UNIVERSITÀ DEGLI STUDI DI PADOVA

UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Biologia

Scuola di Dottorato di Ricerca in Bioscienze e Biotecnologie

Indirizzo Biologia Evoluzionistica

Ciclo XXV

Transcriptomic analysis of the polyploid Adriatic sturgeon (*Acipenser naccarii*)

Direttore della Scuola: Ch.mo Prof. Giuseppe Zanotti

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

Supervisore: dott. Leonardo Congiu

Dottorando : Michele Vidotto

Padova, 31 January 2013

Abstract

Sturgeons are a group of Condrostean fishes with very high evolutionary, economical and conservation interest. The eggs of these living fossils represent a luxury delicacy and are one of the most valuable foods of animal origin. The intense exploitation of wild stocks for the harvesting of caviar caused in the last decades a dramatic decline of their distribution and abundance leading, in 2010, the International Union for Conservation of Nature to list them as the more endangered group of species. As a direct consequence, world-wide efforts have been made to develop sturgeon aquaculture programmes for caviar production. In this context, selective farming of females would increase the economical profits and the characterisation of genes involved in sex determination becomes a major issue. The 454 sequencing of four normalised cDNA libraries from gonads and brain of *A naccarii* and *A. stellatus*, one male and one female full-sib per species, yielded 182,066 and 167,776 reads for *A. naccarii* which after a strict quality control were iteratively assembled together, giving more then 55,000 high quality Expressed sequence tags (ESTs). A total of 184,374 and 169,286 raw reads were instead produced for *A. stellaus* male and female libraries respectively, resulting in 63,606 ESTs after two round assembly. It was estimated the joint assembly of *A. naccarii* was able to cover about 80% of its total transcripts expressed in both gonad and brain with a mean contigs coverage of 4X. Similarly 86% transcriptome coverage was achieved by assembling both sex specific libraries of *A. stellatus,* with 3.6X as mean contig coverage. The Multi-step annotation process finally results in 16% and 15% successfully annotated sequences, with GO terms, respectively in *A. naccarii* and *A. stellatus*. Both transcriptomes were screened for 32 sex related genes and highlighted 5 and 2 genes that are potentially specifically expressed in *A. naccarii* male and female, at the first life stage at which sex is histologically identifiable. The screening in *A. stellatus* is currently at preliminary stage and further filtering steps are required. Both sturgeon transcriptomes were also compared with those of other fish species for which relevant genomic informations were available. Finally, 21,791 putative EST-linked Single Nucleotide Polymorphisms (SNPs) and 5,295 Single Sequence Repeats (SSRs) were identified in *A. naccarii*, while 15,449 and 5,696 were putatively classified in *A. stellatus* assembly. This study represents the first characterisation of transcriptomes from two high endangered sturgeon species. Most of the information acquired for *A. naccarii* were well organised into the public database *AnaccariiBase*, freely available at

http://compgen.bio.unipd.it/anaccariibase/, while the information obtained for *A stellatus* will be released soon. This study represents a precious source of information for more focussed studies aimed at characterising or comparing genes, deciphering molecular mechanisms or genetic pathways in this group of species, or discovering hundreds of EST-linked markers with several possible applications in sturgeon conservation.

Sommario

Gli storioni sono un gruppo di pesci Condrostei, di elevato interesse evoluzionistico, economico e di conservazione. Le uova di questi fossili viventi costituiscono uno degli alimenti di origine animale più preziosi sul mercato. L'intenso sfruttamento delle popolazioni selvatiche per la raccolta del caviale ha causato, negli ultimi decenni, un calo drammatico della loro distribuzione ed abbondanza che ha portato, nel 2010, l'Unione Internazionale per la Conservazione della Natura ad indicarli come il gruppo di specie a maggior rischio di estinzione. Come diretta conseguenza, sono stati compiuti sforzi notevoli, in tutto il mondo, per sviluppare programmi finalizzati alla produzione di caviale via acquacoltura. In questo contesto, l'allevamento selettivo delle femmine aumenterebbe i profitti economici e la caratterizzazione dei geni coinvolti nella determinazione del sesso, diventa decisiva. Il sequenziamento 454 di quattro librerie di cDNA normalizzate, costruite a partire da gonadi e cervello di *A naccarii* e *A. stellatus*, un maschio ed una femmina (fratelli) per specie, ha prodotto 182,066 e 167,776 reads grezze rispettivamente per i due sessi di *A naccarii,* che dopo un severo controllo di qualità sono state assemblate insieme attraverso un processo iterativo, risultando in più di 55,000 Expressed Sequence Tags (ESTs) di qualità elevata. Per la specie *A. stellatus*, invece, sono state prodotte 184,374 reads per la libreria maschile e 169,286 per quella femminile, allineate in 63,606 ESTs dopo due giri di assemblaggio. E' stato stimato che, l'assemblaggio di *A. naccarii* contenga i tag di circa l'80% dei trascritti totali espressi in gonadi e cervello, in questa specie, con una copertura media dei contigs pari a 4X. La copertura del trascrittoma di *A. stellatus* è stata stimata in circa l'86%, con 3.6X come media di copertura dei contigs. Il processo di annotazione multi-fase, ha portato ad annotare correttamente, con termini GO circa 16% ed il 15% delle sequenze, rispettivamente in *A. naccarii* ed *A. stellatus*. Entrambi i trascrittomi sono stati interrogati alla ricerca di 32 geni legati al sesso e sono stati evidenziati 5 geni potenzialmente espressi in modo specifico nel maschio e 2 nella femmina di *A. naccarii*, nel primo stadio di sviluppo, in cui il sesso è istologicamente identificabile. La ricerca nel trascrittoma di *A. stellatus* è attualmente preliminare e sono necessarie ulteriori fasi di filtraggio. Entrambi i trascrittomi sono stati confrontati con quelli di altre specie di pesci, per la quali erano disponibili rilevanti informazioni genomiche. Infine, 21.791 putativi Single Nucleotide Polymorphisms (SNPs) e 5.295 Single Sequence Repeats (SSR) EST-linked sono stati identificati in *A. naccarii*, mentre 15.449 e 5.696 sono stati

rispettivamente classificati nell'assemblaggio di *A. stellatus*. Questo studio rappresenta la prima caratterizzazione dei trascrittomi di due specie di storione ad elevato rischio di estinzione. Gran parte delle informazioni acquisite per la specie *A. naccarii* sono state organizzati all'interno della banca dati pubblica *AnaccariiBase*, liberamente disponibile all'indirizzo http://compgen.bio.unipd.it/anaccariibase/, mentre le informazioni ottenute per *A stellatus* saranno rilasciate al più presto. Questa analisi rappresenta una preziosa fonte di informazioni per ulteriori studi più mirati, volti a caratterizzare o confrontare geni, a decifrare i meccanismi molecolari o le vie genetiche in questo gruppo di specie, o a scoprire centinaia di marcatori associati ad ESTs per diverse applicazioni nella conservazione dello storione.

Acknowledgements

We are most thankful to Sergio Giovannini for providing the samples used in this study; to Prof. Lorenzo Zane and Dr. Alessandro Grapputo from Department of Biology, University of Padova for insightful discussion. A special thanks to Dr. Alessandro Coppe, Dr. Paolo Martini and Dr. Gabriele Sales from the same Department for bioinformatics advice. A special thanks to all members of the Molecular Ecology Group of University of Padova. I also thanks CARIPARO for providing the Ph.D. fellowship and University of Padova - Progetto di Ateneo (grant CPDA087543/08) for providing funds.

"Imagination is more important than knowledge. For knowledge is limited to all we know and understand, while imagination embraces the entire world and all there ever will be to know and understand."

Albert Einstein

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Biological interest

Sturgeons (order: Acipenseriformes, infraclass: Chondrostei) are a very ancient fish group distributed in the Palearctic hemisphere with about 25 species, most of which are considered to be on the brink of extinction (Leonardo Congiu et al. 2011a) Sturgeons are also interesting from a biological standpoint, presenting characteristics that allow us to consider them archaic forms of survivor groups that, in past ages, had a considerable expansion. Often referred to as living fossils, sturgeons very ancient separation from teleosts occurred over 250 Mya (William Bemis 2001), placing them in a key phylogenetic position for evolutionary studies on vertebrates. Since very scarce genomic information is available about sturgeon species and more in general about members of the infraclass Chondrostean, the characterization of a sturgeon transcriptome would represents a significant contribution to studies of comparative evolution. Besides their evolutionary interest, due to the very ancient separation from teleosts (over 200 Mya), these animals present features that make very interesting a first characterization of their transcriptome. The different aspects will be briefly introduced separately.

Conservation interest

Unfertilized sturgeon eggs, sold as caviar, are one of the most valuable products in international food trade (A. Ludwig 2008) and their very high monetary value is the main cause of the extremely-endangered status of most sturgeon species. The commercial value of the different caviars depends not only on taste but mainly on the rarity of the species it belongs to (F Fontana et al. 2001). This is the main reason why, in historical times, the price per Kg of caviar rapidly increased (Error: Reference source not found) (Wuertz et al. 2007; A. Ludwig 2008). Besides caviar, also meat, gelatine, fertilized eggs, fry and adult fish are very appreciated and internationally commercialized. Since 1996 sturgeons have been included in the red list of the International Union for the Conservation of Nature (IUCN) and in 2010 they have been classified as the most endangered group of species worldwide (IUCN; March 2010). Since 1998, international trade in all species of sturgeons has been regulated under CITES (Convention for International Trading of Endangered Species) with the aim of regulate international trade and encourage the employment of

sustainable administration policies.

Caviar	Price per Kg ($)
farmed Siberian (*A. baerii*)	1,000-1,100
farmed Osietra (*A. gueldenstaedtii*)	1,350-1,500
Sevruga (*A. stellatus*)	2,200-2,200
wild Osietra (*A. gueldenstaedtii*)	2,200-2,300
wild Beluga (*H. huso*)	2,600-2,700

Table 1.1. Average prices in U.S. dollars per Kg of caviar from different sturgeon species in the countries where the importation is permitted. Figures for the year 2006.

With regards to the Adriatic sturgeon (*Acipenser naccarii*), up to the early 20th century this was one of the three species occurring in the Italian ichthyofauna together with the European Atlantic sturgeon (*Acipenser sturio*) and the beluga (*Huso huso*). At present the Adriatic sturgeon is the only species still present in the natural habitat. A few decades ago it was relatively common in nearly all tributaries of the north Adriatic Sea. Subsequently its abundance has dramatically decreased as indicated by the catch records. Catch rates exceeded 2000 kg/year at the beginning of the 1970s and declined to about 200 kg/year in 1990 and 1991. In 1993, only 19 specimens were caught (Bronzi et al. 2011). The drastic decline of *A. naccarii* prompted conservation efforts in Italy: in 1977 a stock of Po River specimens was transferred to a fish plant (Azienda Agricola VIP – Orzinuovi, Brescia -Italy) and since 1988 these wild individuals are artificially reproduced allowing the launch of restocking measures (Giovannini et al. 1991; Bronzi et al. 1994). Unfortunately the first genetic support to this *ex-situ* conservation activity started only in the last years (Leonardo Congiu et al. 2011a) and all the genetic investigation performed up to date are based on neutral genetic markers. No EST-Linked marker to be applied to monitor the effects of selective pressures is available for the Adriatic sturgeon neither for other sturgeon species.

For the very low numbers of wild breeders, future restoration efforts must rely on *ex situ* conservation strategies through the set up of long-term breeding programs in which the few wild breeders available are reproduced in captivity and the progenies subsequent reared (Sturgeons 2006).

The availability of a high number of genetic markers to guarantee adequate genetic support to releasing activities, through parental allocation and traceability of the hatchery of origin, become important in this context. Moreover, the availability of EST-linked

markers yielded by a transcriptome characterisation may provide a suitable tool for the identification of footprints of selective pressures in the released stocks, due to natural or anthropogenic stress.

Aquaculture interest

In response to the rapid decline of natural populations, aquaculture production of caviar is rapidly increasing. This would contributed not only to compensate for the decline in production by fishing, but also to fall down market prices so that the illegal trade would become less attractive.

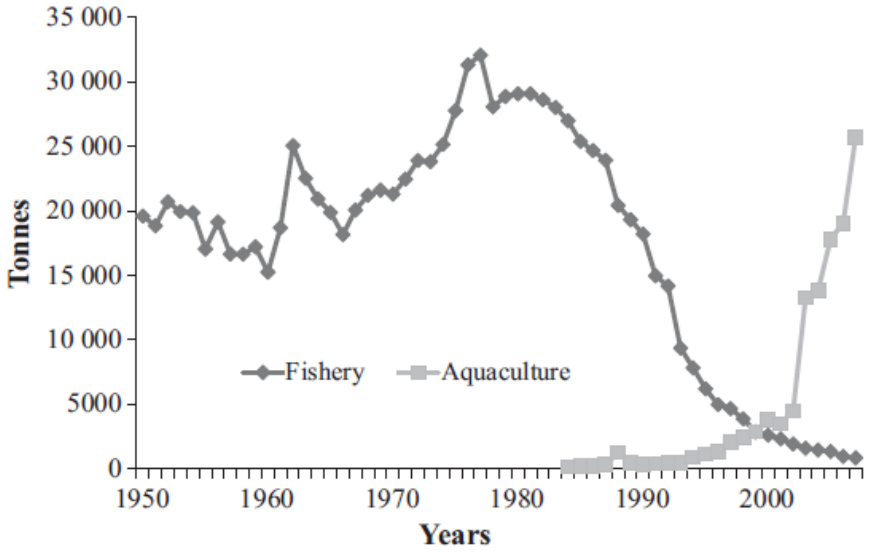

Figure 1.1. Comparison of the official statistics for half a century global sturgeons fishery yields and aquaculture production (all species combined). From (Bronzi et al. 2011).

The rapid growth production through aquaculture of sturgeon occurred at the end of the last century coincides with the massive decline of production coming from natural stocks of the Caspian Sea through years Figure 1.1. Initially the production by aquaculture was not consistent. This industry undergone a rapid expansion after a long lead-time spent in research efforts on cultivation of sturgeon species, which occurred, not only in Russia, but also in North America and Europe in the 1980s. In 2008, the estimated world production of caviar for all farmed species was in the order of 110-120 tonnes, from about 80 companies spread across 16 countries. However, this scenario is changing rapidly due to the rapid advent of market globalization. It is promised that potential new markets will develop in a very diversified manner thus changing the traditional setting (Bronzi et al. 2011).

One of the main problems for aquaculture caviar producers is that 50% of the animals are profitless males which need to be discarded from production as quickly as possible to minimise expenditure and maximise space. Unfortunately sturgeons have no reliable sexual dimorphism and use of external characteristics to determine the gender of juveniles is totally ineffective (Carmona 2009). Moreover "external" methods are unable to asses the development stage of the gametogenesis. Recent approaches using extensive genome screening in different species together with proteomic approaches with the aim of identify sex-specific markers has not, as yet, yielded satisfactory results (Saeed Keyvanshokooh et al. 2009; Wuertz et al. 2006; S. Keyvanshokooh et al. 2007).

So, even today, sex discrimination in sturgeon farming for caviar production can only be performed by ultrasound analysis after 4 or 5 years. Female sturgeon can be identified by having discernible ovarian folds or by the presence of readily discernible oocytes The rearing of males can, thus, represent up to 30% of total farming costs (Wuertz et al. 2006). A genetic identification of the sexes at an early life stage based on PCR techniques could, therefore, contribute to lowering the costs of caviar production in aquaculture and have knock-on effects in both farming and conservation. Aquaculture activity would significantly benefit from this possibility and poaching on natural populations would consequently be reduced.

Although most of the investigation failed to identify sex specific marker to date, the fact that the sex ratio of all documented progenies was not significantly different from 1:1 is a good indications that sex is genetically determined in sturgeon (A. L. Van Eenennaam et al. 1999a; Wuertz et al. 2006). Knowledge of which genes are involved in sex differentiation in sturgeons is limited and analyses at the transcriptome level of the expressed genes at the first stage at which sex can be histologically determined beside expanding the knowledge base, could contribute to the future development of caviar production in aquaculture by reducing the costs.

Functional role of poliploidization

The hight polyploid state of most species is another interesting aspect of sturgeon biology. Different species are characterized by different degrees of ploidy which are the result of multiple and independent duplication events starting from a 60-chromosome common ancestor (A Ludwig et al. 2001; F. Fontana, L. Congiu, et al. 2008; E. Boscari et al. 2011).

Sturgeon species can be divided into two main groups based on their number of chromosomes: the first having approximately 120 and the second approximately 240 chromosomes. The level of ploidy to be ascribed to each chromosome number is still being debated. It is hypothesized that, after a first duplication event, the resulting tetraploid genome (4n=120) underwent a functional diploidization (2n=120). For this reason some authors consider species of this groups to be tetraploid while other attribute the diploid condition to the same group. Similarly, species presenting about 240 chromosomes, originating from a second event of chromosome doubling, are considered to be octoploid by some authors (Birstein & Vasiliev 1987), and functionally tetraploid, by others (F. Fontana, et al. 2008).

To date, all studies aimed to clarify the ploidy level in sturgeons were based on cytogenetic approach or, alternatively, on the allele counting at neutral loci such as microsatellites (A Ludwig et al. 2001). These approaches however present relevant resolution limits mainly due to the high complexity of sturgeon genome. Moreover, the genome-level investigations, do not allow to evaluate the ploidy from a functional perspective.

Recent study based on the relationship between gene expression and the chromosome or chromosome arm duplication proposed a dosage effect or dosage compensation effect after a polyploidization event. which is better monitored by analysing its the transcribed part of the genome (Peng et al. 2008) And, of course different parts of the genome may experiment different degrees of functional ploidy.

The Adriatic sturgeon

Scientific Name: *Acipenser naccarii*, Bonaparte (1836)
Systematic classification:

PHYLUM	CHORDATA
SUBPHYLUM	VERTEBRATA
CLASS	OSTEICHTHYES
SUBCLASS	ACTINOPTERYGII
SUPERORDER	CHONDROSTEI
ORDER	ACIPENSERIFORMES
FAMILY	ACIPENSERIDAE
SUBFAMILY	ACIPENSERINAE

GENUS ACIPENSER

The Adriatic sturgeon (*Acipenser naccarii)* is a tetraploid species endemic of the North Adriatic region. It is currently considered to be at risk of extinction. whilst one time widely distributed in nearly all tributaries of the North Adriatic Sea, This species is included in Appendix II of CITES and its status was updated to ‘‘Critically Endangered’’ in the 2012 by IUCN (Leonardo Congiu et al. 2011a).

A. naccarii, like other sturgeon species, is characterized by delayed a sexual maturity. Males mature at 7-11 years (when about 80 cm in size), while the females mature at 8-15 years (when their total length is at least 1 m). Although males reproduce annually or every two years, each female ovulates every 2-4 years (some females can have intervals between two successive reproductions even longer).

Adriatic sturgeon can be considered virtually extinct in the Po river and its tributaries. In 1977 a breeding program started with the transfer of immature wild specimens from the Po River (Northern Italy) to the fish plant Azienda Agricola VIP (Orzinuovi, Brescia, Italy). Despite the successful reproduction in captivity followed by restocking practices that were conducted in the period 1988-2009 in tributaries of the Po river, recaptures have been scarce and up to now there is no evidence of natural reproduction in the Po river and its tributaries.

The Stellate sturgeon

The anadromous *Acipenser stellatus,* with about 120 chromosomes, (M. Chicca et al. 2002) is considered a diploid species, distributed in Eurasian between the Caspian, Black, Azov and Aegean Seas. Restocking plays a vital role in sustaining stellate sturgeon populations, since this species was heavily exploited for sevruga caviar. Moreover dam construction and environmental pollution has led to a 60% loss of historic spawning grounds as well as range and population size decreases (Doukakis et al. 2002).

1.2 Next Generation Sequencing Technologies: 454 pyrosequencing

General overview

Several Next Generation Sequencing (NGS) technologies has been introduced since 2005. The 454 system (http://www.454.com/) based on pyrosequencing, was the first commercial NGS platform available. This system relies on the measurement of inorganic pyrophosphate released during the incorporation of nucleotides by equivalently converting

it into visible light through a cascade of enzymatic reactions (Metzker 2010).

The Illumina system, initially developed by Solexa (http://www.illumina.com), was the second commercial NGS platform available. This system captures individual molecules on a solid glass surface (similar to a microscope slide) and use *bridge PCR* reaction to amplify single DNA template, into small clusters of identical molecules. These clusters are then sequenced by detecting dye-labelled terminators added at every cycle of the reaction. Since positions for all clusters are known, the sequence of millions of templates can be reconstructed simultaneously producing reads up to 150 bp long. Illumina system today provides the highest throughput (600 Gb per run with HiSeq 2000), good accuracy (98% for 100bp reads length) and lowest reagent costs (Liu et al. 2012).

The Applied Biosystems SOLiD (http://www.appliedbiosystems.com) was the t commercial NGS platform. SOLiD exploits two-base-encoded probes together with a ligase-mediated sequencing approach to determine sequences. The 16 dinucleotide combinations are encoded by four dyes. In this way reads are encoded in *Colour space* that is a unique feature of the SOLiD system. Single molecular templates are linked to modified micro-beads and then amplified by emPCR. The beads are covalently attached on a microscope glass slide. Upon the annealing of a universal primer, a library of 1,2-probes is added. These probes are designed to interrogate the first and second positions adjacent to the hybridized primer. Following four-colour imaging, the ligated 1,2-probes are chemically cleaved. The current ligation cycle is repeated 10 times. Subsequently, the extended primer is stripped out and an *n-1* universal primer is hybridized. Four more ligation cycles are performed in total. Colour calls from the five ligation rounds are then sequentially ordered (that is, the colour space) and aligned to a reference genome to decode the DNA sequence. This method has the primary advantage of improved accuracy in colour calling and Single Nucleotide Variant calling, the latter of which requires an adjacent valid colour change. The read length of SOLiD was initially 35 bp but the last SOLiD 5500xl system improved read length to 85 bp with a declared accuracy of 99.99% (Liu et al. 2012). Before the advent of the most recent Illumina's software and reagents release, this systems held the higher throughput (120 Gb per run with SOLiD v4 reagents).

In 2010 Ion Torrent (http://www.iontorrent.com) commercialised its first Personal Genome Machine (PGM). This sequencing system is similar to the 454 but instead of light beam as sub product of pyrophosphate, hydrogen ions (H^+) released during the polymerisation of DNA, are detected as a change in pH by a CMOS semiconductor chip,

similar to that used in the electronics industry. The throughput of this system is small, (10Mb–100Mb), and it produce sequences up to 200 bp in length. Since no lasers or cameras are needed, low-cost manufacturing microchip can be used and the sample preparation time is short (less than 6 hours for 8 samples in parallel), this platform is the cheapest sequencing system available today.

More recently Pacific Biosciences (http://www.pacificbiosciences.com) introduced to the market the first third generation sequencer. This system sequences individual DNA molecules in real time. Modified DNA polymerases are attached to the surface of a microscope glass consists of millions of zero-mode waveguides (ZMWs). During the reaction the sequence of individual DNA templates can be determined because each dNTP is linked to a unique fluorescent dye that is cleave off when it is incorporated into the complementary strand by the enzyme. A camera capture the signal as a movie of real-time observations. With this system sample preparation is very fast, no PCR amplification step is needed, the turnover rate is quite fast and the average read length provided of 1300 bp is longer than any other sequencing technology (Metzker 2010).

Taken together, the different NGS platforms offer a variety of experimental approaches for characterising a transcriptome including single-end and paired-end cDNA sequencing, tag profiling (3'end sequencing especially appropriate to estimating expression level), methylation assays, small RNA sequencing, sample tagging (i.e. barcoding) to permit small subsample identification, splice variant analyses and more (Glenn 2011). Methods for library construction vary depending on the platform but a common trend among NGS technologies is that the template is immobilised or linked to a solid support or surface. Most of the platforms need a fragment library DNA template (200-1,000 bp) prepared by randomly shearing target DNA into small pieces, and require that each template is linked to forward and reverse *ad-hoc* adaptors. Moreover, template amplification is needed in order to make fluorescent event detectable (in most cases through emulsion PCR or solid-phase amplification).

Roche 454 system

As anticipated the 454 technology is based on the measurement of pyrophosphate released during nucleotide addition. The template preparation protocol requires DNA fragmentation and ligation to both ends of adaptors containing universal priming sites. Following, the DNA is denatured into single strands that are captured by *micro-beads*

under conditions that favour one DNA molecule per bead. An *emulsion* PCR (emPCR), a reaction mixture consisting of an oil–aqueous emulsion is created to seclude bead–DNA complexes into a single aqueous droplets. Subsequently a PCR amplification is performed within these micro-reactors to generate several thousand copies of the same template sequence linked to the same bead. One-two million of beads are loaded onto a *PicoTiteltPlate* (PTP) designed so that each well can accommodate only a single bead. Moreover, smaller beads, which have sulphurylase and uciferase attached to them are packed into the wells surrounding the template beads, to facilitate light production. After the incorporation of a dNTP (dATP, dGTP, dCTP, dTTP) that complement a base of the template strand, the *ATP sulfurylase* exploits released pyrophosphate (ppi) to tranform adenosine 5' phosphosulfate (APS) into ATP. The ATP drives conversion of luciferin into oxyluciferin by *Luciferase* and generates visible light. At the same time, the unmatched bases are degraded by *apyrase.* Then another dNTP is added into the reaction system and the pyrosequencing reaction is repeated. For homopolymeric repeats of up to six nucleotides, the number of dNTPs added is directly proportional to the light signal. The order and intensity of the light peaks are recorded as flowgrams by a *charge-coupled device* (CCD) camera. All template beads are sequenced in parallel by flowing the dNTPs and other reagents in specific order, across the plate. In 2008 Roche launched the 454 GS FLX Titanium system which exploit a titanium-coated PTP design, which substantially increases read length and improves data quality by reducing crosstalk between adjacent PTP wells. Crosstalk happens when light from a given well is scattered into nearby wells thus determining a coupling of light signals that reduce sequence resolution. The 454-FLX Titanium evolution currently produces reads 300-450 bp long, up to 700 bp long for the last 454-FLX Titanium XL+ version, with accuracy 99.9% after filter and output of about 0.7Gb per run that takes only 10 hours from sequencing start till completion (Liu et al. 2012).

454 errors handling

The main limits of 454 technology, is reading through repetitive/homopolymeric regions. With respect to the first 101 reading positions the most common errors found are insertions, followed by deletions, mismatches, and ambiguous base calls. This trend changed when considering longer sequences where mismatch and ambiguous base call errors mostly contribute to global error rate. Thus sequence length affects error rates (Gilles et al. 2011). The main variables that influence the length, quality and amount of 454

sequences, are both sequence-specific and technology-specific, in particular: (i) the position of each base within the sequence, (ii) the primary sequence structure, specifically, the presence of homopolymeric regions, (iii) the length of the sequence generated (iv) the region of the PTP where the bead carrying the sequence is located (v) the distance of the bead from the centre of the PTP, finally (vi) the distance of the bead from the centre of multiple PTP. This variables practically depends on the central position of the CCD camera (edge effect), the reagent washing process (CAFIE effect), the flow of chemicals reagents through the plate, the PTP handling equipment, the quality filters and the base-calling algorithms. The CAFIE effect (carry forward and incomplete extension) happens when the *apyrase* wash is incomplete so a trace amount of not incorporated nucleotide, remains in the wells. These residues cause premature nucleotide incorporations during the next cycle, for specific sequences and add noise to the signal. A reads may result short for two reasons: the polymerization process contingently ends or the quality filter process cuts the reads because of error accumulation. Accumulation of errors manifests when some DNA strands on a bead fail to incorporate the right nucleotide during the appropriate base flow. The strands that fail to incorporate goes out-of-phase with the rest of the strands because it must wait for another flow cycle to add the appropriate nucleotide. This s the reason because the beginning of the sequence is generally more accurate than the end. Taking into account this problem, the GS-FLX system implements a quality filter algorithm that trim sequences starting from the 3' end until the number of valley flows (intermediate signal intensity, i.e., a signal intensity occurring in the valley between the peaks for 1-mer and 2-mer incorporations, or 2-mer and 3-mer, etc.) is < 1.25%. Moreover the 454 sequencing kit provides reference templates as internal controls that are added during the sequencing step and that are modified only by sequencing errors, in order to monitor the occurrence of faults in the process.

454 main choice for transcriptome sequencing

It has been shown that the 454 platform provide an efficient alternative to traditional Sanger (first-generation sequencing technology) sequencing for initial genome and transcriptome characterisation also in complex or repetitive regions (Wicker et al. 2006). Before the introduction of real-time sequencing technology such as Pacific Biosciences, the 454 was the platform that permitted to obtain longer reads. Long sequences limited computational needs for the assembly process allowing to obtain longer and more reliable tags of expressed genes in transcriptomic characterisation. This fact, combined with the

high throughput (compared to first-generation instruments), the cost-effective and the technology maturity have made Roche/454 the main choice for large-scale EST-sequencing projects through the years (Cheung et al. 2006; Vera et al. 2008; Meyer et al. 2009; Pauchet et al. 2010; Franssen et al. 2011; J.-T. Wang et al. 2012).

1.3 Genomic characterisations in sturgeons

Of the about 25 existing species of sturgeon, all have been the object of several genetic investigations. Most characterised loci however are mitochondrial regions (and nuclear neutral markers such as microsatellites (F Fontana et al. 2001). Only few studies focused on expressed genes of physiological relevance, e.g. Ig heavy chain (Dengqiang Wang et al. 2010) and a previously unknown K-dependent calcium binding protein (Viegas et al. 2008). Most of the research efforts focused on the characterisation of genes known to have a key role in the sex determination cascade among vertebrates (Anne Kathrin Hett & Arne Ludwig 2005, p.9; A. K. Hett et al. 2005; Berbejillo et al. 2012), with the hope to identifying molecular markers able to distinguish sexes in these non-dimorphic species, at early development stage. A complete description of Expressed Sequence Tags (ESTs) provides an overview of the transcriptome, those genes expressed (transcribed) in a given tissue at a specific point in time, of an organism. ESTs sequencing have become an invaluable resource for many purposes including gene discovery, genome annotation, alternative splicing, SNP discovery, molecular markers for population analysis, expression analysis in animal, plant, and microbial species (Ewing & Green 2000; Crowhurst et al. 2008; Coppe et al. 2010; S. Yang et al. 2010).

Despite the high evolutionary, commercial and conservation importance of sturgeon species, only a few EST sequencing projects have been completed so far and in very few species. Lazzari and colleagues in 2008 were the first to have made available through a public database called "Sturgeon DB", 2.704 ESTs generated via Sanger sequencing of a non-normalized skin and spleen cDNA library from the American sturgeon *Acipenser transmontanus (Lazzari et al. 2008)*. In 2009, Hale and colleagues published the first transcriptomic of the lake sturgeon *Acipenser fulvescens* through 454 GS-20 platform (Hale et al. 2009). The authors prepared 5 cDNA libraries, by extracting total RNA from gonad biopsies of 5 lake sturgeons: 2 males, 2 females and one individual of unknown sex; all individuals were 3 to 12 years old. Two out of the 5 libraries were normalised, while the remaining 3 were not. Authors claim that the sequencing process yielded 20,741 high quality reads, which were assembled into 1,831 contigs. Although they reported results of

the annotation process and identification of SNPs and xenobiotic organisms, the main purpose of the study was the evaluation of normalized versus native cDNA libraries when using next-generation sequencing platforms to identify novel genes. Noteworthy, the sequences produced by the experiment, are not yet available in GeneBank. In another paper, 3 out of the 5 cDNA libraries from gonads of *A. fulvescens* (1 male and 2 female) were re-analyzed with the aim of identifying specifically sex determining genes and the presence of xenobiotic organisms that may exist as endosymbionts in sturgeon gonads (Hale et al. 2010). Hale and colleagues claimed to have obtained a total of 473,577 reads by 454 GS-20 and GS-FLX titanium sequencing machines. The female sequence libraries assembled together resulted in 32,629 contigs, while the male library yielded 12,791 contigs. The authors documented significant sex differences in the expression of two genes involved in animal sex determination, DMRT1 and TRA-1 but, again, no sequences were disclosed to the scientific community. Finally, in 2010, Cao and colleagues released in GeneBank 2,025 ESTs obtained by Sanger sequencing of a pituitary cDNA library from a 24 years old female Chinese sturgeon (*Acipenser sinensis*) just after its spawning (Cao et al. 2011). In addition to the transcriptome characterization, the authors focused on the analysis of a somatolactin protein

By concluding only data from two transcriptomic characterization exploiting the potential NGS technology are available in the literature to date, for the same sturgeon species (*A. fulvescens*). In both cases no reads were made available to the scientific community. The remaining projects dealt with Sanger sequencing and have released a few ESTs when compared to the coverage and depth today made available by NGS platforms. In particular no gene annotation is available today for the Adriatic sturgeon, so a first characterization of transcribed sequences in this species would be very interesting. This would represent the starting point for a wide range of genomic studies and for the isolation of EST-linked microsatellites and SNPs to be used as markers of selective effects, in association with the neutral markers which are, up to date, the only available for this species.

1.4 Aims of the study

This Ph. D. thesis deals with the first transcriptomic characterization of *Acipenser naccarii* sturgeon of high evolutionary and conservation interest, belonging to the 240 chromosomes group, with the primarily purpose of making most of the massive information acquired, accessible, to the scientific community, through the development of

a public online database.

In order to evaluate the rate of a functional reduction of the ploidy level which was experienced by transcribed component of sturgeon genomes with different kariotypes, the whole transcriptome of *Acipenser stellatus,* having a half number of chromosome compared to *A. naccarii* (120), was also characterized, thus launching the first step towards a comparison between transcriptomes of sister species with different ploidy levels.

A third goal of this project was to identifying sex-specific molecular markers of relevant aquaculture interest for caviar production, which permit to efficiently select females at an early-life stage. Comparative analyses were performed with sequences from other fishes (or more generally from vertebrate), with extensive genomic resources available, with the aim of identify homologous to genes known to be involved in sex determination and sexual development in other species.

Finally, the project aimed to isolate, for the first time in this group of fishes, a relevant number of SNPs (Single Nucleotide Polymorphisms) and EST-linked microsatellites. They represent the most promising classes of markers to detect footprints of natural selection on genomes. These markers will be important tools for analysing selective events in both natural populations and aquaculture stocks, not only for the Adriatic and Stellate sturgeon, but also for the other over 20 species, distributed in the Northern hemisphere and considered to be highly endangered.

Transcriptomes sequencing of both species was realized, by Roche 454 Titanium sequencing platform on the normalized cDNA brain and gonad libraries of two full-sib individuals per species, one male and one female, obtained by artificial reproduction and reared in aquaculture for six months. This is the first stage in which sex could be identified by histological analysis of gonads (Grandi & Milvia Chicca 2008)

Hereafter, the transcriptomic sequence libraries obtained by the two *A naccari* samples are named respectively: cDNA3 (male sample) and cDNA4 (female sample) while libraries from the two *A. stellatus* juveniles are labelled as: cDNA1 (male sample) and cDNA2 (female sample) each one.

1.5 Thesis Overview

In this study I used a purely bioinformatics approach to analyse the transcriptomic sequence data from 454 technology. Analysis pipelines were built according to the

Bioinfotree computational framework (Sales 2008) and by exploiting already accessible software tools while new tools were developed when those available resulted inadequate.

The rest of the thesis is structured as follow. The chapter 2.1 describes the experimental protocol followed for preparation of the four cDNA libraries, the process of normalisation and sequencing. All these steps were not performed by me, but by other members of the Molecular Ecology group in Padova as well as external laboratories. Nevertheless this chapter is essential for the full understanding of the dissertation.

Subsequent chapters explain work done essentially by myself. The transcriptomic data were analyzed fallowing several computational steps. Each stage is subdivided into the three sections that present: an explanation of the performed work flow (chapter 2 "Methods"), followed by the outcomes obtained (chapter 3 "Results") which are separately reported for the two species under study, and finally a comparative discussion of such results (chapter 4 "Discussion"). The "Results" chapter also report the architecture and content of the public database *AnaccariiBase*, developed with the help of Dr. Alessandro Coppe (section 3.8) to share with the scientific community most of the sequence informations obtained for the *A. naccarii* species. To finish, chapter 5 contains the conclusions.

2 Methods

2.1 Preparation of samples, construction of cDNA libraries and sequencing

Two 6-month-old individuals per species were collected from the “Azienda Agricola VIP” farm (Orzinuovi, Brescia, Italy). Their sex was determined by histological analysis of the gonads as being one male and one female. Animals were anaesthetised with chloretone and painlessly killed. Biopsies were performed from gonads and brain for RNA purification and a part of the gonads was used for sex determination according to the procedures described in Grandi and Chicca (Grandi & Milvia Chicca 2008).

An RNeasy mini-column kit (QIAGEN) was used to extract total RNA from 30 mg of each tissue from each individual. Total RNA were checked for integrity, purity and size distribution. RNA samples from each individual were pooled and stored in three volumes of 96% ethanol and 0.1 volume of sodium acetate to obtain 5 µg of pooled RNA in a final volume of 120 µl. Pooled RNA was sent to Evrogen (Moscow, Russia; www.evrogen.com). The SMART (Switching Mechanism At 5' end of RNA Template) kit was used to retrotranscribe total polyadenilated RNA. First-strand cDNA synthesis was performed with SMART Oligo II oligonucleotide (5`-AAGCAGTGGTATCAACGCAGAGTACGCrGrGrG-3`) and CDS-GSU primer (5`-AAGCAGTGGTATCAACGCAGAGTACCTGGAG-d(T)20-VN-3`) using 0.3 µg of total RNA. Double-strand cDNA was obtained from 1 µl of the first-strand reaction (diluted 5 times with TE buffer) by PCR with SMART PCR primer (5`-AAGCAGTGGTATCAACGCAGAGT-3`). Amplified cDNA PCR product was purified using QIAquick PCR purification Kit (QIAGEN, CA). The two SMART prepared libraries were then normalised using the duplex-specific nuclease (DSN) method (Zhulidov et al. 2004). Normalisation included PCR amplification of the normalised fraction.

In order to gain more material, 30 ng of normalised cDNA were used for 100 µl PCR and 7 cycles of PCR amplification with SMART PCR primer were performed. Adapters were trimmed using GsuI (Fermentas) following the standard protocol and cDNA purification was performed with Agencourt AMPure XP (BECKMAN COULTER). BMR Genomics, University of Padua, Italy (http://www.bmr-genomics.it), prepared and sequenced 454 protocol libraries. Approximately 15 µg of normalized cDNA from each library were sequenced in a 1/4 pico-titer plate on a Genome Sequencer FLX instrument

using GS FLX Titanium series reagents.

2.2 Cleaning and Assembly

Assembly algorithms for transcriptomic data

If no reference genome assembly is available, as happens for most of the transcriptomic sequencing projects for non-model organisms, the strategy commonly adopted is to perform a "*de-novo"* assembly. The *de-novo* transcriptome assembly strategy exploits overlaps between redundant short sequencing reads to assemble them into putative *contigs* (i.e. transcripts) (J. A. Martin & Z. Wang 2011). A contig is defined as a peace of genomic sequences with no gaps, in which the order of bases is known with a high confidence.

De-novo assembly of higher eukaryotic transcriptomes in addition to the large size of the data needed, is hampered by the difficulties involved in identifying alternatively spliced variants. Different transcript variants of the same genes can share exons and thus are difficult to disentangle without ambiguity (J. A. Martin & Z. Wang 2011). Moreover the assembly process is frustrated by tandem and sparse repeats, which can be resolved only if a read is sufficiently long to span the repeat. In genome assembly, this issue can be partially circumvented by sequencing at high coverage or by using sequencing depth information to distinguish the repetitive regions in the genome (Finotello et al. 2012). Unfortunately in transcriptomic assembly, this strategy can mistake abundant transcripts for repetitive region because the sequencing depth of different transcripts (in non-normalised data) can vary by several orders of magnitude.

The improvement of data quality and the rapid development of assembly algorithms in the recent years have made it possible to face the above issues. Different assembly software have been developed that differ in the algorithms they use and how they treat individual reads (reads can be split and placed in different contigs), providing different levels of accuracy and completeness of reconstructed transcripts.

Algorithms for *de-novo* assembly fall within a class of NP-hard problems (i.e. no efficient computational solution is known). They can be classified in three main classes: (i) Greedy algorithms, (ii) Overlap-Layout-Consensus (OLC) algorithms and (iii) Eulerian algorithms (Miller et al. 2010). Greedy are optimisation algorithms that build up a solution step by step, always making the locally optimal choice at each stage with the hope to finding the global optimum. In the case of the assembly, problem, it starts from the pair of reads that overlap more and try to extend the alignment by adding on the left or on the right

the other reads with the best overlap. It continues iteratively until no more reads can be assembled. The TIGR (GG Sutton et al. 1995) and CAP3 (X. Huang & Madan 1999) assembler use this approach.

The OLC strategy is arguably the most successful in a practical setting both with long and short reads. Within this approach, read are compared to each other to construct a list of pair-wise overlaps. These informations are then used to construct an overlap graph. The graph is then searched for the path that traverses each node exactly once (*Hamiltonian path*). Finally, the consensus DNA sequence is calculated through a multiple sequence alignment. Software such as the Celera Assembler, the Roche GS De Novo Assembler known as Newbler (454 Life Sciences) and the Mimicking Intelligent Read Assembly alias MIRA, (Chevreux et al. 2004) assembler, use this strategy. Greedy and OLC assemblers share a module called "overlapper" that computes all pairwise alignments between initial reads. This represents the most time-intensive components of their algorithm. This time is proportional to the square of the number of reads $O(n^2)$ or can became simply proportional to the number of reads (polynomial) using indexing strategies, but fortunately overlap computation can be easily parallelized.

Finally Eulerian path (*de Bruijn Graph*) approach was adopted for solve the problem of assembly million of very short reads (in the 25-200bp range) from Solexa (today Illumina) and SOLiD platforms. The algorithm starts by breaking the set of reads into their *k-mer* spectrum (the list of all oligomers of a length *k* that compose each reads). The resulting k-mer spectrum is then used to construct a de Bruijn graph where k-mer are nodes and edges are connection between them based on the sharing of a prefix sequence. The graph is then searched for finding a path (the Eulerian path) that uses every edge in the graph (i.e. every k-mer) in order to reconstruct the genome sequence. The time complexity of this algorithms is estimated to be $O(n \log n)$ were n is the number of k-mer (Zerbino & Birney 2008) so they are faster compared to Greedy and OLC but when applied to solve assembly problems they generally yield short contigs. Assemblers that use this approach are for example Velvet (Zerbino & Birney 2008), Allpaths (MacCallum et al. 2009) and AbySS (Birol et al. 2009).

Since the length of the 454 reads is > 200 bp and the throughput is in the order of hundred of Mb, the algorithms that best fit to *de-novo* assembly of trascriptomic sequences from this platform are the OLC ones.

The proprietary Newbler assembler software is distributed with the 454 sequencing machines and represents the primary choice for many genomic and transcriptomic *de novo* assembly project exploiting 454 sequencing. It provides both a Graphical User Interface (GUI) and a command line interface (CLI). The main advantage over other assemblers is that base-calling and quality value determination for contigs are performed in "flowspace" by processing the flow signals (a continuous variable) at each nucleotide flow of the sequencing run, encoded into the proprietary Standard Flowgram Format (SSF). Noteworthy, Newbler try to solve splicing variants by breaking multiple alignments of overlapping reads that show alignments towards different contigs, into branching structures (called isogroups). Then it traverses the various paths through an isogroup to produce the set of contigs (isotigs) forming it. The reported isotigs are the putative transcripts.

The MIRA 3 assembler represents an excellent alternative due to its freely available source code and en admirable support by the author. It is finely tunable and provide a specific module for assembly EST and RNASeq data. It has been developed with attention to problem of repeats in the assembly. It uses a multi-pass approach, to explore the assembly space, using, in the first steps, a parameterisation that allows some errors in the assembly. Such errors are used to drive a combinatorial approach of hypothesis generation/testing and for training a neural networks, to discern, in the subsequent steps, between sequencing errors and true differences within the multiple alignments of reads with the aim to improve the overall alignment quality. In the assembly of ESTs, MIRA adopts a very conservative approach, meaning that it splits read alignments into different contigs if it has enough evidence that they come from different alleles of the same transcribed gene (Chevreux n.d.).

Assembly strategy

The raw reads from every library were extracted from 454 SFF files through the open source alternative sff_extract 0.2.10 as it is the only known 454 base-calling software which was not under a restrictive license. Sequences and qualities were tagged based on the library of origin. Summary control of raw reads' quality was done with FastQC 0.10.0 (Andrews Simon n.d.). Sequences were cleaned using the est_process module driven by preprocessest.pl script into the est2assembly 1.13 package (Papanicolaou et al. 2009) that perform sequencing adaptor removal, low complexity region masking, quality trimming, and poly A/T detection and removal. This preprocessing routine was built to benefit of NCBI's TraceInfo XML in order to annotate vector and clipping positions. This permits to

use the original (unmasked) data to built assemblies with tools such MIRA which can make intelligent decisions if a clipped region is a false positive.

Since it was highlighted that the standard Newbler 2.3 assembler adopted an approach too restrictive by dismissing an excess of reads, I evaluate its performances compared to that of MIRA 3.2 by independently assembly the 4 cDNA libraries from *A. naccarii* and *A. stellatus* using their default parameters for EST assembly. Parameters used with MIRA were: -job=denovo,est,accurate,454 -fasta -GE:not=2 -LR:fo=no:mxti=yes 454_SETTINGS -DP:ure=0 -CL:cpat=0:qc=0 -ED:ace=1 -OUT:sssip=yes while parameters used with Newbler were: -cdna -nov -notrim -novt -novs.

The quality of the assemblies was then tested based on the number of reads included, the number of contigs produced and other indexes resulting from a BLAST similarity search against four reference databases (e-value < 1e-05 and bit-score 80). Databases were constructed by downloading nucleotide and protein sequences from the American National Centre for Biotechnology Information (NCBI, 2010/10/11): (i) 4,672 cDNA for the genus *Acipenser* mainly from *A. sinensis* (Cao et al. 2011) and *A. transmontanus* (Lazzari et al. 2008), (ii) 57,073 protein sequences from the genome of *Danio rerio*, (iii) 18,782 protein sequences from the genome of *Takifugu rubripes*, (iv) 20 protein sequences for the genus *Acipenser*. Each set was made 100% non-redundant with CD-HIT clustering tool (Fu et al. 2012). Assemblies quality was then tested by counting the total number of positions (amino acids or nucleotides) uniquely identified across the reference databases (coverage) while a redundancy index was calculated by counting the number of times the same position (base or amino acid) was identified between the reference databases and the assemblies.

The effects of MIRA parameters were subsequently evaluated especially concerning the management of repeats, by performing 6 different assemblies of all the libraries varying the parameters that affect the assembly of the repeats (listed in Table 2.1), and evaluating the number of assembled reads, the produced contigs, the number of identified sequences, the coverage and the redundancy level, with respect to the aforementioned database. Assessment of assemblies and parametrization were performed after having hacked the parametrize_assembly module within the est2assembly package (Papanicolaou et al. 2009). To reduce the internal redundancy of sequence assemblies, due to the heuristic nature of the assembly algorithms (Kumar & Blaxter 2010), a routine was developed to execute iterative assembly cycles through MIRA, in which contigs and singletons from the previous round, are reassembled in the next one. Two cycles resulted sufficient to remove

most of the existing redundancy. In the first run (*de-novo* assembly), all cleaned reads were used as input and processed with the following parameters: -job=denovo, est, accurate, 454; -LR:fo=no, -SB:lsd=no, -CL:cpat=0:qc=0, -ED:ace=1, -OUT:sssip=yes, -CO:fnicpst=yes, -LR:mxti=no, -AS:mrpc=1. In the second run the following parameters were used: -job=denovo, est, accurate, 454, -notraceinfo, -LR:fo=no, -CL:cpat=0:qc=0, -ED:ace=1, -OUT:sssip=yes, -CO:fnicpst=yes, -LR:mxti=no, -AS:mrpc=1. After the second assembly step, the native reads of each contig (from the first round) or metacontig (from the second round) were traced back through SSAHA2 2.5.4 (Ning et al. 2001) with specific settings for Roche 454 reads alignment (-rtype 454 -output sam). For each of the above contigs or metacontigs, the alignment of the corresponding reads resulted in a SAM file allowing the calculation of the coverage using SAMtools (H. Li et al. 2009) and BEDtools (Quinlan & Hall 2010).

MIRA settings	Assembly effects
-job=denovo,est,accurate,454 -fasta -GE:not=2 -LR:fo=no:mxti=yes 454_SETTINGS -DP:ure=0 -CL:cpat=0:qc=0 -ED:ace=1 -OUT:sssip=no	***Run 1***: *default setting, MIRA used as assembler.*
-job=denovo,est,accurate,454 -fasta -GE:not=2 -LR:fo=no:mxti=ye**s -CO:asir=yes** 454_SETTINGS -DP:ure=0 -CL:cpat=0:qc=0 -ED:ace=1 -OUT:sssip=no **-CO:fnicpst=yes:rodirs=5 -AL:egp=no:mrs=93**	***Run 2***: *MIRA tuned for a clustering assembly.* Accurate, switch off extra gap penalty, disallow building of completely nonsensical contigs, SNPs are assumed valid base differences.
-job=denovo,est,accurate,454 -fasta -GE:not=2 -LR:fo=no:mxti=yes 454_SETTINGS -DP:ure=0 -CL:cpat=0:qc=0 -ED:ace=1 -OUT:sssip=no **-CL:mbc=1:mbcgs=30:mbcmfg=30:mbcmeg =30 -CO:mrpg=10 -AL:egp=no**	***Run 3***: *increase stringency of SNP and repeats detection.* Accurate: remove extra gap penalty option, minimum number of aligned reads needed to be recognized as repeat marker or a SNPs set to 10.
-job=denovo,est,accurate,454 -fasta -GE:not=2 -LR:fo=no:mxti=yes 454_SETTINGS -DP:ure=0 -CL:cpat=0:qc=0 -ED:ace=1 -OUT:sssip=no **-CL:mbc=1:mbcgs=30:mbcmfg=30:mbcmeg =30 -CO:mrpg=10 -AL:egp=no -ALIGN:bip=20:bmax=120:mo=10**	***Run 4***: *decrease stringency of reads alignments.* Accurate, remove extra gap penalty option, bandwidth of the Smith-Waterman alignment routines to 20%, bmax to 120, minimum overlap decreased to 10.
-job=denovo,est,accurate,454 -fasta -GE:not=2 -LR:fo=no:mxti=yes **-SKIM:nasty_repeat_ratio=15** 454_SETTINGS -DP:ure=0 -CL:cpat=0:qc=0 -ED:ace=1 -OUT:sssip=no **-CL:mbc=1:mbcgs=30:mbcmfg=30:mbcmeg =30 -CO:mrpg=10 -AL:egp=no -ALIGN:bip=20:bmax=120:mo=10**	***Run 5***: *increase the ratio by which repeats are considered nasty.* Accurate, remove extra gap penalty, option, bandwidth defaults to 20%, bmax to 120, minimum overlap decreased to 10, increase to 15 the frequency of k-mer that have to be masked by the initial fast all-against-all read comparison algorithm).
-job=denovo,est,accurate,454 -fasta -GE:not=2 -LR:fo=no:mxti=yes **-SKIM:nasty_repeat_ratio=7** 454_SETTINGS -DP:ure=0 -CL:cpat=0:qc=0 -ED:ace=1 -OUT:sssip=no **-CL:mbc=1:mbcgs=30:mbcmfg=30:mbcmeg =30 -CO:mrpg=10 -AL:egp=no -ALIGN:bip=20:bmax=120:mo=10**	***Run 6***: *decrease the ratio by which repeats are considered nasty.* Accurate, remove extra gap penalty option, bandwidth defaults to 20%, bmax to 120, minimum overlap decreased to 10, frequency of repeated k-mer to be masked decreased to 7.

Table 2.1. Different settings evaluated during MIRA assembler parametrization. The changed parameters with respect to the default are highlighted in bold to the left. A short explanation of the main effects of each setting is given in right column.

2.3 Estimation of sequencing completeness

To test how completely the physical cDNA libraries were sequenced, I adopted the method described in (Franssen et al. 2011), based on saturation curve calculation. From the total cleaned reads pool, increasing subsets of reads were randomly selected and, for each

read, the corresponding contig in which it was assembled was traced back. Detected contigs were blasted against a reference cDNA set using TBLASTX with the e-value cut-off at 1e-03. The best matching subject was recorded for each contig. The sampling was repeated 20 times with a constant increase in sample size, reaching the totality of cleaned reads in the last run, thus identifying, in the end, 20 pools of different reference cDNAs.

The number of matching reference cDNAs at each cycle was plotted against the corresponding reads sample size and a hyperbolic model $y = ax/(b+x)$ was fitted to the points by non-linear regression to assess the parameters "*a*" and "*b*". In order to estimate a confidence interval for each data point and for the model parameters, we repeated the whole process (sampling plus calibration) 10 times. Finally, the number of reads in each sample was plotted against the average of the captured reference cDNAs. The hyperbolic model was fitted on these points giving the final values for "*a*" and "*b*" and the derivative of the function at the higher sample size was calculated. The parameter "*a*" represents the upper limit of the model function, i.e., the maximum theoretical number of reference transcripts identifiable by the initial cDNA libraries if these had been exhaustively sequenced, indicating an upper limit for transcript detection. Moreover, the slope of the hyperbolic curve at maximum sample size gives an evaluation of how quickly the asymptotes "*a*" will be reached, thus indicating the decreasing potential to detect additional transcripts.

For each species, I built saturation curves by sampling cleaned reads from: 1) male only, 2) female only and 3) joint libraries. In all cases, I mapped reads back to the final assembly contigs. A whole cDNA super-set from *Danio rerio* in Ensembl release-66 was chosen as the reference. However, my analysis demonstrated that the fraction of detected reference transcripts, with respect to the maximum estimated, and the slope of the curve at maximum sample size do not substantially change using different cDNA sets as a reference (see Figure 3.1 and Figure 3.3).

2.4 Estimation of transcriptomes completeness

I inferred the total transcripts population size in the two *A. naccarii* and the two *A. stellatus* samples respectively, by estimating the number of transcripts shared by the two independent libraries of each species. By neglecting the differences due to sex-specific transcripts, the two sequence libraries (within each pair), were handled as two independent samples from the same transcripts' population. The fraction of the transcript from the first

library that is also represented in the second one is a direct estimate of the completeness of the second library and vice-versa. The same approach is widely used in ecology to estimate animal population sizes (Chao 1989) and has already been applied to estimate the number of human genes (Ewing & Green 2000). Since within species, each read was labelled with the library of origin before joint assembly, final contigs were classified as being “male_library-specific", “female_library-specific" or “common". The common one is the fraction of contigs composed of reads of both libraries and then represented by transcripts considered to be shared by the two libraries. Figure 2.1 better exemplify the method.

I performed a direct subtraction, i.e. a bidirectional BLASTN, within each pair of libraries to identify the contigs that were not library-specific. Library-specific contigs that align for more than 80% of their length, with e-values below 1e-50 were moved into the common fraction. I also performed an indirect subtraction to take into account contigs representing partially-overlapping or non-overlapping portions of the same long transcript, which had not been assembled together due to the lack of a sufficient link.

Both groups of library-specific contigs, within each species, were searched for similarities, using TBLASTX, against a super-set of all transcripts resulting from Ensembl (release-66) including known-, novel- and pseudo-gene predictions for the following list of reference species (hereafter RS-list): *Petromyzon marinus, Latimeria chalumnae, Danio rerio, Gasterosteus aculeatus, Oryzias latipes, Takifugu rubripes, Tetraodon nigroviridis* and *Homo sapiens*.

All cDNA sequences provided by NCBI-Taxonomy Browser for the genus Acipenser (2011/04/18) were also screened. Protein sequences available for other species were assessed by searching the NCBI nr database (2010/11/02) with BLASTX. The library-specific contigs matching the same subjects, with 1e-06 as the e-value threshold and > 80% query coverage were moved into the common fraction.

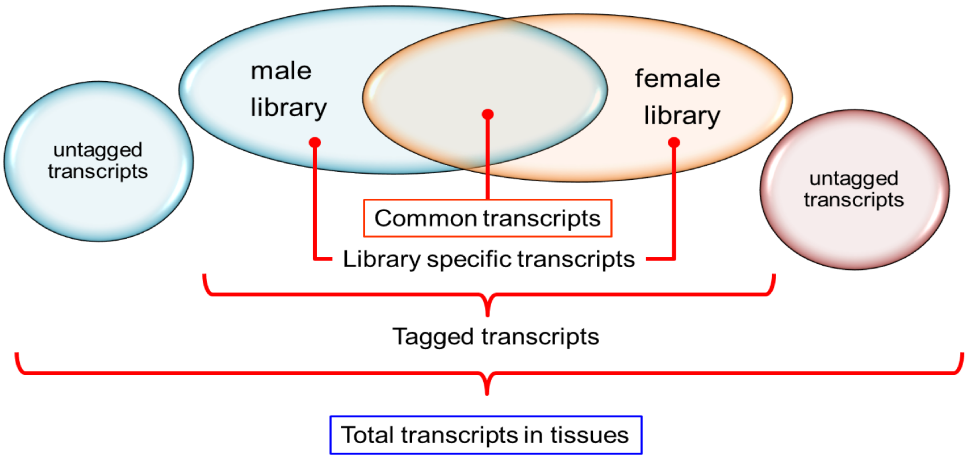

Figure 2.1. Schematic representation of the method used to estimate the transcripts population size in the tissues of origin. The fraction of transcripts from the male library also present in the female library is proportional to the total transcripts in the sequenced tissues.

I exploited the Rcapture R package (Baillargeon & Rivest 2007) to estimate the total transcripts population sizes because it allows the association of a standard error to the obtained estimation. Furthermore, I assessed the completeness of the *A. naccarii* transcriptome by screening for the presence of the 37 genes (22 mt-tRNA, the mt-rRNA 16S, 12S and the 13 polypeptide coding genes) from the complete mitochondrial genome (mt) of *Acipenser transmontanus* (GenBank accession no.: AB042837) using BLASTN with a 1e-10 e-value threshold. I screened the *A. stellatus* assembly for mitochondrial genes also, using its complete mitochondrial genome, already available in GenBank with the accession no.: NC_005795.

2.5 Functional annotations

De novo annotations of *A.naccarii* and *A. stellatus* transcriptomes were performed for both coding and non-coding fractions, through a multi step procedure, starting from similarity search against gender specific nucleotide sequences, main protein and nucleotide databases, full transcribed and protein sequences from other fishes in Ensembl database.

BLASTX 2.2.25+ (Camacho et al. 2009) similarity searches for the entire transcriptomes were conducted locally against a NCBI non-redundant (nr) database (downloaded 2010/10/19) as the first *de novo* annotation step. The Swiss-Prot part of the UniProt database (downloaded on 2012/02/24) was also queried. Local TBLASTX 2.2.25+ similarity searches were conducted locally against (1) 6,088 EST sequences of the genus *Acipenser* downloaded from NCBI-taxonomy (2011/04/18), witch included 2,025 pituitary

sequences obtained from chinese sturgeon *A. sinensis* (Cao et al. 2011) and 2,704 sequeces from *A. transmontanus* (Lazzari et al. 2008) (2) a super-set of all cDNA from RS-list species. In order to identify non-coding sequences, BLASTN 2.2.25+ similarity searches were conducted locally against whole non-coding RNA gene and pseudogene sequences from Ensembl release-66 for the species of the RS-list. A BLASTN similarity search was also performed against the NCBI nucleotide sequence (nt) database (downloaded on 2012/02/24). BLASTX, BLASTN and TBLASTX searches were carried out using default parameters.

Given the high evolutionary distance among the species compared, alignments with an e-value < 1e-03 were considered significant and a maximum of 20 hits were taken into account for each query. The taxonomic classification of annotations was performed by MEGAN 4 (Huson et al. 2011) based on the absolute best BLAST hits. Contigs with multiple best BLAST hits were excluded from the count. The mapping of GO annotations to contigs was achieved with Blast2GO 2.4.7 (Götz et al. 2008). Annotations were conducted only for contigs with significant BLASTX hits below e-value 1e-06, with 55 as the annotation cut-off and 5 as the GO weight. No HSP-hit coverage cut-off was used. InterProScan and KEGG pathways annotation was also conducted via Blast2GO. Obtained information for domains was included to improve global annotations.

2.6 Search for sex-determining genes

I obtained sequences from genes known to be involved in sex determination and sexual development in vertebrates from different species and used them to search the assembled contigs of both species, by similarity in order to investigate the content of library-specific contigs, isolated by *in silico* subtraction in more detail. The genes and gene families considered were: WT1, LHX1, CYP19A1, FHL3, FEM1A, AR, EMX2, DAX1, SOX9, SOX17, SOX1, SOX11, SOX6, SOX14, FOXL2, RSPO, SF1, FGFR2, FGF9, GATA4, LHX9, ATRX, SOX2, SOX4, SOX21, WNT4, SRY, STRA8, FIGLA, AMH, VTG2, DMRT1 (Pala et al. 2009) (Shirak et al. 2006) (McClelland et al. 2012). I obtained sequences in 3 different ways: 1) Ensembl database annotated orthologous and paralogous of the above genes were identified, starting from the well-annotated Zebrafish genome in Ensembl 66, by querying each common name. For each gene, I identified all orthologous and paralogous within Ensembl Compara version 66. Then, for each ortholog and paralog, all alternative transcripts were identified and the corresponding protein sequence downloaded. 2) Clusters of homologs (paralogs and orthologs) of candidate genes were

identified within NCBI HomoloGene Release 66 and corresponding protein sequences were downloaded. 3) Nucleotide sequences for genes FOXL2, DMRT1, and SOX used as references in a previous scientific study aimed at gender identification in the Shovelnose sturgeon (*Scaphirhynchus platorynchus*) (Amberg et al. 2009) together with corresponding sequences from other sturgeons of the genus Acipenser were downloaded from NCBI Genbank (15/10/2012). Each group of paralog and ortholog protein and nucleotide variant representing a gene was searched for similarity in both transcriptomes assembly using TBLASTN and BLASTN respectively. Alignments with an e-value > 1e-03 and fewer than 50 positive matches were discharged. Each different contig that presented a match was extracted for each gene. For each contig (subject) matched by more than one homologue (query), the homologue with the highest alignment bit score was selected. Results obtained by the three approaches were compared for each gene and the more-likely contig was selected based on the following criteria: 1) BLAST alignment bit-score with the query; 2) per-base mean coverage (singletons were discarded); 3) nucleotide alignments between candidates to ensure they actually represented distinct sequences (using MAFFT v6.935b (Katoh & Frith 2012)); 4) alignments between contig translations and corresponding protein queries (using MAFFT); 5) presence of one or more distinctive and important functional domains encoded by the target gene within the translated and aligned fraction of contigs (by searching in Pfam-A version 26.0 (Finn et al. 2009)); 6) the ratio between the length of the translated-aligned fraction and the total contig length; 7) consistency of annotations obtained via blast2GO.

2.7 Discovery of variants

Since mean contig coverage is generally low (< 5X) and the transcriptome comes from different individuals, I adopted a method based on a probabilistic framework, which allows the estimation of uncertainty regarding variants calling, in order to identify SNPs and short INDELs (R. Nielsen et al. 2011). I used Freebayes 0.9.4 (Garrison & G. Marth 2012) which employs Bayesian formulation to calculate the probability that multiple different alleles are present between the reference and the aligned reads. Freebayes is also able to call variants from polyploid pooled samples.

SAM alignments calculated for each contig in the assembly phase were input into Freebayes with the following parameters: probability cut-off of 0.9, 5 as the minimum coverage required to process a site, and each SNP must be supported by at least 2 reads. As it has been shown that improved base-call accuracy can lead to a significant reduction in

false-positive SNP calls, base alignment quality (BAQ) adjustment was applied to the input alignments through SAMTOOLS calmd 0.1.18 (H. Li et al. 2009). Homogenisation of the potential insertion and deletion distribution through reads-independent left realignment to improve the INDELs call was obtained by the bamleftalign tool included in the Freebayes package. It is known that variant calling near repetitive DNA sequences are prone to error, especially in 454 technology where over-calls or under-calls of repetitive stretch, are the most common errors (Brockman et al. 2008). I then filtered out all variants that were beside 4 repetitions of any sample sequence repeats (including homopolymeric regions). For each contig containing SNP, I calculated the number of transitions (Ts), transversions (Tv) and (Ts + 1)/(Tv + 1). As already explained in (Schwartz et al. 2010) the one is added to permit the ratio to be calculated for contigs that have Tv = 0.

The mutation resulting from each SNP was characterized in terms of synonymous (Ks) or non-synonymous (Ka), and location (inside the ORF or in the 5' or 3' UTR regions). For each contig-containing SNP with a BLAST hit against the nr database, the ORF was deduced from the alignment against the best HSP. For contig-containing SNPs without a hit, the ORFs predicted by the ORFpredictor were used. I then calculated the ratio (Ka + 1)/(Ks + 1) where, again, 1 was added to enable the calculation of the ratio even when Ks = 0.

To find both perfect and imperfect microsatellite repeats (SSRs) for di-, tri-, tetra-, penta- and hexa-nucleotides in unit-size, within contigs sequences of each assembly, I adopted the MISA tool version 1.0 (Thiel et al. 2003), with min_repeat specifications of 6, 4, 4, 4, and 4 respectively. These thresholds are in agreement with the minimum lengths recommended for repetitions outlined in (Mudunuri et al. 2010), to allow the polymerase slippage events, which makes the identified microsatellites, potentially polymorphic. I set to 0 the maximal number of nucleotides that interrupt compound microsatellites. Moreover I distinguished SSRs within ORFs predicted by BLAST comparisons or in alternative by ORFpredictor.

2.8 Evaluation of functional ploidy levels

The two sturgeon species under study are characterised by different numbers of chromosomes: 240 in *A. naccari* and 120 in *A. stellatus*. To evaluate if this difference is also detectable at transcriptome level and if the functional ploidy in the two species reflects the difference of nominal ploidy I adopted a strategy conceptually similar to those

described in (Trick et al. 2012; Kaur et al. 2011) and (Trick et al. 2009) by exploiting the MiraSearchESTSNPs assembly pipeline part of the MIRA v. 3.4.0 assembly package.

MiraSearchESTSNPs is an automated pipeline principally designed to cleanly assemble, transcripts coming from different strains. Moreover it allows to reconstruct and keep distinct the different alleles based on the SNPs detected between them. It consists of 3 single assembly steps. In the first steps, reads from all libraries are assembled together. SNPs between individuals are treated as "possible repeat marker bases". This step aims to identify all possible alignments between reads both within and between individuals. Also singletons (sequences that align but that are put aside due to SNPs) are labelled and taken into account for the next steps. In the second step the reads of each individual, marked in the first cycle, are now separately assembled with a higher stringency of overlaps and gap cost, thus enabling transcripts that differ for SNPs to be allocated in separated contigs. This contigs are considered to be different allelic variants of the same transcripts. In the third phase the contigs (alleles) produced in the previous step, within individuals, are assembled together, with the help of alignments from step 1, in order to cluster allelic variants of the different strains, resulting from the same locus. For each contig (shared transcripts) of the third step, I report the number of aligned contigs from the second step (alleles), to estimate the maximum number of alleles owned by each individuals of the two species. Contigs marked as "remains" were not taken into account.

To evaluate the effectiveness of MIRA parameters used in step 2 in order to distinguish allelic variants within individual, I searched for the presence of the 11 constitutively expressed coding genes from *A. transmontanus* complete mitochondrial genome, within contigs of step 3. The genes for tRNAs and rRNAs were discharged, because, resulted in not sufficiently discriminative alignments. The mitochondrial genome is unique per individual and thus constitutes a source of transcripts from single locus genes. For each contig of the phase 3 representing the best match against each mitochondrial gene, the corresponding contig of step 2 (alleles) were counted. Obtained values within individual, were used to determine the error in the alleles count.

3 Results

3.1 Cleaning and Assembly

Test of Mira and Newbler assemblers

When I started the project, it was known that, the public release version 2.3 of Newbler available at the time, was afflicted by several undesirable features some of which were highlighted later by various authors (Papanicolaou et al. 2009) (Kumar & Blaxter 2010) (Finotello et al. 2012). In particular Newbler 2.3 was known to be overly prudent by discharging too much reads. This hypothesis was tested by comparing the quality of the assemblies produced by MIRA 3.2 and Newbler 2.3 using the data from both the two *A. naccarii* and the two *A. stellatus* sequence libraries.

The one quarter picotiter plate of a 454 FLX sequencing run generated 154,882 and 176,703 reads from the *A. naccarii* male (cDNA3) and female (cDNA4) respectively. FastQC (Andrews Simon n.d.) overview of raw sequences showed that mean per-base quality remains above 24 for the first 350 bp and drops rapidly towards the end of the reads (data not shown). The cleaning process was passed by 99% of the reads from each library, yielding a total of 110.25 Mbp cleaned sequences with an average length of 336 bp and mean Phred quality of 28. The main features of the sequences that passed the preprocessing step are summarized in Table 3.1 while their length distribution is plotted in Figure 3.1. The mean GC content calculated for the whole dataset was 37.92%. GC content across sequence length follows a normal distribution thus discarding the hypothesis that systematic bias was present (data not shown). As expected, more than 50% of the total sequences (121,467 sequences) were 400 bp or longer.

Category	Male (cDNA3)	Female (cDNA4)	Total
Total number of raw reads	154,882	176,703	331,585
Total number of cleaned reads	153,215	175,198	328,413
Percentage of cleaned reads	99.00	99.00	99.00
Median length (bp)	376	354	365
Average GC content in percentage	37.77	38.06	37.92
Total length of cleaned reads (Mb)	52.91	57.34	110.25
Average Phred quality of cleaned reads	28	28	28

Table 3.1. Statistics of reads preprocessing for *A. naccarii* libraries.

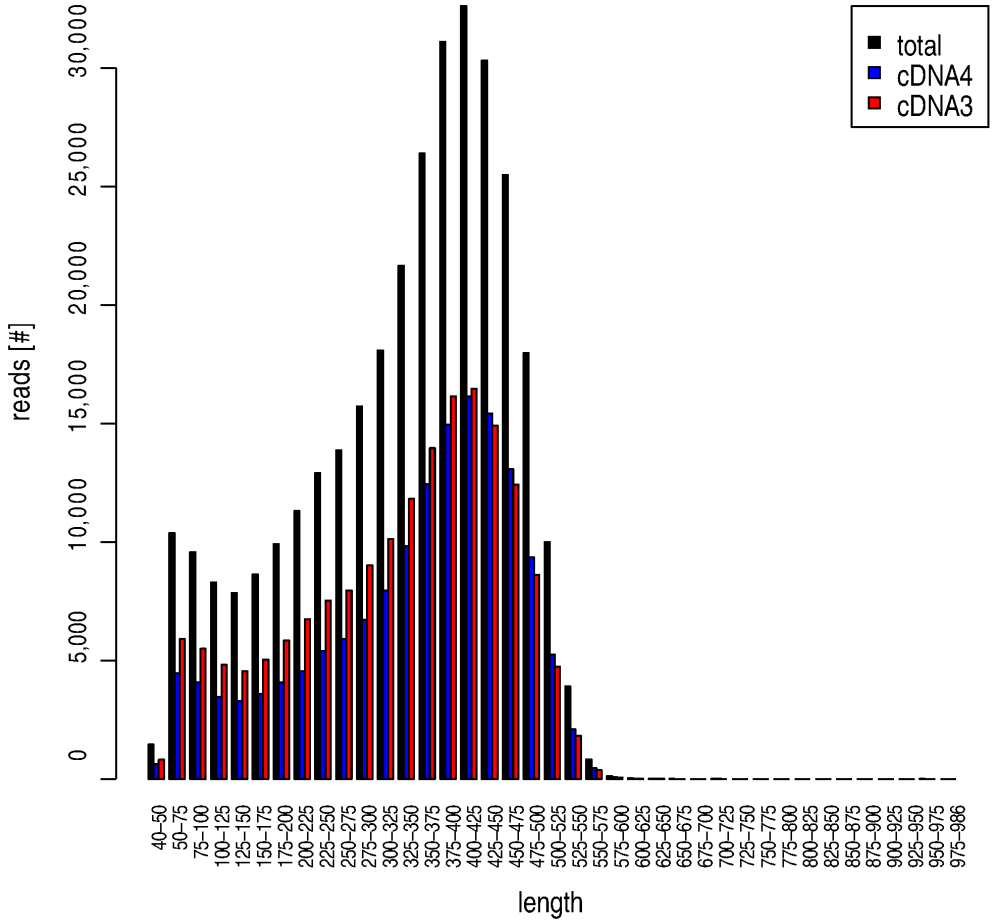

Figure 3.1. Distribution of cleaned-read lengths for the *A. naccarii* male library, *A. naccarii* female library and the joined libraries. Bin intervals are shown along the x-axis.

A total of 184,374 and 169,286 raw reads were generated from the *A. stellatus* male (cDNA1) and female (cDNA2) respectively, from a quarter picotiter plate of a 454 FLX sequencing run. After the preprocessing step, 181,989 male and 167,714 female sequences (99% in both cases) were selected for the assembly process, representing 116.89 Mbp of sequences. The main features of the cleaned sequences are summarized in Table 3.2 while their length distribution is plotted in Figure 3.2.

Category	Male (cDNA1)	Female (cDNA2)	Total
Total number of raw reads	184,374	169,286	353,660
Total number of cleaned reads	181,989	167,714	349,703
Fraction of cleaned reads (%)	99.00	99.00	99.00
Median length (bp)	366	357	361
Average GC conten (%)	37.18	37.10	37.14
Total length of cleaned reads (Mb)	61.29	55.59	116.89
Average Phred quality of cleaned reads	29	29	29

Table 3.2. Statistics of reads preprocessing for *A. stellatus* libraries.

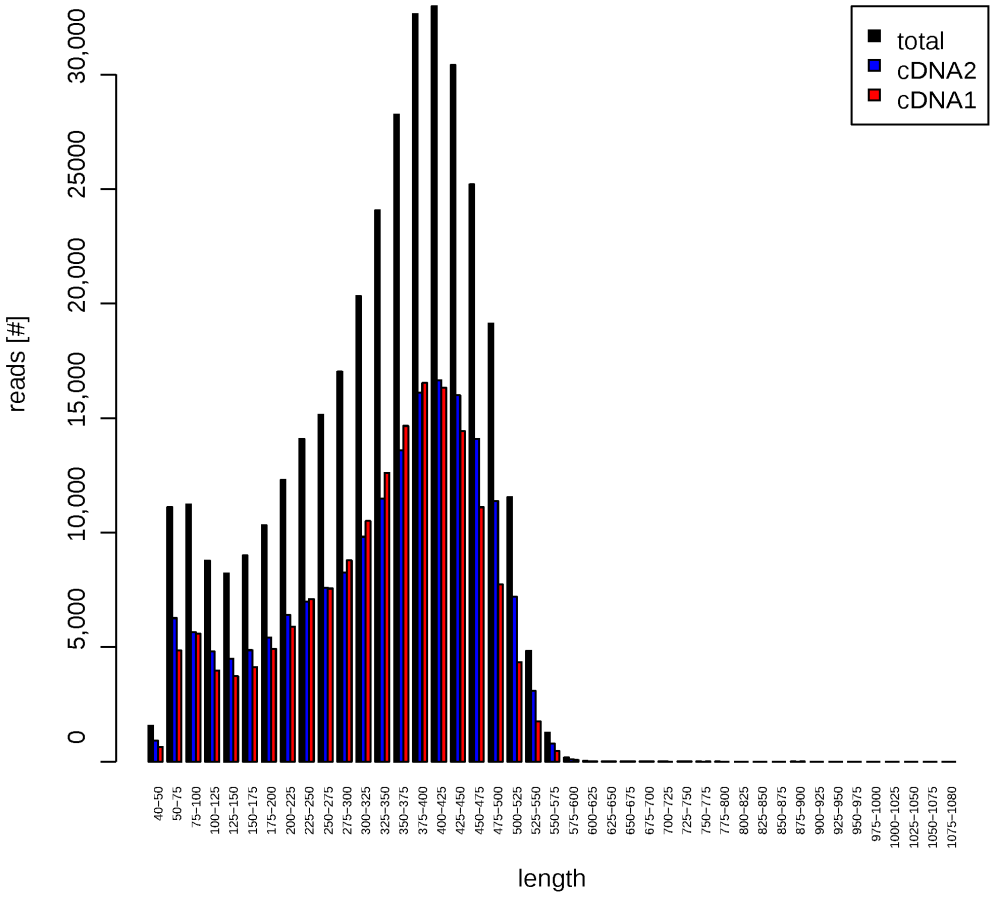

Figure 3.2. Distribution of cleaned read lengths for *A. stellatus* male, female and joined libraries.

The 4 libraries were independently assembled, both with MIRA and Newbler. The number of reads assembled and produced contigs for each library, are plotted in Figure 3.3. The results of the quality control, concerning the number of different sequences tagged and

positions covered summed across the 4 reference databases, are reported in Figure 3.4. The Percentages of assembly internal redundancy, that essentially express how many multiple hits each assembly has versus the same reference databases taken together, are plotted in Figure 3.5. As evident from the data, MIRA actually uses on average 25% more reads then Newbler; this results in many more assembled contigs. The assemblies from MIRA were the most comprehensive: they were able to recognize, on average 36% more sequences and cover 21% more positions of the references, compared to Newbler assemblies. The internal redundancies, were comparable for the two groups. MIRA 3.2 therefore, was chosen in order to obtain the most representative assemblies of the transcriptomes under study.

The MIRA assembling algorithm parameters related to the stringency of the alignments between reads and the handling of repeats, were then explored. The command line option combinations listed in Table 1.1 were tested by assembling reads from all libraries in 6 replicates. In Figure 3.6, are represented the effects on the numbers of reads used and produced contigs. Figure 3.7 represents the fractions of unique hits and positions covered, on the reference databases. Finally, the trends of the calculated internal redundancies are represented in Figure 3.8. The results shown a noteworthy difference in the assemblies when MIRA was set as a clusterer, while different treatments of repeats had negligible effects

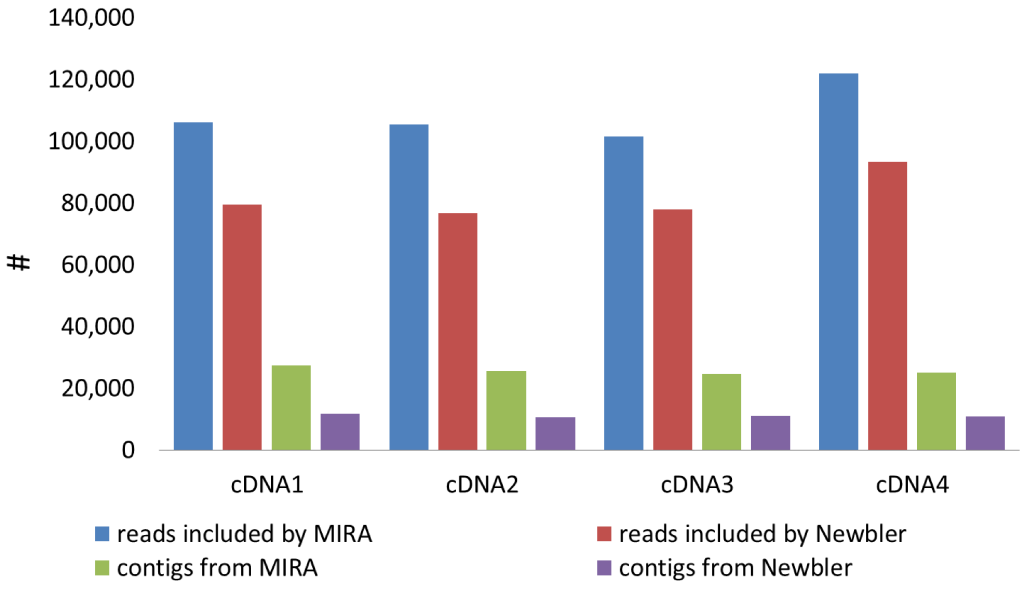

Figure 3.3. Comparison between Newbler 2.3 and MIRA 3.2 with respect to reads included and produced contigs.

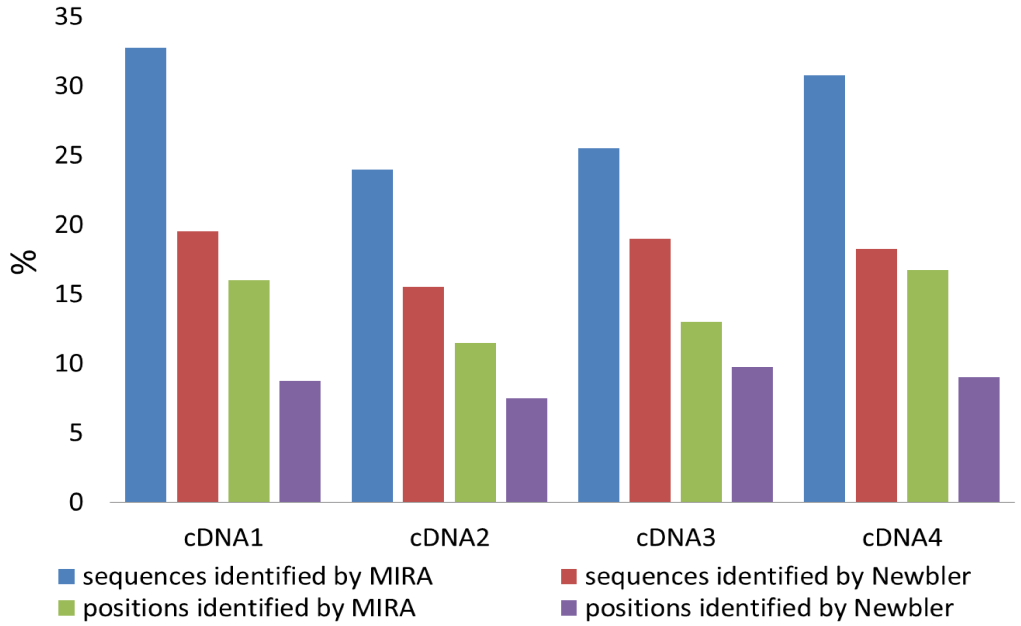

Figure 3.4. Comparison between Newbler 2.3 and MIRA 3.2 with respect to sequences identified and positions covered across reference databases.

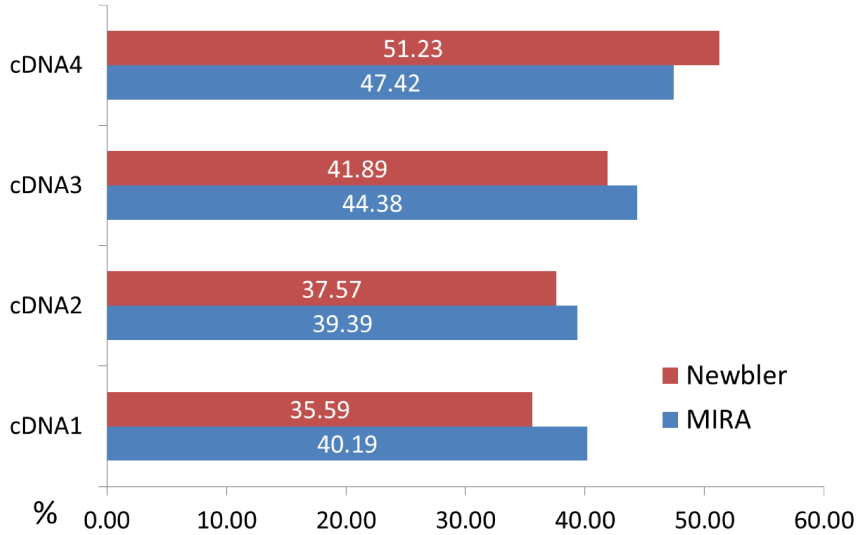

Figure 3.5. Comparison of Newbler 2.3 and MIRA 3.2 with respect to the redundancy index of based on the number of one-to-many hits (summed in both directions) between the reference databases and the assemblies.

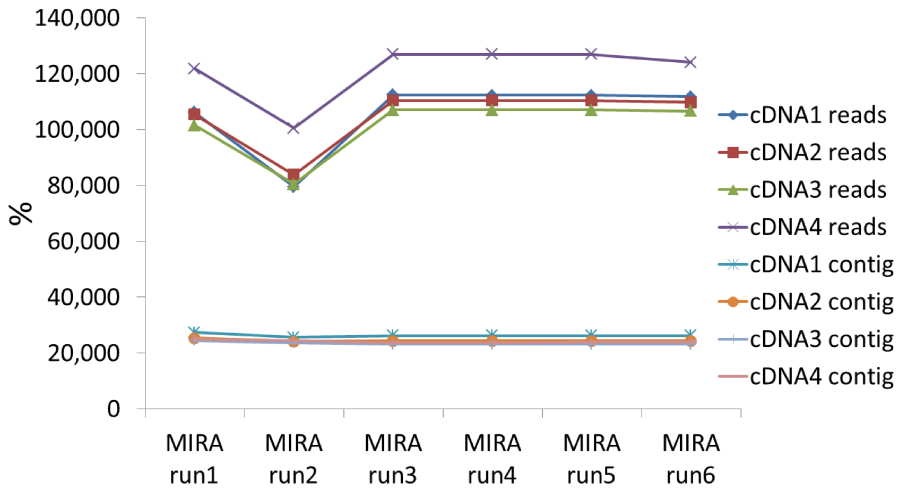

Figure 3.6. Setting effects on the number of reads assembled and contigs obtained.

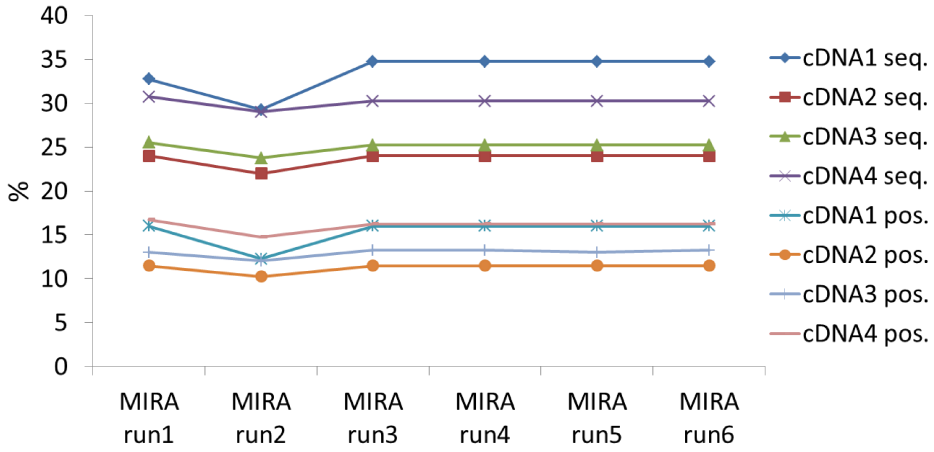

Figure 3.7. Setting effects on the number of sequences and positions identified in the reference databases.

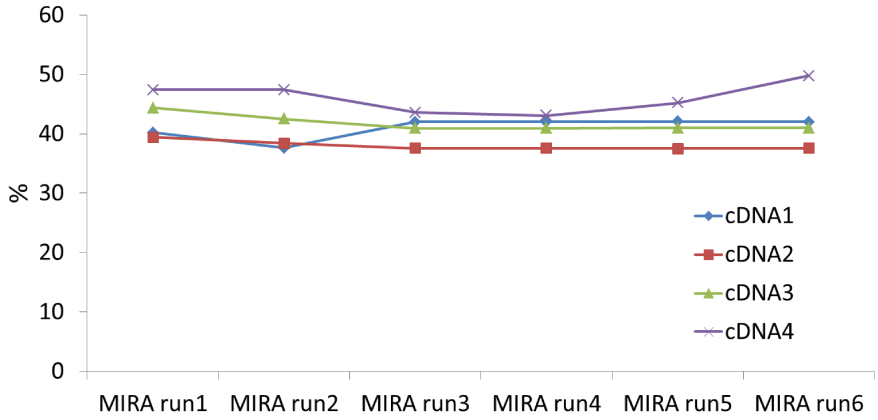

Figure 3.8. Setting effects on the redundancy index.

A. naccarii transcriptome assembly

After preprocessing, *A. naccarii* male (cDNA3) and female (cDNA4) reads were tagged accordingly to sex of origin and jointly assembled, thus allowing contigs to be classified for reads content as being composed by males only, by females only or by both sexes. Two iterative assembly cycles proved to be enough to reduce the majority of the internal redundancy caused by the heuristic nature of the assembly process, as shown in Figure 3.9.

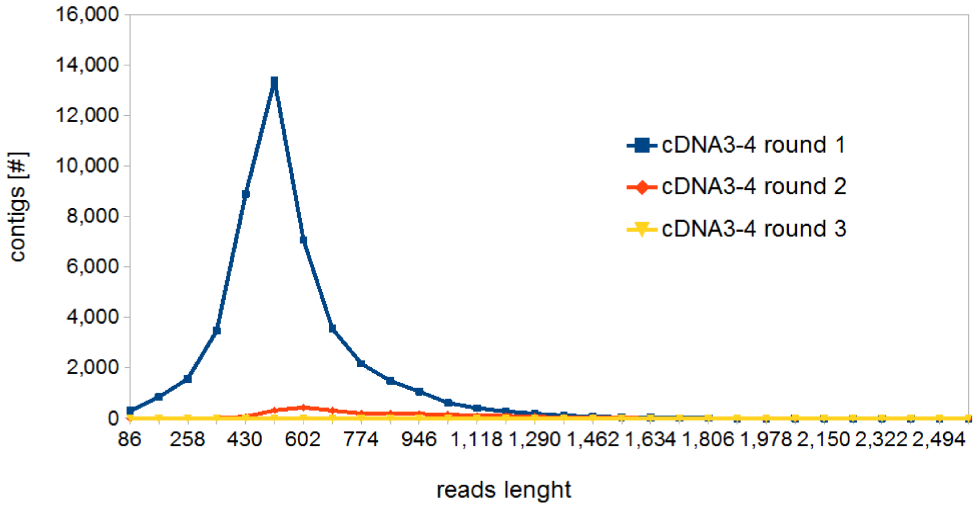

Figure 3.9. New contigs (metacontigs) produced after multiple assembly cycles. Two assembly rounds demonstrated enough to reduce internal sequence redundancy due to the heuristic of the assembly algorithms. As shown, a negligible number of metacondits were produced in the 3th round (19), despite the execution time required was considerable.

The first round of MIRA assembled 256,738 reads (77.43 % of the total cleaned reads) into 44,232 contigs and 16,593 singletons. The first assembly resulted in 27.62 Mbp of total consensus, composed of 60,825 sequences with an average length of 454.14 bp, average Phred quality of 39, a mean GC content of 38.47% and an average coverage of 4.22 reads. More details about the generated contigs and singletons are reported in Table 3.3. In the second round MIRA reassembled 6,242 contigs (14%) and 3,504 singletons (21%) from the previous assembly into 4,203 metacontigs, with an average coverage of 2.32 sequence/metacontig (Table 3.4).

	Round 1	Final assembly
Reads assembled (#)	256,738	256,738
Reads assembled (%)	77.43	77.43
Total contigs (#)	**44,232**	**42,193**
Contigs (%)	72.72	76.32
Total contigs length (Mb)	22.64	21.87
Average contigs length (bp)	511.89	518.29
Average contigs GC content (%)	38.49	38.83
Average contigs quality (Phred)	43	43
Average contigs coverage (bp/position)	3.2	4.09
Total singletons (#)	**16,593**	**13,089**
Singletons (%)	27.28	23.68
Total singleton length (Mb)	4.98	3.91
Average singleton length (bp)	300.19	298.55
Average singleton GC content (%)	38.39	38.46
Average singleton quality (phred)	28	28

Table 3.3. Contigs and singletons summary statistics for first and final *A naccarii* assemblies by MIRA.

Features	Round 2
Total metacontigs (#)	**4,203**
Reassembled contigs (#)	6,242
Percentage of reassembled contigs	14.11
Reassembled singletons (#)	3,504
Percentage of reassembled singletons	21.12
Total consensus metacontig (Mb)	2.95
Metacontig average length (bp)	700.94
Percentage of metacontig average GC content	38.83
Metacontig average consensus quality (Phred)	45
Metacontig average coverage (bp/position)	1.66

Table 3.4. Metacontigs summary statistics for second round *A. naccarii* assembly by MIRA.

Finally the two assembly runs were merged giving a total of 55,282 sequences, 42,193 contigs plus metacontigs (21.87 Mbp) and 13,089 singletons (3.91 Mbp). This resulted in a 9.11% sequence reduction compared to the first assembly as clearly illustrated by Figure 3.10. Overall, the sequences of this final dataset were characterized by a mean length of 466 bp, an average Phred quality of 40 and a mean coverage of 4.64 reads. GC content remained the same as in the first assembly (details relating to contigs and singletons are shown in Table 3.3). Changes in length and quality distribution of contigs from the first to

the second round assembly are shown in Figure 3.11 and Figure 3.12 respectively.

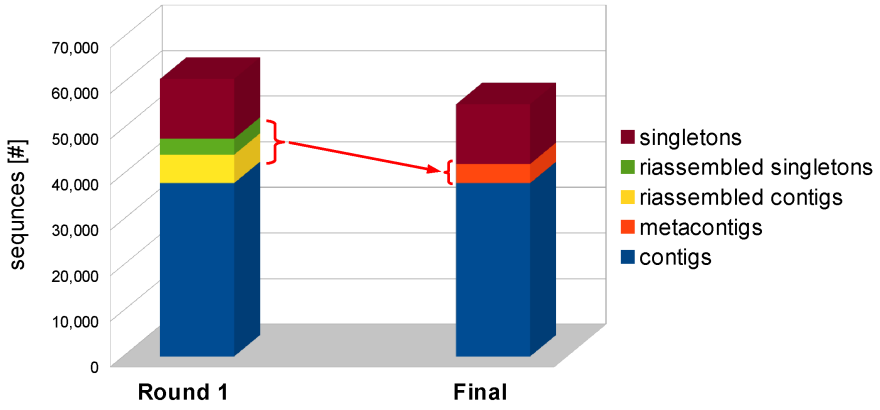

Figure 3.10. Redundancy reduction after two assembly rounds with MIRA for *A. naccarii* data. Graphical representation of the contigs and singletons built in the first assembly round, which were re-assembled as metacontigs in the second round, and then joined to get the final assembly.

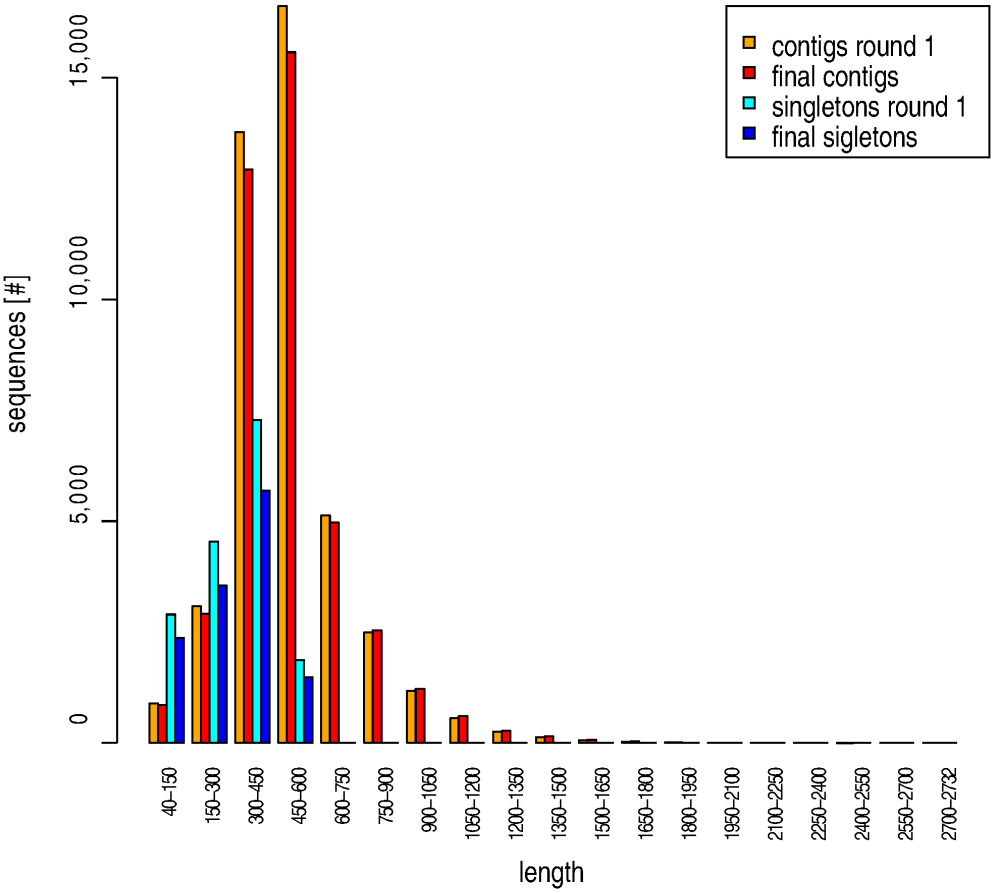

Figure 3.11. Distribution of contig- and singleton- lengths for *A. naccarii* first round and final assemblies. While the average quality of singletons remains between 15 and 40, the average quality of assembled contigs rises to 88.

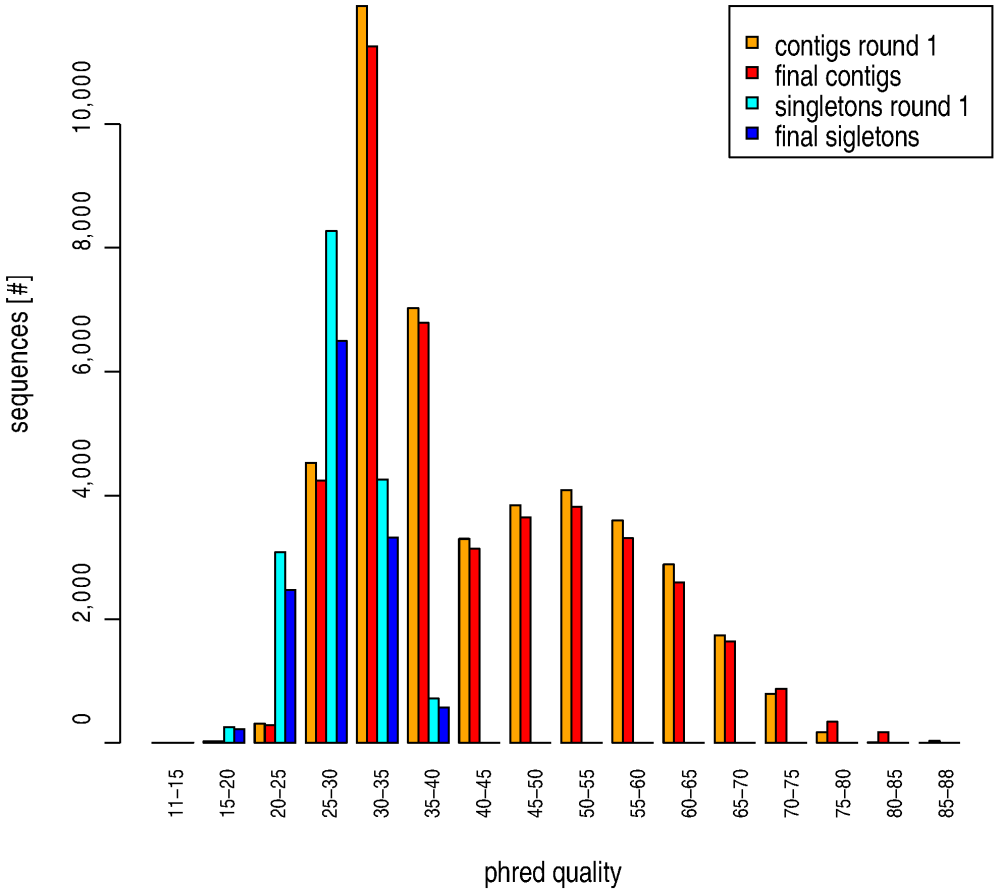

Figure 3.12. Distribution of contigs' and singletons' average quality for *A. naccarii* first round and final assemblies. The figure shows how the number of singletons and contigs resulting from the first assembly (largest contig 2732, N50 contig size 489, N90 contig size 324, N95 contig size 258), is reduced in the final set.

After assembly, all reads of origin were aligned against belonging contigs and metacontigs, obtaining a multiple alignment for each of them. The distribution of the average coverage observed in the contigs and metacontigs from the first and final assemblies are reported in Figure 3.13. Pair-wise relationships between sequence length, number of reads per contig and average sequence quality after the two assemblies are plotted in APPENDIX A. All alignments are provided within a *Anaccarii* database, described in section 3.8.

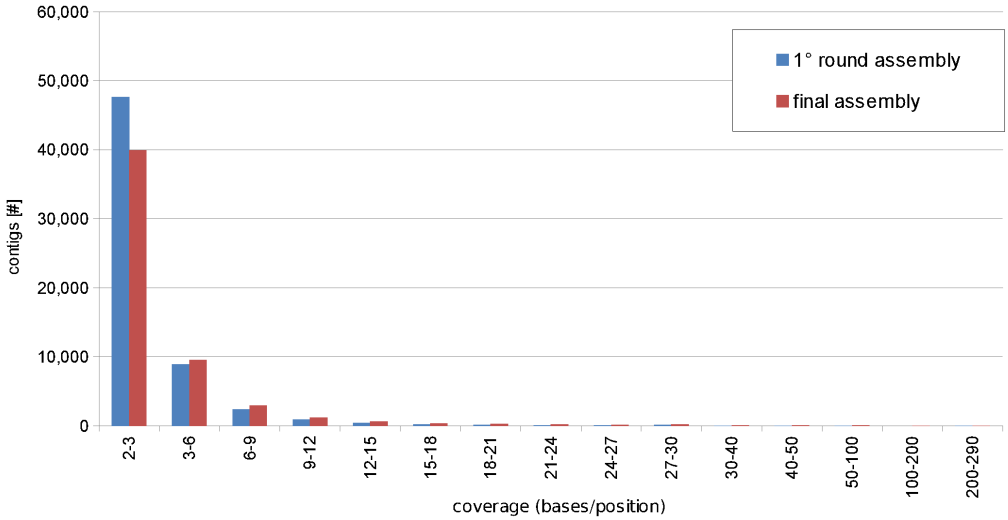

Figure 3.13. Mean contigs coverage distribution for the first and final *A. naccarii* assemblies. The average coverage of the contigs is quite low. As shown on the graph, about 61% of contigs have per base average coverage up to 3, while 93% have per base coverage up to 9. This may be due to the high ploidy in *A. naccarii*, believed to be tetraploid. Thus, the numerous alleles present, which are kept apart by MIRA, were sequenced to low coverage.

A. stellatus transcriptome assembly

Cleaned sequences were labelled according to the library of origin and then assembled employing two rounds of iterative assembly as described in the section methods . In the first round of MIRA, 73.14% of the total cleaned reads (258,661) were assembled in 48,111 contigs of 517 bp average length, and 2.89 base/position average coverage. Twenty thousand, two hundred and three singletons were also selected. A subset of 6,029 contigs and 3,821 singletons were re-assembled in the second round giving 4,522 new contigs (Table 3.5). The final assembly, resulting from the union of the two runs, was composed of 46,604 contigs (24.39 Mbp) and 17,002 singletons (5.14 Mbp). The total number of sequences was 7.73% lower than the total obtained from the first round (Figure 3.14). The average length and coverage of the final contigs increased slightly compared to the first round of assembly to 523 bp and 3.55 base/position as shown in Table 3.6. The length distribution of the contigs and singletons from the first and final assembly rounds is reported in Figure 3.15 while the distribution of average qualities is shown in Figure 3.16 Mean per-base coverage distribution of contigs is plotted in Figure 3.17. Pair-wise relationships between main properties of contigs are summarized in APPENDIX A. From here on, we will no longer make any distinction between contigs and metacontigs in this dissertation, and both will be indicated simply as contigs.

Features	Round 2
Total metacontigs (#)	4,522
Reassembled contigs (#)	6,029
Reassembled contigs (%)	12.53
Reassembled singletons (#)	3,821
Reassembled singletons (%)	18.35
Total consensus metacontig (Mb)	3.189092
Metacontig average length (bp)	705.239
Metacontig average GC content (%)	37.7958
Metacontig average consensus quality (Phred)	43
Metacontig average coverage (bp/position)	1.54368

Table 3.5. Contigs and singletons summary statistics for first and final *A. stellatus* assemblies by MIRA.

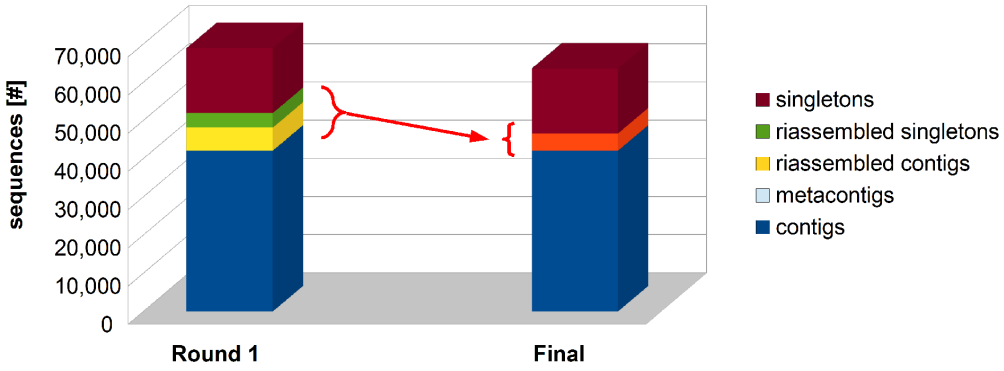

Figure 3.14. Redundancy reduction after two assembly rounds with MIRA for *A. stellatus* data.

Features	Round 1	Final assembly
Reads assembled (#)	258,661	258,661
Reads assembled (%)	73.14	73.14
Total contigs (#)	48,111	46,604
Contigs (%)	69.79	73.27
Total contigs length (Mb)	24.88	24.39
Average contigs length (bp)	517.169	523.302
Average contigs GC content (%)	37.94	37.96
Average contigs quality (Phred)	43.523	43.8934
Average contigs coverage (bp/position)	2.89	3.55
Total singletons (#)	20,823	17,002
Singletons (%)	30.20	26.73
Total singletons length (Mb)	6.27	5.14
Average singleton length (bp)	301.04	302.40
Average singleton GC content (%)	38.07	38.29
Average singleton quality (phred)	28.90	28.92

Table 3.6. Metacontigs summary statistics for second round *A. stellatus* assembly by MIRA.

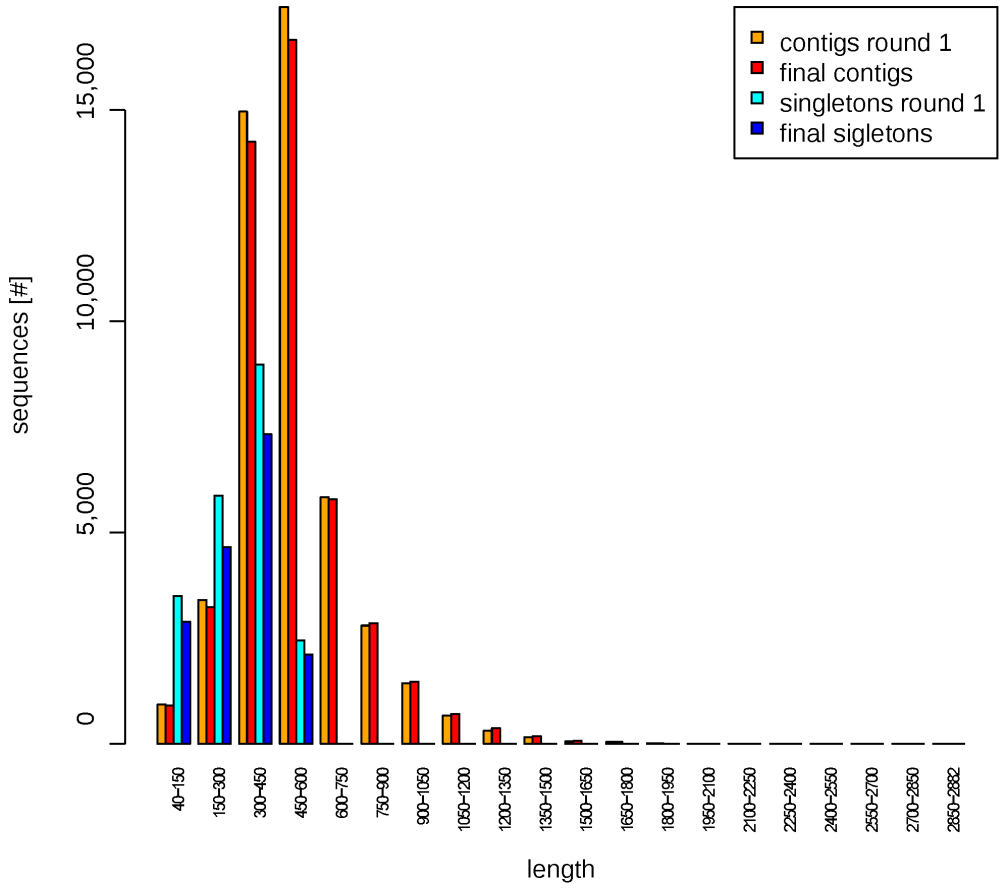

Figure 3.15. Distribution of contig and singleton lengths for *A. stellatus* first round and final unified assemblies.

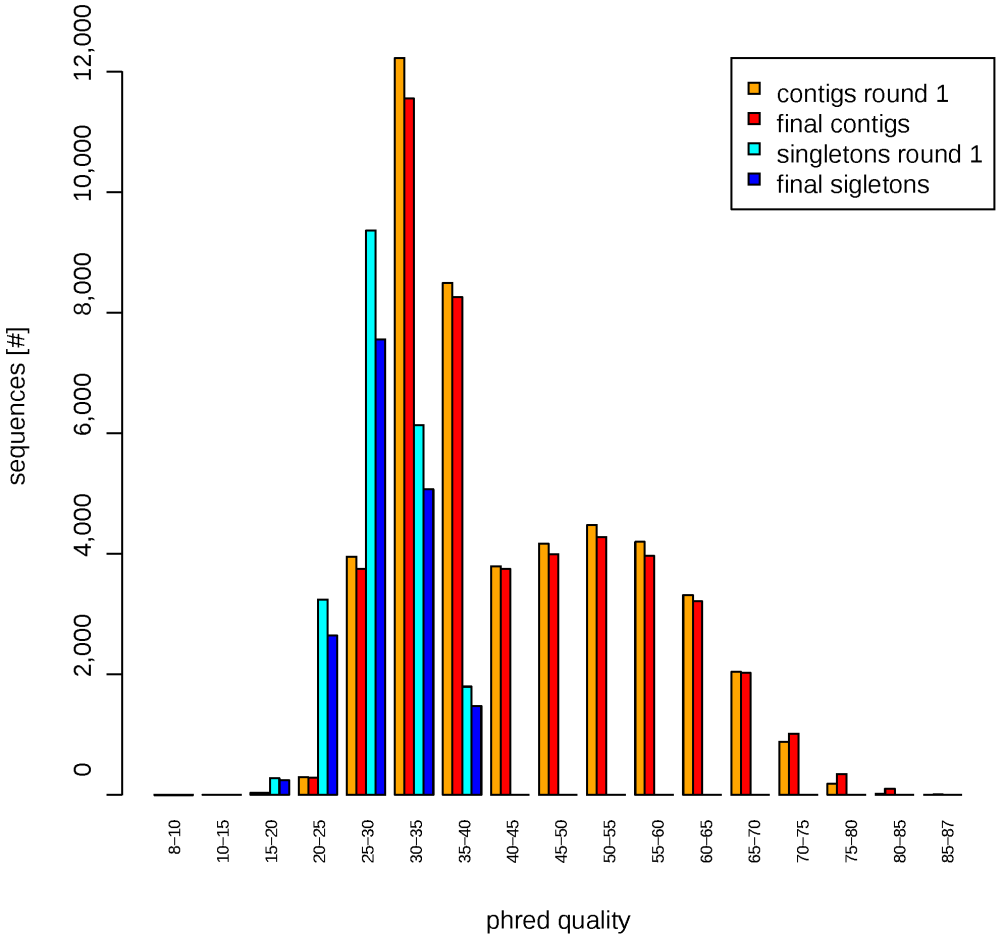

Figure 3.16. Distribution of contig and singleton average quality for *A. stellatus* first round and final unified assemblies.

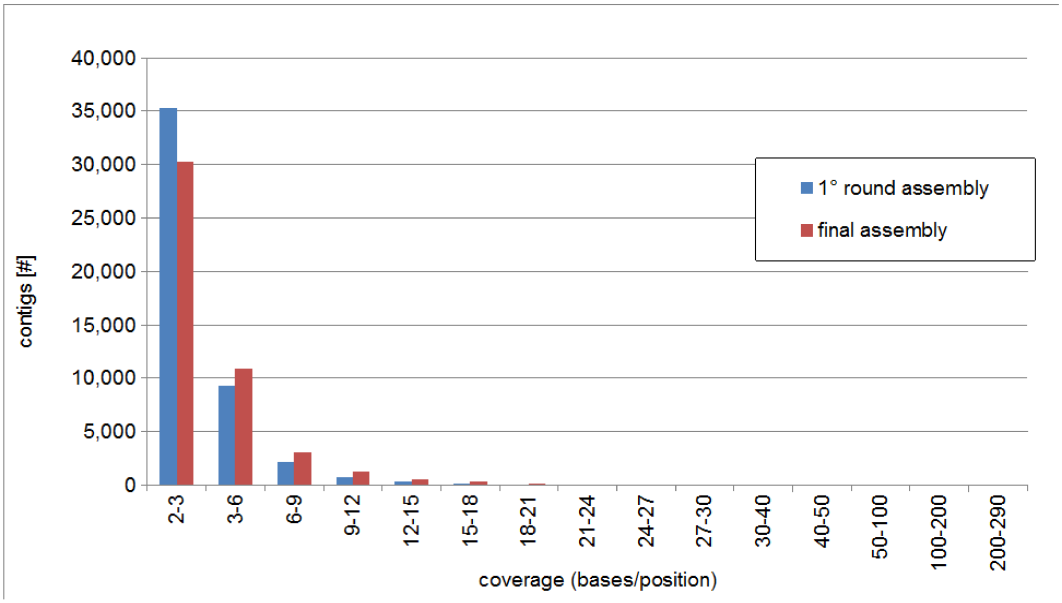

Figure 3.17. Mean contig coverage distribution for first and final assemblies (*A. stellatus*).

3.2 Estimation of sequencing completeness

Sequencing completeness of the A. naccarii cDNA libraries

The extrapolation from the hyperbolic model, fitted onto the average points, obtained by the 10 replications of sampling and reference transcript identification, using reads from the male library only, showed that 7,293 different transcripts were potentially identifiable in the *Danio* cDNA set (asymptote "*a*" of the model function). The 6,043 different transcripts actually identified using all reads represents 83% of the theoretical maximum. The angular coefficient calculated at final read count was 0.157. Using reads from the female library only, the number of transcripts actually identified was 5,989, which, compared to the 7,176 maximum transcripts identifiable at infinite sequencing, represents 83% of the total. The slope at the final read count was 0.145. Finally, putting together reads from both libraries, the model-based extrapolation denoted 8,262 different transcripts potentially identifiable, and the 7,286 actually identified represents 88%. The slope at maximum read count was 0.140. The three extrapolated curves are shown in Figure 3.18. Supplementary saturation curves, were also built using total reads and taking the Ensembl cDNA datasets for the RS-list species as references. These curves are shown in Figure 3.19.

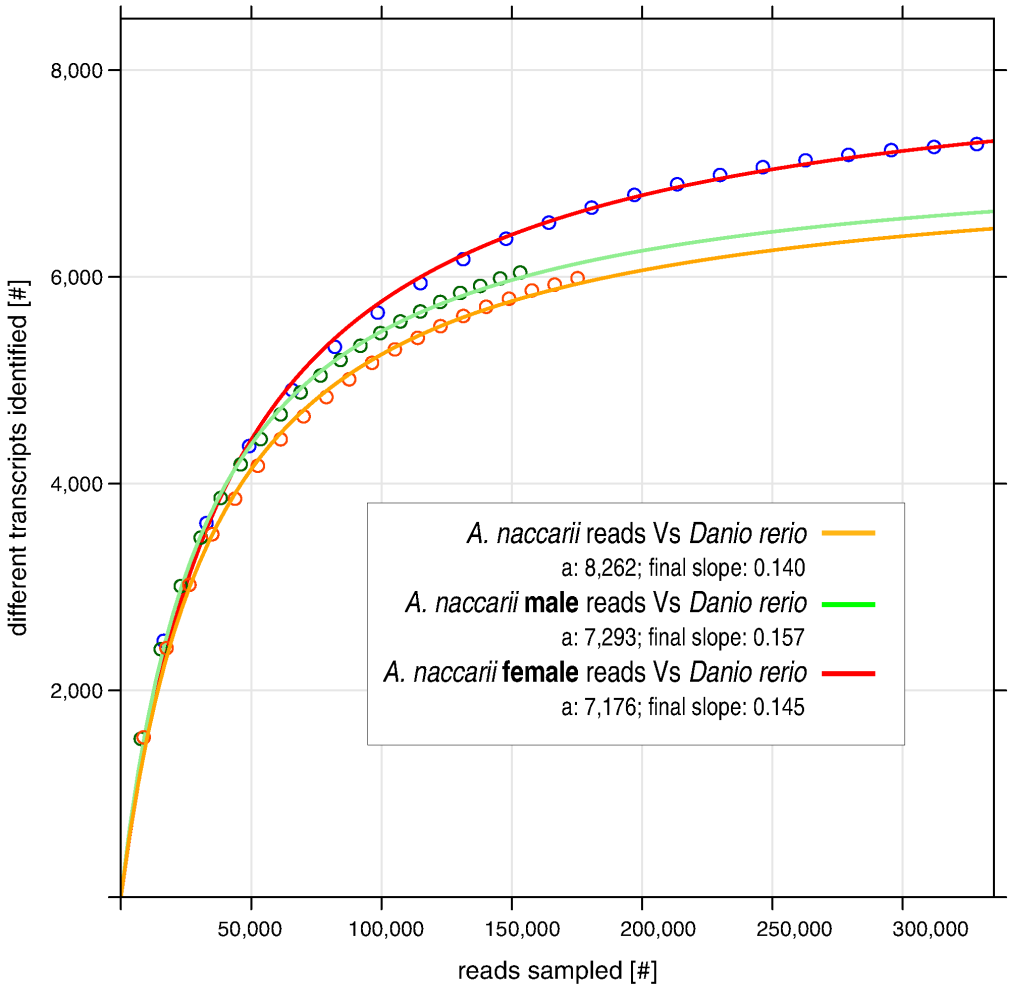

Figure 3.18. Saturation curve for male, female and both *A. naccarii* cDNA libraries.

Read subsets of increasing sample size were randomly extracted from total pool in the library. For each subset, contigs in which reads where assembled were identified. Each contigs pool was used to identify *Danio rerio* cDNAs (TBLASTX 2.2.25+ eval 1e-03). Re-sampling and identification process was repeated 10 times for each sample size. A mean value and a confidence interval for the number of identified Danio cDNAs was calculated for each sample size. Hyperbolic model y = (ax)/(b+x) was fitted on points given by sample size versus average cDNAs hit so that model parameters "*a*" and "*b*" were estimated. The legend shows estimated parameter values obtained by fitting the hyperbolic model on the data. As can be seen, the curves from the single libraries retain the same trend and the difference is mostly due to the different number of reads in each library.

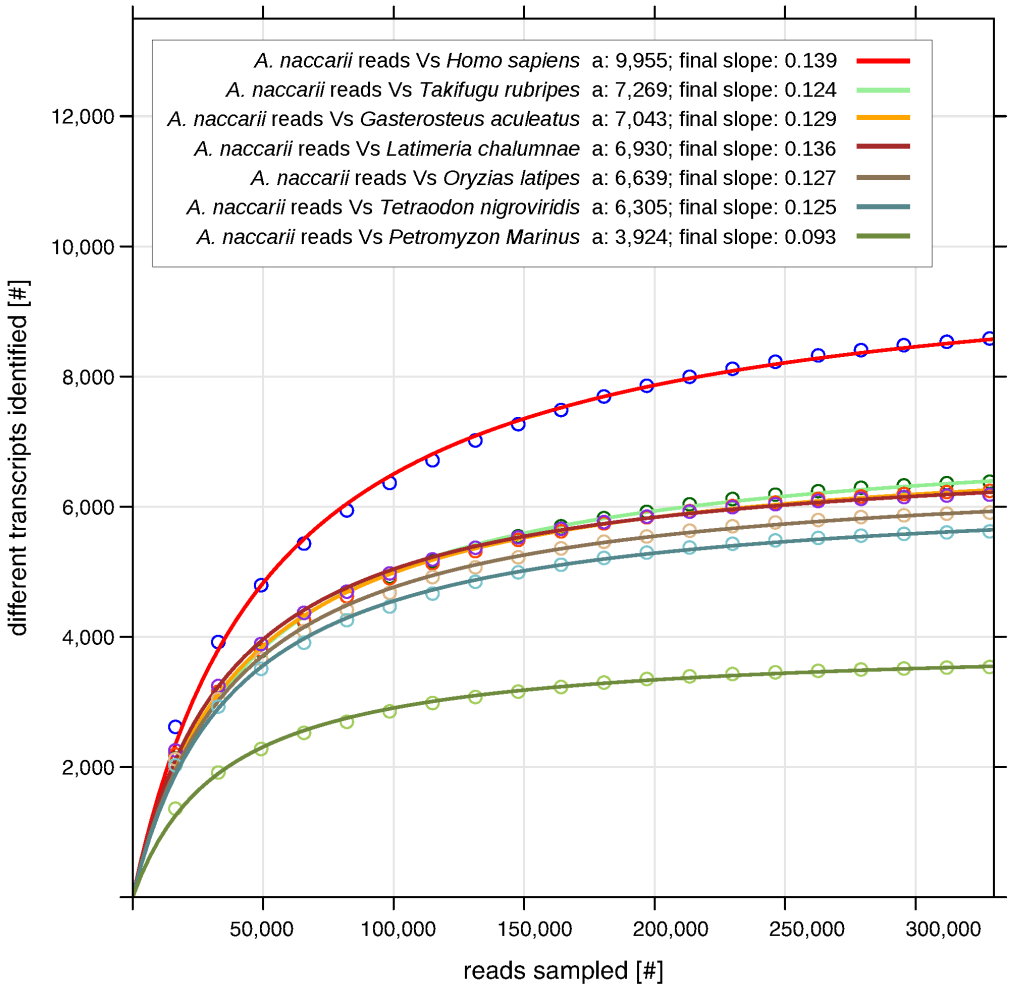

Figure 3.19. Saturation curves plot for joint *A. naccarii* cDNA libraries against cDNA sets from other fishes.

Were constructed several saturation curves, starting from total reads from the two libraries and using different sets of cDNA as a references. cDNA sets used are derived from all transcripts from Ensembl 66 for species in RS-list. The estimated parameters of the curves are reported in the legend.

Sequencing completeness of the A. stellatus cDNA libraries

The saturation curve obtained using reads from the *A. stellatus* male library only achieved the asymptote at the theoretical value of 8,462 transcripts of the *Danio* dataset. The number of transcripts actually identified using all reads was 6,795. This suggests that the sequencing process tagged about 80% of all sequences present in the physical male library. The slope of the curve at maximum sample size was 0.137.

The predicted maximum number of different *Danio* transcripts identifiable for the female library was 8,055 compared to 6,489 actually identified. This means a sequencing

coverage of 81%. The slope at the end of the curve was 0.142. Using reads from both libraries, whole coverage was estimated to be 87% as the asymptote of the curve was 9,045 while 7,857 ESTs had hits against the reference database. The slope dropped slightly to 0.129. The three plots can be shown in Figure 3.20.

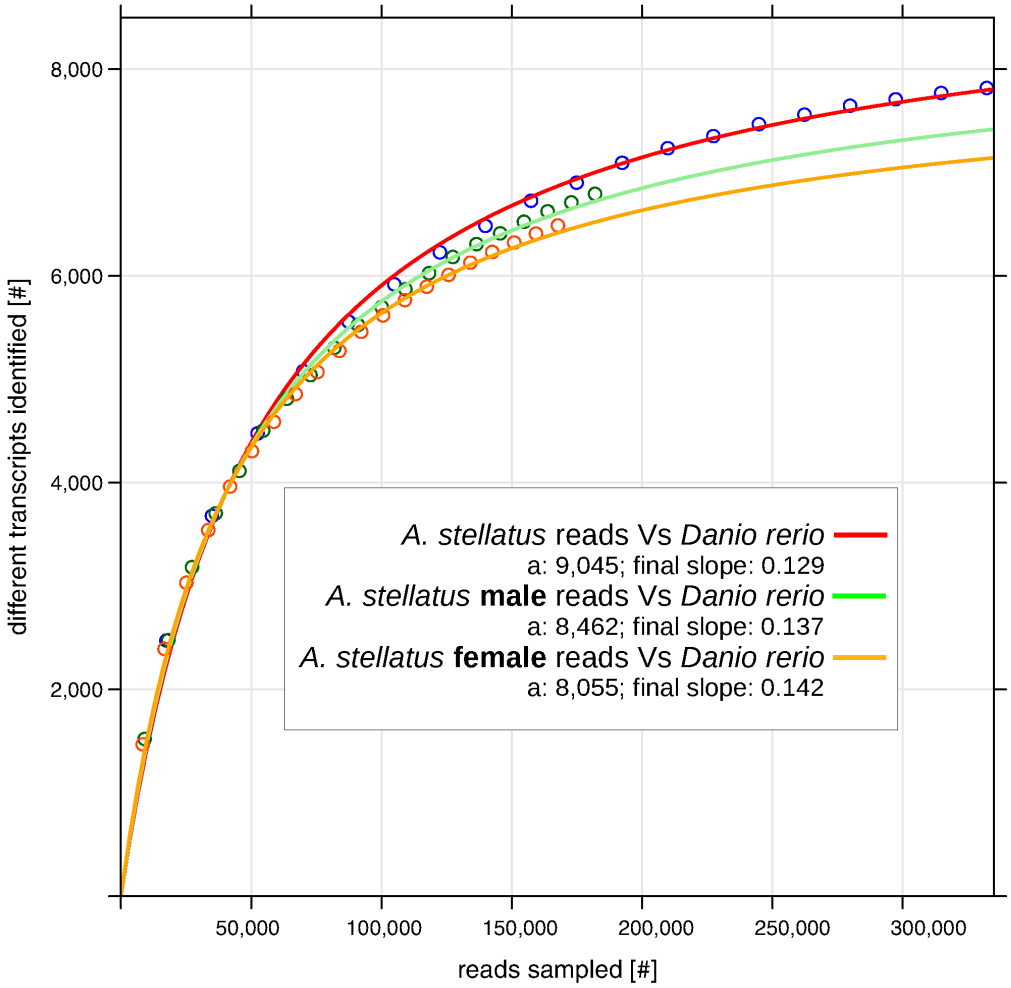

Figure 3.20. Saturation curve for male, female and both *A. stellatus* cDNA libraries.

Further saturation curves, using total reads, were built taking the Ensembl cDNA datasets for the RS-list species as references. These curves are shown in Figure 3.21.

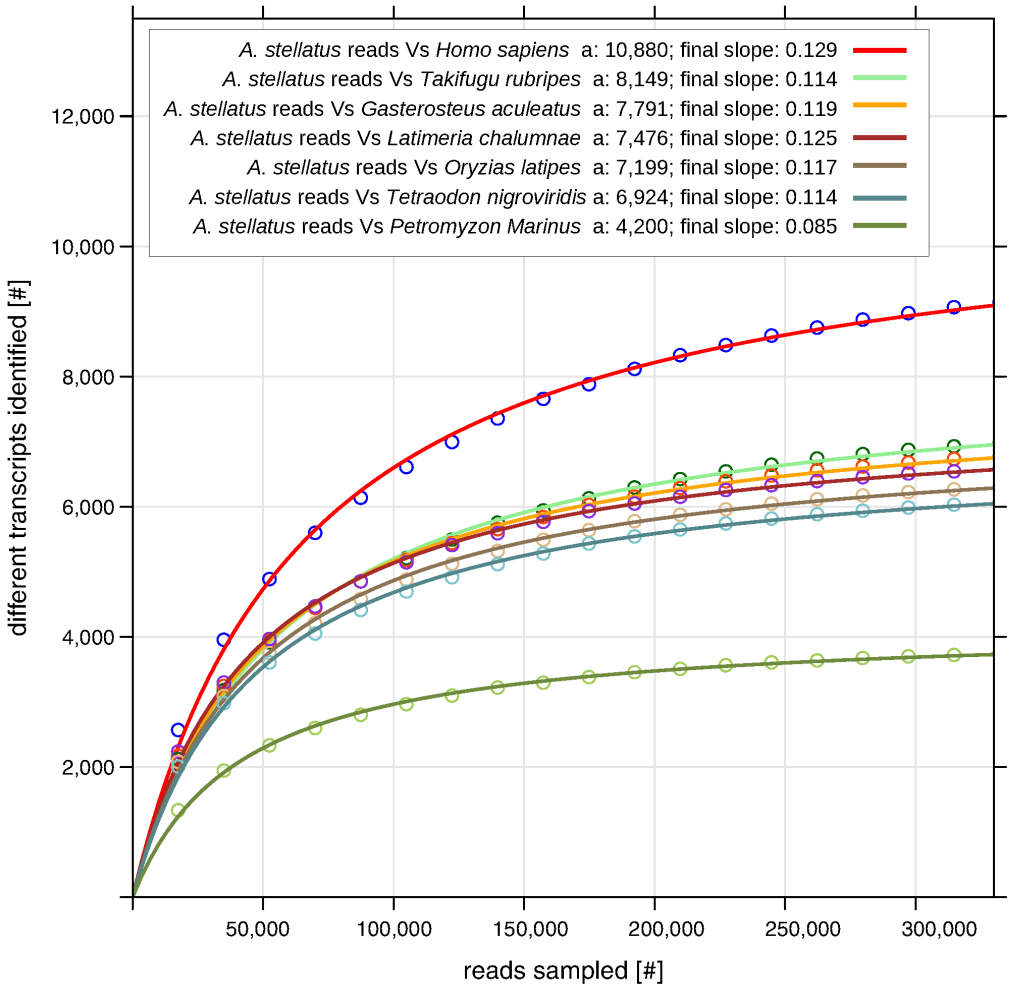

Figure 3.21. Saturation curves plot for joint *A. stellatus* cDNA libraries against cDNA sets from other fishes.

3.3 Transcriptomes completeness estimation

A. naccarii

First, I separated 17,399 cDNA3-specific contigs from the male library (31% of the total) and 17,523 cDNA4-specific contigs from the female library (32% of the total). The direct subtraction between the two groups of library-specific contigs isolated 394 contigs showing mutual alignments from each fraction. The indirect subtraction identified 41 cDNA3-specific and 38 cDNA4-specific contigs, that aligned on 85 common subjects. Finally, using NCBI nr as the common database, I identified an additional 13 cDNA3-specific and 12 cDNA4-specific contigs which map onto the same 10 protein sequences. After all subtractions, 16,951 cDNA3-specific and 17,079 cDNA4-specific contigs remained, that may represent potentially sex-distinctive transcripts.

With the Rcapture R package I estimated the transcripts population size to be 68,904 with a standard error of 210. This means that it was probably sequenced about 80% of the total transcripts in the two tissues of *A. naccarii*.

As explaned in methods, the transcriptome completeness was also evaluated, by searching for constitutively-expressed mitochondrial genes. Of the 37 genes part of the white sturgeon *Acipenser transmontanus* mitochondrial genome, 12 polypeptide coding genes, both 12S and 16S rRNA, and 6 tRNAs out of 22, were found in our assembly. Concerning to polypeptide coding genes, only the gene for ATPase subunit 8 was missing. Contigs that aligned with the sequences of these genes showed between 93 and 100% identity. The number of different contigs that aligned is proportional to the transcription rate of each gene, for example, the mt-mRNAs for COIII, COII and ND2 (notoriously more abundant) aligned with the largest number of different contigs. The even more abundant mt-rRNAs 16S and 12S were, respectively, identified by 16 and 5 different contigs. Curiously, I identified 6 mt-tRNAs in the assembly, with sequence identity close to 100%, matched alongside the region of contigs that aligned with other mt-mRNA genes.

A. stellatus

Assembled contigs were first separated according to their reads composition in: 19,733 cDNA1-specific contigs (from the male library, 31% of the total contigs) and 15,791 cDNA2-specific contigs (from the female library, 25% of the total contigs). In the direct subtraction process, 276 contigs from each specific fraction were taken out and added to the counts of the common fraction. In the process of indirect subtraction against cDNA databases, 32 and 33 contigs were removed from the count of the cDNA1-specific and cDNA2-specific fractions respectively and added to the common one. These contigs showed significant alignments against one or more common best hits, 60 in total. An additional 15 contigs were subtracted from each specific fraction, after the indirect subtraction with nr as the common database. In the end, 19,410 cDNA1-specific and 15,467 cDNA2-specific contigs remained. With these values, the total transcripts population size estimated by the Rcapture package resulted: 74,190±165. This indicated that it was possible to tag about 86% of the total transcripts expressed in both tissues of *A. stellatus* here analysed. Even taking into account the possibility of an overestimation, this value is close to the 80% completeness estimation for *A. naccarii* cDNA sequence libraries.

Of the comprehensive 37 genes present in the *A. stellatus* mitochondrial genome, 25

were identified in our EST collection: 11 mt-tRNA, the ribosomal genes 12S and 16S, and all but one polypeptide coding gene. Again, the ATPase subunit 8 was missing. All alignment identities fell between 95 and 100% identity. As expected, genes for 16 and 12S were detected in the largest number of contigs, followed by polypeptide coding genes ND4 and ND5, each identified in 7 contigs. Again, 3 contigs were found to contain mt-tRNA genes near polypeptide coding genes which again may represent not-yet-processed polycistronic precursors.

3.4 Functional annotation

A. naccarii transcriptome annotation

BLAST against sequences available from the genus Acipenser

The comparison of *A. naccarii* sequences, with 6,088 ESTs for the genus *Acipenser* already available in GenBank, revealed 8,804 *A. naccarii* contigs (15.93%) matching 2,047 different subjects (33.62%). The limited percentage of matching sequences can probably be ascribed to the different tissues of origin: gonad and brain in the Adriatic sturgeon, and mainly pituitary gland, skin and spleen in the reference database.

BLASTX against the main protein sequence databases

The comparison of contigs and singletons to the NCBI non-redundant protein database (nr) using BLASTX, came out with 9,850 contigs and 2,339 singletons (22.05% of total sequences) matching 9,433 different known or predicted proteins. Every query, aligned with a mean alignment length of 87.25 aminoacids. Aligned regions covered on average 48.75% of contig lengths and 30.19% of best hit protein lengths. Mean length, quality and GC content of sequences with or without significant BLASTX hits against nr showed significant differences as shown in Table 3.7. The taxonomic classification of hits from the nr database, by species, is represented in Figure 3.22.

Means	Sequences with hit	Sequences without hit	t-test p-value
length (bp)	569.62	437.03	
quality (Phred)	43.80	38.42	< 2.20E-016
GC (%)	44.13	36.85	

Table 3.7. Two samples t-test to on the average length, GC content and quality of the sequences with and without significant BLAST hit (e-val < 1e-03) against nr database.

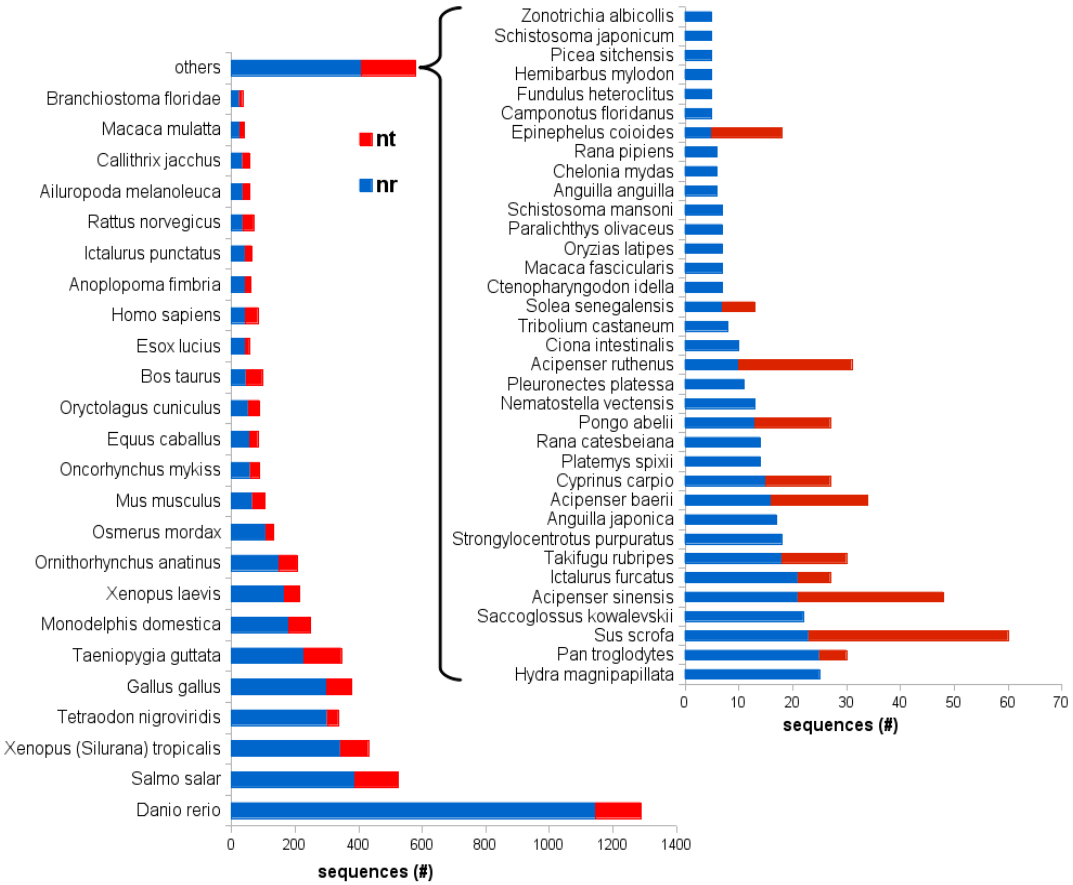

Figure 3.22. **Taxonomic classification of *A. naccarii* contig annotations.** Assignment of annotations obtained from BLASTX and BLASTN comparisons (e-value 1e-03) of contigs against nr and nt databases to different species was performed with MEGAN 4, based on the absolute best BLAST hits. Contigs with multiple best BLAST hits were excluded from the count as they couldn't be assigned to a particular species with assurance. The bar chart shows contigs annotated with the 24 more-represented species in annotations from nr. The contribution of annotations from nt, for the same species, is marked in red. “Others” includes the 34 species less represented in nt annotations.

BLASTX search in Swiss-Prot section of the UniProtKB database, identified 11,088 transcripts (20.06%) with significant matches against 7,111 different well-annotated proteins. Only 179 transcripts not previously annotated emerged, probably due to the different ages of the two databases.

BLASTN against the main nucleotide database

The BLASTN search against the NCBI nucleotide database (nt) identified significant similarity for 10,195 transcripts (18.44%) with 4,509 different subjects. Among sequences with a significant match against nt, 5,366 had not previous matches against nr and Swiss-Prot databases. Considering all the BLAST searches performed so far, a total of 17,734 ESTs obtained at least one hit, representing 32% of the Adriatic sturgeon transcriptome.

Evaluation of the unannotated fraction

A total of 43,093 non-redundant transcripts remained unannotated after the BLAST search against the nr database. ORF prediction showed that 41,935 of these sequences (97.31%) contain a putative open reading frame of 161 bp mean length, suggesting a coding role for these transcripts.

Evolutionary comparison with other fishes

The non-redundant contigs of the two *A. naccarii* libraries were compared to Ensembl release 66 complete cDNA sets for the species of the RS-list. TBLASTX and BLASTX best hit results are collected in Table 3.8 and Table 3.9 respectively. The two non-teleost species on the RS-list, the Sea Lamprey and the Coelacanth, share a higher fraction of genes 33.49% and 30.57% respectively as also confirmed at the protein level (31.97% and 29.24%). Zebrafish, seems to share fewer genes (22.89% through transcripts and 22.42% through proteins) with *A. naccarii* than do other teleosts. Moreover, the fraction of the *A. naccarii* matching ESTs is comparable to other species.

Evaluation of the non-coding RNA component

NcRNA are implicated in every step of gene expression. To discover and annotate potential non-coding RNAs in our transcriptome (miRNA, rRNA, MtrRNA, snoRNA, lncRNA), I searched for genes corresponding to non-coding RNA from genomes of the fish species described above, using BLASTN. Alignment results are collected in Table 3.10. The highest number of alignments was found against miRNA sequences from the 4 teleosts, in particular in Medaka, whose 9 miRNA were found to be homologous in sturgeon. Mitochondrial and ribosomal RNA were next in abundance. Surprisingly, 11 rRNA pseudogenes from humans were found to be homologous in *A. naccarii.* The alignment method used here can underestimate the number of ncRNA detected as different types of ncRNA have different degrees of sequence conservation between species, with miRNA and snoRNA usually well-conserved while longer-functional ncRNA are not (Pang et al. 2006). Moreover, lncRNA elements tend to maintain a consensus secondary structure through compensatory base mutations and, therefore, are difficult to detect by sequence alignments alone (Nawrocki et al. 2009).

	lamprey	coelacanth	danio	stickleback	medaka	fugu	tetraodon	human
reference cDNAs (#)	11,476	21,958	48,636	27,628	24,662	48,003	23,265	180,654
reference genes (#)	10,449	19,174	27,948	20,839	19,687	18,685	19,749	47,266
A. naccarii ESTs with hit on reference cDNAs (#)	7,431	12,428	13,068	11,528	11,270	11,143	10,886	12,740
***A. naccarii* ESTs with hit on reference cDNAs (%)**	**13.44**	**22.48**	**23.64**	**20.85**	**20.39**	**20.16**	**19.69**	**23.05**
reference genes identified (#)	3,499	5,862	6,396	5,710	5,565	5,447	5,429	6,116
reference genes identified (%)	**33.49**	**30.57**	**22.89**	**27.4**	**28.27**	**29.15**	**27.49**	**12.94**

Table 3.8. TBLASTX best hit (e-val < 1e-03) of *A. naccarii* transcriptome against cDNA sequences from Ensembl database. Sequences from known-, novel- and pseudo-gene predictions, from Ensembl realise 66, were collected for the following species: *Petromyzon marinus*, *Latimeria chalumnae*, *Danio rerio*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Takifugu rubripes*, *Tetraodon nigroviridis* and *Homo sapiens*. *A. naccarii* transcriptome sequences were searched against each database. For each sequence, the best hit was annotated (subject).

	lamprey	coelacanth	danio	stickleback	medaka	fugu	tetraodon	human
reference proteins (#)	11,429	21,817	41,693	27,576	24,661	47,841	23,118	97,041
reference genes (#)	10,402	19,033	26,160	20,787	19,686	18,523	19,602	21,860
A. naccarii ESTs with hit on reference proteins (#)	7,052	11,264	11,530	11,010	10,816	10,766	10,359	11,102
***A. naccarii* ESTs with hit on reference proteins (%)**	**12.76**	**20.38**	**20.86**	**19.92**	**19.57**	**19.47**	**18.74**	**20.08**
reference genes identified (#)	3,326	5,565	5,865	5,526	5,403	5,346	5,283	5,478
reference genes identified (%)	**31.97**	**29.24**	**22.42**	**26.58**	**27.45**	**28.86**	**26.95**	**25.06**

Table 3.9. BLASTX best hit (e-val < 1e-03) of *A. naccarii* transcriptome against protein sequences from Ensembl database. Best hits from the alignment of *A. naccarii* transcriptome sequences against all translations from known-, novel- and pseudo-gene predictions in Ensembl release 66 for the different species considered in this work.

	lamprey	coelacanth	danio	stickleback	medaka	fugu	tetraodon	human
reference proteins (#)	11,429	21,817	41,693	27,576	24,661	47,841	23,118	97,041
reference genes (#)	10,402	19,033	26,160	20,787	19,686	18,523	19,602	21,860
A. naccarii ESTs with hit on reference proteins (#)	7,052	11,264	11,530	11,010	10,816	10,766	10,359	11,102
A. naccarii ESTs with hit on reference proteins (%)	12.76	20.38	20.86	19.92	19.57	19.47	18.74	20.08
reference genes identified (#)	3,326	5,565	5,865	5,526	5,403	5,346	5,283	5,478
reference genes identified (%)	**31.97**	**29.24**	**22.42**	**26.58**	**27.45**	**28.86**	**26.95**	**25.06**

Table 3.10. BLASTN best hit (e-val < 1e-03) of *A. naccarii* transcriptomes against non-coding RNA genes from Ensembl database. All non-coding RNA genes and pseudogenes in Ensembl realise 66 for the different species were searched against *A. naccarii* transcriptomes.

GO annotation

I started the GO annotation from the BLASTX results against nr. GO terms were retrieved from the association to best-hit for 10,036 (18.15%) of the overall 55,282 contigs. Protein domains and motif information were retrieved by InterProScan via Blast2GO and corresponding annotations were merged with already existent GO terms. A total of 29,671 contigs provided significant InterProScan information, with only 3,326 of them resulting in GO annotation. After merging, 6,344 unique GO terms (3,811 for biological process, 758 for cellular component, 1,775 for molecular function), were successfully transferred to 8,784 contigs (16%). As expected, the evidence code distribution shows an over-representation of electronic annotations (IEA), although other non-automatic codes, such as Inferred from Direct Assay (IDA) and inferred by mutant phenotype (IMP), were also

well represented (see bar-plot in Figure 3.24).

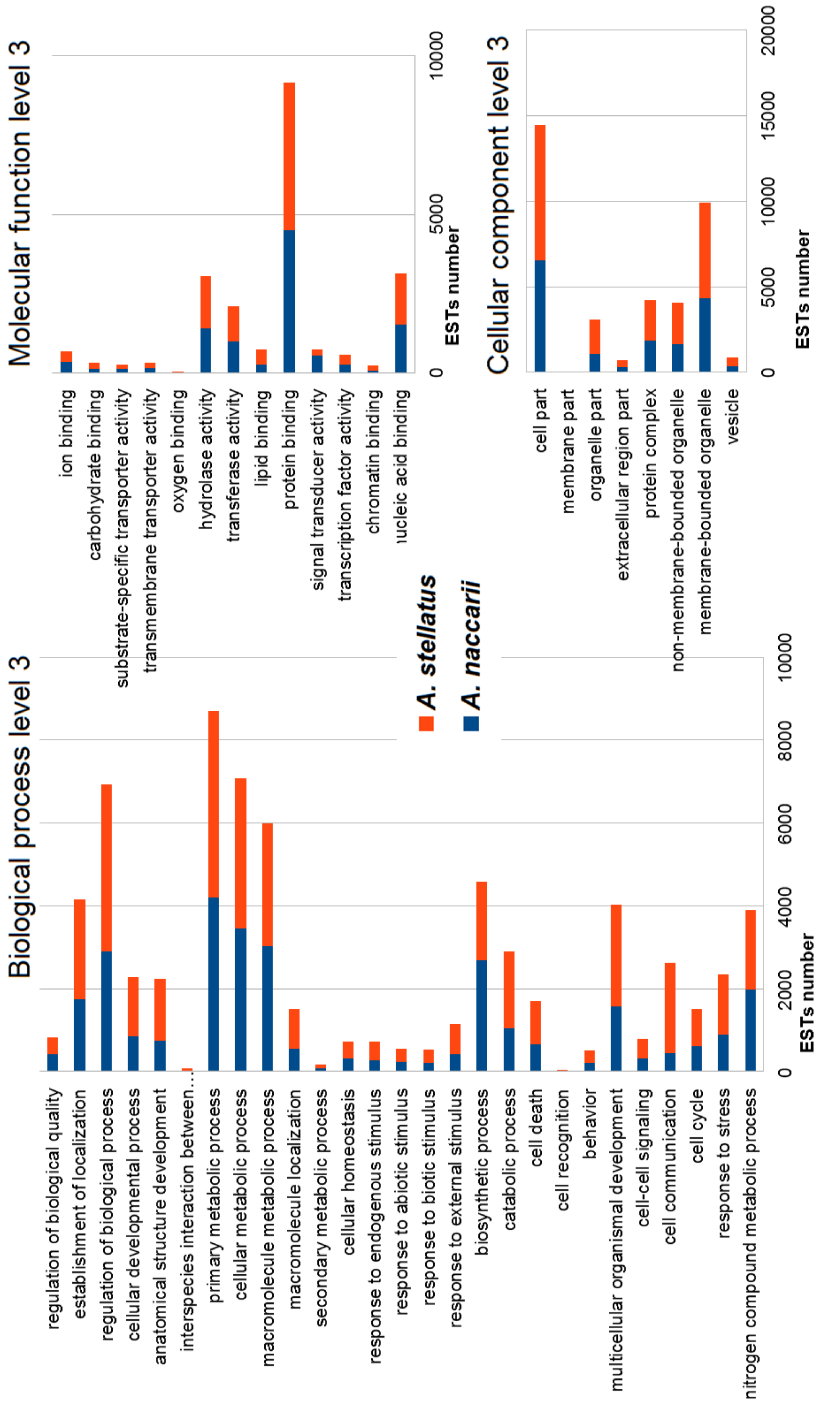

Figure 3.23. Distribution of Gene Ontology categories for *A. naccarii* and *A. stellatus* ESTs, across the tree domains. ESTs of both species were classified into different groups on the basis of generic GO-slim annotations. The bar-plot represents the categories corresponding to level 3 of the DAG graphs built for biological process, molecular function and cellular component domains. The categories in both species were combined and all but 4 resulted present in both species

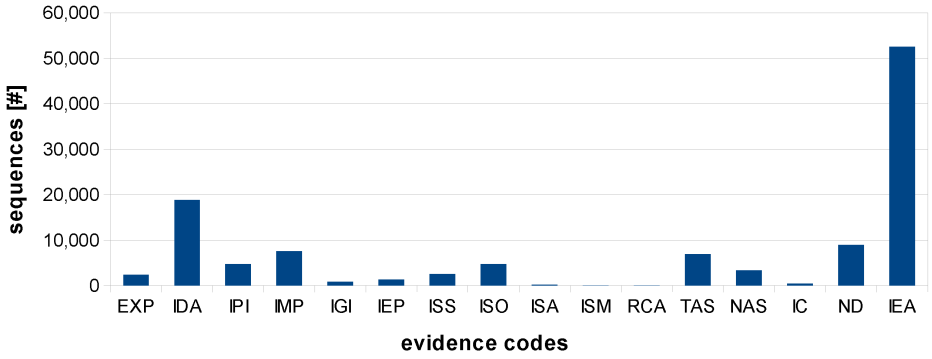

Figure 3.24. Evidence code distribution of the annotation of *A. naccarii* transcriptomes Abundance of sequences, for evidence code. Only the evidence codes assigned to at least one sequence are reported. EXP: Inferred from Experiment, IDA: Inferred from Direct Assay, IPI: Inferred from Physical Interaction, IMP: Inferred from Mutant Phenotype, IGI: Inferred from Genetic Interaction, IEP: Inferred from Expression, Pattern, ISS: Inferred from Sequence or Structural Similarity, ISO: Inferred from Sequence Orthology, ISA: Inferred from Sequence Alignment, ISM: Inferred from Sequence Model, RCA: inferred from Reviewed Computational Analysis, TAS: Traceable Author Statement, NAS: Non-traceable Author Statement, IC: Inferred by Curator, ND: No biological Data available.

A. naccarii ESTs were classified by GO-slims within the biological process, molecular function and cellular component domains and a Direct Acyclic Graph (DAG) of the ontologies was generated. Figure 3.23 shows the number of putative ESTs annotated with high-level GO terms by cutting the DAG graph at level 3 for each of the 3 domains. I also performed enzyme code (EC) annotation through Blast2GO for sequences with GO annotations and retrieved KEGG maps for the metabolic pathways in which they participate. In total 3,634 ESTs were annotated with 448 ECs that identify unique enzymes, participating in 116 different pathways (see APPENDIX B). The most populated pathways are “Purine metabolism” (map 00230) with 33 enzymes involved, “Arginine and proline metabolism” (map 00330) with 25 enzymes and Glycolysis/Gluconeogenesis (map 00010) with 22 enzymes.

A. stellatus transcriptome annotation

BLAST against sequences available from the genus *Acipenser*

By repeating the alignment of the *A. stellatus* ESTs against the sequences of the genus *Acipenser*, I found that 9,739 contigs (15.31%) aligned against 2,156 sequences (35.41%) of other sturgeon species. These values are nearly identical to the ones obtained for *A. naccarii* contigs.

BLASTX against the main protein sequence databases

The alignment against the NCBI nr protein database, using BLASTX, confirmed a match for about 20% of the total sequences in the *A. stellatus* assembly (10,389 contigs and 2,807 singletons), against 10,660 unique sequences in nr. The taxonomic classification of hits by MEGAN 4, is reported in Figure 3.25. The alignments against the protein sequences contained in Swiss-Prot, showed 11,827 transcripts (18.59%) with a match against 7,814 different proteins in this database. Again, a small number of 205 previously unmatched contigs found a hit in Swiss-Prot.

BLASTN against the main nucleotide database

The similarity search within the NCBI non-redundant nucleotide database found 13,200 *A. stellatus* ESTs (21%) that aligned against 4,820 different nucleotide sequences in nt. Of all these ESTs, 7,947 had not previously been matched against nr, nor against Swiss-Prot databases. By adding all the *A. stellatus* sequences found so far with a match, I was able to annotate about 34% of the total assembly. This value is absolutely comparable with the annotated fraction of *A. naccarii* assembly (32%).

Evaluation of the unannotated fraction

I predicted the presence of ORFs inside the 50,410 contigs, which reported no matches against protein sequences in the nr database. Within 49,106 of these (97%), a putative ORF was predicted. The ORFs identified showed an average length of 160 bp.

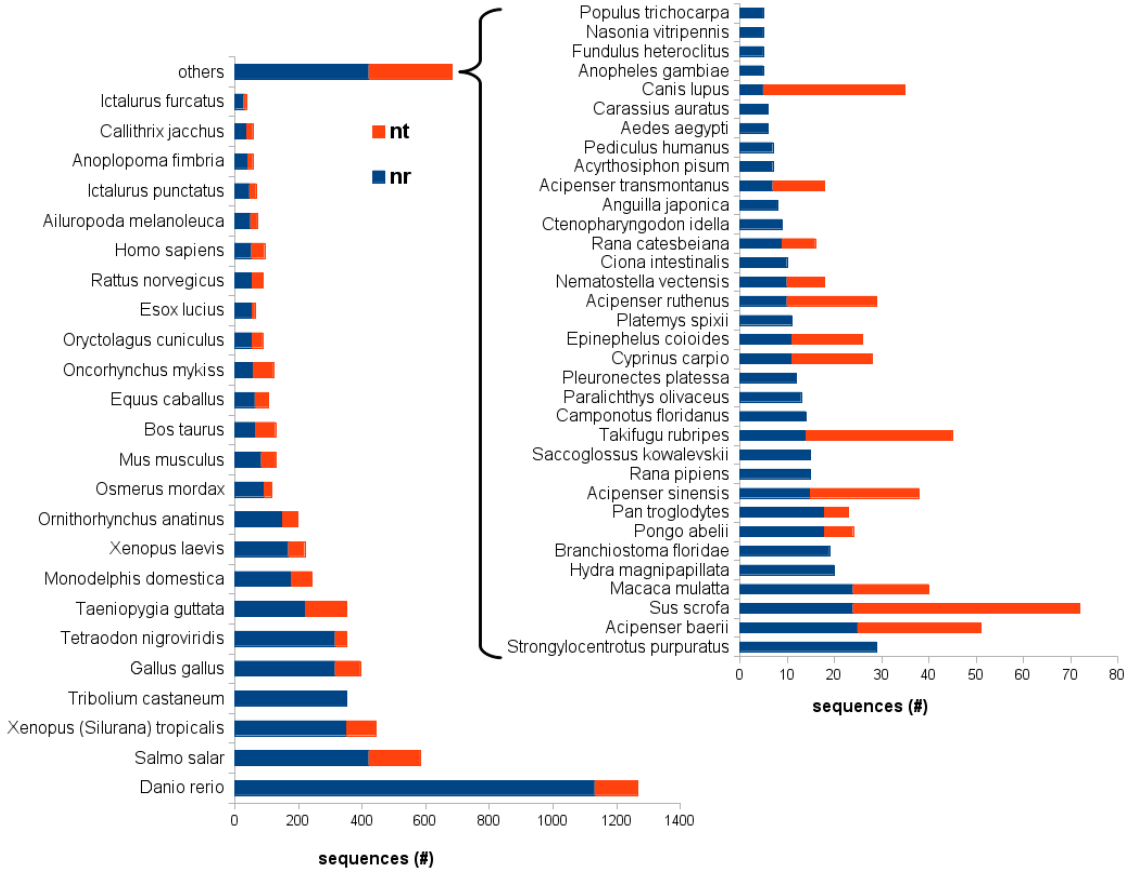

Figure 3.25. Taxonomic classification of *A. stellatus* contig annotations.

Evolutionary comparison with other fishes

TBLASTX and BLASTX best hit results are collected in Table 3.11 and Table 3.12 respectively.

Evaluation of the non-coding RNA component

The search for non-coding RNA genes in the *A. stellatus* transcriptome confirmed the highest number of alignments with genes encoding miRNAs from Medaka, with 5 unique genes found. Again, 12 different rRNA pseudogenes from H*omo sapiens* were discovered. Statistics of all matches are reported in Table 3.13.

	lamprey	coelacanth	danio	stickleback	medaka	fugu	tetraodon	human
reference cDNAs (#)	11,476	21,958	48,636	27,628	24,662	48,003	23,265	180,654
reference genes (#)	10,449	19,174	27,948	20,839	19,687	18,685	19,749	47,266
A. stellatus ESTs with hit on reference cDNAs (#)	7,906	13,350	13,956	12,327	11,984	11,867	11,490	13,584
***A. stellatus* ESTs with hit on reference cDNAs (%)**	**12.43**	**20.99**	**21.94**	**19.38**	**18.84**	**18.66**	**18.06**	**21.36**
reference genes identified (#)	3,710	6,261	6,886	6,229	5,961	5,913	5,865	6,557
reference genes identified (%)	**35.51**	**32.65**	**24.64**	**29.89**	**30.28**	**31.65**	**29.7**	**13.87**

Table 3.11. Table 10. TBLASTX best hit (e-val < 1e-03) of *A. stellatus* transcriptomes against cDNA sequences from Ensembl database.

	lamprey	coelacanth	danio	stickleback	medaka	fugu	tetraodon	human
reference proteins (#)	11,429	21,817	41,693	27,576	24,661	47,841	23,118	97,041
reference genes (#)	10,402	19,033	26,160	20,787	19,686	18,523	19,602	21,860
A. stellatus ESTs with hit on reference proteins (#)	7,522	11,932	12,561	11,646	11,516	11,287	11,142	11,823
***A. stellatus* ESTs with hit on reference proteins (%)**	**11.83**	**18.76**	**19.75**	**18.31**	**18.11**	**17.75**	**17.52**	**18.59**
reference genes identified (#)	3,521	5,941	6,475	5,999	5,762	5,762	5,726	5,923
reference genes identified (%)	**33.85**	**31.21**	**24.75**	**28.86**	**29.27**	**31.11**	**29.21**	**27.1**

Table 3.12. BLASTX best hit (e-val < 1e-03) of *A. stellatus* transcriptomes against protein sequences from the Ensembl database.

	lamprey	coelacanth	danio	stickleback	medaka	fugu	tetraodon	human
reference ncRNA genes (#)	2,628	2,918	4,431	1,617	735	703	813	9,399
A. stellatus ESTs with hit (#)	7	11	25	21	16	10	21	40
A. naccarii ESTs with hit (%)	0.06	0.02	0.02	0.03	0.01	0.04	0.03	0.02
reference ncRNA identified (#)	2	6	9	8	8	4	8	22
reference ncRNA identified (%)	**0.23**	**0.21**	**0.21**	**1.09**	**0.08**	**0.2**	**0.49**	**0.57**

Table 3.13. BLASTN best hit (e-val < 1e-03) of *A. stellatus* transcriptomes against non-coding RNA genes from the Ensembl database.

GO annotation

The GO annotation process allowed the mapping of GO terms associated with the best hits on about 46% of all ESTs (29,243), with a significant match against NCBI nr. Using InterProScan online, via Blast2GO, I also obtained information about functional motifs and conserved domains within them for 33,299 ESTs (52%). After merging, a total of 8,640 different GO terms (5,584 for biological process, 952 for cellular component and 2,104 for molecular function) resulted consistent to annotate 9,846 ESTs representing 15% of the total *A. stellatus* transcriptome. This fraction nearly overlaps the fraction of *A. naccarii* transcriptome annotated (16%).

GO-slim terms, were then mapped onto annotated ESTs. DAG graphs were constructed for each of the three ontology domains and each graph was cut at level three. The number of annotated ESTs with terms belonging to level three for each of the domains are reported in Figure 3.23. The annotation with enzyme codes (EC) gave the following results: 4,204 ESTs were annotated with 493 unique enzyme codes, involving 120 biological pathways. All pathways found are listed in APPENDIX B. Among the most active biochemical

reaction chains I found were "Purine metabolism" (map 00230) with 47 enzymes and "Arginine and proline metabolism" (map 00330) with 25 enzymes.

3.5 Search for sex-determining genes

Putative sex related genes found in *A. naccarii*

I evaluated the presence of 32 candidate genes known to be involved in sex determination and sexual development in vertebrates by queering the *A. naccarii* transcriptome with 3 collections of orthologs and paralogs for those genes (Ensembl Compara, Homologene, Acipenser-specific genes, see methods). After accurate evaluations of results, significant matches were found for 22 out of the 32 genes investigated. The alignments of matching contigs were manually inspected to exclude false-positive matches exclusively due to the presence of widespread protein domains. A complete list with exhaustive descriptions of the best matching contigs considered to have a reliable similarity against the 22 genes recognised is contained in APPENDIX C while a summary list of the same genes is reported in Table 3.14. A similar transcriptomic screening for genes involved in sex differentiation was performed on the lake sturgeon (*A. fulvescens*) (Hale et al. 2010). The authors report positive matches for 12 genes (SOX2, SOX4, SOX17, SOX21, SOX9, DMRT1, RSPO1, WT1, WNT4, FOXL2, TRA-1, FEM1), all but one included in my search list (Table 3.14, APPENDIX C), the exception being TRA-1. All genes were also detected in *A. naccarii* with the exceptions of DMRT1 and WNT4. Positive matches with SOX genes (SOX2, SOX4, SOX21) were discarded after manual inspection, because the same contigs also matched other SOX genes with higher scores. This multiple matching is due to the fact that genes of the SOX family often share the conserved High Mobility Group box domain and assignment based on this domain makes for a less-reliable identification. Special attention was given to the contigs observed to be library-specific. Among the 22 genes detected, only 5 (WT1, LHX1, CYP19A1 (aromatase), FHL3, FEM1A) and 2 (AR, EMX2), were found to be specific to male (cDNA3) or female (cDNA4) libraries. The remaining fifteen genes were detected by contigs belonging to the common fraction.

Preliminary results in *A. stellatus*

The search for sex-determining genes was also performed in the *A. stellatus* transcriptome, but results are still in the form of raw matches that need to be manually screened in order to discharge false positives and select best-candidate contigs for each

gene. The tables in Table 3.14 show the number of hits obtained for each of the three sets of queries used for the screening. The number of contigs with a significant match against each gene were compared with the number of best-candidate contigs selected in the *A. naccarii* transcriptome after manual inspection according to the criteria described in the methods.

A. naccarii				*A. stellatus*			
gene name	cDNA3	cDNA4	common	gene name	cDNA1	cDNA2	common
WT1	1			WT1	1	1	5
WNT4				WNT			
VTG2							
STRA8				TRA	13	10	49
SRY				SRY	1	2	3
SOX9			1	SOX	2	4	7
SOX6			1				
SOX4							
SOX21							
SOX2							
SOX17			1				
SOX14			1				
SOX11			1				
SOX1			1				
SF1			1	SF	19	9	57
RSPO			1	RSPO	1		
LHX9			1	LHX		2	3
LHX1	1						
GATA4			1	GATA			1
FOXL2			1	FOX	1	1	12
FIGLA				FIGLA			1
FHL3	1			FHL	1	2	3
FGFR2			1	FGF	2	3	5
FGF9			1				
FEM1A	1			FEM	3		3
EMX2		1		EMX	1		3
DMRT1				DMRT	1		
DAX1			1	DAX			1
CYP19A1 (aro-matase)	1			CYP	11	6	22
ATRX			1	ATRX		2	2
AR		1		AR	50	41	157
AMH				AMH		2	1

Table 3.14. Preliminary results of the sex-related genes' search in the *A. stellatus* transcriptome.

3.6 Discovery of variants

Variants in *A. naccarii* transcriptome

At 90% Bayesian probability, I were able to identify 23,084 SNPs and 59,150 INDELs. After having filtered out variants beside simple sequence repeats, 21,791 SNPs (94.04%) and 57,996 INDELs (98.05%) were retained from 6,283 and 8,678 contigs respectively. Between contigs-containing variants, the average SNP per contig was 3.5, while the mean

INDELs per contig was 6.7. The mean frequency across all contigs was 1 SNP every 1. Kbp, and 1 every 377 bp for the INDELs. I counted mutations between nucleotides with bases of similar ring structures (Ts: transition from a purine to a purine or from a pyrimidine to a pyrimidine) and mutations between bases of different ring structures (Tv: transversions, from a purine to a pyrimidine or vice-versa). I identified 14,433 Ts and 7,358 Tv, thus confirming that transitions are more common than transversions in *A. naccarii* dataset (Kimura 1980).

I counted 1,237 contigs with Ts/Tv < 1. As transversions are usually less frequent then transitions, these contigs may represent transcribed sequences subjected to diversifying selection. I then classified SNPs that fell in predicted coding regions according to the type of mutations they carried out: if they performed non-synonymous mutations (Ka) that changed the amino-acid sequence, or synonymous mutations (Ks). Of the overall contig-containing SNPs, I was able to identify a putative ORF for 2,482 of them on the basis of the best match against nr database, while for 3,786 an ORF was predicted:15 seemed to contain no ORFs. Of the overall SNPs found in coding regions, 2,750 represented non-synonymous mutations while 1,056 were synonymous.

I found that 1,280 contigs (2.32% of all contigs) had Ka/Ks > 1 thus indicating genes putatively under diversifying selection within this samples. On average, I found 0.73 non-synonymous and 0.28 synonymous SNPs per contig in coding regions; this means one non-synonymous mutation every 9 Kbp of coding portion, and 1 synonymous mutation every 20.7 Kbp. Distribution of SNPs and INDELs across contigs together with distributions Ka/Ks are shown in Figure 3.26. Distribution of Ts/Tv is shown in Figure 3.27.

I also scanned the entire EST set for Sample Sequence Repeats (SSRs, also known as microsatellites), and I identified 5,295 SSRs present in simple formation, within a total of 4,670 (8%) contigs. In particular, I found 1,891 dinucleotides, 2,377 trinucleotides, 1,001 tetranucleotides, 100 pentanucleotides and 45 hexanucleotides. The graph in Figure 3.28 shows the frequency of repeat types found accordingly to unit size. Of the overall contig-containing SSRs, 4,639 also contain a putative ORF. In total 1,779 SSRs are predicted within ORFs (33% of all identified SSRs).

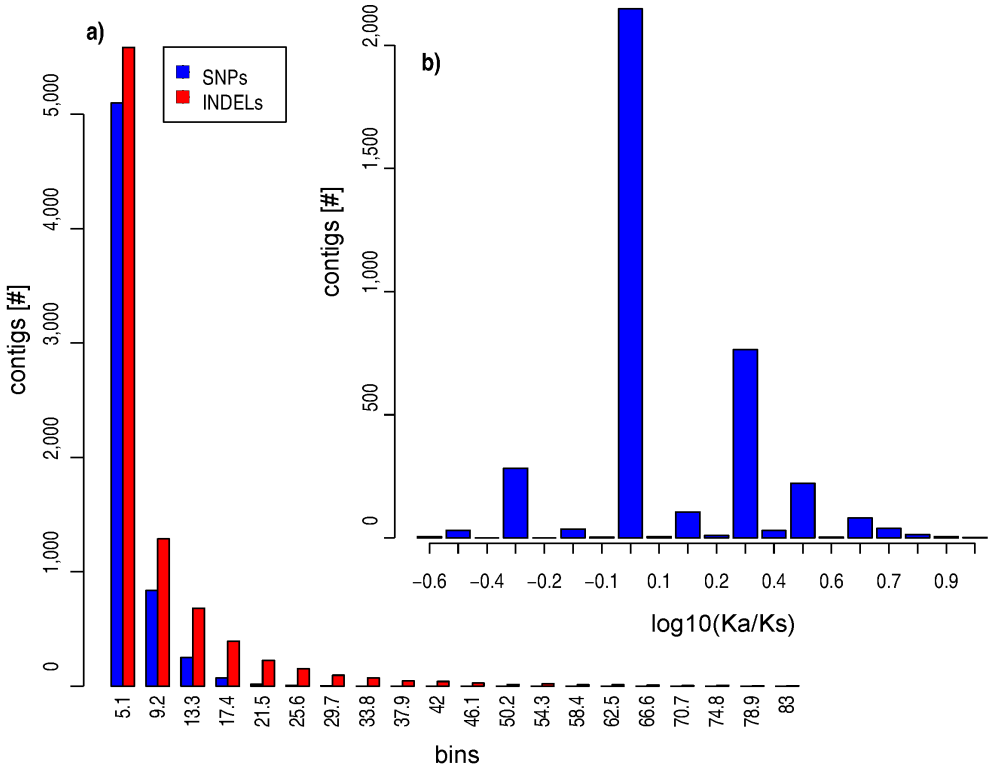

Figure 3.26. Distribution of SNPs and INDELs across *A. naccarii* contigs

a) Bar-plot of the distribution of SNPs and INDELs in contig-containing variants. Most contigs contain up to 5 variants. **b)** Bar-plot of log10 distribution of Ka/Ks for contig-containing SNPs that lie in predicted ORFs. Contigs with Ts/Tv < 1 (log10 < 0) and Ka/Ks > 1 (log10 > 0) are proposed to be under diversifying selection in *A. naccarii*.

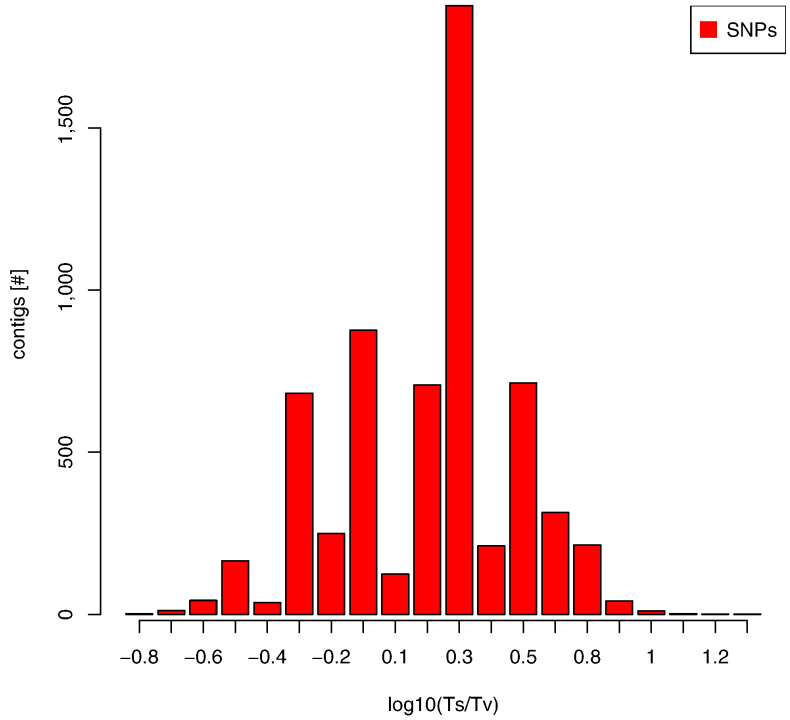

Figure 3.27. Bar-plot of the log10 distribution of Ts/Tv across *A. naccarii* contigs.

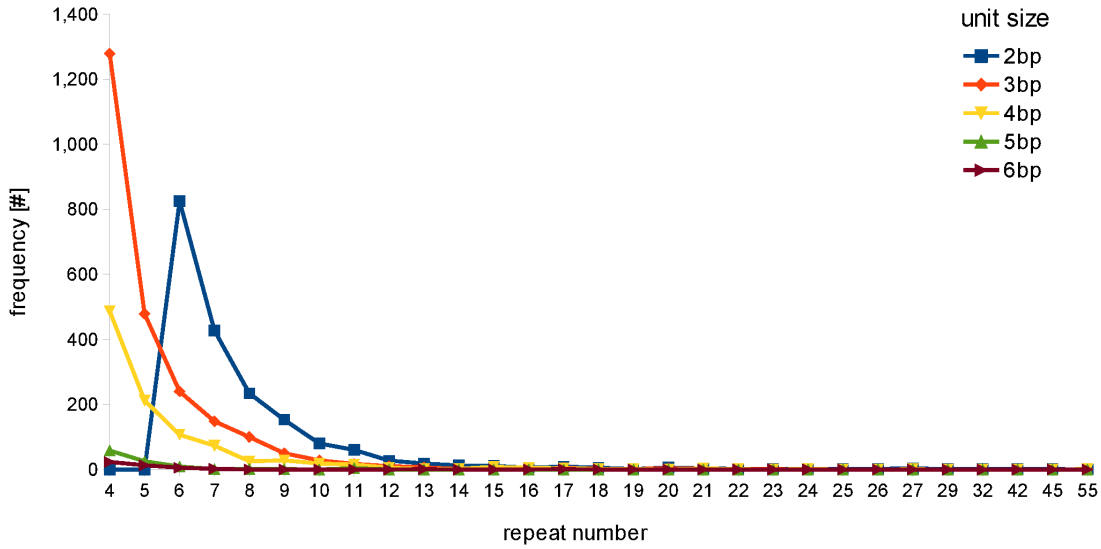

Figure 3.28. Frequency of classified SSR repeat types in *A. naccarii*.

The graph shows the frequency of each repeat motif classified, considering the sum of the frequencies for complementary sequences (for example, the sum of frequencies for the dinucleotides AC and its complementary GT), for the 7,081 total SSRs identified in 6,079 contigs.

Variants in *A. stellatus* transcriptome

With Freebayes scanning, we identified 15,449 SNPs and 46,103 INDELs with high confidence and far from homopolymeric regions. These polymorphisms lie respectively in 8,569 and 5,765 contigs. By focusing on SNPs, a frequency of about 3 SNPs per contig was calculated, i.e. one SNP every 1,58 Kbp. In contig-containing SNPs, 10,337 Ts and 5,112 total Tv were identified, confirming yet again that Ts are more common than Tv. A set of 1,250 contigs (~2%) showed Ts/Tv < 1. These 1,250 contigs represent genes likely under diversifying selection.

Among all contig-containing SNPs, 1,940 obtained a significant match against NCBI nr and thus contain a putative ORF. However, 3,812 did not get any hits, but an ORF was predicted within them. Finally, 13 contig-containing SNPs do not seem to contain any ORFs. Of the 2,925 SNPs lying within predicted ORFs, 2,071 appear to cause non-synonymous mutations in the amino acid sequences encoded by the ORFs, an average of one mutation every 11.78 Kbps of expressed sequences. By analysing the ratio between synonymous and non-synonymous mutations, I identified 1,115 contigs (1.75%) that have Ka/Ks > 1. Again, these sequences may represent genes under diversifying selection in *A. stellatus*. The distribution of SNPs and INDELs in the transcriptome, together with the distribution of Ka/Ks ratios, are shown in Figure 3.29, while barplot of Ts/Tv distribution

is shown in Figure 3.30.

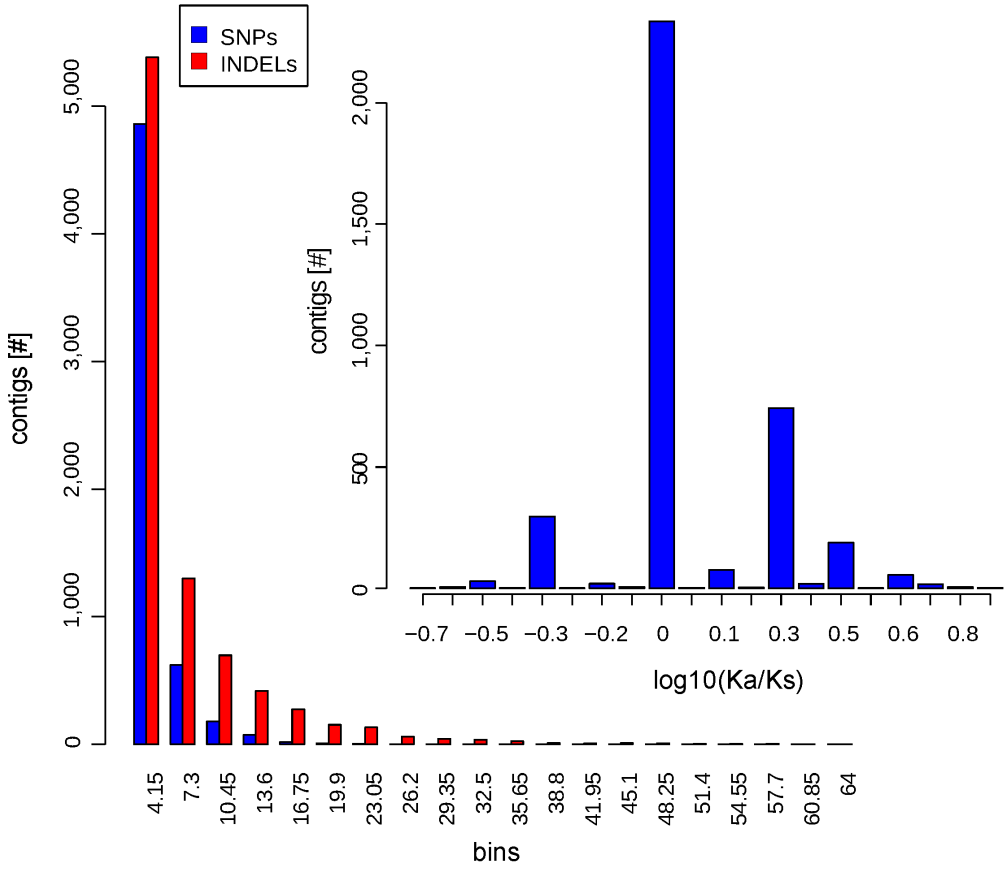

Figure 3.29. Distribution of SNPs and INDELs across *A. stellatus* contigs.

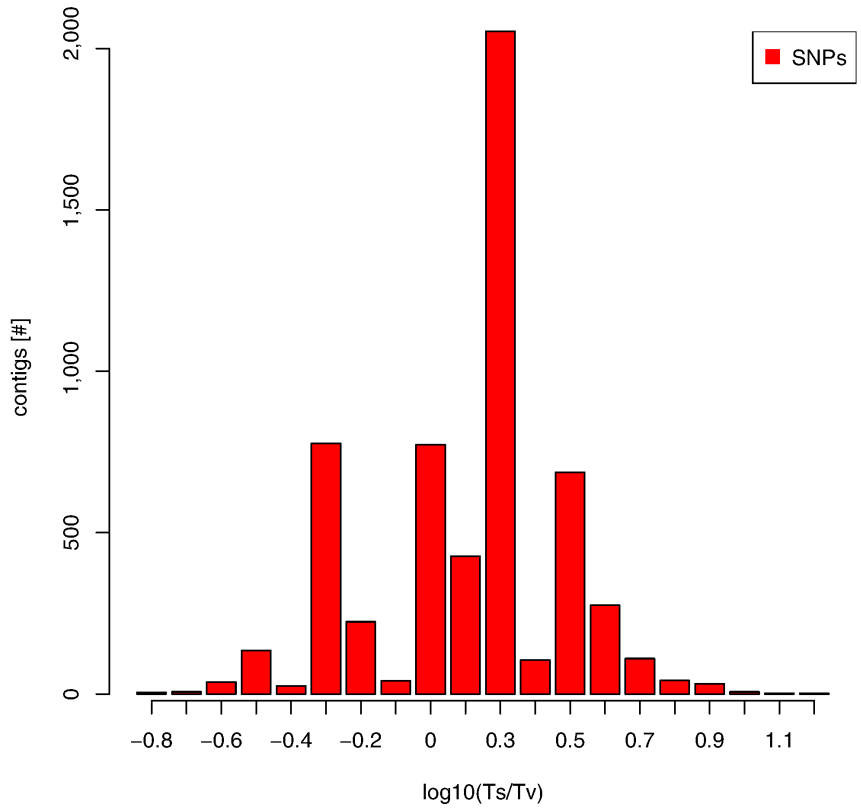

Figure 3.30. Bar-plot of the log10 distribution of Ts/Tv across *A. stellatus* contigs.

In the search for microsatellite repeats, I found 5,696 SSRs in 4,930 contigs (~10% of the total): 2,048 di-, 2,466 tri-, 1,032 tetra-, 124 penta- and 26 hexa-nucleotides. Their frequency according to type is shown in the graph of Figure 3.4. Most of the 4,930 contig-containing SSRs, also contain a putative ORF: 4,906 in total. Within these ORFs are predicted to lie 1,779 SSRs (33% of the total SSR identified).

3.7 Evaluation of functional ploidy levels

After the first step, the assembly pipeline MiraSearchESTSNPs, aligned a total of 519,337 reads from the 4 *A stellatus* and *A. naccarii* sequence libraries, which represent about 77% of the total. The result were 76,875 contigs and 482 singletons. In the second step, 87,274 reads from the male and 86,810 from the female *A. stellatus* libraries, selected during the first step, were individually assembled to produce 19,847 and 18,904 contigs respectively. Similarly 86,810 reads from the male and 96,016 reads from the female *A naccarii* libraries, were assembled into 17,551 and 17.355 contigs respectively. Finally, in the third step, a total of 64,971 contigs from all assemblies of the second step, were

clustered together to return 17,047 transcripts shared between at least two individuals of the two species. A fraction 8,686 transcripts resulted not shared. More details about all assemblies produced by the 3 steps, are reported in Table 3.15.

	Step 1	Step 2				Step 3
		cDNA1	cDNA2	cDNA3	cDNA4	
Reads/contigs aligned (#)	519,337					64,971
***A. stellatus* male (cDNA1)**	130,909	87,274				16,089
***A. stellatus* female (cDNA2)**	127,617		86,810			15,190
***A. naccarii* male (cDNA3)**	120,138			82,170		14,545
***A. naccarii* female (cDNA4)**	140,673				96,016	14,323
"remain" contigs (#)						4,824
Singletons (#)	482	38	29	26	21	8,686
Contigs (#)	76,875	19,847	18,904	17,551	17,355	17,047
Total consensus (bp)	42,512,728	10,504,208	9,677,177	9,431,699	8,621,225	12,313,054
Largest contig (bp)	4,551	2,349	1,822	1,916	2,068	4,465
Average consensus quality (phred)	41	43	40	43	39	63

Table 3.15. Statistics of the assemblies from the 3 steps of MiraSearchESTSNPs pipeline.

The bar chart in Figure 3.31 shows the distribution of the alleles for transcripts shared by at least two individuals. As it can be seen, most of the transcripts exhibit a single allele, for each sample. This flattens the averages of the distributions around the value 1 as clearly demonstrated by the box-plot in Figure 3.32.

Figure 3.33 highlights the allele frequencies in the range 3-12. It can be recognized that transcripts with these allele frequencies increased progressively, in the two *A. naccarii*. For example 0.86% and 0.78% of total shared transcripts in *A. stellatus* male and female respectively, showed 3 alleles while 0.90% showed the same frequency in both *A. naccarii*. The 0.26% and 0.19% of transcripts in *A. stellatus* male and female showed 4 alleles, while 0.38% and 0.31% showed the same frequencies in *A naccarii* male and female. The difference more than doubled in the bin 5 (0.07% cDNA1, cDNA2 0.05%, 0.14% cDNA3, 0.13% cDNA4). The data shows a wake signal of doubled functional ploidy in the Adriatic sturgeon compared to the Stellate one.

I isolated transcripts shared by all sturgeons, which showed at least a double amount of

alleles, in the *A. naccarii* individuals, compared to the *A. stellatus* samples. These were only 241, about 1, 4% of the total. Moreover I estimated the error of the allelic abundance, to be ±1 allele. The results of the corresponding alignments are reported in Table 3.16.

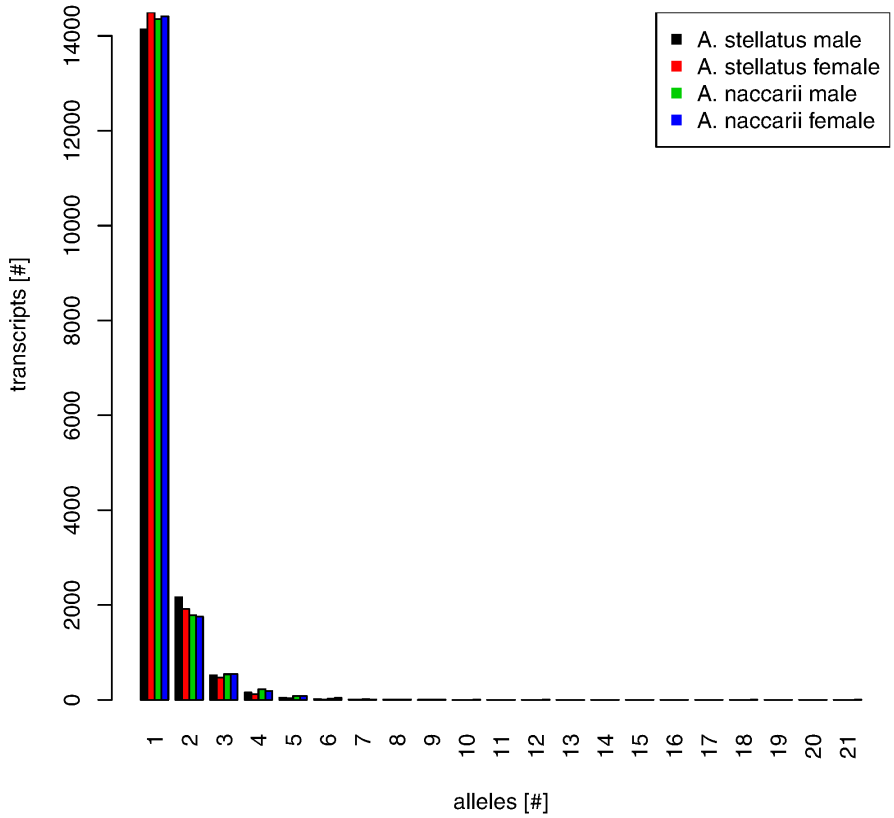

Figure 3.31. Allele distribution for transcripts shared across sturgeon samples.

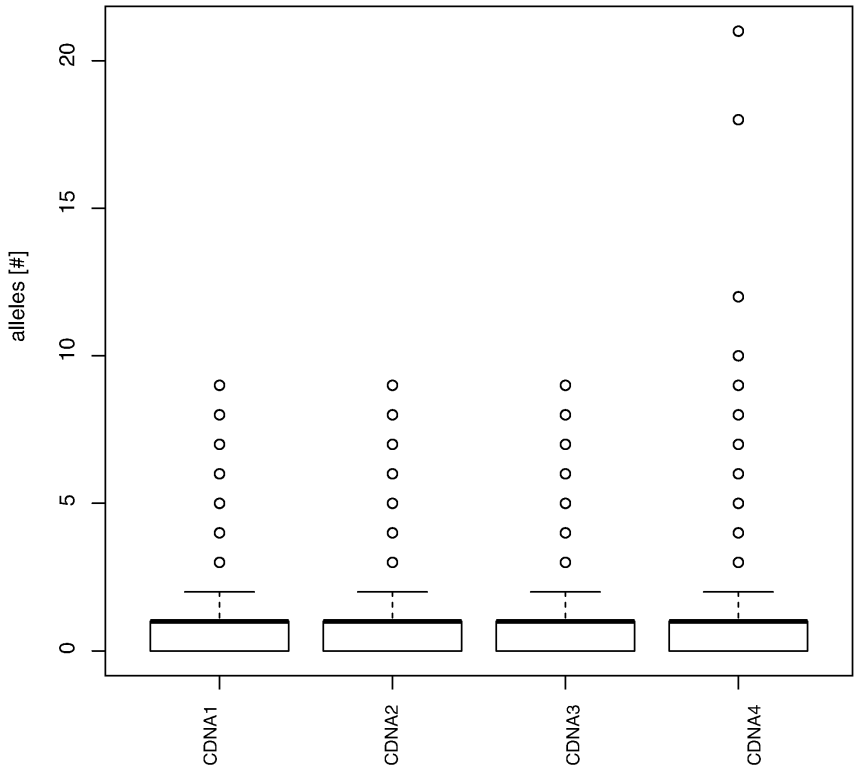

Figure 3.32. Average distribution of alleles for shared transcripts across the 4 transcriptomes.

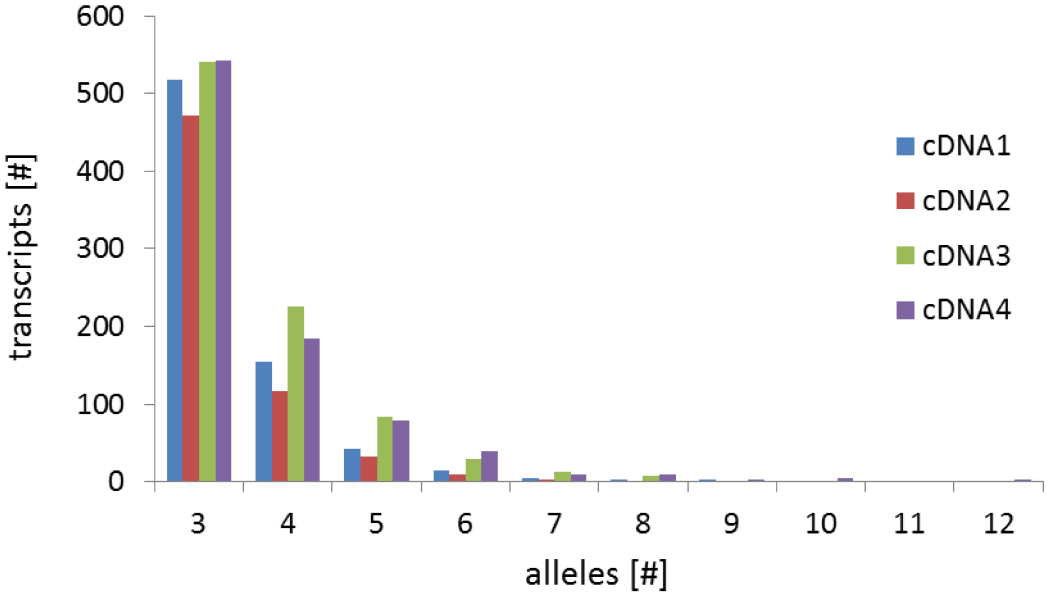

Figure 3.33. Enlargement of allelic abundance in the range 3-12.

target gene	best hit contig	% identity	alleles			
			cDNA1	cDNA2	cDNA3	cDNA4
COIII	step3_c103	93.89	1	1	0	2
ND3	step3_c187	91.12	1	1	1	1
COI	step3_c688	100	2	1	0	0
ND4	step3_c7364	93.5	1	1	1	1
ND5	step3_c76	90.78	3	3	4	2
ND6	step3_c76	92.75	3	3	4	2
ND1	step3_c89	91.21	1	1	1	2
ND2	step3_c9668	93.88	1	0	1	2
ATP_6	step3_c9790	93.24	1	1	1	1
COII	step3_rep_c5804	94.94	1	1	0	1
ND4L	step3_rep_c8707	94.28	1	0	2	1

Table 3.16. Alleles obtained for the best matching third round contigs against the 11 mitochondrial coding genes of *A. transmontanus*.

3.8 AnaccariiBase: the *A. naccarii* transcriptome database

AnaccariiBase is freely available at: http://compgen.bio.unipd.it/anaccariibase/, it contains *A. naccarii* transcriptome informations and results of my bioinformatics analysis, organised in different layers. It has been implemented using MySQL and Django web framework, in collaboration with Dr. Alessandro Coppe who take care of most of the implementation issues. The database is focusses on contig sequences and annotations, and can be searched using contig ID and key-words. Moreover, it allows the user to conduct a local BLAST search on the fly against contigs, using nucleotide or protein sequences, to identify one or more transcript significantly similar to a given query sequence. Furthermore the system provides a customisable data retrieval tool to download large amount of data. The information layers that constitute the implementation are Contig, Assembly, Gene Ontology (GO) and BLAST results, detailed as follows: (1) For each contig, an ID is given together with the FASTA sequence and an informative description, which is defined by the Blast2GO natural language text mining functionality, related to the BLAST hits. The best hit is used when a Blast2GO description is unavailable. (2) The list of the reads assembled into each contig is accessible to the user, together with their sequences. (3) GO terms associated to each transcript are reported, for Biological Process, Molecular Function, and Cellular Component domains, with hyper-link to the GO database. (4) Pre-calculated BLAST results of contigs against the main protein (nr) and

nucleotide (nt) databases, are available with no waiting time, in the classic BLAST output format. Results are hyperlinked to the external databases, and include alignment descriptions and details about the pairwise alignments of each contig with the corresponding BLAST hits.

An example of the results returned by the search system is shown in Figure 3.34 while in Figure 3.35 it can be seen details accessible by clicking on the first result. The “query” page, provides the powerful system that permits to look for specific contigs by: ID, GO annotation term, GO term ID, description and BLAST hits description. The special keyword “HAS” allows to select all sequences featured with a certain field. In addition, the user can choose which informations of the selected subjects to download (ID, description, GO terms associated, GO term IDs, sequences), in the form of TSV file. For the impatient, specific links available at the home page allow to download the complete archives containing all raw reads by library of origin, and all contigs obtained by the joint assembly.

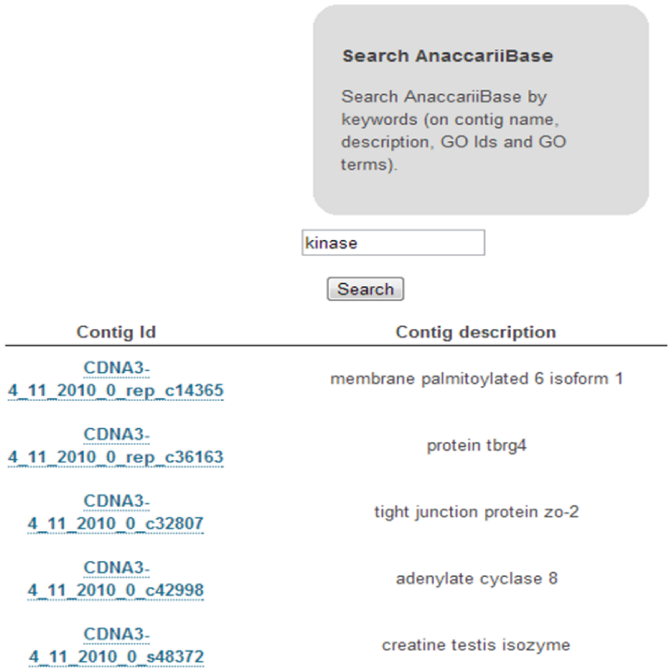

Figure 3.34. Results returned by the AnaccariiBase search system for the key-word "kinase".

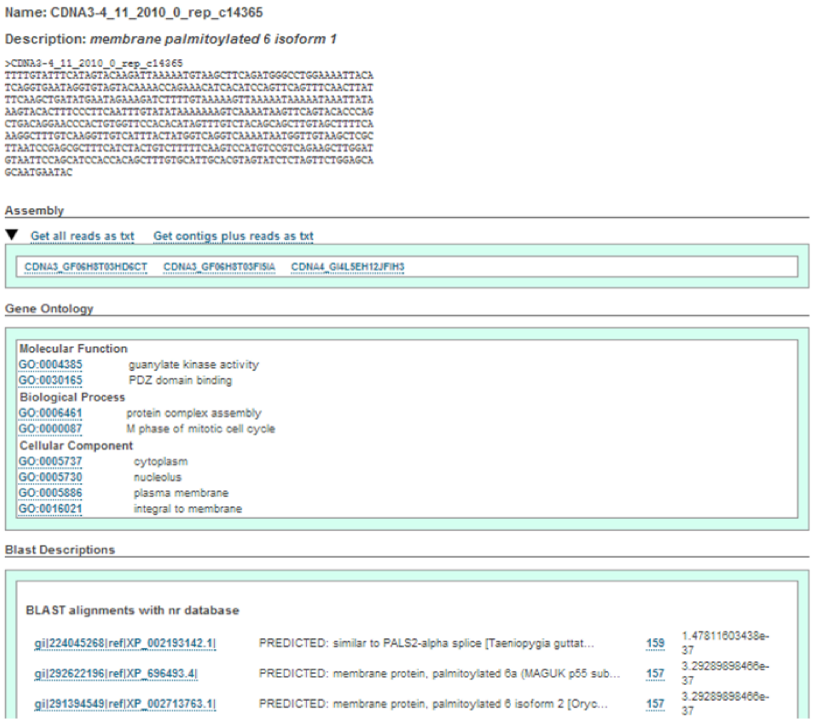

Figure 3.35. Example of features available in "gene-like" form through AnaccariiBase. For each contig, different information are given together with links to external databases.

4 Discussion

4.1 Assembly of the transcriptomes

A theoretically excellent assembly should use all the available reads and should trace reads unambiguously to contigs constructed. It should not contain redundant or widely overlapping contigs from allelic copies of the same transcripts, or data with many errors. This means a large assembly with a lot of redundant contigs could be worse than a shorter assembly with unique sequences. Again, a very good assembly should have the maximum total length of contigs and should not contains chimeric sequences. Moreover contig length distribution should resemble that expected from a real transcriptome (Papanicolaou et al. 2009). Thus, beside the widely used metrics for comparing different assemblies such as: contig length, total number of contigs obtained, total amount of assembled bases, mean contig length distribution and maximum size, a better benchmark of an EST assembly is the identification of proteins and transcribed sequences from one or more phylogenetically closer species, through a sequence alignments (i.e. BLAST). In order to avoid species-specific bias and increase the power of detecting coding sequences it is important that more than one proteome or cDNA database is provided as reference. However, using species too divergent from the one under analysis, could give misleading results. In the case of sturgeons no large proteomes were available for the genus *Acipenser*, while very well characterised proteomes existed for the teleosts *Danio* and Fugu. Accordingly, these species were also used as reference.

My comparison between MIRA 3.2 and Newbler 2.3 assemblers confirmed that Newbler adopted an approach too conservative by discharging a lot of reads, thus giving a smaller number of (carefully) assembled contigs. This unfortunately, lead to the loss of a significant portion of assembled transcriptomes. However, thanks to the reports of several authors (Papanicolaou et al. 2009; Kumar & Blaxter 2010; Finotello et al. 2012), this problem was partially overcame with the later version 2.5. It was demonstrated that the assemblies from MIRA included more reads then that from Newbler, were composed of more contigs that were able to recognize more sequences and cover more positions of the reference databases, while the annotation redundancy, resulted similar to that from the Roche assembler. For these reasons, MIRA was finally chosen.

A high number of reads included in an assembly can improve the probability of discover SNPs or alternative splicing variants through downstream analysis. A higher

proportion of reference databases covered could mean, a greater chance to discover new genes. Conversely, the number of contigs is not always a good prediction of the number of genes sequenced. In fact, non-coding regions such as UTR (UnTranslated Region) or introns from different alleles typically fails to assemble together because they diverge rapidly due to the lack of any selective constraints. This seems to be the major cause of contig inflation. Therefore, for my outbred sturgeon individuals I expected a very high number of contigs within the transcriptomic assemblies, especially for the *A. naccarii* species, considered to be tetraploid. In addition, MIRA in its basic setting is an assembler and not a clusterer, so it splits up read groups into different contigs if it estimate they represent transcripts sharing exons but that substantially differ for real polymorphisms, gaps or sequencing errors. This is the right method to avoid assembling close paralogs (because joining contigs is easier that splitting them) but results in a long computation time when redundant sequences are annotated.

The results of the parametrisation showed that, when configured as a clusterer, MIRA builts a smaller number of contigs using the same number of reads. The smaller number of contigs resulted in a decreased number of hits and positions covered on the reference databases, while redundancy was slightly reduced. No appreciable changes were reported regarding the repeat-management settings. This is consistent with the higher frequency of repeats in intergenomic than in transcribed regions.

To partially reduce the redundancy I performed an iterative assembly process being aware that some degree of assembly accuracy was lost. In fact, by forcing MIRA to resolve ambiguous positions by choosing a consensus, the probability of losing rare transcriptional variants is increased. However, two assembly cycles were performed for three reasons: 1) I were more interested in having a general overview of genes expressed in both *A. naccarii* and *A. stellatus*, 2) information on rare variants can be traced back, realigning all the original reads on the corresponding contigs, and 3) the mean coverage per contig is increased. The latter is a positive result, since the average contig coverage of joint assemblies was low for both species and this was an important limiting factor in SNP prediction and overall quality of the assemblies.

4.2 Estimation of sequencing completeness

Despite two assembly rounds to reduce intrinsic transcriptome redundancy, the estimated average sequencing coverage of contigs remained still low (about 4.09

base/position for *A. naccari*i and 3.55 base/position for *A. stellatus*, see Table 3.3 and Table 3.5). This suggested that transcriptomes under study may have remained a little under-sequenced, especially rare transcripts. So it become important to assess the completeness of both the sequencing and the transcriptome representations.

To evaluate the sequencing completeness of the cDNA libraries, I resort to a rarefaction analysis. Results shown that when reads from the male and female libraries were considered to built separate saturation curves in both species, the potential to identify different *Danio* transcripts, was about 83% of the theoretical maximum for both *A. naccarii* male and female libraries while was 80% for *A. stellatus* male and 81% for *A. stellatus* female libraries respectively.

As expected, when reads from libraries of different sex were merged, the estimated sequenced fraction increased in both species: 88% for *A. naccarii* e 87% for *A. stellatus*, while the slopes at the maximum sample size decreased in both cases (Figure 1.1 and Figure 3.2). This is because the library-specific transcripts present, or been sequenced in one library but not in the other are put together.

Results from further analysis collected in Table 4.1, showed that by using, as reference, cDNA sequences from other species, the absolute value of the potentially identified transcripts, calculated through a saturation model, and those actually identified, change, but the ratio between these quantities remained nearly constant (see Figure 3.1 and Figure 3.3). This ratio therefore, represents a robust value indicating the fraction of the cDNA physical libraries really sequenced.

Target cDNA sets	***A. naccarii* fractions (%)**	***A. stellatus* fractions (%)**
Danio	88	87
Coelacanth	89	88
Fugu	88	86
Human	86	85
Lamprey	90	89
Medaka	89	88
Stickleback	89	87
Tetraodon	89	88

Table 4.1. Sequencing efficiency of the joined male and female libraries, estimated using

cDNAs from different species as reference sequences. The values are consistent.

The comparison with the data from *A. naccarii* libraries suggests that *A. stellatus* physical libraries probably contain a greater number of sequences than *A. naccarii* libraries. Two observations support this idea: the number of raw reads obtained for *A. stellatus* was greater, while the estimated sequencing completeness was the same for the two datasets. Taken together, these data indicated that a half 454 plate is not sufficient to completely characterize the transcriptome of a sturgeon species.

4.3 Transcriptomes completeness estimation

The approach used to estimate the total number of transcripts potentially represented by the two assemblies of *A. naccarii* and *A. Stellatus* respectively, represents an adaptation of the capture-recapture method widely used in ecology to estimate animal population sizes (Chao 1989).

The capture-recapture method requires at least two capture experiments: capture units, mark them, release them, then recapture and count the already marked units. Since the two libraries in each species were constructed from the same tissues from two different animals (two full sibs, one male and one female, of the same age and history), by neglecting the differences due to sex-specific transcripts, it is possible to consider them as two independent samples from the same transcripts population. The number of the recaptured marked units corresponds to the number of transcripts tagged by both libraries, i.e. the number of contigs composed by reads from both libraries. Furthermore, it can be assumed the transcripts population as a closed population, i.e. the total number of transcripts expressed was the same before the construction of the two libraries. Finally, it can be assumed that the probability of sampling a transcript does not change between samples.

Given we are contemplating two capture-recapture experiments corresponding to the two libraries, and the capture probability p can be considered costant, at every capture occasion. The estimator that best fit the data is the "Schnabel estimators" $[(n_1+1)(n_2+1)/(f_2+1)]-1$ that represents the biased-corrected form of the simple "Petersen estimator" $n_1 n_2 / f_2$, where n_i are all the units captured in the ith sample and f_i are the units captured exactly i times in all samples (i.e. the common transcripts) (Chao 1989). The "Petersen estimator" was already applied to estimate the number of human genes by comparing both gene sequences from human chromosome 22 and a non-

redundant set of mRNA sequences, with a set of human EST-derived contigs (Ewing & Green 2000).

This estimation indicated that the joint assemblies the of the *A. naccarii* and *A. stellatus* libraries, represent a comparable fraction of the putatively expressed transcripts in the sequenced tissues. The calculated standard error is negligible. The *A. stellaus* assembly seems to include a slightly higher fraction (86%) compared to the *A. naccarii* assembly (80%). The different ploidy level, between the two species may have contributed to this difference since in *A. naccarii*, tetraploid, more allelic variants could be expressed.

However I'm aware these could be over-estimations of the total transcripts numbers because the presence of 5' and 3' UTR, transcription variants (SNPs, INDEL), splice variants, homologues, and sequencing errors (in particular in homopolymeric regions for 454 technology) during the assembly process may have induced the creation of more contigs than expected. Therefore, additional redundancy could still be present in the common contig fractions (despite attempts to reduce it in the assembly phase), which would have increased the estimated total numbers of transcripts in the tissues of origin, in both species.

The sequencing of the mitochondrial genes, was almost complete for both species; the ATP-ase subunit 8 was missing but for obvious reasons due to sequence shortness. Unlike mt-rRNAs and mt-mRNAs, that are adenylated with, respectively, short and long poly(A) tails at their 3'-terms after processing (Bobrowicz et al. 2008), the mt-tRNAs are not and therefore should not be reverse transcribed during the library preparation process and finally sequenced. Interestingly, 6 of them were identified in our assembly, with sequence identity close to 100%. These mt-tRNAs were matched by the same contigs that aligned with other mt-mRNAs. This finding supports the tRNA punctuation model widely studied in mammals in which the mitochondrial genes are transcribed as a single polycistronic RNA, starting from 2 promoters per strand and tRNAs, placed between other genes, form a cloverleaf structure that acts as a signal for the maturation machinery which recognizes where to cut the precursor (Rorbach & Minczuk 2012). Therefore, contigs that identify both mt-mRNA and mt-tRNA can represent not-yet-processed polycistronic precursors.

4.4 Functional annotation

The alignments of the *A. naccarii* and *A. stellatus* transcriptomes against the reference databases reported comparable results. In both cases it was possible to align the same low

transcript fraction: 34% for the Stellate and 32% for the Adriatic sturgeons. This could be primarily due to the difficulty of finding homologous to sturgeon genes in the present databases, as already experienced by (Hale et al. 2010). Sturgeons belong to Chondrostei lineage of ray-finned fishes for which fewer than 3000 sequences are available. This means there are very few sequences from species closely related to sturgeons in GenBank and that several new genes need to be discover within the transcriptomes of the two sturgeons. Another possible cause is that unannotated sequences are too short and probably lack the conserved functional domains.

The low fraction of aligned transcripts, resulted in an equally low rate of GO terms annotated sequences, in both species: 16% for *A. naccarii* and 15% for *A. stellatus.* However, there is an almost complete sharing of high level ontological categories across all the three GO domains. Furthermore, the number of sequences assigned to each category, does not significantly differs between the two sturgeons. In particular, “protein binding” and “nucleic acid binding” were confirmed as the most assigned categories of the Molecular Function domain, as noted by (Hale et al. 2010). Finally, since purine metabolism is essential to provide the cell with the nucleic acid components and the energy through ATP, I expect it as the most active pathway.

The comparison between the two sturgeons and other fishes for which extensive genomic resources are available (*Petromyzon marinus* (sea lamprey), *Latimeria chalumnae* (Coelacanth), *Danio rerio*, *Gasterosteus aculeatus* (Stickleback), *Oryzias latipes* (Medaka), *Takifugu rubripes, Tetraodon nigroviridis.* and also *Homo sapiens)*, showed that the fraction of *A. naccarii* and *A. stellatus* transcripts that identify putatively orthologous genes in other fishes and humans, reflects the phylogenetic distance between sturgeons and other species, as explained in Figure 4.1.

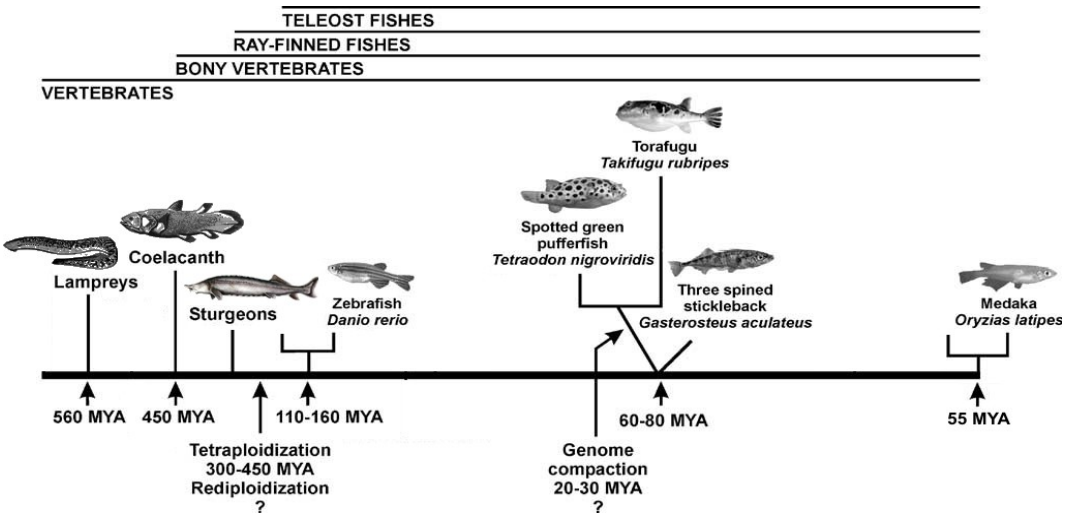

Figure 4.1. Fish lineage evolution. Adapted from (Volff 2004).

It has been postulated that at least one round of whole genome duplication (WGD) occurred before the divergence of jaw-less and jawed vertebrates about 477 Million years ago (Mya, "1R"), on the basis of minimum divergence times estimated based on fossil records (M Kasahara 2007). It was inferred that a second whole genome duplication ("2R") happened after the divergence of jaw-less and jawed vertebrates but before the division between the bony fishes (Osteichthyes) and the cartilaginous fishes (Chondrichthyes) lineage about 450 Mya. More recently, an additional whole genome duplication occurred in the rey-finner fish lineage about 320 Mya, after the separation from Sarcoperygii, about 416 My ago. This fish-specific genome duplication is known as "3R" and it has been hypothesized to be the cause of the rapid speciation in teleosts fishes.

The divergence time of lamprey from jawed vertebrate lineages, was estimated approximately 560 million years ago. Along the lineage of bony fishes, coelacanth have been separed ~450 Mya, from tetrapods and ray-finned fishe (e.g. teleosts) lineages. During all this evolutionary time coelacanth genomes have not experienced WGD or major rearrangement events like teleosts genomes (Amemiya et al. n.d.). Previous studies on the organization and level of duplication in HOX genes in coelacanth, have confirmed that molecular evolution in its genome is going very slow than in tetraposds and teleost fishes.

Teleost fishes are the most divergent and successful group of vertebrates with about 27,000 living species that represents ~50% of extant vertebrates (Kai et al. 2011). In particular Zebrafish specie separates less than 150 Mya and it has been predicted that 30% its genes are present in duplicates (J. H. Postlethwait et al. 2000; I. G. Woods 2005). Medaka and Stickleback diverged at least 55 million and 80 million years ago respectively

(Force et al. 1999),while Fugu and Tetraodon diverged about 44 Mya. It is thought that their genomes undergone a compaction. The Fugu genome is compact, specially because introns are shorter compared with that of other genomes. Moreover, It has been estimated that only ~24% of the ~22,400 genes in Tetraodon, are duplicated.

As for the sturgeons, fossil records and molecular data showed that their lineage separated from teleost lineage more than 250 Mya ago (A Ludwig et al. 2001). It is belived that the common ancestor of all Acipenseriformes had a karyotype of 60 chromosomes. A first WGD duplication events occurred in this ancestor to produce the Acipenseridi lineage whith 120 chromosomes then, a still on going redeploidization may have occurred before the radiation of this orders. Several independent WGD events seems to have evolved more recently in certain sturgeon species.

It is expected in fact, that after WGD the resulting polyploid genome tends to gradually returns in a diploid state throught and extensive gene deletion and only a small part of duplicated genes evolves new function (neofunctionalization) (Jaillon et al. 2004). Alternatively duplicated genes can persist throught partitioning of the ancestral functions between duplicates after complementary degenerative mutations in different regulatory or structural sequences (subfunctionalization) (Volff 2004).

The alignments of the Adriatic and Stellate sturgeons transcriptomes against cDNA and proteins of previously listed fish species showed that the two non-teleost species, the Sea Lamprey and the Coelacanth, share a higher fraction of genes and this was also confirmed at the protein level. A possible explanation for this finding is that these two species separated from the ancestor of teleosts before the WGD. Part of the newly-formed genetic material is known to have persisted after duplication, possibly evolving new functions and thus becoming unrecognisable by sturgeon ESTs. This process of gene diversification results in a reduction of the percentage of detectable genes in teleosts. However, a careful analysis of the matching transcripts and proteins remains be performed to confirm this hypothesis.

In any case, the number of putatively orthologous genes matched by both sturgeon transcripts in other species, is expected to be influenced not only by the genetic similarity among species but also by different parameters such as the accuracy of the genome characterisation in the different species used for the comparison and their evolutionary history in which, for example, different mutation rates may play an important role. In fact,

different lines of Osteichthyes are known to have very different evolutionary rates as a result of different factors such as metabolic features or generation times (Krieger & Fuerst 2002). These differences may deeply affect the number of genes that can be recognized as orthologous among species. Zebrafish seems to share fewer genes (22.89% through transcripts and 22.42% through protein in *A. naccarii* while 24.64% through transcript and 24.75% through protein in *A. stellatus*) with both sturgeons than do other teleosts but, the fraction of the Adriatic and Stellate matching ESTs is comparable to other species. The conclusion is that the *Danio* genome seems to have a higher number of genes. However this could have a different explanation: first, the number of genes is actually higher according to the high level of genes retention after the WGD hypothesized for this species (I. G. Woods 2005) second, more simply, this result is biased by the more complete genome characterization for this model species.

4.5 Search for sex-determining genes

Although there is strong evidence that sturgeon sex is genetically determined (male-to-female ratio 1:1) (A. L. Van Eenennaam et al. 1999b) multitude of molecular genetic approaches used to date, including random screening of polymorphic DNA (RAPDs), single-strand conformational polymorphisms (SSCPs), amplified fragment length polymorphisms (AFLPs), inter-simple sequence repeats (ISSRs) (Wuertz et al. 2006; C. R. McCormick et al. 2008; Yarmohammadi et al. 2012), subtractive hybridization, amplification of regions with high sequence similarity to the sex-determining genes of other fish species (Anne Kathrin Hett & Arne Ludwig 2005; A. K. Hett et al. 2005), and proteomic analysis of male and female gonads (Saeed Keyvanshokooh et al. 2009). None has been successful in the identification of sex-specific DNA in polyploid sturgeon genomes.

In the present study I searched both *A. naccarii* and *A. stellatus* assemblies for the presence of 32 genes known to be involved in sex determination and sexual development in other vertebrates. I arranged 3 collections of orthologs and paralogous of selected genes by selecting sequences from: 1) Ensembl Compara, 2) NCBI Homologene, 3) Acipenser-specific genes from GenBank, (see methods in section 2.6).

The first collection represents the largest variety of annotated homologous (orthologs and paralogs), from sequenced genomes, categorised in Ensembl Compara. The second collection is represented by clusters of more specific orthologs, downloaded from NCBI

HomoloGene (Altenhoff & Dessimoz 2009). The third group of sequences is a collection of complete or partial CDSs from other sturgeon species of the genus Acipenser available in NCBI GenBank. The use of large collections of putative orthologs and paralogs maximizes the possibility of detecting homologues. In contrast, restricted collections of reliable homologues allow a higher confidence on the match they find. If a contig is confirmed as the best subject for a given gene in searches of all trees, then we can be more confident about its identity.

As anticipated, the alignments of matching contigs resulting from the screening of *A. naccarii* assembly, were manually inspected in order to exclude false positive matches, on the contrary, significant alignments found with the transcriptome of *A. stellatus* has not been filtered yet. Therefore, the following discussion is mainly based on the *A. naccari*i results.

What primarily emerged from the study is the absence of the DMRT1 gene from both the *A. naccarii* libraries and also by the preliminary screening of the two *A. stellatus* libraries. This finding is especially interesting and might be due to the incomplete coverage in both species. A second possibility is that this gene is not expressed at the stage at which the samples under study were collected. In fact, the animals analysed for this project were six months old and were at an early stage of gonad differentiation. This is, to my knowledge, the first stage at which sturgeons, which cannot be sexed visually, have unambiguous evidence of gonad differentiation through fine histological investigation (Grandi & Milvia Chicca 2008). The lake sturgeons analysed by Hale and colleagues (Hale et al. 2010) were estimated to be 13 or 14 years old. All characterisations of DMRT1 genes from other sturgeon species have been performed on mature or sub-mature animals (Hale et al. 2010; Amberg et al. 2009). Finally, a low expression of this gene is displayed in the Siberian sturgeon with no evident gonad differentiation (Berbejillo et al. 2012). Thus, the absence of DMRT1 in the transcriptome of the very young *A. naccarii* and *A. stellatus* analysed would suggest that this gene is expressed at a later stage of development in this species (and probably in all sturgeons). DMRT1 in known to play an important role as an activator of the genetic cascade of sex differentiation in some other fishes, such as Medaka (Masahiro Kasahara et al. 2007).

Among genes identified with more confidence I detected SOX9 (Sry box-containing gene 9). It is considered one of the genes at the top of the male sex determination cascade in mammals (McClelland et al. 2012) were it induce testis differentiation by stimulating

the differentiation of Sertoli cells, which then direct testis morphogenesis, while actively suppress genes involved in ovarian development. Its expression in non mammals vertebrates seems to be testis specific and have recently been recently characterised also in fishes included sturgeons (A. K. Hett et al. 2005). I also found traces of other SOX genes (SOX2, SOX4, SOX21) but those were discarded after the manual inspection of *A. naccarii* alignments because same contigs showed multiple matches with different SOX gene with a higher score due to the fact that genes of SOX family share High Mobility Group box and other domains and assignment based on these domains makes the identification less reliable.

Even if most of the genes involved in sex determination are known to act in a dosage-dependent manner (Ferguson-Smith 2007), under the hypothesis that sex differentiation in sturgeon is genetically determined, one could expect that, at the origin of the genetic cascades leading to the different genders, a sex-linked genomic polymorphism occurs. For this reason, special attention was given to the contigs observed to be library-specific. As detailed in APPENDIX C, I observed 5 male specific contigs (from cDNA3) with significant match on genes WT1, LHX1, CYP19A1 (aromatase), FHL3, FEM1A and 2 female specific contigs (from cDNA4) with match against AR, EMX2, in *A. naccarii*. These genes represent, in my opinion, interesting candidate transcripts for experimental validation by PCR amplification.

4.6 Variants discovery

What emerged from the comparison between the transcriptomes of the two species is that approximately 29% more SNPs and 21% more INDELs were found in the *A. naccarii* assembly than in the *A. stellatus* one, within an almost-identical number of contigs. This difference remains constant even after filtration. The number of contig-containing SNPs with a ratio Ts/Tv < 1 remains almost identical between the two species, while 12% more contigs have Ka/Ks > 1 in *A. naccarii*. These discrepancies are, probably, directly related to the different ploidy levels in the two species. In *A. naccarii*, it is likely that very similar reads from the same loci but belonging to a higher number of different chromosomes, were assembled together in a single consensus sequence. The greater number of alleles per locus in *A. naccarii* determined the presence of a greater number of polymorphisms in the *A. naccarii* assembly within a similar number of contigs.

Overall, this availability of a relevant number of EST-linked microsatellites and SNiPs

represents a precious prerequisite for sturgeon conservation genetics in both species under study, thus providing the possibility to monitor the effect of selection on captive and released stocks.

4.7 Evaluation of functional ploidy levels

Polyploidy is known as one of the prominent speciation process in both plants, vertebrates and many other eukaryotes (Wendel 2000). This because genomes evolves by duplication and subsequent divergence of the duplicate parts. However only the most recent duplications are identified as responsible for the "polyploid" speciation events. Various types of polyploid states exist: 1) autopolyploid whose multiple genome copies derive from the genome of a single ancestral species, 2) allopolyploid wich derives from the union of two fully differentiated genomes, and 3) segmental allopolyploid which contains multiple pairs of genomes that share a considerable number of homologous chromosomal segments or even whole chromosomes, but that differs for large number of segments, so that the different genomes produce sterility when present at diploid level.

After a Polyploidisation event, duplicated genes can face different fates. More often one of the duplicated are loss or inactivated through pseudogenization. Alternatively multiple copies can be maintained for long term with a similar if not identical function. Sometimes, with a very low probability genes can evolve new functions or sub-functions because selective pressures are lower in the duplicated copies (Force et al. 1999).

One of the mechanisms responsible for the loss of duplicated genes is the gene silencing. For example, it has been shown that enzyme-encoding genes could respond to polyploidization events with extensive and perhaps repeated episodes of gene silencing (Wendel 2000). Gene silencing can act in an irreversible manner, over a long period of time, or immediately after the duplication event in a reversible way. Point mutations, insertions and deletions, are long-term evolutionary consequences of polyploidy. They can often be the cause of the activity of transposable elements, which can insert into the regulatory regions of the gene. It has been demonstrated, that the polyploid state is associated with an increased activity of transposable elements compared to diploid state, perhaps because of the buffering effect of gene duplication (Wendel 2000). In particular, the high presence and the putative activity of transposable elements in sturgeon genomes, has recently been proven by the characterisation of a new Tc1-like transposable element, putatively expressed in *A. naccarii* (José Martin Pujolar et al. 2012).

A. naccarii and *A stellatus* belong to groups of different ploidy levels. The Adriatic sturgeon (4n=120) has approximately twice the number of chromosomes of the Stellate sturgeon (2n=120). To assess the ploidy level from a functional point of view, through analysis of the transcriptomes, I exploited a multi step assembly approach performed with different levels of alignment stringency in order to: 1) identify transcripts shared by different transcriptomes; 2) identify the number of allelic variants expressed for each transcript, across the two species.

The distinction between different transcripts and alleles, is significantly hard, both due to the complexity of the transcriptome itself and for the presence, especially in polyploid genomes, of paralogous, and the combination of homologous (within genome) and homoeologous (homologous sequences across subgenomes).

This complexity is the cause of a considerable error in the data, that acts by shrinking the differences that exist in the real data. I tried to estimate the error related to allele frequencies, by searching for common transcripts corresponding to mitochondrial coding genes expressed in single copy. The error was very low ±1, however, it is not fully transposable to other genes due to the different expression level that may exist between them.

A. naccarii is considered functionally tetraploid while *A. stellatus* a fully diploid, from the data however, a difference, that could indicate the doubling of the allele frequencies, does not appear. Only by focusing on frequencies greater than 3, it's possible to observe a distinct, even if weak, signal of doubled functional ploidy in the Adriatic sturgeon.

The causes can be several. First of all, the low sequencing coverage may were not sufficient to identify all allelic variants, especially in *A. naccarii*. The assembly parameters used in steps 2 and 3 may were not optimal in order to distinguishing between different transcripts and alleles, in the sturgeon transcriptomes. Moreover, it possible that several portions of the surgeon genomes are experimenting a dosage compensation effect that is temporarily silencing the duplicates. Finally, if confirmed, the activity of transposable elements may still be contributing to gene silencing.

5 Conclusion

The present study provides the first insight into the transcriptome of Adriatic and Stellate sturgeons, More generally, this is also the first massive release of transcriptome information for sturgeon species, shared through a dedicated and searchable database. With nearly 119,000 high quality ESTs, for the two species the information reported represents a significant advance in sturgeon genetics. The apparently limited fraction of successfully annotated sequences with GO terms might be due to the very ancient separation (about 250Mybp) of sturgeons from any other species for which a relevant genomic information is available. (F. Fontana, L. Congiu, et al. 2008). Beside the evolutionary interest of a database obtained from a member of the Chondrostean, certainly applied genetics studies on sturgeons will benefit from this resource. The present study also reports the results of an investigation on genes related to sex differentiation. Out of the 32 genes investigated 7 were detected in only one of the two *A. naccarii* libraries suggesting a possible differential expression between genders at this early stage of gonad differentiation. This result might be affected by the limited coverage of our sequencing and should be considered as a starting point for further investigations. Interestingly, DMRT1, a Master Gene for the sex determination known to be expressed in both sexes in different sturgeon species was not detected, suggesting that, differently from other fish species, DMRT1 is expressed in sturgeons only in latter stages of maturity. A refined analyses of sex determining genre in *A. stellatus* is still to be performed. Finally, the availability of thousands of EST-linked microsatellites makes possible the establishment of a genome-wide genetic markers panel useful to monitor the effect of different selective pressures and to monitor the effects of restocking practices. Restocking of most sturgeon species depends on *ex situ* conservation because of the dramatic decline of natural populations (Leonardo Congiu et al. 2011b). In synthesis, the data provided in the present study and shared through a dedicated website represents the first massive release of information on a sturgeon transcriptome and will hopefully constitute a useful contribution to sturgeon genetics, aquaculture, and conservation.

Bibliography

Altenhoff, A.M. & Dessimoz, C., 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS computational biology*, 5(1), p.e1000262.

Amberg, J.J. et al., 2009. Sexually dimorphic gene expression in the gonad and liver of shovelnose sturgeon (Scaphirhynchus platorynchus). *Fish Physiology and Biochemistry*, 36, pp.923–932.

Amemiya, C.T., Lander, E.S. & Myers, R.M., A white paper for sequencing the genome of a living fossil: the coelacanth, Latimeria chalumnae.

Andrews Simon, *FastQC*, Babraham Bioinformatics. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Baillargeon, S. & Rivest, L.P., 2007. Rcapture: loglinear models for capture-recapture in R. *Journal of Statistcal Software*, 19(5).

Berbejillo, J. et al., 2012. Expression of dmrt1 and sox9 during gonadal development in the Siberian sturgeon (Acipenser baerii). *Fish Physiology and Biochemistry*.

Birol, I. et al., 2009. De novo transcriptome assembly with ABySS. *Bioinformatics*, 25(21), pp.2872–2877.

Birstein, V.J. & Vasiliev, V.P., 1987. Tetraploid-octoploid relationships and karyological evolution in the order Acipenseriformes (Pisces) karyotypes, nucleoli, and nucleolus-organizer regions in four acipenserid species. *Genetica*, 72(1), pp.3–12.

Bobrowicz, A.J., Lightowlers, R.N. & Chrzanowska-Lightowlers, Z., 2008. Polyadenylation and degradation of mRNA in mammalian mitochondria: a missing link? *Biochemical Society Transactions*, 36(Pt 3), pp.517–519.

Boscari, E., Barbisan, F. & Congiu, L., 2011. Inheritance pattern of microsatellite loci in the polyploid Adriatic sturgeon (Acipenser naccarii). *Aquaculture*, 321(3-4), pp.223–229.

Brockman, W. et al., 2008. Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Research*, 18, pp.763–770.

Bronzi, P. et al., 1994. Sturgeon distribution in Italy. Presentation at the International Conference on Sturgeon Biodiversity and Conservation. In the American Museum of Natural History, New York, USA.

Bronzi, P., Rosenthal, H. & Gessner, J., 2011. Global sturgeon aquaculture production: an overview. *Journal of Applied Ichthyology*, 27(2), pp.169–175.

Camacho, C. et al., 2009. BLAST+: architecture and applications. *BMC Bioinformatics*, 10, p.421.

Cao, H. et al., 2011. EST dataset of pituitary and identification of somatolactin and novel

genes in Chinese sturgeon, Acipenser sinensis. *Molecular Biology Reports*.

Carmona, R., 2009. *Biology, Conservation and Sustainable Development of Sturgeons*, Springer.

Chao, A., 1989. Estimating Population Size for Sparse Data in Capture-Recapture Experiments. *Biometrics*, 45(2), pp.427–438.

Cheung, F. et al., 2006. Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics*, 7, p.272.

Chevreux, B., Sequence assembly with MIRA3. The Definitive Guide. Available at: http://mira-assembler.sourceforge.net/docs/DefinitiveGuideToMIRA.html.

Chevreux, B. et al., 2004. Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. *Genome Research*, 14(6), pp.1147 –1159.

Chicca, M. et al., 2002. Karyotype characterization of the stellate sturgeon, Acipenser stellatus by chromosome banding and fluorescent in situ hybridization. *Journal of Applied Ichthyology*, 18(4-6), pp.298–300.

Congiu, Leonardo et al., 2011a. Managing polyploidy in ex situ conservation genetics: the case of the critically endangered Adriatic sturgeon (Acipenser naccarii). *PloS one*, 6(3), p.e18249.

Congiu, Leonardo et al., 2011b. Managing Polyploidy in Ex Situ Conservation Genetics: The Case of the Critically Endangered Adriatic Sturgeon (Acipenser naccarii). *PLoS ONE*, 6(3)

Coppe, A. et al., 2010. Sequencing, de novo annotation and analysis of the first Anguilla anguilla transcriptome: EeelBase opens new perspectives for the study of the critically endangered European eel. *BMC Genomics*, 11, p.635.

Crowhurst, R.N. et al., 2008. Analysis of expressed sequence tags from Actinidia: applications of a cross species EST database for gene discovery in the areas of flavor, health, color and ripening. *BMC Genomics*, 9(1), p.351.

Doukakis, P. et al., 2002. Molecular genetic analysis among subspecies of two Eurasian sturgeon species, Acipenser baerii and A. stellatus. *Molecular Ecology*, 8(s1), pp.S117–S127.

Van Eenennaam, A.L. et al., 1999a. Evidence of female heterogametic genetic sex determination in white sturgeon. *Journal of Heredity*, 90(1), pp.231–233.

Van Eenennaam, A.L. et al., 1999b. Evidence of female heterogametic genetic sex determination in white sturgeon. *Journal of Heredity*, 90(1), pp.231–233.

Ewing, B. & Green, P., 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature genetics*, 25(2), pp.232–234.

Ferguson-Smith, M., 2007. The Evolution of Sex Chromosomes and Sex Determination in

Vertebrates and the Key Role of *DMRT1*. *Sexual Development*, 1, pp.2–11.

Finn, R.D. et al., 2009. The Pfam protein families database. *Nucleic Acids Research*, 38(Database), pp.D211–D222.

Finotello, F. et al., 2012. Comparative Analysis of Algorithms for Whole-Genome Assembly of Pyrosequencing Data. *Briefings in Bioinformatics*, 13(3), pp.269–280.

Fontana, F, Tagliavini, J. & Congiu, L, 2001. Sturgeon genetics and cytogenetics: recent advancements and perspectives. *Genetica*, 111(1-3), pp.359–373.

Fontana, F., Lanfredi, M., et al., 2008. Comparison of karyotypes of Acipenser oxyrinchus and A. sturio by chromosome banding and fluorescent in situ hybridization. *Genetica*, 132(3), pp.281–286.

Fontana, F., Congiu, L., et al., 2008. Evidence of hexaploid karyotype in shortnose sturgeon. *Genome*, 51(2), pp.113–119.

Force, A. et al., 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4), pp.1531–1545.

Franssen, S.U. et al., 2011. Comprehensive transcriptome analysis of the highly complex Pisum sativum genome using next generation sequencing. *BMC Genomics*, 12, p.227.

Fu, L. et al., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28(23), pp.3150–3152.

Garrison, E. & Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*

Gilles, A. et al., 2011. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, 12, p.245.

Giovannini, G. et al., 1991. Growth of hatchery produced juveniles of Italian sturgeon, Acipenser naccarii Bonaparte, reared intensively in fresh water. *P. Williot (ed.) Acipenser, Cemagref Publ., Bordeaux*, pp.401–404.

Glenn, T.C., 2011. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*.

Götz, S. et al., 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, 36(10), pp.3420–3435.

Grandi, G. & Chicca, Milvia, 2008. Histological and ultrastructural investigation of early gonad development and sex differentiation in Adriatic sturgeon (Acipenser naccarii, Acipenseriformes, Chondrostei). *Journal of Morphology*, 269(10), pp.1238–1262.

Hale, M.C. et al., 2009. Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (Acipenser fulvescens): the relative merits of normalization and rarefaction in gene discovery. *BMC Genomics*, 10, p.203.

Hale, M.C., Jackson, J.R. & DeWoody, J. Andrew, 2010. Discovery and evaluation of candidate sex-determining genes and xenobiotics in the gonads of lake sturgeon (Acipenser fulvescens). *Genetica*, 138, pp.745–756.

Hett, A. K. et al., 2005. Characterization of Sox9 in European Atlantic Sturgeon (Acipenser sturio). *Journal of Heredity*, 96(2), pp.150–154.

Hett, Anne Kathrin & Ludwig, Arne, 2005. SRY-related (Sox) genes in the genome of European Atlantic sturgeon (Acipenser sturio). *Genome / National Research Council Canada*, 48(2), pp.181–186.

Huang, X. & Madan, A., 1999. CAP3: A DNA Sequence Assembly Program. *Genome Research*, 9(9), pp.868–877.

Huson, D.H. et al., 2011. Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 21(9), pp.1552–1560.

Jaillon, O. et al., 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431(7011), pp.946–957.

Kai, W. et al., 2011. Integration of the Genetic Map and Genome Assembly of Fugu Facilitates Insights into Distinct Features of Genome Evolution in Teleosts and Mammals. *Genome Biology and Evolution*, 3(0), pp.424–442.

Kasahara, M, 2007. The 2R hypothesis: an update. *Current Opinion in Immunology*, 19(5), pp.547–552.

Kasahara, Masahiro et al., 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 447(7145), pp.714–719.

Katoh, K. & Frith, M.C., 2012. Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics*.

Kaur, S., Francki, M.G. & Forster, J.W., 2011. Identification, characterization and interpretation of single-nucleotide sequence variation in allopolyploid crop species. *Plant biotechnology journal*, 10(2), pp.125–138.

Keyvanshokooh, S., Pourkazemi, M. & Kalbassi, M.R., 2007. The RAPD technique failed to identify sex-specific sequences in beluga (Huso huso). *Journal of Applied Ichthyology*, 23(1), pp.1–2.

Keyvanshokooh, Saeed et al., 2009. Comparative proteomics analysis of male and female Persian sturgeon (Acipenser persicus) gonads. *Animal Reproduction Science*, 111(2-4), pp.361–368.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2), pp.111–120.

Krieger, J. & Fuerst, P.A., 2002. Evidence for a slowed rate of molecular evolution in the order acipenseriformes. *Molecular biology and evolution*, 19(6), pp.891–897.

Kumar, S. & Blaxter, M.L., 2010. Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics*, 11(1), p.571.

Lazzari, B. et al., 2008. A comparative gene index for the white sturgeon Acipenser transmontanus. *Marine Genomics*, 1(1), pp.15–21.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), pp.2078–2079.

Liu, L. et al., 2012. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, 2012, pp.1–11.

Ludwig, A et al., 2001. Genome duplication events and functional reduction of ploidy levels in sturgeon (Acipenser, Huso and Scaphirhynchus). *Genetics*, 158(3), pp.1203–1215.

Ludwig, A., 2008. Identification of Acipenseriformes species in trade. *Journal of Applied Ichthyology*, 24, pp.2–19.

MacCallum, I. et al., 2009. ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biology*, 10(10), p.R103.

Martin, J.A. & Wang, Z., 2011. Next-generation transcriptome assembly. *Nat Rev Genet*, 12(10), pp.671–682.

McClelland, K., Bowles, J. & Koopman, P., 2012. Male sex determination: insights into molecular mechanisms. *Asian Journal of Andrology*, 14(1), pp.164–171.

McCormick, C. R., Bos, D.H. & DeWoody, J. A., 2008. Multiple molecular approaches yield no evidence for sex-determining genes in lake sturgeon (Acipenser fulvescens). *Journal of Applied Ichthyology*.

Metzker, M.L., 2010. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1), pp.31–46.

Meyer, E. et al., 2009. Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics*, 10, p.219.

Miller, J.R., Koren, S. & Sutton, Granger, 2010. Assembly Algorithms for Next-Generation Sequencing Data. *Genomics*, 95(6), pp.315–327.

Mudunuri, S.B. et al., 2010. Comparative analysis of microsatellite detecting software: a significant variation in results and influence of parameters. In *Proceedings of the International Symposium on Biocomputing*. p. 38.

Nawrocki, E.P., Kolbe, D.L. & Eddy, Sean R., 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25(10), pp.1335–1337.

Nielsen, R. et al., 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*, 12(6), pp.443–451.

Ning, Z., Cox, A.J. & Mullikin, J.C., 2001. SSAHA: a fast search method for large DNA

databases. *Genome Research*, 11(10), pp.1725–1729.

Pala, I. et al., 2009. Sex Determination in the Squalius alburnoides Complex: An Initial Characterization of Sex Cascade Elements in the Context of a Hybrid Polyploid Genome I. Dworkin, ed. *PLoS ONE*, 4, p.e6401.

Pang, K.C., Frith, M.C. & Mattick, J.S., 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends in Genetics*, 22(1), pp.1–5.

Papanicolaou, A. et al., 2009. Next generation transcriptomes for next generation genomes using est2assembly. *BMC Bioinformatics*, 10(1), p.447.

Pauchet, Y. et al., 2010. Pyrosequencing the Manduca sexta larval midgut transcriptome: messages for digestion, detoxification and defence. *Insect Molecular Biology*, 19(1), pp.61–75.

Peng, H., Zhang, J. & Wu, X., 2008. The ploidy effects in plant gene expression: Progress, problems and prospects. *Science in China Series C: Life Sciences*, 51(4), pp.295–301.

Postlethwait, J.H. et al., 2000. Zebrafish Comparative Genomics and the Origins of Vertebrate Chromosomes. *Genome Research*, 10(12), pp.1890–1902.

Pujolar, José Martin et al., 2012. Tana1, a new putatively active Tc1-like transposable element in the genome of sturgeons. *Molecular Phylogenetics and Evolution*.

Quinlan, A.R. & Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6), pp.841–842.

Rorbach, J. & Minczuk, M., 2012. The post-transcriptional life of mammalian mitochondrial RNA. *Biochemical Journal*, 444(3), pp.357–373.

Sales, G., 2008. *Functional sequence detection using whole-genome paralogous alignments*. Ph. D. thesis. Università degli studi di Torino.

Schwartz, T.S. et al., 2010. A garter snake transcriptome: pyrosequencing, de novo assembly, and sex-specific differences. *BMC Genomics*, 11, p.694.

Shirak, A. et al., 2006. Amh and Dmrta2 Genes Map to Tilapia (Oreochromis spp.) Linkage Group 23 Within Quantitative Trait Locus Regions for Sex Determination. *Genetics*, 174, pp.1573–1581.

Sturgeons, P. by the participants of the 5th I.S. on, 2006. Ramsar Declaration on Global Sturgeon Conservation. *Journal of Applied Ichthyology*, 22, pp.5–12.

Sutton, GG et al., 1995. TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science & Technology*, 1, pp.9–19.

Thiel, T. et al., 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theoretical and Applied Genetics*, 106(3), pp.411–422.

Trick, M. et al., 2012. Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. *BMC Plant Biology*, 12(1), p.14.

Trick, M. et al., 2009. Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing. *Plant Biotechnology Journal*, 7(4), pp.334–346.

Vera, J.C. et al., 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, 17(7), pp.1636–1647.

Viegas, C.S.B. et al., 2008. Gla-rich Protein (GRP), A New Vitamin K-dependent Protein Identified from Sturgeon Cartilage and Highly Conserved in Vertebrates. *Journal of Biological Chemistry*, 283, pp.36655–36664.

Volff, J.N., 2004. Genome evolution and biodiversity in teleost fish. *Heredity*, 94(3), pp.280–294.

Wang, Dengqiang et al., 2010. Evolution of MHC class I genes in two ancient fish, paddlefish (Polyodon spathula) and Chinese sturgeon (Acipenser sinensis). *FEBS Letters*, 584(15), pp.3331–3339.

Wang, J.-T. et al., 2012. Transcriptome analysis reveals the time of the fourth round of genome duplication in common carp (Cyprinus carpio). *BMC Genomics*, 13(1), p.96.

Wendel, J.F., 2000. Genome evolution in polyploids. *Plant molecular biology*, 42(1), pp.225–249.

Wicker, T. et al., 2006. 454 sequencing put to the test using the complex genome of barley. *BMC Genomics*, 7, p.275.

William Bemis, E.F., 2001. An overview of Acipenseriformes. , pp.25–71.

Woods, I. G., 2005. The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Research*, 15(9), pp.1307–1314.

Wuertz, S. et al., 2006. Extensive screening of sturgeon genomes by random screening techniques revealed no sex-specific marker. *Aquaculture*, 258(1–4), pp.685–688.

Wuertz, S., Belay, M. & Kirschbaum, Frank, 2007. On the risk of criminal manipulation in caviar trade by intended contamination of caviar with PCR products. *Aquaculture*, 269(1–4), pp.130–134.

Yang, S. et al., 2010. Paradigm for industrial strain improvement identifies sodium acetate tolerance loci in Zymomonas mobilis and Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 107(23), pp.10395–10400.

Yarmohammadi, M. et al., 2012. AFLP reveals no sex-specific markers in Persian sturgeon (Acipenser persicus) or beluga sturgeon (Huso huso) from the southern Caspian Sea, Iran. *Progress in Biological Sciences*, 1(1), pp.55–114.

Zerbino, D.R. & Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), pp.821–829.

Zhulidov, P.A. et al., 2004. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Research*, 32(3), p.e37.

APPENDIX A

pairwise relationships between main properties of contigs

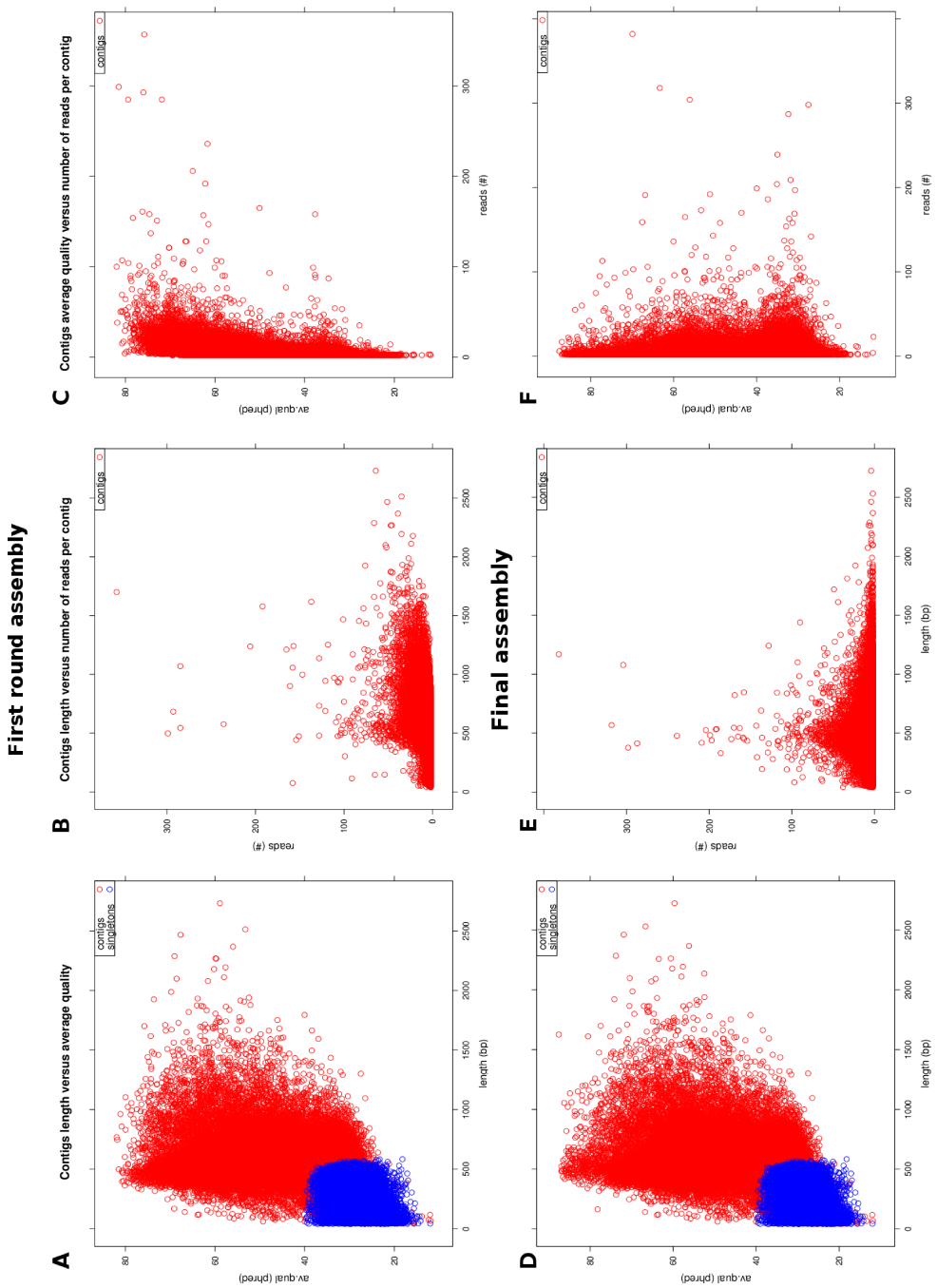

Figure 1. Mean contigs coverage distribution for the first and final assemblies (*A. naccarii*).

The average coverage of the contigs is quite low. As shown on the graph, about 61% of contigs have per base average coverage up to 3, while 93% have per base coverage up to 9. This may be due to the high ploidy in *A. naccari*, believed to be tetraploid. Thus, the numerous alleles present, which are kept apart by MIRA, were sequenced to low coverage.

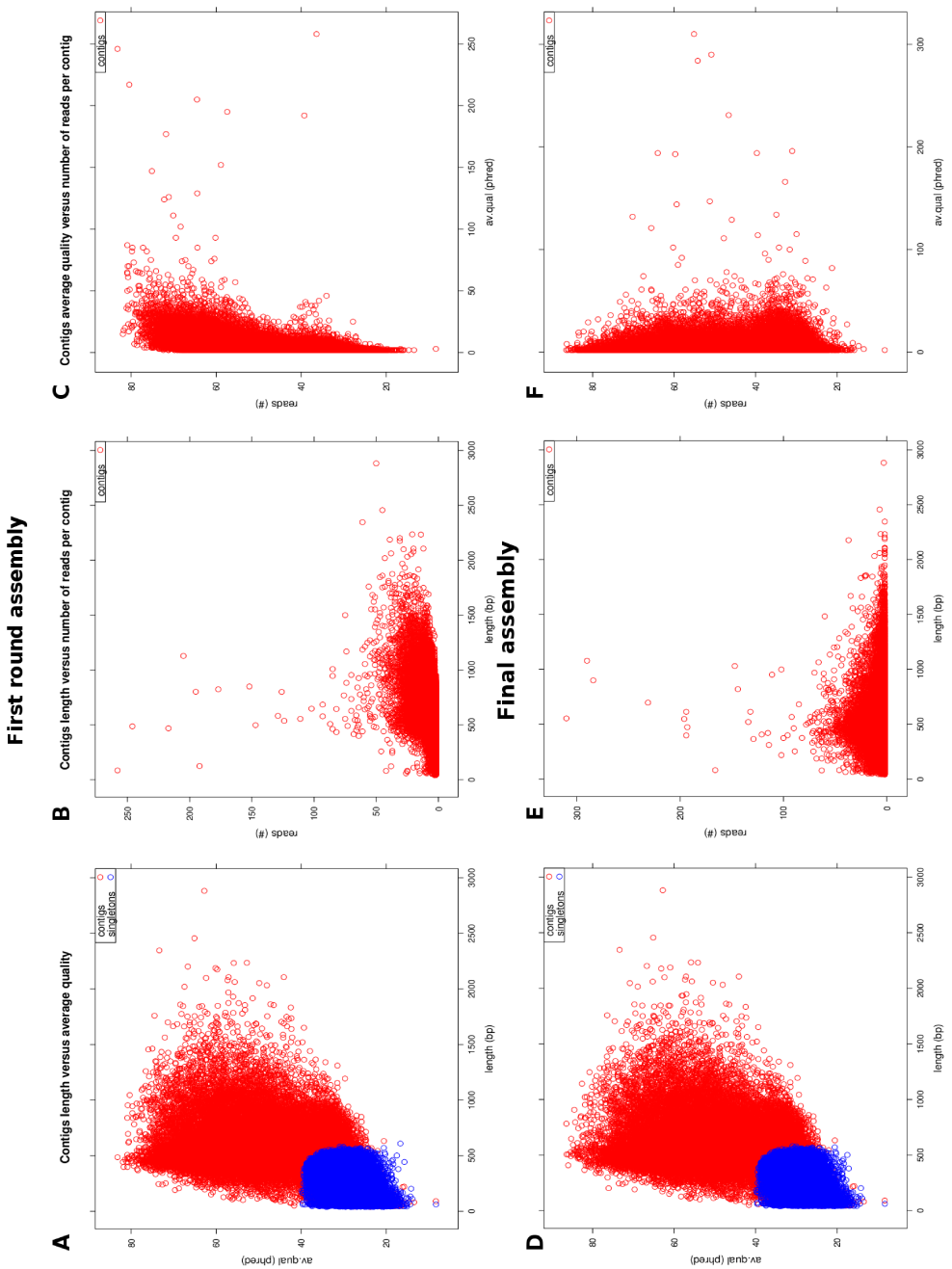

Figure 2. Pair-wise relationships between main properties characterising total contigs obtained by the first and final assemblies (*A. stellatus*).

APPENDIX B

KEGG pathways

Active metabolic KEGG pathways identified in the *A. naccarii* transcriptome		
Different enzymes per pathway (#)	**Pathway name**	**Pathway map ID**
1	Betalain biosynthesis	path:map00965
1	Biosynthesis of ansamycins	path:map01051
1	Biosynthesis of vancomycin group antibiotics	path:map01055
1	Bisphenol degradation	path:map00363
1	C5-Branched dibasic acid metabolism	path:map00660
1	Caffeine metabolism	path:map00232
1	D-Arginine and D-ornithine metabolism	path:map00472
1	D-Glutamine and D-glutamate metabolism	path:map00471
1	Ethylbenzene degradation	path:map00642
1	Glycosaminoglycan biosynthesis - keratan sulfate	path:map00533
1	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	path:map00563
1	Indole alkaloid biosynthesis	path:map00901
1	Lipopolysaccharide biosynthesis	path:map00540
1	Lysine biosynthesis	path:map00300
1	Melanogenesis	path:map04916
1	Mucin type O-Glycan biosynthesis	path:map00512
1	Naphthalene degradation	path:map00626
1	Nitrotoluene degradation	path:map00633
1	Penicillin and cephalosporin biosynthesis	path:map00311
1	Photosynthesis	path:map00195
1	Polyketide sugar unit biosynthesis	path:map00523
1	Vitamin B6 metabolism	path:map00750
2	Butirosin and neomycin biosynthesis	path:map00524
2	Cutin, suberine and wax biosynthesis	path:map00073
2	Drug metabolism - cytochrome P450	path:map00982
2	Fatty acid biosynthesis	path:map00061
2	Limonene and pinene degradation	path:map00903
2	Linoleic acid metabolism	path:map00591
2	Other types of O-glycan biosynthesis	path:map00514
2	Phenylalanine, tyrosine and tryptophan biosynthesis	path:map00400
2	Phenylpropanoid biosynthesis	path:map00940
2	Phosphonate and phosphinate metabolism	path:map00440
2	T cell receptor signaling pathway	path:map04660
2	Thiamine metabolism	path:map00730
2	Toluene degradation	path:map00623
2	Ubiquinone and other terpenoid-quinone biosynthesis	path:map00130
2	Valine, leucine and isoleucine biosynthesis	path:map00290

3	Caprolactam degradation	path:map00930
3	Cyanoamino acid metabolism	path:map00460
3	Geraniol degradation	path:map00281
3	Glycosaminoglycan degradation	path:map00531
3	Glycosphingolipid biosynthesis - ganglio series	path:map00604
3	Glycosphingolipid biosynthesis - globo series	path:map00603
3	Glycosphingolipid biosynthesis - lacto and neolacto series	path:map00601
3	Riboflavin metabolism	path:map00740
3	Styrene degradation	path:map00643
3	Synthesis and degradation of ketone bodies	path:map00072
3	Taurine and hypotaurine metabolism	path:map00430
3	mTOR signaling pathway	path:map04150
4	Benzoate degradation	path:map00362
4	Chloroalkane and chloroalkene degradation	path:map00625
4	Glycosaminoglycan biosynthesis - chondroitin sulfate	path:map00532
4	Glycosaminoglycan biosynthesis - heparan sulfate	path:map00534
4	Metabolism of xenobiotics by cytochrome P450	path:map00980
4	Other glycan degradation	path:map00511
5	Aminobenzoate degradation	path:map00627
5	Folate biosynthesis	path:map00790
5	Primary bile acid biosynthesis	path:map00120
5	Streptomycin biosynthesis	path:map00521
5	alpha-Linolenic acid metabolism	path:map00592
6	Biosynthesis of unsaturated fatty acids	path:map01040
6	Ether lipid metabolism	path:map00565
6	Histidine metabolism	path:map00340
6	Nicotinate and nicotinamide metabolism	path:map00760
6	One carbon pool by folate	path:map00670
6	Phenylalanine metabolism	path:map00360
6	Retinol metabolism	path:map00830
6	Selenocompound metabolism	path:map00450
6	Steroid biosynthesis	path:map00100
6	Sulfur metabolism	path:map00920
7	Ascorbate and aldarate metabolism	path:map00053
7	Porphyrin and chlorophyll metabolism	path:map00860
8	Fatty acid elongation	path:map00062
8	Pantothenate and CoA biosynthesis	path:map00770
8	Pentose and glucuronate interconversions	path:map00040
8	Terpenoid backbone biosynthesis	path:map00900
9	Carbon fixation in photosynthetic organisms	path:map00710
9	Glyoxylate and dicarboxylate metabolism	path:map00630

9	Nitrogen metabolism	path:map00910
9	Oxidative phosphorylation	path:map00190
9	Tyrosine metabolism	path:map00350
9	Various types of N-glycan biosynthesis	path:map00513
10	Butanoate metabolism	path:map00650
10	Drug metabolism - other enzymes	path:map00983
10	Sphingolipid metabolism	path:map00600
10	Steroid hormone biosynthesis	path:map00140
10	Tryptophan metabolism	path:map00380
11	N-Glycan biosynthesis	path:map00510
12	Carbon fixation pathways in prokaryotes	path:map00720
12	Galactose metabolism	path:map00052
13	Arachidonic acid metabolism	path:map00590
13	Glycerolipid metabolism	path:map00561
13	Pentose phosphate pathway	path:map00030
13	beta-Alanine metabolism	path:map00410
14	Lysine degradation	path:map00310
15	Citrate cycle (TCA cycle)	path:map00020
15	Fatty acid metabolism	path:map00071
15	Fructose and mannose metabolism	path:map00051
15	Propanoate metabolism	path:map00640
15	Starch and sucrose metabolism	path:map00500
16	Alanine, aspartate and glutamate metabolism	path:map00250
16	Cysteine and methionine metabolism	path:map00270
16	Glutathione metabolism	path:map00480
16	Glycerophospholipid metabolism	path:map00564
16	Methane metabolism	path:map00680
16	Phosphatidylinositol signaling system	path:map04070
17	Pyruvate metabolism	path:map00620
18	Aminoacyl-tRNA biosynthesis	path:map00970
18	Inositol phosphate metabolism	path:map00562
19	Pyrimidine metabolism	path:map00240
19	Valine, leucine and isoleucine degradation	path:map00280
21	Glycine, serine and threonine metabolism	path:map00260
22	Amino sugar and nucleotide sugar metabolism	path:map00520
22	Glycolysis / Gluconeogenesis	path:map00010
25	Arginine and proline metabolism	path:map00330
33	Purine metabolism	path:map00230

Table 1. KEGG pathways found in the *A. naccarii* transcriptome.

Enzyme Codes were mapped on the sequences with GO annotations through Blast2GO, then metabolic pathway map numbers in which enzymes carry out their function were retrieved, thus identifying 833 different enzymes participating in 116 different pathways.

Active metabolic KEGG pathways identified in the *A. stellatus* transcriptome		
Different enzymes per pathway (#)	**Pathway name**	**Pathway map ID**
1	Biosynthesis of ansamycins	path:map01051
1	Biosynthesis of siderophore group nonribosomal peptides	path:map01053
1	Biosynthesis of vancomycin group antibiotics	path:map01055
1	Bisphenol degradation	path:map00363
1	Cutin, suberine and wax biosynthesis	path:map00073
1	D-Arginine and D-ornithine metabolism	path:map00472
1	D-Glutamine and D-glutamate metabolism	path:map00471
1	Ethylbenzene degradation	path:map00642
1	Flavone and flavonol biosynthesis	path:map00944
1	Flavonoid biosynthesis	path:map00941
1	Glycosaminoglycan biosynthesis - keratan sulfate	path:map00533
1	Glycosphingolipid biosynthesi	
1	Indole alkaloid biosynthesis	path:map00901
1	Limonene and pinene degradation	path:map00903
1	Melanogenesis	path:map04916
1	Naphthalene degradation	path:map00626
1	Nitrotoluene degradation	path:map00633
1	Novobiocin biosynthesis	path:map00401
1	PI3K-Akt signaling pathway	path:map04151
1	Penicillin and cephalosporin biosynthesis	path:map00311
1	Polyketide sugar unit biosynthesis	path:map00523
1	Stilbenoid, diarylheptanoid and gingerol biosynthesis	path:map00945
1	Tetracycline biosynthesis	path:map00253
1	Vitamin B6 metabolism	path:map00750
2	Betalain biosynthesis	path:map00965
2	Butirosin and neomycin biosynthesis	path:map00524
2	C5-Branched dibasic acid metabolism	path:map00660
2	Caffeine metabolism	path:map00232
2	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	path:map00563
2	Linoleic acid metabolism	path:map00591
2	Peptidoglycan biosynthesis	path:map00550
2	Phenylpropanoid biosynthesis	path:map00940
2	Steroid degradation	path:map00984
2	T cell receptor signaling pathway	path:map04660
2	Toluene degradation	path:map00623
2	Tropane, piperidine and pyridine alkaloid	path:map00960

	biosynthesis	
2	Valine, leucine and isoleucine biosynthesis	path:map00290
3	Benzoate degradation	path:map00362
3	Caprolactam degradation	path:map00930
3	Chloroalkane and chloroalkene degradation	path:map00625
3	Geraniol degradation	path:map00281
3	Mucin type O-Glycan biosynthesis	path:map00512
3	Other types of O-glycan biosynthesis	path:map00514
3	Phenylalanine, tyrosine and tryptophan biosynthesis	path:map00400
3	Phosphonate and phosphinate metabolism	path:map00440
3	Synthesis and degradation of ketone bodies	path:map00072
3	Thiamine metabolism	path:map00730
3	mTOR signaling pathway	path:map04150
4	Cyanoamino acid metabolism	path:map00460
4	Folate biosynthesis	path:map00790
4	Glycosaminoglycan biosynthesis - chondroitin sulfate	path:map00532
4	Glycosaminoglycan biosynthesis - heparan sulfate	path:map00534
4	Isoquinoline alkaloid biosynthesis	path:map00950
4	Lysine biosynthesis	path:map00300
4	Riboflavin metabolism	path:map00740
4	Ubiquinone and other terpenoid-quinone biosynthesis	path:map00130
4	alpha-Linolenic acid metabolism	path:map00592
5	Drug metabolism - cytochrome P450	path:map00982
5	Fatty acid biosynthesis	path:map00061
5	Histidine metabolism	path:map00340
5	Primary bile acid biosynthesis	path:map00120
5	Steroid biosynthesis	path:map00100
5	Streptomycin biosynthesis	path:map00521
5	Styrene degradation	path:map00643
5	Taurine and hypotaurine metabolism	path:map00430
6	Ascorbate and aldarate metabolism	path:map00053
6	Glycosaminoglycan degradation	path:map00531
6	Metabolism of xenobiotics by cytochrome P450	path:map00980
6	Other glycan degradation	path:map00511
6	Selenocompound metabolism	path:map00450
6	Terpenoid backbone biosynthesis	path:map00900
7	Aminobenzoate degradation	path:map00627
7	Biosynthesis of unsaturated fatty acids	path:map01040
8	Fatty acid elongation	path:map00062
8	Oxidative phosphorylation	path:map00190

8	Pantothenate and CoA biosynthesis	path:map00770
8	Retinol metabolism	path:map00830
8	Various types of N-glycan biosynthesis	path:map00513
9	N-Glycan biosynthesis	path:map00510
9	Pentose and glucuronate interconversions	path:map00040
10	Ether lipid metabolism	path:map00565
10	Glycerolipid metabolism	path:map00561
10	Sulfur metabolism	path:map00920
10	beta-Alanine metabolism	path:map00410
11	Phenylalanine metabolism	path:map00360
11	Tryptophan metabolism	path:map00380
12	Carbon fixation in photosynthetic organisms	path:map00710
12	Fatty acid metabolism	path:map00071
12	Galactose metabolism	path:map00052
13	Arachidonic acid metabolism	path:map00590
13	Butanoate metabolism	path:map00650
13	Drug metabolism - other enzymes	path:map00983
13	Lysine degradation	path:map00310
13	Nitrogen metabolism	path:map00910
13	One carbon pool by folate	path:map00670
13	Porphyrin and chlorophyll metabolism	path:map00860
14	Tyrosine metabolism	path:map00350
16	Glutathione metabolism	path:map00480
16	Glyoxylate and dicarboxylate metabolism	path:map00630
16	Propanoate metabolism	path:map00640
16	Valine, leucine and isoleucine degradation	path:map00280
17	Starch and sucrose metabolism	path:map00500
17	Steroid hormone biosynthesis	path:map00140
18	Carbon fixation pathways in prokaryotes	path:map00720
18	Citrate cycle (TCA cycle)	path:map00020
18	Fructose and mannose metabolism	path:map00051
18	Methane metabolism	path:map00680
18	Pentose phosphate pathway	path:map00030
18	Phosphatidylinositol signaling system	path:map04070
19	Glycerophospholipid metabolism	path:map00564
19	Inositol phosphate metabolism	path:map00562
20	Alanine, aspartate and glutamate metabolism	path:map00250
20	Aminoacyl-tRNA biosynthesis	path:map00970
20	Pyruvate metabolism	path:map00620
21	Pyrimidine metabolism	path:map00240
22	Glycolysis / Gluconeogenesis	path:map00010
24	Glycine, serine and threonine metabolism	path:map00260
25	Arginine and proline metabolism	path:map00330

26	Amino sugar and nucleotide sugar metabolism	path:map00520
47	Purine metabolism	path:map00230

Table 2. KEGG pathways found in *A. stellatus* transcriptomes.

APPENDIX C

A. naccarii sex determining genes

Table 1. Sex related genes found in the Adriatic sturgeon transcriptome.

The table lists the contigs of the *A. naccarii* transcriptome that best represent 22 of the 32 genes known to be involved in sex determination and sexual development of vertebrates, used for screening. For each recognized gene are shown: the gene full name; the cluster ID representing it in NCBI HomoloGene; the contig that best represents the putative *A. naccarii* orthologous (subject); the assembly fraction the contig belongs to (cDNA3, cDNA4, or common); its mean per-base coverage; the query that caught it; its alignment bit score; the putative *Pfam* domains contained within the translated and aligned fractions of the contig on the query; the contig translated-aligned fractions; and its blast2GO annotation.

Gene Symbol	Gene Name	Homologene ID	Best subject (contig)	Assembly fraction	Contig per-base mean coverage	Best query		Bit score	Domains found (pfam)	Subject aligned fraction	*blast2GO* annotation
WT1	Wilms tumor 1	11536	CDNA3-4_11_2010_0_c27731	CDNA1	2.17	Ensemble	translation:ENSDARP00000097261	124.02	4 Zinc-finger double domains (zf-H2C2_2)	0.68	zinc finger protein 502
LHX1	LIM homeobox 1	4068	CDNA3-4_11_2010_0_c35420	CDNA1	1.59	Homologene	gi:359320451	96.671	2 LIM domains	0.58	lim domain only 1 (rhombotin 1)
						Ensembl	translation:ENSDARP00000044207	278.485	2 LIM domains	0.67	
CYP19A1 (aromatase)	cytochrome P450, family 19, subfamily A, polypeptide 1	30955	CDNA3-4_11_2010_0_c38303	CDNA1	1.83	Homologene	gi:183583540	110.153	Cytochrome P450 domain	0.5	cytochrome p450 aromatase
						Ensembl	translation:ENSTNIP00000016690	117.857	Cytochrome P450 domain	0.5	
FHL3	four and a half LIM domains 3	37928	CDNA3-4_11_2010_0_c27460	CDNA1	1.8	Homologene	gi:345780505	77.026	3 LIM domains	0.79	testis derived transcript
						Ensembl	translation:F25H5.1e	139.813	3 LIM domains	0.85	
FEM1A	fem-1 homolog a (C. elegans)	7713	CDNA3-4_11_2010_0_c40985	CDNA1	1.54	Homologene	gi:114052839	60.847	Ank_2 Ankyrin repeats (3 copies) Family CL0465	0.64	/
						Ensembl	translation:ENSDARP00000006	180.259	Ank_2 Ankyrin repeats (3 copies) Family CL0465	0.99	
AR	androgen receptor	28	CDNA3-4_11_2010_0_c22553	CDNA2	2	Homologene	gi:346421335	135.961	2 Zinc finger, C4 type (two domains)	0.88	glucocorticoid receptor
						Ensembl	translation:ENSDARP00000054263	197.593	2 Zinc finger, C4 type (two domains)	0.88	
EMX2	empty spiracles homeobox 2	3023	CDNA3-4_11_2010_0_rep_c5928	CDNA2	4.12	Homologene	gi:118093051	107.071	1Homeobox domain CL0123	0.26	empty spiracles homolog 1
						Ensembl	translation:ENSDARP00000056747	126.331	1Homeobox domain CL0123	0.26	
DAX1	Nr0b1, nuclear receptor subfamily 0, group B, member 1	403	CDNA3-4_11_2010_0_c36879	common	1.32	Homologene	gi:46048923	56.225	Hormone_recep Ligand-binding domain of nuclear hormone receptor	0.46	nuclear receptor subfamily 2 group c

						Ensembl	translation:ENSLACP00000007913	60.077	Hormone_recep Ligand-binding domain of nuclear hormone receptor	0.45	member 1
SOX9	Sry box-containing gene 9	294	CDNA3-4_11_2010_0_c20466	common	1.69	Homologene	gi:18859409	156.377	HMG (high mobility group) box	0.34	transcription factor sox9
						Ensembl	translation:ENSGACP00000007	161.384	HMG (high mobility group) box	0.33	
						Acipenser-gene	gi:51599118	444.316	HMG (high mobility group) box	0.32	
SOX17	SOX17 SRY (sex determining region Y)-box 17	7948	CDNA3-4_11_2010_0_c37628	common	1.31	Ensembl	translation:ENSLACP00000017352	148.673	Sox C-terminal transactivation domain	0.61	sry-box containing gene 17
SOX1	SRY (sex determining region Y)-box 1	48390	CDNA3-4_11_2010_0_rep_c32497	common	1.79	Ensembl	translation:ENSDARP00000092797	120.168	/	0.49	sry-box containing gene 1a
SOX11	sox11a SRY-box containing gene 11	37733	CDNA3-4_11_2010_0_c20816	common	2.36	Homologene	gi:62234366	195.282	HMG (high mobility group) box Domain CL0114	0.61	sry-box containing gene 11b
						Ensembl	translation:ENSLACP00000016790	206.068	HMG (high mobility group) box Domain CL0114	0.78	
						Acipenser-gene	gi:62240493	316.897	HMG (high mobility group) box Domain CL0114	0.26	
SOX6	SOX6 SRY (sex determining region Y)-box 6	22631	CDNA3-4_11_2010_0_c32494	common	1.86	Homologene	gi:363734209	186.808	HMG (high mobility group) box Domain CL0114	0.97	transcription factor sox-6 isoform 3
						Ensembl	translation:ENSAMEP00000015921	185.652	HMG_box HMG (high mobility group) box Domain CL0114	0.97	
SOX14	SOX14 SRY (sex determining region Y)-box 14	31224	CDNA3-4_11_2010_0_c17363	common	3.31	Homologene	gi:45382127	139.043	/	0.61	sry (sex determining region y)-box 14
						Ensembl	translation:ENSMUSP00000129906	138.658	/	0.31	
FOXL2	forkhead box L2 (FOXL2)	74992	CDNA3-4_11_2010_0_c10724	common	4.21	Ensembl	translation:ENSDARP00000105372	362.073	Fork head domain	0.81	forkhead box c1
RSPO	R-spondin homologue 1	52148	CDNA3-4_11_2010_0_c4644	common	6.61	Homologene	gi:363742268	205.297	VSP Giardia variant-specific surface protein	0.83	r-spondin homolog (xenopus laevis)
						Ensembl	translation:ENSLACP00000019250	218.779	VSP Giardia variant-specific surface protein	0.83	
SF1	Steroidogenic Factor 1 (nuclear receptor subfamily 5, group A, member 1)	3638	CDNA3-4_11_2010_0_c10468	common	3.61	Ensembl	translation:ENSSSCP00000011633	58.921	Ligand-binding domain of nuclear hormone receptor	0.57	nuclear receptor subfamily group member 2
FGFR2	fibroblast growth factor receptor 2	22566	CDNA3-4_11_2010_0_rep_c6815	common	3.75	Homologene	gi:158291261	157.532	Protein tyrosine kinase	0.86	novel protein vertebrate abelson murine leukemia viral oncogene homolog 2 (abelson-related gene)
						Ensembl	translation:ENSDARP00000124224	176.792	Protein tyrosine kinase	0.9	

FGF9	fibroblast growth factor 9	1523	CDNA3-4_11_2010_0_rep_c9232	common	5.97	Ensembl	translation:ENSSHAP00000005334	133.65	Fibroblast growth factor	0.41	fibroblast growth factor 12
GATA4	GATA-binding protein 4	1551	CDNA3-4_11_2010_0_rep_c278	common	11.51	Ensembl	translation:ENSGMOP00000011259	362.844	GATA zinc finger	0.53	gata zinc finger domain-containing protein 1
LHX9	LIM homeobox 9	7816	CDNA3-4_11_2010_0_c17973	common	3.14	Homologene	gi:66792874	117.857	Homeobox domain	0.22	lim homeobox 9
						Ensembl	translation:ENSDARP00000074010	154.836	Homeobox domain	0.29	
ATRX	alpha thalassemia/mental retardation syndrome X-linked	416	CDNA3-4_11_2010_0_c17931	common	5.33	Homologene	gi:297492973	177.178	SNF2 family N-terminal domain CL0023 (PF00176)	0.51	transcriptional regulator atrx
						Ensembl	translation:ENSMODP00000004803	177.563	SNF2 family N-terminal domain CL0023 (PF00176)	0.51	