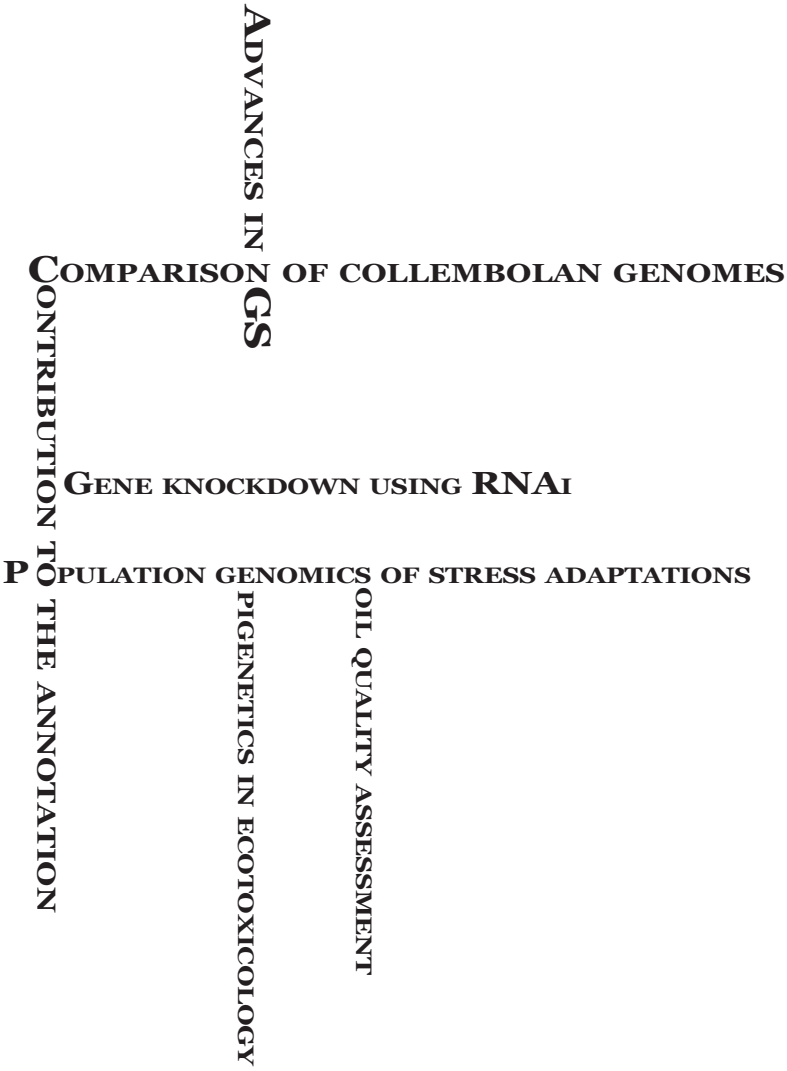
Chapter 5

General Discussion



In this thesis, we investigated two collembolans, *Folsomia candida* and *Orchesella cincta*. These organisms have a great importance for ecotoxicological and soil ecological research, but until recently very limited information was available about their genomes. Genome data are crucial for the further development of the evolutionary ecological and ecotoxicological research lines using these species. To meet these challenges, we have assembled and annotated the genomes and transcriptomes of *F. candida* and *O. cincta* and characterized the gene complement and other aspects of genome content. Since collembolans represent a soil-living lineage of hexapods in between insects and crustaceans, their genes could hold information on the early processes leading to evolution of Hexapoda from a crustacean ancestor. To understand the mechanisms that contributed to hexapod evolution on land, from an aquatic, crustacean ancestor, we analyzed genes showing signs of positive selection in the hexapod and collembolan lineages. Additionally, to have a better perspective on how collembolans evolve and what makes them "collembolans", we investigated the evolution of gene families, i.e., expansions, lineage-specific families, horizontal gene transfer (HGT) genes, and *Hox* genes cluster. Another research purpose was to identify genetic characteristics conferring possible pre-adaptation to the soil environment. To that end, we examined the link between the evolution of gene families and microarray data on metal stress-response and metal tolerance. The final goal was to incorporate all information into a genome browser.

In Chapter 2 we introduce *F. candida* and *O. cincta* transcriptomes and address the mechanisms that contributed to the hexapods evolution on land. The results of this chapter allow us to understand the evolution of collembolans and arthropods at the molecular level. The identified gene families and associated gene ontology categories could represent processes that have played a key role in the divergence of collembolans and hexapods within the pancrustacean clade. Among them are genes involved in stress response, interaction with the environment, metabolism, regulation, and development-related processes. However, in the future, more collembolan and arthropod species should be included in the subsequent analysis to get a better perspective on the molecular mechanisms affected by the positive selection.

Chapter 3 introduces the first genome of a collembolan, *O. cincta*. This chapter gives an overview of the genome content, evolution of gene families, stress response pathways and possible pre-adaptations to metal stress. We discuss expanded and lineage-specific gene families, as well as HGTs in the context of adaptation to soil metal stress. We analyze the association between altered gene expression as a consequence of genetic adaptation to metal stress and lineage-specific/expanded/HGTs gene families. We suggest several genome features (i.e., expansion of cytochromes P450) that may be pre-adaptations to stress and that could enable these animals to respond to high metal exposure by evolving tolerance. We also hypothesize that HGTs could be advantageous for food processing in a soil environment.

In Chapter 4 we present the high-quality reference genome of *F. candida*. An overview of the genome content and evolution of gene families (expansions, lineage-specific families, HGTs) is described in the context of adaptation to life in the soil. Several of the expanded gene families were linked to stress resistance (de Boer *et al.* 2015) and a metal-tolerant phenotype isolated from a French natural *F. candida* population described in an earlier paper (Nota *et al.* 2013). We also observe that the set of HGTs discovered in the genome is enriched for carbohydrate metabolism. Furthermore, in order to understand better how these species evolve we discuss the architecture of *F. candida's Hox* genes cluster, which, in contrast to *O. cincta* is not arranged according to the ancestral arthropod collinearity pattern. This pattern

has most likely evolved as a result of a translocation event placing the *Hox* genes responsible for thoracic patterning in front of the ones that pattern the head. Additionally, we suggest that substantial intragenomic collinearity in the genome reflects the genetic consequences of the animal's parthenogenetic (clonal) lifestyle. Surprisingly, we observe palindromic organization of 13 gene clusters, suggesting that recombination events are counteracted due to inconsistent sister chromosome pairing caused by these palindromes.

In general, in Chapter 2, 3 and 4 we have shed more light on the question of what makes collembolans "collembolans" and how they evolve and adapt to their environmental niche. This research is a foundation for further investigation of these model organisms and their adaptations to soil life.

5.1. Comparison of collembolan genomes

Although *F. candida* and *O. cincta* belong both to the order Collembola, they have evolved in different ways, suggesting an ancient divergence of the two species. When comparing the two collembolans genomes in detail, we note many differences in the gene content between the two springtails. For example, the number of genes differs, as well as the amount of gene family expansions, the number of HGT events, and most strikingly, the gene order of the *Hox* gene cluster. *F. candida* seems to have undergone rearrangements, breaking and disorganizing the synteny with regard to the insect *Hox* clusters. Still, *Folsomia's* embryonic development does not seem to show important deviations, resulting in a properly patterned hexapod body plan. The temporal-spatial timing of gene expression of the *Hox* gene cluster is still very poorly understood. Future research needs to be done to better understand how the differently organized *Hox* genes cluster functions in *F. candida*. For example, by performing transcriptome sequencing of *Hox* genes in each embryonic stage we could obtain a better understanding of the order in which *Hox* genes in *F. candida* are expressed. Moreover, to understand the function of *Hox* genes in *F. candida* knockdown studies could be performed. These studies were done for many organisms, for example, and recently also for the springtail *O. cincta* (Konopova and Akam 2014). Finally, we could use public databases to identify micro-RNAs. Usually, micro RNAs are associated with transcriptional regulation. A temporally organized expression of micro-RNAs could possibly compensate the disorganized genomic structure of the *Hox* genes cluster and silence body developmental genes until genes that are responsible for head development are expressed (Cameron *et al.* 2006). Such integrated approaches could give a better perspective to understand the disorganized *Hox* genes cluster in *F. candida*.

The substantial genomic rearrangements observed in the two genomes prevents us from defining a "reference collembolan genome" that represents the canonical genome structure of this group of animals. Although Collembola is considered a monophyletic group, some autapomorphies do not support this. For example, the possession of a ventral tube and a jumping organ (furca) that are secondary lost in some species (Hopkin 1997). To conclude, the diversity of the collembolan lineage seems to be too large to select one species as a genomic model representative for the whole clade.

5.2. Contribution to the annotation of springtail genes

Obviously, there is a lack of available sequenced collembolan genomes. In this project, we assembled and annotated two collembolan genomes, however, there are lots of sequences that still remain without annotation (see Chapter 2, 3, 4). Nevertheless, we can incorporate our knowledge on RNA-Seq data (Chapter 2) and stress responses in order to have a better idea on which genes are expressed, their functions and cascades they are involved in. Annotations of collembolan genomes will also provide an essential resource for other genome annotation projects and will allow performing more precise gene families' evolutionary analyses in hexapods in general.

5.3. Perspectives of future use of this research

In general, this work is a foundation for further research on molecular mechanisms in collembolans. Bellow I will describe more specific applications of this research.

5.3.1. Population genomics of stress adaptations

Reference genomes of *Folsomia* and *Orchesella* may form the basis for genome-wide association studies (GWAS) that will allow a better understanding of the mechanisms of adaptation, and the genes that are specifically involved in stress adaptation. GWAS typically focuses on associations between single-nucleotide polymorphisms (SNPs) and phenotypic traits. By resequencing and comparing animal genomes from the clean soil with genomes from animals living at polluted sites we can identify SNPs that may be significantly linked to environmental adaptations. For example, in the threespine stickleback, GWAS was performed to identify salt-responsive *loci* (Wang *et al.* 2014). In *Arabidopsis lyrata*, GWAS was performed to identify adaptations to serpentine soil (high content of heavy metals and low calcium-to-magnesium ratio) and to map SNPs responsible for such adaptations. To that end, DNA pools of individuals from serpentine and non-serpentine soils were pooled and re-sequenced (Turner *et al.* 2010). Sequencing of three candidate *loci* in the European subspecies of *A. lyrata* indicated parallel differentiation of the same polymorphism at one locus, confirming ecological adaptation. Also, the study identified different polymorphisms at two additional loci, which may indicate convergent evolution. Similarly, we can compare reference and metal-tolerant phenotype in *O. cincta* to identify genes associated with tolerance.

5.3.2. Epigenetics in ecotoxicology

Epigenetics studies the way in which cell functions are canalized during development, including the mechanisms that enable different cell lines to develop different functions with the same DNA complement. Moreover, epigenetics considers mitotically or meiotically heritable changes in gene function that occur without a change in the DNA. Such changes can be triggered by environmental factors (e.g., metals, organic pollutants, chemicals). The associated changes in gene expression often lead to modifications of an organism's phenotype and these changes could be sometimes transferred to the next generation (Vandegehuchte and Janssen 2011) This is exemplified by a study with fungicide-exposed rats (Anway *et al.* 2005), and water fleas exposed to the drug 5-azacytidine (Vandegehuchte *et al.* 2010).

It is commonly accepted that epigenetic mechanisms provide an additional layer of genome regulation (Cañestro *et al.* 2007; Foret *et al.* 2009; Mattick *et al.* 2009; Suzuki and Bird 2008). Profiling DNA methylation across the genome is vital to understanding the influence of epigenetics. The assembled collembolan genomes allow genome-wide DNA methylation analysis. Although methylation was identified in *O. cincta* (Roelofs, unpublished data), we observe no DNA methylation in *F. candida* (Noordhoek, J., unpublished data). These results suggest that different transcriptional control mechanisms exist among collembolans. An analysis of multi-generational responses of *O. cincta* exposed to different compounds (and different dosages) could be used to develop tools to quantify hazards of these compounds at the epigenetic level.

Similarly, the effects of environmental chemicals on histone methylation could be analyzed. Histone methylation is a critical process for the regulation of gene expression, which allows different cells to express different genes. Depending on which amino acids in the histones are methylated and how many methyl groups are attached, the transcription could be increased or decreased. For example, methylation events that weaken chemical attractions between histone tails and DNA increase transcription (Mosammaparast and Shi 2010).

5.3.3. Soil quality assessment

An increasing number of new chemicals require fast, effective, automated and cheap screening tests. It is suggested that transcriptomics analysis could improve speed, specificity, and sensitivity of the current ecotoxicological tests as compared to the traditional tests that focus on survival, growth, and reproduction (Van Straalen and Roelofs 2008). In a parallel PhD project, Chen (Chen 2016) developed and validated a new tool for environmental risk assessment. However, Chen suggested genetic biomarkers for environmental testing based only on the partial transcriptome of *F. candida*. The desirable properties of these biomarkers were defined as 1) easily measurable, 2) sensitive, and 3) dose-dependent with respect to specific stress factors. The results on two collembolan transcriptomes (introduced in Chapter 2) and predicted gene models (in Chapter 3 and 4) will improve the newly developed tool for soil quality assessment by incorporating all up-to-date information on gene models and gene expression data. In addition, a better perspective on molecular mechanisms triggered by toxic compounds is now feasible. After incorporation of all genes, we can validate the final set of biomarkers to reveal the level of toxicity and potentially identify the toxic compound. In addition, we will be able to get a completer picture of the cascade of molecular effects and how these are correlated with physiological changes and effects on the level of the whole animal. This will allow the establishment of complete adverse outcome pathways, a concept formulated by Ankley *et al.* (Ankley *et al.* 2010) to causally link toxicological information at different levels of biological organization.

5.3.4. Gene knockdown using RNAi

RNA interference (RNAi) is a biological process in which RNA molecules inhibit gene expression by causing the destruction of specific mRNA molecules. It is an important tool to study gene knockdown, thereby generating specific phenotypes from which the functional role of specific genes can potentially be obtained. The technique was successfully developed and applied to knock down gene activity in different invertebrate species such as *Caenorhabditis elegans* (Fire *et al.* 1998), *Tribolium castaneum* (Posnien *et al.* 2009), *Drosophila melanogaster* (Flockhart *et al.* 2006), as well as the collembolan *O. cincta* (Konopova & Akam, EvoDevo 2014). As such, this technique can be applied to study mutant phenotypes and tolerant phenotypes in collembolans. For example, we can confirm whether previously identified candidate genes (i.e., metallothionein, (Sterenborg and Roelofs 2003)) are actually involved in heavy-metal tolerance. We hypothesize that knocking down such genes will revert genetically adapted heavy metal tolerant animals into non-tolerant/metal sensitive phenotypes. Indeed, RNAi was already applied to *O. cincta* to study *Hox* genes (Konopova and Akam 2014). RNAi knockdown of the *Hox* gene ultrabithorax (*Ubx*) resulted in a mutant phenotype with an extra pair of legs that replaced the ventral tube. It was concluded that ventral tube formation (a unique body feature of collembolans) is partly controlled by *Ubx* and that this structure is homologous to leg development. RNAi has not yet been applied to knock down gene activity in *Folsomia*, although we annotated several genes from its genome that are essential for RNA interference (piwi, argonaute, and dicer, data not shown).

5.4. Advances in NGS

Impressive progress has been made in the field of Next Generation Sequencing (NGS) during the last years. As the field of genomics evolves, there is a growing awareness in the scientific community of the importance of to include long-read data for successful and high-quality genome assembly. Our data (especially the Pacific Bioscience data for *Folsomia*) indeed show that long sequence reads improve and simplify *de novo* assembly. Currently, several sequencing platforms are available, which produce long read sequence data. For example, we used Pacific Biosciences (PacBio) Single Molecule Real-Time (SMRT) sequencing technology, which allows direct sequencing of DNA to achieve sequencing reads with lengths of on average > 10,000 bp, with maximum read length up to around 60,000 bp. At the same time read coverage is uniform with a high consensus accuracy (exceeding 99.9%). Moreover, SMRT sequencing does not require an amplification step, increasing uniformity in coverage even more. This allows sequencing through palindromes and low-diversity regions of the genome. Thus, these long reads allow resolving complex regions, but also variants are called with high accuracy and confidence. Although the raw read error rate can be high (~14%), the inclusion of a stochastic error model in combination with a high coverage rate allowed us to achieve high-quality consensus sequences.

Another emerging technology is Oxford Nanopore sequencing (Nanopore Technologies Ltd, Oxford, UK), a recently commercialized small and low-cost single-molecule nanopore sequencer, called MinION, produces long DNA fragments and allows sequencing small genomes in a matter of seconds. The basic principle behind this technology is tunneling of (polymer) molecules through a pore that separates two compartments. The molecule that is passing through the pore causes a temporary change in the potential between the two compartments, which allows for identification of the specific base (Ashkenasy *et al.* 2005).

The error rate of the MinION technique (after base calling) is estimated to be around 38.2%, mean and median read lengths are 2 kbp and 1 kbp respectively, and the longest single read is about 98 kbp (Laver *et al.* 2015). Despite the high error rate, this technology has the potential to revolutionize genomic applications due to its portability, low-cost, real-time analysis and ease of use compared with existing long read sequencing technologies.

The ultimate sequencing platform would work on single DNA or RNA molecules without any (pre-)amplification, without the use of optical steps, generating reads up to Mb to Gb in length, without GC bias, and with high read accuracy. At the same time, such platform should be flexible enough to generate as many sequence reads as are necessary for the specific research question at hand (Buermans and Den Dunnen 2014). In the future, these emerging technologies will reduce, simplify, and automate the post-processing of sequence data and assembly of genomes. We have shown that the current sequencing platforms are already highly instrumental in the ecological research and it is expected that this trend towards integration of high-end genomic data analyses will exponentially increase in fundamental- and applied ecological research.

5.5. Conclusions

This thesis contributes to a wide range of open questions on evolutionary ecological genomics of two collembolans *O. cincta* and *F. candida*.

Firstly, we present a high-quality *de novo* assembly and annotation of two collembolan transcriptomes (Chapter 2). This study provides clues towards the toolkit used to adapt to terrestrialization. The comparative analysis of two springtails to crustaceans and insects revealed functional categories that likely were under selective pressure during divergence of Collembola and subsequent evolution of Hexapoda. Among them are genes involved in stress response against pathogens and toxic compounds; genes involved in interaction with the environment (i.e., osmotic pressure); genes involved in metabolism and more specifically energy metabolism; regulation and development-related processes. It is interesting that many of these genes refer to rather basic metabolic processes. So, terrestrialization of the hexapods is not something that required very specific and easily delineated changes, but called for thorough changes in the whole basal metabolism. However, we could miss several relevant biological processes and molecular functions due to lack of gene annotation, which could hamper clear conclusions.

Secondly, we present the first nuclear and mitochondrial genome of *O. cincta* (Chapter 3). The comparative analysis of gene families suggests that some expanded. Lineage-specific *O. cincta* gene clusters are involved in xenobiotic biotransformation pathways and biotic responses, which may be related to the occupation of the litter layer. In addition, we propose that HGTs in *O. cincta* and other species of springtails could be beneficial for life in an environment rich in decaying organic matter since so many of them are involved in carbohydrate metabolism. Our analysis also suggests that HGT, the evolution of novel genes, and gene duplication could be considered preadaptations facilitating the evolution of stress tolerance in populations living at metal-contaminated sites.

Thirdly, we present nuclear and mitochondrial genome of *F. candida* and the first genome of its endosymbiont, *Wolbachia* (Chapter 4). We suggest that expansion of gene families linked to metal stress and metal tolerance could be a beneficial preadaptation to deal with metals. As

well as for *O. cincta*, we propose that HGT involved in carbohydrate metabolism (specifically in cell wall degradation) could be beneficial for life in the soil since they may allow assessing to plant cell wall degradation products. Another interesting feature of *F. candida* is that the *Hox* gene cluster shows unexpected rearrangements when compared to the ancestral arthropod *Hox* gene cluster. It is not clear whether such rearrangements are associated with adaptation to life in the soil. The intragenomic distribution of collinear gene clusters in palindromes, observed in *F. candida*, may be explained by its parthenogenetic lifestyle, and counteract meiotic recombination events.

Finally, for both genomes, we developed a genome browser and made all genomic and transcriptomic information visual and available for other researchers (www.collembolomics.nl).

In conclusion, this work builds a solid foundation for further comparative genomics of springtails and yields new insights into the evolution of Collembola. This study also provides a foundation for further analysis of evolutionary mechanisms of adaptation to environmental stressors.

5.6. References

Ankley GT, *et al.* 2010. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environmental Toxicology and Chemistry 29: 730-741.

Anway MD, Cupp AS, Uzumcu M, Skinner MK 2005. Epigenetic transgenerational actions of endocrine disruptors and male fertility. Science 308: 1466-1469.

Ashkenasy N, Sánchez-Quesada J, Bayley H, Ghadiri MR 2005. Recognizing a single base in an individual DNA strand: a step toward DNA sequencing in nanopores. Angewandte Chemie 117: 1425-1428.

Buermans H, Den Dunnen J 2014. Next generation sequencing technology: advances and applications. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 1842: 1932-1941.

Cameron RA, *et al.* 2006. Unusual gene order and organization of the sea urchin *Hox* cluster. JOURNAL OF EXPERIMENTAL ZOOLOGY PART B MOLECULAR AND DEVELOPMENTAL EVOLUTION 306: 45.

Cañestro C, Yokoi H, Postlethwait JH 2007. Evolutionary developmental biology and genomics. Nature Reviews Genetics 8: 932-942.

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

de Boer TE, Janssens TK, Legler J, van Straalen NM, Roelofs D 2015. Combined Transcriptomics Analysis for Classification of Adverse Effects As a Potential End Point in Effect Based Screening. Environmental science & technology 49: 14274-14281.

Fire A, *et al.* 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391: 806-811.

Flockhart I, *et al.* 2006. FlyRNAi: the Drosophila RNAi screening center database. Nucleic acids research 34: D489-D494.

Foret S, Kucharski R, Pittelkow Y, Lockett GA, Maleszka R 2009. Epigenetic regulation of the honey bee transcriptome: unravelling the nature of methylated genes. BMC Genomics 10: 1.

Hopkin SP. 1997. Biology of the springtails:(Insecta: Collembola): OUP Oxford.

Konopova B, Akam M 2014. The *Hox* genes Ultrabithorax and abdominal-A specify three different types of abdominal appendage in the springtail *Orchesella cincta* (Collembola). EvoDevo 5: 1.

Laver T, *et al.* 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomolecular detection and quantification 3: 1-8.

Mattick JS, Amaral PP, Dinger ME, Mercer TR, Mehler MF 2009. RNA regulation of epigenetic processes. Bioessays 31: 51-59.

Mosammaparast N, Shi Y 2010. Reversal of histone methylation: biochemical and molecular mechanisms of histone demethylases. Annual review of biochemistry 79: 155-179.

Nota B, de Korte M, Ylstra B, van Straalen NM, Roelofs D 2013. Genetic variation in parthenogenetic collembolans is associated with differences in fitness and cadmium-induced transcriptome responses. Environmental science & technology 47: 1155-1162.

Posnien N, *et al.* 2009. RNAi in the red flour beetle (Tribolium). Cold Spring Harb Protoc 2009: pdb prot5256.

Sterenborg I, Roelofs D 2003. Field-selected cadmium tolerance in the springtail *Orchesella cincta* is correlated with increased metallothionein mRNA expression. Insect Biochemistry and Molecular Biology 33: 741-747.

Suzuki MM, Bird A 2008. DNA methylation landscapes: provocative insights from epigenomics. Nature Reviews Genetics 9: 465-476.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV 2010. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. Nature genetics 42: 260-263.

Van Straalen NM, Roelofs D 2008. Genomics technology for assessing soil pollution. Journal of Biology 7: 1.

Vandegehuchte MB, Janssen CR 2011. Epigenetics and its implications for ecotoxicology. Ecotoxicology 20: 607-624.

Vandegehuchte MB, Lemière F, Vanhaecke L, Berghe WV, Janssen CR 2010. Direct and transgenerational impact on Daphnia magna of chemicals with a known effect on DNA methylation. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology 151: 278-285.

Wang G, *et al.* 2014. Gene expression responses of threespine stickleback to salinity: implications for salt-sensitive hypertension. Frontiers in genetics 5: 312.