)	351,190 (8.47)	3,781,783 (91.2)
61	7,139,646	4,672,784 (65.45)	140,035 (3)	415,955 (8.9)	4,116,794 (88.1)
62	6,474,591	4,410,644 (68.12)	50,447 (1.14)	353,317 (8.01)	4,006,880 (90.85)
63	6,836,270	4,249,431 (62.16)	70,912 (1.67)	435,682 (10.25)	3,742,837 (88.08)
64	6,847,472	4,419,327 (64.54)	54,540 (1.23)	360,586 (8.16)	4,004,201 (90.61)

Table 5.1: Caption for this table is on page...

requires paired-end reads that came from opposite strands and face each other. Moreover, SOLiD™ produces reads in color space and reversing and translating one read is not an easy task.

For these resons, and given the very high coverage of the mate-pair libraries, it was decided to start using these reads as shotgun fragment reads. To evaluate the best parameters to perform the assembly several test was performed on a subset of these reads.

The parameters considered were: the k-mer length and the trimming of the reads in 3'. The k-mer is the "nucleotide word" with which the graph is constructed. The k-mer sizes considered were odd[1] values strating from 21 to 33. The size of the k-mer should be lower than the size of the reads in order to allow the bilding of the graph. Otherwise there will be as many different k-mers as the total number of reads making impossible to find a path between reads.

The trimming at the 3' end of the read is due to the quality values drops at the end of the read. Mate-pair reads were 50 bp long so the tests was performed with 0, 5, 10 and 15 bases removed at the 3'.

Figure 5.1 summarize the results of this test assembly in terms of total bases present in the assembly, numer of contigs produced and N50.

Looking at number of bases assembled, N50 length and to the number of contigs, the better assembly seemed to be the one with reads trimmed of the last 5 bases and with k-mer size of 21.

To decide the proper parameters for the assembly the N50 size should be as higher as possible and the same is true also for the total number of assembled bases. On the contrary, the number of contigs should be relatively low.

Short reads assembly is a complex task. To verify that the resulting contigs are not chimera and that actually represent portions of the genome, they were aligned using BLAST against the contigs of the drat genome assembly of *N. gaditana*. Results are displayed in figure 5.2.

In this graph it is plotted only the longer alignments for each "velvet contig". The distribution shows that the majority of the produced contigs aligns for their entire length on the reference assembly confirming the goodness of the short reads assembly.

[1] In Velvet the k-mer must be an odd value to avoid confusion with the relative reverse complement [42].

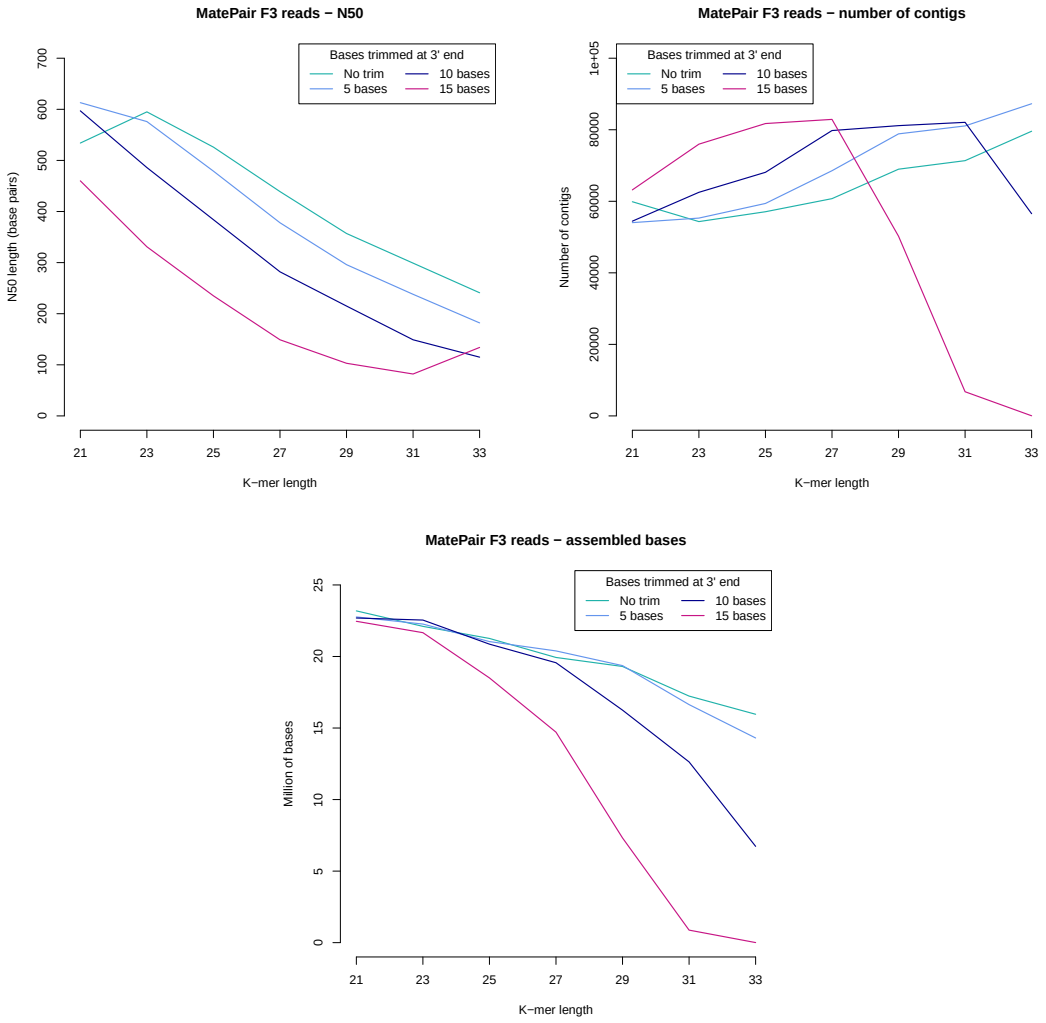


Figure 5.1: Assembled bases, produced contigs and N50

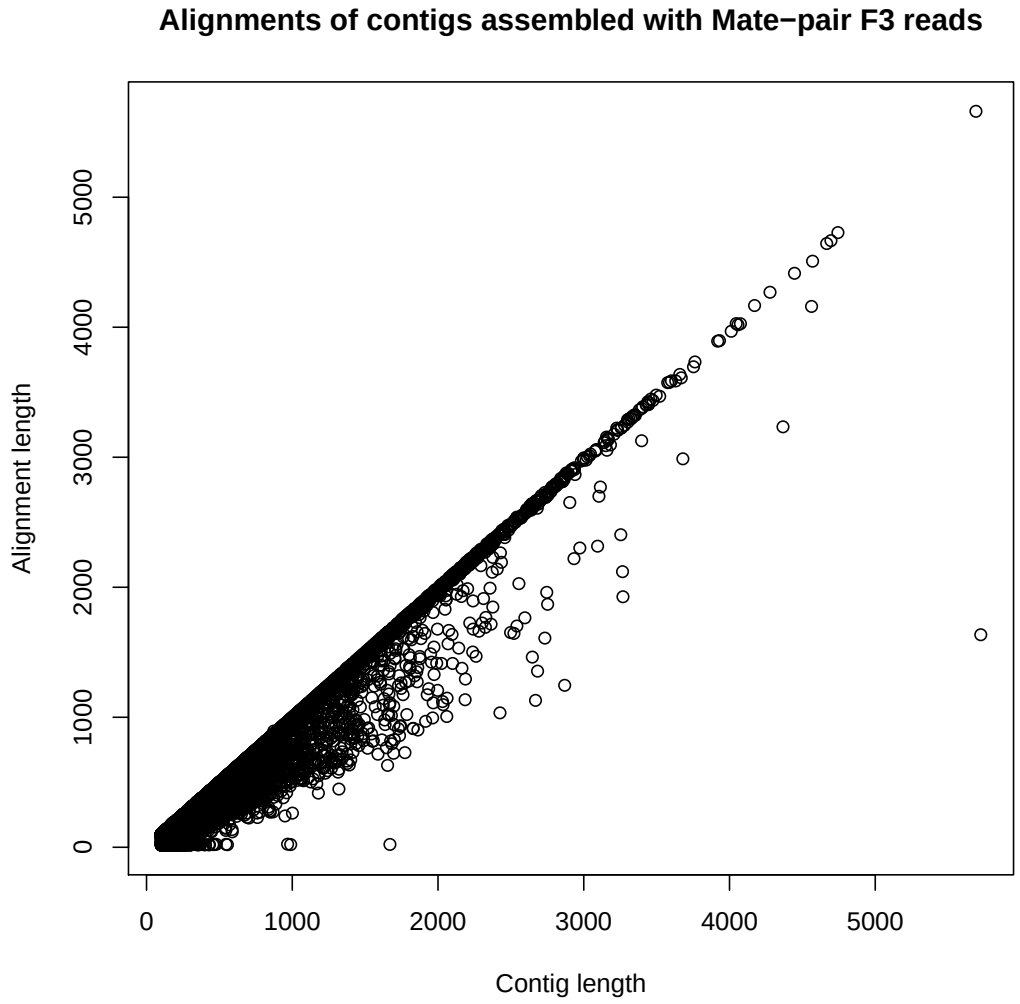


Figure 5.2: The length of the contig is plotted against its maximum length alignment.

Bibliography

[1] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert, and M. Ysebaert. Complete nucleotide sequence of bacteriophage ms2 rna: primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–507, Apr 1976.

[2] F. Sanger, G M Air, B G Barrell, N L Brown, A R Coulson, J C Fiddes, C Hutchison III, and P M Slocombe M Smith. Nucleotide sequence of bacteriophage ϕx174 dna. *Nature*, 265:687–695, Feb 1977.

[3] F Sanger, S Nicklen, and A R Coulson. Dna sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*, 74(12):5463–7, Dec 1977.

[4] R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, and et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd. *Science*, 269(5223):496–512, Jul 1995.

[5] Carol J. Bult, Owen White, Gary J. Olsen, Lixin Zhou, Robert D. Fleischmann, Granger G. Sutton, Judith A. Blake, Lisa M. FitzGerald, Rebecca A. Clayton, Jeannine D. Gocayne, Anthony R. Kerlavage, Brian A. Dougherty, Jean-Francois Tomb, Mark D. Adams, Claudia I. Reich, Ross Overbeek, Ewen F. Kirkness, Keith G. Weinstock, Joseph M.

Merrick, Anna Glodek, John L. Scott, Neil S. M. Geoghagen, Janice F. Weidman, Joyce L. Fuhrmann, Dave Nguyen, Teresa R. Utterback, Jenny M. Kelley, Jeremy D. Peterson, Paul W. Sadow, Michael C. Hanna, Matthew D. Cotton, Kevin M. Roberts, Margaret A. Hurst, Brian P. Kaine, Mark Borodovsky, Hans-Peter Klenk, Claire M. Fraser, Hamilton O. Smith, Carl R. Woese, and J. Craig Venter. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, 273(5278):1058–1073, Aug 1996.

[6] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 274(5287):546–567, Oct 1996.

[7] The C. elegans Sequencing Consortium. Genome sequence of the nematode c. elegans: A platform for investigating biology. *Science*, 282(5396):2012–2018, Dec 1998.

[8] Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M,

Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Sidén-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, and Venter JC. The genome sequence of *Drosophila melanogaster*. *Science*, 287(5461):2185–95, Mar 2000.

[9] Frederick R. Blattner, Guy Plunkett III, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau, and Ying Shao. The complete genome sequence of *Escherichia coli* k-12. *Science*, 277(5331):1453–1462, Sep 1997.

[10] J F Heidelberg, J A Eisen, W C Nelson, R A Clayton, M L Gwinn, R J Dodson, D H Haft, E K Hickey, J D Peterson, L Umayam, S R Gill, K E Nelson, T D Read, H Tettelin, D Richardson, M D Ermolaeva, J Vamathevan, S Bass, H Qin, I Dragoi, P Sellers, L McDonald, T Utterback, R D Fleishmann, W C Nierman, O White, S L Salzberg, H O Smith, R R Colwell, J J Mekalanos, J C Venter, and C M Fraser. Dna sequence of both chromosomes of the cholera pathogen vibrio cholerae. *Nature*, 406(6795):477–83, Aug 2000.

[11] Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815, Dec 2000.

[12] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, N Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French,

D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowski, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[13] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel,

S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. The sequence of the human genome. *Science*, 291(5507):1304–51, Feb 2001.

[14] Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman RD, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison CA 3rd, and Venter JC. The minimal gene complement of *Mycoplasma genitalium*. *Science*, 270(5235):397–403, Oct 1995.

[15] J L Weber and E W Myers. Human whole-genome shotgun sequencing. *Genome Research*, 7(5):401–9, May 1997.

[16] M Boguski, A Chakravarti, R Gibbs, E Green, and R. M Myers. The end of the beginning: The race to begin human genome sequencing. *Genome Research*, 6(9):771–772, Sep 1996.

[17] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, Oct 2004.

[18] Robert H Waterston, Eric S Lander, and John E Sulston. On the sequencing of the human genome. *Proc Natl Acad Sci USA*, 99(6):3712–6, Mar 2002.

[19] Robert H Waterston, Eric S Lander, and John E Sulston. More on the sequencing of the human genome. *Proc Natl Acad Sci USA*, 100(6):3022–4; author reply 3025–6, Mar 2003.

[20] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–62, Dec 2002.

[21] Nature publishing group. Human genome at ten - the sequence explosion. *Nature*, 464(7289):670–671, Apr 2010.

[22] KA Wetterstrand. Dna sequencing costs: Data from the nhgri genome sequencing program (gsp). *Available at: www.genome.gov/sequencingcosts. Accessed on 4 Jan 2013*, -(-):–, - -.

[23] P A Pevzner, H Tang, and M S Waterman. An eulerian path approach to dna fragment assembly. *Proc Natl Acad Sci USA*, 98(17):9748–53, Aug 2001.

[24] M. C Schatz, A. L Delcher, and S. L Salzberg. Assembly of large genomes using second-generation sequencing. *Genome Research*, 20(9):1165–1173, Sep 2010.

[25] R Li, H Zhu, J Ruan, W Qian, X Fang, Z Shi, Y Li, S Li, G Shan, K Kristiansen, S Li, H Yang, J Wang, and J Wang. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20(2):265–272, Feb 2010.

[26] Can Alkan, Saba Sajjadian, and Evan E Eichler. Limitations of next-generation genome sequence assembly. *Nature Methods*, 8(1):61–65, Jan 2011.

[27] H. A Lewin, D. M Larkin, J Pontius, and S. J O'brien. Every genome sequence needs a good map. *Genome Research*, 19(11):1925–1928, Nov 2009.

[28] J Van Oeveren, M De Ruiter, T Jesse, H Van Der Poel, J Tang, F Yalcin, A Janssen, H Volpin, K. E Stormo, R Bogden, M. J. T Van Eijk, and M Prins. Sequence-based physical mapping of complex genomes by whole genome profiling. *Genome Research*, 21(4):618–625, Apr 2011.

[29] The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400):635–641, May 2012.

[30] Romain Philippe, Frédéric Choulet, Etienne Paux, Jan Van Oeveren, Jifeng Tang, Alexander H J Wittenberg, Antoine Janssen, Michiel J T van Eijk, Keith Stormo, Adriana Alberti, Patrick Wincker, Eduard Akhunov, Edwin Van Der Vossen, and Catherine Feuillet. Whole genome profiling provides a robust framework for physical mapping and sequencing in the highly complex and repetitive wheat genome. *BMC Genomics*, 13:47, Jan 2012.

[31] D R Cox, M Burmeister, E R Price, S Kim, and R M Myers. Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science*, 250(4978):245–50, Oct 1990.

[32] P H Dear and P R Cook. Happy mapping: linkage mapping using a physical analogue of meiosis. *Nucleic Acids Research*, 21(1):13–20, Jan 1993.

[33] Author not reported. Classical linkage mapping. *Los Alamos Science*, 20(20):86–93, - 1992.

[34] M. R Miller, J. P Dunham, A. Amores, W. A Cresko, and E. A Johnson. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated dna (rad) markers. *Genome Research*, 17(2):240–8, Feb 2007.

[35] N. A Baird, P. D Etter, T. S Atwood, M. C Currey, A. L Shiver, Z. A Lewis, E. U Selker, W. A Cresko, and E. A Johnson. Rapid snp discovery and genetic mapping using sequenced rad markers. *PLoS ONE*, 3(10):e3376, Oct 2008.

[36] P. A Hohenlohe, S. Bassham, P. D Etter, N. Stiffler, E. A Johnson, and W. A Cresko. Population genomics of parallel adaptation in threespine stickleback using sequenced rad tags. *PLoS Genet*, 6(2):e1000862, Feb 2010.

[37] E Meyer, J K Mckay, S Wang, and M V Matz. 2b-rad: a simple and flexible method for genome-wide genotyping. *Nature Methods*, pages 1–5, May 2012.

[38] R J Elshire, J C Glaubitz, Q Sun, J A Poland, K Kawamoto, E S Buckler, and S E Mitchell. A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. *PLoS ONE*, 6(5):e19379, Jan 2011.

[39] E. Corteggiani Carpinelli, A. Telatin, N. Vitulo, C. Forcato, M. D'Angelo, R. Schiavon, A. Vezzi, G.M. Giacometti, T. Morosinotto, and G. Valle. Chromosome scale genome assembly and transcriptome profiling of *Nannochloropsis gaditana* in nitrogen depletion. *in Pubblication*, -(-):–, - 2013.

[40] Robert R Klein, Daryl T Morishige, Patricia E Klein, Jianmin Dong, and John E Mullet. High throughput bac dna isolation for physical map construction of sorghum, sorghum bicolor. *Plant Molecular Biology Reporter*, 16:351–364, 1998.

[41] D. Campagna, A. Albiero, A. Bilardi, E. Caniato, C. Forcato, S. Manavski, N. Vitulo, and G. Valle. Pass: a program to align short sequences. *Bioinformatics*, 25(7):967–8, Apr 2009.

[42] D. R Zerbino and E Birney. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 18(5):821–829, Feb 2008.

[43] Randor Radakovits, Robert E Jinkerson, Susan I Fuerstenberg, Hongseok Tae, Robert E Settlage, Jeffrey L Boore, and Matthew C Posewitz. Draft genome sequence and genetic transformation of the oleaginous alga *Nannochloropis gaditana*. *Nat Comms*, 3:686, Feb 2012.

[44] Astrid Vieler, Guangxi Wu, Chia-Hong Tsai, Blair Bullard, Adam J Cornish, Christopher Harvey, Ida-Barbara Reca, Chelsea Thornburg, Rujira Achawanantakun, Christopher J Buehl, Michael S Campbell, David Cavalier, Kevin L Childs, Teresa J Clark, Rahul Deshpande, Erika Erickson, Ann Armenia Ferguson, Witawas Handee, Que Kong, Xiaobo Li, Bensheng Liu, Steven Lundback, Cheng Peng, Rebecca L Roston, Sanjaya, Jeffrey P Simpson, Allan Terbush, Jaruswan Warakanont, Simone Zäuner, Eva M Farre, Eric L Hegg, Ning Jiang, Min-Hao Kuo, Yan Lu, Krishna K Niyogi, John Ohlrogge, Katherine W Osteryoung, Yair Shachar-Hill, Barbara B Sears, Yanni Sun, Hideki Takahashi, Mark Yandell, Shin-Han Shiu, and Christoph Benning. Genome, functional gene annotation, and nuclear transformation of the heterokont oleaginous alga *Nannochloropsis oceanica* ccmp1779. *PLoS Genet*, 8(11):e1003064, Nov 2012.