Transcriptome analysis identifies stem cells and immune related genes in the cnidarian *Hydractinia echinata*

Dissertation

submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Jorge Soza Ried

2009

Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Presented by
Jorge Soza Ried
Born in Santiago de Chile
Oral examination*:*

Transcriptome analysis identifies stem cells and immune related genes in the cnidarian *Hydractinia echinata*

Referees: Prof. Dr. Werner A. Müller

PD. Dr. Stefan Wiemann

Acknowledgments

I wish to express my greatest gratitude to Dr. Jörg Hoheisel for giving me the opportunity to be part of his Functional Genome Analysis group. I am deeply thankful for his inspiration and guidance, for the scientific discussions and especially for his friendship, patience and support during difficult moments.

I am sincerely thankful to Prof. Dr. Marcus Frohme for his good advice and guidance, tolerance and continuous help. Thanks for being there not only as a supervisor but also as a friend and to encourage me to realize this interesting project.

I would like to acknowledge Dr. Uri Frank for his unconditionally support, scientific guidance and discussion. Thanks for being such a helpful supervisor and for all the suggestions which substantially improved my work.

I wish to thank Professor Dr. Werner A. Müller for his explanation of *Hydractinia*'s fascinating world, for helping me in my mitomycin experiments, for his continuous constructive remarks and for his time to supervise and evaluate this thesis.

I am also grateful to Dr. Stefan Wiemann, for his help and valuable critical comments which certainly improved this work. Thanks for the time that you spent to read and evaluate this thesis.

I am tempted to individually thank all my friends and close colleagues Achim Stephan, Sarah Schreiber, Sarah Engelhart, Linda Linke, Yasser Riazalhosseini, Amin Moghaddasi, Ole Brandt, Brahim Mali, Rafael Queiroz, Gustave Simo, Christian Busold, Achim Friedrich, Andrea Bauer, Marc Dauber, Anette Jacob, Christoph Schroeder, Michaela Schanne, Sandeep Botla, Mahmoud Youns, Marita Schrenk, Kurt Fellenberg, Karl-Heinz Glatting, Coral del Val, Agnes Hotz-Wagenblatt and all the Functional Genome Analysis group. I thank all of you for being a great support during these years and a great team to work with. Thanks for your help in experiments, protocols, discussions and for all the great moments that we shared. I would also like to thank my dear friends James Robeson, Franz Schmitting, Jean Grisouard, Tewfik Miloud, Caroline Ronzaud, Claudio Diema, Otto Mannherz and Zoran Popovic for your friendship and for being always there with an open hand and a smile.

I wish to thank Dr. Hans Bode and the team of the University of Washington for carrying out the sequencing of the ESTs.

I would like to thank my brothers Cristobal and Cristian, and my sister in law Tati, for their friendship, sympathy, support and unconditionally love. I would like to dedicate this thesis to my mother and father. I miss you so much, especially my father who couldn't see the end of this process but he always believed in me. I admire the strength of my Mom, who despite of the distance always supported and understood me. You were excellent parents and I hope that I will be such a good parent to my daughter.

Finally, I would like to thank my wife Eva and Nora, my little daughter. Eva, there are no words to thank you for being like you are, for your support, encouragement, beauty and love. Nora, thanks for your sweet look, your smile and the little "middle-night" cries, but most to encourage me even more to improve myself.

Contents

Abbreviations

°C	Celsius grad
µ~	Micro
$A_{260\ nm}$	Absorbance at 260 nm
ASW	artificial seawater
AU	Arbitrary units
bHLH	Basic-helix-loop-helix
BLAST	Basic Local Alignment Search Tool
BMP	Bone morphogenetic protein
bp	Base pairs
BRCA2	Breast Cancer type 2
BSA	Bovin Serum Albumin
Bzip	Basic leucine zipper
CA	Correspondence Analysis
CDK	Cyclin-dependent kinase
cDNA	Complementary DNA
CDR	Cysteine rich domain
cm	Centimetre
CTRN	Cnidarian tachylectin-related gene in neurons
dbEST	ESTs databases
DDBJ	DNA Databank of Japan
ddH_2O	Double distilled water
DEPC	Diethylpyrocarbonate
DNA	Deoxyribonucleic acid
dNTP	2'-deoxyribonucleoside-5'-triphosphate
ds -DNA or -RNA	double stranded -DNA or -RNA
e.g.	Exempli gratia (for example)
EG	Embryonic germ cells
EGF	Epidermal growth factor
EMBL	European Molecular Biology Laboratory
ES	Embryonic stem cells
EST	Expressed Sequence Tag
EWS	Ewing's Sarcoma
Ezh2	Enhancer of zeste 2
FAS	Fragment Assembly System
FDR	False discovery rate
FGF	Fibroblast growth factor
FMR	FM (colony) recovered from MMC treatment
FOM	Figure of Merit
g	Gram
GCG	Genetics Computer Group
GO	Gene Ontology
h	Hour

HCL	Hierarchical clustering
HSP70	Heat shock protein 70
HUSAR	Heidelberg Unix Sequence Analysis Resource
HyEED	Embryonic ectoderm development *Hydra* homologue
i.e.	*Id est* (that is)
i-cells	Interstitial cells
IEA	Inferred from electronic annotation
INSD	International Nucleotide Sequence Database
IPTG	Isopropyl β-D-1-thiogalactopyranoside
k~	Kilo
KLF4	Kruppel-like factor 4
KMC	K-means clustering
L	Litre
LB	Luria Bertani
LGT	Lateral gene transfer
Log_2	Logartithm for base 2
LORECs	Libraries of random external controls
LPS	Lipopolysaccharides
LRR	Leucine-rich repeat
LTAs	Lipoteichoic acids
M	Molar
m	Meter
m~	Milli~
MAC	Membrane-attack complexes
MACPF	MAC-perforin domain protein
MAPK	Mitogen-activated protein kinase
M-CHiPS	Multi-Conditional Hybridization Intensity Processing Software
MeV	Multiexperiment Viewer
min	Minute
MMC	Mitomycin-C
mRNA	Messenger RNA
MyD88	Myeloid differentiation factor 88
n~	Nano
NCBI	National Center for Biotechnology Information
NF-κB	Nuclear factor kappa B
Oct3/4	Octamer binding transcription factor 3/4
ORF	Open reading frame
PAMP	Pathogen-associated molecular pattern
PCA	Principal Component Analysis
PCR	Polymerase chain reaction
pf	Post fertilization
pi	Post induction
PIR	Protein Identification Resource
PLA2	Phospholipase A2
PMT	Photomultiplier tube

POR	Points of rejection
PRC2	Polycomb Repressive Complex 2
PRR	Pattern recognition receptor
PSD	Protein Sequence Database
RAG-1	Recombination activation gene 1
RBL	Rhamnose-binding lectin
RIN	RNA integrity number (RIN)
RNA	Ribonucleic acid
RNase	Ribonuclease
rpm	Revolutions per minute
rRNA	Ribosomal RNA
Rsp	Rhamnospondin
RT	Room temperature
RT-PCR	Reverse transcription-PCR
SAM	Significance Analysis of Microarray
Sox2	SRY (sex determining region Y)-box 2
spp.	Species (plural form)
SRF	Serum response factor
SRS	Sequence Retrieval System
ss-DNA/RNA	Single stranded -DNA or –RNA
TFF	Trefoil factor
TGF-β	Transforming growth factor-β
TIR	Toll-interleukin-receptor
TLR	Toll-like receptors
TLS	Transcribed in LipoSarcomas
tRNA	Transfer RNA
TSR	Thrombospondin type 1 domain
UV	Ultraviolet
V	Volts
w/v	Weight in volume
Wnt	Wingless
X-Gal	5-Bromo-4-chloro-3-indolyl-β-D-galactopyranoside

Abstract

An increasing amount of Expressed Sequence Tag (EST) and genomic data predominantly for the cnidarians *Acropora*, *Hydra* and *Nematostella*, reveals that despite being one of the morphologically simplest multicellular animals, cnidarians possess a high genomic complexity. In order to contribute towards a broader coverage of this phylum, an EST project was performed to analyze the transcriptome of *Hydractinia echinata*. Moreover, transcriptional profiling experiments were carried out to characterize the i-cell population and the immune system of the hydroid.

In this work a cDNA-library containing about 20,000 clones was constructed, which covers the entire life cycle of the organism and also represents some stress-induced conditions. After randomly sequencing almost 9,000 clones, EST characterization revealed a broad diversity of genes, with higher sequence similarity to vertebrates than to ecdysozoan invertebrates. Furthermore, a significant number of sequences hitherto unknown in metazoans were detected. The identification of unique *Hydractinia* sequences is consistent with the suggested high diversity and complexity of genes within the phylum. To store all the acquired information a database aimed at making the data widely available was created, which is accessible at www.mchips.org/hydractinia_echinata.html.

To further characterize *Hydractinia* genes, a cDNA-microarray was constructed including the already sequenced ESTs as well as PCR-products from almost 5,000 un-sequenced cDNAs. Genes associated with the i-cell lineage were identified by the analysis of the gene expression profile of colonies depleted from their i-cells using mitomycin-C and colonies after the recovery from the treatment. Microarray normalized data ended up with 162 significant differentially expressed genes. Several growth and transcription factors as well as genes associated with undifferentiated cells were identified including; BMPs, Bzip/Mafl and *CnPL10*. In addition, i-cell depleted organisms exhibited an activation of genes involved in detoxification and wound healing activities. These genes are good candidates to define the i-cell population of *Hydractinia*.

Genes associated with the immune system of *Hydractinia* were identified by the analysis of the expression profile of organisms having a LPS mimicked Gram-negative bacterial infection as well as an allogeneic reaction. 245 candidate genes with a significantly different expression level were identified. Genes associated with an LPS response encode for *e.g.* HSP70, lipocalin-like proteins, serine protease inhibitors, proteins with TSR domains and lectins. In the case of allorecognition, a probable whole genome response with up-and down regulation

of hundreds of genes was observed; demonstrating a complex process. Some of the identified genes encode for *e.g.* minicollagens, transcriptional and growth factors, proteins with a protective function against oxygen metabolites or with potent inflammatory and neurotoxicity effects. Gene expression pattern analysis provided insights towards the function of many genes which are still unknown. In the case of genes with a known functional annotation, the microarray experiments either corroborated their characterization or defined an alternative one for *Hydractinia*.

This project is the first high-throughput effort aimed to identify and characterize the transcriptome of the colonial marine hydroid *Hydractinia echinata*. The combination of the EST dataset, database and the microarray, provides a solid platform to promote and facilitate molecular research not only in *Hydractinia* but also in other cnidarians.

Publications associated with the project

- **Soza-Ried, J.,** Hotz-Wagenblatt, A., Glatting, K.H., Del Val. C., Fellenberg, K., Bode, H., Frank, U., Hoheisel, J.D. & Frohme, M. (2009). The transcriptome of the colonial marine hydroid *Hydractinia echinata*. (*in review*)

- Mali, B., **Soza Ried, J.,** Frohme, M. & Frank, U. (2006). Structural but not functional conservation of an immune molecule: a Tachylectin-like gene in *Hydractinia*. *Dev. Comp. Immunol.* **30**, 275-281.

Zusammenfassung

Die zunehmende Menge an EST und genomischen Daten der Cnidarier, allen voran der *Acropora*, *Hydra* und *Nematostella* zeigt, dass Cnidarier trotz ihrer morphologischen Einfachheit eine große genomische Komplexität aufweisen. Um zu einem tieferen Verständnis des Phylums beizutragen, wurde ein EST Projekt realisiert, mit dessen Hilfe das Transkriptom des Hydroid *Hydractinia echinata* analysiert wurde. Darüberhinaus wurden Experimente zur Analyse der Expressionsmuster durchgeführt, um die I-Zell Population und das angeborene Immunsystem des Hydroids zu charakterisieren.

Im Rahmen dieser Arbeit wurde eine cDNA-Blibliothek mit insgesamt über 20.000 Genen erstellt, die den gesamten Lebenszyklus des Organismus abdeckt, und darüber hinaus einige stress-induzierte Konditionen repräsentiert. Die EST-Charakterisierung zeigt eine breite Vielfalt an Genen, wobei die Sequenzen eine größere Ähnlichkeit mit Vertebraten als mit Invertebraten der Ecdysozoen Gruppe aufweisen. Außerdem konnte eine signifikante Anzahl an Genen detektiert werden, die bisher in Metazoen unbekannt waren. Die Identifizierung der *Hydractinia* spezifischen Sequenzen unterstützt die Annahme einer großen Vielfalt und Komplexität der Gene innerhalb dieses Phylums. Zur Speicherung der gewonnen Informationen wurde eine allgemein zugängliche Datenbank angelegt, die unter www.mchips.org/hydractinia_echinata.html verfügbar ist.

Um die *Hydractinia* Sequenzen näher zu untersuchen, wurde ein cDNA-Microarray erstellt, der die bereits sequenzierten ESTs sowie PCR Produkte von fast 5.000 unsequenzierten cDNAs enthält. Die Identifizierung I-Zell assoziierter Gene erfolgte anhand der Genexpressionsprofile. Hierzu wurden die aufgrund der Mitomycin-C Behandlung I-Zell freien Kolonien mit Kolonien verglichen, die sich nach der Behandlung wieder regeneriert hatten. Aus den normalisierten Daten des Microarrays ergaben sich 162 signifikant unterschiedlich exprimierte Gene. Identifiziert wurden mehrere Wachstums- und Transkriptionsfaktoren, sowie Gene, die im Zusammenhang mit undifferenzierten Zellen stehen, einschließlich BMPs, Bzip/Mafl und *CnPL10*. Darüber hinaus zeigten I-Zell freie Organismen eine Aktivierung der in der Entgiftung und Wundheilung vorkommenden Gene. Diese Gene sind gute Kandidaten, um die I-Zell-Population von *Hydractinia* zu definieren.

Zur Identifizierung der mit dem Immunsystem von *Hydractinia* assoziierten Gene, wurde eine Expressionsanalyse an Tieren durchgeführt, bei denen, mit LPS, eine bakteriellen Infektion nachgeahmt wurde und die eine allogene Antwort zeigten. 245 Kandidatengene mit signifikant unterschiedlicher Expression konnten bestimmt werden. Gene, die mit einer LPS-

Antwort in Verbindung stehen, kodieren für z. B. für HSP70, Lipocalin-ähnliche Proteine, Serin-Protease-Inhibitoren, Proteine mit TSR-Domänen und Lektinen. Im Falle der Allo-Erkennung wurde eine wahrscheinlich das ganze Genom umfassende Reaktion mit Hunderten von positiv und negativ regulierten Genen beobachtet, die einen komplexen Prozess vermuten lässt. Einige der identifizierten Gene kodieren bspw. für Minikollagene, Transkriptions-und Wachstumsfaktoren und für Proteine mit Schutzfunktion gegen Sauerstoff-Metaboliten oder mit stark entzündlichen und neurotoxischen Effekten. Die Analyse der Genexpressionsmuster lieferte Erkenntnisse über die Funktion vieler noch unbekannter Gene. Gene mit bereits bekannter Funktion, wurden durch die Microarray-Experimente entweder in ihre Annotation bestätigt, oder es konnte eine Alternativfunktion für *Hydractinia* definiert werden.

Bei dieser Arbeit wurden erstmals Hochdurchsatztechnologien eingesetzt, um das Transkriptom der kolonialen marinen Hydrozoe, *Hydractinia echinata* zu identifizieren und charakterisieren. Die Kombination aus EST-Datensatz, Datenbank und Microarray, liefert eine zuverlässige Plattform um die molekulare Forschung an *Hydractinia*, aber auch an anderen Cnidariern zu fördern und zu vereinfachen.

Im Rahmen dieser Arbeit erschinene Publikationen

- **Soza-Ried, J.,** Hotz-Wagenblatt, A., Glatting, K.H., Del Val. C., Fellenberg, K., Bode, H., Frank, U., Hoheisel, J.D. & Frohme, M. (2009). The transcriptome of the colonial marine hydroid *Hydractinia echinata*. (*in review*)

- Mali, B., **Soza Ried, J.,** Frohme, M. & Frank, U. (2006). Structural but not functional conservation of an immune molecule: a Tachylectin-like gene in *Hydractinia*. *Dev. Comp. Immunol.* **30**, 275-281.

1. Introduction

1.1 Biological aspects

1.1.1. Cnidarians and genomics

1.1.1.1. The phylum Cnidaria

With fossil records dating back to more than 500 million years, the phylum Cnidaria comprises one of the most ancient living multicellular organisms with true animal features [1]. Its phylogenetic position at the base of the Metazoa allows to consider them as a sister group to the Bilateria, predating the protostome and deuterostome divergence (Fig. 1). The phylum is characterized by a simple body plan with two germ layers -an endo and ectodermal epithelial tissue layer separated by an acellular mesoglea-, a nerve net with sensory and ganglionic nerve cells, and the cnidocytes or stinging cells that give the phylum its name [1, 2].

Cnidarians are divided in four different classes, the Anthozoa, Hydrozoa, Scyphozoa and Cubozoa, 99% of them being marine animals [2]. They can live either as simple solitary or colonial tubes equipped with tentacles, called polyps, as in the case of the anthozoans, including *Nematostella* and *Acropora,* and some hydrozoans such as *Hydra* and *Hydractinia*, or have a life cycle characterized by alternating generations of polyps and a more complex form, the medusa (jellyfish), as in most hydrozoans, scyphozoans and cubozoans [2, 3].

All cnidarian members display a high degree of developmental plasticity, presenting both sexual and asexual reproduction (by budding or fission) and also the possibility to regenerate after injury. Even re-aggregation of single cells or small tissue fragments can regenerate a complete organism under laboratory condition [2, 4]. These features support the use of cnidarians as experimental model organisms, which have been used since Abraham Trembley's experiments with *Hydra* in the 18th century in a variety of biological disciplines [2, 5].

1.1.1.2. Sequencing in Cnidaria

Nowadays, new large-scale sequencing capacities are providing access to the genome sequences from a broad variety of model organisms. Most of the organisms have been selected according to their key phylogenetic position on the evolutionary tree [6]. This information allows the comparison of the genomic data from a wide range of organisms, identifying their differences and similarities between each other, and inferring critical clues about the structure, function and evolution of genomes. For example, cnidarians demonstrate to be particular informative for deciphering the gene content of the last common eumetazoan ancestor (Fig. 1) [7-9]

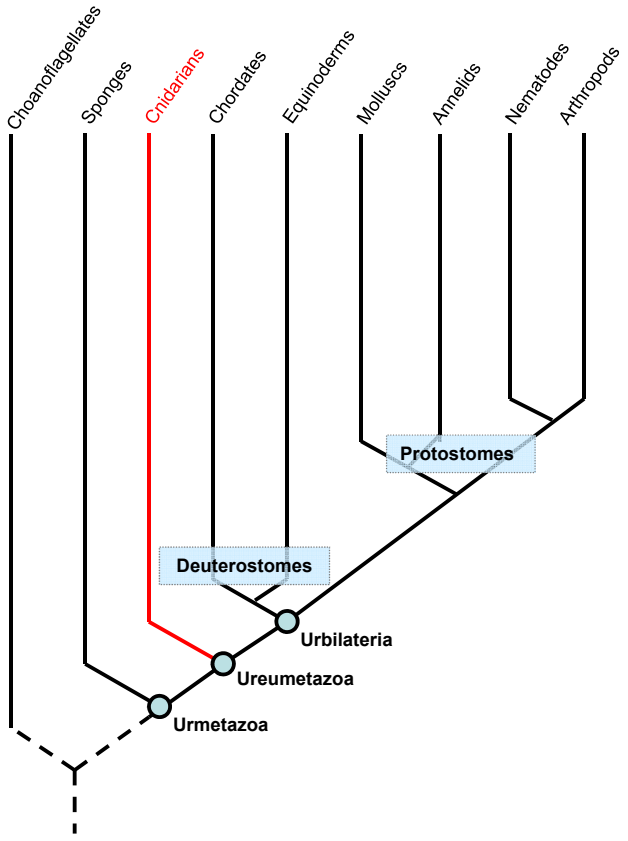


Figure 1 - Phylogenetic representation of the animal kingdom. Phylogenetic analyses suggest a monophyletic origin of the animal kingdom in an organism known as Urmetazoa. Comparisons between deuterostomes and protostomes also suggest a common ancestor of all "higher' animals, the urbilateria. Cnidarians (red) branched off the metazoan stem before the origin of bilaterian. The Ureumetazoa represent the common ancestor of all animals with a tissue grade of organization [10].

First sequencing approaches in cnidarians were made at the transcriptome level, generating thousands of Expressed Sequence Tags (ESTs) [11]. These EST databases are predominantly based on the coral *Acropora millepora*, the freshwater and solitary polyp *Hydra* spp. and the sea anemone *Nematostella vectensis* [8, 12]. At the genomic level, the Joint Genome Institute

(JGI, http://www.jgi.doe.gov) recently released the assembled genome of *Nematostella* [7] and soon will provide the one of *Hydra*.

Despite being regarded as morphologically simple organisms, cnidarian sequencing projects revealed a surprisingly high degree of genetic complexity [7-9, 12, 13]. At first glance, the EST projects on *Acropora*, *Hydra* and *Nematostella* exposed that the genome of cnidarians are likely to contain 25,000 genes, richer than the genomes of the commonly used models *Drosophila* and *Caenorhabditis* [8, 10]. The release of the *Nematostella* genome confirmed that, discarding transposable elements as well as possible pseudogenes and allelic variants, the 2n = 30 chromosome genome contains ~18,000 bona fide genes [7]. Several genes and signalling pathways associated with patterning and developmental processes in bilaterians appeared to be represented in cnidarians, including the components of the wingless (Wnt), transforming growth factor-β (TGF-β) and fibroblast growth factor (FGF) signalling pathways as well as their corresponding secreted ligands and antagonists [7, 10]. Additionally, cnidarian sequences were significantly more similar to vertebrates than to *Drosophila* and *C. elegans* [8]. Indeed, many cnidarian genes presented a vertebrate homologue but were absent from the invertebrate model systems. These confirm the dramatic gene loss observed in the ecdysozoans models, and support that many genes which were considered to be vertebrate innovations were actually present in the last common eumetazoan ancestor [7, 8, 12, 14, 15]. One of the major findings includes the representation in *Nematostella* of all but one of the 12 *Wnt* gene subfamilies known from the chordates genomes. In the case of the invertebrates *Drosophila* and *Caenorhabditis*, only six *Wnt* subfamilies have been identified [9, 10, 14, 16]. Moreover, the genomic organization of cnidarians in terms of intron richness and degree of synteny resembles the one of vertebrates rather than ecdysozoan invertebrates [7, 17].

These genetic complexities allow using cnidarians as an experimental platform for medical research, providing new insights into the genetic and molecular mechanisms underlying human diseases [18]. An example is the cnidarian homologue of the human breast cancer related gene (*BRCA2*). All eight BRC repeats present in the human gene are detectable in *Nematostella*, while the *Drosophila* and *Caenorhabditis* gene only contains three and one complete repeats, respectively [15, 18].

Cnidarians also seem to contain a significant number of protein coding sequences which have not been detected in other animals, indicating that they might be either cnidarian innovations or ancient genes lost in the bilaterian genomes analyzed so far [7, 8].

1.1.2. *Hydractinia echinata* as a model system

1.1.2.1. Cnidarian model organisms

The phylum Cnidaria is a highly diverse group of animals. The available sequencing data suggests a distant relationship between anthozoans and hydrozoans, and is consistent with a high variation in their gene content and gene family diversity [7]. Therefore, for a complete overview of the phylum it is necessary to access more cnidarians genomic data. While anthozoans transcribed genetic data is well represented by the model organisms *Nematostella* and *Acropora*, *Hydra* -as a freshwater solitary polyp- is a poor representative of the class Hydrozoa as most of its members are colonial and marine. Furthermore, *Hydra* sexual reproduction is an erratic and rather unpredictable event, which results in non accessible embryos, limiting developmental biology studies to regeneration experiments [1, 19]. Within the Hydrozoa, the colonial and marine organism *Hydractinia echinata* can be considered as one of the best representative of the phylum. This animal offers attractive features of a good model organism: it is easy to culture and has a short generation time, which enables genetic studies and inbreeding; it reproduces almost daily in highly predictable intervals allowing a continuous access to all different developmental stages; it may easily be subcloned and manipulated in terms of gene expression; and its biology is well studied at the molecular level. Indeed, molecular techniques including transgenic technology are available for *Hydractinia*, which has been a model system to study embryogenesis, metamorphosis, pattern formation and immunity for decades [1, 19-23].

1.1.2.2. The hydroid Hydractinia echinata

The hydroid *Hydractinia echinata* can be found on shallow waters of the northern coast of Europe, frequently growing on the outside of gastropod shells (*e.g. Littorina spp.*, *Buccinum spp.*) inhabited by paguroid hermit crabs (*e.g. Pagurus bernhardus*, *Eupagurus spp.*) [3].

A *Hydractinia* colony is composed of different kinds of polyps connected between each other by a network of gastrovascular canals, the stolons [1]. In a colony growing on a hermit crab shell it is possible to identify four types of polyps; (1) feeding polyps or gastrozooids with upper and lower circles of tentacles; (2) sexual polyps or gonozooids present mainly at the centre or in the dense region of the mat, being the male or female reproductive organs (sexes are separated and probably genetically determined); (3) specialized defensive polyps spread all over the colony known as dactylozooids; and (4) tentaculozooids which probably also

serve for colony protection. The sexual polyps will develop the translucent gonophores, which are basically the sessile states of the medusa form that are morphologically reduced to ball-shaped containers of gametes (gonads) [1].

The life cycle starts with the release of gametes into the surrounding water (Fig. 2). After fertilization, the embryo develops within 72 hours into a metamorphosis-competent planula larva. As in most marine invertebrates, the larva is induced to metamorphosis by external or environmental stimuli. In the case of *Hydractinia*, this process is most probably triggered by bacteria films from the genera *Pseudoalteromonas* and *Alteromonas* present on the surface of the mollusc shell [3, 24]. The result of the metamorphosis is the primary polyp, which grows along the substrate by extension of the tubular stolons. The colony is generated by the branching of the stolons from where new gastrozooids appear in a spatially regulated manner, probably controlled by lateral inhibition influenced by already established polyps [25]. Depending on the individual growth rate of the colony, gonozooids develop within 2-3 months. Therefore, all polyps within a colony share one gastrovascular system, which is necessary for the migration of cells (i-cells), exchange of nutrients and also information [1].

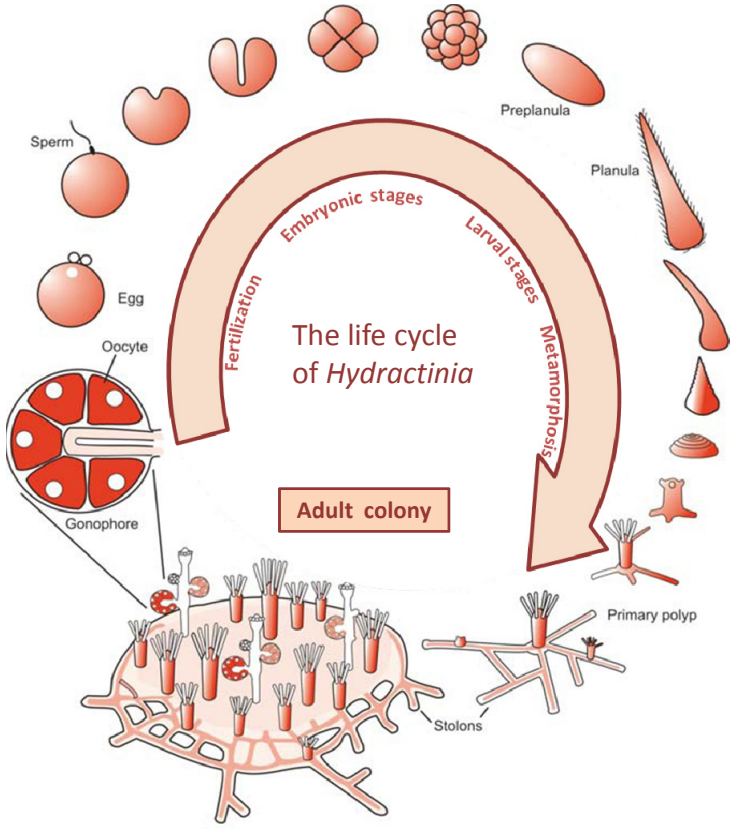


Figure 2 - Life cycle of *Hydractinia echinata*. After fertilization, the embryo will develop into the planula larva, which under certain external stimuli will undergo metamorphosis. The resulting adult morphology (primary polyp) will grow by elongation of the stolons and asexual budding of polyps, ending with the formation of a mature male or female colony (modified from Müller and Leitz [3]).

1.1.3. Stem cells

1.1.3.1. Metazoan stem cells

All metazoans have stem cells, which are undifferentiated cells with a capacity for self renewal and for the production of daughter cells committed to differentiate into products that have a shorter lifespan in comparison to the lifespan of the organism [26, 27]. Stem cells are characterized according to their potential of differentiation. Unipotent stem cells can give rise to only one differentiation product, *e.g.* spermatogonial stem cells. Oligopotent stem cells are restricted to differentiate into cells of a particular lineage, *e.g.* lymphocyte precursor cells. Multipotent stem cells are able to produce different types of cells within a particular organ, system or tissue, *e.g.* hematopoietic stem cells. Pluripotent stem cells can give rise to all differentiation products of the embryo except the extra-embryonic tissue required in mammalian development, *e.g.* embryonic stem cells and embryonic germ cells. Totipotent stem cells can differentiate into all embryonic and extra-embryonic cells types, *e.g.* the zygote and the blastomeres at the 2-cell stage [26-29].

There has been increasing interest over the last years in the use of stem cells for new therapeutic approaches in regenerative medicine. The possibility to culture human ES cells in an undifferentiated state and to regulate their differentiation into many different cell types of all three embryonic gem layers - ectodermal, mesodermal and endodermal - might be the solution to many degenerative diseases and even cancer. However due to the ethical implications involved in the isolation of embryonic stem cells, much of the current work focuses on the developmental potential plasticity of adult stem cells [30]. Besides solving this ethical issue, a treatment using patient-derived adult stem cells would also limit the problem of an immune rejection and the risk of tumour formation associated with ES or EG transplantation [30]. The first plasticity experiments failed in reproducibility as well as to demonstrate if the cell fate switching occurred at the functional level. However, there are concrete examples were environmental cues appear to reprogram precursor or differentiated cells into cells with a less mature state or able to differentiate into a new product [30]. Recent reports revealed that human adult, foetal and neonatal dermal fibroblasts were induced to a pluripotent state by an over-expression of the genes encoding for the transcription factors Oct3/4, Sox2, c-Myc and Klf4 or of a gene set including *Oct4*, *Sox2*, *Nanog* and *Lin28* [31, 32]. The resulted pluripotent cells resembled ES in morphology and epigenetic pattern, and were able to differentiate into derivatives of all three germ layers. To mimic the differentiation of ES in vivo, it is necessary to take into consideration the interplay of signals

from a multilineage plethora of cells at various stages of differentiation. Stem cells fate is regulated by intrinsic regulatory mechanisms and extrinsic signals coming from its microenvironment. It has been suggested that the integration of the extrinsic signals might be mediated by Wnt and Notch pathways, while members of the Polycomb group proteins seem to play an important role for intrinsic signal regulation [33]. It is still necessary to reveal the different developmental signals and the corresponding gene networks used to choose whether the daughter stem cell will self-renew or commit to a particular differentiation product. Understanding stem cells regulation will reveal how the tissues and organs are formed and maintained allowing a concrete biomedical application.

1.1.3.2. Cnidarian stem cells

Studies on *Hydra* revealed that all tissues are self renewing. Cells of the single epithelial layers derive from the ectoderm or endoderm unipotent stem cells. These cells are in a dynamic state; they slowly circulate within the body column and are continuously in the mitotic cycle. All other somatic cell types as well as the gametes, derive from a single, multipotent stem cell referred to as interstitial cells, or shortly, i-cells. I-cells are called so due to their location in the interstitial spaces of the ectoderm and are fast cycling cells distributed along the body column but absent from head and foot regions [26, 27]. Many cell types, derived from i-cells, migrate after acquiring their differentiation commitment into the extremities [33-35].

Despite having continuous mitotic activity the size of the animal remains constant. This is achieved by maintaining equilibrium between gain and loss of cells. Loss of differentiated cells occurs by their displacement into the extremities -tentacles or foot region- and subsequently sloughing them off. Alternatively, cells are lost from the body column into new developing buds, which is the asexual reproduction of *Hydra*, or die by apoptosis [26].

Several studies have been focussed on the identification of stem cell markers in Cnidaria. The two *nanos* (*nos*)-related genes, *Cnos1* and *Cnos2*, are an example of an exclusive expression pattern in germ cells and i-cells. This correlates to the expression profile of their invertebrate homologues where they seem to maintain the germ line [36]. The *vasa* related genes *Cnavas1* and *Cnavas2* are also expressed in germ cells and less strongly in i-cells and ectodermal cells. Genikhovich and colleges showed that the Polycomb Repressive Complex 2 (PRC2) seems to be conserved not only in bilaterians but also in cnidarians [37]. This was suggested after the identification of the embryonic ectoderm development (*EED*) *Hydra* homologue (*HyEED*) which is coo-expressed with the *Hydra* enhancer of zeste 2 (*Ezh2*) in i-cells and precursor

cells but not in terminally differentiated cells [37]. The PRC2 complex covalently modifies the histone tails, generating a dynamic mechanism of epigenetic regulation [38]. It has been suggested that PRC2 is essential for cell differentiation and multipotency maintenance of precursor or later progenitor cells. Other genes that have been identified in cnidarians include among others; the basic-helix-loop-helix (bHLH) transcription factor gene *achaete-scute* homologue *Cnash* which is expressed in nematocytes and neuron precursor cells, a homologue of the zinc finger transcription factor gene *zic/odd-paired* (*Hyzic*) expressed in nematoblasts, the serum response factors genes *HvSRF* in *Hydra vulgaris* and *HeSRF* in *Hydractinia echinata*, and the reprogramming and differentiation factor genes *Sox2*, *Brn3/5* and *c-Myc* [39, 40].

1.1.3.3. Hydractinia i-cell population

In mature colonies, *Hydractinia* i-cells are found predominantly in the stolon mat, specifically between the upper and lower ectodermal epithelium, around the endodermal canals (Fig. 3). The periphery of the colony is often deprived of stem cells. However, using the meshwork of interstitial space that surrounds the base of the epithelial cells, the i-cells can migrate from the central mat to new parts of the colony (*e.g.* growing stolons and newly emerging polyps). Therefore, nematocytes precursors migrate into the feeding polyps, primordial germ cells into the sexual polyps, while precursors of nerve cells settle in all parts of the colony [35].

Experiments with colonies depleted from their i-cell population were already performed by Müller in 1967 [41]. Elimination of i-cells resulted with the time in the immobilization, starvation and finally death of the colony. This probably occurred due to the absence of cnidocytes and therefore, the inability of the colony to capture the prey. Nevertheless it was possible to recover the colony after the addition of histocompatible donor i-cells. Even the phenotype of the colony depleted from i-cells could be changed into the phenotype of the donor, down to the germ line. This experiment showed that donor i-cells were able to provide progenitor nematoblast cells as well as nerve, epithelial and germ cells. In 2004, Müller and co-workers repeated the experiment, and demonstrated that *Hydractinia* i-cells have totipotent capacities [27]. This means that *Hydractinia* i-cells, at least under stress conditions, are able to differentiate into all cell types, including cells of the two epithelial layers. In contrast, *Hydra* i-cell lineages do not differentiate into epithelia [26, 27].

The role of the canonical Wnt pathway in the establishment of the body axis and pattern of the gastrulating embryo seems to be a common feature of all eumetazoans. Particularly in *Hydractinia* the Wnt signalling mediates the polarization of the embryo by maternal

determinants mRNAs, the fate specification of the body regions in the metamorphosing larva and the definition of the oral-aboral axis in the polyp [25, 35]. Teo and co-workers demonstrated, after the characterization of the *Hydractinia frizzled* homologue, that the Wnt/β-catenin pathway also plays a role in the regulation of stem cells [35]. Activation of the Wnt pathway was correlated with a proliferation of a progenitor subpopulation of i-cells but not with an increase of the totipotent i-cell population. After Wnt signal removal, the proliferated products committed to terminal differentiation [28, 35].

With the use of transgenic techniques, Kalthurin and co workers were able to follow up the behaviour of *Hydra* stem cells in vivo [33]. They corroborated the results on *Hydractinia*, observing that Wnt in *Hydra* also controls i-cell differentiation. Additionally, the Notch pathway appears to have a critical influence in the complete differentiation of nematoblasts. This corresponds to its role on neuronal subtype specification in higher vertebrates.

Several examples showing i-cell differentiation induced or affected by their temporal and spatial interaction with surrounding cells as well as their growth as contiguous patches, support the existence of a stem cell niche in Cnidaria [33]. All these data show that there is not only a surprisingly high genomic similarity between cnidarians and higher vertebrates, but also demonstrate that several controlling mechanisms -such as stemness- seem to be conserved at the functional level since the common eumetazoan ancestor.

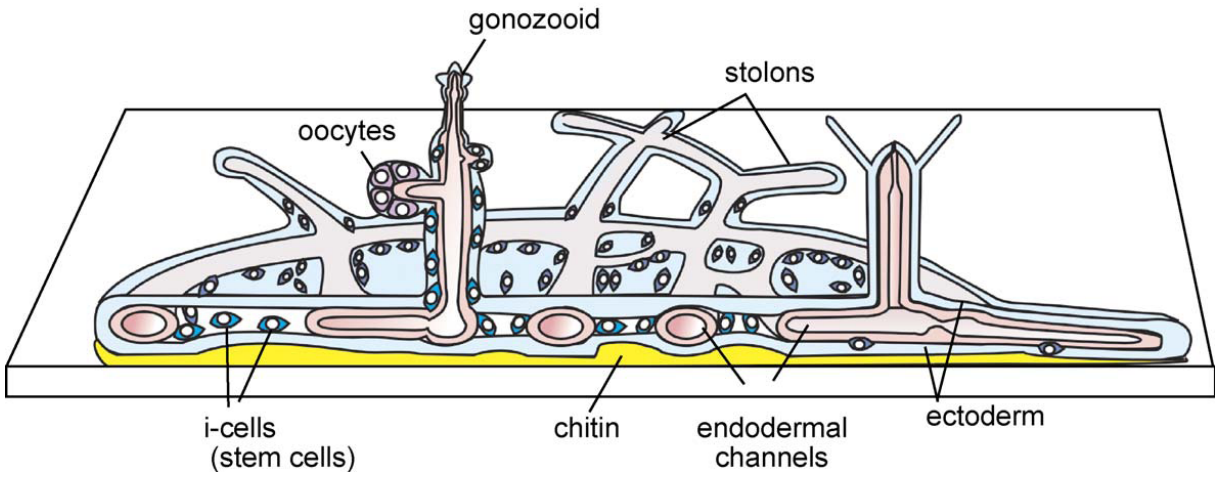


Figure 3 - Schematic representation of the i-cell distribution within a *Hydractinia* colony. I-cells migrate using the interstices spaces of the epithelia (diagram from Müller *et al*, [27]).

1.1.4. Innate immunity

The protection of an organism against pathogenic infections, and recognizing between self and non-self tissues is attributed to its immune systems [42, 43]. Two different systems are known in vertebrates; the innate and adaptive immune system. The adaptive or acquired immune system appeared in the evolutionary tree about 500 million years ago and is confined to jawed vertebrates (cartilaginous and bony fish, amphibians, reptiles, birds and mammals) [44]. The innate response is more ancient and operates in a broad variety of organisms including vertebrates, invertebrates and plants [45]. Therefore, in higher vertebrates both types of immunity coexist where the innate response provides the first line of defence against pathogens. For invertebrates, the innate immune system is the only means by which the organism can detect non-self cells or molecules and eliminate them [44, 46].

Innate immunity was first considered to be of lower specificity, characterized mainly by phagocytosis. However, recent studies revealed that innate responses can specifically discriminate not only between self and pathogens, but also between different classes of pathogens [47]. Microbial metabolic pathways provide many products which are conserved and specific for each microorganism but absent from the host repertoire. These are denoted as pathogen-associated molecular patterns (PAMPs) and include among others lipopolysaccharides (LPS), lipoproteins, peptidoglycans and lipoteichoic acids (LTAs). The innate immune system can recognise this signature and activate an appropriate response. For this, the host uses receptor molecules named pattern recognition receptor (PRR) [47, 48]. Although subtle strain- and specie- specific variations of a pathogen might not be detected, most microorganisms of a given class show a highly conserved common and invariant molecular pattern [48]. Thus, the innate immune system having a limited germ line determined PRRs is able to recognize a great variety of pathogens. However, pathogen unique gene products -the virulence factors- are not recognized by this system. These factors are produced by the pathogen in response to an adaptation within the host or for interacting with it. This means that genes encoding virulence factors have an expression profile according to the state of the infection cycle. In contrast, PRRs are expressed almost continuously and are essential for the survival of the pathogen at all life stages. In terms of evolution, it seems that the low degree of conservation as well as the inducible expression pattern selected against the use of virulence factors as a target for innate immune recognition [48].

There are different kinds of proteins acting as PRR which are expressed within cells, on the cell surface or secreted into the blood stream or tissue fluids. Secreted PRRs includes C-type

lectins, Pentraxins and lipid transfer protein families. Cell surface PRRs include proteins with leucine-rich repeats -*e.g.* Toll/Toll-like receptor (TLR)-, C-type lectins and scavenger receptors. Intracellular PRRs include proteins with dsRNA or protein kinase binding domains, leucine-rich repeats and proteins with nucleotide binding or CARD domains. With these PRRs protein families, the innate immune system is able to opsonise, activate the complement and coagulation cascades, phagocyte, activate pro-inflammatory signalling pathways and induce apoptosis [48].

1.1.4.1. Innate immune system in cnidarians

The acquisition of genomic and EST data from cnidarians does not only confirm the extent of gene loss in the ecdysozoans but also provides an informative system to reveal the immune gene repertoire of the common eumetazoan ancestor [7]. Using bioinformatics approaches, six Toll-interleukin-receptor (TIR) proteins were identified in *Nematostella* and only four in *Hydra*. The *Nematostella* TIR collection includes a myeloid differentiation factor 88 (MyD88) homologue, a Toll/TLR protein (NvTLR-1) which resembles the fly Toll, and three TIRs having immunoglobulin domains in the extracellular portion of the transmembrane protein. In the case of *Hydra*, two of the proteins are homologues to MyD88 and the rest – HyTRR-1 and HyTRR-2 – are atypical Toll-like proteins having a short extracellular domain missing any leucine-rich repeat (LRR) motifs responsible for a pattern recognition function [49]. Despite the presence of these genes, it seems that *Hydra* has no functional Toll/TLR pathway. In addition to lack a Toll/TLR protein, most of the downstream mediators of the signalling pathway that have been identified in *Nematostella* and the coral *Acropora* are missing or substantially diverged in *Hydra*. For example, the nuclear factor kappa B (NF-κB) is present in both *Nematostella* and *Acropora* but has no *Hydra* counterpart. Another example is the complement C3 protein that besides being present in the mentioned anthozoans has also been detected in *Swiftia* [49, 50]. While the complexity of these C3 proteins at the sequence and structural level resemble their deuterostome counterparts, most of their protein domains are missing in *Hydra*. The role of the TIR proteins in *Hydra* is still unknown, but it is expected that they act as a receptor together with unidentified pattern recognition molecules [16, 49]. Furthermore, it seems that pathways leading to the production of antimicrobial compounds have replaced the function of the missing complement system in the hydroid [49].

In vertebrates the complement is composed of more than 20 serum enzymes and cell receptors. After its activation, the complement mediates inflammation, opsonisation of antigenic particles and membrane damages leading to the eventual lysis of the pathogen. The

complement system can be activated via the classical –involving C1 binding to immunoglobulin coated surfaces-, alternative –by the deposition of C3b onto pathogens surface- or the lectin pathway –involving the binding of mannan-binding lectin to carbohydrates- [51]. The three pathways converge on the complement C3 which triggers a cascade of events ending in the formation of membrane-attack complexes (MAC). In cnidarians, the activation of C3 is suggested to be carried out by lectins, while the effectors mechanisms might correspond to the various MAC-perforin (MACPF) domain proteins already identified [49].

The presence in cnidarians of several immune genes, including the Toll/TLR system, a prototypic complement mechanism and even a probable Recombination activation gene 1 (*RAG-1*) related recombinase demonstrate that key components of the mammalian innate immune system were already represented in the common eumetazoan ancestor.

1.1.4.2. Allorecognition in Hydractinia echinata

Allorecognition, the ability to discriminate between self and unrelated genotypes, has been documented in various metazoan groups, from sponges and cnidarians to mammals [1, 52]. Common in many invertebrate organisms, allorecognition is characterized by a series of effector mechanisms induced by conspecific tissue-to-tissue contact, which may be regarded as the ancestral stage of histocompatibility systems of higher vertebrates. Allorecognition has been studied in different taxa, but genetic approaches have been done only in the ascidian *Botryllus* and the hydroid *Hydractinia* [20, 53, 54]. Many sessile, colonial marine invertebrates with indeterminate growth by asexual propagation may come into contact with other organisms, even from the same species, competing for the space. In the case of *Hydractinia*, one or more larvae can settle down on the same substrate. This may result in the contact of compatible colonies forming a chimera, or in case of incompatible ones, the rejection of each other's tissue leading to an antagonistic response [1, 20, 53].

It has been reported that segregation or fusibility of *Hydractinia* colonies is controlled by a set of two linked loci, the *arl1* and *arl2*, with codomonantly expressed alleles [55, 56]. Fusion occurs when the colonies share one or both alleles at both loci, whereas rejection will result when colonies share no allele. If colonies share alleles at only one locus, they transitorily fuse. Thus, considering a high polymorphism for this locus, only closely related animals may share an allele and therefore fuse [56, 57]. By fusion the colonies dissolve their periderm coat, and form a common ectoderm and gastrovascular system. Colony fusion forms a heterogeneous entity, with an increased size and genetic diversity which may have a better chance for

survival than genetically homogeneous organisms. The main disadvantage is proposed to be the germ line parasitism [1, 57]. In the case of rejection, the colonies fail to fuse and each colony competes for its integrity and space resource. There are two types of rejection: passive and active rejection. The passive allogeneic rejection results in the formation of a barrier of non cellular material between both colonies, avoiding cell contact and cell migration. In the aggressive response, there is a massive nematocytes accumulation with nematocyst discharge and abnormal growth of hyperplastic stolons, leading to the destruction of at least one of the involved colonies [1, 53, 57]. It seems that rejection is either passive or aggressive depending on whether the tissue contact is between mat or stolon, respectively [54]. Thus the basis of rejection relies on the genetics of allorecognition, but factors unrelated to the genetics determine the phenotype that will follow.

In transitory fusion the allogeneic tissues fuse but afterwards they follow a variety of incompatibility reactions, with cytotoxic rejection at the original contact area [1, 54, 57]. In the case of transient chimeras, there are no apparent benefits for the involved colonies. When the colonies fuse, there is a slowdown of the growth rate but when they reject there is a massive tissue loss [1, 57]. Similar to rejection responses, the basis of transitory fusion relies on the genetic but the characteristic of the response also depends on several other sources of variation. The different outcomes in transitory fusion might suggest that undiscovered modified loci are at play. Thus, the phenomenon of allorecognition in *Hydractinia* is more complex as originally thought. Recent studies identified an *arl2* candidate gene (putative coding sequence 7, *CDS7*) encoding a putative transmembrane receptor with a highly polymorphic extracellular domain. Sequence analysis showed significant match to proteins of the immunoglobulin superfamily, further supporting a role as allodeterminant in *Hydractinia* [55].

1.2 Technical aspects

1.2.1. Expressed Sequence Tags (ESTs)

The genome of any given organism comprises between 5% and 25% of coding DNA that is subsequently transcribed into mRNA [58]. In vitro it is possible to reverse transcribe mRNA into stable complementary DNA (cDNA) which in turn can be cloned into cDNA-library vectors. The aim of Expressed Sequence Tags (ESTs) approaches is to decipher genome sequences applying a massive cloning of cDNAs and their subsequent sequence characterization. Therefore, the generation of ESTs provides a rapid means of gene discovery, allowing biological analysis prior to the generation of a full genome sequence [59]. Besides the identification of abundantly expressed genes, ESTs can be used to map genes to particular chromosomes and to determine coding regions in genomic sequences [60]. ESTs do not completely replace the need for genomic data but rather complement it, as it is still difficult to predict at the genomic level which sequences are expressed. The full genome sequencing is a resource consuming task, in where EST projects rise as an economic alternative. The EST approach greatly facilitates traditional research strategies and has been particularly useful for complex model genomes including human, rat, mouse, fish and rice [61].

One of the drawbacks of such an approach is that ESTs provide a direct access to a large number of transcribed sequences at the expense of losing the positional information of the genes. These sequencing entries are usually submitted separately to the EST databases and without an extending annotation. This has a high tendency to generate errors, which subsequently can be easily propagated to cross-linked databases [60]. In addition, the selection of the source for the construction of the cDNA library is critical. If the source is limited to a tissue of interest, it will not represent the complete transcriptome of the organism. Even in the case of a representative source selection, rare mRNAs will probably be absent from the library. In a typical somatic cell, it is considered that the mRNA is distributed in three frequently classes. In average, the most prevalent class consists of about 10 mRNA species each represented by 5,000 copies per cell whereas; the class of high complexity comprises 15,000 different species only represented by 1-15 copies. For rare mRNA the numbers are even less promising for being represented in a cDNA library [62].

1.2.2. Searchable databases

Nucleic acid sequences offer a starting point for the understanding of the structure, function, development and evolution of genetically diverse organisms [60]. This has resulted in an explosion of the amount of DNA sequence generated over the past decades, following an exponential growth law. Nowadays, EST data alone accounts to more than 50 million entries, were the majority corresponds to human and mouse sequences. For the storage of such information, biological databanks began to emerge already in the 1980s. The first databanks were the European Molecular Biology Laboratory (EMBL) database (http://www.ebi.ac.uk/embl/) and GenBank (http://www.ncbi.nlm.nih.gov/), followed by the DNA Databank of Japan (DDBJ) (http://www.ddbj.nig.ac.jp/). The basic aim of these databases is to store DNA information in a way that is public, freely accessible and that can be retrieved and used by a third part. The EMBL, GeneBank and DDBJ standardized the process of data collection and annotation, collecting individually only a portion of the whole data produced worldwide. However, the databases synchronically exchange their information on the internet on a daily basis collaborating as part of the International Nucleotide Sequence Database (INSD) (http://www.insdc.org/). This means that sequences submitted in one database will automatically appear in the others, and therefore, the three repositories contain almost the same sets of sequences and with the same quality. In parallel to the DNA databases, also several databases appeared to store protein sequence data. Widely used are the Protein Sequence Database (PSD) of the Protein Identification Resource (PIR) and the Swiss-Prot database. The PSD-PIR is emphasized in protein family classifications, and its automatic annotation is augmented by manual annotation. The Swiss-Prot database has a minimum level of redundancy, with a high level of cross-references to others databases. Protein annotation contains bibliographic references, taxonomic data and function of the protein when possible. This minimizes errors and improves the quality of annotation, but the size of the database is smaller in comparison to other protein sequence repositories. With a lower quality, but extensive in their annotation appeared TrEMBL, which is compiled automatically from the translation of protein coding sequences of EMBL databases. Swiss-Prot and PIR-PSD created an integrated resource which finally provides the most comprehensive repository of protein sequences, called UniProt [60].

1.2.3. Microarray technology

The availability of complete sequenced genomes has opened the genomic era. The discipline of genomics can be characterized as studies dealing with whole sets of genes rather than single genes. Together with the advances in new sequencing technologies appeared experimental techniques that correspondingly worked in a high-throughput way. One of the most important of these is microarray technology, which represent a significant change in how molecular biology and gene regulation studies are done. Initially, this technology was used for the simultaneously measurement of the absolute or relative abundances of nucleic acids in a biological sample for literally several thousands of different genes [60, 63]. Nowadays, microarrays technologies have spread into many different fields. Its success rely not only in the adaptation and flexibility of the microarray technique to a particular case, but also to the allowance of combining it with other techniques [64]. Thus, besides transcriptional profiling, microarrays can be used for detailed analyses of DNA sequences including genotyping, splice variants, gene or exon identification, DNA structure analyses, DNA mapping, re-sequencing, epigenetics studies, etc. At the protein level, microarrays allow the analysis of protein binding sites, protein –DNA or –RNA interactions, structural variations, etc. Furthermore, microarrays can also be used as a device for manufacturing purposes. Although most of the progress in this area is still in a pilot or development phase, efforts are being made to synthesis genomes, genes, RNAs and proteins [64].

The first array approaches were the so called macroarrays, which being a daughter of the Southern technique, involve the immobilization of a DNA probe on a positively charged membrane of nitrocellulose, nylon or polypropylene. In the array construction, different DNA sequences are spotted on the membrane in a regular pattern, were each spot contains several identical DNA copies representing one gene. Radioactively or chemiluminescent labelled target DNAs present in a sample are allowed to hybridize to their complementary sequence on the spot. Macroarrays construction and analysis do not require complicate equipment, resulting in an economic alternative to modern-days microarrays. They are considered user-friendly, involving common hybridization techniques, and sensitive, detecting even low abundance transcripts. However, due to the surface porosity hybridization takes place in larger volumes and longer times, resulting in a limited printing resolution of 100 probes per cm^2 [65].

The nonporous nature of the glass surface of microarrays increases the printing resolution, accessing approximately 5,000 probes per cm^2, and decreases unspecific binding. Using a

standard glass slide of 25 x 75 mm, only a small volume (10-75 µl) of the target sample is needed for the hybridization, resulting in higher target concentration and increased sensitivity. Furthermore, it allows the usage of two-colour fluorescent labelling, avoiding the use of radioactive material. Hence, the targets to be compared can be labelled with different fluorescent dyes and simultaneously hybridized with a microarray in a single reaction, as described below. In contrast, macroarrays need a serial number of parallel reactions to analyse the differential expression profile of two different samples.

There are two types of DNA-microarrays, which can be differentiated according to the preparation of both the array and the sample. Oligo-arrays are produced by spotting on the slide ~25 bases long oligonucleotides using photolithography techniques (http://www.affymetrix.com). It has been shown that short probes may have poor hybridization efficiencies. This probably occurs because they can bind to different regions of a gene yielding different signal intensities [66]. To avoid this, several separate oligos representing the same gene are spotted and conclusion on gene expression is only assessed when almost all of them show the same hybridization pattern. It is suggested that the highest sensitivity and specificity can be obtained using ~70mers to 150mers oligonucleotide arrays [65, 66].

The second type of arrays is the so called cDNA-microarray. cDNA can be synthesized from the mRNA present in cells or obtained directly from cDNA libraries. After its amplification into high concentrations by PCR, the cDNAs are spotted on the slide. Despite their inability to represent the complete gene, the length of cDNA (between 500 and 2,000 bp) provides a good representation of it and allows highly specific hybridizations to the complementary sequence [65, 67].

1.2.3.1. cDNA microarray experimental settings

The slides to be used in a microarray experiment are normally pre-coated with a surface chemical to improve the binding of DNA. Aminosilane or Poly-L-lysine coated slides are the most used ones, offering a homogeneous positive charged surface and high sensitivity (signal to background ratio). The cDNA probes are spotted on the slide using a robotic device. First attachment of the DNA occurs via strong electrostatic attraction between the negatively charged sugar-phosphate backbone of the DNA and the amino groups of the coated slide. For a more stable binding, the slide is subsequently exposed to UV and/or heated, enabling the formation of covalent bonds. In average, spots contain around 10^9 individual molecules, which are able to find its complementary sequence in the hybridization solution. To determine

the amount of mRNA in a sample, it is necessary to take into consideration the hybridization kinetics. Hybridization must be performed on spots containing a large probe excess and in the initial phase of the hybridization reaction where kinetics can be approximated by a linear relationship. Only then, the measured signal in the spot is proportional to the amount of the corresponding molecules in the target [68]. However, only a small fraction of the spotted molecules will hybridize to the corresponding DNA targets. This means that the signal intensity of the spot depends on the duration of hybridization, target concentration and the amount of probe material [68]. The process of array manufacture is less reproducible for spotted arrays than for oligo-arrays, and is difficult to control the real amount of DNA in each spot. Thus, it is not possible to compare absolute intensities between slides. Nevertheless a two-colour labelling system, were the two samples are compared in one slide simultaneously, can solves this problem. For this, total RNA is isolated from the sample tissues or cells that will be compared on the microarray. Then both RNA samples are reverse transcribed into cDNA. During the generation of the first-strand cDNA, each sample will incorporate a different fluorophore. The most frequently used fluorophores are Cy3 and Cy5 [69]. Hence, one of the samples (*e.g.* cDNA of a cancer cell) will be labelled with the green Cy3 dye and the other one (*e.g.* cDNA of a normal or reference cell) with the red Cy5 dye. Then, both labelled samples are mixed and let to hybridize in a competitive manner on the array. The labelled green and red cDNAs should bind to the spots in proportion to their concentration in the complex sample. Using a laser scanner it is possible to measure the intensity of both, the green and red fluorescence from each spot. The detection is done separately in their corresponding channels, 633 nm for Cy5 and 543 nm for Cy3, resulting in two TIF images. Then, the two images are overlaid for visualization and the intensity ratio of Cy3 and Cy5 for each spot is determined [65] (Fig 4).

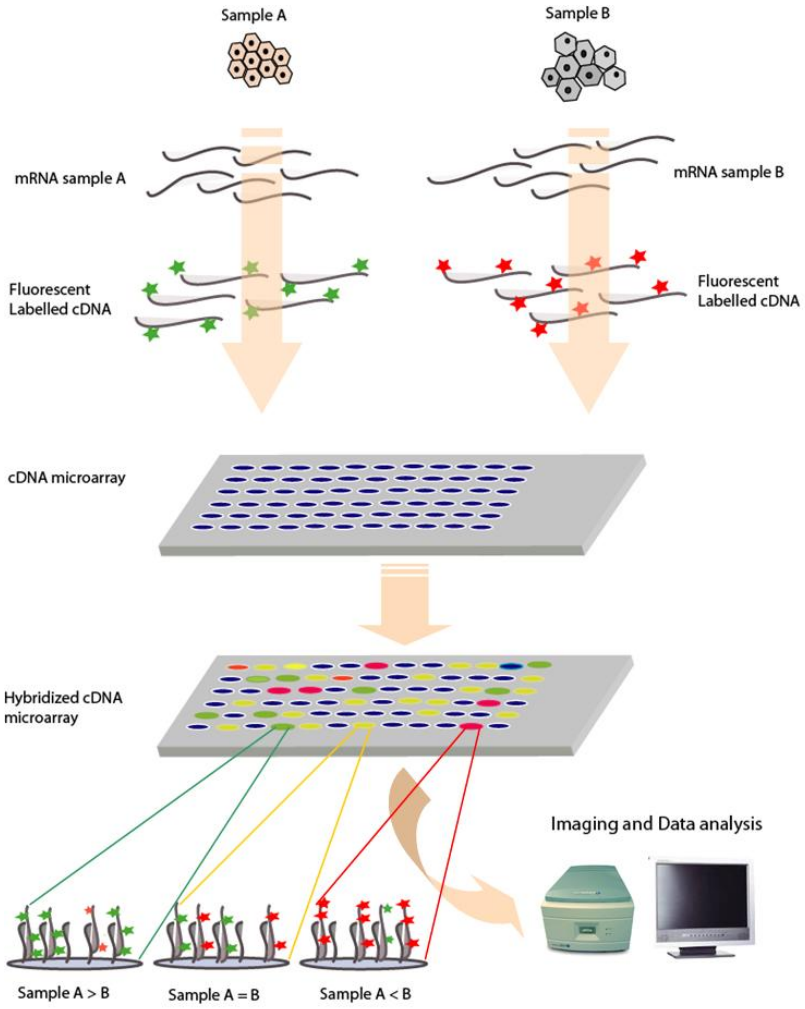


Figure 4 -Workflow of a gene expression analysis using microarrays. mRNA from two different samples (*e.g.* normal and tumour) are differentially labelled (*e.g.* Cy3 and Cy5) and co-hybridized to a cDNA microarray. The array is subsequently scanned in the corresponding wavelengths of the fluorophores. For each cDNA element on the microarray, the ratio of the fluorescent intensities reflects the relative abundance of that mRNA between the compared samples.

1.2.3.2. Processing of microarray data

With image-processing software (*e.g.* Genepix) it is possible to extract the information from a two colour microarray experiment. Microarray raw data consists of foreground and background signal intensities for the green and red channels of each spot on the array. Using the background intensities, the image-software can correct the foreground intensities for local variation on the array surface. The software also helps in assessing the quality of each spot, checking its reproducibility and allowing to flag unreliable spots or arrays. It is possible to generate a graphical display of the raw intensities values, giving an overview of the success of

the experiment, highlighting specific problems, as well as providing an idea of the tools to be chosen for subsequent analysis [65].

For the analysis and interpretation of the microarray data, the green and red intensities should be normalized relatively to one another. Normalization aims to adjust for any bias that arises from technical variation rather than from biological differences among the RNA targets or the printed cDNA probes [70]. Factors that directly or indirectly affect the raw signal intensity values include, among others, different background intensities, unequal amounts of mRNA or cDNA, differences in the labelling reaction efficiencies (called dye bias), variation among spatial positions on a slide or between slides and different hybridization or signal detection behaviours. There are different methods for data normalization which can be done within array or between arrays [65]. After normalization, the green Cy3 and red Cy5 ratios (Cy3/Cy5) should achieve an unbiased representation of the relative abundance of each mRNA in the sample. For such an assumption it is also necessary to have a statistical validation, which is normally achieved by several technical and biological replications of the experiment. [65].

Microarray experiments yield a large amount of data, limiting the identification of genes that have been significantly up- or down- regulated in the test sample relative to the reference. For the selection of differentially expressed genes, besides normalization, statistical filtering tools are applied to extract the subset of genes that may be of interest from the full dataset. These usually corresponds to genes; (a) with a large variance or periodicity within their gene expression, (b) with a high fold induction over a time course, (c) considered significant by a statistical criterion or (d) above a particular threshold, such as exceeding a given percentile rank in the distribution of the ratios [71, 72]. However, observing a list of gene names or ratios provides a poor overview of the real trends or patterns that may exist in the data. Within the different methods that have been developed for displaying microarray data, clustering algorithms or projection methods like Correspondence Analysis (CA) can be applied to study the relations among genes and hybridizations that result from expression profiling experiments [71, 73, 74]. A more detailed description will be discussed in this work.

1.3 Aims of the project

The cnidarian genomic information generated so far, predominantly based on the coral *Acropora*, the sea anemone *Nematostella* and the solitary polyp *Hydra*, is insufficient for the representation of the highly diverse cnidarian phylum. Thus, the genetic design of the common eumetazoan ancestor it is still unclear. Additional sequencing data is needed for revealing the origin and diversification of ancient gene families involved in essential metazoan features such as stemness or immunity, as well as to define their structural and functional conservation into higher metazoans.

1.3.1. General aims:

The general aim of this project is to generate a facility platform to promote molecular research in *Hydractinia* and complement the information of other cnidarian sequencing projects with the final goal to better understand the function and evolution of ancient genes.

1.3.2. Specific aims

The specific aims of this project can be summarized as follows:

- The generation of an EST data set representing a large fraction of the *Hydractinia* transcriptome
- A bioinformatics functional characterization of the generated sequencing reads and comparison of the *Hydractinia* transcriptome to the genomic information available from other cnidarians
- The generation of a database accessible in the web-interface in order to make these data widely available
- To generate a microarray comprising the already analyzed ESTs and un-sequenced cDNAs, in order to identify new candidate's genes for further sequencing and to improve the functional characterization of the already annotated sequences.
- To identify the genetic repertoire associated to the i-cell lineage and the immune system of *Hydractinia*

2. Materials and Methods

2.1 Materials

2.1.1. Chemicals and reagents

The following chemicals and reagents with an analytical research purification grade were use:

Chemical or Reagent	Manufacturer
2'-deoxyadenosine 5'-triphosphate (dATP)	Fermentas, St. Leon-Rot, Germany
2'-deoxycytidine 5'-triphosphate (dCTP)	Fermentas, St. Leon-Rot, Germany
2'-deoxyguanosine 5'-triphosphate (dGTP)	Fermentas, St. Leon-Rot, Germany
2'-deoxythymidine 5'-triphosphate (dTTP)	Fermentas, St. Leon-Rot, Germany
2'-deoxyuridine 5'-triphosphate (dUTP)	Fermentas, St. Leon-Rot, Germany
Acetic acid	Mallinckrodt Baker, Griesheim, Germany
Adenosin 5'-Triphosphat (ATP)	Fermentas, St. Leon-Rot, Germany
Agarose	Sigma, Deisenhofen, Germany
Agarose, Low melting point	Biozym, Oldendorf, Germany
Bacto Agar	Difco, Detroit, MI, USA
Bacto Tryptone	Difco, Detroit, MI, USA
Bacto Yeast Extract	Difco, Detroit, MI, USA
Betain, Monohydrat	Sigma, Deisenhofe, Germany
Bovin Serum Albumin (BSA)	Roth, Karlsruhe, Germany
Bromophenol Blue	Sigma, Deisenhofe, Germany
Caesium chloride	Serva, Heidelberg, Germany
Carbenicilin	Serva, Heidelberg, Germany
Chloroform	Fluka, Deisenhofen, Germany
CTAB (Cetyltrimethylammonium bromide)	Sigma, Deisenhofe, Germany
Cy3-AP3-dCTP	Amersham Biosciences, Freiburg, Germany
Cy5-AP3-dCTP	Amersham Biosciences, Freiburg, Germany
Diethylpyrocarbonate (DEPC)	Roth, Karlsruhe, Germany
Dimethylformamide	Mallinckrodt Baker, Griesheim, Germany
Dimethylsulfoxide (DMSO)	Merck, Darmstadt, Germany
Dithiothreitol (DTT)	Invitrogen, Karlsruhe, Germany
Ethanol	Riedel-de Haen, Seelze, Germany
Ethidium Bromide (EtBr)	Roth, Karlsruhe, Germany

Chemical or Reagent	Manufacturer
Ethylenediaminetetraacetic acid (EDTA)	Roth, Karlsruhe, Germany
Formaldehyde	Merck, Darmstadt, Germany
Formamide	Roth, Karlsruhe, Germany
Giemsa	Merck, Darmstadt, Germany
Glycerol	Roth, Karlsruhe, Germany
Guanidine hydrochloride	Roth, Karlsruhe, Germany
Guanidinium isothiocyanate	Roth, Karlsruhe, Germany
Hydrogen chloride (HCl)	Merck, Darmstadt, Germany
Isopropanol (2-Propanol)	Mallinckrodt Baker, Griesheim, Germany
Isopropyl β-D-1-thiogalactopyranoside (IPTG)	Fermentas, St. Leon-Rot, Germany
Lipopolysaccharide (LPS)	Sigma, Deisenhofe, Germany
Lithium chloride	Merck, Darmstadt, Germany
Magnesium chloride	Fermentas, St. Leon-Rot, Germany
May-Grünwald	Merck, Darmstadt, Germany
Mercaptoethanol (-2)	Merck, Darmstadt, Germany
Methanol	Riedel-de Haen, Seelze, Germany
Mitomycin-C (MMC)	Alexis biochemicals, Lörrach, Germany
Morpholinopropane sulfonic acid (MOPS)	Serva, Heidelberg, Germany
Oligo-$dT_{(12-18)}$	Invitrogen, Karlsruhe, Germany
Phenol	Roth, Karlsruhe, Germany
Phenol,chloroform,isoamylalcohol (PCI, 25:24:1)	Sigma, Deisenhofe, Germany
Potassium Chloride (KCl)	Mallinckrodt Baker, Griesheim, Germany
Sodium Acetate	Merck, Darmstadt, Germany
Sodium azide (NaN_3)	Applichem, Darmstadt, Germany
Sodium Chloride (NaCl)	Mallinckrodt Baker, Griesheim, Germany
Sodium Citrate ($C_6H_5Na_3O_7$)	Roth, Karlsruhe, Germany
Sodium dodecyl sulfate (SDS)	Sigma Chemical Co
Sodium hydroxide (NaOH)	Merck, Darmstadt, Germany
Tris	Roth, Karlsruhe, Germany
Triton X-100	Serva, Heidelberg, Germany
TWEEN-20	Serva, Heidelberg, Germany
X-Gal	Sigma, Deisenhofe, Germany

2.1.2. Solutions, buffers and media

Solution, buffers and media were sterilized by autoclaving. Alternatively, they were filtrated through a 0.22 µm filter. All commercial buffers (*e.g.* kits or enzyme reaction buffers) are described in the methods section.

2.1.2.1. Solutions and buffers

0.5 M EDTA (pH 8.0)
186.1 g EDTA
ddH_2O
Adjusted to pH 8.0 with 20 g NaOH

1 M Tris-HCl (pH 8.3)
242.2 g Tris base
ddH2O
Adjusted to pH 8.3 with concentrated HCl

1% Agarose/TAE
1 g agarose
1X TAE to 100 ml
Warm the mixture until the agarose is dissolved

10X MOPS (pH 7.0)
200 mM morpholinopropane sulfonic acid (MOPS)
50 mM sodium acetate
10 mM EDTA
DEPC treated water

10X PBS
1.4 M NaCl
0.03 M KCl
0.02 M K_3PO_4
0.1 M Na_3PO_4
ddH_2O

10X PCR-Puffer
100 mM Tris-HCl (pH 8.3)
500 mM KCl
ddH_2O

10% SDS
100 g SDS in 900 ml ddH_2O
Heat to 68°C and adjust to pH 7.2 with HCl
Fill to 1 litre with ddH2O

20X SSC
3 M NaCl
0.3 M sodium citratre (pH 7.0)
ddH2O

20% Tween-20
40 ml Tween-20
160 ml ddH_2O or DEPC

3 M NaAc (pH 5.2)
204.14 g sodium acetate
500 ml ddH_2O
Adjusted to pH 5.2 with 22 ml of 37% HCl

35% guanidine hydrochloride
35 g guanidine hydrochloride
ddH2O to 100 ml

50X TAE (pH 7.8)
242 g Tris base
57.1 ml acetic acid
100 ml 0.5 M EDTA
ddH_2O

6X DNA loading buffer
0.25% Bromphenol blue
0.25% xylene cyanol
30% Glycerin
ddH2O

7.8 M NH_4Ac
300.6 g ammonium acetate
ddH_2O to 500 ml

CTAB buffer
2% (w/v) CTAB
2% SDS
0.1 M Tris pH 8.0
1.4 M NaCl
0.02 M EDTA pH 8.0
ddH_2O

DEPC treated water
0.1% (v/v) DEPC in ddH_2O
Mix, incubate overnight and autoclaved twice

Ethidium Bromide Stock Solution
10 mM Tris-HCl
1 mM EDTA
1 mg/ml ethidium bromide
dH_2O

PBST
0.1% (v/v) Tween-20 in PBS
Sterile filtration

Proteinase K stock
10 mg/ml of Proteinase K in H_2O or DEPC
Storage at -20°C

Solution D
4 M guanidium thiocyanate
1 M sodium citrate (pH 7.0)
10% lauryl sarcosine
DEPC treated water

X-Gal solution
20 mg X-Gal
1 ml dimethylformamide
Storage at -20°C

2.1.2.2. Solutions for bacteria culture

10X H.M.F.M.
Solution 1
3 mM $MgSO_4$ ($7H_2O$)
15 mM tri-sodium citrate ($2H_2O$)
70 mM $(NH_4)_2SO_4$
45% glycerin
Addition of 800 ml ddH_2O and filtrate (0.22 µm)
Solution 2
270 mM KH_2PO_4
130 mM K_2HPO_4 ($3H_2O$)
Addition of 200 µl ddH_2O and autoclave
Mix of both solutions before use

2YT freezing media
90% (v/v) 2YT media
10% (v/v) 10X H.F.M.F.
100 µg/ml of carbenicillin

2YT medium
16 g Bacto-tryptone
10 g Bacto-yeast extract
5 g NaCl
For Agar-culture: 15 g Bacto-agar
ddH_2O to 1 litre

2YT or LB-Carbenicillin medium
100 µg/ml Carbenicillin in 1 litre of 2YT or LB medium

LB medium
10 g Bacto-tryptone
5 g Bacto-yeast extract
10 g NaCl
For Agar-culture: 15 g Bacto-agar
ddH_2O to 1 liter

2.1.2.3. Solutions for staining Hydractinia i-cells

6 mM mitomycin-C stock solution
6 mM of mitomycin-C
dissolved in methanol
Stored at -20°C

Lavdovsky's fixative
5 ml of formaldehyde
2 ml of acetic acid
25 ml of ethanol
20 ml of ddH_2O

Sørensen's buffer (0.1 M, pH 7.0)
39 ml of 0.2 M NaH_2PO_4
61 ml of 0.2 M Na_2HPO_4
100 ml of ddH_2O

2.1.2.4. Solutions for the microarrays experiments

1X Spotting solution
3X SSC
150 mM $NaPO_4$-buffer
1.5 M Betain
ddH_2O

Blocking buffer
5X SSC
0.05% (v/v) SDS
1% (w/v) BSA
ddH_2O

$NaPO_4$-buffer
600 mM Na_2HPO_4
600 mM NaH_2PO_4
Adjust to pH 8.5
ddH_2O

Rinsing Solution 1
0.1% (v/v) SDS
ddH_2O

Washing buffer A
2X SSC
0.2% SDS
ddH_2O

Washing buffer B
2X SSC
ddH_2O

Washing buffer C
0.2X SSC
ddH_2O

2.1.3. Enzymes

Enzymes and their respective reaction buffers were purchased from a range of manufacturers.

Enzyme	Manufacturer
Deep Vent®™ DNA Polymerase	NEB, Frankfurt, Germany
DNase (RNase free)	Fermentas, St. Leon-Rot, Germany
E. coli DNA ligase	Invitrogen, Karlsruhe, Germany
E. coli DNA polymerase I	Invitrogen, Karlsruhe, Germany
*Eco R*I	Fermentas, St. Leon-Rot, Germany
Hind III	Fermentas, St. Leon-Rot, Germany
Not I	Invitrogen, Karlsruhe, Germany
Proteinase K	Qiagen, Hilden, Germany
RNase H	Invitrogen, Karlsruhe, Germany
RNase out	Invitrogen, Karlsruhe, Germany
SuperScript™ III Reverse Transcriptase	Invitrogen, Karlsruhe, Germany
T4 DNA ligase	Invitrogen, Karlsruhe, Germany
T4 DNA Polymerase	Invitrogen, Karlsruhe, Germany
Taq DNA Polymerase self-made/commercial	DKFZ / Qiagen, Hilden, Germany

2.1.4. Ladders and oligonucleotides

Specific oligonucleotides were designed using the software Primer3 (v.0.4.0) at: http://frodo.wi.mit.edu.

Nucleic acids ladder	Manufacturer
GeneRuler™ 100 bp DNA ladder	Fermentas, St. Leon-Rot, Germany
GeneRuler™ 1kb DNA ladder	Fermentas, St. Leon-Rot, Germany
GeneRuler™ DNA Ladder Mix	Fermentas, St. Leon-Rot, Germany
RiboRuler™ High Range RNA ladder	Fermentas, St. Leon-Rot, Germany

Standard oligonucleotide	Sequence
M13 (22 mer-) Forward	CCCAGTCACGACGTTGTAAAAC
M13 (23 mer-)Reverse	AGCGGATAACAATTTCACACAGG
SP6 (18 mer-) Reverse	ATTTAGGTGACACTATAG
T7 (20 mer-) Forward	TAATACGACTCACTATAGGG

Specific oligonucleotide	Sequence
Actin-Forward	AAACCCTTTTCCAACCATCCTT
Actin-Reverse	TGGGCCAGATTCATCGTATTCT
HEAB-0034N17-Forward	GCATTGATGTACCTCCACCAC
HEAB-0034N17-Reverse	GCTGTTGCACATCATCAGGTA
HEAB-0042L12-Forward	GCGTCCGCGATTAAGTATCA
HEAB-0042L12-Reverse	GCTGGCGATATGAGGAAGTC
Oligo-$dT_{(15)}$-Not I anchor tag	CTAGTTCTAGATCGCGAGCGGCCGCCC$(T)_{15}$VN
Tai08H10-Forward	GATGATCTTGACCGGCTTGT
Tai08H10-Reverse	CGACAAGGGGAATACCAATG
Tai09B01-Forward	GCAAATCCTTGGGCTGAA
Tai09B01-Reverse	CGAGAGCACAAATGATCGAG
Tai11F02-Forward	TATGGCAGTGGTTGCATCAT
Tai11F03-Reverse	TTCGCGACCACCTAACTTCT
Tai16A08-Forward	AATCCTCAAGCTCGAAGTGG
Tai16A09-Reverse	TTGATGCACCGCATCTTTTG
Tai20D03-Forward	CCCTTTATTCCCCACCTA
Tai20D04-Reverse	GTGAGTATCCTGACTTTGC

2.1.5. Vector and Bacterial strain

Vector and Bacterial strain	Manufacturer
E. coli ElectroMAX™ DH10B T1	Invitrogen, Karlsruhe, Germany
pSPORT1 vector	Invitrogen, Karlsruhe, Germany

2.1.6. Kits

Kit	Manufacturer
Agilent RNA 6000 Nano kit	Agilent tech.,Böblingen, Germany
Dynabeads® Oligo(dT)$_{25}$	Invitrogen, Karlsruhe, Germany
QIAfilter® Plasmid purification kit	Qiagen, Hilden, Germany
QIAprep® Spin Miniprep kit	Qiagen, Hilden, Germany
QIAquick® PCR Purification kit	Qiagen, Hilden, Germany
R.E.A.L.® Prep 96 Plasmid kit	Qiagen, Hilden, Germany
SuperScript™ First-Strand Synthesis System for RT-PCR	Invitrogen, Karlsruhe, Germany
SuperScript™ Plasmid System with Gateway™ technology for cDNA and Cloning	Invitrogen, Karlsruhe, Germany

2.1.7. Technical material and equipment

Technical material	Manufacturer
12-channel-Pipette Biohit Proline®	Biohit, Helsinki, Finland
384-F-well plates (X7001)	Genetix, Dornach, Germany
384-pin metal replicator	Steinbrenner, Wiesenbach, Germany
384-V-well plates (X6004)	Genetix, Dornach, Germany
96-pin metal replicator	Steinbrenner, Wiesenbach, Germany
96-well flat-bottom block	Qiagen, Hilden, Germany
96-well reaction plates	Steinbrenner, Wiesenbach, Germany
Aminosilane coated slides Nexterion® Slide A+	Schott, Louisville, USA
Horizontal electrophoresis chamber (mini)	Renner, Dannstadt, Germany
LifterSlip® coverslides	Erie Scientific c, Portsmouth, USA
Microscopic slides (76x26 cm, glass)	Menzel, Braunschweig, Germany
Owl A2 Large Gel System	Thermo Scientific, New York, USA
Plastic culture plates (22x22 cm)	Nalge Nunc Int., Roskilde, Denmark
PP-tube (14 ml)	Greiner, Frickenhausen, Germany
Self-sealing alu-film	G.Kisker GbR, Stainfurt, Germany
SMP3 stealth pins	TeleChem, CA, USA

Technical equipment	Manufacturer
Agilent 2100 Bioanalyzer	Agilent Tech., Waldbronn, Germany
Centrifuge 6K15	DJB Labcare Ltd, Buckinghamshire, UK
Centrifuge Biofuge *pico*	Heraeus, Hanau, Germany
Centrifuge Megafuge 1.0R	Heraeus, Hanau, Germany
Dry Block heating system	Grant Instruments, Cambridge, UK
E. coli Transporator	BTX, San Diego, USA
Fluorescence microscope Axiovert 200	Zeiss, Göttingen, Deutchland
Gel documentration system Geldoc 1000	Bio Rad, Munich, Germany
Hybridization oven	H.Saur Lab., Reutlingen, Germany
MicroGrid II Array-Roboter	BioRobotics, Cambridge, UK
Nanodrop™ ND-1000	Peqlab, Erlangen , Germany
Power supply E835	Hoefer, CA, USA
Qfill automated microplate filler	Genetix, Dornach, Germany
Qpix Roboter	Genetix, Dornach, Germany
ScanArray® 4000XL	Perkin Elmer, Massachusetts, USA
SlideBooster hybridization station	Advalytix, Munich, Germany
Thermocycler PTC-200	MJ Research Inc., MA, USA
Thermomixer comfort	Eppendorf, Wesseling, Germany
Ultrasonic bath Sonorex RK102	Bandelin electronic, Berlin, Germany
UV-Crosslinker UVC 500	Hoefer, CA, USA
Vacuum concentrator	H.Saur Lab., Reutlingen, Germany
Vacuum manifold QIAvac 96	Qiagen, Hilden, Germany

2.1.8. Softwares

Softwares	Manufacturer
BLAST adapted to GCG/implemented to HUSAR	DKFZ, Germany/NCBI, USA
Composition	HUSAR, DKFZ, Germany
Domainsweep	HUSAR, DKFZ, Germany
Fragment Assembly System (convex-version of GCG/HUSAR)	DKFZ, Germany/Accelerycs, CA,USA
Gel package (convex-version of GCG/HUSAR)	DKFZ, Germany/Accelerycs, CA,USA
GenePix Pro 6.0	Axon Instruments, CA, USA
GOPET	HUSAR, DKFZ, Germany
Matlab Version 7.0	The Math Works Inc, Natick, USA
M-CHiPS	Kurt Fellemberg, DKFZ, Germany
Perl 5.8.9	ActiveState Software Inc., Canada
PostGreSQL	The PostgreSQL consortium
Qsoft Picking software	Genetix, Dornach, Germany
Significance Analysis of Microarrays (SAM)	Stanford University Labs, USA
Spotconverter	Kurt Fellemberg, DKFZ, Germany
TIGR Multiexperiment Viewer (MeV)	The TM4 Consortium

2.1.9. Databases and internet addresses

Databases and additional web pages	Internet addresses
Doe Joint Genome Institute	www.jgi.doe.gov/
Ensembl Genome Browser	www.ensembl.org
European Bioinformatics Institute	www.ebi.ac.uk/
Genome Sequencing Center	genome.wustl.edu/
Heidelberg Unix Sequence Analysis Resource (HUSAR)	husar/menu/biounit/
Hydractinia EST Database	www.mchips.org/hydractinia_echinata.html
National Center for Biotechnology Information	www.ncbi.nlm.nih.gov
National Human Genome Research Institute	www.genome.gov

2.2 Methods

2.2.1. Animal handling

2.2.1.1. Animal culture

Hydractinia mature colonies growing on glass slides were cultured in artificial seawater (ASW) at 18°C under illumination cycles of 14 light- and 10 dark- hours. Colonies were daily fed with 3-4 days old *Artemia salina nauplii* [1]. Fertilized eggs were collected almost daily and maintained in sterile ASW. Embryos and the subsequent larvae were raised for up to 3-5 days. Metamorphosis-competent larvae were induced to metamorphose on glass slides by three hours incubation at 18°C with 116 mM CsCl in seawater, osmotically corrected to 980 mosmol. Primary polyps were examined regularly under the dissecting microscope, and polyps showing abnormal morphology or slow growth rates were removed. Mature colonies were subcloned by cutting out pieces of the stolon plate bearing several feeding polyps. In order to generate explants exhibiting fast growth rates, the cut tissue included part of the soft peripheral area of the stolon plate. The explant pieces were held in the desired location using glass pearls until the tissue resumed growth and adhered to the surface.

2.2.1.2. Mitomycin-C exposure

The antibiotic mitomycin-C (MMC) was dissolved in methanol to a concentration of six-millimolar. From this stock solution, several aliquots were prepared and stored at -20°C. For the treatment, the slides with the colonies were placed in a petri-dish and repeated incubations with increasing concentration of MMC were performed, at RT in the dark. In the first treatment, animals were incubated for 3 hours with 3 µM MMC in seawater. Subsequently, they were washed 4 times with ASW and let to recover overnight. After 24 hours the treatment was repeated and at 48 hours, colonies were incubated overnight with an increased concentration of 15 µM MMC. Animals were washed, fed and let to recover for the next 24 hours. Finally, after 96 hours the colonies were incubated overnight with 30 µM MMC. Then, animals were washed and let to recover under normal culture conditions. To assess for i-cell depletion, explants were extracted at different time-points for cytological examinations. To recover the i-cell depleted colonies, genetically identical donor explants (prepared before the

MMC treatment) were grafted in the middle and edge of the treated colony. The stolons of the explants were allowed to join the gastrovascular system of the treated colony.

2.2.1.3. Lipopolysaccharide (LPS) exposure and allorecognition challenge

Hydractinia colonies were incubated for one hour at 18°C with 100 µg/ml of LPS. Then, the animals were washed 4 times with ASW and let to recover. For the allorecognition experiments, explants from genetically distinct adult colonies were transferred into a common glass slide. They were cultured under normal condition till they grow into contact with each other. Incompatible allogeneic rejections were assessed when colonies started to generate hyperplastic stolons.

2.2.1.4. Staining of i-cells

For cytological examinations, colonies grown on glass slides were fixed with Lavdovsky's fixative overnight at 6°C. After several washes with water, the samples were permeabilised with Sörensen's buffer (pH 7.0) supplemented with 1% Triton-X100 for 1 hour. I-cells were stained with May-Grünwald at RT for 3.5 hours. After washing with Sörensen's buffer for 1 hour, the samples were further stained with Giemsa for 3.5 hours. Final distaining was done overnight at RT with Sörensen's buffer.

2.2.2. Preparation of DNA and RNA samples

2.2.2.1. RNA isolation

Approximately 50 mg of tissue was lysed in 500 µl of solution D (containing 0.7% of β-mercaptoethanol), 500 µl of phenol and 100 µl of 2 M sodium acetate. After the addition of 200 µl of chloroform, the reaction was vigorously shaken for 15 seconds and incubated on ice for 20 minutes. Then, it was centrifuged at 4°C, 14,000 x g for 20 minutes and the aqueous phase containing the nucleic acids was carefully transferred into a RNase free tube. Subsequently, 250 µl of ice-cold isopropanol and 250 µl of 1.2 M NaCl/0.8 M sodium citrate were added. The reaction was incubated at RT for 30 min, shortly vortexed and centrifuged at 4°C, 12,000 x g for 10 min. Then, the pellet was air dried for 15 minutes and incubated with 400 µl of 4 M LiCl at RT for 5 minutes. The mix was centrifuged at 4°C, 5,000 x g for 10 min and the pellet was dissolved in 250 µl of solution D (containing 0.7% β-mercaptoethanol) at 65°C for 5 min. After a briefly centrifugation, the RNA was precipitated adding 250 µl of

100% ice-cold isopropanol, incubating the reaction at -20°C for 30 min and centrifuging at 4°C, 10,000 x g for 10 min. The RNA pellet was washed with 300 µl of 70 % ice-cold ethanol in DEPC-treated water and incubated at RT for 15 min. The solution was centrifuged at 4°C, 10,000 x g for 10 min and the supernatant was removed. The RNA pellet was air-dried at RT for 10-15 min to eliminate fluid traces by evaporation. Finally, the isolated RNA was dissolved in 20-30 µl of RNase free water and stored at -80°C.

2.2.2.2. Isolation of genomic DNA

Approximately 50 mg of tissue was digested in a solution containing 750 µl of CTAB buffer (pH 8.0) and 0.2 mg/ml of proteinase K. After tissue maceration, the mix was incubated at 65°C for 2 hours. Every 30 minutes the solution was softly vortexed. A second batch of proteinase K was added and the solution was incubated overnight. Then, 1 ml of a mix containing phenol, chloroform, isoamylalcohol (PCI) in the ratio 25:24:1 was added and the solution was mixed by inverting the tube. After centrifuging at 10,000 rpm for 10 minutes, the aqueous layer was transferred into a new reservoir and a second PCI separation was performed. Subsequently, one volume of CI (24:1) was added to the aqueous solution and softly mixed. The samples were centrifuged at 12,000 rpm for 12 minutes and the supernatant was collected. The DNA was precipitated by adding 2.5 volumes of 95% ice-cold ethanol and 10% v/v of sodium acetate, incubating the mix at -20°C for 20 minutes. The DNA was spun down at 4°C, 12,000 rpm for 20 minutes and the resulting pellet was washed twice with 750 µl of 70% ice-cold ethanol. After centrifuging at 4°C, 5,000 rpm for 5 minutes, the pellet was air dried at RT for 15 minutes. Finally, the genomic DNA was resuspended in 50 µl of DEPC-treated water and stored at -80°C.

2.2.2.3. Assessing the quality and quantity of the isolated DNA and RNA

The quantity and purification grade of the extracted nucleic acids was determined measuring the absorbance of the sample at different wavelengths using the Nanodrop. Absorvances at 260 nm allowed to calculate the nucleic acids concentration, whereby one $A_{260\ nm}$ unit corresponds to 40 µg/ml of ssRNA or 50 µg/ml of dsDNA. The ratio between the absorbance at 260 nm and 280 nm estimates the purity of the sample. Only samples with a ratio above 1.9 were selected for further analysis.

The integrity of the isolated nucleic acids was assessed by agarose gel electrophoresis. In the case of DNA, a 1% agarose gel was cast and 0.5 µg of sample was loaded. The gel was run in

1X TAE buffer for 45 minutes at 100 volts. In the case of RNA samples, a 1.4% agarose gel with formaldehyde and Morpholinopropane sulfonic acid (MOPS) was cast. An aliquot of 1 µl of the RNA sample was mixed with loading buffer containing ethidium bromide. To denature the RNA, the mix was heated at 70°C for 10 minutes. Then, the sample was chilled on ice, loaded in the gel and ran in 1X MOPS buffer for 30 min at 100 volts. After the electrophoretically separation, the gel was observed under the U. V. transluminator.

Alternatively, the Agilent Bioanalyser was used to analyze the RNA samples. For this, the RNA 6,000 Pico kit was employed following the manufacturer instructions. The assay involves the loading of the RNA samples in a RNA-chip containing an interconnected set of microchannels. These channels work as a gel platform for the electrophoretically separation of nucleic acid fragments based on their size. The assay has a high sensitivity and can determine the integrity of the RNA sample by calculating a RNA integrity number (RIN) which considers the presence or absence of degradation products. RNA samples without degradation were considered when the integrity number was above 6.5.

2.2.2.4. Plasmid DNA preparations

For the purification of the plasmid from the bacteria cells, two different protocols were used. In the analysis of particular clones the plasmid preparation was done with the QIAprep Spin Miniprep Kit, which is based on the alkaline lysis of the bacterial cells followed by a silica-column separation. For this, each clone was cultured in 5 ml of LB media containing 100 µg/ml of carbenicillin at 37°C under shaking overnight. The culture was centrifuged at 1,500 x g for 5 min and the supernatant was removed. From the bacterial pellet, the plasmid DNA was isolated following the protocol described by Qiagen. During the preparation, the optional wash step for the removal of nuclease traces was performed and the elution of the plasmid DNA was done by the addition of 50 µl of 10 mM Tris-HCl (pH 8.5).

In the case of high-throughput plasmid purification, the R.E.A.L. prep 96 Plasmid Kit was used. This procedure is based on modified alkaline lyses of the bacteria cells, followed by the removal of the lysates through vacuum filtration and finally, the purification and concentration of the DNA by isopropanol precipitation. Clones were cultured in 96 deep-well plates containing 1.3 ml of LB or 2YT media with 100 µg/ml of carbenicillin. After the overnight culture at 37°C under shaking, the plate was centrifuged at 1,500 x g for 5 min. The media was removed and the bacterial pellet was used as the starting material in the protocol described by Qiagen. In the preparation, the optional boiling procedure as well as the ice

incubation was omitted. The isopropanol precipitated pellet was dissolved in 100 µl of Tris-HCl (pH 8.5).

2.2.2.5. Restriction digests

To analyze the cDNA insert, the plasmid DNA was digested with the restriction enzymes *EcoR*I and *Hind*III. To ensure a proper digestion, three units of enzyme were added per µg of cDNA in the corresponding buffer (R^+ buffer from Fermentas) and the reaction was incubated at 37°C for a minimum of two hours. BSA was added to the reaction mix to a final concentration of 100 µg/ml. The inactivation of the restriction enzymes was carried out by incubating the reaction at 65°C for 20 minutes.

2.2.2.6. Semi-quantitative reverse transcription polymerase chain reaction (sqRT-PCR)

Semi-quantitative RT-PCR was used to determine the relative amount of mRNA transcripts in a particular sample. For this, total RNA from different *Hydractinia* stages were isolated as described in section 2.2.2.1. Subsequently, 2 µg of total RNA were reversed transcribed into cDNA using the First-strand Synthesis system for RT-PCR from Invitrogen, following the manufactures protocols. Briefly, 2.5 µg of oligo-$dT_{(12-18)}$ and 0.5 mM each of dNTP were added to the RNA sample. After the incubation of the mix at 65°C for 5 minutes and chilling on ice, 10 µl of the reverse transcription reaction mix were added. The reaction mix contained 2 µl of 10X reverse transcription buffer, 2 µl of 25 mM $MgCl_2$, 100 mM dithiothreitol, 40 units of RNase out and 200 units of Superscript III reverse transcriptase. The reaction was incubated at 50°C for 50 minutes. Subsequently, the reaction was terminated incubating the mix at 85°C for 5 minutes and chilling on ice. Finally, 2 units of RNase H were added and the reaction mix was incubated at 37°C for 20 minutes.

Transcripts to be analysed by sqRT-PCR were amplified from 2 µl of cDNA using sequence specific primers. The reaction mix also contained 0.2 mM each of dNTP, 1X PCR reaction buffer containing 15 mM $MgCl_2$ and 1 U of *Taq* DNA polymerase. The PCR protocol described in Table 2 was followed with modifications in the number of amplification cycles. In RT-PCR, product quantification can be performed when the reaction is in the logarithmic phase of amplification. Therefore, for each transcript, the cDNA sample in which the target is highly expressed was selected as template to perform seven PCR reactions having different amplification cycles (15-36). Then, the optimal cycle number to obtain a logarithmic

amplification of the target was defined. Gene expression was compared in the different samples performing the optimized PCR and analysing the amplified products in an agarose gel electrophoresis.

2.2.3. DNA and RNA methods involved in the cDNA library

2.2.3.1. Isolation of RNA for the cDNA library

To maximize the collection of expressed genes, RNA was extracted from different developmental stages as well as from organisms subjected to induction experiments. Subsequently, all RNA samples were pooled in different percentages (Table 1) and used for the library construction. Before any RNA isolation, animals starved for up to two days.

The 10 different developmental stages included were: early embryos at 1-5 hours post fertilization (pf), gastrulating embryos at 24 hours pf, pre-planula and planula larvae at 2 and 3 days pf, respectively, metamorphosing animals at 3, 16, 28 and 72 hours post induction (pi) of metamorphosis with CsCl and finally mature female and male colonies.

Five different types of induction experiments were performed. i) Heat shock treatment: primary polyps were incubated for 30 min at 30°C, washed with ASW and incubated for 1 h at 18°C before RNA isolation. ii) Osmotic shock treatment: mature colonies were incubated for 1h at a salinity of 1.7%, washed with ASW and incubated for 1 h at normal salinity before RNA isolation. iii) Regeneration treatment: polyps were cut and incisions were made in the stolon mat of an adult colony. After 3 hours of recovery, RNA was isolated. iv and v) included a Lipopolysaccharide (LPS) and allorecognition challenge already described in sections 2.1.1.3 and -4.

Table 1. *Hydractinia's* RNA-pooling strategy

RNA source	Pooling %	Total RNA used in pool (µg)
Early embryo (1-5 h)	7.5	16.8
24 h gastrulating embryo	7.5	16.8
48 h pre-planula	7.5	16.8
72 h planula	7.5	16.8
Metamorphosing larvae (3 h pi)	7.5	16.8
Metamorphosing larvae (16 h pi)	7.5	16.8
Metamorphosing larvae (28 h pi)	7.5	16.8
Metamorphosing larvae (72 h pi)	7.5	16.8
Mature colony (male)	15	33.6
Mature colony (female)	15	33.6
Heat shocked animals	2	4.5
Osmotic shocked animals	2	4.5
LPS-treated animals	2	4.5
Allogeneic transplantations	2	4.5
Regeneration	2	4.5
Total	100	224.1

The left column shows the different stages and induction experiments used for the construction of the cDNA library along with the corresponding amount of RNA material used in the pool.

2.2.3.2. cDNA library construction

Poly A+ RNA was isolated from 224 µg of pooled total-RNA using the Dynabeads mRNA purification kit. For this, 300 µl of binding buffer (20 mM Tris-HCl (pH 7.5), 1 mM LiCl, 2 mM EDTA) containing 3 mg of magnetic beads coupled with oligo-dT residues were added to the RNA sample. After several washing steps with the washing buffer (10 mM of Tris-HCl (pH 7.5), 0.15 mM LiCl, 1 mM EDTA), the bound mRNA was capture with the use of a magnet and eluted with 60 µl of 2 mM Tris-HCl (pH 7.5). Subsequently, the mRNA solution was diluted in 240 µl of DEPC-treated water and was again purified with the magnetic beads. Final elution was done with 50 µl of 2 mM Tris-HCl (pH 7.5), and the quality and quantity of the mRNA was analysed as described in section 2.2.2.3.

The cDNA library was constructed from 2.2 µg of poly A+ RNA. For the cDNA synthesis, the SuperScript™ Plasmid System with Gateway™ technology for cDNA and Cloning was used following the manufacturers protocols. The synthesis of the first cDNA strand was primed with 50 µg/ml of oligo-$dT_{(15)}$ carrying a *Not* I anchor tag at the 3´end. The primer was mixed with the 2.2 µg of mRNA, incubated at 70°C for 10 minutes and chilled on ice. Then, 4 µl of 5X first strand buffer (50 mM Tris-HCl (pH 8.3), 75 mM KCl, 3 mM $MgCl_2$), 10 mM dithiothreitol and 0.5 mM of each dNTP were added. The reaction was gently vortexed and

incubated at 45°C for 2 minutes. Then, 600 units of Superscript III were added and the reaction was further incubated at 45°C for 1 hour.

The second strand of the cDNA was synthesized by nick translational replacements, using as template the first cDNA strand. For this, 40 units of *E. coli* DNA polymerase, 2 units of *E. coli* RNase H, 10 units of *E. coli* DNA ligase, 30 µl of 5X second strand buffer (25 mM Tris-HCl (pH 7.5), 100 mM KCl, 5 mM $MgCl_2$, 10 mM $(NH_4)_2SO_4$ and 0.15 mM ß-NAD^+) and DEPC-treated water until a final volume of 150 µl were added. The reaction was incubated at 16°C for 2 hours. To ensure a blunt terminus of the generated cDNAs, 10 units of T4 DNA polymerase were added and the reaction was further incubated for 5 minutes at 16°C. After the addition of 10 µl of 0.5 M EDTA, the DNA was purified by organic extraction (PCI) and precipitated with ethanol.

In order to maximize the ligation of the cDNA into plasmid vectors, 10 µg of adapters with *Sal* I recognition sites were ligated to the blunt end cDNA products. For this, 5 units of T4 DNA ligase with the respective T4 DNA ligase buffer (50 mM Tris-HCl (pH 7.6), 10 mM $MgCl_2$, 1 mM ATP, 5 % PEG 8000, 1 mM DTT) were added. The reaction was incubated overnight at 16°C and the cDNAs were deproteinized by organic extraction (PCI) and precipitated with ethanol. Subsequently, the cDNAs were digested with 60 units of *Not* I restriction enzyme at 37°C for two hours. This resulted in asymmetrically cDNAs, which at the 5' end contained a *Sal*I and at the 3' end a *Not*I recognition site.

Size fractionation of the cDNA was done by column chromatography following the manufacturer protocols. The yield of the first and second strand reaction was determined measuring the amount of precipitable radioactivity. To calculate the amount of incorporated ^{32}P on the cDNA, a scintillantion counter was used. The size ranges of the synthesized products was estimated by loading the radioactive labelled samples in an alkaline agarose gel electrophoresis and their subsequent exposure to x-ray film, as described in the manufacturer's instructions. Only the cDNAs of the largest fractions obtained in the fractionation steps were ligated into the plasmid vector pSPORT1 and electroporated into ElectroMAX^{TM} DH10B T1 phage resistant cells, as described below.

2.2.4. Cloning strategies

2.2.4.1. Ligation into pSPORT1vector

The cDNA inserts were ligated into *Not*I- *Sal*I-Cut pSPORT1 vectors following manufacturer protocols. This was carried out by the addition of 25 ng of vector and three fold molar excess

of the cDNA insert. The reaction was catalyzed by 25 units of T4 DNA ligase in the respective buffer (1 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, 1 mM ATP, 5 % (w/v) PEG 8000, 1 mM DTT). Finally, the reaction was gently mixed and incubated overnight at 4°C.

2.2.4.2. Electrotransformation of E. coli cells and clone culture

Vectors containing cDNA inserts were introduced by electroporation techniques into ElectroMAX ™ DH10B-T1 phage resistance cells. The cDNA sample was mixed on ice with 25 µl of freshly thawed electro-competent cells and transferred to a 0.1 cm electro-cuvette. The mix was incubated on ice for 5 minutes and a pulse with 1.8 kV was applied. Immediately after, the mixture was resuspended in 1 ml of S.O.C media and incubated at 37°C under shaking for 45 min. Then, the culture was spread in plastic plates (22 x 22 cm) containing 200 ml of LB-Agar media with 200 µg/ml of carbenicillin. In order to eliminate the empty vectors from the library, the agar plates contained 500 µl of 20 mg/ml X-Gal and 90 µl of 0.2 g/ml IPTG. Finally, the plates were incubated at 37°C overnight.

2.2.4.3. Colony picking and setting the Hydractinia cDNA library

For colony picking an automatic robotic devise was used. The Qpix robot use a CCD camera for imaging the colonies to be picked by the 96-pin picking head, in which each pin individually samples a single colony. Colony selection is made defining the picking parameters of the QSoft Picking software, for example; in the colour, roundness and diameter of the colony. To avoid picking satellites or double-colonies, only colonies with a diameter between 4-40 pixels were selected. To improve the picking of small colonies, the pin was deeply inserted (~3 mm) into the agar. Sample carryover and cross-contamination was diminished by the sterilization of the pins in 80% ethanol before each new picking round of the robot pin head. The picked colonies were transferred to 384-well microplates previously filled with 50 µl of LB media containing 10 % HFMF freezing solution and 100 µg/ml of carbenicillin. Finally, the plates were incubated overnight at 37°C.

Plates containing more than 15 non-inoculated wells were discarded from the library, and used for colony replacement purposes. Two replicates were made from each plate selected for the library. For this, a 384-pin metal devise was used, which was sterilised between each inoculation by two wash-flame steps with 70% ethanol. The obtained replicates were incubated overnight at 37°C, while the original plates were labelled and stored at -80°C.

During the culture, the replicates were observed to identify slow growing bacteria and empty wells. Finally they were labelled and placed at -80°C.

2.2.4.4. Assembling the Hydractinia-chip library

To avoid redundancy in the *Hydractinia* microarray the previously generated EST library was re-arrayed, transferring the most representative sequences of each 3,808 contigs into new 384-well plates. The *Hydractinia*-chip library was supplemented with 4,992 un-sequenced clones, prepared in the EST project. Additionally, 384 external control DNA sequences -LORECs, artificially generated randomized sequences of 100 bp- were included in the library [75].

2.2.5. DNA and RNA methods involved in the microarray experiments

2.2.5.1. Isolation of RNA for the microarray experiments

For the mitomycin microarray experiment, RNA was isolated from colonies at 96 hours post MMC treatment. 50% of the colony tissue of each biological replicate was used for the RNA extraction. The rest of the colony was allowed to recover. After 4 weeks of the MMC treatment, RNA was isolated from active polyps budding areas of recovering colonies.

For the immune microarray experiments, colonies were induced to different immune responses. In the case of the LPS experiment, RNA extraction was done at 1 and 3 hours after the LPS induction. For the allorecognition experiment, RNA was isolated only from the contact area of colonies showing signs of rejection.

All these isolated RNAs were reverse transcribed into labelled cDNA, as described below, and used in the different microarray hybridization experiments.

2.2.5.2. Target labelling for microarray hybridization

Target for hybridizations were generated from total RNA by incorporation of fluorophor-labelled dCTP during first strand cDNA synthesis. For this, the RNA sample was diluted with DEPC-treated water to a concentration of 7.5 µg/15 µl. To prime the first cDNA strand reaction, 2.5 µg of oligo-$dT_{(12-18)}$ were added and the mix was incubated at 65°C for 10 minutes. After chilling on ice, 22 µl of the reverse transcription reaction mix were added. The reaction mix contained 3 nmoles either of Cy3 or Cy5-dCTP, 0.3 mM each of dATP, dGTP and dTTP, 20 µM dCTP, 100 mM dithiothreitol, 40 units of RNase out and 400 units of Superscript III reverse transcriptase in the buffer provided by the manufacturer. The reaction

was incubated at 42°C for 1 hour. Subsequently, 200 units more of Superscript III reverse transcriptase were added and the mix was further incubated at 42°C for 3 hours. The enzyme was inactivated at 70°C for 15 minutes. The RNA strand of the DNA-RNA hybrid was degraded with the addition of 2 units of RNase H, incubating the reaction mix at 37°C for 20 minutes. Finally, the labelled single stranded cDNAs were purified with the QIAquick PCR purification kit as described below.

2.2.5.3. Purification of the labelled cDNAs

To purify the cDNA fragments from the dye-labelled reaction, one volume of the reaction mix was diluted with 5 volumes of PB buffer (pH ≤ 7.5). After a gently mix, the solution was transferred to a QIAquick column and centrifuged at 13,000 rpm for one minute. The DNA, which is bound to the column, was washed with 0.75 ml of a 35% guanidine hydrochloride solution and centrifuged at 13,000 rpm for one minute. A second wash step was done with the addition of 0.75 ml of buffer PE and the corresponding centrifugation at 13,000 rpm for one minute. To remove rest of the washing solution, another centrifugation for one minute was performed. The DNA was eluted twice with 30 µl of DEPC-treated water (pH 8.5).

2.2.5.4. Determination of the yield of the cDNA synthesis and the Cy3/Cy5 incorporation rates

The Nanodrop™ was used to measure the absorbance of the labelled cDNA sample at four different wavelengths. The absorbance of the sample at 260 nm and 280 nm was used to determine the purity and yield of the cDNA reaction (section 2.2.2.3), while its absorbance at 550 nm and 650 nm allowed to calculate the incorporation rate of Cy3 and Cy5 in the cDNA, respectively. To avoid disturbances due to the A_{260nm} of the dye molecules, the concentration of the single strand cDNA was calculated with a correction factor. This means, that one $A_{260\ nm}$ unit equals to 33 µg/ml of ssDNA minus 0.08 x A_{550} in the case of Cy3 and 0.05 x A_{650} for Cy5. The incorporation rate of the dye molecules was calculated with the following formulas:

$$\boldsymbol{Cy3\ incorporation\ rate} = \left(\frac{\boldsymbol{c[Cy3]}}{\boldsymbol{c[cDNA]}}\right) \times \mathbf{100\%}$$

$$\boldsymbol{Cy5\ incorporation\ rate} = \left(\frac{\boldsymbol{c[Cy5]}}{\boldsymbol{c[cDNA]}}\right) \times \mathbf{100\%}$$

2.2.5.5. Polymerase chain reaction (PCR)

The polymerase chain reaction uses a DNA polymerase from the thermophilic prokaryote *Thermus aquaticus* to amplify a piece of DNA. The uses of specific primers, which hybridize to the target DNA sequence, define the starting point of the DNA polymerase to initiate the DNA synthesis. The generated DNA products in each PCR round are subsequently used as template for further amplification. In the present project, single PCR reactions were used to analyse the insert of particular clones, while whole 96-well plate PCR reactions were used to amplify the probe to be printed on the microarrays (*i.e.* the *Hydractinia*-chip library). In order to have yield and quality homogeneity in all PCR products, logarithmic-phase bacteria cultures were used as PCR templates. For this, all clones were cultured in 96-deep-well blocks containing 1.2 ml of 2YT and 200 µg/ml of carbenicillin at 37°C under shaking for 7-9 hours. Subsequently, an aliquot was used to inoculate a new media for an overnight culture at 37ºC. Then, 5 µl of suspension culture were transferred to a 96-well PCR plate containing 95 µl of ddH_2O. To lyses the bacterial cells and release the plasmid, the mix was heated at 95°C for 10 minutes. Cellular debris was removed by the centrifugation of the PCR plate at 1,200 x g for 3 minutes. Then, 4 µl of the supernatant were transferred to a new 96-well PCR plate to be used as template. In the case of individual PCR reactions, the same procedure was followed in single reaction tubes or alternatively, plasmid DNA was used as template. To the template DNA, 0.5 mM each of T7-forward and SP6-reverse primers were added. Alternatively, the reaction was primed with M13 -forward and -reverse primers. The reaction mix also contained 0.2 mM each of dATP, dCTP, dGTP and dTTP, 1X PCR reaction buffer containing 15 mM $MgCl_2$, 0.5 to 1 U of self made *Taq* DNA polymerase, 0.16 U of Deep $Vent_R$™ DNA polymerase and DEPC-treated water until a final reaction volume of 100 µl.

The inserts of the cDNA-library have high size heterogeneity. Therefore, to optimize the amplification of certain cDNAs, especially those with more than 3 kb, the PCR was performed with some variation in the number of cycles and amount of the reaction reagents; *e.g.* with more units of long range DNA polymerase, higher amount of primer, etc. DNA amplification was performed in a programmable temperature controller Thermocycle able to perform 96 PCR reactions in parallel. The PCR program was adapted depending on the annealing temperature of the primers and the size of the template fragment. The different PCR programs are described below (Table 2).

Table 2 – PCR programs used to amplify clones from the cDNA library

steps of PCR	PCR of fragments below 3 kb	PCR of fragments above 3 kb	PCR of whole 96 plates
Denaturation	94 °C for 30 sec	94 °C for 30 sec	94 °C for 30 sec
Cycles	35	28-31	31
Denaturation	94 °C for 15 sec	94 °C for 15 sec	94 °C for 20 sec
Annealing	48-53 °C for 45 sec	48-53 °C for 30 sec	50 °C for 35 sec
Elongation	68 °C for 3 min	68 °C for 4 min + 0.05 min /cycle	68 °C for 6 min
Extension	72 °C for 7 min	72 °C for 7 min	72 °C for 7 min
Cooling	10 °C for 10 min	10 °C for 10 min	10 °C for 10 min

2.2.5.6. Control of the PCR products

All the PCR products to be used in the microarray were checked by agarose gel electrophoresis using the OWL separation system from Thermo scientific. This system permitted the analysis of 192 PCR products (2 x 96 well PCR plates) in parallel. For this, 2 µl of the PCR products were mixed with loading buffer and transferred to a 1% agarose gel using a 12-channel pipette. Four DNA markers, either the 1kb DNA ladder or the GeneRuler DNA ladder Mix from Fermentas, were loaded in the gel. The DNA was separated in 1X TAE buffer applying 100 volts for 35 minutes. The quality of the fragments was analyzed under the U. V. transluminator. In the case of a PCR plate with an amplification yield below 85%, the PCR reaction of the whole plate was repeated and optimized. Clones with repeated negative PCR results, in most of the cases due to the big size of the insert, were individually treated and re-organize in the original 96-well plates.

2.2.6. Construction of the microarray

2.2.6.1. Preparation of the PCR products for the printing of the microarray

96-well PCR plates with an amplification yield above 85% were selected for the construction of the microarray. From the selected plates, 45 µl of the PCR products were transferred to 384-well plates, whereby four 96-well plates were used to fill out one 384-well plate. The PCR products were dried out in a vacuum concentrator at RT overnight. Subsequently, the PCR pellets were resuspended in 10 µl of spotting buffer and were gently shaken in a plate mixer at 300 rpm for 30 seconds. Before printing the PCR products on the slides, all 384-well plates were centrifuged at 1,000 rpm for 1 minute. Finally, the *Hydractinia* array included 10 *Hydractinia* PCR plates already sequenced and analyzed in the EST project, 13 un-sequenced *Hydractinia* PCR plates and 1 LORECs PCR plate as external control. An additional negative

control plate was added which contained only spotting buffer. The remaining volumes of the PCR reactions were stored at -20 ºC as a backup for further analysis or subsequent microarray fabrications.

2.2.6.2. Printing the PCR products on the aminosilane coated slides

The PCR products of the 25 384-well plates were printed on the surface of the aminosilane coated slides (Nexterion® Slide A+) using the microarray roboter MicroGrid™ II. This was achieved using a print head containing 24 (2 x 12) SMP3 stealth pins. These pins have a liquid reservoir of approximately 100 nl and can deliver drops of about 0.8-1 nl, resulting in spots with a diameter of 80-100 µm. The geometry of the array, distance of the spots, order of the printed PCR, number of pre-spotting and microarray slides were defined with the use of the Microgrid software. For printing, we followed the manufacturer's protocol. Briefly, pins were washed before every source visit by two washing steps in double distilled water of 5 seconds each and by 3 washing cycles in the main washing station of the robot. During the run, pins were regularly checked for uniformly delivery of the probe. In the case of pins with an insufficient performance, they were immediately changed. Pins were allowed to softly touch the slide surface. For the pre-spot and normal microarrays slides the speed of the pins was set at 4 m/s with target height of 1 mm and 0.6 mm, respectively. Pins were refilled with new probe every 103 printed slides and they first printed the primary spots. Once finished, pins were used to print the duplicate spots. The whole printing procedure was performed at RT with 40% humidity.

2.2.6.3. Post-processing of the microarray slides

The printed microarray slides were kept in a vacuum excicator at RT for 12 hours. In order to identify the position of the printed DNA, the microarray area was carefully marked on the back of the slide with a diamond scriber. To increase the DNA-slide binding efficiency, covalent bonds between the probe and the slide amino-groups were generated by irradiating the slides with 250 x 100 µJ/cm^2 UV using a UV-crosslinker. Subsequently, the microarrays were incubated at 80ºC for 4 hours. Finally, the microarrays were stored at 4ºC in a vacuum excicator containing desiccant beads.

2.2.7. Microarray hybridization methods

2.2.7.1. Preparing the array for the hybridization

To avoid unspecific bindings and to remove unbounded DNA or rest of buffer substances, the slides were washed and their active surface was blocked before the hybridization of the target sample. For this, a maximum of five slides were placed in a slide-holder. To avoid blending of the spots, all washing procedures were done quickly by moving the slides up and down in the respective solutions. First, the slides were washed in a bath containing 500 ml of Rinsing solution 1 at RT for 10 seconds. Then, the slides were washed with double distilled water at RT for another 10 seconds, transferred to a third bath containing water at 95ºC and incubated for 3 minutes. Subsequently, the active surface of the slides was blocked by immersing the slides in a fourth bath containing 200 ml of blocking solution at 55ºC under soft shaking for 45 minutes. The slides were washed again with water at RT for 10 seconds and immediately after, were carefully dried with compressed air. The DNA microarrays were ready to be used in the hybridization reaction.

2.2.7.2. Microarray Hybridization

Equal amounts of Cy3- and Cy5- labelled cDNA were allowed to co-hybridize on the microarray. For this, 7.5 µg of each differentially labelled target sample (section 2.2.5.2) were mixed and dried in the dark in a vacuum concentrator. Subsequently, the cDNA pellet was resuspended in 5 µl of 10 mM EDTA (pH 8.0) and softly vortexed for 30 seconds. Meanwhile, a cover-slide with spacers of 0.05 mm thick (LifterSlip) was placed over the printed surface of the microarray. This allowed generating a chamber for hybridization. Then, the microarray was placed in the SlideBooster hybridization station. Advason coupling liquid was added between the slides and the incubation chamber floor. Whereby the surface acoustic waves produced by the microagitation-chips of the chamber floor oscillated the hybridization solution. To maintain a high humidity within the hybridization chamber, 900 µl of AdvaHum were added in the disposed reservoir.

While the microarray was heated until the hybridization temperature, the labelled cDNA was denatured at 95ºC for 1 minute. After spinning down, 65 µl of the hybridization buffer (SlideHyb#1) previously heated at 68ºC were added. The solution was mixed thoroughly and directly pipetted to the microarray, in the edge of the cover-slide. By capillarity, the solution

equally spread all over the microarray hybridization chamber. Hybridization was carried out at 62 ºC for 16 hours.

Afterwards, to remove un-hybridized target cDNAs, the microarray slides were washed at RT with three different buffers preheated at 37ºC. The first wash was done for 10 minutes with buffer A, which contains 2X SSC and 0.2% SDS. Then, they were rinsed for 10 minutes each in buffer B (2X SSC) followed by buffer C (0.2X SSC). All washing steps were performed under softly shaking and in the dark. Finally, the slides were shortly (2 seconds) immersed in isopropanol and immediately dried with compressed air. The fluorescence intensities of the hybridized microarrays were directly measured, otherwise the slides were stored for short periods in a vacuum excicator at RT in the dark.

2.2.7.3. Signal detection

The Cy3- and Cy5- fluorescence signals were measured with the confocal laser scanner Scanarray 4000XL. The microarrays were scanned with a resolution of 10 μm, at a constant laser power, but varying the photomultiplier (PMT) in order to avoid saturation and get the best foreground to background intensity ratio. The excitation of the Cy5-dye was achieved using a laser with wavelengths of 633 nm and its emission was detected by a sensor at 670 nm. For the Cy3-dye, excitation of the molecules occurred with a laser having wavelengths of 543 nm and their emission was detected at 570 nm. A 16-bit TIFF grey-scale image was generated for the emission signals of each dye.

2.2.7.4. Quantification of the signal intensities

The Genepix software was used to quantify the signal intensity of the spots. First, a graphical representation of the scanned image was generated by false-colouring the Cy3 and Cy5 channels in green and red, respectively. Overlying the image from both channels allowed preliminary analysis, whereby yellow spots represented an equal amount of target in the two compared samples and green, red or spots with mixed colours between green and red corresponded to differentially expressed genes. For the quantification of the intensities, a spotting-matrix or grid was placed over the merged image. The grid was semi-automatically adjusted to the spots and blocks of the array. This grid has a Genepix array list format (gal. file) and transfers the gene information to each spot in the array, *e.g.* the clone position in the source plates (cDNA library), gene annotation, etc.

The raw signal intensities of both dyes were calculated for each spot and the generated values were saved and exported as a table file (extension gpr.). In addition, the table contained the average, media and standard deviation of the spots signal intensities, the intensities of the local background, number of pixels, etc.

2.2.8. Bioinformatics methods related to the EST project

2.2.8.1. EST sequencing and Sequence Analysis Pipeline

Single-pass cDNA sequencing from 5'- and/or 3'-ends was conducted at the Washington University Genome Sequencing Center [76]. After removal of vector and ambiguous regions from the raw sequence data, the sequence reads were uploaded to the EST database at the NCBI (National Center for Biotechnology Information [77]). The first step in the sequencing analysis pipeline was a download of the single-pass sequences in FASTA format. Subsequently, the Wisconsin GCG package Fragment Assembly System (FAS) available at the Heidelberg Unix Sequence Analysis Resource (HUSAR, [78]) was initialized. Within FAS, the Gel-package programs were used, starting the assembly project (GelStart), uploading the sequences in GCG format (GelEnter), aligning them into contigs (GelMerge), editing the assembled contigs (GelAssemble), displaying contig structures (GelView), and finally evaluating the created FAS database with respect to quality and statistics (GelStatus and GelAnalyze). The generated consensus sequences, which represented each EST-cluster, were used as a query for BLAST (Basic Local Alignment Search Tool) homology searches against GenBank databases [79, 80].

2.2.8.2. Annotation and subsequent analysis of the Hydractinia sequences

At the DNA level, searches were made by BLASTN algorithm against the non-redundant nucleotide database at NCBI using the default parameters. In the case of insignificant hits, searches were performed against the GeneBank ESTs databases (dbEST) at NCBI. At the protein level, analyses were done using BLASTX against the Swissprotplus database under the Sequence Retrieval System (SRS) [81] at the HUSAR Bioinformatics Lab, which includes the latest full releases both of Swissprot and SpTrembl [82]. Matches with an E-value acceptance threshold of less than 10^{-6} were retrieved from the results page and stored in our local server. For annotations at the nucleotide level only the fourth hit was stored in the

Hydractinia Database, minimizing the chances to retrieve the *Hydractinia* sequences itself which were already annotated on the NCBI nucleotide database.

At the protein level, the first hit was directly linked to the database and used for further analyses. Sequences without any significant annotation or with an un-informative hit -*e.g.* hypothetical, probable, putative or chromosomal annotation- were further analyzed using DomainSweep [83], which allows the identification of domain architectures within a protein sequence. It employs different database search methods to scan a number of protein/domain family databases. A positive match was only considered, when the sequence contained at least two domain hits described in two protein family databases that are members of the same InterPro family/domain or, when there were two blocks or motifs in a correct order already described in the Prints or Blocks dataset. Further functional annotations were made by adding Gene Ontology (GO) terms to the sequences using the Gene Ontology term Prediction and Evaluation Tool, GOPET, available at HUSAR Bioinformatics Lab [84]. Only hits above a confidence threshold of 80% were annotated with GO terms of the two main categories, biological process and molecular function. In subsequent analysis the consensus sequences were compared with TBLASTX against different databases that were downloaded into HUSAR from NCBI, Ensembl Genome Browser [85], and from the Joint Genome Institute. For TBLASTX analysis, significant hits were considered when matches presented an E-value acceptance threshold of less than 10^{-3}. The downloaded databases included: the cnidarian EST databases of *Acropora*, *Hydra* and *Nematostella*, as well as the raw and assembled genome data of *Nematostella*; the new releases of vertebrate cDNA datasets of *Homo sapiens*, *Pan troglodytes*, *Macaca mulatta*, *Canis familiaris*, *Rattus norvegicus*, *Gallus gallus*, *Danio rerio*, *Xenopus tropicalis*; and the ecdysozoan invertebrate cDNA datasets of *Aedes aegypti*, *Anopheles gambiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*.

2.2.8.3. Hydractinia Database

All relevant information about every EST as well as the information generated in the sequence analysis pipeline was automatically integrated into a Database using in-house scripts. The database is a PostgreSQL relational database [86]. For an easy-to-use platform, a web interface was created using Perl/CGI. It can be accessed at:
http://www.mchips.org/hydractinia_echinata.html

2.2.9. Bioinformatics and statistical methods involved in the microarray experiments

2.2.9.1. Normalization and filtering of the signal intensity data

The M-CHiPS (Multi-Conditional Hybridization Intensity Processing Software) software package implemented in Matlab® (MathWorks) was used to analyse the signal intensity data generated in the microarray experiments [72]. First, the table files generated by Genepix (*i.e.* gpr. files of section 2.2.7.4) were separated into green (Cy3) and red (Cy5) intensity channel data. Then, the condition and control measurements, independent of the labelling, were defined for all hybridizations and uploaded in M-CHiPS.

To compensate for systematic variations in the quality of the data, M-CHiPS corrected the raw intensity values based on the logarithmic regression of one condition measurement versus the control measurement [72]. Since in both microarray experiments several repetitions were performed, the median signal intensities of the measurements were used. The log-Cy3 versus log-Cy5 median signal intensities for all genes were plotted and the regression curve was calculated. The algorithm applied a multiplicative and additive corrector to fit the set of genes to the regression curve [72, 74, 87]. A good fitting performance was considered when the regression curve matched the dense region of the plotted data points, with correlation coefficient above 0.8.

M-CHiPS was used to select genes that have a good evidence of being differentially expressed [72]. For this, four different filtering criteria were used. First, all genes with raw intensity level above the detection limits (65,000 AU) were eliminated. Second, all genes having a substantial expression level (*e.g.* fitted intensity level of 1,000-2,000 AU) at least in one condition were selected. Third, the Significance Analysis of Microarray (SAM) program was used to select genes with significant expression changes in at least one of the condition with respect to the control. This program assimilates a set of gene-specific t-tests and assigns to each gene a score based on its change in gene expression with respect to the standard deviation of repeated measurements for that gene [88]. Genes with a score above a defined threshold are considered for further analysis. To estimate the false discovery rate (FDR), 300 permutations of the measurements were performed. All genes with a significant (corrected p-value > 0.05) differential expression level were selected to be filtered by the last filtering criteria. This consisted in the selection of all genes having a two-fold change in expression in at least one of the condition with respect to the control.

2.2.9.2. Correspondence analysis

Correspondence analysis (CA) is an explorative and descriptive method which allows identifying interdependencies between variables in a complex data environment [74]. The data filtered in M-CHiPS was analyzed with correspondence analysis. For this, genes were represented as numerical vectors (points) in a high dimensional space determined by the number of hybridizations (experiments). Conversely, hybridizations vectors were represented in a scenario which dimensionality is equal to the number of genes. Both, genes and hybridization vectors in their respective dimensionalities can be displayed simultaneously in a plot [73]. However, for an easy visualization accounting the main variance of the data, the high-dimensional scenario was scaled down by decomposition into principal axes and projection into a low dimensional space (biplot). Whereby, the χ^2 distances among the represented objects in the two or three dimensional plot resemble their original distances in the high-dimensional space as closely as possible. To simplify the visualization of associations between the variables, virtual genes with an ideal transcriptional profile in a defined condition were included in the plot. The coordinates of such genes are the standard coordinates of the condition were this gene is expressed [72-74].

2.2.9.3. Hierarchical and k-means clustering

The microarray data was further analysed using two different clustering algorithms. First, the M-CHiPS filtered data was exported as a table file to the TIGR Multiexperiment Viewer (MeV) [89]. This table contains the normalized signal intensity median of the selected genes for the condition and control measurements. Then, the Log_2 ratio between the condition and control measurements was calculated. Thus, in all fold-change regulation analysis the expression level of each condition is referenced to the control. Logarithmic conversion was used since it tends to make the data variability more constant, treating the number and their reciprocal symmetrically [71]. Thus, a gene up-regulated by a factor of 2 has a Log_2 ratio of 1 and genes equally expressed (with a ratio of 1) will have a Log_2 ratio equal to zero. A colour-code heat map was used to represent this data.

The microarray data was grouped into expression clusters using Hierarchical clustering (HCL). This agglomerative clustering method works bottom up, by initially assigning each gene to its own cluster. Clusters are then iteratively merged in pairs of clusters, based on their similarity distance, until all groups are hierarchical connected. This algorithm generates a similarity distance-matrix giving high scores to similar patterns of the data [65]. Pearson correlation was used as a distance metrics. Similarity-distances between the clusters were

calculated using the average distance of each member of one cluster to each member of the other cluster (Average linkage method). With an inter-node distance threshold of -0.99, the clustering generated a tree representation of the data.

In addition, the microarray data was clustered using k-means (KMC). First, the number of clusters for the partition of the data was defined using a Figure of Merit (FOM) algorithm. This algorithm removes one sample from the total data set, clusters the remaining data, and then calculates the fit of the removed sample to the previously obtained cluster-pattern [65]. This is done for all samples in the dataset. From each run, a FOM value is returned. These values were plotted versus the number of clusters in order to define the optimal clustering parameters for k-means. Subsequently, k-means was used to group the data. It starts randomly assigning genes into a particular k-cluster. Then, each gene in the data set is compared to the k-clusters-genes and distributed into the cluster with the most similar expression profile. Pearson correlation was used as distance metrics. This procedure is repeatedly refined until its optimization for the clustering criterion [65]. For an easy visualization of the microarray data in each cluster, the Log_2-transformed expression ratios (conditions/control) for all hybridization were plotted with a colour-code GO term annotation.

3. Results

3.1 EST analyses on *Hydractinia echinata*

3.1.1. Generation of the *Hydractinia echinata* ESTs

In a previous work, a size-selected and representative cDNA-library was created. In order to generate an EST dataset covering a large fraction of the *Hydractinia* transcriptome, pooled RNA preparations that were collected from various stages of the complete life cycle were used for the construction of the cDNA-library. The pool was complemented with RNA from several induction experiments (Table 1, section 2.2.3.1). Optimization of the different steps involved in the generation of the clone resource resulted in a library consisting of 21,120 clones, distributed in a 384-well plate format [90]. Enzymatic restriction and PCR analyses revealed that 86% of clones carry a cDNA-insert, with lengths between 0.4 and 5 kb and an average of approximately 1.8 kb (Fig. 5).

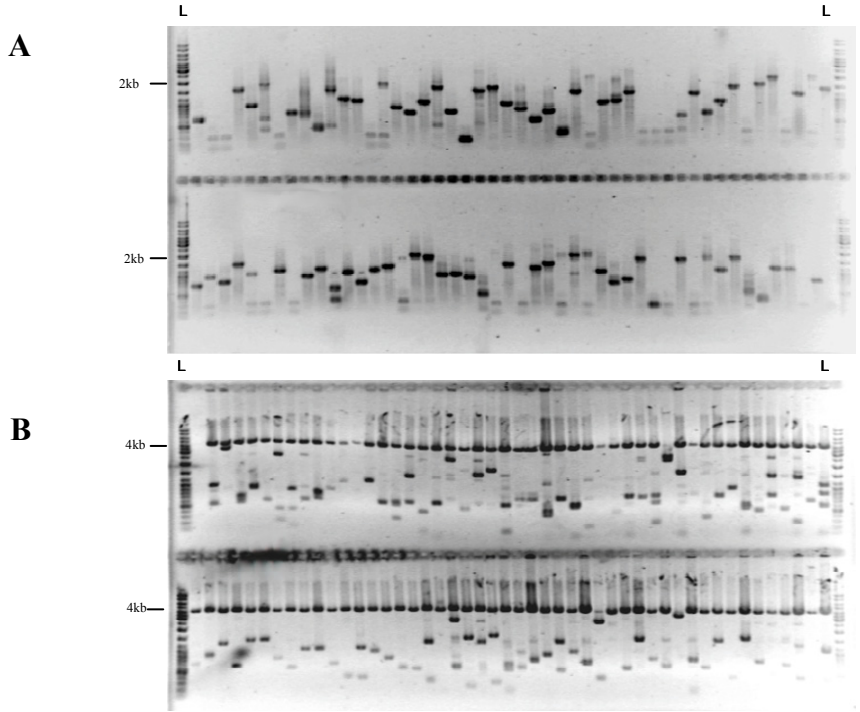


Figure 5 - Quality analysis of the *Hydractinia*-clone library. A. PCR analysis from plate 001. 80% of the inserts were successfully amplified. **B.** Enzymatic restriction analysis of the plate 003. The agarose gel shows that only 8 clones were empty. In both cases high insert size variability was observed.

From the randomly picked clones, 8,151 and 827 sequences were generated from the 5´- and the 3´ -ends respectively. A first clustering was made by physically merging sequence reads derived from clones that were sequenced from both ends. Then, 8,212 sequences were uploaded in the sequence analysis pipeline, as described in section 2.2.8. The sequences were grouped into 3,808 EST clusters including 2,625 singletons and 1,183 clusters of two or more clones comprising 5,587 ESTs (Fig. 6). Finally, consensus sequences with an average length of 439 bp representing each EST cluster were generated and used in the subsequent analyses.

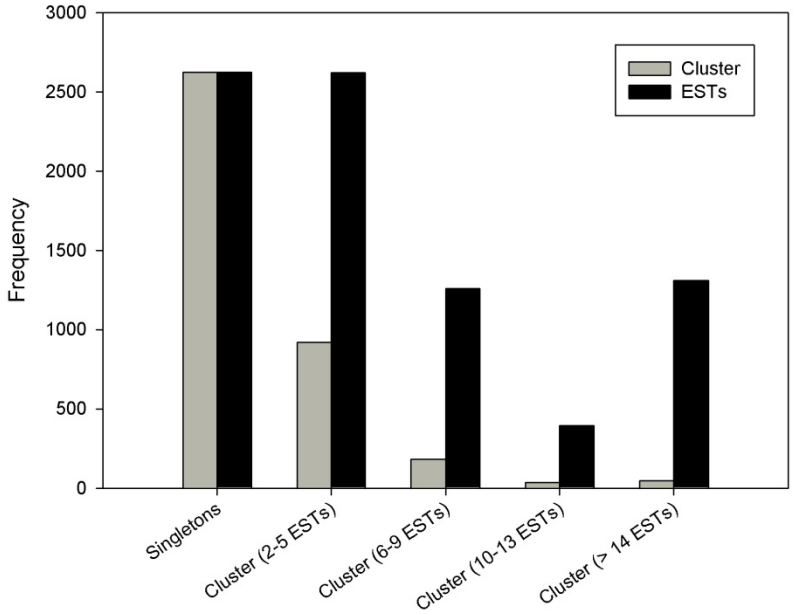


Figure 6 - Prevalence distributions of the EST cluster size. *Hydractinia* ESTs were grouped into 3,808 clusters consisting of 2,625 unique sequences or singletons, 919 clusters of 2-5 ESTs comprising 2,622 ESTs, 182 clusters of 6-9 ESTs containing 1,261 ESTs, 36 clusters of 10-13 ESTs comprising 393 ESTs and 46 clusters of more than 14 ESTs comprising 1,311 ESTs.

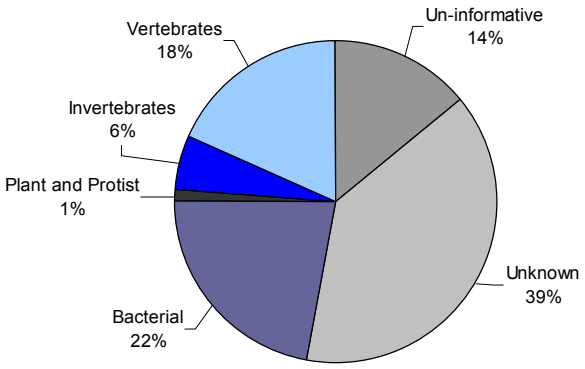
3.1.2. ESTs functional annotation

Analysis by BLASTX showed that, with an E-value acceptance cut-off of less than 10^{-6}, 1,797 *Hydractinia* sequences (47.5%) matched entries in protein databases. The majority of these hits accounted for vertebrate, invertebrate and non-metazoan proteins including fungi, plants, protists and prokaryotes. As expected, a high percentage of ESTs (38.5%, 1,468 sequences) exhibited no similarity to any sequence while 543 sequences (14%) presented an un-informative annotation (Fig. 7A). In order to characterize these ESTs, we searched for known

protein-domain architectures within the sequences. This allowed to assign 267 new functional annotations (Table S1 in Additional data 1, section 6.1).

For an overview of all the different functional classes present in the data, sequences were also annotated with Gene Ontology (GO) terms. Within the ontology *molecular function*, most of the *Hydractinia* sequences were associated with a hydrolase, transferase and binding activity including nucleotide, nucleic acid and protein binding. In the category *biological process*, the majority of the GO term predictions appeared to be related to metabolism (*e.g.* biosynthetic and catabolic processes), cell communication and biogenesis, transport and regulation of biological processes (Fig. 7B).

A Best BLASTX match for *Hydractinia* assembled sequences



B Annotation of *Hydractinia* sequences with Gene Ontology terms of the categories; i) biological process and ii) molecular function

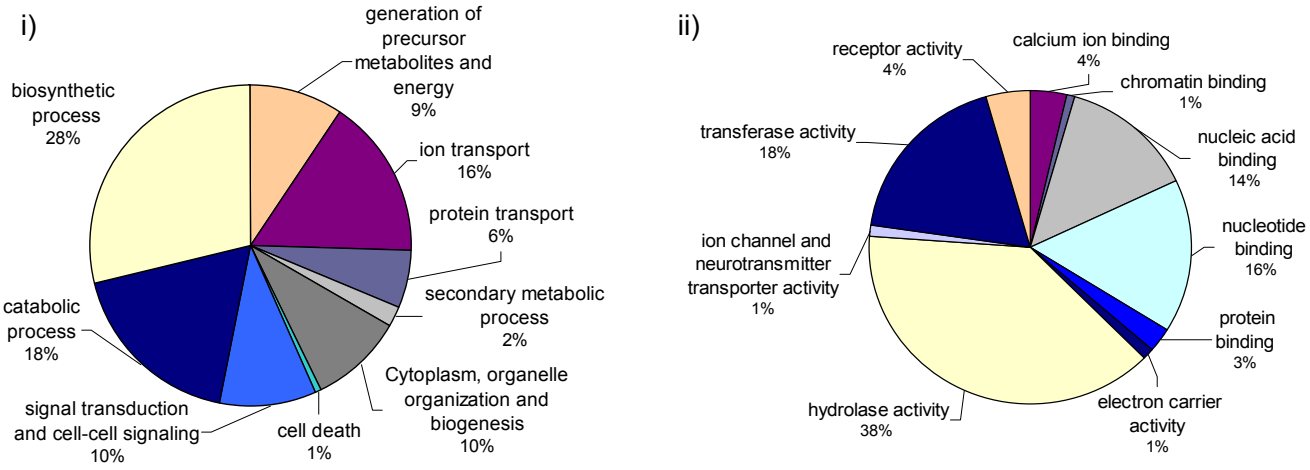


Figure 7 - *Hydractinia* ESTs sequence analyses. A. Distribution of *Hydractinia* best matches. Proportion of consensus sequences showing BLASTX matches in the Swissprotplus database (HUSAR), percentage of sequences without any significant hit, and distribution of best-EST matches to specific organism classes. **B.** Distribution of *Hydractinia* sequences into the two GO functional categories i) biological process and ii) molecular function. All

functional assignments were done at the "inferred from electronic annotation (IEA)" level of evidence.

3.1.3. Non-metazoan hits

In the BLASTX analysis, 22% (844 sequences) of the *Hydractinia* proteins showed a non-metazoan prokaryotic hit, from which 263 and 491 sequences had homologies to bacteria from the beta- and gamma-proteobacteria classes, respectively. Amongst the former, homologies to *Bordetella* spp. and *Burkholderia* spp. accounted for the majority of the hits, while in the latter class, 425 sequences presented homology to *Pseudomonas* spp. To analyze if this is a common feature within cnidarians, the *Hydractinia* sequences were compared by TBLASTX algorithms to the *Acropora*, *Hydra* and *Nematostella* EST datasets as well as to the recently annotated *Nematostella* genome. It was observed that with an E-value acceptance threshold of less than 10^{-3}, 58% (487 sequences) of the prokaryotic protein sequences are represented at least in one of the mentioned datasets, including 331 sequences with a hit on the DNA of *Nematostella*. Analysis at the nucleotide level using BLASTN with the same significance criterion revealed that 201 of these sequences (24%) are common within cnidarians.

The GC profile of the sequences classified as non-metazoan was significantly different to the profile observed in sequences with a metazoan hit. The slope of the metazoan sequence dataset showed almost no overlap with the non-metazoan one, presenting median GC contents of 0.39 and 0.62%, respectively (Fig. 8). Unknown and un-informative sequences presented GC values that ranged from 0.3 to 0.85%. With average and median GC values of 0.43 and 0.40%, unknown sequences showed a tendency to be grouped together with the metazoan group. However, the wide spread distribution of the data did not allow to demonstrate such a tendency in a significant manner. Similar was the case of sequences with un-informative hits, which presented a tendency towards the non-metazoan group (Fig. 8). Comparing the GC composition of *Hydractinia* sequences to different organisms, it was observed that the *Hydractinia* metazoan sequences co-clustered in the range of 39-42% of GC content with the GC profiles of the *Hydra* and *Nematostella* EST datasets as well as with the *Caenorhabditis elegans* cDNAs. In the case of *Hydractinia*'s non-metazoan consensus sequences, their GC content spread out from the previous mentioned profiles together with the GC percentage of bacteria like *Pseudomonas aeruginosa* and *Mycobacterium tuberculosis* [91-94] (Fig. 9).

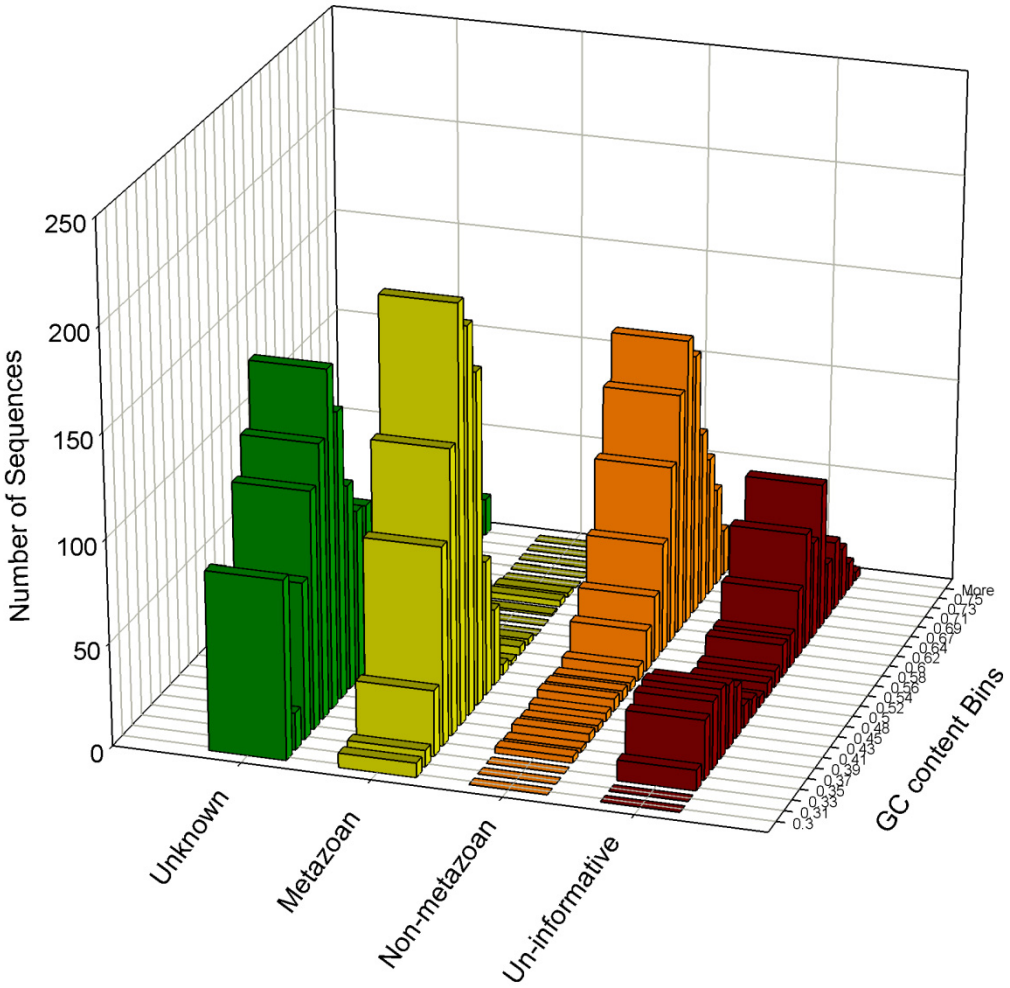


Figure 8 - GC profile of *Hydractinia* sequences. The GC content of all consensus sequences having a length of more than 100 bp was calculated using Composition (HUSAR). Sequences were sub-clustered according to the BLASTX results into metazoan, non-metazoan, un-informative and unknown. The histogram was created by the distribution of the GC content into 24 bins starting at 0.30% GC content. The metazoan sequences' GC content (median 0.39%) was significantly different from the GC content (median 0.63%) of non-metazoan sequences ($p < 0.05$). Unknown and uninformative sequences had median values of 0.40 and 0.60 % respectively.

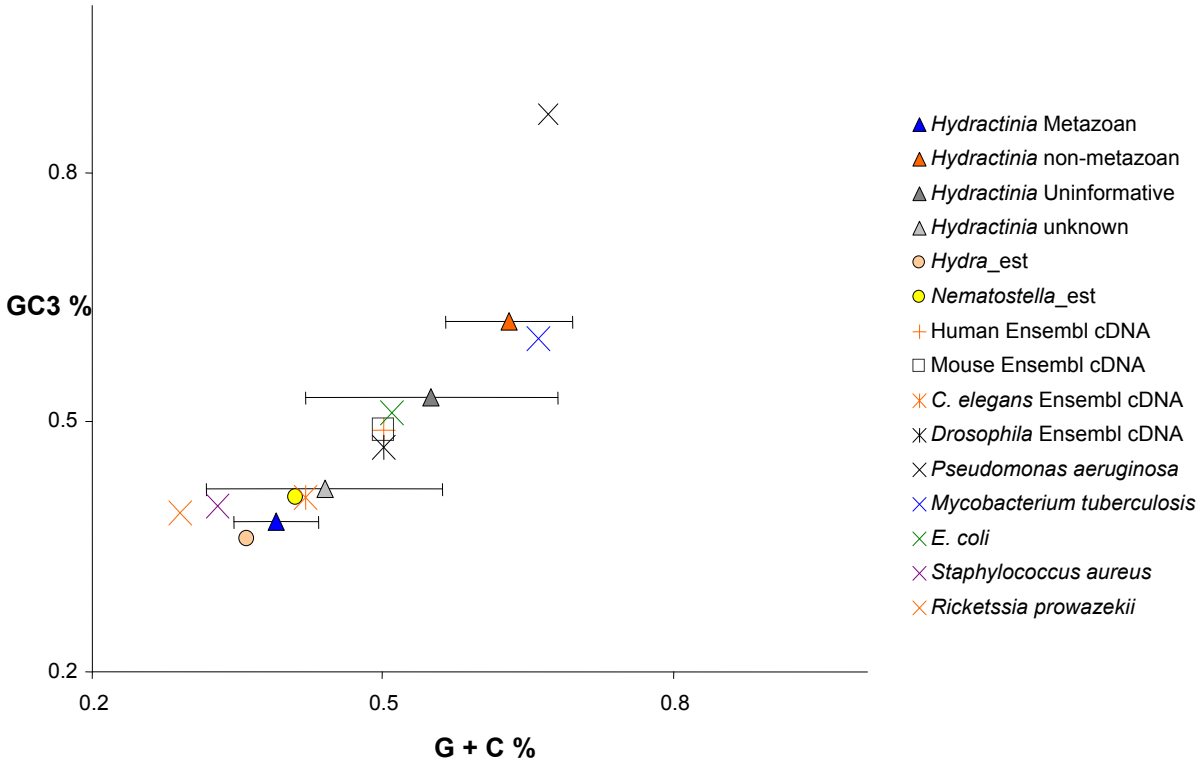


Figure 9 - *Hydractinia* GC profiles compared with EST and cDNA datasets of different organisms. The GC content of metazoan and non metazoan sequences showed a normal distribution with a low standard deviation. Unknown and un-informative sequences with a broad distribution of their GC contents produced a significant standard deviation. The GC distribution of both metazoan and non-metazoan sequences (being significantly different, $p<0.05$) co-cluster with the GC profiles of other cnidarian and metazoan organisms or with other bacterial species (*P. aeruginosa* and *M. tuberculosis*), respectively.

3.1.4. Characteristics of the *Hydractinia* transcriptome

Using TBLASTX, the translated *Hydractinia* sequences were compared with the translated cDNAs of different vertebrate and invertebrate model organisms. When comparing the best hits of the *Hydractinia* sequences with both the vertebrate and the invertebrate datasets, it was observed that 153 consensus sequences presented similarity to the vertebrate sequences with more than 10^{10} fold higher significance than to their invertebrate counterparts, while only 18 sequences appeared to be more similar to invertebrate sequences using the same criteria (Fig. 10). Indeed, 28 consensus sequences with a vertebrate homologue but without any hit in the invertebrate datasets were detected. *Vice versa*, four *Hydractinia* sequences were found only in invertebrates (Table 3).

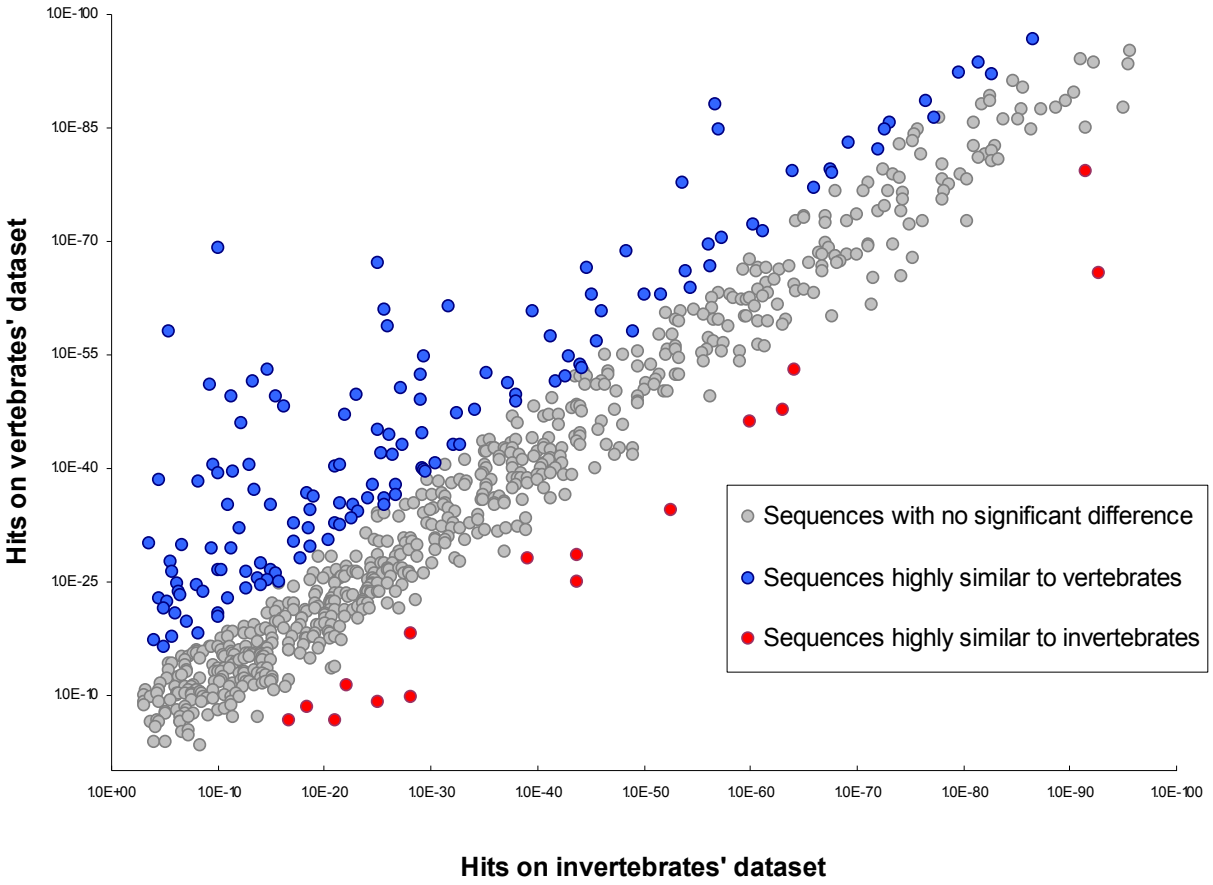


Figure 10 - TBLASTX E-values of the best *Hydractinia* matches to invertebrate and vertebrate cDNA datasets. Considering an acceptance E-value threshold of less than 10^{-3}, the plot includes the consensus sequences best TBLASTX E-values within the range 10^{-3} and 10^{-100} obtained in the comparisons against the vertebrate cDNA datasets of *Macaca mulatta*, *Canis familiaris*, *Rattus norvegicus*, *Gallus gallus*, *Danio rerio*, *Xenopus tropicalis*; and the invertebrate cDNA datasets of *Aedes aegypti*, *Anopheles gambiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*. A significant difference between the E-values was considered when sequences exhibited a 10^{10}-fold more significant similarity to one of the datasets. Sequences present in vertebrates or invertebrates only as well as those with E-values less than 10^{-100} were not shown. This included 28 sequences found only in vertebrates, 4 sequences only from invertebrates and finally 11 and 2 sequences with an E-value less than 10^{-100} with a higher significance similarity to vertebrates or invertebrates, respectively.

Table 3 - *Hydractinia* sequences uniquely shared either with vertebrates or invertebrates

A. *Hydractinia* sequences uniquely shared with vertebrates

Clone name	Ensembl ID entry	Sequence annotation	E-value
HEAB-0024D19	xt_ens_cdna:ENSXETT00000041572	Trefoil factor 1 precursor (pS2 protein) (HP1.A)	1E-28
HEAB-0029B17	dr_ens_cdna:ENSDART00000027368	Hypothetical protein LOC393615	8E-36
HEAB-0031C22	pt_ens_cdna:ENSPTRT00000012985	Glycine amidinotransferase, mitochondrial precursor (EC 2.1.4.1)	8E-59
HEAB-0033J13	hs_ens_cdna:ENST00000331336	KRAB-A domain-containing protein 2.	4E-04
HEAB-0040I14	gg_ens_cdna:ENSGALT00000018072	Adenylate kinase 7	3E-60
HEAB-0040K13	mm_ens_cdna:ENSMUST00000020365	Melanoma associated antigen (mutated) 1	1E-05
tah97g01	xt_ens_cdna:ENSXETT00000046873	AP complex subunit beta 1	4E-10
tah98a12	rn_ens_cdna:ENSRNOT00000004956	Collagen alpha-1(III) chain precursor	7E-20
tah98g04	cf_ens_cdna:ENSCAFT00000028432	Peripheral myelin protein 22	4E-12
tai01f07	xt_ens_cdna:ENSXETT00000001418	Coiled-coil-helix-coiled-coil-helix domain containing 5	4E-14
tai02b03	dr_ens_cdna:ENSDART00000017605	Sperm associated antigen 6	1E-106
tai03d04	dr_ens_cdna:ENSDART00000013591	TRK-fused	7E-52
tai08a11	cf_ens_cdna:ENSCAFT00000037058	Trefoil factor 2 (spasmolytic protein 1)	4E-15
tai08b11	xt_ens_cdna:ENSXETT00000008001	Guanidinoacetate N-methyltransferase.	1E-112
tai08f03	mm_ens_cdna:ENSMUST00000006853	Hypoxia-inducible factor prolyl 4-hydroxylase	3E-08
tai09b12	dr_ens_cdna:ENSDART00000051259	Similar to LOC407707 protein	2E-10
tai11d10	dr_ens_cdna:ENSDART00000010591	WD repeat domain 8.	9E-60
tai13d09	xt_ens_cdna:ENSXETT00000054650	CH41746 (Fragment).	7E-63
tai18f11	xt_ens_cdna:ENSXETT00000047703	DNA (cytosine-5-)-methyltransferase 1	5E-32
tai19h02	xt_ens_cdna:ENSXETT00000027639	Neuropilin-1 precursor (Vascular endothelial cell growth factor 165 receptor)	7E-11
tai20d09	hs_ens_cdna:ENST00000309983	Thiopurine S-methyltransferase (EC 2.1.1.67)	2E-12
tai22h11	mmu_ens_cdna:ENSMMUT00000006486	Hypothetical protein LOC389799	1E-13
tai27a08	xt_ens_cdna:ENSXETT00000039097	Tropomyosin	4E-12
tam53h01	dr_ens_cdna:ENSDART00000087062	Hypothetical protein LOC791183	8E-11
tam54a06	dr_ens_cdna:ENSDART00000062462	Fucolectin precursor	5E-05
tam57f07	xt_ens_cdna:ENSXETT00000016398	Unknown	4E-34
tam59f02	cf_ens_cdna:ENSCAFT00000028627	Uromodulin (uromucoid, Tamm-Horsfall glycoprotein)	2E-05
tam61e02	dr_ens_cdna:ENSDART00000080446	Major vault protein (MVP).	1E-59

B. *Hydractinia* sequences uniquely shared with invertebrates

HEAB-0041C23	ce_ens_cdna:C07E3_1A	Unknown	1E-07
tai25e05	aae_ens_cdna:AAEL003750-RB	Nucleoplasmin	2E-04
tai29g03	dm_ens_cdna:CG9983-RF	Heterogeneous nuclear ribonucleoprotein A1 (hnRNP core protein A1-A)	7E-04
tam61h03	ag_ens_cdna:AGAP009040-RA	Unknown	3E-08

The *Hydractinia* sequences best match on BLASTX comparisons to the vertebrate or the invertebrate cDNA databases are given together with the corresponding E-value. Vertebrate's cDNA datasets included *Macaca mulatta, Canis familiaris, Rattus norvegicus, Gallus gallus, Danio rerio, Xenopus tropicalis*; and the invertebrate's cDNA datasets included *Aedes aegypti, Anopheles gambiae, Caenorhabditis elegans and Drosophila melanogaster*. The ensembl ID entry is provided together with the sequence annotation curated on ensembl and vega databases. Unknown or un-informative sequences presented no protein domain structure annotations.

3.1.5. Unique sequences of *Hydractinia*

Attempting to detect genes present in the *Hydractinia* transcriptome but absent in other cnidarians, the *Hydractinia* sequences were compared by TBLASTX with all available cnidarian sequences including the ESTs of *Acropora millepora*, *Hydra* spp., *Nematostella vectensis* and its genomic DNA data. With an acceptance significance E-value of less than 10^{-3} and excluding all ESTs related to a non-metazoan sequence (discussed above), 23 unique

Hydractinia sequences with a known protein or protein-domain hit were detected (Table 4). Some sequences pointed to the same protein domain hit. However, analysis by specialized BLAST algorithms such as Bl2seq (data not shown) revealed that theses sequences do not have a significant sequence similarity among each other, corroborating that they were not clustered by the Gel assemble program in the sequences analysis pipeline. Regarding the consensus sequences having a non-metazoan match, 393 sequences were uniquely present in the *Hydractinia* dataset.

Table 4 - *Hydractinia echinata* unique sequences

Clone name	Sequence GenBank ID	Protein match ID number at GenBank /Inter Pro	Sequence/domain annotation
HEAB-0027M01	68411965	IPR008412	Bone sialoprotein II
HEAB-0034N17	74135604	IPR002952	Eggshell protein
HEAB-0036J11	74132951	IPR001876	Zinc finger, RanBP2-type
HEAB-0038D19	74134674	IPR005649	Chorion 2
HEAB-0038H17	74134662	IPR006706	Extensin-like region
HEAB-0039H23	74134110	IPR005649	Chorion 2
HEAB-0040M05	74134400	IPR003908	Galanin 3 receptor
HEAB-0042M23	74134684	IPR001841	Zinc finger, RING-type
tah96a10	49453351	IPR006706	Extensin-like region
tah98e04	49451948	IPR002952	Eggshell protein
tah99a03	49453544	IPR007087	Zinc finger, C2H2-type
tai01f07	50347174	gi: 62510506	CHCH5_HUMAN
tai01g09	50347183	IPR006706	Extensin-like region
tai08h10	50351274	IPR000637	HMG-I and HMG-Y, DNA-binding
tai10f09	50348080	IPR007087	Zinc finger, C2H2-type
tai21h03	50351781	IPR005649	Chorion 2
tai32e08	50351456	IPR001152	Thymosin beta-4
tai35e09	50352319	IPR010800	Glycine rich
tai46c12	50697716	IPR007223	Peroxin 13, N-terminal
tam53h06	59829660	IPR007718	SRP40, C-terminal
tam54c10	59829689	IPR002952	Eggshell protein
tam55f08	59829784	IPR006706	Extensin-like region
tam57a05	59829876	IPR007223	Peroxin 13, N-terminal

The left column gives the clone ID number of the *Hydractinia* library. Along is provided identifier ID of the sequence at GenBank as well as the InterPro or Swissprot identifier of the sequences best match and its annotation obtained directly by the BLASTX algorithms (with an E-value acceptance threshold of less than 10^{-6}) or after the Domainsweep analysis.

3.1.6. Searching for genes associated with the marine or colonial characteristics of *Hydractinia*

The few cnidarians that are being used as model systems differ markedly in many aspects of their biology, morphology and life history. Among these cnidarians are solitary as well as colonial species, polyps living in freshwater environment and marine organisms. In addition,

these species have different stem cell systems, reproduce non-sexually or sexually and inhabit different ecological niches. Taking marine *vs.* freshwater and solitary *vs.* colonial as working examples, the cnidarian datasets were analyzed to find genes unique to two different combinations of cnidarians as follows: (i) *Hydra* and *Nematostella* are solitary polyps whereas *Acropora* and *Hydractinia* are colonial. (ii) *Hydra* is a freshwater organism whereas *Hydractinia*, *Nematostella* and *Acropora* are marine animals. Seeking for genes linked to these traits, with TBLASTX algorithms all *Hydractinia* sequences shared with *Acropora* and *Nematostella* but not with *Hydra* were extracted, as well as all sequences present in *Hydractinia* and *Acropora* but missing in the *Hydra* and *Nematostella* datasets. Using the same significance criteria as above (E-values less than 10^{-3}), 11 *Hydractinia* sequences shared by *Acropora* and *Nematostella* were absent in *Hydra*. Only one of the sequences did not have any GO terms or match in the Swissprot-Trembl databases. The 10 remaining sequences were mainly related to metabolism including catalytic activities, protein modification, protein mediated transport, physiological processes, and signal transduction (Table 5A; Table S2A in Additional data 1, section 6.1). In the second analysis, 15 sequences were uniquely found in *Hydractinia* and *Acropora*. Despite the fact that most of these sequences were primarily considered by the BLASTX annotation as unknown or un-informative, Domainsweep and GO analyses helped to determine a functional annotation in some of them. As in the first group of sequences, they were also associated with metabolism (catalytic and biosynthesis), nucleotide binding, signal transduction, and one was related to an intracellular non membrane-bound organelle (Table 5B; Table S2B in Additional data 1, section 6.1).

Table 5 - *Hydractinia* sequences compared to other cnidarians model organisms

Clone name	ID GenBank	Sequence/domain annotation	E-value	GO: Biological Process	GO: Molecular Function
A. *Hydractinia* protein sequences present in *Acropora*, *Nematostella* but not in *Hydra*					
HEAB-0029E05	74134839	Lanin A-related sequence 1 protein	1E-16	GO:0007582	n/a
HEAB-0029J09	74133868	Nuclear protein 1 (p8)	4E-08	n/a	n/a
HEAB-0038N23	74134624	MKIAA0230 protein (Fragment)	1E-41	n/a	GO:0004601
tai09b01	50352378	Guanine nucleotide-binding protein Y-e subunit precursor	2E-09	GO:0008277	GO:0004871
tai11f02	50348136	Malate synthase	1E-91	GO:0008152	GO:0004474
tai11g12	50348149	lysosomal thioesterase ppt2 precursor	2E-45	GO:0006464	GO:0016787
tai20d03	50351692	AP-4 complex subunit sigma-1	2E-08	GO:0016192	n/a
tai33g08	50352245	Isocitrate lyase	2E-72	GO:0008152	GO:0016829
tam56f07	59829849	Cephalosporin hydroxylase family protein	1E-08	n/a	n/a
HEAB-0023B24	68411515	Unknown function	n/a	GO:0005975	GO:0004033
tam53d11	59829628	Unknown function	n/a	n/a	n/a
B. *Hydractinia* protein sequences present in *Acropora* but not in *Nematostella* and *Hydra*					
HEAB-0020F05	68411267	2-c-methyl-d-erythritol 4-phosphate cytidylyltransferase	1E-24	n/a	GO:0008299
HEAB-0024D20	68411599	Response regulator receiver protein	6E-09	n/a	GO:0000166
HEAB-0028A08	68334384	Major facilitator superfamily MFS_1	1E-38	n/a	n/a
HEAB-0028B20	68334404	Fatty-acid desaturase. 2/2007	2E-16	n/a	n/a
HEAB-0037F13	74133658	PcaB-like protein. 2/2007	1E-94	n/a	GO:0016829
HEAB-0039G08	74134978	Signal peptidase I precursor (EC	2E-24	n/a	GO:0000155
HEAB-0042I20	74133750	Glucose-methanol-choline oxidoreductase, N-terminal	n/a	n/a	n/a
HEAB-0020L20	68411323	Unknown function	n/a	n/a	GO:0005884
HEAB-0026O12	68411824	Unknown function	n/a	n/a	n/a
HEAB-0029G01	74134845	Unknown function	n/a	n/a	n/a
HEAB-0036O10	74133537	Unknown function	n/a	GO:0006810	GO:0000166
HEAB-0042L12	74133375	Unknown function	n/a	n/a	n/a
tai07g10	50350972	Unknown function	n/a	n/a	n/a
tai16a08	50352144	Unknown function	n/a	n/a	n/a
tai40g01	50697024	Unknown function	n/a	n/a	n/a

The left column gives the clone ID on the *Hydractinia* library. Along is provided identifier of the sequence at GenBank as well as the annotation obtained by BLASTX comparison to the Swissprot/Sptrembl databases with the corresponding e-value. In case of annotation by domain analyses, the Inter Pro domain ID was provided. The last two right columns show the sequences GO annotations from two main GO categories. For the GO description terms see Supplementary data. Non applicable (n/a) was considered when sequences had no significant match to either domain-Swissprot/Sptrembl or GO databases.

3.1.7. Analysis of selected genes by semi-quantitative RT-PCR

From the unique *Hydractinia* transcripts and the ones shared with one or more cnidarians, a subset was selected to analyze their expression pattern during the life cycle of the hydroid by semi-quantitative RT-PCR. Their transcriptional level was compared by agarose gel electrophoresis and ethidium bromide staining, using actin gene as reference. In the different life stages, the transcripts HEAB-0042L12, Tai20D03 and Tai09B01 showed an invariant and similar amount of mRNA as the reference. The HEAB-0034N17 gene also followed the same transcription pattern, in spite of its low expression level. In the case of Tai16A08, a transcript only shared with *Acropora*, a strong gene expression level was detected in the adult, while a mild one occurred during early embryo, pre-planula and primary polyp stages. Interestingly,

no expression was observed during larvae. Similar was the case of the *Hydractinia* unique transcript Tai08H10 presenting a high abundance in the primary polyp stage, a mild one in early embryo, larva and adult stages, and was undetectable in pre-planula. Finally, the *Hydractinia* transcript Tai11F02, shared only with *Acropora* and *Nematostella*, exhibited a mild expression in pre-planula and adult, a strong one in primary polyp, but was undetectable in early embryo and larva stages (Fig. 11). These three transcripts, having a specific expression pattern, are interesting candidates for further functional analysis.

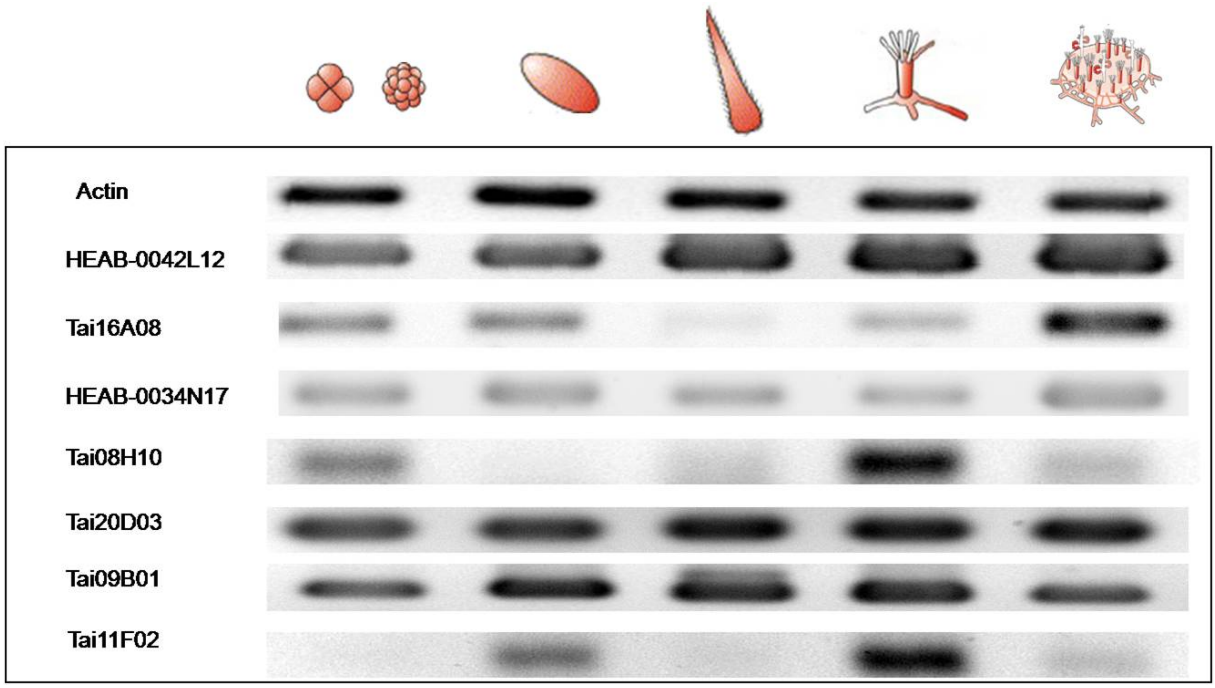
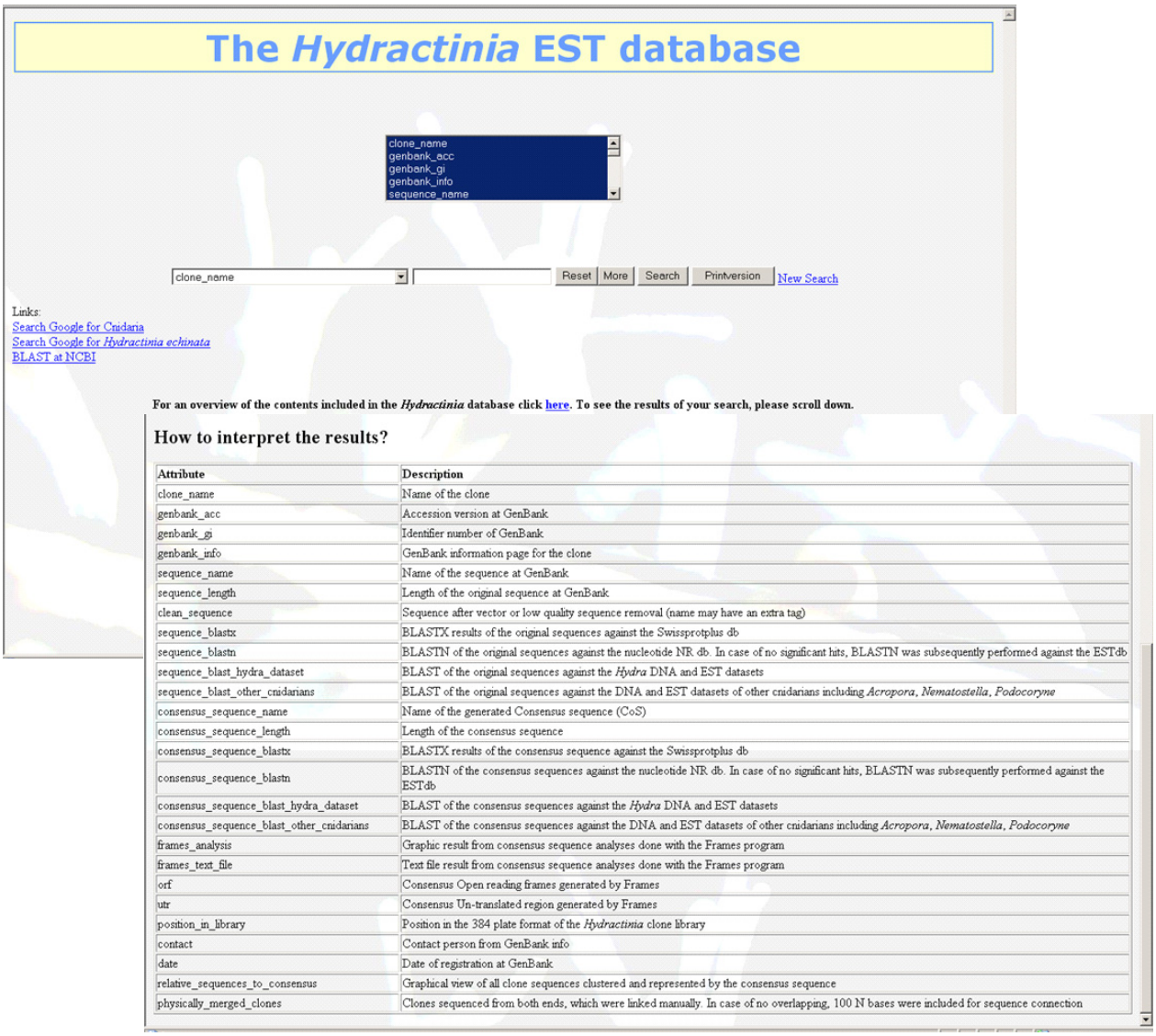


Figure 11 - Semi-quantitative RT-PCR analysis. (sq)RT-PCR products were loaded in an agarose gel to determine their relative level of expression. Actin was considered as reference. HEAB-0042L12 and Tai16A06 sequences are only shared with *Acropora* and the function of both protein products is unknown. HEAB-0034N17 and Tai08H10 are unique sequences of *Hydractinia*, the former carrying an eggshell protein domain and the latter with a DNA-binding activity, respectively. The rest sequences are shared with *Nematostella* and *Acropora* but not *Hydra*. Tai20D03 is annotated as an *AP4S1* homologue, Tai09B01 as a guanine nucleotide-binding protein Y, and Tai11F02 as malate synthase.

3.1.8. *Hydractinia* Database

A database was created in order to optimize the handling of all generated data, including the physical information of each EST clone but also the results of the EST-clustering, the representative consensus sequences, or the BLAST programs (Fig. 12). Searches within the database can be done with GeneBank identification numbers, clones or consensus sequence

names, etc. It is possible to simultaneously query different fields by combining search criteria with "AND" and "OR". Query results are listed on screen, with direct links to the detailed clone or sequence information, which can be easily extracted for further analysis. The EST database can be accessed at: http://www.mchips.org/hydractinia_echinata.html.



Attribute	Description
clone_name	Name of the clone
genbank_acc	Accession version at GenBank
genbank_gi	Identifier number of GenBank
genbank_info	GenBank information page for the clone
sequence_name	Name of the sequence at GenBank
sequence_length	Length of the original sequence at GenBank
clean_sequence	Sequence after vector or low quality sequence removal (name may have an extra tag)
sequence_blastx	BLASTX results of the original sequences against the Swissprotplus db
sequence_blastn	BLASTN of the original sequences against the nucleotide NR db. In case of no significant hits, BLASTN was subsequently performed against the ESTdb
sequence_blast_hydra_dataset	BLAST of the original sequences against the *Hydra* DNA and EST datasets
sequence_blast_other_cnidarians	BLAST of the original sequences against the DNA and EST datasets of other cnidarians including *Acropora*, *Nematostella*, *Podocoryne*
consensus_sequence_name	Name of the generated Consensus sequence (CoS)
consensus_sequence_length	Length of the consensus sequence
consensus_sequence_blastx	BLASTX results of the consensus sequence against the Swissprotplus db
consensus_sequence_blastn	BLASTN of the consensus sequences against the nucleotide NR db. In case of no significant hits, BLASTN was subsequently performed against the ESTdb
consensus_sequence_blast_hydra_dataset	BLAST of the consensus sequences against the *Hydra* DNA and EST datasets
consensus_sequence_blast_other_cnidarians	BLAST of the consensus sequences against the DNA and EST datasets of other cnidarians including *Acropora*, *Nematostella*, *Podocoryne*
frames_analysis	Graphic result from consensus sequence analyses done with the Frames program
frames_text_file	Text file result from consensus sequence analyses done with the Frames program
orf	Consensus Open reading frames generated by Frames
utr	Consensus Un-translated region generated by Frames
position_in_library	Position in the 384 plate format of the *Hydractinia* clone library
contact	Contact person from GenBank info
date	Date of registration at GenBank
relative_sequences_to_consensus	Graphical view of all clone sequences clustered and represented by the consensus sequence
physically_merged_clones	Clones sequenced from both ends, which were linked manually. In case of no overlapping, 100 N bases were included for sequence connection

Figure 12 - The *Hydractinia* EST database web interface and information page. The database can be accessed on the web at http://www.mchips.org/hydractinia_echinata.html.

3.2 *Hydractinia* cDNA-microarray

3.2.1. Construction of the *Hydractinia* cDNA microarray

The *Hydractinia* chip-library was amplified by PCR in a 96-well plate format. To assess the quality and quantity of the PCR products, an aliquot of the total PCR reaction volume (100 µl) was analyzed by agarose gel electrophoresis and ethidium bromide staining (Fig. 13). From the 9,216 clones present in the library, 87 % were successfully amplified.

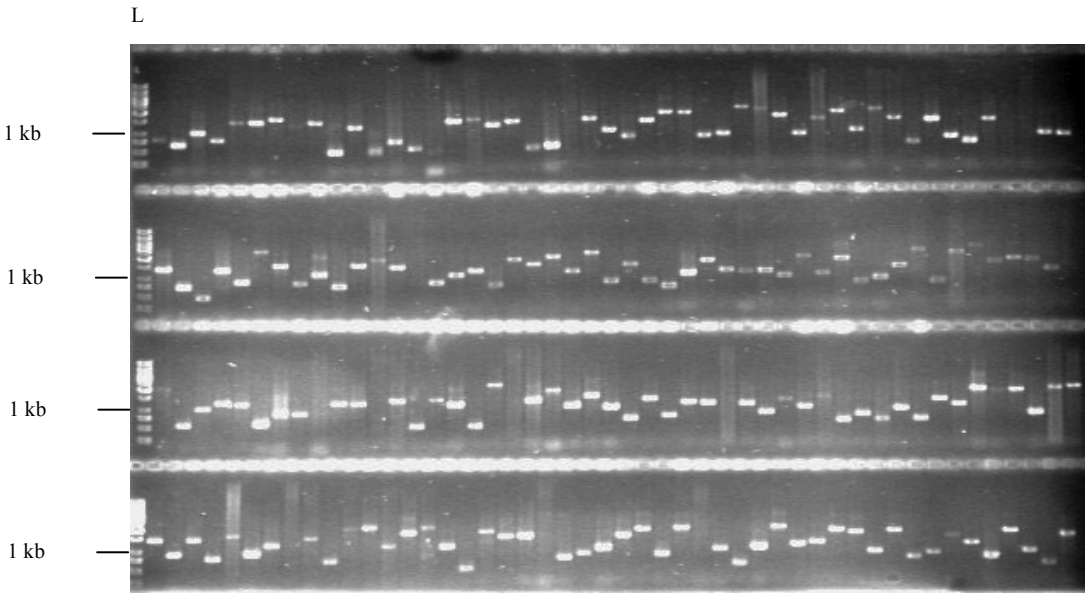
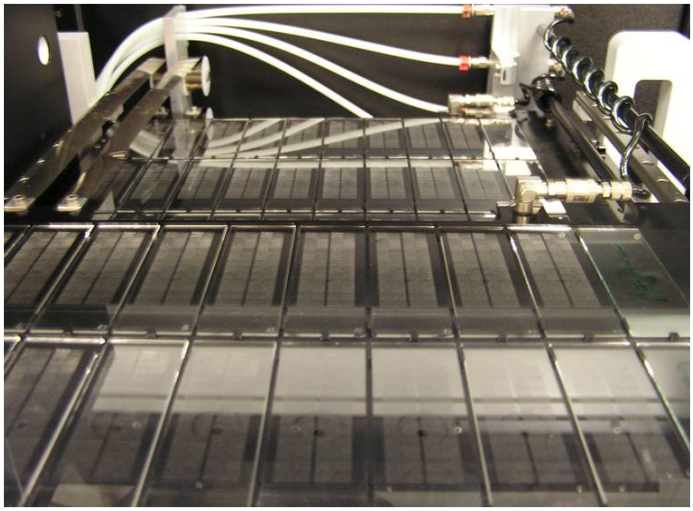


Figure 13 - PCR amplification of the *Hydractinia*-chip library. Two 96-well plates where amplified in a 100 µl PCR reaction, from where 2 µl were loaded in a 1.5% agarose gel. The ladder L used is the GeneRuler 1k DNA ladder.

Considering that each PCR probe was printed out in duplicate, the cDNA microarray comprised 19,200 spots with a diameter of approximately 100 µm and separated from each other by 140 µm (Fig. 14). The spots were organized in 48 blocks of 400 spots each, with 20 spots in X axis and 20 spots in the Y axis (Fig. 14B). To account for local intensity differences within the array, the duplicate (primary and secondary) spots were distributed in different blocks. With the first visit of the pins to the PCR source, all primary spots were delivered on the left side of the slide (blocks 1 to 24). Then, in a second visit, the pins delivered the secondary spots on the right side of the slide (blocks 25 to 48). Thus, in one aminosilane slide of 25 x 75 mm two identical arrays (primary and secondary) were printed out (Fig. 14B). Homogeneity in the size and shape of the spots was achieved with the inclusion of 6 ‘pre-spot slides”, which were visited 8 times by the pins directly after the up-

take of the probe. Per printing run the robot was able to produce 108 cDNA-microarrays slides, from which 80% of them were used for hybridization experiments. The 10 first and last microarrays-slides were used for optimization procedures.

A



B

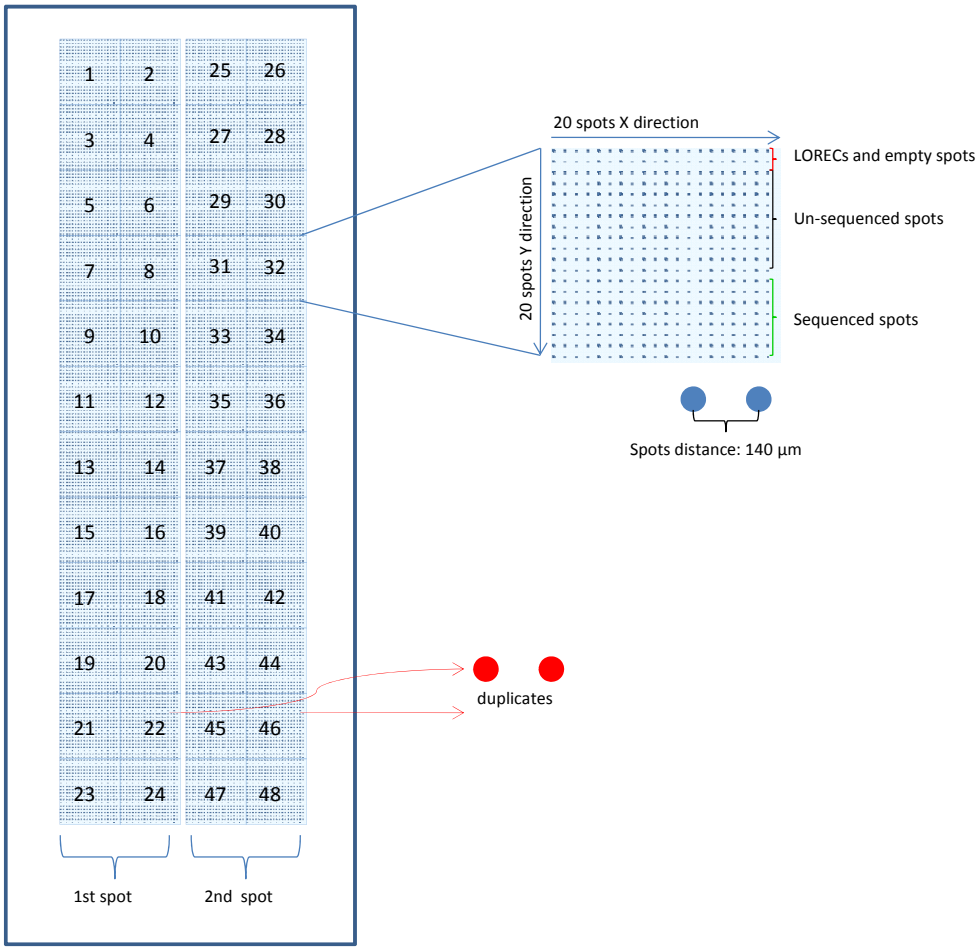


Figure 14 - Construction and design of the *Hydractinia* microarray. A. In one single run, the MicroGrid II robot printed out 108 slides with 9,600 spots in duplicate. **B.** Schematic representation of the array. Within each block, the upper spot lines correspond to the external and null controls, followed by the un-sequenced PCR products. The lower part of the array contains the already sequenced and annotated *Hydractinia* genes.

3.3 Transcription profiling experiments

3.3.1. Searching for i-cell related genes in *Hydractinia* – the mitomycin microarray experiment

3.3.1.1. The mitomycin treatment

The *Hydractinia* microarray was used to analyze the gene expression profile of colonies depleted from their i-cells using the antibiotic mitomycin-C, and after their recovery from the treatment. For this purpose, three female *Hydractinia* colonies -F0, FM and K12, the last two having the same genotype- were incubated with mitomycin-C as described in section 2.2.1.2. A clone member of the K12 colony served as control. For each colony, three small subclones were used as biological replicates.

At 24 hours post-treatment, the first drug effects were observed in the F0 colonies assessed by a continuously contraction of the tentacles. In the following 24 hours the same effect but milder was observed in the K12 colonies, while the FM colonies still looked unaffected. The latter showed a relatively mild effect at 72 hours post-treatment (Fig. 15). By this time the K12 polyps started to resorbe their tentacles, while the rest of the colony seemed intact. However, F0 colonies exhibited a dramatic drug response. Their tentacles completely disappeared and polyps started to be resorbed or degraded. In addition, a significant amount of death tissue detached from the colony (Fig 15).

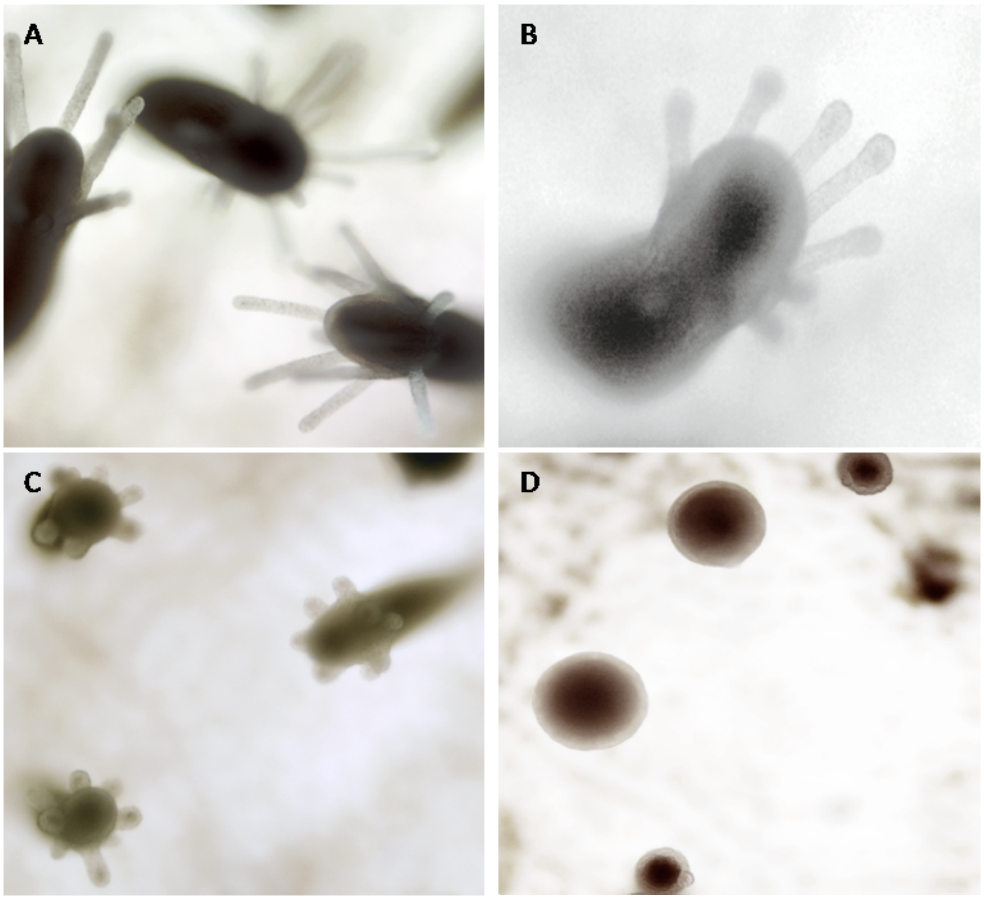


Figure 15 – The mitomycin phenotype effect at 72 hrs post-treatment. A. Wild type **B.** FM colonies showed tentacle-tip contraction, representing a mild response to the drug. **C.** K12 colonies showed a markedly tentacle contraction and some polyps exhibited signs of tentacle resorbtion. **D.** F0 colonies presented polyps depleted of tentacles and signs of polyps resorbtion.

At 96 hours post-treatment, small explants of treated and control colonies were fixed and stained to assess drug i-cell depletion (see methods, section 2.2.1.4). I-cells, rich in ribosomes, were stained with basic blue dyes as May-Gründwald and Giemsa. Cytological examinations of the colonies stolonal compartment revealed that in F0 colonies most if not all i-cells were eliminated. In addition, no nematoblasts were found and a significant amount of apoptotic material was observed (Fig. 16D). This suggests that the harsh treatment did not specifically target the i-cell population of the F0 colonies. In regard to the K12 colonies, the treatment resulted in strong i-cell depletion, identifying only a couple of i-cells in all analyzed fields (Fig. 16C). However, other cell types, as nematoblasts and epithelial cells, seemed unaffected. In FM colonies several i-cells were detected, but its density still presented a significant difference in comparison to the untreated control (Fig. 16A-B).

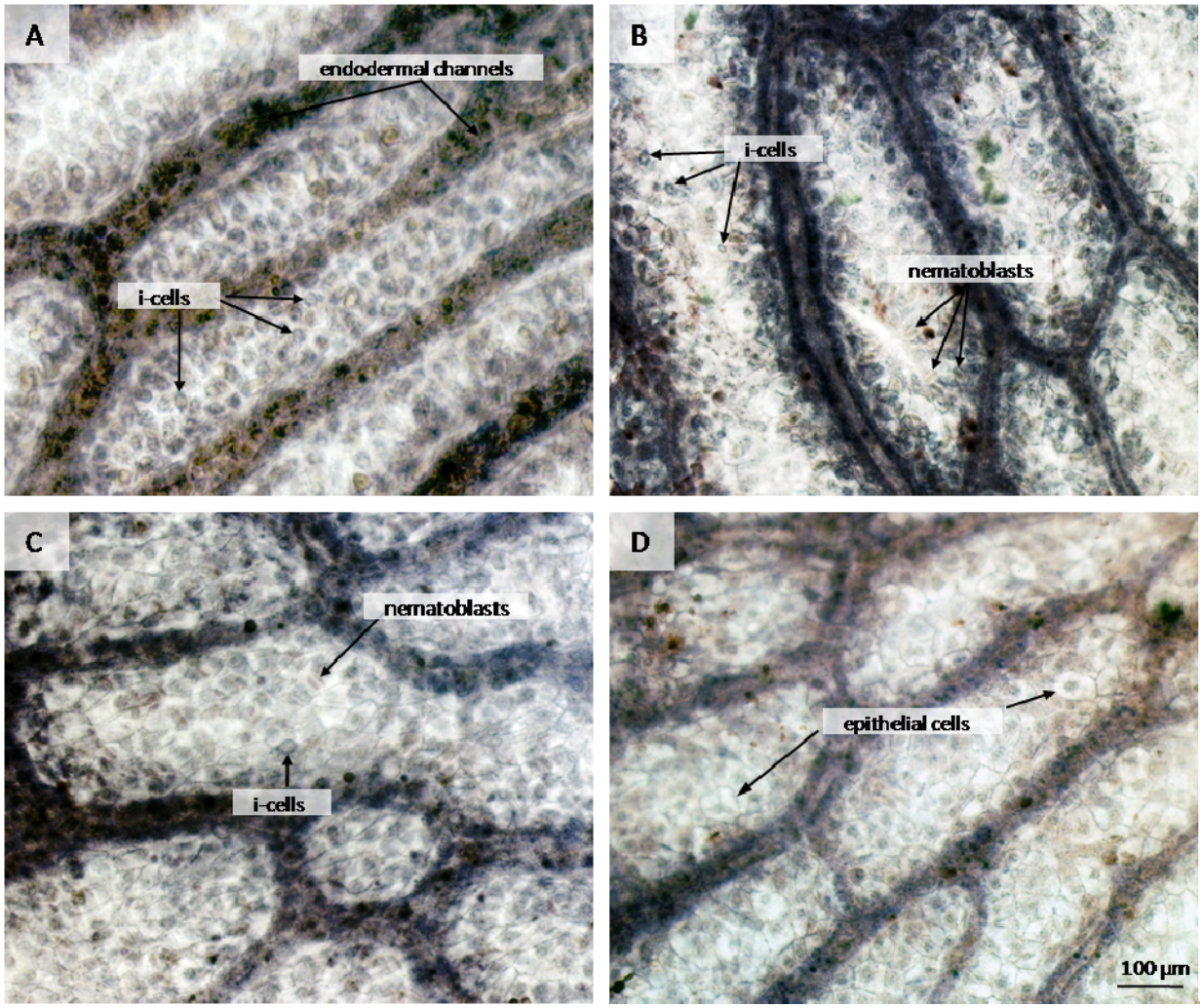


Figure 16 - Mitomycin i-cell depletion at 96 hrs post treatment. Whole mount preparation of the stolon plate of control and mitomycin treated colonies stained with May-Grünwald and Giemsa. **A.** high number of i-cells can be detected throughout the interstitial space of wild type colonies. **B.** FM colonies showed mild i-cell depletion, which is still significant in comparison to the control. **C.** Only few i-cells were detected in K12 stolons. As in the case of the FM colonies, nematoblasts seemed unaffected by the drug. **D.** F0 colonies presented a complete i-cell and nematoblast depletion.

Once the success of the drug treatment was determined, explants from the middle and the edge of each colony were extracted. Immediately, RNA was isolated for the microarray experiments. Then, to recover the wild type phenotype, the corresponding donor-explants were grafted to the central and peripheral region of the treated colonies. After three days, the donor stolons started to weakly fuse with the FM and K12 colonies. However, this did not occur in the F0 colony and therefore, no donor i-cell was delivered. Consequently, the colony followed a complete polyp re-absorption and was unable to bud new ones. The drug treatment was too strong for the F0 colony, which after two weeks, showed signs of stolon plate degradation and necrosis. This colony was discarded for any further analysis.

K12 colonies maintained their condition for up to 3 weeks. Despite that the colonies did not show any signs of degradation, no new polyps budded. The colony-donor junctions were weak and continuously disrupted. This suggests that i-cell migration, if occurred, was not constant. Thus, the polyps that survived the drug treatment were probably able to continue feeding the colony. Nevertheless, at the 4th week the colony started to resorb polyps and degradation of stolon tissue was observed.

In contrast, the FM colony presented stable donor-colony junctions and started to bud new polyps at the 3rd- 4th week. For this colony the drug treatment had a milder effect, and therefore, presented better chances for survival. After 4 weeks, the wild type phenotype of the FM colony was recovered in different parts of the stolon mat (Fig. 17). Interestingly, the budding of new polyps was not exclusive to the place where the donor-explants were grafted. Finally, RNA was only isolated from active polyps budding areas.

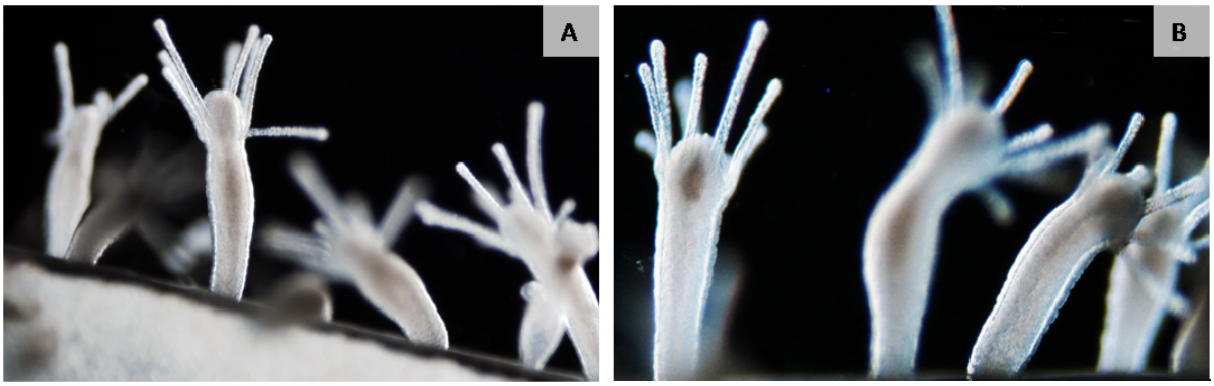


Figure 17 - Recovery of the FM mitomycin treated colony. The FM colony budded new polyps in different spots throughout the colony. At these areas, no difference was observed in comparison to the wild-type colony. **A.** wild type colony. **B.** An active polyps budding area of the recovery FM colonies.

3.3.1.2. Quality control of the isolated RNA

The quality of the RNA sample used in a microarray experiment is critical to obtain meaningful gene expression data. Therefore, only high quality RNA samples with integrity values (RIN) above 6,5 and absorbance ratios of at least 1,9 were selected.

For the mitomycin microarray experiment total RNA was isolated from the control condition (untreated colony) and from the FM milder and K12 stronger i-cell depletion phenotypes at 96 hours post treatment. Three biological replicates were used per sample. In addition, the recovery phenotype of the FM colony (FMR) was represented by RNA isolated from new polyps budding areas at 4 weeks post treatment. To accomplish a higher RNA yield from each

tissue, several RNA isolations were done in parallel (Fig. 18). Finally, for each sample condition the RNAs were pooled and their concentration was determined.

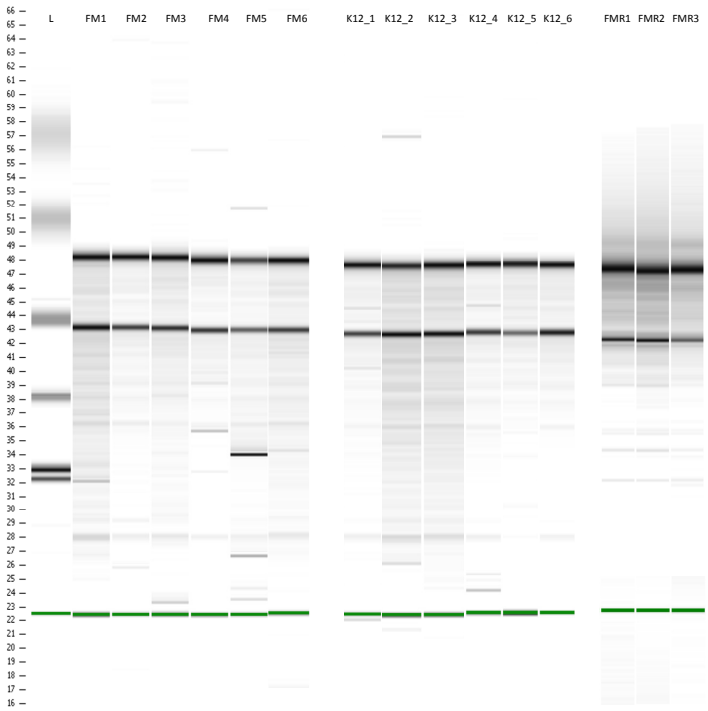


Figure 18 - analysis of the RNA quality. For each condition, RNA samples having a RIN ≥ 6.5 and an abs. ratio (260 nm/280 nm) above 1.9 were pooled. This was done with six FM, six K12 and 3 FM recovery (FMR) samples. In the case of the wild type colonies, 8 high quality samples were pooled (not shown). The two main bands in the figure correspond to the 28s and 18s rRNA.

3.3.1.3. Labelling of RNA samples and microarray experimental design

Each RNA sample to be hybridized into the array was labelled with Cy3-dCTP or Cy5-dCTP through reverse transcription (RT) (see methods, section 2.2.5.2). For the labelling reaction different quantities of RNA were tested, without finding any difference in the product yield if 7,5 or 15 µg were used. Therefore, for all microarray experiments, 7,5 µg of total RNA were defined as the starting material for RT, producing between 1.5 and 2.5 µg of labelled cDNA.

To prevent bias due to preferential label incorporation, an even microarray design including dye-swap was followed (Fig. 19). In addition to the duplicate spots of each gene per array, error measurement and noise were addressed by six technical replications per sample hybridization. This means that in a profiling experiment of two RNA samples, 12 data points per gene were generated for each condition.

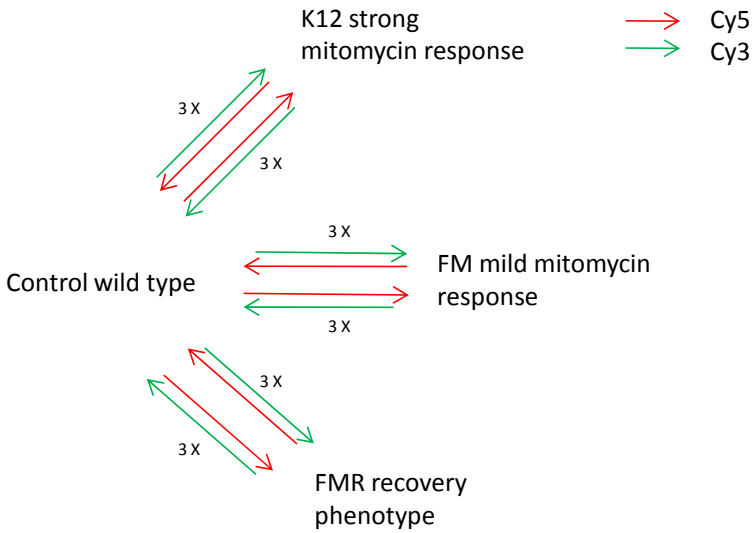


Figure 19 - Double reference design of the mitomycin microarray experiment. Duplicate spots per gene in the array generated 2x data points per gene, which can be used for quality control of individual arrays. In addition, in each competitive comparison, each sample was labelled 3X with Cy3 and 3X with Cy5 (dye swap).

3.3.1.4. Signal detection and quantification of the hybridizations

All hybridizations comparing a sample condition against a control were performed in parallel and with the same amount of labelled material. For a successful hybridization at least 1.5 µg of labelled cDNA per sample were needed. Hybridized slides were scanned at different PMT levels and the signal intensities were quantified and analyzed with the Genepix software. We selected all images in which the brightest pixels were just below the level of saturation. By this, the sensitivity of the image analysis for the less bright pixels was increased. This resulted in the use of different PMT levels for the Cy3 and Cy5 channels.

The Genepix software quantifies the signal intensity of each gene printed on the array. By overlying the image of the Cy3 and Cy5 intensity channels, it was possible to have a rough estimation of the number of differentially expressed genes. In neither of the hybridized slides unspecific hybridization to the external controls LORECs occurred, since all these fragments produced equal signal intensity to the background (Fig. 20).

Background and foreground intensities were analyzed to determine local intensity variation on the array surface. All arrays showing bad quality hybridizations, unreliable spots, artefacts, bad reproducibility, or a strong background were eliminated from further analysis. From the 22 randomized microarray slides hybridized with mitomycin and control labelled material, 16 (73%) scanned images were selected for gene expression analyses. All information, including

signal intensity values of the Cy3 and Cy5 channels, local background data, pixels, saturation levels, etc. were exported as a table (.GPR) for normalization.

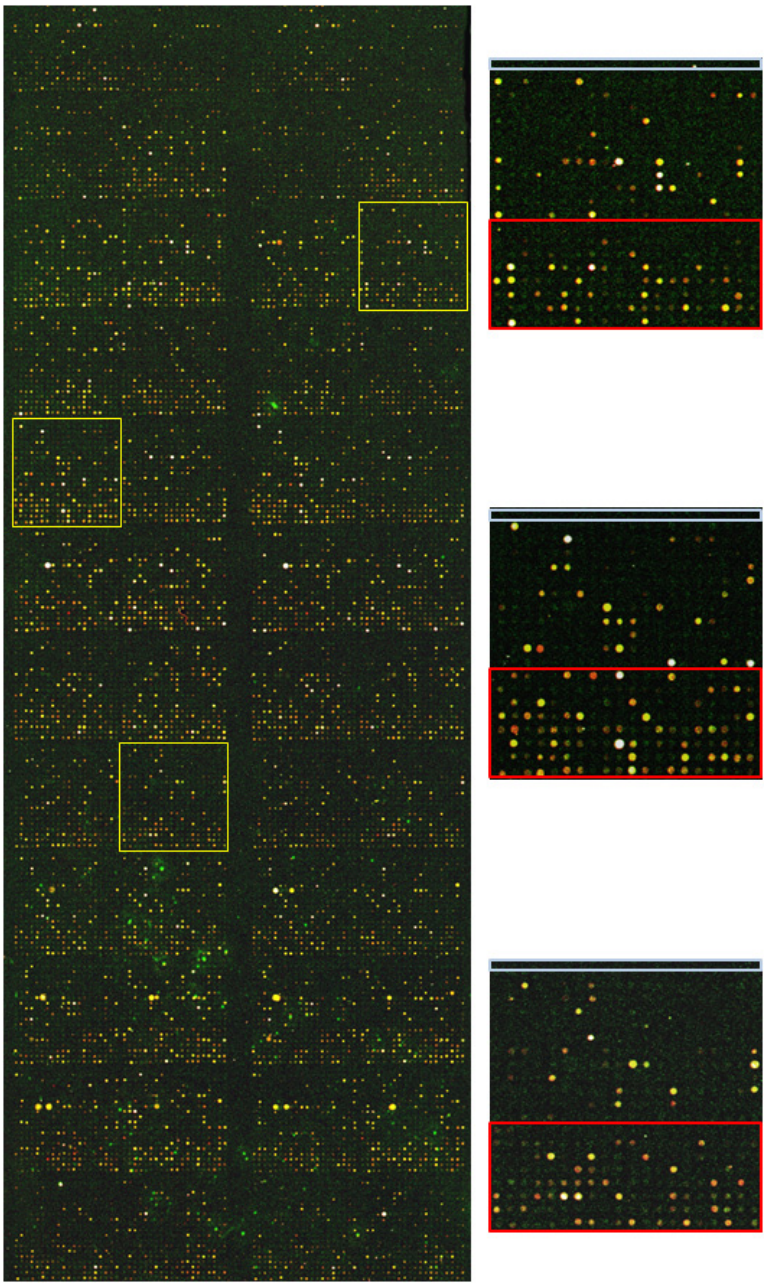


Figure 20 - The *Hydractinia* microarray. Labelled-cDNAs from FMR and wild-type colonies were co-hybridized in the array. The left scanned image correspond to the complete cDNA microarray. Yellow boxes represent the block borders. Within each block it is possible to observe that the sequenced PCR products –red box- are highly hybridized. The blue box corresponds to external and negative controls. White spots resemble signal saturation.

3.3.1.5. Normalization and filtering of the microarray data

In a two-colour array experiment one of the measurements (Cy3 or Cy5) corresponds to the reference. Thus, there are an equal number of control and condition measurements. Since the same reference (wild-type colony) was used in all hybridizations, it was possible to combine all experimental conditions (FM, FMR and K12) in one analysis.

The M-CHiPS software package was used for data normalization and analysis. The entire gene set printed on the array was used for the fitting normalization algorithms, since many of them are not differentially expressed. The performance of the normalization was assessed by plotting the fitted intensities of the condition and the control measurements together with the logarithmic regression curve [65]. All 16 normalized hybridizations presented a regression curve with correlation coefficients above 0.85 (data not shown).

Practically, only a limited number of genes can be followed up in a biological study. Therefore, after normalization, we selected and ranked the genes according to their good evidence of being differentially expressed. First, all genes above the detection limit *i.e.* exhibiting saturation effects, were filtered out. This resulted in the elimination of 40 genes. However, in an array experiment the majority of the genes are not expressed to a measurable level. Thus, in a second filtering step, we selected 2,614 genes exhibiting a considerable absolute expression level (*i.e.* with median of fitted intensities > 1.000 AU) at least in one of the analyzed conditions. The number of replicates in each experimental condition (12 data points per gene per condition) allowed a robust statistical analysis (SAM). This was used to select 167 genes with a significant differential gene expression (corrected p-value < 0.05) between the control measurements and at least one of the other conditions (Fig. 21). Finally, we selected 162 genes (1,8% from the whole gene-set) displaying a minimum of 2-fold difference in their expression level. Plots of the Log_{10} fitted intensities of all three conditions against the control showed, that most of the genes are tightly clustered along the diagonal line (Fig. 21A). This suggests a high correlation between the two channel intensity data. However, there is a significant subset of outlier-genes with nonlinear relationships between the Log intensities. They represent successfully selected transcripts that are significantly differentially regulated in the experimental conditions analyzed (Fig. 21B)

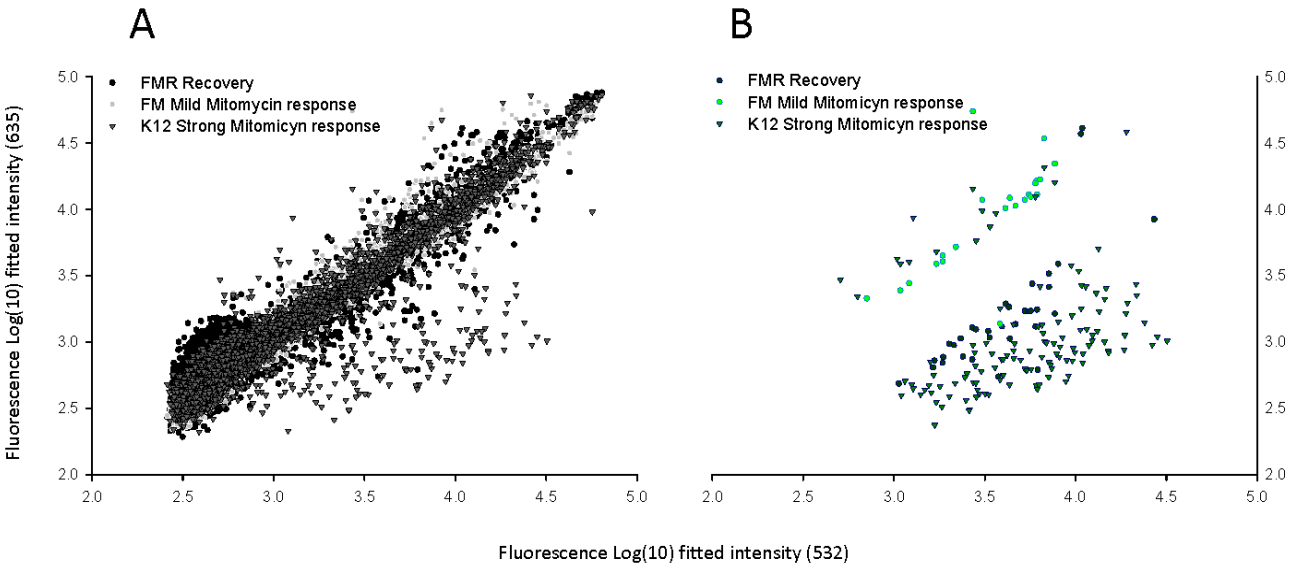


Figure 21 - Pair-wise comparison of the fitted intensities from the conditions FM, FMR and K12 with the reference control. A. Scatter plot of Log_{10} R (red) vs. Log_{10} G (green) including all genes and **B.** only the filtered ones. The differentially expressed genes stand clearly out of the main cloud of un-regulated genes.

3.3.1.6. Correspondence analysis

For data interpretation, the selected differentially expressed genes were analyzed with different applied statistic methods. Correspondence Analysis (CA), available in the M-CHiPS software package, was used to determine the association both within and between genes and conditions (section 2.2.9.2). In Figure 22, each hybridization separated into primary and secondary spot (duplicates within array) is represented as colour boxes. Condition replicates (same colour boxes) were clustered together and each cluster-condition was properly separated between each other.

The plot evidently shows that the K12 condition (pink), representing a strong i-cell depleted phenotype, spread far away from the wild type control (red). FM condition (green) follows another direction, between K12 and FMR hybridizations, correlating to a mild i-cell depletion. The recovery phenotype (FMR, blue) was distributed in a completely opposite direction to the K12 hybridizations, providing the higher distance in the biplot (Fig. 22). In a broad view it is possible to consider a closer relationship between the wild type, FM and FMR phenotypes. For an easy interpretation, lines representing standard coordinates of the measurement conditions were introduced into the plot. These lines were generated by the addition of hypothetical virtual gene-vectors that have an ideal transcriptional profile for each condition. The "real" selected genes are visualized as gray dots. Thus, the association between genes and

conditions can be determined by the position of the genes with respect to the standard coordinate lines and their distance from the centroid [73]. Genes located in the direction of a particular standard coordinate line show a strong expression in such condition. The further away from the centroid, the stronger is the association. Consequently, down-regulated genes appeared in the opposite site of the centroid. In the biplot, it is possible to observe that most of the genes were down-regulated in the K12 condition (Fig. 22).

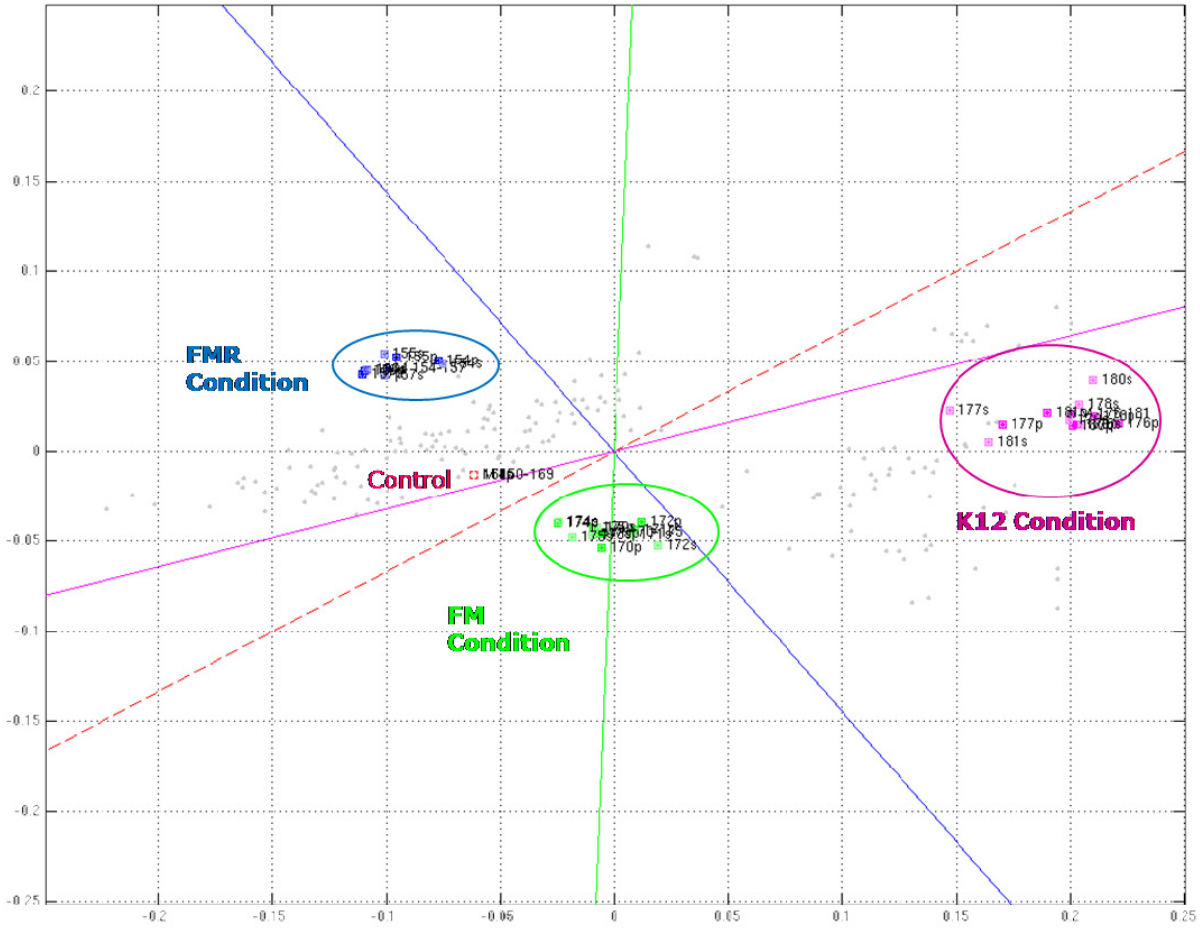


Figure 22 - Correspondence analysis of FM, FMR and K12 conditions. All 162 selected genes are represented as gray dots. The biplot clearly shows four directions corresponding to FMR, FM, K12 and control conditions. The duplicate spots of the array, primary (p) and secondary (s), are displayed with the hybridization numbers and in light and dark colours, respectively. Lines follow the direction of the standard coordinates of the condition medians with the respective colours.

3.3.1.7. Hierarchical clustering

The relationships between the biological samples were analyzed based on their different gene expression patterns, represented by the logarithmic ratio between the median signal intensities of the conditions and control measurements (Conditions/Control) [71]. For this purpose, the TIGR Multiexperiment Viewer (MeV) was used.

In the mitomycin experiment, the values of the Log_2-transformed expression ratios varied between 5 and -5. K12, FM and FMR experimental conditions presented expression ratio medians of -2.1, -0.1 and -0.4, respectively. For an easy visualization of the data, the logarithmic ratios were represented as a colour-code heat map. Relative to the wild type control, green shading indicates decreased gene expression, red shading indicates increased gene expression and black was used in case of no regulation (Fig. 23).

Based on the premise that co-expression is a result of co-regulation, genes were grouped into expression clusters. Ordering genes into meaningful groups provides the genetic fingerprint of a particular condition, allowing the extraction of gene networks and the functionality of even unknown genes. The selected microarray data was clustered using two different methods.

First, a hierarchical clustering (HCL) algorithm using Pearson correlation as a distance metrics was applied. The HCL resulted in a tree representation of the data, whose leaves are the input patterns and the nodes represent a hierarchy of groupings [65]. In order to reduce the complexity of the tree and extract meaningful information of co-regulated genes, an inter-node distance threshold of -0.99 was applied. This resulted in 15 main nodes or clusters (Fig. 23). Elements on nodes, with distances below this threshold, can be considered as one entity. Correlating the results from the CA biplot, most of the genes can be grouped into two main clusters with a high down-regulation expression profile for the K12 condition. These two clusters differentiate between each other, because one of them presented a marginal gene up-regulation in the FM condition. The rest of the genes can be subdivided into small groups of two to 13 genes per clusters with different expression profiles. Experimental conditions were separated into the FMR node and the FM - K12 cluster, again supporting the CA (Fig. 22-23).

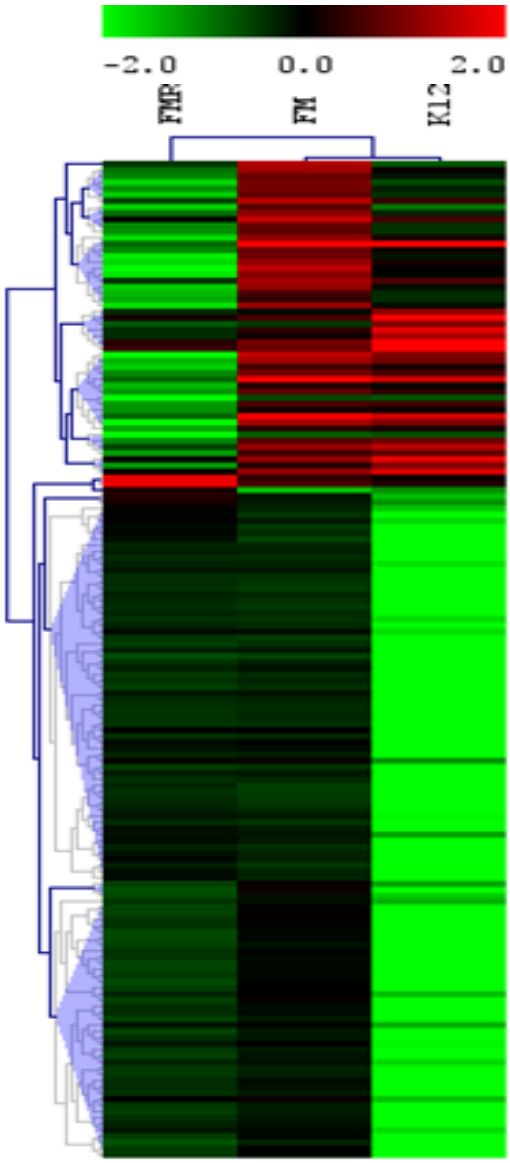


Figure 23 - HCL analysis of the mitomycin microarray data. All 162 selected genes are individually ordered as rows, while the samples are ordered in columns. HCL was used to group the data into 15 main clusters or nodes. They are represented by dark blue branches and a translucent wedge from that node to all enclosed elements. Sub-nodes, below the cluster distance threshold are shown as light gray branches. The length of the branches is proportional to the distance between the nodes. The scale followed up the median of the data, between -2 and 2.

3.3.1.8. Figure of Merit algorithm and k-means clustering

We sustained the analysis of the microarray data with k-means clustering algorithm (KMC). First, the predictive power of the KMC in generating the clusters was assessed using a Figure of Merit (FOM).

FOM values were calculated for 40 clusters using KMC and both variables were plotted to define the optimal clustering parameters for k-means (Fig. 24). In the first k-means runs, the

value of the adjusted FOM drastically decreased. At 9 clusters, the slope of the curve changed to a smooth profile. However, the FOM number still presented a significant difference if 10 or 20 clusters were used. For this microarray data, it was considered that KMC performs optimally for 15 clusters. Additional clusters will not provide meaningful information. This was also empirically demonstrated, by testing k-means using 20 and 25 clusters. This resulted in several single-gene (singletons) or even empty clusters.

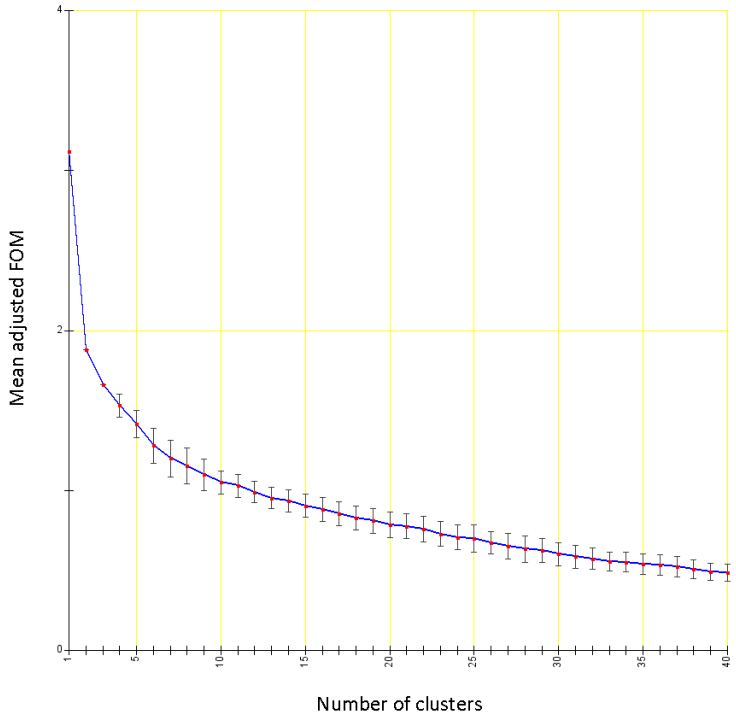


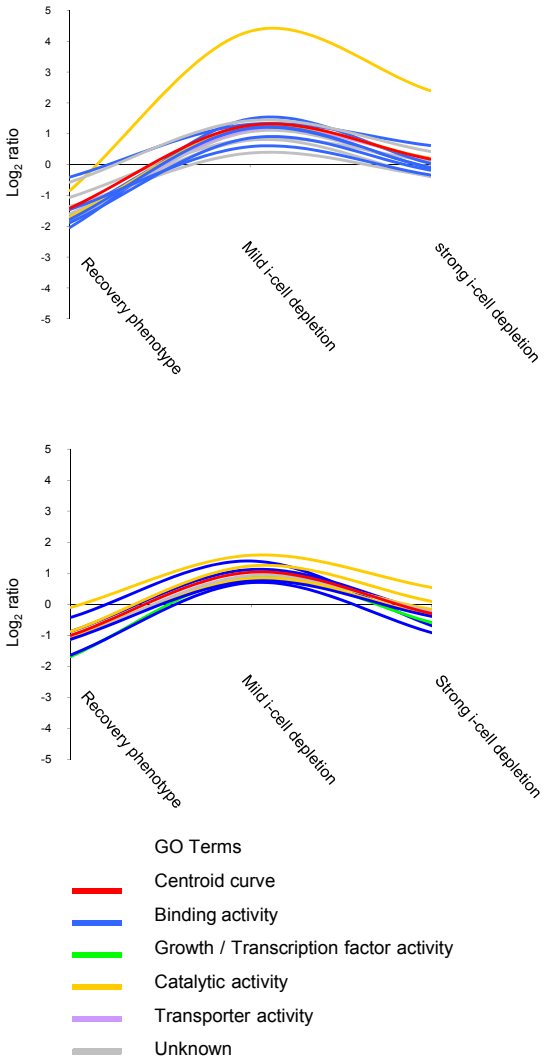
Figure 24 - FOM analysis for k-means algorithm. FOM helps to define the best parameters to use for the k-means clustering algorithm. The lower is the FOM value, the higher is the predictive power of k-means. 200 FOM iterations were performed in the analysis. The mean FOM values are showed as red dots with the corresponding standard deviation.

In KMC there is a partition of the dataset into defined clusters. From the FOM analysis we defined 15 clusters for k-means (Fig. 24). Again Pearson correlation as distance metrics was used. For an easy visualization of the clustered data, the Log_2-transformed expression ratios (conditions/control) were plotted for the recovery (FMR), mild (FM) and strong i-cell depletion (K12) phenotypes (Fig. 25-28). K-means successfully distributed the genes into clusters having a similar transcription profile. Pearson correlation grouped genes with different intensity levels but with same expression patterns. Besides confirming the results of HCL, k-means showed a better distribution of the clustered genes. Redundant genes were grouped in the same cluster or distributed in clusters with a similar expression pattern, only differentiated due to a small variation in the weight of the cluster curves. This is the case of RNA-binding protein 12b, mini-collagens or genes with egg-shell domains in clusters 1, 8 and

5 respectively (Fig. 25-27). In order to extract all possible information from the expression kinetics of the three conditions analyzed, a GO-term colour-code annotation was added to the graphic representation of the data. Subsequently, clusters with genes exhibiting a similar and specific expression pattern for a particular condition were grouped for individual analysis. A selection of the most interesting clusters is described below. The rest of the clusters are provided in the Additional data 2 (section 6.2).

3.3.1.9. Genes up-regulated in organisms mildly depleted from i-cells (FM condition)

A total of 24 genes, grouped into two clusters, were highly regulated in the FM condition (Fig. 25). In cluster 1, the centroid curve (red) – which represents the mean expression profile of the cluster– reached a 3-fold up-regulation. This average expression level could be affected by the extreme activation of the Gluthation S- tranferase gene. In cluster 12, most of the genes followed an expression profile curve with a 2-fold up-regulation in the FM condition (Fig. 25). In both clusters, gene expression profiles were accompanied by a down-regulation in the FMR phenotype and almost no regulation in the K12 condition. GO annotation revealed that the majority of the genes presented a binding activity. Interestingly, several lectin genes appeared in these clusters and a putative Bone morphogenetic protein-4 (*Bmp-4*) homologue was identified.



Cluster 1 (14 genes)

Clone ID	Sequence Annotation
HEAB-0035E15g1.con	RNA-binding protein 12b
tai04c03x1_PT_O.con	Glutathione S-transferase
tai47g07y1.con	GO:0000166 F nucleotide binding
HEAB-0020F15g1_PT.con	RNA-binding protein 12b
HEAB-0044B02	RNA-binding protein 12b
HEAB-0045D05	IPR002952 Eggshell protein
HEAB-0046D08	RNA-binding protein 12b
tai12h10_J.con	Rhamnose binding lectin STL4
tai23d10y1.con	GO:0016787 F hydrolase activity
tai48h01y1.con	RNA-binding protein 12b
tai26c04y1.con	Rhamnospondin 2 (Fragment)

Cluster 12 (10 genes)

Clone ID	Sequence annotation
HEAB-0027G10	Predicted protein IPR006210; EGF
HEAB-0028P12g1.con	Protein putative kinase
HEAB-0029K06g1.con	Putative Bone morphogenic protein 4
HEAB-0034H12g1.con	L-rhamnose-binding lectin
HEAB-0041M08g1.con	Rhamnose-binding lectin OLL
HEAB-0041P11g1.con	Conodipine-M alpha chain
HEAB-0049C03	IPR016090 Phospholipase_A2
tai15f07y1.con	Protein kinase
HEAB-0020E07g1.con	8(R)-lipoxygenase

Figure 25 - Genes highly expressed in FM condition. Expression level of each condition was referenced to the control using Log_2 ratio. Recovery phenotype corresponds to FMR condition. Mildly and a strongly i-cell depletion represent FM and K12 condition, respectively. Logarithmic is in units of 2-fold changes, *e.g.* Log_2 ratio = 0 corresponds to an equal gene expression between the condition and the control. The gene identification ID and annotations are listed in the table. Gene annotation was performed through BLAST, GO and Domain analysis. Unknown genes were plotted in gray colour, but are not listed in the table.

3.3.1.10. Genes highly down-regulated in organisms strongly depleted from i-cells (K12 condition)

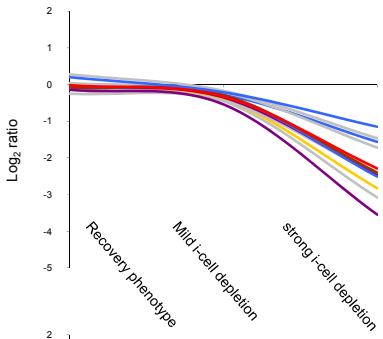
Most of the selected genes (46%) were distributed in four clusters exhibiting a dramatic decrease in their expression level in the K12 condition, reaching in average a 6-fold down-regulation with respect to the control. In contrast, the expression profiles of the FM and FMR conditions were not affected (Fig. 26). All these sequences were also detected with the HCL algorithms, but grouped into just two main clusters (Fig. 23). Thus, k-means was more sensitive to identify similar patterns between these sequences and provided a better

representation of the data. These clusters contained several collagen related genes; including genes encoding for collagen like proteins, mini-collagens, and genes with collagenase domains. The different identified mini-collagen genes showed no similarity between their sequences. In clusters 2 and 8; most of the sequences presented a binding, transport or structural activity, while in cluster 15; genes with a catalytic activity were highly represented.

3.3.1.11. Genes down-regulated in the recovery (FMR) and strongly i-cell depleted phenotype (K12)

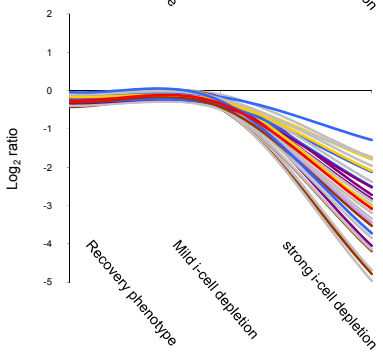
The expression profiles represented in the following clusters are quite similar to the ones of clusters 2, 8, 10 and 15. However, by careful analysis of the expression patterns, it is possible to detect their differences. First, down-regulation in gene expression in the K12 condition was milder, exhibiting a mean of 4-fold change. In addition, in the recovery phenotype all genes showed a down-regulation effect of about 1 fold-change, having a clear difference to their expression level in the FM condition (Fig. 27). This was not the case for the clusters of Figure 26, where no change in expression was observed in the FM and FMR conditions.

Only four of the six clusters exhibiting this expression pattern are shown in Figure 27 (for the rest of the clusters see Fig. S1 in Additional data 2, section 6.2). Most of the gene sequences presented no match to the protein, protein-domain or GO databases. From the genes with a known functional annotation, we detected the *CnPL10* homologue which encodes a protein that belongs to the DEAD-box RNA helicase family [95]. The *Hydractinia* homologue of *RAD23*, encoding a protein involved in DNA repair, was identified in cluster 4. Again, several genes carrying similar functional information appeared in the same clusters which corroborates the quality of the microarray data and clustering algorithm. An example is provided by the three genes bearing the same eggshell protein-domain grouped in cluster 5.



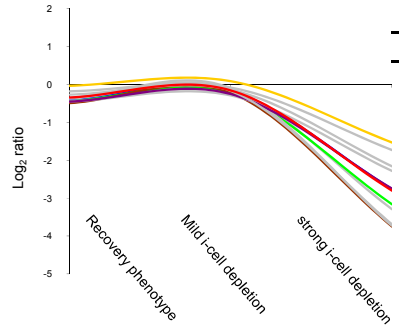
Cluster 2 (11 genes)

Clone ID	Sequence annotation
HEAB-0020F03g1.con	RNA-binding protein cabeza
HEAB-0028L10g1.con	GO:0016787 F hydrolase activity
HEAB-0050E11	Putative uncharacterize methyltransferase
HEAB-0053I14	Predicted regulator of arylsulfatase activity
HEAB-0054J22	Mini-collagen
tai05d09_J.con	E3 ubiquitin-protein ligase DZIP3
tai07h09_O.con	Minicollagen 1
tam59e09y2.con	Collagen-like protein 7



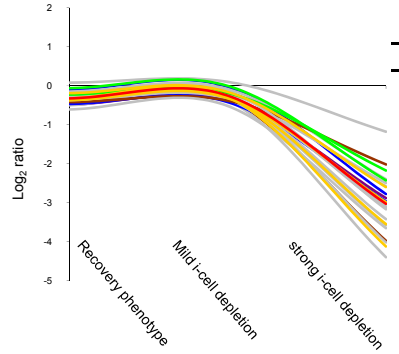
Cluster 8 (29 genes)

Clone ID	Sequence annotation
HEAB-0049L07	Minicollagen 1
HEAB-0050P20	GO:0016787 F hydrolase activity
HEAB-0052F07	Minicollagen 1
tai39f05y1.con	Uncharacterized histidine rich protein
HEAB-0044G19	Mini-collagen
HEAB-0048G21	Mini-collagen
HEAB-0048H01	Minicollagen 1
HEAB-0049B20	IPR010515 Collagenase NC10-endostatin
HEAB-0053F05	ABC transporter
HEAB-0053F14	L-lactate permease
HEAB-0054E08	IPR009305 DUF962
tai04h01_O.con	GO:0003676 F nucleic acid binding
tai17g04y1.con	Mini-collagen
tai33d09y1.con	GO:0016787 F hydrolase activity
tai42a10y1.con	RNA-binding region RNP-1
tai43a12y1.con	HSC70-interacting protein (hip)
HEAB-0020D15g1.con	IPR010515 Collagenase NC10-endostatin
HEAB-0053L23	UDP-N-acetylmuramate



Cluster 10 (11 genes)

Clone ID	Sequencing annotation
HEAB-0052B04	Cytochrome c-type biogenesis protein
tah96f12x1_PT_O.con	Predicted protein
tai20h08y1.con	Mini-collagen precursor
HEAB-0020M09g1.con	Minicollagen 3-4a
HEAB-0049C18	Minicollagen 1
tai51b04y1.con	Minicollagen 3-4a
HEAB-0050O12	IPR002952 Eggshell protein



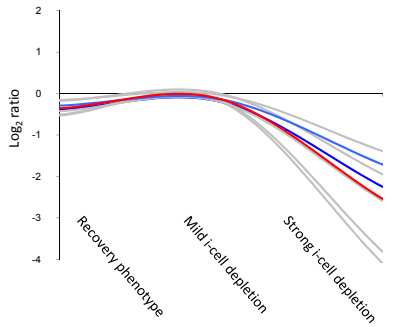
Cluster 15 (23 genes)

Clone ID	Sequence annotation
HEAB-0050L06	Mini-collagen
HEAB-0055N07	IPR010515 Collagenase NC10-endostatin
tai15e03y1.con	Heterogeneous nuclear ribonucleoprot. H
HEAB-0029J13g1.con	Oxidative stress
HEAB-0030L03g1.con	Minicollagen 3-4a
HEAB-0046M07	Minicollagen 1
HEAB-0049O19	Minicollagen 3-4a
HEAB-0053N22	Transcriptional regulator LysR family
tah96a04_J.con	trna-splicing endonuclease 2-like
tai39f07y1.con	GO:0016787 F hydrolase activity
tai44b07y1_PA.con	GO:0016787 F hydrolase activity
HEAB-0035D10g1.con	GO:0016787 F hydrolase activity

GO Terms
- Centroid curve
- Binding activity
- Catalytic activity
- Nutrient reservoir activity
- Structural molecule activity
- Growth / Transcription factor activity
- Transport activity
- Unknown

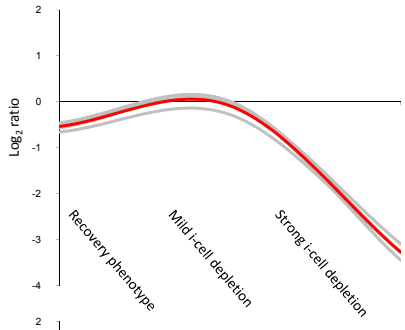
Figure 26 - Genes highly down-regulated in K12 condition. The expression level of the 74 genes in each condition was referenced to the control using Log_2 ratio. List of genes with the corresponding annotation are provided in the table. For an easy overview, a GO colour code

annotation for each gene is provided in the transcriptional profiling curve. Unknown genes were plotted in gray colour, but are not listed in the table.



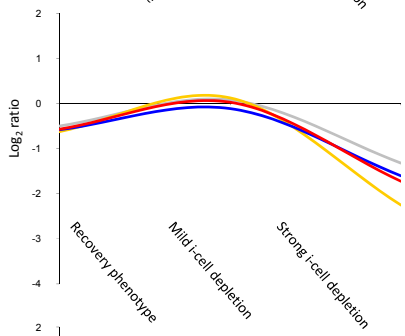
Cluster 4 (8 genes)

Clone ID	Sequence annotation
HEAB-0053F18	RAD23 homolog
HEAB-0054H17	Minicollagen 1
tah99a08_J.con	Uncharacterized protein



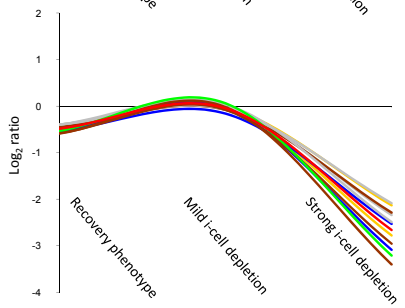
Cluster 5 (3 genes)

Clone ID	Sequence annotation
tam57a05y2.con	IPR002952 Eggshell protein
HEAB-0026C19g1.con	IPR002952 Eggshell protein
tah96a01y1.con	IPR002952 Eggshell protein



Cluster 6 (4 genes)

Clone ID	Sequence annotation
tah96c10y1.con	Predicted protein
HEAB-0053F19	DNA polymerase iv
tai30c12y1.con	Molecular chaperone MRJ



Cluster 11 (11 genes)

Clone ID	Sequence annotation
HEAB-0047O01	Amidase
HEAB-0028J22g1.con	PL10-related protein CnPL10
HEAB-0054A01	Integrase
HEAB-0054J19	IPR010515 Collagenase NC10-endostatin
tai49b10y1.con	IPR006706 Extensin-like region
tai02e09_J.con	3-oxoacyl-(Acryl-carrier-protein) reductase

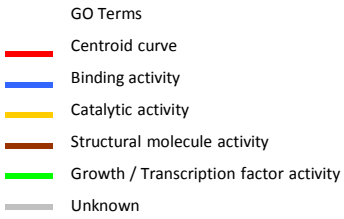


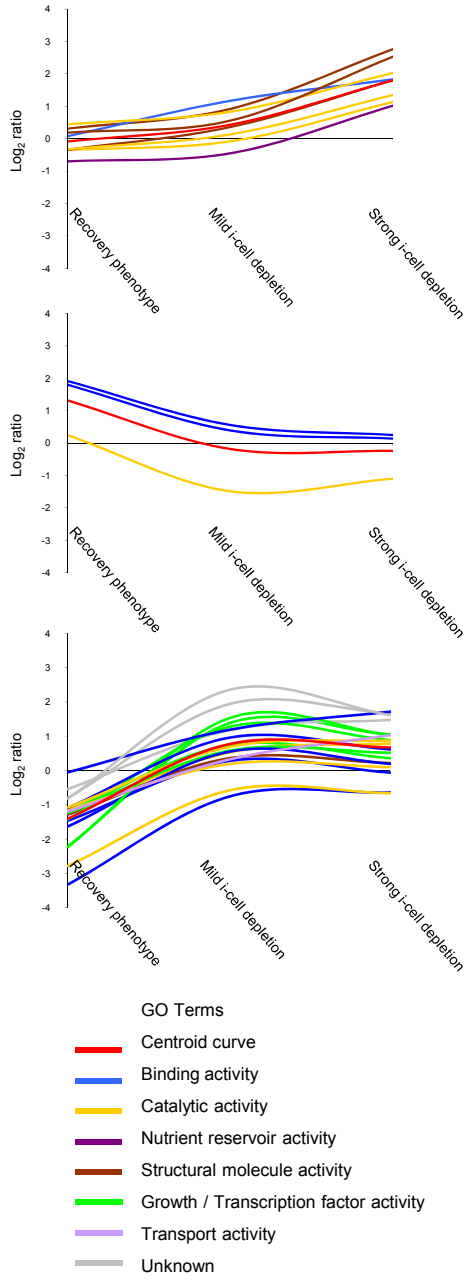
Figure 27 - Genes down-regulated in FMR and K12 conditions. From the six clusters (34 genes in total) following this expression pattern, only four are shown. Expression level of each condition was referenced to the control using Log_2 ratio. List of genes with the corresponding annotation are provided in the table. For an easy overview, a GO colour code annotation for each gene is provided in the transcriptional profiling curve. Unknown genes were plotted in gray colour, but are not listed in the table.

3.3.1.12. Clusters with other gene transcriptional profiles

Only cluster 3 grouped genes, which were up-regulated by a factor of 2 in the K12 condition (Fig. 28). The expression of most of these genes seemed unaffected in the FMR phenotype while in FM, they showed a slight up-regulation. Only the mini-collagen gene (purple curve) was slightly down-regulated, of about 1- fold, in the recovery condition. The genes of this cluster were mainly related to structural or enzymatic activities. Interesting is the up-regulation of a gene encoding for Cathepsin-L, an enzyme with a major role in protein catabolism using substrates as collagen and elastin [96].

An up-regulation of the gene expression in the FMR condition was only observed in cluster 7 (Fig. 28). The gene encoding for Heat shock protein 90 was up-regulated by a factor of 4. This gene was represented by two clones with an almost equal expression level. One of these genes was previously sequenced in the EST project and the second gene comes from the un-sequenced part of the library.

Cluster 9 grouped 19 genes mostly down-regulated in the recovery phenotype and activated in the FM condition. A few of them were also activated in the K12 condition (Fig. 28). In this cluster it is possible to appreciate how Pearson correlation works; where all genes following the same transcription pattern - *i.e.* same curve slope- are clustered in spite of their different expression levels. A gene with a chitin-binding domain provided the lowest expression level in the FMR condition, with a 7-fold down-regulation. In addition, genes encoding for transcription factors, like Bzip/MafL and trefoil factors, were detected. The gene encoding for Astacin-3 presented a 2-fold down-regulation in the FM condition, and an increased expression of about 1-fold in FMR and K12 phenotypes. The profile of the growth factor Bone morphogenetic protein 2 (BMP-2) was similar, but without the expression change in the K12 condition.



Cluster 3 (8 genes)

Clone ID	Sequence annotation
HEAB-0051J21	Amidohydrolase 2
HEAB-0054F03	Mini-collagen
tai04b12x1_PT_O.con	Collagen IV alpha 1 chain
tai08b10_J.con	Clathrin, heavy polypeptide (Hc)
tai18b05y1.con	Rhamnose-binding lectin
tai27f06y1.con	Fibrillar collagen precursor. 2
HEAB-0034B23g1.con	Cathepsin L
HEAB-0040L01g1.con	IPR008160 Collagen/ IPR008161 Clg_helix

Cluster 7 (3 genes)

Clone ID	Sequence analysis
tam55d09y2.con	GO:0016787 F hydrolase activity
HEAB-0049H11	Heat shock protein 90
tah96b01_J.con	Cytosolic-type hsp90

Cluster 9 (19 genes)

Clone ID	Sequence annotation
tai02e07_J.con	Bzip transcription factor MafL
tai08a11_J.con	Trefoil factor 2 precursor
tai33b04y1.con	IPR002557 Chitin binding protein, peritrophin-A
tai38b11y1.con	GO:0016787 F hydrolase activity
HEAB-0018H01g1.con	Predicted protein
HEAB-0024D19g1.con	Trefoil factor
HEAB-0033H17g1.con	Zinc finger protein 578
HEAB-0036F06g1.con	Bone morphogenetic protein 2 precursor (BMP-2)
HEAB-0037B17g1.con	IPR000719; Prot_kinase_core
HEAB-0043A13	Putative uncharacterized protein
HEAB-0049G05	Trefoil factor
HEAB-0052O10	Trefoil factor
HEAB-0053P02	IPR007116 6-pyruvoyl tetrahydropterin synthase
HEAB-0055O16	Heat Shock protein DnaJ
tai05h01x1_PT_O.con	Astacin 3
tam54a08y2.con	Rhamnose binding lectin
HEAB-0023E14g1.con	Zinc finger protein 578
tai46g02y1.con	Lipocalin family protein

Figure 28 - Genes with different expression profiles. Approximately, 20% of the genes were distributed in these three clusters. Expression level of each condition was referenced to the control using Log_2 ratio. List of genes with the corresponding annotation are provided in the table. For an easy overview, a GO colour code annotation for each gene is provided in the transcriptional profiling curve. Unknown genes were plotted in gray colour, but are not listed in the table.

3.3.2. Searching for allorecognition and immune related genes in *Hydractinia* –the immune microarray experiment

The microarray was used to identify the *Hydractinia* immune gene repertoire involved in infection and allogeneic reactions. For this purpose, the following experiments were performed; four adult colonies with the same genotype (termed K4) were incubated with Lipopolysaccharide (LPS) as described in methods (section 2.2.1.3). Subsequently, RNA was isolated from two colonies, each at 1 and 3 hours after treatment. To asses for allorecognition related genes, four adult K4-clones were allowed to grow into contact with a genetically distinct colony. Then, RNA was isolated only from the contact area that exhibited signs of rejection. As a reference control, RNA was isolated from untreated K4 colonies. The quality of the RNA was analyzed with the Agilent Bioanalyzer and by spectrophotometric readings as described in section 2.2.2.3. For each sample condition, only high quality isolated RNAs were pooled (Fig. 29). In the microarray experiment, allorecognition was represented by 4 biological replicates, while each time point after the LPS-infection was represented by 2 biological replicates.

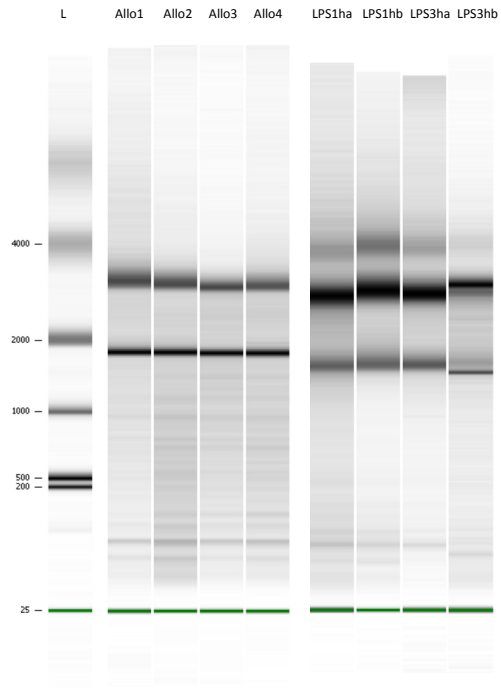


Figure 29 – Analysis of the RNA quality. All RNA samples presented a RIN ≥ 6.5 and an abs. ratio (260nm/280 nm) above 1.9. For each condition, all samples were pooled. In the case of the wild-type colonies, 4 high quality RNA samples were isolated and pooled (data not shown). The two main bands in the figure correspond to the 28s and 18s rRNA.

3.3.2.1. Generation of the microarray data

The optimized protocols for the labelling reaction defined in the mitomycin microarray experiment were used, producing between 1,5 and 2,5 µg of labelled cDNA. Then a double reference design was followed (Fig. 30). Six technical replications (including dye-swap) were performed for each comparison between a condition and the reference. Therefore, together with the duplicate spots within the microarray, 12 data points per gene were generated in each hybridization. To diminish any additional technical biases, all replicates were hybridized in parallel and with the same amount of labelled material. The slides were hybridized, scanned and analyzed with Genepix software as described in methods. Genepix graphic display showed a good hybridization quality in the scanned arrays with respect to the background intensity level, reproducibility and physical characteristics of the spots. Thus, signal data coming from 16 hybridized slides were exported as a .GPR file for further analysis.

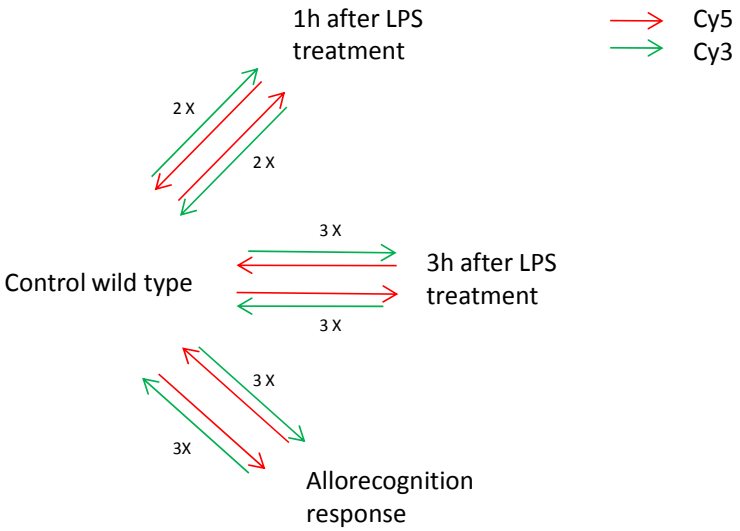


Figure 30 - Double reference design of the immune microarray experiment. For within array quality control, duplicate spots were included in the array. Six competitive hybridizations between condition (applying dye-swap) and the reference were performed for each experiment. This resulted in the acquisition of 12 data points per gene, allowing the use of a robust statistic analysis. Only in the condition 1h after LPS treatment, 4 hybridizations were considered, *i.e.* 8 data points per gene.

3.3.2.2. Normalization and filtering of the microarray data

First, the green and red intensity channel data was extracted from the .GPR table file. Then, the conditions and reference measurements were defined for all hybridizations, and all experimental conditions (allorecognition, LPS; 1 and 3 hours after induction) were analyzed together using the same wild-type control as reference. Again, normalization was done with the M-CHiPS software package using the original intensities of each measurement and the

logarithmic regression as fitting method algorithm (section 2.2.9.1). Plotting the fitted intensities of the condition and control measurements resulted in an adjusted regression curve matching the main body of the data-point cloud, with a correlation coefficient above 0.85 in all 18 hybridizations (data not shown).

Genes with a strong evidence of being differentially expressed were identified from the normalized data. This was done following a similar four-step filtering criterion used in the mitomycin microarray experiment (section 3.3.1.5). First, 56 genes showing saturated intensities were subtracted from the data set. The second filtering step resulted in the selection of 1,565 genes having a substantial absolute expression level (> 2,000 AU) in at least one of the conditions. Using SAM statistics, we extracted 284 genes with a significant difference in their transcription level (corrected p-value < 0.05). Finally, the fourth filtering criterion resulted in the selection of 245 transcripts (2.6 % from the whole gene-set) with a minimum of a 2-fold expression change.

Fitted intensity plots of all conditions compared to the reference control clearly demonstrated that in spite of the high correlation between the two channel data, the normalization and filtering steps were successful in the selection of a substantial number of differentially regulated genes (Fig. 31).

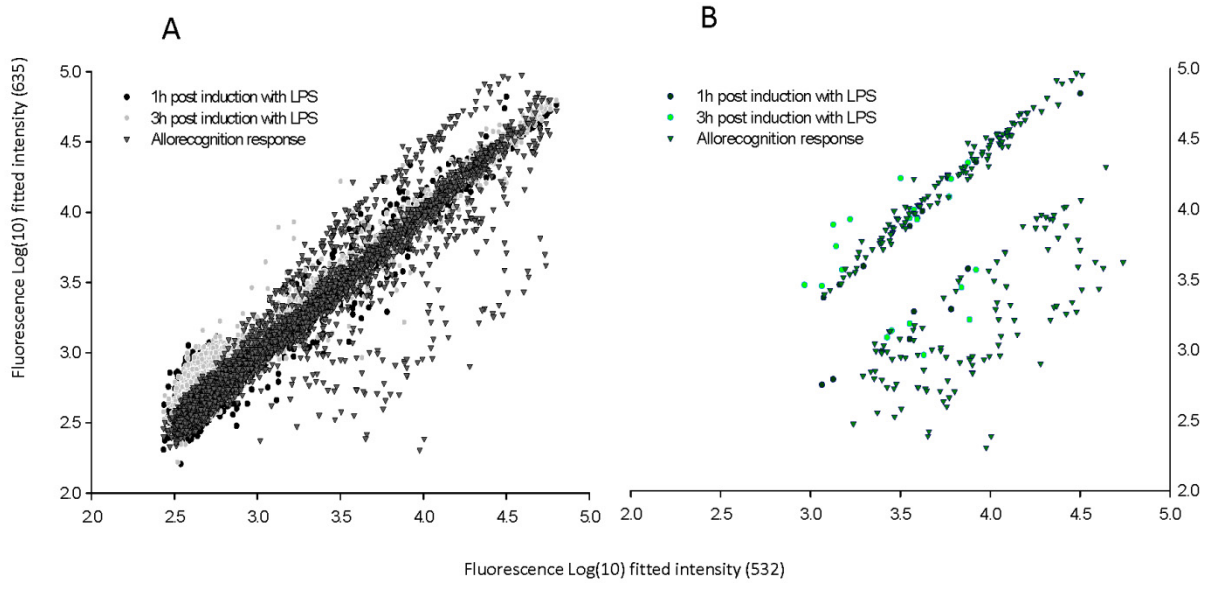


Figure 31 - Pair-wise comparison of the fitted intensities of the LPS (1h and 3h post induction) and allorecognition challenges refereed to the control A. Scatter plot of Log R (red) vs. Log G (green) including all genes and **B.** only the filtered ones. Filtering allowed the selection of the best 245 differentially expressed genes, which clearly stand out from the data points cloud.

3.3.2.3. Correspondence analysis

To reveal the relationships among and between genes and conditions, the differentially expressed genes identified in LPS and allorecognition experiments were analyzed using correspondence analysis (for CA descriptions see section 2.2.9.2). For this, the distance among these variables were displayed in a low dimensional biplot (Fig. 32). Replicate hybridizations of each experimental condition (represented by colour boxes) were successfully clustered. The fitted intensity values of the primary and secondary spots, which are within array controls, presented a minimal variation. This sustains a high experimental reproducibility. The plot showed a clear separation of all different conditions, supporting a good normalization and filtering of the data. Allorecognition response measurements (pink) branched out from the central region of the biplot, where both LPS hybridizations grouped together with the control (red). A dense cloud of genes, with up- and down-regulation expression profiles, were associated to the allorecognition response. These results demonstrate that allogeneic reactions generated the highest change in expression in the analyzed data. In contrast, few genes were specific to the LPS treatment. No gene dots were distributed in the centroid of the plot, supporting the elimination of non-regulated genes by the filtering criteria.

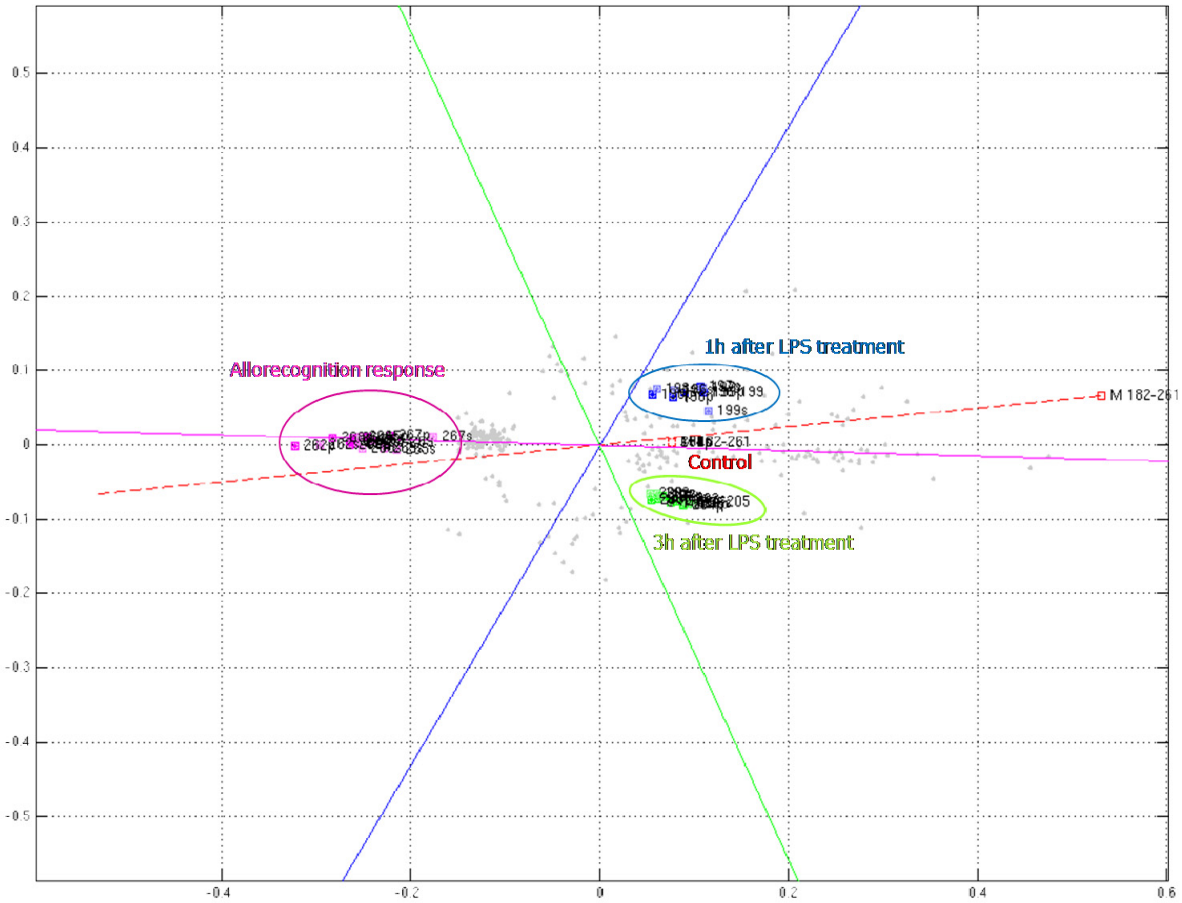


Figure 32 - Correspondence analysis of LPS (1 and 3 hours after treatment) and allorecognition conditions. The biplot clearly shows four directions corresponding to the two LPS treated conditions, the allorecognition phenotype and the control reference. The 245 selected genes are represented as gray dots. Dots distributed in the centroid of the biplot correspond to non-regulated genes. The duplicate spots of the array, primary (p) and secondary (s), are displayed with the hybridization numbers and in light and strong colours, respectively. Lines follow the direction of the standard coordinates of the condition medians with the respective colours.

3.3.2.4. Hierarchical clustering

As in the mitomycin microarray experiment, the M-CHiPS results were exported to the TIGR Multiexperiment Viewer (MeV). Again, the logarithmic ratios between the median signal intensities of the conditions and control measurements were used to determine differential gene expression patterns. Log_2-transformed expression ratios varied between -5.5 and 2.5, with a median of zero for both LPS induced conditions and a slightly higher value (0.21) for the allorecognition phenotype. An easy visualization of the logarithmic data was obtained displaying all the information in a colour-coded heat map and applying HCL algorithms in order to cluster the genes and conditions according to their expression pattern. The results observed in the CA biplot were confirmed, showing that most of the selected genes were

differentially regulated in the allogeneic provoked organisms. Considering a threshold of 2-fold expression change, 108 and 116 genes were down- and up- regulated in this condition, respectively. The two other conditions, 1 and 3 hours after LPS treatment, presented a similar expression pattern as the control; having only seven and 13 genes with a 2-fold down- and up-regulated profile, respectively. However, the transcription pattern of the few regulated genes at 3 hours after the LPS treatment was similar to the one observed in an allogeneic reaction. This can be seen in the genes of sub-clusters I and II, with a down-regulated (green) profile for this two conditions while having an up-regulation (red) in the 1h after LPS induction phenotype, or *vice versa* (Fig. 33). Therefore, HCL -using Pearson correlation as distance metrics- clustered allorecognition with the 3 hours after LPS treatment condition, rather than the two LPS induced organisms together. In contrast, CA grouped both LPS conditions quite close to the control and centroid of the biplot (Fig. 32). Thus, while CA clustering was influenced by the high number of genes regulated in allorecognition, giving the maximum weight in the plot, the profile of the few LPS regulated genes substantially influenced the HCL clustering. With a threshold node distance of -0.99, HCL created 21 main nodes or clusters (Fig. 33). Interesting to mention is that most of the generated clusters carried several sequences except in the case of the sub-clusters I and II, were the clusters contained between 2 and 5 sequences each.

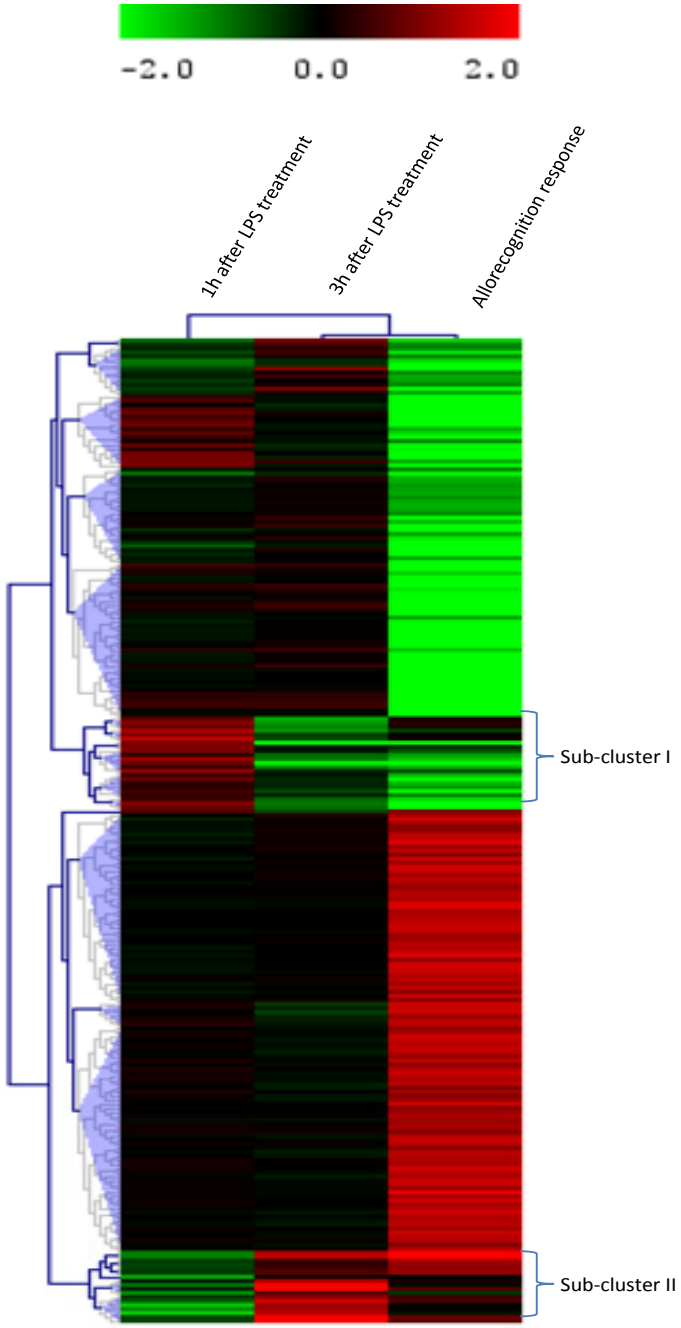


Figure 33 - HCL analysis of the differentially expressed genes in the immune microarray experiment. All 245 selected genes are ordered individually in rows, while the samples are ordered in columns. HCL resulted in 21 main clusters or nodes. They are represented by dark blue branches and a translucent wedge from that node to all enclosed elements. Sub-nodes, below the cluster distance threshold are shown as light gray branches. The length of the branches is proportional to the distance between the nodes. The scale followed up the median of the data, between -2 and 2.

3.3.2.5. Figure of Merit algorithm and k-means clustering

Before applying k-means clustering, the optimal clustering parameters were defined using the Figure of Merit algorithm (see section 2.2.9.3). FOM values were calculated for 40 clusters using k-means and both variables were subsequently plotted (Fig. 34). The plot shows that the

FOM values drastically decreased in the first k-means runs. Then, the curve slope smoothly decreased with each new added cluster to the algorithm. As in the case of the FOM analysis in the mitomycin microarray data, it was difficult to determine whether 20, 25 or more clusters improve the representation of the data. For a better support in defining the number of clusters to be used in KMC, the performance of k-means was empirically tested with 15, 20, 23, 25 and 30 clusters (data not shown). The use of more than 23 clusters resulted in the generation of several singletons and empty clusters. With less than 20 clusters, k-means resulted in highly populated clusters, limiting the analysis of the expression profiling data. Finally, supporting the analysis done with the HCL clustering, k-means was best performed with 21 clusters.

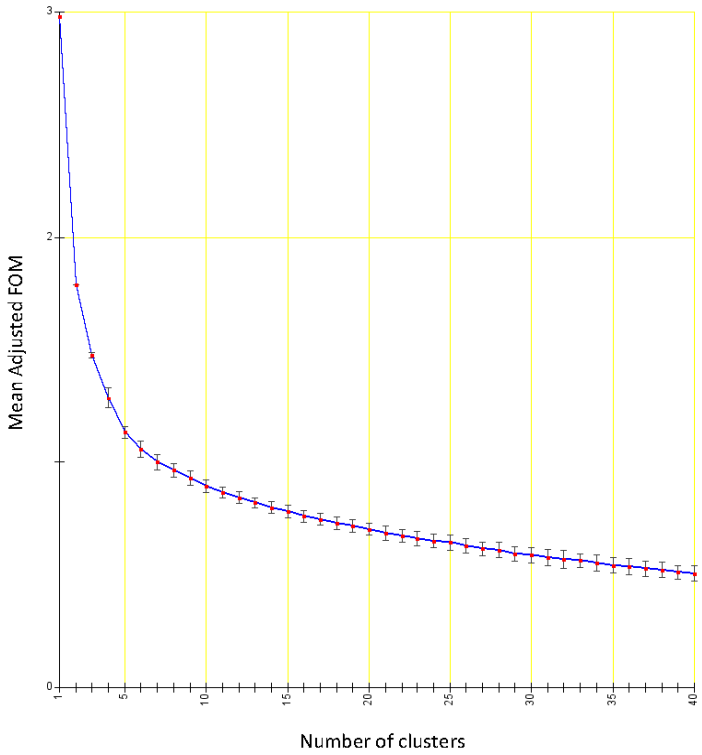


Figure 34 - FOM analysis for k-means algorithm. FOM helps to define the predictive power of the k-means algorithm in generating clusters. The lower is the FOM value, the higher is the predictive power of k-means. The mean FOM values are showed as red dots with the corresponding standard deviation.

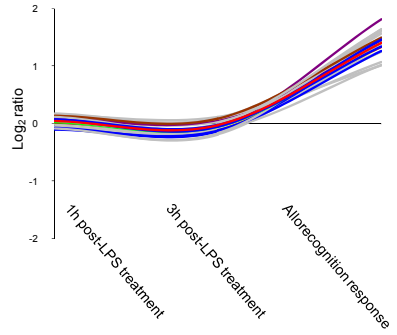
Pearson correlation was used as distance metrics for the k-means clustering algorithm. For each cluster, the Log_2-transformed expression ratios (conditions/control) of all the analyzed conditions –the two LPS and allorecognition provoked phenotypes- were plotted. As in the analysis of the mitomycin microarray data, k-means produced clusters containing genes with similar transcription profiles but different expression levels. The good performance of the clustering method was confirmed by the detection of redundant genes within a cluster or in

different clusters having a similar expression pattern. Thus, gene replicates were equally or similarly affected in the analyzed conditions. This was the case for mini-collagens, present in all clusters exhibiting an up-regulation in the gene expression during an allogeneic reaction (Fig. 35). Also lectins were grouped together but mainly in clusters showing a down-regulation pattern in the allorecognition condition (Fig. 36).

As in the mitomycin KMC analysis, GO terms were added to the graphic display of the clusters and, clusters showing a similar and specific expression pattern for a certain condition were grouped for individual analysis. A selection of the most interesting clusters is described below. The rest of the clusters are available in Additional data, section 6.2.

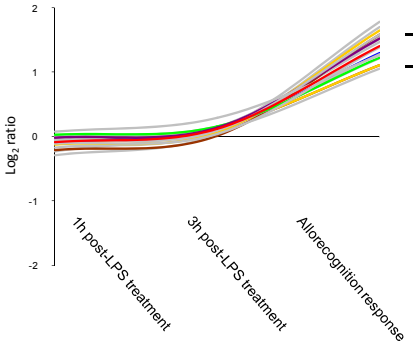
3.3.2.6. Genes specifically up-regulated in an allogeneic reaction

K-means grouped 109 genes (44%) into 5 clusters showing in average a 2.8-fold up-regulation in the allorecognition condition (Fig. 35). These genes presented no change in their expression profile when the animals were treated with LPS. In HCL, these allorecognition specific genes were grouped in three main clusters. This means, that k-means was more effective in finding similarity patterns between the expression profiles, allowing a better characterization of the genes. In spite of the broad variety of GO terms through all the clusters, the majority of genes were related to a binding, structural or catalytic activity. This included several genes encoding for mini-collagen or collagen–like peptides as well as the already identified *RAD23* and *CnPL10* genes (Fig. 35 and Fig S2 in Additional data 2). A total of 38 genes did not present any similarity match in the databases used for annotations (gray curves). Below are shown the clusters 5, 6 and 15. The two other clusters exhibiting this expression pattern are available in Additional data 2 (Fig. S2, section 6.2).



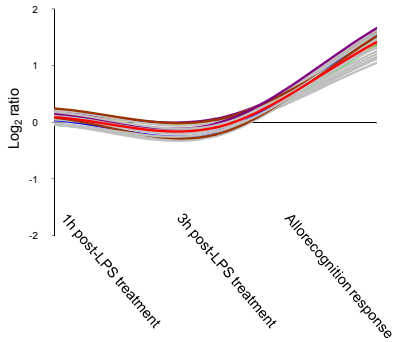
Cluster 5 (26 genes)

Clone ID	Sequence annotation
HEAB-0047O01	Amidase
HEAB-0020D15g1.con	IPR010515 Collagenase NC10- endostatin
HEAB-0048G21	Mini-collagen
HEAB-0049L07	Minicollagen 1
HEAB-0049O19	Minicollagen 3-4a
HEAB-0050D08	Minicollagen 3-4a
HEAB-0050L06	Mini-collagen
HEAB-0053F18	RAD23 homolog
tah99a08_J.con	Uncharacterized protein
tai20h08y1.con	Mini-collagen precursor
tam59e09y2.con	Collagen-like protein 7
HEAB-0052B04	Cytochrome c-type biogenesis protein
tai43a12y1.con	HSC70-interacting protein (hip)



Cluster 6 (18 genes)

Clone ID	Sequence annotation
HEAB-0029A21g1.con	IPR002952 Eggshell protein
HEAB-0035E14g1.con	General secretion pathway protein C
HEAB-0050G02	IPR002952 Eggshell protein
HEAB-0054J19	IPR010515 Collagenase NC10, endostatin
tah96f12x1_PT_O.con	Predicted protein
tai15e03y1.con	hnRNP H
tai33d09y1.con	GO:0016787 F hydrolase activity
HEAB-0053N22	Transcriptional regulator LysR family
HEAB-0054J22	Mini-collagen
tam54c10y2.con	IPR002952 Eggshell protein



Cluster 15 (26 genes)

Clone ID	Sequence annotation
HEAB-0053F05	ABC transporter
HEAB-0055N07	IPR010515 Collagenase NC10-endostatin
tah96a04_J.con	trna-splicing endonuclease subunit sen3
tah96c09y1.con	IPR000480 Glutelin
tai17g04y1.con	Mini-collagen
tai39f05y1.con	Uncharacterized histidinine rich protein
tai44b07y1_PA.con	GO:0016787 F hydrolase activity
tai51b04y1.con	Minicollagen 3-4a
HEAB-0029J13g1.con	Oxidative stress
HEAB-0030L03g1.con	Minicollagen 3-4a
HEAB-0046M07	Minicollagen 1
HEAB-0049C18	Minicollagen 1
HEAB-0050O12	IPR002952 Eggshell protein
HEAB-0053I14	Predicted regulator of arylsulfatase activity

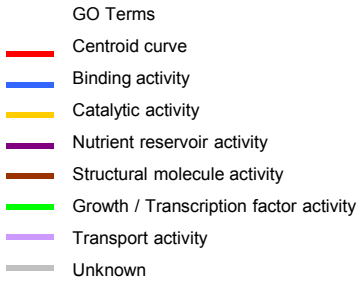


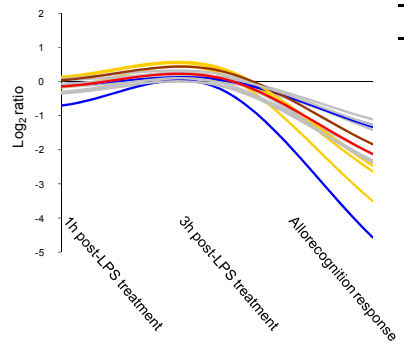
Figure 35 - Genes specifically up-regulated in an allogeneic reaction. The expression level of the 109 genes in each condition was referenced to the control using Log_2 ratio (Condition/Control). Logarithms are in units of 2-fold changes, *e.g.* Log_2 ratio = 0 corresponds to an equal gene expression between the condition and the control. In the table are listed the gene identification ID and their annotation. Gene annotation was performed through BLAST, GO and Domain analysis. For an easy overview, a GO colour-code annotation for each gene is provided in the transcriptional profiling curve. Unknown genes were plotted in gray colour but are not listed in the table.

3.3.2.7. Genes specifically down-regulated in an allogeneic reaction

A total of 62 genes, distributed in four clusters, presented in average a 5-fold down-regulation in the allorecognition condition. As in the clusters analysed in section 3.3.2.6, these genes were un-affected during an LPS treatment and most of them were associated to a binding or catalytic activity (Fig. 36). In accordance with this, Rhamnose-binding lectin (*RBL*) genes and Rhamnospondin-2 (*Rsp*-2) -a gene that contains one RBL and eight thrombospondin type 1 domains (TSR)- were detected mainly in cluster 21. Interestingly, four genes encoding for enzymes of the Phospholipase A2 (PLA2) family (two of them correspond to Conodipine-M alpha chain, a novel class of PLA2), were reported. In addition, three different genes encoding for protein kinases were detected including; the Pantothenate kinase (*PANK*) and the *C. elegans* Mitogen-activated protein kinase (MAPK) homologue *mpk-1* (Fig. 36 and Fig. S3 in Additional data 2, section 6.2). Only four growth and transcription factors were found through all the clusters, including the fibroblast growth factor 2 and trefoil factors. Again, a significant number of the clustered genes (22) presented no functional sequence annotation. They were displayed as gray curves in the plots (Fig. 36).

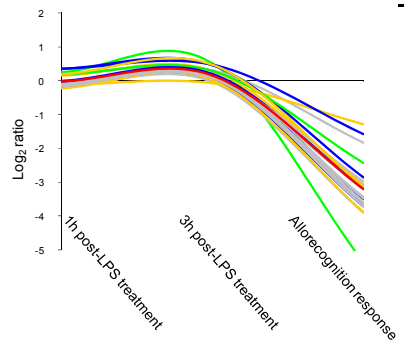
3.3.2.8. Genes up-regulated immediately after LPS treatment

Six clusters grouped a total of 42 genes exhibiting an enhanced transcription directly after the LPS treatment. The majority of these genes were distributed in two clusters, presenting in average a 1.2 fold up-regulation (Fig. 37A). At 3 hours post LPS induction, these genes were equally expressed as in the untreated organism. However, allogeneic encounters resulted in a drastically 5-fold down-regulation in the gene expression. GO annotation revealed that most of these genes were associated with a binding activity. In cluster 19, the putative genes encoding for the growth factor BMP-2 and the cell cycle regulator G2/M cyclin were detected. Cluster 2 and 12 included seven genes with a slightly different expression pattern (Fig. 37B). They presented an expression level with a 2-fold up-regulation directly after the LPS incubation but, in contrast to the clusters of Figure 37A, a 1-fold down-regulation in the following 3 hours and no alteration after an allorecognition challenge. Two genes encoding for histone H1 and H1/H5 domains were reported in these clusters. From the genes up-regulated at 1 hour after LPS treatment, 20 were unknown or annotated with non-informative terms. Cluster 10, 11 and 20 are available in Fig. S4 in Additional data 2 (section 6.2).



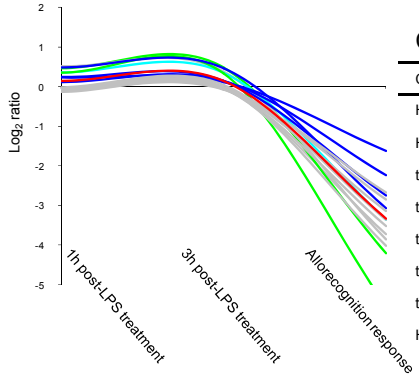
Cluster 7 (15 genes)

Clone ID	Sequence annotation
HEAB-0053N12	Putative uncharacterize protein
tai06g05_J.con	FGF-2 binding protein
HEAB-0049C03	IPR016090 Phospholipase_A2
HEAB-0052O21	IgA-binding beta antigen
tam53h06y2.con	Surface protein PspC
HEAB-0028C03g1.con	Multiple epidermal growth factor-like domains-6
HEAB-0041P11g1.con	Conodipine-M alpha chain
HEAB-0043F19	Predicted protein
HEAB-0052G04	Actin, non-muscle 6.2
HEAB-0054C18	IPR005028 Herpes virus intermediate



Cluster 18 (23 genes)

Clone ID	Sequence annotation
HEAB-0049G05	Trefoil factor
HEAB-0053H21	Type III pantothenate kinase
HEAB-0049I14	Conodipine-M alpha chain
HEAB-0055C06	Fluorescent protein raspberry
HEAB-0044H22	Putative IPR016090 Phospholipase_A2
tai12h10_J.con	Rhamnose binding lectin STL4
tai29e11y1.con	GO:0016787 F hydrolase activity
HEAB-0024D19g1.con	Trefoil factor
HEAB-0048H16	Transmembrane sensor
HEAB-0055O16	IPR003095 Heat schock protein DnaJ
tah98b07_J.con	IPR012677; a_b_plait_nuc_bd / IPR000504; RRM_RNP1
tai12d06_O.con	Rhamnose binding lectin
HEAB-0046H17	Aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit B
tai38b11y1.con	GO:0016787 F hydrolase activity



Cluster 21 (15 genes)

Clone ID	Sequence annotation
HEAB-0052O10	Trefoil factor
HEAB-0034H12g1.con	L-rhamnose-binding lectin
tah97a06y1.con	Rhamnose-binding lectin
tai08a11_J.con	Trefoil factor 2 precursor
tai15f07y1.con	Protein kinase
tai26c04y1.con	Rhamnospondin 2 (Fragment)
tam54d02y2.con	Rhamnospondin 2
HEAB-0021K03g1.con	IPR002350 Prot_inh_Kazal

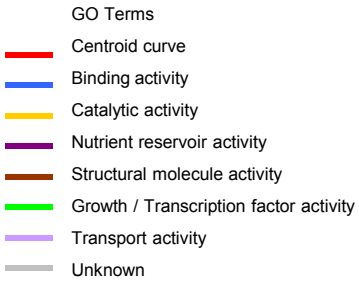
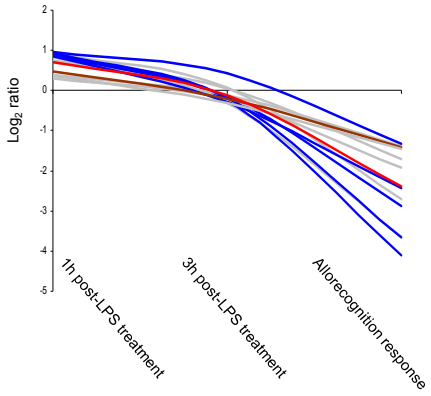


Figure 36 - Genes specifically down-regulated in an allorecognition challenge. The plots show the expression level of the 62 genes in each condition referred to the control using Log_2 ratio (Condition/Control). Log_2 ratio = 0 corresponds to an equal gene expression between the condition and the control. The gene identification ID and annotations are listed in the table. For an easy overview, a GO colour-code annotation for each gene is provided in the transcriptional profiling curve. Unknown genes were plotted in gray colour but are not listed in the table.

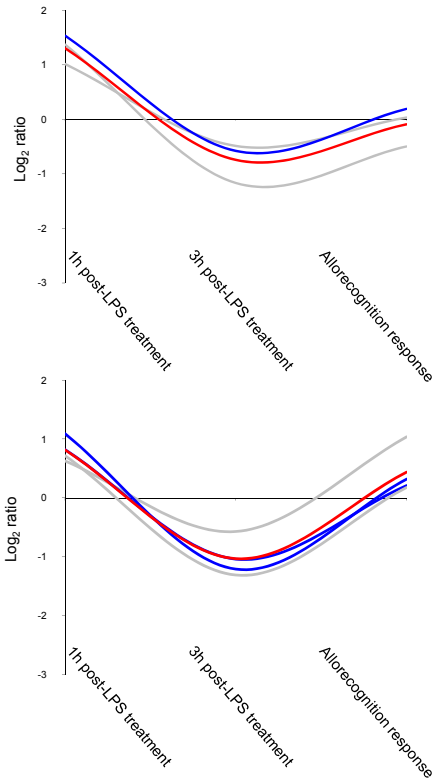
A



Cluster 19 (12 genes)

Clone ID	Sequence annotation
tai35c05y1.con	g2/mitotic-specific cyclin-b 70
HEAB-0035E15g1.con	RNA-binding protein 12b
HEAB-0036F06g1.con	Putative Bone morphogenetic protein 2 precursor
HEAB-0045D05	IPR002952 Eggshell protein
HEAB-0046D08	RNA-binding protein 12b
HEAB-0048D18	ORF 2
tah98g01y1.con	Eukaryotic translation elongation factor 1 beta 2
tai48h01y1.con	RNA-binding protein 12b
HEAB-0039G19g1.con	Putative uncharacterized protein
HEAB-0037E10g1.con	ORF 2

B



Cluster 2 (3 genes)

Clone ID	Sequence annotation
tam59e12y2.con	Translation elongation factor 1B gamma subunit

Cluster 12 (4 genes)

Clone ID	Sequence annotation
HEAB-0055M09	IPR005818 Histone H1/H5 / IPR009435 Acid shock
tai08a12_O.con	Histone H1.0
HEAB-0048A23	Predicted protein

GO Terms
Centroid curve
Binding activity
Structural molecule activity
Unknown

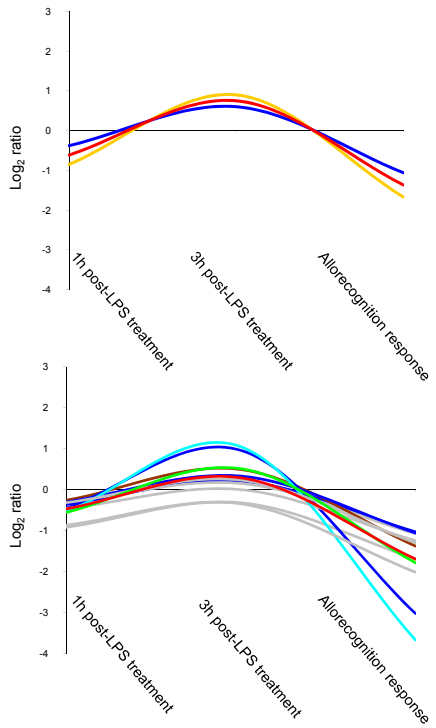
Figure 37 - Genes up-regulated immediately after LPS induction. Few genes were up-regulated after an LPS treatment. In **A.** genes up-regulated immediately after LPS induction were down-regulated at 3 hours post LPS induction and in an allorecognition response. In **B.** genes up-regulated at 1 hour after LPS induction followed a down-regulation at 3 hour post LPS treatment and were not affected in an allorecognition response. Each condition referred to the control using Log_2 ratio (Condition/Control). The gene identification ID and annotations are listed in the table. Unknown genes were plotted in gray colour but are not listed in the table.

3.3.2.9. Genes up-regulated at three hours after LPS treatment

The next four clusters display 28 genes activated after 3 hours of the LPS induction. In spite of having similar profile patterns, the clusters showed significant differences in their gene expression levels. Clusters 14 and 16 presented smoothly profile curves, with genes down- and up- regulated by 1-fold at 1 and 3 hours post-LPS treatment, respectively. After an allorecognition response, these genes were significantly affected, reaching in average a 3-fold down-regulation (Fig. 38A). Some of these genes encode for Heat shock protein 70 (HSP70). In cluster 16, two genes reached a 2-fold up-regulation at 3 h post LPS induction. One of these genes presented a high sequence similarity to Tachylectin. The second gene contains a Thrombospondin type 1 domain, also observed in several lectin proteins. As in the mitomycin microarray analysis, again a putative gene encoding for bone morphogenic protein 4 (BMP-4) was reported.

A slightly different expression profile pattern exhibited the genes of cluster 1 (Fig. 38B) and 13 (not shown). These genes showed a smoothly down-regulation directly after the LPS induction but an enhanced 3-fold up-regulation in the next 3 hours and no regulation during allorecognition. In cluster 13, all sequences are unknown and therefore, are good candidates for the identification of novel pathogen-induced specific genes. In contrast, all sequences of cluster 1 were annotated with a catalytic activity. Some of the identified genes encode for the metalloprotease Astacin and the serine protease inhibitor Antistasin.

A



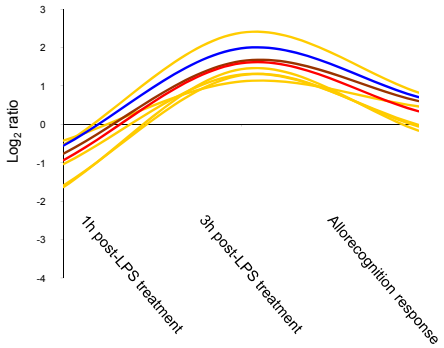
Cluster 14 (2 sequences)

Clone ID	Sequence annotation
HEAB-0020E07g1.con	8(R)-lipoxygenase
HEAB-0050J15	Heat shock protein 70

Cluster 16 (12 sequences)

Clone ID	Sequence annotation
HEAB-0044K02	Heat shock protein 70
HEAB-0050H23	Heat shock protein 70
HEAB-0055I08	IPR001442 Type 4 procollagen, C-terminal repeat
HEAB-0031J15g1.con	Cnidarian tachylectin-related protein
HEAB-0041J22g1.con	IPR007151 Mpp10 protein
HEAB-0048D02	Heat shock protein 70
HEAB-0055P10	IPR000884 Thrombospondin, type 1 repeat
HEAB-0029K06g1.con	Putative bone morphogenic protein 4

B



Cluster 1 (7 sequences)

Clone ID	Sequence annotation
tai04c03x1_PT_O.con	Glutathione S-transferase
tai05h01x1_PT_O.con	Astacin 3
tai07b02x1_PT_O.con	Antistasin
tai08b10_J.con	Clathrin, heavy polypeptide (Hc)
tai46g02y1.con	Lipocalin family protein
tam55d09y2.con	GO:0016787 F hydrolase activity
tam64d08y2.con	RNA-binding region RNP-1

GO Terms
- Centroid curve
- Binding activity
- Catalytic activity
- Nutrient reservoir activity
- Structural molecule activity
- Growth / Transcription factor activity
- Transport activity
- Unknown

Figure 38 - Genes specifically up-regulated at three hours after LPS treatment. In total 28 genes (11% from the selected genes) showed an up-regulation at 3 hours post LPS induction. In **A.** this up-regulation profile was accompanied by a down-regulation in both at 1h after LPS treatment and allorecognition condition. In **B.** the up-regulation profile at 3 hours post LPS induction was accompanied only by a down-regulation at 1 hour post LPS induction. The transcription level of each condition was referenced to the control using Log_2 ratio. List of genes with the corresponding annotation are provided in the table. Unknown genes were plotted in gray colour but are not listed in the table.

4. Discussion

4.1 The *Hydractinia echinata* EST project

4.1.1. The *Hydractinia* EST dataset

The quality of EST collections highly depends on the selection of the RNA sources employed for the generation of the cDNA library. It is well described that in standard libraries it is difficult to discover rarely expressed genes. The yield in gene discovery can be increased by in-depth sequencing or by broadening the diversity of source materials [97, 98]. In the case of *Hydractinia*, its complex life cycle provides a broad spectrum of temporarily and spatially regulated genes. To obtain a more complete representation of the *Hydractinia* transcriptome, a RNA pooling strategy was used for the construction of the cDNA library (see section 2.2.3.1). Using this approach, the information related to gene expression at any particular stage was lost, but all life stages were covered. Thus, the chance to include rare transcripts in the library was increased. Despite having a non-normalized library, EST clustering resulted in 60% of the ESTs being singletons or grouped in clusters of 2-5 sequences (Fig. 6). Only relatively few ESTs were highly redundant. They mainly correspond to housekeeping genes. The 3,808 consensus sequences generated by FAS may be considered as an overestimation of the real number of unique transcripts isolated. EST end-sequencing usually does not retrieve the complete cDNA sequence of a clone, because genes are too long to be covered or because there is a decrease in the quality at the end of the reads. This complicates assembly and clustering, which may result in different unique consensus sequences carrying the same information. On the other hand, it is also possible to have an under-representation of the real number of unique sequences in case of members of closely related gene families [98]. With the availability of genome data, it might be possible to test and improve the EST assembly, but this information has not been generated in *Hydractinia* so far [99]. However, the quality of the assembly was assessed by two different ways. At the nucleotide level, BLASTN comparison of the consensus sequences against all *Hydractinia* ESTs corroborated the physical clustering done by the FAS programs (data not shown). At the protein level, BLASTX comparison to different protein databases revealed a redundancy of 1.6 % in all consensus sequences with a significant hit. These redundant consensus sequences represent

different parts of genes and could therefore not be clustered by FAS for the lack of overlapping sequences. Most of these genes encode ribosomal, actin and lectin proteins, or proteins related to an enzymatic activity.

4.1.2. Functional annotation of the ESTs

Despite the success of having cDNA inserts with an average length of 1.8 kb, suggesting sequences with open reading frames (ORFs), a significant number of them could not be annotated. Consequently, these sequences were considered as unknown or with an un-informative description (Fig. 7A). Analyses of the unknown sequences revealed a lower average sequence length of approximately 300 bp with a median of 160 bp. Thus, it is reasonable to assume that the majority of these sequences do not represent an ORF, but correspond mainly to the 3` rather than 5` non-coding region of a gene [11]. In contrast, sequences with a positive match in the protein databases presented an average and median of 639 bp and 629 bp, respectively. In addition, a better characterization of these sequences was possible since more than 60% of the reads corresponded to ORFs. The inclusion of a protein domain annotation step allowed characterizing 55% of the *Hydractinia* consensus sequences.
The program GOPET, which can perform an organism-independent GO annotation [84, 100], successfully assigned GO terms to the *Hydractinia* sequences. This classification was supported with a generic GO slim, which by limiting the level of detail of the GO-specific fine terms, revealed a broad range of functions and processes in the *Hydractinia* dataset (Fig. 7B). GO classification correlated to the BLAST gene product predictions, assessing the accuracy and quality of the sequence annotation. Nevertheless, some sequences encoding members of gene families with known function were not annotated with GO terms. This was probably because most of these sequences were of a short-length. In such cases, the function was assigned based on fairly weak matches (close to the BLASTX cut-off E-value of 10^{-6}). Conversely, sequences considered as unknown by BLASTX analyses were annotated with GO terms. GOPET uses BLAST approaches that, besides searching in the SwissProt database, include searches in 16 GO-mapped protein databases of different model organisms. Hence, using an organism-independent prediction performance and with a prediction quality assessed by assigned confidence values, GOPET provides a rich reference platform for annotation [84]. Improvements in the functional annotation of *Hydractinia* genes may be achieved with an increasing amount of EST reads. This may allow larger consensus sequences representing nearly complete coding sequences to be generated, providing more accurate annotations

[101]. In addition, the ongoing cnidarian sequencing projects as well as the improvements of the GO annotation of other organisms will provide better platforms for sequence comparisons [7, 8].

Another possible explanation for the unknown sequences is that they could be cnidarian, or even smaller taxon specific genes (*i.e.* absent even from *Hydra* and *Nematostella*). These taxon-specific genes may either be the result of the conservation of ancient genes, lost in all other animals, or evolutionary novelties. For example, cnidarians possess many unique features such as their stinging cells, known as nematocytes or cnidocytes, which are not found in any other group of animals.

4.1.3. *Hydractinia* sequences with non-metazoan hits

A significant amount of the *Hydractinia* consensus sequences presented a non-metazoan hit in the protein databases (Fig 7A). The majority corresponded to bacterial sequences with a high GC content that was significantly different to the amount of GC observed in sequences with a metazoan match (Fig. 8). Therefore, based on the GC content, the annotated *Hydractinia* EST dataset seems to contain two physically different kinds of sequences. This was confirmed by comparing the GC profiles of the *Hydractinia* sequences to those observed in other organisms including bacteria, cnidarians, invertebrates and vertebrates (Fig. 9) [91-94]. In the case of sequences without a functional annotation, the broad range of GC percentage suggests that some of them may have a GC composition characteristic of bacterial sequences. However, for the group of unknown sequences, the majority exhibited a low GC percentage suggesting a higher relationship to metazoan rather than to bacterial proteins. In contrast, most of the sequences with un-informative terms seem to be described by a bacterial GC profile. This can be expected since several bacterial annotations on the protein databases contain un-informative terms (Fig. 8).

These sequences are unlikely to represent a bacterial contamination, since a poly A+ selection and oligo-dT priming was used for mRNA isolation and cDNA construction, respectively. *Hydractinia* sequences with a bacterial hit can be divided into two different groups. The first group consists of 487 sequences, which were also found in the ESTs of *Acropora*, *Hydra* and/or *Nematostella* genome. Approximately two thirds of them might be present in the genome of *Hydractinia*, since 331 sequences were identified in the genome of *Nematostella*. At the nucleotide level, fewer bacterial sequences (24%) were shared within cnidarians, probably due to the suspected sequence divergence between anthozoans and hydrozoans [7].

Almost half of the sequences exhibited a best match to a particular class of bacteria (*Pseudomonas spp*.), which suggests that their appearance in the cnidarians may have been the result of lateral gene transfer (LGT) events. The fact that these sequences are being shared by *Hydractinia*, *Hydra*, *Nematostella* and *Acropora* suggests that the LGTs predate the Anthozoa-Hydrozoa divergence. Perhaps this transfer occurred only in ancient Cnidaria, or the transferred sequences were subsequently lost in other animal lines. Therefore, despite the fact that recent LGTs have already been observed in Cnidaria, an ancient common origin for the majority of these sequences is the favoured hypothesis [8, 102]. In accordance with the analyses done by Technau *et al*. [8] on *Acropora* and *Nematostella*, *Hydractinia* non-metazoan sequences containing introns (data not shown) and sequences with homologues in diverse organisms were also found, which argues against recent LGT events [8, 103].

The second group of non-metazoan sequences consists of 357 sequences with a bacterial hit and no counterparts in other cnidarians. It is possible to consider them as unique *Hydractinia* sequences, taking into account the suggested substantial variation in gene content within the Cnidaria [7]. In contrast to the majority of cnidarian sequencing projects done so far, adult material was included in the *Hydractinia* cDNA library. This may have resulted in the discovery of expressed genes related to an adult condition, for example genes related to nutrition or reproduction, which could not be detected in the other EST projects carried out using embryos [104, 105]. Consequently, the majority of these non-metazoan sequences were related to enzymatic activities. Nevertheless, for *Hydractinia* bacterial-like sequences without a clear genomic cnidarian representation, it is not possible to exclude symbiotic, parasitic or epiphytic bacterial sources. Commensal microbes or microbes living epiphytically on the exoskeleton are common in adult cnidarians as well as in higher metazoans [104-107].

4.1.4. Characteristics of the *Hydractinia* transcriptome and its contribution defining the cnidarian gene repertoire

Hydractinia homology analyses against other bilaterian organism revealed a substantial number of ESTs with a significantly higher sequence similarity to vertebrate sequences rather than to their fly, mosquito or nematode counterparts (Fig. 10). Additionally, 28 sequences with only a vertebrate homologue were found (Table 3A). Thus, despite having a small dataset, the *Hydractinia* ESTs corroborate a cnidarian ancestral genetic complexity, providing more examples of gene loss or secondary sequence modification in ecdysozoans [7, 8, 12, 14]. In contrast, fewer sequences possessed a higher similarity or were even uniquely identified in

the invertebrates analyzed (Table 3B). Apparently, we are also faced with genes that have been lost or are highly diverged in vertebrates.

One of the objectives in the generation of *Hydractinia* ESTs is a complementation of the information obtained from others cnidarian genome projects, identifying the genes maintained or added during the evolution of cnidarians. Comparing the *Hydractinia* ESTs to all other available cnidarian datasets, a list of 23 unique *Hydractinia* genes with known protein domain architectures were identified (Table 4). Despite the fact that some genes shared protein domains, their sequences did not overlap and were considered unique *Hydractinia* sequences. Examples of these are the six sequences showing a chorion or eggshell protein domain. These protein families are associated with a tissue- and temporal-specific gene expression pattern in ovaries, and are highly conserved in evolution [108, 109]. Their presence in our cDNA library may result from the inclusion of sexual mature female colonies in the mRNA pool rather than being *Hydractinia*-specific. Some of the identified putative proteins are unexpected and their functions are hard to interpret at present. For example, a sequence homologue to the vertebrate bone sialoprotein was found. This protein seems to be involved in bone mineralization and remodelling [110]. Another example is the Galanin receptor. In vertebrates this receptor is expressed in the peripheral and central nervous system, activating K^+ channels by coupling G proteins [110, 111]. In addition, several unknown sequences appeared to be unique to *Hydractinia*. For this result, two interpretations can be considered. First, as previously described, it is expected that several of these sequences represent short ORFs or non-coding sequences, resulting in fewer sequences that can be matched by BLAST. This holds true not only for the *Hydractinia* ESTs in question but also for the other EST databases that were used for comparison. Second, we may reconsider the option that the divergence of the Anthozoa and Hydrozoa is expected to be as extent as the protostomia and deuterostomia split. This implies large genetic differences and gene family diversity within the Cnidaria [7]. Indeed, there are marked differences in cnidarian morphology and physiology. Trying to extract genes which might be related to such differences, comparison of the databases resulted in a list of sequences probably linked either to physiological demands due to the environment (*e.g.* sea or fresh water) or to the colonial phenotype displayed by *Hydractinia* and *Acropora*. Despite the fact that most of the sequences identified in the first analysis showed an enzymatic (reductase, hydrolase) activity, which may correspond to the regulation of intracellular osmolarity, it is not possible to satisfactorily conclude a direct relation of these sequences to such physiological functions (Table 5A). The same holds true for the *Hydractinia* sequences shared only with *Acropora* (Table 5B). As most of these sequences are

unknown or associated with a diverse functionality, it is not possible to establish a firm linkage to colonial growth using only the bioinformatics tools currently available. However, such a linkage can be considered as a working hypothesis for further analyses.

To support the previous approach, semi-quantitative (sq) RT-PCR was performed on genes selected from the previous list, and some of them showed a specific expression pattern in the different life stages of *Hydractinia* (Fig. 11). For example, the unknown transcript Tai16A08 was highly expressed in the adult phenotype while exhibiting a relatively low abundance of other stages or even absence in the case of larvae. This indicates that the gene is probably related to development and is especially linked to the colonial condition. This result sustains the hypothesis which, by detecting a homologue only in the *Acropora* dataset, considers this transcript as a good candidate associated with the colonial phenotype. Another example is the sequence Tai11F02 encoding for malate synthase. (sq) RT-PCR showed a high transcriptional activity in primary polyp and a milder one in pre-planula and adult stages (Fig. 11). This correlates to previous observations done in *C. elegans* where this protein, specifically localized in the differentiating intestinal and body-wall muscle cells, exhibited an increased activity during embryogenesis but decayed in larval stage [112]. This enzyme is involved in the conversion of acetyl CoA into succinate. Thus, it might be related to the metabolic requirements in the developing embryo, pre-planula and primary polyp. The bioinformatics approach showed that this sequence was shared by all analyzed cnidarians except *Hydra*. Based on this, it is possible to speculate that this enzyme is functionally associated to a seawater physiological condition. For a satisfactory explanation it will be necessary to perform *in situ* hybridization analysis in order to better characterize the gene and its association to marine cnidarians.

4.1.5. The combination of bioinformatics and molecular tools leads to a better functional annotation

It is well accepted that sequence comparison between different organisms can be used to designate the function of unknown genes and even provide their evolutionary history [113]. However, it was previously demonstrated that the combination of bioinformatics and molecular biological approaches can lead to more solid functional statements. Especially in the case of organisms poorly represented in the public databases, sequence characterization with solely bioinformatics might lead to functional annotation bias. An example is the cnidarian tachylectin-related gene in neurons (*CTRN*) identified in the *Hydractinia* dataset

[114]. Tachylectin proteins belong to the group of pattern recognition molecules and its function in innate immunity is evolutionary conserved, from sponges to vertebrates [115-117]. Sequence comparison against other tachylectins and related genes revealed that *CTRN* presented a highly conserved structure. Therefore, solely from the bioinformatics analysis it will be logical to conclude that the *CTRN* gene is involved in immunity, but this does not seem to be the case in *Hydractinia*. Semi-quantitative RT-PCR performed for the different developmental stages of *Hydractinia* showed that *CTRN* is expressed after metamorphosis, with undetectable mRNA levels in the embryo and larval stages. This feature has already been observed in other immune molecules. However, it was unexpected that after LPS induction the mRNA expression level was not affected (this point will be further discussed in section 4.3.8). So far, no specialized immune cells have been identified in cnidarians, but it is reasonable to expect that cells producing immune molecules should have an efficient accessibility to potential pathogens, like in endodermal cells. In contrast, *in situ* hybridization analysis revealed that *CTRN* have a specific expression in the ectodermal tissues, restricted to particular neurons and their precursor cells around the mouth (Fig. S5 in Additional data 3, section 6.3) [114]. Thus, at the functional level the *CTRN* gene showed some discrepancies with respect to its homologue immune genes; having rather than a function related to immunity a role in neuronal development [114].

4.2 Technical aspects of the *Hydractinia echinata* microarray

Gene expression profiling experiments provide a straightforward approach to assign the functionality for many thousands of genes in a single assay [64, 71]. In the case of un-sequenced organisms, it allows a rapid identification of interesting transcripts for sequencing. This offers an economical alternative to redundant whole library sequencing methods. Therefore, in order to extract the maximal information from the generated transcriptome dataset, the EST project was supported with a microarray comprising the most representative cDNA sequences for each 3,808 generated EST clusters but also ~5,000 un-sequenced cDNA clones.

Microarray experiments, from the experimental design until the final analysis of the data, are a time consuming approach. It involves several distinct stages that have a direct impact in the final results. The biggest disadvantage is that the various factors affecting the construction, handling, target labelling and hybridization of the array are difficult to address before the final statistical analyses. Thus, it is necessary to repeat the experimental tests several times to optimize each laboratory protocol, in order to improve the quality of the generated data.

4.2.1. Construction of the cDNA microarray

As one of the fundamental microarray components, the amplified cDNAs printed on the array are the principal determinants of the hybridization outcomes. They are used to query the pool of differentially labelled targets in order to determine the relative expression level of each gene [71]. As explained before in sections 3.1.1 and 3.2.1, the *Hydractinia*-chip library used for the production of the probe contains cDNA inserts with a high size heterogeneity (Fig. 5 and 13). This highlights some of the pitfalls of cDNA microarrays. First, clone-collection handling is expensive, time consuming and can be a source of contamination. Second, long fragments are limiting steps for clone culture and PCR amplification [65]. To solve these problems all reactions were managed in a 96-well plate format, which together with the avoidance of plasmid preparation diminished the handling work and cross-contamination. Optimization in the culture of the clones and PCR protocols allowed the amplification of cDNA inserts with more than 6 kb. Clones with repeatedly negative amplification were separately handled or alternatively, other EST clones from the same contig, *i.e.* carrying the

same information, were selected as probe. This resulted in the successful amplification of 87% of the library-cDNAs with an average concentration of 150-300 ng/µl.

Size heterogeneity of the amplified probes can generate bias in the hybridization results due to for example; non-uniformity of the spots, differences in the available amount of probe and the annealing melting temperatures [65]. While short probes have poor hybridization efficiencies, longer fragments can have secondary structures, which might affect the binding kinetics [66]. Furthermore, it is necessary to consider the specificity of the probe. In case of family related genes or alternatively spliced variants, both short but principally longer probes, are prone to cross hybridization. It has been reported that if different targets have more than 70% sequence similarity to the cDNA probe, they can indiscriminately hybridize to such spots. Cross or missing hybridization particularly affects the analysis of transcripts expressed at low levels, since a small variation will be enough to provide significant false signals [118].

The quality of the spot directly influences the hybridization kinetics. If we consider a minimum PCR yield of ~150 ng/µl and a pin delivery of ~1 nl, each spot should contain approximately 0.75 ng of cDNA. Under this condition, and specially having long fragments, it is expected that GC rich regions can form secondary structures directly affecting the binding of complementary sequences. The use of Betain in the spotting solution diminished such structures equilibrating the effect of AT and GC base pairs in the stability of the DNA [119]. This additive also increases the viscosity, allowing the cDNA probes to be equally distributed on the spot surface. Betain also minimizes the evaporation rate resulting in longer ionic-binding reaction times between the negatively charged phosphate groups of the cDNA and the positively charged glass surface. All this helped to provide a maximum concentration of probe and a good homogeneity in the morphology of the spot, allowing the maintenance of a linear relationship between the detected signal intensity of a gene and their expression rate in the analyzed sample.

The selection of the highly hydrophobic aminosilane-slides as glass surface allowed, in spite of the hydrophilic feature of the printing solution, a relatively small spot diameter of 100 µm and a spot-to-spot distance of 140 µm (Fig. 14). 108 slides with 19,200 spots were produced, whereby 6 slides corresponded to “pre-spot slides”. The pre-spot run is essential for a continuous delivery of the same sample volume and for maintaining the spot uniformity. It induces the elimination of air bubbles within the pin reservoir or extra-drops in the tip borders, normally occurring in the probe up-take.

4.2.2. Hybridization of the cDNA microarray

An efficient labelling method is a critical parameter to acquire high quality microarray images [120]. For labelling the target RNA, the direct labelling approach was used due to its simplicity. With the incorporation of fluorescently modified deoxynucleotides during the first strand cDNA synthesis, molecule labelling is done in a single and cost effective step [120]. The enzymatic reaction was initialized with Oligo (dT) primers for two reasons. First, it focused on the labelling of mRNA molecules and not ribosomal RNA, which constitute approximately 90% of total RNA. Second, the EST project revealed a significant amount of sequences with bacterial hits, which may have come from a contamination source. Thus, poly (A) priming also diminished the chances to label such bacterial materials in the target sample. A disadvantage of this enzymatic approach is that the labelled cDNA predominantly represents the 3' ends of the mRNA, due to the limited processivity of the reverse transcriptase. In addition, incomplete denaturation of secondary RNA structures can shorten the cDNA copies [121].

To assess for dye imbalance incorporation in the reaction resulting in different product yields [122], always the same amount of labelled target was used competitively on the array.

Taking into consideration both array experiments, 80% of the hybridizations were of high quality, demonstrating a good optimization of the spotting, hybridization, washing, blocking and scanning protocols (see section 2.2.6 and 2.2.7). Preliminary Genepix analysis showed, besides the high quality foreground but low background intensities, a high reproducibility and a good spot morphology. However, approximately 30% of the chip presented a low or even absent signal (Fig. 20). In most cases those spots corresponded to the un-sequenced clones of the array. In all performed experiments, negative controls showed no cross hybridization, having an equal signal intensity level to empty spots or even background. This demonstrates that LORECs sequences are unique and can be successfully used as negative or spiking controls [75].

Signal intensities are proportional to the target concentration, but also to the hybridization time and the amount of immobilized material. It is described that only 0.1 to 1% of the immobilized molecules in each spot are still bound to the labelled target at the end of the experiment [68]. Thus, it might be possible to obtain more signals if the labelled starting material is increased. Glass microarrays have a sensitivity of approximately 2 X 10^7 molecules, relatively to the mRNA abundance; which means that this corresponds to the minimum number of molecules of a given sequence in the starting sample that can be detected

after hybridization [68]. Assuming no background effects, the signals are multiplied by a factor of ten if the starting material is ten-fold concentrated. As mentioned in section 3.3.1.3, different amounts of starting target-materials were tested without improving the detection sensitivity. Yet, better signals might be obtained if more than 15 µg of total RNA are used. However, in most cases and including our experiments, sample availability is the limitation factor.

4.2.3. Experimental design

The design layout of a microarray experiment is essential to estimate the precision and statistical power of the analysis. For both, mitomycin and immune microarray experiments, an even double reference design was followed (Fig 19 and 30). This permitted to indirectly compare every condition with each other, as they are directly compared to the same reference. Moreover, in this even design dye bias is not affecting the estimates of gene expression because every sample is labelled with both dyes, and each differently labelled sample is used equally often in the experimental layout [65]. Technical replicates of the assay and spot duplication resulted in 12 data points per gene for each compared condition, allowing robust statistical analysis. For the representation of the biological replicates, a pooling strategy was followed. This approach requires less RNA material to be hybridized in the array. Pooling equalizes the variability of the samples (*e.g.* different genotype, age, *etc.*). Yet, that variability cannot be measured and therefore, it is not possible to determine how such variability affect the final results [65]. Nevertheless, organisms with the same genotype (clones) were used in the mitomycin and LPS experiments. In contrast, in the allorecognition challenge sample material was only extracted from the contact area between clone members of the previously used colony and a genetically distinct one. Therefore, it is necessary to take into account that the genetic variability of the colonies is not assessed in this experiment. To provide higher robustness and diminish the influence of a different genotype in the expression profiles of allorecognition induced organisms, this data was analyzed together with the one obtained from the LPS treated organisms as a multiconditional experiment in M-CHiPS [72, 73]. This resulted in the selection of genes related to both experimental conditions.

4.2.4. Analysis of signal intensities

To adjust for any bias affecting the average ratio of the two dyes due to technical variation rather than biological differences, M-CHiPS re-scaled the raw intensity values [71, 73]. There are different strategies to normalize gene expression data, but all start with the selection of genes for the fitting normalization algorithms. All the genes present in the array were used, since they represent the complete transcriptome of *Hydractinia* and many of them have an invariant expression level in the condition and reference samples. This can be observed directly from Fig. 21 and 31, where the majority of the plotted gene intensities of the different samples clustered along the diagonal line. One of the advantages in normalizing with all spotted genes is that the algorithm properly estimates the spatial and intensity-dependent trends of the data [70].

The intensity data was standardized based on a logarithmic regression model. The distortion of the measured intensities influenced by the background was corrected subtracting an additive constant or offset. To account for the multiplicative factors affecting the gene intensities due to, for example, different labelling rates; the medians of the intensity ratios (Cy3/Cy5) from equally expressed genes were used as an adjusting factor, such that the ratio for these genes becomes one [71, 73, 74, 87]. In all mitomycin and immune array hybridizations, the normalization data presented good fitting performance with correlation values between 0.85 and 0.98.

After normalization, M-CHiPS was used to remove all genes with an invariant expression profile or a poor reproducibility, while selecting those with a significant evidence of being differentially expressed across the conditions. The first filtering criteria eliminated all genes with saturation effects. Signal saturation proportionally increases the bias estimation of the gene expression level. However, it is important to consider that saturation is directly related to the settings of the photomultiplier tube (PMT) and RNA abundance, and it is not necessarily associated with a poor quality of the spot [123]. This suggests that just highly expressed genes are being eliminated, which might be particular informative for a certain condition. To diminish the saturation effects, the lowest PMT values were selected while maintaining a good signal-to-noise ratio for spots having a low intensity. In all microarray experiments only 1% of the genes still showed saturation, and most of them presented such an effect in both, condition and control measurements.

Spots with low intensities, similar to the background, may display notable ratios due to measurement fluctuations. To avoid this bias a second filtering criteria was used, which

selected all genes having at least in one of the conditions a considerable absolute expression level. This filtering step eliminated approximately 70% of the microarray data. It is important to mention that in two-channel data, low spot signal intensity could occur not only because of a low concentration of the corresponding mRNA in the sample but also due to the saturation of all binding possibilities of the probe by one of the target mRNAs. Therefore, it is necessary to treat the normalized intensities as ratios of the relative mRNA expression in the condition with respect to the control [73]. In subsequent analysis, the logarithm of the expression ratio was used. In contrast to the solely expression ratio, the logarithmic function treats the values symmetrically and does not limit down-regulation genes to a scale between 0 and 1 [71].
Quality filtering steps were applied as final selection criterion. A statistical method specifically adapted for microarray (SAM) was used, based on the assimilation of a set of gene-specific t-tests [88]. The aim was to identify a large number of differentially expressed genes with a minimum of false positives. Genes were selected if they exhibited a significant and reproducible (corrected p-value < 0.05) different expression level between at least one condition and the control. In addition, only genes with a 2-fold change in expression were considered. In both microarray experiments, this filtering resulted in the selection of 1 to 3% of all analyzed data (Fig. 21 and 31). They represent the best candidate-genes to be associated with a mitomycin treatment or an LPS and allorecognition response.

4.2.5. Finding genes with common expression patterns

The final and most important goal of microarray experiments is to analyze several hybridizations in order to identify genes having a common expression pattern [71]. This suggests that those genes are more likely co-regulated under a particular condition and therefore, perform similar or related biological functions. There are different methods to cluster microarray data. In the present project, data was first analyzed in M-CHIPS using Correspondence Analysis (CA). This projection method, similar to Principal Component Analysis (PCA), successfully represented in a two dimensional scenario all possible associations between and within the genes and hybridizations (Fig. 22 and 32). However, one of the drawbacks of CA is the low precision to define the borders of the clusters [124, 125]. CA analysis was complemented with two other methods, Hierarchical clustering and k-means algorithms. In general we can consider that both methods confirmed the CA results, but slight differences were observed in the generated clusters. It is necessary to take into account that clustering results are sensitive to the used algorithms, normalization and distance metrics [71,

126]. Hierarchical clustering has the advantage of being a simple approach with an easy visualization (Fig. 23 and 33). But this method is also imprecise in defining the clusters, and there are no confidence values to support the performance of the algorithm. In addition, this is an agglomerative method and therefore, wrong assignment in the first clusters cannot be corrected. This means that subsequent clusters are constructed based on false assumptions [71]. This could be one of the reasons why several genes having the same function were distributed in different clusters.

Alternatively, the k-means clustering approach was tested. This partitional algorithm orders the genes into a fixed number of clusters which are internally similar but externally dissimilar and therefore, avoid the bias discussed above [71]. Nevertheless, in this case it is necessary to determine the number of clusters in which the data will be distributed. To define the optimal number of clusters for k-means, a figure of merit algorithm was used. The analysis showed for both arrays that in the first KMC generated clusters the calculated FOM values drastically decreased (Fig. 24 and 34). This correlates to the CA biplot, where in a first view the data can be mainly distributed in few main clusters. However, both cases also suggest that additional clusters might provide a better overview of the relationships of the expression data. The CA biplots clearly show the presence of sub-clusters, especially in outlier genes (Fig. 22 and 33). Correspondingly, the curve representing the calculated FOM values continuously decreased, whereby the predicting power of k-means increased (Fig. 24 and 34). Different numbers of clusters for k-means were empirically tested, finally deciding that it performs optimally for 15 and 21 clusters in the mitomycin and immune array, respectively. The addition of a colour-coded GO term annotation in the graphic representation of the k-means clusters provided maximal functional information of the candidate genes.

While there is no perfect clustering classification method, there are some that are more appropriated for a certain data-set [71]. In this project it is possible to confirm that the combination of different approaches improves the detection of data relationships, resulting in a powerful tool to group genes with similar expression patterns.

4.3 *Hydractinia* microarray experiments

The unexpected representation of most vertebrate gene families in cnidarians resulted in an emerging interest to trace the evolutionary origin of different metazoan features, like the regeneration of stinging tissues by stem cells, and the response to infection or allogeneic reactions. While bioinformatics approaches have already started to reveal such traits in cnidarians, the acquired EST information was combined with a microarray to directly address the genetic repertoire of the stem cell and innate immune system of *Hydractinia*.

4.3.1. The use of mitomycin-C to target the i-cell population

Stem cell properties of the interstitial-cell lineage have attracted significant attention since decades. Recent studies reported that several metazoan stem- and germ cell genes are represented in cnidarians [19, 40]. However, it is still unclear if the identified cnidarian homologues are also functionally conserved. Furthermore, the complete gene repertoire and the associated molecular pathways involved in the maintenance and differentiation of cnidarians stem cells is unknown.

The first step to approach this problem was to identify genes associated with the i-cell lineage. To achieve this, following the experiments of Müller and colleagues, the i-cell population was partially and completely depleted from *Hydractinia* colonies using the antibiotic mitomycin-C (MMC) [27, 41]. This drug is normally used in the treatment of cancer and other tumours due to its capacity to interfere with DNA replication, leading to rapidly induced cell death [127]. Thus, all *Hydractinia* cells having a high division rate were targeted by MMC. These included proliferating stem cells and uni-potent germ cells. In addition, cells committed to differentiation were also targeted since previous studies suggested that these cells undergo one or more cell divisions before complete differentiation [26]. So far, the fraction of the i-cell population that actually functions as stem cells and the number of the different stem cell sub-types is unknown. It is expected that a subpopulation of progenitor cells, which are proliferating cells in between the multipotent and the committed state, are present in *Hydractinia* [26, 28]. Accordingly, Giemsa/May-Grünwald staining of i-cells showed that in spite of their relative homogeneous morphology, cells presented slightly differences in size (Fig. 16). It has been proposed that large cells correspond to stem cells, while relatively small cells resemble the presence of intermediates or differentiating cells [19].

4.3.2. Microarray analysis of colonies treated with mitomycin

Drug doses were adjusted to specifically target the i-cell population but not the differentiated cells. To achieve this, the treatment started with a lower amount of drug compared to the one suggested in previous publications, but always maintaining a periodical delivery [27]. Only after applying higher doses (30 μM) the first phenotypical changes were observed in the colonies (Fig. 15). As expected, colonies with different genotypes varied in the response to the drug. In the F0 clones the drug rapidly eliminated all dividing cells, resulting in severe apoptosis and necrosis all over the colony. Cytological examinations confirmed the probably complete but not exclusive elimination of i-cells (Fig. 16D). Based on these results, F0 colonies were discarded for further analysis. In contrast, drug responsiveness of the FM and K12 colonies was milder, but more accentuated in the latter. K12 animals were clones of a FM colony and therefore, genetically identical organisms. The different age, size, thickness of the stolon mat and the chitin-layer may have differentially affected drug uptake, resulting in the slight time-shift drug response observed in these two clones. At 96 hours post treatment, significant phenotypic changes with respect to the control were observed (section 3.3.1.1). Cytological examinations corroborated the exclusive elimination of most i-cells (~90 %) from the K12 clones. In the case of the FM colonies, the amount of remaining i-cells (~52 %) was significantly lower than the control (Fig. 16A-C).

Recovery from the mitomycin treatment was only achieved in the FM colonies (Fig. 17). These were the only clones having a stable junction with the grafted donor-explants. Generation of new tissues and polyps was not exclusive to the grafted regions, taking place all over the colony. It is known that i-cells can migrate a considerable distance [27]. However, recovery may also have occurred due to the survival of a significant number of stem cells rather than from the newly populating stem cells acquired from the donor.

With the microarray, i-cells related genes were identified by comparing these three distinct phenotypes -the mild and strong i-cell depleted colonies and a recovered condition- against the same reference control (Fig. 19). From this first analysis it is not possible to determine which particular cells are expressing those genes and therefore, distinguish different cell types within the i-cell population. Yet, 162 good candidate genes could be identified which might differentially mark such cells.

The comparison of the logarithmic fitted intensity from all conditions with respect to the reference, demonstrated that the expression level of most of the genes spotted on the array were not affected during the treatment (Fig. 21). This not only supports the fact that the array

represents a broad variety of genes, but also that the drug treatment affected a particular population of cells rather than the complete organism. As expected, most of the differentially expressed genes were associated to and down-regulated in the strongly i-cell depleted phenotype (K12, Fig. 21-22). Few genes were differentially transcribed in the other two conditions and even less were up-regulated.

Correspondence analysis allowed a clear overview of the transcription profile of each different condition. Correlating to the phenotypic changes observed after the MMC treatment, organisms differed in their expression pattern proportional to their drug response. This can be observed in the clockwise distribution of the condition-clusters in the plot, where most of the detected expression changes occurred in the K12 phenotype and were closer to FM (mild depletion of i-cells) than FMR (recovery). Subsequently, k-means was used to analyze this data in detail and distribute the genes into clusters of similar expression pattern.

4.3.3. Genes associated with organisms having a mild response to mitomycin

After drug treatment, FM colonies still presented a substantial amount of cells from the interstitial lineage. The array results support the hypothesis that, mainly those surviving i-cells committed to recover the colony, regenerating new stolon tissues and polyps. K-means analysis showed that some of the genes specific to the FM condition are related to a metabolic or detoxification function. For example, a gene encoding for glutathione S- transferase was identified (cluster 1, Fig. 25). This family of enzymes, besides their function in cell signalling and S-glutathiolation, is involved in the detoxification of lipid peroxidation products as well as in anti-neoplastic drug resistance of carcinogens and xenobiotics in germ- and other cells [128-130]. In addition, several genes encoding for RNA binding proteins were identified, suggesting that the cells of the FM condition exhibited active mechanisms of post-transcriptional regulation, e. g. splicing [131]. Metabolic regulation is also supported by the presence of kinases and other catalytic proteins (cluster 12, Fig. 25).

Genes with a high expression in the FM phenotype were also distributed in cluster 9. In contrast to the previously described clusters, these genes presented a broader spectrum of expression with a strong inactivation in FMR but some of them highly expressed in K12 (Fig. 28). Nevertheless, these clusters shared the identification of several growth and transcription factors. For example, putative genes encoding for BMP-2 and -4 were identified. Despite the relatively low score of their BLAST similarity match, domain analysis identified a TNRF cysteine-rich domain at the N-terminal region and a TGFβ/Netrin-module (non TIMP type)

domain at the C-terminal part of these sequences. BMP proteins can be considered as multifunctional growth factors of the TGFβ superfamily associated with the development of different tissues [132-134]. In cnidarians, BMPs are involved in axial patterning, neurogenesis and epithelial differentiation. In addition, it has been shown that BMP 2/4 expression levels increased in stress responses, *e.g.* wounding [135].

Several genes encoding for Trefoil factors (TFF) were also identified. In higher vertebrates, these peptides are involved in the integrity of the mucous epithelia, influencing migration of cells and actively participating in tissue healing by a process called restitution [136, 137]. After damage, TFF peptides and epidermal growth factors (EGF) are highly expressed in the nearby tissue. This suggests that the interaction of these peptides is beneficial in the defence and repair of the mucosa [137]. Moreover, in vitro studies demonstrated that under the presence of glutathione, human-Trefoil factors are able to promote proliferation. Interestingly, our list of genes in cluster 1, 9 and 12 includes not only the *Hydractinia* homologue to *TFF-peptide* but also genes carrying EGF domains and encoding for glutathione. The presence of genes involved in the protection and regeneration of damaged tissue is also sustained by the identification of several Rhamnose-binding lectins (Fig. 25). In cnidarians, the function of these molecules is still unclear but it is suggested that beside their role as PRRs they might act in tissue remodelling and repair [138].

The active response of i-cells in detoxification, proliferation, wound healing, and cell fate determination (including epithelial muscle or nerve cells) is further supported with the identification of the basic leucine zipper (Bzip) transcription factor Mafl, zinc finger proteins and astacin metalloproteinases [139-142].

4.3.4. Genes associated with organisms having a strong response to mitomycin

As expected, colonies exclusively and almost completely depleted from their i-cells exhibited a down-regulation in the expression of several genes (Fig. 26 and 27). Indeed, only one cluster with 8 genes showed activation in the gene expression in this phenotype (cluster 3, Fig. 28). The identification of genes without or with a decreased expression in the K12 condition with respect to the control should directly reflect that those genes are functional associated with i-cells. As previously mentioned, the i-cell population comprises a heterogeneous group of cells including stem and differentiating cells. This was corroborated by the microarray results, showing that the MMC treatment eliminated cells committed to nematocytes production. This was extrapolated from the identification of more than 15 genes

encoding for three different types of minicollagens. These are small collagens-like proteins, major components of the capsule wall and tubule of nematocyst and therefore, specific markers of nematogenesis [143, 144]. It is also possible to speculate that neuron precursor cells were targeted since it is thought that bipotent stem cells give rise to both nematocytes and gland cells [145].

As in the FM condition, several genes encoding for metabolic proteins with probable roles in the development of new tissues were identified [146]. This included the RNA-protein binding cabeza, RNA-binding region RNP A1, RAD23 and the E3 ubiquitin protein ligase (Fig. 26, clusters 2 and 8). Interestingly, the RNA-protein binding cabeza transcript showed high sequence similarity to two human genes, *TLS* (Transcribed in LipoSarcomas) and *EWS* (Ewing's Sarcoma). It has been shown that if these proteins fuse to transcription factors (*e.g.* C/EBP), they can induce chromosomal translocations leading to tumour formation [147].

Another gene following this expression pattern was the homologue of the *Hydra magnipapillata CnPL10*. PL10s are members of the DEAD-box RNA helicase protein family and have an important role in translational control. The identified sequence is relatively short representing the N-terminal region of the transcript and is probably the reason why there is no match to the *Hydractinia* homologue already identified by Mochizuki and Fujisawa. The down-regulation of *CnPL10* in the K12 colonies correlates to the observations done in *Hydra*, where this gene is expressed in multipotent and germline stem cells, ectodermal epithelial cells in the body column, and differentiating cells of the interstitial lineage [95].

It is appropriate to mention that, despite the harsh MMC treatment, in the FM or K12 conditions only one activated gene was associated with an apoptotic function. This was the gene encoding for Cathepsin, which besides its function in degradation and digestion of phagocytosis products it has also been associated to programmed cell death [148]. Indeed, most up-regulated genes exhibited a catalytic function -like amidohydrolase- or a structural activity –including fibrillar collagen precursor or clathrin proteins.

4.3.5. Transcriptional profile of the recovery FMR phenotype

It was expected that expression profile analysis of recovering colonies will identify genes associated with migratory donor i-cells and to the ones actively regenerating damaged tissues; including stem-cells and cells in their differentiation process. In comparison to the control, such genes should be up-regulated in this condition and correspondingly down-regulated in the i-cell depleted phenotypes. As discussed above, the recovery of FM may have occurred

due to the migration of donor i-cells but also because a significant amount of stem cells survived the MMC treatment. Unfortunately, there was not enough material to determine the amount of donor or host i-cells in the FMR organisms by cytological examinations. Recovery occurred in small spots spread all over the colony, and despite that RNA was isolated just from these regions, the amount of new emerging polyps or stolon tissue was significantly lower than the untreated colony. This is supported by the fact that almost all k-means clusters showed a down-regulation profile in the FMR condition. Only one cluster included up-regulated genes encoding stress related proteins (*e.g.* Heat shock proteins).

The FMR phenotype helped performing a robust multiconditional experiment analysis. However, most of the information obtained in the mitomycin experiment was retrieved from the FM and K12 conditions. The detection sensitivity of our experimental design did not allow to identify the stem- or germ- cell markers already known in cnidarians; like *Sox*, *Nanos* or *Vasa* [19, 40, 95]. Probably stem- or germ- cell related genes are not expressed at high levels and therefore are difficult to be represented in a RNA sample [95]. Considering that a colony comprises cells in heterogeneous stages, the pooling of RNA material will diminish the possibility to detect lower expressed transcripts. The previous fact does not only affect the target sample but also the RNA and cDNA material used as a probe for the array.

Nonetheless, this first array analysis identified several genes associated to the i-cell population of *Hydractinia*, including transcription or growth factors. In addition, various genes involved in detoxification and wound healing were found. More interesting is the fact that many of the identified sequences are still unknown. For a better characterization of these genes, it is necessary to acquire their complete open reading frame sequence and combine bioinformatics functional annotation with *in situ* hybridization experiments. This will not only describe the action of previously unknown genes but together with the already annotated ones will help to define and characterize the i-cell population of *Hydractinia*.

4.3.6. Identification of genes associated with the *Hydractinia* immune system

Genomic data suggests that the eumetazoan ancestor had a complex tool kit for defence against pathogens and aggressors. However, with the available information it is still unclear to which extent the innate immune system of higher metazoans derives from cnidarians. The genetic diversity of the phylum Cnidaria further complicates this analysis, since not all the identified genes are equally represented in the different cnidarian classes (section 1.1.4.1) [16]. Moreover, not all cnidarians have demonstrated the highly specific system of allo- and

xeno- recognition observed in anthozoans and hydrozoans. This particular feature allows considering *Hydractinia* as a suitable model organism to define the immune gene repertoire of the cnidarian ancestor [1, 138].

To identify genes involved in the immune system of *Hydractinia*, the microarray was used to analyze the expression profiles of animals with two different immune responses (Fig. 30). First, the colonies were incubated with LPS mimicking a Gram-negative bacterial infection. Expecting a fast immune reaction in the host, two different time points were analyzed after the LPS induction. Second, colonies were challenged by allogeneic contact. To mainly represent genes involved in an allogeneic reaction in the target-sample, the RNA used in the array experiments was exclusively isolated from the contact area of allogeneic colonies showing signs of rejection.

First analysis of the microarray data showed that most spots correspond to housekeeping genes (Fig. 31). This allowed a good normalization of the intensity values, which together with a robust statistical analysis identified 245 genes being significantly differentially expressed in the analyzed condition (Fig. 31). While some genes were associated with a LPS response, the allorecognition reaction seems to affect the expression level of several hundreds of genes demonstrating a complex process. CA not only sustained this result but also reported the high reproducibility and quality of the produced microarray data. In CA, all LPS hybridizations clustered quite close to the control and centroid of the biplot (Fig. 32). This suggests that LPS has a mild impact on the gene expression level of few genes. In contrast, the allorecognition condition spread out from the centroid of the biplot, providing the major difference in the expression data. This is also clearly observed in the amount of either up or down-regulated genes associated with this condition. It was also possible to identify two sub-clusters with genes that were initially up-regulated after 1 hour of the LPS induction but were down-regulated in the subsequent 3 hours. The reverse situation was also observed (sub-cluster I and II, Fig. 33). Interestingly, in these two sub-clusters, animals at 3 hours post LPS induction displayed a similar expression pattern as the allogeneic challenged colonies (Fig. 33). K-means clustering was used for the detailed analysis of the gene expression patterns.

4.3.7. Genes associated with organisms undergoing allorecognition

The ability of multicellular organisms to discriminate between self and alloantigens should be modulated by the antigens itself, recognition molecules and positive or negative reaction pathways [54, 149]. This suggests that several genes must be involved in both, fusion and

rejection processes. Colony rejection resembles an inflammatory response, where the interacting tissues swell due to the active migration of nematocytes aimed to discharge toxins and damage the competitor [144, 148]. In the allorecognition condition, an increased production and activity of nematoblasts was confirmed by the up-regulation of the nematoblast marker *CnPL10* and several different genes encoding for mini-collagens (Fig. 35, and Fig. S2 in Additional data 2). The diversity in nematocysts morphology is achieved by the different types of minicollagens and their disulfide association with an additional capsule protein, NOWA [143]. Hydrozoans have the largest repertoire of nematocysts within the phylum, with the representative *Hydra* comprising 17 different minicollagens [143, 144]. Based on their N and C terminal cysteine rich domains (CDR), minicollagens are organized in three groups. In our list of genes, the *Hydractinia* homologue to the *Hydra* minicollagen 1 (*Hm-Ncol1*) was identified; which with identical cysteine pattern in their N- and C-terminal CDRs belongs to the group 1. In addition, a gene encoding for the *Hydra* minicollagen 3/4 was found; which with their variable CDR motifs are organized in group 3. While additional minicollagen related genes were identified in the array, it is necessary to acquire their complete cDNA sequence for a proper characterization. Other collagen-like peptides were detected, of which their associations with nematocysts must be further analyzed (Fig. 35).

The activation of genes involved in the mitochondrial electron transporter chain, like the ABC transporter or cytochrome c-type, suggests that allorecognition is a highly demanding energy process [148]. Active metabolic regulation occurred since several genes with a nucleic acid binding or splicing related activity were identified (Fig. 35). It is expected that in an antagonistic response oxidative stress may play an important role in cell damage and correlating with this, we found several enzymes with a protective function against reactive oxygen metabolites including; oxidoreductases, the antioxidant zinc-metalloproteinase and the slightly activated glutathione-S transferase (Fig. 35, Fig. S2 in Additional data 2) [128, 148]. While this kind of protective activity was taking place, nematocysts toxins delivery was probably supported by preliminary digestion of the opponent tissue using degrading enzymes such as collagenase (cluster 1, Fig. 35) [150, 151]. In addition, a gene encoding for a deoxyribonuclease was found in our data. This protein has been reported in nematocyst extracts and it is thought that at least in the anemone, they exhibit a hemolytic effect [150].

A significant amount of genes were down-regulated during the allorecognition response (Fig. 36). Six of them were lectins. In *Hydractinia*, they probably act as secreted PRRs binding conserved surface epitopes of microorganism invading the colony. Alternatively, their function might be also related to development, tissue remodelling and repair [114, 138, 148,

152]. Previous analysis done in the tunicate *Botryllus schlosseri* have shown an up-regulation of lectins following allogeneic contact and suggested that these genes are actively involved in allorecognition [148]. In *Hydractinia*, different lectins, *e.g.* Rhamnospondin (*Rsp*) and Tachylectins, are constitutively expressed in a ring-like pattern around the polyp's hypostomes (see section 4.1.5) [114, 138]. This restricted expression pattern and the fact that the RNA for allorecognition sample only included rejecting stolon tissue bearing just a few polyps, leads to the assumption that the latter sample contains less amount of lectin mRNA with respect to a colony highly populated with polyps. Therefore, in this particular case it is not possible to satisfactorily define the expression profile of these genes during an allorecognition process. But we can confirm the presence of different lectins, suggesting that a high variability exists in the immune-molecule repertoire of *Hydractinia* [138] .

With the same down-regulation profile during allorecognition, genes encoding for Phospholipase (PLA2) and Conodipine were found (Fig. 36). These proteins have been isolated from a wide variety of venoms as well as from mammalian pancreatic and inflammation fluids. Besides their catalytic activities on lipid metabolism, some of the PLA2s show potent pro-inflammatory, antimicrobial or neurotoxicity effects [153-155]. According to these descriptions and to their presence in nematocysts, it would theoretically be expected that these genes are activated in allorecognition [150]. Instead, their down-regulation suggests that in nematocytes, the mRNA level of such venom related genes does not correlate to the amount of their final protein products. It has been described by immunocytochemical approaches that in the resting nematocysts this toxins are stored in the outer membrane of the inverted tubule. Upon discharge, the toxin is translocated to the internal surface of the everting tubule and its delivery by the dart-like spines occurs in the nanosecond scale [156]. Under these conditions, one could expect that nematocysts reaching the points of rejection (POR) carry toxins ready for delivery. In addition, because of the explosive discharge of the toxins, slightly different time frames might vary significantly the transcriptome analyses. Alternatively, as already observed in *Hydra*, such toxins might also be found in tissues devoid of nematocytes. This suggests that such proteins perform multiple roles in the organism. Particularly, PLA2s fulfil the criterion of a bioactive protein involved in toxic and digestive roles [143, 150].

Tissues in the periphery of the colony are devoid of undifferentiated cells [27]. Therefore, the use of such tissue in the allorecognition experiment resulted in a down-regulation of several genes encoding for transcription and growth factors; including the already identified trefoils, BMP-4 and fibroblast growth factor-2 (FGF-2) [157]. Moreover, several kinases with the

same expression profile were identified; including PANK, which have been shown to physically regulate the intracellular concentration of coenzyme A (CoA), and mpk-1 involved in a variety of cellular functions including proliferation, differentiation, development and transcriptional regulation [158, 159].

When colonies get into contact, incompatibility reactions are observed in the reactive stolons but not in adjacent tissues. It is known that prolonged allogeneic contacts affect the entire colony, but in the initial phase of the reaction, allogeneic responses are rather local. Although some tissues are actively delivering nematocytes, others are recruiting several genes to recover the colony, like for example *RAD23* or the antioxidants genes previously described. Thus, allorecognition seems to be a complex process that includes a whole genome response with up- and down-regulation of several genes [148]. Highly interesting are those 60 genes from our list which are still unknown. In order to better characterize those, and further analyze the rejection processes, the candidate's genes must be completely sequenced and *in situ* hybridization analysis has to be applied.

4.3.8. Genes associated with organisms having an LPS challenge

Lipopolysaccherides (LPS), peptidoglycans and glucans have been successfully used to activate the invertebrates and vertebrates immune system [160, 161]. Particularly in lower metazoans, Gram-negative bacterial infections mimicked by the use of purified LPS resulted in the identification of pattern recognition receptors and antimicrobial peptides [138, 162, 163]. In the current analysis 70 genes being specifically associated with a response to LPS exposure were identified, from which 42 were slightly up-regulated directly after the immune challenge (Fig. 37). From these, half of the genes are unknown and most of the ones with a functional annotation are related to a binding activity. For example RNA-binding molecules and translation elongation factors were detected, which suggest the activation of post-transcriptional regulation mechanisms. In addition, the presence of histone genes indicates that chromatin modifications also took place during an LPS response. It is well described that histone H1 facilitates the condensation of the chromatin fiber and therefore, regulates the expression of specific genes [164]. Recent reports demonstrated that besides the well accepted transcriptional repressor activity, H1 can also up-regulate gene expression [165]. In our analysis, histone activation was further demonstrated with the identification of a gene carrying a H5 domain which in mammals can replace the function of H1 in certain cells [166]. These histones seem to affect the expression of a variety of genes with functions related to *e.g.* cell cycle, cellular development, growth and proliferation, cell-cell signalling, drug and nucleic

acid metabolism [165]. In addition to transcriptional and translational regulation, the cell-cycle was probably actively regulated by the high expression of G2 cyclin (cluster 19, Fig. 37). Cyclins are able to bind cyclin-dependent kinases (CDKs), and the resulting complex is involved in all cell cycle transitions. In *Hydra*, cyclins transcription expression level dropped down immediately after injury but subsequently, increased especially in regions populated with highly proliferating cells [167].

At a first glance, the genes described above have a relatively broad functionality and therefore, do not satisfactorily describe a specific response to LPS. Moreover, these genes followed non- or a down- regulation in their expression level at 3 hpi of LPS. This supports the idea that they merely represent a stress response to the treatment (Fig. 37). However, 28 genes that were initially not affected or even down-regulated by the LPS treatment, started to be actively expressed at 3 hpi of LPS (Fig. 38). Sequence analysis demonstrated that most of these genes have a clear functional relation to an immune response. For example, genes encoding the Heat shock protein 70 were redundantly found (cluster 16, Fig. 38). In addition to their primary function as molecular chaperones, Heat shock proteins from the HSP60, HSP70 and HSP90 families, are potent activators of the innate immune system [168, 169]. It has been shown that inflammatory responses mediated by LPS accumulate Hsp70 and plasma pro-inflammatory cytokines [170]. Hsp70 and cytokine effects are probably mediated via Toll-like receptor signal transduction pathways towards the activation of NF-κB and MAPKs [169]. In addition to these results, a gene encoding for a lipocalin-like protein was identified (cluster 1, Fig. 38). Lipocalins or lipid binding proteins have diverse functions, including among others nutrient and pheromone transport, control of cell cycle and synthesis of prostaglandins. But they are also involved in the innate immune response to bacterial infections [171-173]. In higher vertebrates it has been described that LPS activation of Toll like receptors resulted in a stimulation of transcription, translation and secretion of Lipocalin 2. The secreted protein mediates the host defence against infection by sequestrating iron, which is essential for the growth and activity of most bacteria [172, 174, 175].

Antistasin, a serine protease inhibitor, exhibited a similar expression profile as the previously described genes (Fig. 38). Former analyses in cnidarians (*e.g. Hydra*) suggested that this anticoagulant might function in digestion of prey but also in the protection of the mucous cells from its own digestive enzymes [176]. Alternatively, as already observed in other metazoans, this gene might be directly involved in the regulation of inflammation [177].

The LPS challenge also stimulated the expression of growth and transcription factors as well as enzymes involved in detoxification, wound healing and cell differentiation (Fig. 38).

Although PRRs were expected to be activated after the exposure of the colony to LPS, most of the genes encoding for lectins showed no significant changes in their expression level. This result correlates to previous reports done in cnidarians, where tachylectin and rhamnospondin (*Rsp*) genes presented an invariant expression profile after the incubation with LPS, bacteria or fungi, respectively [114, 138]. It is thought that the constitutively secreted Rsp molecules act as a mouth immunological filter, but they also may have functional roles in remodelling and repair. In the case of tachylectin, its involvement in neuronal development is suggestive, based on its spatial and temporal expression pattern (4.1.5, see Fig. S5 in Additional data 3) However, the possibility that tachylectin molecules are able to specifically identify others PAMPs and not LPS cannot be excluded. Interestingly, two genes encoding for a cnidarian like tachylectin were identified in the microarray experiment. One of them (HEAB-0028L01) confirmed the previous analysis and presented an invariant expression level after the LPS exposure, but the other (HEAB-0031J15) showed a 2 fold up-regulation at 3 hpi of LPS (Fig. 38 and Fig. S3 in Additional data 2). These sequences are ~70 % identical among each other and 63-66% identical to the already identified *CTRN* (section 4.1.5). From this first analysis, it is possible to speculate that the several point mutations identified in the sequences resulted in the polymorphism of the gene. This suggests that these genes are specialized to bind different targets. Further genetic and functional analysis should be done to clarify their diversification and functionality.

The gene with a Thrombospondin type 1 repeat (TSR) also deserves special attention (Cluster 16, Fig. 38). The TSR domains are present in a large number of proteins involved in cell-cell and cell-extracellular matrix interactions. Several TSR proteins also have a direct role in immunity. Indeed, the *Hydractinia* Rsp molecule contains eight tandemly repeated TSR domains which probably act as an extracellular signalling tag for cross-communication with other immune molecules or receptors [138].

This first microarray analysis identified several immune related genes for a detailed functional characterization. Particularly interesting are those transcripts which still have an unknown function. In order to better describe how *Hydractinia* responds against an infection, the complete gene sequences will be acquired and whole mount *in situ* hybridizations will be performed.

4.4 Conclusion and future perspectives

This project is the first high-throughput effort aimed to identify and characterize the transcriptome of the colonial marine hydroid *Hydractinia echinata*. The generated EST dataset, supported with a database harbouring all the acquired information and a microarray for transcriptional profiling analysis, provides a platform to promote and facilitate molecular research not only in *Hydractinia* but also in other cnidarians.

The *Hydractinia* ESTs confirmed that cnidarians have a remarkable genetic complexity, with a pattern of a high gene-sequence maintenance and relatedness to vertebrates rather than to ecdysozoan invertebrates. In addition, the *Hydractinia* non-metazoan sequences found in the different cnidarians corroborates the present view that a substantial number of ancient prokaryotic genes have been maintained in cnidarians' genomes and were either cnidarian-specific or lost in other metazoans [7, 8, 11]. The detection of genes specific to *Hydractinia* demonstrates that the cnidarians analyzed to date do not represent all the features present in the phylum and therefore, a better overview of cnidarians might be possible with additional sequencing data from different basal metazoans.

To characterize the genetic repertoire associated with the i-cell population and the innate immune system of *Hydractinia*, transcriptional profiling experiments were performed which identified 162 and 245 good candidate genes being significantly related to such traits, respectively. Gene expression pattern in the different experiments provided insights into the function of many genes which are still unknown. In the case of genes with a known functional annotation, the microarray experiments either corroborated their characterization or defined an alternative one for *Hydractinia*. This demonstrates that the *Hydractinia* platform provides a straightforward approach for functional analysis and the discovery of new genes.

These microarray results should be confirmed by Real-time PCR. Furthermore, genes' functional characterization will be improved by acquiring the complete ORF sequences and combining bioinformatics functional annotation with *in situ* hybridization experiments. Additional microarray experiments can be performed to increase and improve the acquired data. For example; the results of the immune microarray suggest that the analysis of more time points in the LPS challenge or with different PAMPs can be quite informative. Additional expression profiling analysis in colonies having different types of allogeneic responses (passive or active rejection, transitorily fusion, etc.) might help to unravel the allorecognition system of *Hydractinia*. In the case of the mitomycin array, genes that can

mark particular undifferentiated cells could be used to enrich such population and analyse it on the array.

All the sequences and functional information that are generated in cnidarians together with the ongoing genome projects in other unusual model organisms (*e.g.* sponges, chætognath or lophotrochozoans), is helping to reconstruct the genetic design of the common metazoan ancestor and provides further insight into the maintenance, loss or divergence of genes in the vertebrates [7, 8, 16, 17, 178].

5. References

1. Frank U, Leitz T, Muller WA: **The hydroid Hydractinia: a versatile, informative cnidarian representative**. *Bioessays* 2001, **23**(10):963-971.
2. Galliot B, Schmid V: **Cnidarians as a model system for understanding evolution and regeneration**. *Int J Dev Biol* 2002, **46**(1):39-48.
3. Muller WA, Leitz T: **Metamorphosis in the Cnidaria**. *Cand J Zool* 2002, **80**:1735-1754.
4. Ryan JF, Finnerty JR: **CnidBase: The Cnidarian Evolutionary Genomics Database**. *Nucleic Acids Res* 2003, **31**(1):159-163.
5. Lenhoff SG, Lenhoff HM: **Hydra and the Birth of Experimental Biology, 1744: Abraham Trembley's Memoires Concerning the Polyps**: The Boxwood press; 1986.
6. **National Human Genome Research Institute** [http://www.genome.gov/]
7. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV *et al*: **Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization**. *Science* 2007, **317**(5834):86-94.
8. Technau U, Rudd S, Maxwell P, Gordon PM, Saina M, Grasso LC, Hayward DC, Sensen CW, Saint R, Holstein TW *et al*: **Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians**. *Trends Genet* 2005, **21**(12):633-639.
9. Miller DJ, Ball EE, Technau U: **Cnidarians and ancestral genetic complexity in the animal kingdom**. *Trends Genet* 2005, **21**(10):536-539.
10. Guder C, Philipp I, Lengfeld T, Watanabe H, Hobmayer B, Holstein TW: **The Wnt code: cnidarians signal the way**. *Oncogene* 2006, **25**(57):7450-7460.
11. Adams MD, Soares MB, Kerlavage AR, Fields C, Venter JC: **Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library**. *Nat Genet* 1993, **4**(4):373-380.
12. Kortschak RD, Samuel G, Saint R, Miller DJ: **EST analysis of the cnidarian Acropora millepora reveals extensive gene loss and rapid sequence divergence in the model invertebrates**. *Curr Biol* 2003, **13**(24):2190-2195.
13. Sullivan JC, Ryan JF, Watson JA, Webb J, Mullikin JC, Rokhsar D, Finnerty JR: **StellaBase: the Nematostella vectensis Genomics Database**. *Nucleic Acids Res* 2006, **34**(Database issue):D495-499.
14. Kusserow A, Pang K, Sturm C, Hrouda M, Lentfer J, Schmidt HA, Technau U, von Haeseler A, Hobmayer B, Martindale MQ *et al*: **Unexpected complexity of the Wnt gene family in a sea anemone**. *Nature* 2005, **433**(7022):156-160.
15. Sullivan JC, Reitzel AM, Finnerty JR: **Upgrades to StellaBase facilitate medical and genetic studies on the starlet sea anemone, Nematostella vectensis**. *Nucleic Acids Res* 2008, **36**(Database issue):D607-611.
16. Miller DJ, Hemmrich G, Ball EE, Hayward DC, Khalturin K, Funayama N, Agata K, Bosch TC: **The innate immune repertoire in cnidaria--ancestral complexity and stochastic gene loss**. *Genome Biol* 2007, **8**(4):R59.
17. Miller DJ, Ball EE: **Cryptic complexity captured: the Nematostella genome reveals its secrets**. *Trends Genet* 2008, **24**(1):1-4.
18. Sullivan JC, Finnerty JR: **A surprising abundance of human disease genes in a simple "basal" animal, the starlet sea anemone (Nematostella vectensis)**. *Genome* 2007, **50**(7):689-692.

19. Rebscher N, Volk C, Teo R, Plickert G: **The germ plasm component Vasa allows tracing of the interstitial stem cells in the cnidarian Hydractinia echinata**. *Dev Dyn* 2008, **237**(6):1736-1745.
20. Cadavid LF, Powell AE, Nicotra ML, Moreno M, Buss LW: **An invertebrate histocompatibility complex**. *Genetics* 2004, **167**(1):357-365.
21. Muller WA: **Autoaggressive, multi-headed and other mutant phenotypes in Hydractinia echinata (Cnidaria: Hydrozoa)**. *Int J Dev Biol* 2002, **46**(8):1023-1033.
22. Seipp S, Schmich J, Kehrwald T, Leitz T: **Metamorphosis of Hydractinia echinata--natural versus artificial induction and developmental plasticity**. *Dev Genes Evol* 2007, **217**(5):385-394.
23. Seipp S, Schmich J, Leitz T: **Apoptosis--a death-inducing mechanism tightly linked with morphogenesis in Hydractina echinata (Cnidaria, Hydrozoa)**. *Development* 2001, **128**(23):4891-4898.
24. Leitz T, Wagner T: **the marine bacterium *Alteromonas espejiana* induces metamorphosis of the hydroid *Hydractinia echinata***. *Marine Biology* 1993, **115**:173-178.
25. Muller W, Frank U, Teo R, Mokady O, Guette C, Plickert G: **Wnt signaling in hydroid development: ectopic heads and giant buds induced by GSK-3beta inhibitors**. *Int J Dev Biol* 2007, **51**(3):211-220.
26. Bode HR: **The interstitial cell lineage of hydra: a stem cell system that arose early in evolution**. *J Cell Sci* 1996, **109 (Pt 6)**:1155-1164.
27. Muller WA, Teo R, Frank U: **Totipotent migratory stem cells in a hydroid**. *Dev Biol* 2004, **275**(1):215-224.
28. Teo R: **The Wnt Cascade and Stem Cell Fate in a Basic Metazoan**. *PhD thesis.* Heidelberg: University of Heidelberg; 2005.
29. Wagers AJ, Weissman IL: **Plasticity of adult stem cells**. *Cell* 2004, **116**(5):639-648.
30. Raff M: **Adult stem cell plasticity: fact or artifact?** *Annu Rev Cell Dev Biol* 2003, **19**:1-22.
31. Takahashi K, Yamanaka S: **Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors**. *Cell* 2006, **126**(4):663-676.
32. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R *et al*: **Induced pluripotent stem cell lines derived from human somatic cells**. *Science* 2007, **318**(5858):1917-1920.
33. Khalturin K, Anton-Erxleben F, Milde S, Plotz C, Wittlieb J, Hemmrich G, Bosch TC: **Transgenic stem cells in Hydra reveal an early evolutionary origin for key elements controlling self-renewal and differentiation**. *Dev Biol* 2007, **309**(1):32-44.
34. Miljkovic-Licina M, Chera S, Ghila L, Galliot B: **Head regeneration in wild-type hydra requires de novo neurogenesis**. *Development* 2007, **134**(6):1191-1201.
35. Teo R, Mohrlen F, Plickert G, Muller WA, Frank U: **An evolutionary conserved role of Wnt signaling in stem cell fate decision**. *Dev Biol* 2006, **289**(1):91-99.
36. Mochizuki K, Sano H, Kobayashi S, Nishimiya-Fujisawa C, Fujisawa T: **Expression and evolutionary conservation of nanos-related genes in Hydra**. *Dev Genes Evol* 2000, **210**(12):591-602.
37. Genikhovich G, Kurn U, Hemmrich G, Bosch TC: **Discovery of genes expressed in Hydra embryogenesis**. *Dev Biol* 2006, **289**(2):466-481.
38. Chamberlain SJ, Yee D, Magnuson T: **Polycomb repressive complex 2 is dispensable for maintenance of embryonic stem cell pluripotency**. *Stem Cells* 2008, **26**(6):1496-1505.

39. Hoffmann U, Kroiher M: **A possible role for the cnidarian homologue of serum response factor in decision making by undifferentiated cells**. *Dev Biol* 2001, **236**(2):304-315.
40. Hemmrich G, Bosch TC: **Compagen, a comparative genomics platform for early branching metazoan animals, reveals early origins of genes regulating stem-cell differentiation**. *Bioessays* 2008, **30**(10):1010-1018.
41. Muller WA: **Differenzierungspotenzen und Geschlechtsstabilität der I-Zellen von Hydractnia echinata**. *Roux' Archiv für Entwicklungsmechanik* 1967, **159**:412-432.
42. Murphy KM, Travers P, Walport M: **Janeway's immunobiology**; 2008.
43. Janeway CA, Jr.: **How the immune system protects the host from infection**. *Microbes Infect* 2001, **3**(13):1167-1171.
44. Pancer Z, Cooper MD: **The evolution of adaptive immunity**. *Annu Rev Immunol* 2006, **24**:497-518.
45. Janeway CA, Jr., Medzhitov R: **Innate immune recognition**. *Annu Rev Immunol* 2002, **20**:197-216.
46. Beutler B: **Innate immunity: an overview**. *Mol Immunol* 2004, **40**(12):845-859.
47. Fujita T, Matsushita M, Endo Y: **The lectin-complement pathway--its role in innate immunity and evolution**. *Immunol Rev* 2004, **198**:185-202.
48. Medzhitov R: **Toll-like receptors and innate immunity**. *Nat Rev Immunol* 2001, **1**(2):135-145.
49. Hemmrich G, Miller DJ, Bosch TC: **The evolution of immunity: a low-life perspective**. *Trends Immunol* 2007, **28**(10):449-454.
50. Dishaw LJ, Smith SL, Bigger CH: **Characterization of a C3-like cDNA in a coral: phylogenetic implications**. *Immunogenetics* 2005, **57**(7):535-548.
51. Thiel S, Vorup-Jensen T, Stover CM, Schwaeble W, Laursen SB, Poulsen K, Willis AC, Eggleton P, Hansen S, Holmskov U *et al*: **A second serine protease associated with mannan-binding lectin that activates complement**. *Nature* 1997, **386**(6624):506-510.
52. Grosberg RK, Hart MW, Levitan DR: **Is allorecognition specificity in Hydractinia symbiolongicarpus controlled by a single gene?** *Genetics* 1997, **145**(3):857-860.
53. Muller WA: **Experimentele Untersuchungen über Stockentwicklung und Sexualchimären bei Hydractinia echinata**. *Roux' Arch für Entwicklungsmechanik* 1964, **155**:181-268.
54. Powell AE, Nicotra ML, Moreno MA, Lakkis FG, Dellaporta SL, Buss LW: **Differential effect of allorecognition loci on phenotype in Hydractinia symbiolongicarpus (Cnidaria: Hydrozoa)**. *Genetics* 2007, **177**(4):2101-2107.
55. Nicotra ML, Powell AE, Rosengarten RD, Moreno M, Grimwood J, Lakkis FG, Dellaporta SL, Buss LW: **A hypervariable invertebrate allodeterminant**. *Curr Biol* 2009, **19**(7):583-589.
56. Lakkis FG, Dellaporta SL, Buss LW: **Allorecognition and chimerism in an invertebrate model organism**. *Organogenesis* 2008, **4**(4):236-240.
57. Gild S, Frank U, Mokady O: **Allogeneic interactions in Hydractinia: is the transitory chimera beneficial?** *Int J Dev Biol* 2003, **47**(6):433-438.
58. Stürzenbaum SR, Parkinson J, Blaxter, Morgan AJ, Kille P, Georgiev O: **The earthworm EST sequencing project.** . *Pedobiologia* 2003, **5**:447-451.
59. Zeng S, Gong Z: **Expressed sequence tag analysis of expression profiles of zebrafish testis and ovary**. *Gene* 2002, **294**(1-2):45-53.
60. Higgs PG, Attwood TK: **Comparative and Functional Genomics**. Oxford, UK: Blackwell; 2005.
61. Li L, Brunk BP, Kissinger JC, Pape D, Tang K, Cole RH, Martin J, Wylie T, Dante M, Fogarty SJ *et al*: **Gene discovery in the apicomplexa as revealed by EST**

sequencing and assembly of a comparative gene database. *Genome Res* 2003, **13**(3):443-454.
62. Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L, Efstratiadis A: **Construction and characterization of a normalized cDNA library**. *Proc Natl Acad Sci U S A* 1994, **91**(20):9228-9232.
63. DeRisi JL, Iyer VR: **Genomics and array technology**. *Curr Opin Oncol* 1999, **11**(1):76-79.
64. Hoheisel JD: **Microarray technology: beyond transcript profiling and genotype analysis**. *Nat Rev Genet* 2006, **7**(3):200-210.
65. Brownstein MJ, Khodursky AB: **Functional genomics; methods and protocols**: HUMANA PR; 2003.
66. Chou CC, Peck K: **Design and fabrication of spotted long oligonucleotide microarrays for gene expression analysis**. *Methods Mol Biol* 2007, **381**:213-225.
67. Bae JW, Park YH: **Homogeneous versus heterogeneous probes for microbial ecological microarrays**. *Trends Biotechnol* 2006, **24**(7):318-323.
68. Granjeaud S, Bertucci F, Jordan BR: **Expression profiling: DNA arrays in many guises**. *Bioessays* 1999, **21**(9):781-790.
69. Shalon D, Smith SJ, Brown PO: **A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization**. *Genome Res* 1996, **6**(7):639-645.
70. Smyth GK, Speed T: **Normalization of cDNA microarray data**. *Methods* 2003, **31**(4):265-273.
71. Quackenbush J: **Computational analysis of microarray data**. *Nat Rev Genet* 2001, **2**(6):418-427.
72. Fellenberg K, Hauser NC, Brors B, Hoheisel JD, Vingron M: **Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis**. *Bioinformatics* 2002, **18**(3):423-433.
73. Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, Vingron M: **Correspondence analysis applied to microarray data**. *Proc Natl Acad Sci U S A* 2001, **98**(19):10781-10786.
74. Fellenberg K: **Storage and analysis of microarray data**. university of Cologne; 2002.
75. Koschmieder A: **Klonbibliotheken mit randomisierten artifiziellen DNA-Inserts in Microarray-Anwendungen**. University of Applied Sciences Jena; 2005.
76. **Genome Sequencing Center** [http://genome.wustl.edu/]
77. **National Center for Biotechnology Information** [http://www.ncbi.nlm.nih.gov]
78. **Heidelberg Unix Sequence Analysis Resource (HUSAR)** [http://husar/menu/biounit/]
79. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.
80. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
81. Etzold T, Ulyanov A, Argos P: **SRS: information retrieval system for molecular biology data banks**. *Methods Enzymol* 1996, **266**:114-128.
82. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I *et al*: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**. *Nucleic Acids Res* 2003, **31**(1):365-370.

83. del Val C, Ernst P, Falkenhahn M, Fladerer C, Glatting KH, Suhai S, Hotz-Wagenblatt A: **ProtSweep, 2Dsweep and DomainSweep: protein analysis suite at DKFZ**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W444-450.
84. Vinayagam A, del Val C, Schubert F, Eils R, Glatting KH, Suhai S, Konig R: **GOPET: a tool for automated predictions of Gene Ontology terms**. *BMC Bioinformatics* 2006, **7**:161.
85. **Ensembl Genome Browser** [http://www.ensembl.org]
86. **PostgreSQL** [http://www.postgresql.org]
87. Beissbarth T, Fellenberg K, Brors B, Arribas-Prat R, Boer J, Hauser NC, Scheideler M, Hoheisel JD, Schutz G, Poustka A *et al*: **Processing and quality control of DNA array hybridization data**. *Bioinformatics* 2000, **16**(11):1014-1022.
88. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proc Natl Acad Sci U S A* 2001, **98**(9):5116-5121.
89. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M *et al*: **TM4: a free, open-source system for microarray data management and analysis**. *Biotechniques* 2003, **34**(2):374-378.
90. Soza-Ried J: **Construction and first analyses of a *Hydractinia echinata* (Fleming, 1928) cDNA library**. Mannheim: University of applied Sciences Mannheim; 2004.
91. Xu HX, Kawamura Y, Li N, Zhao L, Li TM, Li ZY, Shu S, Ezaki T: **A rapid method for determining the G+C content of bacterial chromosomes by monitoring fluorescence intensity during DNA denaturation in a capillary tube**. *Int J Syst Evol Microbiol* 2000, **50 Pt 4**:1463-1469.
92. Wang HC, Badger J, Kearney P, Li M: **Analysis of codon usage patterns of bacterial genomes using the self-organizing map**. *Mol Biol Evol* 2001, **18**(5):792-800.
93. Belle EM, Duret L, Galtier N, Eyre-Walker A: **The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny**. *J Mol Evol* 2004, **58**(6):653-660.
94. Lobry JR, Sueoka N: **Asymmetric directional mutation pressures in bacteria**. *Genome Biol* 2002, **3**(10):RESEARCH0058.
95. Mochizuki K, Nishimiya-Fujisawa C, Fujisawa T: **Universal occurrence of the vasa-related genes among metazoans and their germline expression in Hydra**. *Dev Genes Evol* 2001, **211**(6):299-308.
96. **Genecards** [http://www.genecards.org/]
97. Carninci P, Shibata Y, Hayatsu N, Sugahara Y, Shibata K, Itoh M, Konno H, Okazaki Y, Muramatsu M, Hayashizaki Y: **Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes**. *Genome Res* 2000, **10**(10):1617-1630.
98. Forment J, Gadea J, Huerta L, Abizanda L, Agusti J, Alamar S, Alos E, Andres F, Arribas R, Beltran JP *et al*: **Development of a citrus genome-wide EST collection and cDNA microarray as resources for genomic studies**. *Plant Mol Biol* 2005, **57**(3):375-391.
99. Jain M, Shrager J, Harris EH, Halbrook R, Grossman AR, Hauser C, Vallon O: **EST assembly supported by a draft genome sequence: an analysis of the Chlamydomonas reinhardtii transcriptome**. *Nucleic Acids Res* 2007, **35**(6):2074-2083.
100. Vinayagam A, Konig R, Moormann J, Schubert F, Eils R, Glatting KH, Suhai S: **Applying Support Vector Machines for Gene Ontology based gene function prediction**. *BMC Bioinformatics* 2004, **5**:116.
101. Whitfield CW, Band MR, Bonaldo MF, Kumar CG, Liu L, Pardinas JR, Robertson HM, Soares MB, Robinson GE: **Annotated expressed sequence tags and cDNA**

microarrays for studies of brain and behavior in the honey bee. *Genome Res* 2002, **12**(4):555-566.

102. Steele RE, Hampson SE, Stover NA, Kibler DF, Bode HR: **Probable horizontal transfer of a gene between a protist and a cnidarian**. *Curr Biol* 2004, **14**(8):R298-299.
103. Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, Brown JR: **Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates**. *Nature* 2001, **411**(6840):940-944.
104. Rosenberg E, Koren O, Reshef L, Efrony R, Zilber-Rosenberg I: **The role of microorganisms in coral health, disease and evolution**. *Nat Rev Microbiol* 2007, **5**(5):355-362.
105. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, Boffelli D, Anderson IJ, Barry KW, Shapiro HJ *et al*: **Symbiosis insights through metagenomic analysis of a microbial consortium**. *Nature* 2006, **443**(7114):950-955.
106. Kuo J, Chen MC, Lin CH, Fang LS: **Comparative gene expression in the symbiotic and aposymbiotic Aiptasia pulchella by expressed sequence tag analysis**. *Biochem Biophys Res Commun* 2004, **318**(1):176-186.
107. Merle PL, Sabourault C, Richier S, Allemand D, Furla P: **Catalase characterization and implication in bleaching of a symbiotic sea anemone**. *Free Radic Biol Med* 2007, **42**(2):236-246.
108. Orr-Weaver TL: **Drosophila chorion genes: cracking the eggshell's secrets**. *Bioessays* 1991, **13**(3):97-105.
109. Rodrigues V, Chaudhri M, Knight M, Meadows H, Chambers AE, Taylor WR, Kelly C, Simpson AJ: **Predicted structure of a major Schistosoma mansoni eggshell protein**. *Mol Biochem Parasitol* 1989, **32**(1):7-13.
110. Sasaguri K, Ganss B, Sodek J, Chen JK: **Expression of bone sialoprotein in mineralized tissues of tooth and bone and in buccal-pouch carcinomas of Syrian golden hamsters**. *Arch Oral Biol* 2000, **45**(7):551-562.
111. Branchek TA, Smith KE, Gerald C, Walker MW: **Galanin receptor subtypes**. *Trends Pharmacol Sci* 2000, **21**(3):109-117.
112. Liu F, Thatcher JD, Barral JM, Epstein HF: **Bifunctional glyoxylate cycle protein of Caenorhabditis elegans: a developmentally regulated protein of intestine and muscle**. *Dev Biol* 1995, **169**(2):399-414.
113. Koonin EV: **Orthologs, paralogs, and evolutionary genomics**. *Annu Rev Genet* 2005, **39**:309-338.
114. Mali B, Soza-Ried J, Frohme M, Frank U: **Structural but not functional conservation of an immune molecule: a tachylectin-like gene in Hydractinia**. *Dev Comp Immunol* 2006, **30**(3):275-281.
115. Schroder HC, Ushijima H, Krasko A, Gamulin V, Thakur NL, Diehl-Seifert B, Muller IM, Muller WE: **Emergence and disappearance of an immune molecule, an antimicrobial lectin, in basal metazoa. A tachylectin-related protein in the sponge Suberites domuncula**. *J Biol Chem* 2003, **278**(35):32810-32817.
116. Beisel HG, Kawabata S, Iwanaga S, Huber R, Bode W: **Tachylectin-2: crystal structure of a specific GlcNAc/GalNAc-binding lectin involved in the innate immunity host defense of the Japanese horseshoe crab Tachypleus tridentatus**. *EMBO J* 1999, **18**(9):2313-2322.
117. Galliano M, Minchiotti L, Campagnoli M, Sala A, Visai L, Amoresano A, Pucci P, Casbarra A, Cauci M, Perduca M *et al*: **Structural and biochemical characterization of a new type of lectin isolated from carp eggs**. *Biochem J* 2003, **376**(Pt 2):433-440.
118. Wren JD, Kulkarni A, Joslin J, Butow RA, Garner HR: **Cross-hybridization on PCR-spotted microarrays**. *IEEE Eng Med Biol Mag* 2002, **21**(2):71-75.

119. Musso M, Bocciardi R, Parodi S, Ravazzolo R, Ceccherini I: **Betaine, dimethyl sulfoxide, and 7-deaza-dGTP, a powerful mixture for amplification of GC-rich DNA sequences**. *J Mol Diagn* 2006, **8**(5):544-550.
120. Do JH, Choi DK: **Normalization of microarray data: single-labeled and dual-labeled arrays**. *Mol Cells* 2006, **22**(3):254-261.
121. Gupta V, Cherkassky A, Chatis P, Joseph R, Johnson AL, Broadbent J, Erickson T, DiMeo J: **Directly labeled mRNA produces highly precise and unbiased differential gene expression data**. *Nucleic Acids Res* 2003, **31**(4):e13.
122. Mujumdar RB, Ernst LA, Mujumdar SR, Lewis CJ, Waggoner AS: **Cyanine dye labeling reagents: sulfoindocyanine succinimidyl esters**. *Bioconjug Chem* 1993, **4**(2):105-111.
123. Dodd LE, Korn EL, McShane LM, Chandramouli GV, Chuang EY: **Correcting log ratios for signal saturation in cDNA microarrays**. *Bioinformatics* 2004, **20**(16):2685-2693.
124. Dudoit S, Fridlyand J: **A prediction-based resampling method for estimating the number of clusters in a dataset**. *Genome Biol* 2002, **3**(7):RESEARCH0036.
125. Dudoit S, Fridlyand J: **Bagging to improve the accuracy of a clustering procedure**. *Bioinformatics* 2003, **19**(9):1090-1099.
126. Gibbons FD, Roth FP: **Judging the quality of gene expression-based clustering methods using gene annotation**. *Genome Res* 2002, **12**(10):1574-1581.
127. Jamieson D, Tung AT, Knox RJ, Boddy AV: **Reduction of mitomycin C is catalysed by human recombinant NRH:quinone oxidoreductase 2 using reduced nicotinamide adenine dinucleotide as an electron donating co-factor**. *Br J Cancer* 2006, **95**(9):1229-1233.
128. Rao AV, Shaha C: **Role of glutathione S-transferases in oxidative stress-induced male germ cell apoptosis**. *Free Radic Biol Med* 2000, **29**(10):1015-1027.
129. Dehari H, Tchaikovskaya T, Rubashevsky E, Sellers R, Listowsky I: **The proximal promoter governs germ cell-specific expression of the mouse glutathione transferase mGstm5 gene**. *Mol Reprod Dev* 2009, **76**(4):379-388.
130. Djadid ND, Barjesteh H, Raeisi A, Hassanzahi A, Zakeri S: **Identification, sequence analysis, and comparative study on GSTe2 insecticide resistance gene in three main world malaria vectors: Anopheles stephensi, Anopheles culicifacies, and Anopheles fluviatilis**. *J Med Entomol* 2006, **43**(6):1171-1177.
131. Kim YJ, Zuo P, Manley JL, Baker BS: **The Drosophila RNA-binding protein RBP1 is localized to transcriptionally active sites of chromosomes and shows a functional similarity to human splicing factor ASF/SF2**. *Genes Dev* 1992, **6**(12B):2569-2579.
132. Chen D, Zhao M, Mundy GR: **Bone morphogenetic proteins**. *Growth Factors* 2004, **22**(4):233-241.
133. Bowers RR, Kim JW, Otto TC, Lane MD: **Stable stem cell commitment to the adipocyte lineage by inhibition of DNA methylation: role of the BMP-4 gene**. *Proc Natl Acad Sci U S A* 2006, **103**(35):13022-13027.
134. Soza-Ried C, Bleul CC, Schorpp M, Boehm T: **Maintenance of thymic epithelial phenotype requires extrinsic signals in mouse and zebrafish**. *J Immunol* 2008, **181**(8):5272-5277.
135. Reber-Muller S, Streitwolf-Engel R, Yanze N, Schmid V, Stierwald M, Erb M, Seipel K: **BMP2/4 and BMP5-8 in jellyfish development and transdifferentiation**. *Int J Dev Biol* 2006, **50**(4):377-384.
136. Botzler C, Oertel M, Hinz M, Hoffmann W: **Structure of the Xenopus laevis TFF-gene xP4.1, differentially expressed to its duplicated homolog xP4.2**. *Biochim Biophys Acta* 1999, **1489**(2-3):345-353.

137. Poulsom R, Begos DE, Modlin IM: **Molecular aspects of restitution: functions of trefoil peptides**. *Yale J Biol Med* 1996, **69**(2):137-146.
138. Schwarz RS, Hodes-Villamar L, Fitzpatrick KA, Fain MG, Hughes AL, Cadavid LF: **A gene family of putative immune recognition molecules in the hydroid Hydractinia**. *Immunogenetics* 2007, **59**(3):233-246.
139. Mohrlen F, Maniura M, Plickert G, Frohme M, Frank U: **Evolution of astacin-like metalloproteases in animals and their function in development**. *Evol Dev* 2006, **8**(2):223-231.
140. Kelly KF, Daniel JM: **POZ for effect--POZ-ZF transcription factors in cancer and development**. *Trends Cell Biol* 2006, **16**(11):578-587.
141. Yet SF, McA'Nulty MM, Folta SC, Yen HW, Yoshizumi M, Hsieh CM, Layne MD, Chin MT, Wang H, Perrella MA *et al*: **Human EZF, a Kruppel-like zinc finger protein, is expressed in vascular endothelial cells and contains transcriptional activation and repression domains**. *J Biol Chem* 1998, **273**(2):1026-1031.
142. Seipel K, Yanze N, Muller P, Streitwolf R, Schmid V: **Basic leucine zipper transcription factors C/EBP and MafL in the hydrozoan jellyfish Podocoryne carnea**. *Dev Dyn* 2004, **230**(3):392-402.
143. David CN, Ozbek S, Adamczyk P, Meier S, Pauly B, Chapman J, Hwang JS, Gojobori T, Holstein TW: **Evolution of complex structures: minicollagens shape the cnidarian nematocyst**. *Trends Genet* 2008, **24**(9):431-438.
144. Kurz EM, Holstein TW, Petri BM, Engel J, David CN: **Mini-collagens in hydra nematocytes**. *J Cell Biol* 1991, **115**(4):1159-1169.
145. Denker E, Manuel M, Leclere L, Le Guyader H, Rabet N: **Ordered progression of nematogenesis from stem cells through differentiation stages in the tentacle bulb of Clytia hemisphaerica (Hydrozoa, Cnidaria)**. *Dev Biol* 2008, **315**(1):99-113.
146. Kreft SG, Nassal M: **hRUL138, a novel human RNA-binding RING-H2 ubiquitin-protein ligase**. *J Cell Sci* 2003, **116**(Pt 4):605-616.
147. Stolow DT, Haynes SR: **Cabeza, a Drosophila gene encoding a novel RNA binding protein, shares homology with EWS and TLS, two genes involved in human sarcoma formation**. *Nucleic Acids Res* 1995, **23**(5):835-843.
148. Oren M, Douek J, Fishelson Z, Rinkevich B: **Identification of immune-relevant genes in histoincompatible rejecting colonies of the tunicate Botryllus schlosseri**. *Dev Comp Immunol* 2007, **31**(9):889-902.
149. Ben-Shlomo R: **The molecular basis of allorecognition in ascidians**. *Bioessays* 2008, **30**(11-12):1048-1051.
150. Sher D, Knebel A, Bsor T, Nesher N, Tal T, Morgenstern D, Cohen E, Fishman Y, Zlotkin E: **Toxic polypeptides of the hydra--a bioinformatic approach to cnidarian allomones**. *Toxicon* 2005, **45**(7):865-879.
151. Sher D, Fishman Y, Zhang M, Lebendiker M, Gaathon A, Mancheno JM, Zlotkin E: **Hydralysins, a new category of beta-pore-forming toxins in cnidaria**. *J Biol Chem* 2005, **280**(24):22847-22855.
152. Grasso LC, Maindonald J, Rudd S, Hayward DC, Saint R, Miller DJ, Ball EE: **Microarray analysis identifies candidate genes for key roles in coral development**. *BMC Genomics* 2008, **9**:540.
153. Burke JE, Dennis EA: **Phospholipase A2 structure/function, mechanism and signaling**. *J Lipid Res* 2008.
154. Murakami M, Kudo I: **Phospholipase A2**. *J Biochem* 2002, **131**(3):285-292.
155. McIntosh JM, Ghomashchi F, Gelb MH, Dooley DJ, Stoehr SJ, Giordani AB, Naisbitt SR, Olivera BM: **Conodipine-M, a novel phospholipase A2 isolated from the venom of the marine snail Conus magus**. *J Biol Chem* 1995, **270**(8):3518-3526.

156. Lotan A, Fishman L, Zlotkin E: **Toxin compartmentation and delivery in the Cnidaria: the nematocyst's tubule as a multiheaded poisonous arrow**. *J Exp Zool* 1996, **275**(6):444-451.
157. Yang H, Xia Y, Lu SQ, Soong TW, Feng ZW: **Basic fibroblast growth factor-induced neuronal differentiation of mouse bone marrow stromal cells requires FGFR-1, MAPK/ERK, and transcription factor AP-1**. *J Biol Chem* 2008, **283**(9):5287-5295.
158. Leacock SW, Reinke V: **Expression profiling of MAP kinase-mediated meiotic progression in Caenorhabditis elegans**. *PLoS Genet* 2006, **2**(11):e174.
159. Lehane AM, Marchetti RV, Spry C, van Schalkwyk DA, Teng R, Kirk K, Saliba KJ: **Feedback inhibition of pantothenate kinase regulates pantothenol uptake by the malaria parasite**. *J Biol Chem* 2007, **282**(35):25395-25405.
160. O'Neill L: **The Toll/interleukin-1 receptor domain: a molecular switch for inflammation and host defence**. *Biochem Soc Trans* 2000, **28**(5):557-563.
161. Kurata S, Ariki S, Kawabata S: **Recognition of pathogens and activation of immune responses in Drosophila and horseshoe crab innate immunity**. *Immunobiology* 2006, **211**(4):237-249.
162. Bosch TC, Augustin R, Anton-Erxleben F, Fraune S, Hemmrich G, Zill H, Rosenstiel P, Jacobs G, Schreiber S, Leippe M *et al*: **Uncovering the evolutionary history of innate immunity: the simple metazoan Hydra uses epithelial cells for host defence**. *Dev Comp Immunol* 2009, **33**(4):559-569.
163. Wiens M, Korzhev M, Krasko A, Thakur NL, Perovic-Ottstadt S, Breter HJ, Ushijima H, Diehl-Seifert B, Muller IM, Muller WE: **Innate immune defense of the sponge Suberites domuncula against bacteria involves a MyD88-dependent signaling pathway. Induction of a perforin-like molecule**. *J Biol Chem* 2005, **280**(30):27949-27959.
164. Brown DT: **Histone variants: are they functionally heterogeneous?** *Genome Biol* 2001, **2**(7):REVIEWS0006.
165. Bhan S, May W, Warren SL, Sittman DB: **Global gene expression analysis reveals specific and redundant roles for H1 variants, H1c and H1(0), in gene expression regulation**. *Gene* 2008, **414**(1-2):10-18.
166. Ramakrishnan V, Finch JT, Graziano V, Lee PL, Sweet RM: **Crystal structure of globular domain of histone H5 and its implications for nucleosome binding**. *Nature* 1993, **362**(6417):219-223.
167. Scheurlen I, Hoffmeister SA, Schaller HC: **Presence and expression of G2 cyclins in the coelenterate hydra**. *J Cell Sci* 1996, **109 (Pt 5)**:1063-1069.
168. Kramer S, Queiroz R, Ellis L, Webb H, Hoheisel JD, Clayton C, Carrington M: **Heat shock causes a decrease in polysomes and the appearance of stress granules in trypanosomes independently of eIF2(alpha) phosphorylation at Thr169**. *J Cell Sci* 2008, **121**(Pt 18):3002-3014.
169. Tsan MF, Gao B: **Heat shock protein and innate immunity**. *Cell Mol Immunol* 2004, **1**(4):274-279.
170. Lunin SM, Khrenov MO, Novoselova TV, Parfenyuk SB, Novoselova EG: **Thymulin, a thymic peptide, prevents the overproduction of pro-inflammatory cytokines and heat shock protein Hsp70 in inflammation-bearing mice**. *Immunol Invest* 2008, **37**(8):858-870.
171. Berger T, Togawa A, Duncan GS, Elia AJ, You-Ten A, Wakeham A, Fong HE, Cheung CC, Mak TW: **Lipocalin 2-deficient mice exhibit increased sensitivity to Escherichia coli infection but not to ischemia-reperfusion injury**. *Proc Natl Acad Sci U S A* 2006, **103**(6):1834-1839.

172. Saiga H, Nishimura J, Kuwata H, Okuyama M, Matsumoto S, Sato S, Matsumoto M, Akira S, Yoshikai Y, Honda K *et al*: **Lipocalin 2-dependent inhibition of mycobacterial growth in alveolar epithelium**. *J Immunol* 2008, **181**(12):8521-8527.
173. Flower DR: **The lipocalin protein family: structure and function**. *Biochem J* 1996, **318 (Pt 1)**:1-14.
174. Flo TH, Smith KD, Sato S, Rodriguez DJ, Holmes MA, Strong RK, Akira S, Aderem A: **Lipocalin 2 mediates an innate immune response to bacterial infection by sequestrating iron**. *Nature* 2004, **432**(7019):917-921.
175. Holland DB, Bojar RA, Farrar MD, Holland KT: **Differential innate immune responses of a living skin equivalent model colonized by Staphylococcus epidermidis or Staphylococcus aureus**. *FEMS Microbiol Lett* 2009, **290**(2):149-155.
176. Holstein TW, Mala C, Kurz E, Bauer K, Greber M, David CN: **The primitive metazoan Hydra expresses antistasin, a serine protease inhibitor of vertebrate blood coagulation: cDNA cloning, cellular localisation and developmental regulation**. *FEBS Lett* 1992, **309**(3):288-292.
177. Joo SS, Won TJ, Kim JS, Yoo YM, Tak ES, Park SY, Park HY, Hwang KW, Park SC, Lee do I: **Inhibition of coagulation activation and inflammation by a novel Factor Xa inhibitor synthesized from the earthworm Eisenia andrei**. *Biol Pharm Bull* 2009, **32**(2):253-258.
178. Marletaz F, Gilles A, Caubit X, Perez Y, Dossat C, Samain S, Gyapay G, Wincker P, Le Parco Y: **Chaetognath transcriptome reveals ancestral and unique features among bilaterians**. *Genome Biol* 2008, **9**(6):R94.

6. Appendix

6.1 Additional data 1

Supplementary data related to the sequence analysis pipeline.

Table S1. Annotation of *Hydractinia* un-informative and unknown sequences using Domainsweep

Clone name	Inter Pro annotation number	Sequence annotation
HEAB-0018B03	IPR007116	6-pyruvoyl tetrahydropterin synthase
HEAB-0020B05	IPR001680/IPR007190	WD-40 repeat/Periodic tryptophan protein-associated region
HEAB-0020C15	IPR000719	Protein kinase
HEAB-0020D15	IPR001442	Type 4 procollagen, C-terminal repeat
HEAB-0020L12	IPR002086	Aldehyde dehydrogenase
HEAB-0021B13	IPR003660	Histidine kinase, HAMP region
HEAB-0021B17	IPR001202	WW/Rsp5/WWP
HEAB-0023G03	IPR001283	Allergen V5/Tpx-1 related
HEAB-0024F14	IPR002323	Cytochrome c, class IE
HEAB-0024L07	IPR000905	Peptidase M22, glycoprotease
HEAB-0024L20	IPR001497	Methylated-DNA-[protein]-cysteine S-methyltransferase, active site
HEAB-0025A24	IPR007087	Zinc finger, C2H2-type
HEAB-0025F04	IPR005028	Herpes virus intermediate/early protein 2/3
HEAB-0026I15	IPR010528	TolA
HEAB-0027F04	IPR005116	TOBE
HEAB-0027F07	IPR001754	Orotidine 5'-phosphate decarboxylase, core
HEAB-0027I05	IPR001564	Nucleoside diphosphate kinase
HEAB-0027J22	IPR004012	RUN
HEAB-0027M01	IPR008412	Bone sialoprotein II
HEAB-0028B05	IPR003780	Cytochrome oxidase assembly
HEAB-0028B06	IPR002364/IPR011597	Quinone oxidoreductase/zeta-crystallin/GroES-related
HEAB-0028B17	IPR003236	Mitochondrial ribosomal protein L5
HEAB-0028D13	IPR001650/IPR012541	Helicase, C-terminal/DBP10CT
HEAB-0028D14	IPR001506	Peptidase M12A, Astacin
HEAB-0028E20	IPR000892	Ribosomal protein S26E
HEAB-0028F19	IPR000695/IPR004714	H+ transporting ATPase, proton pump/Cytochrome oxidase maturation protein cbb3-type
HEAB-0028G18	IPR002123	Phospholipid/glycerol acyltransferase
HEAB-0028H01	IPR001611	Leucine-rich repeat
HEAB-0028J07	IPR005119	LysR, substrate-binding
HEAB-0028K15	IPR002155	Thiolase
HEAB-0028L21	IPR003122/IPR004089/IPR004090	Ligand binding Tar/Bacterial chemotaxis sensory transducer/Chemotaxis methyl-accepting protein/Histidine kinase
HEAB-0028N06	IPR006108/IPR006176	3-hydroxyacyl-CoA dehydrogenase, C-terminal/3-hydroxyacyl-CoA dehydrogenase, NAD-binding
HEAB-0028N08	IPR004089/IPR004090	Bacterial chemotaxis sensory transducer/Chemotaxis methyl-accepting protein/Histidine kinase
HEAB-0028O11	IPR001638/IPR003439/IPR005074	Bacterial extracellular solute-binding protein, family 3/ABC transporter related/Peptidase C39, bacteriocin processing
HEAB-0028O22	IPR003439/IPR005116/IPR008779	ABC transporter related/TOBE/Plasmodium histidine-rich
HEAB-0029C14	IPR006674	Metal-dependent phosphohydrolase, HD region, subdomain
HEAB-0029C20	IPR005829	Sugar transporter superfamily
HEAB-0029E04	IPR001303	Class II aldolase/adducin, N-terminal
HEAB-0029G07	IPR001190	Speract/scavenger receptor
HEAB-0029H20	IPR003042	Aromatic-ring hydroxylase
HEAB-0029I17	IPR011603	2-oxoglutarate dehydrogenase, E1 component
HEAB-0029K24	IPR001404	Heat shock protein Hsp90
HEAB-0029L13	IPR007838	Protein of unknown function DUF710
HEAB-0029L19	IPR000281/IPR001347	Helix-turn-helix protein RpiR/Sugar isomerase (SIS)

Clone name	Inter Pro annotation number	Sequence annotation
HEAB-0029N13	IPR000289	Ribosomal protein S28e
HEAB-0029O15	IPR002823	Protein of unknown function DUF112, transmembrane
HEAB-0030C13	IPR000524	Bacterial regulatory protein GntR, HTH
HEAB-0030I07	IPR001123	Lysine exporter protein (LYSE/YGGA)
HEAB-0030M14	IPR002792/IPR013848	Deoxyribonuclease/rho motif-related TRAM/Protein of unknown function UPF0004, N-terminal
HEAB-0030N07	IPR002652	Importin-alpha-like, importin-beta-binding region
HEAB-0031B09	IPR000515/IPR001638	Binding-protein-dependent transport systems inner membrane component/Bacterial extracellular solute-binding protein, family 3
HEAB-0031C21	IPR003768	Prokaryotic chromosome segregation and condensation protein ScpA
HEAB-0031D20	IPR003439	ABC transporter related
HEAB-0031F21	IPR000160	GGDEF
HEAB-0031F22	IPR001176/IPR004113/IPR004838 /IPR006094	aminocyclopropane-1-carboxylate synthase/FAD linked oxidase, C-terminal/Aminotransferases class-I pyridoxal-phosphate-binding site
HEAB-0031I13	IPR006143	Secretion protein HlyD
HEAB-0031N18	IPR000481	Pheromone B alpha-1 receptor
HEAB-0031N22	IPR007087	Zinc finger, C2H2-type
HEAB-0031P18	IPR003122/IPR003660/IPR004010 /IPR004089/IPR004090	Ligand binding Tar/Histidine kinase, HAMP region/Cache/Bacterial chemotaxis sensory transducer /Chemotaxis methyl-accepting protein
HEAB-0032P06	IPR000988	Ribosomal protein L24E
HEAB-0033C06	IPR001387	Helix-turn-helix type 3
HEAB-0033C24	IPR001452	Src homology-3
HEAB-0033D24	IPR011712	Histidine kinase, dimerisation and phosphoacceptor region
HEAB-0033E23	IPR003115	ParB-like nuclease
HEAB-0033F09	IPR002371	Flagellar hook-associated protein
HEAB-0033G05	IPR003397	Mitochondrial import inner membrane translocase, subunit Tim17/22
HEAB-0033I16	IPR000449	Ubiquitin-associated/Translation elongation factor EF1B, N-terminal
HEAB-0033I17	IPR003778	Allophanate hydrolase subunit 2
HEAB-0033K02	IPR000884	Thrombospondin, type I
HEAB-0033M09	IPR012987	ROK, N-terminal
HEAB-0033O24	IPR002347	Glucose/ribitol dehydrogenase
HEAB-0034A23	IPR000873	AMP-dependent synthetase and ligase
HEAB-0034B11	IPR002925	Dienelactone hydrolase
HEAB-0034B16	IPR003920	Cellulose synthase, subunit B
HEAB-0034B24	IPR001611	Leucine-rich repeat
HEAB-0034E15	IPR002504	ATP-NAD/AcoX kinase
HEAB-0034E16	IPR001442/IPR002541	Type 4 procollagen, C-terminal repeat/Cytochrome c assembly protein
HEAB-0034I24	IPR000015	Fimbrial biogenesis outer membrane usher protein
HEAB-0034J18	IPR000754	Ribosomal protein S9
HEAB-0034K20	IPR004827	Basic-leucine zipper (bZIP) transcription factor
HEAB-0034N17	IPR002952	Eggshell protein
HEAB-0034O22	IPR001188	Bacterial periplasmic spermidine/putrescine-binding protein
HEAB-0035D08	IPR007087	Zinc finger, C2H2-type
HEAB-0035E21	IPR000595/IPR001808	Cyclic nucleotide-binding/Bacterial regulatory protein, Crp
HEAB-0035G21	IPR000951/IPR001221/IPR001709/ IPR001834/IPR008333	Phthalate dioxygenase reductase, FPNCR module Phenol hydroxylase reductase /Flavoprotein pyridine nucleotide cytochrome reductase
HEAB-0035I13	IPR001638/IPR003439	Bacterial extracellular solute-binding protein, family 3/ABC transporter related
HEAB-0035J12	IPR001545	Gonadotropin, beta chain
HEAB-0035L05	IPR000759/IPR000960	Adrenodoxin reductase/Flavin-containing monooxygenase FMO
HEAB-0035L08	IPR000592	Ribosomal protein S27E
HEAB-0035N14	IPR003219	Cytochrome c, alcohol dehydrogenase-like subunit
HEAB-0036G06	IPR004358	Histidine kinase related protein, C-terminal
HEAB-0036G16	IPR005064	Bordetella uptake gene
HEAB-0036J11	IPR001876	Zinc finger, RanBP2-type
HEAB-0036J18	IPR005835	Nucleotidyl transferase
HEAB-0036N05	IPR003711/IPR005118	Transcription factor CarD/TRCF
HEAB-0037A22	IPR007476	Putative exonuclease, RdgC
HEAB-0037F23	IPR001789	Response regulator receiver

Clone name	Inter Pro annotation number	Sequence annotation
HEAB-0037I16	IPR000160	GGDEF
HEAB-0037J03	IPR007087	Zinc finger, C2H2-type
HEAB-0037L12	IPR001419	HMW glutenin
HEAB-0038C07	IPR000515	Binding-protein-dependent transport systems inner membrane component
HEAB-0038C12	IPR005565	Hemolysin activator HlyB
HEAB-0038D19	IPR005649	Chorion 2
HEAB-0038G01	IPR001387	Helix-turn-helix type 3
HEAB-0038G06	IPR000196	Ribosomal protein L19e
HEAB-0038H11	IPR008168	Cytochrome c, class IC
HEAB-0038H17	IPR006706/IPR007223/IPR010800	Extensin-like region/Peroxin 13, N-terminal/Glycine rich
HEAB-0038I20	IPR000719	Protein kinase
HEAB-0038O03	IPR001283/IPR002413	Allergen V5/Tpx-1 related/Ves allergen
HEAB-0039H08	IPR000504	RNA-binding region RNP-1 (RNA recognition motif)
HEAB-0039H23	IPR005649	Chorion 2
HEAB-0039K09	IPR001673/IPR001881/IPR002212	Dictyostelium (slime mold) repeat/EGF-like calcium-binding/Matrix fibril-associated
HEAB-0039L13	IPR006686	MscS Mechanosensitive ion channel, middle
HEAB-0039N14	IPR001005	SANT, DNA-binding
HEAB-0039P05	IPR002198	Short-chain dehydrogenase/reductase SDR
HEAB-0040C22	IPR001680	WD-40 repeat
HEAB-0040H07	IPR001023	Heat shock protein Hsp70
HEAB-0040I24	IPR008638	Filamentous haemagglutinin, N-terminal, bacterial
HEAB-0040K05	IPR000914	Bacterial extracellular solute-binding protein, family 5
HEAB-0040L16	IPR004358	Histidine kinase related protein, C-terminal
HEAB-0040L19	IPR000637	HMG-I and HMG-Y, DNA-binding
HEAB-0040M05	IPR003908	Galanin 3 receptor
HEAB-0040N02	IPR001635	Flagellar hook-length control protein
HEAB-0041A23	IPR003439/IPR005116	ABC transporter related/TOBE
HEAB-0041C11	IPR001442	Type 4 procollagen, C-terminal repeat
HEAB-0041I12	IPR003172	MD-2-related lipid-recognition
HEAB-0041N22	IPR002078/IPR002197	RNA polymerase sigma factor 54, interaction/Helix-turn-helix, Fis-type
HEAB-0041O11	IPR002078	RNA polymerase sigma factor 54, interaction
HEAB-0042A02	IPR000037	SmpB protein
HEAB-0042C06	IPR001789	Response regulator receiver
HEAB-0042E11	IPR000708	Prostanoid EP1 receptor
HEAB-0042I20	IPR000172/IPR007867	Glucose-methanol-choline oxidoreductase, N-terminal/Glucose-methanol-choline oxidoreductase, C-terminal
HEAB-0042J08	IPR011704	ATPase associated with various cellular activities, AAA-5
HEAB-0042L09	IPR003265	HhH-GPD
HEAB-0042L21	IPR006650	Adenosine/AMP deaminase active site
HEAB-0042M23	IPR001841	Zinc finger, RING-type
HEAB-0042N17	IPR007630	RNA polymerase sigma-70 region 4
HEAB-0042P17	IPR003453	Protein of unknown function DUF140
HEAB-0042P22	IPR001912/IPR002942	Ribosomal protein S4/RNA-binding S4
tah96a10	IPR006706/IPR007223	Extensin-like region/Peroxin 13, N-terminal
tah96c11	IPR007087	Zinc finger, C2H2-type
tah96d03	IPR000194/IPR000790	ATPase, F1/V1/A1 complex, alpha/beta subunit, nucleotide-binding/ATPase, F1 complex, alpha subunit, C-terminal
tah96e04	IPR000623	Shikimate kinase
tah97c04	IPR001506	Peptidase M12A, Astacin
tah97g12	IPR000772	Ricin B lectin
tah98a10	IPR000008	C2 calcium-dependent membrane targeting
tah98a11	IPR004045	Glutathione S-transferase, N-terminal
tah98b04	IPR000301	CD9/CD37/CD63 antigen
tah98c10	IPR001680	WD-40 repeat
tah98d10	IPR004000	Actin/actin-like

Clone name	Inter Pro annotation number	Sequence annotation
tah98e04	IPR002952/IPR007223	Eggshell protein/Peroxin 13, N-terminal
tah99a03	IPR007087	Zinc finger, C2H2-type
tah99b02	IPR001650	Helicase, C-terminal
tah99d04	IPR000297/IPR001202/IPR002349	Peptidyl-prolyl cis-trans isomerase, PpiC-type/WW/Rsp5/WWP/WW
tah99f03	IPR005448	P/Q-type voltage-dependent calcium channel alpha 1 subunit
tah99h06	IPR002478	PUA
tah99h09	IPR001638	Bacterial extracellular solute-binding protein, family 3
tai01a04	IPR002007	Haem peroxidase, animal
tai01g09	IPR006706	Extensin-like region
tai02a07	IPR001305	DnaJ central region
tai02c10	IPR000626	Ubiquitin
tai02h08	IPR001662/IPR004045	Translation elongation factor EF1B, gamma chain, conserved/Glutathione S-transferase, N-terminal
tai03g11	IPR001993/IPR002067/IPR002113	Mitochondrial substrate carrier/Mitochondrial carrier protein/Adenine nucleotide translocator 1
tai04c07	IPR002048	Calcium-binding EF-hand
tai05b07	IPR000048	IQ calmodulin-binding region
tai05d07	IPR009828	Protein of unknown function DUF1394
tai05e10	IPR001650	Helicase, C-terminal
tai05h01	IPR001506	Peptidase M12A, Astacin
tai06f03	IPR002049	EGF-like, laminin
tai07d01	IPR001380	Ribosomal protein L13e
tai07g11	IPR001752	Kinesin, motor region
tai08b05	IPR000608	Ubiquitin-conjugating enzyme, E2
tai08b10	IPR000547	7-Fold repeat in clathrin and VPS proteins
tai08c10	IPR000608	Ubiquitin-conjugating enzyme, E2
tai08h10	IPR000637	HMG-I and HMG-Y, DNA-binding
tai09d11	IPR005028	Herpes virus intermediate/early protein 2/3
tai09h02	IPR012541	DBP10CT
tai10a03	IPR004116	Amelogenin
tai10f05	IPR000926	GTP cyclohydrolase II
tai10f09	IPR007087	Zinc finger, C2H2-type
tai11f02	IPR011076	Malate synthase-like
tai12a08	IPR006706	Extensin-like region
tai13f04	IPR014038	Translation elongation factor EF1B, beta and delta chains, guanine nucleotide exchange
tai14a08	IPR006885	ETC complex I subunit conserved region
tai15b12	IPR001633	EAL
tai16a07	IPR000585/IPR001506/IPR001818 /IPR006026	Hemopexin/Peptidase M12A, astacin/Peptidase M10A and M12B, matrixin and adamalysin/Peptidase, metallopeptidases
tai16g08	IPR001806/IPR002041/IPR003577	Ras GTPase/Ran GTPase/Ras small GTPase, Ras type
tai16h12	IPR006706	Extensin-like region
tai17e07	IPR000504	RNA-binding region RNP-1 (RNA recognition motif)
tai17h06	IPR007087	Zinc finger, C2H2-type
tai18e07	IPR006706	Extensin-like region
tai19a12	IPR000719/IPR001772/IPR003527 /IPR008350	Protein kinase/Kinase-associated KA1/MAP kinase/ERK3/4 MAP kinase
tai19b07	IPR006649	Like-Sm ribonucleoprotein, eukaryotic and archaea-type, core
tai19d10	IPR000010/IPR003243	Proteinase inhibitor I25, cystatinProteinase inhibitor I25A and I25B, type 2 and phytocystatins
tai21a09	IPR000772/IPR006706	Ricin B lectin/Extensin-like region
tai21h03	IPR005649/IPR006706	Chorion 2/Extensin-like region
tai22d02	IPR010674/IPR012973	Nucleolar GTP-binding 1/NOG, C-terminal
tai22g03	IPR002035	von Willebrand factor, type A
tai22g06	IPR001952	Alkaline phosphatise
tai25b10	IPR004000	Actin/actin-like
tai25h08	IPR007783	Eukaryotic translation initiation factor 3, subunit 7
tai26g03	IPR002007	Haem peroxidase, animal
tai27c11	IPR002672	Ribosomal L28e protein

Clone name	Inter Pro annotation number	Sequence annotation
tai27f06	IPR008197	Whey acidic protein, core region
tai27h09	IPR001442/IPR002952	Type 4 procollagen, C-terminal repeat/Eggshell protein
tai28c11	IPR001993	Mitochondrial substrate carrier
tai28h10	IPR000741	Fructose-bisphosphate aldolase, class-I
tai30c12	IPR001305/IPR001623/IPR002939/ IPR012895	DnaJ central region/Heat shock protein DnaJ, N-terminal/Chaperone DnaJ, C-terminal/HSCB oligomerisation, C-terminal
tai31e08	IPR007932	Phage tail fibre adhesin Gp38
tai31e09	IPR000594	UBA/THIF-type NAD/FAD binding fold
tai32c09	IPR001680/IPR006692	WD-40 repeat/Coatomer WD associated region
tai32e08	IPR001152	Thymosin beta-4
tai32e08	IPR001152	Thymosin beta-4
tai32h11	IPR006652	Kelch repeat
tai35d12	IPR001196	Ribosomal protein L15
tai35e09	IPR010800	Glycine rich
tai36f10	IPR003594/IPR014762	ATP-binding region, ATPase-like/DNA mismatch repair, MutL/HexB/PMS1
tai37b04	IPR006662	Thioredoxin-related
tai37d12	IPR007087	Zinc finger, C2H2-type
tai37f09	IPR000169	Peptidase, cysteine peptidase active site
tai38f07	IPR000837	Fos transforming protein
tai39e03	IPR001680	WD-40 repeat
tai39f05	IPR002952/IPR007223	Eggshell protein/Peroxin 13, N-terminal
tai39g07	IPR001680	WD-40 repeat
tai40e06	IPR002048	Calcium-binding EF-hand
tai40g04	IPR001506/IPR003582	Peptidase M12A, astacin/Metridin-like ShK toxin
tai41h12	IPR007087	Zinc finger, C2H2-type
tai42a10	IPR000504/IPR012956	RNA-binding region RNP-1 (RNA recognition motif)/CBF, N-terminal
tai42b07	IPR008610	Eukaryotic rRNA processing
tai42d03	IPR012580	NUC153
tai42h05	IPR005127	Giardia variant-specific surface protein
tai44a02	IPR006706	Extensin-like region
tai44e01	IPR001214	SET
tai45a08	IPR001878	Zinc finger, CCHC-type
tai46c12	IPR007223	Peroxin 13, N-terminal
tai46g02	IPR013208	Lipocalin-like
tai46h04	IPR000626	Ubiquitin
tai49a01	IPR000163/IPR001107	Prohibitin/Band 7 protein
tai49b05	IPR007087	Zinc finger, C2H2-type
tai52b09	IPR000738	WHEP-TRS
tam53a02	IPR003338/IPR003960/IPR011546	AAA ATPase VAT, N-terminal/AAA ATPase, subdomain/Peptidase M41, FtsH extracellular
tam53e08	IPR005651	Protein of unknown function DUF343
tam53e12	IPR013061	Tryptophan/tryrosine permease
tam53h04	IPR001007	von Willebrand factor, type C
tam53h06	IPR007718	SRP40, C-terminal
tam54c10	IPR002952/IPR006706	Eggshell protein/Extensin-like region
tam54d03	IPR004825	Insulin/IGF/relaxin
tam55b06	IPR002048	Calcium-binding EF-hand
tam55f08	IPR006706	Extensin-like region
tam55f11	IPR002735/IPR003307	Translation initiation factor IF2/IF5/eIF4-gamma/eIF5/eIF2-epsilon
tam56b03	IPR000885/IPR001442	Fibrillar collagen, C-terminal/Type 4 procollagen, C-terminal repeat
tam56d01	IPR007502/IPR011709	Helicase-associated region/Domain of unknown function DUF1605
tam57a05	IPR007223	Peroxin 13, N-terminal
tam57c12	IPR001680	WD-40 repeat
tam57e07	IPR002823	Protein of unknown function DUF112, transmembrane
tam57h04	IPR004033	UbiE/COQ5 methyltransferase

Clone name	Inter Pro annotation number	Sequence annotation
tam58a04	IPR001506	Peptidase M12A, Astacin
tam59b09	IPR000504	RNA-binding region RNP-1 (RNA recognition motif)
tam60b12	IPR003409	MORN motif
tam60d03	IPR003084	Histone deacetylase
tam61d05	IPR000169/IPR000668/IPR012599 /IPR013201	Peptidase, cysteine peptidase active site/Peptidase C1A, papain C-terminal/ Proteinase inhibitor I29, cathepsin propeptide
tam61d06	IPR001650/IPR012541/IPR001304	Helicase, C-terminal/DBP10CT/C-type lectin
tam62b10	IPR010304	Survival motor neuron
tam62f09	IPR006662	Thioredoxin-related
tam64e08	IPR000948/IPR001921/IPR004037	Ribosomal protein HS6/Ribosomal protein L7A/L7AE
tam64f08	IPR001298	Filamin/ABP280 repeat
tam64g08	IPR007593	Interferon-induced transmembrane protein

The left column contains the clone ID number of the *Hydractinia* library. Furthermore, the table lists the sequence domain match along with the corresponding InterPro ID number. Different domains within sequences and their corresponding annotations were separated by ”/”.

Table S2. GO annotation of *Hydractinia* sequences shared with other cnidarians

A. *Hydractinia* sequences GO terms annotation shared with *Acropora*, *Nematostella* but not in *Hydra*

Clone name	Biological process		Molecular Function	
	GO number	GO terms	GO number	GO terms
HEAB-0029E05	GO:0007582	physiological process	n/a	n/a
HEAB-0029J09	n/a	n/a	n/a	n/a
HEAB-0038N23	n/a	n/a	GO:0004601	peroxidase activity
tai09b01	GO:0008277	regulation G-protein coupled receptor protein	GO:0004871	signal transducer activity
tai11f02	GO:0008152	metabolism	GO:0004474	malate synthase activity
tai11g12	GO:0006464	protein modification	GO:0016787	hydrolase activity
tai20d03	GO:0016192	vesicle-mediated transport	n/a	n/a
tai33g08	GO:0008152	metabolism	GO:0016829	lyase activity
tam56f07	n/a	n/a	n/a	n/a
HEAB-0023B24	GO:0005975	carbohydrate metabolism	GO:0004033	aldo-keto reductase activity
tam53d11	n/a	n/a	n/a	n/a

B. *Hydractinia* sequences GO terms annotation shared with *Acropora* but not in *Nematostella* and *Hydra*

HEAB-0020F05	n/a	n/a	GO:0008299	isoprenoid biosynthesis
HEAB-0024D20	n/a	n/a	GO:0000166	nucleotide binding
HEAB-0028A08	n/a	n/a	n/a	n/a
HEAB-0028B20	n/a	n/a	n/a	n/a
HEAB-0037F13	n/a	n/a	GO:0016829	lyase activity
HEAB-0039G08	n/a	n/a	GO:0000155	two-component sensor activity
HEAB-0042I20	n/a	n/a	n/a	n/a
HEAB-0020L20	n/a	n/a	GO:0005884	actin filament
HEAB-0026O12	n/a	n/a	n/a	n/a
HEAB-0029G01	n/a	n/a	n/a	n/a
HEAB-0036O10	GO:0006810	transport	GO:0000166	nucleotide binding
HEAB-0042L12	n/a	n/a	n/a	n/a
tai07g10	n/a	n/a	n/a	n/a
tai16a08	n/a	n/a	n/a	n/a
tai40g01	n/a	n/a	n/a	n/a

The left column provides the clone ID of the *Hydractinia* library. Furthermore, GO IDs are listed along with the corresponding GO terms of the two main GO categories Biological process and Molecular function. Not applicable (n/a) was considered when sequences presented no significant match to any GO term.

6.2 Additional data 2

Supplementary data related to the microarray experiments.

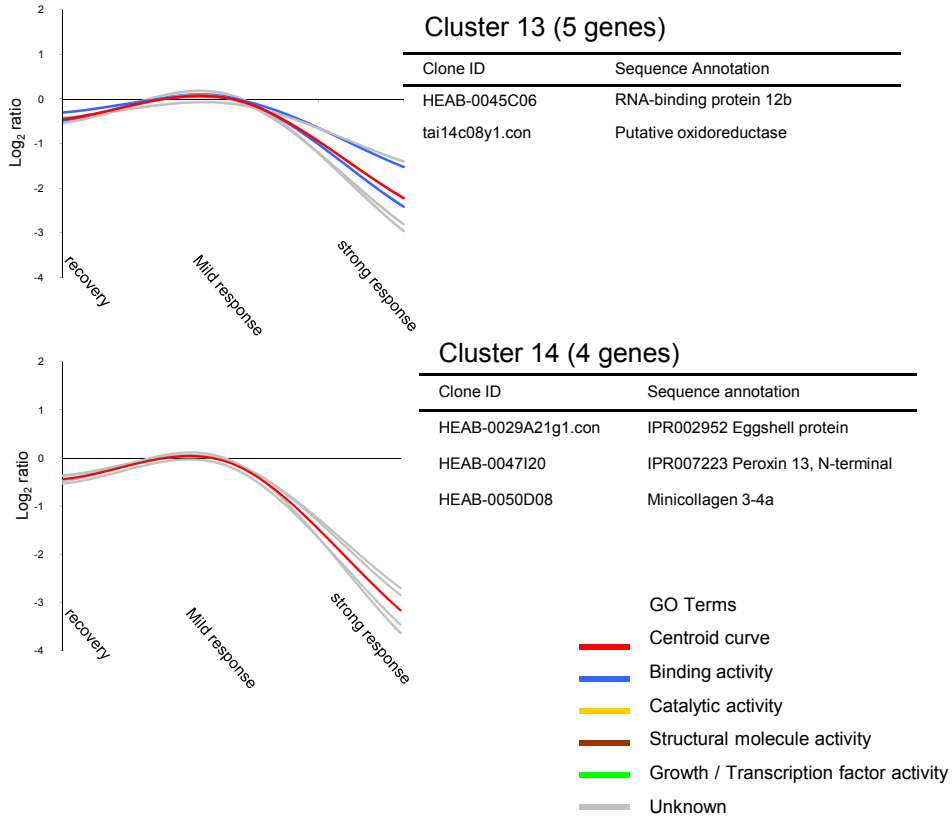
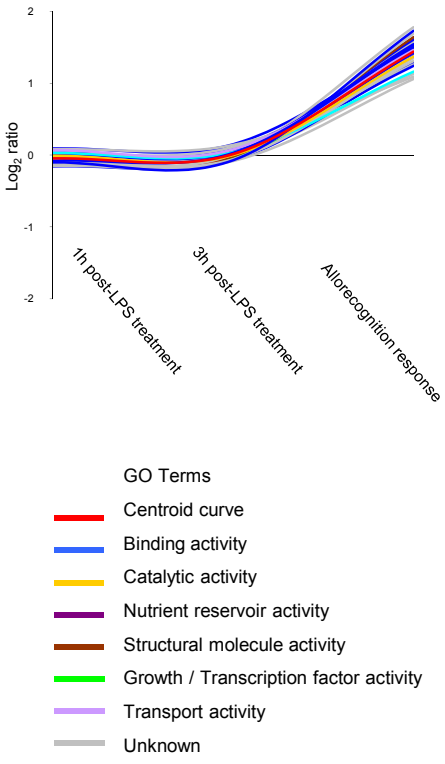
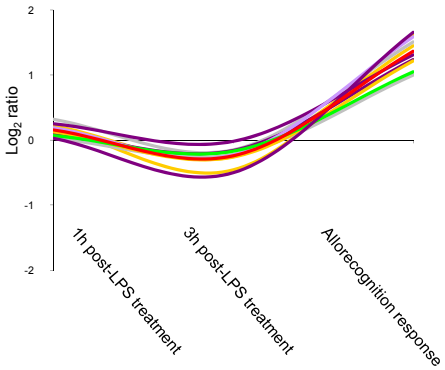


Figure S1 – Additional clusters with genes down regulated in FMR and K12 conditions. Expression level of each condition was referenced to the control using Log_2 ratio. List of genes with the corresponding annotation are provided in the table. Gene annotation was performed through BLAST, GO and Domain analysis. For an easy overview, a GO colour code of each annotated gene is provided in the transcriptional profiling curve. Unknown genes were plotted in gray colour but are not listed in the table



Cluster 4 (29 genes)

Clone ID	Sequence annotation
HEAB-0047I20	IPR007223 Peroxin 13, N-terminal
HEAB-0053F19	DNA polymerase iv
HEAB-0054A01	Integrase
HEAB-0054E08	IPR009305 DUF962
tah96a01y1.con	IPR002952 Eggshell protein
tah96c10y1.con	Predicted protein
tai02e09_J.con	3-oxoacyl-(Acyl-carrier-protein) reductase
tai04h01_O.con	GO:0003676 F nucleic acid binding
tai05d09_J.con	E3 ubiquitin-protein ligase DZIP3
tai20h05y1.con	RNA binding protein b12
tai30c12y1.con	Molecular chaperone MRJ
tai49b10y1.con	IPR006706 Extensin-like region
tam56c11y2.con	Rhamnose binding lectin
tam57a05y2.con	IPR002952 Eggshell protein
tam59a06y2.con	Minicollagen 3-4a
HEAB-0020M09g1.con	Minicollagen 3-4a
HEAB-0026C19g1.con	IPR002952 Eggshell protein
HEAB-0028J22g1.con	PL10-related protein CnPL10
HEAB-0034G03g1.con	Putative Erythrocyte membrane protein 1
HEAB-0035D10g1.con	GO:0016787 F hydrolase activity
HEAB-0044G19	Mini-collagen
HEAB-0045C06	RNA binding protein b12
tai14c08y1.con	Putative oxidoreductase



Cluster 8 (10 genes)

Clone ID	Sequence annotation
HEAB-0046F02	Zinc metalloproteinase nas-30
tai39f07y1.con	GO:0016787 hydrolase activity
tam59f09y2.con	IPR000480 Glutelin
HEAB-0048H01	Minicollagen 1
HEAB-0053F14	L-lactate permease
HEAB-0053L23	UDP-N-acetylmuramate ligase
tai07b03_J.con	Putative mini-collagen
tai07h09_O.con	Minicollagen 1

Figure S2 – Additional clusters with genes specifically up-regulated in an allorecognition challenge. Gene expression level was referenced to the control using Log_2 ratio (Condition/Control). The gene identification ID and annotations are listed in the table. A GO colour code of each annotated gene is provided in the transcriptional profiling curve. Unknown genes were plotted in gray colour but are not listed in the table.

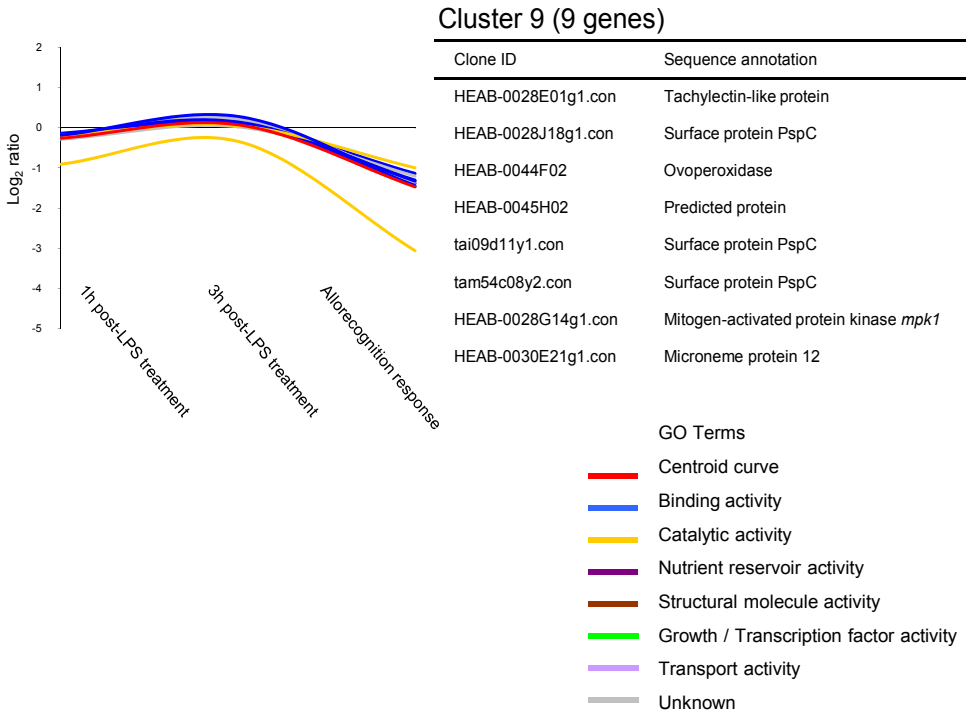
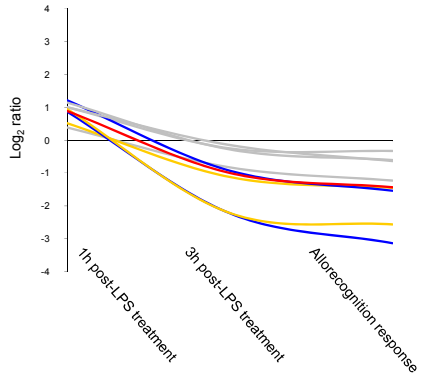


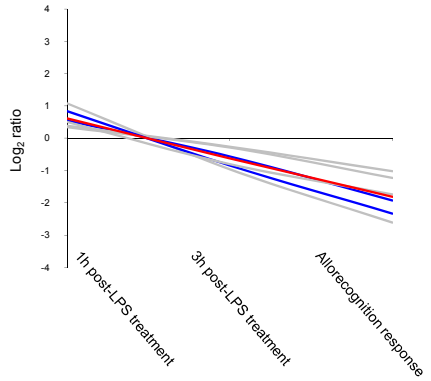
Figure S3 - Additional cluster with genes specifically down-regulated in an allorecognition challenge. Gene expression level was referenced to the control using Log_2 ratio (Condition/Control). The gene identification ID and annotations are listed in the table. A GO colour code annotation for each gene is provided in the transcriptional profiling curve. Unknown genes were plotted in gray colour but are not listed in the table.



Cluster 10 (8 genes)

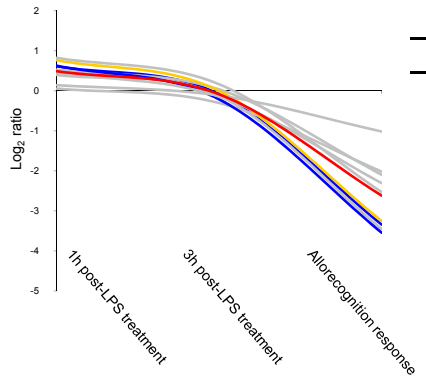
Clone ID	Sequence annotation
HEAB-0050P05	IPR007934 AbfB
HEAB-0054D11	IPR002557 Chitin-bd_peritrophin-A
tai33b04y1.con	IPR002557 Chitin bd_peritrophin-A
HEAB-0018H01g1.con	Predicted protein

GO Terms
- Centroid curve
- Binding activity
- Catalytic activity
- Nutrient reservoir activity
- Structural molecule activity
- Growth / Transcription factor activity
- Transport activity
- Unknown



Cluster 11 (6 genes)

Clone ID	Sequence annotation
tah97e08_J.con	Putative uncharacterized protein
tai06c07y1.con	Putative uncharacterized protein
tai49b05y1.con	Predicted protein
tam63e08y2.con	Sperm nuclear basic protein PL-I isoform PLIa



Cluster 20 (9 genes)

Clone ID	Sequence annotation
HEAB-0044B02	RNA-binding protein 12b
HEAB-0044B01	Putative uncharacterized protein
tai23d10y1.con	GO:0016787 F hydrolase activity
HEAB-0020F15g1_PT.con	RNA-binding protein 12b
HEAB-0044J23	IPR000258 Bacterial ice-nucleation
HEAB-0021D19g1.con	Uncharacterize IPR010310 DUF909_bac
tai01h10x1_PT_O.con	Protein phosphatase 1 regulatory subunit 3B

Figure S4 – Additional clusters with genes up-regulated immediately after LPS induction. Few genes were up-regulated after an LPS treatment. Gene expression level was referenced to the control using Log_2 ratio (Condition/Control). The gene identification ID and annotations are listed in the table. A GO colour code annotation for each gene is provided in the transcriptional profiling curve. Unknown genes were plotted in gray colour but are not listed in the table.

6.3 Additional data 3

Supplementary data related to the analysis of *CTRN* gene. The following *in situ* hybridization was performed by Dr. Brahim Mali and is published in [114].

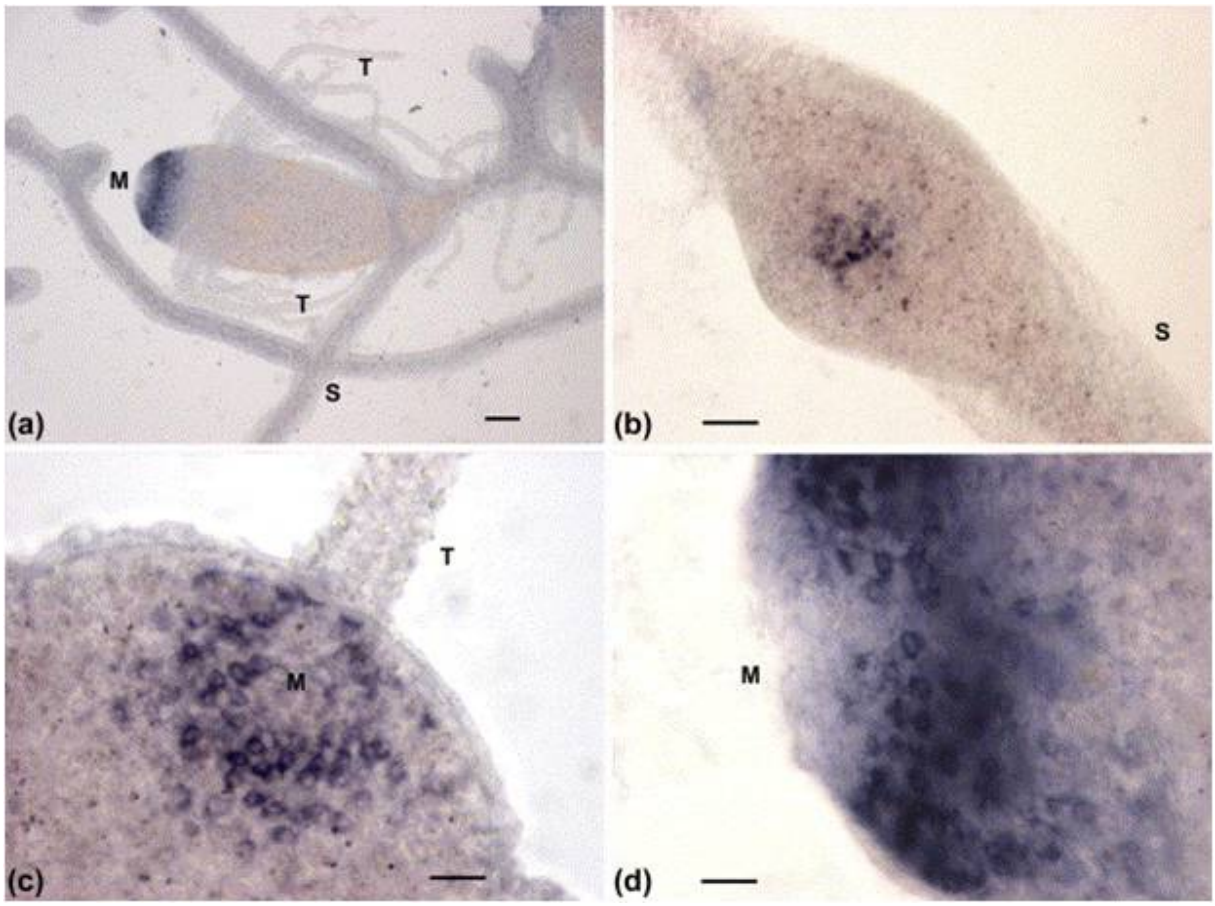


Figure S5 – Expression of the transcript *CTRN* in *Hydractinia*. The *in situ* hybridization of *CTRN* cRNA was done using a digoxigenin-labeled RNA probe. (a) An overview of a young colony. *CTRN* positive cells form a ring around the mouth. (b) An early polyp bud developing from a stolon. CTRN expressing neurons form a ring-like structure around the area of the future mouth. (c) A higher magnification of the mouth region of a polyp showing *CTRN* expressing neurons. (d) The same like the previous picture, view from the side. M, mouth; S, stolon; T, tentacle. Scale bars approximately 50 mm.