Western Graduate&PostdoctoralStudies

Western University

Scholarship@Western

Electronic Thesis and Dissertation Repository

9-22-2017 11:00 AM

Bioinformatics and Next Generation Sequencing: Applications of Arthropod Genomes

Zaichao Zhang, *The University of Western Ontario*

Supervisor: Miodrag Grbic, *The University of Western Ontario*
Joint Supervisor: Yves Van de Peer (Ghent University), *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Biology
© Zaichao Zhang 2017

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Bioinformatics Commons

Recommended Citation

Zhang, Zaichao, "Bioinformatics and Next Generation Sequencing: Applications of Arthropod Genomes" (2017). *Electronic Thesis and Dissertation Repository*. 4920.
https://ir.lib.uwo.ca/etd/4920

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

Abstract

Over the past decade, the Next Generation Sequencing (NGS) technology has been broadly applied in many areas such as genomics, medical diagnosis, biotechnology, virology, biological systematics, forensic biology, and anthropology. Taken together, it has offered us brilliant insights into life sciences. Most of the work presented in this thesis describes NGS applications on genome assembly, genome annotation, and comparative genomics, using arthropods as case studies: (1) by sequencing and analyzing the genomes of three *Tetranychus* spider mites with three completely different feeding behaviors, we uncovered genomic signature variations and indicative of pest adaptations; (2) we sequenced, assembled and annotated five *Brevipalpus* flat mite genomes and their corresponding endosymbiont *Cardinium* genomes. Comparative genomics reveals herbivorous pest adaptations and parthenogenesis; (3) the complete genomic analysis of parasitoid wasp *Copidosoma floridanum* indicates the mechanism of polyembryony of such primary parasite of moths. By bioinformatics and genomics approaches, my study provides the genomic basis and establishes the hypotheses for the future biology in pest and arthropod researches. These NGS applications of arthropod genomes will offer new insights into arthropod evolution and plant-herbivore interactions, open unique opportunities to develop novel plant protection strategies, and additionally, provide arthropod genomic resources as well.

Keywords

Next Generation Sequencing (NGS), Genome Assembly, Genome Annotation, Arthropod, Spider Mite, Flat Mite, Wasp, Optical map, F-box, DNMT

Samevatting

In de afgelopen tien jaar is de Volgende Generatie Sequencing een essentiële applicatie geworden in vele gebieden, zoals genomica, medische diagnose, biotechnologie, virologie, biologische systematica, forensische biologie en antropologie en bood ons briljante inzichten in de biowetenschappen. Veel van het werk dat in dit proefschrift wordt gepresenteerd, beschrijft genoomsamenstelling, genoom annotatie en vergelijkende genomica waarbij artropoden als casestudies worden gebruikt: (1) door de genen van drie *Tetranychus* spinmijten te sequentiëren en analyseren Die drie volledig verschillende voedingsgedrag vertegenwoordigen, het onthult de genomische handtekeningvariaties en belangrijke agrarische plaagaanpassingen; (2) wij sequenced, assembled en annotated vijf *Brevipalpus* platte mijt genomen hun overeenkomstige endosymbiont *Cardinium* genomen. Vergelijkende genomica onthult herbivore pestaanpassingen en parthenogenese; (3) de volledige genomische analyse van parasitoïde wesp *Copidosoma floridanum* duidt op het mechanisme van polyembryonie van deze primaire parasitoïde van motten. Door de benaderingen van bioinformatica en genomica zouden deze NGS-applicaties op artropoden-genen nieuwe inzichten bieden in arthropod-evolutie en plantaardige herbivoorinteracties, unieke mogelijkheden bieden om nieuwe plantenbeschermingsstrategieën te ontwikkelen en arthropod genoom middelen te geven.

Trefwoorden

Volgende Generatie Sequencing, Genoom Assemblage, Genoom Annotatie, Geleedpotige, Spinmijt, Platte Mijt, Wesp, Optische kaart, F-doos, DNMT

Co-Authorship Statement

Chapter 1 and 8: Introduction, summary, and perspectives

Zaichao Zhang: wrote the manuscript.

Yves Van de Peer and Miodrag Grbić: revised the manuscript.

Chapter 2: The NGS toolkit and pipeline

Zaichao Zhang: summarized and optimized the typical workflow of NGS and wrote the manuscript.

Stephane Rombauts and Vojislava Grbić: revised the manuscript.

Chapter 3: Update *T. urticae* genome

Zaichao Zhang: updated *T. urticae* assembly using optical mapping data *in silico* and re-annotated the genome using multiple complementary hybrid data; analyzed the data, and wrote the manuscript.

Vladimir Zhurov, Stephane Rombauts, Vojislava Grbić, Yves Van de Peer and Miodrag Grbić: collected samples and performed the optical mapping experiment.

Chapter 4: The three spider mite genomes project

Zaichao Zhang: revised, modified and improved the three genome annotation data; analyzed the data including genome assembly validation and synteny, screening contaminated scaffolds, clustering gene families, analyzing transposable elements; drafted the manuscripts.

Stephane Rombauts, Vojislava Grbić, Vladimir Zhurov, Yves Van de Peer and Miodrag Grbić: developed the experimental design, assembled and annotated the genomes.

Toni Gabaldon: constructed the phylogenetic trees for the three genomes.

Chapter 5: The Novel F-box Genes in *Tetranychus*

Zaichao Zhang: discovered the novel F-box gene family in *Tetranychus* and manually curated these genes, designed the experiment, analyzed all the data *in silico* and wrote the manuscript.

Pengyu Jin, Vojislava Grbić, Vladimir Zhurov, Yves Van de Peer and Miodrag Grbić: designed and performed the experiment *in situ* and provided input for the bioinformatics analyses.

Chapter 6: *Brevipalpus* genomes and their endosymbiont *Cardinium* genomes

Zaichao Zhang: participated in the experimental design and analyzed the data on *Brevipalpus* genomes, especially on screening contamination, *Cardinium* genome assembly, the *Brevipalpus* genome annotation, SNP calling analysis, genome visualization on ORCAE and comparative genomics, and drafted the manuscript.

Phuong Le, Yao-Cheng Lin, and Yves Van de Peer: developed the experimental design, assembled *Brevipalpus* genomes, annotated *Cardinium* genomes, constructed phylogenetic trees, and co-drafted the manuscript.

Thomas Van Leeuwen and Hans Breeuwer: provided *Brevipalpus yothersi* (Amsterdam strain), *B. californicus*, *B. papayensis* sample data; suggestions on the experimental design and analyzing the reproduction asexuality of mites and their symbionts.

Denise Navia, Juliana Freitas-Astua, and Priscila Grynberg: provided *Brevipalpus yothersi* (Brazilian strain) samples and input for developing the experimental design and analyzing the data.

Chapter 7: The *Copidosoma* genome project

Zaichao Zhang: annotated *Copidosoma floridanum* genome focusing on DNMT gene families, identified and manually revised the DNMT1, DNMT2, and DNMT3 families, discovered and annotated DNMT3-like genes, and provided input for the manuscript.

Vladimir Zhurov, Vojislava Grbić, Stephane Rombauts, Tyler Alioto, Simon Heath, Toni Gabaldon, Stephan Ossowski, Paolo Ribeca, Richard Clark, Roderic Guigó, Yves Van de Peer and Miodrag Grbić: developed the experimental design, collected samples and data, assembled and annotated the genome, DNA methylation transcriptome profiling analysis, methylation analysis, phylogeny studies and drafted the manuscript.

Acknowledgments

This dissertation is dedicated to all whose love, encouragement and support made it possible.

To Miodrag Grbić and Yves Van de Peer, thank you for guiding me the way to be a scientist in Northern America and Europe, for bucking me up whenever necessary, for allowing me in different genome projects, and for offering me the possibilities and opportunities to explore the West World.

To Stephane Rombauts, thank you for being a great advisor and intimate mentor, academically and socially. Because of you, so many of us are blessed still, not only in research but also in Belgian life. Thank you for being my no-matter-what over the years. When I look back at all my best in Belgium, inside all of it, I see you.

To Vojislava Grbić and Yao-Cheng Lin, thank you for all the advice and guidance through my Ph.D. studies. Vava, your patience and gentleness cheered me up when I was down. Yao-Cheng, you strengthened my oriental values and guided me to an efficient way to switch between diverse cultures and thinking perspectives.

To David Smith, Shiva Singh, Pierre Rouzé, and Wannes Dermauw, thank you for advising me and letting me learn how to change negative attitude to positive happiness in doing research; To Toni Gabaldon and Stephan Ossowski, thank you for offering me the amazing internship at CRG in Barcelona; To my colleagues across different project consortiums, thank you for all the assistance and feedback!

To friends and colleagues, without your support, I could not make this far. Vlad, thanks for being like a big brother to me and again for all the help; Nico, you showed me French romance and humour as a handsome and sunshine boy; Kristie, my dear teacher, thanks for improving my accent and correcting my writings; Bilijana, your motherly-like love offered us a sense of peace and security in this exotic land; Maria, my beloved Spanish girl, your smile always echoes in my heart and without your encouragement, I could not finish *Camino de Santigo de Compostela*; Peng, thanks for the inspiring scientific ideas; Tara, thank you again for all the advice when I joined this group; Hooman Jan and Golnaz, thanks for sharing good lab/lunch/beer-time; Yanju and the Xis family, thanks for being my family here in London; Arniban, thanks for all the talks during our Tim Hortons' time; Xi, thanks for being my Chinese pal at Western; Arzie, Carol, Diane, and Sophie, thank you again for all the administrative assistance; The Brazles, Luk and Holy, thanks for having me as a family member in Belgium; Yue-Chen, thanks for your fancy ideas; Shanshuo and Jiejie, thanks for the great time we had in Gent; Yun-Shu, I appreciate you being my lucky sister in Belgium. Because of you, my personal life was brightly colored; Shu-min, you have a brilliant IT mind and thanks for all the help in Linux; Shubada and Sri, thanks for releasing my '*little stressful imp*' during tea-breaks at VIB; Lieven, thank you for the ORCAE database and beverage platforms that benefited us so much in the lab; Oren, thank you for being my Chair at UGent and all the encouragement during my hard times.

To VIB, thank you for the best practice of research facilities, great working environments, amazing seminars and workshops as well as cool IT genius for high-performance computational support; You've indeed achieved '*From science to value for society*'; To UGent, through 200 years of historical suffering and growing, you are now full of mature and you motivated me to the vision of '*Dare to Think*'; To Western, you are not only having the most beautiful campus but the deepest comprehension of '*Veritas et Utilitas*'. Thank you for offering me this opportunity to live and work in this great country.

To my family, for the years behind us, for the years ahead of us, with love side-by-side; especially to my grandfather, I know through eight years of witness, you would be so proud when you look at your grandson with amazing grace from the peaceful paradise.

Finally, to everyone whose bright smile and friendly assistance made this journey possible! Highly appreciated and thanks again! THANK YOU!

This thesis was supported by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-046) and Ontario Research Fund - Global Leadership in Genomics and Life Sciences (GL2-01-035) and (ORF RE8-067), Ruth Horner Arnold Fellowship, Western Graduate Research and Teaching Fellowship, Mitacs Global Link Research Award (IT04121) and Dual-PhD degree BOF Funding (01SF0314) from Ghent University.

Contents

Abstract i

Samevatting ii

Co-Authorship Statement iii

Acknowledgments vi

Contents viii

List of Tables xv

List of Figures xvii

List of Plates xxi

List of Abbreviations xxii

List of Appendices xxvi

Preface xxvii

Chapter 1 1

1 Introduction 1

1.1 NGS and Bioinformatics 2

1.2 Genome assembly and genome annotation 7

1.3 Comparative genomics 8

1.4 Arthropods 10

1.5 The *Tetranychus* spider mites 12

1.6 The *Brevipalpus* flat mites 16

1.7 The wasp *Copidosoma floridanum* 21

1.8 References 23

Chapter 2 32

2 A beginner's guide to NGS genome projects: from genome sequencing, assembly to annotation 32

2.1 Project investigation 34
2.1.1 Project objectives 34
2.1.2 Species survey and complexity 34
2.1.3 *De novo* sequencing and resequencing 37
2.1.4 Expertise and facility 37
2.2 Genome sequencing 38
2.2.1 Sample collection and DNA extraction 38
2.2.2 Library preparation 38
2.2.3 Sample fragmentation 40
2.2.4 Adapters addition 41
2.2.5 Size selection and PCR 41
2.3 Sequencing technologies 41
2.4 Genome assembly 45
2.4.1 Quality Control (QC) 45
2.4.2 Assembly algorithms 45
2.4.3 Scaffolding and gapfilling 51
2.4.4 Resequencing assembly 52
2.4.5 Contaminants identification 52
2.4.6 Assembly quality assessment 55
2.5 Genome annotation 57
2.5.1 Genome masking 59
2.5.2 Annotation for protein-coding genes 63
2.5.3 Structural prediction 63
2.5.4 Functional prediction 67
2.5.5 Inspection and modification 67

2.5.6 Annotation quality assessment 68
2.5.7 Annotation for ncRNA 70
2.5.8 Annotation for pseudogenes 70
2.5.9 Genomic statistics 71
2.5.10 Genome visualization, maintenance, and update 71
2.6 Perspective 72
2.7 Supplementary Links 74
2.8 Reference 75
Chapter 3 87
3 Improvement of the *Tetranychus urticae* genome using optical mapping and cumulative hybrid data 87
3.1 Background 89
3.2 Data description 90
3.2.1 Update genome assembly 90
3.2.2 Assess the OM assembly 93
3.2.3 Re-annotation by cumulative evidence 94
3.3 Discussion 98
3.4 Materials and Methods 99
3.4.1 OM data and re-assembly 99
3.4.2 Re-assembly assessment 100
3.4.3 Re-annotation 100
3.4.4 Re-annotation assessment 100
3.5 Reference 101
Chapter 4 105
4 Comparative genomics of three spider mites reveals genome evolution and genomic signatures of adaptation to different feeding modes 105

4.1 Introduction 106

4.2 Results and Discussions 108

4.2.1 Genomic statistics of the three spider mites 108

4.2.2 A recent divergence of the three mites 111

4.2.3 Feeding and detoxification 113

4.2.4 Protein-binding and transmembrane genes 115

4.2.5 Gene families expanded in *T. lintearius* 116

4.2.6 TE expansion 119

4.3 Conclusion 123

4.4 Materials and Methods 124

4.4.1 Strain selection and DNA preparation. 124

4.4.2 Genome sequencing and assembly 124

4.4.3 Removal of contaminated scaffolds 124

4.4.4 Assembly assessment for the genome completeness 125

4.4.5 Synteny analysis and visualization 126

4.4.6 Structural annotation 126

4.4.7 Functional annotation 127

4.4.8 ncRNA annotation 127

4.4.9 TE annotation, visualization, and statistics 127

4.4.10 Orthologous gene identification 128

4.5 Supplementary Information 130

4.6 Reference 134

Chapter 5 140

5 Evolutionary dynamics of a massively expanded novel F-box gene family in polyphagous pest *Tetranychus urticae* 140

5.1 Introduction 141

5.2 Results and Discussion 142

5.2.1 Conventional F-box genes are conserved 142

5.2.2 The novelty and assessment of NFB genes in *Tetranychus*.................... 144

5.2.3 The extreme expansion of NFB genes in *T. urticae* 149

5.2.4 Structure and Domain Organization of NFB 151

5.2.5 NFB expanded by tandem duplication.. 153

5.2.6 Negative selection 156

5.2.7 TE might be a motivation of expansion 158

5.2.8 Transcriptome profiling 158

5.2.9 Co-expression network of NFB genes with SKP1 160

5.2.10 Low GC-content.. 160

5.3 Summary 162

5.4 Materials and Methods 162

5.4.1 NFB gene discovery and annotation 162

5.4.2 Alignment and phylogeny analyses 163

5.4.3 Transcriptome profiling and network analysis 163

5.4.4 Evolutionary selection pressure analysis 164

5.4.5 Transposable element dynamics 164

5.4.6 Supplementary data.. 165

5.5 Reference 166

Chapter 6 170

6 The genomes of *Brevipalpus* flat mites and their *Cardinium* endosymbionts offer insights into herbivorous pest adaptation and parthenogenesis.. 170

6.1 Introduction.. 171

6.2 Results and discussion 173

6.2.1 Genome statistics of the *Brevipalpus* stains... 173

6.2.2 Phylogeny shows a recent divergence of the five genomes.................... 178

6.2.3 Linkage-specific gene families 181

6.2.4 Genomes statistics of the *Cardinium* stains 183

6.2.5 Phylogeny of *Cardinium* strains 187

6.2.6 Genomic database visualization.. 191

6.3 Conclusion 199

6.4 Materials and Methods.. 199

6.4.1 *Brevipalpus* genomes 199

6.4.2 *Cardinium* genomes 208

6.5 Supplementary Information 211

6.6 Reference 214

Chapter 7 220

7 Genome annotation of DNMT gene families in parasitoid wasp *Copidosoma floridanum*....220

7.1 Introduction.. 221

7.2 Results and discussions.. 222

7.2.1 Genomic statistics of *C. floridanum* 222

7.2.2 Genome annotation of the DNMT gene families.................................... 224

7.2.3 Evolutionary analysis of DNMT gene families 226

7.2.4 DNMT1.. 229

7.2.5 DNMT2.. 232

7.2.6 DNMT3 and DNMT3-like.. 235

7.3 Summary... 239

7.4 Materials and Methods... 239

7.4.1 Genome sequencing and assembly 239

7.4.2 Genome annotation and DNMT gene identification................................ 240

7.4.3 Phylogeny analysis 240
7.5 Supplementary Information 241
7.6 Reference 242
Chapter 8 245
8 Conclusion and perspectives 245
8.1 NGS: opportunities and challenges 246
8.2 TGS: the next NGS? 247
8.3 Arthropod genomics 252
8.4 Perspectives 252
8.5 Reference 253
Appendices 258
Links 259
Supplementary Tables 260
Supplementary Figures 269
Supplementary Protocol 270
NGS Term Box 273
Curriculum Vitae 275

List of Tables

Table 1: Biological comparison between spider mites and flat mites. 20

Table 2: An overview of primary sequencing technologies. 43

Table 3: Tools for genome assembly. 49

Table 4: Tools for repeats identification and genome masking. 60

Table 5: Main tools for genome annotation. 65

Table 6: Statistics categories for genome assembly and annotation. 73

Table 7: OM results and final assembly. 92

Table 8: Statistics of annotation improvement. 96

Table 9: Genomic statistics of the three genomes. 110

Table 10: Key expanded gene families across the three genomes. 114

Table 11: TE distribution across the three genomes. 120

Table 12: TE statistics in the flanking region of 10 kb of key gene families. 122

Table 13: TE statistics in the flanking region of 5 kb of key gene families. 131

Table 14: TE statistics in the flanking region of 15 kb of key gene families. 132

Table 15: TE statistics in the flanking region of 20 kb of key gene families. 133

Table 16: Evidence for confirming NFB genes. 145

Table 17: GC-content in the genome of *T. urticae*. 161

Table 18: The 42 NFB genes highly correlated to SKP1 165

Table 19: Genomic statistics of the *Brevipalpus* strains. 174

Table 20: TE distribution across five flat mite genomes. 176

Table 21: SNP calling using GATK with *B. yothersi* (Brazilian strain) as a reference. 177

Table 22: The raw OrthoMCL results of gene families. 182

Table 23: Genomic features of *Cardinium* strains. 184

Table 24: Sequencing data of the five *Brevipalpus* genomes. 200

Table 25: The assessment of *Cardinium* genome assemblies. 211

Table 26: Assembly and annotation statistics of *C. floridanum* genome. 223

Table 27: DNMT comparison between VIB and CRG versions 225

Table 28: The evolutionary pressure analysis of DNMT3-like genes. 237

Table 29: DNMT annotation note for both VIB and CRG versions. 241

Table 30: Characteristics of TGS technologies and three mapping platforms. 249

Table 31: A general comparison between NGS and TGS. 251

Table 32: Published animal genomes. 260

Table 33: The arthropod genome datasets used in this thesis. 267

List of Figures

Figure 1: The history of sequencing technologies and representative genomes. 3

Figure 2: Statistics of sequencing cost and data. 4

Figure 3: Major NGS and TGS sequencing technologies. 6

Figure 4: Eukaryotic gene structure and gene transcription processes. 9

Figure 5: Phylogeny of arthropods and evolutionary time. 11

Figure 6: The spider mite is feeding the content of leaf cells using its stylet. 13

Figure 7: An overview of the three spider mites. 15

Figure 8: The body shape of flat mite. 19

Figure 9: The life cycle and development of *C. floridanum* in its host *Trichoplusia ni.* 22

Figure 10: A typical NGS workflow from sample collection, assembly to annotation. 33

Figure 11: A workflow of sample library preparation. 39

Figure 12: Four assembly approaches in NGS. 48

Figure 13: Four types of libraries including SE, PE, MP, and TGS reads. 54

Figure 14: Genome assembly assessment using coverage and long reads. 58

Figure 15: A comprehensive workflow for a *de novo* genome masking. 62

Figure 16: EUGENE combines comprehensive evidence for genome annotation. 66

Figure 17: Common structural prediction errors. 69

Figure 18: The circos overview of updated genome of *T. urticae* 88

Figure 19: The OM results from BioNano and OpGen for the *T. urticae* assembly. 91

Figure 20: Assessment methods for superscaffolds and primary gene model errors. 95

Figure 21: Newly predicted genes in *T. urticae* genome. 97

Figure 22: An overview of comparative analysis results of the three genomes. 109

Figure 23: Phylogenetic analysis shows a recent divergence of the three genomes. 112

Figure 24: Heatmap of gene family number across the three genomes. 118

Figure 25: HCSP transmembrane structure of tetur09g07110. 130

Figure 26: An overview of all F-box genes. 143

Figure 27: Genomic synteny of *T. urticae* scaffold 1. 147

Figure 28: Genomic reads coverages of *T. lintearius* and *T. evansi* mapped to *T. urticae*. 148

Figure 29: Phylogenetic tree of NFB genes across the three genomes. 150

Figure 30: The conserved structure and domains of NFB genes in cluster 5. 152

Figure 31: NFB clusters in *T. urticae* genome on the top three scaffolds. 154

Figure 32: Estimates of evolutionary divergence between NFB sequences. 155

Figure 33: Evolutionary pressure analysis of NFB genes in *T. urticae*. 157

Figure 34: Phylogeny of NFB genes and their expression profiles. 159

Figure 35: The relationship of the five flat mites using SNP calling results. 179

Figure 36: Phylogeny of 31 arthropod genomes using single copy gene dataset. 180

Figure 37: The Venn graph indicates their conserveness at the gene level. 185

Figure 38: The synteny of *Cardinium* strains in five *Brevipalpus* genomes. 186

Figure 39: Phylogeny of 39 single copy genes from 84 *Bacteroidete* genomes. 188

Figure 40: Phylogeny of *Cardinium* strains using single copy genes. 189

Figure 41: The phylogenetic tree of different *Cardinium* strains of gyrB genes. 190

Figure 42: ORCAE interface for *Brevipalpus* genomes. 192

Figure 43: Welcome page for *B. yothersi* (Brazilian strain) on ORCAE. 193

Figure 44: Gene locus and functional description on ORCAE. 194

Figure 45: Expression profile, gene ontology, and domain information. 195

Figure 46: Protein homology and alignment in various databases. 196

Figure 47: Gene structure, CDS, and protein sequences. 197

Figure 48: GenomeView offers various annotated information. 198

Figure 49: Self-mapping to validate assembly and screen bacterial contamination. 203

Figure 50: The pipeline of Splice Machine and training EUGENE parameters. 205

Figure 51: SNP calling pipeline using GATK method across the five flat mite strains. 207

Figure 52: Phylogenies of the five genomes using 5k random SNP sites. 212

Figure 53: Divergent time estimation. 213

Figure 54: Phylogenetic distribution of DNA methylation in insects. 227

Figure 55: Prevalence of the DNMT family across taxa. 228

Figure 56: The protein sequence alignment of DNMT1 genes in *C. floridanum*. 230

Figure 57: Phylogenetic tree of DNMT1 genes across arthropod genomes. 231

Figure 58: The mutation point of DNMT2 genes in *C. floridanum*. 233

Figure 59: Phylogenetic tree of the two DNMT2 genes in arthropod context............................. 234

Figure 60: The phylogenetic tree of DNMT3 and the expansion of DNMT3-like genes............ 236

List of Plates

NA

List of Abbreviations

3C	Chromosome Conformation Capture
AAEL	*Aedes aegypti*
ABC	ATP-biding-cassette transporters
ACEP	*Atta cephalotes*
ACYPI	*Acyrthosiphon pisum*
ADAC	*Anopheles darlingi*
AGAP	*Anopheles gambiae*
AMA1	Apical Membrane Antigen 1
BAC	Bacterial Artifical Chromosome
BGIBMGA	*Bombyx mori*
BIG	Beijing Institute of Genomics
BioNano	BioNano Genomics Lnc.
breci	*Brevipalpus californicus* infected
brecu	*Brevipalpus californicus* uninfected
brepa	*Brevipalpus papayensis*
breya	*Brevipalpus yothersi* Amsterdam
breyb	*Brevipalpus yothersi* Brazil
BUSCO	Benchmarking Universal Single-Copy Orthologs
CAS	Chinese Academy of Sciences
cBcal1	*Cardinium* strain in *Brevipalpus californius* infected
cBcal2	*Cardinium* strain in *Brevipalpus californius* uninfected
cBpap1	*Cardinium* strain in *Brevipalpus papayensis*
cByotA1	*Cardinium* strain in *Brevipalpus yothersi* (Amsterdam)
cByotB1	*Cardinium* strain in *Brevipalpus yothersi* (Brazil)
CCE	Carboxyl/cholinesterases
CDNA	Coding DNA
CDS	Coding Sequence
CEGMA	Core Eukaryotic Genes Mapping Approach
CFB	Conventional F-box
CI	Cytoplasmic Incompatibility
CPIJ	*Culex quinquefasciatus*
CSM	*Stegodyphus mimosarum*
CTR	Chain Termination Reaction
DBG	De Bruijn Graph
DNA	Deoxyribonucleic Acid
DNMT	DNA-methyltransferase
DNMT1	DNA-methyltransferase 1
DNMT2	DNA-methyltransferase 2
DNMT3	DNA-methyltransferase 3

dsDNA	double strand DNA
dUTPase	dUTP diphosphatase
eDNA	extrachromosomal DNA
EFX	*Daphnia pulex*
EHJ	*Danaus plexippus*
ENN	*Dendroctonus ponderosae*
EST	Expressed Sequence Tag
FAANG	Functional Annotation of Animal Genomes
FBpp	*Drosophila melanogaster*
GB	*Apis mellifera*
GDL	D-glucono-1,5-lactone
gDNA	chromosomal DNA
GF	Gene Family
GO	Gene Ontology
GSH	Glutathione
GST	Glutathione-S-transferases
HCSP	Hypothetical Cell Surface Protein
HEBUST	Hebei University of Science and Technology
HGT	Horizontal Gene Transfer
HMEL	Heliconius melpomene
HMM	Hidden Markov Model
HNV	*Nasonia vitripennis*
HSINV	*Solenopsis invicta*
IAP	Inhibitor of Apoptosis Protein
INDEL	Insertion and Deletion
IGV	Integrative Genomics Viewer
ISCW	*Ixodes scapularis*
LBA	Long Branch Attraction
lncRNA	long non-coding RNA
LP	*Limulus polyphemus*
LRR	Leucine-Rich Repeat
miRNA	microRNA
ML	Maximum Likelihood
MMa	*Mesobuthus martensii*
MP	Mate Pair reads
mRNA	messager RNA
MSMAR	*Strigamia maritima*
MYA	Million Years Ago
NA	Not Applicable
NCBI	National Center for Biotechnology
NFB	Novel F-box

NGS	Next Generation Sequencing
nHdt	*Hypsibius dujardini*
NJ	Neighbor-Joining Method
OLC	Overall Layout Consensus
OM	Optical Mapping
ONT	Oxford Nanopore Technology
OpGen	OpGen lnc.
ORCAE	Online Resource for Community Annotation of Eukaryotes
PacBio	Pacific Biosciences
PCA	Principal Component Analysis
PE	Pair End reads
PHUM	*Pediculus humanus*
piRNA	piwi-interacting RNA
PPI	Protein-Protein Interaction
PSB	Center of Plant Systems Biology
Px	*Plutella xylostella*
QC	Quality Control
RNA	Ribonucleic Acid
RNAi	RNA interference
SBL	Sequencing by Ligation
SBS	Sequencing by Synthesis
SCF	SKP1-Cullin1-F-box protein
SE	Single End reads
SKP1	S-phase kinase-associated protein 1
SMRT	Single Molecule Real Time
snoRNA	small nucleolar RNA
SNP	Single Nucleotide Polymorphism
SP	Signal Peptide
SSGP	Secreted Salivary Gland Protein
SSR	Simple Sequence Repeat
TC	*Tribolium castaneum*
TE	Transposable Elements
tetev	*Tetranychus evansi*
tetli	*Tetranychus lintearius*
tetur	*Tetranychus urticae*
TGS	Third Generation Sequencing
TRDMT1	tRNA (cytosine-5-)-methyltransferase
tRNA	transfer RNA
UCAS	University of Chinese Academy of Sciences
UCSC	University of California Santa Cruz
UGent	Ghent University

UGT	Glucuronosyltransferase
UTR	Untranslated Region
UV	Ultraviolet
VIB	Vlaams Insititute of Biotechnology
Western	The University of Western Ontario
WGS	Whole Genome Sequencing

List of Appendices

NA

Preface

NA

To Love & Freedom

Chapter 1

1 Introduction

The advancement of Next Generation Sequencing (NGS) technology has brought a striking revolution in life science. Using NGS data and bioinformatics approaches, in Chapter 1, I initially introduce the background of NGS and several arthropods that I used as model organisms for genomic case studies. Chapter 2 illustrates a best-practice NGS toolkit with key concepts and essential steps, aiming to provide NGS beginners a comprehensive understanding of a genome project. From Chapters 3 to 7, I discuss several NGS applications on arthropod genomes. In Chapter 3, I report the updated assembly and annotation of *Tetranychus urticae* genome and how it will provide more opportunities for chromosomal-level and structural-level studies. In Chapter 4, the comparative analysis of three spider mite genomes representing three completely different feeding models is presented. Increased understanding of these spider mite genomes not only offers us new insights into arthropod evolution and plant-herbivore interactions but also provides unique opportunities to develop novel plant protection strategies against these agricultural pests. Interestingly, a novel F-box gene family was found to be expanded with over 220 copies in *T. urticae* but only about thirty copies in two other mite species. The comprehensive analysis of this novel gene family is presented in Chapter 5. Furthermore, I report the five assembled and annotated genomes of flat mites (*Brevipalpus*), another group of major agricultural pests feeding on important economic crops such as citrus and strawberry. In Chapter 6, I present a comparative genomics study of five *Brevipalpus* genomes. This is further followed by Chapter 7, where I discuss the wasp (*Copidosoma floridanum*) genome, focusing on DNA-methyltransferase (DNMT) gene family annotation, as a partial investigation of polyembryony. In the last Chapter 8, I quickly review state-of-the-art NGS technologies and future perspectives in genomic applications.

1.1 NGS and Bioinformatics

DNA is the secret of life (Marks, 2003). With the aim to decode DNA sequences, sequencing technologies have been developed, improved and revolutionized the field of life sciences over the past four decades, as shown in Figure 1 (Heather and Chain, 2016). Initiated at the beginning of the 1970s, the first DNA sequence was identified by Wu and Taylor at Cornell University (Wu and Taylor, 1971) and the first complete gene was sequenced by Min Jou and his colleagues at Ghent University (Min Jou *et al.*, 1972). Five years later, the first virus genome was sequenced by Sanger *et al.* (Sanger *et al.*, 1977a; Sanger *et al.*, 1977b). In 1995, *Haemophilus influenza* was sequenced as the first complete bacterial genome (Fleischmann *et al.*, 1995). Later, the first complete eukaryotic genome, *Saccharomyces cerevisiae,* was released in 1996 (Goffeau *et al.*, 1996). Coming to the 21st century, the first arthropod genome of *Drosophila melanogaster,* the first plant genome of *Arabidopsis thaliana*, and the first draft human genome were released (Adams *et al.*, 2000; Arabidopsis Genome Initiative, 2000; Lander *et al.*, 2001). All these important genomes mentioned above were sequenced using the first-generation sequencing technology - the Sanger Chain Termination method (Sanger *et al.*, 1977b), which is highly demanding of time, expense and labor work. Consequently, only large consortia with substantial funding could accomplish these sorts of endeavors in priority organisms.

A breakthrough came in 2004 with the introduction of NGS, represented by 454 pyrosequencing leading to rapid advances and subsequently followed by Illumina and SOLiD sequencing technologies (Margulies *et al.*, 2005; Shendure and Ji, 2008). From the first-generation sequencing to NGS, only over a few years, technologies have reduced the sequencing costs by several orders of magnitude and have been accelerating diverse fields in bioinformatics and genomics. Currently, NGS has sharply decreased the sequencing expense for a human genome, from ten million US dollars to approximately one thousand US dollars (Figure 2a, data source: https://www.genome.gov/sequencingcosts/, last access on Jan 2016).

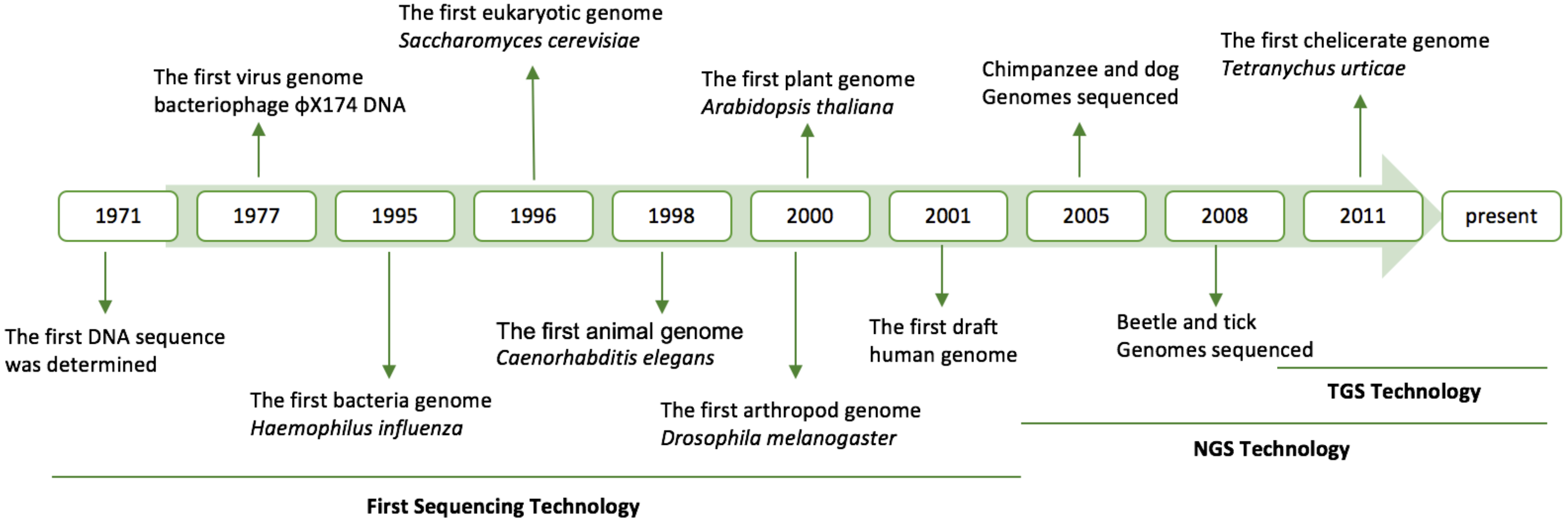


Figure 1: The history of sequencing technologies and representative genomes. TGS: the Third Generation Sequencing.

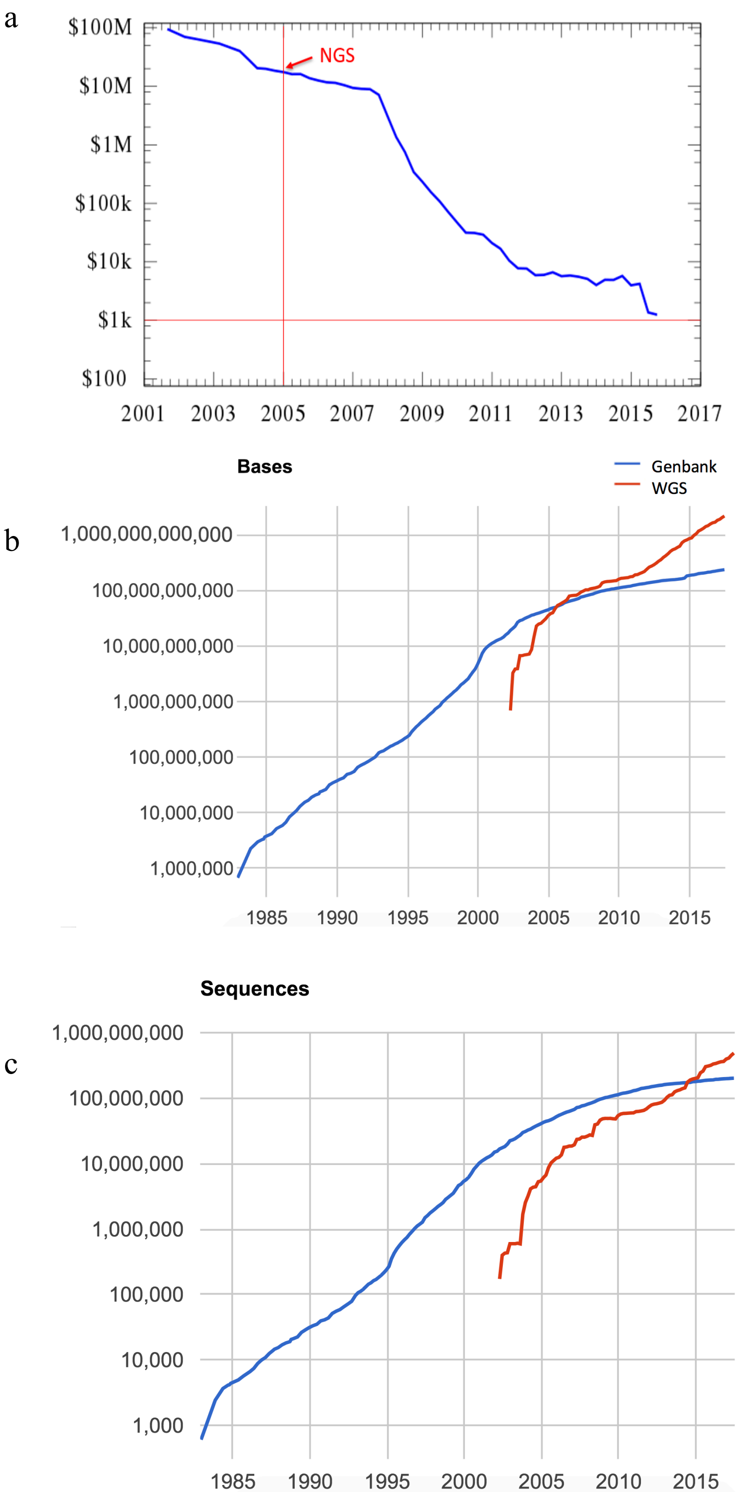


Figure 2: Statistics of sequencing cost and data.

a: the costs for sequencing a human genome; b and c: GenBank and WGS statistics for annual sequenced bases and sequences. GenBank data in blue and WGS data in red. Data source: https://www.ncbi.nlm.nih.gov/genbank/statistics/, last access on Jan 2016.

Compared with the first-generation sequencing technology, genome projects using NGS technologies have become more affordable and time-efficient, bringing genomics within the reach of individual laboratories. Advances in these technologies also allow greater numbers and varieties of organisms to be studied. This has led to numerous prokaryotic and eukaryotic species being sequenced; their genomic data is being released at a steady, ever-increasing speed (Figure 2b and 2c). To date, nearly 300 animal genomes have been released in public databases (Supplementary data: Table 32).

Genome sequencing has been widely applied in a great range of fields and proliferated and deepened our understanding of life sciences. For instance, in agricultural studies, it has offered insights into pest-control strategies and genetically modified organisms (Drosophila 12 Genomes Consortium *et al.*, 2007; Grbic *et al.*, 2011a; Ngoc *et al.*, 2016; Willems *et al.*, 2016). In clinical studies, genome sequencing also has provided the possibility of faster, safer and more precise diagnoses (Boland *et al.*, 2015; Crowgey *et al.*, 2015; Ang *et al.*, 2016; Au *et al.*, 2016; Duke *et al.*, 2016).

Overall, Sanger sequencing technology has been generally supplanted by NGS technology, which is currently dominant in the global sequencing market. The major NGS supplier is Illumina because of its low expense and massive productivity. Recently, the Third-Generation Sequencing (TGS) technologies are also coming on hand, for example, PacBio RS SMRT and Oxford Nanopore Minion as shown in Figure 3, they produce longer genomic reads, but their high cost and high error rates, nevertheless, still mitigate longer read advantage to NGS.

The accumulation of massive genomic datasets is useless without data mining and statistics in bioinformatics. At no other point in history has our ability to understand the complexities of life been so dependent on data analysis skills to decode these data. Thus, bioinformatics, the interdisciplinary field of science combining computer science, statistics, mathematics, and engineering, develops methods and software tools to understand biological data. Here in the following chapters, I will apply bioinformatics as a major approach to decoding high-throughput NGS data using arthropod genomes as case studies, concentrating on genome assembly, genome annotation, and comparative genomics.

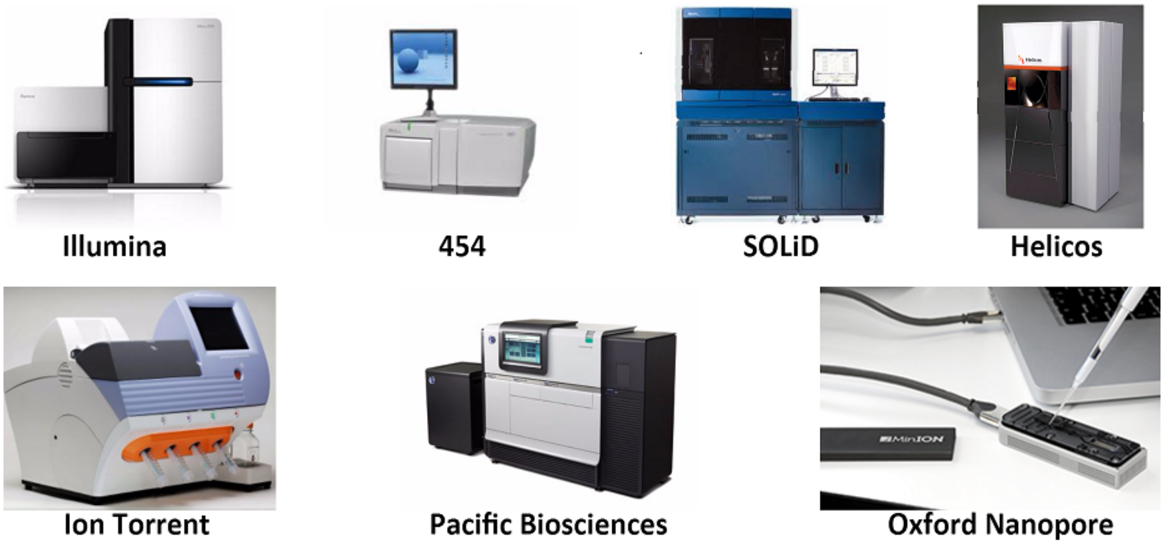


Figure 3: Major NGS and TGS sequencing technologies.
Top: NGS technologies are known as high-throughput and fast sequencing; bottom: TGS technologies are the latest advanced method by producing longer reads at least 10 kb.

1.2 Genome assembly and genome annotation

Genome assembly and genome annotation are two essential steps for NGS downstream analyses.

Genome assembly is simply the genome sequence produced after chromosomes have been fragmented, those fragments have been sequenced, and the resulting sequences have been put back together. This is due to the limitation that an entire chromosome cannot be read by any current sequencing technologies. Therefore, chromosomes must be split into much smaller pieces (100 bp-100 kb), known as **reads**, for sequencers to accommodate and sequence. Therefore, these short reads need to be assembled back to, ideally, original chromosomal-level. However, it is difficult to accomplish these challenges introduced by the limitations of the computational platform, and most importantly, the short length of these reads in a context where genomes contain many short and repetitive sequence motifs.

Normally, only model organisms have a finished or complete genome with the coverage of more than 95% (e.g. the human genome, *Drosophila* genome and *Arabidopsis* genome), but most of published genomes are still only at the level of draft genomes, even if published in top journals (Adam, 2002; Check, 2002; Dehal *et al.*, 2002; Dennis, 2003; Xia *et al.*, 2004; Kasahara *et al.*, 2007; Ming *et al.*, 2008; Green *et al.*, 2010; Bos *et al.*, 2011; Jex *et al.*, 2011; Jia *et al.*, 2013; Ling *et al.*, 2013; International Wheat Genome Sequencing, 2014). The details for the draft genome, complete genome and finished genome are listed in NGS TERM BOX.

Once a genome is assembled to a certain level (i.e. impossible to improve its assembly using all the available datasets), then the repetitive regions across the whole genomic sequences need to be masked, known as **genome masking**. After a genome is masked, the following genome annotation process will not predict genes in the masked regions. However, **over masking** would lead to a bad genome annotation because it would miss genes in the masked loci while **under masking** would lead to over predicted genomes.

Genome annotation is a process of identifying the locations of genes and all of the coding regions in a genome and determining the function of these genes. It consists of three layers: *where*, *what* and *how* (Stein, 2001). First, *where* are the genes across the genome? Genes are hard to determine and define because of the complexity of eukaryotic gene structures, including various regions such as promoter, enhancer, TATA box, 5'UTR, start codon, exons, introns, stop codon and 3'UTR, as shown in Figure 4. Second, once these loci are located, we need to understand these predicted genes at the protein-level: *what* their functions are. Third, we ask *how* these genes act at process-level, e.g., how these associated proteins function in the cell, or even more, in the complexity of life activities.

1.3 Comparative genomics

Comparative genomics, a new branch of bioinformatics and genomics, provides a highly detailed view of how organisms are related to each other from the genomic perspective (Ellegren, 2008; Rubin and Moreau, 2016). Comparative genomics can uncover a wide range of genomic features including DNA sequences, genome structures, gene orders, and likely gene regulatory networks (Xia, 2013). Basically, comparative genomics can be performed at three levels: 1) population genomics - within the same species (i.e. pan-genomics in microbiology) (Ledford, 2008; Romiguier *et al.*, 2014; Allentoft *et al.*, 2015); 2) wide-ranging comparative genomics - across related species (Drosophila 12 Genomes Consortium *et al.*, 2007; Green *et al.*, 2014); 3) meta-genomics - across different diverse species that may vary in their phylogenetic relatedness (i.e. environmental genomics, eco-genomics or community genomics) (Vieira-Silva and Rocha, 2010; Chen *et al.*, 2011; Quraishi *et al.*, 2011; Roux *et al.*, 2014);

A great number of comparative genomics studies have revealed genetic variations, providing valuable information on human diseases or evolution across different species (Varki and Altheide, 2005; Lefebure and Stanhope, 2007; Kuehn, 2008; Tettelin *et al.*, 2008; Alfoldi and Lindblad-Toh, 2013). Furthermore, comparative genomics also promises a closer look at eukaryotic evolutionary mechanisms, adaptations, and diseases (Grbic *et al.*, 2011a; Gulia-Nuss *et al.*, 2016).

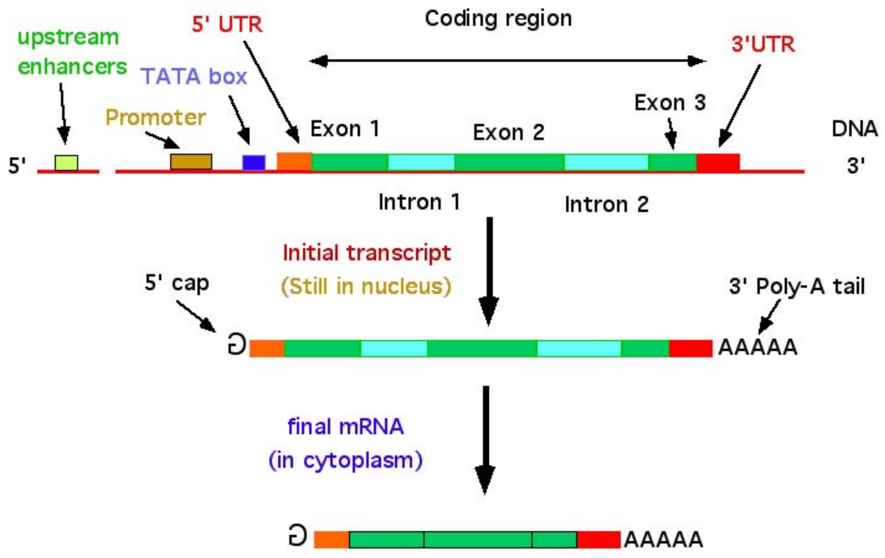


Figure 4: Eukaryotic gene structure and gene transcription processes.
From DNA sequence to pre-mRNA transcription and alternative splicing into mature mRNA, used for protein translation. Photo credit:
http://nitro.biosci.arizona.edu/courses/EEB600A-2003/lectures/lecture24/lecture24.html

1.4 Arthropods

Arthropods, as the name suggests as 'jointed legs', are invertebrate animals with three anatomical parts: an exoskeleton, a segmented body, and paired appendages. Because arthropods' body plan consists of rigid cuticle that inhibits growth, they must periodically replace the body cuticle by molting, also known as ecdysis. To date, the number of arthropods is estimated over 1 million species, encompassing over eighty percent of all described living animal species from insects, arachnids, myriapods, and crustaceans. (Odegard, 2000).

Arthropods are quite ancient and the fossil records first reveal their presence about 550 Million Years Ago (MYA), compared with dinosaurs 240 MYA and humans 6 MYA, as shown in Figure 5. A recent study also has revealed arthropod phylogenies and provided a statistically well-supported phylogenetic framework for the largest animal phylum (Regier *et al.*, 2010). Arthropoda includes Chelicerata, Myriapoda, and Pancrustacea (comprising all crustaceans and hexapods), of which, the most well-studied insects belong to Hexapoda in Pancrustacea (Zrzavý and Štys, 1997; Rota-Stabelli *et al.*, 2010).

Over thirty arthropod genomes have been sequenced in the past decade (Supplementary data: Table 33). These genomes have tremendously enhanced our knowledge of arthropod genetics and genomics, either in natural populations of a given species or across different species (Adams *et al.*, 2000; Carlton *et al.*, 2002; Holt *et al.*, 2002; Waterston *et al.*, 2002; Hardison, 2003; Ivanova *et al.*, 2003; Xia *et al.*, 2004; Lindblad-Toh *et al.*, 2005; Ullmann *et al.*, 2005; Honeybee Genome Sequencing, 2006; Grbic *et al.*, 2011a; Sanggaard *et al.*, 2014).

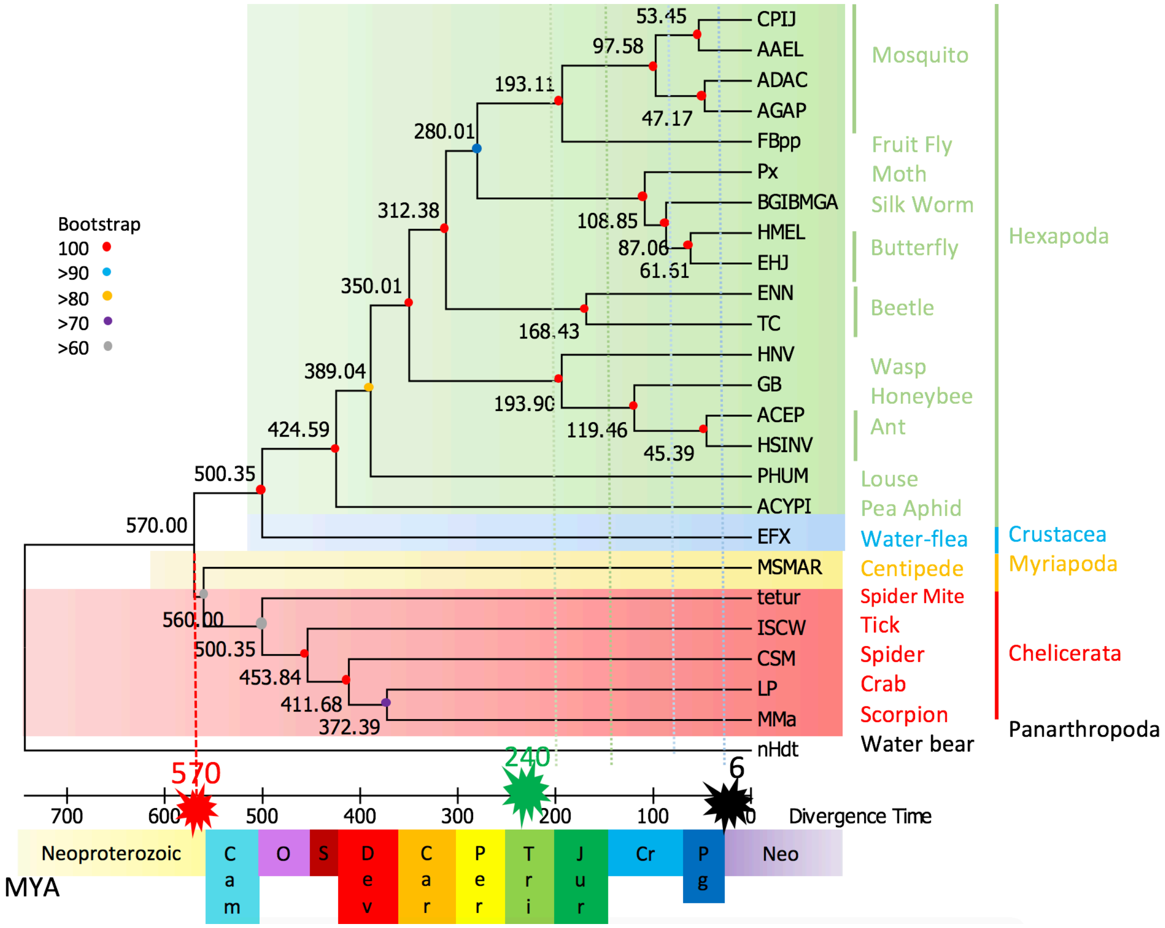


Figure 5: Phylogeny of arthropods and evolutionary time.
The estimated times for the first arthropod, first dinosaur and first hominin are marked in red, green and black, respectively. Abbreviations: Np, Neoproterozoic; Cam, Cambrian; O, Ordovician; S, Silurian; Dev, Devonian; Car, Carboniferous; Pe, Permian; Tr, Triassic; Ju, Jurassic; Cr, Cretaceous; Pg, Paleogene; N, Neogene. Values in the abscissa are millions of years. Details for this phylogeny are in Chapter 6.

1.5 The *Tetranychus* spider mites

Mites belong to Chelicerata, a basal branch of Arthropoda and also the second largest group of terrestrial animals (Regier *et al.*, 2010; Misof *et al.*, 2014). The most diverse chelicerate clade Acari (including ticks and mites) have a wide range of lifestyles from parasitic to predatory and herbivory (Dunlop and Selden, 2009). The *Tetranychus urticae* is one of the most economically important species due to a high feeding potential that can destroy various agricultural plants worldwide (Walter, 2011), especially among greenhouse crops (e.g., tomatoes, peppers, cucumbers, roses, and carnations), annual field crops (e.g., maize, soybeans, and sugar beets) and perennial crops (e.g., strawberries, grapes, apples, and pears) (Bolland *et al.*, 1997). Because of their small body size, spider mites normally disperse using the wind as a vehicle to travel from plant to plant. Mites live in colonies, mostly on the underside of the leaves, probably to avoid the UV light. They feed by piercing leaf tissues using stylets and sucking up plant cell contents (Figure 6). Feeding marks usually show up as light or gray dots on the leaves. As feeding continues, these plant leaves turn yellow and may dry up or drop off.

Spider mites are extremely small (less than 0.5mm) and can hardly be seen by the naked eye. Male spider mites are smaller than females. The sex ratio of spider mites is female-biased: there are approximately 3 females to 1 male. They have four major developmental stages: egg, larvae, nymph, and adult. It takes about one week to ten days from hatching the egg to adult at room temperature. Because spider mites prefer hot and dry conditions, it takes less time to grow up in the hot wild environment. Most mite species overwinter as eggs on the leaves or on the bark of host plants. Once the temperature gets warm, the tiny six-legged larvae begin hatching, and after a few days, they molt into the nymph stage. Nymphs have eight legs and after two more rounds of molting, they grow up and become mature adults. Normally, spider mites reproduce dramatically unless they suffer diapause during bad environmental conditions. Once the weather improves better, a female adult spider mite can lay dozens of eggs per day, which hints that the spider mite population number increases approximately at a rate of one generation per week (assuming that a female mite reproduces 20 eggs/day and 2 weeks-production-capability/generation with one week grown up from egg to adult).

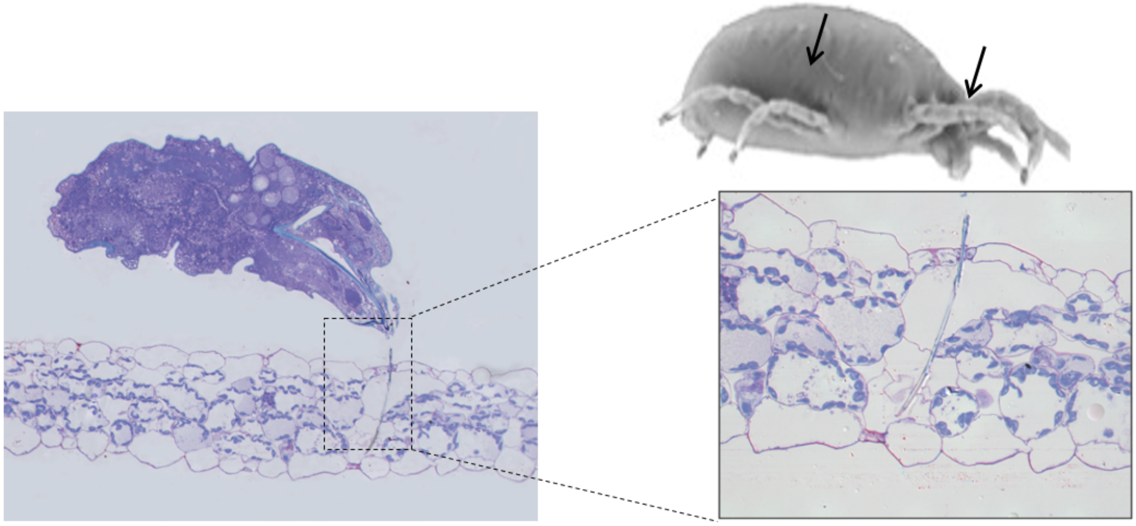


Figure 6: The spider mite is feeding the content of leaf cells using its stylet.
The left arrow indicates the gut and the right arrow indicates the stylet of the spider mite. Microscopy photo credit: Nicolas Bensoussan.

The spider mite can produce silks, a potential chelicerate nano-biomaterials. In fact, the name of ‘spider mite’ highlights their ability to produce silk-like webbing that is used to establish a colonial micro-habitat, protect against abiotic agents, shelter from predators, communicate via pheromones and provide a vehicle for dispersion (Grbic *et al.*, 2011a). Spider mites have a unicellular gland that extends from each palp back to the central nervous mass, which is almost filled with vacuoles containing a proteinaceous secretory product. Silk production in spider mites represents *de novo* evolution of silk-spinning relative to silk production in spiders (Sabelis, 1987), but spider mite silk fibers are thinner 54 ± 3 nm (adult silk, Figure 7b) and 23.3 ± 0.9 nm (larval silk) (Grbic *et al.*, 2011a), i.e., 435-185 times thinner than the silk fibers of the spider *Nephila clavipes* (Kluge *et al.*, 2008). Consequently, evolutionary innovation in the process of *T. urticae* silk production will extend the repertoire of potential chelicerate biomaterials.

T. urticae is not the only spider mite feeding on plants. It has been reported in the book *World Catalogue of the Spider Mite Family* that spider mite family includes over 1,200 species (Bolland *et al.*, 1997). For instance, one of the specialist spider mites (also called monophagous mites) *Tetranychus lintearius* originated from Europe, feeding on one host plant, *Ulex europeus* (gorse). *T. lintearius* is native to parts of Europe and recently became an invasive species due to its high productivity of silk (Figure 7d). Another spider mite, *Tetranychus evansi* is native to South America and has been accidentally introduced to other parts of the world. *T. evansi* is an oligophagous pest, feeding on *Solanaceous* plants such as tomato, potato and tobacco (Qureshi *et al.*, 1969; Tsagkarakou *et al.*, 2007; Gotoh *et al.*, 2010; Boubou *et al.*, 2011; Van Leeuwen *et al.*, 2013; Antonious *et al.*, 2014).

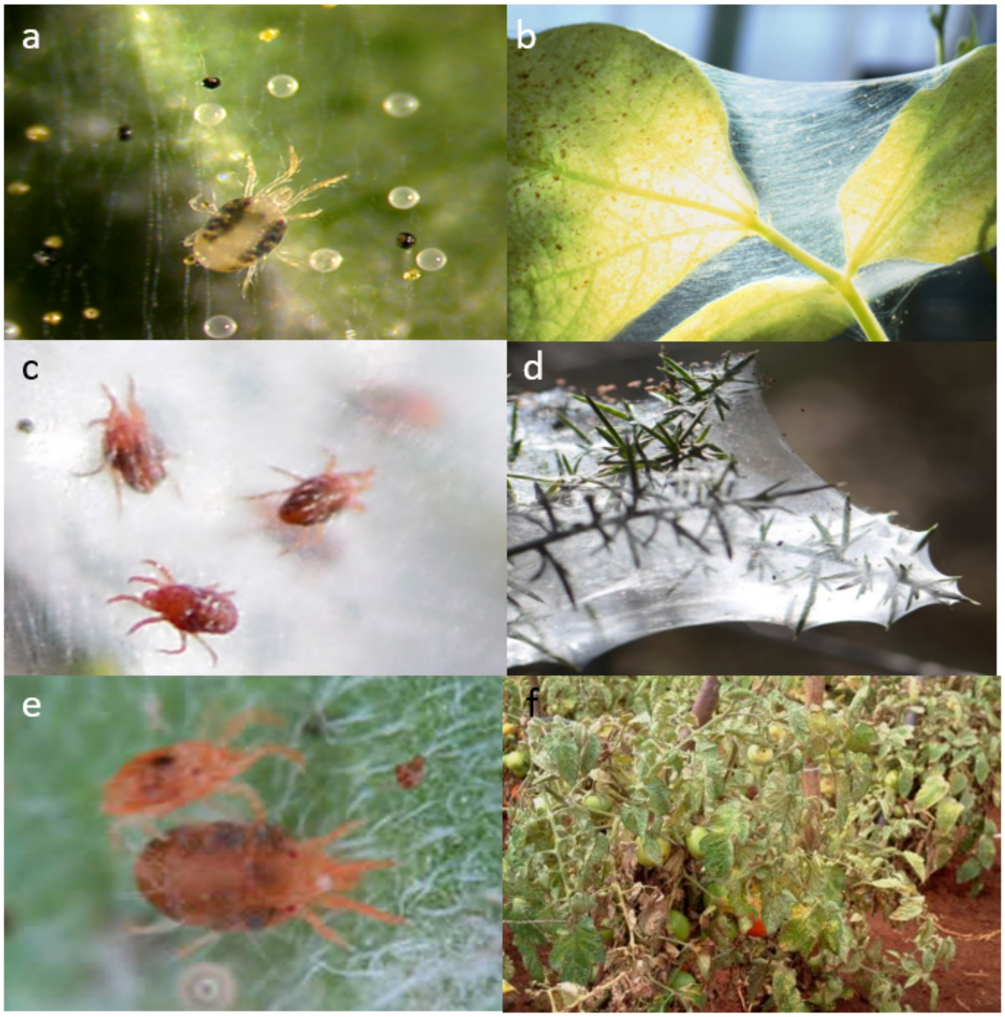


Figure 7: An overview of the three spider mites.
a&b: the two-spot *T. urticae* spider mites and their silk on soybean leaves; c&d: *T. lintearius* and their silk on gorse; e&f: *T. evansi* and their damage on tomato plants. Photo credit: (a&b) The Grbic Lab; (c) Monique and Daniel Blogger; (e) A. Migeon; (f): http://www.infonet-biovision.org/PlantHealth/MinorPests/Spider-mites.

As such, whether specialists or global generalists, these mite pests have had a huge economic impact on agriculture. Therefore, pest biological control and crop protection such as damage assessment, host-plant resistance, and pesticide resistance are severe. Phytoseiid predators and pesticides have been conventionally used to control spider mites (Oliveira *et al.*, 2007). However, pesticides can encourage the spread of spider mites by killing their predators and meanwhile, mites are also known to develop quick resistance to various pesticides (Van Leeuwen *et al.*, 2010). Chemical control often causes a broad cross-resistance within and between pesticide classes, resulting in resistance to novel pesticides within 2 to 4 years. Many biological aspects of spider mites, including rapid development, high fecundity, and haplodiploid sex determination, seem to facilitate rapid evolution of pesticide resistance (Grbic *et al.*, 2011a). Therefore, it is necessary to use effective natural and biological methods to develop alternative pest control strategies for sustainable agriculture (Skirvin and de Courcy Williams, 1999; Easterbrook *et al.*, 2001; Skirvin and Fenlon, 2001; Fraulo and Liburd, 2007; Abad-Moyano *et al.*, 2009; Davies *et al.*, 2009; Grbic *et al.*, 2011a; Hardman *et al.*, 2013; Howell and Daugovish, 2013; Navajas *et al.*, 2013b; Van Leeuwen *et al.*, 2013; Woods *et al.*, 2014; Gigon *et al.*, 2016). Thanks to the advancement of NGS, comparative genomics provides a powerful tool for gleaning further insight into pest control studies. Spider mites are convenient experimental subjects in a broader context and might become the best model to study resistance evolution and plant interactions on a genomic scale (Van Leeuwen *et al.*, 2013).

1.6 The *Brevipalpus* flat mites

Brevipalpus mites (Acari: Tenuipalpidae) are commonly known as false spider mites or flat mites because of their inability of producing silk and flat-shaped body (Figure 8). They also represent one of the most economically important mites in the world, partly due to their association with the transmission of plant viruses, the most economically damaging of which is *Citrus Leprosis Virus* (Rodrigues *et al.*, 2003). Over forty plant species have been reported infected with plant viruses that are transmitted through flat mites (Beard *et al.*, 2015). Although they are not as agriculturally important as spider mites, flat mites are of sufficient concern to warrant investigations of their biology and control (McMurtry and Croft, 1997; Rossi-Zalaf *et al.*, 2008).

Flat mites physically range from 0.25-0.4 mm in size, with a diversity of body shapes from round, elongate, pyriform, ovoid, and triangular in cross-section with a flat venter or flat dorsum (Figure 8). Meanwhile, they also vary a great deal in color: red being the most common but many latest reported mites from Asia, Africa, and Australia show a great range of diversity in color variation from yellow, orange, green to brown, as extensively demonstrated on the website entitled *Flat Mites of The World*, which is accessible at the following link: http://idtools.org/id/mites/flatmites/#sthash.X68x3Oty.dpuf

Genetically, male flat mites are haploid while females are diploid. However, there are some differences between spider mites and flat mites*,* as indicated in Table 1. Flat mites do not spin webbing and their lifespan is generally longer than spider mites*,* but with fewer generations. Spider mites prefer hot and dry environments whereas flat mites like humid habitations, hiding in more shaded areas on their hosts in a humid environment to avoid higher temperature conditions. Flat mites have a broader appetite for diverse plant tissues. For example, spider mites feed on plant leaves while flat mites feed not only on the plant leaves but also plant buds, stems, and fruit.

Several species of flat mites are currently recognized as the most important economic pests within the genus of *Brevipalpus* (Childers and Rodrigues, 2011). *Brevipalpus californicus*, known as a vector for the orchid fleck virus, has a wide range of host plants, causes spots and rings on orchid leaves, and can form galls on bitter orange (Childers and Rodrigues, 2011). *Brevipalpus phoenicis* feeds on tea plant leaves and thus can reduce tea yields. It has also been observed on tangerine. Additionally, it is also a vector for Cilevirus, a plant virus that causes citrus leprosis. *B. phoenicis* is also a vector of passion fruit green spot virus and coffee ringspot virus (Childers and Rodrigues, 2011). *Brevipalpus obovatus* is another global agricultural pest feeding on ornamentals (Miranda *et al.*, 2007). Moreover, there are also some other important flat mites such as *Brevipalpus papayensis*, *Brevipalpus chilensis* and *Brevipalpus lewisi* (Childers and Rodrigues, 2011).

In addition, flat mites have a very important biological feature - the parthenogenetic with thelytokous reproduction, which is a type of parthenogenesis in which females are reproduced from unfertilized eggs. This is because of the presence of feminizing bacterial symbionts of the genus *Cardinium* that induce haploid thelytoky in most clones of three closely related flat mites (Groot and Breeuwer, 2006). Interestingly, infected females can produce offspring with either infected or uninfected males. However, uninfected females can only have descendants with uninfected males. The mechanism behind this endosymbiosis is little known yet.

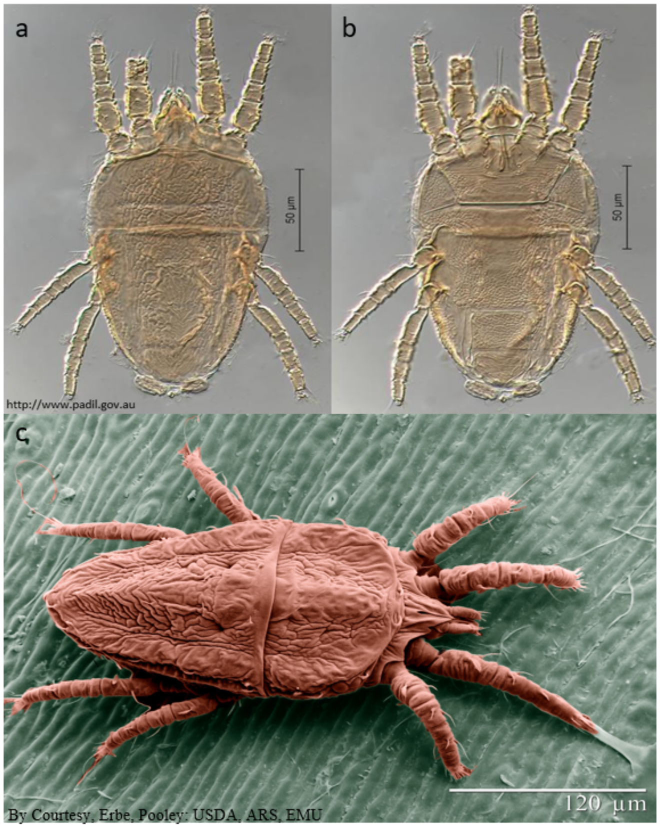


Figure 8: The body shape of flat mite.
a: diagnostic dorsal - adult, magnification 40×; b: diagnostic images ventral - adult magnification 40×; c: the flat mite (*Brevipalpus phoenicis*) on citrus. Photo credit: (a&b) http://www.padil.gov.au; (c) Courtesy, Erbe, Pooley from USDA, ARS, EMU.

Table 1: Biological comparison between spider mites and flat mites.

	Spider mites	Flat mites
Physical size	0.1 to 0.5mm	~0.25 to 0.4mm
Sex determination	Male (haploid), female (diploid)	Male (haploid), female (diploid)
Longevity	~30 days or ~4 weeks	41.68 ± 5.92 days
Lifecycle	Four Stages	Four Stages
Eggs/female	Hundreds (~20/day)	50-60 (in a lifetime)
Webbing	Yes	No
Target Tissue	Primarily leaves	Leaves, stems, fruits or nuts
Body color	Green or red	Various
Preferable condition	Dry and hot	Humid

1.7 The wasp *Copidosoma floridanum*

C. floridanum, taxonomically belonging to Insecta, is a parasitoid wasp of moths. *C. floridanum* has the largest record of brood with over 3,055 individuals (Alvarez, 1997). It has a fascinating developmental mode, as shown in Figure 9 (Zhurov *et al.*, 2004, 2007). Briefly, a female adult wasp initially lays two eggs into a suitable host, usually one male and one female; Subsequently, each egg divides repeatedly and finally develops into a brood of multiple individuals with two major morphogenesis castes. This process is known as polyembryony. The two major morphogenesis castes are different: one is reproductive caste consisting of more than three-quarters of all larvae; the other is precocious caste, primarily involving in adjusting sex ratio by killing males, including both reproductive and precocious males. The host's moulting cycle plays a significant role in determining the identity of precocious and reproductive larvae. More specifically, the *C. floridanum* young mature in synchrony with specific phases within the moth's molting cycle. In the early stages of embryonic development, changes within the host's developmental program intrinsically influence caste determination (Strand *et al.*, 1997). The precocious larvae will die in their host while the reproductive larvae keep feeding on the tissues of their host. Eventually, the reproductive wasps become imagoes (the final and fully developed adult stage of an insect, typically winged) and fly away.

Despite its significance to agriculture as a method of pest control, the mechanism of wasp morphogenesis is currently poorly understood, in part because of lack of corresponding genomic and methylomic data. Only until recently, studies have categorized differentially expressed genes in *C. floridanum* castes that code for classifiable proteins that the sterile soldiers share. Soldiers and reproductive larvae express enzymes with the differential usage of proteinase inhibitors and ribosomal proteins (Donnell and Strand, 2006).

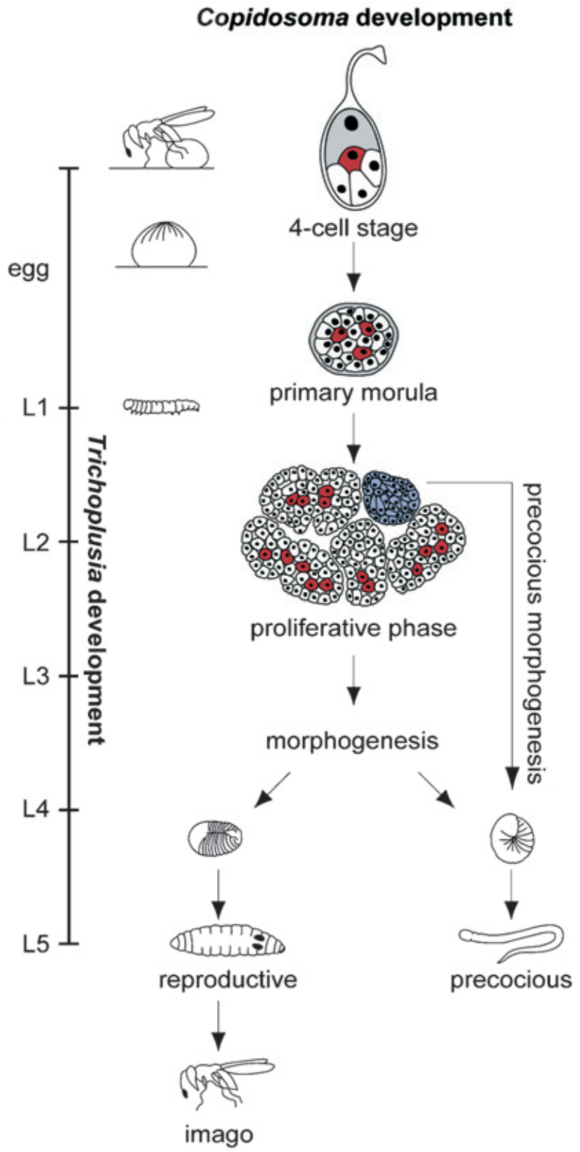


Figure 9: The life cycle and development of *C. floridanum* in its host *Trichoplusia ni.* Photo credit: Dr. Vladimir Zhurov at Western University, Canada.

1.8 References

Abad-Moyano, R., Pina, T., Ferragut, F., and Urbaneja, A. (2009). Comparative life-history traits of three phytoseiid mites associated with Tetranychus urticae (Acari: Tetranychidae) colonies in clementine orchards in eastern Spain: implications for biological control. Experimental & applied acarology *47*, 121-132.

Adam, D. (2002). Draft cow genome heads the field. Nature *417*, 778.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F.*, et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Alfoldi, J., and Lindblad-Toh, K. (2013). Comparative genomics as a tool to understand evolution and disease. Genome research *23*, 1063-1068.

Allentoft, M.E., Sikora, M., Sjogren, K.G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlstrom, T., Vinner, L.*, et al.* (2015). Population genomics of Bronze Age Eurasia. Nature *522*, 167-172.

Alvarez, J.M.A. (1997). Chapter 26: Largest Parasitoid Brood. In Book of Insect Records.

Ang, S.F., Lim, S.C., Tan, C., Fong, J.C., Kon, W.Y., Lian, J.X., Subramanium, T., and Sum, C.F. (2016). A preliminary study to evaluate the strategy of combining clinical criteria and next generation sequencing (NGS) for the identification of monogenic diabetes among multi-ethnic Asians. Diabetes Res Clin Pract *119*, 13-22.

Antonious, G.F., Kamminga, K., and Snyder, J.C. (2014). Wild tomato leaf extracts for spider mite and cowpea aphid control. J Environ Sci Health B *49*, 527-531.

Arabidopsis Genome Initiative, A. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796-815.

Au, C.H., Wa, A., Ho, D.N., Chan, T.L., and Ma, E.S. (2016). Clinical evaluation of panel testing by next-generation sequencing (NGS) for gene mutations in myeloid neoplasms. Diagn Pathol *11*, 11.

Beard, J.J., Ochoa, R., Braswell, W.E., and Bauchan, G.R. (2015). Brevipalpus phoenicis (Geijskes) species complex (Acari: Tenuipalpidae)--a closer look. Zootaxa *3944*, 1-67.

Boland, P.M., Ruth, K., Matro, J.M., Rainey, K.L., Fang, C.Y., Wong, Y.N., Daly, M.B., and Hall, M.J. (2015). Genetic counselors' (GC) knowledge, awareness, understanding of clinical next-generation sequencing (NGS) genomic testing. Clin Genet *88*, 565-572.

Bolland, H.R., Gutierrez, J., and Flechtmann, C.H.W. (1997). World Catalogue of the Spider Mite Family. 1-3.

Bos, K.I., Schuenemann, V.J., Golding, G.B., Burbano, H.A., Waglechner, N., Coombes, B.K., McPhee, J.B., DeWitte, S.N., Meyer, M., Schmedes, S., *et al.* (2011). A draft genome of Yersinia pestis from victims of the Black Death. Nature *478*, 506-510.

Boubou, A., Migeon, A., Roderick, G.K., and Navajas, M. (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.

Carlton, J.M., Angiuoli, S.V., Suh, B.B., Kooij, T.W., Pertea, M., Silva, J.C., Ermolaeva, M.D., Allen, J.E., Selengut, J.D., Koo, H.L., *et al.* (2002). Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. Nature *419*, 512-519.

Check, E. (2002). Draft mouse genome makes public debut. Nature *417*, 106.

Chen, C., Zhang, Z., Ding, A., Wu, J., Xiao, J., and Sun, Y. (2011). Bar-Coded Pyrosequencing Reveals the Bacterial Community during Microcystis water Bloom in Guanting Reservoir, Beijing. Procedia Engineering *18*, 341-346.

Childers, C.C., and Rodrigues, J.C.V. (2011). An overview of Brevipalpus mites (Acari: Tenuipalpidae) and the plant viruses they transmit. Zoosymposia *6*, 180-192.

Crowgey, E.L., Kolb, A., and Wu, C.H. (2015). Development of Bioinformatics Pipeline for Analyzing Clinical Pediatric NGS Data. AMIA Jt Summits Transl Sci Proc *2015*, 207-211.

Davies, J.T., Ireson, J.E., and Allen, G.R. (2009). Pre-adult development of Phytoseiulus persimilis on diets of Tetranychus urticae and Tetranychus lintearius: implications for the biological control of Ulex europaeus. Experimental & applied acarology *47*, 133-145.

Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D.M., *et al.* (2002). The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science *298*, 2157-2167.

Dennis, C. (2003). Draft guidelines ease restrictions on use of genome sequence data. Nature *421*, 877-878.

Donnell, D.M., and Strand, M.R. (2006). Caste-based differences in gene expression in the polyembryonic wasp Copidosoma floridanum. Insect biochemistry and molecular biology *36*, 141-153.

Drosophila 12 Genomes Consortium, Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., *et al.* (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature *450*, 203-218.

Duke, J.L., Lind, C., Mackiewicz, K., Ferriola, D., Papazoglou, A., Gasiewski, A., Heron, S., Huynh, A., McLaughlin, L., Rogers, M., *et al.* (2016). Determining performance characteristics of an NGS-based HLA typing method for clinical applications. HLA *87*, 141-152.

Dunlop, J.A., and Selden, P.A. (2009). Calibrating the chelicerate clock: a paleontological reply to Jeyaprakash and Hoy. Experimental & applied acarology *48*, 183-197.

Easterbrook, M.A., Fitzgerald, J.D., and Solomon, M.G. (2001). Biological control of strawberry tarsonemid mite Phytonemus pallidus and two-spotted spider mite Tetranychus urticae on strawberry in the UK using species of Neoseiulus (Amblyseius) (Acari: Phytoseiidae). Experimental & applied acarology *25*, 25-36.

Ellegren, H. (2008). Comparative genomics and the study of evolution by natural selection. Molecular ecology *17*, 4586-4596.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M.*, et al.* (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496-512.

Fraulo, A.B., and Liburd, O.E. (2007). Biological control of twospotted spider mite, Tetranychus urticae, with predatory mite, Neoseiulus californicus, in strawberries. Experimental & applied acarology *43*, 109-119.

Gigon, V., Camps, C., and Le Corff, J. (2016). Biological control of Tetranychus urticae by Phytoseiulus macropilis and Macrolophus pygmaeus in tomato greenhouses. Experimental & applied acarology *68*, 55-70.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M.*, et al.* (1996). Life with 6000 genes. Science *274*, 546, 563-547.

Gotoh, T., Sugimoto, N., Pallini, A., Knapp, M., Hernandez-Suarez, E., Ferragut, F., Ho, C.C., Migeon, A., Navajas, M., and Nachman, G. (2010). Reproductive performance of seven strains of the tomato red spider mite Tetranychus evansi (Acari: Tetranychidae) at five temperatures. Experimental and Applied Acarology *52*, 239-259.

Grbic, M., Van Leeuwen, T., Clark, R.M., Rombauts, S., Rouze, P., Grbic, V., Osborne, E.J., Dermauw, W., Ngoc, P.C., Ortego, F.*, et al.* (2011a). The genome of Tetranychus urticae reveals herbivorous pest adaptations. Nature *479*, 487-492.

Green, R.E., Braun, E.L., Armstrong, J., Earl, D., Nguyen, N., Hickey, G., Vandewege, M.W., St John, J.A., Capella-Gutierrez, S., Castoe, T.A.*, et al.* (2014). Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science *346*, 1254449.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.*, et al.* (2010). A draft sequence of the Neandertal genome. Science *328*, 710-722.

Groot, T.V., and Breeuwer, J.A. (2006). Cardinium symbionts induce haploid thelytoky in most clones of three closely related Brevipalpus species. Experimental & applied acarology *39*, 257-271.

Gulia-Nuss, M., Nuss, A.B., Meyer, J.M., Sonenshine, D.E., Roe, R.M., Waterhouse, R.M., Sattelle, D.B., de la Fuente, J., Ribeiro, J.M., Megy, K*., et al.* (2016). Genomic insights into the Ixodes scapularis tick vector of Lyme disease. Nature communications *7*, 10507.

Hardison, R.C. (2003). Comparative genomics. Plos Biol *1*, 156-160.

Hardman, J.M., van der Werf, W., Blatt, S.E., Franklin, J.L., Karsten, R., and Teismann, H. (2013). Simulating effects of environmental factors on biological control of Tetranychus urticae by Typhlodromus pyri in apple orchards. Experimental & applied acarology *60*, 181-203.

Heather, J.M., and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics *107*, 1-8.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R*., et al.* (2002). The genome sequence of the malaria mosquito Anopheles gambiae. Science *298*, 129-149.

Honeybee Genome Sequencing, C. (2006). Insights into social insects from the genome of the honeybee Apis mellifera. Nature *443*, 931-949.

Howell, A.D., and Daugovish, O. (2013). Biological control of Eotetranychus lewisi and Tetranychus urticae (Acari: Tetranychidae) on strawberry by four phytoseiids (Acari: Phytoseiidae). J Econ Entomol *106*, 80-85.

International Wheat Genome Sequencing, C. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science *345*, 1251788.

Ivanova, N., Sorokin, A., Anderson, I., Galleron, N., Candelon, B., Kapatral, V., Bhattacharyya, A., Reznik, G., Mikhailova, N., Lapidus, A*., et al.* (2003). Genome sequence of Bacillus cereus and comparative analysis with Bacillus anthracis. Nature *423*, 87-91.

Jex, A.R., Liu, S., Li, B., Young, N.D., Hall, R.S., Li, Y., Yang, L., Zeng, N., Xu, X., Xiong, Z*., et al.* (2011). Ascaris suum draft genome. Nature *479*, 529-533.

Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., Appels, R., Pfeifer, M., Tao, Y., Zhang, X*., et al.* (2013). Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature *496*, 91-95.

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y*., et al.* (2007). The medaka draft genome and insights into vertebrate genome evolution. Nature *447*, 714-719.

Kluge, J.A., Rabotyagova, U., Leisk, G.G., and Kaplan, D.L. (2008). Spider silks and their applications. Trends Biotechnol *26*, 244-251.

Kuehn, B.M. (2008). 1000 Genomes Project promises closer look at variation in human genome. JAMA : the journal of the American Medical Association *300*, 2715.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Ledford, H. (2008). Population genomics for fruitflies. Nature *453*, 1154-1155.

Lefebure, T., and Stanhope, M.J. (2007). Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. Genome biology *8*, R71.

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., Zody, M.C.*, et al.* (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature *438*, 803-819.

Ling, H.Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., Fan, H., Li, D., Dong, L., Tao, Y.*, et al.* (2013). Draft genome of the wheat A-genome progenitor Triticum urartu. Nature *496*, 87-90.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z.*, et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376-380.

Marks, A.R. (2003). DNA: The Secret of Life. Journal of Clinical Investigation *112*, 972-972.

McMurtry, J.A., and Croft, B.A. (1997). Life-styles of Phytoseiid mites and their roles in biological control. Annu Rev Entomol *42*, 291-321.

Min Jou, W., Haegeman, G., Ysebaert, M., and Fiers, W. (1972). Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. Nature *237*, 82-88.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J.H., Senin, P., Wang, W., Ly, B.V., Lewis, K.L.*, et al.* (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature *452*, 991-996.

Miranda, L.C., Navia, D., and Rodrigues, J.C. (2007). Brevipalpus mites Donnadieu (Prostigmata: Tenuipalpidae) associated with ornamental plants in Distrito Federal, Brazil. Neotrop Entomol *36*, 587-592.

Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G.*, et al.* (2014). Phylogenomics resolves the timing and pattern of insect evolution. Science *346*, 763-767.

Navajas, M., de Moraes, G.J., Auger, P., and Migeon, A. (2013b). Review of the invasion of Tetranychus evansi: biology, colonization pathways, potential expansion and prospects for biological control. Experimental & applied acarology *59*, 43-65.

Ngoc, P.C., Greenhalgh, R., Dermauw, W., Rombauts, S., Bajda, S., Zhurov, V., Grbic, M., Van de Peer, Y., Van Leeuwen, T., Rouze, P*., et al.* (2016). Complex Evolutionary Dynamics of Massively Expanded Chemosensory Receptor Families in an Extreme Generalist Chelicerate Herbivore. Genome biology and evolution *8*, 3323-3339.

Odegard, F. (2000). How many species of arthropods? Erwin's estimate revised. Biological Journal of the Linnean Society *71*.

Oliveira, H., Janssen, A., Pallini, A., Venzon, M., Fadini, M., and Duarte, V. (2007). A phytoseiid predator from the tropics as potential biological control agent for the spider mite Tetranychus urticae Koch (Acari: Tetranychidae). Biological Control *42*, 105-109.

Quraishi, U.M., Murat, F., Abrouk, M., Pont, C., Confolent, C., Oury, F.X., Ward, J., Boros, D., Gebruers, K., Delcour, J.A*., et al.* (2011). Combined meta-genomics analyses unravel candidate genes for the grain dietary fiber content in bread wheat (Triticum aestivum L.). Funct Integr Genomics *11*, 71-83.

Qureshi, A.H., Oatman, E.R., and Fleschne.Ca (1969). Biology of Spider Mite, Tetranychus Evansi. Ann Entomol Soc Am *62*, 898-&.

Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W., and Cunningham, C.W. (2010). Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature *463*, 1079-1083.

Rodrigues, J.C., Kitajima, E.W., Childers, C.C., and Chagas, C.M. (2003). Citrus leprosis virus vectored by Brevipalpus phoenicis (Acari: Tenuipalpidae) on citrus in Brazil. Experimental & applied acarology *30*, 161-179.

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N*., et al.* (2014). Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature *515*, 261-263.

Rossi-Zalaf, L.S., Alves, S.B., and Vieira, S.A. (2008). [Effect of culture media on virulence of Hirsutella thompsonii (Fischer) (Deuteromycetes) to control Brevipalpus phoenicis (Geijskes) (Acari: Tenuipalpidae)]. Neotrop Entomol *37*, 312-320.

Rota-Stabelli, O., Kayal, E., Gleeson, D., Daub, J., Boore, J.L., Telford, M.J., Pisani, D., Blaxter, M., and Lavrov, D.V. (2010). Ecdysozoan Mitogenomics: Evidence for a Common Origin of the Legged Invertebrates, the Panarthropoda. Genome biology and evolution *2*, 425-440.

Roux, S., Hawley, A.K., Torres Beltran, M., Scofield, M., Schwientek, P., Stepanauskas, R., Woyke, T., Hallam, S.J., and Sullivan, M.B. (2014). Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. eLife *3*, e03125.

Rubin, B.E., and Moreau, C.S. (2016). Comparative genomics reveals convergent rates of evolution in ant-plant mutualisms. Nature communications *7*, 12679.

Sabelis, W.H.a.M.W. (1987). Spider mites: Their biology, natural enemies and control (World crop pests vols 1A and 1B).

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., and Smith, M. (1977a). Nucleotide sequence of bacteriophage phi X174 DNA. Nature *265*, 687-695.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977b). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America *74*, 5463-5467.

Sanggaard, K.W., Bechsgaard, J.S., Fang, X., Duan, J., Dyrlund, T.F., Gupta, V., Jiang, X., Cheng, L., Fan, D., Feng, Y*., et al.* (2014). Spider genomes provide insight into composition and evolution of venom and silk. Nature communications *5*, 3765.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nature biotechnology *26*, 1135-1145.

Skirvin, D.J., and de Courcy Williams, M. (1999). Differential effects of plant species on a mite pest (Tetranychus utricae) and its predator (Phytoseiulus persimilis): implications for biological control. Experimental & applied acarology *23*, 497-512.

Skirvin, D.J., and Fenlon, J.S. (2001). Plant species modifies the functional response of Phytoseiulus persimilis (Acari: Phytoseiidae) to Tetranychus urticae (Acari: Tetranychidae): implications for biological control. Bull Entomol Res *91*, 61-67.

Stein, L. (2001). Genome annotation: from sequence to biology. Nature reviews Genetics *2*, 493-503.

Strand, M.R., Rivers, D., and Grbic, M. (1997). Caste formation in the polyembryonic wasp Copidosoma floridanum (Hymenoptera: Encyrtidae): in vivo and in vitro analysis. Journal of insect physiology *43*, 553-565.

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. Current opinion in microbiology *11*, 472-477.

Tsagkarakou, A., Cros-Arteil, S., and Navajas, M. (2007). First record of the invasive mite Tetranychus evansi in Greece. Phytoparasitica *35*, 519-522.

Ullmann, A.J., Lima, C.M., Guerrero, F.D., Piesman, J., and Black, W.C.t. (2005). Genome size and organization in the blacklegged tick, Ixodes scapularis and the Southern cattle tick, Boophilus microplus. Insect molecular biology *14*, 217-222.

Van Leeuwen, T., Dermauw, W., Grbic, M., Tirry, L., and Feyereisen, R. (2013). Spider mite control and resistance management: does a genome help? Pest management science *69*, 156-159.

Van Leeuwen, T., Vontas, J., Tsagkarakou, A., Dermauw, W., and Tirry, L. (2010). Acaricide resistance mechanisms in the two-spotted spider mite Tetranychus urticae and other important Acari: a review. Insect biochemistry and molecular biology *40*, 563-572.

Varki, A., and Altheide, T.K. (2005). Comparing the human and chimpanzee genomes: searching for needles in a haystack. Genome research *15*, 1746-1758.

Vieira-Silva, S., and Rocha, E.P. (2010). The systemic imprint of growth and its uses in ecological (meta)genomics. PLoS genetics *6*, e1000808.

Walter, D.E. (2011). Invasive Mite Identification: Tools for Quarantine and Plant Protection.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P*., et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Willems, S., Fraiture, M.A., Deforce, D., De Keersmaecker, S.C., De Loose, M., Ruttink, T., Herman, P., Van Nieuwerburgh, F., and Roosens, N. (2016). Statistical framework for detection of genetically modified organisms based on Next Generation Sequencing. Food Chem *192*, 788-798.

Woods, J.L., James, D.G., Lee, J.C., Walsh, D.B., and Gent, D.H. (2014). Development of biological control of Tetranychus urticae (Acari: Tetranychidae) and Phorodon humuli (Hemiptera: Aphididae) in Oregon hop yards. J Econ Entomol *107*, 570-581.

Wu, R., and Taylor, E. (1971). Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. Journal of molecular biology *57*, 491-511.

Xia, Q., Zhou, Z., Lu, C., Cheng, D., Dai, F., Li, B., Zhao, P., Zha, X., Cheng, T., Chai, C*., et al.* (2004). A draft sequence for the genome of the domesticated silkworm (Bombyx mori). Science *306*, 1937-1940.

Xia, X. (2013). Comparative Genomics (Springer-Verlag Berlin Heidelberg).

Zhurov, V., Terzin, T., and Grbic, M. (2004). Early blastomere determines embryo proliferation and caste fate in a polyembryonic wasp. Nature *432*, 764-769.

Zhurov, V., Terzin, T., and Grbic, M. (2007). (In)discrete charm of the polyembryony: evolution of embryo cloning. Cell Mol Life Sci *64*, 2790-2798.

Zrzavý, J., and Štys, P. (1997). The basic body plan of arthropods: insights from evolutionary morphology and developmental biology. Journal of Evolutionary Biology *10: 353–367*.

Chapter 2

2 A beginner’s guide to NGS genome projects: from genome sequencing, assembly to annotation

A NGS genome project is a comprehensive and complex study, including project design, sample collection, DNA library preparation, sequencing, genome assembly, genome annotation, comparative genomics and downstream biological experimentation. The background and history of NGS have been elaborated in Chapter 1. Although high-throughput data is rapidly cumulating, analyzing these data is yet not a point-and-click process. Although state-of-the-art, genome assembly and genome annotation keep encountering practical challenges and theoretical issues. Therefore, this chapter aims to offer a best-practice toolkit to NGS beginners, especially to recruit bioinformaticians. Here we are using eukaryotic genomes and Illumina sequencing data to demonstrate a typical NGS genome project, focusing on genome sequencing, assembly algorithms and annotation methods. The workflow is indicated in Figure 10. I hope that beginners in this field can quickly grasp the essence of dealing with NGS data and find their way more smoothly and efficiently. The key terms in bold that used in this Chapter are listed in NGS TERM BOX section of this thesis.

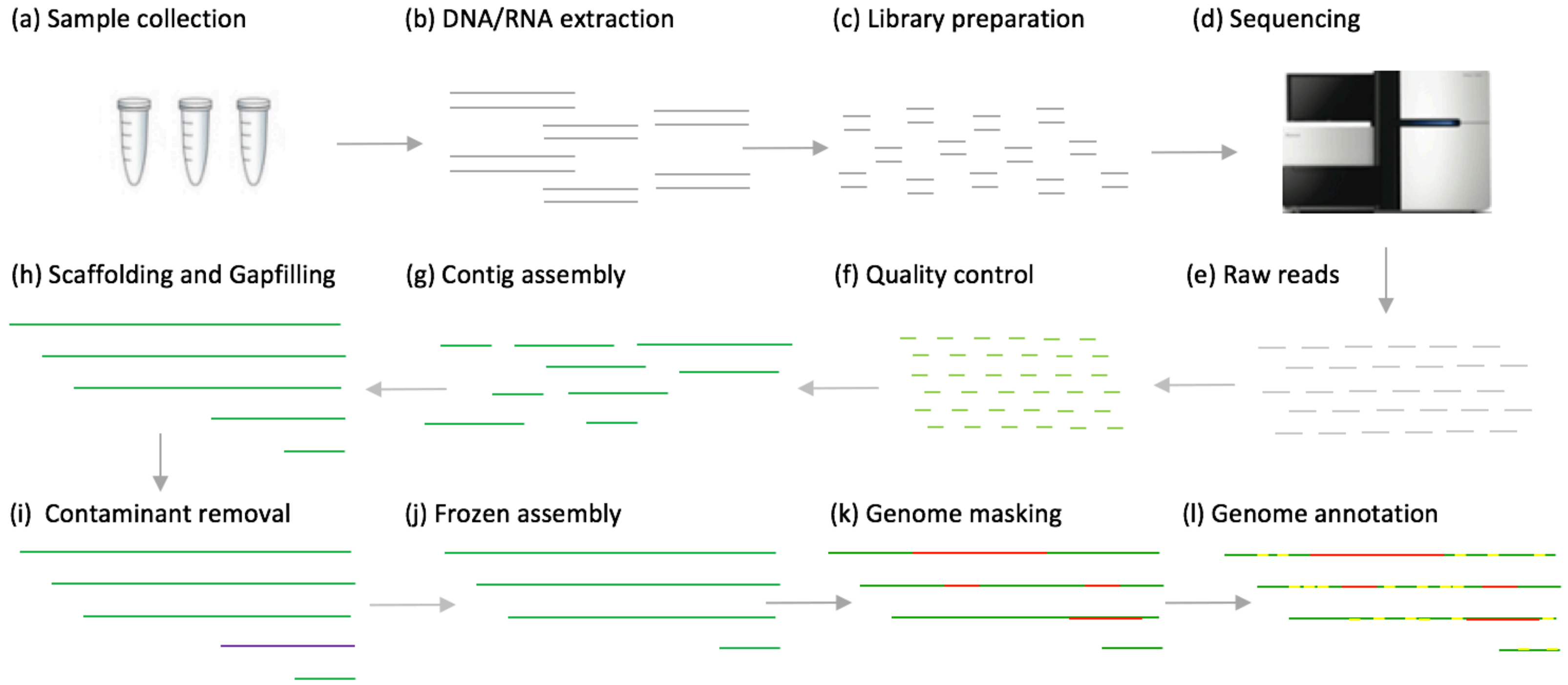


Figure 10: A typical NGS workflow from sample collection, assembly to annotation.
Gray line: sample DNA; green line: genomic sequence; purple line: contaminants; red line: repeats; yellow line: genes.

2.1 Project investigation

Before initiating a genome project, some concerns need to be considered and double-checked in terms of the applicability and reliability of the genome project.

2.1.1 Project objectives

These objectives might be of minor importance to bioinformaticians but are of major importance for the genome project, especially for the project leaders. These initial issues will determine subsequent sequencing methods, data quality, and outcomes. Sequencing method and reads coverage are important funding concerns in a genome project. High **depth coverage** and **breadth coverage** provide a more precise genome assembly but require high costs and many computational resources (Sims *et al.*, 2014). In terms of time, in practice, a relatively big eukaryotic genome project usually takes years, starting from experiment design, species generation inbreeding for heterozygosity purification, DNA sample extraction, sequencing, genome assembly, genome annotation to downstream analyses and further experiments. Each procedure might take months even years to be accomplished. Meanwhile, given that more and more genomes are being published, a good NGS genome publication requires not only a high quality of genome assembly but also additional wet laboratory experiments to validate hypotheses for the corresponding biological significance (Grbic *et al.*, 2011a; Olsen *et al.*, 2016). Otherwise, the paper of a draft assembly without further biological significance probably will be relegated from a high impact journal article into a “Genome Reports or Genome Announcements” (Smith, 2013, 2017).

2.1.2 Species survey and complexity

Basic biological information of the species needs to be understood. For instance, genome size, heterozygosity rate, GC-content, and repeat content will determine sequencing approaches.

Genome size estimation. It is important to estimate genome size because it can determine reads quantity to be obtained for sufficient sequencing coverage (Haridas *et al.*, 2011). In general, the bigger genome size is, the more reads are required. Genome size can be estimated by the following several methods. For bioinformaticians, we can apply **K-mer** estimation using the formula (Li and Waterman, 2003):

$$N=M*L/(L-K+1)$$

$$G=T/N$$

N is the depth of reads coverage; M is the average of K-mer coverage; L is the read length; K is the K-mer size; G is the genome size; T is the total number of bases. For biologists, a variety of methods such as flow cytometry, pulsed-field gel electrophoresis, PCR, Feulgen densitometry can be used for genome size estimation (Sun *et al.*, 2001; Wilhelm *et al.*, 2003; Rasch *et al.*, 2004; Pellicer and Leitch, 2014).

Homozygosity and heterozygosity. **Heterozygosity** is a cumulative result of genome mutation and hybridization. The rate will affect assembly complexity and accuracy. Higher rates of heterozygosity cause assembly problems because reads with SNPs are hard to be assigned back to the correct loci. It is difficult to obtain a well-purified sample species but it is possible to inbreed some short-life-cycle species in order to purify the rate of heterozygosity. Genomics reads from homozygous species are relatively easy to assemble because reads with fewer SNPs (i.e., low allele frequency) are more easily aligned with overlapping regions. However, current assembly tools cannot easily process high heterozygosity genome for a good assembly. In practice, when heterozygosity is less than 0.5%, it is extremely effective for downstream genome assembly, otherwise, it is recommended that deeper coverage or even, longer reads (e.g., PacBio SMRT reads) are deployed for the benefit of the assembly as well as genomic structural studies.

Haploidy and polyploidy. **Haploidy** as a single set of unpaired chromosomes, as in a germ cell, such as an egg or a sperm. **Polyploid** (including diploid) species have more than one set of chromosomes, such as the hexaploid wheat with three pairs of component genomes (A, B and D). Therefore, it is hard to assign reads to each component chromosomes because little information is provided for these reads to their own chromosomes (Brenchley *et al.*, 2012a).

The number of chromosomes. **Chromosome number** can provide information on determining how good the final scaffolds are; i.e., can we make the assembly to the chromosome-level? For example, it is suspicious if the species has ten chromosomes in reality, while the final assembly only has seven. In theory, the number of assembled scaffolds is supposed to be identical to the number of chromosomes. However, in practice, most of the published genomes are far away from their ideal chromosome number, except a few well-studied genomes such as the human genome, *Drosophila* genome, and *Arabidopsis* genome (Adams *et al.*, 2000; Arabidopsis Genome Initiative, 2000; Lander *et al.*, 2001).

Repetitive elements, also known as **repeats**, vary widely across different genomes. Almost half the human genome is represented by repetitive elements and the maize genome strikingly reaches up to 90% (SanMiguel *et al.*, 1996; Mills *et al.*). It is much more difficult to assemble reads from species with a high repetitive element composition by short Illumina reads because if a repetitive sequence is longer than a read, then coverage can never compensate and therefore, all copies of that sequence will produce gaps in the assembly (Schatz *et al.*, 2010).

Evolutionarily related genomes. A genome from a related species can help genome assembly as a mapping reference, without which, the assembly needs to start from scratch, known as ***de novo* assembly**.

Genomes can be categorized into two groups based on the genomic survey: regular genomes and complex genomes. Technically, there is no standard characterization for them. However, in general, **regular genomes** are referred as haploid or diploid genomes which have less 0.5% heterozygosity rate, less than 3Gb genome size, and 35%-65% GC-content. Examples of regular genomes are *Tetranychus urticae* (a small genome of 90Mb, low repeats, male haploid and female diploid) in animals and *Arabidopsis thaliana* (a small diploid genome of 135 Mb) in plants. Regular genomes require relatively less sequencing data and smaller libraries. **Complex genomes** can manifest different attributes - higher (>0.5%) heterozygosity rate, or GC-content less than 35% or greater than 65%, or repeat content higher than 50% or polyploid genomes. For instance, human genome (big genome size and high repeats), wheat genome (a big polyploid genome) and maize genome (a big sized genome with high repeats) are all considered as complex genomes. Because Illumina sequencing technology is sensitive to genomes with high GC-content, a complementary sequencing technology needs to be considered when encountering with high GC-content genomes. Complex genomes need both small libraries and big libraries (of different insert sizes) for assembly to deal with repeats issues (Green, 2001; Jurka *et al.*, 2007). Additionally, complex genomes also require higher reads coverage to detect SNPs and INDELs. Therefore, the above genomic survey is quite important to subsequent sequencing plans.

2.1.3 *De novo* sequencing and resequencing

***De novo* sequencing** is to assemble genomic reads into contigs and scaffolds in the absence of a reference genome, *i.e.*, the genome is assembled from scratch (Li *et al.*, 2010; Seo *et al.*, 2016b). This generally requires more and longer reads to generate a good assembly. *De novo* sequencing can be used to obtain new genomic sequence, identify genomic rearrangements and structural variations. **Resequencing** is to map genomic reads directly to a reference genome, skipping over the assembly process. Resequencing can improve genomic assembly, and investigate polymorphisms as well as genome structure variances (Rubin *et al.*, 2010). However, recent studies have revealed some unexpected drawbacks of resequencing such as failing to detect true genomic structure variations (Zapata *et al.*, 2016; Chen *et al.*, 2017).

2.1.4 Expertise and facility

Even though NGS is within the reach of small laboratories, no laboratory is the jack of all trades. A genome project involves various expertise and collaborations across a broad range of disciplines: taxonomists identify and categorize organisms; biologists collect samples and perform experiments; sequencing centers produce high-throughput genomic data; bioinformaticians assemble genomic reads, annotate genomes and analyze downstream data; IT scientists assist high-performance computational facilities. Therefore, it is indispensable to take diverse expertise and collaborations into consideration before starting a genome project.

2.2 Genome sequencing

Genome sequencing is the first crucial step once a genome project begins. Initially, biologists need to prepare DNA samples and send them to a sequencing center. However, to give NGS beginners in bioinformatics an impression of how genome sequencing works, here I will go over the essential steps from sample collection and DNA extraction, template library preparation and sequencing strategies.

2.2.1 Sample collection and DNA extraction

DNA extraction, also known as DNA isolation, is a process of purification of DNA from collected samples using physical and chemical methods. The methods of breaking cell wall (plants, fungi, and bacteria but not animals), cell membrane (using lysozyme), proteins (using protease) and RNAs (using RNase) are not identical. Note two kinds of DNA need to be clearly distinguished: gDNA is **chromosomal DNA** and it is distinct from **extrachromosomal DNA** (eDNA) such as plasmid DNA and mitochondrial DNA. Most genome projects require gDNA from the nucleus, which is easier because a good strong lysis to release the gDNA into solution is all that is required. A universal method of extracting gDNA is to purify proteins, RNA, reagents and other cell contents by cell lysis. A detailed introduction of how to identify and extract gDNA is described in (Dahm, 2008).

2.2.2 Library preparation

In general, there are four steps to prepare sample libraries: sample fragmentation, adapter addition, size selection and PCR (Van Dijk *et al.*, 2014; Simpson and Pop, 2015), as shown in Figure 11.

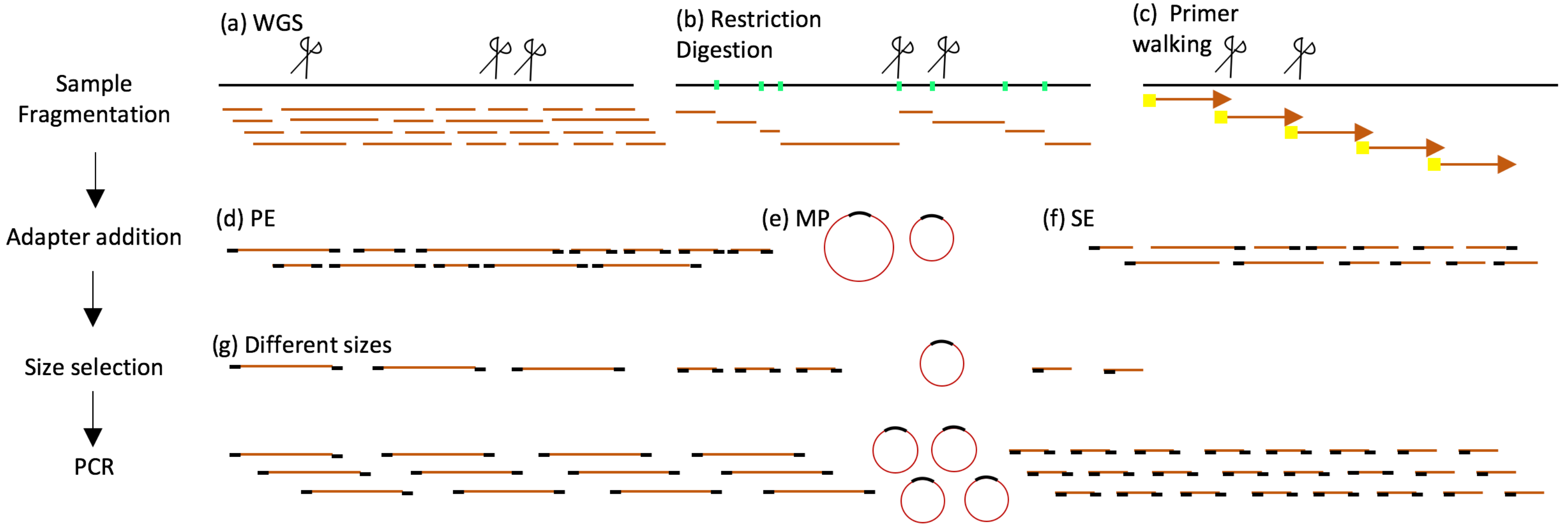


Figure 11: A workflow of sample library preparation.
Black line: DNA sequence; dark red line: DNA fragment; green spot: restriction enzyme digest spot; yellow spot: a primer; black block: adapter; PE: pair end, MP: mate pair, SE: single end.

2.2.3 Sample fragmentation

Long DNA samples are fragmented into smaller pieces because NGS technologies can barely handle longer pieces (700 bp from 454-sequencer is the longest reads in NGS and 100 bp-500 bp from Illumina). Fragmentation can be performed by one of three strategies (Figure 11a-c).

The first strategy, **Whole Genome Shotgun** (WGS), is the most widely applied strategy. It is quite simple and straightforward. The original DNA sequences are randomly sheared into smaller pieces for subsequent sequencing (Figure 11a). If not randomly but orderly sheared, the sheared sites are hard to be concatenated during assembly and thus more gaps will appear. If randomly, these cutting sites in theory always have other reads that cover these sites, which is quite useful for the subsequent assembly because there are more overlapping reads.

The second strategy is **enzyme restriction** to digest certain DNA sites and split the sequence into smaller fragments (Figure 11b). Sequencing starts at the terminus of these fragments. This strategy is quite applicable on small genomes and is feasible for assembly but it requires time and effort on digestion preparation.

The third strategy is **primer walking** (Figure 11c). Specific locations are primed using specific known sequence primers. Prerequisite sequences at the end of the reads permit the design of the subsequent sequence primer. Like restriction digestion, primer walking also requires elaborate preparations for primer and additional experimental design.

The latter two methods, enzyme restriction, and primer walking, have a big common drawback that either the digested loci or primer starting loci are hard to be concatenated because of lack of covered reads on these cutting sites. Therefore, they are more applicable on small genomes like bacteria rather than large eukaryotes, particularly used in resequencing, not ideally in *de novo* sequencing.

2.2.4 Adapters addition

After the samples are fragmented, the indexed adapters are added at the end of fragments (Figure 11d-f). Adapters are short single nucleotide sequences (>12 bases) for fixing DNA fragments on a solid surface by complementary tag sequences (e.g. bead-based, solid-state, or DNA nanoball). These index tags are like barcoding the samples so that multiple DNA libraries can be mixed tightly into one sequencing lane, known as **multiplexing**. Illumina provides three types of adapter addition: **Single End** (SE), **Pair End** (PE) and **Mate Pair** (MP), as shown in Figure 11d-f. SE is sequenced from only one end of a sequence fragment. PE consists of two reads (Read1 & Read2) connected by an insert of different size. The insert size usually is 100 bp-500 bp. Thanks to the inserts, PE and MP can provide an additional layer of evidence that can improve the quality of assembly. MP also has two reads but it needs a completely different preparation protocol using circularized molecules via internal adapter (Figure 11e), and it has two ends with a longer insert size 2 kb-20 kb, which is helpful in scaffolding because MP reads encompass larger continuous spans.

2.2.5 Size selection and PCR

Once the indexed adapters are added, fragment sizes are selected (Figure 11f). Depending on sequencing technologies and insert size, appropriate fragment sizes will be selected. Different types of insert sizes will benefit assembly because they provide more bridging information for the short contigs or reads. Finally, these fragments are amplified by PCR.

2.3 Sequencing technologies

Sequencing technologies have shown an extraordinary progress since the completion of human genome project (Goodwin et al., 2016). Table 2 shows a list of sequencing technologies with diverse features. The first sequencing generation, represented by precise, expensive but slow Sanger method, dominated sequencing market for over 30 years (Metzker, 2005). However, introduced at the beginning of the new century, NGS technologies have made tremendous progress in throughput, speed, capacity, accuracy, and expense per base (Goodwin et al., 2016).

The technology of 454 pyrosequencing was released in 2004 as the first NGS technology (Margulies *et al.*, 2005). Even though 454 pyrosequencing is still expensive and slow, it produces longer reads and higher throughput across all other NGS sequencing providers. Since 2005, several NGS technologies ensued such as SOLiD, Solexa and Illumina. Currently, NGS technologies are dominant by Illumina, represented by HiSeq series, MiSeq and very recent NextSeq and NovaSeq Series (a complete list can be seen at (Goodwin *et al.*, 2016)). Illumina provides various categories in SE, PE, and MP. Despite short reads, Illumina sequencing has been widely used due to its efficiency, cost-effectiveness per base and high throughput.

However, short reads have a severe limit. For instance, eukaryotic genomes with a high content of repetitive elements can fail to assemble well, because Illumina short reads are too short to distinguish repeats (i.e., reads are not longer than repeats). The longest NGS reads produced by 454 is 700 bp. Such length is sufficient for prokaryotic genome assembly but still difficult and insufficient for a eukaryotic genome with a high content of repeats.

Although NGS technologies are currently dominating the sequencing market, the Third-Generation Sequencing (TGS) technologies are lurking. TGS offers more potential for genome assembly, particularly for large genomes with a high proportion of repeat elements. **TGS technologies**, represented by PacBio SMRT cell and Oxford Nanopore, produce read length, on average, up to 15 kb (max 200 kb claimed by a MinION user) at the cost of over 10% error rate (Laver *et al.*, 2015). These high error rate (>15%) and high expense still prevent TGS to be extensively utilized. However, more and more genome projects are being performed using TGS data as complementary information for a better assembly (Gordon *et al.*, 2016b; Zapata *et al.*, 2016). For instance, using PacBio (SMRT sequencing) to build up scaffolds and then using high-quality Illumina data to correct low-quality bases, also known as **hybrid sequencing**.

Table 2: An overview of primary sequencing technologies.

Method	Generation	Read length	Accuracy	Reads per run	Time/run	Cost/Mb**	Advantages	Disadvantages
Pacific Biosciences RS SMRT	TGS	10,000 bp to 15,000 bp avg (14,000 bp N50); maximum read length >40,000 bases	87% single-read accuracy	50,000 per SMRT cell, or 500–1000 megabases	30 minutes to 4 hours	$0.13–$0.60	Longest read length. Fast. Detects 4mC, 5mC, 6mA.	Moderate throughput and expensive.
Oxford Nanopore Sequencing	TGS	Dependent on library prep, not the device (up to 20 kb ***)	~92–97% single read (up to 99.96% consensus)	dependent on read length selected by user	data streamed in real time. 1 min to 48 hrs	$500–999 per Flow Cell	Very long reads, Portable (Palm sized)	Lower throughput and high error rates.
Ion semiconductor (Ion Torrent sequencing)	NGS/TGS	up to 400 bp	98%	up to 80 million	2 hours	$1	Less expensive, fast.	Homopolymer errors.
Pyrosequencing (454)	NGS	700 bp	99.90%	1 million	24 hours	$10	Long read size. Fast.	Runs are expensive. Homopolymer errors.
Sequencing by synthesis (Illumina)	NGS	50 bp ~ 500 bp #	99.9% (Phred30)	1 million ~ 3 billion ##	in days, depending upon sequencer and read length	$0.05 to $0.15	High output and cheap, applicable for big genomes.	Equipment can be very expensive. Requires high concentrations of DNA.
Sequencing by ligation (SOLiD sequencing)	NGS	50+35 or 50+50 bp	99.90%	1.2 to 1.4 billion	1 to 2 weeks	$0.13	Low cost per base.	Slower and issues with palindromic sequences.
Chain termination (Sanger sequencing)	1st	400 to 900 bp	99.90%	N/A	20 minutes to 3 hours	$2,400	Long individual reads. Useful for many applications.	Expensive and impractical for big genomes. Time-consuming.

Notes for Table 2 (previous page):

*this table and data are modified from https://en.wikipedia.org/wiki/DNA_sequencing and (Liu *et al.*, 2012; Quail *et al.*, 2012)(Escalona et al., 2016; Goodwin et al., 2016; Mardis, 2008; Nagarajan and Pop, 2013)

** in US dollars

***the longest read reported by a MinlON user, accessed on August 5th, 2016 at https://www.nanoporetech.com/

# MiniSeq, NextSeq: 75-300 bp; MiSeq: 50-600 bp; HiSeq 2500: 50-500 bp; HiSeq 3/4000: 50-300 bp; HiSeq X: 300 bp

## MiniSeq/MiSeq: 1-25 Million; NextSeq: 130-00 Million, HiSeq 2500: 300 million - 2 billion, HiSeq 3/4000 2.5 billion, HiSeq X: 3 billion

2.4 Genome assembly

I have discussed NGS prerequisite knowledge and sequencing technologies. Hereafter, the second key step for a genome project: genome assembly *in silico*. In practice, bioinformaticians anticipate the best possible reads data as a start for genome assembly, including quality control, contig assembly, scaffold assembly (scaffolding), gap filling and contaminant removal (Figure 10e-j). Here I use *de novo* genome assembly to demonstrate the workflow.

2.4.1 Quality Control (QC)

QC, or clean-up low-quality bases, is the first step for genome assembly. Despite Illumina having an error rate of less than 0.1%, QC is still indispensable because raw sequencing reads contain remnant adapters and low-quality bases (particularly at the both ends of a read due to the imaging sensor is not stable at these two stages). Popular QC tools, including FastQC and fastq_screen (http://www.bioinformatics.babraham.ac.uk), can identify low-quality nucleotides, which can be chopped off by Trimmomatic (designed for Illumina NGS data) (Bolger *et al.*, 2014b).

2.4.2 Assembly algorithms

The development of sequencing technologies subsequently brought four major assembly algorithms - one conventional approach **Greedy** and three graph-based approaches **Overlap-Layout-Consensus** (OLC), **DeBruijn Graph** (DBG) and **String**, respectively (Miller *et al.*, 2010; Nagarajan and Pop, 2013). Choosing an appropriate assembly algorithm is based on reads type and computational competency. For example, OLC and Greedy were originally designed for long reads (e.g., Sanger reads and 454 reads) while DBG is more appropriate for short reads (e.g. Illumina reads). Here I demonstrate the essentials of the four algorithms (Figure 13).

Greedy, represented by PHRAP, TIGR assembler, and CAP3 toolkits, is an initial assembly algorithm designed for assembling Sanger reads (Table 3). This algorithm seeks overlapped consensus regions and extends the sequence length. It has a good approximation and simplicity but bad performance on large repeats (Huang and Madan, 1999; Zhang *et al.*, 2000; de la Bastide and McCombie, 2007). Briefly, the Greedy algorithm first sets the longest read as an initial contig, and then merges itself with another overlapping reads/contig, which needs to have the largest overlap with the initial one, into a new contig. Repeat the same procedures above until all contigs are exhaustive. In sum, the greedy algorithm works well on long reads. However, it is incapable to handle repeats because of infinite loops that are generated by assembling identical repetitive reads.

OLC algorithm was proposed by Staden in the 1980s and subsequently improved in the past two decades (Staden, 1980). OLC was primarily used for Sanger assemblies by overlapping all detected reads in pairwise and concatenating overlapping reads iteratively until no overlapping reads can be found (Li *et al.*, 2012). Briefly, OLC requires an all-against-all comparison of reads with three steps: first, identify all pairs of reads that overlap sufficiently; second, layout reads that align to each other and organize them into a graph; third, construct a consensus by concatenating reads. OLC has a great performance on small genomes while it is still difficult to generate overlapping graphs with highly repetitive sequences. Representative assembly tools for OLC are Arachne, Celera assembler (updated continuously), Newbler, Minimus (Myers *et al.*, 2000; Batzoglou *et al.*, 2002; Sommer *et al.*, 2007).

DBG, invented with the emergence of short reads sequencers (Illumina), is primarily designed for short reads assembly. The DBG algorithm initially corrects reads errors and then cut them into **K-mer**. K-mer is a trade-off between specificity and sensitivity of genome assembly. A large K-mer is good for assembly specificity but might be resulted in short scaffolds. Small K-mer offers higher sensitivity by joining more fragments but may fail to resolve suspicious overlaps and result in more genomic gaps. These chopped K-mer fragments are listed in a path graph. The DBG algorithm chooses the best path to walk through most the reads. Then DBG builds up the K-mer hash table, and then tracks the graph by overlaps and finally walks the path through the table to generate the assembly. Alike to OLC, DBG also extends fragments exclusively. The resolvable reads (fragments) are assembled as contigs while the left unresolvable reads (particularly repeats) are being left out, then broken into fragments and reassembled again (Treangen and Salzberg, 2012). It is necessary to run several K-mer trials to compare the potential assemblies, suggested K-mer size can be from 27 to 63 (usually it is an odd number, reason details in NGS TERM BOX). DBG requests high computational resources because of the vast number of K-mer strings. Widely used DBG assemblers are ABySS, Velvet, SOAPdenovo2, AllPaths, ClC_assembler (a commercial tool at http://www.clcbio.com/) (Zerbino and Birney, 2008; Maccallum *et al.*, 2009; Simpson *et al.*, 2009; Luo *et al.*, 2012).

String Graph is a recently developed memory-efficient algorithm that operated by removing contained reads and transitive edges (Myers, 2005). In brief, it uses a compressed representation of DNA sequence reads to calculate per-base error rates, insert distributions and coverage metrics in the absence of a reference genome. Meanwhile, it estimates genome features such as repeats and heterozygosity. Using compressed reads structures from string graph, Edena Assembler and String Graph Assembler (very memory-efficient) are efficient tools for large genome assembly (Hernandez *et al.*, 2008; Simpson and Durbin, 2012).

In summary, assemblers of the same algorithm usually have similar procedures. For example, OLC and DBG assemblers construct a graph and reduce non-intersecting paths. They collapse polymorphism-induced fragments, tangle simplification and finally convert paths into contigs. DBG and string graph assemblers detect and correct errors before splitting into K-mer. Again, choosing an appropriate assembler mainly depends on read types. OLC assemblers are more suitable for longer reads such as Sanger reads, 454, TGS reads. DBG assemblers can better handle short reads because there is no requirement for the long reads information.

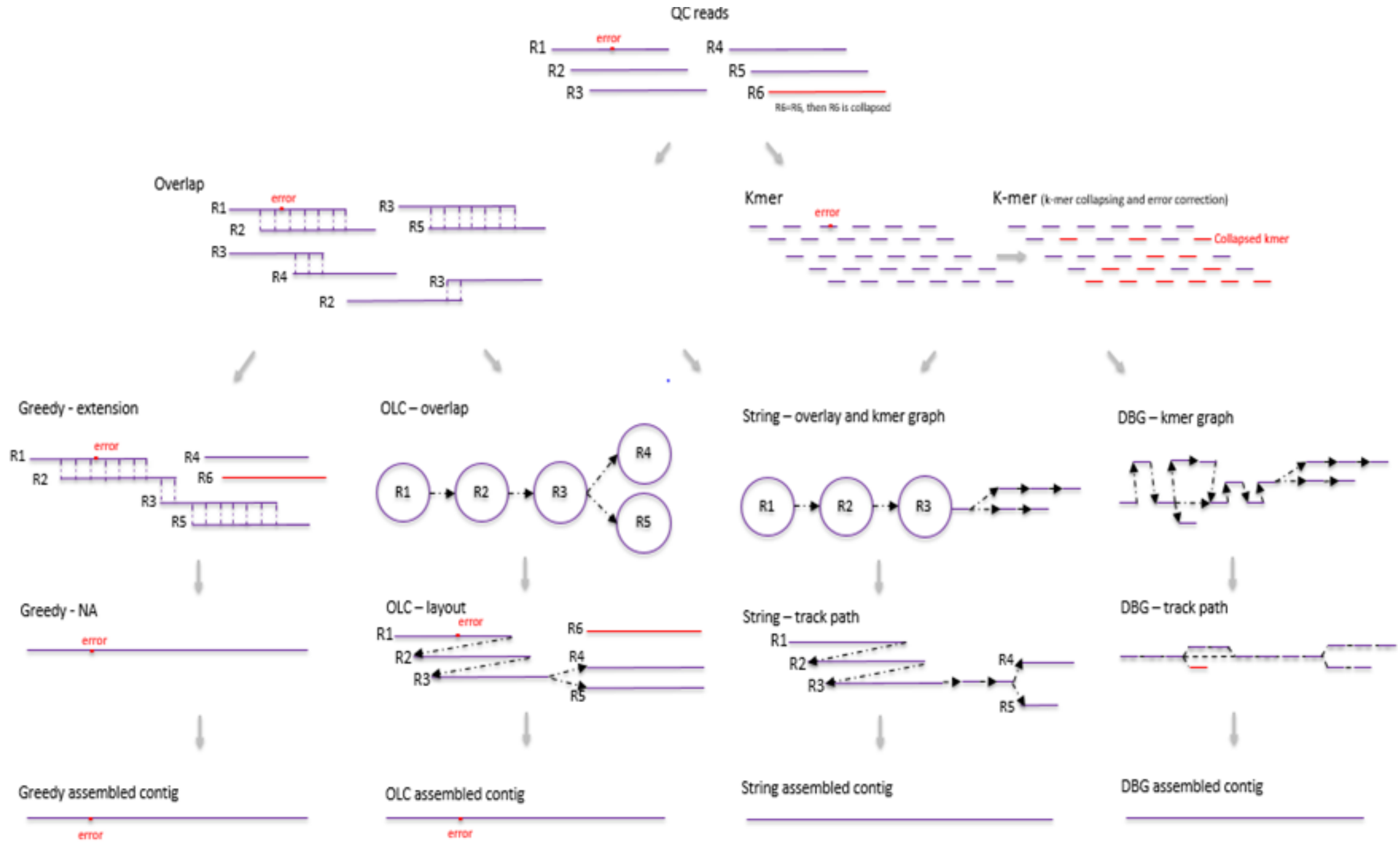


Figure 12: Four assembly approaches in NGS.
Long purple lines represent reads, short purple lines represent read fragments, black dash line with arrows represent connections between reads; R means reads; red dot lines represent alternative connections; red dot is error base; black dash line without arrows represent overlapped regions (Chaisson *et al.*, 2015b; Simpson and Pop, 2015).

Table 3: Tools for genome assembly.

	Tools	***Remarks***	***Citation or Website***
Read quality reporter	FastQC	Check reads quality and visualization	http://www.bioinformatics.babraham.ac.uk
	Fastq_screen	Check reads quality and screen a library of sequences into FastQ format	http://www.bioinformatics.babraham.ac.uk
	NGS QC Toolkit	Check reads quality and screen high-quality data	http://www.nipgr.res.in/ngsqctoolkit.html
Quality Control	Trimomatic	Chop off low-quality nucleotides	http://www.usadellab.org/cms/index.php?page=trimmomatic (Bolger *et al.*, 2014a)
	ngsShoRT	Pre-process SE/PE/MP reads in FastQ format or Illumina's native QSEQ format	http://research.bioinformatics.udel.edu/genomics/ngsShoRT/
	bbduk	Trim and filter adapters. Fast, scalable, and memory-efficient	http://www.geneious.com/plugins/bbduk
	Fastq_quality_trimmer	Trim sequences based on quality	http://hannonlab.cshl.edu/fastx_toolkit/
DNA or RNA Assembler	Phrap/Phred	Use greedy method to assembly WGS data, especially for Sanger reads, produce long contigs,	http://www.phrap.com/
	Oases	Assemble transcripts in absence of a reference genome	http://www.ebi.ac.uk/~zerbino/oases/
	CAP3	Based on error auto-correction, easy to use in scaffolding	http://doua.prabi.fr/software/cap3 (Huang and Madan, 1999)
	String Graph Assembler	Use string graph to assemble a genome. Memory efficient	https://github.com/jts/sga
	Edena v3	Use string graph, fast	
	Cufflinks	Assemble transcripts using RNAseq data without a reference genome	http://cole-trapnell-lab.github.io/cufflinks/
	Celera Assembler	Originally designed for Sanger reads, a *de novo* WGS assembler, also it supports NGS hybrid assemblies	http://wgs-assembler.sourceforge.net/wiki/index.php?title=Main_Page
	AbySS	Assemble short genomic sequences using DBG	(Simpson *et al.*, 2009)
	SOAPdenovo2	A *de novo* draft assembler for large genomes using DBG, especially for Illumina reads, but requires huge memory, very flexible	(Luo *et al.*, 2012)
	Arachne	Designed for long reads using OLC algorithm, good performance on assembling many genomes with large and highly repetitive.	(Batzoglou *et al.*, 2002)
	CLC assembler	Commercial tool, high-performance on *de novo* assembling of NGS data, faster than SOAP*denovo*	http://www.clcbio.com/products/clc-assembly-cell/

	Minimus	A fast assembler using lightweight memory, good for small genomes	(Sommer *et al.*, 2007)
	Newbler	A *De novo* assembler for 454 data (or other pyrosequencing data)	www.my454.com
	Trinity	Illumina RNAseq data assembler for transcripts	https://github.com/trinityrnaseq/trinityrnaseq/wiki (Haas *et al.*, 2013)
	Trans-Abyss	*De novo* assembly of RNA-Seq data and generate fragmented transcriptomes	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss (Simpson *et al.*, 2009)
	AllPaths(LG)	Use DBG algorithm to assemble both DNA and RNA short reads (not Sanger or 454 reads or mixed), good performance on large genomes. It requires high coverage, easy to run incompatible libraries	(Butler *et al.*, 2008; Maccallum *et al.*, 2009)
	IDBA-UP	A *de novo* assembler for single-cell and metagenomic sequencing data	(Peng *et al.*, 2012)
	Velvet	Use DBG and leverage very short reads in combination with PE	(Zerbino and Birney, 2008)
	SGA	OLC assembler for large genomes	(Simpson and Durbin, 2012)
Aligner	HISAT2	RNAseq data mapper/aligner, fast speed	(Kim *et al.*, 2015)
	TopHat2	Align RNAseq data to a reference genome, slower than HISAT2	(Kim *et al.*, 2013)
	SOAPaligner	Align reads to a *de novo* assembly, check breadth and depth coverage	(Li *et al.*, 2009b)
	MUMMER	A fast alignment toolkit for large genomes, especially for genome-wide alignment	(Kurtz *et al.*, 2004)
	Bowtie2	An ultrafast, memory-efficient short read aligner, little memory, applicable to large genomes	(Langmead and Salzberg, 2012)
	BWA	Map low-divergent sequences to a large reference genome, fast and accurate	(Li and Durbin, 2009)
Gap filling	Gapfiller	Close gaps within pre-assembled scaffolds using NGS PE data	(Nadalin *et al.*, 2012b)
	GapCloser	Designed to fill the gaps of SOAP*denovo* assembly	http://soap.genomics.org.cn/about.html
	Sealer	Close gaps within assembly scaffolds by navigating DBG paths	(Paulino *et al.*, 2015)

2.4.3 Scaffolding and gapfilling

Scaffolding is a process of assembling contigs into scaffolds (Figure 10h). By using PE, MP or long-reads as bridging information, contigs are sorted and concatenated into scaffolds with “N” as bridges. The number of consecutive ‘N’ suggests estimated gap size between two joint contigs. Different scaffolding tools can tackle different genomic data. For example, SSPACE is a stand-alone scaffolding tool using PE reads as input while SSPACE_LG is designed as a hybrid assembly tool using PacBio long reads (Boetzer *et al.*, 2011; Boetzer and Pirovano, 2014). SSPACE_LG can assess the order, distance, and orientation of contigs and subsequently concatenate these contigs into scaffolds. Another example is ALLPATHS-LG that can solve repeats problem by modeling MP reads and concatenate contigs (Gnerre *et al.*, 2011).

In additional to PE, MP and TGS data, some other methods can also be applied in scaffolding. **Optical mapping** is a recently advanced technology in improving genome scaffolding assembly (Neely *et al.*, 2011; Mendelowitz and Pop, 2014), particularly for high repeat content genomes (Kawahara *et al.*, 2013; Luo *et al.*, 2016). Optical mapping sorts relative contigs in order for subsequent scaffolding (Chen *et al.*, 2013). The principle is simple: to align the visualized beam spot patterns that were inserted in the sequences. Current leading optical mapping providers are OpGen and BioNano. They employ similar methods except OpGen uses double-stranded DNA while BioNano uses single-stranded DNA. These marked DNA sequences go through a nanochannel for visualization and alignment. Optical mapping has been readily applied across a wide range of organisms for improving their genome assembly (Perry *et al.*, 2011; Dong *et al.*, 2013; Kawahara *et al.*, 2013). **Chromosome Conformation Capture** (3C) is another latest technique to improve genome assemblies by mapping to derive the linear order of sequences across the pericentromeric space and to investigate the spatial organization of chromatin in the nucleus at megabase resolution. Because it also has a high requirement on DNA integrity thus it works better on animal genomes rather than plants because breaking up plant cell walls has a negative effect on DNA integrity (Van Berkum *et al.*, 2010; Burton *et al.*, 2013; Korbel and Lee, 2013). However, a recent study of the high-quality barley genome was successfully assembled through this technique (Mascher *et al.*, 2017).

Gapfilling, or gap closing, is to fill gaps (N, not A, T, C or G) across all scaffolds. It initially aligns genomic reads to scaffolds and replaces 'N's by informative nucleotides (A, G, C or T). Normally, several iterations are required before most of the gaps in all the scaffolds are filled. Genomes with many gaps would affect gene structural annotation, leading an intact gene to a truncated gene caused by gaps. Popular scaffolding tools are, for instance, GapFiller and Sealer (Nadalin *et al.*, 2012b; Paulino *et al.*, 2015).

2.4.4 Resequencing assembly

Resequencing assembly is relatively easy because reads are directly mapped to a reference genome to form a new assembly (Martin and Wang, 2011). However, a risk of resequencing assembly is engendered by genome structures such as gene translocations or transpositions. Because, in most cases, the aim of resequencing a genome is normally to seek for structural variations such as INDELs and SNPs (Xia *et al.*, 2009), it is a challenge to do resequencing assembly by mapping genomic reads to a reference genome. Theoretically, it is necessary to deep resequence a genome for a greater coverage and subsequently, assemble the genome from scratch (i.e., *de novo*). As for RNAseq/transcriptome assembly, it also depends on the availability and quality of the reference genome assembly. If the reference is available, it is feasible to align reads and build up alternative splicing graph. Popular mapping tools are TopHat2, Bowtie and a recently published faster tool HISAT2 (Langmead, 2010; Kim *et al.*, 2013; Kim *et al.*, 2015). However, if a reference genome is absent, tools like CuffLinks, AbySS, SOAPdenovo2, and Trinity can assembly transcriptome from scratch (Simpson *et al.*, 2009; Grabherr *et al.*, 2011; Luo *et al.*, 2012; Trapnell *et al.*, 2012).

2.4.5 Contaminants identification

It is indispensable to remove contaminants after scaffolding (Figure 10i). This is because collected DNA samples or tissues were possibly contaminated: in most cases, eukaryotic species carry bacteria within their body (e.g., in guts, skins or even endosymbionts). Contaminants cannot be checked by reads QC because raw reads are too short to be distinguished whether they are contamination.

In practice, most prokaryotic reads (mainly bacterial reads rather than archaeal reads) will be assembled into contigs even scaffolds. These bacteria scaffolds are easy to be identified using BLAST against prokaryotic or bacterial databases. Scaffolds that were assembled from prokaryotic reads usually have bizarre reads coverages. In general, regions that have coverage twice higher or lower than the normal coverage need to be set up an alert flag for further validation. Therefore, by checking reads coverage of all scaffolds can leave out these contaminants with bizarre coverages.

However, genes from **horizontal gene transfer** (HGT) events are not supposed to be identified as contaminants. HGT genes are bacterial genes scattered in eukaryotic genomes. They are not only an important contributor in genome evolution but also have a big influence on adaptation and behavior of related eukaryotes (Raymond and Blankenship, 2003; Keeling and Palmer, 2008; Monier *et al.*, 2009; Grbic *et al.*, 2011a; Soanes and Richards, 2014). Scaffolds with both prokaryotic and eukaryotic BLAST hits can either be HGT genes, or contaminants or even assembly errors. Thus, they must be scrutinized. PCR, **Bacterial Artificial Chromosomes** (BACs), or TGS long reads can be used to validate such regions (Figure 12). In addition, these ambiguous scaffolds that contain BLAST hits from both eukaryotes and prokaryotes can also be chopped into small pieces (e.g. 2.5 kb) and then double-checked against the BLAST hits of each piece at a smaller scale. Be very cautious of these scaffolds, otherwise, there will be strong impacts on the conclusion drawn from the results (Boothby *et al.*, 2015; Koutsovoulos *et al.*, 2016).

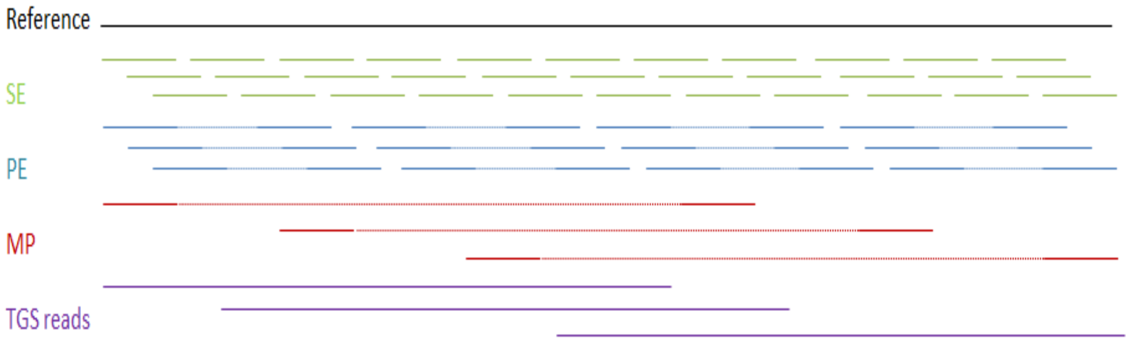


Figure 13: Four types of libraries including SE, PE, MP, and TGS reads.
This figure demonstrates different NGS reads including SE, PE and MP mapping on a reference sequence. TGS long genomic reads are also mapped to this reference sequence.

2.4.6 Assembly quality assessment

Genome assembly usually reaches up to three levels: draft genome, complete genome and finished genome. **Draft genome** is at some points useful to perform certain analyses, even though it possibly has short **scaffold N50** (N50 size of contigs or scaffolds was calculated by sorting all sequences and then adding the lengths from the longest to the shortest until the summed length exceeded 50% of the total length of all sequences.) and low genome coverage (Bos *et al.*, 2011; Ling *et al.*, 2013; Sanggaard *et al.*, 2014). However, it must meet the minimum submission requirement to a public database (Chain *et al.*, 2009). **Complete genome**, despite a few gaps, usually reflects high genome coverage (>90%) with high accuracy and long N50. These complete genomes usually have a completely continuous representation and no further sequencing needs to be done in such cases in spider mite and Neanderthal genomes (Grbic *et al.*, 2011a; Prufer *et al.*, 2014). **Finished genome** has a complete coverage (>99%) and each base in the genome has a very high quality (Yandell and Ence, 2012). Model organisms are usually finished genomes for gene model building and other precise studies (Collado-Vides *et al.*, 2003; Rogers, 2003).

The quality of an assembly is a milestone for a genome project and thus to assess the quality is a must. Unfortunately, there is neither a clear boundary across the three assembly levels nor a gold standard to validate assembly quality. Evaluation becomes more difficult when most state-of-the-art genome assemblies are non-trivial. Current genome papers preferably use scaffolds N50 to demonstrate the quality of an assembly. For instance, if N50 is longer than median gene length, it means at least half of genes from the whole genome are located on a single scaffold. Otherwise, it hints more than half of the genes are truncated because of short scaffolds and thus it is hard to perform subsequent analyses because of the gene truncation issue. In the case that N50 is smaller than median length, it is recommended to sequence more reads data for better assembly before downstream analyses.

However, longer scaffolds are not a determining factor to a good assembly. Accuracy, contiguity, and completeness can also indicate the quality of an assembly (Li *et al.*, 2010; Lee *et al.*, 2016). First, **genome completeness**, to some extent, is indicated by the difference between assembled genome size and actual estimated genome size (Li *et al.*, 2010). Second, **genome accuracy** implies nucleotide resolution at every base. High depth coverage provides more weight in base accuracy to clearly distinguish SNPs in population genomics and genome-wide association studies (Ledford, 2008; Romiguier *et al.*, 2014; Birney and Soranzo, 2015). **Genome contiguity**, the third indicator, represents the presentation of the scaffolds. The length and order of scaffolds are important to contiguity in genome structure and variation studies (Marchini and Howie, 2010).

To apply the three measurements in genome assembly assessment, normally it is good to use longer sequences as assistant evidence. BACs are regularly used to assess the large-scale and local assembly accuracy. Long BAC can be aligned back to assembled scaffolds to see whether the assembly has obvious errors through misassembly (Li *et al.*, 2010). Similar to the BACs, resequencing on a small region of interest using MP, PE and TGS is applicable as well, very similar to the previously mentioned approach of searching contaminants.

These methods of assessing assembly quality are all quite straightforward. Briefly, first use mapping tool (e.g. BWA or CLC_mapper) to align all genomic short reads back to assembled scaffolds, and then calculate reads depth coverage (by Bedtools kit) for all scaffolds at a flexible window size (e.g., 10 kb, a lower number will offer a better resolution but may offset the odds of appearance of contaminants) (Quinlan and Hall, 2010). Collapsed reads (regions with much higher coverage) and gap regions (including regions with much lower coverage) on each scaffold can be easily visualized (Figure 14a, b, d, and e). These non-average-coverage regions, i.e., regions with higher/lower coverage or even gaps need to be alert and inspected using long sequences as supplementary evidence (Figure 14, long reads, MP or PE). Figure 14a is a typical gap and no read coverage is found in this region. The long reads also show it is a gap and MP reads confirms that the scaffolding in this region is correct. Figure 14b is a low coverage region but MP reads and long reads have validated its accuracy. Figure 14c, however, is an error assembled loci because no long reads can be aligned and MP evidence clearly shows the two fragments need to be switched. Figure 14d-e are high repeated regions. It is easy to collapse short reads in Figure 14d and thus lead Figure 14e to a lower coverage region.

Evidence from both long reads and MP reads shows this assembly is correct despite Figure 14d has an extremely high bar and Figure 14e is extremely low.

In addition, there are some databases designed to assess assembly by searching certain sequences. For example, BUSCO can assess genome assembly completeness with benchmarking universal single-copy orthologs (Simao *et al.*, 2015). The regions (gaps, high-coverage or low-coverage region) of the non-average-coverage need to be double-checked as well. These regions are possibly caused by repetitive sequences or even contaminated sequences from other species (Figure 14d-e).

2.5 Genome annotation

Genome assembly is worthless if it cannot be deciphered and interpreted; therefore, efforts to describe, or 'annotate', genome annotation begins as soon as a **frozen assembly** (no more assembly required and set up as a final assembly) becomes available (Mudge and Harrow, 2016). Gene features and functions in a genome are essential questions in genome projects. **Genome annotation** (also called computational gene prediction, gene-building, gene-calling) is a process of searching gene models *in silico* in a well-assembled genome and predicting these gene model functions, which will be propagated into downstream analyses. Genome annotation mainly includes genome masking, structural prediction, functional prediction, manual curation, genome update, and database maintenance.

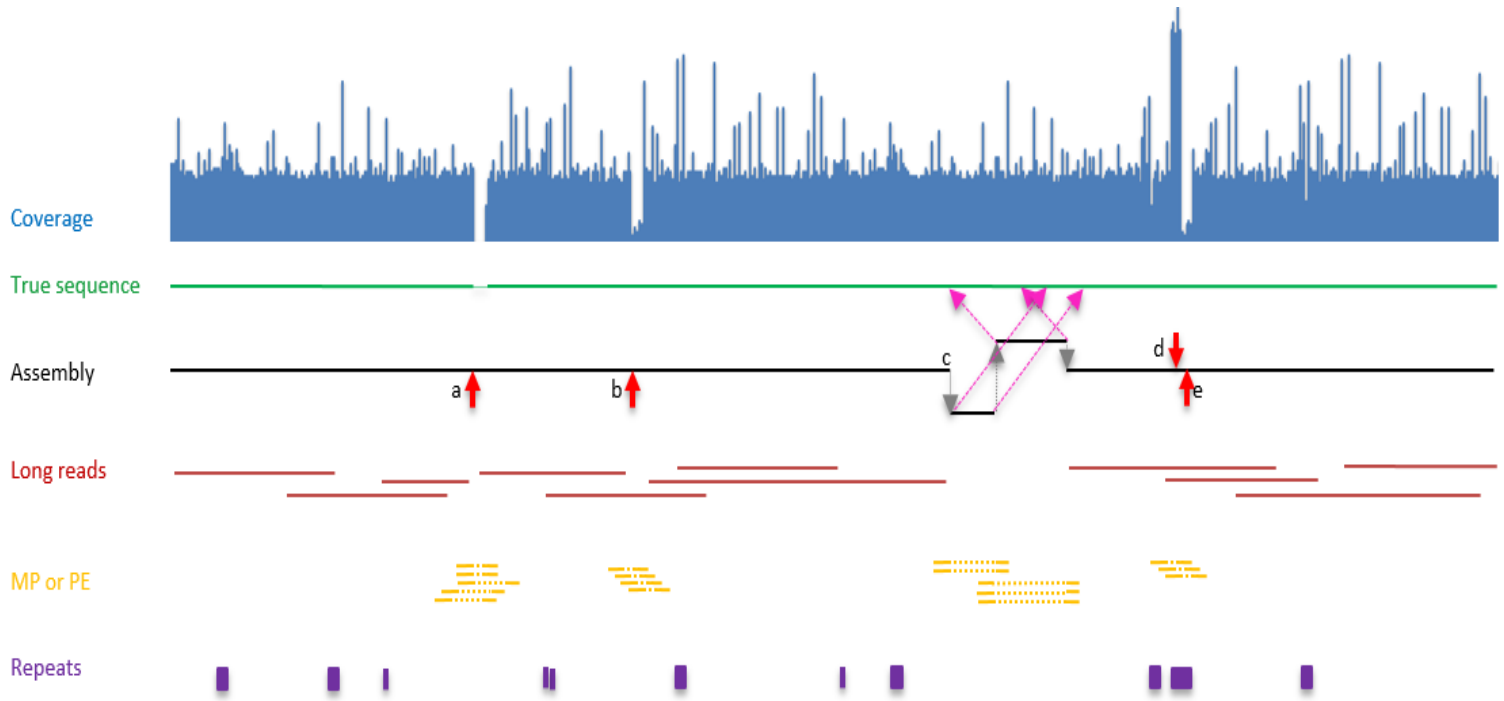


Figure 14: Genome assembly assessment using coverage and long reads.
(a) a real gap; (b) and (e) low coverage; (c) an assembly error - the two fragments need to be switched; (d) high coverage.

2.5.1 Genome masking

Genome masking, or masking repetitive sequences, is an initial step for genome annotation (Figure 10k). **Repetitive sequences** are usually poorly conserved and have a huge impact on genome structure and size (Feschotte *et al.*, 2009). This is because some repeats like transposons can jump over along with flanking genes by cut-and-paste or copy-and-paste mechanisms. Normally, these repeats represent themselves as certain patterns of nucleic acids in multiple copies dispersed across a genome. If the genome is masked inappropriately, protein-coding genes would possibly be annotated in these repetitive regions. Repeats can be categorized into three types: terminal repeats, tandem repeats (including microsatellite, minisatellite and satellite DNA) and interspersed repeats (or **transposable elements** (TE), consists of DNA transposons and retrotransposons). Each type of repeats presents diverse proportions in different genomes. Large eukaryotic genomes often consist of the high content of repetitive elements. For instance, human and maize genomes have about 50% and 90% repetitive elements, respectively (Lander *et al.*, 2001; Zhou *et al.*, 2009).

Genome masking includes hard masking and soft masking. **Hard masking** is known as transforming each nucleotide in repeat region into an ‘N’. **Soft masking** can transform these regions into low case letters a, c, g or t. Soft masking is more sequence-friendly and these masked regions can be easily traced back. Hard masking removes the sequence information and makes no difference between repeated regions and gaps (e.g., they are all ‘N’s). Hence, soft masking is preferable in state-of-the-art genome masking method. Once a genome is masked, these masked regions will be skipped during annotation, which means no genes are supposed to be predicted on these masked loci.

Table 4: Tools for repeats identification and genome masking.

Tools	***Remarks***	***Citation or website***
RepeatMasker	Repeat searching tool	http://www.repeatmasker.org/ (Tarailo-Graovac and Chen, 2009)
RepeatModeler	Build repeat library	http://www.repeatmasker.org/ (Smit and Hubley, 2008-2015; Tarailo-Graovac and Chen, 2009)
Piler	Identify and classify repeats	http://www.drive5.com/piler/ (Edgar and Myers, 2005)
LTR_FINDER	Find full-length LTR retrotransposons	http://tlife.fudan.edu.cn/ltr_finder/ (Xu and Wang, 2007)
TRF	A public database of tandem repeats	http://tandem.bu.edu/trf/trf.html (Benson, 1999)
DUST	Mask low-complexity sequences using BLAST	(Morgulis *et al.*, 2006)
LTR_STRUC	Find LTR retrotransposon structure	http://www.mcdonaldlab.biology.gatech.edu/ltr_struc.html (McCarthy and McDonald, 2003)
LTR_harvest	*De novo* detection of full-length LTR	(Ellinghaus *et al.*, 2008)
RepeatScout	*De novo* detection of repeat families in large genomes	http://bix.ucsd.edu/repeatscout/ (Price *et al.*, 2005)
RepeatRunner	A CGL-based tool integrates RepeatMasker and BLASTX	http://www.yandell-lab.org/software/repeatrunner.html
REPET pipELine	Two main pipelines (Tedenovo and Teannot) for finding repeats	https://urgi.versailles.inra.fr/Tools/REPET

There are two basic approaches for genome masking: *De novo* and homology-based. ***De novo*** **masking** is to mask a genome from scratch. It requires a good repeat library. RepeatModeler and RepeatScout are widely used to build up the *de novo* library by using consensus sequences (Price *et al.*, 2005). Because *de novo* library might include protein-coding genes or transposons sequences, thus after a draft repeats library is built, it is necessary to filter out protein-coding genes (e.g., using UniProt Database). Then employ REPCLASS or RepBase to assign these anonymous sequences TE categories because sequences in the repeat library are anonymous (Feschotte *et al.*, 2009). When a TE library is finalized, repeat element masker tools such as RepeatMasker can mask these repeated regions (Tarailo-Graovac and Chen, 2009).

Compared with *de novo* prediction, the **homology-based prediction** is relatively easy. It skips the step of library building and directly uses a built-up repeat library to mask consensus sequences across the whole genome. Be advised that once a genome is masked, it is a must to inspect if these masked regions have been overlapping with RNAseq or transcriptome. If overlapped, it hints this genome is probably over-masked. Given that inappropriate masking (**over-masking** or **under-masking**) might lead to the failure of gene prediction in the masked regions, therefore, be cautious of repeats even after genome annotation is finished. Here I list commonly used tools for genome masking in Table 4 and show a comprehensive method to build up a good *de novo* genome TE library in Figure 15.

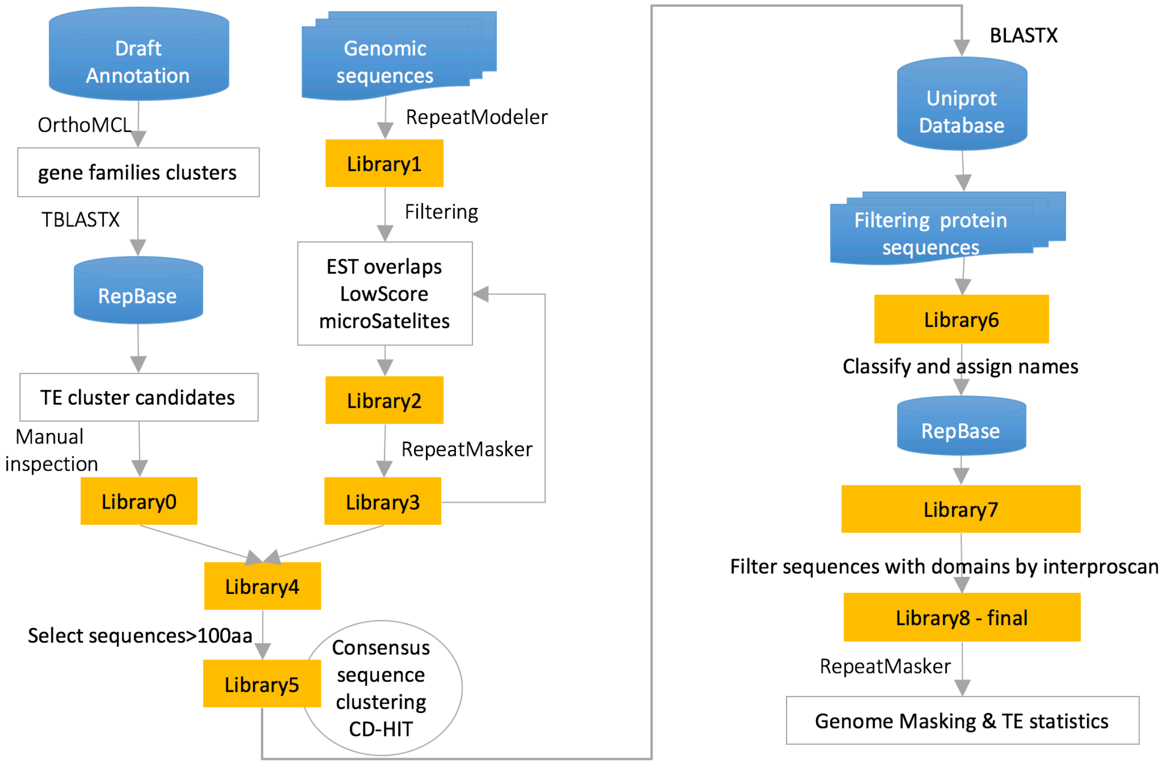


Figure 15: A comprehensive workflow for a *de novo* genome masking.
Mis-annotated TE can be clustered thus using RepBase to search gene family cluster is essential. Library4 keeps sequences longer than 300 nucleotides or 100 amino acids. To remove redundant sequences, Library5 needs consensus sequence clustering. Using UniProt and InterProScan can remove misannotated protein-coding sequences. However, be advised that TE can also be reverse transcriptase, which is often encoded by the TE itself. If in this case that transcriptase is detected, this sequence needs to be categorized as TE. In theory and practice, two excellent genome masking cases are recommended from spider mite genome and pig genome (Grbic *et al.*, 2011a; Groenen *et al.*, 2012).

2.5.2 Annotation for protein-coding genes

Annotation for protein-coding genes includes two steps: structural prediction and functional prediction. EUGENE is used here as an example to demonstrate a comprehensive workflow of gene annotation (Foissac *et al.*, 2008). EUGENE is a sensitive and comprehensive gene finder, which can distinguish non-coding sequences by probabilistic models such as Hidden Markov Models (HMM), which is a statistical Markov model where the system being modeled is assumed to be a Markov process with unobserved (hidden) models. EUGENE also can discriminate effective splicing sites from false splicing sites using various mathematical models in both eukaryotic and prokaryotic genomes. I summarize state-of-the-art public available tools for genome annotation in Table 5, which is an entry point for exploring annotation in greater detail and is not intended to be comprehensive, owing to space limitations.

2.5.3 Structural prediction

Structural prediction is a process of predicting gene structures across the entire genome. Here protein-coding gene structural prediction is used as a demonstration. The structural prediction has two approaches: evidence-based and *ab initio*. **Evidence-based prediction** uses extrinsic evidence including RNAseq for junction prediction and reference genomes for weight prediction. EUGENE and GenomeScan are typical evidence-based tools (Burge and Karlin, 1997; Foissac *et al.*, 2008). As for ***ab initio* prediction***,* prediction tools use intrinsic features without any extra data. Augustus and GeneMark-ES were designed for *ab initio* prediction (Lukashin and Borodovsky, 1998; Stanke *et al.*, 2004; Besemer and Borodovsky, 2005; Borodovsky and Lomsadze, 2011). However, because of the accumulation of genomic data, the evidence-based approach offers more evidence for structural prediction. Here I demonstrate an evidence-based structural approach by following three major steps: initial draft gene structure detection, data training, and structural re-prediction.

First, one should run EUGENE with default parameters to obtain an initial draft structural prediction (Figure 16, in blue and yellow). Initially, all the evidence-based resources (such as RNAseq, ESTs, junctions or protein sequences) will be aligned to the assembly and then EUGENE uses its default parameters to predict a draft structural prediction. The draft prediction offers an overview of gene models, even though final prediction will have some difference from this initial version. The draft gene models also will be used in the subsequent data training process. The evidence-based resources are ideally required to be as comprehensive as possible in EUGENE to obtain a good draft structural prediction.

Second, EUGENE parameters need to be specifically trained for a *de novo* genome (Figure 16, in purple). It is a machine learning process to determine the potential gene structures. Data training consists of both training the Splice Machine and EUGENE parameters. This is a win-or-die battle in current genome annotation projects and it requires a lot of manual work. In eukaryotic genomes, splicing sites have an impressive effect on the quality of structural prediction and thus machine learning-based predictions for exonic variants is quite important. In brief, RNAseq reads that are aligned to the whole genomic assembly can offer junction evidence, from where the Splice Machine will learn donor/acceptor weights across the whole genome. Then, select a number of flanking splicing gene models from draft predictions with strong junction support and manually correct them in genome editors such as GenomeView, Artemis or IGV browser (Rutherford *et al.*, 2000; Robinson *et al.*, 2011b; Abeel *et al.*, 2012). Ideally, at least 100 sets of neighbor genes (at least two genes in one set) with good junction data support need to be manually curated. Later, these curated genes are used as input for the training dataset, which assists in evaluating and optimizing EUGENE parameters. EUGENE offers ‘fitness’ to represent sensitivity, specificity, and accuracy of prediction. It takes several trials until a good fitness (a parameter to assess annotation - the higher fitness, the better annotation) can be reached. For instance, if fitness reaches 70%, it means at least 70% gene models can be exactly predicted.

Third, *ab initio* prediction is retreated by re-running EUGENE for structural prediction using trained parameters. Plus by using EvidenceModeler (Haas *et al.*, 2008), the best gene models will be obtained by combining of intrinsic and extrinsic evidence, as shown in Figure 16 (in yellow).

Table 5: Main tools for genome annotation.

	METHOD	***TOOL****	***REMARK***
STRUCTURAL PREDICTION	*ab initio* (Intrinsic approach)	Augustus	Designed for eukaryotic genome prediction based on HMM without using external evidence, also applicable for RNAseq data
		GeneID	Predict genes in anonymous genomic sequences designed with a hierarchical structure, efficient in speed and memory usage, compatible with multiple sources
		GeneMark	Gene Prediction in bacteria, fungi, archaea, metagenomes and metatranscriptomes, eukaryotes, transcripts and viruses, phages and plasmids
		Glimmer	Designed to find genes in microbial DNA, especially the genomes of bacteria, archaea, and viruses using IMMs
		Gnomon	A combination of homology searching with ab initio modeling using HMM
		BRAKER1	A pipeline for unsupervised RNAseq-based genome annotation combining GeneMark-ET and AUGUSTUS not require pre-trained parameters
		FGENESH	Designed from Fgene (pattern-based human gene prediction) and Fgenesh (hidden Markov model(HMM)) based gene prediction with Drosophila gene parameters, and it is organism-specific and now available at soft berry
	Evidence-based (Extrinsic approach)	EUGENE	An open integrative gene finder for eukaryotic and prokaryotic genomes, using extrinsic and intrinsic data.
		TwinScan	Exploit homology between two related genomes using separate probability models, currently available for mammals, worms, dicot plants and Cryptococci.
		GeneScan	Online resource for predicting the locations and exon-intron structures of genes
		GenomeScan	Identify exon-intron structures and sequencing similarity
FUNCTIONAL PREDICTION	homolog or domain-based	BLAST2GO	Obtain gene ontology based on data similarity searches with statistical analysis
		InterPro kits	InterProScan and interpro2GO, analyze protein functions and predict domain and important protein signatures
		Pfam	A large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs)
		Phobius	A combined transmembrane topology and signal peptide predictor
		PANNZER	A fully automated service for functional annotation of prokaryotic and eukaryotic proteins of novel function
		TMHMM	Prediction of transmembrane helices in proteins
		SignalIP	A server predicts the presence and location of signal peptide cleavage sites in amino acid sequences

* A full list of website links and citations is in the supplementary section of this chapter.

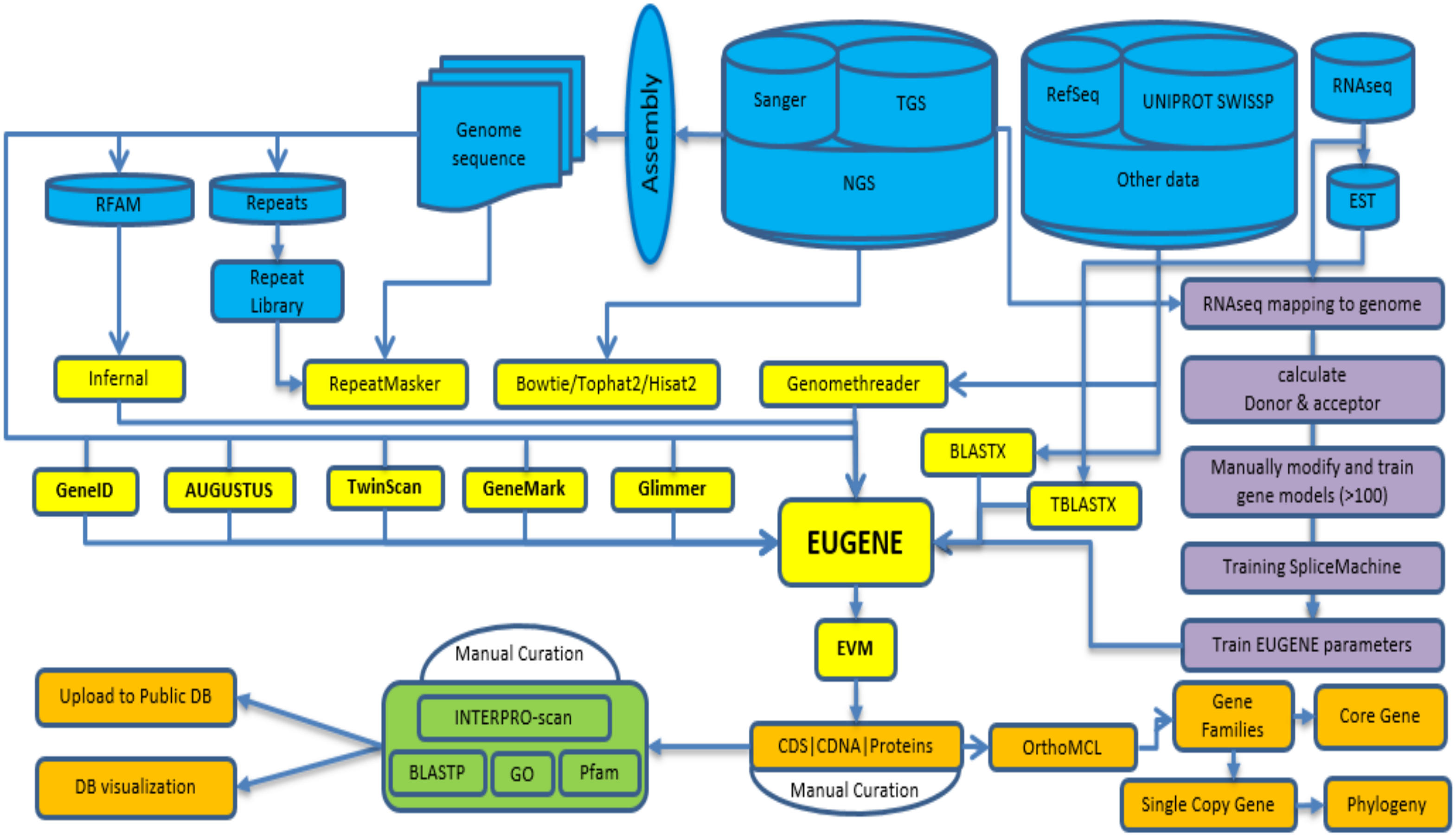


Figure 16: EUGENE combines comprehensive evidence for genome annotation.
Database (evidence) in blue, structural annotation in yellow, training dataset in purple, functional annotation in green, other analyses in orange.

2.5.4 Functional prediction

Functional prediction *in silico* is a process of predicting gene functional descriptions, i.e., what their functions are, using related evidence from known homologous proteins, protein domains or motifs. Widely used functional prediction tools are including BLAST2GO and InterProScan (Conesa *et al.*, 2005; Quevillon *et al.*, 2005; Mulder and Apweiler, 2007; Jones *et al.*, 2014). These tools use homologous domain sequences to assign gene models functional descriptions (Figure 16, in green). Some other databases are also available for functional prediction. For instance, UniProtKB and SwissProt can annotate newly identified proteins (Bairoch and Apweiler, 2000; Apweiler *et al.*, 2004). Preferably, functional annotation requires the sources be as comprehensive as possible to determine the best functional descriptions. Without detecting any domains, novel, or short, or orphan genes possibly end up as hypothetical genes. Of course, an accurate functional prediction needs to be validated by experimentation, which takes huge effort and time (Mudge and Harrow, 2016).

2.5.5 Inspection and modification

This is simply a must after automatic structural prediction. Genome annotation is a never-ending job because no annotation software can manage a perfect prediction without any errors and misannotations. Several common annotation errors are shown in Figure 17b: neighbor genes concatenation, gene splitting, genes with additional extension or genes without extension. A good training data-set and good prediction software can reduce these errors but still cannot eliminate them completely. Modifying these errors requires a good sense to gene models and good additional data supports.

An example of a good gene model is shown in Figure 17a, supported by various lines of evidence such as reference genes, RNAseq assemblies, junctions and blast hits (Figure 17c). However, this final step is quite labor-intensive and many model organism genomes require human effort on revisiting each gene to decide the best gene model.

There are some essential curation principles:

a) Gene starts with a start codon ATG (M);

b) Gene ends up with a stop codon TAA or TAG or TGA;

c) No stop codon inside (unless it is a pseudogene or a sequencing or assembly error);

d) Donor GT;

e) Acceptor AG (AG|GT); *

f) BLAST evidence;

g) RNAseq and junction evidence;

h) Public resources for reference genomes (Mudge and Harrow, 2016);

*In most cases, donor is GT and acceptor is AG but other types of donor and acceptor are rare but possible, especially in prokaryotic genomes;

2.5.6 Annotation quality assessment

The quality of genome annotation has an essential influence on downstream analyses and experimental hypotheses. Poorly annotated genomes can barely be used to explore biological significance. It is acceptable that most genes are well annotated with a few over-predicted or mispredicted genes. BUSCO and CEGMA are often employed to test the completeness of annotated gene set (Parra *et al.*, 2007; Simao *et al.*, 2015). They offer insights into possible unpredicted/missing genes. Check the completeness of annotation since it is possible that some core genes or single-copy genes are missing. Another way is to detect the domains or motifs of the protein-coding genes. If >30% of protein-coding genes have no detected domains or motifs, it is more likely that the prediction is not good, rather than a burst of real novel proteins.

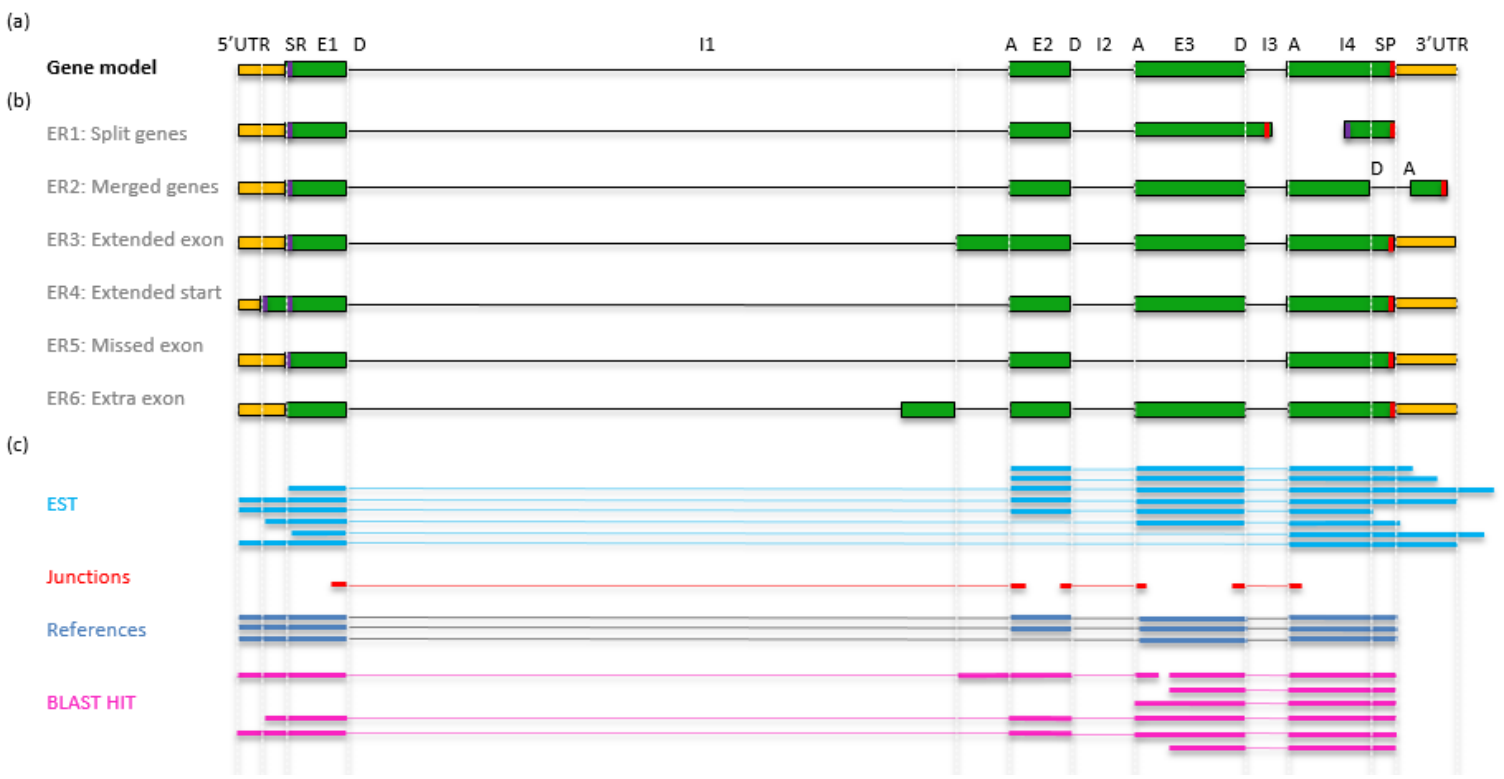


Figure 17: Common structural prediction errors.

(a) gene model; (b) six possible structural predictions; (c) supportive evidence; Orange bars are 5'UTR and 3'UTR. Green bars are exons connected by dark solid lines (introns). ER: error; SR: start codon; SP: stop codon; E: exon; I: intron; D: donor; A: acceptor; Purple bar is start codon; red bar is stop codon. ER5 is also probably an alternative splicing gene. However, the third exon (E3) is supposed to be annotated in the gene model.

2.5.7 Annotation for ncRNA

The ncRNA genes contribute an important proportion of RNAs, including lncRNA and small RNA (such as tRNA, rRNA, piRNA, miRNA, and snoRNA). However, ncRNA genes play important roles in genome regulation and network (Griffiths-Jones, 2007; Kim *et al.*, 2009). Therefore, annotation for ncRNA also presents a substantial challenge in a genome project.

Some ncRNA present themselves as clusters while some others disperse across the whole genome. For example, rRNA genes usually present as an integrated cluster but tRNA genes are scattered across the genome. Some conserved secondary structures and motifs can also be utilized as signatures to identify ncRNA. The lncRNA genes can be annotated (e.g., using PLAR) according to non-protein-coding transcripts (Hezroni *et al.*, 2015). The miRBase can be used to annotate high-confidence miRNAs (Kozomara and Griffiths-Jones, 2014). The tool tRNAscan-SE is used to predict tRNA (Lowe and Eddy, 1997; Lowe and Chan, 2016). As for other ncRNA genes, they can be annotated by homologous sequences from the public database as well. For instance, Rfam is a database of 2,450 types of ncRNA (last access on May 20th, 2016) (Nawrocki *et al.*, 2015) including lncRNA, tRNA, rRNA, sRNA, snRNA, miRNA, and snoRNA. Infernal is a fast and precise tool to predict ncRNA using Rfam database and it has been successfully applied in many ncRNA studies (Nawrocki and Eddy, 2013; El Korbi *et al.*, 2014; Nawrocki, 2014; Barquist *et al.*, 2016). However, ncRNA annotation is at a cutting-edge era because of their poorly conserved primer structures. Nevertheless, they are quite conserved at the level of secondary structure. Therefore, it is a conventional approach to identify ncRNA using secondary structure by ncRNA-specific database and tools such as miRbase and tRNA-scan (Lowe and Eddy, 1997; Kozomara and Griffiths-Jones, 2014).

2.5.8 Annotation for pseudogenes

Pseudogenes, also known as genomic fossils, originate from genome duplication or retrotransposition, which leads to frameshifts, large INDELs or nonsense mutations in various species (Mighell *et al.*, 2000; Zhang *et al.*, 2006). Some pseudogenes are terminated in the middle of the protein-coding sequence by stop codons. It is likely some pseudogenes might still play a function if any domain exists in the coding region. However, during genome annotation, many genes fail into 'pseudogenes' because of sequence gaps, truncated scaffolds or even artificial sequencing errors. Thus, they are usually difficult to identify. Pseudogenes prediction tools (i.e., PseudoPipe) integrates a combination of criteria including homologous proteins, intron-exon structures, and the existence of stop codons and frameshifts (Zhang *et al.*, 2006). Nevertheless, predicted pseudogenes must be carefully treated and validated for higher confidence.

2.5.9 Genomic statistics

Once a genome is ready to be submitted and published, statistics for assembly and annotation are required to offer an overview of the genome assembly as well as essential genomic features. Key statistical categories are listed in Table 6. For example, scaffold/contig N50 suggests the continuity of the genome assembly. Genome size and gene number show how big the genome is and the gene density across the whole genome. Other analyses such as gene families, lineage-specific genes, and likely gene regulation reveal biological significance.

2.5.10 Genome visualization, maintenance, and update

Periodic genomic database maintenance and update are important for biologists. In theory, genomic sequence and annotation are supposed to be submitted to a public database. Some important model organisms such as Human, *Drosophila*, and *Arabidopsis* have their own scientific communities for data access, preliminary analysis, where users can modify and update genomes (Adams *et al.*, 2000; Arabidopsis Genome Initiative, 2000; Lander *et al.*, 2001). An excellent example of the eukaryotic genome community is ORCAE, offering users comprehensive tools and evidence to genomic datasets (Sterck *et al.*, 2012). Currently, over 20 eukaryotic genomes are publicly available on ORCAE, which supports the viewing of most genomic information such as functional description, gene locus and structure, homologous genes, protein domains and expression profiles. Genome annotation always works in progress. Even after fifteen years of the human genome project, scientists are continuously improving its annotation by looking at more RNAseq data to improve gene models or to detect alternative spliced genes. Therefore, as more NGS data is cumulating, it is necessary to update genome database periodically, even after acceptance of the respective genome paper.

2.6 Perspective

Genome sequencing, assembly, and annotation are fundamental steps for a genome project. I have gone through the essential steps of a genome sequencing project with key workflow, algorithms, technologies, and terminologies. However, genome assembly and annotation could not possibly have been accomplished without the aid of NGS technologies. NGS, no doubt, has changed our knowledge in life science and helped us to uncover more information in various genomes. This information can be applied in agriculture, clinical studies, personalized precision medicine and so forth. Consequently, the NGS market is becoming quite competitive. In 2016, global NGS market was dominated by Illumina, Thermo Fisher Scientific, and Pacific Biosciences. These companies provide the essential NGS platforms in the world. However, with the emergence of TGS, short reads become a disadvantage of NGS. Nevertheless, TGS still requires high upfront expense despite its long reads productivity. To compensate for this problem, more and more genome projects start to apply the hybrid sequencing method using both NGS short reads and TGS long reads (Gordon *et al.*, 2016b; Zapata *et al.*, 2016). NGS short reads are applied to correct the precision of the bases while TGS long reads can overcome the assembly issues caused by repetitive sequences and scaffold gaps.

Devising these novel methods, algorithms and strategies for the biological interpretation of massively parallel sequencing data will be the next step for NGS goals. I anticipate that one day, genomic sequences will be read at the chromosomal level and no more assembly will be required. With the development of more accurate reference genes, faster high-performance computational platforms, and more precise annotation pipelines, I believe a more advanced genomic era is quite within reach.

Table 6: Statistics categories for genome assembly and annotation.

Assembly	**Annotation**
Genome size	Number of genes
N50/N90	Gene density
L50/L90	Average length of genes
Largest scaffold	Median length genes
Average length of scaffold	Number of exons
Number of contigs	Total exon length
Largest contig	Average length of exons
Average contig length	Median length of exons
Gaps (>50N)	Longest exons
Longest/shortest CDS	Average exon number per gene
GC-content	Gene with most exons

2.7 Supplementary Links

Tool	**Website or Citation**
Augustus	http://augustus.gobics.de (Stanke et al., 2004)
GeneID	http://genome.crg.es/software/geneid/index.html (Parra et al., 2000)
GeneMark	http://exon.gatech.edu/GeneMark/ (Besemer et al., 2001; Borodovsky and Lomsadze, 2014)
Glimmer	https://ccb.jhu.edu/software/glimmer/ (Delcher et al., 1999; Aggarwal and Ramaswamy, 2002; Delcher et al., 2007)
Gnomon	http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml
BRAKER1	http://bioinf.uni-greifswald.de/bioinf/braker/ (Hoff et al., 2016)
FGENESH	www.softberry.com (Salamov and Solovyev, 2000)
EUGENE	http://eugene.toulouse.inra.fr/ (Foissac et al., 2008)
TwinScan	http://mblab.wustl.edu/software.html (Korf et al., 2001)
GeneScan	http://genes.mit.edu/GENSCAN.html
GenomeScan	http://genes.mit.edu/genomescan.html (Burge and Karlin, 1998)
BLAST2GO	http://www.blast2go.de/ (Conesa et al., 2005; Conesa and Gotz, 2008)
InterPro kits	https://www.ebi.ac.uk/interpro/ (Mulder and Apweiler, 2007; Jones et al., 2014)
Pfam	http://pfam.xfam.org/ (Bateman et al., 2002)
Phobius	http://phobius.sbc.su.se/ (Kall et al., 2004)
PANNZER	http://ekhidna.biocenter.helsinki.fi/pannzer/ (Koskinen et al., 2015)
TMHMM	http://www.cbs.dtu.dk/services/TMHMM/ (Krogh et al., 2001)
SignalP	http://www.cbs.dtu.dk/services/SignalP/ (Petersen et al., 2011)

2.8 Reference

Abeel, T., Van Parys, T., Saeys, Y., Galagan, J., and Van de Peer, Y. (2012). GenomeView: a next-generation genome browser. Nucleic acids research *40*, e12.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F.*, et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M.*, et al.* (2004). UniProt: the Universal Protein knowledgebase. Nucleic acids research *32*, D115-119.

Arabidopsis Genome Initiative, A. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796-815.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic acids research *28*, 45-48.

Barquist, L., Burge, S.W., and Gardner, P.P. (2016). Studying RNA Homology and Conservation with Infernal: From Single Sequences to RNA Families. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] *54*, 12 13 11-12 13 25.

Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., and Lander, E.S. (2002). ARACHNE: a whole-genome shotgun assembler. Genome research *12*, 177-189.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research *27*, 573-580.

Besemer, J., and Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic acids research *33*, W451-454.

Birney, E., and Soranzo, N. (2015). Human genomics: The end of the start for population sequencing. Nature *526*, 52-53.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578-579.

Boetzer, M., and Pirovano, W. (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC bioinformatics *15*, 211.

Bolger, A.M., Lohse, M., and Usadel, B. (2014a). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.

Bolger, A.M., Lohse, M., and Usadel, B. (2014b). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.

Boothby, T.C., Tenlen, J.R., Smith, F.W., Wang, J.R., Patanella, K.A., Nishimura, E.O., Tintori, S.C., Li, Q., Jones, C.D., Yandell, M., *et al.* (2015). Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. Proceedings of the National Academy of Sciences of the United States of America *112*, 15976-15981.

Borodovsky, M., and Lomsadze, A. (2011). Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] *Chapter 4*, Unit 4 6 1-10.

Bos, K.I., Schuenemann, V.J., Golding, G.B., Burbano, H.A., Waglechner, N., Coombes, B.K., McPhee, J.B., DeWitte, S.N., Meyer, M., Schmedes, S., *et al.* (2011). A draft genome of Yersinia pestis from victims of the Black Death. Nature *478*, 506-510.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G.L., D'Amore, R., Allen, A.M., McKenzie, N., Kramer, M., Kerhornou, A., Bolser, D., *et al.* (2012a). Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature *491*, 705-710.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. Journal of molecular biology *268*, 78-94.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nature biotechnology *31*, 1119-1125.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C., and Jaffe, D.B. (2008). ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome research *18*, 810-820.

Chain, P.S., Grafham, D.V., Fulton, R.S., Fitzgerald, M.G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D.C., Buhay, C., *et al.* (2009). Genomics. Genome project standards in a new era of sequencing. Science *326*, 236-237.

Chaisson, M.J., Wilson, R.K., and Eichler, E.E. (2015b). Genetic variation and the de novo assembly of human genomes. Nature reviews Genetics *16*, 627-640.

Chen, L., Liu, P., Evans, T.C., Jr., and Ettwiller, L.M. (2017). DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. Science *355*, 752-756.

Chen, Y.M., Yu, C.H., Hwang, C.C., and Liu, T. (2013). OMACC: an Optical-Map-Assisted Contig Connector for improving de novo genome assembly. BMC systems biology *7 Suppl 6*, S7.

Collado-Vides, J., Medrano-Soto, A., and Tusie-Luna, M.T. (2003). With the finished human genome in hand, what next? Genome biology *4*, 328.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics *21*, 3674-3676.

Dahm, R. (2008). Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. Human genetics *122*, 565-581.

de la Bastide, M., and McCombie, W.R. (2007). Assembling genomic DNA sequences with PHRAP. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] *Chapter 11*, Unit11 14.

Dong, Y., Xie, M., Jiang, Y., Xiao, N., Du, X., Zhang, W., Tosser-Klopp, G., Wang, J., Yang, S., Liang, J.*, et al.* (2013). Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). Nature biotechnology *31*, 135-141.

Edgar, R.C., and Myers, E.W. (2005). PILER: identification and classification of genomic repeats. Bioinformatics *21 Suppl 1*, i152-158.

El Korbi, A., Ouellet, J., Naghdi, M.R., and Perreault, J. (2014). Finding instances of riboswitches and ribozymes by homology search of structured RNA with Infernal. Methods in molecular biology *1103*, 113-126.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC bioinformatics *9*, 18.

Feschotte, C., Keswani, U., Ranganathan, N., Guibotsy, M.L., and Levine, D. (2009). Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome biology and evolution *1*, 205-220.

Foissac, S., Gouzy, J., Rombauts, S., Mathe, C., Amselem, J., Sterck, L., Van de Peer, Y., Rouze, P., and Schiex, T. (2008). Genome annotation in plants and fungi: EuGene as a model platform. Curr Bioinform *3*, 87-97.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S.*, et al.* (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences of the United States of America *108*, 1513-1518.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nature reviews Genetics *17*, 333-351.

Gordon, D., Huddleston, J., Chaisson, M.J.P., Hill, C.M., Kronenberg, Z.N., Munson, K.M., Malig, M., Raja, A., Fiddes, I., Hillier, L.W.*, et al.* (2016b). Long-read sequence assembly of the gorilla genome. Science *352*, 52-+.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q*., et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology *29*, 644-652.

Grbic, M., Van Leeuwen, T., Clark, R.M., Rombauts, S., Rouze, P., Grbic, V., Osborne, E.J., Dermauw, W., Ngoc, P.C., Ortego, F*., et al.* (2011a). The genome of Tetranychus urticae reveals herbivorous pest adaptations. Nature *479*, 487-492.

Green, E.D. (2001). Strategies for the systematic sequencing of complex genomes. Nature reviews Genetics *2*, 573-583.

Griffiths-Jones, S. (2007). Annotating noncoding RNA genes. Annual review of genomics and human genetics *8*, 279-298.

Groenen, M.A.M., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J*., et al.* (2012). Analyses of pig genomes provide insight into porcine demography and evolution. Nature *491*, 393-398.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M*., et al.* (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols *8*, 1494-1512.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome biology *9*, R7.

Haridas, S., Breuill, C., Bohlmann, J., and Hsiang, T. (2011). A biologist's guide to de novo genome assembly using next-generation sequence data: A test with fungal genomes. Journal of microbiological methods *86*, 368-375.

Hernandez, D., Francois, P., Farinelli, L., Osteras, M., and Schrenzel, J. (2008). De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome research *18*, 802-809.

Hezroni, H., Koppstein, D., Schwartz, M.G., Avrutin, A., Bartel, D.P., and Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep *11*, 1110-1122.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome research *9*, 868-877.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G*., et al.* (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236-1240.

Jurka, J., Kapitonov, V.V., Kohany, O., and Jurka, M.V. (2007). Repetitive sequences in complex genomes: structure and evolution. Annual review of genomics and human genetics *8*, 241-259.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S.*, et al.* (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice *6*, 4.

Keeling, P.J., and Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. Nature reviews Genetics *9*, 605-618.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nature methods *12*, 357-360.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology *14*, R36.

Kim, V.N., Han, J., and Siomi, M.C. (2009). Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol *10*, 126-139.

Korbel, J.O., and Lee, C. (2013). Genome assembly and haplotyping with Hi-C. Nature biotechnology *31*, 1099-1101.

Koutsovoulos, G., Kumar, S., Laetsch, D.R., Stevens, L., Daub, J., Conlon, C., Maroon, H., Thomas, F., Aboobaker, A.A., and Blaxter, M. (2016). No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. Proceedings of the National Academy of Sciences of the United States of America *113*, 5053-5058.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic acids research *42*, D68-73.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome biology *5*, R12.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] *Chapter 11*, Unit 11 17.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.

Laver, T., Harrison, J., O'Neill, P.A., Moore, K., Farbos, A., Paszkiewicz, K., and Studholme, D.J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. Biomolecular detection and quantification *3*, 1-8.

Ledford, H. (2008). Population genomics for fruitflies. Nature *453*, 1154-1155.

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., McCombie, W.R., and Schatz, M. (2016). Third-generation sequencing and the future of genomics. bioRxiv.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y.*, et al.* (2010). The sequence and de novo assembly of the giant panda genome. Nature *463*, 311-317.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009b). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966-1967.

Li, X., and Waterman, M.S. (2003). Estimating the repeat structure and length of DNA sequences using L-tuples. Genome research *13*, 1916-1922.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B.*, et al.* (2012). Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Briefings in functional genomics *11*, 25-37.

Ling, H.Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., Fan, H., Li, D., Dong, L., Tao, Y.*, et al.* (2013). Draft genome of the wheat A-genome progenitor Triticum urartu. Nature *496*, 87-90.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. J Biomed Biotechnol *2012*, 251364.

Lowe, T.M., and Chan, P.P. (2016). tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic acids research *44*, W54-57.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic acids research *25*, 955-964.

Lukashin, A.V., and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic acids research *26*, 1107-1115.

Luo, M.C., Deal, K.R., Murray, A., Zhu, T., Hastie, A.R., Stedman, W., Sadowski, H., and Saghbini, M. (2016). Optical Nano-mapping and Analysis of Plant Genomes. Methods in molecular biology *1429*, 103-117.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., *et al.* (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience *1*, 18.

Maccallum, I., Przybylski, D., Gnerre, S., Burton, J., Shlyakhter, I., Gnirke, A., Malek, J., McKernan, K., Ranade, S., Shea, T.P., *et al.* (2009). ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. Genome biology *10*, R103.

Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nature reviews Genetics *11*, 499-511.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376-380.

Martin, J.A., and Wang, Z. (2011). Next-generation transcriptome assembly. Nature reviews Genetics *12*, 671-682.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S.O., Wicker, T., Radchuk, V., Dockter, C., Hedley, P.E., Russell, J., *et al.* (2017). A chromosome conformation capture ordered sequence of the barley genome. Nature *544*, 427-433.

McCarthy, E.M., and McDonald, J.F. (2003). LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics *19*, 362-367.

Mendelowitz, L., and Pop, M. (2014). Computational methods for optical mapping. GigaScience *3*, 33.

Metzker, M.L. (2005). Emerging technologies in DNA sequencing. Genome research *15*, 1767-1776.

Mighell, A.J., Smith, N.R., Robinson, P.A., and Markham, A.F. (2000). Vertebrate pseudogenes. FEBS letters *468*, 109-114.

Miller, J.R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. Genomics *95*, 315-327.

Mills, R.E., Bennett, E.A., Iskow, R.C., and Devine, S.E. (2007b). Which transposable elements are active in the human genome? Trends in Genetics *23*, 183-191.

Monier, A., Pagarete, A., de Vargas, C., Allen, M.J., Read, B., Claverie, J.M., and Ogata, H. (2009). Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. Genome research *19*, 1441-1449.

Morgulis, A., Gertz, E.M., Schaffer, A.A., and Agarwala, R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. Journal of computational biology : a journal of computational molecular cell biology *13*, 1028-1040.

Mudge, J.M., and Harrow, J. (2016). The state of play in higher eukaryote gene annotation. Nature reviews Genetics *17*, 758-772.

Mulder, N., and Apweiler, R. (2007). InterPro and InterProScan: tools for protein sequence classification and comparison. Methods in molecular biology *396*, 59-70.

Myers, E.W. (2005). The fragment assembly string graph. Bioinformatics *21 Suppl 2*, ii79-85.

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A.*, et al.* (2000). A whole-genome assembly of Drosophila. Science *287*, 2196-2204.

Nadalin, F., Vezzi, F., and Policriti, A. (2012b). GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC bioinformatics *13 Suppl 14*, S8.

Nagarajan, N., and Pop, M. (2013). Sequence assembly demystified. Nature reviews Genetics *14*, 157-167.

Nawrocki, E.P. (2014). Annotating functional RNAs in genomes using Infernal. Methods in molecular biology *1097*, 163-197.

Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J.*, et al.* (2015). Rfam 12.0: updates to the RNA families database. Nucleic acids research *43*, D130-137.

Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933-2935.

Neely, R.K., Deen, J., and Hofkens, J. (2011). Optical mapping of DNA: single-molecule-based methods for mapping genomes. Biopolymers *95*, 298-311.

Olsen, J.L., Rouze, P., Verhelst, B., Lin, Y.C., Bayer, T., Collen, J., Dattolo, E., De Paoli, E., Dittami, S., Maumus, F.*, et al.* (2016). The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea. Nature *530*, 331-335.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics *23*, 1061-1067.

Paulino, D., Warren, R.L., Vandervalk, B.P., Raymond, A., Jackman, S.D., and Birol, I. (2015). Sealer: a scalable gap-closing application for finishing draft genomes. BMC bioinformatics *16*, 230.

Pellicer, J., and Leitch, I.J. (2014). The application of flow cytometry for estimating genome size and ploidy level in plants. Methods in molecular biology *1115*, 279-307.

Peng, Y., Leung, H.C., Yiu, S.M., and Chin, F.Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics *28*, 1420-1428.

Perry, D.A., Morrison, H.G., and Adam, R.D. (2011). Optical map of the genotype A1 WB C6 Giardia lamblia genome isolate. Molecular and biochemical parasitology *180*, 112-114.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics *21 Suppl 1*, i351-358.

Prufer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C*., et al.* (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43-49.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC genomics *13*, 341.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic acids research *33*, W116-120.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Rasch, E.M., Lee, C.E., and Wyngaard, G.A. (2004). DNA-Feulgen cytophotometric determination of genome size for the freshwater-invading copepod Eurytemora affinis. Genome / National Research Council Canada = Genome / Conseil national de recherches Canada *47*, 559-564.

Raymond, J., and Blankenship, R.E. (2003). Horizontal gene transfer in eukaryotic algal evolution. Proceedings of the National Academy of Sciences of the United States of America *100*, 7419-7420.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011b). Integrative genomics viewer. Nature biotechnology *29*, 24-26.

Rogers, J. (2003). The finished genome sequence of Homo sapiens. Cold Spring Harbor symposia on quantitative biology *68*, 1-11.

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N*., et al.* (2014). Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature *515*, 261-263.

Rubin, C.J., Zody, M.C., Eriksson, J., Meadows, J.R., Sherwood, E., Webster, M.T., Jiang, L., Ingman, M., Sharpe, T., Ka, S., *et al.* (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. Nature *464*, 587-591.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. Bioinformatics *16*, 944-945.

Sanggaard, K.W., Bechsgaard, J.S., Fang, X., Duan, J., Dyrlund, T.F., Gupta, V., Jiang, X., Cheng, L., Fan, D., Feng, Y., *et al.* (2014). Spider genomes provide insight into composition and evolution of venom and silk. Nature communications *5*, 3765.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., *et al.* (1996). Nested retrotransposons in the intergenic regions of the maize genome. Science *274*, 765-768.

Schatz, M.C., Delcher, A.L., and Salzberg, S.L. (2010). Assembly of large genomes using second-generation sequencing. Genome research *20*, 1165-1173.

Seo, J.S., Rhie, A., Kim, J., Lee, S., Sohn, M.H., Kim, C.U., Hastie, A., Cao, H., Yun, J.Y., Kim, J., *et al.* (2016b). De novo assembly and phasing of a Korean human genome. Nature.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210-3212.

Simpson, J.T., and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. Genome research *22*, 549-556.

Simpson, J.T., and Pop, M. (2015). The Theory and Practice of Genome Sequence Assembly. Annual review of genomics and human genetics *16*, 153-172.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome research *19*, 1117-1123.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nature reviews Genetics *15*, 121-132.

Smit, A., and Hubley, R. (2008-2015). RepeatModeler Open-1.0.

Smith, D.R. (2013). Death of the genome paper. Frontiers in genetics *4*, 72.

Smith, D.R. (2017). Goodbye genome paper, hello genome report: the increasing popularity of 'genome announcements' and their impact on science. Briefings in functional genomics *16*, 156-162.

Soanes, D., and Richards, T.A. (2014). Horizontal gene transfer in eukaryotic plant pathogens. Annual review of phytopathology *52*, 583-614.

Sommer, D.D., Delcher, A.L., Salzberg, S.L., and Pop, M. (2007). Minimus: a fast, lightweight genome assembler. BMC bioinformatics *8*, 64.

Staden, R. (1980). A new computer method for the storage and manipulation of DNA gel reading data. Nucleic acids research *8*, 3673-3694.

Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic acids research *32*, W309-312.

Sterck, L., Billiau, K., Abeel, T., Rouze, P., and Van de Peer, Y. (2012). ORCAE: online resource for community annotation of eukaryotes. Nature methods *9*, 1041.

Sun, L.V., Foster, J.M., Tzertzinis, G., Ono, M., Bandi, C., Slatko, B.E., and O'Neill, S.L. (2001). Determination of Wolbachia genome size by pulsed-field gel electrophoresis. Journal of bacteriology *183*, 2219-2225.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] *Chapter 4*, Unit 4 10.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols *7*, 562-578.

Treangen, T.J., and Salzberg, S.L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature reviews Genetics *13*, 36-46.

Van Berkum, N.L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L.A., Dekker, J., and Lander, E.S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. Journal of visualized experiments : JoVE.

Van Dijk, E.L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. Trends in genetics : TIG *30*, 418-426.

Wilhelm, J., Pingoud, A., and Hahn, M. (2003). Real-time PCR-based method for the estimation of genome sizes. Nucleic acids research *31*, e56.

Xia, Q., Guo, Y., Zhang, Z., Li, D., Xuan, Z., Li, Z., Dai, F., Li, Y., Cheng, D., Li, R.*, et al.* (2009). Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). Science *326*, 433-436.

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic acids research *35*, W265-268.

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nature reviews Genetics *13*, 329-342.

Zapata, L., Ding, J., Willing, E.M., Hartwig, B., Bezdan, D., Jiao, W.B., Patel, V., Velikkakam James, G., Koornneef, M., Ossowski, S.*, et al.* (2016). Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. Proceedings of the National Academy of Sciences of the United States of America *113*, E4052-4060.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research *18*, 821-829.

Zhang, Z., Carriero, N., Zheng, D., Karro, J., Harrison, P.M., and Gerstein, M. (2006). PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics *22*, 1437-1439.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. Journal of computational biology : a journal of computational molecular cell biology *7*, 203-214.

Zhou, S., Wei, F., Nguyen, J., Bechner, M., Potamousis, K., Goldstein, S., Pape, L., Mehan, M.R., Churas, C., Pasternak, S.*, et al.* (2009). A single molecule scaffold for the maize genome. PLoS genetics *5*, e1000711.

Chapter 3

3 Improvement of the *Tetranychus urticae* genome using optical mapping and cumulative hybrid data

The spider mite *Tetranychus urticae* is a generalist herbivore of key ecological and agricultural importance. Published in 2011, the complete genome of *T. urticae* was initially released with 640 scaffolds and 18,414 protein-coding genes (Grbic *et al.*, 2011a). However, pest-control and genetics studies would better serve if provided with a better genome assembly and annotation. Therefore, here we present a new version of the *T. urticae* genome with significantly improved assembly and genome annotation. We accomplished this by the advances of optical mapping, the availability of accumulated RNAseq data, and manual curation for thousands of gene models. Briefly, the 640 scaffolds were assembled into six major super-scaffolds using optical mapping (OM) data. Subsequently, based on these six super-scaffolds, *T. urticae* genome was re-annotated using EUGENE and EvidenceModeler (EVM). The revised version of *T. urticae* genome annotation has a total number of the protein-coding gene of 19,042, of which 1,809 extra new genes were also recently predicted using additional RNAseq data. Of these new genes, 83.4% are supported by RNAseq and 39.2% of them were assigned functional descriptions. Over 29% genes show hallmarks of the transmembrane function, although these genes could not be clustered into one family. This suggests that these transmembrane-associated genes are probably fulfilling different roles. These extra protein-coding genes are relatively short, as 62.8% of them have a length between 50-100 amino acids. Only 12.5% of these are longer than 200 amino acids. The updated assembly for *T. urticae* genome will provide more opportunities for chromosomal and structural level studies and the updated annotation will offer more insights into spider mite genomics studies as well.

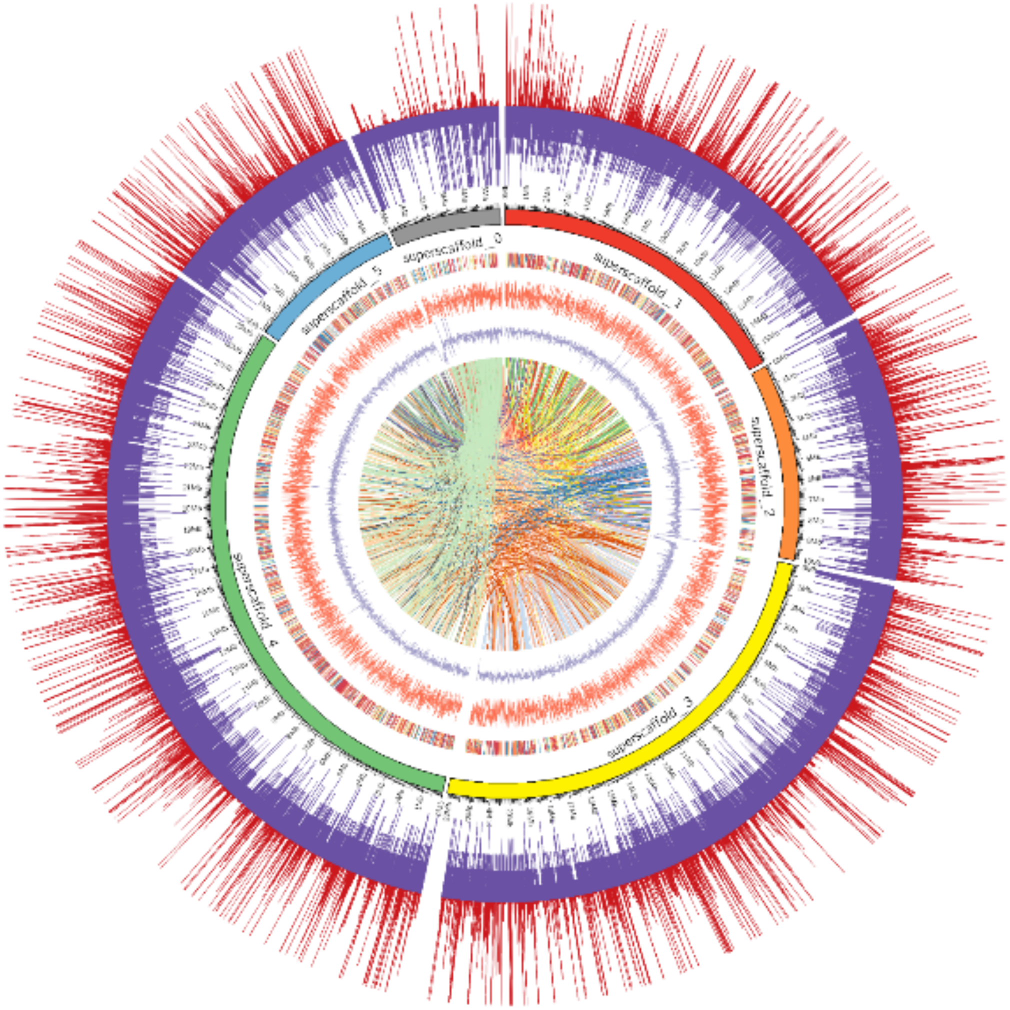


Figure 18: The circos overview of updated genome of *T. urticae*
From outside to inside: RNAseq coverage, genomic reads coverage, six superscaffolds, gene density, GC-skew, GC-content and genomic synteny (window 10 kb); Heatmap color was used spectral-7-div and the color ranging from blue to red suggests gene density from big to small, respectively. Superscaffold_0 has many collapsed regions because of unplaced reads and beginning of a scaffold also possibly old assembly artifact. Genomic reads coverage is usually high because multiple reads can support a beginning but usually low at the end due to no reads can be extended at the end of a scaffold. Plus, the telomere is a repetitive region to protect chromosome from deterioration or fusion with other chromosomes thus telomere regions are also often hard to assemble which leads artifacts.

3.1 Background

The spider mite *T. urticae* is a cosmopolitan agricultural pest, feeding on more than 1,100 plants and leading to significant economic damage worldwide (Migeon *et al.*, 2006). It has an extensive plant host range, an extreme record of pesticide resistance, a rapid life cycle, and an accelerating reproductive capability. Therefore, *T. urticae* has been established as a candidate pest-model for pest-plant-interactions. Its genome has revealed herbivorous pest adaptation and improved our understanding of the chelicerate genome (Grbic *et al.*, 2011a).

OM technique is a non-PCR-based approach to generate genome-wide restriction enzyme maps. Because it is not subject to cloning, amplification, hybridization or sequencing bias, it is ideally suited to the improvement of fragmented genome assemblies that can no longer be improved by conventional approaches. Therefore, OM has been widely applied in comparing the structures of bacterial genomes, completing bacterial genome assembly and correcting eukaryotic genome assembly errors (Lim *et al.*, 2001; Chen *et al.*, 2006b; Zhou *et al.*, 2007; Nagarajan *et al.*, 2008; Wei *et al.*, 2009; Wu *et al.*, 2009). Several large vertebrate genomes have also been successfully assembled by OM data, solving the problematic issues of repetitive elements and short reads (Neely *et al.*, 2011; Perry *et al.*, 2011; Dong *et al.*, 2013; Mendelowitz and Pop, 2014). Therefore, a chromosomal level assembly is helpful for genomic organization studies, which can shed light on species' evolutionary dynamics.

OM technique consists of the following steps: DNA extraction, labeling, massive parallelization, and imaging. By collecting long-range information on genomic sequences and visualizing beam spot pattern, the OM technique extends scaffolds by estimating the gap length between scaffolds and combines them into longer sequences without adding extra bases. Currently, there are two major optical mapping suppliers: OpGen and BioNano. The former uses restriction enzymes to sequence-specifically cleave two DNA strands. The latter, however, cuts only one DNA strand and generates shorter DNA. Through massive alignment of numerous beam spot patterns, optical mapping thus offers the possibility of longer scaffold assembly.

In this study, by using OM technique (both OpGen and BioNano data), the initial 640 scaffolds of *T. urticae* genome were assembled into six superscaffolds, based on which, the *T. urticae* genome was re-annotated using EUGENE and EVM with cumulative RNAseq data sequenced over the past six years. To the best of our knowledge, it is the first invertebrate genome assembled by OM data. Figure 18 shows an overview result of this updated genome assembly, indicating the RNAseq coverage, genomic reads coverage, and GC-content.

3.2 Data description

3.2.1 Update genome assembly

The initially released assembly of this complete genome (89.6 Mb) has 640 scaffolds and 18,414 protein-coding genes (Grbic *et al.*, 2011a). OpGen assembled six large Maptigs with a total size of 88 Mb. Out of six Maptigs, there are five potential complete chromosomal arms. Four Maptigs show a similar repetitive pattern at one end, suggesting these four Maptigs may come from two chromosomes. If the big fragment end of the Maptig truly represents the telomere region, we estimate the chromosome number to be three, otherwise, it would be four (possibly an additional tiny one). The chromosome was determined when each end of the Maptig reached either a big fragment region or a highly repetitive region that couldn’t be crossed further by assembly process.

Using these OM data, 43 scaffolds were finally assembled five major superscaffolds (85.77 Mb in total, taking up 96.4% of the whole *T. urticae* genome). The results are shown in Figure 19 and Table 7. The superscaffold_0 was concatenated by the rest 597 short scaffolds from long to short order with 1 kb “N” as bridges. Scaffold 1, 2, 4 and 8 were split and reversed as stated in the OM results. The longest superscaffold_4 is 29.86 Mb, taking up a proportion of 32% of the whole genome in size. The average scaffold length of the OM assembly is now increased ten folds compared with the initial assembly, from 141,899 bp to 15,242,415 bp.

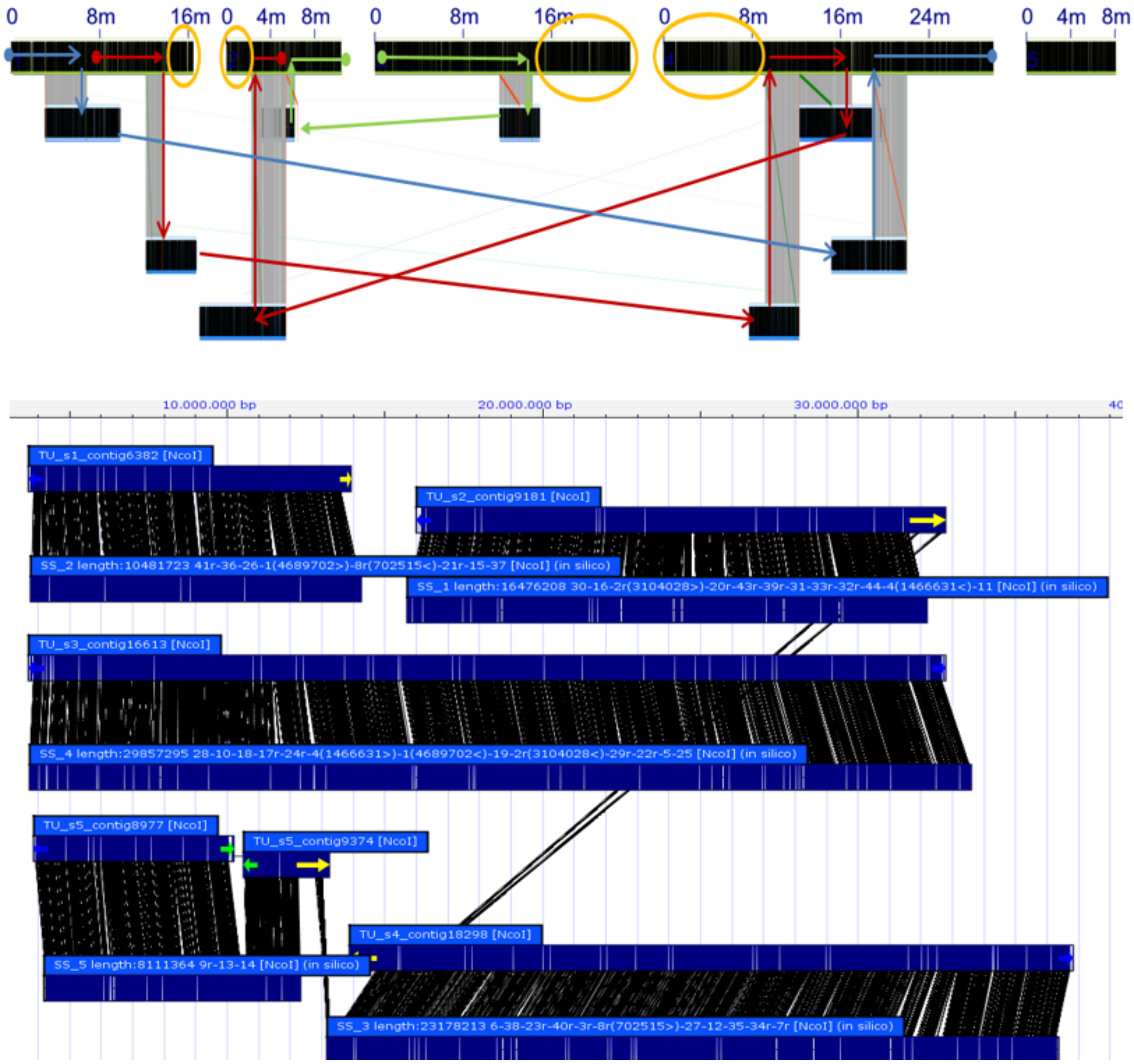

0
8m
16m
0
4m
8m
0
8m
16m
0
8m
16m
24m
0
4m
8m
10.000.000 bp
20.000.000 bp
30.000.000 bp
TU_s1_contig6382 [NcoI]
TU_s2_contig9181 [NcoI]
SS_2 length:10481723 41r-36-26-1(4689702>)-8r(702515<)-21r-15-37 [NcoI] (in silico)
SS_1 length:16476208 30-16-2r(3104028>)-20r-43r-39r-31-33r-32r-44-4(1466631<)-11 [NcoI] (in silico)
TU_s3_contig16613 [NcoI]
SS_4 length:29857295 28-10-18-17r-24r-4(1466631>)-1(4689702<)-19-2r(3104028<)-29r-22r-5-25 [NcoI] (in silico)
TU_s5_contig8977 [NcoI]
TU_s5_contig9374 [NcoI]
TU_s4_contig18298 [NcoI]
SS_5 length:8111364 9r-13-14 [NcoI] (in silico)
SS_3 length:23178213 6-38-23r-40r-3r-8r(702515>)-27-12-35-34r-7r [NcoI] (in silico)

Figure 19: The OM results from BioNano and OpGen for the *T. urticae* assembly.
Top: BionanoGenomics maps were generated with less long DNA molecules (extraction protocol is still under development). Therefore, maps were joined helped with the scaffolds; Plus, the three colored lines present 3 big chromosomes and yellow regions are unsolved. Bottom: the OpGen result after the protocol was optimized. Breakpoints are known and need to be confirmed in future studies.

Table 7: OM results and final assembly.

(a) The OM results - the order of initial scaffolds.

OM assembly	Superscaffold size (bp)	OM results
Superscaffold_1	16,476,208	30-16-2r(3104028>)-20r-43r-39r-31-33r-32r-44-4(1466631<)-11
Superscaffold_2	10,481,723	41r-36-26-1(4689702>)-8r(702515<)-21r-15-37
Superscaffold_3	23,178,213	6-38-23r-40r-3r-8r(702515>)-27-12-35-34r-7r
Superscaffold_4	29,857,295	28-10-18-17r-24r-4(1466631>)-1(4689702<)-19-2r(3104028<)-29r-22r-5-25
Superscaffold_5	8,111,364	9r-13-14

(b) The comparison between the initial assembly and OM assembly.

	Initial assembly	OM assembly
Genome size (bp)	90,815,494	91,454,494
Largest scaffold length (bp)	7,801,961	29,219,295
Scaffold number	640	6
Average scaffold in size (bp)	141,899	15,242,415

The comparison between the two assembly versions is listed in Table 7b. These superscaffolds have an average size of 15.24 Mb, ten-fold larger than the initial scaffolds. Given that we transposed the gene annotation from scaffolds to superscaffolds accordingly, the actual genomic sequences between the two versions have not changed because neither extra informative bases (A, T, C or G) were added (except concatenating 1 kb gaps) nor were gaps filled.

3.2.2 Assess the OM assembly

To validate this OM assembly, two approaches were used to confirm the continuity of these six superscaffolds. First, the synteny (the conservation of blocks of order within two sets of chromosomes that are being compared with each other) of the six superscaffolds was analyzed and the result indicates that no obvious blocks can be observed from superscaffolds 1 to 5, as shown in Figure 18. This suggests no large sequences were used as repeats. It is detected in that only in superscaffold_0, the synteny density is much higher. This is because these small scaffolds could not be placed by OM. Given that superscaffold_0 was assembled by concatenating unplaced small scaffolds, it is unavoidable to have relatively small similar sequences (Figure 20a). To further assess these superscaffolds, the initial Sanger reads for the *T. urticae* assembly were aligned back to these superscaffolds. Similarly, no obvious collapsed region or large gaps could be found in the Circos coverage map (Figure 20b), which hints that no large sequence is repetitively applied in this OM assembly. Regarding superscaffold_0, again due to the brevity of initial scaffolds, the coverage is not as good compared with the other superscaffolds. Conversely, due to the low reads coverage for these short scaffolds, they could not be assembled better in the first place. Meanwhile, because 596 gaps were used to bridge these unplaced small scaffolds, each 1 kb "N" was added in between (the gap number 596 is from 640 total scaffolds - 43 placed scaffolds - 1). To concatenate these short scaffolds, the coverage density is relatively less high than the no-large-gapped superscaffolds 1 to 5.

3.2.3 Re-annotation by cumulative evidence

Genomes are periodically re-annotated when new evidence becomes available (e.g. RNAseq data) or when a new assembly is released. Over the past six years after the initial release of *T. urticae* genome, more transcriptome data had accumulated and we (including experts in spider mite consortium) manually curated approximately 3,000 *T. urticae* genes (e.g. ABC transporters, chemosensory genes, and many hypothetical genes) (Dermauw *et al.*, 2013a; Ngoc *et al.*, 2016). Of these, 54% were merged from at least two separated flanking genes into one gene, 5% were chopped with extended starting codon, 30% were extended because of pre-terminating codons and nearly 5% were split into at least two individual genes (Figure 20c-d). Besides, there were also some previous annotation errors such as wrong splice sites, missing or overpredicted exons.

In addition to that, with the guidance of recently predicted genomes of the other two spider mite species (*T. lintearius* and *T. evansi*, details in Chapter 4), we used EUGENE and EVM to re-annotate the genome using the six superscaffolds from the updated OM assembly (Foissac *et al.*, 2008; Haas *et al.*, 2008). We also matched the previous and current gene models annotated at the overlapping genomic loci. Previous models were taken into consideration when double-checking these improved gene models. The updated annotation was synchronized in the ORCAE *T. urticae* database.

In this updated annotation version, we found over 1,800 additional genes, within which 83.4% of these extra genes are supported by RNAseq and 39.2% of them have assigned functional annotation. Of these genes, 29% (526 out of 1809) are related to the function of transmembrane-associated proteins but these genes could not be clustered in one family. This suggests that these transmembrane-associated genes are probably playing different roles in spider mites. We notice that these extra protein-coding genes are relatively short, as 62.8% of them have a length between 50-100 amino acids. Only 12.5% of these are longer than 200 amino acids.

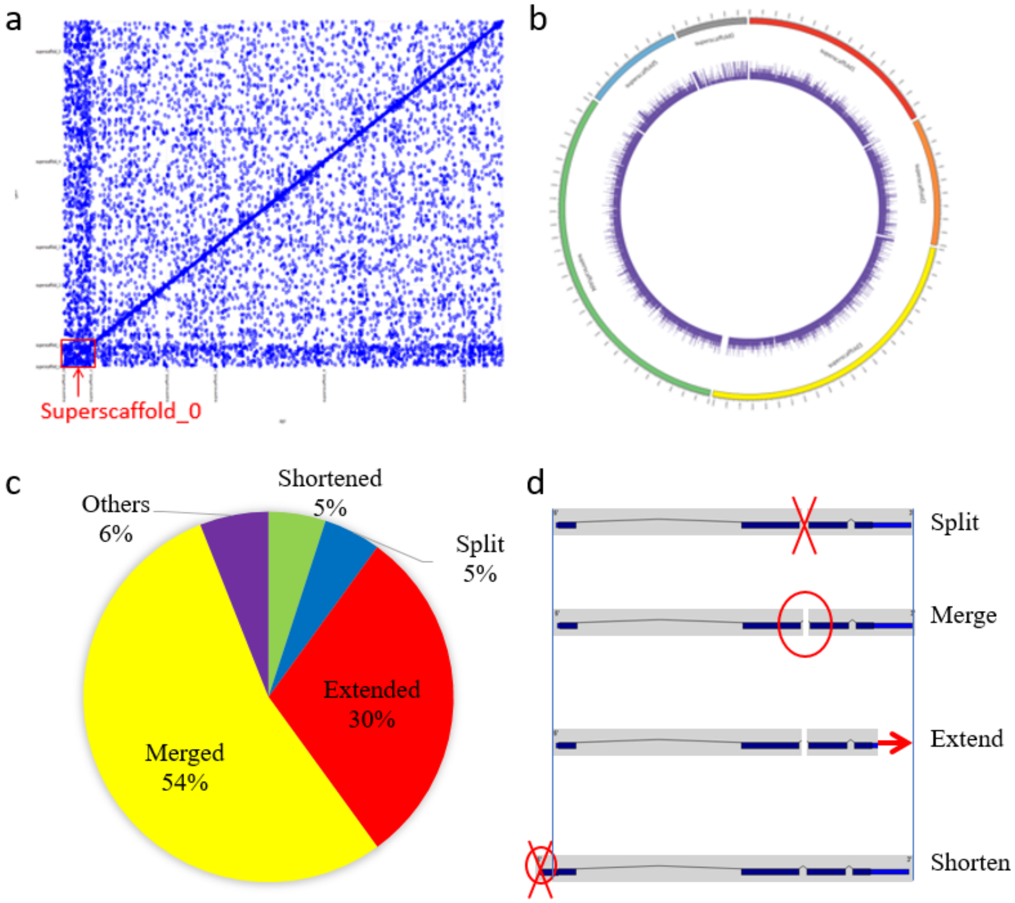


Figure 20: Assessment methods for superscaffolds and primary gene model errors. a: genomic synteny of *T. urticae* superscaffold assembly; b: initial genomic reads coverage mapping to the six superscaffolds; c, statistics of improved gene models; d, four major types of gene model errors.

Table 8: Statistics of annotation improvement.

	Previous version* V.S. Updated version
Extra predicted genes	1809
Improved gene models	1473 (manual curation)
TE	33
Others (inactive, truncated and pseudo)	279

*This was the latest version before updating. The initial version (Nov 2011) is not applicable because of manual curation on numerous gene models. Therefore, here we compared previous version (Feb 2016) with the latest updated version.

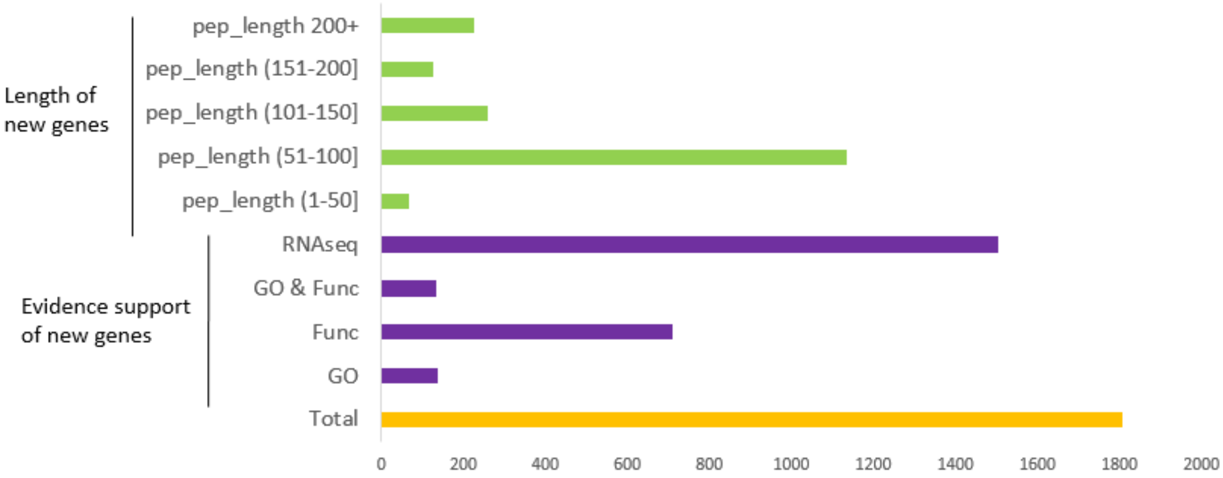


Figure 21: Newly predicted genes in *T. urticae* genome.
The numbers in brackets are the length of amino acids; Func is short for function; RNAseq is for RNAseq data evidence that supports gene models; Pep stands for protein.

3.3 Discussion

In this study, the assembly and annotation of the *T. urticae* genome were updated using both OM data and RNAseq hybrid data. This update is important for pest-plant interaction studies and the improved genome assembly will provide more insights into spider mite genomic structure because a chromosome-level assembly can reveal the extent of translocation and inversion polymorphism (Li *et al.*, 2016; Zapata *et al.*, 2016). OM technique can compensate for the low accuracy and high expense problems of the TGS long-reads sequencing methods.

Despite advances made in OM techniques, there are still some remain unresolved problems and challenges. The mapping data obtained are of relatively low resolution (Howe and Wood, 2015). In this study, we managed to assemble about 95% the scaffolds but some small breakpoints are known and need to be confirmed. For example, PCRs are supposed to be applied on detecting amplicons, which is a piece of DNA or RNA that is the source and/or product of natural or artificial amplification or replication events. It is reported that all spider mites have a haplodiploid sex-determination and the chromosome numbers are low ranging from n=2 to n=7 (Helle *et al.*, 1972; Bolland and Helle, 1981). These OM results suggest *Tetranychus urticae* has three or four chromosomes (if it is four, it consists 3 large chromosomes and another tiny one). Thus, it is still a challenge to assemble the *T. urticae* genome into three or four chromosomes. Additionally, we still have 1.8 Mb gaps detected by OM technique in the assembled genome (including the 596x1 kb and 42x1 kb concatenating gaps in superscaffold_0 and superscaffolds 1-5, respectively).

The spider mite genome, together with the favorable biological feature of the spider mites as a laboratory model including short generation cycle, easy breeding and established tools for gene analysis, has provided a novel genomic resource for studies of pest-plant interactions and development of alternative tools for plant protection (Grbic *et al.*, 2011a; Altincicek *et al.*, 2012; Dermauw *et al.*, 2013a; Ahn *et al.*, 2014; Martel *et al.*, 2015). In this study, the OM data has significantly improved the assembled scaffolds for the spider mite genome. The updated *T. urticae* OM assembly will facilitate genome-wide studies, especially comparative analyses across arthropods genomes at chromosome levels. Meanwhile, this updated annotation by cumulative RNAseq data and other reference data can offer more insights into gene model prediction methods, new genomic features as well as more evidence into a better accurate structural and functional annotation of the spider mite genome. Similar strategies of genome assembly and annotation will be applied to other genome assemblies as well in the future.

3.4 Materials and Methods

3.4.1 OM data and re-assembly

The OM protocols of both OpGen and BioNano were applied for the sample preparation. In short, DNA molecules on slides were stretched and fixed. OpGen processed the DNA after the protocol was optimized. Digestion with restriction enzymes was applied to relax of DNA and this allows visualization of gaps. OpGen collected seventeen high-density MapCards totaling 707,282 molecules with molecule size, average fragment size, and gap metrics all consistent with predicted metrics from the feasibility analysis of the NcoI enzyme. Thirteen and six linking maps were obtained from BioNano and OpGen, respectively. Bionano maps were generated with shorter DNA molecules (extraction protocol is still under development). The OM results from OpGen and BioNano have a few conflicts but these were mainly resolved by aligning BioNano maps on the five OpGen consensus maps. Thus, all the maps were joined with these scaffolds. Final OM data shows that one BioNano map joined two OpGen maps.

The initial scaffold_1, scaffold_2, scaffold_4 and scaffold_8 were split and reserved corresponding to OM data. We finally concatenated 42 scaffolds into five completely covered superscaffolds. We used 1 kb gap to concatenate the unmapped scaffolds from long to short order as superscaffold_0 with a genomic size of 5,657,691 bp. The placed scaffolds in the optical map were also concatenated using 1 kb “N” as gap bridges.

3.4.2 Re-assembly assessment

To validate the superscaffolds by OM, the initial Sanger genomic reads and Illumina RNAseq reads were mapped back to the superscaffolds by BWA and HISAT2, respectively (Li and Durbin, 2009; Kim *et al.*, 2015). Bedtools kit and in-house Perl scripts were used to analyze the genome coverage (Quinlan and Hall, 2010). GC-content and GC-skew were calculated by Perl with a sequence window of 10 kb. The super-scaffold synteny was aligned by MUMMER with default parameter (Kurtz *et al.*, 2004). The final figure was drawn by Circos for the overall genomic visualization (Krzywinski *et al.*, 2009).

3.4.3 Re-annotation

Previous annotation data were retrieved from ORCAE-MySQL database and converted into embl files (Sterck *et al.*, 2012). We kept all the annotated gene IDs unchanged between scaffolds to superscaffolds for further check-ups. After the preset, we used EUGENE and EVM to re-annotated the superscaffolds, combined with our cumulative RNAseq data, reference sequences from other two spider mites (details in Chapter 4) as well as previous *T. urticae* annotation (Foissac *et al.*, 2008; Haas *et al.*, 2008). Briefly, we employed optimized EUGENE pipeline and added BLASTX (protein reference of related species and the other two mites) as well as BLASTN (EST data, full-length cDNA, RNAseq-assembly (500 nt) and curated gene models (CDS)). RefSeq from old predictions was also used as references. As for the latest RNAseq data, we transformed them using Tophat2 as more precise junction data for splicing site prediction (Kim *et al.*, 2013). Additionally, NCBI BLAST hits were mapped to the superscaffolds as a reference by GenomeThreader (Gremme *et al.*, 2005). EvidenceModeler was employed to choose the best-predicted gene models (Haas *et al.*, 2008).

3.4.4 Re-annotation assessment

After the automatic structural re-annotation, we compared the new version with the previous version of the whole genome to correct small errors and mistakes. We also manually curated the biased gene models. The additionally predicted genes went through Blast2GO and InterProScan to detect GO, domains and predicted functions (Conesa *et al.*, 2005; Quevillon *et al.*, 2005; Mulder and Apweiler, 2007; Jones *et al.*, 2014). All the genome information such as homologs, domain, structure and function description were transposed to the updated OM assembled genome. All the data was formatted and submitted to ORCAE database at http://bioinformatics.psb.ugent.be/orcae.

3.5 Reference

Ahn, S.J., Dermauw, W., Wybouw, N., Heckel, D.G., and Van Leeuwen, T. (2014). Bacterial origin of a diverse family of UDP-glycosyltransferase genes in the Tetranychus urticae genome. Insect biochemistry and molecular biology *50*, 43-57.

Altincicek, B., Kovacs, J.L., and Gerardo, N.M. (2012). Horizontally transferred fungal carotenoid genes in the two-spotted spider mite Tetranychus urticae. Biology letters *8*, 253-257.

Bolland, H.R., and Helle, W. (1981). A survey of chromosome complements in the Tenuipalpidae. International Journal of Acarology *7*.

Chen, Q., Savarino, S.J., and Venkatesan, M.M. (2006b). Subtractive hybridization and optical mapping of the enterotoxigenic Escherichia coli H10407 chromosome: isolation of unique sequences and demonstration of significant similarity to the chromosome of E. coli K-12. Microbiology *152*, 1041-1054.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics *21*, 3674-3676.

Dermauw, W., Osborne, E.J., Clark, R.M., Grbic, M., Tirry, L., and Van Leeuwen, T. (2013a). A burst of ABC genes in the genome of the polyphagous spider mite Tetranychus urticae. BMC genomics *14*, 317.

Dong, Y., Xie, M., Jiang, Y., Xiao, N., Du, X., Zhang, W., Tosser-Klopp, G., Wang, J., Yang, S., Liang, J.*, et al.* (2013). Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). Nature biotechnology *31*, 135-141.

Foissac, S., Gouzy, J., Rombauts, S., Mathe, C., Amselem, J., Sterck, L., Van de Peer, Y., Rouze, P., and Schiex, T. (2008). Genome annotation in plants and fungi: EuGene as a model platform. Curr Bioinform *3*, 87-97.

Grbic, M., Van Leeuwen, T., Clark, R.M., Rombauts, S., Rouze, P., Grbic, V., Osborne, E.J., Dermauw, W., Ngoc, P.C., Ortego, F*., et al.* (2011a). The genome of Tetranychus urticae reveals herbivorous pest adaptations. Nature *479*, 487-492.

Gremme, G., Brendel, V., Sparks, M.E., and Kurtz, S. (2005). Engineering a software tool for gene structure prediction in higher organisms. Information and Software Technology *47*, 965-978.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome biology *9*, R7.

Helle, W., Boll, H.R., and Gutierrez, J. (1972). Minimal chromosome number in false spider mites Experientia *28*, 707-707.

Howe, K., and Wood, J.M. (2015). Using optical mapping data for the improvement of vertebrate genome assemblies. GigaScience *4*, 10.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G*., et al.* (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236-1240.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nature methods *12*, 357-360.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology *14*, R36.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome research *19*, 1639-1645.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome biology *5*, R12.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, J., Bian, C., Hu, Y., Mu, X., Shen, X., Ravi, V., Kuznetsova, I.S., Sun, Y., You, X., Qiu, Y*., et al.* (2016). A chromosome-level genome assembly of the Asian arowana, Scleropages formosus. Sci Data *3*, 160105.

Lim, A., Dimalanta, E.T., Potamousis, K.D., Yen, G., Apodoca, J., Tao, C., Lin, J., Qi, R., Skiadas, J., Ramanathan, A*., et al.* (2001). Shotgun optical maps of the whole Escherichia coli O157:H7 genome. Genome research *11*, 1584-1593.

Martel, C., Zhurov, V., Navarro, M., Martinez, M., Cazaux, M., Auger, P., Migeon, A., Santamaria, M.E., Wybouw, N., Diaz, I.*, et al.* (2015). Tomato Whole Genome Transcriptional Response to Tetranychus urticae Identifies Divergence of Spider Mite-Induced Responses Between Tomato and Arabidopsis. Molecular plant-microbe interactions : MPMI *28*, 343-361.

Mendelowitz, L., and Pop, M. (2014). Computational methods for optical mapping. GigaScience *3*, 33.

Migeon, A., Nouguier, E., and Dorkeld, F. (2006). Spider Mites Web: A comprehensive database for the Tetranychidae. Paper presented at: 12 international congrès of Acarology (Amsterdam, Netherlands: Springer).

Mulder, N., and Apweiler, R. (2007). InterPro and InterProScan: tools for protein sequence classification and comparison. Methods in molecular biology *396*, 59-70.

Nagarajan, N., Read, T.D., and Pop, M. (2008). Scaffolding and validation of bacterial genome assemblies using optical restriction maps. Bioinformatics *24*, 1229-1235.

Neely, R.K., Deen, J., and Hofkens, J. (2011). Optical mapping of DNA: single-molecule-based methods for mapping genomes. Biopolymers *95*, 298-311.

Ngoc, P.C., Greenhalgh, R., Dermauw, W., Rombauts, S., Bajda, S., Zhurov, V., Grbic, M., Van de Peer, Y., Van Leeuwen, T., Rouze, P.*, et al.* (2016). Complex Evolutionary Dynamics of Massively Expanded Chemosensory Receptor Families in an Extreme Generalist Chelicerate Herbivore. Genome biology and evolution *8*, 3323-3339.

Perry, D.A., Morrison, H.G., and Adam, R.D. (2011). Optical map of the genotype A1 WB C6 Giardia lamblia genome isolate. Molecular and biochemical parasitology *180*, 112-114.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic acids research *33*, W116-120.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Sterck, L., Billiau, K., Abeel, T., Rouze, P., and Van de Peer, Y. (2012). ORCAE: online resource for community annotation of eukaryotes. Nature methods *9*, 1041.

Wei, F., Zhang, J., Zhou, S., He, R., Schaeffer, M., Collura, K., Kudrna, D., Faga, B.P., Wissotski, M., Golser, W.*, et al.* (2009). The physical and genetic framework of the maize B73 genome. PLoS genetics *5*, e1000715.

Wu, C.W., Schramm, T.M., Zhou, S., Schwartz, D.C., and Talaat, A.M. (2009). Optical mapping of the Mycobacterium avium subspecies paratuberculosis genome. BMC genomics *10*, 25.

Zapata, L., Ding, J., Willing, E.M., Hartwig, B., Bezdan, D., Jiao, W.B., Patel, V., Velikkakam James, G., Koornneef, M., Ossowski, S., *et al.* (2016). Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. Proceedings of the National Academy of Sciences of the United States of America *113*, E4052-4060.

Zhou, S., Bechner, M.C., Place, M., Churas, C.P., Pape, L., Leong, S.A., Runnheim, R., Forrest, D.K., Goldstein, S., Livny, M., *et al.* (2007). Validation of rice genome sequence by optical mapping. BMC genomics *8*, 278.

Chapter 4

4 Comparative genomics of three spider mites reveals genome evolution and genomic signatures of adaptation to different feeding modes

Spider mites are major agricultural pests that cause millions of dollars' economic losses worldwide. The first spider mite genome of *Tetranychus urticae* has improved our understanding of pest-host adaptation and plant-herbivore interactions (Grbic *et al.*, 2011b; Zhurov *et al.*, 2014b). Across over 1,200 reported spider mite species, *T. urticae* displays polyphagous feeding lifestyle attacking more than different 1,100 plant species. In addition, there are other two types of feeding modes within this genus. *Tetranychus lintearius* is a monophagous species feeding on one host plant, gorse (*Ulex europaeus*) while *Tetranychus evansi* feeds almost exclusively on solanaceous plants (e.g. tomato and potato) displaying oligophagous feeding lifestyle. To date, *T. urticae* is the only completely sequenced *Tetranychus* genome and little is known about other *Tetranychus* species. Therefore, to understand the genetic and genomics variations across the three different mites, the genomes of *T. lintearius* and *T. evansi* were sequenced and annotated.

Here we report a comparative genomics study of three spider mites *T. urticae*, *T. lintearius* and *T. evansi*, associated with three respective feeding models: polyphagy, monophagy, and oligophagy, to dissect genomic basis of different feeding style taking advantage of their close phylogenetic relationship. Phylogenetic analysis shows the three mites diverged quite recently, only approximately three million years ago (MYA). The genomic sequences of the three mites are quite conserved in micro-synteny while transposable elements might play a role in shuffling and expanding some gene families and genome structure. Gene families that are associated with feeding and detoxification (e.g. chemosensory genes, P450 and ABC transporters), to some extent, proliferated in *T. urticae*. Moreover, some other gene families also have expanded or were lost during the evolutionary divergence of the three species.

The three *Tetranychus* genomes will markedly advance our understanding of genome evolution associated with pest feeding adaptations, agricultural plant-herbivore interaction studies and may further accelerate the development of environmentally sound pest control strategies that reduce environmental pollution and energy consumption in agriculture.

4.1 Introduction

Mites belong to the Chelicerata, representing a basal branch of arthropods. Mites exhibit tremendous variations in lifestyle ranging from parasitic to predatory to plant-feeding. Some mites (e.g., allergy-causing dust mites, scabies mites and mite vectors of scrub typhus) are of major concern to human health (Walter and Proctor, 1999). Some other mites (e.g., herbivorous spider mites and flat mites), however, are of great importance to agricultural crops (McCulloch, 1947; Sabelis, 1987; Bolland *et al.*, 1997; Flechtmann and Noronha, 2013; Van Leeuwen *et al.*, 2013).

The capability of herbivorous mites to feed on different host plants is due to their detoxification and digestion systems. Multiple genes associated with feeding and detoxification have been uncovered in the recent years. Cytochromes, also called P450, can metabolize most (lipophilic) xenobiotic compounds (Danielson, 2002). Glutathione-S-transferases (GST) can catalyze the conjugation of the reduced form of glutathione (GSH) to xenobiotic substrates in process of detoxification. Carboxyl/cholinesterases (CCE) have pivotal roles in dietary detoxification, pheromone or hormone degradation and neurodevelopment (Tsubota and Shiotsuki, 2010). ATP-biding-cassette transporters (ABC) utilize the energy of ATP binding and hydrolysis to transport various substrates across cellular membranes (Jones and George, 2004).

The polyphagous *T. urticae* was originally native to Eurasia but has acquired a cosmopolitan distribution (Donald M and Edward W, 1968). It was not until recently that global climate change has led to an emergence of this cosmopolitan agricultural pest *T. urticae*. It is known feeding on more than 1,100 host plants including many significantly economical plants (e.g., soy, maize and cotton), greenhouse crops (e.g., tomato, peppers and cucumbers) and horticultural plants (e.g., apple, pear and strawberry) (Grbic *et al.*, 2011a; Cazaux *et al.*, 2014). Its rapid life cycle and accelerating reproductive capability lead to significant agricultural economic damage worldwide, and thus it has become an established and emerging pest model on various crops (Muller-Scharer *et al.*, 2004; Lim *et al.*, 2011; Clotuche *et al.*, 2013). The recently sequenced genome of the two-spotted spider mite *T. urticae* has offered insights into herbivorous pest adaptation and plant-herbivore interactions (Grbic *et al.*, 2011a).

The monophagous *T. lintearius* originates from Europe and only feeds on one host plant *Ulex europaeus* (gorse), an important weed in some European countries. Because of its host specificity, *T. lintearius* is thus referred to as a monophagous spider mite. Heavy mite activity reduces flowering and stunts the development of the branches. *T. lintearius* has been introduced as a biological control agent to control gorse in New Zealand and Australia where this introduced plant proliferated producing problems in agriculture.

The oligophagous *T. evansi* is native to South America and has been accidentally introduced to other parts of the world, mainly spreading within Mediterranean countries as well as Africa. *T. evansi* prefers *Solanaceous* crops (e.g., tomato, potato and tobacco). However, it is also been found in several other vegetables (e.g. beans, citrus, and cotton) and ornamental crops (e.g. roses and cactus), as well as on many weed species (e.g., horseweeds, wall barleys and black nightshades) (Qureshi *et al.*, 1969; Tsagkarakou *et al.*, 2007; Gotoh *et al.*, 2010; Boubou *et al.*, 2011; Onyambus *et al.*, 2011; Navajas *et al.*, 2013a).

Traditional chemical methods often fail in controlling mites because the accelerated reproductive rate of spider mites allows their populations to quickly spread and develop resistance to pesticides, especially because of global warming. Indeed, *T. urticae* is considered a record-breaker in the development of pesticide resistance where it is recorded to be resistant on more than 90 chemical compounds. Often, after exposure to pesticides, *T. urticae* develops resistance in a period of 2-4 years after exposure (Van Leeuwen *et al.*, 2010; Dermauw *et al.*, 2013b; Van Leeuwen *et al.*, 2013). Thus, chemical control methods become less applicable when the same pesticide is used over a prolonged period. Therefore, genomic studies should shed light on the impact of pest feeding and detoxification mechanisms and lead to novel techniques in pest control against spider mites. It is hypothesized that there would be various combinations of genomic signatures associated with feeding and detoxication in spider mites. Here in this study, the genomes of *T. lintearius* and *T. evansi* were sequenced and annotated. We performed the comparative analysis of the three *Tetranychus* genomes to dissect genomic signatures of feeding mode evolution as well as to understand evolutionary forces that are shaping genome evolution. This study will not only provide genetic materials for arthropod genomic resources but also offer insights into pest-plant interactions and the development of new pest control tools based on genomic studies.

4.2 Results and Discussions

4.2.1 Genomic statistics of the three spider mites

The initial *T. urticae* genome (strain London) was sequenced using Sanger method to 8.05× coverage and assembled into 640 scaffolds covering 89.6 Mb genome size with 18,414 protein-coding gene models (Grbic *et al.*, 2011a). For *T. evansi* and *T. lintearius*, NGS Illumina short reads sequencing technology was applied for genome sequencing. The two spider mites have the same genome size of about 90 Mb. The largest scaffolds for *T. evansi* and *T. lintearius* are 1.4 Mb and 1.6 Mb, respectively. Multiple genomic characteristics of the *T. evansi* and *T. lintearius* correlate with their compact sizes: small transposable element content, low microsatellite density, and high gene density, which are all quite close to *T. urticae* (Grbic *et al.*, 2011a). Genomic synteny analysis shows that the three genomes are conserved in the microscale (10 kb, see Figure 22d). However, currently little is known about their synteny status at chromosome level due to lack of longer reads/assembly.

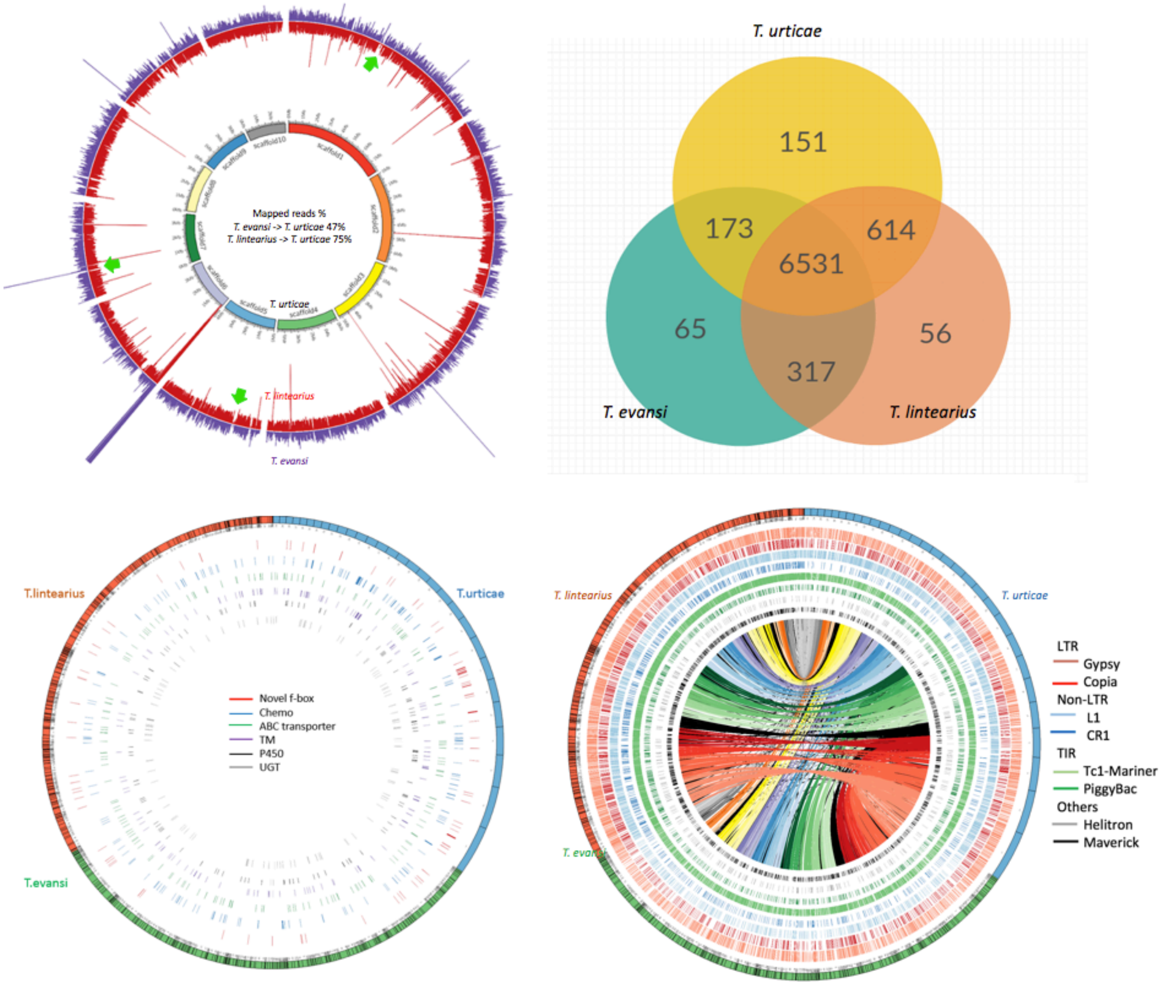


Figure 22: An overview of comparative analysis results of the three genomes.
a: reads mapping of *T. evansi* and *T. lintearius* to the assembly of *T. urticae*. This extremely long bar is a telomere region and thus it has many repeats leading to collapsed reads; b: Venn graph of gene family numbers across the three genomes; c: examples of expanded gene families locating on the three genomes; d: transposable elements expansion and genomic synteny in the three genomes.

Table 9: Genomic statistics of the three genomes.

	T. urticae	*T. evansi*	*T. lintearius*
genome size (scaffolds)	90,815,494 nt	91,505,123 nt	88,801,182 nt
genome size (contigs)	89,600,102 nt	82,282,823 nt	84,457,164 nt
largest scaffold	7,801,961 nt	1,473,105 nt	1,693,225 nt
av. scaffold length	141,899.21 nt	29,489.24 nt	47,640.12 nt
number of contigs	2,035	12,902	8,496
largest contig	929,118 nt	360,470 nt	277,343 nt
av. contig length	44,029.53 nt	6,377.52 nt	9,940.81 nt
gaps (>50N)	1,395 (1,215,392 nt)	5,548 (9,222,300 nt)	3,863 (4,344,018 nt)
Scaffold N50	2,993,488 bp	346,923 bp	374,049 bp
Scaffold L50	10	80	70
nr.big_introns	68	73	27
nr_loci (exons+introns)	19,043	15,376	15,028
av.length.loci	2,323.97 nt	2,895.36 nt	2,430.22 nt
loci density	4,705.39 nt/gene	5,351.38 nt/gene	5,619.99 nt/gene
nr_genes	19,042	15,376	15,028
gene density	212.52 genes/Mb	186.87 genes/Mb	177.94 genes/Mb
av.length.genes	1,108.23 nt	1,128.89 nt	1,099.06 nt
median.length.genes	825 nt	810 nt	798 nt
nr_exons	64,947	58,561	50,866
%GC of CDS	37.62	37.39	37.73
cumul_exon_length	21,102,957 nt	17,357,859 nt	16,5167,35 nt
av.length.exons	324.93 nt	296.41 nt	324.71 nt
median.length.exons	158 nt	143 nt	162 nt
longest.exons	45,659 nt (tetur30g00590.4)	42,418 nt (tetev263g00020.1)	14,619 nt (tetli109g00370.1)
av.nr.exons/gene	3.41	3.81	3.38
most exons/gene	55, tetur04g02800	41, tetev124g00030	36, tetli26g02110
cumul_CDS_length	20,307,982 nt	15,977,777 nt	16,516,735 nt
av.length.CDS	1,066.48 nt	1,039.14 nt	1,099.06 nt
cumul_intron_length	19,724,877 nt	22,288,029 nt	17,070,674 nt
av.length.intron	432.86 nt	524.75 nt	484.77 nt
median.length.intron	94 nt	113 nt	103 nt
%GC of intron	29.78	29.26	29.69

The annotation for *T. evansi* and *T. lintearius* were accomplished by an optimized EUGENE pipeline with a reference of updated *T. urticae* genome dataset (Chapter 3). The initial annotations of both genomes were compared with the *T. urticae* genome and over 7,000 genes across the three genomes were manually inspected and curated. The current version of *T. urticae* genome has 19,042 protein-coding genes while the other two spider mites have over 15,000. The protein-coding gene numbers across the three genomes have a subset of about 3,000 genes, most of which are hypothetical short genes in *T. urticae* without any detectable domains (50-100 amino acids). Mites (including *Tetranychus* and *Brevipalpus* genomes, details in Chapter 6) usually have a higher gene density (over 150 genes/Mb) in such compacted genomes, compared with other arthropods - much lower gene density ranging from *Stegodyphus mimosarum*'s 11 genes/Mb to *Pediculus humanus*' 98 genes/Mb, except *Drosophila melanogaster's* gene density at 181 genes/Mb (149Mb and protein-coding gene number 26,950) and *Daphnia pulex*'s 155 gene/Mb (details see Table S2).

4.2.2 A recent divergence of the three mites

Using the single copy genes obtained from the three genomes, the phylogenetic tree for the three mites was constructed using *Tribolium castaneum* (beetle) as an outgroup (Figure 23). Mites belong to the Acariformes with the earliest fossils dating 410 MYA (Hirst, 1923; Dubinin, 1962; Grbic *et al.*, 2011a). The phylogeny suggests that the three spider mites diverged about 3 MYA and additionally, *T. urticae* and *T. lintearius* have a more recent divergence, approximately 0.85 MYA. *T. evansi* is more ancient than *T. urticae* and *T. lintearius*, which hints that *T. urticae* probably has gained the capability of polyphagy during evolution after divergence while *T. lintearius* evolved into monophagy focusing on one host plant. It is assumed that during the rapid evolutionary process of three spider mites in such short period of time, gene gain and loss across various gene families would play a key role in terms of their feeding behaviors adapting to different host plants and the fast-changing environments (Jame, 1990; Magalhaes *et al.*, 2007; Dermauw *et al.*, 2013b).

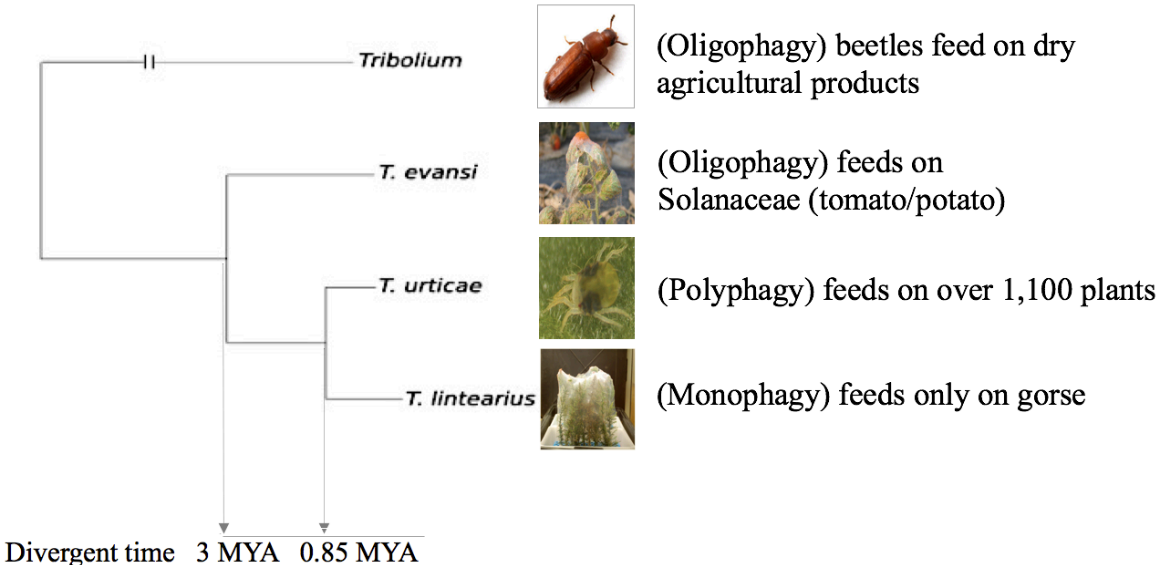

Tribolium
(Oligophagy) beetles feed on dry agricultural products
T. evansi
(Oligophagy) feeds on Solanaceae (tomato/potato)
T. urticae
(Polyphagy) feeds on over 1,100 plants
T. lintearius
(Monophagy) feeds only on gorse
Divergent time
3 MYA
0.85 MYA

Figure 23: Phylogenetic analysis shows a recent divergence of the three genomes. This phylogeny was constructed by Dr. Toni Gabaldon at the Center for Regulation Genomics in Barcelona, Spain.

4.2.3 Feeding and detoxification

The three genomes share 6,531 gene families, representing a majority of gene families determined in spider mite genomes (Figure 22b). Each spider mite has a few unique gene families, consisting of mostly short hypothetical sequences barely with known functions and often lacking RNAseq data supports. Previous studies have described that *T. urticae* is one of the most striking examples of polyphagy among herbivores and it has an unmatched ability to develop resistance to pesticides (Sabelis, 1987; Van Leeuwen *et al.*, 2010). Some essential gene families implicated in digestion, detoxification, and transport of xenobiotics have a unique composition in the genome of *T. urticae.* These feeding and detoxification gene families are often expanded when compared with insects (Grbic *et al.*, 2011a). In contrast to *T. evansi* and *T. lintearius*, we observed that these previously reported feeding and detoxification gene families have also proliferated in *T. urticae* (Table 11). For example, cytochrome P450, a protein that metabolizes most (lipophilic) xenobiotic compounds, is almost doubled in gene number in *T. urticae* in contrast to the other two mites, 86 compared with 41 and 35, respectively. Meanwhile, the gene copy numbers of GST and cholinesterase also have increased in *T. urticae* genome.

Chemosensory genes, especially of the perception of taste and smell, are important to animals in process of finding food. They primarily include gustatory receptors, olfactory receptors, and ionotropic receptors. A striking example of gene family proliferation in *T. urticae* is 689 gustatory receptor genes while there are only 227 and 258 in the other two mites (Phuong, 2014). These proliferated gene families probably hint that *T. urticae* has an unmatched ability to adapt to feeding upon more plants through rapid evolution than the other two mites.

Table 10: Key expanded gene families across the three genomes.

Note	*T. urticae*	*T. evansi*	*T. lintearius*	Gene Family/Function
Feeding and detoxification associated genes	86	41	35	P450
	16	13	11	Intradiol ring-cleavage dioxygenase
	32	18	20	Glutathione S-transferase
	689	227	258	Chemosensory-related gustatory genes
	71	59	53	CCE carboxyl/cholinesterases
	103	101	102	ABC-transporters
Expanded genes in *T. urticae*	95	57	65	UDP-Glycosyltransferase
	234	38	36	Novel F-box genes (NFB)
	88	9	16	Hypothetical Cell Surface Protein (HCSP)
	168	7	17	Hypothetical, not Glutathione S-transferase
	98	46	50	Dehydrogenase/reductase SDR
	83	14	5	BTB/Kelch-associated
	115	32	21	Apple-like transmembrane
Expanded genes in *T. lintearius*	26	14	63	Inhibitor of apoptosis (IAP)
	33	6	60	dUTPase

4.2.4 Protein-binding and transmembrane genes

As previously mentioned, there is a spider-mite-specific expansion of known gene families contributing to the ability of spider mites to overcome host defenses. Among these, genes with the most extreme expression fold-changes can encode putative secreted proteins or lipid-binding proteins, suggesting the extracellular binding and transport of small ligands are therefore likely to be important in further dissecting mite-plant interactions (Grbic *et al.*, 2011a). In addition to the known feeding and detoxification gene families that proliferated in *T. urticae*, we also observed some other proliferated gene families associating with protein-binding and transmembrane signaling process.

A novel F-box gene family expanded in *T. urticae* with over 230 copies, of these 188 are intact. Conventionally, F-box genes play a role in protein-protein-interaction and protein degradation based on the ubiquitination. In *Drosophila*, F-box proteins function in various cellular settings such as tissue development, cell proliferation, and cell death (Ho *et al.*, 2006). However, the function of this novel F-box family in *T. urticae* is unknown, but these expanded F-box genes could function as a mite response to toxic plant defenses (details see Chapter 5).

BTB-Kelch-associated genes, containing an N-terminus BTB domain and the C-terminus Kelch motifs, have a copy number of 83 in *T. urticae*, but only 14 and 5 in *T. evansi* and *T. lintearius*, respectively. These genes facilitate protein binding and dimerization. Kelch domains form a tertiary structure of β-propellers that have a role in extracellular functions, morphology, and binding to other proteins (Dhanoa *et al.*, 2013). The BTB-ZF proteins are encoded by at least 49 genes in mouse and man and commonly serve as sequence-specific silencers of gene expression (Siggs and Beutler, 2012). The large expansion of this gene family in *T. urticae* suggests more cellular function activity and protein binding-related process are required in *T. urticae.*

In addition, some gene families are implicated in transmembrane functions also have expanded in *T. urticae*. These transmembrane associated gene families function as gateways to permit the transport of specific substances and signals across the biological cell membranes. In *T. urticae,* there are 115 genes containing Apple domain (in the shape of an apple and has been accordingly called apple domain) while the other two mites only contain 32 and 21, respectively. The apple domain is a subset of the PAN domain superfamily and is widely detected in various organisms, including bacteria, apicomplexans, filamentous fungi, plants, nematodes, amphibians, avians, and mammals. The PAN/Apple domain mediates protein-protein or protein-carbohydrate interactions (Tordai *et al.*, 1999). The PAN proteins have especially been studied in apicomplexans (e.g. Plasmodium and Toxoplasma) where they play a critical role in host invasion (Brown *et al.*, 2001; Carruthers and Tomley, 2008). It is reported that the apple domains of plasma prekallikrein can mediate its binding to high molecular weight kininogen (Herwald *et al.*, 1996). The apple domains of factor Xi can also bind to factor XIIa, platelets, kininogen, factor IX and heparin (Ho *et al.*, 1998). The PAN family members have no documented homologs in arthropods, and currently little is known about their function.

A gene family entitled hypothetical cell surface proteins (HCSPs) with a structure of single exon is expanded with a gene copy number of 88 in *T. urticae*. These HCSPs only have one detected transmembrane domain at C-terminus region (Supplement: Figure 25, e.g. tetur09g07110 using TMHMM (Krogh *et al.*, 2001)). These HCSPs may act in signaling transduction in spider mites.

There are 18 phospholipid scramblase proteins found in *T. urticae*, compared with 2 and 3 in *T. evansi* and *T. lintearius*, respectively. These scramblases are normally in the cell membrane and transporting (scramble) the negatively charged phospholipids from the inner leaflet to the outer leaflet and vice versa (Bevers and Williamson, 2010). The expansion of phospholipid scramblase proteins in *T. urticae* may have the original function of these membrane proteins (Yu *et al.*, 2015; Bevers and Williamson, 2016).

4.2.5 Gene families expanded in *T. lintearius*

A few expanded gene families in *T. lintearius* also emerged in our gene family cluster analysis. Respectively, the inhibitor of apoptosis proteins (IAP) and the dUTP diphosphatase (dUTPase) have almost doubled in *T. lintearius* in contrast to those in *T. urticae*. The IAP gene family serves as endogenous inhibitors of programmed cell death, called apoptosis. The dUTPase proteins can remove dUTP from the deoxynucleotide pool, which reduces the probability of this base being incorporated into DNA by DNA polymerases. Lack or inhibition of dUTPase action leads to harmful perturbations in the nucleotide pool, resulting in increased uracil content of DNA that activates a hyperactive futile cycle of DNA repair (Vassylyev and Morikawa, 1996; Vertessy and Toth, 2009). Both IAP and dUTPase serve the functionality of cell and DNA maintenance. However, there is no evidence to date suggesting that *T. lintearius* has an increased longevity, compared with *T. urticae* and *T. evansi*.

No gene families have proliferated in *T. evansi* (Figure 24). It is probable that *T. evansi* represents an ancestral state while *T. urticae* and *T. lintearius* have been evolving somehow more rapidly, thus both of them have dynamic gene gain and gene loss that detected in their genomes.

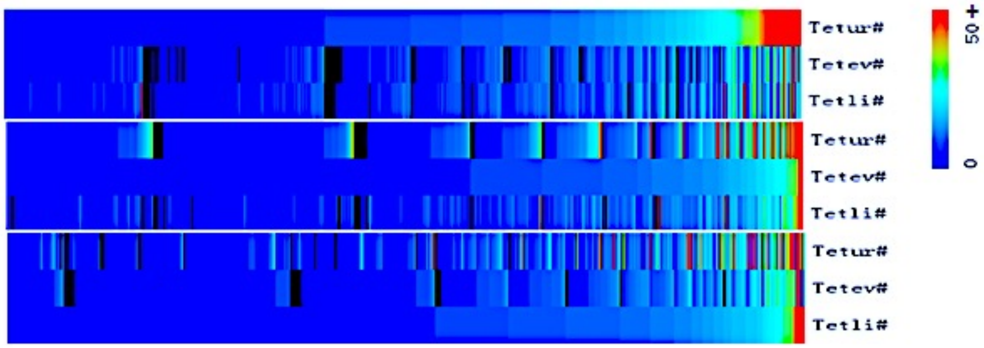


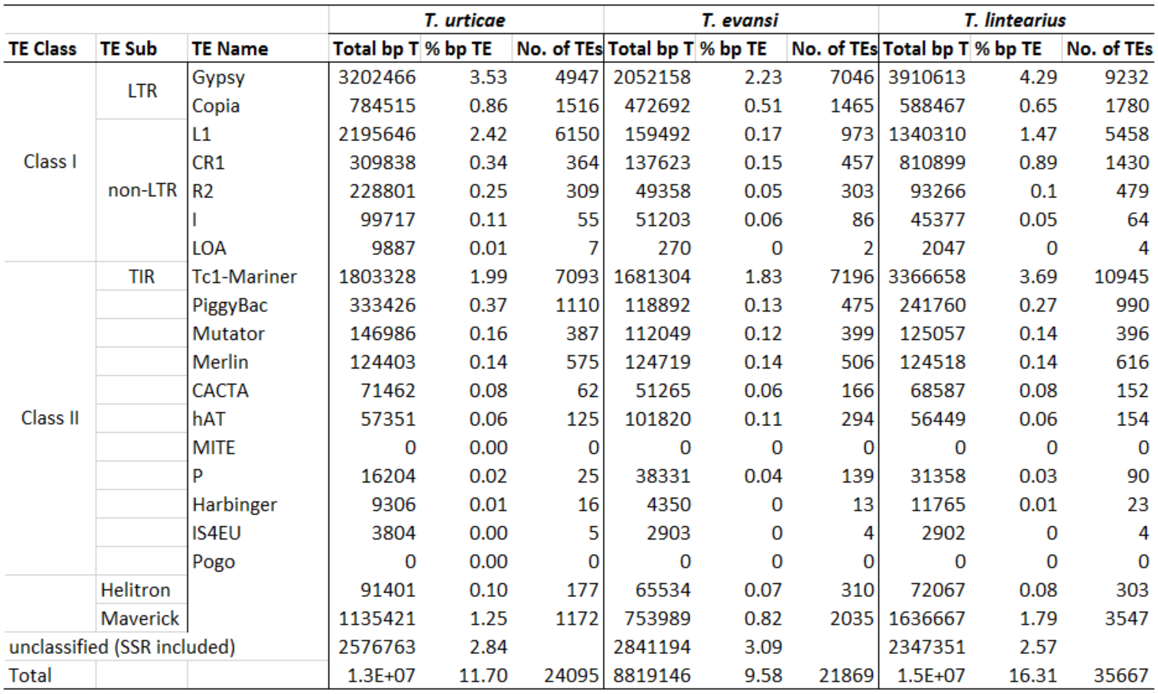
Figure 24: Heatmap of gene family number across the three genomes.
Top 500 gene families are sorted by total number by *T. urticae*, *T. evansi,* and *T. lintearius* from top to bottom, respectively. The top figure shows the large expansion of *T. urticae* while the last two figures show rare expansion in *T. evansi* and *T. lintearius*; Tetur – *T. urticae*; Tetli – *T. lintearius*; Tetev – *T. evansi*; # - gene number;

4.2.6 TE expansion

The previous study reports that *T. urticae* has a TE proportion of 9.09 Mb, taking up 10.15% of the whole genome (Grbic *et al.*, 2011a). With the trained TE library from that study, we masked the genomes of *T. lintearius* and *T. evansi*, in which there are 16.31% and 9.58% TE, respectively. In contrast to *T. lintearius*, *T. evansi* and *T. urticae* have quite similar TE proportions across the whole genomes (35,667 copies in *T. lintearius*, 24,095 in *T. urticae* and 21,869 in *T. evansi*).

Strikingly, TE class I Gypsy in *T. lintearius* has a larger number, almost doubled compared with that in *T. urticae* (9,232 vs 4,947). Gypsy belongs to Long Terminal Repeat (LTR) retrotransposons, which range from over 100 bp to over 5 kb in size. Gypsy is found in high copy number (up to a few million copies per haploid nucleus) in animals, fungi and plants genomes. They encode at least four protein domains in the following order: protease, reverse transcriptase, ribonuclease H, and integrase. In *Drosophila*, Gypsy is the cause of numerous spontaneous mutations (Peifer and Bender, 1988; Dorsett *et al.*, 1989; Flavell *et al.*, 1990). Gypsy also encodes putative gene products which are homologous to retroviral proteins (Marlor *et al.*, 1986). The high content of Gypsy in *T. lintearius* might accelerate the duplication of certain gene families, because Gypsy in Class I, as retrotransposon, may be actively involved in genome evolution. It is tempting to propose that these increased content of Gypsy in *T. lintearius* could shuffle its genomic structures through insertion and deletion, shaping the evolution of *T. lintearius* genome.

Table 11: TE distribution across the three genomes.



			T. urticae			*T. evansi*			*T. lintearius*		
TE Class	**TE Sub**	**TE Name**	**Total bp T**	**% bp TE**	**No. of TEs**	**Total bp T**	**% bp TE**	**No. of TEs**	**Total bp T**	**% bp TE**	**No. of TEs**
Class I	LTR	Gypsy	3202466	3.53	4947	2052158	2.23	7046	3910613	4.29	9232
		Copia	784515	0.86	1516	472692	0.51	1465	588467	0.65	1780
	non-LTR	L1	2195646	2.42	6150	159492	0.17	973	1340310	1.47	5458
		CR1	309838	0.34	364	137623	0.15	457	810899	0.89	1430
		R2	228801	0.25	309	49358	0.05	303	93266	0.1	479
		I	99717	0.11	55	51203	0.06	86	45377	0.05	64
		LOA	9887	0.01	7	270	0	2	2047	0	4
Class II	TIR	Tc1-Mariner	1803328	1.99	7093	1681304	1.83	7196	3366658	3.69	10945
		PiggyBac	333426	0.37	1110	118892	0.13	475	241760	0.27	990
		Mutator	146986	0.16	387	112049	0.12	399	125057	0.14	396
		Merlin	124403	0.14	575	124719	0.14	506	124518	0.14	616
		CACTA	71462	0.08	62	51265	0.06	166	68587	0.08	152
		hAT	57351	0.06	125	101820	0.11	294	56449	0.06	154
		MITE	0	0.00	0	0	0	0	0	0	0
		P	16204	0.02	25	38331	0.04	139	31358	0.03	90
		Harbinger	9306	0.01	16	4350	0	13	11765	0.01	23
		IS4EU	3804	0.00	5	2903	0	4	2902	0	4
		Pogo	0	0.00	0	0	0	0	0	0	0
	Helitron		91401	0.10	177	65534	0.07	310	72067	0.08	303
	Maverick		1135421	1.25	1172	753989	0.82	2035	1636667	1.79	3547
unclassified (SSR included)			2576763	2.84		2841194	3.09		2347351	2.57	
Total			1.3E+07	11.70	24095	8819146	9.58	21869	1.5E+07	16.31	35667

*This table was accomplished by RepeatMasker using *T.urticae* trained TE library in the default parameter. It is possible that *T. lintearius* had transposition and translocation by TE expansion genome structure has been shuffled, probably for adaptations in gorse.

Except for Gypsy, some other types of TE such as CR1, Tc1-Mariner, and Maverick, also have slightly increased in *T. lintearius*. Tc1-Mariner belongs to TE class II and acts as a cut-and-paste function, which in many cases, aids genomic sequences to ‘jump over’ across the whole genome, and consequently facilitates genomic rearrangements (Cordaux and Batzer, 2009). It is also reported that TE may have played a role in the observed structural complexity of some large gustatory clusters in *T. urticae* (Ngoc *et al.*, 2016). In sum, it is assumed that these expansions of TE subclasses would be a drive for the genomic shuffling in *T. lintearius* during evolution.

We analyzed different TE types and copy number variations around the flanking regions (using 5 kb, 10 kb, 15 kb and 20 kb as region window, respectively) of gene families of interest in this study, Interestingly, we also found that the expanded gene families across the three genomes typically have a higher Gypsy number around the flanking regions (Table 12 and Supplement Table 14-16). Especially, the Novel F-box gene family has stronger evidence that Gypsy number is extremely higher than those homologous genes in the other two genomes. Chemosensory gene families, however, have an increased copy number of Gypsy in *T. lintearius*, rather than in other two genomes. It might be suggesting that Gypsy might not only play a gene expansion function (copy and paste) but also other functions (e.g. shuffling the genome structure).

Table 12: TE statistics in the flanking region of 10 kb of key gene families.

Flanking region 10 kb	*Gene Family*	*Gypsy*	*Copia*	*L1*	*CR1*	*Mariner*	*PiggyBac*	*Helitron*	*Average TE Per Gene*
T. evansi	Chemo	1.6	1	0.17	0.08	3.24	0.27	0.08	6.43
	P450	1	0.2	0.49	0.07	2.27	0.07	0.05	4.15
	NFB	0.53	0.24	1.06	0.06	2.71	0.35	0	4.94
	ABC	0.81	0.32	0.23	0.04	1.39	0.1	0.04	2.93
	dUTPase	3.33	0	0.17	0.17	3.33	0	0.17	7.17
	IAP	3.86	0.79	0.07	0	1.79	0	0.21	6.71
	IDRCD	1.58	0	0.5	0	2.25	0.17	0.08	4.58
	SDR	0.63	0.11	0.3	0.04	1.26	0.07	0.07	2.48
	HCSP	8	0.56	0.67	0	1.44	0	0	10.67
T. lintearius	Chemo	6.1	3.08	2.5	0.27	5.4	0.85	0.12	18.33
	P450	1.66	0.4	2	0.2	4.2	0	0.06	8.51
	NFB	3	0.35	2.41	0.35	3.24	0.18	0	9.53
	ABC	1.48	0.38	2.28	0.25	2.49	0.24	0.09	7.22
	dUTPase	2.65	0.32	0.5	0.2	2.98	0.15	0.12	6.92
	IAP	4.43	1.11	0.48	0.13	3.79	0.27	0.11	10.32
	IDRCD	1.4	0.2	1.3	1.2	3.3	0.2	0.1	7.7
	SDR	2.4	0.3	1.28	0.68	2.74	0.16	0.04	7.6
	HCSP	6.38	1.06	2.44	0.38	2.38	0.19	0.25	13.06
T. urticae	Chemo	1.79	2.43	1.54	0.09	2.89	1.17	0.02	9.91
	P450	1.16	0.35	3.07	0.26	3.84	0.21	0.02	8.9
	NFB	5.73	0.56	2.45	0.24	2.64	0.44	0	12.06
	ABC	1.14	0.35	3.34	0.02	1.8	0.23	0.02	6.89
	dUTPase	1.21	0.15	0.79	0.06	0.82	0.12	0.03	3.18
	IAP	1.35	0.73	0.88	0.08	2.69	0.23	0.04	6
	IDRCD	0.88	0.62	2.69	0.06	1.5	0.25	0	6
	SDR	1.14	0.31	1.68	0.07	1.09	0.01	0	4.31
	HCSP	11.47	0.91	2.18	0.01	1.95	0.15	0.05	16.72

4.3 Conclusion

Two *Tetranychus* spider mite genomes were sequenced, annotated, and comparatively analyzed their genomic organization with *T. urticae* to explore the three different feeding behaviors: polyphagy, monophagy, and oligophagy. The results show that the three spider mites diverged 3 MYA and *T. evansi* may represent the ancestral state of the three mites. It also shows that feeding and detoxification associated gene families in polyphagous *T. urticae* expanded at different levels, compared with monophagous *T. lintearius* and oligophagous *T. evansi*. The three genomic assemblies show a conserved synteny from the micro-scale. However, little is known whether they also share a conserved synteny from the macro-scale due to lack of longer assembly. Nevertheless, it is observed that the TE contents in *T. lintearius* apparently are higher than in the other two spider mites, suggesting these TE probably played a role in shuffling the spider mite genome structure and accelerating the divergence of the three species. Interestingly, TE density around expanded gene families, in general, is also higher than around the non-expanded gene families. This implies the hypothesis that TE might be a drive for these gene families' expansion.

The characterization for the feeding and detoxification associated gene families adds to a growing body of evidence that lineage-specific expansions of genes in this polyphagous herbivore *T. urticae,* associated with polyphagous feeding strategy. Its populations have also been documented to vary in host plant adaptations (Fellous *et al.*, 2014). This provides an exciting opportunity to understand the micro-evolutionary forces at the population level that would underline diversification in a genetically tractable herbivore (Ngoc *et al.*, 2016). The sequencing of two additional spider mite genomes will not only provide opportunities to understand the evolution of pest plant interactions but also genomic tools necessary for the development of new pest control techniques and approaches.

4.4 Materials and Methods

4.4.1 Strain selection and DNA preparation.

Briefly, we collected about 0.5 mL of spider mite eggs and followed the Illumina protocol for DNA preparation. (Note details can be found in the Supplementary Protocol section of this thesis. This biological part was done by my colleagues).

4.4.2 Genome sequencing and assembly

T. evansi was sequenced using Illumina short reads with mate-pair (5 kb) and pair-end (300 bp and 500 bp). *T. lintearius* was also sequenced by Illumina but with single reads and mate-paired reads. The total coverage for *T. evansi* and *T. lintearius* are over 100x, respectively. We employed the commercial tools CLCBio (https://www.qiagenbioinformatics.com/) to assemble the paired-end reads and SSPACE for scaffolding with the mate-pairs (Boetzer *et al.*, 2011).

To validate our assembly for *T. evansi* and *T. lintearius*, we mapped the genomic reads of these two genomes to the assembly of *T. urticae* and used the top 10 scaffolds of *T. urticae* and reads coverage by the other two genomes. No obvious large gaps or collapsed loci can be found, which suggests our assembly for *T. evansi* and *T. lintearius* have no apparent artifacts. Although we do observe a few small gaps and bars from the mapped coverage, they have possibly expanded gene families or assembly technical problems that will be improved and double-checked with longer sequencing reads in future.

4.4.3 Removal of contaminated scaffolds

To identify the contaminated scaffolds of the three genomes (version 20160229), we applied three approaches, from both scaffold level and protein-coding gene level. First, all the raw scaffolds were scanned against NCBI nrDNA database (BLASTN, version 20160402). If a scaffold (e-value < 1-5e) returns all hits from the prokaryotic origin, then this scaffold was discarded. Second, as for the scaffolds that are potentially remotely homologous to prokaryotes, we ran BLASTX against the nrProt database (version

20160409) to search hits from prokaryotic genes. When all hits were returned, and if no other hits could be shown as being from eukaryotic species, then this scaffold was labeled as prokaryotic. Third, protein-coding genes from the three mites (version 20160229) were compared to the NCBI Protein Database (BLASTP, nrProt version_20160317). If most of the genes on a scaffold whose best hits (e-value < 1-5e) are from prokaryotic genomes, then we manually inspected such scaffold before discarding it. These scaffolds should have only genes with a prokaryotic signature (single exon, no introns, etc.). Most of the scaffolds in the assembly found this way, are quite short and only consist of two or three predicted prokaryotic genes on the whole scaffold. By these three approaches, confirmed contaminated scaffolds with their genes were discarded from the draft assembly.

4.4.4 Assembly assessment for the genome completeness

All the quality controlled genomic reads of *T. lintearius* & *T. evansi* were mapped back to *T. urticae* genome by CLCbio tool CLC_mapper (http://www.clcbio.com/) (Cock, 2013). The circos shows top 10 scaffolds of *T. urticae* and mapped reads coverage of other two genomes. In brief, we applied CLC_mapper to map the raw reads of *T. lintearius* and *T. evansi* to *T. urticae* genome. The percentages of mapped reads were calculated by Samtools stats and Plot-bamstats (Li *et al.*, 2009a) (http://bamstats.sourceforge.net/). Then we used the Samtool sort and Bedtools Genomecov to calculate the mapped reads number (Li *et al.*, 2009a; Quinlan and Hall, 2010). In-house Perl scripts were used to extract and format the data for Circos visualization (Krzywinski *et al.*, 2009).

Similarly, all the quality controlled RNAseq and *de novo* transcripts data were mapped back to *T. evansi* and *T. lintearius* genomes, respectively. Briefly, we used HISAT2 to map the quality controlled reads back to each genome with a max-intron length of 90 kb by its sensitive single reads mapping method (Kim *et al.*, 2015). The output bam files of HISAT2 were also calculated by Samtools stats and Plot-bamstats (Li *et al.*, 2009a) (http://bamstats.sourceforge.net/).

To check the completeness of the assemblies, BUSCO were run both at the genomic sequence level and gene set level (Simao *et al.*, 2015). Because BUSCO arthropod database is biased to insects rather than chelicerate genomes, our results can only show the assembly completeness for the three genomes by a similar percentage, either from genomic level or gene set level. BUSCO predicted 542 missing genes in *T. urticae* (out of 2,675). We used the hardware-accelerated Decypher-blast algorithm (version decypher/x86_64/2, eval 1-5e) to blast these 542 missing genes by BLASTP and TBLASTN. Respectively, 305 genes and 311 small genomic fragments were retrieved. We used InterProScan to annotate the function of these missing genes, most of which only have general functional descriptions (Jones *et al.*, 2014).

4.4.5 Synteny analysis and visualization

We employed the *T. urticae* genome sequence as a reference and assigned scaffold synteny of the other two genomes to its top 30 scaffolds using i-adhore in scaffold-scale (Proost *et al.*, 2012). Then we used 10 kb windows to investigate small synteny by NUCMER (Delcher *et al.*, 2002). Lacking any information for longer scaffolds for *T. evansi* and *T. lintearius*, we focused on potential micro-synteny of three genomes also by using Circos as a visualization tool (Krzywinski *et al.*, 2009). We assigned 30 colors for the top 30 scaffolds of *T. urticae*, matching its corresponding 10 kb window sequences to the loci of other two genomes.

4.4.6 Structural annotation

We used trained gene set from *T. urticae* for the structural prediction of *T. evansi* and *T. lintearius* by our optimized EUGENE pipeline (Foissac *et al.*, 2008). Coding-potential was modeled with Hidden Markov Model. RNAseq and EST datasets from *T. evansi* and *T. lintearius* were used as BLASTN input for EUGENE. Typically, protein database uniport was used for quality improvement as well (UniProt, 2015; Dogan *et al.*, 2016). RepeatLib from *T. urticae* (Spidermite_TElib_300310.nt.tfa) was performed as a repeat masking tool for EUGENE (Foissac *et al.*, 2008; Grbic *et al.*, 2011a). The latest *T. urticae* gene annotation version was applied to a RefSeq genome dataset for the evaluation of structure annotation. Aiding in automatically annotated by EUGENE, approximate 7,000 genes across the three genomes with mispredicted structure were identified in accordance with gene alignments. To better improve the structural annotation of three genomes, we reviewed these genes and curated them manually based on the transcriptomic data, gene structure (exons and introns) and sequence alignments. Additionally, to those genes of interests, we manually checked these genes in each gene family for a precise and correct gene model, especially for pseudogenes. We also used BLASTN and BLASTP to search the whole genome to fish missing annotated genes as well as genes of interest.

4.4.7 Functional annotation

We used BLAST2GO and InterProScan for the functional annotation for *T. evansi* and *T. lintearius (Aparicio et al., 2006; Gotz et al., 2008)*. The raw results were filtered by in-house Perl scripts to assign genes functional descriptions.

4.4.8 ncRNA annotation

All the ncRNA genes were screened using Infernal and Rfam databases across the *T. evansi* and *T. lintearius* genomes using default parameters (Griffiths-Jones *et al.*, 2003) (updated on May 20, 2016). The raw result then was filtered by in-house Perl script after quality control. The in-house script assigned the predicted ncRNA genes into seven categories as rRNA, tRNA, sRNA, snRNA, miRNA, snoRNA, spliceosomal RNA and the rest were assigned to 'other RNA types'. All the genomic information for the three genomes is available at http://bioinformatics.psb.ugent.be/orcae on the ORCAE database (Sterck *et al.*, 2012).

4.4.9 TE annotation, visualization, and statistics

Initially, we annotated all the TEs across the three genomes by RepeatMasker based on *T. urticae* TE trained library (Smit and Hubley, 2008-2015; Tarailo-Graovac and Chen, 2009) (http://www.repeatmasker.org). All the TE IDs are currently set as the format of tetxx##te##### (## means the scaffold number and ##### represents five digitals of the gene ID on the corresponding scaffold). Then we double-checked these suspicious genes to see if they were contamination, TEs or hypothetical genes resembling TEs. Most TE and hypothetical protein-coding genes can be possibly mixed in one gene family. If a gene is sitting in a TE locus but has no clear TE-related domain, we define it a hypothetical gene. If many copies of these hypothetical genes can also be found un-unattached to TEs, we assume these the genes have "hitchhiked" with TEs.

To further investigate if more TEs were mistakenly annotated as protein-coding genes, we scanned all the genes from the three genomes through IPRSCAN nrProt and RepBase according to the following protocol (Jurka *et al.*, 2005; Jones *et al.*, 2014): the RepBase is relatively small and can cause a bias, probably returning a hit when the initial query is not necessarily a true TE. Therefore, an extra filtering needs to be included relying on the gene-family composition table: if that query is a member of a gene family that overall has no similarity to TE, then this member will remain as a 'hypothetical gene'. Finally, all the confirmed TEs are marked as inactive genes in current ORCAE database (Sterck *et al.*, 2012). We selected Gypsy, Copia, L1, CR1, Mariner, PiggyBac and Helitron to visualize their position and proportion on Circos (Krzywinski *et al.*, 2009).

We investigated TE flanking some gene families by counting the number of TE in the 5 kb, 10 kb, 15 kb and 20 kb (in the case of the diverse lengths of different TE types) flanking region (both forward and backward) each gene of interest. If a TE is shared by more than one gene, then we assigned this TE for each gene independently (i.e., count twice).

4.4.10 Orthologous gene identification

We downloaded all the protein-coding genes of the three mites (version 20160408) and built the gene family table based on the standard protocol of OrthoMCL (Li *et al.*, 2003). Briefly, all the protein-coding genes were filtered by quality control (min_length 10, max_percent_stops 20). All the defined "good protein" genes were compared against themselves (BLASTP, all-against-all, e-value < 1-5e) using the hardware-accelerated decypher-blast algorithm (version decypher/x86_64/2). The "query-hit" pairs were further processed using the OrthoMCL pipeline to cluster the gene families (Fischer *et al.*, 2011). A Venn graph was drawn to show overlapping gene families across the three data sets.

OrthoMCL-DB ID and weight of each gene were calculated (Chen *et al.*, 2006a; Fischer *et al.*, 2011) as follows: initially we retrieved the latest OrthomclDB (20160415) and employed decipher for BLASTP (e-value < 1-5e) all the genes. The best hit with an orthomcl ID was assigned to each gene. We counted the number of ID as weight in each orthomcl ID.

After the gene family table was set and finalized, we matched all gene IDs to the function annotation information (version 20160405). In the study, we used the top 30 scaffolds of *T. urticae* as a reference and showed six expanded gene families from *T. urticae* compared with the other two genomes by Circos (Krzywinski *et al.*, 2009). The higher density of clusters on *T. urticae*, for instance, indicates higher expanded gene families.

4.5 Supplementary Information

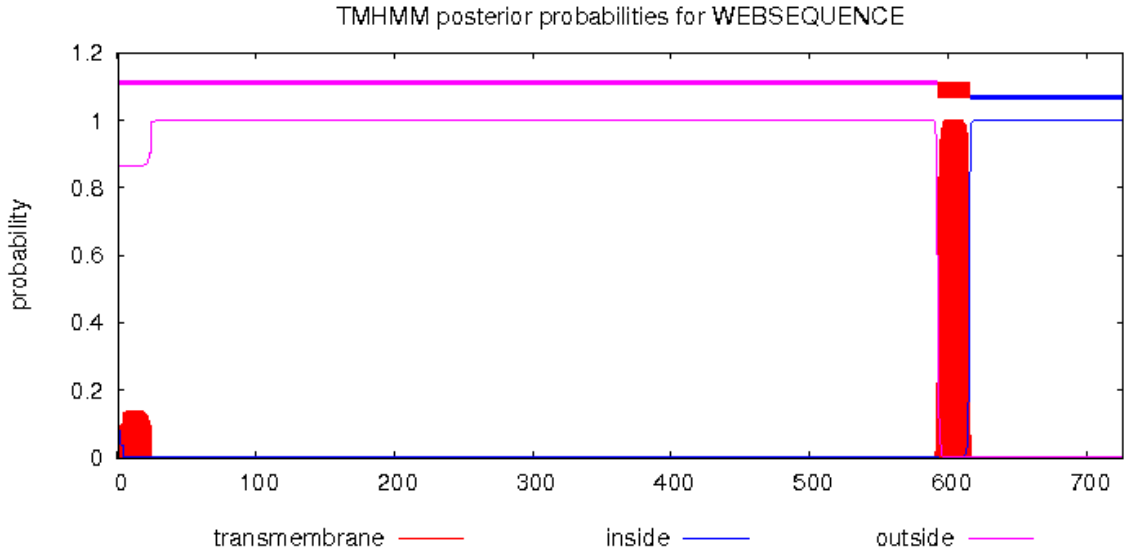


Figure 25: HCSP transmembrane structure of tetur09g07110.

This figure was updated on June 9, 2017.

```
# WEBSEQUENCE Length: 726
# WEBSEQUENCE Number of predicted TMHs: 1
# WEBSEQUENCE Exp number of AAs in TMHs: 25.78436
# WEBSEQUENCE Exp number, first 60 AAs: 2.95923
# WEBSEQUENCE Total prob of N-in: 0.13693
WEBSEQUENCE          TMHMM2.0      outside    1 592
WEBSEQUENCE          TMHMM2.0      TMhelix  593 615
WEBSEQUENCE          TMHMM2.0      inside     616 726
```

```
> tetur09g07110
MFIHLLLIIWTFQFCLLIQETYSFRTPKHDSLFYYKTHASVFLTNPLNHTTSYGIYVAGQTITVDIPTS
VANVFDILNWKVINVKKNQLMFVHQNKPYILINQTIISEMEYSGELSDSIIAFGDNEALHVPTIFNPN
LTKPIPDWNYIELLHFDDQSEKVSVSRFLPWLKDDWKFIKEWNMTDYIHFDNKLYLAIKRSIWNEK
SAKVTQEISIVRLCLDKGSELISSAVEIHFTQEAFENNKIIDLFFVFLSGPLITENQRYQLHTTQSQPSN
FTIYYIYFIYDIVSLFEQTSNECASGFGNITLLRHHLRSEIGKCKKTSYQSCSTKANIVPSKNVSLIVTG
QIPDLLDGALYGLAIFMPKPQFVTLPSPFDRAAILIRAKPFFLTKICKYRNLFSVPLECINLHANSISPD
DISEFNEADFHTNKLPYGAVYVTKETNKILFIPIEVCSRLKTCTQCIMYGLNSGCIWFTSICVHDNQP
KNKVTLTVDHCFKIMNISPLILNSSSPTILTIELDKPLIMASQEQLVIQAGDNHCTDIAMNGQFINCSM
RLTKSGEFNIDVSLRNDRYADTSIISAVSSDKVHIFASDSDYTLIIISVLFSCLIINSFAFIVYFRKCNKK
HLNRSKKVSRPRKVKQFVGTLSDKKFIKFFEPKKQTDLSAITPVKAQIVSSTMATLDDSRIINETSSE
QASLWITMRSVPRQIFPRRKLLQSKPKQRPNDFSQLD*
```

Table 13: TE statistics in the flanking region of 5 kb of key gene families.

Flanking region 5 kb	*Gene Family*	*Gypsy*	*Copia*	*L1*	*CR1*	*Mariner*	*PiggyBac*	*Helitron*	*Average TE Per Gene*
T. evansi	Chemo	1.11	0.71	0.13	0.04	1.88	0.2	0.08	4.16
	P450	0.44	0.1	0.44	0.02	1.68	0.07	0	2.76
	NFB	0.35	0.12	0.94	0	1.53	0.24	0	3.18
	ABC	0.42	0.25	0.17	0.02	0.79	0.05	0	1.7
	dUTPase	2.33	0	0.17	0	2.33	0	0.17	5
	IAP	2.14	0.43	0	0	1	0	0.21	3.79
	IDRCD	0.83	0	0.33	0	1.33	0.08	0	2.58
	SDR	*0.13*	0.07	0.17	0.04	0.54	0	0.02	0.98
	HCSP	5.78	0.44	0.67	0	1	0	0	7.89
T. lintearius	Chemo	4.97	2.69	1.85	0.14	4.24	0.74	0.07	14.69
	P450	1.31	0.34	1.37	0.14	2.8	0	0	5.97
	NFB	1.41	0.29	1.41	0.06	2.24	0.18	0	5.59
	ABC	0.88	0.17	1.46	0.08	1.68	0.13	0.09	4.48
	dUTPase	1.18	0.05	0.27	0.08	1.42	0.05	0.03	3.08
	IAP	*2.49*	0.75	0.13	0.11	2	0.13	0.1	5.7
	IDRCD	0.7	0	1.2	0	1	0.1	0	3
	SDR	1.26	0.2	0.68	0.16	1.46	0.14	0	3.9
	HCSP	4.56	0.44	1.94	0.12	1.25	0.06	0.25	8.62
T. urticae	Chemo	0.93	1.48	0.91	0.04	1.57	0.75	0	5.68
	P450	0.73	0.22	1.82	0.09	2.07	0.12	0.02	5.07
	NFB	3.51	0.3	1.36	0.16	1.49	0.25	0	7.06
	ABC	0.64	0.18	2.22	0.01	1.12	0.11	0.01	4.29
	dUTPase	0.39	0.09	0.39	0	0.33	0.03	0.03	1.27
	IAP	0.96	0.42	0.38	0	0.62	0.12	0.04	2.54
	IDRCD	0.38	0.5	1.62	0	0.56	0.19	0	3.25
	SDR	0.56	0.2	0.96	0.05	0.6	0	0	2.38
	HCSP	6.95	0.6	1.25	0.01	1.18	0.1	0.02	10.12

Table 14: TE statistics in the flanking region of 15 kb of key gene families.

Flanking region 15 kb	*Gene Family*	*Gypsy*	*Copia*	*L1*	*CR1*	*Mariner*	*PiggyBac*	*Helitron*	*Average TE Per Gene*
T. evansi	Chemo	1.98	1.18	0.2	0.11	3.94	0.38	0.1	7.9
	P450	1.41	0.22	0.54	0.07	2.68	0.12	0.05	5.1
	NFB	1.24	0.29	1.06	0.06	3.29	0.35	0	6.29
	ABC	1.06	0.37	0.47	0.08	2.03	0.15	0.05	4.21
	dUTPase	5	0	0.17	0.17	4.67	0	0.17	10.17
	IAP	5.71	1	0.07	0	3.29	0	0.21	10.29
	IDRCD	2.08	0	0.5	0	2.83	0.17	0.08	5.67
	SDR	0.76	0.17	0.54	0.07	1.76	0.11	0.07	3.48
	HCSP	8.78	0.67	0.67	0	1.78	0	0	11.89
T. lintearius	Chemo	7.24	3.51	2.83	0.38	6.19	0.94	0.17	21.26
	P450	2.09	0.51	2.6	0.43	5.2	0.06	0.06	10.94
	NFB	4.88	0.47	2.82	0.47	4.18	0.18	0	13
	ABC	2.38	0.5	3.43	0.6	3.23	0.31	0.2	10.65
	dUTPase	4.32	0.6	0.72	0.28	4.25	0.27	0.13	10.57
	IAP	6.56	1.48	0.89	0.13	6.05	0.38	0.16	15.63
	IDRCD	2.3	0.3	1.6	2.1	4.7	0.2	0.1	11.3
	SDR	3.1	0.58	2.22	0.92	3.96	0.24	0.04	11.06
	HCSP	8.25	1.44	2.62	0.38	3	0.5	0.25	16.44
T. urticae	Chemo	2.63	3.29	2.11	0.13	4.08	1.44	0.03	13.7
	P450	1.58	0.52	3.97	0.31	4.67	0.34	0.04	11.44
	NFB	7.6	0.78	3.46	0.34	3.59	0.61	0	16.38
	ABC	1.57	0.44	4.52	0.02	2.36	0.32	0.05	9.29
	dUTPase	1.79	0.24	1.39	0.06	1.18	0.39	0.03	5.09
	IAP	1.88	1.19	1.31	0.08	4.19	0.35	0.04	9.04
	IDRCD	1.25	0.62	2.81	0.06	2.62	0.38	0.06	7.81
	SDR	1.52	0.39	2.59	0.11	1.54	0.06	0.01	6.22
	HCSP	15.45	1.28	2.77	0.01	2.91	0.22	0.05	22.69

Table 15: TE statistics in the flanking region of 20 kb of key gene families.

Flanking region 20 kb	*Gene Family*	*Gypsy*	*Copia*	*L1*	*CR1*	*Mariner*	*PiggyBac*	*Helitron*	*Average TE Per Gene*
T. evansi	Chemo	2.47	1.27	0.22	0.13	4.53	0.45	0.11	9.19
	P450	1.66	0.27	0.63	0.07	3.1	0.2	0.05	5.98
	NFB	1.82	0.47	1.41	0.24	3.71	0.41	0	8.06
	ABC	1.33	0.48	0.56	0.11	2.38	0.21	0.09	5.16
	dUTPase	5.33	0	0.17	0.17	5	0	0.17	10.83
	IAP	6.29	1.14	0.07	0.07	4.29	0.07	0.21	12.14
	IDRCD	2.5	0.17	0.5	0	3.17	0.17	0.08	6.58
	SDR	0.96	0.28	0.59	0.17	2.13	0.17	0.11	4.41
	HCSP	8.78	0.67	0.67	0	2.44	0	0	12.56
T. lintearius	Chemo	8.11	3.9	2.97	0.55	7.08	1.09	0.2	23.9
	P450	2.6	0.69	2.69	0.46	6.51	0.29	0.06	13.29
	NFB	6.65	0.59	4	0.53	5	0.29	0	17.06
	ABC	2.93	0.57	4.01	0.7	4.17	0.38	0.2	12.95
	dUTPase	5.37	0.88	1.22	0.37	5.67	0.38	0.17	14.05
	IAP	8.1	1.71	1.3	0.25	7.87	0.59	0.22	20.05
	IDRCD	3.7	0.3	1.8	2.1	5.4	0.3	0.1	13.7
	SDR	3.56	0.78	2.44	1.08	5	0.32	0.06	13.24
	HCSP	10.12	1.5	2.75	0.5	4.38	0.88	0.25	20.38
T. urticae	Chemo	3.36	4.09	2.52	0.19	5.14	1.7	0.05	17.06
	P450	2.09	0.67	4.74	0.38	5.69	0.37	0.04	13.98
	NFB	9.44	0.99	4.29	0.4	4.27	0.74	0	20.11
	ABC	1.98	0.55	5.61	0.04	3.09	0.37	0.08	11.71
	dUTPase	2.7	0.42	1.67	0.06	2.61	0.48	0.12	8.06
	IAP	2.54	1.85	1.73	0.15	5.73	0.5	0.04	12.54
	IDRCD	1.88	0.62	3.06	0.12	3.06	0.44	0.06	9.25
	SDR	1.77	0.42	3.02	0.11	2.26	0.08	0.02	7.67
	HCSP	19	1.57	3.45	0.01	3.57	0.31	0.07	27.98

4.6 Reference

Aparicio, G., Gotz, S., Conesa, A., Segrelles, D., Blanquer, I., Garcia, J.M., Hernandez, V., Robles, M., and Talon, M. (2006). Blast2GO goes grid: developing a grid-enabled prototype for functional genomics analysis. Studies in health technology and informatics *120*, 194-204.

Bevers, E.M., and Williamson, P.L. (2010). Phospholipid scramblase: an update. FEBS letters *584*, 2724-2730.

Bevers, E.M., and Williamson, P.L. (2016). Getting to the Outer Leaflet: Physiology of Phosphatidylserine Exposure at the Plasma Membrane. Physiol Rev *96*, 605-645.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578-579.

Bolland, H.R., Gutierrez, J., and Flechtmann, C.H.W. (1997). World Catalogue of the Spider Mite Family. 1-3.

Boubou, A., Migeon, A., Roderick, G.K., and Navajas, M. (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.

Brown, P.J., Gill, A.C., Nugent, P.G., McVey, J.H., and Tomley, F.M. (2001). Domains of invasion organelle proteins from apicomplexan parasites are homologous with the Apple domains of blood coagulation factor XI and plasma pre-kallikrein and are members of the PAN module superfamily. FEBS letters *497*, 31-38.

Carruthers, V.B., and Tomley, F.M. (2008). Microneme proteins in apicomplexans. Subcell Biochem *47*, 33-45.

Cazaux, M., Navarro, M., Bruinsma, K.A., Zhurov, V., Negrave, T., Van Leeuwen, T., Grbic, V., and Grbic, M. (2014). Application of two-spotted spider mite Tetranychus urticae for plant-pest interaction studies. Journal of visualized experiments : JoVE.

Chen, F., Mackey, A.J., Stoeckert, C.J., Jr., and Roos, D.S. (2006a). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic acids research *34*, D363-368.

Clotuche, G., Turlure, C., Mailleux, A.C., Detrain, C., and Hance, T. (2013). Should I lay or should I wait? Egg-laying in the two-spotted spider mite Tetranychus urticae Koch. Behavioural processes *92*, 24-30.

Cock, P.J.A. (2013). Galaxy wrapper for the CLC Assembly Cell suite from CLCbio (http://toolshed.g2.bx.psu.edu/view/peterjc/clc_assembly_cell).

Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nature reviews Genetics *10*, 691-703.

Danielson, P.B. (2002). The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. Curr Drug Metab *3*, 561-597.

Delcher, A.L., Phillippy, A., Carlton, J., and Salzberg, S.L. (2002). Fast algorithms for large-scale genome alignment and comparison. Nucleic acids research *30*, 2478-2483.

Dermauw, W., Wybouw, N., Rombauts, S., Menten, B., Vontas, J., Grbic, M., Clark, R.M., Feyereisen, R., and Van Leeuwen, T. (2013b). A link between host plant adaptation and pesticide resistance in the polyphagous spider mite Tetranychus urticae. Proceedings of the National Academy of Sciences of the United States of America *110*, E113-122.

Dhanoa, B.S., Cogliati, T., Satish, A.G., Bruford, E.A., and Friedman, J.S. (2013). Update on the Kelch-like (KLHL) gene family. Hum Genomics *7*, 13.

Dogan, T., MacDougall, A., Saidi, R., Poggioli, D., Bateman, A., O'Donovan, C., and Martin, M.J. (2016). UniProt-DAAC: domain architecture alignment and classification, a new method for automatic functional annotation in UniProtKB. Bioinformatics *32*, 2264-2271.

Donald M, T., and Edward W, B. (1968). Spider Mites of Southwestern United States and a Revision of the Family Tetranychidae (Tucson, University of Arizona Press).

Dorsett, D., Viglianti, G.A., Rutledge, B.J., and Meselson, M. (1989). Alteration of hsp82 gene expression by the gypsy transposon and suppressor genes in Drosophila melanogaster. Genes Dev *3*, 454-468.

Dubinin, V. (1962). Class Acaromorpha: mites or gnathosomic chelicerate arthropods. Fundamentals of Palaeontology, 447-473.

Fellous, S., Angot, G., Orsucci, M., Migeon, A., Auger, P., Olivieri, I., and Navajas, M. (2014). Combining experimental evolution and field population assays to study the evolution of host range breadth. J Evol Biol *27*, 911-919.

Fischer, S., Brunk, B.P., Chen, F., Gao, X., Harb, O.S., Iodice, J.B., Shanmugam, D., Roos, D.S., and Stoeckert, C.J., Jr. (2011). Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] *Chapter 6*, Unit 6 12 11-19.

Flavell, A.J., Alphey, L.S., Ross, S.J., and Leigh-Brown, A.J. (1990). Complete reversions of a gypsy retrotransposon-induced cut locus mutation in Drosophila melanogaster involving jockey transposon insertions and flanking gypsy sequence deletions. Mol Gen Genet *220*, 181-185.

Flechtmann, C.H., and Noronha, A.C. (2013). A new species of the genus Tenuipalpus (Prostigmata: Tenuipalpidae) with remarks on a conceivable ovipositor in flat mites. Zootaxa *3681*, 493-499.

Foissac, S., Gouzy, J., Rombauts, S., Mathe, C., Amselem, J., Sterck, L., Van de Peer, Y., Rouze, P., and Schiex, T. (2008). Genome annotation in plants and fungi: EuGene as a model platform. Curr Bioinform *3*, 87-97.

Gotoh, T., Sugimoto, N., Pallini, A., Knapp, M., Hernandez-Suarez, E., Ferragut, F., Ho, C.C., Migeon, A., Navajas, M., and Nachman, G. (2010). Reproductive performance of seven strains of the tomato red spider mite Tetranychus evansi (Acari: Tetranychidae) at five temperatures. Experimental and Applied Acarology *52*, 239-259.

Gotz, S., Garcia-Gomez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talon, M., Dopazo, J., and Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research *36*, 3420-3435.

Grbic, M., Van Leeuwen, T., Clark, R.M., Rombauts, S., Rouze, P., Grbic, V., Osborne, E.J., Dermauw, W., Ngoc, P.C., Ortego, F.*, et al.* (2011a). The genome of Tetranychus urticae reveals herbivorous pest adaptations. Nature *479*, 487-492.

Grbic, M., Van Leeuwen, T., Clark, R.M., Rombauts, S., Rouze, P., Grbic, V., Osborne, E.J., Dermauw, W., Phuong, C.T.N., Ortego, F.*, et al.* (2011b). The genome of Tetranychus urticae reveals herbivorous pest adaptations. Nature *479*, 487-492.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S.R. (2003). Rfam: an RNA family database. Nucleic acids research *31*, 439-441.

Herwald, H., Renne, T., Meijers, J.C., Chung, D.W., Page, J.D., Colman, R.W., and Muller-Esterl, W. (1996). Mapping of the discontinuous kininogen binding site of prekallikrein. A distal binding segment is located in the heavy chain domain A4. The Journal of biological chemistry *271*, 13061-13067.

Hirst, S. (1923). XLVI.—On some Arachnid remains from the Old Red Sandstone (Rhynie Chert Bed, Aberdeenshire). Journal of Natural History *12*, 455-474.

Ho, D.H., Badellino, K., Baglia, F.A., and Walsh, P.N. (1998). A binding site for heparin in the apple 3 domain of factor XI. The Journal of biological chemistry *273*, 16382-16390.

Ho, M.S., Tsai, P.I., and Chien, C.T. (2006). F-box proteins: the key to protein degradation. J Biomed Sci *13*, 181-191.

Jame, D.F. (1990). Evolutionary adaptation to host plants in a laboratoty population of the phytophagous mite Tetranychus urticae Koch. Oecologia *83*, 568.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., *et al.* (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236-1240.

Jones, P.M., and George, A.M. (2004). The ABC transporter structure and mechanism: perspectives on recent research. Cell Mol Life Sci *61*, 682-699.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res *110*, 462-467.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nature methods *12*, 357-360.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology *305*, 567-580.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome research *19*, 1639-1645.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009a). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research *13*, 2178-2189.

Lim, E.G., Roh, H.S., Coudron, T.A., and Park, C.G. (2011). Temperature-dependent fumigant activity of essential oils against twospotted spider mite (Acari: Tetranychidae). Journal of economic entomology *104*, 414-419.

Magalhaes, S., Fayard, J., Janssen, A., Carbonell, D., and Olivieri, I. (2007). Adaptation in a spider mite population after long-term evolution on a single host plant. J Evol Biol *20*, 2016-2027.

Marlor, R.L., Parkhurst, S.M., and Corces, V.G. (1986). The Drosophila melanogaster gypsy transposable element encodes putative gene products homologous to retroviral proteins. Mol Cell Biol *6*, 1129-1134.

McCulloch, R. (1947). The adaptation of military scrub typhus mite control to civilian needs. Med J Aust *1*, 449-452.

Muller-Scharer, H., Schaffner, U., and Steinger, T. (2004). Evolution in invasive plants: implications for biological control. Trends in ecology & evolution *19*, 417-422.

Navajas, M., de Moraes, G.J., Auger, P., and Migeon, A. (2013a). Review of the invasion of Tetranychus evansi: biology, colonization pathways, potential expansion and prospects for biological control. Experimental and Applied Acarology *59*, 43-65.

Ngoc, P.C., Greenhalgh, R., Dermauw, W., Rombauts, S., Bajda, S., Zhurov, V., Grbic, M., Van de Peer, Y., Van Leeuwen, T., Rouze, P.*, et al.* (2016). Complex Evolutionary Dynamics of Massively Expanded Chemosensory Receptor Families in an Extreme Generalist Chelicerate Herbivore. Genome biology and evolution *8*, 3323-3339.

Onyambus, G.K., Maranga, R.O., Gitonga, L.M., and Knapp, M. (2011). Host plant resistance among tomato accessions to the spider mite Tetranychus evansi in Kenya. Experimental and Applied Acarology *54*, 385-393.

Peifer, M., and Bender, W. (1988). Sequences of the gypsy transposon of Drosophila necessary for its effects on adjacent genes. Proceedings of the National Academy of Sciences of the United States of America *85*, 9650-9654.

Phuong, C.T.N. (2014). Genome annotation and evolution of chemises receptors in spider mites. In VIB Department of Plant Systems Biology (Gent University).

Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van de Peer, Y., and Vandepoele, K. (2012). i-ADHoRe 3.0--fast and sensitive detection of genomic homology in extremely large data sets. Nucleic acids research *40*, e11.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Qureshi, A.H., Oatman, E.R., and Fleschne.Ca (1969). Biology of Spider Mite, Tetranychus Evansi. Ann Entomol Soc Am *62*, 898-&.

Sabelis, W.H.a.M.W. (1987). Spider mites: Their biology, natural enemies and control (World crop pests vols 1A and 1B).

Siggs, O.M., and Beutler, B. (2012). The BTB-ZF transcription factors. Cell Cycle *11*, 3358-3369.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210-3212.

Smit, A., and Hubley, R. (2008-2015). RepeatModeler Open-1.0.

Sterck, L., Billiau, K., Abeel, T., Rouze, P., and Van de Peer, Y. (2012). ORCAE: online resource for community annotation of eukaryotes. Nature methods *9*, 1041.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] *Chapter 4*, Unit 4 10.

Tordai, H., Banyai, L., and Patthy, L. (1999). The PAN module: the N-terminal domains of plasminogen and hepatocyte growth factor are homologous with the apple domains of the prekallikrein family and with a novel domain found in numerous nematode proteins. FEBS letters *461*, 63-67.

Tsagkarakou, A., Cros-Arteil, S., and Navajas, M. (2007). First record of the invasive mite Tetranychus evansi in Greece. Phytoparasitica *35*, 519-522.

Tsubota, T., and Shiotsuki, T. (2010). Genomic analysis of carboxyl/cholinesterase genes in the silkworm Bombyx mori. BMC genomics *11*, 377.

UniProt, C. (2015). UniProt: a hub for protein information. Nucleic acids research *43*, D204-212.

Van Leeuwen, T., Dermauw, W., Grbic, M., Tirry, L., and Feyereisen, R. (2013). Spider mite control and resistance management: does a genome help? Pest management science *69*, 156-159.

Van Leeuwen, T., Vontas, J., Tsagkarakou, A., Dermauw, W., and Tirry, L. (2010). Acaricide resistance mechanisms in the two-spotted spider mite Tetranychus urticae and other important Acari: a review. Insect biochemistry and molecular biology *40*, 563-572.

Vassylyev, D.G., and Morikawa, K. (1996). Precluding uracil from DNA. Structure *4*, 1381-1385.

Vertessy, B.G., and Toth, J. (2009). Keeping uracil out of DNA: physiological role, structure and catalytic mechanism of dUTPases. Acc Chem Res *42*, 97-106.

Walter, M.E., and Proctor, H.C. (1999). Mites: Ecology, Evolution and Behaviour (CABI Publishing).

Yu, K., Whitlock, J.M., Lee, K., Ortlund, E.A., Cui, Y.Y., and Hartzell, H.C. (2015). Identification of a lipid scrambling domain in ANO6/TMEM16F. eLife *4*, e06901.

Zhurov, V., Navarro, M., Bruinsma, K.A., Arbona, V., Santamaria, M.E., Cazaux, M., Wybouw, N., Osborne, E.J., Ens, C., Rioja, C.*, et al.* (2014b). Reciprocal Responses in the Interaction between Arabidopsis and the Cell-Content-Feeding Chelicerate Herbivore Spider Mite. Plant Physiol *164*, 384-399.

Chapter 5

5 Evolutionary dynamics of a massively expanded novel F-box gene family in polyphagous pest *Tetranychus urticae*

F-box proteins are known in animals for playing various functions, from the immune response, cell cycle, signaling cascades to developmental programs. They combine with SKP1 and Cullin1 to form SCF complex that further mediates protein ubiquitination and degradation. Here we present a class of novel F-box (NFB) genes extremely expanded in the polyphagous herbivore *T. urticae*. To the best of our knowledge, this NFB gene family has never been reported and no homologs can be found in public database. This NFB gene family significantly proliferated in *T. urticae* (234 copies including 188 intact genes and 96% have transcripts' support), compared with these in *T. evansi* and *T. lintearius* (38 and 36 copies, respectively). Meanwhile, 12 (5%) of NFB genes are pseudogenes in *T. urticae*, in contrast to 188 (77.6%) intact genes. The NFB genes evolved as tandem duplication events in big clusters. It is also observed that these NFB clusters are highly dispersed by transposable element (TE), which suggests transposable elements would play an important role in shuffling and expanding this gene family. Transcriptome profiling and network analyses show NFB genes also have a strong correlation with the SKP1 gene, suggesting NFB genes have similar functions as conventional F-box genes, but their binding proteins are unknown yet.

5.1 Introduction

F-box proteins belong to a large gene family that regulates the cell cycle, signaling cascades, and developmental programs by targeting proteins for ubiquitination. This process is operated by F-box-SKP1-Cullin1 (SCF complex) that mediates ubiquitination of proteins for degradation. An F-box protein, by definition, contains an F-box domain, a protein structural motif around 50 amino acids in size. F-box proteins are subdivided into three major classes according to the presence of additional domains (Ho *et al.*, 2006). The first class, WD40, also known as WD or beta-transducing repeat, contains a short motif of 40 amino acids. Approximately 4 to 16 tandem copies form a circularized beta-propeller that plays a variety of functions including signaling, regulating cell cycle, autophagy, and apoptosis. The second class, Leucine-Rich Repeat (LRR), composed of nearly 30 amino acids, forms a beta strand and alpha helix structure. Many such repeats constitute of a horseshoe shape and they are frequently involved in protein-protein interactions (PPI) (Rothberg *et al.*, 1990; Gay *et al.*, 1991; Kobe and Kajava, 2001). The third class contains miscellaneous domains or motifs and the functions of most of these proteins have not yet been identified (Kipreos and Pagano, 2000).

Many studies have shown that F-box proteins generally tend to evolve through massive waves of duplication either in both plants and animals (Xu *et al.*, 2009; Navarro-Quezada *et al.*, 2013; Wang *et al.*, 2014; Zhao *et al.*, 2015). For example, there are 692, 337, and 779 F-box genes in *Arabidopsis*, poplar, and rice, forming one of the largest multi-gene superfamilies in plants. These plant F-box genes can be further classified into 42 minor families and they have experienced dramatically different modes of sequence divergence, apparently resulting in adaptive changes in function (Xu *et al.*, 2009). F-box genes are relatively less frequent in animals than that in plants. There are 11 F-box proteins in budding yeast, 326 predicted in *Caenorhabditis elegans*, a minimum 20 in *Drosophila*, and at least 38 in humans (Kipreos and Pagano, 2000). Only recently, a large class of F-box genes with LRR and signal peptide (SP) was identified as an extreme expansion in the wheat pest Hessian fly (Zhao *et al.*, 2015). These F-box proteins are supposed to enable Hessian flies to hijack the plant proteasome to directly produce nutritive tissue and additionally to defeat basal plant immunity. They were first identified as hundreds of related transcripts in the insect's salivary gland and were termed as secreted salivary gland proteins (SSGPs). At that point, neither transcripts nor associated genes appeared to have sequence similarities to other genes. This class of F-box genes has a total number of 426 copies, which is one-eighth of the genes that encode putative gall effectors. Interestingly, these genes have an SP at the beginning of the sequence, followed by an F-box domain and 13 LRRs, suggesting these proteins can be exported from the cells.

In this study, a novel F-box (NFB) protein family was found in spider mite *T. urticae*, a polyphagous herbivore that feeds on over 1,100 plants, most of which are important agricultural crops such tomatoes, potatoes, berries, corn and citruses (Grbic *et al.*, 2011a; Cazaux *et al.*, 2014). This type of NFB proteins has not been described in any previous studies. These NFB genes have an extreme expansion in the genome of polyphagous pest *T. urticae* in contrast to oligophagous *Tetranychus evansi* and monophagous *Tetranychus lintearius*. Because of F-box proteins' roles in protein ubiquitination and degradation, it is important to highlight this novel gene family and furthermore, to investigate how and why the polyphagous pest *T. urticae* possesses so many NFB genes

5.2 Results and Discussion

5.2.1 Conventional F-box genes are conserved

In contrast to NFB, we initially investigated the conventional F-box (CFB) genes, which have been widely studied in other organisms like *Drosophila*. These CFB genes in three *Tetranychus* spider mites and *Drosophila melanogaster* were analyzed. The results show that the CFB genes are relatively conserved across these four species and no obvious expansion occurred from chelicerates to insects (Figure 26b), even though some studies show CFB genes in mammals (WD40 and LRR) have slightly increased, compared with these in arthropods (Wang *et al.*, 2014).

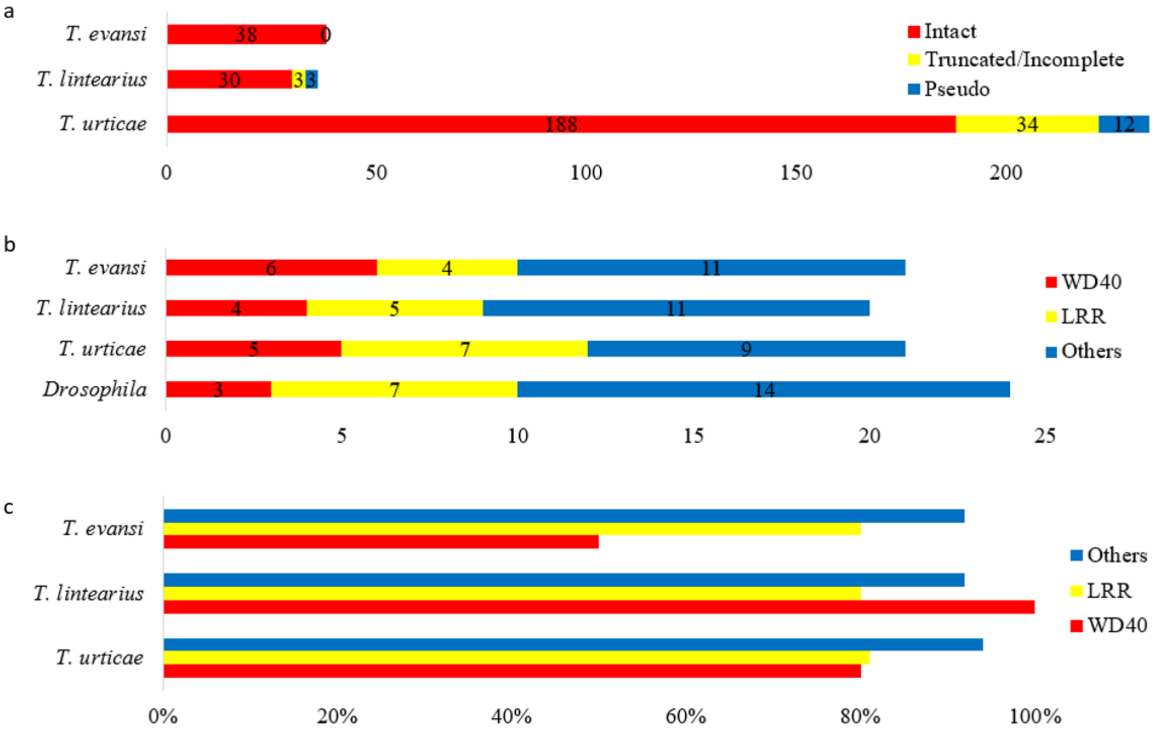


Figure 26: An overview of all F-box genes.
a: NFB copy number in three genomes; b: CFB copy number in four species; c: expression profiles of CFB genes in three *Tetranychus* spider mites.

5.2.2 The novelty and assessment of NFB genes in *Tetranychus*

We observed 234 NFB genes in *T. urticae* including 188 intact genes, 34 truncated/incomplete genes and 12 pseudogenes. In contrast, *T. evansi* and *T. lintearius* only contain 32 and 33 NFB genes, respectively, as shown in Figure 26a.

This NFB gene family has a conserved structure with primarily two subfamilies: subfamily A (a small cluster of 12 genes: 4 exons and about 320 aa in length) and subfamily B (180+ genes of 6 exons and about 360 aa in length). The F-box domains of CFB genes are not necessarily located at N-terminus of the sequence. However, the F-box domains of NFB genes always locate at the N-terminus, and the other part of the sequence contains LRR repeats. No differences are observed in LRR between CFB and NFB genes. Unlike SSGPs, no signal peptide was detected from NFB protein sequences using SignalP 4.0 (Petersen *et al.*, 2011; Zhao *et al.*, 2015), which indicates that these NFB proteins probably cannot be secreted and transported outside of endoplasmic reticulum membranes, instead, they would play unknown functions. Compared with CFB genes (LRR class), both F-box and LRR domains in NFB genes are more conserved and F-box domains typically locate at first 50 amino acids of the N-terminus.

Table 16: Evidence for confirming NFB genes.

(a) BLASTN hits of NFB genes in six other *Tetranychus* (transcriptomic data).

Six other *Tetranychus* strain	Returned hits number
Tetranychus malaysiensis	1
Tetranychus truncates	2
Tetranychus kanzawai	2
Tetranychus pueraricola	3
Tetranychus ludeni	2
Tetranychus phaselus	5

(b) Reassessment of NFB genes in pseudo scaffolds by TBLASTN.*

Tetur_10k_split	206	tetli_10k_split	65	tetev_10k_split	44	tetli_ont_10k_split	56
tetur_5k_split	271	tetli_5k_split	73	tetev_5k_split	54	tetli_ont_5k_split	62
tetur_2k_split	392	tetli_2k_split	91	tetev_2k_split	68	tetli_ont_2k_split	81

* tetur-*T. urticae*; tetli-*T. lintearius*; tetev-*T. evansi*; ont-Oxford Nanopore Technology; Table (b) credit: Dr. Vladmir Zhurov.

We applied several approaches to validate the novelty of these NFB genes. First, no homologs (except genes from only *Tetranychus*) were found in the NCBI protein database using NFB proteins as baits. Second, we also used BLASTN and TBLASTN to search in flat mite *Brevipalpus yothersi* genome (unpublished data, Chapter 6) and found only one sequence fragment (E-value < 1e-5), which only possessed on LRR domain. No F-box domain of NFB was found in flat mite *Brevipalpus yothersi* genome, suggesting probably no NFB genes are present in flat mite. Third, we used NFB sequences and BLASTN (e-value < 1e-5) to search six *Tetranychus* (unpublished data: *Tetranychus malaysiensis, Tetranychus truncates, Tetranychus kanzawai, Tetranychus pueraricola, Tetranychus ludeni, Tetranychus piercei* and *Tetranychus phaselus*) transcriptome assembly data. Table 16a shows that other *Tetranychus* strains generally have a few NFB hits in the fragment, suggestive of no large expansion in other *Tetranychus* species.

Additionally, pseudo scaffolds of *T. urticae*, *T. evansi* and *T. lintearius* (both NGS assembly and TGS assembly) were also searched for NFB genes. Briefly, each genome FASTA files were concatenated and then split again into pseudo-scaffolds of 2, 5 or 10 kb (no overlaps, 1x coverage). TBLASTN was performed against these shredded genomes with *T. urticae* F-box domain sequences (E-value < 1e-6). Unique pseudo-scaffold ID's were extracted and their number used as an indicator of several potential loci. The same method was applied on *T. lintearius* (NGS assembly) and *T. lintearius* (TGS assembly). Oxford Nanopore Technology (ONT) assembly produced very similar results (Table 16b).

To confirm this expansion is not due to an artifact of assembly, we further investigated the genomic region of *T. urticae* assembly. The results show that the regions where novel F-box gene clusters are located have no artifact caused by repeated raw reads or overlapping assembly (Figure 27). Moreover, no identical sequences are found across this NFB gene family.

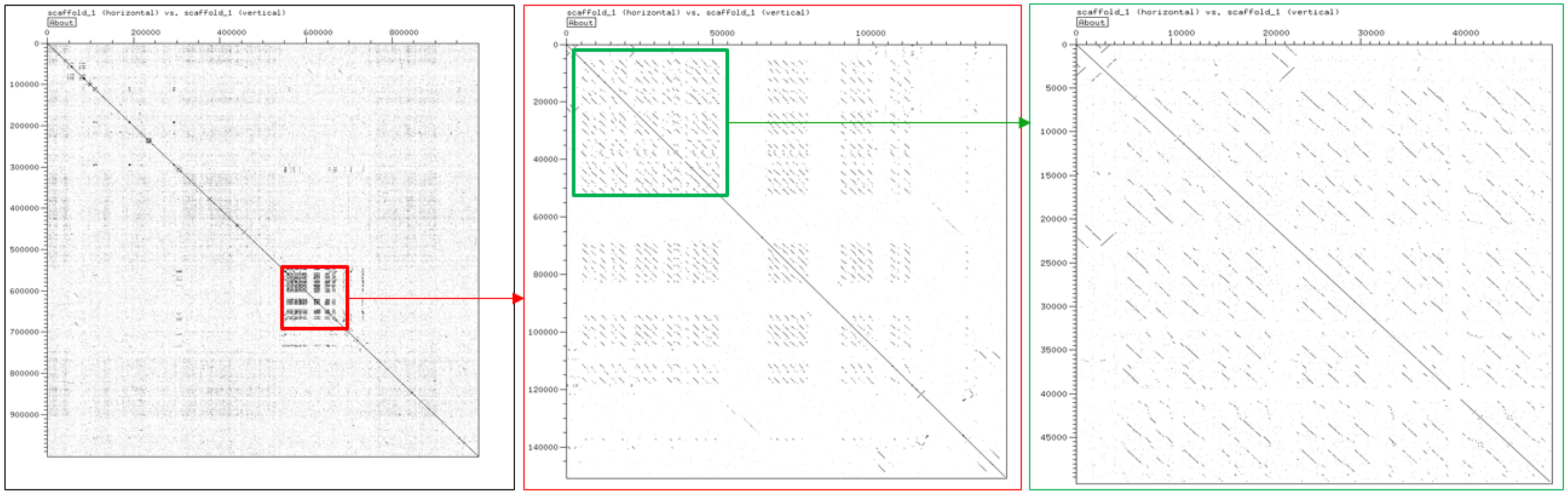


Figure 27: Genomic synteny of *T. urticae* scaffold 1.
Left: 3Mb-4Mb; middle: 3.55 Mb-3.7 Mb; right: 3.55 Mb-3.6 Mb. No straight continuous long diagonals are found and the short fragmental diagonals represent the similarity of NFB sequences.

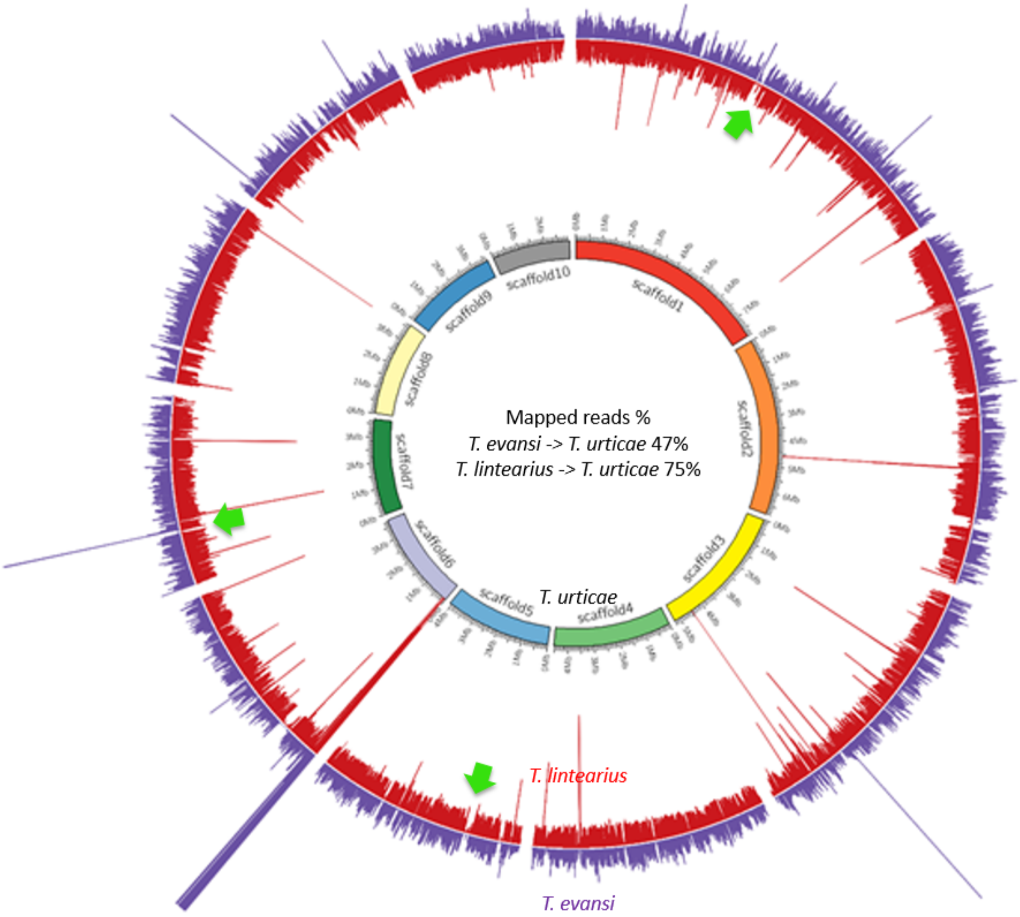


Figure 28: Genomic reads coverages of *T. lintearius* and *T. evansi* mapped to *T. urticae*.

Inner Circle: Top ten *T. urticae* scaffolds; red circle: the coverage of *T. lintearius* genomic reads mapping to *T. urticae*; purple circle: the coverage of *T. evansi* genomic reads mapping to *T. urticae*. Green Arrow: loci of expanded NFB clusters corresponding to *T. urticae.*

To further confirm NFB genes are not an assembly error, we also mapped the genomic reads of *T. lintearius* and *T. evansi* to the *T. urticae* genome. It is observed that the NFB proliferated loci (Figure 28, marked in the green arrow on scaffold 1, scaffold 5 and scaffold 7) are barely covered by genomics reads. This suggests that these NFB genes are a true expansion, otherwise, these loci will be covered by genomics reads from the other two genomes as well.

5.2.3 The extreme expansion of NFB genes in *T. urticae*

To understand the evolution of this NFB gene family across the three *Tetranychus* spider mites, we performed phylogenetic tree analysis using F-box domain sequences of the intact NFB genes (256 sequences) from *T. urticae*, *T. lintearius,* and *T. evansi*. Although the bootstrap values for many branches were low because of a large number of sequences and the small size of the F-box domain, the topology was generally reasonable because protein sequences with high similarities usually clustered together, as demonstrated in the phylogenetic relationship of F-box proteins in plants (Xu *et al.*, 2009).

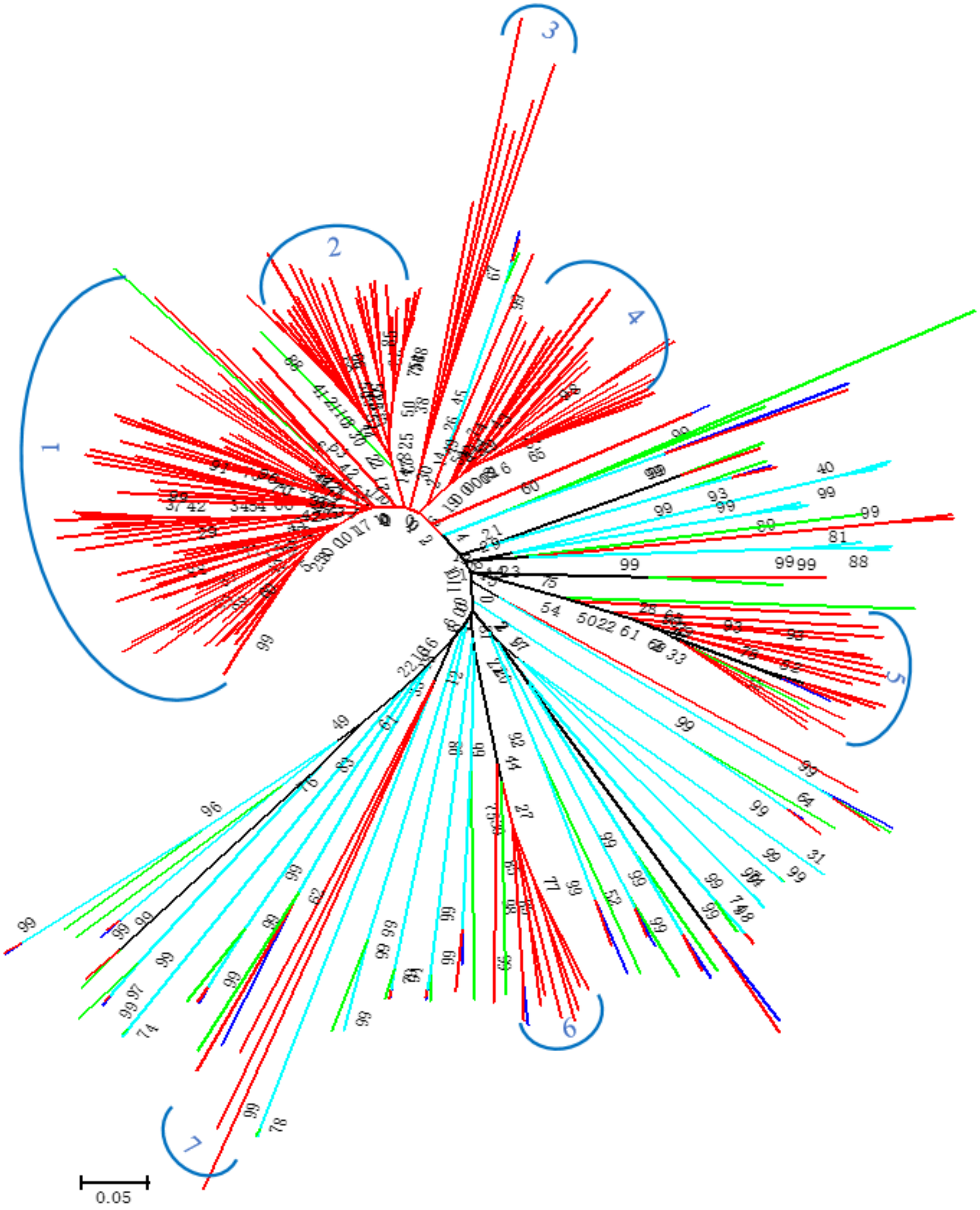


Figure 29: Phylogenetic tree of NFB genes across the three genomes.
Red: *T. urticae*; green: *T. evansi*; blue: *T. lintearius*; cyan: 3 mites; cluster numbers 1-7 start from the biggest cluster (from the top to the left clockwise).

Based on the phylogenetic relationships and domain organizations, we divided the F-box gene family into two subfamilies. Subfamily A has an exon number of four and its protein sequence about 320 amino acids in size while subfamily B has six exons and the protein length around 360 amino acids. Phylogenetic analysis provides the opportunity to identify evolutionarily conservative and divergent F-box genes (Figure 29). Basically, we grouped the NFB genes into seven clusters. Obviously, cluster 1, 2, 3 and 4 belong to a super-cluster and derived quite recently. However, cluster 5, 6 and 7 were derived from relatively ancient NFB genes.

5.2.4 Structure and Domain Organization of NFB

The structure and the domains of these NFB genes were further investigated using NFB cluster 5 as an example (Figure 30), which shows that NFB genes have more similar structure and domain distribution if they are close on the phylogeny

The F-box domain is located at the N-terminus of the protein sequences (about 50 amino acids), consist of the first exon of NFB 88 nt (i.e., 29 amino acids) and the second exon 60 nt (i.e., 20 amino acids). F-box domain alignment across these NFB genes clearly shows that these first two exons are quite conserved, compared with the rest region (C-terminus/LRR domain), as shown in Figure 30a.

We observed both N-terminus (F-box domain region) and C-terminus (the last 100 amino acids) are much conserved in NFB gene family (Figure 30b). The LRR region is relatively divergent, suggestive of LRR domains may bind different targeting proteins while both terminus may bind SKP1. To investigate the potential functions of this NFB protein family, we compared NFB protein sequences with the human SKP2 gene, which is also an F-box protein. We randomly selected a list of NFB genes and aligned them. Figure 30c shows the arrowheads-marked amino acids positions are potentially important sites for SKP1 and F-box (human SKP2) interactions (Zheng *et al.*, 2002).

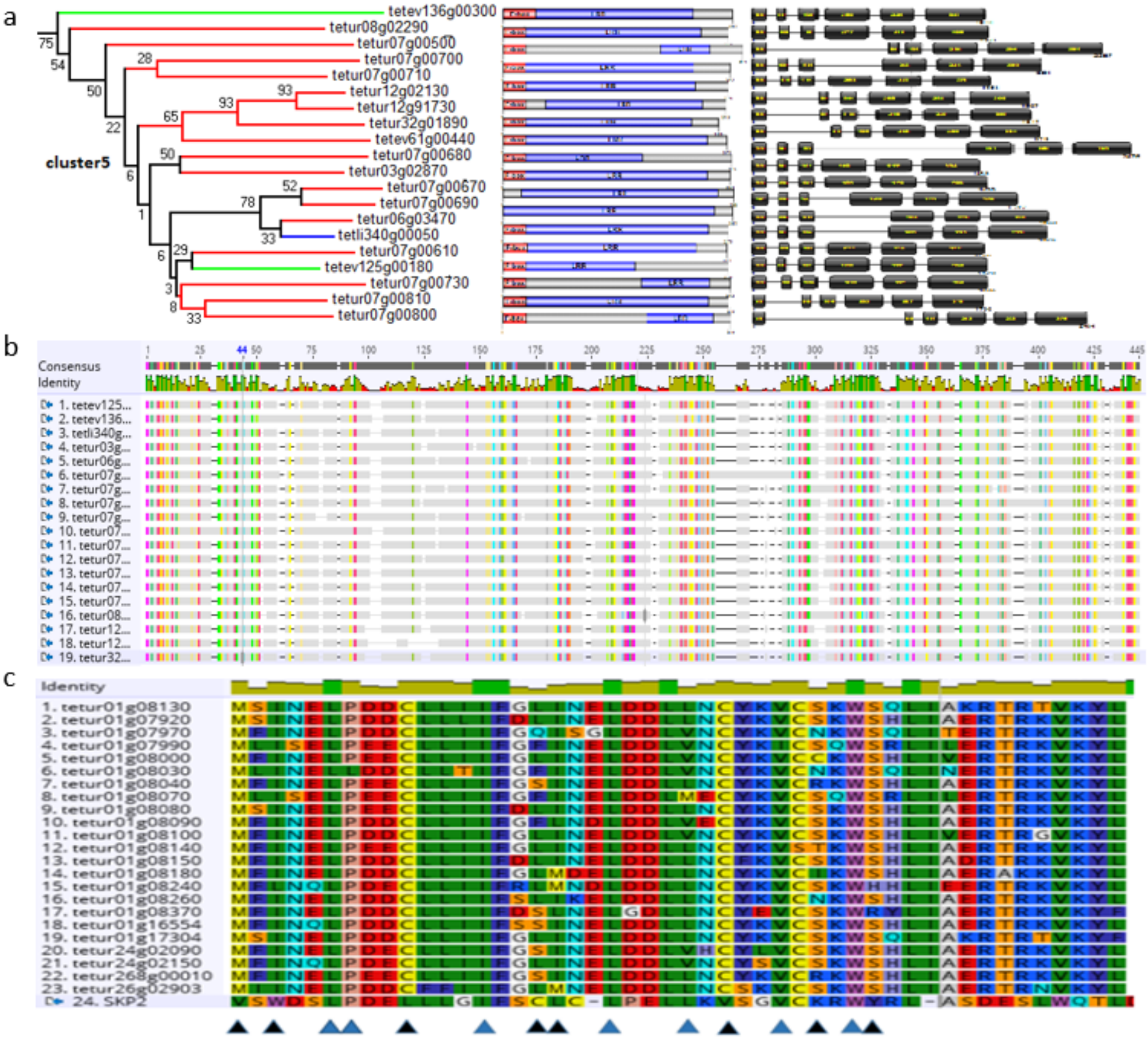


Figure 30: The conserved structure and domains of NFB genes in cluster 5.
a: phylogenetic relationships, domain organization, and exon-intron structure of NFB cluster 5; b: a global view of protein sequence alignment of NFB cluster 5 with a threshold of 90%; c: alignment shows NFB genes have some conserved spots with human SKP2.

5.2.5 NFB expanded by tandem duplication

Previous studies have suggested that F-box genes could be present as tandem arrays in the same chromosomal regions, suggestive of tandem duplication (Gagne *et al.*, 2002; Jain *et al.*, 2007; Xu *et al.*, 2009). To investigate the contribution of tandem duplication in terms of NFB expansion across the whole genome, we concentrated on the major clusters in scaffolds 1, 5 and 7, shown in Figure 31.

NFB genes located in the same clusters have a higher similarity in domain organization and exon-intron structure. NFB genes in the same cluster have a higher identity (Figure 32, identity >70% is shown in the yellow region; the colored bars indicate different scaffolds), suggesting they probably have proliferated through tandem duplication.

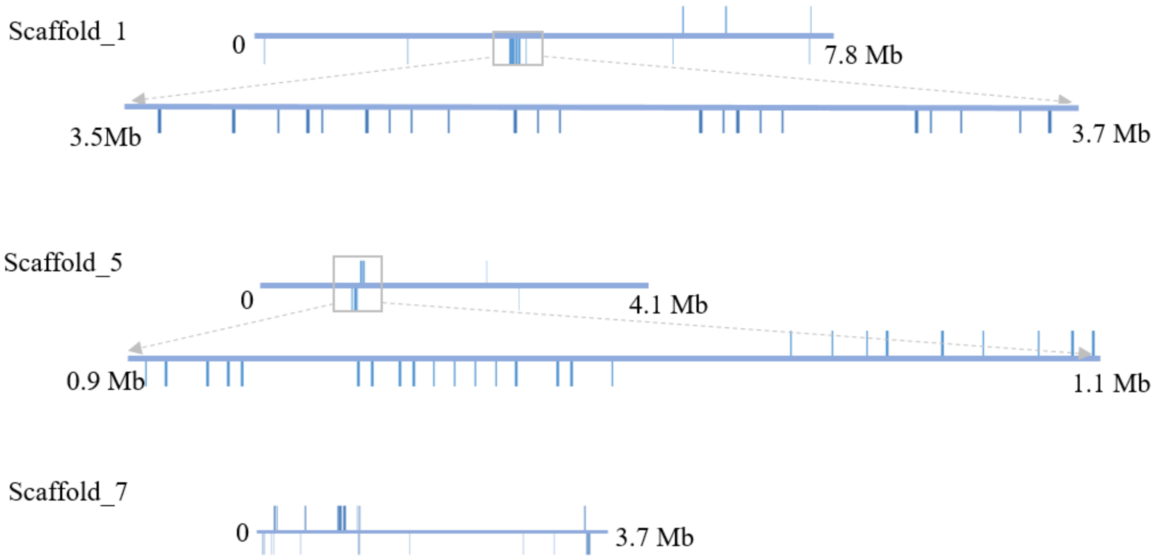


Figure 31: NFB clusters in *T. urticae* genome on the top three scaffolds.
The thick blue horizontal bars present scaffolds and the thin vertical blue bars mean NFB genes on scaffolds. Here we only show that the top three large expanded clusters of NFB genes in the genome of *T. urticae*.

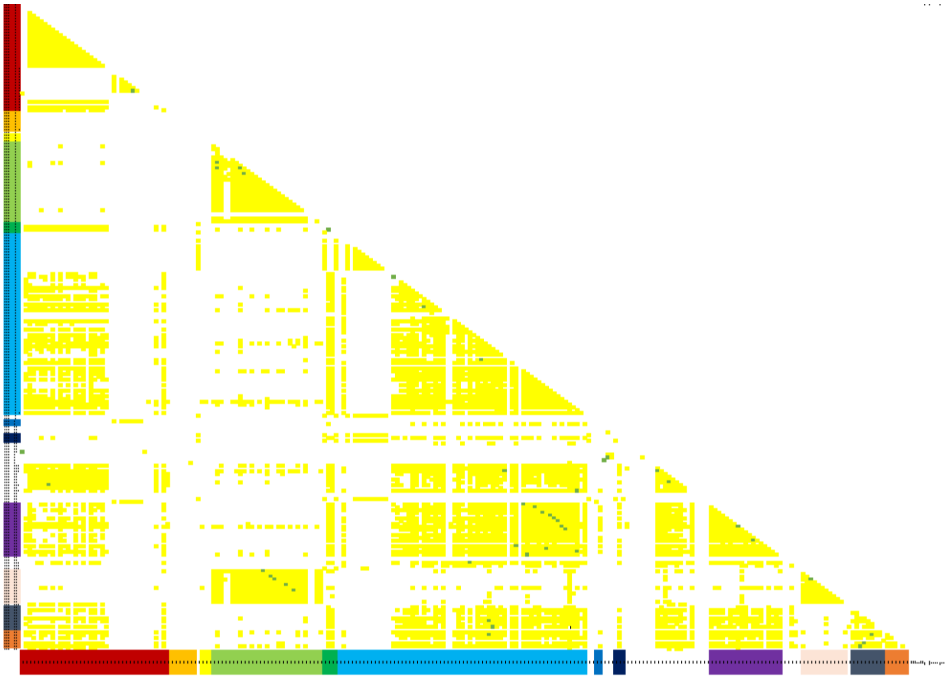


Figure 32: Estimates of evolutionary divergence between NFB sequences.
The presence of 'not applicable' spots (in the blank) in the results denote the cases in which it was not possible to estimate evolutionary distances. The NFB genes were sorted by their loci on each scaffold, which is in each distinctive color (red, orange, yellow, light green, green, blue, white, dark blue, white, black, white, purple, white, light grey, dark grey, deep orange and white, from the top to the bottom, from the left to the right, respectively). The yellow and green spots indicate the evolutionary distance is less than 0.05 and 0.02, respectively. Primary scaffolds of highly expanded NFB are marked as scaffolds 1, 5 and 7. Photo credit: Pengyu Jin.

The differences of amino acid per site from among sequences are shown in Figure 32. This analysis involved 234 amino acid sequences (all NFB genes from *T. urticae* genome). All ambiguous positions were removed for each sequence pair. There was a total of 1,082 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar *et al.*, 2016). In addition to tandem duplication, segmental duplications may also play a role in the expansion of the NFB gene family because of small clusters of NFB genes, or even, individual NFB genes scatter across the whole genome. The paralogous gene pairs in different clusters display clear similarities across different clusters.

5.2.6 Negative selection

The nonsynonymous (Ka) and synonymous (Ks) substitution rate (Ka/Ks) is of great significance in understanding evolutionary dynamics of protein-coding sequences across closely related and recently diverged species (Fay and Wu, 2003). The peak rate of NFB genes is around 0.7 (below 1), suggesting most them were evolving as experiencing negative/purifying selection. They maintain the long-term stability of biological structures by removing deleterious mutations.

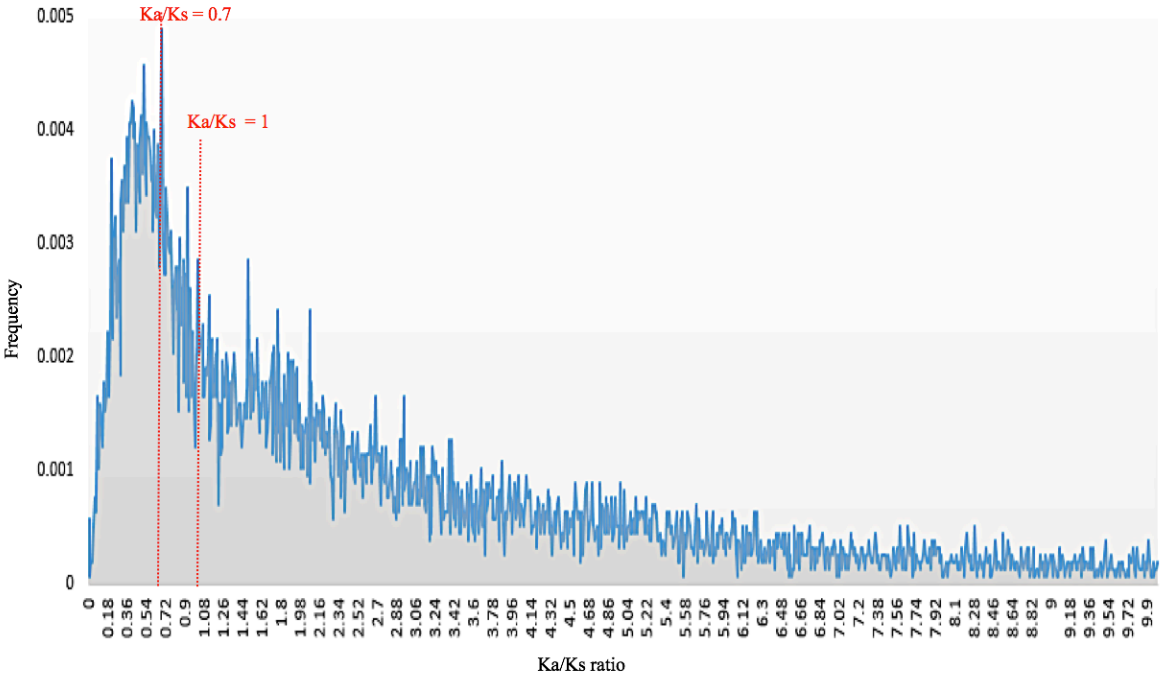


Figure 33: Evolutionary pressure analysis of NFB genes in *T. urticae*.

5.2.7 TE might be a motivation of expansion

A great number of TE in the flanking regions of these NFB clusters were observed (Table 11, Chapter 4). The average number of Gypsy for each NFB gene in *T. urticae* is higher than that of *T. lintearius* and *T. evansi*. Gypsy is one type of retrotransposons, being transcribed from DNA to RNA and then reversibly transcribed back to DNA. This mechanism is also called "copy-and-paste", a principle drive for gene duplication, gene translocation, and gene transposition. In *D. melanogaster*, Gypsy elements are infectious and they encode putative gene products homologous to retroviral proteins (Marlor *et al.*, 1986; Kim *et al.*, 1994; Song *et al.*, 1994). Gypsy is a superfamily of LTR retrotransposons of approximately 7.5 kb in length and widely distributed among animals, fungi, protists, and plants. Its activity can be transferred among *Drosophila* strains by microinjection of egg plasma into embryos or by exposing larvae to viral particles (Huang *et al.*, 2012). It also has been reported that in crocodilian genomes, TE played an important role in the divergence of mammals and reptiles at 310-330 MYA because TE provides the evidence of an extraordinarily low rate of crocodilian genome evolution (Green *et al.*, 2014). Many chemosensory-associated genes that proliferated in *T. urticae* are rich in TE (Ngoc *et al.*, 2016). Since TE can facilitate genomic rearrangement, they may have played an essential role in the observed structural complexity of some large expanded gene family clusters. Meanwhile, TE insertions are generally deleterious for host genes via coding sequence disruption or effects on expression (Cordaux and Batzer, 2009). Consequently, it is assumed that these TE might play an important role in shuffling the *Tetranychus* genomes and thus driving rapid evolution.

5.2.8 Transcriptome profiling

We performed the transcriptome profiling analysis using all the available RNAseq data from the three spider mites. Our results show that only the ancient NFB genes shared by the three mites have a high expression value while those expanded ones are in a relatively low expression status (Figure 34). Those recently proliferated NFB genes might be expressed when *T. urticae* are transferred to different host plants. This hypothesis can be tested with more RNAseq data from *T. urticae* responding to different plant hosts.

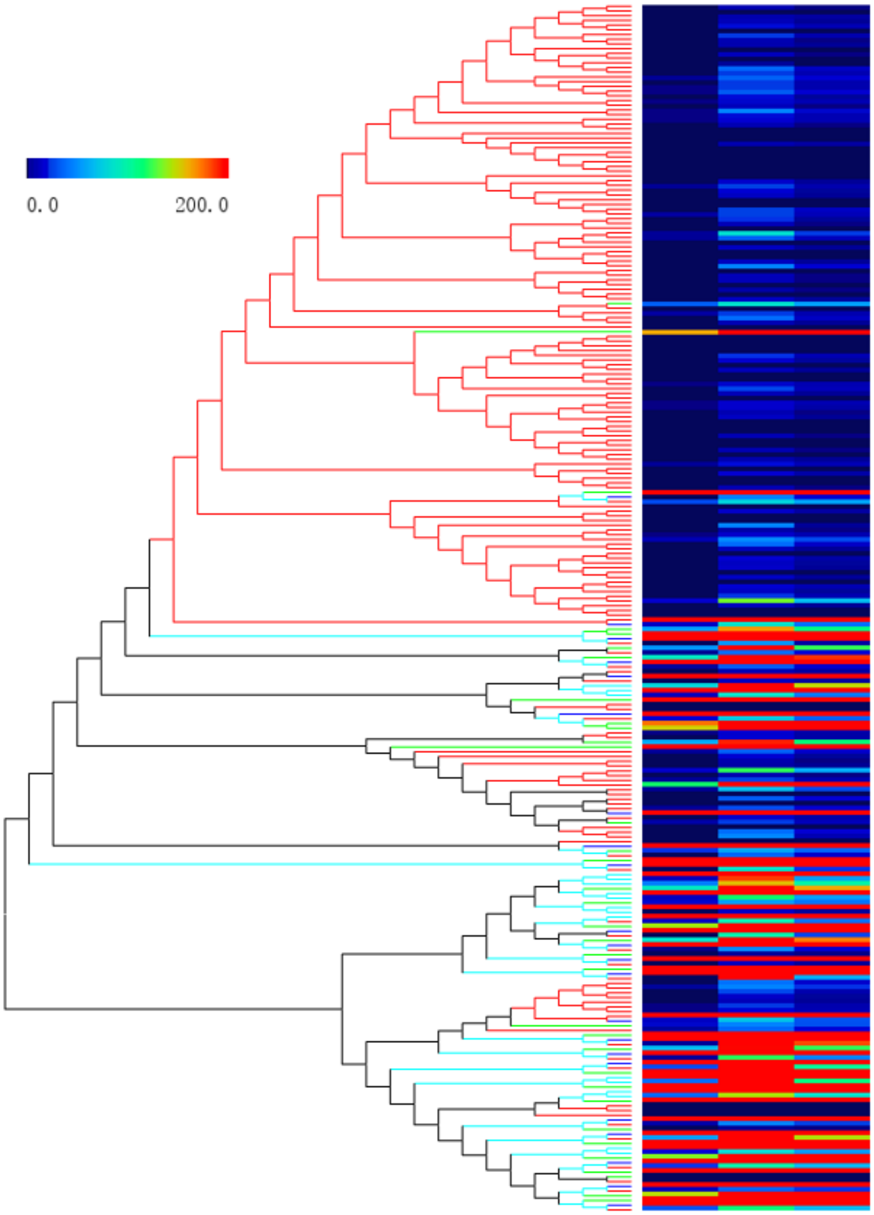


Figure 34: Phylogeny of NFB genes and their expression profiles.
Red: *T. urticae*; green: *T. evansi*; blue: *T. lintearius*; cyan: 3 mites; the three columns of the heat map from the left to the right are the expression data (normalized reads count) in average, maximum and minimum scale, respectively;

5.2.9 Co-expression network of NFB genes with SKP1

To investigate the networks of NFB, CFB, SKP1, and Cullin1, we used the normalized RNAseq data from a previous study (Zhurov *et al.*, 2014a). The gene of Cullin1 tetur17g00940 has an extremely high expression profile, thus the co-expression network excludes it. In total, 3,703 genes correlated to SKP1; of these, 37.5% genes (42 out of 112 expressed NFB genes) having a strong co-relation with SKP1. However, only 24% of CFB genes can be detected as co-expressed with SKP1, using Pearson correlation method in CoExpNetViz (Tzfadia *et al.*, 2015). The result suggests NFB genes probably facilitate with SKP1 to target other proteins. SKP1 and its 42 co-related NFB genes (listed in Table 18) are being tested in subsequent RNAi studies.

5.2.10 Low GC-content

Gene families in *T. urticae* have similar GC-content at about 38%. However, NFB genes have a lower GC-content (34%) but close to the GC-content in the whole *T. urticae* genome (32%). GC-rich DNA sequences are more stable than sequences with lower GC-content, which indicates that these regions are more easily broken due to less hydrogen bond energy. When GC-rich regions form secondary structures, particularly hairpin loops, they are very stable and thus they persist around and accumulate. GC-low DNA may be more flexible and more easily wrapped nucleosomes than GC-rich DNA (Katan-Khaykovich and Struhl, 2002). GC-rich chromatin displays lower interaction frequencies than AT-rich chromatin (Dekker, 2007), suggesting these low GC-content NFB proteins, reversely, probably have more interactions with other proteins.

Table 17: GC-content in the genome of *T. urticae*.

Note	GC-content
CFB	0.3781
NFB	0.3397
ABC transporter	0.3693
P450	0.3631
CCE	0.38
GST	0.3936
All CDS	0.3762

5.3 Summary

This study reports an extreme expansion of a novel gene family NFB in polyphagous pest *T. urticae*. All the current evidence suggests the ancient NFB genes, most likely, actively interact with other proteins. It is hypothesized that such proteins of this gene family would be produced in mite's salivary gland as well and associated with the feeding process. It is also observed that these NFB genes that expanded in *T. urticae* have relatively conserved structures. No detected homologs in any other species make them novel. They also might be expanded through TE meditation. Biological experimentation is required to investigate the biological functions of such novel proteins.

Future studies will be focused the expression of these NFB genes, for example, using RNAi to observe the phenotype of spider mites or by differentially expressed genes of spider mites transferring from various plants, for further validate the biological function of these novel genes.

5.4 Materials and Methods

5.4.1 NFB gene discovery and annotation

Initially, when comparatively analyzed the three genomes, we found there is a large gene family proliferated in *T. urticae* with a copy number of over 200 while only about 20 in other two mites. The precise function of this expanded gene family is unknown and InterProScan domain analysis shows they have F-box and LRR domains (Jones *et al.*, 2014). This unknown F-box gene family triggered our curiosity and thus, we carefully searched extensively across the whole genomes and annotated them manually. Briefly, we chose several of these NFB sequences, which have good RNAseq data support as queries using BLASTP against the three genomes (e-value < 1e-5 and score >50 and amino acid length >50). The BLASTP results were filtered by checking F-box domain using Pfam, and InterProScan (Bateman *et al.*, 2002; Jones *et al.*, 2014). The filtered sequences were then applied as for 1st query set to TBLASTN against the three genomes. The unpredicted genes, fragmental genes and automatically annotated genes with errors were double-checked and manually curated. The final NFB gene family set consists of intact genes, truncated/incomplete genes and pseudogenes (probably caused by frameshifts or silent mutations).

The CFB genes in three spider mites and *Drosophila melanogaster* (downloaded the longest transcripts from http://flybase.org/ on Dec 1st, 2015) were searched extensively using BLASTP and the hits were checked manually. We used InterProScan to search the protein domains to assign the three categories as WD40, LRR and others (Jones *et al.*, 2014).

5.4.2 Alignment and phylogeny analyses

Given the F-box domain of NFB proteins locates at the N-terminus, we extracted the 60 amino acids at the beginning of N-terminus of NFB protein sequences. In total 256 intact NFB F-box domain sequences across the three spider mite genomes were extracted and MUSCLE was applied to align these sequences (Edgar, 2004). We followed a similar approach to construct the phylogenetic tree (Xu *et al.*, 2009). In brief, the evolutionary history was inferred using the Neighbor-Joining method (Saitou and Nei, 1987). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test as 1000 replicates (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the p-distance method as a substitution model (Nei and Kumar, 2000) in the units of the number of amino acid differences per site. All ambiguous positions were removed (i.e. Homogeneous pattern among lineages and pairwise deletions) for each sequence pair and the final uniform rates are among 120 sites. Evolutionary analyses were conducted in MEGA7 (Kumar *et al.*, 2016).

5.4.3 Transcriptome profiling and network analysis

To investigate the expression profiles and networks of NFB, CFB, SKP1, and CULLIN1, we normalized the RNAseq data from the previous study (Zhurov *et al.*, 2014a). The RNAseq data were analyzed by the online tool CoExpNetViz to check the network of these genes (Tzfadia *et al.*, 2015). The SKP1 tetur07g01590 (only skp1 gene in *T. urticae* genome) was used as a bait to fish the correlations across all the genes from *T. urticae* genome by Pearson Correlation Coefficient (using a threshold of lower percentile rank 0.1 and upper percentile rank 0.9). The final heatmap of transcriptome profile was visualized by TM4 MeV (Howe *et al.*, 2011).

5.4.4 Evolutionary selection pressure analysis

To understand the evolutionary selection pressure of these NFB genes, we analyzed the Ka/Ks for the NFB gene families. In short, all F-box genes were extracted from the three genomes' database, both nucleotide and protein sequences. We used CD-HIT (version /cd-hit/x86_64/4.6.1) and clustalw2 (version /clustalw/x86_64/2.1) to cluster and align these protein sequences, respectively (Larkin *et al.*, 2007; Fu *et al.*, 2012). Then all the protein alignments were transferred back to nucleotide alignments accordingly using in-house Perl scripts. We used PAML package phyml (version /paml/x86_64/4.4c) to calculate the selection pressure, or Ka/Ks rate, for NFB families (Yang, 2007). Finally, all the results were extracted and analyzed by in-house Perl scripts.

5.4.5 Transposable element dynamics

We retrieved the TE annotation from ORCAE background MySQL database. The annotation contains TE ID, TE type and loci information. We counted the TE number and type in both flanking regions for each gene, with a window of 10 kb. One TE would be repeatedly counted if it appears within the 10 kb flanking region of two or more genes.

5.4.6 Supplementary data

Table 18: The 42 NFB genes highly correlated to SKP1.

tetur01g08030	tetur01g14990	tetur07g00540	tetur07g02650	tetur07g07890	tetur20g01200
tetur01g08070	tetur01g16190	tetur07g00680	tetur07g02660	tetur07g07950	tetur34g00330
tetur01g08100	tetur02g02430	tetur07g02210	tetur07g02690	tetur07g07970	tetur34g01253
tetur01g08180	tetur02g04990	tetur07g02230	tetur07g03280	tetur07g08169	tetur34g01263
tetur01g11180	tetur02g05450	tetur07g02270	tetur07g03810	tetur120g00010	tetur36g01130
tetur01g11610	tetur05g02140	tetur07g02600	tetur07g06280	tetur16g02170	tetur36g01140
tetur01g11810	tetur06g03470	tetur07g02630	tetur07g07880	tetur20g01180	tetur65g00060

5.5 Reference

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. (2002). The Pfam protein families database. Nucleic acids research *30*, 276-280.

Cazaux, M., Navarro, M., Bruinsma, K.A., Zhurov, V., Negrave, T., Van Leeuwen, T., Grbic, V., and Grbic, M. (2014). Application of two-spotted spider mite Tetranychus urticae for plant-pest interaction studies. Journal of visualized experiments : JoVE.

Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nature reviews Genetics *10*, 691-703.

Dekker, J. (2007). GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. Genome biology *8*, R116.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research *32*, 1792-1797.

Fay, J.C., and Wu, C.I. (2003). Sequence divergence, functional constraint, and selection in protein evolution. Annual review of genomics and human genetics *4*, 213-235.

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution *39*, 783-791.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150-3152.

Gagne, J.M., Downes, B.P., Shiu, S.H., Durski, A.M., and Vierstra, R.D. (2002). The F-box subunit of the SCF E3 complex is encoded by a diverse superfamily of genes in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America *99*, 11519-11524.

Gay, N.J., Packman, L.C., Weldon, M.A., and Barna, J.C. (1991). A leucine-rich repeat peptide derived from the Drosophila Toll receptor forms extended filaments with a beta-sheet structure. FEBS letters *291*, 87-91.

Grbic, M., Van Leeuwen, T., Clark, R.M., Rombauts, S., Rouze, P., Grbic, V., Osborne, E.J., Dermauw, W., Ngoc, P.C., Ortego, F.*, et al.* (2011a). The genome of Tetranychus urticae reveals herbivorous pest adaptations. Nature *479*, 487-492.

Green, R.E., Braun, E.L., Armstrong, J., Earl, D., Nguyen, N., Hickey, G., Vandewege, M.W., St John, J.A., Capella-Gutierrez, S., Castoe, T.A.*, et al.* (2014). Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science *346*, 1254449.

Ho, M.S., Tsai, P.I., and Chien, C.T. (2006). F-box proteins: the key to protein degradation. J Biomed Sci *13*, 181-191.

Howe, E.A., Sinha, R., Schlauch, D., and Quackenbush, J. (2011). RNA-Seq analysis in MeV. Bioinformatics *27*, 3209-3210.

Huang, C.R., Burns, K.H., and Boeke, J.D. (2012). Active transposition in genomes. Annu Rev Genet *46*, 651-675.

Jain, M., Nijhawan, A., Arora, R., Agarwal, P., Ray, S., Sharma, P., Kapoor, S., Tyagi, A.K., and Khurana, J.P. (2007). F-box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress. Plant physiology *143*, 1467-1483.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G.*, et al.* (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236-1240.

Katan-Khaykovich, Y., and Struhl, K. (2002). Dynamics of global histone acetylation and deacetylation in vivo: rapid restoration of normal histone acetylation status upon removal of activators and repressors. Genes Dev *16*, 743-752.

Kim, A., Terzian, C., Santamaria, P., Pélisson, A., Purd'homme, N., and Bucheton, A. (1994). Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America *91*, 1285-1289.

Kipreos, E.T., and Pagano, M. (2000). The F-box protein family. Genome biology *1*, REVIEWS3002.

Kobe, B., and Kajava, A.V. (2001). The leucine-rich repeat as a protein recognition motif. Current opinion in structural biology *11*, 725-732.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Molecular biology and evolution *33*, 1870-1874.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R.*, et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

Marlor, R.L., Parkhurst, S.M., and Corces, V.G. (1986). The Drosophila melanogaster gypsy transposable element encodes putative gene products homologous to retroviral proteins. Mol Cell Biol *6*, 1129-1134.

Navarro-Quezada, A., Schumann, N., and Quint, M. (2013). Plant F-box protein evolution is determined by lineage-specific timing of major gene family expansion waves. PloS one *8*, e68672.

Nei, M., and Kumar, S. (2000). Molecular Evolution and Phylogenetics (Oxford University Press).

Ngoc, P.C., Greenhalgh, R., Dermauw, W., Rombauts, S., Bajda, S., Zhurov, V., Grbic, M., Van de Peer, Y., Van Leeuwen, T., Rouze, P., *et al.* (2016). Complex Evolutionary Dynamics of Massively Expanded Chemosensory Receptor Families in an Extreme Generalist Chelicerate Herbivore. Genome biology and evolution *8*, 3323-3339.

Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature methods *8*, 785-786.

Rothberg, J.M., Jacobs, J.R., Goodman, C.S., and Artavanistsakonas, S. (1990). Slit - an Extracellular Protein Necessary for Development of Midline Glia and Commissural Axon Pathways Contains Both Egf and Lrr Domains. Gene Dev *4*, 2169-2187.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution *4*, 406-425.

Song, S.U., Gerasimova, T., Kurkulos, M., Boeke, J.D., and Corces, V.G. (1994). An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus. Genes Dev *8*, 2046-2057.

Tzfadia, O., Diels, T., De Meyer, S., Vandepoele, K., Aharoni, A., and Van de Peer, Y. (2015). CoExpNetViz: Comparative Co-Expression Networks Construction and Visualization Tool. Frontiers in plant science *6*, 1194.

Wang, A., Fu, M., Jiang, X., Mao, Y., Li, X., and Tao, S. (2014). Evolution of the F-box gene family in Euarchontoglires: gene number variation and selection patterns. PloS one *9*, e94899.

Xu, G., Ma, H., Nei, M., and Kong, H. (2009). Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. Proceedings of the National Academy of Sciences of the United States of America *106*, 835-840.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution *24*, 1586-1591.

Zhao, C., Escalante, L.N., Chen, H., Benatti, T.R., Qu, J., Chellapilla, S., Waterhouse, R.M., Wheeler, D., Andersson, M.N., Bao, R., *et al.* (2015). A massive expansion of effector genes underlies gall-formation in the wheat pest Mayetiola destructor. Curr Biol *25*, 613-620.

Zheng, N., Schulman, B.A., Song, L., Miller, J.J., Jeffrey, P.D., Wang, P., Chu, C., Koepp, D.M., Elledge, S.J., Pagano, M., *et al.* (2002). Structure of the Cul1-Rbx1-Skp1-F boxSkp2 SCF ubiquitin ligase complex. Nature *416*, 703-709.

Zhurov, V., Navarro, M., Bruinsma, K.A., Arbona, V., Santamaria, M.E., Cazaux, M., Wybouw, N., Osborne, E.J., Ens, C., Rioja, C.*, et al.* (2014a). Reciprocal responses in the interaction between Arabidopsis and the cell-content-feeding chelicerate herbivore spider mite. Plant physiology *164*, 384-399.

Chapter 6

6 The genomes of *Brevipalpus* flat mites and their *Cardinium* endosymbionts offer insights into herbivorous pest adaptation and parthenogenesis

The *Brevipalpus* flat mites are major agricultural pests primarily feeding on citrus, grapes and other fruit plants. They have risen from near obscurity to that of considerable economic importance over the past decades (Lal, 1979; Childers and Derrick, 2003; Childers *et al.*, 2003; Kitajima *et al.*, 2003; Groot *et al.*, 2005; De Carvalho Mineiro *et al.*, 2008; Rodrigues and Childers, 2013; Beard *et al.*, 2015). Interestingly, *Cardinium* symbionts can induce haploid thelytoky in flat mites through infecting females to reproduce only female progeny, and non-*Cardinium*-infected females reproduce a few male progenies. They also have an effect on the feminization of flat mites (Chigira and Miura, 2005; Groot and Breeuwer, 2006).

To address genomic feeding signatures and host plants adaptation of flat mites as well as *Cardinium* endosymbiont mechanisms, five different *Brevipalpus* genomes (*Brevipalpus yothersi* – both Brazillian and Amsterdam strains, *Brevipalpus californicus* – both infected and uninfected, and *Brevipalpus papayensis*) and their corresponding symbionts *Cardinium* genomes were sequenced and analyzed. The flat mites have the smallest genome size of arthropod genomes reported so far, with an average of 70 Mb. The gene family analysis shows that known gene families associated with digestion and detoxification are often expanded. The *Cardinium* species in their host might originate from different strains, which suggests that host mites could harbor either multi-infections or integration of *Cardinium* into their hosts. Moreover, *Cardinium*-infected *Brevipalpus* species undergo parthenogenesis while *Cardinium*-infected *Bemisia* species or *Encarsia* species undergo cytoplasmic incompatibility. This study on *Brevipalpus* genomes along with the endosymbionts will highlight the pest feeding genomic signatures, host-symbiont coevolution, symbiont motility and the differences in the asexuality of different mite clonal lineages.

6.1 Introduction

The genus *Brevipalpus* flat mites are commonly referred as flat mites or false spider mites because of the flat-shaped body and inability to spin webs. They are recognized as serious economic plant pests since they can be a vector of one or more cytoplasmic or nuclear type plant viruses including coffee ringspot, green spot on passion fruit, orchid fleck viruses as well as citrus leprosis disease, an important viral disease affecting citrus crops. Such emerging disease is widely distributed in South and Central America, from Argentina to Mexico (Childers and Rodrigues, 2011). Flat mites feed on plants by inserting their body-size mouthparts into the plant tissues, injecting toxic saliva and sucking plant cell contents. Distinguished from spider mites primarily that feed on plant leaves, flat mites feed on plant leaves as well as plant fruits, stems, nuts, and buds.

Flat mites have a body color from dark green to red-orange. Males are more wedge-shaped than females. Flat mites are the approximately half the size of spider mites. Male flat mites are haploid (n=2 chromosomes) while females are diploid (2n=4 chromosomes) (Childers *et al.*, 2003). Flat mites have four active stages in their life cycle including egg, larva, nymph, and adult. A female flat mite lays approximately 50 eggs in a lifetime, less than over hundreds of eggs reproduced by a spider mite. These eggs hatch in 8 to 16 days before becoming larvae (Childers and Rodrigues, 2011). Their longevity usually lasts longer than spider mites, which is over one month or about 5 to 7 weeks.

The *Brevipalpus* consists of several species that are among the most important economic pests in flat mite family (Childers and Rodrigues, 2011). *B. californicus*, sometimes called omnivorous mite, has an extensive host range feeding on citrus, oranges, and mandarins (Salinas-Vargas *et al.*, 2016). It is also known vector of orchid fleck virus was found in Asia, Australia, Europe and America (Kondo H, 2006). *B. papayensis*, known as citrus leprosis mite or passion-vine mite, is a global pest of economic crops such as citrus, tea, papaya, and coffee. Another type of flat mites, *B. yothersi,* is reported to transmit viruses associated with two major cytopathology groups (Rodrigues *et al.*, 2003; Adams *et al.*, 2015). *B. yothersi* is also another crucial pest that has a strong impact on all citrus species (Salinas-Vargas *et al.*, 2016).

Many flat mites harbor inherited bacterial endosymbionts (an endosymbiont or endobiont is any organism that lives within the body or cells of another organism) that are maternally transmitted and have an impact on their hosts' biology, ecology, and evolution. Endosymbionts can be a key to host survival under specific environmental conditions, such as parasitoid attack, climate change, or insecticide pressure. One of the most common phenotypes of facultative symbionts appears to be cytoplasmic incompatibility (CI), a type of reproductive failure, in which bacteria modify sperms of male flat mites in a way that reduces the reproductive success with uninfected female mates. Furthermore, reproductive manipulator symbionts may also be useful in pest management for suppression or transformation of pests (Zabalou *et al.*, 2004; Walker *et al.*, 2011; Blagrove *et al.*, 2012).

Wolbachia is a type of well-studied endosymbionts *(Sun et al., 2001; Walker et al., 2011; Blagrove et al., 2012; Ros et al., 2012)*. Other than that*, Cardinium,* a symbiont of tiny parasitic wasps, is a recently discovered maternally transmitted bacterial endosymbiont and causes CI in arthropods (Zhang *et al.*, 2010). However, CI is evolved independently in *Wolbachia* and *Cardinium* (Penz *et al.*, 2012). One of the most striking phenomena is that *Cardinium* can be associating with *Brevipalpus* female-only colonies. *Cardinium* can induce abnormal phenotypes of *Brevipalpus* reproduction including CI, parthenogenesis, and feminization. Furthermore, *Cardinium* can influence the fitness of its host in addition to manipulating the reproduction of the hosts (Chigira and Miura, 2005; Groot and Breeuwer, 2006; Kitajima *et al.*, 2007). *Cardinium* genomes have a large proportion of transposable elements, leading to gene inactivation, chromosomal rearrangements, and duplication (Santos-Garcia *et al.*, 2014). Phylogenetic evidence shows that these bacteria must have been laterally transferred between mite clonal lineages and may facilitate the lateral gene transfer between mite hosts (Ros *et al.*, 2012; Santos-Garcia *et al.*, 2014).

Previous studies preferably focused on CI, genome reduction, symbiont mobility, and settlement of *Cardinium* in insects, but rarely on mites (Nakamura *et al.*, 2009; Penz *et al.*, 2012; Santos-Garcia *et al.*, 2014). No flat mite genomes have yet been sequenced and reported. Additionally, *Cardinium* genomes and the mechanism of them hosting in *Brevipalpus* is still poorly understood. Therefore, in this study, five Brevipalpus strains were sequenced and analyzed: two *B. californicus* (infected strain and uninfected strain which means this strain was treated using antibiotics to kill bacteria), two *B. yothersi* (Amsterdam strain and Brazilian strain) and *B. papayensis*. The aim of this study is to understand feeding mechanisms of different flat mites. The *Cardinium* endosymbionts from each *Brevipalpus* genome were also assembled and analyzed for the investigation on the mechanisms of endosymbiosis and feminization.

6.2 Results and discussion

6.2.1 Genome statistics of the *Brevipalpus* stains

The *B. yothersi* (Brazilian strain) was sequenced by hybrid reads datasets (454 sequencing technology with SE reads and MiSeq technology with SE, PE, MP reads) with an average coverage of 42x. The other four *Brevipalpus* strains were sequenced using Illumina PE reads (2x250 bp) with an average coverage of 230x. Respectively, the genome sizes of *B. yothersi* Brazilian strain and Amsterdam strain were 71.2 Mb and 71.9 Mb (Table 19). The other three strains are smaller, only about 67 Mb in size. We identified 12,777 protein-coding genes in *B. yothersi* (Brazilian strain) and the similar gene numbers in the other four strains. The complete genome datasets of these genomes are stored at http://bioinformatics.psb.ugent.be/orcae.

With an assembled genome size of 70 Mb, *Brevipalpus* genome is the smallest arthropod genome sequenced so far, smaller than *Tetranychus urticae* (90 Mb) (Grbic *et al.*, 2011a). The genome sizes of other chelicerates are much larger (Table 32), except for the recently published tick genome (*Ixodes scapularis*) 1.8 Gb (Gulia-Nuss *et al.*, 2016). Similar to the genome of *T. urticae*, multiple characteristics of the *Brevipalpus* genomes correlate with their compact size: small transposable element content (10%) and increased gene density (180 genes/Mb).

Table 19: Genomic statistics of the *Brevipalpus* strains.

Category*	*B. yothersi Brazil*	*B. yothersi Amsterdam*	*B. papayensis*	*B. californicus*
genome size (scaffolds)	71,162,551 nt	71,923,315 nt	67,204,343 nt	66,670,430 nt
genome size (contigs)	70538551 nt	71845897 nt	67077365 nt	66,575,954 nt
largest scaffold	766,678 nt	524,526 nt	477,117 nt	366,163 nt
av. scaffold length	84,315.82 nt	43,093.66 nt	35,539.05 nt	39,567.02 nt
number of contigs	3,444	3,091	4,446	3,419
largest contig	247,827 nt	318,950 nt	255,439 nt	329,840 nt
av. contig length	20,481.58 nt	23,243.58 nt	15,087.13 nt	19,472.35 nt
N50 length (kb)	170	139	80	84
nr_loci (exons+introns)	12,777	13,448	12,499	12,476
av.length.loci	2,401.43 nt	2,310.34 nt	2,219.65 nt	2,353.08 nt
loci density (nt/gene)	5,520.74	5,342.50	5,366.62	5,336.32
nr_genes	12,777	13,448	12,499	12,476
gene density (genes/Mb)	181.13	187.18	186.34	187.39
av.length.genes	1,474.65 nt	1,447.57 nt	1,452.46 nt	1,576.43 nt
median.length.genes	1,131 nt	1,119 nt	1,110 nt	1,215 nt
nr_exons	43,224	44,233	41,129	41,986
%GC of CDS	40.05	39.99	40.54	40.48
cumul_exon_length	18,841,567 nt	19,466,892 nt	18,154,272 nt	19,667,520 nt
av.length.exons	435.91 nt	440.10 nt	441.40 nt	468.43 nt
median.length.exons	218 nt	222 nt	221 nt	243 nt
av.nr.exons/gene	3.38	3.29	3.29	3.37
most exons/gene	27 bryot164g00100	31 brpho63g00350	27 brobo288g00010	25 brcal594g00020
av.length.CDS	1,474.65 nt	1,447.42 nt	1,452.23 nt	1,463.73 nt
cumul_intron_length	10,310,514 nt	10,916,528 nt	9,089,164 nt	9,337,867 nt
av.length.intron	342.04 nt	355.54 nt	318.72 nt	317.39 nt
median.length.introns	100 nt	101 nt	106 nt	103 nt
%GC of intron	34.5	34.54	34.74	35.3

**B. californicus* (the infected stain) is not included in this table.

The Brazilian stain and Amsterdam strain of *B. yothersi* contain 9.77% and 9.68% TEs, respectively (Table 20). TE proportions in *B. californicus* are slightly lower, about 7.09% and 7.36% in uninfected strain and infected strain. *B. papayensis* genome was masked with 8.42% TE. Overall, TE in *Brevipalpus* flat mites is less than that in *Tetranychus* spider mite (11%) (Grbic *et al.*, 2011a), but still considerably less than TE proportion in tick genome (70%), human genome (44%) or maize genome (90%) (SanMiguel *et al.*, 1996; Mills *et al.*, 2007a; Gulia-Nuss *et al.*, 2016).

Table 20: TE distribution across five flat mite genomes.

% TE	*B. yothersi (Brazilian strain)*	*B. californiucs (infected)*	*B. californiucs (uninfected)*	*B. papayensis*	*B. yothersi (Amsterdam strain)*
otherLTR	0.31	0.16	0.16	0.21	0.29
Gypsy	1.61	0.66	0.68	0.79	1.54
Copia	0.61	0.23	0.23	0.21	0.59
LINE	0.03	0.01	0.01	0.01	0.03
L1	0.36	0.18	0.19	0.24	0.37
CR1	0.15	0.05	0.06	0.06	0.15
R2	0.03	0	0.01	0.01	0.04
I	0.1	0.04	0.04	0.03	0.08
LOA	0.05	0	0	0.01	0.06
SINE	0	0	0	0	0
Kolobok	0.05	0.02	0.02	0.03	0.04
Kiri	0.03	0.01	0.01	0.01	0.03
Jockey	0.05	0.02	0.02	0.05	0.05
ISL2EU	0.69	0.35	0.37	0.34	0.73
hAT	0.19	0.09	0.09	0.11	0.19
Mariner	0.32	0.21	0.21	0.18	0.37
PiggyBac	0.03	0.01	0.01	0.03	0.03
Merlin	0.01	0	0	0	0.01
CACTA	0	0	0	0	0
hAT	0.19	0.09	0.09	0.11	0.19
MITE	0	0	0	0	0
Harbinger	0.07	0.02	0.02	0.04	0.09
Penelope	0.06	0.02	0.02	0.03	0.07
Polinton	0.17	0.09	0.09	0.11	0.18
Helitron	0.18	0.07	0.07	0.05	0.19
Maverick	0	0	0	0	0
unclassified(SSR* incl.)	4.48	4.76	4.96	5.76	4.36
Total	**9.77**	**7.09**	**7.36**	**8.42**	**9.68**

*SSR – Simple Sequence Repeat, also known as microsatellite repeat, ranges in length (from 2 bp to 5 bp) with, typically, 5 to 50 times repeats (Turnpenny and Ellard, 2012).

Table 21: SNP calling using GATK with *B. yothersi* (Brazilian strain) as a reference.

VCF-statistics*	*B. yothersi (Amsterdam)*	*B. obovatus*	*B. yothersi (Brazil) (self-calling)*	*B. californicus uninfected*	*B. californicus infected*
hom_AA_count	401,844	2,474,513	86	2,460,736	2,488,989
het_RA_count	86,705	93,785	9,543	136,812	87,001
snp_count	489,878	2,571,176	9,639	2,600,153	2,578,567
ref	86,705	93,785	9,543	136,812	87,001
private	400,097	1,472,155	5,117	50,367	61,625
missing	4,093,913	2,012,615	4,574,152	1,983,638	2,005,224
het_AA_count	1,329	2,878	10	2,605	2,577
unphased	489,878	2,571,176	9,543	2,600,153	2,578,567
ref_count	86,705	93,785	9,639	136,812	87,001

*hom_AA: homozygous for a single alternate allele (eg. both alleles have the same mutation); het_AA: both alleles are non-reference but they are not the same allele (e.g. one has the S98A mutation and the other has the L206P mutation). Think of it as het_A1A2 if that helps; hom_RR: homozygous reference; het_RA: one reference allele, one alternate allele.

6.2.2 Phylogeny shows a recent divergence of the five genomes

The SNP calling results (Table 21) show *B. yothersi* (Amsterdam strain) has 489,878 SNPs, using *B. yothersi* (Brazilian strain) as a reference genome. It is slightly less than the 542,600 SNPs in *T. urticae* that differentiate the London strain and Montpellier strain (Grbic *et al.*, 2011a). *B. californicus*, both infected and uninfected strains, have 2,578,567 and 2,600,153 SNPs, respectively. This hints approximately 3.5% nucleotide variances across different *Brevipalpus* strains. Principle Component Analysis (PCA) in Figure 35a shows *B. yothersi* strains (Brazilian and Amsterdam strains) are close but *B. californicus* strains have a slight evolutionary distance. However, in theory, *B. californicus* strains are supposed to be genetically closer. This might be due to the lack of 3D visualization of this PCA graph. To confirm their relationships, the phylogeny across the five mites was built up (Figure 35) and now it clearly shows both *B. yothersi* strains are close and *B. californicus* strains are close as well. To further verify this, 5,000 SNP sites were randomly selected to rebuild phylogenetic trees, which also validate the relationships of the five flat mites (Supplementary data: Figure 52).

B. papayensis evolutionarily locate aside of *B. yothersi* and *B. californicus*, which implies that *B. papayensis* was probably evolved after their ancestors, but the evolutionary status of all the five *Brevipalpus* is unknown yet. To address this question, a phylogeny analysis was performed using single copy genes from 31 arthropod genomes and one tardigrade genome as an outgroup. The result suggests *B. papayensis* is closer to *B. yothersi* than *B. californicus*, which might be the ancestor of the five strains. Previous evidence shows mites belong to Acari with the earliest fossils dating 410 MYA. The phylogeny (Supplementary data: Figure 53) shows *Brevipalpus* and *Tetranychus* diverged around 140 MYA, much more ancient than the divergence within their species. In addition, the phylogenetic tree shows that ticks and mites diverged about 260 MYA, which is supported by the study of ticks diverged as far back as 300 ± 27 MYA (Jeyaprakash and Hoy, 2009). The tick genome size is twenty-fold of mite genomes, which suggests Acari has diverged quite differently over the past two hundred million years (Dunlop and Selden, 2009; Dunlop, 2010).

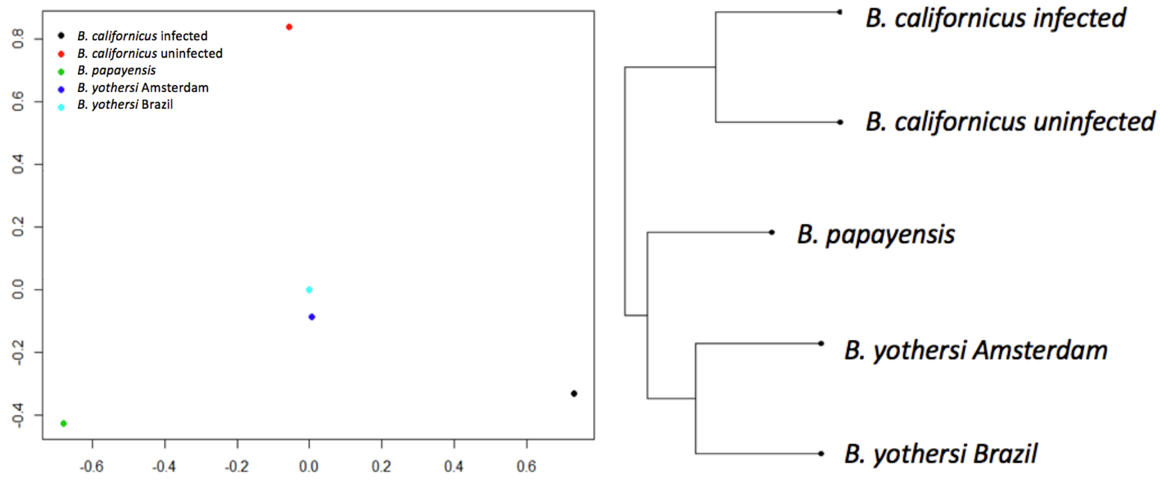


Figure 35: The relationship of the five flat mites using SNP calling results. (left) PCA analysis; (right) phylogeny using all SNP sites;

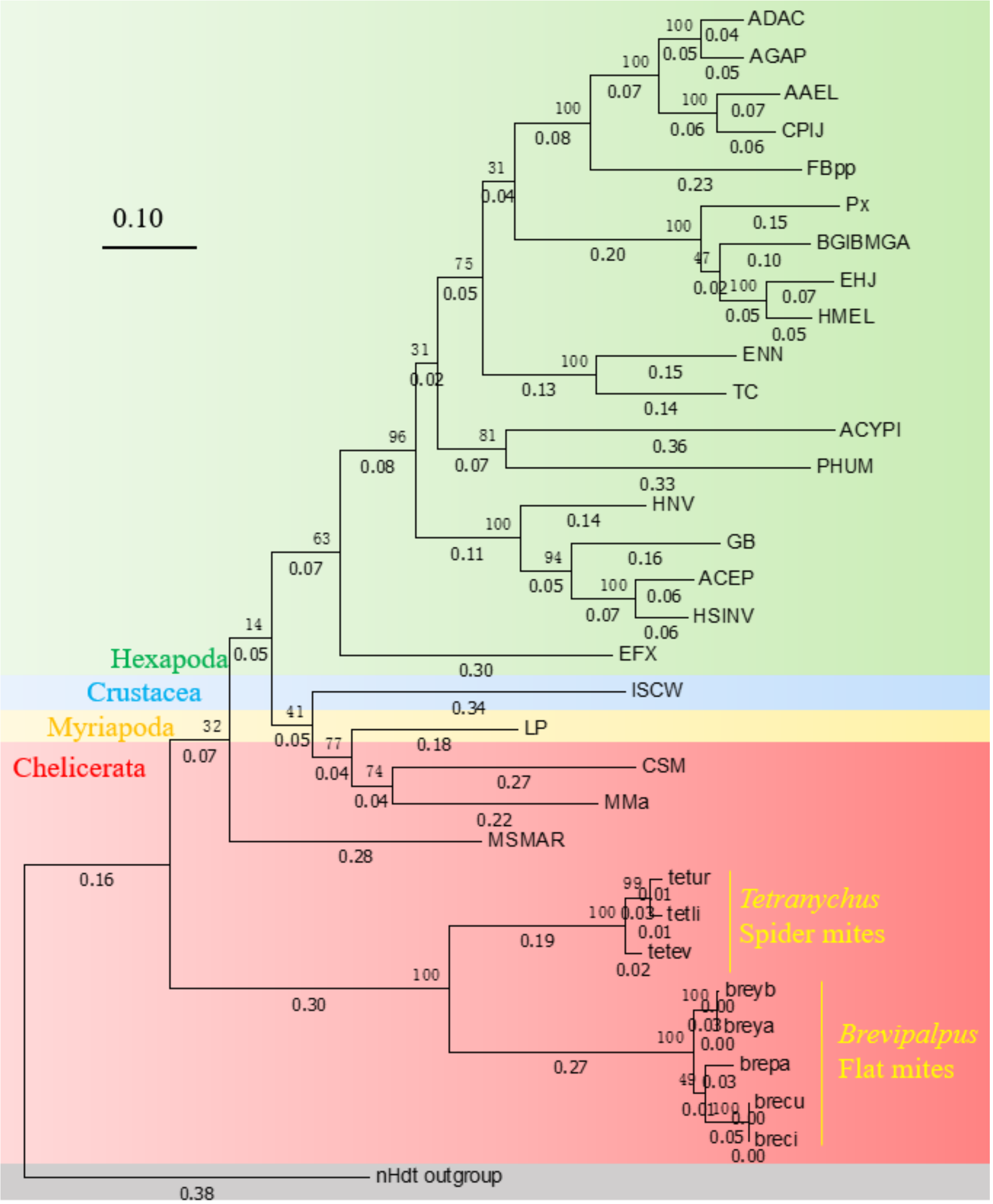


Figure 36: Phylogeny of 31 arthropod genomes using single copy gene dataset. The full names in this phylogenetic tree can be found in the *List of Abbreviations* section on page xxii. The numbers at branch cross indicate bootstrap value and the numbers below the branches indicate branch length.

6.2.3 Linkage-specific gene families

Comparative analysis of gene families across *Brevipalpus* and *Tetranychus* show that feeding and detoxification genes are at different expansion levels (Table 22). *T. urticae* shares almost the same amount of P450 with *Brevipalpus* but *Breviaplus* has more Glutathione S-transferases (GST). Strikingly, the result shows that two gene families in *Brevipalpus* have extremely expanded: glucose dehydrogenase and leukocyte elastase inhibitor. Glucose dehydrogenase is an enzyme that has two substrates (D-glucose and acceptor), whereas its two products are D-glucono-1,5-lactone (GDL) and reduced acceptor. GDL is commonly found in honey, fruit juices, personal lubricants and wine. That would be an explanation why flat mites prefer citrus because these expanded glucose dehydrogenase genes may help in their digestion systems. The leukocyte elastase inhibitor is also known as serpin B1 that regulates the activity of neutrophil serine proteases such as elastase, Catharpin G and proteinase-3. Leukocyte elastase in human is released during inflammation and damage the homeostasis the inhibitor may inhibit the release little is known in arthropods. It may play a regulatory role to limit inflammatory damage due to proteases of cellular origin (Cooley *et al.*, 2001).

Table 22: The raw OrthoMCL results of gene families.

breya#*	breca#	breyb#	brepa#	tetur#	tetev#	tetli#	Function
112	133	115	99	132	65	58	P450
109	131	114	126	136	106	105	ABC transporter
77	76	77	78	69	42	42	Carboxyl/cholinesterase
11	19	12	14	41	19	22	Glutathione S-transferase
1	1	1	1	204	18	14	Novel F-box in spider mite
53	46	51	48	20	17	16	glucose dehydrogenase
47	45	48	43	18	4	10	Leukocyte elastase inhibitor

*breya - *B. yothersi* (Amsterdam strain); breyb - *B. yothersi* (Brazilian strain); breca - *B. californiucs* (uninfected); brepa - *B. papayensis*; tetur-*T. urticae*, tetev - *T. evansi*; tetli - *T. lintearius*. The numbers (#) in this table are directly extracted from OrthoMCL raw results.

6.2.4 Genomes statistics of the *Cardinium* stains

The *Cardinium* genomes from each *Brevipalpus* strains were assembled and annotated (Table 23). Surprisingly, *B. californicus* uninfected strains still contain *Cardinium* genome. This *B. californicus* uninfected strain was treated with antibiotics to kill bacteria inside but it still contains *Cardinium*. This suggests either the ancestors of *B. californicus* already had *Cardinium* or the *Cardinium* was laterally transferred into *B. californicus* genomes.

The *de novo* assembly of cByotB1 (abbreviations in methods and materials) resulted in 1 Mb with 37% GC-content, close to the genome size of two published *Cardinium* strains cPer1 and cBtQ1 (Penz *et al.*, 2012; Santos-Garcia *et al.*, 2014). The completeness of contigs demonstrates our *Cardinium* strains cover about 60% single copy BUSCO genes, while cPer1 and cBtQ1 have around 54% BUSCO genes (Supplementary data: Table 25). The results suggest these assemblies were reliable for downstream analyses.

The genomic reads of cByotB1 were mapped to cBtQ1 and cEper1, respectively. The average genomic coverage is 29%, suggesting the genomes of cBtQ1 and cEper1 might not phylogenetically close to cByotB1 because of the following reasons: first, our genomic reads are from *B. yothersi* scaffold_1 and scaffold_88, not the extraction of a complete *Cardinium* genome assembly; Second, this 29% also hints that cByotB1 might possibly be a different species, comparing with cBtQ1 and cEper1.

The five *Cardinium* genomes share 502 genes (Figure 37), which take up over half their genomes, from the lowest 55% to the highest 71%. It indicates these *Cardinium* still have a majority of house-keeping genes to maintain basic cell life activities. The synteny of the five genomes shows (Figure 38) that, again, cBcal1 and cBal2 are quite consistent in the assembly while cByotB1 and cByotA1 are slightly different, probably due to genomic rearrangements.

Table 23: Genomic features of *Cardinium* strains.

	cByotB1	cByotA1	cBcal1	cBcal2	cBpap1	cBtQ1	cEper1
Genome size (kp)	1,087	1,150	1,050	1,052	1,057	1,033	887
Plasmids (kb)	NA	NA	NA	NA	NA	52	58
GC-content	36.7	36.7	36.9	36.9	36.6	35	32
CDS	920	966	888	891	865	709	841
rRNA	5	10	10	7	9	3	3
tRNA	35	36	36	34	34	35	37

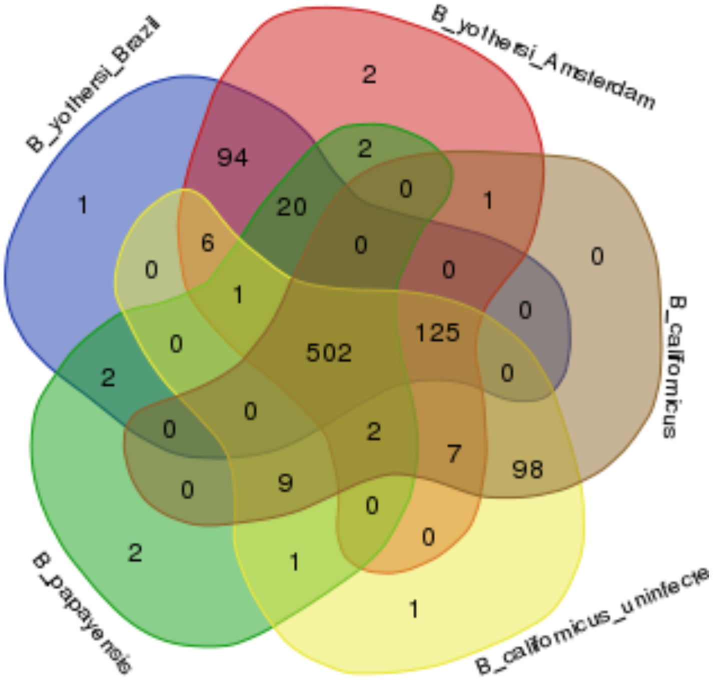


Figure 37: The Venn graph indicates their conserveness at the gene level.
Graph credit: Dr. Phuong Le at VIB, Gent, Belgium.

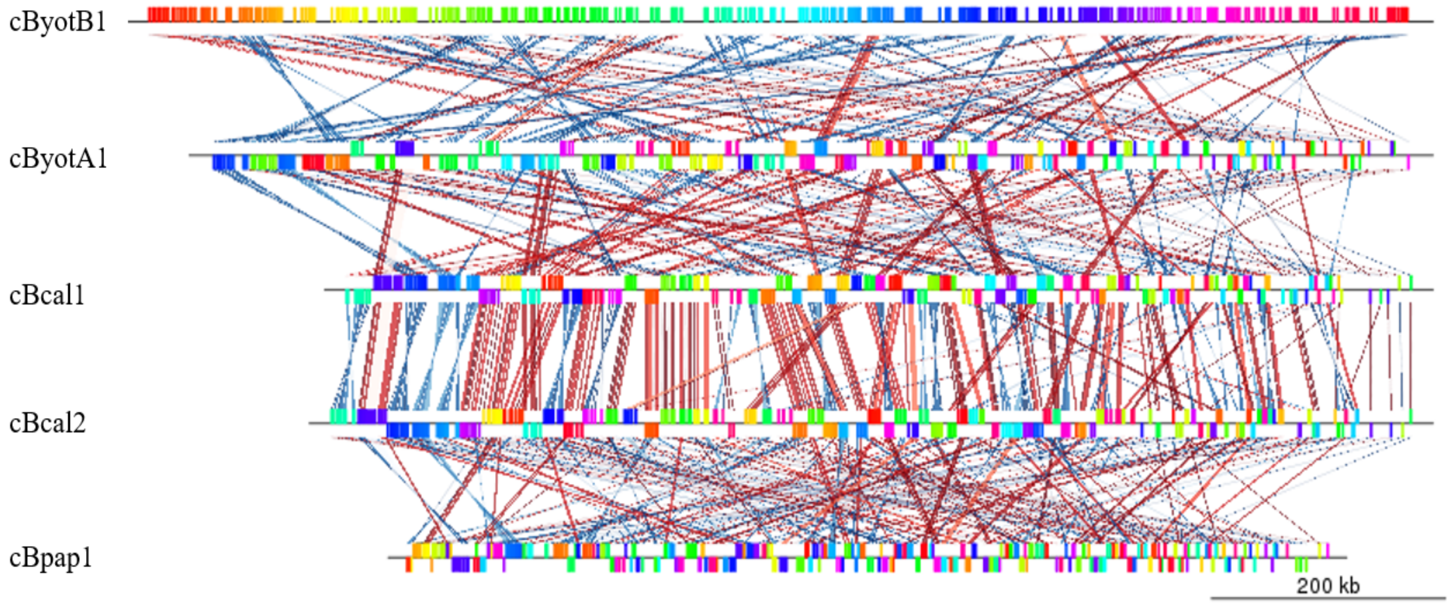


Figure 38: The synteny of *Cardinium* strains in five *Brevipalpus* genomes.
The published two genomes were excluded because of low synteny. Figure credit: Dr. Phuong Le at VIB, Gent, Belgium.

6.2.5 Phylogeny of *Cardinium* strains

Our phylogenetic tree based on 39 single copy genes of Bacteroidetes shows that the five *Cardinium* genomes in this study formed a clade with cBtQ1 and cEper1 with well-supported bootstrap values (100%) (Figure 39). However, cBpap1 did not cluster with the other four *Cardinium* (cByotB1, cByotA1, cBcal1, cBcal2), but cBpap1 clustered with *Cardinium* cPer1 and cBtQ1. Similarly, the phylogenetic tree of 192 single core genes of seven *Cardinium* genomes shows *Cardinium* cBpap1 formed a cluster with two *Cardinium* cEper1 and cBtQ1 (Figure 40). Our phylogeny of *Cardinium* strains shows that cByotB1, cByotA1, cBcal1, and cBcal2 were clustered together with public data, at a high bootstrap of 95, shown in Figure 40. The cBpap1 strain was nearest to *Cardinium* of the genus of *Tetranychus* with a bootstrap score of 72. The cBpap1 strain did not form a monophyletic group with other *Cardinium* species.

The phylogenetic analysis of symbionts within *Cardinium* species, based on the analysis of different phylogenetic trees including gyrB (DNA gyrase subunit B) and single copy genes, shows two separate clades. The cByotB1, cByotA1, cBcal1, and cBcal2 belong to the one clade whereas the cBpap1 belongs to another. Using gyrB gene sequences, we constructed a phylogenetic tree to determine the phylogenetic position of *Cardinium* found in *Brevipalpus* hosts. The phylogeny shows, again, cBcal1 and cBcal2 are closer, so are cByotA1 and cByotB1, as shown in Figure 41. Based on this phylogeny evidence, it hints that *Cardinium* genomes may have a co-evolution with *Brevipalpus* genomes.

There are two possible explanations. First, the bacteria may have been horizontally transmitted among the three species and has similarity been recognized between species within a genus. Second, an ancestral *Brevipalpus* species were infected by *Cardinium*. The phylogeny of *Brevipalpus* is congruent to the phylogeny of *Cardinium*, leading to the latter hypothesis that *Cardinium* infected the ancestral *Brevipalpus*. The close phylogenetic relationship of *Cardinium* genomes and the phylogenomic reconstruction of the Bacteroidetes clade, after the divergence of two *Cardinium* endosymbionts, force us to make the conclusion that the most plausible scenario was an ancestral infection of *Cardinium* endosymbionts.

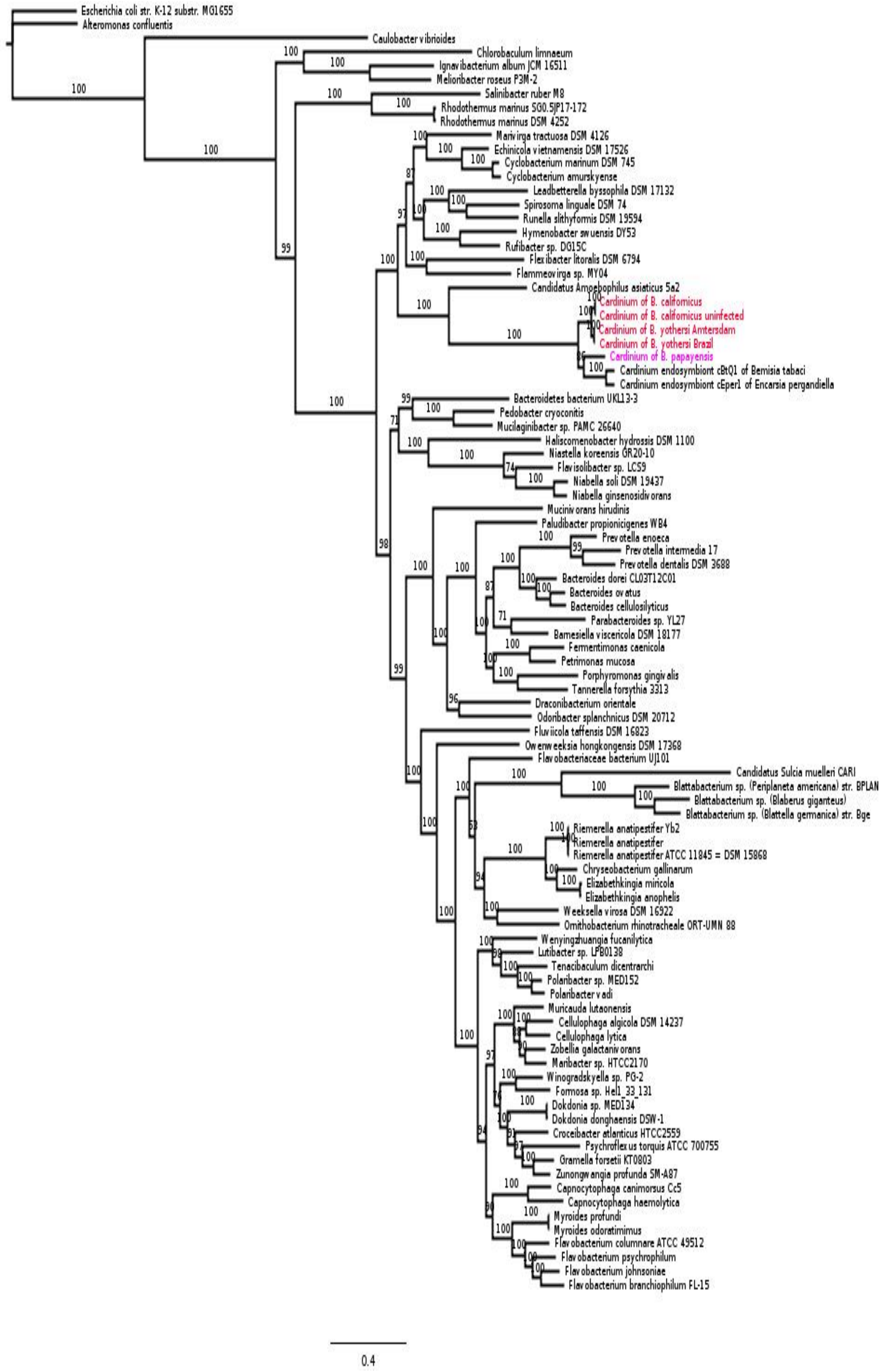

Escherichia coli str. K-12 substr. MG1655
Alteromonas confluentis
Caulobacter vibrioides
Chlorobaculum limnaeum
Ignavibacterium album JCM 16511
Melioribacter roseus P3M-2
Salinibacter ruber M8
Rhodothermus marinus SG0.5JP17-172
Rhodothermus marinus DSM 4252
Marivirga tractuosa DSM 4126
Echinicola vietnamensis DSM 17526
Cyclobacterium marinum DSM 745
Cyclobacterium amurskyense
Leadbetterella byssophila DSM 17132
Spirosoma linguale DSM 74
Runella slithyformis DSM 19594
Hymenobacter swuensis DY53
Rufibacter sp. DG15C
Flexibacter litoralis DSM 6794
Flammeovirga sp. MY04
Candidatus Amoebophilus asiaticus 5a2
Cardinium of B. californicus
Cardinium of B. californicus uninfected
Cardinium of B. yothersi Amtersdam
Cardinium of B. yothersi Brazil
Cardinium of B. papayensis
Cardinium endosymbiont cBtQ1 of Bemisia tabaci
Cardinium endosymbiont cEper1 of Encarsia pergandiella
Bacteroidetes bacterium UKL13-3
Pedobacter cryoconitis
Mucilaginibacter sp. PAMC 26640
Haliscomenobacter hydrossis DSM 1100
Niastella koreensis GR20-10
Flavisolibacter sp. LCS9
Niabella soli DSM 19437
Niabella ginsenosidivorans
Mucinivorans hirudinis
Paludibacter propionicigenes WB4
Prevotella enoeca
Prevotella intermedia 17
Prevotella dentalis DSM 3688
Bacteroides dorei CL03T12C01
Bacteroides ovatus
Bacteroides cellulosilyticus
Parabacteroides sp. YL27
Barnesiella viscericola DSM 18177
Fermentimonas caenicola
Petrimonas mucosa
Porphyromonas gingivalis
Tannerella forsythia 3313
Draconibacterium orientale
Odoribacter splanchnicus DSM 20712
Fluviicola taffensis DSM 16823
Owenweeksia hongkongensis DSM 17368
Flavobacteriaceae bacterium UJ101
Candidatus Sulcia muelleri CARI
Blattabacterium sp. (Periplaneta americana) str. BPLAN
Blattabacterium sp. (Blaberus giganteus)
Blattabacterium sp. (Blattella germanica) str. Bge
Riemerella anatipestifer Yb2
Riemerella anatipestifer
Riemerella anatipestifer ATCC 11845 = DSM 15868
Chryseobacterium gallinarum
Elizabethkingia miricola
Elizabethkingia anophelis
Weeksella virosa DSM 16922
Ornithobacterium rhinotracheale ORT-UMN 88
Wenyingzhuangia fucanilytica
Lutibacter sp. LPB0138
Tenacibaculum dicentrarchi
Polaribacter sp. MED152
Polaribacter vadi
Muricauda lutaonensis
Cellulophaga algicola DSM 14237
Cellulophaga lytica
Zobellia galactanivorans
Maribacter sp. HTCC2170
Winogradskyella sp. PG-2
Formosa sp. Hel1_33_131
Dokdonia sp. MED134
Dokdonia donghaensis DSW-1
Croceibacter atlanticus HTCC2559
Psychroflexus torquis ATCC 700755
Gramella forsetii KT0803
Zunongwangia profunda SM-A87
Capnocytophaga canimorsus Cc5
Capnocytophaga haemolytica
Myroides profundi
Myroides odoratimimus
Flavobacterium columnare ATCC 49512
Flavobacterium psychrophilum
Flavobacterium johnsoniae
Flavobacterium branchiophilum FL-15
0.4

Figure 39: Phylogeny of 39 single copy genes from 84 *Bacteroidete* genomes.
B. yothersi Amsterdam and Brazilian strains, *B. californicus* infected and uninfected strains are marked in red. *B. papayensis* is in pink. Figure credit: Dr. Phuong Le at VIB, Gent, Belgium.

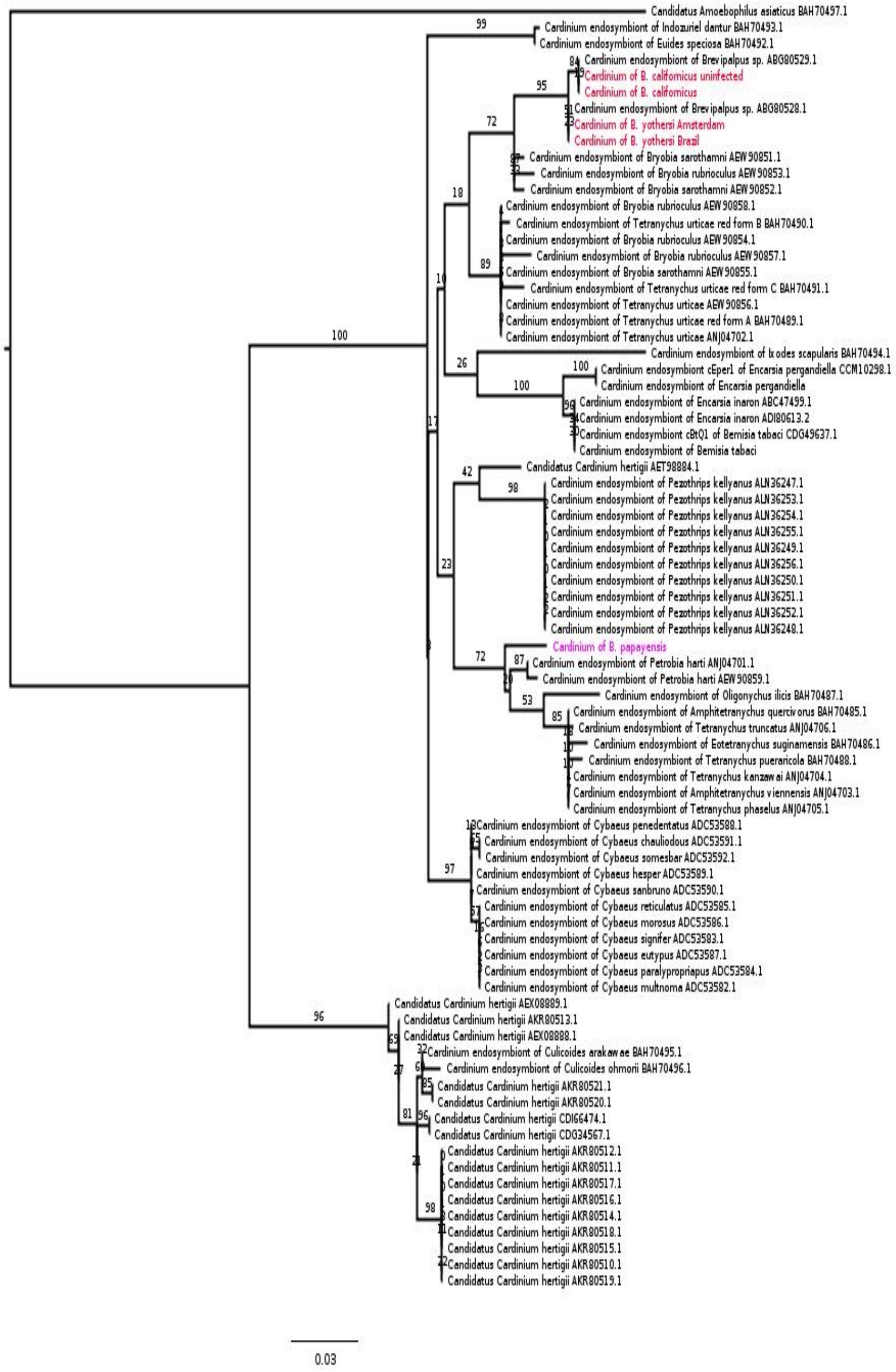

Candidatus Amoebophilus asiaticus BAH70497.1
99
Cardinium endosymbiont of Indozuriel dantur BAH70493.1
Cardinium endosymbiont of Euides speciosa BAH70492.1
84
Cardinium endosymbiont of Brevipalpus sp. ABG80529.1
19
Cardinium of B. californicus uninfected
95
Cardinium of B. californicus
51
Cardinium endosymbiont of Brevipalpus sp. ABG80528.1
72
23
Cardinium of B. yothersi Amsterdam
Cardinium of B. yothersi Brazil
Cardinium endosymbiont of Bryobia sarothamni AEW90851.1
Cardinium endosymbiont of Bryobia rubrioculus AEW90853.1
Cardinium endosymbiont of Bryobia sarothamni AEW90852.1
18
Cardinium endosymbiont of Bryobia rubrioculus AEW90858.1
Cardinium endosymbiont of Tetranychus urticae red form B BAH70490.1
Cardinium endosymbiont of Bryobia rubrioculus AEW90854.1
Cardinium endosymbiont of Bryobia rubrioculus AEW90857.1
89
Cardinium endosymbiont of Bryobia sarothamni AEW90855.1
10
Cardinium endosymbiont of Tetranychus urticae red form C BAH70491.1
Cardinium endosymbiont of Tetranychus urticae AEW90856.1
Cardinium endosymbiont of Tetranychus urticae red form A BAH70489.1
100
Cardinium endosymbiont of Tetranychus urticae ANJ04702.1
Cardinium endosymbiont of Ixodes scapularis BAH70494.1
26
100
Cardinium endosymbiont cEper1 of Encarsia pergandiella CCM10298.1
100
Cardinium endosymbiont of Encarsia pergandiella
96
Cardinium endosymbiont of Encarsia inaron ABC47499.1
34
Cardinium endosymbiont of Encarsia inaron ADI80613.2
17
30
Cardinium endosymbiont cBtQ1 of Bemisia tabaci CDG49637.1
Cardinium endosymbiont of Bemisia tabaci
42
Candidatus Cardinium hertigii AET98884.1
98
Cardinium endosymbiont of Pezothrips kellyanus ALN36247.1
Cardinium endosymbiont of Pezothrips kellyanus ALN36253.1
Cardinium endosymbiont of Pezothrips kellyanus ALN36254.1
Cardinium endosymbiont of Pezothrips kellyanus ALN36255.1
Cardinium endosymbiont of Pezothrips kellyanus ALN36249.1
23
Cardinium endosymbiont of Pezothrips kellyanus ALN36256.1
Cardinium endosymbiont of Pezothrips kellyanus ALN36250.1
Cardinium endosymbiont of Pezothrips kellyanus ALN36251.1
Cardinium endosymbiont of Pezothrips kellyanus ALN36252.1
Cardinium endosymbiont of Pezothrips kellyanus ALN36248.1
Cardinium of B. papayensis
72
87
Cardinium endosymbiont of Petrobia harti ANJ04701.1
20
Cardinium endosymbiont of Petrobia harti AEW90859.1
53
Cardinium endosymbiont of Oligonychus ilicis BAH70487.1
85
Cardinium endosymbiont of Amphitetranychus quercivorus BAH70485.1
Cardinium endosymbiont of Tetranychus truncatus ANJ04706.1
Cardinium endosymbiont of Eotetranychus suginamensis BAH70486.1
Cardinium endosymbiont of Tetranychus pueraricola BAH70488.1
Cardinium endosymbiont of Tetranychus kanzawai ANJ04704.1
Cardinium endosymbiont of Amphitetranychus viennensis ANJ04703.1
Cardinium endosymbiont of Tetranychus phaselus ANJ04705.1
Cardinium endosymbiont of Cybaeus penedentatus ADC53588.1
65
Cardinium endosymbiont of Cybaeus chauliodous ADC53591.1
Cardinium endosymbiont of Cybaeus somesbar ADC53592.1
97
Cardinium endosymbiont of Cybaeus hesper ADC53589.1
Cardinium endosymbiont of Cybaeus sanbruno ADC53590.1
57
Cardinium endosymbiont of Cybaeus reticulatus ADC53585.1
Cardinium endosymbiont of Cybaeus morosus ADC53586.1
Cardinium endosymbiont of Cybaeus signifer ADC53583.1
Cardinium endosymbiont of Cybaeus eutypus ADC53587.1
Cardinium endosymbiont of Cybaeus paralypropriapus ADC53584.1
Cardinium endosymbiont of Cybaeus multnoma ADC53582.1
Candidatus Cardinium hertigii AEX08889.1
96
Candidatus Cardinium hertigii AKR80513.1
65
Candidatus Cardinium hertigii AEX08888.1
32
Cardinium endosymbiont of Culicoides arakawae BAH70495.1
27
61
Cardinium endosymbiont of Culicoides ohmorii BAH70496.1
85
Candidatus Cardinium hertigii AKR80521.1
Candidatus Cardinium hertigii AKR80520.1
81
96
Candidatus Cardinium hertigii CDI66474.1
Candidatus Cardinium hertigii CDG34567.1
Candidatus Cardinium hertigii AKR80512.1
21
Candidatus Cardinium hertigii AKR80511.1
Candidatus Cardinium hertigii AKR80517.1
Candidatus Cardinium hertigii AKR80516.1
98
Candidatus Cardinium hertigii AKR80514.1
11
Candidatus Cardinium hertigii AKR80518.1
Candidatus Cardinium hertigii AKR80515.1
22
Candidatus Cardinium hertigii AKR80510.1
Candidatus Cardinium hertigii AKR80519.1
0.03

Figure 40: Phylogeny of *Cardinium* strains using single copy genes.
B. yothersi Amsterdam and Brazilian strains, *B. californicus* infected and uninfected strains are marked in red. *B. papayensis* is in pink. Figure credit: Dr. Phuong Le at VIB, Gent, Belgium.

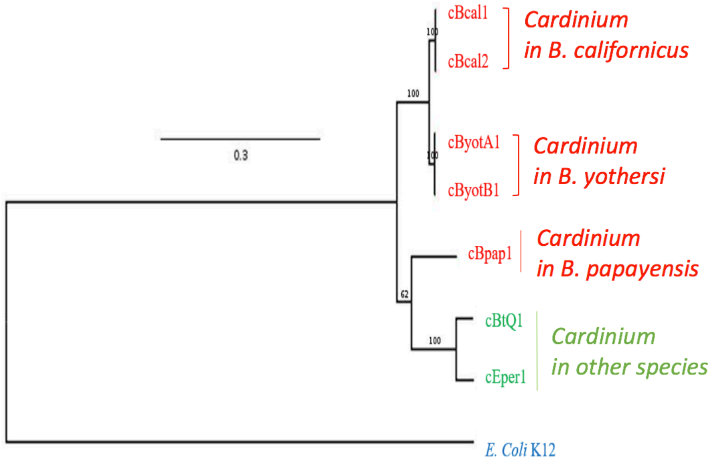
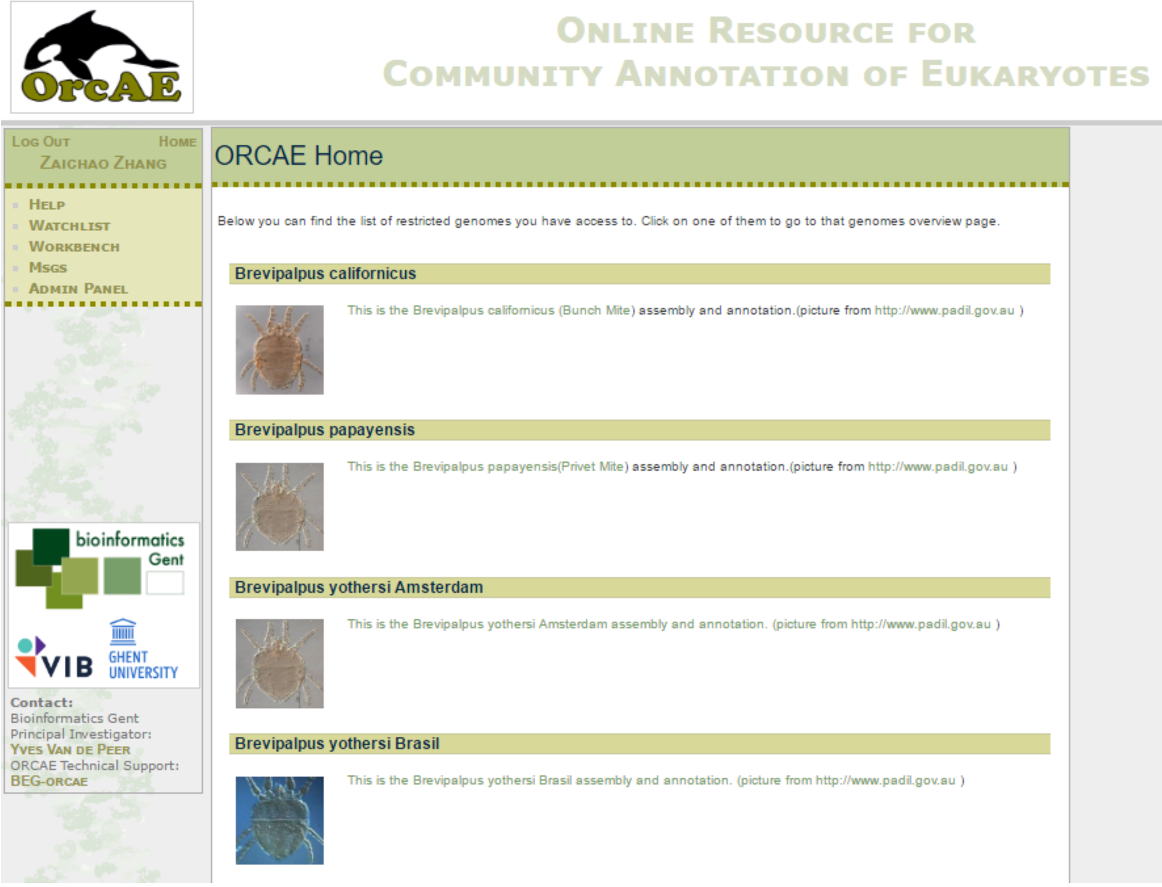


Figure 41: The phylogenetic tree of different *Cardinium* strains of gyrB genes. It suggests a possible scenario of co-evolution of Cardinium genomes with their hosts.

6.2.6 Genomic database visualization

To provide a friendly web-interface and an easy access to these genomes, we prepared all the *Brevipalpus* genomic data on ORCAE database. Multiple functions, such as BLAST and DOWNLOAD, are available for users (Figures 42-48). Here, the visualization for ORCAE is demonstrated using *B. yothersi* (Brazilian strain) gene ID bryot07g00870 at http://bioinformatics.psb.ugent.be/orcae/annotation/Bryot/current/bryot07g00870.



Contact:
Bioinformatics Gent
Principal Investigator:
Yves Van de Peer
ORCAE Technical Support:
BEG-orcae

ORCAE Home

Below you can find the list of restricted genomes you have access to. Click on one of them to go to that genomes overview page.

Brevipalpus californicus

This is the Brevipalpus californicus (Bunch Mite) assembly and annotation.(picture from http://www.padil.gov.au)

Brevipalpus papayensis

This is the Brevipalpus papayensis(Privet Mite) assembly and annotation.(picture from http://www.padil.gov.au)

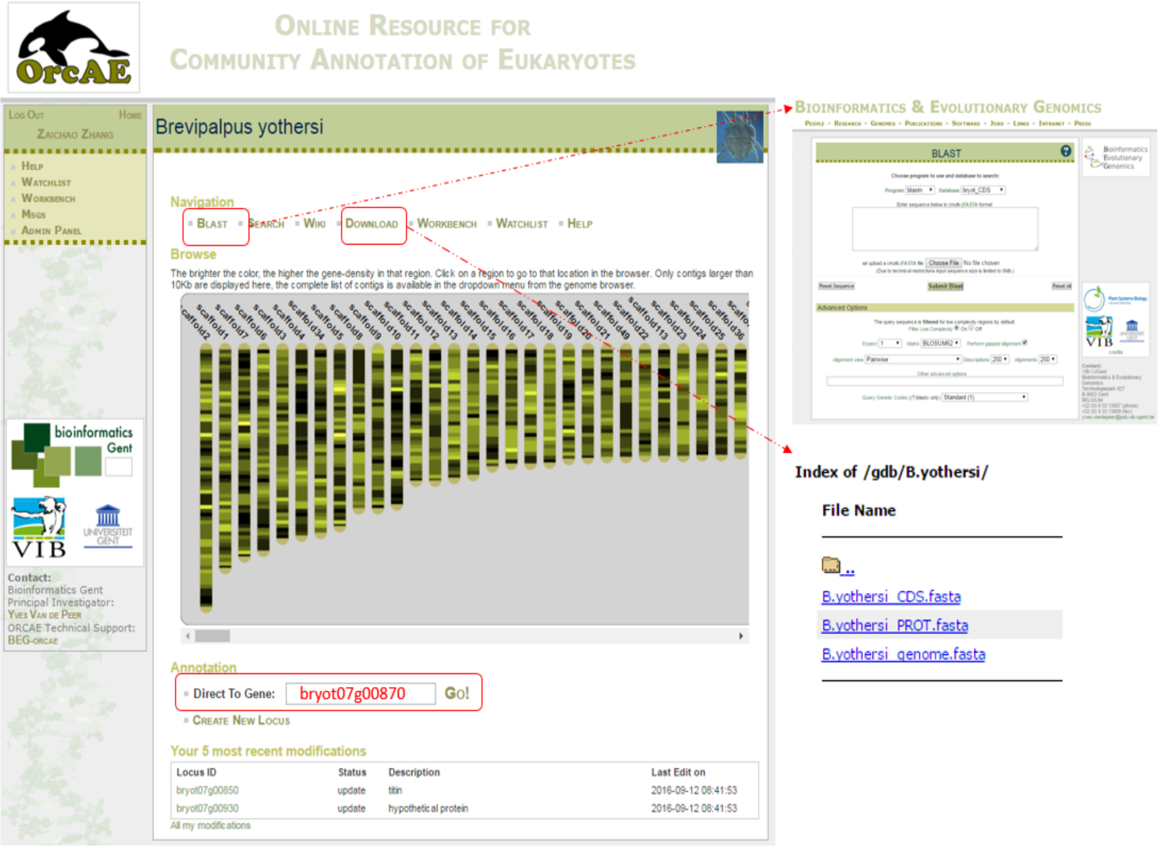
Brevipalpus yothersi Amsterdam

This is the Brevipalpus yothersi Amsterdam assembly and annotation. (picture from http://www.padil.gov.au)

Brevipalpus yothersi Brasil

This is the Brevipalpus yothersi Brasil assembly and annotation. (picture from http://www.padil.gov.au)

Figure 42: ORCAE interface for *Brevipalpus* genomes.



The brighter the color, the higher the gene-density in that region. Click on a region to go to that location in the browser. Only contigs larger than 10Kb are displayed here, the complete list of contigs is available in the dropdown menu from the genome browser.

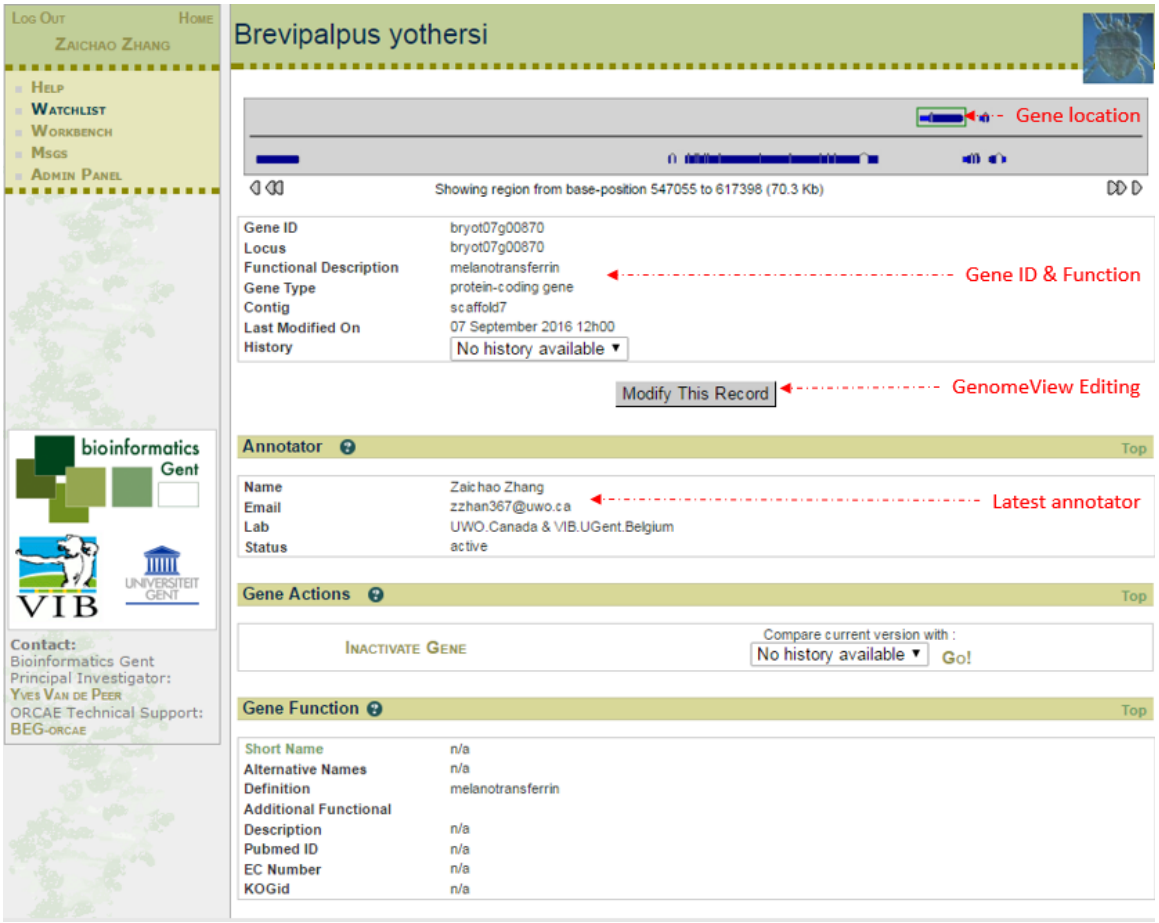
Direct To Gene: bryot07g00870 Go!

Your 5 most recent modifications

Locus ID	Status	Description	Last Edit on
bryot07g00850	update	titin	2016-09-12 08:41:53
bryot07g00930	update	hypothetical protein	2016-09-12 08:41:53

All my modifications

Figure 43: Welcome page for *B. yothersi* (Brazilian strain) on ORCAE.



Brevipalpus yothersi

Gene location

Showing region from base-position 547055 to 617398 (70.3 Kb)

Gene ID	bryot07g00870
Locus	bryot07g00870
Functional Description	melanotransferrin
Gene Type	protein-coding gene
Contig	scaffold7
Last Modified On	07 September 2016 12h00
History	No history available ▾

Gene ID & Function

GenomeView Editing

Annotator

Name	Zaichao Zhang
Email	zzhan367@uwo.ca
Lab	UWO.Canada & VIB.UGent.Belgium
Status	active

Latest annotator

Gene Actions

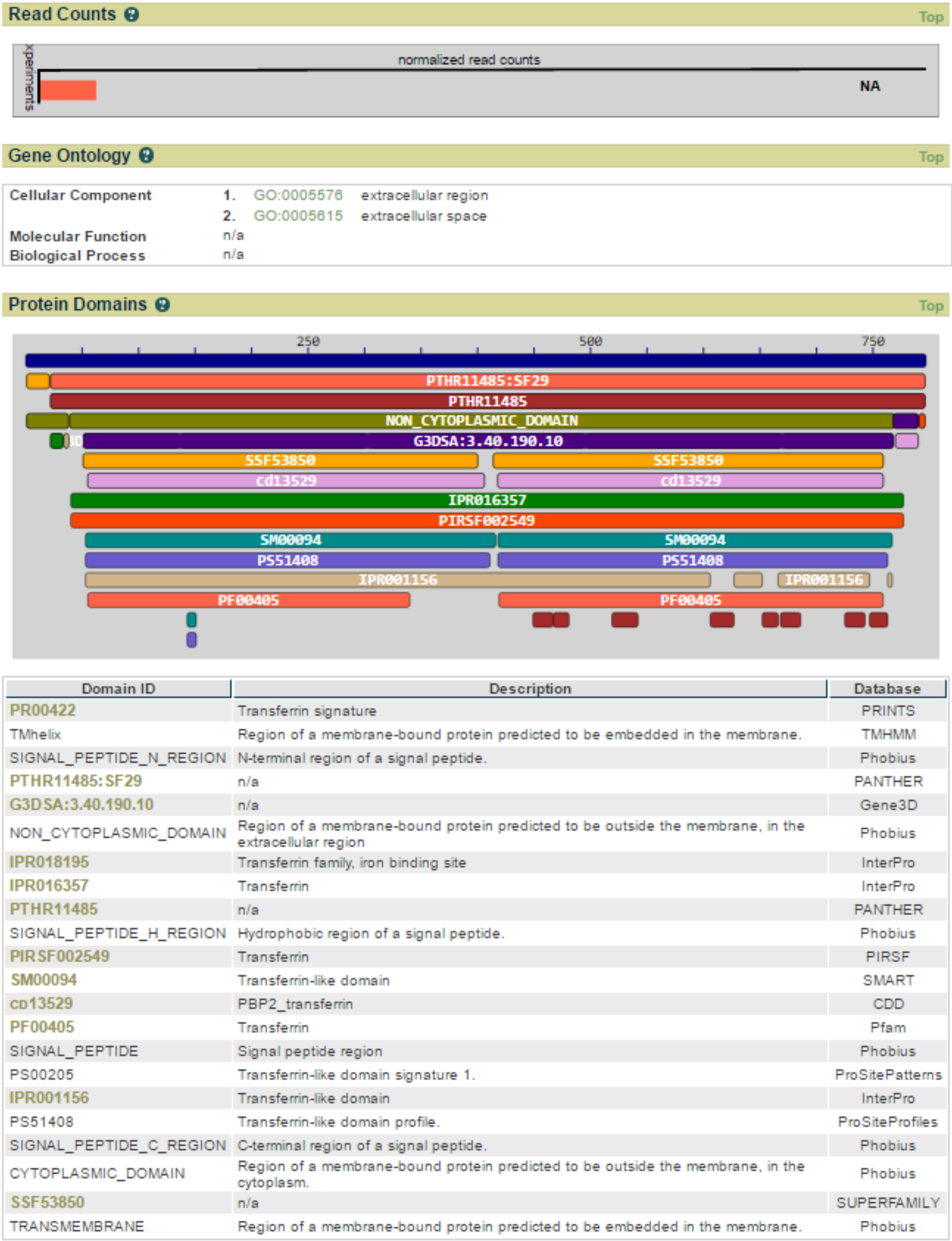
Compare current version with :
No history available ▾

Gene Function

Short Name	n/a
Alternative Names	n/a
Definition	melanotransferrin
Additional Functional Description	n/a
Pubmed ID	n/a
EC Number	n/a
KOGid	n/a

Contact:
Bioinformatics Gent
Principal Investigator:
Yves Van de Peer
ORCAE Technical Support:
BEG-orcae

Figure 44: Gene locus and functional description on ORCAE.



Read Counts

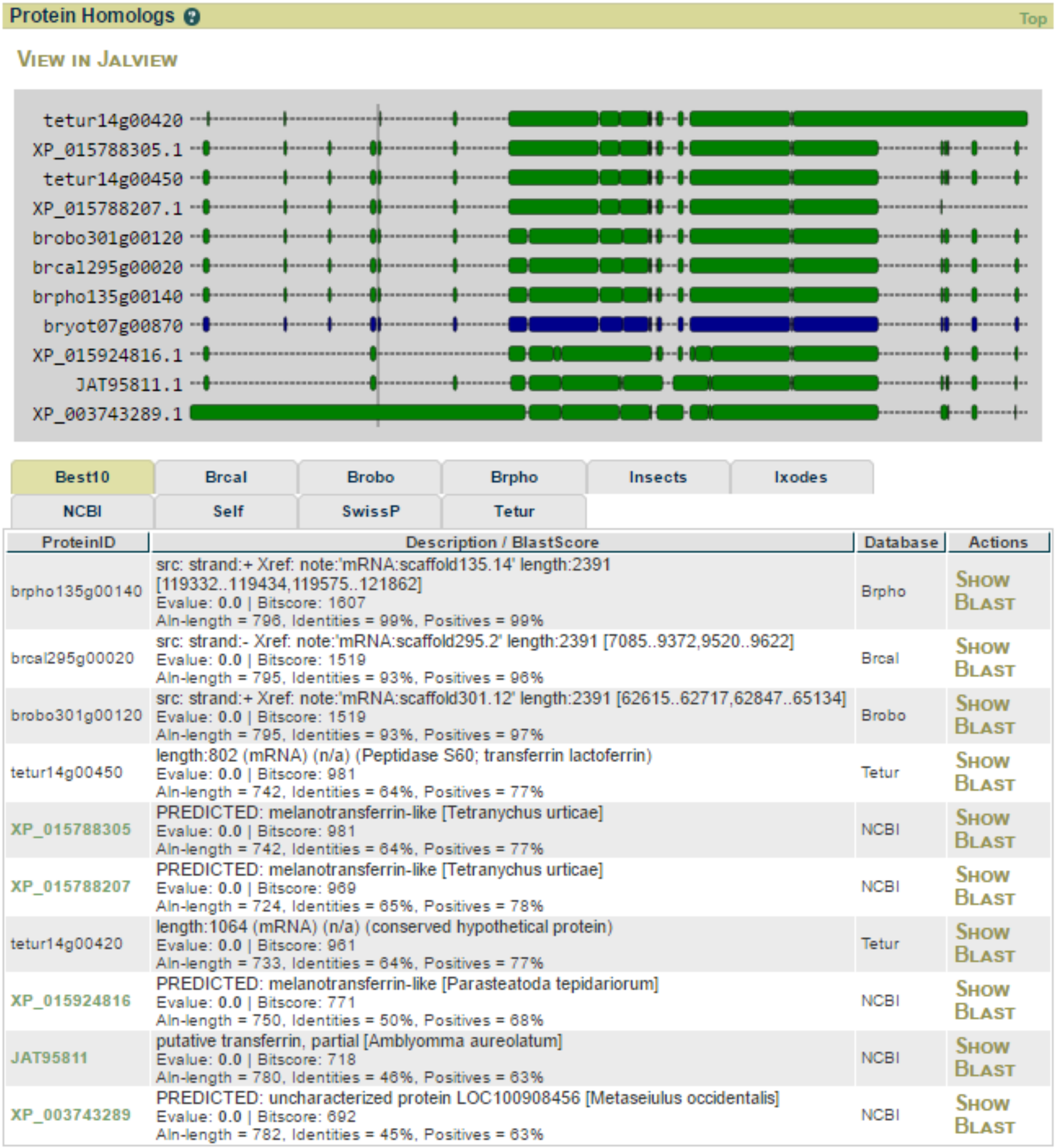
Gene Ontology

Cellular Component	1.	GO:0005576	extracellular region
	2.	GO:0005615	extracellular space
Molecular Function	n/a		
Biological Process	n/a		

Protein Domains

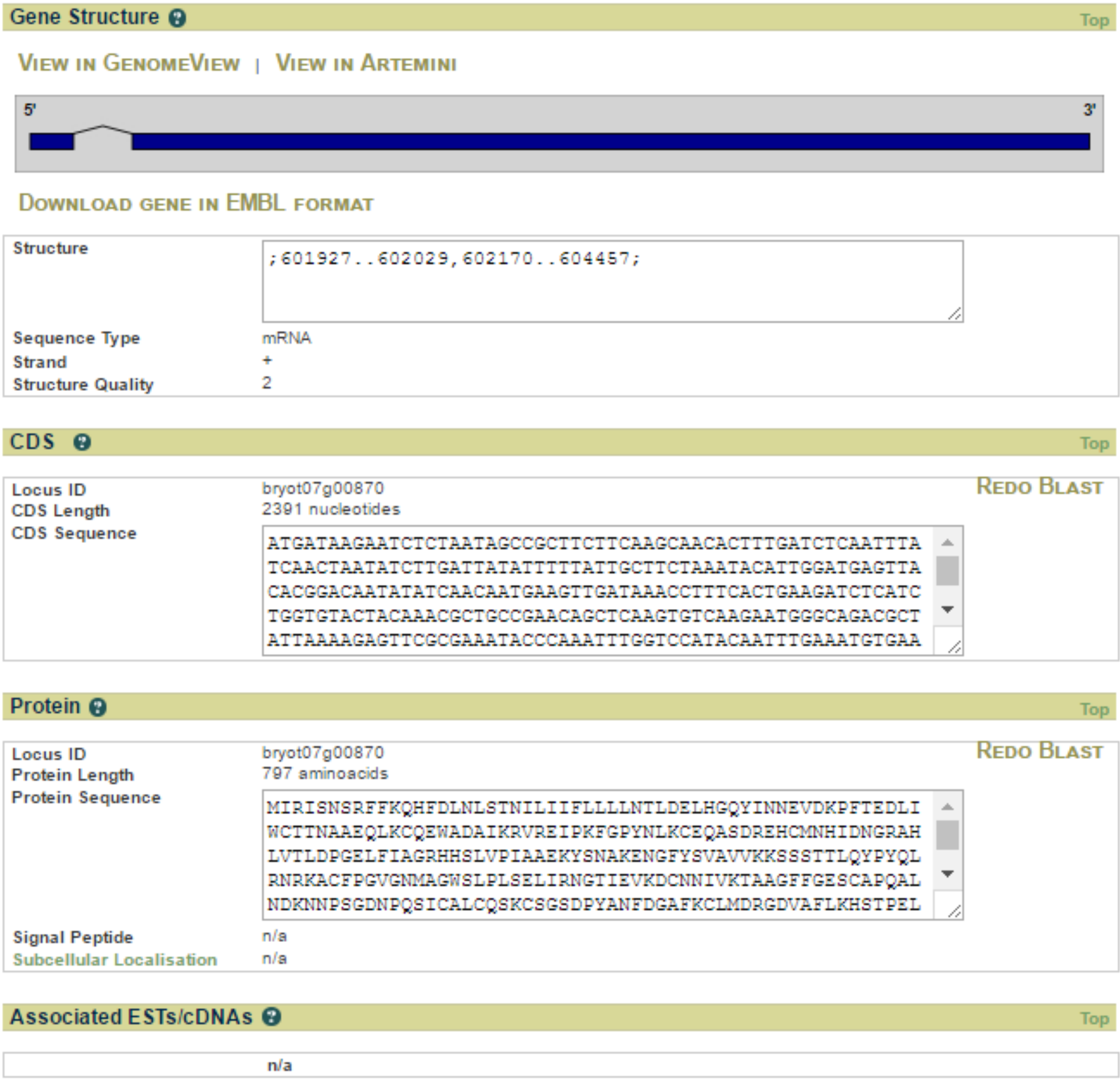
Domain ID	Description	Database
PR00422	Transferrin signature	PRINTS
TMhelix	Region of a membrane-bound protein predicted to be embedded in the membrane.	TMHMM
SIGNAL_PEPTIDE_N_REGION	N-terminal region of a signal peptide.	Phobius
PTHR11485:SF29	n/a	PANTHER
G3DSA:3.40.190.10	n/a	Gene3D
NON_CYTOPLASMIC_DOMAIN	Region of a membrane-bound protein predicted to be outside the membrane, in the extracellular region	Phobius
IPR018195	Transferrin family, iron binding site	InterPro
IPR016357	Transferrin	InterPro
PTHR11485	n/a	PANTHER
SIGNAL_PEPTIDE_H_REGION	Hydrophobic region of a signal peptide.	Phobius
PIRSF002549	Transferrin	PIRSF
SM00094	Transferrin-like domain	SMART
cd13529	PBP2_transferrin	CDD
PF00405	Transferrin	Pfam
SIGNAL_PEPTIDE	Signal peptide region	Phobius
PS00205	Transferrin-like domain signature 1.	ProSitePatterns
IPR001156	Transferrin-like domain	InterPro
PS51408	Transferrin-like domain profile.	ProSiteProfiles
SIGNAL_PEPTIDE_C_REGION	C-terminal region of a signal peptide.	Phobius
CYTOPLASMIC_DOMAIN	Region of a membrane-bound protein predicted to be outside the membrane, in the cytoplasm.	Phobius
SSF53850	n/a	SUPERFAMILY
TRANSMEMBRANE	Region of a membrane-bound protein predicted to be embedded in the membrane.	Phobius

Figure 45: Expression profile, gene ontology, and domain information.



ProteinID	Description / BlastScore	Database	Actions
brpho135g00140	src: strand:+ Xref: note:'mRNA:scaffold135.14' length:2391 [119332..119434,119575..121862] Evalue: **0.0** | Bitscore: 1607 Aln-length = 796, Identities = 99%, Positives = 99%	Brpho	SHOW BLAST
brcal295g00020	src: strand:- Xref: note:'mRNA:scaffold295.2' length:2391 [7085..9372,9520..9622] Evalue: **0.0** | Bitscore: 1519 Aln-length = 795, Identities = 93%, Positives = 96%	Brcal	SHOW BLAST
brobo301g00120	src: strand:+ Xref: note:'mRNA:scaffold301.12' length:2391 [62615..62717,62847..65134] Evalue: **0.0** | Bitscore: 1519 Aln-length = 795, Identities = 93%, Positives = 97%	Brobo	SHOW BLAST
tetur14g00450	length:802 (mRNA) (n/a) (Peptidase S60; transferrin lactoferrin) Evalue: **0.0** | Bitscore: 981 Aln-length = 742, Identities = 64%, Positives = 77%	Tetur	SHOW BLAST
XP_015788305	PREDICTED: melanotransferrin-like [Tetranychus urticae] Evalue: **0.0** | Bitscore: 981 Aln-length = 742, Identities = 64%, Positives = 77%	NCBI	SHOW BLAST
XP_015788207	PREDICTED: melanotransferrin-like [Tetranychus urticae] Evalue: **0.0** | Bitscore: 969 Aln-length = 724, Identities = 65%, Positives = 78%	NCBI	SHOW BLAST
tetur14g00420	length:1064 (mRNA) (n/a) (conserved hypothetical protein) Evalue: **0.0** | Bitscore: 961 Aln-length = 733, Identities = 64%, Positives = 77%	Tetur	SHOW BLAST
XP_015924816	PREDICTED: melanotransferrin-like [Parasteatoda tepidariorum] Evalue: **0.0** | Bitscore: 771 Aln-length = 750, Identities = 50%, Positives = 68%	NCBI	SHOW BLAST
JAT95811	putative transferrin, partial [Amblyomma aureolatum] Evalue: **0.0** | Bitscore: 718 Aln-length = 780, Identities = 46%, Positives = 63%	NCBI	SHOW BLAST
XP_003743289	PREDICTED: uncharacterized protein LOC100908456 [Metaseiulus occidentalis] Evalue: **0.0** | Bitscore: 692 Aln-length = 782, Identities = 45%, Positives = 63%	NCBI	SHOW BLAST

Figure 46: Protein homology and alignment in various databases.



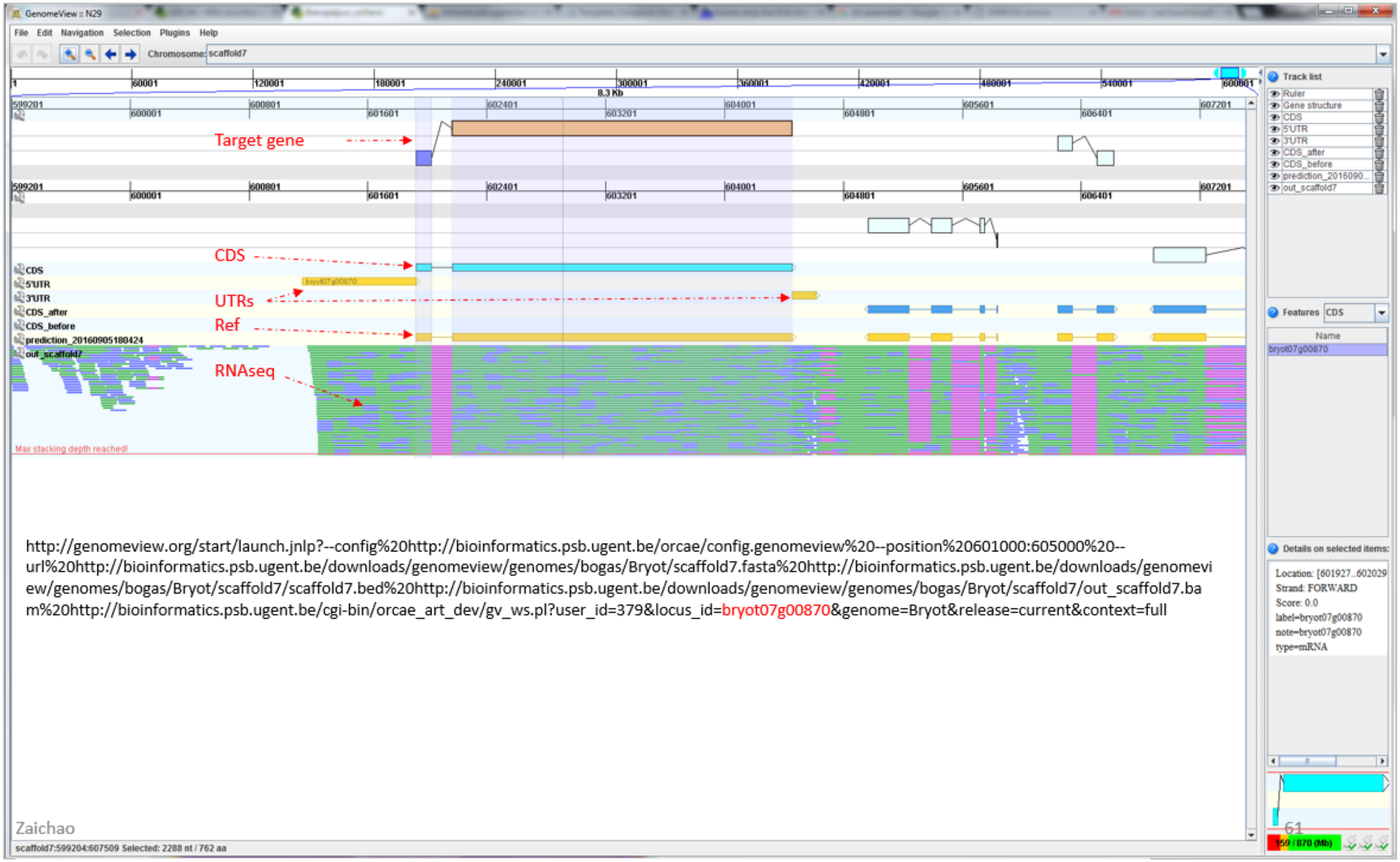
Structure	;601927..602029,602170..604457;
Sequence Type	mRNA
Strand	+
Structure Quality	2

Locus ID	bryot07g00870
CDS Length	2391 nucleotides
CDS Sequence	ATGATAAGAATCTCTAATAGCCGCTTCTTCAAGCAACACTTTGATCTCAATTTA TCAACTAATATCTTGATTATATTTTTATTGCTTCTAAATACATTGGATGAGTTA CACGGACAATATATCAACAATGAAGTTGATAAACCTTTCACTGAAGATCTCATC TGGTGTACTACAAACGCTGCCGAACAGCTCAAGTGTCAAGAATGGGCAGACGCT ATTAAAAGAGTTCGCGAAATACCCAAATTTGGTCCATACAATTTGAAATGTGAA

Locus ID	bryot07g00870
Protein Length	797 aminoacids
Protein Sequence	MIRISNSRFFKQHFDLNLSTNILIIFLLLLNTLDELHGQYINNEVDKPFTEDLI WCTTNAAEQLKCQEWADAIKRVREIPKFGPYNLKCEQASDREHCMNHIDNGRAH LVTLDPGELFIAGRHHSLVPIAAEKYSNAKENGFYSVAVVKKSSSTTLQYPYQL RNRKACFPGVGNMAGWSLPLSELIRNGTIEVKDCNNIVKTAAGFFGESCAPQAL NDKNNPSGDNPQSICALCQSKCSGSDPYANFDGAFKCLMDRGDVAFLKHSTPEL
Signal Peptide	n/a
Subcellular Localisation	n/a

n/a

Figure 47: Gene structure, CDS, and protein sequences.



```
http://genomeview.org/start/launch.jnlp?--config%20http://bioinformatics.psb.ugent.be/orcae/config.genomeview%20--position%20601000:605000%20--url%20http://bioinformatics.psb.ugent.be/downloads/genomeview/genomes/bogas/Bryot/scaffold7.fasta%20http://bioinformatics.psb.ugent.be/downloads/genomeview/genomes/bogas/Bryot/scaffold7/scaffold7.bed%20http://bioinformatics.psb.ugent.be/downloads/genomeview/genomes/bogas/Bryot/scaffold7/out_scaffold7.bam%20http://bioinformatics.psb.ugent.be/cgi-bin/orcae_art_dev/gv_ws.pl?user_id=379&locus_id=bryot07g00870&genome=Bryot&release=current&context=full
```

Location: [601927..602029
Strand: FORWARD
Score: 0.0
label=bryot07g00870
note=bryot07g00870
type=mRNA

scaffold7:599204:607509 Selected: 2288 nt / 762 aa

Figure 48: GenomeView offers various annotated information.

6.3 Conclusion

The flat mites are major agricultural pests feeding primarily on citrus in Central Europe and South America. We sequenced the first *Brevipalpus* genomes and deeply investigated the SNPs, INDELs and genome structure variation, genes associated with feeding and detoxification process and evolutionary scenario. We found glucose dehydrogenase genes are highly expanded in flat mites. These genes participate in pentose phosphate pathway, a metabolic pathway parallel to glycolysis. It generates NADPH and pentoses (5-carbon sugars) as well as ribose 5-phosphate, the last one a precursor for the synthesis of nucleotides. While it does involve oxidation of glucose, its primary role is anabolic rather than catabolic. This might be why flat mites prefer citrus as a sweet citrus adaptation. Additionally, we also sequenced and comparatively analyzed the endosymbiont *Cardinium* genomes across different *Brevipalpus* strains., suggesting they had co-evolution with their hosts. In all, our genomic assemblies and annotations will not only expand the arthropod genetic toolkit but also provide the fundamentals for mite-plant interaction studies as well.

6.4 Materials and Methods

6.4.1 *Brevipalpus* genomes

The *B. yothersi* (Brazilian strain) was collected in Brazil and the other strains were collected in Amsterdam, Netherlands.

Table 24: Sequencing data of the five *Brevipalpus* genomes.

Species	***Reads***	***Assembly (Mb)***	***Scaffolds***	***N50(kb) scaffold***	***L50 scaffold***	***EST coverage***
B. yothersi (Brazil)	454 PE (2x150 bp) MP(2x1.5 kb)	72.2	849	175.1	132	42,130
B. yothersi (Amsterdam)	~130M PE (2x125 bp)	75.5	15,934	47.6	439	40,777
B. californicus	~130M PE (2x125 bp)	68.9	8,971	38.9	488	22,686
B. californicus uninfected	~120M PE (2x125 bp)	67.5	7,221	41.6	448	24,699
B. papayensis	~120M PE (2x125 bp)	66.5	6,977	37.4	510	22,390

6.4.1.1 Genome assembly

The *B. yothersi* (Brazilian strain) genome was sequenced by multiple libraries including Roche 454 GS FLX SE, Illumina MiSeq (SE 250 bp, PE 150 bp, MP 1.5 kb). Briefly, low-quality reads were removed by Trimmomatic and the best K-mer was estimated by the Kmergenic tool (Bolger *et al.*, 2014a; Chikhi and Medvedev, 2014). We initially used Newbler to assemble reads from hybrid data sets with an estimated total coverage of 42x (Zhang *et al.*, 2012). Then we used SSPACE to assemble the hybrid data sets of MP (length 1.5 kb) and draft assembly (Boetzer *et al.*, 2011; Boetzer and Pirovano, 2014). GapFiller was used to fill the gaps from the *de novo* assembly (Nadalin *et al.*, 2012a).

We further sequenced the four other genomes (Amsterdam strains) using Illumina sequencing technology with an average coverage of 230x (Table 24). In short, *B. yothersi* (Amsterdam strain) and *B. papayensis* were both sequenced with an output of 130M PE reads (2x125 bp), respectively. *B. californicus* (both infected and uninfected strain) were sequenced with an output of 120M PE reads (2x125 bp), respectively. We employed Newbler, SSPACE (v3.0) and GapFiller to finish assembling these four genomes (Boetzer *et al.*, 2011; Nadalin *et al.*, 2012a; Zhang *et al.*, 2012).

To validate these five assemblies, we mapped both RNAseq PE reads (using HISAT2) and EST data (using gmap-gsnap) from each genome back to each genome assembly accordingly (Wu and Watanabe, 2005; Wu and Nacu, 2010; Kim *et al.*, 2015). The RNAseq and EST coverages are from 93.4% to 98.7% and from 94.09% to 99.35%, respectively. As such, it suggests they are complete assemblies, instead of drafts.

To verify whether the assembled genome contains contamination from another source of DNA such as bacteria, we divided the assembled scaffolds into 2.5 kb fragments as query sequences and TBLASTX against the RefSeq protein database(https://www.ncbi.nlm.nih.gov/refseq/, e-value < 1e-5 and max target hits 10) (Altschul *et al.*, 1990). If the query sequence had more than 50% of the hits coming from non-arthropod proteins, the scaffold was flagged as potential contamination and was subjected to further manual inspection. Furthermore, the abnormal average sequencing read coverage per scaffold could also be a hint of potential contamination. That is, the average coverage of *B. yothersi* (Brazilian strain) assembly is 42x, if the coverage of the scaffold in question is differ from two times of the standard deviation of the average coverage, this scaffold might be contamination or assembly error. Scaffolds in question were examined manually and subsequently removed in the final assemblies.

Because the *B. yothersi* (Brazilian strain) has the best assembly due to hybrid datasets, we further mapped the genomic reads of five strains to this assembly using BWA and Samtools (Li and Durbin, 2009; Li *et al.*, 2009a). The genomic coverages for *B. yothersi* Brazilian strain and Amsterdam strain are quite close, about 99.2% (self-mapping) and 95.8%, respectively. The other three strains cover 58.3% to 61.2% of *B. yothersi* (Brazilian strain) assembly.

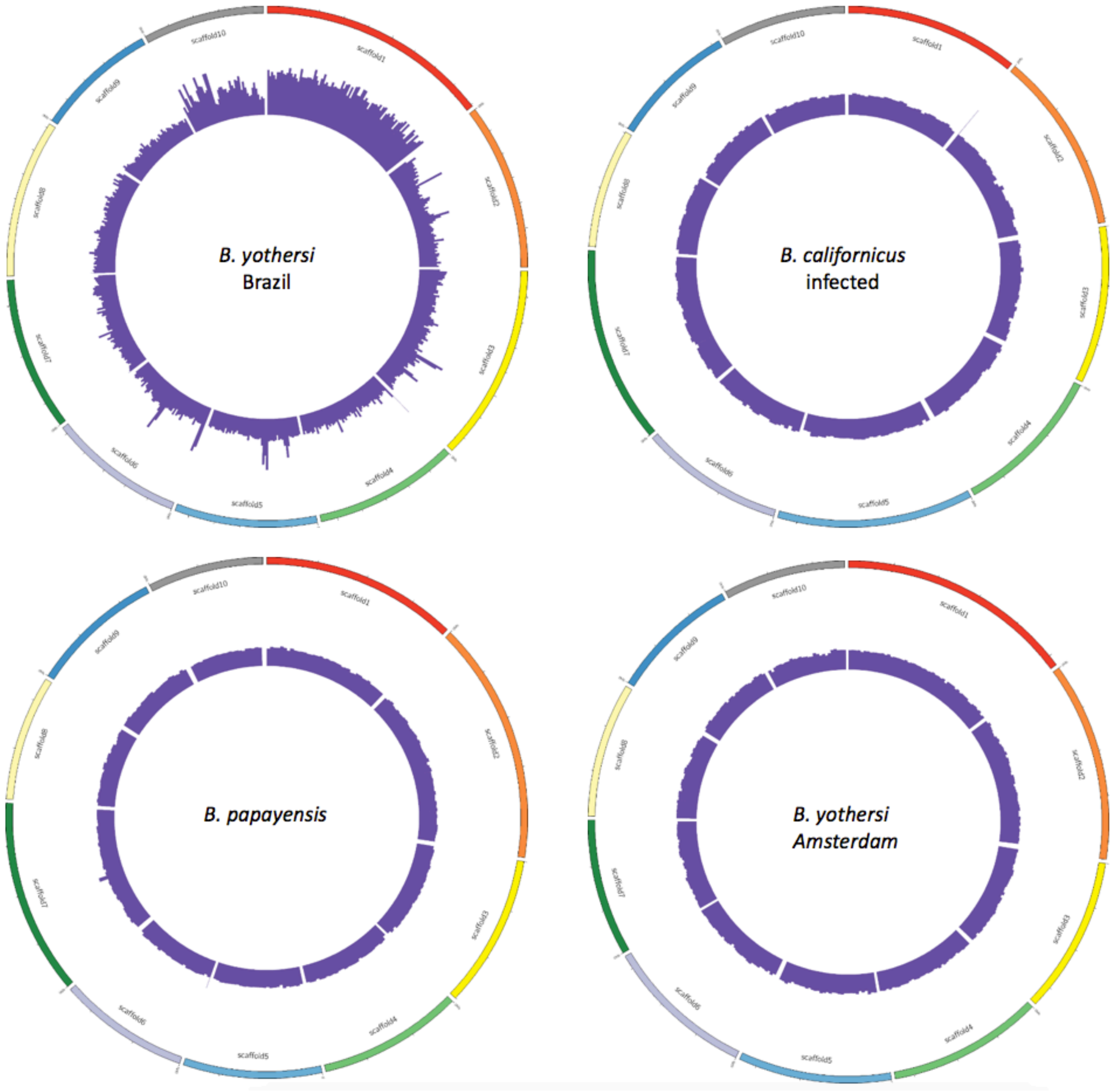


Figure 49: Self-mapping to validate assembly and screen bacterial contamination. We demonstrate four *Brevipalpus* genomic reads coverage using their top 10 scaffolds to check the assembly and contamination. The *B. yothersi* (Brazilian strain) obviously have different coverage on scaffold 1 and scaffold 10, which need to be manually checked. The average coverage of the rest three genomes has few collapsed and gap regions.

6.4.1.2 Genome masking and TE annotation

A customized TE library was created by applying multiple complementary tools and procedures. Initially, we employed RepeatModeler to build up the *de novo* repeat library using *B. yothersi* (Brazilian strain) genome (Smit and Hubley, 2008-2015). We filtered the TE clusters by using gene family clustering method and RepBase (Bao *et al.*, 2015). Meanwhile, all the initial TE from RepeatModeler were also filtered based on their EST overlaps, low score and microsatellites annotations (Smit and Hubley, 2008-2015). Both libraries from the previous two steps were combined and the short sequences (less than 100 nt) were removed. To avoid the homologous sequences, we used CD-HIT to cluster all the consensus sequences (Li and Godzik, 2006; Fu *et al.*, 2012). Later the consensus library was filtered potential protein sequences using UniProt database by BLASTX (Apweiler *et al.*, 2004; Pundir *et al.*, 2016). Finally, we applied RepBase again to classify and assign TE categories and InterProScan to filter sequences with domains (Jones *et al.*, 2014; Bao *et al.*, 2015). The final repeat library is 1,967,002 bp in size and contains 4,414 TE sequences. Five genome assemblies were masked by RepeatMasker using the customized repeat library (Smit and Hubley, 2008-2015).

6.4.1.3 Genome annotation

We initially trained the Splice Machine and also used EUGENE for the structural annotation (Degroeve *et al.*, 2005; Foissac *et al.*, 2008), shown in Figure 50. In short, multiple datasets were used as inputs such as three *Tetranychus* protein databases, both UniProt and Swiss-Prot databases (version 20160822), RNAseq and EST data from *B. yothersi* (Brazilian strain), Insect protein database (version jan2015.49.protein.faa) and *T. urticae* mRNA data (version 20160811) as a reference genome. We then applied InterProScan to predict the domain information for these gene models and finally used our in-house Perl script to assign each gene functional descriptions (Jones *et al.*, 2014). Infernal and Rfam (version 20150608) were used to annotate ncRNA across the five strains (Nawrocki and Eddy, 2013; Nawrocki, 2014; Nawrocki *et al.*, 2015). These genomes are available ORCAE at http://bioinformatics.psb.ugent.be/orcae/.

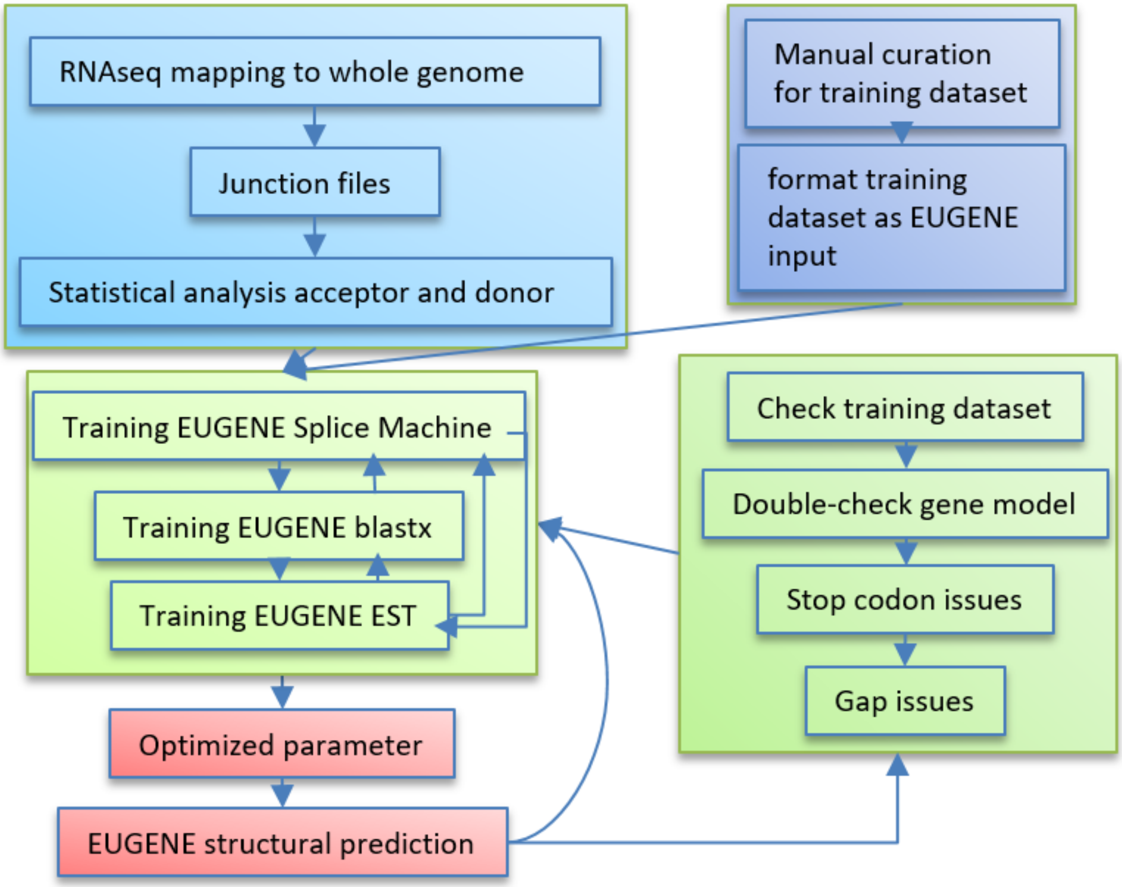

RNAseq mapping to whole genome
Junction files
Statistical analysis acceptor and donor
Manual curation for training dataset
format training dataset as EUGENE input
Training EUGENE Splice Machine
Training EUGENE blastx
Training EUGENE EST
Check training dataset
Double-check gene model
Stop codon issues
Gap issues
Optimized parameter
EUGENE structural prediction

Figure 50: The pipeline of Splice Machine and training EUGENE parameters.

6.4.1.4 SNP calling

To understand the SNP variations across the five strain flat mites, we performed the following analysis using GATK method as the pipeline shown in Figure 51 (McKenna *et al.*, 2010). The *B. yothersi* (Brazilian strain) was used as a reference genome because of its best assembly. The PCA analysis and phylogeny were done using in-house R scripts.

6.4.1.5 Phylogenetic analyses

To construct a phylogenetic tree for the five *Brevipalpus* strains, we collected all the detected SNP sites (exclude INDEL) and randomly 5000 SNP sites. We used the neighbor-joining method with 500 bootstrap replications to run three trials using MEGA (Tamura *et al.*, 2013).

Arthropod genomes and outgroup *Hypsibius dujardini* were retrieved from the following databases: The genomes of *Ixodes scapularis, Dendroctonus ponderosae, Tribolium castaneum, Daphnia pulex , Aedes aegypti, Anopheles darlingi, Anopheles gambiae, Culex quinquefasciatus , Drosophila melanogaster, Acyrthosiphon pisum, Apis mellifera, Atta cephalotes, Nasonia vitripennis, Solenopsis invicta, Bombyx mori, Danaus plexippus, Heliconius melpomene, Strigamia maritima* and *Pediculus humanus* were from ftp://ftp.ensemblgenomes.org on July 24, 2014. The sources of other six genomes are: *Hypsibius dujardini*, http://badger.bio.ed.ac.uk/H_dujardini/home/download, peptide Version 2.3.1, on Feb 29, 2016; *Tetranychus urticae*, http://bioinformatics.psb.ugent.be/, on Feb 29, 2016; *Plutella xylostella*, http://iae.fafu.edu.cn/DBM/download.php, Protein sequences of OGSv1.0 on Feb 29, 2016; *Mesobuthus martensii*, http://lifecenter.sgst.cn/main/en/scorpion.jsp, on March 1, 2016; *Stegodyphus mimosarum*, http://www.ncbi.nlm.nih.gov/protein, on March 1, 2016; *Limulus polyphemus*, http://www.ncbi.nlm.nih.gov/protein/?term=Limulus+polyphemus, on March 1, 2016.

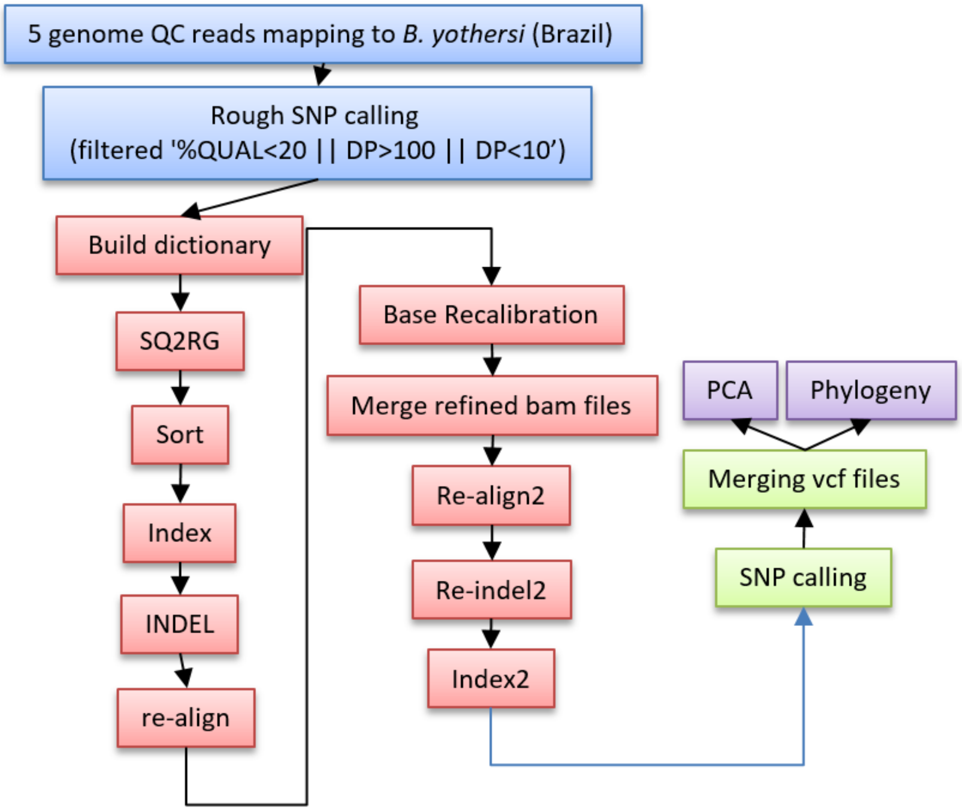


Figure 51: SNP calling pipeline using GATK method across the five flat mite strains. It consists of three major steps: alignments, SNP calling and filtering.

We employed Decypher (TimeLogic® Tera-BLAST™ algorithm, e-value < 1e-10) and OrthoMCL to calculate the homologous genes across 32 genomes (Li *et al.*, 2003). The initial protein-coding gene sequences from 32 genome datasets were concatenated into one FASTA file. To avoid the artifacts of annotation and short proteins, we filtered the poor protein sequences by OrthoMCL (orthomclFilterFasta, cut-off: min_length 10, max_percent_stops 20) and formatted the good protein sequences in as Decypher subject database (Li *et al.*, 2003). Then we performed all-against-all BLASTP approach as search method to calculate the homologous genes. In this way, we collected all the single copy genes and built up the arthropod phylogeny using MEGA with 500 bootstrap replicates and NJ partial deletion site coverage cutoff 90% (Tamura *et al.*, 2013). The arthropod divergence times are based on molecular estimates, as described in (Misof *et al.*, 2014).

6.4.2 *Cardinium* genomes

6.4.2.1 Nomenclature of *Cardinium* strains

We adopted the nomenclature of *Cardinium* cEper1 for the *Cardinium* strains in *Brevipalpus* (Penz et al., 2012). The genome strain is cByotB1, where 'c' refers to *Cardinium*, 'Byot' refers to the host *B. yothersi*, 'B' refers to Brazil strain, and '1' simply denotes the first named strain from this host. The same rule applies to others *Cardinium* strains: cByotA1 refers to *Cardinium* in *B. yothersi* str. Amsterdam, cBcal1 refers to *Cardinium* in *B. californicus*, cBcal2 refers to *Cardinium* in *B. californicus* (uninfected strain treated by Tetracycline), and cBpap1 refers to *Cardinium* in *B. papayensis*.

6.4.2.2 Genome identification and assembly

Among the potentially contaminated genome assemblies, two scaffolds from *B. yothersi* str. Brazil was confirmed derived from *Cardinium*. We retrieved reads originated from two scaffolds and mapped to two published genomes including *Cardinium* cBtQ1 and cPer1 using BWA (Groot and Breeuwer, 2006; Li and Durbin, 2009; Penz *et al.*, 2012). It confirmed the *Brevipalpus* infected bacteria were *Cardinium*. We then used the cByotB as a reference to retrieve the potential *Cardinium* sequences from four other *Brevipalpus* genomes. The reads coming from contaminated scaffolds of four other *Brevipalpus* were used to align back to cByotB1, and the outputs were assembled de novo applying CLC Genomics Workbench (v 4.4.2, default parameters). The completeness of the assembled contigs was verified by employing BUSCO (Simao *et al.*, 2015). All *Cardinium* endosymbiont species have around 60% of core genes, which is as same as two other cEPer1 and cBQt1. This may be because BUSCO is designed preferably for insect genomes instead of bacterial genomes.

These fished reads were assembled by clc_assembly_cell (version x86_64/4.4.2) with default parameters (https://www.qiagenbioinformatics.com/products/clc-assembly-cell/). To compare the *Cardinium* assemblies in the five *Brevipalpus* with two other *Cardinium* genomes: *Cardinium* cBtQ1, a facultative bacterial endosymbiont of *Bemisia tabaci* (silver leaf whitefly) and *Cardinium* cEper1, endosymbionts of amoeba and wasps (Penz *et al.*, 2012; Santos-Garcia *et al.*, 2014), we employed MAUVE to investigate and visualize the synteny across the three strains (Darling *et al.*, 2004). Meanwhile, we mapped the raw genomic reads of *Cardinium* in Brazilian strain to cBtQ1 and cEper1 using BWA and Samtools (Li and Durbin, 2009; Li *et al.*, 2009a) and consequently, visualized these mapping data using Bedtools and Circos (Krzywinski *et al.*, 2009; Quinlan and Hall, 2010).

6.4.2.3 Synteny comparison

The five *Cardinium* genome sequences with contigs in decreasing order were used to compute nucleotide synteny blocks with progressive Mauve aligner (Darling *et al.*, 2004). The cByotB1 strain was set as the reference due to its best assembly, and the alignment was plotted with the genoPlotR package (Guy *et al.*, 2010).

6.4.2.4 Genome annotation and phylogeny analysis

The *Cardinium* genomes were annotated using RAST platform (Aziz *et al.*, 2008). The automatic annotation of CDS was further refined by BLASTP against NRprot database using an E-value of 1e-3, a minimum amino acid identity of 30%, and minimum alignment overlap of 30% as a threshold.

The tRNA genes were annotated using tRNAscan-SE v1.31 (Lowe and Eddy, 1997). All other features were searched for using Infernal v1.1 against Rfam V12 (Nawrocki *et al.*, 2015). All hits with E-value <1e-3 were considered and manually curated. Protein domains were predicted using Pfam and SMART (Schultz *et al.*, 2000; Bateman *et al.*, 2002).

The amino acid sequences of the gyrB gene of *Cardinium* species, proteomes of Bacteroidetes, and three non-Bacteroidetes genomes (*E. coli* str. K12 MG1655, *Alteromonas confluentis*, and *Caulobacter vibriodes*) were downloaded from NCBI database (Jan 2017). The gene families of Bacteroidetes were built using OrthoMCL (proteins superior to 20 amino acids, BLASTP with E-value < 1e-5, inflation value 1.5), aligned with MUSCLE (Edgar, 2004). Protest3 gave the gamma-distributed rates across sites (JTT+G) as the best evolutionary model for gyrB gene and the single core gene of Bacteroidetes (Abascal *et al.*, 2005). The latest COG database was used to assign ORFs to functions applying BLASTP (E-value < 1e-5 and identity > 70%) (Galperin *et al.*, 2015).

6.4.2.5 Orthologous gene identification

The five *Cardinium* genomes in this study and two published genomes (cEper1 and cBtQ1) were used for orthologous gene identification. OrthoMCL (amino acid > 20, Inflation value 1.5) and COG profile assignment were run (BLASTP with E-value 1e-5, identity 70%). Gene clusters may contain zero, one, two, or more genes in each genome. Orthologous genes across seven *Cardinium* were classified as core genes (shared by seven organisms), dispensable genes (shared by two or three organisms) and unique genes (strain-specific). The Venn diagram was drawn by the online tool at http://bioinformatics.psb.ugent.be/webtools/Venn/.

6.5 Supplementary Information

Table 25: The assessment of *Cardinium* genome assemblies.

	Complete and single-copy (%)	Complete and duplicated (%)	Fragmented (%)	Missing (%)
cByotB1	57.4	0	4.7	37.9
cByotA1	57.4	0	4.7	37.9
cBcal2	60.1	0	3.4	36.5
cBcal1	60.8	0	3.4	35.8
cBpap1	52.7	0	2.7	44.6
cEper1	54.7	0	5.4	39.9
cBtQ1	54.8	0	4.7	40.6

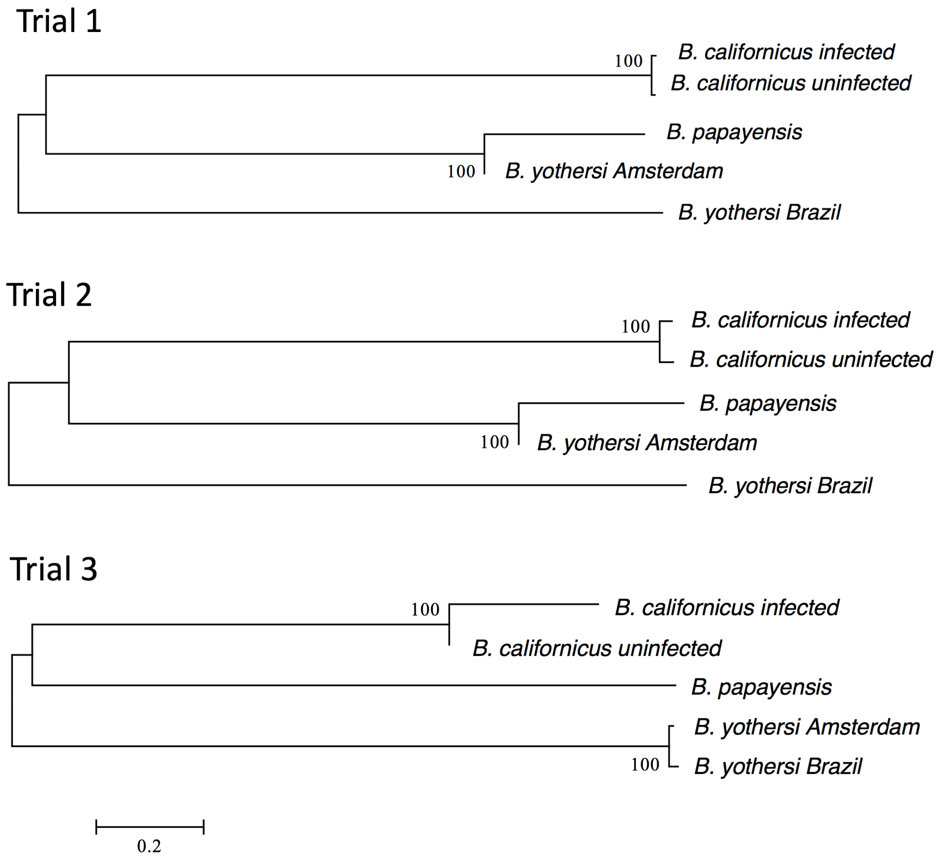


Figure 52: Phylogenies of the five genomes using 5k random SNP sites.

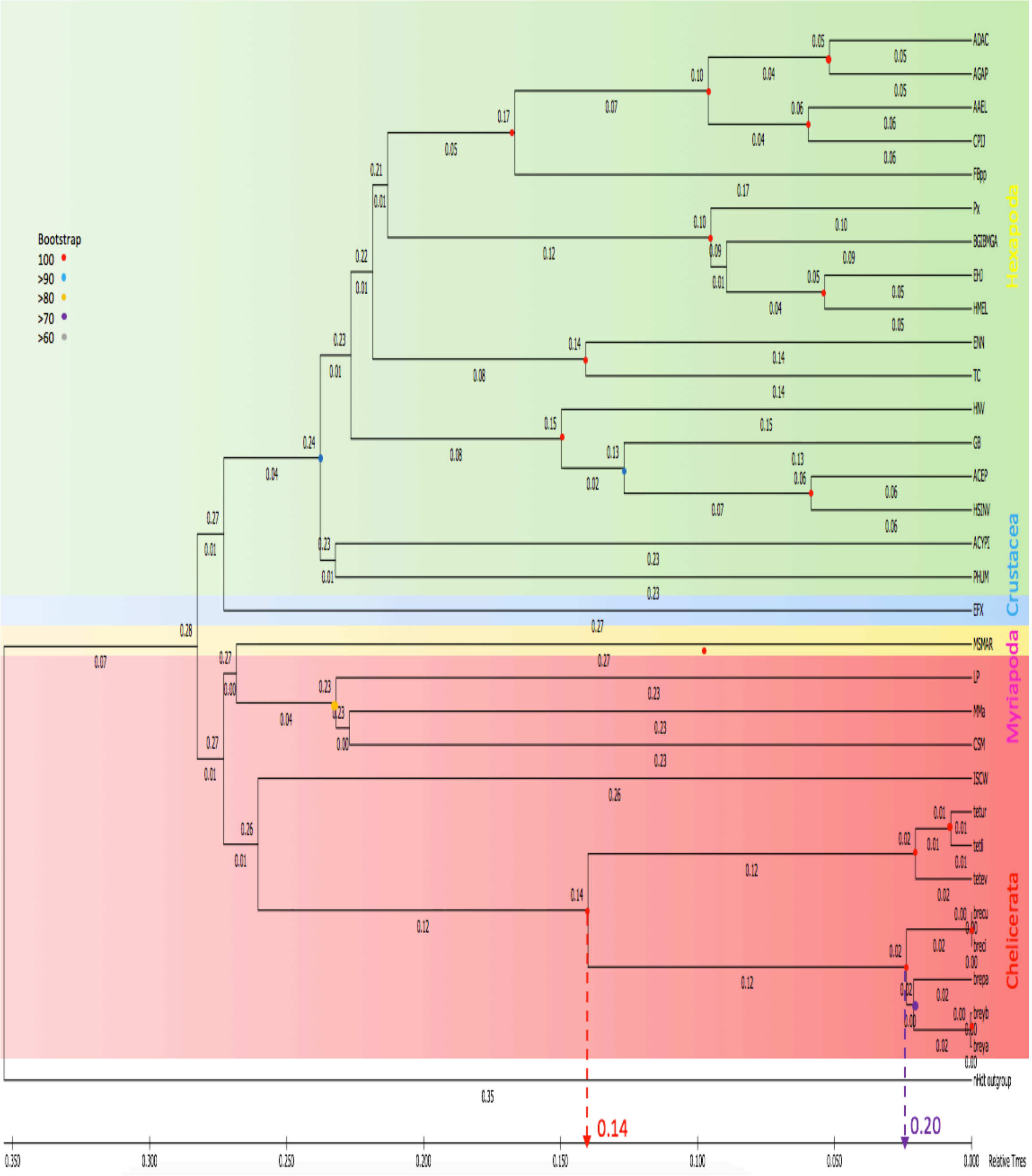


Figure 53: Divergent time estimation.

The red spot is the divergent time (140 MYA) of *Brevipalpus* and *Tetranychus*; the Purple spot is the divergent time (20 MYA) of *Brevipalpus* strains. Abbreviations are on page xxii.

6.6 Reference

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. Bioinformatics *21*, 2104-2105.

Adams, M.J., Lefkowitz, E.J., King, A.M., Bamford, D.H., Breitbart, M., Davison, A.J., Ghabrial, S.A., Gorbalenya, A.E., Knowles, N.J., Krell, P.*, et al.* (2015). Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2015). Arch Virol *160*, 1837-1850.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology *215*, 403-410.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M.*, et al.* (2004). UniProt: the Universal Protein knowledgebase. Nucleic acids research *32*, D115-119.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M.*, et al.* (2008). The RAST Server: Rapid Annotations using Subsystems Technology. BMC genomics *9*, 75.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA *6*, 11.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. (2002). The Pfam protein families database. Nucleic acids research *30*, 276-280.

Beard, J.J., Ochoa, R., Braswell, W.E., and Bauchan, G.R. (2015). Brevipalpus phoenicis (Geijskes) species complex (Acari: Tenuipalpidae)--a closer look. Zootaxa *3944*, 1-67.

Blagrove, M.S., Arias-Goeta, C., Failloux, A.B., and Sinkins, S.P. (2012). Wolbachia strain wMel induces cytoplasmic incompatibility and blocks dengue transmission in Aedes albopictus. Proceedings of the National Academy of Sciences of the United States of America *109*, 255-260.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578-579.

Boetzer, M., and Pirovano, W. (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC bioinformatics *15*, 211.

Bolger, A.M., Lohse, M., and Usadel, B. (2014a). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.

Chigira, A., and Miura, K. (2005). Detection of 'candidatus Cardinium' bacteria from the haploid host Brevipalpus californicus (Acari: Tenuipalpidae) and effect on the host. Experimental & applied acarology *37*, 107-116.

Chikhi, R., and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. Bioinformatics *30*, 31-37.

Childers, C.C., and Derrick, K.S. (2003). Brevipalpus mites as vectors of unassigned rhabdoviruses in various crops. Experimental & applied acarology *30*, 1-3.

Childers, C.C., French, J.V., and Rodrigues, J.C. (2003). Brevipalpus californicus, B. obovatus, B. phoenicis, and B. lewisi (Acari: Tenuipalpidae): a review of their biology, feeding injury and economic importance. Experimental & applied acarology *30*, 5-28.

Childers, C.C., and Rodrigues, J.C.V. (2011). An overview of Brevipalpus mites (Acari: Tenuipalpidae) and the plant viruses they transmit. Zoosymposia *6*, 180-192.

Cooley, J., Takayama, T.K., Shapiro, S.D., Schechter, N.M., and Remold-O'Donnell, E. (2001). The serpin MNEI inhibits elastase-like and chymotrypsin-like serine proteases through efficient reactions at two active sites. Biochemistry *40*, 15762-15770.

Darling, A.C., Mau, B., Blattner, F.R., and Perna, N.T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome research *14*, 1394-1403.

De Carvalho Mineiro, J.L., Sato, M.E., Raga, A., and Arthur, V. (2008). Population dynamics of phytophagous and predaceous mites on coffee in Brazil, with emphasis on Brevipalpus phoenicis (Acari: Tenuipalpidae). Experimental & applied acarology *44*, 277-291.

Degroeve, S., Saeys, Y., De Baets, B., Rouze, P., and Van de Peer, Y. (2005). SpliceMachine: predicting splice sites from high-dimensional local context representations. Bioinformatics *21*, 1332-1338.

Dunlop, J.A. (2010). Geological history and phylogeny of Chelicerata. Arthropod Struct Dev *39*, 124-142.

Dunlop, J.A., and Selden, P.A. (2009). Calibrating the chelicerate clock: a paleontological reply to Jeyaprakash and Hoy. Experimental & applied acarology *48*, 183-197.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research *32*, 1792-1797.

Foissac, S., Gouzy, J., Rombauts, S., Mathe, C., Amselem, J., Sterck, L., Van de Peer, Y., Rouze, P., and Schiex, T. (2008). Genome annotation in plants and fungi: EuGene as a model platform. Curr Bioinform *3*, 87-97.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150-3152.

Galperin, M.Y., Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic acids research *43*, D261-D269.

Grbic, M., Van Leeuwen, T., Clark, R.M., Rombauts, S., Rouze, P., Grbic, V., Osborne, E.J., Dermauw, W., Ngoc, P.C., Ortego, F.*, et al.* (2011a). The genome of Tetranychus urticae reveals herbivorous pest adaptations. Nature *479*, 487-492.

Groot, T.V., and Breeuwer, J.A. (2006). Cardinium symbionts induce haploid thelytoky in most clones of three closely related Brevipalpus species. Experimental & applied acarology *39*, 257-271.

Groot, T.V., Janssen, A., Pallini, A., and Breeuwer, J.A. (2005). Adaptation in the asexual false spider mite Brevipalpus phoenicis: evidence for frozen niche variation. Experimental & applied acarology *36*, 165-176.

Gulia-Nuss, M., Nuss, A.B., Meyer, J.M., Sonenshine, D.E., Roe, R.M., Waterhouse, R.M., Sattelle, D.B., de la Fuente, J., Ribeiro, J.M., Megy, K.*, et al.* (2016). Genomic insights into the Ixodes scapularis tick vector of Lyme disease. Nature communications *7*, 10507.

Guy, L., Roat Kultima, J., and Andersson, S.G.E. (2010). genoPlotR: comparative gene and genome visualization in R. Bioinformatics *26*, 2334-2335.

Jeyaprakash, A., and Hoy, M.A. (2009). First divergence time estimate of spiders, scorpions, mites and ticks (subphylum: Chelicerata) inferred from mitochondrial phylogeny. Experimental & applied acarology *47*, 1-18.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G.*, et al.* (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236-1240.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nature methods *12*, 357-360.

Kitajima, E.W., Chagas, C.M., and Rodrigues, J.C. (2003). Brevipalpus-transmitted plant virus and virus-like diseases: cytopathology and some recent cases. Experimental & applied acarology *30*, 135-160.

Kitajima, E.W., Groot, T.V., Novelli, V.M., Freitas-Astua, J., Alberti, G., and de Moraes, G.J. (2007). In situ observation of the Cardinium symbionts of Brevipalpus (Acari: Tenuipalpidae) by electron microscopy. Experimental & applied acarology *42*, 263-271.

Kondo H, M.T., Shirako Y, Tamada T (2006). Orchid fleck virus is a rhabdovirus with an unusual bipartite genome. Journal of General Virology *Aug 87(Pt 8):2413-21*.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome research *19*, 1639-1645.

Lal, L. (1979). Biology of Brevipalpus phoenicis (Geijskes) (Tenuipalpidae: Acarina). Acarologia *20*, 97-101.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009a). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research *13*, 2178-2189.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658-1659.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic acids research *25*, 955-964.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.*, et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research *20*, 1297-1303.

Mills, R.E., Bennett, E.A., Iskow, R.C., and Devine, S.E. (2007a). Which transposable elements are active in the human genome? Trends in genetics : TIG *23*, 183-191.

Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G.*, et al.* (2014). Phylogenomics resolves the timing and pattern of insect evolution. Science *346*, 763-767.

Nadalin, F., Vezzi, F., and Policriti, A. (2012a). GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC bioinformatics *13*, S8.

Nakamura, Y., Kawai, S., Yukuhiro, F., Ito, S., Gotoh, T., Kisimoto, R., Yanase, T., Matsumoto, Y., Kageyama, D., and Noda, H. (2009). Prevalence of Cardinium bacteria in planthoppers and spider mites and taxonomic revision of "Candidatus Cardinium hertigii" based on detection of a new Cardinium group from biting midges. Applied and environmental microbiology *75*, 6757-6763.

Nawrocki, E.P. (2014). Annotating functional RNAs in genomes using Infernal. Methods in molecular biology *1097*, 163-197.

Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J., *et al.* (2015). Rfam 12.0: updates to the RNA families database. Nucleic acids research *43*, D130-137.

Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933-2935.

Penz, T., Schmitz-Esser, S., Kelly, S.E., Cass, B.N., Muller, A., Woyke, T., Malfatti, S.A., Hunter, M.S., and Horn, M. (2012). Comparative genomics suggests an independent origin of cytoplasmic incompatibility in Cardinium hertigii. PLoS genetics *8*, e1003012.

Pundir, S., Martin, M.J., O'Donovan, C., and UniProt, C. (2016). UniProt Tools. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] *53*, 1 29 21-15.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Rodrigues, J.C., and Childers, C.C. (2013). Brevipalpus mites (Acari: Tenuipalpidae): vectors of invasive, non-systemic cytoplasmic and nuclear viruses in plants. Experimental & applied acarology *59*, 165-175.

Rodrigues, J.C., Kitajima, E.W., Childers, C.C., and Chagas, C.M. (2003). Citrus leprosis virus vectored by Brevipalpus phoenicis (Acari: Tenuipalpidae) on citrus in Brazil. Experimental & applied acarology *30*, 161-179.

Ros, V.I., Fleming, V.M., Feil, E.J., and Breeuwer, J.A. (2012). Diversity and recombination in Wolbachia and Cardinium from Bryobia spider mites. BMC microbiology *12 Suppl 1*, S13.

Salinas-Vargas, D., Santillan-Galicia, M.T., Guzman-Franco, A.W., Hernandez-Lopez, A., Ortega-Arenas, L.D., and Mora-Aguilera, G. (2016). Analysis of Genetic Variation in Brevipalpus yothersi (Acari: Tenuipalpidae) Populations from Four Species of Citrus Host Plants. PloS one *11*, e0164552.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z.*, et al.* (1996). Nested retrotransposons in the intergenic regions of the maize genome. Science *274*, 765-768.

Santos-Garcia, D., Rollat-Farnier, P.A., Beitia, F., Zchori-Fein, E., Vavre, F., Mouton, L., Moya, A., Latorre, A., and Silva, F.J. (2014). The genome of Cardinium cBtQ1 provides insights into genome reduction, symbiont motility, and its settlement in Bemisia tabaci. Genome biology and evolution *6*, 1013-1030.

Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., and Bork, P. (2000). SMART: a web-based tool for the study of genetically mobile domains. Nucleic acids research *28*, 231-234.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210-3212.

Smit, A., and Hubley, R. (2008-2015). RepeatModeler Open-1.0.

Sun, L.V., Foster, J.M., Tzertzinis, G., Ono, M., Bandi, C., Slatko, B.E., and O'Neill, S.L. (2001). Determination of Wolbachia genome size by pulsed-field gel electrophoresis. Journal of bacteriology *183*, 2219-2225.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular biology and evolution *30*, 2725-2729.

Turnpenny, P.D., and Ellard, S. (2012). Emery's elements of medical genetics (Philadelphia, PA : Elsevier/Churchill Livingstone).

Walker, T., Johnson, P.H., Moreira, L.A., Iturbe-Ormaetxe, I., Frentiu, F.D., McMeniman, C.J., Leong, Y.S., Dong, Y., Axford, J., Kriesner, P.*, et al.* (2011). The wMel Wolbachia strain blocks dengue and invades caged Aedes aegypti populations. Nature *476*, 450-453.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873-881.

Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics *21*, 1859-1875.

Zabalou, S., Riegler, M., Theodorakopoulou, M., Stauffer, C., Savakis, C., and Bourtzis, K. (2004). Wolbachia-induced cytoplasmic incompatibility as a means for insect pest population control. Proceedings of the National Academy of Sciences of the United States of America *101*, 15042-15045.

Zhang, K., Xie, R., and Hong, X.Y. (2010). Research progress of endosymbiont Cardinium. Journal of Naming Agricultural University *2010-05*.

Zhang, T., Luo, Y., Chen, Y., Li, X., and Yu, J. (2012). BIGrat: a repeat resolver for pyrosequencing-based re-sequencing with Newbler. BMC research notes *5*, 567.

Chapter 7

7 Genome annotation of DNMT gene families in parasitoid wasp *Copidosoma floridanum*

Copidosoma floridanum is a wasp parasitoid of moths. *C. floridanum* exhibits polyembryonic development producing over 2,000 individuals from a single egg, as well as two distinct larval castes – reproductive and precocious (or soldier) larvae. As a cosmopolitan species, *C. floridanum* is distributed worldwide. It is of great significance to pest control strategy development as well as the phylogenic relationship with other important insects. However, little is known about the evolutionary dynamics and molecular mechanisms behind its developmental novelties. Here we sequenced the first *C. floridanum* genome to investigate the polyembryony in this parasitoid wasp. Given that DNA methylation plays a key role in wasp casting (Zwier *et al.*, 2012; Wang *et al.*, 2013; Mukherjee *et al.*, 2015), we found two copies of DNMT1, three copies of DNMT2 and one copy of DNMT3. Strikingly, we also discovered that some potential additional copies DNMT3-like proteins, some of which are expressed. Our results will not only add the arthropod genetic resource and DNMT genetic toolkit but offer insights into studies on wasp polyembryony and adaptation.

7.1 Introduction

Copidosoma floridanum, a parasitoid wasp, is distributed worldwide and broadly used as agricultural pest control (Watanabe *et al.*, 2012). It has the largest record of a brood of any parasitical insect of 3,055 siblings. By the mechanism of polyembryony, a female parasitoid wasp lays one or two eggs into a suitable host and afterward, each egg developed into over 2,000 genetically identical individuals, who later develop into a brood of two major castes. Polyembryony studies in parasitic wasps can offer insights into the evolution of a novel mode of development (Grbic, 2003).

The life cycle of wasps is fascinating. An adult female wasp initially oviposits one or two eggs inside an egg of a host moth. As the moth egg develops into a caterpillar, the wasp egg starts to develop into thousands of wasp larvae siblings with identical genetic information. These larvae feed on the caterpillar's tissues. Simultaneously, wasps adjust their caste ratio to generate interspecific competition, creating a trade-off between reproduction and defense (Harvey *et al.*, 2000). Nearly a quarter of the larvae take on 'snakelike soldier forms' that attack larvae from other wasps or from rival eggs of their siblings. The surviving larvae (not killed by the soldiers) devour their host and later form pupae. Eventually, these pupae hatch, break and fly away from the mummified host, leaving the soldier larvae trapped inside to eventually die (Zhurov *et al.*, 2007).

Recent studies have revealed the evolution of polyembryony is associated with the evolution of developmental novelties such as total cleavage, the early specification of embryonic and cell proliferation phases, and sibling rivalry and brood sex ratio adjustment (Grbic *et al.*, 1992; Zhurov *et al.*, 2004, 2007). However, the mechanisms leading to polyembryony are still poorly understood. What evolutionary dynamics shaped the evolution of polyembryony and which mechanistic changes in the development underlie the embryo cloning process is little known (Zhurov *et al.*, 2007). In this study, we sequenced and annotated the first *C. floridanum* genome, focusing on DNMT gene families, to explore the polyembryony and host adaptations in parasitic wasps.

7.2 Results and discussions

7.2.1 Genomic statistics of *C. floridanum*

The genome of *C. floridanum* is 526 Mb with 8,028 scaffolds, shown in Table 26. The longest scaffold is 11.33 Mb and N50 length is 1.9 Mb. The genome size is relatively larger than *Drosophila* assembled genomes (111 Mb - 187 Mb) (Drosophila 12 Genomes Consortium *et al.*, 2007) but close to *Acyrthosiphon pisum,* pea aphid (464 Mb) (International Aphid Genomics, 2010) and *Culex quinquefasciatus,* southern house mosquito (579 Mb) (Arensburger *et al.*, 2010).

We used EUGENE gene annotation pipeline and predicted 21,050 protein-coding genes, which is higher than the *C. quinquefasciatus* repertoire of 18,883 genes but lower than pea aphid with a total gene number of 34,604 (this estimate is likely to exceed the true number of protein-coding genes), as described in (International Aphid Genomics, 2010). A significant fraction of the assembled *C. floridanum* genome was composed of 58% transposable element (TE) (i.e., 307 Mb TE out of the 526 Mb genome size), which is greater than the TE fractions of *C. quinquefasciatus* (29%), *A. pisum* (37.8%) (International Aphid Genomics, 2010), *Nasonia* (30%) (Werren *et al.*, 2010) and *Ae.* aegypti (42 to 47%) (Nene et al., 2007; Arensburger et al., 2010), suggesting an increased level of TE activity or reduced intensity of selection though TE activities in *C. floridanum*.

All the genomic data (note: in this thesis I used VIB assembly and annotation while the current updated genome version in ORCAE is CRG version) is available on ORCAE database (Sterck et al., 2012) at http://bioinformatics.psb.ugent.be/orcae/overview/Copfl.

Table 26: Assembly and annotation statistics of *C. floridanum* genome.

Category	Info
genome size (scaffolds)	530,269,664 nt
genome size (contigs)	520,344,150 nt
largest scaffold	11,336,592 nt
av. scaffold length	66,052.52 nt
number of contigs	17,839
largest contig	1,375,424 nt
av. contig length	29,168.91 nt
gaps (>50N)	4,939 (9,925,514 nt)
nr_loci (exons+introns)	21,050
av.length.loci	9,636.51 nt
loci density	24,719.44 nt/gene
nr_genes	21,050
gene density	40.45 genes/Mb
av.length.genes	1,241.99 nt
median.length.genes	842 nt
nr_exons	101,701
%GC of CDS	46.89
cumul_exon_length	26,143,993 nt
av.length.exons	257.07 nt
median.length.exons	182 nt
longest.exons	30,924 nt (COPFL5399g00240.1.1)
av.nr.exons/gene	4.83
most exons/gene	142 COPFL6713g00100.6
cumul_CDS_length	23,991,140 nt
av.length.CDS	1,139.72 nt
cumul_intron_length	50,130,747 nt
av.length.intron	644.27 nt
median.length.introns	94 nt
%GC of intron	35.23
longest CDS	52,326 nt
shortest CDS	84 nt

7.2.2 Genome annotation of the DNMT gene families

Given the genome of a single zygote stays the same during duplication and all the siblings share identical genomic information, it is possible that epigenetic information is inherited from one parental generation to the next to specify the caste fates.

Initially, we prepared two assembly versions (VIB Gent version and CRG Barcelona version). We annotated the DNMT gene family across both assembly versions and compared DNMT genes, which later showed that the two assemblies primarily have the same DNMT1, 2 and 3 genes. Table 27 listed the annotated DNMT genes in the two assemblies.

Table 27: DNMT comparison between VIB and CRG versions

	VIB assembly and annotation	CRG assembly and annotation
DNMT1	2 intact genes	2 intact 1 fragment
DNMT2	3 intact genes	3 intact genes
DNMT3	1 truncated gene	1 truncated gene
DNMT3-like	9 intact genes 11 pseudogenes 1 fragment 3 pseudo&fragments	11 intact genes 8 pseudogenes 1 truncated

7.2.3 Evolutionary analysis of DNMT gene families

The distribution of DNMT1, 2 and 3 families in insects is known to be patchy (Glastad et. al. 2011). Most of the DNMT3 genes across the DNMT gene families seem lost in species where DNA methylation is absent. Some multiple copies of DNMT1 (Nasonia, Apis, Aphids) and DNMT3 (Aphids) have been reported (International Aphid Genomics, 2010; Werren *et al.*, 2010). During our annotation for *C. floridanum* genome, we have found potential additional copies of DNMT3-related proteins (named as DNMT3-like), some of them expressed, which prompted a phylogenetic analysis of these DNMT gene families. In all cases, we searched for homologs in annotated genomes of other sequenced insects and *Daphnia pulex* (as out-group) and reconstructed phylogenies using Maximum Likelihood approach.

Figure 54 shows the relationships and approximate divergence times of major insect lineages and an outgroup crustacean, *Daphnia pulex* (Gaunt and Miles, 2002; Grimaldi and Engel, 2005). Branches are named for insect orders, with representative species for which DNA methylation information has been obtained listed below. Dots represent the number of DNMTs found in a sequenced genome and the presence of methyl-CpG-binding domain proteins (MDBs: absence indicates no DNMTs of a given family, whereas question marks indicate no data is applicable). The putative DNMT loss is marked on branches based on currently available data.

As the diagram shown in Figure 55 (Werren *et al.*, 2010), *Nasoina* harbors a toolkit of methyltransferase genes like vertebrates. *Drosophila*'s diminutive toolkit comprised solely of DNMT3, thus illustrating the usefulness of *Nasonia* as an insect model for mammalian-style methylation.

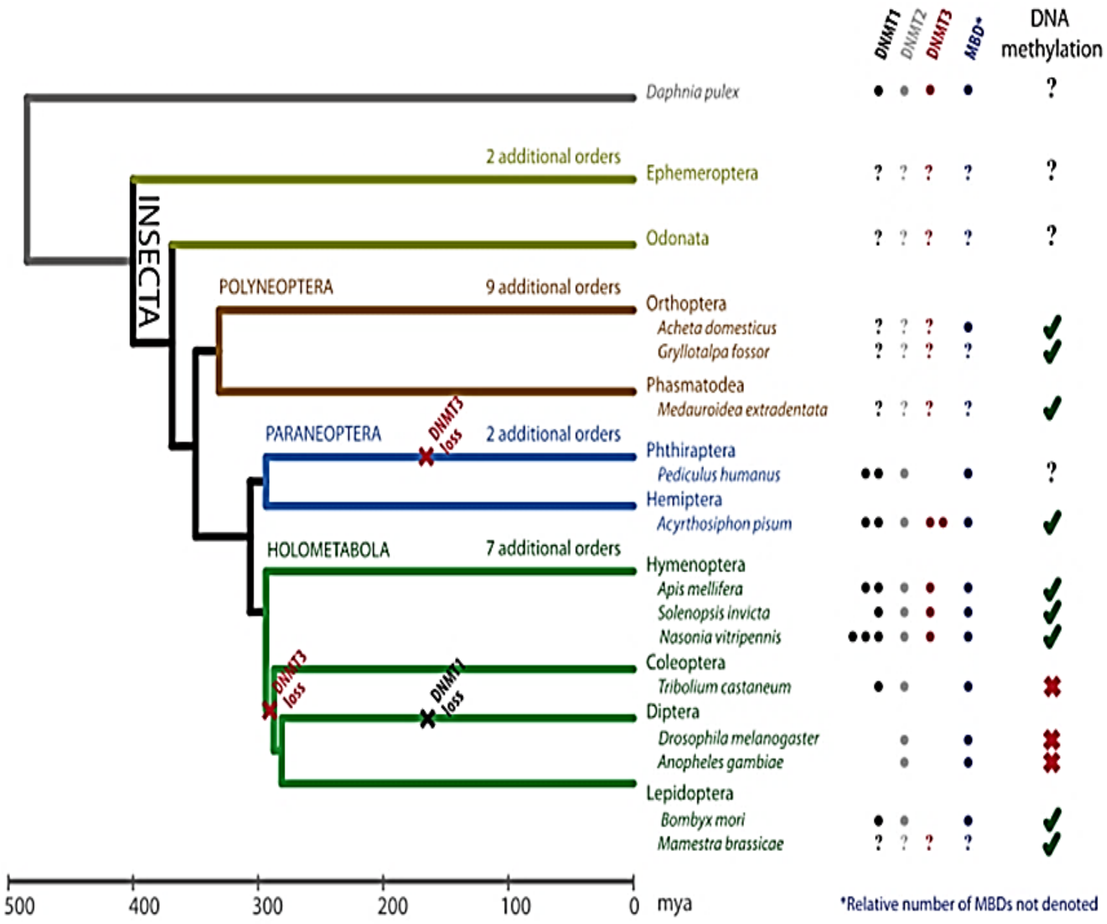


Figure 54: Phylogenetic distribution of DNA methylation in insects.
The detection of DNA methylation is indicated by a check mark and the validation of a near-total lack of DNA methylation is indicated by an 'X' with references provided in the text.

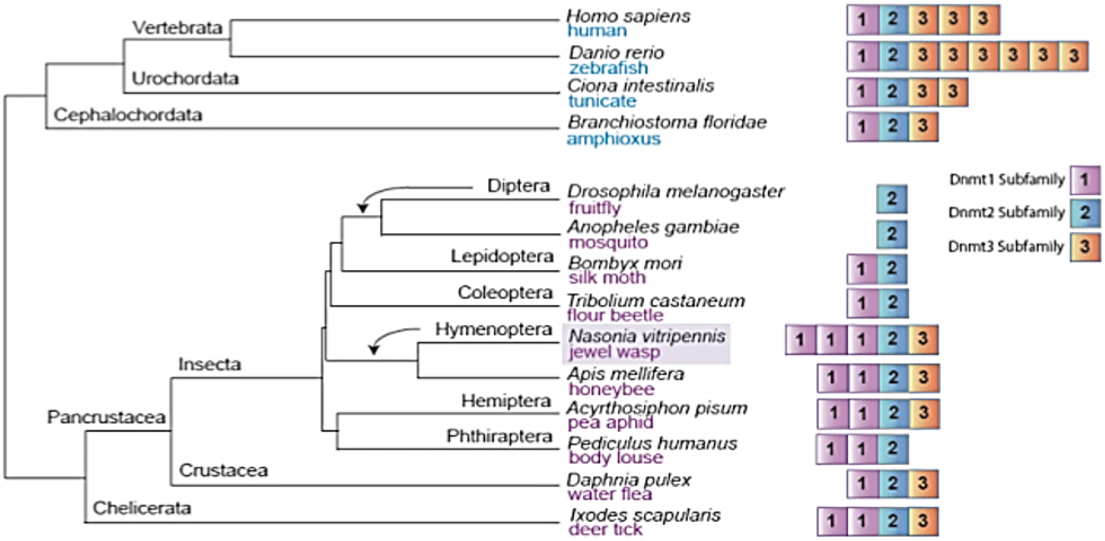
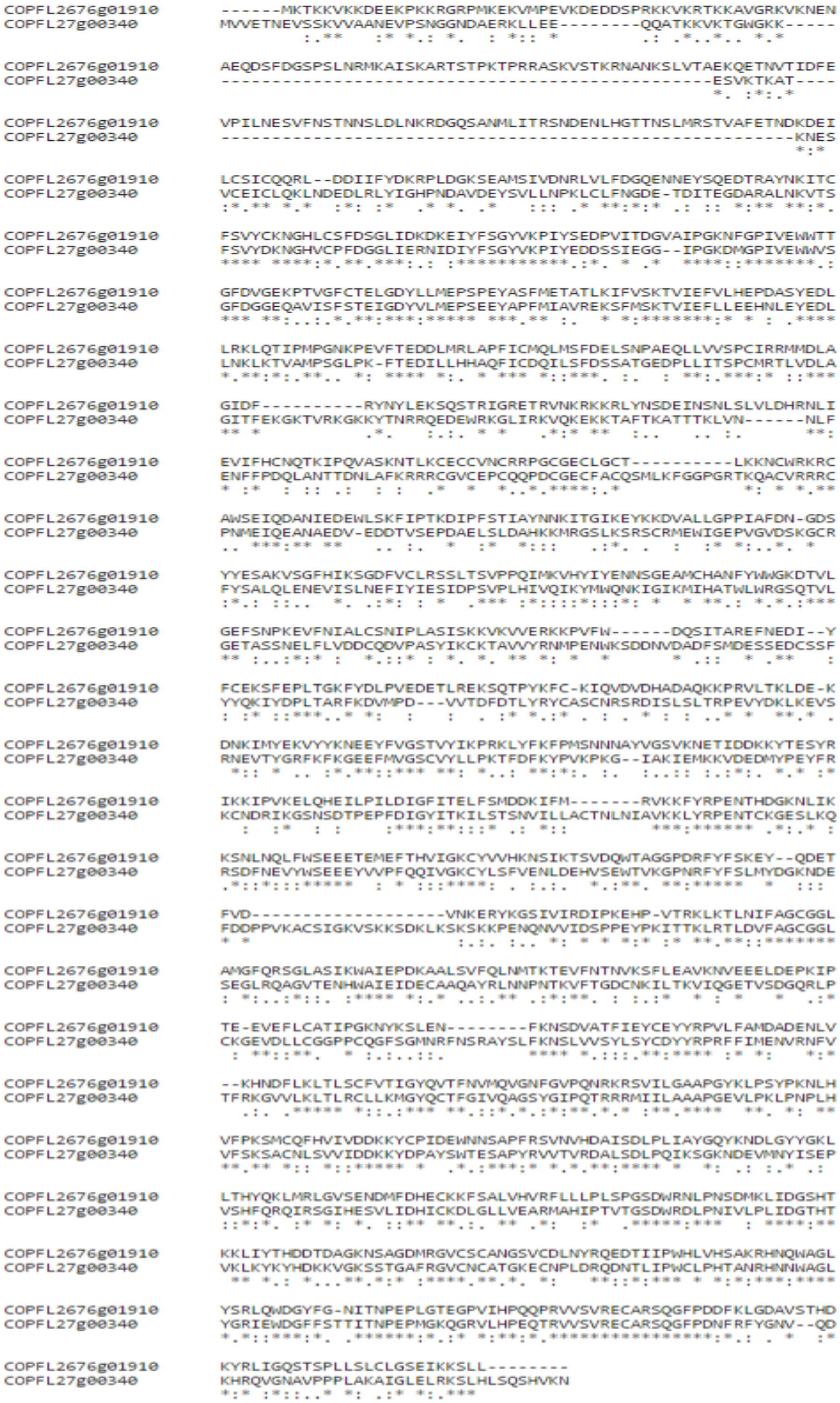


Figure 55: Prevalence of the DNMT family across taxa.

7.2.4 DNMT1

DNMT1 is typically considered a maintenance methyltransferase that involved in preserving consistent methylation across cell divisions and generations (Lyko and Maleszka, 2011). Our annotation finds two copies of DNMT1 genes, which correspond to two out of three DNMT1 genes presented in *Nasonia*. These genes originated through independent duplications in wasps (Werren *et al.*, 2010). Parallel duplications are observed that affect bees and ants, although, interestingly, one subfamily was lost in ants and only retained in bees, as shown in Figure 54. COPFL2676g01910 (CRG ID COPFL2676g03170) and COPFL27g00340 (CRG ID COPFL27g00460) are quite divergent at the protein level. Additionally, we observe that aphid and *Pediculus* lineages also show independent duplications. We assume that the DNMT1 family was duplicated independently in several insect lineages. Strikingly, the lineages in which DNMT1 presents but is not duplicated (*Tribolium* and *Bombyx*) show a large degree of sequence divergence and are wrongly placed in the tree (Figure 57), suggesting they might have a different function. Therefore, *C. floridanum* is possibly a normal wasp in this respect; it is possible that one copy is missed either in assembly or annotation (but unlikely).



```
COPFL2676g01910     ------MKTKKVKKDEEKPKKRGRPMKEKVMPEVKDEDDSPRKKVKRTKKAVGRKVKNEN
COPFL27g00340       MVVETNEVSSKVVAANEVPSNGGNDAERKLLEE--------QQATKKVKTGWGKK-----
                            :.**   :* *.: *.  : *:: *        .: .*..*.. *.*

COPFL2676g01910     AEQDSFDGSPSLNRMKAISKARTSTPKTPRRASKVSTKRNANKSLVTAEKQETNVTIDFE
COPFL27g00340       ------------------------------------------------ESVKTKAT----
                                                                    *. :*:.*

COPFL2676g01910     VPILNESVFNSTNNSLDLNKRDGQSANMLITRSNDENLHGTTNSLMRSTVAFETNDKDEI
COPFL27g00340       --------------------------------------------------------KNES
                                                                            *:*

COPFL2676g01910     LCSICQQRL--DDIIFYDKRPLDGKSEAMSIVDNRLVLFDGQENNEYSQEDTRAYNKITC
COPFL27g00340       VCEICLQKLNDEDLRLYIGHPNDAVDEYSVLLNPKLCLFNGDE-TDITEGDARALNKVTS
                    :*.** *.*  :*: :*  .* *. .*   ::: .* **:*:* .: :: *:** **:*.

COPFL2676g01910     FSVYCKNGHLCSFDSGLIDKDKEIYFSGYVKPIYSEDPVITDGVAIPGKNFGPIVEWWTT
COPFL27g00340       FSVYDKNGHVCPFDGGLIERNIDIYFSGYVKPIYEDDSSIEGG--IPGKDMGPIVEWWVS
                    **** ****:*.**.***:.: :***********.:*. * .*  ****::*******.:

COPFL2676g01910     GFDVGEKPTVGFCTELGDYLLMEPSPEYASFMETATLKIFVSKTVIEFVLHEPDASYEDL
COPFL27g00340       GFDGGEQAVISFSTEIGDYVLMEPSEEYAPFMIAVREKSFMSKTVIEFLLEEHNLEYEDL
                    *** **:..:.*.**:***:***** ***.** :.  * *:*******:* * : .****

COPFL2676g01910     LRKLQTIPMPGNKPEVFTEDDLMRLAPFICMQLMSFDELSNPAEQLLVVSPCIRRMMDLA
COPFL27g00340       LNKLKTVAMPSGLPK-FTEDILLHHAQFICDQILSFDSSATGEDPLLITSPCMRTLVDLA
                    *.**:*:.**.. *: **** *:. * *** *::***. :.  : **:.***:* ::***

COPFL2676g01910     GIDF----------RYNYLEKSQSTRIGRETRVNKRKKRLYNSDEINSNLSLVLDHRNLI
COPFL27g00340       GITFEKGKTVRKGKKYTNRRQEDEWRKGLIRKVQKEKKTAFTKATTTKLVN------NLF
                    ** *          .*.   :.:. * *   .*:* **  :..   .. :.      **:

COPFL2676g01910     EVIFHCNQTKIPQVASKNTLKCECCVNCRRPGCGECLGCT----------LKKNCWRKRC
COPFL27g00340       ENFFPDQLANTTDNLAFKRRRCGVCEPCQQPDCGECFACQSMLKFGGPGRTKQACVRRRC
                    * :*  : :: .:  : :  .*  *  *..*.****:.*            *: * *.**

COPFL2676g01910     AWSEIQDANIEDEWLSKFIPTKDIPFSTIAYNNKITGIKEYKKDVALLGPPIAFDN-GDS
COPFL27g00340       PNMEIQEANAEDV-EDDTVSEPDAELSLDAHKKMRGSLKSRSCRMEWIGEPVGVDSKGCR
                    .. ***:** **   .. :.  *  :*  *:::   .:*. .  :  :* *:..*. *

COPFL2676g01910     YYESAKVSGFHIKSGDFVCLRSSLTSVPPQIMKVHYIYENNSGEAMCHANFYWWGKDTVL
COPFL27g00340       FYSALQLENEVISLNEFIYIESIDPSVPLHIVQIKYMWQNKIGIKMIHATWLWRGSQTVL
                    :*.: ::..  *. .:*: : *  .*** :*::::*:::*: *  * **.: *.*.:***

COPFL2676g01910     GEFSNPKEVFNIALCSNIPLASISKKVKVVERKKPVFW------DQSITAREFNEDI--Y
COPFL27g00340       GETASSNELFLVDDCQDVPASYIKCKTAVVYRNMPENWKSDDNVDADFSMDESSEDCSSF
                    ** :..:*:* :  *.::* : *. *. ** *: *  *      * .::  * .**   :

COPFL2676g01910     FCEKSFEPLTGKFYDLPVEDETLREKSQTPYKFC-KIQVDVDHADAQKKPRVLTKLDE-K
COPFL27g00340       YYQKIYDPLTARFKDVMPD---VVTDFDTLYRYCASCNRSRDISLSLTRPEVYDKLKEVS
                    : :* ::***..* *:  :   :  . :* *.:* . : . * : : ..* *  **.* .

COPFL2676g01910     DNKIMYEKVYYKNEEYFVGSTVYIKPRKLYFKFPMSNNNAYVGSVKNETIDDKKYTESYR
COPFL27g00340       RNEVTYGRFKFKGEEFMVGSCVYLLPKTFDFKYPVKPKG--IAKIEMKKVDEDMYPEYFR
                     *:: *  .. :*.**::*** **: *..: **:*:. :.  :..:: :.:*:. *.* :*

COPFL2676g01910     IKKIPVKELQHEILPILDIGFITELFSMDDKIFM-------RVKKFYRPENTHDGKNLIK
COPFL27g00340       KCNDRIKGSNSDTPEPFDIGYITKILSTSNVILLACTNLNIAVKKLYRPENTCKGESLKQ
                      :  :*  : :    :***:**:::* .: *::        ***:****** .*:.* :

COPFL2676g01910     KSNLNQLFWSEEETEMEFTHVIGKCYVVHKNSIKTSVDQWTAGGPDRFYFSKEY--QDET
COPFL27g00340       RSDFNEVYWSEEEYVVPFQQIVGKCYLSFVENLDEHVSEWTVKGPNRFYFSLMYDGKNDE
                    .*::*:::*****  : * :::****: . :.:.  *.:**. **:*****  *  :::

COPFL2676g01910     FVD-------------------VNKERYKGSIVIRDIPKEHP-VTRKLKTLNIFAGCGGL
COPFL27g00340       FDDPPVKACSIGKVSKKSDKLKSKSKKPENQNVVIDSPPEYPKITTKLRTLDVFAGCGGL
                    * *                    :.:. :.. *: * * *:* :* **.**::*******

COPFL2676g01910     AMGFQRSGLASIKWAIEPDKAALSVFQLNMTKTEVFNTNVKSFLEAVKNVEEELDEPKIP
COPFL27g00340       SEGLRQAGVTENHWAIEIDECAAQAYRLNNPNTKVFTGDCNKILTKVIQGETVSDGQRLP
                    : *:..:*::. :**** *:.* ..:.** .:*:**. : :.:*  * : *   *  .:*

COPFL2676g01910     TE-EVEFLCATIPGKNYKSLEN--------FKNSDVATFIEYCEYYRPVLFAMDADENLV
COPFL27g00340       CKGEVDLLCGGPPCQGFSGMNRFNSRAYSLFKNSLVVSYLSYCDYYRPRFFIMENVRNFV
                     : **::**.  * :.:..::.        **** *.:::.**:**** :* *:   *:*

COPFL2676g01910     --KHNDFLKLTLSCFVTIGYQVTFNVMQVGNFGVPQNRKRSVILGAAPGYKLPSYPKNLH
COPFL27g00340       TFRKGVVLKLTLRCLLKMGYQCTFGIVQAGSYGIPQTRRRMIILAAAPGEVLPKLPNPLH
                      .:. .***** *::.:*** **.::*.*.:*:**.*.* :**.****  **. *: **

COPFL2676g01910     VFPKSMCQFHVIVDDKKYCPIDEWNNSAPFRSVNVHDAISDLPLIAYGQYKNDLGYYGKL
COPFL27g00340       VFSKSACNLSVVIDDKKYDPAYSWTESAPYRVVTVRDALSDLPQIKSGKNDEVMNYISEP
                    **.** *:: *::***** *  .*.:***:* *.*.**:**** *  *: .: :.* .:

COPFL2676g01910     LTHYQKLMRLGVSENDMFDHECKKFSALVHVRFLLLPLSPGSDWRNLPNSDMKLIDGSHT
COPFL27g00340       VSHFQRQIRSGIHESVLIDHICKDLGLLVEARMAHIPTVTGSDWRDLPNIVLPLIDGTHT
                    ::*:*. :* *: *. ::** **.:. ** .*:  :*  .*****:***  : ****:**

COPFL2676g01910     KKLIYTHDDTDAGKNSAGDMRGVCSCANGSVCDLNYRQEDTIIPWHLVHSAKRHNQWAGL
COPFL27g00340       VKLKYKYHDKKVGKSSTGAFRGVCNCATGKECNPLDRQDNTLIPWCLPHTANRHNNWAGL
                     ** *.: *...**.*:* :****.**.*. *:   **::*:*** * *:*:***:****

COPFL2676g01910     YSRLQWDGYFG-NITNPEPLGTEGPVIHPQQPRVVSVRECARSQGFPDDFKLGDAVSTHD
COPFL27g00340       YGRIEWDGFFSTTITNPEPMGKQGRVLHPEQTRVVSVRECARSQGFPDNFRFYGNV--QD
                    *.*::***:*. .******:*.:* *:**:*.****************:*.: . *  :*

COPFL2676g01910     KYRLIGQSTSPLLSLCLGSEIKKSLL--------
COPFL27g00340       KHRQVGNAVPPPLAKAIGLELRKSLHLSQSHVKN
                    *:* :*::..* *: .:* *:.***
```

Figure 56: The protein sequence alignment of DNMT1 genes in *C. floridanum*.

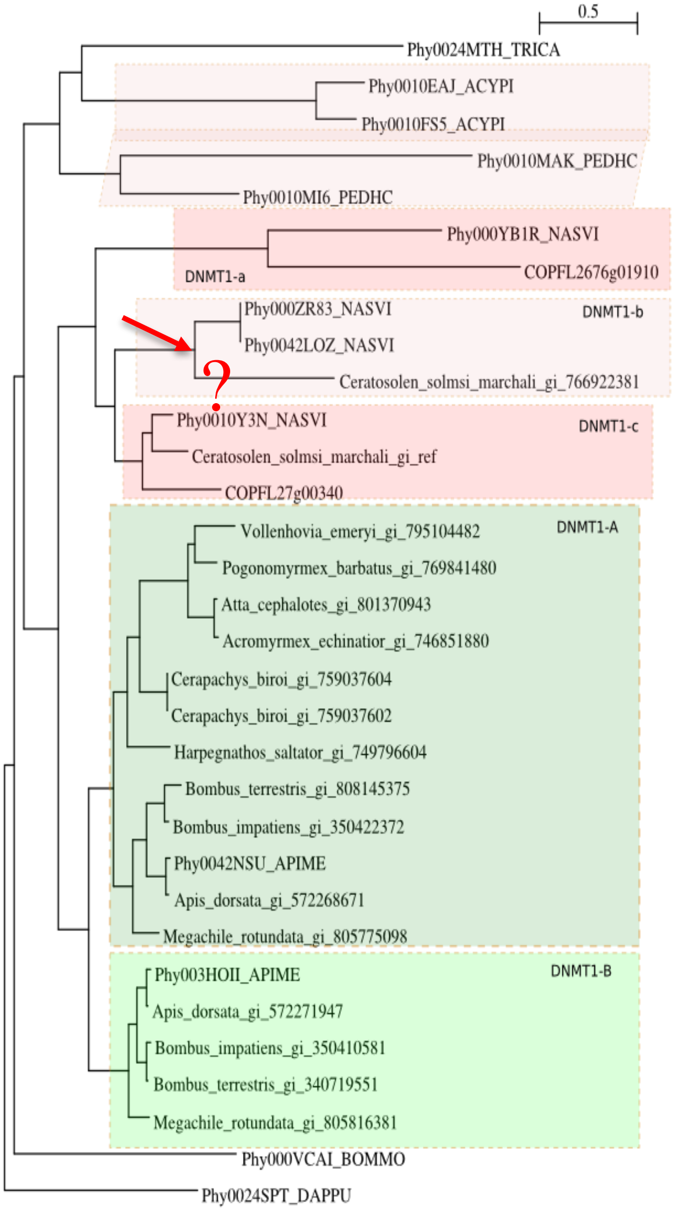
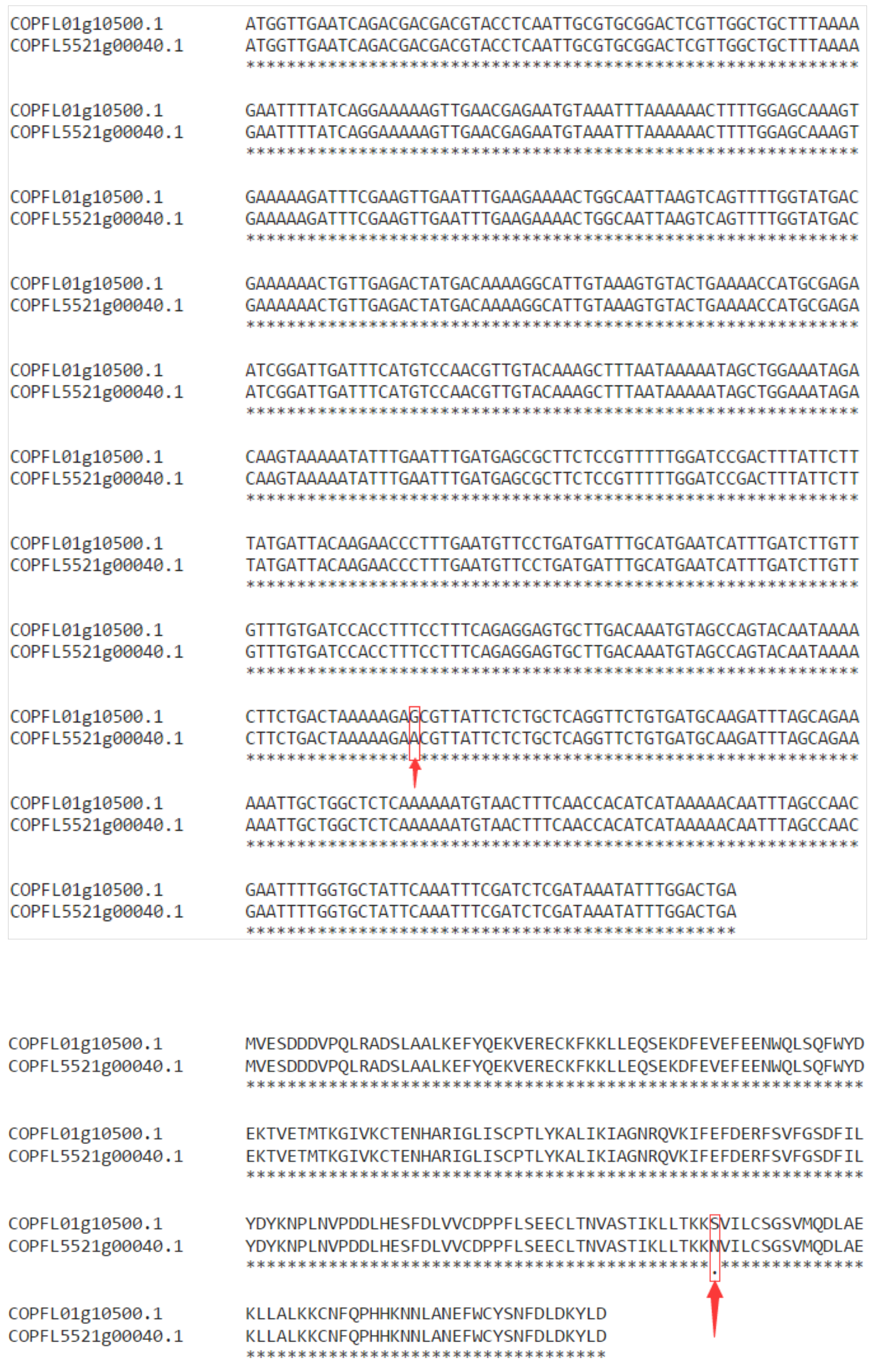

0.5
Phy0024MTH_TRICA
Phy0010EAJ_ACYPI
Phy0010FS5_ACYPI
Phy0010MAK_PEDHC
Phy0010MI6_PEDHC
Phy000YB1R_NASVI
DNMT1-a
COPFL2676g01910
Phy000ZR83_NASVI
DNMT1-b
Phy0042LOZ_NASVI
?
Ceratosolen_solmsi_marchali_gi_766922381
Phy0010Y3N_NASVI
DNMT1-c
Ceratosolen_solmsi_marchali_gi_ref
COPFL27g00340
Vollenhovia_emeryi_gi_795104482
DNMT1-A
Pogonomyrmex_barbatus_gi_769841480
Atta_cephalotes_gi_801370943
Acromyrmex_echinatior_gi_746851880
Cerapachys_biroi_gi_759037604
Cerapachys_biroi_gi_759037602
Harpegnathos_saltator_gi_749796604
Bombus_terrestris_gi_808145375
Bombus_impatiens_gi_350422372
Phy0042NSU_APIME
Apis_dorsata_gi_572268671
Megachile_rotundata_gi_805775098
Phy003HOII_APIME
DNMT1-B
Apis_dorsata_gi_572271947
Bombus_impatiens_gi_350410581
Bombus_terrestris_gi_340719551
Megachile_rotundata_gi_805816381
Phy000VCAI_BOMMO
Phy0024SPT_DAPPU

Figure 57: Phylogenetic tree of DNMT1 genes across arthropod genomes.
The question mark indicates that there might be one missed DNMT1 gene in the genomes of *C. floridanum*, possibly due to assembly artifacts.

7.2.5 DNMT2

DNMT2 is known to be involved in the methylation of transfer RNAs (Dong *et al.*, 2001). A previous study in human genome shows that DNMT2 strongly binds to DNA but it does not display methyltransferase activity. Instead, it can methylate cytosine 38 in the anticodon loop of aspartic acid transfer RNA, thus it is also called tRNA (cytosine-5)-methyltransferase (TRDMT1) (Okano *et al.*, 1998; Goll *et al.*, 2006).

We annotated three possible DNMT2 genes, of which, one is extremely divergent and excluded from further analyses because it would result in extremely long branches that were placed near the root and thus we did not include it in our phylogenetic analysis. The other two remaining genes are nicely placed in the phylogenetic tree, as a sister branch to the ortholog in *Nasonia*. They may result from a very recent duplication in *C. floridanum.* Many other species tend to have two highly related copies, probably they are isoforms. Genes COPFL5521g00040 (VIB and CRA share the same ID) and COPFL01g10500 (VIB version ID COPFL01g07190) are located on scaffold 5521 and scaffold 1, respectively. They are located separately in the genomic loci, which suggest these two genes are possibly evolved from gene conversion rather than from tandem duplication (if it is not an incorrect assembly), or they might be just an allelic variant. These two genes are both 648 nt in size but with only one nucleotide difference, thus leading one amino acid changes (S - N) (Figure 58).



```
COPFL01g10500.1         ATGGTTGAATCAGACGACGACGTACCTCAATTGCGTGCGGACTCGTTGGCTGCTTTAAAA
COPFL5521g00040.1       ATGGTTGAATCAGACGACGACGTACCTCAATTGCGTGCGGACTCGTTGGCTGCTTTAAAA
                        ************************************************************

COPFL01g10500.1         GAATTTTATCAGGAAAAAGTTGAACGAGAATGTAAATTTAAAAAACTTTTGGAGCAAAGT
COPFL5521g00040.1       GAATTTTATCAGGAAAAAGTTGAACGAGAATGTAAATTTAAAAAACTTTTGGAGCAAAGT
                        ************************************************************

COPFL01g10500.1         GAAAAAGATTTCGAAGTTGAATTTGAAGAAAACTGGCAATTAAGTCAGTTTTGGTATGAC
COPFL5521g00040.1       GAAAAAGATTTCGAAGTTGAATTTGAAGAAAACTGGCAATTAAGTCAGTTTTGGTATGAC
                        ************************************************************

COPFL01g10500.1         GAAAAAACTGTTGAGACTATGACAAAAGGCATTGTAAAGTGTACTGAAAACCATGCGAGA
COPFL5521g00040.1       GAAAAAACTGTTGAGACTATGACAAAAGGCATTGTAAAGTGTACTGAAAACCATGCGAGA
                        ************************************************************

COPFL01g10500.1         ATCGGATTGATTTCATGTCCAACGTTGTACAAAGCTTTAATAAAAATAGCTGGAAATAGA
COPFL5521g00040.1       ATCGGATTGATTTCATGTCCAACGTTGTACAAAGCTTTAATAAAAATAGCTGGAAATAGA
                        ************************************************************

COPFL01g10500.1         CAAGTAAAAATATTTGAATTTGATGAGCGCTTCTCCGTTTTTGGATCCGACTTTATTCTT
COPFL5521g00040.1       CAAGTAAAAATATTTGAATTTGATGAGCGCTTCTCCGTTTTTGGATCCGACTTTATTCTT
                        ************************************************************

COPFL01g10500.1         TATGATTACAAGAACCCTTTGAATGTTCCTGATGATTTGCATGAATCATTTGATCTTGTT
COPFL5521g00040.1       TATGATTACAAGAACCCTTTGAATGTTCCTGATGATTTGCATGAATCATTTGATCTTGTT
                        ************************************************************

COPFL01g10500.1         GTTTGTGATCCACCTTTCCTTTCAGAGGAGTGCTTGACAAATGTAGCCAGTACAATAAAA
COPFL5521g00040.1       GTTTGTGATCCACCTTTCCTTTCAGAGGAGTGCTTGACAAATGTAGCCAGTACAATAAAA
                        ************************************************************

COPFL01g10500.1         CTTCTGACTAAAAAGAGCGTTATTCTCTGCTCAGGTTCTGTGATGCAAGATTTAGCAGAA
COPFL5521g00040.1       CTTCTGACTAAAAAGAACGTTATTCTCTGCTCAGGTTCTGTGATGCAAGATTTAGCAGAA
                        **************** *******************************************

COPFL01g10500.1         AAATTGCTGGCTCTCAAAAAAATGTAACTTTCAACCACATCATAAAAACAATTTAGCCAAC
COPFL5521g00040.1       AAATTGCTGGCTCTCAAAAAAATGTAACTTTCAACCACATCATAAAAACAATTTAGCCAAC
                        ************************************************************

COPFL01g10500.1         GAATTTTGGTGCTATTCAAATTTCGATCTCGATAAATATTTGGACTGA
COPFL5521g00040.1       GAATTTTGGTGCTATTCAAATTTCGATCTCGATAAATATTTGGACTGA
                        ************************************************
```

```
COPFL01g10500.1         MVESDDDVPQLRADSLAALKEFYQEKVERECKFKKLLEQSEKDFEVEFEENWQLSQFWYD
COPFL5521g00040.1       MVESDDDVPQLRADSLAALKEFYQEKVERECKFKKLLEQSEKDFEVEFEENWQLSQFWYD
                        ************************************************************

COPFL01g10500.1         EKTVETMTKGIVKCTENHARIGLISCPTLYKALIKIAGNRQVKIFEFDERFSVFGSDFIL
COPFL5521g00040.1       EKTVETMTKGIVKCTENHARIGLISCPTLYKALIKIAGNRQVKIFEFDERFSVFGSDFIL
                        ************************************************************

COPFL01g10500.1         YDYKNPLNVPDDLHESFDLVVCDPPFLSEECLTNVASTIKLLTKKSVILCSGSVMQDLAE
COPFL5521g00040.1       YDYKNPLNVPDDLHESFDLVVCDPPFLSEECLTNVASTIKLLTKKNVILCSGSVMQDLAE
                        *********************************************.**************

COPFL01g10500.1         KLLALKKCNFQPHHKNNLANEFWCYSNFDLDKYLD
COPFL5521g00040.1       KLLALKKCNFQPHHKNNLANEFWCYSNFDLDKYLD
                        ***********************************
```

Figure 58: The mutation point of DNMT2 genes in *C. floridanum*.

Top: nucleotide sequences; Bottom: protein sequences. Both nucleotide and protein sequences were aligned using MUSCLE.

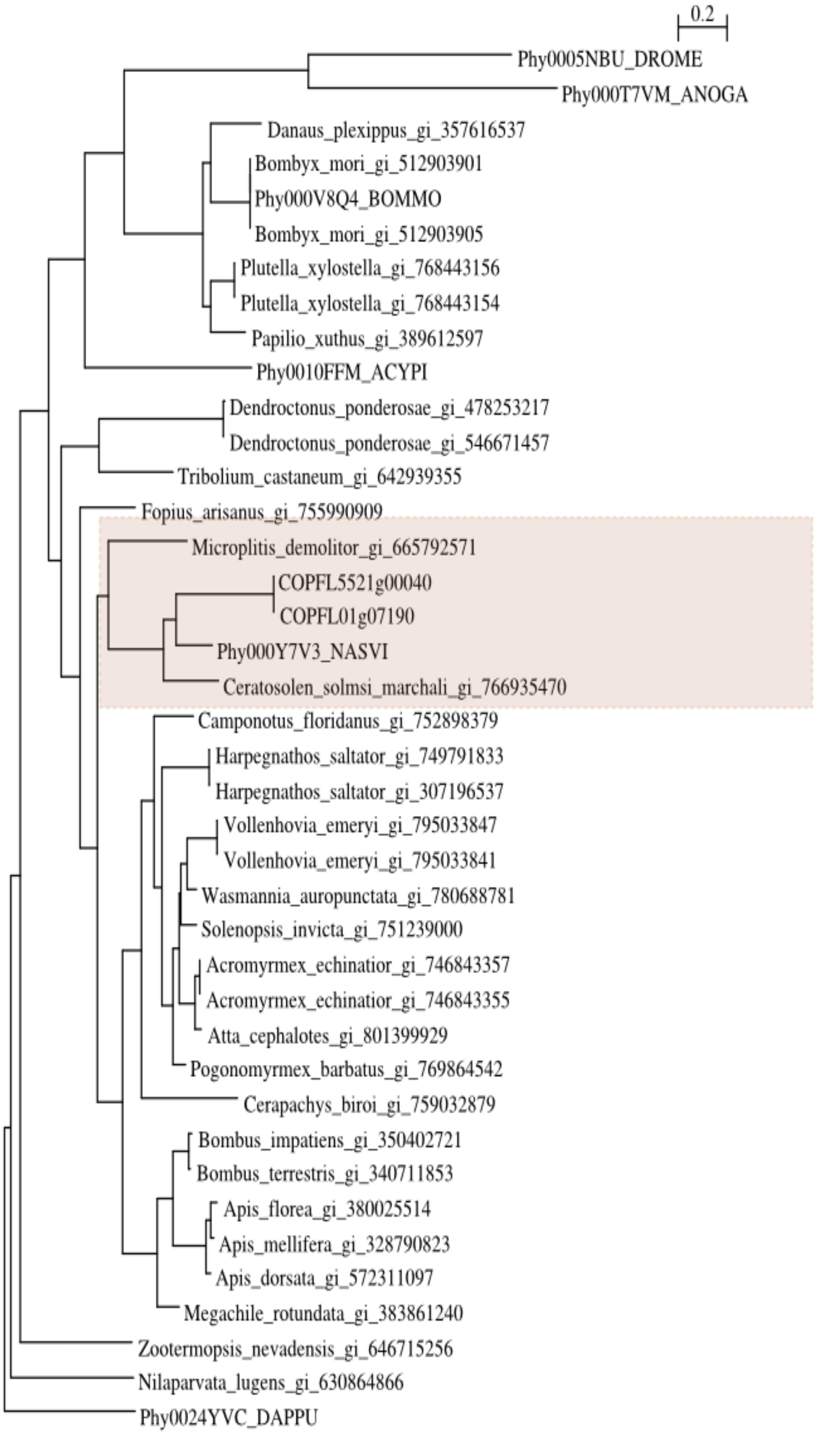

0.2
Phy0005NBU_DROME
Phy000T7VM_ANOGA
Danaus_plexippus_gi_357616537
Bombyx_mori_gi_512903901
Phy000V8Q4_BOMMO
Bombyx_mori_gi_512903905
Plutella_xylostella_gi_768443156
Plutella_xylostella_gi_768443154
Papilio_xuthus_gi_389612597
Phy0010FFM_ACYPI
Dendroctonus_ponderosae_gi_478253217
Dendroctonus_ponderosae_gi_546671457
Tribolium_castaneum_gi_642939355
Fopius_arisanus_gi_755990909
Microplitis_demolitor_gi_665792571
COPFL5521g00040
COPFL01g07190
Phy000Y7V3_NASVI
Ceratosolen_solmsi_marchali_gi_766935470
Camponotus_floridanus_gi_752898379
Harpegnathos_saltator_gi_749791833
Harpegnathos_saltator_gi_307196537
Vollenhovia_emeryi_gi_795033847
Vollenhovia_emeryi_gi_795033841
Wasmannia_auropunctata_gi_780688781
Solenopsis_invicta_gi_751239000
Acromyrmex_echinatior_gi_746843357
Acromyrmex_echinatior_gi_746843355
Atta_cephalotes_gi_801399929
Pogonomyrmex_barbatus_gi_769864542
Cerapachys_biroi_gi_759032879
Bombus_impatiens_gi_350402721
Bombus_terrestris_gi_340711853
Apis_florea_gi_380025514
Apis_mellifera_gi_328790823
Apis_dorsata_gi_572311097
Megachile_rotundata_gi_383861240
Zootermopsis_nevadensis_gi_646715256
Nilaparvata_lugens_gi_630864866
Phy0024YVC_DAPPU

Figure 59: Phylogenetic tree of the two DNMT2 genes in arthropod context.

7.2.6 DNMT3 and DNMT3-like

DNMT3 can methylate DNA and it has been assumed to be related more to environmentally responsive DNA methylation that happens within the lifetime of an individual (Lyko and Maleszka, 2011). We observe the DNMT3 gene family in *C. floridanum* is by far the most extreme example even though we find only one DNMT3 gene with truncated gene structure. However, there are up to 9 additional DNMT3-like protein-coding genes (plus up to 15 pseudogenes) discovered through manual annotation, some of which are expressed. When placed in phylogenies they group in a sister clade to the DNMT3 from *Nasonia/Copidosoma/Ceratosolen*, suggesting DNMT3-like genes result from duplication at the base of Apocrita, a suborder of insects in the order Hymenoptera. However, the low support of that sister relationship and the long branches within the DNMT3-like clade make it possible that this position is the result of Long Branch Attraction (LBA) artifact, known to pull long branches towards positions closer to the root of the trees.

Additionally, there are three DNMT3-like genes COPFL53355g00312, COPFL6729g00401, and COPFL078g0010 with shorter terminal branches in the phylogeny. They seem to have similar levels of synonymous and non-synonymous divergence rates with respect to *Nasonia* DNMT3 compared with what is observed in *Copidosoma* DNMT3 (Table 28), which may suggest they have some constraints at the protein level. The most conserved region corresponds to the DNMT3 domain, and despite significant amino acid substitution some short stretches of conserved residues are present.

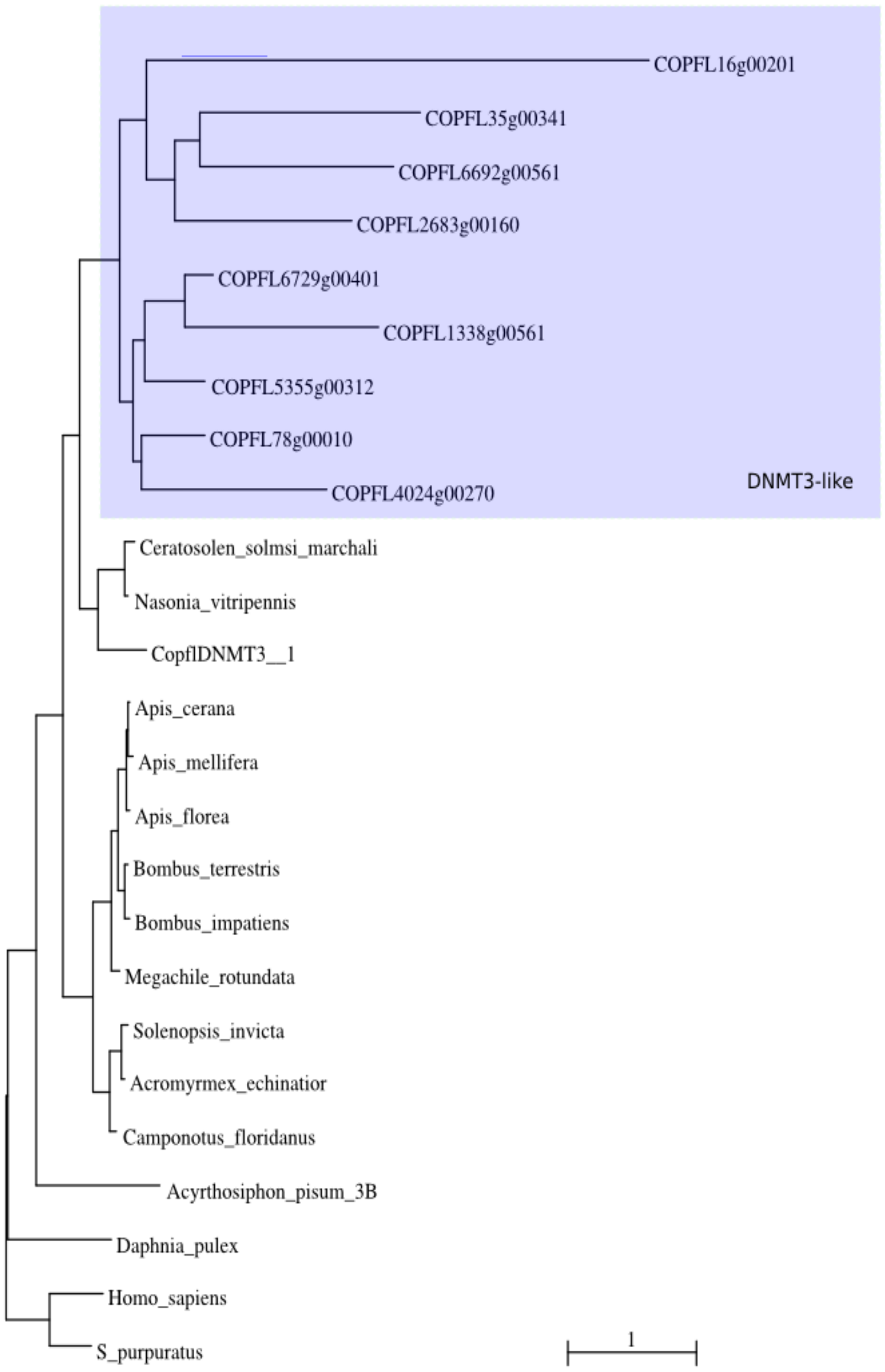


Figure 60: The phylogenetic tree of DNMT3 and the expansion of DNMT3-like genes.

Table 28: The evolutionary pressure analysis of DNMT3-like genes.

	Ratio*	**Ka**	**Ks**
DNMT3	0.1279	0.4604	3.59
COPFL2683g00160	0.16	0.52	3.25
COPFL5355g00312	0.106	0.38	3.6
COPFL6729g00401	0.1472	0.35753	3.17
COPFL16g00201	0.27	0.7	2.54
COPFL6692g00561	0.2	0.614	3.06
COPFL1338g00561	N/D	0.45	7.6
COPFL35g00341	N/D	0.722	19.08
COPFL078g0010	0.112	0.35	3.31
COPFL4024g00270	N/D	0.6	10.4

*it implies purifying or stabilizing selection (acting against change) and N/D indicates ratio is less than 0.1. DNMT3 is the gene from *C. floridanum* and the rest gene IDs are DNMT3-likes.

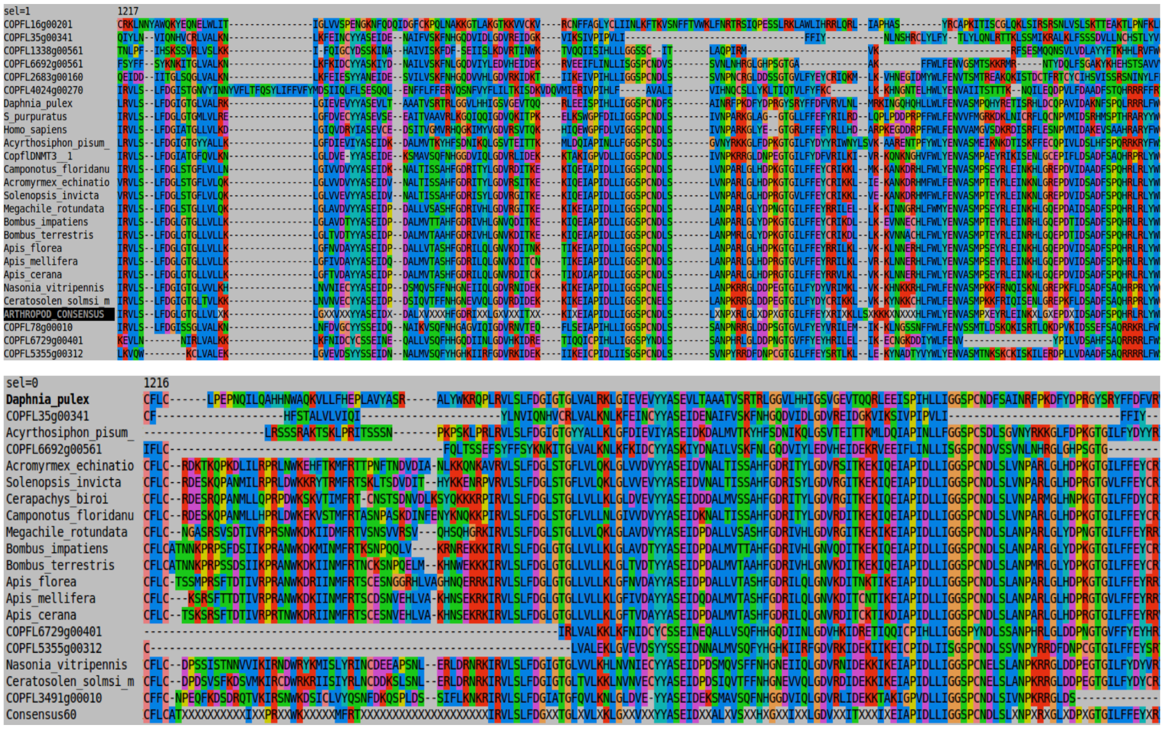


Figure 61: The protein sequence alignment of DNMT3-domain region.

Top: alignment DNMT3-domain region (695-896 residues in DNMT3); bottom: the regions of the alignment where DNMT3-like and other DNMT3 are most similar.

7.3 Summary

The genome of *C. floridanum* was sequenced and analyzed in this study, and the key results include the identification of a functional DNA methylation toolkit for wasp studies, as well as materials for the evolutionary and developmental genetics. This study also provides genomic resources for parasitoid biology as well as knowledge for further increasing the utility of parasitoids as pest-control agents.

7.4 Materials and Methods

7.4.1 Genome sequencing and assembly

Sequencing and draft assemblies: we initially prepared libraries PE-275, PE-330 and PE-800 for Illumina sequencing, and assembled them using multiple tools CLCbio (https://www.qiagenbioinformatics.com/) and Newbler contig assemblies (Zhang *et al.*, 2012; Nederbragt, 2014). Several assemblies were attempted and merged. Merging was successful, but further scaffolding with existing data did not improve the best assembly. Therefore, we prepared longer insert PE libraries, both 5 kb and 10 kb MP, before finalizing the draft assembly.

Whole genome alignment mis-assembly detection: whole genome alignment to a close reference using assembly fragmentation and BLAT was performed (Kent, 2002). Alignment chaining was done with syntenic alignment blocks and dynamic programming algorithm. Assembly fragmentation was achieved using synteny breakpoints. Re-scaffolding was done by ABySS scaffolder using PE and MP libraries (Simpson *et al.*, 2009).

Consistency-based mis-assembly detection: first, raw reads were mapped to the assembled genome. We selected the best scoring pair or pairs that fit fragment size distribution and intervals. Second, when both ends are mapped but in inconsistent order and orientation, we define the intervals where the other end should have mapped. Then we assigned the scoring intervals (+1 for consistent intervals and -1 for inconsistent intervals; The score is divided by a number of mappings in the case of multimaps). Later, all the scores summed at each position of the genome. Third, we determined intervals of positive and negative values.

7.4.2 Genome annotation and DNMT gene identification

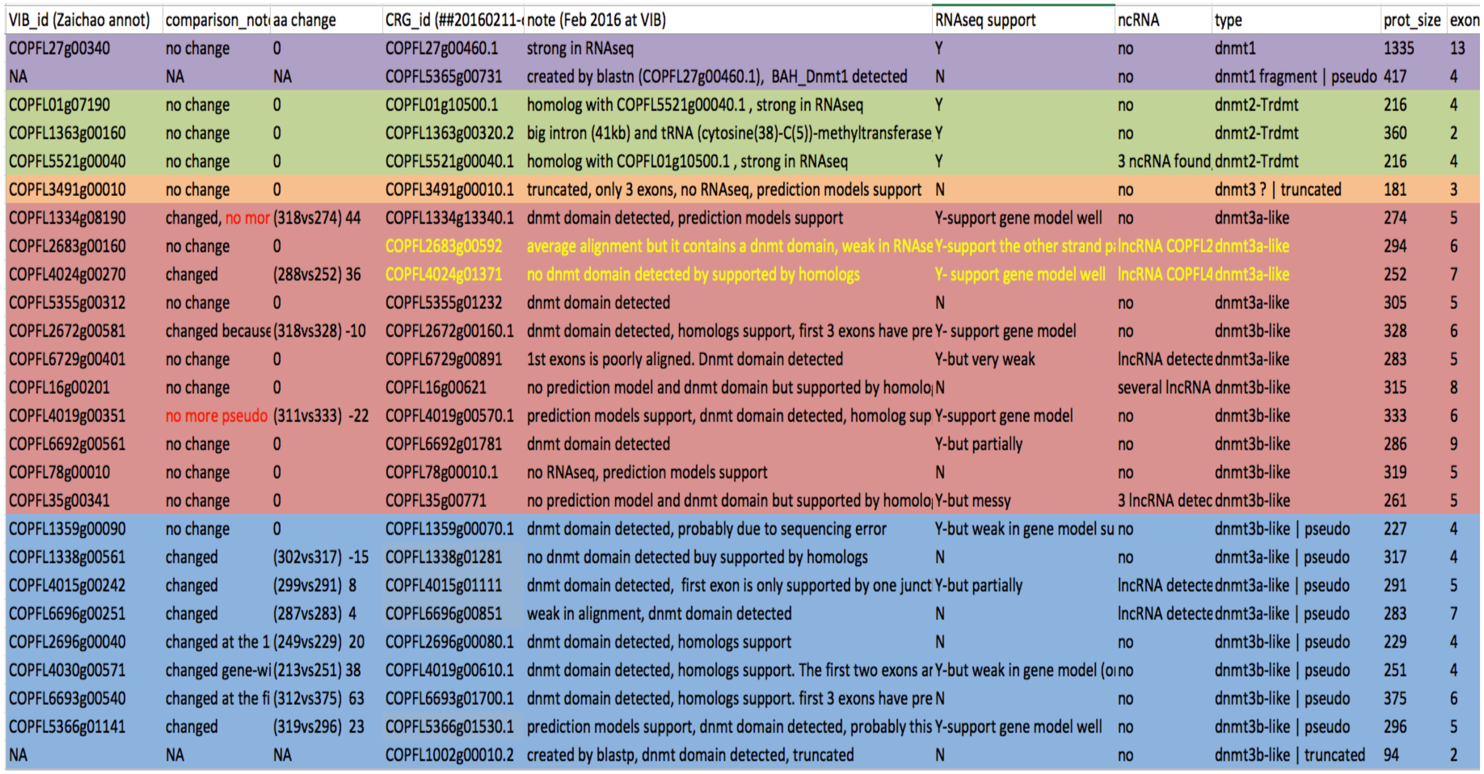
The final assembled genomic sequences (both VIB Gent version and CRG Barcelona version) were annotated using EUGENE pipeline with *Nasonia* protein database as a reference (version downloaded 20140610) (Foissac *et al.*, 2008; Werren *et al.*, 2010). The genome was masked using RepeatMasker with in-house built TE library by RepeatModelor (Smit and Hubley, 2008-2015; Tarailo-Graovac and Chen, 2009). The draft annotation then was further applied in the annotation for DNMT gene families. We used DNMT genes from Nasonia as baits to BLASTN, TBLASTN and BLASTP against *C. floridanum* genomic sequences and protein database, respectively (Werren *et al.*, 2010). The returned hits were manually verified in GenomeView using RNAseq data as supporting evidence (Abeel *et al.*, 2012).

7.4.3 Phylogeny analysis

The phylogenies of DNMT gene families were constructed using PhylomeDB at Center for Regulation Genomics in Barcelona, Spain (Huerta-Cepas *et al.*, 2014).

7.5 Supplementary Information

Table 29: DNMT annotation note for both VIB and CRG versions.



VIB_id (Zaichao annot)	comparison_not	aa change	CRG_id (##20160211-(	note (Feb 2016 at VIB)	RNAseq support	ncRNA	type	prot_size	exon
COPFL27g00340	no change	0	COPFL27g00460.1	strong in RNAseq	Y	no	dnmt1	1335	13
NA	NA	NA	COPFL5365g00731	created by blastn (COPFL27g00460.1), BAH_Dnmt1 detected	N	no	dnmt1 fragment | pseudo	417	4
COPFL01g07190	no change	0	COPFL01g10500.1	homolog with COPFL5521g00040.1 , strong in RNAseq	Y	no	dnmt2-Trdmt	216	4
COPFL1363g00160	no change	0	COPFL1363g00320.2	big intron (41kb) and tRNA (cytosine(38)-C(5))-methyltransferase	Y	no	dnmt2-Trdmt	360	2
COPFL5521g00040	no change	0	COPFL5521g00040.1	homolog with COPFL01g10500.1 , strong in RNAseq	Y	3 ncRNA found,	dnmt2-Trdmt	216	4
COPFL3491g00010	no change	0	COPFL3491g00010.1	truncated, only 3 exons, no RNAseq, prediction models support	N	no	dnmt3 ? | truncated	181	3
COPFL1334g08190	changed, no mor	(318vs274) 44	COPFL1334g13340.1	dnmt domain detected, prediction models support	Y-support gene model well	no	dnmt3a-like	274	5
COPFL2683g00160	no change	0	COPFL2683g00592	average alignment but it contains a dnmt domain, weak in RNAse	Y-support the other strand pa	lncRNA COPFL2	dnmt3a-like	294	6
COPFL4024g00270	changed	(288vs252) 36	COPFL4024g01371	no dnmt domain detected by supported by homologs	Y- support gene model well	lncRNA COPFL4	dnmt3a-like	252	7
COPFL5355g00312	no change	0	COPFL5355g01232	dnmt domain detected	N	no	dnmt3a-like	305	5
COPFL2672g00581	changed because	(318vs328) -10	COPFL2672g00160.1	dnmt domain detected, homologs support, first 3 exons have pre	Y- support gene model	no	dnmt3b-like	328	6
COPFL6729g00401	no change	0	COPFL6729g00891	1st exons is poorly aligned. Dnmt domain detected	Y-but very weak	lncRNA detecte	dnmt3a-like	283	5
COPFL16g00201	no change	0	COPFL16g00621	no prediction model and dnmt domain but supported by homolo	N	several lncRNA	dnmt3b-like	315	8
COPFL4019g00351	no more pseudo	(311vs333) -22	COPFL4019g00570.1	prediction models support, dnmt domain detected, homolog sup	Y-support gene model	no	dnmt3b-like	333	6
COPFL6692g00561	no change	0	COPFL6692g01781	dnmt domain detected	Y-but partially	no	dnmt3b-like	286	9
COPFL78g00010	no change	0	COPFL78g00010.1	no RNAseq, prediction models support	N	no	dnmt3b-like	319	5
COPFL35g00341	no change	0	COPFL35g00771	no prediction model and dnmt domain but supported by homolo	Y-but messy	3 lncRNA detec	dnmt3b-like	261	5
COPFL1359g00090	no change	0	COPFL1359g00070.1	dnmt domain detected, probably due to sequencing error	Y-but weak in gene model su	no	dnmt3b-like | pseudo	227	4
COPFL1338g00561	changed	(302vs317) -15	COPFL1338g01281	no dnmt domain detected buy supported by homologs	N	no	dnmt3a-like | pseudo	317	4
COPFL4015g00242	changed	(299vs291) 8	COPFL4015g01111	dnmt domain detected, first exon is only supported by one junct	Y-but partially	lncRNA detecte	dnmt3a-like | pseudo	291	5
COPFL6696g00251	changed	(287vs283) 4	COPFL6696g00851	weak in alignment, dnmt domain detected	N	lncRNA detecte	dnmt3a-like | pseudo	283	7
COPFL2696g00040	changed at the 1	(249vs229) 20	COPFL2696g00080.1	dnmt domain detected, homologs support	N	no	dnmt3b-like | pseudo	229	4
COPFL4030g00571	changed gene-wi	(213vs251) 38	COPFL4019g00610.1	dnmt domain detected, homologs support. The first two exons ar	Y-but weak in gene model (o	no	dnmt3b-like | pseudo	251	4
COPFL6693g00540	changed at the fi	(312vs375) 63	COPFL6693g01700.1	dnmt domain detected, homologs support. first 3 exons have pre	N	no	dnmt3b-like | pseudo	375	6
COPFL5366g01141	changed	(319vs296) 23	COPFL5366g01530.1	prediction models support, dnmt domain detected, probably this	Y-support gene model well	no	dnmt3b-like | pseudo	296	5
NA	NA	NA	COPFL1002g00010.2	created by blastp, dnmt domain detected, truncated	N	no	dnmt3b-like | truncated	94	2

7.6 Reference

Abeel, T., Van Parys, T., Saeys, Y., Galagan, J., and Van de Peer, Y. (2012). GenomeView: a next-generation genome browser. Nucleic acids research *40*, e12.

Arensburger, P., Megy, K., Waterhouse, R.M., Abrudan, J., Amedeo, P., Antelo, B., Bartholomay, L., Bidwell, S., Caler, E., Camara, F.*, et al.* (2010). Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics. Science *330*, 86-88.

Dong, A., Yoder, J.A., Zhang, X., Zhou, L., Bestor, T.H., and Cheng, X. (2001). Structure of human DNMT2, an enigmatic DNA methyltransferase homolog that displays denaturant-resistant binding to DNA. Nucleic acids research *29*, 439-448.

Drosophila 12 Genomes Consortium, Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W.*, et al.* (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature *450*, 203-218.

Foissac, S., Gouzy, J., Rombauts, S., Mathe, C., Amselem, J., Sterck, L., Van de Peer, Y., Rouze, P., and Schiex, T. (2008). Genome annotation in plants and fungi: EuGene as a model platform. Curr Bioinform *3*, 87-97.

Gaunt, M.W., and Miles, M.A. (2002). An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. Molecular biology and evolution *19*, 748-761.

Goll, M.G., Kirpekar, F., Maggert, K.A., Yoder, J.A., Hsieh, C.L., Zhang, X., Golic, K.G., Jacobsen, S.E., and Bestor, T.H. (2006). Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. Science *311*, 395-398.

Grbic, M. (2003). Polyembryony in parasitic wasps: evolution of a novel mode of development. Int J Dev Biol *47*, 633-642.

Grbic, M., Ode, P.J., and Strand, M.R. (1992). Sibling rivalry and brood sex ratios in polyembryonic wasps. Nature *360*, 254-256.

Grimaldi, D., and Engel, M.S. (2005). Evolution of the Insects (Cambridge Evolution Series).

Harvey, J.A., Corley, L.S., and Strand, M.R. (2000). Competition induces adaptive shifts in caste ratios of a polyembryonic wasp. Nature *406*, 183-186.

Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L.P., Marcet-Houben, M., and Gabaldón, T. (2014). PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic acids research *42*, D897-D902.

International Aphid Genomics, C. (2010). Genome sequence of the pea aphid Acyrthosiphon pisum. PLoS biology *8*, e1000313.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome research *12*, 656-664.

Lyko, F., and Maleszka, R. (2011). Insects as innovative models for functional studies of DNA methylation. Trends in genetics : TIG *27*, 127-131.

Mukherjee, K., Twyman, R.M., and Vilcinskas, A. (2015). Insects as models to study the epigenetic basis of disease. Progress in Biophysics and Molecular Biology *118*, 69-78.

Nederbragt, A.J. (2014). On the middle ground between open source and commercial software - the case of the Newbler program. Genome biology *15*, 113.

Nene, V., Wortman, J.R., Lawson, D., Haas, B., Kodira, C., Tu, Z.J., Loftus, B., Xi, Z., Megy, K., Grabherr, M.*, et al.* (2007). Genome sequence of Aedes aegypti, a major arbovirus vector. Science *316*, 1718-1723.

Okano, M., Xie, S., and Li, E. (1998). Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells. Nucleic acids research *26*, 2536-2540.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome research *19*, 1117-1123.

Smit, A., and Hubley, R. (2008-2015). RepeatModeler Open-1.0.

Sterck, L., Billiau, K., Abeel, T., Rouze, P., and Van de Peer, Y. (2012). ORCAE: online resource for community annotation of eukaryotes. Nature methods *9*, 1041.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] *Chapter 4*, Unit 4 10.

Wang, X., Wheeler, D., Avery, A., Rago, A., Choi, J.H., Colbourne, J.K., Clark, A.G., and Werren, J.H. (2013). Function and evolution of DNA methylation in Nasonia vitripennis. PLoS genetics *9*, e1003872.

Watanabe, K., Nishide, Y., Roff, D.A., Yoshimura, J., and Iwabuchi, K. (2012). Environmental and genetic controls of soldier caste in a parasitic social wasp. Scientific reports *2*, 729.

Werren, J.H., Richards, S., Desjardins, C.A., Niehuis, O., Gadau, J., Colbourne, J.K., Nasonia Genome Working, G., Werren, J.H., Richards, S., Desjardins, C.A.*, et al.* (2010). Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. Science *327*, 343-348.

Zhang, T., Luo, Y., Chen, Y., Li, X., and Yu, J. (2012). BIGrat: a repeat resolver for pyrosequencing-based re-sequencing with Newbler. BMC research notes *5*, 567.

Zhurov, V., Terzin, T., and Grbic, M. (2004). Early blastomere determines embryo proliferation and caste fate in a polyembryonic wasp. Nature *432*, 764-769.

Zhurov, V., Terzin, T., and Grbic, M. (2007). (In)discrete charm of the polyembryony: evolution of embryo cloning. Cell Mol Life Sci *64*, 2790-2798.

Zwier, M.V., Verhulst, E.C., Zwahlen, R.D., Beukeboom, L.W., and van de Zande, L. (2012). DNA methylation plays a crucial role during early Nasonia development. Insect molecular biology *21*, 129-138.

Chapter 8

8 Conclusion and perspectives

In this final Chapter, I primarily discuss the advantages and disadvantages of current NGS and coming TGS technologies. Subsequently, I quickly give an overview of state-of-the-art applications of sequencing technologies on arthropod genomes.

8.1 NGS: opportunities and challenges

NGS, starting with 454 Pyrosequencing in 2004 and followed by Illumina sequencing technology, has revolutionized genomic sequencing by reducing cost and increasing throughput exponentially over Sanger sequencing. Over the past decade, NGS has achieved great success and explosively advanced our understanding in various fields such as diagnostics, drug discovery, biomarker discovery, precision medicine, agriculture and animal research (Lander *et al.*, 2001; Yu *et al.*, 2002; International Chicken Genome Sequencing, 2004; Potato Genome Sequencing *et al.*, 2011; Brenchley *et al.*, 2012b; 2012; Olsen *et al.*, 2016). Represented by Illumina, NGS is fast advancing in producing massively unprecedented throughput data that empowers new levels of genomic possibilities. The latest released sequencers NovaSeq Series (5000 and 6000 systems), whose flow cell types are PE 2x50 bp, 2x100 bp, and 2x150 bp and runtime is within less than 48 hours (including cluster generation, sequencing, and base calling for a dual S2 flow cell run on the NovaSeq 6000 system) but the output is up to 3Tb (Illumina, 2017).

However, decoding NGS data still presents several emerging problems and challenges, concerning trade-off between effort, budget and result accuracy. First, NGS requires amplification of source DNA before sequencing, leading to amplification artifacts and biased coverage of the genome related to the chemical-physical properties of the DNA (Dohm *et al.*, 2008; Niu *et al.*, 2010). DNA damage is a pervasive cause of sequencing errors. A recent study claims that mutagenic damage accounts for the majority of the erroneous identification of variants with low to moderate (1% to 5%) frequency (Chen *et al.*, 2017). The extent of this damage directly confounds the determination of somatic variants in these data sets. Secondly, because of relatively short reads (i.e., 100-500 bp for Illumina and nearly 700 bp for 454), assembly quality is always a burning issue since genome assembly is critical to downstream bioinformatic analyses, and even further, to our understanding of evolution and genetic variation. Whole-genome assembly of large eukaryotic genomes remains problematic because of the presence of repetitive DNA (Gordon *et al.*, 2016a). Current assemblers produce a high degree of variability between output assemblies, which suggests that different tools might be particularly useful for certain read types and even the best assemblers make numerous and unexpected errors

(Salzberg *et al.*, 2012; Bradnam *et al.*, 2013). A recent chromosome-level assembly study reveals the extent of translocation and inversion polymorphisms which re-sequencing or small-scale assembly failed to detect (Zapata *et al.*, 2016). Additionally, the *de novo* genome assemblies using NGS reads can cause considerable genetic information loss (Alkan *et al.*, 2011) and the shorter the reads from NGS technologies cause the higher error rates from the relatively short insert libraries occurred (Bentley *et al.*, 2008; Wheeler *et al.*, 2008). Therefore, these high-quality assemblies must be considered in conjunction with NGS data for genomics analyses, otherwise, there would be huge errors in genomics, caused by short sequence reads (Alkan *et al.*, 2011).

To sum up, in the best genome assembly scenario, full-length chromosome level assemblies are ideally necessary, compared with short scaffold level assemblies. If required, base quality can be further improved by polishing with complementary NGS reads (Berlin *et al.*, 2015). For the remaining gaps, long-read assemblies could be paired with super-long linking information as generated by OM data or chromatin interaction maps (Schwartz *et al.*, 1993; Burton *et al.*, 2013; Kaplan and Dekker, 2013). These complementary scaffolding approaches could be used to span centromeres, resolve whole chromosomes and phase haplotypes to produce truly complete assemblies. Long reads have the capability of producing better assemblies, even at a relatively low coverage, as reported that a 10-20x Sanger assembly is better than 1,100x Illumina assembly despite the expense difference (Schatz *et al.*, 2010; Gnerre *et al.*, 2011).

8.2 TGS: the next NGS?

To compensate the shortcoming of NGS short reads, as mentioned, more cost-effective and longer-reads sequencing technologies are required. Over the past several years, TGS technologies have been creating a renaissance in high-quality genome sequencing even though it is currently still under development (Bleidorn, 2016). TGS, presented by PacBio RS SMRT and Oxford Nanopore Minion, was designed to improve accuracy, increase the length and decrease cost. Take PacBio Sequel System for example, it delivers long reads, high consensus accuracy and uniform coverage that enable more complete, accurate and contiguous assemblies for these large and complex genomes. The latest Sequel chemistry can produce over 5Gb per SMRT Cell with reduced input SMRT cell libraries. Read length ranging 10 kb to 15 kb can be routinely accomplished, with the longest reads >60 kb. Furthermore, 50% of usable reads are greater than 20 kb (Sisneros *et al.*, 2017).

Furthermore, to address another NGS problem, TGS technologies require no amplification. Meanwhile, they can reduce compositional bias and produce longer sequences, demonstrating TGS technologies unparalleled advantages (Eid *et al.*, 2009; Schadt *et al.*, 2010; Chin *et al.*, 2011). Assembling large genomes from single-molecules using TGS data, has been generally adopted in recent studies and accompanied with this, new assembly methods, including error correction and reduction of the assembly complexity (Koren *et al.*, 2013; Lee *et al.*, 2014), are also emerging and being improved (Chaisson and Tesler, 2012; Berlin *et al.*, 2015). Even the human genome assembly has been recently resolved using single-molecule sequencing (Chaisson *et al.*, 2015a).

Table 30: Characteristics of TGS technologies and three mapping platforms.

*Note**	*Technology*	*Mean Length*	*Raw Error Rate*	*Costs/GB*	*Time/GB*	*Human Metrics*
Illumina TruSeq Synthetic Long Reads (2012)	Barcoded & Amplified Synthetic long reads	3-5k bp	0.10%	~$2500*	2-3 days*	0.5M bp Haplotype phasing N50
Pacific Biosciences (2010)	Single Molecule Real Time Sequencing	10-15k bp	10-15%	~$500†	2-3 hours	26.9M bp Contig N50
Oxford Nanopore (2014)	Nanopore Sequencing	5-10k bp	10-30%	~$1000†	1-2 days	NA
BioNano Genomics	Optical mapping of fluorescent probes	100-250k bp	Fragile sites, incomplete labeling	NA	NA	31.1M bp Scaffold N50
10X Genomics	Barcoded "Read Clouds"	30-100k bp	Barcode reuse, Short read mapping	NA	NA	21.6M bp Haplotype phasing N50
Dovetail cHiCago	Chromatin mate-pairs	25-100k bp	Variable span, short read mapping	NA	NA	29.9M bp Scaffold N50

*this table is adopted and modified from Lee et al (Lee *et al.*, 2016) and all the prices subject to change, please see https://www.dugsim.net/estimate_cost for current estimates.

However, TGS is not perfect because of two defects: high cost and high error rate. Compared with Illumina Hiseq2000 ~$41/Gb and Miseq ~$502/Gb, PacBio RS is too expensive at a price of ~$2000/Gb. Again take PacBio Sequel System for example, this instrument generates reads with an average at only ~85% nucleotide accuracy and uniformly distributed errors dominated by INDELs (Chin *et al.*, 2011; Rasko *et al.*, 2011). Consequently, this low accuracy not only obscures the alignments but also complicates the downstream analyses because the pairwise difference between two reads is approximately two times of their individual error rate (Margulies *et al.*, 2005; Miller *et al.*, 2008; Koren *et al.*, 2012; Salzberg *et al.*, 2012; Goodwin *et al.*, 2015). Nevertheless, there is a great potential advantage for the TGS long reads because of recently developed and improved algorithms that overcome the limitations of high error rates and unlock its full potential for a *de novo* assembly (Koren *et al.*, 2012; Goodwin *et al.*, 2015). These algorithms and tools can improve the assemblies with fewer errors and gaps, which will drive down the expensive cost of genome sequencing. Moreover, TGS will offer more accurate genomic data for downstream analyses. In summary, TGS technologies are undergoing active improvement, especially on the high error rates. In the recent years, TGS has shown its strength, for instance, its applications in assembling large genomes and clinical genomics (Qiao *et al.*, 2016; Seo *et al.*, 2016a; Shi *et al.*, 2016; Avni *et al.*, 2017; Merker *et al.*, 2017; Zhao *et al.*, 2017).

Table 31: A general comparison between NGS and TGS.

Generations	**Pros**	**Cons**
NGS technology	High-throughput High accuracy Less expensive Fast speed	Short reads Amplification and synthesis Require better platform and algorithm for assembly
TGS technology	Long reads Portable and easy	High error rates Relatively expensive

8.3 Arthropod genomics

Arthropods, as the largest genus of terrestrial animals on Earth, have revealed the biological diversity and offered us valuable biological materials. To date, an estimated number of arthropod species up to 10 million, and probably they account for over 80% of all known living animal species (Odegard, 2000). Currently, we only have characterized of the tiny tip of the iceberg of arthropod biology. The phylogenomic analysis of nuclear protein-coding sequences revealed arthropod relationships and offers insight into the arthropod evolution (Regier *et al.*, 2010).

In the recent years, more and more arthropod genomes are being decoded. The i5k Initiative, also known as the 5k Insect Genome Project, was launched in 2011, was aiming to sequence the genomes of 5,000 insects and other arthropods over the next five years (Robinson *et al.*, 2011a). The project has not officially finished yet, but many other important arthropods have been sequenced and released such as Centipede (Chipman *et al.*, 2014), Hessian fly (Zhao *et al.*, 2015) and Asian long-horned beetle (McKenna *et al.*, 2016). Other genomes are in working progress such as Turnip Sawfly and Water Strider (https://www.hgsc.bcm.edu/arthropods). These arthropod genomes will offer us more opportunities to insecticide resistance, for developing new pesticides, for understanding transmission of disease, and for agricultural pest control studies in the future.

8.4 Perspectives

Now is a watershed moment in genomics. In 2005, the editor of *The Evolution of Genome* T. Ryan Gregory stated: “the growth of genomics shows no sign of slowing - indeed, all indications suggest it will continue to accelerate for the foreseeable future” (Gregory, 2005). Over the last decade, we have witnessed so many achievements in the field of genomics and bioinformatics. Undoubtedly, genomics and bioinformatics will provide even more exciting and unexpected findings in the next decade. By great improvement and advancement of NGS and future TGS studies, we can be sure that the next stage promises to be another era of extraordinary biological discovery.

8.5 Reference

Alkan, C., Sajjadian, S., and Eichler, E.E. (2011). Limitations of next-generation genome sequence assembly. Nature methods *8*, 61-65.

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S.O., Gundlach, H., Hale, I., Mascher, M., Spannagl, M., Wiebe, K*., et al.* (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science *357*, 93-97.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R*., et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53-59.

Berlin, K., Koren, S., Chin, C.S., Drake, J.P., Landolin, J.M., and Phillippy, A.M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nature biotechnology *33*, 623-630.

Bleidorn, C. (2016). Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. Systematics and Biodiversity *14*, 1-8.

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J.A., Chapuis, G., Chikhi, R*., et al.* (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience *2*, 10.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G.L.A., D/'Amore, R., Allen, A.M., McKenzie, N., Kramer, M., Kerhornou, A., Bolser, D*., et al.* (2012b). Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature *491*, 705-710.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nature biotechnology *31*, 1119-1125.

Chaisson, M.J., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M*., et al.* (2015a). Resolving the complexity of the human genome using single-molecule sequencing. Nature *517*, 608-611.

Chaisson, M.J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC bioinformatics *13*, 238.

Chen, L., Liu, P., Evans, T.C., Jr., and Ettwiller, L.M. (2017). DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. Science *355*, 752-756.

Chin, C.S., Sorenson, J., Harris, J.B., Robins, W.P., Charles, R.C., Jean-Charles, R.R., Bullard, J., Webster, D.R., Kasarskis, A., Peluso, P.*, et al.* (2011). The origin of the Haitian cholera outbreak strain. N Engl J Med *364*, 33-42.

Chipman, A.D., Ferrier, D.E., Brena, C., Qu, J., Hughes, D.S., Schroder, R., Torres-Oliva, M., Znassi, N., Jiang, H., Almeida, F.C.*, et al.* (2014). The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede Strigamia maritima. PLoS biology *12*, e1002005.

Consortium, T.G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature *485*, 635-641.

Dohm, J.C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic acids research *36*, e105.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B.*, et al.* (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133-138.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S.*, et al.* (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences of the United States of America *108*, 1513-1518.

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M.C., and McCombie, W.R. (2015). Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome research *25*, 1750-1756.

Gordon, D., Huddleston, J., Chaisson, M.J., Hill, C.M., Kronenberg, Z.N., Munson, K.M., Malig, M., Raja, A., Fiddes, I., Hillier, L.W.*, et al.* (2016a). Long-read sequence assembly of the gorilla genome. Science *352*, aae0344.

Gregory, T.R. (2005). The Evolution of the Genome (Elsevier Inc.).

Illumina, I. (2017). NovaSeq System Specifications.

International Chicken Genome Sequencing, C. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature *432*, 695-716.

Kaplan, N., and Dekker, J. (2013). High-throughput genome scaffolding from in vivo DNA interaction frequency. Nature biotechnology *31*, 1143-1147.

Koren, S., Harhay, G.P., Smith, T.P., Bono, J.L., Harhay, D.M., McVey, S.D., Radune, D., Bergman, N.H., and Phillippy, A.M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome biology *14*, R101.

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D., *et al.* (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature biotechnology *30*, 693-700.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lee, H., Gurtowski, J., Yoo, S., Marcus, S., McCombie, W.R., and Schatz, M. (2014). Error correction and assembly complexity of single molecule sequencing reads. bioRxiv.

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., McCombie, W.R., and Schatz, M. (2016). Third-generation sequencing and the future of genomics. bioRxiv.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376-380.

McKenna, D.D., Scully, E.D., Pauchet, Y., Hoover, K., Kirsch, R., Geib, S.M., Mitchell, R.F., Waterhouse, R.M., Ahn, S.J., Arsala, D., *et al.* (2016). Genome of the Asian longhorned beetle (Anoplophora glabripennis), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface. Genome biology *17*, 227.

Merker, J.D., Wenger, A.M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., Waggott, D., Utiramerur, S., Hou, Y., Smith, K.S., *et al.* (2017). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genetics in medicine : official journal of the American College of Medical Genetics.

Miller, J.R., Delcher, A.L., Koren, S., Venter, E., Walenz, B.P., Brownley, A., Johnson, J., Li, K., Mobarry, C., and Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. Bioinformatics *24*, 2818-2824.

Niu, B., Fu, L., Sun, S., and Li, W. (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC bioinformatics *11*, 187.

Odegard, F. (2000). How many species of arthropods? Erwin's estimate revised. Biological Journal of the Linnean Society *71*.

Olsen, J.L., Rouze, P., Verhelst, B., Lin, Y.C., Bayer, T., Collen, J., Dattolo, E., De Paoli, E., Dittami, S., Maumus, F., *et al.* (2016). The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea. Nature *530*, 331-335.

Potato Genome Sequencing, C., Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., *et al.* (2011). Genome sequence and analysis of the tuber crop potato. Nature *475*, 189-195.

Qiao, W., Yang, Y., Sebra, R., Mendiratta, G., Gaedigk, A., Desnick, R.J., and Scott, S.A. (2016). Long-Read Single Molecule Real-Time Full Gene Sequencing of Cytochrome P450-2D6. Hum Mutat *37*, 315-323.

Rasko, D.A., Webster, D.R., Sahl, J.W., Bashir, A., Boisen, N., Scheutz, F., Paxinos, E.E., Sebra, R., Chin, C.S., Iliopoulos, D*., et al.* (2011). Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med *365*, 709-717.

Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W., and Cunningham, C.W. (2010). Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature *463*, 1079-1083.

Robinson, G.E., Hackett, K.J., Purcell-Miramontes, M., Brown, S.J., Evans, J.D., Goldsmith, M.R., Lawson, D., Okamuro, J., Robertson, H.M., and Schneider, D.J. (2011a). Creating a Buzz About Insect Genomes. Science *331*, 1386-1386.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., Roberts, M*., et al.* (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome research *22*, 557-567.

Schadt, E.E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. Human molecular genetics *19*, R227-240.

Schatz, M.C., Delcher, A.L., and Salzberg, S.L. (2010). Assembly of large genomes using second-generation sequencing. Genome research *20*, 1165-1173.

Schwartz, D.C., Li, X., Hernandez, L.I., Ramnarain, S.P., Huff, E.J., and Wang, Y.K. (1993). Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. Science *262*, 110-114.

Seo, J.S., Rhie, A., Kim, J., Lee, S., Sohn, M.H., Kim, C.U., Hastie, A., Cao, H., Yun, J.Y., Kim, J*., et al.* (2016a). De novo assembly and phasing of a Korean human genome. Nature *538*, 243-247.

Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., Fu, A., Li, Q., Li, N., Gong, S*., et al.* (2016). Long-read sequencing and de novo assembly of a Chinese genome. Nature communications *7*, 12065.

Sisneros, N., Chakraborty, S., Kingan, S., Hall, R., Wilson, J., Lambert, C., Eng, K., Hatas, E., and Baybayan, P. (2017). Best Practices for Whole Genome Sequencing Using the Sequel System.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T*., et al.* (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872-876.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., *et al.* (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science *296*, 79-92.

Zapata, L., Ding, J., Willing, E.M., Hartwig, B., Bezdan, D., Jiao, W.B., Patel, V., Velikkakam James, G., Koornneef, M., Ossowski, S., *et al.* (2016). Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. Proceedings of the National Academy of Sciences of the United States of America *113*, E4052-4060.

Zhao, C., Escalante, L.N., Chen, H., Benatti, T.R., Qu, J., Chellapilla, S., Waterhouse, R.M., Wheeler, D., Andersson, M.N., Bao, R., *et al.* (2015). A massive expansion of effector genes underlies gall-formation in the wheat pest Mayetiola destructor. Curr Biol *25*, 613-620.

Zhao, L., Deng, L., Li, G., Jin, H., Cai, J., Shang, H., Li, Y., Yang, A.X., Chen, F., Zhao, Z., *et al.* (2017). Resequencing the Escherichia coli genome by GenoCare single molecule sequencing platform. bioRxiv.

Appendices

NA

Links

Website	**Links**
UniProt	http://www.uniprot.org/
SwissProt	http://web.expasy.org/docs/swiss-prot_guideline.html
NCBI	https://www.ncbi.nlm.nih.gov/
UCSC	https://genome.ucsc.edu/
ORCAE	http://bioinformatics.psb.ugent.be/orcae/
GENCODE	http://www.gencodegenes.org
Ensembl	http://www.ensembl.org
RefSeq	http://www.ncbi.nlm.nih.gov/refseq
Giga DB	http://gigadb.org/
Animal Genome Size Database	http://www.genomesize.com
BAMStats	http://bamstats.sourceforge.net/
i5k	http://arthropodgenomes.org/wiki/i5K

Supplementary Tables

Table 32: Published animal genomes.

Latin Name#	**Genome Size****	**Journal**	Published Date *
Caenorhabditis elegans	97Mb	Science	199812
Drosophila melanogaster	120Mb	Science	200003
Homo sapiens	3.2Gb	Nature	200102
Oikopleura dioica	65Mb	Science	200112
Anopheles gambiae	280Mb	Science	200201
Takifugu rubripes	333Mb	Science	200208
Ciona intestinalis	150Mb	Science	200212
Mus musculus	2.5Gb	Nature	200212
Fugu rubripes	380Mb	Science	200212
Caenorhabditis briggsae	104Mb	PloS Biology	200311
Tetraodon nigroviridis	340Mb	Nature	200401
Rattus norvegicus	2.75Gb	Nature	200404
Gallus gallus	1.05Gb	Nature	200412
Gallus sonneratii	1.06Gb	Nature	200412
Bombyx mori	428.7Mb	Science	200412
Drosophila pseudoobscura	139M	Genome research	200501
Trypanosoma cruzi	67Mb	Science	200507
Pan troglodytes	2.7Gb	Nature	200509
Canis familiaris	2.5Gb	Nature	200512
Apis mellifera	236Mb	Nature	200601
Strongylocentrotus purpuratus	1Gb	Science	200611
Caenorhabditis remanei	135Mb	Trends in Genetics	200703
Ciona savignyi	174Mb	Genome Biology	200703
Callorhinchus milii	0.91Gb	PloS Biology	200704
Macaca mulatta	2.87Gb	Science	200704
Monodelphis domestica	3.4Gb	Nature	200705
Aedes aegypti	1376Mb	Science	200706
Oryzias latipes	700Mb	Nature	200706
Nematostella vectensis	357Mb	Science	200707
Brugia malayi	90Mb	Science	200707
12 Drosophila (10 new species)	111Mb~176Mb	Nature	200711
Felis catus	2.7Gb	Genome Research	200711
Tribolium castaneum	204Mb	Nature	200804
Ornithorhynchus anatinus	1.84Gb	Nature	200805
Branchiostoma floridae	520Mb	Nature	200806
Trichoplax adhaerens	104Mb	Nature	200808
Meloidogyne incognita	86Mb	Nature Biotechnology	200808
Pristionchus pacificus	169Mb	Nature Genetics	200809
Mammuthus primigenius	4.7Gb	Nature	200811
Bos Taurus	2.87Gb	Science	200904

Schistosoma mansoni	360Mb	Nature	200907
Schistosoma japonicum	397Mb	Nature	200907
Equus caballus	2.7Gb	Science	200911
Ailuropoda melanoleura	2.25Gb	Nature	201001
Nasonia vitripennis, N.giraulti, N.longicornis	295Mb	Science	201001
Acyrthosiphon pisum.	517Mb	PLoS Biol.	201002
Hydra	1.05Gb	Nature	201003
Taeniopygia guttata	1.2Gb	Nature	201004
Xenopus tropicalis	1.7Gb	Science	201004
Pediculus humanus	110Mb	PNAS	201007
Amphimedon queenslandica	190Mb	Nature	201008
Camponotus floridanus, Harpegnathos saltator	240Mb, 330Mb	Science	201008
Meleagris gallopavo	1.1Gb	PLoS Biol.	201009
Culex quinquefasciatus	540Mb	Nature	201010
Caenorhabditis angaria	80Mb	Genome Research	201010
Oikopleura	148Mb	Science	201011
Linepithema humile	250.8Mb	PNAS	201101
Pogonomyrmex barbatus	250?284	PNAS	201101
Pongo abelii, Pongo pygmaeus	3.09Gb	Nature	201101
Solenopsis invicta	484.2Mb	PNAS	201101
Daphnia pulex	200Mb	Science	201102
Atta cephalotes	300Mb	PLoS Genetics	201102
Trichinella spiralis	64Mb	Nature Genetics	201102
Sarcophilus harrisii	3.3Gb	PNAS	201106
Acromyrmex echinatior	313Mb	Genome Research	201106
Python molurus bivittatus	1.4Gb	Genome Biology	201107
Acropora digitifera	420Mb	Nature	201107
Macropus eugenii	2.9Gb	Genome Biology	201108
Gadus morhua	830Mb	Nature	201108
Anolis carolinensis	1.78G	Nature	201109
Bursaphelenchus xylophilus	74.5Mb	PloS pathoggens	201109
Pteropus vampyrus	1.84Gb	Nature	201110
Tursiops truncatus	2.3Gb	Nature	201110
Clonorchis sinensis	516M	Genome Biology	201110
Heterocephalus glaber	2.6G	Nature	201111
Macaca fascicularis, Macaca mulattalasiotaNature	2.84 Gb, 2.85Gb	Nature Biotechnology	201111
Ascaris suum	272M	Nature	201111
Danaus plexippus	273M	Cell	201111
Tetranychus uritcae	90M	Nature	201111
Ictalurus punctatus	1G	BMC Genomics	201112
Daubentonia madagascariensis	3G	Genome Biology and Evolution	201112
Crocodylus siamensis	2.5G	Genome Biology	201201
Schistosoma haematobium	385M	Nature Genetics	201201
Pinctada fucata	1150M	DNA Research	201202

Gorilla gorilla	3.04G	Nature	201203
Gasterosteus aculeatus	463M	Nature	201204
Heliconius melpomene	269M	Nature	201205
Pan paniscus	2.7G	Nature	201206
Melopsittacus undulatus	1.2G	Nature Biotechnology	201207
Ursus maritimus	2.53G	PNAS	201207
Bos grunniens	2.66G	Nature Genetics	201207
Geospiza fortis	1.07Gb	Giga Science	201208
Plasmodium cynomolgi	26.2Mb	Nature Genetics	201208
Plasmodium vivax	28-29Mb	Nature Genetics	201208
Crassostrea gigas	559Mb	Nature	201209
Ficedula albicollis	1.13Gb	Nature	201210
Drosophila mauritiana MS17	113.3Mb	Genome Research	201210
Ficedula albicollis, Ficedula hypoleuca	1.1Gb	Nature	201211
Camelus bactrianus	2.38Gb	Nature Comm	201211
Sus scrofa (Wuzhishan)	2.64Gb	Giga Science	201211
Sus scrofa (Dormastic)	2.6Gb	Nature	201211
Dirofilaria immitis	84.2Mb	FASEB journal	201211
Pteropus alecto, Myotis davidii	2.00Gb, 1.94Gb	Science	201212
Capra hircus	2.92G	Nature Biotechnology	201212
Lottia gigantea, Capitella teleta, Helobdella robusta	348Mb,324Mb,228Mb	Nature	201212
Columba livia	1.3Gb	Science	201301
Plutella xylostella	343Mb	Nature Genetics	201301
Tupaia belangeri	3.2Gb	Nature Comm	201302
Petromyzon marinus	816Mb	Nature Genetics	201302
Pseudopodoces humilis	1.1Gb	Genome Biology	201303
Falco peregrinus, Falco cherrug	1.2Gb	Nature	201303
Camelus bactrianus	1.6Gb	Journal of Heredity	201303
Echinococcus multilocularis, E. granulosus, Taenia solium, Hymenolepis microstoma	115-141Mb	Nature	201303
Chrysemys pictabellii	2.59Gb	Genome Biology	201303
Chrysemys picta bellii	2.59Gb	Genome Biology	201303
Dendroctonus ponderosae (Hopkins)	208Mb	Genome Biology	201303
Xiphophorus maculatus	750-950Mb	Nature Genetics	201303
Loa loa	91.4Mb	Nature Genetics	201303
Danio rerio	1.4Gb	Nature	201304
Pelodiscus sinensis	2.22Gb	Nature Genetics	201304
Chelonia mydas	2.24Gb	Nature Genetics	201304
Latimeria chalumnae	2.86Gb	Nature	201304
Pelodiscus sinensis, Chelonia mydas	2.Gb,2.24Gb	Nature Genetics	201304
Ara macao	1.11-1.16G bp	PLoS ONE	201305
Pantholops hodgsonii	2.75Gb	Nature Comm	201305

Parus humilis	1.08Gb	Nature Comm	201306
Anas platyrhynchos	1.2Gb	Nature Genetics	201306
Anopheles darlingi	201Mb	Nucleic Acids Research	201306
Thunnus orientalis	800Mb	PNAS	201306
Parus humilis	1.08G	Nature Comm	201307
Adineta vaga	244Mb	Nature	201307
Heterorhabditis bacteriophora	80Mb	PLoS ONE	201307
Cricetulus griseus	2.33Gb	Nature Biotechnology	201308
Alligator sinensis	2.3Gb	Cell research	201308
Myotis brandtii	2Gb	Nature Comm	201308
Haemonchus contortus	320Mb	Genome Biology	201308
Panthera uncia	108Gb	Nature Comm	201309
Echinococcus granulosus	151.6Mb	Nature Genetics	201309
Panthera tigris	2.4G	Nature Comm	201309
Panthera tigris altaica	203Gb/84Gb	Nature Comm	201309
Panthera leo krugeri	84Gb	Nature Comm	201309
Panthera tigris tigris	86Gb	Nature Comm	201309
Panthera leo	98Gb	Nature Comm	201309
Mesobuthus martensii	1.3G	Nature Comm	201310
Megaderma lyra	2Gb	Nature	201310
Pteronotus parnellii	2Gb	Nature	201310
Eidolon helvum	2Gb	Nature	201310
Rhinolophus ferrumequinum	2Gb	Nature	201310
Balaenoptera acutorostrata	2.44G	Nature Genetics	201311
Balaenoptera physalus	2.44Gb	Nature Genetics	201311
Ophiophagus hannah	1.66Gb	PNAS	201312
Reticulomyxa filosa	1.6Gb	Current Biology	201312
Mnemiopsis leidyi	2.5Gb	Science	201312
Romanomermis culicivorax	270Mb	BMC Genomics	201312
Necator americanus	244Mb	Nature Genetics	201401
Cerapachys biroi	214Mb	Current Biology	201402
Tetrao tetrix	1.02Gb	BMC Genomics	201403
Neocaridina denticulata	1.2Gb	Marine Drugs	201403
Globodera pallida	124Mb	Genome Biology	201403
Meloidogyne hapla	53Mb	Genome Biology	201403
Oncorhynchus mykiss	1.9Gb	Nature Comm	201404
Ursus maritimus	2.25Gb	Cell	201405
Stegodyphus mimosarum	2.55Gb	Nature Comm	201405
Limulus polyphemus	2.7Gb	GigaScience	201405
Acanthoscurria geniculata	6.5Gb	Nature Comm	201405
Pleurobrachia bachei	156Mb	Nature	201406
Electrophorus electricus	533Mb	Science	201406
Cryptobranchus alleganiensis	55Gb	Genome Biology and Evolution	201406
Trichuris trichiura	75Mb	Nature Genetics	201406
Trichuris suis	81Mb/76Mb	Nature Genetics	201406
Trichuris muris	85Mb	Nature Genetics	201406

Callithrix jacchus	2.26Gb	Nature Genetics	201407
Opisthorchis viverrini	634.5Mb	Nature Comm	201407
Esox lucius	824Mb	PLoS ONE	201407
Cyprinus carpio	1.83Gb	Nature Genetics	201409
Chironomus tentans	200Mb	BMC Genomics	201409
Musca domestica	691Mb	Genome Biology	201410
Mustela putorious furo	1.83Gb	Nature Biotechnology	201411
Strigamia maritima	290Mb	Genome Biology	201411
Rhinopithecus roxellana	3Gb	Nature genetics	201411
Acanthisitta chloris	1.05Gb	Science	201412
Tinamus guttatus	1.05Gb	Science	201412
Merops nubicus	1.06Gb	Science	201412
Nestor notabilis	1.06Gb	Science	201412
Pterocles gutturalis	1.07Gb	Science	201412
Buceros rhinoceros	1.08Gb	Science	201412
Colius striatus	1.08Gb	Science	201412
Apaloderma vittatum	1.08Gb	Science	201412
Chlamydotis macqueenii	1.09Gb	Science	201412
Manacus vitellinus	1.12Gb	Science	201412
Haliaeetus albicilla	1.14Gb	Science	201412
Balearica regulorum gibbericeps	1.14Gb	Science	201412
Opisthocomus hoazin	1.14Gb	Science	201412
Phoenicopterus ruber	1.14Gb	Science	201412
Fulmarus glacialis	1.14Gb	Science	201412
Tyto alba	1.14Gb	Science	201412
Antrostomus carolinensis	1.15Gb	Science	201412
Cariama cristata	1.15Gb	Science	201412
Cuculus canorus	1.15Gb	Science	201412
Gavia stellata	1.15Gb	Science	201412
Leptosomus discolor	1.15Gb	Science	201412
Podiceps cristatus	1.15Gb	Science	201412
Phalacrocorax carbo	1.15Gb	Science	201412
Phaethon lepturus	1.16Gb	Science	201412
Cathartes aura	1.17Gb	Science	201412
Tauraco erythrolophus	1.17Gb	Science	201412
Pelecanus crispus	1.17Gb	Science	201412
Picoides pubescens	1.17Gb	Science	201412
Chaetura pelagica	1.1Gb	Science	201412
Eurypyga helias	1.1Gb	Science	201412
Mesitornis unicolor	1.1Gb	Science	201412
Calypte anna	1.1Gb	Science	201412
Struthio camelus	1.23Gb	Science	201412
Pygoscelis adeliae	1.25Gb	Giga Science	201412
Corvus brachyrhynchos	1.26Gb	Science	201412
Charadrius vociferus	1.2Gb	Science	201412
Egretta garzetta	1.2Gb	Science	201412
Haliaeetus leucocephalus	1.4Gb	Science	201412

Nipponia nippon	1.6Gb	Science	201412
Alligator mississippiensis, Crocodylus porosus, Gavialis gangeticus	2.17Gb, 2.12Gb, 2.88Gb	Science	201412
Boleophthalmus pectinirostris	827Mb	Nature Comm	201412
Aptenodytes forsteri	1.39Gb	Giga Science	201413
Serinus canaria	1.3Gb	Genome Biology	201501
Aedes albopictus	1.967Gb	PNAS	201501
Balaena mysticetus	2.87Gb	Cell report	201501
16 Anopheles mosquitoes	134Mb-375Mb	Science	201502
Toxocara canis	317Mb	Nature Comm	201502
Papilio glaucus	376Mb	Cell Reports	201502
Papilio glaucus	376Mb	Cell report	201502
Nanorana parkeri	2.3Gb	PNAS	201503
Ancylostoma ceylanicum	313Mb	Nature Genetics	201503
Ophiosaurus gracilis	1.78Gb	GigaScience	201504
Bombus terrestris, Bombus impatiens	249Mb, 247Mb	Genome Biology	201504
Ctenopharyngodon idellus	0.9Gb, 1.07Gb	Nature Genetics	201505
Anser cygnoides	1.12Gb	Genome Biology	201505
Apis mellifera, Apis florea, Eufriesea mexicana, Bombus terrestris, Bombus impatiens, Melipona quadrifasciata, Habropoda laboriosa, Megachile rotundata, Lasioglossum albipes, Dufourea novaeangliae	234Mb-1Gb	Science	201505
Apteryx mantelli	1.59Gb	Genome Biology	201507
Plasmodium falciparum	23Mb	BMC Genomics	201507
Octopus bimaculoides	2.7Gb	Nature	201508
Equus przewalskii	2.36Gb	Scientific Report	201509
Aiptasia pallida	260Mb	PNAS	201509
Eisenia fetida	1.05Gb	Genome Biology and Evolution	201510
Aegypius monachus	1.13Gb	Genome Biology	201510
Philomachus pugnax	1.23Gb	Nature Genetics	201511
Saccoglossus kowalevskii, Ptychodera flava	1Gb	Nature	201511
Saccaglossus kowalevskii	1Gb	Nature	201511
Gekko japonicus	2.55Gb	Nature Comm	201511
Hypsibius dujardini	212.3Mb	PNAS	201511
Nothobranchius furzeri	1Gb	Cell	201512
Panthera pardus	2.45Gb	Genome Biology	201512
Kudoa iwatai	22.5Mb	PNAS	201512
Rhodnius prolixus	702Mb	PNAS	201512
Arachis duranensis and Arachis ipaensis	1.2Gb, 1.5Gb	Nature Genetics	201602
Ixodes scapularis	2.1Gb	Nature Comm	201602

Cimex lectularius	650Mb	Nature Comm	201602
Lepisosteus oculatus	945Mb	Nature Genetics	201603
Gorilla gorilla gorilla	3.1Gb	Science	201604
Salmo salar	2.97Gb	Nature	201605
Giraffa camelopardalis. tippelskirchi	2.9Gb	Nature Comm	201605
Okapi johnstoni	3.3Gb	Nature Comm	201605
Ictalurus punctatus	783Mb	Nature Comm	201606
Mola mola	730Mb	GigaScience	201609
Deinagkistrodon acutus	1.43 Gb	Nature Communications	201610
Phormia regina	550Mb	BMC Genomics	201610
Panthera pardus	2.45Gb	Genome Biology	201611
Onchocerca volvulus	97Mb	Nature Microbiology	201611
Anoplophora gladbripennis	981Mb	Genome Biology	201611
Hippocampus comes	502Mb	Nature	201612
Paralichthys olivaceus	546M	Nature Genetics	201612
Castor canadensis	2.486Gb	G3: Genes, Genomes, Genetics	201701
Bathymodiolus platifrons	1.64Gb	Nature Ecology & Evolution	201704
Modiolus philippinarum	2.38Gb	Nature Ecology & Evolution	201704
Biomphalaria glabrata	916 Mb	Nature Communications	201705
Gopherus agassizii	2.4Gb	PloS ONE	201705

# we tried to include all the published animal gneomes as extensive as possible and we apologize if any important genomes are missed in this list.
* published date is either online date or paper-version date (updated 2017.06).
** data resource: NCBI, Google Scholar, Giga DB, related-journals.

Table 33: The arthropod genome datasets used in this thesis.

Name	Genome Size (Mb) *	Gene Num	Gene Density (per Mb)	Data source and date
Stegodyphus mimosarum	2550	27235	11	http://www.ncbi.nlm.nih.gov/protein 20160301
Ixodes scapularis	1770	20486	12	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Aedes aegypti	1383	17156	12	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Limulus polyphemus	1830	23660	13	http://www.ncbi.nlm.nih.gov/protein/?term=Limulus+polyphemus 20160301
Mesobuthus martensii	1128	32016	28	http://lifecenter.sgst.cn/main/en/scorpion.jsp 20160301
Culex quinquefasciatus	579	19032	33	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Bombyx mori	398	14623	37	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Solenopsis invicta	396	16569	42	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Plutella xylostella	394	18073	46	http://iae.fafu.edu.cn/DBM/download.php Protein sequences of OGSv1.0 20160229
Heliconius melpomene	269	12829	48	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Dendroctonus ponderosae	253	13457	53	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Atta cephalotes	317	18093	57	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Nasonia vitripennis	296	17174	58	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Danaus plexippus	273	16254	60	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Anopheles gambiae	236	14697	62	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Apis mellifera	245	15314	63	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Acyrthosiphon pisum	542	36195	67	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Anopheles darlingi	137	10457	76	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Tribolium castaneum	210	16526	79	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Strigamia maritima	176	15008	85	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Hypsibius dujardini	252	23021	91	http://badger.bio.ed.ac.uk/H_dujardini/home/download peptide Version 2.3.1 , 20160229 (outgroup)
Pediculus	110	10788	98	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ ,

humanus				20140724
Daphnia pulex	197	30611	155	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Tetranychus lintearius	89	15028	169	NA
Tetranychus evansi	91	15376	169	NA
Brevipalpus yothersi Brazil	72	12492	174	NA
Drosophila melanogaster	149	26950	181	ftp://ftp.ensemblgenomes.org/pub/release-23/metazoa/fasta/ , 20140724
Brevipalpus californicus uninfected	66	12476	189	NA
Brevipalpus yothersi Amsterdam	71	13448	189	NA
Brevipalpus californicus infected	66	12537	190	NA
Brevipalpus papayensis	67	12750	190	NA
Tetranychus urticae	90	19042	212	http://bioinformatics.psb.ugent.be/ 20160229

* these data are either from NCBI or original publication.

Supplementary Figures

NA

Supplementary Protocol

Protocol for the preparation of HMW DNA from spider mite eggs:

1. Collect about 0.5 mL of spider mite eggs, add a small volume of PBS to make a pipette-able slurry.
2. Transfer to Dounce homogenizer, add 5-10 mL of DHBS and homogenized with 5-7 strokes of pestle A then 5-7 strokes of pestle B.
3. Centrifuge suspension at 200 RCF, 5 minutes, 4° in 50 mL Falcon tube to pellet large pieces. Transfer supernatant to clean 50 mL tube. Try to avoid large bits by leaving some solution on top of the pellet.
4. Centrifuge at 1000 RCF, 15 minutes, 4° to pellet cells.
5. Discard supernatant and gently re-suspend pellet in 1 mL of DHBS.
6. Aliquot into batches of 0.4 ml. At this stage, you can check how much useful material is retained. See notes.
7. Briefly warm suspension to 37°. Add 0.4 mL of LDS, mix gently and incubate at 37° for 30 minutes with occasional gentle mixing.
8. Add 0.2 mL of 2.5% low-melting-point agarose in DHB at 37°, mix well and dispense into 100 ul plug molds. Use wide bore tips for mixing and aliquoting or cut a regular tip.
9. Cool plugs on ice for 20 minutes or until solid.
10. Push 20 plugs into 45 mL of LDS in 50 mL Falcon tube.
11. Incubate on a rocker at 37° for 1 hour.
12. Replace LDS solution and incubate on a rocker at 37° for 1 hour.
13. Replace LDS solution and incubate on a rocker at 37° overnight.
14. Replace LDS with 25 mL of 0.2X NDS with Pro-K. Incubate on a rocker at 50° for 24 hours.
15. Replace NDS + Pro-K solution with 45 mL 0.2X NDS and incubate on a rocker at room temperature for 2-4 hours.
16. Equilibrate plugs with 50 mm EDTA. 5 washes of 20-30 minutes each at room temperature.
17. Store plugs in 50 mm EDTA at 4° for up to several months.

Solutions:

DHB, DNA Homogenization Buffer:
0.1 M NaCl
10 mM EDTA
10 mM Tris-HCl, p H8.0
filter sterilize, store at 4°C.

DHBS:
DHB with 0.2 M sucrose

LDS:
1% (w/v) LiDS (lauryl sulfate, lithium salt)
10 mM Tris-HCl, pH 8.0
100 mM EDTA, pH 8.0
Filter, store at room temperature.

NDS, 1X:
0.5 M EDTA, sodium salt
10 mM Tris base
1% (w/v) N-lauroylsarcosine, sodium salt
Combine EDTA and Tris base in dH20. Adjust to a pH greater than 8.0 with solid NaOH pellets. Add N-lauroylsarcosine. Adjust pH to 9.5 with concentrated NaOH. Filter and store at room temperature.

NDS with Pro-K:
20 uL of 20 mg/mL Pro-K per 1 mL of 0.2X NDS.

Notes:

To check whether the good amount of material was collected take 20 uL of suspension and add 20 mL of LDS. Mix by pipetting and observe viscosity of lysate. You should get a viscous lysate that can be pooled into about 0.5-1 cm thread/column.

Low-melting-point agarose:
SeqPlaque Low Melting Temperature Agarose (Lonza, catalog number 50101).

Disposable plug molds:
Bio-Rad catalog number 170-3713

NGS Term Box

Draft genome is at some points useful to perform certain analyses, even though it possibly has short scaffold N50 and low genome. However, it must meet the minimum submission requirement to a public database.

Complete genome, despite a few gaps, usually reflects high genome coverage (>90%) with high accuracy and long N50. These complete genomes usually have a completely continuous representation and no further sequencing needs to be done.

Finished genome has a complete coverage (>99%) and each base in the genome has a very high quality.

***De novo* sequencing** typically accomplished by assembling genomic reads into scaffolds without any prior knowledge of the genomic sequence and therefore, the genome needs to be assembled from scratch.

Resequencing is to re-sequence a known genome with by mapping reads to the reference sequence.

Genome assembly is the computational reconstruction of a long genomic sequence from small sequence reads.

Genome annotation is to find gene structures in assembled genomic sequences and predict these gene functional descriptions. Generally, the annotation is synonymous to prediction. However, in this review, annotation represents both structural prediction and functional prediction.

Single-end (SE) is a read sequenced from only one end without any inserts.

Pair-end (PE) is a paired read (read1 and read2) sequenced at both ends of a single molecule with an insert of (100 bp-500 bp).

Mate-pair (MP) also consists of a paired read but usually has longer insert (2-20 kb) than PE by circularized molecule via an internal adapter.

Depth coverage means the number that one nucleotide locus is covered by reads. Most genome papers use coverage to represent depth coverage.

Breadth coverage is the percentage that the region was covered by all reads across the whole genomic assembly.

Contig is a gap-free sequence assembled from DNA reads.

Scaffold is a DNA sequence that concatenated by organized contigs and gaps.

Scaffolding is a process of concatenating sorted contigs into scaffolds, using gaps as bridges.

Gapfilling is a process of filling gaps by aligning reads back to scaffolds.

N50 size of contigs or scaffolds was calculated by sorting all sequences and then adding the lengths from the longest to the shortest until the summed length exceeded 50% of the Total length of all sequences.

L50 is the number of N50 contig or scaffold.

K-mer is a small string chopped from reads for DBG graph, which normally is set up as an odd number from 27 to 63 because an odd number avoids palindrome sequences.

Repeats and TE: Repeats are a group of repetitive elements dispersed in a genome. Repeats can be genes but reversely not. TE is a type of repeats which can transpose sequences (with flanking genes) from one locus to another locus across a whole genome.

Mapping usually represents mapping reads or contigs or scaffolds to reference genomic sequences.

Alignment means sequences comparison, mostly, it is irrelevant with genomic reads.

Frozen assembly is a dataset of genomic sequences that no more assembly needs to be done and thus set up as a final genome assembly.

Curriculum Vitae

Name: Zaichao Zhang

Post-secondary Education and Degrees:

The University of Western Ontario
2013-2017 Dual-Degree Ph.D. in Biology, London Ontario Canada

Ghent University & Center for Plant Systems Biology VIB
2014-2017 Dual-Degree Sc.D. in Bioinformatics, Gent Belgium

University of Oxford
Department for Continuing Education
2015, Certificate in Effective Writing for Life Sciences Research, Oxford, UK

Beijing Institute of Genomics
& University of Chinese Academy of Sciences
2009-2012 M.Sc. in Bioinformatics, Beijing China

University of Chinese Academy of Sciences
School of Economics and Management
2009-2011, Graduate Certificate in Management of Engineering &Technology, Beijing, China

Hebei University of Science and Technology
School of Electrical Engineering
2005-2009 B.Eng. in Biomedical Engineering, China

Honors and Awards:

BOF Funding
2015-2016, Full Scholarship, Ghent University, Belgium
2014-2015, Top-up Funding, Ghent University, Belgium

Graduate Research Assistant Fellowship
2013-2015, 2017, The University of Western Ontario, Canada
2009-2012, Beijing Institute of Genomics, China

Ruth Horner Arnold Fellowship
2014, The University of Western Ontario, Canada

Mitacs Global Link Research Award
2014, Mitacs, Canada

Excellent Master Student Scholarship
2010, University of Chinese Academy of Sciences, China

Related Work Experience:

Doctoral Researcher
2014-2016, Center for Plant Systems Biology VIB, Gent, Belgium

Teaching Assistant
2013-2017, The University of Western Ontario, London Canada
2009-2012, Beijing Institute of Genomics, Beijing China

Student Internship
2013, Center Regulation Genomics (CRG), Barcelona Spain
2012-2013, Beijing Institute of Genomics, Beijing China
2012, Novogene Bioinformatics Institute, Beijing China

Publications:

A hitchhiker's guide of NGS: from sequencing, genome assembly to annotation (a review in preparation)

A Massive Expansion of Novel F-box Genes in polyphagous pest *Tetranychus urticae* (in preparation)

Update *Tetranychus urticae* genome using optical mapping (Ready for submission)

Comparative genomics of symbionts *Cardinium* reveals the sexuality of the major agricultural pest *Brevipalpus* host flat mites (in preparation)

Comparative genomics of three spider mite genomes reveals the feeding modes of three major agricultural pests (ready for submission)

Copidosoma genome and methylome reveal the polyembryony mechanism of wasp (ready for submission)

Zaichao Zhang, Zhong Jin, Yongbing Zhao, Zhewen Zhang, Rujiao Li, Jingfa Xiao, Jiayan Wu. A systematic study on GPCR prototypes: did they really evolve from prokaryotic genes? IET Systems Biology, doi:10.1049/iet - syb.2013.0037

Zaichao Zhang, Jiayan Wu, Jun Yu, Jingfa Xiao. A Brief Review on Evolution of GPCRs: Conservation and Diversification, Open Journal of Genetics, doi:10.4236/ojgen. 2 (2012) 11 – 17

Chen Cheng, Zaichao Zhang, Aizhong Ding, Jiayan Wu, Jingfa Xiao, Yujiao Sun. Bar-Coded Pyrosequencing Reveals the Bacterial Community during Microcystis water Bloom in Guanting

Reservoir, Beijing. Procedia Engineering (ISSN: 1877 - 7058) doi:10.1016/j.proeng.2011.11.054

Zaichao Zhang, Jiayan Wu, Jingfa Xiao. Evolutionary Analysis of the Multicopy Genes in Human Chromosome X. Collection of Abstracts of "The First Harbin International Symposium on Salmonella and Other Enteric Bacteria: Genomics and Biology". May 2011

Presentations:

NGS: Applications on Arthropod Genomes, Western University, Canada, 2017

Comparative Genomics of Three *Tetranychus* Mites with Different Feeding Habits, San Diego USA, 2017

Comparative Genomics of *Brevipalpus* Genomes, VIB Gent Belgium, 2016

Finalizing the Three Spider Mite Genomes, VIB Gent Belgium, 2016

Updating *T. urticae* genome using Optical mapping, VIB Gent Belgium, 2015

The *Tetranychus* gene family expansion and novel F-box gene identification, VIB Gent Belgium, 2015

Three genome comparisons, the ultimate combat: manual genomics, Ibiza, Spain, 2014

Comparative analysis of three spider mite genomes project, London Canada, 2013

Posters:

Comparative genomics of symbiont *Cardinium* species reveals the asexuality of *Brevipalpus* flat mites, Belgium (March 2017)

Comparative Genomics of *Tetranychus* Genomes, Western University, Canada (Jan 2017)

The three spider mite genomes reveal the feeding mechanism of major agricultural pests, VIB, Belgium (Jun 2016)

Genome annotation and comparative analysis: finishing three *Tetranychus* genomes, VIB, Belgium (May 2015)