CARLOS MIGUEL ESTEVENS VIEIRA ROLO GARCIA

ASSEMBLY AND ANNOTATION OF THE SARDINE (*Sardina pilchardus*) TRANSCRIPTOME

UNIVERSIDADE DO ALGARVE

Faculdade de Ciências e Tecnologias

2018

CARLOS MIGUEL ESTEVENS VIEIRA ROLO GARCIA

ASSEMBLY AND ANNOTATION OF THE SARDINE (*Sardina pilchardus*) TRANSCRIPTOME

Mestrado em Biotecnologia

Trabalho efetuado sob a orientação de:

Drª Deborah Power

Dr Bruno Louro

UNIVERSIDADE DO ALGARVE

Faculdade de Ciências e Tecnologias

2018

ASSEMBLY AND ANNOTATION OF THE SARDINE (*Sardina pilchardus*) TRANSCRIPTOME

Declaração de autoria de trabalho

Declaro ser o autor deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

Assinatura:

Acknowledgments

I would like to thank Professor Deborah Power for the opportunity to participate on this project and for her guidance.

I would like to express my gratitude to my coordinator Doctor Bruno Louro for he’s guidance, knowledge, patience and dedication.

A thank you for my family, specially my parents who kept supporting with a special thanks to my grandmother Gracinda.

I would also like to thank my friends for encouraging me to keep going and my colleagues for distracting me when I needed to relax.

Abstract

The European sardine (*Sardina pilchardus*) is a fish of high cultural and economic importance in Portugal and current stock assessment studies report an alarming stock biomass decrease due to overfishing and/or environmental change. For better management of the sardine fisheries, there is an urgent need to understand the causal factors leading to the historically low level of the sardine stock in Portuguese waters. Important biological questions such as population diversity level, structure and migrations can be tackled with the development and usage of genomic tools. The ability to answer such important biological questions will be valuable and can be integrated into stock assessment data modelling and aid data-based policy making for better biological resource management. Eleven tissues were sequenced and curated to assemble the transcriptome. Through the comparison of different approaches, the best seemed to go through a quality control step with Trim Galore and a *de novo* assembly. A post-assembly quality control with Transrate seemed to be better when assembling a group of different tissues rather than one specific ones. The assembly with reads from all the tissues studied contained 170,478 contigs and had an N50 value of 486. Before this project almost no genomic/genetics resources existed to assist biological studies of the sardine and the species genome and transcriptome are cornerstone resources needed to translate applied scientific genetic data into management measures. In this project, a reference transcriptome of the sardine was assembled and functionally annotated.

Keywords: sardine, transcriptome, annotation, bioinformatics.

Resumo

A sardinha europeia (*Sardina pilchardus*) é um peixe de grande importância cultural e económica em Portugal e os atuais estudos de avaliação das unidades populacionais mostram uma diminuição preocupante da biomassa das unidades populacionais devido à sobrepesca e / ou alterações ambientais. Para uma melhor gestão da pesca da sardinha, existe uma necessidade urgente de compreender os fatores que levam ao baixo nível histórico do estoque de sardinha nas águas portuguesas. Questões importantes biológicas, como níveis de diversidade populacional, estrutura e migrações, podem ser abordadas com o desenvolvimento e uso de ferramentas genómicas. A capacidade de responder a essas importantes questões biológicas será valiosa e poderá ser integrada à modelagem de dados de avaliação de estoques e à criação de políticas baseadas em dados de ajuda para um melhor gerenciamento dos recursos biológicos.

Onze tecidos foram sequenciados e tratados para montar o transcriptoma. Através da comparação de diferentes abordagens, os melhores pareciam passar por uma etapa de controlo de qualidade com o Trim Galore e uma montagem *de novo*. Um controlo de qualidade pós-montagem com o Transrate parecia ser melhor quando se montava um grupo de diferentes tecidos, em vez de um único específico. A montagem com leituras de todos os tecidos estudados continha 170 478 contigs e tinha um valor de N50 de 486.

Através da comparação do controlo de qualidade executado pelo Trim Galore com o Trimmomatic, notou-se uma melhor qualidade de leituras após o Trimmomatic com pontuações de qualidade acima de 32 e percentagens de leituras removidas entre os 0,28 e 0,44 % em contraste com pontuações de qualidade de 28 e percentagens de leituras removidas entre os 5,77 e 8,08 % resultantes do Trim Galore, ambas as abordagens originaram em percentagens de guanina-citocina entre os 49 e 55 %. No entanto, devido a sequências menores do que 30 pares de base inesperadas e percentagens de leituras removidas maiores do que o esperado resultantes do Trimmomatic o projeto procedeu com as leituras resultantes do Trim Galore.

Entre as duas abordagens para a montagem do transcriptoma com o Trinity, como a montagem guiada pelo genoma originou valores de N50 mais baixos para o primeiro tecido testado nos dois métodos de alinhamento (local e de ponta-a-ponta) mais nenhum tecido foi testado e o projeto procedeu com as montagens *de novo*. As montagens *de novo* passaram por outro passo de controlo de qualidade feito pelo Transrate que reteve entre 44 e 80 % de sequências com medias de comprimento entre os 425,98 e 686,88 pares de base e valores de N50 entre os 474 e 1 039. O Transrate diminui os valores de N50, o que não era esperado, mas diminuiu também o número de contigs para um valor mais realista para os tecidos tendo assim ter sido escolhidas para a anotação as montagens de *novo* após tratadas pelo Transrate.

Através do Trinotate, entre 14,66 e 38,07 % dos contigs foram deduzidos em regiões codificadoras com TransDecoder; 25,49 a 44,77 % e 11,56 a 31,71 % dos contigs foram anotados com homologias de sequências via Sprot blastx Sprot blastp, respetivamente. Com base na sequência SwissProt ID obtida e no banco de dados SQL do Trinotate, 20,92 a 39,63 % anotados com homologias de sequências via BLAST + tiveram a anotação de Kegg, 19,70 a 39,20 % de eggNOG, 24,81 a 44,11 % de GO blast. Foram identificados 9,70 a 25,05 % de domínios proteicos com HMMER / PFAM e, consequentemente, 5,90 a 15,00 % anotados com GO com base nos domínios Pfam. No geral, o banco de dados que anotou o maior número de transcritos foi eggNOG, enquanto o que anotou o menor foi com SignalP, mostrando apenas uma pequena percentagem (1,02 a 1,94 % de peptídeos de sinal) dos transcritos representam proteínas que são secretadas a partir da célula, seguido por proteínas transmembranares identificadas com tmHMM, com 2,73 a 5,46 % de domínios transmembranares encontrados. Comparando a anotação antes das montagens passarem pelo Transrate, foram também anotadas as montagens do tecido da barbatana caudal e da montagem com todos os tecidos notando-se no geral uma diminuição de percentagem de transcritos anotados após o Transrate, o que não deveria acontecer. As isoformas dos genes foram retiradas para novos cálculos das percentagens para perceber se era o motivo da diminuição, com esta forma a percentagem de genes anotados diminuíram menos.

Uma quantificação de transcritos fornecida pelo Trinity determinou 12 747 genes e 13 732 transcritos expressos entre 10 e 100 TPM (transcritos por milhão), dos quais 26 053 genes e 28 211 transcritos são expressos por pelo menos 10 TPM.

Foram considerados entre 64 a 1189 genes específicos de tecidos dos quais foram anotados por volta de 64 % quando os genes tinham uma expressão total de 95 % nesse tecido. A anotação dos 10 genes específicos mais significantes por tecido permitiu a verificação de genes que correspondiam com a função de cada tecido e onde seriam mais expressos como também a verificação de genes duplicados. Após estes genes duplicados terem sido analisados notou-se que apenas existia uma cópia destes antes dos teleostes e entres os teleostes era possível verificar mais do que uma, confirmando assim um evento de duplicação de genoma inteiro nos teleostes.

Pelo website REVIGO foram gerados gráficos de dispersão e tabelas com GOs de processos biológicos e funções moleculares que correspondiam com a função de cada tecido para os quais foram gerados.

Antes, quase não existiam recursos genómicos / genéticos para auxiliar os estudos biológicos, e o genoma e o transcriptoma das espécies são recursos fundamentais necessários para

transformar dados genéticos científicos aplicados em manejo. Neste projeto, o transcriptoma representativo da sardinha foi montado e funcionalmente.

Palavras-chave: sardinha, transcriptoma, anotação, bioinformática.

Abbreviations

DNA: deoxyribonucleic acid
RNA: ribonucleic acid
mRNA: messenger RNA
tRNA: transfer RNA
rRNA: ribosomal RNA
miRNA: microRNA
siRNA: small interfering RNA
RNA-seq: RNA sequencing
NGS: Next Generation Sequencing
cDNA: complementary DNA
eggNOG: evolutionary genealogy of genes: Non-supervised Orthologous Groups
GO: Gene Ontology
Kegg: Kyoto enciclopedia of genes and genomes
kg: kilogram
cm: centimetres
EPPO: European and Mediterranean Plant Protection Organization
Gi: Gill + Branchial Arch
Lv: Liver
Sp: Spleen
Gn: Gonad (female)
Mg: Midgut
WM: White Muscle
RM: Red Muscle
Kd: Kidney
HKd: Head Kidney
Br: Brain + Pituitary
CF: Caudal Fin (Skin + Cartilage + Bone)
DNase: desoxyribonuclease
Poly(A): Polyadenine
PE: paired end
Prep: Preparation
RIN: RNA integrity number
M: million
R1: read 1
R2: read 2
HPC: High Performance Computing
INCD: Infraestrutura Nacional de Computação Distribuída
CCMAR: Centro de Ciências do Mar
bp: base pair
fa: FASTA
G: gigas
RAM: Random Access Memory
CPU: Central Processing Unit
fq: FASTQ
BAM: Binary Alignment Map
SAM: Sequence Alignment Map
vs: versus
PCR: Polymerase Chain Reaction
%GC: Percentage of Guanine-Cytosine content
Seqs: Sequences
sprot: Swissprot
TPM: transcripts per million
ORF: open reading frame
OrcAE: Online Resource for Community Annotation of Eukaryotes
FDR: False Discovery Rate

Glossary

DNA (deoxyribonucleic acid) – A polymer of nucleotides that has the entire organism's biological information. Its nucleotides are composed of one subunit of a sugar (deoxyribose) and one nucleobase (Adenine, Timine, Guanine or Cytosine).

RNA (ribonucleic acid) – A polymer of nucleotides formed when DNA is expressed and is fundamental for the organism's biological roles. The nucleotides of RNA consist of a sugar (ribose) subunit and one nucleobase (Adenine, Uracil, Guanine or Cytosine).

mRNA (messenger RNA) – Subtype of RNA single-stranded molecules responsible for carrying information from the DNA to ribosomes necessary for protein synthesis.

Non-coding RNA – Functional subtype of RNA that is not translated into a protein, that includes tRNA (transfer RNA) and rRNA (ribosomal RNA).

Small RNA – Subtype of RNA with less than 200 nucleotides most often non-coding, usually responsible for RNA silencing. These include miRNA (microRNA) and siRNA (small interfering RNA). If longer than 200 nucleotides it is considered lncRNA (long non coding RNA).

RNA-seq (RNA sequencing) – Transcriptome sequencing through Next Generation Sequencing (NGS) to get information of RNA in a specific physiological condition and time.

Unix – Multitasking and multiuser computer operating system that manages hardware and software resources.

Software – Sequence of computer instructions that are executed in a way to achieve a certain goal.

Pipeline – Arranged chain of processes where the output of one process will be the input of the next one.

Transcriptome assembly – Reconstruction/Rearrangement of transcript sequences from RNA-seq reads to create a full transcriptome.

Table of Contents

List of Tables and Figures

1 Introduction

Biotechnology is defined as the exploitation of biological systems, living organisms, or their derivatives in technological applications to make or modify products or processes generally for societal benefit [5]. To engage with the public and to have an easy to understand system, biotechnology can be divided into colour codes. White biotechnology is related to industrial processes, red biotechnology is related to medical processes, green biotechnology is related to agricultural processes and blue biotechnology is related to marine and aquatic processes. In this project, the organism in question is a fish therefore the present project fits within blue biotechnology, which includes marine aquaculture and ocean farming. The present project is focused on the assembly and annotation of the European sardine (*Sardina pilchardus*) or sardine transcriptome and the project will form the basis of future studies.

The transcriptome is derived from an organism's DNA and contains the information necessary for all the proteins required to build and sustain an organism. Transcriptomics provides a wealth of information that can be used for a diversity of possible studies on a given organism. But in order to study the transcriptome the first thing that needs to be done is to sequence it. Recent development in sequencing, or Next Generation Sequencing (NGS) methods are leading to an unprecedented increase in available transcriptomes and underpins the remarkable big data results. To manage this big data, it is essential to have access to bioinformatic tools supported by high computing processes and much of this thesis will be based on the application of bioinformatics to assemble and analyse the transcriptomes generated from a number of sardine tissues.

The sardine is the target species because of it elevated economic and cultural importance for the Mediterranean and due to an unexplained and unexpected decline in the stock there are several species conservation concerns. Relatively few molecular resources are available for the sardine, which is a non-model organism and the factors underlying the population decline are unknown. The availability of molecular resources for the sardine is a priority as it will be a tool that will enable studies of biotic factors that may explain the population decline. The present project aims to assemble and annotate the transcriptome of the sardine and in this way, increase the molecular (both genomic and transcriptome) data for studies of this species.

1.1 Transcriptome

An organism's entire biological information is coded in its genome, which consists of DNA that gains structure in the chromosomes through its interaction with histones and other molecules [6]. Through the process of transcription, DNA is copied into RNA with the help of RNA polymerases, helicases and transcription factors. RNA polymerase binds to a sequence of DNA located before what is going to be expressed called the promoter and starts building an RNA strand, complementary to the DNA, which will serve as the template for the other RNA strand, while helicase is responsible for breaking the hydrogen bonds of DNA, separating the two strands. In turn, the RNA gets translated into proteins that are frequently important in modulating or determining the phenotype of an organism [7]. The transcriptome is most simply defined for any specific stage of every physiological condition as every RNA molecule in a cell or tissue, and the quantity. Transcriptomics aims to catalogue all the transcripts, including mRNAs, non-coding RNAs and small RNAs, determine the transcriptional structure of genes and to quantify the expression levels of each transcript. To extract total RNA, it is necessary to take into consideration that RNA molecules are very labile and degrade rapidly over time so the quality of the RNA being used for transcriptomics needs to be checked. In most cases to analyse expression data, the RNA coding information is passed into a more stable molecule (cDNA) via reverse transcription. In the RNA-seq wet lab process the RNA coding information is passed into cDNA during the RNA-seq sequencing library construction [6].

1.2 Sequencing

There are a number of different types of biological sequences, but the present study is most directed at the nucleic acid and proteomic sequences. DNA nucleotide sequences differ from RNA since the base thymine is substituted by uracil and the sugar in the nucleotide in DNA is deoxyribose and in RNA is ribose. Sequencing provides information about the sequence of bases or amino acids in a nucleic acid or protein, respectively and thus yields the primary structure of the sequence (Mardis 2008).

Sequencing of nucleic acids started with the Sanger method. Sanger sequencing consists in running four reactions in parallel, one for each of the bases, and generating one long sequence at a time. The sequencing reaction ends with a radioactively or fluorescent labelled dideoxy nucleotide. Analysing the size of the fragments originated from the sequencing reaction allows the order of the bases in a sequence to be established. The advance of technology means that Sanger sequencing has been replaced by Next Generation Sequencing (NGS) [9]. NGS

techniques are based on massive parallel sequencing where millions of small fragments of DNA are simultaneously amplified. Most techniques emit some kind of signal that represents and indicates a base has been added. Different techniques vary in relation to the signal produced or the sequencing reaction itself [8]. The signal is captured by a computer that saves the information as RNA-seq data.

While NGS is the most recent method it is still sometimes preferable to use the Sanger method due to its cost effectiveness. The Sanger method is best used to sequence up to 100 fragments with a higher accuracy and lower cost than NGS. The actual NGS RNA-seq output comes in the format of FastQ files, structured in four text lines per read, the first consists of the name of the read, the second the read's nucleotide base sequence and in the fourth their respective base qualities in Phred33 coding, a code in each character, there is a quality value assigned to it [10]. Handling these big files is a challenge and requires high throughput processing computers and bioinformatics knowhow in the Unix environment. The reason Unix is preferred is due to the powerful use of its tools that allow big file editing and the ability to connect to servers where the commands can be run.

1.3 Bioinformatics

The application of tools of computation and analyses to capture and interpret biological data is the core of Bioinformatics. This approach uses sophisticated software, pipelines and platforms connected through the internet or machines to build genomes and transcriptomes.

The advance of technologies in Genomics led to an era in the early 90's of "*Large Data Acquisition*" where biological data kept accumulating at a fast pace and there was a need to deal with the high amounts. As more data is accumulated and stored we have now entered the "*Big Data era*". With the continuous evolution of high throughput methods the challenge is now how to compute the data and efficiently store, transfer, secure and process the large amounts of data, while minimising the errors [11], [12].

Different types of data in bioinformatics can be defined, the main five are: gene expression (transcriptome study); DNA, RNA and protein sequence; protein-protein interaction; pathway; gene ontology [13]. For a transcriptome study identifying novel transcripts from annotated genes, splicing isoforms and gene-fusion transcripts is fundamental. Three major steps are required for transcriptome studies, these steps can be performed with different tools depending on factors like the sequencing technique used, the type of organism in study and the goal of the analyses but consist on the same general principles. Data pre-processing is required to remove

sequencing adaptors, insertions resulting from library preparation and near-identical reads to correct sequencing errors and improve the read quality. A transcriptome assembly strategy can be reference-based (using the genome sequence as the reference), *de novo* assembled or a combination of both [14]. Once assembled a range of tools are available to search for sequence similarities using various public databases, such as Swiss-Prot, Pfam, eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups), GO (Gene Ontology) and Kegg (Kyoto enciclopedia of genes and genomes), to analyse the contigs from the assembled transcriptome and annotate them.

The three main methodological steps in the present project were, 1) processing of the raw read quality, 2) the assembly of the sequence reads and 3) the annotation.

Quality control methodology

Quality control tools are required to assess and edit raw sequence read data, as adapter contaminations and low-quality bases can pose a real problem depending on the library preparation and downstream application. Two approaches will be tested in the thesis project to consistently apply quality and adapter trimming to FastQ files, first the Trim Galore software (www.bioinformatics.babraham.ac.uk/projects/trim_galore/), and secondly the Trimmomatic (github.com/timflutre/trimmomatic) plus FastQC. Trim Galore is a wrapper tool around Cutadapt and FastQC. Cutadapt (github.com/marcelm/cutadapt/) finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from high-throughput sequencing reads. FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/) is a quality control tool for high throughput sequence data producing an overall report of the edited reads. FastQC is also used in the second approach to confirm the edited output of Trimmomatic. Trimmomatic is a fast, multithreaded command line tool that can be used to trim and crop Illumina (FASTQ) data as well as to remove adapters and is similar to the program Cutadapt.

Assembly methodology

Assembly tools process large volumes of RNA-seq reads, the assembly procedures may be different depending on the software and whether or not there is prior genomic reference data. If there is a reference genome available it can be genome-guided, where the reads are first aligned to the genome, if there isn't a genome it must be a *de novo* assembly procedure and reads are assembled when they overlap each other, it is also possible to use a combination of the methods. Various software are freely available including Oases, Trinity, trans-ABySS and Cufflinks.

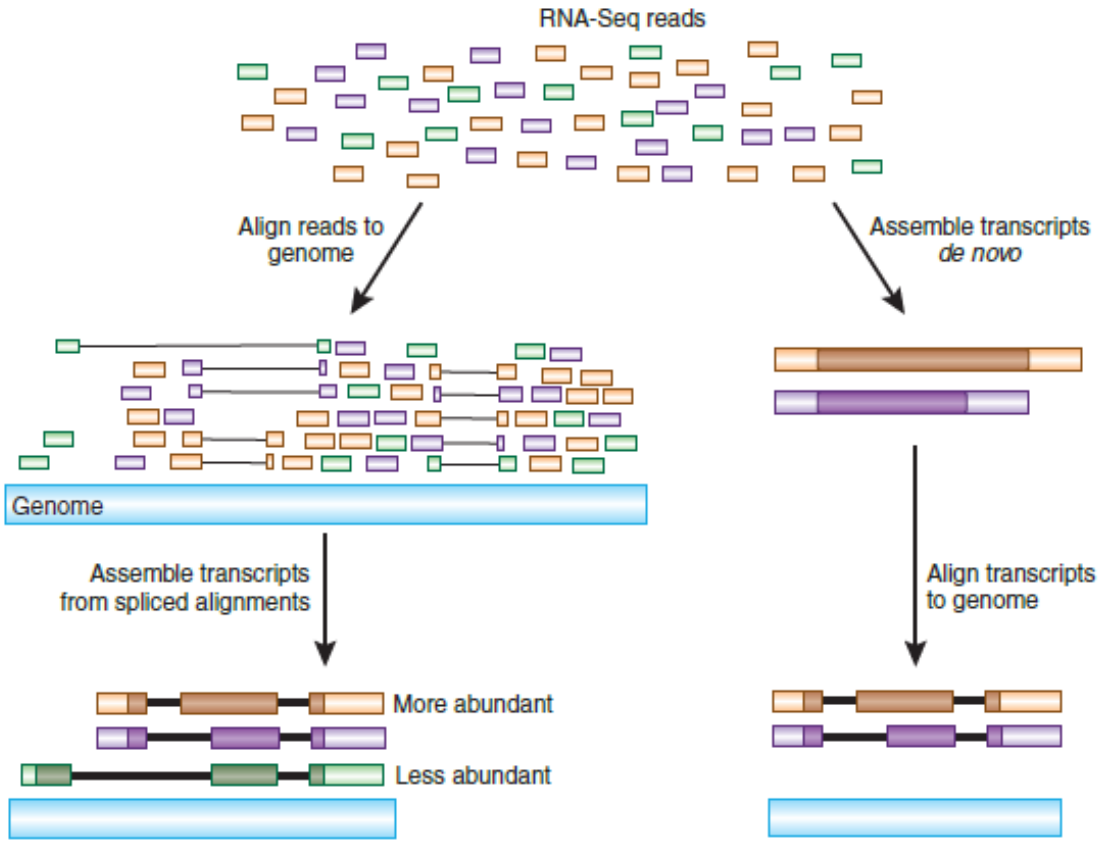

Figure 1.1: Strategies for reconstructing transcripts from RNA-Seq reads [1].

In the present thesis the software used for assembly was Trinity (github.com/trinityrnaseq/trinityrnaseq/wiki) (Grabherr, M. et all 2013), an efficient and robust *de novo* reconstruction of transcriptomes from RNA-seq data and also more recently a genome guided reconstruction. Trinity combines three independent software modules: Inchworm, Chrysalis, and Butterfly. Inchworm assembles the RNA-seq data into the unique potential sequence as contigs resulting from the kmers extensions combinations, Chrysalis clusters the Inchworm contigs into clusters when they overlap and constructs complete de Bruijn graphs for each cluster, and finally Butterfly then processes the individual graphs in parallel, tracing the paths that reads and pairs of reads taken within the graph, ultimately reporting full-length transcripts for alternatively spliced isoforms, and separating transcripts that corresponds to paralogous genes (Figure 1.2). Bowtie2 (bowtie-bio.sourceforge.net/bowtie2/index.shtml) is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences and is used in the present project to allow genome guided transcriptome assembly. Two approaches will be tested, Trinity *de novo* and Trinity genome guided assemblies. The metrics

of the assemblies generated will be compared to decide which output generated will be annotated. Before the annotation, a post-assembly quality control is implemented using Transrate (http://hibberdlab.com/transrate/) [16], a software for transcriptome assembly quality analysis.

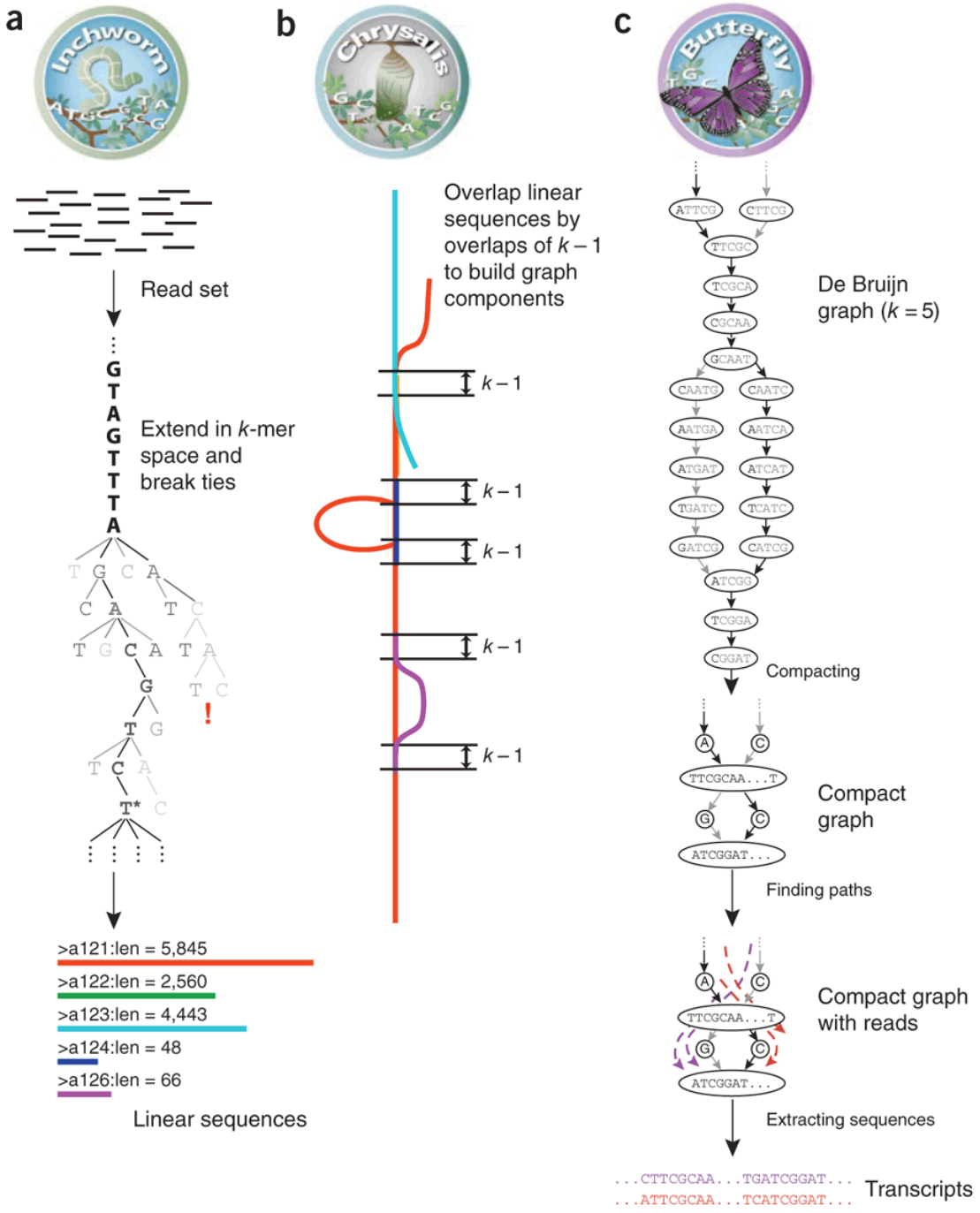

Figure 1.2: Overview of Trinity [2].

Annotation methodology

Annotation is the next step after assembly and consists of identifying assembled transcripts by comparison with other transcripts in public databases and assigning their function based on the most similar transcripts found, whether it belongs to the same species or not, considering their primary structure, corresponding proteins and domains. This makes it possible to separate putative proteins based on their involvement in cellular components, biological processes and molecular functions.

Annotation tools identify coding sequences by similarity and composition searches in databases. There are many automated pipelines created for this purpose such as Annocript, Trinotate, Blast2GO and TRAPID that go through different databases. In the present study Trinotate (trinotate.github.io/) - comprehensive annotation suite designed for automatic functional annotation of transcriptomes, particularly *de novo* assembled transcriptomes, from model or non-model organisms was used. Trinotate makes use of a number of different well referenced methods for functional annotation including homology searches to known sequence data (BLAST+/SwissProt), protein domain identification (HMMER/PFAM), protein signal peptide and transmembrane domain prediction (signalP/tmHMM), and also leverages various annotation databases (eggNOG/GO/Kegg databases). All functional annotation data derived from the analysis of transcripts is integrated into a SQLite database which allows fast efficient searching for terms with specific qualities related to a desired scientific hypothesis or as a means to create a whole annotation report for a transcriptome.

1.4 Sardine

The sardine (*Sardina pilhardus*) is a subtropical small pelagic fish distributed along the north-eastern Atlantic Ocean and in the Mediterranean Sea belonging to the Clupeidae family. It is the most important fish in terms of catch biomass with the biggest fishery occurring in Morocco [17]. Catches of the sardine have been increasing over the past years, in 1960, 487 900 tons were captured and in 2010 the total capture for the sardine has risen to 1 245 956 tons 2010 (FAO 2017). Fish stock management studies indicate that the capture of sardine is no longer sustainable and to avoid the collapse of the stocks the European community has lowered the allowable catch and in Portugal the volume of landings for this fishery has rapidly decreased. This has led to considerable problems within the sector and has changed the sardine from a “poor mans” food as market prices have soared from 1-2 euros/kg to 8-10 euros/kg. As a consequence of this decline in the available sardine biomass in Portuguese waters there has

been a rise in interest in establishing why the stock has collapsed. Concern has been raised in relation to the conservation of the sardine. Some fishermen take this concern seriously while others think there is no danger to the population (H. O. Braga et al. 2017). Recent studies found that the biomass of the sardine is currently declining along with its harvest, however the reason for the decline in the population isn't clearly known.

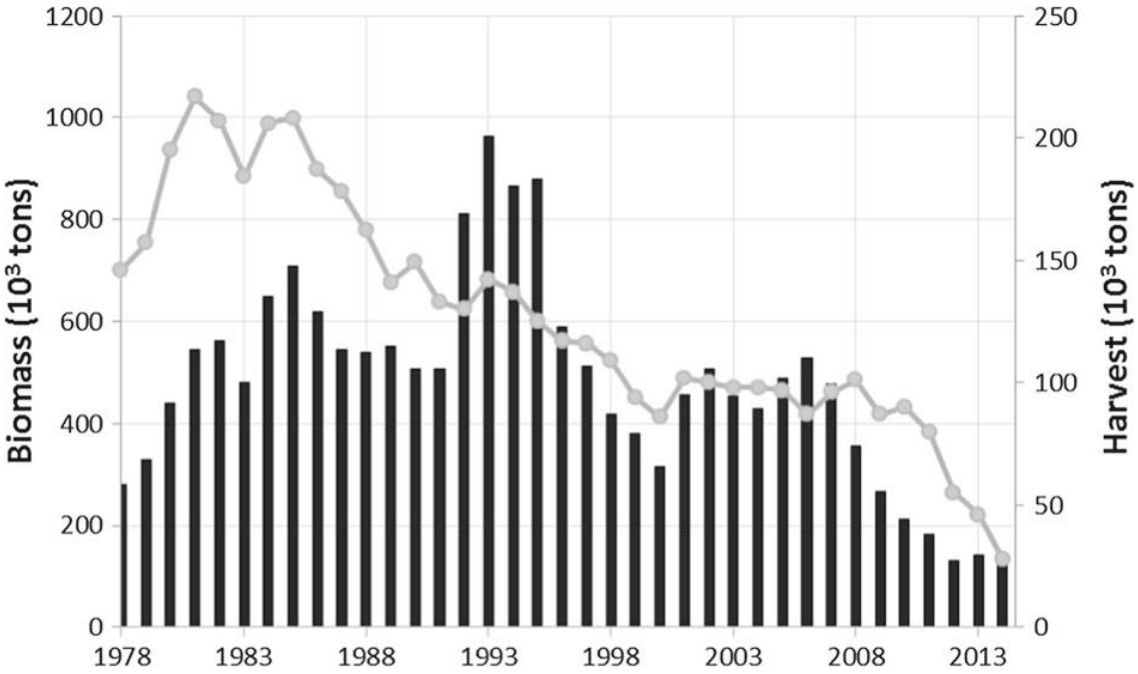

Figure 1.3: Sardine historical landings (line) and biomass (columns) [3].

Sardine biology and ecology

Sardines have grey subcylindrical bodies with a rounded belly and dark spots, the last 2 anal fin rays are enlarged and grow up to 25 cm. The sardine is a serial spawner and starts breeding at one year old throughout the year, mostly in the winter, near the shore and it has a maximum reported life span of 15 years, reaching maturity at 1 year old when it is around 8 cm and it swims in large schools. The sardine mostly eats planktonic crustaceous and occupy a basal position on the food chain, transferring energy from plankton and small organisms to larger fishes, sea birds and marine mammals with great influence over the health of the animals above the sardine in the food chain (Jawad 2015, FAO 2017). Females tend to be slightly larger, heavier and less abundant than males and sardines from offshore tend to be bigger than those that are inshore [21], [22]. They are also of interest due to some distinct biological characteristics they possess such as rapid growth and resistance to algal blooms.

Rationale for transcriptome characterization

The decrease in the sardine population is a recent debated concern and the difficulties in its conservation can possibly have a negative consequence in the ecosystem. To know what changes this may cause and maybe spread awareness there needs to be more studies. Since there isn't much information available on this specific fish because it isn't a model organism, one starting point is to characterize its transcriptome and then annotate the transcripts. The long-term final objective is to determine the population structure and dynamics and separate them based on biological borders and not geopolitical.

On the NCBI website there are 57.196 entries on the nucleotide database of Expressed Sequence Tags and Genome Survey Sequences for all the clupeiformes and most of the reads (39,344) are for *Clupea harengus* (Atlantic herring) and only 566 are for *Sardina pilchardus* and the latter sequences are mainly cytochrome related (NCBI December 2017). During the current project, the sardine's genome was assembled and reported in a manuscript entitled "A haplotype-resolved draft genome of the European sardine (*Sardina pilchardus*)" (http://bioinformatics.psb.ugent.be/orcae/overview/Spil, Genebank Bioproject PRJEB26757) [23]. The information generated in this project may contribute to studies trying to establish the unique characteristics of the sardine by comparing its transcriptome with other closely related vertebrate fishes and help better understand evolution since the sardine belongs to the clupeiformes order which is situated in an older position in the phylogenetic tree relative to other orders with more studied fishes that went through more recent speciation events (Figure 1.4).

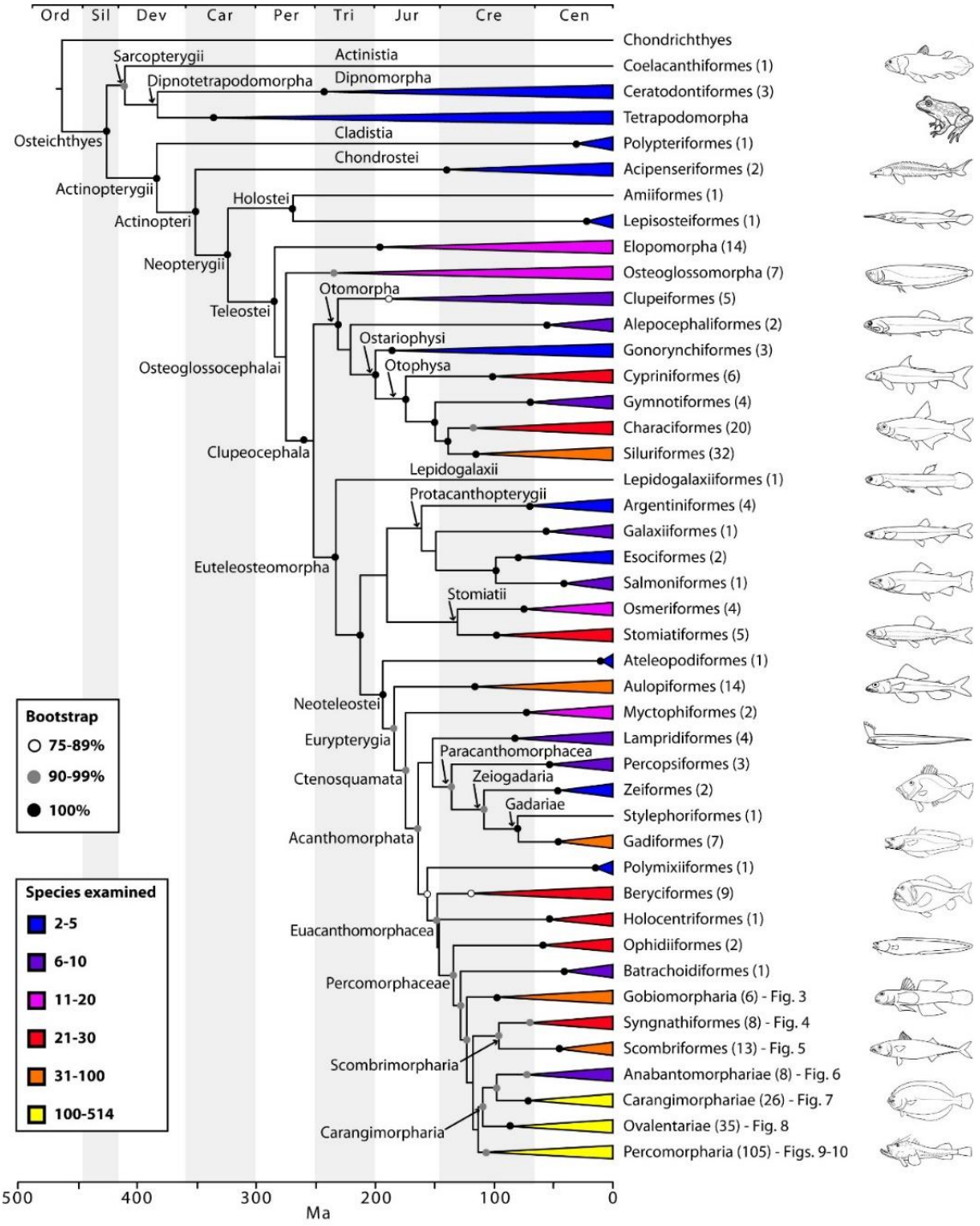

Figure 1.4: Main phylogenetic hypothesis of bony fish groups collapsed to depict higher-level clades [4].

1.5 Objective

This project intends to develop a representative transcriptome for the adult sardine with the help of bioinformatic tools to serve as a resource for future biological and genetics studies of the sardine, which is a non-model organism, that benefit from, or require looking at its transcriptome. This is achievable by using sequencing data for several tissues and executing a series of algorithms that go through important stages to get the most accurate transcriptome as possible and will be run in servers with high performance computing. Different bioinformatics methodologies will also be approached for the transcriptome analysis. The different stages will serve to:

- Trim the reads and exclude the small and poor-quality reads using 2 different tools.
- Assemble the resulting reads into a mapped transcriptome using 2 different ways;
- Annotate the possible genes by comparing them with transcripts of other species in data bases.

2 Material and Methods

The methodology of the bioinformatic workflow applied in this study is represented in figure 2.1.

2.1 Sampling

Eleven tissues were collected from a single female sardine, blood was also sampled for the purpose of genome sequencing. The female was fished off the shore from Olhão, on May 2016 and maintained in Pilot Station of Aquaculture in Olhão (EPPO) until it was sacrificed by an overdose of anaesthesia (1:250, 2-phenoxyethanol) followed by euthanasia by cervical section in September 2017. The tissues were then preserved in RNA*later*® and stored at -20 ºC.

In the context of the genome sequencing project in which the transcriptome is integrated, a bioproject (PRJEB27990) has been created in the European Nucleotide Archive (ENA) where the samples have the accession numbers listed in table 2.1.

Table 2.1: List of the tissues used as a source of RNA respective abbreviations and sample accession number in ENA archive.

Tissue	Abbreviation	Sample Accession
Gill + Branchial Arch	Gi	SAMEA4809353
Liver	Lv	SAMEA4809357
Spleen	Sp	SAMEA4809360
Gonad (female)	Gn	SAMEA4809354
Midgut	Mg	SAMEA4809358
White Muscle	WM	SAMEA4809350
Red Muscle	RM	SAMEA4809359
Kidney	Kd	SAMEA4809356
Head Kidney	HKd	SAMEA4809355
Brain + Pituitary	Br	SAMEA4809351
Caudal Fin (Skin + Cartilage + Bone)	CF	SAMEA4809352

2.2 Sequencing

The total RNA from the eleven tissues was extracted using a Maxwell® 16 Total RNA Purification Kit after they were mechanically disrupted. Total RNA was then double-treated using a DNA-free kit with DNase, quantified by NanoDrop 1000 Spectrophotometer and stored at -80 ºC as described in the genome manuscript "A haplotype-resolved draft genome of the European sardine (*Sardina pilchardus*)"[23].

The mRNA was isolated from the total RNA using a NEBNext® Poly(A) mRNA Magnetic Isolation Module kit and sequenced using Illumina – HiSeq4000 PE 150 bp Cycle with the NEBNext® Ultra™ Directional RNA Library Prep kit. The quality control of the mRNA was made with Qubit, Tapestation and the quality was above 8 RIN. This part was done by the Admera Health company and generated stranded paired-end reads. The adapters used for the sequencing and considered in the quality control procedures for the removal of these sequences from the resulting reads by the software were the following:

Table 2.2: Illumina sequencing adapters used in the sequencing. R1 and R2 are forward and reverse reads, respectively of the paired-end reads. Sequence orientation 5' to 3'.

Read	Adapter Sequence	Reverse complement
R1	AGATCGGAAGAGCACACGTCTGAACTCCAGTCA	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
R2	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT

2.3 Computational usage

Most algorithms ran on the "High Performance Computing" (HPC) of INCD (Infraestrutura Nacional de Computação Distribuída) servers located in Lisbon through the Unix environment. Processes that demanded high memory to run were queued with qsub that orderly runs batch job submissions in different parted machines (Examples of the codes and R scripts used in appendices 6.1).

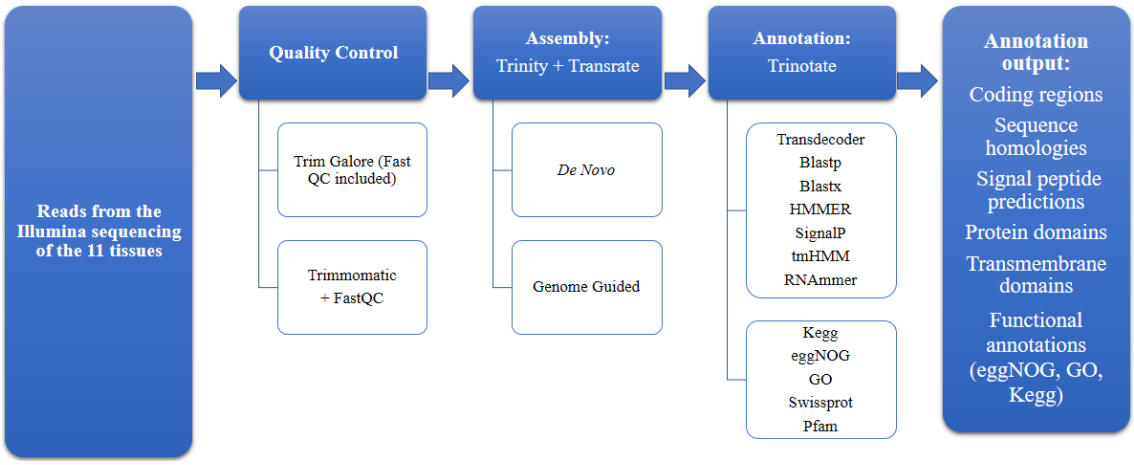

Figure 2.1: Flowchart of the methodology. The original reads go through a quality control check, followed by the assembly with Trinity and another quality control step with Transrate and finnaly annotated with Trinotate that uses various search tools for protein and transcript analyses in different databases.

2.4 Quality Control and Reads Editing

The quality control and reads trimming were performed with two alternative software pipelines: Trim Galore version 0.4.5 and Trimmomatic version 0.36 plus FastQC version 0.10.1, for methodological comparative purposes.

2.4.1 Trim Galore

With Trim Galore, a wrapper of Cutadapt and FastQC, the code used had parameters to trim quality ends from reads if the score was below 20 (default), indicate to use ASCII+33 quality scores as Phred scores, remove 1 bp from the 5' end of read 1 and 2, a stringency of 3 bp overlapping with adapter sequence required to trim a sequence, allow a maximum error rate of 0.1 (default). After Cutadapt trimmed the reads, it discarded reads that became shorter than 30 bp and unpaired reads and the output was edited FastQ files compressed with gzip. Edited reads were then analysed by FastQC to generate the quality report with descriptive statistics represented in a graphical format for better decision making of subsequent procedures.

2.4.2 Trimmomatic plus FastQC

The Trimmomatic command used had parameters to specify the input type as paired-end reads, use 4 threads, trim reads if the average quality of a window across 4 bases was below 20, keep reads of a minimum of 30 bp, find and remove Illumina adapters specified in the "adapter.fa" file with a maximum of 2 mismatches and remove 1 base from the start of the read. Edited reads were also analysed by FastQC downstream to visualize the quality of the output the same way as with Trim Galore.

2.5 Assembly

The transcriptome assembly was achieved with Trinity version 2.5. A post assembly quality control step was made using Transrate version 1.0.3 by comparing the assemblies with the raw reads. Transrate ran only for the assemblies generated with the *De Novo* approach. Assemblies were generated for each tissue and an additional assembly containing the reads from all the tissues.

2.5.1 *De Novo* assembly

The Trinity command used for *de novo* assembly had parameters to indicate the reads were in FASTQ format, had a maximum memory usage of 184G of RAM, 8 CPU and the orientation of R1 and R2 were the reverse forward read, respectively. Additional parameters specified were minimum contig length of 200 bp (default) and run normalization separately for each pair of FASTQ files, then one final normalization that combined the individual normalized reads with a maximum read coverage of 50 (default). To obtain the metric statistics of the assemblies a script (TrinityStats.pl) provided by Trinity was run.

The code for Transrate had parameters to use 8 threads (default) and to indicate the left and right reads in FASTQ format in addition to the assembly input in FASTA format.

2.5.2 Genome Guided

Before the genome guided assembly could be run, an alignment of the RNA reads against the genome sequence was required, this alignment was made with Bowtie2 (version 2.3.4) which forms output in the form of a Sequence Alignment Map (SAM) file. This alignment procedure had two alternatives approaches, one that makes local alignments and the other that performs end-to-end alignments, so both were tested. The parameters of Bowtie2 specified in the command were that the reads were in FASTQ format, the orientation of R1 was Reverse and the orientation of R2 was Forward, and to use 14 CPU. The parameters specified for the input data were all RNA FastQ files from the present study and the preliminary draft genome assembly of the sardine. After the alignment, a pipeline of Samtools (version 1.1) utilities converted the SAM files to Binary Alignment Map (BAM) files with the command view and then sorted them with the command sort.

The trinity genome guided command uses solely the sorted BAM to retrieve sequence information for the assembly, the specified parameters to indicate a maximum intron length of 10,000 bp (for the end-to-end alignment based) and 25,000 bp (for the local alignment based), a maximum memory of 114G of RAM and to use 14 CPU were used.

To get the metric statistics of the assemblies generated a script “TrinityStats.pl” provided by Trinity was ran on both assemblies FASTA files.

2.6 Functional Annotation

The functional annotation of the transcriptome assemblies obtained was done using the Trinotate version 3.1.1 annotation pipeline, and the REVIGO (revigo.irb.hr/) [24], a web server that summarizes long, unintelligible lists of GO terms by finding a representative subset of the terms using a simple clustering algorithm that relies on semantic similarity measures. Software's required for the Trinotate pipeline included TransDecoder (5.0.2), SQLite (3.6.20), NCBI BLAST+ (2.7.1), HMMER (3.1 and 2.3 for the RNAMER), tmHMM (2.0), SignalP (1.05, with Perl (5.8.8)) and RNAMMER (1.2).

2.6.1 Trinotate

The obtained deduced proteins FASTA file "transdecoder.pep" was used as input for several annotation programs that belong to the Trinotate pipeline; the "SignalP" to predict signal peptides, the "tmHMM" to predict transmembrane regions, the "HMMER" (hmmscan) to identify protein domains of the PFAM v30.0 database ("Pfam-A.hmm", 15th Feb 2018)
Blastp to identify protein homologies of Swissprot/Uniprot database (accessed 14th Feb 2018).
The transcriptome nucleotide FASTA file "good.trinity.fasta" was used as input for ribosomal RNA using RNAMER program and query the same Swissprot/Uniprot database but using blastx.
All output of the several annotation analysis was then integrated into single SQLlite database already populated with GO, eggNOG, and KEGG pathways via the Trinotate utility.
This procedure as done to all transcriptomes curated by Transrate, for comparison purposes two Transrate non-curated transcriptomes (Caudal Fin and all tissues) were also annotated following the same procedure.

2.6.2 Transcript Quantification

To count the overall transcripts being expressed, several Trinity scripts were used with the assembly of the reads from all the tissues after going through Transrate. Firstly, the "align and estimate abundances.pt" script aligned and estimated abundances and had parameters to indicate the input as a FASTA file, paired-end, the use of the RSEM method and the use of the bowtie2 alignment method. Secondly, the "abundance_estimates_to_matrix.pl" script used the estimates to create matrices of counts and of normalized expression values and had parameters to the use of the RSEM method. Thirdly, the "count_matrix_features_given_MIN_TPM_threshold.pl" script counted the expression numbers on the matrices. Finally, using R language, a linear regression was calculated and

plotted, within the range of 10-100 TPM to obtain the intersect representing the total count of genes being expressed. an Rplot was edited to provide the graphs for total gene and transcript expression.

2.6.3 Tissue-specific genes

For the prediction of the specific genes from each tissue with a threshold of 95% of total expression, normalised counts matrix of each tissue vs the all the others were created with a Trinity Differential Expression script "run_DE_analysis.pl" specifying the parameters of 0.4 dispersion and the use of RSEM method. A heatmap for the predicted tissue-specific genes was generated using the Trinity Differential Expression scripts "analyse_diff_expr.pl" using the Transrate non-curated all tissues assembly.

2.6.4 REVIGO

GO enrichment analysis were performed with GOseq scripts, the first from Trinotate toolkit and the others from Trinity. "extract_GO_assignment_from_Trinotate_xls.pl" created a file with the GO annotations, "fasta_seq_lenght-pl" created a file with the transcript lengths with the use of the transcript lengths, "TPM_weighted_gene_lenght.py" created a file with the gene lengths, and finally, "run_GOseq.pl" created files with enriched and depleted categories. The resulting list of GO terms with their associated over representative p-values was used for REVIGO analysis with an allowed similarity of 0.9, 0.7 and 0.5 for biological process and molecular function. The scatterplot based on REVIGO semantic similarity results were plotted using R language. This procedure was done for all assemblies.

3 Results and Discussion

The results of each section of the project determined which method was adopted before moving onto the next pipeline step as detailed in figure 3.1.

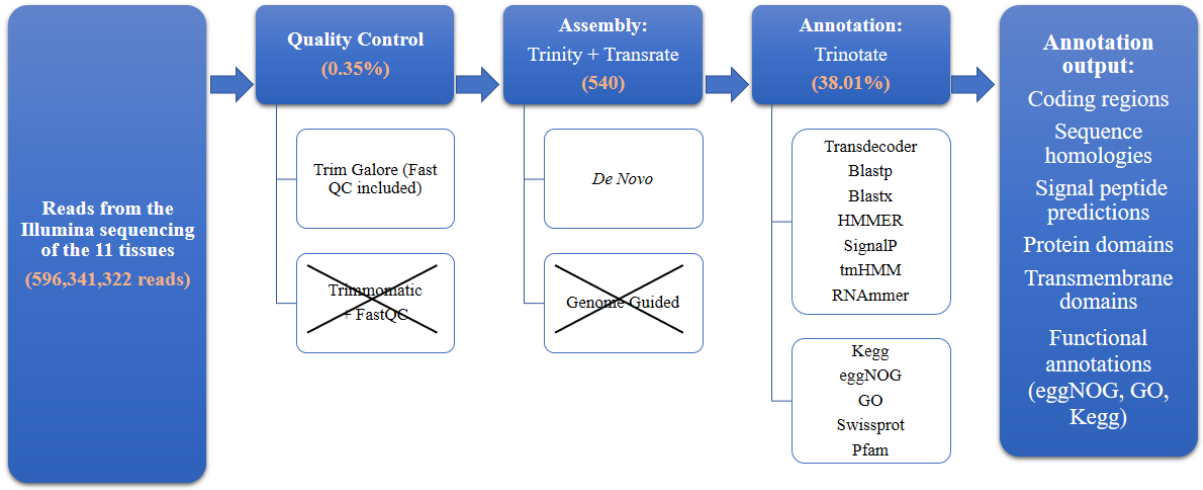

Figure 3.1: Methodology decision making flowchart and respective results. The original 593,341,322 reads went through a quality control check using Trim Galore and Trimmomatic which were compared with FastQC. The results from Trim Galore, with an average of 0.35% reads removed, proceded to the assembly with Trinity by *de novo* and genome guided. The results from the *de novo* assembly plus Transrate proceed to the annotation with Trinotate. N50 and SwissProt annotation results of the caudal fin de novo assembly were 540 and 38.01%, respectively. The results for all assemblies and respective annotations are described in tables 3.5, 3.6, 3.7 and 3.8.

3.1 Quality Control and Reads Editing

The sequencing of the 11 tissues generated in total about 600 million (593,341,322) paired-end reads that were submitted to the ENA archive under the accession run numbers from ERR2720641 to ERR2720651, that had to be trimmed for better quality.

Two approaches were compared for the quality control and reads editing. The parameters used for both the Trim Galore and the Trimmomatic were as similar as possible to better compare them without different variables. The indication of Illumina adapters wasn't needed for Trim Galore as it recognized them automatically. Reads and the respective pair were discarded if they were less than 30 bp in length. Quality editing of the reads is essential for the assembly procedures, as adapter contaminations and low-quality reads lead to poor assembly's outcomes with many artefacts that can prevent the assembler software from being able to deal with data. Results from the Trim Galore gave reads above 28 quality score (Figures in appendices 6.2) from 30 to 149 bp, GC percentages from 49 to 55 % in line with coding sequences/GC richer relationship and removed from 0.28 to 0.44 % of the original reads sequences.

Table 3.1: Trim Galore edited results of the raw paired-reads from the 11 sequenced sardine tissue libraries.

Tissue	Raw paired-reads	Trim galore results				
		Surviving Reads	Removed Reads	Removed Reads Percentage (%)	Sequence Length	%GC
Gill + Branchial Arch	**29,783,994**	29,700,355	83,639	0.28	30-149	50
Liver	**33,479,471**	33,372,568	106,903	0.32	30-149	55
Spleen	**25,634,530**	25,537,817	96,713	0.38	30-149	55
Gonad	**22,241,327**	22,149,330	91,997	0.41	30-149	52
Midgut	**28,016,117**	27,891,625	124,492	0.44	30-149	55
White Muscle	**24,409,160**	24,326,763	82,397	0.34	30-149	53
Red Muscle	**30,653,774**	30,544,700	109,074	0.36	30-149	53
Kidney	**27,861,879**	27,763,700	98,179	0.35	30-149	52
Head Kidney	**25,280,960**	25,208,409	72,551	0.29	30-149	52
Brain + Pituitary	**24,467,352**	24,376,337	91,015	0.37	30-149	49
Caudal Fin	**26,342,097**	26,255,854	86,243	0.33	30-149	52
All Tissues	**298,170,661**	297,127,458	1,043,203	0.35	30-149	53
Mean	**27,106,423.73**	27,011,587	94,837	0.35	30-149	53

Results from the Trimmomatic granted reads above 32 quality score from 1 to 149 bp and GC percentages from 49 to 55 % and removed from 5.77 to 8.08 % of the original sequences.

Table 3.2: Trimmomatic edited results of the raw paired-reads from the 11 sequenced sardine tissue libraries.

Tissue	Raw paired-reads	Trimmomatic results				
		Surviving Reads	Removed Reads	Removed Reads Percentage (%)	Sequence Length	%GC
Gill + Branchial Arch	**29,783,994**	28,065,613	1,718,381	5.77	1-149	50
Liver	**33,479,471**	31,305,340	2,174,131	6.49	1-149	55
Spleen	**25,634,530**	23,690,270	1,944,260	7.58	1-149	55
Gonad	**22,241,327**	20,482,936	1,758,391	7.91	2-149	52
Midgut	**28,016,117**	25,757,615	2,258,502	8.06	2-149	55
White Muscle	**24,409,160**	22,446,925	1,962,235	8.04	3-149	53
Red Muscle	**30,653,774**	28,176,295	2,477,479	8.08	2-149	53
Kidney	**27,861,879**	25,767,818	2,094,061	7.52	4-149	52
Head Kidney	**25,280,960**	23,522,879	1,758,081	6.95	2-149	52
Brain + Pituitary	**24,467,352**	22,734,872	1,732,480	7.08	1-149	49
Caudal Fin	**26,342,097**	24,648,080	1,694,017	6.43	1-149	52
All Tissues	**298,170,661**	276,598,643	21,572,018	7.23	1-149	53
Mean	**27,106,423**	25,145,331	1,961,092	7.23	2-149	53

Trimmomatic results revealed that it removed more sequences than CutAdapt from Trim Galore, a difference of overall 6.88 % as it considered more reads to be of low quality, and yielded reads smaller than 30 bp in discordance with the size threshold defined in the Trimmomatic command. The GC percentage of the surviving reads were the same for both assembly approaches, around 53 %, which seems to be appropriated to gene expression products (Tables 3.1 and 3.2).

Results from Trimmomatic granted a per base sequence quality slightly higher than Trim Galore (Figures in appendices 6.2). Every other aspect of the assemblies remained very similar or equal. Ordering the tissues by their removed reads percentages yielded similar lists and the gill plus branchial arch tissue was always the tissue with the lower removed reads percentages in both approaches.

To proceed to the assembly, the output of the Trim Galore was chosen because of the unpredicted behaviour of Trimmomatic with regards to retaining reads smaller than 30 bp, and the higher than expected ratio of low quality reads removal, even though Trimmomatic showed slightly better quality scores.

3.2 Assembly

Two approaches were compared for the assembly. The Trinity *de novo* and Trinity genome guided assemblies that could be done with either local or end-to-end alignments. Besides assembling the reads from each tissue, a reference assembly was made using the reads from the 11 tissues for a global analysis. Of the many statistics obtained, the ones considered more relevant were the GC-content, mean length of sequence, number of contigs assembled and N50. A higher N50 and mean length values meant there was a higher chance of the assembly being less fragmentated and more likely to cover the genes full coding region in an ideal scenario. However N50 might not be completely accurate as artefacts such as contigs concatemerizations might and wrongly generated longer contigs lead to biased inflated N50´s values, due to very similar genomic regions of duplicated genes. Since the teleosts have gone through 3 whole genome duplication events there is high probability of existing very similar sequence regions, for example from paralogous genes and DNA copy number variations, that could be assembled in the same contig [25], [26]. These artefact contigs lead to a higher N50 that does not necessary mean that it belongs to a better assembly.

The assembly from the reads of all the tissues after Transrate was used in modelling the gene predictions for the genome annotation project, improving the detection of genomic features such as untranslated regions and organism specific genes[27]. The RNA-seq assembly,

repetitive elements and protein homology and ad initio gene prediction were used as inputs for the software's MAKER and SNAP (Semi-HMM-based Nucleic Acid Parser) to generate the gene models. For validation and assessment of genome completeness, the RNA-seq reads should have a high alignment rate with the genome as this indicates a good quality assembly, in these study all raw reads aligned with 90 % of the sardine draft genome (data not shown) indicating a good genomic assembly.

After Transrate assembly curation, 44 to 80 % contigs were retained from the initial Trinity *de novo* assembly, with a mean length varying from 425.98 to 686.88 bp and N50 from 474 to 1,039 (Table 3.3). According to the definition of the transcriptome, gene expression captured in sampling is dependent of tissue specificity, organismal developmental stage and physiological state in response to environmental stimulus.

Table 3.3: Statistic results from the Trinity *De Novo* and Transrate.

Tissues	Transrate	N Seqs	Seqs retained %	N Bases	Mean Length	Mean ORF %	N50
Gill + Branchial Arch	Before	113,560	55.06	69,178,936	609.18	61.06	945
	After	62,526		31,897,660	510.15	63.55	660
Liver	Before	99,177	53.54	58,513,424	589.99	63.88	899
	After	53,104		26,641,277	501.68	65.99	643
Spleen	Before	107,688	61.68	67,022,297	622.37	64.60	1,005
	After	66,419		35,016,057	527.20	66.81	694
Gonad	Before	84,447	50.35	78,951,254	934.92	61.67	1,757
	After	42,521		29,206,715	686.88	63.26	1,039
Midgut	Before	98,324	77.07	56,809,315	577.78	62.63	872
	After	75,782		42,003,712	554.27	63.11	770
White Muscle	Before	65,873	74.79	44,656,815	677.92	62.80	1,101
	After	49,266		29,177,911	592.25	64.57	810
Red Muscle	Before	69,238	80.70	39,118,150	564.98	63.02	818
	After	55,873		30,465,796	545.27	63.59	735
Kidney	Before	114,576	51.93	84,190,774	734.80	61.12	1,301
	After	59,495		33,533,130	563.63	63.28	779
Head Kidney	Before	109,603	60.12	77,073,001	703.20	60.92	1,206
	After	65,888		37,805,855	573.79	62.54	805
Brain + Pituitary	Before	132,824	56.93	71,241,852	536.36	61.42	769
	After	75,620		34,786,848	460.02	64.23	557
Caudal Fin	Before	98,276	65.97	44,894,190	456.82	65.61	540
	After	64,832		27,616,895	425.98	66.49	474
All	Before	385,373	44.24	258,928,743	671.89	61.11	1,370
	After	170,478		73,525,673	431.29	63.24	486

Mean	Before	123,246.58	61.03	79,214,895.92	640.02	62.49	1,048.58
	After	70,150.33		35,973,127.42	531.03	64.22	704.33

Transrate overall lowered every metric of statistic except for the mean ORF percentage, which was unexpected. Even though the contig N50 values obtained were lower after the Transrate filtration, the deflation of contig counts to values more similar to the expected value of genes being expressed in a given tissue, time and condition, was the main reason that the decision was taken to proceed to the annotation step with Transrate curated assemblies. All the assemblies prior to Transrate were kept for comparison of the annotation results.

Assemblies that got a higher number of contigs, like the midgut tissue assembly, have the chances to contain better assembled transcripts, but also more non-real contigs, while a lower number of contigs, like the gonad tissue assembly, might have less noise, but worse assembled transcripts.

The assembly containing reads from all the tissues could have more non-real contigs than the other assemblies, but a higher number of contigs would be expected as it represents more than one tissue, furthermore the sum of contigs from each tissue assembly yields a value bigger than the number of contigs the assembly from all the tissues got meaning the same contig may be represented across different tissue assemblies.

For methodological comparison purposes the caudal fin tissue reads were assembled with Trinity genome guided. Results from the genome guided assembly on that first tested tissue granted noticeable differences from the Trinity *De Novo* and in the different alignment methods, so no more other tissue went through all the approaches, neither did Transrate filtered the genome guided results. In the overall granted descriptive statistics, it indicated a lower quality assembly, with the exception of a higher median contig length than the obtained with the *de novo* approach, with the one from the end-to-end alignment being the highest. Every other metric of statistic was lower, with the ones from end-to-end alignment being the lowest (Table 3.4).

Table 3.4: Statistic Trinity results for the caudal fin tissue from the *De Novo* and genome guided based on local (GG local) and end-to-end (GG end) alignment approaches.

Caudal Fin	Total trinity genes	Total trinity transcripts	Stats based on ALL transcript contigs			
			N bases	Mean length	Contig N50	Median contig length
De Novo	78,584	98,276	44,894,190	456.82	540	287
GG local	72,485	79,561	33,866,744	425.67	464	292
GG end	62,761	67,338	28,106,139	417.39	444	302

Based on the comparison of the assembly results, one of a specific tissue (caudal fin) indicating the genome guided approaches was yielding lower quality assemblies based mainly on the assessment of N50 values, the decision to move forward with the *de novo* assembly strategy for all the remaining tissues transcriptomes assemblies was taken. The genome sequence used as a reference for the guided assemblies was still an initial preliminary draft, highly fragmented, this probably hindered on the effectiveness to obtain a better quality assembly in comparison with the *de novo* strategy. (Table 3.4). Several fine tunings of the local alignment based genome guided assembly parameters were tested, such as a bigger maximum intron length leading to some improvement of the assembly but not enough to surpass the *de novo* approach.

3.3 Functional Annotation

The transcriptome assembled is a valuable genomic resource for future biological studies such as physiological experimental studies via differential expression analyses, or evolutionary studies among others. For that it is required that a transcriptome is annotated in order that biologically meaningful information can be retrieved from the usage of the transcriptome in such studies.

3.3.1 Trinotate

Results from the Trinotate granted plenty of information and reports in the form of tables which were analysed for better representation of the biological value.

Table 3.5: Percentages of annotated transcripts per tissue.

Tissues	Sprot blastx	Deduced CDS	Sprot blastp	Pfam	SignalP	TmHMM	eggNOG	Kegg	GO blast	GO Pfam	Total contigs
Gill + Branchial Arch	38.60	29.31	23.99	18.69	1.56	4.18	33.32	33.84	38.08	11.26	62,526
Liver	40.07	29.66	25.06	19.39	1.54	4.11	34.85	35.29	39.52	11.67	53,104
Spleen	40.41	31.61	25.83	20.90	1.80	4.66	33.84	34.96	39.61	12.66	66,419
Gonad	42.50	38.07	31.71	25.05	1.94	5.46	37.14	37.85	41.88	15.00	42,521
Midgut	39.46	30.95	25.53	20.36	1.85	4.63	33.64	34.23	38.79	12.39	75,782
White Muscle	44.77	35.44	30.23	23.29	1.64	4.51	39.20	39.63	44.11	14.23	49,266
Red Muscle	42.08	30.31	25.75	20.39	1.27	3.74	36.14	36.74	41.34	12.61	55,873
Kidney	37.28	30.81	25.08	19.43	1.89	4.44	32.13	32.50	36.59	11.45	59,496
Head Kidney	38.40	32.20	26.41	20.61	1.92	4.84	33.08	33.58	37.78	12.30	65,888
Brain + Pituitary	37.09	24.47	20.21	15.13	1.02	3.56	31.75	32.43	36.27	9.06	75,620
Caudal Fin	38.01	23.89	19.73	14.99	1.03	3.01	32.72	33.41	37.49	9.06	64,832
All	25.49	15.88	11.56	9.70	1.22	2.73	19.70	20.92	24.81	5.90	170,478

Table 3.6: Percentages of annotated transcripts before and after Transrate filtration on the caudal fin tissue.

Caudal Fin	Sprot blastx	Deduced CDS	Sprot blastp	Pfam	SignalP	TmHMM	eggNOG	Kegg	GO blast	GO Pfam	Total contigs
Before Transrate	37.29	24.17	19.34	15.16	1.37	3.30	31.48	32.31	36.67	9.24	64,832
After Transrate	38.01	23.89	19.73	14.99	1.03	3.01	32.72	33.41	37.49	9.06	98,276

Table 3.7: Percentages of annotated transcripts contigs before and after Transrate filtration on the assembly containing reads from all the tissues.

All Tissues	Sprot blastx	Deduced CDS	Sprot blastp	Pfam	SignalP	TmHMM	eggNOG	Kegg	GO blast	GO Pfam	Total contigs
Before Transrate	33.89	26.95	21.11	18.46	2.42	5.14	27.37	28.32	32.98	11.49	385,373
After Transrate	25.49	15.88	11.56	9.70	1.22	2.73	19.70	20.92	24.81	5.90	170,478

Table 3.8: Percentages of annotated gene contigs before and after Transrate filtration on the assembly containing reads from all the tissues.

All Tissues	Sprot blastx	Deduced CDS	Sprot blastp	Pfam	SignalP	TmHMM	eggNOG	Kegg	GO blast	GO Pfam	Total contigs
Before Transrate	27.86	16.00	12.17	10.28	1.16	2.87	21.94	23.07	27.00	6.38	247,300
After Transrate	24.17	12.95	9.50	7.78	0.88	2.24	18.78	20.09	23.47	4.78	143,335

Between 15.88 to 38.07 % of the contigs were deduced to be coding sequences with TransDecoder; 25.49 to 44.77 % and 11.56 to 31.71 % of the contigs were annotated based on sequence homologies via Sprot blastx and Sprot blastp, respectively. Based on the sequence SwissProt ID's obtained and the Trinotate relational SQL database, 20.92 to 39.63 % of the contigs annotated with sequence homologies via BLAST+ were annotated using Kegg, 19.70 to 39.20 % using eggNOG, 24.81 to 44.11 % using GO blast. 9.70 to 25.05 % protein domains were identified with HMMER/PFAM and consequently, 5.90 to 15.00 % were annotated with GO based on the Pfam domains.

Overall, the database that annotated the most number of transcripts was eggNOG while the one that annotated the least number of transcripts was SignalP, that gave only a small percentage (1.02 to 1.94 % signal peptides) of the transcripts annotated and therefore presumably representing proteins secreted from cells, followed by transmembrane proteins identified with tmHMM, with 2.73 to 5.46 % transmembrane domains found. The assembly containing reads from all the tissues had an even lower percentage of annotation with the preceding annotation approaches. The tissue with the most total number of annotated genes was the midgut, while the tissues with the highest percentage of annotated transcripts was the gonad, closely followed by white muscle, both of these tissues had the lowest number of total genes. On the other hand, the tissue with lowest annotation percentage was brain plus pituitary, and this was the tissue with the second highest total number of gene transcripts, just behind midgut.

After the filtration of Transrate, in the case of a specific tissue, the caudal fin, some percentages of transcripts annotated by the different software's became higher while others became lower. In the case of the assembly containing reads from all the tissues all the percentages of annotation irrespective of the software were lower. Transrate lowering percentages of annotated transcripts means that before the filtering there were more transcripts that would get annotated. To further test Transrate, the assembly containing reads from all the tissues was also annotated before going through Transrate and had higher percentages of annotated transcripts than after

Transrate. As what would be expected was the opposite, all the isoforms removed to see if isoforms from annotated transcripts were the reason for the such lower percentages. Comparing tables 3.7 and 3.8, the difference in percentages with only 1 isoform was lower, meaning it had some impact on the decrease of the percentages, but it couldn't be the only reason (Tables 3.7 and 3.8).

Through Trinotate, it was possible to annotate several transcripts. The assembly containing all the tissues had generally lower percentages of annotated transcripts since it had more transcripts that might not correspond to actual coding genes. It may be just caused by the number of total genes it expresses or different isoforms for the genes Trinotate found. The midgut tissue assembly has captured the highest genes expressed in comparison to all the other tissues studied, gonad and white muscle tissues assemblies had the most genes annotated via similarity against SwissProt curated database as opposed to the brain plus pituitary tissue assembly with the least (Table 3.5). The difference in percentage of annotated transcripts before and after Transrate indicates a lot of possible pre-mRNA or contaminations were filtered. Some real genes that are specific to the teleosts/clupeiforms that the sardine belongs to might not be present in the SwissProt database. Transrate also seemed to better benefit assemblies of various tissues mixed then of specific tissues as some percentages for the caudal fin tissue assembly got higher (Tables 3.6 and 3.7). Relative small percentages might be the result of not many similar fishes having their transcriptome fully annotated on the searched databases. For the rest of the results, different tissues were assumed as different conditions although it is not the norm and was merely to better visualise the differences and the assembly from all the tissues was used.

3.3.2 Transcript Quantification

Results from the Trinity Transcript Quantification scripts allowed the plotting of a matrix of TPM values. A linear regression was then made for total gene and transcripts to determinate the expression between 10 and 100 TPM, granting 12747expressed genes and 13732 expressed transcripts, and a line across the neg_min_tpm was drawn across -10, granting 26053 genes and 28211 transcripts, indicating the number of genes and transcripts respectively expressed by at least 10 TPM. Heatmaps were generated for all the tissues against each other but the one better representing the cluster of tissues was with the white muscle tissue compared against the others.

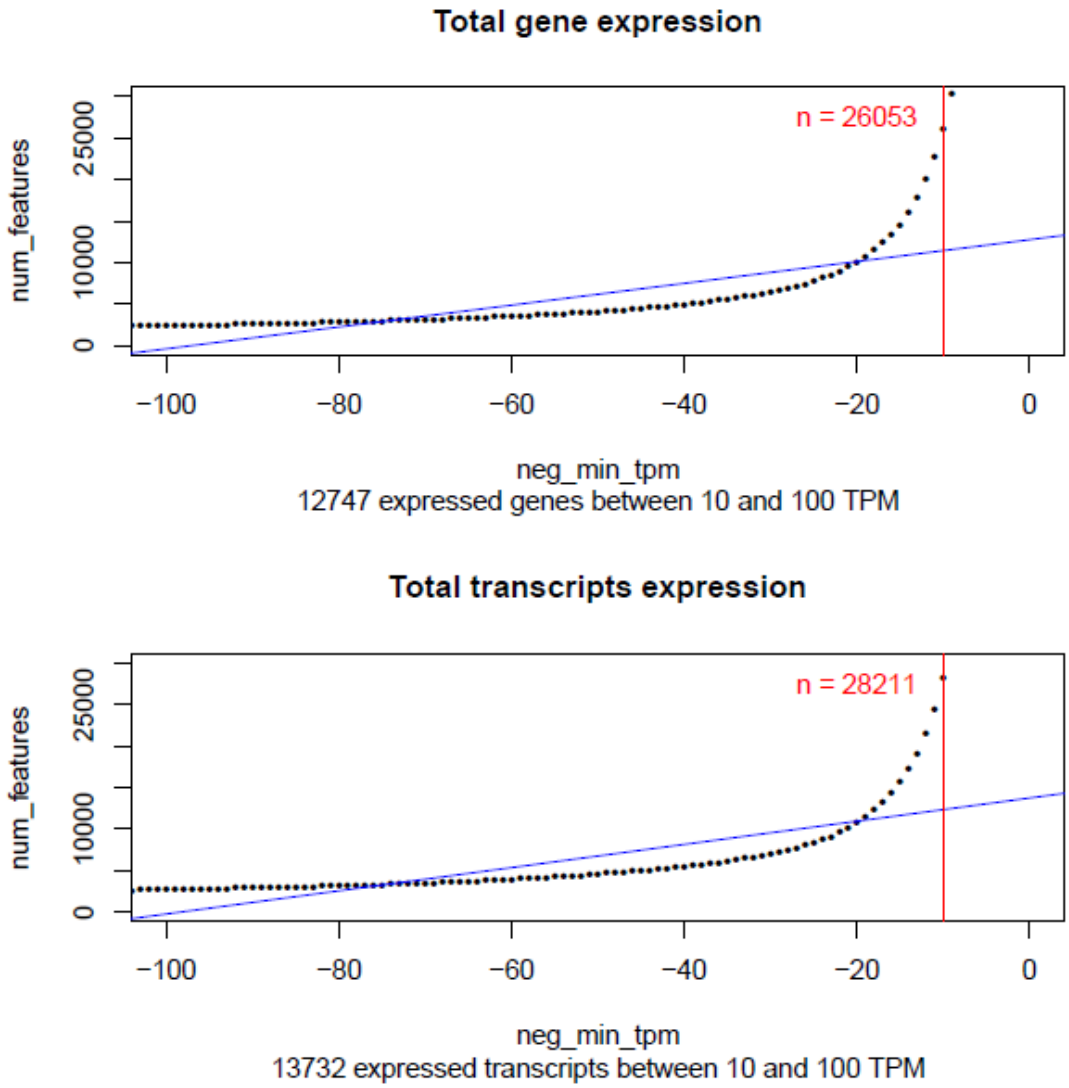

Figure 3.2: Trinity Transcript Quantification. Filtered to show up to 100 TPM. Linear regression between 10 TPM to 100 TPM in blue, 10 TPM in red.

While the Trinity Transcript Quantification give us only a possible exaggerated estimate it can still give us a trusty value on the number of genes and transcripts that are expressed by between 10 and 100 TPM. As expected there are more transcripts than genes expressed on a given TPM, for a gene may have different transcript isoforms (Figure 3.2).

3.3.3 Tissue-specific genes

Genes considered with tissue-specific expression ranged from 64 to 1189 per tissue, and around 64% of the specific genes found were annotated. The brain plus pituitary tissue stood out with the most tissue-specific genes, while the kidney had the least tissue-specific genes. Some of the predicted tissue-speficic genes on the heatmap appear to have low expression in other tissues, since the treashold considered was 95% of the total expression.

Table 3.9: Number of tissue-specific genes predicted in different tissues, and respective annotated and non-annotated but with ORF detected.

Tissues	Specific genes	Annotated	Non-annotated (ORF detected)
Gill + Branchial Arch	580	372	81
Liver	314	225	29
Spleen	132	85	20
Gonad	924	763	82
Midgut	589	361	59
White Muscle	393	267	42
Red Muscle	222	145	17
Kidney	64	36	7
Head Kidney	122	83	12
Brain + Pituitary	1189	619	117
Caudal Fin	411	224	56

Of the tissue-specific, the first most significant 10 genes that matched a gene from swissprot from each tissue were used to create a list with their corresponding False Discovery Rate (FDR). The name of the corresponding protein was obtain through the website/database Uniprot, along with its information such as function and expression.

Table 3.10: List of the top most significant 10 annotated tissue-specific genes of each tissue with the gene ID from UniProtKB and the respective protein name and FDR.

Tissue	UniProtKB entry	Identified protein	FDR

Gill + Branchial Arch	PL8L1_MOUSE	PLAC8-like protein 1	2.95e-47
	CAH6_BOVIN	Carbonic anhydrase 6	1.44e-45
	MFAP4_BOVIN	Microfibril-associated glycoprotein 4	1.82e-43
	MFAP4_BOVIN	Microfibril-associated glycoprotein 4	1.73e-42
	K1C13_ONCMY	Keratin, type I cytoskeletal 13	2.03e-40
	CAH4_BOVIN	Carbonic anhydrase 4	2.50e-39
	MFAP4_BOVIN	Microfibril-associated glycoprotein 4	1.37e-36
	MFAP4_BOVIN	Microfibril-associated glycoprotein 4	1.40e-35
	OIT3_BOVIN	Oncoprotein-induced transcript 3 protein	5.03e-33
	ACBG2_XENLA	Long-chain-fatty-acid--CoA ligase ACSBG2	2.81e-32
Liver	INTLP_ALLMI	Intelectin-like protein	3.20e-57
	SHBG_RABIT	Sex hormone-binding globulin	1.93e-56
	MFAP4_BOVIN	Microfibril-associated glycoprotein 4	7.85e-56
	CO8B_ONCMY	Complement component C8 beta chain	1.70e-55
	THRB_BOVIN	Prothrombin	4.49e-54
	HABP2_HUMAN	Hyaluronan-binding protein 2	1.87e-53
	ITIH4_BOVIN	Inter-alpha-trypsin inhibitor heavy chain H4	3.92e-53
	C1QL4_MOUSE	Complement C1q-like protein 4	4.32e-53
	FA10_CHICK	Coagulation factor X	6.63e-53
	ANT3_MOUSE	Antithrombin-III	8.70e-53
Spleen	AICDA_BOVIN	Single-stranded DNA cytosine deaminase	2.47e-14
	KLK7_MOUSE	Kallikrein-7	1.09e-12
	LDLR_MOUSE	Low-density lipoprotein receptor	1.91e-12
	ACM2_MOUSE	Muscarinic acetylcholine receptor M2	9.52e-11
	IFI44_HUMAN	Interferon-induced protein 44	3.52e-09
	IRF4_HUMAN	Interferon regulatory factor 4	1.26e-08
	MUCM_ICTPU	Ig mu chain C region membrane-bound form	2.05e-08
	GIMA7_HUMAN	GTPase IMAP family member 7	1.06e-07
	CLC4E_HUMAN	C-type lectin domain family 4 member E	1.83e-07
	AA3R_RABIT	Adenosine receptor A3	2.03e-07
Gonad	TMPS4_MOUSE	Transmembrane protease serine 4	6.55e-34
	M10L1_HUMAN	RNA helicase Mov10l1	8.40e-30
	SPIR2_MOUSE	Protein spire homolog 2	3.27e-29

	HENMT_DANRE	Small RNA 2'-O-methyltransferase	4.36e-29
	BRACA_DANRE	T-box transcription factor T-A	4.43e-29
	ASZ1_MACEU	Ankyrin repeat, SAM and basic leucine zipper domain-containing protein 1	8.60e-29
	PATL2_XENLA	Protein PAT1 homolog 2	3.80e-28
	ARHG5_MOUSE	Rho guanine nucleotide exchange factor 5	3.80e-28
	4F2_RABIT	4F2 cell-surface antigen heavy chain	4.35e-28
	STK31_HUMAN	Serine/threonine-protein kinase 31	7.39e-28
Midgut	RN128_PONAB	E3 ubiquitin-protein ligase RNF128	3.51e-42
	ANS4B_HUMAN	Ankyrin repeat and SAM domain-containing protein 4B	3.00e-41
	ACE2_MOUSE	Angiotensin-converting enzyme 2	2.03e-40
	C1QL4_MOUSE	Complement C1q-like protein 4	2.97e-40
	TRYP_PLEPL	Trypsin	3.95e-40
	CLCA1_BOVIN	Calcium-activated chloride channel regulator 1	2.38e-38
	ABCG2_HUMAN	ATP-binding cassette sub-family G member 2	6.86e-38
	MFAP4_MOUSE	Microfibril-associated glycoprotein 4	7.28e-38
	MFAP4_BOVIN	Microfibril-associated glycoprotein 4	8.45e-38
	REL3_MOUSE	Relaxin-3	1.02e-37
White Muscle	IGFN1_HUMAN	Immunoglobulin-like and fibronectin type III domain-containing protein 1	4.21e-35
	MT21E_HUMAN	Putative methyltransferase-like protein 21E pseudogene	4.11e-28
	XIRP2_MOUSE	Xin actin-binding repeat-containing protein 2	4.35e-28
	GATM_DANRE	Glycine amidinotransferase, mitochondrial	1.57e-27
	TECR_RAT	Very-long-chain enoyl-CoA reductase	1.57e-27
	SC4AB_DANRE	Sodium channel protein type 4 subunit alpha B	3.24e-27
	CMYA5_HUMAN	Cardiomyopathy-associated protein 5	2.82e-26
	TM233_MOUSE	Transmembrane protein 233	1.13e-24
	YD023_HUMAN	Putative uncharacterized protein FLJ45035	6.13e-24
	CLCN1_CANLF	Chloride channel protein 1	1.40e-23
Red Muscle	MYPC3_CHICK	Myosin-binding protein C, cardiac-type	2.97e-58
	MYLK2_RABIT	Myosin light chain kinase 2, skeletal/cardiac muscle	6.54e-46
	IGFN1_MOUSE	Immunoglobulin-like and fibronectin type III domain-containing protein 1	1.37e-44
	ASB18_HUMAN	Ankyrin repeat and SOCS box protein 18	4.74e-39

	MYH7_HORSE	Myosin-7	3.37e-37
	A33_PLEWA	Zinc-binding protein A33	1.86e-35
	ASB15_BOVIN	Ankyrin repeat and SOCS box protein 15	1.22e-34
	MYH7_HUMAN	Myosin-7	6.22e-34
	KLH34_HUMAN	Kelch-like protein 34	1.26e-32
	RYR3_RABIT	Ryanodine receptor 3	2.55e-31
Kidney	S12A3_RAT	Solute carrier family 12 member 3	6.02e-22
	CLCKB_XENLA	Chloride channel protein ClC-Kb	2.66e-06
	PEPC_CALJA	Gastricsin	8.76e-06
	CHIT1_HUMAN	Chitotriosidase-1	9.42e-06
	MUC2_MOUSE	Mucin-2	5.85e-05
	AGRE2_CANLF	Adhesion G protein-coupled receptor E2	6.41e-05
	LRP2_HUMAN	Low-density lipoprotein receptor-related protein 2	6.86e-05
	BSND_MOUSE	Barttin	7.55e-05
	TM147_DANRE	Transmembrane protein 147	8.78e-05
	LRP2_HUMAN	Low-density lipoprotein receptor-related protein 2	1.56e-04
Head Kidney	CP11A_ONCMY	Cholesterol side-chain cleavage enzyme, mitochondrial	1.97e-06
	ENDD1_MOUSE	Endonuclease domain-containing 1 protein	4.79e-06
	CFA52_HUMAN	Cilia- and flagella-associated protein 52	5.99e-06
	GTR5_SHEEP	Solute carrier family 2, facilitated glucose transporter member 5	6.35e-06
	CD045_HUMAN	Uncharacterized protein C4orf45	6.35e-06
	RFT2_SALSA	Riboflavin transporter 2	1.36e-05
	PCKGC_CHICK	Phosphoenolpyruvate carboxykinase, cytosolic [GTP]	1.57e-05
	PTHD3_MOUSE	Patched domain-containing protein 3	1.61e-05
	IRX5_XENTR	Iroquois-class homeodomain protein irx-5	1.62e-05
	IRX3_XENTR	Iroquois-class homeodomain protein irx-3	1.93e-05
Brain + Pituitary	VIME_PANTR	Vimentin	4.86e-52
	S6A11_RAT	Sodium- and chloride-dependent GABA transporter 3	4.64e-49
	CBLN1_MOUSE	Cerebellin-1	9.25e-47
	SNAB_MOUSE	Beta-soluble NSF attachment protein	1.29e-45
	NFM_PIG	Neurofilament medium polypeptide	2.74e-44
	FABP7_HUMAN	Fatty acid-binding protein, brain	9.55e-44
	SN25B_CARAU	Synaptosomal-associated protein 25-B	3.93e-40

	C1QT4_MOUSE	Complement C1q tumor necrosis factor-related protein 4	6.59e-40
	PXN1_XENLA	Pentraxin fusion protein	1.69e-39
	NPTX1_RAT	Neuronal pentraxin-1	1.88e-39
Caudal Fin	PPN_MOUSE	Papilin	2.35e-40
	EMIL2_HUMAN	EMILIN-2	4.42e-31
	CO6A3_CHICK	Collagen alpha-3(VI) chain	5.29e-25
	HPLN1_RAT	Hyaluronan and proteoglycan link protein 1	4.18e-23
	PGS1_XENLA	Biglycan	2.66e-22
	CO2A1_MOUSE	Collagen alpha-1(II) chain	1.34e-18
	COCA1_MOUSE	Collagen alpha-1(XII) chain	1.18e-17
	LIPHB_XENLA	Lipase member H-B	2.01e-17
	PRRX1_HUMAN	Paired mesoderm homeobox protein 1	3.26e-16
	PPN_HUMAN	Papilin	1.75e-14

Most of the top 10 annotated tissue-specific genes has its function expected from each tissue and its highest expression levels in the respective tissues across taxa (Table 3.10). Some gill plus branchial arch specific genes have their highest expression levels in the lungs of the organisms (mammals) used for its annotation and although gills and lung are totally different organs they share one common main function, to obtain oxygen for the organism. Expression levels were also checked on the website Expression Atlas (www.ebi.ac.uk/gxa/home). The red muscle tissue and the white muscle tissue, although similar, slightly differ in their function. While the red muscle is used for sustained swimming speed, the white muscle is used for prolonged high swimming speed[28]. So, their tissue-specific genes differ based on their utility, for example the myosin-7 red muscle specific protein is a constituent of slow muscles and the very-long-chain enoyl-CoA reductase white muscle specific protein regulates fast muscles[29].

The microfibril-associated glycoprotein 4 (MFAP4) gene stood out as specifically expressed multiple times across tissues, especially in the gill plus branchial arch tissue. According to Uniprot annotations, its function seems related with calcium-dependent cell adhesion or intercellular interactions and its highest expression level in the mammal organisms is in the lungs. MFAP4 also seems to have a role on the immune system in fishes and several copies of the gene with different expression levels across tissues[30].

The myosin-7 (MYH7) and papilin (PPN) genes also appear two times in the same tissue, red muscle and caudal fin respectively, annotated from different organisms.

To analyse this particularity, the Ensembl website (Ensembl release 94) was used to look for paralogues and orthologues of the MFAP4 gene of the organism that was used for annotation of most of this gene across the list, the cow (*Bos taurus*). The numbers of the species represented belong to the species with annotation in the Ensembl database (Table 3.11).

Table 3.11: Summary of orthologues of the MFAP4 gene (Figure adapted from http://Oct2018.archive.ensembl.org/Bos_taurus/Gene/Compara_Ortholog?db=core;g=ENSBTAG00000006187;r=19:34685357-34687892;t=ENSBTAT00000008130).

Species set	With 1:1 orthologues	With 1 : many orthologues
Primates (*24 species*) Humans and other primates	24	0
Rodent and related species (*24 species*) Rodents, lagomorphs and tree shrews	23	1
Laurasiatheria (*16 species*) Carnivores, ungulates and insectivores	14	0
Placental Mammals (*69 species*) All placental mammals	64	2
Sauropsida (*7 species*) Birds and Reptiles	2	0
Fish (*48 species*) Ray-finned fishes	3	41

The information obtained from the gene gain/loss tree confirms that MFAP4 gene is present in a single copy across the Tetrapods but it varied across the fishes. The sardine was situated in a new branch from a node before the Clupeocephala node. The reference gene obtained in the SwissProt annotation was from the cow that belongs to the Tetrapods superclass branch (Figure 3.3).

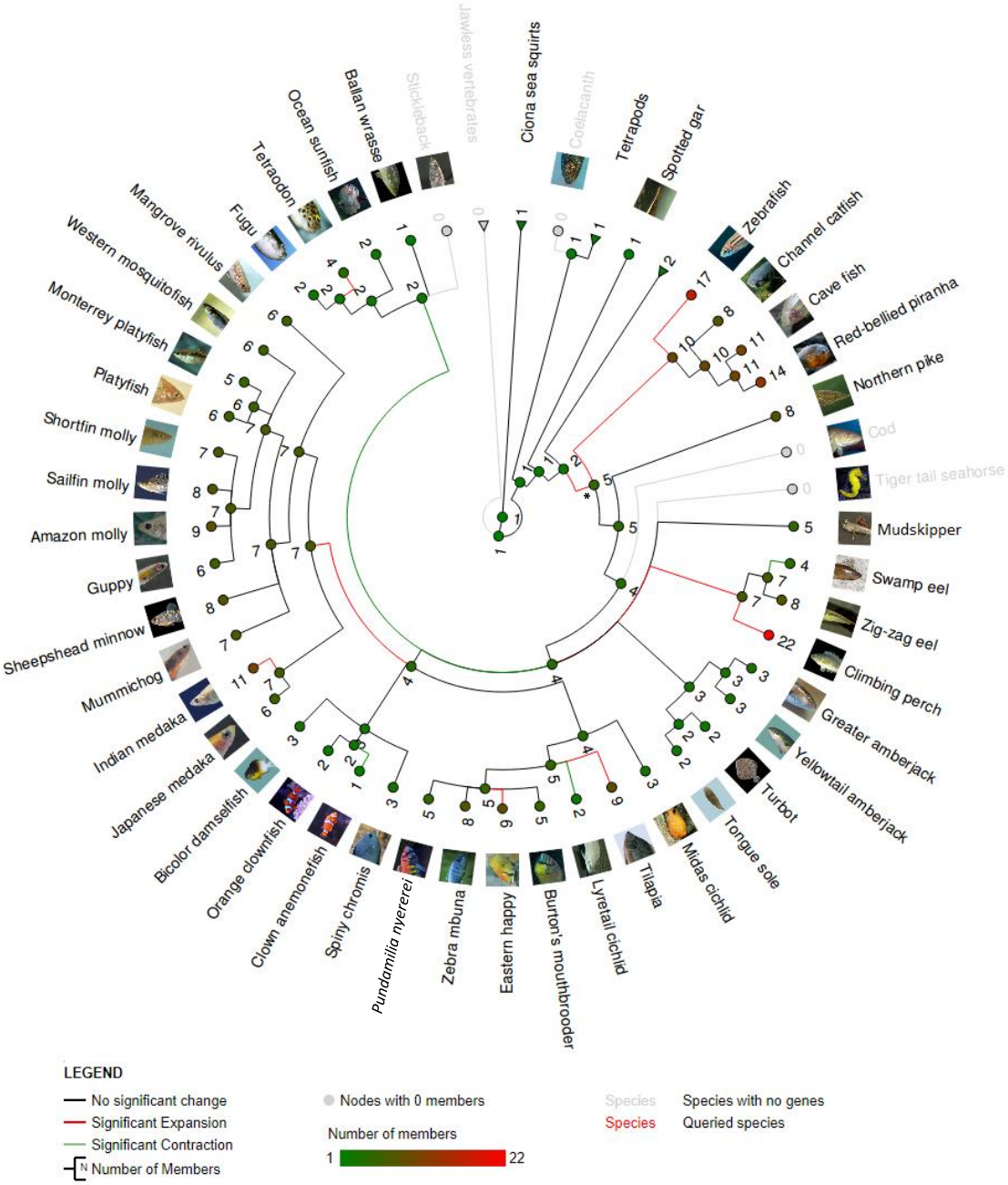

Figure 3.3: MFAP4 gene gain/loss tree. This gene family has significant gene gain or loss events (p-value for the gene family is 0, as computed by CAFE). *Clupeocephala node. (Figure adapted from http://oct2018.archive.ensembl.org/Bos_taurus/Gene/SpeciesTree?db=core;g=ENSBTAG00000006187;r=19:34685357-34687892;t=ENSBTAT00000008130).

The fact that teleost fish contain more than one copy of the MFAP4 gene supports a whole genome duplication event prior to the evolution of teleosts and that several species-specific multiple gene duplications occurred. As for the MYH7 gene, most of the organisms displayed

in the gene gain/loss tree contained only one copy of the gene until the teleost fishes which contained up to 4 copies, although some fish species lost the duplicated gene and contained only a single copy, but the species that contained more than one gene don't seem to have a pattern and the duplication of the gene appears in species across different groups (Figure 3.4). The PPN gene also appears to have been duplicated in the teleost fishes as most contain 2 copies of the gene with some species exceptions containing 3 gene copies while tetrapods only contain one with a few rare exceptions containing 2 (Figure 3.5). The organisms used for the visualization of the gene gain/loss tree for the MYH7 and the PPN genes were horse (*Equus caballus*) and the mouse (*Mus musculus*) respectively.

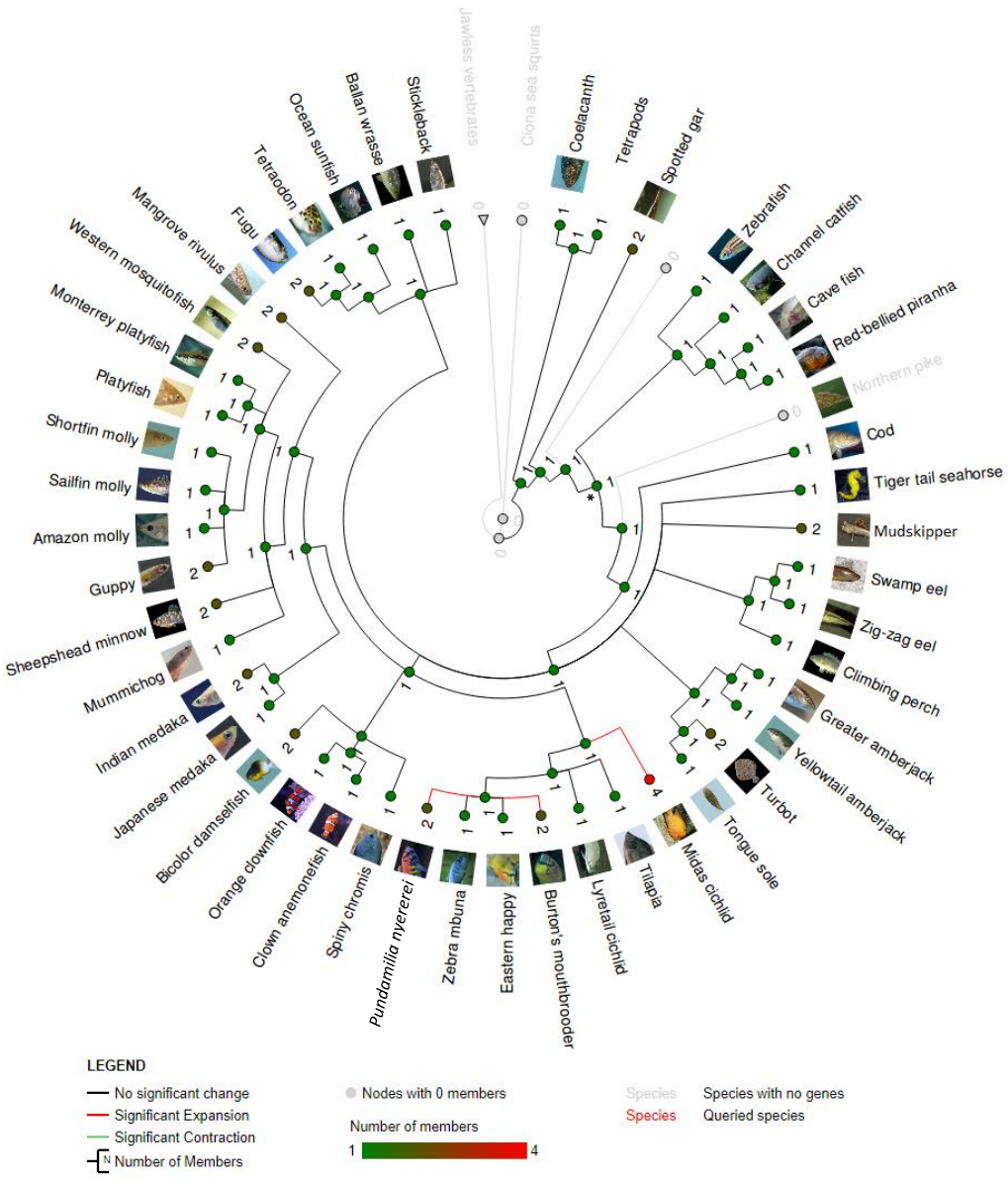

Figure 3.4: MYH7 gene gain/loss tree. This gene family does not have any significant gene gain or loss events (p-value for the gene family is 0.692, as computed by CAFE). *Clupeocephala node. (Figure adapted from http://oct2018.archive.ensembl.org/Equus_caballus/Gene/SpeciesTree?db=core;g=ENSECAG00000019844;r=1:161506573-161527244)

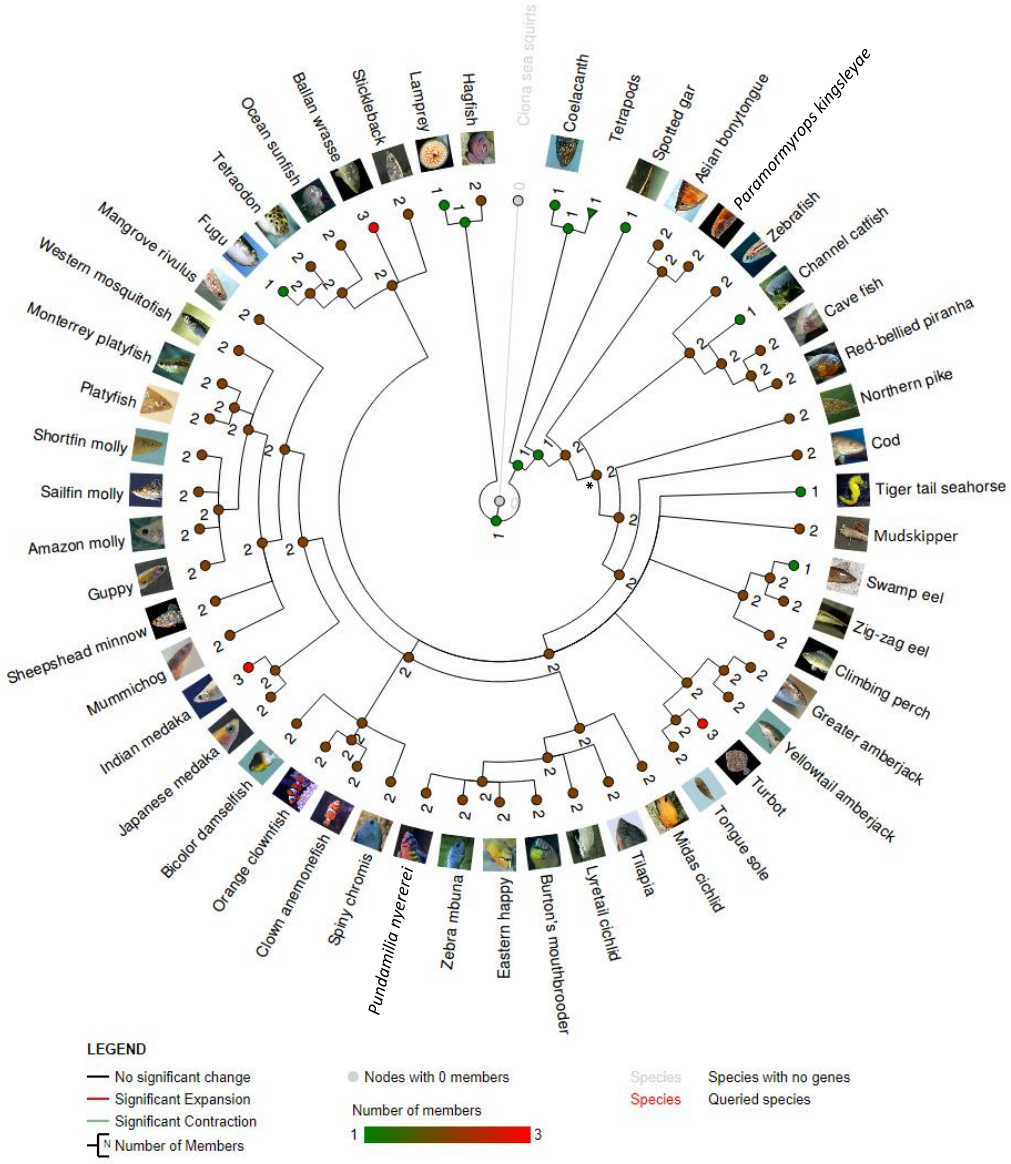

Figure 3.5: PPN gene gain/loss tree. This gene family does not have any significant gene gain or loss events (p-value for the gene family is 0.995, as computed by CAFE). *Clupeocephala node. (Figure adapted from http://oct2018.archive.ensembl.org/Mus_musculus/Gene/SpeciesTree?family=PTHR13723_SF20;g=ENSMUSG00000021223;r=12:83763634-83792382)

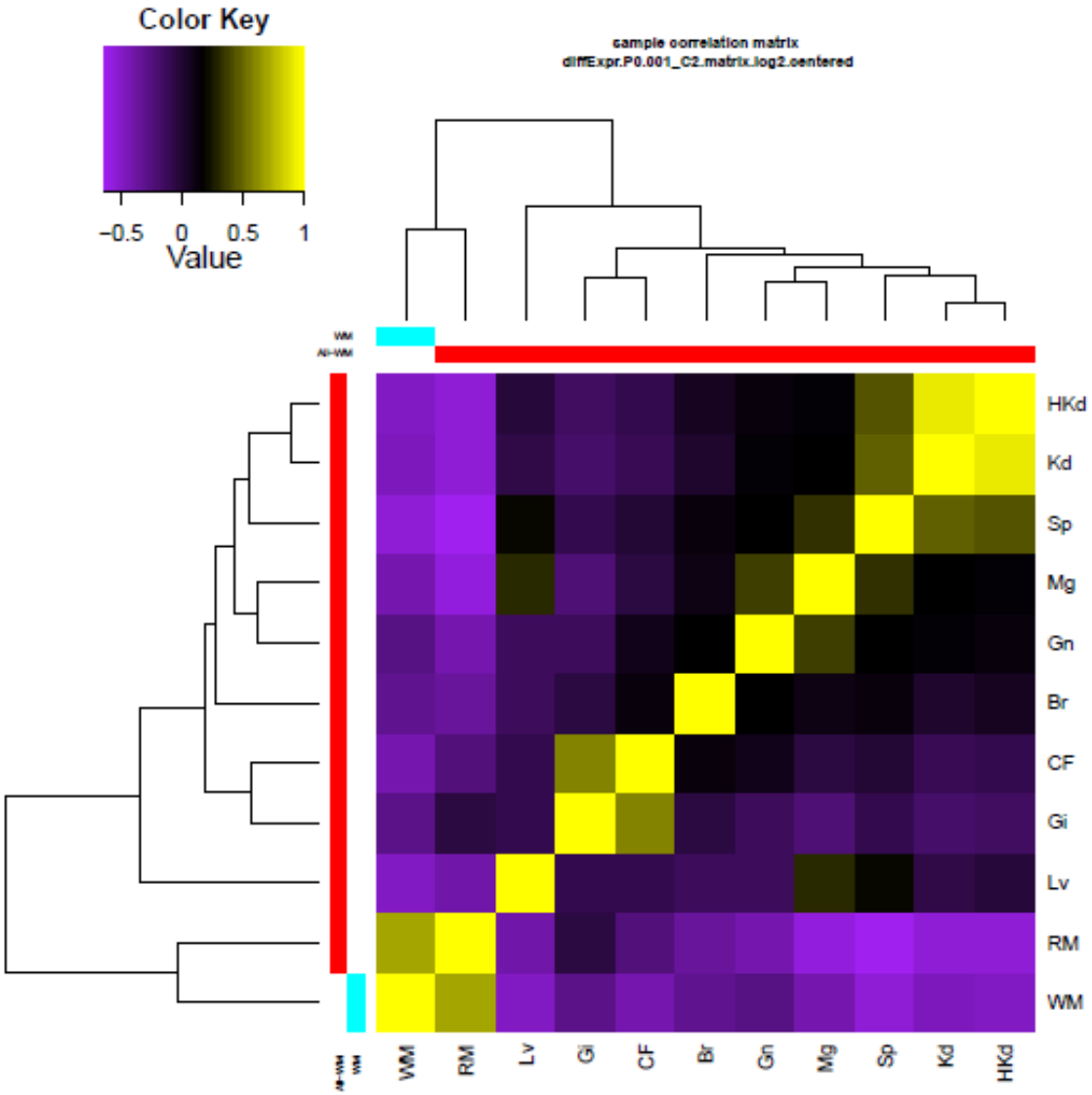

Figure 3.6: Heatmap of the correlation of the different tissues. (Colour represents the value of similarity with purple representing the least similarity and yellow the most. Tissues were clustered according to their similarity and the white muscle tissue was the considered outlier. Abbreviations of the tissues are extended on the list of abbreviations)

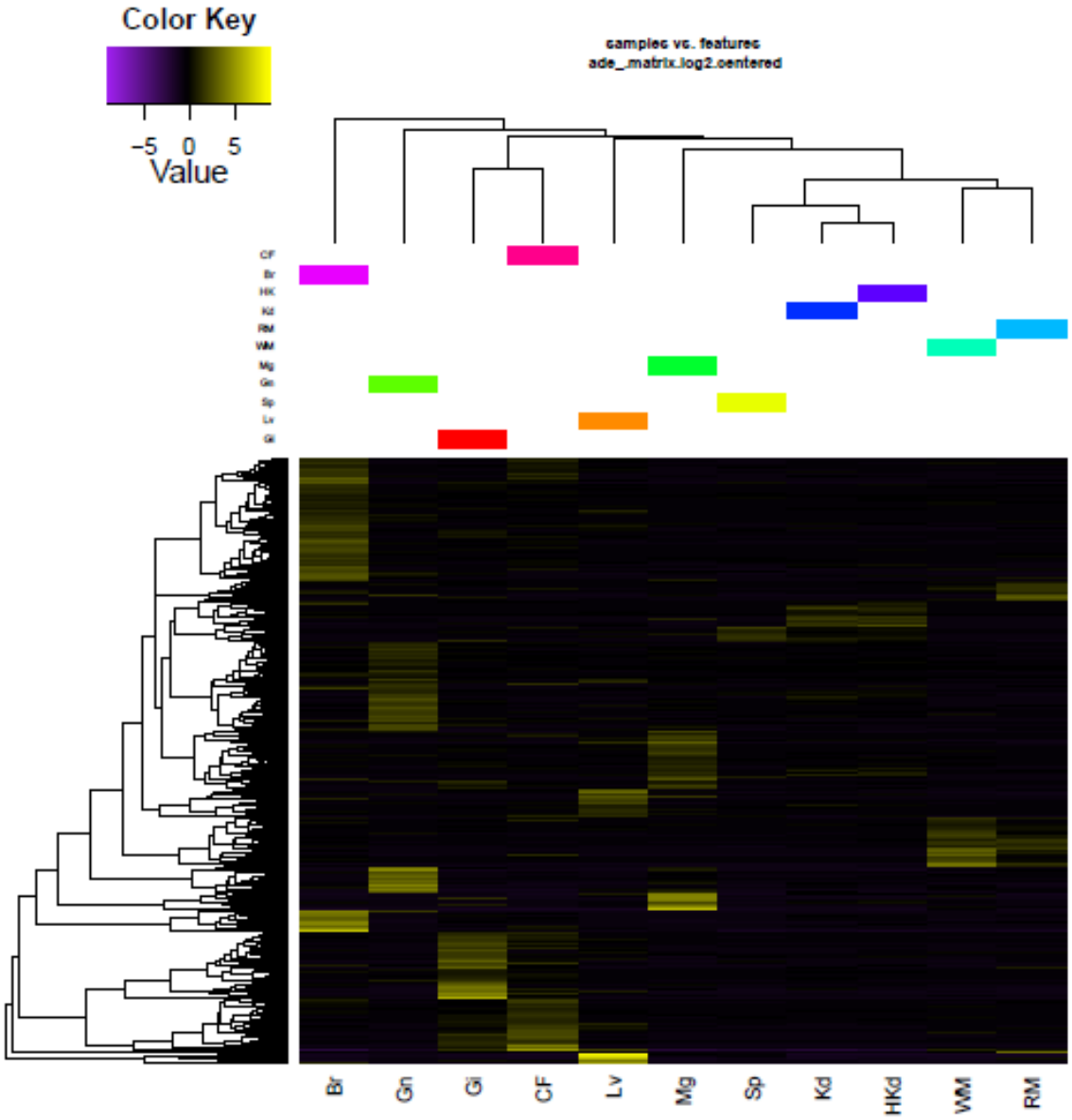

Figure 3.7: Heatmap of tissue-specific genes predicted in different tissues. (Colour represents the value of expression on each tissue with purple representing the least expression and yellow the most. Tissues and genes were clustered according to their expression pattern similarities. Abbreviations of the tissues are extended on the list of abbreviations)

Tissue-specific genes were predicted having these only 11 tissues, which means some specific genes predicted for these tissues may also be expressed on tissues not analysed. The predictions show that the brain plus pituitary tissue had more specific genes as it would be expected from a highly specialized tissue (Table 3.9). The heatmap with the tissue-specific genes cluster the tissues slightly different from the heatmap of correlation of the tissues but maintaining the most similar tissues clustered together, like red muscle with white muscle and kidney with head kidney, and confirming that the brain plus pituitary tissue to be a more specialized tissue as it didn't correspond with a high similarity colour with any other tissue (Figures 3.6 and 3.7).

3.3.4 REVIGO

REVIGO generated scatterplots of gene ontologies (GO) of semantic similarity with colours according to their log10 p-values and sizes according to their log sizes. The tables of the top 10 GOs were sorted from the least dispensability and not assigned in a cluster, some of the GOs on the tables are identified in the scatterplots. In the scatterplots, bubble colours indicate the p-value provided and plot sizes indicate the frequency of the GO terms in the database, the indicated ID term represents the term with the lowest p-value of the cluster and the axis are scales of a matrix of the GO terms' semantic similarities.

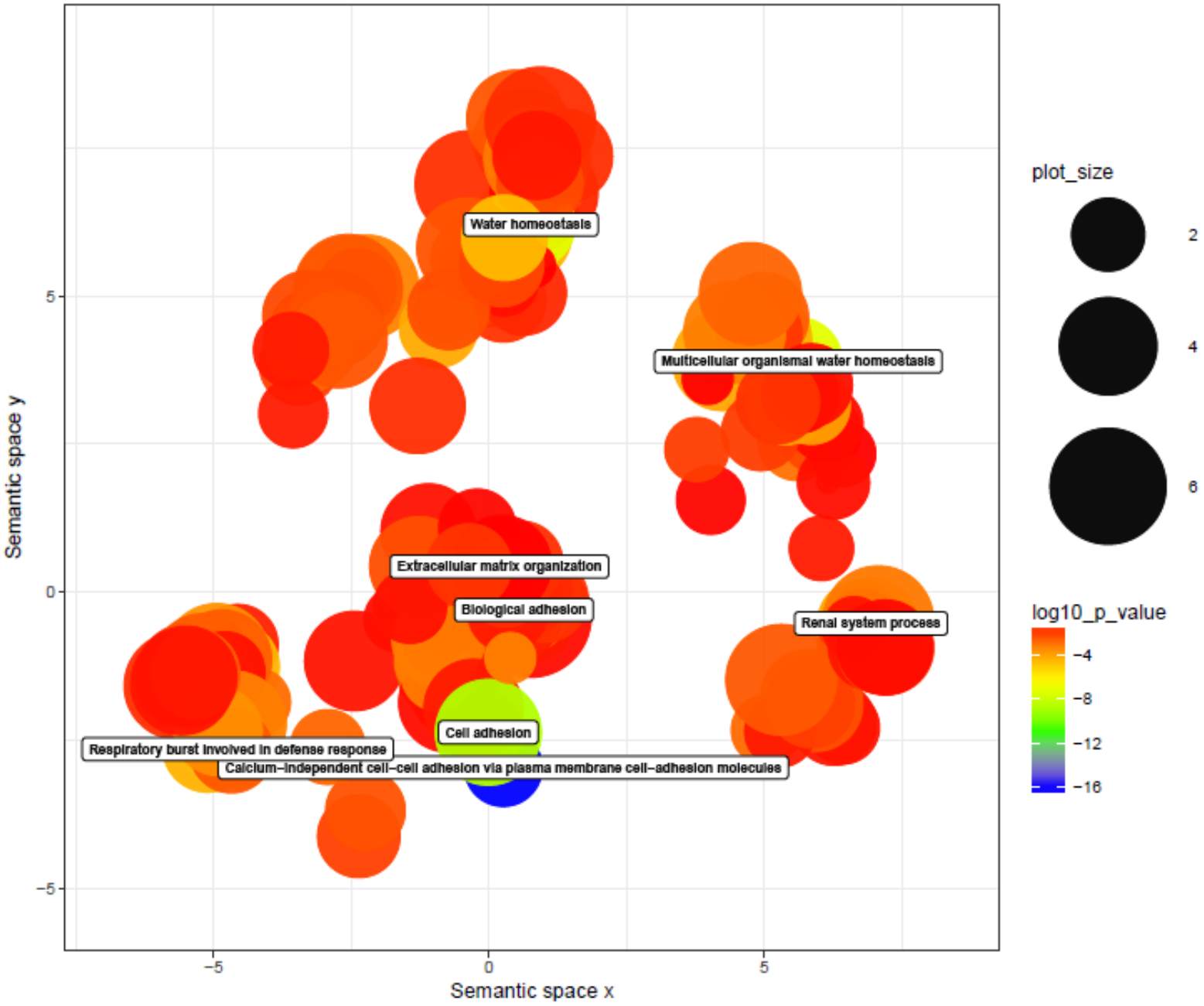

Figure 3.8: Gene ontology scatterplot generated with REVIGO for the gill plus branchial arch tissue.

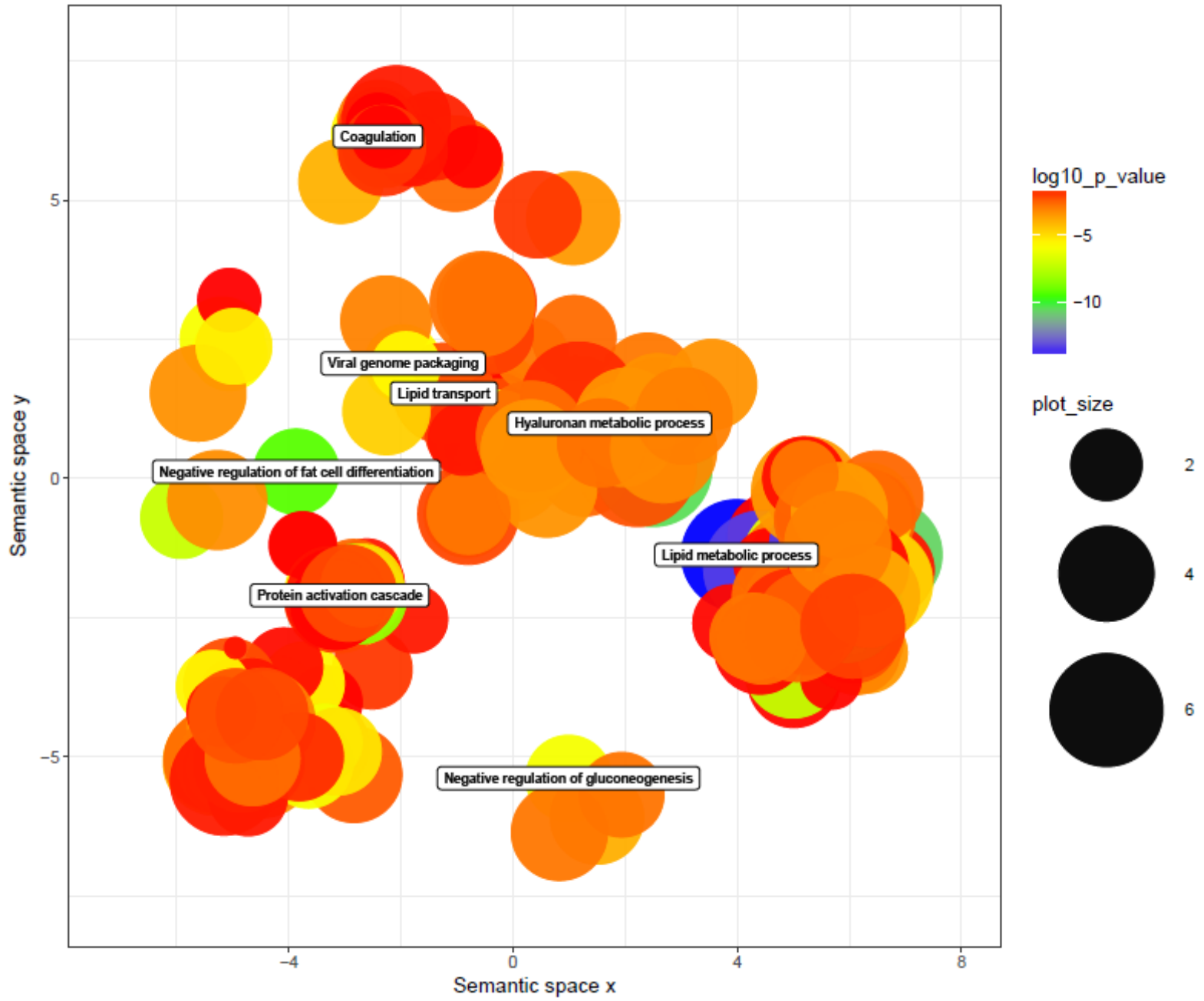

Figure 3.9: Gene ontology scatterplot generated with REVIGO for the liver tissue.

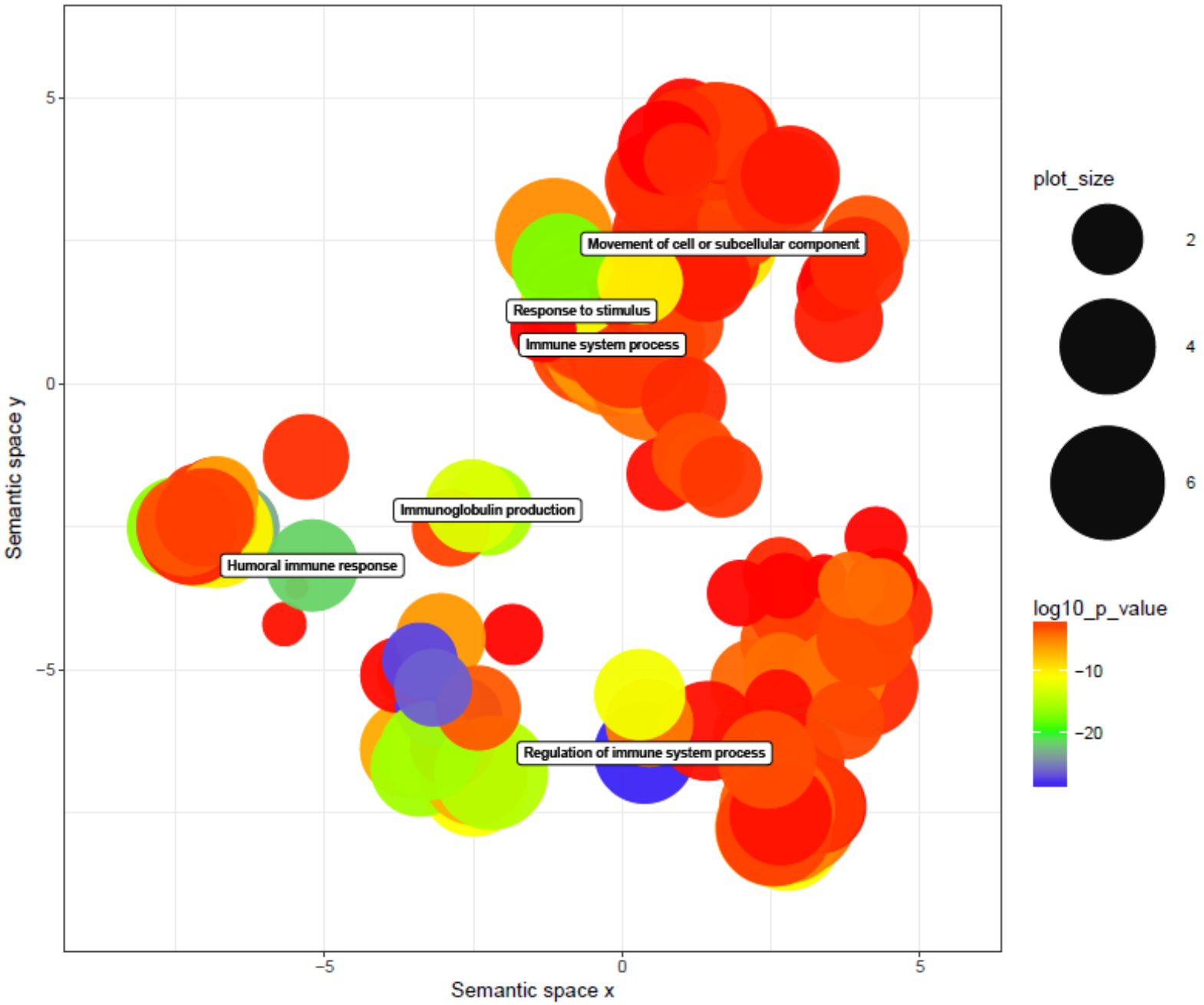

Figure 3.10: Gene ontology scatterplot generated with REVIGO for the spleen tissue.

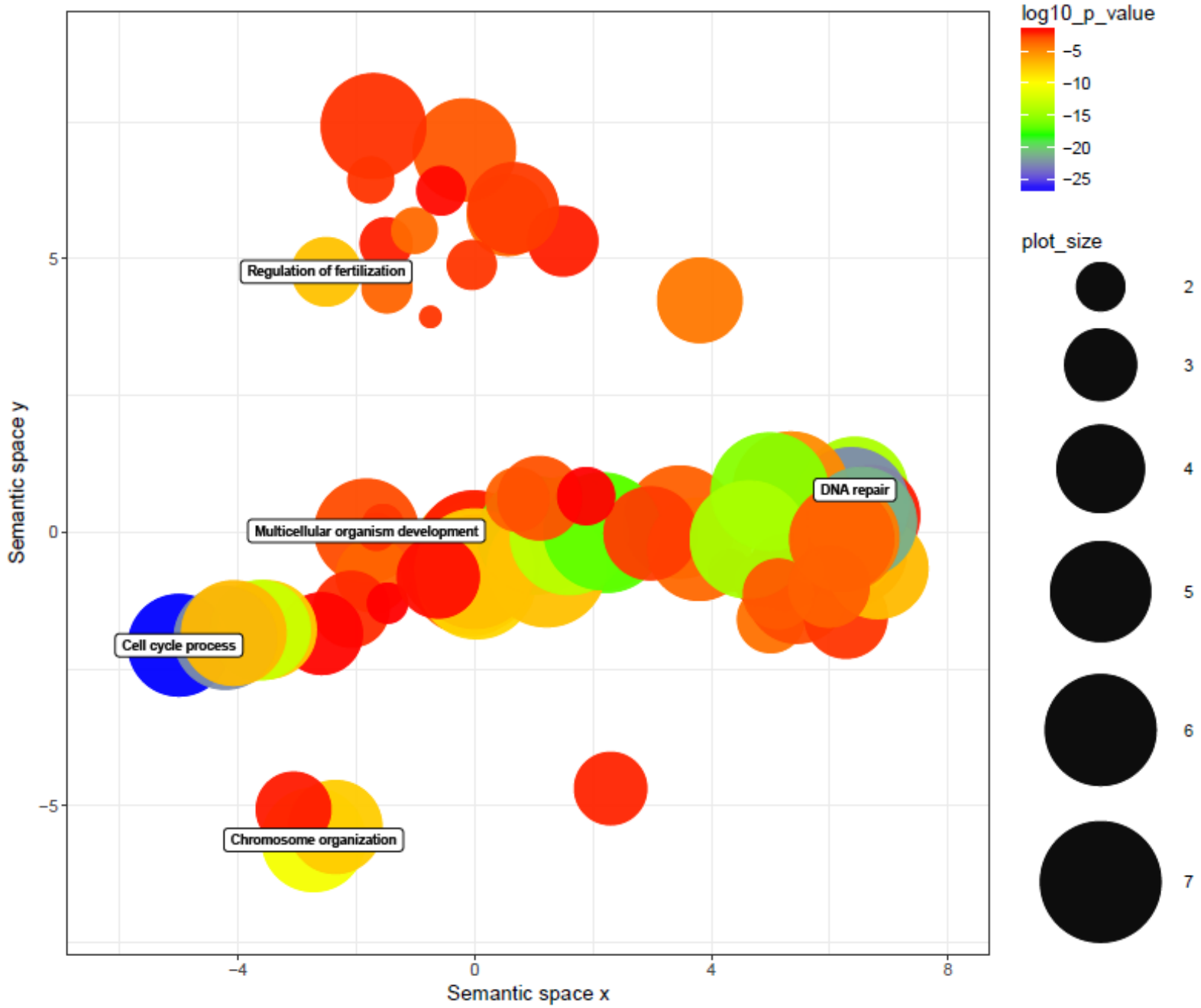

Figure 3.11: Gene ontology scatterplot generated with REVIGO for the gonad tissue.

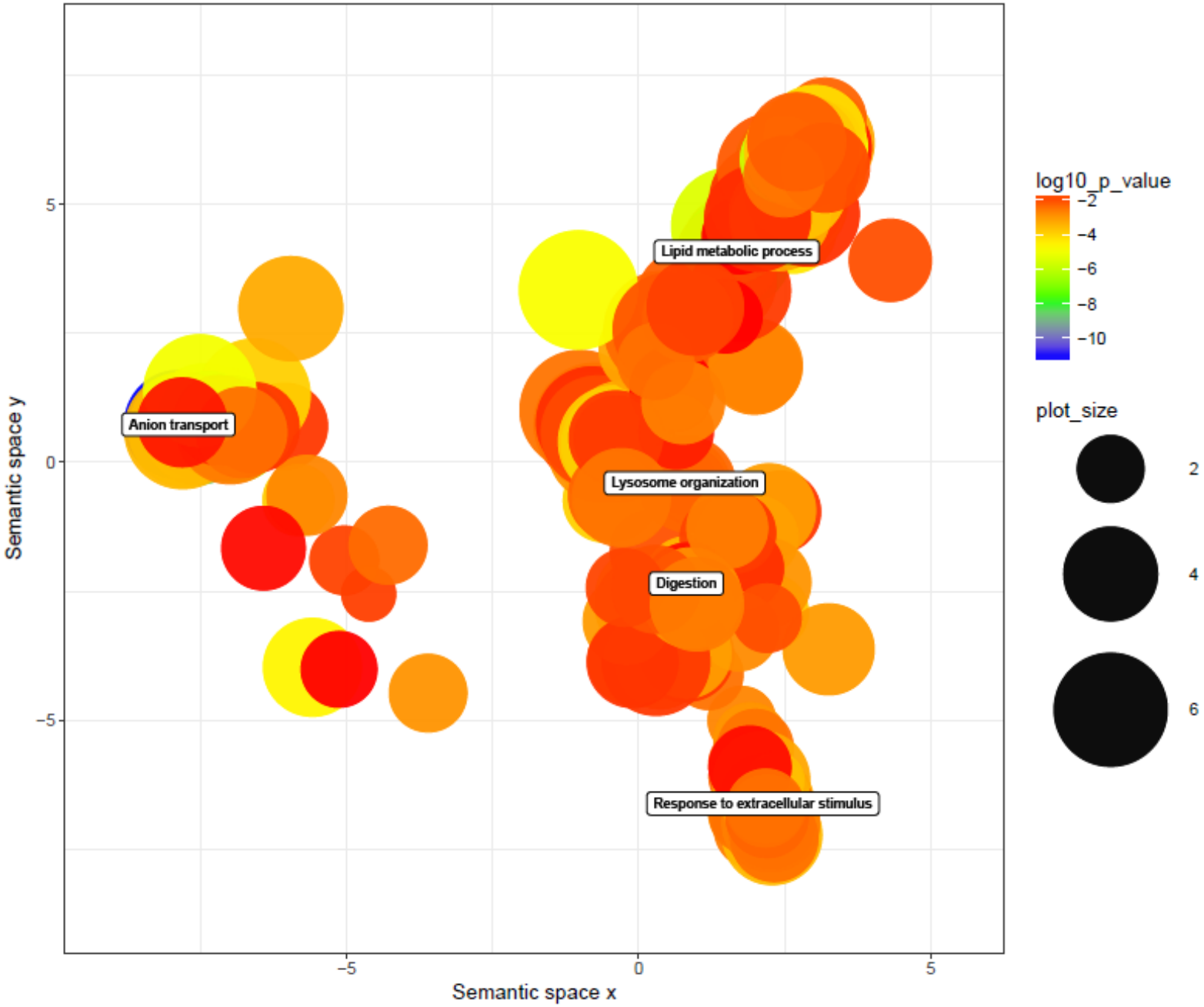

Figure 3.12: Gene ontology scatterplot generated with REVIGO for the midgut tissue.

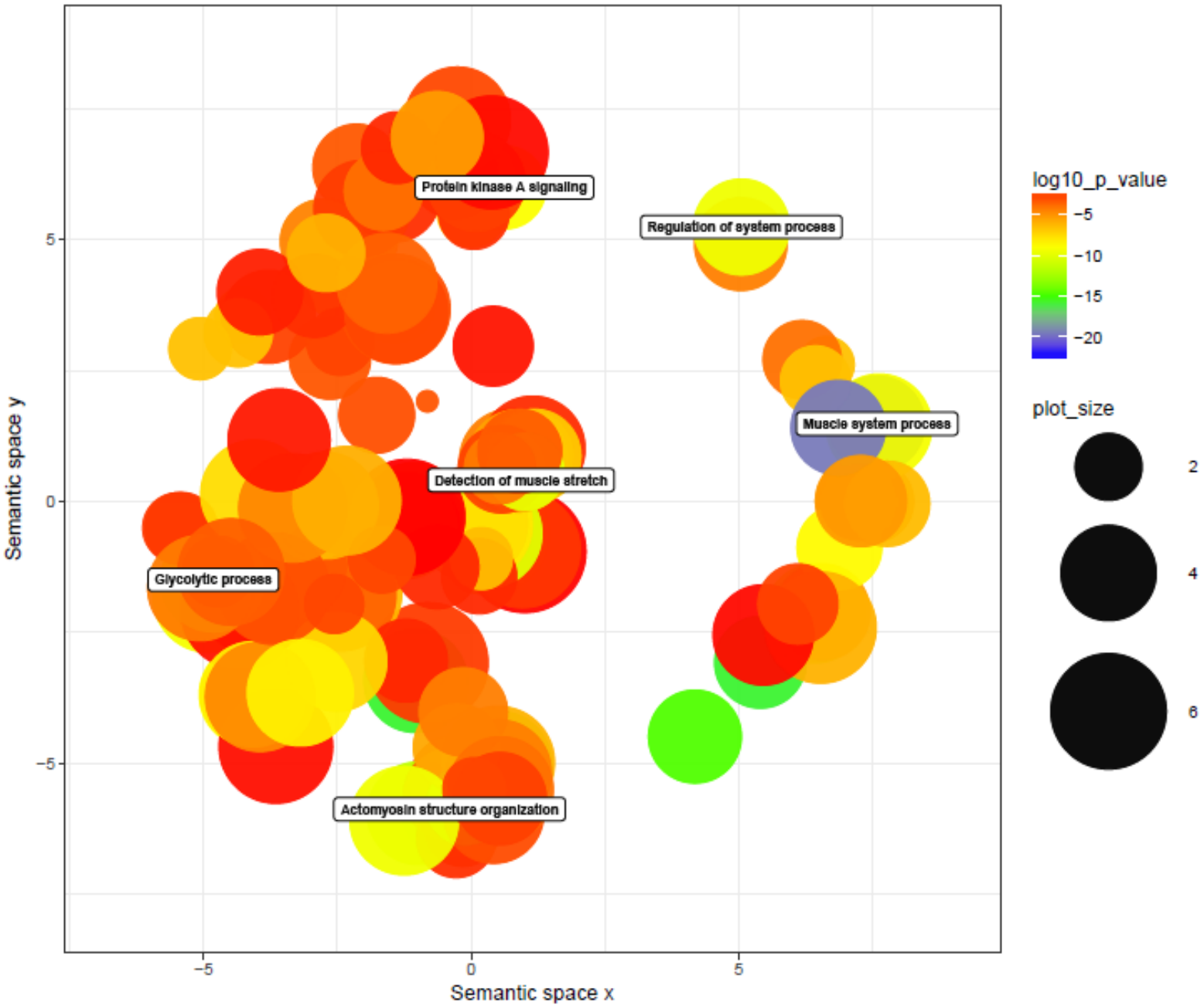

Figure 3.13: Gene ontology scatterplot generated with REVIGO for the white muscle tissue.

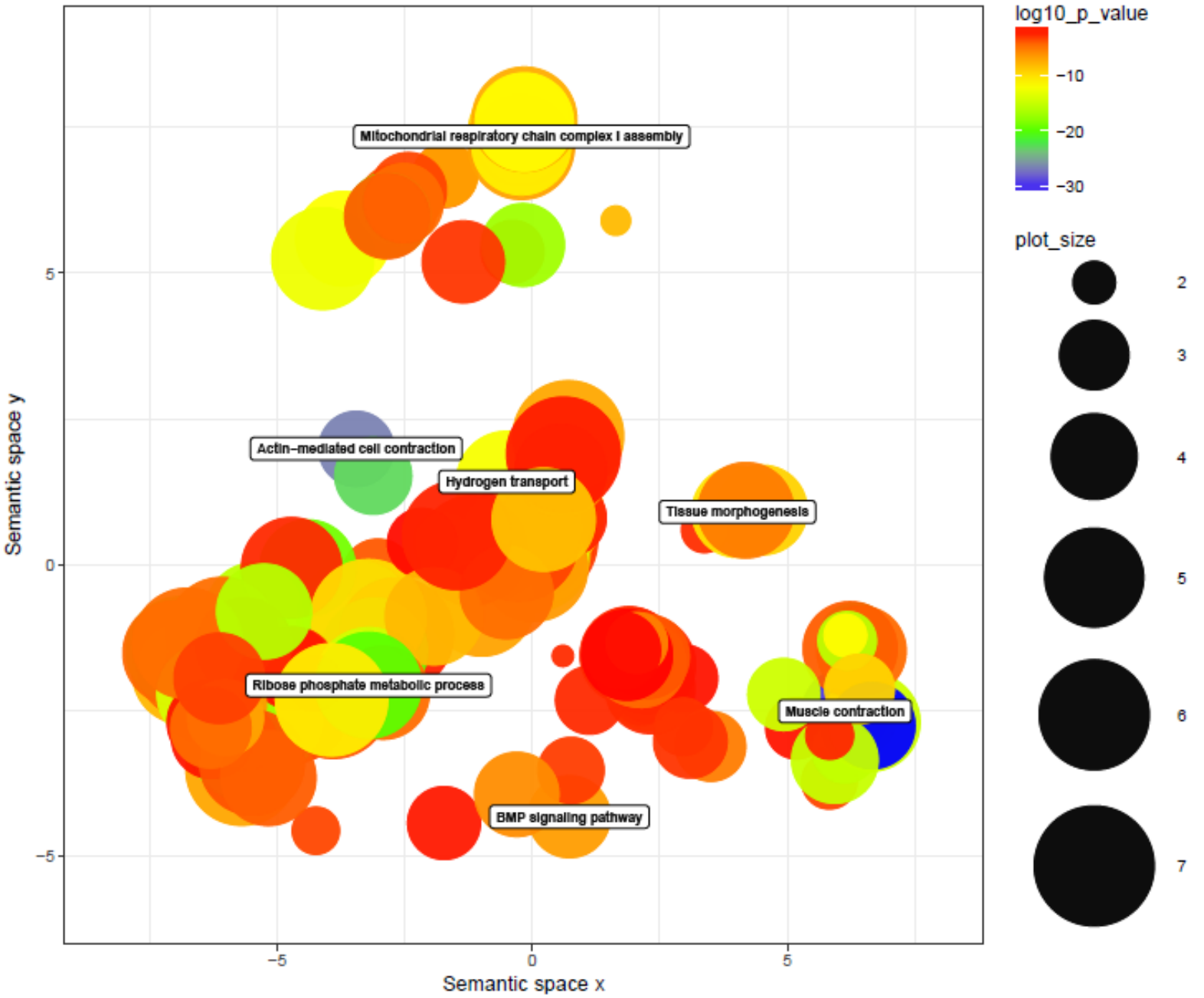

Figure 3.14: Gene ontology scatterplot generated with REVIGO for the red muscle tissue.

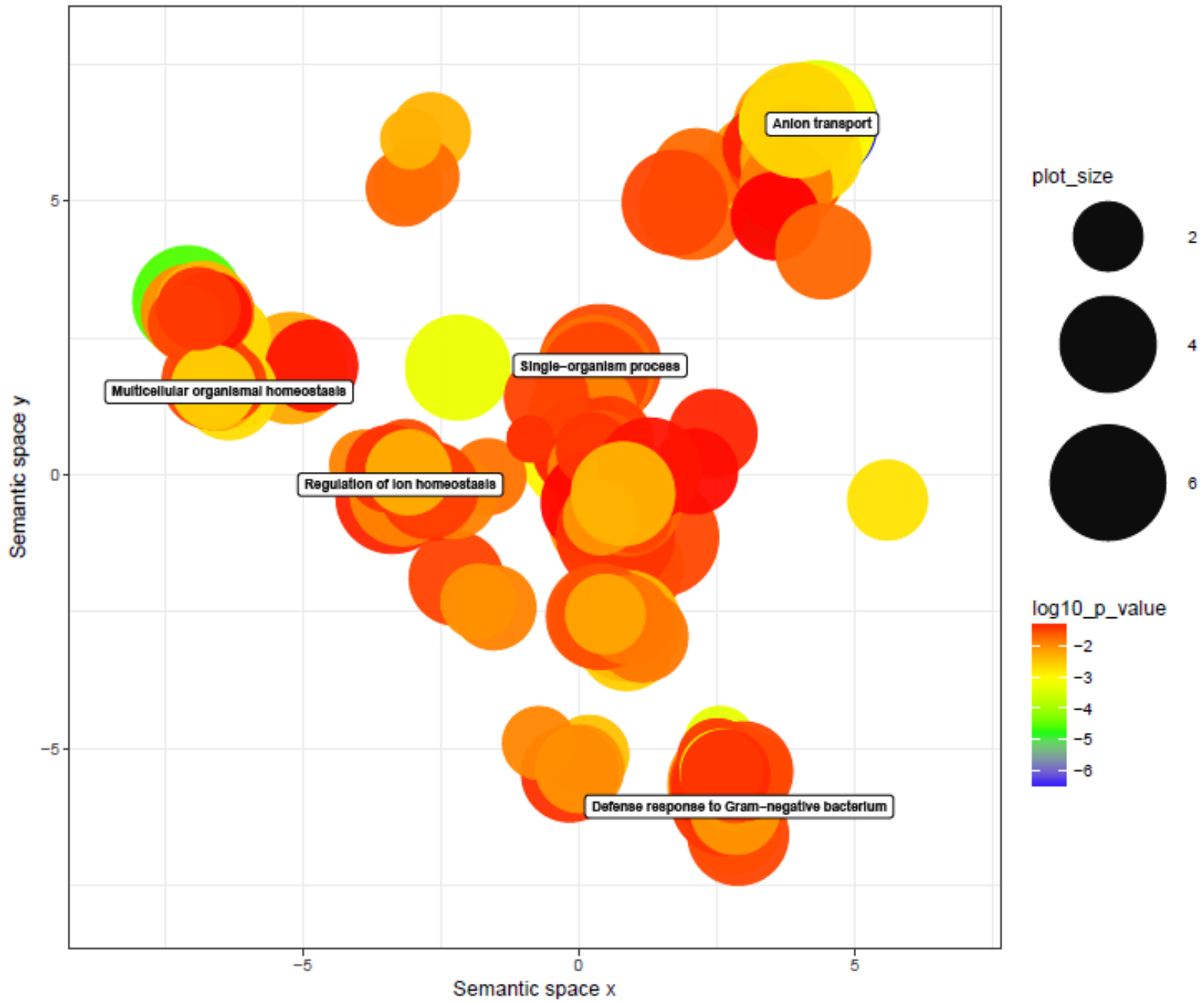

Figure 3.15: Gene ontology scatterplot generated with REVIGO for the kidney tissue.

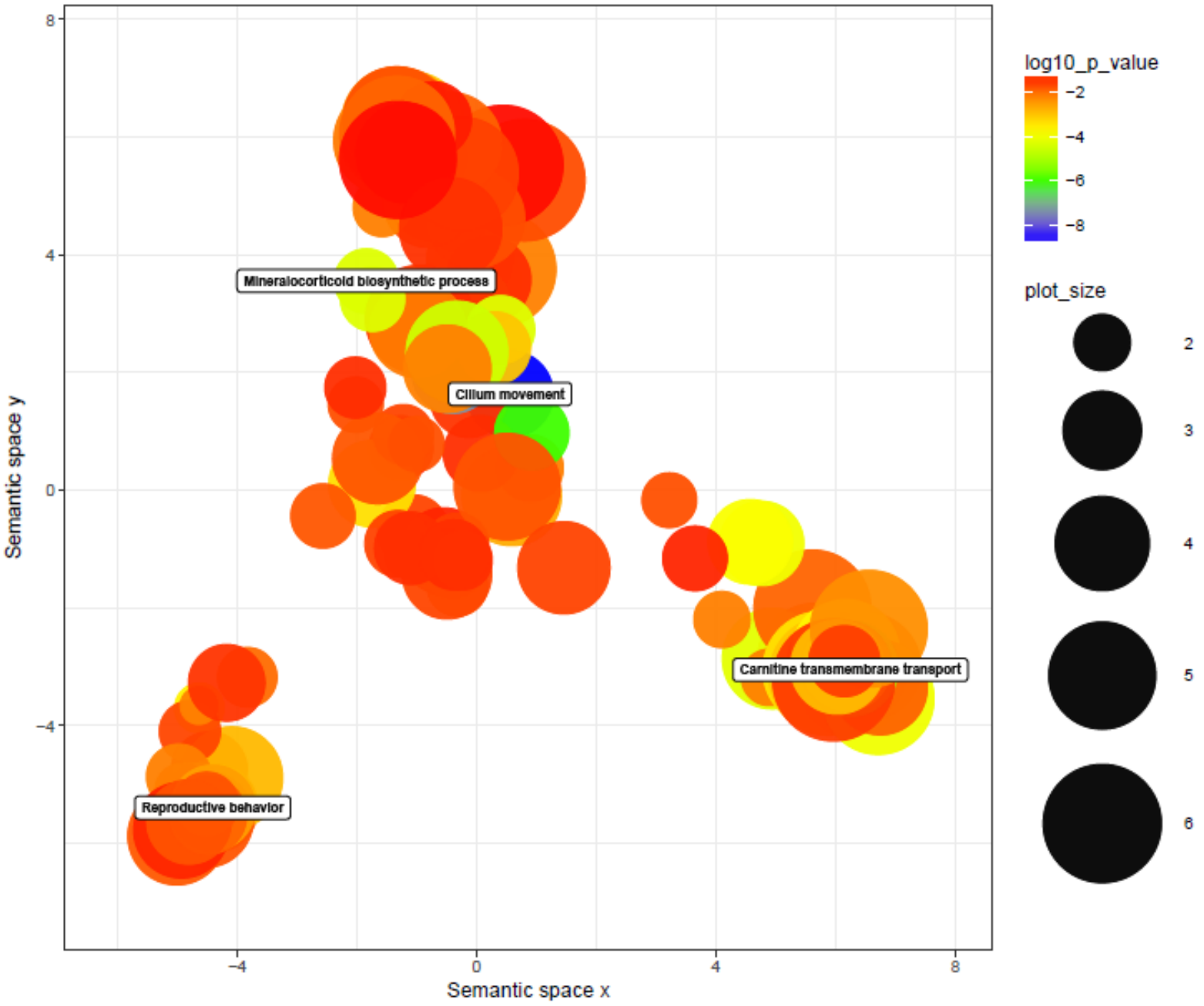

Figure 3.16: Gene ontology scatterplot generated with REVIGO for the head kidney tissue.

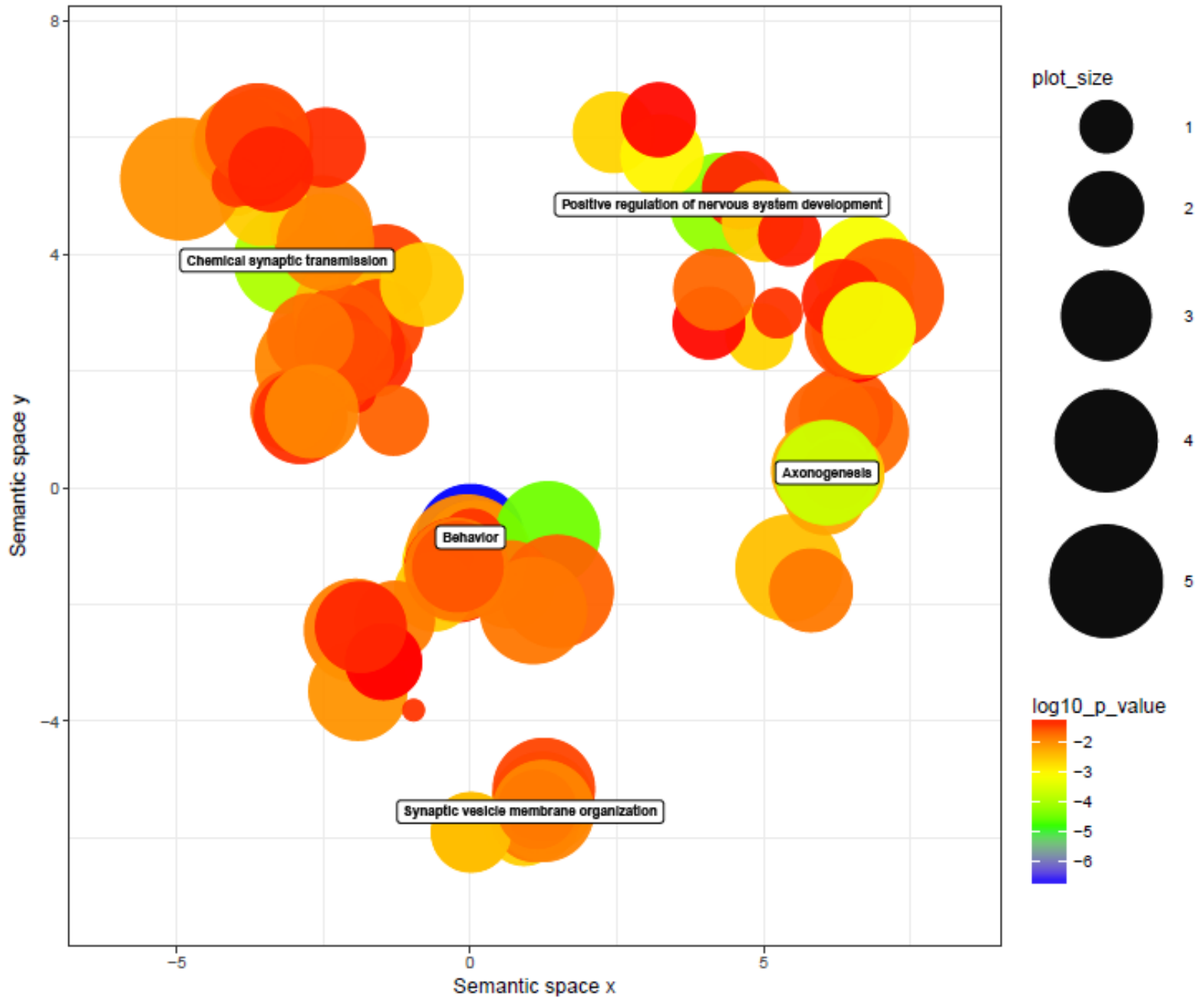

Figure 3.17: Gene ontology scatterplot generated with REVIGO for the brain plus pituitary tissue.

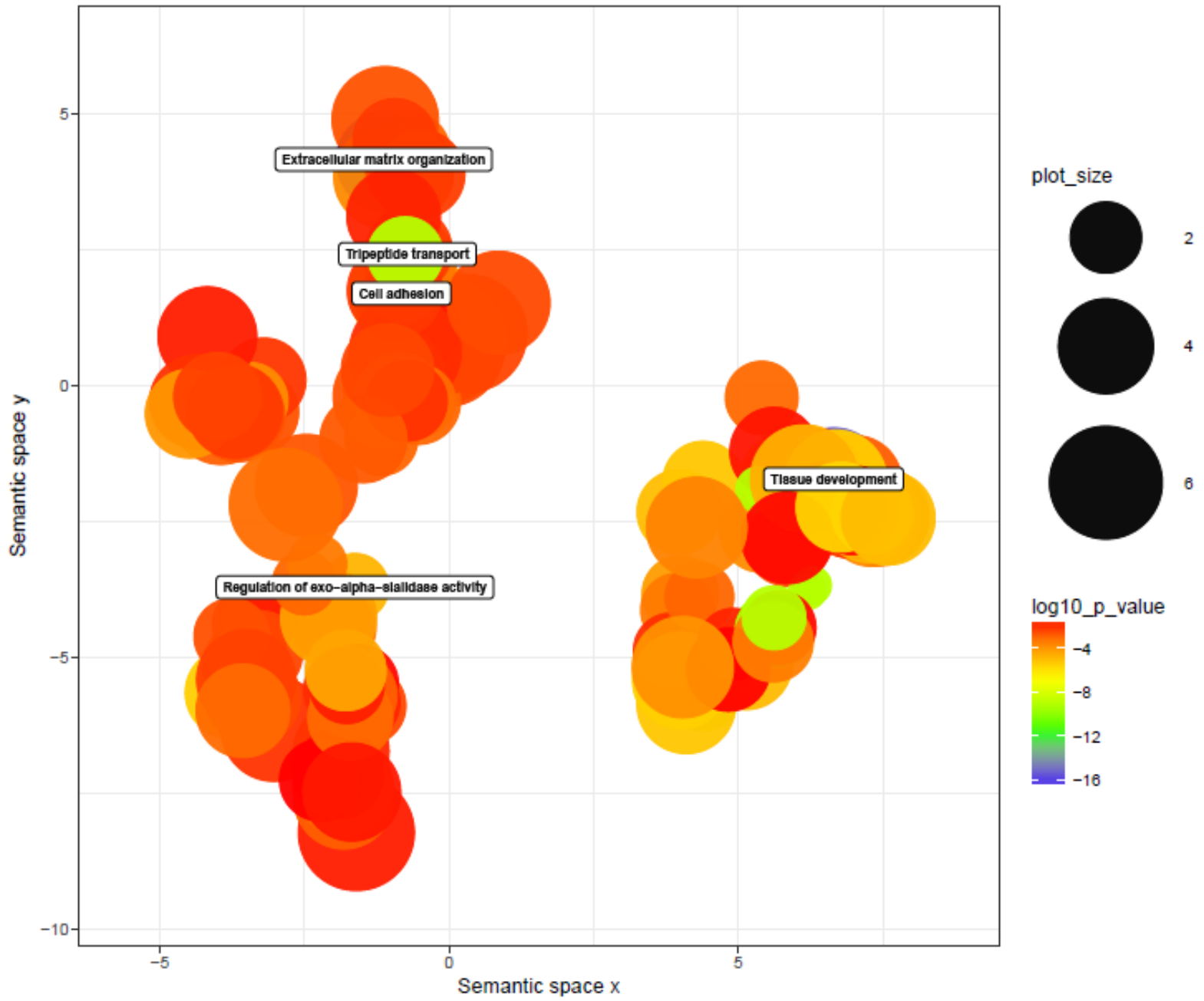

Figure 3.18: Gene ontology scatterplot generated with REVIGO for the caudal fin tissue.

Table 3.12: Top 10 gene ontologies according to their dispensability thoughout the studied tissues with their respective GO identifications, descpription and log10 p-values.

Term ID	Description	Log10 *p*-value
	Gill + Branchial Arch	
GO:0001906	Cell killing	-1.5993
GO:0002376	Immune system process	-2.6007
GO:0006821	Chloride transport	-1.6240
GO:0016338	Calcium-independent cell-cell adhesion via plasma membrane cell-adhesion molecules	-16.0350
GO:0022610	Biological adhesion	-8.5336
GO:0032501	Multicellular organismal process	-2.8593
GO:0032502	Developmental process	-1.7920
GO:0040011	Locomotion	-1.7932
GO:0042703	Menstruation	-3.1856
GO:0044699	Single-organism process	-1.5927
	Liver	
GO:0006629	Lipid metabolic process	-13.8762
GO:0008150	Biological process	-3.1101
GO:0008152	Metabolic process	-2.6076
GO:0019072	Viral genome packaging	-5.6109
GO:0032501	Multicellular organismal process	-2.6022
GO:0044699	Single-organism process	-4.8963
GO:0051704	Multi-organism process	-2.6678
GO:0009056	Catabolic process	-2.1195
GO:0006790	Sulphur compound metabolic process	-4.7816
GO:1902224	Ketone body metabolic process	-1.6929
	Spleen	
GO:0002376	Immune system process	-24.0259
GO:0002682	Regulation of immune system process	-28.1425
GO:0006898	Receptor-mediated endocytosis	-21.0645
GO:0006928	Movement of cell or subcellular component	-9.4835
GO:0008150	Biological process	-5.5223

GO:0008152	Metabolic process	-3.0439
GO:0008897	Cellular process	-3.9831
GO:0040011	Locomotion	-13.2869
GO:0044699	Single-organism process	-4.5660
GO:0050896	Response to stimulus	-17.3886
	Gonad	
GO:0008150	Biological process	-4.7417
GO:0008152	Metabolic process	-6.6040
GO:0009987	Cellular process	-7.3832
GO:0022402	Cell cycle process	-26.2354
GO:0031503	Protein complex localization	-2.8562
GO:0032259	Methylation	-3.7147
GO:0065007	Biological regulation	-2.1056
GO:0071840	Cellular component organization or biogenesis	-3.8303
GO:0006807	Nitrogen compound metabolic process	-10.2216
GO:0006281	DNA repair	-14.1016
	Midgut	
GO:0000003	Reproduction	-1.5974
GO:0002376	Immune system process	-1.7883
GO:0002474	Antigen processing and presentation of peptide antigen via MHC class I	-3.9646
GO:0006820	Anion transport	-11.0403
GO:0007040	Lysosome organization	-4.2468
GO:0007586	Digestion	-8.3621
GO:0008150	Biological process	-1.9607
GO:0009991	Response to extracellular stimulus	-3.2784
GO:0032501	Multicellular organismal process	-1.7423
GO:0042445	Hormone metabolic process	-4.5130
	White Muscle	
GO:0003012	Muscle system process	-22.1114
GO:0031032	Actomyosin structure organization	-20.6625
GO:0032501	Multicellular organismal process	-3.3406
GO:0032502	Developmental process	-7.1526

GO:0033058	Directional locomotion	-6.4609
GO:0035995	Detection of muscle stretch	-11.1153
GO:0036309	Protein localization to M-band	-3.2780
GO:0040007	Growth	-2.4182
GO:0044699	Single-organism process	-2.2163
GO:0071840	Cellular component organization or biogenesis	-2.6721
	Red Muscle	
GO:0006936	Muscle contraction	-30.1038
GO:0008150	Biological process	-2.6527
GO:0008152	Metabolic process	-1.9604
GO:0009987	Cellular process	-2.5790
GO:0032501	Multicellular organismal process	-4.2361
GO:0032502	Developmental process	-3.8335
GO:0043462	Regulation of ATPase activity	-7.2515
GO:0044699	Single-organism process	-8.8881
GO:0070252	Actin-mediated cell contraction	-26.7524
GO:0071840	Cellular component organization or biogenesis	-4.6402
	Kidney	
GO:0002376	Immune system process	-1.8025
GO:0006820	Anion transport	-6.5101
GO:0032501	Multicellular organismal process	-5.4164
GO:0032502	Developmental process	-1.9481
GO:0042435	Indole-containing compound biosynthetic process	-2.0654
GO:0042756	Drinking behaviour	-2.0829
GO:0044699	Single-organism process	-1.5311
GO:0044707	Single-multicellular organism process	-4.5893
GO:0048856	Anatomical structure development	-2.2398
GO:0050829	Defense response do Gram-negative bacterium	-6.4032
	Head Kidney	
GO:0003341	Cilium movement	-8.5993
GO:0035902	Response to immobilization stress	-1.6007
GO:0006577	Amino-acid betaine metabolic process	-3.6240
GO:0019614	Catechol-containing compound catabolic process	-2.0350

GO:1902603	Carnitine transmembrane transport	-4.5336
GO:0097164	Ammonium ion metabolic process	-2.8593
GO:0006705	Mineralocorticoid biosynthetic process	-4.7920
GO:0060012	Synaptic transmission, glycinergic	-2.7932
GO:0007028	Cytoplasm organization	-1.1856
GO:0006629	Lipid metabolic process	-2.5927
	Brain + Pituitary	
GO:0002213	Defence response to insect	-1.3940
GO:0007155	Cell adhesion	-2.7930
GO:0007610	Behaviour	-6.6852
GO:0022610	Biological adhesion	-2.7119
GO:0023052	Signalling	-2.0037
GO:0044708	Single-organism behaviour	-4.5036
GO:0046189	Phenol-containing compound biosynthetic process	-2.1186
GO:0006595	Polyamine metabolic process	-1.3748
GO:0048499	Synaptic vesicle membrane organization	-1.0268
GO:0051962	Positive regulation of nervous system development	-4.2251
	Caudal Fin	
GO:0007155	Cell adhesion	-8.0661
GO:0008150	Biological process	-2.3034
GO:0009888	Tissue development	-15.3856
GO:0022610	Biological adhesion	-7.8262
GO:0023052	Signalling	-2.9961
GO:0030198	Extracellular matrix organization	-16.7599
GO:0032501	Multicellular organismal process	-3.8466
GO:0032502	Developmental process	-12.3429
GO:0042107	Cytokine metabolic process	-3.8347
GO:0042939	Tripeptide transport	-8.9773

REVIGO generated scatterplots with enriched gene ontologies over the other tissues and those represented GOs seem appropriated for each tissue, alongside the tables with the 10 GOs with the least dispensability for biological process (Figures 3.8 to 3.18 and Table 3.12) and for molecular function (Figures and Table on appendices 6.3). The function of each tissue can be

confirmed with the GOs seen in the REVIGO results as we see biological process GOs related to muscle contractions and actomyosin structure organization on the white and red muscles, immune system processes on the spleen and, previously seen with the MFAP4 gene, calcium-independent cell-cell adhesion via plasma membrane cell-adhesion molecules on the gill plus branchial arch.

4 Conclusions

In this project, the European sardine transcriptome was assembled and annotated for the first time to be used as a cornerstone for following studies due to a concern for the conservation of populations given its economic and ecosystem importance and the lack of genomic information available.

The same female sardine was used for the genome assembly on a parallel project, along with the transcriptome assembly results from this study as it helps combining contigs. Even though the assembly of the transcriptome covered only 11 tissues of the sardine it now can be used as reference on future more specific studies, alongside the assemblies of the specific tissues and the rest of the results granted from this study.

High throughput Illumina sequencing of 11 tissues was edited for quality control with Trim Galore over Trimmomatic, due to unpredicted smaller reads regardless of the quality from Trimmomatic, and assembled via *de novo* over genome guided, due to smaller N50 values from genome guided assemblies in both different alignment methods, with Trinity to generate a high-quality draft of the sardine transcriptome.

Transrate didn't work as expected but still granted more realistic number of contigs for each assembly and seemed to help the filtration on the assembly of a mixture of reads from various tissues better then on assemblies of specific tissues when comparing the percentage of annotated genes. Annotation of the assemblies with Trinotate yielded results that corresponded with each tissue, as genes from each tissue show some part of its function. The number of contigs and tissue-specific genes and the tissue enriched GOs represented in general the tissue the assembly and annotation represented. Analysis on the tissue expression profile predicted tissue-specific genes with some duplicated genes, confirming a whole genome duplication event on teleost fishes when comparing to the number of orthologues other species sets have.

The data generated here may help conducting future studies ultimately figuring out the decline in sardine populations and its consequences and gene expression evolution.

5 Bibliography

[1] B. J. Haas and M. C. Zody, "Advancing RNA-Seq analysis," *Nat. Biotechnol.*, vol. 28, no. 5, pp. 421–423, 2010.

[2] M. G. Grabherr *et al.*, "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nat. Biotechnol.*, vol. 29, no. 7, pp. 644–652, 2011.

[3] R. Rosa, J. Vaz, R. Mota, and A. Silva, "Preference for Landings' Smoothing and Risk of Collapse in Optimal Fishery Policies: The Ibero-Atlantic Sardine Fishery," *Environ. Resour. Econ.*, pp. 1–21, 2017.

[4] R. Betancur-R. *et al.*, "The Tree of Life and a New Classification of Bony Fishes," *Tree Life*, pp. 1–54, 2013.

[5] P. Kafarski, "Rainbow code of biotechnology," *Chemik*, vol. 66, no. 8, pp. 814–816, 2012.

[6] R. A. Paselk, "Physical biochemistry, applications to biochemistry and molecular biology, second edition (Freifelder, David)," *J. Chem. Educ.*, vol. 60, no. 11, p. A321, 1983.

[7] F. Crick, "Central Dogma of Molecular Biology," vol. 227, no. 6, pp. 6–8, 1970.

[8] J. S. Reis-Filho, "Next-generation sequencing," *Breast Cancer Res.*, vol. 11, no. S3, p. S12, 2009.

[9] E. R. Mardis, "The impact of next-generation sequencing technology on genetics," *Trends Genet.*, vol. 24, no. 3, pp. 133–141, 2008.

[10] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nat. Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[11] V. Jongeneel *et al.*, "EXPRESSED SEQUENCE TAGS (ESTs)," *Bioinforma. A Pract. Guid. to Anal. Genes Proteins*, vol. 10, no. 1, pp. 57–63, 2001.

[12] P. Nemade and H. Kharche, "Big Data in Bioinformatics & the Era of Cloud Computing," *IOSR J. Comput. Eng.*, vol. 14, no. 2, pp. 2278–661, 2013.

[13] H. Kashyap, H. A. Ahmed, N. Hoque, S. Roy, and D. K. Bhattacharyya, "Big Data Analytics in Bioinformatics: A Machine Learning Perspective," vol. 13, no. 9, pp. 1–20, 2015.

[14] J. A. Martin and Z. Wang, "Next-generation transcriptome assembly," *Nat. Rev. Genet.*, vol. 12, no. 10, pp. 671–682, 2011.

[15] M. G. . Grabherr, N. Brian J. Haas, Moran Yassour Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W., and and A. R. Friedman, "Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data," *Nat. Biotechnol.*, vol. 29, no. 7, pp. 644–652, 2013.

[16] R. Smith-Unna, C. Boursnell, R. Patro, J. M. Hibberd, and S. Kelly, "TransRate: Reference-free quality assessment of *de novo* transcriptome assemblies," *Genome Res.*, vol. 26, no. 8, pp. 1134–1144, 2016.

[17] W. Abderrazik, A. Baali, Y. Schahrakane, and O. Tazi, "Study of reproduction of sardine, *Sardina pilchardus* in the north of Atlantic Moroccan area," *AACL Bioflux*, vol. 9, no. 3, pp. 507–517, 2016.

[18] H. O. Braga, U. M. Azeiteiro, H. M. F. Oliveira, and M. A. Pardal, "Evaluating fishermen's conservation attitudes and local ecological knowledge of the European sardine (*Sardina pilchardus*), Peniche, Portugal," *J. Ethnobiol. Ethnomed.*, vol. 13, no. 1, pp. 1–12, 2017.

[19] H. de O. Braga, M. Â. Pardal, and U. M. Azeiteiro, "Sharing fishers' ethnoecological knowledge of the European pilchard (*Sardina pilchardus*) in the westernmost fishing

community in Europe," *J. Ethnobiol. Ethnomed.*, vol. 13, no. 1, pp. 1–13, 2017.
[20] L. A. Jawad, "Biology and Ecology of Sardines and Anchovies," *J. Fish Biol.*, vol. 87, no. 4, pp. 1127–1128, 2015.
[21] B. Mustać and G. Sinovčić, "Reproduction, length-weight relationship and condition of sardine, *Sardina pilchardus* (Walbaum, 1792), in the eastern Middle Adriatic Sea (Croatia)," *Period. Biol.*, vol. 112, no. 2, pp. 133–138, 2010.
[22] G. Sinovčić, V. Č. Keč, and B. Zorica, "Population structure, size at maturity and condition of sardine, Sardina pilchardus (Walb., 1792), in the nursery ground of the eastern Adriatic Sea (Krka River Estuary, Croatia)," *Estuar. Coast. Shelf Sci.*, vol. 76, no. 4, pp. 739–744, 2008.
[23] A. M. V. Louro, B.; De Moro, G.; Garcia, C. M. E. V. R. ; Cox, C. ; Veríssimo, A.; Sabatino, S.; Santos, A.M.; Canário, "A haplotype-resolved draft genome of the European sardine (*Sardina pilchardus*)," *Gigascience*, 2018.
[24] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, "Revigo summarizes and visualizes long lists of gene ontology terms," *PLoS One*, vol. 6, no. 7, 2011.
[25] J. Inoue, Y. Sato, R. Sinclair, K. Tsukamoto, and M. Nishida, "Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling," *Proc. Natl. Acad. Sci.*, vol. 112, no. 48, pp. 14918–14923, 2015.
[26] T. Makino, A. McLysaght, and M. Kawata, "Genome-wide deserts for copy number variation in vertebrates," *Nat. Commun.*, vol. 4, pp. 1–10, 2013.
[27] M. S. Campbell, C. Holt, B. Moore, and M. Yandell, *Genome Annotation and Curation Using MAKER and MAKER-P*, vol. 2014, no. December. 2014.
[28] L. J. Magnoni, D. Crespo, A. Ibarz, J. Blasco, J. Fernández-Borràs, and J. V. Planas, "Effects of sustained swimming on the red and white muscle transcriptome of rainbow trout (*Oncorhynchus mykiss*) fed a carbohydrate-rich diet," *Comp. Biochem. Physiol. - A Mol. Integr. Physiol.*, vol. 166, no. 3, pp. 510–521, 2013.
[29] B. Glancy and R. S. Balaban, "Protein composition and function of red and white skeletal muscle mitochondria," no. 48, pp. 1280–1290, 2011.
[30] D. Niu *et al.*, "Microfibrillar-associated protein 4 (MFAP4) genes in catfish play a novel role in innate immune responses," *Dev. Comp. Immunol.*, vol. 35, no. 5, pp. 568–579, 2011.

5.1 Web page list

Trim Galore - www.bioinformatics.babraham.ac.uk/projects/trim_galore/
Trimmomatic - github.com/timflutre/Trimmomatic
Cutadapt - github.com/marcelm/cutadapt/
FastQC - www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trinity - github.com/trinityrnaseq/trinityrnaseq/wiki
Bowtie2 - bowtie-bio.sourceforge.net/bowtie2/index.shtml
Transrate - hibberdlab.com/transrate/
Trinotate - trinotate.github.io/
REVIGO - revigo.irb.hr/
FAO - www.fao.org/fishery/species/2910/en
NCBI-(clupeiformes)--
www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=32446&lvl=3&lin=f&keep=1&srchmode=1&unlock
ENA - www.ebi.ac.uk/ena

OrcAE - bioinformatics.psb.ugent.be/orcae/overview/Spil
Uniprot - www.uniprot.org/
Ensembl - www.ensembl.org/
Expression Atlas - www.ebi.ac.uk/gxa/home

6 Appendices

6.1 Code

Trim galore:

```
paste <(ls *R1_001.fastq.gz) <(ls *R2_001.fastq.gz) | while read args ; do
trim_galore -q 20 --phred33 --fastqc --clip_R1 1 --clip_R2 1 --stringency 3 -e 0.1 --gzip --length 30 -o ./qc --paired $args
done
```

Trimmomatic:

```
for f in $(ls *.fastq.gz | sed -e 's/1_001.fastq.gz//' -e 's/2_001.fastq.gz//' | sort -u)
do
java -jar trimmomatic.jar PE -threads 4 ${f}1_001.fastq.gz ${f}2_001.fastq.gz \
./fastq_edited/${f}1_paired.fastq.gz ./fastq_unpaired/${f}1_unpaired.fastq.gz \
./fastq_edited/${f}2_paired.fastq.gz ./fastq_unpaired/${f}2_unpaired.fastq.gz \
SLIDINGWINDOW:4:20 MINLEN:30 ILLUMINACLIP:adapter.fa:2:20:10 HEADCROP:1
fastqc ./fastq_edited/${f}1_paired.fastq.gz ./fastq_edited/${f}2_paired.fastq.gz -o ./QC_trimmo
done
```

Trinity *de novo*:

```
Trinity --seqType fq --max_memory 184G --samples_file tissue.table --SS_lib_type RF --CPU 24 --min_contig_length 200 \
--output /data/ccmar/sardinha/analyses/trinity_all --verbose --normalize_max_read_cov 50 --normalize_by_read_set
```

Trinity genome guided:

bowtie2 -q --rf --threads 14 -x spil_75h -1 17109R-01-42_S0_L001_R1_001_val_1.fq.gz -2 17109R-01-42_S0_L001_R2_001_val_2.fq.gz | samtools view -bS - > aligned_end_42.bam
bowtie2 -q --local --rf --threads 14 -x spil_75h -1 17109R-01-42_S0_L001_R1_001_val_1.fq.gz -2 17109R-01-42_S0_L001_R2_001_val_2.fq.gz | samtools view -bS - > aligned_local_42.bam
samtools sort -o aligned_end_42.bam -@ 14 aligned_end_sorted_42.bam
samtools sort -o aligned_local_sorted2_42.bam -@ 14 aligned_local_42.bam

Trinity --genome_guided_bam aligned_local_sorted_42.bam --genome_guided_max_intron 25000 --max_memory 114G --CPU 14 --output ./trinity_42gglocal
Trinity --genome_guided_bam aligned_end_sorted_42.bam --genome_guided_max_intron 10000 --max_memory 114G --CPU 14 --output ./trinity_42ggend

Transrate:
transrate --assembly Trinity.fasta \
--left 17109R-01-42_S0_L001_R1_001_val_1.fq \
--right 17109R-01-42_S0_L001_R2_001_val_2.fq \
--threads 8 --output transrate_42

Trinotate:
TransDecoder.LongOrfs -t ../../../trinity_42/transrate_42/good.Trinity.fasta --gene_trans_map ../../../trinity_42/Trinity.fasta.gene_trans_map
blastp -query good.Trinity.fasta.transdecoder_dir/longest_orfs.pep \
-db ../../blastdb/uniprot_sprot.pep -max_target_seqs 1 \
-outfmt 6 -evalue 1e-5 -num_threads 8 > blastp_42g.outfmt6
hmmscan --cpu 8 --domtblout pfam_42g.domtblout ../../blastdb/Pfam-A.hmm \
good.Trinity.fasta.transdecoder_dir/longest_orfs.pep
TransDecoder.Predict -t ../../../trinity_42/transrate_42/good.Trinity.fasta --retain_pfam_hits pfam_42g.domtblout --retain_blastp_hits blastp_42g.outfmt6
mv good.Trinity.fasta.transdecoder.pep transdecoder_42g.pep
blastp -query transdecoder_42g.pep -db ../../blastdb/uniprot_sprot.pep -num_threads 8 -max_target_seqs 1 -outfmt 6 > blastp_42g.outfmt6
hmmscan --cpu 8 --domtblout pfam_42g.domtblout ../../blastdb/Pfam-A.hmm transdecoder_42g.pep \

> ../pfam.log

signalp -f short -n signalp_42g.out transdecoder_42g.pep

tmhmm --short < transdecoder_42g.pep > tmhmm_42g.out

RnammerTranscriptome.pl --transcriptome ../../../trinity_42/Trinity.fasta --path_to_rnammer ~/bin/RNAMMER/rnammer --org_type euk

mv Trinity.fasta.rnammer.gff rnammer_42g.gff

Trinotate Trinotate_42g.sqlite init --gene_trans_map ../../../trinity_42/Trinity.fasta.gene_trans_map --transcript_fasta ../../../trinity_42/transrate_42/good.Trinity.fasta --transdecoder_pep transdecoder_42g.pep

Trinotate Trinotate_42g.sqlite LOAD_swissprot_blastp blastp_42g.outfmt6

Trinotate Trinotate_42g.sqlite LOAD_pfam pfam_42g.domtblout

Trinotate Trinotate_42g.sqlite LOAD_tmhmm tmhmm_42g.out

Trinotate Trinotate_42g.sqlite LOAD_signalp signalp_42g.out

Trinotate Trinotate_42g.sqlite LOAD_swissprot_blastx blastx_42g.outfmt6

Trinotate Trinotate_42g.sqlite LOAD_rnammer rnammer_42g.gff

Trinotate Trinotate_42g.sqlite report -e 1e-5 --pfam_cutoff DNC > trinotate_42g_annotation_report.xls

6.2 Fast QC Report from Trim Galore: Per base sequence quality

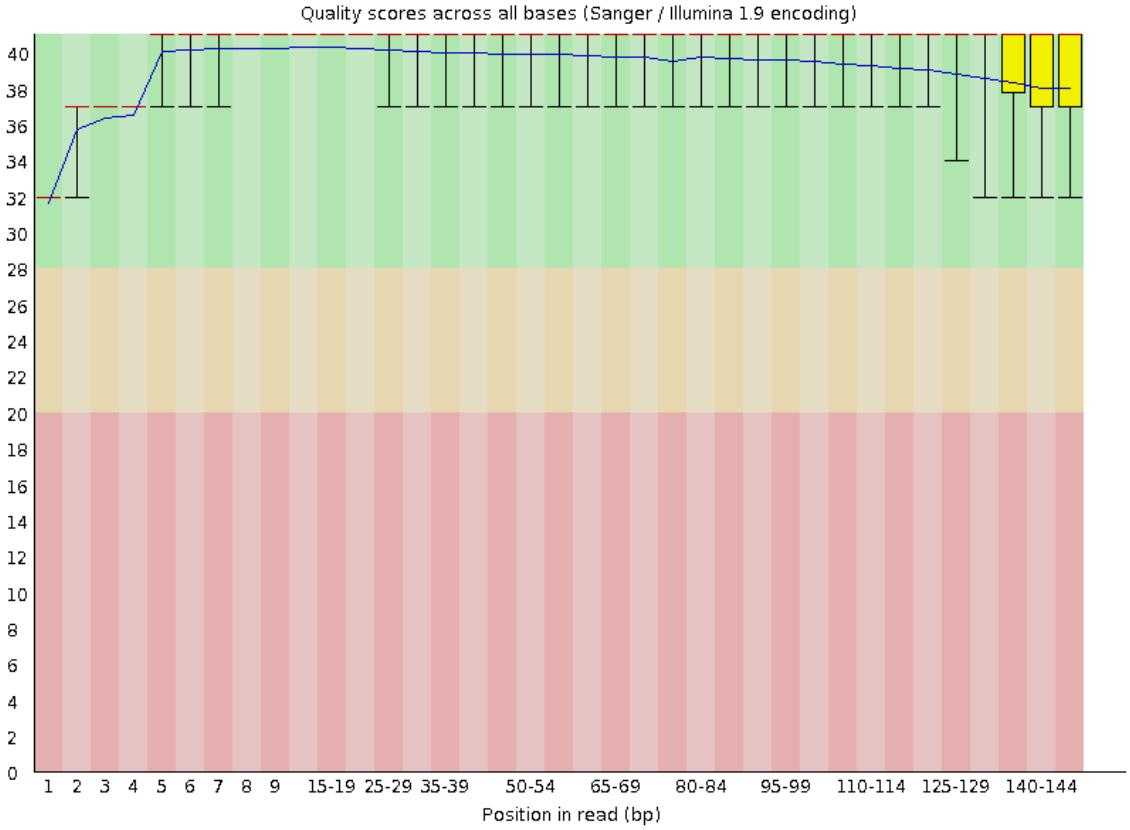

Figure 6.1: Per base sequence quality of gill plus branchial arch reads.

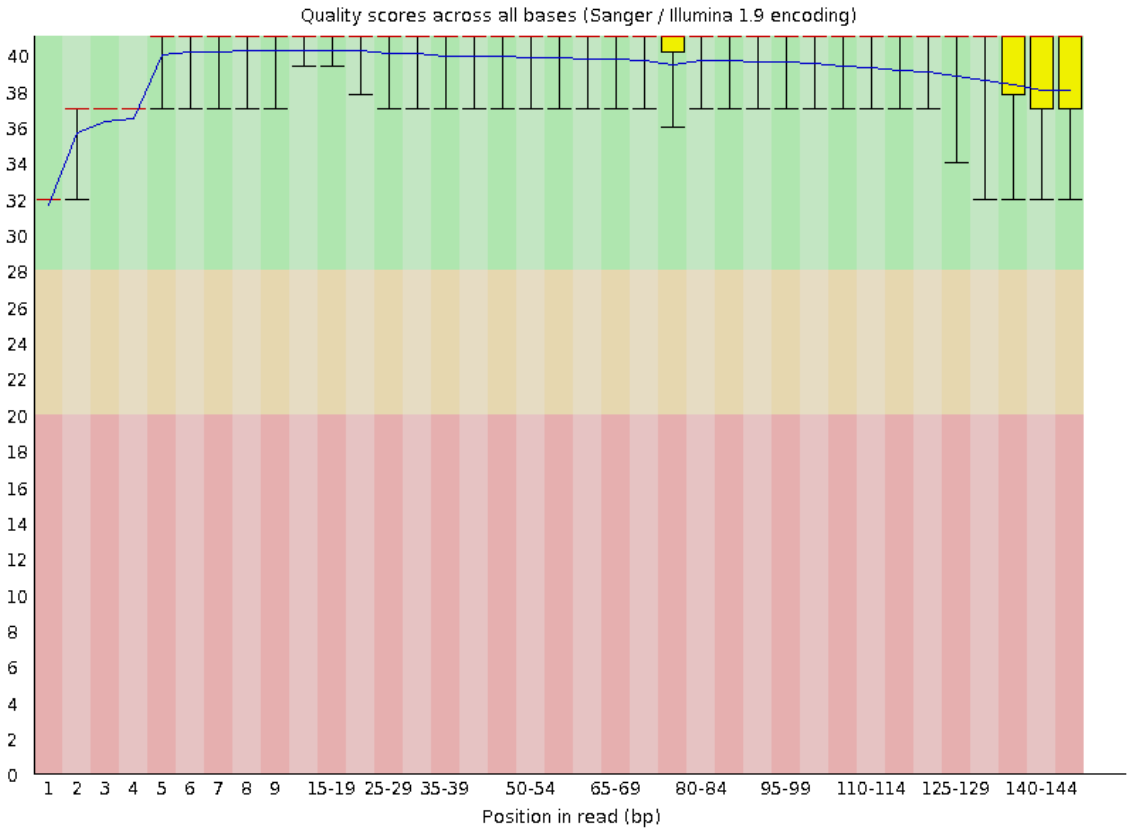

Figure 6.2: Per base sequence quality of liver reads.

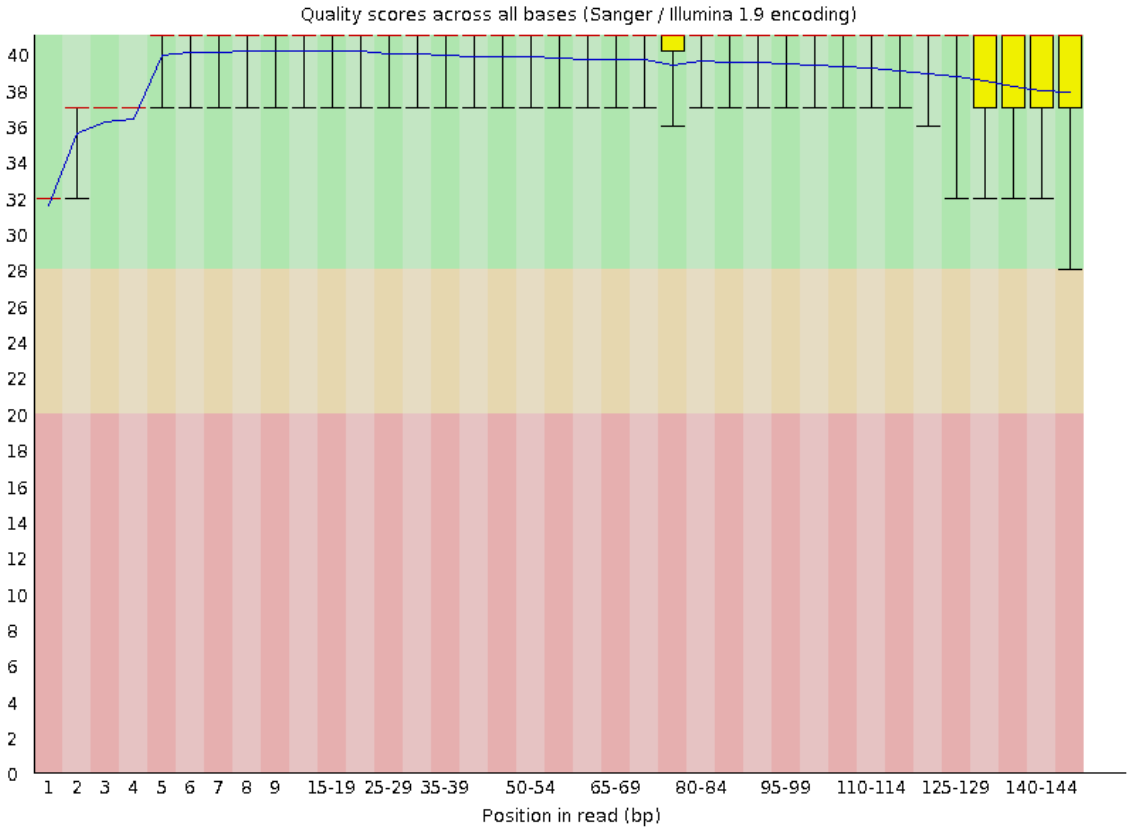

Figure 6.3: Per base sequence quality of spleen reads.

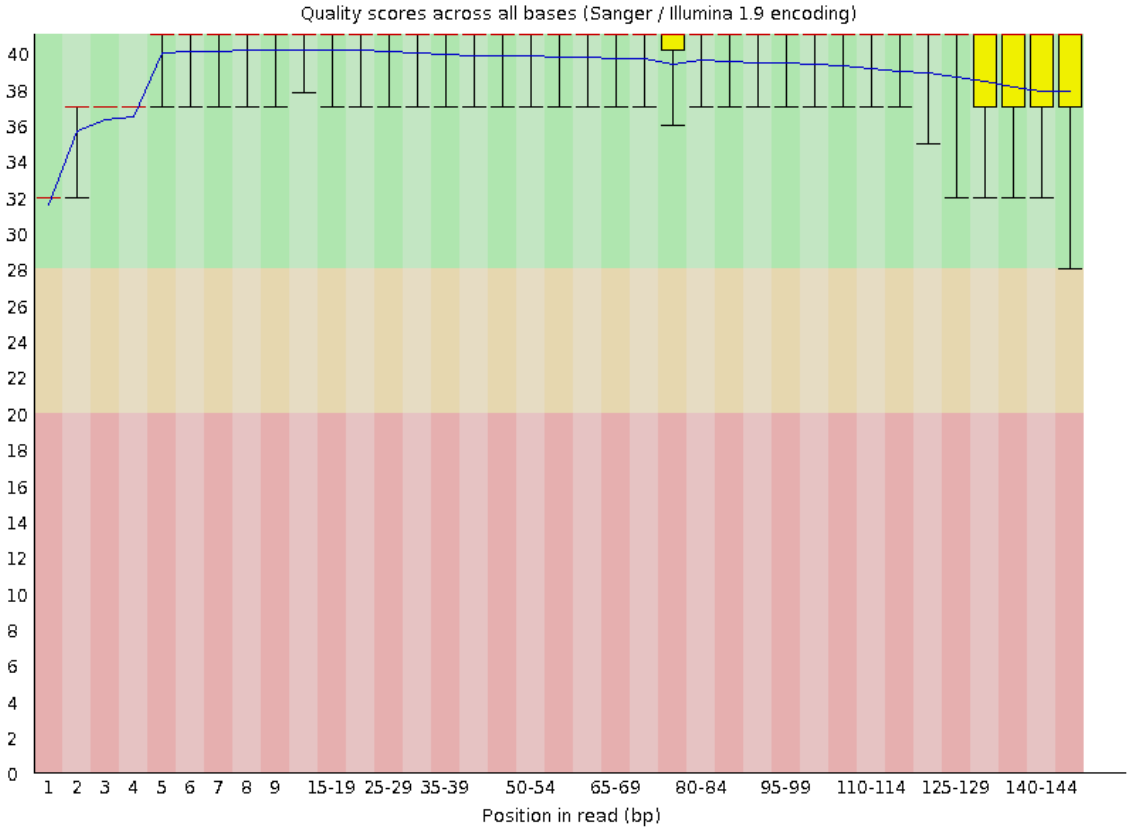

Figure 6.4: Per base sequence quality of gonad reads.

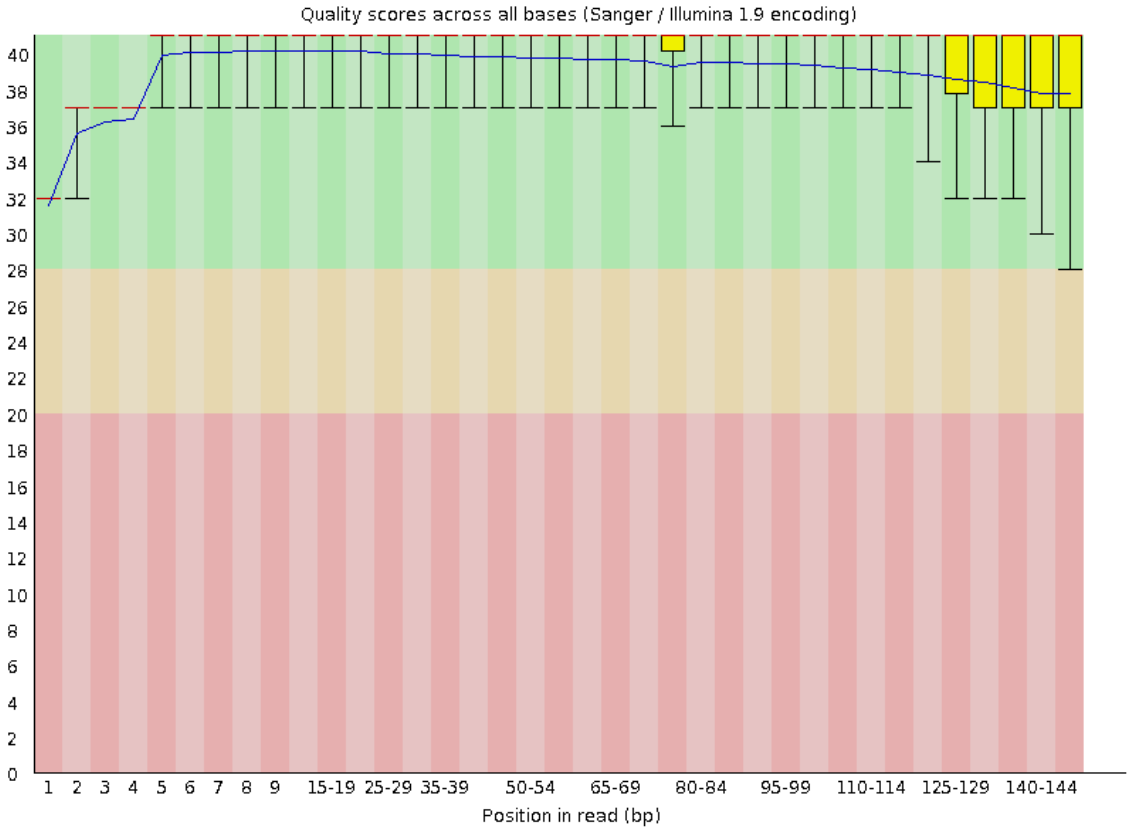

Figure 6.5: Per base sequence quality of midgut reads.

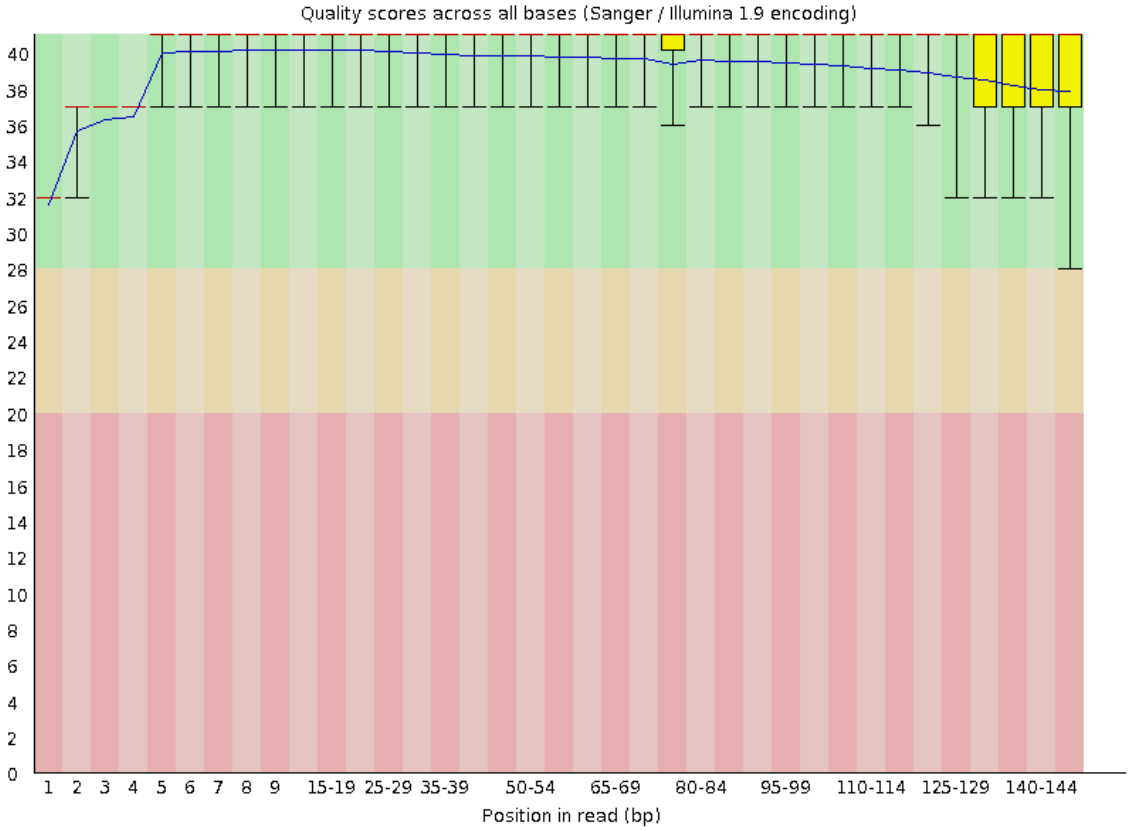

Figure 6.6: Per base sequence quality of white muscle reads.

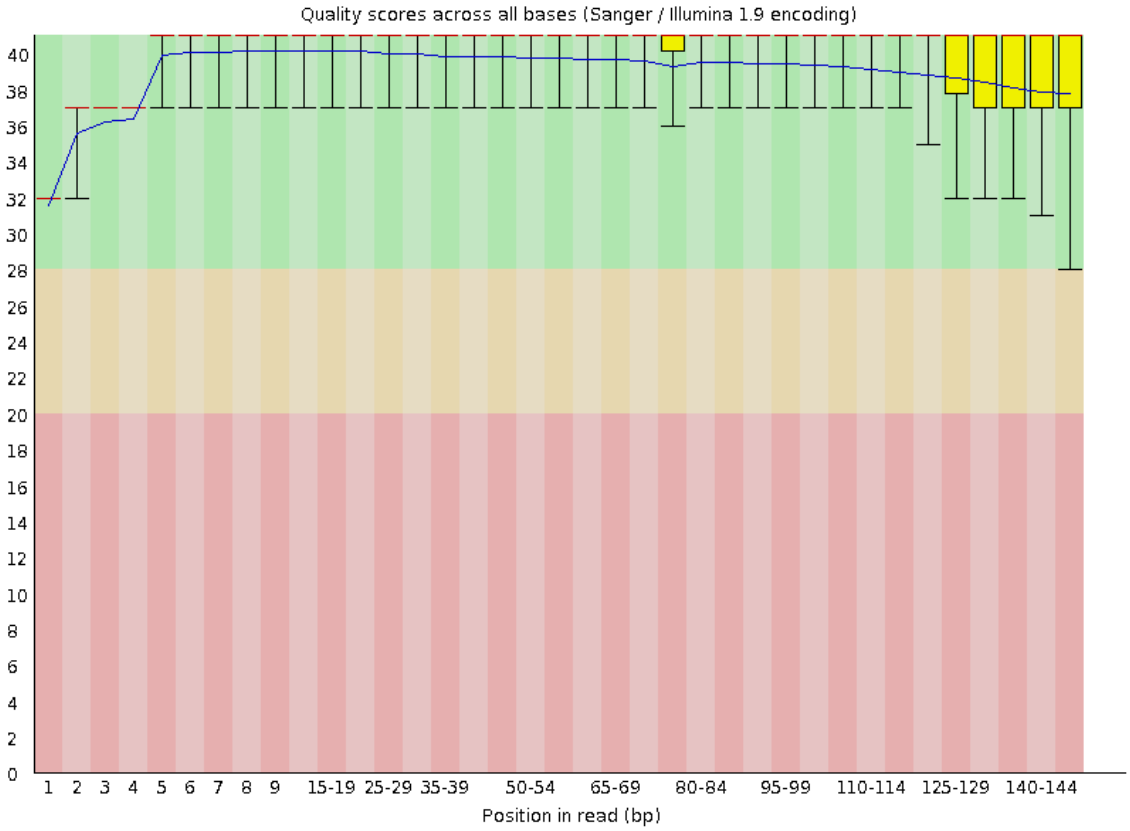

Figure 6.7: Per base sequence quality of red muscle reads.

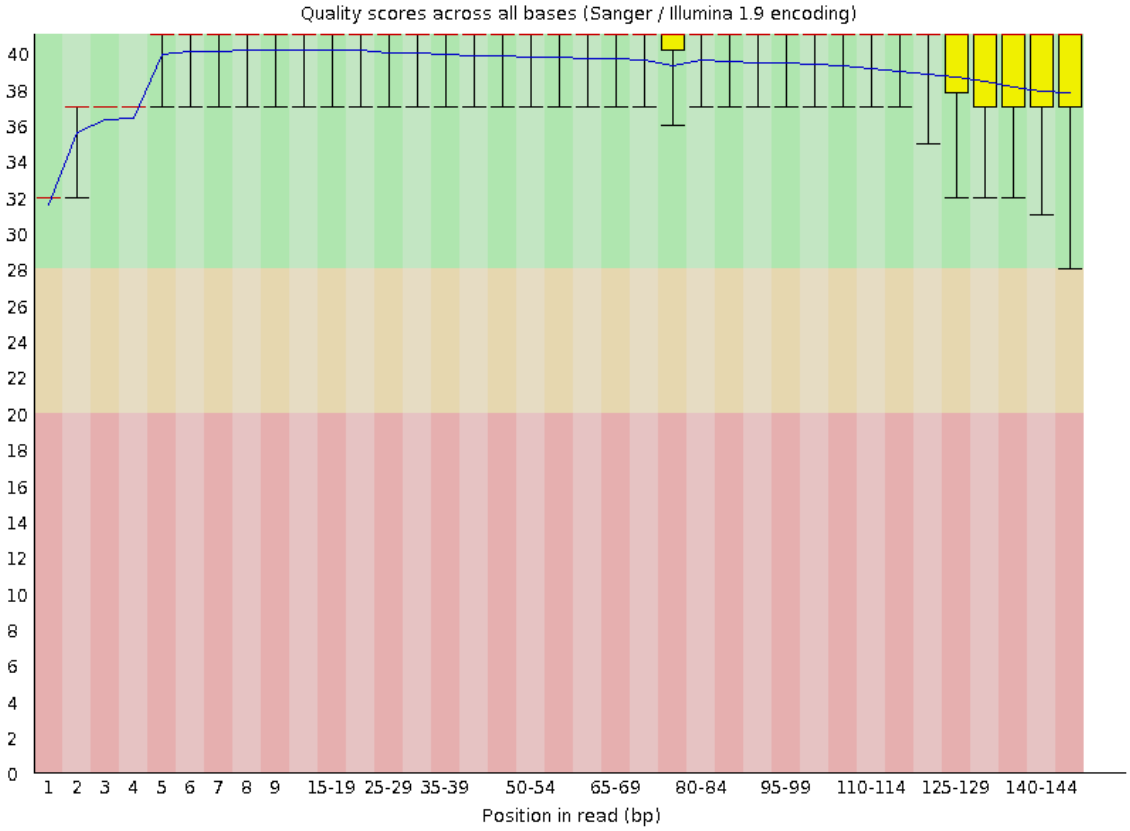

Figure 6.8: Per base sequence quality of kidney reads.

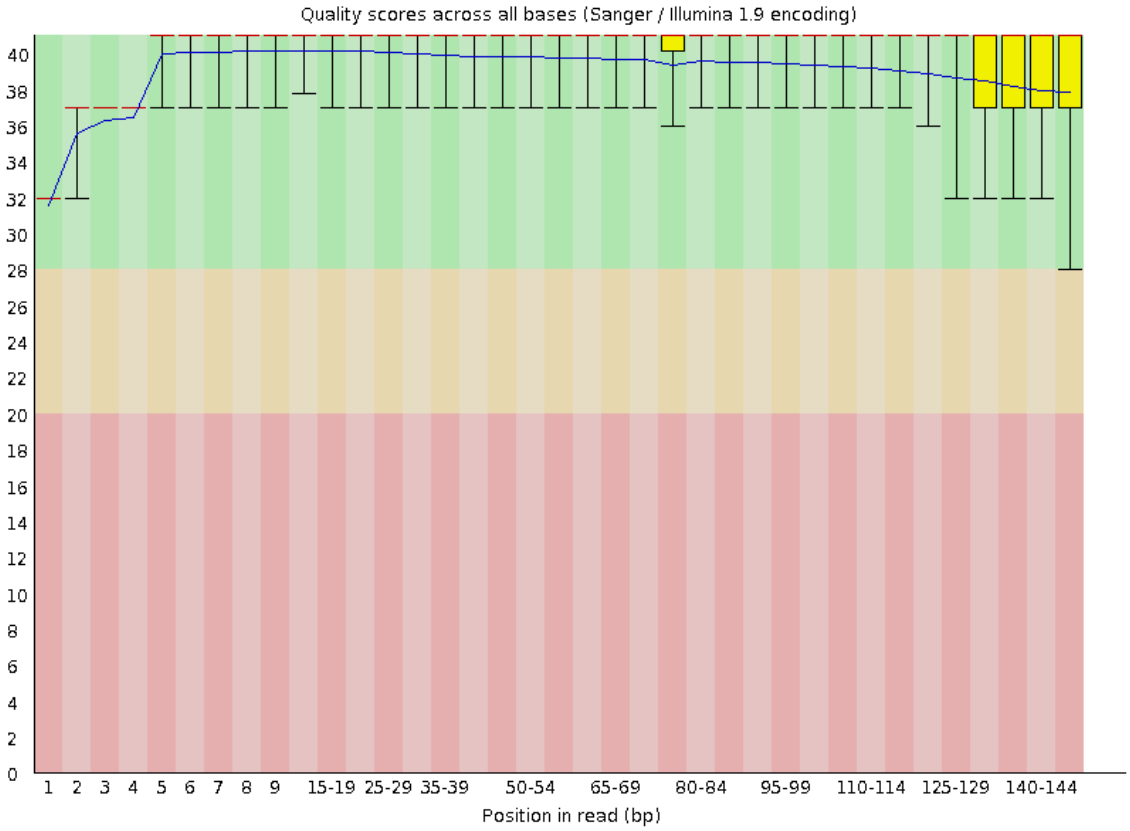

Figure 6.9: Per base sequence quality of head kidney reads.

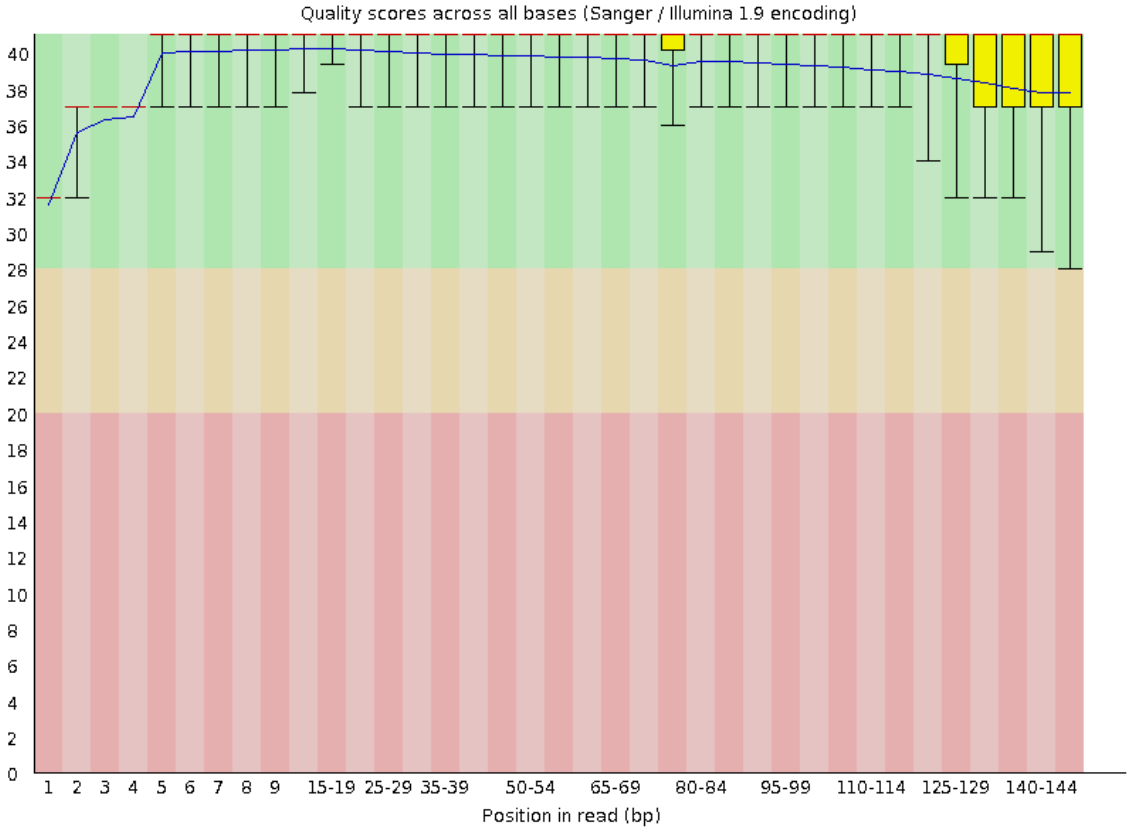

Figure 6.10: Per base sequence quality of brain plus pituitary reads.

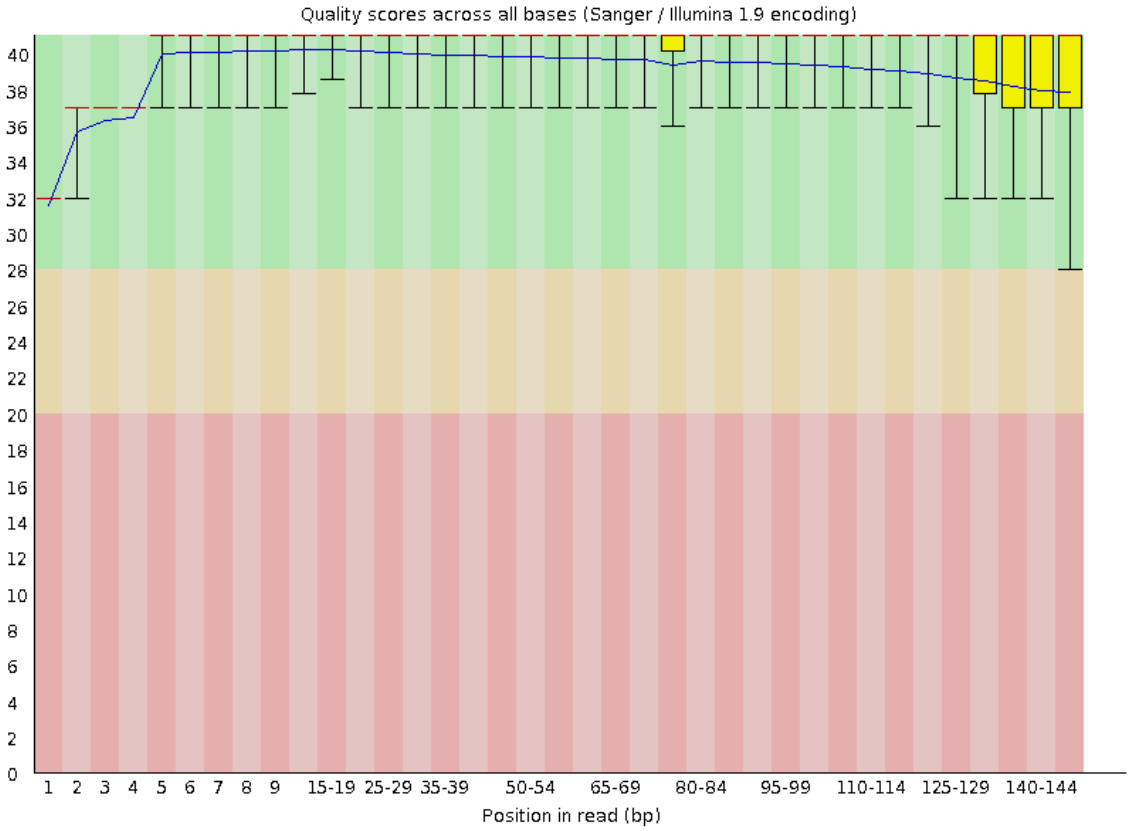

Figure 6.11: Per base sequence quality of caudal fin reads.

6.3 Molecular function REVIGO results

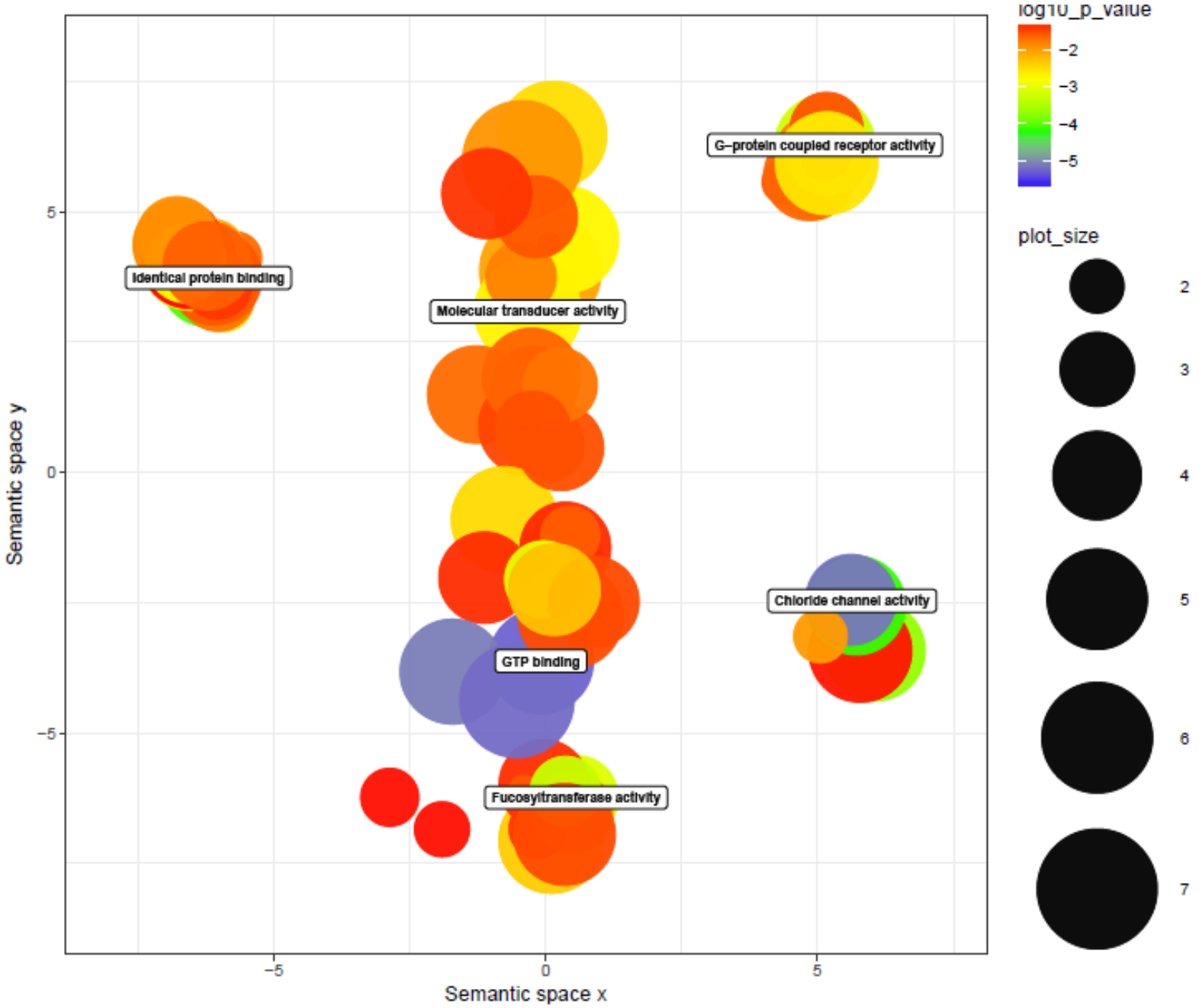

Figure 6.12: Gene ontology scatterplot generated with REVIGO for the gill plus branchial arch tissue.

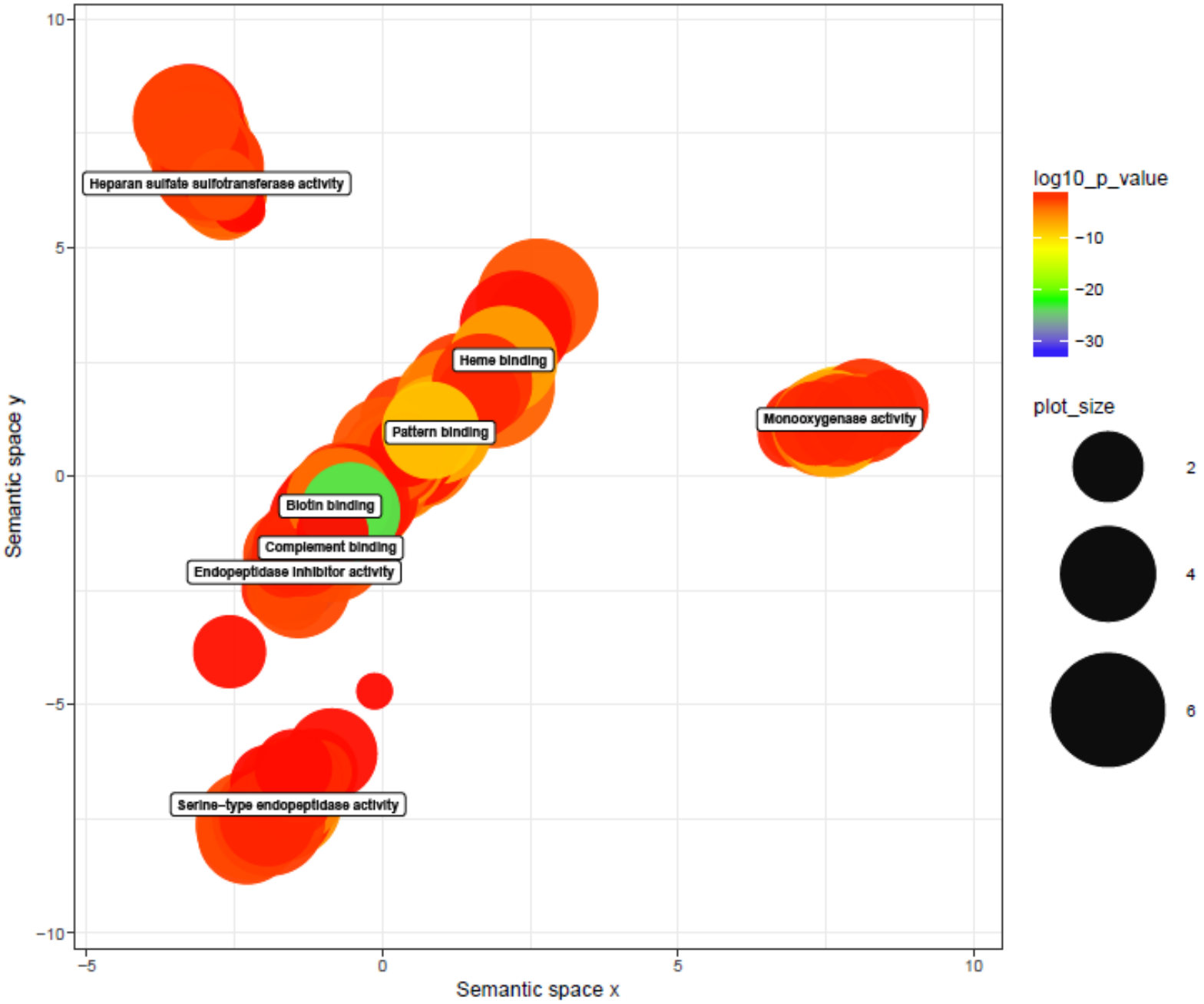

Figure 6.13: Gene ontology scatterplot generated with REVIGO for the liver tissue.

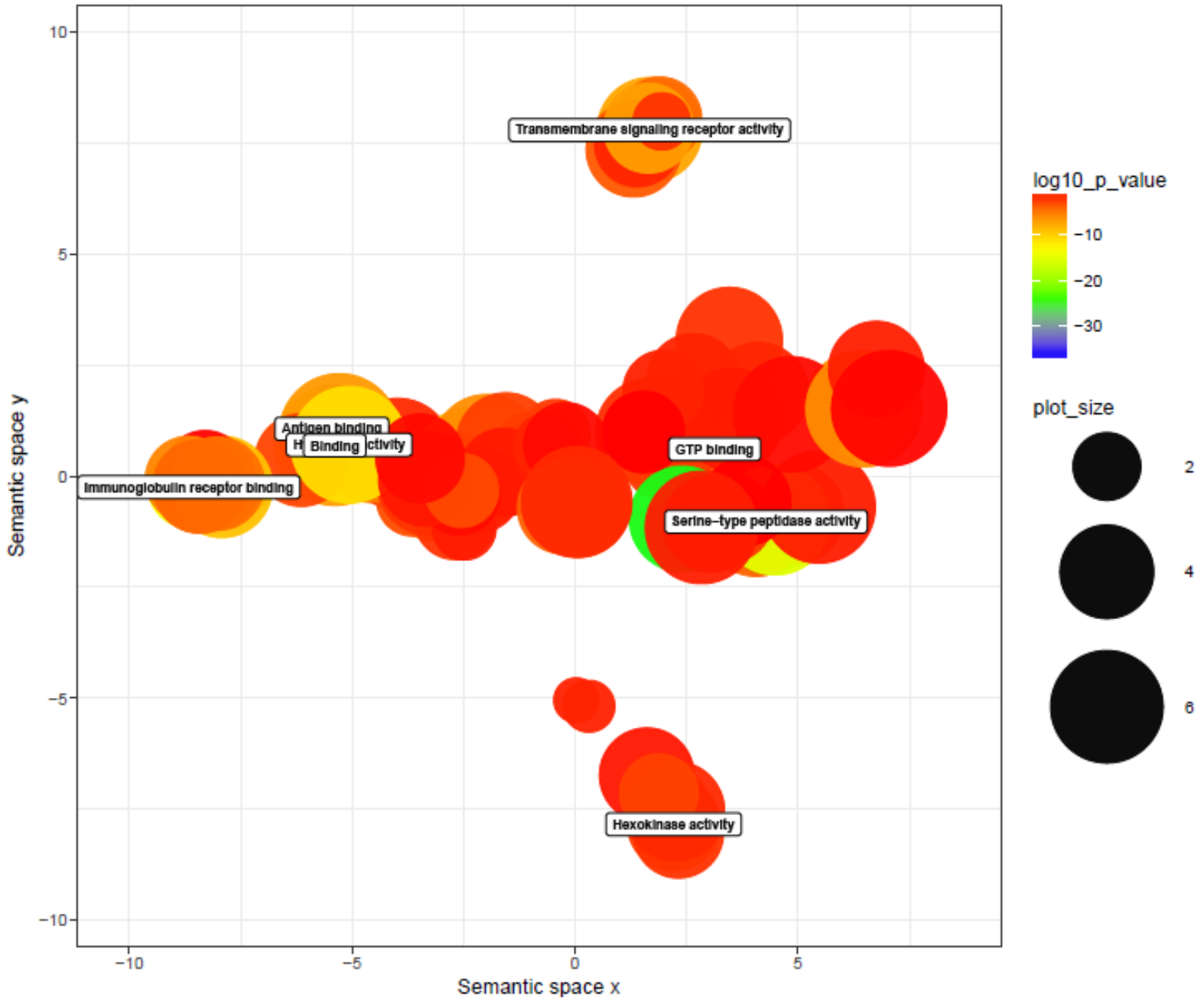

Figure 6.14: Gene ontology scatterplot generated with REVIGO for the spleen tissue.

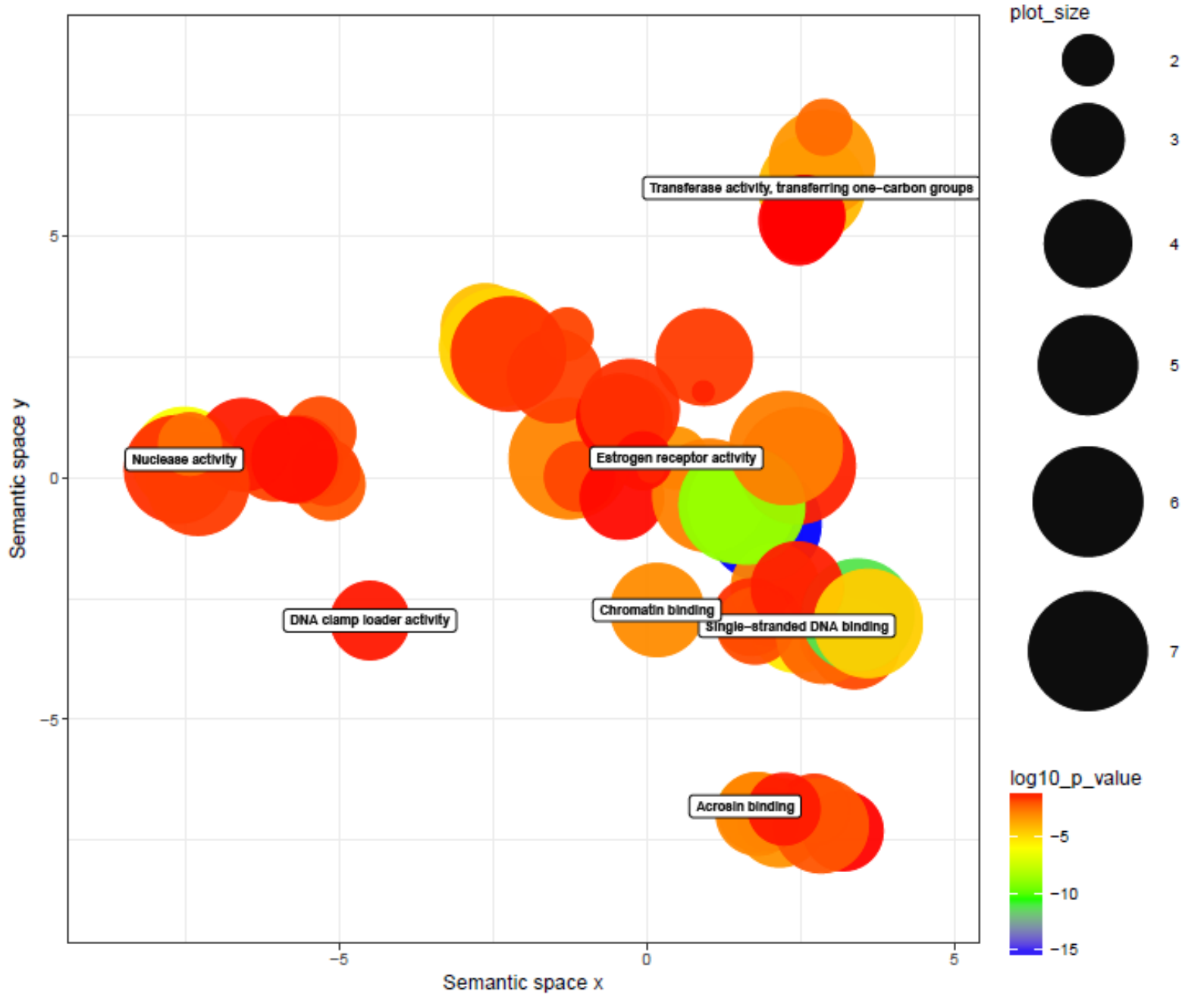

Figure 6.15: Gene ontology scatterplot generated with REVIGO for the gonad tissue.

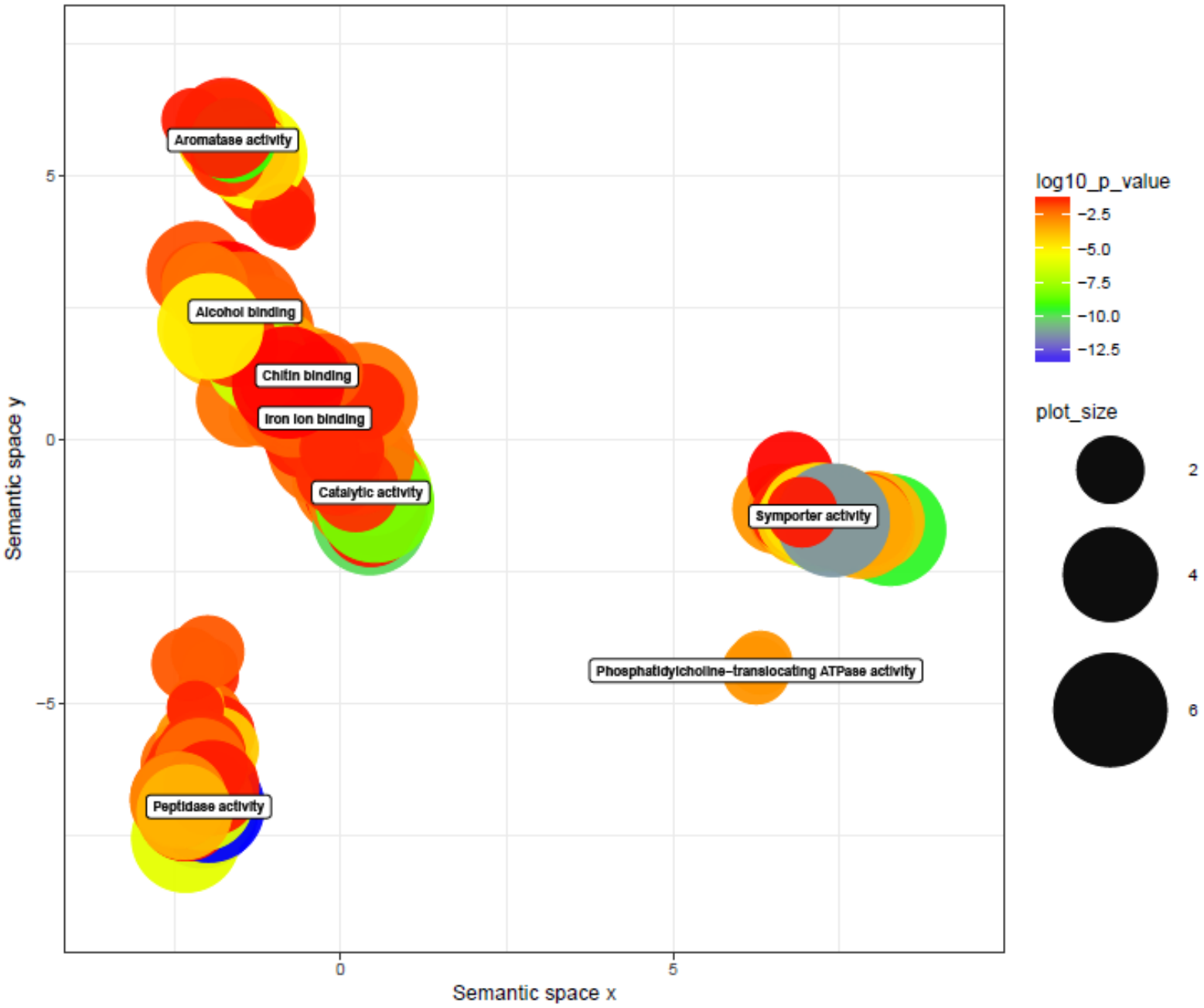

Figure 6.16: Gene ontology scatterplot generated with REVIGO for the midgut tissue.

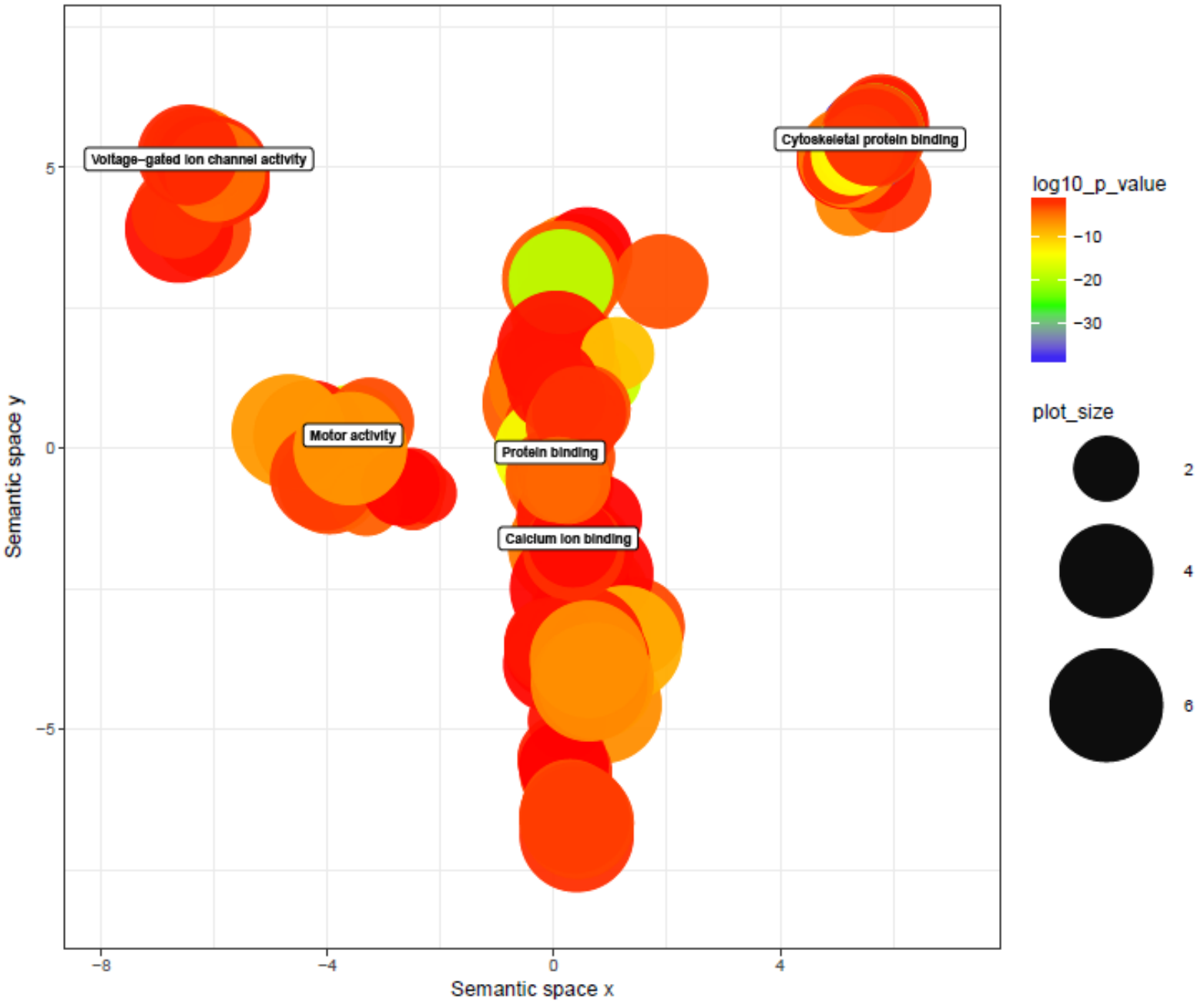

Figure 6.17: Gene ontology scatterplot generated with REVIGO for the white muscle tissue.

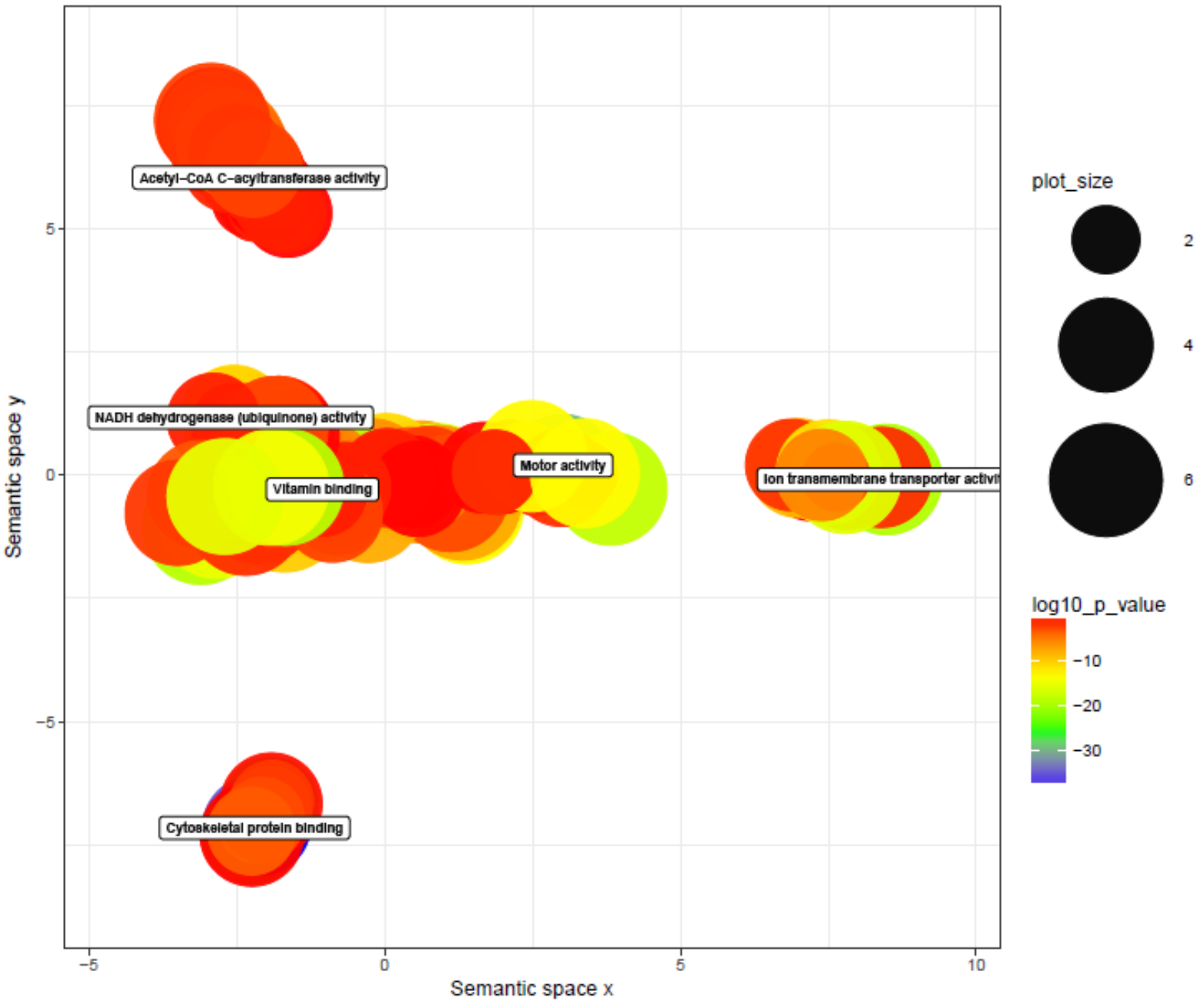

Figure 6.18: Gene ontology scatterplot generated with REVIGO for the red muscle tissue.

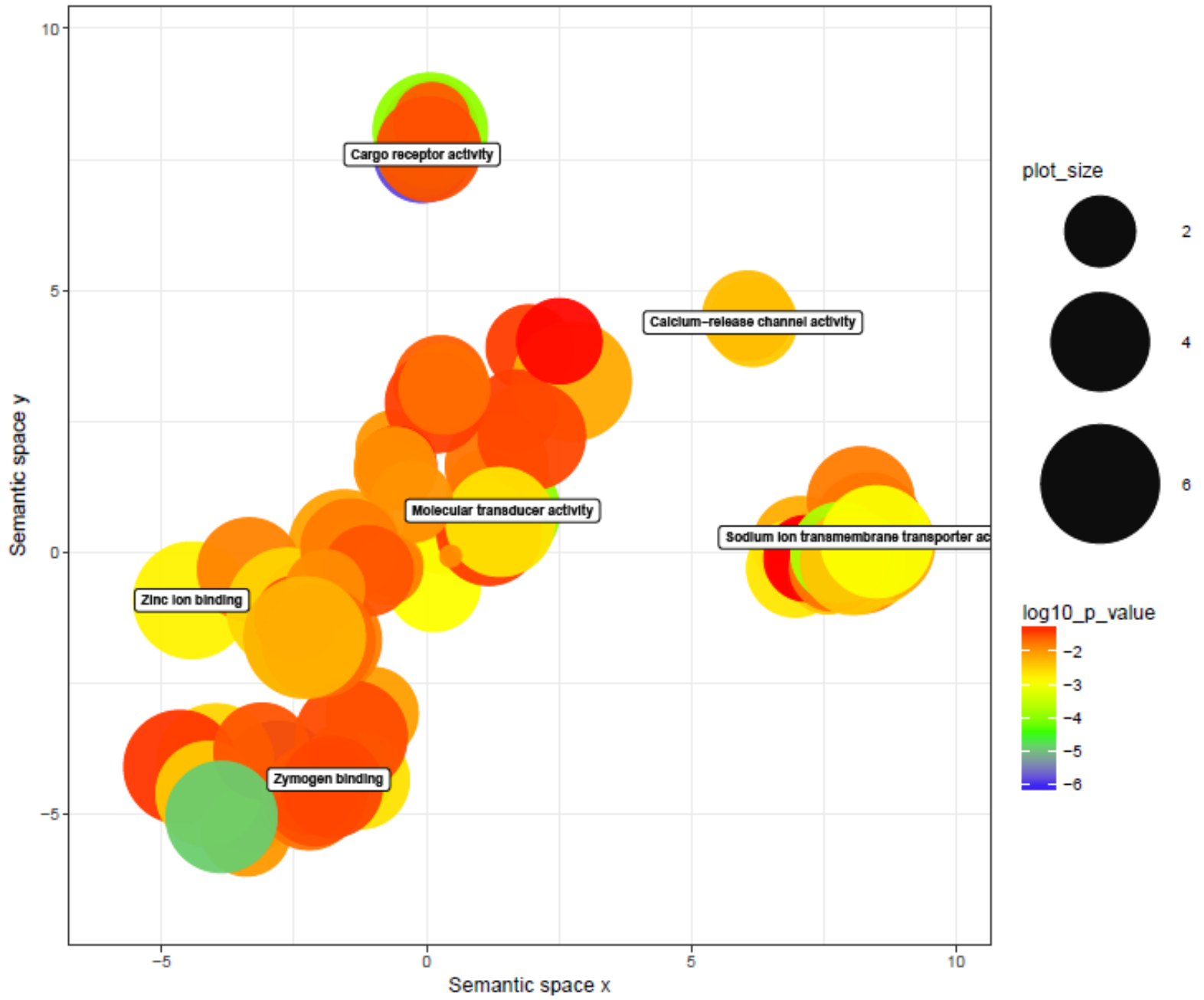

Figure 6.19: Gene ontology scatterplot generated with REVIGO for the kidney tissue.

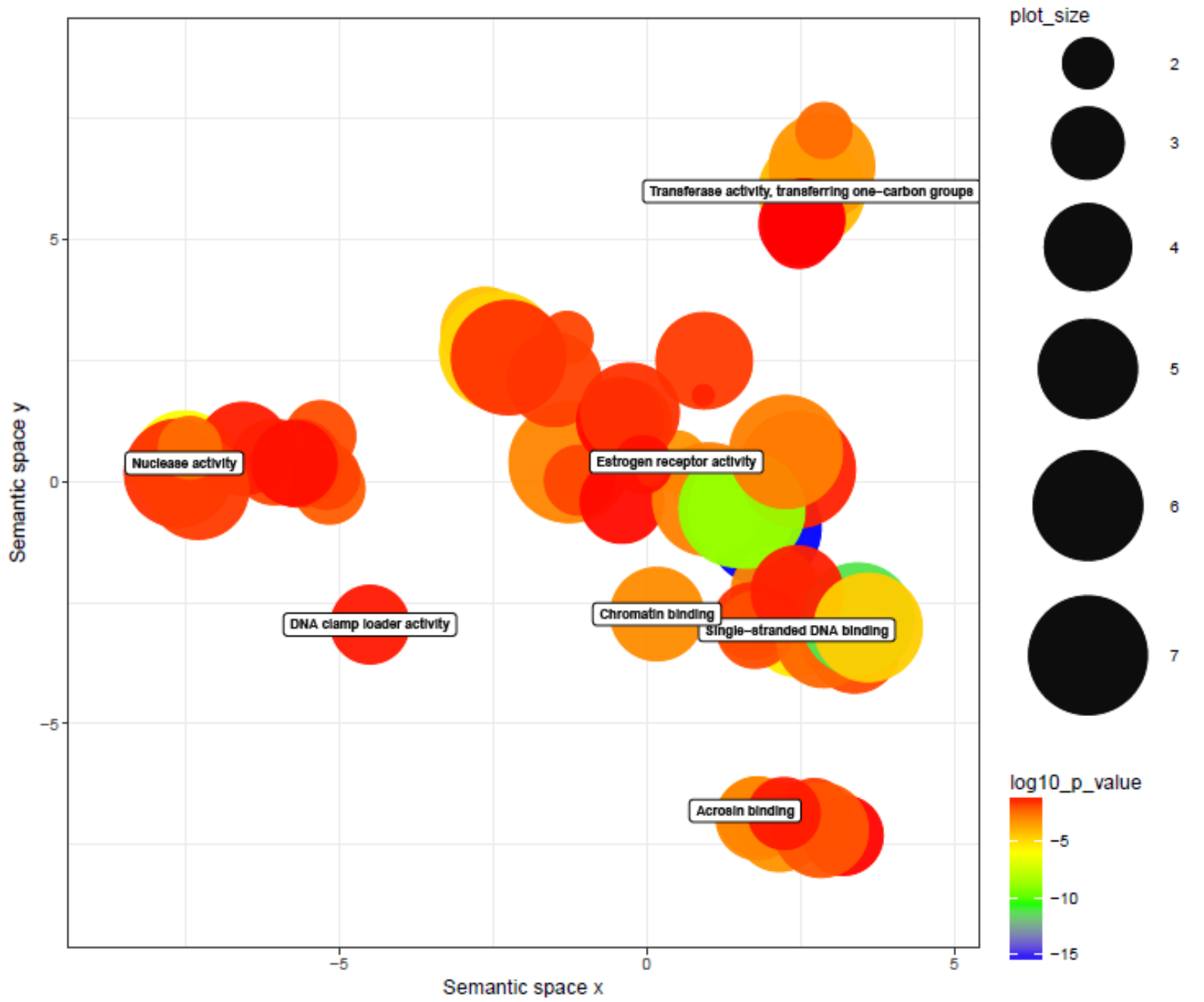

Figure 6.20: Gene ontology scatterplot generated with REVIGO for the head kidney tissue.

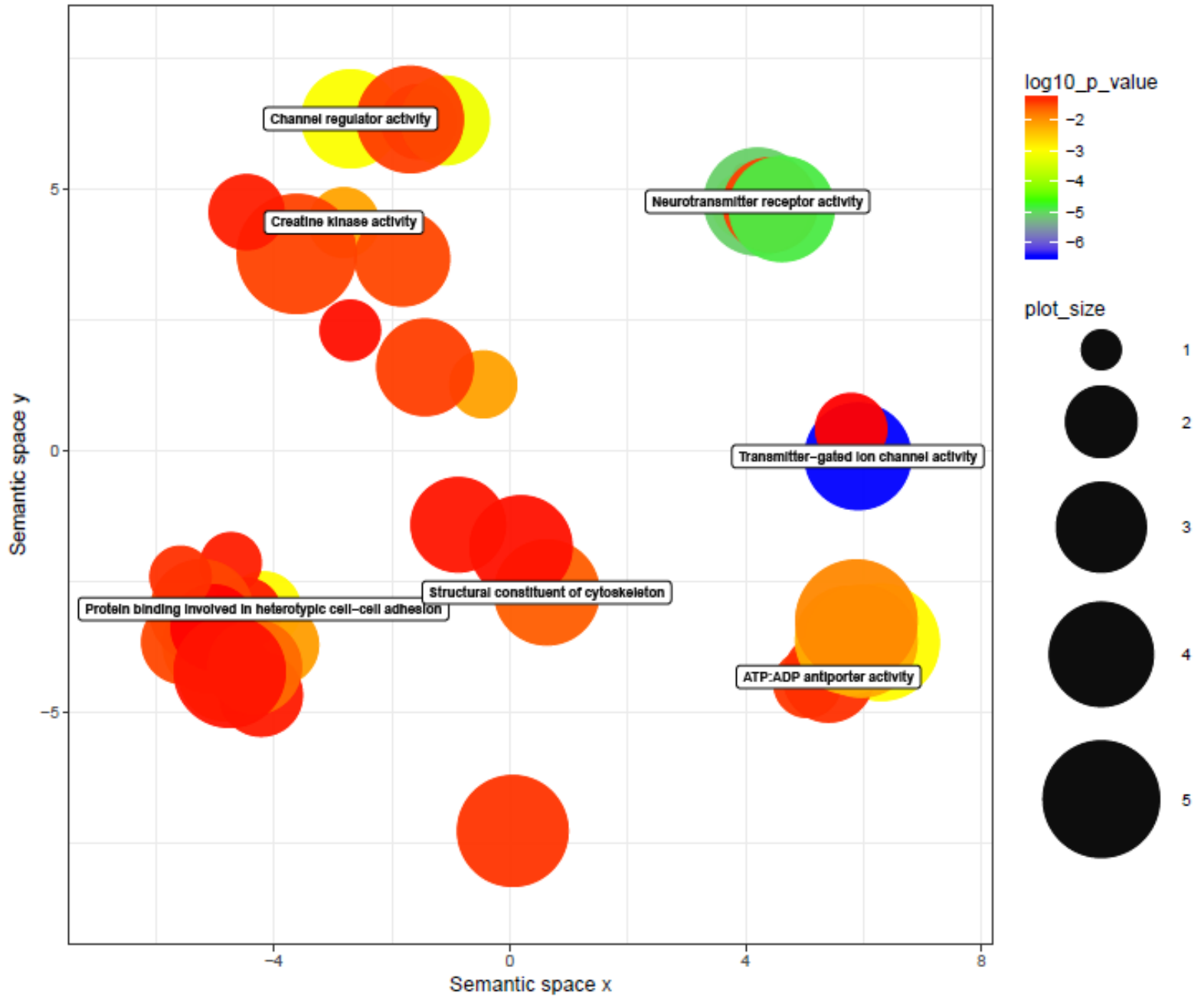

Figure 6.21: Gene ontology scatterplot generated with REVIGO for the brain plus pituitary tissue.

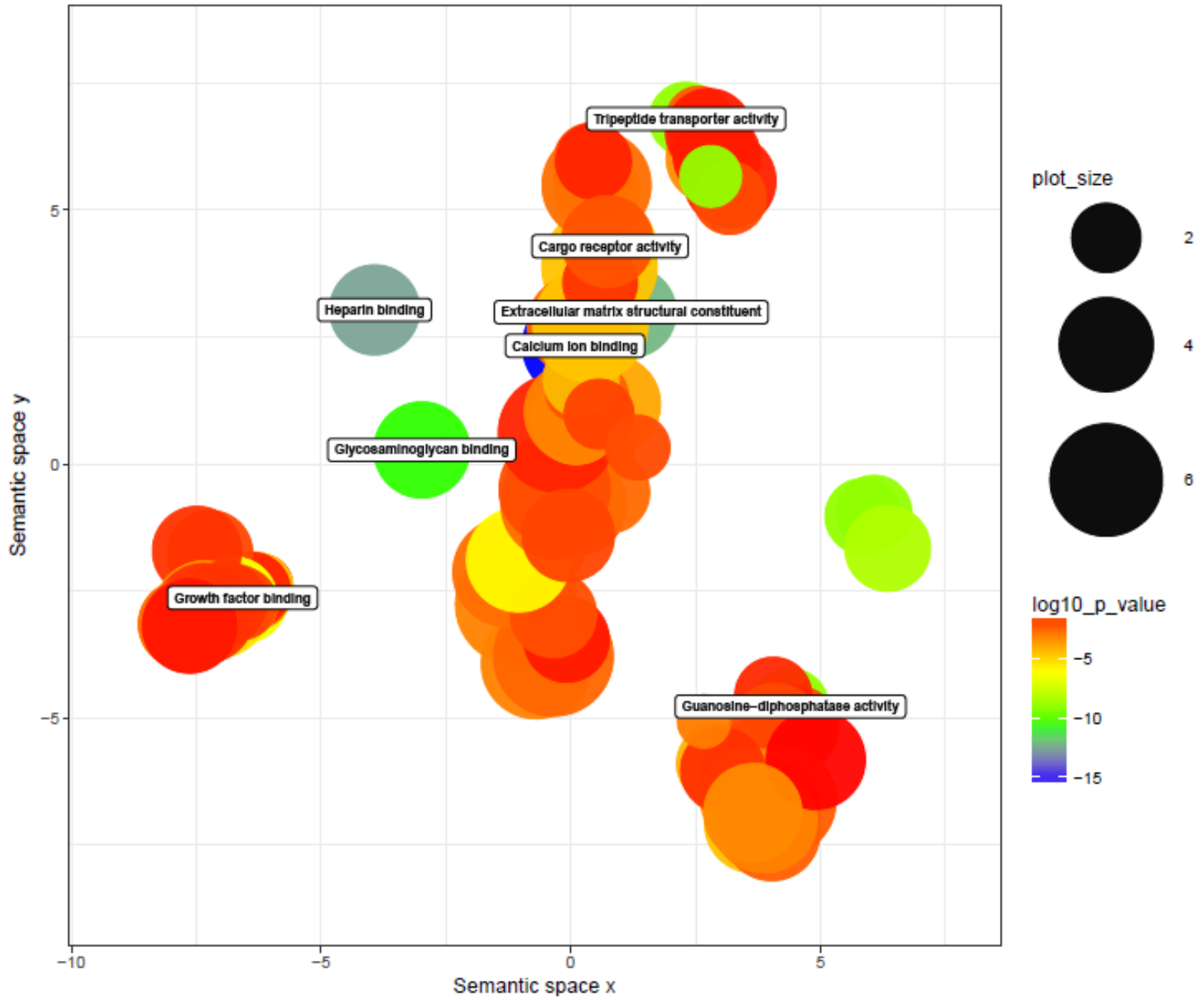

Figure 6.22: Gene ontology scatterplot generated with REVIGO for the caudal fin tissue.

Table 6.1: Top 10 gene ontologies according to their dispensability thoughout the studied tissues with their respective GO identifications, descpription and log10 *p*-values.

Term ID	Description	Log10 *p*-value
Gill + Branchial Arch		
GO:0003674	Molecular function	-2.0408
GO:0004871	Signal transducer activity	-2.7187
GO:0004930	G-protein coupled receptor activity	-3.1987
GO:0005198	Structural molecule activity	-2.5023
GO:0005254	Chloride channel activity	-5.6869
GO:0005488	Binding	-1.9421
GO:0005525	GTP binding	-5.2744
GO:0008417	Fucosyltransferase activity	-3.0894
GO:0030296	Protein tyrosine kinase activator activity	-1.7262
GO:0060089	Molecular transducer activity	-2.6782
Liver		
GO:0001848	Complement binding	-11.7678
GO:0003674	Molecular function	-3.5133
GO:0003824	Catalytic activity	-2.8902
GO:0004497	Monooxygenase activity	-8.4881
GO:0004866	Endopeptidase inhibitor activity	-32.4708
GO:0004872	Receptor activity	-2.2016
GO:0015269	Calcium-activated potassium channel activity	-2.0978
GO:0060089	Molecular transducer activity	-2.2016
GO:0098772	Molecular function regulator	-5.4091
GO:0004736	Pyruvate carboxylase activity	-6.7557
Spleen		
GO:0000982	Transcription factor activity, RNA polymerase II core promoter proximal region sequence-specific binding	-4.0998
GO:0001071	Nucleic acid binding transcription factor activity	-1.823
GO:0003674	Molecular function	-8.5615
GO:0003823	Antigen binding	-36.3099
GO:0003824	Catalytic activity	-6.7292

GO:0004871	Signal transducer activity	-4.8418
GO:0004888	Transmembrane signalling receptor activity	-7.4646
GO:0005344	Oxygen transporter activity	-3.3524
GO:0005488	Binding	-6.7272
GO:0008236	Serine-type peptidase activity	-24.9918
	Gonad	
GO:0003674	Molecular function	-3.0068
GO:0004518	Nuclease activity	-6.0794
GO:0005200	Structural constituent of cytoskeleton	-4.4164
GO:0005488	Binding	-4.8747
GO:0022841	Potassium ion leak channel activity	-1.8362
GO:0030284	Estrogen receptor activity	-3.3379
GO:0032190	Acrosin binding	-8.3438
GO:0043027	Cysteine-type endopeptidase inhibitor activity involved in apoptotic process	-1.8251
GO:0016671	Oxidoreductase activity, acting on a sulphur group of donors, disulphide as acceptor	-1.9334
GO:0003916	DNA topoisomerase activity	-1.7576
	Midgut	
GO:0003674	Molecular function	-2.4785
GO:0003824	Catalytic activity	-7.0359
GO:0004871	Signal transducer activity	-1.4276
GO:0005201	Extracellular matrix structural constituent	-4.0918
GO:0005215	Transporter activity	-10.0287
GO:0015075	Ion transmembrane transporter activity	-9.5018
GO:0020037	Heme binding	-5.5905
GO:0034188	Apolipoprotein A-I receptor activity	-3.0854
GO:0060089	Molecular transducer activity	-2.5257
GO:0061134	Peptidase regulator activity	-1.3331
	White Muscle	
GO:0003674	Molecular function	-3.7038
GO:0003774	Motor activity	-16.3302
GO:0003824	Catalytic activity	-2.5221

GO:0005198	Structural molecule activity	-5.4739
GO:0005244	Voltage-gated ion channel activity	-6.6367
GO:0005488	Binding	-5.1478
GO:0008092	Cytoskeletal protein binding	-38.85
GO:0008307	Structural constituent of muscle	-17.6977
GO:0017080	Sodium channel regulator activity	-3.8633
GO:0030374	Ligand-dependent nuclear receptor transcription coactivator activity	-1.8558
	Red Muscle	
GO:0003674	Molecular function	-7.7557
GO:0003774	Motor activity	-29.4809
GO:0003824	Catalytic activity	-16.7441
GO:0005198	Structural molecule activity	-2.9348
GO:0005215	Transporter activity	-13.3334
GO:0005488	Binding	-6.4375
GO:0008092	Cytoskeletal protein binding	-35.012
GO:0008307	Structural constituent of muscle	-11.3186
GO:0008426	Protein kinase C inhibitor activity	-3.227
GO:0009055	Electron carrier activity	-8.8509
	Kidney	
GO:0001134	Transcription factor activity, transcription factor recruiting	-1.6283
GO:0004222	Metalloendopeptidase activity	-6.0987
GO:0005212	Structural constituent of eye lens	-1.4517
GO:0005215	Transporter activity	-2.1819
GO:0015081	Sodium ion transmembrane transporter activity	-4.4874
GO:0035375	Zymogen binding	-5.8742
GO:0038024	Cargo receptor activity	-5.8171
GO:0060089	Molecular transducer activity	-3.9128
GO:0060228	Phosphatidylcholine-sterol O-acyltransferase activator activity	-2.0346
GO:0004305	Ethanolamine kinase activity	-2.088
	Head Kidney	
GO:0004499	N,N-dimethylaniline monooxygenase activity	-3.78

GO:0005215	Transporter activity	-5.0102
GO:0005496	Steroid binding	-2.3833
GO:0015291	Secondary active transmembrane transporter activity	-9.397
GO:0039660	Structural constituent of virion	-1.4308
GO:0004769	Steroid delta-isomerase activity	-1.8265
GO:0004067	Asparaginase activity	-1.8461
GO:0004550	Nucleoside diphosphate kinase activity	-2.3366
GO:0004611	Phosphoenolpyruvate carboxykinase activity	-1.8422
GO:0030165	PDZ domain binding	-2.373
	Brain + Pituitary	
GO:0001190	Transcriptional activator activity, RNA polymerase II transcription factor binding	-1.3412
GO:0098811	Transcriptional repressor activity, RNA polymerase II activating transcription factor binding	-1.3412
GO:0004111	Creatine kinase activity	-2.1739
GO:0005200	Structural constituent of cytoskeleton	-1.6445
GO:0016247	Channel regulator activity	-3.0348
GO:0022824	Transmitter-gated ion channel activity	-6.4015
GO:0005230	Extracellular ligand-gated ion channel activity	-5.0896
GO:0015276	Ligand-gated ion channel activity	-2.6381
GO:0004970	Ionotropic glutamate receptor activity	-4.8631
GO:0022835	Transmitter-gated channel activity	-6.4015
	Caudal Fin	
GO:0003674	Molecular function	-2.9932
GO:0004382	Guanosine-diphosphatase activity	-8.9773
GO:0005198	Structural molecule activity	-2.9746
GO:0005201	Extracellular matrix structural constituent	-12.2462
GO:0005488	Binding	-2.3707
GO:0005509	Calcium ion binding	-15.3665
GO:0030414	Peptidase inhibitor activity	-1.8947
GO:0038024	Cargo receptor activity	-5.3617
GO:0042937	Tripeptide transporter activity	-8.9773
GO:0060089	Molecular transducer activity	-2.5691