Hacking the genomes of soil arthropods

Faddeeva-Vakhrusheva, A.

2017

document version
Publisher's PDF, also known as Version of record

Chapter 1

General Introduction

1.1. Genome Analysis and Evolution

Evolutionary processes give rise to biodiversity at all levels of biological organization (Hall and Hallgrímsson 2008). The principles of evolution by natural selection are known for a long time; Charles Darwin first described them in his famous book "On the Origin of Species" published in 1859. The first factor for natural selection to drive evolution is the presence of genetic variation. The insights on the molecular mechanisms that generate variation have been greatly enhanced by the rise of biotechnology including DNA sequencing and a variety of methods to analyze gene function. Especially the genomics revolution starting in 1995 and receiving new momentum with the application of Next Generation Sequencing (NGS) methods around 2005, has been significant. It has become possible to compare genomes across species and within species and thus to analyze variation in a genome-wide manner. It also has become possible to analyze genomes from less-investigated invertebrate species that are not considered to be classical genetic models. This has given rise to new insights into the tree of life, into the nature of genetic variation and has delivered many surprises, e.g. on the number of "foreign" genes in a genome, acquired by horizontal gene transfer from other species. Stress as a selective agent is an important driver of speciation and evolution. Obviously, genotypes differ in their response to stress and associated fitness consequences. Following the genomics revolution, we can now generate more knowledge on molecular mechanisms underlying stress defense and adaptive evolution of stress tolerance.

This thesis capitalizes on the developments sketched above and aims to generate genome-wide information on the soil invertebrates *Folsomia candida* and *Orchesella cincta*, in order to understand better their evolution and ecology, specifically their responses to stress. These two species belong to the order of Collembola and have been used intensively in research on stress responses at the molecular, physiological and population levels. Moreover, they have been the focus of studying the mechanism underlying the adaptive evolution of stress tolerance. However, until recently, genome information on these animals was anecdotal and fragmented. With the recent advances and increasingly lowering costs of genome sequencing, it became feasible to generate whole genome sequence information for this intriguing group of animals. This thesis describes the research on genome structure of *O. cincta* and *F. candida*, with regard to functional annotation, gene family evolution, horizontal gene transfer, and ecological niche preference in a comparative genomic context. In this introductory chapter, I will discuss themes that are relevant for the thesis followed by a short outline of the thesis.

1.2. Comparative Genome Analysis

Comparative genomics compares genome sequences and related features among strains/species and describes their similarities and differences. Genomes can be explored on different levels, ranging from large-scale genomic rearrangement analysis to the comparison of genes, coding sequences, RNAs, regulatory regions or small-scale variation like single nucleotide polymorphisms (SNPs). Although most comparative studies focus on conserved features, inter-species differences could be a good indicator of evolutionary history and divergence, providing insights on species-specific adaptations. The main purpose of comparative genomics at the gene level is to link differences in the gene content of the species analyzed to their phenotypical features.

In order to perform a comparative analysis, a complete overview of homology relations

across species, among genes and inside gene families is essential. This is obtained by constructing orthologs gene clusters.

A homologous gene is a gene inherited in two species by a common ancestor. There are two types of homologs – orthologs and paralogs. Orthologous genes are genes in different species that originated by vertical descent from a single gene of the last common ancestor. It is assumed, and often observed, that after the split orthologs maintain identical gene functions. Paralogs are homologous genes that derived from gene duplication in an ancestral species. There are two types of paralogs: in-paralogs (within one organism, duplication after a speciation event) and out-paralogs (duplication before a speciation event). Figure 1 shows the early duplication of globin genes that gave rise to *α*-globin and *β*-globin in a species ancestral to amphibians, birds, and mammals. The *α*-globin genes of frog, chick, and mouse are orthologs, as are the *β*-globin genes of frog, chick, and mouse. The *α*-globin gene of the mouse is the paralog of the *β*-globin gene of the mouse. Similarly, *α*-globin and *β*-globin are paralogs in chick and frog.

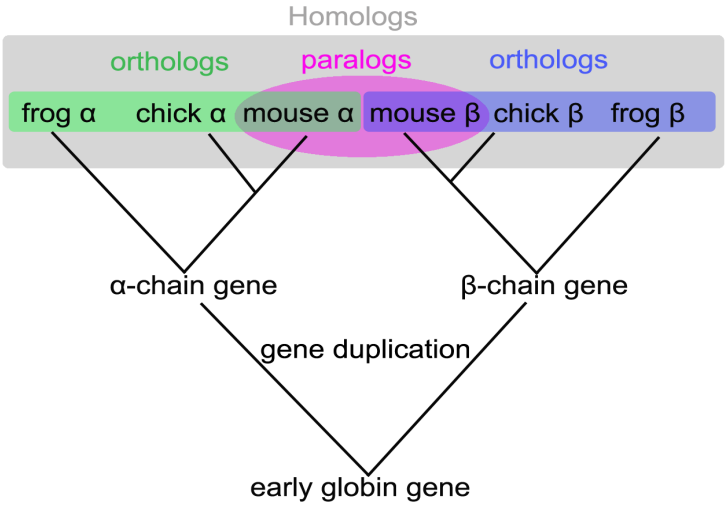

Figure 1. Evolution of globin genes.

There are different approaches for orthology analysis and prediction. The methods can be categorized into two groups: (1) graph-based methods (i.e. OrthoMCL), which cluster orthologs based on protein sequence similarity, and (2) tree-based methods (i.e. TreeFam), which not only cluster but also reconcile the protein family tree with a species tree. Tree-based methods are computationally expensive and often fail due to the complexity of the family or lack of a substantial number of species in the comparison (Kristensen *et al.* 2011). The OrthoMCL considers all-against-all Basic Local Alignment Search Tool P (BLASTP) comparisons for a set of protein sequences from organisms of interest. Afterwards, reciprocal best similarity pairs between species (putative orthologs) and within species (recent paralogs) are defined. The resulting graph is represented by a symmetric similarity matrix to which the MCL algorithm is applied to define orthologous groups with (recent) paralogs (Li *et al.* 2003). Methods that construct orthologous groups based on BLAST scores without taking into account gene length could result in missing genes in the orthologous groups that contain short genes and incorrectly clustered genes in orthologous groups that contain long genes.

Therefore, a new method, called OrthoFinder, became available recently (Emms and Kelly 2015). It solves fundamental biases in whole genome comparisons and dramatically improves orthologous group inference accuracy by gene length and phylogenetic distance normalization of BLAST bit scores.

However, some pitfalls remain, like the issue with recent gene duplications (after speciation), in this case, there is a one-to-many or many-to-many orthologs relationship. In such cases, it is non-trivial (and usually impossible) to determine which of the orthologs is functionally equivalent to the orthologs in the other species.

After orthology mapping, similarities (conservation) and differences (divergence) can be defined. Conservation of gene sequence is maintained by purifying selection. This often applies to genes with a high functional importance. Divergence could be nonfunctional (by genetic drift) and functional adaptation (due to positive selection) or loss of function (due to lack of selection) (Tirosh *et al.* 2007). How to distinguish between these categories will be explained further in section 3 (Adaptive evolution), as well as the principle to use ortholog groups in the investigation of gene family evolution.

1.3. Stress Response

By means of the cellular stress response, organisms can tolerate a certain stress impacting the ecological niche they inhabit. This involves induction of stress-combating proteins. Some of these proteins have very long evolutionary histories while others are specific to the phylogenetic lineage or even the organism. Stress is a major reason of natural selection. Genotypes that are more resistant to the stress are favored and replace less adapted ones (Van Straalen and Roelofs 2012).

The study of stress responses on the genomic level allows understanding the mechanisms that enable organisms to survive harsh environments. There could be different types of stress factors i.e., heat, cold, drought, salt, hypoxia, ROS (reactive oxygen species), herbivory/ microbial infection, toxic substances (heavy metals, pesticides). Although there are different stress factors it is suggested that there is a uniformity in the stress responses of different species to different factors (Van Straalen and Roelofs 2012). Korsloot *et al.* (Korsloot *et al.* 2004) suggested the following types of stress defense systems in arthropods:

- Basal signal transduction systems (i.e., MAPK)
- Stress proteins (i.e., heat-shock)
- The oxidative stress response (i.e, glutathione, catalase)
- Metallothionein and associated systems
- Mixed function oxygenase (cytochrome P450, glutathione-S-transferase)

Korsloot *et al.* also suggested that stress defense systems are interconnected with each other and they work as a single, integrated, cellular stress defense system (Korsloot *et al.* 2004). In general, stress signals activate transcription factors (*trans*-acting factors), which bind to specific DNA sequences (*cis*-regulatory elements) in the promoter regions of stress-induced genes (Van Straalen and Roelofs 2012). Often many genes have promoters that have regions responsive to different stimulus (i.e., metallothionein has metal-responsive, antioxidant-responsive, and steroid hormone receptor-binding sites). Also often genes could

have the same transcription factor-binding site in the promoters or they are regulated by the same promoter and as a result, they could be induced by the same transcription factor.

1.4. Adaptive Molecular Evolution

1.4.1. Protein Structure

Polymorphisms in the DNA of an organism are a potential source of phenotypic variation and they could contribute to microevolution and speciation. However, it is often assumed that non-adaptive changes (i.e. mutation, recombination, genetic drift) are dominating genome architecture, while natural selection is playing a minor role on the molecular level (Lynch and Conery 2003). Others argue against this statement (Protas *et al.* 2006; Yim *et al.* 2014; Zakon *et al.* 2006; Zhang 2006). Deleterious mutations are expected to be quickly eliminated by purifying selection, while slightly deleterious mutations are not quickly removed and could even be fixed due to random genetic drift.

Adaptive changes may be caused by various mechanisms of gene evolution. One of the mechanisms of adaptive evolution is altering the structure of proteins through mutations in the coding regions of genes. If such mutations make the organisms more resistant against stress, selective pressure may be strong and the population as a whole may evolve an enhanced stress tolerance.

The selective pressure in protein coding genes can be detected within the framework of comparative genomics. The selective pressure is assumed to be defined by the ratio $\omega = d_N/d_S$, where d_S represents the rate of synonymous substitutions (keeping the amino acid) and d_N the rate non-synonymous substitutions (changing the amino acid). In the absence of evolutionary pressure, the synonymous rate and the non-synonymous rate are equal, so the d_N/d_S ratio is equal to 1. Under purifying selection, natural selection prevents the replacement of amino acids, so the d_N will be lower than the d_S, and $d_N/d_S < 1$. And under positive selection, the replacement rate of the amino acid is favored by selection and $d_N/d_S > 1$.

There are different approaches to detect positive selection e.g. approximate methods, counting methods and maximum-likelihood methods. Approximate methods involve the following steps:

(1) Counting the number of synonymous and nonsynonymous sites in the two sequences, or estimating this number by multiplying the sequence length by the proportion of each class of substitution;

(2) Counting the number of synonymous and nonsynonymous substitutions; and

(3) Correcting for multiple substitutions.

The maximum-likelihood approach uses probability theory to complete all three steps simultaneously (e.g. CodeML from PAML). It estimates critical parameters, including the divergence between sequences and the transition/transversion ratio, by deducing the most likely values to produce the input data (Yang and Bielawski 2000).

There are different codon models that exist in CodeML: the branch models that estimate different d_N/d_S among lineages, the site models that estimate different d_N/d_S among sites, and the branch-site models that estimate different d_N/d_S among sites and among branches.

To define positive selection, two models are computed: a null model, in which the foreground

branch may have different proportions of sites under neutral selection to the background (i.e. relaxed purifying selection), and an alternative model, in which the foreground branch may have a proportion of sites under positive selection. A chi-squared test is applied to distinguish between the null model and the observed results (alternative model). A significant result with the branch-site codon model means that positive selection has affected a subset of sites along the specified lineage during a specific evolutionary time (also called episodic model of protein evolution).

An example of adaptive evolution is immune system genes such as immunoglobulin (*Ig*) and major histocompatibility complex (*MHC*). These genes tend to evolve faster to protect the host from ever changing parasites (viruses, bacteria, etc.) (Nei 2007).

Another example of conserved genes are the *HOX* genes. The high conservation of these genes is explained by the high degree of purifying selection rather than a low mutation rate. It is suggested that 99.7% of nonsynonymous mutations are eliminated by purifying selection in homeobox regions (Nei 2007).

1.4.2. Transcriptional Regulation

Altering the amount of protein through changes in transcription is another mechanism of adaptive evolution. This could be achieved by substitutions in the gene's promoter (*cis*-regulatory change) or by altering the structure or number of transcriptional regulatory proteins (*trans*-regulatory change). It is assumed that promoter sequences (a stretch of DNA located upstream of a transcription initiation site) evolve much faster than coding sequences since they are not limited by the strict evolutionary constraints on open reading frames. It is less likely that these mutations cause lethal phenotypes, and it is therefore expected that the variation in promoters is larger than in coding sequences (Van Straalen and Roelofs 2012). Indeed, Janssens *et al.* showed that extensive genetic variation in the promoter of the metallothionein gene of *O. cincta* (a gene encoding for metal binding protein) influences the metallothionein protein expression and contributes to cadmium tolerance (Janssens *et al.* 2007). TATA box promoters could be especially important for evolutionary adaptations and enhanced stress tolerance since they are associated with genes that have variable gene expression and changes in such promoters could provide a sudden burst of gene expression under environmental stress (López-Maury *et al.* 2008; Roelofs *et al.* 2010).

1.4.3. Gene Families

Another mechanism of adaptive evolution is to change the size dynamics of gene families (expansions/gain, contractions/loss). Gain and loss of gene families is a major source of genetic variation and novel gene functions and it allows for future adaptations. In *Daphnia pulex*, it is suggested that extensive expansions of gene families in metabolic pathways (non-random gene amplification) are the main mechanism of environmental adaptation (Colbourne *et al.* 2011). It is suggested that amplification of gene families in *Daphnia* allows the coordinated induction of expression of paralogs (Colbourne *et al.* 2011; Shaw *et al.* 2007), a process called super functionalization (Van Straalen and Roelofs 2012).

Another category of raw material for the evolution of new gene functions are orphan genes. They may play a very important role in lineage-specific adaptations (Tautz and Domazet-Loso 2011). Genes can be classified as orphan genes if they lack detectable similarity to genes in other species and therefore no clear signals of common descent (i.e., homology) can be inferred (Wissler *et al.* 2013). BLAST is a common tool to estimate nucleotide or protein sequence similarity. However, short and rapidly evolving genes could be missed by BLAST (Moyers

and Zhang 2014). Alternatively, phylostratigraphy (Domazet-Lošo *et al.* 2007), a method for dating the evolutionary emergence of a gene or gene family, could be used to detect orphan genes. Phylostratigraphy generates a phylogenetic tree in which the homology is calculated between all genes of a focal species and the genes of other species, and the earliest common ancestor for a gene determines the age of the gene (phylostratum). Orphan genes are those genes located within the highest phylostratum.

1.4.4. Horizontal Gene Transfer

Another mechanism that contributes to diversification and change in genome content and structure during animal evolution is horizontal gene transfer (HGT). HGT is defined as the transfer of genes between phylogenetically unrelated organisms (absence of a common ancestor). It is very common and well-studied in bacteria and unicellular eukaryotes, where it plays an important role in adaptive evolution (Polz *et al.* 2013). In contrast, it was assumed that HGT is rare in multicellular organisms. Nowadays, there are more and more studies that suggest that HGT also happens in these organisms (Boto 2014; Crisp *et al.* 2015; Hotopp 2011; Keeling and Palmer 2008; Scholl *et al.* 2003).

In prokaryotic genetics, it is generally assumed that horizontal gene transfer is limited to operational genes, that is, genes conferring specific resistance phenotypes that are often encoded on plasmids. Informational genes, which define the basic biochemical pathways of a species, and its cellular structure, are assumed to be less prone to horizontal gene transfer. As a consequence, the species concept in prokaryotic biology can still be maintained (although with wide margins), because the informational genes that define the species do not easily skip from one species to another. This is known as the complexity theory.

Indeed, some well-known examples of genetic adaptation to biotic stress (predation) are due to HGT, including transfer of carotenoid biosynthesis genes to the pea aphid (resulting in beneficial pigmentation, because the predator cannot recognize the aphids anymore on a certain substrate) (Moran and Jarvik 2010), stress adaptation to toxic compounds: transfer of a cysteine synthase into the arthropod lineage (resulting in detoxification of plant produced cyanide) (Wybouw *et al.* 2014). However, most HGT cases in eukaryotes are observed between symbionts and their hosts, parasites and their hosts, and between animals and their food (Van Straalen and Roelofs 2012), and there is not always clear evidence that a HGT-derived gene confers a fitness advantage. As an example, consider genes from the bacterium *Wolbachia* that are transferred to invertebrates, most probably due to a long-term intimate association (Hotopp *et al.* 2007; Kondo *et al.* 2002), although a new function in their host is not always clear. The most common fate of horizontally transferred genes in the new host is a loss of function.

1.5. Scope of the thesis

As mentioned in the beginning of this chapter, this thesis focuses on two collembolans, *F. candida* and *O. cincta*, their genomes and transcriptomes, and comparative analysis with other organisms. We discuss in more details different evolutionary events i.e., genes under positive selection, gene family expansion, gene family gain/loss, horizontal gene transfer and biological processes associated with them. We also pay special attention to families that are linked with stress response and metal tolerance (in *O. cincta*) and speculate on the signatures of Collembola genome evolution in relation to living inside soil or soil litter layer.

This thesis work is performed in the context of the supporting BE-Basic Foundation, which I will introduce shortly below. Subsequently, I describe the importance to monitor soils, giving a short overview of traditional and innovative (developed during BE-BASIC) methods of soil quality assessment and describe the relevance of collembolans models for soil risk assessment. Finally, I present aims and outline of the thesis and sketch the perspective for each of the chapters.

1.6. BE-BASIC

This PhD project was conducted within the BE-Basic Foundation (Bio-based Ecologically Balanced Sustainable Industrial Chemistry Consortium), which is an international public-private partnership that develops industrial bio-based solutions to build a sustainable society. It initiates the collaborations between academia and industry, between the Netherlands and abroad to work on fundamental science and industrial challenges required for the transition to a bio-based economy. Innovative projects on bio-chemicals, bio-materials, bio-construction concepts and bio-based monitoring tools are the main focus of BE-BASIC. Research projects in BE-BASIC are organized in twelve Flagships; each of them is addressing a major scientific/socio-economic challenge. This project is a part of "Sustainable soil management and upstream processing" Flagship. Flagship 8 aims at the development of innovative tools for environmental quality and human health. This is important because these tools will monitor the sustainable production of chemical and biofuels. By generating genome information on less investigated invertebrates, their use as test organisms for environmental quality can be improved (Chen 2016). In addition, genomes of less-well investigated invertebrates may reveal hitherto unknown genes that can be candidates for new biosynthesis pathways. In this way, the flagship provides causal evidence that the production methods are really bio-based (do not impact man and environment and do not exhaust natural resources). To that end, several tools were developed; their features and relation to classical test methods are discussed below.

1.7. Soil Quality Assessment

A soil is an important natural resource. It is a mediator of biochemical processes that supply plants with nutrients and water and allowing them to grow, a habitat for many organisms and a storehouse of water, minerals and fossil resources. Soils influence natural ecosystems, agricultural productivity, water quality, carbon, and nutrient cycling, air quality, and the global climate. Pollution is one of the major threats to soil function (Rodrigues *et al.* 2009). Protecting and improving soil quality is, therefore, crucial.

The combination of Chemistry, Toxicity and Ecology is used for environmental risk assessment (Jensen *et al.* 2006; Long and Chapman 1985). For soil quality assessment bioassays are commonly used to measure the ecotoxicological impact of compounds on soil-living organisms. During the traditional "inverse" test, model species (validated in international standard tests) are exposed to the soil containing different concentrations of the compound and their survival, reproduction, and growth are evaluated. These parameters allow to define dose-response relationship and to estimate 10% and 50% effect concentrations (EC10 and EC50 correspondingly) for the prediction of the maximal acceptable concentration of a chemical that is safe for the environment. The disadvantage of the method is that for the experiments, a standardized soil is used (i.e., LUFA2.2, sand-clay-peat mixture recommended by OECD),

which have properties different from field soils and this could affect the measurements.

Another "forward" type of bioassays was suggested to evaluate soils from potentially contaminated sites and to compare biological responses of model organisms exposed to tested soils and to clean reference soils. A disadvantage of this method is that it is difficult to find reference soil with the same properties and without contamination. In general, bioassays are time-consuming and laborious. Furthermore, only two parameters are measured (survival and reproduction), and the outcome does not give insight into specific toxic mechanisms potentially caused by toxic compounds.

To improve soil quality monitoring tools the iSQ (invertebrate soil quality) chip was designed in Animal Ecology Department at the Vrije Universiteit (VU) Amsterdam on the basis of the *F. candida* EST database (Nota *et al.* 2009; Nota *et al.* 2008). In contrast, the exposure time in the iSQ test has been decreased to 2 days, and the generated gene expression profiles provide mechanistic insights into the biological effects exerted by the potentially toxic compounds, it reveals the level of toxicity and potentially identifies the toxic compound that is causing these adverse effects. Although the ISQ chip is a significant progress in soil ecotoxicogenomics, the current version contains only 5.069 unique gene probes. Partial availability of *F. candida* transcript sequences hampers an in-depth analysis of stress response pathways in a soil ecological relevant setting. That is why the complete genome and transcriptome information is needed.

1.8. Collembola: Their Biology and Role as Model in Soil Quality Assessment

Collembola (springtails) are a soil-living lineage of wingless hexapods that play an important ecological role as decomposers of organic matter in the soil. They share the most recent common ancestor with insects (Misof *et al.* 2014). *F. candida* belongs to the family Isotomidae and *O. cincta* to Entomobryidae.

Members of Collembola are normally less than 6 mm long, have six or fewer abdominal segments and possess a tubular appendage (ventral tube) on the first abdominal segment, used for osmoregulation. Both *Orchesella* and *Folsomia* have an abdominal, tail-like appendage, the furcula, that is used for jumping when the animal is disturbed. The furcula, as well as ventral tube, are features unusual for hexapods. Although collembolans are a sister group to insects, they lack some insect features such as wings, malpighian tubules, and cessation of moulting in the adult stage. Collembolans moult repeatedly during their entire life, they also shed the gut epithelium at each moult, which helps them to get rid of metals.

Features different between *Folsomia* and *Orchesella* are summarized in the Table 1 below:

Table 1. Features different between *F. candida* and *O. cincta*.

	F. candida	*O. cincta*
Habitat	Inside soil	On top of soil (litter layer)
Reproduction	Parthenogenetic, rarely sexual	Sexual
Endosymbiont	*Wolbachia* demonstrated	Low frequencies reported in some populations
Morphology features	Cuticle lacks pigment, eyes are not developed, short appendages	Cuticle is pigmented, body is covered with hair, eyes are developed, long legs and antennae
Body size	~2.5 mm	~6 mm
Diet	Bacteria, fungi, algae	Algae, lichens

At the Animal Ecology Department at the VU Amsterdam *F. candida* and *O. cincta* gene expression profiles are used to design classifiers for ecotoxicological testing on specific soil pollutants. *F. candida* was chosen as a model organism for soil quality control (Fountain and Hopkin 2005; Organisation for Economic Co-operation and Development 2009; The International Organization for Standardization (ISO) 1999) and *O. cincta* as a model organism for studying long-term pollution of the soil (Posthuma *et al.* 1993; Roelofs *et al.* 2009; Roelofs *et al.* 2007).

In general, earthworms, enchytraeids, and collembolans are the most widely used organisms for soil risk assessment because they are easy to culture and they have a relatively short generation time (Achazi *et al.* 1997; Fountain and Hopkin 2005; Ronday and Houx 1996). Most researchers used *F. candida* as a test organism for soil risk assessment; as a result, the species was recommended by the International Standards Organization (ISO) as model test organism (The International Organization for Standardization (ISO) 1999). *Folsomia* has short reproduction cycle, it has a parthenogenetic mode, it is easy to culture, and it is very sensitive to toxic compounds, which makes it a good model organism for ecotoxicological research (Fountain and Hopkin 2005). It has also been employed as a model for studying cold tolerance, quality as a prey item, and effects of microarthropod grazing on pathogenic fungi and mycorrhizae

The other springtail, *O. cincta*, has both sensitive and metal-tolerant populations; therefore, it is an appropriate model to address the effect of pollutants on the stress response system and for identifying genes that are possibly involved in metal tolerance (Roelofs *et al.* 2009; Roelofs *et al.* 2007). In contrast to *F. candida*, it is easily found in the field and population densities can be estimated by core sampling. This makes *O. cincta* a good model to perform classical ecological population studies in the field including experiments in contaminated soils. It is also an interesting species for mate preference and sexual selection, due its method of indirect sperm transfer without physical contact: females are fertilized after picking a spermatophore that has been deposited on the substrate by a male (Zizzari *et al.* 2009). However, the fragmentary EST information makes ecological analysis difficult, so there is also a need for the complete transcriptome and genome of *O. cincta*.

1.9. Aim and Outline of the Thesis

In this thesis, we analyzed two collembolan species, *F. candida* and *O. cincta*. Since they are commonly used for ecotoxicological research, there is a need to know more about these organisms, about their genome and transcriptome, the evolution of gene families and mechanisms of adaptations to the soil environment. This information could extend our general knowledge on these organisms, understand their evolution within the hexapod lineage, allowing further research of specific genes and gene families and, consequently, improve the ecotoxicological tests that are available nowadays.

The main research tasks of the thesis were:

1) to assemble and annotate two collembolan transcriptomes and genomes;
2) to characterize the genome content;
3) to investigate the evolution of gene families
4) to study the stress response pathways and pre-adaptations to stress; and
5) to develop a genome browser for both species, so the genomic/transcriptomic information would be visual and available for other researchers.

We also addressed the following, more fundamental questions:

1) How did collembolans evolve? In other words, what makes collembolans "collembolans"?
2) What are the mechanisms that contributed to hexapod evolution on land, from an aquatic, crustacean ancestor?
3) What are the genetic characteristics conferring pre-adaptation to the soil environment?

In Chapter 2 we present transcriptomes of two collembolans *F. candida* and *O. cincta*, which are key organisms in the evolutionary link between crustaceans and insects. To answer a question about the molecular evolution of collembolans and hexapods and their possible adaptation to terrestrial lifestyle, we employed state-of-the-art computational methods to analyze these two transcriptomes along with other arthropod genomes in an evolutionary context. We applied robust tests of positive selection, based on the branch-site codon model and subsequent gene ontology enrichment analysis to identify functional categories that are overrepresented among the gene sets that show positive selection. We discussed these functional categories in more details.

In Chapter 3 we present the first genome of a collembolan, *O. cincta*. The genome was assembled from Illumina and Pacific Bioscience sequence data. We analyzed gene family expansions, lineage-specific families and horizontally transferred genes. Previously, it was reported, that *O. cincta* has evolved stress tolerance against metal pollution in the soil caused by anthropogenic activity. To answer the question about mechanisms involved in pre-adaptations to soil environment we checked if some of expanded, lineage-specific gene families and horizontally transferred genes were involved in metal tolerance based on previous gene expression data. We discuss gene families linked to metal stress response and

the associated biological processes in the context of adaptation to the soil environment and associated adverse effects.

Chapter 4 presents the reference genome of another springtail, *F. candida*. In order to describe specific genomics features linked to the living in a soil, we discuss gene family expansions, lineage-specific families, horizontal gene transfer (HGT), and *HOX* genes that are responsible for morphological development. We also discuss gene family expansions in the context of metal stress response and metal tolerance. Furthermore, to know more about genome organization of *F. candida*, we additionally performed collinearity analysis.

The final Chapter 5 is a general discussion on results from previous chapters.

1.10. References

Achazi R, Chroszcz G, Mierke W 1997. Standardization of test methods with terrestrial invertebrates for assessing remediation procedures for contaminated soils. Eco-Informa 12: 284-289.

Boto L 2014. Horizontal gene transfer in the acquisition of novel traits by metazoans. Proc. R. Soc. B, Biol Sci. 281: 20132450.

Chen G 2016. New Tools for Assessment of Soil Toxicity towards the Bio-based Economy. [The Netherlands]: Vrije University Amsterdam.

Colbourne JK, *et al.* 2011. The ecoresponsive genome of *Daphnia pulex*. Science 331: 555-561.

Crisp A, Boschetti C, Perry M, Tunnacliffe A, Micklem G 2015. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. Genome Biology 16: 50.

Domazet-Lošo T, Brajković J, Tautz D 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends in Genetics 23: 533-539.

Emms DM, Kelly S 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology 16: 1-14.

Fountain MT, Hopkin SP 2005. *Folsomia candida* (COLLEMBOLA): A "Standard" Soil Arthropod. Annu Rev Entomol 50: 201-222.

Hall BK, Hallgrímsson B. 2008. Strickberger's evolution: Jones & Bartlett Learning.

Hotopp JCD 2011. Horizontal gene transfer between bacteria and animals. Trends in Genetics 27: 157-163.

Hotopp JCD, *et al.* 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317: 1753-1756.

Janssens TK, *et al.* 2007. Recombinational micro-evolution of functionally different metallothionein promoter alleles from *Orchesella cincta*. BMC evolutionary biology 7: 88.

Jensen J, *et al.* 2006. Ecological risk assessment of contaminated land-Decision support for site specific investigations: RIVM.

Keeling PJ, Palmer JD 2008. Horizontal gene transfer in eukaryotic evolution. Nature Reviews Genetics 9: 605-618.

Kondo N, Nikoh N, Ijichi N, Shimada M, Fukatsu T 2002. Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. Proceedings of the National Academy of Sciences 99: 14280-14285.

Korsloot A, van Gestel CA, Van Straalen NM. 2004. Environmental stress and cellular response in arthropods: CRC Press.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV 2011. Computational methods for Gene Orthology inference. Briefings in bioinformatics 12: 379-391.

Li L, Stoeckert CJ, Jr., Roos DS 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-2189.

Long ER, Chapman PM 1985. A sediment quality triad: measures of sediment contamination, toxicity and infaunal community composition in Puget Sound. Marine Pollution Bulletin 16: 405-415.

López-Maury L, Marguerat S, Bähler J 2008. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. Nature Reviews Genetics 9: 583-593.

Lynch M, Conery JS 2003. The origins of genome complexity. Science 302: 1401-1404.

Misof B, *et al.* 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346: 763-767.

Moran NA, Jarvik T 2010. Lateral transfer of genes from fungi underlies carotenoid production in aphids. Science 328: 624-627.

Moyers BA, Zhang J 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. Molecular Biology and Evolution 32:258-67.

Nei M 2007. The new mutation theory of phenotypic evolution. Proceedings of the National Academy of Sciences 104: 12235-12242.

Nota B, Bosse M, Ylstra B, Van Straalen NM, Roelofs D 2009. Transcriptomics reveals extensive inducible biotransformation in the soil-dwelling invertebrate *Folsomia candida* exposed to phenanthrene. BMC Genomics 10: 236.

Nota B, *et al.* 2008. Gene expression analysis of collembola in cadmium containing soil. Environmental science & technology 42: 8152-8157.

Organisation for Economic Co-operation and Development O. 2009. Test No. 232: Collembolan Reproduction Test in Soil: OECD Publishing.

Polz MF, Alm EJ, Hanage WP 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends in Genetics 29: 170-175.

Posthuma L, Hogervorst RF, Joosse EN, Van Straalen NM 1993. Genetic variation and covariation for characteristics associated with cadmium tolerance in natural populations of the springtail *Orchesella cincta* (L.). Evolution: 619-631.

Protas ME, *et al.* 2006. Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. Nat Genet. 38: 107-111.

Rodrigues SM, Pereira ME, da Silva EF, Hursthouse A, Duarte A 2009. A review of regulatory decisions for environmental protection: Part I—Challenges in the implementation of national soil policies. Environment International 35: 202-213.

Roelofs D, *et al.* 2009. Adaptive differences in gene expression associated with heavy metal tolerance in the soil arthropod *Orchesella cincta*. Mol Ecol 18: 3227-3239.

Roelofs D, Marien J, van Straalen NM 2007. Differential gene expression profiles associated with heavy metal tolerance in the soil insect *Orchesella cincta*. Insect Biochemistry and Molecular Biology 37: 287-295.

Roelofs D, Morgan J, Stürzenbaum S 2010. The significance of genome-wide transcriptional regulation in the evolution of stress tolerance. Evolutionary Ecology 24: 527-539.

Ronday R, Houx N 1996. Suitability of seven species of soil-inhabiting invertebrates for testing toxicity of pesticides in soil pore water. Pedobiologia 2: 106-112

Scholl EH, Thorne JL, McCarter JP, Bird DM 2003. Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach. Genome Biology 4: 1.

Shaw JR, *et al.* 2007. Gene response profiles for *Daphnia pulex* exposed to the environmental stressor cadmium reveals novel crustacean metallothioneins. BMC Genomics 8: 1.

Tautz D, Domazet-Loso T 2011. The evolutionary origin of orphan genes. Nat Rev Genet 12: 692-702.

The International Organization for Standardization (ISO). 1999. Guideline 11267 Soil quality - inhibition of reproduction of Collembola (*Folsomia candida*) by soil pollutants. In.

Tirosh I, Bilu Y, Barkai N 2007. Comparative biology: beyond sequence analysis. Current Opinion in Biotechnology 18: 371-377.

Van Straalen NM, Roelofs D. 2012. An introduction to ecological genomics: Oxford University Press.

Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E 2013. Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol Evol 5: 439-455.

Wybouw N, *et al.* 2014. A gene horizontally transferred from bacteria protects arthropods from host plant cyanide poisoning. Elife 3: e02365.

Yang Z, Bielawski JP 2000. Statistical methods for detecting molecular adaptation. Trends in ecology & evolution 15: 496-503.

Yim HS, *et al.* 2014. Minke whale genome and aquatic adaptation in cetaceans. Nat Genet. 46: 88-92.

Zakon HH, Lu Y, Zwickl DJ, Hillis DM 2006. Sodium channel genes and the evolution of diversity in communication signals of electric fishes: convergent molecular evolution. Proc Natl Acad Sci U S A 103: 3675-3680.

Zhang J 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat Genet. 38: 819-823.

Zizzari ZV, Braakhuis A, van Straalen NM, Ellers J 2009. Female preference and fitness benefits of mate choice in a species with dissociated sperm transfer. Animal Behaviour 78: 1261-1267.