Ghent University
Faculty of Sciences

Genome annotation and evolution of chemosensory receptors in spider mites

---

Cao Thi Ngoc Phuong

Promoter: Prof. Dr. Yves Van de Peer

Department of Plant Biotechnology and Bioinformatics
VIB Department of Plant Systems Biology
Bioinformatics & Systems Biology
Technologiepark 927, B-9052 Gent
BELGIUM

Dissertation submitted in fulfillment of the requirement for the degree
Doctor (PhD) in Sciences, Bioinformatics
Academic year 2013-2014

Examination Committee

Prof. Yves Van de Peer (Promoter): Ghent university

Prof. Geert De Jaeger (Chair): Ghent university

Dr. Pierre Rouzé (reading committee): Ghent university

Dr. Pieter De Bleser (reading committee): Ghent university

Prof. Miodrag Grbic (reading committee): Western University, Canada

Prof. Thomas Van Leeuwen (reading committee): University of Amsterdam

Prof. Kathleen Marchal: Ghent university

Acknowledgements

I would like to thank my promoter, Yves Van de Peer for providing me the opportunity to work as PhD student in your lab, and for guiding, supporting and encouraging me during more than 5 years.

I would like to thank my thesis committee members, for reading my thesis and evaluating my work. I greatly appreciate your guidance and comments.

I would next like to thank all my colleagues in the Bioinformatics group. Stephane has helped me since my first day in Ghent, Belgium. I am very grateful to Pierre for his scientific advice and knowledge and many insightful discussions and suggestions. I had the pleasure time with my friends – Vanessa, Yao-Cheng, Bing, Sara, Ying, Anagha, Evangelia, Jeffrey, Yao Yao, Elisabeth, Cindy…

Especially, I would like to thank my family (my parents, my older sister, and my younger brother), my parent-in-law, my husband and my son for the love, support, and constant encouragement. I undoubtedly could not have done my PhD road without you. I also thank my Vietnamese friends (Giang+Frank, Ut+Tho, Duong+Hieu, Thong, Anh Phong+Chi Tuyen, Be Viet Anh, Be Thao…) for providing support and friendship throughout my time in Ghent.

Ghent, June 2014

Cao Thi Ngoc Phuong

Table of Contents

List of Figures

List of Tables

Summary

Understanding the evolution of species and speciation, the mechanism producing the diversity of life on Earth, has always fascinated scientists. In recent years, advances in next generation sequencing techniques, together with the development of data analyzing software tools, allow us to sequence and analyze genomes of many species and reconstruct their evolutionary history. We can detect the evolutionary changes of a group of species or of different populations of a single species. In this thesis, we perform studies on three spider mite genomes, *Tetranychus urticae*, *Tetranychus evansi* and *Tetranychus lintearius*. The spider mites belong to the Chelicerata, the second largest group of arthropods after insects. While many insect genomes were sequenced and analyzed already, *Tetranychus urticae* represents the first complete chelicerate genome.

This thesis has been organized into five chapters.

The introductory Chapter 1 provides an overview of the explosion of genome sequences in times of the fast development of next generation sequencing techniques, describes genome annotation information, methods and pipelines to give biological meaning to these genomes, and explains the importance of genome based research for the evolution of arthropod-plant interactions. In addition, a short overview of the chemosensory receptors is provided since in the thesis we have particularly studied the annotation and evolution of this gene family in three different spider mites. Chapter 2 provides the results of annotation and analysis of the *Tetranychus urticae* genome (London strain). *T. urticae* represents one of the most polyphagous arthropod herbivores, feeding on more than 1,100 plant species including species known to produce toxic compounds. We have annotated the *T. urticae* genome with support of RNA-seq data and made it publicly available to the research community. The *T. urticae* genome sequence reveals herbivorous pest adaptations with strong signatures of polyphagy and detoxification in gene families associated with feeding on different hosts and in new gene families acquired by lateral gene transfer. Moreover, how this pest responds to a changing host environment is shown

by deep transcriptome analysis of *T. urticae* feeding on different plants. Thus, the *T. urticae* genome sequence opens up new avenues for understanding the evolution of arthropods as well as the fundamentals of plant–herbivore interactions.

The next two chapters (Chapter 3 and Chapter 4) present studies on the annotation and evolution of chemosensory receptors (CRs) in three different spider mites. Chemosensory receptors help animals to detect certain chemical components in their environment to find food, to locate shelter, mates and offspring, and to avoid danger. In Chapter 3, starting from *Daphnia* and insect chemosensory receptors, we describe mining the *T. urticae* genome for putative chemosensory receptors, including the ones related to insect gustatory receptors (GRs), the ionotropic receptors (IRs) and the epithelial Na+ channels (ENaCs). *T. urticae* has a huge repertoire of GRs, many more than the total number of GRs and odorant receptors (ORs) found to date in any other arthropod. Similar to *Daphnia pulex*, we observed the complete lack of ORs in *T. urticae*. This is consistent with the hypothesis that ORs are an insect-specific class of GR-related chemosensory receptors. Futhermore, we compare chemosensory receptor genes among three strains (London, Montpellier, and EtoxR). We find that GR genes that are intact in some *T. urticae* populations appeared to be inactived in other populations. Next, in Chapter 4, we describe the annotation of GR genes in *T. evansi* and *T. lintearius*, and the evolutionary analysis of this gene family in the three spider mites. We identify many GR gene expansions in the polyphagous *T. urticae*, a few gene expansions and many gene losses in the oligophagous *T. evansi*, and no gene expansion but also many gene losses in the monophagous *T. lintearius*. Finally, general remarks are discussed in the Chapter 5.

Chapter 1

Introduction

1.1. The explosion of genome projects provides deeper insight into the evolution of organisms

Life on Earth has been developing for many billions of years under changing environment conditions, from the first prokaryotes with simple cell structures to the first eukaryotes and higher organisms such as the plants, animals and human beings. Each organism is specified by its genome, containing the biological information to build and maintain the life of that organism. Therefore the evolution of the genome over time forms the foundation of the complexity of life on Earth. Recently, due to revolutions in DNA sequencing technologies, more and more genomes are being sequenced and analyzed. This not only allows scientists to study the evolutionary history of different species at the molecular level, based on changes in the genome, but also allows deciphering the evolutionary relationship between species, revealing many mysteries.

1.1.1. Evolution is the mechanism producing the diversity of Life

The origin, history and diversity of life on Earth have always been interesting issues for scientists. In the book “Cradle of Life: The Discovery of Earth's Earliest Fossils” [1], J. William Schopf, UCLA paleobiologist, condensed life’s 4.5 billion year history on Earth into a single 24-hour day. If Earth was formed during the first second past midnight, life began around 4:00 A.M. The oldest fossils were entombed at 5:30 in the morning. Then early-evolving plant like microbes chemically joined with oceanic iron to form rusty sediments that accumulated slowly at about 2:00 in the afternoon. Floating single-celled algae with cell nuclei and chromosomes appeared by 6:00 P.M. Larger multicellular seaweeds entered Earth at about 8:30 in the afternoon, and a few minutes later did jellyfish and worms. At about 9:00 at night, larger organisms appeared. To date, organisms from

small bacteria to the largest land animals are found everywhere, from the darkest depths of the ocean/buried kilometers deep in the Earth's crust to the highest mountains, from the hottest volcanic mud to the frozen surface of the Antarctic… Camilo More [2] estimated that there are ~8.7 million eukaryotic species in the range of 3 to 100 million species in the world [3], containing 2.21 million species in the oceans, including ~7.77 million species of animals, ~298,000 species of plants, ~611,000 species of fungi, ~36,400 species of protozoa, and ~27,500 species of chromists. Known species form only a small fraction on land (~14%) and in the ocean (~9%). Since ancient times, scientists have pursued the ideal to systematically classify all living organisms. In 1758, Carolus Linnaeus, as the father of biological classification, published his "Systema naturae". By comparing species' anatomy, he described the formal classification system of species showing the relationship between organisms. In 1859, Charles Darwin, with the publication of "the origin of species" proposed that all organisms on Earth evolved from a common ancestor by natural selection. Organisms that are well adapted to their environment have better chances of survival and reproduction. To emphasize the importance of Darwin's work, American philosopher Daniel Dennett in his book "Darwin's Dangerous Idea" claimed that "If I were to give an award for the single best idea anyone has ever had, I'd give it to Darwin… In a single stroke, the idea of evolution by natural selection unifies the realm of life, meaning, and purpose with the realm of space and time, cause and effect, mechanism and physical law". Evolution is thus the mechanism producing the diversity of life [4]. The discovery of DNA's structure (1953) [5], the first DNA sequencing (1968) [6], as well as the central dogma of molecular biology [7, 8] have opened the era of molecular biology and allowed to research evolution at the molecular level. The intellectual concepts from the study of evolution have changed many other fields of study because of Dobzhansky's famous statement "nothing in biology makes sense except in the light of evolution" [9].

1.1.2. Advances of DNA sequencing methods

In 1977, two DNA sequencing methods were published: the chain termination method of Sanger [10] in which DNA polymerase is used to synthesize the new chain based on the template sequence and 2', 3'-dideoxy and arabinonucleoside analogues of the normal deoxynucleoside triphosphates acting as specific chain terminating inhibitors of DNA

polymerase; and the chemical degradation method of Maxam and Gilbert [11] in which chemical agents break a terminally labeled DNA sequence at adenine, guanine, cytosine or thymine and the lengths of the labeled fragments then identify the positions of that base. From then to now, the Sanger method has played a vital role in sequencing the genomes of many model organisms. The Sanger method was used to sequence the lambda phage genome (48.5kb) in 1982. In 1995, two complete genome sequences of bacterial species, namely those of *Haemophilus influenzae* (1.83Mb) [12] and *Mycoplasma genitalium* [13], were reported by Craig Venter's group at TIGR, in which the *M. genitalium* genome contains the minimal set of genes required for cellular life. Since then, many model organisms were sequenced. For instance, *Escherichia coli* [14] and several other strains were sequenced to obtain information about bacterial evolution and pathogenicity [15, 16]. *Saccharomyces cerevisiae* (12Mb) was the first eukaryotic organism to be sequenced [17]. The *Bacillus subtilis* genome (4.2Mb), the best-characterized member of the Gram-positive bacteria, was published in 1997 [18]. Next, the first animal genome, *Caenorhabditis elegans*, was reported in Science in 1998 [19]. Because DNA sequencing was truly automated in 1996 with the first commercial DNA sequencer that used capillary electrophoresis rather than a slab gel, other model organisms with more complex and larger genome sequences were sequenced such as *Drosophila melanogaster* genome [20], *Arabidopsis thaliana* [21] and of course the human genome [22, 23].

DNA sequencing technologies have been developed for the last 40 years, mostly relying on versions of the Sanger dideoxy terminator sequencing. In 1986, the first automated DNA sequencing machine collecting and storing sequencing data directly to a computer without autoradiography of the sequencing gel was published by the laboratory of Leroy Hood at Caltech in collaboration with Applied Biosystems (ABI). The 3730XL generation machine, provided by ABI, is the newest machine based on the Sanger method. This method can be applied to achieve read-lengths of up to ~1000bp and per-base 'raw' accuracies as high as 99.999%, with sequencing costs at about $0.50 per kilobase. With increasingly large and complex genomes, cheaper and faster sequencing methods have been required [24]. Next generation sequencing methods, therefore, have provided

sequencing throughput at a low price in a short amount of time with accepted accuracy [25].

1.1.2.1. Next generation sequencing platforms

The next generation sequencing (NGS) technologies commercially available today include the 454 pyrosequencing from 454 Life Sciences, Illumina Genome Analyzer, SoLid from ABI and Heliscope from Helicos [26, 27].

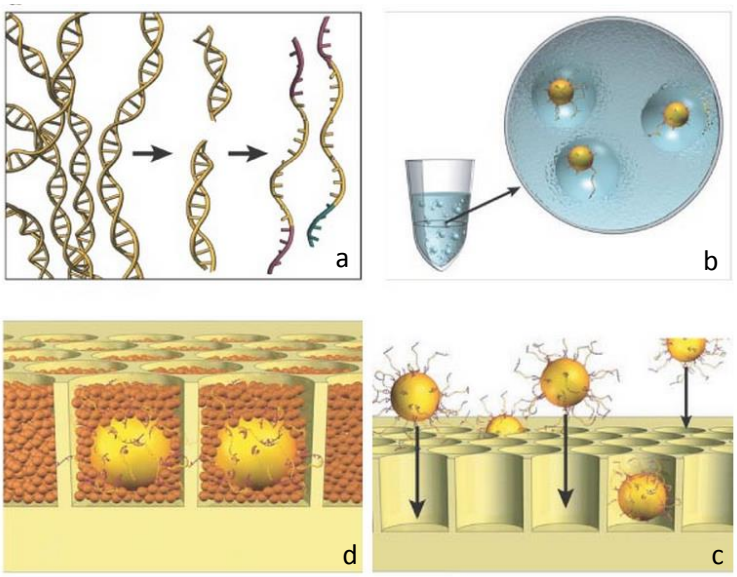


Figure 1.1. The method used by the Roche/454 sequencer (taken from [28]). (a) Genomic DNA is isolated, fragmented, ligated to adapters and separated into single strands. (b) Fragments are bound to beads under conditions that favour one fragment per bead. The beads are captured in the droplets of a PCR-reaction-mixture-in-oil emulsion. (c) Beads carrying single-stranded DNA clones are deposited into wells of a fibre-optic slide. (d) Smaller beads carrying immobilized enzymes required for pyrophosphate sequencing are deposited into each well.

In 2005, the Roche GS FLX sequencing system [28], developed by 454 Life Sciences, was the first next generation sequencing system on the market based on the pyrosequencing method [29]. The sequencing process includes three steps [28] (Figure 1.1): DNA library preparation, emulsion PCR for DNA amplification, and pyrosequencing. In the DNA library preparation step, the fragments created by shearing DNA randomly are ligated to specialized common adaptors and separated into single

strands. Then each fragment is captured on an agarose bead whose surface carries oligonucleotide complementary to the adapter sequence. The fragment and bead complexes are isolated into individual oil:water micelles that also contain reactants for emulsion PCR, producing about one million copies of each DNA fragment on the surface of each bead. In pyrosequencing, on the picotiter plate, one of dNTPs (dATP, dGTP, dCTP, dTTP) will complement to the base of the template strand and release a pyrophosphate. The ATP transformed from pyrophosphate by ATP sulfurylase drives the luciferin into oxyluciferin and generates visible light, which is detected by a charge-coupled device imaging system. At the same time, unmatched dNTPs are degraded by apyrase, and next cycle, another dNTPs are added into the system.

The Solexa technology first appeared in 2008 with the purpose of resequencing to create shorter reads compared to a reference genome to identify intraspecies genetic variation [30]. This method generates several billion bases of accurate nucleotide sequence per experiment at low cost and also contains 3 steps: sequencing library preparation, solid support amplification and sequencing using fluorophore labeled reversible terminator nucleotides. DNA samples are sheared randomly into DNA fragments. Then these fragments are joined to a pair of oligonucleotides in a forked adaptor configuration. After ligation, the fragments are amplified using two oligonucleotide primers, resulting in double-stranded blunt-ended DNA molecules with a different adaptor sequence on either end (Figure 1.2a). Next, the fragments are denatured and single strands are annealed to complementary oligonucleotides on the flow-cell surface. A new strand is copied from the original strand and then the original strand is removed by denaturation. The adaptor sequence at the 3' end of each copied strand is annealed to a new surface bound complementary oligonucleotide, forming a bridge and generating a new site for synthesis of a second strand. Many cycles are repeated, creating DNA "clusters' from each fragment (Figure 1.2b). The DNA in each cluster is linearized by cleavage within one adaptor sequence and denatured, generating single-stranded template for sequencing by synthesis to obtain a sequence read (read 1). To perform paired-read sequencing, the products of read 1 are removed by denaturation, the template is used to generate a bridge, the second strand is re-synthesized and the opposite strand is then cleaved to provide the template for the second reads (Figure 1.2c). Each sequencing cycle includes the

simultaneous addition of a mixture of four fluorescent labels and a reversibly terminating moiety at the 3' hydroxyl position, extension of primed sequencing features followed by imaging in the four channels, and cleavage of both the fluorescent labels and the terminating moiety.

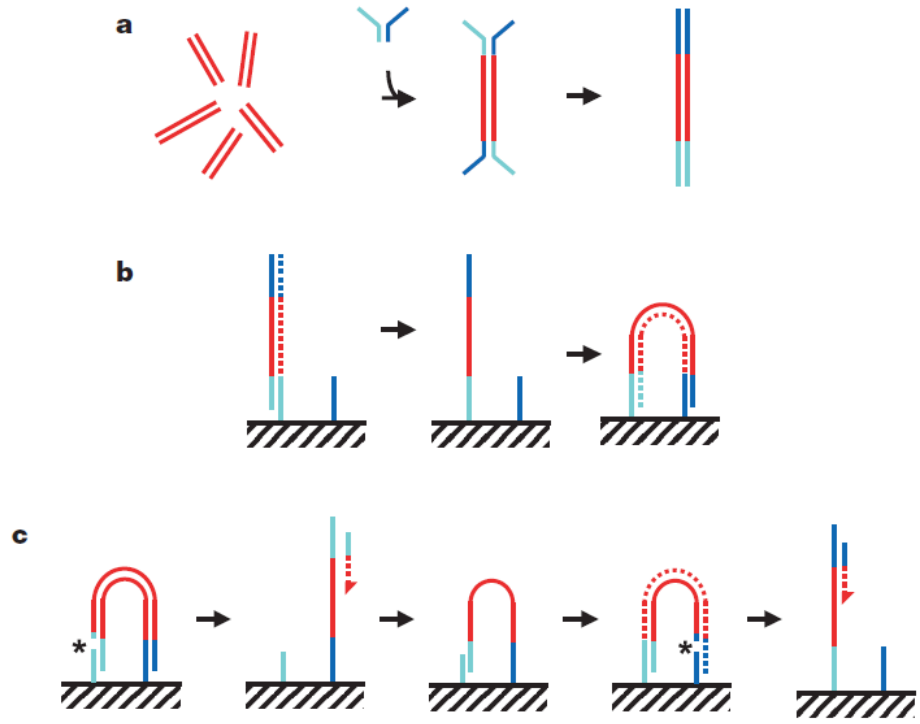


Figure 1.2. Preparation of samples for the Illumina sequencing (taken from [30]). (a) Preparing the double-stranded blunt-ended DNA molecules. (b) Amplifying and creating DNA "clusters' from each molecule. (c) Sequencing paired reads.

The AB Solid method, initially applied to sequence a bacterial genome [31], uses an adapter-ligated fragment library similar to those of the other next-generation methods, and an emulsion PCR approach with small magnetic beads (Figure 1.3). Unlike the other methods, Solid uses a different approach with DNA ligase to sequence the amplified fragments. Each sequence cycle introduces a partially degenerate population of fluorescently labeled octamers. After ligation and imaging in four channels, the labeled portion of the octamer is cleaved via a modified linkage between bases 5 and 6, leaving a free end for another cycle of ligation.

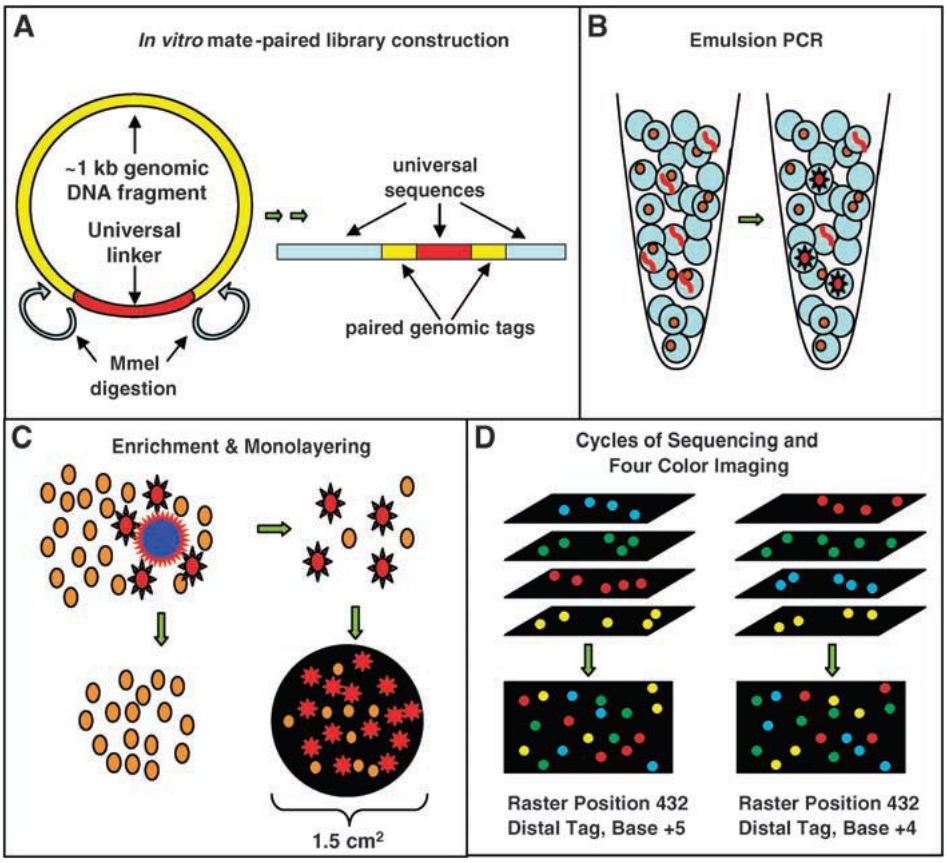


Figure 1.3. The method used by the AB Solid sequencer (taken from [31]). (A). Sheared, size-selected genomic fragments (yellow) are circularized with a linker (red) bearing Mme I recognition sites. Subsequent steps, which include a rolling circle amplification, yield the 134- to 136-bp mate-paired library molecules shown at right. (B) ePCR yields clonal template amplification on 1-µm beads. (C) Hybridization to nonmagnetic, low-density "capture beads" (dark blue) permits enrichment of the amplified fraction (red) of magnetic ePCR beads by centrifugation. Beads are immobilized and mounted in a flow cell for automated sequencing. (D) At each sequencing cycle, four-color imaging is performed across several hundred raster positions to determine the sequence of each amplified bead at a specific position in one of the tags.

The Helicos sequencer, based on the method from Quake and colleagues [32], sequences the DNA samples without amplification. Instead, a highly sensitive fluorescence detection system is used to directly interrogate single DNA molecules via sequencing by synthesis (Figure 1.4). Poly-A-tailed DNA fragments, labeled with Cy3, are attached to the flow cell surface-tethered poly-T oligomers. At each cycle, DNA polymerase and a

single species of fluorescently labeled nucleotide are added, resulting in template-dependent extension. After fluorescence imaging of the full array, chemical cleavage and release of the fluorescent label permits the next cycle. Several hundred cycles of single-base extension create average read-lengths of 25bp or greater.

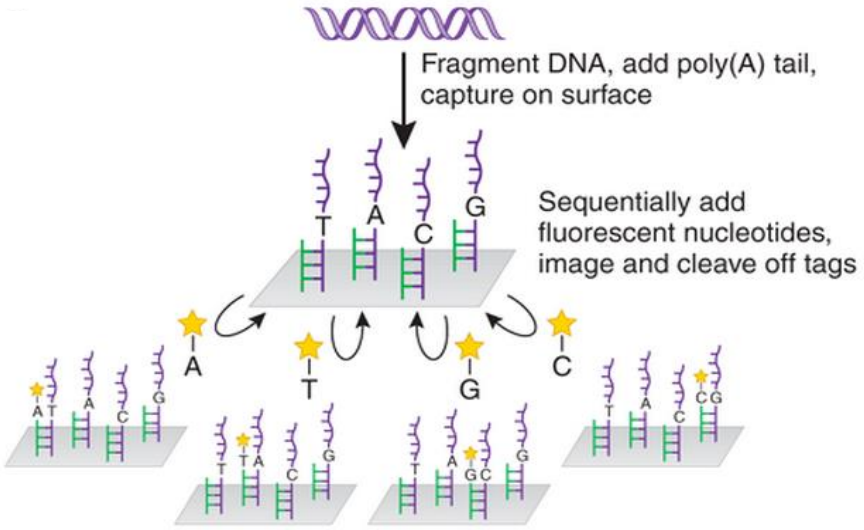


Figure 1.4. Single-molecule sequencing as implemented by the Helicos sequencer (taken from [33]).

All technologies listed above are called second generation sequencing methods, while the Sanger method is referred to as first generation sequencing technology. Both approaches vary significantly with regard to their throughput, read-length, and operating cost. The length of sequence reads from second generation sequencers is much shorter than from first generation sequencers and each read type has a unique error model. The comparison of next generation sequencing with the Sanger method is shown in Table 1.1. Many different bioinformatics tools have been developed for analyzing the output data of these methods. With the high demand for low cost technologies, third generation sequencing (next-next generation sequencing) technologies have been developed [34, 35], including Pacific Biosciences Single Molecule Real Time (SMRT) sequencing [36], Nanopore Sequencing [37], and Ion Torrent sequencing [38].

Table 1.1. Comparison of DNA sequencing methods (taken from [39])

Sequencer	454 GS FLX	HiSeq 2000	SOLiDv4	Sanger 3730xl
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding	Dideoxy chain termination
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp or 50 + 50 bp	400~900 bp
Accuracy	99.9%*	98%, (100PE)	99.94% *raw data	99.999%
Reads	1 M	3 G	1200~1400 M	—
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9~84 Kb
Time/run	24 Hours	3~10 Days	7 Days for SE 14 Days for PE	20 Mins~3 Hours
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput

1.1.2.2. Application of next generation sequencing methods

Next generation sequencing methods have a wide variety of applications such as the 'de novo' whole-genome sequencing, whole-genome/targeted resequencing and transcriptome sequencing for quantification of gene expression and alternative splicing and transcript annotation. In addition, NGS is being used to study the epigenome for interaction between transcription factors and their direct targets, genomic profiles of histone modification, DNA methylation and genomic profiles of nucleosome positions. Finally, NGS is used for metagenomics to study, for example, environmental communities and the human microbiome [40].

Whole-genome sequencing using the Sanger method requires time and resources. Therefore, NGS sequencers with the ability to produce large volumes of data in short of time have become the preferred tools for whole-genome sequencing. At first, genome sequencing projects still utilized a hybrid Sanger/pyrosequencing approach, for example to sequence the genome of *Vitis vinifera* grape, a highly heterozygous and large eukaryote genome [41], and marine microbial genomes [42]. An approximately 32.5Mb draft genome sequence for the forest pathogen *Grosmannia clavigera*, an ascomycete fungus, was sequenced by combining Illumina, 454 and Sanger sequence data [43]. The draft genome sequence of *Cucumis sativus L*. was reported by assembling Sanger and Illumina sequence data [44]. However, Li and colleages [45] claimed to have successfully sequenced and assembled the giant panda genome using only Illumina Genome Analyser

sequencing technology. The assembled contigs (2.24Gb) cover approximately 94% of the whole genome, and the remaining gaps seem to contain carnivore-specific repeats and tandem repeats. This result has demonstrated that the next-generation sequencing technologies can be applied to sequence large eukaryotic genomes for accurate, cost-effective and rapid de novo assemblies.

With the increase of high quality reference genome sequences, whole genomes, genomic regions, or genes in multiple individuals can be resequenced to study generic variation by aligning to the appropriate reference [25]. For examples, 10 *Caenorhabditis elegans* mutation-accumulation lines were resequenced to understand mutation processes [46], 40 silkworm genomes were resequenced to decipher domestication processes [47], and many genome sequences of individual humans were also resequenced by Sanger dideoxy technology, Illumina paired-end sequencing method [48-53]. Because about 2% of the human genome consists of exons, many methods have been developed to capture particularly these regions [54-56] for sequencing with next-generation sequencing technologies and for studies of human diseases such as deafness and other genetic disabilities [57, 58].

The transcriptome is the complete set of transcripts and their quantity in a cell. NGS methods have been applied to sequence cDNAs of a specific developmental stage or physiological condition, resulting in so-called RNA-seq data or RNA-seqs. The RNA-seqs have effectively been used for researching transcriptomics [59-67]. RNA-seq data are also heavily used to improve gene prediction, because ESTs sequenced by the Sanger method detect only about 60% of transcripts in the cell [68], to reveal splicing isoforms of known genes, and to map 5’ and 3’ boundaries of many genes, as well as identifying novel transcribed regions in genomes [69].

In addition, NGS methods have been applied to identify known and novel small RNAs as well. Profiling the small RNAs with a 454 sequencer was reported for organisms such as *Physcomitrella patens* [70], and *Arabidopsis thaliana* [71, 72]. Especially, a novel class of small RNAs, termed Piwi-interacting RNAs, was discovered [73-75]. The research on small RNAs with an Illumina sequencer was also reported in human embryonic stem cells

[76], developing chicken embryo [77], *Gossypium hirsutum L* [78].

The epigenome is the study of heritable gene regulation without altering the DNA sequence, through biochemical modifications such as DNA methylation patterns, post-translational modifications of histone proteins. Bisulfite DNA sequencing [79] was improved by combining with 454 sequencing to analyze DNA methylation patterns in multiple gene promoters of human cells [80]. Combining the bisulfite sequencing with Illumina ultra-high throughput sequencing was performed to measure cytosine methylation at a genome-wide scale of *Arabidopsis thaliana* [81, 82]. This combined approach was reported for other genomes as well [83-85].

Traditionally, the ChIP-chip approach, i.e. chromatin immunoprecipitation (ChIP) followed by microarray analysis, was used to determine the association of proteins with specific genomic sequences *in vivo* [86]. The Chip-seq approach, i.e. ChIP followed by sequencing, has recently been developed for genome-wide scanning of DNA-protein interaction of nucleosomes [87-90].

Metagenomics is the genomic analysis of microorganisms by direct extraction of DNA from an uncultured ensemble of microbial communities. For instance, Edwards and colleagues have used 454 pyrosequencing to generate environmental genome sequences from two sites in the Soudan Mine, Minnesota, United States [91]. Since then, many metagenomics have been studied, including the Human Microbiome project (HMP) [92] in which microbial communities of various parts of the human body have been studied, including the gut [93].

1.1.3. Explosion of genome projects

With the introduction of NGS technologies, many international genome projects, both 'de novo' and resequencing projects have recently been launched. The genome sequence of *Arabidopsis thaliana* became available by Sanger method in 2000 [21], and since then, many other plant genomes were published. NGS techniques have generated genome, transcriptome and epigenome data for many model sequences and important crop species that have permitted deep inferences into plant biology [94] as well as studying the history

of plant domestication to accelerate crop improvement [95]. Medicinal plant genome projects have set up the foundation for the development of natural medicines and the selection of cultivars with good agricultural traits [96]. In addition, with sharply reduced costs of NGS, thousands of *A. thaliana* individuals have been resequencing to decipher genomic differences at the population level [97, 98].

The Genome 10K project has been launched to sequence whole genomes of 10,000 different vertebrate species, including some species that recently became extinct [99]. The main purpose of the project is to study the evolution of vertebrate species derived from a common ancestor that lived between 500 and 600 million years ago (MYA), before the Cambrian explosion of animal life. In the 100K pathogen genome sequencing project (http://100kgenome.vetmed.ucdavis.edu/), next generation sequencing approaches have been applied to sequence important pathogens, the results of which have been used to study increased food security. The i5K project has started in 2011 and aims to sequence the genomes of 5,000 insects to better understand insect biology, to prevent the transmission of human diseases, and to protect important crops or livestock (http://arthropodgenomes.org/wiki/i5K).

The 1,000 Genomes Project is the first project to sequence the genomes of a large number of humans, and to provide a comprehensive resource on human genetic variation [100]. The genomes of 1,092 individuals from 14 different populations were determined using a combination of low-coverage whole-genome and exome sequencing to build a resource to decipher the genetic contribution to diseases [101]. The Genome-Environment-Trait Evidence (GET-Evidence) system has been developed to automatically process personal genomes and to make these publicly available [102, 103]. In January 2014, Illumina announced that an individual's entire genome could be sequenced for $1000 or less [104]. There are 53,262 sequencing projects, of which 6,328 have been completed, in the Genomes OnLine Database (GOLD) (http://www.genomesonline.org/index) in May 2014, versus 11,472 in September 2011 [105] (Figure 1.5).

Conclusively, there has been an explosion of genomes sequenced with commercially available 454 pyrosequencing followed by Illumina, SOLiD, Helicos and now even third

generation sequencing. However, these can be very poor quality genomes because of inherent errors in the sequencing technologies, and the inability of assembly programs to fully address these errors. Therefore the current gold standards need to be applied to distinguish between draft genomes and finished genomes [106]. Followed by genome assembly, genome annotation is the next important step of a genome project.

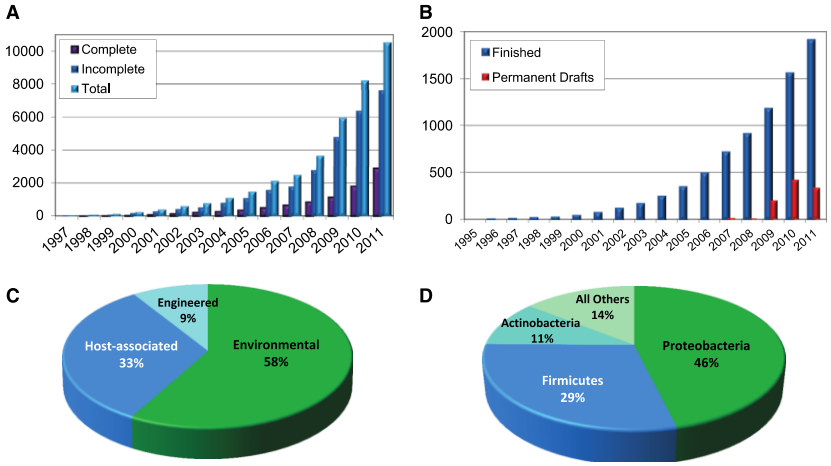


Figure 1.5. Statistical information from GOLD data as of September 2011 (taken from [105]. (A) Evolution of the complete, incomplete and total number of projects monitored in GOLD. Genome projects in GOLD: 11,472. (B) Evolution of the complete projects monitored in GOLD separated into finished and permanent drafts. Complete genome projects in GOLD: 2,907. (C) Distribution of the 340 metagenome projects in GOLD across the three major metagenome classification categories. (D) Phylogenetic distribution of the 8,448 bacterial genome projects.

1.2. Genome annotation

The more DNA sequencing technologies have been developed, the more complete genomic sequences have become available. These raw sequences however only are A, T, G and C strings on the computer or on paper if they are not annotated. Genome annotation is the first and most important step to give biological meaning to the genome. The aim of the annotation is to assign as much information as possible to the raw sequence of genomes with an emphasis on the location and structure of the genes. Genome annotation includes two processes [107]. Structural genome annotation is the process of determining the structure of all of the genes, while functional genome annotation is the process of

identifying their functions. Gene prediction for eukaryotes is more complex than for prokaryotes, and requires not only the identification of the position of the start and stop codons for each gene as with prokaryotes, but also the position of all the gene's introns, which vary tremendously in size and number even within a single species. Besides, prokaryotes have higher gene density than eukaryotes. For example, 85% of the *Haemophilus influenzae* genome is coding sequence, while less than 25% of the fly and worm genome encodes proteins and only about 2% of the human genome contains genes encoding proteins. Although methods for predicting protein coding regions in the genomic DNA sequences have been developed since the 1980's [108], the first real software tools to predict genes only first appeared in the early 1990's. GeneModeler was the first one for eukaryotes [109]. Computational methods, tools and resources for genome annotation, especially for eukaryotes, have been evolving rapidly [110]. Where originally gene prediction programs used only intrinsic features of the genome sequence itself to produce a prediction, it became clear that exploiting extrinsic evidence, lead to a substantial gain in accuracy. However, with more and more sequenced genomes, genome annotation has become more challenging [107].

1.2.1. Gene prediction information

Gene prediction is naturally based on the information on the sequence, which can be divided into intrinsic information and extrinsic information [111]. Prediction based on intrinsic information only uses certain properties of the target sequence itself, while prediction based on extrinsic information includes other external information [112]. Moreover, information on the sequence is also subdivided into two different types: 'signals', and 'contents'. 'Signals' refer to functional sites, i.e. short sequence motifs, specific to a gene, while 'contents' refer to larger regions in the DNA, such as coding sequences (regions) and non-coding regions. Combining intrinsic and extrinsic information has the potential of enhancing the reliability of the results and extracting maximum information from genomic sequences.

1.2.1.1. 'Signals'

As stated previously, gene prediction in prokaryotes is easier than in eukaryotes because of the higher gene density, and the absence of introns. However, genes in prokaryotes

may often overlap with each other and the translation starts can be difficult to predict correctly. Minimum intrinsic 'signals' for predicting prokaryotic genes include the start codon and stop codon of coding sequences, while for eukaryotic genes, the donor and acceptor splice sites for each intron are also important characteristics. Besides, there are some other signals such as branch points, polyadenylation sites, CpG islands, motifs in promoters, such as TATA boxes, transcription factor binding sites, ribosomal binding sites, and terminators.

The stop codons are TGA, TAG or TAA for both prokaryotes and eukaryotes. In eukaryotes, the translation initiation codon is almost always ATG, although translation can begin with a different codon. However, in prokaryotes, there are more non-ATG start codons than in eukaryotes. For instance, in *E. coli*, start codons are ATG, GTG and TTG [14]. Based on the analysis of 211 genes [113] and 699 genes [114], respectively, Kozak showed that (GCC) GCCA/GCCATGG has been the consensus sequence upstream from the translational start site in eukaryotic mRNAs. However, only 0.2% of 2,595 vertebrate mRNA sequences contain precisely this sequence [115]. There is a diversity of nucleotide sequences around the ATG codon in eukaryotes [116].

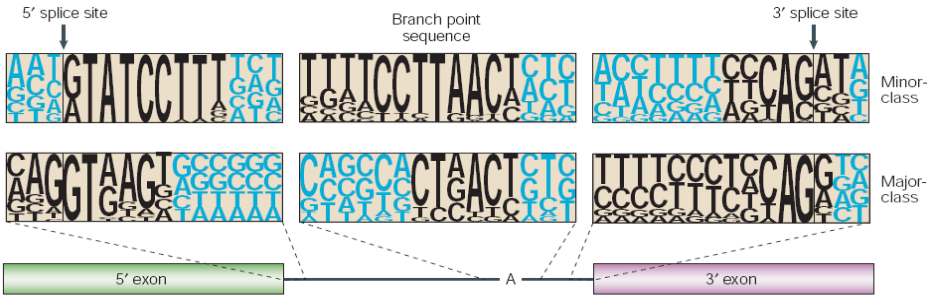


Figure 1.6. Consensus sequences of major-class and minor-class introns [117]. The consensus sequences of the 5′ splice site, branch site and 3′ splice site are shown from left to right for minor-class introns (upper row) and for major-class introns (lower row). The letter heights at each position represent the frequency of occurrence of the corresponding nucleotides at that position. The positions that are thought to be involved in intron recognition are shown in black; other positions are shown in blue. Frequencies were derived from a set of U12-type introns from various plant and animal species and from a set of mammalian U2-type introns (taken from [117]).

Most introns contain the consensus sequences near 5' end (donor splice site) and branch point, which are recognized by spliceosomal components and required for spliceosome formation [117]. Major class introns, i.e. U2-type introns using U2 and analogous U2 snRNPs such as U1, U4, U5 and U6, have the GT donor site and AG acceptor site. Minor class introns, i.e. U12-type introns using U12 snRNP, have the AT donor site and AC acceptor site. While donors and acceptors of U2-type introns are almost never AT-AC, donors and acceptors of U12-type introns can be GT-AG (Figure 1.6). In vertebrates, the frequency of occurrence of U12-type introns is in the range of 0.15-0.34% compared to U2 type introns but is lower in other metazoan taxa.

To discover the 'signals' in a genomic sequence, DNA patterns and consensus pattern search algorithms are applied. However, they cannot represent all information of 'signals' because 'signals' are not universally conserved. Therefore, weight matrices are usually used. In a weight matrix, a score is assigned to each possible nucleotide at each possible position of the signal. The score of a potential signal is defined as the sum of the positional weights of the constituent nucleotides. Kozak developed the first weight matrix for eukaryotic start codons [114]. Weight matrices to identify donor and acceptor sites are much more reliable than a consensus sequence but still predict a large excess of incorrect sites. Start codons are much more difficult to predict than stop codons because start codons can, in principle, be any codon encoded by a methionine in the protein. However, if the start codon is predicted correctly, then the stop codon can easily be predicted. Many approaches applied to predict start codons in eukaryotes and prokaryotes are based on artificial neural networks [118, 119], support vector machines [120], and Gaussian models [121].

1.2.1.2. 'Contents'

So-called 'contents' on genome sequences include coding regions (exons) and non-coding regions (introns, intergenic regions and un-translated regions). In eukaryotes, exons (coding regions) can be classified into four classes: single exons that begin with a start codon and end with a stop codon, initial exons that begin with a start codon and end with a donor site, terminal exons that begin with an acceptor site and end with a termination codon, and internal exons that begin with an acceptor site and end with a

donor. The most obvious indicators of coding and non-coding sequences are: trinucleotide (codon) or hexamer (dicodon) frequencies, compositional bias, codon usage, base occurrence periodicity, and G+C content [122]. Many coding measures measure the 'codingness' of the sequence (reviewed in [108]): codon usage, methods related to the encoded amino acid sequence, base compositional bias between codon positions, imperfect periodicity in base occurrences, and so on. Codon usage is the most fundamental of these coding measures and is based on the effects of unequal usage of codons. It has been widely used in gene prediction, such as frequency counts for the occurrence of successive codon pairs, also called dicodon usage measure or hexamer-0 measure. Initial and terminal exons are difficult to predict by content measures because they are less informative and often much shorter than internal exons.

Extrinsic information of ‘contents’ usually relates to similarity between a genomic sequence region and a protein or a DNA sequence present in a biological database to determine whether the region is coding (exon) or non-coding [123]. The obvious disadvantage of this method is that when no homologues of the new gene are found in the databases, similarity searches will yield little or no useful information. However, with the development of sequencing techniques, we often have access to experimental sequences of the transcriptome (cDNA, ESTs, RNA-seqs) and sometimes of the proteome as well, through mass spectrometry [124]. Mass spectrometry based proteomics approaches directly measure peptides arising from expressed proteins. These peptides can be potentially integrated into the genome annotation process to improve genome annotation quality by verifying protein coding genes, identifying missed protein coding genes, confirming the expression of alternative splice variants in eukaryotic genomes, and correcting stop and start sites and reading frames. Sequences of full transcripts (cDNA) which provide the ideal information for gene modeling are occasionally available. For many years ESTs which are single pass partial sequences of cDNAs have been obtained through Sanger sequencing. The coverage of ESTs sequences, ranging from a few hundreds bp to 1kb, is often rather low, being completely informative for only a fraction of the genes. RNA-seqs are short reads of cDNA (currently 60-100bp with Illumina, several hundreds bp with 454) with allow a better coverage of the transcriptome. Illumina RNA-seqs which are the most common ones nowadays are often inaccurate but offer a

very deep coverage of the transcriptome, depending on individual transcript abundance. RNA-seq data have the great potential to improve the accuracy of gene predictions. RNA-seq data can be assembled *de novo* and used as ESTs, or aligned directly to genome [125, 126]. RNA-seqs are also a great help in identifying alternative splicing [127]. Hitherto, RNA-seqs have proven the existence of hundreds of genes earlier missing in genome annotations [128, 129].

1.2.2. Gene prediction methods

Gene prediction methods are generally divided into two main categories: *ab initio* gene prediction and evidence driven gene prediction [130].

Table 1.2. The lists of *ab initio* gene prediction and evidence-driven gene prediction (taken from [130])

Program	Web page	Evidence
Genscan	http://genes.mit.edu/GENSCAN.html	No
GeneID	http://www1.imim.es/geneid.html	No
SNAP	http://homepage.mac.com/iankorf/	No
GlimmerHMM	http://www.cbcb.umd.edu/software/GlimmerHMM/	No
GeneMark	http://exon.gatech.edu/GeneMark/eukhmm.cgi	No
AUGUSTUS	http://augustus.gobics.de/	ESTs, cDNAs, and proteins
SGP2	http://genome.imim.es/software/sgp2/sgp2.html	TBLASTX hits
GENOMESCAN	http://genes.mit.edu/genomescan.html	BLASTX hits
TWINSCAN	http://mblab.wustl.edu/nscan/submit/	BLASTN hits and ESTs
GENOMINER	http://bl209.caspur.it/Gminer/	Complete genomes
ENSEMBL	http://www.ensembl.org/	ESTs, cDNAs, and proteins
N-SCAN	http://mblab.wustl.edu/nscan/submit/	ESTs, complete genomes
EXOGEAN	http://www.biologie.ens.fr/dyogen/spip.php?rubrique4&lang=en	ESTs, cDNAs, and proteins
GENEWISE	http://www.ebi.ac.uk/Wise2/index.html	Proteins
ASPIC	http://t.caspur.it/ASPIC/	ESTs and cDNAs
Eugene	http://www.inra.fr/mia/T/EuGene/	ESTs, cDNAs, and proteins
GAZE	http://www.sanger.ac.uk/Software/analysis/GAZE/	All available + *ab initio*
JIGSAW	http://www.cbcb.umd.edu/software/jigsaw/	All available + *ab initio*

Ab initio methods only use genomic sequences and apply mathematical models for coding regions and “signals” to identify genes and to determine their exon-intron structures. In

contrast, evidence-driven gene prediction (homology method) performs similarity search procedures between the genome against sequence databases or experimental data including expressed sequence tags (ESTs), full-length complementary DNAs (cDNAs), and even data from microarray hybridization experiments. The evidence-driven gene prediction is usually able to detect only a limited number of genes (low sensitivity) due to the lack of known mRNAs, whereas the gene-level sensitivity of *ab initio* gene prediction can approach 100%, but its accuracy is usually much lower, ~60–70%. Therefore, the trend in recent years is the combination of the two methods to improve the accuracy of gene prediction. Many gene prediction programs have been developed that rely on these methods and are listed in Table 1.2. The evidence-driven gene prediction programs often include *ab initio* gene prediction.

1.2.2.1. *Ab initio* gene prediction

Ab initio gene predictors were first developed for prokaryotic genomes. These programs used different statistical approaches on signal and content information to identify the coding region starting with the ATG codon and ending with a termination codon (TGA, TAA, TAG). Some algorithms were based on machine learning approaches such as neural networks, support vector machines, Fourier transforms, dynamic programming and Markov models, in which Markov models are among the most successful for gene finding in both prokaryotes and eukaryotes. A markov model (MM) is a stochastic model that assumes that the probability of a particular nucleotide occurring at a given position depends only on the k previous nucleotides. In this case k is the order of the MM, and the larger k the finer the MM can characterize dependencies between adjacent nucleotides. Markov models have been in use for decades as a natural method for modelling sequences. Several different types of Markov models have been used in order to capture the compositional differences among coding regions and noncoding regions as listed in Table 1.3 [131, 132].

Table 1.3. An overview of Markov models in gene prediction [132]

Positional weight matrices (PWM)	The simplest MMs are homogeneous zero order MMs which assume that each base occurs independently with a given frequency. Such simple models are often used for non-coding regions.
Weight array model (WAM)	An inhomogeneous higher order MM capable of capturing potential dependencies between adjacent positions of a signal.
Three-periodic Markov model	Characterize coding sequence. Coding regions are defined by three MMs, one for each position inside a codon.
Interpolated Markov model (IMM)	IMMs combine statistics from several MMs, from order zero to a given order k (typically k=8), according to the information available.
Hidden Markov model (HMM)	HMMs allow for insertions and deletions and so variation in signal length.
Generalized Hidden Markov model (GHMM)	GHMMs allow a string, rather than a single symbol, as the output of a state.
Semi-Markov conditional random field (SMCRF)	A more flexible variation of GHMM which allows a wider range of biological features to be incorporated with fewer technical concerns.
Evolutionary Hidden Markov model (EHMM)	EHMMs model molecular evolution as a Markov process in two dimensions: a substitution process over time at each site in the aligned genomes, which is guided by a phylogenetic tree; and a process by which the rate of evolution changes from one site to the next.

In 1994, a Hidden Markov model was built to predict the gene structure in *E. coli* [133]. The HMM includes 'states' that model the codons and their frequencies in *E. coli*, as well as the patterns found in the intergenic regions, including repetitive palindromic sequences and the Shine-Delgarno motif. In the same year, Stormo and Haussler showed that HMM could be generalized to allow a string, rather than a single symbol, as the output of a state in the model [134]. Gene prediction for eukaryotes involves not only the prediction of coding sequences, as in prokaryotes, but also the prediction of introns and the need to differentiate both from intergenic sequences. In 1996, Kulp described the first Generalized Hidden Markov model (GHMM) in the Genie program for eukaryotic gene prediction [135], with two neural networks for splice site prediction.

It is to be noted that *ab initio* prediction is lineage-specific and thus needs training to be done for every new genome or borrowed from a very closely related genome. Popular *ab initio* gene prediction programs are Genscan [136] capturing the general and specific compositional properties of exon, intron, splice site, promoter (TATA box, cap site), GeneID [137], SNAP [138], Glimmer [139], and GeneMark [140]. GeneID was designed

with a hierarchical structure. First, signals (splice sites, start and stop codons) are predicted and scored using position weight matrices (PWMs). Next, coding exons are constructed from these signals and classified into four categories: single, initial, internal, and terminal. Finally, the gene structures are assembled using a dynamic programming algorithm to maximize the sum of the score of the exons. Genscan, GeneMark, and GeneID distinguish coding from noncoding regions using three-periodic Markov models of order four or five on hexamer usage. Glimmer (Gene Locator and Interpolated Markov ModelER) uses interpolated Markov models (IMMs) to identify the coding regions and to distinguish them from noncoding DNA in microbial DNA. GENSCAN, SNAP, GlimmerHMM, and AUGUSTUS model the sequence via GHMM with specific state diagrams and use the Viterbi algorithm to produce a reliable gene architecture.

1.2.2.2. Evidence driven gene prediction

To improve the accuracy of gene prediction, many *ab initio* gene prediction programs such as AUGUSTUS [141, 142], SGP2 [143], Twinscan [144], and N-scan [145] can use external evidence including similarity search results and expression data (EST, cDNA, protein, RNA-seqs). AUGUSTUS has used GHMM to model the sequence with exons, introns, intergenic regions and so on, corresponding to states in the model, and applied a new method to model the intron length distribution. AUGUSTUS+ was built based on AUGUSTUS and incorporated extrinsic information to predict genes. This extrinsic information can be ESTs, cDNAs, second syntenic genomic sequence from another species, or RNA-seqs. SGP2 integrated the *ab initio* gene prediction GENEID with TBLASTX searches between two genome sequences. Twinscan extended the probability model of GENSCAN, allowing it to exploit homology between two related genomes. N-scan has predicted gene structures in one or more genomic sequences based on an alignment of the sequences. N-SCAN can model the phylogenetic relationships between the aligned genome sequences, including context dependent substitution rates, and insertions and deletions. In addition, there are some combined programs such as Eugene [146], Gaze [147], Jigsaw [148]. Eugene is a gene finder for eukaryotic organisms, using IMMs to model coding region in the same way as the Glimmer program. A specific characteristic of Eugene is its ability to integrate at ounce arbitrary sources of information into its prediction process, including the information obtained from several signals (splice

site, translation start...) prediction software tools, information obtained from similarity with existing sequences (EST, mRNA, 5'/3' EST from full length mRNA, proteins, genomic homologous sequences) and the output of other gene finders. Also Jigsaw can combine the outputs from gene finders, splice site prediction programs and sequence alignments to predict gene models.

1.2.3. Genome annotation pipeline

Although genome annotation pipelines differ in their details, the structure of a genome annotation pipeline has usually the following parts (as also shown in Figure 1.7 [149]): 1) predicting genes, 2) assigning the function to the genes, 3) storing the annotation data, and 4) visualizing the annotation data. In the gene prediction step, first, repeat regions in the genomic sequence are identified and masked. Next the extrinsic information (ESTs, cDNA, protein, RNA-seqs) is aligned to the sequence. After that, a gene prediction program is used to identify the gene structure of the sequence. In the functional annotation, the similarity of the encoded protein sequences of the genes to proteins in public databases is reported as well as the occurrence of documented domains, the putative gene ontology is searched, after which the annotation data is stored into a structured database. Finally, the annotation data is distributed via website annotation delivery system. From this website, the experts can curate their preferred genes.

During my PhD, I focused on the gene prediction and genome annotation of the spider mite, an arthropod that belongs to the chelicerates. A large amount of my time has been dedicated to the annotation of spider mite genomes, including *Tetranychus urticae*, *Tetranychus evansi*, and *Tetranychus linearius*. I have used EuGene (discussed above), a gene prediction platform for eukaryotes that combines several sources of evidence including RNA-seqs, to annotate these genomes. All information of annotation is stored in a local database and is accessible via ORCAE [150], a web application built by the Bioinformatics and Evolutionary Genomics group from the Department of Plant Systems Biology (VIB). The method is detailed in section 2.6.1. The spider mite consortium has used the annotation data to decipher arthropod-plant interaction in greater detail (see Chapter 2).

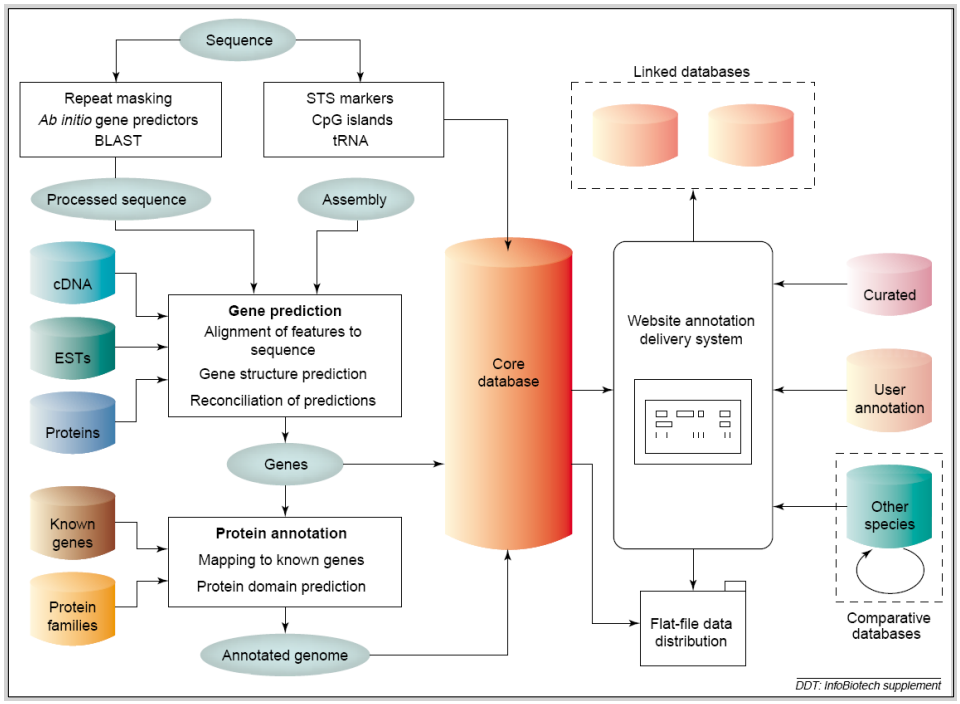


Figure 1.7. The generic structure of an automatic genome annotation pipeline and delivery system. The diagram can be simply partitioned into the analysis pipeline stage on the left and delivery system on the right, where data is distributed between the two stages via the central, core database. Raw, sequence data enters the pipeline and flows down through the gene prediction analyses before being stored in the core database. The website annotation delivery system is a key feature as it serves to integrate the gene predictions and features from the core database alongside multiple, external databases (taken from [149]).

1.3. Evolution of Arthropod-Plant Interaction

Most living organisms have evolved to exist and reproduce by using not only a combination of their own genetic machinery but also that of one or more other species with whom they interact [151]. For example, many plants would quickly become extinct without their animal pollinators. Among species interactions, interactions between plants and arthropods dominate the terrestrial ecology on the Earth [152]. These interactions can be useful for both plants and arthropods: plants provide shelter, oviposition sites, and food to arthropods while arthropods protect plants and help them to reproduce by pollination. However, many arthropods act as plant pests, and can be extremely harmful to plants. Arthropods are remarkably diverse when compared to other eukaryotic organisms. They

exist for more than 500 million years and come in all shapes and sizes. Depending on the range of host plants, arthropod herbivores are classified into generalists and specialists. Generalist arthropods are poliphagous and feed on many host plants from different families. Specialist arthropods are monophagous or oligophagous and feed on one or only few plants from the same plant family. Plant phylogenetic diversity has promoted diversity and abundance of herbivorous and predatory arthropods [153, 154] such that the diversity of herbivorous arthropods probably results from their specialization to different host plants. On the contrary, J. Daniel Hare [155] described how insect herbivores can drive the evolution of plants: "the presence or absence of particular herbivore species influences which plant genotypes are favored by natural selection". Thus the interaction of plants and arthropods is an important cause to create diversity in both the evolution of plants and arthropods. Understanding the evolution of these interactions can help us to control herbivores on crops as well as to understand the diversity of life. To date, with many plant genomes sequenced, many arthropod projects developed, and the new era of sequencing, studies on the evolution of species interactions at a molecular level promise to reveal much about the interrelationships of plant-arthropods.

1.3.1. Arthropod plant interactions in natural ecosystems

The complexity and dynamics of interactions between plants and herbivorous arthropods are shown in Figure 1.8 [156]. When herbivores attack a plant, the plant perceives herbivore-associated molecular patterns (HAMPs) by transmembrane pattern recognition receptors [157-160] and reacts by building direct defenses including physical barriers such as leaf toughness and trichomes, defensive proteins or secondary metabolites with toxic, repellent or anti-digestive effects on arthropod pests, and indirect protection as emitting volatile compounds to attract predators of arthropod pests. In turn, arthropod herbivores respond to these plant defenses by producing detoxification and digestion enzymes.

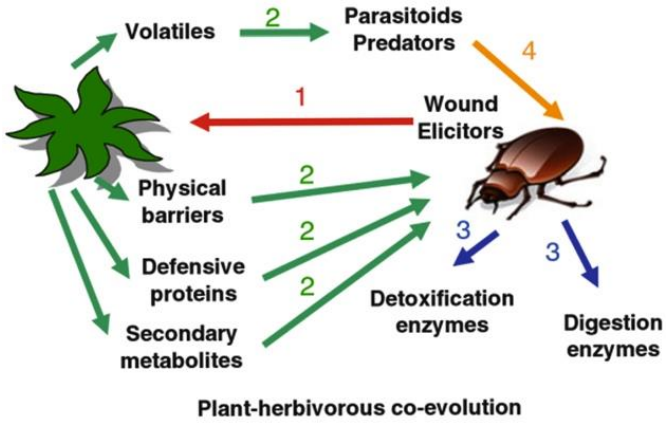


Figure 1.8. Arthropod-plant interactions (taken from [156])

Plants have evolved during ≈ 410 million years to defend against these herbivores [161]. For instance, the jasmonate family plays a central role in promoting defense gene expression as well as regulating defense responses of plant to herbivorous arthropods [162]. Some classes of plant secondary metabolites with defensive properties are studied well, such as terpenoids, alkaloids, furanocoumarins, cardenolides, tannins, saponins, glucosinolates, and cyanogenic glycosides. Defense proteins include lectins, ribosome-inactivating proteins, enzymes inhibitors, arcelins, chitinases, ureases, and modified storage proteins [163]. Herbivore-induced plant volatiles (HIPVs) are emitted from leaves, flowers, and fruits into the air or from roots into the soil in response to herbivores and attract natural enemies of herbivores as parasitoids and predators [164-166]. The main classes of plant volatiles such as terpenoids, phenylpropanoids/benzenoids, fatty acid derivatives, and amino acid derivatives are involved in plant defense against herbivores [167].

The plant disease resistance triggered by transmembrane receptors that recognize pathogen-associated molecular patterns (PAMPs, HAMPs in herbivores) belongs to the so-called basal defense. All plants have this defense. As a second line of disease resistance, resistance (R) gene mediated defense (also called gene-for-gene plant

resistance), is more specific and is only found in certain plant species [168]. Plant disease R proteins [169, 170] detect effector molecules (also called avirulence (Avr) proteins) that pathogens (herbivores) secrete into plant cells to counteract or weaken host defense. The plant disease resistance proteins are divided into five main classes (I-V): *Pto*, NBS-LRR, TIR-LR, *Cf*, *Xa21*. Nucleotide-Binding Site plus Leucine-Rich Repeat (NBS-LRR) is the largest class of R genes related to innate immune response proteins in animals [171]. R proteins share striking structural similarities although they confer resistance to diverse groups of organisms, such as bacteria, virus, fungi, oomycetes, nematodes, and insects. Surprisingly, the tomato *Mi-1* gene that belongs to the NBS–LRR class of R genes, can recognize phylogenetically distinct pathogens including potato aphid (*Macrosiphum euphorbiae*), whitefly (*Bemisia tabaci*), and root-knot nematodes (*Meloidogyne spp.*) [172]. Besides disease resistance, disease susceptibility is an opposite plant response to pathogens. An example of a plant disease susceptibility gene is *PMR6*, a pectate lyase-like gene, required for powdery mildew susceptibility in Arabidopsis [173]. A loss-of-function mutation in *PMR6* gene conferred resistance to the powdery mildew. A disease resistance gene and a disease susceptibility gene can share identity. For instance, *LOV1*, a disease susceptibility gene in *Arabidopsis thaliana* to the fungus *Cochliobolus victoriae*, is a member of the NBS-LRR resistance gene family [174].

In response to the plant defense, specific gene families have been evolved in arthropod herbivores with the ability to metabolize and detoxify plant chemicals. The genes directly involved in feeding can change their expression pattern in response to the host [156]. Moreover, herbivores often manipulate their host plants to feed on them in a better way. As a result of this manipulation, attacked plants may become even better resources for herbivores than non-damaged plants. For example, spider mite *Tetranychus evansi* manipulates its host (tomato) by interfering with signaling pathways involved in its defense mechanism [175].

1.3.2. Genome based research for the evolution of arthropod plant interaction

As described above, the genes involved in plant-arthropod interactions were discovered because of the development of molecular biology, genetics, genomics, electrophysiology, and biochemistry. To date, with the rapid advances of molecular biology, especially next

generation sequencing methods, not only many genomes of herbivores and their host plants are becoming available [176], but researchers can also survey an entire transcriptome under a variety of experimental and field conditions [177] to study the evolutionary ecology of arthropod-plant interactions [178]. For example, the completed *Acyrthosiphon pisum* and *Medicago truncatula* genomes allow performing genetic and genomic studies on both sides of the interaction [179, 180]. The genome sequence of *Tetranychus urticae* leads to a better understanding of plant mite interaction [181]. In May 2014, the completed or ongoing genome projects of 472 arthropods and 1968 plants were stored in GOLD [105]. These resources allow detailed studies on genes involved in plant-herbivous interaction. In addition, by genetic mapping on the insect side of the interaction using SHOREmap [182] with data genome-wide genotyping and candidate-gene sequencing, genes underlying host plant choice can be identified.

The spider mite consortium has studied the interaction between spider mites and plants using the spider mite genomes and transcriptomes of spider mite feeding on different plants at different development stages to gain new insights into the interaction, and to develop novel plant protection strategies. The *Tetranychus urticae genome paper*, to which I made an important contribution (see chapter 2), was published in 2011. Subsequent to the annotation of the spider mite genomes, we focused on the annotation and evolution analysis of chemosensory receptors (CRs) to study the relationship between the evolution of CRs and the range of host plants of three spider mites, *Tetranychus urticae*, *Tetranychus evansi* and *Tetranychus lintearius* (see chapter 3 and chapter 4).

1.4. Chemosensory receptors in arthropods

Animals use their chemosensory systems to detect and discriminate among chemical cues in the surroundings, such as odorants, tastants, and pheromones. By recognizing these chemical cues, animals locate food resources, find mates, avoid predators, and modulate communication with conspecifics. The molecular and cellular basis of chemosensory perception in insects largely based on studies in *Drosophila melanogaster* has recently been revealed, due to the identification of gene families for olfactory receptors and insect olfactory receptor neurons [183]. Two large multi-gene families were discovered in 1999 and 2000, and expressed in olfactory and gustatory organs of *Drosophila melanogaster*,

respectively [184, 185]. Since then, many odorant receptors (ORs) and gustatory receptors (GRs) have been described in insects, including species such as *Anopheles gambiae*, *Heliothis virescens*, *Aedes aegypti*, *Tribolium castaneum*, *Bombyx mori*, *Acyrthosiphon pisum*, and *Nasonia vitripennis*. Besides insects, only GRs, but no ORs, were identified in *Daphnia pulex*, a model species from the closely related arthropod sub-phylum crustacea [186]. Although ORs and GRs in insects have the seven-helical trans-membrane domain structure like the ones in mammals [187] and in nematodes [188], they clearly have a different origin. This is evident from the lack of sequence similarity, but more importantly from the protein domain organization being reversed in insects, the N-terminus located intracellularly and the C-terminus located extracellularly (Figure 1.9). While most chemoreceptors in mammals and nematodes are slow acting metabotropic receptors indirectly activating ion channels through second messengers, chemoreception in insects by ionotropic receptors including ORs and GRs being ligand-gated ion channels allows a much faster reaction [189] (Figure 1.9). Insect GRs comprise highly divergent sequences with 8-12% amino acid identity. This suggests that they could cover a broad range of tastants.

In 2004, an evolutionary independent taste receptor was discovered in *Drosophila*, called (D)mX receptor [190]. DmXR is an insect orphan G-protein-coupled receptor that has partially diverged in its ligand-binding pocket from the metabotrophic glutamate receptor family (mGluR). The mGluR structure is divided into three regions: the extracellular region, the seven-transmembrane spanning region, and the cytoplasmic region. The extracellular region is further divided into the ligand binding domain (LBD) and the cysteine-rich region. In mammals and insects, mGluRs, activated by the neurotransmitter glutamate, play different roles in the central nervous system. DmXR differs from mGluRs in the distal part of the LBP, so this receptor is not activated by glutamate or any other standard amino acids. However, LBD of DmXR and mGluR still share the crucial residues necessary to bind a ligand with amino acid structural properties. In 2009, L-Canavanine, a non-proteinogenic α-amino acid found in the seeds of many legumes, was proved as the ligand of DmXR [191]. Thus, the DmX receptor fulfills a gustatory function necessary to avoid eating a natural toxin as L-Canavanine.

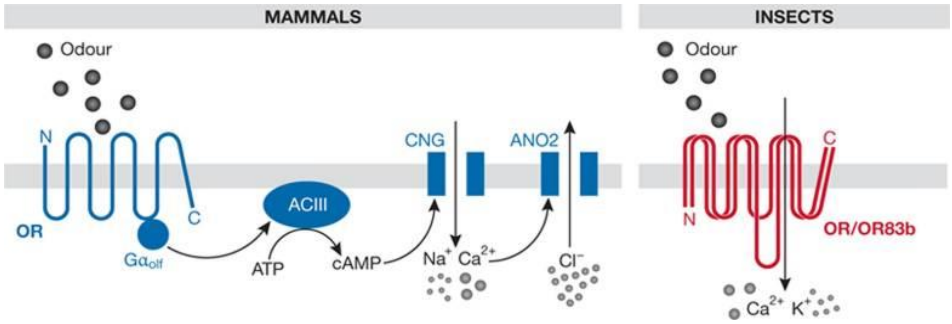


Figure 1.9. Signalling mechanisms of mammalian and insect odorant receptors (taken from [189]). A schematic of the molecular basis of olfactory signal transduction in the mouse and fruit fly. ACIII, type III adenylyl cyclase; ATP, adenosine triphosphate; cAMP, cyclic adenosine monophosphate; ANO2, anoctamin 2 channel; CNG, cyclic nucleotide-gated channel;.$G\alpha_{olf}$, olfactory G protein α-subunit; OR, odorant receptor.

In addition, in 2009, members of a large expansion of the ionotropic glutamate receptor (iGluR) gene family were identified in *Drosophila* as a novel class of chemosensory receptors and named ionotropic receptors (IRs) [192]. These IRs are not closely related to members of the canonical families of iGluRs (AMPA, kainate, NMDA, or delta). IRs are divergent, exhibiting overall amino acid sequence identities of 10%-70%. However, they have a similar domain structure comprising an extracellular N terminus, a ligand-binding domain (LBD) whose two lobes (S1 and S2) are separated by an ion channel pore formed by two transmembrane segments and a re-entrant pore loop, and a short cytoplasmic C terminus. IRs have divergent LBDs that lack some or all known glutamate-interacting residues, supporting their distinct classification from iGluRs. The comparison on chemosensory organs, receptors and putative ligands in the mouse and the fruit fly is shown in Figure 1.10.

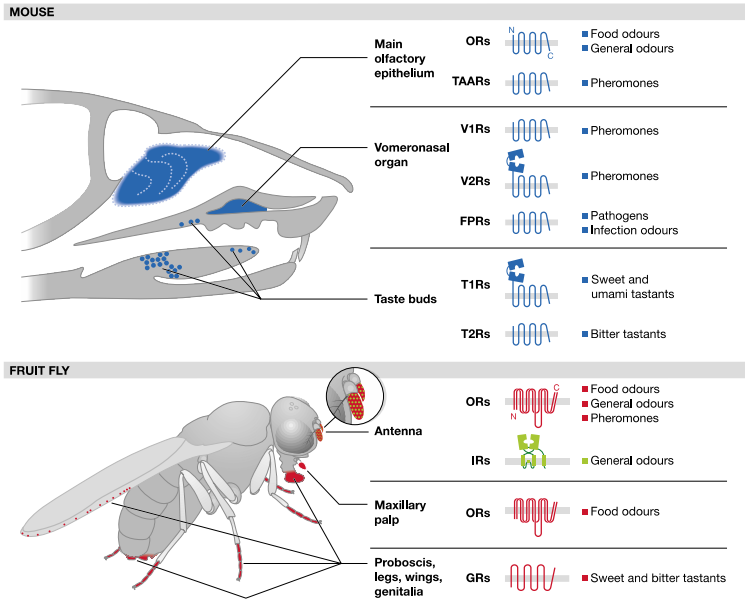


Figure 1.10. The main chemosensory organs, receptors and putative ligands in the mouse and the fruit fly. FPRs, formyl peptide receptors; GRs, gustatory receptors; IRs, ionotropic receptors; ORs, odorant receptors; T1Rs, taste receptors type 1; T2Rs, taste receptors type 2; TAARs, trace amine-associated receptors; V1Rs, vomeronasal receptors type 1; V2Rs, vomeronasal receptors type 2 (taken from [189]).

Like insects, spider mites are chemically sensing their environment by detecting a characteristic set of attractants, repellents and pheromones. Spider mites are repelled by 2,3-dihydrofarnesoic acid from glandular trichomes of *Solanum habrochaites*, a wild tomato which they avoid as a host [193]. They are also repelled by C15-C18 alkanes and to a lower degree by homologous alkenes [194]. Another wild tomato, *Solanum pimpinellifolium*, deters spider mite feeding by producing acylsucrose on its trichomes [195]. The African spider plant (*Gynandropsis gynandra*) is a strong repellent towards spider mites on roses, acetonitrile being the main volatile emitted by this plant [196]. On the reverse, males of *T. urticae* are attracted to pharate females, and would guard them until hatching when they would eventually mate. This attraction is mediated by a range of compounds from these pharate females, namely E,Z-farnesol, cis-nerolidol, geraniol and citronellol [197]. Interestingly, *Tetranychus urticae* specific predator *Phytoseiulus persimilis*, a blind mite, is strongly attracted by the odors of plants infested by *T. urticae*.

These herbivore-induced plant volatiles (HIPV) are a blend of chemicals that varies in composition and ratios between species and strains, including linalool, (E)-β-ocimene, nerolidol, (E)-DMNT and methyl salicylate [198]. Plant volatiles, induced or not, may nevertheless have ambiguous impacts, attracting both herbivores and their predators [199]. It even appears that spider mites may depress plant defense and the release of HIPV, facilitating further infestation [200] which tends to contradict previous report of HIPV acting as repellent [201]. Next to volatiles, chemicals laid on surfaces, e.g. in the very case of spider mites on silk, have to be considered too. Besides their role as a shield, silk threads are used as guiding trails and are indeed a mean of chemical communication among spider mite individuals [202] including kin recognition [203] but also a way for their predators to track them on a long range [204]. These reports altogether strongly argue for chemosensing to be a major aspect in the ecology of mites generally speaking and spider mites more specifically.

Chapter 2

The genome of *Tetranychus urticae* reveals herbivorous pest adaptations

Miodrag Grbić, Thomas Van Leeuwen, Richard M. Clark, Stephane Rombauts, Pierre Rouzé, Vojislava Grbić, Edward J. Osborne, Wannes Dermauw, Phuong Cao Thi Ngoc, Félix Ortego, Pedro Hernández-Crespo, Isabel Diaz, Manuel Martinez, Maria Navajas, Élio Sucena, Sara Magalhães, Lisa Nagy, Ryan M. Pace, Sergej Djuranović, Guy Smagghe, Masatoshi Iga, Olivier Christiaens, Jan A. Veenstra, John Ewer, Rodrigo Mancilla Villalobos, Jeffrey L. Hutter, Stephen D. Hudson, Marisela Velez, Soojin V. Yi, Jia Zeng, Andre Pires-daSilva, Fernando Roch, Marc Cazaux, Marie Navarro, Vladimir Zhurov, Gustavo Acevedo, Anica Bjelica, Jeffrey A. Fawcett, Eric Bonnet, Cindy Martens, Guy Baele, Lothar Wissler, Aminael Sanchez-Rodriguez, Luc Tirry, Catherine Blais, Kristof Demeestere, Stefan R. Henz, T. Ryan Gregory, Johannes Mathieu, Lou Verdon, Laurent Farinelli, Jeremy Schmutz, Erika Lindquist, René Feyereisen, Yves Van de Peer

Redrafted from Nature 479 (2011), 487-492

Author contribution

This chapter is redrafted from an article published in Nature. I was responsible for annotating protein coding genes, setting up the annotation portal under guidance of Stephane Rombauts and Yves Van de Peer (section 2.6.1) and helped write parts of section 2.3.1 concerning the transposable element and predicted gene information. I annotated, analyzed the transcription factors and wrote sections 2.3.8, 2.6.2 under the supervision of Yves Van de Peer. I drew Figure 2.2, Figure 2.8, Figure 2.10, and Table 2.2.

2.1. Abstract

The spider mite *Tetranychus urticae* is a cosmopolitan agricultural pest with an extensive host plant range and an extreme record of pesticide resistance. Here we present the completely sequenced and annotated spider mite genome, representing the first complete chelicerate genome. At 90 megabases *T. urticae* has the smallest sequenced arthropod genome. Compared with other arthropods, the spider mite genome shows unique changes in the hormonal environment and organization of the Hox complex, and also reveals evolutionary innovation of silk production. We find strong signatures of polyphagy and detoxification in gene families associated with feeding on different hosts and in new gene families acquired by lateral gene transfer. Deep transcriptome analysis of mites feeding on different plants shows how this pest responds to a changing host environment. The *T. urticae* genome thus offers new insights into arthropod evolution and plant–herbivore interactions, and provides unique opportunities for developing novel plant protection strategies.

2.2. Introduction

Mites belong to the Chelicerata, the second largest group of terrestrial animals. Chelicerates represent a basal branch of arthropods. Subsequent to their origin in the Cambrian period, arthropods radiated into two lineages: the Chelicerata and the Mandibulata (comprising the Myriapoda and the Pancrustacea (which includes both crustaceans and insects)) [205, 206]. Extant lineages of chelicerates include Pycnogonida, Xiphosura (horseshoe crabs) and Arachnida (a large group comprising scorpions, spiders and the Acari (ticks and mites) [207, 208] (Figure 2.1). Within the Acari, *T. urticae* belongs to the Acariformes with the earliest fossils dating from the Lower Devonian period (410 million years ago). The Acari represent the most diverse chelicerate clade, with over 40,000 described species that exhibit tremendous variations in lifestyle, ranging from parasitic to predatory to plant-feeding. Some mites are of major concern to human health and include allergy-causing dust mites, scabies mites and mite vectors of scrub typhus [209].

The two-spotted spider mite, *Tetranychus urticae*, is a cosmopolitan agricultural pest [210] belonging to an assemblage of web-spinning mites. The name ‘spider’ highlights

their ability to produce silk-like webbing used to establish a colonial micro-habitat, protect against abiotic agents, shelter from predators, communicate via pheromones and provide a vehicle for dispersion [211].

Tetranychus urticae represents one of the most polyphagous arthropod herbivores, feeding on more than 1,100 plant species belonging to more than 140 different plant families including species known to produce toxic compounds. It is a major pest in greenhouse production and field crops, destroying annual and perennial crops such as tomatoes, peppers, cucumbers, strawberries, maize, soy, apples, grapes and citrus. The recent introduction of the related species *Tetranychus evansi* to Europe and Africa from South America demonstrates the invasive nature of these pests in global agriculture [212]. Computer modelling suggests that with intensifying global warming, the detrimental effects of spider mites in agriculture will markedly increase [213] due to accelerated development at high temperatures.

Tetranychus urticae is known for its ability to develop rapid resistance to pesticides. Among arthropods it has the highest incidence of pesticide resistance [214]. Chemical control often causes a broad cross-resistance within and between pesticide classes, resulting in resistance to novel pesticides within 2–4 years. Many aspects of the biology of the spider mite, including rapid development, high fecundity and haplo-diploid sex determination, seem to facilitate rapid evolution of pesticide resistance. Control of multi-resistant mites has become increasingly difficult and the genetic basis of such resistance remains poorly understood [215].

As the first completed chelicerate genome, the comparison of the *T. urticae* genome with the genomes of insects and the crustacean *Daphnia pulex* expands the arthropod genetic toolkit. At the same time, the very compact *T. urticae* genome has unique attributes among arthropod genomes with remarkable instances of gene gains and losses. The completion of the *T. urticae* genome sequence opens new avenues for understanding the fundamentals of plant–herbivore interactions, developing novel pest-management strategies and producing new biomaterials on the nanometre scale.

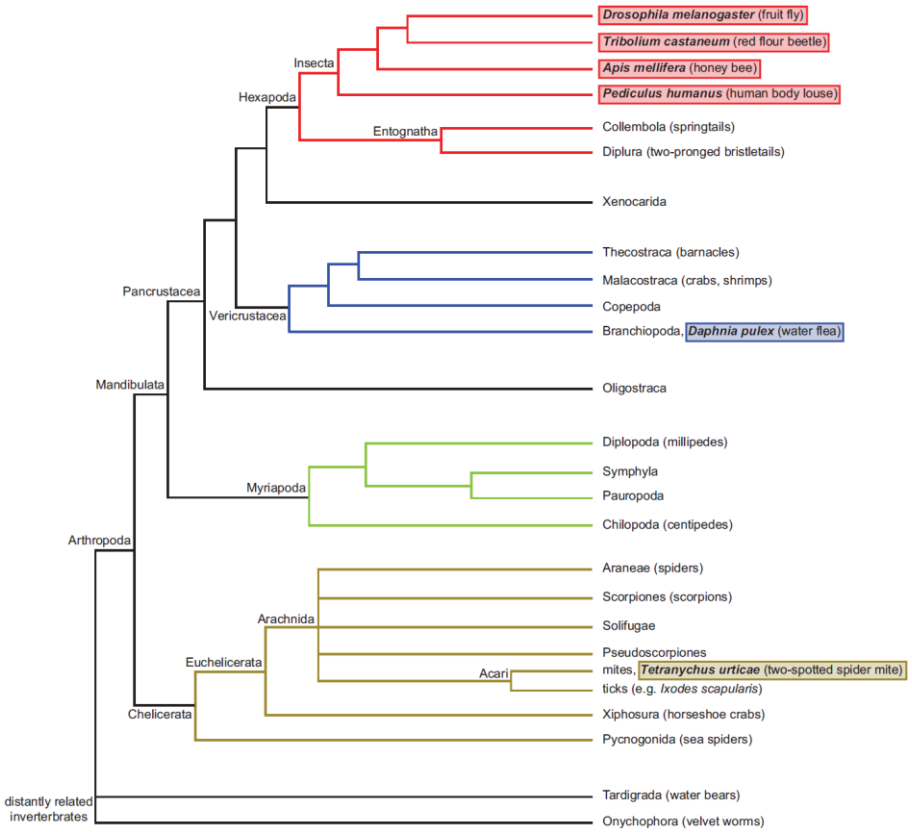


Figure 2.1. Phylogenetic position of the spider mite, *Tetranychus urticae* within the phylum Arthropoda. The tree represents generally accepted arthropod relationships as described in [206]. Species and clades mentioned in the text are represented, and fully sequenced genomes are in bold.

2.3. Results and Discussions

2.3.1. The small genome of *T. urticae*

The *T. urticae* genome (strain London) was sequenced (Sanger) to 8.05X coverage and assembled into 640 scaffolds covering 89.6 megabases (Mb). 70,778 Sanger expressed sequence tag (EST) sequences from embryos, larvae, nymphs and adults were generated, and further complemented with RNA-seq data on matching samples. We identified 18,414 protein-coding gene models (n), of which 84% (15,397) are supported by EST (8,243), protein homology (11,433) and/or RNA-seq data (14,545) (Figure 2.2A). From alignments of ~43-million paired-end Illumina reads from a second *T. urticae* strain

(Montpellier) to the London sequence, 542,600 single nucleotide polymorphisms and small indels were predicted. The complete genome annotation of *T. urticae* is available at the ORCAE website [150]. With an estimated genome size of about 90Mb, the *T. urticae* genome is the smallest arthropod genome sequenced so far. The genomes of other chelicerates are much larger (565–7,100Mb), with the unfinished genome of the tick *Ixodes scapularis* estimated at 2,100Mb [216]. Multiple characteristics of the *T. urticae* genome correlate with its compact size: small transposable element content and microsatellite density, increased gene density and holocentric chromosomes.

Transposable elements totalled 9.09Mb, putting *T. urticae* together with *D. pulex* and *Apis mellifera* as arthropods with 10% or less of their genomes comprised of transposable elements. Long terminal repeat (LTR) retrotransposons, and in particular Gypsy-like elements, were the most abundant type of transposable elements. L1-like Long interspersed elements (LINEs), Tc1/Mariner-like DNA transposons, and Maverick (Polinton) elements were also detected (Table 2.2). Deep sequencing of small RNAs (~19–30 nucleotides) across developmental stages identified 226,829 unique RNAs that mapped to 676,266 different loci in the genome. The number of unique small RNA counts per size category shows a peak at 21 and 26 nucleotides. These two peaks include short interfering RNAs and Piwi-interacting RNAs, respectively, similar to what is observed in *Drosophila melanogaster* [217]. Their alignments to the genome indicate that both probably silence diverse transposable elements. Included among ~21-nucleotide small RNAs are 52 predicted microRNAs (miRNAs). On the basis of the identity of their seed regions (nucleotides 2–7 of the miRNA sequence), the *T. urticae* miRNAs can be grouped into 43 families. Half of the predicted miRNAs were not conserved when compared to annotated miRNAs and available genomes of other arthropods [218], suggesting that they might be *T. urticae*- or lineage-specific.

The microsatellite density in the *T. urticae* genome is among the lowest observed for arthropods, consistent with the expectation that repeat content of genomes typically scales with genome size. The *T. urticae* microsatellite classes have a distinct profile: mono-nucleotide repeats are virtually non-existent, and di-nucleotide repeats, normally the most abundant type of microsatellites, are found significantly less often than tri-nucleotides, as

in *Tribolium castaneum* [219]. The gene density is twice as high compared to *D. melanogaster*, with 205 versus 92 genes per Mb, respectively. The mean number of exons per gene was low and similar to that found in *D. melanogaster* (~3.8 exons per gene). The size distribution of introns was typically skewed with a mean intron size of 400bp and a median of 96bp (Figure 2.2B, Table 2.1). The holocentric nature of *T. urticae* chromosomes [220] (the absence of centromeres and the diffuse nature of the kinetochores) is correlated with a lack of large tracts of gene-poor heterochromatin. The uniformly distributed gene density contrasts with the human body louse (*Pediculus humanus*), concentrated in only 55Mb of the 110-Mb genome [221].

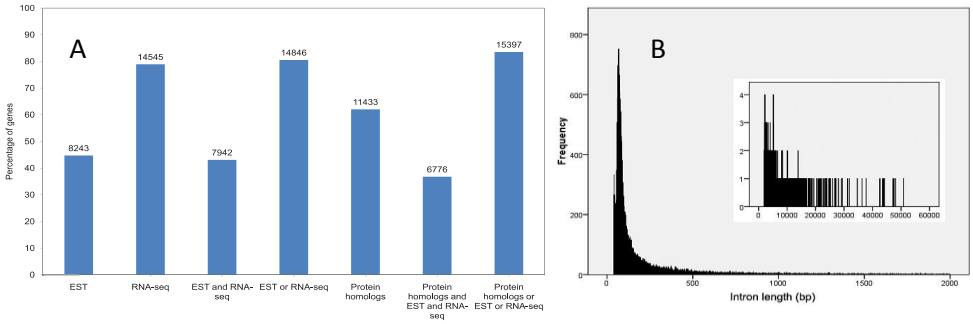


Figure 2.2. A. Predicted *T. urticae* genes supported by protein homologs, ESTs or RNA-seq reads/splice junctions. Protein homologs were determined from BLASTP [222] results of the predicted proteins against a protein database (E-value cutoff e^{-3}). Genes supported by ESTs were identified by GenomeTheader [223]. Genes supported by RNA-seq reads/splice junctions were identified by Bowtie [224] and Tophat [225]. Genes were regarded as supported by RNA-seq reads/splice junctions if having at least three RNA-seq reads/splice junctions aligning to it. B. Intron length distribution for genes of *T. urticae*. The minimum intron size is about 40bp with 267 introns. About 70% of introns have a length between 40bp to 150bp.

Table 2.1. Comparison of genome and annotation statistics for the draft sequence of the spider mite *T. urticae* genome and genomes of *D. melanogaster* and *T. castaneum*.

		T.urticae	%*	*D.melanogaster*	%*	*T.castaneum*	%*
nr. of loci (exons + introns)		18,414		14,861		14,460	
av. length of loci	nt	2,652		6,328		5,422	
loci density	nt/gene	4,866		12,160		10,471	
	genes/Mb	205.51		91.53		95.55	
nr. of genes		18,414		13,353		14,452	
av. length of gene	nt	1,428		1,506		1,371	
median length of genes	nt	1,138		1,139		1,011	
nr. exons		70,405		38,648		65,479	
cumul. exon length	nt	26,292,088	29.34	20,111,395	12.39	19,818,063	13.10
av. length of exon	nt	374		520		303	
median length of exon:	nt	178		312		195	
longest exon	nt	45,659[a]		27,510		26,331	
av. nr. exons/gene		3.82		2.89		4.53	
most exons/gene		55[b]		81		105	
cumul CDS length	nt	19,505,397	21.77	17,825,960	10.98	19,784,616	13.07
av. length of CDS	nt	1,060		1,335		1,369	
longest CDS	nt	54,762[c]		68,916		63.354	
shortest CDS	nt	63[d]		2		60	
%GC of CDS		37.8		53.2		44.3	
cumul. intron length	nt	20,681,179	23.08	14,956,521	9.21	34,101,322	22.53
av. length of intron	nt	400		597		711	
median length of intron	nt	96		67		53	
nr. of big introns (>20 kb)		36		198		121	
longest intron	nt	50,833[e]		131,739		98,797	
%GC of intron		29.7		40.3		32.1	
genome size (scaffolds)	nt	90,815,494		168,736,537		210,566,138	
genome size (contigs)	nt	89,600,102		162,367,812		151,333,735	
largest scaffold	nt	7,801,961		29,004,656		38,791,480	
av. scaffold length	nt	141,899		11,249,102		97,394	
number of contigs		2,035		137		8,828	
largest contig	nt	929,118		27,905,053		597,263	
av. contig length	nt	44,030		1,185,167		17,142	
gaps (>50N)		1,395		119		6,660	
percent of the genome involved in protein encoding transcripts (exon + intron)			52.43		21.60		35.63

*percent of total genome, [a] exon: tetur30g00590.4, [b] gene: tetur04g02800, [c] gene: tetur30g00590, [d] 84 genes in total, [c] exon: tetur07g02140.1, [d] exon: tetur07g02140.2.

Table 2.2. Composition of transposable elements (TEs) in the *T. urticae* genome. A TE is regarded as complete when its sequence shows at least 90% coverage in length with a similar TE.

	Total bp TE	% bp TE	Nu. of TEs	Total bp complete TE	% bp complete TE	Nu. of complete TEs
Transposable element	**9,089,640**	**10.0**	**13,552**	**5,350,678**	**58.87**	**2,243**
Class I: retrotransposon	5,657,281	6.23	6,738	3,512,918	62.10	1,169
LTR retrotransposon	3,510,815	3.87	3,459	2,343,456	66.75	427
Gypsy	2,827,124	3.11	2,594	1,912,013	67.63	348
Copia	683,691	0.75	865	431,443	63.10	79
Non-LTR retrotransposon	2,146,466	2.36	3,279	1,169,462	54.48	742
LINE	2,146,466	2.36	3,279	1,169,462	54.48	742
L1	1,536,281	1.69	2,853	695,386	45.26	624
CR1	296,083	0.33	222	225,829	76.27	77
R2	201,874	0.22	145	177,115	87.74	24
I	102,101	0.11	53	65,606	64.26	15
LOA	10,127	0.01	6	5,526	54.57	2
Class II: DNA transposon	3,432,025	4.00	6,813	1,837,760	53.55	1,074
TIR	2,290,988	2.52	6,016	1,157,831	50.54	1,002
Tcl-Mariner	1,487,149	1.64	3,983	748,908	50.36	656
PiggyBac	291,499	0.32	661	92,499	31.73	3
Mutator	138,067	0.15	253	103,021	74.62	70
Merlin	115,975	0.13	411	38,953	33.59	46
CACTA	71,309	0.08	41	61,547	86.31	29
hAT	56,031	0.06	99	45,957	82.02	30
MIKE	86,975	0.10	473	41,020	47.16	146
P	16,223	0.02	20	13,335	82.20	10
Harbinger	9,383	0.01	14	7,743	82.52	8
IS4EU	2,776	0.00	3	2,776	100.00	3
Pogo	15,601	0.02	58	2,072	13.28	1
Helitron	78,741	0.09	91	53,040	67.36	22
Maverick	1,062,296	1.17	706	626,889	59.01	50
unclassified	334	0.00	1	0	0.00	0

2.3.2. Comparative genomics

As the first completely sequenced and annotated chelicerate genome, the *T. urticae* genome expands the set of arthropod genomes beyond Pancrustacea and provides an important out-group for comparative genomics. Comparison of the coding gene repertoire of *T. urticae* with the arthropods *T. castaneum*, *D. melanogaster*, *Nasonia vitripennis* and *D. pulex*, the chordate *Homo sapiens*, and the cnidarian *Nematostella vectensis* (Figure 2.3) resulted in 2,667 shared gene families. Almost 3,000 gene families are common to

the arthropods sampled, whereas 5,038 gene families (8,329 genes) are unique to *T. urticae*. Of those, 622 gene families (1,398 genes) have homologues in species other than those listed above, most of which belong to other arthropods. Homologues of 74 gene families (93 genes) were found in the unfinished genomes of tick [216] and/or *Varroa destructor* [226] and are probably chelicerate, rather than specific to *T. urticae*. Therefore, 4,416 gene families (6,609 genes) were found to be unique to *T. urticae*. A gene gain/loss analysis of these genomes showed a gain of about 700 new gene families in the lineage leading to *T. urticae*, plus almost 4,300 genes that are single copy (orphans). More than 1,000 gene families, still present in other arthropods, were lost in *T. urticae*. The 58 gene families are significantly (*z*-score >2) expanded in *T. urticae* compared to the other arthropods.

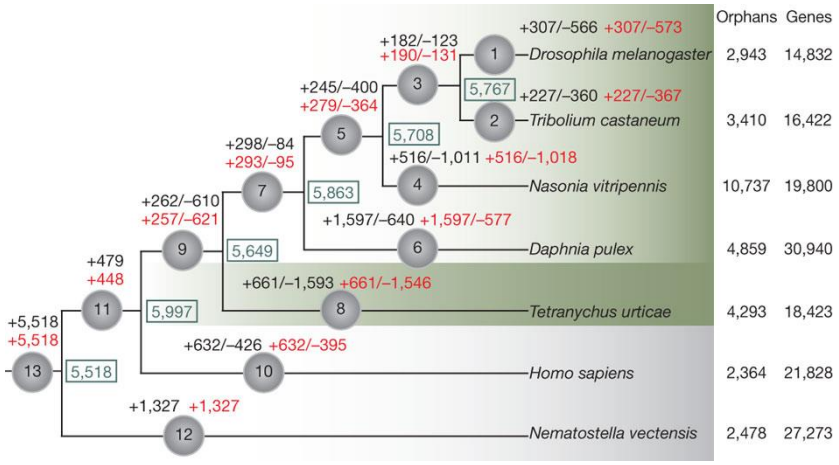


Figure 2.3. Gene family history. At each time point (grey circles), the number of gains (1) and losses (2) of gene families is indicated as inferred by DOLLOP (black) and CAFE´ (red) programs. The inferred ancestral number of gene families, according to DOLLOP, is shown in green boxes.

2.3.3. Feeding and detoxification

Tetranychus urticae is one of the most striking examples of polyphagy among herbivores and it has an unmatched ability to develop resistance to pesticides [210, 215]. We discovered that known gene families implicated in digestion, detoxification and transport of xenobiotics had a unique spider mite composition, and were often expanded when

compared to insects. This included a threefold proliferation of cysteine peptidase genes, particularly C1A papain and the C13 legumain genes, consistent with proteolytic digestion based mostly on cysteine peptidase activity [227]. Eighty-six cytochrome P450 (CYP) genes were detected in the *T. urticae* genome, a total number similar to insects but with an expansion of *T. urticae*-specific intronless genes of the CYP2 clan. The carboxyl/cholinesterases (CCEs) gene family contained 71 genes, with a single acetylcholinesterase gene (*Ace1*) but two new clades at the root of the neurodevelopmental class of CCEs, representing 34 and 22 CCEs, respectively. A notable case of expansion was found within the family of 32 glutathione *S*-transferases (GSTs) that include a group of 12 Mu-class GSTs that were, until now, believed to be vertebrate-specific. Finally, we discovered 39 multidrug resistance proteins belonging to the ATP-binding cassette (ABC) transporters (class C). The repertoire from this class of ABC transporters far exceeds the number (9–14) found in crustaceans, insects, vertebrates and nematodes. Few of the genes involved in detoxification had close insect homologues, and only four of the CYP genes could clearly be assigned as orthologues of insect and crustacean CYP genes.

The involvement of these gene families and their spider-mite-specific expansion in host plant adaptation is markedly illustrated by RNA-seq transcriptome profiling of spider mite feeding on its preferred host, bean (*Phaseolus vulgaris*), and on hosts to which the London strain is not adapted: *Arabidopsis thaliana* and tomato (*Solanum lycopersicum*) (Figure 2.4). We found 24% of all genes to be differentially expressed upon host transfer (Figure 2.4a-c); relative to bean, more genes were differentially expressed on tomato than on *A. thaliana*, but responses were nonetheless correlated (Figure 2.4b,c). Genes in the detoxification and peptidase families exhibited the most profound changes (Figure 2.4a-c), with expression of nearly half of P450 genes affected by the host plant, including 19 of 39 genes in the intronless CYP392 family and the CYP389 family. These subfamilies are spider-mite-specific P450 expansions that define lineage-specific expansions [228]. This finding is unprecedented. In humans, only up to one-third of P450 genes are metabolizing xenobiotics [229], and in *D. melanogaster* only one-third of the CYP genes are inducible by xenobiotics [230]. The proportion of P450 genes responding to the chemical environment is much greater in the spider mite. Similar patterns were also found

within other families (Figure 2.4c). For GSTs and CCEs, the expression of Mu and Delta GSTs and the two spider-mite-specific CCE clades were most affected and about one-third of cysteine peptidases, the C1A papains and C13 legumains, were overexpressed after transfer to tomato. More than two-thirds of the CYP and GST genes affected by the host plant are present in clusters of (multiple) tandem duplicated genes. Co-regulation of the majority of tandem duplicates strongly indicates that the ancestral gene was already plant-responsive before duplication, and that a role in plant adaptation may have favoured duplicate retention.

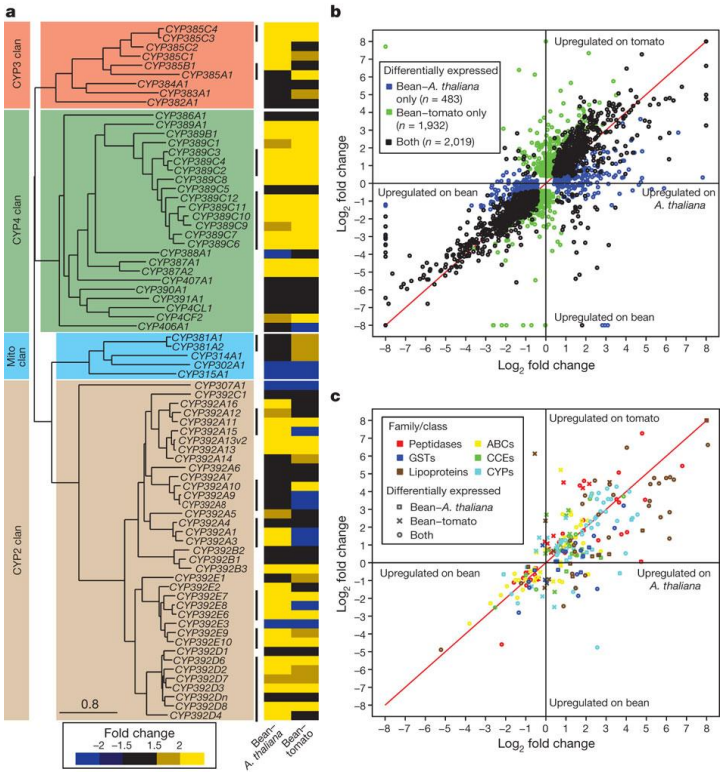


Figure 2.4. Gene expression changes when mites are shifted from *P. Vulgaris* (bean) to *A. thaliana* or to *S. lycopersicum* (tomato). a, A phylogeny of the cytochrome P450 (CYP) genes and heat map of the response of CYP genes to host transfer. Two-thirds of the genes that are tandemly duplicated or that form clusters (indicated by black vertical lines) are co-regulated. b, Global changes in gene expression after host shift. c, Fold changes of important gene family members in digestion and detoxification are colour coded. The analysis of differential expression (b and c) is with a5% false discovery rate as assessed with RNA-seq data collected in biological triplicate (fold changes between mean values are plotted).

Although these data indicate that spider-mite-specific expansion of known gene families contributes to the ability of spider mites to overcome host defences, many genes differentially regulated upon host transfer lack homology to genes of known function. Notably, among those with the most extreme expression fold-changes are genes that encode putative secreted proteins or lipid-binding proteins. Understanding extracellular binding and transport of small ligands is therefore likely to be important in further dissecting spider mite–plant interactions.

2.3.4. Lateral gene transfer

Our search for genes related to detoxification and digestion also revealed the existence and surprising expansion of intradiol ring cleavage dioxygenases, genes previously unreported from metazoan genomes but characteristic for bacteria and fungi [231]. We annotated 16 functional genes in this family in *T. urticae*, whereas bacterial genomes usually carry only 1 to 7. They have an average sequence similarity of 43% with the homologue of *Streptomyces avermitilis* and share the conserved 2 His 2 Tyr non-haem iron(III) binding site. These dioxygenases might have evolved to metabolize aromatic compounds found in plant allelochemicals. Other clear instances of lateral gene transfers include (1) the presence of a cobalamin-independent methionine synthase (*MetE*) gene with four predicted introns and up to 58% sequence identity to the *MetE* gene of soil Bacilli (this sequence has not previously been reported in any animal species); (2) two very similar levanase-encoding genes of probable bacterial origin that encode secreted exo-fructosidases upregulated upon feeding on tomato; and (3) a cyanate lyase-encoding gene that might be involved in feeding on cyanogenic plants.

We detected two clusters of carotenoid biosynthesis genes in *T. urticae* representing homologues of genes from zygomycete fungi and aphids. The latter are the only animal carotenoid biosynthesis genes known so far, thought to be derived from fungal genes by lateral gene transfer [232]. The unique intron–exon structure of the spider mite and aphid genes and their clustering in phylogenetic analyses is strong evidence that the genes from fungi were transferred only once to arthropods (Figure 2.5). The sequence and orientation of the two spider mite clusters indicate that they are the result of an ancient transfer followed by duplications, rearrangements and divergence. They also suggest that a

second, more recent transfer occurred between a spider mite and an aphid ancestor, although the sequence of the two transfers remains speculative. Carotenoids are known to have a role in diapause induction in spider mites [233] and our findings indicate that they can also synthesize them.

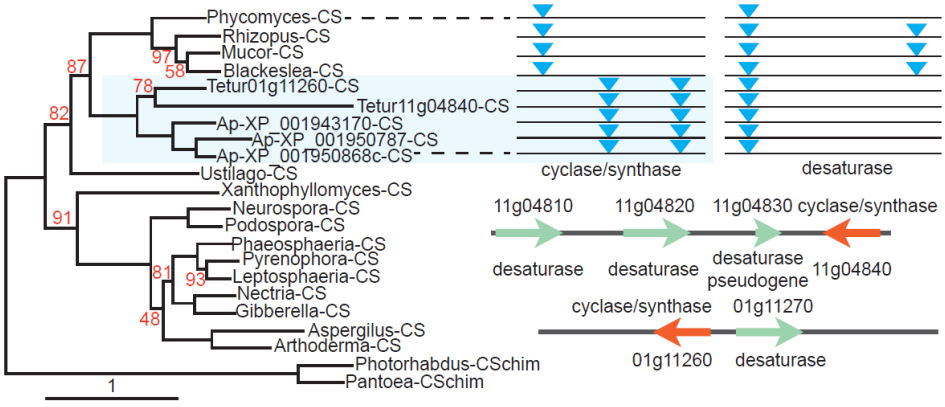


Figure 2.5. Maximum likelihood phylogeny of the fungal and arthropod carotenoid cyclase/synthase (CS) fusion proteins. The out-group comprises chimaeric assemblies (CSchim) of the closest bacterial sequences of cyclases and synthases. The *T. urticae* and *Acyrthosiphon pisum* sequences form amonophyletic group closely related to the zygomycete sequences. Evidence for a single lateral gene transfer event is also shown by the common intron positions in the cyclase/synthase (orange) and desaturase (green) genes (upper right panel). Two clusters of carotenoid biosynthesis genes are found in *T. urticae*: a tail-to-tail arrangement on scaffold 1 as seen in zygomycetes and aphids, and a more complex head-to-head (re)arrangement on scaffold 11 (bottom right).

2.3.5. Ponasterone A as moulting hormone

Ecdysteroid control of moulting is one of the defining features of arthropods. We detected gene orthologues coding for ecdysteroid biosynthesis enzymes [228]. Surprisingly, the *T. urticae* genome lacks two P450 genes, *CYP306A1* and *CYP18A1*, encoding, respectively, the biosynthetic C25 hydroxylase and a C26 hydroxylase/oxidase involved in hormone inactivation. The absence of CYP306A1 indicates that the spider mite uses the ecdysteroid 25-deoxy-20-hydroxyecdysone (ponasterone A) as the moulting hormone, instead of the typical arthropod 20E. This was confirmed by biochemical analysis of

spider mite extracts by HPLC–enzyme immunoassay and liquid chromatography/mass spectrometry that identified ponasterone A. *CYP306A1* and *CYP18A1* form a head-to-head cluster in all insect and crustacean genomes studied so far, therefore their absence from the *T. urticae* genome indicates that they were lost together, affecting both biosynthesis and inactivation pathways of the spider mite moulting hormone. Ponasterone A has been previously identified in some decapod crustaceans, albeit always coincident with 20E [234], and it is a high potency ligand of all known ecdysteroid receptors.

2.3.6. Reduced Hox cluster

Hox genes are a conserved set of homeobox-containing transcription factors typically found clustered within the genome and used to establish region-specific identity during early development. The body plan of mites consists of an anterior prosoma and posterior opisthosoma and is further distinguished by an extremely reduced body plan presumably achieved through the fusion of segments (Figure 2.6b). The ancestral arthropod is predicted to have a Hox cluster with 10 genes [235]. The *T. urticae* genome contains 8 of the canonical 10 genes. The *ftz* gene is present in duplicate, in two closely linked copies; orthologues of *Hox3* and *abdominal A* (*abdA*) were not found (Figure 2.6a). This is unusual among chelicerates: all 10 canonical Hox genes are present in the wandering spider [236]. The absence of *abdA* in *T. urticae* correlates with the spider mite's reduced opisthosomal segmentation. Consistent with the absence of *abdA* and a reduced opisthosoma, only two opisthosomal stripes of the segment polarity gene *engrailed* (typically expressed in each arthropod segment) are detected in the developing embryo (Figure 2.6c), in contrast to five *engrailed* stripes detected in the opisthosoma of the wandering spider [237]. Although numerous examples correlate morphological variation in arthropods with changes in Hox gene expression, this is the first example that correlates morphological evolution with the loss of a Hox gene within a fully sequenced Hox cluster.

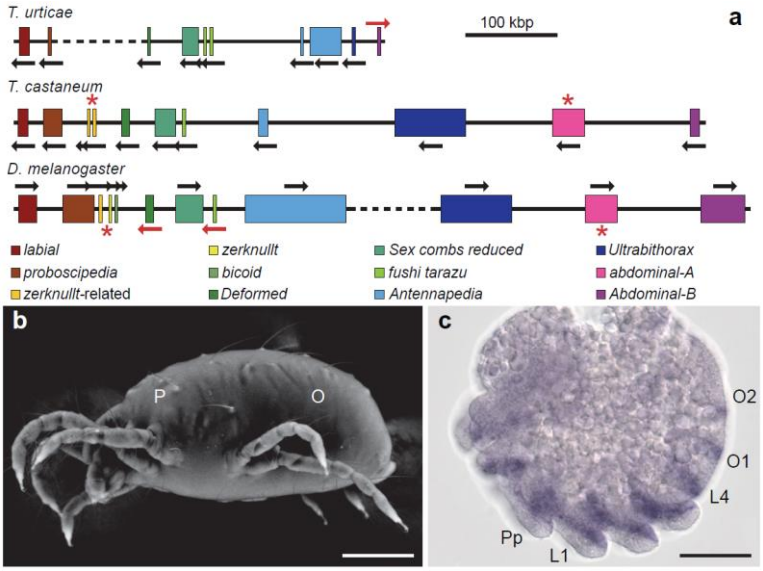


Figure 2.6. Comparative organization of Hoxclusters and expression pattern of the *T. urticae* engrailed gene. a, *T. urticae*, *T. castaneum* and *D. melanogaster* Hox clusters. Gene sizes and intergenic distances are shown to scale. Dashed lines represent breaks in the cluster >1Mb. In *T. urticae*, *fushi tarazu* and *Antennapedia* are present in duplicate whereas abdominal-A and Hox3/zerknullt are missing (red asterisk). b, Variable pressure scanning electron microscopy (SEM) image of adult *T. urticae* with two main body regions indicated: P, prosoma; O, opisthosoma. c, *T. urticae* engrailed (en) expression pattern. *en* transcripts are detected in five prosomal stripes that correspond to future pedipalpal (Pp), four walking leg (L1–L4) and two opisthosomal (O1 and O2) segments. Scale bars: b, 0.125mm; c, 40 mm.

2.3.7. Nanometre dimensions of *T. urticae silk*

Silk production in spider mites (Figure 2.7a,b) represents a *de novo* evolution of silk-spinning relative to silk production in spiders. Spiders typically spin silk from a complex glandular abdominal spinneret, whereas *T. urticae* uses paired silk glands connected to the mouth appendages (pedipalps) [238]. Seventeen fibroin genes were uncovered in the genome of *T. urticae* encoding fibroins of unusually high (27–39%) serine content. We performed mechanical testing on fibres deposited by adult and larval mites with an atomic force microscope. This technique measures the Young's modulus of the fibres, which is the ratio of applied stress (tension per cross-sectional area) to the resulting strain (fractional change in length) and describes the stiffness of the material. Young's modulus

was higher than or comparable to other natural materials, but *T. urticae* silk fibres are thinner—54±3nm (adult silk, Figure 2.7c) and 23.3±0.9nm (larval silk), that is, 435–185 times thinner—than the silk fibres of the spider *Nephila clavipes* [239].

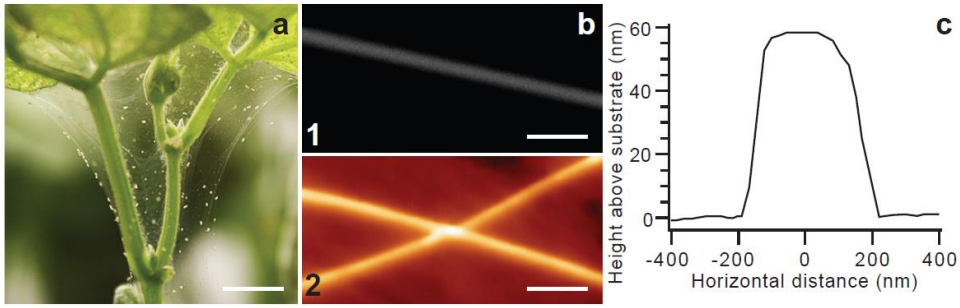


Figure 2.7. *T. urticae* silk structure and dimensions. a, Spider mite colony on a bean plant forming characteristic silk webbing. b, SEM image of the spider mite larval silk filament (top), and atomic force microscopy (AFM) image of two larval spider mite silk filaments (bottom). c, Height profile of the adult spider mite silk filament obtained from the AFM image. Scale bars: a, 0.75 cm; b, 1 µm.

2.3.8. Transcription factors

Using the method in section 2.6.2, we found a total of 772 TFs in the genome of *T. urticae*, comprising ~4.2% of all *T. urticae* genes. In eukaryotes, approximately 3-5% of all genes usually encode TFs. Of the 772 TFs, 734 TFs are similar to *D. melanogaster* TFs, 33 TFs are specific to arthropods, 8 TFs are specific to insects, and 462 TFs are specific to animals. *T. urticae* TFs are divided into 49 families compared to 50 TF families of *D. melanogaster* (Figure 2.8). BESS TFs appear to be specific to *D. melanogaster* and some other insects while these are missing from *T. urticae.* TFs with the zf-C4 DBD are more expanded in *T. urticae* than in *D. melanogaster.* On the contrary, TFs with the zf-C2H2 DBD are remarkably reduced in *T. urticae* compared *D. melanogaster.*

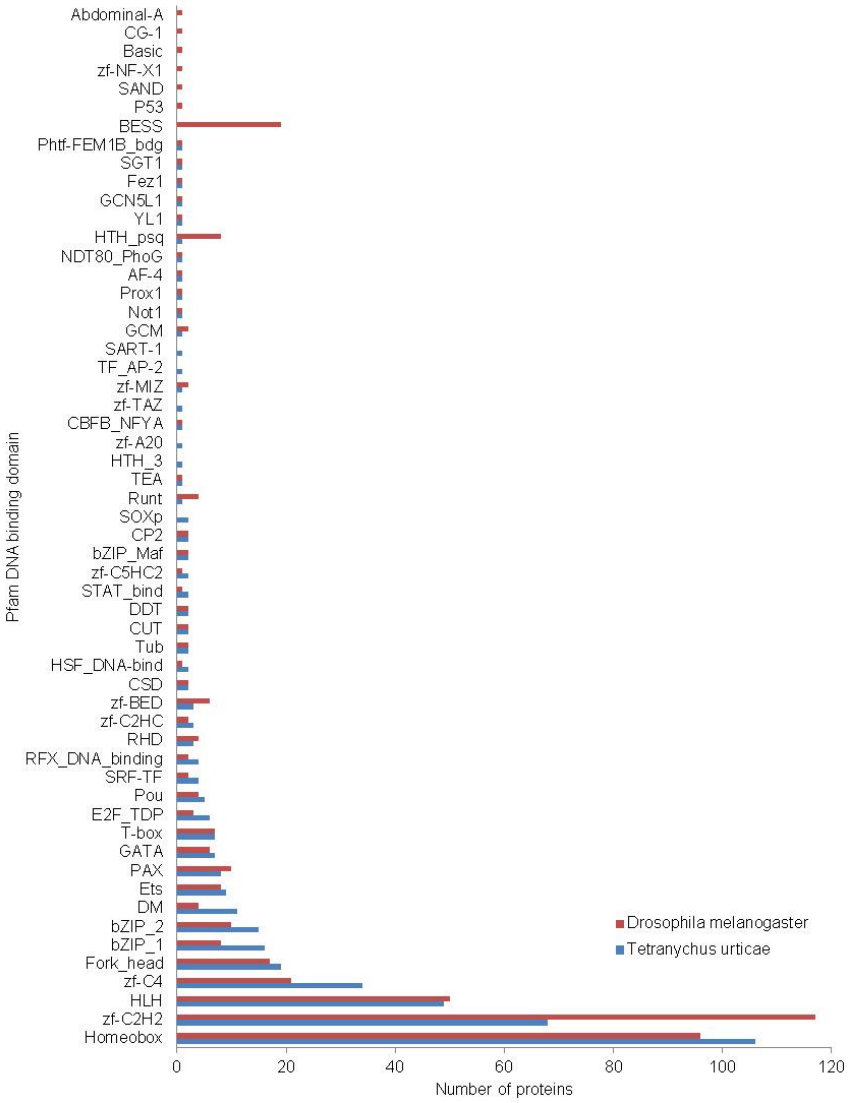


Figure 2.8. Transcription factor families in *T. urticae* and *D. melanogaster*.

2.4. Concluding remarks

Our analysis of the *T. urticae* genome also included nuclear receptors and neuropeptide genes, immunity-related genes and RNA interference, cuticle protein genes, and DNA methylation. The first complete genome of a chelicerate species provides the opportunity for a detailed phylogenomic analysis of arthropods, the most diverse group of animals on

Earth. The *T. urticae* genome illustrates the specialized life history of this polyphagous herbivorous pest. Striking gene gains include lineage-specific expansions within detoxification gene families and lateral transfer of genes from fungi and bacteria that further expanded in *T. urticae*. The functional significance of these innovations is supported by the upregulation of many of these genes in response to feeding on less preferred host plants.

The genome of the two-spotted spider mite, together with the favourable biological features of the spider mite as a laboratory model including short generation time, easy rearing and tools for gene analysis and gene silencing [240], provide a novel resource for agriculture that should allow the dissection of pest–plant interactions and development of alternative tools for plant protection. Finally, evolutionary innovation in the process of *T. urticae* silk production expands the repertoire of potential chelicerate biomaterials (such as the well-known spider silk) with a natural biomaterial at the nanometre scale.

2.5. Methods

All genomic sequencing reads were collected with standard Sanger sequencing protocols. RNA sequencing was performed with Illumina RNA-seq protocols. Annotation of the *T. urticae* genome was done using the gene prediction platform EuGene. The complete genome annotation is available at http://bioinformatics.psb.ugent.be/orcae/. The *T. urticae* (London) genome project was registered under the INSDC project ID 71041.

2.6. Supporting information

2.6.1. Annotation of protein coding genes

Annotation of the *T. urticae* genome was done using the gene prediction platform EuGene. This gene prediction platform is designed to be able to integrate many different sources of extrinsic evidence as well as *ab initio* prediction results. For intrinsic gene prediction, different software modules need to be trained and a number of parameters need to be estimated [146]. For instance, splice sites were identified using the SpliceMachine [241] signal sensor components trained specifically on *T. urticae* data. To this end, a ‘positive’ set of 2,690GT donor and 1,455AG acceptor sites was constructed in windows of 402 bp (200 bp up- and downstream of either the donor or the acceptor

site). The negative set consisted of windows of the same size but with 23,905GT and 23,854AG dinucleotides known not to be splice sites as based on EST alignments. These windows were thus exclusively derived from either exon or intron regions. The SpliceMachine models gave specificity scores of 84.3% for donor and 65.2% for acceptor. The content sensor used by EuGene to recognize coding sequences is an interpolated Markov model that was trained on 15,887 *T. urticae* conserved coding regions that were collected genome-wide using BLASTX on proteins in SWISSprot. For the non-coding part of the IMM, we extracted 6697 introns based upon the spliced alignments obtained by ESTs. We only considered introns that were confirmed by at least six ESTs. The same intron data was used to extract donor and acceptor sites (see above). Training EuGene also requires the estimation of scaling parameters from known *T. urticae* genes within their genomic context. As such, 211 genomic *T. urticae* sequences that each contained abutting genes were constructed and used to train EuGene. After training, we obtained a sensitivity of 93% and specificity 81.4% on a set of 211 manually curated genes.

For extrinsic annotation, the following data sources were integrated into our annotation system: 1) protein sets from the latest Flybase [242] release, other arthropods and Uniprot-Swissprot; and 2) ESTs generated from the *T. urticae* for four different developmental stages as well as a large number of Illumina RNA-seq reads resulting from different developmental stages and several feeding experiments performed on different hosts (Arabidopsis and bean). The fully integrated annotation obtained with EuGene yielded 18,414 protein encoding nuclear genes (Table 2.1). Following gene prediction and manual annotation, RNA-seq reads were aligned back to predicted gene models with Bowtie/Tophat to assess expression. Of the 18,414 predicted protein-encoding genes, 8243 genes were supported by ESTs, while 14,545 genes were supported by RNA-seq data; 11,433 had a protein homologue in the non- redundant protein database (Figure 2.2A and Figure 2.9). Sequence patterns of GT, GC donors and AG acceptors of these protein-coding genes showed in Figure 2.10.

To compare the genome features (genes, exons and introns) with reference insect genomes a custom Perl script was used, parsing the GFF3 and genomic fasta files

describing the genomes of *D. melanogaster* (FlyBase-r5.2911), *T. castaneum* (Tcas2.012) and *T. urticae.* The genome features of all three organisms are presented in the Table 2.1.

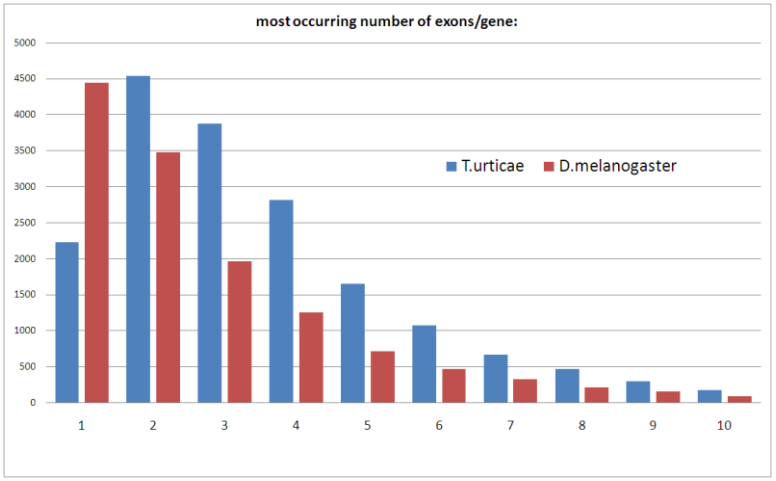


Figure 2.9. Distribution of the number of exons per gene (cut-off is 10 exons) for *T. urticae* and *D. melanogaster*. 2966 genes contain no introns. 529 genes contain 11 exons or more.

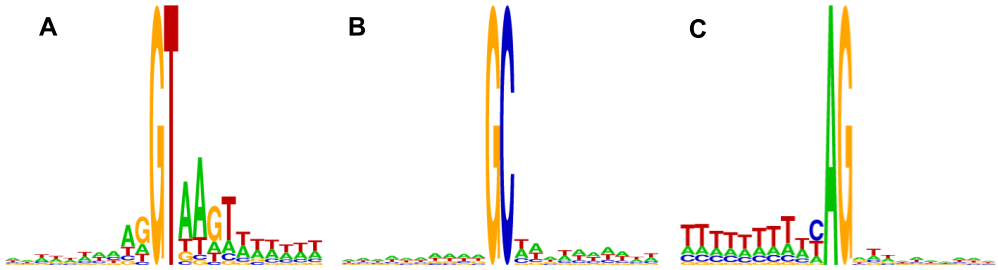


Figure 2.10. Sequence logos of donor and acceptor sites. A) Sequence logo of GT donorsite (plus 10 bp up- and downstream) created from 43,732 sequences. B) Sequence logo ofGC donor sites created from 1,647 sequences. C) Sequence logo of AG acceptor sites createdfrom 45,384 sequences.

2.6.2. Identification of transcription factors

T. urticae transcription factors (TFs) were predicted with 167 DNA binding domains (DBDs) described in Pfam (v24.0), 38 DNA binding families of 'Superfamily' [243] and transcription-related GO terms. All protein sequences were searched against the Pfam DBD HMMs by pfam_scan.pl with Pfam GA cut off, and scanned against InterPro by IPRscan. Only putative TFs matching a described Pfam DBD, a DNA binding family, and transcription-related GO term were extracted. However, the collection thus obtained can still contain false positives such as proteins from the basal transcriptional apparatus (DNA polymerases), chromatin alterations, DNA packaging (histones), etc. To delete false positives, all putative TFs were searched against the non-redundant protein database and D. melanogaster TFs by BlastP (E- value cutoff e-6) and annotated with the functional description based on the protein homologue with the lowest E-value. If the function description of the TF contained a "keyword" not related to transcription factors (for example: polymerase, histone, splicing factor, etc.) and there was no homology to any of the *D. melanogaster* TFs [244], the TF was discarded. The remaining TFs homologous to D. melanogaster TFs were manually checked. Moreover, to obtain true negatives, we manually checked the *T. urticae* proteins not predicted as TFs but homologous to *D. melanogaster* TFs. Finally, gene structures of *T. urticae* TFs were manually curated.

2.6.3. Acknowledgements

M.G. and V.G. acknowledge support from NSERC Strategic Grant STPGP 322206-05, Marie Curie Incoming International Fellowship, OECD Co-operative Research Programme: Biological resource management for Sustainable Agricultural Systems JA00053351, and Ontario Research Fund–Global Leadership in Genomics and Life Sciences GL2-01-035. The genome and transcriptome sequencing projects were funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-046), JGI Community Sequencing Program grant 777506 to M.G., a University of Utah SEED grant (to R.M.C.), and National Science Foundation (NSF) grant 0820985 (to R.M.C., Principal Investigator L. Sieburth); work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under contract No. DE-AC02-05CH11231. Y.V.d.P. acknowledges support from the Belgian Federal Science Policy Office IUAP P6/25

(BioMaGNet), the Fund for Scientific Research Flanders (FWO), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), and Ghent University (MRP N2N). T.V.L. is a post-doctoral fellow of the FWO. We acknowledge the work of J. Boore, T. Negrave, A. Migeon, P. Auger, L. Swevers and H. Van Langenhove. M.G. and V.G. thank D. Weigel, G. Schäfer, M. Gerberding, R. Sommer, J. Felix and T. Nuernberger for discussions and support. The genome annotation of *T. urticae* is available at the VIB Department of Plant Systems Biology, Ghent University (http://bioinformatics.psb.ugent.be/orcae).

Chapter 3

The first chelicerate genome illustrates evolutionary innovation, fine tuning and adaptative plasticity in the arthropod chemoreceptor gene repertoire

Phuong Cao Thi Ngoc, Thomas Van Leeuwen, Robert Greenhalgh, Stephane Rombauts, Richard M. Clark, Miodrag Grbić, Yves Van de Peer and Pierre Rouzé

Manuscript under preparation

Author contribution

I performed the analyses and made the figures under the supervision of Pierre Rouzé and Yves Van de Peer. Richard M. Clark and Robert Greenhalgh helped with Figure A.7. Pierre Rouzé drafted the text with my contribution, especially parts describing the methods and results regarding the analyses.

3.1. Abstract

The spider mite *Tetranychus urticae* is one of most polyphagous arthropod herbivores. Chemosensory receptors that have important roles in chemical interation between spider mites and plants are potential targets in alternative plant protection strategies. Here we mined the *T. urticae* genome for putative chemosensory receptors, including the ones related to insect gustatory receptors (GRs), the ionotropic receptors (IRs) and the epithelial Na+ channels (ENaCs). While the number of TuIRs was lower than in insects, and comparable to *Drosophila* for TuENaCs, we identified a huge repertoire of 690 TuGRs, intact and pseudogenes, which is much more than the total number of GRs and odorant receptors (ORs) found to date in any other arthropod. When TuIRs are all expressed, significant level of transcription was observed for a few tens TuGRs only, and surprisingly some on the sense strand and others as anti-sense. Interestingly, several GR genes that are intact in some *T. urticae* populations appeared to be inactived in other populations. Added to the unusually large GR repertoire, this pseudogenization can be seen as a dynamic mechanism fine-tuning of the adaptation of *T. urticae* spider mites to their diverse environment with the ability of feeding on a uniquely large number of plant species.

3.2. Introduction

Chemoreception is the process by which animals perceive their environment and tune their behavior according to the smell and taste of chemicals they are confronted to. By recognizing chemical cues, they locate food sources, find mates, avoid predators and modulate communication with conspecifics [245, 246]. The olfactory system is best understood in vertebrates and insects showing a distinct organization in each group. While vertebrates use G-protein coupled receptors (GPCR) to initiate the intracellular signaling cascade leading to the downstream opening of ion channels (metabotropic signaling), insects are using an unrelated set of proteins that form a heteromeric complex of ionotropic receptors that are directly gated by odorants [246, 247]. Another distinction between these two groups is the strong difference in size of the gene repertoire encoding chemosensory receptors (CRs). Vertebrate species display a plethora of chemosensory receptor genes culminating in the elephant with ca. 4000 genes, including pseudogenes [248]. In contrast, insects display a smaller CR repertoire. The fruit fly genome only

revealed 62 ORs, 70 genes encoding gustatory receptor (GRs) [249, 250] and 66 for ionotropic receptors (IRs) related to ionotropic glutamate receptors (iGluRs) [251].

Our understanding of the evolution of olfaction and taste in arthropods is still incomplete and largely biased towards insects. GRs and ORs in arthropods are evolutionary related, encoding seven transmembrane (7TM) proteins with an inverted topology compared to GPCRs [252]. It was proposed that ORs were an insect-specific expansion of the GR family, as an adaptation to terrestriality from an aquatic ancestor [249]. This hypothesis is supported by the complete lack of Ors in the crustacean genome of the waterflee, *Daphnia pulex*, while a total of 58 Grs are documented [186]. The more recently discovered IRs fall within another family of chemoreceptors which are unrelated to ORs and GRs, being divergent homologs of iGluRs involved in sensory processes [192]. The olfactory system in animals is undergoing very dynamic evolutionary changes. The comparative study of genomes of 12 *Drosophila* species revealed that chemoreceptor evolution is dominated by a birth/death process with evidence of positive selection [250].

Insight from more basal arthropod taxons, chelicerates and myriapods, is needed to understand the evolution of olfaction and taste in arthropods. Chelicerates represent a basal arthropod taxon that diverged from other arthropod lineages more than 600MYA. While insects sense the volatile molecules with antenna and possess wings allowing them to travel a long-distance to find mates, food or oviposition sites, chelicerates have not evolved such traits, displaying a more primitive morphological architecture and limited mobility with possible repercussion to the evolution of chemosensing. Acari (ticks and mites) are the most diverse clade within the chelicerates (horseshoe crabs, scorpions, spiders, acari), with over 40,000 described species with different lifestyles, ranging from parasitic to predatory and plant feeding [253]. Within acari, the two-spotted spider mite *Tetranychus urticae* is a major agricultural pest of world-wide distribution that feeds on over 1,100 different plant species. Protecting crops against spider mites is a major challenge for agriculture. Acaricides to which resistance is building very fast in spider mites remain up to now the most efficient way to control infestation. A better understanding of spider mite-plant interaction was one of the main incentives for the sequencing of the *T. urticae* genome [253]. Chemosensory genes that recognize chemical

cues in the surroundings to locate food sources are obvious players in this host-pest interaction and potential targets in the search for alternative control strategies. Here we describe the exhaustive mining and annotation in the *T. urticae* genome of genes encoding chemosensory receptors (TuCRs), gathering homologs of pancrustacean GRs and IRs (TuGRs, TuIRs) together with homologs of *Drosophila* ppk genes encoding epithelial sodium channels (ENaCs), many of which having been shown to actually act as chemosensory receptors [254]. This led us to observe an important expansion of chemosensory receptors in this mite with a repertoire of 690TuGRs, exceeding the total number of GRs and ORs found to date in any other arthropod. We also report variation in pseudogenization of TuCRs between genetically distant *T. urticae* populations, revealing a dynamic evolutionary process within this TuCR family. Finally, using transcriptome profiling we observed significant expression of many TuCRs and more unexpectedly almost as often cis-antisense transcription as well.

3.3. Results

3.3.1. *T. urticae* has a huge repertoire of chemosensory receptors related to insect GRs

Iterative mining of the *T. urticae* genome initiated by documented GRs from arthropods and followed by manual checking, correction or implementation of gene models (see section 3.5.1) revealed a total of 690TuGRs, in which 449 intact genes, 219 pseudogenes and 22 partial genes (Table A.1). The intact TuGRs are predicted to be 7TM proteins with the expected N-inside:C-outside topology [252] and are the only gene products with this feature in the genome, showing the exhaustivity of TuGR mining. GRs being fast evolving proteins, most TuGRs did not show significant sequence similarity with pancrustacean GRs.

TuGRs (intact genes, exception for 10 highly diverse genes, and pseudogenes with single/two events having full C-terminals) were compared to each other by alignment of the more conserved three C-terminal transmembrane sequences, which allowed to construct a phylogenetic tree of TuGRs rooted with five sugar chemoreceptors from *D. melanogaster* (Figure A.1). Investigation of phylogenic relationships revealed that intact TuGRs generally cluster into two main groups, TuGR-A (Figure A.3) and TuGR-B

(Figure A.4), that are both very distantly related to pancrustacean GRs. In contrast, a few TuGRs which are more closely related to GRs from insects, daphnia, and *Ixodes scapularis* (Figure A.2) cluster independently of TuGR-A and TuGR-B (Figure 3.1). These TuGRs and highly diverse genes were group into TuGR-C. Number of intact genes, partial genes and pseudogene of each class was shown in Table 3.1. In addition to sequence similarity, genes in either TuGR-A or TuGR-B share other group-specific features, such as the number and position of introns, the length of the protein, and the occurrence of specific conserved residues in it (Table 3.2), which further indicates that they are part of two separate gene family expansions.

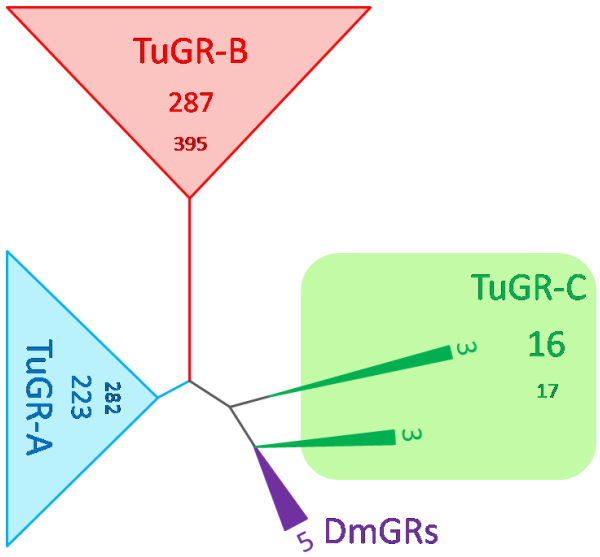


Figure 3.1. Phylogenetic tree of the TuGRs (intact genes – except the 10 most divergent ones in TuGR-C and pseudogenes with single or two mutation events) from *T. urticae*. The tree was rooted by five sugar receptors from *Drosophila melanogaster*. The total number of genes in each class including partial genes and pseudogenes is given in smaller fonts.

Table 3.1. Statistics of TuGRs in *T. urticae*

	TuGRs		TuGR-A		TuGR-B		TuGR-C	
	N	%	N	%	N	%	N	%
Total	690	100.0	279	40.4	394	57.1	17	2.5
Intact	449	65.1	188	67.4	245	62.2	16	94.1
Partial	22	3.2	6	2.2	16	4.1	0	0.0
Pseudogenes	219	31.7	85	30.5	133	33.8	1	5.9
> single event	64	29.2	28	32.9	36	27.1	0	0.0
> two events	22	10.0	10	11.8	12	9.0	0	0.0
> more/many	133	60.7	47	55.3	85	63.9	1	100.0

Table 3.2. Specific features in the two main groups TuGR-A and TuGR-B.
R212, E309, G323 refer to positions of specific residues in the sequence of TuGR274 taken as reference for TuGR-A group, and (102,103,104), G396 and E413 to positions and residues from TuGR100 taken as reference for TuGR-B.

	TuGR-A	TuGR-B
Number of intact genes	188	245
Average length	364 aa	432 aa
Number of introns	1 phase-0 intron conserved between TM-6 and TM-7 (I_2 on Figure 3.2) (exception: additional phase-2 intron at the N-terminus of TuGR253)	2 or 3 phase-0 intron conserved between TM-6 and TM-7 for all genes (I_1 on Figure 3.2) second phase-0 intron conserved in TM-7 for all genes (I_3 on Figure 3.2) third phase-0 intron conserved between TM-6 and TM-7 (I_2) for 211 genes fourth intron in 4 TuGRs
Conserved residues	1. R212 in the intracellular loop-2 between TM-4 and TM-5 (exception: TuGR310) 2. E309 in the next intracellular loop-3 before the intron (exception : 4 TuGRs with Q) 3. G323 in the same intracellular loop-3 (except A in TuGR226)	1. E413 in TM-7 (exception: TuGR51 with G). TM7 as a whole is rather conserved. 2. conserved triad (102,103,104) in the extracellular loop-1 between TM-1 and TM-2 3. G396 in the intracellular loop-3 is often conserved

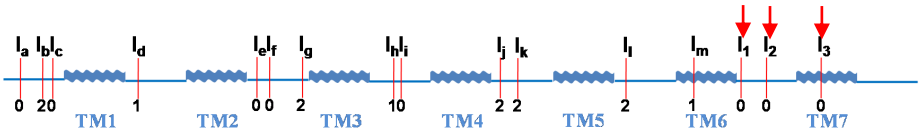


Figure 3.2. Location and phase of introns in the TuGRs of *T. urticae* with the 7TMs serving as topological reference. Phase-0 C-terminal introns I_1, I_2, and I_3 are conserved in TuGR-B group (except for a family including 30 TuGRs only with two conserved introns I_1 and I_3). Phase-0 C-terminal intron I_2 is conserved in TuGR-A group. Phase-0 I_f appears in 2 particular TuGRs. All other introns (I_a, I_b,…, I_m) appear only once in a particular TuGR.

Although TuGRs are very divergent in sequence with not a single conserved position when compared to GRs from other arthropods, it is noteworthy that the location of GR ancestral introns is conserved in TuGRs. Indeed, the three phase-0 introns located at the C-terminus of GRs from insect [249] and daphnia [186], in the last transmembrane helix (TM7) and in the upstream extracellular loop, are very likely homologous to the three phase-0 ones observed in TuGR-B (Figure 3.2), the middle one being the single phase-0 one found in TuGR-A. Lastly, the most divergent TuGR's, TuGR1, TuGR2, and TuGR3, the only TuGR's for which homologs are found in the genome of the tick *I. scapularis*, do resemble a group of pancrustacean GRs (Figure A.5) having the TYxxxxxQ motif in the last transmembrane helix, TM7. This sub-family is therefore suggested to be the most ancient, as it appears to be the only one conserved in all arthropods. It is important to note that no TuGR was identified with similarity to insect ORs and furthermore that no ortholog of the universal OR co-receptor from insects (Orco) was found in *T. urticae* genome. The complete lack of ORs was also observed in *D. pulex* [186] and is consistent with the hypothesis that ORs are an insect-specific class of GR-related chemosensory receptors [249, 255]. We also noticed extensive pseudogenization of chemosensory receptors in the *T. urticae* genome. 219 pseudo-genes were discovered (Table 3.1), indicating dynamic evolution and selection acting on the chemosensory gene families (see section 3.3.5).

The genes encoding GRs are dispersed all over the genome. However, many genes encoding GR-like TuGRs (449 TuGRs out of 690) are indeed found in 50 clusters (25 clusters without intervening genes, 25 clusters with 1, 2, 3, and 4 intervening genes)

ranging in size from 2 to 39 genes indicating gene proliferation driven by tandem duplication. The largest cluster on scaffold 17 with 39 TuGRs and 4 non-TuGRs was ~158kb long. Other clusters had a size in the range 10kb to 79kb. Interestingly, 59 TuGR genes are nested in large introns of 19 hosting genes, ranging in number from one 1 to 17 genes per host (tetur08g08289).

3.3.2. *T. urticae* has fewer IRs chemosensory receptors than insects

Ionotropic Receptors (IRs) are divergent variants of ionotropic Glutamate Receptors (iGluR) which are not anymore implicated in synaptic transmission but in chemosensing in protostomes [251, 192]. Mining of the iGluR family in the spider mite genome uncovered 19 members, including one pseudogene (Table A.3). A phylogenetic comparison of these proteins with iGluRs and IRs from arthropods clearly shows that four members in this family are closely related to known chemosensory IRs, with TuIR1, TuIR3 and TuIR4 being related to IR25a, and TuIR2 being related to IR93a (Figure 3.3). The repertoire of IRs in *T. urticae* is thus smaller than the 66 IRs identified in *D. melanogaster*. This corroborates the suggestion that no chelicerate-specific expansion of IRs has occurred [251] and fits with the observation that IR25a and IR93a are the most ancient IR members, the only ones to be found in *Daphnia* as well [251].

Similarly, the search for metabotropic glutamate receptors (mGluRs) revealed the existence of 4 members in this family, one in each group (I, II & III) of neurotransmitter mGluR's and one related to the L-canavanine receptor XR from insects [191] (Figure 3.4); the latter homology suggests that *T. urticae* is using this TuXR to sense L-canavanine, a toxic amino acid not only stored primarily in some legume seeds but also found in all tissues of the living plant.

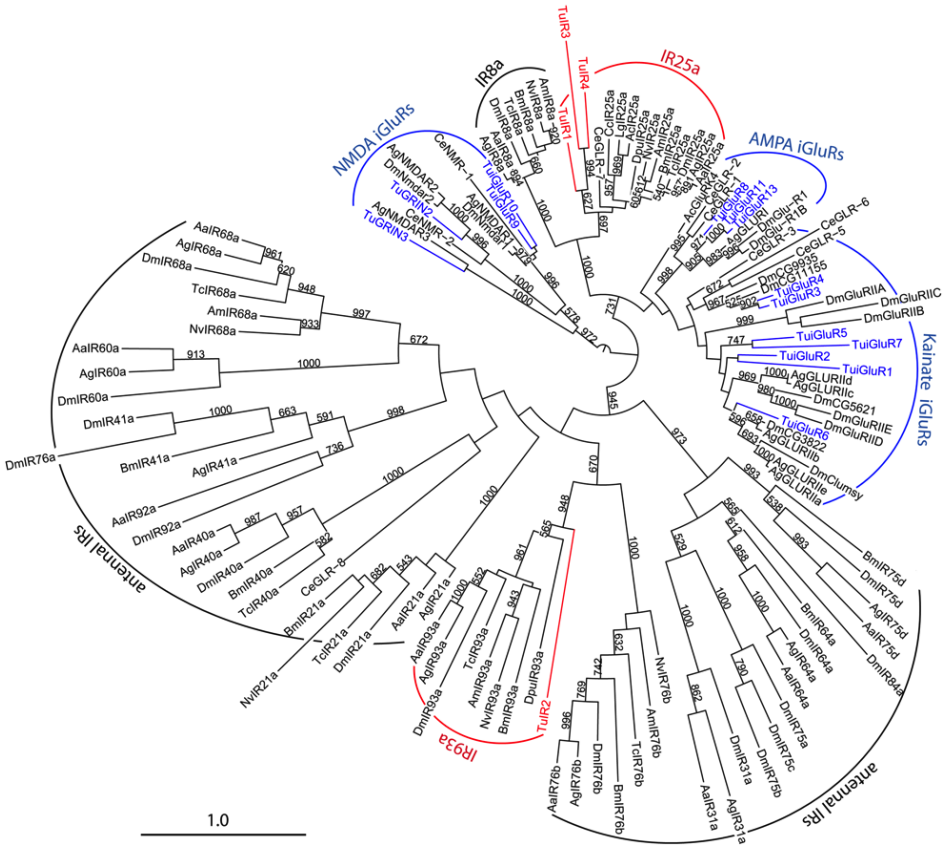


Figure 3.3. Evolutionary relationships of ionotropic glutamate receptors (iGluRs, blue) and their related chemosensory receptors (IRs, red) in *T. urticae* and in a few protostome species (Aa – *Aedes aegypti*, Ac – *Aplysia california*, Ag – *Anopheles gambiae*, Am – *Apis mellifera*, Bm – *Bombyx mori*, Cc – *Capitella capitata*, Ce – *Caenorhabditis elegans*, Dm – *Drosophila melanogaster*, Dpu – *Daphnia pulex*, Lg – *Lottia gigantea*, Nv – *Nasonia vitripennis*, Tc – *Tribolium castaneum*, Tu – *Tetranychus urticae*). The phylogenetic tree was built by PhyML with bootstrap 1000 and rooted with NMDA receptors.

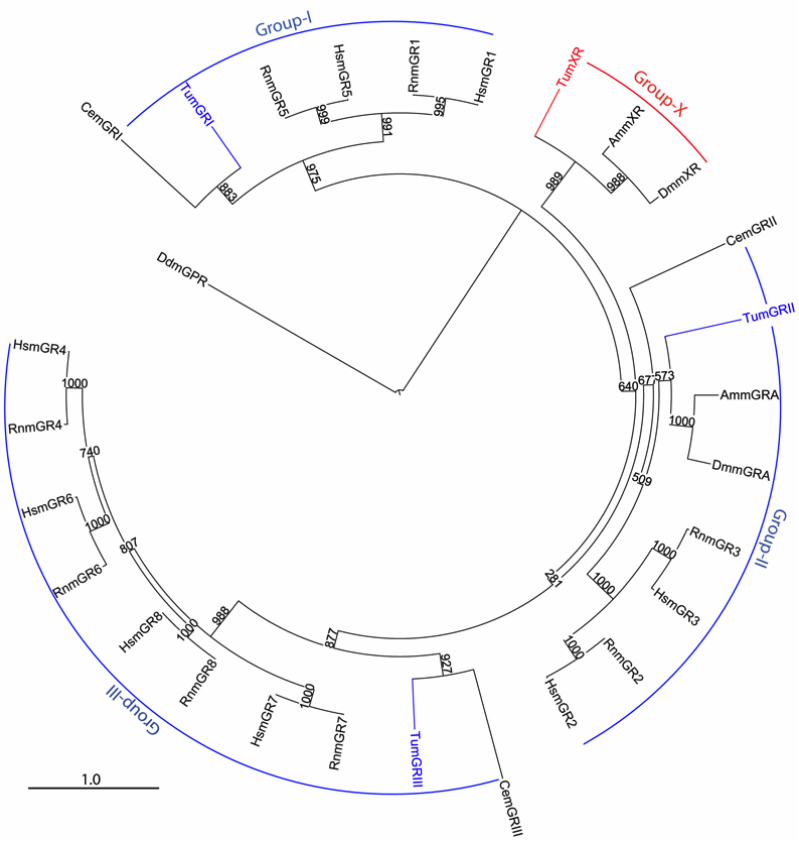


Figure 3.4. Evolutionary relationships of mGluRs (blue) and mXRs (red) in *T. urticae* and in other representative species (Am – *Apis mellifera*, Ce – *Caenorhabditis elegans*, Dd – *Dictyostelium discoideum*, Dm – *Drosophila melanogaster*, Hs – *Homo sapiens*, Rn – *Rattus norvegicus*, Tu – *Tetranychus urticae*). The phylogenetic tree was built by PhyML with bootstrap 1000 and rooted with DdmGPR.

3.3.3. Gene expansion of ENaCs in *T. urticae* unveils candidates for chemosensory function

Epithelial Na+ Channels (ENaCs) represent a gene family of ion channels involved in various cell functions in metazoans including cell volume regulation, nociception, mechanosensation and taste perception as well [256]. For this reason we mined the *T.urticae* genome for ENaCs using members of the *D. melanogaster* and *C. elegans* ENaC protein family, ending up in the finding of 27 TuENaCs (Table A.2), including 2 pseudogenes and 2 older relics. Interestingly, TuENaCs are clustered in two separate groups, a group of two proteins related to *C. elegans* UNC8 and UNC105 with a large

extracellular loop, and a second group with 23 more divergent members (TuENaC01-23). Alignment with ASIC1, an acid sensor for which a 3D structure has been established [257], allows the domains and functionally important residues of these TuENaC01-23 proteins to be unambiguously identified.

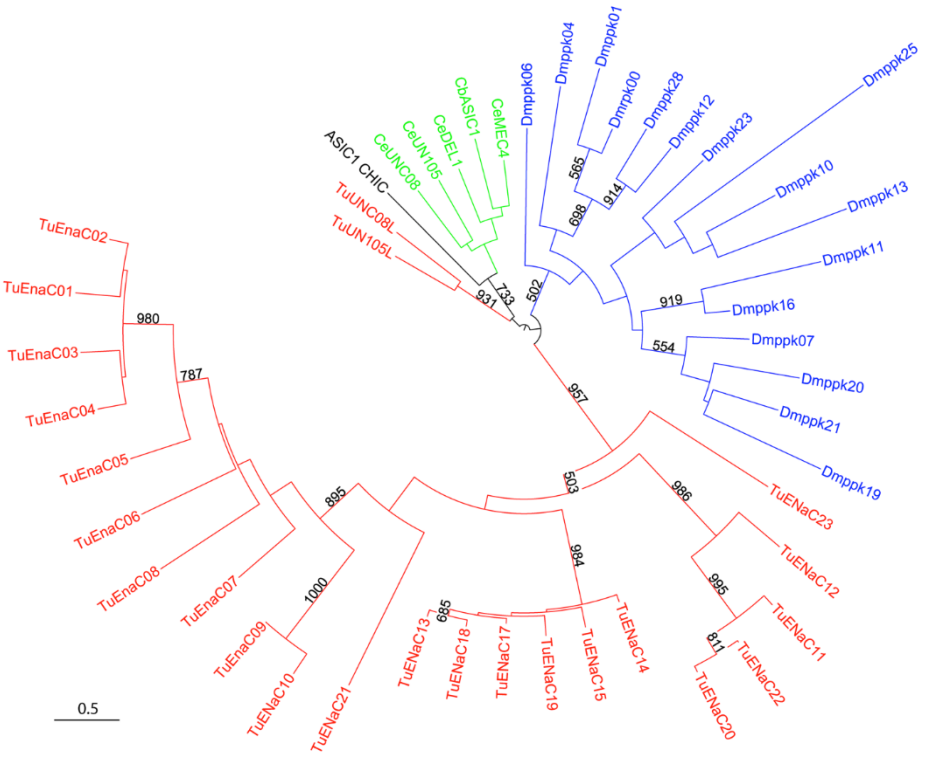


Figure 3.5. Phylogenetic tree of ENaCs from *T. urticae* (Tu) compared to ppk/ENaCs from *D. melanogaster* (Dm), selected ENaCs from *C. elegans* (Ce), *C. briggsae* (Cb) and ASIC1 from Chicken. The phylogenetic tree was rooted by midpoint rooting.

A phylogenetic reconstruction based on the conserved trans-membrane domains of ENaCs (Figure 3.5) shows that the TuENaC01-23 group resulted from a gene expansion that is reminiscent to the one observed for *D. melanogaster* PPKs, but which occurred independently. In *C. elegans* UNC8 and UNC105 are involved in movement coordination [258], suggesting a similar role in mechanosensing or proprioception for their orthologs in *T. urticae*. Conversely, it is likely that some if not all the proteins in the TuENaC01-23 group would be chemosensory receptors, as shown for several members of the

pickpocket PPK family in *Drosophila* [254]. Seven TuENaCs out of 27 are found on the genome in gene clusters indicating the expansion of this gene family by tandem duplication, as for TuGRs.

3.3.4. Chemosensory receptor sense and antisense expression

T. urticae is known to adapt to different host plants but it remains unknown how is the host switching regulated at the molecular level. To test whether transfer from the original host plant (bean) on which the London population of the spider mite was grown to different host plants (Arabidopsis and tomato) would affect gene expression we analyzed the expression of the chemosensory genes 12 hours after the *T. urticae* larval transfer. In addition transcript levels have been measured in spider mites at various developmental stages [253]. From these experimental data the relative expression of candidate chemosensory receptor genes was investigated (Table A.2, Table A.3). Whereas TuIRs are all expressed, albeit at low levels, only a few tens of the GR-like genes showed significant expression, with the vast majority of the other GR genes showing undetectable transcript levels. This data represent a conservative estimate of gene expression, since we used whole animals for the transcriptome profiling. It is known that the chemoreceptors are expressed only in individual neurons at the low level and analysis of whole animals inevitably dilutes the expression level. Interestingly, expression of some of these genes was induced following transfer to different host plants. For example, tetur37g00370 (TuGR251) showed increased expression when mites were fed on tomato (Table A.3). Besides sense expression we often observed clear-cut cases of expression of antisense transcripts overlapping TuGR genes on the opposite strand (Figure A.6). Initially, antisense transcription was discovered in bacteria [259], then found in eukaryotes [260, 261]. It is being recognized as an important regulator of gene expression through the act of transcription or through the non-coding RNA that is produced [262]. Antisense transcripts might be part of self-regulatory circuits that allow sense genes to regulate their own expression [262]. It is noteworthy that in most cases sense and *cis*-antisense expression are mutually exclusive, i.e. either the TuGR transcript (sense transcript) or antisense transcript is significantly expressed (Figure A.7). To our knowledge, expression of antisense transcripts has not been reported before for chemosensory receptors. If we accept the suggestion that antisense expression would cause the silencing of the TuGR

sense gene, one would expect reducting or abolishing the ability of the mite to recognize a chemical even if present as a consequence. Such a combination of positive and negative transcript regulation may provide a sophisticated fine tuning for chemical reception, allowing or forbidding sensing to occur under specific environmental or developmental stimuli.

3.3.5. Fast evolution of chemosensory receptors in *T. urticae*

It is known that chemosensory receptor genes undergo rapid evolution [245, 246, 263] with genes being gained, lost or pseudogenized and diverging fast in closely related species, ending up in loosing or gaining capability to sense a peculiar chemical in the specific environment of a given species. The genome analysis of 12 Drosophilid species inhabiting different habitats revealed e.g. that the OR and GR genes vary greatly in number, experiencing lineage-specific duplication and pseudogenization. Most loci undergo purifying selection, only a few being positively selected. It was also observed that genes which are the most often duplicated had relaxed constraints, allowing evolutionary divergence in receptor function [250]. However, it remains unknown how CR genes are evolving within large populations of a single species. *T. urticae* might be an interesting study subject to address this question, as large genetically distant populations adapted to different host plants can be easily collected. Thus, in addition to the genome sequencing of the London *T. urticae* population (isolated in Canada, originally from an apple orchard) we used resequenced genomes of two other strains of *T. urticae*, the Montpellier strain (from Scotland, originally collected on strawberries) and the EtoxR strain (from Japan), and the reads were mapped onto the assembled scaffolds from the London strain, which was used for the initial genome initiative [253]. A search was done to identify and assemble the orthologs of each CR gene in these two strains. These genes were then compared between the three strains, paying special attention to nucleotide changes that would end up in turning a gene into a pseudogene or restore functionality of a pseudogene (Table A.4). Surprisingly, twenty-one TuGRs that were pseudogenes in the London strain appear to be intact genes in either the Montpellier strain or the EtoxR strain, or in both. In a few cases, allelic variation was even observed within these strains (which are not inbred), i.e. being intact genes for some reads and pseudogenes for others. Conversely, ten valid TuGRs genes in the London strain appear

to be pseudogenes or to show allelic variation in the Montpellier and EtoxR strains. Similarly, ENaCs also show allelic variation. Indeed, ENaC13 is an intact gene in the London strain and ENaC18 is a pseudogene, whereas both show allelic variation in the Montpellier strain, with a mix of pseudogenes and intact genes. Similar pseudogenization events have been observed for chemosensory receptors in *Drosophila* [264] and vertebrates [265, 266] and a discussion is ongoing on the effects of such allelic variation on feeding ecology. Nevertheless, the variations described in vertebrates occur among species that diverged 20MYA or more, whereas here we report on pseudogenization events that affect several genes in different populations within the same species, which suggests that such variability could be a dynamic mechanism that allows the adaptation of *T. urticae* isolates to various environmental constraints.

3.4. Discussion

This investigation into chemosensory receptors of *T. urticae* from genome sequence data of the London strain, from sequence data from two other strains and from RNA-seq expression data allowed to build not only an exhaustive repertoire of putative CRs, as it has been done previously from genome mining of for several species of insects, but also to have some insight into short-term evolution in populations adapted to a given environment and to which genes are actually expressed at different developmental stages and when feeding on different host plants.

The results we gathered are interesting in several respects. First it appears that the spider mite has more CRs than insects and any other arthropod up to now. This suggests that chemical sensing does play an important role in the behavior of these mites possibly related to the specific life style of *T. urticae*, which is able to adapt to diverse environmental niches by feeding on a uniquely large number of plant species. It now matters to understand which molecules are ligands of these many CRs and which specific cells and organs are expressing these CRs. We confirmed the absence of OR genes sensu insecta. This nevertheless does not mean that among the many TuGRs that we discovered, some if not many, would not be receptors for odorant molecules.

We observed that if many CR genes are expressed, only a few do so at high levels. TuGR6 and TuGR253, which are among the ones with highest expression in almost all tested conditions, may be encoding proteins with a structural role, being e.g. subunits in multimeric chemosensory receptors, similarly to the one played by ORco for insect dimeric odorant receptors. As far as expression is concerned, the discovery of many antisense transcripts overlapping CR genes and their observed significant expression is an important new feature. In most cases those transcripts are not from neighboring protein-encoding genes but from *cis*-acting non-coding RNA genes, suggesting that their function would be the specific regulation of the cognate GR gene. This may suggest that the transcription of GR genes is under the control of sense and antisense regulation.

It is striking to see that chemosensory receptors do show variation at the level of population, with CR genes coding for active receptors in some populations and inactive in others. It raises interesting questions on the adaptive function of these gains and losses, and especially if it has something to do with host choice. For a polyphagous species like *T. urticae* that has been shown to be able to feed on more than 1,100 host plants, it matters to understand if there would be sub-species variation or if we are dealing with a plastic continuum of individuals which can quickly adapt as a population if a suitable niche would build up. We are currently addressing this question at the species level, comparing the genome of *T. urticae* to the ones of *Tetranychus* species that are oligophagous or monophagous.

3.5. Materials and Methods

3.5.1. Chemosensory receptor gene annotation

Daphnia and insect gustatory and odorant receptors whose sequences have been entered into Genbank were used to perform TBlastN searches (E-value < 1) for similar regions in the *T. urticae* genome. From these regions *T. urticae* CR gene models were checked, corrected or constructed through GenomeView [267] and updated manually using ORCAE. To find CRs exhaustively, annotated CRs were used in iterative rounds of TBlastN searches.

To search for iGluRs, the PF00060 domain from Pfam database [268], was used to scan the *T. urticae* proteins with HMMER [269]. All significant hits with E-value < e^{-5} have be assigned to iGluRs or IRs of *T. urticae*. No further candidate was returned through TBlastN with the identified genes. mGluRs were retrieved using the IRs identified by the Benton team [192] as baits for BlastP and TBlastN, together with a few iGluRs from insects and vertebrates. Similarly ENaCs were mined using *D. melanogaster* PPKs and a range of ENaCs from vertebrates and invertebrates. All the identified CR gene models as well as the complete gene annotation for the *T. urticae* (London) genome sequence [253] are available through the ORCAE website (http://bioinformatics.psb.ugent.be/ORCAE/) [150].

3.5.2. Phylogenetic analysis

The GR-like CRs from *T. urticae* and five sugar receptors from *D. melanogaster* were aligned using ClustalW [270]. The alignment was annotated based on the seven trans-membrane domains predicted by TMHMM [271] as well as on intron positions, and edited in Jalview [272]. The divergent N-terminal region (TM1-4), the short C-terminal tail as well as the major gaps between trans-membranes were removed to obtain the final alignment for tree construction, based on 117 unambiguously aligned sequence positions. To build the phylogenetic tree for this highly divergent protein family, we applied a similar procedure as published previously for insects and *Daphnia* [249, 186]. Amino acid distances were corrected for multiple amino acid replacements by using the BLOSUM62 amino acid exchange matrix in TREE-PUZZLE v5.0 [273]. The heuristic search based on corrected distances with tree-bisection and reconnection branch swapping in PAUP*v4 [274] was used to build the phylogenetic tree. Bootstrap analysis was performed by analyzing 1000 neighbor-joining replications with uncorrected distances.

The sequences of iGluRs/IRs in *T. urticae* and in protostome species [192] were aligned by PROBCONS [275] and edited in Jalview. Trans-membranes PFAM domains (PF00060, PF10613, PF01094) were edited manually to obtain the highly conserved C–terminal region. The sequences of mGluRs/mXRs in *T. urticae*, *D. melanogaster*, *A. mellifera*, *C. elegans*, *H. sapiens*, *R. norvegicus* and DdmGluPR from *D. discoideum* were

aligned by PROBCONS. This last one was included because phylogenetic analysis suggested that DdmGluPR diverged after the mGluR family-GABAB receptors split but before mGluR family divergence [276]. Similarly, the phylogenetic tree was also built from the edited alignment of ENaCs from *T. urticae*, ppk/ENaCs from *D. melanogaster*, selected ENaCs from *C. elegans*, *C. briggsae*, and ASIC1 from chicken. The phylogenetics trees of GluRs and ENaCs were built by PhyML [277] with bootstrap set to 1000.

Chapter 4

The molecular evolution of chemoreceptors related to insect gustatory receptors in *Tetranychus urticae*, *Tetranychus evansi* and *Tetranychus lintearius* spider mites

Phuong Cao Thi Ngoc, Stephane Rombauts, Pierre Rouzé and Yves Van de Peer

Author contribution

I performed all analyses and made the figures, and wrote the chapter under the supervision of Pierre Rouzé and Yves Van de Peer and with the contribution of Stephane Rombauts for section 4.5.1.

4.1. Abstract

Chemosensory receptors play an important role in the chemical interaction of animals with their environment. Here we analyzed the size of the repertoires of chemoreceptors related to insect gustatory receptors (GRs) among the genomes of three species of spider mites, the polyphagous *Tetranychus urticae*, the oligophagous *Tetranychus evansi*, and the monophagous *Tetranychus lintearius*. We identified 226 GRs in *T. evansi*, 257 GRs in *T. lintearius*, compared to 690 GRs in *T. urticae* (Chapter 3). From the common ancestor of these mites, the GR gene family has evolved though gene duplication, gene loss and pseudogenization. The many gene duplications that occurred in *T. urticae* might be related to the ability of this species to feed on more than a thousand plant hosts. On the opposite, many genes losses occurred in the closely related *T. lintearius*, which might explain why this species only feeds on common gorse. In the bit more distant *T. evansi*, many gene losses and a few gene duplications have maintained the GR repertoire in the range of the sum of insect GRs and ORs.

4.2. Introduction

All animals detect certain chemical components in their environment via chemosensory receptors (CRs) to find food, to locate shelter mates and offspring, and to avoid danger [245]. As olfactory receptors (ORs) in vertebrates [187] and in nematodes [188] were discovered to be members of the superfamily of G-protein-coupled receptors (GPCRs) with seven trans-membrane domains, insect ORs and gustatory receptors (GRs), which are distantly related to each other, were initially thought to be novel GPCRs which later on appeared to be wrong [184, 185, 249]. This family indeed not only shares no sequence similarity to vertebrate and nematode ORs, but more importantly has a reversed seven trans-membrane domain topology, as compared to GPCRs, with a N-terminus located intracellularly and a C-terminus located extracellularly [252]. As a consequence, the biochemical function of ORs (and GRs) and the signaling mechanism downstream differ in between vertebrates (plus nematodes) and insects. While most chemoreceptors in mammals, including ORs, are slow acting metabotropic receptors indirectly activating ion channels through second messengers, chemoreception in insects is ionotropic, receptors including ORs and GRs being ligand-gated ion channels allowing a much faster reaction [189]. Besides insects, GRs, but no ORs were identified in the crustacean

Daphnia pulex, fitting with the view that OR is a lineage evolved in insects or their hexapod ancestors. The numbers of CRs vary enormously among the genomes of different animal species [245]. This variation was explained by positive selection in adaptation of organisms to different environments and by genomic drift, a random process of gene duplication and deletion. In insects, the population size of CRs can be used as an index of the complexity in insect-environment interaction [278]. *Drosophila sechellia*, a specialist on the fruit of *Morinda citrifolia*, has lost several functional CRs as a consequence of host specialization [279].

Spider mites (Tetranychidae) belong to chelicerates, the second largest group of arthropods after the insects. This family comprises 1,250 phytophagous species forming an important group in agriculture with more than a hundred of them as pests and about ten as major pests [280]. Among major pests, the two-spotted spider mite, *Tetranychus urticae*, is an extremely polyphagous mite able to feed on more than 1,100 plant species from more than 140 different plant families [210]. On the contrary, *Tetranychus evansi* [281] is an oligophagous mite feeding mainly on *solanaceous* plants, but also reported occasionally on plants of 37 other families. *Tetranychus lintearius* [282] is a monophagous mite feeding only on common gorse plants (*Ulex europaeus*). *Tetranychus evansi* is an increasing concern for agriculture because being native from South America, it now spreads to various parts of the world including southern USA, sub-Saharan Africa, Mediterranean Basin and East Asia and affects the production of crops such as tomatoes [281].

Tetranychus urticae, is being used as a chelicerate genomic model because of its small genome size, short generation time, the easy maintenance in the laboratory, and its economical impact as a cosmopolitan agricultural pest [240]. The completely sequenced and annotated *T. urticae* genome, representing the first complete chelicerate genome, offers new insight into arthropod evolution and plant-herbivore interaction [253]. When compared to insects, gene families associated with feeding on different hosts were expanded in *T. urticae*, including cytochrome P450, carboxyl/cholinesterases, ATP-binding cassette transporters, and glutathione S-transferases. In addition, *T. urticae* has many more chemoreceptors related to insect gustatory receptors (GRs) (690 genes) than

any other arthropod up to now (Chapter 3). This seems to indicate that chemical sensing does play an important role in the interaction of *T. urticae* with its environment in with a range of more than 1,100 host plants. Here we performed a comparative analysis on the chemoreceptors of these *Tetranychus* species to study the ecological adaptation of *T. urticae*, *T. evansi*, and *T. lintearius*. We did not include IRs in the analysis because there are only four IRs in each *Tetranychus* species. We identified 226 GRs (168 intact genes, 58 pseudogenes/partial genes) in the genome of *T. evansi* and 257 GRs (127 intact genes, 130 pseudogenes/partial genes) in the genome of *T. lintearius*.

4.3. Results and discussion

4.3.1. Genome annotation of *T. evansi* and *T. lintearius*

The overall assembly sizes of *T. evansi* and *T. lintearius* are 91.5Mb and 91.2Mb, respectively (see Materials and Methods, section 4.5.1 and Table 4.1). Phylogeny of the three spider mites shows that *T. urticae* and *T. lintearius* are closely related species where *T. evansi* is more distantly related (Figure 4.1) with the following speciation dates: 2.9MYA for *T. evansi*/*T. urticae*, 3.0MYA for *T. evansi*/*T. lintearius*, and 0.85MYA for *T. urticae*/*T. lintearius* (Toni Gabaldon, personal communication).

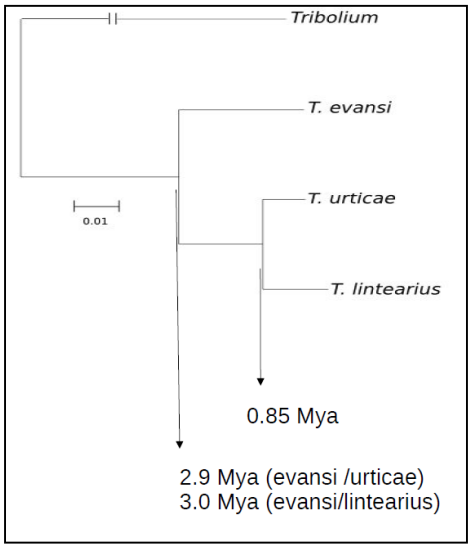


Figure 4.1. Phylogeny of three spider mites (taken from Toni Gabaldon at the Ibiza meeting of the Spider Mite Consortium, personal communication)

Table 4.1. Genome comparison and annotation statistics for the spider mite genomes of *T. evansi*, *T. lintearius*, and *T. urticae* (ORCAE [150]: 04/10/2013)

		T. evansi*	***T. lintearius****	***T. urticae***
Genome size (scaffolds)	Nt	91,474,574	91,247,315	90,815,494
Number of scaffold		3,318	2,610	640
Largest scaffold	Nt	1,473,105	1,693,225	7,801,961
Scaffold N50		79	70	9
Av. length of scaffolds	Nt	27,704	35,201	141,899
gaps (>50N)		5,641	4,497	1,395
Nr. of genes		**18,230**	**17,551**	**18,550**
Gene density	genes/Mb	198.32	192.58	204.26
Av. length of genes	Nt	1,010	988	1,438
Median length of genes	Nt	678	666	1,157
Nr of Exons		63,893	55,073	70,584
cumul. exon length	Nt	18,404,262	17,339,732	26,675,051
%cumul. exon length		20.02	19.03	29.37
Av. length of exons	Nt	288	315	378
Median length of exons	Nt	150	174	189
Longest exon	Nt	24,195	9,307	45,659
Av.nr. exons/gene		4	3	4
most exons/gene		41	36	55
cumul. CDS length	Nt	16,943,354	17,339,732	20,081,908
%cumul. CDS length		18.43	19.03	22.11
Av. length of CDSs	Nt	929	988	1,083
%GC of CDS		37.15	37.48	37.57
cumul. intron length	Nt	25,412,707	18,119,172	23,322,223
%cumul. intron length		27.65	19.88	25.68
Av. length of introns	Nt	557	483	448
Median length of introns	Nt	113	102	96
Longest intron	Nt	52,739	69,917	59,291
%GC of intron		28.44	29.13	29.79

*The genomes *T. evansi* and *T. lintearius* are still under manual curation.

We predicted (see Materials and Methods, section 4.5.2) 17,551 and 18,230 protein coding genes in *T. evansi* and *T. lintearius*, respectively (Table 4.1). The complete genome annotations of these genomes are available at the ORCAE website [150]. The size of both genomes and proteomes is about the same as the ones for *T. urticae* (Table

4.1). The genes of *T. evansi* and *T. lintearius* are still under manual curation by the Spider Mite Consortium.

4.3.2. The GR repertoire in *T. evansi* and *T. lintearius*

Based on *T. urticae* GRs, we identified a total of 226 GR genes in *T. evansi* (168 intact genes, 58 pseudogenes/partial genes) (Table A.5), and 257 GR genes in *T. lintearius* (127 intact genes, 130 pseudogenes/partial genes) (Table A.6). The size of the GR repertoires (intact genes, pseudogenes and partial genes) of three spider mites is shown in Table 4.2. The total number of GRs in the polyphagous *T. urticae* is more than twice the total number of GRs in the monophagous *T. lintearius* and three times the total number of GRs in the oligophagous *T. evansi*. The total number of GRs in *T. lintearius* is higher than in *T. evansi*, but 45% of these are pseudogenes and the number of intact GR genes is then lower in *T. lintearius* than in *T. evansi*.

Similarly to *T. urticae* GRs, GRs in *T. evansi* and in *T. lintearius* could be subdivided into three A, B, and C classes with the same typical features (see Chapter 3): TeGR-A & TlGR-A, TeGR-B & TlGR-B, and TeGR-C & TlGR-C (Table 4.2). Sixteen genes of class C in three spider mites have a 1:1:1 orthologous relationship (Table 4.3). Class B contains more than 50% of the total number of GRs in each species. The variation in GR number is much bigger in class A with a ratio of intact GRs (Tu:Te:Tl) being roughly 6:2:1 for class A when the ratio is only 12:5:4 for class B. This means that class A is under stronger selection pressure than class B in three spider mites. Interestingly, the variation in number of pseudogenes differs from the one of intact genes with Tu:Te:Tl ratios of 5:1:4 for class A, and 5:1:2 for class B. On average, there are two intact TuGRs for each pseudogene in *T. urticae*, for both GR-A and GR-B classes, but a higher ratio of intact genes vs. pseudogenes in *T. evansi*, especially in the GR-B class. On the reverse, *T. lintearius* has a higher ratio of pseudogenes, especially in the GR-A class (Table 4.2). The tendancy for TeGRs and TlGRs to be more often turned into pseudogenes in GR-A compared to GR-B fits with the previous observation that class A is under stronger selection pressure than class B, with selection operating both at the number of genes and in pseudogenization.

Table 4.2. Comparison on number of GRs in *T. urticae*, *T. evansi* and *T. lintearius*

	T. urticae		***T. evansi***		***T. lintearius***	
	N	%	N	%	N	%
Total of GRs	690	100.0	226	100	257	100
GR-A	279	40.4	64	28.3	95	37.0
GR-B	394	57.1	146	64.6	145	56.4
GR-C	17	2.5	16	7.1	17	6.6
Intact	449	40.4	168	74.3	127	49.4
GR-A	188	41.9	46	27.4	29	22.8
GR-B	245	54.6	107	63.7	84	66.1
GR-C	16	3.6	15	8.9	14	11.0
Partial	22	3.2	15	6.6	13	5.1
GR-A	6	27.3	2	13.3	2	15.4
GR-B	16	72.7	13	86.7	10	76.9
GR-C	0	0.0	0	0.0	1	7.7
Pseudogenes	219	31.7	43	19.0	117	45.5
GR-A	85	38.8	16	37.2	64	54.7
GR-B	133	60.7	26	60.5	51	43.6
GR-C	1	0.5	1	2.3	2	1.7
Intact/pseudogenes	2.1		3.9		1.1	
GR-A	2.2		2.9		0.5	
GR-B	1.8		4.1		1.6	

Table 4.3. GR Orthology in the GR-C class
*: pseudogene, @: partial gene.

	GR ID	*T. urticae*	*T. evansi*	*T. lintearius*
1	GR1	tetur19g02720	tetev08g01380	tetli22g01240
2	GR2	tetur06g01480	tetev183g00090	tetli41g00100
3	GR3	tetur07g04370	tetev14g01020	tetli88g00280
4	GR4	tetur02g09560	tetev79g00250	tetli92g01140
5	GR5	tetur11g00460	tetev175g00110	tetli221g00110@
6	GR33	tetur06g00770	tetev100g00420	tetli44g00420*
7	GR210	tetur01g06380	tetev66g00500	tetli57g00140
8	GR212	tetur20g00810	tetev88g00910	tetli116g00730
9	GR213	tetur03g02210	tetev07g00290*@	tetli01g00170
10	GR214	tetur10g05170	tetev18g00660	tetli142g00170
11	GR215	tetur22g02600	tetev377g00080	tetli117g00390
12	GR216	tetur05g00780	tetev56g02470	tetli12g01470
13	GR510	tetur11g00450	tetev175g00120	tetli221g00100
14	GR511	tetur05g08450	tetev49g00320	tetli143g00170
15	GR528	tetur04g06520	tetev44g00750	tetli07g00450
16	GR543	tetur11g06470	tetev175g00300	tetli221g00075

*the frameshift in tetli44g00420 can be an intact gene because of sequence error.

4.3.3. Phylogenetic trees of insect GR-like chemoreceptors in the three spider mites

Because of the high divergence of GRs, it is difficult to align all GRs from three spider mites and build a single exhaustive tree. Therefore, we built two trees; one for class A and one for class B. The tree for class C was excluded because of its highly diverse genes with 1:1:1 orthologous relationship. For class A, we built a phylogenetic tree including 188 intact genes, and 35 pseudogenes (the ones for which the sequence of their intact copy could safely be anticipated) from *T. urticae*; 46 intact genes, and 1 pseudogene from *T. evansi*; and 29 intact genes, and 1 pseudogene from *T. lintearius* to build the phylogenetic tree (Figure A.8). Similarly, for class B, we built a tree including 245 intact genes, 42 pseudogenes from *T. urticae*; 107 intact genes, 2 pseudogenes from *T. evansi*; and 84 intact genes and 1 pseudogene from *T. lintearius* (Figure A.9).

Different evolutionary trends can be observed in these trees for GRs in *T. urticae*, *T. evansi* and *T. lintearius*. An individual GR gene can indeed be conserved as a single copy in all three species, i.e. with Tu:Te:Tl orthology being (1:1:1), or lost in one or several other species (1:1:0, 1:0:1, 1:0:0,...), or duplicated in one or several other species too (N:1:0, N:0:1, N:N:0, N:0:0, N:N:N, N:N:1,…) as exemplified in (Figure 4.2). We counted gene duplication, pseudogenization, and gene loss in each species, according to parsimony analysis for each subtree manually (see Materials and Methods, section 4.5.4). Following this analysis, the predicted gain and loss patterns for GR-A and GR-B are shown in Figure 4.3. This analysis reveals that the common ancestor of the three spider mites was likely having a total of ca. 270 GRs. From this ancestor, *T. evansi* and the ancestor of *T. lintearius* and *T. urticae* experienced a few cases of GR expansion (ca. 30). In *T. evansi* 130 GRs were then lost or pseudogenized. From their common ancestor, *T. urticae* and *T. lintearius* evolved into opposite directions, with *T. urticae* almost doubling his GR repertoire and losing only a few, whereas *T. lintearius* gaining no GR and losing or pseudogenizing 149 GRs. It should be noted that expansions in *T. evansi* and *T. urticae* happened independently.

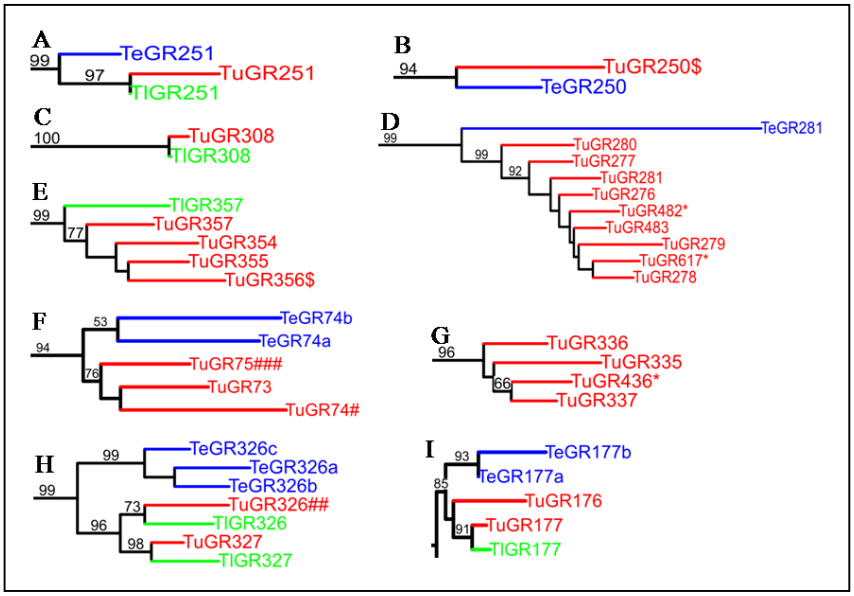


Figure 4.2. The different evolutionary scenarios of GRs in *T. urticae*, *T. evansi*, and *T. lintearius*.

GR with suffix "*" after protein name is pseudogene. TuGR and TeGR with suffix "#" after protein name has homologous pseudogene(s)/partial gene(s) with many changes (blast cutoff 1e-25) in *T. evansi* and *T. urticae*, respectively. TuGR and TlGR with suffix "$" after protein name has homologous pseudogene(s)/partial gene(s) with many changes (blast cutoff 1e-25) in *T. lintearius* and *T. urticae*, respectively.

A, 1:1:1. The GR251 gene is conserved as a single copy in the 3 species. Note that the gene tree fits with the species tree, i.e. TeGR branching earlier than TuGR and TlGR.
B, 1:1:0. The GR250 gene has been pseudogenizated in *T. lintearius*.
C, 1:0:1. The GR308 gene has been lost in *T. evansi*.
D, N:1:0. The GR281 has been lost in *T. lintearius* and expanded in nine copies in *T. urticae*, two of which being pseudogenes.
E, N:0:1. The GR357 has been lost in *T. evansi* and expanded in four copies in *T. urticae*.
F, N:N:0. The GR74 has been duplicated in *T. evansi* and independently triplicated in *T. urticae* but lost in *T. lintearius*.
G, N:0:0. The GR336 has been expanded into 4 copies in *T. urticae*, one of which being a pseudogene.
H, N:N:N. A rare case where GR336 has been independently duplicated in *T. lintearius* and in *T. urticae* and triplicated in *T. evansi*.
I, N:N:1. The GR177 has been duplicated in *T. evansi* and has been independently duplicated in the ancestor of *T. lintearius* and *T. urticae*, with one copy being lost later in *T. lintearius*.

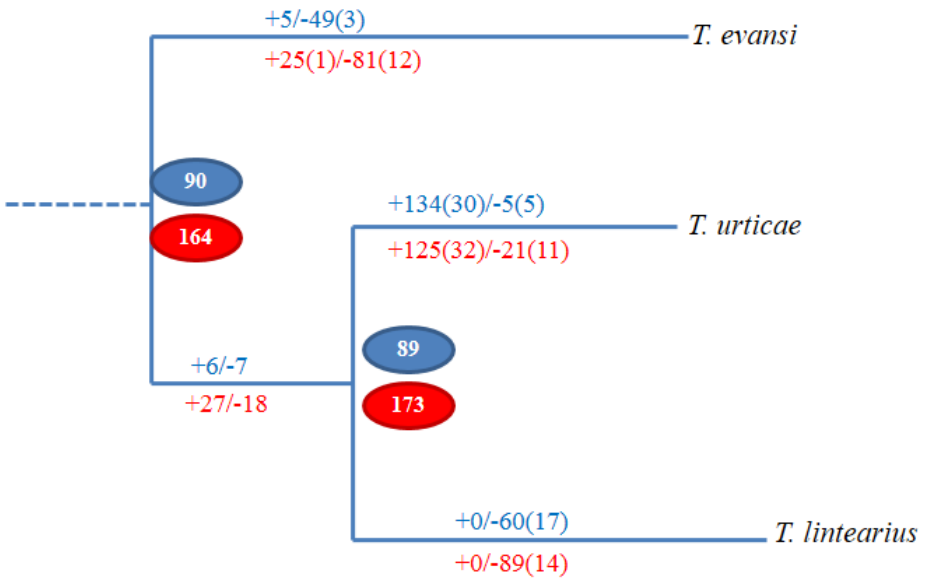


Figure 4.3. Predicted pattern of GR gain, loss and pseudogenization in the three spider mite genomes.
The numbers of GR gains and losses are indicated on each branch, together with pseudogenization in parenthesis (e.g. +25(1)/-81(12) refers to 25 GR gains in which 1 pseudogene, 81 losses in which 12 GR pseudogenizations). Colour code: GR-A, blue, GR-B, red. The ovals are the inferred numbers of ancestral GRs in each class.

The phylogenies and figures in Table 4.4 confirm the differential evolution of the three GR classes. GR-B has many more 1:1:1 orthologous relationships than GR-A, i.e. single copy GRs that are conserved in the three mites. Class A on the opposite is showing many more variations than class B, especially more duplications (Table 4.4).

Twenty clusters of TuGRs have orthologous TeGRs and TlGRs on different scaffolds of *T. evansi* and *T. linterarius* genome. 38 TeGRs out of 226 are indeed found in 11 clusters (5 clusters without intervening genes, 6 clusters with 1 intervening gene) ranging in size from 2 to 10 genes (Figure 4.4). The largest cluster on scaffold 46 with 9 TeGRs and 1 non-TeGRs was 67kb long. Other clusters had a size in the range 10kb to 19kb. 60 TlGRs

out of 257 are indeed found in 12 clusters (5 clusters without intervening genes, 7 clusters with 1, and 2 intervening genes) ranging in size from 2 to 10 genes. The largest cluster on scaffold 61 with 10 TlGRs and 1 non-TlGRs was 33kb long. Other clusters had a size in the range 10kb to 32kb. No cluster among TuGR, TeGR and TlGR shares the same size because of duplication of TuGR, and loss of TeGR (TlGR). Microsyntenies of GRs among three spider mites were showed in the Table A.7. There are 36 microsyntenies among three species, 30 between *T. urticae* and *T. evansi*, 12 between *T. evansi* and *T. lintearius*, and 50 between *T. urticae* and *T. lintearius* (Table A.7). These results are consistent with *T. evansi* being diverged much earlier than *T. urticae* from *T. lintearius*.

Table 4.4. Major features in the evolution of GR in three spider mites

GR-A	**GR-B**
1:1:1 (Tu:Te:Tl) orthologous relationship	
10 genes	36 genes
The largest expansions in TuGR (≥8)	
21 genes (263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 475, 476, 569, 668, 669, 670, 672, 673)	13 genes (63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 401, 402, 404)
14 genes (315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 415, 480, 481)	13 genes (163, 484, 572, 573, 578, 579, 580, 582, 583, 596, 625, 627, 629)
13 genes (338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 414, 417)	11 genes (158, 159, 169, 574, 581, 588, 594, 593, 620, 622, 623,)
9 genes (276, 277, 278, 279, 280, 281, 482, 483, 617)	8 genes (155, 156, 157, 561, 562, 565, 567, 568)
9 genes (388, 390, 391, 392, 393, 394, 395, 396, 644)	
8 genes (282, 283, 284, 285, 286, 287, 288, 289)	
The expansions in TeGR	
3 genes (326a, 326b, 326c)	7 genes (70a, 70b, 72a, 72b, 72c, 72d, 72f)
2 genes (364a, 364b)	4 genes (659a, 659b, 659c, 659d)
2 genes (364c, 364d)	4 genes (182b, 182d, 182e, 182f)
2 genes (287a, 287c)	3 genes (6a, 6b, 6c)
	3 genes (171, 172, 175)
	3 genes (126, 127, 130)
	2 genes (74a, 74b)
	2 genes (536b, 536c)
	2 genes (75a, 75b)
	2 genes (177a, 177b)
	2 genes (195a, 195b)
	2 genes (182a, 182c)
	2 genes (165, 586)

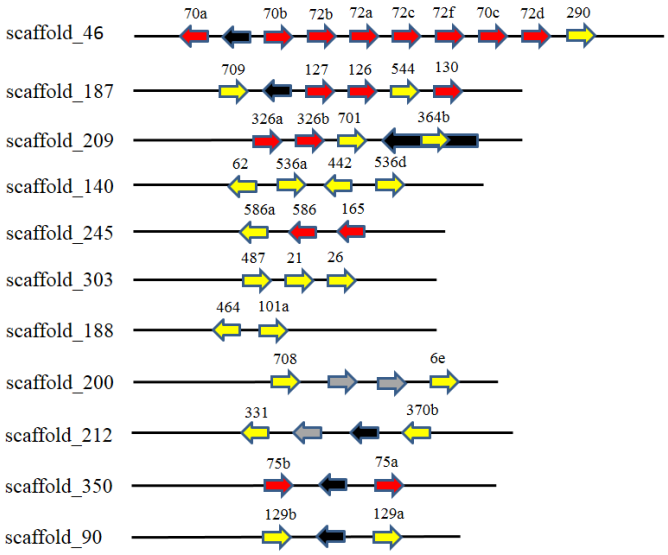


Figure 4.4. GR clusters in *T. evansi*. GRs from a sub-family expansion: red, GRs from different sub-families: yellow, intervening gene: black, TE/repeat: grey.

4.4. Conclusion

This research focused on the comparison of the repertoire of GR chemoreceptors in *T. urticae*, *T. evansi*, and *T. lintearius* as well as the molecular evolution of this highly divergent protein family. As reported in Chapter 3, *T. urticae* with its huge repertoire of 690 TuGRs has many more GRs than any insect or other arthropod described up to now. The results from this study showed a striking difference in GR gene number between the three species, GR abundance in *T. evansi* and *T. lintearius* being ca. one third of the one in *T. urticae*. The huge number of GRs in *T. urticae* unique to this polyphagous species results from the high occurrence of tandem duplications. In particular compared to the GRs of the oligophagous *T. evansi* where duplication is rare and to the monophagous *T. lintearius* where it never happens. The GR gene family has evolved in different ways in the three spider mite species though gene duplication, pseudogenization, and gene loss, the common ancestor of these mites having probably a repertoire of ca. 270 GR genes, which stays in the range of the sum of insect GRs and ORs. While *T. urticae* GRs had many lineage-specific expansions of particular gene subfamilies, with some copies occasionally inactivated by pseudogenization, GRs in the closely related *T. lintearius*

evolved through gene loss and pseudogenization, while not a single GR gene seems to have been duplicated. In the more distant *T. evansi*, the GR repertoire is also small, due to gene loss and rather few gene duplications. This denotes the high dynamic status of this GR family evolving through positive selection, which translates into a huge expansion in *T. urticae*. This fits with the need for *T. urticae* to cope with an increased range of plant hosts through its ability to recognize the new chemicals that these plants would produce, when on the reverse, the feeding of *T. evansi* mainly on *solanaceae* and of *T. lintearius* only on *Ulex europaeus* ends up in inactivation and loss of useless GR genes. These results tell that GRs do play an important role on the ecological adaptation of *T. urticae*, *T. evansi*, and *T. lintearius* to their feeding environment. We do not know which chemicals are recognized in the GR sub-families that show expansion or loss, and this of course should be the objective of further investigation, which would lead to better understanding of pest-host interplay and to prospects to control those pests. Having this objective in mind, the observation that one GR class, class A, is much more affected by this adaptive evolution suggests that GRs from this class should be given priority.

4.5. Materials and Methods

4.5.1. Sequencing and Genome assembly

Unlike *T. urticae*, which was still sequenced using whole genome shotgun sequencing with Sanger technology [253], *T. evansi* and *T. lintearius*, were sequenced using Illumina technology (2012). The reads generated where, for both, 2 paired-end read libraries with respectively 300nt and 500nt as targeted insert size. These were complemented with a mate-pair library targeting an insert size of 5,000nt. These data were evaluated using the FASTQC program (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). In every case the general quality was good, but as expected the quality at the end of the reads diminishes. Therefore, quality trimming was applied to all the libraries with a cut-off of quality score 20 and minimum length of 65nt. The read data has been assembled using the CLCBio assembly-cell software (version 4.06beta) (CLC Inc, Aarhus, Denmark). This software uses the DeBruijn graph algorithm to assemble the paired-end reads into contigs. The mate-pairs were used here only to scaffold within the CLCbio software. The CLCbio software delivers good quality contigs but is too stringent to exhaustively scaffold the contigs together to larger scaffolds. The obtained assembly from CLCBio

yielded 11,201 contigs of *T. evansi* (87.5Mb), and 40,412 contigs of *T. lintearius* (94.4Mb), with N50 of 216 and 400 respectively.

To improve the number of scaffolded contigs, and thus the N50, we used the software SSPACE (BaseClear) [283]. This software uses the reads that map on different contigs to link them together. This procedure improved the N50 to 85 for *T. evansi* (3,720 scaffolds, 95.7Mb) and 70 for *T. lintearius* (2,706 scaffolds, 91.5Mb). The gapped scaffolds were then threated with GapFiller (BaseClear) [284]. This software uses paired read where 1 read is anchored on a contig border while the other is not mapping at all. It assumed that the not mapping read, can be placed in the neighboring gap depending on the insert size. It is therefore important to have a paired-end library with an as tight insert size as possible. The approach of gap-filling is thus to extend contig end in an iterative manner by performing local assemblies. Gap-filling doesn't influence the N50, but allows incorporation of reads in more difficult regions of the genome, like repeats. The resulting final assemblies for *T. evansi* and *T. lintearius* are given in Table 4.1.

4.5.2. Genome annotation

Similar to the annotation of *T. urticae* genome (Chapter 2), we have also used EuGene [146], a gene prediction platform for eukaryotes that combines several sources of evidence, to annotate the genome of *T. evansi* and *T. lintearius* based on the proteome of *T. urticae.* All parameters from *T. urticae* genome annotation were used for *T. evansi* and *T. lintearius*. The transposable elements of *T. urticae* were mapped on *T. evansi* and *T. lintearius* scaffolds to mask repeat regions in *T. evansi* and *T. lintearius*. For extrinsic annotation, the following data sources were integrated into the annotation system: 1) protein sets from the latest Flybase release [242] and Swissprot; 2) protein set from *T. urticae*; and 3) a large number of Illumina RNA-seq reads of *T. evansi* and *T. lintearius*. Similarity with proteins from Flybase and Swissprot was obtained through BLASTX and passed on to Eugene. *T. urticae* proteins were mapped to *T. evansi* and *T. lintearius* scaffolds using GenomeThreader [223]. The output of GenomeThreader with intron-exon boundaries was reformatted into GFF3 for Eugene and used as highly reliable data. The illumina RNA-seq reads were quality trimmed and filtered using the FASTA tools. The

good RNA-seqs were aligned to *T. evansi*, and *T. lintearius* scaffolds using Bowtie and Tophat to identify spliced reads (junctions).

4.5.3. Insect GR like chemoreceptor annotation in *T. evansi* and *T. lintearius*

The GRs from *T. urticae* (Chapter 3) have been used as a reference set to performTBLASTN searches (E-value < 1) for similar regions in *T.* evansi and *T. lintearius* scaffolds. From similar regions, through GenomeView [267], GR gene models in *T. evansi* and *T. lintearius* were checked, corrected for Eugene predicted genes or constructed for unpredicted genes and updated manually using ORCAE website [150]. The minimum length allowed for a pseudogene to be annotated was 20% of shortest intact GR in *T. urticae*.

Orthologous pairs between *T. urticae* and *T. evansi* GRs, *T. urticae* and *T. lintearius* GRs were identified by reciprocal best hits (RBH), and microsynteny by checking the adjacent upstream/downstream neighbors (if the neighbor is GR, the next neighbor was taken). Genes for each orthologous group were identified by reciprocal BLASTs. *T. evansi* and *T. lintearius* GRs were named based on *T. urticae* GR (TuGR) numbering. If GRs of *T. urticae* and *T. evansi*/*T. lintearius* were 1:1 orthology relationship, the name of *T. evansi*/*T. lintearius* GR was TeGR/TlGR with numbering from *T. urticae* GR (e.g. TeGR366, and TlGR366 for the orthologs of *T. urticae* GR366). If GRs of *T. urticae* and *T. evansi*/*T. lintearius* were 1:many orthology relationship, the name of *T. evansi*/*T. lintearius* GR was TeGR/TlGR with numbering from *T. urticae* GR and a,b,c,d…suffix. In case of novelty, name of *T. evansi*/*T. lintearius* GR was TeGR/TlGR with new numbering.

4.5.4. Loss and gain parsimony analysis

Illustrations of parsimony analysis to GR gain and loss to infer the evolutionary history of GRs were showed in the Figure 4.5. Parsimonious scenario is one that is best consistent with the topology of the species tree. For example, in the subtree A (Figure 4.5A), a first duplication predating Tu-Te-Tl speciation gave rise to GR251 on the one side and (GR250/GR252) on the other. GR251 has the expected 1:1:1 topology (with *T. evansi* older) as the topology of the species tree. Then, a another duplication occurred before Tu-

Te-Tl speciation with GR250 on the one side, followed by pseudogenization of GR250 in *T. lintearius* and GR252 on the other, followed by loss of GR252 in *T. evansi*, and pseudogenization of GR252 in *T. lintearius*. In the subtree B (Figure 4.5B), a first duplication predating Tu-Te-Tl speciation gave rise to GR308 on the one side, followed by loss of GR308 in *T. evansi*, and (GR303/GR305) on the other, followed by a next duplication before Tu-Te-Tl speciation. GR305 has the expected 1:1:1 topology (with *T. evansi* older) as the topology of the species tree. TlGR303 was pseudogenizated, and TuGR304 was duplicated from TuGR303. In the subtree C (Figure 4.5C), there were 2 duplications in *T. urticae*, 1 duplication in *T. evansi*, and 1 loss in *T. lintearius*. In the subtree D (Figure 4.5D), a first duplication predating Tu-Tl speciation gave rise to GR379 on the one side, and GR541 on the other, followed by loss in *T. lintearius*, and two duplications, 1 pseudogenization in *T.urticae*.

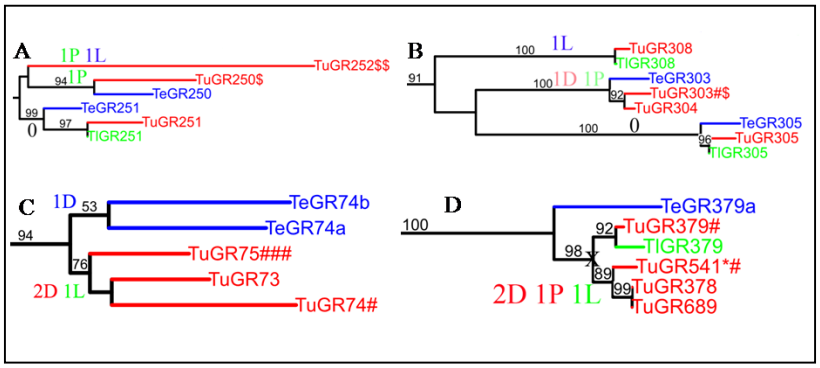


Figure 4.5. Inferring loss and gain of GR genes according to species tree topology (*T. urticae* (red), *T. evansi* (blue) and *T. lintearius* (green)). D: Gene Duplication, L: Gene Loss, P: Pseudogenization, X: gene duplication predating *T. urticae-T. lintearius* speciation. GR with suffix "*" after protein name is pseudogene. TuGR and TeGR with suffix "#" after protein name has homologous pseudogene(s)/partial gene(s) with many changes (blast cutoff 1e-25) in *T. evansi* and *T. urticae*, respectively. TuGR and TlGR with suffix "$" after protein name has homologous pseudogene(s)/partial gene(s) with many changes (blast cutoff 1e-25) in *T. lintearius* and *T. urticae*, respectively.

4.5.5. Phylogenetics trees

The insect GR-like chemoreceptors (intact genes, and pseudogenes with one or two events) from *T. urticae*, *T. evansi* and *T. lintearius* of each class (class A, and class B) were aligned using ClustalW [270]. The alignments were annotated based on the seven trans-membrane domains predicted by TMHMM [271] as well as on intron positions, and

edited in Jalview [272]. The divergent N-terminal regions (TM1-4), the short C-terminal tails as well as the major gaps between trans-membranes were removed to obtain the final alignments for tree construction, based on 129 and 127 unambiguously aligned sequence positions, for class A and class B respectively. To build the phylogenetic tree for each class in this highly divergent protein family, we applied a similar procedure as published previously for insects and *Daphnia* [249, 186]. Amino acid distances were corrected for multiple amino acid replacements by using the BLOSUM62 amino acid exchange matrix in TREE-PUZZLE v5.0 [273]. The heuristic search based on corrected distances with tree-bisection and reconnection branch swapping in PAUP*v4 [274] was used to build the phylogenetic tree. Bootstrap analysis was performed by analyzing 1000 neighbor-joining replications with uncorrected distances.

Chapter 5

Conclusions and perspectives

5.1. Genome annotation in times of fast development of next and next-next generation sequencing

Recent advances in high throughput next generation sequencing techniques allow the production of huge amounts of biological data in an acceptable time frame and at a low cost [285]. As a result, many genome projects have been initiated (and finished), and have led to a large number of whole genome sequences that needed to be annotated. Furthermore, in parallel with genome sequencing, RNA sequencing based on next generation sequencing technology has provided tens to hundreds of millions of RNA-seq "reads" and information on billions of individual bases of RNAs inside a cell at various growth conditions and time points [286]. These data sources are used not only to study gene expression but also to help with the genome sequence annotation. In this thesis, I showed how to integrate RNA-seq data into the Eugene gene prediction program to annotate the spider mite genomes of *Tetranychus urticae*, *Tetranychus evansi* and *Tetranychus lintearius*. Below, I will discuss the benefits from using RNA-seq reads for genome annotation as well as challenges that still need to be solved. And I will also briefly discuss the quality of current genome annotation.

5.1.1. Challenges

Traditionally, EST, cDNA, and protein sequences have been used for evidence driven gene prediction to improve the accuracy of genome annotations. These resources have to be mapped on a genomic sequence to identify the exon-intron structures of the sequence. Similarly, in recent years, new tools, such as Bowtie and Tophat, have allowed mapping RNA-seq reads on the genomic sequence to identify splice junctions between exons. Thus, RNA-seq reads can be used as ESTs to update existing annotations as well as to annotate newly sequenced genomes [287]. However, the genome annotation making use

of next generation sequencing data is still challenging because of sequencing errors and lack of specialized tools [107]. In fact, to predict genes with RNA-seq reads, first these reads must be aligned to the genome to identify intron/exon structures and splice sites. Next, these results must be post-processed before they can be passed on to a gene finder. All these tasks require specialized tools. Because of the large number of sequence errors of short RNA-seq reads, it can be difficult to align them unambiguously on scaffolds. Therefore, tools need to be available or developed to create the best alignments. Moreover, it is difficult to assemble short reads into longer scaffolds of a genome sequenced and assembled by next generation sequencing techniques, thus genes on the border of the scaffolds can be split up in the gene prediction. Finally, existing gene finders can be overloaded by a huge number of RNA-seqs. Practically when including RNA-seq data into the Eugène platform we have been facing several issues. First we ended up in the wrong prediction of many small genes which were seemingly supported by very few spurious RNA-seq reads. Second, mapping was far from perfect and longer transcripts were often split, especially for low expressed genes. For these reasons, we tested two scenarios, which improved the prediction: using RNA-seqs as assembled contigs directly and only integrating splice junctions into Eugene. As a result, based on the comparison of different versions of the gene predictions and feedback from the spider mite community, the only splice junctions were used as ESTs to pass into Eugene. Despite existence of the above challenges, I believe that, in the near future, with cheaper, faster and more accurate sequencing techniques as well as better/updated gene prediction software tools, the challenges described above can be overcome and RNA-seq data will continue to provide really good evidence to increase accuracy of genome annotation.

5.1.2. Bad quality versus high quality

When a genome sequence is determined, genome annotation is first step to analyze the genome and to bridge the gap from the sequence to the biology of the organism [288]. An incorrect annotation can lead to erroneous results for follow-up studies and downstream analyses. In addition, the errors will spread when other genome projects use the annotation for annotating their own genomes. Therefore, quality is the most important criterion of a genome annotation. However, whereas more and more genome sequences have been produced, in many cases, the annotation quality has gone down. Traditionally,

a genome project is run by a consortium, including a number of research groups with broad experience and knowledge, to create a high quality annotation. With next generation sequencing techniques, the production of genome sequences by a research group is easier, faster and cheaper. Conveniently, the genomes are usually only annotated and checked by the group that has done the sequencing. However, often the research group lacks expertise in annotation or biological knowledge on various gene families to check the correctness of the gene models, which often lead to bad quality annotations. To maintain high quality annotations in these times of genome sequence explosion, I think that genome projects, especially of new model organisms, should continue to be performed through collaboration of many research groups. However, to decrease the time to obtain high quality annotations of a genome sequence, better online tools to improve the interaction between bioinformaticians and biologists should be built, ORCAE being one of this kind [150]. In general, a reference genome should be annotated by expert bioinformaticians and checked carefully by biologists.

5.2. The next generation of arthropod genomics

About one decade ago, whole genome sequencing was only performed for model organisms like *Drosophila melanogaster*, *Mus musculus*, *Arabidopsis thaliana* and *Caenorhabditis elegans* because of limitations of DNA sequencing methods. Thanks to recent advances of next generation sequencing techniques, it is now possible to sequence transcriptomes and genomes at the population scale of all species on Earth [289]. For examples, the 1,000 Human Genome Project was launched in January 2008 [100] and 1,092 genomes were already sequenced and published as a result [290]; the 1,000 Plant Genome Project has sequenced genomes and transcriptomes of 1000 different plant species around the world. The 1,001 *Arabidopsis thaliana* genome project focused on different populations across different places on our Planet [97]; the 10,000 vertebrate species project was also launched in 2009 [99]. Currently, we are in the early stages of i5K arthropod genome project launched in 2012 [291]. The project has opened many studies at genomic level on arthropods, the most diverse and successful branch of metazoan evolution, with millions of extant species.

In this thesis, it has been shown that the genome of *Tetranychus urticae* reveals herbivorous pest adaptation and offers new insights into arthropod evolution and plant–herbivore interactions, and provides unique opportunities for developing novel plant protection strategies (Chapter 2). Recent research on the scorpion *Mesobuthus martensii*, grouped together with *T. urticae* in the arachnid clade and known as 'living fossils' that maintain an ancient anatomy and are adapted to have survived extreme climate changes, reveal a unique adaptation model of arthropods and offer new insights into the genetic bases of living fossils [292]. Thus, the planned sequencing of 5,000 arthropod genomes will provide general insights into the evolution of arthropod genomes and will allow inferring the basis processes of evolution, development, physiology, reproduction, and survival. Understanding the genetics of how organisms adapt to different environments has been a fundamental issue of evolutionary genetics for decades. However, studies were limited by the lack of a genomic perspective. Therefore, with 5,000 arthropod genomes, ecologists and evolutionary biologists will have an opportunity to study evolutionary mechanisms leading to the adaptation of arthropods that are able to live everywhere on Earth.

However, bioinformaticians do need to develop or update tools to turn the genomic resources into biological knowledge. Assembling the reads produced by automatic sequencing machines still pose challenges because of short-read lengths, sequencing errors, and genomic repeats [293]. Current assemblers depend on the sequencing platform, error model, sequence reads, etc. Thus, a general assembler that can handle the sequences generated by different sequencing platforms needs to be constructed. In addition, tools to control the quality of NGS data are extremely important for meaningful downstream analysis [294], although next–next generation DNA sequencing systems which have been developed promise to overcome above-mentioned challenges because of much longer reads and reduced errors [39]. Recently, structural genetic variations have focused on genetic differences in the form of short sequence fragments or structural rearrangements in populations or close relatives of a single species. The basic structural variations include insertions, deletions, duplications, translocations and inversions. Balanced variations (translocations and inversions) do not change the total DNA content, whereas unbalanced variations (insertions, deletions, and duplications) change the total

DNA content. With the advent of next-generation sequencing, new methods are being developed to detect structural variations in genomes [295, 296]. In addition, visualization for the structural variations, comparative genomics etc., for biologists has been developed. Besides, experimental biologists also need to develop genomic methods allowing genome-wide studies. Further collaboration between bioinformaticians and biologists to study genomic data from the i5K and related projects will undoubtedly provide much further insight into many of the mysteries of arthropod evolution.

5.3. Chemosensory receptors in spider mites: What are the next steps?

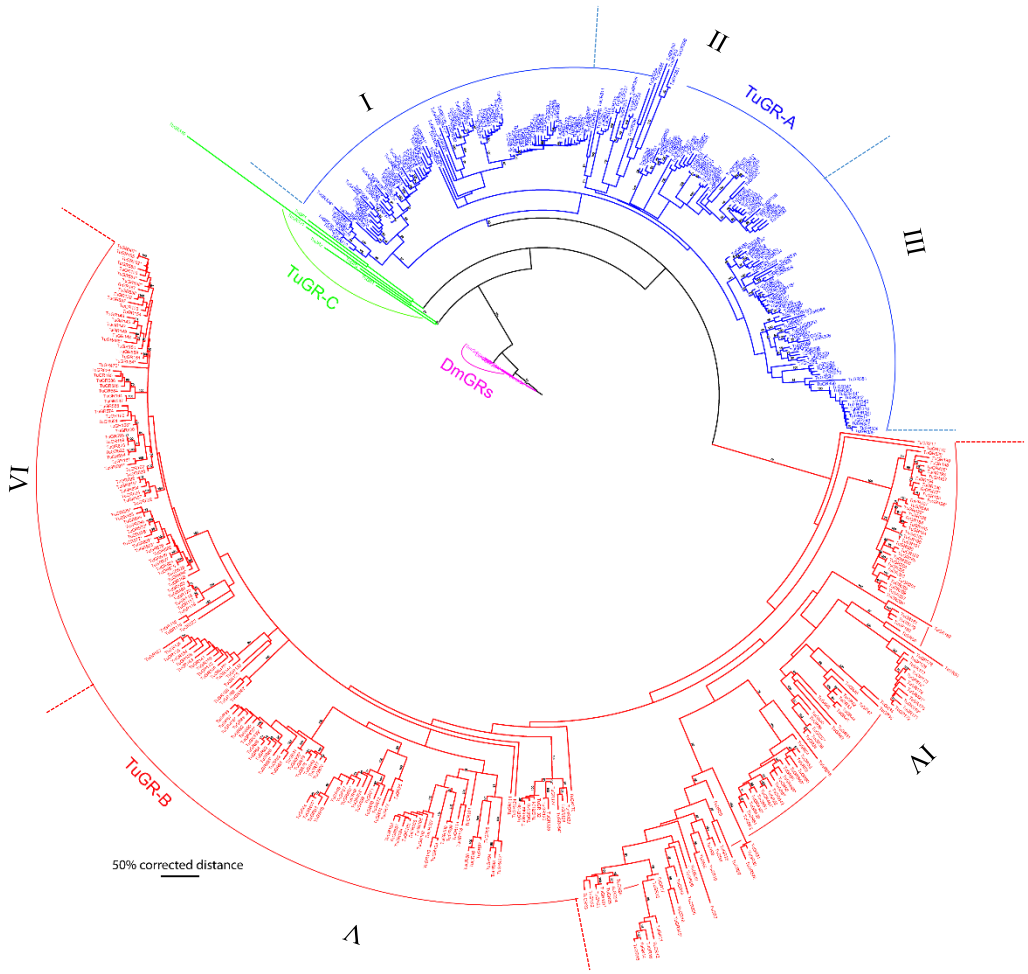
Tetranychus urticae not only has many more GR-like chemoreceptors than insects and any other arthropod studied up to now but also shows adaptive selection at the level of population of this protein family (Chapter 3). When compared to GRs of the oligophagous *Tetranychys evansi* and the monophagous *Tetranychus lintearius*, GRs were expanded remarkably in the polyphagous *T. urticae* with many lineage-specific duplications of particular gene subfamilies (Chapter 4). However, we do not know the natural ligands for these GRs. Therefore, in the next important step, expression and functional analysis of the GRs need to be performed to identify receptor ligands and to map the receptors to functional classes of receptor neurons. In the last decade, the molecular and cellular basis of chemosensory reception, allowing to recognize and discriminate attractive and repulsive odorants and tastants, and make behavioral decisions accordingly, has been remarkably well studied and understood in *Drosophila melanogaster*, a model organism of insects [297]. This basis can be used to guide studies on the molecular and cellular mechanisms of chemoreceptors in spider mites which link to differences in spider mite behaviour and ecology. If we can fully understand these mechanisms, spider mite chemoreceptors, like insect ones [298], may be future potential targets in the search for alternative control strategies.

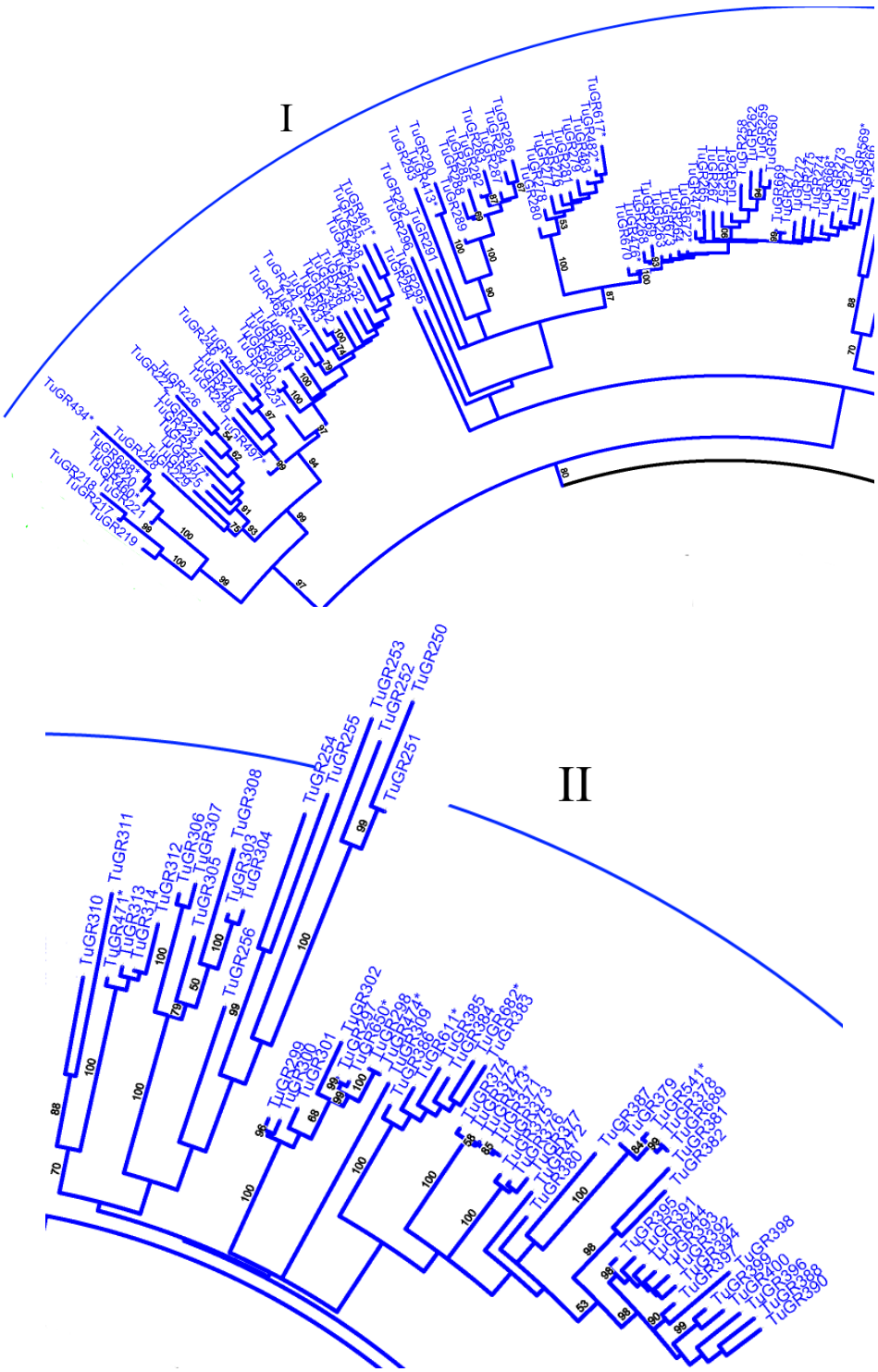
Appendix

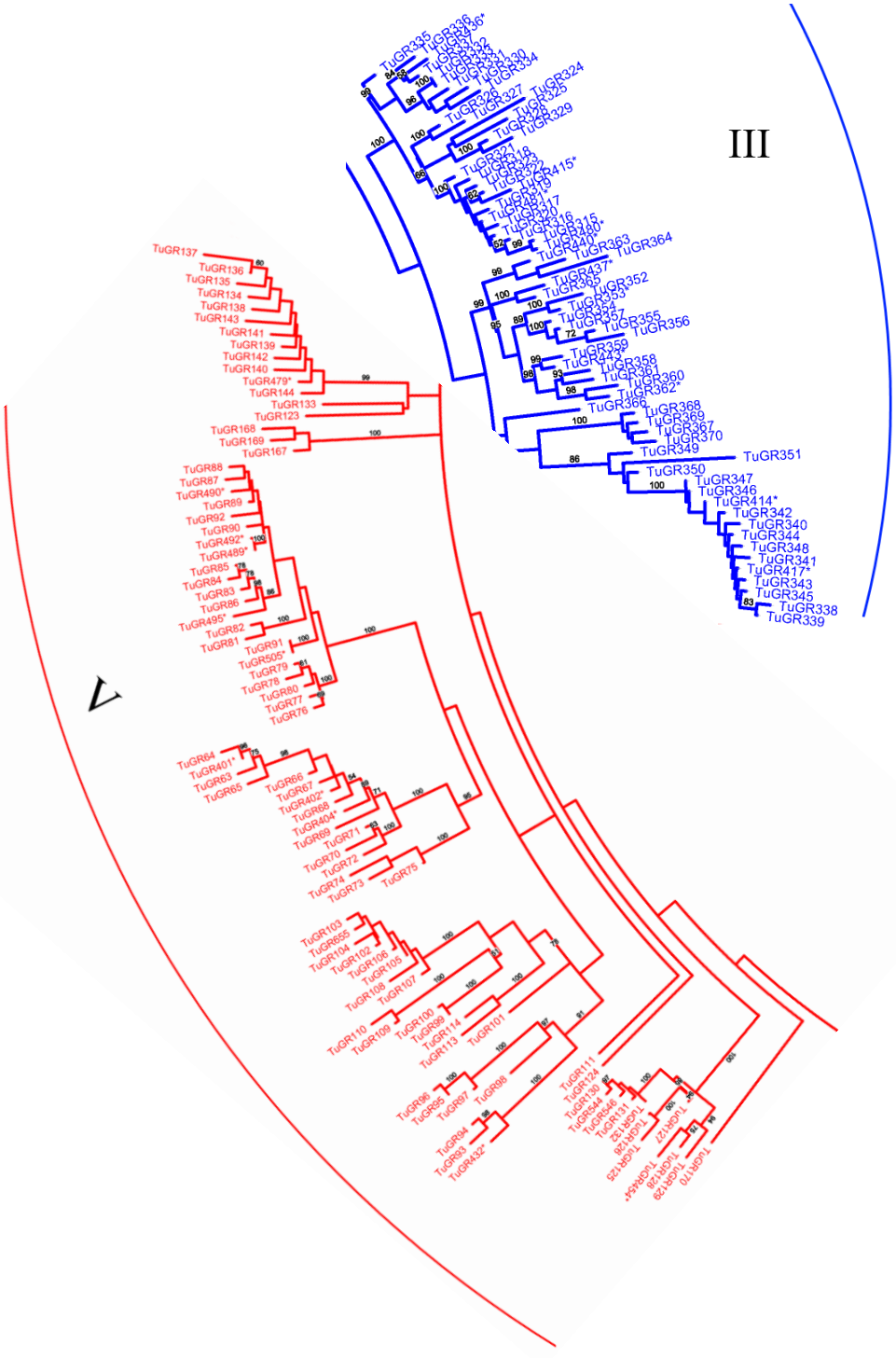
Supplementary figures and tables

A.1. *Tetranychus urticae* chemosensory receptors

A.1.1. Phylogenetic trees of the chemosensory receptors (TuGRs) from *T. urticae*




I
II


III
V


IV

Figure A.1. Phylogenetic tree of the TuGRs (intact genes and pseudogenes with single or two events) from *T. urticae* (exception 10 highly diverse genes of TuGR-C) with bootstrap values ≥50/100 from 1000 replications of uncorrected distance analysis. This corrected distance tree was rooted with five sugar receptors from *Drosophila melanogaster*. Pseudogenes have suffix "*" after protein names.

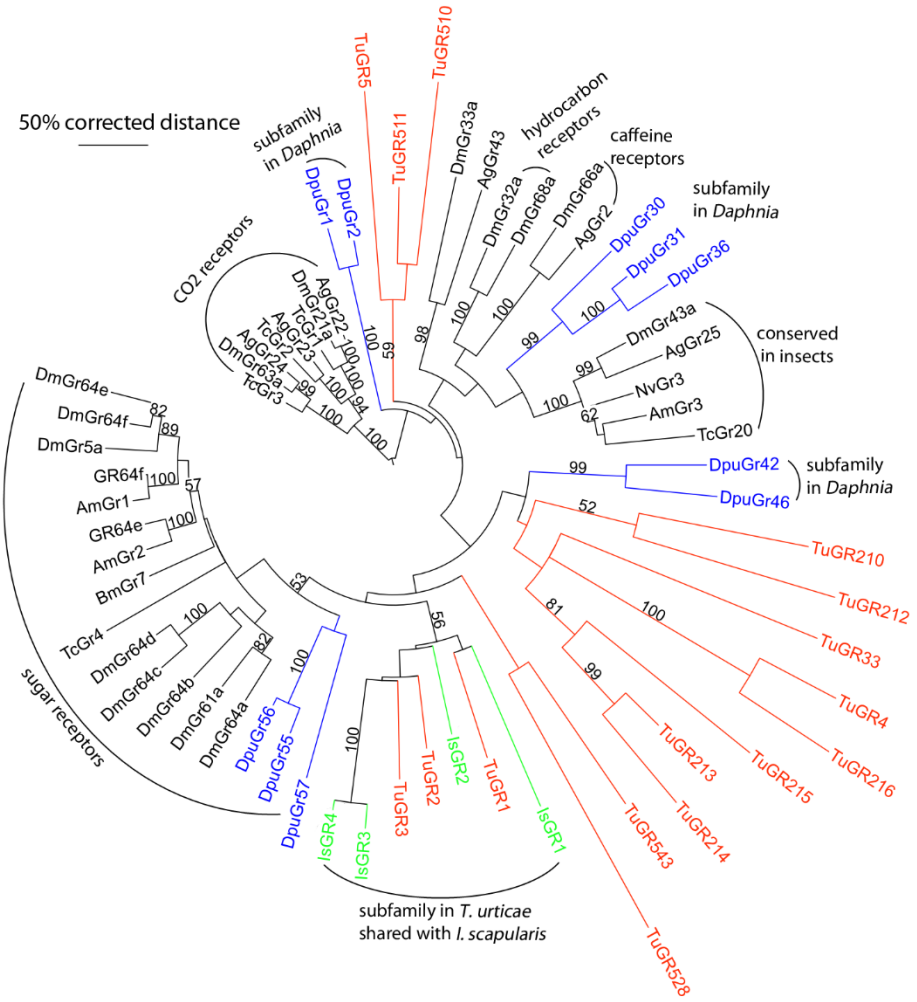


Figure A.2. Phylogenetic tree of 16 divergent TuGRs (red), 4 IsCRs (green) and representative gustatory receptor subfamilies from *Daphnia pulex* (blue) and insect (black) (Ag - *Anopheles gambiae*, Am – *Apis mellifera*, Bm – *Bombyx mori*, Dm - *Drosophila melanogaster*, Dpu - *Daphnia pulex*, Is – *Ixode scapularis*, Nv – *Nasonia vitripennis*, Tc – *Tribolium castaneum*, Tu – *Tetranychus urticae*) with bootstrap values ≥50/100 from 1000 replications of uncorrected distance analysis. The tree was rooted with highly conserved CO_2 receptors.

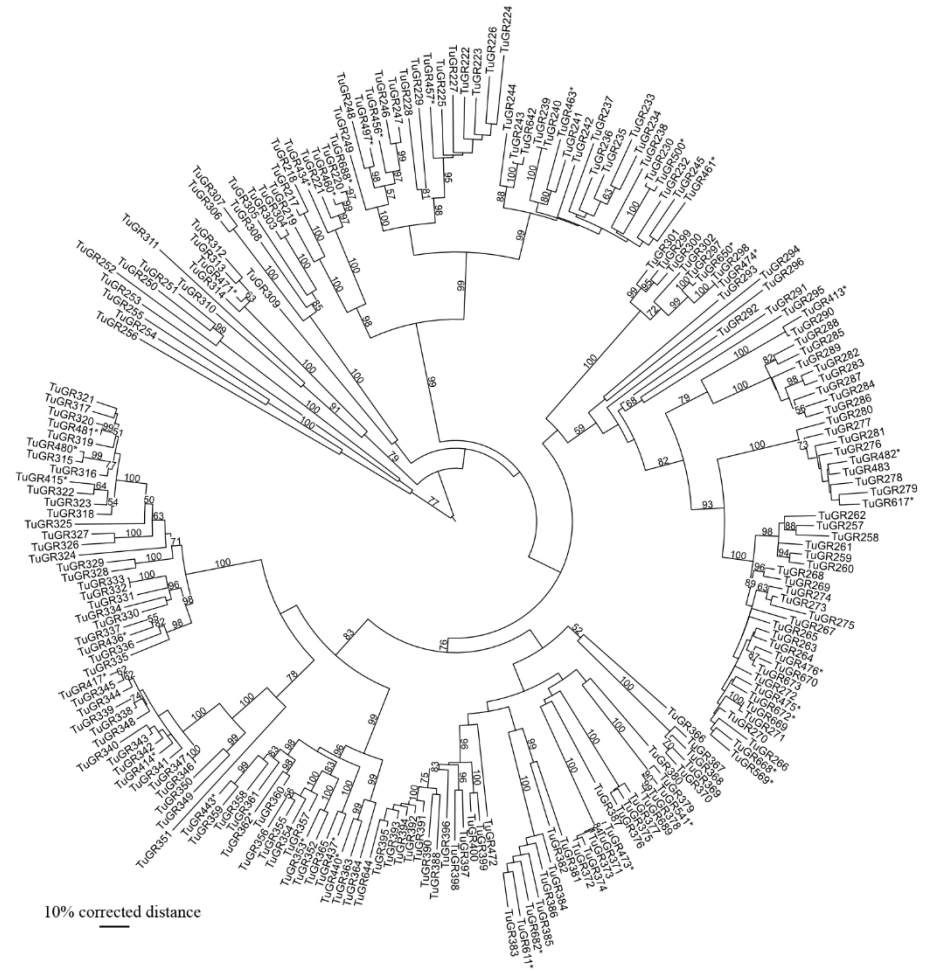


Figure A.3. Phylogenetic tree of TuGR-A gustatory receptor group (intact genes and pseudogenes with single or two events) from *T. urticae* with bootstrap values ≥50/100 from 1000 replications of uncorrected distance analysis. The tree was rooted by midpoint rooting. Pseudogenes have suffix "*" after protein names.

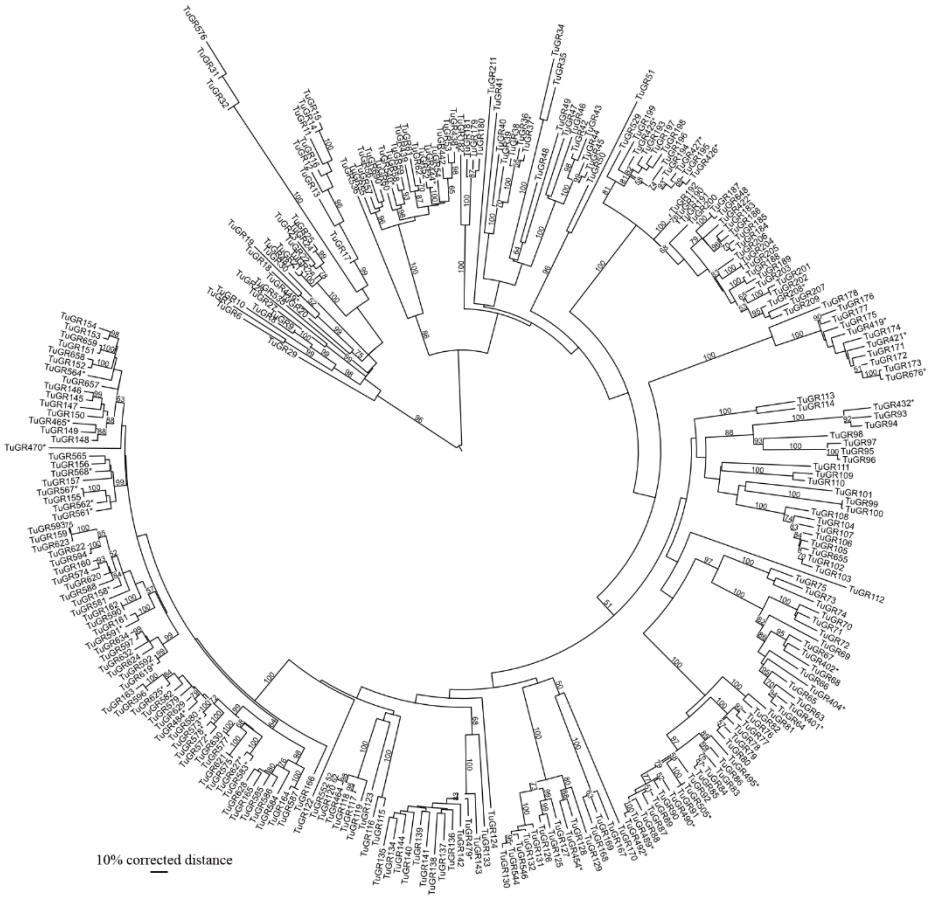


Figure A.4. Phylogenetic tree of TuGR-B gustatory receptor group (intact genes and pseudogenes with single or two events) from *T. urticae* with bootstrap values ≥50/100 from 1000 replications of uncorrected distance analysis. The tree was rooted by midpoint rooting. Pseudogenes have suffix "*" after protein names.

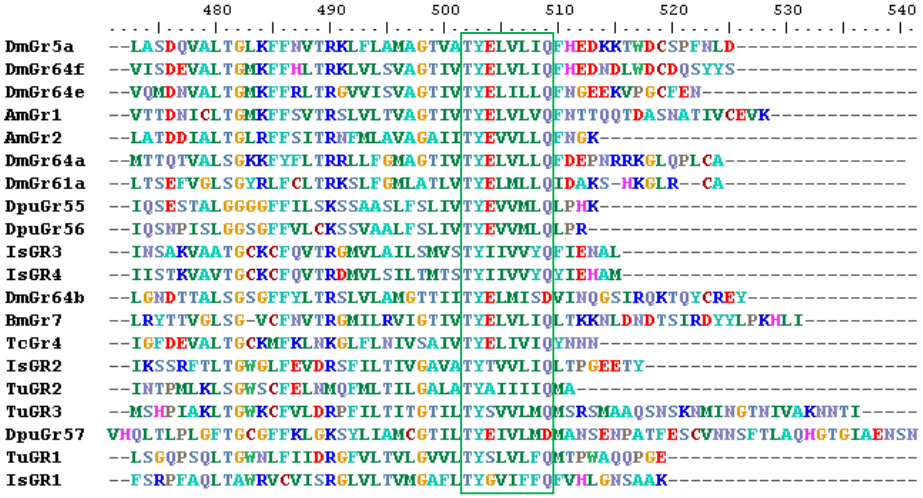


Figure A.5. The last transmembrane helix (TM7) of the most divergent TuGR's, their homologs in the genome of the tick *I. scapularis* and a group of pancrustacean GRs. The alignment shows motif TYxxxxxQ.

A.1.2. TuGRs in *T. urticae*

Table A.1. The TuGRs in *T.urticae*

N	GR_ID	gene_ID	status	length	Introns	change vs. intact	group
1	tetur12g00810	TuGR217	intact	378	1		A
2	tetur34g00390	TuGR218	intact	377	1		A
3	tetur06g05120	TuGR219	intact	387	1		A
4	tetur12g04763	TuGR220	intact	383	1		A
5	tetur35g00340	TuGR221	intact	383	1		A
6	tetur12g01660	TuGR222	intact	370	1		A
7	tetur12g04661	TuGR223	intact	372	1		A
8	tetur12g01680	TuGR224	intact	371	1		A
9	tetur12g04903	TuGR225	intact	370	1		A
10	tetur12g01670	TuGR226	intact	370	1		A
11	tetur12g01690	TuGR227	intact	339	1		A
12	tetur12g04743	TuGR228	intact	390	1		A
13	tetur27g01820	TuGR229	intact	378	1		A
14	tetur12g04010	TuGR230	intact	378	1		A
15	tetur12g04783	TuGR232	intact	377	1		A
16	tetur12g04893	TuGR233	intact	387	1		A
17	tetur66g00120	TuGR234	intact	373	1		A
18	tetur12g04833	TuGR235	intact	378	1		A
19	tetur12g04853	TuGR236	intact	378	1		A
20	tetur12g04823	TuGR237	intact	366	1		A
21	tetur12g04843	TuGR238	intact	379	1		A
22	tetur12g04020	TuGR239	intact	372	1		A
23	tetur66g00010	TuGR240	intact	372	1		A
24	tetur66g00090	TuGR241	intact	364	1		A
25	tetur12g04090	TuGR242	intact	376	1		A
26	tetur12g04773	TuGR243	intact	372	1		A

27	tetur12g04793	TuGR244	intact	373	1		A
28	tetur12g04803	TuGR245	intact	378	1		A
29	tetur34g00400	TuGR246	intact	378	1		A
30	tetur34g00410	TuGR247	intact	383	1		A
31	tetur12g04703	TuGR248	intact	380	1		A
32	tetur12g04753	TuGR249	intact	373	1		A
33	tetur14g02850	TuGR250	intact	381	1		A
34	tetur37g00370	TuGR251	intact	385	1		A
35	tetur13g04410	TuGR252	intact	375	1		A
36	tetur01g16774	TuGR254	intact	404	1		A
37	tetur13g00950	TuGR255	intact	391	1		A
38	tetur24g00910	TuGR256	intact	415	1		A
39	tetur01g16734	TuGR257	intact	376	1		A
40	tetur01g16854	TuGR258	intact	367	1		A
41	tetur01g16784	TuGR259	intact	361	1		A
42	tetur01g16844	TuGR260	intact	360	1		A
43	tetur01g16864	TuGR261	intact	361	1		A
44	tetur12g04913	TuGR262	intact	376	1		A
45	tetur16g03956	TuGR263	intact	359	1		A
46	tetur16g04106	TuGR264	intact	356	1		A
47	tetur16g03916	TuGR265	intact	359	1		A
48	tetur16g04096	TuGR266	intact	338	1		A
49	tetur16g03976	TuGR267	intact	358	1		A
50	tetur16g03996	TuGR268	intact	345	1		A
51	tetur16g04006	TuGR269	intact	360	1		A
52	tetur16g03966	TuGR270	intact	348	1		A
53	tetur16g04086	TuGR271	intact	359	1		A
54	tetur16g03986	TuGR272	intact	380	1		A
55	tetur16g04046	TuGR273	intact	354	1		A
56	tetur16g04066	TuGR274	intact	357	1		A
57	tetur16g04056	TuGR275	intact	355	1		A
58	tetur21g03360	TuGR276	intact	346	1		A
59	tetur21g03380	TuGR277	intact	345	1		A
60	tetur21g03370	TuGR278	intact	364	1		A
61	tetur21g03350	TuGR279	intact	344	1		A
62	tetur04g09587	TuGR280	intact	345	1		A
63	tetur21g03390	TuGR281	intact	345	1		A
64	tetur07g04980	TuGR282	intact	351	1		A
65	tetur07g05040	TuGR283	intact	348	1		A
66	tetur07g08117	TuGR284	intact	349	1		A
67	tetur07g08097	TuGR285	intact	350	1		A
68	tetur07g04990	TuGR286	intact	350	1		A
69	tetur07g08107	TuGR287	intact	353	1		A
70	tetur07g05030	TuGR288	intact	346	1		A
71	tetur07g05000	TuGR289	intact	350	1		A
72	tetur01g16834	TuGR290	intact	349	1		A
73	tetur13g04726	TuGR291	intact	349	1		A
74	tetur03g07910	TuGR292	intact	371	1		A
75	tetur03g07800	TuGR293	intact	338	1		A
76	tetur31g02042	TuGR294	intact	380	1		A
77	tetur03g04270	TuGR295	intact	358	1		A
78	tetur02g08860	TuGR296	intact	341	1		A
79	tetur02g14340	TuGR297	intact	361	1		A
80	tetur02g14350	TuGR298	intact	353	1		A
81	tetur12g00150	TuGR299	intact	372	1		A
82	tetur12g00160	TuGR300	intact	370	1		A

83	tetur12g00170	TuGR301	intact	370	1		A
84	tetur12g00130	TuGR302	intact	363	1		A
85	tetur07g04200	TuGR303	intact	426	1		A
86	tetur14g00940	TuGR304	intact	424	1		A
87	tetur14g01040	TuGR305	intact	378	1		A
88	tetur19g01180	TuGR306	intact	384	1		A
89	tetur19g01190	TuGR307	intact	389	1		A
90	tetur13g04736	TuGR308	intact	412	1		A
91	tetur25g02112	TuGR309	intact	403	1		A
92	tetur06g06711	TuGR310	intact	376	1		A
93	tetur15g00440	TuGR311	intact	372	1		A
94	tetur15g00420	TuGR312	intact	378	1		A
95	tetur398g00010	TuGR313	intact	373	1		A
96	tetur15g00430	TuGR314	intact	378	1		A
97	tetur02g06680	TuGR315	intact	355	1		A
98	tetur02g06720	TuGR316	intact	352	1		A
99	tetur02g06730	TuGR317	intact	355	1		A
100	tetur02g06750	TuGR318	intact	352	1		A
101	tetur02g06740	TuGR319	intact	349	1		A
102	tetur02g15237	TuGR320	intact	355	1		A
103	tetur02g06770	TuGR321	intact	355	1		A
104	tetur02g06820	TuGR322	intact	355	1		A
105	tetur02g06870	TuGR323	intact	355	1		A
106	tetur08g08150	TuGR324	intact	351	1		A
107	tetur08g08265	TuGR325	intact	354	1		A
108	tetur08g01080	TuGR326	intact	355	1		A
109	tetur08g08100	TuGR327	intact	354	1		A
110	tetur08g01070	TuGR328	intact	342	1		A
111	tetur08g01090	TuGR329	intact	350	1		A
112	tetur08g08170	TuGR330	intact	342	1		A
113	tetur08g08220	TuGR331	intact	353	1		A
114	tetur11g05510	TuGR332	intact	344	1		A
115	tetur46g00120	TuGR333	intact	344	1		A
116	tetur08g00750	TuGR334	intact	345	1		A
117	tetur08g08160	TuGR335	intact	338	1		A
118	tetur08g08210	TuGR336	intact	346	1		A
119	tetur08g00780	TuGR337	intact	346	1		A
120	tetur02g06690	TuGR338	intact	349	1		A
121	tetur20g03320	TuGR339	intact	353	1		A
122	tetur02g06710	TuGR340	intact	352	1		A
123	tetur02g06780	TuGR341	intact	349	1		A
124	tetur02g06760	TuGR342	intact	349	1		A
125	tetur02g06810	TuGR343	intact	349	1		A
126	tetur02g06800	TuGR344	intact	348	1		A
127	tetur02g15227	TuGR345	intact	348	1		A
128	tetur02g06830	TuGR346	intact	345	1		A
129	tetur02g06860	TuGR347	intact	345	1		A
130	tetur02g06840	TuGR348	intact	349	1		A
131	tetur08g00530	TuGR349	intact	355	1		A
132	tetur08g00540	TuGR350	intact	355	1		A
133	tetur40g00402	TuGR351	intact	361	1		A
134	tetur08g00840	TuGR352	intact	351	1		A
135	tetur14g01230	TuGR354	intact	358	1		A
136	tetur14g01250	TuGR355	intact	360	1		A
137	tetur14g01260	TuGR356	intact	352	1		A
138	tetur14g01240	TuGR357	intact	357	1		A

139	tetur08g00870	TuGR358	intact	355	1		A
140	tetur08g00910	TuGR359	intact	353	1		A
141	tetur08g00880	TuGR360	intact	354	1		A
142	tetur08g00860	TuGR361	intact	353	1		A
143	tetur08g08110	TuGR363	intact	354	1		A
144	tetur08g08130	TuGR364	intact	354	1		A
145	tetur08g00770	TuGR365	intact	354	1		A
146	tetur12g01060	TuGR366	intact	361	1		A
147	tetur08g08200	TuGR367	intact	359	1		A
148	tetur08g08230	TuGR368	intact	359	1		A
149	tetur08g00800	TuGR369	intact	359	1		A
150	tetur08g00790	TuGR370	intact	359	1		A
151	tetur09g02300	TuGR371	intact	370	1		A
152	tetur167g00020	TuGR372	intact	365	1		A
153	tetur09g02310	TuGR373	intact	365	1		A
154	tetur09g02320	TuGR374	intact	365	1		A
155	tetur09g03640	TuGR375	intact	374	1		A
156	tetur09g03660	TuGR376	intact	374	1		A
157	tetur09g06809	TuGR377	intact	382	1		A
158	tetur09g06799	TuGR378	intact	373	1		A
159	tetur09g06819	TuGR379	intact	372	1		A
160	tetur13g04716	TuGR380	intact	371	1		A
161	tetur18g03721	TuGR381	intact	370	1		A
162	tetur18g03761	TuGR382	intact	362	1		A
163	tetur18g03731	TuGR383	intact	370	1		A
164	tetur18g03811	TuGR384	intact	391	1		A
165	tetur18g03831	TuGR385	intact	369	1		A
166	tetur18g03751	TuGR386	intact	376	1		A
167	tetur01g16754	TuGR387	intact	374	1		A
168	tetur18g03470	TuGR388	intact	366	1		A
169	tetur18g03821	TuGR390	intact	363	1		A
170	tetur18g03771	TuGR391	intact	366	1		A
171	tetur18g03791	TuGR392	intact	365	1		A
172	tetur18g03801	TuGR393	intact	369	1		A
173	tetur18g03741	TuGR394	intact	369	1		A
174	tetur18g03480	TuGR395	intact	368	1		A
175	tetur18g03841	TuGR396	intact	376	1		A
176	tetur17g00720	TuGR397	intact	368	1		A
177	tetur18g03781	TuGR398	intact	364	1		A
178	tetur01g16744	TuGR399	intact	368	1		A
179	tetur17g00730	TuGR400	intact	367	1		A
180	tetur15g03430	TuGR472	intact	375	1		A
181	tetur21g03400	TuGR483	intact	345	1		A
182	tetur265g00001	TuGR642	intact	372	1		A
183	tetur305g00020	TuGR644	intact	362	1		A
184	tetur144g00002	TuGR669	intact	359	1		A
185	tetur144g00003	TuGR670	intact	359	1		A
186	tetur144g00006	TuGR673	intact	359	1		A
187	tetur371g00001	TuGR689	intact	373	1		A
188	tetur06g05440	TuGR253	intact	375	2		A
1	tetur01g01400	TuGR6	intact	400	2		B
2	tetur27g01920	TuGR7	intact	451	2		B
3	tetur41g00690	TuGR8	intact	423	2		B
4	tetur41g00720	TuGR9	intact	423	2		B
5	tetur19g03411	TuGR10	intact	406	2		B
6	tetur01g16714	TuGR11	intact	394	2		B

7	tetur29g01290	TuGR12	intact	395	2		B
8	tetur01g07520	TuGR13	intact	400	2		B
9	tetur07g08067	TuGR14	intact	394	2		B
10	tetur07g08077	TuGR15	intact	386	2		B
11	tetur07g08057	TuGR16	intact	389	2		B
12	tetur17g03040	TuGR17	intact	396	2		B
13	tetur09g03430	TuGR18	intact	385	2		B
14	tetur09g06789	TuGR19	intact	384	2		B
15	tetur09g06728	TuGR20	intact	428	2		B
16	tetur30g02429	TuGR21	intact	381	2		B
17	tetur30g02479	TuGR22	intact	384	2		B
18	tetur30g02469	TuGR23	intact	378	2		B
19	tetur30g02439	TuGR24	intact	384	2		B
20	tetur30g02449	TuGR25	intact	384	2		B
21	tetur30g02459	TuGR26	intact	386	2		B
22	tetur04g08780	TuGR27	intact	412	2		B
23	tetur08g00330	TuGR28	intact	414	2		B
24	tetur16g00670	TuGR29	intact	403	2		B
25	tetur24g01040	TuGR30	intact	410	2		B
26	tetur17g03860	TuGR31	intact	411	2		B
27	tetur17g03870	TuGR32	intact	411	2		B
28	tetur01g05060	TuGR34	intact	419	3		B
29	tetur14g00710	TuGR35	intact	416	3		B
30	tetur03g09110	TuGR36	intact	445	3		B
31	tetur03g09100	TuGR38	intact	440	3		B
32	tetur10g05784	TuGR39	intact	431	3		B
33	tetur03g09090	TuGR40	intact	424	3		B
34	tetur11g06420	TuGR41	intact	480	3		B
35	tetur13g03050	TuGR42	intact	452	3		B
36	tetur13g04696	TuGR43	intact	449	3		B
37	tetur13g03090	TuGR44	intact	437	2		B
38	tetur13g03110	TuGR45	intact	451	3		B
39	tetur19g03401	TuGR46	intact	454	3		B
40	tetur10g02840	TuGR47	intact	417	3		B
41	tetur26g02843	TuGR48	intact	453	3		B
42	tetur11g04930	TuGR49	intact	418	3		B
43	tetur04g02760	TuGR50	intact	428	3		B
44	tetur11g01980	TuGR51	intact	446	3		B
45	tetur08g01370	TuGR52	intact	430	3		B
46	tetur08g01390	TuGR53	intact	429	3		B
47	tetur08g01420	TuGR54	intact	437	3		B
48	tetur08g01400	TuGR55	intact	429	3		B
49	tetur08g08329	TuGR56	intact	430	3		B
50	tetur08g08299	TuGR57	intact	416	3		B
51	tetur08g01380	TuGR58	intact	430	3		B
52	tetur08g01440	TuGR59	intact	430	3		B
53	tetur08g01510	TuGR60	intact	430	3		B
54	tetur08g01450	TuGR61	intact	424	3		B
55	tetur08g08309	TuGR62	intact	422	3		B
56	tetur01g14830	TuGR63	intact	430	3		B
57	tetur01g14820	TuGR64	intact	433	3		B
58	tetur01g16614	TuGR65	intact	441	3		B
59	tetur01g14840	TuGR66	intact	431	3		B
60	tetur01g14870	TuGR67	intact	423	3		B
61	tetur01g14880	TuGR68	intact	423	3		B
62	tetur01g14900	TuGR69	intact	457	3		B

63	tetur01g14950	TuGR70	intact	453	3		B
64	tetur01g16694	TuGR71	intact	438	3		B
65	tetur01g14970	TuGR72	intact	432	3		B
66	tetur44g00080	TuGR73	intact	417	3		B
67	tetur44g00110	TuGR74	intact	425	3		B
68	tetur44g00090	TuGR75	intact	422	3		B
69	tetur11g06380	TuGR76	intact	432	3		B
70	tetur33g01000	TuGR77	intact	432	3		B
71	tetur33g00990	TuGR78	intact	453	3		B
72	tetur33g01010	TuGR79	intact	447	3		B
73	tetur33g00980	TuGR80	intact	432	3		B
74	tetur33g01020	TuGR81	intact	451	3		B
75	tetur33g01030	TuGR82	intact	433	3		B
76	tetur32g02210	TuGR83	intact	432	3		B
77	tetur33g00070	TuGR84	intact	432	3		B
78	tetur32g02190	TuGR85	intact	432	3		B
79	tetur33g00100	TuGR86	intact	429	3		B
80	tetur33g00130	TuGR87	intact	435	3		B
81	tetur33g00150	TuGR88	intact	432	3		B
82	tetur33g00140	TuGR89	intact	432	3		B
83	tetur33g00090	TuGR90	intact	432	3		B
84	tetur33g00010	TuGR91	intact	433	3		B
85	tetur33g00110	TuGR92	intact	435	3		B
86	tetur04g02560	TuGR93	intact	432	3		B
87	tetur04g02570	TuGR94	intact	432	3		B
88	tetur17g00030	TuGR95	intact	425	3		B
89	tetur24g02550	TuGR96	intact	428	3		B
90	tetur17g00060	TuGR97	intact	426	3		B
91	tetur17g00070	TuGR98	intact	431	3		B
92	tetur04g03360	TuGR99	intact	428	3		B
93	tetur119g00010	TuGR100	intact	428	3		B
94	tetur19g02640	TuGR101	intact	421	3		B
95	tetur04g05470	TuGR102	intact	445	3		B
96	tetur04g05500	TuGR103	intact	445	3		B
97	tetur04g09504	TuGR104	intact	450	3		B
98	tetur04g05530	TuGR105	intact	445	3		B
99	tetur04g05590	TuGR106	intact	445	3		B
100	tetur04g05610	TuGR107	intact	446	3		B
101	tetur05g07870	TuGR108	intact	425	3		B
102	tetur24g00240	TuGR109	intact	454	3		B
103	tetur24g00260	TuGR110	intact	438	3		B
104	tetur28g00180	TuGR111	intact	430	3		B
105	tetur44g00271	TuGR112	intact	422	3		B
106	tetur19g02650	TuGR113	intact	437	3		B
107	tetur19g02700	TuGR114	intact	432	3		B
108	tetur07g01770	TuGR115	intact	424	3		B
109	tetur07g01820	TuGR116	intact	439	3		B
110	tetur131g00010	TuGR117	intact	428	3		B
111	tetur19g02630	TuGR118	intact	428	3		B
112	tetur19g02670	TuGR119	intact	439	3		B
113	tetur19g02610	TuGR120	intact	428	3		B
114	tetur02g00190	TuGR122	intact	440	3		B
115	tetur02g15287	TuGR123	intact	449	4		B
116	tetur02g08930	TuGR124	intact	403	3		B
117	tetur02g04180	TuGR125	intact	429	3		B
118	tetur02g04200	TuGR126	intact	429	3		B

119	tetur19g01600	TuGR127	intact	435	3		B
120	tetur10g05340	TuGR128	intact	435	3		B
121	tetur14g02630	TuGR129	intact	431	3		B
122	tetur22g00920	TuGR130	intact	432	3		B
123	tetur22g00910	TuGR131	intact	432	3		B
124	tetur22g00930	TuGR132	intact	432	3		B
125	tetur11g05240	TuGR133	intact	440	3		B
126	tetur17g03050	TuGR134	intact	427	3		B
127	tetur17g02940	TuGR135	intact	447	3		B
128	tetur17g02990	TuGR136	intact	443	3		B
129	tetur17g03070	TuGR137	intact	443	3		B
130	tetur17g03000	TuGR138	intact	445	3		B
131	tetur17g03020	TuGR139	intact	447	3		B
132	tetur17g03080	TuGR140	intact	447	3		B
133	tetur17g03980	TuGR141	intact	427	3		B
134	tetur17g02980	TuGR142	intact	455	3		B
135	tetur17g03090	TuGR143	intact	435	3		B
136	tetur17g03060	TuGR144	intact	422	3		B
137	tetur13g04560	TuGR145	intact	446	3		B
138	tetur13g04580	TuGR146	intact	447	3		B
139	tetur13g01070	TuGR147	intact	434	3		B
140	tetur13g04570	TuGR148	intact	448	3		B
141	tetur13g04670	TuGR149	intact	446	3		B
142	tetur13g04590	TuGR150	intact	446	3		B
143	tetur13g04600	TuGR151	intact	440	3		B
144	tetur13g04620	TuGR152	intact	444	3		B
145	tetur13g04630	TuGR153	intact	441	3		B
146	tetur13g04640	TuGR154	intact	441	3		B
147	tetur13g04650	TuGR155	intact	438	3		B
148	tetur13g04660	TuGR156	intact	457	3		B
149	tetur13g01100	TuGR157	intact	438	3		B
150	tetur17g04000	TuGR159	intact	435	3		B
151	tetur24g02737	TuGR160	intact	458	3		B
152	tetur24g02747	TuGR161	intact	442	3		B
153	tetur24g02757	TuGR162	intact	437	3		B
154	tetur24g02697	TuGR163	intact	438	3		B
155	tetur24g02707	TuGR164	intact	434	3		B
156	tetur24g02717	TuGR165	intact	444	3		B
157	tetur19g00260	TuGR166	intact	438	3		B
158	tetur17g01060	TuGR167	intact	433	3		B
159	tetur17g01070	TuGR168	intact	430	3		B
160	tetur17g01210	TuGR169	intact	430	3		B
161	tetur02g02230	TuGR170	intact	452	3		B
162	tetur03g08950	TuGR171	intact	463	3		B
163	tetur03g08960	TuGR172	intact	436	3		B
164	tetur03g09010	TuGR173	intact	455	3		B
165	tetur03g08990	TuGR174	intact	437	3		B
166	tetur03g08910	TuGR175	intact	445	3		B
167	tetur03g08920	TuGR176	intact	436	3		B
168	tetur03g08930	TuGR177	intact	436	3		B
169	tetur03g08790	TuGR178	intact	435	3		B
170	tetur05g09335	TuGR179	intact	431	3		B
171	tetur05g09345	TuGR180	intact	443	3		B
172	tetur05g04830	TuGR181	intact	410	3		B
173	tetur05g03560	TuGR182	intact	419	3		B
174	tetur04g01350	TuGR183	intact	449	3		B

175	tetur04g09577	TuGR184	intact	450	3		B
176	tetur04g01360	TuGR185	intact	427	3		B
177	tetur04g01310	TuGR186	intact	440	3		B
178	tetur04g01320	TuGR187	intact	436	3		B
179	tetur20g02810	TuGR188	intact	426	3		B
180	tetur20g02820	TuGR189	intact	455	3		B
181	tetur17g00110	TuGR190	intact	417	3		B
182	tetur17g00130	TuGR191	intact	417	3		B
183	tetur24g02670	TuGR192	intact	417	3		B
184	tetur04g04020	TuGR193	intact	456	3		B
185	tetur04g04070	TuGR194	intact	440	3		B
186	tetur04g04120	TuGR195	intact	440	3		B
187	tetur04g04090	TuGR196	intact	456	3		B
188	tetur04g04060	TuGR197	intact	425	3		B
189	tetur04g04080	TuGR198	intact	427	3		B
190	tetur16g02540	TuGR199	intact	437	3		B
191	tetur310g00010	TuGR200	intact	427	3		B
192	tetur04g02860	TuGR201	intact	426	3		B
193	tetur17g00250	TuGR202	intact	448	3		B
194	tetur17g00260	TuGR203	intact	427	3		B
195	tetur10g05440	TuGR204	intact	453	3		B
196	tetur10g05450	TuGR205	intact	453	3		B
197	tetur17g00320	TuGR206	intact	456	3		B
198	tetur17g00270	TuGR207	intact	448	3		B
199	tetur17g00340	TuGR209	intact	450	3		B
200	tetur27g01230	TuGR211	intact	443	3		B
201	tetur08g01490	TuGR439	intact	435	4		B
202	tetur08g08319	TuGR442	intact	430	3		B
203	tetur131g00030	TuGR464	intact	436	3		B
204	tetur04g09618	TuGR529	intact	440	3		B
205	tetur08g07160	TuGR535	intact	407	2		B
206	tetur08g08379	TuGR536	intact	417	3		B
207	tetur121g00010	TuGR544	intact	432	3		B
208	tetur121g00030	TuGR546	intact	432	3		B
209	tetur131g00050	TuGR552	intact	437	3		B
210	tetur150g00001	TuGR565	intact	438	3		B
211	tetur17g00010	TuGR574	intact	443	3		B
212	tetur17g00040	TuGR575	intact	435	3		B
213	tetur17g03880	TuGR576	intact	408	2		B
214	tetur17g04010	TuGR577	intact	437	3		B
215	tetur17g04030	TuGR579	intact	447	3		B
216	tetur17g04040	TuGR580	intact	438	3		B
217	tetur17g04050	TuGR581	intact	443	3		B
218	tetur17g04060	TuGR582	intact	439	3		B
219	tetur17g04080	TuGR584	intact	450	3		B
220	tetur17g04090	TuGR585	intact	444	3		B
221	tetur17g04100	TuGR586	intact	444	3		B
222	tetur17g04110	TuGR587	intact	434	3		B
223	tetur17g04120	TuGR588	intact	441	3		B
224	tetur17g04150	TuGR590	intact	437	3		B
225	tetur17g04160	TuGR591	intact	443	3		B
226	tetur17g04170	TuGR592	intact	444	3		B
227	tetur17g04180	TuGR593	intact	440	3		B
228	tetur17g04190	TuGR594	intact	439	3		B
229	tetur17g04210	TuGR596	intact	438	3		B
230	tetur17g04220	TuGR597	intact	443	3		B

231	tetur24g02777	TuGR620	intact	443	3		B
232	tetur24g02787	TuGR621	intact	435	3		B
233	tetur24g02797	TuGR622	intact	439	3		B
234	tetur24g02807	TuGR623	intact	438	3		B
235	tetur24g02827	TuGR624	intact	446	3		B
236	tetur24g02867	TuGR628	intact	434	3		B
237	tetur24g02877	TuGR629	intact	439	3		B
238	tetur24g02887	TuGR630	intact	436	3		B
239	tetur24g02917	TuGR632	intact	446	4		B
240	tetur24g02937	TuGR634	intact	443	3		B
241	tetur373g00010	TuGR648	intact	436	3		B
242	tetur629g00010	TuGR655	intact	445	3		B
243	tetur85g00001	TuGR657	intact	442	3		B
244	tetur85g00002	TuGR658	intact	443	3		B
245	tetur85g00003	TuGR659	intact	440	3		B
1	tetur19g02720	TuGR1	intact	409	0		C
2	tetur06g01480	TuGR2	intact	403	4		C
3	tetur01g06380	TuGR210	intact	403	2		C
4	tetur20g00810	TuGR212	intact	410	2		C
5	tetur03g02210	TuGR213	intact	453	3		C
6	tetur10g05170	TuGR214	intact	456	3		C
7	tetur22g02600	TuGR215	intact	437	3		C
8	tetur05g00780	TuGR216	intact	429	3		C
9	tetur07g04370	TuGR3	intact	476	3		C
10	tetur06g00770	TuGR33	intact	434	2		C
11	tetur02g09560	TuGR4	intact	412	4		C
12	tetur11g00460	TuGR5	intact	377	0		C
13	tetur11g00450	TuGR510	intact	397	0		C
14	tetur05g08450	TuGR511	intact	361	0		C
15	tetur04g06520	TuGR528	intact	489	1		C
16	tetur11g06470	TuGR543	intact	401	0		C

1	tetur103g00001	TuGR542	partial	133	1	N-ter missing	A
2	tetur398g00020	TuGR649	partial	155	1	N-ter missing	A
3	tetur144g00007	TuGR674	partial	179	0	N-ter missing	A
4	tetur265g00011	TuGR684	partial	337	1	N-ter missing	A
5	tetur617g00001	TuGR691	partial	120	1	C-terminus	A
6	tetur618g00001	TuGR692	partial	272	0	int.seq. missing	A
1	tetur445g00010	TuGR121	partial	360	3	N-ter missing	B
2	tetur13g01130	TuGR467	partial	423	3	N-ter missing	B
3	tetur24g02600	TuGR506	partial	358	0	C-ter missing	B
4	tetur438g00010	TuGR507	partial	161	3	N-ter missing	B
5	tetur456g00010	TuGR508	partial	363	3	N-ter missing	B
6	tetur84g00060	TuGR509	partial	375	0	C-ter missing	B
7	tetur13g04846	TuGR559	partial	302	1	int.seq. missing	B
8	tetur24g02957	TuGR636	partial	305	3	int.seq. missing	B
9	tetur32g02250	TuGR645	partial	354	3	N-ter missing	B
10	tetur310g00020	TuGR686	partial	251	0	C-ter missing	B
11	tetur150g00002	TuGR566	partial	395	2	C-ter missing	B
12	tetur545g00001	TuGR653	partial	75	1	C-terminus	B
13	tetur592g00001	TuGR654	partial	428	3	N-ter missing	B
14	tetur85g00004	TuGR660	partial	169	3	N-ter missing	B
15	tetur85g00005	TuGR661	partial	326	0	C-ter missing	B
16	tetur529g00010	TuGR675	partial	191	3	N-ter missing	B

1	tetur08g08190	TuGR353	pseudo	349	1	1n < frameshift	A

2	tetur08g08180	TuGR362	pseudo	352	1	RT insertion	A
3	tetur01g16824	TuGR413	pseudo	362	1	premature stop	A
4	tetur02g06700	TuGR414	pseudo	350	1	1n < frameshift	A
5	tetur02g15247	TuGR417	pseudo	350	1	premature stop	A
6	tetur08g00740	TuGR436	pseudo	347	1	premature stop	A
7	tetur08g00760	TuGR437	pseudo	351	1	1n< frameshift	A
8	tetur08g08120	TuGR440	pseudo	326	1	int.seq. deletion	A
9	tetur08g08349	TuGR443	pseudo	353	1	premature stop	A
10	tetur12g00820	TuGR456	pseudo	377	1	1n< frameshift	A
11	tetur12g01700	TuGR457	pseudo	368	1	premature stop	A
12	tetur12g04733	TuGR460	pseudo	383	1	9n< frameshift	A
13	tetur12g04813	TuGR461	pseudo	381	1	premature stop	A
14	tetur12g04883	TuGR463	pseudo	370	1	1n< frameshift	A
15	tetur15g00450	TuGR471	pseudo	378	1	premature stop	A
16	tetur167g00010	TuGR473	pseudo	370	1	1n< frameshift	A
17	tetur16g04016	TuGR475	pseudo	350	1	RT insertion	A
18	tetur16g04116	TuGR476	pseudo	358	1	premature stop	A
19	tetur20g03310	TuGR480	pseudo	358	1	premature stop	A
20	tetur21g03040	TuGR482	pseudo	345	1	1n< frameshift	A
21	tetur34g01223	TuGR497	pseudo	375	1	1n > frameshift	A
22	tetur47g00160	TuGR499	pseudo	336	0	int.seq. deletion	A
23	tetur66g00020	TuGR500	pseudo	378	1	1n< frameshift	A
24	tetur09g06830	TuGR541	pseudo	373	1	1n< frameshift	A
25	tetur18g03853	TuGR611	pseudo	369	1	25bp insertion	A
26	tetur21g03432	TuGR617	pseudo	348	1	4n< frameshift	A
27	tetur47g00230	TuGR650	pseudo	356	1	MITE insertion	A
28	tetur144g00005	TuGR672	pseudo	359	1	2n< frameshift	A
1	tetur02g06850	TuGR415	pseudo	354	0	2 frameshifts	A
2	tetur07g03770	TuGR434	pseudo	385	1	2 frameshifts	A
3	tetur07g05010	TuGR435	pseudo	315	1	FS & C-ter del	A
4	tetur16g03820	TuGR474	pseudo	388	1	2 frameshifts	A
5	tetur20g03330	TuGR481	pseudo	354	1	2 frameshifts	A
6	tetur16g04117	TuGR569	pseudo	360	1	RTI& FS	A
7	tetur144g00001	TuGR668	pseudo	347	1	2 frameshifts	A
8	tetur167g00040	TuGR681	pse/partial	293	0	C-ter miss & FS	A
9	tetur17g04272	TuGR682	pseudo	370	1	2 frameshifts	A
10	tetur329g00001	TuGR688	pse/partial	249	1	N-ter miss & PS	A
1	tetur02g14290	TuGR416	pseudo	359	1	2FS & 1PS	A
2	tetur06g04870	TuGR433	pseudo	384	0	2 FS & 1 PS	A
3	tetur08g08256	TuGR441	pseudo	343	1	3FS & I.S.Del	A
4	tetur12g03990	TuGR458	pseudo	164	0	N-ter relics	A
5	tetur12g04030	TuGR459	pseudo	339	1	FS & I.S.del	A
6	tetur12g04873	TuGR462	pseudo	108	0	relics	A
7	tetur34g01213	TuGR496	pseudo	309	0	N-ter fragment	A
8	tetur66g00100	TuGR501	pseudo	78	0	relics	A
9	tetur66g00110	TuGR502	pseudo	184	0	N-ter relics	A
10	tetur66g00130	TuGR503	pseudo	71	1	relics	A
11	tetur07g05020	TuGR504	pse/partial	295	0	4PS & N gap	A
12	tetur01g16934	TuGR512	pseudo	247	1	relics	A
13	tetur02g15347	TuGR514	pseudo	209	1	relics	A
14	tetur02g15349	TuGR516	pseudo	347	1	no start, del &TEI	A
15	tetur02g15350	TuGR517	pseudo	133	0	relics	A
16	tetur02g15351	TuGR518	pseudo	252	0	relics	A
17	tetur02g15352	TuGR519	pseudo	187	1	relics	A
18	tetur02g15353	TuGR520	pseudo	126	0	relics	A
19	tetur02g15354	TuGR521	pseudo	68	0	relics	A

20	tetur02g15355	TuGR522	pseudo	109	1	relics	A
21	tetur03g10153	TuGR523	pseudo	81	0	relics	A
22	tetur06g06773	TuGR532	pseudo	96	0	relics	A
23	tetur07g08119	TuGR534	pseudo	295	0	2 INS & 2 FS	A
24	tetur08g08389	TuGR537	pseudo	176	1	relics	A
25	tetur12g04914	TuGR547	pseudo	124	0	N-ter relics	A
26	tetur12g04915	TuGR548	pseudo	69	0	N-ter relics	A
27	tetur12g04916	TuGR549	pseudo	175	1	N-ter relics	A
28	tetur16g04118	TuGR570	pseudo	279	0	N-ter relics	A
29	tetur16g04119	TuGR571	pseudo	104	0	C-ter relics	A
30	tetur18g03852	TuGR610	pseudo	364	1	3 frameshifts	A
31	tetur18g03854	TuGR612	pseudo	372	1	major changes	A
32	tetur18g03855	TuGR613	pseudo	77	0	C-ter relics	A
33	tetur21g03430	TuGR616	pseudo	231	1	C-ter relics	A
34	tetur21g03433	TuGR618	pseudo	104	0	relics	A
35	tetur24g02958	TuGR637	pseudo	129	0	relics	A
36	tetur24g02959	TuGR638	pseudo	253	0	N-ter relics	A
37	tetur24g02960	TuGR639	pseudo	139	0	relics	A
38	tetur24g02961	TuGR640	pseudo	150	0	relics	A
39	tetur27g02599	TuGR643	pseudo	113	0	N-ter relics	A
40	tetur520g00010	TuGR651	pseudo	95	1	C-ter relics	A
41	tetur520g00021	TuGR652	pseudo	352	1	3 FS & 3 PS	A
42	tetur18g03865	TuGR665	pseudo	371	0	4 frameshifts	A
43	tetur18g03885	TuGR666	pseudo	69	1	C-ter relics	A
44	tetur144g00004	TuGR671	pseudo	292	1	major changes	A
45	tetur20g03457	TuGR683	pseudo	105	0	N-ter relics	A
46	tetur305g00030	TuGR685	pseudo	90	1	C-ter relics	A
47	tetur66g00141	TuGR693	pseudo	87	0	N-ter relics	A

1	tetur32g02220	TuGR505	pse/partial	343	3	N del/missing	B
2	tetur03g09120	TuGR37	pseudo	444	2	4n > frameshift	B
3	tetur24g02727	TuGR158	pseudo	441	3	1n < frameshift	B
4	tetur17g00280	TuGR208	pseudo	448	3	4n < frameshift	B
5	tetur01g14810	TuGR401	pseudo	446	3	premature stop	B
6	tetur01g14850	TuGR402	pseudo	450	4	RT insertion	B
7	tetur01g16594	TuGR404	pseudo	429	3	premature stop	B
8	tetur03g08970	TuGR419	pseudo	435	3	5n < frameshift	B
9	tetur03g09000	TuGR421	pseudo	437	3	premature stop	B
10	tetur04g01330	TuGR422	pseudo	426	3	premature stop	B
11	tetur04g04050	TuGR425	pseudo	424	3	1n < frameshift	B
12	tetur04g04100	TuGR426	pseudo	452	3	1n > frameshift	B
13	tetur04g04110	TuGR427	pseudo	440	3	premature stop	B
14	tetur04g09597	TuGR432	pseudo	428	3	1n< frameshift	B
15	tetur08g01430	TuGR438	pseudo	431	3	premature stop	B
16	tetur08g08359	TuGR444	pseudo	429	3	1n > frameshift	B
17	tetur09g03680	TuGR449	pseudo	394	2	1n > frameshift	B
18	tetur10g05330	TuGR454	pseudo	445	2	premature stop	B
19	tetur13g01060	TuGR465	pseudo	445	3	1n< frameshift	B
20	tetur13g04610	TuGR470	pseudo	429	3	premature stop	B
21	tetur17g02970	TuGR479	pseudo	445	3	1n< frameshift	B
22	tetur24g02520	TuGR484	pseudo	439	3	1n< frameshift	B
23	tetur30g00870	TuGR487	pseudo	381	2	1n< frameshift	B
24	tetur32g02200	TuGR488	pseudo	302	3	int.seq. deletion	B
25	tetur32g02230	TuGR489	pseudo	431	2	G del at 3' of I-2	B
26	tetur32g02240	TuGR490	pseudo	435	3	1n< frameshift	B
27	tetur33g00120	TuGR495	pseudo	432	3	MITE insertion	B

28	tetur05g00020	TuGR531	pseudo	240	0	N-ter moeity	B
29	tetur08g08402	TuGR540	pseudo	355	3	int.seq. deletion	B
30	tetur13g04866	TuGR561	pseudo	457	3	1n< frameshift	B
31	tetur13g04886	TuGR562	pseudo	457	3	RT insertion	B
32	tetur150g00003	TuGR567	pseudo	438	3	1n< frameshift	B
33	tetur172g00030	TuGR573	pseudo	287	3	C-ter moiety	B
34	tetur17g04070	TuGR583	pseudo	461	3	MITE insertion	B
35	tetur24g02590	TuGR619	pseudo	400	3	int.seq. deletion	B
36	tetur598g00020	TuGR677	pse/partial	413	2	int.seq. deletion	B
1	tetur24g02570	TuGR485	pse/partial	425	3	PS & N gap	B
2	tetur33g00030	TuGR492	pseudo	434	3	2 frameshifts	B
3	tetur13g04906	TuGR564	pseudo	434	3	2 frameshifts	B
4	tetur150g00004	TuGR568	pseudo	438	3	2 frameshifts	B
5	tetur172g00020	TuGR572	pseudo	447	3	2 frameshifts	B
6	tetur17g04020	TuGR578	pseudo	447	3	FS & PS	B
7	tetur24g02837	TuGR625	pse/partial	335	3	N-ter mis & -1FS	B
8	tetur24g02857	TuGR627	pseudo	439	3	PS & 13bp del	B
9	tetur686g00010	TuGR656	pse/partial	246	1	N&C-ter mis, FS	B
10	tetur85g00006	TuGR662	pse/partial	283	0	N&C-ter mis, PS	B
11	tetur585g00010	TuGR676	pse/partial	267	2	N-ter miss &TEI	B
12	tetur454g00001	TuGR690	pseudo	464	3	2 frameshifts	B
1	tetur01g14940	TuGR403	pseudo	323	1	I.S.Del & PS	B
2	tetur01g16604	TuGR405	pseudo	444	3	many PS & FS	B
3	tetur01g16624	TuGR406	pseudo	202	0	many changes	B
4	tetur01g16634	TuGR407	pseudo	186	0	C-ter del & FS	B
5	tetur01g16644	TuGR408	pseudo	86	0	relics	B
6	tetur01g16654	TuGR409	pseudo	435	3	many changes	B
7	tetur01g16664	TuGR410	pseudo	160	0	many changes	B
8	tetur01g16704	TuGR412	pseudo	410	2	many changes	B
9	tetur03g08830	TuGR418	pseudo	381	1	many changes	B
10	tetur03g08980	TuGR420	pseudo	346	3	many changes	B
11	tetur04g04030	TuGR423	pseudo	373	3	I.S.Del & FS	B
12	tetur04g04040	TuGR424	pseudo	354	3	many changes	B
13	tetur04g05570	TuGR428	pseudo	250	2	int.seq. deletion	B
14	tetur04g05580	TuGR429	pseudo	243	1	N-ter missing	B
15	tetur04g05630	TuGR430	pseudo	415	2	FS & C-ter del	B
16	tetur04g09547	TuGR431	pseudo	72	0	N-ter relics	B
17	tetur09g03420	TuGR445	pseudo	328	2	N FS & I.S.Del	B
18	tetur09g03450	TuGR446	pseudo	193	1	many changes	B
19	tetur09g03470	TuGR447	pseudo	95	2	relics	B
20	tetur09g03490	TuGR448	pseudo	107	0	relics	B
21	tetur09g06749	TuGR450	pseudo	148	1	relics	B
22	tetur09g06759	TuGR451	pseudo	238	1	major changes	B
23	tetur09g06769	TuGR452	pseudo	78	0	relics	B
24	tetur09g06779	TuGR453	pseudo	55	0	N-ter relics	B
25	tetur11g05530	TuGR455	pseudo	262	4	major changes	B
26	tetur13g01080	TuGR466	pseudo	223	6	major changes	B
27	tetur13g01150	TuGR468	pseudo	388	3	Nter-Del & FS	B
28	tetur13g03100	TuGR469	pseudo	358	3	FS & 2 Dels	B
29	tetur17g00290	TuGR477	pseudo	430	3	major changes	B
30	tetur17g00310	TuGR478	pseudo	459	3	RTI, PS & FS	B
31	tetur29g01796	TuGR486	pseudo	119	1	relics	B
32	tetur33g00020	TuGR491	pseudo	56	0	relics	B
33	tetur33g00040	TuGR493	pseudo	91	0	N-ter relics	B
34	tetur33g00080	TuGR494	pseudo	137	3	C-ter relics	B
35	tetur46g00160	TuGR498	pseudo	439	3	many FS	B

36	tetur02g15348	TuGR515	pseudo	107	0	relics	B
37	tetur03g10154	TuGR524	pseudo	72	0	relics	B
38	tetur03g10155	TuGR525	pseudo	69	0	relics	B
39	tetur03g10156	TuGR526	pseudo	168	0	N-ter relics	B
40	tetur03g10157	TuGR527	pseudo	148	0	relics	B
41	tetur04g09619	TuGR530	pseudo	249	2	Ndel, PS & FS	B
42	tetur07g08118	TuGR533	pseudo	77	0	relics	B
43	tetur08g08400	TuGR538	pseudo	93	3	C-ter relics	B
44	tetur08g08401	TuGR539	pseudo	92	0	relics	B
45	tetur121g00020	TuGR545	pseudo	136	0	N-ter relics	B
46	tetur131g00020	TuGR550	pseudo	127	0	N-ter relics	B
47	tetur131g00040	TuGR551	pseudo	270	1	major changes	B
48	tetur131g00052	TuGR553	pseudo	108	0	N-ter relics	B
49	tetur13g04776	TuGR554	pseudo	306	0	N-ter relics	B
50	tetur13g04786	TuGR555	pseudo	203	2	C-ter relics	B
51	tetur13g04796	TuGR556	pseudo	494	3	Repeat Ins. & FS	B
52	tetur13g04806	TuGR557	pseudo	112	0	relics	B
53	tetur13g04836	TuGR558	pseudo	110	0	relics	B
54	tetur13g04856	TuGR560	pseudo	101	0	relics	B
55	tetur17g04130	TuGR589	pseudo	430	3	3 frameshifts	B
56	tetur17g04200	TuGR595	pseudo	438	2	major changes	B
57	tetur17g04230	TuGR598	pseudo	199	0	N-ter moiety	B
58	tetur17g04231	TuGR599	pseudo	180	3	C-ter moiety	B
59	tetur17g04232	TuGR600	pseudo	96	0	N-ter relics	B
60	tetur17g04233	TuGR601	pseudo	79	0	N-ter relics	B
61	tetur17g04234	TuGR602	pseudo	93	0	relics	B
62	tetur17g04235	TuGR603	pseudo	119	0	N-ter relics	B
63	tetur17g04236	TuGR604	pseudo	157	0	relics	B
64	tetur17g04237	TuGR605	pseudo	451	3	3 frameshifts	B
65	tetur17g04238	TuGR606	pseudo	90	0	N-ter relics	B
66	tetur17g04239	TuGR607	pseudo	184	0	relics	B
67	tetur17g04240	TuGR608	pseudo	429	0	N-terM & TEI	B
68	tetur17g04241	TuGR609	pseudo	156	0	relics	B
69	tetur19g02620	TuGR614	pseudo	125	0	N-ter relics	B
70	tetur19g03472	TuGR615	pseudo	76	0	N-ter relics	B
71	tetur24g02847	TuGR626	pse/partial	413	3	PS, N gap & del	B
72	tetur24g02897	TuGR631	pseudo	226	0	relics	B
73	tetur24g02927	TuGR633	pseudo	225	0	N-ter relics	B
74	tetur24g02947	TuGR635	pseudo	193	0	relics	B
75	tetur24g02963	TuGR641	pseudo	148	0	N-ter relics	B
76	tetur33g01720	TuGR646	pseudo	421	3	major changes	B
77	tetur33g01721	TuGR647	pseudo	251	0	N-ter relics	B
78	tetur11g06472	TuGR663	pseudo	166	0	relics	B
79	tetur17g04252	TuGR664	pseudo	117	1	C-ter relics	B
80	tetur141g00001	TuGR667	pseudo	118	0	relics	B
81	tetur601g00010	TuGR678	pseudo	93	1	relics	B
82	tetur68g00040	TuGR679	pseudo	118	1	relics	B
83	tetur150g00014	TuGR680	pseudo	118	0	relics	B
84	tetur85g00017	TuGR687	pseudo	168	0	relics	B
85	tetur679g00001	TuGR694	pseudo	92	1	relics	B
1	tetur01g16944	TuGR513	pseudo	123	0	relics	C

A.1.3. Expression of chemoreceptor genes in *T. urticae*

Table A.2. The ENaCs-encoding genes and their expression in *T. urticae*

Gene	Entry_ID	pseudo	size aa	N_exons	Expres.S	Expres.Q	Remark	
ENaC01	tetur01g14980		493	1	N	0	downstream of 2 neighbors tetur01g14970 & 60 : CR genes	
ENaC02	tetur25g00590		524	1	AraTom	1,36	in between the two Jouberin genes Ahi1 & Ahi1L	
ENaC03	tetur01g15350		487	1	Tom	13,9	divergent neighbor gene: ABC transporter (detox)	
ENaC04	tetur06g03730		484	1	Tom	8,9		
ENaC05	tetur43g00480		487	1	AraTom	0,96		
ENaC06	tetur41g00660		461	1	N	0	upstream a Cys-loop neurotransmitter-gated ion-channel	
ENaC07	tetur13g00530		468	1	N	0	inserted inside the H2A-H2B-MZT2 gene cluster	
ENaC08	tetur08g03130		431	1	Bean	1,23		
ENaC09	tetur24g00040		473	1	N	0	allele : tetur538g00010	
EnaC10	tetur24g01660		472	1	N	0	allele: tetur577g00020	
EnaC11	tetur04g03870		481	1	Tom	0,34	3' divergent to hydroxyacylglutathione hydrolase (detox)	
EnaC12	tetur19g03391		481	1	N	0	upstream a gene for a small secreted protein	
ENaC13	tetur01g01020		530	1	Kon	0,26		
ENaC14	tetur01g01030		476	1	N	0		
ENaC15	tetur01g01040		496	1	N	0		
ENaC16	tetur01g01050	partial	274	1	N	0		
ENaC17	tetur01g01070		479	1	N	0		
ENaC18	tetur01g01080	ps:fs	494	1	N	0		
ENaC19	tetur01g01090	ps:st	489	1	N	0		
ENaC20	tetur01g02290		487	1	N	0		
ENaC21	tetur03g07530		427	1	Bla	0,2	divergent gene: Receptor Protein Tyrosine Phosphatase	
ENaC22	tetur04g02740		485	1	Kon	0,1	tetur04g02760 very near: CR gene, expressed on tomato	
ENaC23	tetur04g03920		484	3	Emb	0,78	3' divergent to a gene for a small secreted protein	rel. UNCs
UNC105L	tetur05g03680		785	9	Tom	0,8	homolog of UNC105 from *C. elegans* (muscle contraction)	
UNC8L	tetur09g00460		712	9	Bean	0,27	homolog of UNC8 from *C. elegans* (proprioception: Coordonit.)	
UNC8p2	tetur09g00430	relic	199	3			tandem copy of UNC8L, embedded in tandem copies of ADAM	
UNC8p1	tetur09g06717	relic	279	4			tandem copy of UNC8L, embedded in tandem copies of ADAM	

Table A.3. Expression of TuGRs*, TuIRs, TuXR and their related genes in *T.urticae*. * Only the TuGR genes showing expression higher than 1 fpkm are shown in this table

tetur geneID	S_Name	P	type	length	N.Int.	bean	beanx	arab. bla	arab. blax	arab. konx	tom.	adult	embr.	larv.	nymp.	AS
tetur02g09360	TuGRIN2		NMDA-2	1137	12	0.4509	0.0986	0.53	0.16	0.08	0.524	0.049	0.529	0.55	0.266	
tetur03g03430	TuiGluR8		AMPA/RI	891	16	0.851	0.025	1.17	0.12	0.08	0.981	0.033	0.605	0.59	0.229	
tetur04g00420	TuiGluR12	pseudo	AMPA/RII	865	11	0	0	0	0	0	0	0	0	0	0	
tetur04g00480	TuiGluR11		AMPA/RII	883	10	0.5096	0.3645	0.54	0.92	0.88	0.587	0.617	0.279	1.17	0.866	
tetur04g01860	TuiGluR7		Kainate/RII	929	12	1.0838	0.2851	1.3	0.06	0.11	3.487	0.122	0.253	0.16	0.13	
tetur04g01930	TuiGluR4		Kainate/RII	903	12	0.1905	0.0263	0.24	0.03	0.03	0.159	0.005	0.1	0.24	0.078	
tetur06g03760	TuiGluR6		Kainate/RII	932	11	0.1773	0.0901	0.23	2.22	1.34	0.299	0	8.602	6.47	1.392	
tetur07g05440	TuiGluR5		Kainate/RII	933	12	0.1241	0	0.15	0.88	0	0.371	0	0.128	0.89	0.156	
tetur07g06860	TuiGluR2		Kainate/RII	910	10	0.3146	0	0.6	0.15	0	0.482	0.166	0.46	0.45	0.187	
tetur09g02270	TuiGluR1		Kainate/RII	925	10	2.5483	0.4459	2.71	0.93	0.97	2.805	0.296	2.101	1.99	1.217	
tetur12g00180	TuiGluR3		Kainate/RII	870	12	0.1715	0.057	0.26	0.08	0.03	0.28	0.047	0.145	0.24	0.108	
tetur22g01330	TuGRIN3		NMDA-3	1072	11	0.7966	0.7204	1	0.55	0.7	0.862	0.475	0.712	0.71	0.722	
tetur31g00200	TuiGluR9		NMDA-1	1235	6	16.388	5.9997	15.9	8.66	8.3	21.37	3.859	7.01	11.1	5.899	
tetur43g00460	TuiGluR10		NMDA-1	1099	5	0.9552	0.3287	0.9	0	0	0.776	0	0.625	0.58	0.676	
tetur110g00050	TuiGluR13		AMPA/RII	892	10	0.0058	0	0.06	0.02	0	0.183	0.04	0.022	0.06	0.027	
tetur02g05540	TuIR2		IR93A homolog	1020	6	6.469	1.7045	7.26	3.94	3.55	6.992	1.181	2.842	4.04	2.019	
tetur03g07510	TuIR3		TuIR1 paralog	938	8	0.0437	0	0.04	0	0	0	0	0	0	0	
tetur13g03980	TuIR4		TuIR1 paralog	940	8	1.1589	1.5512	0.85	2.31	1.58	1.344	0.746	0.459	0.32	0.824	
tetur30g01550	TuIR1		IR25A homolog	942	8	0.814	0.3943	1.15	0.83	0.21	0.972	0	0.278	0.36	0.332	
tetur04g01970	TumGRI		mGluR-I	1380	9	0.1062	0.0208	0.15	0.02	0.03	0.125	0.03	0.103	0.14	0.081	

tetur06g06400	TumGRIII		mGluR-III	893	7	0.3001	0.0937	0.33	0.06	0.07	0.346	0.075	0.582	0.15	0.05	
tetur40g00040	TumGRA		mGluR-II	879	7	0.4866	0.4694	0.39	0.45	0.46	0.492	0.423	0.476	0.36	0.618	
tetur04g09010	TumXR		mXR (MTT)	1107	7	0.3435	0	0.33	0.04	0.07	0.247	0.023	0.173	0.21	0.045	
tetur01g01400	TuGR6	intact	GR-A	400	2	4.8408	1.5008	4.54	0.99	1.02	3.651	0.079	2.94	2.29	1.238	
tetur01g05060	TuGR34	intact	GR-B	419	3	0.1523	0.7302	0.33	1.05	0.43	0.37	0	1.249	0.17	0.509	
tetur01g16774	TuGR254	intact	GR-A	404	1	2.8055	4.138	4.91	3.33	3.72	5.4	0.632	5.884	3.53	2.826	Y
tetur02g00190	TuGR122	intact	GR-B	443	3	0.8262	0.136	0.98	0	0	1.688	0	0.401	0.32	0.418	
tetur02g02230	TuGR170	intact	GR-B	452	3	0.4716	0	0.68	0	0.09	0.329	0	1.16	0.35	0	
tetur02g08860	TuGR296	intact	GR-A	341	1	7.2657	1.7217	5.4	3.07	3.81	5.116	0	8.012	5.76	3.241	
tetur02g09560	TuGR4	intact	GR-C	412	4	0.8082	0.679	0.67	1.05	0.86	1.497	0.134	0.421	0.55	0.396	
tetur03g02210	TuGR213	intact	GR-C	453	3	1.0617	0.1299	1.32	1.57	0	2.292	0.145	0.16	0.73	0.155	
tetur03g09090	TuGR40	intact	GR-B	424	3	0.6825	0	1.2	0	0	1.189	0	0	0.36	0.071	
tetur05g03560	TuGR182	intact	GR-B	453	3	1.2458	0.1126	1.79	0.41	0.38	1.361	0	1.143	0.42	0.284	Y
tetur05g04830	TuGR181	intact	GR-B	410	3	1.9422	1.6951	2.93	3.05	2.09	2.728	0.833	0.941	1.28	1.742	Y
tetur05g09335	TuGR179	intact	GR-B	431	3	1.6365	2.5223	3.42	2.99	2.67	3.174	0.814	0.728	1.52	2.764	Y
tetur05g09345	TuGR180	intact	GR-B	443	3	1.5225	2.5505	2.88	1.65	2.02	2.726	2.001	0.208	1.57	2.773	Y
tetur05g07870	TuGR108	intact	GR-B	425	3	1.8097	0.5402	2.19	0.92	1.65	4.619	0.109	3.651	3.22	1.62	
tetur06g01480	TuGR2	intact	GR-C	403	4	3.4476	0.1868	2.57	0.86	0.44	2.082	0	0.397	0.92	0.633	Y
tetur06g05440	TuGR253	intact	GR-A	375	2	3.6426	2.5749	4.53	3.95	2.64	3.572	0	2.19	0.73	1.842	
tetur07g08097	TuGR285	intact	GR-A	350	1	0.4835	0.1661	1.53	0	0	0.149	0	0	0	0	
tetur09g02300	TuGR371	intact	GR-A	370	1	1.6339	0.3529	1.73	0.69	0.2	1.798	0	0.912	1.27	0.221	
tetur10g05784	TuGR39	intact	GR-B	431	3	0.8747	0	1.18	0.19	0	1.678	0	0	0	0.069	
tetur10g05170	TuGR214	intact	GR-C	456	3	0.3894	0	0.05	0	0	0.105	0	0.478	1.15	0.38	
tetur11g00460	TuGR5	intact	GR-C	377	0	0.3779	0	0.71	0.11	0.12	1.181	0	0.078	0.43	0.089	
tetur12g01060	TuGR366	intact	GR-A	361	1	0.996	0.8547	0.72	1.08	0.78	0.732	0.897	0.272	0.68	0.899	
tetur13g03110	TuGR45	intact	GR-B	451	3	0.0772	1.1418	0	1.46	1.64	0	0	0	0	0.906	Y

tetur13g04410	TuGR252	intact	GR-A	375	1	1.6199	0.387	1.85	0.12	0.18	1.793	0.337	0.516	2.19	0.319
tetur17g02940	TuGR135	intact	GR-B	447	3	2.2204	0.14	0.24	0.27	0.48	0.172	0	0.225	0.8	0.239
tetur19g01190	TuGR307	intact	GR-A	389	1	0.6077	0.981	0.94	0.33	0.29	0.725	0.075	0.596	0.44	1.072
tetur19g03411	TuGR10	intact	GR-B	406	2	1.8497	1.5135	2.01	1.61	0.95	1.718	1.088	0.906	1.04	1.841
tetur19g02720	TuGR1	intact	GR-C	409	0	0.7176	2.1287	1.08	2.14	2.49	1.703	1.999	1.441	1.3	2.479
tetur20g00810	TuGR212	intact	GR-C	410	2	2.5066	0.6445	4.53	2.06	2.94	3.947	0.21	0.619	1.11	1.686
tetur24g00910	TuGR256	intact	GR-A	415	1	6.7417	0.2046	6.23	1.44	1.52	5.852	0	4.02	2.96	1.308
tetur24g01040	TuGR30	intact	GR-B	410	2	1.7145	0.1218	0.59	1.12	0.81	1.381	0	0.552	0.4	0.531
tetur25g02112	TuGR309	intact	GR-A	403	1	0.3834	0	1.04	0	0	0.761	0	0	0	0.049
tetur28g00180	TuGR111	intact	GR-B	430	3	0.3212	0	0.85	0.3	0.12	1.27	0.192	0.715	0.85	0.14
tetur37g00370	TuGR251	intact	GR-A	385	1	0.3648	0	0.89	0	0.12	2.095	0	0.389	0	0.114

	very high expression; FPKM > 5
	high expression; 5> FPKM >2
	lower expression 2 > FPKM > 0.5

A.1.4. Anti-sense expression of several chemoreceptor genes

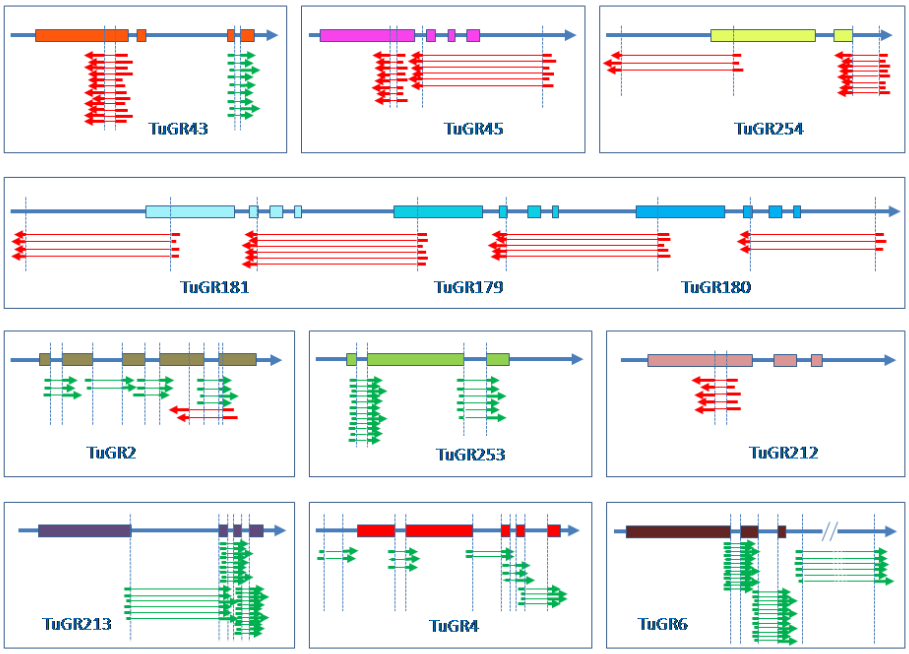


Figure A.6. Many chemoreceptor genes have anti-sense expression. On top: gene models in the sense orientation; coloured boxes: coding exons. Below: RNA-seq junction reads location in sense orientation (green), or anti-sense orientation (red).

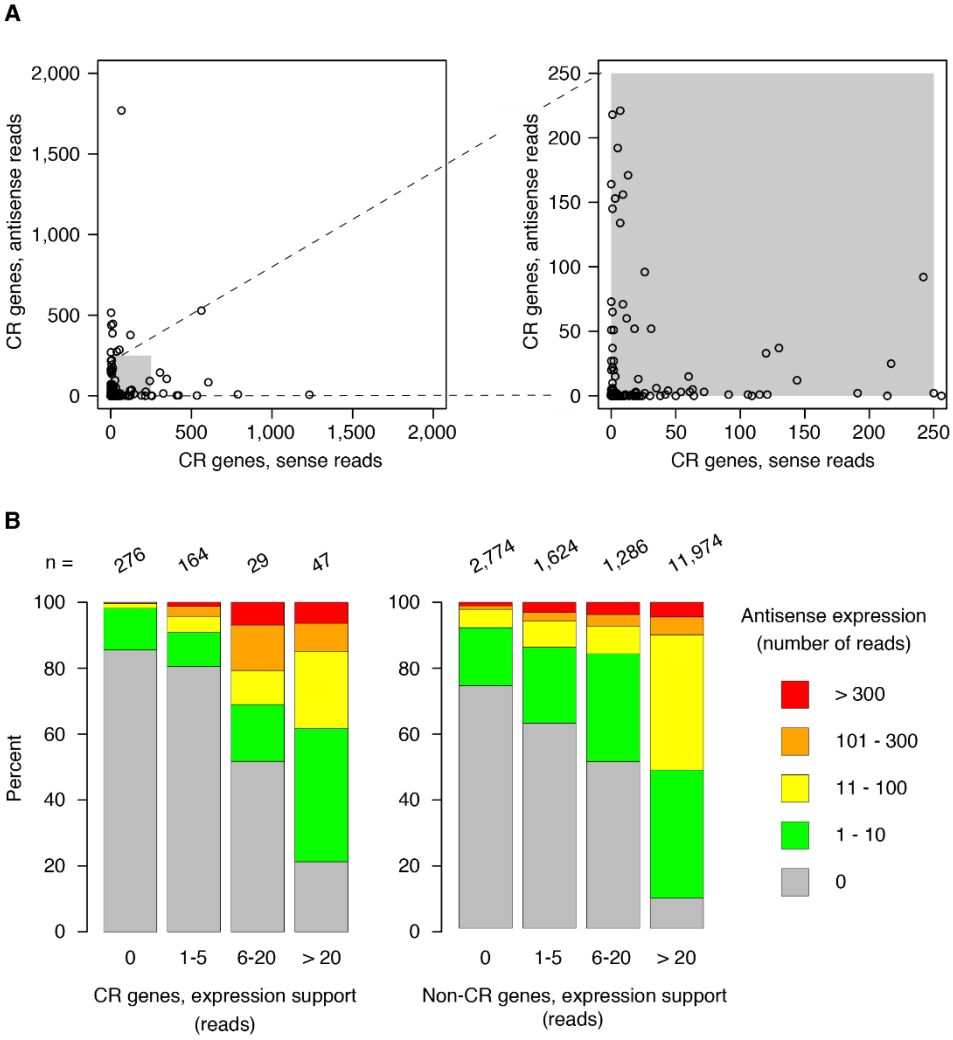


Figure A.7. The CR and antisense expression with the strand-specific data (from Richard's lab, personal communication). A, many CR genes that are expressed have little antisense expression, and in many cases where the CR is not expressed (or lowly expressed) there is moderate or substantial antisense expression. B, the very lowly expressed CRs or CRs with no expression support whatsoever have comparatively little antisense. In contrast, those with higher expression, on average, have more antisense even though a given CR may still have very low expression relative to antisense expression. The pattern for non-CR genes is similar.

A.1.5. Chemoreceptor-encoding genes in London – EtoxR – Montpellier strains

Table A.4. Chemoreceptor-encoding genes differ remarkably in among the London – EtoxR – Montpellier strains

N	GR_ID	gene_ID	str	status in London strain		Change London > Montpellier		Change London > EtoxR	
1	TuGR417	tetur02g15247	+	pseudo	1 stopcodon	intact	TAA > TTA 2/2	Unkn.	No read
2	TuGR421	tetur03g09000	+	pseudo	1 stopcodon	intact	TAA > TAT 23/23	intact	TAA > TAT 19/19 reads
3	TuGR422	tetur04g01330	+	pseudo	1 stopcodon	intact	TAG > TGG 13/13	intact	TAG > TGG 5/5 reads
4	TuGR424	tetur04g04040	-	pseudo	1 stopcodon	intact	TAG > TGG 7/7	pseudo	complex
5	TuGR435	tetur07g05010	+	pseudo	1 stopcodon	intact	TAA > GAA 28/28	heter	TAA 13/14, GAA1/14
6	TuGR436	tetur08g00740	-	pseudo	1 stopcodon	heter	TAA 1/29, GAA 28/29	intact	TAA > GAA 29/29 reads
7	TuGR438	tetur08g01430	+	pseudo	1 stopcodon 1 frameshift	pseudo	TAG 11/14, GAG 3/14; FS 12/12	intact	TAG > GAG 3/3 reads; insertion C (4/4 reads)
8	TuGR443	tetur08g08349	-	pseudo	1 stopcodon	intact	TAA > TCA 13/13	intact	TAA > TCA 27/27 reads
9	TuGR454	tetur10g05330	-	pseudo	1 stopcodon	intact	TAA > TTA 4/5 >TCA 1/5	intact	TAA > TTA 15/15 reads
10	TuGR457	tetur12g01700	+	pseudo	1 stopcodon	intact	TAA > TAT 22/22	intact	TAC 1/10, TAT 9/10
11	TuGR460	tetur12g04733	-	pseudo	1 frameshift	pseudo	FS 8/8	intact	insertion: AGTGAAAT 3/5, AATAGTGA 2/5
12	TuGR463	tetur12g04883	-	pseudo	1 frameshift	pseudo	FS 5/5	intact	insertion A (9/9 reads)
13	TuGR476	tetur16g04116	-	pseudo	1 stopcodon	pseudo	TAA 7/7	intact	No read
14	TuGR480	tetur20g03310	-	pseudo	1 stopcodon	heter	TGA 7/8, AGA 1/8	intact	TGA > AGA 5/5 reads
15	TuGR487	tetur30g00870	+	pseudo	1 frameshift	intact	TC > TCT 6/6	intact	insertion T (13/13 reads)
16	TuGR490	tetur32g02240	-	pseudo	1 frameshift	intact	No FS 5/5	heter?	insertion A (16/17 reads) 1/17 no
17	TuGR497	tetur34g01223	-	pseudo	1 frameshift	Unkn	no read	intact	deletion C 13/13 reads
18	TuGR541	tetur09g06830	+	pseudo	1 frameshift	Unkn	no read	intact	No FS 5/5
19	TuGR567	tetur150g00003	-	pseudo	1 frameshift	Unkn	no read	intact	No FS 1/1
20	TuGR568	tetur150g00004	-	pseudo	2 frameshift	Unkn	FS1: no read, No FS2 6/6	intact	No FS1 9/10, No FS2 20/20
21	TuGR591	tetur17g04160	+	pseudo	1 frameshift	intact	GG > AG 5/5	intact	GG > AG 11/11
1	TuGR106	tetur04g05590	-	intact		heter.	TTA > TAA 9/20 = TTA 11/20	heter	TTA > TAA 2/12 = TTA 10/12
2	TuGR259	tetur01g16784	-	intact		heter. ?	TCA 26/27, TCA > TAA 1/27	heter.	TCA > TAA 13/19 ; TCA 6/19
3	TuGR160	tetur24g02737	+	intact		heter.	CGA > TGA 4/9, CGA 5/9	heter.	CGA > TGA 14/22 ; CGA 8/22
4	TuGR326	tetur08g01080	+	intact		UnKn	no read	heter.	GAA > TAA 12/15 ; AAA 3/15
5	TuGR308	tetur13g04736	+	intact		pseudo	CAG > TAG 4/4	heter	CAG > TAG 2/6, CAG 4/6
6	TuGR592	tetur17g04170	+	intact		heter.	FS 2/4, no FS 2/4	heter	FS 3/5, no FS 2/5
7	TuGR594	tetur17g04190	+	intact		heter.	TCA > TAA 10/14 = TCA 4/14	intact	TCA 11/11
8	TuGR622	tetur24g02797	+	intact		heter.	FS 2/7, no FS 5/7	intact	No FS 15/15
9	TuGR632	tetur24g02917	+	intact		pseudo	insertion T 5/5	Unkn	no read
10	TuGR644	tetur305g00020	+	intact		pseudo	FS 5/5	intact	no FS 9/9
1	ENaC13	tetur01g01020	+	intact		heter.	AGA > TGA 2/12 = AGA 10/12; TAT > TAA 1/13 = TAT 12/13	intact	AGA 22/22, TAT 24/24
2	ENaC18	tetur01g01080	+	pseudo	1 frameshift	heter.	FS: 1/3, no FS: 2/3 (insertion T)	pseudo	FS: 11/11

A.2. Chemosensory receptors in three spider mites

A.2.1. Insect GR like chemoreceptors in *T. evansi* and *T. lintearius*

Table A.5. The GRs in *T. evansi*

N	GR_ID	gene_ID	status	L	I	change vs. intact	C	putative ortholog TuGR		
								RBH	microsynteny	blast (cut-off 1e-25)
1	tetev54g00645	TeGR217	intact	380	1		A	TuGR217	TuGR217	TuGR217
2	tetev54g00640	TeGR220	intact	386	1		A	TuGR220		TuGR220
3	tetev36g01040	TeGR223	intact	375	1		A			TuGR223
4	tetev90g00210	TeGR250	intact	388	1		A	TuGR250	TuGR250	TuGR250
5	tetev129g00065	TeGR251	intact	386	1		A	TuGR251	TuGR251	TuGR251
6	tetev154g00370	TeGR253	intact	381	2		A	TuGR253	TuGR253	TuGR253
7	tetev23g00655	TeGR254	intact	405	1		A			TuGR254
8	tetev13g00600	TeGR255	intact	398	1		A	TuGR255	TuGR255	TuGR255
9	tetev98g00690	TeGR256	intact	414	1		A			TuGR256
10	tetev25g00315	TeGR260a	intact	364	1		A			TuGR260
11	tetev84g00105	TeGR260b	intact	358	1		A			TuGR260
12	tetev89g00460	TeGR281	intact	343	1		A	TuGR281		TuGR281
13	tetev63g00600	TeGR282	intact	347	1		A			TuGR282
14	tetev63g00530	TeGR287a	intact	353	1		A	TuGR287		TuGR287
15	tetev63g00860	TeGR287b	intact	351	2		A			TuGR287
16	tetev63g00565	TeGR287c	intact	363	1		A			TuGR287
17	tetev46g00095	TeGR290	intact	350	1		A	TuGR290		TuGR290
18	tetev12g01390	TeGR291a	intact	354	1		A	TuGR291		TuGR291
19	tetev116g00105	TeGR291b	intact	356	1		A			TuGR291
20	tetev11g01255	TeGR292	intact	378	1		A	TuGR292	TuGR292	TuGR292
21	tetev52g00625	TeGR294	intact	380	1		A	TuGR294	TuGR294	TuGR294
22	tetev157g00020	TeGR295	intact	358	1		A			TuGR295
23	tetev03g00550	TeGR301	intact	359	1		A	TuGR301		TuGR301
24	tetev96g00530	TeGR302	intact	367	1		A			TuGR302
25	tetev368g00060	TeGR303	intact	423	1		A			TuGR303
26	tetev94g00040	TeGR305	intact	382	1		A	TuGR305	TuGR305	TuGR305
27	tetev146g00625	TeGR309	intact	407	1		A	TuGR309		TuGR309
28	tetev15g01601	TeGR310	intact	382	1		A	TuGR310		TuGR310
29	tetev226g00085	TeGR311	intact	379	1		A			TuGR311
30	tetev226g00082	TeGR313	intact	379	1		A			TuGR313
31	tetev226g00087	TeGR314	intact	378	1		A			TuGR314
32	tetev209g00195	TeGR326a	intact	350	1		A			TuGR326
33	tetev209g00200	TeGR326b	intact	354	1		A			TuGR326
34	tetev867g00001	TeGR326c	intact	355	1		A			TuGR326
35	tetev212g00030	TeGR331	intact	347	1		A	TuGR331		TuGR331
36	tetev212g00090	TeGR350a	intact	357	1		A			TuGR350
37	tetev97g00275	TeGR364a	intact	354	1		A	TuGR364		TuGR364
38	tetev209g00211	TeGR364b	intact	354	1		A			TuGR364
39	tetev209g00030	TeGR364c	intact	354	1		A	TuGR353		TuGR364
40	tetev640g00010	TeGR364d	intact	354	1		A			TuGR364
41	tetev54g00460	TeGR366	intact	362	1		A	TuGR366	TuGR366	TuGR366
42	tetev116g00480	TeGR372	intact	366	1		A	TuGR372		TuGR372
43	tetev165g00470	TeGR376	intact	375	1		A	TuGR376		TuGR376
44	tetev71g00405	TeGR379a	intact	386	1		A	TuGR379		TuGR379
45	tetev13g01240	TeGR380	intact	374	1		A	TuGR380	TuGR380	TuGR380
46	tetev89g00480	TeGR514	intact	336	1		A	TuGR514	TuGR514	TuGR514
1	tetev01g01620	TeGR122	intact	442	4		B	TuGR122	TuGR122	TuGR122
2	tetev01g02220	TeGR12	intact	394	2		B	TuGR12		TuGR12
3	tetev03g00540	TeGR29	intact	383	2		B	TuGR29	TuGR29	TuGR29
4	tetev06g00500	TeGR116	intact	438	3		B	TuGR116	TuGR116	TuGR116
5	tetev06g02020	TeGR182b	intact	416	3		B			TuGR182
6	tetev08g01240	TeGR101b	intact	430	3		B			TuGR101
7	tetev08g01250	TeGR114b	intact	431	3		B			TuGR114
8	tetev08g01410	TeGR114a	intact	432	3		B	TuGR114		TuGR114
9	tetev104g00130	TeGR47	intact	418	3		B	TuGR47	TuGR47	TuGR47
10	tetev105g00590	TeGR109	intact	439	3		B	TuGR109	TuGR109	TuGR109
11	tetev108g00360	TeGR35	intact	417	3		B	TuGR35	TuGR35	TuGR35
12	tetev10g00240	TeGR20	intact	411	2		B	TuGR20	TuGR20	TuGR20
13	tetev113g00410	TeGR170	intact	458	3		B	TuGR170		TuGR170
14	tetev113g00420	TeGR123	intact	496	3		B	TuGR123	TuGR123	TuGR123
15	tetev116g00490	TeGR19	intact	400	2		B	TuGR19		TuGR19
16	tetev11g01020	TeGR48	intact	429	3		B	TuGR48	TuGR48	TuGR48
17	tetev122g00400	TeGR99	intact	430	3		B	TuGR99		TuGR99
18	tetev124g00420	TeGR6b	intact	405	2		B			TuGR6
19	tetev125g00560	TeGR27	intact	406	2		B	TuGR27	TuGR27	TuGR27
20	tetev13g00460	TeGR659b	intact	442	3		B			TuGR659
21	tetev13g00470	TeGR659c	intact	439	3		B			TuGR659

22	tetev13g00570	TeGR659d	intact	442	3	B			TuGR659
23	tetev13g00580	TeGR659a	intact	442	3	B	TuGR659		TuGR659
24	tetev140g00090	TeGR62	intact	417	3	B			TuGR62
25	tetev140g00095	TeGR536a	intact	418	3	B	TuGR536	TuGR536	TuGR536
26	tetev140g00100	TeGR442	intact	436	3	B	TuGR442	TuGR442	TuGR442
27	tetev143g00020	TeGR22	intact	388	2	B	TuGR22	TuGR22	TuGR22
28	tetev144g00100	TeGR82	intact	441	3	B	TuGR82	TuGR82	TuGR82
29	tetev144g00110	TeGR79	intact	426	3	B	TuGR79	TuGR79	TuGR79
30	tetev151g00150	TeGR28	intact	410	2	B	TuGR28	TuGR28	TuGR28
31	tetev15g00670	TeGR15	intact	393	2	B			TuGR15
32	tetev161g00080	TeGR51	intact	448	3	B	TuGR51	TuGR51	TuGR51
33	tetev170g00190	TeGR50	intact	431	3	B	TuGR50	TuGR50	TuGR50
34	tetev171g00020	TeGR98	intact	433	3	B	TuGR98		TuGR98
35	tetev171g00025	TeGR594	intact	436	3	B	TuGR594	TuGR594	TuGR594
36	tetev171g00040	TeGR191	intact	418	3	B	TuGR191	TuGR191	TuGR191
37	tetev171g00190	TeGR200	intact	425	3	B	TuGR200		TuGR200
38	tetev179g00310	TeGR182c	intact	421	3	B			TuGR182
39	tetev179g00320	TeGR182a	intact	420	3	B	TuGR182	TuGR182	TuGR182
40	tetev184g00350	TeGR211	intact	445	3	B	TuGR211	TuGR211	TuGR211
41	tetev187g00030	TeGR127	intact	440	3	B			TuGR127
42	tetev187g00040	TeGR126	intact	442	3	B	TuGR126		TuGR126
43	tetev187g00050	TeGR544	intact	433	3	B	TuGR544		TuGR544
44	tetev187g00055	TeGR130	intact	433	3	B			TuGR130
45	tetev188g00190	TeGR464	intact	429	3	B	TuGR464		TuGR464
46	tetev188g00200	TeGR101a	intact	422	3	B	TuGR101	TuGR101	TuGR101
47	tetev18g00140	TeGR188a	intact	451	3	B	TuGR188		TuGR188
48	tetev18g00700	TeGR74b	intact	422	3	B			TuGR74
49	tetev191g00030	TeGR171	intact	438	3	B			TuGR171
50	tetev191g00110	TeGR40	intact	420	3	B	TuGR40	TuGR40	TuGR40
51	tetev199g00150	TeGR41	intact	480	3	B	TuGR41		TuGR41
52	tetev21g00330	TeGR49	intact	418	3	B			TuGR49
53	tetev245g00100	TeGR586	intact	425	3	B			TuGR586
54	tetev245g00110	TeGR165	intact	432	3	B	TuGR165		TuGR165
55	tetev24g00880	TeGR143	intact	428	3	B	TuGR143	TuGR143	TuGR143
56	tetev256g00100	TeGR112	intact	412	3	B	TuGR112	TuGR112	TuGR112
57	tetev29g01010	TeGR177b	intact	431	3	B			TuGR177
58	tetev303g00095	TeGR487	intact	412	2	B	TuGR487	TuGR487	TuGR487
59	tetev303g00100	TeGR21	intact	389	2	B			TuGR21
60	tetev303g00105	TeGR26	intact	379	2	B	TuGR26	TuGR26	TuGR26
61	tetev307g00050	TeGR177a	intact	431	3	B	TuGR177		TuGR177
62	tetev307g00080	TeGR175	intact	438	3	B			TuGR175
63	tetev307g00110	TeGR172	intact	438	3	B	TuGR172		TuGR172
64	tetev324g00080	TeGR536c	intact	442	3	B			TuGR536
65	tetev34g01201	TeGR529	intact	441	3	B			TuGR529
66	tetev350g00030	TeGR75b	intact	426	3	B			TuGR75
67	tetev350g00050	TeGR75a	intact	441	3	B			TuGR75
68	tetev357g00050	TeGR535	intact	407	2	B			TuGR535
69	tetev369g00081	TeGR11	intact	394	2	B			TuGR11
70	tetev370g00050	TeGR74a	intact	424	3	B			TuGR74
71	tetev375g00040	TeGR536b	intact	414	3	B			TuGR536
72	tetev38g00020	TeGR18	intact	383	2	B			TuGR18
73	tetev39g00300	TeGR169	intact	430	3	B			TuGR169
74	tetev40g00390	TeGR188b	intact	433	3	B			TuGR188
75	tetev42g00150	TeGR166	intact	424	3	B			TuGR166
76	tetev42g00255	TeGR13	intact	405	2	B			TuGR13
77	tetev46g00015	TeGR70a	intact	454	3	B			TuGR70
78	tetev46g00030	TeGR70b	intact	431	3	B			TuGR70
79	tetev46g00040	TeGR72b	intact	447	3	B			TuGR72
80	tetev46g00050	TeGR72a	intact	440	4	B			TuGR72
81	tetev46g00060	TeGR72c	intact	439	3	B			TuGR72
82	tetev46g00090	TeGR72d	intact	430	3	B			TuGR72
83	tetev48g00290	TeGR181	intact	417	3	B	TuGR181		TuGR181
84	tetev48g01060	TeGR182f	intact	422	3	B			TuGR182
85	tetev497g00010	TeGR113	intact	432	3	B	TuGR113		TuGR113
86	tetev504g00010	TeGR182d	intact	425	3	B			TuGR182
87	tetev504g00040	TeGR182e	intact	372	3	B			TuGR182
88	tetev50g00800	TeGR195b	intact	438	3	B			TuGR195
89	tetev50g00810	TeGR195a	intact	441	3	B	TuGR195		TuGR195
90	tetev532g00010	TeGR6c	intact	407	2	B			TuGR6
91	tetev57g00040	TeGR39	intact	404	3	B	TuGR39		TuGR39
92	tetev58g00220	TeGR30a	intact	405	2	B	TuGR30	TuGR30	TuGR30
93	tetev58g00580	TeGR34	intact	420	3	B	TuGR34		TuGR34
94	tetev61g00760	TeGR84	intact	434	3	B	TuGR84	TuGR84	TuGR84
95	tetev61g00770	TeGR90a	intact	433	3	B			TuGR90
96	tetev61g00780	TeGR90b	intact	433	3	B			TuGR90
97	tetev67g00100	TeGR7	intact	416	2	B	TuGR7	TuGR7	TuGR7
98	tetev69g00730	TeGR111	intact	430	3	B	TuGR111	TuGR111	TuGR111
99	tetev73g00550	TeGR108a	intact	415	3	B			TuGR108
100	tetev75g00240	TeGR10	intact	430	2	B	TuGR10		TuGR10
101	tetev75g00641	TeGR46	intact	460	3	B			TuGR46

102	tetev769g00020	TeGR6a	intact	428	2		B			TuGR6
103	tetev85g00240	TeGR124	intact	404	3		B			TuGR124
104	tetev86g00400	TeGR43	intact	454	3		B	TuGR43	TuGR43	TuGR43
105	tetev86g00415	TeGR45	intact	453	3		B	TuGR45	TuGR45	TuGR45
106	tetev90g00380	TeGR129a	intact	436	3		B	TuGR129	TuGR129	TuGR129
107	tetev95g00400	TeGR32	intact	411	2		B	TuGR32	TuGR32	TuGR32
1	tetev08g01380	TeGR1	intact	410	0		C	TuGR1		TuGR1
2	tetev183g00090	TeGR2	intact	404	4		C	TuGR2	TuGR2	TuGR2
3	tetev66g00500	TeGR210	intact	404	2		C	TuGR210	TuGR210	TuGR210
4	tetev88g00910	TeGR212	intact	411	2		C	TuGR212	TuGR212	TuGR212
5	tetev18g00660	TeGR214	intact	458	3		C	TuGR214		TuGR214
6	tetev377g00080	TeGR215	intact	438	3		C	TuGR215		TuGR215
7	tetev56g02470	TeGR216	intact	430	3		C	TuGR216		TuGR216
8	tetev14g01020	TeGR3	intact	478	3		C	TuGR3	TuGR3	TuGR3
9	tetev100g00420	TeGR33	intact	439	2		C	TuGR33	TuGR33	TuGR33
10	tetev79g00250	TeGR4	intact	413	4		C	TuGR4	TuGR4	TuGR4
11	tetev175g00110	TeGR5	intact	380	0		C	TuGR5	TuGR5	TuGR5
12	tetev175g00120	TeGR510	intact	400	0		C	TuGR510	TuGR510	TuGR510
13	tetev49g00320	TeGR511	intact	362	0		C	TuGR511		TuGR511
14	tetev44g00750	TeGR528	intact	489	1		C	TuGR528		TuGR528
15	tetev175g00300	TeGR543	intact	402	0		C	TuGR543		TuGR543
1	tetev1045g00001	TeGR343	partial	270	0	N,C-ter missing	A			TuGR343
2	tetev89g00010	TeGR323	partial	151	1	N-ter missing	A			TuGR323
1	tetev96g00560	TeGR133a	partial	405	3	int.seq. missing	B	TuGR133		TuGR133
2	tetev24g00865	TeGR17a	partial	376	2	N gap	B			TuGR17
3	tetev799g00001	TeGR110	partial	360	0	C-ter missing	B			TuGR110
4	tetev1600g00001	TeGR93a	partial	352	0	int.seq, C-ter missing	B			TuGR93
5	tetev24g00870	TeGR17b	partial	319	1	int.seq. missing	B			TuGR17
6	tetev318g00130	TeGR93c	partial	306	3	N-ter missing	B			TuGR93
7	tetev200g00120	TeGR6g	partial	305	0	C-ter missing	B	TuGR6		TuGR6
8	tetev1961g00001	TeGR72e	partial	300	1	N,C-ter missing	B			TuGR72
9	tetev1845g00001	TeGR6d	partial	257	0	N,C-ter missing	B			TuGR6
10	tetev173g00250	TeGR94	partial	129	0	int.seq. missing	B			TuGR94
11	tetev1975g00001	TeGR133c	partial	128	2	N-ter missing	B			TuGR133
12	tetev372g00110	TeGR133b	partial	112	0	C-ter missing	B			TuGR133
13	tetev200g00010	TeGR708	partial	72	1	relics	B			
1	tetev71g00520	TeGR350b	pseudo	345	1	1 frameshift	A			TuGR350
2	tetev89g00005	TeGR342	pseudo	313	0	int.seq. deletion	A			TuGR342
3	tetev262g00200	TeGR397	pse/partial	293	2	N-ter missing, int.seq.deletion	A	TuGR397		TuGR397
4	tetev1010g00001	TeGR370a	pse/partial	290	1	N-ter missing, int.seq.deletion	A			TuGR370
5	tetev212g00200	TeGR370b	pseudo	279	1	int.seq. deletion	A	TuGR370		TuGR370
6	tetev423g00090	TeGR541	pse/partial	191	2	N,C-ter missing	A			TuGR541
7	tetev28g01360	TeGR695	pseudo	190	0	many changes	A			
8	tetev446g00080	TeGR379b	pseudo	175	1	relics	A			TuGR379
9	tetev25g01700	TeGR696	pseudo	170	0	relics	A			
10	tetev640g00030	TeGR698	pseudo	155	0	relics	A			
11	tetev85g00730	TeGR699	pseudo	149	0	relics	A			
12	tetev17g00980	TeGR700	pseudo	130	0	relics	A			
13	tetev209g00221	TeGR701	pseudo	127	0	relics	A			
14	tetev67g00690	TeGR703	pseudo	109	0	relics	A			
15	tetev89g00791	TeGR704	pseudo	99	0	relics	A			
16	tetev2835g00001	TeGR706	pseudo	94	1	relics	A			
1	tetev245g00090	TeGR586a	pseudo	450	3	3 frameshifts	B			TuGR586
2	tetev46g00080	TeGR70c	pseudo	430	3	5 frameshifts	B			TuGR70
3	tetev90g00590	TeGR129b	pse/partial	419	3	1 FS & gap	B			TuGR129
4	tetev73g00580	TeGR108b	pseudo	417	3	1 frameshift	B			TuGR108
5	tetev46g00070	TeGR72f	pseudo	417	3	1 frameshift	B			TuGR72
6	tetev137g00185	TeGR197	pseudo	399	3	many FS & PS	B			TuGR197
7	tetev13g00465	TeGR152	pseudo	398	3	many FS & PS	B			TuGR152
8	tetev73g00510	TeGR108c	pseudo	396	1	int.seq. deletion	B			TuGR108
9	tetev147g00330	TeGR30b	pse/partial	324	0	2 FS, 2PS & C-ter missing	B			TuGR30
10	tetev68g01380	TeGR75c	pseudo	319	0	many changes, int. seq.deletion	B			TuGR75
11	tetev777g00001	TeGR6h	pseudo	293	1	int.seq. deletion	B			TuGR6
12	tetev140g00400	TeGR536d	pseudo	270	0	N-ter deletion	B			TuGR536
13	tetev117g00490	TeGR178	pseudo	268	1	int. seq. deletion	B			TuGR178
14	tetev200g00040	TeGR6e	pseudo	266	1	many changes	B			TuGR6
15	tetev189g00311	TeGR128	pseudo	264	0	many changes	B			TuGR128
16	tetev200g00080	TeGR6f	pseudo	263	0	many changes	B			TuGR6

17	tetev61g01150	TeGR490b	pseudo	207	0	many changes	B			TuGR490
18	tetev143g00400	TeGR23	pse/partial	205	2	N-ter msising, int.seq. deletion	B			TuGR23
19	tetev40g00970	TeGR188c	pseudo	176	0	relics	B			TuGR188
20	tetev173g00260	TeGR93b	pseudo	173	0	N-ter relics	B			TuGR93
21	tetev191g00140	TeGR38	pseudo	152	0	relics	B			TuGR38
22	tetev2976g00001	TeGR490a	pseudo	121	0	relics	B			TuGR490
23	tetev48g01070	TeGR182g	pseudo	110	0	N-ter relics	B			TuGR182
24	tetev453g00051	TeGR705	pseudo	95	0	N-ter relics	B			
25	tetev73g00780	TeGR707	pseudo	93	0	relics	B			
26	tetev187g00320	TeGR709	pseudo	70	0	relics	B			
1	tetev07g00290	TeGR213	pse_partial	396	2	C-ter missing, 1FS. tetev1751g00002 is the central part of this gene.	C			TuGR213

Table A.6. The GRs in *T. lintearius*. Pseudogene TlGR33 (Class TlGR-C) can be an intact gene because of sequencing error.

N	GR_ID	gene_ID	status	L	I	change vs. intact	C	putative ortholog TuGR		
								RBH	microsynteny	blast (cut-off 1e-25)
1	tetli64g00760	TlGR251	intact	386	1		A	TuGR251	TuGR251	TuGR251
2	tetli33g00220	TlGR253	intact	376	2		A	TuGR253	TuGR253	TuGR253
3	tetli51g00670	TlGR255	intact	392	1		A	TuGR255	TuGR255	TuGR255
4	tetli104g00360	TlGR256	intact	416	1		A	TuGR256	TuGR256	TuGR256
5	tetli36g00295	TlGR258	intact	362	1		A	TuGR258		TuGR258
6	tetli40g00985	TlGR290a	intact	350	1		A	TuGR290		TuGR290
7	tetli26g02230	TlGR298	intact	354	1		A	TuGR298		TuGR298
8	tetli225g00080	TlGR305	intact	379	1		A	TuGR305	TuGR305	TuGR305
9	tetli03g01150	TlGR307a	intact	390	1		A	TuGR307		TuGR307
10	tetli06g01390	TlGR308	intact	413	1		A	TuGR308	TuGR308	TuGR308
11	tetli232g00180	TlGR309	intact	404	1		A	TuGR309	TuGR309	TuGR309
12	tetli293g00250	TlGR310	intact	377	1		A	TuGR310		TuGR310
13	tetli123g00280	TlGR311	intact	373	1		A	TuGR311	TuGR311	TuGR311
14	tetli61g00768	TlGR326	intact	356	1		A	TuGR326		TuGR326
15	tetli61g00753	TlGR327	intact	356	1		A	TuGR327		TuGR327
16	tetli177g00040	TlGR350	intact	356	1		A	TuGR350	TuGR350	TuGR350
17	tetli160g00260	TlGR357	intact	364	1		A	TuGR357		TuGR357
18	tetli61g00755	TlGR363a	intact	352	1		A			TuGR363
19	tetli240g00190	TlGR366	intact	362	1		A	TuGR366		TuGR366
20	tetli39g00890	TlGR371	intact	371	1		A	TuGR371	TuGR371	TuGR371
21	tetli260g00100	TlGR376	intact	375	1		A	TuGR376		TuGR376
22	tetli151g00355	TlGR379	intact	373	1		A	TuGR379		TuGR379
23	tetli51g00465	TlGR380	intact	374	1		A	TuGR380		TuGR380
24	tetli66g00700	TlGR381	intact	363	1		A	TuGR381	TuGR381	TuGR381
25	tetli01g01325	TlGR386	intact	381	1		A	TuGR386	TuGR386	TuGR386
26	tetli26g01280	TlGR387	intact	375	1		A	TuGR387	TuGR387	TuGR387
27	tetli55g00305	TlGR397	intact	369	1		A	TuGR397	TuGR397	TuGR397
28	tetli123g00275	TlGR471	intact	379	1		A	TuGR471	TuGR471	TuGR471
29	tetli175g00381	TlGR642	intact	373	1		A	TuGR642		TuGR642
1	tetli03g01610	TlGR10	intact	407	2		B	TuGR10	TuGR10	TuGR10
2	tetli294g00140	TlGR101	intact	423	3		B	TuGR101		TuGR101
3	tetli78g00500	TlGR108	intact	427	3		B	TuGR108	TuGR108	TuGR108
4	tetli106g00001	TlGR11	intact	395	2		B	TuGR11	TuGR11	TuGR11
5	tetli170g00160	TlGR111	intact	431	3		B	TuGR111	TuGR111	TuGR111
6	tetli193g00190	TlGR112	intact	403	3		B	TuGR112		TuGR112
7	tetli354g00050	TlGR114	intact	433	3		B	TuGR114	TuGR114	TuGR114
8	tetli25g00090	TlGR116	intact	437	3		B	TuGR116	TuGR116	TuGR116
9	tetli29g01460	TlGR12	intact	396	2		B	TuGR12		TuGR12
10	tetli29g02120	TlGR122	intact	445	3		B	TuGR122	TuGR122	TuGR122
11	tetli65g00380	TlGR123	intact	484	3		B	TuGR123	TuGR123	TuGR123
12	tetli92g00500	TlGR124	intact	404	3		B	TuGR124	TuGR124	TuGR124
13	tetli54g00710	TlGR129	intact	432	3		B	TuGR129	TuGR129	TuGR129
14	tetli113g00320	TlGR13	intact	401	2		B	TuGR13	TuGR13	TuGR13
15	tetli12g00370	TlGR130	intact	433	3		B	TuGR130	TuGR130	TuGR130
16	tetli238g00090	TlGR133	intact	441	3		B	TuGR133		TuGR133
17	tetli01g02290	TlGR141	intact	427	3		B	TuGR141	TuGR141	TuGR141
18	tetli01g02335	TlGR142	intact	454	3		B	TuGR142	TuGR142	TuGR142
19	tetli01g02280	TlGR143	intact	436	3		B	TuGR143	TuGR143	TuGR143

20	tetli31g01035	TlGR15	intact	394	2		B	TuGR15	TuGR15	TuGR15
21	tetli109g00003	TlGR154	intact	441	3		B			TuGR154
22	tetli31g01020	TlGR16	intact	391	2		B	TuGR16	TuGR16	TuGR16
23	tetli03g01981	TlGR166	intact	439	3		B	TuGR166	TuGR166	TuGR166
24	tetli295g00130	TlGR168	intact	431	3		B	TuGR168		TuGR168
25	tetli305g00010	TlGR169	intact	431	3		B	TuGR169		TuGR169
26	tetli01g02300	TlGR17	intact	397	2		B	TuGR17	TuGR17	TuGR17
27	tetli130g00040	TlGR170	intact	453	3		B	TuGR170	TuGR170	TuGR170
28	tetli76g00480	TlGR171	intact	434	3		B			TuGR171
29	tetli111g00260	TlGR172	intact	464	3		B	TuGR172	TuGR172	TuGR172
30	tetli335g00070	TlGR177	intact	437	3		B	TuGR177		TuGR177
31	tetli190g00010	TlGR182	intact	420	3		B	TuGR182	TuGR182	TuGR182
32	tetli60g00450	TlGR185	intact	428	3		B	TuGR185		TuGR185
33	tetli70g00410	TlGR188	intact	427	3		B	TuGR188	TuGR188	TuGR188
34	tetli172g00010	TlGR18a	intact	386	2		B	TuGR18		TuGR18
35	tetli01g01210	TlGR192	intact	418	3		B	TuGR192	TuGR192	TuGR192
36	tetli194g00080	TlGR197	intact	426	3		B	TuGR197	TuGR197	TuGR197
37	tetli45g00250	TlGR199	intact	430	3		B	TuGR199	TuGR199	TuGR199
38	tetli161g00261	TlGR20	intact	433	2		B	TuGR20		TuGR20
39	tetli91g00510	TlGR201	intact	427	3		B	TuGR201		TuGR201
40	tetli90g00280	TlGR22	intact	381	2		B	TuGR22	TuGR22	TuGR22
41	tetli283g00195	TlGR23a	intact	379	2		B	TuGR23		TuGR23
42	tetli283g00200	TlGR26	intact	387	2		B	TuGR26		TuGR26
43	tetli138g00400	TlGR28	intact	415	2		B	TuGR28	TuGR28	TuGR28
44	tetli14g00750	TlGR29	intact	403	2		B	TuGR29	TuGR29	TuGR29
45	tetli104g00430	TlGR30	intact	411	2		B	TuGR30	TuGR30	TuGR30
46	tetli36g00490	TlGR34	intact	420	3		B	TuGR34		TuGR34
47	tetli56g00320	TlGR35	intact	417	3		B	TuGR35	TuGR35	TuGR35
48	tetli35g00140	TlGR39	intact	434	3		B	TuGR39		TuGR39
49	tetli111g00130	TlGR40	intact	428	3		B	TuGR40		TuGR40
50	tetli21g01360	TlGR41	intact	481	3		B	TuGR41		TuGR41
51	tetli06g01260	TlGR42	intact	453	3		B	TuGR42	TuGR42	TuGR42
52	tetli111g00280	TlGR421	intact	438	3		B	TuGR421	TuGR421	TuGR421
53	tetli194g00050	TlGR427	intact	442	3		B	TuGR427		TuGR427
54	tetli151g00400	TlGR449	intact	394	2		B	TuGR449		TuGR449
55	tetli06g00920	TlGR45	intact	452	3		B	TuGR45		TuGR45
56	tetli65g01031	TlGR46	intact	455	3		B			TuGR46
57	tetli294g00151	TlGR464	intact	429	3		B	TuGR464		TuGR464
58	tetli55g01640	TlGR47	intact	418	3		B	TuGR47	TuGR47	TuGR47
59	tetli45g00350	TlGR48	intact	448	3		B	TuGR48	TuGR48	TuGR48
60	tetli90g00220	TlGR487	intact	386	2		B	TuGR487	TuGR487	TuGR487
61	tetli133g00200	TlGR49	intact	410	3		B	TuGR49	TuGR49	TuGR49
62	tetli91g00320	TlGR50	intact	429	3		B	TuGR50	TuGR50	TuGR50
63	tetli72g00200	TlGR51	intact	447	3		B	TuGR51	TuGR51	TuGR51
64	tetli61g00758	TlGR52	intact	426	3		B	TuGR52		TuGR52
65	tetli313g00081	TlGR53	intact	427	3		B	TuGR53		TuGR53
66	tetli61g00757	TlGR536	intact	418	3		B	TuGR536		TuGR536
67	tetli226g00065	TlGR575	intact	433	3		B	TuGR575		TuGR575
68	tetli01g01570	TlGR576	intact	409	2		B	TuGR576	TuGR576	TuGR576
69	tetli309g00020	TlGR580	intact	434	3		B	TuGR580		TuGR580
70	tetli11g00505	TlGR6	intact	401	2		B	TuGR6		TuGR6
71	tetli309g00041	TlGR628	intact	435	3		B	TuGR628		TuGR628
72	tetli143g00500	TlGR655	intact	413	3		B	TuGR655		TuGR655
73	tetli422g00012	TlGR657	intact	430	3		B	TuGR657		TuGR657
74	tetli422g00011	TlGR659a	intact	442	3		B	TuGR659		TuGR659
75	tetli422g00010	TlGR659b	intact	438	3		B			TuGR659
76	tetli164g00460	TlGR75	intact	427	3		B	TuGR75		TuGR75
77	tetli95g00400	TlGR77	intact	433	3		B	TuGR77	TuGR77	TuGR77
78	tetli47g00730	TlGR7a	intact	450	2		B	TuGR7	TuGR7	TuGR7
79	tetli84g00350	TlGR90	intact	436	3		B	TuGR90	TuGR90	TuGR90
80	tetli91g00020	TlGR93	intact	433	3		B	TuGR93	TuGR93	TuGR93
81	tetli91g00050	TlGR94a	intact	428	3		B		TuGR432	TuGR94
82	tetli226g00068	TlGR98	intact	431	3		B	TuGR98		TuGR98
83	tetli91g01030	TlGR99	intact	429	3		B	TuGR99		TuGR99
84	tetli422g00013	TlGR659d	intact	442	3	partial gene (5') to be assembled with tetli109g00001 (3') to end up with a complete TlGR	B			TuGR659

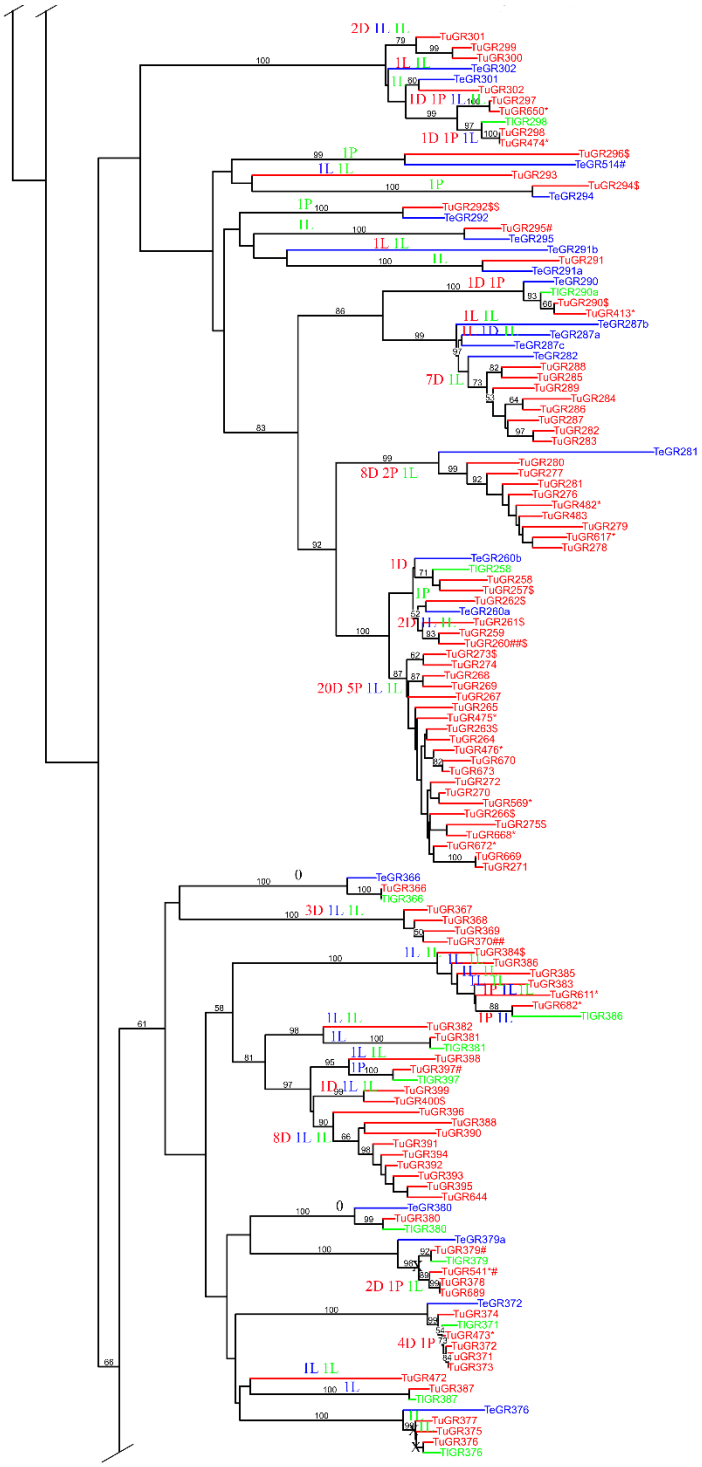
1	tetli22g01240	TlGR1	intact	410	0		C	TuGR1		TuGR1
2	tetli41g00100	TlGR2	intact	404	4		C	TuGR2	TuGR2	TuGR2
3	tetli57g00140	TlGR210	intact	404	2		C	TuGR210	TuGR210	TuGR210
4	tetli116g00730	TlGR212	intact	411	2		C	TuGR212	TuGR212	TuGR212
5	tetli01g00170	TlGR213	intact	454	3		C	TuGR213	TuGR213	TuGR213
6	tetli142g00170	TlGR214	intact	457	3		C	TuGR214		TuGR214

7	tetli117g00390	TlGR215	intact	438	3		C	TuGR215	TuGR215	TuGR215
8	tetli12g01470	TlGR216	intact	430	3		C	TuGR216		TuGR216
9	tetli88g00280	TlGR3	intact	478	3		C	TuGR3	TuGR3	TuGR3
10	tetli92g01140	TlGR4	intact	413	4		C	TuGR4		TuGR4
11	tetli221g00100	TlGR510	intact	398	0		C	TuGR510	TuGR510	TuGR510
12	tetli143g00170	TlGR511	intact	362	0		C	TuGR511	TuGR511	TuGR511
13	tetli07g00450	TlGR528	intact	490	1		C	TuGR528		TuGR528
14	tetli221g00075	TlGR543	intact	402	0		C	TuGR543	TuGR543	TuGR543
1	tetli319g00095	TlGR219b	partial	221	1	N-ter missing	A			TuGR219
2	tetli240g00335	TlGR460a	partial	342	1	N-ter missing	A			TuGR460
1	tetli01g01600	TlGR31	partial	393	1	int.seq. missing	B	TuGR31	TuGR31	TuGR31
2	tetli01g01580	TlGR32	partial	357	2	int.seq. missing	B			TuGR32
3	tetli513g00011	TlGR38	partial	342	0	N,C-ter missing	B			TuGR38
4	tetli61g00095	TlGR149	partial	282	0	int.seq. missing	B			TuGR149
5	tetli109g00002	TlGR156a	partial	231	3	N-ter missing	B			TuGR156
6	tetli707g00011	TlGR23b	partial	198	2	N-ter missing	B			TuGR23
7	tetli1356g00002	TlGR138b	partial	108	3	N-ter missing	B			TuGR138
8	tetli61g00090	TlGR725	partial	101	0	C-ter missing	B			
9	tetli445g00001	TlGR732	partial	82	0	C-ter missing	B			
10	tetli61g00751	TlGR738	partial	77	0	C-ter missing	B			
1	tetli221g00110	TlGR5	partial	368	0	N-ter missing	C	TuGR5	TuGR5	TuGR5
1	tetli19g01111	TlGR254	pseudo	405	1	4 PS	A			TuGR254
2	tetli239g00201	TlGR434	pseudo	383	1	1 FS	A			TuGR434
3	tetli287g00115	TlGR217	pseudo	379	1	3 FS & 4 PS	A			TuGR217
4	tetli80g00080	TlGR303	pseudo	378	0	C-ter deletion	A			TuGR303
5	tetli239g00202	TlGR456a	pseudo	372	1	many changes	A			TuGR456
6	tetli89g00590	TlGR294	pseudo	370	1	4FS & 3PS	A	TuGR294		TuGR294
7	tetli108g00100	TlGR252a	pseudo	364	1	1FS, int.seq. deletion	A	TuGR252	TuGR252	TuGR252
8	tetli97g00385	TlGR221	pseudo	360	0	many changes	A			TuGR221
9	tetli26g02460	TlGR257	pseudo	356	0	many changes	A			TuGR257
10	tetli160g00205	TlGR356	pseudo	353	0	2FS & 1 PS	A			TuGR356
11	tetli36g00850	TlGR260	pseudo	352	1	5 PS	A			TuGR260
12	tetli119g00065	TlGR457	pseudo	351	0	1FS, int.seq. deletion	A			TuGR457
13	tetli01g01500	TlGR384	pse/partial	350	0	int.seq. missing	A			TuGR384
14	tetli175g00452	TlGR245	pseudo	348	0	many changes	A			TuGR245
15	tetli36g00860	TlGR261	pseudo	337	1	many changes	A	TuGR261		TuGR261
16	tetli11g01640	TlGR266	pseudo	333	0	many changes	A	TuGR266		TuGR266
17	tetli319g00080	TlGR219a	pseudo	325	0	C-ter deletion	A	TuGR219		TuGR219
18	tetli61g00756	TlGR363b	pseudo	319	0	2FS, 1PS & N,int.seq. deletion	A			TuGR363
19	tetli55g01811	TlGR400	pseudo	318	0	many changes	A			TuGR400
20	tetli213g00190	TlGR710	pseudo	317	0	many changes	A			
21	tetli175g00382	TlGR236	pseudo	309	0	1FS, 3PS & C-ter deletion	A			TuGR236
22	tetli175g00462	TlGR232b	pseudo	308	0	many changes	A			TuGR232
23	tetli226g00141	TlGR365	pseudo	307	1	int.seq. deletion	A			TuGR365
24	tetli92g01260	TlGR296	pseudo	302	1	many changes	A	TuGR296	TuGR296	TuGR296
25	tetli240g00410	TlGR248	pseudo	301	0	many changes	A			TuGR248
26	tetli175g00442	TlGR230	pseudo	294	0	many changes	A			TuGR230
27	tetli61g00788	TlGR325	pseudo	281	0	many changes	A	TuGR325		TuGR325
28	tetli03g01982	TlGR307b	pseudo	277	1	1 FS, 1PS & N,C-ter, int.seq deletion	A			TuGR307
29	tetli43g00490	TlGR351	pseudo	271	0	many changes	A	TuGR351	TuGR351	TuGR351
30	tetli54g00741	TlGR250	pse/partial	266	0	2FS & C-ter missing	A			TuGR250
31	tetli61g00828	TlGR329	pseudo	250	1	many changes	A			TuGR329
32	tetli175g00392	TlGR232a	pseudo	239	1	N-ter, int.seq. deletion	A			TuGR232
33	tetli226g00151	TlGR352	pseudo	233	1	many changes	A			TuGR352
34	tetli61g00754	TlGR363c	pseudo	231	0	many changes	A			TuGR363
35	tetli239g00170	TlGR460b	pseudo	221	0	N,C-ter deletion	A			TuGR460
36	tetli175g00412	TlGR233	pseudo	212	0	many changes	A			TuGR233
37	tetli239g00212	TlGR247	pseudo	212	1	many changes	A			TuGR247
38	tetli239g00222	TlGR456b	pseudo	207	0	many changes	A			TuGR456
39	tetli47g00810	TlGR229	pseudo	189	0	many changes	A			TuGR229
40	tetli108g00510	TlGR252b	pseudo	185	0	many changes	A			TuGR252
41	tetli119g00250	TlGR223	pseudo	172	0	N-ter relics	A			TuGR223
42	tetli61g00808	TlGR324	pseudo	170	0	N-ter relics	A			TuGR324
43	tetli119g00240	TlGR224	pseudo	164	0	relics	A			TuGR224

44	tetli33g01830	TlGR292a	pseudo	163	0	relics	A			TuGR292
45	tetli239g00200	TlGR228	pseudo	157	1	C-ter relics	A			TuGR228
46	tetli1363g00002	TlGR275	pse/partial	157	0	relics	A			TuGR275
47	tetli73g01010	TlGR263	pseudo	155	0	relics	A			TuGR263
48	tetli268g00150	TlGR292b	pseudo	152	0	N-ter relics	A			TuGR292
49	tetli430g00010	TlGR262	pse/partial	148	1	relics	A			TuGR262
50	tetli40g01220	TlGR290b	pseudo	148	0	relics	A			TuGR290
51	tetli226g00171	TlGR714	pseudo	141	0	relics	A			
52	tetli240g00420	TlGR715	pseudo	140	0	N-ter relics	A			
53	tetli1363g00001	TlGR273	pse/partial	132	1	relics	A			TuGR273
54	tetli61g00778	TlGR716	pseudo	131	0	relics	A			
55	tetli61g00818	TlGR717	pseudo	123	0	relics	A			
56	tetli319g00131	TlGR219c	pse/partial	121	0	N-ter deletion, C-ter missing	A			TuGR219
57	tetli177g00310	TlGR719	pseudo	114	0	relics	A			
58	tetli226g00161	TlGR722	pseudo	108	0	N-ter relics	A			
59	tetli118g00910	TlGR723	pseudo	103	1	relics	A			
60	tetli73g01020	TlGR730	pseudo	86	0	C-ter relics	A			
61	tetli01g03670	TlGR731	pseudo	82	0	C-ter relics	A			
62	tetli01g03660	TlGR733	pseudo	81	1	C-ter relics	A			
63	tetli123g00690	TlGR734	pseudo	80	0	N-ter relics	A			
64	tetli73g01030	TlGR735	pseudo	80	0	relics	A			
1	tetli01g02320	TlGR138a	pseudo	445	3	2 premature stops	B	TuGR138		TuGR138
2	tetli276g00150	TlGR211	pseudo	440	3	4FS &7 PS	B	TuGR211	TuGR211	TuGR211
3	tetli91g00040	TlGR94c	pseudo	433	3	2 FS & 3 PS	B			TuGR94
4	tetli1287g00010	TlGR94b	pseudo	432	3	2 FS & 3 PS	B	TuGR94		TuGR94
5	tetli66g00720	TlGR203	pseudo	427	3	1 PS	B	TuGR203		TuGR203
6	tetli81g00155	TlGR126a	pseudo	419	1	3 FS, int.seq. deletion	B			TuGR126
7	tetli101g00115	TlGR178	pseudo	405	1	many changes	B			TuGR178
8	tetli95g00491	TlGR80	pseudo	372	1	many changes	B			TuGR80
9	tetli01g01205	TlGR632	pseudo	350	0	2FS, N, C-ter deletion	B			TuGR632
10	tetli158g00285	TlGR180b	pseudo	326	0	many changes	B			TuGR180
11	tetli109g00004	TlGR659c	pseudo	323	0	many changes	B			TuGR659
12	tetli1356g00001	TlGR137	pse/partial	309	0	2 FS, 1PS & C-ter missing	B			TuGR137
13	tetli158g00280	TlGR180a	pse/partial	300	1	1 FS & int.seq. missing	B	TuGR180		TuGR180
14	tetli03g01983	TlGR127	pseudo	295	3	1FS & N-ter deletion	B			TuGR127
15	tetli20g03350	TlGR535	pseudo	288	2	many changes	B			TuGR535
16	tetli132g00470	TlGR27	pseudo	281	1	many changes	B	TuGR27	TuGR27	TuGR27
17	tetli01g02295	TlGR134	pseudo	280	1	many changes	B			TuGR134
18	tetli309g00042	TlGR583	pseudo	269	0	many changes	B			TuGR583
19	tetli61g00200	TlGR148	pseudo	265	2	many changes	B			TuGR148
20	tetli335g00020	TlGR175	pseudo	263	0	many changes	B			TuGR175
21	tetli226g00067	TlGR96	pseudo	263	1	int.seq. deletion	B			TuGR96
22	tetli06g01250	TlGR43a	pseudo	232	0	C-ter deletion	B			TuGR43
23	tetli309g00062	TlGR581	pseudo	232	0	many changes	B			TuGR581
24	tetli295g00141	TlGR167	pseudo	225	1	many changes	B			TuGR167
25	tetli84g01240	TlGR85a	pseudo	201	2	int.seq.deletion	B			TuGR85
26	tetli94g00300	TlGR126b	pseudo	178	0	many changes	B			TuGR126
27	tetli109g00610	TlGR712	pseudo	173	1	many changes	B			
28	tetli309g00052	TlGR713	pseudo	173	1	many changes	B			
29	tetli607g00040	TlGR174	pseudo	171	0	relics	B			TuGR174
30	tetli111g00310	TlGR37	pseudo	163	0	N-ter relics	B			TuGR37
31	tetli01g03680	TlGR140	pseudo	156	0	relics	B			TuGR140
32	tetli109g00020	TlGR156b	pseudo	153	0	relics	B			TuGR156
33	tetli84g00320	TlGR85b	pseudo	138	0	N-ter relics	B			TuGR85
34	tetli101g00610	TlGR718	pseudo	115	0	relics	B			
35	tetli111g00320	TlGR720	pseudo	111	0	relics	B			
36	tetli109g00640	TlGR721	pseudo	108	0	relics	B			
37	tetli309g00072	TlGR724	pseudo	103	0	relics	B			
38	tetli109g00620	TlGR726	pseudo	99	0	relics	B			
39	tetli06g00900	TlGR43b	pseudo	98	2	C-ter relics	B			TuGR43
40	tetli23g02190	TlGR727	pseudo	96	0	N-ter relics	B			
41	tetli61g00798	TlGR728	pseudo	96	0	relics	B			
42	tetli47g00820	TlGR7b	pseudo	94	0	relics	B			TuGR7
43	tetli422g00033	TlGR729	pseudo	89	0	relics	B			
44	tetli23g01250	TlGR736	pseudo	78	2	C-ter relics	B			
45	tetli111g00120	TlGR737	pseudo	77	2	C-ter relics	B			
46	tetli78g00670	TlGR739	pseudo	77	0	relics	B			
47	tetli109g00630	TlGR740	pseudo	76	0	relics	B			

48	tetli422g00023	TIGR741	pseudo	75	2	C-ter relics	B			
49	tetli47g00830	TIGR742	pseudo	73	0	relics	B			
50	tetli61g00752	TIGR743	pseudo	73	0	relics	B			
51	tetli101g00620	TIGR525	pseudo	69	0	N-ter relics	B		TuGR525	TuGR525
1	tetli44g00420	TIGR33	pseudo	440	2	1FS in a long polyA stretch	C	TuGR33	TuGR33	TuGR33
2	tetli19g01112	TIGR711	pseudo	199	0	many changes	C			

A.2.2. Phylogenetic trees of GRs in three spider mites


2D 1L 1L
TuGR301
TuGR299
TuGR300
TeGR302
TeGR301
TuGR302
1D 1P 1L 1L
TuGR297
TuGR650*
TIGR298
TuGR298
1D 1P 1L
TuGR474*
1P
1L 1L
TuGR296$
TeGR514#
TuGR293
TuGR294$
1P
TeGR294
1P
TuGR292$$
TeGR292
1L
TuGR295#
TeGR295
1L 1L
TeGR291b
1L
TuGR291
TeGR291a
1D 1P
TeGR290
TIGR290a
TuGR290$
1L 1L
TuGR413*
TeGR287b
1L 1D 1L
TeGR287a
TeGR287c
TeGR282
7D 1L
TuGR288
TuGR285
TuGR289
TuGR284
TuGR286
TuGR287
TuGR282
TuGR283
TeGR281
8D 2P 1L
TuGR280
TuGR277
TuGR281
TuGR276
TuGR482*
TuGR483
TuGR279
TuGR617*
TuGR278
1D
TeGR260b
TIGR258
TuGR258
TuGR257$
1P
TuGR262S
TeGR260a
2D 1L 1L
TuGR261S
TuGR259
TuGR260##$
TuGR273$
TuGR274
20D 5P 1L 1L
TuGR268
TuGR269
TuGR267
TuGR265
TuGR475*
TuGR283S
TuGR264
TuGR476*
TuGR670
TuGR673
TuGR272
TuGR270
TuGR569*
TuGR266$
TuGR275S
TuGR668*
TuGR672*
TuGR669
TuGR271
0
TeGR366
TuGR366
TIGR366
3D 1L 1L
TuGR367
TuGR368
TuGR369
TuGR370##
1L 1L
TuGR384$
TuGR386
TuGR385
TuGR383
TuGR611*
TuGR682*
TIGR386
1L 1L
TuGR382
1L
TuGR381
TIGR381
1L 1L
TuGR398
1P
TuGR397#
TIGR397
1D 1L 1L
TuGR399
TuGR4005
8D 1L 1L
TuGR396
TuGR388
TuGR390
TuGR391
TuGR394
TuGR392
TuGR393
TuGR395
TuGR644
0
TeGR380
TuGR380
TIGR380
TeGR379a
TuGR379#
TIGR379
TuGR541*#
2D 1P 1L
TuGR378
TuGR689
TeGR372
TuGR374
TIGR371
4D 1P
TuGR473*
TuGR372
TuGR371
TuGR373
1L 1L
TuGR472
1L
TuGR387
TIGR387
TeGR376
1L
TuGR377
TuGR375
TuGR376
TIGR376

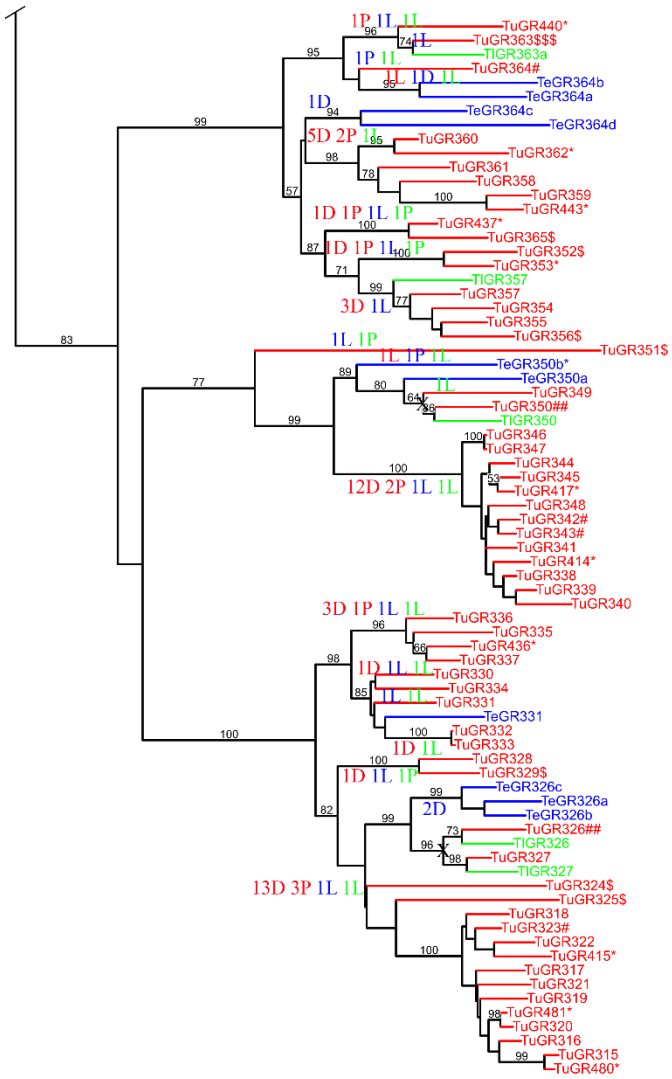
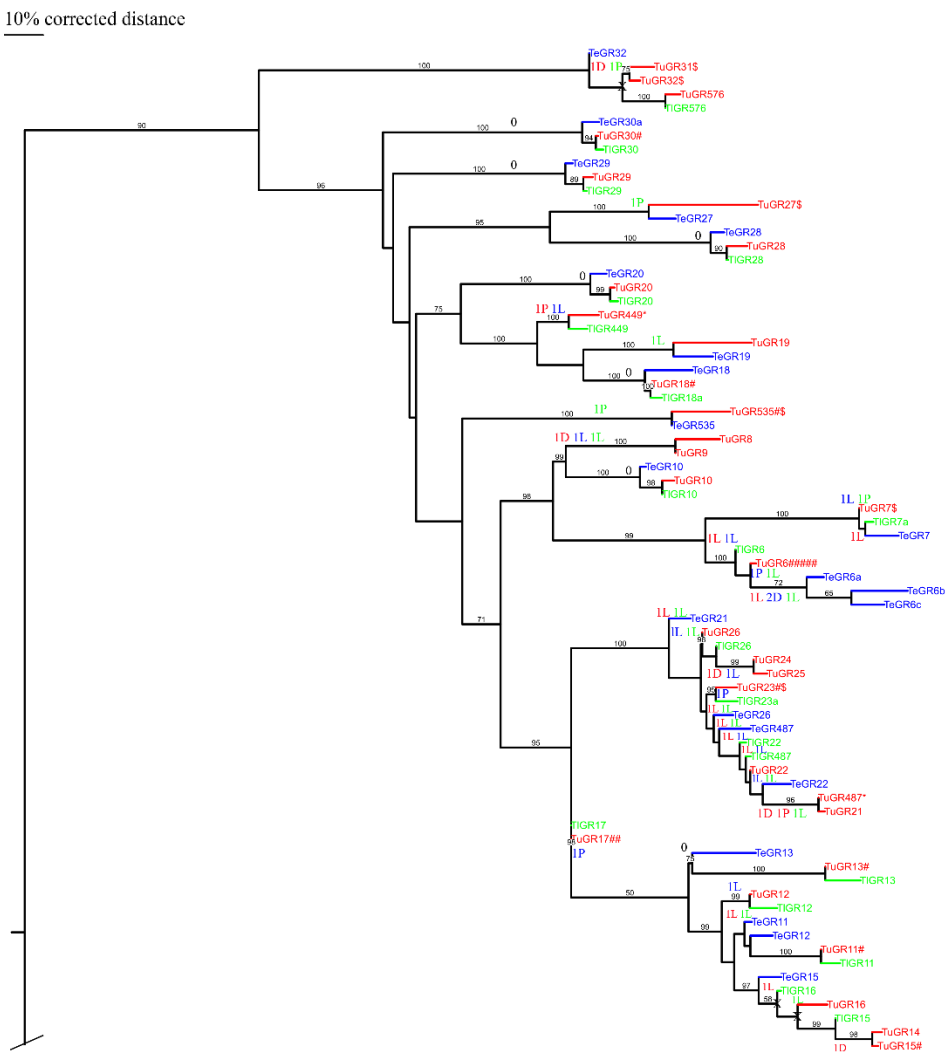
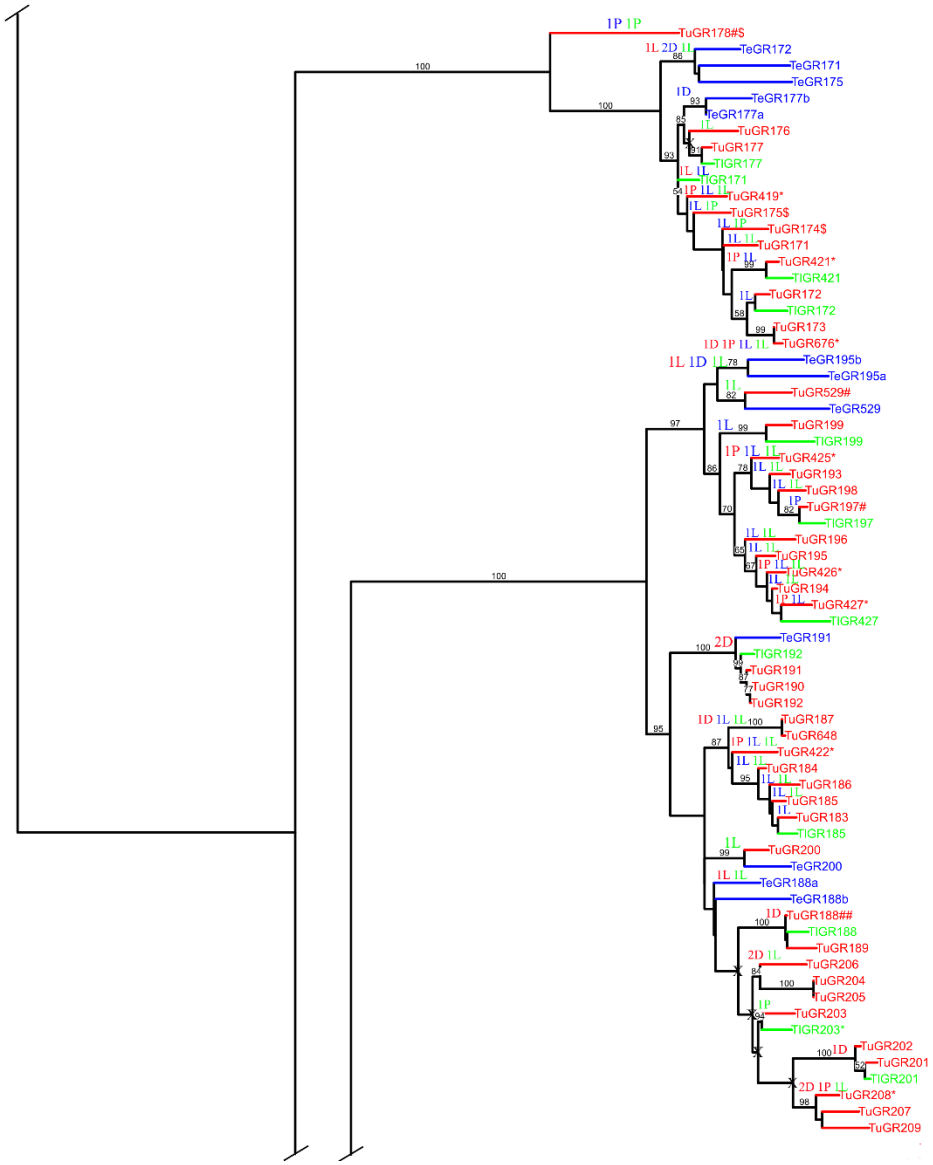
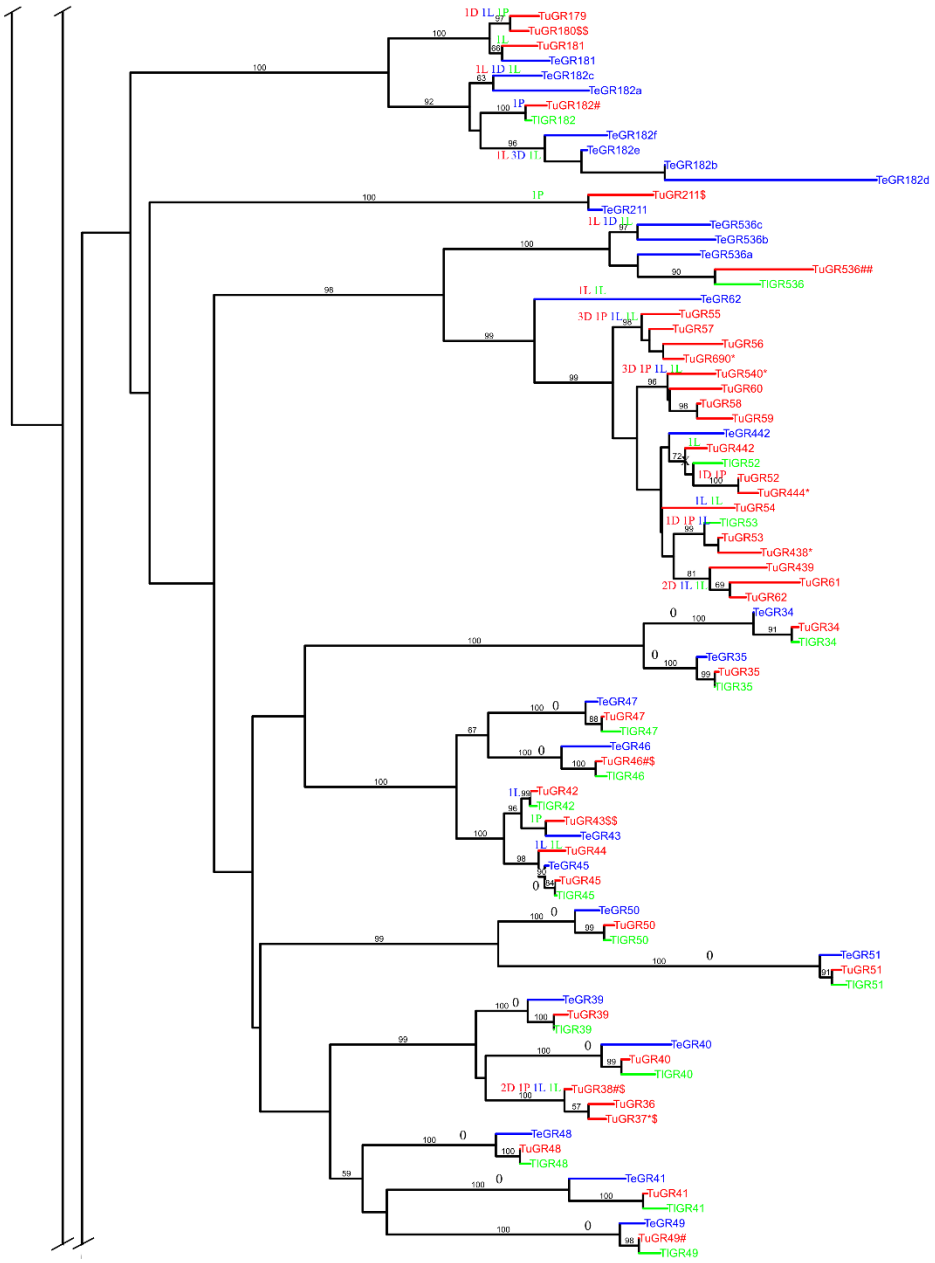
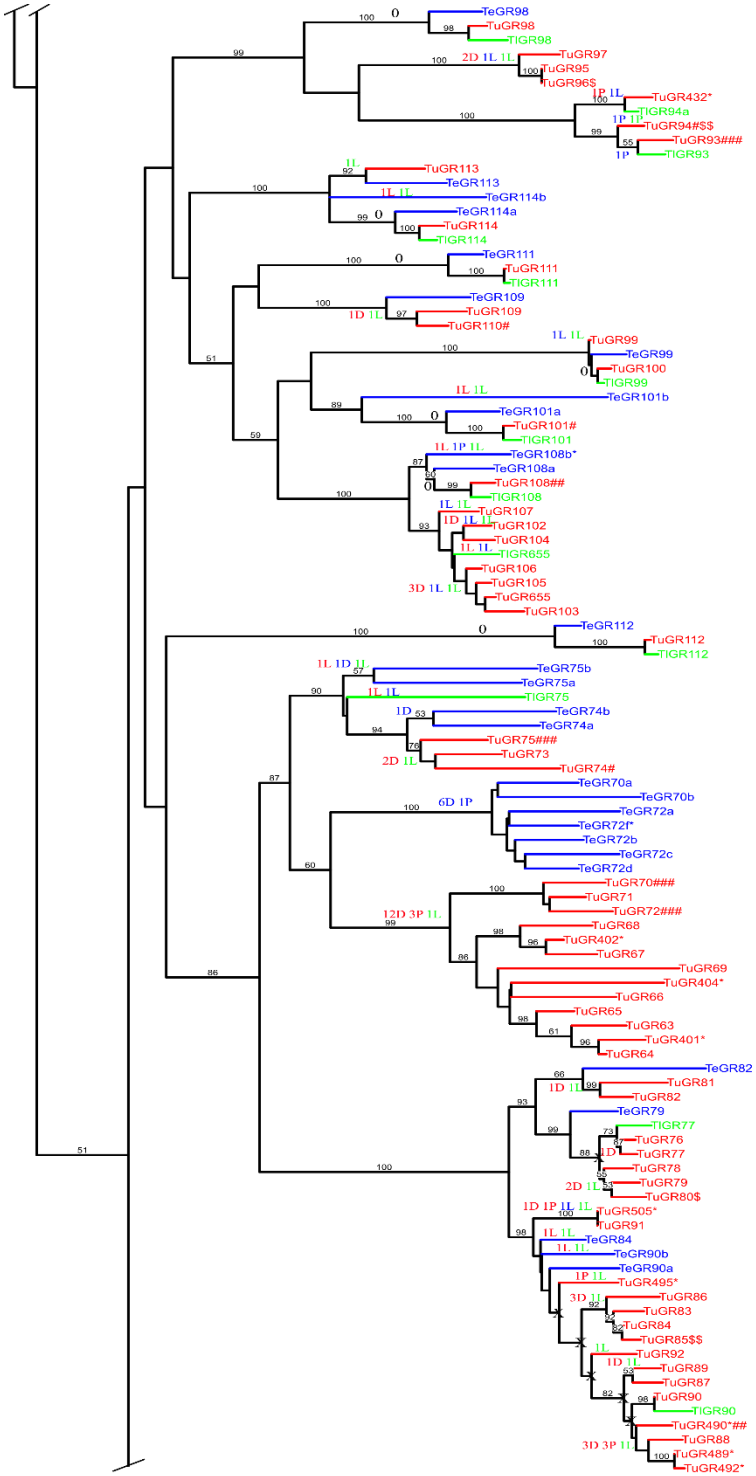


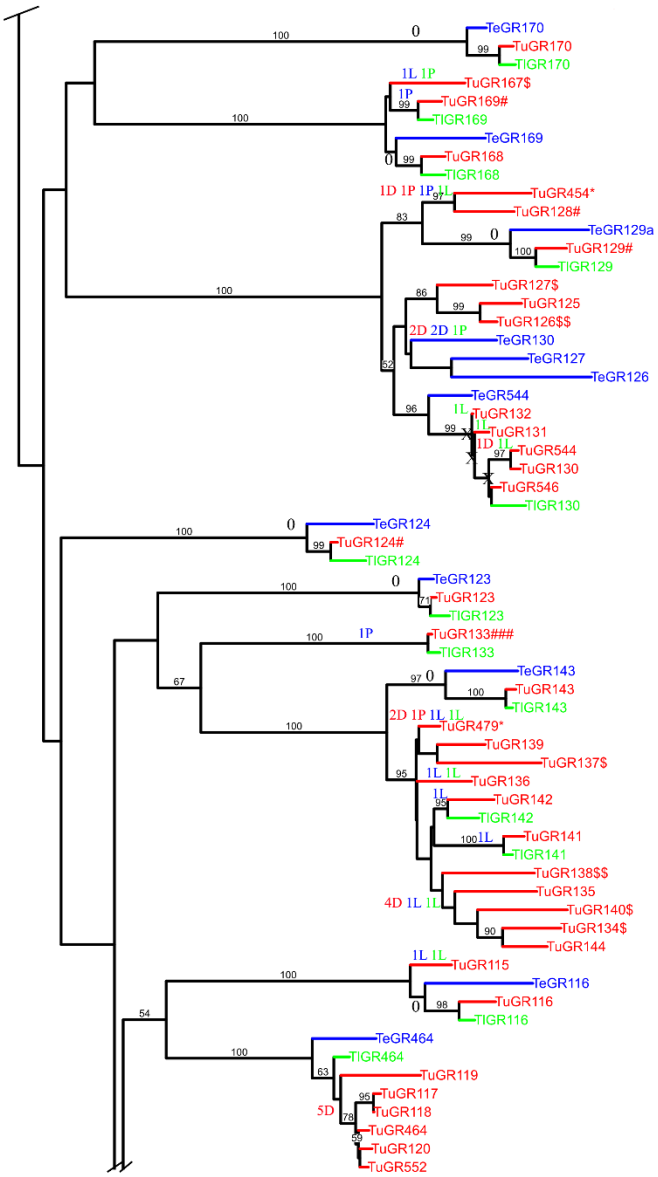
Figure A.8. Phylogenetic tree of the class A GRs (intact genes and pseudogenes with one or two events of class A) from *T. urticae* (red), *T. evansi* (blue) and *T. lintearius* (green) with bootstrap values ≥ 50/100 from 1000 replications of uncorrected distance analysis. This tree was rooted by midpoint rooting. D: Gene Duplication, L: Gene Loss, P: pseudogenization, X: gene duplication predating *T. urticae*-*T. lintearius* speciation. GR with suffix "*" after protein name is pseudogene. TuGR and TeGR with suffix "#" after protein name has homologous pseudogene(s)/partial gene(s) with many changes (blast cutoff 1e-25) in *T. evansi* and *T. urticae*, respectively. TuGR and TlGR with suffix "$" after protein name has homologous pseudogene(s)/partial gene(s) with many changes (blast cutoff 1e-25) in *T. lintearius* and *T. urticae*, respectively.


10% corrected distance
TeGR32
TuGR31$
TuGR32$
TuGR576
TlGR576
TeGR30a
TuGR30#
TlGR30
TeGR29
TuGR29
TlGR29
TuGR27$
TeGR27
TeGR28
TuGR28
TlGR28
TeGR20
TuGR20
TlGR20
TuGR449*
TlGR449
TuGR19
TeGR19
TeGR18
TuGR18#
TlGR18a
TuGR535#$
TeGR535
TuGR8
TuGR9
TeGR10
TuGR10
TlGR10
TuGR7$
TlGR7a
TeGR7
TlGR6
TuGR6#####
TeGR6a
TeGR6b
TeGR6c
TeGR21
TuGR26
TlGR26
TuGR24
TuGR25
TuGR23#$
TlGR23a
TeGR26
TeGR487
TlGR22
TlGR487
TuGR22
TeGR22
TuGR487*
TuGR21
TlGR17
TuGR17##
TeGR13
TuGR13#
TlGR13
TuGR12
TlGR12
TeGR11
TeGR12
TuGR11#
TlGR11
TeGR15
TlGR16
TuGR16
TlGR15
TuGR14
TuGR15#


1P 1P
TuGR178#$
1L 2D 1L
TeGR172
TeGR171
TeGR175
1D
TeGR177b
TeGR177a
1L
TuGR176
TuGR177
TIGR177
1L 1L
TIGR171
1P 1L 1L
TuGR419*
1L 1P
TuGR175$
1L 1P
TuGR174$
1L 1L
TuGR171
1P 1L
TuGR421*
TIGR421
1L
TuGR172
TIGR172
TuGR173
1D 1P 1L 1L
TuGR676*
1L 1D 1L
TeGR195b
TeGR195a
1L
TuGR529#
TeGR529
1L
TuGR199
TIGR199
1P 1L 1L
TuGR425*
1L 1L
TuGR193
1L 1L
TuGR198
1P
TuGR197#
TIGR197
1L 1L
TuGR196
1L 1L
TuGR195
1P 1L 1L
TuGR426*
1L
TuGR194
1P 1L
TuGR427*
TIGR427
TeGR191
2D
TIGR192
TuGR191
TuGR190
TuGR192
1D 1L 1L
TuGR187
1P 1L 1L
TuGR648
1L 1L
TuGR422*
1L 1L
TuGR184
1L 1L
TuGR186
1L 1L
TuGR185
1L
TuGR183
TIGR185
1L
TuGR200
TeGR200
1L 1L
TeGR188a
TeGR188b
1D
TuGR188##
TIGR188
TuGR189
2D 1L
TuGR206
TuGR204
TuGR205
1P
TuGR203
TIGR203*
1D
TuGR202
TuGR201
TIGR201
2D 1P 1L
TuGR208*
TuGR207
TuGR209


1D 1L 1P
TuGR179
TuGR180$$
1L
TuGR181
TeGR181
1L 1D 1L
TeGR182c
TeGR182a
1P
TuGR182#
TIGR182
TeGR182f
TeGR182e
1L 3D 1L
TeGR182b
TeGR182d
1P
TuGR211$
TeGR211
1L 1D 1L
TeGR536c
TeGR536b
TeGR536a
TuGR536##
TIGR536
1L 1L
TeGR62
3D 1P 1L 1L
TuGR55
TuGR57
TuGR56
TuGR690*
3D 1P 1L 1L
TuGR540*
TuGR60
TuGR58
TuGR59
TeGR442
1L
TuGR442
TIGR52
1D 1P
TuGR52
TuGR444*
1L 1L
TuGR54
1D 1P 1L
TIGR53
TuGR53
TuGR438*
TuGR439
2D 1L 1L
TuGR61
TuGR62
TeGR34
TuGR34
TIGR34
TeGR35
TuGR35
TIGR35
TeGR47
TuGR47
TIGR47
TeGR46
TuGR46#$
TIGR46
1L
TuGR42
TIGR42
1P
TuGR43$$
TeGR43
1L 1L
TuGR44
TeGR45
TuGR45
TIGR45
TeGR50
TuGR50
TIGR50
TeGR51
TuGR51
TIGR51
TeGR39
TuGR39
TIGR39
TeGR40
TuGR40
TIGR40
2D 1P 1L 1L
TuGR38#$
TuGR36
TuGR37*$
TeGR48
TuGR48
TIGR48
TeGR41
TuGR41
TIGR41
TeGR49
TuGR49#
TIGR49


TeGR98
TuGR98
TlGR98
2D 1L 1L
TuGR97
TuGR95
TuGR96$
1P 1L
TuGR432*
TlGR94a
1P 1P
TuGR94#$$
TuGR93###
1P
TlGR93
1L
TuGR113
TeGR113
1L 1L
TeGR114b
TeGR114a
TuGR114
TlGR114
TeGR111
TuGR111
TlGR111
TeGR109
1D 1L
TuGR109
TuGR110#
1L 1L
TuGR99
TeGR99
TuGR100
TlGR99
1L 1L
TeGR101b
TeGR101a
TuGR101#
TlGR101
1L 1P 1L
TeGR108b*
TeGR108a
TuGR108##
TlGR108
1L 1L
TuGR107
1D 1L 1L
TuGR102
TuGR104
1L 1L
TlGR655
TuGR106
3D 1L 1L
TuGR105
TuGR655
TuGR103
TeGR112
TuGR112
TlGR112
1L 1D 1L
TeGR75b
TeGR75a
1L 1L
TlGR75
1D
TeGR74b
TeGR74a
TuGR75###
2D 1L
TuGR73
TuGR74#
TeGR70a
TeGR70b
6D 1P
TeGR72a
TeGR72f*
TeGR72b
TeGR72c
TeGR72d
TuGR70###
TuGR71
12D 3P 1L
TuGR72###
TuGR68
TuGR402*
TuGR67
TuGR69
TuGR404*
TuGR66
TuGR65
TuGR63
TuGR401*
TuGR64
TeGR82
1D 1L
TuGR81
TuGR82
TeGR79
TlGR77
TuGR76
1D
TuGR77
TuGR78
2D 1L
TuGR79
TuGR80$
1D 1P 1L 1L
TuGR505*
TuGR91
1L 1L
TeGR84
1L 1L
TeGR90b
TeGR90a
1P 1L
TuGR495*
3D 1L
TuGR86
TuGR83
TuGR84
TuGR85$$
1L
TuGR92
1D 1L
TuGR89
TuGR87
TuGR90
TlGR90
TuGR490*##
TuGR88
3D 3P 1L
TuGR489*
TuGR492*


TeGR170
TuGR170
TlGR170
1L 1P
TuGR167$
1P
TuGR169#
TlGR169
TeGR169
TuGR168
TlGR168
1D 1P 1P 1L
TuGR454*
TuGR128#
TeGR129a
TuGR129#
TlGR129
TuGR127$
TuGR125
TuGR126$$
2D 2D 1P
TeGR130
TeGR127
TeGR126
TeGR544
1L
TuGR132
1L
TuGR131
1D 1L
TuGR544
TuGR130
TuGR546
TlGR130
TeGR124
TuGR124#
TlGR124
TeGR123
TuGR123
TlGR123
TuGR133###
1P
TlGR133
TeGR143
TuGR143
TlGR143
2D 1P 1L 1L
TuGR479*
TuGR139
TuGR137$
1L 1L
TuGR136
1L
TuGR142
TlGR142
1L
TuGR141
TlGR141
TuGR138$$
TuGR135
4D 1L 1L
TuGR140$
TuGR134$
TuGR144
1L 1L
TuGR115
TeGR116
TuGR116
TlGR116
TeGR464
TlGR464
TuGR119
TuGR117
5D
TuGR118
TuGR464
TuGR120
TuGR552

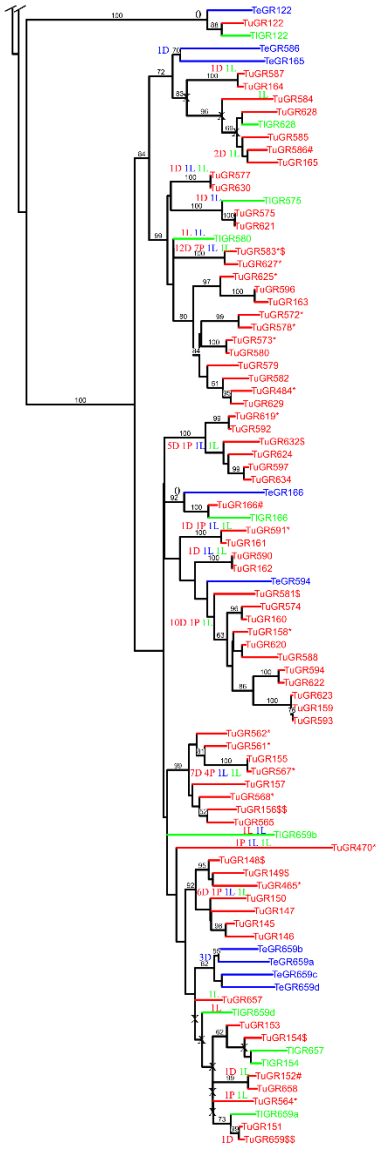


Figure A.9. Phylogenetic tree of the class B GRs (intact genes and pseudogenes with one or two events of class B) from *T. urticae* (red), *T. evansi* (blue) and *T. lintearius* (green) with bootstrap values ≥ 50/100 from 1000 replications of uncorrected distance analysis. This tree was rooted by midpoint rooting. D: Gene Duplication, L: Gene Loss, P: pseudogenization, X: gene duplication predating *T. urticae*-*T. lintearius* speciation. GR with suffix "*" after protein name is pseudogene. TuGR and TeGR with suffix "#" after protein name has homologous pseudogene(s)/partial gene(s) with many changes (blast cutoff 1e-25) in *T. evansi* and *T. urticae*, respectively. TuGR and TlGR with suffix "$" after protein name has homologous pseudogene(s)/partial gene(s) with many changes (blast cutoff 1e-25) in *T. lintearius* and *T. urticae*, respectively.

A.2.3. Microsyntenies

Table A.7. Microsyntenies of GRs in *T. urticae*, *T. evansi* and *T. lintearius*

N	*T. urticae*	*T. evansi*	*T. lintearius*
1	tetur28g00170-TuGR111-tetur28g00190	tetev69g00740-TeGR111-tetev69g00720	TlGR111-tetli170g00150
2	tetur07g01840-TuGR116	tetev06g00490-TeGR116	tetli25g00100-TlGR116
3	tetur02g00200-TuGR122-tetur02g00170	tetev01g01610-TeGR122-tetev01g01640	TlGR122-tetli29g02110
4	tetur02g02280-TuGR123-tetur02g02270	tetev113g00310-TeGR123-tetev113g00300	tetli65g00370-TlGR123
5	tetur02g08940-TuGR124-tetur02g08920	TeGR124-tetev85g00250	tetli92g00490-TlGR124-tetli92g00510
6	tetur17g03100-TuGR143	tetev24g00890-TeGR143	tetli01g02270-TlGR143
7	tetur19g00250-TuGR166	tetev42g00140-TeGR166	tetli03g00160-TlGR166-tetli03g00150
8	tetur17g03100-TuGR17	tetev24g00890-TeGR17a	tetli01g02270-TlGR17
9	tetur05g03550-TuGR182	tetev179g00300-TeGR182a	tetli190g00020-TlGR182
10	tetur06g01490-TuGR2-tetur06g01460	TeGR2-tetev183g00100	tetli41g00110-TlGR2-tetli41g00080
11	tetur01g06390-TuGR210-tetur01g06370	tetev66g00510-TeGR210-tetev66g00490	tetli57g00150-TlGR210-tetli57g00130
12	tetur27g01220-TuGR211	tetev184g00140-TeGR211	tetli276g00060-TlGR211
13	tetur20g00820-TuGR212-tetur20g00800	tetev88g00920-TeGR212-tetev88g00900	tetli116g00720-TlGR212-tetli116g00740
14	tetur22g02620-TuGR215-tetur22g02590	tetev377g00070-TeGR215-tetev377g00090	tetli117g00400-TlGR215-tetli117g00380
15	tetur30g00900-TuGR22-tetur30g00890	tetev143g00030-TeGR22	tetli90g00290-TlGR22-tetli90g00270
16	tetur37g00380-TuGR251-tetur37g00360	tetev129g00060-TeGR251-tetev129g00070	tetli64g00750-TlGR251-tetli64g00770
17	tetur06g05450-TuGR253-tetur06g05430	tetev154g00380-TeGR253-tetev154g00360	tetli33g00210-TlGR253-tetli33g00230
18	tetur13g00940-TuGR255-tetur13g00960	tetev13g00610-TeGR255	tetli51g00680-TlGR255-tetli51g00660
19	tetur24g00920-TuGR256	tetev98g00700-TeGR256	tetli104g00370-TlGR256
20	tetur04g08790-TuGR27-tetur04g08770	tetev125g00550-TeGR27-tetev125g00570	tetli132g00460-TlGR27-tetli132g00480
21	tetur08g00320-TuGR28-tetur08g00340	tetev151g00160-TeGR28-tetev151g00140	tetli138g00390-TlGR28-tetli138g00410
22	TuGR29-tetur16g00660	tetev03g00560-TeGR29-tetev03g00530	tetli14g00740-TlGR29-tetli14g00760
23	tetur03g07900-TuGR292	tetev11g01260-TeGR292	tetli33g01240-TlGR292a
24	tetur07g04380-TuGR3-tetur07g04360	TeGR3-tetev14g01010	tetli88g00260-TlGR3-tetli88g00290
25	tetur14g01030-TuGR305-tetur14g01050	tetev94g00050-TeGR305	tetli225g00070-TlGR305-tetli225g00090
26	tetur06g00760-TuGR33-tetur06g00780	tetev100g00430-TeGR33-tetev100g00410	tetli44g00430-TlGR33-tetli44g00410
27	tetur14g00700-TuGR35	tetev108g00370-TeGR35-tetev108g00350	tetli56g00330-TlGR35-tetli56g00310
28	tetur10g02830-TuGR47-tetur10g02860	tetev104g00150-TeGR47-tetev104g00120	tetli55g01630-TlGR47-tetli55g01650
29	tetur26g02720-TuGR48-tetur26g02730	tetev11g01010-TeGR48-tetev11g01030	tetli45g00340-TlGR48-tetli45g00360
30	tetur30g00850-TuGR487	tetev303g00090-TeGR487	tetli90g00230-TlGR487
31	tetur11g04920-TuGR49-tetur11g04940	tetev21g00340-TeGR49-tetev21g00320	tetli133g00210-TlGR49-tetli133g00190
32	tetur11g00470-TuGR5-tetur11g00440	tetev175g00100-TeGR5-tetev175g00130	tetli221g00120-TlGR5-tetli221g00090
33	tetur04g02770-TuGR50	tetev170g00180-TeGR50	tetli91g00330-TlGR50
34	tetur11g01970-TuGR51-tetur11g01990	tetev161g00090-TeGR51-tetev161g00070	tetli72g00190-TlGR51-tetli72g00210
35	tetur11g00470-TuGR510-tetur11g00440	tetev175g00100-TeGR510-tetev175g00130	tetli221g00120-TlGR510-tetli221g00090
36	tetur27g01930-TuGR7	tetev67g00090-TeGR7	tetli47g00740-TlGR7a
1	tetur19g02600-TuGR101	tetev188g00180-TeGR101a	
2	tetur24g00230-TuGR109	tetev105g00600-TeGR109	
3	tetur44g00060-TuGR112	tetev256g00060-TeGR112	
4	tetur14g02620-TuGR129	tetev90g00390-TeGR129a	
5	tetur07g06690-TuGR15-tetur07g06700	tetev15g00680-TeGR15-tetev15g00660	
6	tetur17g00140-TuGR191	tetev171g00050-TeGR191	
7	tetur09g06729-TuGR20-tetur09g06550	tetev10g00230-TeGR20-tetev10g00250	
8	tetur12g00800-TuGR217	tetev54g00650-TeGR217	
9	tetur14g02860-TuGR250	tetev90g00200-TeGR250	
10	tetur30g00850-TuGR26	tetev303g00090-TeGR26	
11	tetur31g01840-TuGR294	tetev52g00620-TeGR294	
12	tetur03g04260-TuGR295	tetev157g00030-TeGR295	
13	tetur24g01030-TuGR30	tetev58g00210-TeGR30a	
14	tetur17g03850-TuGR32	tetev95g00390-TeGR32	
15	tetur08g00520-TuGR350	tetev212g00100-TeGR350a	
16	tetur12g01050-TuGR366	tetev54g00470-TeGR366	
17	tetur13g00340-TuGR380-tetur13g00350	tetev13g01250-TeGR380-tetev13g01230	
18	tetur02g09570-TuGR4	tetev79g00240-TeGR4	
19	tetur03g09080-TuGR40	tetev191g00100-TeGR40	
20	tetur13g03040-TuGR43	tetev86g00390-TeGR43	
21	tetur08g01360-TuGR442	tetev140g00110-TeGR442	
22	tetur13g03040-TuGR45	tetev86g00390-TeGR45	
23	tetur15g00460-TuGR471	tetev226g00080-TeGR313	
24	tetur02g07810-TuGR514-tetur02g07830	tetev89g00470-TeGR514-tetev89g00490	
25	tetur04g04290-TuGR529-tetur04g04300	tetev34g00060-TeGR529-tetev34g00070	
26	tetur08g01360-TuGR536	tetev140g00110-TeGR536a	
27	tetur17g00140-TuGR594	tetev171g00050-TeGR594	
28	tetur33g01070-TuGR79	tetev144g00090-TeGR79	
29	tetur33g01070-TuGR82	tetev144g00090-TeGR82	
30	tetur33g00160-TuGR84	tetev61g00790-TeGR84	
1		tetev75g00250-TeGR10	tetli03g01620-TlGR10
2		tetev73g00560-TeGR108a	tetli78g00510-TlGR108
3		tetev369g00050-TeGR11	tetli106g00070-TlGR11
4		tetev171g00050-TeGR191	tetli01g01220-TlGR192
5		tetev18g00670-TeGR214	tetli142g00160-TlGR214
6		tetev54g00650-TeGR220	tetli239g00190-TlGR434
7		tetev54g00455-TeGR366	tetli240g00180-TlGR366
8		tetev79g00260-TeGR4	tetli92g01130-TlGR4

9		tetev86g00390-TeGR43	tetli06g01270-TlGR42
10		tetev188g00180-TeGR464	tetli294g00070-TlGR464
11		tetev144g00090-TeGR79	tetli95g00390-TlGR77
12		tetev61g00790-TeGR84	tetli84g00360-TlGR90
1	tetur19g01780-TuGR10		tetli03g01600-TlGR10
2	tetur05g07860-TuGR108-tetur05g07880		tetli78g00490-TlGR108-tetli78g00510
3	tetur01g10770-TuGR11		tetli106g00070-TlGR11
4	tetur19g02710-TuGR114		tetli354g00040-TlGR114
5	tetur14g02620-TuGR129		tetli54g00700-TlGR129
6	tetur01g07530-TuGR13		tetli113g00310-TlGR13
7	tetur22g00940-TuGR130		tetli12g00380-TlGR130
8	tetur17g03100-TuGR141		tetli01g02270-TlGR141
9	tetur17g02950-TuGR142		tetli01g02340-TlGR142
10	tetur07g06690-TuGR15-tetur07g06700		tetli31g01040-TlGR15-tetli31g01030
11	tetur07g06700-TuGR16		tetli31g01030-TlGR16
12	tetur03g08940-TuGR172		tetli111g00290-TlGR172
13	tetur03g08780-TuGR178		tetli101g00120-TlGR178
14	tetur20g02800-TuGR188		tetli70g00400-TlGR188
15	tetur24g02680-TuGR192		tetli01g01220-TlGR192
16	tetur04g04010-TuGR197		tetli194g00070-TlGR197
17	tetur16g02530-TuGR199-tetur16g02550		tetli45g00870-TlGR199-tetli45g00260
18	tetur03g02200-TuGR213-tetur03g02220		tetli01g00180-TlGR213-tetli01g00160
19	tetur13g04400-TuGR252		tetli108g00110-TlGR252a
20	tetur01g03800-TuGR257		tetli26g00310-TlGR257
21	tetur02g08870-TuGR296		tetli92g00190-TlGR296
22	tetur24g01030-TuGR30-tetur24g01050		tetli104g00420-TlGR30-tetli104g00440
23	tetur07g04210-TuGR303		tetli80g00090-TlGR303
24	tetur13g02930-TuGR308-tetur13g02950		tetli06g01400-TlGR308-tetli06g01360
25	tetur25g01930-TuGR309-tetur25g01940		tetli232g00190-TlGR309-tetli232g00170
26	tetur17g03900-TuGR31		tetli01g01550-TlGR31
27	tetur15g00410-TuGR311-tetur15g00460		tetli123g00290-TlGR311-tetli123g00270
28	tetur08g00550-TuGR350		tetli177g00050-TlGR350
29	tetur40g00040-TuGR351-tetur40g00030		tetli43g00180-TlGR351-tetli43g00160
30	tetur09g06715-TuGR371		tetli39g00900-TlGR371
31	tetur18g03460-TuGR381		tetli66g00690-TlGR381
32	tetur18g03520-TuGR386		tetli01g01330-TlGR386
33	tetur01g02900-TuGR387		tetli26g01290-TlGR387
34	tetur17g00710-TuGR397		tetli55g00300-TlGR397
35	tetur17g00710-TuGR400		tetli55g00300-TlGR400
36	tetur13g03040-TuGR42		tetli06g01270-TlGR42
37	tetur03g08940-TuGR421		tetli111g00290-TlGR421
38	tetur04g02580-TuGR432		tetli91g00060-TlGR94a
39	tetur12g00800-TuGR456		tetli239g00190-TlGR456a
40	tetur12g01710-TuGR457		tetli119g00060-TlGR457
41	tetur19g02130-TuGR46-tetur19g02140		tetli65g01010-TlGR46-tetli65g01000
42	tetur12g00830-TuGR460		tetli240g00330-TlGR460a
43	tetur15g00460-TuGR471-tetur15g00410		tetli123g00270-TeGR471-tetli123g00290
44	tetur05g08460-TuGR511-tetur05g08440		tetli143g00180-TlGR511-tetli143g00160
45	tetur03g08820-TuGR525		tetli101g00110-TlGR525
46	tetur11g00430-TuGR543		tetli221g00080-TlGR543
47	tetur17g03900-TuGR576		tetli01g01550-TlGR576
48	tetur33g01070-TuGR77		tetli95g00390-TlGR77
49	tetur33g00160-TuGR90		tetli84g00360-TlGR90
50	tetur04g02550-TuGR93		tetli91g00010-TlGR93

List of abbreviations

aa	amino acid
AMPA	α-Amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
bp	base pair
cDNA	complementary DNA
CDS	Coding sequence
ChIP	Chromatin ImmunoPrecipitation
CRs	Chemosensory Receptors
dATP	deoxyadenosine triphosphate
DBDs	DNA binding domains
dCTP	deoxycytidine triphosphate
dGTP	deoxyguanosine triphosphate
DNA	Deoxyribonucleic acid
dNTPs	deoxynucleotide triphosphates
dTTP	deoxythymidine triphosphate
ENaCs	Epithelial Na+ Channels
ESTs	Expressed Sequence Tag
GHMM	Generalized Hidden Markov Model
GOLD	Genomes OnLine Database
GPCRs	G-protein Coupled Receptors
GRs	Gustatory Receptors
HAMPs	Herbivore-Associated Molecular Patterns
HIPVs	Herbivore-induced plant volatiles
HMM	Hidden Markov Model
HMP	Human Microbiome project
iGluRs	ionotropic Glutamate Receptors

IMMs	Interpolated Markov Models
IRs	Ionotropic Receptors
LBD	Ligand-Binding Domain
mGluRs	metabotropic Glutamate Receptors
MM	Markov Model
mRNA	messenger RNA
MYA	Million Years Ago
NGS	Next Generation Sequencing
NMDA	N-methyl-D-aspartate
ORCAE	Online Resource for Community Annotation of Eukaryotes
ORs	Odorant Receptors
PAML	Phylogenetic Analysis by Maximum Likelihood
PCR	Polymerase Chain Reaction
PHYLIP	Phylogeny Inference Package
PWM	Positional weight matrices
RNA	Ribonucleotide
RNA-seq	RNA Sequencing
snRNPs	small nuclear RiboNucleic Particles
Te	Tetranychus evansi
TEs	Transposable Elements
TFs	Transcription Factors
TIGR	The Institute for Genomic Research
Tl	Tetranychus lintearius
TM	Transmembrane
Tu	Tetranychus urticae
UCLA	University of California, Los Angeles

Curriculum Vitae

CAO THI NGOC PHUONG

Date of Birth: 10 September 1981
Place of Birth: Thanh hoa, Vietnam

VIB Department of Plant Systems Biology, Ghent University
Technologiepark 927, 9052 Gent, BELGIUM

Email: phuong.thingoccao@psb.vib-ugent.be / ctnphuong@hcmus.edu.vn

Education

Bachelor degree in Biotechnology, University of Natural Sciences, VNU-HCMC, Vietnam (1999-2003). Dissertation: "Building chitinase protein database".

Master degree in Microbiology, University of Natural Sciences, VNU-HCMC, Vietnam (2004-2006). Dissertation: "Building the program to design protein and applying to increase stability of hG-CSF protein *in silico*".

PhD student, Ghent University, Belgium (2009-2014) (Promoter: Prof. Dr. Yves Van de Peer). PhD thesis "Genome annotation and evolution of chemosensory receptors in spider mites".

Publications

1. Grbic, M., Van Leeuwen, T., Clark, T. G., Rombauts, S., Rouzé, P., Grbic, V., Osborne, E., Dermauw, W., **Thi Ngoc Cao, P.**, Ortego, F., Hernandez-Crespo, P., Diaz, I., Martinez, N.J., Navajas, M., Sucena, E., Magalhaes, S., Nagy, L., Pace, N.R.,

Djuranovic, S., Smagghe, G., Iga, M., Christiaens, M., Veenstra, J., Ewer, J., Mancilla Villalobos, R., Hutter, J., Hudson, A., Velez, M., Yi, S., Zeng, Q., Pires-daSilva, A., Roch, F., Cazaux, M., Navarro, M., Zhurov, V., Acevedo, G., Bjelica, A., Fawcett, J., Bonnet, E., Martens, C., Baele, G., Wissler, L., Sanchez-Rodriguez, A., Tirry, L., Blais, C., Demeestere, K., Henz, SR., Gregory, R., Mathieu, J., Verdon, L., Farinelli, L., Schmutz, J., Lindquist, E., Feyereisen, R., Van de Peer, Y. (2011) The genome of Tetranychus urticae reveals herbivorous pest adaptations. Nature 479,487-492.

2. Phuong Cao Thi Phuong, Thomas Van Leeuwen, Robert Greehalgh, Stephane Rombauts, Richard M. Clark, Miodrag Grbic, Yves Van de Peer and Pierre Rouze, The first chelicerate genome illustrates evolutionary innovation, fine tuning and adaptative plasticity in the arthropod chemoreceptor gene repertoire (In preparation).

Oral presentations

1. "Transcription factors in *T. urticae* genome", and "Exploring Prokaryotic metagenomics in *T. urticae* genome", the first meeting of the spider mite genome consortium, Logroño, Spain, October 25-28, 2009.
2. "Transcription factor families in *Tetranychus urticae* genome", and "Microflora in spider mite", the second meeting of the spider mite genome consortium, Tivat, Montenegro, September 20-23, 2010.
3. "The chemosensory receptors in the spider mite genome", the third meeting of the spider mite genome consortium, Tarragona, Spain, September 26-29, 2011.

Poster presentations

"The First Complete Chelicerate Genome of Tetranychus urticae", Kick-Off meeting MRP Bioinformatics: From Nucleotides to Networks (N2N), 4th May, 2011, Ghent, Belgium.

References

[1] Schopf JW (2001). Cradle of Life: The Discovery of Earth's Earliest Fossils. Princeton University Press.

[2] Mora C, Tittensor DP, et al. (2011). How many species are there on Earth and in the ocean?. PLoS Biol 9(8): e1001127.

[3] May RM (2010). Ecology. Tropical arthropod species, more or less?. Science 329: 41-2.

[4] Hanson B, Chin G, et al. (1999). The diversity of evolution. Science 25: 2105.

[5] Watson JD, Crick FH (1953). The structure of DNA. Cold Spring Harb Symp Quant Biol 18:123-31.

[6] Wu R, Kaiser AD (1968). Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. J Mol Biol 35: 523-37.

[7] Crick FH (1958). On protein synthesis. Symp Soc Exp Biol 12: 138-63.

[8] Crick FH (1970). Central dogma of molecular biology. Nature 227: 561-3.

[9] Ayala FJ (1977). "Nothing in biology makes sense except in the light of evolution": Theodosius Dobzhansky: 1900-1975. J Hered 68: 3-10.

[10] Sanger F, Nicklen S, et al. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463-7.

[11] Maxam AM, Gilbert W (1977). A new method for sequencing DNA. Proc Natl Acad Sci USA 74: 560-4.

[12] Fleischmann RD, Adams MD, et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269: 496-512.

[13] Fraser CM, Gocayne JD, et al. (1995). The minimal gene complement of *Mycoplasma genitalium*. Science 270: 397-403.

[14] Blattner FR, Plunkett G, et al. (1997). The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453-62.

[15] Perna NT, Plunkett G, et al. (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409: 529-33.

[16] Welch RA, Burland V, et al. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci USA 99: 17020-4.

[17] Goffeau A, Barrell BG, et al. (1996). Life with 6000 genes. Science 274: 546, 563-7.

[18] Kunst F, Ogasawara N, et al. (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390: 249-56.

[19] *C. elegan* sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282: 2012-8.

[20] Adams MD, Celniker SE, et al. (2000). The genome sequence of *Drosophila melanogaster*. Science 287: 2185-95.

[21] The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796-815.

[22] Venter JC, Adams MD, et al. (2001). The sequence of the human genome. Science 291: 1304-51.

[23] Lander ES, Linton LM, et al. (2001). Initial sequencing and analysis of the human genome. Nature 409: 860-921.

[24] Imelfort M, Batley J, et al. (2009). Genome sequencing approaches and successes. Methods Mol Biol 513: 345-58.

[25] Bentley DR (2006). Whole-genome re-sequencing. Curr Opin Genet Dev 16: 545-52.

[26] Shendure J, Ji H (2008). Next-generation DNA sequencing. Nat Biotechnol 26: 1135-45.

[27] Pareek CS, Smoczynski R, et al. (2011). Sequencing technologies and genome sequencing. J Appl Genet 52: 413-35.

[28] Margulies M, Egholm M, et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-80.

[29] Ronaghi M, Karamohamed S, et al. (1996). Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 242: 84-9.

[30] Bentley DR, Balasubramanian S, et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53-9.

[31] Shendure J, Porreca GJ, et al. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309: 1728-32.

[32] Harris TD, Buzby PR, et al. (2008). Single-molecule DNA sequencing of a viral genome. Science 320: 106-9.

[33] Li Y, Wang J (2009). Faster human genome sequencing. Nat Biotechnol 27(9):820-1.

[34] Schadt EE, Turner S, Kasarskis A (2010). A window into third-generation sequencing. Hum Mol Genet. 19(R2):R227-40.

[35] Munroe DJ, Harris TJ (2010). Third-generation sequencing fireworks at Marco Island. Nat Biotechnol. 28(5):426-8.

[36] Eid J, Fehr A, Gray J, et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science 323(5910):133–138.

[37] Branton D, Deamer DW, Marziali A, et al. (2008). The potential and challenges of nanopore sequencing. Nat Biotechnol. 26(10):1146-53.

[38] Rothberg JM, Hinz W, et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature 475: 348-52.

[39] Liu L, Li Y, et al. (2012). Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012: 251364.

[40] Zhou X, Ren L, et al. (2010). The next-generation sequencing technology and application. Protein Cell 1: 520-36.

[41] Velasco R, Zharkikh A, et al. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS One 2: e1326.

[42] Goldberg SM, Johnson J, et al. (2006). A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci USA 103: 11240-5.

[43] Diguistini S, Liao NY, et al. (2009). De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biol 10(9): R94.

[44] Huang S, Li R, et al. (2009). The genome of the cucumber, Cucumis sativus L. Nat Genet 41: 1275-81.

[45] Li R, Fan W, et al. (2010). The sequence and de novo assembly of the giant panda genome. Nature 463: 311-7.

[46] Denver DR, Dolan PC, et al. (2009). A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. Proc Natl Acad Sci USA 106: 16310-4.

[47] Xia Q, Guo Y, et al. (2009). Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). Science 326: 433-6.

[48] Levy S, Sutton G, et al. (2007). The diploid genome sequence of an individual human. PLoS Biol 5: e254.

[49] Wheeler DA, Srinivasan M, et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872-6.

[50] Lupski JR, Reid JG, et al. (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med 362: 1181-91.

[51] Pushkarev D, Neff NF, et al. (2009). Single-molecule sequencing of an individual human genome. Nat Biotechnol 27: 847-50.

[52] Wang J, Wang W, et al. (2008). The diploid genome sequence of an Asian individual. Nature 456: 60-5.

[53] Ahn SM, Kim TH, et al. (2009). The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res 19: 1622-9.

[54] Hodges E, Xuan Z, et al. (2007). Genome-wide in situ exon capture for selective resequencing. Nat Genet 39: 1522-7.

[55] Porreca GJ, Zhang K, et al. (2007). Multiplex amplification of large sets of human exons. Nat Methods 4: 931-6.

[56] Turner EH, Ng SB, et al. (2009). Methods for genomic partitioning. Annu Rev Genomics Hum Genet 10: 263-84.

[57] Harismendy O, Ng PC, et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10: R32.

[58] Lin X, Tang W, et al. (2012). Applications of targeted gene capture and next-generation sequencing technologies in studies of human deafness and other genetic disabilities. Hear Res 288: 67-76.

[59] Wang Z, Gerstein M, et al. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63.

[60] Cloonan N, Forrest AR, et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5: 613-9.

[61] Mortazavi A, Williams BA, et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621-8.

[62] Sugarbaker DJ, Richards WG, et al. (2008). Transcriptome sequencing of malignant pleural mesothelioma tumors. Proc Natl Acad Sci USA 105: 3521-6.

[63] Sultan M, Schulz MH, et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321: 956-60.

[64] Nagalakshmi U, Wang Z, et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344-9.

[65] Wilhelm BT, Marguerat S, et al. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453: 1239-43.

[66] Cloonan N, Forrest AR, et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods. 5(7):613-9.

[67] Wang Y, Zeng X., et al. (2012). Exploring the switch grass transcriptome using second-generation sequencing technology. PLoS One 7: e34225.

[68] Brent MR (2008). Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat Rev Genet 9: 62-73.

[69] Nagalakshmi U, Wang Z, et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344-9.

[70] Axtell MJ, Jan C, et al. (2006). A two-hit trigger for siRNA biogenesis in plants. Cell 127: 565-77.

[71] Henderson IR, Zhang X, et al. (2006). Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning. Nat Genet 38: 721-5.

[72] Lu C, Kulkarni K, et al. (2006). MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. Genome Res 16: 1276-88.

[73] Houwing S, Kamminga LM, et al. (2007). A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell 129: 69-82.

[74] Girard A, SachidanandamR, et al. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature 442: 199-202.

[75] Lau NC, Seto AG, et al. (2006). Characterization of the piRNA complex from rat testes. Science 313(5785): 363-7.

[76] Morin RD, O'Connor MD, et al. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells." Genome Res 18: 610-21.

[77] Glazov EA, Cottee PA, et al. (2008). A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach. Genome Res 18: 957-64.

[78] Pang M, Woodward AW, et al. (2009). Genome-wide analysis reveals rapid and dynamic changes in miRNA and siRNA sequence and expression during ovule and fiber development in allotetraploid cotton (*Gossypium hirsutum* L.). Genome Biol 10: R122.

[79] Grunau C, Clark SJ, et al. (2001). Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. Nucleic Acids Res 29: E65-5.

[80] Taylor KH, Kramer RS, et al. (2007). Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing." Cancer Res 67: 8511-8.

[81] CokusSJ, Feng S, et al. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452: 215-9.

[82] Lister R, O'Malley RC, et al. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133: 523-36.

[83] Bormann Chung CA, Boyd VL, et al. (2010). Whole methylome analysis by ultra-deep sequencing using two-base encoding. PLoS One 5: e9320.

[84] Costello JF, Krzywinski M, et al. (2009). A first look at entire human methylomes. Nat Biotechnol 27: 1130-2.

[85] Smith ZD, Gu H, et al. (2009). High-throughput bisulfite sequencing in mammalian genomes. Methods 48: 226-32.

[86] Aparicio O, Geisberg JV, et al. (2005). Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. Curr Protoc Mol Biol, Chapter 21: Unit 21 3.

[87] Ren B, Robert F, et al. (2000). Genome-wide location and function of DNA binding proteins. Science 290: 2306-9.

[88] Johnson DS, Mortazavi A, et al. (2007). "Genome-wide mapping of *in vivo* protein-DNA interactions." Science 316: 1497-502.

[89] Robertson G, Hirst M, et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4: 651-7.

[90] Johnson SM, FJ Tan, et al. (2006). Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. Genome Res 16: 1505-16.

[91] Edwards RA, Rodriguez-Brito B, et al. (2006). Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7: 57.

[92] Turnbaugh PJ, Ley RE, et al. (2007). The human microbiome project. Nature 449: 804-10.

[93] Qin J, R Li, et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59-65.

[94] Hamilton JP, Buell CR (2012). Advances in plant genome sequencing. Plant J 70: 177-90.

[95] Morrell PL, ES Buckler, et al. (2011). Crop genomics: advances and applications. Nat Rev Genet 13: 85-96.

[96] Chen S, Xiang L, et al. (2011). An introduction to the medicinal plant genome project. Front Med 5: 178-84.

[97] Weigel D, Mott R (2009). The 1001 genomes project for *Arabidopsis thaliana*. Genome Biol 10: 107.

[98] Cao J, Schneeberger K, et al. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet 43: 956-63.

[99] Haussler D, O'Brien SJ, et al. (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10000 vertebrate species. J Hered. 100(6):659-74.

[100] Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. (2010) A map of human genome variation from population-scale sequencing. Nature. 467(7319):1061-73.

[101] Abecasis GR, Auton A, et al. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

[102] Jones B (2012). Genomics: personal genome project. Nat Rev Genet 13: 599.

[103] Ball MP, Thakuria JV, et al. (2012). A public resource facilitating clinical use of genomes. Proc Natl Acad Sci USA 109: 11920-7.

[104] Hayden EC (2014) Technology: The $1,000 genome. Nature. 507(7492):294-5.

[105] Pagani I, Liolios K, et al. (2012). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 40: D571-9.

[106] Chain PS, Grafham DV, et al. (2009). Genomics. Genome project standards in a new era of sequencing. Science 326: 236-7.

[107] Yandell M, Ence D (2012). A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13: 329-42.

[108] Fickett JW, Tung CS (1992). Assessment of protein coding measures. Nucleic Acids Res 20: 6441-50.

[109] Fields CA, Soderlund CA (1990). gm: a practical tool for automating DNA sequence analysis. Comput Appl Biosci 6: 263-70.

[110] Fickett JW (1996). Finding genes by computer: the state of the art. Trends Genet 12: 316-20.

[111] Borodovsky M, Rudd KE, et al. (1994). Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. Nucleic Acids Res 22: 4756-67.

[112] Zhang MQ (2002). Computational prediction of eukaryotic protein-coding genes. Nat Rev Genet 3: 698-709.

[113] Kozaks M (1984). Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucleic Acids Res 12: 857-72.

[114] Kozak M (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res 15: 8125-48.

[115] Cavener DR, Ray SC (1991). Eukaryotic start and stop translation sites. Nucleic Acids Res 19: 3185-92.

[116] Nakagawa S, Niimura Y, et al. (2008). Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. Nucleic Acids Res 36: 861-71.

[117] Patel AA, Steitz JA (2003). Splicing double: insights from the second spliceosome. Nat Rev Mol Cell Biol 4: 960-70.

[118] Stormo GD, Schneider TD, et al. (1982). Characterization of translational initiation sites in *E. coli*. Nucleic Acids Res 10: 2971-96.

[119] Pedersen AG, Nielsen H (1997). Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. Proc Int Conf Intell Syst Mol Biol 5: 226-33.

[120] Li H, Jiang T (2005). A class of edit kernels for SVMs to predict translation initiation sites in eukaryotic mRNAs. J Comput Biol 12: 702-18.

[121] Li G (2005). Translation initiation sites prediction with mixture Gaussian models in human cDNA sequences. IEEE Trans Knowl Data Eng 17:1152-1160.

[122] Duret L, Mouchiroud D, et al. (1995). Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J Mol Evol 40: 308-17.

[123] Gish W, States DJ (1993). Identification of protein coding regions by database similarity search. Nat Genet 3: 266-72.

[124] Ansong C, Purvine SO, et al. (2008). Proteogenomics: needs and roles to be filled by proteomics in genome annotation. Brief Funct Genomic Proteomic 7: 50-62.

[125] Tisserant E, Da Silva C, et al. (2011). Deep RNA sequencing improved the structural annotation of the *Tuber melanosporum* transcriptome. New Phytol 189: 883-91.

[126] Mizrachi E, Hefer CA, et al. (2010). De novo assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. BMC Genomics 11: 681.

[127] Coleman, SJ, Zeng Z, et al. (2010). Structural annotation of equine protein-coding genes determined by mRNA sequencing. Anim Genet 41: 121-30.

[128] Martin J, Zhu W, et al. (2010). *Bacillus anthracis* genome organization in light of whole transcriptome sequencing. BMC Bioinformatics 11: S10.

[129] Passalacqua KD, Varadarajan A, et al. (2012). Strand-specific RNA-seq reveals ordered patterns of sense and antisense transcription in *Bacillus anthracis*. PLoS One 7: e43350.

[130] Picardi E, Pesole G (2010). Computational methods for *ab initio* and comparative gene finding. Methods Mol Biol 609: 269-84.

[131] Wang Z, Chen Y, et al. (2004). A brief review of computational gene prediction methods." Genomics Proteomics Bioinformatics 2: 216-21.

[132] Sleator RD (2010). An overview of the current status of eukaryote gene prediction strategies. Gene 461: 1-4.

[133] Krogh A, Mian IS, et al. (1994). A hidden Markov model that finds genes in *E. coli* DNA. Nucleic Acids Res 22: 4768-78.

[134] Stormo GD, Haussler D (1994). Optimally parsing a sequence into different classes based on multiple types of evidence. Proc Int Conf Intell Syst Mol Biol 2: 369-75.

[135] Kulp D, Haussler D, et al. (1996). A generalized hidden Markov model for the recognition of human genes in DNA. Proc Int Conf Intell Syst Mol Biol 4: 134-42.

[136] Burge C, Karlin S (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol 268: 78-94.

[137] Guigó R, Knudsen S, Drake N, Smith T (1992). Prediction of gene structure. J Mol Biol 226(1):141-57.

[138] Korf I (2004). Gene finding in novel genomes. BMC Bioinformatics 5: 59.

[139] Delcher AL, Harmon D, et al. (1999). Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27: 4636-41.

[140] Besemer J, Borodovsky M (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 33: W451-4.

[141] Stanke M, Steinkamp R, et al. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 32: W309-12.

[142] Stanke M, Schoffmann O, et al. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7: 62.

[143] Parra G, Agarwal P, et al. (2003). Comparative gene prediction in human and mouse. Genome Res 13: 108-17.

[144] Yeh RF, Lim LP, et al. (2001). Computational inference of homologous gene structures in the human genome. Genome Res 11: 803-16.

[145] Gross SS, Brent MR (2006). Using multiple alignments to improve gene prediction. J Comput Biol 13: 379-93.

[146] Foissac S, et al. (2008). Genome annotation in plants and fungi: EuGene as a model platform. Curr. Bioinform. 3, 87-97.

[147] Howe KL, Chothia T, Durbin R (2002). GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. Genome Res. 12(9):1418-27.

[148] Allen JE, Salzberg SL (2005). JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics. 21(18): 3596-603.

[149] Rust AG, Mongin E, et al. (2002). Genome annotation techniques: new approaches and challenges. Drug Discov Today 7: S70-6.

[150] Sterck L, Billiau K, et al. (2012). ORCAE: online resource for community annotation of eukaryotes. Nat Methods. 9(11):1041.

[151] Thompson JN (1999). The evolution of species interactions. Science 284: 2116-8.

[152] Jander G, Howe G (2008). Plant Interactions with Arthropod Herbivores: State of the Field Plant Physiol. 146(3): 801–803.

[153] Dinnage R, Cadotte MW, et al. (2012). Diversity of plant evolutionary lineages promotes arthropod diversity. Ecol Lett 15: 1308-17.

[154] Ferrier SM, Bangert RK, et al. (2013). Unique arthropod communities on different host-plant genotypes results in greater arthropod diversity. Arthropod-Plant Interactions 6(2): 187-195.

[155] Hare JD (2012). Ecology. How insect herbivores drive the evolution of plants. Science 338: 50-1.

[156] Guy Smagghe, Isabel Diaz (2012). Arthropod-plant interactions: Novel insights and approaches for IPM.

[157] Mithofer A, Boland W (2008). Recognition of herbivory-associated molecular patterns. Plant Physiol 146: 825-31.

[158] Wu J, Baldwin IT (2010). New insights into plant responses to the attack from insect herbivores. Annu Rev Genet 44: 1-24.

[159] Felton GW, Tumlinson JH (2008). Plant-insect dialogs: complex interactions at the plant-insect interface. Curr Opin Plant Biol 11: 457-63.

[160] Wu J and Baldwin IT (2009). Herbivory-induced signalling in plants: perception and action. Plant Cell Environ 32: 1161-74.

[161] Howe GA, Jander G (2008). Plant immunity to insect herbivores. Annu Rev Plant Biol 59: 41-66.

[162] Browse J, GA Howe (2008). New weapons and a rapid response against insect attack. Plant Physiol 146: 832-8.

[163] Carlini CR, Grossi-de-Sa MF (2002). Plant toxic proteins with insecticidal properties. A review on their potentialities as bioinsecticides. Toxicon 40: 1515-39.

[164] Kessler A, Baldwin IT (2001). Defensive function of herbivore-induced plant volatile emissions in nature. Science 291: 2141-4.

[165] Dicke M, van Loon JJ, et al. (2009). Chemical complexity of volatiles from plants induced by multiple attack. Nat Chem Biol 5: 317-24.

[166] War AR, Sharma HC, et al. (2011). Herbivore induced plant volatiles: their role in plant defense for pest management. Plant Signal Behav 6: 1973-8.

[167] Dudareva N, Negre F, et al. (2006). Plant volatiles: recent advances and future perspectives. Crit Rev Plant Sci25: 417–440.

[168] Jones JD, Dangl JL (2006). The plant immune system. Nature. 444(7117):323-9.

[169] Hammond-Kosack KE, Jones JD (1997). Plant disease resistance genes. Annu Rev Plant Physiol Plant Mol Biol. 48:575-607.

[170] Van der Biezen EA, Jones JD (1998). Plant disease-resistance proteins and the gene-for-gene concept. Trends Biochem Sci. 23(12):454-6.

[171] DeYoung BJ, Innes RW (2006). Plant NBS-LRR proteins in pathogen sensing and host defense. Nat Immunol. 7(12):1243-9.

[172] Kaloshian I (2004). Gene-for-gene disease resistance: bridging insect pest and pathogen defense. J Chem Ecol. 30(12):2419-38.

[173] Vogel JP, Raab TK, Schiff C, Somerville SC (2002). PMR6, a Pectate Lyase–Like Gene Required for Powdery Mildew Susceptibility in Arabidopsis. Plant Cell. 14(9):2095-106.

[174] Lorang JM, Sweat TA, Wolpert TJ. Plant disease susceptibility conferred by a "resistance" gene. Proc Natl Acad Sci U S A. 104(37):14861-6.

[175] Sarmento R, Lemos F, et al. (2011). A herbivore that manipulates plant defense. Ecol Lett. 14(3): 229-236.

[176] Whiteman NK, Jander G (2010). Genome-enabled research on the ecology of plant-insect interactions. Plant Physiol 154: 475-8.

[177] Wise RP, Moscou MJ, et al. (2007). Transcript profiling in host-pathogen interactions. Annu Rev Phytopathol 45: 329-69.

[178] Johnson MTJ (2011). Evolutionary ecology of plant defences against herbivores. Funct Ecol 25: 305–311.

[179] Gao LL, Klingler JP, et al. (2008). Characterization of pea aphid resistance in *Medicago truncatula*. Plant Physiol 146: 996-1009.

[180] Jaquiery J, Stoeckel S, et al. (2012). Genome scans reveal candidate regions involved in the adaptation to host plant in the pea aphid complex. Mol Ecol 21: 5251-64.

[181] Zhurov V, Navarro M, et al. (2014). Reciprocal responses in the interaction between Arabidopsis and the cell-content feeding chelicerate herbivore spider mite. Plant Physiol.164(1):384-99.

[182] Schneeberger K, Ossowski S, et al. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Methods 6: 550-1.

[183] Tunstall NE, Warr CG (2012). Chemical communication in insects: the peripheral odour coding system of *Drosophila melanogaster*. Adv Exp Med Biol. 739:59-77.

[184] Clyne PJ, Warr CG, et al. (1999). A novel family of divergent seven-transmembrane proteins: candidate odorant receptors in *Drosophila*. Neuron 22: 327-38.

[185] Clyne PJ, Warr CG, et al. (2000). Candidate taste receptors in *Drosophila*. Science 287: 1830-4.

[186] Peñalva-Arana DC, Lynch M, et al. (2009). The chemoreceptor genes of the waterflea *Daphnia pulex*: many Grs but no Ors. BMC EvolBiol 9:79.

[187] Buck L, Axel R (1991). A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell 65: 175-87.

[188] Troemel ER, Chou JH, et al. (1995). Divergent seven transmembrane receptors are candidate chemosensory receptors in *C. elegans*. Cell 83: 207-18.

[189] Silbering AF, Benton R (2010). Ionotropic and metabotropic mechanisms in chemoreception: 'chance or design'? EMBO Rep 11: 173-9.

[190] Mitri C, Parmentier ML, et al. (2004). Divergent evolution in metabotropic glutamate receptors. A new receptor activated by an endogenous ligand different from glutamate in insects. J Biol Chem 279: 9313–9320.

[191] Mitri C, et al. (2009). Plant insecticide L-canavanine repels *Drosophila* via the insect orphan GPCR DmX. PLoS Biol 7(6):e1000147.

[192] Benton R, Vannice KS, et al. (2009). Variant ionotropic glutamate receptors as chemosensory receptors in *Drosophila*. *Cell* 136:149-162.

[193] Snyder JC, Guo Z, et al. (1993). 2,3-Dihydrofarnesoic acid, a unique terpene from trichomes of *Lycopersicon hirsutum*, repels spider mites. J Chem Ecol. 19(12):2981-97.

[194] Snyder JC, Antonious GF, Thacker R (2011). A sensitive bioassay for spider mite (*Tetranychus urticae*) repellency: a double bond makes a difference. Exp Appl Acarol. 55(3):215-24.

[195] Alba JM, Montserrat M, Fernández-Muñoz R (2009). Resistance to the two-spotted spider mite (*Tetranychus urticae*) by acylsucroses of wild tomato (*Solanum pimpinellifolium*) trichomes studied in a recombinant inbred line population. Exp Appl Acarol. 47(1):35-47.

[196] Nyalala SO, Petersen MA, Grout BWW (2011). Acetonitrile (methyl cyanide) emitted by the African spider plant (*Gynandropsis gynandra* L. (Briq)): Bioactivity against spider mite (*Tetranychus urticae* Koch) on roses. Sci Hortic 128(3):352-356.

[197] Regev S, Cone WW (1980). The monoterpene citronellol, as a male sex attractant of the two spotted spider mite, *Tetranychus urticae* (Acarina: Tetranychidae). Environ Entomol 9:50-52.

[198] Kappers IF, Hoogerbrugge H (2011). Variation in herbivory-induced volatiles among cucumber (*Cucumis sativus* L.) varieties has consequences for the attraction of carnivorous natural enemies. J Chem Ecol. 37(2):150-60.

[199] Unsicker SB1, Kunert G, Gershenzon J. (2009). Protective perfumes: the role of vegetative volatiles in plant defense against herbivores. Curr Opin Plant Biol. 12(4):479-85.

[200] Sarmento RA, Lemos F, et al. (2011). A herbivorous mite down-regulates plant defence and produces web to exclude competitors. PLoS ONE 6(8): e23757.

[201] Dicke M. (1986). Volatile spider-mite pheromone and host-plant kairomone, involved in spaced-out gregariousness in the spider mite *Tetranychus urticae*. Physiol. Entomol. 11:251–262.

[202] Yano S (2008). Collective and solitary behaviors of two-spotted spider mite (Acari: Tetranychidae) are induced by trail following. Ann Entomol Soc Am 101(1): 247–252.

[203] Le Goff G1, Mailleux AC (2009). Spatial distribution and inbreeding in *Tetranychus urticae*. C R Biol. 332(10):927-33.

[204] Shinmen T, Yano S, Osakabe M (2010). The predatory mite *Neoseiulus womersleyi* (Acari: Phytoseiidae) follows extracts of trails left by the two-spotted spider mite *Tetranychus urticae* (Acari: Tetranychidae). Exp Appl Acarol. 52:111–118.

[205] Edgecombe GD (2010). Arthropod phylogeny: An overview from the perspectives of morphology, molecular data and the fossil record. Arthropod Struct.Dev. 39, 74–87.

[206] Regier JC, et al. (2010). Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463, 1079–1083.

[207] Dunlop JA, Selden PA (2009). Calibrating the chelicerate clock: a paleontological reply to Jeyaprakash and Hoy. Exp. Appl. Acarol. 48, 183–197.

[208] Dunlop JA (2010). Geological history andphylogeny of Chelicerata. Arthropod Struct.Dev. 39, 124–142.

[209] Walter DE, Proctor HC (1999). Mites: Ecology, Evolution and Behaviour (CABI Publishing).

[210] Jeppson LR, Keifer HH, Baker EW (1975). Mites Injurious to Economic Plants (Univ. California Press).

[211] Gerson U (1985). Spider Mites: their Biology, Natural Enemies and Control. Vol. 1A (eds Helle, W, Sabelis MW) 223–232 (Elsevier).

[212] Boubou A, Migeon A, et al. (2011). Recent emergence and worldwide spread of the red tomato spider mite, *Tetranychus evansi*: genetic variation and multiple cryptic invasions. Biol. Invasions 13, 81–92.

[213] Migeon A, et al (2009). Modelling the potential distribution of the invasive tomato red spider mite, *Tetranychus evansi* (Acari: Tetranychidae). Exp. Appl. Acarol. 48, 199–212.

[214] Van Leeuwen T, Vontas J, et al. (2010). Acaricide resistance mechanisms in the two-spotted spider mite *Tetranychus urticae* and other important Acari: a review. Insect Biochem. Mol. Biol. 40, 563–572.

[215] Khajehali J, Van Nieuwenhuyse P, et al. (2011). Acaricide resistance and resistance mechanisms in *Tetranychus urticae* populations from rose greenhouses in the Netherlands. Pest Manag. Sci. 67, 1424–1433.

[216] VectorBase (2008). Ixodes scapularis Wikel annotation, IscaW1 Æhttp://iscapularis.vectorbase.orgæ.

[217] Lau NC, et al. (2009). Abundant primary piRNAs, endo-siRNAs, and microRNAs in a Drosophila ovary cell line. Genome Res. 19, 1776–1785.

[218] Kozomara A, Griffiths-Jones S (2011). miRBase: integrating microRNAannotation and deep-sequencing data. Nucleic Acids Res. 39, D152–D157.

[219] Richards S, et al. (2008). The genome of the model beetle and pest *Tribolium castaneum*. Nature 452, 949–955.

[220] Oliver JH (1977). Cytogenetics of mites and ticks. Annu. Rev. Entomol. 22, 407–429 (1977).

[221] Kirkness EF, et al. (2010). Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc. Natl Acad. Sci. USA 107, 12168–12173.

[222] Altschul SF, Gish W, et al. (1990). Basic local alignment search tool. J Mol Biol. 215(3):403-10.

[223] Gremme G, Brendel V, et al. (2005). Engineering a software tool for gene structure prediction in higher organisms. Information and Software Technology, 47(15):965-978.

[224] Langmead B, Trapnell C, et al. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10(3):R25.

[225] Trapnell C, Pachter L, et al. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 25(9):1105-11.

[226] Cornman SR, et al (2010). Genomic survey of the ectoparasitic mite *Varroa destructor*, a major pest of the honey bee *Apis mellifera*. BMC Genom. 11, 602.

[227] Carrillo L, et al (2011). Expression of a barley cystatin gene in maize enhances resistance against phytophagous mites by altering their cysteine-proteases. Plant Cell Rep. 30, 101–112.

[228] Feyereisen R (2011). Arthropod CYPomes illustrate the tempo and mode in P450 evolution. Biochim. Biophys. Acta 1814, 19–28.

[229] Guengerich FP, Wu ZL, et al. (2005). Function of human cytochrome P450s: characterization of the orphans. Biochem. Biophys. Res. Commun. 338, 465–469.

[230] Giraudo M, Unnithan GC, at el. (2010). Regulation of cytochrome P450 expression in Drosophila: Genomic insights. Pestic. Biochem. Physiol. 97, 115–122.

[231] Vaillancourt FH, Bolin JT, et al. (2006). The ins and outs of ring-cleaving dioxygenases. Crit. Rev. Biochem. Mol. Biol. 41, 241–267.

[232] Moran NA, Jarvik T (2010). Lateral transfer of genes from fungi underlies carotenoid production in aphids. Science 328, 624–627.

[233] Veerman A, Helle W (1978). Evidence for functional involvement of carotenoids in photoperiodic reaction of spider mites. Nature 275, 234.

[234] Chung JS (2010). Hemolymph ecdysteroids during the last three molt cycles of the blue crab, *Callinectes sapidus*: quantitative and qualitative analyses and regulation. Arch. Insect Biochem. Physiol. 73, 1–13.

[235] Grenier JK, Garber TL, et al. (1997). Evolution of the entire arthropod Hox gene set predated the origin and radiation of the onychophoran/arthropod clade. Curr. Biol. 7, 547–553.

[236] Schwager EE, Schoppmeier M, et al. (2007). Duplicated Hox genes in the spider *Cupiennius salei*. Front. Zool. 4, 10.

[237] Damen WGM, Hausdorf M, et al. (1998) .A conserved mode of head segmentation in arthropods revealed by the expression pattern of Hox genes in a spider. Proc. Natl Acad. Sci. USA 95, 10665–10670.

[238] Mothes U, Seitz KA (1981). Fine-structure and function of the prosomal glands of the 2-spotted spider-mite, *Tetranychus urticae* (Acari, Tetranychidae). Cell Tissue Res. 221, 339–349.

[239] Kluge JA, Rabotyagova U, et al. (2008). Spider silks and their applications. Trends Biotechnol. 26, 244–251.

[240] Grbic M, et al. (2007). Mity model: *Tetranychus urticae*, a candidate for chelicerate model organism. Bioessays 29, 489–496.

[241] Degroeve S, Saeys Y, et al. (2005). SpliceMachine: predicting splice sites from high-dimensional local context representations. Bioinformatics 21**,** 1332-1338.

[242] Tweedie S, Ashburner M, et al. (2009). FlyBase: enhancing *Drosophila* Gene Ontology annotations. Nucleic Acids Res. 37(Database issue):D555-9.

[243] Kummerfeld SK, Teichmann SA, (2006). DBD: a transcription factor prediction database. Nucleic Acids Res. 34, D74-81.

[244] Pfreundt U, et al. (2010). FlyTF: improved annotation and enhanced functionality of the *Drosophila* transcription factor database. Nucleic Acids Res. 38, D443-447.

[245] Nei M, Niimura Y, et al. (2008). The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat Rev Genet 9: 951–963.

[246] Cande J, et al. (2012). Smells like evolution: the role of chemoreceptor evolution in behavioral change. Curr Opin Neurobiol http://dx.doi.org/10.1016/

[247] Kaupp UB (2010). Olfactory signalling in vertebrates and insects: differences and commonalities. Nat Rev Neurosci 11:188-200.

[248] Dong D, Jin K, et al. (2012). CRDB: database of chemosensory receptor gene families in vertebrates. PLoS One 7(2):e31540.

[249] Robertson HM, Warr CG, et al. (2003). Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. Proc Natl Acad Sci USA 100:14537-14542.

[250] Gardiner A, Barker D, et al. (2008). *Drosophila* chemoreceptor gene evolution: selection, specialization and genome size. *Mol Ecol* 17:1648–1657.

[251] Croset V, et al. (2010). Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. PLoS Genet 6:e1001064.

[252] Benton R, Sachse S, et al. (2006). A typical membrane topology and heteromeric function of *Drosophila* odorant receptors *in vivo*. PLoS Biol 4:0240–0257.

[253] Grbic M, et al. (2011). The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature* 479:487-492.

[254] Ben-Sahar Y (2011). Sensory Functions for Degenerin/Epithelial Sodium Channels (DEG/ENaC). Adv Genet 76:1–26.

[255] Vieira FG, Rozas J (2011). Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. Genome Biol Evol 3:476-90.

[256] Kellenberger S, Schild L (2002). Epithelial sodium channel/degenerin family of ion channels: a variety of functions for a shared structure. Physiol Rev. 82:735-67.

[257] Jasti J, Furukawa H, et al. (2007). Structure of acid-sensing ion channel 1 at 1.9 A resolution and low pH. Nature 449:316-23.

[258] Tavernarakis N, Shreffler W, et al. (1997). unc-8, a DEG/ENaC family member, encodes a subunit of a candidate mechanically gated channel that modulates C. elegans locomotion. Neuron 18:107-119.

[259] Wagner EG, Simons RW (1994). Antisense RNA control in bacteria, phages, and plasmids. Annu Rev Microbiol. 48:713-42.

[260] Katayama S, Tomaru Y, et al. (2005). Antisense transcription in the mammalian transcriptome. Science. 309(5740):1564-6.

[261] David L, Huber W, et al. (2006). A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci USA. 103(14):5320-5.

[262] Pelechano V, Steinmetz LM. Gene regulation by antisense transcription. Nat Rev Genet. 14(12):880-93.

[263] Sánchez-Gracia A, Vieira FG, et al. (2009). Molecular evolution of the major chemosensory gene families in insects. *Heredity* 103:208-216.

[264] Wisotsky Z, Medina A, et al. (2011). Evolutionary differences in food preference rely on gr64e, a receptor for glycerol. *Nat Neurosci* 14:1534-1541.

[265] Zhao H, et al. (2010). Evolution of the sweet taste receptor gene Tas1r2 in bats. Mol Biol. Evol 27:2642–2650.

[266] Jiang P, et al. (2012). Major taste loss in carnivorous mammals. Proc Natl Acad Sci USA 109:4956–4961.

[267] Abeel T, Van Parys T, et al. (2012). GenomeView: a next-generation genome browser. Nucleic Acids Res 40:e12.

[268] Finn RD, et al. (2010). The Pfam protein families database. Nucleic Acids Res 38: D211-222.

[269] Eddy SR (2009). A new generation of homology search tools based on probabilistic inference. Genome Inform. 23:205-211.

[270] Larkin MA, et al. (2007). ClustalW and ClustalX version 2.0. Bioinformatics 23:2947-2848.

[271] Krogh A, Larsson B, et al. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305(3):567-80.

[272] Waterhouse AM, Procter JB, et al. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9):1189-91.

[273] Schmidt HA, Strimmer K, et al. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18(3):502-4.

[274] Swofford DL (2002). PAUP* 4.0: Phylogenetic Analysis Using Parsimony. Sinauer Associates.

[275] Do CB, Mahabhashyam MS, et al. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res 15(2):330-40.

[276] Taniura H, Sanada N, et al. (2006). A metabotropic glutamate receptor family gene in *Dictyostelium discoideum*. J Biol Chem 281(18):12336-43.

[277] Guindon S, Dufayard JF, et al. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59(3):307-21.

[278] Matsuo T (2008). Genes for Host-Plant Selection in Drosophila. J Neurogenet. 22(3):195-210.

[279] Dworkin I, Jones CD (2009). Genetic Changes Accompanying the Evolution of Host Specialization in *Drosophila sechellia*. Genetics. 181(2):721-36.

[280] Alain MIGEON, Franck DORKELD (2006-2013). Spider Mites Web: a comprehensive database for the Tetranychidae. http://www.montpellier.inra.fr /CBGP/spmweb.

[281] Navajas M, de Moraes GJ, et al. (2013). Review of the invasion of *Tetranychus evansi*: biology, colonization pathways, potential expansion and prospects for biological control. Exp Appl Acarol. 59(1-2):43-65.

[282] Hill RL, O'Donnell DJ (1991) The host range of *Tetranychus lintearius* (Acarina: Tetranychidae). Exp. Appl. Acarol., 11, 253–69.

[283] Boetzer M, Henkel CV, et al. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 27(4):578-9.

[284] Nadalin F, Vezzi F, Policriti A (2012). GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinformatics. 13 Suppl 14:S8.

[285] Mardis ER (2013). Next-generation sequencing platforms.Annu Rev Anal Chem (Palo Alto Calif).6:287-303.

[286] Ozsolak F, Milos PM (2011). RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. 12(2):87-98.

[287] Denoeud F, Aury JM, et al. (2008). Annotating genomes with massive-scale RNA sequencing. Genome Biol. 9(12):R175.

[288] Stein L. (2001). Genome annotation: from sequence to biology. Nat Rev Genet. 2(7):493-503.

[289] Ellegren H (2014). Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol.29(1):51-63.

[290] Altshuler D, Durbin RM, et al. (2012). An integrated map of genetic variation from 1092 human genomes. Nature. 491(7422):56-65.

[291] Evans JD, Brown SJ, et al (2012). The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment. J Hered.104(5):595-600.

[292] Cao Z, Yu Y, et al. (2013). The genome of *Mesobuthus martensii* reveal a uniqua adaptation model of arthropods. Nat Commun.4:2602.

[293] El-Metwally S, Hamza T, Zakaria M, Helmy M (2013). Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges. PloS Comput Biol. 9(12):e1003345.

[294] Patel RK, Jain M (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One. 7(2):e30619.

[295] Medvedev P, Stanciu M, Brudno M (2009). Computational methods for discovering structural variation with next-generation sequencing. Nat Methods. 6(11 Suppl):S13-20.

[296] Pavlopoulos GA, Oulas A, Iacucci E, Sifrim A, Moreau Y (2013). Unraveling genomic variation from next generation sequencing data. BioData Min. 6(1):13.

[297] Tunstall NE, Warr CG (2012). Chemical communication in insects: the peripheral odour coding system of Drosophila melanogaster. Adv Exp Med Biol. 2012;739:59-77.

[298] van der Goes van Naters W, Carlson JR. Insects as chemosensors of humans and crops. Nature. 444(7117):302-7.