Summary

Soil invertebrate communities are remarkably diverse and essential for soil ecosystem health. Collembola are wingless hexapods that support soil mineralization by feeding on and processing dead organic matter (detritus). However, virtually nothing is known about the genomic make-up of these organisms, despite the fact that they are intensively studied in the context of ecotoxicology and soil ecology. This thesis focuses on two collembolan models: *Folsomia candida* and *Orchesella cincta*.

The research aims were to assemble and annotate the transcriptomes and genomes of *F. candida* and *O. cincta*, to characterize genome content, to investigate how collembolans have evolved, and to better understand specific features and possible adaptations and pre-adaptations related to soil life and environmental stress.

In Chapter 2 we assembled and annotated transcriptomes of *F. candida* and *O. cincta*, and performed a comparative analysis with protein-coding gene sequences of three crustaceans together with three insects to identify adaptive signatures associated with the evolution of hexapods within the pancrustacean clade. Subsequently, we predicted orthologous clusters among eight species and applied the branch-site test to detect episodic positive selection in the Hexapoda and Collembola lineages. A subset of 250 genes showed significant positive selection along the Hexapoda branch and 57 in the Collembola lineage. Gene Ontology categories enriched in these genes include metabolism, stress response (i.e. DNA repair, immune response), ion transport, ATP metabolism, regulation and development-related processes (i.e. eye development, neurological development). We suggest that the identified gene families represent processes that have played a key role in the divergence of hexapods within the pancrustacean clade that eventually evolved into the most species-rich group of all animals, the hexapods. Furthermore, we suggest that some adaptive signatures in collembolans may provide valuable clues to the understanding of the evolution of hexapods on land.

In Chapter 3 we present a draft genome of an *O. cincta* reference strain with an estimated size of 286.8 Mbp, containing 20,249 genes. In total, 446 gene families are expanded and 1,169 gene families evolved specific to this lineage. Besides these gene families involved in general biological processes, we observed the expansion of gene clusters participating in xenobiotic biotransformation. Furthermore, we identified 253 cases of horizontal gene transfer. Although the largest percentage of them originated from bacteria (37.5%), we observed an unusually high percentage (30.4%) of such genes of fungal origin. The majority of foreign genes are involved in carbohydrate metabolism and cellulose degradation. Moreover, some foreign genes (e.g. bacillopeptidases) expanded after horizontal gene transfer. We hypothesize that horizontally transferred genes could be advantageous for food processing in a soil environment that is full of decaying organic material. Finally, we identified several lineage-specific genes, expanded gene families, and horizontally transferred genes, associated with altered gene expression as a consequence of genetic adaptation to metal stress. We suggest that these genome features may be pre-adaptations allowing natural selection to act on. This genome study provides a solid foundation for further analysis of evolutionary mechanisms of adaptation to environmental stressors.

In Chapter 4 we describe a reference genome for *F. candida*, which is 221.7 Mbp in size and contains 28,734 genes. The assembly comprises 162 scaffolds with an N50 of 6.5 Mbp and a maximum sequence length of 28.5 Mbp. To reveal specific genomics features linked to soil living, we studied gene family expansions, lineage-specific gene families, genes acquired by

horizontal gene transfer (HGT), and *Hox* genes related to the typical soil-adaptive morphology of *F. candida*. We found 368 gene family expansions, 74 lineage-specific gene clusters, as well as 10,080 genes that could not be assigned to gene families. Specific gene families are linked to stress resistance and, in particular, to a metal-tolerant phenotype. Furthermore, we identified 809 (2.8%) cases of HGT. The main sources of HGT are bacterial (39.9%), followed by fungal and protists donors. The majority of HGTs are involved in carbohydrate metabolism, a pattern that was also observed in *O. cincta*. *F. candida* contains 11 *Hox* genes within a single cluster. However, the gene order is disorganized with respect to the ancestral arthropod synteny. Additionally, we identified extensive regions of colinearity between some genomic scaffolds of *F. candida*. Finally, we were able to assemble the genome of *Folsomia's* endosymbiont, *Wolbachia*, which turned out to be significantly larger than any *Wolbachia* genome investigated to date.

The expansion of gene families linked to stress responses suggests that stress defense is an important characteristic allowing colonization of soil. Also, the relatively large number of HGT genes related to carbohydrate metabolism, in particular, lignocellulose degradation, could be beneficial in soil, which is full of cell wall degradation products. The colinearity analysis of the *F. candida* genome suggests some features that could be related to parthenogenesis, a common mode of reproduction among soil-living hexapods. In *F. candida* parthenogenetic reproduction is most likely imposed by *Wolbachia*. Moreover, the very large *Wolbachia* genome is remarkable. In most cases, endosymbiotic genomes are significantly reduced due to gene loss and horizontal gene transfer of genes to the host genome. The genome of *F. candida* provides an essential resource for further research on this model organism and its adaptations to soil life.

This PhD thesis builds a solid foundation for further comparative genomics of springtails and provides new insights into the evolution of Collembola. This study also provides a foundation for further analysis of evolutionary mechanisms of adaptation to environmental stressors and to life in the soil. Finally, for both genomes, we developed a genome browser and made all genomic and transcriptomic information available for other researchers (www.collembolomics.nl).