UVIC UNIVERSITAT DE VIC

Master of Science in Omics Data Analysis

Master Thesis

Study of tergal glands morphogenesis through an integrative analysis of genomic data

Author:
Guillem Ylla Bou

Supervisor:
Dr. Xavier Belles, Evolution of Insect Metamorphosis Group IBE (CSIC-UPF)

Department of Systems Biology

University of Vic – Central University of Catalonia

September 2014

UNIVERSITY OF VIC

Abstract

Study of tergal glands morphogenesis through an integrative analysis of genomic data

by Guillem Ylla Bou

About 50% of living species are holometabolan insects. Therefore, unraveling the origin of insect metamorphosis from the hemimetabolan (gradual metamorphosis) to the holometabolan (sudden metamorphosis at the end of the life cycle) mode is equivalent to explaining how all this biodiversity originated. One of the problems with studying the evolution from hemimetaboly to holometaboly is that most information is available only in holometabolan species. Within the hemimetabolan group, our model, the cockroach *Blattella germanica*, is the most studied species. However, given that the study of adult morphogenesis at organismic level is still complex, we focused on the study of the tergal gland (TG) as a minimal model of metamorphosis. The TG is formed in tergites 7 and 8 (T7-8) in the last days of the last nymphal instar (nymph 6). The comparative study of four T7-T8 transcriptomes provided us with crucial keys of TG formation, but also essential information about the mechanisms and circuitry that allows the shift from nymphal to adult morphogenesis.

Acknowledgements

I would like to thank all the lab mates for the discussions and feedback that I obtained from them, as well as for the tips and advices that they were always giving me. I want to especially mention Carolina Santos, who has always been so helpful and provided the SEM pictures.

To my supervisors, Xavier Belles and M. Dolors Piulachs, that always helped and encouraged me, and whom I can consider my scientific parents.

I also would like to thank to Aleksandra Cybulska, who helped me with the English and corrected the biggest English mistakes from this work.

And finally, to the professors of the UVic for all that they taught me since I started to study biotechnology until now, when I am finishing my master project.

Contents

Abstract **i**

Acknowledgements **ii**

Contents **iii**

List of Figures **v**

List of Tables **vi**

Abbreviations **vii**

0 Introduction **1**

0.1 Insects importance 1

0.2 Background in insects evolution 2

0.3 Historical view of metamorphosis 3

0.4 Current view of metamorphosis 5

0.5 *Blattella germanica*'s tergal gland as a minimal model of metamorphosis . 6

1 mRNA-Librerires Preparation and Sequencing **8**

1.1 Collection of samples 8

1.2 Sequencing 9

1.3 Sequencing output 10

1.4 Other samples 10

2 Reference Database **11**

2.1 De-novo assembly 11

2.2 Assembly metrics 12

2.3 Functional annotation of reference transcriptome 13

2.4 Gene ontology terms obtention 13

2.5 Database preparation 14

3 Analysis of differential Gene Expression **15**

3.1 Processing reads 15

3.2 Reads mapping and construction of an abundance matrix 15

3.3 Differential gene expression 16

4 Functional Analysis of Transcriptomes 18

4.1 Gene ontology enrichment analysis . . . 18

4.2 Transcription factors search . . . 20

4.2.1 PFAM motifs search . . . 20

4.2.1.1 TFs quantitative analysis . . . 21

4.2.1.2 TFs qualitative analysis . . . 21

4.3 General discussion of TFs function . . . 23

5 Gene Regulatory Networks 24

5.1 *D. melanogaster* genes homology . . . 24

5.2 Transcription factor interactions . . . 25

5.3 Network Reconstruction . . . 25

5.4 Network topology analysis . . . 25

5.4.1 Scale-free topology . . . 26

5.5 Networks comparison . . . 27

5.5.1 N6Ecd vs N5Ecd . . . 28

5.5.2 N6D1C vs N6D1T . . . 29

6 Conclusions 31

6.1 Working without a reference genome . . . 31

6.2 Differential expression analysis without replicates . . . 31

6.3 Functional annotation . . . 31

6.4 Database and web interface . . . 32

6.5 The role of transcription factors in tergal glands formation . . . 32

6.6 Gene regulatory networks . . . 33

6.7 Beyond this work . . . 33

Appendix A PFAM Motifs related with TFs 34

Appendix B TFs-Motifs abundance 35

Appendix C GO-terms enriched in N6Ecd 36

Appendix D GO-terms enriched in N6D1C 38

Appendix E Graphs 40

Appendix F Annimation Frames 44

Bibliography 48

List of Figures

0.1 Insects diversity . . . 2
0.2 Insects phylogenetic . . . 3
0.3 Scarab pendant . . . 4
0.4 Types of Metamorphosis . . . 5
0.5 *B. germanica's* Tergal glands . . . 7

1.1 Sample collection stages . . . 9
1.2 Roche's 454-Junior machine . . . 10

2.1 Database accession . . . 14

3.1 Alignment visualization . . . 16
3.2 DEGs plot . . . 17

4.1 GO tree . . . 19
4.2 Venn-Diagram . . . 22

5.1 Barabasi network . . . 26
5.2 network degree distribution . . . 27
5.3 Vertices overlap . . . 28
5.4 Annimation of JH remodeling the GRN . . . 30

E.1 N6Ecd Graph . . . 40
E.2 N5Ecd Graph . . . 41
E.3 N6D1C Graph . . . 42
E.4 N6D1T Graph . . . 43

F.1 Annimation: frame 1 . . . 44
F.2 Annimation: frame 2 . . . 45
F.3 Annimation: frame 3 . . . 46
F.4 Annimation: frame 4 . . . 47

List of Tables

1.1 Output number of reads . . . 10

2.1 Processed reads . . . 12

4.1 TFs in transcriptomes . . . 21

5.1 Graph parameters . . . 28

A.1 PFAM Motifs related with TFs . . . 34

B.1 TFs-Motifs abundance . . . 35

C.1 N6Ecd Enriched GO-terms . . . 36
C.2 N6Ecd Enriched GO-terms . . . 37

D.1 N6D1C Enriched GO-terms . . . 38
D.2 N6D1C Enriched GO-terms . . . 39

Abbreviations

20E	20-hydroxyecdysone
DEG	Differentially Expressed Gene.
GRN	Gene Regulatory Network.
GO	Gene Ontology.
N5Ecd	Nymph 5 Ecdysone peak transcriptome when TG is not being formed.
N6Ecd	Nymph 6 Ecdysone peak transcriptome when TG is being formed.
N6D1C	Nymph 6 day 1 when the insect starts the “adult genetic program”.
N6D1T	Nymph 6 day 1, treated with juvenile hormone, which inhibits metamorphosis.
T7	Tergite 7.
T8	Tergite8.
TF	Transcritpion Factor.
TG	Tergal Gland.
ORF	Open Reading Frame.

Chapter 0

Introduction

The main purpose of this chapter is to contextualize this work inside the historical view that humans had about metamorphosis and explain why the cockroach *Blattella germanica* is as a good model organism for evolutionary studies.

0.1 Insects importance

Insects are the most diverse lineage of all life in number of species, and ecologically they dominate terrestrial ecosystems (Engel and Grimaldi, 2004). With more than one million of described species, insects represent about the 90% of the total number of animal species (Belles, 2013). From all the described insect species, 90% of them are metamorphic and the 83% of them have Holometabolous metamorphosis (Figure 0.1).

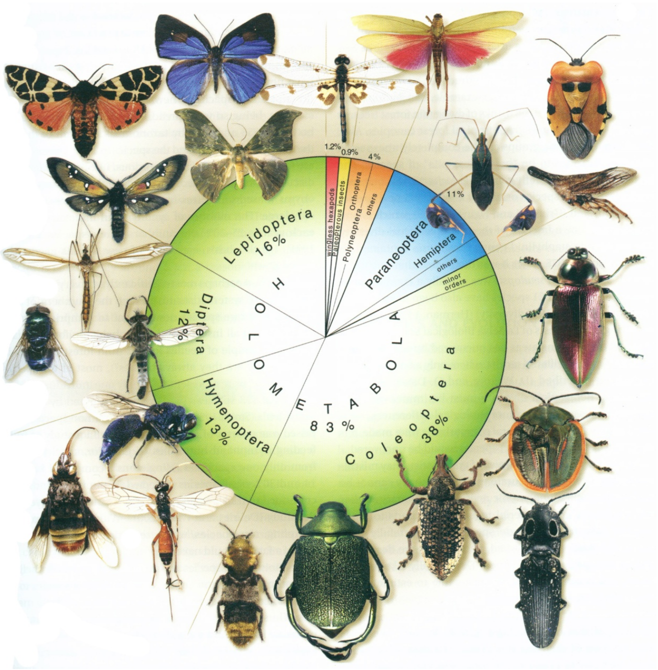


Figure 0.1: Insects diversity, image from Grimaldi and Engel (2005).

0.2 Background in insects evolution

Nowadays, holometabolous metamorphosis is the most common type of metamorphosis among insects and includes a huge diversity of species with a wide range phenotypes. The holometabolous metamorphosis, however, is the most modified kind of metamorphosis that derived from the basal type that is the "hemimetabolous". Therefore, explaining how hemimetabolous metamorphosis derived to the holometabolus one is equivalent to explaining how all this huge biodiversity within insects emerged.

The most studied insect is *Drosophila melanogaster*, which has been used during lots of years as a model organism to study holometabolous metamorphosis among other subjects. Within the hemimetabolous group of insects, really few work – comparing with *D. melanogaster* – has been done, with *B. germanica* being the most well known organism. The phylogenetic tree of hexapods (Figure 0.2) shows the order that contains the *B. germanica* and its evolutionary relation with the other orders.

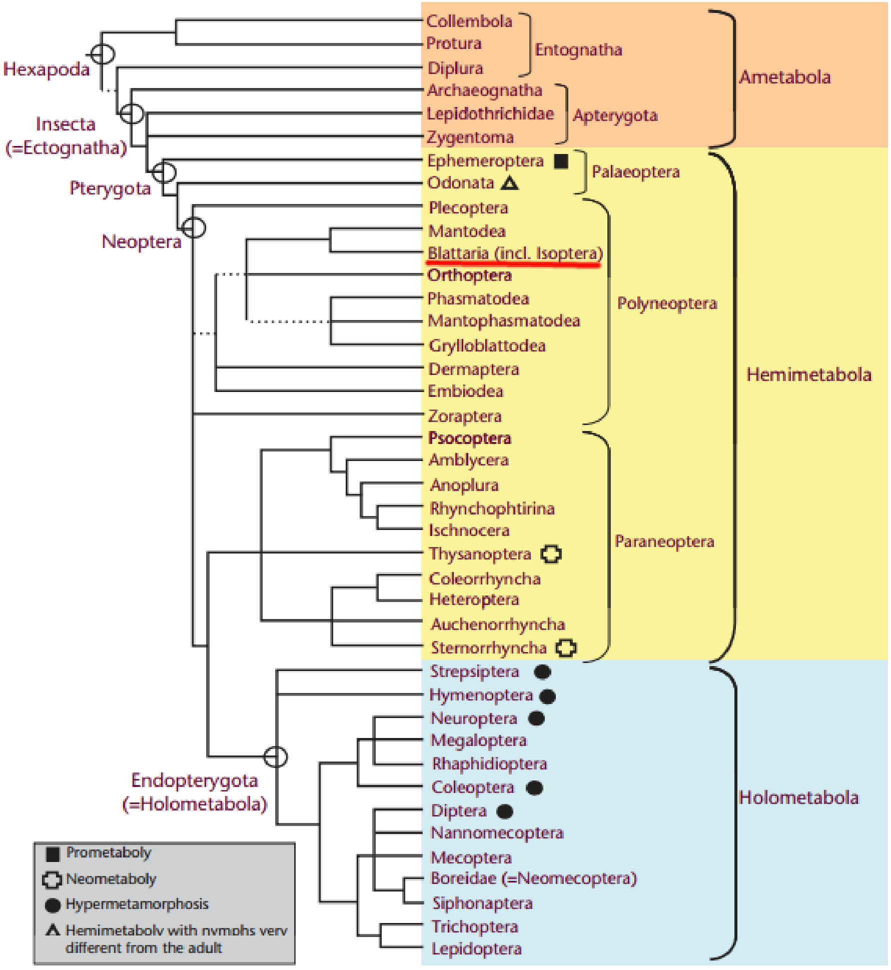


Figure 0.2: Holometabolous insects are part of a monophyletic group evolved from hemimetabolan insects. Modified from Wheeler (2001) and Belles (2011).

Notice that *Blattaria* belongs to a paraphyletic group that corresponds to hemimetabola, a group that contains "Polyneoptera", "Paraneoptera" and "Paleoptera", that are the ancestors of holometabolan linage.

0.3 Historical view of metamorphosis

One of the first great human civilizations, the ancient Egypt civilization, was already fascinated by metamorphosis. In fact, the beetles became one of the most important symbols of ancient Egypt, and were represented in architectural ornaments of important places like all Pharaoh tombs as well as in their jewellery (Figure 0.3).

Later on, Aristotle was also intrigued by metamorphosis and he described the caterpillar as a continuation of embryonic life; “caterpillar is nothing more than a soft egg” he wrote.

Figure 0.3: Scarab pendant from Egypt dated at 12th Dynasty, around 1890 BC. It is made of electrum inlaid with carnelian, green feldspar and lapis lazuli. *(image obtained from British museum)*

The first important formal theory about metamorphosis was written in 1651 by William Harvey, where he hypothesized that embryo is forced to hatch before the complete development due to the scarce amount of nutrients in the egg. After this theory, the studies of Jan Swammerdam pointed that larva is not a sort of egg but a transitional stage between egg and adult. He also categorized the insects in terms of four metamorphosis types; insects that grow without transformation (exemplified by lice), insects that develop wings progressively without quiescent stage (locusts), insects that develop under the larva cuticle through a quiescent pupa stage (butterflies and beatles), and the fourth group that includes insects that pass the pupal stage under the skin of last larva stage (flies) (Belles, 2011).

Nevertheless, metamorphosis continued to be surrounded by mystery during a lot of years for a wide part of the population. For example, in the middle of nineteenth century, as Darwin explains in his book “The Voyage of the Beagle” (Darwin, 1839), that during a visit in Chile the authorities had arrested a man called Renous accused of witchcraft because he was capable of transforming disgusting worms into beautiful butterflies.

0.4 Current view of metamorphosis

Fortunately, nowadays thanks to the modern science, the vision that we have about metamorphosis completely changed comparing to nineteenth century. The classification of metamorphosis classes, however, remains quite similar to the one described by Jan Swammerda in the seventeenth century.

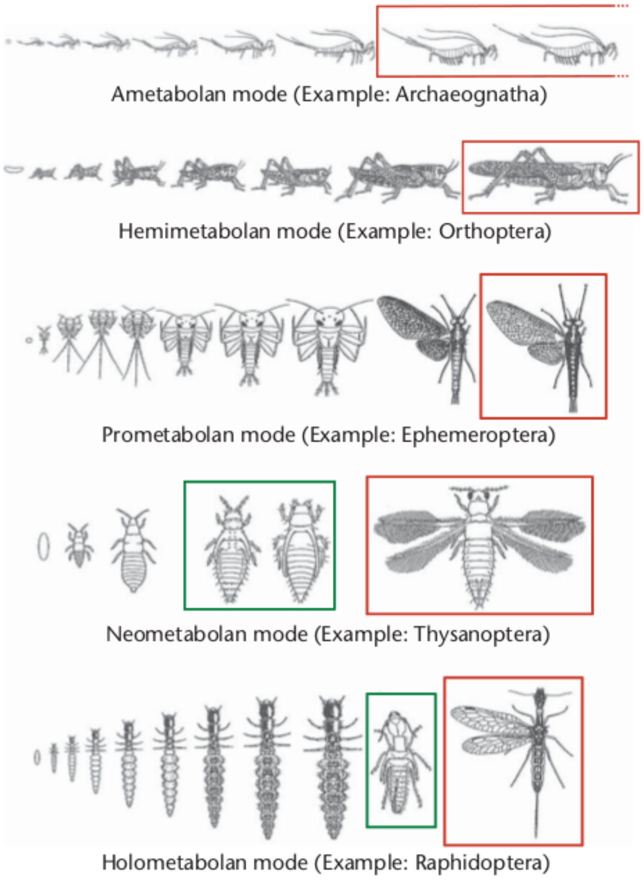

Ametabolan mode (Example: Archaeognatha)
Hemimetabolan mode (Example: Orthoptera)
Prometabolan mode (Example: Ephemeroptera)
Neometabolan mode (Example: Thysanoptera)
Holometabolan mode (Example: Raphidoptera)

Figure 0.4: The main types of insect metamorphosis: ametabolan, hemimetabolan, and holometabolan. Also the subtypes of hemimetabolan: prometabolan and neometabolan. Quiescent stages are showed in a green square. Adult, reproductively competent stages are showed in a red square. In the ametabolans, the red square is open because the adult continues moulting. Modified from Belles (2011).

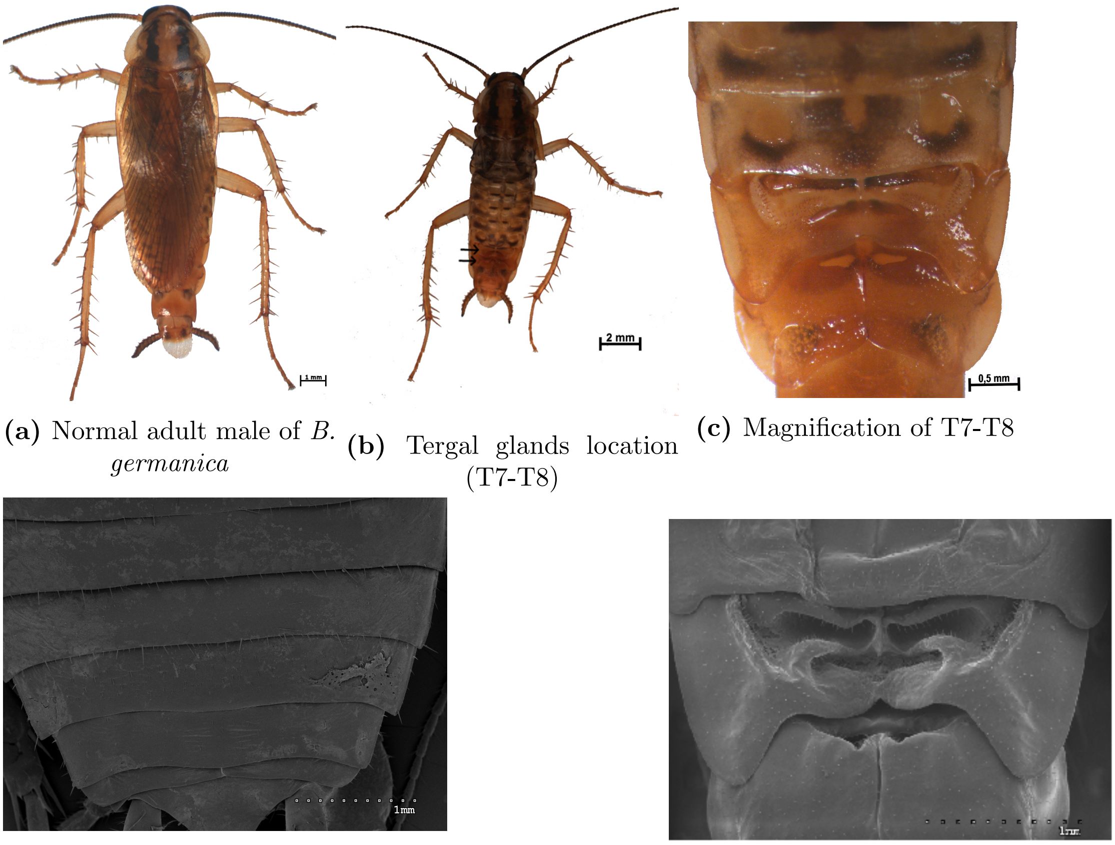
At this moment, scientific community classifies insect metamorphosis in three main types of insect: Ametabolan, which are insects without metamorphosis; the hemimetabolan, which are doing the basal methamorphosis; and the holometabolan that are doing the most derived methamorphosis. Within hemimetabolans, two subtypes of metamorphosis (prometabolan and neometabolan) are differentiated (Figure 0.4).

Our comprehension about the metamorphic transitions is still low. Even though nowadays we are aware of the importance of transcriptional regulation of genes in order to control the metamorphosis process, the exact mechanism remains unknown. It has been shown that hormones play an important role in order to initiate and control a metamorphic changes (Riddiford, 2008), but the downstream mechanisms are not yet clear.

0.5 *Blattella germanica*'s tergal gland as a minimal model of metamorphosis

Blattella germanica – the most well known hemimetabolous insect from an endocrinological point of view – has 6 nymphal instars followed by an adult instar. The end of each nymphal instar its signaled by a peak of the hormone 20-hydroxyecdysone (20E), which induces the molt through the activation of a transcriptional cascade (Hirn et al., 1979). In the sixth nymphal instar, apart from the ecdysone peak, the Juvenil Hormone (JH) that was present in all nymphal stages dramatically decays. This lack of JH triggers the formation of an adult that will emerge in the next molt.

In this molt to adult, a gland that was not present in the nymphal instars is formed in tergits 7 and 8 (Figure 0.5). This gland, whose function is to attract females in order to mate, is called Tergal Gland (TG). Since the metamorphosis is a really complex and complicated process, we used this small gland that is formed from scratch in the adult cockroach as minimal model of insect metamorphosis.



(a) Normal adult male of *B. germanica*

(b) Tergal glands location (T7-T8)

(c) Magnification of T7-T8

(d) Tergites 7-8 of a N6 male viewed with Scanning electron microscope (SEM)

(e) Tergal glands viewed with Scanning electron microscope (SEM)

Figure 0.5: A) Picture of a normal male of *B. germanica*. B) Male with cut wings and arrows pointing to tergites 7-8, where TG are located. C) Amplification of tergites 7 and 8 with clear view of the tergal glands. D) Picture obtained by SEM of tergites 7 and 8 in nymph 6 male (when no tergal glands are present). E) Picture obtained by SEM of adult tergites 7 and 8 (with tergal glands).

Unraveling how the formation of this gland is regulated with such a precision might help us to understand the basic mechanism of hemimetabolan metamorphosis regulation. This mechanism could be compared with the well known holometabolan metamorphosis ones in order understand this evolutionary transition.

Chapter 1

mRNA-Librerires Preparation and Sequencing

In order to reveal how the formation of tergal gland is regulated, 4 transcriptomes of tergites 7 and 8 in key developmental stages were obtained. In this chapter, it is briefly explained how the RNA-seq data was obtained in the laboratory previously of my incorporation to the project.

1.1 Collection of samples

A colony of German cockroach, *Blattella germanica*, was maintained in environmental chambers at 28°C under L:D=16:8 h conditions. Dog chow and water were provided *ad libitum.*

The tergal glands are formed in the molt to adult, which means that in the end of the 6th nymphal instar the gland should be almost totally constructed. For this reason, one of the developmental stages to obtain samples was chosen in the middle of nymph 6, during the Ecdysone peak when the formation of the tergal gland should be occurring. In order to identify which genes are responsible of forming the tergal gland, another transcriptome in the 5th nymphal instar – where instead of constructing a tergal gland, the insect is just a forming a nymph with flat tergites – was obtained.

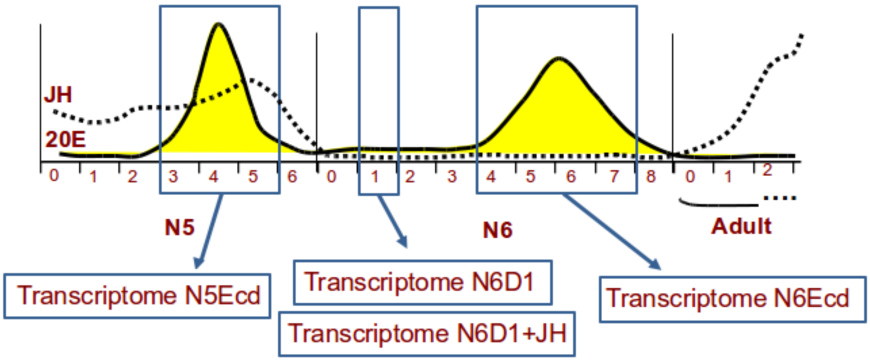


Figure 1.1: The 4 samples for RNA-seq were collected in 3 developmental stages (plus a treatment) important for their hormonal framework.

At the beginning of the 6th nymphal instar, the expression of the early genes of the TG construction should be being happening. In order to detect these early genes, 2 transcriptomes more were done: One in nymph 6 day 1, and another in the same stage but in animals previously treated with Juvenil hormone, which inhibits the molt to adult giving a supernumerary nymph 7 without tergal gland (Figure 1.1).

From each chosen stage a pool of insects was selected and the dissections of tergites 7 and 8 were performed. The mRNA of the samples was obtained and processed for the 454-Junior RNA-sequencing using the Roche protocols.

1.2 Sequencing

The samples were send to the PRBB genomic services, where they sequenced the samples with Roche's 454-Junior technology (Figure 1.2). For each sample a ".sff" file was obtained containing raw reads and their quality scores.

Figure 1.2: Roche's 454-Junior machine

1.3 Sequencing output

The number of reads obtained in each sample is shown in Table 1.1.

Table 1.1: Number of raw reads obtained from each sample

Transcriptome	# reads
N6Ecd	131,3297
N5Ecd	102,019
N6D1C	100,140
N6D1T	82,279

1.4 Other samples

Apart from the previous mentioned 4 samples, the group already had 7 transcriptomes sequenced with the same technology and procedure but from different tissues and experimental conditions. The raw reads from this 7 transcriptomes were also used in order to generate a reference transcriptome as it is described in the following chapter.

Chapter 2

Reference Database

Since *Blattella germanica*'s genome has not been sequenced, we have no reference where to map the reads. Furthermore, the amount of publicly available genomic information about this organism remains low. That is why by using all the previously generated data in the group (in total 11 transcriptomes; 4 for this project plus 7 from previously projects) the assembly of the reads and the further annotation was performed following the below described pipeline.

2.1 De-novo assembly

In order to generate a "reference transcriptome" where to map our reads for assessing gene expression levels, we used 11 transcriptomes previously available in the laboratory to perform the assembly.

In order to perform the assembly, we used the Roche's proprietary software Newbler v2.5p1 – which shows the best performance for assembling de novo 454 reads (Kumar and Blaxter, 2010) – . Before the assembly, Newbler trim the sequencing adapters and also can filter contaminant sequence if a fasta file with possible contaminants is passed as a parameter.

We prepared this "fasta" file with possible contaminants including: Five genomes of species of Blattella's endosymbiont *Blattabacterium* (Ref seq. NC_013454; NC_017924.1; NC_016621.1; NC_013418.2; NC_016146.1); the whole genome of *Blattella germanica* *densovirus* (Ref seq. NC_005041.2); *Blattella germanica mitochondrion genome* (Ref seq. NC_012901.1); *Caenorhabditis elegans* nuclear genome WBcel215 (GCA_000002985.1); And the *Escherichia coli* genome (AE014075.1);

After processing, 1,306,009 long-reads (446,021,729 bases) where included to the assembly with the Newbler standard parameters ("minimum read length"=20; "minimum overlap length"=40; "minimum overlap identity"=90%).

Table 2.1: Number of reads of the 11 transcriptomes used in the assembly before and after processing. **OT** refers to the 7 other transcriptomes previously obtained in the lab.

Dataset	# raw Reads	# Reads after processing	# raw Bases	# after processing
N6Ecd	131,329	128,976	50,601,265	49,759,967
N5Ecd	102,019	100,268	40,697,795	40,087,037
N6D1C	100,140	97,844	38,263,300	37,416,662
N6D1T	82,279	80,604	34,388,508	33,739,945
OT1	15,4932	144,829	40304096	37,244,372
OT2	139,895	135,144	60,435,716	58,413,046
OT3	88,487	85,430	37,167,556	35,841,093
OT4	212,195	210,973	52,334,374	51,796,827
OT5	61,219	60,856	1,849,1286	1,835,0726
OT6	157,564	156,227	40,762,859	4,0203,489
OT7	106,552	104,858	4,3783,056	43,168,565
SUM	**1,336,631**	**1,306,009**	**457,229,811**	**446,021,729**

The assembly process requires a high computational power; in our case we ran the Newbler during several days with more than 50Gb of RAM memory and 3 cores in the informatics cluster of the institute.

2.2 Assembly metrics

In total, 715,597 reads (222,616,399 bases) were assembled in 32,606 contigs (23,002,710 bases) from which 17,980 where more than 500bp with N50 contig size about 1,165bp. The largest contig obtained was 11,024 bases while the shortest one 97 bases. The average contig size was 706.54 bases and the median size 529 bases. The 32,606 contigs contained 23,002,710 bases, so the empirical per base coverage depth (Sims et al., 2014) was: 19.39X

The reference transcriptome created was functionally annotated by homology with the home-made pipeline described below.

2.3 Functional annotation of reference transcriptome

The functional annotation of the 32.606 contigs was performed by homology using the Blasts algotithm (Altschul et al., 1990). The running time of Blast is really influenced by the database size, so that, the following series of Blast with smaller databases allow to annotate the reference transcriptome reducing the running time compared with doing a single Blast**x** run of the transcriptome against all ncbi sequences.

1. Firstly, all the contigs were compared by Blast**n** to all manually curated sequences available in the laboratory. If a blast hit (e-value threshold of 0.001) was obtained for a contig, the name of the homologous sequence was transfered to the query contig.
2. All the contigs that did not give a blast hit before were compared by Blast**x** against all “arthropod“ all sequences available in ncbi.
3. All the contigs that were not annotated by homology in the previous steps were compared by Blast**x** with all available sequences in ncbi.
4. All the contigs that in this point were not annotated by homology were labeled as ”Unknown sequences“.

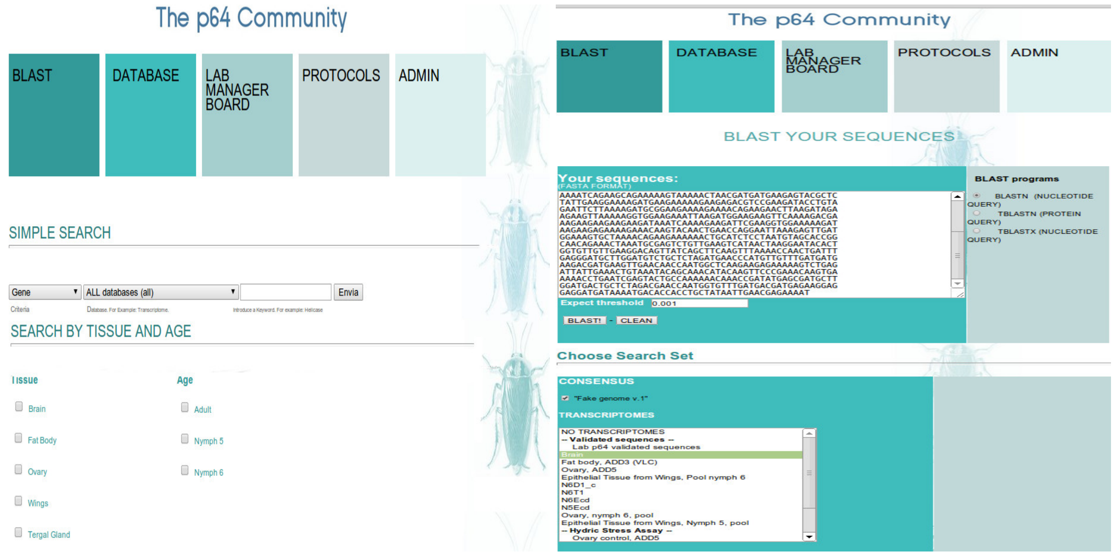
This pipeline provides better annotation accuracy than doing single Blast**x** of the transcriptome against all ncbi sequences and its short running time makes it suitable for periodically updates of the annotation. Following this approach, we identified homologous sequences for 15,624 contigs.

2.4 Gene ontology terms obtention

With each sequence annotated with an hortologous one, we used the orthologous accession code to retrieve its GO-terms (Gene and Consortium, 2000) and associated them to our reference sequence. In this way we retrieved 2,922 GO-terms.

2.5 Database preparation

The reference transcriptome generated together with the best blast hit descriptor, the accession code to the blast hit and an internal accession number were uploaded to a MySQL database. In order to make the database available to all the research group, and in a while to all scientific community, a php-html webpage was created with the collaboration of Aníbal de Horna. The webpage offers two main ways to access the transcriptomes information: The first one is a search by "gene name" with some filter criteria available (Figure 2.1a) that shows if a genes is present in our transcriptomes and in which abundance in each one. The second way is a blast search, where inserting a query sequence a blastn, tblastn and tblastx can be performed against each one of the transcriptomes (Figure 2.1b).



Gene ▾ ALL databases (all) ▾

Tissue: □ Brain □ Fat Body □ Ovary □ Wings □ Tergal Gland

Age: □ Adult □ Nymph 5 □ Nymph 6

```
AAAATCAGAAGCAGAAAAAGTAAAAACTAACGATGATGAAGAGTACGCTC
TATTGAAGGAAAAGATGAAGAAAAAGAAGAGACGTCCGAAGATACCTGTA
GAATTCTTAAAAGATGCGGAAGAAAAGAAAACAGAAGAACTTAAGATAGA
AGAAGTTAAAAAGGTGGAAGAAATTAAGATGGAAGAAGTTCAAAAGACGA
AAGAAGAAGAAGAAGATAAATCAAAAGAAGATTCGAAGGTGGAAAAAGAT
AAGAAGAGAAAAGAAACAAGTACAACTGAACCAGGAATTAAAGAGTTGAT
GGAAAGTGCTAAAACAGAAGAAAAAAACTGCATCTCCTAATGTAGCACCGG
CAACAGAAACTAAATGCGAGTCTGTTGAAGTCATAACTAAGGAATACACT
GGTGTTGTTGAAGGACAGTTATCAGCTTCAAGTTTAAAACCAACTGATTT
GAGGGATGCTTGGATGTCTGCTCTAGATGAACCCATGTTGTTTGATGATG
AAGACGATGAAGTTGAACAACCAATGGCTCAAGAAGAGAAAAAGTCTGAG
ATTATTGAAACTGTAAATACAGCAAACATACAAGTTCCCGAAACAAGTGA
AAAACCTGAATCGAGTACTGCCAAAAAACAAACCGATATGAGCGATGCTT
GGATGACTGCTCTAGACGAACCAATGGTGTTTGATGACGATGAGAAGGAG
GAGGATGATAAAATGACACCACCTGCTATAATTGAACGAGAAAAT
```

Expect threshold 0.001

● BLASTN (NUCLEOTIDE QUERY)
○ TBLASTN (PROTEIN QUERY)
○ TBLASTX (NUCLEOTIDE QUERY)

☑ "Fake genome v.1"

NO TRANSCRIPTOMES
-- **Validated sequences** --
Lab p64 validated sequences
Brain
Fat body, ADD3 (VLC)
Ovary, ADD5
Epithelial Tissue from Wings, Pool nymph 6
N6D1_c
N6T1
N6Ecd
N5Ecd
Ovary, nymph 6, pool
Epithelial Tissue from Wings, Nymph 5, pool
-- **Hydric Stress Assay** --
Ovary control, ADD5

(a) Database search by gene name

(b) Database Blast page

Figure 2.1: Database accession methods.

Chapter 3

Analysis of differential Gene Expression

For a differential expression genes (DEG) analysis, we first need to process reads and then map them to the reference. This chapter presents the steps that we followed.

3.1 Processing reads

The four files from the 454-Junior sequencer output were transformed from the Roche format ".sff" to ".fastq" format with homemade scripts. Subsequently, all the reads where cleaned from the adapter "ˆGACT". Since RSEM (Li and Dewey, 2011) takes in account the quality values, we do not need to do a quality control of the reads.

3.2 Reads mapping and construction of an abundance matrix

In order to do the mapping step, we chose to use the RSEM package (Li and Dewey, 2011) since it has previously shown good performance in estimating abundances when reads are mapped in a rna-seq de novo assembly (Haas et al., 2013). The reference transcriptome was indexed and formated to be used with RSEM, with the RSEM provided command "rsem-prepare-reference".

All the reads were mapped against the reference transcriptome (generated in the chapter 2) with the “rsem-calculate-expression” tool from the RSEM package. An abundance matrix, which has for each reference sequence the number of counts in each trasncriptome, was obtained. The same matrix was created with RPKMs (Mortazavi et al., 2008) normalized abundances.

The “rsem-calculate-expression” tool also provides a “.bam” file as output. The “.bam” file is a binary file that contains the sequence alignment data and can be visualized by an alignment visualization software. We used the software Tablet (Milne et al., 2010) in order to access the “.bam” file (Figure 3.1).

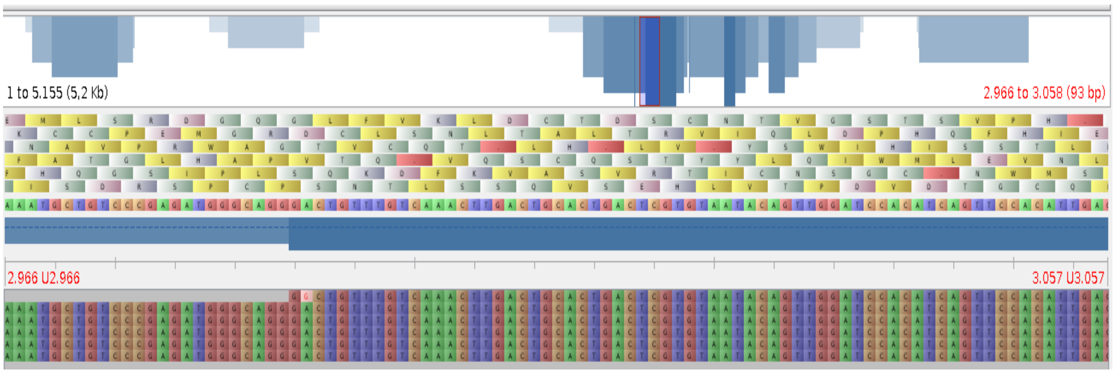


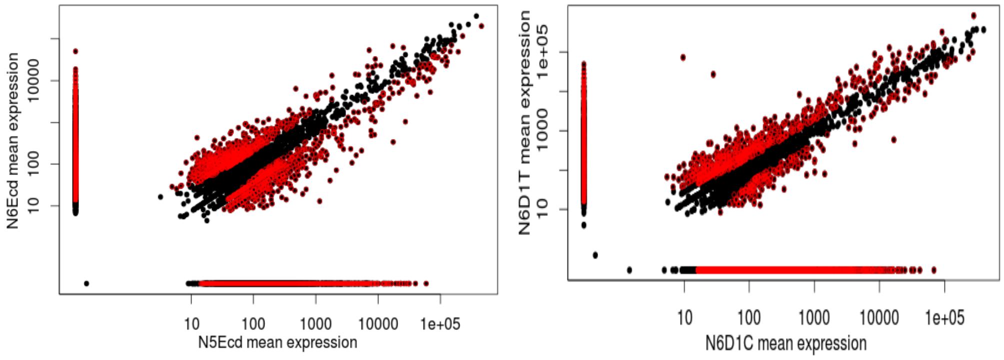
Figure 3.1: Six reads were mapped to the reference. One of the reads has a mismatch that is highlighted with red letter. Image obtained with the Tablet (Milne et al., 2010).

3.3 Differential gene expression

A wide number of programs is available for this purpose, however not so many allows to do a differentially expression analysis with neither an annotated reference genome (we only have rna-seq assembly without features) nor replicates. We used NoiSeq (Ferrer et al., 2011), which is based in a data-adaptive and non-parametric algorithm and allows to compute differential expressed genes even though no replicates are available. It was ran with the following parameters: pnr (size of the simulated samples)=0.2; nss (number of replicates to be simulated)=5; v(variability)=0.02; replicates=“no”.

Working with low coverage and without replicates do not allow us to prove that a specific gene is differentially expressed comparing different conditions, however we can find some groups of genes with a trend to be up or down regulated when comparing two samples (Anders et al., 2013).

We performed a differential expression test comparing N6Ecd vs N5Ecd in order to find the genes that are involved in the formation of the TG, the ones over-expressed in N6Ecd. We also applied a DE analysis comparing the samples of N6D1C against N6D1T. In this comparison we expected to find the genes that are initiating the TG formation, which would be over-expressed in N6D1C.



(a) Expression plot N5Ecd vs N6Ecd **(b)** Expression plot N6D1C vs N6D1T

Figure 3.2: The expression value of each gene in each sample is plotted in black, the genes with probability higher than 90% are highlighted in red.

Once the DEG are computed it is interesting to plot the expression values in each condition and highlight the features declared as differentially expressed (Figure 3.2). A total of 5,615 and 2,621 genes appeared to be over-expressed in N6Ecd and N6D1C respectively, with a NoiSeq probability threshold of 90%.

Chapter 4

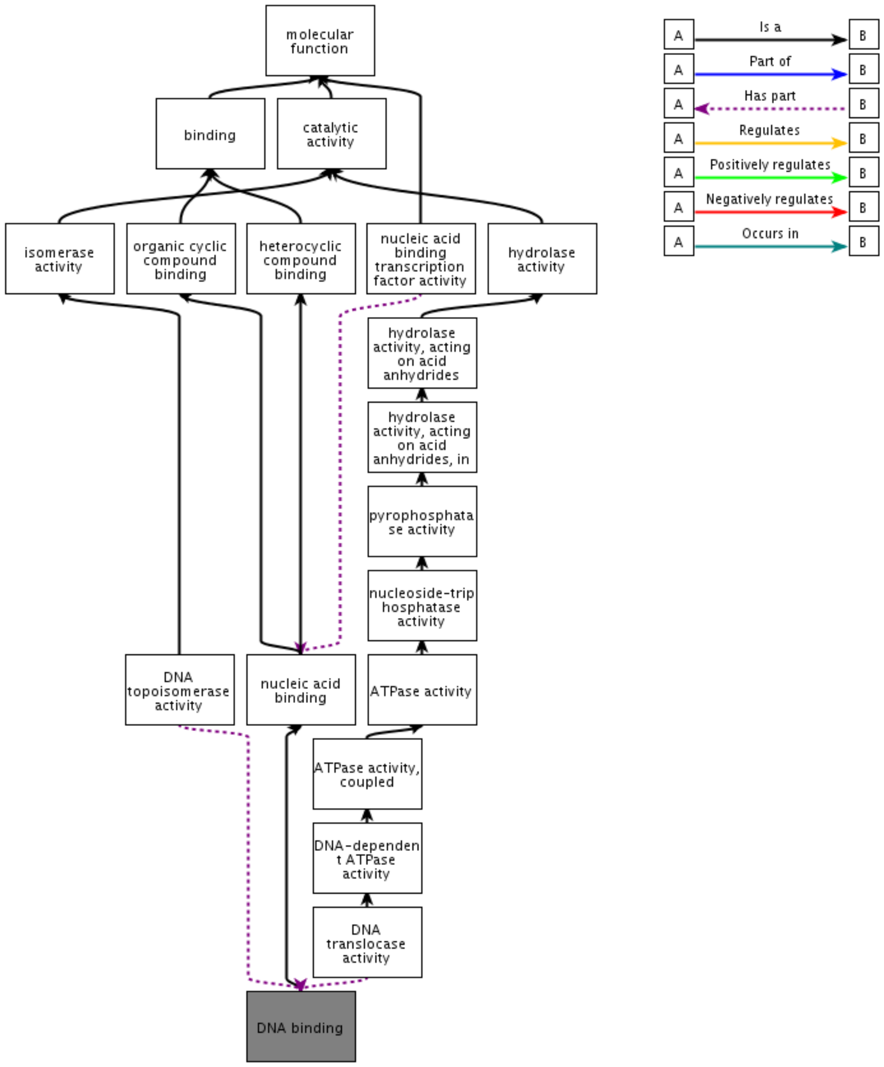
Functional Analysis of Transcriptomes

In this chapter we show the functional analysis of the transcriptomes by integrating the previous obtained data as well as adding new information of different types.

4.1 Gene ontology enrichment analysis

With the 5,615 DEG found in N6Ecd and the 2,621 found in N6D1C we did a GO-enrichment analysis with the R (Team, 2003) package "TopGO" (Alexa and Rahnenfuhrer, 2010) available in Bioconductor (Gentleman et al., 2004).

The lists of enriched GO-terms in over-expressed sequences are shown in Appendix C ans Appendix D, together with their p-values obtained with the fisher test.



QuickGO - http://www.ebi.ac.uk/QuickGO

Figure 4.1: Go chart from QuickGO web portal (Binns et al., 2009).

The analysis showed the GO:0003676 "nucleic acid binding" enriched with the p-value of 0.0388 from the Fisher's exact test in N6Ecd. This GO-term was especially interesting since include all the transcription factors (including all the DNA binding proteins Figure 4.1). Moreover, the GO:0046872 "metal ion binding", which appears as significant up-regulated, is also an indicator of enrichment of TFs since most of them bind to to metal ions, with the Zn-fingers being the most clear example. The correlation between these two GO-terms is also shown in the "co-ocurring terms" tab, of the GO:0003676 and GO:0046872 QuickGO web portal (Binns et al., 2009).

In the comparison between the transcripomes of nymph 6 day 1 control vs treated with JH, the GO-term "nucleic acid binding" did not appear as significantly enriched in

N6D1C, however it was within the TOP-40 most enriched GO terms, as well as its child term “sequence-specific DNA binding”.

4.2 Transcription factors search

From the results of the Gene Ontology Enrichment Analysis, we hypothesized that transcription factors could be the main drivers in the tergal gland formation. For that purpose we focused in the study of this kind of proteins in a two different ways: qualitative and quantitaive.

4.2.1 PFAM motifs search

In order to identify which of our sequences are giving transcription factors in the most reliable way, we chose a direct method non based in homology but in protein domain identification.

First of all, all the reference transcriptome was translated to protein with the 6 possibles open reading frames (ORFs) with the “Transeq” package from the EMBOSS suite version 6.4.0.0 (Rice et al., 2000). After it, with the standalone version of the PfamScan and PFAM-A database (PFAM-A contains entries with high quality and manually curated) (Bateman et al., 2004), we predicted all the PFAM domains contained in the translated sequences.

Due to the sequencing errors and ambiguities in the assembly process (some sequences can have insertions, deletions or substitutions), some sequences may not be entire in the same open reading frame. That is why we gave to each reference sequences all the PFAM domains detected with an E-value lower than 0.05, in all its ORFs.

A total of 12,071 PFAM domains (2,821 different types) contained in 9,743 different reference sequences were retrieved. In order to identify which of them correspond to TF activity, we used the manually accurate list of PFAM-domains directly related to transcription factor activity published by de Mendoza et al. (2013) and available at Appendix A.

For each one of the TF-PFAM domains, we summed the number of “expected counts” from the RSEM output in each trasncriptome obtaining a TF-PFAM domains abundance table. The abundance of each domain was divided by library size to obtain relative abundances (see Appendix B).

4.2.1.1 TFs quantitative analysis

In Table 4.1 the normalized proportion of counts corresponding to TFs is shown.

Table 4.1: Transcription factors in the different transcriptomes

Transcriptome	% of expected counts containing TF domain
N6Ecd	1.8
N5Ecd	1.06
N6D1C	1.5
N6D1T	1.98

We applied a chi-square with Yates continuity correction to compare the proportions of expected counts corresponding to TFs in N6Ecd vs N5Ecd, and N6D1C vs N6D1T . The statistic tests comparing the proportions of TFs in N5Ecd vs N6Ecd and in N6D1C vs N6D1T, gave significant p-value in both two comparisons.

This points out that the proportion of reads corresponding to TFs is enriched in the N6Ecd, as we previously observed, however here we found that the proportion of reads corresponding to TFs is larger in N6D1T than in N6D1C.

Besides this result could seem that contradicts the GO-enrichment analysis, where N6D1C was slightly enriched with TFs related GO-terms, it does not. The GO analysis do not take into count the abundance of each GO term, just a number of times that a GO-term appears. Due to the above mentioned facts, a high number of low abundance of TFs would make appear TFs GOs enriched while a few TFs with really high abundance no.

4.2.1.2 TFs qualitative analysis

After considering all the issues mentioned above, we wanted to check for differences in types of TFs and not just to refer to abundances. We cannot directly use the previously obtained abundances (Appendix B) to check for presence/absence of each TF type since the libraries have different size. If we observe the largest library, N6Ecd, its also the library with more kinds of TFs but we cannot know if it is due a biological reasons or just because it has higher sequencing depth.

In order to avoid such a bias, we subsampled the reads (Huson et al., 2009) selecting 82,279 random reads – the size of the smallest library – from each library. The subsampled reads were mapped against the reference genome and following the same steps performed before we obtained a new TF domains abundance table that was converted to a presence/absence table (presence if a single read for a given TF domain was found in a transcriptome, absence if any read detected). The presence/absence table is shown as a Venn-Diagram (Figure 4.2) for better visualization with the R package "VennDiagram" (Chen and Boutros, 2011) .

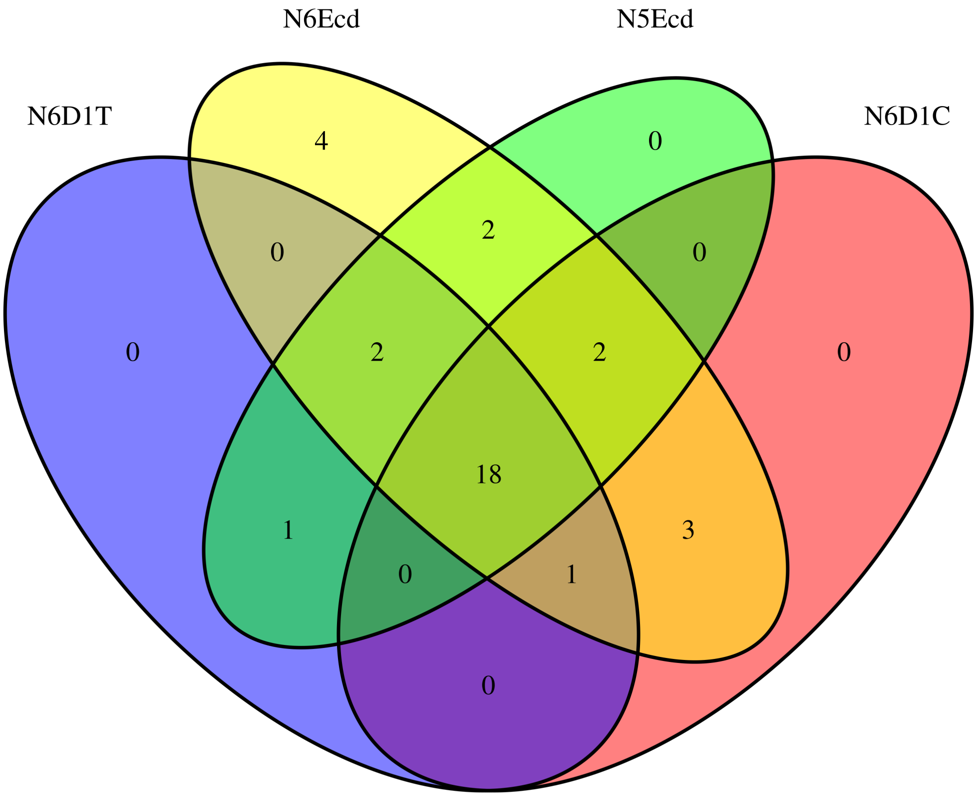


Figure 4.2: The diagram shows the number of different TF-PFAM domain types that are common in each possible comparison of the 4 transcriptomes. We should remark that N6Ecd has 8 TF-pfam-types not detected in the N5Ecd transcriptome. It is also remarkable the highest overlap – 24 types in common – between the transcriptomes where TG is being formed.

4.3 General discussion of TFs function

N6Ecd showed the highest diversity of TF-types and highest relative abundance of them, especially when compared with N5Ecd. There is just one TF-type (the PF09271-LAG1-DNAbind) detected in N5Ecd and not in N6Ecd while 8 types in N6Ecd are not present in N5Ecd. Since in N6Ecd the new tergite is being constructed, – also happening in N5Ecd – plus a tergal gland, it makes sense that the same TFs families that in N5Ecd plus some others are needed, as well as more amount of TFs.

In the other side, N6D1T shows more abundance of TFs, but less types than N6D1C. N6D1C had 5 TF-types that were not found in N6D1T, while N6D1T had 3 TF-types that were not present in N6D1C. The data suggest that JH activates TFs repressors of the adult program, and these repressing TFs are different types of TFs than the ones that are constructing the TG. There are 3 types of TF-pfam-motifs that we only found it in the samples where the TG is being formed.

Instead of continuing analyzing factors and groups of factors independently, in order to go an step further in the understanding of how TFs are regulating this complex process, we will represent the transcriptional regulatory networks in the next chapter.

Chapter 5

Gene Regulatory Networks

Most biological characteristics arise from complex interactions between the cell's numerous constituents, such as proteins, DNA, RNA and small regulatory molecules. That is why reductionism failed to explain big transitions in evolution. Last discoveries indicate that cellular networks are governed by universal laws and offer a new conceptual framework that could potentially revolutionize our view of biology in the twenty-first century (Barabási and Oltvai, 2004).

One of these cellular networks are the gene regulatory networks (GRNs). They represent the relationship between transcription factors and their target genes. They can be represented as directed graphs where the vertices are transcription factors and genes, and the relations between them are the edges. These regulatory networks usually display a scale-free topology where the degree distribution usually follows a power law (Barabási, 1999) or an exponential function (Guelzim et al., 2002). Transcriptional regulatory networks are usually producing regulatory hubs as well as other substructures such as motifs and modules (Babu et al., 2004).

The *B. germanica* tergal gland GRNs reconstructions, were another attempt to understand the metamorphic changes and how are they regulated during the metamorphosis.

5.1 *D. melanogaster* genes homology

Since we do not have information about which TF regulates which genes in *B. germanica*, neither a reference genome available that could allow us to do other experiments such

Chip-seq, we based the networks on homology with *D. melanogaster*.

The first step was to annotate all the reference transcriptome sequences with "*D. melanogaster* FlyBase accession codes" through a BlastX query against all the CDS sequences from FlyBase (St Pierre et al., 2014).

5.2 Transcription factor interactions

From the public database DroID (Yu et al., 2008) all the transcription factors and their relations with other genes (database: TF⟺Genes) in *D. melanogaster*, was downloaded. The downloaded database consists in a tab delimited text file with pairwise relation of genes. Each line have several information about 2 genes, the first gene is the TF and the second is the target. This order is important since it is what defines the direction of the edge.

5.3 Network Reconstruction

After obtaining the *D. melanogaster* homologue genes and the TF-gene information, we used the Bioconductor package Igraph (Csardi and Nepusz, 2006) in order to depict the regulatory networks. We obtained a regulatory network for each one of the four transcriptomes by converting all the genes present in the given transcriptome in to vertices. After it, using the information from DroID database we established their relations as directed edges. The four reconstructed networks were plotted for visualization purpose as graphs (Appendix E).

5.4 Network topology analysis

A number of studies have revealed that gene regulatory networks have many interesting structural and mutational features such as their scale-free topology, mutational robustness and evolvability. However, how these features have emerged from evolution remains unknown (Tsuda and Kawata, 2010).

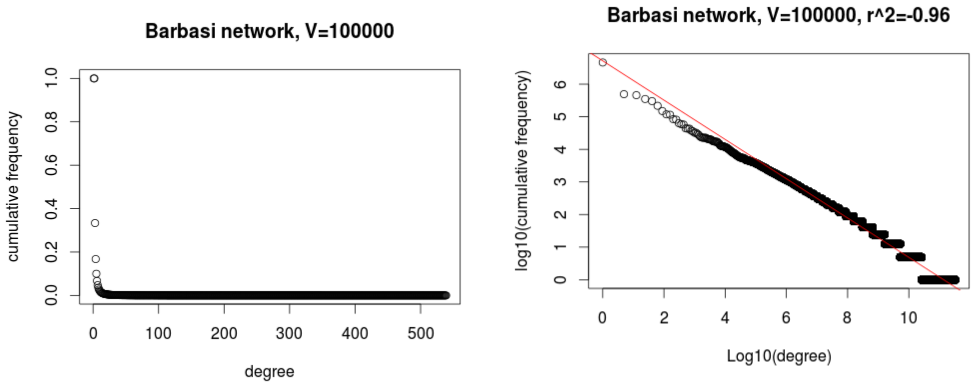
We checked if our reconstructed networks are really following a free scale-free topology as a quality control of the reconstructions.

5.4.1 Scale-free topology

In order to examine whether approximate scale-free topology is satisfied, Zhang and Horvath (2005) propose to use the square of correlation between *log(p(k))* and *log(k)* from the power law function (Equation 5.1).

$$p(k) \sim K^{-\gamma} \tag{5.1}$$

We tested whether Zhang and Horvath approach works appropriately by first generating a scale-free graph with the Barabasi-Albert model (Barabási, 1999) with 100,000 vertices. Then, plotting the degree distribution (Figure 5.1a) of edges that shows the typical long tail of power low function. Finally applying the log10 in both axis (Figure 5.1b) and calculating the linear regression coefficient $R^2 = -0.961$.

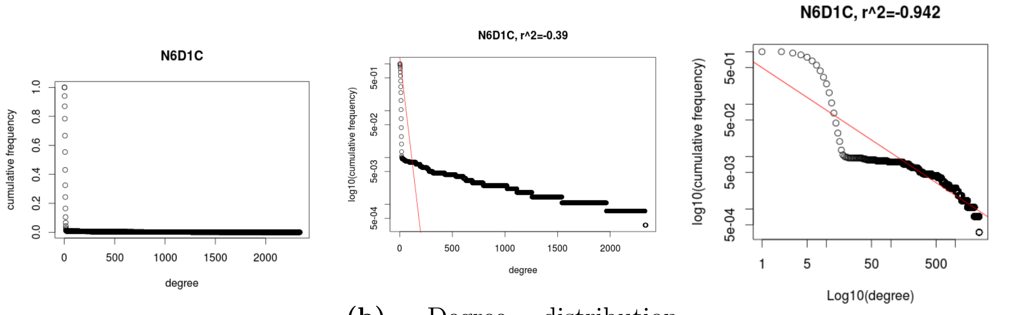


(a) Barabasi degree distribution plot.

(b) Barabasi degree distribution log-log plot.

Figure 5.1: Degree distribution plots from the ideal scale-free Barbasi network.

After showing that it works good for an ideal scale-free topology network, we applied the same procedure to our networks. When we plot the degree distribution frequencies (Figure 5.2a) the plot presents, like in the previous case, the long tail of the power law function. When logarithms were applied (Figure 5.2c) the linear regression line does not seem to really fit the data, however the correlation coefficient is acceptable $R^2 = -0.942$. The semi-log plot was also performed in order to check if better fits an exponential distribution, also common in this kind of networks (Guelzim et al., 2002), rather than a power law (Figure 5.2b). (In this paragraph we only show the results for the N6D1C network as representative since for the other 3 networks the results are mostly the same).



(a) Degree distribution plot.

(b) Degree distribution semi-log plot.

(c) Degree distribution log-log plot.

Figure 5.2: Example plots from the N6D1C network. The four networks follow the same pattern.

Zhang and Horvath (2005) pointed out that many times real data, only satisfy the scale-free topology approximately and can show an “exponentially truncated power law” . In this case, seems that we have 2 differentiated group of nodes, probably due to the differences of connectivity between transcription factors and final targets. The nodes with less connectivity, the final targets, seems to better fit an exponential function. On the other hand, in the case of nodes with high connectivity, usually transcription factors, seems to fit better the power law.

The fact that our networks are following a “exponentially truncated power law“ and clearly not a random network – random network degree follows a Poisson distribution – indicates that the reconstruction may have some biological sense.

5.5 Networks comparison

After assessing the topology of the four networks, which in all cases was the same, we checked for differences among networks. The main objective of this point was to find the differences in the networks that could explain the biological differences of the samples, that is the same that finding how the GRN of T7-T8 change in order to construct a TG.

5.5.1 N6Ecd vs N5Ecd

The networks were plotted as graphs in the Appendix E (Figure E.1 and Figure E.2 for N6Ecd and N5Ecd graphs respectively). Due to the size of the graph it is difficult to appreciate differences when are plotted in a A4 paper, that is why we obtained their numerical descriptors (Table 5.1).

Table 5.1: Numerical description of N6Ecd and N5Ecd GRNs.

	N6Ecd	**N5Ecd**
# Vertices	2,788	1,757
# Edges	17,249	10,547
Density	0.002219	0.003418
Transitivity	0.01228	0.01586

As is shown in the Table 5.1, the GRN of N6Ecd appeared to be much larger, in number of vertices as well as in number of edges, than the N5Ecd. However, the network density (ratio of the number of edges and the number of possible edges) is larger in N5Ecd than in N6Ecd as well as the transitivity (probability that the adjacent vertices of a vertex are connected. Sometimes also called the clustering coefficient).

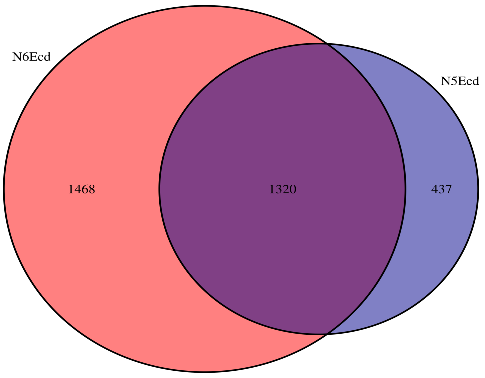


Figure 5.3: Vertices overlapping between N6Ecd and n5Ecd graphs.

Most of the N5Ecd vertices are also present in the N6Ecd, but most of the N6Ecd vertices are not in the N5Ecd network (Figure 5.3).

The comparison of these two networks pointed out that the N6Ecd GRN contains almost the whole N5Ecd GRN. Furthermore, the N6Ecd GRN is much larger since contains an extra set of genes that are constructing the TG. These TG formation genes seem to be less interconnected than the others (N6Ecd has lower density and transitivity Table 5.1). This decay of the density and transitivity probably happens due to the fact that the networks not only include TFs but also all the final products of the regulatory cascade. These highly diverse final products that were not present in N5Ecd are structural and functional constituents of the TG and are connected by just a few edges, basically are only connected by the "input edges" coming from the TFs that are activate their expression.

The most interesting genes are those with more connections – also known as hubs in graph theory – that are present in N6Ecd and not in N5Ecd. These selected hubs are probably the TFs that are regulating main process of the TG formation, take into account that some of them have more than 2,000 connections. (The list of hubs is not shown because currently work is being done with them).

5.5.2 N6D1C vs N6D1T

In this comparison the main point was to assess the role of the JH in the GRN remodeling. The GRN remodeling can be visualized in the animated Figure 5.4 (animation available in the pdf version with Adobe Reader 9.5 or later). For paper readers or problems reproducing the animation, all the frames are shown in Appendix F.

In this remodeling of the GRN by the JH effect, the most interesting genes where those that disappears when JH hormone is present, since are the ones that are triggering the TG formation. From the 942 genes that are disappearing we ordered them in decreasing number of connections and the top genes in the list – the most connected ones are probably important regulatory hubs – were selected for further functional experiments (Data not shown, other members of the laboratory are currently working with them).

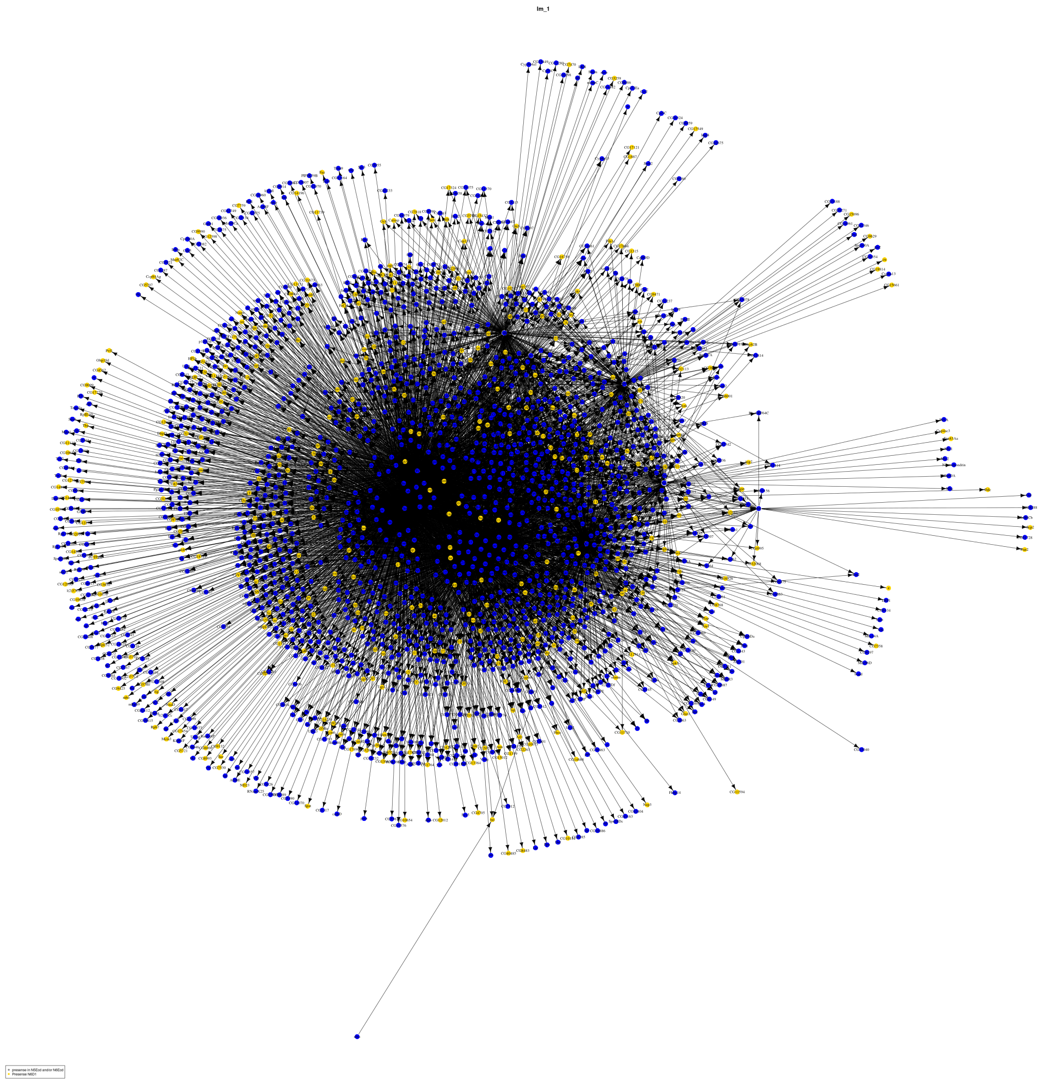


Figure 5.4: Effect of JH in the GRN remodeling (animation available in the pdf version with Adobe Reader 9.5 or later.) First the GRN with the genes present in N6D1C is shown, the blue vertices represents all the genes that are not specific of nymph 6 day 1 (all the genes also present in N5Ecd and/or N6Ecd) and in yellow the specific genes of N6D1. In the second frame, the genes that are disappearing when JH is injected (genes not present in N6D1T) are colored in red. The next frame remove all edges connecting red vertices and finally all the red vertices are removed showing how the network had simplified.

Chapter 6

Conclusions

6.1 Working without a reference genome

Although it is difficult to work with RNA-seq when no reference genome is available, thanks of the 454 technology and the high computational power of the IBE's informatic cluster, we could build a reference transcriptome that has been really useful for our analysis.

6.2 Differential expression analysis without replicates

The DE analysis without replicates did not allow us to confirm that a particular gene was really differentially expressed in a sample respect to another. Nevertheless, the DE analysis plus the GO-enrichment gave us a crucial clue showing that TFs may be playing a key role in the TG formation process.

6.3 Functional annotation

The GO-terms were useful to obtain a general idea about which kind of function are over-represented in a subset of sequences, however in order to asses more specifically the function of a single sequence the PFAM motifs search gave us the most reliable result.

Moreover, the sequence descriptors (the fasta header) of the homologous can be in some cases useful in order to have a first idea about the function or the name of a given sequence. However, due to the big discordance giving gene names in different organisms and the non-controlled language of this field, is not recommended to work with it for functional purposes.

6.4 Database and web interface

In order to properly work with such a huge amount of sequences, different nomenclatures and other data, without losing information it is important to have a good organized and powerful database. In this case a good defined relational database on a MySQL server was really useful for organizing all the data produced along this work.

The creation of an easy to access web interface linked to the MySQL server was really welcomed for all the biologists of the laboratory.

Furthermore, the fast to run and automatized pipeline described in chapter 2 for the database annotation, was especially useful to periodically update the annotation. The periodic updates allows to have the annotation every time more curated since more information is added to public databases.

6.5 The role of transcription factors in tergal glands formation

Our analysis points out that TFs are playing a crucial role in the regulation of the TG. Is not just a fact of the amount of them, but also about the different types of TF that are being expressed.

Actually, after reconstructing the GRNs was evident that N6Ecd network was indeed the N5Ecd network that dramatically grow thanks of the expression of new TFs types.

6.6 Gene regulatory networks

Although the GRNs reconstructions were based on *Drosophila*, they gave us a better idea about the complexity of the morphogenesis process.

Genes that appear as hubs in the networks where the TG is being formed are very interesting since are probably the TFs with main roles in the TG formation. The lists of these genes are not available here since other members of the laboratory are currently working with these genes from a functional point of view.

6.7 Beyond this work

This work allowed to retrieve a list of candidate genes that could be acting as the main regulators of the TG formation. The function of these candidate genes must be experimentally validated as well as their precise mechanisms of action.

This work showed that with the expression of new transcription factors a new organ can be formed from scratch. This let us to hypothesize that the transcription factors may have played an important role in the evolutionary transition from hemimetabola to holometabola, since in this last type of metamorphosis most of the body plan must be build in a single stage transition.

Appendix A

PFAM Motifs related with TFs

Table A.1: Manually accurate list of DNA-binding domains (BDB) and their PFAM-number that are directly related with transcription factor activity. Table from de Mendoza et al. (2013)

DBD	PF number
AP2	PF00847
ARID	PF01388
B3	PF02362
bZIP_1	PF00170
bZIP_2	PF07716
bZIP_Maf	PF03131
CBFB_NFYA	PF02045
CG-1	PF03859
Copper_first	PF00649
CP2	PF04516
CSD	PF00313
CUT	PF02376
Dict-STAT-coil	PF09267
DM	PF00751
E2F_TDP	PF02319
EIN3	PF04873
Ets	PF00178
FLO_LFY	PF01698
Fork_head	PF00250
Fungal_trans	PF04082
Fungal_trans_2	PF11951
GATA	PF00320
GCM	PF03615
GCR1_C	PF12550
GRAS	PF03514
HLH	PF00010
HMG_box	PF00505
Homeobox	PF00046
Homeobox_KN	PF05920
HSF_DNA-bind	PF00447
HTH_psq	PF05225
IRF	PF00605
LAG1-DNAbind	PF09271
MADF_DNA_bdg	PF10545
MAT_Alpha1	PF04769
MH1	PF03165
Myb_DNA-binding	PF00249
NAM	PF02365
NDT80_PhoG	PF05224
P53	PF00870
PLATZ	PF04640
RFX_DNA_binding	PF02257
RHD	PF00554
Runt	PF00853
S1FA	PF04689
SAND	PF01342
SBP	PF03110
SRF-TF	PF00319
STAT_bind	PF02864
STE-like TF	PF02200
T-box	PF00907
TBP	PF00352
TCP	PF03634
TEA	PF01285
TF_AP-2	PF03299
TSC22	PF01166
Tub	PF01167
Whirly	PF08536
WRKY	PF03106
YABBY	PF04690
YL1 nuclear protein	PF05764
zf-BED	PF02892
zf-C2H2	PF00096
zf-C2HC	PF01530
zf-C4	PF00105
zf-Dof	PF02701
zf-GRF	PF06839
zf-MIZ	PF02891
zf-NF-X1-type	PF01422
zf-TAZ	PF02135
Zn_clus	PF00172

Appendix B

TFs-Motifs abundance

Table B.1: Relative abundances of all motifs related with transcription factor function across the transcriptomes.

TFMotifs	**N5Ecd**	**N6Ecd**	**N6D1C**	**N6D1T**
PF03131 bZIP_Maf	0,0002894	0,0007115	0,0005528	0,0009985
PF00313 CSD	0,0001447	0,0002287	0,0004054	0,0003495
PF00249 Myb_DNA-binding	0,0006132	0,0007341	0,0002211	0,0005492
PF02892 zf-BED	0,0003979	0,0003049	0,0001843	0,0000999
PF00505 HMG_box	0,0002894	0,0007019	0,0002948	0,0007020
PF04516 CP2	0,0003979	0,0005077	0,0006634	0,0004993
PF00010 HLH	0,0006873	0,0014505	0,0009582	0,0015477
PF07716 bZIP_2	0,0003093	0,0003812	0,0002580	0,0005492
PF03165 MH1	0,0013023	0,0021115	0,0013637	0,0010484
PF00170 bZIP_1	0,0002532	0,0004066	0,0003686	0,0002996
PF00096 zf-C2H2	0,0026122	0,0065222	0,0070748	0,0100092
PF02319 E2F_TDP	0,0001085	0,0000508	0,0002580	0,0002496
PF01166 TSC22	0,0000362	0,0000762	0,0000369	0,0000499
PF00554 RHD	0,0002170	0,0004320	0,0000369	0,0002496
PF00352 TBP	0,0000723	0,0000254	0,0000369	0,0000000
PF01388 ARID	0,0001085	0,0001525	0,0000737	0,0001997
PF05225 HTH_psq	0,0002532	0,0001779	0,0000000	0,0002996
PF00046 Homeobox	0,0005426	0,0008892	0,0007371	0,0010484
PF00250 Fork_head	0,0005426	0,0004317	0,0001843	0,0004993
PF02135 zf-TAZ	0,0002170	0,0001016	0,0000369	0,0000000
PF05920 Homeobox_KN	0,0000723	0,0000254	0,0000000	0,0000000
PF09271 LAG1-DNAbind	0,0000723	0,0000793	0,0000000	0,0003495
PF00105 zf-C4	0,0001085	0,0003304	0,0001106	0,0000499
PF01285 TEA	0,0001809	0,0001525	0,0000000	0,0000000
PF10545 MADF_DNA_bdg	0,0003256	0,0002287	0,0001106	0,0001498
PF00178 Ets	0,0004703	0,0006099	0,0007371	0,0004493
PF06839 zf-GRF	0,0000000	0,0000254	0,0001843	0,0000000
PF02045 CBFB_NFYA	0,0000000	0,0000508	0,0000000	0,0000000
PF00907 T-box	0,0000000	0,0000762	0,0000000	0,0000000
PF02257 RFX_DNA_binding	0,0000000	0,0001016	0,0002580	0,0000000
PF02891 zf-MIZ	0,0000000	0,0000254	0,0000000	0,0000000
PF01530 zf-C2HC	0,0000000	0,0000508	0,0000000	0,0000000
PF02376 CUT	0,0000723	0,0000762	0,0000369	0,0000000
PF03299 TF_AP-2	0,0000000	0,0000254	0,0000000	0,0000000
Summatory	**0,0106972861**	**0,0182563209**	**0,0151830155**	**0,0202470142**

Appendix C

GO-terms enriched in N6Ecd

Table C.1: N6Ecd Enriched GO-terms

GO.ID	Term	p-value
GO:0005509	calcium ion binding	0.0032
GO:0016881	acid-amino acid ligase activity	0.0194
GO:0003676	nucleic acid binding	0.0388
GO:0016879	ligase activity, forming carbon-nitrogen...	0.0414
GO:0043169	cation binding	0.0466
GO:0046872	metal ion binding	0.0466
GO:0003743	translation initiation factor activity	0.0572
GO:0003755	peptidyl-prolyl cis-trans isomerase acti...	0.0572
GO:0004180	carboxypeptidase activity	0.0572
GO:0004435	phosphatidylinositol phospholipase C act...	0.0572
GO:0005544	calcium-dependent phospholipid binding	0.0572
GO:0008135	translation factor activity, nucleic aci...	0.0572
GO:0008158	hedgehog receptor activity	0.0572
GO:0016859	cis-trans isomerase activity	0.0572
GO:0016874	ligase activity	0.0721
GO:0008238	exopeptidase activity	0.0922
GO:0016853	isomerase activity	0.0922
GO:0004842	ubiquitin-protein ligase activity	0.1353
GO:0005326	neurotransmitter transporter activity	0.1444
GO:0005328	neurotransmitter:sodium symporter activi...	0.1444
GO:0015081	sodium ion transmembrane transporter act...	0.1444
GO:0015293	symporter activity	0.1444
GO:0015294	solute:cation symporter activity	0.1444
GO:0015370	solute:sodium symporter activity	0.1444
GO:0034450	ubiquitin-ubiquitin ligase activity	0.1444
GO:0003723	RNA binding	0.1448
GO:0015077	monovalent inorganic cation transmembran...	0.1526
GO:0019787	small conjugating protein ligase activit...	0.1799
GO:0000030	mannosyltransferase activity	0.2396
GO:0003827	alpha-1,3-mannosylglycoprotein 2-beta-N-...	0.2396
GO:0003916	DNA topoisomerase activity	0.2396
GO:0003917	DNA topoisomerase type I activity	0.2396
GO:0003918	DNA topoisomerase type II (ATP-hydrolyzi...	0.2396
GO:0003950	NAD+ ADP-ribosyltransferase activity	0.2396
GO:0004000	adenosine deaminase activity	0.2396
GO:0004177	aminopeptidase activity	0.2396
GO:0004181	metallocarboxypeptidase activity	0.2396

Table C.2: N6Ecd Enriched GO-terms

GO:0004185	serine-type carboxypeptidase activity	0.2396
GO:0005158	insulin receptor binding	0.2396
GO:0005344	oxygen transporter activity	0.2396
GO:0008146	sulfotransferase activity	0.2396
GO:0008375	acetylglucosaminyltransferase activity	0.2396
GO:0016763	transferase activity, transferring pento...	0.2396
GO:0016782	transferase activity, transferring sulfu...	0.2396
GO:0016814	hydrolase activity, acting on carbon-nit...	0.2396
GO:0031625	ubiquitin protein ligase binding	0.2396
GO:0032403	protein complex binding	0.2396
GO:0044389	small conjugating protein ligase binding	0.2396
GO:0061505	DNA topoisomerase II activity	0.2396
GO:0070008	serine-type exopeptidase activity	0.2396
GO:0008565	protein transporter activity	0.2440
GO:0015291	secondary active transmembrane transport...	0.2440
GO:0022890	inorganic cation transmembrane transport...	0.2574
GO:0016791	phosphatase activity	0.2784
GO:0005488	binding	0.2878
GO:0046873	metal ion transmembrane transporter acti...	0.2952
GO:0042578	phosphoric ester hydrolase activity	0.3014
GO:0004888	transmembrane signaling receptor activit...	0.3185
GO:0038023	signaling receptor activity	0.3185
GO:0004872	receptor activity	0.3317
GO:0004386	helicase activity	0.3400
GO:0004620	phospholipase activity	0.3452
GO:0004629	phospholipase C activity	0.3452
GO:0008094	DNA-dependent ATPase activity	0.3452
GO:0016298	lipase activity	0.3452
GO:0016758	transferase activity, transferring hexos...	0.3452
GO:0046914	transition metal ion binding	0.3606
GO:0008236	serine-type peptidase activity	0.3700
GO:0016757	transferase activity, transferring glyco...	0.3700
GO:0017171	serine hydrolase activity	0.3700
GO:0004221	ubiquitin thiolesterase activity	0.3806
GO:0008324	cation transmembrane transporter activit...	0.3806
GO:0036459	ubiquitinyl hydrolase activity	0.3806
GO:0022892	substrate-specific transporter activity	0.3901
GO:0008270	zinc ion binding	0.4031
GO:0003725	double-stranded RNA binding	0.4220
GO:0003993	acid phosphatase activity	0.4220
GO:0003995	acyl-CoA dehydrogenase activity	0.4220
GO:0003997	acyl-CoA oxidase activity	0.4220
GO:0004714	transmembrane receptor protein tyrosine ...	0.4220
GO:0005102	receptor binding	0.4220
GO:0005261	cation channel activity	0.4220
GO:0005262	calcium channel activity	0.4220
GO:0008194	UDP-glycosyltransferase activity	0.4220
GO:0008235	metalloexopeptidase activity	0.4220
GO:0015085	calcium ion transmembrane transporter ac...	0.4220
GO:0016634	oxidoreductase activity, acting on the C...	0.4220
GO:0016849	phosphorus-oxygen lyase activity	0.4220
GO:0017056	structural constituent of nuclear pore	0.4220
GO:0019239	deaminase activity	0.4220
GO:0019899	enzyme binding	0.4220
GO:0051536	iron-sulfur cluster binding	0.4220
GO:0051540	metal cluster binding	0.4220
GO:0072509	divalent inorganic cation transmembrane ...	0.4220
GO:0005506	iron ion binding	0.4415
GO:0016790	thiolester hydrolase activity	0.4421
GO:0016887	ATPase activity	0.4855

Appendix D

GO-terms enriched in N6D1C

Table D.1: N6D1C Enriched GO-terms

GO.ID	Term	p-value
GO.ID	Term	p-value
GO:0005509	calcium ion binding	0.0032
GO:0016881	acid-amino acid ligase activity	0.0194
GO:0003676	nucleic acid binding	0.0388
GO:0016879	ligase activity, forming carbon-nitrogen...	0.0414
GO:0043169	cation binding	0.0466
GO:0046872	metal ion binding	0.0466
GO:0003743	translation initiation factor activity	0.0572
GO:0003755	peptidyl-prolyl cis-trans isomerase acti...	0.0572
GO:0004180	carboxypeptidase activity	0.0572
GO:0004435	phosphatidylinositol phospholipase C act...	0.0572
GO:0005544	calcium-dependent phospholipid binding	0.0572
GO:0008135	translation factor activity, nucleic aci...	0.0572
GO:0008158	hedgehog receptor activity	0.0572
GO:0016859	cis-trans isomerase activity	0.0572
GO:0016874	ligase activity	0.0721
GO:0008238	exopeptidase activity	0.0922
GO:0016853	isomerase activity	0.0922
GO:0004842	ubiquitin-protein ligase activity	0.1353
GO:0005326	neurotransmitter transporter activity	0.1444
GO:0005328	neurotransmitter:sodium symporter activi...	0.1444
GO:0015081	sodium ion transmembrane transporter act...	0.1444
GO:0015293	symporter activity	0.1444
GO:0015294	solute:cation symporter activity	0.1444
GO:0015370	solute:sodium symporter activity	0.1444
GO:0034450	ubiquitin-ubiquitin ligase activity	0.1444
GO:0003723	RNA binding	0.1448
GO:0015077	monovalent inorganic cation transmembran...	0.1526
GO:0019787	small conjugating protein ligase activit...	0.1799
GO:0000030	mannosyltransferase activity	0.2396
GO:0003827	alpha-1,3-mannosylglycoprotein 2-beta-N-...	0.2396
GO:0003916	DNA topoisomerase activity	0.2396
GO:0003917	DNA topoisomerase type I activity	0.2396
GO:0003918	DNA topoisomerase type II (ATP-hydrolyzi...	0.2396
GO:0003950	NAD+ ADP-ribosyltransferase activity	0.2396
GO:0004000	adenosine deaminase activity	0.2396

Table D.2: N6D1C Enriched GO-terms

GO:0004177	aminopeptidase activity	0.2396
GO:0004181	metallocarboxypeptidase activity	0.2396
GO:0004185	serine-type carboxypeptidase activity	0.2396
GO:0005158	insulin receptor binding	0.2396
GO:0005344	oxygen transporter activity	0.2396
GO:0008146	sulfotransferase activity	0.2396
GO:0008375	acetylglucosaminyltransferase activity	0.2396
GO:0016763	transferase activity, transferring pento...	0.2396
GO:0016782	transferase activity, transferring sulfu...	0.2396
GO:0016814	hydrolase activity, acting on carbon-nit...	0.2396
GO:0031625	ubiquitin protein ligase binding	0.2396
GO:0032403	protein complex binding	0.2396
GO:0044389	small conjugating protein ligase binding	0.2396
GO:0061505	DNA topoisomerase II activity	0.2396
GO:0070008	serine-type exopeptidase activity	0.2396
GO:0008565	protein transporter activity	0.2440
GO:0015291	secondary active transmembrane transport...	0.2440
GO:0022890	inorganic cation transmembrane transport...	0.2574
GO:0016791	phosphatase activity	0.2784
GO:0005488	binding	0.2878
GO:0046873	metal ion transmembrane transporter acti...	0.2952
GO:0042578	phosphoric ester hydrolase activity	0.3014
GO:0004888	transmembrane signaling receptor activit...	0.3185
GO:0038023	signaling receptor activity	0.3185
GO:0004872	receptor activity	0.3317
GO:0004386	helicase activity	0.3400
GO:0004620	phospholipase activity	0.3452
GO:0004629	phospholipase C activity	0.3452
GO:0008094	DNA-dependent ATPase activity	0.3452
GO:0016298	lipase activity	0.3452
GO:0016758	transferase activity, transferring hexos...	0.3452
GO:0046914	transition metal ion binding	0.3606
GO:0008236	serine-type peptidase activity	0.3700
GO:0016757	transferase activity, transferring glyco...	0.3700
GO:0017171	serine hydrolase activity	0.3700
GO:0004221	ubiquitin thiolesterase activity	0.3806
GO:0008324	cation transmembrane transporter activit...	0.3806
GO:0036459	ubiquitinyl hydrolase activity	0.3806
GO:0022892	substrate-specific transporter activity	0.3901
GO:0008270	zinc ion binding	0.4031
GO:0003725	double-stranded RNA binding	0.4220
GO:0003993	acid phosphatase activity	0.4220
GO:0003995	acyl-CoA dehydrogenase activity	0.4220
GO:0003997	acyl-CoA oxidase activity	0.4220
GO:0004714	transmembrane receptor protein tyrosine ...	0.4220
GO:0005102	receptor binding	0.4220
GO:0005261	cation channel activity	0.4220
GO:0005262	calcium channel activity	0.4220
GO:0008194	UDP-glycosyltransferase activity	0.4220
GO:0008235	metalloexopeptidase activity	0.4220
GO:0015085	calcium ion transmembrane transporter ac...	0.4220
GO:0016634	oxidoreductase activity, acting on the C...	0.4220
GO:0016849	phosphorus-oxygen lyase activity	0.4220
GO:0017056	structural constituent of nuclear pore	0.4220
GO:0019239	deaminase activity	0.4220
GO:0019899	enzyme binding	0.4220
GO:0051536	iron-sulfur cluster binding	0.4220
GO:0051540	metal cluster binding	0.4220
GO:0072509	divalent inorganic cation transmembrane ...	0.4220
GO:0005506	iron ion binding	0.4415
GO:0016790	thiolester hydrolase activity	0.4421
GO:0016887	ATPase activity	0.4855

Appendix E

Graphs

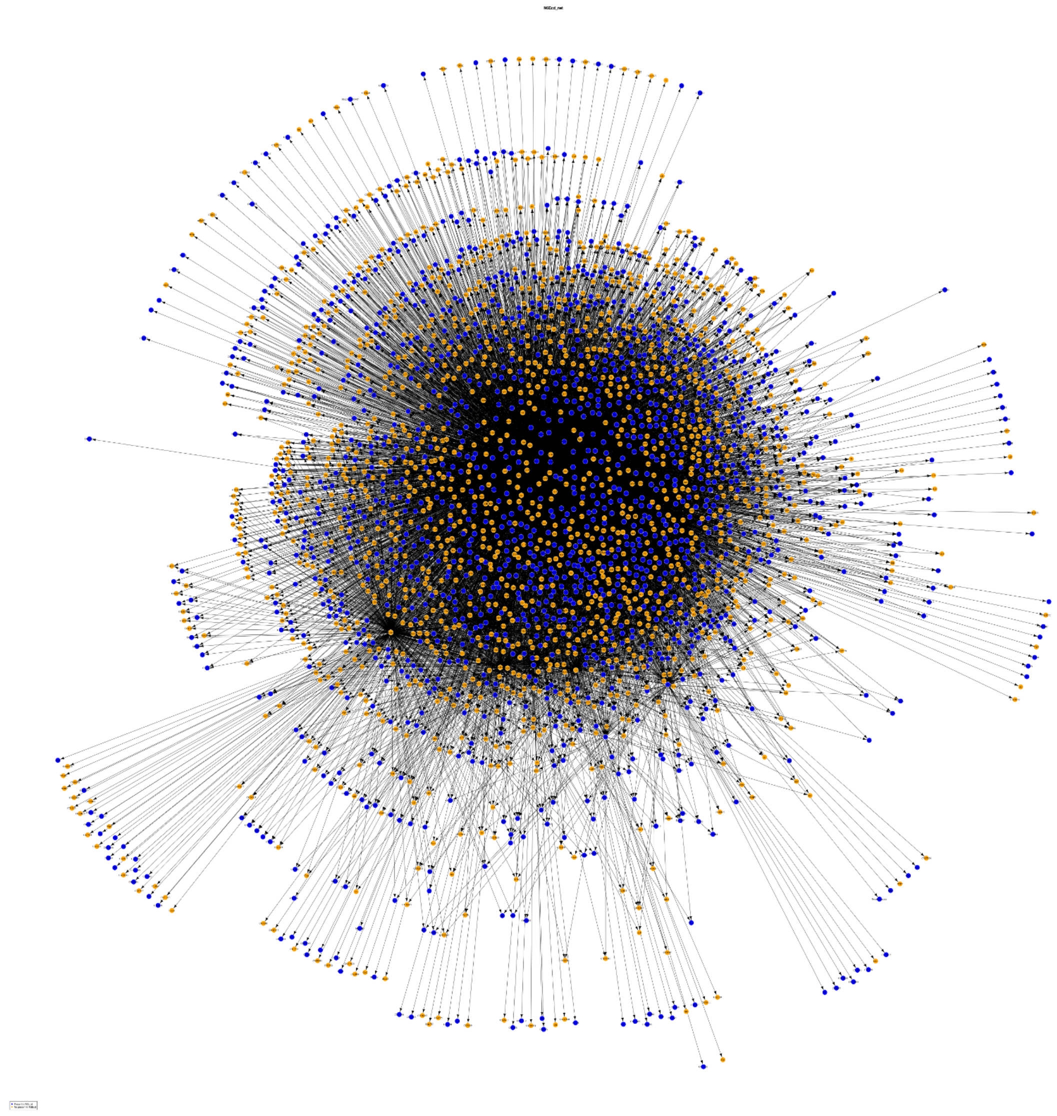


Figure E.1: Graph of all genes present in N6Ecd. Blue vertices are genes also present in N5Ecd.

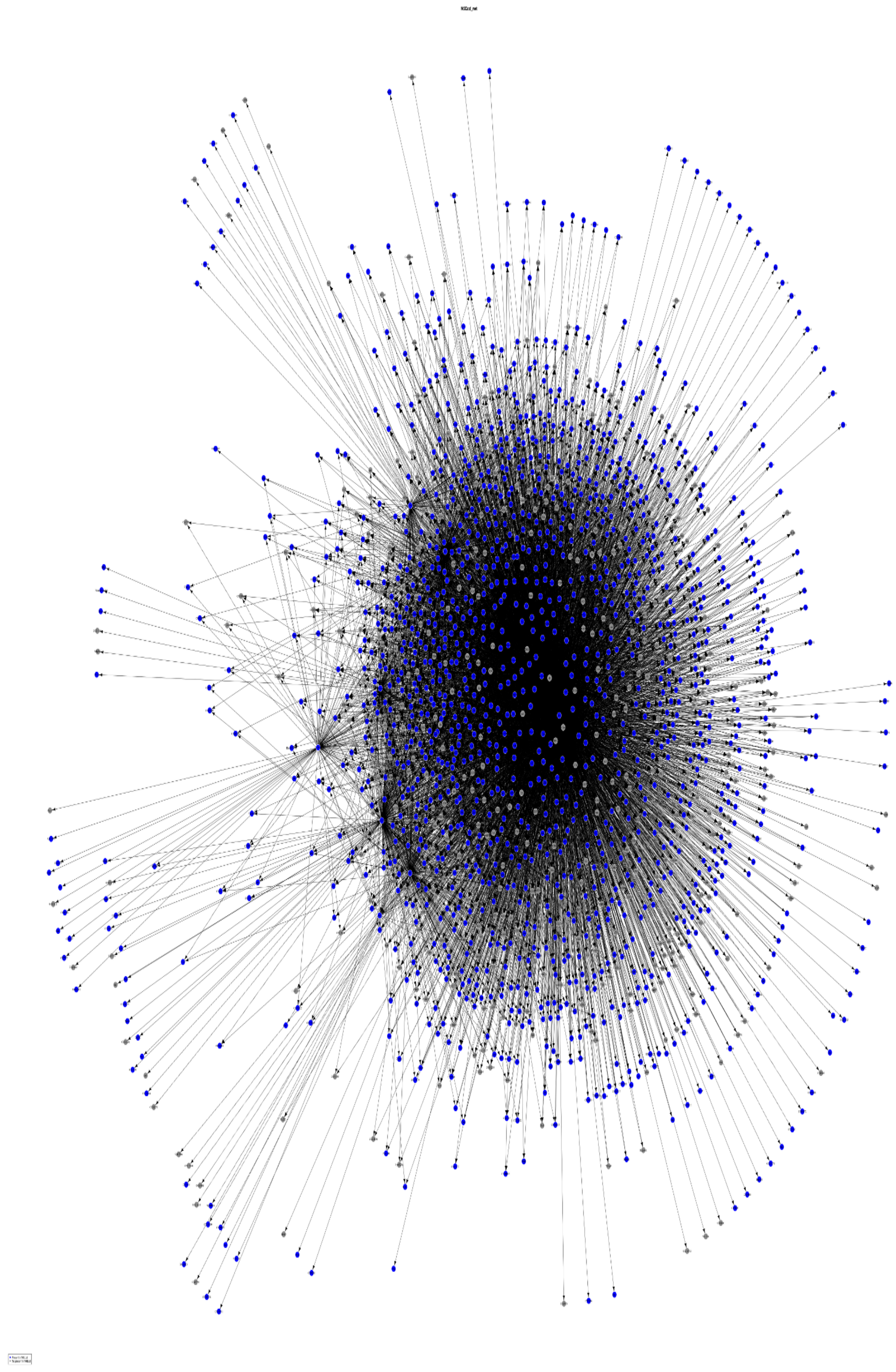


Figure E.2: Graph of all genes present in N5Ecd. Blue vertices are genes also present in N6Ecd.

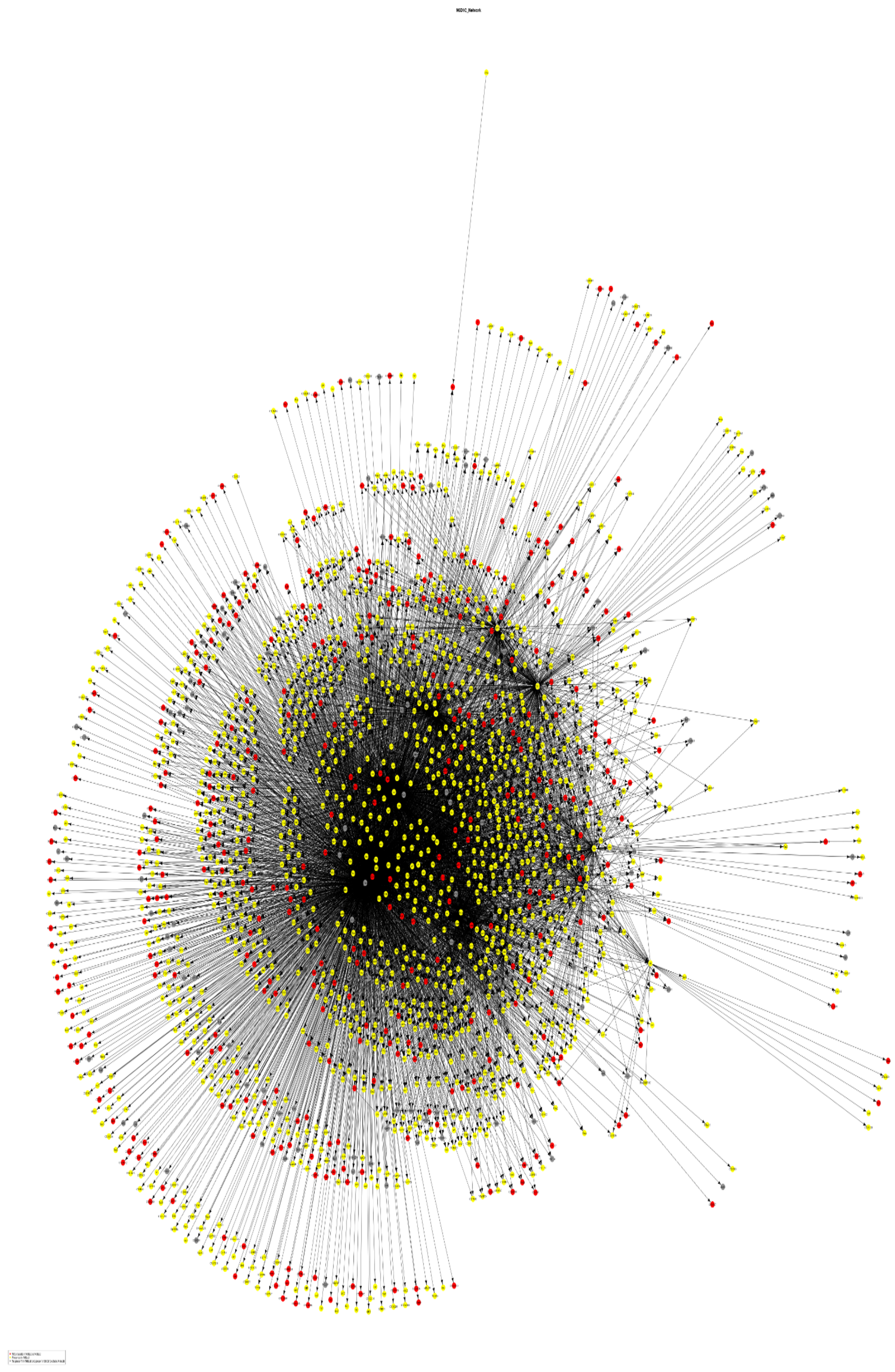


Figure E.3: Graph of all genes present in N6D1C. Yellow vertices are also expressed in N6Ecd. Red vertices are not present in N6Ecd nor N5Ecd. Grey ones are also expressed in N5Ecd but not in N6Ecd.

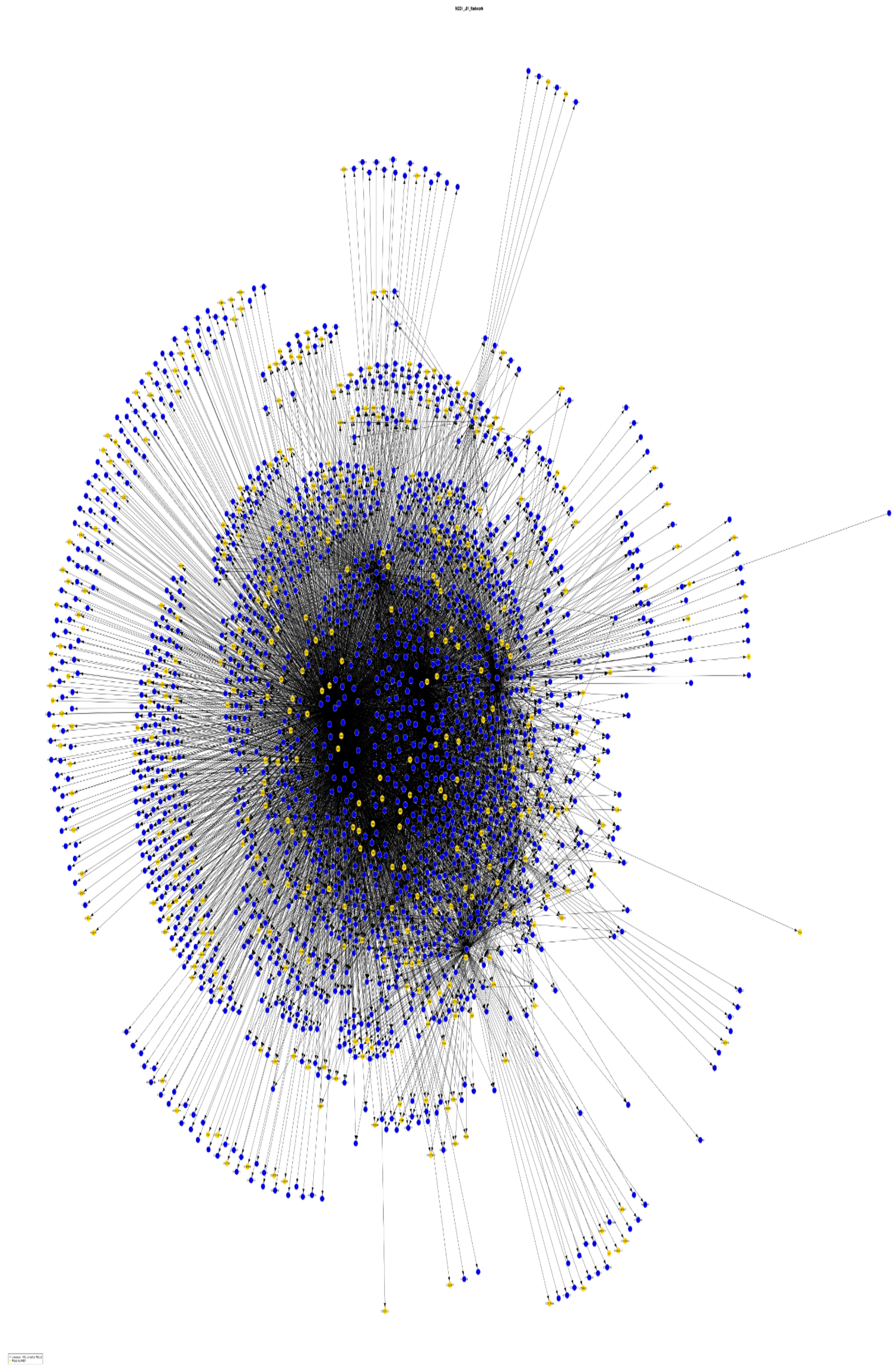


Figure E.4: Graph of all genes present in N6D1T. Yellow vertices are also expressed in N5Ecd and/or N6Ecd.

Appendix F

Annimation Frames

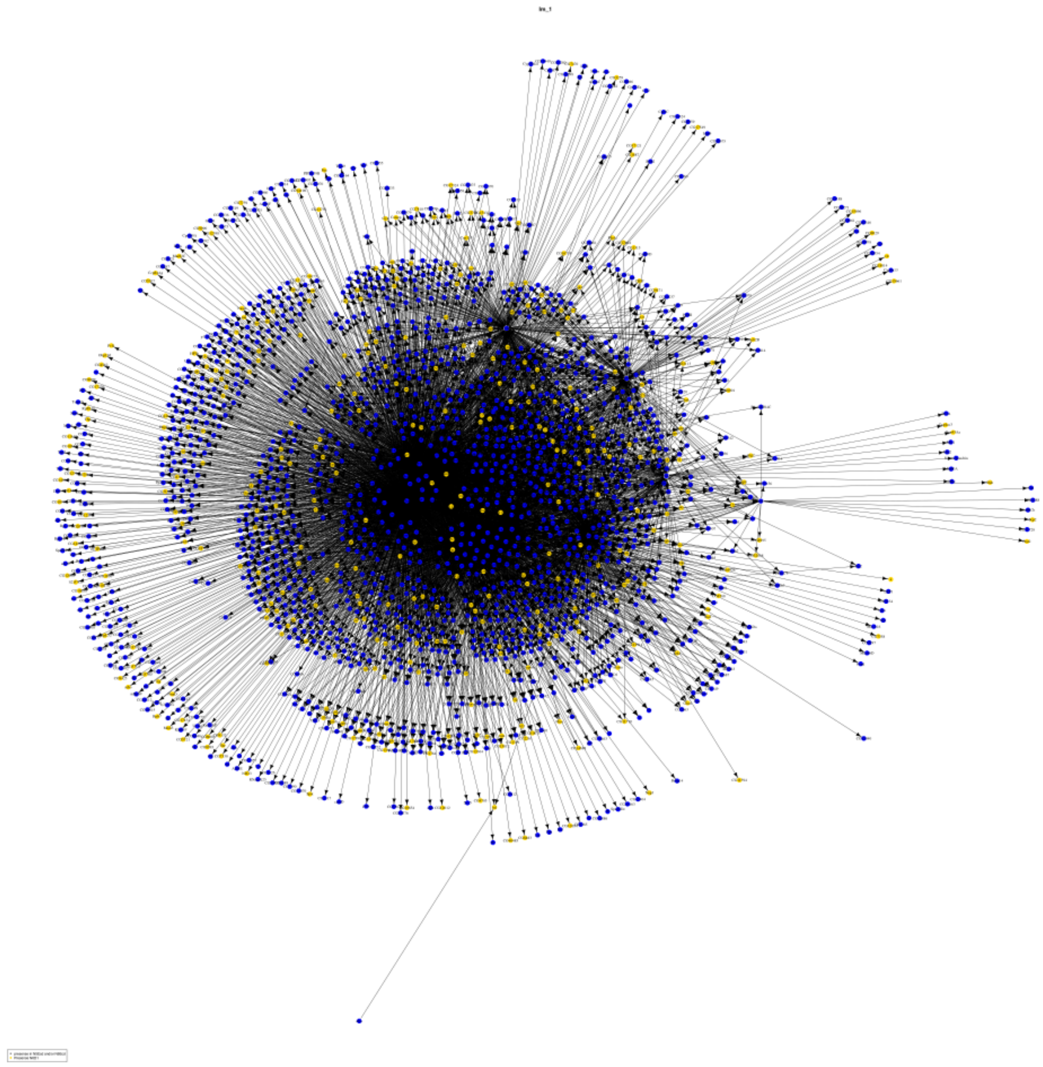


Figure F.1: (Frame 1) First the GRN with the genes present in N6D1C is shown, the blue vertices represents all the genes that are not specific of nymph 6 day 1 (all the genes also present in N5Ecd and/or N6Ecd) and in yellow the specific genes of N6D1

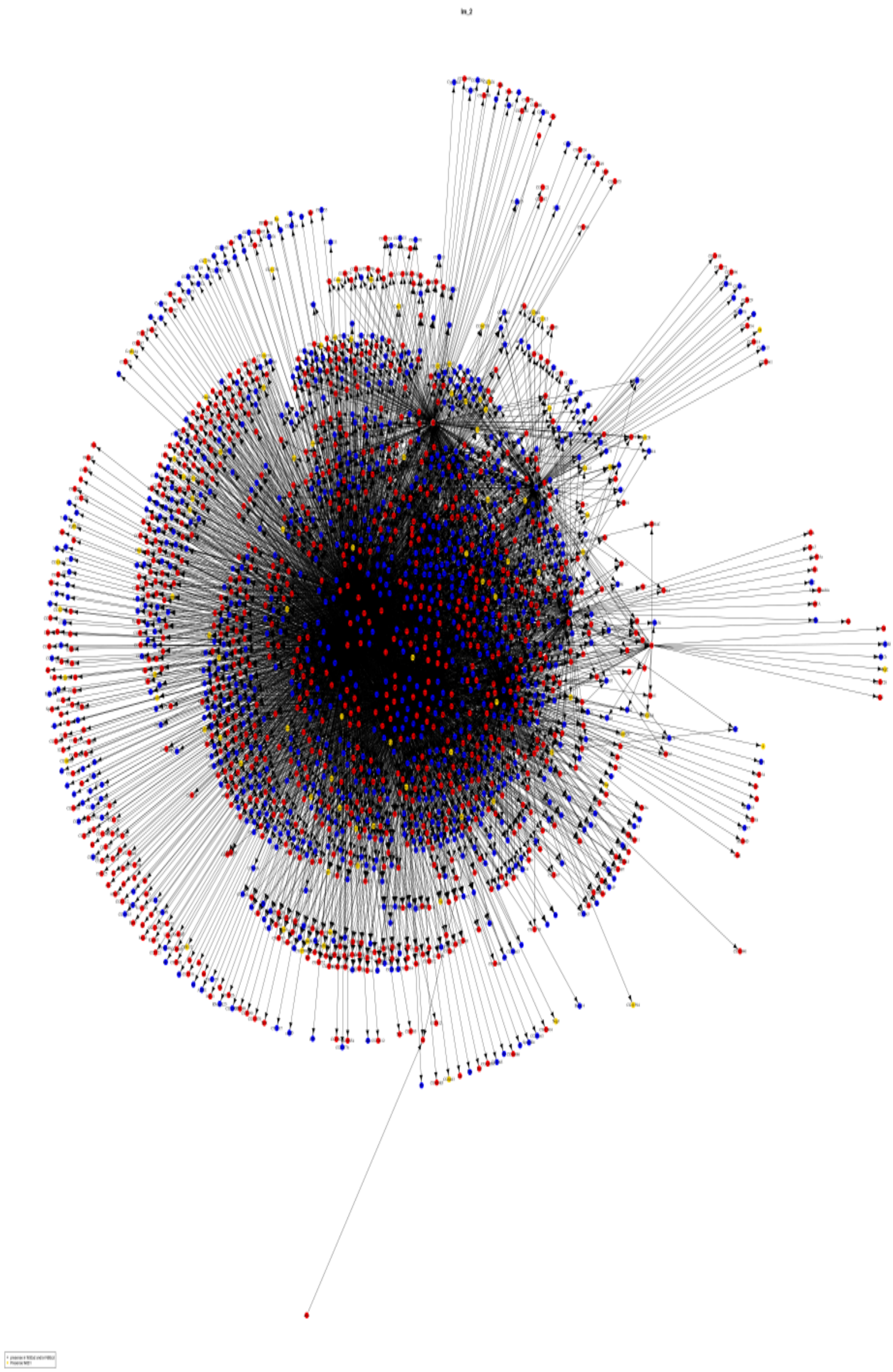


Figure F.2: (Frame 2) The genes that are disappearing when JH is injected (genes not present in N6D1T) are colored in red.

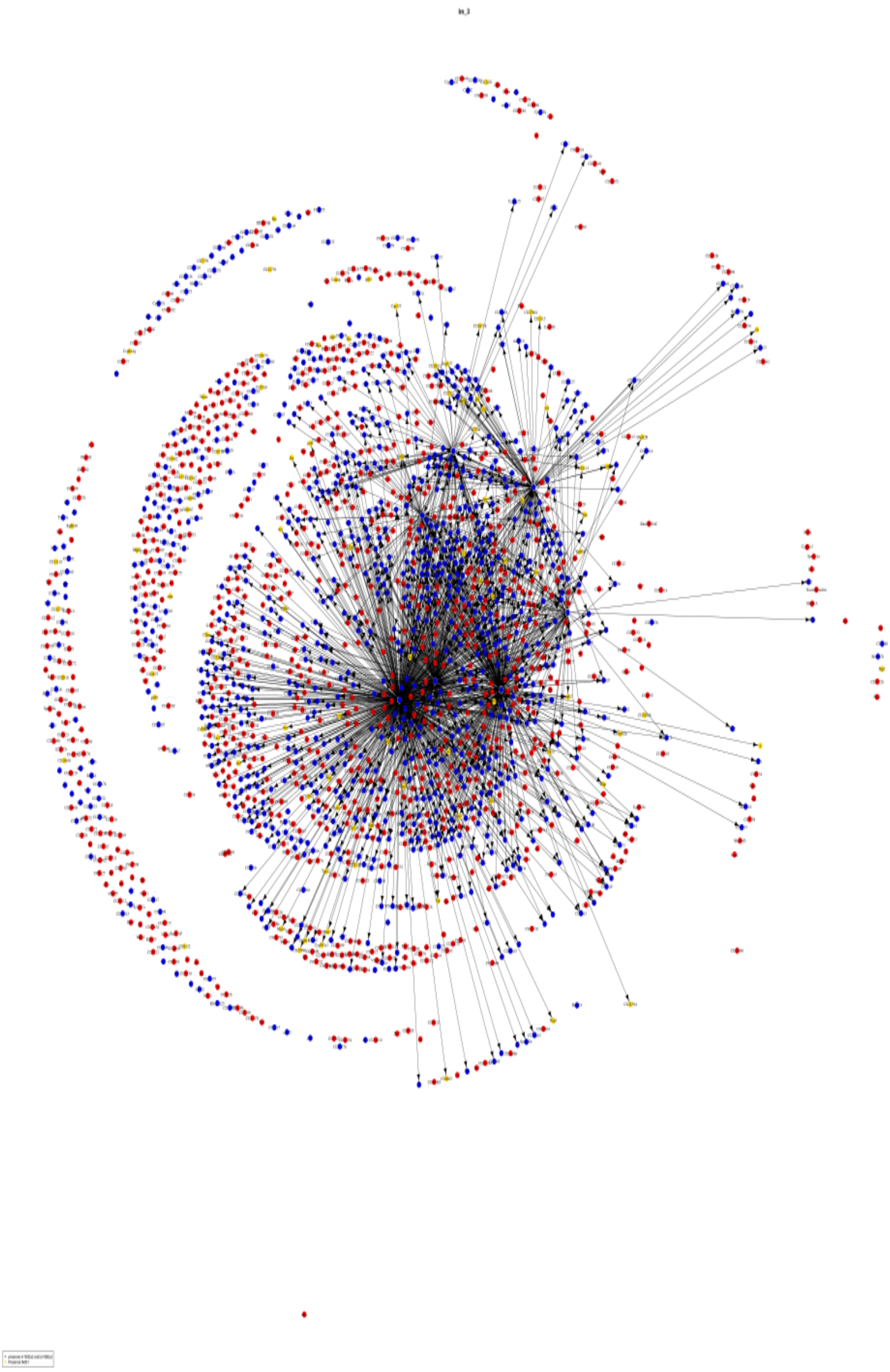


Figure F.3: (Frame 3) All edges connecting red vertices were removed.

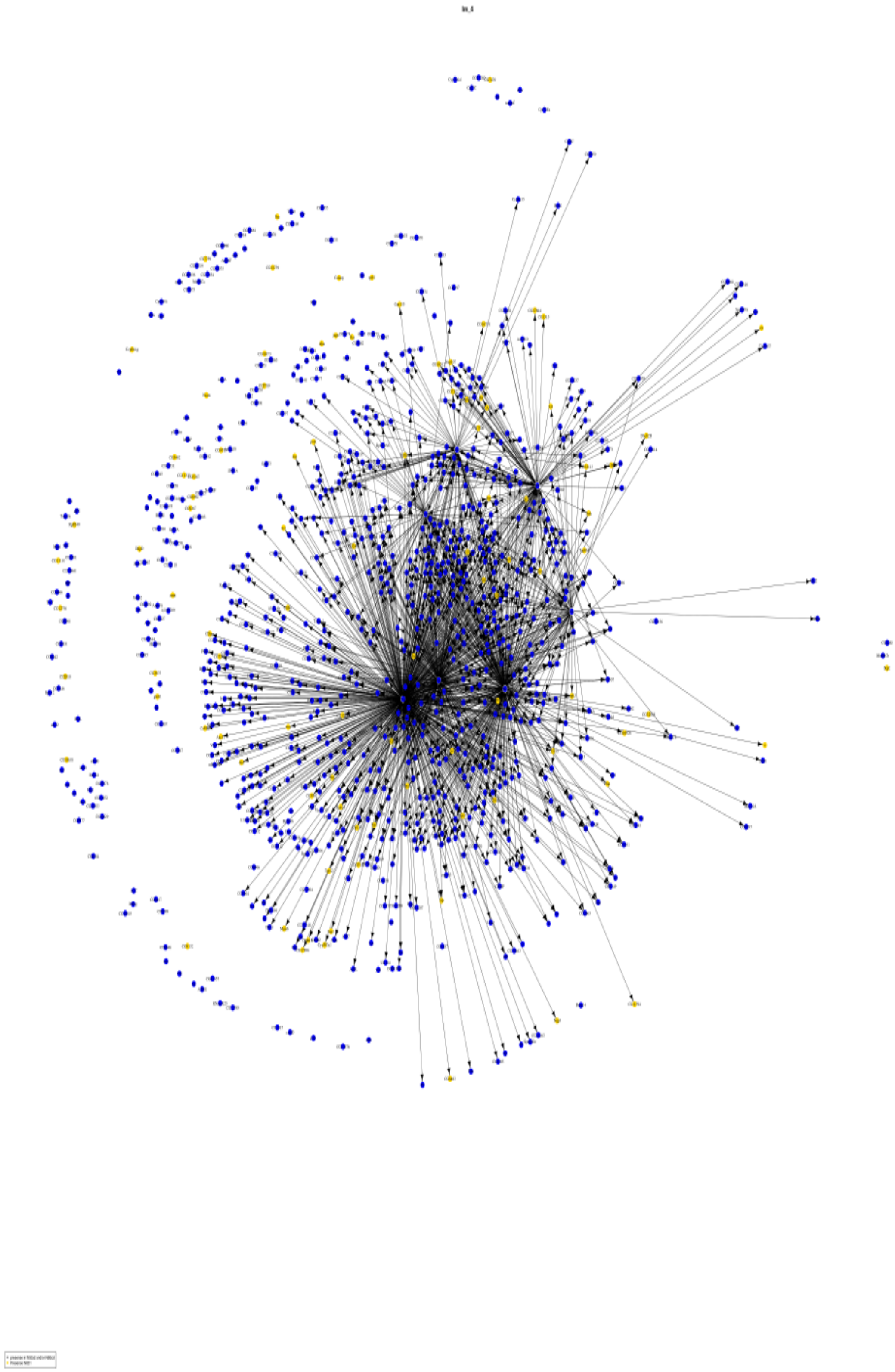


Figure F.4: (Frame 4) All the red vertices were removed showing how the network simplified.

Bibliography

Alexa, A. and Rahnenfuhrer, J. (2010). topGO: enrichment analysis for gene ontology. *R Packag. version 2.8.*

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410.

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., and Robinson, M. D. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.*, 8(9):1765–1786.

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, 14(3):283–91.

Barabási, A. (1999). Emergence of Scaling in Random Networks. *Science (80-.).*, 286(5439):509–512.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, 5(2):101–13.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Res.*, 32(Database issue):D138–41.

Belles, X. (2011). Origin and Evolution of Insect Metamorphosis. *Encycl. Life Sci.*, pages 1–11.

Belles, X. (2013). *la metamorfosis de los insectos.* Consejo Superior de Investigaciones Cientificas, Madrid.

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., and Apweiler, R. (2009). QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, 25(22):3045–3046.

Chen, H. and Boutros, P. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics*, 12(1):35.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, (Complex Systems):1695.

Darwin, C. (1839). *The Voyage of the Beagle.*

de Mendoza, A., Sebé-Pedrós, A., Sestak, M. S., Matejcic, M., Torruella, G., Domazet-Loso, T., and Ruiz-Trillo, I. n. (2013). Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc. Natl. Acad. Sci. U. S. A.*, 110(50):E4858—-66.

Engel, M. S. and Grimaldi, D. A. (2004). New light shed on the oldest insect. *Nature*, 427(6975):627–30.

Ferrer, A., Tarazona, S., and Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. pages 2213–2223.

Gene, T. and Consortium, O. (2000). Gene Ontology: tool for the unification of biology. 25(may):25–29.

Gentleman, R. C., Carey, V. J., Bates, D. M., and Others (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.*, 5:R80.

Grimaldi, D. and Engel, M. S. (2005). *Evolution of the Insects.* Cambridge Evolution Series. Cambridge University Press.

Guelzim, N., Bottani, S., Bourgine, P., and Képès, F. (2002). Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.*, 31(1):60–3.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., LeDuc, R. D., Friedman, N., and Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.*, 8(8):1494–1512.

Hirn, M., Hetru, C., Lagueux, M., and Hoffmann, J. (1979). Prothoracic gland activity and blood titres of ecdysone and ecdysterone during the last larval instar of Locusta migratoria L. *J. Insect Physiol.*, 25(3):255–261.

Huson, D., Richter, D., Mitra, S., Auch, A., and Schuster, S. (2009). Methods for comparative metagenomics. *BMC Bioinformatics*, 10(Suppl 1):S12.

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

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12:323.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., and Marshall, D. (2010). Tablet–next generation sequence assembly visualization. *Bioinformatics*, 26(3):401–2.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5(7):621–8.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.*, 16(6):276–277.

Riddiford, L. M. (2008). Juvenile hormone action: a 2007 perspective. *J. Insect Physiol.*, 54(6):895–901.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet*, 15(2):121–132.

St Pierre, S. E., Ponting, L., Stefancsik, R., and McQuilton, P. (2014). FlyBase 102–advanced approaches to interrogating FlyBase. *Nucleic Acids Res.*, 42(Database issue):D780–8.

Team, R. D. C. (2003). R: a language and environment for statistical computing. *Vienna, Austria R Found. Stat. Comput.*

Tsuda, M. E. and Kawata, M. (2010). Evolution of gene regulatory networks by fluctuating selection and intrinsic constraints. *PLoS Comput. Biol.*, 6(8):22.

Wheeler, W. (2001). The Phylogeny of the Extant Hexapod Orders. *Cladistics*, 17(2):113–169.

Yu, J., Pacifico, S., Liu, G., and Finley, R. L. (2008). DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics*, 9(1):461.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, 4:Article17.