

Università degli Studi di Padova

Scuola di Dottorato di Ricerca in Bioscienze e Biotecnologie

Indirizzo Biotecnologie

XXIII Ciclo

**Design and implementation of novel algorithms
to integrate different DNA sequencing technologies
for *de novo* genome assembly: *Nannochloropsis* as a test case**

Direttore della Scuola: Ch.mo Prof. Giuseppe Zanotti

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

Supervisore: Ch.mo Prof. Giorgio Valle

Dottorando: Andrea Telatin

A. A. 2010-2011



*genome assembly is an **endless** puzzle.*

Abstract

The advent of next generation sequencing technologies marked the beginning of a new era in the production of genomic data; nonetheless it also offered novel challenges to the bioinformatics community. While re-sequencing of genomes was made relatively easy and cheap, *de novo* assembly of eukaryotic genomes still presents significant hurdles.

In this thesis we attempted the application of a mixed-techniques approach to the *de novo* sequencing of a small eukaryotic genome, that could allow us to take advantage of both the relatively long reads obtainable using the Roche 454 and the incredibly high coverage of short reads allowed by SOLiD sequencer. The application of a hierarchical approach based on the production of reliable contigs using the 454 and the assembly of these contigs in scaffold using the SOLiD mate pairs, could represent a cost effective strategy to address this important issue.

To realize this project a contig-centered data repository, called 4NGS, was produced that allowed the real time interrogation of partially assembled data as well as the evaluation of the assembly and the design of new experiments. Moreover I designed and implemented a scaffolding algorithm, ScaMP (Scaffolding with Mate Pairs), that uses the SOLiD mate-paired reads aligned to the reference contigs, to produce and store scaffolds in the 4NGS database.

To further improve the assembly results, a gap closure pipeline was also developed that allows joining adjacent contigs using the SOLiD short sequences.

I assessed the performance of both programs using as a test case the genome of a microalga, *Nannochloropsis gaditana*, which received pressing attention from the scientific community for its potential for biofuel production. The genome (that has an estimated size spanning between 30 and 40 Mbp) has been sequenced with a low-coverage 454 (that produced more than 12,000 contigs) and with SOLiD Mate Paired libraries.

Scaffolding performed with my platform produced 95 scaffolds that include 26.8 Mbp of the genome and have an average size of 285,594 bp.

The gap filling pipeline closed more than 3,000 gaps between adjacent contigs, and gave best results for scaffolded regions (the largest scaffold, composed by 140 contigs, had 106 gaps closed raising N50 of its contigs from 8.3 kbp to 77.4 kbp).

My study fulfilled the expectation that for a small eukaryotic genome, *de novo* assembly starting from next generation data alone is feasible, cheap and efficient; that a mixture of SOLiD and 454 sequencing substantially improves the assembly; and that the quality of the resulting genome draft is enough to support further analysis of comparative genomics and to obtain a valuable framework to design the application of recombinant techniques.

A good quality draft of *N. gaditana* genome was produced in this thesis, meeting the need of the scientific community for valuable tools able to boost the application of the new genomics resources to a vast plethora of biological problems and to serve new and interesting biotechnological applications.

Riassunto in italiano

L'avvento e la rapida evoluzione dei sequenziatori di nuova generazione (NGS) ha abbattuto il costo ed il tempo necessario alla produzione dei dati. La fase di assemblaggio di un genoma che porta ad ottenere la corretta sequenza genomica a partire dalle singole sequenze prodotte dai sequenziatori è sempre stato un processo complesso, e l'aumento della mole di dati prodotti non è corrisposto ad una nostra aumentata capacità di analisi degli stessi.

In questa tesi si presenta un approccio misto di sequenziamento che combina i benefici di due sequenziatori di nuova generazione (il 454 di Roche che fornisce le sequenze più lunghe ed il SOLiD di Applied Biosystems che fornisce una massiva produzione di sequenze, ciascuna di lunghezza ridotta) al fine di ottenere le informazioni per il sequenziamento di un genoma.

La strategia è stata testata sul genoma della microalga eucariote *Nannochloropsis gaditana*, un organismo che negli ultimi anni ha ricevuto notevole attenzione dalla comunità scientifica per la sua capacità di immagazzinare energia luminosa sotto forma di acidi grassi (fino al 70% del suo peso). Questa caratteristica rende *Nannochloropsis* un valido candidato per le ricerche su fonti di energie alternative a quelle di origine fossile. La stima della dimensione del suo genoma varia tra i 30 ed i 40 milioni di paia di basi.

Il rapido miglioramento delle tecnologie di sequenziamento non è corrisposto ad una altrettanto rapida evoluzione dei programmi di analisi dei dati, che spesso risultano inadeguati a gestire la nuova mole di dati o a sfruttarne le potenzialità.

Per questo ho deciso di progettare ed implementare una collezione di programmi per l'assemblaggio di genomi con dati misti (SOLiD e 454).

Le sequenze ottenute da un sequenziamento di tipo shotgun con il 454 vengono assemblate per produrre un insieme di porzioni genomiche

denominate *contig*. Per il genoma di *Nannochloropsis* ne sono stati prodotti 7 035 di dimensioni superiori alle 500 paia di basi.

Sfruttando le informazioni delle librerie “mate-paired” del SOLiD, che prevedono il sequenziamento combinato di paia di sequenze ad una distanza nota nel genoma ho sviluppato un programma (ScaMP) che permette di produrre liste ordinate di *contig* (dette *scaffold*).

Il programma ha prodotto 95 *scaffold* di dimensione media pari a 285 594 paia di basi, incorporandovici 26,8 milioni di nucleotide in totale.

L’elevato numero di sequenze prodotte con il SOLiD permette anche, una volta ottenuti gli *scaffold*, di completare le sequenze mancanti fra un *contig* ed il successivo (dette *gap*). A tal fine ho sviluppato un ulteriore programma che estrae dall’insieme di sequenze SOLiD il sottoinsieme di quelle adiacenti ad un *contig*, ed effettua un assemblaggio locale che viene infine utilizzato per colmare *gap*. Su uno *scaffold* di 140 *contig* ha eliminato 106 regioni *gap*, portando il numero di *contig* a 36 ed aumentando la dimensione media da 8 300 a 77 400 paia di basi.

I risultati ottenuti confermano che l’approccio combinato di SOLiD e 454 permette di ottenere un buon assemblaggio di un genoma eucariotico limitando al contempo i costi di sequenziamento.

I risultati ottenuti sono stati validati tramite il sequenziamento di estremità di inserti BAC successivamente allineati contro il dataset di *scaffold*. I programmi sviluppati hanno dimostrato di essere un valido sistema di assemblaggio affidabile e di colmare una lacuna nel panorama dei programmi bioinformatici per il sequenziamento de novo con tecniche di nuova generazione.

List of abbreviations

| | |
|--------------------|--|
| B (kB, MB, GB, TB) | Byte (kilo-, Mega-, Giga-, Tera-) |
| BAC | Bacterial Artificial Chromosome, cloning vector for large inserts |
| bp (kbp, Mbp, Gbp) | Base pair (kilo-, Mega-, Giga-) |
| CCD | Charge-coupled device (electronic sensor for digital imaging) |
| Chl | Chloroplast |
| dNTP | Deoxy-Nucleoside Tri-Phosphate |
| emPCR | Emulsion PCR |
| EST | Expressed Sequence Tag |
| FOSS | Free and Open Source Software |
| gDNA | Genomic (nuclear) DNA |
| INDEL | Insertion/Deletion. «A collective abbreviation to describe relative gain or loss of a segment of one or more nucleotides in a genomic sequence...typically used to denote relatively small-scale variants» — from Scherer <i>et al.</i> 2007 |
| IR | Inverted Repeat |
| MP | Mate-Pairs, specifically referring to SOLiD v3 Long Mate-Paired Libraries (Applied Biosystems) |
| mtDNA | Mitochondrial DNA |
| N50 | «Given a set of sequences the N50 length is defined as the length N for which 50% of all bases in the sequences are in a sequence of length $L < N$ » — from Miller <i>et al.</i> 2010 |
| NGS | Next-Generation Sequencing |
| OS | Operating System |
| PCR | Polymerase Chain Reaction |
| polyA+ | mRNA preparation performed polyadenylated transcript enrichment |
| SNP | Single Nucleotide Polymorphism |
| WGS | Whole Genome Shotgun |

Table of Contents

| | |
|--|-----------|
| Abstract..... | v |
| Riassunto in italiano | vii |
| List of abbreviations | ix |
| 1 Introduction..... | 1 |
| 1.1 Genome sequencing | 1 |
| 1.1.1 Shotgun assembly | 3 |
| 1.1.2 Scaffolding: ordering contigs | 4 |
| 1.1.3 Opportunities and challenges from technical advances..... | 5 |
| 1.2 Advent of “Next-Generation Sequencing” | 6 |
| 1.2.1 Emulsion PCR | 7 |
| 1.2.2 Roche 454: pyrosequencing..... | 8 |
| 1.2.3 Applied Biosystems SOLiD: sequencing by ligation | 8 |
| 1.2.4 SOLiD Mate-Paired libraries | 10 |
| 1.2.5 Genome sequencing with NGS technologies: benefits and issues | 10 |
| 1.3 A mixed approach for genome sequencing | 12 |
| 1.4 <i>N. gaditana</i> genome project..... | 15 |
| 1.4.1 <i>N. gaditana</i> samples for DNA and RNA-Seq | 16 |
| 2 Material and Methods..... | 19 |
| 2.1 Biological sample preparation..... | 19 |
| 2.2 Hardware and OS | 19 |
| 2.3 Interpreters and web servers..... | 20 |
| 2.4 Bioinformatic packages..... | 20 |
| 2.4.1 PASS v.1.65 and PASS-Pair | 20 |
| 2.4.2 Newbler 2.5.3 (January 2011) | 21 |
| 2.4.3 Velvet 1.2.01 | 21 |
| 2.4.4 Other packages used..... | 21 |
| 2.5 Custom tools: technical specifications..... | 22 |
| 2.5.1 SOLiD mate-paired reads analysis..... | 22 |
| 2.5.2 4NGS platform | 24 |
| 2.5.3 Gap closure pipeline..... | 24 |
| 2.5.4 RNA-Seq tracks and chloroplast map | 25 |

| | | |
|----------|--|-----------|
| 3 | Results and Discussion | 27 |
| 3.1 | Sequencing data for the genome of <i>N. gaditana</i> | 28 |
| 3.1.1 | 454 whole genome shotgun and Newbler assembly..... | 28 |
| 3.1.2 | SOLiD mate-paired libraries..... | 31 |
| 3.1.3 | BAC-ends..... | 32 |
| 3.2 | New bioinformatics tools..... | 33 |
| 3.2.1 | 4NGS: a user-friendly data repository | 33 |
| 3.2.2 | ScaMP: a tool for automatic scaffolding | 34 |
| 3.2.3 | BAC-Validate: scaffold validation and super-scaffolding..... | 37 |
| 3.2.4 | Manual finishing assistant | 38 |
| 3.2.5 | PatchGap: a pipeline for gap-closure via local assemblies..... | 39 |
| 3.3 | <i>N. gaditana</i> genome scaffolding | 41 |
| 3.3.1 | ScaMP testing with selected seeds..... | 41 |
| 3.3.2 | <i>N. gaditana</i> genome scaffolding | 42 |
| 3.3.3 | BAC-ends for scaffolds validation and superscaffolding..... | 43 |
| 3.3.4 | Gap closure results..... | 45 |
| 3.4 | Chloroplast genome of <i>N. gaditana</i> | 47 |
| 3.5 | Wheat: an independent test set..... | 49 |
| 4 | Conclusion | 51 |
| 5 | Bibliography | 55 |
| 6 | Supplementary material | 59 |

1 Introduction

1.1 Genome sequencing

The importance of having the complete genome sequence of an organism became evident long before that DNA sequencing technology could sustain the amount of work required for this kind of projects. The Human Genome Project[1] itself was proposed by several leading scientists (among them J. D. Watson and R. Dulbecco) in the '80s, when no automatic sequencing machine was available, and started in 1990 with a colossal roadmap ahead and the involvement of several laboratories from all over the world that were undertaking a decades long project.

DNA sequencing – no matter which technology is used – allows to determinate the correct nucleotide sequence of a limited fraction of a DNA molecule, thus several steps divide the set of sequences produced (referred to as “**reads**”) from the complete genomic sequence.

The “*International Human Genome Sequencing Consortium*” adopted a complex strategy that involved the preparation of several BAC libraries, the use of a physical map to determine the pool of BAC to be sequenced to avoid excessive overlap between them and finally the sequencing of each selected BAC with a shotgun approach (shearing the DNA, sequencing all fragments and finally assembling them). This approach reduces the complexity of assembly but also minimizes the amount of DNA sequencing required that was still limiting at that time, even if compared with the impressive amount of laboratory work needed.

It was clear that a “whole genome shotgun” approach could become a feasible strategy for large genomes only with important advances in both sequencing and bioinformatics technologies.

A strong supporter of this approach has been J. C. Venter who was able to sequence *H. influenzae* (1.8 Mbp) with this approach[2]. Venter became popular for pushing this strategy to the highest level with the human genome sequencing[3] carried on with his company, Celera (see Figure 1).

Venter's company started its Human Genome project several years later than the public consortium and decided to adopt a *whole genome shotgun* approach also thanks to the small advances in DNA automation and the reduction of sequencing[4].

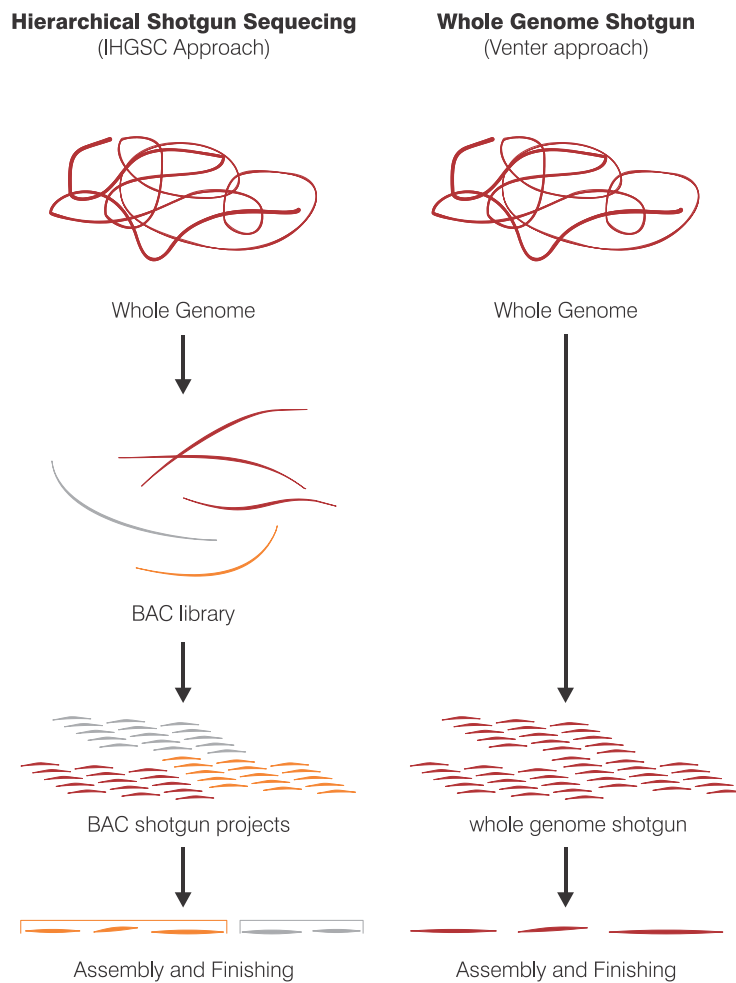


Figure 1

J.C. Venter claimed in 1998 to be able to carry on a «*whole genome shotgun*» for the Human genome, an approach that heavily relies on robust DNA sequencing technology and assembly capabilities.

It should be taken into account, however, that Venter had access to the publicly available data from the IHGS Consortium (while keeping confidential his own data), including physical mapping, thus vanishing his claim to pursue a real WGS strategy.

1.1.1 Shotgun assembly

When sequencing a large piece of DNA with a shotgun approach, the problem is how to rebuild the original sequence starting from the small fragments sequenced (a process called «*de novo* assembly»).

By comparing a sequence with all the other it's possible to find overlapping regions and merge them together in a progressive way (assembly via overlaps, that has been the traditional approach). There are two key aspects in this process: sequencing coverage and sequence repetitiveness.

The sequencing of a G bp long molecule, using an instrument giving reads that are L bp long, can be described using a Poisson distribution model[5].

We call «**sequence coverage**» the quantity $c = NL/G$ (*i. e.* the average number of times a single nucleotide has been sequenced). If we produce N reads such as $NL = G$ (1X sequence coverage) have little probability of having sampled the whole genome at least ones. Using the Poisson distribution the probability that a single nucleotide of the genome was not sequenced is:

$$P_0 = e^{-c} = 0.367879 \quad (63.21\% \text{ of the genome was sampled})$$

On the other hand if we sequence 10-fold the whole genome ($c = 10$):

$$P_0 = e^{-c} = 0.000045 \quad (99.99\% \text{ of the genome was sampled})$$

Thus, theoretically, a 10X sequence coverage should suffice for the complete determination of most of the genome.

A critical aspect in sequence assembly is the presence of **repeated regions** in DNA sequences, and because of this read length acquires importance when trying to assemble DNA reads: longer reads can overcome longer repeated regions. The presence of repeats makes the output of assembly programs being set of contiguous sequences (referred to as «**contigs**») rather than a single sequence. This resulting fragmentation of the genome is a major concern in downstream analysis as gene prediction and genome annotation. The length of contigs depends on sequence coverage and read

length, as well on the structure and complexity of the genome and the number and length of repeated regions (see Figure 2 for a simulation).

Reference sequence (G=76):

THE-QUICK-BROWN-FOX-JUMPED-OVER-THE-LAZY-DOG-THAT-JUMPED-OVER-THE-OLD-ROCK.



Shotgun (L=5, N=114)

THE-QU K-BROW FOX-JU ED-OVE THE-LA -DOG-T T-JUMP D-OVRR IE-OLD-
 HE-QCI -BROWN OX-JUM D-OVER HE-LAZ DOG-TH -JUMPE OVER-T :-OLD-R
 E-QUIC BROWN- X-JUMP -OVER- E-LAZY OG-THE JUMPED OVER-TH :OLD-RC
 -QUICK ROWN-F -JUMPE OVER-T -LAZY- G-THAT UMPED- ER-THE)LD-ROC
 QUICK- AWN-FO JUMPED VER-TH LAZY-D -THAT- MPED-O R-THE- .D-ROCK
 UICK-B WN-FOX UMPED- ER-THE AZY-DO THAT-J PED-OV (-THE-O)-ROCK.
 ICK-BR N-FOX- MPED-O R-THE- ZY-DOG HAT-JU AD-OVE THE-OL
 CK-BRO -FOX-J PED-OV -THE-L Y-DOG- AT-JUM D-OVER EH-OLD
 K-BROW FOX-JU ED-OVE THE-LA -DOG-T T-JUMP -OVER- IE-OLD-
 -BROWN OX-JUM D-OVER HE-LAZ DOG-TH -JUMPE OVER-T -OLD-R
 BROWN- X-JUMP -OVER- E-LAZY OG-THA JUMPED VER-TH OLD-RO



Assembled «contigs»



Figure 2

Sequence assembly. An English sentence is treated as a DNA molecule and exposed to “sequence shotgun” with read length of 5 letters. The presence of a repeated part (“jumped-over-the”), which is longer than the single reads, impairs the whole sentence reconstruction. An assembly program would return 4 “contigs”, one with a doubled coverage being a region repeated twice.

1.1.2 Scaffolding: ordering contigs

A widely adopted strategy to overcome the technical limitations in DNA sequencing is to shear the genome in pieces much larger than the read length, and to sequence them both from the 3’ and from the 5’ (see Figure 3). In the past century this strategy involved the cloning of large DNA fragments into BACs and using universal primers for “BAC-ends sequencing”. Cloning-free approaches are now used to achieve the same result with “next-generation sequencing”.

Aligning the two paired-reads against contigs can help sorting them: if the two sequences match into two different contigs they connect them with a peculiar orientation, forming a virtual bridge between them.

A sorted array of contigs (e.g. joined via paired reads alignment) is called **scaffold**. The regions between contigs are called **gaps** because they are often non-sampled parts of the genome, or parts not included in the assembly. It can happen, however, that the length of a gap is zero.

While «**sequence coverage**» measures how many sequences cover a certain position, when dealing with mate-paired reads or pair ends we can also consider the coverage obtained by the whole fragment that generated the two pairs, that is called «**physical coverage**» (in the example below the word “fox” has 1X sequence coverage, with only one read, and 2X physical coverage).

Large fragments sequenced from both ends:



One scaffold:



Figure 3

Three large fragments of the sentence used in Figure 2 were sequenced from both ends. Mapping the ends to the previously assembled contigs allow for contig order determination, resulting in a single scaffold. Reads are in red, while the physical coverage is represented by a dashed line (gray).

1.1.3 Opportunities and challenges from technical advances

The advent of next-generation sequencing technologies (see next paragraph) has been absolutely beneficial in terms of number of sequences per run, but with a considerable disadvantage in terms of read length, a key factor in *de novo* sequencing.

Very short reads and very high coverage make assembly via sequence overlap detection very difficult: partly because short sequences may have limited overlap with other, but mainly because an impressive all-against-all comparison is required and computationally too hard to be completed with the impressive coverage produced by NGS. A different approach has been

introduced using a mathematical structure called «De Bruijn graph»[6], that reduces the complexity of the input dataset (*i. e.* all the reads) to a set of k -mers generated parsing input reads using a k -long window, and incrementing a counter for each k -mer. A robust implementation of this principle is Velvet[7], and the Ph. D. thesis of its author, D. Zerbino, is a crystal clear reference on the topic[8].

1.2 Advent of “Next-Generation Sequencing”

The first genome projects (*S. cerevisiae*, *H. sapiens*, *A. thaliana*...) were all based on di-deoxynucleotides chain terminating chemistry, proposed by Friedrich Sanger in 1977[9]. This method was greatly improved through the years, from the original version based on radio-labeled bases and manual loading of product on polyacrylamide gels to the final fluorophores-labeled nucleotides and the introduction of automatic capillary electrophoresis, yet the overall throughput was limited by two factors: the need of bacterial cloning to amplify the input material when performing genome shotgun, and the gel-electrophoresis step. State-of-the-art Sanger sequencers could produce sequence as long as 1,000 bp, but with a poor parallelization (96 reactions loaded simultaneously)[4].

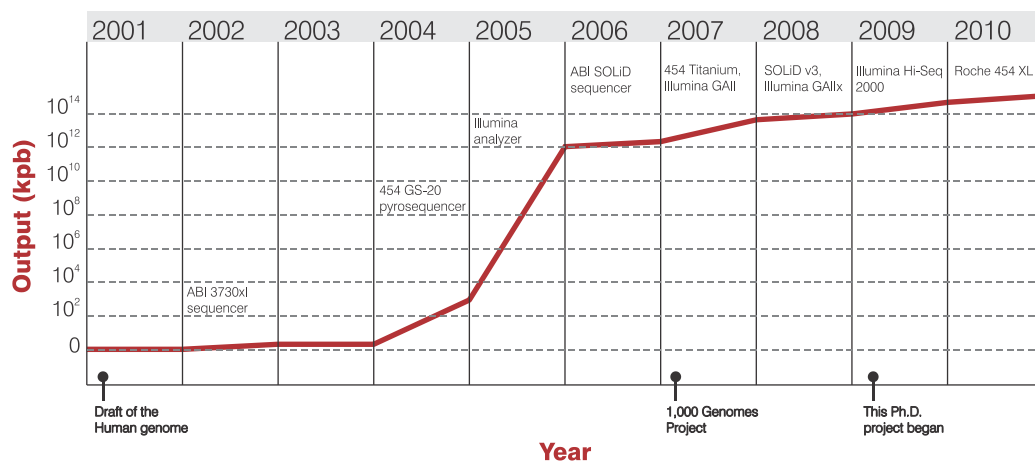


Figure 4

Increase in sequencing output during last decade. It is clear that the advent of NGS machines in 2005 provided an unsurpassed boost in DNA sequencing capacity (adapted from [4]).

In the first years of this century new sequencing methods started to be proposed that avoided both molecular cloning and electrophoresis, and

they are currently referred to as «**next-generation sequencing**»[10] (abbreviated NGS). The first implementations were little or not impressive, in particular for the very short reads produced, but the research to improve them was greatly enhanced in 2006 when the X Foundation offered a prize (10 million dollars) for “*the first team that can build a device and use it to sequence 100 human genomes [...] at a recurring cost of no more than \$10,000 per genome.*”[11].

There are currently three major NGS sequencers available in the market:

- **454 XL by Roche** that currently sequences 1 Gbp in 7 hours, average read length of 700 bp (thus comparable with traditional Sanger sequencing);
- **SOLiD 5500XL by Applied Biosystems** that can provide a higher throughput, 200 Gbp per run, with a maximum read length of 75 bp;
- **Hi-Seq 2000 by Illumina** that sequences 200 Gb per run (25 Gb per day) with each single read 100 bp long.

Each sequencing machine has its advantages and disadvantages, and found a peculiar niche of applications. They have become so popular and sequencing costs are so low to date that each company proposes a “bench-top version” of their machines (e. g. the “454 junior” from Roche) for small-scale sequencing and diagnostics.

This thesis focuses on *de novo* genome sequencing using both the 454 by Roche and the SOLiD by Applied Biosystems, thus it is worthy to briefly introduce their chemistry, and then to highlight the improvements of *de novo* sequencing with NGS (§1.2.5 on page 10).

1.2.1 Emulsion PCR

Both the 454 and the SOLiD make use of a method called *emulsion PCR*[12] (abbreviated emPCR) to avoid cloning into bacterial vectors and allowing for library amplification in a single tube (see Figure 5).

All the DNA fragments to be sequenced (as in the case of sheared genomic DNA) are ligated to two adaptors. An enrichment step discards molecules that have the same adaptor at both ends or no adaptor at all, and finally all the molecules are amplified in an emulsion, prepared to minimize the

chance that two molecules fit in the same aqueous droplet of the emulsion. The aqueous phase contains the reaction mix. A peculiar difference with standard PCR is the use of a primer-coated bead instead of a free primer, making easy to recover the PCR product after emulsion breaking.

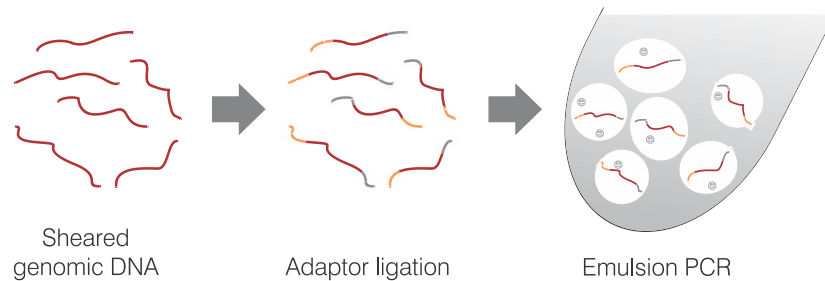


Figure 5

Simplified overview of library amplification using emulsion PCR. Adaptors (in gray and orange) are ligated to end-repaired DNA fragments. Ligation is followed by removal of molecules that ligated the same primer (or no primer) on both ends. Finally the library is added to a PCR reaction mix and emulsified, trying to minimize the event of two templates per water droplet.

1.2.2 Roche 454: pyrosequencing

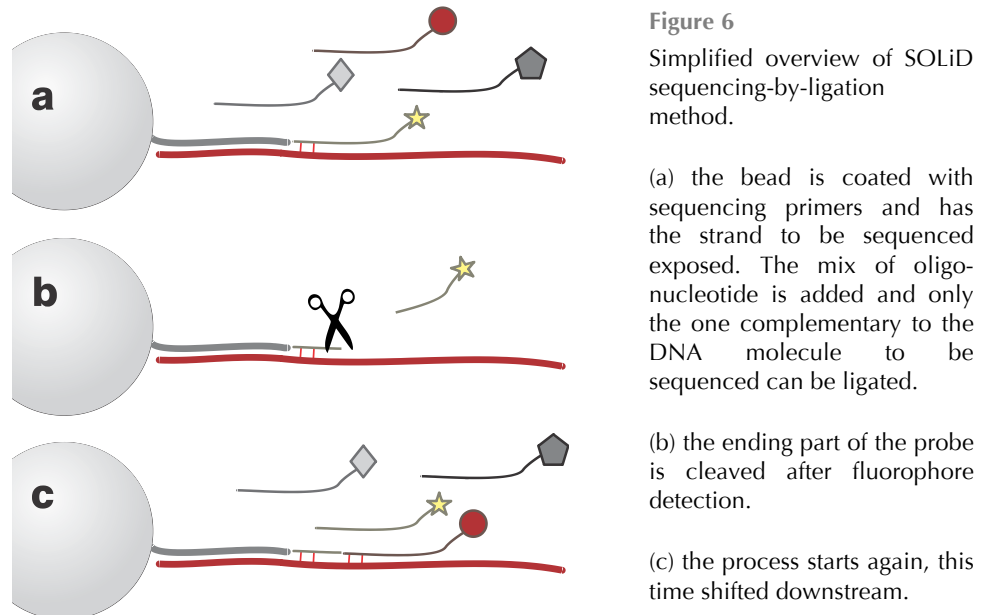
The emulsion PCR is loaded into a sequencing plate, with a bead system that ensures that each well accommodate just a DNA coated bead. Sequencing happens priming the polymerization of a strand, with the use of standard deoxynucleotides that are added one per time. The release of pyrophosphate (P_i) is coupled with light production by luciferase enzyme and thanks to a hi-resolution CCD camera, all the beads of the plate are monitored in real time[13]. When a homopolymeric stretch is found the flash of light is higher, and quantitating the light allow for an approximate detection of the number of subsequent equal bases, yet this lack of accuracy in homopolymeric stretches is one of the weak point of the technology.

Roche 454 provides the longest reads among all NGS machines, and it is a *de facto* standard for genome sequencing. A notable drawback of this solution is the relatively small throughput (just 1 Gb per plate) and the highest price per nucleotide in the market.

1.2.3 Applied Biosystems SOLiD: sequencing by ligation

The SOLiD system uses a completely different sequencing chemistry based on oligonucleotide ligation, using a special mix of oligonucleotides

composed by all possible sequences, having each probe labeled with one out of four fluorophores associated to the first two bases. This means that there are four possible sets of oligonucleotides, characterized by the color of a fluorophore, and each set can start with four different di-nucleotides (see Figure 6).



This probe mix is ligated to the sequencing primer (see Figure 6a) and a color is detected, referring to position 1 and 2 of the DNA molecule. The probe is cleaved and the process repeated (Figure 6b and c), this time probing positions 6 and 7 of the DNA molecule. At the end of the process all the ligated probes are striped away and the sequencing restarted at position -1 (changing the sequencing primer). This means that each single nucleotide is probed twice (from the dinucleotide $n, n+1$ and $n-1, n$).

As for the 454 the sequence detection is in real time, but the sequence of color detected is not directly linked to the DNA sequence as it was in Sanger sequencing. The peculiar sequence encoding (called «color space») adds an extra layer to bioinformatic pipelines that have to deal with it, but at the same time each nucleotide is called twice, enhancing the accuracy and making it easier to discriminate between sequencing errors and SNPs, because the latter involve the change of two colors, not just one, when comparing the sequence with a reference).

Color space encoded reads are easy to align against a reference (converting it into color space), yet still difficult to manipulate because of the complex

rules for color space to base space conversion, especially when dealing with SNPs or INDELs.

1.2.4 SOLiD Mate-Paired libraries

Applied Biosystems provide its own kit for sequencing of large DNA fragments' ends, and they call this approach "*Long Mate-Paired libraries*" (in this thesis referred to as "Mate-Paired" or MP), as they use the "Paired ends" term for another similar approach.

Even if the general principle of MP libraries is the same as BAC-ends sequencing, the protocol is completely different.

SOLiD Long Mate-Paired Library preparation

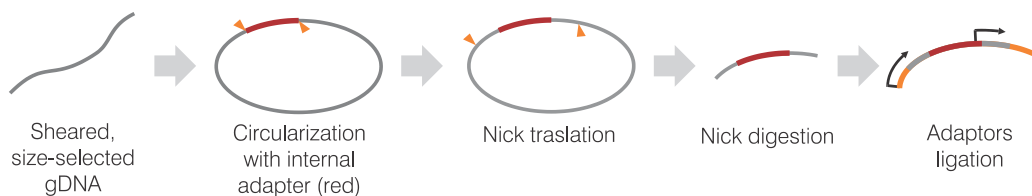


Figure 7

SOLiD Long Mate-Paired library preparation (simplified scheme). Large fragments of DNA after size selection (*gray*) are ligated to an internal (*red*) adaptor and circularized. After nick traslation and digestion of nicked DNA the chimeric fragment carrying the two "mates" is ligated to amplification adaptors. Sequencing primers (*black*) are in the same strand.

Genomic DNA is sheared and end-repaired and large fragments are circularized with an internal adaptor, then a short incubation with *E. coli* polymerase I translates downstream the nicks created with adaptor ligation that are used to break the DNA with T7 exonuclease and S1 nuclease.

Resulting fragments are ligated to adaptors used for the emulsion PCR. One of these adaptors and the internal adaptor are used for sequencing: thus the two mate-paired sequences are in the same strand.

1.2.5 Genome sequencing with NGS technologies: benefits and issues

The unsurpassed throughput obtained with NGS technologies (see Figure 4) has been a major push in genome sequencing. With current technology even single laboratories are enabled to have the genome of their model organism to be sequenced at an affordable price.

Eliminating bacterial libraries and with real-time imaging the whole process can be carried on in a couple of month (see Figure 8).

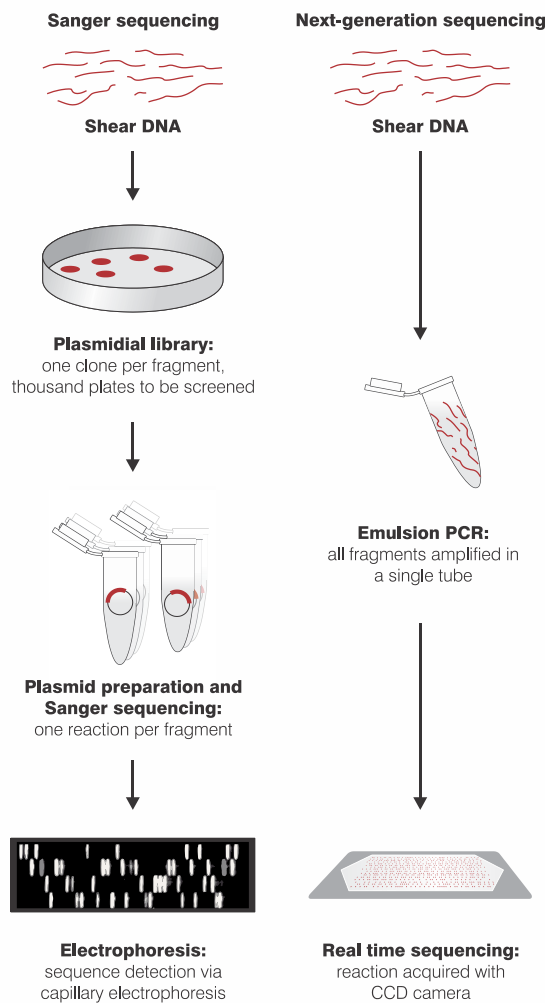


Figure 8

Comparison of whole genome shotgun approach carried on with traditional Sanger approach and Next-Generation Sequencing. Sanger sequencing involves molecular cloning and electrophoretic separation of sequencing products: two time consuming steps removed from NGS approaches.

This significant advance in timing, combined with a massive parallelization gave a major boost to the overall throughput.

As an example, the research group I worked in during my Ph.D. completely sequenced the genome of a deep-sea bacterium (2 chromosomes, 6 Mb total) in 2004, using a traditional Sanger approach[14]. It took more than a year to produce the ~3X coverage and another year for the finishing step. With a single 454 run (two weeks from library preparation to raw data) it would be easy to obtain a much higher coverage (~50X with half plate).

The limiting step to date is the amount of computational power required to handle impressive sequencing outputs and the bioinformatics necessary to make sense of genomic data.

1.3 A mixed approach for genome sequencing

All the NGS machines available have their advantages and disadvantages in terms of total throughput, average read length and running costs. As mentioned before, for *de novo* genome assembly read length plays a pivotal role. This made 454 the ideal instrument for this task, even if its running cost are much higher than those of the competitors.

It should be mentioned that with the information content of mate-paired reads, even Illumina and SOLiD could be competitive because of the higher coverage produced and the much smaller cost per base pair.

I thus propose to combine the benefits of the two platforms both in term of assembly accuracy (the SOLiD being more robust with homopolymeric stretches and SNP detection, the 454 providing longer reads) and in terms of sequencing costs. There is currently a lack of bioinformatic tools able to handle short reads for genome assembly, scaffolding and gap closure.

The aim of this thesis is to fill this gap, designing and implementing novel software tools to assist the whole process.







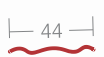

| | Roche 454 | | | AB SOLiD | | |
|------|--|---|---|--|--|---|
| |  | | |  | | |
| |  Throughput |  Read Length |  1X cost |  Throughput |  Read Length |  1X cost |
| 2008 | 500 Mb/run | 350 bp | 1,500 \$ | 30 Gb/run | 50+50bp | 30 \$ |
| 2010 | 1 Gb/run | 600 bp | 800 \$ | 100 Gb/run | 65+65bp | 5 \$ |

Figure 9

Comparison of sequencing costs and output for the two platforms tested in this thesis. The cost for 1X refers to the cost to sequence 35 Mbp, the estimated genome size for the case study of this project, *Nannochloropsis gaditana*. Sequencing technology evolves at a fast rate, thus here I report data available at the begin of the project (2008) and data referred to last data produced for the project (2011).

A low coverage 454 genome shotgun is cost effective in producing a set of contigs that with a mate-paired library sequenced with the SOLiD could become a good quality draft of the genome (Figure 9).

A whole-454 approach is feasible, broadly adopted yet expensive, while a whole-SOLiD approach, despite inexpensive, requires a much more complex bioinformatic analysis and has a computational demand achievable by an *élite* of groups.

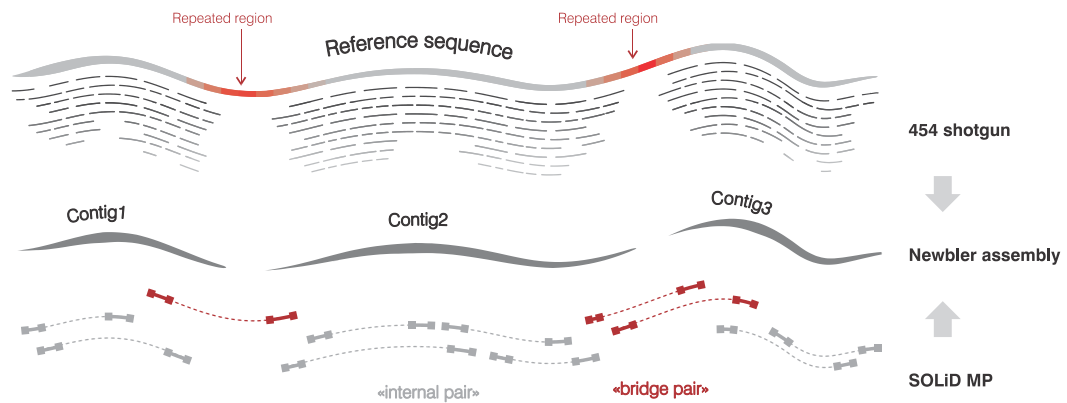


Figure 10

Schematic overview of a mixed approach using Roche 454 shotgun reads to generate a set of contigs (using the software provided by Roche, Newbler), and then one or more long mate-paired libraries sequenced with the SOLiD System used for scaffolding (when the two mates align on different contigs), contig validation and gap closure.

The proposed approach involves a low-coverage 454 sequencing, using a shotgun approach, combined with sequencing of SOLiD mate-paired libraries (Figure 10).

454 reads are assembled with an overlap-based program (I choose Newbler that is supplied with the instrument). A low coverage leaves several non-sampled regions, thus breaking the assembly in many positions and producing a large amount of contigs.

The SOLiD mate-paired reads (that are strand specific coming from the same strand of the DNA insert) are aligned against Newbler contigs and then the alignment file for both pairs are combined together. There are three possible alignment results, as summarized in Figure 11.

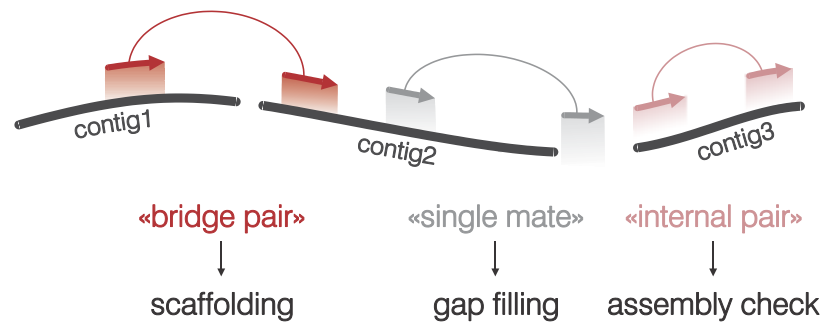


Figure 11

SOLiD mate-paired reads aligned against a set of contigs (gray). There are three main categories arising from pairing of alignments that are more suitable for different applications.

When both mates align uniquely within the same contig they can be used to confirm the contig itself, as long as their match is in the same strand and their distance plausible with the library insert size.

1.4 *N. gaditana* genome project

Our group, in collaboration with the Photosynthesis Group headed by Prof. G. M. Giacometti, decided to sequence the genome of the microalga *Nannochloropsis gaditana* because of its interesting biotechnological potential in biofuel production and because it belongs to a poorly known genus that has an intriguing phylogeny, since it was originated after two endosymbiotic events[15, 16].

Nannochloropsis genus is composed by six species of microalgae (their diameter being less than 5µm). The majority of these species populates marine environments, but fresh water species are also found. Their morphology, either with light or electron microscopy, is not peculiar and their classification is mostly performed via *rbcL* (that encodes the large subunit of the RuBisCO enzyme) and 18S gene sequencing[17].

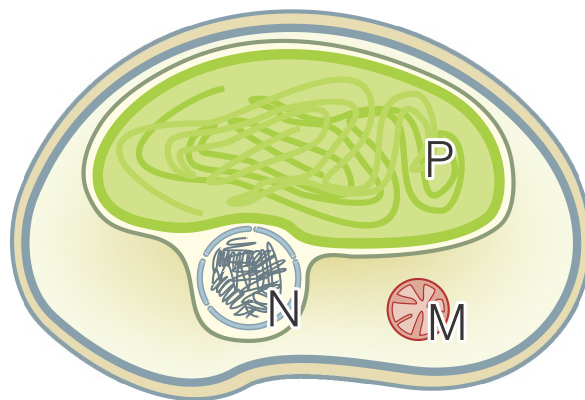


Figure 12

Schematic representation of a typical *Nannochloropsis* cell.

The single chloroplast (*P*) accounts for a large fraction of overall cellular volume, and it is included with the nucleus (*N*) in a membrane.

A mitochondrion (*M*) is shown.

Nannochloropsis, when exposed to stressing environments as nitrogen-deprived media, is able to store solar energy into lipid droplets. These lipid droplets were found to mainly contain triacylglycerols[18], which are the molecules of choice for the production of biodiesel. Even if this behavior is common among other algae, *Nannochloropsis* has been reported to store in lipids the impressive amount of 70% of the overall dry mass[19].

The genome size of *Nannochloropsis* sp. was estimated to be between 30 and 40 Mbp[20], and as we wanted to test the feasibility of the mixed

approach previously described with a relatively small eukaryotic genome, *N. gaditana* appeared to be an excellent choice.

One of the goals of the sequencing project was the production of a good annotation of *Nannochloropsis* in order to identify the set of genes involved in lipid synthesis and accumulation. Moreover there was a great interest on the genes differentially expressed in conditions that led to lipid synthesis comparing to the normal growth conditions. Therefore, to describe the pathways involved in stress sensing and in lipids biosynthesis, RNA-Seq experiments were also conducted.

1.4.1 *N. gaditana* samples for DNA and RNA-Seq

Dr. Elisa Corteggiani Carpinelli prepared both gDNA and mRNAs for the project, preparing cultures with both standard medium and nitrogen-depleted medium. During growth, cultures were tested for the presence of neutral lipids by staining with the fluorescent dye Nile Red and measuring the average signal per cell by fluorometry.

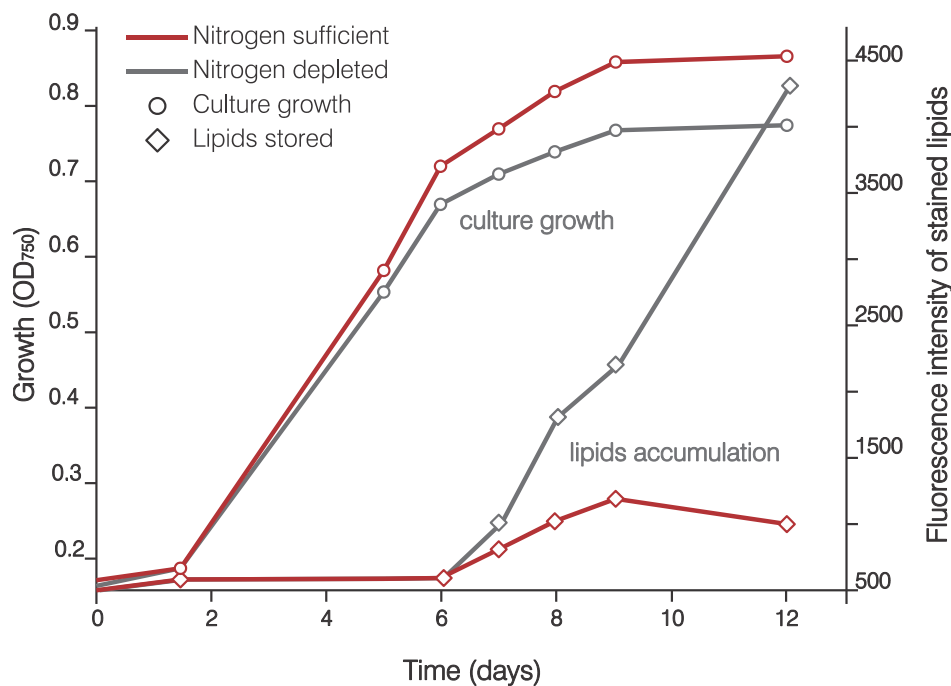


Figure 13

Lipids accumulation (detected via Nile Red staining) and culture density (OD₇₅₀) of *N. gaditana* grown in complete medium (dark brown) and nitrogen depleted medium (pink). Nitrogen depletion slightly affects cell growth (dots), but greatly enhance lipids accumulation (boxes) [16].

Stressed cells were also observed with a confocal microscope (Figure 14), after staining with Nile Red. Observation showed evident lipid droplets in the stressed cells that were absent in the control.

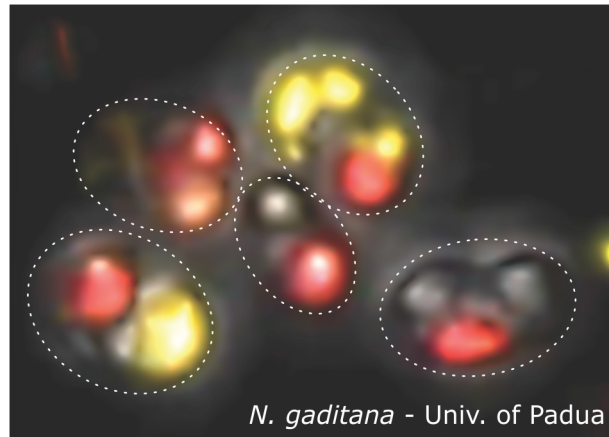


Figure 14

Confocal microscopy of *N. gaditana* cells grown in nitrogen deficient medium, from Dr. Corteggiani Carpinelli's Ph. D. thesis [16]. Nile red stains lipids (yellow) and lipid bodies are clearly present in many cells, while they are not visible in control cultures (data not shown). Chloroplast autofluorescence was also recorded (red).

2 Material and Methods

2.1 Biological sample preparation

Cell cultures and subsequent genomic DNA, total RNA and mRNA isolation and mate-paired libraries were performed by Dr. Elisa Corteggiani Carpinelli and described in detail in her Ph. D. thesis[16].

GDNA SEQUENCING: SOLiD v.3+ (December 2009) sequencing was performed in our group by the sequencing core (Dr. M. D'Angelo, Dr. R. Zimbello and Dr. R. Schiavon).

A full run of 454 Titanium was performed in November 2009 by BMR Genomics srl (Italy), while an additional half plate of 454 XL+ was performed on October 2011 by the Ramaciotti Center at the University of New South Wales (Australia).

RNA-SEQ: cells cultured with standard medium and with nitrogen-deprived medium[16] were collected for transcriptome analysis. mRNA was prepared both via polyA+ enrichment and rRNA subtraction (the former for higher performance, the latter to enable plastidial mRNAs sequencing).

A **BAC LIBRARY** with an average insert size of 120 kbp was prepared by "Bio S&T" (Canada).

2.2 Hardware and OS

One of the aims of this project was to enable genome assembly on commonly available workstations and using when possible Free and Open Source Software (FOSS). When not otherwise stated the development and testing of software packages was performed on a workstation manufactured on 2009: Intel Core 2 Quad Q9300 (2.5GHz, 6MB cache) with 8 GB RAM running GNU/Linux (Ubuntu 8.04 LTS later updated to 10.04).

For Newbler assembly a DELL server with 72GB RAM was used (running Debian Etch).

For Velvet assembly with large datasets an HP server with 8 processors and 2 TB RAM was used (running CentOS 6.2). This server is part of the CRIBI

HPC cluster and tasks have to be submitted via a job scheduler (OpenPBS) installed into a “masternode” server.

2.3 Interpreters and web servers

All the scripts and packages coded for this project are cross-platform and have been tested both in GNU/Linux and Mac OS X. Version used of the interpreter for these languages are reported below.

Relational database: MySQL (5.1.44 on Mac, 5.1.49 on Ubuntu)

Scripting languages: Perl v5.10 (5.10.0 on Mac, 5.10.1 on Ubuntu)
PHP (5.3.2 on Mac, 5.3.3 on Ubuntu)

Web server: Apache 2.0.63 on Mac, Lighttpd 1.4.26 on Ubuntu

2.4 Bioinformatic packages

2.4.1 PASS v.1.65 and PASS-Pair

PASS is a multithreaded program for short reads alignment, with native color space implementation. It was used with a minimum identity of 90%, seed word pattern “-p 11111100111111” (as a SNP results in two mismatches in color space)[21]. Trimming was automatically optimized by PASS to maximize alignment. Prior to alignment reads were converted to FASTQ[22] format with the “csfasta_to_fastq” program provided with the suite.

Alignments were stored both in GFF format and in SAM format (as the latter was introduced later and most tool written for this project were adapted for SAM format just in a second time).

PASS-Pair is a tool of the suite that combines alignment results from both the “forward” and “reverse” of a paired end or mate-paired library. It produces several output files according to the alignment results, in particular for this project:

| | |
|-----------------|---|
| UNIQUE_PAIR | Both reads align uniquely within the same reference sequence (“ internal pairs ” in this report) with the correct mutual orientation and distance. |
| UNIQUE_PAIR_OUT | Both contig align uniquely, but in two distinct reference sequences (“ bridge pairs ” in this report). |

These two files are in the same format alignment format provided (GFF or SAM) and the pairing information is stored in the order: each odd line is paired with the subsequent even line.

2.4.2 Newbler 2.5.3 (January 2011)

Software package for *de novo* DNA sequence assembly developed and distributed with a commercial license by 454 Life Sciences (Roche)[23]. Newbler uses an overlap detection approach.

2.4.3 Velvet 1.2.01

Software package developed by Daniel Zerbino that uses De Bruijn graph to perform *de novo* assembly with huge number of short reads[7, 8]. Not developed for color space, input has to be provided in «double encoded» format (*i. e.* translating each color {0, 1, 2, 3} to a letter {A, C, G, T} even if the translation is unrelated to the original base space).

2.4.4 Other packages used

ARTEMIS is a Java program to display and annotate DNA sequences[24]. Nucleotide tracks can be added to the sequence (the format required is an integer value per line, one line for each nucleotide of the reference).

BLAST was compiled from sources for x86_64 architectures and used with multithreading support, but not with the OpenMPI implementation[25].

CIRCOS, a program to produce circular maps, has been used for chloroplast and mitochondrion genome maps[26].

CGVIEW has been used for chloroplast and mitochondrion genome maps[27].

GRAPHVIZ (<http://www.graphviz.org/>) is open source graph visualization software interpreting the DOT language. It has been used for scaffold visual representation.

PRIMER3 is a command-line program for primer design[28, 29].

SAMTOOLS were used for SAM to BAM conversion, sorting and indexing[30].

SOPRA is a scaffolding program based on paired reads. It has been tested with default parameters on *N. gaditana* data[31].

2.5 Custom tools: technical specifications

Programs written for this thesis are explained in the “Results and Discussion” chapter, technical details about them are reported below.

2.5.1 SOLiD mate-paired reads analysis

Alignments of SOLiD reads (encoded in color space[32]) were performed with the PASS and then paired with PASS-Pair (§ 2.4.1). Internal pair and bridge pair files were converted to a more compact format using, respectively, `uniquepair-compact.pl` and `upo-compact.pl`.

The compact files store in one line the name of the matching reference and starting and ending position of both mates, their size is usually ~10% of the original and being a one-line format they can be easily sorted without losing pairing information.

A. UNIQUE_PAIR COMPACT: FILE FORMAT SPECIFICATION

The GFF input from PASS-Pair is a set of lines providing alignment information in GFF format, having each line followed by the alignment of the other pair. An example:

```
contig00015    pass    match    29422    29453    32    -    .    [..]ReadName_F3[..]Hits=1;
contig00015    pass    match    32420    32451    32    -    .    [..]ReadName_R3[..]Hits=1;
```

The output stores the contig, starting position of the first mate and ending position of the second, as well as alignment strand:

```
contig00015    29422    32451    -
```

B. UNIQUE_PAIR_OUT COMPACT: FILE FORMAT SPECIFICATION

Similarly for what happens for the “internal pairs” file, the “bridge pair” is a GFF with paired lines:

```
contig00015    pass    match    29422    29453    32    +    .    [..]ReadName_F3[..]Hits=1;
contig00211    pass    match    92      127     35    -    .    [..]ReadName_R3[..]Hits=1;
```

The output keeps information about both alignment in one line, sorting alphabetically the two contigs so that all connection between two contigs can be easily found sorting the file. The above example is converted to:

```
contig00015    contig00211    +-      29442    32      92      35
```


C. PIPELINE FROM RAW READS TO MYSQL DATABASE

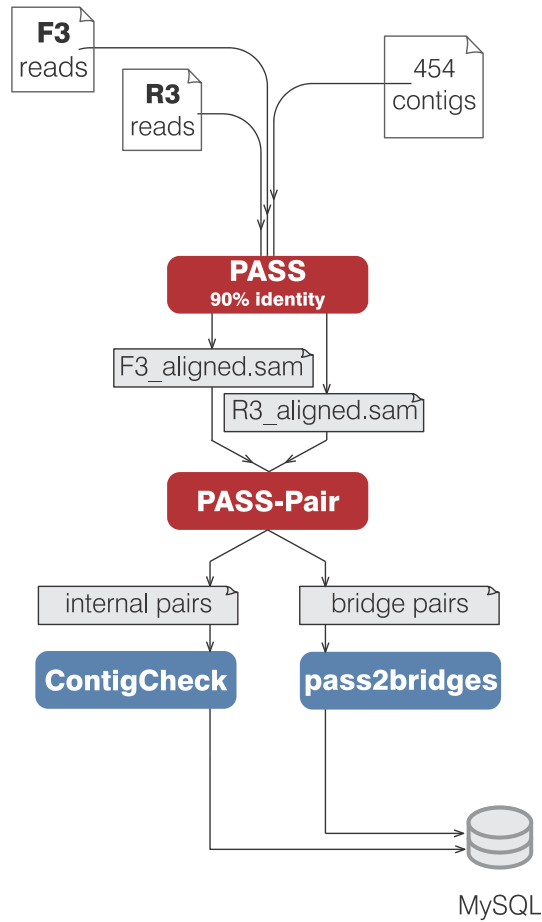


Figure 15

SOLiD mate-paired reads processing pipeline. SOLiD reads (F3 and R3 are the tag for the two mates) are aligned against reference contigs using the PASS program and then paired with PASS-Pair.

Two files (UNIQUE_PAIR and UNIQUE_PAIR_OUT) are used to verify the absence of misassemblies in reference contigs (with the ContigCheck tool) and to create a set of connections between contigs for scaffolding (pass_2_bridges tool).

The simplified files are then parsed to check the presence of misassemblies and to create bridges with tools described later.

D. CONTIGCHECK SCRIPT

The script parses the “internal pairs” file (sorted compact version) and analyzes contig-by-contig the physical coverage of the mate-paired reads (in red in the picture below). Summing the physical coverage of all mates we obtain a global plot (Figure 16, right) that should be bell-shaped for consistent contigs otherwise it is reasonable that the contig was misassembled.



Figure 16

E. **PASS2BRIDGES SCRIPT**

The script parses the “bridge pairs” file (compact version) and for each pairs aligned increment a counter and associate several other information: direction of the connection (stored as four counters: one for each possibility and finally saving the most frequent) and the positions covered. The scripts also store the covered positions in the contig, because wrong connections can arise from small duplicated regions within a contig. Output format is a tabular file and a script loads bridges into the “bridges” table of user’s MySQL database.

2.5.2 **4NGS platform**

The web-based repository was coded in PHP and MySQL. Two Perl scripts were coded to import into MySQL both information about **contigs** (“454contigs_2_sql.pl”) and the “**bridge pairs**” produced by the “gff_2_bridges.pl” script (“bridges_2_sql.pl”).

Database access parameters and navigation bar are stored in a configuration file. The interface uses CSS 2.0 style sheets.

2.5.3 **Gap closure pipeline**

A. **CSX: A CUSTOM SORTED FILE FORMAT FOR SHORT READS**

Color space reads are usually stored in MultiFASTA format, with the disadvantage of making it difficult to search for a particular read given its name. I introduced a custom file format that stores both name and sequence (and optionally the quality) in the same line, separating each field with a pipe character. The script that converts the original MultiFASTA file (csfasta_2_csx.pl) immediately sorts the produced one-lined file.

B. **READS EXTRACTION**

Alignment files (either in SAM or GFF format) are sorted by subject sequence name (contig name), then a script (gff_2_reads.pl) saves all reads matching on each contig into a separate file, converting it in double encoded format for Velvet assembly.

C. VELVET ASSEMBLY AND GAP-CLOSING CONTIG SELECTION

Using k -mer size of 31 and default parameters, all reads extracted for each contig of a scaffold are assembled together.

Velvet output (contigs) is aligned with BLAST against reference contigs (produced by Newbler with 454 shotgun). A script (`blast_2_patch.pl`) parses BLAST output in order to identify those newly assembled contigs matching two Newbler contigs (see Figure 30 on page 40 for a graphical representation). BLAST output allows identifying the missing region, and extracts it from the query.

2.5.4 RNA-Seq tracks and chloroplast map

Data from transcriptome sequencing was added to the 4NGS framework as strand specific coverage tracks. A pipeline converting alignment results to coverage tracks and saving the track in multiple formats has been written in Perl.

For organellar genomes, and chloroplast in particular, the output is a circular map produced with Circos (Figure 17); for the 4NGS pages bitmaps were produced using the GD module Perl; for visualization in Artemis, a simple-text track was saved in Artemis format.

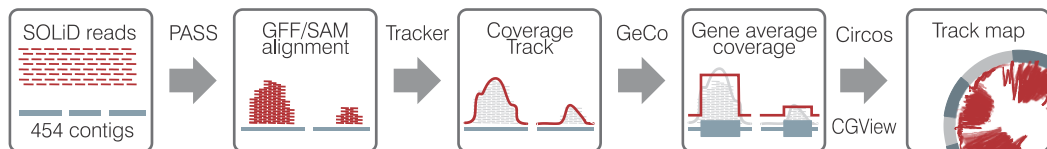


Figure 17

Reads to RNA-Seq coverage track pipeline. Two *ad hoc* scripts were written to save an Artemis-compatible gene expression track, and to produce the input for Circos and CGView.

3 Results and Discussion

This section describes the sequencing data available for the project (454 shotgun, SOLiD MP libraries and BAC-ends), the bioinformatic programs developed for genome scaffolding and gap closure, and finally results obtained by the programs when applied to the sequencing data.

3.1 Sequencing data for the genome of *N. gaditana*

The mixed approach for genome assembly (see §1.3) requires a set of contigs generated by whole genome shotgun with the 454 by Roche, and a set of mate-paired libraries sequenced with the SOLiD by Applied Biosystems. The sequencing of a library of BAC ends was performed to validate the scaffolding procedure and to join separated scaffolds thus creating “superscaffolds”.

3.1.1 454 whole genome shotgun and Newbler assembly

A. SEQUENCING

A first shotgun was performed on November 2009 using the Roche 454 Titanium. Raw output were 741,399 reads (accounting for a total 203 Mbp) with a median read length of 400 bp. A second run using (for a half-plate) latest upgrade (XL+) was performed on September 2011 producing 715,763 reads for a total 806 Mbp (median: 1,102 bp). Read length distribution of both is plotted on Figure 18.

It should be noted that the last sequencing run was performed on late 2011, thus several analysis are still incomplete unlike the Titanium dataset that has been extensively analyzed.

The importance of read length when dealing with *de novo* sequencing has been stressed in the introduction, and will be confirmed comparing assembly performance of the Newbler package with the two datasets analyzed independently.

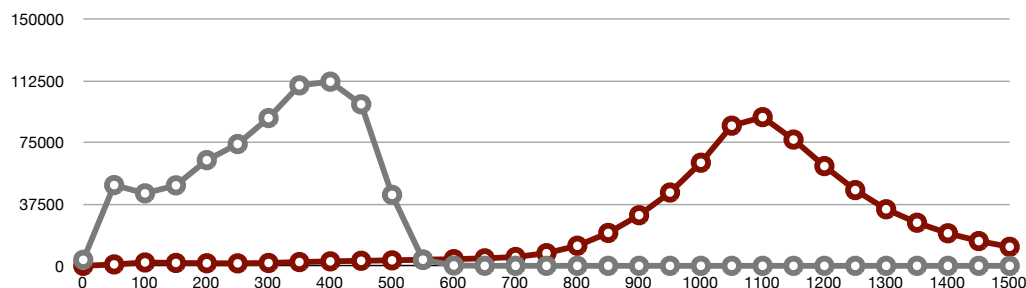


Figure 18

Read length distribution of the 454 Titanium runs: standard kit (gray) was run in late 2009 while latest XL+ kit (dark red) was done in late 2011.

Assuming a genome size of 35 Mb we can calculate a sequence coverage of ~6X for the Titanium run, ~15X for the XL and ~21X for both.

For a bacterial genome a 10X coverage is barely sufficient to produce a good draft, but when dealing with longer eukaryotic genomes it is common to start with a higher sequence coverage[33] (the manual of a *de novo* assembler, ABySS[34], suggests starting from a 40X coverage).

B. NEWBLER ASSEMBLY

Reads were assembled with the Newbler package using default parameters for the two sequencing run (a full run using 454 Titanium kit and half a slide using 454 XL+) and combining the two input together. It should be noted that the deep difference in read-length distribution of the two run affects the assembly results. Better performances are registered with homogeneous input data. Results obtained for the three datasets are summarized in Table 1.

A popular indicator of assembly performance is the «N50» index, that indicates that half of the genome is included in contigs that are greater than the index itself[23].

Table 1

Newbler assembly results for the three 454 datasets. «Titanium-09»: reads produced by a full run using 454 Titanium in 2009. «XL-11»: reads produced sequencing half a slide with the XL+ kit. «Both» refers to the two sequencing data combined.

| Dataset | Estimate | Contigs produced | | N50 | Total |
|--------------------|----------|------------------|-----------|-------|-------|
| | (Mbp) | (>500 bp) | (>100 bp) | (kbp) | (Mbp) |
| Titanium-09 | 48 | 10,246 | 12,045 | 4.6 | 28.4 |
| XL-11 | 29 | 3,410 | 4,862 | 25.5 | 27.2 |
| Both | 273 | 7,035 | 10,271 | 31.4 | 32.1 |

A deeper insight to assembly performance is given by the contig size distribution (Figure 19): the first dataset could not produce any contig longer than 50 kbp (the two 50 kbp contigs are pieces of chloroplast, that being present in multiple copies per cell, has an impressive 200X coverage). Assembly of the reads produced with the XL kit gives (in *pink* in the chart) much better results.

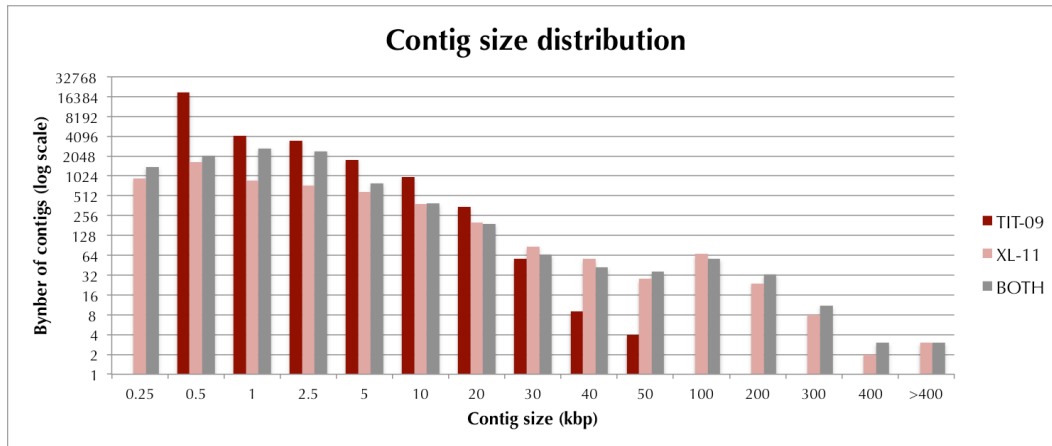


Figure 19
Contig size distribution chart for the three datasets mentioned in Table 1.

C. ASSEMBLY ACCURACY

The CheckContig script, that verifies the physical coverage obtained with MPs, detected only three misassembled contigs (contig12452, contig09180, contig07916) on the Titanium-09 dataset (an example in Figure 20). The shortest MP library, spanning from 1.5 to 3.0 kpb, does not give the necessary resolution power for a dataset with a short average contig size (7,494 contigs are shorter than 1.0 kbp). A first survey on other datasets showed that there isn't any evidence of misassembly. However a deeper analysis has to be performed to confirm this evidence.

When I tested the scaffolding program with and without misassembly correction, it was evident that between the few wrong scaffolds, the majority was due to misassembled contigs given in input. Therefore, the future implementation of the program that creates “bridges” will also check the coverage of MP aligned prior to pass the information to ScaMP.

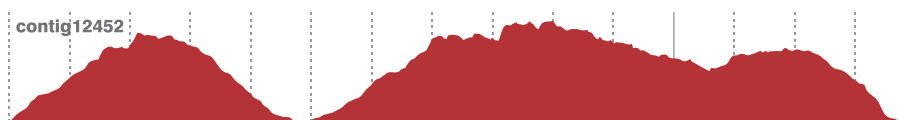


Figure 20
An example of contig misassembly in a contig of the Tit-09 dataset.

D. GENOME SIZE ESTIMATE

We expected a genome size for *Nannochloropsis* raging from 30 to 40 Mbp[20]. Newbler gives an estimate of each assembly, reported in Table 1, which is however affected by the presence of high-coverage plastidial contigs. The remarkable difference emerging when combining the two

datasets, resulting in a completely wrong estimate of almost 300 Mbp, seems to be an artifact due to the deep difference between the two datasets.

3.1.2 SOLiD mate-paired libraries

Two mate-paired libraries were prepared: one with an insert size range of 1.5–3.0 kbp and a second with an insert size range of 3.0–5.0 kbp. Both libraries were sequenced in a SOLiD v.3+ slide divided in four lanes (the two lane model was not produced). This produced four data sets whose size and sequence coverage is reported below.

Table 2

Reads produced sequencing two mate-paired libraries

| Mate Paired Set | Number of reads | Data produced | Coverage |
|---------------------|-----------------|---------------|----------|
| 1.5-3.0kbp_A | 74.749.807 | 7.47 Gbp | 213X |
| 1.5-3.0kbp_A | 69.418.621 | 6.94 Gbp | 198X |
| 3.0-5.0kbp_B | 68.334.726 | 6.83 Gbp | 195X |
| 3.0-5.0kbp_B | 78.164.673 | 7.82 Gbp | 223X |

Average quality was 17.5 for the first color and decreases to 12.0 for the last position (50th).

As part of the scaffolding pipeline the mate-paired reads were aligned with PASS (alignment statistics are reported in Table 6 on page 68) against reference contigs, and then the alignments of the two mates were paired using the Pass_pair tool.

The “internal pairs” can be used to have a downstream estimate of libraries insert size. Both libraries appears to be slightly shifted towards lower values, but while the short library has 99% of pairs within declared boundaries, the larger only 60%, appearing to be a 2.0–4.0 kbp (Figure 21).

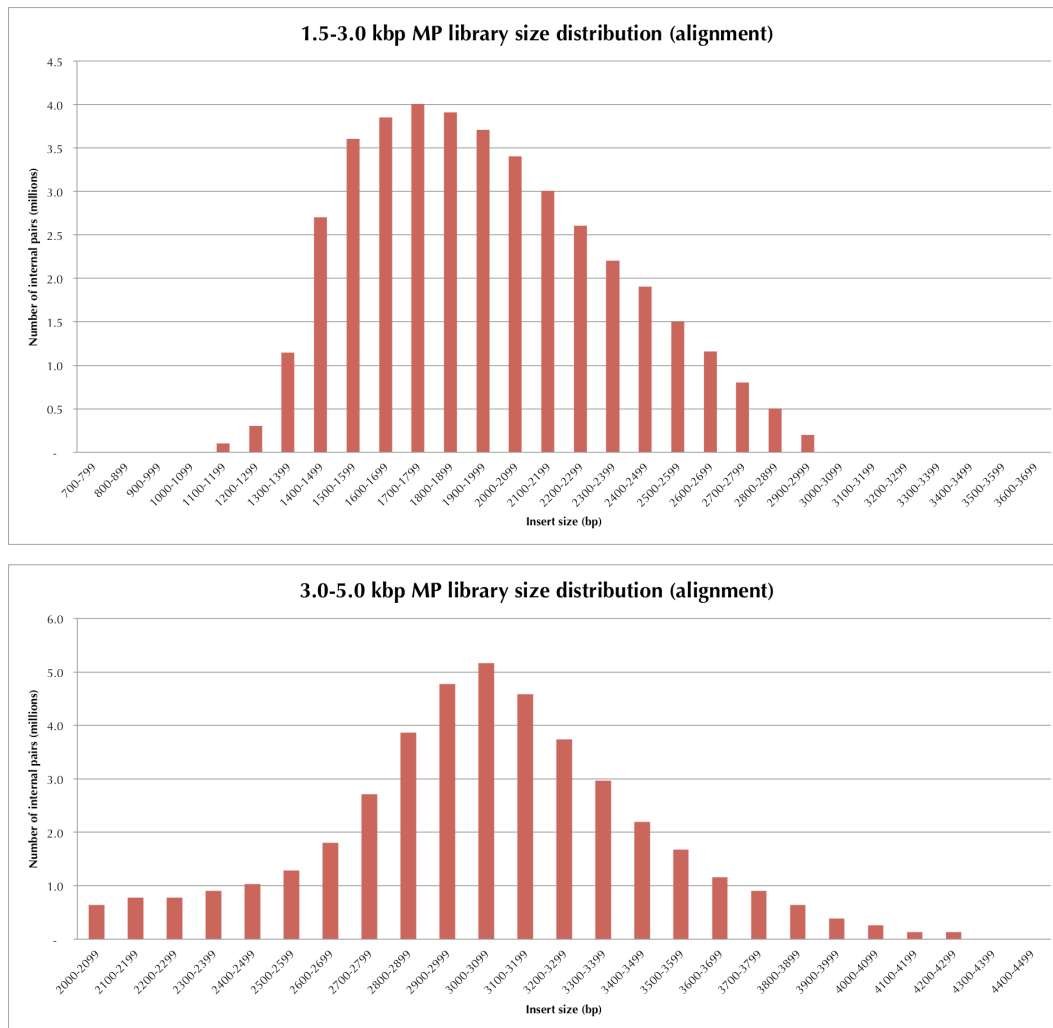


Figure 21

Distribution of MP libraries insert size as appearing on alignment against reference contigs (XL-11 dataset). 1.5–3.0 kbp library (*top*) and 3.0–5.0 kbp library (*bottom*).

3.1.3 BAC-ends

The genome of *Nannochloropsis* is being used in my research group also to test a novel method of physical mapping. A first step of that project required the production of a BAC library (having the average insert size of 120 kbp), thus we decided to have a 7 96-well plates of BAC ends sequenced with traditional Sanger method accounting for a total 665 BAC-end pairs. 76 sequences failed during sequencing thus reducing the number of valid pairs to 393, accounting for a total 1.3 X physical coverage.

3.2 New bioinformatics tools

Aim of my project was the design and implementation of bioinformatic tools for genome scaffolding and finishing to assist the mixed approach described in §1.3. In this section I describe these programs while I remand to §3.3 for their performance on *N. gaditana* genome assembly.

3.2.1 4NGS: a user-friendly data repository

A first necessity arisen from this project was a repository to store data from genome and transcriptome sequencing. It's a common habit to set up a genomic browser at the end of a genome project, but what we were lacking was a tool to share (with co-workers) genomic data as it was produced, to enable cooperation and immediate access on that data.

The platform is conceived with contig-centric model (screenshot in Figure 22), showing for each contig its connection with others, the physical coverage track, RNA-Seq tracks and basic information about the contig itself (size and sequence coverage).

The interface allows for manual scaffolding following the links between on contig end and the other, thus integrates a system for manual scaffold annotation. After data production we decided to perform some manual scaffolding in order to verify its feasibility (*i.e.* that the low 454 coverage was enough).

BLAST, a query system to select contigs and a primer-design tool for scaffolding verification via PCR were also integrated. The "scaffolds" section allow for scaffold structure visualization, editing and manual annotation.

Beside this front-end, 4ngs database has two tables, «contigs» and «bridges», that are used by the scaffolding program I wrote to make scaffolds.

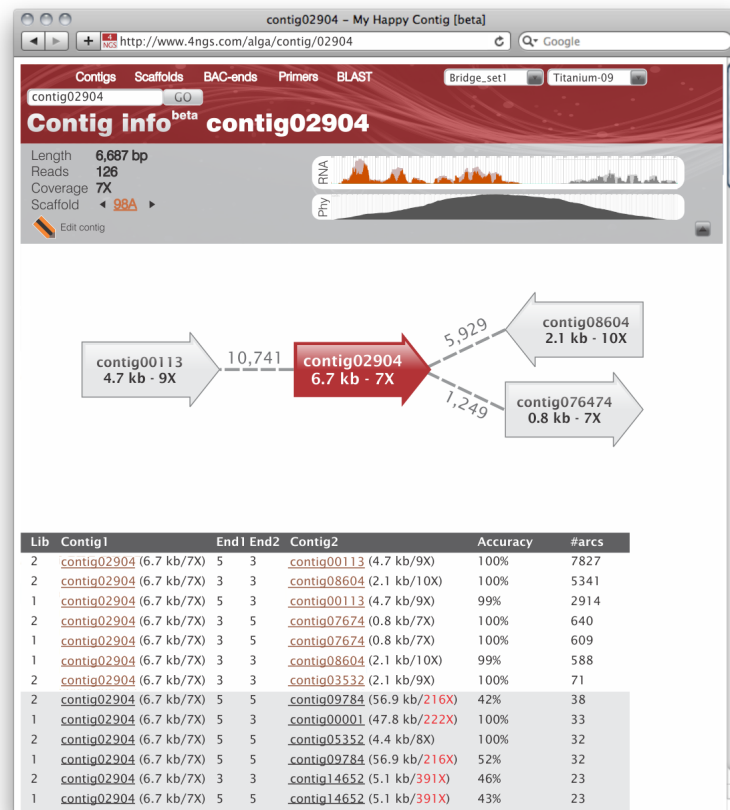


Figure 22

Screenshot of a contig page from 4NGS. The complete list of bridges is presented in a table at the bottom, while high-connection bridges are summarized in the scheme in the middle.

3.2.2 ScaMP: a tool for automatic scaffolding

I developed a program called ScaMP and a pipeline for automatic scaffolding based on it. The core program starting from a seed contig ("seed") and crawling to a specified direction, and a pipeline extend the procedure genome-wide.



Figure 23

ScaMP (Scaffolding with Mate-Pairs), program logo

A. INPUT PREPARATION: CONTIGS AND BRIDGES IN A MYSQL TABLE

Data for scaffolding are stored into a MySQL database (to increase speed and reduce memory usage) composed by two main tables: contigs and bridges. The former contains name, length, coverage of all contigs produced by Newbler and is populated by a Perl script that parses the MultiFASTA output of Newbler.

The latter is populated by a script that parses the compact version of `UNIQUE_PAIR_OUT` (containing “bridge pairs”) file from Pass. Each pair alignment between two distinct contigs is counted, recording the direction of the alignment. For each “bridge” the program saves the amount of alignment that confirm that connection, the direction of the alignment (in terms of contigs extremities connected: “5-3” means that the 5’ end of the first contig is connected to the 3’ end of the other) and the consistency of alignments (all mate-pairs should connect the two contigs with the same orientation, the program saves the percentage of the prevalent orientation).

B. RECURSIVE SCAFFOLDING FROM A “SEED” CONTIG

The main program of ScaMP starts scaffolding from a given contig (called “seed”) using extension algorithm accesses the MySQL database previously populated, and continues extension as long as possible.

A simplified scheme of the core function is shown in Figure 24: retrieving all connections from a contig to the desired direction (*i. e.* 5’ or 3’) and selecting a proper contig to continue.

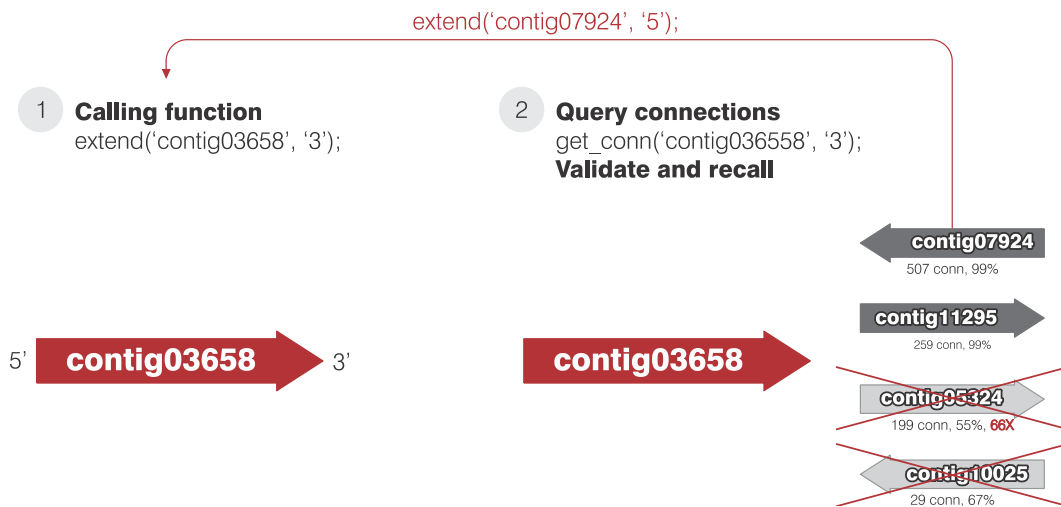


Figure 24

Scheme of the basic functioning of ScaMP. Extension continues only if all possible paths (after discarding less plausible connections) converge. If this happens, the core function “extend” is called with the new contig found and the new extension direction. Some connections are discarded *a priori* either because of a much too high coverage, or because of a low consistency of direction (reliable connections have it $\geq 98\%$).

The list of “bridges” is *a priori* filtered discarding connections with few arcs (a suggested threshold is contig specific and expressed as $t = \frac{1}{10} \cdot T$, where T

is the highest number of mates composing a connection from the contig of interest) and/or with a low direction concordance (suggested setting: $\geq 97\%$). After this filter if there is only one possible connection the program proceeds, if more than one connection are still present the program extends recursively all possibilities, and if they collapse within n recursion steps (suggested $n = 5$, maximum $n = 8$) the programs tries solving the path if possible and proceeds. Figure 25 displays two exemplification schemes of: a completely solved paths (example on the top); a small “bubble” (on the bottom), that is a set of contigs whose mutual position cannot be solved due to missing connections.

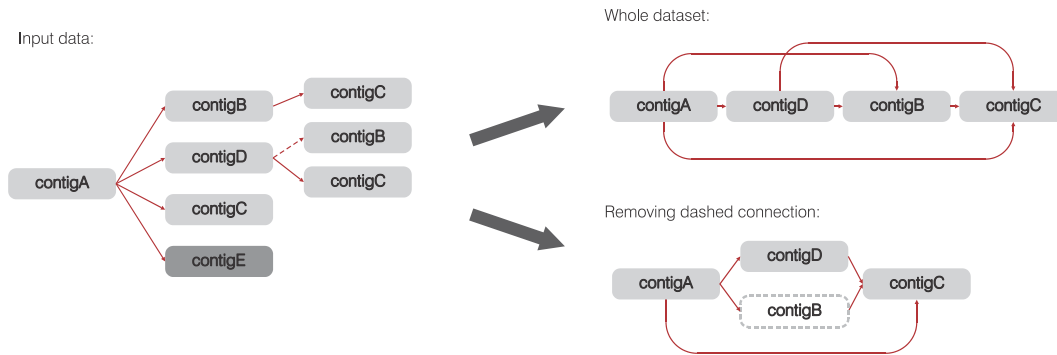


Figure 25

A double example of extension by ScaMP. Connections starting from “contigA” are shown, and the one pointing to “contigE” is discarded *a priori*. In the first example let’s consider all the connection shown in the left panel: it is possible to unravel the nodes and to determine the correct order of contigs (right panel, top), but if we suppose not having the connection from “contigD” to “contigC” (dashed line) we obtain a “bubble”, meaning that we can’t know if “contigB” precedes or succeeds “contigD”. ScaMP takes the longest and ignore the other (in the example, “contigB”).

The program stops the extension when there are no more arcs, when the different paths starting from the last contig do not collapse together or when it finds ahead a high coverage contig (coverage threshold is user defined, and usually it’s safe to set it in terms of average contig coverage, C_{avg} , and its standard deviation σ : $C_{max} = C_{avg} + 4 \cdot \sigma$).

C. WHOLE GENOME SCAFFOLDING

A pipeline for whole genome scaffolding has been implemented in BaSH/Perl. A query to the “contigs” table retrieve a list of good seeding contigs (meaning that the coverage is between $C_{avg} - 2 \cdot \sigma$ and $C_{avg} + 2 \cdot \sigma$) and with a minimum length, if desired. All the seeding contigs are extended as described above: keeping track of contigs added by the scaffolding process

and removing them from the seeds list, if they were already included in a scaffold.

This generate a set of scaffolds that can still overlap, thus a Perl script performs a polishing process.

All data about scaffold is added to the MySQL database.



Figure 26

Whole genome scaffolding pipeline.

3.2.3 BAC-Validate: scaffold validation and super-scaffolding

BAC ends sequences are a valuable tool for scaffold validation and to connect adjacent scaffolds. The intimate logic of BAC ends is exactly the same as for mate-paired reads, but with consistent technical difference. BAC ends falls on opposite strands, while the two mate-paired reads are sequenced in the same strand, but much more important is the insert size that for BAC ends exceeds 100 kbp, allowing for resolution of virtually all sorts of repeats.

A fully automated pipeline, integrated into the 4NGS platform, processes the chromatograms and extracts the sequence, that is aligned against reference chromosomes using BLAST and if a single match is found the program associates the contig and its scaffold to the sequence.

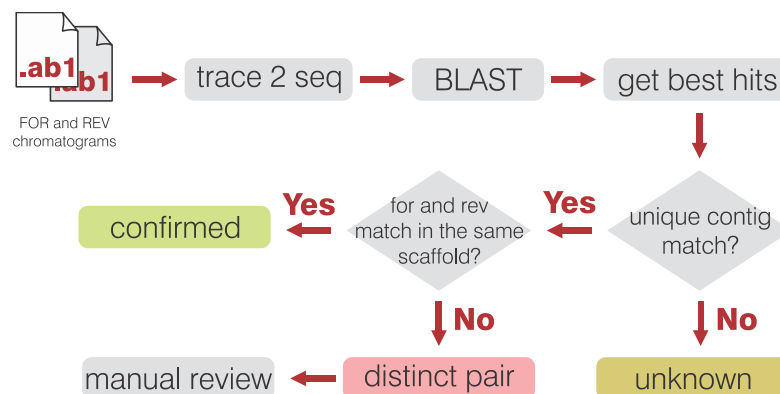


Figure 27

BAC-ends sequence analysis workflow. See Figure 34 for a review on the relationship between BAC-ends alignment and scaffold integrity.

Each pair is classified based on BLAST result: “unknown” if one of both sequences failed or their alignment gave no match or multiple matches,

“confirmed” if both forward and reverse sequence align within the same scaffold, and *“distinct pairs”* if the two sequences aligned against different scaffolds (Figure 27).

BAC-ends falling in the latter class could either join two different scaffolds or be a hint of a misscaffolding: thus are loaded into a section of the 4NGS platform for manual review, consisting in browsing from the contig matching with the forward sequence to the reverse sequence best hit. When this is possible the two scaffolds are joined together, but sometimes there is a lack of coverage (*i. e.* there are no more connection in the desired direction) leading to the creation of a super-scaffold: it is known that the two scaffold should be joined together but it's not possible to verify this via mate-paired reads.

The platform ranks connections to be verified counting the number of independent and provide a graphical representation of the physical coverage of scaffolds: lack of coverage in the middle of a scaffold could suggest a possible misassembly (even if it should be noted that the number of BAC-ends sequenced is too low, accounting for a 1X physical coverage).

3.2.4 Manual finishing assistant

A scaffold is a set of contigs and gaps. To solve a gap and join the adjacent contigs it is possible to design a specific PCR and having it sequenced (a single step strategy is possible whenever the gap is smaller than the sequencing capacity of the Sanger sequencing. This condition is rarely verified when the sequence coverage is high enough).

I implemented in the 4NGS repository a tool for primer design with an Ajax interface. The user inputs the starting contig and receives a set of suggestions based on the “bridges” table (mate-paired reads). Once that starting and ending contigs, and their mutual orientation, are chosen the program invokes Primer3, retrieve a set of primer couples and aligns them with BLAST against the contigs. The user can finally add the desired primer pair into a wish list for cumulative orders.

A screenshot of the primer design results is shown in Figure 28.

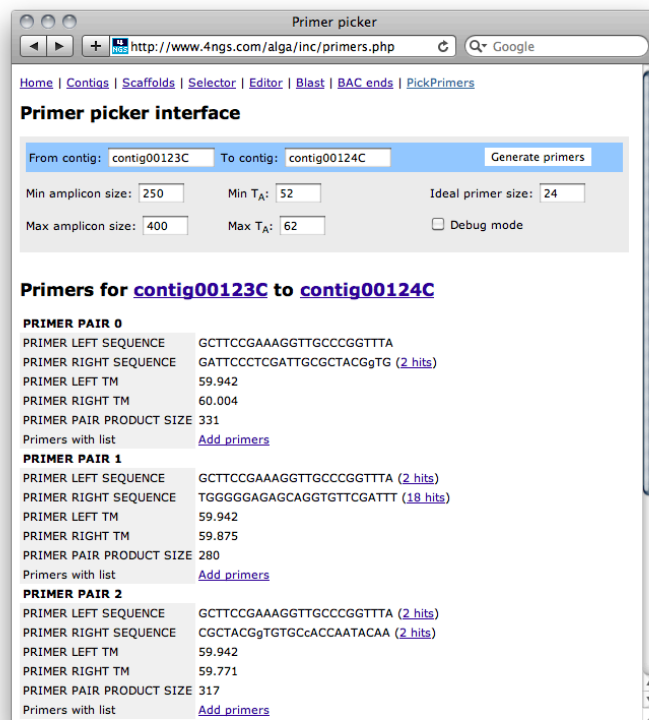


Figure 28

Web interface for primer design. Ajax implementation helps the user choosing the correct destination contig and its orientation.

3.2.5 PatchGap: a pipeline for gap-closure via local assemblies

Scaffolding is the fundamental process in genome assembly that makes order among genome pieces. Several analyses are made on sequence level (gene prediction and annotation, regulatory elements discovery, etc.) and they are affected by the fragmentation of a genome into several contigs.

Making use of the MP libraries it should be possible to “close the gaps” between contigs, using the “single mates” (see Figure 11 on page 2), the pairs that had just one mate aligned uniquely, because the other falls in a not assembled region, a gap. Collecting these reads and performing a *de novo* assembly should help recovering these missing parts of the genome, because the complexity of assembly is greatly reduced.

The «PatchGap» pipeline that I developed aims closing gaps between contigs gathering short reads not present in the reference contigs by means of their mate-paired reads that do.

A naïve approach could involve the retrieval of all the mates aligned in the contigs surrounding the gap to perform the local assembly, but when the

size of MP library is comparable or bigger than the size of contigs this could be a less effective strategy. A simple example is depicted in Figure 29 where a gap of interest (“gapBC”) is covered by MPs that don’t start from adjacent contigs.

Mate-paired covering a gap:

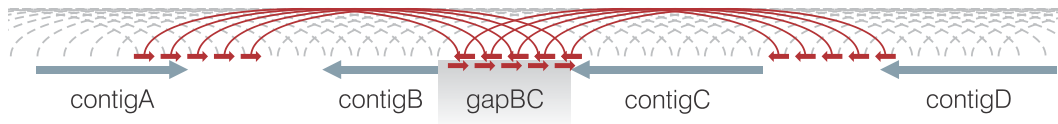


Figure 29

A set of four adjacent contigs (*air force blu*) and the MP covering them (*gray, dashed*). Explanation in the text. In the top panel MP falling in the gap are highlighted in *dark red*.

I prepared a general pipeline that, for each contigs, saves all the reads aligning in it and their mates, then they are used for local assemblies: all reads connected to contigs part of a scaffold are assembled together, while reads connected to contigs not part of a scaffold are assembled independently.

All the resulting Velvet contigs are aligned with BLAST against Newbler contigs, and a program looks for Velvet contigs matching with two Newbler contigs. I refer to these type of contigs as «**patches**» (see Figure 30).

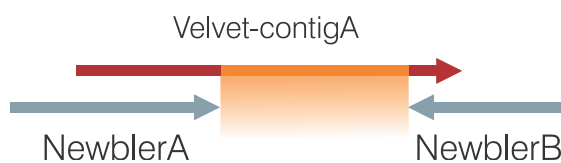


Figure 30

A «patch» is a contig assembled locally by Velvet using MP reads (*red*) that act as a bridge between two contigs assembled by Newbler (*gray*). Highlighted region (*orange*) contains the “gap” sequence.

3.3 *N. gaditana* genome scaffolding

3.3.1 ScaMP testing with selected seeds

ScaMP has been extensively tested to spot possible causes of misassemblies, that were implemented in the algorithm and in the choice of parameters in order to have a conservative and less error-prone tool, even if it could break good scaffolds when facing less clear situations: high coverage contigs (could lead to a repeated region), or non-converging paths.



Figure 31

A graphical representation of the largest scaffold produced by ScaMP using the contigs generated with the Titanium dataset (magnification provided in the inset). This scaffold, 1.2 Mb long, includes 292 contigs.

The program has been tested comparing its scaffolding results with manual work performed browsing through the 4NGS interface: more than 20 large scaffolds (*i. e.* containing more than 20 contigs) have been compared. ScaMP never produced misassembled scaffolds in this small test set, but, occasionally, it interrupted the extension progress a few contigs before the manual curator. This was mainly due to the presence of regions with low coverage, where the number of “connections” between contigs was lower than the fixed threshold, and the program was therefore forced to stop.

ScaMP was implemented into 4NGS so that the user can extend a scaffold starting from any contig and tune parameters to get best results. It can print

a graphical representation of the scaffold using the GraphViz program: Figure 31 shows the graphical output of the scaffold produced from a seed. Testing of the program showed promising performance in terms of speed (from a fraction of a second to few seconds depending on the length of the final product and the number of connections to be explored), in term of number of scaffold produced but in particular in term of accuracy. ScaMP has been developed to be conservative and several events trigger the exit instead of continuing the extension of the scaffold.

A critical aspect for good scaffolding is the starting dataset: misassembled contigs (*chimeras*) lead to unfaithful scaffolding, and an even more important data is the reliability of “bridges”. Being created clustering the output of an alignment program they can make use of extensive information about each read mapped, thus its is possible to improve scaffolding with a more robust alignment parser.

3.3.2 *N. gaditana* genome scaffolding

The ScaMP pipeline was run with the three Newbler datasets, all results were stored in the 4NGS framework.

Scaffolding performance was more than satisfactory for all the datasets, with substantial differences: on the Titanium-09 contigs 20.9/28.4 Mbp (72%) were included in scaffolds, 26.4/27 Mbp (97%) on XL-11 contigs and 84% for the combined datasets.

The first dataset is more fragmented, with an average contig size of just 2.7 kbp, has many contigs per scaffold (see Figure 32), and a long list of small scaffolds with just a few contigs. Total number of scaffolds for this dataset is 312, even if the first 20 scaffolds includes 11 Mb, a half of the whole scaffolds.

As emerged from Newbler assembly (Table 1 on page 29), the second dataset is more robust in terms of number of contigs (small) and contigs size (N50 of 25 kbp), but at the expense of genome sampled (only 27 Mb).

If we consider the size of scaffolds expressed as sum of their contigs length (ignoring gaps) the quality of XL-11 assembly gives better results, both alone and in the combined dataset (Figure 33).

The Titanium-09 dataset has been used for algorithm design and extensively tested to tune the parameters. It has not been possible, yet, to perform analysis with the same level of accuracy for the last sequencing run (October 2011), but is planned to have it done soon.

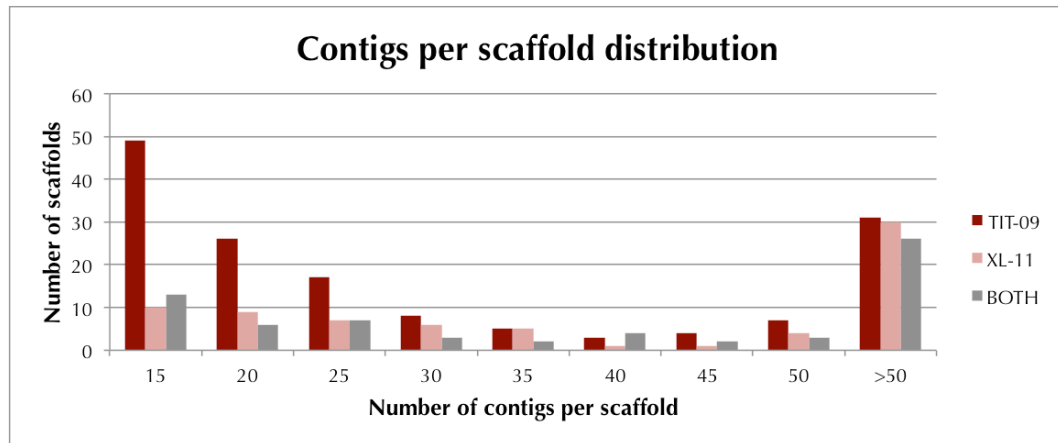


Figure 32

Number of contigs per scaffold added by ScaMP. The Titanium-09 dataset is more fragmented, thus yields scaffolds with the highest contigs number.

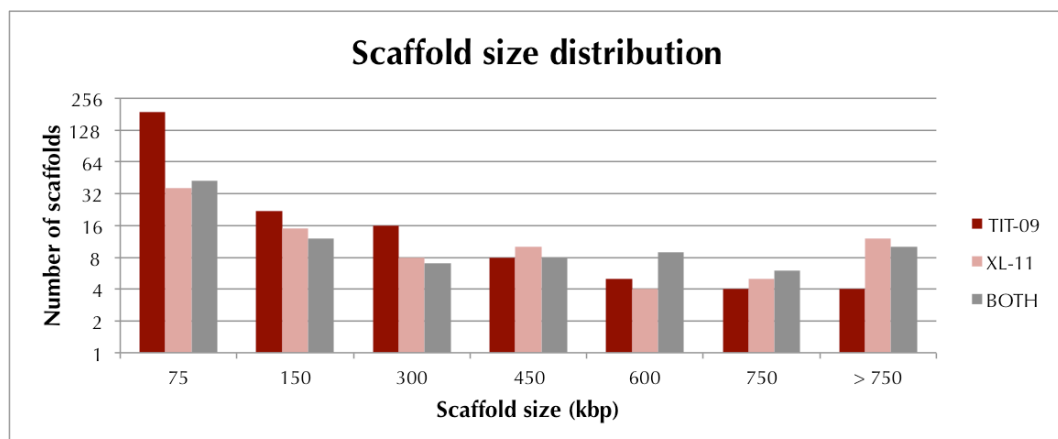


Figure 33

Size of scaffolds expressed as sum of their contigs length (plotted in *logarithmic* scale).

3.3.3 BAC-ends for scaffolds validation and superscaffolding

BAC-ends sequenced with Sanger are a valuable tool for scaffolding, but – thanks to their average insert size of 120 kb – they have been used as a testing tool for ScaMP output.

When the two ending sequences of a BAC insert match against two contigs of the same scaffold (and the sum of the contigs size between them is compatible with BAC library), it is possible to have an independent proof of the correctness of the scaffold.

On the other hand when two BAC ends falls on different scaffold and the sum of the contigs between them largely exceeds the average insert size of BAC ends it is a strong evidence of a misassembly.

When a region of a large contig has no physical coverage it is marked for manual verification: it could be a misassembly but also a lack of coverage, having just a 1X physical coverage with BAC ends.

Beside their usefulness as testing tool, BAC ends can be implemented in scaffolding by joining two independent scaffolds (forming a so-called «**superscaffold**»). All these events are summarized in Figure 34.

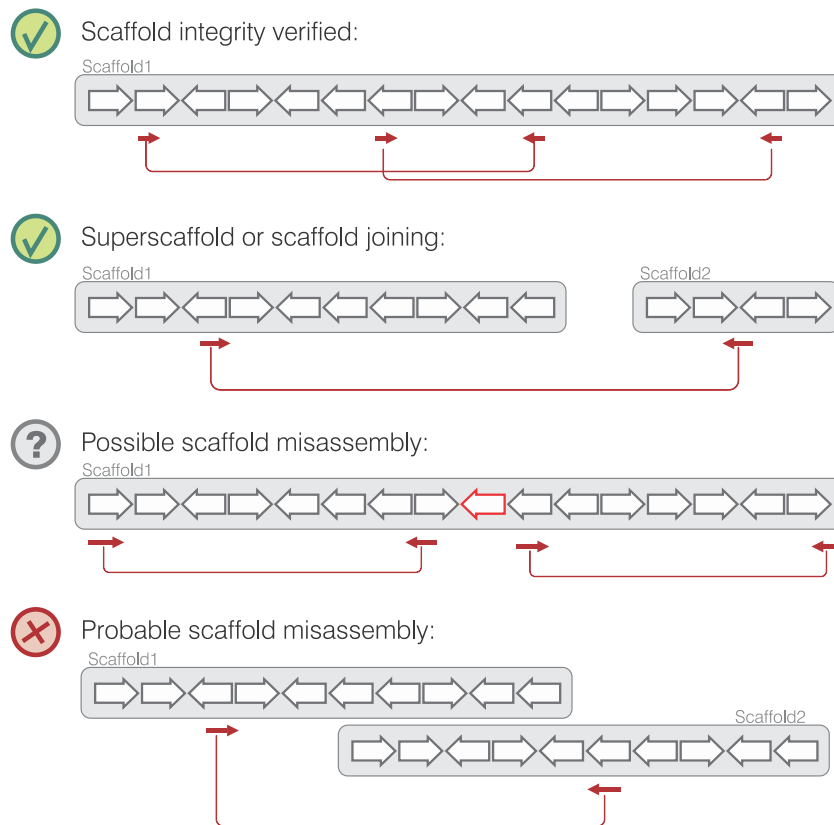


Figure 34

BAC-ends alignment against scaffolds: they can be used to confirm an existing scaffold or to join together independent scaffolds (two top panels). If there is a lack of physical coverage in the middle of a scaffold it could be caused by the low coverage of BAC-ends or because there is a misassembly, thus such regions have been manually controlled

A. SCAFFOLDS VALIDATION

From the alignment of BAC ends against the three datasets we had on average good results. When comparing the contig matching with the two ends, 9% (46/512) of BAC ends confirm a contig sequence for the Titanium-

09 dataset (that has shorter contigs, thus with a lower probability of being confirmed via BAC ends). With the XL-11 dataset this fraction raises to 22% (117/522) and reach the 24% with the combined set.

When comparing the scaffold found via alignment by the two sequences of the pair, 64% confirm a scaffold for the Titanium-09 dataset, 87% for the XL-11 and 85% for the combined dataset.

B. SUPERSCAFFOLDS

Using the connections between scaffolds obtained with BAC-ends we were able to produce 23 superscaffolds out of 98 scaffolds. These superscaffolds include 12.2 Mbp (one third of the whole genome).

It is reasonable to think that some of this could be whole chromosomes or chromosome-arms (from pulsed-field gel electrophoresis we noticed that biggest chromosomes are less than 2 Mbp long, which is the approximate size of biggest superscaffolds).

An interesting example, shown in Figure 35, is “superscaffold1” that has been originated joining five scaffolds. Scaffold136 and Scaffold122 were separated by a single contig 100 bp long and with an impressive 2000X coverage, that could be a centromeric repeat collapsed in a short sequence.

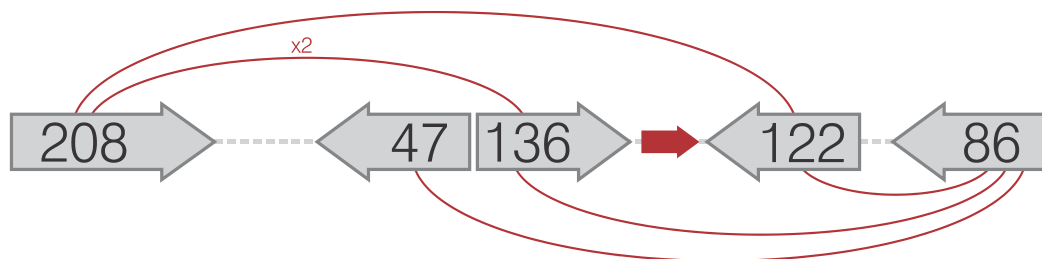


Figure 35

A superscaffold obtained joining five different scaffolds (gray) by means of BAC-ends sequences (red arcs). An interesting 100 bp contig with 2000X coverage (red arrow) joins two scaffolds: could be a centromeric repeat collapsed.

3.3.4 Gap closure results

The gap closure pipeline has been tested for the Titanium-09 dataset.

It should be noted that all the assembly was performed on a desktop computer with 8 Gb RAM (four assembly at the time, being a four-cores system) because local assemblies are little resource demanding.

The `blast_2_patches.pl` script identified 3,262 contigs assembled by Velvet that could fill a gap, and 2,686 of them (82%) were found to connect two contigs of the same scaffold and only 58 connected contigs belonging to different scaffolds (patches of this kind could be misassembly, correct patches joining contigs non in the correct scaffold or repeated, or correctly joining two scaffolds). This small fraction of patches requires a manual validation that will be performed soon.

The remaining 16% of patches connects two contigs (both or one of the two) that were not included in any scaffold.

Gap size distribution (Figure 36) shows a remarkable fraction of small gaps, with a 11% of all gaps shorter than 10 bp.

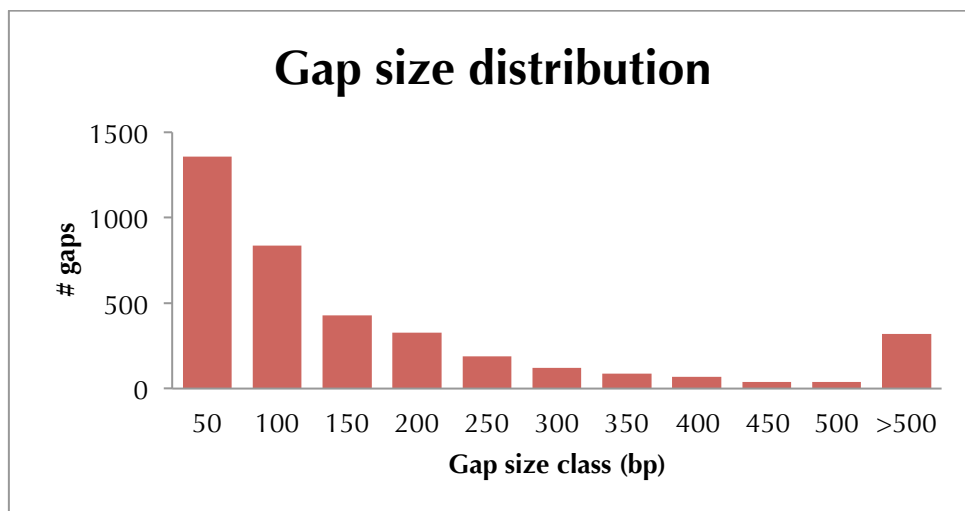


Figure 36

Gap size distribution. Almost 50% of gaps identified is shorter than 50 bp (and an 11% is shorter than 10 bp).

When performing gap closure on whole scaffolds (Titanium-09 dataset) it has been noted that most patches join clusters of adjacent contigs and that long scaffolds had a higher fraction of gaps closed.

Gaps filled in Scaffold230



Figure 37

Scaffold230 as an array of contigs (*black boxes*, not in scale). Gap filling joined nine clusters of contigs (*red boxes*) raising the N50 value from 8.3 to 77.4 kbp.

As an example Scaffold230 is composed by 140 contigs with an average size of 5.3 kbp (N50: 8.3 kbp). The gap closure pipeline identified 141 patches that resulted the number of contigs to 34, and raised the average contigs size to 21.9 kbp (N50: 77.4 kbp). A schematic representation is shown in Figure 37.

3.4 Chloroplast genome of *N. gaditana*

Nannochloropsis has a single chloroplast with multiple copies of plastidial genome, thus resulting in much higher sequence coverage than that of the nuclear genome. Among the high-coverage contigs three were found to be plastidial via NCBI BLAST queries:

- contig09847 (56.9 kbp, 216X coverage) includes the RuBisCO large subunit coding sequence;
- contig00001 (47.8 kbp, 222X coverage) includes the *psA* gene, part of Photosystem I;
- contig14652 (5.1 kbp, 392X) includes a ribosomal operon related to other chloroplast, that because of its coverage could be the typical chloroplast IR.

Beside the presence of plastidial genes there were a relevant similarity to the plastidial DNA of *H. akashiwo* and *T. pseudonana*.

Using information from MP alignments we proposed a model (shown in Figure 38, outer ring) that was verified via PCR, designing primers spanning the four junctions between the three contigs. All the PCR were positive, confirming the model, and were sequenced via Sanger. As expected there were small gaps between the contigs (except for one out of four junctions) that have been identified and used to produce the complete sequence of the plastidial genome.

A preliminary gene prediction has been performed combining *ab initio* ORF finding and alignment of *T. pseudonana*'s genes (Figure 38, middle ring), while data from RNA-Seq (inner ring) has not been implemented yet.

It is relevant to report that RNA-Seq libraries were prepared both via PolyA+ enrichment, that is a proven and effective method to get rid of rRNAs, and with rRNA depletion. The latter method preserves polyA- mRNAs, also

including plastidial transcripts. It is reasonable to think that important metabolic pathways connected with photosynthesis could be under control of plastidial genes.

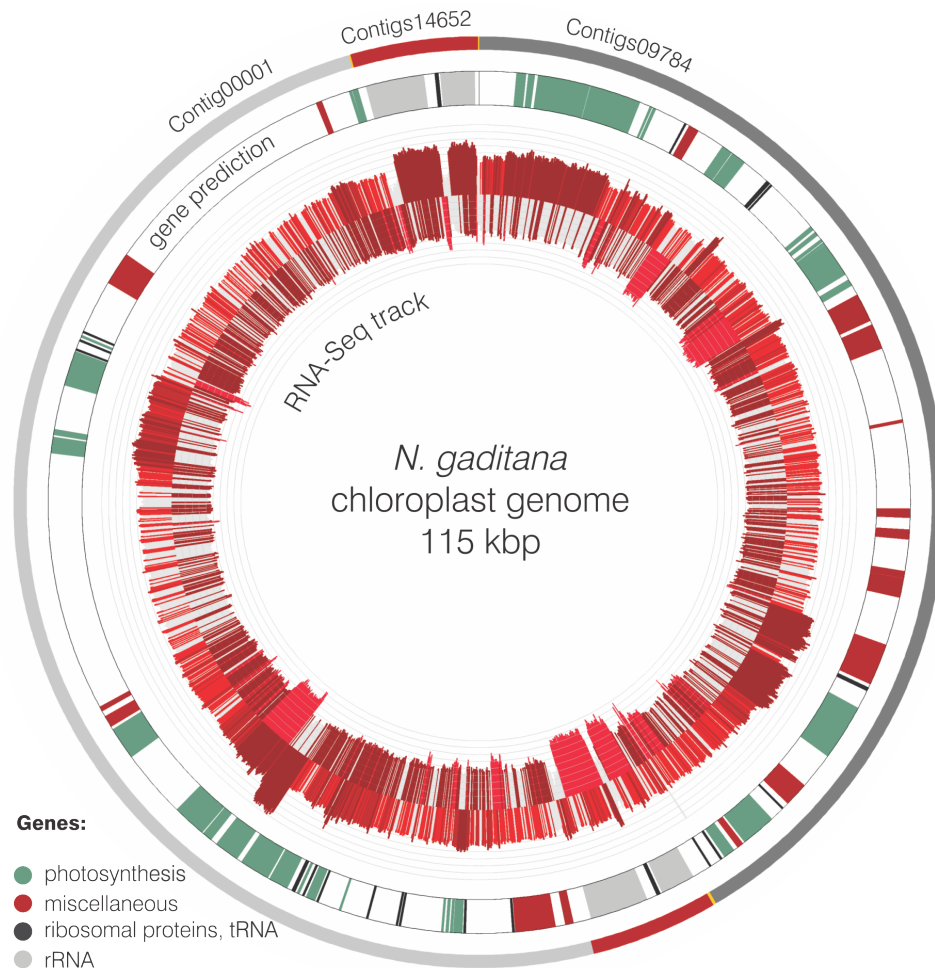


Figure 38

Chloroplast genome of *N. gaditana*. *Outer ring*: the three contigs composing the plastidial genome. *Middle ring*: gene prediction. *Inner ring*: RNA-Seq track (log scale) performed with strand specific sequencing.

3.5 Wheat: an independent test set

ScaMP was designed in the context of *N. gaditana* genome sequencing, working with a very high coverage of MPs. Our group joined the international consortium for Wheat genome sequencing (for Chromosome 5A, ~500 Mbp) and I tested the ScaMP pipeline on data available for this project: a set of contigs made with a 454 shotgun (2X coverage, 229,594 contigs), but a very low coverage of SOLiD MP (approximately 1X). The whole genome size is ~16 Gbp. I worked using MP generated from whole genome preparation and contigs from a Chromosome 5A shotgun. The major problem of this dataset is the extremely poor MP coverage, which prevents the preparation of a robust datasets of “bridges”.

The program produced 660 scaffolds, of which only one included 53 contigs while the remainders only 11 or less.

Dr. Nicola Vitulo aligned against the scaffolds a database of ESTs sequences. This approach provided a partial yet independent validation of 149 scaffolds. An example is reported in Figure 39.

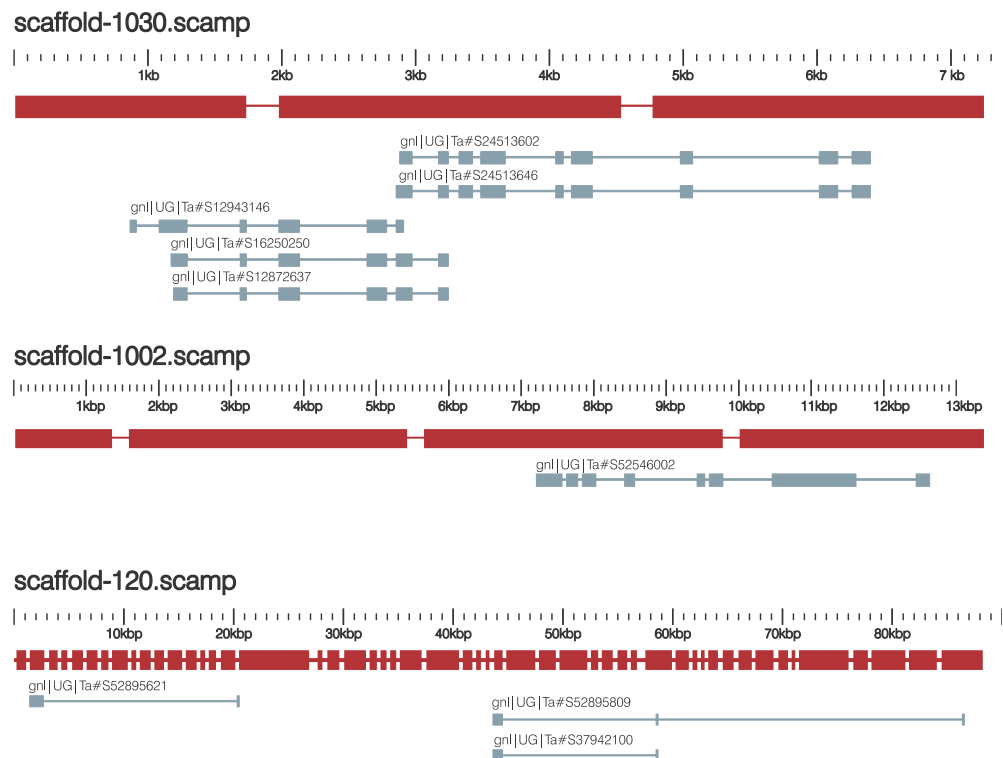


Figure 39

EST alignment against scaffolds made using datasets from Wheat, chromosome 5A.

4 Conclusion

It's difficult to underestimate the radical change in today's biology that came with the advent of NGS. When I joined the Genomics Group on 2007 the sequencing core hosted four Sanger sequencers (3730XL by Applied Biosystems) that were used to produce a 2X coverage of wine grape genome and it took more than a year – having the four machines operating at full capacity – to complete the shotgun sequencing, not to mention the high cost of consumables and operators to load the instruments.

With NGS sequencing the whole process from extracted DNA to sequences is straightforward, fast and much cheaper than in the past.

Shotgun sequencing for the *Nannochloropsis* genome required approximately two months. SOLiD MP libraries have been more time consuming but yielded an impressive coverage in about six months.

Comparing the assembly results from the sequencing of a full plate in 2009 with the Titanium kit and the half plate sequenced in 2011 with the XL+ kit we can appreciate the important advance in 454 sequencing both in terms of throughput and as average read length. For complex genomes this important advance is still not enough to produce a fairly assembled draft.

A. BENEFITS OF A MIXED APPROACH

Even though 454 sequencing costs are decreasing, they are still a bottleneck especially for larger genomes (>100 Mbp). In these projects (as for Wheat, \$3.5) a mixed approach is a cost effective sequencing strategy because the SOLiD MP libraries can provide a high coverage that can be used both for scaffolding and gap closure.

My project and the programs developed for it confirm the power of MP libraries in genome scaffolding and gap closure. Moreover when aligning local assemblies of MP reads against reference contigs, it has been evident that the SOLiD is more accurate in solving homopolymeric stretches. This suggest that the information content of MP libraries could be further exploited in the pipeline to remove small errors in reference.

B. SCAFFOLDING USING SOLiD MATE-PAIRED LIBRARIES

SOLiD MP libraries provide a valuable tool in genome sequencing. The protocol allows choosing the desired insert size and combining more insert size length can help overcome short and long repeated regions.

ScaMP, the program developed to produce scaffolds converting MP reads to directed connections between contigs is probably one of the first tools for genome scaffolding with color space reads, and addresses a need in the SOLiD community as emerged when presenting the whole pipeline at the “International SOLiD User Meeting” held in Treviso on August 2010.

Scaffolding with mate-paired reads has been proved to be effective even with a poor dataset (the low-coverage 454 Titanium made in 2009) for which it included one third of the genome into 20 scaffolds.

The highly fragmented contig dataset produced with the 454 Titanium kit (~14,000 contigs) combined with the two MP libraries gave good overall results: 77% of the sequenced genome was included in scaffold, 80 scaffold longer than 50 kbp and the N50 value of 323 kb.

ScaMP core algorithm seems valid and a paper is under preparation to release the program to the scientific community.

C. GAP CLOSURE

A remarkable advantage of using MP libraries in genome sequencing is the possibility to close gaps between contigs performing local assemblies of short MP reads.

The gap closure pipeline developed for this project can fill gaps between contigs in base space assembling selected subsets of color space reads, and Gap closure results in a raise of average contig length that is beneficial for downstream analysis as gene prediction and annotation, and it's possible to reduce the complexity of the task so that a standard desktop computer can perform it.

D. FUTURE PERSPECTIVES

The program has been developed and tested on a small genome with two MP libraries of comparable size, so no modeling of “bridge” size was implemented. It will be crucial, however, to have a correct modeling of

“bridges” size for larger genome making use of different MP libraries (e. g. for the Tomato genome project our group sequenced a 25 kbp MP library). Bridge creation starting from alignment result can be further strengthened modeling the distribution of mates alignment and comparing the model with actual alignments: most artifacts in bridges can be discriminated because of their uneven distribution along the contig.

The current gap closure pipeline produce contigs with Velvet using a high coverage of MP libraries, but they are used only to recover the missing portion of the genome laying between contigs, while they could be used also for error correction of Newbler contigs as the first appears to be more accurate, not only because of the higher coverage, but also because the SOLiD chemistry is less error prone in homopolymeric stretches.

5 Bibliography

1. Lander ES, Linton LM, Birren B, et al.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860–921.
2. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.** *Science* 1995, **269**:496–512.
3. Venter JC, Adams MD, Myers EW, et al.: **The sequence of the human genome.** *Science* 2001, **291**:1304–1351.
4. Mardis ER: **A decade's perspective on DNA sequencing technology.** *Nature* 2011, **470**:198–203.
5. Lander ES, Waterman MS: **Genomic mapping by fingerprinting random clones: a mathematical analysis.** *Genomics* 1988, **2**:231–239.
6. Pevzner PA: **An Eulerian path approach to DNA fragment assembly.** *Proceedings of the National Academy of Sciences* 2001, **98**:9748–9753.
7. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res.* 2008, **18**:821–829.
8. Zerbino D: **Genome assembly and comparison using de Bruijn graphs.** *Ph. D. thesis* 2009:1–164.
9. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors.** *Proc. Natl. Acad. Sci. U.S.A.* 1977, **74**:5463–5467.
10. Metzker ML: **Sequencing technologies — the next generation.** *Nat. Rev. Genet.* 2009, **11**:31–46.
11. **The Archon Genomics X PRIZE** [<http://genomics.xprize.org/competition-details/prize-overview>].
12. Williams R, Peisajovich SG, Miller OJ, Magdassi S, Tawfik DS, Griffiths AD: **Amplification of complex gene libraries by emulsion PCR.** *Nat. Methods* 2006, **3**:545–550.
13. Ronaghi M, Uhlén M, Nyren P: **A sequencing method based on real-time pyrophosphate.** *Science* 1998, **281**:363, 365.
14. Vezzi A, Campanaro S, D'Angelo M, Simonato F, Vitulo N, Lauro FM, Cestaro A, Malacrida G, Simionati B, Cannata N, Romualdi C, Bartlett DH, Valle G: **Life at depth: *Photobacterium profundum* genome sequence and expression analysis.** *Science* 2005, **307**:1459–1461.

15. Gouveia L, Oliveira AC: **Microalgae as a raw material for biofuels production.** *J. Ind. Microbiol. Biotechnol.* 2009, **36**:269–274.
16. Corteggiani Carpinelli E: *Going ultra deep to unravel the secret recipe of biofuel.* LAP LAMBERT Academic Publishing; 2011:1–129.
17. Andersen R, Brett R, Potter D: **Phylogeny of the Eustigmatophyceae based upon 18S rDNA, with emphasis on Nannochloropsis.** *Protist* 1998.
18. **Triacylglycerol profiling of marine microalgae by mass spectrometry** [<http://www.jlr.org/content/early/2011/08/11/jlr.D018408.short?rss=1>].
19. Boussiba S, Vonshak A, Cohen Z, Avissar Y: **Lipid and biomass production by the halotolerant microalga Nannochloropsis salina.** *Biomass* 1987.
20. **Cellular DNA content of marine phytoplaknton using two new fluorochromes: taxonomic and ecological implications** [<http://onlinelibrary.wiley.com/doi/10.1111/j.0022-3646.1997.00527.x/abstract>].
21. Campagna D, Albiero A, Bilardi A, Caniato E, Forcato C, Manavski S, Vitulo N, Valle G: **PASS: a program to align short sequences.** *Bioinformatics* 2009, **25**:967–968.
22. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.** *Nucleic Acids Res.* 2010, **38**:1767–1771.
23. Miller JR, Koren S, Sutton G: **Assembly algorithms for next-generation sequencing data.** *Genomics* 2010, **95**:315–327.
24. Mural RJ: **ARTEMIS: a tool for displaying and annotating DNA sequence.** *Brief. Bioinformatics* 2000, **1**:199–200.
25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J. Mol. Biol.* 1990, **215**:403–410.
26. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics.** *Genome Res.* 2009, **19**:1639–1645.
27. Stothard P, Wishart DS: **Circular genome visualization and exploration using CGView.** *Bioinformatics* 2005, **21**:537–539.
28. Koressaar T, Remm M: **Enhancements and modifications of primer design program Primer3.** *Bioinformatics* 2007, **23**:1289–1291.
29. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods Mol. Biol.* 2000, **132**:365–386.
30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup: **The Sequence**

Alignment/Map format and SAMtools. *Bioinformatics* 2009, **25**:2078–2079.

31. Dayarian A, Michael TP, Sengupta AM: **SOPRA: Scaffolding algorithm for paired reads via statistical optimization.** *BMC Bioinformatics* 2010, **11**:345.

32. *Applied Biosystems* [<http://www3.appliedbiosystems.com>].

33. Haridas S, Breuill C, Bohlmann J, Hsiang T: **A biologist's guide to de novo genome assembly using next-generation sequence data: A test with fungal genomes.** *Journal of Microbiological Methods* 2011, **86**:368–375.

34. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I: **ABYSS: A parallel assembler for short read sequence data.** *Genome Res.* 2009, **19**:1117–1123.

6 Supplementary material

Table 3

Chloroplast genes. For each predicted ORF the table indicate coordinates (start, end), gene symbol and strand of the ORF.

| Start | End | Gene | Strand |
|-------|-------|------------|--------|
| 1585 | 1998 | psbV | + |
| 2094 | 2360 | petJ | + |
| 2512 | 4755 | psaA | + |
| 4784 | 6991 | psaB | + |
| 7451 | 7546 | petG | - |
| 7602 | 7733 | psbK | - |
| 9097 | 9169 | trnW-CCA | + |
| 9320 | 9742 | rpl11 | + |
| 10974 | 11453 | petD | - |
| 11489 | 12133 | petB | - |
| 13356 | 13427 | trnM-CAU | + |
| 13446 | 13532 | trnS-GCU | + |
| 13588 | 13661 | trnD-GUC | + |
| 16043 | 16237 | psaE | + |
| 16344 | 16535 | psbH | - |
| 16736 | 16873 | psbN | + |
| 16916 | 17011 | psbT | - |
| 17029 | 18555 | psbB | - |
| 17035 | 18555 | psi_psbT | - |
| 18809 | 19105 | petF | + |
| 19557 | 20783 | tufA | - |
| 20974 | 21438 | rps7 | - |
| 21465 | 21716 | rps12_3end | - |
| 21720 | 21833 | rps12_5end | - |
| 21840 | 22061 | rpl31 | - |
| 25222 | 25332 | rpl36 | - |
| 29031 | 29393 | rpl14 | - |
| 29926 | 30321 | rpl16 | - |
| 31691 | 31972 | rps19 | - |
| 32007 | 32828 | rpl2 | - |
| 34923 | 36710 | dnaK | + |
| 36880 | 36946 | trnF-GAA | + |
| 36965 | 37035 | trnC-GCA | + |
| 36968 | 36996 | trnQ-UUG | + |
| 36980 | 37026 | trnF-GAA | + |
| 37066 | 37085 | trnL-GAG | + |
| 37067 | 37097 | trnL-CAA | + |
| 37852 | 38919 | psbD | + |
| 38831 | 40276 | psbC | + |
| 41867 | 42907 | ycf59 | - |
| 43168 | 43239 | trnN-GUU | - |
| 43846 | 45360 | chlB | - |
| 45464 | 45724 | ycf66 | - |
| 45930 | 46415 | psaF | - |
| 46533 | 46618 | trnI-CAU | - |
| 47099 | 47172 | trnP-GGG | + |
| 47100 | 47170 | trnP-UGG | + |
| 47137 | 47166 | trnM-CAU | + |
| 47989 | 49463 | rrn16 | + |
| 49567 | 49640 | trnI-GAU | + |
| 49644 | 49717 | trnA-UGC | + |

| | | | |
|---------------|--------|-----------|---|
| 50168 | 52676 | rrn23 | + |
| 53472 | 53798 | rpl20 | - |
| 54247 | 55947 | ilvB | + |
| 55996 | 56075 | trnY-GUA | - |
| 58082 | 58155 | trnG-UCC | + |
| 58219 | 58461 | psbE | + |
| 58467 | 58595 | psbF | + |
| 58627 | 58734 | psbL | + |
| 58813 | 58929 | psbJ | + |
| 59025 | 59117 | psaI | - |
| 60856 | 60927 | trnQ-UUG | - |
| 60988 | 61058 | trnR-ACG | - |
| 60993 | 61054 | trnR-CCG | - |
| 62283 | 62356 | trnH-GUG | + |
| 63477 | 63563 | petN | + |
| 64511 | 64581 | trnfM-CAU | + |
| 64667 | 65062 | psaD | + |
| 65157 | 65241 | trnS-UGA | + |
| 65387 | 65488 | psbI | + |
| 65592 | 65663 | trnV-UAC | + |
| 65666 | 65736 | trnR-UCU | + |
| 65961 | 66803 | chlL | + |
| 66885 | 68183 | chlN | + |
| 68415 | 69491 | psbA | + |
| 69789 | 70205 | rbcS | - |
| 70267 | 71721 | rbcL | - |
| 74673 | 76091 | atpB | - |
| 76235 | 76720 | ycf3 | - |
| 76995 | 77186 | rpl33 | - |
| 88333 | 89037 | atpI | + |
| 89120 | 89362 | atpH | + |
| 91277 | 92788 | atpA | + |
| 92834 | 92906 | trnE-UUC | - |
| 93251 | 93322 | trnG-GCC | - |
| 93298 | 93315 | trnN-GUU | - |
| 93429 | 93524 | psbY | - |
| 93648 | 93719 | trnK-UUU | - |
| 96026 | 97444 | ycf24 | - |
| 107883 | 108131 | rpl27 | + |
| 109373 | 109495 | psaJ | - |
| 109515 | 109757 | psaC | - |
| 110115 | 112623 | rrn23 | - |
| 113093 | 113166 | trnA-UGC | - |
| 113170 | 113243 | trnI-GAU | - |
| 113326 | 114800 | rrn16 | - |

Table 4

Scaffolds made with ScaMP using three reference datasets.

| Titanium (TITAN-09) | | | XL (XL-11) | | | Combined (BOTH) | | |
|---------------------|-----|-----------|---------------|-----|-------------|-----------------|-----|-------------|
| Scaffold name | # | Len. (bp) | Scaffold name | # | Length (bp) | name | # | Length (bp) |
| Scaffold190AT | 294 | 1,203,882 | contig00003 | 69 | 1,543,172 | 00003 | 59 | 1,546,607 |
| Scaffold246 | 201 | 871,519 | contig00002 | 88 | 1,355,202 | 00028 | 87 | 1,427,206 |
| Scaffold220AT | 158 | 798,265 | contig00012 | 74 | 1,348,784 | 00009 | 87 | 1,367,352 |
| 230AT | 141 | 753,407 | contig00007 | 103 | 1,129,217 | 00294 | 59 | 1,089,509 |
| Scaffold208AT | 148 | 691,365 | contig00001 | 70 | 1,068,942 | 00065 | 99 | 1,002,688 |
| Scaffold241 | 110 | 631,398 | contig00006 | 48 | 896,680 | 00022 | 62 | 968,996 |
| Scaffold156AT | 132 | 623,974 | contig00019 | 61 | 883,126 | 00005 | 13 | 925,566 |
| Scaffold171AT | 134 | 618,976 | contig00004 | 12 | 844,389 | 00203 | 138 | 920,669 |
| Scaffold245 | 154 | 546,432 | contig00087b | 46 | 800,132 | 00002 | 39 | 900,935 |
| Scaffold236 | 100 | 535,193 | contig00014 | 20 | 790,684 | 00020 | 52 | 812,876 |
| Scaffold8A | 187 | 503,332 | contig00005 | 57 | 762,108 | 00070 | 76 | 739,916 |
| Scaffold196B | 84 | 484,252 | contig01367 | 107 | 750,208 | 00010 | 80 | 702,347 |
| Scaffold243 | 115 | 465,417 | contig00029 | 84 | 734,505 | 00004 | 36 | 674,438 |
| Scaffold237 | 103 | 435,215 | contig00031 | 108 | 712,444 | 00018 | 75 | 624,626 |
| Scaffold195AT | 97 | 434,978 | contig00013 | 95 | 692,313 | 00301 | 51 | 622,381 |
| Scaffold235 | 91 | 407,200 | contig00065b | 64 | 630,078 | 00049 | 137 | 612,401 |
| Scaffold183AT | 102 | 398,199 | contig00038 | 61 | 602,110 | 00368 | 88 | 599,604 |
| Scaffold232 | 74 | 389,026 | contig00053 | 105 | 589,186 | 00277 | 59 | 588,168 |
| Scaffold231 | 72 | 348,377 | contig00033 | 93 | 544,704 | 00194 | 69 | 550,553 |
| Scaffold229 | 70 | 323,528 | contig00043 | 150 | 510,792 | 00015 | 79 | 548,095 |
| Scaffold234 | 89 | 317,878 | contig00040 | 19 | 474,513 | 00011 | 112 | 538,831 |
| Scaffold226 | 66 | 279,688 | contig00077 | 76 | 446,720 | 00053 | 50 | 505,214 |
| Scaffold223 | 59 | 259,869 | contig00028 | 44 | 442,654 | 00019 | 63 | 502,457 |
| Scaffold218 | 54 | 253,741 | contig00017 | 66 | 438,898 | 00064 | 101 | 477,448 |
| Scaffold233 | 74 | 253,034 | contig00171 | 50 | 430,089 | 00038 | 16 | 475,950 |
| Scaffold222 | 57 | 225,212 | contig00076 | 85 | 416,819 | 00236 | 49 | 440,935 |
| Scaffold213 | 49 | 221,076 | contig00024 | 17 | 379,540 | 00210 | 103 | 440,514 |
| Scaffold216 | 52 | 220,548 | contig00067 | 53 | 363,065 | 00084 | 97 | 432,671 |
| Scaffold71AT | 88 | 196,934 | contig01158 | 77 | 341,921 | 00048 | 45 | 389,720 |
| Scaffold211 | 47 | 195,940 | contig00102 | 34 | 313,442 | 00061 | 38 | 360,714 |
| Scaffold214 | 50 | 186,254 | contig00009 | 24 | 308,643 | 00091 | 89 | 332,530 |
| Scaffold225 | 64 | 174,942 | contig00062 | 56 | 296,458 | 00013 | 25 | 312,964 |
| Scaffold217 | 54 | 172,127 | contig00082b | 80 | 294,387 | 00014 | 12 | 303,392 |
| Scaffold210 | 47 | 168,478 | contig00073 | 54 | 294,127 | 00189 | 53 | 278,315 |
| Scaffold172 | 22 | 154,797 | contig00101 | 83 | 279,130 | 00036 | 3 | 249,716 |
| Scaffold68AT | 40 | 154,094 | contig00090 | 62 | 253,855 | 00112 | 34 | 237,271 |
| Scaffold198 | 35 | 150,476 | contig00604 | 74 | 210,468 | 00123 | 9 | 199,397 |
| Scaffold212 | 49 | 143,558 | contig01021 | 26 | 189,971 | 00081 | 29 | 192,629 |
| Scaffold86AT | 46 | 135,653 | contig00134 | 25 | 152,342 | 00192 | 30 | 168,090 |
| Scaffold215 | 51 | 132,562 | contig00237 | 33 | 147,552 | 00408 | 22 | 154,061 |
| Scaffold189 | 27 | 117,950 | contig00018 | 58 | 147,440 | 00138 | 30 | 131,888 |
| Scaffold185 | 25 | 116,066 | contig00328 | 69 | 141,445 | 00298 | 59 | 129,173 |
| Scaffold192 | 27 | 114,166 | contig00210 | 28 | 125,380 | 00145 | 14 | 128,153 |
| Scaffold205 | 40 | 113,818 | contig00266 | 53 | 121,919 | 00117 | 25 | 122,971 |
| Scaffold209 | 47 | 112,708 | contig00172 | 26 | 119,752 | 00418 | 57 | 112,951 |

| | | | | | | | | |
|--------------|----|---------|-------------|----|---------|--------|----|---------|
| Scaffold174 | 22 | 110,215 | contig00385 | 20 | 108,967 | CHL | 3 | 109,988 |
| Scaffold2 | 3 | 109,822 | contig00252 | 24 | 106,275 | 00050 | 2 | 101,062 |
| Scaffold207 | 44 | 105,480 | contig00456 | 15 | 98,778 | 00105 | 16 | 97,260 |
| Scaffold182 | 24 | 103,395 | contig00450 | 47 | 98,383 | 00313 | 37 | 86,687 |
| Scaffold204 | 40 | 97,401 | contig00335 | 29 | 94,539 | 00446 | 46 | 84,011 |
| Scaffold206 | 42 | 96,802 | contig00232 | 30 | 94,336 | 00197 | 13 | 83,930 |
| Scaffold200 | 35 | 96,733 | contig00319 | 32 | 88,964 | 00524 | 44 | 76,346 |
| Scaffold197 | 30 | 93,564 | chl | 2 | 87,356 | 00140 | 6 | 73,903 |
| Scaffold143 | 17 | 87,477 | contig00162 | 17 | 84,174 | 00217 | 7 | 69,968 |
| Scaffold184 | 25 | 83,377 | contig00248 | 10 | 74,864 | 00234 | 8 | 69,348 |
| Scaffold161 | 20 | 82,936 | contig00439 | 13 | 72,625 | 00579 | 23 | 66,290 |
| Scaffold201 | 42 | 81,842 | contig00340 | 32 | 71,671 | 00216 | 12 | 66,090 |
| Scaffold160 | 19 | 76,999 | contig00606 | 32 | 63,958 | 00338 | 15 | 62,095 |
| Scaffold187 | 26 | 76,864 | contig00507 | 17 | 63,177 | 00196 | 10 | 59,478 |
| Scaffold46AT | 32 | 74,882 | contig00433 | 19 | 61,365 | 00322 | 16 | 56,896 |
| Scaffold193 | 29 | 74,850 | contig00283 | 10 | 58,901 | 00299 | 23 | 56,397 |
| Scaffold130 | 15 | 74,819 | contig00673 | 21 | 58,632 | 00153 | 4 | 50,923 |
| Scaffold170 | 22 | 72,856 | contig00382 | 13 | 58,542 | 00339 | 17 | 50,553 |
| Scaffold152 | 18 | 71,862 | contig00537 | 38 | 57,010 | 00048b | 4 | 49,009 |
| Scaffold199 | 35 | 70,618 | contig00363 | 15 | 54,243 | 00747 | 32 | 45,716 |
| Scaffold133 | 15 | 70,118 | contig00349 | 22 | 51,906 | 00572 | 21 | 44,871 |
| Scaffold180 | 24 | 67,438 | contig00368 | 15 | 48,084 | 00365 | 15 | 43,472 |
| Scaffold186 | 25 | 66,767 | contig01082 | 27 | 45,519 | 00308 | 16 | 42,938 |
| Scaffold121 | 14 | 65,549 | contig00858 | 19 | 45,290 | MIT | 1 | 42,216 |
| Scaffold181 | 24 | 64,116 | contig00659 | 20 | 40,896 | 00017 | 11 | 41,382 |
| Scaffold163 | 20 | 63,234 | contig00613 | 25 | 40,439 | 00582 | 23 | 39,595 |
| Scaffold176 | 22 | 63,017 | contig00463 | 25 | 37,277 | 00024 | 18 | 37,961 |
| Scaffold145 | 17 | 62,927 | contig01401 | 13 | 35,662 | 00332 | 7 | 36,837 |
| Scaffold1CC | 25 | 61,200 | contig00418 | 12 | 32,954 | 00219 | 12 | 35,247 |
| Scaffold151 | 18 | 60,578 | contig00487 | 11 | 21,044 | 00384 | 9 | 30,847 |
| Scaffold414A | 32 | 56,827 | contig01000 | 11 | 16,300 | 00474 | 11 | 27,931 |
| Scaffold147 | 18 | 56,305 | contig00826 | 4 | 15,535 | 00508 | 14 | 25,937 |
| Scaffold153 | 18 | 55,793 | contig00751 | 8 | 14,984 | 00399 | 4 | 23,504 |
| Scaffold175 | 22 | 52,756 | contig00585 | 6 | 14,696 | 00433 | 10 | 22,664 |
| Scaffold165 | 21 | 52,679 | contig00802 | 8 | 14,517 | 00647 | 14 | 21,941 |
| Scaffold167 | 21 | 52,536 | contig05405 | 6 | 9,692 | 00573 | 8 | 14,979 |
| Scaffold194 | 29 | 52,133 | contig00661 | 4 | 8,292 | 00706 | 7 | 14,172 |
| Scaffold158 | 19 | 51,593 | contig00842 | 7 | 7,594 | 00001 | 8 | 7,508 |
| Scaffold136 | 16 | 49,584 | contig00078 | 10 | 5,555 | 00708 | 4 | 7,289 |
| Scaffold149 | 18 | 49,489 | contig00872 | 3 | 5,514 | 00696 | 1 | 6,304 |
| Scaffold138 | 15 | 49,118 | contig00063 | 10 | 5,401 | 00074 | 13 | 5,537 |
| ScaffoldA007 | 29 | 48,614 | contig00508 | 4 | 3,346 | 00030 | 8 | 2,741 |
| Scaffold177 | 23 | 48,109 | contig01457 | 4 | 3,266 | 00007 | 7 | 2,160 |
| Scaffold148 | 18 | 46,318 | contig00042 | 4 | 2,854 | 00264 | 7 | 2,160 |
| Scaffold45 | 8 | 46,240 | contig00010 | 7 | 2,182 | 00067 | 3 | 1,805 |
| Scaffold134 | 15 | 46,133 | | | | 00032 | 4 | 1,785 |
| Scaffold89 | 10 | 45,045 | | | | 00034 | 4 | 1,748 |
| Scaffold356 | 5 | 44,976 | | | | 00600 | 2 | 1,187 |
| Scaffold139 | 16 | 44,975 | | | | 00006 | 1 | 248 |
| Scaffold188 | 26 | 44,833 | | | | | | |
| Scaffold123 | 14 | 43,563 | | | | | | |

| | | | | | |
|--------------|----|--------|--|--|--|
| Scaffold117 | 13 | 43,209 | | | |
| Scaffold1 | 3 | 41,496 | | | |
| Scaffold91 | 11 | 40,465 | | | |
| Scaffold155 | 19 | 40,359 | | | |
| Scaffold110 | 12 | 40,041 | | | |
| Scaffold85 | 10 | 39,091 | | | |
| Scaffold178 | 23 | 38,855 | | | |
| Scaffold84 | 10 | 38,784 | | | |
| Scaffold131 | 15 | 38,648 | | | |
| Scaffold157 | 19 | 37,171 | | | |
| Scaffold109 | 12 | 36,396 | | | |
| Scaffold135 | 15 | 35,707 | | | |
| ScaffoldA005 | 42 | 35,674 | | | |
| Scaffold100 | 11 | 35,604 | | | |
| Scaffold112 | 13 | 34,974 | | | |
| Scaffold126 | 15 | 34,584 | | | |
| Scaffold179 | 24 | 33,963 | | | |
| Scaffold93 | 11 | 33,865 | | | |
| Scaffold164 | 20 | 33,735 | | | |
| Scaffold20 | 6 | 33,536 | | | |
| Scaffold132 | 15 | 33,464 | | | |
| Scaffold168 | 20 | 32,802 | | | |
| Scaffold142 | 17 | 32,630 | | | |
| Scaffold98 | 11 | 31,954 | | | |
| Scaffold119 | 14 | 31,427 | | | |
| Scaffold36 | 7 | 30,870 | | | |
| Scaffold150 | 18 | 30,666 | | | |
| Scaffold122 | 14 | 30,130 | | | |
| Scaffold137 | 16 | 28,702 | | | |
| Scaffold411a | 15 | 28,549 | | | |
| Scaffold159 | 19 | 28,421 | | | |
| Scaffold81 | 9 | 27,961 | | | |
| Scaffold104 | 12 | 27,557 | | | |
| Scaffold124 | 14 | 27,469 | | | |
| Scaffold95 | 11 | 27,398 | | | |
| ScaffoldA512 | 13 | 26,391 | | | |
| Scaffold154 | 18 | 26,171 | | | |
| Scaffold101 | 12 | 26,046 | | | |
| Scaffold129 | 15 | 25,766 | | | |
| Scaffold70 | 9 | 25,425 | | | |
| Scaffold113 | 13 | 24,691 | | | |
| Scaffold146 | 17 | 24,198 | | | |
| Scaffold141 | 16 | 24,040 | | | |
| Scaffold82 | 10 | 23,936 | | | |
| Scaffold62 | 8 | 23,839 | | | |
| Scaffold80 | 9 | 23,603 | | | |
| ScaffoldA006 | 15 | 23,171 | | | |
| ScaffoldA004 | 13 | 22,774 | | | |
| Scaffold359 | 5 | 22,766 | | | |
| Scaffold115 | 13 | 21,757 | | | |
| Scaffold128 | 15 | 21,189 | | | |
| Scaffold108 | 12 | 21,034 | | | |
| Scaffold96 | 11 | 20,867 | | | |
| Scaffold28 | 7 | 20,848 | | | |
| Scaffold361 | 5 | 20,824 | | | |

| | | | | | |
|--------------|----|--------|--|--|--|
| ScaffoldA010 | 10 | 20,765 | | | |
| Scaffold118 | 13 | 20,533 | | | |
| Scaffold390 | 3 | 20,377 | | | |
| Scaffold106 | 12 | 20,363 | | | |
| Scaffold116 | 13 | 20,173 | | | |
| ScaffoldA009 | 19 | 20,142 | | | |
| Scaffold103 | 12 | 19,974 | | | |
| Scaffold111 | 13 | 19,916 | | | |
| Scaffold56 | 8 | 19,798 | | | |
| Scaffold88 | 10 | 19,728 | | | |
| Scaffold47 | 8 | 19,589 | | | |
| Scaffold77 | 9 | 19,534 | | | |
| ScaffoldA001 | 9 | 19,069 | | | |
| Scaffold107 | 12 | 18,905 | | | |
| Scaffold392 | 4 | 18,882 | | | |
| Scaffold99 | 11 | 18,716 | | | |
| Scaffold90 | 10 | 18,344 | | | |
| Scaffold60 | 8 | 18,278 | | | |
| Scaffold76 | 9 | 18,087 | | | |
| Scaffold127 | 15 | 18,049 | | | |
| Scaffold94 | 11 | 18,027 | | | |
| Scaffold87 | 10 | 17,737 | | | |
| Scaffold72 | 9 | 17,682 | | | |
| Scaffold79 | 9 | 17,674 | | | |
| Scaffold114 | 13 | 17,576 | | | |
| Scaffold33 | 7 | 17,532 | | | |
| Scaffold65 | 9 | 17,127 | | | |
| Scaffold43 | 7 | 17,101 | | | |
| Scaffold83 | 10 | 16,619 | | | |
| Scaffold41 | 7 | 16,532 | | | |
| Scaffold323 | 6 | 16,357 | | | |
| Scaffold105 | 12 | 16,242 | | | |
| Scaffold58 | 8 | 16,125 | | | |
| Scaffold49 | 8 | 15,961 | | | |
| Scaffold92 | 11 | 15,637 | | | |
| Scaffold35 | 7 | 15,611 | | | |
| Scaffold353 | 5 | 15,532 | | | |
| Scaffold97 | 11 | 14,790 | | | |
| Scaffold40 | 7 | 14,751 | | | |
| Scaffold336 | 5 | 14,626 | | | |
| Scaffold6 | 6 | 14,568 | | | |
| Scaffold350 | 5 | 14,401 | | | |
| Scaffold368 | 4 | 14,329 | | | |
| ScaffoldA002 | 6 | 14,160 | | | |
| Scaffold120 | 14 | 14,114 | | | |
| Scaffold332 | 5 | 14,023 | | | |
| Scaffold51 | 8 | 13,917 | | | |
| Scaffold366 | 4 | 13,847 | | | |
| Scaffold25 | 7 | 13,270 | | | |
| Scaffold351 | 5 | 13,175 | | | |
| Scaffold385 | 4 | 13,172 | | | |
| Scaffold52 | 8 | 12,946 | | | |
| Scaffold400 | 4 | 12,912 | | | |
| Scaffold73 | 9 | 12,873 | | | |
| Scaffold102 | 12 | 12,826 | | | |

| | | | | | |
|--------------|---|--------|--|--|--|
| ScaffoldA003 | 7 | 12,728 | | | |
| Scaffold287 | 6 | 12,664 | | | |
| Scaffold456 | 2 | 12,645 | | | |
| Scaffold24 | 7 | 12,597 | | | |
| Scaffold379 | 4 | 12,308 | | | |
| Scaffold38 | 7 | 12,221 | | | |
| Scaffold372 | 4 | 12,023 | | | |
| Scaffold203 | 8 | 11,991 | | | |
| Scaffold395 | 4 | 11,860 | | | |
| Scaffold32 | 7 | 11,750 | | | |
| Scaffold347 | 5 | 11,278 | | | |
| Scaffold5 | 6 | 11,193 | | | |
| Scaffold23 | 7 | 11,117 | | | |
| Scaffold61 | 8 | 11,014 | | | |
| Scaffold460 | 4 | 10,989 | | | |
| Scaffold16 | 6 | 10,911 | | | |
| Scaffold373 | 4 | 10,727 | | | |
| Scaffold469 | 2 | 10,727 | | | |
| Scaffold29 | 7 | 10,624 | | | |
| Scaffold18 | 6 | 10,562 | | | |
| Scaffold357 | 5 | 10,393 | | | |
| Scaffold17 | 6 | 10,227 | | | |
| Scaffold399 | 4 | 10,165 | | | |
| Scaffold358 | 5 | 10,105 | | | |
| Scaffold31 | 7 | 10,098 | | | |
| Scaffold37 | 7 | 10,084 | | | |

Table 5

Superscaffolds composition in terms of scaffold number and their mutual orientation (C = Complemented, U = Uncomplemented).

| Superscaffold | Scaffolds | Orientation |
|---------------|-----------|-------------|
| 1 | 169 | C |
| | 351 | C |
| | 368 | U |
| | 233 | C |
| | 356 | U |
| | | |
| 2 | 176 | C |
| | 495 | U |
| | 197 | C |
| | 235 | U |
| | 226 | U |
| | 20 | C |
| | 56 | U |
| | | |
| 3 | 366 | U |
| | 210 | C |
| | 174 | U |
| | 232 | C |
| | 41 | C |
| | 166 | U |
| | | |
| 4 | 68AT | C |
| | 377 | U |
| | 135 | C |
| | | |
| 5 | 204 | C |
| | 218B | C |
| | 196B | C |
| | 211A | U |
| | 187 | C |
| | 143 | U |
| | | |
| 6 | 185 | U |
| | 213 | C |
| | 189 | C |
| | 218A | U |
| | 241 | C |
| | 207 | U |
| | 100 | C |
| | | |
| 7 | 238 | U |
| | 202 | U |
| | | |
| 8 | 165 | U |
| | 148 | U |
| | 43 | C |
| | 13 | U |
| | 214 | U |
| | 133 | C |
| | 216 | U |
| | 113 | U |
| | | |
| 9 | 212 | U |
| | 171AT | U |
| | | |
| 10 | 208AT B | U |
| | 47 | U |
| | 136 | U |
| | 122 | C |
| | 86AT | C |

| | | |
|-----------|---------|---|
| | | |
| 11 | 70 | C |
| | 61 | C |
| | 206 | U |
| | 223 | C |
| 12 | 126 | U |
| | 373 | U |
| | 88 | U |
| | 186 | U |
| | 110 | U |
| | 198 | C |
| 13 | 236 | C |
| | 172 | C |
| | 183AT | U |
| | 215 | C |
| 14 | 112 | C |
| | 192 | C |
| 15 | 405 | U |
| | 336 | C |
| | 352 | C |
| | 231 | U |
| | 368 | C |
| | 89 | C |
| | 205 | U |
| 16 | 201 | C |
| | 243 | U |
| 17 | 184 | C |
| | 181 | U |
| 18 | 39 | C |
| | 195AT | C |
| 19 | 170 | U |
| | 359 | C |
| 20 | 106 | U |
| | 400 | C |
| | 323 | C |
| | 24 | C |
| | 208AT A | U |
| 21 | 237 | U |
| | 379 | C |
| | 161 | U |
| 22 | 17 | C |
| | 116 | U |
| | 139 | C |
| | 384 | U |
| | 217 | U |
| 23 | 211 B | C |
| | 199 | C |

Table 6

Alignment statistics of the two MP libraries against the three Newbler assemblies.

Table 4

Alignment statistics: SOLiD mate-paired libraries aligned against Newbler contigs (three datasets: Titanium, XL and the two combined).

Mate-Paired reads aligned against Newbler datasets

| Dataset / Reads | Total reads | Filtered | (%) | Remaining | Aligned | (%) | Unique | (%) |
|------------------|-------------|------------|-----|------------|------------|-----|------------|-----|
| XL-2011 | | | | | | | | |
| 1.5 - 3.0 kb For | 68,876,674 | 9,868,642 | 14% | 59,008,032 | 40,270,376 | 58% | 38,381,784 | 95% |
| 1.5 - 3.0 kb Rev | 68,881,266 | 11,373,759 | 17% | 57,507,508 | 42,859,202 | 62% | 40,893,142 | 95% |
| 3.0 - 5.0 kb For | 76,457,240 | 9,558,954 | 13% | 66,898,286 | 45,894,168 | 60% | 44,229,271 | 96% |
| 3.0 - 5.0 kb Rev | 76,692,245 | 8,933,333 | 12% | 67,758,912 | 51,669,558 | 67% | 49,949,338 | 97% |
| TITANIUM-2009 | | | | | | | | |
| 1.5 - 3.0 kb For | 68,399,694 | 13,364,758 | 20% | 55,034,936 | 35,694,271 | 52% | 34,014,883 | 95% |
| 1.5 - 3.0 kb Rev | 68,876,674 | 13,217,262 | 19% | 55,659,412 | 38,529,736 | 56% | 36,594,619 | 95% |
| 3.0 - 5.0 kb For | 74,749,807 | 13,348,884 | 18% | 61,400,924 | 40,584,897 | 54% | 39,163,389 | 96% |
| 3.0 - 5.0 kb Rev | 74,932,512 | 10,534,613 | 14% | 64,397,899 | 46,326,028 | 62% | 44,819,323 | 97% |
| BOTH | | | | | | | | |
| 1.5 - 3.0 kb For | 68,876,674 | 9,868,642 | 14% | 59,008,032 | 40,914,908 | 59% | 38,998,161 | 95% |
| 1.5 - 3.0 kb Rev | 68,399,694 | 13,847,508 | 20% | 54,552,186 | 40,480,384 | 59% | 38,447,298 | 95% |
| 3.0 - 5.0 kb For | 74,749,807 | 7,699,419 | 10% | 67,050,388 | 47,402,218 | 63% | 45,741,930 | 96% |
| 3.0 - 5.0 kb Rev | 74,932,512 | 8,016,680 | 11% | 66,915,832 | 52,529,133 | 70% | 50,820,679 | 97% |