Università degli Studi di Padova

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Biologia

SCUOLA DI DOTTORATO DI RICERCA IN BIOSCIENZE E BIOTECNOLOGIE

INDIRIZZO: BIOLOGIA EVOLUZIONISTICA CICLO XXV

Gene expression study in the non-model organism Botryllus schlosseri through SOLiD RNA-seqs

Direttore della Scuola : Ch.mo Prof. Giuseppe Zanotti

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

Supervisore: Ch.mo Prof. Loriano Ballarin

Co-supervisore: Ch.mo Prof. Giorgio Valle

Dottorando : Davide Campagna

INDEX

ABSTRACT

Botryllus schlosseri is a colonial ascidian widespread in temperate, shallow seas of the world. The organism is widely used for the study, in an evolutionary perspective, of a variety of biological processes ranging from sexual and asexual reproduction to regeneration, allorecognition and immune responses. However, despite its importance as model organism, no sequenced genome is available today. We undertook the analysis of the transcriptome of *B. schlosseri* in various colonial developmental phases and during tunic regeneration. The asexual reproduction by continuous palleal budding and the vascular system regeneration were the target of our RNA-seq experiments. In the first experiment, 3 different phases of the colonial blastogenetic cycle were considered: the mid-cycle, where colonies are metabolically very active; a phase immediately before the generation change (take-over), where colonies are getting ready to the generation change, and the take-over phase where adult zooids die and are replaced by buds which reach adulthood. Total RNA was extracted from various colonies for each of considered experimental condition. In the second experiment, the tunic of some colonies was cut and let to regenerate for 2 days. After tunic regeneration, total RNA was extracted. cDNA libraries were built according to SOLiD protocols and they were sequenced using SOLiD 4 and SOLID 5500 sequencers.

In the absence of a reference genome, the gene expression analysis requires a *de novo* assembly of RNA-seq experiments. In this thesis a method to assemble RNAseq reads produced by SOLiD sequencers is described for the first time. The analysis of simulated data allowed us to improves the overall method.

Gene expression data and gene annotation have been stored in a database and they can be managed in a compact structure which is directly and quickly accessed by a developed Web interface. The Web interface makes possible the analysis of many experimental conditions and their comparison, to highlight

expression differences, through a common Web browser. Many thousands of differentially expressed transcripts were found and some of these are involved in natural apoptosis. This biological process was described in details using morphological studies: during take-over, tissues of adult zooids undergo apoptosis and zooids are replaced by primary buds that grow to become the new adult generation.

RIASSUNTO

Botryllus schlosseri è un ascidia coloniale diffusa in tutti i mari temperati del mondo. L'organismo è oggi ampiamente usato per lo studio, in prospettiva evolutiva, di un'ampia varietà di processi biologici che vanno dalla riproduzione sessuale e asessuale alla rigenerazione, all'alloriconosciemnto e alle risposte immunitarie. Tuttavia, nonostante la sua importanza come organismo modello, non è, a tutt'oggi, disponibile il genoma sequenziato. Abbiamo intrapreso uno studio del trascrittoma di *B. schlosseri* in diverse fasi di sviluppo coloniale e durante la rigenerazione della tunica. La riproduzione asessuale mediante gemmazione palleale e la rigenerazione del sistema vascolare sono stati i bersagli dei nostri esperimenti di RNA-seq. Nel primo esperimento, 3 diverse fasi del ciclo blastogenetico coloniale sono state prese in considerazione: il “mid-cycle”, quando le colonie sono molto attive metabolicamente; una fase immediatamente precedente il cambio di generazione (take-over), quando le colonie si stanno preparando al cambio di generazione, e la fase di “take-over”, quando gli zoidi adulti muoiono e vengono sostituiti dalle gemme che diventano funzionalmente mature. L'RNA totale è stato estratto da diverse colonie in ciascuna delle condizioni sperimentali considerate. Nel secondo esperimento, la tunica di alcune colonie è stata asportata marginalmente e lasciata rigenerare per 2 giorni. Dopo la rigenerazione, l'RNA totale è stato estratto. Librerie di cDNA sono state ottenute seguendo i protocolli SOLiD e sono state sequenziate usando sequenziatori SOLiD 4 e SOLID 5500.

In mancanza di un genoma di riferimento è stato necessario produrre un assemblaggio del trascrittoma. In questa tesi viene descritto il primo metodo per assemblare dati di esperimenti RNA-seq adattato alla tecnologia di sequenziamento SOLiD. Le analisi condotte sui dati simulati hanno permesso di ottimizzare il metodo e minimizzare l'effetto delle isoforme di trascritti sulla qualità dell'assemblaggio.

I dati di espressione e di annotazione genica sono stati archiviati in un database ad accesso rapido e gestiti in una struttura compatta direttamente accessibile tramite un'interfaccia Web. Attraverso l'uso di un comune Web browser, è possibile analizzare le diverse condizioni sperimentali e comparare i risultati ottenuti. I risultati hanno rivelato molte migliaia di trascritti differenzialmente espressi dei quali alcuni sono coinvolti nell'apoptosi naturale. Questo processo biologico è stato descritto a livello morfologico: durante il take-over, i tessuti degli zoidi vanno in apoptosi e questi ultimi sono sostituiti dalle gemme primarie che divengono il nuovo stadio adulto.

1. INTRODUCTION

1.1 BACKGROUND

The transcriptome is a subset of active genes in tissues of selected species. Understanding the transcriptome dynamics is essential to interpret the phenotypic variation caused by the combination of genotypic and environmental factors. Massive parallel sequencing of RNA has made possible to characterize the transcriptome with unprecedented sensitivity and depth, revolutionizing gene expression study. High-throughput sequencing (also called next-generation sequencing) technologies are effective for both cost and work, so the range of studied organisms is expanding. The RNA sequencing (RNAseq) of non-model organisms can provide new insights into the mechanisms underlying the diversity of life in our planet. In animals and plants, the “innovations” that cannot be examined in the common model organisms, include mimicry, mutualism, parasitism, and asexual reproduction.

During the last years a novel bioinformatics method has allowed a *de novo* transcriptome assembly directly by sequenced data [Birol et al. 2009). When the reference-based methods are not possible this is the preferred method for the study of non-model organisms.

Species of the class Ascidiacea (phylum Chordata, subphylum Urochordata or Tunicata) are sessile animals, spread all over the word, especially in shallow, tropical and temperate waters. Approximately 3,000 species have been reported so far, both solitary and colonial. In recent years, *Ciona intestinalis, Ciona savignyi, Halocyinthia roretzi* has emerged as model organisms for the study of embryogenesis and differentiation of specific cell lines and their genome has been partially or completely sequenced.

Although colonial ascidians have been poorly studied at molecular level, they offer the opportunity to compare, in the same organism at various levels (morphological, biochemical and molecular), different pathways of development (embryogenesis, regeneration and blastogenesis). It is very interesting to note that, among the chordates, the great capacity of regeneration and asexual reproduction was maintained only in tunicates.

In the compound ascidian *Botryllus schlosseri*, palleal budding occurs continuously in a colony, in an ordered and synchronized way, and cyclical changes of adult generations occurs. The interval of time from one generation change to the next one is defined as the colonial blastogenetic cycle.

In the above species, we undertook RNAseq experiments with the aim to get information on the transcriptome of colonies at different phases of the blastogenetic cycle and put in evidence differently-transcribed genes. The lack of a reference genome oriented us towards a *de novo* transcriptome assembly.

Unlike sequence coverage levels of a genome, which can vary randomly as a result of repeat content in non-coding DNA intron regions, transcripts sequence coverage levels can be directly indicative of gene expression. The repeated sequences create ambiguities in the formation of contigs in genome assembly, while ambiguities in transcriptome assembly usually correspond to spliced isoforms, or minor variation among members of a gene family. These problems must be addressed by bioinformatics analysis and have a central role in this project.

1.2 PHYLOGENETIC CONTEXT OF ASCIDIANS

As reported before, ascidians belong to the phylum Chordata, sub-phylum Urochordata (Tunicata). The chordates represent the largest deuterostomes phylum; they have bilateral symmetry and share at least four main features: i) the permanent or temporary presence of the notochord in the form of a backbone, rod-like structure that prevents the elastic shortening of the body when longitudinal muscles contract, ii) a dorsal hollow neural tube, slightly enlarged in the front end, iii) a ventral bowel that form, a pharynx at its anterior end, provided with gill slits or pharyngeal pouches and a ventral glandular structure capable of fixing iodine (endostyle/thyroid), iv) one muscular tail (post-anal part of the body).

The majority of chordates belongs to the sub-phylum vertebrates (around 47000 species), while the invertebrate Chordata represent about 3% of the species of this Phylum and are commonly grouped under the name of Protochordates which includes Urochordata (Tunicates) and Cephalochordata. Unlike vertebrates that are widely spread on land, fresh water and sea water and are either predators or herbivores, protochordates are marine filter feeders and many of them are characterized by a sedentary lifestyle.

Tunicates or Urochordata are marine filter feeders, solitary or colonial, benthic or pelagic. Normally they have a larval muscular tail equipped with notochord (hence the name Urochordata) and a dorsal neural tube present only in larval stages. The pharynx is well developed in the adult and normally occupies most of the volume of the body. The body is covered by the tunic that gives its name to the sub-phylum. This coating consists of an amorphous matrix which contains cells, rich in water, salts and fibrous components, with tunicine (a polysaccharide similar to cellulose) as main component, and proteins cross-linking tunicine fibers. The tunic, the consistency of which varies from soft tough, is produced by the epidermis and is frequently crossed by blood vessels; some circulatory cells, are probably involved in its synthesis. The tunic anchors the animal to the substrate, and provides both protection and support. In some species it contains calcareous secretions (spicules) of various shapes. Tunicates are traditionally divided in three classes: Ascidiacea, Thaliacea and Larvacea.

Ascidians are sessile, marine invertebrates found throughout the world in the shallow waters of temperate and tropical seas. They include about 3000 species with solitary and colonial forms. Usually, solitary individuals (zooids) are larger (up to 20 cm in length) than colonial ones. The latter can form very large colonies consisting of hundreds of individuals who share the same tunic and may be interconnected by a common circulatory system. All ascidians are simultaneous hermaphrodites, proterandrous or proterogynic. On the basis of the position of the gonads, we can define two ascidian orders: Enterogona and Pleurogona [Burighel and Cloney, 1997]. Enterogone ascidians have the gonads in close association (inside or behind) with the intestinal loop, far from the atrial siphon and equipped with a long gonoduct. In these species, sperm and eggs can be stored in gonoducts before being released. In the Pleurogone ascidians the gonads are located on the side of the body wall and are equipped with short gonoducts.

Cross-fertilization is the rule and the swimming, tadpole-like larva emerges from ovular envelopes at the end of embryo-genesis. Metamorphosis takes place after a period ranging from few minutes to few days and is preceded by the inversion of tropisms. The tail and part of the larval nervous system are reabsorbed and the oozooid begins to feed once the siphons open [Burighel and Cloney, 1997].

Ascidians have a well-developed open circulatory system. The blood circulates inside lacunae, in the cavity of the mantle and in the pharyngeal wall, which are extensions of the epidermal tunic and inside vessels lined by a thin epithelial layer, in the tunic. In some species (e.g.: *B. schlosseri*) blind, sac-like endings, named ampullae, emerge from peripheral vessels . The heart is a double-walled tube with an outer pericardium and an internal myocardium, arising from the introflection of the outer layer. The two layers define a thin pericardial cavity considered the remainder of the coelomic cavity of ascidians. The heartbeat occurs in the form of contractile waves that start at one end of the myocardium. A peculiarity of the ascidian heart is that it reverses its beat every 2-3 minutes.

The blood of ascidians consists of a colorless plasma, isotonic with seawater and different cell types (hemocytes) that can be grouped into at least four categories: 1. undifferentiated cells; 2. phagocytes (hyaline amoebocytes and macrophage-like cells, uni or multi-vacuolated); 3. vacuolated cytotoxic cells (granular amoebocytes and morula cells); 4. Storage cells (pigment cells and nephrocytes). Blood plays a minor role in gas exchange and does not contain respiratory pigments as many exchanges can take place by simple diffusion.

1.3 Botryllus schlosseri

B. schlosseri is a cosmopolitan colonial ascidian which form encrusting colonies. Colony-forming zooids are organized in star shaped systems of 6-12 zooids completely enveloped by the common tunic. Each zooid is provided with two siphons: the oral siphon located at the periphery of the system body side and the atrial siphon that opens in a common cloacal opening at the center of each system. While individual zooids are long from 1 to 1.5 mm, a colony can be formed by many systems and cover the surfaces of several square centimeters. The common tunic is covered by a network of blood vessels that interconnect the zooids and ensure the maintenance at the same developmental stage. The hemolymph passes from the heart to two large sinuses, the dorsal sinus and the endostilar sinus, and reaches the vessels of the tunic and the numerous ampullae.

The blood of *B. schlosseri*, like other ascidians, includes numerous types of hemocytes that, according to traditional classifications [Goodbody et al. 1974] and [Wright, 1981], integrated with other studies [Rowley et al., 1984; Ballarin et al., 1994 , 1998, Cima et al., 2001; Ballarin and Cima, 2005], can be grouped into three categories:

- lymphocyte-like cells;
- immunocytes (including: i) phagocytes, both hyaline amoebocytes and macrophage-like cells; ii) cytotoxic cells, i.e., morula cells and their precursors), granular amebocytes.
- storage cells, which include pigment cells and nephrocytes.

Lymphocyte-like cells owe their name to the morphological similarity with lymphocytes of vertebrates. They have a spherical shape with a diameter of 3-5 µm and a nucleus that occupies most of the cell volume. Around the nucleus there is a thin ring of hyaline cytoplasm in which few mitochondria, polyribosomes and the endoplasmic reticulum are immersed [Wright, 1981]. It is widely believed that all other blood cells derive from this cell type [Sabbadin, 1955] since, in this species, no haemopoietic organs, comparable to the lymph nodes on the wall of the gills of some solitary ascidian [Ermak, 1976], are present. Recent data indicates positive labeling of these cells to the anti-CD34 antibody, raised against a vertebrate stem cells antigen [Cima et al., 2001; Ballarin and Cima, 2005], in agreement with the putative role of the circulating undifferentiated cells.

The hyaline amebocytes have variable shape: their size ranges between 5 and 10 µm and are characterized by a round nucleus and a cytoplasm rich in small granules that cannot be resolved by optical microscope. They contain many ribosomes, a well-developed Golgi apparatus and glycogen granules [Milanesi and Burighel, 1978]. These cells represent the active phagocytes, capable of amoeboid movements and ingesting foreign material. They are also involved in the processes of coagulation and transport of nutrients [Goodbody, 1974]. The macrophage-like cells are very large cells (diameter between 10 and 15 µm), that have a variable shape and an ovoidal nucleus. Organelles are scarce, lysosomes are present in the cytoplasm as well as large digestive vacuoles containing phagocytized material. These cells represent the terminal stage of the differentiation of phagocytes that, after ingesting foreign material, withdraw the cytoplasmic extensions and assume a spherical shape [Goodbody, 1974; Wright, 1981; Ballarin et al., 1994].

Morula cells (MC) have a spherical shape with a diameter of 8-16 µm. They have numerous vacuoles of 2-3 um of diameter that gives a lumpy appearance to these cells, once fixed. Reducing substances, various oxidative enzymes, including the pro-enzyme pro-phenoloxidase, and polyphenols, which act as substrates for the latter enzyme, are present within the vacuoles. In various ascidians, MC are active in the defence reactions and can cause the lysis of foreign cells [Parrinello, 1996; Cammarata et al., 1997; Ballarin et al., 1993, 1994, 1995, 1998] and the encapsulation of foreign bodies [Anderson, 1971]. *B. schlosseri* MC are also involved in the formation of the necrotic masses which characterize the process of rejection between incompatible colonies [Taneda and Watanabe, 1982a, b, c; Scofield and Nagashima, 1983; Sabbadin et al., 1992; Ballarin el al. 1995.1998, 2002b; Rinkevich et al., 1998; Hirose et al., 2002; Cima et al., 2004]. They also have other functions, such as tissue repair, transport of nutrients, accumulation of iron within the vacuoles.

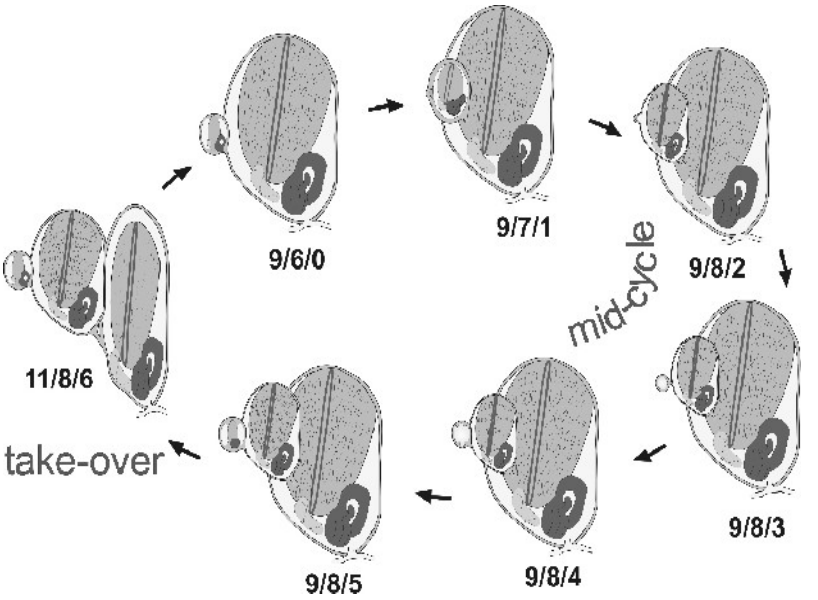
Granular amoebocytes contain small spherical vacuoles (size 0.4 to 1.5 µm) and are believed to be the MC precursors with which they share many features of the vacuolar content [Ballarin et al., 1993; Ballarin and Cima, 2005].

Storage cells include pigment cells and nephrocytes. The former have a diameter of 10-20 µm; with large vacuoles containing crystalline granules of pigment (orange, yellow, red, blue) in Brownian motion that contribute to the color of the colonies. Nephrocytes are vacuolated cells with a size comparable to pigment cells with large vacuoles containing birefringent granules of urate crystals in Brownian motion [Sabbadin and Tontodonati, 1967; Milanesi and Burighel, 1978].

1.4 BLASTOGENETIC CYCLE

In this species, budding occurs continuously in an ordered and synchronized way, so that, in a colony, three blastogenetic generations coexist together: adults (filtering zooids), their palleal (primary) buds and budlets (secondary buds) on buds. The development of buds and zooids is highly synchronized, so that cyclic generation changes occur during which adult zooids are absorbed and replaced by primary buds, which opens their siphons within 24-36 h, while secondary buds become primary buds and give rise to a new budlet generation [Sabbadin et al., 1975]. At the generation change or take-over, old zooids contract and their tissues undergo massive, diffuse apoptosis (in parallel with some necrosis in the digestive tube) and cells and corpses are cleared by phagocytes [Burighel and Schiavinato, 1984, Lauzon at al. 1993, Cima at al. 2003]. The colonies, which cannot feed until the new adult generation open their siphons, rely on recycling of components of dying zooids, which assure the growth of the developing buds [Sabbadin, 1956; Lauzon at al., 2002].

The interval of time from one change of generation to the next one is referred to as the colonial blastogenetic cycle. A blastogenetic cycle starts with the opening of the siphons of new adult zooids, and ends with the take-over, when the next blastogenetic generation reach functional maturity. The entire cycle takes one week at 19°C.



The developmental stages can be represented using a formula of three numbers separated by slashes (e.g., 9/8/6), as defined by Sabbadin (1955). Each number refers to the development of the coexisting generation in the colony, the first to adult filtering zooids, the intermediate to primary buds, and the last to secondary buds, respectively. Each colonial developmental stage is, therefore, directly related to the development of zooids, buds and budlets. In optimal condition the cycle starts with 9/6/0 (a brief interval following the take- over during which colonies contain newly formed adults and their buds without budlet generation) and continues with 9/7/1 and 9/8/2-5 until 9-11/8/6, when take-over occurs again.

Secondary bud: stages **1-6**

The early buds appears as a small disc like thickening of the peribranchial wall; the bud arches symmetrically become an hemisphere, and then skews toward the anterior end of the parent (stage 2+). Next, the inner layer of the hemisphere folds into a a sealed vesicle enclosed by an epidermal vesicle (stage 3). The two epithelial layers and the connected tissues between them form the mantle; during bud development, the hearth, gonads, blood sinuses, muscles, neural complex, and nerves form and locate in it. While the inner vesicle loses its connections with the peribranchial wall, the outer one remains associated with the parent epidermis through a hollow peduncle. The bud then elongates along an anteroposterior axis (stage 3+). The bud now is ready to begin organogenesis, which involves the inner vesicle and the mantle; the outer vesicle constitutes the bud epidermis.

Primary Bud: Stages **7**/1-**8**/6

The passage from stages 6 to **7**/1 occurs during take-over, when the bud is ready to produce a new generation of budlets and becomes primary bud. It rotates along its longitudinal axis, becoming oriented like its parent. The primary bud complete its morphogenesis and undergoes cytodifferentiation. At stage **8**/6, the buds are ready to substitutes adult zooids, which are undergoing take-over. Concurrently with take-over, buds arrange themselves into new systems: their atrial siphons elongate until they join (stage 9).

2. AIM OF THE THESIS

B. schlosseri is considered a reliable model organism for the study of a variety of biological phenomena ranging from sexual and asexual reproduction [Sabbadin et al., 1975; Laird et al., 2005; Tiozzo et al., 2005], to regeneration [Zaniolo and Trentin, 1987, Laird et al., 2005], immune responses and allorecognition [Ballarin et al., 1994, 2000, 2001, 2002a, 2006], apoptosis and the clearance of senescent cells [Lauzon et al., 1992, 1993, 2002; Cima et al., 2003]. However, in this species, few studies have been conducted at genomic and transcriptomic level. The lack of a reference genome limits a molecular biology approach to Botryllus studies.

General aim:

Produce reliable transcriptomes from colonies at different blastogenetic phases, useful to identify differentially transcribed genes.

Specific aims:

(1) Genes are expressed at different levels and often transcribed into many isoforms. Independently by used sequencing technology the assembly of transcripts is still difficult. While several published methods allow to address ILLUMINA RNA-seq assembly, the absence of methods for SOLiD data represents a strong limitation for the scientific community. The first objective of this thesis is finalized to set a method that allows to produce a *de novo* transcriptome assembly using SOLiD RNA-seqs.

(2) The blastogenetic cycle of *B. schlosseri* is one target of the RNA-seq experiments. Three development phases were considered: a phase far away the

cyclic generation change (mid-cycle) indicated as 9/8/2, a phase immediately before the take-over which is indicated as 9/8/6 and take-over phase indicated as 11/8/6. Other RNA-seq experiments regard the vascular system regeneration. The second objective of this thesis is focused on gene expression study of RNA-seq experiments which have been considered.

(3) The high number of RNA-seq experiments and biological replicas require an efficient method to search differentially expressed genes basing on statistical significance and logical criteria. Furthermore, biological function of each transcript can be inferred through bioinformatics. The third objective of this thesis is oriented to develop a Web-based interface finalized to make possible this important task.

3. MATERIALS

3.1 Animals sampling and breeding

B. schlosseri forms flat, variously pigmented colonies, extending on submerged vegetal and rocky substrates. Colonies, collected in the lagoon of Venice in September-November 2011, were left to adhere to and grow on glass slides. They were initially kept at the Marine Station of the Department of Biology in Chioggia into a 500-l tank with continuous seawater flow. Each colony was fragmented in three to five sub-clones. Colonies were then transferred to the aquaria of the Department of Biology in Padova in January 2012 and reared under controlled conditions, according to Sabbadin's technique, in thermostated rooms, at the temperature of 15°C.

3.2 RNA-seq experiments

Once total mRNA was extracted and purified from cells, it was sent to a high-throughput sequencing facility, than fragmented and selected by length before the construction of cDNA library. The paired-end libraries were sequenced at the CRIBI center of Padua using SOLiD 4 and 5500 sequencers. The experiments that regards the vascular system regeneration are referred to the first run, while the experiments inherent the blastogenetic cycle are referred to the second run. Some experiments of SOLiD run 1 are technical replicas used to test the reproducibility of the sequencing output. All experiments were designed considering biological replicas.

<u>SOLiD RUN1</u>

<u>Experiment 1</u>: mRNA sample extracted from colony A at the achievement of the 9/8/2 phase.

<u>Experiment 2</u>: technical replica of the experiment n. 1.

Experiment 3: mRNA sample extracted from colony A in which a portion of the tunic was removed 2 days before reaching mid-cycle (9/8/2) phase, when the mRNA was extracted.

Experiment 4: technical replica of the experiment n. 3.

Experiment 5: mRNA sample extracted from colony A when take-over phase (11/8/6) was reached.

Experiment 6: technical replica of the experiment n. 5.

Experiment 7: biological replica of the experiment n. 1.

Experiment 8: biological replica of the experiment n. 5.

SOLiD RUN2

Experiment 1: mRNA sample extracted from a colony at the achievement of the 9/8/2 phase.

Experiment 2: biological replica of the experiment n. 1.

Experiment 3: biological replica of the experiment n. 1.

Experiment 4: biological replica of the experiment n. 1.

Experiment 5: biological replica of the experiment n. 1.

Experiment 6: mRNA sample extracted from a colony at the achievement of the 9/8/5 phase.

Experiment 7: biological replica of the experiment n. 5.

Experiment 8: biological replica of the experiment n. 5.

Experiment 9: biological replica of the experiment n. 5.

Experiment 10: biological replica of the experiment n. 5.

Experiment 11: mRNA sample extracted from a colony at the achievement of the 11/8/6 phase.

Experiment 12: biological replica of the experiment n. 11.

Experiment 13: biological replica of the experiment n. 11.

Experiment 14: biological replica of the experiment n. 11.

Experiment 15: biological replica of the experiment n. 11.

3.3 SOLiDTM 4 and SOLiDTM 5500 sequencers

Two different version of sequencers were used for sequencing: the SOLiD 4 and its most advanced version SOLiD 5500. The manufacturer reports many advantages arising from the use of its sequencer which may be listed in 5 points:

- high quality of the data
- high throughput for a single run (300 Gb)
- 80% reduction of the lab work (automated work flow)
- directional paired-end sequencing
- barcoding system which allows to reduce costs and time to prepare libraries

SOLiD^{TM} 4 specifications

The SOLiD^{TM} 4 has a throughput greater than 100 GB and can produce 1.4 billion of reads per run. The size of the reads depending on the type of library to be sequenced: 50 bases for fragment libraries, 50 + 35 bases for paired-end libraries and 50 + 50 bases for mate-paired libraries. The sequencer has a base accuracy of 99.94% while the accuracy of the consensus of each base at 15x coverage is 99.999%. The multiplexing system can handle 96 and 48 barcode respectively for DNA and RNA libraries. In addition, more than 80% of the bases have a quality ≥ QV30 (as reported by the manufacturer).

SOLiD^{TM} 5500 specification

The sequencer SOLiDTM 5500 has a throughput of sequences greater than 300 GB and can produces about 5 billion of reads per run. The size of the reads depends on the type of the library to be sequenced: 75 bases for fragment libraries, 75 + 35 bases for paired-end libraries and 60 + 60 bases for mate-paired libraries. The sequencer has a base accuracy of 99.94% while the consensus accuracy at 15x coverage is 99.999%. The multiplexing system can handle 96 barcodes for both RNA and DNA libraries. In addition, more than 80% of the bases have a quality ≥ QV30 (as reported by the manufacturer).

3.4 Computing cluster

The computing cluster system available at the CRIBI center is a cluster of 32 servers grouped as follow: 4 servers with 12 CPU and 96Gb of RAM; 24 servers with 12 CPU and 48Gb of RAM; 4 servers with 12 CPU and 24 Gb of RAM; 1 server with 64 CPU and 2Tb of RAM; 2 servers mounting Tesla GPUs. Two servers called “master node” manage the main needs of the file system “Lustre”, the local network and the submitted jobs. The file system Lustre [http://www.lustre.org/] is a parallel distributed file system, mostly used for large scale computing cluster accessible under the GNU GPL (v2 only), that supply a high performance for computer clusters. Because Lustre has high performance and open licensing, it is frequently used in supercomputers. Furthermore, Lustre is scalable and can support tens of thousands of client systems, tens of petabytes (PB) of storage, and hundreds of gigabytes per second (GB/s) of aggregate I/O throughput.

3.5 BLAST PROGRAM

In bioinformatics, Basic Local Alignment Search Tool, or BLAST [Altschul at al., 1990], is an algorithm for comparing primary biological sequence content, such as the amino-acid sequences of different proteins or nucleotides of DNA sequences. This program enables researchers to compare a query sequence with a database of sequences. BLAST identifies the sequences belonging the database that resemble the query above certain thresholds. Different BLAST programs are available according to the type of the query or database to be analysed (nucleotides or proteins).

3.6 CD-HIT-EST PROGRAM

CD-HIT-EST [Weizhong Li at al.] is part of a suite of programs planned to quickly group sequences. CD-HIT-EST groups nucleotide sequences (without introns) into clusters that meet a user-defined similarity threshold. Input is a fasta file of nucleotide sequences. Output is a file of (non-redundant) representative sequences and a file containing a list of sequence names of each cluster.

3.7 PASS PROGRAM

PASS [Campagna at al., 2009] is a program to align short sequences coming from new sequencing technologies. It has been developed with an innovative strategy to perform fast gapped and ungapped alignment onto a reference sequence. It supports several data formats and allows the user to modulate very finely the sensitivity of the mapping. The program is designed to handle huge amounts of short reads generated by ILLUMINA, SOLiD and Roche-454 technology. The optimization of the internal data structure and a filter based on precomputed short-word alignments allow the program to skip false positives in the extension phase, thus reducing the execution time without loss of sensitivity. The final alignment is performed by dynamic programming.

3.8 BLAST2GO TOOL

Blast2GO (B2G) [Conesa et al., 2005] is a extensive bioinformatics tool for the functional annotation and analysis of gene or protein sequences. The tool was originally improved to provide a user-friendly interface for Gene Ontology (The gene ontology consortium annotation – 2008). Recent change of state have considerably increased the annotation practicality of the tool and presently, Enzyme code (EC), KEGG Maps and InterPro motifs are also supported [Hunter et al. 2011]. In addition the application offers a panoramic array of graphical and analytical tools for annotation handling and data mining. Blast2GO utilizes local or remote BLAST searches to discovery similar sequences to one or various input sequences. The program extracts the GO terms connected to each of the obtained hits and returns the GO annotation for the query sequence(s). Enzyme code number are acquired by mapping from equivalent Gos terms, while InterPro motifs are immediately queried at the InterProScan web service. GO annotation can be envisioned thought the structure of the Gene Ontology relationships and Ecs are high spot on KEGG maps. A exemplary use of Blast2GO essentially consists of 5 steps: blasting, mapping, annotation, statistics analysis and visualization.

3.9 C++ LANGUAGE

The C++ is a statically typed, free-form, compiled, general-purpose programming language. It is regarded as an intermediate-level language, and it consists of both high-level and low-level language features. Developed by Bjarne Stroustrup starting in 1979 at Bell Labs, it supply object oriented features, such as classes, and other enhancements to the C programming language. Originally called C with Classes, the language was renamed C++ in 1983, as a wordplay involving the increment operator. C++ is one of the most popular programming languages and it is implemented on a wide variety of hardware and operating system platforms. As an effective compiler to native code, its application domains include systems software, device drivers, application software, embedded software, client applications and, high-performance server. Various groups give both free and proprietary C++ compiler software, including the GNU Project, Microsoft, Intel and Embarcadero Technologies.

3.10 BIOLOGICAL DATA

1. CIPRO 2.5 protein database [Endo T. at al.] : The *Ciona intestinalis* protein database (CIPRO) is an integrated protein database for the tunicate C. intestinalis. The current database is based on a recently developed KH model containing 36,034 unique sequences, but for higher usability it covers 89,683 all known and predicted proteins from all gene models for this species. Of these sequences, more than 10,000 proteins have been manually annotated. Furthermore, to establish a community-supported protein database, these annotations are open to evaluation by users through the CIPRO website. CIPRO 2.5 is freely accessible at http://cipro.ibio.jp/2.5.

2. FM1-filtered models protein database of *Ciona intestinalis* produced by the JGI - U.S. Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA and available at the Web path: http://genome.jgi-psf.org/Cioin2/Cioin2.download.ftp.html. "FilteredModels" is the filtered set of models representing the best gene model for each locus.

3. *Ciona intestinalis* Unigene dataset. These are the *C. intestinalis* transcript sequences derived both known genes and ESTs that have been partitioned into clusters and could be downloaded at: ftp://ftp.ncbi.nih.gov/repository/-UniGene/Ciona_intestinalis/.

4. nr database. The nr database is a 'non-redundant' database (i.e. with duplicated sequences removed). nr contains non-redundant sequences from GenBank translations (i.e. GenPept) together with sequences from other databanks (Refseq, PDB, SwissProt, PIR and PRF).

4. ASSEMBLY METHOD VALIDATION

4.1 SIMULATED DATA

The multiple-kmer method adapted to color space sequences could be validated and optimized using simulated data. For further details about the method please see chapter 5. The *C. intestinalis* genome annotation database and its associated reference genome (v2.0) are required to extract the coding sequences and the UTR regions from *C. intestinalis* mRNA species. These data were downloaded at the UCSC ftp site. The genome annotation includes 8639 transcripts; 5082 are primary mRNA and 3557 represent minor isoforms that share 1 or more exons. This simulation plans to cover a range of 1 to 100X sequence coverage in order to simulate a large range of expression levels linearly distributed to all mRNA species. The expression levels of the minor isoforms were set to 1/5 of the associated primary mRNA in order to understand their impact in the assembling of chimeric contigs. According to these conditions, the color space simulated reads were generated using the dwgsim-0.1.8 program from the samtool package (http://sourceforge. net/apps/mediawiki/dnaa/index.php? title=Whole_Genome_Simulation).

The per base/color/flow error rate and rate of mutation was set to the default values (respectively: 0.02 and 0.001).

Parameter used to generate reads:

```
dwgsim  -y 0 -z 0 -d 100 -S 2 -c 1 -1 76 -2 35 -C coverage
```

where “coverage” is the sequence coverage according to a specified expression level.

The simulated reads were selected to simulate directed cDNA library and used to produce a *de novo* transcriptome assembly. The “Velvet” assembly program [Zerbino at al. 2008] supports the assembly of the double encoded reads;

actually this program represents the only way to assemble color space sequences. First, the color space reads were encoded, and then they were used to generate assemblies for each of the following Kmer size: 21, 23, 25, 27, 29, 31 and 33 bases. In order to understand the ability of the assembly program to assemble color space reads, the same analysis was repeated using also simulated base space reads (a typical ILLUMINA strategy). The comparison of the results excluded post process analysis interferences to unspecific results and allowed to understand the weight of each step on assembly errors. Base space and color space assembly statistics is reported as well as other analysis.

4.2 Redundancy removing: two evaluated strategies

The removing of sequence redundancy is necessary when the same transcript portions is assembled many times as resulting by multiple-kmer assembly. Two possible strategies are analysed: i) in the first one, each assembly must be color space converted, then added to form a pool of base space contigs and finally, redundancy must be removed. ii) in the second strategy, the color space assemblies should be pooled, the contigs redundancy removed and then color space conversion applied to not redundant sequences.

The “cd-hit-est” program [Weizhong Li at al.] is designed to process base space sequences. Consequently, this program is affected by the presence of polymorphisms only in the case of color space sequences. Each polymorphism changes at least two contiguous colors while in the case of base space sequences, only one. Polymorphisms are not the only problem.

Despite the assembled contigs from “directed RNA-seq libraries” should be produced in the same orientation, theoretically forward and reverse strands of color space sequences should be treated in different way than the base space ones.

Considering these problems, the second strategy is not interesting for color space managing however it is interesting for its small time-demanding. The two strategies are compared in terms of mapped contigs and fraction of contigs not mapped on reference mRNAs. The similarity threshold for clustering analysis was set to 0.9 according to the proposed method [Beide at all. 2012]. The minimal percentage of identity was set to 90% and the mapping was performed using a global alignment. The results indicates that 90.37% of total and non-redundant contigs produced in the first strategy were mapped on the reference mRNAs. In the second strategy the 90.19% of total contigs were mapped on the reference mRNAs using the same conditions used in the first strategy. The two strategies seem to be qualitatively equivalent.

4.3 Base space and color space assembly statistics

On the other way, the strategy applied to base space assembly is similar to the second strategy for color space data with the only exception of color space conversion. The assembly of base space simulated data indicates that 90.83% of total and non-redundant contigs were mapped on the reference mRNAs. Base space and color space assembly did not show relevant differences on results (90.37% vs 90.83%) so both sequencing platforms indicate that 10% of non-redundant contigs seems to be incorrectly assembled. The analysis of non-mapped contigs of each assembly revealed a maximum of 15% in length of badly assembled sequences. The 90% of non-mapped sequences were successful aligned using a local alignment. This result indicates the presence of chimeric contigs formed by parts of different transcripts. Furthermore, no significant differences was found between color space and base space data analysis. A consistent number of redundant contigs (2,5 to 6%) seems to be badly assembled and do not map on the reference mRNAs. The statistics about color space and base space assemblies are reported in the following figures.

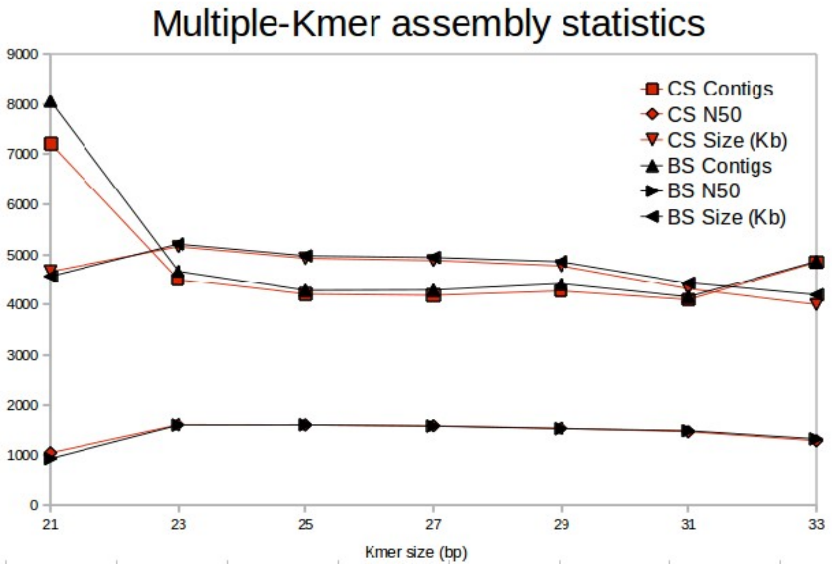


Figure 4.3.1: The figure shows the assembly statistics for both color space (CS) and base space data (BS). X axis: Kmer length; Y axis: number of contigs , N50 value and size of each assembly.

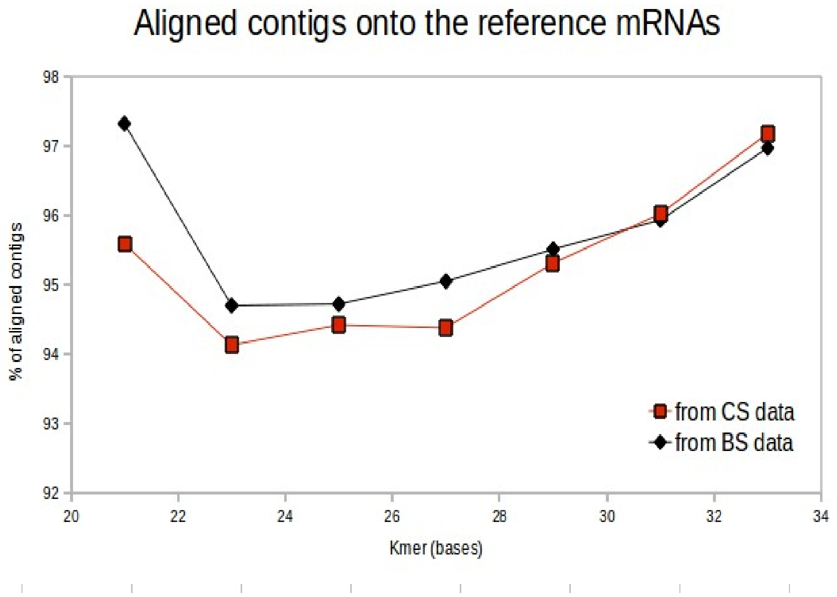


Figure 4.3.2: Percentage of aligned contig for each considered Kmer assembly for both color space (CS) and base space data (BS). The color space contigs were converted to base space through MCSC (see chapter 5). The mapping was performed using the PASS program with global alignment. The minimal percentage of identity was set to 90%.

4.4 Assembly errors and sequence coverage

Two assembly errors are greatly represented in RNA-seq assembly. The first one is represented by chimeric contigs formed by different assembled transcripts; this is the consequence of the assembly interferences of sequence variants as isoforms or conserved domain shared in more genes. In this case, the mapping analysis of paired-ends has revealed a number of mapped single-ends closest to the mis-assembling junctions greater than the same ones closest to other regions. This situation is caused by the impossibility to map the reads of the pair in the wrong sequence context which is located the error. In contrast, paired-end alignments are expected to be uniformly distributed in both assembled regions outside mis-assembly junction. In a second type of assembly error, a small portion or the entire contig, is completely incoherent. This situation gives no chance to map paired-ends onto the sequence context, but it is expected an increased number of mapped single-ends. Generally, in a more practical way, assembly errors produce an incoherent distribution of mapped reads that theoretically, could be recognized on the basis of sequence coverage. The mean contig coverage and the percentage of covered contigs should be calculated to measure the correlation coefficient. It is reasonable to think that, the correlation coefficient relative to these two variables should be different for both good and wrong assembled contigs.

As shown in fig. 4.4.1 the correlations are always high in all cases, although there is a small difference between well and badly assembled contigs. As resulting from the analysis, a great fraction of badly assembled contig are good sequences that have small and localized misassemblies. This type of errors generate a small mapping incoherence that is not sufficient for an error discrimination. Concluding, a strategy based only on mapped reads seems to be not adequate to recognize and minimize assembly errors.

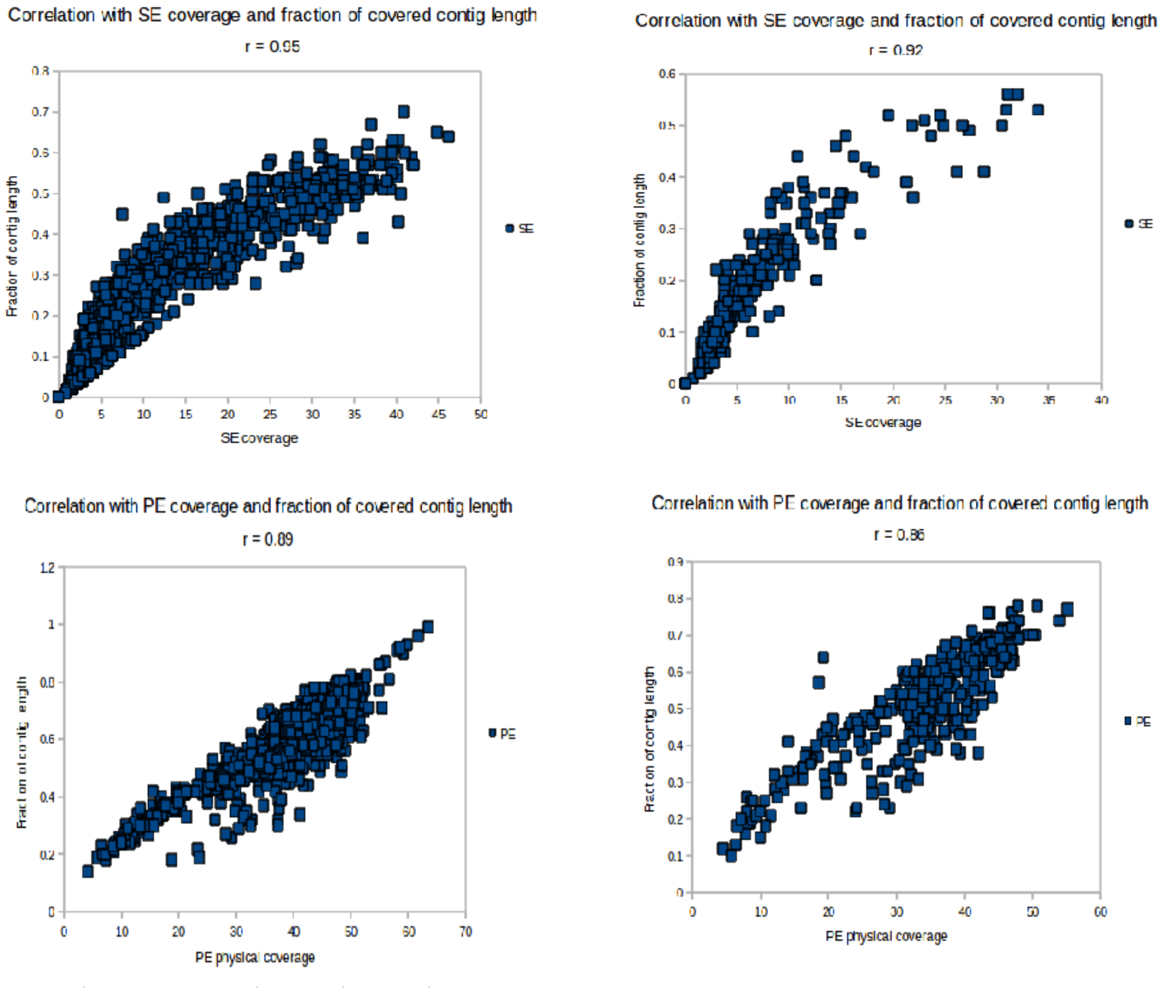


Figure 4.4.1: Correlation of mean coverage and percentage of contig positions that have a coverage > mean coverage. The 2 plots placed on the left are referred to the correct contigs, while the other two on the right are referred to the badly assembled contigs. In both cases single-end (SE) and paired-ends (PE) alignments were considered separately to emphasize this kind of information.

4.5 MATRA - *Minimizing Assembly errors Through Redundancy Analysis*

The multiple-Kmer method makes possible the assembly of low and high covered transcripts resulting from the heterogeneous gene expression level of tissues or experimental conditions. Different Kmers allow to assemble different portion of transcripts and give us a specific sensitivity for different sequence coverage. It is likely to think that these differences are not given also for correctly assembled contigs but they also include specific assembly errors. In the majority of the cases, the multiple-Kmer assembly produces the same assembled transcript portions which will be redundant. In other words, the suggested method [Surget-Groba at al. 2010] considers both wrong and correct consensus generated using different kmer size. Consequently, if there is a specificity of assembly errors produced by a specific Kmer, the simple selection of redundant regions should be the key to recognize assembly errors as less redundant sequences.

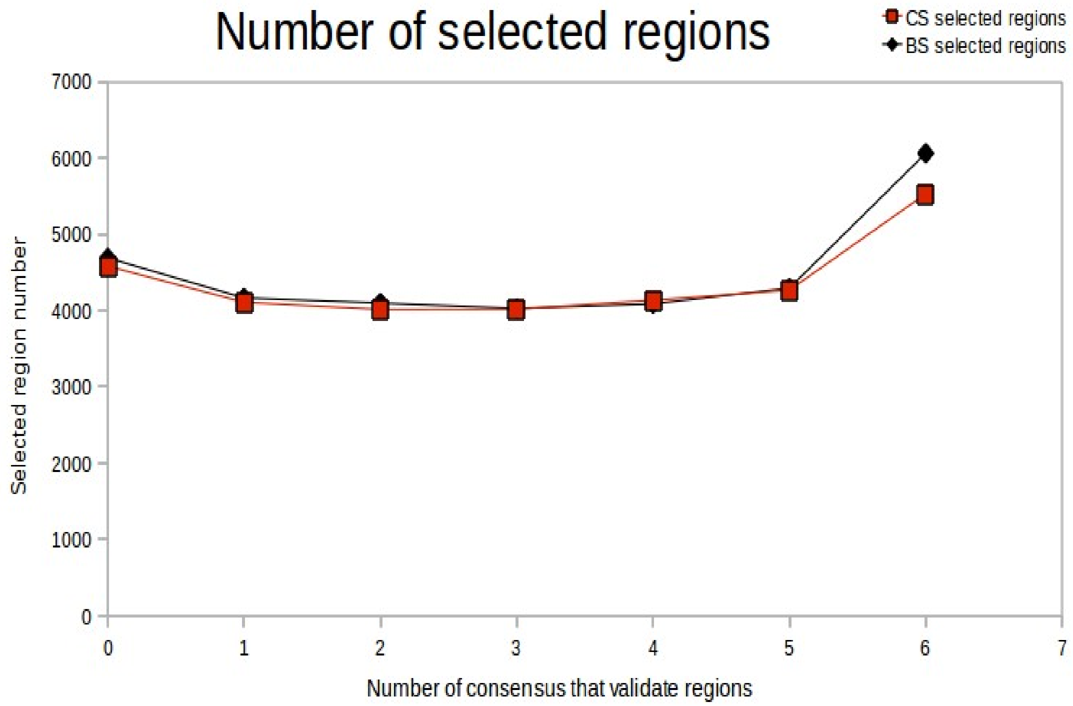


Figure 4.5.1: Number of selected regions obtained using different setting. The red line indicates the analysis of color space data while the black line is referred to base space one. The X axis represents the minimal number of consensus that validate the same selected region while the Y axis the number of selected regions obtained for each set condition.

In order to evaluate this possibility, a specific program was developed to analyze the assembly of simulated reads. The reduction of assembly errors was quantified through a mapping of redundant consensus on the reference mRNAs (global alignments). The mapped consensus are selected considering a threshold number of minimal redundant consensus coming from different assemblies that confirmed the same region. The composition of unmapped contigs were studied to understand the properties of assembly error as well as the mapping statistics.

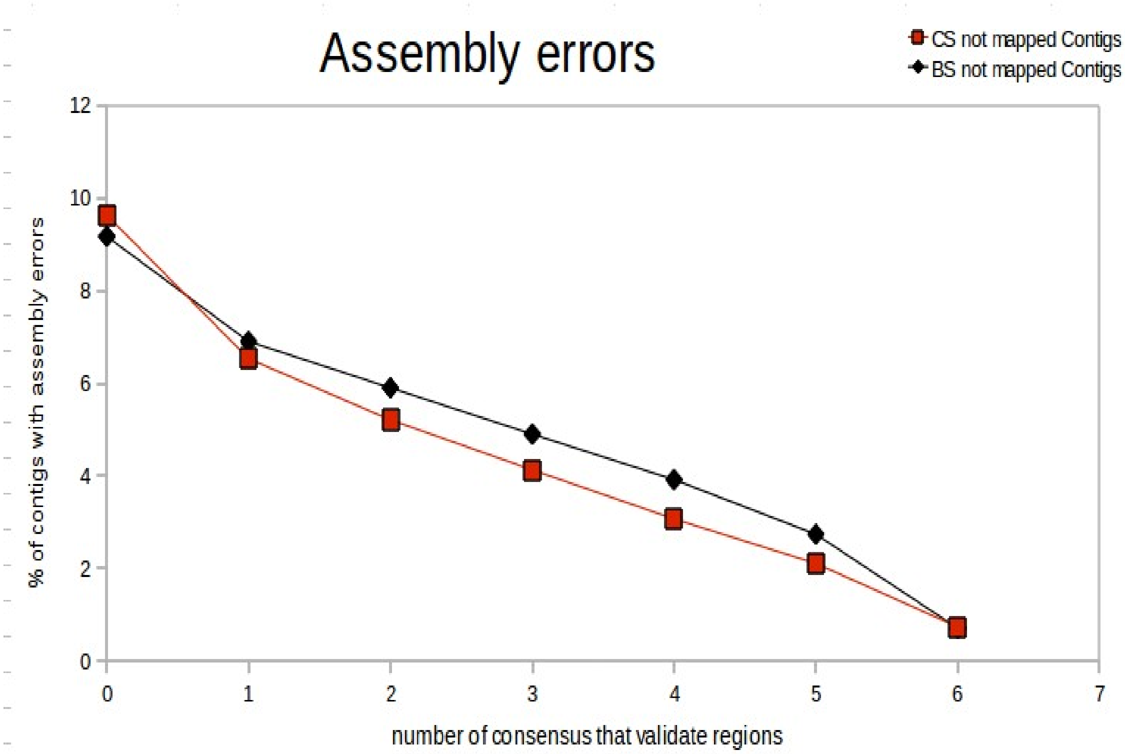


Figure 4.5.2: Percent of assembly errors obtained using different setting. Consensus were globally mapped using PASS program. Red line: color space data; black line: base space data. X axis: minimal number of consensus that validate the same region; Y axis: percentage of assembly errors obtained by each set condition.

As shown in figure 4.5.2 assembly errors could be reduced from 10% to < 1%.

If sequence quality is low, MATRA can strongly reduce the assembly size, so caution must be taken to set the redundancy threshold. Under this condition, the representativeness of the assembled transcripts will strongly depends by setting.

No significant differences could be appreciated in the simulation of the two evaluated sequencing platforms that seem to be equivalent for number and quality of results.

4.6 Contigs elongation

After redundancy removing, many contigs are orphan contigs and many other belonging to the same transcript could be assembled because they have overlapped ends. The contigs elongation requires 2 operations described below: i) contigs were mapped one against to each other with the aim to cluster overlapped ends [PASS at al. 2009]; ii) each cluster was analysed to assemble contigs with compatible ends using the CAP3 program [Huang at al. 1999]. In order to have high specificity, the program checks the following conditions: i) the considered overlapped ends must have the same orientation because they are assembled using directed cDNA library, ii) the cluster information are saved and used to avoid chimeric associations in the following steps; iii) in order to minimize the number of false positives alignments, the minimal percentage of identity of the overlapped ends was set to 95% and the minimum length of the same alignments was set to 70 bases in size.

The total contigs generated by color space and base space assemblies were mapped on the reference mRNAs after contigs elongation. Applying MATRA method before contig elongation, 99.29% of the total contigs from base space assembly were successfully global aligned on the reference mRNAs as those referred to color space data (99.26%). The quality of the assembly has revealed a minimal difference with the percent of mapped contigs in the previous step (<0,2%). Basing on this set, the contigs elongation seems to have high specificity to improve the assembly and does not represents a critical point of the method.

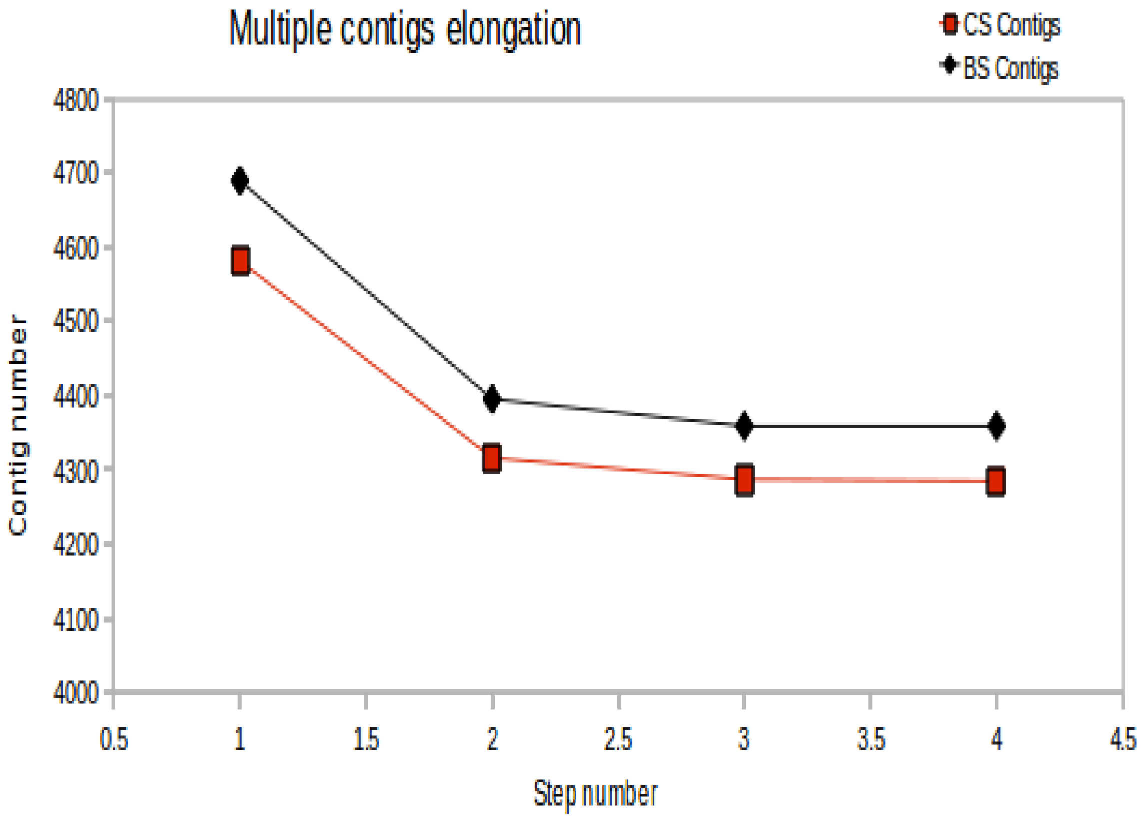


Figure 4.6.1: Multiple contig elongation statistics. During assembly steps the contig number reaches the saturation indicating that no more supercontig could be assembled. In this analysis MATRA processing is not considered.

4.7 PE scaffolding

The assembly subjected to contig elongation was scaffolded using paired-end scaffolding. The developed program is similar to SSPAPE program [Boetzer at al. 2010] but it is designed to manage also color space assembly. The threshold ratio *R* of the two best associations (number of paired-ends that link a contig to another one) is the main parameter to modulate sensitivity and specificity of both programs. In this analysis *R* was set to 0.3 and the results indicate that 99.31% of total contigs obtained from base space simulation and 99.27% of total contigs referred to color space simulation, were mapped using global alignment on the reference mRNAs. The difference with the previous step is less than 0.1%. On the basis of this set, the PE scaffolding seems to improve assembly with high specificity and does not represents a critical point of the method.

4.8 Conclusion

All steps of the method, except for Velvet assembly, seem to have a great efficiency to assemble correct supercontigs. As resulting from the analyses, using the setting of the designed simulation, there is no practical difference for the compared sequencing technologies. The quality of the assembly seems to be greatly influenced by the Velvet assembly that produced about 10% of non-redundant badly assembled contigs.

On the basis of our results, it is possible to remove about 93% of the assembly errors using a novel method called MATRA. MATRA allows to split chimeric contigs to selected regions confirmed by a minimal number of consensus sequences. The contigs produced using base space data represent 74% and 17% of primary and secondary isoforms belonging to the considered data set, respectively, while the contigs coming from color space data represent 72% and 17%. At the end of the process the resulted chimeric contigs were less than 1% for both the considered sequencing platforms.

5. METHODS AND RESULTS

5.1 RNA extraction

Total RNA was extracted from five different colonies (corresponding to five different genotypes); from each colony three sub-clones at various colony developmental phases were utilized. Developmental phases were carefully resolved *in vivo* following the recently revised Sabbadin stadiation and chosen in order to study differentially gene expression during the colonial blastogenetic cycle: I) the colony phase 9/8/5 identifies the phase immediately preceding the take-over, when the colony is preparing to the generation change; ii) the colony phase 11/8/6 identifies the take-over, when adult zooids are absorbed and replaced by new ones, and iii), the colony phase 9/8/2 identifies the mid-cycle phase, when buds and zooids coexist together and no generation change occurs. It is to note that during the take-over several morphological changes occurs in the colony, and a recently dedicated sub-stadiation (11^1 to 11^4/8/6) was followed, permitting to better recognize the progress of the generation change; the selected colony phase was 11^2/8/6. Colonies at the appropriate developmental phase were immersed in liquid nitrogen and kept frozen at -80°C until use.

RNA was obtained from a *B. schlosseri* colony sub-clones, averaging 230 mg (110 to 350), minced for 2 min with a frosted glass pestle in 15 ml tubes filled with approximately 2,5 ml of an heated (65°C) extraction buffer composed of CTAB Lysis buffer (Applichem, cat. n. A4150) and 2% β-mercaptoethanol. Samples were then maintained for 1,5 h at 65°C in a waterbath shaking strongly for few seconds every 20 min, and cooled for 2 min in ice. Then, 3 volumes of a solution of chloroform-isoamyl Alcohol (24:1) was added and the sample was thoroughly mixed shaking the tube with hands to get an emulsion. A centrifugation step of 15 min at 3500 rpm, at 4°C, was performed so that the three layers (top: aqueous phase, middle: debris and proteins, bottom: chloroform) formed. The aqueous phase was quickly collected into a new 1.5 ml tube and total RNA was isolated and concentrated from it. An equal volume of 96% ethanol was then added and the total RNA was purified using the columns of the SV Total RNA Isolation Kit (Promega, cat. n. Z3100). The Invitrogen Qubit Fluorometer and the Thermo Scientific Nanodrop ND-1000 Spectrophotometer were used to check RNA purity and concentration. Ethidium bromide-stained agarose gel (1%) and the Agilent 2100 Bioanalyzer were used to determine RNA sizing and integrity.

5.2 SOLiD RNA-seq library preparation

After total RNA quality checking, beads linked to oligo(dT) are used to enrich poly(A) mRNA from total RNA. Than the RNA was fragmented and used to prepare paired-end cDNA libraries in accordance with SOLiD protocol. Two adapters (called by Life Technologies P1 and P2) which contain the “sequencing primer sites” were linked to each fragment to form a cDNA construct to be sequenced (figure 5.2.1).

The SOLiD sequencing of a “fragment library” allows to produce sequences of 75 bp in length starting from the P1 adapter (50bp with SOLiD 4 sequencer). These sequences are labelled with the abbreviations “F3” or “FWD1”.

The reverse sequences are only 35 bp in length, due to the loss of the ligation efficiency after many ligation cycles. The reverse reads are sequenced starting from the opposite end of the cDNA construct closest to the P2 adapter (figure 5.2.1) and they are labeled using “F5” or “REV1” tags.

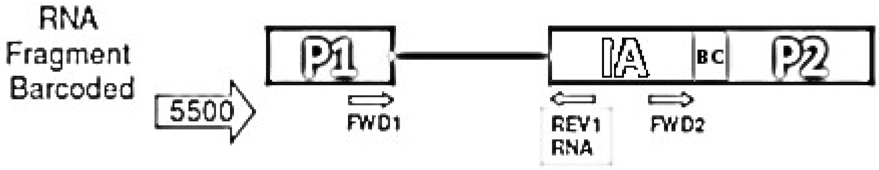


Figure 5.2.1: Simplified diagram of the DNA construct designed for paired-end sequencing. The SOLiD libraries have an internal adapter of 20 bp and a variable sequence of 5-10 bp in length called “barcode adapter” (BC). The figure shows the internal adapter IA, the adapter P2 and the “barcode sequence” highlighted with the label “BC”. FWD1, REV1 and FWD2 represent the sequencing primers.

During DNA amplification, each fragment that contains both “priming sites” will be amplified. This not happens for fragments who have only one of the 2 incorporated adapters. The fragment libraries may be “multiplexed” if barcode are used, however a maximum of 96 barcode are allowed. The Applied Biosystem suggests to use only the indicated barcode because they ensure the color balance during the sequencing. The barcode sequences are numerated and grouped into several categories. Groups 1-4, 5-8, 9-12 and 13-16 are preferable because they require 5 additional bases to be sequenced, while the barcode 17-96 require the sequencing of 10 additional bases. The barcode size does not reduce the number of produced reads because they are treated as a separate sequencing run, however more time is demanding. The software of the sequencer will distinguish the reads coming from each sequenced sample, through the barcode recognition. The SOLiD cDNA library requires the production of a population of “cloned beads” (McKernan et al. 2009) and the procedure could be summarized on 4 points: i) the emulsion PCR allows to amplify the DNA inside the microreactors that contain all necessary reagents; ii) the amplified DNA is denatured and the beads saturate the template sequences of each species; iii) the templates are bound to the universal adapter sequence P1 and P1 will bind the surface of the bead; iv) each DNA construct will be modified at the 3' end in order to have a covalent bind to a slide. Finally the magnetic beads will be deposited on the slide, ready to be sequenced. The slide could be split into 1, 4 or 8 sections. The capability of the slide to bind the beads at high density strongly affects the number of sequenced reads.

5.3 Sequencing by ligation

At the beginning of each ligation cycle, sequencing primers hybridize with the adapter P1 that is linked to the template molecule. A set of 4 fluorescent probes compete for the same hybridization at the same position of the DNA construct.

The specificity of the reaction is guaranteed by the hybridization of the bases at position 4 and 5 of the probe.

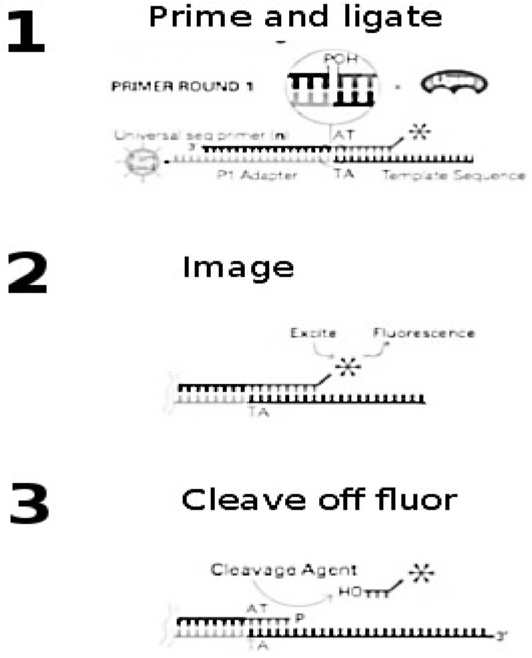


Figure 5.3.1: For each ligation step, a base is sequenced 5 positions far away to the previous one (1). A fluorescent signal is detected (2) and fluorophore is cleaved off after light emission (3).

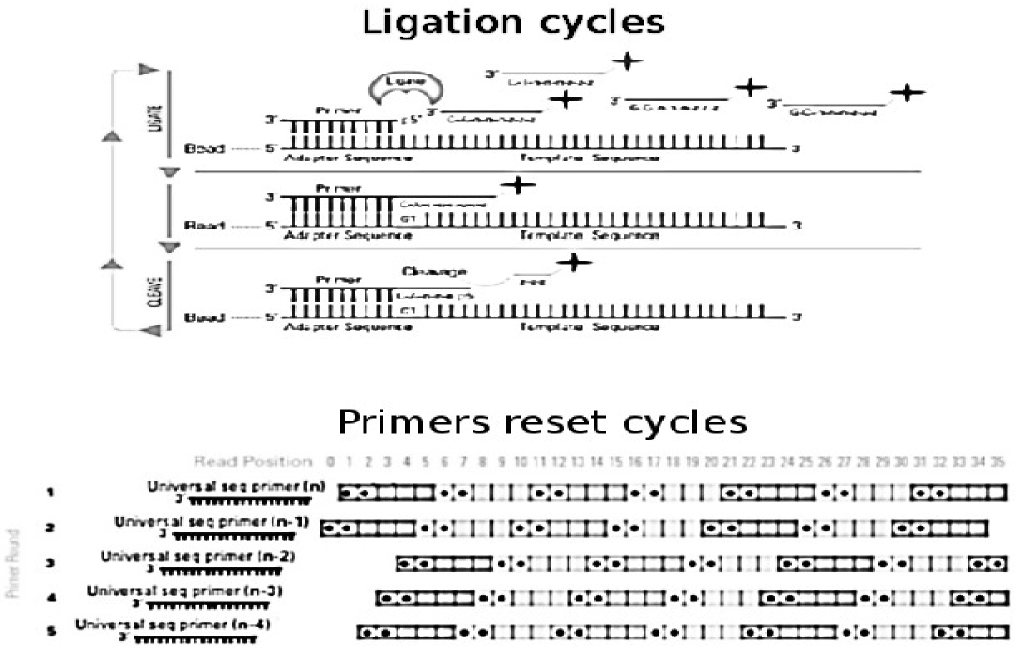

Ligation cycles
LIGATE
CLEAVE
Ligase
Primer
Bead
5'
3'
Adapter Sequence
Template Sequence
Cleavage
Primers reset cycles
Read Position 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
Primer Round
1 Universal seq primer (n)
2 Universal seq primer (n-1)
3 Universal seq primer (n-2)
4 Universal seq primer (n-3)
5 Universal seq primer (n-4)

Figure 5.3.2: SOLiD sequencing ligation cycle. As consequence of the primer reset each base will be sequenced in two different ligation cycles. For instance, the base at position 5 will be sequenced in the second ligation cycle using the primer number 2 and, in the third ligation cycle using the primer number 1.

Multiple cycles of ligations, detection and cleavage provide and determine the length of the sequenced reads. During each cycle, the extension products will be removed and the templates will be initialized for the next binding at position n-1: a subsequent ligation cycle starts.

The entire sequencing process requires five rounds of primer initialization (primer reset). As consequence each base will be sequenced in two independent reactions from 2 different primers. The chemistry of the SOLiD sequencer is not affected by the presence of homopolymers.

5.4 2 BASE ENCODING

The 2 base encoding is a technique used to map color space reads produced by SOLiD sequencers [McKernan et al. 2009]. The SOLiD sequencer exploits a technology based on 16 probes associated to 4 colors emitted by corresponded fluorophores. Each pool that shares the same fluorophore, is formed by 4 probes associated to the same emitted color that will recognize 4 specific dinucleotides (figure 5.4.1).

The 2-base encoding can be described using a 4-color schema (figure 5.4.3) because each base is effectively probed in two different ligations. As consequence of the double interrogation, each sequenced base has a different measurement of the sequencing error that can be reflected in the overall quality of the sequenced reads. Furthermore the 2 base encoding allows to distinguish the sequencing errors (single color difference) by real mutations (at least 2 variants of consecutive colors).

The color space introduced with the SOLiD technologies and the traditional base space consist of 4 elements: 4 colors indicated as 0, 1, 2, 3 and 4 bases indicated as A, C, G, T. Color-space reads can be efficiently mapped using mappers specifically designed to manage color space rules.

Generally, mappers decode the reference sequence into color-space, using the 16 colors schema of 2 base encoding (figure 5.4.3) and then try to align directly the color space reads. Finally, the produced alignments are converted into base space. The color-space mapping is absolutely preferable to the “double encoded” technique which is used by other applications that manage SOLiD reads not considering the color space rules. Another method take in consideration, the direct decoding of a color-space sequence to base-space. Also this technique is not advisable because each sequencing error would cause a frame-shift error in the color conversion. In fact using 2 base encoding, 4 possible base space variants could be possible: only one is right.

The 2 base encoding does not represents a system of error correction but a way to process the error. Considering a read size of 50 bp in length the likelihood of two adjacent errors can be estimated. There are 49 ways to make adjacent changes of a sequence 50 bases long in size and 1225 ways to make changes that are not consecutive in the same string: i.e. 1 against 25. Simplistically, if we assume completely random errors, only 49 can be candidates mutations of 1225 variants. Furthermore, only 1 of 3 possible variants can be a candidate mutation, so the ratio is lowered down to (1 / 75). The statistical effect is clear, especially for single base polymorphisms subjected to low sequence coverage (Smith et al. 2008).

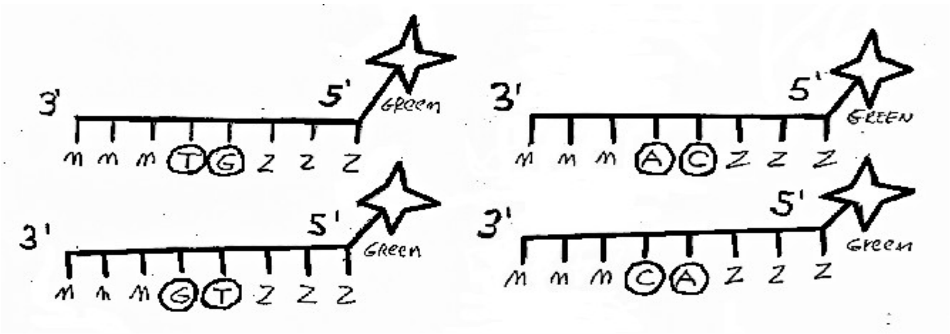


Figure 5.4.1: Each probe consists of 8 bases. The first 3 bases are degenerated (n) and the last 3 are universal (z). The base number 4 and 5 need to be queried. Therefore the recognition of a single color is limited to 4 of 16 possible dinucleotides. In this case the green signal represents the AC, CA, TG, GT dinucleotides.

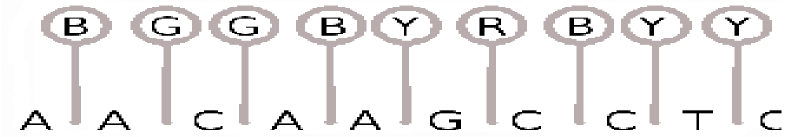
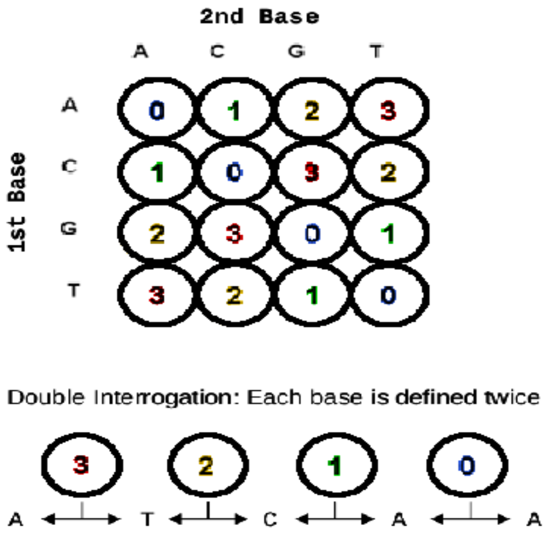


Figure 5.4.2: The principle of the double interrogation of each base. Each color may encode 4 possible dinucleotides. For instance, the first blue is 'AA' while the third one is 'CC'. Practically, there are 4 different ways to convert a color space sequence: only one is right.



1st Base \ 2nd Base	A	C	G	T
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0

Figure 5.4.3: 2 base encoding matrix. If 1 base is known, a color space sequence can be converted to base space. This figure shows how each base is defined by two adjacent colors. The example reports the logic of the color space conversion as the result of a double interrogation.

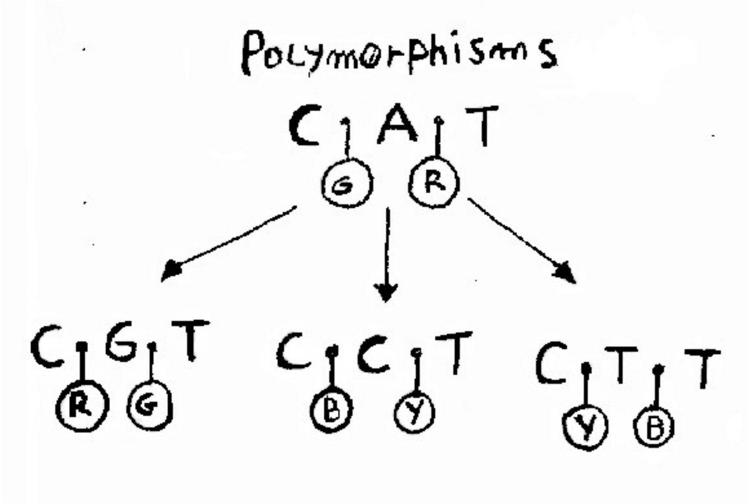


Figure 5.4.4: If a SNP is present there are only three possible outcomes: CGT, CCT, CTT. Only 3 combinations of dinucleotides are allowed while the other represents errors, since each base is defined by 2 colors. For example CA and AT produce two adjacent changes. B, Y, G, R represent the blue, yellow, green and red colors respectively.

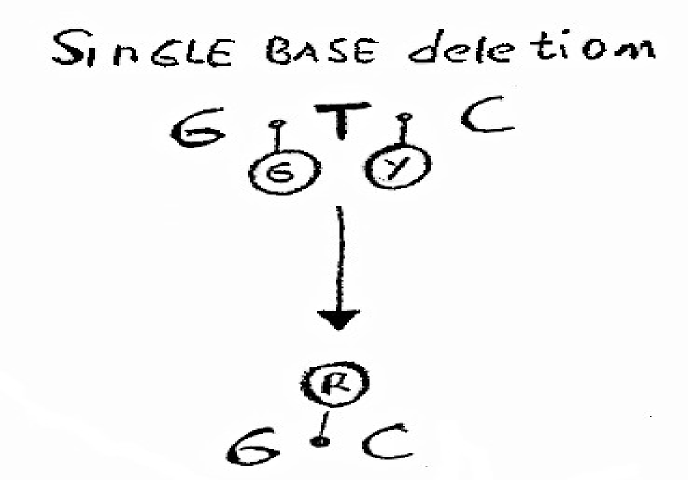


Figure 5.4.5: this figure shows a single deletion in the GTC sequence. The number of observed transitions has decreased from 2 down to 1 which correspond to the sequenced bases G and C. B, Y, G, R represent the blue, yellow, green and red colors respectively.

5.5 SOLiD READS CLEANING

During the library preparation some contaminants could be introduced due to the non-perfect efficiency of the RNA enrichment phase. The cleaning process by contaminants is necessary for two reasons: (i) contaminants adversely affect bioinformatics analysis, (ii) it is necessary when you want to publish the sequenced data. It was estimated that at least 0.4% of the sequences stored in the primary biological database contains various type of contaminants [Falgueras et al. 2010].

Furthermore, if the molecules to be sequenced are smaller than the size of the produced reads, the adapter which is linked to the 3' end of the DNA construct will be sequenced. As result some of the sequenced reads could contain both the sequence of interest and adapter. This phenomenon mostly affects micro-RNA experiments where RNA species are smaller than the reads size, but also in the case of nonspecific hybridization where dimers could be generated due to the library preparation.

There are many tools able to filter contaminants and adapters, such as *SeqTrim* [Falgueras et al. 2010] and *cutadapt* (a program developed by Marcel Martin) that works for base-space data. The SOLiD RNA-seq reads was cleaned using an implementation of PASS program [Campagna et al. 2009]. In figure 5.5.2 is reported a simple work-flow of the described pipeline.

Concluding, the presence of small DNA inserts in a SOLiD RNA-seq library can involve in the sequencing of the adapters in accordance with DNA constructs. Sequencing always starts from FWD1 (figure 5.5.1) and proceeds until reaching the “IA” adapter; the sequencing from primer REV1 proceeds until reaching P1 adapter so IA and P1 will be sequenced together with the insert.

In the cleaning process both P1 and IA are recognized and removed from sequencing products but, if the size of the cleaned reads is lower than a certain threshold, the reads are discarded (see figure 5.5.2).

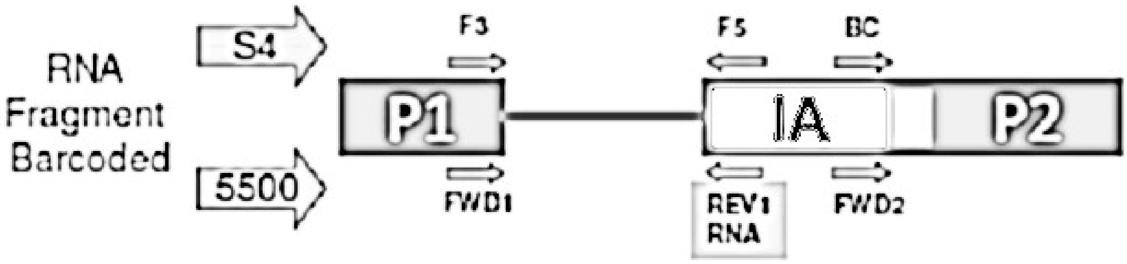


Figure 5.5.1: Simplified diagram of the DNA construct designed for paired-end sequencing. Firstly, the insert is linked to the adapter P1 then the IA (internal adapter) and the adapter P2 was bound to the 3' end of the insert. The “barcode” sequence (BC) allows the association of each DNA construct to its cDNA library. FWD1, REV1 and FWD2 represent the sequencing primers.

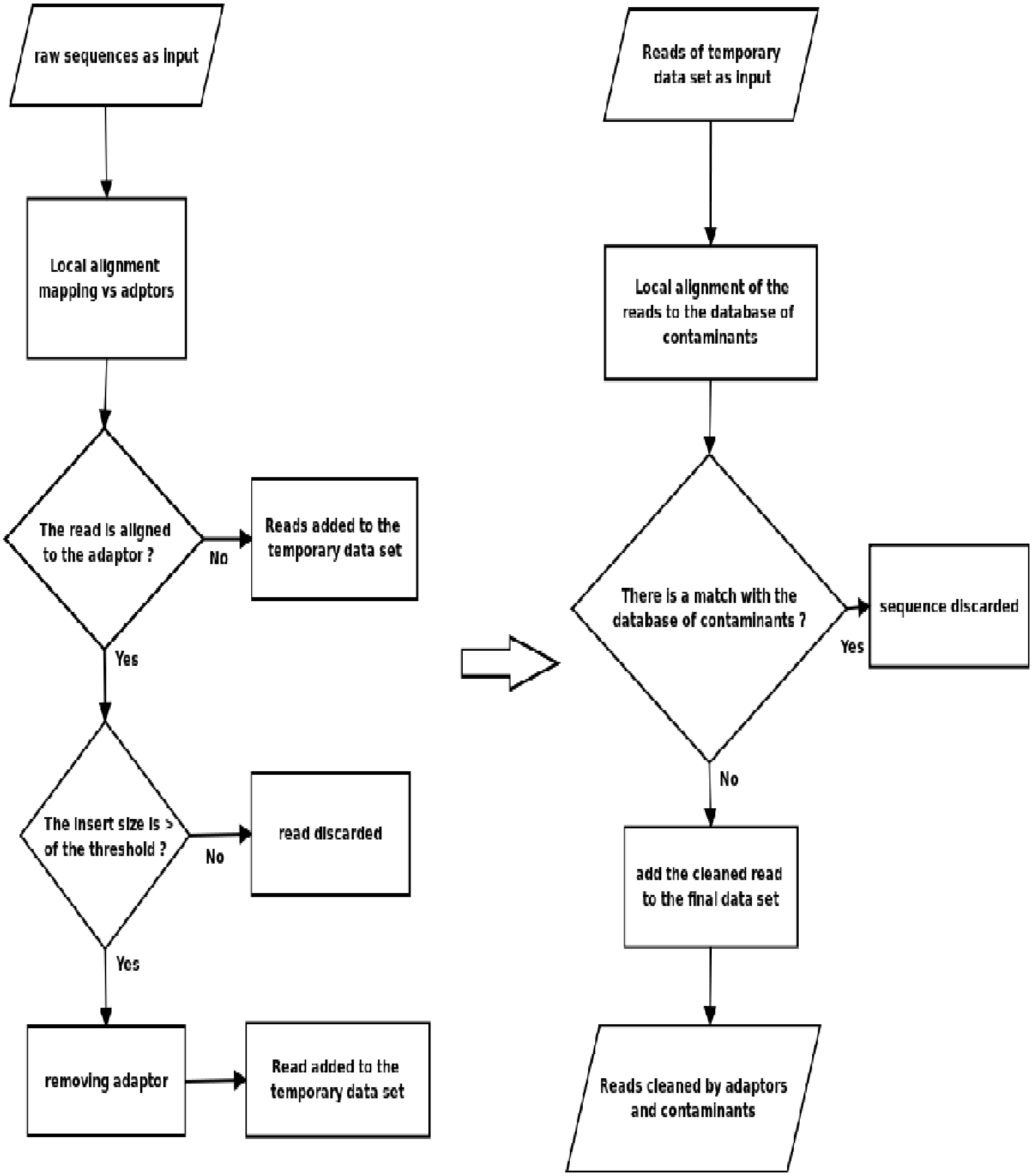

raw sequences as input
Local alignment mapping vs adptors
The read is aligned to the adaptor ?
No
Reads added to the temporary data set
Yes
The insert size is > of the threshold ?
No
read discarded
Yes
removing adaptor
Read added to the temporary data set
Reads of temporary data set as input
Local alignment of the reads to the database of contaminants
There is a match with the database of contaminants ?
Yes
sequence discarded
No
add the cleaned read to the final data set
Reads cleaned by adaptors and contaminants

Figure 5.5.2: Work-flow describing how PASS removes adapters and contaminants by SOLiD data. In the first step (left part of the diagram), the reads containing the adapter were cleaned and stored in a temporary data set. In the second step (right part of the diagram) the saved reads were checked for the presence of various contaminants: ribosomal, mitochondrial, tRNA, primers and cloning vectors.

Cleaning results of the sequenced SOLiD run 1

Sample	Raw	Passed	Cleaned	Discarded by size	Discarded by quality
F3_982-6-1	59.17M	39.68M	1.04M	14.84M	3.60M
F5_982-6-1	59.17M	41.85M	0.17M	13.38M	3.77M
F3_982-6-2	101.97M	87.35M	1.80M	7.36M	5.44M
F5_982-6-2	101.97M	89.36M	0.17M	6.61M	5.81M
F3_1186-6-1	106.12M	68.55M	1.55M	30.41M	5.58M
F5_1186-6-1	106.12M	72.94M	0.25M	27.84M	5.07M
F3_1186-6-2	65.26M	40.06M	1.12M	21.14M	2.93M
F5_1186-6-2	65.26M	42.23M	0.19M	19.80M	3.02M
F3_1186-7-1	74.91M	66.33M	1.64M	3.01M	3.91M
F5_1186-7-1	74.91M	67.86M	0.12M	2.42M	4.50M
F3_982-RG-6-1	94.62M	35.31M	3.93M	49.63M	5.72M
F5_982-RG-6-1	94.62M	42.16M	0.74M	47.58M	4.13M
F3_982-RG-6-2	53.30M	47.94M	0.34M	2.17M	2.84M
F5_982-RG-6-2	53.30M	47.38M	0.04M	1.99M	3.87M

Table 5.5.3: Statistics of the cleaned reads in the SOLiD run (07/2011). The first column represents the name of the sample, the second column the total number of sequenced reads for each sample (Raw), the third column the number of reads without sequenced adapter (passed), the fourth column the number of reads cleaned by adapter (cleaned), the fifth column the number of filtered sequences by minimal size (discarded by size 20 bp) and the last column the number of filtered reads by low overall quality (discarded by quality).

Sample	***E. coli***	**tRNA**	**ribosomal**	**mitochondrial**	**UniVec**
F3_982-6-1	0	835	210079	11	56
F5_982-6-1	0	227	116034	0	32
F3_982-6-2	0	1781	496704	66	207
F5_982-6-2	0	432	295747	22	40
F3_1186-6-1	0	1596	648561	14	150
F5_1186-6-1	0	381	378957	0	38
F3_1186-6-2	0	848	411207	4	57
F5_1186-6-2	0	245	229081	1	44
F3_1186-7-1	0	935	893692	139	37
F5_1186-7-1	0	187	497837	21	13
F3_982-RG-6-1	0	1848	86704	362	28
F5_982-RG-6-1	0	831	48442	37	15
F3_982-RG-6-2	0	976	204569	30	67
F5_982-RG-6-2	0	239	113586	2	31

Table 5.5.4: Statistics of the removed contaminants in the SOLiD run (7/2011). The first column represents the considered sample; the second column the number of recognized sequences as part of *E. coli* genome; the third column the number of reads recognized as tRNA; the fourth column represents the number of reads recognized as ribosomal sequences; the fifth column the number of reads recognized as part of mitochondrial genome and finally the number of reads recognized as UniVec sequences.

Cleaning results of the sequenced SOLiD run 2

Sample	Raw	Passed	Cleaned	Discarded by size	Discarded by quality
F3_1186-1	48.59 M	38.82 M	5.62 M	4.14 M	0.01 M
F5_1186-1	48.59 M	44.41 M	0.11 M	4.07 M	0.01 M
F3_982-1	34.49 M	29.55 M	3.36 M	1.59 M	0.00 M
F5_982-1	34.49 M	32.85 M	0.07 M	1.57 M	0.01 M
F3_985-1	62.03 M	51.02 M	8.32 M	2.68 M	0.01 M
F5_985-1	62.03 M	59.41 M	0.10 M	2.51 M	0.01 M
F3_1186-2	49.92 M	43.60 M	4.95 M	1.37 M	0.01 M
F5_1186-2	49.92 M	48.47 M	0.08 M	1.37 M	0.01 M
F3_982-2	43.42 M	37.09 M	4.91 M	1.40 M	0.01 M
F5_982-2	43.42 M	41.98 M	0.08 M	1.35 M	0.01 M
F3_985-2	40.91 M	29.34 M	8.03 M	1.81 M	0.00 M
F5_985-2	40.91 M	39.57 M	0.09 M	1.26 M	0.01 M
F3_1186-3	42.69 M	35.02 M	5.61 M	2.06 M	0.01 M
F5_1186-3	42.69 M	40.55 M	0.13 M	2.01 M	0.01 M
F3_982-3	45.11 M	38.64 M	4.55 M	1.90 M	0.01 M
F5_982-3	45.11 M	43.17 M	0.10 M	1.84 M	0.01 M
F3_985-3	41.17 M	34.46 M	4.63 M	2.07 M	0.01 M
F5_985-3	41.17 M	39.05 M	0.11 M	2.00 M	0.01 M
F3_1186-4	62.4 M	52.11 M	6.66 M	3.62 M	0.01 M
F5_1186-4	62.4 M	58.65 M	0.14 M	3.61 M	0.01 M
F3_982-4	45.57 M	37.90 M	5.89 M	1.78 M	0.01 M

F5_982-4	45.57 M	43.66 M	0.14 M	1.76 M	0.01 M
F3_985-4	59.48 M	50.14 M	7.72 M	1.61 M	0.01 M
F5_985-4	59.48 M	57.81 M	0.12 M	1.55 M	0.01 M
F3_1186-5	45.25 M	38.96 M	4.42 M	1.87 M	0.01 M
F5_1186-5	45.25 M	43.26 M	0.11 M	1.87 M	0.01 M
F3_982-5	62.29 M	52.67 M	6.96 M	2.66 M	0.01 M
F5_982-5	62.29 M	59.48 M	0.17 M	2.63 M	0.01 M
F3_985-5	70.68 M	60.95 M	7.65 M	2.07 M	0.01 M
F5_985-5	70.68 M	68.54 M	0.10 M	2.02 M	0.02 M

Table 5.5.5: Statistics of the cleaned reads in the SOLiD run (08/2012). The first column represents the name of the sample, the second column the total number of sequenced reads for each sample (Raw), the third column the number of reads without sequenced adapter (passed), the fourth column the number of reads cleaned by adapter (cleaned), the fifth column the number of filtered sequences by minimal size (discarded by size 20 bp) and the last column the number of filtered reads by low overall quality (discarded by quality).

Sampl	***E. coli***	**tRNA**	**ribosomal**	**mitochondrial**	**UniVec**
F3_1186-1	0	1620	683720	18	15386
F5_1186-1	1	209	118226	4	8214
F3_982-1	1	1073	1386856	18	7679
F5_982-1	4	247	478396	7	3599
F3_985-1	0	1423	915149	19	13741
F5_985-1	1	230	132948	4	6391
F3_1186-2	1	1208	333396	18	5545
F5_1186-2	0	299	110318	5	2578
F3_982-2	1	1016	2042014	34	6201
F5_982-2	0	176	829234	3	2873

F3_985-2	1	1035	1161296	24	5417
F5_985-2	0	212	466012	6	2560
F3_1186-3	1	847	1158338	25	9273
F5_1186-3	0	238	362944	6	4714
F3_982-3	1	1264	1571633	23	9136
F5_982-3	1	395	509583	7	4494
F3_985-3	0	979	764450	26	10202
F5_985-3	0	332	169670	2	5228
F3_1186-4	2	702	824824	11	15124
F5_1186-4	1	184	130367	9	8374
F3_982-4	0	680	1876017	42	6874
F5_982-4	0	138	711538	8	3328
F3_985-4	1	1188	3193477	32	8900
F5_985-4	1	301	1263021	9	4141
F3_1186-5	3	1334	921390	18	11580
F5_1186-5	0	370	214438	7	5856
F3_982-5	1	1842	2827875	24	13889
F5_982-5	1	392	1079439	14	5973
F3_985-5	2	1061	607096	44	12940
F5_985-5	1	164	188309	10	5865

Table 5.5.6: Statistics of the removed contaminants in the SOLiD run (8/2012). The first column represents the considered sample; the second column the number of recognized sequences as part of E. coli genome; the third column the number of reads recognized as tRNA; the fourth column represents the number of reads recognized as ribosomal sequences; the fifth column the number of reads recognized as part of mitochondrial genome and finally the number of reads recognized as UniVec sequences.

5.6 Color space encoding

The “Velvet” assembly program [Zerbino et al. 2008] supports the color encoded reads (color encoding not to be confused with 2 base encoding). The color encoding was originally implemented to use base space programs with color space sequences and it is based on the simple substitution of the colors indicated as '0', '1', '2' and '3' to the symbols 'A' , 'C', 'G' and 'T'. Normally, these changes allow mappers designed for base space data to map color space sequences in a simplistic way, if also the reference genome is 2 base encoded (double encoding).

In contrast with mappers, the “Velvet” program conceive the encoded reads as real color-space data, so it supports the main rules of the color space technology. For example a typical mapping problem is represented by reads which should be complemented and inverted in the case of base space assembly. In the case of color space they should be only inverted.

The color encoding program developed for this project considers all information that will be applied in the following steps: the first base, the first color and colors that must be encoded to ACGT space. For instance, the 2-base encoding of the color space sequence “T13210312031023010320103213210102301” produces a “G” that represents the first base of the sequenced read obtained from 2 base encoding of “T” and “1”. Furthermore the colors represented by '0', '1', '2' and '3' will be replaced with 'A', 'C', 'G' and 'T' respectively.

<u>Example</u>

color-space sequence:

```
>NNN_NNN_NNN
T13210312031023010320103213210102301
```

encoded sequence:

```
>NNN_NNN_NNN G
TGCATCGATCAGTACATGACATGCTGCACAGTAC
```

The “converted base G” added to the end of the read name, will be recovered by another program used in the “multiple color space conversion” (see paragraph 5.9). Each read will contribute to this process.

The decreasing of the ligation efficiency during the sequencing cycles cause a high number of sequencing errors, especially at the 3' end of the SOLiD reads. The negative effect of sequencing errors can be minimized if the encoded reads are adequately trimmed as shown in figure 5.6.1. The best trimming was obtained analyzing the higher N50 assembly acquired in different set conditions. The assemblies were generated by Velvet program [Zerbino et al. 2008] using the same SOLiD data set.

The encoding and subsequent trimming of color space reads are addressed by the developed program “*2csfastq_1csfastq”*. The sequenced RNA-seq library are analysed using the following set.

Encoding setting:

```
2csfastq_1csfastq \
-csfastq1 reads_file1 -csfastq2 reads_file2 \
-tags _F3 F5-RNA -t1 30 -t2 0 -double-encoded \
> outputfile
```

where “*-csfastq1 reads_file1” indicates the first fastq file to set as input, “-csfastq2 reads_file2” indicates the second fastq file to set as input, “-tags _F3 F5-RNA” set the tag of the paired-end names, “-t1 30 -t2 0” set the number of bases to trim at 3' end, “-double-encoded” indicates that the reads must be color encoded and “outputfile” is the output file containing the encoded reads.*

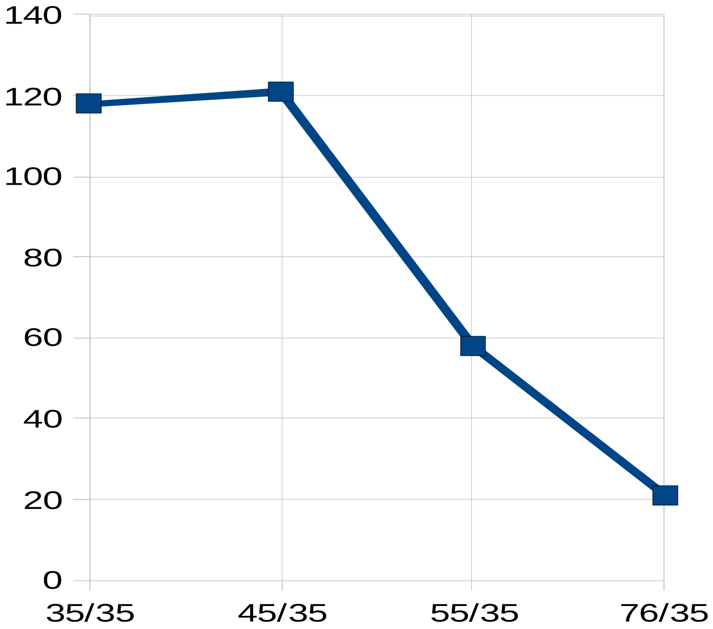


Figure 5.6.1: Effect of trimming on N50 assembly. The trimming regards only the reads of 75bp in length and it is referred to the 3' end. In the category 35/35 the reads were trimmed of 40bp; in the category 45/35 the reads were trimmed of 30bp; in the category 55/35 the reads were trimmed of 20bp and finally, in the category 76/35 the reads were not trimmed. The Velvet program [Zerbino at al. 2008] was used to assembly the color space data set.

5.7 Multiple-Kmer assembly

The assembly was performed using the program Velvet (Zerbino et al. 2008). This program is able to assembly color space reads basing on De Bruijn graph which are mathematical structures used to model relationships between pair of objects of a certain collection. The De Bruijn graph allows to reduce the complexity of the assembled data, so it represents a mathematical way to compact data structures. Velvet employs directed graphs where edges are directed from one node to another one, in a compact representation based on short words (k-mers). During the assembly process the program selects the short sequences basing on the overall base quality, then it calculates and evaluates the graph and finally, it saves the output relative to the assembled color space contigs.

As indicated in the name of the method, the multiple-Kmer assembly [Surget-Groba at al. 2010] takes into consideration several size of *k*mer, each one used in a separate assembly process. Thus the transcriptome of a non model organism must be assembled several times as the number of chosen k-mers. Generally, the best *k*mer for a given assembly depends by the sequencing depth, sequence error rate and, the complexity of the genome/transcriptome to be assembled [Simpson et al. 2009]. The Kmer must be chosen in a wide range of size but could be limited by the computing resource.

In the RNA-seq assembly the expected coverage of each transcript is not constant. Zerbino and Birney have demonstrated that many assemblies obtained using different Kmer size are more representative of the transcriptome complexity than the better one from a single Kmer (Zerbino at al. 2008). The authors evidenced how higher *k*-mer size will theoretically result in a more contiguous assembly of highly expressed transcripts while poorly expressed transcripts will be better assembled with small *k*mer size.

The chose of the range size depends by two factors: the read length and computing resources. Computing resources should be adequate for RAM and number of required CPUs (especially for small k-mer < 23 bases). The reads size represents the maximum set of k-mer size. SOLiD paired-end sequencing produces forward reads of 75 bp and reverse reads of 35 bp; basing on result obtained from the assembly tests the inferior limit of Kmer size was set to 21, while the superior limit was set to 33.

The entire SOLiD run has too much data to be assembled in a single step, so it was divided into three parts, each one corresponds to a considered phase of the blastogenetic cycle: 9/8/2, 9/8/6 and 11/9/6. Since the considered k-mers are: 21, 23, 25, 27, 29, 31 and 33, the total number of assemblies is 21.

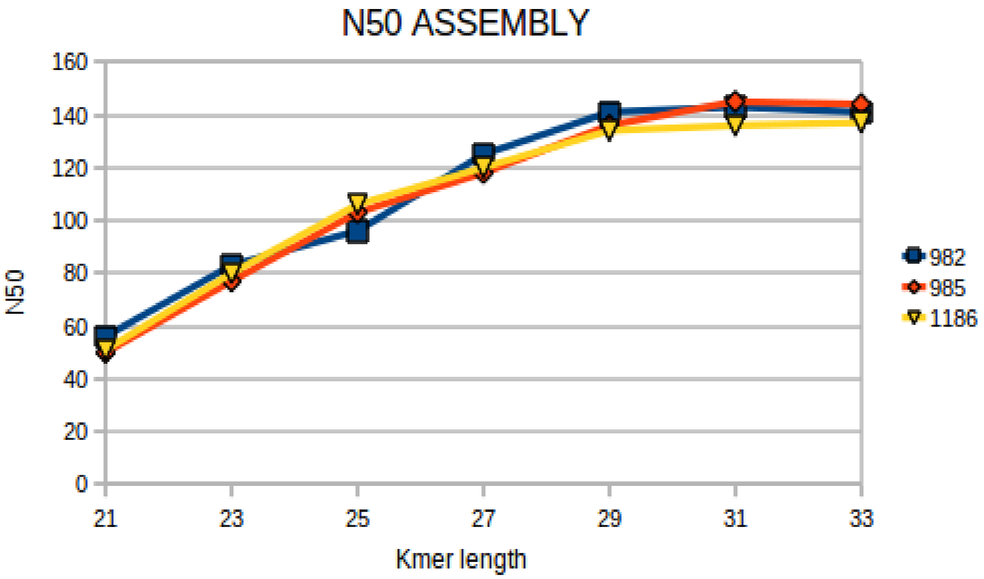


Figure 5.7.1: N50 assembly values obtained using the Kmer size of 21, 23, 25, 27, 29, 31 and 33 bases. The results are referred to the blastogenetic phase 9/8/2. The reads were assembled using the Velvet program.

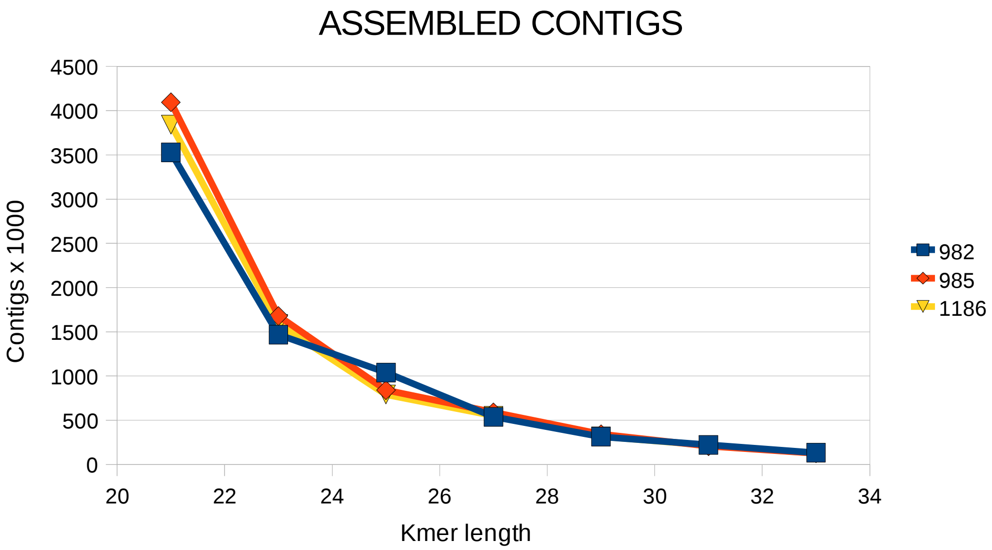


Figure 5.7.2: Number of assembled contigs obtained using the Kmer size of 21, 23, 25, 27, 29, 31 and 33 bases. The results are referred to the blastogenetic phase 9/8/2. The reads were assembled using the Velvet program.

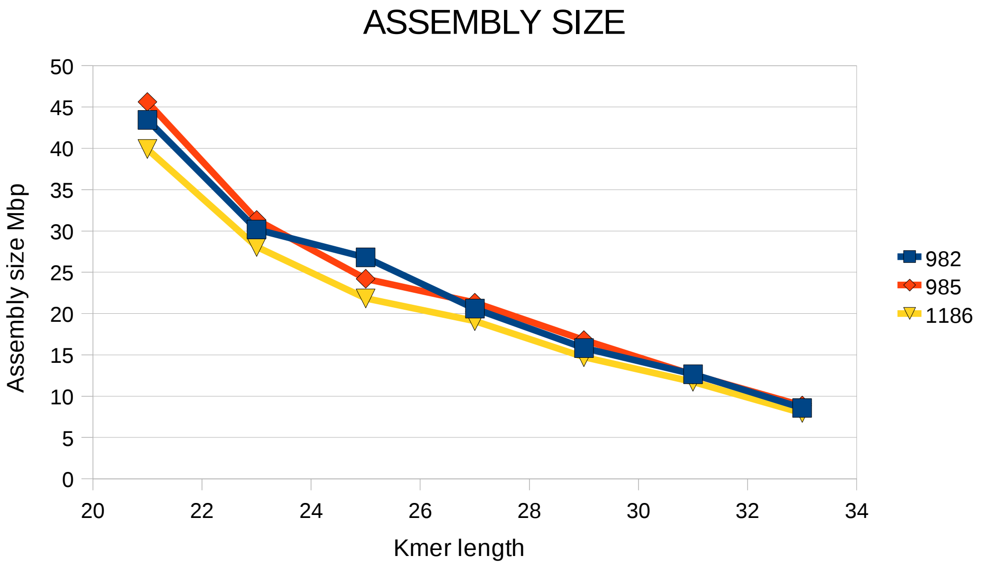


Figure 5.7.3: Values of assembly size obtained using the Kmer size of 21, 23, 25, 27, 29, 31 and 33 bases. The results are referred to the blastogenetic phase 9/8/2. The reads were assembled using the Velvet program.

Setting of Velvet program

velveth_de OUTPUT_PATH/ -strand_specific -shortPaired -fastq encoded_reads.fastq

velvetg_de OUTPUT_PATH/ -cov_cutoff auto -exp_cov auto -min_contig_lgth 100 -ins_length 300 -clean yes

For setting details please see the manual of the program.

5.8 Assembly errors reduction using MATRA

According to the simulation (see Chapter 4) a consistent fraction of contigs resulting by de novo transcriptome assembly are expected to be badly assembled. Assembly errors depend by many factors as the presence of sequencing errors, low complexity sequences, repeated domains and finally, transcript isoforms that are involved in the assembly of chimeric contigs.

The most of consensus are well assembled in more assemblies, while errors seem to be more specific to Kmer setting. In this sense, the assembly errors from multiple-Kmer assembly could be recognized as less redundant sequences less representative of the good ones. MATRA is the acronym of (*Minimizing Assembly errors Through Redundancy Analysis*) and represents a new method to minimize assembly errors through a selection of redundant sequences (see Chapter 4 for further details about the method).

The pool of contigs obtained from the multiple-Kmer assemblies was analysed using MATRA: 436.172 redundant consensus were selected and confirmed by at least 6 contigs coming from different assemblies. After redundancy selection the consensus were subjected to redundancy removing and *Multiple Color Space Conversion.*

5.9 Multiple color space conversion (MCSC)

After MATRA processing the contigs was clustered basing on their similarity. Substantially, at the end of the procedure, the redundant contigs coming from different assemblies were removed and replaced by the more representative contig of each cluster. This step was performed using the CD-HIT-EST program [Li et al. 2006] set to the default values (for details about the strategy please see “Chapter 4 – Validation of assembly method”). The “multiple color-space conversion” is a method based on read coverage analysis. The sequence coverage allows to identify the reliable start sites necessary to execute the 2 base encoding but also the undesired colors which are not supported by a sufficient number of clues. While the first information allows to know which are the start sites of the conversion, the second one ensures the colors that are confirmed by many reads. The multiple color-space conversion is substantially a multiple 2-base encoding where 2 base encoding is performed from different positions of the same color space contig (the start sites). The start sites can be inferred through the mapping of the color encoded reads on the color space assembly.

Color encoded reads mapping

The encoded reads were mapped on the color space assembly using PASS program [Campagna at al. 2009]. Through this step the following information are available:

- ✓ The coordinates of each alignment.
- ✓ The strand of the alignment mapped onto a reference contig.
- ✓ The converted bases as part of the read name.
- ✓ The structure of each alignment at a specified contig position.

Coverage threshold

Genes are expressed with different expression levels and the sequence coverage of the assembled contigs reflects this biological feature.

The following formula describes how to calculate the Tc_n "coverage threshold" basing on standard deviation and mean coverage. For a considered contig *n:*

$$\boldsymbol{Tc_n = Mean_n - k * Dev.Standard_n}$$

where n represents the contig n, "Tc_n" the threshold coverage, "Mean_n" the mean coverage, "$\mathrm{Dev.Standard}_n$" the standard deviation and finally, k represents the z-score that allows to parametrically modulate Tc_n.

According to the statistical analysis, the Tc_n allows to discriminate which colors can be converted and which ones are non informative. As consequence of this fact the color space conversion is more consistent with the true expected results.

Selection of start sites

Only reliable bases are used as potential start sites of 2 base encoding because the wrong start sites leads to incoherent results.. Under this restriction It is very important to check the coherence of 2 base encoding among the mapped reads. More precisely, a converted base is associated to its start site and those information are used to perform a 2 base encoding from a certain position. If 2-base encoding from different start sites produces the same results (multiple 2 base encoding are phased) than the corresponded “converted bases” could be selected as valid information because they are confirmed by multiple conversions.

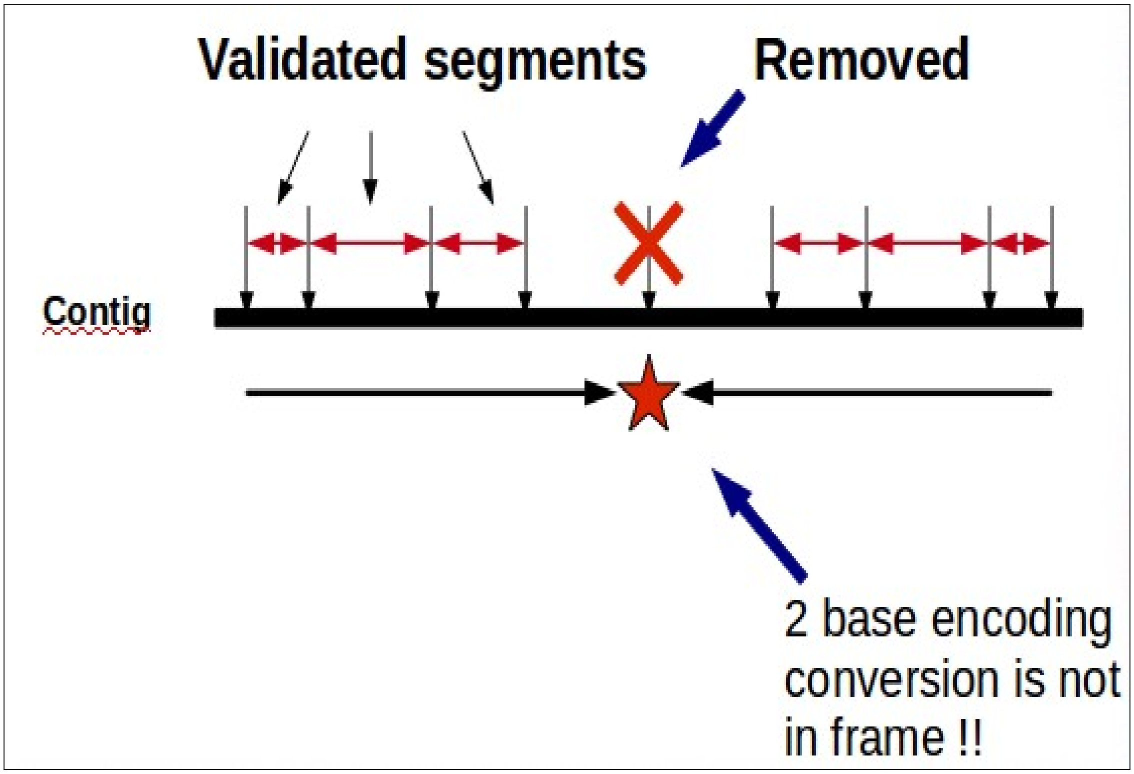


Figure 5.9.1: Starting point selection. For each putative starting point starts a bidirectional 2-base encoding. If the 2 base encoding frame is not phased with the other ones then the starting site is removed from the list.

Practically, the selection of start sites tries to minimize wrong conversions (see figure 5.9.1). A mis-assembly could be localized if two adjacent conversions will lead to different results (see figure 5.9.2). In this cases the program defines which positions could be converted and / or corrected through sequence coverage analysis.

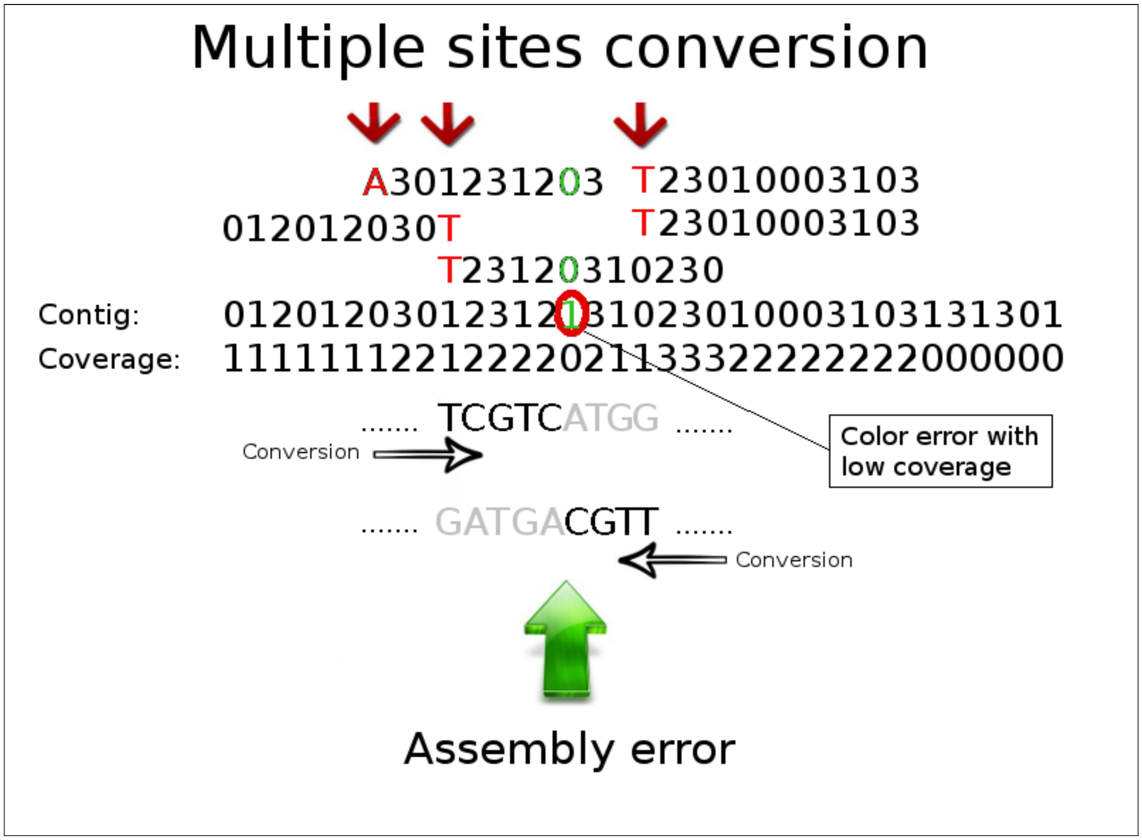


Figure 5.9.2: MSCC algorithm description. The red arrows indicate the starting points. Below is represented the assembled sequence and the coverage for each position. Despite the coverage represents the number of clues of each color position the strategy requires the overcoming of a minimal number of clues (in this case is 2) that confirms the converted bases (red color). In the example the 'A' base is confirmed by one sequence and it is discarded. The conversion proceeds on both directions up to the achievement of the minimal coverage.

5.10 MCSC method validation

The multiple color site conversion was tested for its ability to recognize assembling errors. The effect of different error types on results was evaluated as well as the efficiency of the conversion. This test should be seen as an indication of 'probability' to find errors. We can also assume that the probability to obtain real results (close to reality) is mainly related to the type of the test, and not all tests achieve the same probability. In this simulation an equal proportion of 3 types of errors: substitutions, deletions and insertions was generated. All these errors could affect the color-space conversion. In the simulation 1000 random base space sequences, which have a range of size of 200 to 650 bases, represented the data set to be analysed. The following table reports number and types of the considered errors.

Contigs	Substitutions	Deletions	Insertions
100	334	34	322

Table 5.10.1: Number and type of considered errors (one per contig). Four different coverage conditions were considered: 10X, 20X, 50X e 100X.

Firstly, the base space sequences without errors, were passed to dwgsim program in order to generate a data set of color space reads cleaned by inserted errors. Each base space sequence was converted to color space through a reverse 2-base encoding and then one error per contig was inserted at the middle of each sequence (table 5.10.1). This data set simulates the worst condition that includes an high frequency of different assembly errors.

Errors generate color space inconsistency that should be recognized and corrected by multiple color site conversion. The subsequent base space sequences resulted by MCSC were aligned onto the original error-free base space data set. Mapping results are evaluated in order to understand the efficiency of the program to threat assembly errors.

The following work-flow describes the overall strategy step by step.

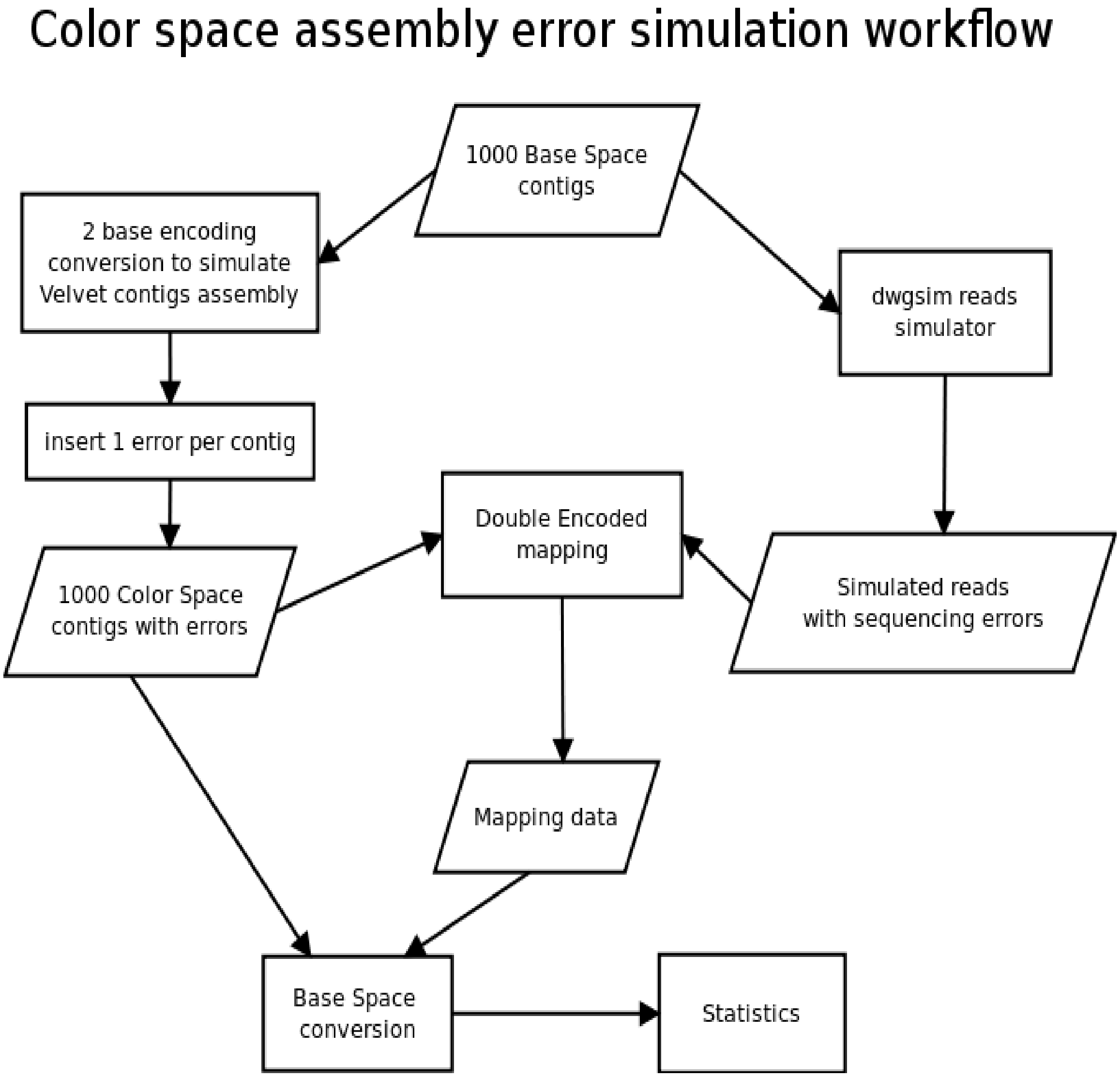

Color space assembly error simulation workflow
1000 Base Space contigs
2 base encoding conversion to simulate Velvet contigs assembly
dwgsim reads simulator
insert 1 error per contig
Double Encoded mapping
1000 Color Space contigs with errors
Simulated reads with sequencing errors
Mapping data
Base Space conversion
Statistics

Figure 5.10.2: Overall simulation work-flow. 1000 random sequences was converted to color-space using 2-base encoding and then 1 error per sequence was inserted to generate a color-space incoherence. The same base space contigs were used to generate 4 test set of simulated reads at 10X, 20X, 50X, 100X coverage. The encoded reads were mapped using PASS program [Campagna et. al. 2009]. Mapping data and modified color-space contigs have been used as input to finalize MCSC. All results were subjected to statistical analysis.

Sensitivity and specificity are based on the following definitions:

True positives

number of sequences on which the error was recognized

False positives

number of sequences on which the error was not recognized

False negatives

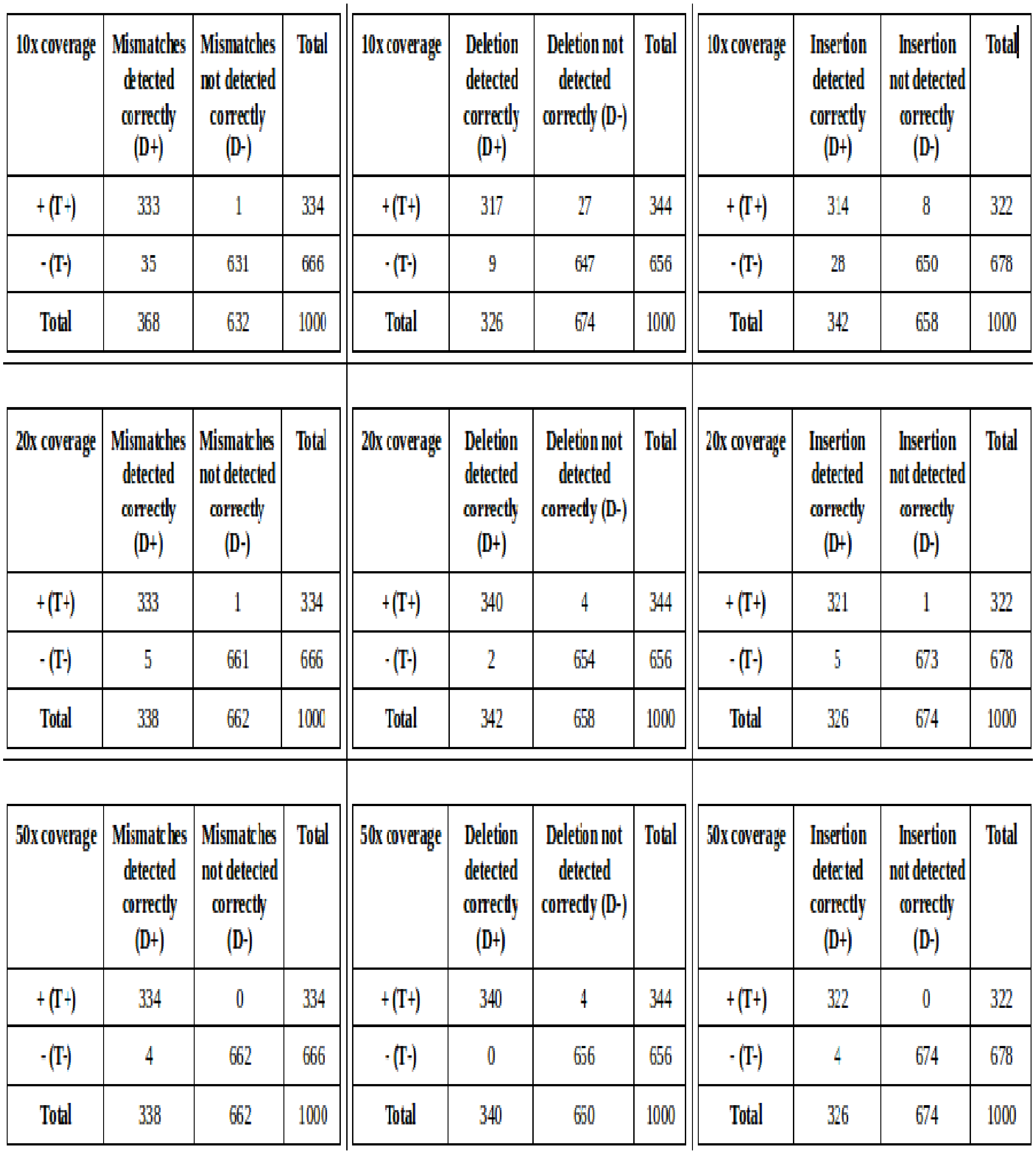
number of sequences on which a false error was not recognized

True negatives

number of sequences on which a false error was recognized.

Color space assembly subjected to MCSC was mapped (global alignment) onto the original “error-free” base-space sequences and results were analysed to distinguish assembly errors. Error positions and error types inferred by alignments, were compared with a list of inserted errors. True positives, false positives, false negatives and true negatives were counted to estimate the sensitivity and specificity to recognize errors.

Conclusion: the analysis reveals high specificity for all the considered sequence coverage conditions. The sensitivity can vary from 0.90 to 0.97 (depends by error type) at 10x coverage, from 0.98 to 0.99 at 20x coverage and from 0.988 to 1 at 50x coverage. The high fraction of insertions and deletions, represents the worst situation that should put in crisis the MCSC; this scenario is low probable considering real data. It is also noticeable that the inability of the mapper to recognize the 100% of insertions/deletions may cause an underestimation of the statistics of MCSC real performance.



10x coverage	Mismatches detected correctly (D+)	Mismatches not detected correctly (D-)	Total
+ (T+)	333	1	334
- (T-)	35	631	666
Total	368	632	1000

10x coverage	Deletion detected correctly (D+)	Deletion not detected correctly (D-)	Total
+ (T+)	317	27	344
- (T-)	9	647	656
Total	326	674	1000

10x coverage	Insertion detected correctly (D+)	Insertion not detected correctly (D-)	Total
+ (T+)	314	8	322
- (T-)	28	650	678
Total	342	658	1000

20x coverage	Mismatches detected correctly (D+)	Mismatches not detected correctly (D-)	Total
+ (T+)	333	1	334
- (T-)	5	661	666
Total	338	662	1000

20x coverage	Deletion detected correctly (D+)	Deletion not detected correctly (D-)	Total
+ (T+)	340	4	344
- (T-)	2	654	656
Total	342	658	1000

20x coverage	Insertion detected correctly (D+)	Insertion not detected correctly (D-)	Total
+ (T+)	321	1	322
- (T-)	5	673	678
Total	326	674	1000

50x coverage	Mismatches detected correctly (D+)	Mismatches not detected correctly (D-)	Total
+ (T+)	334	0	334
- (T-)	4	662	666
Total	338	662	1000

50x coverage	Deletion detected correctly (D+)	Deletion not detected correctly (D-)	Total
+ (T+)	340	4	344
- (T-)	0	656	656
Total	340	660	1000

50x coverage	Insertion detected correctly (D+)	Insertion not detected correctly (D-)	Total
+ (T+)	322	0	322
- (T-)	4	674	678
Total	326	674	1000

Table 5.10.3: MCSC results obtained from different coverage conditions. (+(T+), D+) represents the true positives , (+(T+), D-) represents the false positives, (-(T-), D+) represents the false negatives and (-(T-), D-) the true negatives. True positive, false positive , false negative and true negative are described in the previous page.

Type	10x coverage	20x coverage	50x coverage
Sens. substitutions	0.905	0.985	0.988
Spec. substitutions	0.998	0.998	1
Sens. deletions	0.972	0.994	1
Spec. deletions	0.960	0.994	0.994
Sens. insertions	0.918	0.985	0.987
Spec. insertions	0.988	0.998	1

Table 5.10.4: Summarized table of the calculated sensitivity (sens.) and specificity (Spec.) for 3 different errors (substitutions, deletions and insertions) and 3 different sequence coverage (10x, 20x, 50x). Sensitivity and specificity are calculated basing on results of table 5.10.3.

Setting description:

The simulated reads were generated by dwgsim-0.1.8 program from the samtool package (http://sourceforge.net/apps/mediawiki/dnaa/index.php?title=Whole_-Genome_Simulation).

Dwgsim-0.1.8 setting:

```
dwgsim  -y 0  -z 0  -d 100  -S 2 -c 0 \
-1 50 -2 50 -C COVERAGE_VALUE \
modified_chr10 \
simulated_reads
```

The "-c 0" is used to produce base space reads while "-c 1" is used for color space reads. The per base/color/flow error rate and the rate of mutation was set to the default values (respectively: 0.02 and 0.001). All simulated test sets were produced using the same seed, so they are comparable for number of reads, position and strand.

PASS program setting:

```
pass_v2.0_I+ \
-cpu 12 -double_encoded -flc 4 \
-d CS_ASSEMBLY.fasta \
-fastq ENCODED.fastq -fid 90 \
-sam -query_size 200 -b  > ENCODED.sam
```

where “-cpu” indicates the core processor number, “-double_encoded” enables the color encoded reads mapping, “-flc” set the low complexity filter, “-d” set the input color-space assembly, “-fastq” set the encoded reads to map, “-fid” set the minimal percentage of identity of the alignments, “-sam” set the SAM output format, “-query_size” set the max query size, “-b” allows to output only the best hit alignments and ENCODED.sam represents the output file containing the mapped reads.

Multiple Color Site Conversion program:

```
cs2bs_assembly \
-fasta CS_ASSEMBLY.fasta \
-sam ENCODED.sam \
-l 100 -n 0.1 -z 2 -C 1 \
> BS_ASSEMBLY.fasta
```

where “-fasta” allows to pass the input color-space assembly in FASTA format, “-sam” set the input alignments resulted by mapping, “-l” set the minimal length of the converted sequences to output, “-n” allows to filter the contigs having more than 10% of undefined bases (Ns), “-z” set the z-score useful to calculate the coverage threshold, “-C” set the minimal coverage for each converted base and finally, BS_ASSEMBLY.fasta represents the output file containing the base-space sequences.

MCSC applied to RNA-seq assembly

The de novo transcriptome assembly was converted using MCSC. The statistics of the results are reported in the following table:

Number of color space contigs	: 117.922
Color space assembly Size	: 18.779.887
Color space assembly undefined bases	: 0
Number of base space contigs	: 112.389
Max contig length	: 1.021
Basespace assembly size	: 17.967.040
Base space assembly undefined bases	: 11.444
Corrected bases around assembly errors	: 360.143
N50 of basespace assembly	: 160

5.11 Contigs elongation

After redundancy removing and color space conversion, the assembly could be treated as a base-space sequence. At this point of the analysis, many contigs of the assembly are orphan contigs and many other, belonging the same transcript, have overlapped ends that could be assembled.

The TGI Clustering tools (TGICL) (http://compbio.dfci.harvard.edu/tgi/software/) allows to cluster and assembly large datasets of EST or contigs typically from multiple-Kmer assembly. The clustering is performed using a slightly modified version of the NCBI's *megablast* and, clusters are assembled using CAP3 assembly program [Huang at al. 1999]. Several clustering and assembling iterations are necessary. For this specific purpose a new program was developed. It supports both color space and base space assemblies and allows to modulate very finely the sensitivity and specificity of the elongated contigs. For details about the efficiency of the program please see "Chapter 4 – Validation of the method".

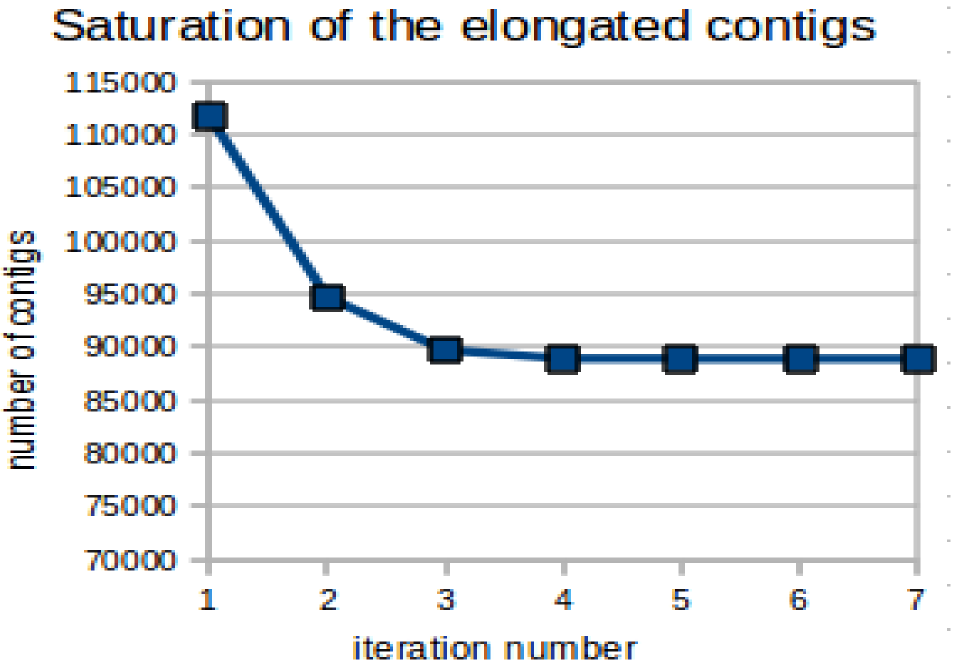


Figure 5.11.1: Effect of several iterations of contigs elongation. The X axis represents the number of assembly iteration, the Y axis represents the number of assembled contigs. As evidenced in the figure, the contigs elongation goes to saturation after 4 iterations.

Basically, the CS contigs elongation requires 3 operations described as follow: i) the contig redundancy must be removed using the cd-est-program; ii) the remaining contigs must be mapped using PASS [Campagna at al. 2009] one against to each other at the purpose to cluster consensus; iii) each cluster must be analysed in order to assemble the overlapped contigs which have compatible ends.

Basically, the efficiency of contigs elongation depends by 3 important points:

1) Clustering process is affected by the sequence similarity among contig ends.

2) The quality of the contigs elongation strongly depends on clustering.

3) The oriented paired-ends coming from a directed cDNA library should be assembled in the same orientation of their transcripts. In other words the specificity of the clustering process is guaranteed by the sequence identity, size and orientation of the alignments. The clusters must be cleaned by unwanted sequences.

Contigs elongation applied to RNA-seq assembly

As described in figure 5.11.1 the analysis regards a pool of 112.389 contigs with N50 of 158 bp. After 4 assembly iterations, the assembly was reduced to 88.906 contigs and N50 was increased up to 169 bases.

<u>Setting of the used programs:</u>

CD-HIT-EST program:

```
cd-hit-est -g 0 -T 12 \
-c 0.90 -n 8 \
-i pooled_assembly.fasta or elongated.fasta \ -o not_redundant_assembly.fasta
```

where “*-c 0.90” indicates a similarity threshold of 90%, “-n 8” set the sensitivity of the clustering analysis, “-i” set the input fasta file and “-o” sets the output fasta file.*

<u>PASS program:</u>

```
pass_v2.0_I+ \
-fasta  not_redundant_assembly.fasta \
-p 111111101111111 \
-d not_redundant_assembly.fasta \
-fid 90 \
-query_size 50000 \
-check_block 1000 \
-l -fle 70 -flc 3 -seeds_step 5 \
-sam -cpu 12 \
> alignments.sam
```

where “-cpu” indicates the number of core processors, “-flc” set the low complexity filter, “-d” set the input color-space assembly, “-fasta” set the input color-space assembly to map into itself, “-fid” set the minimal percent of identity of the alignments, “-sam” set the output format (SAM), “-query_size” sets the max query size, “-b” allows to output only the best hit alignments and alignments.sam represents the output file containing the alignments of mapped contigs.

Assembly program:

```
MULTIK-contigs \
-sam alignments.sam \
-fasta not_redundant_assembly.fasta \
> elongated.fasta
```

where “-sam alignments.sam” set the input file containing the alignments, “-fasta not_redundant_assembly.fasta” set the input file containing the contigs to be elongated and “elongated.fasta” is the output of the improved assembly.

5.12 Paired-end scaffolding

Scaffolding is the use of higher order information to group, orient and connect the assembled contigs. Many information could be used as paired-end sequences, mate-pair sequences, restriction maps, clone (cosmid/fosmid/BAC/PAC) maps. Furthermore, scaffolding may also be done using related genome information.

The developed program is based on the following work steps:

- ✓ The SOLiD paired-ends are mapped onto *de novo* transcriptome assembly using PASS program [Campagna at al. 2009).
- ✓ The analysis of the mapped paired-ends allows to recognize uniquely mapped reads on different contigs. The unique paired-end associations "links" are the main information to group contigs in scaffolds.
- ✓ The presence of transcript isoforms or conserved domains could generate wrong links that negatively affects scaffolding. Similarly to SSPACE program (*Boetzer at al. 2011)*, two conditions are checked: i) each association should have a minimal expected unique paired-ends; ii) if more associations are found, only ones which have a threshold ratio among second and first best link (in terms of unique paired-end number) less than a set value will be considered in the scaffolding. This set determines the possibility to modulate specificity and sensitivity of the scaffolding basing on mapping information.
- ✓ Links which have passed selection allow contig scaffolding. if possible, overlapped contigs are assembled using CAP3 program [Huang at al. 1999].

<u>Statistics of PE scaffolding</u>

The SOLiD reads coming from all experiments were mapped onto "*de novo* assembly" and alignments who map uniquely on different contigs, were saved into the same file: 12.702.776 links are found. The paired-end scaffolding of the pre-assembled *de novo* assembly (88.906 contigs with N50 of 169 bp) followed by redundancy removing has greatly improved the assembly. The following table summarizes the results.

	Total BASES	***N50***
Scaffolds	13.982.766 bp	283

	ITEM COUNT	***RANGE OF LENGTH***
Scaffolds	61.214	30 to 2882

Mapping setting:

```
pass_v2.2_I+ \
-csfastq reads_file -d assembly.fasta \
-fid 90 -sam -p 11111100111111 \
-b -cpu 12 \
>mapped.sam
```

where “-cpu” indicates the number of core processors, “-d” set the input file of the assembly, “-fastq” set the input file of color-space reads to map, “-fid” represents the minimal percent of identity of the alignments, “-sam” set the output format, “-b” set the output of best hit alignments, “-p” set the structure of the seed pattern, and “mapped.sam” represents the output file containing the alignments in SAM format.

Pairing setting:

```
pass_v1.7_I+ -program pairing \
-cpu 12 \
-sam1 mapped.sam \
-range 0 500 500 \
-pe_type 0 \
-tags _F3 _F5-RNA \
-append -no_header \
-ref assembly.fasta \
-unique_pair_out 1 \
-unique_pair 1 \
-o PAIRING/ \
```

“-program pairing” set the pairing function, “-cpu” indicates the number of core processors, “-sam1” set the input file of alignments, “range 0 500 500” set the expected max distance of mapped paired-ends, “-pe_type 0” indicates paired-end libraries, “-tags _F3 _F5-RNA” set the read tag of paired-ends, “-append -no_header” removes header from output and data are appended to the saved file, “-ref assembly.fasta” set the input assembly file for scaffolding, “-unique_pair_out 1” set the type of output (only unique links) and “-o PAIRING/” set the directory path to save data.

PE scaffolding setting

```
PE-scaffolding \
-tags _F3 _F5-RNA \
-sam UNIQUE_PAIR_OUT \
-fasta assembly.fasta \
-min_connect 2 \
-ambiguity 30 \
-not_included PE-level3-NI-1.fasta \
> level4.fasta \
```

“-tags _F3 _F5-RNA” set the tags of the paired-end names, “-sam UNIQUE_PAIR_OUT” set the input file containing of alignments, “-fasta assembly.fasta” set the file containing the assembled sequences, “-min_connect

2” set the minimal accepted links, “-ambiguity 30” set the threshold ratio (second best hit link number / best hit link number) *100 and “-not_included PE-level3-NI-1.fasta” set the output file where orphan contigs have been saved.

5.13 STM scaffolding

The RNA-seq assembly can be substantially improved by STM scaffolding. STM scaffolding allows to sort, orient and assemble the contigs of the same transcript basing on protein similarity. This is a very interesting alternative which requires the only condition to use a reference proteome of a specie phylogenetically close to the studied organism. This method, called STM (Scaffolding using Translation Mapping) uses the information obtained from the translation of the assembled contigs into amino acids to identify orthologous protein regions [Surget-Groba et al. 2010]. The proteome can be used to associate contigs among the same coding sequence of the two species. In this way, the translated contig mapped in the reference proteome, can be assembled into supercontigs and scaffolds, if overlap and quality conditions are checked. Furthermore, if the size of the reads used to assembly the transcriptome are greater than 70 bp the authors suggest to improve assembly adding the orphan reads (singleton). In this case the method is called (STM +). On the Contrary reads smaller than 70 bp should be removed from the final assembly (STM-) as the case of SOLiD data.

The STM scaffolding program was developed according to the description of the published algorithm [Surget-Groba et al. 2010].

Statistics of STM scaffolding

The number of initial contigs obtained from the multiple color-space conversion (MCSC) was 61.214. The contigs were mapped on a collection of the following protein database: I) FM1-filtered models protein database of *Ciona intestinalis; ii)* CIPRO 2.5 protein database.

The STM scaffolding has produced associations with 8747 of 61.214 contigs and scaffolds with 6624 of 7604 recognized proteins.

Mapping analysis indicated that the proteins were covered by contigs with the following percent in length: 34.74% of the proteins were covered more then 30% of their length, 18.31% of the reference proteins were covered more then 50% of their length, 9.29% of the reference proteins were covered more then 70% of their length. Concluding *B. schlosseri* transcriptome assembly and *C. intestinalis* proteome have low similarity. STM scaffolding gave small improvements to the final assembly that has reached 56234 contigs and N50 of 305 bp.

Program setting

```
STM-scaffolding \
-fasta level1.fasta \
-blastx level1.blast \
-percent_contig_len 70 \
-e_value 100 \
-not_included level2-orfans.fasta \
> level2-STM.fasta \
```

“-fasta” set the input file of assembly to be scaffolded, “-blastx” set the file of alignments resulted by blastx mapping, “-percent_contig_len” set the minimal percent contig length that must be covered by a protein sequence, “-not_included” allow to save the un-scaffolded contigs in a separate file.

5.14 Reads mapping

The sequenced reads of each RNA-seq library were mapped into *de novo* assembly using PASS program [Campagna et al. 2009]. Prior to mapping, PASS proceeds to execute several learning steps which allow to evaluate the best parameters that maximize the number of mapped reads.

This task is performed through mapping iterations of a small data set using different setting. Each mapping step cause a different trimming of the reads that must meet certain quality conditions. The statistical evaluation of each step allows to determine the best setting.

23 RNA-seq libraries were mapped onto *de novo* transcriptome assembly of *B. schlosseri* and a total of 545 million reads were aligned successfully.

5.15 COG and gene annotation

The cluster of orthologous groups (COGS) contain clusters of prokaryotes (COGS) and eukaryotes (KOGs). It was built comparing proteins (both predicted and characterized) encoded by complete genomes. Precisely, clusters of orthologous groups derives from 66 unicellular and 7 eukaryotic organisms including *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophilia melanogaster, Homo sapiens*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Encephalitozoon cuniculi [Tatusov et al. 2000]*. Furthermore they represents about 54% of the gene products and are classified into functional categories. The mapping of coding sequences against COG allows to group sequences through the functional transfer of COG proteins.

The assembled contigs were mapped against KOG database using BLASTX program [Altschul et al. 1990]. The alignments selected under certain conditions (e-value = 10^{-3} , the same protein must cover 90% of the contig length) were saved into a database of preprocessed data.

In this way contigs and scaffolds have been classified into functional categories. In addition, the assembled contigs were subjected to gene annotation using BLAST2GO program [Conesa at al. 2008]. COG alignments and gene annotation data were integrated into a database designed to be fast both for accessing and searching data.

5.16 Web-based interface for gene expression study

The web interface (WI) is part of a package that includes many programs written in C, C++ and perl that play different roles both for analyses and processing the data. Paired-ends mapping, genes/contigs classification into functional categories, RPKM analysis to discover false counts (exploiting directed cDNA library), graphical display, gene expression analysis, logical comparison of results, management of server/client connections, are some important tasks addressed by the developed tool.

The WI is thought in a compact structure that could be directly and quickly accessed from the network through a common Web browser. Just few seconds are required to elaborate results which will be presented in a well-designed graphical format. The analyses could be replicated using different statistical significance and/or including biological replicas or comparing results of different experiments. The statistical analyses of the experiments are directly and quickly accessible using a common Web browser.

After experiment's selection and setting, the WI recovers the required information from a database of preprocessed data. Consequently it elaborates information and finally it outputs the results.

Six important tasks:

(1) Gene expression analysis based on published work [Wang et al. 2010].

(2) Possibility to investigate for chimeric assembled contigs.

(3) Classification of differentially expressed genes into functional categories that allows a powerful overview of the genetic changes of different experimental conditions.

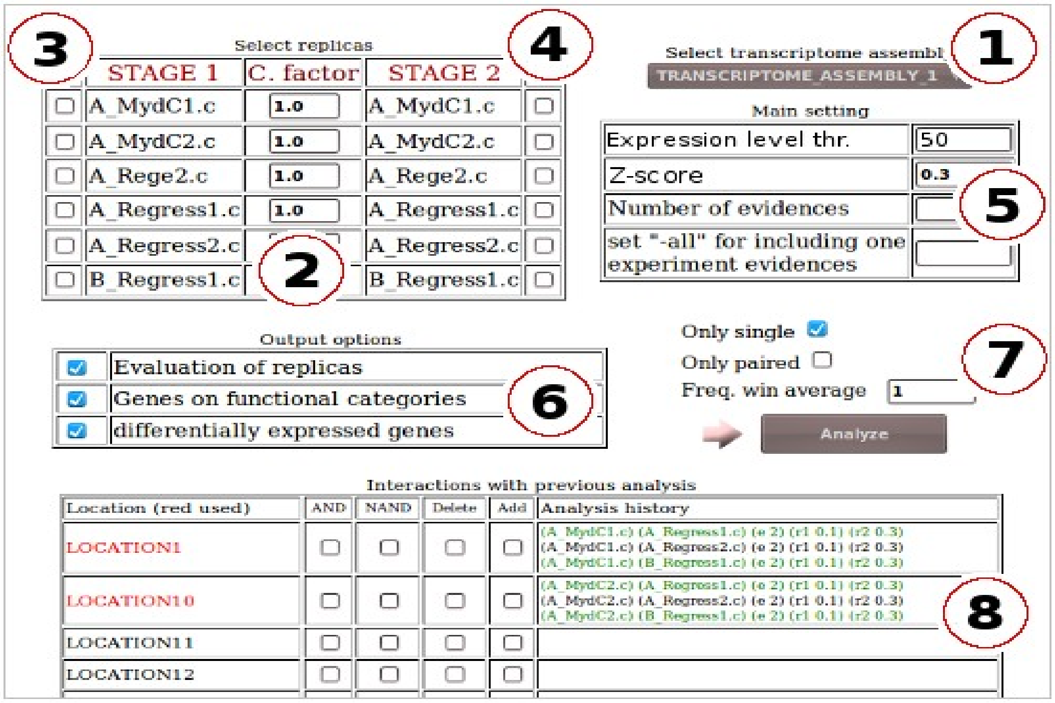
(4) Specific scoring system based on statistical significance of differentially expressed genes.

(5) It make possible to compare differentially expressed genes obtained from different RNA-seq analyses using logical operators.

(6) It allows to set the main parameters involved in the statistical analysis.

Setting

The WI consists of several input panels that should be used to set parameters and experiments to be analysed. Some parameters allow to set the statistical significance of the analysis, other parameters allow to set graphic information and other ones allow to compare results from different analyses. The following figure shows the appearance of the panels. The number inside red circles is associated to a function description.



Select replicas

	STAGE 1	C. factor	STAGE 2	
□	A_MydC1.c	1.0	A_MydC1.c	□
□	A_MydC2.c	1.0	A_MydC2.c	□
□	A_Rege2.c	1.0	A_Rege2.c	□
□	A_Regress1.c	1.0	A_Regress1.c	□
□	A_Regress2.c		A_Regress2.c	□
□	B_Regress1.c		B_Regress1.c	□

Select transcriptome assembl

TRANSCRIPTOME_ASSEMBLY_1

Main setting

Expression level thr.	50
Z-score	0.3
Number of evidences	
set "-all" for including one experiment evidences	

Output options

☑	Evaluation of replicas
☑	Genes on functional categories
☑	differentially expressed genes

Only single ☑

Only paired □

Freq. win average 1

Interactions with previous analysis

Location (red used)	AND	NAND	Delete	Add	Analysis history
LOCATION1	□	□	□	□	(A_MydC1.c) (A_Regress1.c) (e 2) (r1 0.1) (r2 0.3) (A_MydC1.c) (A_Regress2.c) (e 2) (r1 0.1) (r2 0.3) (A_MydC1.c) (B_Regress1.c) (e 2) (r1 0.1) (r2 0.3)
LOCATION10	□	□	□	□	(A_MydC2.c) (A_Regress1.c) (e 2) (r1 0.1) (r2 0.3) (A_MydC2.c) (A_Regress2.c) (e 2) (r1 0.1) (r2 0.3) (A_MydC2.c) (B_Regress1.c) (e 2) (r1 0.1) (r2 0.3)
LOCATION11	□	□	□	□	
LOCATION12	□	□	□	□	

(1) Choosing the assembly

The user can select the transcriptome assembly listed on the right of the Web page from the field list "Select transcriptome assembly". Through this function different assemblies could be compared.

(2) *Correction factors*

It is possible to set the correction factor of each replicas from panel "Select replicas" on the central column "C. factor". The normalized count of each transcript are altered by this set. The correction factor represents the correction of frequency ratio of the selected experimental conditions. All counts of the selected replicas will be normalized and than corrected before executing statistical analysis. If there aren't valid reasons to change this parameter, the suggested value is 1.

(3) and (4) *<u>Selection of the replicas</u>*

Before starting the analysis, the user should select the replicas checking the proper check box. STAGE 1 and 2 represent two experimental conditions. The statistical analysis is referred to the STAGE number 2. For example if we obtain some negative differentially expressed genes for a couple of experiments they are negative differentially expressed in the experiment associated to the STAGE 2. It is possible to combine the counts from two or more replicas as a single one. This function is enabled from the check-box located in the same row of “STAGE 1” and “STAGE 2” red tags (the check-boxes are not visible in the figure because they are covered by numbers inside red circles).

(5) *<u>Main setting</u>*

Expression level threshold: The differentially expressed genes will be selected on the basis of their expression levels. The WI evaluates which one is the major expression level of the 2 possible status and considers the higher one of each transcript. Firstly, expression levels are ordered by their normalized frequency and then compared to select the higher one. For example, if you set the value to 70, it means that you will select all differentially expressed genes having an expression level equal or greater than the 70 percent of the total transcripts.

Z-score: In statistics, a standard score indicates how many standard deviations an observation or datum is above or below the mean. It is a dimensionless quantity derived by subtracting the population mean from an individual raw score and then dividing the difference by the standard deviation. The highly differentially expressed genes have an high z-score and the better statistical significance (suggested 1.64).

number of evidences: The Web interface will select only differentially expressed genes confirmed simultaneously by a certain number of selected replicas (analysed separately in all possible combinations). This set could be very important in order to filter false positives; the default values is equal to the number of selected replicas that represents the superior limit.

all evidences: You need to add **-all** if you want to include also the expressed genes which are not present in both experiments. This set regards the genes expressed at very low level in only one of the two considered experimental conditions. The RNA-seq sensitivity in not sufficient to infer the expression level that it is forced to the minimum (non-zero), that allows gene expression analysis. This manipulation of the data allows to select differentially expressed genes that should not be considered because the lacking of data. By default this functionality is disabled.

(6) *Output options*

The output includes 3 sections: the statistics of gene expression analysis; the differentially expressed genes grouped into functional categories (summarized by a graph plus the associated table) and finally, the table of differentially expressed genes which reports many information. It is possible to include or exclude one of these sections through the proper check box.

(7) *Other sets*

This table includes some input fields inherent the plot setting and some parameters that should be used in the statistical analysis.

Only single: This set is used in the statistical analysis. The selection of this check box will excludes paired-end RPKM.

Only paired: This set is used in the statistical analysis. The selection of this check box will excludes single-end RPKM.

Interaction with previous analysis

This function allows to compare previous analyses with new ones. Logical operators make possible to compare and selects the differentially expressed genes obtained from different combination of RNA-seq experiments.

Saving and deleting: Before starting the analysis, you should select the “add” check box of the desired “LOCATION”. At the end of the process the differentially expressed genes will be saved into the selected record and then a description of the parameters appears into the field "analysis history". The next analysis saved in the same record will cause the merging of new data to the previous one. In order to erase a record content, it is necessary to select the “delete” check box. The content of the record will be erased after clicking the “analysis” button.

Logical operations: The differentially expressed genes among different analyses could be selected using a logical “AND”. The logical operator “NAND” executes the opposite operation of “AND”. Both logical operators could be simultaneously selected in the same query for all saved analyses.

Results

The analysis starts when you click the "analysis" button. A page containing several graphs and tables appears. It is mainly formed by 3 sections: i) the MA plot; ii) the graph of functional categories that includes a related table; iii) the differentially expressed genes that includes a lot of information.

MA plot:

The MA plot is useful to understand the dispersion of the data among different experiments or replicas.

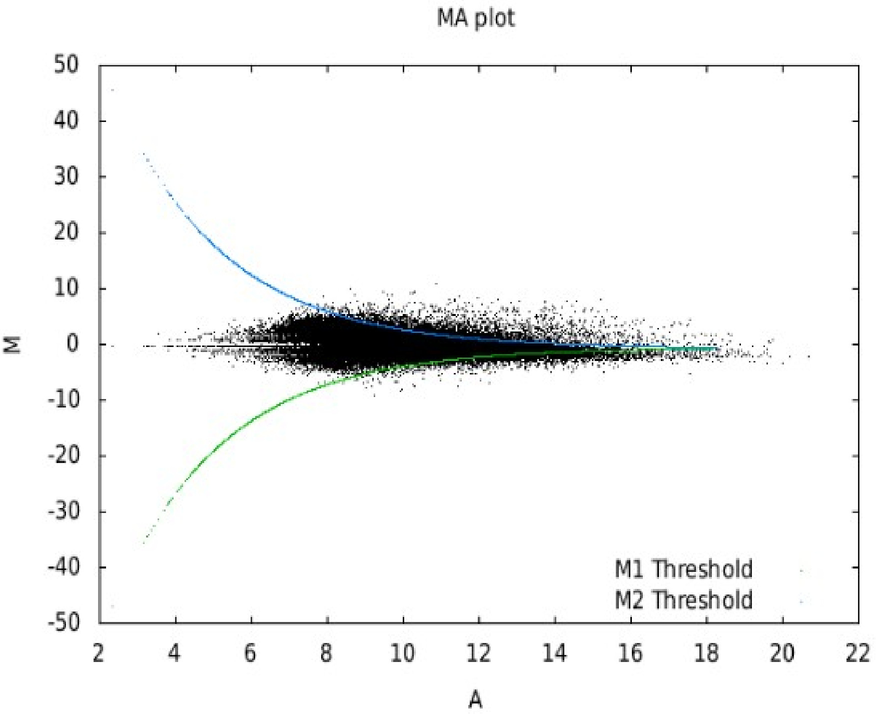


Figure 5.16.1: MA plot generated by the Web Interface. The Z-score threshold is set to 1.64 (Green and blue lines). The selected experiments for the statistical analysis are referred to the 9/8/2 and 11/8/6 phases of the blastogenetic cycle. It is shown how transcripts with low expression level have a strong dispersion of the signals.

The Y axis indicates the log2 of the difference among gene expression levels (M) and the X axis indicates the log2 of the average among gene expression levels (A). In the case of the MA-plot the differentially expressed genes produce a "points cloud". The points outside the Z-score thresholds, which are indicated with blue and green lines, represent the putative differentially expressed genes (see figure 5.16.1).

Functional categories

"COG" stands for Cluster of Orthologous Groups of proteins. The proteins included in each COG are assumed to have evolved from a single ancestral protein and, therefore, are either orthologs or paralogs. Ortholog proteins are molecules from different species that evolved by vertical descent (speciation), and typically retain the same function as the original. Paralog proteins derived from gene duplication, and may evolve new functions that are related to the original one. COGs were identified using an all-against-all sequence comparison of the proteins encoded in completely-sequenced genomes. In considering a protein from a given genome, this comparison would reveal the protein from each of any other genomes most similar to the other. Each of these proteins are in turn considered. If a reciprocal best-hit relationship between these proteins (or a subset) is revealed, then those that are reciprocal best-hits will form a COG. Thus, a member of a COG will be more similar to other members of the COG than to any other protein from the compared genomes, even if the absolute similarity is low. The use of the best-hit rule, without the constraint of an arbitrarily-chosen statistical cut-off, therefore accommodates both slow- and fast-evolving proteins. However, one constraint that was imposed is that a COG must include one protein from at least three phylogenetically distant genomes.

Using COGs, there are three general kinds of information:

1. Annotation of proteins. Known functions (and two- or three-dimensional structures) of one COG member can often be directly attributed to the other members of the COG. Caution must be used here, however, since some COGs contain paralogs whose function may not precisely correspond to that of the known protein.

2. Phylogenetic patterns. These show the presence or absence of proteins from a given organism in a specific COG. Used systematically, such patterns can identify whether a particular metabolic pathway exists in an organism.

3. Multiple alignments. Each COG page includes a link to a multiple alignment of COG members, which can be used to identify conserved sequence residues and analyze evolutionary relationships between member proteins.

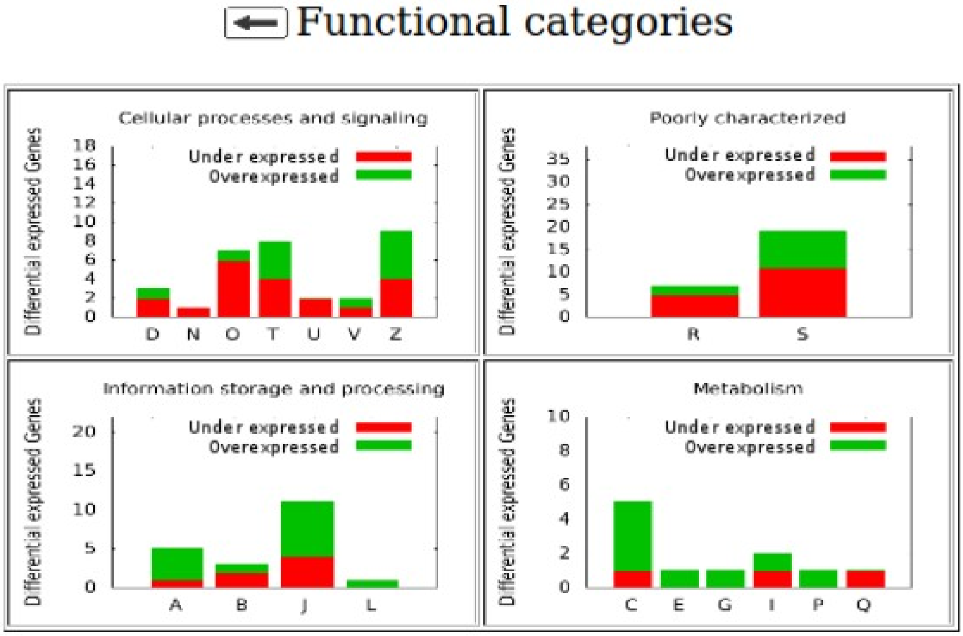


Figure 5.16.2: A graphic representation of differentially expressed genes grouped into COG categories. The differentially expressed genes grouped into COG or classified using gene ontology could reveal potential changes in particular metabolic pathways. The related table (not shown) list the hyper-test links containing the main information. Clicking on the hyper-test link will appear a list of differentially expressed genes belonging the selected category.

Differentially expressed genes

The differentially expressed genes are ordered basing on 2 factors: i) the number of replicas that confirms the same result; ii) the *p-value*.

These 2 factors are combined to calculate a gene penalty based on the statistical significance of each gene.

Fields description

*(1) Order number***:** This number is associated to the related gene basing on field 2 “gene penalty”

(2) Gene penalty: It is calculated considering the number of replicas that confirm the same transcript and the better *p-value among the analysed replicas*.

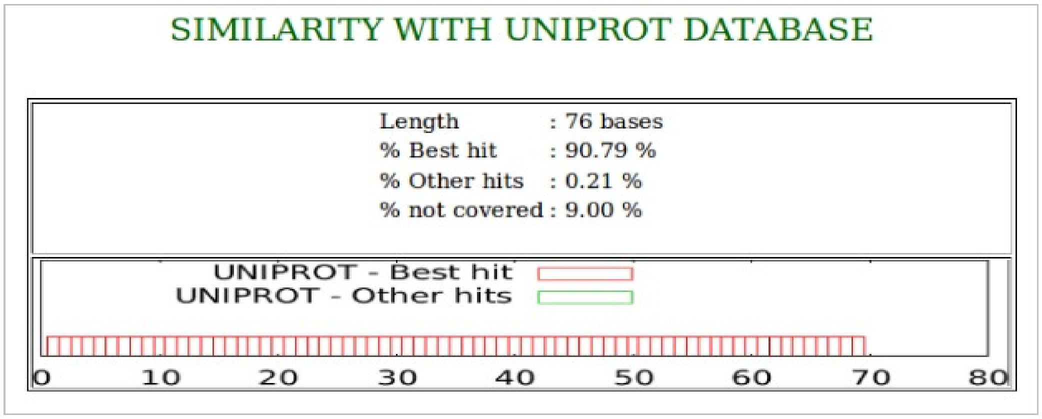
(3) P-value: p-value coming from statistical analysis.

(4) expression level: this value indicates the percentage of genes that have the expression level lower than considered transcript.

(5) Replicas confirming under expression: how many replicas have confirmed the differentially under-expressed transcript.

(6) Replicas confirming over expression: how many replicas have confirmed the differentially over-expressed gene.

(7) Reference name: Clicking on this link, a page with several charts appears as showed in the figures 5.16.4 and 5.16.5. The graphs show the information about the sequence coverage and *nr* proteins similarity.



SIMILARITY WITH UNIPROT DATABASE

Length : 76 bases
% Best hit : 90.79 %
% Other hits : 0.21 %
% not covered : 9.00 %

Figure 5.16.4: Contig similarity with nr protein database. The contig has a strong similarity with a non-redundant protein (< 1e-3) from nr database [Pruit at al. 2007]. This protein covers the 90.79% of the contig length; only 9% of the contig is uncovered. The region covered by the best hit protein is evidenced by a red area and could be compared with the regions covered by paired-ends (see figure 5.16.5).

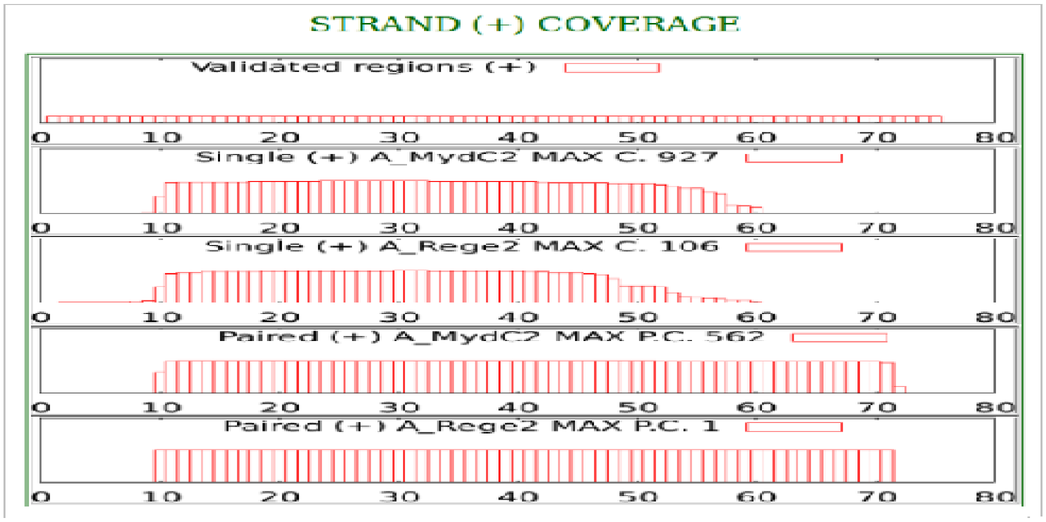


Figure 5.16.5: The first track summarize the confirmed region by paired-end (all replicas), second and third tracks show the coverage of mapped reads of each considered experiment (unpaired or single) and finally, fourth and fifth tracks show the physical coverage of mapped paired.ends of each experiment. All information are reported separately for strands + and – (- is not shown).

Situations where genes are well supported by clues:

(i) the region covered by the best hit protein and the same one confirmed by paired-end should overlap one to each other.

(ii) The mapped paired-ends should cover physically the entire contig without discontinuities.

(iii) The mapped reads of directed cDNA libraries should map only in one of the two possible strands.

Situations where genes are not well supported by clues:

(i) If a contig is not well covered by mapped paired-ends.

(ii) gaps should be found close to the junction between different aligned proteins.

(iii) The protein that match into a contig region is not coherent to the mapped paired-ends (location and/or strand).

(iv) The mapped reads of directed cDNA libraries seem to be mapped on both strands.

Ambiguous situations:

(i) If the contig length is approximately equal to the insert size of the paired-end library. This situation involves to underestimate the number of mapped paired-ends.

(ii) In the case of differentially expressed genes that are low expressed and slightly exceed the Z-score threshold.

(8) Functional category symbol: The associated symbol of functional category.

(9) NR identification code: This code identify a specific protein in the database. Normally, identifiers are simply accession, accession.version or gi's.

(10) Protein description: description of the similar protein mapped into the selected transcript or description coming from gene annotation. The associated *gene identification code* is linked to the NCBI search tool. Clicking on this link a lot of information will be reported from different biological databases.

5.17 A preliminary study of natural apoptosis in B. schlosseri

SOLiD reads were assembled into 56234 contigs. 52% have high similarity with known proteins in the protein database nr. *B. schlosseri* transcriptome assembly was annotated through gene ontology (GO) and cluster of orthologous groups (COG). The gene expression analysis (p-value of 0.05) has revealed more than 4 thousand of differentially expressed unigenes in the comparison of the blastogenetic phases (11/8/6 vs 9/8/2), about 3 thousand for the comparison (9/8/2 vs 9/8/5) and (9/8/5 vs 11/8/6). In a preliminary study 10 genes involved in the apoptosis pathways were analysed and selected basing on two criteria: i) at least 2 biological replicas must confirm the same result; ii) the biological replicas must agree for the same differential expression sign. Furthermore, to better understand the relation between genes, they were classified into 5 categories on the basis of their inferred biological function. The classification is reported below.

Category 1: early stage inducers (4 genes)

(1) similar to death-associated protein: Death-associated protein kinase (DAPk) is a family of Ser/Thr kinases, whose members not only share cell death-associated functions but possess significant homology in their catalytic domains

[Bialik at al. 2006]. It seems to be involved in the activation of cysteine-type endopeptidase (caspase) activity involved in apoptotic process. They may play a role in the early stages of bud differentiation or in apoptosis.

(2) similar to cellular apoptosis susceptibility protein - CAS: A nucleocytoplasmic transport protein that binds to alpha karyopherins and RAN GTP binding protein inside the cell nucleous. It seems to participates in their export into cytoplasm [Schroeder et al. 1999]. It is also associated with the regulation of apoptosis and microtubule assembly. CAS bind strongly to nuclear localization signal (NLS)-free importin-alpha, and this binding is released in the cytoplasm by the combined action of RANBP1 and RANGAP1. In addition, the encoded protein may play a role in both apoptosis and cell proliferation. Proteins that carry a NLS are transported into the nucleus by the importin-alpha/beta heterodimer. Importin-alpha binds the NLS, while importin-beta mediates translocation through the nuclear pore complex.

(3) similar to death inducer-obliterator 1: In mice, the death inducer-obliterator-1 gene is up regulated by apoptotic signals and encodes a cytoplasmic protein that translocates to the nucleus upon apoptotic signal activation. When overexpressed, the mouse protein induced apoptosis in cell lines growing *in vitro*. This gene is similar to the mouse gene and therefore is thought to be involved in apoptosis [Futterer at al. 2005].

(4) similar to programmed cell death 2: its over expression suppresses AP1, CREB, NFAT, and NF-kB transcriptional activation, and delays cell cycle progression at S phase [Chan at al. 2008].

Category 2: inducers (2 genes)

(5) similar to apoptosis-inducing factor (AIF)-like mitochondrion-associated inducer of death: This gene encodes a flavoprotein oxidoreductase that binds

single stranded DNA and probably it is involved in the apoptosis in the presence of bacterial and viral DNA. The expression of this gene is also found to be induced by tumour suppressor protein p53 in colon cancer cells [Joza at al. 2009] [*Candé at al. 2002*] .

(6) apoptosis-inducing factor 3-like: involved in the activation of cysteine-type endopeptidase activity by cytochrome C in the apoptotic process. It induces apoptosis through a caspase dependent pathway. It seems to have a role in the reduction of the mitochondrial membrane potential (gene function given is not based on experimental findings but by similarity information).

Category 3: death signal (2 genes)

(7) CAF74916 apoptosis-linked gene 2: is a Ca^{2+}-binding protein that has been implicated in T cell receptor-, Fas-, and glucocorticoid-induced cell death [Wiens at al. 2004].

(8) GAA55569 death domain-containing protein 1: DD is related in sequence and structure to the death effector domain (DED) and the caspase recruitment domain (CARD), which work in similar pathways and show similar interaction properties. DD binds to other DDs forming oligomers. Mammals have numerous and various DD-containing proteins. Some DD-containing proteins are involved in the regulation of apoptosis and inflammation through their activation of caspases and NF-kB, which typically involves interactions with TNF (tumour necrosis factor) receptors (gene function given is not based on experimental findings but by similarity information).

Category 4: inhibitor (1 gene)

(9) apoptosis inhibitor 5-like isoform 1: Antiapoptotic factor that may have a role in protein assembly. Negatively regulates ACIN1. The inhibitor 5-like isoform 1 binds ACIN1, and suppresses ACIN1 cleavage by caspase 3, consequently it

suppresses the ACIN1-mediated DNA fragmentation [Kim J.W. At al. 2000; Morris E.J at al. 2006; Rigou P. at al. 2009]. Also known to efficiently suppress E2F1-induced apoptosis.

Category 5: Defender against apoptosis signal (1 gene)

(10) similar to defender against apopototic cell death 1 isoform 2: a better description is not available (gene function given is not based on experimental findings but by similarity information).

Analysis

Figure 5.17.1 shows a schematic diagram of the differentially expression trend among the compared blastogenetic phases. Early inducers of apoptosis have positive differential expression in the comparison (11/8/6-9/8/2) while inducers in the comparison (9/8/5-11/8/6) as the DD genes. The gene involved in the inhibition of apoptosis seems to have positive differential expression in the comparison (11/8/6-9/8/2) while negative in (9/8/2-9/8/5). The defender against apoptosis gene is positive differentially expressed in the comparison (9/8/2-9/8/5).

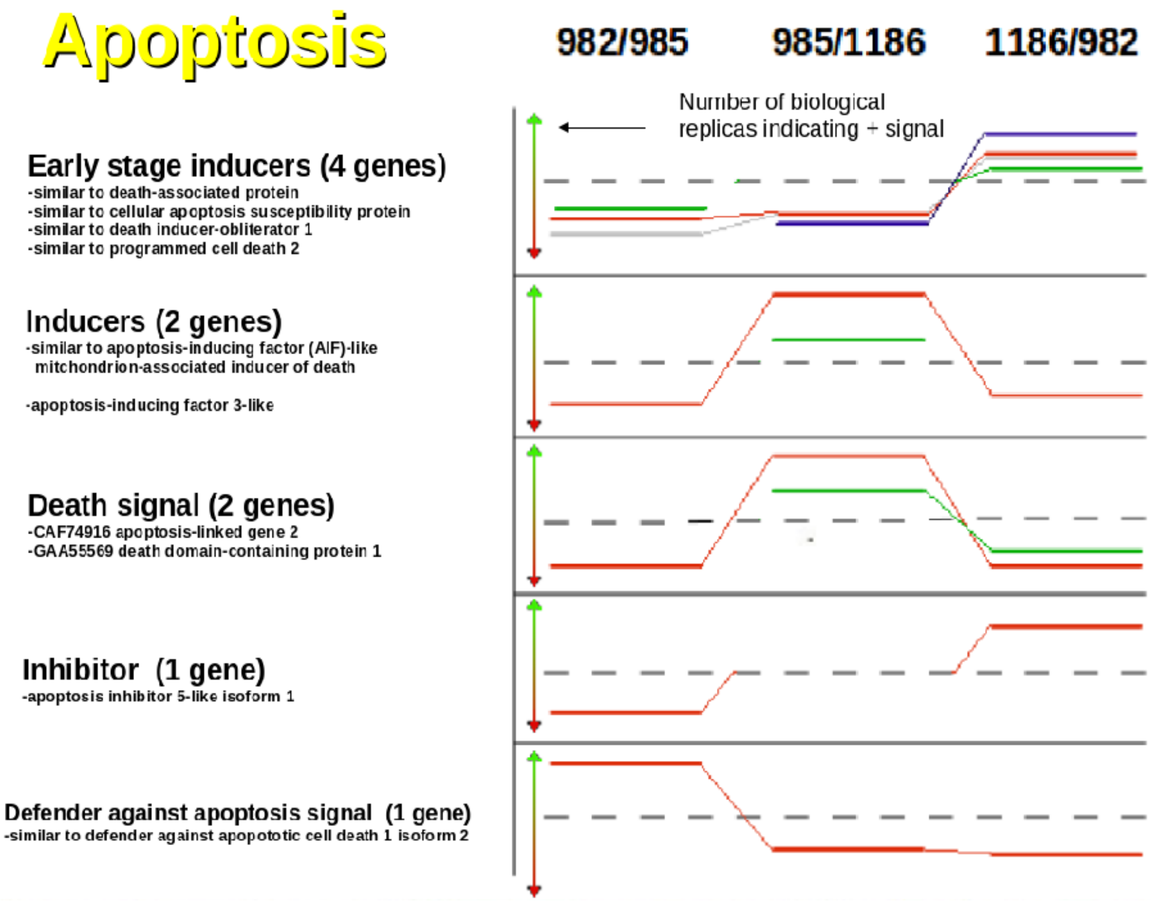


Figure 5.17.1: The compared RNA-seq experiments regard the following couple of experimental conditions: 9/8/2 vs 9/8/5, 9/8/5 vs 11/8/6, 11/8/6 vs 9/8/5 which are reported on X axis. The Y axis is proportional to the biological replicas that confirm the same result. Above the dashed line the signal is positive and indicates differentially over-expression (for instance: 9/8/5 vs 11/8/6 inducers). Below the dashed line, the signal is negative and indicates differentially under-expression. Different color of the lines are referred to different genes. Some signals overlap to each other and some lines are not distinguishable. In some case the signal lack and no lines are plotted. This fact indicates low statistical significance or lacking of differential expression.

5.18 Vascular system regeneration

The gene expression analysis of this experiment produced no significant results.

6. CONCLUSION

1. A new method to assembly SOLiD RNA-seq data from a non model organisms is described in details. The developed method consists of several analyses that could be grouped in two categories:

- ✓ *Color-space reads managing*: actually SOLiD reads can be assembled only using the Velvet program (Zerbino at al. 2008). The input of Velvet is not given by color space sequences in the original format but they have to be modified. A developed program has been specifically design to convert color-space assembly into base-space using a novel method called MCSC (Multiple Color Space Conversion). This method is based on statistical analysis of reads coverage and 2-base encoding extended to multiple sites. A simulation of MCSC was performed in order to understand the ability of the program to recognize assembly errors. Sensitivity and specificity appear high also in the worst conditions. The quality of color-converted assembly seems to be very similar to the assemblies coming from base-space reads.

- ✓ De novo *assembly optimization*: low sequence coverage and transcript isoforms are responsible of many assembly errors. Whole de novo assembly method has been validated using a simulation to simulate the expression level of transcripts and their isoforms. The result of the simulation has allowed to set a new method called MATRA. MATRA allows to select redundant sequence of assembled contigs coming from different K-mer assembly. This process removes unspecific assembly errors as non redundant sequences resulting by multiple K-mer assembly. If sequence quality is low, MATRA can strongly reduce the assembly size, so caution must be taken to set the redundancy threshold. When applied to simulated data at its maximum stringency, this method has reduced the assembly errors by 93%. Contigs elongation, PE-scaffolding and STM-scaffolding are specifically designed to optimize the overall method. RNA-seq experiments have been assembled to produce a *de novo* transcriptome assembly of *B. schlosseri* based on MATRA method.

2. A preliminary analysis of the RNA-seq experiments has revealed interesting and hopeful results. Natural apoptosis should act on cells of dying adult zooids at phase 11/8/6. In order to confirm this hypothesis, ten genes involved in natural apoptosis were grouped into five categories. Such categories were related to the biological function of each gene. Genes were selected if they had multiple confirmations in different biological replicas without conflicting results. The gene expression analysis regards the following experimental conditions: (9/8/2 vs 9/8/5), (9/8/5 vs 11/8/6) and (11/8/6 vs 9/8/2), ordered by a temporal succession. The study of the differentially expressed signals and their sign have allowed us to deduce the gene expression level of each considered experiment. This analysis led to the following experimental evidence: early inducers of apoptosis are highly expressed in phase 9/8/2. On the contrary, late inducers genes seem to be high expressed during phase 11/8/6 as well as other genes which have the death signal domain. The inhibitor of apoptosis was highly expressed during phase 9/8/2 while it seems low expressed during phases 9/8/5 and 11/8/6.

Primary buds become new adult zooids during take-over and they are constituted of many proliferating cells. The gene that seems to have a defence role against the apoptosis was highly expressed during phase 9/8/5 but its expression level decreased progressively during phases 11/8/6 and 9/8/2. Probably this gene is expressed in such proliferating cells and could be very interesting to confirm such speculation with new experiments.

3. The gene expression study should be supported by tools which allow to compare and localize interesting information. In order to accomplish this important aim we have developed a Web-based interface that interacts with a database of preprocessed data. The Web interface is thought in a compact structure that could be directly and quickly accessed from the network through a common Web browser. The analyses could be replicated using different statistical significance and/or including biological replicas or comparing results of different experiments. The differentially expressed genes are grouped basing on the Gene ontology and COG information.

References

Anderson R. *Cellular responses to foreign bodies in the tunicate Molgula manhattensis (DeKay). Biol. Bull. 141:91–98. 1971*

Altschul S, Gish W, Miller W, Myers E, Lipman D. *"Basic local alignment search tool." J. Mol. Biol. 215:403-410 ,1990.*

Ballarin L, Cima F, Sabbadin A. *Histoenzymatic staining and characterization of the colonial ascidian Botryllus schlosseri hemocytes. Boll. Zool. 60: 19-24. 1993***.**

Ballarin L, Cima F, Sabbadin A. *Phagocytosis in the colonial ascidian* Botryllus schlosseri*. Dev. Comp. Immunol. 18:467–481, 1994*

Ballarin L, Cima F, Sabbadin A. *Morula cells and histocompatibility in the colonial ascidian* Botryllus schlosseri*. Zool. Sci. 12:757–764, 1995.*

Ballarin L, Cima F, Sabbadin A. *Phenoloxidase and cytotoxicity in the compound ascidian Botryllus schlosseri .Dev. Comp. Immunol. 22: 479-492, 1998.*

Ballarin L, **Tonello C, Sabbadin A.** *Humoral opsonin from the colonial ascidian Botryllus schlosseri as a member of the galectin family. Mar. Biol.* ***136****: 813-822, 2000.*

Ballarin L, Cima F, Floreani M, Sabbadin A. *Oxidative stress induces cytotoxicity during rejection reaction in the compound ascidian* Botryllus schlosseri*. Comp. Biochem Physiol. 133C: 411-418, 2002*

Ballarin L, **Scanferla M, Cima F, Sabbadin A.** *Phagocyte spreading and phagocytosis in the compound ascidian Botryllus schlosseri: evidence for an* *integrin-like, RGD-dependent recognition mechanism. Dev. Comp. Immunol. 26: 39-48, 2002.*

Ballarin L, Cima F. *Cytochemical properties of Botryllus schlosseri haemocytes: indications for morpho-functional characterisation. Eur. J. Histochem. 49: 255-264, 2005.*

Ballarin L, **Cima F.** *Apoptosis and recognition of apoptotic cells in colonial ascidians. Caryologia 59: 354-357, 2006.*

Ballarin L, Schiavon F, Manni L. *Natural apoptosis during the blastogenetic cycle of the colonial ascidian* Botryllus schlosseri*: a morphological analysis. Zool Sci 27: 96-102, 2010.*

Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJM. De novo *transcriptome assembly with AbySS. Bioinformatics 25: 2872-2877, 2009.*

Bialik S, Kimchi A. *The death-associated protein kinases: structure, function, and beyond. Annu Rev Biochem. 75:189-210, 2006*

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. *Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27: 578-579, 2011*

Burighel P, Cloney A. *Urochordata: Ascidiacea. In F.W. Harridon (Ed.): Microscopic Anatomy of Invertebrates. Wiley-Liss. Vol 15, pp. 221-247, 1997.*

Burighel P, **Schiavinato A.** *Degenerative regression of the digestive tract in the colonial ascidian Botryllus schlosseri (Pallas). Cell Tissue Res. 235: 309-318, 1984*

Cammarata M, Arizza V, Parrinello N, Candore G, Caruso C. *Phenoloxidase* *dependent cytotoxic mechanism in ascidian (*Styela plicata*) hemocytes against erythrocyte and K562 tumor cells. Eur. J. Cell. Biol. 74: 302–307, 1997*

***Campagna D, Albiero A, Bilardi A, Caniato E, Forcato C, Manavski S, Vitulo N, Valle G**. PASS: a program to align short sequences. Bioinformatics 25: 967-968, 2009*

***Cima F, Perin A, Burighel P, Ballarin L**. Morpho- functional characterisation of haemocytes of the compound ascidian Botrylloides leachi (Tunicata, Ascidiacea). Acta Zool. 82: 261-274, 2001.*

***Cima F, Basso G, Ballarin L**. Apoptosis and phosphatidylserine-mediated recognition during the take-over phase of the colonial life-cycle in the ascidian Botryllus schlosseri. Cell Tissue Res. 312: 369-376, 2003.*

***Cima F, Sabbadin A, Ballarin L.** Cellular aspects of allorecognition in the compound ascidian Botryllus schlosseri. Dev. Comp. Immunol. 28: 881-889, 2004.*

***Chan PP, Lowe TM**. GtRNAdb: A database of transfer RNA genes detected in genomic sequence. Nucl. Acids Res. 37: D93-D97, 2009.*

***Chen Q, Yan C, Yan Q, Feng L, Chen J, Qian K**. The novel MGC13096 protein is correlated with proliferation. Cell Biochem. Funct. 26:141-145, 2008.*

***Cochrane GR, Galperin MY.** “The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources” (in eng). Nucleic Acids Res. 38: D1-4, 2010.*

***Cole JR, Chai B, Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, McGarrell DM, Bandela AM, Cardenas E, Garrity GM, Tiedje JM.** The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res. 35: D169-D172, 2007.*

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. *Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676, 2005.*

Endo T, Ueno K, Yonezawa K, Mineta K, Hotta K, Satou Y, Yamada L, Ogasawara M, Takahashi H, Nakajima A, Nakachi M, Nomura M, Yaguchi J, Sasakura Y, Yamasaki C, Sera M, Yoshizawa AC, Imanishi T, Taniguchi H, Inaba K. *CIPRO 2.5:* Ciona intestinalis *protein database, a unique integrated repository of large-scale omics data, bioinformatic analyses and curated annotation, with user rating and reviewing functionality. Nucleic Acids Res. 39: D807-814, 2011.*

Falgueras J, ***Lara A***, ***Fernández-Pozo N***, ***Cantón F***, ***Pérez-Trabado G, Claros M***. *SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. BMC Bioinformatics 11:38, 2010.*

Futterer A, Campanero MR, Leonardo E, Criado LM, Flores JM, Hernandez JM, San Miguel JF, Martinez-A C. *Dido gene expression alterations are implicated in the induction of hematological myeloid neoplasms. J. Clin. Invest. 115:2351-2362, 2005.*

Goodbody I. *The physiology of ascidians. Adv. Mar. Biol. 12: 1-149, 1974.*

Hirose E, 2003. *Colonial allorecognition, hemolytic rejection, and viviparity in botryllid ascidians. Zool. Sci. 20: 387–394, 2003.*

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D. *InterPro in 2011: New developments in the family and domain prediction database. Nucleic Acids Research 40: D306–D312. 2011.*

Huang X, Madan A. *CAP3: A DNA Sequence Assembly Program. Genome Res., 9: 868–877, 1999.*

Joza N, Pospisilik JA, Hangen E, Hanada T, Modjtahedi N, Penninger JM, Kroemer G. *AIF: not just an apoptosis-inducing factor. Ann. N. Y. Acad. Sci. 1171: 2–11, 2009.*

Candé C, Cohen I, Daugas E, Ravagnan L, Larochette N, Zamzami N, Kroemer G. *Apoptosis-inducing factor (AIF): a novel caspase-independent death effector released from mitochondria. Biochimie 84 (2–3): 215–22, 2002.*

Kim JW, Cho HS, Kim JH, Hur SY, Kim TE, Lee JM, Kim IK, Namkoong SE. *AAC-11 overexpression induces invasion and protects cervical cancer cells from apoptosis. Lab. Invest. 80: 587-594, 2000.*

Laird DJ, Chang WT, Weissman IL, Lauzon RJ*, Identification of a novel gene involved in asexual organogenesis in the budding ascidian Botryllus schlosseri. Dev Dyn 234: 997-1005, 2005.*

Lauzon RJ, Ishizuka KJ, Weissman IL. *Cyclical, developmentally-regulated death phenomenon in a colonial urochordate. Dev. Dyn. 194: 71-83, 1992.*

Lauzon RJ, Patton CW, Weissman IL. *A morphological and immunohistochemical study of programmed cell death in Botryllus schlosseri (Tunicata, Ascidiacea). Cell Tissue Res. 272: 115-127, 1993.*

Lauzon RJ, Shizuka KJ, Weissman IL*. Cyclical generation and degeneration of organs in a colonial urochordate involves crosstalk between old and new: a model for development and regeneration. Dev. Biol. 249: 333-348, 2002.*

Li W, Godzik A. *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658-1659, 2006.*

Manni L, Zaniolo G, Cima F, Burighel P, Ballarin L. Botryllus schlosseri*: a model ascidian for the study of asexual reproduction. Dev. Dyn. 236: 335-352, 2007.*

Mckernan K, Blanchard A, Kotler L, Costa G. *Reagents,Methods and Libraries for Bead-Based Sequencing. United States Patent 8329404, 2012.*

Milanesi C, Burighel P. *Blood cell ultrastructure of the ascidian Botryllus schlosseri. Hemoblast, granulocytes, macrophage, morula cell and nephrocyte. Acta Zool. 59: 135-147, 1978.*

Millar RH, Goodbody I. *New species of ascidians from the West Indies. Studies on the fauna of Curacao* and other Caribbean islands *45: 142-161, 1974*

Morris EJ, Michaud WA, Ji JY, Moon NS, Rocco JW, Dyson NJ. *Functional identification of Api5 as a suppressor of E2F-dependent apoptosis* in vivo*. PLoS Genet. 2: E196-E196, 2006.*

Parrinello N. *Cytotoxic activity of tunicate hemocytes. In "Invertebrate Immunology", B. Rinkevich and W. E. G. Müller eds, Springer-Verlag, Berlin. pp. 190–217, 1996.*

Pruitt K, Tatusova T, Maglott D. *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35: D61–D65, 2007.*

Rigou P, Piddubnyak V, Faye A, Rain JC, Michel L, Calvo F, Poyet JL. *The antiapoptotic protein AAC-11 interacts with and regulates Acinus-mediated DNA fragmentation. EMBO J. 28: 1576-1588, 2009.*

Rinkevich B, Tartakover S, Gershon H. *Contribution of morula cells to allogeneic responses in the colonial urochordate Botryllus schlosseri. Mar. Biol. 131: 227-236, 1998.*

Rowley AF, Rhodes CP, Ratcliffe NA. *Protochordate leucocytes: a review. Zool. J. Linn. Soc. (London) 80: 283-295, 1984.*

Sabbadin A. *Osservazioni sullo sviluppo, l'accrescimento e la riproduzione di Botryllus schlosseri (Pallas), in condizioni di laboratorio. Boll Zool 22:243 – 263, 1955.*

Sabbadin A. *Il ciclo biologico di Botryllus schlosseri (Pallas) [Ascidiacea] nella laguna di Venezia. Arch. Oceanogr. Limnol. 10: 219-231, 1955.*

Sabbadin A*. Effetti dell'estirpazione delle gemme sulla durata del ciclo vitale in Botryllus schlosseri (Pallas). Boll. Zool. 23: 331-342, 1956.*

Sabbadin A. *Ulteriori osservazioni sull'allevamento e sulla biologia dei Botrilli in condizioni di laboratorio. Arch Oceanogr Limnol 12:97-107, 1960.*

Sabbadin A, Tontodonati A. *Nitrogenous excretion in the compound ascidians* Botryllus schlosseri *(Pallas) and* Botrylloides leachi *(Savigny). Monit. Zool. Ital. 1: 185-190, 1967.*

Sabbadin A, Zaniolo G, Majone F. *Determination of polarity and bilateral asimmetry in palleal and vascular buds of the ascidian* Botryllus schlosseri*. Dev. Biol. 46: 79-87, 1975.*

Sabbadin A, Zaniolo G, Ballarin L. *Genetic and cytological aspects of histocompatibility in ascidians. Boll. Zool. 59: 167-173, 1992.*

Schroeder AJ, Chen XH, Xiao Z, Fitzgerald-Hayes M. *Genetic evidence for interactions between yeast importin alpha (Srp1p) and its nuclear export receptor, Cse1p. Mol Gen Genet. 261(4-5):788-95, 1999*

Scofield VL, Nagashima LS. *Morphology and genetics* of *rejection reactions between oozooids from the tunicate* Botryllus schlosseri*.Biol. Bull. 165:733-744. 1983.*

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. *AbySS: a* *parallel assembler for short read sequence data. Genome Res. 19(6): 1117-23, 2009.*

Surget-Groba Y, Montoya-Burgos J. *Optimization of* de novo *transcriptome assembly from next-generation sequencing data . Genome Res. 20: 1432-1440, 2010.*

Smith D, Quinlan A, Peckham H. *Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res. 18(10): 1638–1642, 2008.*

Taneda V, Watanabe H. *Studies on colony specificity in the compound ascidian, Botryllus primigenus Oka. I. Initiation of "nonfusion" reaction with special reference to blood cells infiltra tion. Dev.Comp.Immunol. 6: 43-52, 1982.*

Tatusov RL, Galperin MY, Natale DA, Koonin EV. *The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28:33–36, 2000.*

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. *The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29:22–28, 2001.*

Tiozzo S, Christiaen L, Deyts C, Manni L, Joly JS, Burighel P. *Embryonic vs. blastogenetic development in the compound ascidian* Botryllus schlosseri*: insights from Pitx expression patterns. Dev. Dyn. 232: 469-479, 2005.*

Wright RK. *Urochordates. In: "Invertebrate Blood Cells", Ratcliffe NA and* Rowley AF*, eds, , Vol. 2, Academic Press, New York, London, pp 565-626, 1981.*

Wright RK. *Urochordates. In: Ratcliffe NA and Rowley AF (Eds), Invertebrate Blood Cells, Academic Press - New York - London, 2: 565- 626, 1981.*

DR Zerbino, E Birney*. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18: 821–829, 2008.*

Zaniolo G, Trentin P*. Regeneration of the tunic in the colonial ascidian,* Botryllus schlosseri*. Acta Embryol. Morphol. Exp. 8: 173-180, 1987.*