KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Science

# Phase transition in the desert locust *Schistocerca gregaria*: a transcriptomic approach

**Rik Verdonck**

Supervisors:
Prof. dr. J. Vanden Broeck
Prof. dr. S.R. Ott

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor in Science: biology

July 2017

# Phase transition in the desert locust *Schistocerca gregaria*: a transcriptomic approach

**Rik VERDONCK**

Examination committee:
Prof. dr. R. Huybrechts, chair
Prof. dr. J. Vanden Broeck, supervisor
Prof. dr. S.R. Ott, co-supervisor
(University of Leicester, UK)
Prof. dr. L. Moons
dr. H. Verlinden
dr. D. Cullen

dr. S.M. Rogers
(University of Cambridge, UK)

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor in Science: biology

July 2017

© 2017 KU Leuven – Faculty of Science
Uitgegeven in eigen beheer, Rik Verdonck, Naamsestraat 59, B-3000 Leuven (België)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

# Dankwoord

Om dit doctoraat tot een goed einde te brengen, heb ik naast een aantal wetenschappelijke uitdagingen, vooral mezelf moeten overwinnen. Ik heb veel geleerd, en ben dankbaar dat ik dit avontuur heb mogen beleven met fantastische mensen aan mijn zijde. Aan het einde van de rit sta ik er dan ook op om deze mensen te bedanken.

Jef, dankjewel om mij de kans te geven deel uit te maken van een geweldige onderzoeksgroep! Toen ik aan dit project begon, had ik hoegenaamd niet de bagage om een RNA-Seq experiment tot een goed einde te brengen. Volgens mijn oorspronkelijke projectaanvraag ging ik dan ook micro-array doen, maar de technologie evolueerde snel. Je hebt me gesteund in mijn leerproces en hebt me het vertrouwen gegeven om zelf aan de slag te gaan met mijn eigen transcriptomics data. Dankjewel voor je geduld, voor de interessante discussies, en voor de ontspannende momenten op congres in Florida.

Swidi, thank you for sparking my interest in locusts! Our journey started almost 9 years ago, and you have been an incredible mentor ever since. You taught me how to dissect and video-track but, more importantly, you taught me the importance of first carefully designing my experiments before rushing into the lab. You have always generously shared your ideas and knowledge, and I have learned a lot from our numerous discussions over Skype. Many thanks for hosting me in your lab, and at your dinner table.

I would also like to express my sincere gratitude to the other members of my jury: Steve, Darron, Heleen, Lieve and Roger. I realise that reading and assessing a dissertation is hard work, and I am grateful for all critical notes and suggestions. They helped me finish this manuscript, and to order my thoughts about some lingering questions. I really enjoyed our discussions before, during and after my preliminary defence, and I'm looking forward to even more discussions in our future encounters.

Dit hele project had nooit tot stand kunnen komen zonder de hulp van een batterij enthousiaste en behulpzame collega's. Jullie zorgden mee voor de sprinkhanen en dissecteerden mee voor het transcriptoom. Wanneer ik bovendien weer eens verloren

aan het lopen was in het labo, waren jullie er altijd om mij te vertellen waar ik tipjesdozen of erlenmeyers kon vinden. Ik heb van jullie geleerd, en heb altijd genoten van de geweldige sfeer in het zoölogisch instituut! Een aantal collega's wil ik in het bijzonder bedanken. Dankjewel professor De Loof: jij knoopte me in mijn tweede masterjaar in de oren dat ik maar eens FWO moest aanvragen. Liliane, dankjewel om me klaar te stomen voor het echte werk, en voor de vruchtbare samenwerking rond peptidomics. Rut, jij bood het perfecte tegengewicht tegen mijn chaos. Samen hebben we echt bergen verzet op die derde verdieping, en een aanzienlijk deel van de data die ik in dit project gebruikt heb, heb ik aan jou te danken. Liesbeth: jij was mijn steun en toeverlaat wanneer het op bioinformatica aankwam. Ik heb je sinds je vertrek nog heel vaak gemist. Michiel: jij was mijn *partner in crime* bij het kietelen van sprinkhanen. Dankjewel om vrijwillig je tijd te komen slijten in een snikhete kamer ter grootte van een toilethokje. Wouter: ook al hebben de peptidomics uiteindelijk mijn doctoraat niet gehaald: we hebben de klus toch maar mooi geklaard. Dankjewel om mij keer op keer terug geduldig uit te leggen hoe mixed models nu weer juist in elkaar zitten. Roger en Evelien: heel veel dank voor de goede zorgen voor de beestjes.

I also owe a huge thanks to a number of people who have been of immense help in various aspects of this work. Thanks to Katja S. for getting me acquainted with the behavioural assay in Cambridge. Thanks Alejandro, for your continuous support during my journey in the world of bioinformatics. Special thanks to Katja N., Jens and Rohit for selflessly sharing your expertise and supercomputer. Thanks to all the people who made it possible to organise the yearly bioinformatics course in Leipzig, and the fantastic conference in Porto. I learned so much from you guys, I could barely imagine where I would have been without you!

Ik heb hard gewerkt, maar heb samen met mijn collega's ook heel veel plezier gemaakt. Ik dank hierbij met name de leden van de gouden generatie die samen met mij aan dit avontuur begonnen zijn. *You know who you are.* Vooral in de eerste jaren zijn we wel eens met kleine oogjes aan onze dag moeten beginnen omdat we tot diep in de nacht over ons onderzoek hadden vergaderd. Heel veel dank aan de organisatoren van ons jaarlijks laboweekend, het kerstetentje, en de vele andere sociale activiteiten. Many thanks to my office mates Heleen, Elisabeth, Rut, Isabel, Kamila, Darron, Jan, Yoshiki and Koen, for your support, humour, entertainment and generosity. Thanks to you, our office in the museum has always felt like home to me. Dankjewel ook aan de leden van zwemclub de walvis. Dankzij jullie zitten er bij de tien kilogram die ik ben bijgekomen tenminste nog een beetje spieren.

Ik wil ook graag mijn vrienden, familie en schoonfamilie in de bloemetjes zetten. Gedurende al deze jaren heb ik beroep kunnen doen op jullie, ook al was ik zelf niet altijd even beschikbaar. Jullie hebben kleur gegeven aan de waardevolle momenten van ontspanning, en hebben samen met mij gereisd, gekampeerd, gesport, geschaakt, gewandeld, gekaart, gemusiceerd, gegeten, gepintelierd en gepalaverd. We gingen samen naar festivals en op bedevaart, plukten oesters en boleten, zongen uit volle

borst, gingen lunchen om ons hart te luchten, en gingen bij elkaar dineren om de wereld te verbeteren. Ik mocht bij jullie binnenvallen; thuis en op jullie bureau. Jullie gaven me zelfs een heus steun-pakket om de laatste loodjes wat te verlichten. We vierden nieuwjaar dicht bij huis, en aan de andere kant van de wereld. We vierden jullie doctoraten en huwelijken, we wuifden jullie vrijgezellenbestaan uit, en we verwelkomden jullie kinderen. We hebben samen gelachen en we hebben samen gehuild. Want ook toen waren jullie er voor mij.

Mama en papa, ik ben jullie eindeloos dankbaar. Ik heb altijd kunnen rekenen op jullie financiële en morele steun. Maar vooral: jullie hebben het zaadje geplant van mijn interesse in biologie. Ik ben doorgaans een koele minnaar van indoctrinatie, maar voor jullie knijp ik graag een oogje dicht. Het was toen jullie mij leerden over boerenwormkruid en *Dytiscus marginalis*, toen jullie mij verhaaltjes vertelden over de tijgerhaai, toen jullie mij leerden lezen uit Grzimek, toen we fossielen gingen zoeken aan de voet van Cap Blanc-nez, toen ik mocht spelen met apenschedels, en toen we samen door de bino naar een ontwikkelend steur-embryo keken, dat de fundamenten gelegd werden van mijn grote passie voor de natuur. Die passie maakt me tot wie ik ben.

Leen, de laatste maanden van mijn doctoraat zijn zwaar geweest, niet het minst omdat ze drie jaar geduurd hebben. Duizend keer heb ik gezegd dat het "bijna klaar" was. Ik heb je geduld eindeloos op de proef gesteld, en toch stond je keer op keer voor me klaar. Ik kan je hiervoor niet genoeg bedanken. Het is als wetenschapper niet mijn gewoonte om stellingen te poneren op basis van ongerepliceerde metingen, maar ook voor jou knijp ik graag een oogje dicht. Je bent het beste lief ter wereld!

Graag wil ik tenslotte het Fonds Wetenschappelijk Onderzoek (FWO) bedanken voor de financiële steun en de regelmatige aanmoedigingen om vaker als eerste of laatste auteur te publiceren.

Saint-Girons, juni 2017

# Abstract

Locusts are grasshoppers that exhibit two density-dependent phenotypes called phases. When local population density rises, solitarious animals transition to the gregarious phase through a series of intermediates. Some aspects of this phenotypic change happen slowly, such as colour changes over successive moults. However, the locust's behaviour changes in a matter of hours. In the desert locust, changes in activity and attraction towards conspecifics take place after only a few hours of forced crowding. Two main sensory input pathways are involved in this process: a thoracic pathway with tactile input from the hindlegs, and a cephalic pathway with visual and olfactory input from the eyes and antennae.

In this thesis, I present a transcriptomics (RNA-Seq) experiment focusing on the $5^{th}$ instar desert locust brain during early behavioural phase transition. For two hours, solitarious locusts were subjected to crowding, to isolated tactile or visual/olfactory stimuli that are known to induce behavioural gregarization, and to a control treatment. Transcriptional responses were weak and non-specific: out of 13 transcripts that responded to crowding or crowding-mimicking treatments, 11 also responded to the control treatment that did not induce behavioural gregarization. Remarkably, 12 transcripts were up-regulated as a response to a 10-minute observation in an arena containing other locusts, including *takeout*: a transcript involved in early behavioural change in the migratory locust (*Locusta migratoria*).

As a first prerequisite to this experiment, a reference transriptome was constructed based on mRNA extracted from different tissues from animals in different life stages under a variety of conditions. Illumina-sequenced reads were assembled into 170,066 contigs. This assembly was filtered to a final reference of 21,595 contigs that should contain a single representative of each gene that was expressed in the brain. In this brain-specific reference transcriptome, 10,276 contigs were annotated with at least one gene ontology term, among which 7,164 contigs with at least one hit in all three gene ontology domains.

In order to study phase transition on the level of the individual's behaviour, a video-based arena-assay was used. I developed a new tracking software, written in the statistical programming language R, and successfully tracked locusts in more than a thousand videos. Based on the tracks, 50 behavioural variables were recorded. Multivariate techniques (hierarchical clustering, principal component analysis and exploratory factor analysis) were used to explore the main axes of variation and correlations between all recorded behaviours. Two logistic regression models and one random forest were fitted on a training set of observations of long-term solitarious and gregarious animals. These models were later used for the assessment of the individual phase state of animals during crowding, or under the crowding-related experimental conditions of the transcriptomics experiment.

# Beknopte samenvatting

De woestijnsprinkhaan *Schistocerca gregaria* kent, afhankelijk van de populatiedensiteit, twee verschillende verschijningsvormen (fasen). Wanneer lokaal de populatiedensiteit stijgt, ondergaan solitaire dieren een transitie naar de gregaire fase. Terwijl de meeste aspecten van deze "fasetransitie"relatief traag verlopen, zijn gedragsveranderingen al zeer snel merkbaar. Wanneer solitaire sprinkhanen samen gedwongen worden op een kleine oppervlakte, zijn ze in eerste instantie passief en vermijden ze elk contact. Na enkele uren worden ze actiever en gaan ze meer met elkaar interageren. De belangrijkste prikkels die dit proces aansturen zijn enerzijds aanrakingen ter hoogte van de achterpoten, en anderzijds het zien en ruiken van andere sprinkhanen.

In deze thesis beschrijf ik een transcriptomics-experiment (RNA-Seq) dat focust op de hersenen van de woestijnsprinkhaan gedurende de eerste uren van de fasetransitie. Solitaire sprinkhanen werden gedurende twee uur gestimuleerd door hen gedwongen samen te brengen, of door het aanbieden van geïsoleerde tactiele of gecombineerde visuele en olfactorische stimuli. Over het algemeen reageerde het transcriptoom zwak en aspecifiek op deze stimuli. Slechts 13 transcripts kwamen differentieel tot expressie, waarvan er 11 ook reageerden op een controlebehandeling die geen gregair gedrag veroorzaakte. Opmerkelijk was dat 12 transcripts opgereguleerd waren na slecths 10 minuten observatie in een arena waarin zich nog sprinkhanen bevonden. Eén van deze genen was *takeout*: een transcript dat ook betrokken is bij de vroege fasetransitie van de Afrikaanse treksprinkhaan (*Locusta migratoria*).

Een belangrijke benodigdheid bij dit experiment was een referentietranscriptoom. Ik maakte een assembly van 170066 contigs vertrekkende van gesequeneerd (Illumina sequencing) cDNA op basis van geëxtraheerd mRNA uit verschillende weefsels van dieren van verschillende leeftijden, opgekweekt onder verschillende condities. Deze assembly werd verder gefilterd tot een referentietranscriptoom van 21595 contigs, met een representatief transcript voor de meerderheid van de genen die in de hersenen tot expressie komen. Voor 10276 van deze contigs werd een annotatie gevonden met minsens 1 "gene ontology"(GO) term. 7164 contigs hadden zelfs een annotatie in elk

van de drie GO domeinen.

Om het gedrag van de individuele sprinkhanen te bestuderen tijdens hun fasetransitie, werd een toets gebruikt waarbij de sprinkhaan in een arena gefilmd werd. Ik ontwikkelde software om uit deze videos het traject van de sprinkhaan te bepalen en interpreteerde zo meer dan duizend filmpjes. Op basis van de trajecten werden 50 gedragsvariabelen gemeten die vervolgens geanalyseerd werden met multivariate statistiek (hiërarchische clustering, hoofdcomponentanalyse en factoranalyse) om zo de belangrijkste correlatiestructuren bloot te leggen. Op basis van observaties van solitaire en gregaire sprinkhanen werden twee logistische regressiemodellen en een "random forest"getraind. Deze modellen werden later gebruikt om de progressie van fasetransitie te bepalen onder bovengenoemde experimentele omstandigheden.

# Abbreviations

| | |
|---|---|
| **5-HT** | 5-hydroxytryptamine, serotonin |
| **AIC** | Akaike information criterion |
| **ALRC** | Anti locust research centre |
| **AUC** | Area Under the Receiver-Operator-Curve |
| **BAM** | Binary Sequence Alignment/Map file format |
| **BIC** | Bayesian information criterion |
| **BH** | Benjamini-Hochberg |
| **BLAST** | Basic local alignment search tool |
| **cDNA** | complementary DNA |
| **CDS** | Coding sequence |
| **CNS** | Central nervous system |
| **COPR** | Centre for Overseas Pest Research |
| **CP** | Cuticle protein |
| **CREB** | cAMP response element-binding protein |
| **CSP** | chemosensory protein |
| **DE** | Differentially expressed |
| **DEG** | Differentially expressed gene |
| **DNA** | Deoxyribonucleic acid |
| **dNTP** | deoxy Nucleoside triphosphate |
| **dsDNA** | double stranded DNA |
| **DSN** | Duplex specific nuclease |
| **EFA** | Exploratory factor analysis |
| **eggNOG** | evolutionary genealogy of genes: Non-supervised Orthologous Groups |
| **EST** | Expressed sequence tag |
| **etOH** | Ethanol |
| **FAO** | Food and Agriculture Organisation of the united nations |
| **FC** | Fold change |
| **FD** | Fractal dimension |
| **FPR** | False positive rate |
| **fps** | Frames per second |

| | |
|---|---|
| **gb** | gigabase |
| **GB** | gigabyte |
| **GOF** | Goodness-of-fit |
| **GO** | gene ontology |
| **HMM** | Hidden Markov Model |
| **ICE** | Max Planck Institute for Chemical Ecology (Jena, Germany) |
| **JH** | Juvenile Hormone |
| **KEGG** | Kyoto Encyclopedia of genes and genomes |
| **KU Leuven** | Katholieke universiteit Leuven |
| **KUL** | Katholieke universiteit Leuven |
| **L1, L2, ...** | First instar stage, second instar stage, ... |
| **LR** | Logistic regression |
| **mb** | megabase |
| **MB** | megabyte |
| **MS** | Mass spectrometry |
| **MSA** | Multiple sequence alignment |
| **mtDNA** | Mitochondrial DNA |
| **MWU** | Mann-Whitney-U test |
| **N** | Number |
| **NCBI** | National Centre of Biotechnology |
| **NDE** | Not differentially expressed |
| **NGS** | Next-generation sequencing |
| **NP** | Neuropeptide |
| **OOB** | out-of-bag estimate |
| **nt** | nucleotides (length of contigs, reads, or DNA or RNA fragments) |
| **ORF** | Open reading frame |
| **PCA** | Principal component analysis |
| **PC** | Principal component |
| **PCR** | Polymerase chain reaction |
| **PE** | Pair(ed) end |
| **Pfam** | Protein family (database) |
| **Pgreg** | The probability of being gregarious |
| **PKA** | Protein kinase A |
| **QC** | Quality control |
| **RIN** | RNA integrity number |
| **RF** | Random forest |
| **RNA** | Ribonucleic acid |
| **RNAi** | RNA interference |
| **RNA-Seq** | RNA sequencing, usually in a quantitative context |
| **ROC** | Receiver-operator-curve |
| **RT** | Room temperature, 20 °C |
| **rRNA** | ribosomal RNA |

| | |
|---|---|
| **SAM** | Sequence Alignment/Map file format |
| **SE** | Single end |
| **SOG** | Sub-oesophagal ganglion |
| **STAR** | Spliced Transcript Alignment to a Reference |
| **TB** | Terabyte |
| **TF** | Time fraction |
| **TG** | Thoracic ganglion |
| **TMM** | Trimmed mean of M-values |
| **TO** | Takeout |
| **TPR** | True positive rate |
| **UniProt** | Universal Protein Resource |
| **UTR** | Untranslated region |

# Contents

**Abstract** v

**Abbreviations** ix

**Contents** xiii

**1 Introduction** 1

1.1 Locust phase polyphenism . . . . . 2

1.2 Phenotypic plasticity and polyphenism . . . . . 2

1.3 The desert locust *Schistocerca gregaria* . . . . . 3

1.3.1 *Schistocerca gregaria* (Forskål) . . . . . 3

1.3.2 Differences between extreme phases . . . . . 5

1.3.3 Phenotypic plasticity during phase transition . . . . . 6

1.3.4 Monitoring and control . . . . . 6

1.4 The ethology of phase polyphenism: the Tinbergen questions . . . . 8

1.4.1 The mechanisms of phase transition . . . . . 9

1.4.2 Why locust swarm, sometimes . . . . . 10

1.4.3 The development of phase polyphenism . . . . . 12

1.4.4 The evolution of locust phase polyphenism . . . . . 13

1.5 Context and objectives of this work . . . . . 16

1.5.1 Context . . . 16
1.5.2 Objectives . . . 17

**2 General Methods 19**

2.1 Introduction . . . 20
2.2 Locust breeding . . . 20
2.3 Tissue collection . . . 22
2.4 General lab procedures . . . 22
2.4.1 RNA extraction . . . 22
2.4.2 mRNA purification . . . 23
2.4.3 cDNA synthesis and sequencing library preparation . . . 24
2.4.3.1 Clontech SMARTer™ . . . 24
2.4.3.2 TruSeq® . . . 25
2.4.4 Determining the quality and concentration of nucleic acid samples . . . 25
2.5 Next generation sequencing . . . 26
2.5.1 Sequencing by synthesis: Illumina® . . . 26
2.5.2 The FASTQ file format and base call quality . . . 29
2.6 Bioinformatics . . . 32
2.6.1 Hardware and operating systems . . . 32
2.6.2 Scripts . . . 32
2.6.3 Software . . . 33
2.6.3.1 FastQC . . . 33
2.6.3.2 BLAST . . . 33
2.6.3.3 Multiple sequence alignment . . . 34
2.6.3.4 Mappers . . . 35
2.6.3.5 SAM, BAM and SAMtools . . . 35
2.6.3.6 BEDtools . . . 36

2.6.3.7 The Trinity transcriptome assembler . . . . . 36
2.6.3.8 Trimmomatic . . . . . 38
2.6.3.9 FLASH . . . . . 39
2.6.3.10 insilico_read_normalization.pl . . . . . 39
2.6.3.11 RepeatMasker . . . . . 40
2.6.3.12 HMMER, TMHMM RNAmmer and signalP . . . . . 40
2.6.3.13 TransDecoder . . . . . 41
2.6.3.14 Trinotate . . . . . 41
2.6.3.15 CD-HIT and CD-HIT-EST . . . . . 41
2.6.4 Multivariate analysis . . . . . 42
2.6.4.1 Hierarchical clustering . . . . . 42
2.6.4.2 Principal component analysis . . . . . 42
2.6.4.3 Exploratory factor analysis . . . . . 43
2.6.5 Predictive modelling . . . . . 44
2.6.5.1 Binary logistic regression . . . . . 44
2.6.5.2 Random forests . . . . . 45

**3 A new *Schistocerca gregaria* reference transcriptome 47**

3.1 Introduction . . . . . 48
3.1.1 Why sequence a reference transcriptome? . . . . . 48
3.1.2 Why not sequence a reference genome? . . . . . 48
3.1.3 Current orthopteran reference genomes and transcriptomes . 48
3.1.4 De novo transcriptome assembly . . . . . 49
3.1.5 Transcript and gene annotation . . . . . 49
3.2 Aims . . . . . 50
3.3 Wetlab procedures . . . . . 50
3.3.1 Overview of the samples . . . . . 50

3.3.2 Sample preparation for set 5 . . . 51
3.3.3 Sequencing in three attempts . . . 53
3.3.4 Third time lucky: set 5a2 . . . 54
3.3.4.1 The Clontech libraries . . . 54
3.3.4.2 The TruSeq® libraries . . . 55
3.3.4.3 Further library preparation with TruSeq® . . . 56
3.3.4.4 *In vitro* Library normalization . . . 57
3.3.4.5 Loading samples onto the sequencer . . . 59
3.4 Bioinformatics . . . 59
3.4.1 The assembly pipeline . . . 60
3.4.1.1 Acquiring and pre-processing the read data . . . 60
3.4.1.2 Preliminary contaminant analysis . . . 61
3.4.1.3 Estimating the distribution of insert sizes . . . 62
3.4.1.4 Trimming-merging-trimming . . . 63
3.4.1.5 Redundancy and *In silico* read normalization . . . 64
3.4.1.6 Assembly with Trinity . . . 68
3.4.1.7 Summary of the entire assembly pipeline . . . 69
3.4.2 Filtering and curation of the assemblies . . . 71
3.4.2.1 Why filtering? . . . 71
3.4.2.2 Generating a reference set for RNA-Seq . . . 71
3.4.2.3 Manual curation . . . 74
3.4.3 Annotation of the reference set . . . 75
3.4.4 Neuropeptide precursor recovery and annotation . . . 76
3.5 Results . . . 76
3.5.1 Input data for assembly 5 . . . 77
3.5.2 Assembly 5: quality and contents . . . 77
3.5.3 The reference for the RNA-Seq experiment . . . 78

3.5.4 Neuropeptide precursors . . . 79
3.6 Discussion . . . 84
3.6.1 Technical considerations . . . 84
3.6.1.1 Library preparation . . . 84
3.6.1.2 Roche FLX+ 454 reads . . . 85
3.6.1.3 Contig numbers . . . 85
3.6.1.4 Stratification of ordered contig output . . . 87
3.6.2 The reference transcriptome . . . 87
3.7 Acknowledgements . . . 88

**4 An automated behavioural assay and quantitative analysis of phase-related behaviour 91**

4.1 Introduction . . . 92
4.1.1 Measuring behaviour . . . 92
4.1.2 Video tracking . . . 93
4.1.3 A history of quantitative studies in locust behaviour . . . 94
4.2 Aims . . . 97
4.3 Methods . . . 98
4.3.1 The observed locusts . . . 98
4.3.2 Automated behavioural assay . . . 98
4.3.3 The arena . . . 99
4.3.4 Video recording . . . 99
4.3.5 Metadata recording . . . 101
4.3.6 Video tracking . . . 102
4.3.6.1 Swistrack . . . 102
4.3.6.2 Generating auxiliary files. . . . 102
4.3.6.3 The tracker . . . 103
4.3.7 Segmenting tracks into behavioural variables . . . 106

4.3.7.1 tandy.R . . . 106
4.3.7.2 Input parameters for tandy.R . . . 110
4.3.8 Exploration of behavioural variables and their contribution to phase . . . 110
4.3.8.1 The data . . . 110
4.3.8.2 Selection and transformation . . . 111
4.3.8.3 Hierarchical clustering . . . 116
4.3.8.4 Principal component analysis . . . 116
4.3.8.5 Exploratory factor analysis . . . 117
4.3.9 Classification, model building and model validation . . . 117
4.3.9.1 Logistic regression . . . 117
4.3.9.2 Random forests . . . 118
4.3.9.3 Model validation . . . 119
4.3.10 Comparison between time points . . . 119
4.4 Results . . . 121
4.4.1 Evaluation of the new pipeline . . . 121
4.4.1.1 Performance of the tracker . . . 121
4.4.1.2 Performance of tandy.R . . . 121
4.4.2 A first comparison between long term phases . . . 123
4.4.3 Comparison between manual and automated assay . . . 129
4.4.4 Hierarchical clustering . . . 131
4.4.5 Correlations between observed variables . . . 134
4.4.5.1 The "automated training" data . . . 134
4.4.5.2 The "automated total" data . . . 134
4.4.6 Principal component analysis . . . 136
4.4.6.1 PCA on the training data . . . 136
4.4.6.2 Projection of the "automated total" data . . . 141

4.4.7 Exploratory factor analysis . . . 144
4.4.8 Models for behavioural phase state . . . 148
4.4.8.1 Logistic regression models . . . 148
4.4.8.2 Random forest models . . . 148
4.4.8.3 A final selection of 3 models . . . 149
4.4.8.4 Validation on a small test set . . . 151
4.4.9 The effects of crowding . . . 151
4.4.10 Some interesting observations . . . 153
4.4.10.1 Metadata without evidence for effect . . . 153
4.4.10.2 Metadata with effect . . . 155
4.5 Discussion . . . 160
4.5.1 Evaluation of the new tracker . . . 160
4.5.2 Multivariate analysis of behaviour . . . 161
4.5.3 Modelling extreme phases . . . 163
4.5.4 Comparison between old and new set-up . . . 165
4.5.5 Gregarization . . . 166
4.6 Acknowledgements . . . 167

**5 Transcriptomic signatures of different sensory input pathways during early behavioural gregarization 169**

5.1 Introduction . . . 170
5.1.1 State of the art . . . 170
5.1.1.1 Transcriptomic studies in locusts . . . 170
5.1.1.2 The sensory pathways during early behavioural change . . . 172
5.1.2 Design of the experiment . . . 174
5.1.3 Finding differentially expressed genes . . . 175
5.1.3.1 Mapping and counting . . . 175

5.1.3.2 Statistical inferences . . . 176
5.2 Aims . . . 178
5.3 Methods . . . 179
5.3.1 Animals . . . 179
5.3.2 Treatments . . . 179
5.3.3 Behavioural observation and analysis . . . 181
5.3.4 Dissections . . . 181
5.3.5 Pooling samples . . . 181
5.3.6 Sample preparation . . . 181
5.3.7 Quality, concentration and extra samples . . . 182
5.3.8 Sequencing . . . 182
5.3.9 Quality and filtering of sequences . . . 185
5.3.10 Mapping and counting mappings . . . 186
5.3.11 Exploring, normalizing and processing count data . . . 186
5.3.12 Differentially expressed genes . . . 187
5.4 Results . . . 188
5.4.1 Behaviour . . . 188
5.4.2 Transcriptomics . . . 194
5.4.2.1 RNA quality . . . 194
5.4.2.2 Sequence quality and contamination . . . 194
5.4.2.3 Mapping and read counting . . . 195
5.4.2.4 Differentially expressed genes . . . 195
5.5 Discussion . . . 200
5.5.1 Behaviour . . . 200
5.5.2 Transcriptomics . . . 201
5.5.2.1 Technical considerations . . . 201
5.5.2.2 Biological signal . . . 204

5.5.2.3 Noteworthy findings . . . . . 205
5.6 Acknowledgements . . . . . 207

**6 Conclusions and perspectives 209**

6.1 The new desert locust reference transcriptome is a valuable resource 210
6.2 Contributions to a quickly developing discipline . . . . . 211
6.3 The transcriptomic response to two hours of crowding is weak and non-specific . . . . . 211
6.4 Concerns about earlier studies . . . . . 212
6.5 An R-based behavioural assay . . . . . 213
6.6 $P_{greg}$ does not say it all . . . . . 213

**References 215**

# Preface

> So Moses stretched out his staff over Egypt, and the Lord made an east wind blow across the land all that day and all that night. By morning the wind had brought the locusts; they invaded all Egypt and settled down in every area of the country in great numbers. Never before had there been such a plague of locusts, nor will there ever be again. They covered all the ground until it was black. They devoured all that was left after the hail — everything growing in the fields and the fruit on the trees. Nothing green remained on tree or plant in all the land of Egypt.
>
> Book of Exodus 10: 13-15, NIV

Since ancient times, people in North Africa and the Middle East have feared locust outbreaks, and still also today, these voracious insects can cause crop losses that put millions of people at risk of famine. In 2004, swarms of desert locusts infested 20 countries from Morocco all the way east to Israel. Despite hundreds of millions of dollars invested in control operations, harvest losses amounted to an estimated 250 - 500 million USD [21].

Less than 10 years later I was on a trip in Morocco. In Ouarzazate, the *door of the desert*, I asked our guide if he could show me some of those famous locusts. He explained to me that they were hard to find and even harder to catch. We were in luck, and after a couple of hours walking in the awe-inspiring semi-desert, we found a single adult female. It took me a couple of attempts to catch her: she was pretty well camouflaged against the background of rock and sand, and time after time she remained motionless until the moment she was almost within my reach. Then she would jump up and fly about ten metres to land back on the arid desert soil: a typical behaviour for a solitarious desert locust.

It seems hard to imagine that this solitary, inconspicuous grasshopper I found in Morocco belongs to the same species as the billions of bright yellow swarming locusts that visited the area in 2004. Yet it does.

# Chapter 1

# Introduction

> Thus, the periodicity of locust invasions is caused entirely by the wonderful phenomenon of the transformation of a swarming locust into a solitary, harmless grasshopper. Of course, the outline here sketched is necessarily rough, and the actual proceedings are far more complicated, but the theory seems to me to be the best possible in the circumstances
>
> Boris P. Uvarov

## 1.1 Locust phase polyphenism

The fact that swarming locusts are actually conspecific to their shy and solitary counterparts, was not described in Western scientific literature until the beginning of last century, when the Russian-British entomologist Boris Uvarov (1886 - 1970) [92] postulated the theory of locust *phases* [264]. In the years before the First World War, Uvarov was sent out to install a crop protection programme in the Northern Caucasus, a region that was heavily afflicted by grasshopper pests at the time. After detailed field observations of the common grasshopper species, he came to the conclusion that the migratory locust *Locusta migratoria* and the solitary grasshopper *Locusta danica*, were actually one single species, and that the *solitaria* phase transitioned to the *gregaria* phase during outbreaks. Uvarov was not the first person to witness this phenomenon[1], but he was the first one to postulate the hypothesis that the migratory locust was a single species capable of producing two extremely different forms. He called these forms "phases": the *solitaria* phase was typically solitary, inconspicuous and cryptically coloured, while the more contrasting coloured *gregaria* phase was inclined to aggregate and was highly mobile. Uvarov also described the transition between both phases, stressing that it was neither seasonal nor predictable.

In later works [263], the phenomenon of *phase transition* was also reported for a number of other species, leading to a series of taxonomic revisions, including the merging of *Schistocerca gregaria* (Forskål) and *Schistocerca flaviventris* (Burmeister) into a single species: *Schistocerca gregaria* (Forskål) [267]. In the meantime we know that many grasshopper species show some degree of population-density-dependent plasticity [238], but only those species where the high-density *phase* engages in collective movements and swarming, qualify as *locusts* [195, 238].

## 1.2 Phenotypic plasticity and polyphenism

*Phenotypic plasticity* is the ability of a single genotype to produce multiple phenotypes in response to the environment [98, 139]. The availability of multiple phenotypes that can develop according to environmental cues, confers obvious benefits compared to a rigid genotype-phenotype relationship, notably in heterogeneous environments. As a result, phenotypic plasticity is widespread in the animal kingdom. Some animals develop (presumably costly) anti-predation mechanisms upon presentation of predator kairomones: different species of water flea (*Daphnia*) develop spines and helmets [164, 244, 29], and tadpoles of different frog species develop shorter tails or aposematic colouration [146, 225, 142]. In other cases, plasticity is induced by social cues, such

[1] In 1914, V. Plotnikov, reported that in captivity, *Locusta migratoria* could be bred from *Locusta danica* and *vice versa* [264]. Plotnikov's hypothesis was that he was witnessing speciation in real-time, and that *Locusta danica* was currently splitting off from the primitive species *Locusta migratoria*.

as the development of secondary sex characteristics depending on social status in orang-utans [230, 262], or even sex reversals in coral reef fish, depending on the relative abundance of both sexes [96].

Since phenotypic variation within a species can also be the result of genetic variation, Ernst Mayr proposed the distinction between *polymorphism* and *polyphenism* in his book *Animal Species and Evolution* [175]. He defined polyphenism as "the occurrence of several phenotypes in a population, the differences between which are not the result of genetic differences".

Locusts are a textbook example of such a polyphenism: depending primarily on local population density, a single genome can give rise to two contrasting phenotypes that are adapted either to living in a crowd, or to living a solitary life. Between these two polarities, a spectrum of intermediate phenotypes can be observed [140, 216, 256].

## 1.3 The desert locust *Schistocerca gregaria*

### 1.3.1 *Schistocerca gregaria* (Forskål)

The desert locust, *Schistocerca gregaria* (Forskål) is a short-horned grasshopper (suborder *Caelifera*) in the family of *Acrididae*, subfamily *Cyrtacanthacridinae*. In this work, "desert locust" and *Schistocerca gregaria* will refer to the nominate subspecies *Schistocerca gregaria gregaria* (Forskål), that lives in North Africa and the Middle East. The subspecies *Schistocerca gregaria flaviventris* (Burmeister) from South-West Africa is less inclined to swarm, and seems to be genetically isolated from *Schistocerca gregaria gregaria* (Forskål) [45].

The solitarious desert locust inhabits semi-arid and arid areas at low population densities. Generally it does not present a threat to agriculture. It has a range (the recession area) of about 16 million $km^2$, from Mauritania to Western India, mostly between 15°N and 30°N, but also including the horn of Africa and the entire Arabian peninsula. Gregarious locusts form swarms that may leave the recession area and extend this range over a land area of an additional 13 million $km^2$ (fig. 1.2) [170, 252].

In many respects, the desert locust is one of the most pronounced locusts: the two extreme phases look drastically different, yet during the transition from the *solitaria* phase to the *gregaria* phase, some aspects of their behaviour and physiology change in a matter of hours [266, 214, 219, 71, 5, 195, 187].

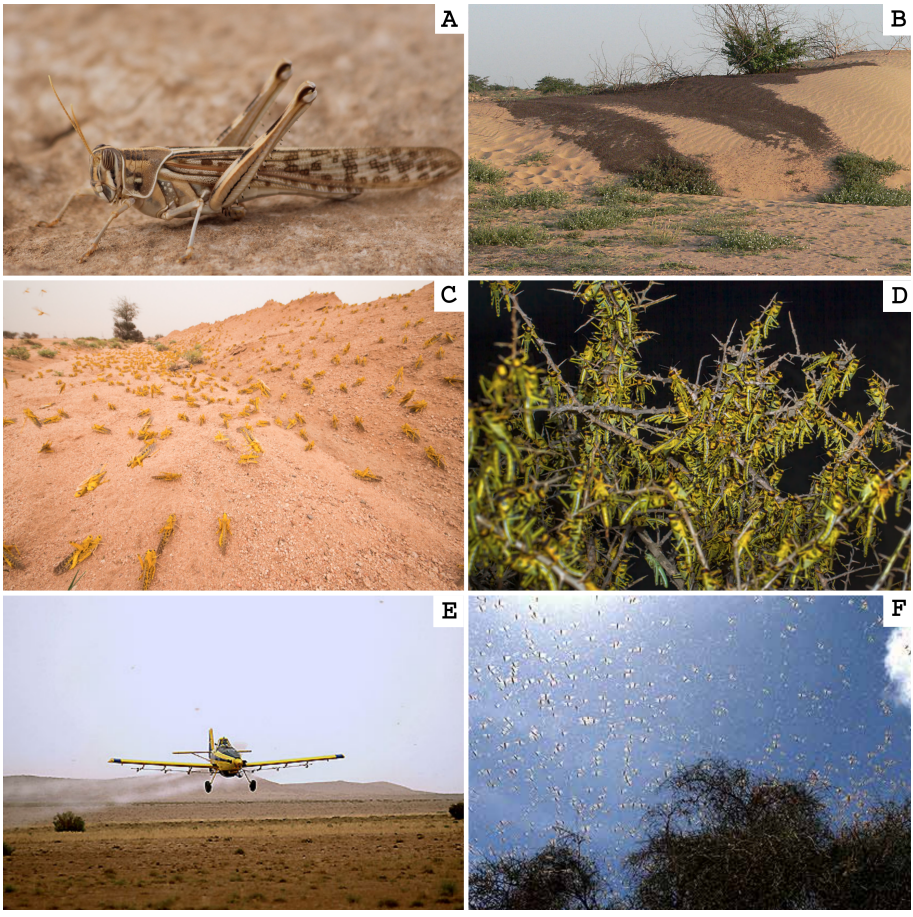



**Figure 1.1:** A: an adult female solitarious desert locust, near Oarzazate, Morocco (Rik Verdonck). B: dense *hopper bands* of $3^{rd}$ instar stage nymphs on the coast of the Red Sea near Aeterba, Sudan (wikimedia commons). C: yellow coloured gregarious adults assemble to mate (© Yazeed Alsahli). D: gregarious desert locust nymphs aggregate on a bush at night. Note the single solitarious phase nymph on the bottom center of the image. Negev desert, Israel (© Gil Wizen). E: aerial control operations in Morocco (FAO). F: airborne swarm near Kaedi, Mauritania (FAO).

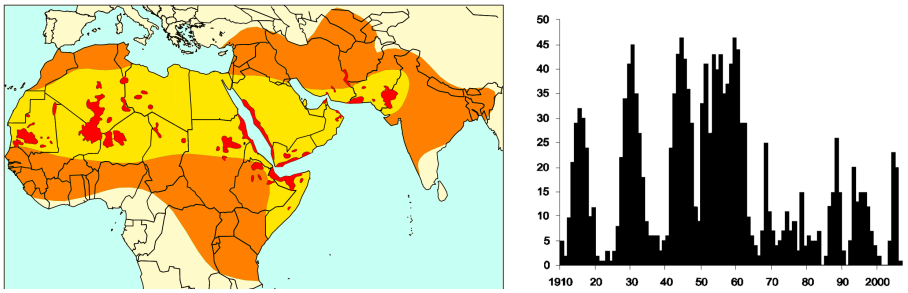



**Figure 1.2:** Left: distribution map of *Schistocerca gregaria gregaria* (Forskål). Low density solitarious populations are mainly restricted to the recession area (yellow). Outbreaks often nucleate at hot-spots (red) that are characterized by relatively high levels of rainfall, such as the semi-desert of southern Mauritania, or the coast of the Red Sea [252]. Once swarming, desert locust may invade regions where the species usually does not occur (invasion area, orange) (adapted from Lecoq, 2004). Right: number of countries annually affected by desert locust swarms, (adapted from Magor *et al.* 2008).

### 1.3.2 Differences between extreme phases

The extreme (long-term) phases (henceforth referred to as the gregarious and solitarious phase) are adapted to radically different lifestyles, which is reflected in many aspects of their biology [266, 195] (fig. 1.3).

The nymphs of the solitarious phase range from bright green to straw colour, with limited or no black patterning. Gregarious nymphs have a black-yellow pattern that shows little variation among individuals. After the final moult, the solitarious adult is beige-brown while gregarious adults are pink at first, turning yellow when sexual maturity is reached. Adult males of both phases are about equal in size. In both phases females are the larger sex, but sexual dimorphism is most pronounced in the solitarious phase, where the female is about 65% larger than the male (in surface area [266]). Solitarious animals have relatively longer and heavier legs that allow them to jump further despite their heavier weight [221]. The distribution of tactile hairs over the body differs between both phases, suggesting differences in sensitivity to tactile stimulation. In the gregarious phase, the brain is about 30% larger than in the solitarious phase. These differences are not isometric: the proportional size of many neuropiles differs between phases. Brain regions involved in primary sensory processing are relatively well developed in solitarious animals, while in gregarious locusts, the mid-brain and central complex are larger, suggesting the need for more higher integration, which may be more important in the context of high mobility and intra-specific competition [186].

In captivity, solitarious desert locusts live about a month longer than gregarious

animals (81 days into adulthood versus 54), which allows them to produce an extra clutch of eggs [28]. Their clutch size is larger, but the eggs are smaller than the ones produced by gregarious mothers. Despite their lower *per capita* fecundity, gregarious locusts have faster rates of population increase thanks to synchrony, faster maturation and higher survival chances of their larger hatchlings [47].

Apart from differences in appearance, brain anatomy and life history, phases also exhibit important differences in behaviour. When faced with other locusts, solitarious locusts avoid contact, while gregarious locusts interact with each other and engage in collective movements. The social aspect is not the only aspect of behaviour that depends on phase: food preference [71] and circadian rhythm [254] also differ considerably.

### 1.3.3 Phenotypic plasticity during phase transition

When phases transition in either direction, phase-related traits change each at their pace. Some traits, such as morphometrics (e.g. the ratio of femur length over head width), change gradually over multiple generations [140, 122], while others, such as colouration, change over successive moults.

Behaviour changes over a much shorter time. Depending on the strain and the assay used, solitarious desert locusts have been shown to shift towards more gregarious-like behaviour upon 2 to 4 hours of crowding. They show increased activity and lose their repulsion from conspecifics [214]. They also modulate their food preference towards plants that are toxic to their predators [71, 235]. Behavioural change upon isolation of the gregarious phase is a slower process [214, 243], but this has been less intensely studied. Behaviours acquired during early stages of phase transition are labile at first, and are easily lost upon re-isolation or re-crowding [214].

### 1.3.4 Monitoring and control

The Food and Agriculture Organization of the United Nations (FAO) exercises close monitoring of desert locust populations, with a monthly *desert locust bulletin* since 1978, and with regular updates on their website. They employ a specific vocabulary for the varying degrees of aggregation, with a colour coded map indicating the corresponding threats to agriculture. During *outbreaks*, locust numbers rise: hoppers may aggregate locally, and small, low-density swarmlets may occur. Occasionally, larger swarms may form as smaller aggregations coalesce because of convergent winds. Without any control operations, outbreaks may result in *upsurges* after some generations of successful breeding. The total population grows exponentially, leading to the formation of coherent hopper bands and swarms consisting of hundreds of

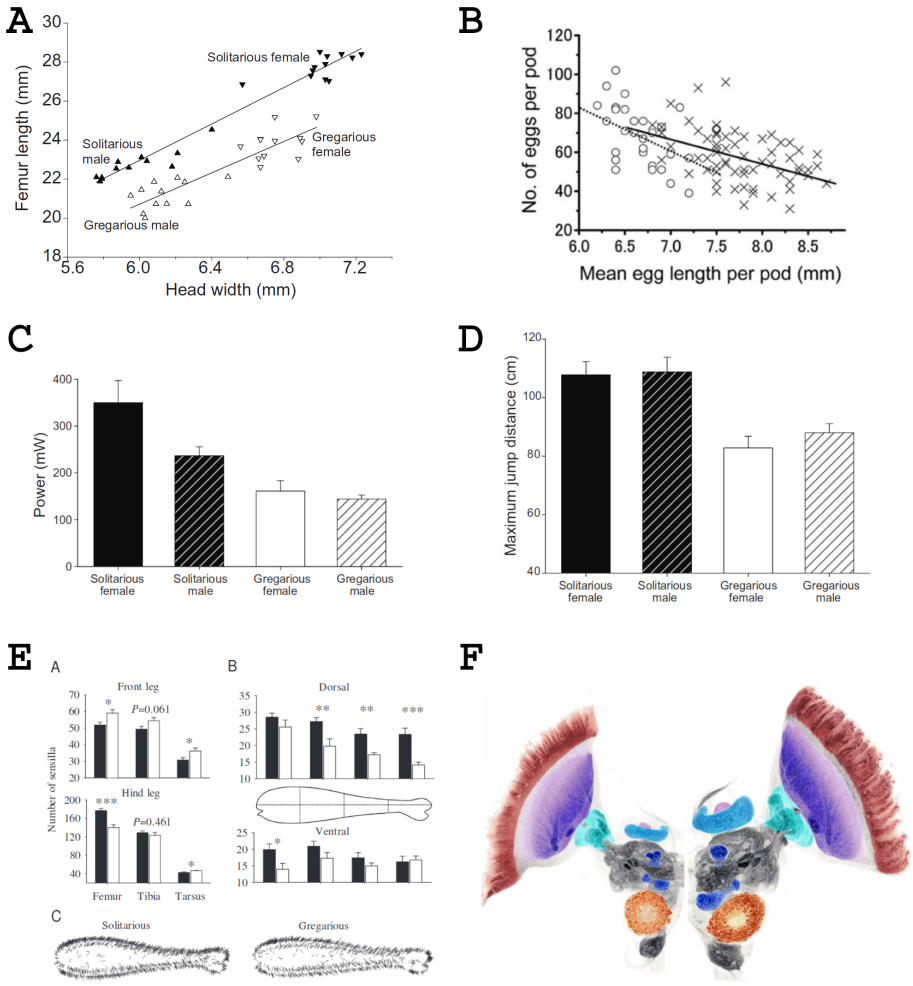



**Figure 1.3:** Some examples of differences between long term phases of the desert locust. A: among many morphometric differences, solitarious desert locusts have a higher femur length over head width ratio (from Rogers *et al.* 2016 [221]). B: solitarious females (circles and dashed regression line) lay more and smaller eggs than gregarious (crosses and solid regression line), reflecting differences in life-history choices (from Maeno *et al.* 2013 [168]). C and D: when stimulated to jump, solitarious desert locusts generate more power, helping them to jump further in spite of their higher body weight (from Rogers *et al.* 2016 [221]). E: solitarious locusts (black bars) have less tactile hairs on their front legs, but more on their hind femur, which is the most touch-sensitive body part during early phase transition (from Rogers *et al.* 2003 [218]). F: Comparison of a solitarious (left), and gregarious (right) brain. Note that the gregarious brain is not just larger, but there are also allometric differences in the relative size of different neuropiles (adapted from Ott & Rogers 2010 [186]).

millions of individuals, covering several square kilometres [61]. Widespread upsurges in consecutive years, or on multiple localities, are referred to as *plagues* [254].

In 1965, the FAO started organized chemical pest control programs, with the use of improved techniques such as aerial pesticide treatments. Thanks to close population monitoring, outbreaks can often be controlled before population growth leads to upsurges. The arrival of migrant swarms can also be predicted to a certain degree because the direction of movement is heavily influenced by the prevailing winds [140]. However, sometimes locusts seem to collectively wait for wind from a given direction before getting airborne [254].

Small-scale outbreaks still occur several times per year within the recession area, but upsurges and plagues have lowered in frequency and intensity compared to the first half of the last century (fig. 1.2) [170].

## 1.4 The ethology of phase polyphenism: the Tinbergen questions

The study of animal behaviour is called ethology. There are many sub-disciplines, each studying behaviour from a different point of view. It is not always easy to communicate between these sub-disciplines because of their specific vocabularies. A suitable framework to get oneself oriented was developed by Nikolaas Tinbergen [260]. Tinbergen argued that for each behaviour, there is a set of 4 complementary explanations: the mechanisms, the utility in the current environment, the development, and the evolution [260].

In what follows, I will formulate the Tinbergen questions for some aspects of phase polyphenism. I will highlight some key studies that address these questions, but it is in no way my intention to summarize the entire literature on locust phase polyphenism[2].

The problem of applying the Tinbergen questions to phenotypic plasticity has been pointed out before [18]. Depending of whether the plasticity itself, or the alternative phenotypes are considered the adaptation under study, the formulation of the questions is different. Phase polyphenism encompasses having two contrasting phases, and having the ability to transition between both. As a result, the exact

[2]There has been about a century of published research in phase polyphenism, ranging from direct observations of natural populations to gene expression studies in lab strains. A great body of information can be found in the Anti-Locust Bulletins, that were published between 1948 and 1971 by the Anti-Locust Research Centre (ALRC), headed by Boris Uvarov, and in Uvarovs cardinal work "Grasshoppers and Locusts: a Handbook of General Acridology" published in 1966 and 1977 [265, 266]. A comprehensive review about locust phase polyphenism was written by Meir Paul Pener and Stephen J. Simpson in 2009 [195], and multiple smaller reviews have since been published [277, 238, 38, 282, 86]. Since locusts are a potential threat to food security, there is also a large body of information available through the FAO and CIRAD.

formulation of the two first questions depends on the level on which we focus. I chose to zoom in on the transition itself for the first question, while focusing more on the long-term phases for the third question. These are my four "Tinbergen questions" for locust phase polyphenism:

1. How does the transition from the solitarious to gregarious phase state work mechanistically?
2. Why do locusts undergo phase transition and how are extreme phases best adapted to their environment?
3. How do long-term phases develop their typical phase-related traits, and how are these traits transmitted over generations?
4. How did phase polyphenism evolve and what are the phylogenetic relationships between locusts?

### 1.4.1 The mechanisms of phase transition

The question of how phase transition works mechanistically, is studied from multiple points of view, ranging from the physiology of the individual to the collective movement of the crowd. The bigger picture is that phase transition works through the positive feedback between all these levels.

The process starts with an increase in local population density, caused by an interplay between population dynamics, weather and landscape architecture. When resources are plentiful enough to support the population, densities may rise without phase transition. However, when resources dwindle, hungry locusts concentrate on the remaining patches of vegetation. At first, they are repelled from their conspecifics. However, the sensory input of the sight, smell and touch of other hoppers leads to a cascade of molecular and physiological changes in the central nervous system (CNS), resulting in a whole series of behavioural changes. Being less and less repelled from each other, the number of encounters increases, generating positive feedback that ultimately leads to a solid gregarious phase state.

Upsurges do not happen randomly, but typically nucleate at hot-spots within the recession area, both on a large geographical scale [11, 68] and on the scale of the landscape [70, 67]. The critical density at which phase transition occurs depends on the landscape structure. This has been observed in wild populations [140, 279, 216, 266], in field experiments [140, 67], in the lab [57, 70], and with computer simulations [57]. The quality of food is also important: on optimal diets, hoppers spend more time in postprandial quiescence, lowering the intensity of gregarizing stimuli [69].

In *Schistocerca gregaria*, the main sensory inputs are either the combined sight and smell of other locusts [213], or being touched by other locusts [213]. The hind legs are the most sensitive body part [233, 218], but this is not necessarily the case for every locust species [63]. Upon sensing an increase in local population density, the CNS reacts within minutes to hours, with altered titers of a variety of neurochemicals [219], and notably, an increase in the total amount of serotonin (5-hydroxytryptamine, 5-HT) in the thoracic ganglia (TG) as measured by HPLC in tissue extracts [5]. Anstey *et al.* (2009) found that in the desert locust, this serotonin burst is necessary and sufficient for early gregarious behaviour to unfold, but already after less than 24 hours, serotonin levels drop below solitarious baseline levels. Also, in the thoracic CNS, the neurons responsible for the early burst are distinct from the neurons that constitutively produce higher amounts of serotonin in solitarious locusts [220]. These findings suggest that different factors are involved in the initiation and consolidation of gregarious behaviour. Serotonin typically signals through G-protein coupled receptors, firing a signal transduction cascade leading to downstream effects such as modulation of the behaviour of neurons, and altered gene expression. A good candidate effector in this serotonergic signalling cascade is protein kinase A (PKA), which is known to be involved in memory formation in a variety of animals. Although a *direct* link between 5-HT signalling and PKA has not yet been demonstrated in the desert locust, a central role for PKA in the acquisition of gregarious behaviour was demonstrated by Ott *et al.* (2012). Pharmacological inactivation or RNAi mediated knockdown of the catalytic subunit PKAC1 led to decreased gregarization response, while knockdown of the regulatory subunit PKAR1, which is an inhibitor of PKAC1, led to an increased gregarization response [187].

How exactly these *immediate* and *early* responses lead to behavioural change is unknown. We know that PKA-mediated signal transduction can alter gene expression by activating transcription factors, such as *CREB* (*cAMP response element-binding protein*), a conserved protein with known roles in long-term memory formation. Transcriptomic responses to only a few hours of crowding have been demonstrated in *Schistocerca* [12] and *Locusta* [109]. In the latter, the gene products of *takeout* (*TO1*) and *chemosensory protein* (CSP-3) were found to increase as a function of isolation and crowding respectively. Knockdown of these gene products also influenced the rate of repulsion loss during crowding experiments [109].

### 1.4.2 Why locust swarm, sometimes

Collective behaviour is widespread in the animal kingdom, functioning as a foraging strategy, a safe means of travelling long distances, or a way to lower individual predation risk. Though often co-occurring with social interactions, collective behaviours need not necessarily be cooperative or altruistic.

As far as we know, there is no predictable periodicity in the outbreaks of locusts. In the desert locust, the majority of generations are solitarious[3], which raises the question why most of the time, solitarious behaviour is a premium survival strategy, while under specific circumstances, the choice for gregarious behaviour is made.

The prime reason why locusts engage in collective behaviour, is thought to be the reduction of the risk for each individual to fall victim to predation. Foraging theory predicts that in a landscape composed of potential food patches, predation is only sustainable if after depleting a food patch, there is a sufficiently high chance of finding a new one within the range the predator can easily move. If a critical proportion of potential food patches turn out to be empty on arrival, the food patch network is too interrupted to sustainably support the predator [207]. Consider a clump of vegetation containing locusts above a critical density to be a potential food patch for predators such as lizards, small mammals and birds. With a growing solitarious locust population, the proportion of colonized clumps of vegetation rises. This increases the connectivity of the food patch network of the predators, allowing them to "percolate" through the landscape with high predation success. In a theoretical study, Reynolds *et al.* (2009) [207] developed a model that predicts the existence of a critical population density at which each individual locust experiences a higher predation risk when retaining a dispersed spatial distribution, than when assuming a new, extremely clumped distribution, abruptly distorting the connectivity of the predators' food patch network.

The newly assumed clumped distribution presents a range of challenges to gregarious locusts. They, on their turn, need to be able to percolate through the landscape between potential food patches, and given their numbers, they quickly deplete every patch on their path. The behaviour of the crowd is an *emergent property* of collective stereotypic movements and behaviours. One of the direct drivers of movement may be cannibalism, especially in nymphs [20, 114]. The main driver of coherence is the tendency to align the body axis with direct neighbours, while keeping a certain perimeter of "personal space" [36]. It was proven that the mere alignment of the body to neighbours provides an efficient way of transferring information (e.g. about an approaching predator, or about the location of a food source) through a swarm [247].

Density-dependent aposematism is commonly seen in insects, including non-swarming grasshopper species [6], and simulations show that warning colouration is most advantageous under crowded conditions in the presence of chemical defence [72]. When desert locusts nymphs aggregate, they quickly start investing in chemical defence by eating toxic or unpalatable plants [71], while their cuticle colour can only change over multiple successive moults. Lizards learn to avoid foul-tasting black-yellow hoppers after only one contact [248, 253], but fail to do so with green nymphs [253]. Green nymphs may be able to deter predators by regurgitating their

[3] Although outbreaks may have been more common in times before pest control.

toxic gut contents [249], but the full effectiveness of aposematism is only reached when warning colouration is in place.

### 1.4.3 The development of phase polyphenism

Locust hatch with a predisposition towards the parental phase state [127], which can be considered the "best bet" for optimal survival chances. However, upon changes in their social environment, mothers are able to modify their offspring's phase state until the late stages of oogenesis, as was demonstrated by the shift in phase state of hatchlings from long-term solitarious mothers that were crowded only in the hours before oviposition [128, 167]. Proposed mechanisms underlying these seemingly inherited phase characteristics are factors in the egg foam [177], the size of the eggs [168], and epigenetic "imprinting" [28].

There is good evidence that in the desert locust, the gregarious behaviour acquired after short periods of crowding is both physiologically and molecularly distinct from the consolidated long-term gregarious phase state: different serotonergic neurons are activated [220], and different regions of the genome are methylated [172]. Also, behavioural solitarization of gregarious animals is a slow, two-stage process, while recently acquired gregarious behaviour is lost more easily, though there are conflicting observations about the rate of behavioural re-solitarization after isolation of previously crowded solitarious locusts [104, 214, 32, 243].

Some phase-related traits, such as body form or colouration, are physically constrained from undergoing rapid changes. The greater longevity of solitarious animals is also fixed before or during nymphal stages: if solitarious animals are crowded after their ultimate moult, they may acquire gregarious behaviour, yet they will retain their solitarious-like longevity, outliving their gregarious counterparts by about a month [28].

The above observations show that not all phase-related traits are equally labile, and they suggest that behavioural aspects of early phase transition are initiated through labile modifications of the CNS, while differences between extreme phases are more stable. Both in *Locusta migratoria* and in *Schistocerca gregaria*, differences between extreme phases rely on large-scale alternative gene expression throughout the lifetime of the animal [48, 14, 165, 281]. While there is overlap in the gene ontology (GO) terms of genes that react to crowding [109], it is presently unknown whether the same molecular pathways are involved in the initiation, consolidation and maintenance of phase state.

In the desert locust, endocrine differences between long-term phases include differences in juvenile hormone [194], corazonin [246], pyrokinins [276], neuroparsins [53] and a phase-related peptide of unknown function [203]. Microarray studies

comparing long-term phases in the migratory locust [165] and the desert locust [14] have found phase-dependent differential expression of genes involved in immunity, longevity, DNA methylation, metabolic rate, stress response, sensory processing and hormone signalling. At least part of the differences in gene expression between long-term phases seem to be due to epigenetic modifications, but little is known about the timing of this regulation, the identity of epigenetically regulated genes, or the relative contribution of mechanisms like histone modification or DNA methylation.

### 1.4.4 The evolution of locust phase polyphenism

The species that are commonly referred to as locusts belong to 6 different subfamilies within *Acrididae*, and they do not form a monophyletic group [237, 195]. One would think that there must either be a tendency to easily evolve the "locust syndrome", or it must have been the ancestral state that got lost multiple times. The latter explanation would not be parsimonious, given that less than half a percent of known *Acrididae* species are locusts, and phylogenetic analysis suggests large scale convergent evolution [237].

However, scattered through the phylogeny of *Acrididae*, we see varying degrees and combinations of density dependent changes in appearance and behaviour in species that are not typically considered locusts. Many non-swarming grasshopper species show plastic aposematic warning colouration, while the Australian plague locust *Chortoicetes terminifera* undergoes phase transition without obvious changes in colour [107]. There is also intraspecific variation: a population of *Locusta migratoria* from Israel is known to swarm, but does not change colour [118]. In the non-swarming *Schistocerca lineata*, populations mainly feeding on the toxic *Ptelea* show stronger colour changes than populations feeding on the non-toxic *Rubus* [250].

The mechanisms underlying phase transition in different locust species are overlapping, but not identical. Desert locusts gregarize easily by touching their hind legs [233] while in *C. terminifera*, the main tactile input pathway is through the antennae [63]. *Locusta* and *Schistocerca* differ in their time course of phase transition, and different biogenic amines play central roles in their early behavioural gregarization.

Different aspects of phase polyphenism are also under independent physiological control. The band-winged grasshopper *Oedaleus asiaticus* exhibits colour variation that correlates only weakly to behavioural change in a high population density regime [42]. In several locust species, the manipulation of corazonin influences melanization and morphometrics, but not behaviour [123, 258, 255]. In the desert locust, melanization can be induced by the smell of other locusts alone, while the combination of sight and smell is necessary for behavioural phase transition to

complete [152]. This may explain why in high-density desert locust populations[4], varying degrees of black patterning may already be present prior to behavioural phase transition [256].

These findings do not support the idea that phase transition would evolve as a syndrome, but rather that in different lineages, density-dependent differences in colour, morphology, reproduction and gregarious behaviour have independent evolutionary histories. The species that we call locusts, are those species in which these aspects co-occur, and might have become more tightly coupled throughout evolution. As a result, the question of whether a grasshopper qualifies as a locust is to some extent semantic, and moreover, judgement is often hampered by incomplete knowledge of the biology of the grasshoppers in question, because controlled experiments are lacking, or field and lab observations contradict each other [238].

The genus *Schistocerca* comprises about 50 species, of which *Schistocerca gregaria* is the only Old World representative. Among the New World species, varying degrees of polyphenism are seen, and in at least two species (*S. cancellata* and *S. piceifrons*), the gregarious phase shows the whole suite of gregarious behaviours, including hopper band formation, swarming, group mating and group oviposition [237, 162].

The phylogeny of the genus *Schistocerca* and hence the origin of their biogeographical separation have long been controversial [237, 162, 240], but consensus has settled that *gregaria* is a sister group to one monophyletic clade that contains all New World *Schistocerca* species [162, 239]. The presumed scenario is a colonization of the Americas after a trans-Atlantic flight by an ancestral *gregaria-like* species, followed by an adaptive radiation [162]. Such long distance flights over thousands of kilometres have been observed multiple times, including a trans-Atlantic flight as recently as 1989 [61]. The basal position of the desert locust within the genus also implies that phase polyphenism is plesiomorphic in *Schistocerca*, and that it was lost in the majority of New World species.

[4] And perhaps also in our lab.

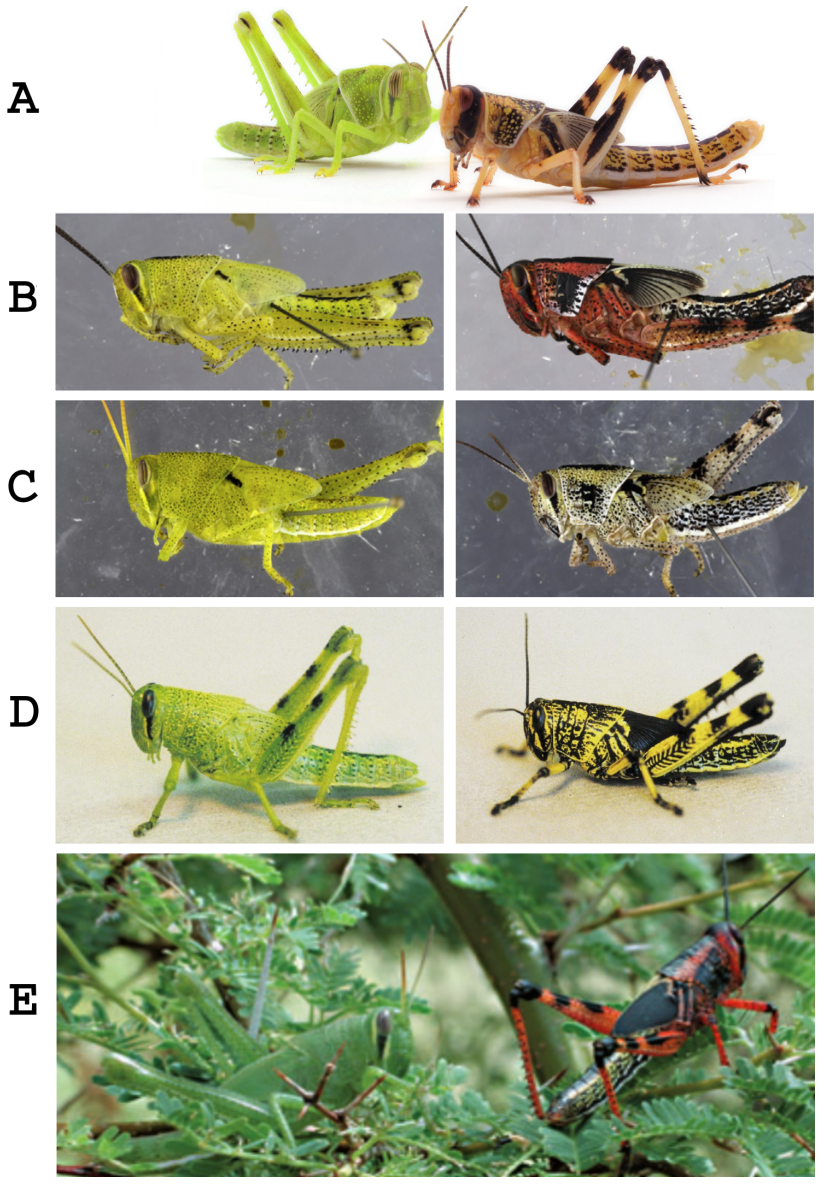



**Figure 1.4:** Final instar nymphs of five *Schistocerca* species, with the isolated/solitarous phase on the left, and the crowded/gregarious phase on the right. A: *Schistocerca gregaria*, the only Old World species, has extreme phenotypic plasticity (© Boris Godfroid). B and C: *Schistocerca americana* from the Southern US and *Schistocerca serialis cubense* from the Caribbean and Florida are not known to swarm, but show a certain degree of behavioural plasticity in the lab [251, 239] (from Gotham & Song 2013 [105]). D: *Schistocerca lineata* from North America does not swarm, and shows varying degrees of aposematism depending on the population [250] (from Sword 2002 [250]). E: *Schistocerca piceifrons* from central and South America is a swarming locust that can cause significant crop losses [117] (from Song & Wenzel 2008 [240]).

## 1.5 Context and objectives of this work

### 1.5.1 Context

The project presented here focuses on the early hours of phase transition in the gregarizing desert locust. Although many aspects of this transition have been studied, major questions are still outstanding. Many of these questions revolve around the syndrome-like nature of phase change. Though in long-term phases, many traits are simultaneously present, they each change at their own pace during phase transition, and seem to be under physiologically distinct control [103, 136, 152, 66, 240].

A range of behaviours are subject to change during phase transition, but it is unknown to what extent they share the same molecular or physiological pathways (fig. 1.5). This is an important consideration when setting up experiments that should shed light on the mechanisms underlying behavioural phase transition, because the interpretation of behavioural outcomes depends on the extent to which different aspects of behaviour are mechanistically coupled. In many studies, behavioural gregarization has been quantified using logistic regression models based on parameters related to at least two important categories of behaviours that can be observed in an arena assay: activity, and attraction/repulsion [215, 214, 213, 233, 123, 218, 152, 5, 63, 64, 187, 242, 243]. This method does not explicitly require both aspects of behaviour to change at a similar pace or to the same extent, which has sparked discussion about the interpretation of e.g. the role of 5-HT in phase transition [257, 217].

In the desert locust, a very direct causal relationship has been observed between a

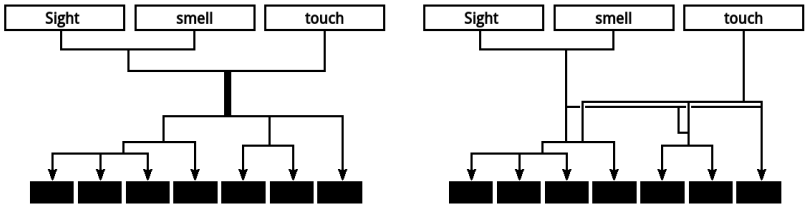



**Figure 1.5:** Schematic representation of two possible scenarios for the mechanisms underlying changes in behaviour. Black boxes represent behaviours that are subject to change during phase transition, such as food preference, the propensity to walk, or attraction towards conspecifics. Connecting lines represent pathways. These may include endocrine signalling pathways, neuronal circuits in the brain, or even molecular signal transduction pathways within single cells. In the first case, the two main sensory input pathways that are known to provoke behavioural changes converge into a single pathway. In the second case, the same sensory information leads to different aspects of phase-related behaviour through independent pathways.

few specific sensory inputs, and a change both in activity and attraction or repulsion. Under laboratory conditions, the combination of sight and smell of other locusts causes behavioural phase transition [213], as does tactile stimulation of the hind legs [233, 218]. However, the presentation of either sight or smell, or touch of other body parts does not have this effect. Sensory inputs can be by-passed on several levels: electro-stimulation of sensory input neurons of the hind legs has a gregarizing effect [218], as does the administration of 5-HT to the thoracic ganglia [5].

These experimental data strongly suggest that in the desert locust, sensory information from different channels is integrated in a process that ultimately leads to changes in at least two aspects of phase-related behaviour. It is unknown whether this pathway is mostly governed by physiological neuro-modulation, or whether it depends on large-scale changes in gene expression, or both. Though 5-HT and PKA have been identified as candidate mediators [5, 187], it is not known at which level of this putative core pathway they act.

Based on microarray studies, there is evidence for crowding-induced changes in gene expression in the heads of $4^{th}$ instar *Locusta* [109]. After 1 and 4 hours of crowding, 93 and 165 genes changed their expression. An unpublished microarray study on pooled $5^{th}$ instar *Schistocerca gregaria* CNS samples has also identified 48 and 33 genes that respond to 2 and 4 hours of crowding [12]. However, in neither of both cases it is clear whether differentially expressed genes are involved in a specialized core molecular pathway that leads to very specific phase-related behavioural changes, or whether they are part of a broader response to manipulation and changes in sensory input.

### 1.5.2 Objectives

The principal objective of this thesis is to gain a better understanding of early behavioural change, both in terms of the manifested behaviours, and the molecular mechanisms. In order to create the right conditions to experimentally address these two aspects, a number of sub-aims are defined:

1. To construct a good quality reference transcriptome that can serve as a source of sequence information for future experiments.
2. To further develop a reliable behavioural assay, based on improved insights into the nature of behavioural change.
3. To design and conduct an experiment that provides a better understanding of the molecular mechanisms that lead to behavioural change upon the perception of gregarizing stimuli.

# Chapter 2

# General Methods

> Je suis de ceux qui pensent que la science est d'une grande beauté. Un scientifique dans son laboratoire est non seulement un technicien : il est aussi un enfant placé devant des phénomènes naturels qui l'impressionnent comme des contes de fées. Nous ne devrions pas laisser croire que tout progrès scientifique peut être réduit à des mécanismes, des machines, des rouages, quand bien même de tels mécanismes ont eux aussi leur beauté.
>
> Marie Skłodowska Curie

## 2.1 Introduction

In this chapter I present the methods used in this work. Since this is the first RNA-Seq project on the desert locust in our lab, there is unavoidably a prominent role for software and bioinformatic techniques. I have intentionally conserved some rather elaborate descriptions in the main text, in the hope that this chapter may be used as a reference for future experiments by colleagues and myself.

## 2.2 Locust breeding

The *Schistocerca gregaria* colony in our laboratory was originally isolated from a wild population in Nigeria, and kept in captivity in the Aquazoo of Düsseldorf (Germany). In 1985 it was brought to our lab. Throughout the course of this project, two separate laboratory colonies (solitarious and gregarious) have permanently been kept, alongside a colony of gregarious migratory locusts (*Locusta migratoria*). Except during a couple of black-outs, there has been a constant 14h light, 10h dark regime, with temperatures within the range of 28-31$°C$.

Our animals were fed on a diet of fresh collard greens (*Brassica oleracea*) cultivated on a field outside of Leuven, cabbage (*Brassica oleracea*) from the fresh market (Flandria), and rolled oats (*Avena sativa*). I have no hard data on the performance and fitness of the colony during long periods of feeding either collards or cabbage, but my general impression has been that locusts are healthier and more fecund when fed collards. Perhaps they have superior nutritional quality, or perhaps the cabbage sometimes contained traces of insecticides. In any case, preference was given to our self cultivated collards whenever they were available. Feeding took place every morning between 8 and 11 a.m..

In order to ensure the maintenance of the gregarious phase state through all life stages, gregarious locusts were kept with at least 100 animals in acrylic plastic cages (40 x 43 x 85 cm) cages. Solitarious locusts were kept in rooms separate from the gregarious colony, and upstream of their air supply to prevent olfactory contamination. Solitarious lines were regularly supplemented with solitarized animals from the gregarious colony, in order to prevent both lines from genetically drifting apart. Only solitarious animals that were at least $3^{rd}$ generation isolated were used in experiments.

At the beginning of the project all solitary locusts were kept as described by Hoste *et al.* (2002) [122]. We noticed that, according to the assay that was used back then, phase transition in our solitarious locusts would sometimes be slow and incomplete, compared to the time course reported from animals of other colonies (e.g. in Cambridge or Oxford) [5, 25, 187, 243]. I hypothesized that the problem was a lack

of total olfactory isolation, which would desensitize their responses to conspecifics. The containers were also small, restricting the locust's opportunity for locomotion, which may have been reflected during behavioural assays. In order to overcome these problems, and to increase our solitary locust breeding capacity, we designed a new breeding facility that would provide total visual and olfactory isolation, and more space per animal.

Our new solitary locust breeding facility consisted of two white PVC cabinets, each consisting of 8 rows of individual cages with individual transparent perspex windows in the front, and double bottoms of which the upper layer consisted of a perforated stainless steel sheet. Because of their resemblance to blocks of flats, the cages were called "apartments". Each apartment was individually ventilated through a hole in the back, allowing air to flow in from a plenum that was directly connected to the room ventilation, creating a constant moderate over-pressure. The first unit, called "the small apartments", contained 80 individual cages of 10 x 20 x 14 cm. These were mainly used for early instars. The second unit contained 40 larger individual cages of 20 x 20 x 14 cm (fig. 2.1).

Since usually the entire capacity of the apartments was used for experiments, most breeding still took place in the old facility, and animals were transferred to the small apartments after their second instar moult, from where they would usually be transferred to the large apartments around their $4^{th}$ instar moult. The large apartments

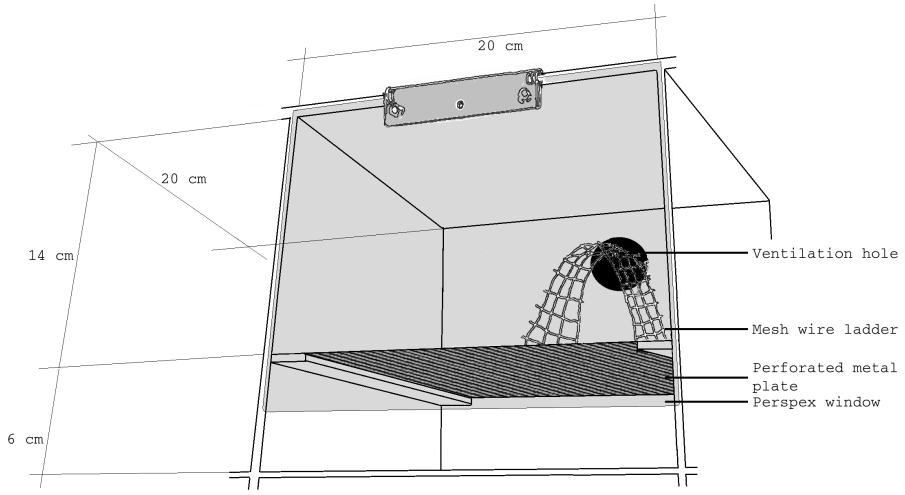



**Figure 2.1:** Drawing of one large apartment unit. Small apartments were half the width but otherwise identical.

were preferentially used for experimental animals, but because of the double capacity of the small apartments, we have also kept experimental animals there until $5^{th}$ instar. From the moment the new facility was operational, only solitary locusts that had spent at least the last two weeks in this facility were used for experiments.

## 2.3 Tissue collection

In this project, many different tissues were isolated from locusts. The reference transcriptome contains transcripts from a variety of tissues, and I dissected out brains and thoracic ganglia for the experiments described in chapter 5. All dissections were executed with the dissection dish sitting on ice, and under saline between 0 and 4°C (Ringer solution without added sugars: 8.766g $NaCl$, 0.188g $CaCl_2$, 0.746 g $KCl$, 0.407 g $MgCl_2$, 0.336 g $NaHCO_3$ in 1L miliQ®, pH 7.2). The main focus in this work will be on the brain (fig. 2.2). To dissect out the brain, the animal was decapitated with scissors and the head was placed in the dissection dish facing upwards and pinned down through the mandibles. Next, the antennae were cut away and an incision was made between the eyes. With a fine pair of scissors, the frontal exoskeleton was detached by cutting through the compound eye, downwards to the clipeus, following the line between clipeus and frons. After removing the exoskeleton and air sacs, the connectives were cut about 1 mm under the tritocerebrum. With a pair of forceps, the optic lobes were gently scooped out of the eye exoskeleton to fully detach the brain from the rest of the head. For the experiments described in chapter 5, the optic lobes were removed by cutting them at their narrowest point of the optic stalk (under the lobula), and the retina was removed by gently pulling it apart from the lamina with fine forceps.

## 2.4 General lab procedures

All procedures were executed by me, in our lab (Naamsestraat 59, Leuven), at room temperature (RT, 20°C) unless specified otherwise.

### 2.4.1 RNA extraction

In order to minimize RNA degradation, all dissections were executed on ice, and all samples were snap-frozen in liquid nitrogen. Pre-extraction samples were always stored at -80°C.

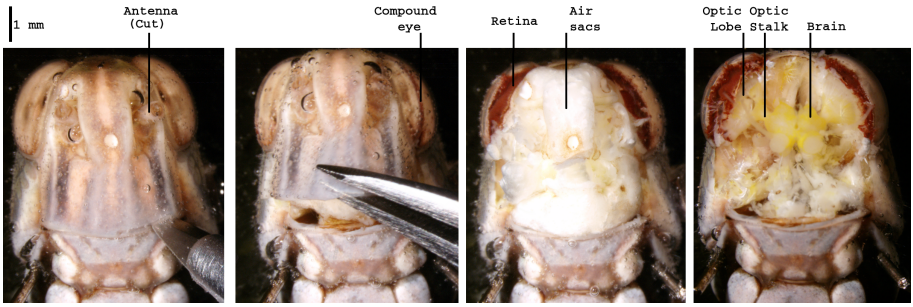



**Figure 2.2:** Dissection of a locust brain (an adult in this example). From left to right: the head is pinned down. The frontal exoskeleton is detached and removed. On the inside of the eye, the retina (red) is clearly visible. After removal of the air sacs (white structures), the brain appears. The arrow points at the protocerebrum, but the entire yellow structure is the brain (see also figure 1.3). The optic lobes are removed at the narrowest point of the optic stalk.

All RNA extractions were done with the QIAGEN RNeasy Lipid Tissue Mini Kit. Standard protocols were always followed, and the extra DNase digestion step was always executed. Briefly: samples were introduced in 2.0 ml MagNA lyser™ Green Beads vials (t1) with 1 ml QIAzol lysis reagent buffer, which contains phenol. Next, vials were placed in the MagNA lyzer™ instrument, and vigorously shaken (6000 x g) for 30 seconds, allowing the "green beads" ($\sim$ 1 mm ceramic beads) to thoroughly homogenize the sample. Next, 1 ml chloroform was added, the vials were shaken for 15 s and left to incubate for 3 min. After centrifugation (15 min., 5000 x g, 4 °C), the upper, colourless aqueous phase (about 2 - 3 ml) was transferred to a fresh tube and mixed (vortex) with 3 ml 70% ethOH. Next, samples were loaded (4 ml at a time) on RNeasy Midi spin columns and centrifuged (5 min., 5000 x g, RT) and an on-column DNase digestion was executed according to standard protocols. Column membranes were washed twice with 2.5 ml Buffer RPE (ethOH buffer) (5min., 5000 x g, RT). Finally, purified RNA was eluted in 50-80 µl RNAse free water (incubate for 1 min., spin at 5000 x g for 5 min, RT). In order to not further disturb samples post extraction, a 10 µl aliquot was always separated for quality control.

### 2.4.2 mRNA purification

Some RNA samples were further enriched in mRNA with the Illustra™ QuickPrep Micro mRNA Purification Kit (GE Healthcare). This kit is designed to directly extract mRNA from tissues, but purified RNA samples in RNAse free water can also be used as starting material. The basic principle of this method is to immobilize 3' polyadenylated RNA molecules onto an oligo-dT cellulose resin, while any RNA without poly-A, or

other contaminants such as DNA, can be washed away. The final step is the elution of poly-A enriched RNA into a volume of 0.2 ml elution buffer.

### 2.4.3 cDNA synthesis and sequencing library preparation

Two kits were used for cDNA synthesis: Illumina® TruSeq® and Clontech SMARTer™. The latter is not a full Illumina® library preparation kit, so after cDNA synthesis, samples underwent the library preparation steps provided by Illumina® TruSeq®.

#### 2.4.3.1 Clontech SMARTer™

Clontech SMARTer™ is a cDNA synthesis kit that is specialized in yielding cDNA that covers the full length of RNA transcripts. It makes use of the SMARTScribe™ reverse transcriptase that is capable of template switching. Double strand cDNA synthesis is a two-step process where after first strand synthesis, the reverse transcriptase adds additional nucleotides to the 3' end of the newly formed first strand of cDNA (hereby actually exerting terminal transferase activity). These non-template overhanging nucleotides base-pair with an oligonucleotide of known sequence (II A Oligonucleotide) creating a template that the reverse transcriptase replicates until the end of the oligonucleotide. The result is still an RNA - DNA hybrid, but with a small stretch of dsDNA with a well recognizable sequence at the 5' end of the RNA template. This "anchor" sequence along with the 3' poly A sequence serve as universal priming sites for end-to-end cDNA amplification, generating double stranded cDNA. Since terminal transferase activity is virtually absent unless the transcriptase is situated at the 5' end of the template RNA molecule, this technique should deliver a high proportion of full 3' to 5' sequences, as long as input RNA molecules are intact. For second strand synthesis and PCR amplification, the reference manual strongly recommends using the Advantage 2 PCR Kit (which is not part of the SMARTer kit). The TITANIUM® *Taq* polymerase in this kit should be especially useful for the amplification of a wide size range of DNA fragments, which is a defining quality of a good cDNA library with as many full sequences as possible. The kit also contains the neutralizing monoclonal antibody TaqStart™ that recognizes the N-terminal of *Taq* polymerase. This antibody inhibits any non-specific enzymatic activity during PCR set-up at ambient temperatures. Only when thermal cycling starts, the antibody is denatured and the inhibition is completely and irriversibly lost. This mechanism is called "hot start" and should increase both sensitivity and specificity. The Advantage 2 kit comes with two buffers, but with little information about their differences in formulation (Advantage 2, which is the standard choice and Advantage 2 AS, which should be used in the case of non-specific background amplification).

#### 2.4.3.2 TruSeq®

TruSeq® is the standard sample preparation kit provided by Illumina®. The protocol starts with a poly-A-based RNA purification using oligo-dT attached magnetic beads. Next, the purified mRNA is chemically fragmented (300-700 bp) using divalent metal cation-dependent nucleophilic phosphotransesterification under elevated temperature. Most buffers and reagents provided by Illumina® come with little information about their composition. The manual does not specify which cation is used here, but I assume it is either magnesium or zinc. This strategy is different from the Clontech workflow, where cDNA synthesis precedes shearing. Here, cleaved RNA fragments serve as input for the first strand cDNA synthesis, which is done using random hexamer primers and SuperScript™ II reverse transcriptase, an engineered version of Moloney Murine Leukemia Virus Reverse Transcriptase (M-MLV RT) with reduced RNase H activity and increased thermal stability. Next, the second strand is synthesized by DNA Polymerase I and Ribonuclease H. In case the samples cannot immediately be further processed, the resulting double stranded cDNA is considered a safe stopping point, and samples can be stored at -20 °C for up to seven days.

### 2.4.4 Determining the quality and concentration of nucleic acid samples

In this project, many different methods were used to assess the quality, integrity and concentration of nucleic acids. Both in the KU Leuven genomics core and in our lab, extracted RNA quality and cDNA size distributions were measured with the Agilent 2100 Bioanalyzer, which is a microcapillary gel-electrophoresis on a chip. Different chips are available for specific kinds of samples. RNA 6000 Nano chips were used for RNA extracts and DNA 1000 chips and DNA *high sensitivity* for the assessment of cDNA quality. Samples (1 µl) and ladder solution are loaded into small wells on the chip, the chip is placed in the device where nucleic acids are separated over a voltage gradient while an intercalating dye is visualized by laser-induced fluorescence. The result called an *electropherogram*, which is represented either as a series of bands, or as an intensity spectrum. For total RNA extracts, the Agilent expert software calculates a quality measure called RNA integrity number (RIN-score) [227, 126], which is based on the entire electrophoretic trace, including the pattern of ribosomal peaks and possible breakdown products. The RIN-score is considered the golden standard in the field[1], and it is also the preferred method for sample quality assessment for the Beijing Genomics Institute (BGI) that sequenced our RNA-Seq samples. However, it has turned out to be unreliable for insect RNA, which commonly shows particular rRNA peaks because of alternative ribosomal breakpoints [286].

[1]Which I really do not grasp, because the algorithm is proprietary, and hence nobody really knows what the RIN number actually is.

Next to the fragment length-based QC, I used spectrophotometry to further assess sample quality. On the NanoDrop™ 1000 (Thermo Scientific™), a droplet of sample (1 µl) is applied onto a small optical surface, and absorbance is measured in a broad wavelength spectrum. The absorbance at 260 nm is used to calculate nucleic acid concentration, and the absorbance in other wavelengths is used to determine possible contaminations, such as phenolic agents or proteins. For RNA samples, $A_{260nm}/A_{280nm}$ ratios above 1.8 were considered good, and values above 2 excellent.

Apart from the techniques mentioned above, the genomics core unit uses several other techniques. They use two assays that are based on a detection dye that fluoresces upon binding to RNA: the Quant-iT™ RiboGreen® RNA assay and the Qubit™ fluorometer (Thermo Fisher Scientific). Finally, before sequencing, cDNA libraries were quantified using the qPCR-based KAPA library quantification kit.

## 2.5 Next generation sequencing

### 2.5.1 Sequencing by synthesis: Illumina®

Although the era of next generation sequencing started with the invention of pyrosequencing [222] (Roche 454), real high throughput sequencing only became fully accessible to the broader community with the advent of Solexa® [22] (later Illumina®) dye sequencing by synthesis [40]. During my experimental period, the maximum available sequence length was 2x100 bp (paired-end sequencing), but now readlengths of 2x150 bp are common. This is shorter than the reads generated by pyrosequencing (up to 700 bp), but this is compensated by a thousandfold deeper sequencing, and a more random error pattern and lower error rate (about 1%) [179, 283, 226]. In this project, transcriptome assemblies were mainly based on Illumina® paired end reads, while shorter single end reads were used for quantitative RNA-Seq. In order to understand the bioinformatics that are going to constitute a major part of the chapters to follow, it is therefore important to explain in some more detail how these reads are generated.

Illumina® sequencing by synthesis is a staged procedure involving the formation of clonal colonies on a flow cell, and their sequencing by reading out colour codes from incorporated fluorescently labelled nucleotides. The exact procedures are explained at www.illumina.com and illustrated in figures 2.3 and 2.4.

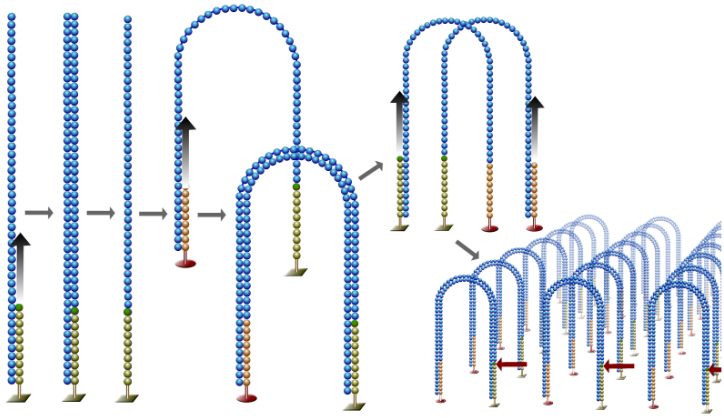


**Figure 2.3:** Schematic overview of colony formation through bridge amplification. First, the 3' sequencing adapter on the fragment anneals to an oligo attached to the surface, and a duplex is created. Next, the original fragment is washed away and the newly formed 3' end anneals to a second oligo and a second duplex is created. Now if this duplex melts, two 5' ends are attached to the surface and the 3' ends can now anneal to new oligos, starting another cycle. (wikimedia commons)

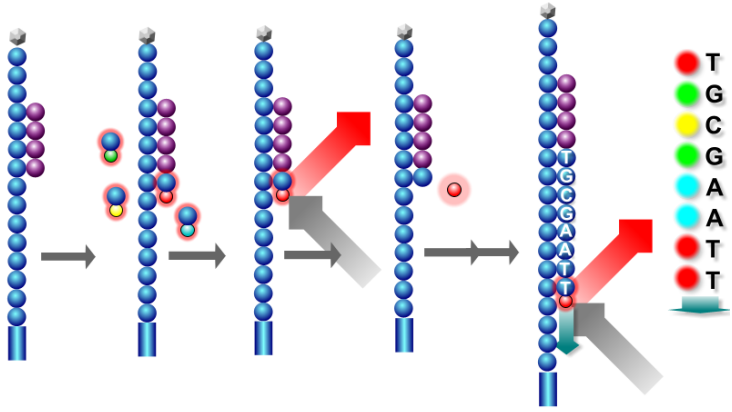



**Figure 2.4:** Schematic overview of sequencing by synthesis. After a sequencing primer has annealed, labelled bases compete for incorporation into the new strand. After exactly one complementary nucleotide is allowed to hybridise on each strand, the fluorescent label is excited with a laser, upon which it emits a light signal with a nucleotide-specific colour. Next, the label is washed away, creating a new free 3' hydroxyl group ready for the next cycle. (wikimedia commons)

Since this technique is not suited to directly sequence RNA, reverse transcription to a cDNA intermediate is always needed for transcript sequencing. In a first step, both ends of the dsDNA fragments are ligated with an adapter that contains the complement of the sequencing primers, terminal sequences for hybridization to flow cell oligos, and indices for multiplexing (if necessary, see further). Next, the DNA is denatured and loaded into a flow cell that is densely coated with two specific oligonucleotides that are complementary to the two terminal adapter sequences that are ligated to the cDNA. The fragments attach in a random pattern to the surface of the flow cell by hybridization of their 3' terminals to the oligos. These are the nuclei of the would-be colonies. Here it is important that the density of attached DNA strands is high enough to yield a sufficiently high number of sequences, but low enough to prevent interference between neighbouring colonies during amplification or sequence read-out. Now, these hybridized ssDNA strands will be amplified in a process called bridge amplification. The first oligo to which the ssDNA is hybridized serves as a primer for the synthesis of a complementary strand, and the original template strand is washed away. The oligo to which the template first hybridized is now part of the remaining complementary single strand. Now this single strand "bridges" its free 3' end towards the bottom of the flow cell, where in its turn it hybridizes to another oligo that is complementary to the second sequencing adapter (the free 3' end). Now, this oligo serves as the primer for a subsequent complementary strand synthesis. In a last step of the cycle, the resulting double strand DNA, that stands as a bridge between the two different oligos, is denatured. The resulting two neighbouring complementary single strands can now both serve as templates for a next cycle of bridge amplification. This process is repeated up to 35 cycles, resulting in several million immobilized colonies that each contain millions of ssDNA fragments. These colonies are not clonal yet, since they are a mixed population of original and complementary sequence. Therefore, the reverse strands are cleaved from the original oligos, and washed away. The resulting clonal colonies of only original forward strands are now ready for sequencing.

Sequencing primers are hybridized to the adapter sequences at the free 3' end of the ssDNA fragments. Next, sequencing starts with a replication-like process. Fluorescently labelled nucleotides compete for addition to the growing chain at the 3' end of the sequencing primer. When a correct base pairing is achieved, the nucleotide is incorporated and further extension is inhibited by the presence of the fluorescent tag. After all fragments in every colony are allowed to have their chains extended by exactly one nucleotide, the fluorescent tags are excited with a light source, and a fluorescent signal with a nucleotide-specific colour is emitted while the bond between the nucleotide and tag breaks, allowing a next cycle to begin. This process is repeated until the desired read length is achieved, which can take up to several days. The light signal is recorded simultaneously for all colonies in a series of consecutive images with differently coloured "spots". The computational translation of these colour signals, along with their intensities and qualities, is called "base calling". A

bright and unambiguously uni-coloured spot will be easy to assign an identity to, whereas a colony with a less pure light signal will have a certain degree of uncertainty in its base calling. Ambiguity in the light signal can be caused by the lagging behind of some strands after unsuccessful base incorporations during previous cycles. This *dephasing* accumulates with each cycle, causing the typical decay of base call quality towards the 3' end, which is an important limiting factor for the maximum achievable read length [182].

In case of single reads, the process stops here. In the case of paired-end reads, the complementary strands that were formed during sequencing (the read products) are washed away, and the free 3' ends of the already sequenced forward strands are again hybridized to the complementary oligos, generating a bridge again. In one round of bridge amplification, the complementary strands are formed again, and this time the original strands are cleaved off and washed away. Finally, the reverse strand is sequenced in the same manner as the original strand.

In the case of multiplexing, when multiple libraries are loaded per flow cell, indices (also called barcodes) are used. These 6 to 8 nucleotides are sequenced in a small separate sequencing reaction that precedes actual read sequencing.

### 2.5.2 The FASTQ file format and base call quality

After base calling and de-multiplexing (assigning a library identity to each read, based on the readouts of the indices), the final output format of the Illumina® sequencer (and of most sequencing platforms used today) is the FASTQ format (with either .fastq or .fq file extension) [56]. This format combines a FASTA-like sequence output [193] with base calling qualities. Some small variations may exist, but without exception, a FASTQ file has four lines per sequence: a description starting with "@", the sequence, a plus (+) sign where sometimes the description is repeated, and the quality scores. As an example, the first sequence of one of our own libraries looks like this:

```
1 @HWI-ST571:185:D111MACXX:1:2314:1400:2144 1:N:0:CTTGTA
2 CACAACTGTGGTGGCTCCATCCTCTCCAGCAATGGCCGTCGCCGGAGACCACGA
3 +
4 @@@FFDDDHHGHIJJII<GHHIJJIEEJGFBDEADDDDDDBB<9>B?########
```

The description is an identifier line that contains a lot of information separated by colons. This example is only one of many possible formats that are in use, and depending on the sequencing device, they can be built up quite differently. However, all FASTQ files generated in our own experiments have the same formulation of description lines. The description usually starts with a unique identifier of the sequencing device used (HWI-ST571, which is the HiSeq2000 sequencer at genomics core Leuven). This should be the same for the entire FASTQ file, unless results from different sequencers are merged together. The device identifier is followed by the

unique run identifier (185), the flowcell identifier (D111MACXX), the flowcell lane identifier (1), the tile number within the lane (2314), and the x- and y-coordinates of the colony within the tile. This is enough information to trace back every read to its original coloured spot on the images before base calling. Next, the description contains the membership of a pair (1 in this case and for all single-end reads, 2 if it were the second pair in case of paired-end reads), N or Y to indicate if the reads are filtered (N, not filtered), a number indicating if the read is an internal control (0 if not, a higher number according to the quality of the control), and finally the index, if present. These indices can be used for demultiplexing, if multiple samples were loaded onto the same flowcell lane. The sequence line is analogous to the sequence line in a FASTA file, and is just a string of a five letter alphabet: the four DNA bases (C, A, T, G), and N for a failed base call, starting from the 5' end of the read. The quality line is what really discriminates a FASTQ file from a FASTA. Qualities are encoded as ASCII characters, where each character corresponds to a certain quality score "Q" (ASCII characters have numbers). Multiple encodings are in use, but today the most generally used is Phred+33. This encoding is based on the original quality scoring protocols for Sanger sequencing by the program "Phred" [87, 88]. It starts off at ASCII 33 (the exclamation mark) and ends at ASCII 126 (the tilde) to encode quality values ranging from 0 to 93. Since base calling is a matter of colour signal and colour noise, the Q values are in Decibel units, which is a logarithmic scale for signal/noise ratios. They are formulated as follows: $Q = -10 * log_{10}(p)$ where p is the probability of the base call being incorrect. In usual Illumina® reads, Phred+33 quality scores above 50 (corresponding to 0.1 ppm per nucleotide error rate) are extremely rare, but when reads are merged into an assembly, well supported positions can get higher quality scores.

In the example, the first base has "@" quality score, which corresponds to ASCII number 64, and hence the Phred+33 score is 64 - 34 = 30, which comes down to a probability of $10^{-3}$, which is 1 in 1000. Note that the last couple of bases have quality score "#", which corresponds to ASCII 35, and hence their Phred score is only 2, which indeed comes down to a probability 63% of a wrong base call. This is a typical pattern that is often seen in Illumina® reads, where base call quality decreases exponentially at the 3' end of the read. This is mainly caused by the accumulation of sequencing reactions that lag behind within a colony. Low quality reads are usually clipped off during data pre-processing (see further).

Quality scores in a FASTQ file are only based on the signal to noise ratio in the readout of light signals (or other signals in case of other sequencing techniques) and their corresponding base calls. Several other sources of errors can influence the sequence before the actual readout. For example, replication errors may occur during library preparation, or in the first stage of bridge amplification. These sources of error are particularly dangerous, because they give rise to wrong sequences with high quality scores.

Reads can appear in a variety of orientations compared to the original template strand (in this case a transcript). A first source of variation is the cDNA synthesis process and the initial loading of ssDNA onto the flowcell (fig. 2.5). This generates a 50/50 proportion of original sequence and its complement creating colonies. This source of variation can be eliminated by using strand-specific protocols (not used in this project). Next, the first read that is sequenced, will be the complement of the complement of the original ssDNA, and the second read will be the complement of the original. In any case, two paired-end reads will always be pointing towards each other on opposite strands. In the special case where the input fragments are shorter than two times the read length, the paired-end reads will overlap. Since their orientation in the FASTQ file is 3' to 5' for both, the overlap will be situated at the "end" of both reads, and will be reverse-compliment.

In the case of paired-end sequencing, usually two FASTQ files are generated (although interleaved and merged style FASTQ files exist as well), with in the first one all the first reads (in some software packages also called "left" read) of the pair, and in the

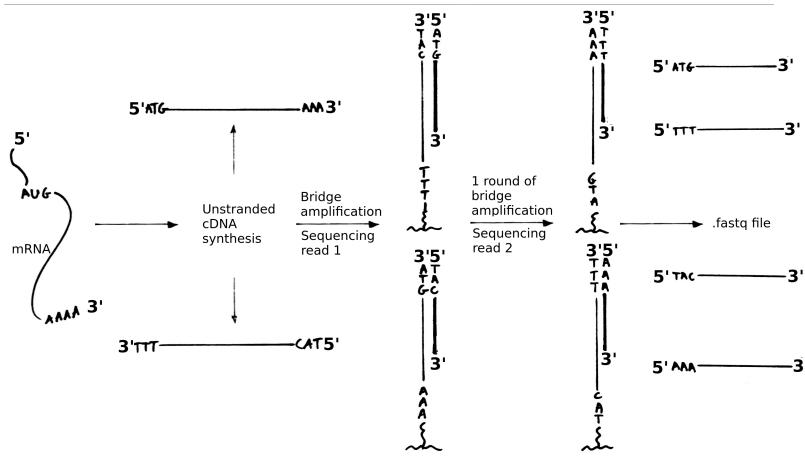



**Figure 2.5:** Schematic overview to demonstrate the origin of read polarity in the final FASTQ files in the case of Illumina® pair end sequencing with unstranded cDNA. In this simplified example, the cDNA is exactly capturing a stretch from start codon all the way down to the 3' poly-A tail. Sequencing adapters and primers are ignored for clarity. Note that in the final FASTQ, reads corresponding to opposite ends of the original cDNA strand are reverse complemented. In cases where the insert is smaller than twice the read length (so in this case, if the distance between start codon and poly-A was smaller than 200 nucleotides), the 3' ends of both reads would base-pair in the "innie" orientation. In the unlikely case the insert was smaller than 100 nucleotides, the 5' of the first read would overlap with the second read, and the other way around. This would be called the "outie" orientation.

second one all the second reads ("right" reads). These files should contain exactly the same number of reads in exactly the same order. If a read is discarded from either of two files, for example because of quality issues, it must be discarded from the other as well, in order to maintain compatibility between the two files.

## 2.6 Bioinformatics

### 2.6.1 Hardware and operating systems

During this project, two computers were used. My own workstation is a self-assembled desktop computer with Intel Core i7 3770 (3.4 GHz, 4 cores, 8 threads, 8 MB cache) processor on a Gigabyte GA-H77-DS3H motherboard, with 8 GB RAM. It runs on *Bio-Linux* (version 7.0) [31], which is a Linux distribution dedicated for bioinformatics, with many common applications and scripting languages pre-installed. This computer was used for all computations except for the memory intensive work on RNA sequences. The second computer is a computation server at the university of Leipzig (Germany). This server with Intel Xeon E7 8860 (2.27GHz, 80 cores) and 1 TB RAM runs on another Linux distribution called Fedora (release 20).

### 2.6.2 Scripts

In this project, I have extensively used shell scripts (Bash and zsh) and the statistical programming environment R. All scripts used in the analysis of locust behaviour and RNA-Seq are included in the supporting materials, with extensive annotations to aid future implementation. In case an illustration code snippet is needed in the manuscript itself it will be printed in a *codeblock*, as demonstrated here for the code to log in into the Leipzig server. In all scripting languages that I have used, the hash character (#) precedes a comment, which is a line that is not executed.

**Codeblock 2.1:** The code to log in to the Leipzig server. Secure shell (ssh) is an encrypted network protocol that allows users to run commands on remote machines over a network (in this case over the internet). This protocol is standard installed on most Linux distributions. Here, I log in with my user name "rik" on the server called "k61". The -X parameter allows forwarding of graphical content (windows) from the server to the receiving computer.

```
1 # Log in to Leipzig server:
2 ssh -X rik@k61.bioinf.uni-leipzig.de
```

### 2.6.3 Software

For most bioinformatic applications, multiple dedicated software packages are available. Often they perform very well, but many of them are difficult to install, and some of them have bugs. The choice between alternatives usually comes down to reading manuals, installing dependencies, and a lot of trial and error. The operating system Bio-Linux conveniently offers many pre-installed applications that work more or less out of the box. They are usually also the most tested and trusted alternatives according to the developers of the operating system [31]. In my experience, this especially smoothed getting started with bioinformatics without too much background. Whenever I could, I have therefore used the applications that were present in Bio-Linux.

#### 2.6.3.1 FastQC

A widely used tool for the assessment of read quality is FastQC[2]. This tool takes a FASTQ file as input, and computes a couple of statistics that summarize the quality of the data. During my entire analysis, I used FastQC at every step that involves FASTQ files. In this manner, the quality of the input material for sequence assembly was closely monitored. FastQC references to a configuration file that contains sequences of common contaminants. This file was supplemented with the *Schistocerca gregaria* mitochondrial genome, some sequences that are known from the ribosomes, and all adapters and primers that might have been used throughout sample preparation.

**Codeblock 2.2:** Standard command to execute FastQC on 20 threads (-t) with a quality histogram output per position (–nogroup)

```
1 fastqc -q -t 20 --nogroup file.fq &
```

#### 2.6.3.2 BLAST

Basic local alignment search tool (BLAST [1]) is one of the most commonly used tools in bioinformatics. It is a search algorithm to quickly compare a query nucleotide or amino acid sequence to a large database of sequences, and to identify those database entries that show similarity to the query sequence. BLAST can be executed on servers such as the one offered by NCBI[3] (which I have extensively used) or it can be run locally with a graphical user interface or from the command line. In this project I used the *Blastall* application and the NCBI BLAST® command line application. In both applications, an indexed database needs to be constructed based on a FASTA

[2] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[3] https://blast.ncbi.nlm.nih.gov/Blast.cgi

file with sequences. Next, the indexed database can be queried with one or multiple query sequences. I usually used tabular output, which puts hits in a tab separated data frame that can be read and further processed in R.

An example workflow with the NCBI BLAST® command line application may look like this:

**Codeblock 2.3:** Some commonly used commands when working with BLAST. Generally, it is convenient to work with single line FASTA files, where sequences are not chopped up in blocks. After generating a single line FASTA file, the database is created. This indexing step greatly speeds up future searches. BLAST comprises a variety of programs depending on the kind of query and database. In this case for example, blastn is used, which is a nucleotide versus nucleotide search. Other common BLAST programs such as blastp, blastx, tblastn and tblastx are available as well. Notice that the output format parameter 'outfmt' is used to specify exactly which output is desired. In a final phase, interesting sequences can be extracted from the database with a search with the function 'grep'.

```
1
2  # Convert database to single line FASTA:
3  awk '/^>/ {printf("\n%s\n",$0);next; } { printf("%s",$0);} END {printf(
       "\n");}' < db.fasta
4
5  # Create a database:
6  makeblastdb -in db.fasta -parse_seqids -dbtype nucl
7
8  # Now BLAST:
9  blastn -query query.fasta -db db.fasta -outfmt '6 qseqid sseqid slen
       pident length mismatch gapopen evalue bitscore' -out allhits.
       output
10
11 # Extract a transcript from a database.
12 # This only works with single line FASTA.
13 grep -A1 "name_of_contig_in_database" db.fasta
```

### 2.6.3.3 Multiple sequence alignment

For multiple sequence alignment (MSA), I have mainly used the command line version of MAFFT (multiple alignment program for amino acid or nucleotide sequences) [138]. This aligner scores well in benchmark tests both for nucleotide and amino acid sequence alignment [259, 188]. Moreover, for the alignment of large sets of unsorted nucleotide sequences, it has the option to re-organize a mix of sense and anti-sense sequences by orienting them all in their most likely sense direction.

#### 2.6.3.4 Mappers

Mapping is one of the most common tasks in bioinformatics. It is used in reference-guided assembly, in the assessment of assembly quality, in variant calling and in expression analysis [95]. Much like in (multiple) sequence alignment or BLAST, a mapper looks for stretches of similarity between two sequences. However, in the case of alignment, the input sequences are usually more or less of the same length, while mapping usually involves many short reads and one longer reference (such as a reference genome or transcriptome). The rules by which similarity is scored, are usually a lot simpler as well: while in (multiple) sequence alignment an evolutionary model is used, and a high degree of modification is possible (indels, mutations), to sometimes almost force the sequences into their most likely alignment, mapping goes by much simpler rules, usually allowing only a limited number of mismatches or indels. If a read does not find a place in the reference by these rules, it remains unmapped. In some cases these rules have to be loosened, for example when mapping to a reference genome of another species, or in cases where many mutations or variants are to be expected. A possible factor of complexity is split-read mapping, where a read (or a pair of PE reads) is split-up and mapped onto different positions in the reference. This is usually the case when mapping transcript reads to multiple exons in a reference genome [95].

Many different mappers are available, and they use a variety of algorithms to tackle the challenges pointed out above. Apart from the algorithms under the hood, they also differ in their ability to take care of indels or split-reads, or to take into account base-call quality (FASTQ format). The main players in the field today are BWA (Burrows-Wheeler aligner) [153, 154], which is based on the Burrows-Wheeler transform, a data compression algorithm that was originally developed for text compression. Another popular choice is Bowtie [148] (and its newer version Bowtie2 [149]) that also uses BW-transform. Originally, BW-transform-based mappers performed badly with indels and split-reads, but these issues have now mostly been taken care of.

I have tried all the above mappers, but finally decided to work with STAR [74] (Spliced Transcript Alignment to a Reference), which was relatively new at the time, but scored consistently high in benchmark comparisons in terms of speed and accuracy. This comes at the cost of consuming a lot of memory. STAR gained popularity and credibility after it was used as the prime mapper for the "encyclopedia of DNA elements" (ENCODE) project [59].

#### 2.6.3.5 SAM, BAM and SAMtools

After reads have been mapped onto a reference, they are typically outputted in the Sequence Alignment/Map (SAM) format [155]. A typical SAM file consists of an

optional header section that contains information about the reference contigs, and a multi-column tab delimited "alignment section" that has 11 mandatory field, and a range of optional columns. Each line in the alignment section represents a mapping, and contains information about the identity of the mapping read, the place where it mapped in the reference (contig, coordinates, strand,...), the quality of the mapping, whether it mapped anywhere else and much more. A SAM file for a large mapping project may become very large, and usually it is compressed to a binary format called BAM, containing exactly the same information, but not human-readable. The most commonly used toolset to interact with SAM/BAM files, is SAMtools [155]. SAMtools can compress SAM to BAM, decompress BAM to SAM, and sort, edit or interpret SAM/BAM files in different ways. It can also index SAM/BAM files, allowing faster look-up and counting.

#### 2.6.3.6 BEDtools

BEDtools [202] is a commonly used tool for comparative or arithmetic operations on certain statistics over intervals of a stretch of sequence, like averaging mapping depth over an interval of a contig, or counting variants. I used BEDtools to generate position-specific mapping counts for each contig starting from BAM files.

#### 2.6.3.7 The Trinity transcriptome assembler

Many software packages [228, 112, 51] are available for graph-based *de novo* transcriptome assembly[4] and among the many people I asked for advice, there seemed to be consensus that Trinity [112, 106] is reliable and relatively easy to use. At this moment, Trinity is presumably the most widely used and trusted *de novo* transcript assembler for large transcriptomes [51]. As a result, it has a large community of users with a lively on-line discussion group. As its name implies, it originally consisted of three modules (Inchworm, Chrysalis and Butterfly) that together made up a pipeline from clean input all the way down to assembled contigs. Thanks to this three-step procedure, the most computationally demanding part of De Bruijn graph-based sequence assembly gets partitioned into several smaller sub-processes that can be run in parallel. The composition of the Trinity program and the exact *modus operandi* have changed multiple times since the original release of the program in 2011. However, not all these changes have been published, and many of them are poorly documented despite their dramatic influence on how the program works and on the output it generates. The best information can be found in the release notes of all consecutive versions of the program on the code-sharing platform GitHub, but this is very technical and sometimes difficult to understand. Another source of help and information can be

[4]See supporting materials for an introduction to graph-based *de novo* assembly.

found on the dedicated Google discussion group, albeit with the caveat that anybody can contribute, and hence not all information is reliable.

The Trinity developers have an update paper in the pipeline (Brian Haas, pers. comm.), but for the time being, I here present my digest of how *de novo* transcriptome assembly in Trinity version 2.06 works.

**phase 1:**
This is the phase where the large number of unstructured reads gets partitioned into multiple smaller subsets. This will later greatly reduce the computational complexity of the graph traversing part. In a very first step, a program called Jellyfish builds a catalog (a hash-table) of all k-mers, along with the number of times they occur in all input reads. By default, a single k-mer length of 25 is used. Next, the program Inchworm takes the most abundant k-mer and starts extending it in both directions with a *greedy* algorithm: it extends the initial "seed" k-mer on both sides with those k-mers that show perfect k-1 (24 bp) overlap. In the case of multiple candidates, the most common k-mer is used. After the k-mers have been merged with the growing contig, they are removed from the hash table. The pre-contig is ready when no perfectly overlapping k-mers are left, and the process is repeated starting with the new most abundant k-mer, until no k-mers are left. Since every k-mer can only be incorporated once, this process generates linear pre-contigs that do not share a single k-mer. Sometimes this may lead to full-length *bona fide* transcripts that may well be the best representative for a number of alternative variants that share k-mers due to alternative splicing, paralogy or allelic variation. However, in other cases the pre-contigs may only represent fragments of real transcripts, and no information about isoforms is present at this stage. This is where a second program called Chrysalis comes in: it clusters Inchworm-contigs based on perfect overlaps of length k-1, called junctions. If a junction is supported by a sufficient number of reads, the two contigs are clustered. In early versions of Trinity, the next step would be to directly construct De Bruijn graphs of these clusters and proceed to assembly (Butterfly), but in newer versions, the process described above is repeated within each cluster.

**phase 2:**
In phase 2, the entire suite of Trinity programs is run within each of the clusters generated in phase 1. Since many smaller hash tables have to be constructed in parallel, another k-mer counting algorithm is used in stead of Jellyfish. Upon completion of k-mer counts per cluster, the greedy contig extension and clustering procedures are repeated within each cluster. Next, Chrysalis builds a separate De Bruijn graph for each resulting subcluster and partitions all input reads by assigning each read to the subcluster it shares the highest number of k-mers with.

In a last step, the program Butterfly will reconstruct full-length linear transcripts within these subclusters. This is the point where the actual reads come into play in an iterative processes of pruning and graph compaction. Firstly, read support is used

to resolve small ambiguities that are likely due to errors or rare alleles. This pruning step further simplifies the graph by merging near-identical nodes and removing low-support edges. In a next step, the graphs are compacted by merging consecutive nodes of low degree into single nodes representing unambiguous longer linear stretches. This process drastically reduces the number of nodes, but also abandons the strict rule of equal sequence length among nodes that defines a De Bruijn graph. After pruning and compaction are reiterated until convergence, the graphs should have become small enough to score all possible paths with dynamic programming procedures. It is important to note that these scores are based on their support by actual reads and read pairs rather than by k-mers alone. Since reads (and especially pairs) are longer than the k-mer length, they provide much more power to resolve combinatorial ambiguities for the exons that constitute different isoforms, which should lead to more reliable splice variant recovery compared to purely k-mer-based strategies. This is Trinity's strategy to tackle the specificity-sensitivity trade-off while sticking to a single k.

Upon completion of Butterfly, the final transcriptome can be found in a file called Trinity.fasta. Throughout its history, Trinity has used many different FASTA accession line formats in this file. Trinity version 2.0.6 outputs each contig with an accession line that contains information about the identity and origin of the contig. Each sequence has a "TR" number that corresponds to the initial read clusters generated during phase 1, followed by a (sub)cluster number (c) and a gene number (g) from Chrysalis and Butterfly in phase 2. These "gene" specific numbers are followed by isoform numbers (i). Finally, the accession line also contains information about the read length, and the path through the modified De Bruijn graph that was traversed to obtain the contig.

#### 2.6.3.8 Trimmomatic

In this project, the program Trimmomatic [30] was used for quality trimming of raw reads. This tool reads FASTQ files and removes adapters and trims off stretches of lower quality base calls with a sliding window approach, and it does this particularly fast thanks to multithreading. Trimmomatic is also interesting because of its "palindrome" mode. Here, reads are checked for overlap, and when more than complete overlap (read-through, a.k.a. "outie" orientation, see explanation fig. 2.5) is detected, adapter sequences in the first read can be clipped off with a threshold as low as a single nucleotide, thanks to the extra information provided by the overlapping mate. Next, the overlapping second read can be discarded because it contains no new information. Trimmomatic uses a configuration file containing all possible contamination motives that the data have to be filtered for. I extended the standard configuration file with all possible sequencing primers and adapters for Clontech, Nextera, TruSeq 2 and TruSeq 3 library preparation protocols.

#### 2.6.3.9 FLASH

FLASH (fast length adjustment of short reads) [169] is a tool to merge paired-end reads with overlaps. It splits up a library into non- (or little) overlapping paired-ends, and merged single reads. In the first place, this software looks at the "innie" orientation that is expected from Illumina® paired-end sequencing. However, it can also screen for the "outie" orientation. One parameter that should be carefully selected is the maximal overlap parameter (-M) that demarcates the expected 90% percentile of read overlap sizes (10% of the reads have longer overlaps). It is not used as a hard filter, but fragments with overlaps larger than this value are scored on a more stringent scale in order to avoid biases. When FLASH notices that many fragments appear to have overlaps larger than this parameter, it will throw a warning. FLASH generates three streams of output: a size histogram of the merged fragments, a single FASTQ that contains merged reads, and two paired FASTQ files that contain the presumed non overlapping pairs. Note that the latter can contain many reads that do overlap in reality, but by a quantity smaller than the pre-set number of nucleotides (default 10), or with a mismatch proportion larger than a pre-set value (default 25%). In the case a mismatch passes filtering, a consensus nucleotide is chosen based on the base call quality (Phred score), where the winner takes it all. The new quality score will be re-calculated as the absolute value of the difference in quality scores (with a minimum of 2). Since quality scores are on a logarithmic scale, this comes down to dividing the lowest error probability by the highest. This rule helps to maintain sufficient quality scores for situations with a clear sequencing error, while clearly "flagging" positions with low quality scores or with higher scores but no obvious "winner".

#### 2.6.3.10 insilico_read_normalization.pl

Trinity needs roughly one gigabyte memory per million reads to be assembled, regardless of their content. This means that reads of the most frequent transcripts will increase memory requirements and calculation times without contributing new information.

A good workaround to redundancy is *in silico* read normalization, also known as digital normalization. Here, a hash table of all k-mers and their counts is made, and reads are discarded from the data based on the frequency of their k-mers. In this project, I used the insilico_read_normalization.pl script that ships together with Trinity. Briefly: first, a catalog with k-mers (k=25) is made using Jellyfish and a hash table is constructed with their total counts. Next, reads are parsed and their k-mer abundances are calculated by looking up the abundances of all k-mers along their length in the hash table. For each read, the median, mean and standard deviation of the k-mer abundances along the length of the read are recorded. In a final step, reads are discarded with a probability proportional to their median k-mer abundance.

One potential problem with normalization, is the relative enrichment in sequencing errors, since errors create unique k-mers. To overcome this problem, reads with a standard deviation of their k-mer coverage above a certain percentage of the mean coverage are also discarded. The default setting for the *max_pct_stdev* parameter is 200. I keep my reservations here, for the simple fact that in the case of a k-mer size of 25 and a read size of 100, a single nucleotide sequencing error cannot possibly cause a drop in standard deviation substantial enough to make it surpass the mean by a factor two, while on the other hand, it is easily proven that rare variants (e.g. splice variants or paralogues) that share most of their k-mers with a more common variant would be discarded when this parameter is set to 200. (see supporting materials for a demonstration). I chose to set this parameter to 300, which in my simulations allowed to retain variants with a difference of a factor 100 in sequencing depth. This will yield little error correction, but I assume by keeping the maximum coverage at a high enough level, the correct reads will outnumber the sequencing errors by a high enough factor. The rest should be dealt with by the assembler.

#### 2.6.3.11 RepeatMasker

RepeatMasker [236] is a program that screens nucleic acid sequences for a variety of interspersed repeats and low complexity regions. Hidden Markov models are trained based on annotated repeat databases such as Repbase [17], and used to detect elements in new sequences. The output is a FASTA file where repeats are masked, accompanied by a list of hits per sequence, with the identity of the putative repeat element with a probability score.

#### 2.6.3.12 HMMER, TMHMM RNAmmer and signalP

High throughput sequence analysis for motifs or orthology is a computational challenge. BLAST is one option, but especially when specific domains have to be recognized, or when probabilistic inferences are needed, the algorithms of choice are usually based on *profile hidden Markov models* (pHMM) [75]. HMMER [76, 94] uses pHMMs to search "profile databases" such as Pfam [199, 93], which is a large database of curated multiple sequence alignments of proteins of a variety of gene families, with their own HMMs.

TMHMM [143] and RNAmmer [147] are similar applications that use HMMs for the prediction of trans-membrane helices and ribosomal RNA.

SignalP [196] recognizes secretory signal peptides using a combination of HMMs and artificial neural networks.

#### 2.6.3.13 TransDecoder

TransDecoder, which is one of the scripts that ships with Trinity, is a program for likelihood-based coding sequence (CDS) prediction in large sets of sequences.

For CDS predictions in a transcriptome, TransDecoder runs in two steps: first, all input sequences are translated into all 6 possible open reading frames (ORF), and a HMM is trained on the 500 longest CDSs. At this stage, all potential CDSs can serve as input to a BLAST or HMMER search in databases of known proteins or domains. In a second step, likelihood scores for all possible CDSs are calculated based on the HMM, and BLAST or HMMER hits if present, and a best candidate CDS is selected for each transcript above a likelihood cut-off.

#### 2.6.3.14 Trinotate

Annotation pipelines typically consist of two major stages. In a first stage, unknown sequences are matched to databases of known sequences. Often, nucleotide and translated protein sequences are queried in parallel. The main tools here are pHMMs, usually used for recognizing motifs or domains. In parallel, alignment-based tools like BLAST [1] are used to find evidence for one-to-one orthologues. In a second stage, inferred membership to gene families, or orthology to genes in better studied organisms, is used to assign putative functional information to as many sequences as possible, usually involving standardized vocabularies such as GO-terms [9].

Many software packages are available for this task [180, 58]. I experimented with a couple of options and finally opted for the Trinotate pipeline that ships together with Trinity [106], which is one of the most complete annotators. It also recognizes the structure of a Trinity transcriptome by its FASTA headers.

First, Trinotate uses TransDecoder, combining likelihood-based CDS prediction with BLAST [1] hits and domain and protein family searches with and HMMER [94], a HMM-based domain search algorithm. Next, putative amino acid sequences are used in a HMMER search in Pfam, and a BLAST search in UniProt. Other structures can also be predicted, such as rRNA [147], signal peptides [196] and transmembrane helices [143]. After all searches are finished, all hits are combined and used to attribute functional annotations like GO-terms to each sequence. This information is then finally stored in a SQL database.

#### 2.6.3.15 CD-HIT and CD-HIT-EST

Sets of contigs may have a certain degree of redundancy, for example caused by genetic variation, or by the incomplete assembly of UTR sequences. CD-HIT-EST

[156] is a program that clusters nucleotide sequences based on their similarity, and merges them into best representatives. CD-HIT [156, 97] does the same for amino acid sequences.

### 2.6.4 Multivariate analysis

In this work, datasets often contain multiple variables. These may be different observed behaviours, or the expression of different genes. Multivariate statistics provide the tools to reveal relationships between variables based on the observation of multiple samples, and between samples based on the observation of multiple variables.

#### 2.6.4.1 Hierarchical clustering

In large datasets, a given observation may have a lot in common with another observation, while being totally different from yet another one. This makes it interesting to cluster observations based on multiple variables. If this is done in an unsupervised way (the computer is not informed about the identity of the observations) and clusters correspond to relevant subpopulations, such as samples that share a treatment, there is likely strong signal in the data.

Hierarchical clustering is a group of hard clustering methods (observations belong to a single cluster) that are based on multi-variate distances between observations [284]. The process always starts with the clustering of the two observations that are closest to each other. This generates a first cluster. For this cluster, the distance to all other points can be re-calculated according to several criteria, such as the average distance of all elements of the cluster, or the distance with the nearest or even the furthest element that is part of the cluster. In a next round, the two closest points (which now can be single observations or clusters of observations) are clustered again and distances are again re-calculated. This process is iterated until all observations are clustered. A hierarchical clustering is typically visualized with a dendrogram.

In this work, I used the R functions *dist* and *hclust* for hierarchical clustering.

#### 2.6.4.2 Principal component analysis

Principal component analysis (PCA) is a geometric transformation of a set of observed variables into a set of orthogonal components, based on the covariance structure in the data. Mathematically, PCA is a singular value decomposition of the (scaled)

data matrix, which is equivalent to an eigenvalue-eigenvector decomposition of its correlation (scaled) or covariance (unscaled) matrix [192, 124].

PCA defines a new set of orthogonal axes within the space defined by all variables, along which the variance is largest. These axes are defined as linear combinations of all variables. It is a lossless procedure resulting in a rotation from which the original data can always be re-calculated. Hence, it is not a dimensionality reduction method. However, it can be used to estimate the real dimensionality of a set of variables. In the case of high degrees of redundancy (multicolinearity), only a few principal components will capture the large bulk of the variance within a dataset. A principal component, even if it explains a good deal of variance, simply accounts for a maximal amount of variance without the obligation of being interpretable.

In this work, I used the R function *prcomp* for PCA.

#### 2.6.4.3 Exploratory factor analysis

In a large set of variables, correlations are often caused by latent factors. These are variables that are not directly measured, but rather witnessed through the observed variables that they influence. These latent factors do not necessarily correspond to principal components, especially when they are themselves correlated to each other (non-orthogonal).

Exploratory factor analysis (EFA) also addresses the non-interpretability problem typical to PCA, by defining factors as linear combinations of all variables, maximizing correlations with those variables that are related to the latent factor, while minimizing correlations with unrelated variables (cross-loading) [133]. In many respects, EFA is similar to PCA, and in some special cases it will be virtually equivalent. However, while PCA is a descriptive method that aims to account for the *total variance* in the data, EFA is a structural model of the data, only accounting for the *common variance* shared between variables. A pre-defined number of factors are calculated while the error variance is captured in error terms much like in linear models. This makes EFA a real dimensionality reduction, from which the original data cannot be re-calculated. By omitting the constraint of orthogonality, factors can also be allowed to be correlated with each other. While PCA has one unique solution for each dataset, an infinite number of factor rotations are possible, from which the user chooses the one with the highest interpretability.

In this work, I used the R function *fa* from the *psych* package [206].

### 2.6.5 Predictive modelling

In this work, I explored two broadly used predictive modelling methods. In both cases, the model was "trained" on a set of observations where the outcome variable (the phase state of the animals) was known, in order to later serve as a tool to predict the same outcome variable from new observations.

#### 2.6.5.1 Binary logistic regression

Binary logistic regression is a parametric method in which the probability of membership to either of two classes is expressed as a function of a finite number of predictor variables [60].

It is a generalized linear model with a logit link function. This means that a linear regression is fitted using as an outcome variable the logit of the probability, which is defined as follows:

$$logit(p) = ln(\frac{p}{1-p}) \tag{2.1}$$

For a binary outcome this translates to the natural logarithm of the ratio of probabilities for both class memberships.

Now, given an $n \times m$ matrix X with $n$ "training" observations, and $m$ predictor variables, the fitted outcome variable for the i$^{th}$ observation is defined as:

$$logit(p_i) = X_i\beta + \varepsilon_i = \beta_0 + \beta_1 x_{1,i} + ... + \beta_m x_{m,i} + \varepsilon_i \tag{2.2}$$

With $X_i$ the vector of length $m$ that contains the values for all predictor variables for observation i, $\beta$ the vector of fitted parameter estimates (including an intercept), and $\varepsilon$ the observation-specific error term.

Now, given the same set of parameter estimates $\beta$, the logit of class membership can be calculated for a new observation:

$$ln(\frac{p}{1-p}) = X\beta \tag{2.3}$$

From this follows the probability:

$$p = \frac{e^{X\beta}}{1 + e^{X\beta}} \tag{2.4}$$

The $P_{greg}$ score that we know from the phase transition literature is such a probability. In this case, X contains a selection of behavioural variables. Often, more variables are at hand than are necessary to fit a model. In this case, different alternative models can be compared to each other and an optimal model is selected.

#### 2.6.5.2 Random forests

The fields of multivariate statistics, classification and machine learning are quickly expanding, and the possibilities for classification are too numerous to expand on each one of them. As an illustration of the power of machine learning, a *random forests* (RF) will be used to construct a classifier for locust phase. Out of the overabundance of machine learning techniques, I selected RF because of its simplicity, and its strong performance with interacting variables. RF is a classification tree-based method [34] that gained enormous momentum in machine learning and is now one of the most widely used techniques in artificial intelligence [91]. It is a non-parametric method that constructs a "forest" of multiple classification trees, each of which is constructed based on a subset of the variables and a subset of the data. Since no distances have to be calculated in high-dimensional space, it has a low computational cost. Also, this method is rather resistant against over-fitting because only a subset of the data is used for each individual tree. Moreover, during training, each tree can be directly cross-validated using the observations that were left out, providing the *out-of-the bag* (OOB) estimates of accuracy. RF can be used for mere classification as well as for regression [34, 159, 171, 91, 145, 144]. In the case of behavioural assays, this may provide an interpolation framework of crowded locusts between long term phases, much like in logistic regression.

# Chapter 3

# A new *Schistocerca gregaria* reference transcriptome

> [...] the uncontrolled causes which may influence the result are always strictly innumerable.
>
> Ronald A. Fisher

# 3.1 Introduction

## 3.1.1 Why sequence a reference transcriptome?

At the start of this project the only *Schistocerca gregaria* sequence data present was the EST database [13] that was generated in our lab. This transcriptome consisted of 12,709 unique sequences that were assembled from a total of 34,672 EST clones. Although this database has been successfully used as the basis for developing microarray studies [14], we decided that constructing an improved reference transcriptome would be a prerequisite for the ambitious transcriptomics experiments that we planned. Furthermore, with the advent of next generation sequencing technologies such as pyrosequencing [173] and sequencing by synthesis [40], a more substantial sequencing project aiming at covering the entire transcriptome with a good coverage and sequencing depth, became affordable.

## 3.1.2 Why not sequence a reference genome?

At the time of writing, the *Schistocerca gregaria* genome sequence has not been published yet, but strong interest was expressed by many members of our research community recently during ICE 2016, the International congress of Entomology in Orlando (Florida, US). Hopefully, with the ever increasing efficiency of genome sequencing, this will be achieved in a matter of a couple of years at most.

For my project, a reference genome would have had many added benefits, but a genome project would also have been of a totally different scale, especially given the large genome size (8.55 gigabase [39]) and the presumed high content of repetitive elements [281, 39].

The publication of the *Locusta migratoria* reference genome, the first orthopteran to have its genome sequenced, gives us a first insight in structure and content of grasshopper genomes [281]. However, it is not ideal for gene finding because of its poor assembly and incomplete documentation.

## 3.1.3 Current orthopteran reference genomes and transcriptomes

Hemimetabolous insects have long been a under-represented group in databases of reference genomes and transcriptomes, and still today, out of 230 published[1] insect

[1] https://www.ncbi.nlm.nih.gov/genome/browse/

nuclear genomes, only 36 belong to hemimetabolous insects, and only 1 belongs to an orthopteran [281].

Apart from the *Locusta migratoria* nuclear genome and some mitochondrial genomes, all publicly available orthopteran sequences are currently in transcriptomes. 2016 has been a productive year in the publication of transcriptomes, and I know from personal communication that much more is in the pipeline. The currently published ensiferan transcriptomes comprise *Laupala kohalensis* [65], *Gryllus bimaculatus* [289], *Gryllus firmus* [275], *Gryllus rubens* [23], *Teleogryllus oceanicus* [16] and *Teleogryllus commodus* [137]. The currently published caeliferan transcriptomes are *Locusta migratoria* [166], *Schistocerca gregaria* [13], *Tetrix japonica* [200] and *Chorthippus biguttulus* [24].

Depending on the kind of transcriptome (tissue specific, whole body, life stages, ...) and the techniques used (sequencer, assembly, filtering,...), these transcriptomes contain between 19,874 and 74,657 contigs.

Other orthopteran transcriptomes are reported in a number of studies, but their sequences are not publicly available [4, 291, 132, 201].

### 3.1.4 De novo transcriptome assembly

Assembly is the construction of consensus sequences (contigs) out of reads that are typically only fragments of the original sequences. In the case of genome sequences, reads are much smaller than a chromosome, and assembled contigs ideally correspond to entire chromosomes. In the case of transcriptome sequencing, the ultimate goal is to assemble contigs that correspond to entire RNA molecules. Compared to genome assembly, transcriptome assembly has an important added layer of complexity caused by non-uniformity of sequencing depth. Many strategies and software packages are available for transcriptome assembly [181, 197, 50, 43, 290, 106, 228, 112, 51, 183], with graph-based algorithms that can broadly be subdivided in overlap-layout-consensus, and De Bruijn graph-based methods [157] (see supporting materials for more information).

### 3.1.5 Transcript and gene annotation

After a set of new sequences has been acquired, predictions about their possible functions can be made based on presumed orthology with better studied organisms. Since orthology is hard to infer in high throughput, most methods are based on identity scores or motifs, which is a typical task for Hidden Markov Models [37,

75, 245, 35, 174, 76, 94]. In order to avoid confusion with colloquial gene names[2], standardized vocabularies are in use, like eggNOG [198], KEGG [135] and GO [9].

High throughput annotation is challenging and in my opinion it deserves more critical attention than it usually gets[3]. Even bracketing out the difficulties in distinguishing between orthology, paralogy and identity by chance, every inference about function based on what is known from another organism, comes with a chance of making mistakes. Notably, the standardized vocabularies assume strong functional conservation, sometimes over millions of generations of independent evolution. While this may be the case for the majority of genes, the most interesting genes are often the ones that function most differently, even between closely related species. For example, only a few members of the genus *Schistocerca* are locusts [240], and therefore the gene products that are involved in phase polyphenism must in some way act differently, be it in function or in quantity, between swarming and non-swarming grasshopper species, regardless of their close relatedness. Without any differences in gene function between two species, they would not be different species in the first place.

## 3.2 Aims

**1.** To assemble a reference transcriptome for *Schistocerca gregaria*, in which transcripts of multiple tissues under multiple physiological states are represented.

**2.** To explore the possibility of combining reads from different origins, obtained with different sequencing technologies into one single assembly.

**3.** To explore the contents of the *Schistocerca gregaria* transcriptome with an emphasis on neuropeptide precursors.

## 3.3 Wetlab procedures

### 3.3.1 Overview of the samples

This sequencing project started in September 2011 while in parallel, similar efforts were undertaken in other groups. I expressed my interest in trying to combine all sequencing projects into one assembly, and data were kindly shared. Because of technical difficulties, multiple batches of Illumina® sequences were generated (see further) and also one sample was sequenced with 454 pyrosequencing.

[2] Or, to avoid confusion with Microsoft Excel changing them into dates and floating point numbers [292].

[3] It gets some more attention in the supporting materials.

As a result, the starting material consisted of multiple independently sequenced batches of Illumina® reads, one batch of 454 reads and our original Sanger-sequenced EST reads. The final aim was to assemble all this sequence information into one transcriptome as complete and correct as possible.

Six different batches of RNA were used in this project. Except for set 4, all samples were taken from animals raised in our own colony. Only set 5 contains samples collected and processed by myself.

**Set 1:** Central nervous system tissues of mixed sex and age solitarious and gregarious locusts from the KU Leuven colony. These were the samples used for the Sanger-sequenced reads as published in Badisco *et al.* (2011) [13].

**Set 2:** Brain from adult solitarious locusts of the KU Leuven colony. RNA was extracted by Bart Boerjan using the Illustra™ QuickPrep Micro mRNA Purification Kit (GE Healthcare). Libraries were prepared with standard Illumina® TruSeq® protocols and sequenced on illumina® HiSeq™ 2000 at the university of Heidelberg (Germany). These are the samples discussed by Falckenhayn *et al.* (2013) [89].

**Set 3:** Brain from adult gregarious locusts of the KU Leuven colony, from the same study by Falckenhayn *et al.* (2013). [89], prepared and sequenced as in set 2.

**Set 4:** Balanced mix of all tissues of mixed sex and age gregarious locusts, supplemented with heat shock exposed animals, animals fed on different host plants and toxins and animals exposed to high levels of bacteria and fungi. Animals from the Max Planck Institute for Chemical Ecology (ICE), Jena (Germany). Sequenced on one lane of Illumina® HiSeq™ 2000.

**Set 5:** Balanced mix of all tissues of mixed sex and age solitarious and gregarious locusts from the KU Leuven colony supplemented with protease inhibitor fed and starved animals. This set contains two independently sequenced lanes of Illumina® HiSeq™ 2000 (set 5a1 and set 5a2) and one batch of Roche GS FLX+ 454 sequences (set 5b). Multiple library preparation methods were used.

In the following sections, the focus will be with the preparation of set 5, which was the only batch that I personally followed starting from dissections, proceeding with RNA extractions, purifications, library preparation and finally sequencing.

### 3.3.2 Sample preparation for set 5

Animals of mixed sex and age were used. In total, the following 14 different samples were prepared:

**sample 1:** Whole first and second instars, mixed sex and phase
**sample 2:** L3, L4 and L5, mixed sex and phase: brain, SOG and TG
**sample 3:** L3, L4 and L5, mixed sex and phase: antenna, leg, cuticule, flight muscle
**sample 4:** L3, L4 and L5, mixed sex and phase: fat body and salivary glands
**sample 5:** L3, L4 and L5, mixed sex and phase: cleaned digestive tract
**sample 6:** Adult mixed age, sex and phase: brain, SOG and TG
**sample 7:** Adult mixed age, sex and phase: antenna, leg, cuticule and flight muscle
**sample 8:** Adult mixed age, sex and phase: fat body and salivary glands
**sample 9:** Adult mixed age, sex and phase: cleaned digestive tract
**sample 10:** Adult mixed age, sex and phase: gonads (incl. eggs)
**sample 11:** Adult mixed age and sex gregarious, starved for 24 and 48 h: brain, SOG and TG
**sample 12:** Adult mixed age and sex gregarious, starved for 24 and 48 h: cleaned digestive tract
**sample 13:** Adult mixed age and sex gregarious, protease inhibitor[4] fed: brain, SOG and TG
**sample 14:** Adult mixed age and sex gregarious, protease inhibitor fed: cleaned digestive tract

Samples were taken directly in MagNA lyser™ Green Beads vials (Roche Diagnostics) and were continuously kept in liquid nitrogen during sampling and in the -80 freezer in between sampling days. From each sample, RNA was extracted using the QIAGEN RNeasy Lipid Tissue Mini Kit (standard protocol with DNAse digestion step). Samples were eluted in 50 µl RNAse free water and concentrations were measured using NanoDrop™ 1000 (Thermo Scientific™). Next, a balanced mixture of these 14 samples was made with a calculated 5 µg RNA per original sample. This yielded a mix of 141 µl solution with a theoretical RNA concentration of 500 ng/µl. This mix will now be referred to as master sample 5. Master sample 5 was stored in 5 µl aliquots at -80°C.

Master sample 5 was analyzed using the Agilent 2100 Bioanalyzer. The concentration estimate (627 ng/µl) was in the same order of magnitude as the calculated RNA concentration based on NanoDrop™ 1000 (Thermo Scientific™) measurements[5].

[4]One group of locusts was fed with the combination of soybean bowman birk inhibitor (SBBI) and soybean trypsin inhibitor (SBTI), the other one with *Schistocerca gregaria* pacifastin like inhibitor 1 and 2 (SGPI-1 and SGPI-2), a trypsine and chymotrypsine inhibitor respectively. The sample is a mix of both conditions. These conditions are not directly relevant to the current project, but they are added because the physiological response to protease inhibitors is a research line in our group, and interesting new transcripts may be discovered.

[5]It should be noted that concentration estimates of insect RNA by Agilent 2100 Bioanalyzer are only approximate, since the calibration is based on mammalian RNA [286], see general methods.

**Table 3.1:** Overview of concentrations, absorbance and volume of each sample that was added to master sample 5

| Sample nr | conc $(ng/\mu l)$ | $A_{260/280}$ | volume used in mix $(\mu l)$ |
|---|---|---|---|
| 1 | 280.2 | 1.87 | 18 |
| 2 | 502.1 | 1.97 | 10 |
| 3 | 278.4 | 2.03 | 18 |
| 4 | 701.3 | 2.13 | 7 |
| 5 | 1,264.2 | 2.19 | 4 |
| 6 | 904.8 | 1.96 | 6 |
| 7 | 310.2 | 1.98 | 16 |
| 8 | 547.6 | 2.04 | 9 |
| 9 | 3,047.5 | 2.06 | 2 |
| 10 | 265.3 | 2.07 | 19 |
| 11 | 293.2 | 2.11 | 17 |
| 12 | 1,504.0 | 2.21 | 3 |
| 13 | 512.6 | 1.98 | 10 |
| 14 | 2,019.1 | 2.07 | 2 |

### 3.3.3 Sequencing in three attempts

In September 2011, a first unnormalized library was prepared from master sample 5 by the local staff of genomics core using the standard Illumina® TruSeq® protocol (see general methods). Sequencing on one lane of Illumina® HiSeq™ 2000 yielded 186.8 million paired-end (2 x 100 nt) reads, amounting to about 37 gigabase. This set of reads will henceforth be referred to as set 5a1.

After preliminary analysis of set 5a1, severe quality issues were observed (see results). Also, it was estimated that about 75 % of the pairs overlapped, meaning that the median insert size was smaller than 200 bp (see below for more details on the estimation of insert size distributions). Genomics core went back to their intermediate quality checks and indeed informed us that the library had a skewed insert size distribution of 150 - 600 nt with median smaller than 200. About half of the reads mapped back onto the *Schistocerca gregaria* mitochondrial genome, presumably caused by contaminations, either with mtDNA, or with 16S rRNA, which is encoded in the mitochondrial genome.

In January and February 2012, two new libraries were separately prepared starting from master sample 5. They were prepared according to the standard Illumina® TruSeq® protocol, but this time with a selection of larger fragments, and with the inclusion of a library normalization step (see below). Both these libraries were sequenced at 10% of an Illumina® HiSeq™ 2000 lane, but yielded no more than

30 and 80 megabase, where about 3 gigabase would be expected. These reads were discarded.

### 3.3.4 Third time lucky: set 5a2

Since initially I did not personally follow up the library preparations, it was difficult to determine if the observed problems were mainly due to contamination, or if something went wrong during library preparation. In May 2012, we turned to a new strategy: I would now closely follow all procedures in collaboration with the genomics core staff[6]. In order to spread the risk of library preparation failures, 4 separate libraries were prepared, each sequenced with a barcode in multiplex mode. Library preparations also included 4 GS FLX+ libraries (mainly because consumables were kindly sponsored by Roche Diagnostics Belgium). Since non-mRNA contaminations were a major problem in the first trials, purified mRNA was used, obtained from the same stock RNA mixed sample. This purification was executed in our own lab with the Illustra™ QuickPrep Micro mRNA Purification Kit (GE Healthcare, see general methods), and yielded two samples of 70 ng/µl of presumably ribosome free RNA. This purified mRNA was used as input to all subsequent library preparations.

In total, eight libraries were prepared: 4 for GS FLX+ sequencing (set 5b), and 4 for Illumina® HiSeq™ 2000 sequencing (set 5a2). In what follows, the focus will be on the preparation for Illumina® sequencing. The following libraries were prepaired:

**set5a2c:** non-normalized cDNA library prepared with Clontech SMARTer™

**set5a2cn:** normalized cDNA library prepared with Clontech SMARTer™

**set5a2t:** non-normalized cDNA library prepared with Illumina® TruSeq®

**set5a2tn:** normalized cDNA library prepared with Illumina® TruSeq®

#### 3.3.4.1 The Clontech libraries

Quant-iT™ RiboGreen® RNA Assay confirmed the starting material had a concentration of 74.4 ng/µl. Two samples of 3.5 µl each (260 ng mRNA) were taken, along with a positive control (Mouse Liver Total RNA) and a negative control sample. cDNA was synthesized according to standard protocols using SMARTScribe™ reverse transcriptase. Briefly: RNA was incubated with 3' SMART CDS Primer II A for 3 min. at 72°C, followed by 2 min at 42°C. Here, RNA is fully denatured and the forward primer anneals. Next, a master mix containing DTT, RNAse inhibitor, dNTPs, SMARTer™ II

[6]Which is a service lab, so I was allowed to follow, but not to execute the procedures myself.

A Oligonucleotide and SMARTScribe™ reverse transcriptase was added. Next, the tubes were incubated at 42°C for 90 min. which is the time needed to generate a sufficient proportion of full-length cDNAs. The reverse transcription reaction was terminated by further heating to 70°C for 10 min. Now each tube contained 10 µl first strand (hybrid RNA-DNA) product solution. This was diluted with 190 µl TE buffer, and a subsample of 4 µl was subjected to PCR amplification, using specific primers that hybridize to the anchor sequence and the poly-A tail. The standard Advantage 2 protocol was followed, but with 6 min. elongation instead of 3 (as suggested in the SMARTer™ manual for applications requiring longer cDNA molecules). A series of 4 µl aliquots were PCR amplified for 15, 18, 21, 24 and 27 cycles, with 6 min. elongation instead of 3 (as suggested in the manual for applications requiring longer cDNA molecules). Both Advantage 2 and Advantage 2 SA buffer were tested. All amplified libraries were quality checked with Agilent 2100 Bioanalyzer, and finally it was decided that 17 cycles[7] with the normal Advantage 2 buffer would be the best strategy for library preparation. As expected, positive controls gave signal comparable to the locust libraries and negative controls did not yield any observable cDNA.

Since the Clontech SMARTer™ protocol is optimized to yield longer fragments, but sequencing libraries have fragment length restrictions, the cDNA products were sheared. First, libraries were diluted with TE buffer up to a volume of 100 µl and sonicated with the Diagenode Bioruptor® Pico sonication device in 7 cycles (30” on, 90” off). Next they were purified with the QIAGEN Min Elute kit and eluted in 50 µl QIAGEN EB buffer.(10 mM Tris-Cl, pH 8.5).

#### 3.3.4.2 The TruSeq® libraries

Since the starting material consisted of purified mRNA, the purification step was omitted. RiboGreen confirmed the starting material had a concentration of 70.4 ng/µl. Two 5.7 µl samples (400 ng mRNA) were topped up to a volume of 20 µl EPF (Elute, prime, fragment mix by Illumina®) buffer. Next, the standard TruSeq® protocol was executed. Briefly: RNA was chemically sheared by heating the samples to 94°C for 2 min. in EPF buffer. Next, samples were transferred to a Hard-SHell 96-well PCR (0.3 ml) plate (HSP-9601) and 8 µl of FSM (first strand master mix, containing SuperScript™ II, nucleotides and random hexamers) was added. Next, the plate was sealed and placed in a pre-heated thermal cycler for the $1^{st}$ Strand program (pre-heat lid at 100°C, 10 min. at 25°C, 50 min. at 42°C, 15 min. at 70°C and hold at 4°C). Immediately after the first strand synthesis, 25 µl of Second Strand Master Mix was added to each well, followed by 60 min. incubation at 16°C. Finally, cDNA was purified with Agencourt® AMPure® XP magnetic beads. First 90 µl beads solution was added to each well and

[7] The protocol prescribes to substract one cycle from the optimal number of cycles, which is defined as the number of cycles where the yield of PCR product stops increasing and before smears appear in the high-molecular-weight region, indicating "overcycling".

placing the plate on a magnetic stand for 5 min., allowing all beads to bind to the side of the wells. Next, the supernatant was discarded and the wells were washed twice for 30 seconds with 200 µl 80% EtOH with the plate still on the magnetic stand. Next, the plate was allowed to dry on the magnetic stand at room temperature for 15 min. Next, the plate was taken from the magnetic stand and 52.5 µl resuspension buffer was added and allowed to incubate for 2 min. at room temperature. Finally, the plate was placed on the magnetic stands for 5 min. and 50 µl supernatant containing purified cDNA was recovered.

#### 3.3.4.3 Further library preparation with TruSeq®

The cDNA prepared as described in the two previous sections was not ready for sequencing yet. As described earlier, Illumina® sequencing depends on a the ligation of oligos onto the cDNA fragments. These procedures start with end-repair, where possible overhangs resulting from fragmentation are converted into blunt ends, either by digesting 3' ends, or by filling 5' overhangs. Next, a single deoxyadenosine monophosphate ("A" nucleotide) is added to the 3'-end, to which complementary multiple indexing adapters are hybridized and ligated.

Since Clontech SMARTer™ is not a full Illumina® library preparation protocol, the last part of the TruSeq® protocol was applied to all four cDNA samples (2 Clontech and 2 TruSeq®). Briefly:

**Generation of blunt ends and ligation of 3' "A" overhang**
50 µl cDNA was pipetted (4 samples) into an IMP (Insert Modification Plate) together with 10 µl resuspension buffer and 40 µl end repair mix and incubated for 30 min. at 30°C. Next, 160 µl Agencourt® AMPure® XP magnetic beads solution was added and after 15 min. incubation at room temperature, the IMP was placed on the magnetic stand and liquid was allowed to clear (5 min.). Supernatant was removed and the wells were washed 3 times with 80% etOH. After the plate was allowed to dry for 15 min., 17.5 µl resuspension buffer was added and incubated at room temperature for 2 min., followed by 5 min. on the magnetic stand. Finally 15 µl of supernatant was transferred to a new 0.3 ml HSP plate labelled with ALP (Adapter Ligation Plate) barcodes. In a next step, 12.5 µl freshly thawed A-Tailing Mix was added and the ALP plate was placed in a pre-heated thermal cycler for ATAIL70 program (pre-heat lid at 100°C, 30 min. at 37°C, 5 min. at 70°C and hold at 4°C).

**Adapter ligation**
2.5 µl Ligation Mix and 2.5 µl freshly thawed RNA adapter index (with adapters that have a single "T" nucleotide overhang) was added (one unique index per well), and incubated at 30°C for 10 min. The ligation reaction was stopped by adding 5 µl Stop Ligation buffer, followed by two steps of Agencourt® AMPure® XP magnetic beads

purification (42 µl and 50 µl beads solution). Finally, 20 µl supernatant was transferred to a new 0.3 ml PCR plate.

**Enrichment of correctly ligated fragments**
5µl PCR primer cocktail and 25 µl PCR Master Mix was added to each well, and the plate was placed in a pre-heated thermal cycler for the PCR program (pre-heat lid at 100°C, 30 sec. at 98°C, 15 cycles of 10 sec. at 98°C, 30 sec at 60°C and 30 sec. at 72°C, 5 min. at 72°C and hold at 10°C). PCR products were cleaned with Agencourt® AMPure® XP magnetic beads (50 µl beads solution), resuspended in 32.5 µl Resuspension Buffer and transferred to a new 0.3 ml PCR plate.

#### 3.3.4.4 *In vitro* Library normalization

*In vitro* library normalization is a procedure where a skewed distribution of the relative abundances of sequences in a library is shifted towards a more symmetrical distribution. We used the Illumina® DSN Normalization protocol, which is based on the Duplex-Specific thermostable nuclease (DSN) enzyme (Evrogen), purified from Kamchatka crab (*Paralithodes camtschaticus*) hepatopancreas. This endonuclease is practically inactive towards ssDNA, while having a high affinity for dsDNA, especially when both strands are perfectly matching. By denaturing the DNA and allowing it to re-anneal under kinetically limiting conditions, this enzyme can be used to selectively digest those molecules that most often find their complement. In this manner, the library is depleted in its most frequent sequences, resulting in the relative enrichement of fragments originating from uncommon transcripts. This procedure was applied to one sample from each library preparation. Briefly:

**Digestion with DSN**
In a nuclease-free 200 µl PCR tube, 13.5 µl sample library was combined with 4.5 µl Hybridization buffer (0.2 M HEPES, 2M NaCl) and incubated in a thermal cycler (2 min. at 98°C and 5 hours at 68°C). Next, 20 µl pre-heated (68°C) 2X DSN buffer (100 mM Tris-HCl, pH 8.0; 10 mM MgCl2; 2 mM DTT) was added twice, with incubations in the thermal cycler for 10 min. at 68°C. Next, 1 µl[8] DSN enzyme mix (prepared according to the manufacturer protocol) was added (in two times 0.5 µl) and incubated at 68°C for 25 min. Finally, the reaction was quenched by adding 40 µl 2X DSN stop solution (10 mM EDTA) and placing the tubes on ice.

**Purification of DSN treated templates**
80 µl of each sample was transferred to a 96-well V-bottom plate, and 128 µl Agencourt® AMPure® XP magnetic beads solution was added. Next, the plate was placed on a magnetic stand for 8 min. at room temperature. After removing the supernatant, the plate was washed twice with 180 µl 80% EtOH. Next, the plate was

[8]The protocol prescribes 4 µl, but we did not have enough enzyme for all reactions.

centrifuged at 1000 rpm for 10 sec. and placed back on the magnetic stand for 30 sec. After removing the supernatans, the plate was heated to 37°C until the pelets were dry. Finally, the pellet was resuspended in 30 µl QIAGEN EB buffer (10 mM Tris-Cl, pH 8.5) and incubated for 1 min. on the magnetic stand before samples (30 µl) were transferred to nuclease-free 200 µl PCR tubes.

**Enrichment PCR**
A PCR reaction was run with Phusion™ High-Fidelity DNA Polymerase (Finnzymes, New England BioLabsn Inc.) and PCR primers that are complementary to the ends of the ligated adapters, using the following cycling conditions: 30 sec. at 98°C, 12 cycles of 10 sec. at 98°C, 30 sec. at 65°C and 30 sec. at 72°C, 5 min. at 72°C and hold at 10°C.

**Purification of PCR products**
50 µl sample was transferred to a new 96-well V-bottom plate, and the purification procedures as described above were repeated, but with 80 µl Agencourt® AMPure® XP magnetic beads solution, and resuspension in 20 µl QIAGEN EB with 2 min. incubation. After leaving the plate on the magnetic stand for 1 min., all samples (20 µl) were transferred to new nuclease-free 200 µl PCR tubes and stored at -20°C.

The four libraries, all with their own multiplexing indices (MID9 and MID10 for the Clontech libraries and MID11 and MID12 for the TruSeq® libraries), were loaded on the Agilent 2100 Bioanalyzer DNA1000 chip and their concentration was measured with the Qubit fluorometer. Next, libraries MID9 and MID11 were selected based on their higher concentration, and normalized according to the Illumina® DSN protocol, followed by analysis on a DNA1000 chip.

**Table 3.2:** Preparation of 4 libraries. Fragment lengths were estimated based on electrophenograms. Concentrations in ng/µl were measured using Qubit fluorometer. Concentrations in mol/µl were calculated based on estimated fragment length and concentration.

| Set | Index | estim. frag. length | Concentration | | |
|---|---|---|---|---|---|
| | | | pre norm. ng/µl | post norm. ng/µl | conc. nM |
| 5a2tn | MID9 | 340 | 85.4 | 11.2 | 51 |
| 5a2t | MID10 | 298 | 55.5 | | 287 |
| 5a2cn | MID11 | 387 | 60.0 | 12.9 | 51 |
| 5a2c | MID12 | 400 | 52.0 | | 200 |

#### 3.3.4.5 Loading samples onto the sequencer

The optimal concentration for a library to be loaded on Illumina® HiSeq™ 2000 is 2 nM. This molar concentration was estimated based on the concentration in ng/µl as measured with the Qubit™ fluorometer (Thermo Fisher Scientific), and the estimated average fragment length. Next, the concentration was calculated based on the molecular weight of DNA (649 g/mol/bp). Based on this estimated molar concentration, libraries were further diluted in QIAGEN EB buffer (10 mM Tris-Cl, pH 8.5) supplemented with 0.1% Tween to the final desired concentration of 2 nM in 50 µl and loaded onto the Illumina® HiSeq™ 2000 according to standard protocols.

## 3.4 Bioinformatics

The assembly of the transcriptome has been a process where methods were regularly updated based on what I learned about the data and the techniques I was using. It is therefore hard to break down this chapter into mere methods and results. Most intermediate results feed back directly into downstream operations. Hence, the following sections, although being primarily methodological, inevitably contain a lot of intermediate results.

Initially I planned to include Sanger-sequenced, Illumina®-sequenced and pyrosequenced (454) reads in a single assembly. However, transcriptome assembly software has turned out to be rather unaccommodating for reads of different origin and length: I did not find software that could incorporate a stream of 454 reads in an assembly otherwise dominated by Illumina® reads. Also the error structure of 454 reads (homopolymer errors), and the file type they are stored in, did not allow them to be processed in the same pipeline as Illumina®-sequenced reads. Actually, the only program specifically designed for the assembly of 454 reads is Newbler, which takes into account raw base calls in order to better accommodate for potential homopolymer errors [43]. Unfortunately, Newbler is a rather expensive proprietary software package.

Luckily, the input of 8 Illumina® libraries with slightly different histories and FASTQ formats was enough of a challenge on its own. The pipeline outlined below focuses uniquely on Illumina® reads with their specific length distribution, overlap profiles and sequencing adapters. The Sanger-sequenced reads were considered "trusted", and only come into play at the very end. Both the input reads and all intermediates in the pipeline were carefully quality-checked with FastQC. Special attention was paid to over-represented sequences (which may be indicative of contamination) and base call quality (Phred scores).

All software that was used, is described in the general methods chapter. All commands and FastQC reports can be found in the supporting materials.

## 3.4.1 The assembly pipeline

### 3.4.1.1 Acquiring and pre-processing the read data

The four libraries of set 5a2 (Clontech SMARTer™, Clontech SMARTer™ normalized, Illumina® TruSeq® and Illumina® TruSeq® normalized) were sequenced on one lane of 2 x 100 bp on Illumina® HiSeq™ 2000, which took about two weeks. After sequencing, reads were *demultiplexed* based on their index barcodes. During this process, which was executed at genomics core Leuven, all reads were sorted into separate FASTQ files based on index sequences (the 6 nucleotide barcodes MID9 - MID12), allowing a maximum of one error per index. A first observation was that, although libraries had been loaded onto the sequencer in equimolar concentrations, the *in vitro* normalized libraries produced less than one percent of the total number of reads. An interesting second observation was a much higher mismatch proportion in barcodes of *in vitro* normalized libraries. In both unnormalized libraries, more than 90% of the barcodes were without error, while this was less than 30% in either of the normalized libraries. See supporting materials for details.

After acquiring reads from all involved parties, the entire collection of read data was gathered on the computer in Leipzig in zipped format (.fq.gz). Set 4, which was the only Phred+64 formatted set, was converted to Phred+33 using the program seqret[9]. The quality of the initial libraries was variable. In the overview of base-calling quality per tile over the length of the read, set 2 and set 3 showed tile-specific error patterns especially in the second mate. Set 4 was generally of lower quality, showing a severe decay towards the end of the read. All members of set 5 were of good quality, except for set 5a1, where the first mate showed severe cycle specific error patterns spread over multiple tiles (fig. 3.1).

[9] http://emboss.sourceforge.net/apps/release/6.6/emboss/apps/seqret.html

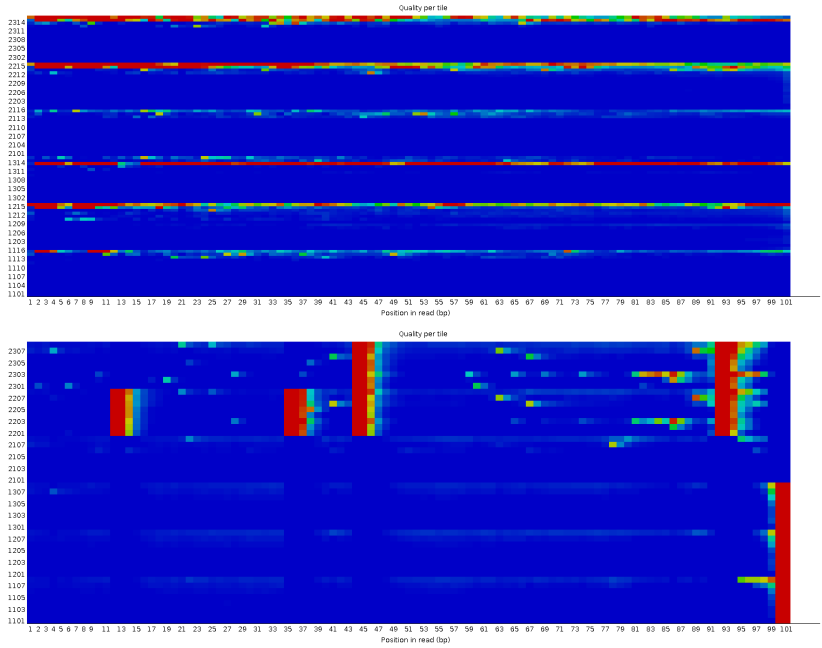



**Figure 3.1:** Examples of tile and base specific error patterns as outputted by FastQC. Unfortunately, FastQC does not provide an annotation of the colour scale, but any colour "hotter" than blue points to severe drops in Phred score. Top: tile-specific error patterns in the second mate of set 3. Bottom: position-specific disturbances over multiple tiles, as seen in the first mate of set 5a1.

#### 3.4.1.2 Preliminary contaminant analysis

Before any quality trimming, databases were screened for obvious contaminations. The over-represented sequences detection function of FastQC revealed varying amounts of contamination with ribosomal RNA, potentially mitochondrial DNA and sequencing adapters. Especially the Clontech SMARTer™ adapters were present in high proportions (in up to 7% of the reads of the Clontech libraries, but also in trace amounts in set 5a2t and 5a2tn). Set 5a1 showed over-representation of 16S ribosomal sequences, and set 5a2t showed moderate contaminations with cytosolic ribosomes.

To further quantify the content of the input libraries, a sample of 1000 reads (both pairs) was mapped onto two potential sources of biological contamination: the *Schistocerca gregaria* mitochondrial genome, and to the *Gregarina niphandrodes* reference genome and predicted transcriptome, using STAR. None of the input libraries had more than 10% of their reads mapping on the mitochondrial genome, except set 5a1, where

about 45% of the reads found a hit. The vast majority of these hits were in the 16S ribosomal RNA gene, suggesting the contamination was ribosomal RNA rather than mitochondrial DNA (see supporting materials). Mappings to *Gregarina* were rare or absent in all libraries.

#### 3.4.1.3 Estimating the distribution of insert sizes

All input libraries contained significant proportions of overlapping reads[10] (where the insert size was shorter than twice the read length), also in "outie" orientation (where the insert size was shorter than the read length). In order to optimally merge these overlapping reads with FLASH, input parameters had to be individually adjusted for each library, based on a preliminary estimate of insert size. This was mainly important to optimize the maximal overlap parameter (-M) for FLASH, which corresponds to the 90% percentile of overlap length in the "innie" orientation.

For the libraries that we prepared ourselves, we had an idea about the distribution of insert sizes based on electropherograms (see supporting materials and table 3.2), but this was not the case for libraries that were prepared elsewhere. I used FLASH itself with standard parameters (except for allowing the "outie" orientation) to get a preliminary idea of the distribution of insert sizes for each library. Next, the histogram output of all merged fragment sizes was used to make an educated guess about insert sizes. Since in the histogram output, only merged pairs can be observed, part of the distributions were hidden and had to be estimated only from their left tail. Log-normal distributions were fitted based on the estimated overlap proportion and the visible left tail of the distribution of fragment sizes (see supporting materials for R scripts and supplementary figures). Based on the estimated geometric mean and geometric standard deviation, estimates were made of the mean and median insert sizes of the libraries. The larger the proportions of overlap, the more information was available to estimate these parameters, and hence the uncertainty was largest for the libraries with the smallest overlap.

Inferred proportions of overlapping reads ranged from 21 to 73% and the proportion of "outie" orientation overlaps was low (1 - 3%) in all datasets, but considerable enough to be of concern for consecutive steps in the pipeline. Estimated overlap proportions and corresponding estimates of mean fragement size and 10% percentiles are shown in table 3.3.

The actual DNA fragments observed in e.g. electropherograms are about 120 bp longer than the inserts, because of the presence of Illumina® sequencing adapters. If this is taken into account, the original insert sizes of sets 2 and 3 were estimated to have been between 270 and 280 nulcleotides, while the publication [89] mentions 250. A

[10] Since overlaps of less than 10 bp cannot reliably be estimated, overlap was defined as "at least 10 bp overlap".

comparison between the fits and the observed insert sizes (Agilent 2100 Bioanalyzer) from our own library preparations (set 5a2) (based on electropherograms, see table 3.5), shows that the estimates of the mean insert sizes are not far off the estimates based on empirical fragment sizes provided by the Agilent 2100 expert software (table 3.2).

#### 3.4.1.4 Trimming-merging-trimming

In this pipeline, two steps of quality trimming were applied, both with the program Trimmomatic. Before any trimming was done, proportions of overlap were estimated as described above. Next, with the raw data as input, only mild quality trimming was applied, keeping poor base calls on board. These could potentially be recovered after merging of overlapping reads, and if this was not the case, they would be removed in the second trimming step. In this first step it was mainly important to trim off all adapter and primer sequences from any source and in any orientation, because once FLASH has merged reads, some of the critical pairing information is lost and Trimmomatic would no longer work in palindrome mode.

After adapter clipping, and mild quality trimming, the surviving pair ends were merged using FLASH. Unlike the first time, where FLASH was used just for exploratory purposes, this time the merged reads had to serve as input for downstream operations. Therefore, the family-wise error rate had to be controlled, and I decided not to use the aggressive "factory settings", but rather to apply more stringent criteria for read merging. Assuming random sequences, the likelihood for two sequences to have a perfect overlap of n nucleotides is $1/4^n$. Given the estimates of overlap proportions, posterior odds can be calculated for any overlap as $prior\ odds\ X\ likelihood\ ratio$. Unfortunately it is difficult to set a fixed family wise error using FLASH, because it does not make use of a probabilistic model, and hence there is no one likelihood function that can be optimized over the range of possible overlaps[11]. In analogy to calculating a Bayesian posterior probability, a more stringent cut-off was used for datasets with a low prior probability of overlap (e.g. set 4), than in data where almost all reads overlap. I therefore used a minimal trusted overlap of 12 nucleotides for set 2 and set 3, 14 nucleotides for set 4 and 13 nucleotides for all sets 5. These values were a consensus between sensitivity and accuracy given the prior based on previously estimated distributions (see supporting materials for R scripts).

In FLASH, a maximal mismatch proportion of 7% was allowed, which for a minimal trusted overlap of 12 allowed for one error starting from 15 nucleotides onwards.

[11] The tool "leeHom" does exactly this, and uses maximal posterior probabilities to co-estimate overlap proportions and the presence of adapters. I think it is probably a good (perhaps even better) alternative to the trimming-merging-trimming strategy. Unfortunately, I discovered this tool to late.

Finally, the "outie" orientation was allowed once again, albeit this time only to check if any were still present (they should have been gone after adapter clipping).

After merging all well supported overlapping pairs, a second round of trimming was executed. Now "outies" were no longer expected, and most overlapping reads should have been merged, palindrome mode was no longer of any use. Since the quantity of input data was large to start with, quality trimming parameters could be set to stringent values. A sliding window over 3 nucleotides was used with a cut-off for an average Phred score of 30.

The aim of the trimming-merging-trimming approach is that merged pairs with a mismatch can pass quality filtering if the consensus nucleotide was derived from a clear sequencing error (so the "winner" had a high Phred score), while regions with low consensus or low quality will now be filtered out and discarded. Although in theory adapters should have been removed earlier, adapter clipping is still active just in case merged pairs contain previously unnoticed adapters.

Finally, merged reads were selected at a minimal length of at least 100 nucleotides after trimming (otherwise no new information was incorporated) and single ends at least 75 nucleotides.

#### 3.4.1.5 Redundancy and *In silico* read normalization

A first way to cope with the problem of redundancy is to *in vitro* normalize cDNA prior to sequencing. Some of our own cDNA libraries were normalized in this way (set 5a2cn and set 5a2tn), but the yield was disappointing. As a result, the large part of input reads were not normalized, and a good deal of redundancy was to be expected.

Prior to any operation on the data, sequence redundancy was studied with estimates of the percentage of sequences remaining after de-duplication, using FastQC. Secondly, for each input library (after trimming-merging-trimming), a hash table of all 25-mers was constructed with the program Jellyfish, and k-mer count distributions were compared (fig. 3.2). Uniqueness estimates ranged from 20 to 90%, with the smaller and normalized libraries showing higher uniqueness. Table 3.3 summarizes these figures for the left and right input read of the untreated data, and for the merged reads after quality trimming (see further). Note how the "left" read of set 5a1 has a remarkably higher uniqueness than the "right". This may be explained by the poor quality of the former (fig. 3.1).

For the assembly based on all Illumina® data (the *big assembly*, see further), all input reads were *in silico* normalized with the insilico_read_normalization.pl script with a k-mer size of 25 and a maximal median k-mer coverage of 50. However, in the assembly based only on set 5, (*assembly 5*, see further) a k-mer size of 30 was used, and a median k-mer occurence cut-off of 100. As expected, the largest libraries generally showed

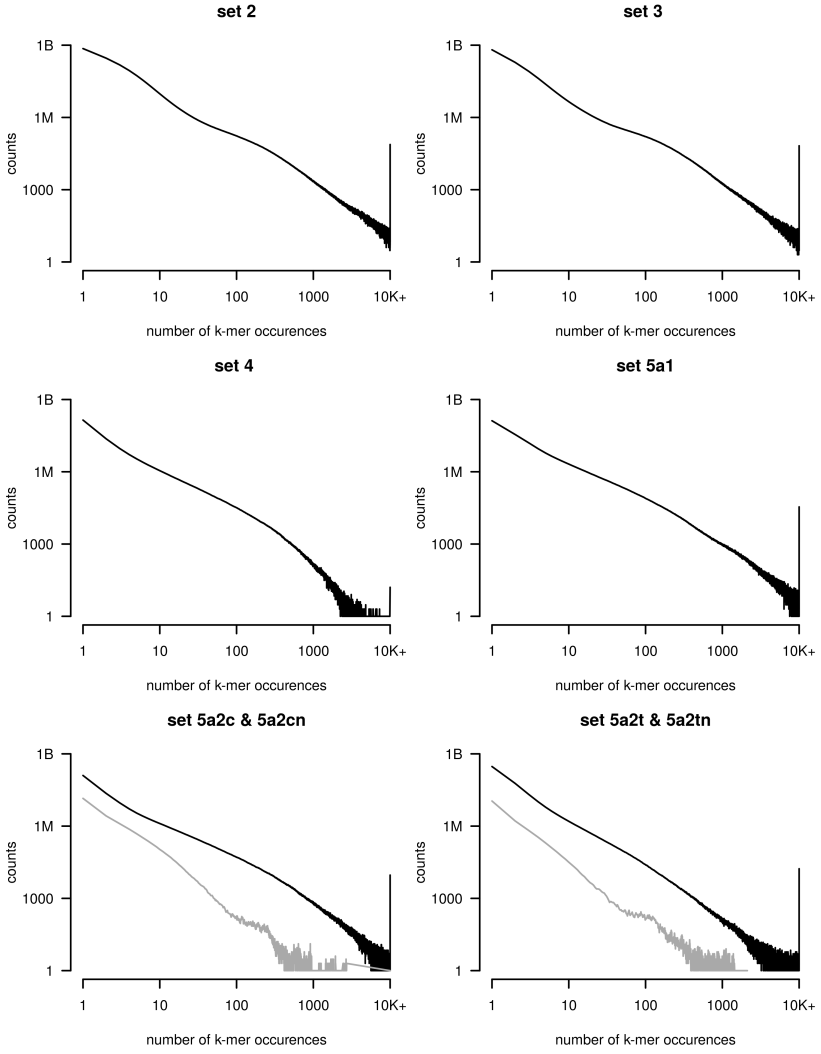



**Figure 3.2:** K-mer occurences for all libraries. The two largest sets, set 2 and set 3, both have more than half a billion unique k-mers, and more than 50,000 k-mers that occur more than 10,000 times. Note how the normalized libraries (grey colour) barely have k-mers that occur more than 1000 times.

the highest degree of redundancy. The normalized libraries had little redundancy, but given their small size it was difficult to quantify the effect of normalization. Next to merging all overlapping reads, between 0.08 and 1.26% of the reads were discarded as likely aberrant based on coverage profiles.

After *in silico* normalization, the paired reads (that were normalized as one concatenation) got reduced from 96 million to 27.6 reads. The most drastic reduction was seen in set 5a1 that got reduced by about 90%. The *in vitro* normalized libraries had the highest uniqueness, and were only reduced by around 10% to 20%. Since the amount of data in all single and merged reads was too large (412 million reads, a 112 gigabyte concatenation), it was impossible to normalize them all at once, so they were normalized one-by-one and afterwards the normalized sets were concatenated (71.7 million reads).

**Table 3.3:** Overview of all libraries used in this study, with their most important summary statistics. See main text for details. The size of set 1 was variable because some reads had already been assembled into longer contigs. The original size distribution of set 1 was between 200 and 1000 bp. Note the big differences between raw number of reads or total bases between different libraries. The base counts of single, merged and paired are post filtering. The percentage of bases left after filtering depends on two factors: read quality and the amount of overlap between pairs. For example: sets 2 and 3 were of good quality, but showed high levels of overlap, while set 4 had relatively little overlap, but was of poorer quality. The percentages of "outie" orientation overlaps is the proportion relative to the number of overlapping reads. The 10% percentile and mean estimated insert sizes are the values used for setting parameter values of FLASH. The uniqueness measures are values provided by FastQC for the filtered and trimmed sets. Finally, the percentage of reads that are left after normalization is relative to the number of reads after quality checks.

| Set | colony | sequenced in | lib. prep. | In vitro norm. | sequencing technology | read length | N reads | N bases (gb) | single + merged (gb) | paired (gb) | % bases post filter | % overlap | % outies | 10% %ile | estim. mean ins. size | % uniq. L | % uniq. R | % uniq. single + merged | % reads left post norm. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | KUL | Illinois | EST | yes | Sanger | 200-5,000 | 13 K | 0.01 | 0.01 | 0 | 100 | na | na | na | na | na | na | na | na |
| 2 | KUL | Heidelberg | TruSeq | no | Ill. HiSeq | 2 x 101 | 256 M | 51.7 | 18.35 | 6.89 | 49 | 71 | 2 | 83 | 166 | 38 | 34 | 41 | 26 |
| 3 | KUL | Heidelberg | TruSeq | no | Ill. HiSeq | 2 x 101 | 243 M | 49.1 | 14.4 | 4.58 | 39 | 73 | 1.7 | 83 | 163 | 34 | 32 | 40 | 22 |
| 4 | ICE | Jena | TruSeq | no | Ill. HiSeq | 2 x 100 | 41 M | 8.2 | 1.17 | 0.94 | 26 | 21 | 4.2 | 48 | 218 | 59 | 59 | 91 | 32 |
| 5a1 | KUL | Leuven | TruSeq | no | Ill. HiSeq | 2 x 101 | 187 M | 37.8 | 11.32 | 3.46 | 39 | 73 | 3.8 | 87 | 161 | 46 | 21 | 20 | 10 |
| 5a2t | KUL | Leuven | TruSeq | no | Ill. HiSeq | 2 x 101 | 32 M | 6.5 | 2.22 | 1.81 | 62 | 50 | 6.4 | 83 | 198 | 37 | 38 | 39 | 19 |
| 5a2tn | KUL | Leuven | TruSeq | yes | Ill. HiSeq | 2 x 101 | 245 K | 0.05 | 0.015 | 0.02 | 70 | 39 | 2.8 | 77 | 222 | 82 | 87 | 84 | 72 |
| 5a2c | KUL | Leuven | SMARTer | no | Ill. HiSeq | 2 x 101 | 43 M | 8.7 | 2.79 | 2.06 | 56 | 35 | 2.2 | 78 | 229 | 41 | 42 | 63 | 23 |
| 5a2cn | KUL | Leuven | SMARTer | yes | Ill. HiSeq | 2 x 101 | 437 K | 0.09 | 0.024 | 0.04 | 71 | 33 | 2.7 | 75 | 236 | 88 | 91 | 91 | 80 |
| 5b | KUL | Leuven | 454TM | yes | Roche FLX+ | 2 x 500 | 323 K | 0.32 | na | na | na | na | na | na | na | na | na | na | na |

#### 3.4.1.6 Assembly with Trinity

After all input reads were cleaned and normalized, the execution of Trinity was just a one-liner. Trinity was always run with a maximum allowed memory usage of 200 GB and a maximum use of 30 processors. Many input parameters can be adjusted, and the parameter with the highest impact is probably –min_kmer_cov, which is a filter parameter setting the minimal k-mer coverage for reads to be retained in the hash table constructed by Jellyfish. The default value is 1, which allows all reads to enter the assembly process. Setting this parameter to a value of 2 drastically drops the number of input reads (e.g. set 2 and 3 both had more than half a billion unique k-mers.) and results in lower contig numbers in the final assembly, at the cost of potentially getting a more fragmented picture of rarer transcripts. Other parameters that I explored either modify how Chrysalis glues together similar pre-contigs, or influence the way Butterfly merges paths during graph traversing.

Trinity is able to incorporate longer well-trusted sequences in the assembly through the –long_reads flag. From personal communication with the trinity developers I know that these longer sequences are only helpful in a limited number of cases, but they may certainly be beneficial in the recovery of transcripts that are otherwise impossible to assemble. The long sequences are mostly treated like normal reads, but they have a special role during read cluster definition by Chrysalis and for the resolution of complex isoforms or repeats by Butterfly. I have experimented with the incorporation of a set of trusted sequences containing the following elements:

**1.** All original assembled ESTs by Badisco *et al.* (2011) [13] filtered for reads originating from the common apple *Malus domestica*, a contamination I discovered during re-annotation of this transcriptome. The contigs were filtered based on a BLAST search (standard parameters of blastall) in the apple genome [274], with a bit-score cut-off of 150 bits (see supporting materials). This yielded 42 candidate apple sequences out of which 34 indeed originated from apple rather than from locust, based on a BLAST search against the entire nucleotide collection on NCBI (see supporting materials).

**2.** 141 *Schistocerca gregaria* sequences as found in NCBI genbank version May $30^{th}$ 2016 with the following search details: *("Schistocerca gregaria"[Organism] OR Schistocerca gregaria[All Fields]) AND "Schistocerca gregaria"[porgn] AND biomol_mrna[PROP]*

**3.** Sequences of transcripts that have been subject to study in our lab and that were confirmed with Sanger sequencing.

Since along the way, the input libraries would turn out to each have their particular issues, different inputs were explored alongside different parameter settings for

Trinity. Most assemblies were based on all Illumina® reads, sometimes combined with the trusted sequences that also contained the Sanger-sequenced reads. In all cases where Trinity was run with the –min_kmer_cov flag set to 2, the input file with trusted sequences was modified to contain every sequence twice. Because of time and computing constraints (A single run of Trinity uses up to 200 gb RAM for multiple days.), not all parameter and input combinations could systematically be tested.

The resulting assemblies were compared in terms of transcript numbers, N50 statistics, back-mapping coverage and sequence content. In many cases, unrealistically large numbers of contigs were outputted, ranging from one to two million, depending on the exact parameters. For the assembly that would serve as a reference for the RNA-Seq experiment of chapter 6, it was decided to run Trinity only on set 5, with default parameters (except for –min_kmer_cov set to 2), and with the inclusion of the trusted sequences. This assembly, that yielded 175,066 contigs, will henceforth be referred to as *assembly 5* and will be discussed further in the results section. Finally, a similar assembly was made of set 2 only, mainly to contrast its structure and contents to assembly 5.

#### 3.4.1.7 Summary of the entire assembly pipeline

**1. Input data organization and standardization:** All datasets were given shorter and more uniform filenames, and FASTQ headers were reformatted to a common format.

**2. Estimation of insert size per dataset:** For every dataset, the distribution of overlap between pairs was estimated based on histogram output of FLASH. Since in this step only the distributions of insert sizes were important, output FASTQ files were discarded.

**3. Trimming-merging-trimming:** All datasets underwent mild quality trimming and adapter clipping with Trimmomatic, yielding pair ends and orphan single ends. Next, the pair ends went through read merging (FLASH) with dataset-specific parameters based on estimated overlap proportions, yielding merged reads and pairs. Next, all resulting datasets (orphans, merged and unmerged pairs) were subjected to a second round of more stringent quality trimming, yielding trimmed versions of the original orphans and merged libraries, as well as the original pairs and a new series of orphans originating from the pairs.

**4. *In silico* read normalization:** For assemblies containing all Illumina® reads, this procedure worked in two parallel circuits: pairs of all datasets were concatenated and normalized as one file, while orphans and merged reads were also concatenated and normalized together. This latter turned out to be such a large number of reads, that it had to be done in two consecutive rounds where in a first round all orphans and

merged reads were concatenated per dataset and normalized, and next the normalized data were all concatenated together and normalized a second time. The standard k-mer size of 25 was used and the –max_cov parameter was set to 50. For the assembly based on set 5 alone, pairs and singles/merges were also normalized separately, with k = 30 and –max_cov = 100.

**5. Trinity assembly:** a mixture of read pairs, orphans, merged reads and longer, "trusted" reads was shoehorned into a single Trinity assembly.

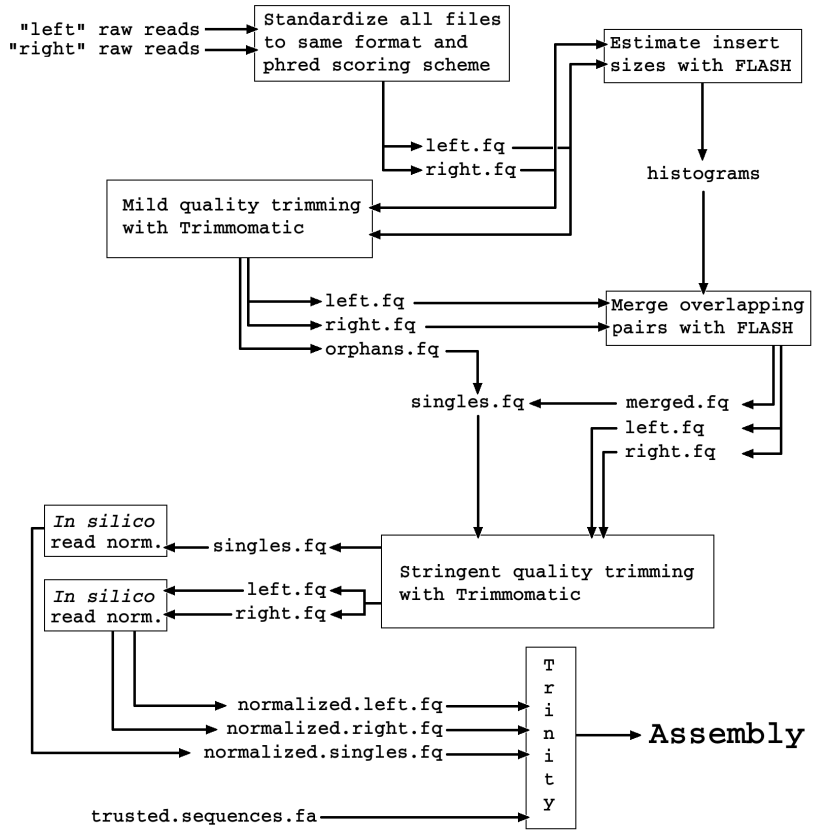



**Figure 3.3:** Schematic overview of the different steps of the pipeline. In between each two steps, a quality check of all FASTQ files was done with the program FastQC. All code can be found in the supporting materials.

## 3.4.2 Filtering and curation of the assemblies

### 3.4.2.1 Why filtering?

Even when every input read is subject to close scrutiny and the most stringent quality filters are applied, an assembly will contain spurious contigs. There are many possible sources of error, ranging from contamination with DNA or non-subject RNA, to sequencing errors that pass the quality criteria for base calls. To make matters worse, redundancy reduction with *in silico* normalization potentially comes with the cost of enriching data for errors. The rather short k-mer size of 25 may also lead to lack of resolution of alternative splicing events, potentially yielding non-genuine isoforms. The Trinity developers claim that their software is particularly suited for the resolution of splice variants, because of the special role for reads in graph compaction and traversing [112]. Finally, the more data go in, and the more different sources of input data are being used, the higher the expected number of contigs, comprising both genuine transcripts due to deeper sampling and allelic variation, and nonsense because of deeper sequencing of contaminations.

Among transcriptome assemblers, Trinity is known to output very high numbers of contigs, without providing much information about their support by input reads. The "too many contigs" problem is always hot topic in their discussion group. The Trinity developers mitigate this problem by providing the following explanation [sic]: *"Be cautious in filtering transcripts solely based on expression values, as you can easily discard biologically relevant transcripts from your data. Ideally, any filtering would consider a combination of expression values and functional annotation data, and filtering is currently more of an art than a science, and again, simply not needed in most circumstances unless you have a very clear objective in doing so."*

A pipeline for contig filtering, based on back-mapping rates of input reads, is available at the Trinity website, but it is solely based on mapping counts, which is a crude statistic. Also, there is no attention for sequence content.

### 3.4.2.2 Generating a reference set for RNA-Seq

assembly 5 was used as a starting point to generate a reference set that was going to be used in chapter 5. However, this assembly contained 170,066 contigs, and a more concise reference was absolutely critical[12].

I wanted to use sequence content as a filtering criterion, but these methods are computationally demanding, and I could not execute them on all 175,066 contigs. For

[12]With the original 170,066 contigs, statistical analyses on mappings just did not work. Furthermore, homology searches (BLAST or HMMER) on all potential CDSs just took too long.

example: the TransDecoder software offers a coding likelihood-score-based method for best candidate CDS detection [108], but it has a high error rate. The proposed solution is to incorporate BLAST hit information, but this is difficult for large datasets.

To overcome this problem, a first crude filtering was done based on mappings alone: all input reads were mapped onto the reference with STAR, and coverage per position was calculated using bedtools. I did not only concentrate on back-mapping counts, but tried alternative summary statistics that I considered more informative about contig quality, such as the homogeneity of back-mapping. Also the 30 libraries from the brain transcriptomics experiment described in chapter 5 were mapped. Based on mapping statistics, the assembly was reduced to a subset of 59,665 contigs, which is a more workable number for the computationally demanding methods that rely on sequence content.

This is a stepwise summary of how assembly 5 was filtered down to a reference for RNA-Seq. All code and additional details can be found in supporting materials.

1. **Back-mapping:** All sequences that served as an input to the assembly were mapped (STAR) back to the assembly. Coverage was calculated with bedtools.

2. **Mapping of RNA-Seq reads:** All reads from the RNA-Seq experiment in chapter 5 were mapped (STAR) to the assembly, and counts per sample per contig were recorded with SAMtools.

3. **Remove contigs with gaps:** 1,514 contigs that contained internal back-mapping coverage gaps (zero coverage) were excluded.

4. **Only keep contigs that get mappings:** Mapping counts of the 30 RNA-seq libraries were used to select those contigs that received 3 or more mappings from at least 27 libraries, and 5 or more mappings from at least 20 libraries. This selection of 59,665 contigs (from 43,964 "genes") was extracted from the original FASTA file with SAMtools.

5. **Get open reading frames:** First, all candidate CDSs above 50 amino acids were detected with TransDecoder in all possible open reading frames. In case the putative CDS would appear 5' truncated, it was either reported as such, or in case truncation at the first potential start codon resulted in less than 25% length reduction, it was reported as "complete" starting at the start codon.

6. **Best candidate CDS:** All detected CDSs served as input to a BLASTp search in a database of arthropod UniProtKB/Swiss-Prot and UniProtKB/TrEMBL[13] sequences, and a a HMMER scan (hmmscan command) in Pfam [199, 93].

[13] Swiss-Prot contains reviewed and manually annotated entries, while TrEMBL is a computer-annotated supplement.

Finally, the best candidate CDS was detected with a combination of the classic coding likelihood score [27], and demonstrated homology to known proteins or domains.

7. **BLASTp:** These best candidate CDSs served as an input to a BLASTp search in a protein sequence database of 20 organisms (table 3.4.) that contained several arthropods alongside some potential contaminants (downloaded from UniProt[14]), such as plant material (feed) or pathogens. All BLAST hits with e-value under 0.01 were retained.

8. **BLASTn:** All 59,665 nucleotide sequences were BLASTed in a list of trusted sequences, and their scores were recorded.

9. **RepeatMasker:** All 59,665 nucleotide sequences were screened for any kind of repeat or transposon-related pattern with the program RepeatMasker.

10. **CD-HIT-EST:** CD-HIT-EST was run on all 59,665 nucleotide sequences, (95% identity cut-off). No contigs were discarded, but it was recorded which contigs would have been discarded.

11. **Proceed filtering in R:** All information obtained in steps 1 - 10 was read in in an R session.

12. **Representative transposons:** A representative set of 2053 transposon-derived contigs was selected based on RepeatMasker output. Next, all clusters with good evidence for transposon content or repeats were removed. 52,753 contigs were left.

13. **Protein coding genes:** 15,532 contigs with putative coding regions were selected based on back-mapping quality, length of CDS and BLASTp scores with high bit scores for arthropods but low bit scores for contaminants. These were all single best isoforms per gene, based on bit scores, CDS length and mapping coverage.

14. **Non coding:** After filtering out transposon-associated clusters, and coding-gene associated "genes", 26,891 contigs in 24,294 "genes" were left. Here, 5,687 best candidate non-coding contigs[15] were selected, based on BLAST hits in trusted sequences, back-mapping and mapping by the RNA-Seq sets.

15. **Generate new FASTA:** In steps 11, 12 and 13, lists of names of selected contigs were outputted by R. Now, these 23,272 contigs were extracted from the original assembly with SAMtools.

[14] ftp://ftp.uniprot.org/pub/databases/uniprot/

[15] or at least, contigs without any strong *proof* for protein coding

16. **Remove redundant contigs:** Trinity sometimes outputted virtually identical contigs under different "genes" or even "clusters". To remove this redundancy, CD-HIT-EST was run on the 15,532 coding contigs merging contigs above 99% nucleotide sequence identity (down to 14,586 contigs), and on the 7,740 non-coding or transposon-derived contigs merging contigs above 95% nucleotide sequence identity (down to 6,948 contigs). The final file reference.fasta contained 21,534 contigs.

**Table 3.4:** Overview of the protein databases used to BLAST contigs for filtering. All proteomes were downloaded from UniProt and concatenated into one large database. Contigs with bit scores that were higher in arthropods than in potential contaminants were considered good candidates for locust protein coding transcripts.

| **Species name** | **Common name** | **# sequences** |
|---|---|---|
| **Arthropods** | | |
| *Acyrthosiphon pisum* | pea aphid | 35,514 |
| *Tribolium castaneum* | red flour beetle | 16,554 |
| *Drosophila melanogaster* | fruit fly | 13,705 |
| *Apis mellifera* | honey bee | 15,147 |
| *Zootermopsis nevadensis* | dampwood termite | 14,534 |
| *Pediculus humanus* | body louse | 10,754 |
| *Daphnia pulex* | water flea | 30,109 |
| **Potential contaminants and controls** | | |
| *Brassica rapa* | Chinese cabbage | 40,758 |
| *Triticum aestivum* | wheat | 94,896 |
| *Mus musculus* | mouse | 22,218 |
| *Caenorhabditis elegans* | nematode worm | 20,078 |
| *Homo sapiens* | human | 21,002 |
| *Gallus gallus* | chicken | 15,782 |
| *Metarhizium acridum* | parasitic fungus | 9,830 |
| *Gregarina niphandrodes* | septate eugregarine | 6,341 |
| *Nosema apis, bombycis* and *ceranae* | microsporidian parasite | 11,942 |
| *Pseudomonas aeruginosa* | bacteria | 5,550 |
| *Escherichia coli* | bacteria | 4,306 |

#### 3.4.2.3 Manual curation

A total of 186 sequences, originating from more than 80 genes, were manually curated. These were mainly genes that were of special interest, such as neuropeptide precursors,

or sequences that were already known from earlier studies in our lab. For each gene, all available *Schistocerca gregaria* and *Locusta migratoria*, and some *Gryllus rubens*[23] sequences were collected using BLASTn, BLASTp and a literature search.

A database was constructed with 5 Trinity assemblies made with a variety of parameters, an assembly of the 454 reads made with Newbler, and a handful of assemblies that were constructed with Velvet/Oases during my exploration of transcriptome assemblers.

For each of the target sequences, all hits in the *Schistocerca* transcriptomes were aligned and compared. In case of disagreement between assemblies, alignments and translated CDSs were investigated manually and sequences were selected with the best supported CDS. Also, to discriminate between genuine splice variants and assembly errors, k-mer counts (jellyfish) for the ambiguous regions were used to select best-supported contigs. In this manner, a set of 186 high-fidelity, manually curated sequences was created.

For those genes whose sequences were present in assembly 5, the original sequences were removed from the assembly, and the resulting set was concatenated to the manually curated set.

### 3.4.3 Annotation of the reference set

The final selection that was going to be used as a reference for RNA-Seq (chapter 5) contained 21,595 sequences. These sequences were first translated in 6 ORFs with TransDecoder. Next, all potential CDSs were used as a query for a HMMER scan (hmmscan command) in Pfam [199, 93], and a BLASTp in UniProtKB/Swiss-Prot. Standard parameters were used, except for a BLAST e-value cut-off of $e^{-4}$ i.s.o. $e^{-5}$. Next, based on likelihood scores [27] and hits by BLAST or HMMER, TransDecoder retained at least one CDS above 35 amino acids in 20,577 contigs (37,582 CDSs in total).

All predicted peptides were again subject to BLASTp and HMMER searches in the same databases (this time all standard parameters), and on top, untranslated nucleotide sequences were used as query for a BLASTx search in UniProtKB/Swiss-Prot.

All predicted peptides were queried for signal peptide sequences with SignalP (-f short), for ribosomal sequences with RNAmmer and for transmembrane helices with TMHMM.

All sequences were organized in a sqlite database (fig. 3.9) using a script called Trinotate (developed by the Trinity developers). In this database, every contig is represented by its gene- and transcript ID, nucleotide sequence and presumed peptide sequence. It contains all information about putative structure elements

(signal peptides, transmembrane helices, rRNA) and presumed homology (BLAST and Pfam). Based on the latter, it contains Gene Ontology terms [9], and two alternative vocabularies: eggNOG [198] and KEGG [135].

A text-based summary of the database was outputted and GO annotations were counted for "biological process", "molecular function" and "cellular component". Within biological process, annotations were counted for signal transduction (GO:0007165 and all nested GO terms, see supporting materials).

### 3.4.4 Neuropeptide precursor recovery and annotation

A database with known *Locusta* and *Zootermopsis* peptide precursors was constructed, mostly based on the supplementary data of Veenstra (2014) [272] and Hou *et al.* (2015) [125]. Since neuropeptide precursors may be difficult to assemble due to repetitiveness or ubiquitous alternative splicing [272], all contigs from all assemblies I ever made were queried in this database with BLASTx. Amino acid sequences of candidate neuropeptide hits were aligned and compared with sequences in other insects.

For each gene, the best supported sequences were maintained and translated with TransDecoder. Next, they were screened for signal peptides with SignalP 4.1 Server, and for di-basic cleavage sites and other candidate processing sites [271]. Resulting predicted neuropeptides were organized in a database that can be used for mass-matching in future mass spectrometry (MS) experiments (supporting materials).

## 3.5 Results

Many different assemblies were made. Some of them are discussed in the supporting materials. Briefly: assemblies resulting from the total input resulted in non-realistically high numbers of contigs (between 1.2 and 2 million "transcripts"). The overall quality of these assemblies was poor: although many expected sequences were present, they were often represented multiple times by contigs that were only marginally different from each other, but not represented within the same "gene". Also, these assemblies often contained many smaller contigs with low back-mapping coverage. In what follows, little focus will be on these large assemblies. An example of an assembly based on all input data is worked out in the supporting materials. Briefly: this assembly was generated based on all input reads, with standard Trinity parameters except for the min_kmer_cov parameter with a value of 2. This assembly contained 827,367 clusters with 1,208,777 ‘genes’ and a total of 1,330,397 contigs. After clustering with CD-HIT-EST at identity cut-off 95%, the number of contigs was down to 1,143,910. This assembly will henceforth be referred to as the "big assembly".

After *quite some trial and error*, it was decided that an assembly based on a single input set was the safest solution. Hence, in what follows, the focus will be on *assembly 5*: the assembly that was based on the reads of set 5 and a set of "trusted sequences". This set of contigs was used as the basis for a selection that served as a reference for the RNA-Seq experiment

### 3.5.1 Input data for assembly 5

After trimming-merging-trimming, the total input contained 14.5 million pairs, with a modulus average Phred score of 39, and 13.7 million singles with a length distribution between 75 and 190 bp (with a sharp peak around 100 bp), and a modulus average Phred score of 39.

### 3.5.2 Assembly 5: quality and contents

The assembly took 47 hours and yielded a total of 175,066 contigs, distributed over 140,093 genes in 107,942 clusters. N50 based on the longest isoform per gene was 1,106 and GC content was 43.33. The distribution of isoform numbers is shown in figure 3.4: 129,774 genes had a single isoform, and the highest number of isoforms for one gene was 157 (TR91546|c0_g1). Contigs belonging to this "gene" were generally very long (mean 8,986), and were deviant in many respects (e.g. very high standard deviation of back-mapping depth). I was unable to identify these sequences with BLAST.

The median of the mean back-mapping depth per position was 8.1, and since reads containing common k-mers were *in silico* normalized, more than 99% of contigs had mean back-mapping depths under 150 (and many were close to that number, see fig. 3.5). Nonetheless, the highest mean back-mapping depth was as high as 6,710, seen in a contig belonging to a cluster of transposon derived sequences (TR99767|c95). Most contigs that got high mean mapping rates were transposon derived.

A remarkable observation was an enigmatic pattern (fig. 3.6) in several summary statistics when contigs were plotted in the order they were created by Trinity[16]. Contigs were divided in two groups. The first group contained roughly 34,000 contigs, the second one roughly 141,000. Contigs in the first group were shorter, had smaller CDSs, and lower back-mapping density. They also typically only had one isoform per gene. It seemed like the second group was actually a mixture of contigs that looked much like the ones in the first group, and contigs that were generally longer with higher back-mapping coverage. During the quality filtering that led to the reference set, the filtered assembly became relatively enriched for contigs of the second group.

[16] A similar, but even more dramatic stratification was seen in the big assembly. See supporting materials.

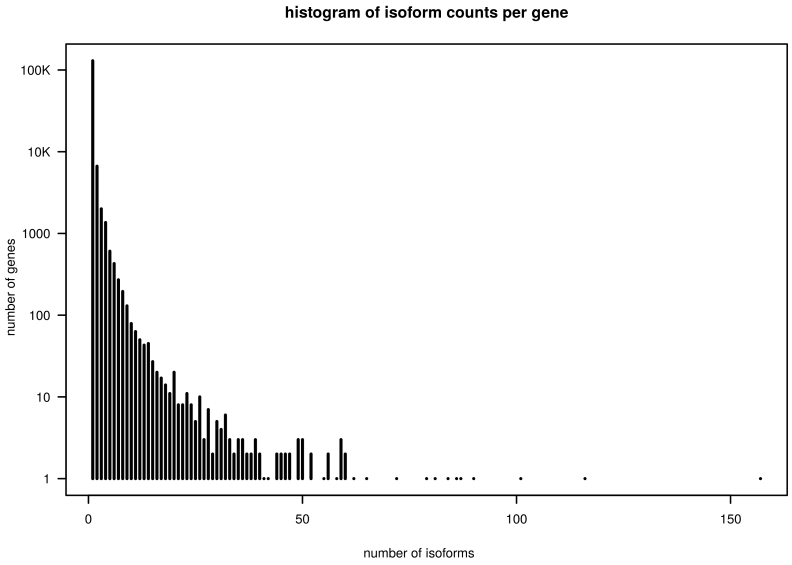



**Figure 3.4:** Histogram of the number of isoforms counted per gene.

### 3.5.3 The reference for the RNA-Seq experiment

The final reference contained 21,595 contigs, with a total GC content of 39.32%, a median contig length of 1,174 and an N50 of 2,200 (fig. 3.7). 18164 contigs had a CDS over 50 amino acids. 16,051 over 75, 12,832 over 100, and 447 over 1000 (fig. 3.8). The longest CDS was 8,441 amino acids long and coded for a member of the projectin/twitchin family of muscle proteins, that are known to be as large as 1,000 kDa [10].

10,276 contigs got at least one GO annotation (BLAST e-value cut-off: 0.00001, and Pfam with default domain noise cut-off). 8,738 had at least one biological process-related GO term, 8855 had at least one molecular function-related GO term and 9103 had at least one cellular component-related GO term. 7,164 contigs had at least one hit in all three GO term domains. 1,801 contigs had a signal transduction-related GO term. All this information is present in the annotation database (fig. 3.9) that is available in supporting materials.

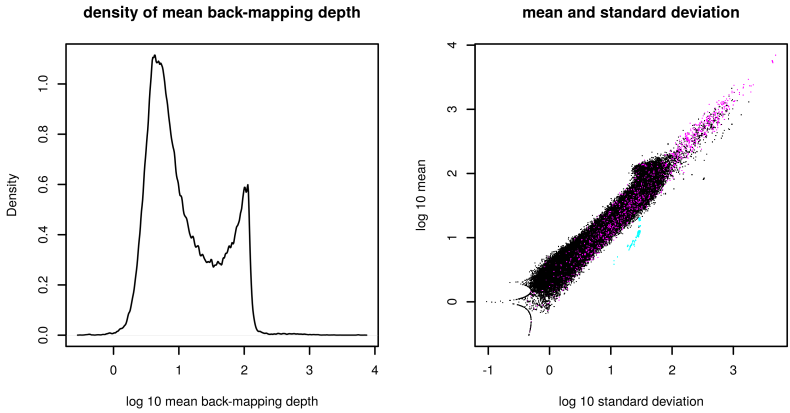



**Figure 3.5:** Left: the mean back-mapping depth seems to follow a bi-modal distribution, but actually the second "bump" is the result of *in silico* read normalization, which truncates the distribution by eliminating reads with common k-mers. Right: generally, the standard deviation of back-mapping depth is in the same order of magnitude as the mean. The data points in cyan are the 157 isoforms of gene "TR91546|c0_g1", that seem to have higher standard deviations than would be expected from their mean mapping depth. The data points in magenta are putative transposon derived sequences.

### 3.5.4 Neuropeptide precursors

A total of 51 NP precursors were found: 39 with complete CDS and 12 partial. For 12 precursors, two or more splice variants were found. For 37 of the 44 peptides described by Clynen & Schoofs [55], the precursor sequence is now known, but also many additional precursors were found.

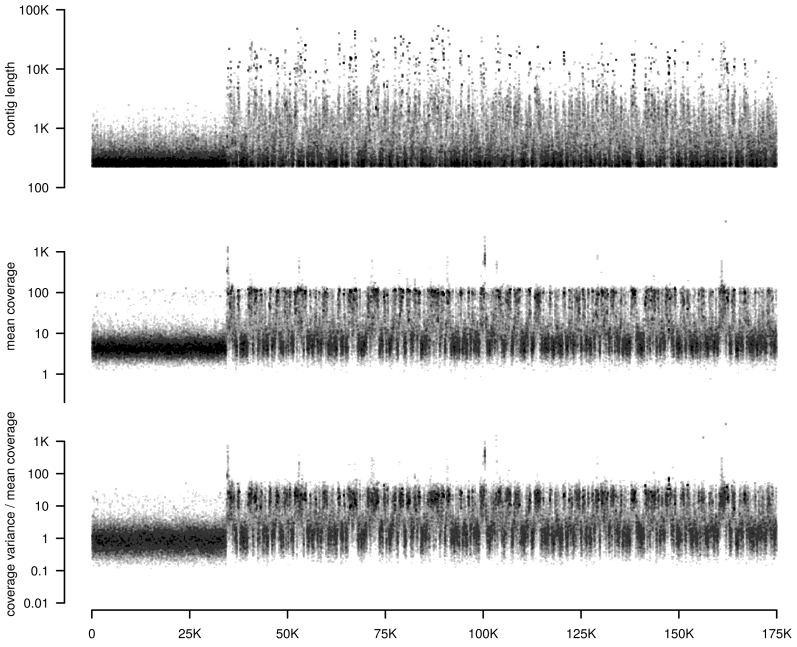



**Figure 3.6:** When summary statistics of contigs are plotted in the order they were created by Trinity, a stratification appears: the first (approximately) 34,000 contigs clearly have lower length, lower mean coverage and a lower variance per mean coverage.

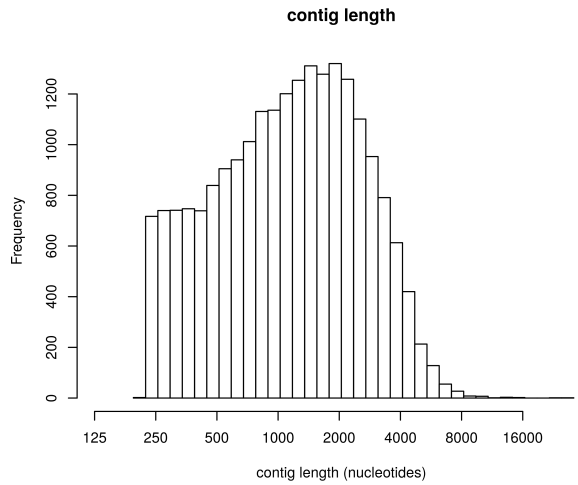



**Figure 3.7:** Frequency distribution of contig lengths in the 21,595 selected contigs.

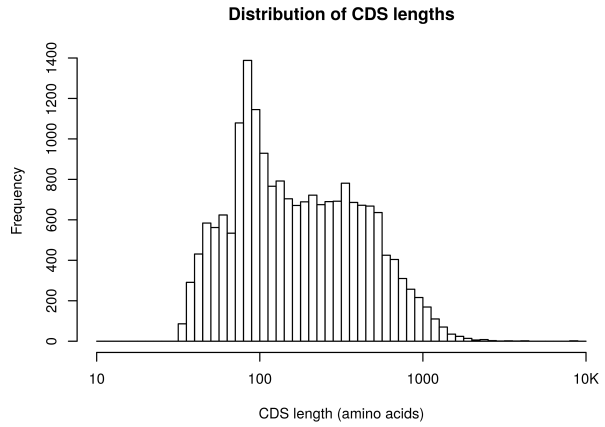



**Figure 3.8:** Frequency distribution of CDS lengths in the reference transcriptome used in the RNA-Seq experiment (see chapter 5). Minimal recorded CDS length was 35. Possibly, some predicted CDSs do not correspond to real proteins, especially in the shorter CDS range.

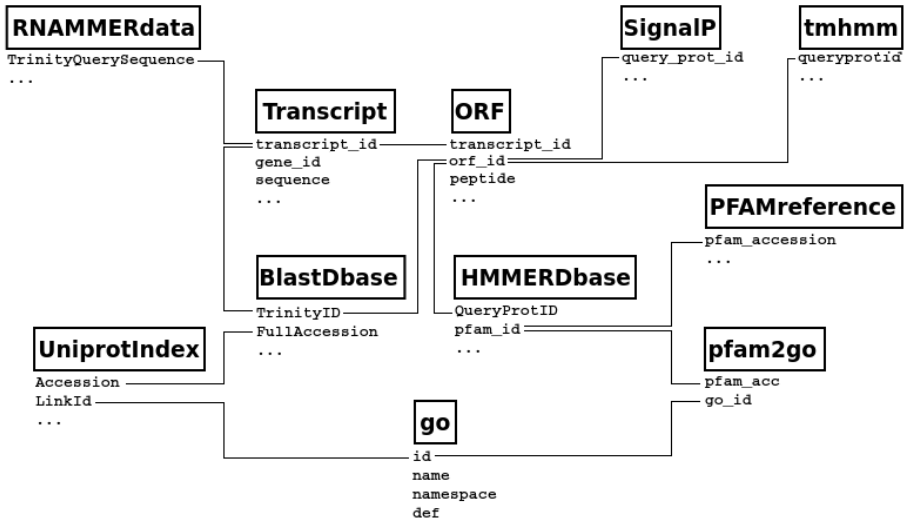



**Figure 3.9:** Structure of the relational database that contains all annotation information. Centrally, contigs (*Transcript*) are present with their sequence, and their detected CDS (*ORF*). These two tables link to a range of other tables either through the transcript identifier, or through a specific CDS identifier. Hits in a BLAST-database or HMMER-database link to tables that contain elaborate information (*UniprotIndex* and *PFAMreference*), that in their turn link to gene ontology terms (*go*). Dots represent extra information that does not link to other tables.

**Table 3.5:** Neuropeptide precursors found in the *Schistocerca gregaria* transcriptome. The references are those publications containing sequences used in homology search: *Schistocerca* peptide amino acid sequences obtained with mass spectrometry (Clynen & Schoofs 2009 [55]), *Schistocerca* peptide precursor nucleotide sequences (mainly Badisco *et al.* 2011 [13]), or *Locusta* peptide precursor nucleotide sequences (mainly Veenstra 2014 [272] and Hou *et al.* 2015 [125]).

| Name | # isoforms | references | comments |
|---|---|---|---|
| **COMPLETE CDS** | | | |
| adipokinetic hormone 1 | 1 | [55, 13, 272] | |
| adipokinetic hormone 2 | 1 | [55, 13, 272] | |
| adipokinetic hormone 4 | 1 | [272] | |
| allatostatin A | 1 | [55, 13] | |
| allatostatin B | 4 | [272] | |
| allatostatin C | 2 | [272, 273] | Weak signal peptide |
| allatostatin CCC | 1 | [13, 272, 273] | Called AST-CC by Badisco [13] |
| allatotropin | 1 | [55, 13] | |
| baratin | 2 | [185] | |
| calcitonin b | 2 | [272] | |
| calcitonin like diuretic hormone | 2 | [272] | |
| CCH amide | 2 | [13] | |
| $[His^7]$corazonin | 1 | [55] | |
| crustacean cardioactive peptide | 1 | [55, 13] | |
| eclosion hormone 1 | 1 | [272] | |
| eclosion hormone 2 | 1 | [272] | |
| ecdysis triggering hormone | 1 | [272] | |
| elevenin | 1 | [272] | |
| glycoprotein hormone alpha 2 | 1 | [272, 125] | |
| glycoprotein hormone beta 5 | 1 | [272] | |
| hypertrehalosemic hormone | 1 | [55, 13] | |
| insulin related peptide | 1 | [55, 272] | |
| ion transport peptide | 2 | [55, 13] | |
| ITG containing peptide | 2 | [13] | *Apis* ITGQGNRIF orthologue |
| leucokinin | 1 | [125] | |
| myosuppressin FLRF-amide | 2 | [55, 272] | |
| neuroparsin | 4 | [130, 13, 272] | |
| neuropeptide like peptide | 1 | [13, 272] | Formerly wrongly fused with NPF |
| neuropeptide F 1 | 2 | [55, 13, 269, 272] | Formerly wrongly fused with NLP |
| neuropeptide F 2 | 1 | [272] | |
| orcokinin | 1 (2?) | [55, 272] | Very repetitive, difficult to assemble. |
| ovary maturating parsin / CRF-like DH | 1 | [13, 272] | |
| pigment dispersing factor | 1 | [125] | |
| CAP2/myotropin | 2 | [55, 13, 272] | |
| RY-amide | 2 | [272] | |
| short neuropeptide F | 2 | [55, 73, 272] | |
| SIF-amide | 1 | [272] | |
| SMY-amide | 1 | [272] | |
| tachykinin | 1 | [55, 13] | |
| **PARTIAL CDS** | | | |
| allatostatin CC | 1 | [273] | |
| bursicon alpha | 1 | [272, 125] | |
| bursicon beta | 1 | [13, 272, 125] | |
| CNM-amide | 1 | [272] | |
| FMRF-amide | 1 | [272] | |
| natalisin | 3 | [125] | |
| PBAN/hugin | 1 | [55, 272] | Middle bit missing |
| CAPA | 1 | [55, 272] | Middle bit missing |
| tryptopyrokinin 2 | 1 | [272] | |
| salivary gland salivation stimulating | 1 | [272] | Not sure if this is orthologous to the *Locusta* version |
| sulfakinin | 1 | [55, 272] | Middle bit missing |
| trissin | 1 | [125] | |

## 3.6 Discussion

### 3.6.1 Technical considerations

At the beginning of this project I set the objective to generate a reference transcriptome as complete as possible, with representation of transcripts from different tissues under a variety of conditions. In the supposition that for me and my colleagues such a reference would become a highly valued resource, I may have failed to timely recognize that generating an assembly based on sequences from different sequencing projects, with different sequencing technologies and even with subjects from different colonies, was not just excessively ambitious, but also not very clever.

Many of the technical issues are explained and extensively discussed in the supporting materials. In this first part of the discussion, I will briefly highlight some important or interesting technical topics, but most observations are *post hoc* and hence are not ideally suited to draw conclusions.

#### 3.6.1.1 Library preparation

The first library prepared at KU Leuven (set 5a1) was of poor quality, containing high levels of ribosomal sequences, and having position-specific low base call quality in the first sequenced mate (fig. 3.1), which indicates that a disturbance may have occurred at a specific moment during sequencing. The authors of FastQC suggest that this kind of pattern may occur due to bubbles or debris moving through the flow cell.

The second attempt for library preparation (set 5a2) was more successful and yielded a high number of good quality reads with relatively low levels (< 10%) of rRNA, which indicates that the introduction of an mRNA purification step was probably the right thing to do.

Whether the *in vitro* library normalization was of much added benefit is hard to tell, because these normalized libraries supplied less than one percent of the total number of reads in spite of their equimolar concentrations when they were loaded onto the sequencer. On top of that, they had much higher rates of mismatches in their barcodes, with less than 30% having the 6 nucleotides of the index completely right. The only explanation I can think of[17] is that the process of *in vitro* normalization enriches barcode mismatches because perfectly matching sequencing adapters and barcodes are preferentially digested by the duplex specific nuclease. This observation raises the question of whether similar processes may be happening on the level of the insert, enriching the actual reads for rare variants or errors introduced during early library preparation steps.

[17] With thanks to Alejandro Sifrim.

It may well be that during demultiplexing, many reads from the normalized library were "missed" because they had even more errors in their barcodes, but this definitely cannot explain the orders of magnitude difference in read numbers between normalized and unnormalized libraries.

#### 3.6.1.2 Roche FLX+ 454 reads

At the moment when the sequencing took place, the addition of 454 reads just looked like one of many challenges, but at the time of writing, the discussion has reached consensus that 454 reads are of little added value in this kind of projects[18]. No De Bruijn graph-based assembler was designed to incorporate reads of such different length, and no assembler other than Newbler was designed to account for the typical homopolymer error pattern seen in 454 reads.

Given the methods that are currently available, the only way I could imagine 454 reads to contribute, is to assemble them separately and funnel them in through the "trusted sequences" route in Trinity. However, I did not trust the 454 reads enough, except for the well supported contigs that would be the easiest to assembly anyway.

#### 3.6.1.3 Contig numbers

Whenever multiple input libraries were combined, the resulting numbers of contigs were high. The parameter with the most dramatic influence on the total number of output contigs was min_kmer_cov, where a value of 2 (as opposed to the default 1) reduced the output by about 40%. This corresponds quite well to the observation that in the input set, about 45% of the 25-mers occur only once. However, even after this reduction, contig numbers were still high. There are multiple possible explanations, and with the information at hand now, I can only speculate as to the real cause.

A first possibility is genetic variation: some contigs showed minor differences that may well have been alleles, often in different Trinity "genes". In this respect, the inclusion of set 4 was definitely a risk, because the colony in Jena is undoubtedly genetically distinct from ours. This is especially problematic because both populations are most probably quite inbred, meaning that in neighbouring loci, alternative allele combinations may have been fixed [25], which would make contigs appear more different than they actually are.

However, in assemblies based solely on animals from our colony, some pairs of contigs also looked like alternative alleles. Perhaps some of these alternatives can be explained by sequencing errors, but I probably underestimated the genetic variation, both in

[18] In the meantime, Roche has discontinued the production of FLX+ sequencers altogether

terms of magnitude, and in its influence on reference assembly. The actual mRNA samples would not bring us very far in formally quantifying hererozygosity because they all consist of pools of individuals. The mathematical derivation of the expected genetic variation of our colony would be a challenge as well, given that little is known about the founder population, and no historical records of the effective population size are available. In a study on genetic variation in 3 commonly studied locust species, heterozygosity of *Schistocerca gregaria* was quantified based on nine microsatellite loci, and proven to be much lower in our laboratory population compared to wild populations [25]. However, the residual genetic variation seems to have been high enough to preserve alternative alleles for a substantial number of transcripts.

One striking example was the family of chemosensory proteins (CSP) that was included for manual curation given their presumed importance in early phase transition [109]. In assembly 5, Trinity produced 110 contigs that had clear sequence identity to 18 known *Locusta* CSP sequences. After manual curation, these 110 contigs were reduced to 19 "genes" with multiple subtle variants. Without closer investigation it is hard to exclude recent duplication events, especially since it is known that this gene cluster has expanded recently in grasshoppers (pers. comm. Tyler Raszick). On the other hand, these genes may have been particularly diverse in the founder population, since genes involved in olfaction and chemosensation can be subject to balancing selection [62].

To me, allelic variation is the more parsimonious explanation for the high number of variants, but it is certainly not the only explanation for the inflated contig counts. When the 1,330,937 contigs of the big assembly were clustered with CD-HIT-EST with a 95% identity cut-off, only a reduction of about 15% was seen. A second explanation may be library-specific contaminations. The big assembly contained 1,330,397 contigs, but only 135,935 got back-mappings by all input libraries. Input sets 2 and 3 were brain-only samples, but even adding the back-mappings common between set 4 and set 5, this number was still lower than 160,000. Conversely, 417,139 contigs only got mappings from set 2 and/or set 3, and these contigs were usually short, low coverage and without any proof for protein coding. Moreover: an assembly on sets 2 and 3 only contained 891,217 contigs, which is a high number for a single-tissue transcriptome. A closer look into the publication of Falckenhayn *et al.* (2013) [89] is not informative: transcriptome data are barely shown and there is no mention of mapping rates against the *Schistocerca gregaria* EST database [13]. My suspicion is that these sets actually contained quite high levels of genomic DNA that, given the limited sequencing depth, assembled only in small fragments.

Another library-specific idiosyncrasy was the presence of *Gregarina* genetic material. *Gregarina* is a parasitic apicomplexan alveolate related to *Plasmodium* and *Toxoplasma* that is known to be present in the intestines of locusts [41]. In the big assembly, about 5000 contigs were *Gregarina* derived. They got the vast majority of their back-mappings from sets 4 and 5: the samples that contained digestive tract. Since sets 2

and 3 were brain-only, they did not contain *Gregarina* sequences. *Gregarina* contigs were mainly long fragments with high coverage, and about half of them had more than 5 independent CDSs longer than 125 amino acids, suggesting that the contamination with was at least partly genomic.

#### 3.6.1.4 Stratification of ordered contig output

Both in the big assembly and in assembly 5, clear stratification was observed in statistics such as contig length or back mapping coverage when plotted in the order the contigs were created. For assembly 5, two groups were observed. Rather surprisingly, contigs in the first group (about 34,000 contigs) had lower back-mapping density than the second group (about 141,000 contigs), alongside other hallmarks of lower quality, such as shorter length, smaller CDSs and a gradual relative depletion during quality filtering. This is unexpected, because Trinity works with the Inchworm "seed" k-mer prioritization scheme, greedily building contigs in the order of the most abundant k-mer, and clustering contigs that show k-1 overlap. Hence, I would expect Trinity to actually output the best quality contigs first. I opened a discussion in the Trinity users Google discussion group, where Brian Haas suggested the observations may be an artefact from *in silico* normalization, where the maximal number of back-mappings is capped.

This cannot be the answer however, because similar patterns are observed in contig lengths, or in the number of mappings from reads that were never normalized, such as the RNA-Seq reads from chapter 5.

### 3.6.2 The reference transcriptome

The final reference transcriptome should in theory be a good representation of the transcripts present in the brain, with each protein coding gene represented by the isoform coding for the longest protein. In the distribution of CDS lengths (fig. 3.8), the high peak in the shorter length range probably contains false positives, but the likelihood of obtaining an ORF by chance quickly declines as a function of CDS length, so the larger part of detected CDSs must be genuine.

1,801 contigs, including 51 neuropeptide precursors, had a signal transduction-related GO term, which is a more than 16 fold increase compared to the number of 108 mentioned by Badisco *et al.* (2011) [13]. However, I count 364 contigs with signal transduction-related GO terms in the original annotation file, which makes me believe that in the original study, nested terms were not counted.

But even 364 contigs is still a five-fold difference, that I attribute at least partly to improvements in annotation algorithms, and the much higher amount of sequence

and annotation data present in public databases compared to five years ago. In the original EST database, only 3887 contigs were classified according to at least one of the GO vocabulary domains, while in this updated reference transcriptome, 10,276 contigs got at least one GO annotation, among which 7164 contigs with classifications in all three GO term domains.

The neuropeptide precursor sequences will be a valuable resource for future research. Out of 39 complete precursors, 19 were hitherto unknown. The sequence of the neuropeptide F precursor that was previously reported [13] also contained the precursor sequence of neuropeptide like peptide. Moreover, Van Wielendaele *et al.* (2013) reported this sequence to be confirmed after re-sequencing the cDNA [269]. However, not any of my assemblies contained such a chimeric precursor, and there is no report of such a gene fusion anywhere in literature. Hence, I think this must have been an assembly error.

In the last comprehensive review of known *Schistocerca* peptides [54], 44 neuropeptides were known from MS. For some of those, the precursor sequence is still partly unknown. This includes sequences coding for 7 MS-predicted peptides. On the other hand, 84 additional neuropeptides were predicted, and 56 potentially associated peptides. It should now be possible to design PCR primers to recover missing stretches in some of the partially assembled precursors, such as sulfakinin, PBAN/hugin and CAPA, given that both ends of the CDS are known. Sulfakinin is an interesting case. The active peptide has been observed in *Schistocerca* and *Locusta* [55], and there is good evidence that in the latter, a fully fledged sulfakinin signalling system is present [288]. However, no complete gene sequence could be found in the *Locusta* reference genome, neither could the complete precursor sequence be assembled from transcriptome reads [272]. Perhaps this precursor contains a stretch of sequence that is extremely difficult to amplify (high GC) or assemble (repetitive).

Some neuropeptides that are expected to be present in locusts have not yet been found in *Schistocerca gregaria*. These include accessory gland myotropin, inotocin, proctolin and orcokinin B, which in *Locusta* is a splice variant of the orcokinin A that I did recover [272]. Obviously, absence of evidence is no evidence for absence, and I expect most of these precursors, and by extension genes of many gene families, to be sequenced either in future RNA-Seq projects that target specific organs or life stages, or in the reference genome sequence that the community eagerly anticipates.

## 3.7 Acknowledgements

Reads of set 2 and 3 were shared by Liliane Schoofs (KU Leuven) and Frank Lyko (Deutsches Krebsforschungszentrum, Heidelberg, DE). Reads of set 4 were shared by Heiko Vogel (Max Planck Institute for Chemical Ecology, Jena, DE). Libraries of set 5

were prepared by Vanesse Brys and sequenced by Jeroen Van Houdt. All computations were executed on the computing server at the university of Leipzig that I was kindly granted access to by Katja Nowick and colleagues. Special thanks to Alejandro Sifrim, Liesbeth Badisco, Rohit Kolora, Gero Doose, Jens Steuck and Alexander Donath for helping me with the bioinformatics throughout this project.

# Chapter 4

# An automated behavioural assay and quantitative analysis of phase-related behaviour

> All models are wrong; some models are useful.
>
> George E. P. Box

## 4.1 Introduction

In this chapter I discuss and develop methods to study phase-related behaviour of a locust in an arena assay, with a focus on automated video tracking. Next, these methods will be used to explore a range of observable behaviours in the light of phase state and behavioural phase transition.

### 4.1.1 Measuring behaviour

Our understanding of animal behaviour would be dire without methods to describe it in terms of variables. In an experiment, we can only draw inferences about the effects of treatments when we have access to trustworthy quantitative measures of behaviour.

Some variables may be directly observable, like the number of times a bird visits its nest to feed the young [26], or the proportion of time spent at a certain place in function of the availability of food [116]. These observations can be done in real-time, with manual annotations or with event logging software. In other cases, behaviours may be a lot more difficult to directly observe, requiring specialized devices or computer programs.

Typically, video or audio recordings are used in these cases, and dedicated software extracts all relevant information. Sometimes this software is specifically designed for a certain type of animal or experiment, such as SonoBat™ and Batsound®, programs for the analysis of bat calls. But, also more general purpose programs are available, such as Noldus EthoVision® and Swistrack [160] for the tracking of a variety of animals in videos, or Noldus The Observer® for event logging and behavioural coding and segmentation.

In recent years, the field of quantitative ethology has been pushed forward by progress in computer algorithms that can extract and interpret large numbers of variables from video or sound recordings of a range of animals, often in real-time. Especially with the advent of machine learning, and the availability of more powerful computers, the possibilities have expanded tremendously [134]. For example: it is now possible to individually video-track interacting fruit flies in a group [204], to monitor the movements of individual body parts of a mouse [184], or to record and interpret animal sounds against a background of environmental noise [270].

Some of the above mentioned software is open source and free [160, 134], but often these programs are proprietary and expensive. Hence, it is still useful to research and develop open source methods that can be freely redistributed and further improved by the scientific community. In this chapter, I describe a behavioural assay that has

been used extensively in the field of locust phase transition, and I present a pipeline of open source tools that are specifically designed for the analysis of this assay.

### 4.1.2 Video tracking

In the last few years, automated video tracking has entered the field of locust phase transition research [107, 64, 243]. The tracking of a moving object in a video usually follows the same basic steps. First, the object has to be discriminated from the background, and areas in a video frame (e.g. pixels) need to be classified as belonging to the object or to the background. Many methods are available, ranging from simple background subtraction to complex pattern recognition algorithms. Next, the position of the detected object needs to be recorded. This can range from a simple reduction to a single point (centroid) in the 2D plane, to records of surface area and shape, or even the delineation of the entire body in 3D space [184]. Finally, positional information is extrapolated to real spatial coordinates based on earlier calibration, or camera-angle arithmetic, and the entire series of coordinates is integrated over time in order to yield a coherent track that describes the movement of the object.

In the case of background subtraction-based methods (like the ones used in locust behavioural assays [242, 243]), there are several possible sources of error in obtaining the coordinates of the centroid of a tracked object. A first possible source of error is variation in the contrast between the object and the background. This variation may be caused by heterogeneity of illumination, resulting in differences in contrast within one video, or from colour variations between subjects, causing differences in contrast between videos. A second source of error originates from small body movements that are not part of a whole-body progression, for example grooming behaviour in locusts. These movements may cause instability of the centroid while the animal does not actually engage in real locomotion [120]. This "body wobble" has already proven to be a challenge in the automated tracking of locusts in an arena assay [64, 243]. A third source of error is *precision level noise*, which is caused by the discrete nature of a pixellated image: if the boundaries of the tracked object do not fall on pixel boundaries, the centroid may oscillate between neighbouring pixels [120, 64].

Much of the final quality of the track primarily depends on the quality of the videos. For example: the higher the frame rate, the more information is available to filter a stable track out of a series of centroid coordinates. Also, the higher the resolution of the image frames that make up the video recording, the lower will be the precision level noise. Lossy video compression may drastically lower the potential of a video to be tracked correctly. Apart from mere video quality, the experimental set-up is also of importance. For instance: lighting conditions and background should be chosen such that the animal to be tracked contrasts maximally while minimally interfering with its natural behaviours.

### 4.1.3 A history of quantitative studies in locust behaviour

Early laboratory experiments confirmed that extreme phases of this same species could be reproduced in captivity by varying the rearing density [90] but these experiments were focused on long term phases and only looked into changes in morphometrics and colouration. The first person to really apply a quantitative framework to behavioural phase change was Peggy E. Ellis [77, 78, 79, 80, 81, 83, 84], who was collaborating with the London anti-locust research centre in the fifties. In a series of studies on *Locusta migratoria* nymphs, Ellis and co-workers developed the first behavioural assay that could be used to quantify some aspects of phase-related behaviour. The set-up was rather simple: groups of hoppers were put together in a cage, and a subset of marked individuals were closely observed. The main observed behavioural variable was marching activity, defined as continuous walking with the antennae held erect [77].

In a first series of experiments, the marching activity of five marked individuals was recorded for 10 minutes, every half hour for 6 to 8 hours [77, 79]. In other experiments,

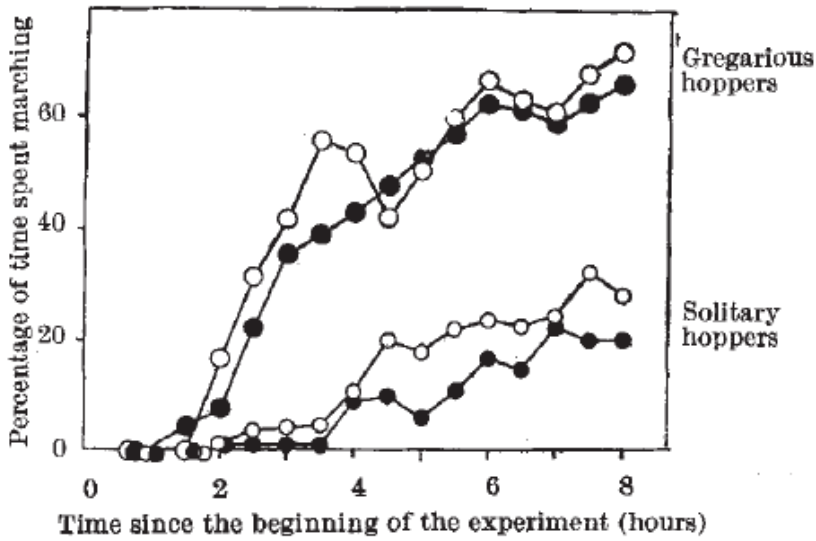



**Figure 4.1:** The first accounts of marching behaviour of locusts (*Locusta migratoria*) under laboratory conditions, from Ellis 1950 [77]. 18 $4^{th}$ instar hoppers, either fully gregarious or second generation solitarious, were placed together in an arena and the marching behaviour of marked individuals was recorded. Apart from noting that the solitarious hoppers spent less time marching, it was also observed that they marched more slowly.

marching behaviour was either scored during a continuous one-hour observation [79], or in one-minute time bins spread over 3.5 hours of crowding. This experimental set-up was used to demonstrate that gregarious locusts have the strongest tendency to march, and that nymphs reared in visual and olfactory isolation showed stereotypic low activity levels upon crowding, followed by a change in their propensity to march during the following hours. A similar set-up was used to demonstrate that gregarious animals were more inclined to march when put in larger or denser hopper bands [79].

Another experimental set-up employed by Ellis, consisted of a ring-like test cage divided in sectors. The readout of the assay was the count of hoppers per sector of the ring, which would give a measure of group coherence and the degree of concerted movement (marching) [80, 82]. A good part of what we know today about phase transition, was first demonstrated using the circular arena assay. However, this assay was less suited for larger scale experimentation, because there was no read-out of phase state for each individual animal, and hence large numbers of locusts were needed for each experimental replicate.

Ellis also used annular test galleries in individual-based assays [79]. In its most simple form, the assay was just a single ring in which the path of a test locust was observed against a "background" of 50 marching hoppers. Numerous other set-ups were used, ranging from an arena with decoys [85], consisting of immitation locusts, or live locusts of either phase tethered to the arena floor with fine wire, all the way to assays where a test hopper was presented the sight of taxidermized hoppers in a merry-go-round mounted on a bicycle wheel. Read-outs ranged from mere activity-based variables such as marching, to attraction/repulsion-related behaviour through the counting of interactions with decoys.

Ellis had a remarkable sense of innovation, and never really settled with one assay. Rather, she seemed to prefer variations that were explicitly designed to probe a specific aspect of phase-related behaviour. Group-based circular assays have remained a prime tool in phase transition research until this day [100, 101, 102, 119, 8, 7].

A major breakthrough in the assaying of phase-related behaviour was achieved when in 1993, Roessingh *et al.* [215] devised a behavioural assay where a single locust was observed in a rectangular arena with on one side a chamber containing a stimulus group of gregarious locusts, and an identical but empty chamber on the other side. This polarized design provided a way to simultaneously quantify some measures of activity and the animal's tendency to seek or avoid the company of other locusts. Moreover, compared to the aggregation-based assays, this individual-based set-up allowed for the observation of many more replicates while using a smaller number of locusts. In this manner, more specific (and often labour-intensive) experimental treatments could be tested.

In the earliest work by Ellis, the small set of observed behaviours were directly

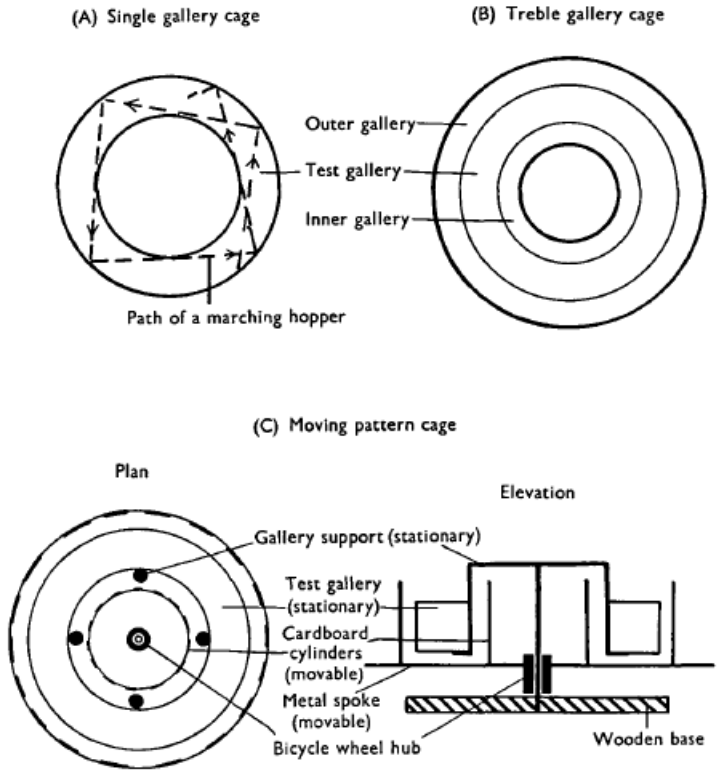



**Figure 4.2:** Some of the more adventurous individual-based annular test galleries used by Ellis and co-workers. The test galleries were 8 cm across with an inner diameter of 28 cm. Either they contained hoppers themselves (A), or decoys or hoppers were presented in an outer gallery (B and C).Figure from Ellis 1953 [79]

compared between experimental conditions, without much formal inference towards phase state. In the group-based assay, Poisson statistics of hoppers per arena sector were used, and in later work, an "associative index" was calculated based on pictures of hopper bands in the arena [118]. With the Roessingh-style assay, a larger number of variables came available, sampling the locust's activity and attraction towards its peers, alongside small body movements such as grooming. The probability of belonging to a given phase ($P_{greg}$) was expressed in function of a set of observed variables, both activity and attraction/repulsion-related. Based on the assumption that phase transition entails a series of concerted changes in the variables that make up the model, logistic regression has been used extensively to assess the "gregariousness" ($P_{greg}$) in different species of locusts under a variety of experimental conditions.

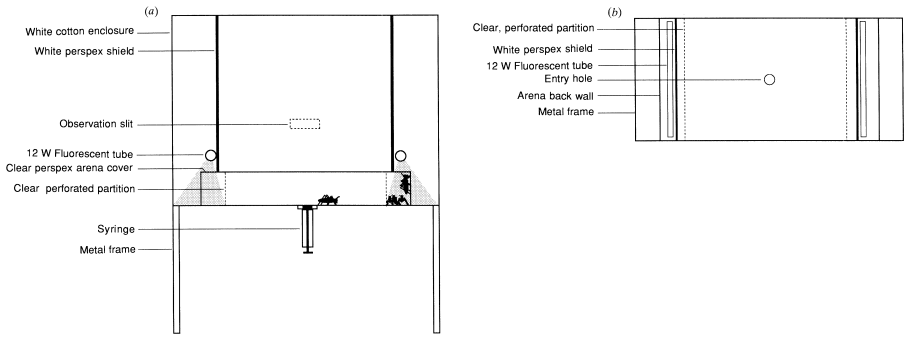



**Figure 4.3:** The original arena as presented by Roessingh *et al.* (1993) [215].

Recently, the assumption of concerted changes was challenged by Tanaka & Nishide (2013) [257], who used yet another behavioural assay: $5^{th}$ instars were introduced in a rounded rectangular arena[1], with two transparent perforated plastic cups spaced about 50 cm apart, one of which containing 12 gregarious nymphs. The attraction of the subject towards other locusts was measured based on 3 photographs taken after 20, 30 and 40 min.. Both the frequencies of visits to four sectors of the arena, and the number of times the locust reached the stimulus cup were compared between long term phases, and between experimental conditions. In a response, Rogers *et al.* (2014) [217] re-analysed their data from previous studies and argued that attraction/repulsion behaviours do change together with activity-related variables, and that changes in activity alone are not sufficient to explain the differences in $P_{greg}$ between experimental conditions.

## 4.2 Aims

1. To develop a procedure for video tracking locusts in the arena, and to segment the tracks into behavioural parameters.

2. To use this procedure to characterize the differences in behaviour between long term phases, and changes in behaviour observed during phase transition.

3. To compare the performance and outputs of the new method to older data from manual analysis, and to contrast my findings with what is known from literature.

4. To develop a reliable predictor that can be used to quantify gregariousness in an experimental setting.

[1]They call it an elliptical arena, but their figures show a rounded rectangle rather than an ellips.

## 4.3 Methods

All scripts, ImageJ macros, Swistrack component files, the code of the tracker and of *tandy.R* can be found in the supporting materials, with extensive annotations for future use.

### 4.3.1 The observed locusts

All animals were reared as described in the general methods chapter. All multivariate analysis and model building was based on a set of mixed sex $5^{th}$ instar nymphs containing 125 first-time observations of untreated gregarious animals and 110 first-time observations of untreated solitarious animals that had all been raised in the apartments.

A total of 81 animals (including 26 dsGFP injected control animals from an experiment by Rut Vleugels) were crowded for 24 hours and observed before crowding, after 2 hours, after 4 hours and after 24 hours. Crowding was done in a 30cm x 30cm x 30cm perspex box together with 25 mixed sex $5^{th}$ instar gregarious locusts. A maximum of 13 solitarious locusts was simultaneously crowded. During crowding, fresh feed was present in order to prevent cannibalism.

### 4.3.2 Automated behavioural assay

In our lab, behavioural analyses were previously based on the methods described by Roessingh *et al.* (1993) [215]. Videos were analyzed by watching them in real time and using event logging software (The Observer®, Noldus) to manually record a set of eighteen different behavioural parameters. It should be noted that although the manual assay was labour intensive, and there was a risk for observer-dependent biases, it was still a good method for the recording of a range of behaviours, including subtle movements such as grooming and swaying.

In November and December 2012, I visited the lab of Malcolm Burrows and Swidbert Ott at the university of Cambridge to get a training in automated video tracking. Together with Katja Stettin, Dr. Ott developed a method where a locust is tracked in a video using the open-CV-based tracker Swistrack [160], generating a track of coordinates and body axis positions from which a series of behavioural parameters are extracted with a MATLAB/Octave script called *tandy*. These original methods are well described in the PhD dissertation of Katja Stettin [243].

Back home in Leuven a new arena was built and all necessary software was set up for this system to also work for our videos. Apart from a struggle to get Swistrack

installed, and some minor adjustments of the parameters in Swistrack and *tandy*, this set-up was quickly up and running in Leuven, and has been used and described in the PhD dissertation of Rut Vleugels. However, when I wanted to further develop the methods, the limitations of Swistrack became a constraint. Also, my familiarity with the MATLAB/Octave syntax was too limited to further develop *tandy*. Therefore, I decided to convert the entire locust tracking project to the R environment.

### 4.3.3 The arena

The arena that was used for our previous behavioural assays was a copy of the arena described by Roessingh *et al.* (1993) [215]: a rectangle of 57 cm x 30 cm with a Cartesian grid on the floor, where the test locust was separated from the stimulus group by a single perforated perspex window. Both the grid and the interference by the stimulus group made our old videos difficult to video track. To overcome these problems, we based the design of a new arena on the arena used in Cambridge.

The new arena had a 42 cm x 28 cm blank floor, and a double perforated perspex separation between test locust and stimulus group, removing the latter from the field of view of the camera. Also, the edges of the arena were rounded in order to overcome the locust's strong preference for residing in the corners. Two daylight mimicking fluorescent tubes (Lucky Reptile Daylight Sun T5 Eco, 8W) were positioned on top of the stimulus and control side. In order to guarantee steady illumination, the arena was placed in a wooden cupboard away from any external light source. Both arenas are compared in figure 4.4.

### 4.3.4 Video recording

A color camera (Panasonic WV-CP480/G) was installed 125 cm above the arena floor, high enough to make the image almost rectilinear. Next, the objective was focused one single time and the set-up was fixed in this position for the entire duration of the project. During all recordings, the temperature was between 28°C and 31°C. At the beginning of each recording session, the arena floor was covered with a clean sheet of white paper with a hole cut in the middle to allow the test animal to enter the arena. A plastic chamber with a stimulus group of 30 $5^{th}$ instar long-term gregarious locusts was placed on one randomly selected side of the arena, while an identical but empty chamber was placed on the other side. 10 min. prior to the assay, the locust was placed in an opaque modified 20 ml syringe, deprived from most sensory input. The locust was introduced in the assay straight from this syringe and filmed for at least 500 seconds. Video files were compressed in MPEG-4 format (codec Divx 5.x 6.x) and stored on a desktop (Microsoft windows) computer.

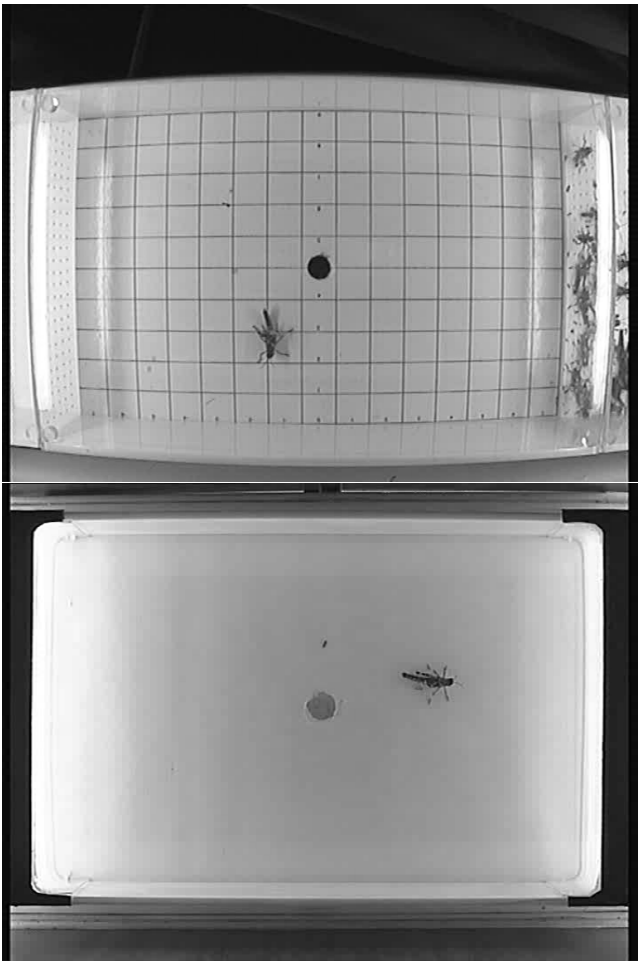


**Figure 4.4:** Top: screenshot of a video with the old arena. The stimulus group is clearly visible on the right hand side. Bottom: the new arena: the Cartesian grid has disappeared, the corners have been rounded and the stimulus group is masked from the camera's field of view. The image is also less optically distorted.

Videos were recorded in frames of 720 x 576 pixels at 25 frames per second (fps). The MPEG-4 codec that was used to compress the videos resulted in a significant loss of information. The intra-frame compression, which lowers the resolution of the images much like in JPEG compression, compressed the colour channels (U and V) to blocks of 8 x 8 pixels, which resulted in a locust-coloured haze around the actual locust in the picture. The luma channel (brightness, Y) was less compressed.

There was also inter-frame compression, with B-frames recalculated based on neighbouring reference frames (I-frames). Every second frame was an I-frame, which is a higher than usual density.

### 4.3.5 Metadata recording

For each video, the following metadata were recorded:

**1.** Information about the circumstances of recording: the name of the operator handling the locust, the day of the week and date, the side of the stimulus group and the time since the stimulus group was introduced in the arena.

**2.** A unique identifier for the locust, its phase at birth, date of birth and sex.

**3.** Information about the locust at the time of video recording: its instar stage and the date of its last molt, its color and amount of body black (see figure 4.5), when it had last been fed and where it had lived in the week preceding the observations (pots, small apartments or large apartments, plus on which floor), and the number of previous observations.

**4.** Treatment

**5.** Remarks, especially checking if all legs and antennae were intact

I wrote a small program that interrogated the observer, and already auto-completed all the increments and the formerly recorded information in case of multiple observations. This helped keeping all metadata in a consistent format, and prevented errors with identifiers, dates etc. For example: it produced a warning when the date of birth, instar stage and date of last ecdysis were in an illogical combination.

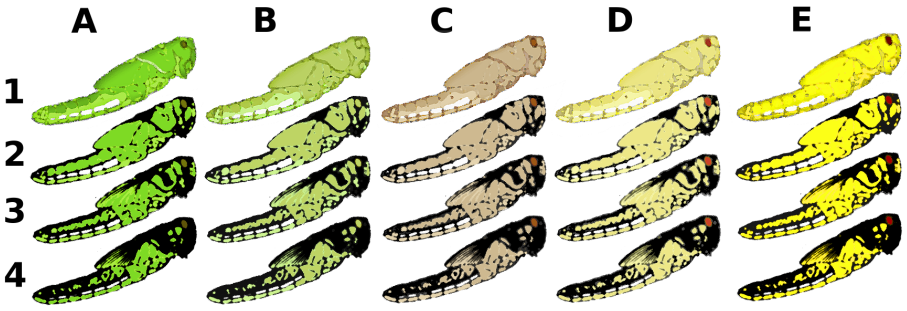



**Figure 4.5:** Chart to score the level of black and background colour of desert locusts. Black patterning is loosely based on Stower, 1959 and colouration is based on my own observations of animals from our colony. Images were deliberately stylized to avoid distractions from details.

## 4.3.6 Video tracking

### 4.3.6.1 Swistrack

SwisTrack [160] is a powerful free[2] software for tracking moving bodies from a camera or a recorded video input source. Written in C++, it uses Intel's OpenCV library for fast image processing, providing an interface for AVI files.

The original pipeline relied on Swistrack[3], but it had shortcomings that made me decide to write my own tracker. First, Swistrack was difficult to install, and often I found the software broken after a system update on my computer. Second, I wanted to be able to adjust and optimize the tracker further than Swistrack would allow me to. For example, I wanted to use dynamic thresholding during background subtraction: something that Swistrack cannot do.

### 4.3.6.2 Generating auxiliary files.

Before starting the tracker, some auxiliary files were generated for all videos. First, all first frames were extracted and converted to mask files with a custom imageJ macro. These binary images would later tell the tracker where it is allowed to look for objects to track. Next, the tracker also needed calibration files that translate pixel coordinates to real-world coordinates. The more precise the calibration, the better the discrimination between e.g. sitting near or on the window. Since the arena position

[2]The program is available under the GNU General Public License (GPLv3), which grants the user the right to freely use the software both for research purposes and industrial application.

[3]In the supporting materials I give a detailed overview of the entire Swistrack pipeline.

could shift a bit between different videos, an individual calibration file was generated for each video with another imageJ macro and an R script that together extracted the coordinates of some landmarks in the first frames.

#### 4.3.6.3 The tracker

R cannot parse video, so the tracker starts with a call to the program ffmpeg[4] to decompose the video in a gallery of separate image files. Since the color channels (U and V) were of lower resolution, only the Y-channel (luminance) was extracted, and saved in a lossless format (png). Given that the B-frames were of lower quality than the I frames, only the latter were extracted. As a result, the frame rate was halved to 12.5 fps. After creation of the image gallery, the first frame was read in as the background, assuming no locust was present at the start of video recording.

Also, much like in the Swistrack pipeline, a mask that shades the regions where the tracker should not look for a locust, was read in. Since the frames were grey-scale only, they were internally coded as 720 x 576 matrices of values between 0 (black) and 1 (white). This allowed for easy arithmetic like subtraction, multiplication or re-scaling. The tracker parses the image gallery one frame at a time, executing a series of operations that are illustrated in figures 4.6 and 4.7.

**1.** Subtract the background from the frame, take the absolute value and multiply by a mask. This results in an almost black image with a white locust.

**2.** Threshold the 250 brightest pixels. This results in a pure black-white image.

**3.** Narrow down the search frame by calculating the marginals (row and column sums) of the matrix, fitting a smooth spline, and reducing the matrix to the 100 x 100 pixels around the maxima of these splines. Now, this smaller sub-frame should contain the locust. This operation sits upstream of the more intense computations, and by reducing the matrix to only a fraction of its original size, a significant speed gain is obtained.

**4.** In the background subtracted image apply a Gaussian smoother to the search frame defined in step 3.

**5.** Threshold the blurred background subtracted image at the brightest 500 pixels, which is about the size of an average $5^{th}$ instar locust in my videos.

**6.** Find all *connected components* (blobs) in the picture, in the case of multiple blobs, only select the largest one. Determine its center of gravity and transform the coordinates back to the original axes of the full-size picture.

[4]https://ffmpeg.org/

7. Determine the angle of the long axis of the blob relative to the x-axis using Eigenvectors.

During this loop, not only the coordinates and blob orientation are recorded, but also the size of the search frame window, the blob size, the two grey scale thresholds and the variance between neighbouring frames (as a means of motion detection). Finally, also the image intensity of the background subtraction in a small sector around the introduction hole in the middle of the arena is recorded. These recordings are used to make an educated guess about the first frame in which a locust appears. This is

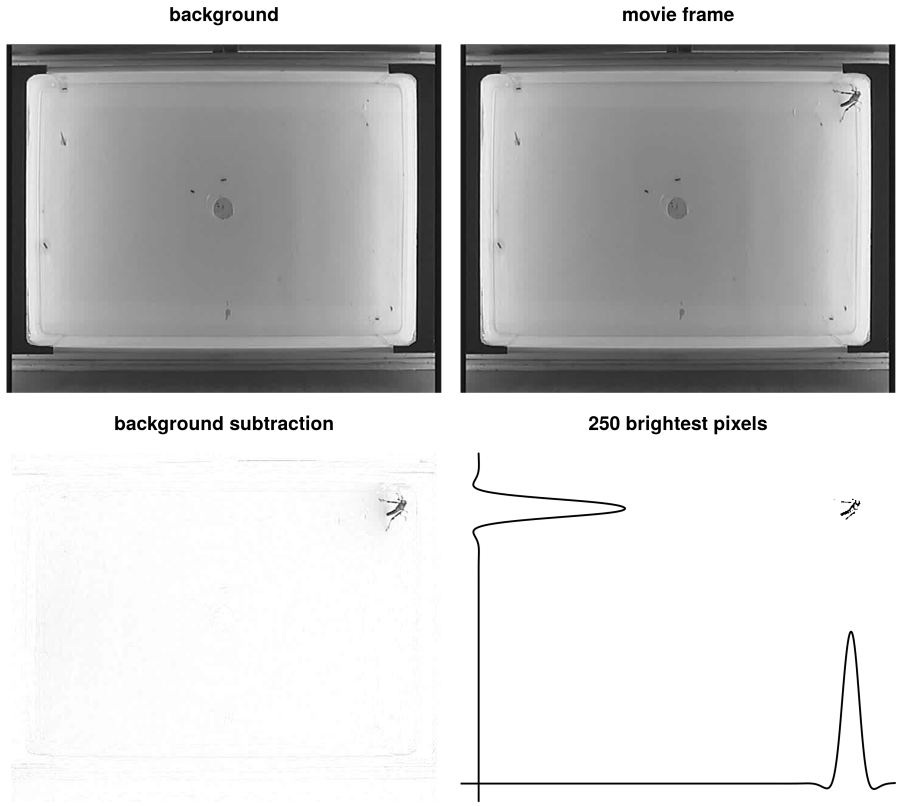



**Figure 4.6:** Phase 1 of the tracker: each video frame is subtracted from the background. Next, the 250 brightest pixels are selected, assuming that most of them belong to the locust. Finally, a smooth spline is fitted through the marginals of this thresholded picture (here printed in negative), resulting in a presumptive centroid, around which an area of 100 x 100 pixels is selected.

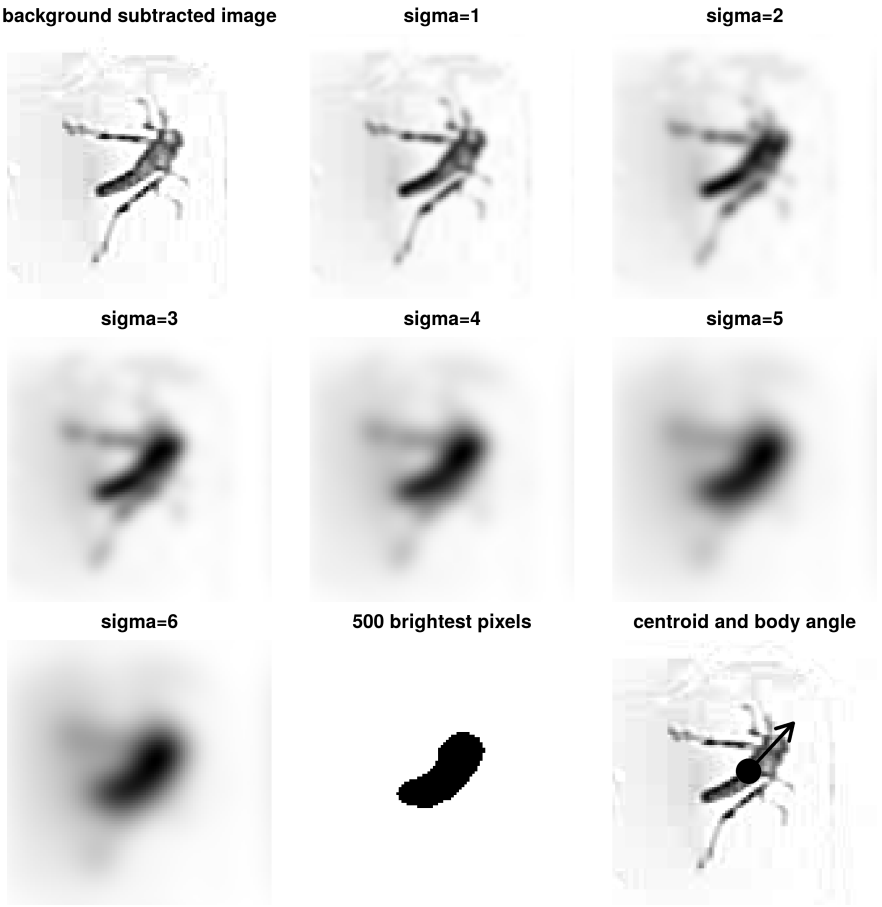



**Figure 4.7:** Phase 2 of the tracker: the image is now zoomed in on an area of 100 x 100 pixels around the putative centroid. Now, the image is blurred before thresholding it at the 500 brightest pixels. Here, 6 different blurs are shown for illustration purposes, but in the final tracker, a sigma value of 4 was used. The last panel shows the final position of the centroid and the observed body angle. Note that all pictures are printed in negative for illustration purposes.

marked by the tracker to be the putative beginning of the track. In a final step, a linear model is used to convert the pixel coordinates to real-world coordinates, using the same landmark-based coordinate sets as used in the original Swistrack pipeline.

The tracker generates two outputs: an image file with graphs that give a quick idea about the quality of the track, allowing manual filtering of obvious errors, and a table with coordinates (both in pixels and in real-world), the body angle, and the indication of the putative beginning of the track.

### 4.3.7 Segmenting tracks into behavioural variables

A track consists of a series of coordinates and body angles, and the real challenge is to extract behavioural variables from this information. Some variables can be derived directly from the coordinates, such as the time spent in a certain sector of the arena. Other variables, such as rest time fraction, have to be derived from the velocity. Yet others, such as climbing, require the assessment of a combination of position and movement.

As mentioned earlier, a program called *tandy* was developed in Cambridge to execute this task, and we have successfully used this in our lab. The original *tandy* script, written in MATLAB/Octave, was well adapted for the set-up in Cambridge, and could be adjusted to an acceptable level to also fit our set-up in Leuven. In the beginning I used the parameters as they were optimized in Cambridge, except for some minor adjustment to accommodate for the differences in framerate and arena dimensions between the Leuven and Cambridge system. This is the way videos were analyzed for the PhD dissertation of Rut Vleugels.

However, after writing my own tracker with its own specific output, and after changing and tweaking the parameters of *tandy* until the point where I had to make changes in the source code, I decided to further develop the program to my needs. Since I am not proficient enough in MATLAB/Octave, I converted the code to the R environment.

#### 4.3.7.1 tandy.R

Although the R script was more or less written from scratch, most concepts and the general flow of logic are entirely borrowed from the original *tandy* script that was written by Swidbert Ott. Below, the outlines of my version of *tandy* will be presented, with a focus on the differences with the original.

**Reading in the coordinates as points in the complex plane:**
*tandy.R* starts with reading in the track files generated by the tracker, and extracting

the first 500 seconds following the introduction of the locust. All operations are executed on the real-world coordinates, internally coded as a position vector with the x and y coordinates as complex numbers. By describing the coordinates as single complex numbers in the complex plane, most relevant information can be obtained with straightforward mathematical operations. For example, the velocity is the ordered set of all differences between neighbouring coordinates, which in its turn can be decomposed in speed and direction by calculating respectively the absolute value and the argument of the set.

**High speed movement:**
Unlike the original *tandy*, *tandy.R* focuses on high speed movement before looking at positional information. The rationale is that a logical vector of "high speed movement" can be used as a mask in downstream steps, avoiding the recording of positional information during jumps, or, more importantly, during erroneous episodes of fast movements that are actually due to e.g. a brief trembling of the arena, light flicker, or the presence of a smear or dropping on the arena floor. Here, the track may jump to some far corner and back to the locust. Obviously, this should not be counted as positional information, nor should it be considered to be any kind of genuine movement like jumping.

The first subcategory of activity that is recorded is "jumps", defined as movement above a jump speed threshold. Desert locusts' jump speed is in the order of magnitude of 1 m $s^{-1}$ [221]. However, given the low frame rate of the videos, a lower threshold was chosen. This was rather straightforward since jumps really spike above the background by a large margin. The threshold for jump speed was set to the equivalent of 2 cm per frame, which corresponds to 25 cm $s^{-1}$. Apart from mere high speed, an additional condition for a movement to be considered a real jump, is that it should result in a net movement of the center of gravity of the locust by at least a certain distance. This may lead to not counting jump-bump-jumps against walls or windows, but the distance threshold is important to account for spikes where the tracker jumps away and back to the locust, which does not result in net movement. Jumps starting from walls or windows are not considered jumps, because it can just be the locust falling on the floor, leading to an overestimation of the number of jumps for locusts that climb a lot.

**Segmentation of the track into episodes of activity:**
After recording high-speed movement, the track is segmented in "active" versus "passive". This has turned out to be the most challenging part of the entire pipeline because of the difficulty to discriminate between jitter of one or two pixels (precision level noise) and genuine small movements while the body is motionless (body wobble). Several criteria are simultaneously applied to segment a track in "active" and "passive" episodes.

A first criterion is the number of pixels that are visited within a given number of frames, using a sliding window. The resulting vector of counts can then be used as a

noise filter: precision level noise mostly jumps up-and-down between two or three positions, while genuine movements typically visit more unique pixels, even if the locust stays more or less in place. I applied this filter with a minimum of at least 5 unique pixels over a window of 10 frames.

A second criterion is the speed between neighbouring frames. Here, speed is defined as the absolute value of (complex) velocity, meaning that no directionality is taken into account. Even at a frame rate as low as 12.5 fps, speed shows a spiky pattern during most movements. Also, because speed has no direction component, body wobble and other sources of jitter just look like extended periods of moderate speed. A way around this problem is to calculate velocity over an interval of multiple frames. This is mathematically equivalent to a convolution of the velocity vector with a flat gliding mean before taking its absolute value[5]. The advantage of using a convolution rather than calculating velocity over a multiple-frame interval, is the possibility to first mask high-speed episodes. A downside of working with convolved speed, is that the thresholds set for walk speed correspond less directly to measurable quantities. Finally, a bout of activity will usually become a little bit wider than its real duration, leading to a slight overestimation of activity time, and in some cases also the conjoining of multiple distinct bouts of activity.

The classical behavioural models of phase transition often contained subtle body movements, such as grooming [215], or body-swaying [123]. The annotation of these subtle movements has proven to be difficult. This may be the downside of using a Gaussian smoother in the tracker, which is especially harsh for small movements in the periphery of the blob, resulting in fewer visited pixels. However, the smoother is also necessary to yield a robust centroid that does not dance around on the locust because of illumination flicker or precision level noise. I have tested different combinations of tracker parameters and motion detection methods, but when aligning the automated segmentation with my own manual annotations of the videos, it turned out time after time that small body movements drown in the background noise: if there is no actual net movement, it seems to be impossible to tell apart grooming from jitter. I even experimented with Hidden Markov Models to classify spiky patterns of activity into blocks of *bona-fide* activity, but this was to no avail. In the end I gave up on the annotation of subtle movements and decided to segment the track into "passive" versus "obviously moving". Hence, it is important to remember that the behavioural variable *rest time fraction* inevitably includes time spent grooming or swaying.

**Positional information:**
Only after the track has been segmented into jumping, activity and resting, the positional information is taken into account. Positional information is important to quantify potential attraction towards the stimulus group, but also to further segment

[5] I have worked on the optimization of track segmentation looking into convolution and speed over a larger number of frames as two competing strategies, until I found out they are mathematically equivalent and indeed yield exactly the same result.

activity into more specific behaviours such as walking or climbing. The positional information is also used to measure the locusts' track length and the ground it covers, as well as its path tortuosity which is expressed as the fractal dimensions of these quantities (suffix .FD).

**Speed and bout durations:**
After the segmentation of activity into walking and climbing, the average and median walk and climb speed are measured, defined as the mean or median velocity during the entire duration of all walks or climbs. These measures are complemented with walk and climb bout speeds: the medians and means of all net-movement components of walks and climbs. Since locusts tend to walk along the walls and windows a lot, it is sometimes difficult to precisely define walks and climbs. For details I refer to the comments in the source code of *tandy.R*.

Many more variables are recorded, all explained in detail in the comments of the *tandy.R* script. At the end of each run, *tandy.R* outputs an ethogram that shows the basic segmentation of the track into rest, walking, fast walking, climbing, jumping and turning, alongside a list of observed variables, as detailed in tables 4.1 to 4.3. Most variables are expressed in terms of a time fraction (suffix .TF) or counts (suffix .N)

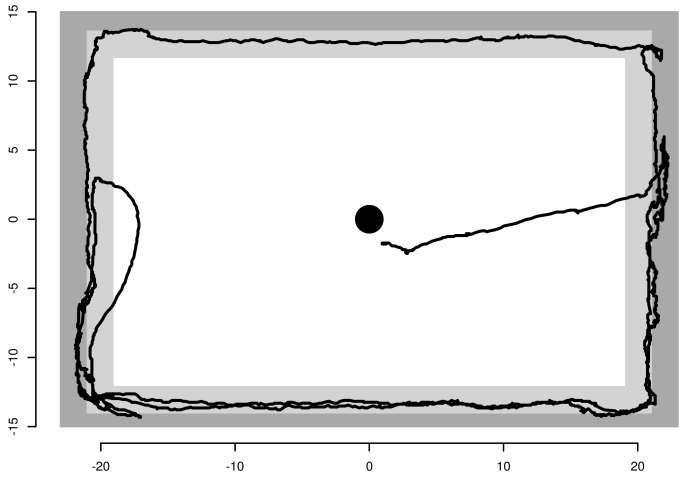



**Figure 4.8:** Video *greg_AA32_model_1* was one of the videos used to precisely calibrate the arena edges. Black is the path of the locust, dark grey regions are walls and windows, light grey shade is up to a body length away from edges, and white is the region where the locust is considered to be "flat on the floor".

#### 4.3.7.2 Input parameters for tandy.R

A set of 23 input parameters determine the exact behaviour of *tandy.R*. Some of them modulate the output format, others inform the program about the dimensions of the arena, and 12 parameters directly control the definitions of behavioural variables, as summarized in table 4.4. Parameter values in *tandy.R* are intended to correspond as much as possible to real-world values that can be directly observed. However, it is still a challenge to pick the right values in order to obtain realistic measures of behaviours that can be used to tell apart solitarious and gregarious locusts. At first sight, it may seem alluring to optimize all input parameters (and hence the definition of behavioural variables) for their performance in separating both phases, but this strategy may come with a catch. First, there is the risk of over-fitting models on particularities of the training data that do not generalize to test sets. Second, some combinations of parameters may accidentally produce spurious differences between solitarious and gregarious locusts that can be solely attributed to their differences in appearance, for example creating artefacts when the locust has a very pale colour.

In order to fine-tune the parameters, I wrote a graphical interface where the input parameters can be adjusted using tick boxes and sliders, changing the output variables in real-time. In this way, parameter values could be tested while directly comparing the output to the actual video. For the segmentation of the tracks into stretches of "rest", "walking", "fast walking" and "climbing", a manually annotated set of 60 randomly selected fragments of 1 min. each was used. For the calibration of turns, 5 carefully selected representative turns in gregarious animals were used. The exact values for the margins between floor, walls and windows were determined using videos where the locust regularly switches between climbing and walking (fig. 4.8)

### 4.3.8 Exploration of behavioural variables and their contribution to phase

#### 4.3.8.1 The data

Principally, this chapter is about the analysis of the data issued from the automated assay described above. However, it is interesting to contrast the results of a new technique with its predecessor (the manual assay). Throughout the larger part of the analysis, the following 3 datasets will be used:

**Manual data:** A set of 192 observations from the previous behavioural assay, containing 95 solitarious and 97 gregarious mixed sex $5^{th}$ instar nymphs. All observations by Heleen Verlinden and Rut Vleugels. Note that the solitarious locusts

in this set were all raised in the plastic containers, which may have influenced their behaviour and even their phase state.
**Automated training:** A training set of mixed sex $5^{th}$ instar nymphs. In order to have them as uniform and "naive" as possible, only first time observations were used. All solitarious animals were raised in the apartments, and all animals were well fed in the morning. A total of 110 solitarious and 125 gregarious animals met these criteria. All observed in the automated setting.
**Automated total:** A set of 586 observations of mixed sex $5^{th}$ instar nymphs which includes the training set, 81 animals crowded for 2, 4 and 24 hours, (some of them are control animals injected with dsGFP from the PhD project of Rut Vleugels), and a small supplemental test set of both gregarious, crowded and solitarious locusts. This set also contained repeated measurements of some solitarious animals.

#### 4.3.8.2 Selection and transformation

After tracking with the new tracker, a total of 53 behavioural parameters were extracted by *tandy.R*. These were all visually inspected for their distribution, possible outliers and their capacity to separate phases. The variables attraction speed (mean and median) and laterality were discarded not only because of their poor performance to separate phases, but also because of the many missing values they contain.

Missing values typically occurred when the animal is very passive. For example: it is impossible to estimate the walking speed if the animal does not walk. Unsurprisingly, this problem is biased towards solitarious locusts, since sitting still is a hallmark of their behaviour. In the case of walk and climb bout duration, the null values were left untouched, since according to my observations, solitarious locusts tend to walk or climb in shorter bouts. In this manner, the zeros seemed to connect quite naturally with the range of observed values, and the benefits of imputing zero values outweigh the costs of discarding two variables.

However, the case of walk bout speed and overall walk speed is much more troublesome, because (much to my own surprise) in my observations, solitarious locusts show the tendency to walk faster than gregarious locusts while others do not walk at all during the entire observation. This results in a bimodal distribution of their walk speeds, with one peak at a value higher than the average gregarious walk speed, and a sharp peak on exactly zero. Here, zero clearly means "unobserved" rather than "slow". This kind of problem can often be tackled by imputation based on regression of correlated parameters [129], but the problem here is that if the locust did not move in the first place, there is little information to impute from. Hence, I decided not to use walk speed and walk bout speed in any model or classification tool. I did keep them in my data for multivariate analysis, where I just replaced the zero values with the class means of solitarious and gregarious locusts. Finally, walk

bout length is a bit of an intermediate, being composed of both duration and speed. This makes it difficult to impute, but since it does not really discriminate between solitarious and gregarious locusts, I just left this variable out of the analysis.

Among the remaining variables, many had more symmetric distributions after a variance stabilizing transformation. Hence, all time fractions were arc sine square root transformed, and all counts were incremented by 1 and log-2 transformed. These operations made the data more symmetrically distributed, and decreased the differences in variance between the two phases. Transformations towards more symmetrical distributions are especially necessary for the multivariate techniques described in the next paragraph, since their outcome critically depends on the variance structure of the data.

This leaves us with exactly 50 variables. 10 of these variables are 5 means and 5 medians of the same variables. In order to not artificially blow up redundancy in the data, I only worked with the means (except for modelling with *random forests*, see later) which reduced the set to 45 variables.

**Table 4.1:** Activity-related variables. TF = time fraction, N = number, FD = fractal dimension.

| | |
|---|---|
| rest.TF | TF spent passive. |
| walk.TF | TF spent walking, defined as movement on the floor with convolved speed above the walk threshold. |
| fastwalk.TF | TF spent walking fast, with convolved speed above fastwalk threshold. |
| climb.TF | TF spent climbing, defined as movement on walls and windows with convolved speed above the walk threshold. |
| jump.N | Number of jumps, defined as movement with speed above *jumpthresh_speed*, covering at least a distance of *jumpthresh_length*. |
| walk.N | Number of walks, defined as activity bouts with speed above *walkthresh* that are at least 25 % on the floor. |
| border.N | Number of walks along the border, defined as walking less than a body length away from walls or windows at least half of the time. |
| climb.N | Number of climbs, defined as activity bouts with speed above *walkthresh* that are at least 10 % on walls or windows. |
| longwalk.N | Number of walks taking longer than the *duration.longwalkbout* threshold. |
| longclimb.N | Number of climbs taking longer than the *duration.longwalkbout* threshold. |
| turn.N | Number of turns, defined as Hamming-filtered angular velocity above *turnthresh* while speed is lower than *walkspeed*, while flat on the floor. |
| walkspeed | Mean or median walk speed, defined as the mean convolved speed while walking. |
| walkboutspeed | Mean or median walk bout speed, defined as end position - startposition per time for each walk bout that is at least 80 % on the floor. |
| walkboutlength | Mean or median distance of walkbouts that are at least 80 % on the floor. |
| walkboutduration | Mean or median duration of a walk bout. In the case of hybrid walk-climb bouts, only the part walking is counted. |
| climbboutduration | Mean or median duration of a climb bout. In the case of hybrid walk-climb bouts, only the part climbing is counted. |
| ground_covered | Number of unique 1 cm x 1 cm quadrants visited. |
| ground_covered.FD | The fractal dimension of ground covered, defined as the logarithm of the proportional increase in number of visited quandrants, measured with 7 exponentially decreasing quadrant areas. |
| path_length | The total length of the track, but not while jumping. |
| path_length.FD | The fractal dimension of path length, defined as the logarithm of the proportional increase in path length measured with 7 exponentially decreasing time slice widths. |

**Table 4.2:** Position-related variables. TF = time fraction.

| | |
|---|---|
| hole.TF | TF spent less than two body lengths away from the introduction hole. |
| onfloor.TF | TF spent on the floor of the arena. |
| flatonfloor.TF | TF spent on the floor at least a body length away from walls or windows. |
| wall_window.TF | TF spent on walls and windows, regardless of activity. |
| border.TF | TF spent within a body length of walls and windows, but not on them. |
| window_stim.TF | TF spent at the stimulus window. |
| window_cont.TF | TF spent at the control window. |
| corner_stim.TF | TF spent in the corners at the stimulus side, defined as the centroid less than a body length away from the corner coordinates. |
| corner_cont.TF | TF spent in the corners at the control side. |
| wall.TF | TF spent on the wall. |
| border_stim.TF | TF spent within a body length from, but not on the stimulus window. |
| border_cont.TF | TF spent within a body length from, but not on the control window. |
| border_wall.TF | TF spent near and on walls. |
| third_stim.TF | TF spent in the third of the arena nearest to the stimulus group. |
| third_cont.TF | TF spent in the third of the arena nearest to the control side. |
| half_stim.TF | TF spent in the stimulus half of the arena. |
| middlethird.TF | TF spent in the middle third of the arena. |
| region_stim.TF | TF spent near or on the stimulus window and in the stimulus corners. |
| x_distance.final | Mean or median distance from the stimulus group in the last 10 seconds of the track. |
| x_distance | Mean or median distance from the stimulus group during the entire track. |
| x_distance.max | The closest the locust approaches the stimulus group during the entire assay. |

**Table 4.3:** Combinations and others. TF = time fraction, N = number.

| | |
|---|---|
| climb_stim.TF | TF spent climbing on the stimulus window. |
| climb_cont.TF | TF spent climbing on the control window. |
| climb_stim.N | Number of climbs on the stimulus window. |
| climb_cont.N | Number of climbs on the control window. |
| kick_in | Time needed to first leave the introduction hole area. |
| attractionspeed | Mean or median base 2 logarithm of the mean ratio of velocity components towards and away from the stimulus group. |
| laterality | Base 2 logarithm of the mean ratio of sideways over forwards velocity components while walking. |

**Table 4.4:** 12 most defining input parameters for tandy.R, with their default values.

| | | |
|---|---|---|
| bodylength | 2 cm | Used to determine if a locust is near walls or windows, or in corners etc. 1.75 and 1.5 cm for $4^{th}$ and $3^{th}$ instars. |
| activitythresh | 0.1 cm $s^{-1}$. | Convolved speed below which the locust is considered to be resting. |
| walkthresh | 0.3 cm $s^{-1}$. | Convolved speed above which movement on the floor is considered walking. |
| noisefilter | 10, 5 | Noise filter sliding window of 10 frames, within which 5 unique pixels have to be visited. On top of this, every bout of activity has to visit at least 5 unique pixels, also setting its minimal duration to 5 frames (0.4 seconds). |
| jumpthresh_speed | 25 cm $s^{-1}$. | Minimal unconvolved speed for a jump. |
| jumpthresh_length | 2 cm | Minimal distance covered for a jump to be counted. |
| turnthresh | $2\pi$ rad $s^{-1}$. | Hamming filtered convolved radial velocity above which the locust is considered turning. May not really correspond to real-world radial velocity because of the Hamming filter. |
| length.velocitykernel | 7 | Number of frames over which velocity is convolved. |
| extrasmoother | True | Before starting the analysis, a single Tukey smoother is ran along the raw input x- and y-coordinates of the track, with the default value of the function *smooth* in R. |
| duration.longwalkbout | 3 s. | Minimal duration for a walk bout to be considered a long walk bout. It turns out that few solitarious locusts walk or climb for durations longer than 3 seconds, while gregarious locusts do. |
| length.hamming | 25 | Number of frames used to Hamming filter angular velocity. |
| attractionstart | 90 s. | Time to wait before starting to record attraction/repulsion-related parameters. This is important because the solitarious locusts behaved erratically at the beginning of the assay. |

#### 4.3.8.3 Hierarchical clustering

All training animals (125 gregarious, 110 solitarious) were clustered based on the observed behavioural variables. The less informative variables *kick_in* and *hole.TF* were left out, as well as all walk speed-related variables.

Centroid-based methods such as k-means are not suitable here, because they require the number of clusters to be specified in advance, and they have difficulties with clusters of unequal size. Hence, I opted for hierarchical clustering based on the euclidean distance of the z-scores of the transformed data. For the clustering algorithm (*hclust* function in R) itself, I explored complete linkage, UPGMA and WPGMA, and Ward's algorithm.

Not only the observations, but also the variables were clustered, based on their correlation matrices. Here, all observed variables were clustered with complete linkage clustering based on the euclidean distances of their Pearson correlations.

#### 4.3.8.4 Principal component analysis

Two separate principal component analyses were conducted: in a first phase, only the training data were used to calculate the components: long term gregarious and solitarious $5^{th}$ instar locusts that got observed for the first time. This should provide the best image of how the behavioural variables differ between long term phases, and how they correlate. As a matter of extrapolation, the set of control animals and 81 animals that were crowded for 2, 4 and 24 hours, were then projected in the above defined rotation.

In a second phase, the same analyses were repeated on the set containing all 5th instar locusts (586 animals). Because of the presence of many intermediates, the latent variable "phase" is expected to be less pronounced, which may facilitate the interpretation of orthogonal components.

The analyses described above were first executed on the set of 45 variables, and after a deeper look into the correlation structure, it was repeated on a subset of 37 variables, excluding variables that showed strong off-diagonal correlations in the ordered correlation matrix (see results section for hierarchical clustering of variables based on correlation). The following variables were excluded: climb_stim.N, climb_stim.TF, border_cont.TF, corner_cont.TF, climb_cont.N, climb_cont.TF, window_cont.TF and flatonfloor.TF. All analyses were executed both on the transformed and untransformed data.

#### 4.3.8.5 Exploratory factor analysis

Factor analysis was executed with the most commonly used estimation methods and rotations for 2, 3 and 4 factors. After this initial exploration, it was concluded that 3 or 4 factor solutions would best describe the structure of the data. In a next step, factor analyses by minimal residual and weighted least squares were executed both for 3 and 4 factor solutions, and rotated with the obligatory orthogonal *varimax* rotation and the possibly oblique *oblimin* rotation. Since cross-loading variables hampered factor estimation, the variable with the highest cross-loading was removed and EFA was repeated until no variable had a cross-correlation higher than 0.4.

### 4.3.9 Classification, model building and model validation

#### 4.3.9.1 Logistic regression

All logistic regression models were fitted using the *glm* function in R. Models were built based on observed behavioural variables, principal components and factor scores. Unfortunately, many of the "canonical" regression models were excluded, since they contained the parameter "grooming", which was not recorded in our assay, and walk speed-related parameters, which I did not want to use because these variables contain imputed values and the difference in walk speed between both phases went in the opposite direction of what had formerly been described in literature [5, 187, 217].

With so many variables at hand, selection of appropriate models is not straightforward. Variable selection should always be done in a statistically correct way, while never loosing touch with the biological relevance of the model. In this process, I have combined classical model building methods (stepwise procedures, goodness-of-fit (GOF) testing) with my insights based on previous multivariate analysis. The first statistics that were taken into account were information criteria, which provide a relative estimate of how much information would be lost if the presumed data-generating process had to reproduce the data. The less information would be lost, the better the model explains the data. The two most commonly used criteria are the Aikaike information criterion (AIC) and the Bayesian information criterion (BIC), which only differ in the way they penalize the addition of parameters. Information can be considered generalizations of likelihoods that allow for comparison of non-nested models. Hence, just like the commonly used likelihood ratio test, information criteria only give a relative measure of fit quality between competing models. A more absolute estimate of goodness-of-fit can be provided by a test based on the division of the data into subsets within which the predicted number of outcomes is compared to the observed number of outcomes. Since expected numbers of outcomes follow known distributions, a test statistic can be used for formal hypothesis testing, where

typically the null hypothesis is a good fit. I decided to work with the *le Cessie-van Houwelingen-Copas-Hosmer unweighted sum of squares test* for global goodness-of-fit, which uses smoothing techniques to overcome the inconsistencies caused by dividing the data into an arbitrary number of bins like in the better known, but heavily debated Hosmer-Lemeshow test statistic [121].

As a first exploration, regression models were built for all single variables. Next, the performance of variables as model parameters was assessed in a brute-force random permutation of 4 to 6 parameter models without interactions. In a parallel control for over-fitting, the same models were fitted with one of the model parameters randomized.

A selection of 27 best performing variables was made based on their occurrence in the top 1% models (with the lowest AIC), and earlier observation from multivariate analysis. These variables were used in forward and backward stepwise model selection procedures (both based on AIC and BIC), without modelling interactions (R function *step*).

A first crude selection of 16 models was made, primarily based on the outcome of the stepwise procedures. Also re-classification rate, parsimony and goodness-of-fit were taken into account. Only models with a good balance between activity- and attraction/repulsion-related variables were considered.

#### 4.3.9.2 Random forests

The *randomForest* function of the *randomForest* package in R has two main modes: classification of categorical variables, or regression through a gradient. In both strategies, a forest of classification trees is grown, but the decision rules are different. In classification mode, the splits are based on the maximization of homogeneity of the subsets, while in regression mode, the split is based on the minimization of the sum of squared errors between regressed and actual values. In R, some defaults (with which I worked) are also different between both modes. In regression trees, one third (13 in this case) of the variables are sampled at each split, while in classification, the square root (6 in this case) is used. Also, classification trees are typically grown to completion, while in regression, the tree is grown until all terminal nodes contain less than 6 cases. Since the model was trained on two discrete classes, binary classification may seem to be the method of choice at first sight. However, since the model will be used to derive intermediate forms (based on the assumption that the intermediates are intermediate in all latent dimensions), regression trees actually provide a more valid inference framework, where the proportion of assignments to a certain class is a consistent estimator of the probability to belong to said class [171, 91, 145, 144].

Random forests were trained on all variables (containing both the mean and medians)

except for path_length.FD and speed or duration-related variables, because these were imputed. Default parameters were used both for classification and regression trees, except for the forest size (5,000 trees) and the class weights (equal priors).

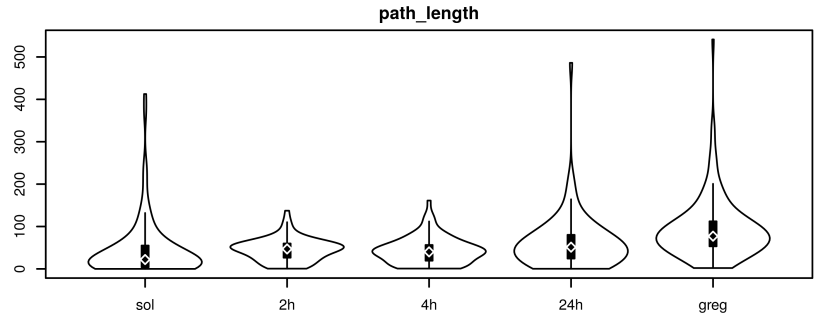
#### 4.3.9.3 Model validation

All logistic regression models were validated based on information criterions and GOF tests. Re-classification rate was calculated for all regression models based on predicted values for the training data, alongside out-of-the-bag (OOB) estimates for the random forest models. Additionally, model performance was tested on a small test set of 10 gregarious and 26 solitarious animals that had been used as controls in the experiments by Rut Vleugels on biogenic amine receptors. All these animals had been injected with dsRNA of GFP (green fluorescent protein) but did not undergo any further treatment. As a further measure for model performance, receiver operator characteristic (ROC) curves were plotted for all re-classifications and OOB estimates, and their area under the curve was calculated. The ROC curve is a plot of the true positive rate (TPR, sensitivity) versus the false positive rate (FPR, a measure of specificity) at all thresholds of a classifier. A curve with a very steep slope towards high TPR is characteristic for a performant classifier, and hence the higher the area under the curve (AUC), the better the model.

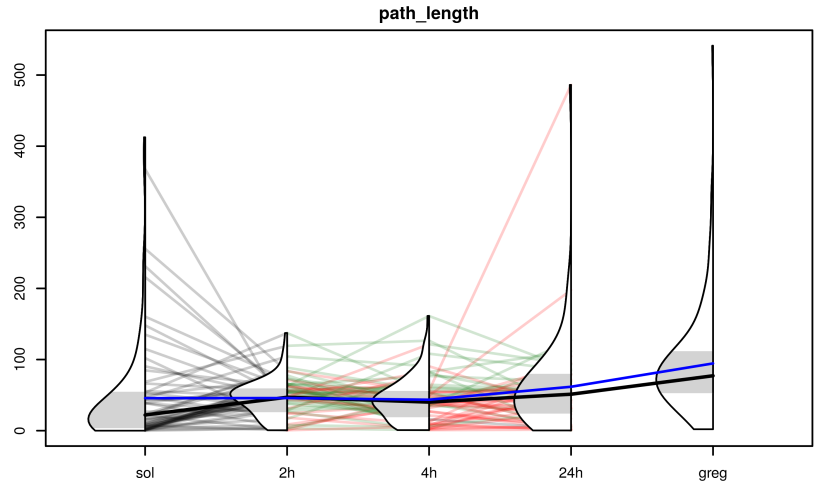
### 4.3.10 Comparison between time points

ANOVA with Tukey HSD post hoc tests and MWU tests (Benjamini-Hochberg corrected) were done between time points (crowding durations) for predicted values by each model or EFA. The observations that served for model training were excluded from this part of the analysis. For detailed statistics, see supporting materials. Also, time series were plotted (this time including training animals, since some of them were part of the crowding experiment), and analyzed with paired t-tests, Wilcoxon tests and linear models (supporting materials).

## Violin plots




Violin plots are comparable to boxplots, but apart from a measure for median and inter-quantile range (the small boxplots), they also show the kernel density estimate of the data, which is a non-parametric method for the approximation of a density distribution. Depending on the kernel sizes used, the estimation yields more ragged or smooth curves. The *vioplot* function in R has an inbuilt optimization for finding an optimal smoothing level for the visualization.




My adaptation of violin plots aims to summarize as much information as possible in one plot, hopefully without becoming too cluttered. I adapted the *vioplot* function such that the violins are filled according to the quartiles (here with the interquartile range coloured in grey, which is the equivalent of the box of a box plot). Further more, only half violins are plotted, which leaves one side open to show how individual locusts change their behaviour between time points. Red lines mean that in the previous interval the value dropped, green means it increased, and grey means it was constant, or there was no previous interval. The solid blue line connects the sample means and the black line connects the sample medians.

# 4.4 Results

## 4.4.1 Evaluation of the new pipeline

### 4.4.1.1 Performance of the tracker

The most important parameter to optimize was the number of standard deviations (sigma) for the Gaussian smoother applied on the zoomed-in image of the locust. After tracking a random selection of 58 videos with smoothers ranging from 2 to 5 standard deviations, it was concluded that the most accurate tracks were obtained with a sigma value of 4 (fig. 4.7). Lower values tended to have increased jitter around the centroid, while higher smoother values showed the tendency to pull the centroid towards edges of walls and windows.

A total of 1,112 videos of 9 to 10 min. were tracked, taking between 7 and 12 min. per video, depending on the efficiency of zooming in during phase 1 (fig. 4.6). Among 1,112 tracks, about 100 tracks needed editing. In the majority of cases these were minor adjustments like shifting the start position of the track by a couple of frames. Only 19 tracks needed more substantial editing:

**9 videos:** background subtraction failed because the locust was already present in the first frame.
**5 videos:** the locust was not recognized because it was partially masked while sitting on top of the introduction hole.
**2 videos:** the track jumped between the locust and a stain on the floor.
**3 videos:** the track was disturbed by trembling of the arena.

### 4.4.1.2 Performance of tandy.R

With default parameters, the analysis of a single track takes less than a second, and upon visual inspection of the ethograms, no obvious mistakes were found. In many cases, comparable results could be obtained using different parameter combinations. Especially speed thresholds could be lowered to virtually zero if the noise filtering was set to more stringent values, and velocity was convolved with longer kernels.

Based on my observations of a set of well annotated turns, a typical turn for a gregarious locust was usually around $180^\circ$ ( 3 radians) and took between 1 and 3 seconds to complete. This corresponds to a turn speed of 1 to 3 radians per second. However, full turns taking up to 10 seconds were also observed. After convolution with a Hamming window of length 25 frames (2 seconds, equivalent to the 61 frame hamming window used in the original tandy.m, which was optimized for 30 fps

videos), the convolved angular velocity cut-off was chosen such that 5 video fragments containing representative turns were successfully annotated without leaving too much margin (fig. 4.9).

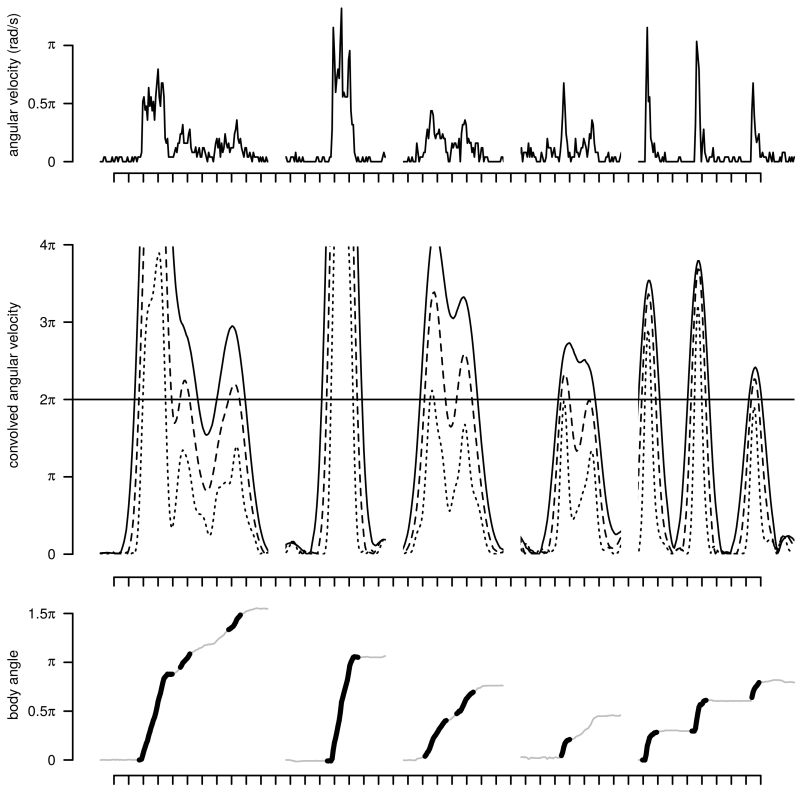



**Figure 4.9:** 5 representative fragments containing turns with decreasing turn angles, with time scale in seconds. The first case is a 270° turn with decreasing turn speed, the second is a fast half-turn, cases 3 and 4 are more subtle turns, and the last case is a half turn in 3 discrete steps. Top panel: the angular velocity in radians per second for the five turns. Note how all of them at some point have a turn speed of more than 90° per second. Middle panel: convolved angular velocity, with Hamming windows ranging from about 3 seconds (solid line), 2 seconds (dashed line) and 1 second (dotted line), and the cut-off value of 2 $\pi$ radians per second. Bottom panel: body angle relative to the start of the turn. The bold black line indicates the parts of the turn that were successfully detected using a Hamming window of 25 frames, and a convolved angular velocity cut-off of 360° per second.

### 4.4.2 A first comparison between long term phases

Apart from path_length.FD, 6 variables related to the time spent in some "unpolarized" arena sectors (flatonfloor.TF, wall.TF, border_wall.TF, border.TF, middlethird.TF and kick_in) and 2 variables related to positions away from the stimulus side (window_cont.TF, corner_cont.TF), all observed behavioural variables showed a significant difference between long term phases, both for the raw data (BH adjusted $p < 0.01$, Mann Whitney U test) and the transformed data (BH adjusted $p < 0.01$, Welch t-test). The observed differences were all in line with what is known from literature with the notable exception of all variables related to walk speed, where the highest speeds were observed in solitarious locusts. A selection of observed variables is shown in figures 4.10 and 4.11. All data can be found in the supporting materials.

Based on re-classification of training data, some univariate logistic regression models showed relatively good accuracy, especially in the transformed variables. For example, more than 82 % of locusts from both phases got classified correctly based solely on the time fraction spent resting or climbing. The attraction/repulsion variables did not perform this well, and their classification accuracy was often asymmetric because solitarious animals quite consistently avoid the stimulus side, but gregarious ones did not necessarily avoid the control side. This asymmetry was also reflected in the boxplots shown in figure 4.11: if there was a genuine difference between both groups, almost invariably, one had a wide distribution and the other one was skewed, even after variance stabilizing transformation. But still, figures 4.11 and 4.12 leave little doubt about a strong difference in attraction/repulsion between the two long-term phases.

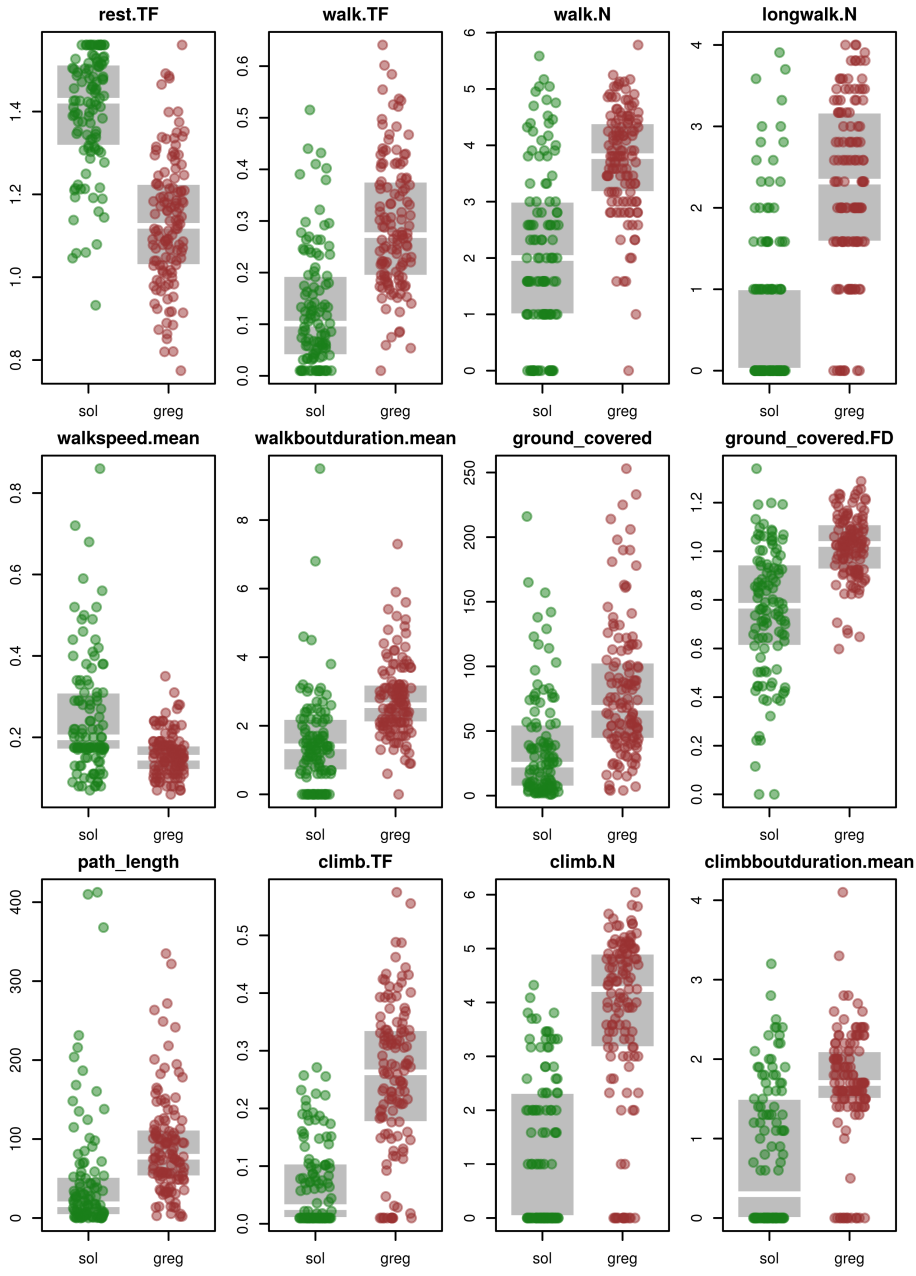



**Figure 4.10:** Activity-related variables, compared between long term phases. Time fractions (TF) are arcsine transformed, counts (N) are log2 transformed.

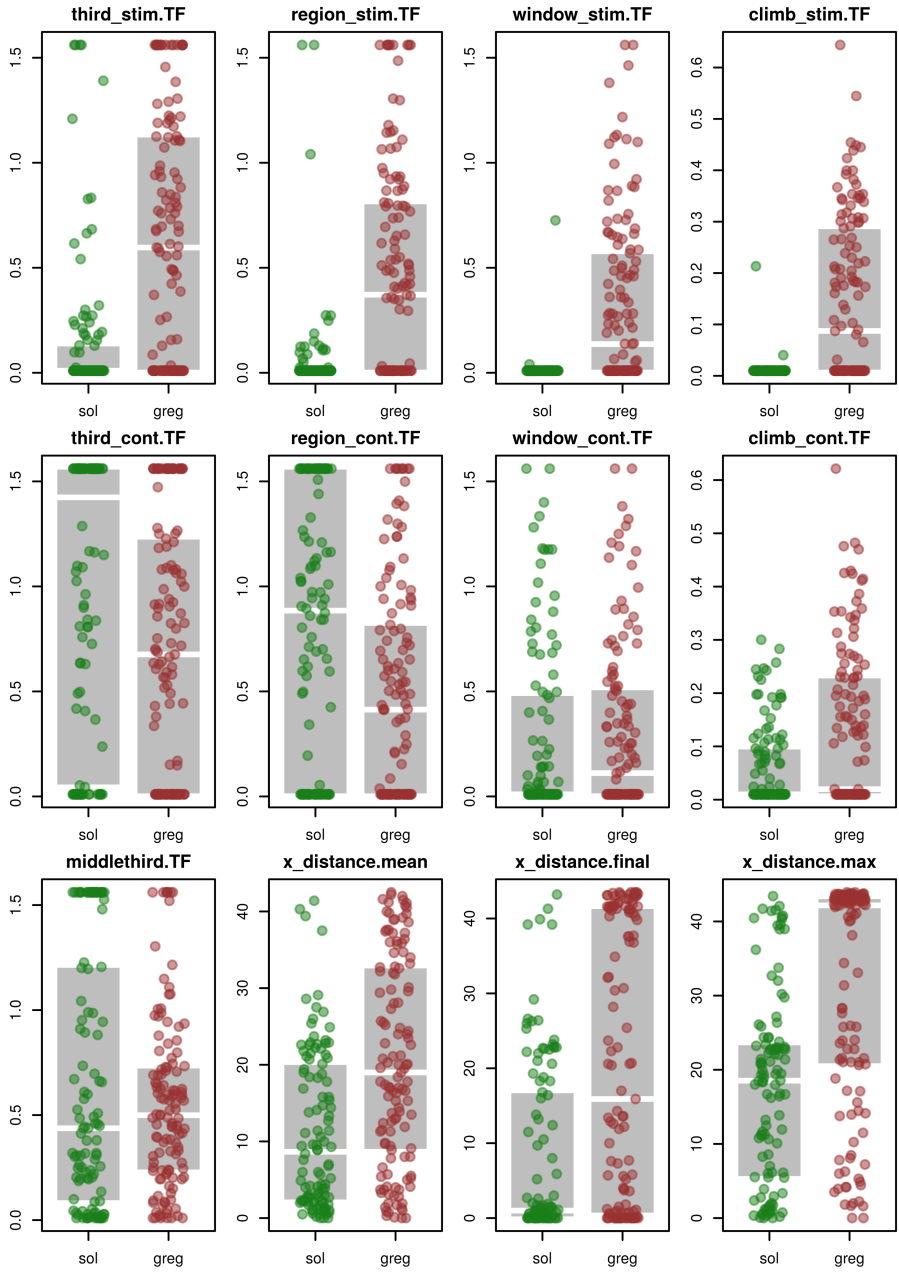



**Figure 4.11:** Attraction/repulsion-related variables, compared between long term phases. Time fractions (TF) are arcsine transformed.

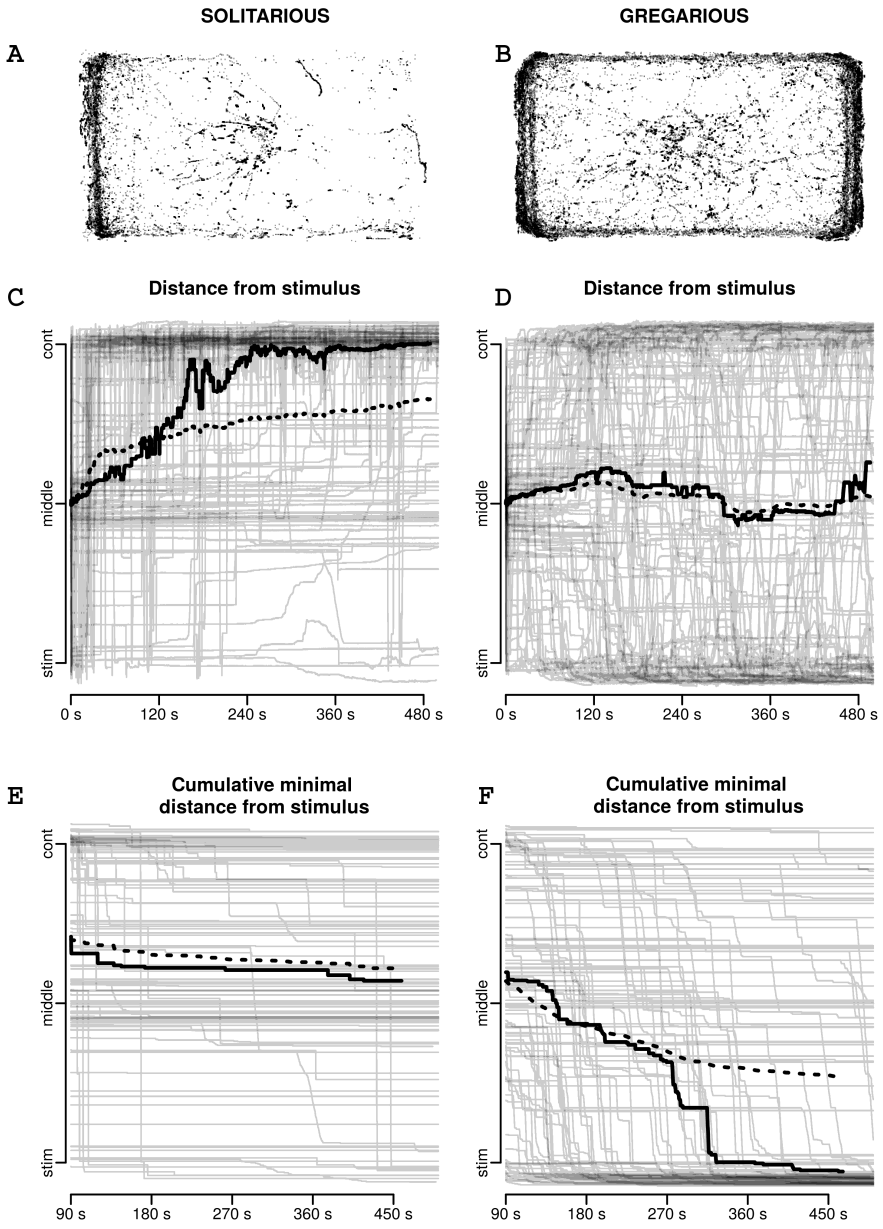



**Figure 4.12:** Upper row: scatter plots of the positions of all solitarious (A) and gregarious (B) training locusts in the arena. Stimulus group on the right hand side. Middle and bottom row: distance from the stimulus group for all extreme phase locusts (125 gregarious, 110 solitarious). Solid line indicates the median position, dotted line indicates mean position. Middle row (C, D): distance from the stimulus group over the length of the entire recording (500 seconds). Note how solitarious locusts need some time before settling at a large distance away from the stimulus group. Bottom row (E, F): cumulative minimal distance from stimulus group. The first 90 seconds are ignored in order to skip the erratic period observed in (C). Note that the majority of gregarious animals approach the stimulus group closely at some point, while this is only the case for a small fraction of solitarious locusts.

**Table 4.5:** Classification performance for untransformed variables. P values are for Mann-Whitney U test and a Welch t-test between 125 gregarious and 110 solitarious animals, and for the coefficient in a univariate linear regression, for which also the z-score, standard error, AIC and coefficient estimate are shown. The last two columns show the classification rate by the regression model based on predicted values for the training data.

| variable | p MWU | p t-test | p regr. | Z | SE | AIC | estim. | class. sol | class. greg |
|---|---|---|---|---|---|---|---|---|---|
| rest.TF | * * * | * * * | * * * | -7.78 | 2.504 | 206 | -19.5 | 83.6 | 80.0 |
| hole.TF | 0.002 | 0.956 | 0.956 | 0.06 | 1.422 | 329 | 0.08 | 0.0 | 100.0 |
| onfloor.TF | * * * | * * * | * * * | -4.58 | 0.570 | 304 | -2.61 | 70.9 | 65.6 |
| flatonfloor.TF | 0.422 | 0.054 | 0.051 | -2.02 | 0.388 | 325 | -0.78 | 36.4 | 78.4 |
| border.TF | 0.578 | 0.082 | 0.076 | -1.83 | 0.408 | 325 | -0.75 | 40.9 | 76.0 |
| window_stim.TF | * * * | * * * | * * * | 3.43 | 5.335 | 260 | 18.3 | 99.1 | 48.0 |
| window_cont.TF | 0.422 | 0.782 | 0.780 | 0.30 | 0.458 | 329 | 0.137 | 0.0 | 100.0 |
| corner_stim.TF | * * * | 0.010 | 0.01 | 2.64 | 2.902 | 318 | 7.67 | 95.5 | 46.4 |
| corner_cont.TF | 0.201 | * * * | * * * | -3.45 | 0.978 | 313 | -3.38 | 31.8 | 83.2 |
| wall.TF | * * * | 0.256 | 0.264 | 1.15 | 1.740 | 327 | 2.01 | 0.0 | 100.0 |
| border_stim.TF | * * * | * * * | 0.002 | 3.12 | 1.538 | 311 | 4.80 | 93.6 | 48.0 |
| border_cont.TF | * * * | * * * | * * * | -5.14 | 0.499 | 296 | -2.57 | 48.2 | 83.2 |
| border_wall.TF | * * * | 0.046 | 0.05 | 2.03 | 0.541 | 324 | 1.101 | 64.5 | 60.8 |
| third_stim.TF | * * * | * * * | * * * | 5.79 | 0.535 | 279 | 3.096 | 89.1 | 57.6 |
| third_cont.TF | * * * | 0.002 | 0.002 | -3.18 | 0.315 | 318 | -1.00 | 58.2 | 64.0 |
| half_stim.TF | * * * | * * * | * * * | 3.81 | 0.330 | 313 | 1.257 | 67.3 | 64.8 |
| middlethird.TF | 0.882 | 0.138 | 0.13 | -1.56 | 0.371 | 326 | -0.58 | 34.5 | 80.8 |
| walk.TF | * * * | * * * | * * * | 6.18 | 3.673 | 261 | 22.7 | 78.2 | 67.2 |
| fastwalk.TF | * * * | * * * | * * * | 3.53 | 6.477 | 312 | 22.9 | 74.5 | 64.8 |
| climb.TF | * * * | * * * | * * * | 7.10 | 7.693 | 200 | 54.64 | 85.5 | 75.2 |
| climb_stim.TF | * * * | * * * | 0.002 | 3.22 | 37.03 | 243 | 119.2 | 99.1 | 51.2 |
| climb_cont.TF | 0.003 | * * * | * * * | 3.91 | 5.304 | 303 | 20.72 | 77.3 | 46.4 |
| region_stim.TF | * * * | * * * | * * * | 4.78 | 1.143 | 272 | 5.459 | 97.3 | 56.0 |
| region_cont.TF | * * * | * * * | * * * | -4.04 | 0.343 | 312 | -1.39 | 54.5 | 72.0 |
| jump.N | * * * | * * * | * * * | -3.54 | 0.093 | 287 | -0.33 | 32.7 | 96.0 |
| walk.N | * * * | * * * | * * * | 5.89 | 0.019 | 281 | 0.111 | 78.2 | 69.6 |
| border.N | * * * | * * * | * * * | 3.64 | 0.021 | 313 | 0.077 | 72.7 | 64.0 |
| climb.N | * * * | * * * | * * * | 7.02 | 0.031 | 198 | 0.215 | 83.6 | 76.0 |
| climb_stim.N | * * * | * * * | * * * | 3.39 | 0.146 | 250 | 0.497 | 99.1 | 51.2 |
| climb_cont.N | 0.002 | * * * | * * * | 3.92 | 0.027 | 302 | 0.105 | 72.7 | 44.0 |
| longwalk.N | * * * | * * * | * * * | 6.25 | 0.066 | 259 | 0.410 | 83.6 | 68.0 |
| longclimb.N | * * * | * * * | * * * | 6.06 | 0.143 | 246 | 0.868 | 79.1 | 72.0 |
| turn.N | * * * | * * * | * * * | 4.78 | 0.330 | 285 | 1.576 | 91.8 | 47.2 |
| x_distance.final | * * * | * * * | * * * | 5.32 | 0.009 | 295 | 0.049 | 69.1 | 56.8 |
| x_distance.mean | * * * | * * * | * * * | 5.10 | 0.012 | 299 | 0.061 | 61.8 | 64.8 |
| walkspeed.mean | * * * | * * * | * * * | -5.17 | 2.013 | 286 | -10.4 | 45.5 | 84.0 |
| walkboutspeed.mean | * * * | * * * | * * * | -4.82 | 0.287 | 288 | -1.39 | 30.9 | 81.6 |
| walkboutdur.mean | * * * | * * * | * * * | 6.31 | 0.152 | 270 | 0.96 | 71.8 | 77.6 |
| climbboutdur.mean | * * * | * * * | * * * | 6.74 | 0.181 | 271 | 1.22 | 63.6 | 86.4 |
| x_distance.median | * * * | * * * | * * * | 4.67 | 0.010 | 304 | 0.047 | 58.2 | 64.0 |
| walkspeed.median | * * * | * * * | * * * | -4.81 | 2.199 | 292 | -10.58 | 45.5 | 84.0 |
| walkboutspeed.median | * * * | * * * | * * * | -4.26 | 0.277 | 298 | -1.182 | 29.1 | 80.0 |
| walkboutdur.median | * * * | * * * | * * * | 5.35 | 0.182 | 292 | 0.974 | 70.0 | 72.0 |
| climbboutdur.median | * * * | * * * | * * * | 6.16 | 0.211 | 282 | 1.3 | 59.1 | 88.0 |
| kick_in | 0.192 | 0.616 | 0.611 | -0.54 | 0.004 | 328 | -0.002 | 0.9 | 98.4 |
| ground_covered | * * * | * * * | * * * | 5.64 | 0.004 | 283 | 0.021 | 72.7 | 67.2 |
| ground_covered.FD | * * * | * * * | * * * | 6.91 | 1.031 | 243 | 7.123 | 67.3 | 85.6 |
| path_length | * * * | * * * | * * * | 4.30 | 0.003 | 304 | 0.011 | 76.4 | 73.6 |
| path_length.FD | 0.028 | 0.138 | 0.13 | 1.56 | 1.875 | 326 | 2.92 | 40.0 | 80.8 |
| x_distance.max | * * * | * * * | * * * | 6.35 | 0.010 | 281 | 0.063 | 65.5 | 69.6 |

**Table 4.6:** Classification performance for transformed variables. P values are for Mann-Whitney U test and a Welch t-test between 125 gregarious and 110 solitarious animals, and for the coefficient in a univariate linear regression, for which also the z-score, standard error, AIC and coefficient estimate are shown. The last two columns show the re-classification rate by the regression model based on predicted values for the training data.

| variable | p MWU | p t-test | p regr. | Z | SE | AIC | estim. | class. sol | class. greg |
|---|---|---|---|---|---|---|---|---|---|
| rest.TF | * * * | * * * | * * * | -8.25 | 1.307 | 197 | -10.8 | 82.7 | 82.4 |
| hole.TF | 0.002 | 0.568 | 0.565 | 0.59 | 0.915 | 328 | 0.542 | 0.0 | 100.0 |
| onfloor.TF | * * * | * * * | * * * | -5.35 | 0.416 | 295 | -2.23 | 67.3 | 73.6 |
| flatonfloor.TF | 0.422 | 0.045 | 0.041 | -2.09 | 0.283 | 324 | -0.59 | 33.6 | 82.4 |
| border.TF | 0.577 | 0.280 | 0.267 | -1.15 | 0.304 | 327 | -0.35 | 31.8 | 88.0 |
| window_stim.TF | * * * | * * * | * * * | 3.97 | 2.648 | 246 | 10.52 | 99.1 | 52.0 |
| window_cont.TF | 0.422 | 0.560 | 0.557 | 0.62 | 0.318 | 328 | 0.197 | 0.0 | 100.0 |
| corner_stim.TF | * * * | * * * | * * * | 4.29 | 1.544 | 298 | 6.619 | 94.5 | 48.0 |
| corner_cont.TF | 0.201 | 0.004 | 0.004 | -2.95 | 0.515 | 319 | -1.52 | 37.3 | 77.6 |
| wall.TF | * * * | 0.012 | 0.015 | 2.48 | 0.964 | 322 | 2.393 | 73.6 | 53.6 |
| border_stim.TF | * * * | * * * | * * * | 4.47 | 0.928 | 295 | 4.150 | 86.4 | 54.4 |
| border_cont.TF | * * * | * * * | * * * | -4.90 | 0.330 | 300 | -1.62 | 52.7 | 79.2 |
| border_wall.TF | * * * | 0.013 | 0.015 | 2.50 | 0.374 | 322 | 0.934 | 61.8 | 63.2 |
| third_stim.TF | * * * | * * * | * * * | 5.86 | 0.351 | 279 | 2.058 | 87.3 | 59.2 |
| third_cont.TF | * * * | 0.001 | 0.001 | -3.29 | 0.211 | 318 | -0.70 | 56.4 | 66.4 |
| half_stim.TF | * * * | * * * | * * * | 3.75 | 0.223 | 314 | 0.835 | 65.5 | 68.0 |
| middlethird.TF | 0.882 | 0.186 | 0.175 | -1.40 | 0.261 | 327 | -0.37 | 31.8 | 84.0 |
| walk.TF | * * * | * * * | * * * | 7.32 | 1.519 | 242 | 11.11 | 74.5 | 76.8 |
| fastwalk.TF | * * * | * * * | * * * | 5.14 | 1.925 | 295 | 9.897 | 68.2 | 75.2 |
| climb.TF | * * * | * * * | * * * | 8.18 | 1.769 | 208 | 14.47 | 80.0 | 80.8 |
| climb_stim.TF | * * * | * * * | * * * | 3.80 | 6.509 | 237 | 24.75 | 98.2 | 52.0 |
| climb_cont.TF | 0.003 | * * * | * * * | 4.01 | 1.290 | 309 | 5.180 | 63.6 | 49.6 |
| region_stim.TF | * * * | * * * | * * * | 5.45 | 0.656 | 267 | 3.571 | 93.6 | 56.0 |
| region_cont.TF | * * * | * * * | * * * | -3.90 | 0.234 | 313 | -0.91 | 54.5 | 72.0 |
| jump.N | * * * | * * * | * * * | -4.75 | 0.138 | 297 | -0.65 | 42.7 | 83.2 |
| walk.N | * * * | * * * | * * * | 7.23 | 0.134 | 250 | 0.965 | 71.8 | 82.4 |
| border.N | * * * | * * * | * * * | 5.36 | 0.094 | 296 | 0.506 | 64.5 | 74.4 |
| climb.N | * * * | * * * | * * * | 8.41 | 0.102 | 218 | 0.861 | 78.2 | 82.4 |
| climb_stim.N | * * * | * * * | * * * | 4.04 | 0.360 | 241 | 1.456 | 99.1 | 51.2 |
| climb_cont.N | 0.002 | * * * | * * * | 3.72 | 0.083 | 314 | 0.309 | 61.8 | 48.8 |
| longwalk.N | * * * | * * * | * * * | 7.89 | 0.154 | 235 | 1.214 | 75.5 | 82.4 |
| longclimb.N | * * * | * * * | * * * | 7.19 | 0.189 | 246 | 1.359 | 79.1 | 72.0 |
| turn.N | * * * | * * * | * * * | 5.42 | 0.342 | 282 | 1.854 | 91.8 | 47.2 |
| x_distance.final | * * * | * * * | * * * | 5.32 | 0.009 | 295 | 0.049 | 69.1 | 56.8 |
| x_distance.mean | * * * | * * * | * * * | 5.10 | 0.012 | 299 | 0.061 | 61.8 | 64.8 |
| walkspeed.mean | * * * | * * * | * * * | -5.17 | 2.013 | 286 | -10.4 | 45.5 | 84.0 |
| walkboutspeed.mean | * * * | * * * | * * * | -4.82 | 0.287 | 288 | -1.39 | 30.9 | 81.6 |
| walkboutdur.mean | * * * | * * * | * * * | 6.31 | 0.152 | 270 | 0.956 | 71.8 | 77.6 |
| climbboutdur.mean | * * * | * * * | * * * | 6.74 | 0.181 | 271 | 1.219 | 63.6 | 86.4 |
| x_distance.median | * * * | * * * | * * * | 4.67 | 0.010 | 304 | 0.047 | 58.2 | 64.0 |
| walkspeed.median | * * * | * * * | * * * | -4.81 | 2.199 | 292 | -10.6 | 45.5 | 84.0 |
| walkboutspeed.median | * * * | * * * | * * * | -4.26 | 0.277 | 298 | -1.18 | 29.1 | 80.0 |
| walkboutdur.median | * * * | * * * | * * * | 5.35 | 0.182 | 292 | 0.974 | 70.0 | 72.0 |
| climbboutdur.median | * * * | * * * | * * * | 6.16 | 0.211 | 282 | 1.299 | 59.1 | 88.0 |
| kick_in | 0.192 | 0.591 | 0.587 | -0.54 | 0.004 | 328 | -0.002 | 0.9 | 98.4 |
| ground_covered | * * * | * * * | * * * | 5.64 | 0.004 | 283 | 0.021 | 72.7 | 67.2 |
| ground_covered.FD | * * * | * * * | * * * | 6.91 | 1.031 | 243 | 7.123 | 67.3 | 85.6 |
| path_length | * * * | * * * | * * * | 4.30 | 0.003 | 304 | 0.011 | 76.4 | 73.6 |
| path_length.FD | 0.028 | 0.140 | 0.133 | 1.56 | 1.875 | 326 | 2.918 | 40.0 | 80.8 |
| x_distance.max | * * * | * * * | * * * | 6.35 | 0.010 | 281 | 0.063 | 65.5 | 69.6 |

### 4.4.3 Comparison between manual and automated assay

Only a subset of observed behaviours had definitions similar enough to be directly comparable between the old and the new arena set-up, while others may be qualitatively equivalent at most. Since nothing changed in the gregarious colony between both series of observations, differences in values for gregarious animals point to differences between the assays. Differences for solitarious animals can either be attributed to differences in rearing conditions, or differences between the assays. In general, the differences between both phases were more clear-cut (and statistically more significant) in the newer assay.

The observed time fraction spent resting was a lot higher in the new assay, both for gregarious and solitarious locusts. I think this is caused by the conservative segmentation of the track into passive versus active. Consequently, the time fraction spent walking and especially climbing was now lower. The latter also owing to the fact that the animals spent more time at the windows in the old assay, both climbing and perching, while time spent on walls was low in both settings.

The number of walk bouts and jumps were about equal between the old and new assay in gregarious animals. There were slight differences between solitarious locusts: especially the number of jumps was significantly higher in the new assay. In the old assay, more turns were counted for both phases. The automated assay shows high counts for climb bouts, which in combination with the lower time fraction, may suggest that the automated assay splits climbs that were lumped in the manual assay.

Other notable differences could be observed at the windows, where in the old assay, a high proportion of solitarious locusts passed part of their time at the windows, with only a modest difference between stimulus and non-stimulus side, while the gregarious animals showed a strong preference for the stimulus window. However, in the new assay, only 2 out of 110 solitarious animals paid a visit to the stimulus window[6], while 50 (about half) reached the control window at least once. Gregarious locusts on the other hand, failed to show a pronounced preference for the stimulus window compared to the control window. The control window was visited in equal frequencies by both phases.

[6]The first 90 seconds were left out only for the automated assay, but the figures do not change starting at time point zero.

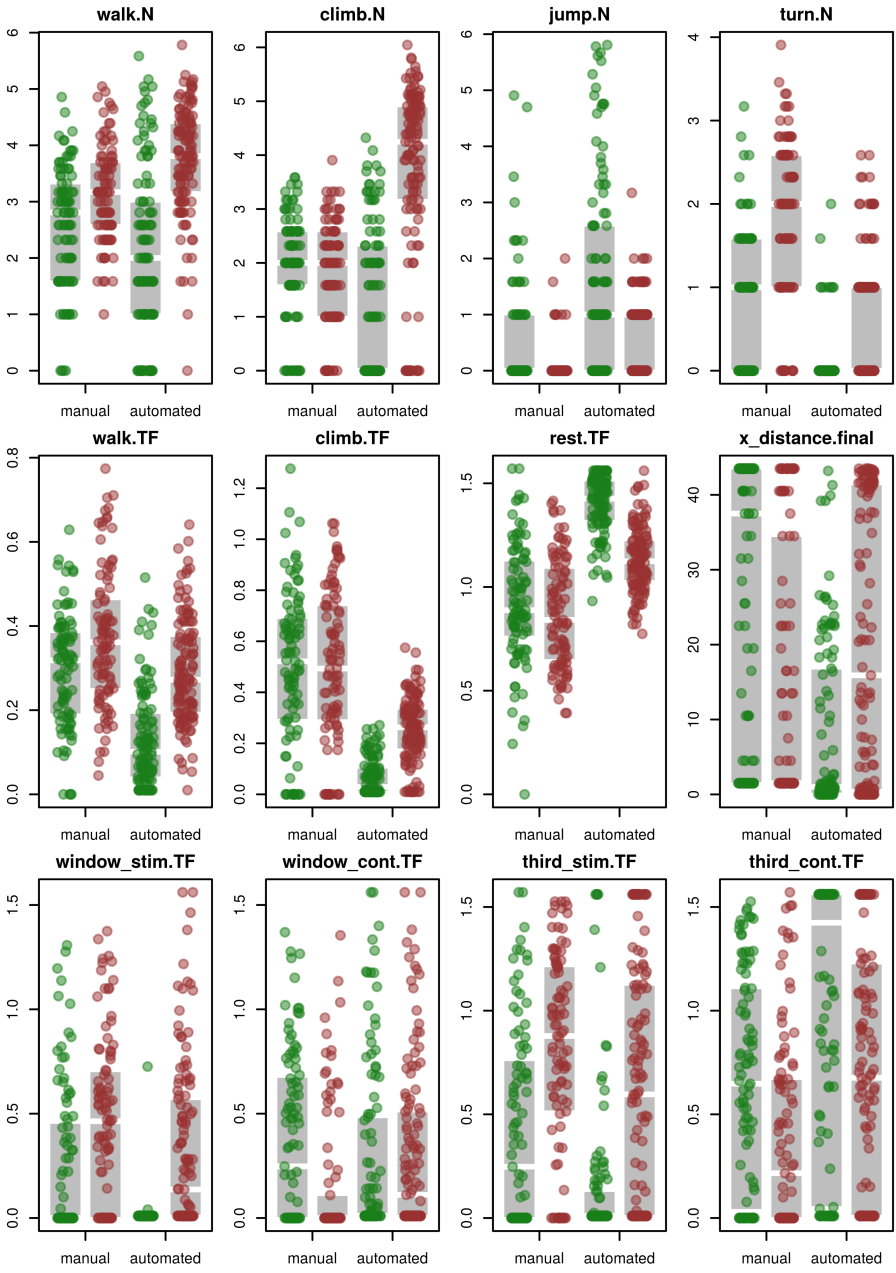



**Figure 4.13:** Variables that were more or less directly comparable between the old and the new set-up. Counts (N) are log2 transformed and time fractions (TF) are arcsine transformed. Green: long term solitarious, red: long term gregarious.

### 4.4.4 Hierarchical clustering

Most clustering methods yielded at least some substantial clusters with a predominant phase. However, in most cases the interpretation was difficult. Often two dominant clusters were formed, more or less containing all "obviously" gregarious or solitarious observations, alongside a single intermediate cluster within which no pure subclusters could be found. In other cases, just two sub-clusters of moderate purity were formed. For example, the WPGMA clustering based on the euclidean distances of the ranks resulted in two sub-clusters that were 82.4% and 84.5% pure. However, because of the homogeneous branching depth, no clear sub-clusters could be identified, and the impurities were scattered over the entire dendrogram.

The most informative clustering was achieved with complete linkage (farthest neighbour joining) (fig. 4.13). For the training data, this clustering revealed a structure with 3 main clusters. Cluster A contained only one observation: a single solitarious locust (sol_AB_004_model_1). This animal held the record of sitting on the wall and sitting in the corner near the stimulus group, which given the very skewed distribution of these variables, resulted in extreme z-scores and hence in a large multivariate distance from any other datapoint. This individual was also one of the only solitarious animals in the training set that spent a significant amount of their time at the stimulus side. On the other hand, it displayed characteristics of solitariousness, such as a high number of jumps. This paradoxical combination of behaviours made it the most deviant animal of the set.

The two other clusters divided more or less on activity, with cluster B generally containing animals that barely walked or climbed, and hence had a high rest time fraction, and cluster C with observations of more active individuals. Since hierarchical clustering is based on multivariate distances, no set of simple rules can be constructed to determine the grouping into sub-clusters (unlike in decision trees). For example, there was overlap in all activity-related variables between clusters C and B.

Cluster B contained 123 observations, among which 89 solitarious. It contained 81% of all solitarious locusts in the training data. As mentioned earlier, animals in this cluster generally had high rest time fractions, and walked and climbed little. Cluster C contained 112 observations, among which 91 gregarious. It contained 73% of all gregarious locusts in the training data.

The impossibility of pinpointing clustering structure to single variables was further illustrated in the substructures of clusters B and C, where combinations of attraction - repulsion, jumping- and climbing-related variables played a role in the subdivision into sub-clusters. The fact that multivariate distances are seldom directly readable, complicated the definition of sub-clusters, making it a bit of an arbitrary choice. Often, sub-clusters are defined based on a certain *depth* at which the dendrogram should be chopped up. However, in this case, the phase state of the members of a candidate

sub-cluster could be used as extra information. On the one hand there is little use in chopping up pure clusters, however large, while on the other hand it may be interesting to isolate smaller clusters based on their purity, or the lack thereof. With this in mind, I think there were about a dozen relevant sub-clusters (fig. 4.9), but other interpretations are certainly possible.

Some sub-clusters were almost pure, consisting mostly of animals of the same phase. These were the animals that behaved in the most stereotypic way. For example, sub-cluster C2 contained all active locusts that paid a visit to the stimulus side, and it contained 97% gregarious locusts. Sub-cluster B1b.2 contained passive animals that stayed on the control side and did not climb a lot, and as a result contained 90% solitarious animals. An interesting sub-cluster was C1b: a purely solitarious cluster in the middle of the "active" animals. These were the rather active animals that favoured the control side and jumped a lot.

Finally, some sub-clusters were mixed, containing about equal proportions of both phases. These were the clusters with animals that showed more ambiguous behaviours: either they showed intermediate levels of all behaviours, or they combined typical gregarious and solitarious behaviours, such as cluster B2b that contained passive animals that tended to spend some time on the stimulus side.

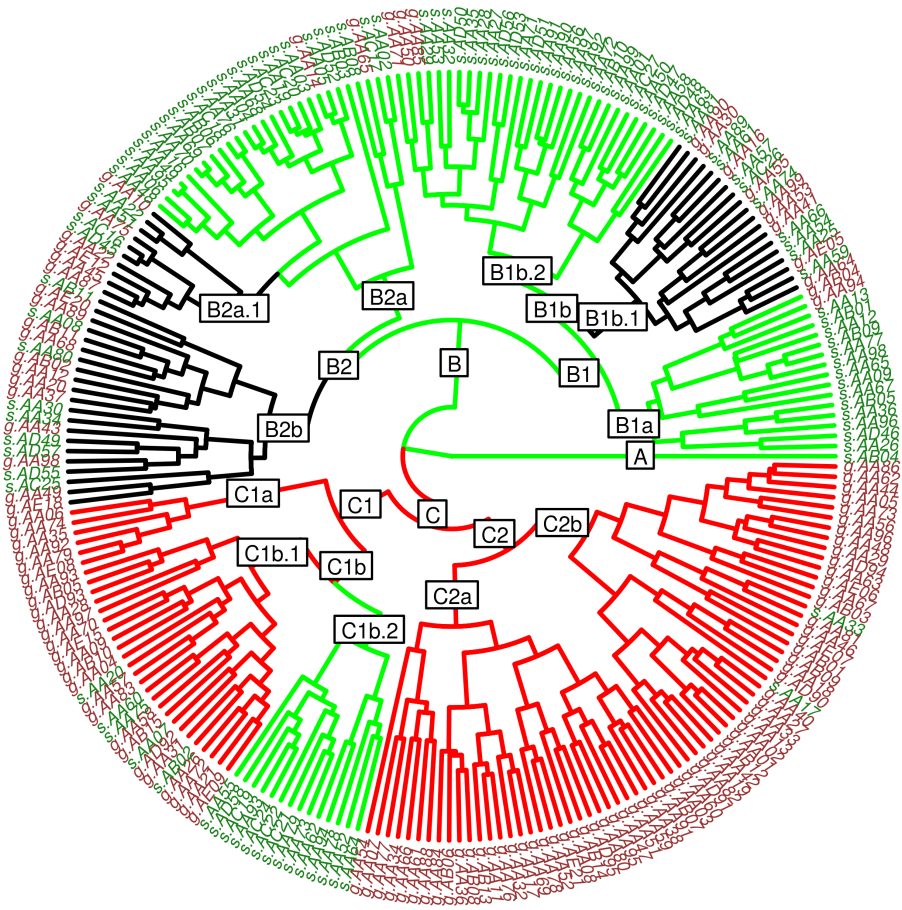



**Figure 4.14:** Dendrogram of hierarchical (unsupervised) clustering based on euclidean distances of z-scores of behaviours, clustered based on complete linkage. Edge colors represent the predominant phase within (sub)clusters. Red stands for predominantly (> 80%) gregarious, green for solitarious, and black stands for ambiguous.

## 4.4.5 Correlations between observed variables

### 4.4.5.1 The "automated training" data

Within the training set, most variables showed strong correlations with at least some other variables. The heat map in figure 4.15 was based on the the absolute value of the Pearson correlation matrix of the transformed training data. The variables were clustered (euclidean distance, complete linkage) according to their correlations, resulting in blocks of similar and correlated variables along the diagonal of the heat map. Clearly, variables related to activity were strongly correlated, generally splitting up in climbing- and walking-related variables (lower left block), while on the other hand, variables related to attraction/repulsion show strong correlations as well. If the hierarchical clustering of the variables has successfully captured the correlational structure of the different variables, off-diagonal blocks can be an important sign for variables that cross-load on multiple latent variables. Obviously, the off-diagonal elements had a certain degree of correlation, which was expected given the underlying latent variable *phase*. But, some of the "compound" variables, such as climbing on the stimulus window, showed high correlation with both climbing-related variables and attraction/repulsion-related variables. These strong off-diagonal elements may become problematic in PCA or EFA, where they may obfuscate the interpretability of the components and factors.

### 4.4.5.2 The "automated total" data

Both the correlation matrix and the variable clustering of the total data (586 observations containing training data and crowding time series) looked similar to the training data, albeit generally with slightly weaker correlations (see supporting materials).

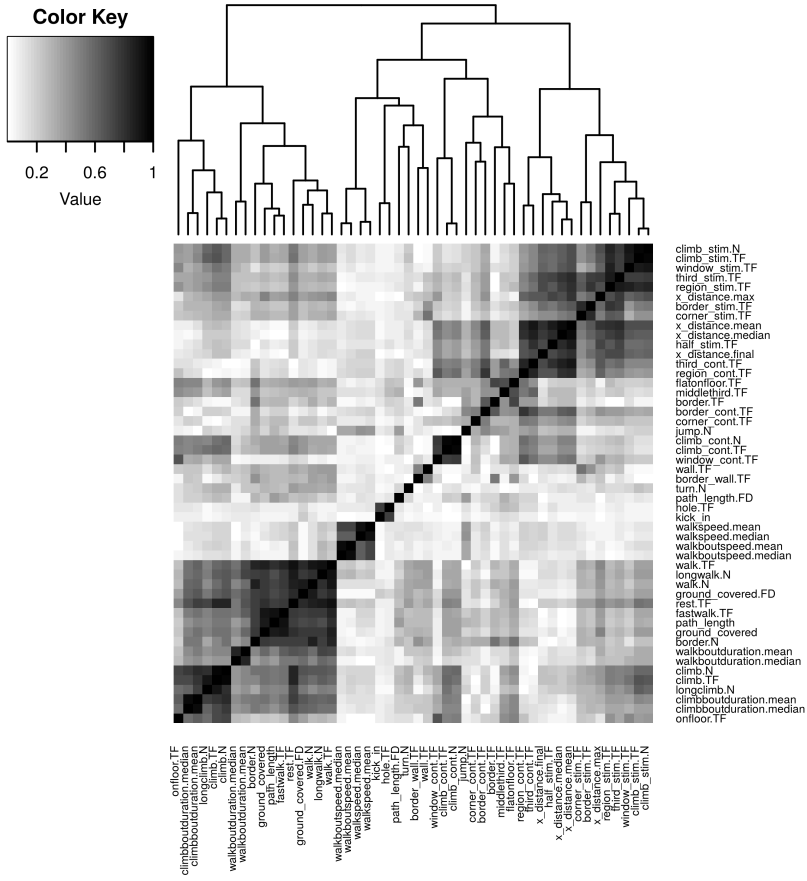



**Figure 4.15:** Heat map of the absolute values of the correlation matrix of variables in the training set. Note that all variables have a Pearson correlation of 1 with themselves (diagonal elements of the matrix). The dendrogram is a hierarchical clustering of the variables based on the correlations. Blocks of strongly correlated elements on the diagonal point to clusters with strong correlations, while off-diagonal blocks point to variables that crosslink the different dimensions within the data.

## 4.4.6 Principal component analysis

### 4.4.6.1 PCA on the training data

The two first principal components captured about 50% of the variance, with the transformed variables performing slightly better than the untransformed ones. After removal of the potentially cross-loading variables, the scree plots (fig. 4.16) remained virtually the same apart from the first principal component becoming slightly more dominant. The scree plots showed a clear break after two principal components (fig. 4.16). The first component loads highest on activity-related variables, while the second is more correlated to variables related to attraction/repulsion. However, although the highest loading variables of these components were virtually orthogonal in this space, they were rotated with respect to the principal components and hence the components did not seem to correspond directly to latent variables. The effect of omitting potentially cross-loading variables was of minor influence, only causing a slight rotation of the high-loading variables towards the principal components (fig. 4.16). Within the space of these first two components, the solitarious and gregarious observations formed two "horns" with overlap at their bases. These horns were slightly more compact when only the selection of variables was used. In the following sections, the focus will be with the results obtained with the PCA on the selection of transformed variables, but both other analyses are comparable, and can be found in supporting materials.

Observations were not homogeneously distributed in the plane of the first pair of components (fig. 4.16). Also, there seemed to be a "forbidden" area running along the edge between the first and fourth quadrant, where moderately active animals show signs of attraction. There was also a clear-cut boundary where simultaneous low values for both first components seemed to be "forbidden". This boundary became even clearer when looking in 3 dimensions (adding PC3, fig 4.17)), and it even held true for projections of any member of the entire dataset of 1122 observations. The 3 solitarious locusts sitting on top of the gregarious "horn" (fig 4.16), were bound to be misclassified. As a matter of fact, in the hierarchical clustering, these were the only two solitarious animals from cluster C2b (sol_AA33 and sol_AA17), and sol_AA04, the single member of cluster A. Furthermore, the regions where solitarious and gregarious observations showed overlap, corresponded well to the ambiguous clusters B2b, B2a.1 and B1b.1. (see supporting materials). The remarkable sub-cluster C1b.2, which was a pure sub-cluster of solitarious animals in the middle of gregarious observations, corresponded to the basis of the solitarious "horn" (fig 4.16). As described in the previous section, these were indeed animals that were rather active, yet showed strong signs of repulsion.

Long term behavioural phases separated quite well in the space defined by the first 4 principal components, and also the hierarchical clusters described in the previous

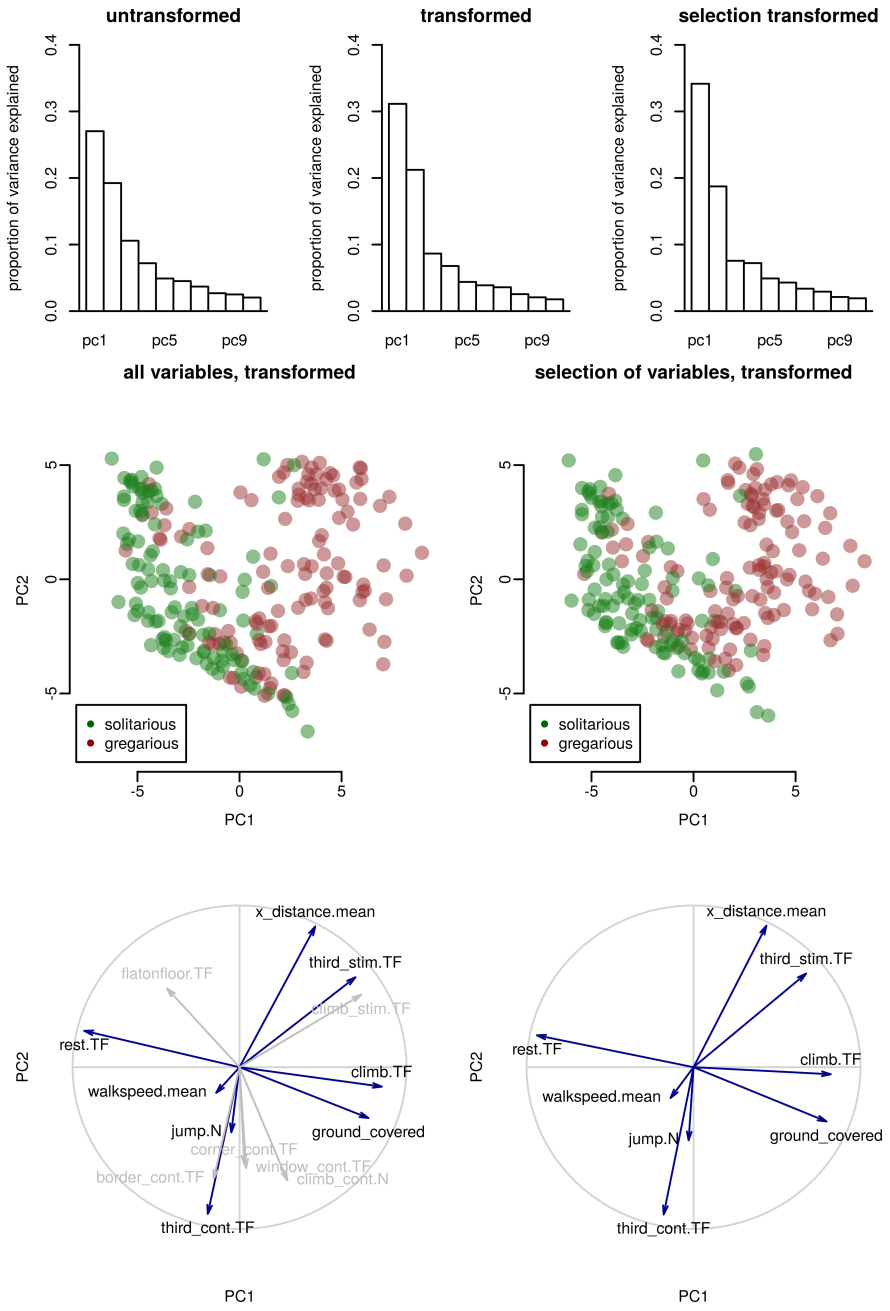



**Figure 4.16:** Top: Scree plot for 3 principal component analyses. Transformation leads to the capture of slightly more variance in the first two principal components. The removal of putative crossloading variables moved some captured variance from the second to the first component. Middle: projection of the observations onto the first two principal components for the total set of variables and the selection of variables, both transformed. Bottom: projection of a selection of high loading variables onto the first two principal components in a correlation circle. Grey circle is at unit distance from the origin. The variables plotted in grey in the first plot are the ones that are omitted in the second analysis.

section were clearly observable in this space. The third component loaded high on the number of jumps, walk speed and some position-related variables, while loading negative on climbing-related variables. The fourth component loaded high on variables related to sitting in the middle of the arena, while loading negative on variables involved in walk speed and sitting near the windows, regardless of the side of the arena (fig. 4.17 and 4.18).

Based on re-classification of the fitted values of a logistic regression without interactions (table 4.7), the two phases segregated well, albeit not completely on the first PC. Adding PC2 did not improve classification, while PC3 did. A logistic model based on the first 4 principal components had about 90% re-classification accuracy.

**Table 4.7:** re-classification rates for logistic models based on principal components

| Model | PCA on training data | | PCA on total data | |
|---|---|---|---|---|
| | class.greg | class.sol | class.greg | class.sol |
| PC1 | 80.0 | 78.2 | 80.1 | 76.4 |
| PC1 + PC2 | 76.8 | 82.7 | 76.0 | 81.8 |
| PC1 + PC3 | 84.8 | 84.5 | 86.4 | 84.5 |
| PC1 + PC4 | 81.6 | 79.1 | 85.6 | 80.0 |
| PC1 + PC2 + PC3 | 86.4 | 87.3 | 85.6 | 84.5 |
| PC1 + PC2 + PC4 | 78.4 | 81.8 | 83.2 | 83.6 |
| PC1 + PC3 + PC4 | 86.4 | 84.5 | 87.2 | 84.5 |
| PC1 + PC2 + PC3 + PC4 | 89.6 | 89.0 | 87.2 | 89.1 |

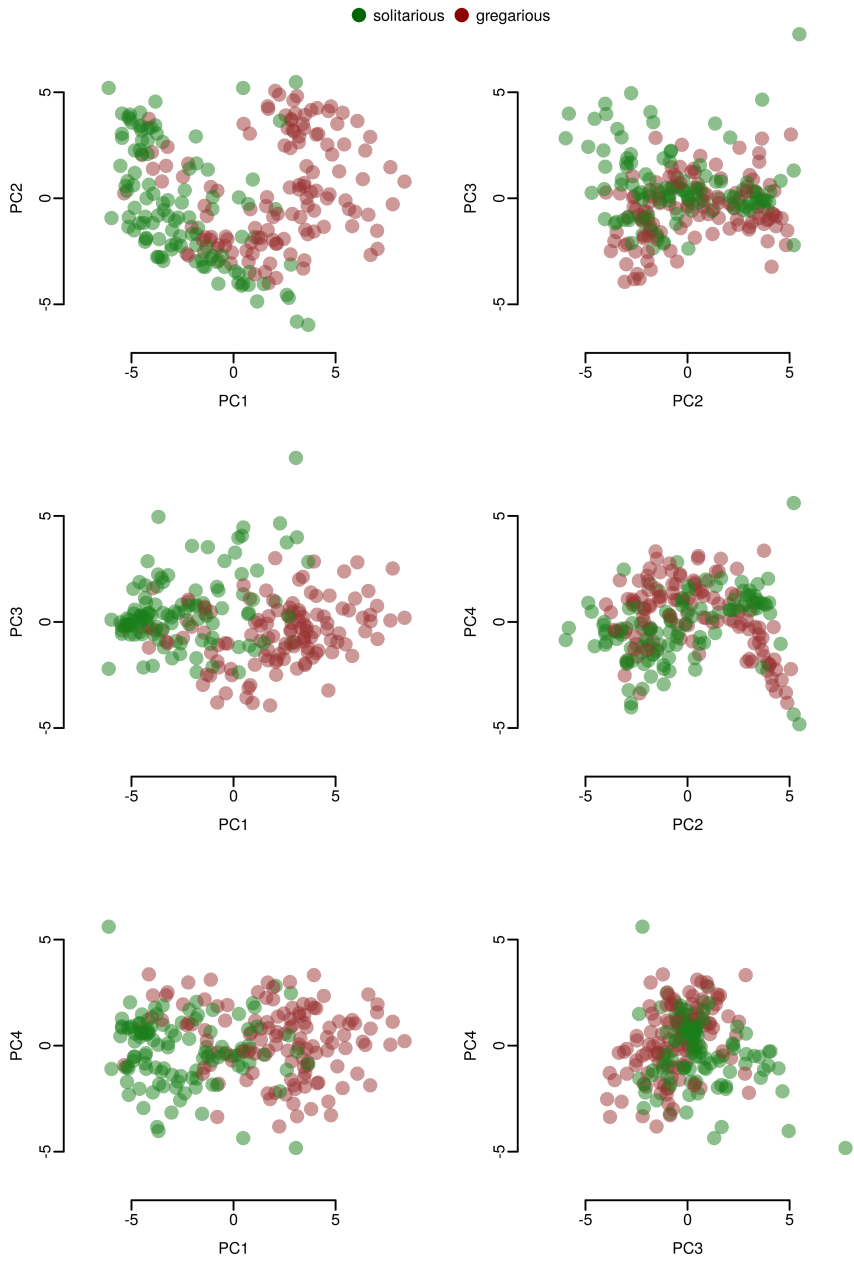



**Figure 4.17:** Scatterplots of the training data in the first four principal components originating from PCA based on the training data alone. Note how the majority of phase-related signal was present in PC1.

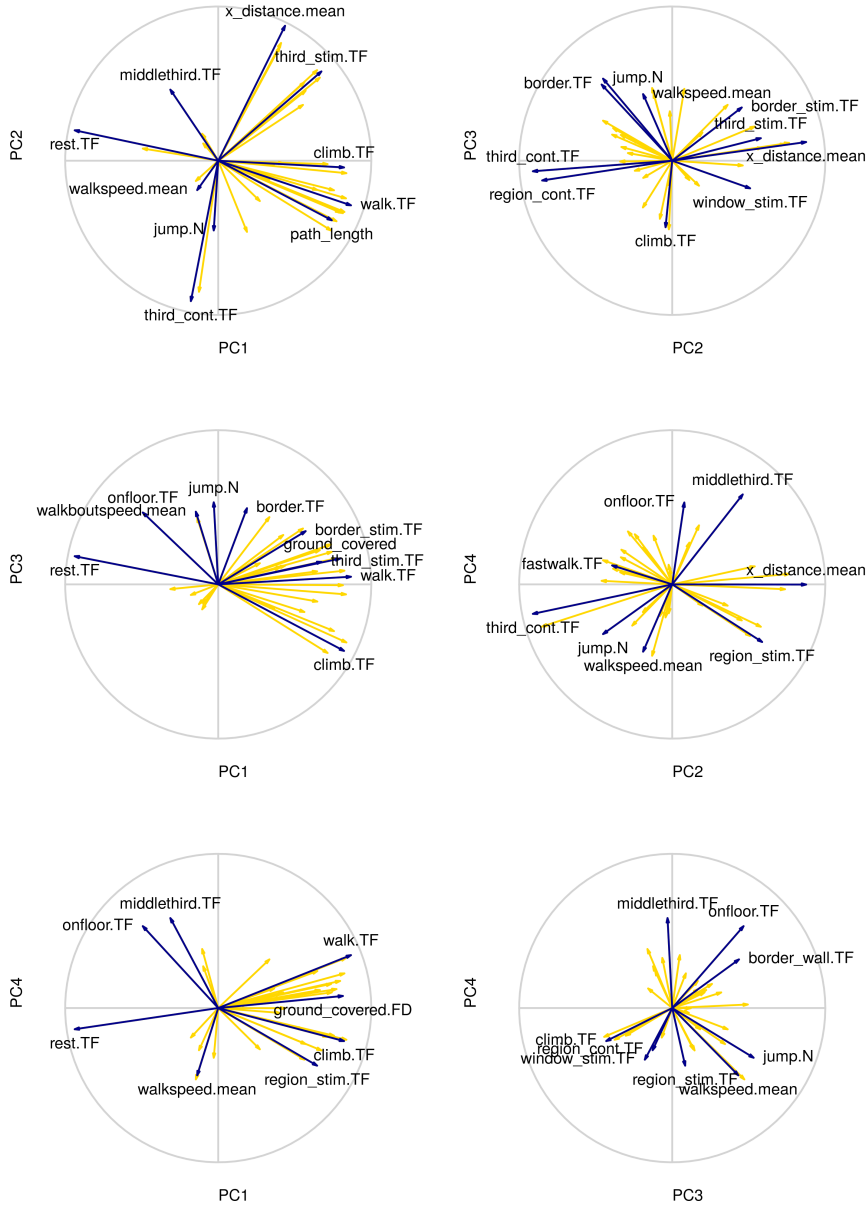



**Figure 4.18:** Projections of the variables with high loadings in the space of the first four components for PCA based on the training data alone. Since not all variable names can be plotted, the blue arrows represent illustrative variables. It is clear that PC1 primarily loads on activity, but also higher components captured part of the activity-related variance.

#### 4.4.6.2 Projection of the "automated total" data

When the observations from the crowding series and a small number of extra solitarious and gregarious animals were projected into the space defined by the 4 first components of the (above described) PCA based on the training data (fig 4.20 and 4.21), the following observations were made:

1. Neither the control animals nor the crowded ones projected on values outside the range that was observed for the training animals, except for one gregarious animal that scoreed very high on PC1. I double checked the video and it was indeed an extremely active animal.

2. Solitarious $5^{th}$ instars that were observed for the second, third or fourth time (always with at least a day in between) all seemed to fall within the distribution of first-time observations (not formally tested here, see later).

3. PC1 showed the clearest trend during behavioural gregarization, with a clear effect after 2 hours. There was no evidence for a clean linear trend within individual (fig. 4.19).

4. PC2, that loaded mostly on attraction/repulsion-related variables, increased most between 2 and 4 hours of crowding. A full day of crowding pulled the majority of animals over the edge towards gregarious-like values for PC2.

5. In PC4, transient states scored higher values than both long-term phases.

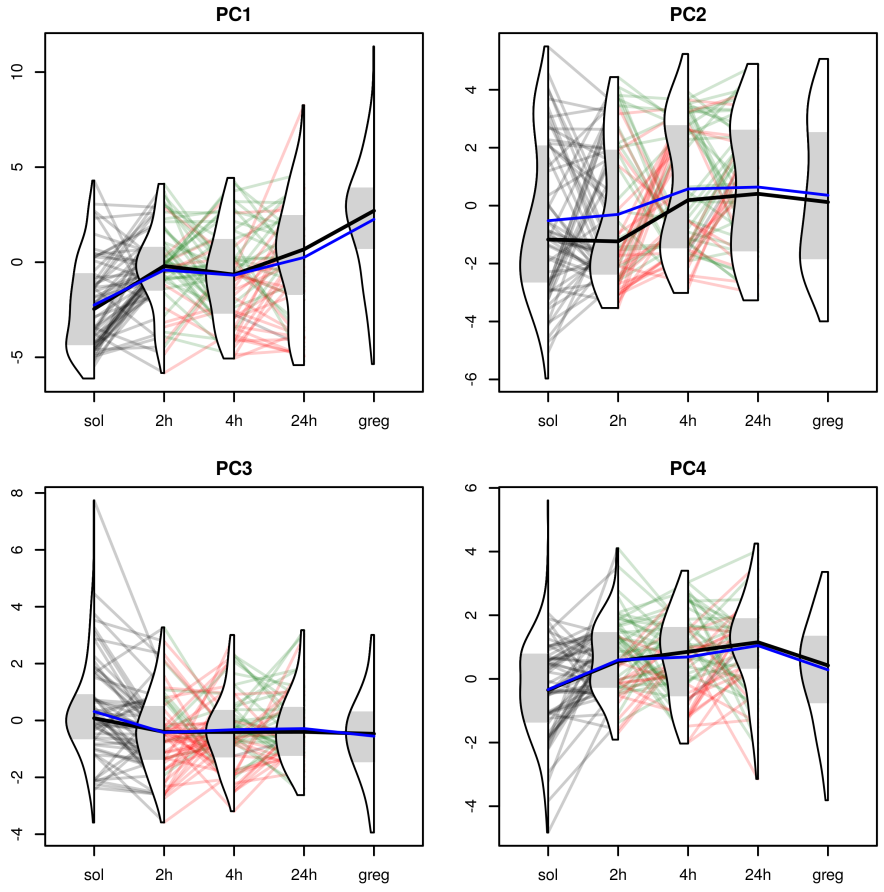



**Figure 4.19:** Violin plots for all 586 observations in the "automated total" set, projected in the space of the first 4 principal components defined by the training set. Connection lines only available for the 81 animals in the crowding series.

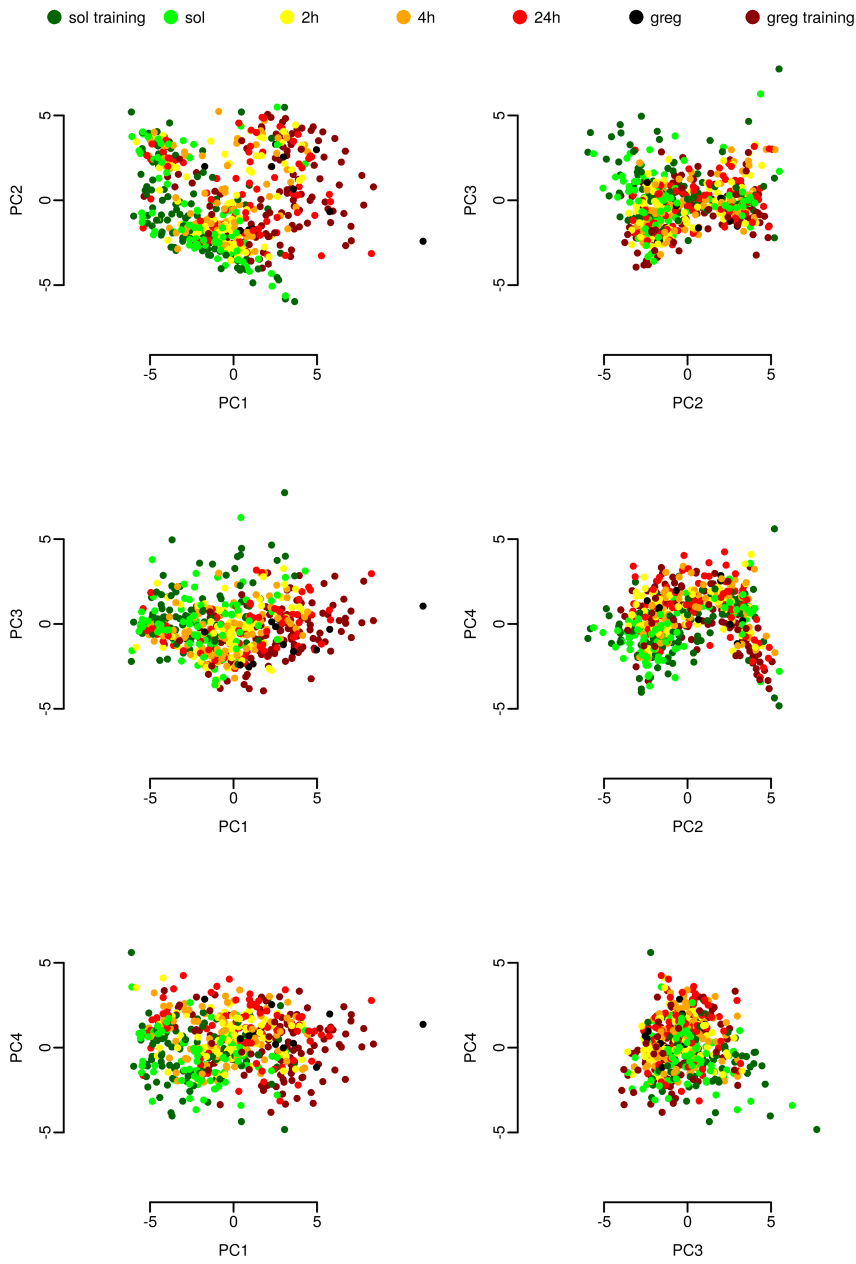



**Figure 4.20:** Scatterplot of the projections of all data into the space defined by the principal components calculated from the training data.

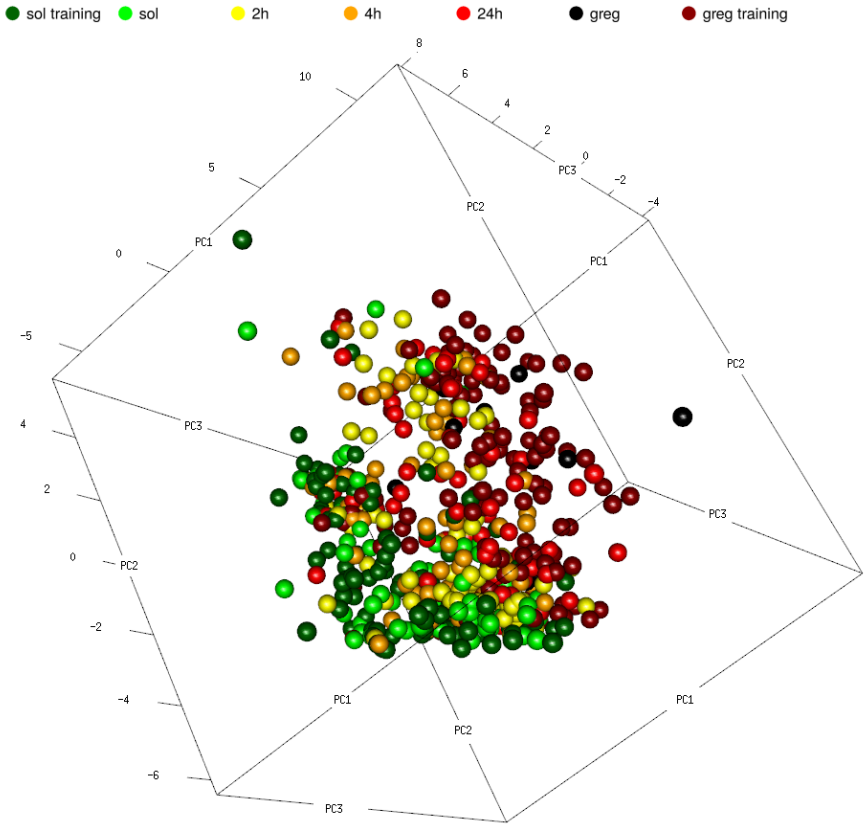



**Figure 4.21:** 3-D projection of all data onto the first 3 principal components.

### 4.4.7 Exploratory factor analysis

The plethora of possible estimation and rotation techniques did not yield dramatically different results. Since candidate latent factors such as "climbing" and "walking" were correlated, obligue rotations yielded cleaner factors with fewer cross loading variables. However, regardless of the rotation, cross-loading remained a problem until the variables were trimmed down to the following set: walk.TF, fastwalk.TF, walk.N, border.N, longwalk.N, ground_covered, ground_covered.FD, path_length, x_distance.max, third_stim.TF, third_cont.TF, half_stim.TF, region_stim.TF, region_cont.TF, x_distance.final, x_distance.mean, onfloor.TF, climb.TF, climb.N, longclimb.N, jump.N and turn.N.

The cleanest solution was obtained with 3 factors estimated with the minimal residual

method and rotated with *oblimin*. This yielded 3 clear-cut factors corresponding to "walking", "climbing" and "attraction/repulsion". The correlation between "walking" and "climbing" was 0.54, which already suggests that this structure would never have been recovered in a merely orthogonal rotation. Attraction/repulsion and climbing were also mildly correlated (r=0.22). Cross loadings were minimal (r < 0.25), except for x_distance.max that loaded on activity (0.39) and region_stim.TF that loaded on "climbing" (0.32). Jumping and turning were not well captured in this 3-factor solution. If a fourth factor was allowed, it loaded faintly on jump.N (r=0.51) and on variables related to sitting on or near windows, regardless of activity. Since the loadings on this factor were low and the interpretability was poor, the 3 factor solution was chosen as best describing the data.

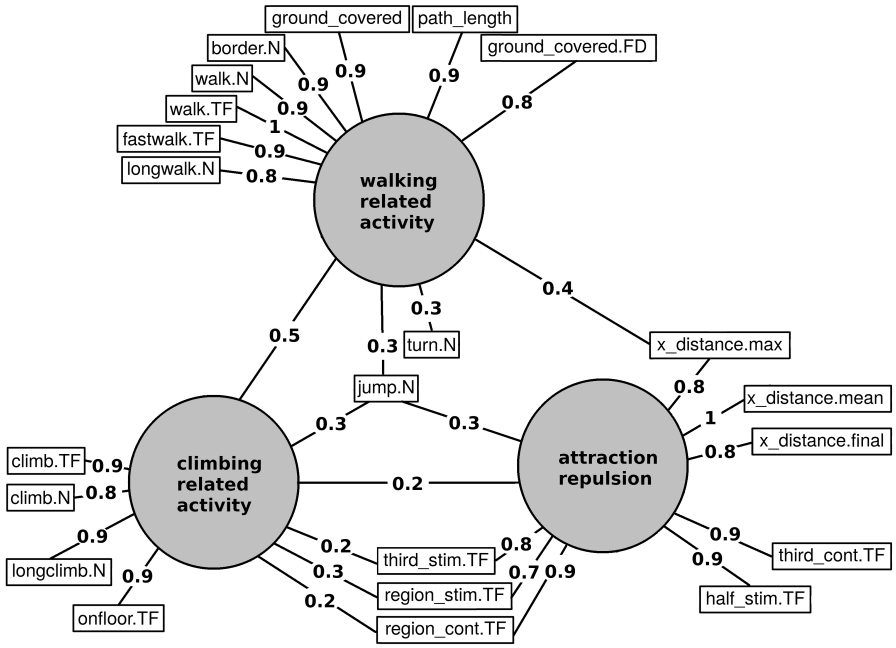



**Figure 4.22:** Diagram for the 3 factor solution. Absolute values of all factor correlations above 0.2 are shown. Precise values can be found in table 4.8.

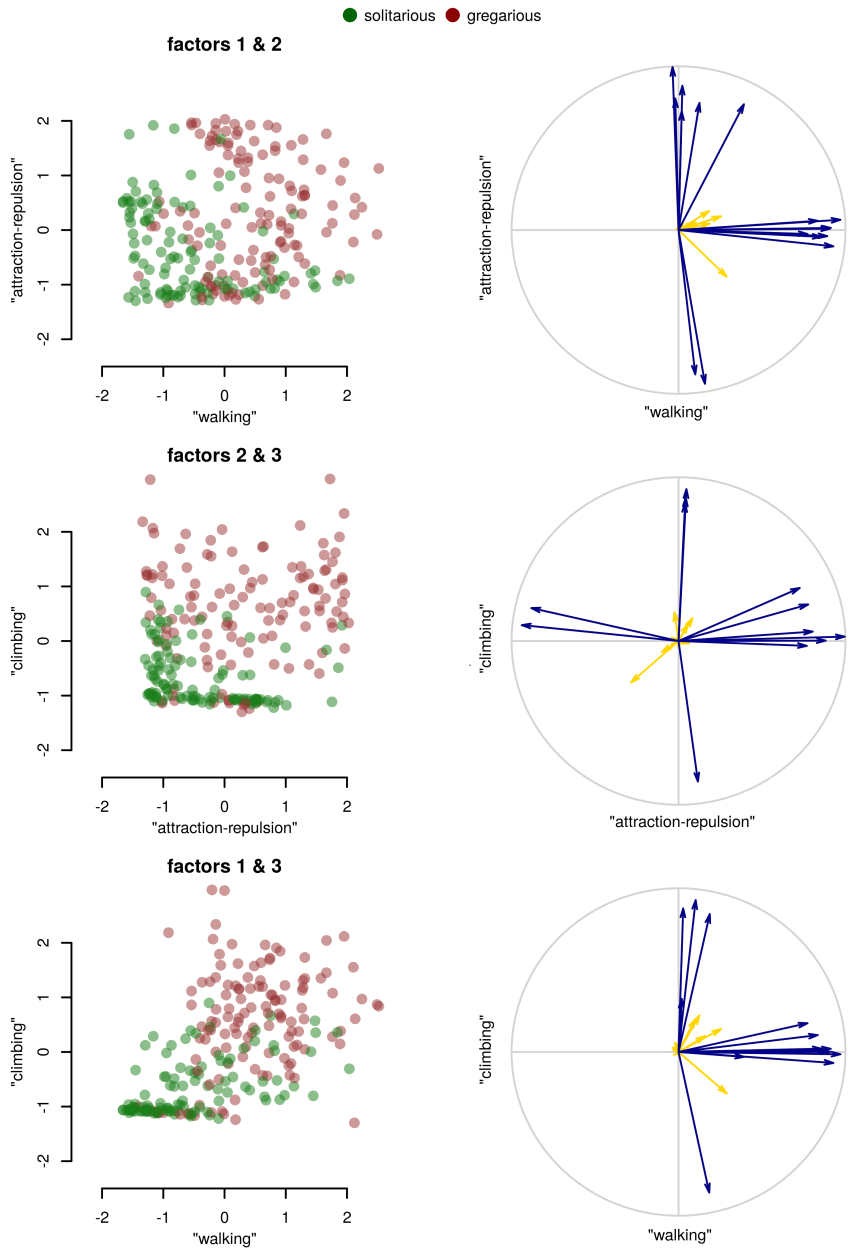



**Figure 4.23:** Exploratory factor analysis: component and factor scores for the 3 factor solution. No names were plotted on the factor scores because of printing reasons, but they clearly group well, and only factors with minor loadings (yellow, Pearson correlation < 0.4) deviate somewhat more from orthogonality. Solitarious locusts that show a certain degree of activity, also show strong repulsion, while in gregarious locusts, only animals that happen to be passive have low factor scores for attraction/repulsion.

**Table 4.8:** Summary of *oblimin* rotation of minimal residual 3 factor model. All correlations above 0.2 are shown.

| Variable | Factor loadings | | | communality | uniqueness |
|---|---|---|---|---|---|
| | "walking" | "att/rep" | "climbing" | | |
| walk.TF | 0.97 | | | 0.94 | 0.0595 |
| fastwalk.TF | 0.89 | | | 0.80 | 0.2048 |
| walk.N | 0.91 | | | 0.85 | 0.1469 |
| border.N | 0.93 | | | 0.79 | 0.2090 |
| longwalk.N | 0.83 | | | 0.81 | 0.1867 |
| ground_covered | 0.91 | | | 0.83 | 0.1743 |
| ground_covered.FD | 0.77 | | | 0.77 | 0.2308 |
| path_length | 0.86 | | | 0.75 | 0.2533 |
| x_distance.max | 0.39 | 0.77 | | 0.78 | 0.2192 |
| third_stim.TF | | 0.78 | 0.22 | 0.79 | 0.2072 |
| third_cont.TF | | -0.94 | | 0.86 | 0.1372 |
| half_stim.TF | | 0.88 | | 0.78 | 0.2188 |
| region_stim.TF | | 0.73 | 0.32 | 0.74 | 0.2572 |
| region_cont.TF | | -0.88 | 0.20 | 0.75 | 0.2452 |
| x_distance.final | | 0.80 | | 0.66 | 0.3352 |
| x_distance.mean | | 0.99 | | 1.00 | 0.0047 |
| onfloor.TF | | | -0.86 | 0.57 | 0.4262 |
| climb.TF | | | 0.93 | 1.00 | 0.0049 |
| climb.N | | | 0.84 | 0.93 | 0.0672 |
| longclimb.N | | | 0.88 | 0.81 | 0.1863 |
| jump.N | 0.29 | -0.28 | -0.25 | 0.16 | 0.8361 |
| turn.N | 0.26 | | | 0.14 | 0.8598 |

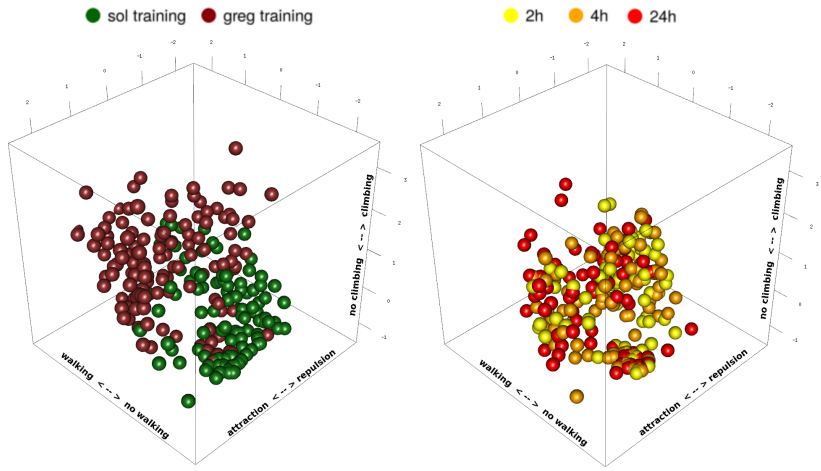



**Figure 4.24:** 3D projection of the training data (left) and the crowded animals (right) onto the EFA factors. Most observations for long term phases are situated on two opposite corners of the space. As mentioned earlier, the read-out of attraction/repulsion is clearest in those animals that show some degree of walking or climbing. Observations from the progression from solitarious (green) to gregarious (dark red) behaviour upon crowding, occupy the intermediate space about evenly after 2 hours (yellow), 4 hours (orange) and 24 hours (red) of crowding.

## 4.4.8 Models for behavioural phase state

### 4.4.8.1 Logistic regression models

The top 1% logistic regression models based on brute-force random permutation of all variables reached accuracies as high as 93% based on predicted values of the training data, although it should be noted that based on the permutation tests, these were almost certainly over-fits, especially in the case of 6-parameter models. The best-performing parameters were x_distance.mean, x_distance.final, x_distance.max, region_cont.TF, third_cont.TF, middlethird.TF, third_stim.TF, region_stim.TF, window_stim.TF, half_stim.TF, corner_stim.TF, rest.TF, walk.TF, fastwalk.TF, climb.TF, walk.N, climb.N, longwalk.N, longclimb.N, path_length, ground_covered, ground_covered.FD, climb_stim.TF, climb_cont.TF, onfloor.TF, jump.N and turn.N.

16 models were selected for further consideration (see table 4.9 and supporting materials). These models were based on stepwise procedures with the selection of best-performing parameters, but also I attempted to select models that covered the relevant variability within the set of variables as well as possible within 2 to 6 parameters. Equivalent models were compared based on AIC, BIC and GOF. Since the permutation test and the BIC suggested that 6-parameter models were typically over-fits, the selection was narrowed down to models with a maximum of 5 parameters. In the 5-parameter models, it was difficult to find a good balance between attraction/repulsion-based and activity-based parameters without getting poor GOF values (low p-values for le Cessie-van Houwelingen-Copas-Hosmer unweighted sum of squares test), even when the information criterions were low. The 16 models were complemented with models based on PCA and EFA. The PCA models are just for illustrative purposes, since the principal components contained some variables that were imputed with zeros. The logistic regression model based on factor scores did not show a particularly convincing re-classification, neither did it have low information criteria.

### 4.4.8.2 Random forest models

Based on the OOB estimates, the correct re-classification rate at cut-off 0.5 was 89.8% in classification mode and 90.2% in regression mode. There was a strong linear correlation between the OOB estimates of both models (pearson r = 0.996).

Looking at the increase of node purity for the regression model, and the mean decrease in Gini coefficient in the classification model, the most informative variables turned out to be rest time fraction and the number and time fractions of walking and climbing. Attraction/repulsion-related variables were present in the top 10, but generally weighed less on the classification than activity. For example: in the

regression model the increase in node purity is 9.6 for rest.TF (rank: 1) and 6.3 for climb.TF (rank: 2), while it is 3.1 for border_cont.TF (rank: 5) and only 1.3 for x_distance.max (rank: 10).

#### 4.4.8.3 A final selection of 3 models

After careful comparison of all summary statistics, re-classification rate and biological relevance, the selection of logistic regression models was narrowed down to two models. For random forests, the regression model was chosen. These are the 3 models that will be used in all further inferences about phase state:

**Model 2** is a simple bivariate logistic regression model, based only on the rest time fraction and the time spent in the control area (near or on the window). These two variables perform well as a proxy for activity (or the lack thereof) and repulsion. Consisting of only two parameters without interaction, it is unlikely to over-fit, and it offers simplicity without severely compromising accuracy (with a re-classification rate of 87%).

**Model 5** is the model that, in my opinion, optimally samples the underlying latent dimensions without being over-parameterized. It consists of three activity-related variables with relatively low correlations, alongside two attraction/repulsion-related parameters. The fractal dimension of ground covered combines a measure of activity with path tortuosity, the time spent climbing focuses on the other major axis within activity, and the number of jumps supplements the model with a typical solitarious activity. Next, the closest distance from the stimulus window is a good measure for attraction, while the time spent in the neighbourhood of the control window is a typical repulsion parameter. This model has a sensitivity of 85% for solitarious and 90% for gregarious locusts with a $P_{greg}$ cut-off at 0.5, or 88% and 90% with the cut-off set at 0.53. It has low information criterions and easily passes the goodness-of-fit test.

**Model RF** is a random forest based on all untransformed observed variables except path_length.FD , and variables related to walk speed or bout duration. It consists of 5,000 regression trees, each node based on a random sample of 13 variables and a random sample with replacement of 225 observations. Based on OOB estimates, its re-classification rate exceeds 90%. Since the relative weights of the latent variables was not explicitly fixed in the model, the weight of activity may be higher than attraction/repulsion.

A total of 44 training locusts was misclassified by at least one model, but only 13 were misclassified by all. These were 7 gregarious animals that were passive and showed no signs of attraction, and 6 solitarious locusts that climbed and walked without

showing strong repulsion. Model 2 had the most unique misclassifications, 5 of each phase, with all misclassified solitarious animals from the same subcluster C1b.2.

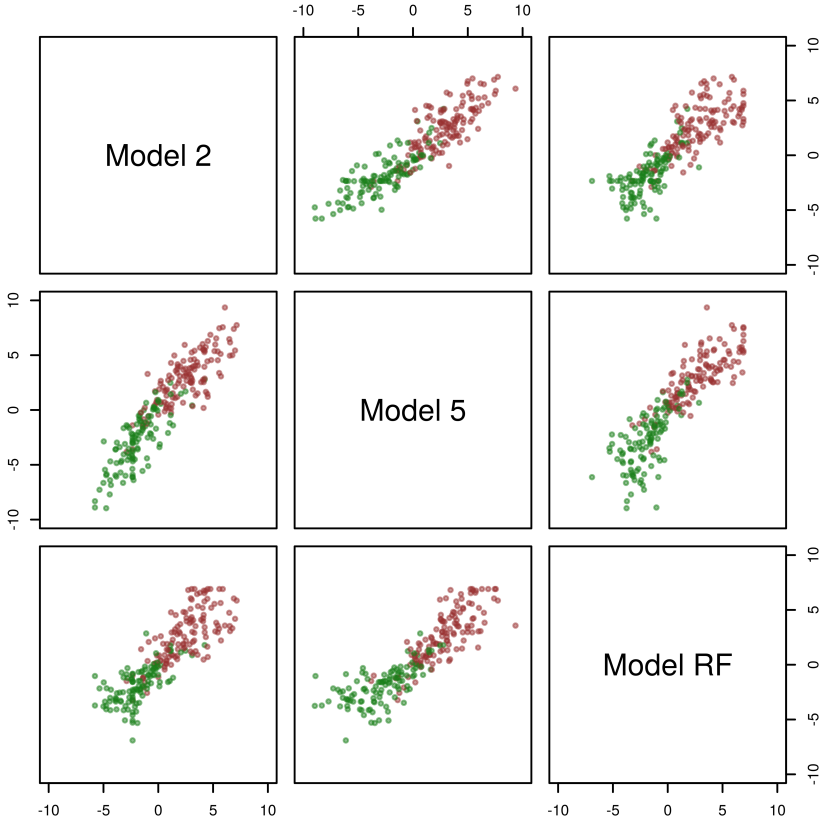



**Figure 4.25:** Scatterplot matrix of predicted $P_{greg}$ scores for training animals for the 3 selected models, all on logit scale. Red dots represent gregarious animals, green dots solitarious. The outcomes of the simple two parameter model and the more elaborate 5 parameter model are well correlated, with the 5 parameter model having extremer values, pointing to a higher degree of certainty in the classification of more obvious cases. Since a random forest is over-confident in the reclassification of the training set, the OOB estimates are shown here. Overall, they correspond well to the $P_{greg}$ values obtained with the logistic regression models.

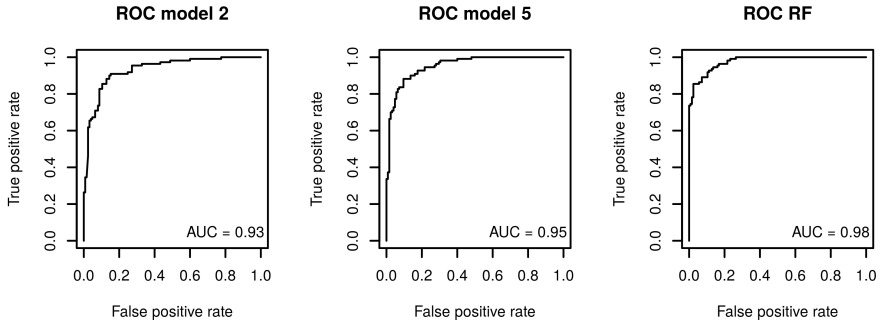



**Figure 4.26:** ROC for the 3 models, all with a high area under the curve (AUC). However, since these are all based on predicted scores of the training data (logistic regression models) or OOB estimates (random forests), the trade-off between sensitivity and specificity is likely to be slightly worse in reality.

#### 4.4.8.4 Validation on a small test set

The 3 models were validated on a small test set: 10 gregarious and 26 solitarious animals that were used as control animals in the experiments by Rut Vleugels on biogenic amine receptors. All these animals had been injected with dsRNA of GFP (green fluorescent protein). For all 3 models, the groups were perfectly separable (and obviously significantly different), but at the $P_{greg}$ cut-off of 0.5, 3 solitarous animals were misclassified by model 2, and 1 by model 5 and model RF (fig. 4.27). Factor scores for activity, attraction/repulsion and climbing were significantly different ($p<0.01$, Welch Two Sample t-test) between phases as well.

### 4.4.9 The effects of crowding

84 solitarious locusts were gregarized by crowding during an entire day and observed beforehand and after 2, 4 and 24 hours of crowding. For most activity-related variables, and hence for the EFA factors "walking" and "climbing", a quick response was seen after 2 hours, followed by a stagnation at values intermediate between both phases (fig. 4.28). The quick response was clearest in climbing-related variables, peaking at 2 hours. The number of jumps dropped to gregarious levels starting from 2 hours of crowding onwards, although it should be noted that also most solitarious locusts never jumped.

Attraction/repulsion-related variables reacted slightly more slowly, with a tipping point between 2 and 4 hours of crowding. There was no significant shift in the EFA

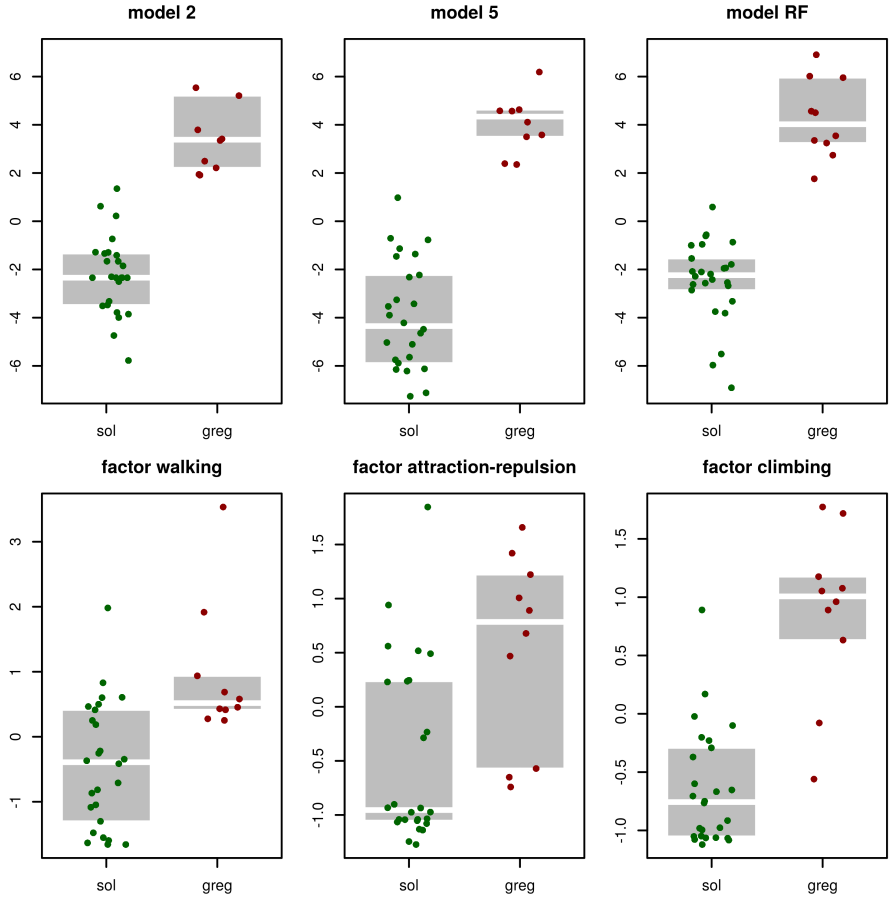



**Figure 4.27:** Top: comparison of $P_{greg}$ (logit scale) for the 3 compared models on test data. Bottom: comparison of factor scores between two phases.

factor "attraction/repulsion" after 2 hours, but it was definitely above solitarious levels from 4 hours onwards, reaching near-gregarious levels after 24 hours of crowding. The arena distribution plots (fig. 4.29, 4.30 and 4.31) show the same difference between 2 and 4 hours of crowding. While after 2 hours, the median position in the arena was still near the control window, it had shifted towards the middle of the arena after 4 hours (fig 4.30). Also, from 4 hours onwards, more than half of the animals visited the stimulus half of the arena at least once, and increasing proportions approached the stimulus window. However, even after 24 hours, levels of attraction/repulsion-related variables were still below gregarious values.

Correlations between the factor scores were similar to the ones seen in the training animals, both when calculated for all crowded animals, or per crowding duration (0.39 - 0.53 between walking and climbing, 0.12 - 0.27 between climbing and attraction/repulsion, and 0 - 0.05 between walking and attraction/repulsion).

All models differentiated well between long term phases and any time point of crowding (fig. 4.28). The difference between 2 hours crowding and 4 hours of crowding was not significant for any model, and between 2 or 4 hours and 24 hours it was significant for the logistic regression models only. Generally $P_{greg}$ reacted to crowding within 2 to 4 hours.

### 4.4.10 Some interesting observations

More than 15 potential nuisance variables were recorded as metadata. Since the combinations of treatments and metadata are not random, it is difficult to screen the entire set for the strongest effects, especially because these effects could potentially be found in 50 behavioural parameters plus a series of derived variables such as principal components, factors and models.

Because of the impossibility to model all possible effects, and the risks that come with multiple comparisons, only a limited set of hypotheses was tested. The results of all formal tests can be found in supporting materials.

#### 4.4.10.1 Metadata without evidence for effect

No significant effects on $P_{greg}$ or either of 3 factors were found for sex or stimulus side. The appearance of solitarious locusts (body black and colour) was of no influence to their observed phase state. There were no significant differences between the small and large apartments. There is no evidence that subsequent observations of solitarious locusts in the arena (up to 3 observations, always at least a day in between) had an effect.

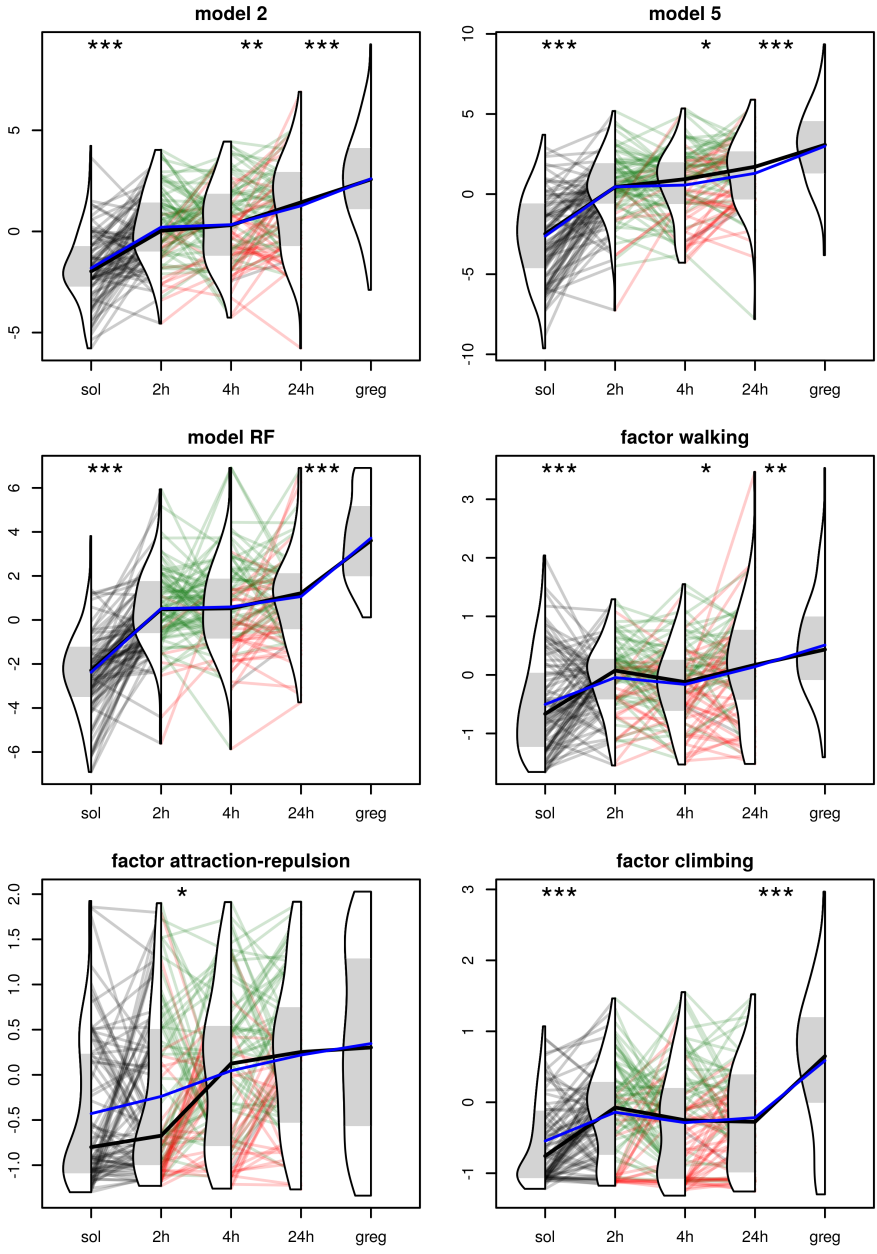



**Figure 4.28:** Violin plots for or models and factors during the time course of gregarization. Unpaired MWU tests were done between all neighbouring classes (* $p<0.05$, ** $p<0.1$, *** $p<0.01$).

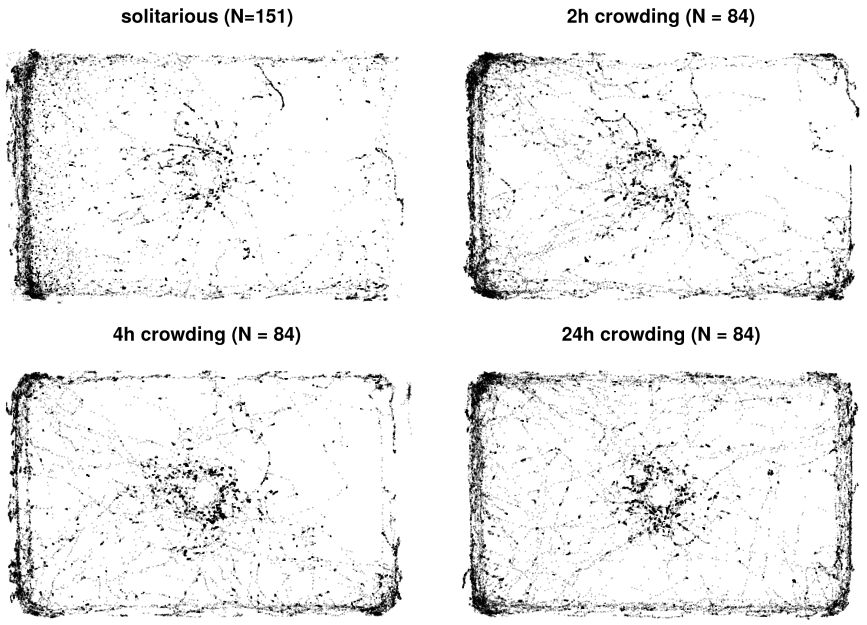



**Figure 4.29:** Scatter plots of the positions of hoppers in the arena for the crowding series. Stimulus group on the right hand side.

#### 4.4.10.2 Metadata with effect

There were suggestions of minor observer effects, but since observers were inevitably collinear with some conditions, they were difficult to account for.

The most striking effect was the difference in phase state in first-time observations of solitarious animals from the apartments versus the old cultures (fig. 4.32). All models classified the latter as more gregarious than solitarious, and the difference in $P_{greg}$ and all factor scores with the control solitarious locusts (26 animals raised in apartments but not used for model building) was significant (BH corrected MWU p-values all 0.012 or smaller, BH corrected Welch t-test for $P_{greg}$ all smaller than 0.01, but factor scores 0.016, 0.054 and 0.054). Unfortunately, this is based on the observation of only 5 animals, all within 1 day, so any inference is tentative.

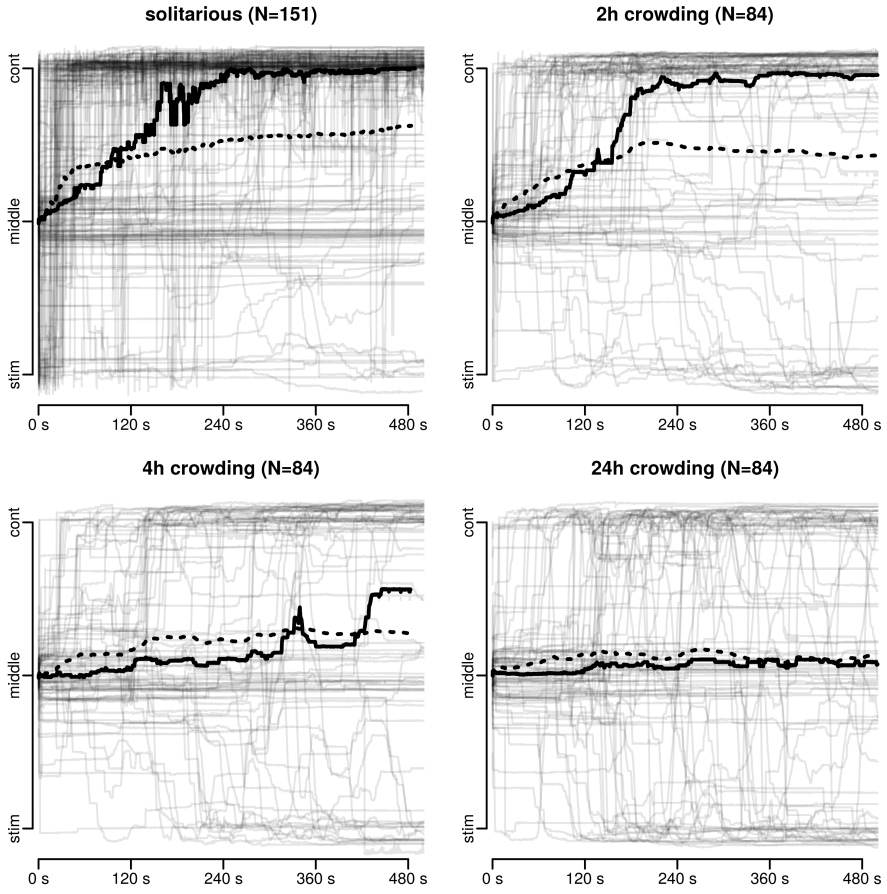



**Figure 4.30:** Distance from the stimulus group throughout the arena assay for the crowding series. Little happened after 2 hours, while after 4 hours the mean (dotted line) and median (solid line) position have clearly shifted.

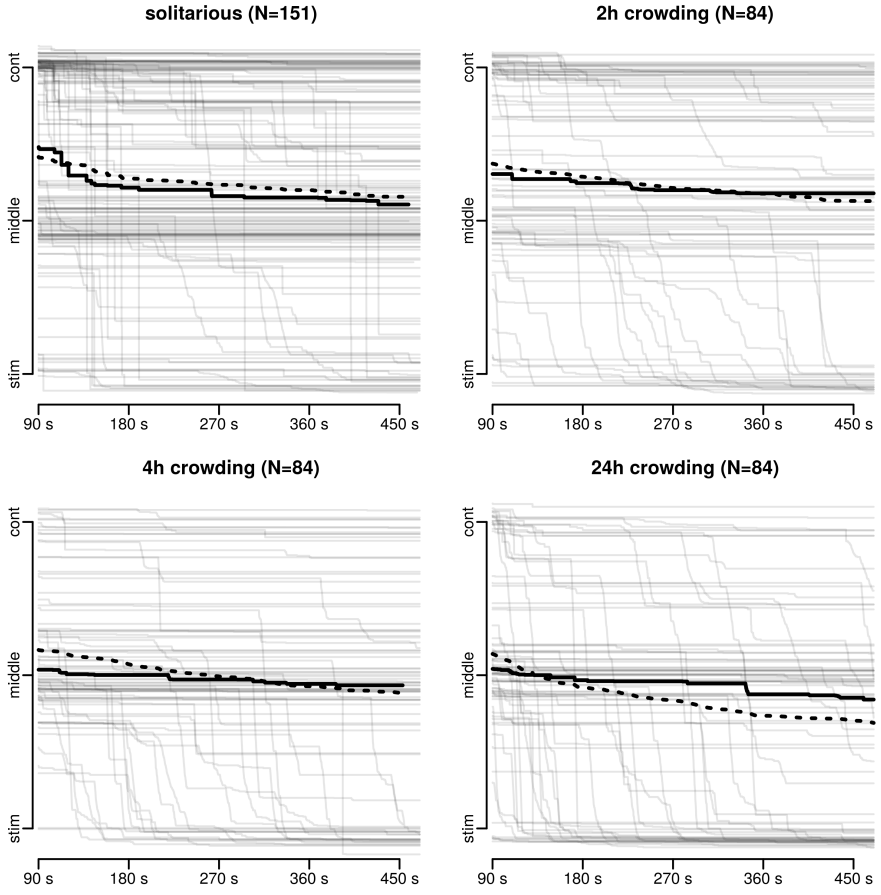



**Figure 4.31:** Cumulative closest distance to the stimulus group for the crowding series, starting after 90 seconds. There is a gradual shift of the mean (dotted line) and median (solid line) closest distance, but even after 24 hours of crowding the picture is still different from real gregarious animals (see fig. 4.12).

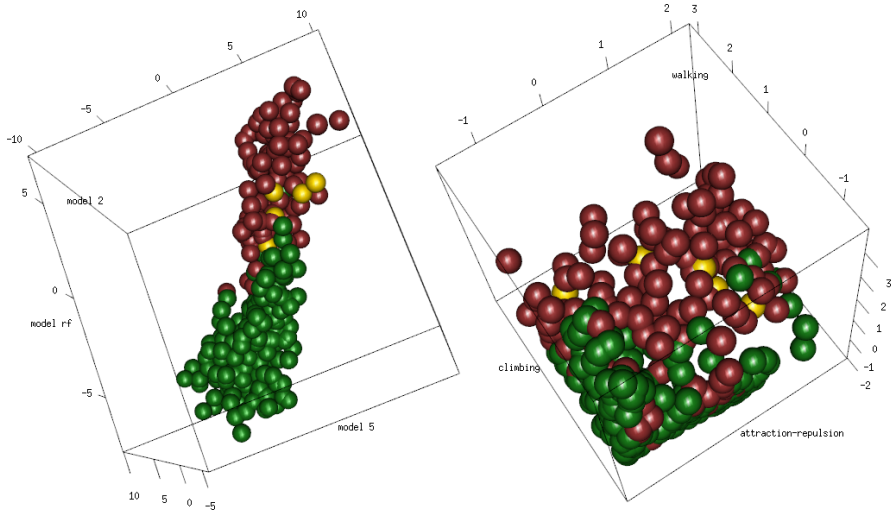



**Figure 4.32:** The position of the five solitarious animals that were raised in the pot cultures (golden color) relative to all other observations of long term phases. Left: the three models on logit scale. Right: the 3 EFA factors.

**Table 4.9:** All considered models with their accuracy based on predicted values of the training data (at $P_{greg}$ = 0.5), Aikake (AIC) and Bayesian (BIC) information criteria, significance level for the GOF test and mean squared errors (Brier score). The last two rows are random forest classifiers based on all variables except path_length.FD and all speed and bout duration variables. Here, the re-classification percentages are based on OOB estimates, because by definition, reclassification equals 100% for random forests. PC = principal component, r.tf = rest time fraction, rc.tf = time fraction spent in the control region (very close or on window), rs.tf = time fraction spent in stimulus region, w.tf = walk time fraction, lw.N = number of long walks (longer than 3 s.), gc.fd = fractal dimension of ground covered, x.max = maximal distance from control side, x.mean = mean distance from control side, c.tf = climb time fraction, j.N = number of jumps, t.N = number of turns, cs.tf = climb time fraction at stimulus window, cc.tf = climb time fraction at control window, 3s.tf,3c.tf and 3m.tf = time fractions spent in the stimulus, control and middle third of the arena. Fit 4e is closest to the model described in the PhD dissertation of Katja Stettin [243]. The models retained for further analysis are printed in boldface.

| Name | Parameters | accuracy | | AIC | BIC | p(GOF) | Brier |
|---|---|---|---|---|---|---|---|
| | | sol | greg | | | | |
| fit.pca.1 | PC1 | 78.2 | 80.0 | 208 | 215 | 0.177 | 0.141 |
| fit.pca.12 | PC1 + PC2 | 82.7 | 76.8 | 200 | 210 | 0.251 | 0.134 |
| fit.pca.13 | PC1 + PC3 | 84.5 | 84.8 | 173 | 184 | 0.314 | 0.112 |
| fit.pca.123 | PC1 + PC2 + PC3 | 87.3 | 86.4 | 159 | 173 | 0.849 | 0.100 |
| fit.pca.134 | PC1 + PC2 + PC4 | 84.5 | 86.4 | 163 | 177 | 0.750 | 0.100 |
| fit.pca.1234 | PC1 + PC2 + PC3 + PC4 | 89.1 | 89.6 | 153 | 171 | 0.738 | 0.092 |
| fit.pca.1i234 | PC1 x PC2 + PC3 + PC4 | 89.1 | 91.2 | 143 | 164 | 0.156 | 0.082 |
| fit.efa.123 | factor1 + factor2 + factor3 | 82.7 | 83.2 | 190 | 204 | 0.275 | 0.130 |
| **fit.2** | **r.tf + rc.tf** | **87.3** | **87.2** | **161** | **171** | **0.465** | **0.101** |
| fit.3a | r.tf + rc.tf + x.max | 85.5 | 90.4 | 155 | 169 | 0.245 | 0.094 |
| fit.3b | r.tf + rc.tf + x.mean | 86.4 | 88.8 | 159 | 173 | 0.155 | 0.107 |
| fit.4a | w.tf + c.tf + rc.tf + j.N | 90.9 | 86.4 | 147 | 164 | 0.355 | 0.097 |
| fit.4b | r.tf + rc.tf + x.max + j.N | 84.4 | 87.2 | 152 | 169 | 0.223 | 0.091 |
| fit.4c | w.tf + c.tf + rc.tf + j.N | 88.2 | 88.0 | 146 | 163 | 0.029 | 0.083 |
| fit.4d | gc.fd + c.tf + rc.tf + j.N | 87.3 | 89.6 | 144 | 161 | 0.811 | 0.099 |
| fit.4e | lw.N + gc.fd + rs.tf + j.N | 84.4 | 85.6 | 166 | 184 | 0.371 | 0.116 |
| fit.5a | w.tf + c.tf + rc.tf + j.N + t.N | 90.0 | 88.8 | 136 | 157 | 0.290 | 0.088 |
| fit.5b | w.tf + c.tf + rc.tf + j.N + x.mean | 90.9 | 89.6 | 142 | 163 | 0.033 | 0.081 |
| fit.5c | w.tf + c.tf + rc.tf + j.N + x.max | 90.9 | 89.6 | 142 | 162 | 0.042 | 0.081 |
| **fit.5d** | **gc.fd + c.tf + rc.tf + j.N + x.max** | **85.4** | **90.4** | **143** | **163** | **0.882** | **0.096** |
| fit.6a | w.tf + c.tf + rc.tf + j.N + x.max + t.N | 91.8 | 92.0 | 133 | 157 | 0.026 | 0.072 |
| fit.6b | w.tf + c.tf + rc.tf + j.N + x.mean + t.N | 92.7 | 88.8 | 133 | 157 | 0.017 | 0.072 |
| fit.6c | w.tf + cs.tf + cc.tf + j.N + x.mean + t.N | 91.8 | 86.4 | 145 | 170 | 0.889 | 0.097 |
| fit.6d | w.tf + cs.tf + cc.tf + j.N + rc.tf + x.max | 93.6 | 88.8 | 140 | 165 | 0.051 | 0.080 |
| fit.att.rep | rc.tf + rs.tf + x.mean + 3s.tf + 3c.tf + 3m.tf | 83.6 | 71.2 | 258 | 282 | 0.000 | 0.178 |
| fit.act.1 | r.tf + path_length + gc.fd | 88.2 | 90.4 | 168 | 182 | 0.009 | 0.108 |
| fit.act.2 | r.tf + path_length + c.tf | 85.5 | 89.6 | 168 | 182 | 0.102 | 0.100 |
| rf.class | Random forest on entire set | 89.0 | 90.4 | N/A | N/A | N/A | 0.100 |
| **rf.regr** | **Random forest on entire set** | **90.0** | **89.6** | **N/A** | **N/A** | **N/A** | **0.020** |

## 4.5 Discussion

### 4.5.1 Evaluation of the new tracker

I think an important benefit of the new tracker and *tandy.R* is the fact that the entire pipeline is now written in R. There are still some dependencies on other software (ffmpeg, imageJ), but the entire analysis starting from videos all the way down to the final statistical inferences can now be run from within a single environment.

It is difficult to quantify the difference in accuracy between Swistrack and the new tracker. The tracks obtained from Swistrack were often fragmented, while not a single fragmented track was generated by the new tracker. The few obvious errors almost all originated from external disturbances such as shaking of the arena or the locust already being present in the first frame. Although good results have been obtained with Swistrack, I have never managed to find a universal set of parameters that could track all our videos, despite my efforts to standardize them in terms of illumination conditions. A single cut-off for the contrast between locust and background also presented a risk for biases because differences in contrast between the animal and its background were collinear with phase. This problem has been solved with dynamic thresholding: the new tracker has little problems tracking all videos with a single set of parameters, except for the expected size of the locust in pixels, which shows a lot less variation within one age cohort.

The behavioural segmentation by *tandy* had already proven to provide reliable readings of a variety of behavioural variables [243, 242], and the adjustments I made were mostly subtle. My version of *tandy* outputs more behavioural variables than the original, some of which have turned out to be interesting. For example, x_distance.max is the attraction/repulsion-related variable that suffers least from unequal distributions between long term phases. On the other hand, other newly defined variables did not show any phase-related signal, but perhaps some day they will turn out to be important in another context.

I introduced some extra options for filters to cope with the typical problems encountered when differentiating between noise and genuine movement. This may present the risk of "feature bloat", where the number of adjustable parameters becomes too large for the user to optimize. To overcome this problem, I included a graphical user interface in which the filters can be tested in real time. It was reassuring that plausible results were obtained for a wide range of filter settings. I have tried kernels of sizes between 5 and 11 frames (for 12.5 fps videos). The default value of 7 frames corresponds to about 0.56 seconds, which is relatively high compared to the original value of 11 frames for a 30 fps video that was used in Cambridge. However, in conjunction with the noise filter, this value turned out to reliably tell apart genuine movement from noise even for very sensitive activity thresholds. Just

like in earlier attempts with EthoVision® [107] or *tandy* [243, 242], *tandy*.R fails to correctly identify small body movements such as grooming and swaying. The only account of successful automatic identification of small body movements, was an assay on *Chortoicetes terminifera* using EthoVision® [64], but here, all non-locomotive movements, including turning, were binned into one variable, so it is unclear to what extent a difference could be made between e.g. grooming, turning and swaying.

There is always room for improvement, and I think many problems would be solved by better video quality. Both the video camera and the accompanying software were present at the time of my arrival in the lab. The resolution and frame rate of the camera were well suited for the behavioural assays at that time, but the codec used to compress the video files was not ideal[7]. With higher resolution or lower compression, the amplitude of precision level noise relative to genuine movements would be reduced, which would greatly aid in the detection of subtle movements. Also, a sharper image would allow head-tail discrimination, or even the identification of body posture, which would be helpful to more precisely describe behaviours such as turning or climbing. For example, identification of turns was successful in my assay, but it was not possible to tell whether the animal turned towards or away from the stimulus. It would be interesting to find out if the direction the animal is facing is a component of attraction/repulsion behaviour.

### 4.5.2 Multivariate analysis of behaviour

To the best of my knowledge, this is the first detailed account of a multivariate analysis the arena assay for phase-related behaviour since its original publication in 1993 [215]. In a way, all multivariate analyses presented here, are belabouring the obvious: a set of behavioural variables is carefully defined to represent a number of latent factors, followed by the recovery of these very same factors by a variety of multivariate techniques. Nevertheless, each analysis has contributed insights in some aspects of the assay. Since all exploratory analyses were unsupervised, it was reassuring that overall, long term phases separate well, more or less regardless of the type of analysis. Yet, some more subtle traits only surfaced in a specific type of analysis. This rigorous exploration of the data has helped me constructing robust logistic regression models that are both logically and statistically sound.

Thanks to clustering, the predominant types of readings from extreme phases could be characterized. It would have been uninteresting to work with two deep branching clusters that separate both phases reasonably well, with impurities evenly scattered around. This is the reason why I have chosen to use the "complete" linkage function that yields rather spherical clusters of comparable size (which is confirmed when

[7] This surprises me since this whole system was set up by Noldus, a company specialized in behavioural analysis.

plotting the clusters in PCA space, see supplementary materials), with homogeneous branching depth. In retrospect, the clusters turned out to be more or less separable in other multivariate spaces as well (both in the first 3 components and factors, see supplementary data), but one could never have guessed them.

In the first place, hierarchical clustering has helped define stereotypes. For example: C1b.2 is a pure cluster of 14 rather active solitarious animals that showed no attraction and did not climb. The neighbouring clusters C1b.1 and C1a have about the same profile, but these animals do climb. These are almost purely gregarious. So, if a solitarious animal is active, it will actively avoid the stimulus side, while a gregarious locust that *happens* to show little attraction, will divulge its identity by climbing. Other stereotypical and more or less pure clusters are B2a (solitarious: passive animals that neither show strong attraction or repulsion) and C2 (C2a and C2b both show no repulsion, but differ in their activity).

Cluster analysis also discloses the weak points of models by showing the type of observations that are consistently misclassified. The clusters that are most mixed are B2b, B2a.1 and B1b.1. B2b shows animals that are generally average in most behaviours, which is also reflected in their position right in the middle of the PC1-PC2 plane, in between the two "horns". Within this cluster, gregarious animals have a higher $P_{greg}$ than solitarious animals, but still about half the solitarious animals get misclassified. B2a.1 is a group of passive animals that fail to demonstrate any signs of attraction or repulsion. This is the typical kind of observation where the read-out of phase state is hampered by a lack of information. Especially model 2 seems to be badly informed about this, misclassifying 4 out of 6 gregarious observations in this cluster. B1b.1 contains animals that show moderate activity both walking and climbing, but do not show any sign of attraction. Because of their activity, these animals get intermediate $P_{greg}$ values, and the random forest does the best job in classifying them correctly (16 out of 19 correct).

PCA was only partly successful in disentangling the layeredness of phase. Since *phase* was such a strong latent factor in the training data, it confounded the recovery of the next level of latent factors such as *attraction* or *activity*. This behaviour is caused by the obligation of principal components to be orthogonal linear combinations. PC1 may be interpretable as *phase*, but this does not mean that all variance captured in the higher components is unrelated to phase, as is exemplified by the fact that the addition of higher components to a logistic regression model greatly improved the model and its accuracy. PC1 loaded mostly on activity-related variables, and the loadings for attraction/repulsion-related variables were somewhat diffusely partitioned over the first 4 principal components, with PC1 perhaps containing the phase-related part, and PC2 catching the part that is not predictive (which is interesting). In PC3-PC4 space, stimulus and control side even fell exactly onto each other. This is likely the "residual" part of movement in the X-axis of the arena that has little to do with attraction/repulsion. That the first couple of components did not capture the majority

of variance, need not be a problem. It may just have to do with the high number of variables, each sampling slightly different aspects and each contributing some random error. Compared to earlier analyses, more behaviours that have little to do with phase were recorded, such as the extra fractal dimensions, the kick-in, and the subdivision of the arena in more sectors like borders and corners.

I think EFA managed better to provide an understandable image of the composition of behaviours in the arena, thanks to the possibility of oblique rotation, which confirmed that the two most important latent factors are correlated. Unlike the principal components, these factors are highly interpretable. In the space of the 3 factors, the majority of observations of the extreme phases occupy two opposite corners, with the space in between less densely populated. Upon crowding, the animals move into this intermediate space without moving much outside the distributions covered by the long term phases.

A striking observation was the failure to build a good logistic regression model based on EFA factors. Even when allowing interactions, the information criteria stayed high above values seen in trivariate regression models based on manifest variables or principal components. Perhaps it has to do with the way factor analysis treats error: the variance captured by the first principal components may be more purely phase-related, because error is captured in the higher components, while in the factors, all phase-related error is still present. Hence, for the mere representation of the dimensions that make up phase, factor analysis is most useful, while PCA, although it owes us no interpretability, has been most powerful in the recovery of *phase*.

### 4.5.3 Modelling extreme phases

Although some variables that were used in earlier models were not at hand (swaying, grooming), performant logistic regression models could be built. No walkspeed-related variables were used either, because of their high degree of missingness. This is not a new problem: Peggy Ellis [79] also had to resort to taking averages of travelled distances for the active hoppers only in her famous annular arena assay. I have had many discussions about what to do with variables like walk speed, and I argue that there is no problem in comparing them between groups for the sake of interest (like Ellis did), but if their value is set to 0 just because they were never observed, they may not be the best parameters in regression models.

It is difficult to get a good balance between activity- and attraction/repulsion-related variables, not only because activity has more signal, but also because attraction/repulsion-related variables have more awkward distributions. Solitarious locusts prefer to reside in one sector of the arena, while for gregarious locusts, the arena is less polarized. This is reflected in the distributions of many attraction/repulsion-related variables, and although the distributions were clearly

(and significantly) different between long term phases, their asymmetric overlap makes them poor classifiers. This is the reason why I introduced the x_distance.mean and x_distance.max variables, which seem to be less asymmetric. The time spent near the control window (region_cont.TF) is certainly related to repulsion, and it performs especially well in combination with climbing. This illustrates that a variable need not necessarily be a good separator on its own in order to perform well as a parameter in a multivariate model.

I am certainly not the first one to observe and acknowledge that in the arena setting, repulsion seems to be a stronger force than attraction. This same phenomenon was described in the original publication of the assay [215], and has later been observed in other assays with *Schistocerca gregaria* [217] and *Chortoicetes terminifera* [64]. I agree with the interpretation that in the first place, the arena may be too small for gregarious animals to really manifest attraction, and secondly, that in individual-based assays, the gregarious locust does not get the chance to exhibit its full gregarious behaviour because of the impossibility to interact, align and march with its peers. In the old (manual) assay, gregarious animals showed stronger signs of attraction. Perhaps this is no coincidence: in the old assay, the stimulus group could be approached until only the window separated them, while in the new assay, there was always a couple of centimeters separation.

In this work I have deliberately chosen to include as many subjects as possible in the training set. This was a necessity to pick up signal from behaviours that relatively few animals exhibit, like jumping. The drawback of this approach was the difficulty to accurately estimate model performance, because either estimates were based on the fitted values of the training data (which is by definition blind for overfitting), or on the classification of a set that was too small to discriminate between marginally different models. To compensate for these shortcomings, I used goodness of fit tests and information criterions, combined with my understanding of the system based on multivariate analyses[8].

Based on the multivariate statistics, I think it is reasonable to assume that for each class, about 10 animals would never be classified correctly because in any multivariate analysis they fall right in the middle of a region densely populated by the opposite phase (and indeed, 13 animals got consistently misclassified by all models). Next, the interface between both clouds in the 3 factor plot also predicts difficulties for another dozen observations. Hence, between 85% and 90% accuracy would be reasonable, but higher values should raise suspicion. The 3 proposed models all have estimated sensitivity in this order of magnitude, and also the classifications of the test set points in this direction (but it was too small to draw solid conclusions).

For most observations, the 3 proposed models agree, and most misclassifications

[8] Although it could be argued that since these multivariate statistics were also executed on the same data, they would be of little use in noticing overfitting.

that were specific to one or two models came from intermediate or paradoxical observations. However, each model also seemed to have its specific weakness. The most striking shortcomings were seen in model 2, which was the only model to misclassify observations from the purely solitarious subcluster C1b.2 (5 out of 14). Apparently, the heavy dependence on rest time fraction makes this model prone to the misclassification of relatively active solitarious animals. As mentioned earlier, model 2 was also the model that struggled most with passive gregarious animals (cluster B2a.1).

Model 2 may be elegant because of its simplicity, but its shortcomings point to a strong dependency of a single measure of activity (rest time fraction). Hence, I prefer model 5 because of its good representation of the entire suite of observed behaviours, which should make it less prone to bias. If any, the "blind spot" of model 5 is subcluster B2b: animals that are intermediate in all dimensions get a slight bias towards higher $P_{greg}$ values.

The random forest model performs well, albeit with the same bias in subcluster B2b as model 5. It has the strongest performance in the mixed subcluster B1b.1. Unfortunately it is difficult to understand what happens under the hood. Based on the mean decrease in Gini coefficient per parameter, a lot of decision power lies with activity-related variables, but this does not mean that attraction/repulsion-related variables may not be critical in cases where all things are equal with respect to locomotion. Actually random forests should be the eminent method to deal with classifiers that are only informative under certain condition, but it is always difficult to make head or tail of a black box.

### 4.5.4 Comparison between old and new set-up

The new and old set-up differed on many levels that may influence the behaviour of the observed animals. The new arena was smaller, the floor and windows were made of a different material, and the illumination was different. For the new assay, the solitarious animals were also raised in a new facility. Hence, next to technical differences, there may also be differences of biological relevance. Unfortunately, constantly giving priority to other experiments, I did not have sufficient numbers of locusts to formally test for differences in solitariousness between the pot cultures and the apartments. As a consequence, it is sometimes difficult to attribute meaning to my observations: did variations in observed behaviour reflect differences in solitariousness, or were they due to technical differences between the assays?

Only five solitarious locusts originating from the pot cultures were observed with the new set-up, and regardless of the model, their $P_{greg}$ differed significantly from the model population. This difference was not merely due just to increased activity, proven by the significant difference in their score for EFA factor "attraction/repulsion".

It is dangerous to draw conclusions based on the observation of 5 animals, all within 1 day (however small the p-values), but there are more clues that point into the direction of important differences in those variables that can be directly compared between the old and the new assay. The observed difference in numbers of jumps is unlikely to be merely technical, and it is also hard to believe that the differences in numbers of climb and walk bouts are just coincidentally present only in solitarious locusts. Other interesting trends are observed in variables related to attraction and repulsion: solitarious locusts prove to be strongly repelled by the stimulus group in the new assay, while in the older assay this effect was clearly less pronounced. On the other hand, the failure of gregarious locusts to show strong attraction, was less of a problem for the older assay. Here, again, it will be difficult to draw definitive conclusions because of the differences between both arenas, and the absence of animals that were observed in both settings.

Personally, I do not doubt that animals raised in the new facility are more solitarious. While most of the pots in the old facility were provided with an opaque haze by the washing machine, preventing the animals from directly seeing each other, the room was constantly filled with the smell of locusts, as evidenced by my swift allergic reaction upon entrance[9]. These perhaps rather weak, but ever present stimuli may have been sufficient to partly stimulate the cephalic pathway of phase transition. Moreover: I even speculate that this situation may actually have led to a desensitization of this very pathway, preventing the swift further transition to the gregarious phase state. This is in line with our observation that in earlier studies, the solitarious locusts from the Leuven colony did not easily gregarize, while in the later work by Rut Vleugels and myself, an unmistakable phase shift is observed after no more than two hours of crowding. Of course, the question remains to what extent these differences are not merely a reflection of two distinct behavioural models that were fitted on animals with two different degrees of "baseline" solitariousness.

### 4.5.5 Gregarization

There is general consensus that during phase transition in *Schistocerca gregaria*, a wide array of behaviours change. However, the jury still seems to be out when it comes to the relative pace of these changes. Personally, I am not convinced by the assay of Tanaka *et al.* (2013) [257]. First of all, they looked after 1 hour and after 1 day, skipping the time frame in which the first changes in attraction/repulsion. are typically seen. Also, although the animals are observed for a longer time, only 3 snapshots of their behaviour towards their conspecifics are taken. Next, the absence of strong evidence for changes in position relative to the stimulus is used as some kind of evidence for the absence of a quick response in attraction/repulsion. This

[9] A bio-assay as reliable as it gets

flawed reasoning was answered with a meta-analysis by Rogers *et al.* (2014) [217], and new data presented by Stettin (2014) [243].

In this work, I have provided further evidence that phase transition happens fast, and that repulsion may drop as fast as after 2 to 4 hours. It could be argued that after 2 hours, there is a confounding effect of increased activity on the reading of repulsion, but with the factor analysis, I have proven that the effect on attraction/repulsion is unmistakable the latest after 4 hours. This significant difference with solitarious animals does however not imply that the behaviour of advanced transient animals is equal to gregarious behaviour. All multivariate analyses are also in line with the idea that during early gregarization, the observed behaviours move along intermediate values, without moving in hyperspace regions that were not occupied by the training data. There is however no evidence for *concerted* changes, because activity and attraction/repulsion-related behaviours are uncorrelated in crowded animals.

## 4.6 Acknowledgements

I got an excellent training in video tracking and behavioural analysis by Swidbert Ott and Katja Stettin at Cambridge university, UK. The entire pipeline, and especially *Tandy.R* is closely based on their work. Videos discussed in this chapter were recorded by myself, Rut Vleugels, Heleen Verlinden, Senne Dillen, Elisabeth Marchal and Pieter Van Wielendaele. Steven Rombauts has been of great help in getting Swistrack to work on my computer. The arena was built and installed by me, Rut Vleugels and Roger Vranckx, and embedded in the impeccable carpentry work of Gert Heselmans.

# Chapter 5

# Transcriptomic signatures of different sensory input pathways during early behavioural gregarization

> Unfortunately, nature seems unaware of our intellectual need for convenience and unity, and very often takes delight in complication and diversity.
>
> Santiago Ramón y Cajal

# 5.1 Introduction

## 5.1.1 State of the art

### 5.1.1.1 Transcriptomic studies in locusts

Several studies have analyzed the transcriptome during locust phase change, or on transcriptomic differences between long-term phases, mostly in *Locusta migratoria*, which was the first locust species to have its transcriptome (EST) published [136, 166]. In the earliest studies, some general differences between long term phases were demonstrated, albeit with a method that does not require biological replication [241], but also with independent qPCR validations. This publication was followed by several papers that described phase-related expression patterns of specific transcripts (heat shock proteins [280], transposons [110, 131] and small RNAs [285]), but not in high-throughput.

In 2010, the EST-based *Locusta* transcriptome was updated with Illumina® sequencing [49], providing an improved coverage and the recovery of previously unknown transcripts. Long term phases were compared based on one replicate per life stage. Differences between phases increased gradually, peaking at the $4^{th}$ instar stage. General trends included an increasing investment in transcription and gene expression-related pathways, biosynthesis and general metabolism in solitarious animals, while the gregarious phase was more active in signal transduction, suggesting an increased need for processing of environmental information. In 2011, two papers were published with microarray studies in *Locusta migratoria*. The first one [165] contrasted long-term phases of $4^{th}$ instars, and had a strong focus on the catecholamine pathway, including functional experiments with knockdowns and pharmacology. The other one [109] was the first transcriptomics study with a focus on behavioural change rather than long term phases, looking both at gregarization and solitarization. In the head transcriptome, a total of 1,444 genes were found to be differentially expressed during the time course, although in my opinion, the majority of differential expression may be due to circadian rhythm, given the periodicity of changes after intervals of 24 hours[1]. 453 genes were found to respond in opposite ways to crowding or isolation, including chemosensory proteins (CSP), linked to the gregarious phase and the *takeout* gene product, linked to the solitarious phase. Both these gene products were also tested with qPCR, and found to react to crowding and isolation, especially in the antennae. Finally, the genes *LmigCSP3* and *LmigTO1* were knocked down with RNAi and animals were tested with an arena assay where both knockdowns turned out to influence the *attraction index*, a measure for the propensity of the animal to reside close to the stimulus group.

[1] Look at their supplementary figure S3 for more details.

In 2014, the draft reference genome of *Locusta migratoria* was published [281], alongside a large scale transcriptomic profiling of the brain during gregarization and solitarization (0, 4, 8, 16 and 32 hours). No less than 4893 genes came out to be differentially expressed at least at one time point compared to the starting point. Although I think this number may have been an overconfident claim[2], it provided a large set of transcripts that could be used for the study of pathways, and GO enrichment analysis. Here again, a contrast could be observed between investment in metabolism (solitarious) versus signal transduction (gregarious), alongside a change in expression of cytoskeletal components, suggesting re-modelling of neuronal circuits, which is a typical feature of neuroplasticity. Finally, this study was the first one to report on alternative splicing (45 genes).

A first EST-based transcriptome of *Schistocerca gregaria* was published in 2011 [13], followed by a microarray study focusing on differences between long term phases in the CNS transcriptome of adult animals [14]. The general picture of differences between long term phases was comparable to what was known from *Locusta*: genes involved in sensory processing, stress and immunity showed stronger expression in long term gregarious locusts, while solitarious animals invested more in metabolic processes (especially anabolic renewal), and processes that promote longevity, such as detoxification and anti-oxidant systems.

We are a long way from understanding the molecular underpinnings of phase change: studies have been done in different life stages, in different tissues, for different species, and although some general patterns have emerged, it is too early to make hard assertions or predictions about expected changes in gene expression upon early crowding in *Schistocerca gregaria*. Only long-term differences have been published for this species, and while many parallels can be drawn between the long-term phase differences in *Locusta* and *Schistocerca*, the transition is remarkably different between both species, suggesting that different molecular mechanisms may be involved [14].

Luckily, we have some prior knowledge from an unpublished microarray experiment (From the dissertation of Liesbeth Badisco: a crowding time course of $5^{th}$ instars, brain samples [12]) that was done in our group in cooperation with Cambridge university. In this study, crowded animals (2, 4 and 24 hours) were all compared to a solitarious reference sample. After two hours of crowding, 15 genes were detected as significantly up-regulated, and 33 as down-regulated. Up-regulated transcripts included two *heat-shock proteins*, two *rhodopsins*, *calmodulin-binding protein*, *chaoptin*, *stathmin*, *cathepsin L* and two *helicases* involved in RNA processing. Down-regulated transcripts included the putative chaperone *resistant to inhibitors of cholinesterase*, two *adipokinetic hormone* precursors, two *NADH dehydrogenase* subunits, *NADH ubiquinone oxidoreductase*, *cytochrome b*, some subunits of *cytochrome c*, *ATP synthase F0* subunit 6, and a *reverse transcriptase*.

[2]The paper is vague about the exact methods, the number of replicates etc.

Some of these patterns of differential expression were stable also after 4 and 24 hours of crowding, while others showed only transient peaks. In any case this study suggested that also in the desert locusts, some components of processes that are known to be different between long term phases, react to short-term crowding as well. *Usual suspects* include genes involved in stress response, functioning and organization of the nervous system and oxidative phosphorylation.

#### 5.1.1.2 The sensory pathways during early behavioural change

In *Schistocerca gregaria*, both the sensory input channels and the actual behaviours have a thoracic and a cephalic component. The cephalic pathway is activated by the combined sight and smell of other locusts [213, 152], while the thoracic sensory input pathway is activated upon tactile stimulation of the legs [233, 218]. In different species, different body parts probe for tactile information, and in the desert locust, the hind legs are by far the most sensitive spot [233]. In order to avoid undesirable stimulation caused by wind or sand, or by movement through vegetation, the stimulus needs to combine touch of the tactile hairs of the anterior surface of the hind femur with an inwards pressure on the coxal joint, which is typically caused by animals bumping and pushing into each other. For more details, see figure 5.2.

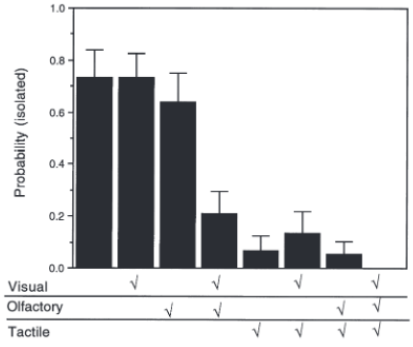



**Figure 5.1:** Figure from Roessingh *et al.* 1998 [213]: mean values (+SEM) of the probability of being solitarious based on the arena assay [215]. All treatments took 4 hours, the tactile stimulus was provided by moving paper balls. The triple treatment reliably gregarized all test animals, as exemplified by the absence of standard error (although this is actually an illustration that standard errors should not be used on probabilities), while tactile stimuli, or the combination of sight and smell were almost as effective. Sight or smell on their own, however, seem to have little to no effect after 4 hours.

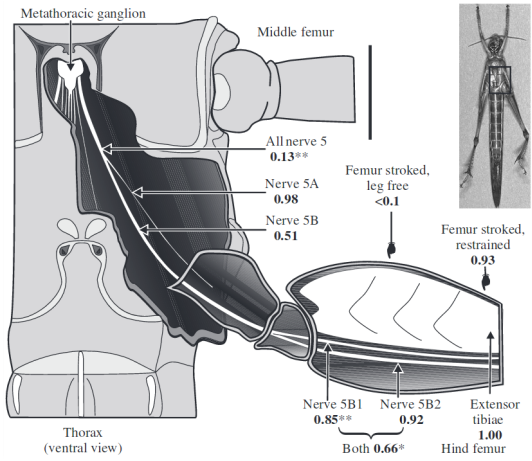



**Figure 5.2:** Figure from Rogers *et al.* 2003 [218]. Ventral view of a locust thorax and hind leg, with indications of the extent of gregarization caused by electrical stimulation of different sensory neurons, or stroking of a free moving leg in a solitary locust (4 hour treatments). Lower values indicate gregarious behaviour, and stars indicate a significant difference compared to solitarious control. The strongest effects were seen when metathoracic nerve 5 was stimulated close to the ganglion, or when a free-moving leg was stroked on the femur. Although stimulation of nerve 5A did not cause gregarization, it may well contribute, because stimulation of the distal proportion of nerve 5B has a more modest effect. Even more distally, nerve 5B splits, and only the combined stimulation of 5B1 and 5B2 had effects of a magnitude similar to the stimulation of the entire nerve. These findings are in line with the observation that a leg needs to move about the coxal joint in an orientation perpendicular to the normal walking movement, because while nerve 5B1 mostly conveys exteroceptive information from tactile hairs on the anterior femur, at least part of the proprioreceptive signal caused by inwards movement of the leg comes from strain on the cuticle, perceived by the campaniform sensilla, and conveyed through nerve 5B2.

Stimulation through either the thoracic or cephalic pathway has the potential to lead to the full set of gregarious behaviours, which implies that sensory input from different channels should converge at some point, with downstream effects that are independent of the initial sensory pathway. This integration should be a rapid process, given the observation that even during early behavioural change, both activity and attraction towards other locusts change together in a concerted manner [5, 217].

An important aspect of behavioural change is positive feedback [252]: the early boost in activity and loss of repulsion increase the likelihood of receiving more gregarizing stimuli of all kind. These early changes have both thoracic and cephalic components: increased locomotory activity is conceivably governed by the thoracic ganglia, while

increased attraction towards conspecifics must have an important cephalic component, given the need for visual and olfactory information about the position of other locusts.

The earlier stages of phase transition may well be a good candidate target for pest control, given that they precede part of the feedback that aids in the consolidation of the gregarious phase state. The same may be true from an experimental point of view: during early phase transition, some key processes may have started before secondary effects caused by feedback are noticeable. Some of the observed processes may even be specific for a certain sensory input pathway, while others may be part of the early integration of sensory information coming from both pathways.

### 5.1.2 Design of the experiment

The aim of my experiment was to disentangle the different sensory input channels and picture their convergence into a common pathway that ultimately leads to behavioural change. Animals underwent crowding-mimicking treatments through the presentation of subsets of naturally co-occurring stimuli. Their behaviour was tested in the arena, and observed changes in their transcriptome would tell us which processes were at work. These processes may either be specific to a certain sensory pathway, or they may be common between animals that underwent drastically different treatments, suggesting a more central role.

The high price of RNA sequencing puts a limit on the number of samples that can reasonably fit into one single experiment. This leads to a trade-off between the number of hypotheses that can be tested (conditions and their controls), versus the desired inference power (replicates). Part of the solution is to pool many individuals within one sample, which may reduce variation if its main source precedes pooling, which includes biological variation. This reduced variation can increase inference power, allowing a lower number of replicates. However, it is generally advised not to reduce the number of replicates too drastically, since a lack of statistical power can spoil an entire experiment, especially in the light of multiple testing problems. Based on similar experiments, I decided to use five replicates, which would allow me to test and compare six conditions in 30 samples. Since not enough animals were available to assay the behaviour and gene expression in different subjects, the samples for RNA-Seq had to be taken after observation in the arena.

A solitarious baseline was critical as a reference point, and in order to be able to quantify the effect of the arena, two solitarous conditions were created: completely naive, and after ten min. of assaying. The thoracic pathway was targeted like in the experiment by Rogers *et al.* (2003) [218], with the animals mounted in modelling clay, and their hind legs being tickled with a paint brush. At least one control was needed to account for the stress that the animals experienced while sitting in modelling clay, like being mounted without being tickled, or being tickled with a restrained hind leg.

Finally, two conditions were created with stimuli by real gregarious hoppers: one where the subject could not engage in physical contact with the stimulus hoppers (sight and smell only), and one where the subject was placed in the crowd, allowing the gregarious hoppers to physically interact with the subject.

It was not easy to decide on the appropriate *place and time*: the time course of behavioural change differs between lab strains, but also within the same strain, differences have been observed in the pace at which phase transition unfolds. I decided to look after two hours of stimulation, with the expectation that this would be a moment where transcriptomic signatures of the different sensory inputs start to surface, while some of the earliest integration may be taking place already. In later stages, beyond a certain "tipping point", I would expect the picture to become obscured by a whole range of downstream effects and positive feedback.

Now the time point was chosen, it still had to be decided which part of the central nervous system the focus would be on, because both the thoracic ganglia [219, 220] and the brain [219, 187] are known to play pivotal roles in the transmission of sensory input and in the manifestation of gregarious behaviour. Since the brain is considered the highest center of neuronal integration, and even early gregarious behaviour entails attraction towards other locusts (requiring visual input about the position of other animals), it was decided to work with the brain, which in itself is a complex tissue with many nuclei that may well show heterogeneous patterns of gene expression.

## 5.1.3 Finding differentially expressed genes

### 5.1.3.1 Mapping and counting

Counting mappings per reference contig would be a trivial task if every read mapped uniquely. However, references, be it genomes or transcriptomes, usually contain duplicate sequences. These may be repetitive elements, splice variants, allelic variants or paralogues. In case a read maps to multiple contigs in the reference with similar support, it is impossible to know for sure to which (if any) reference contig it actually belongs.

Several strategies are available to tackle this problem. First, efforts can be taken to reduce the redundancy of the reference: a single isoform or allele selected for each gene, repetitive regions masked, etc. A second straightforward measure may be to just discard any non-uniquely mapping read, or even to randomly assign it to a single reference contig. However, all these strategies are oversimplifications that may potentially lead to the loss of valuable information, for example about alternative splicing.

There are methods that provide an estimate about the most likely origin of a multi-mapping read: impossible as it may be to hunt down the origin of a single multi-mapper, with knowledge of the entire set of mappings at hand, information can be "borrowed" from other reads mapping on the same contigs, especially those that map on unique portions. A number of programs can do such estimations, with a variety of underlying methods [158, 191, 190, 33, 261]. I decided to work with RSEM [158], which uses an *expectation maximization* algorithm that assigns to each mapping a posterior probability that is conditional on all other "best guess" mappings.

#### 5.1.3.2 Statistical inferences

The statistical analysis of differential gene expression has presented a major challenge since the advent of high throughput technologies like microarrays. With these methods, the cost per tested gene is low, but the cost per total replicate is high. This results in analyses consisting of thousands of tests with limited numbers of replicates, which presents a major statistical challenge.

Classical multiple testing corrections (like Bonferroni or Benjamini-Hochberg procedures), assume independence of tests, which is not to be expected in the case of gene expression. Hence, these procedures are probably too conservative. However, without detailed prior knowledge about interactions in gene expression, no alternative is really available within the classical frequentist framework.

A major problem, especially with low numbers of replicates, is the estimation of variance. Typically, the distribution of variance estimates based on few observations is skewed, and the estimates are imprecise. In order to overcome this problem, information from other genes that show similar expression patterns can be used to better estimate the variance. This *shrinkage* was popularized in the R package *limma*[208], but in the meantime, numerous programs and R packages are available that all have their own methods to increase statistical power and robustness using information sampled from the entire set of genes [211, 209, 115, 151, 161]. Typically, these methods have an "empirical Bayesian" approach: for certain variance-related statistics, a prior distribution is estimated based on the entire dataset, followed by the gene-by-gene calculation of a posterior estimate based on the prior and an individually calculated likelihood score. I do think that under the right circumstances, these methods may indeed provide an increase in power, but I am afraid that in other cases, they may mainly provide overconfident inferences. Some packages even offer the possibility of working without replicates whatsoever. Personally, I put very little trust in these methods.

In order to be able to formally test for differential expression in a parametric way, some kind of distribution of read counts needs to be assumed. In other words: if the expression of a single gene is measured in an infinite number of biologically equivalent

samples, what is the expected distribution of counts? And, if the expression of this gene changes, which parameters of the distribution is expected to change? The theoretical distribution of read counts is debated [15, 163], but it is generally accepted that it can be approximated reasonably well by a negative binomial distribution [163, 211]. This distribution can be parametrized in different ways, always with two parameters. The most intuitive parametrization, I think, is the over-dispersed Poisson distribution. Imagine a gene of which the expression is measured multiple times from the same RNA sample. In theory, if the sampling technique (library preparation, sequencing, mapping etc.) is perfect, the only source of variance should be "shot noise" and a Poisson distribution of counts is expected, with the mean equal to the variance by definition. However, upon repeating this exercise with multiple RNA samples from individuals raised under identical circumstances, biological variation is introduced (typically modelled with a gamma distribution), and hence, the variance of the sampling distribution is expected to increase. The amount by which the Poisson variance increases, is called "overdispersion". The more biological variation within a condition, the larger this overdispersion factor (though it should be noted that sequencing techniques are not perfect, so part of the overdispersion is expected to be technical). If, finally, a group of individuals is subjected to a treatment that influences the expression of this gene, the mean counts are expected to change, together with the Poisson-part of the variance. Whether the overdispersion factor changes as well, is debated, and often it is assumed to be constant along conditions, which reduces the number of parameters to be estimated.

Another important aspect in the analysis of differential expression, is the normalization of counts. In the first place, counts have to be normalized within each sample (*i.e.* without looking at the distribution of counts in other samples). If the total number of counts differs between samples, counts of specific genes will differ accordingly. Likewise, if a certain contamination (like ribosomal RNA) is present in different proportions, the number of reads available to sample the expression of genes of interest differs as a function of the amount of this contamination (see figure 5.3). The same phenomenon can manifest itself more subtly, in cases where a few highly-expressed genes differ substantially between samples. Several methods are available for within-sample normalization [176]. In its most simple form, for each reference contig, counts are divided by the total number of mappings in the sample. In other cases, quantiles or median expression are used. These latter are more likely to account for idiosyncrasies that are independent of mere library size. Finally, some more advanced (but still single scalar) normalizers have recently been proposed, like relative log expression (RLE) [161] and trimmed mean of M-values (TMM) [210].

Normalization factors can be used in different ways. All counts can be divided by a scalar, converting them to measures like counts per million, or even to counts per kilobase per million (which is only interesting for the comparison of expression between different genes). A complicating factor here, is that normalized counts no

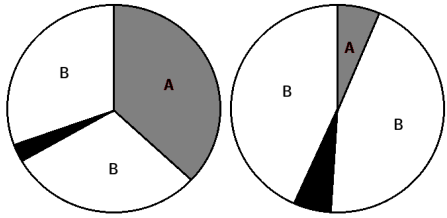



**Figure 5.3:** Pie charts of two hypothetical RNA-Seq datasets with identical sequencing depth (surface of the pie). The library on the left hand side has a severe contamination, e.g. with ribosomal sequences (surface A) while this same contamination is more modest in the other library. Note how only surface "B" can provide counts of other interesting transcripts. A randomly chosen transcript (black) may take the same proportion of surface B, while the raw counts may differ a lot. This kind of situation can be dealt with by normalising, within each sample, for the surface of B.

longer follow a negative binomial distribution. Some packages accommodate for this problem by simply rescaling counts and rounding them again to integer numbers, while others take more complicated approaches, involving the scaling factors as an extra parameter in their statistical models.

In some cases, normalization within sample can be combined with, or replaced by between-sample normalization. Here, counts are re-scaled such that the distribution of counts along a wide range of expression levels becomes more similar between different samples [176]. Rather than the single scaling factor used in within-run normalization, these methods apply a function along the continuum of mapping counts. Typical methods are quantile regression, or local regressions like LOESS.

## 5.2 Aims

1. To confirm and further characterize the effect of brief treatments with crowding, touch, or the combination of sight and smell on the behaviour of desert locusts.

2. To characterize changes in the brain transcriptome at the early stages of behavioural change.

3. To investigate whether these changes depend on the sensory channel through which the locust perceives changes in population density.

4. To look for a transcriptomic footprint of the integration of the cephalic and thoracic pathway.

## 5.3 Methods

### 5.3.1 Animals

All animals used in this experiment were mixed sex $5^{th}$ instar (day 3 to day 5 post ecdysis) solitarious locusts raised in the apartments. Only healthy looking locusts with intact legs and antennae were used. All locusts were fed fresh cabbage or collards in the morning.

### 5.3.2 Treatments

Treatments were chosen to optimally discriminate between sensory input pathways, and controls were chosen to represent the sensory input and stress the animals may have experienced by the treatments, without the gregarizing component. The treatments were the following:

**sol blank (sb)**: subjects were immediately sacrificed. This condition served as a control for arena effects.

**sol arena (sa)**: subjects were immediately filmed in the arena and then sacrificed. This condition served as the solitarious baseline for behavioural analysis and gene expression.

**crowded (cr)**: subjects were introduced in a 20 x 15 x 15 cm perspex box together with 25 mixed sex $5^{th}$ instar gregarious locusts. Some feed was present to prevent cannibalism. After 2 hours, subjects were taken out of the crowd, filmed in the arena and sacrificed.

**sight and smell (ss)**: subjects were individually put in a transparent 10 x 10 x 10 cm perspex box containing chicken wire to sit on and a leaf to feed on. The box had a fly screen lid and various perforations to allow free air flow. The box was placed in the $5^{th}$ instar cage in the gregarious facility. After two hours, subjects were taken out, filmed in the arena and sacrificed.

**touch (t)**: subjects were mounted on their back in modelling clay, with a bridge over the thorax between the middle and hind legs allowing all legs to move more or less freely (fig. 5.4). A piece of plastic was provided such that the locust could grasp onto something with its front and middle legs. The entire assembly was placed in a plastic tray to prevent visual distractions. Every minute, the femur of a randomly chosen hind leg was tickled for five seconds with a soft paint brush. After two hours, subjects were filmed in the arena and sacrificed.

**touch restricted (tr)**: subjects were mounted as described above, but both hind legs were also fixed into the modelling clay, with the dorsal sides of the hind femora facing upwards, and with a slight inward pressure, firmly immobilizing the thorax-coxa and coxa-trochanter joints. Just like in the "touch" condition, the femur of a random hind leg was tickled with a soft paint brush for five seconds every minute. After two hours, subjects were filmed in the arena and sacrificed.

**touch control (tc)**: subjects were mounted as described in the "touch" condition, but not tickled. After two hours, subjects were filmed in the arena and sacrificed.

Thanks to the help of a co-worker dealing with the treatments, it was workable to maintain a tight schedule where every 10 min. a new locust was introduced into one of the treatment groups. This allowed me to simultaneously observe a locust in the arena, and dissect the animal that had just been observed. All treatments took two hours except for the solitarious conditions. Dissections were all conducted between noon and 17h00.

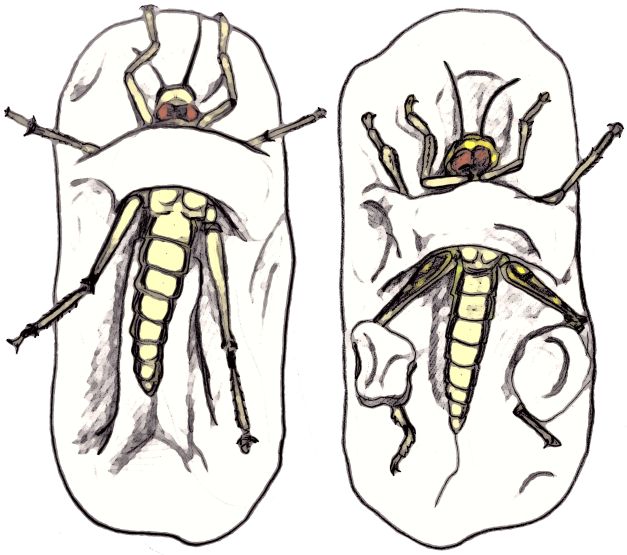


**Figure 5.4:** Drawing of the experimental set-up for the touch-related conditions. The animals are immobilized on their back in modelling clay, with a bridge over their thorax. The front legs can grasp a piece of plastic that is mounted above the head (not shown). Left: in the "touch" and "touch control" condition, the animal is able to freely move its hind legs. Right: in the "touch restricted" condition, the hind legs are immobilized dorsal side up.

### 5.3.3 Behavioural observation and analysis

Every locust, except for the "sb" condition, was observed in the arena according to the methods described in chapter 4. Since often many animals had to be observed and dissected in one day, the same schedule was always used: a subject was introduced in the treatments every 10 min., and two hours later every 10 min. a subject was put in the plastic tube while the next one was introduced in the arena and observed for 500 seconds, leaving about a minute to remove it from the arena for dissection and introduce the next animal for assaying. Videos were analysed as described in chapter 4 (25-31 observations per condition), and $P_{greg}$ values were calculated based on the logistic regression and random forest models described in chapter 4. $P_{greg}$ values were transformed to log-odds, which are expected to asymptotically follow a normal distribution, and all conditions were compared with ANOVA and Tukey's HSD (honest significant difference) *post-hoc* test. Factor scores were fitted based on the EFA presented in chapter 4. They had more awkward distributions, especially for attraction/repulsion, and in the absence of an expected (theoretical) distribution, conditions were compared with pairwise Wilcoxon tests with Benjamini-Hochberg correction.

### 5.3.4 Dissections

Every 10 min., a new locust came out of the arena and was immediately placed on ice. The brain, optic lobes and thoracic ganglia were all extracted following to the procedures described in the general methods. All dissections were executed in ice-cold locust saline, and tissues were separately snap frozen in 30 µl locust saline using liquid nitrogen. Samples were stored at -80°C until the moment of RNA extraction.

### 5.3.5 Pooling samples

For the time being, only brain samples were further processed. All samples were simultaneously thawed on ice, and pooled in a fresh tube per 6 subjects (3 male, 3 female) of the same condition, following the order in which the samples were taken.

### 5.3.6 Sample preparation

RNA was extracted in 3 batches of 10 samples each, with the QIAGEN RNeasy Lipid Tissue Mini Kit according to standard protocols, including a DNase digestion step, as described in the general methods chapter. Each sample was eluted in 50 µl and 30 µl

RNase-free water in an RNase free test tube. A 10 µl aliquot was immediately kept apart for later concentration and quality measurements.

### 5.3.7 Quality, concentration and extra samples

RNA concentration and integrity, and potential contamination were measured with NanoDrop™ 1000 (Thermo Scientific™) and Agilent 2100 Bioanalyzer. One sample was of insufficient quality (sample 24, "tc"), and three samples (samples 22, 27 and 28) did not contain sufficient amounts of RNA. Samples 22 and 28 (both "ss") were pooled, and three extra samples were taken for "sb" (4 animals), "tc" (6 animals) and "ss" (6 animals). The latter two served to replace sample 24 and the replicate that was lost by pooling 22 and 28, while the former was pooled with sample 27. Samples 22+28 and 27 + 31 were precipitated to a volume of 70 µl. Since the extra samples had to be taken a while after the anticipated end of the experiment, the breeding of large numbers of solitarious animals had been discontinued, and I was forced to include animals that had been raised in the old cultures.

### 5.3.8 Sequencing

Samples were packed on dry-ice, and were shipped (World courier, dry-ice guaranteed) to the Beijing Genomics Institute (BGI Tech Solutions), Hongkong on December 2, 2014. In Hongkong, their quality was measured, and next, 50bp SE libraries were prepared according to their standard protocols, using Illumina® TruSeq®. The following protocol was provided:

1. DNase I digestion.
2. mRNA isolation using poly-T-oligo-attached magnetic beads.
3. mRNA fragmentation with ultra-sonication (covaris).
4. cDNA first strand synthesis with random hexamer-primed reverse transcription.
5. Second strand synthesis and cDNA purification.
6. End repair with 3‘ adenylation.
7. Adapter ligation to the 3’ adenylated cDNA fragments.
8. Adapter ligated cDNA purification on TAE-agarose gel.
9. Many rounds*[sic]* of PCR amplification.

**10.** Library quality control with Agilent 2100 Bioanalyzer and ABI StepOnePlus Real-Time PCR System.

The sequencing was finished on January 16th 2015. After some initial quality checks, it turned out that between 20% and 51% (mean 25%) mapped to the *Schistocerca gregaria* mitochondrial genome, with the vast majority mapping to the 16S rRNA gene. This would only leave 13-14 million interesting reads per sample rather than the 20 million that were agreed on. When I reported this finding back to BGI, they first denied the problem, backed by an analysis by their own bioinformatics staff with their own software. They claimed that contaminations with 16S rRNA were minor, suggesting the contamination was mainly mtDNA, and hence would be our fault. In a reply I argued that DNA contaminations would be surprising given the fact that a DNAse digestion was included in my sample preparations and should have been repeated in their own library preparations. I also sent them a report of mappings to the reference mitochondrial genome (see figure 5.5 and supporting materials), which left little doubt about the origin of the contamination. In their reply, BGI apologized for mistakenly having claimed the contamination was genomic, now acknowledging it was indeed ribosomal, which they allegedly found out after a re-analysis which also confirmed the proportions of rRNA that I reported earlier. The bioinformatics staff of BGI hypothesized that the contamination was due to an exceptionally high adenine content in the *Schistocerca gregaria* 16S rRNA, which -I had to understand- was a species-specific anomaly that was not their responsibility. In a reply, I compared the 16S rRNA of the desert locust with that of some commonly sequenced species including fruit fly and human, many of which actually turned out to have equally long or even longer stretches of pure poly-A ribosomal RNA. Running out of excuses, BGI finally proposed to re-sequence each library to top up clean mRNA reads to at least 95% of the 20 million reads we agreed on. For my convenience, they already pre-calculated the number of extra reads they would sequence for each sample. After explaining to them that during our 3-month e-mail conversation, the rRNA contaminations would not spontaneously have disappeared from the cDNA libraries and hence the expected 18-40% contamination would have to be taken into account, they re-calculated the necessary read numbers and re-sequenced the libraries. On April 30th, the data were ready.

We ended up with more or less the appointed clean read numbers, albeit divided over two batches that will henceforth be referred to as *batch 1* and *batch 2*.

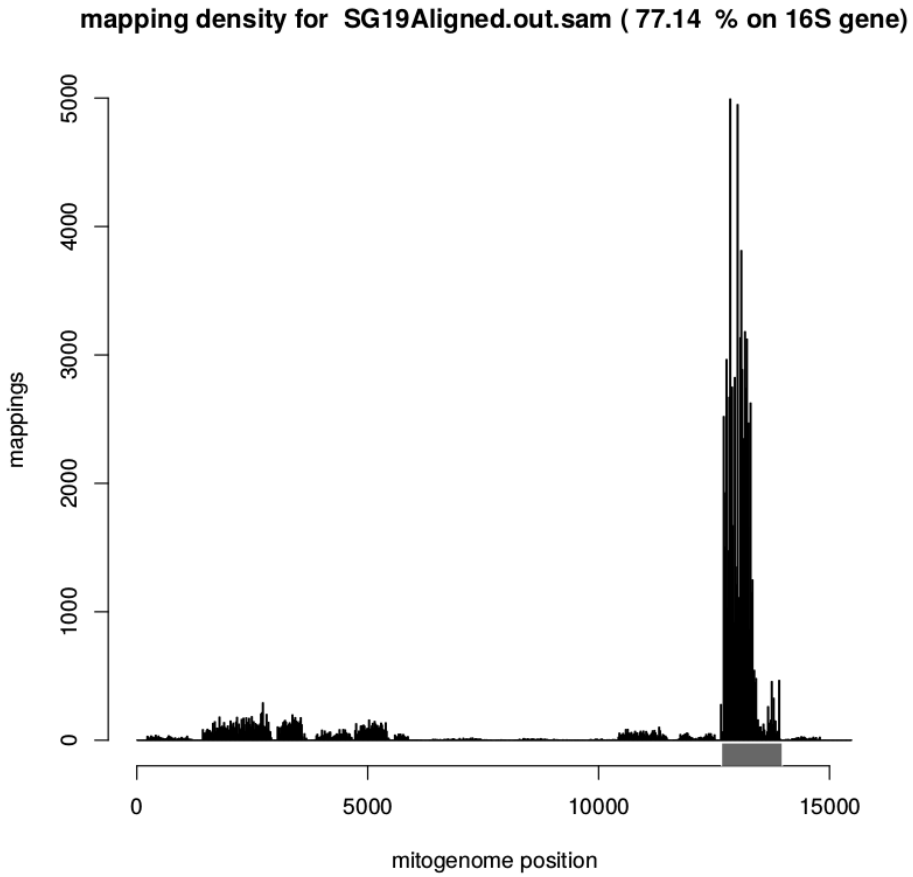



**Figure 5.5:** For each library (for example library SG19), all sequences were mapped against the *Schistocerca gregaria* reference mitochondrial genome. The grey block corresponds to the 16S rRNA gene. Obviously there are also some mappings on mitochondrial genes such as *COX1*, but the majority clearly maps to a single hot-spot.

**Table 5.1:** Overview of samples, with their presumed contaminations and read numbers sequenced for batch 1 and batch 2.

| Library | condition | % reads mapping to mito-genome | N reads batch1 | N reads batch 2 |
|---|---|---|---|---|
| SG0100 | sa | 20.9 | 22,310,400 | 4,196,300 |
| SG0200 | cr | 23.0 | 22,305,500 | 7,163,300 |
| SG0300 | sb | 29.5 | 22,211,300 | 12,545,800 |
| SG0400 | ss | 19.8 | 22,124,100 | 4,667,500 |
| SG0500 | t | 22.2 | 22,300,300 | 5,936,900 |
| SG0600 | tc | 21.4 | 22,316,300 | 6,033,200 |
| SG0700 | sa | 24.4 | 22,299,500 | 8,699,700 |
| SG0800 | cr | 20.1 | 22,322,100 | 5,027,300 |
| SG0900 | sb | 21.0 | 22,247,800 | 3,685,800 |
| SG1000 | ss | 22.8 | 22,297,500 | 5,814,900 |
| SG1100 | t | 25.0 | 22,324,100 | 7,531,100 |
| SG1200 | tc | 20.2 | 22,301,200 | 3,206,800 |
| SG1300 | sa | 33.3 | 22,228,500 | 13,080,300 |
| SG1400 | cr | 24.5 | 22,276,500 | 6,709,300 |
| SG1500 | sb | 24.1 | 22,003,400 | 6,433,400 |
| SG1600 | ss | 21.4 | 22,282,800 | 3,432,200 |
| SG1700 | t | 31.5 | 22,240,900 | 11,388,900 |
| SG1800 | tc | 19.6 | 22,313,900 | 2,321,100 |
| SG1900 | sa | 29.4 | 22,270,200 | 10,609,700 |
| SG2000 | cr | 20.8 | 22,263,900 | 3,372,200 |
| SG2100 | sb | 22.9 | 22,309,300 | 5,875,900 |
| SG2228 | ss | 20.8 | 22,312,800 | 3,346,400 |
| SG2300 | t | 20.2 | 22,301,300 | 4,475,600 |
| SG2500 | sa | 20.5 | 22,311,200 | 4,595,000 |
| SG2600 | cr | 22.4 | 22,330,100 | 5,244,600 |
| SG2731 | sb | 31.4 | 22,323,500 | 12,747,400 |
| SG2900 | t | 21.8 | 22,263,500 | 3,540,000 |
| SG3000 | tc | 17.9 | 22,130,600 | 2,401,600 |
| SG3200 | ss | 50.7 | 22,309,200 | 33,879,800 |
| SG3300 | tc | 43.0 | 22,322,900 | 27,253,800 |

### 5.3.9 Quality and filtering of sequences

The quality of all sequences was assessed with FastQC. All sequences were mapped (with STAR) against the *Schistocerca gregaria* mitochondrial genome, and the mapping density was visualized to screen for sources of contamination. Next, all sequences

were mapped (with STAR) to a reference containing the *Schistocerca* mitochondrial genome, and all known *Schistocerca* ribosomal sequences, supplemented by ribosomal sequences of other orthopterans that I thought may have been missing from the *Schistocerca* reference. Only reads that did not map to this reference were used in subsequent analyses.

### 5.3.10 Mapping and counting mappings

RSEM works well with a Trinity-formatted reference, but has little support for mappers other than bowtie and bowtie2[3]. I used bowtie2 to map all 60 libraries (30 replicates, two batches) to the reference transcriptome, with mapping parameters suggested by the RSEM developers, optimized for short SE reads mapping onto a transcriptome reference.

### 5.3.11 Exploring, normalizing and processing count data

All following procedures were executed on the gene-level counts. First, the data were explored with special attention for run- or batch-specific patterns, and condition-dependent trends. Hierarchical clustering was used to assess whether equivalent samples from batch 1 and batch 2 clustered together, while PCA and heatmaps of the correlation structure were combined to look for deviant samples, or condition-dependent grouping of samples.

Samples SG2731 (sb), SG32 (ss) and SG33 (tc) were left out of the analysis because they deviated from all others in terms of rRNA content and correlation between genes. The rationale behind this decision is worked out in the supporting materials.

Since there was little condition-dependent signal in the data, multiple normalizations were explored. I experimented with between-run normalizations, using functions from the Bioconductor *limma* package, but without convincing results. Four scalar within-run normalizations were used: re-scaling the upper 75% quantile of counts, the median count, the trimmed mean of M-values (TMM) method of Robinson and Oshlack [211] that is present in the Bioconductor *edgeR* package, and a TMM-based normalization based only on seven housekeeping genes [268].

Each time after normalization, data were filtered for extreme outliers: in the log-transformed data, for each gene, the distance of each data-point from the mean of all others was measured (in standard deviations, calculated excluding the data-point), and each gene containing at least one outlier further than 5 standard deviations from

[3] I tried to combine it with STAR, which (I think) worked, but it required some patching and included many manual scripting steps.

the mean, was omitted entirely. I attempted something similar within condition, but this turned out to be highly erratic with only 3 or 4 replicates available to estimate standard deviations. After outlier removal, normalization and outlier detection were repeated to measure the effect of outliers on the general structure of the count tables. These effects were minor.

### 5.3.12 Differentially expressed genes

I experimented with many R packages that are written specifically for the analysis of RNA-Seq experiments [211, 209, 115, 151, 161], and some more basic analyses to explore the gain in power I got from using these specialized methods. All these procedures and some of their results can be found in the supporting materials.

A problem that I found difficult to tackle with most software packages, was the *post-hoc* assessment of differentially expressed genes. The tests often stated that "something" was different, without detailing which biologically relevant pattern was observed, or which particular contrast was actually significant. This was different for the two Bayesian packages that I explored: *baySeq* [115] and *EBSeq* [151]. Both these packages apply an empirical Bayesian approach to the estimation of negative binomial distribution parameters and prior probabilities of differential expression. They go by quite similar philosophies, with neatly formulated hypotheses, but use different mathematical approaches for parameter estimation and normalization. The package *baySeq*, has the most readable documentation and publication, but it was extremely slow (days), and often analyses did not run to completion because of errors. The package *EBSeq* was a little faster, and analyses ran to completion, with parameter estimates converging into logical values that neatly fitted the data. Although *EBSeq* is one of the few packages that offers isoform-resolution analysis, I executed all procedures on gene-level counts (see discussion). Since I was not sure what the influence was of pooling the batches that had been sequenced at different times, all analyses were executed both on batch 1 only, and on the sum of counts from batch 1 and batch 2. The four scalar normalizations described in the previous section were all used.

In *EBSeq*, the hypotheses that are contrasted are formulated in terms of which conditions are similar or different from each other. For example: the null hypothesis is that all expression values are drawn from the same negative binomial distribution, regardless of condition. One of the alternative hypotheses was: all conditions are drawn from the same distribution, except for the "t" and "tc" conditions, because sitting in modelling clay may have an influence on the expression of some genes.

*EBSeq* optimizes 3 "hyperparameters". While fitting a negative binomial distribution on each gene for each hypothesis, it models the distribution of q-values (the complement of the p parameter of the negative binomial distribution, see supporting materials),

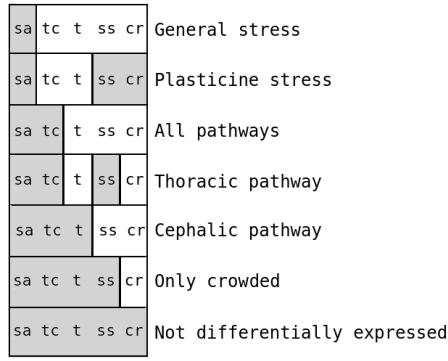



**Figure 5.6:** The set of hypotheses that were tested with *EBSeq*. In a first round, all genes were screened for an effect of the arena only ("sb" versus "sa"). Next, these 7 mutually exclusive hypotheses were tested. For each hypothesis, separate negative binomial distributions are fitted for conditions with the same background colour. For the last hypothesis, only one distribution is fitted for all. This can be considered a "null hypothesis", and is expected do best describe most of the genes. sa = sol arena, tc = touch control, t = touch, ss = sight and smell, cr = crowded.

with a two-parameter beta distribution. The alpha- and beta-parameter of this beta distribution are estimated together with the prior probabilities for all hypotheses, using the *expectation maximization* algorithm. After these three hyperparameters have converged, posterior probabilities are calculated for each hypothesis for each gene, using the Bayes theorem. I worked with a two-step procedure, where first, only the "sb" and "sa" conditions were contrasted with each other, and next, "sb" samples were ignored and hypotheses were formulated regarding the "sa", "tc", "t", "ss" and "cr" condition (see figure).

## 5.4 Results

### 5.4.1 Behaviour

The solitarious and crowded animals had $P_{greg}$ scores (both for the regression models and the random forest) and EFA factor scores within the range of earlier observations. The arena distributions (fig. 5.7 - 5.9) suggest that solitarious animals were somewhat slower in settling at a distance from the stimulus group, but the mean and median closest distance reached after 8 min. was similar.

All treatments, including both controls, had $P_{greg}$ scores significantly (p adj. < 0.05) higher than solitarious levels (fig. 5.10 and table 5.2), except for "tc" according to

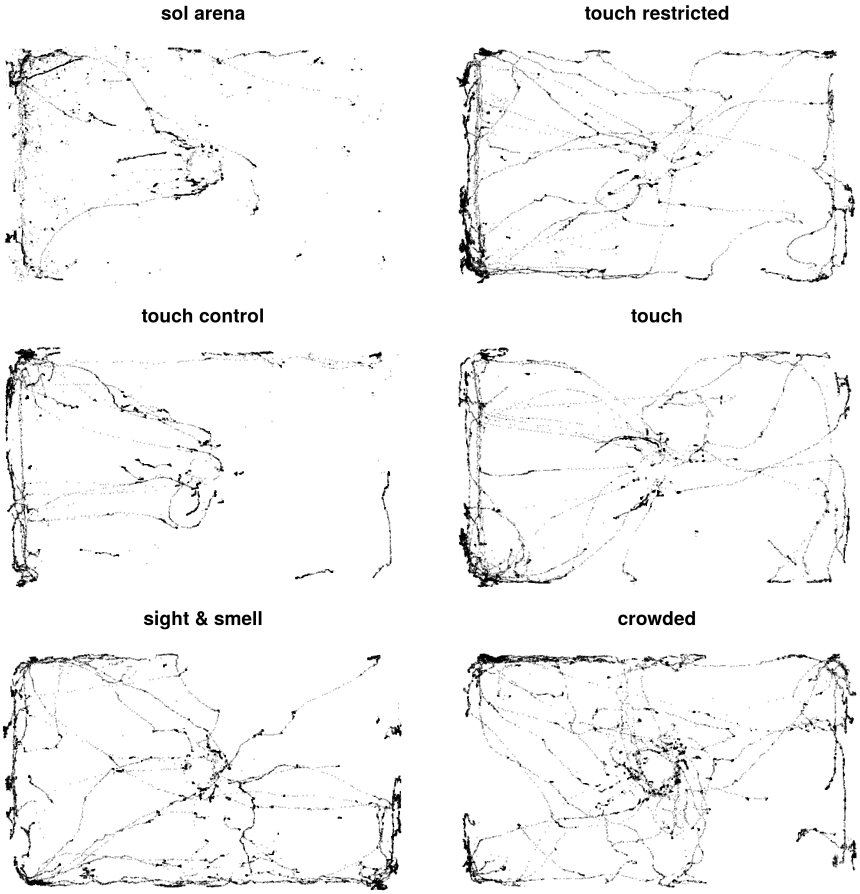



**Figure 5.7:** Scatterplots of the positions of hoppers in the arena. Stimulus group on the right hand side. These figures seem to show more trails than the figures presented in chapter 4, due to the fact that fewer animals were observed, and hence more points are printed per animal.

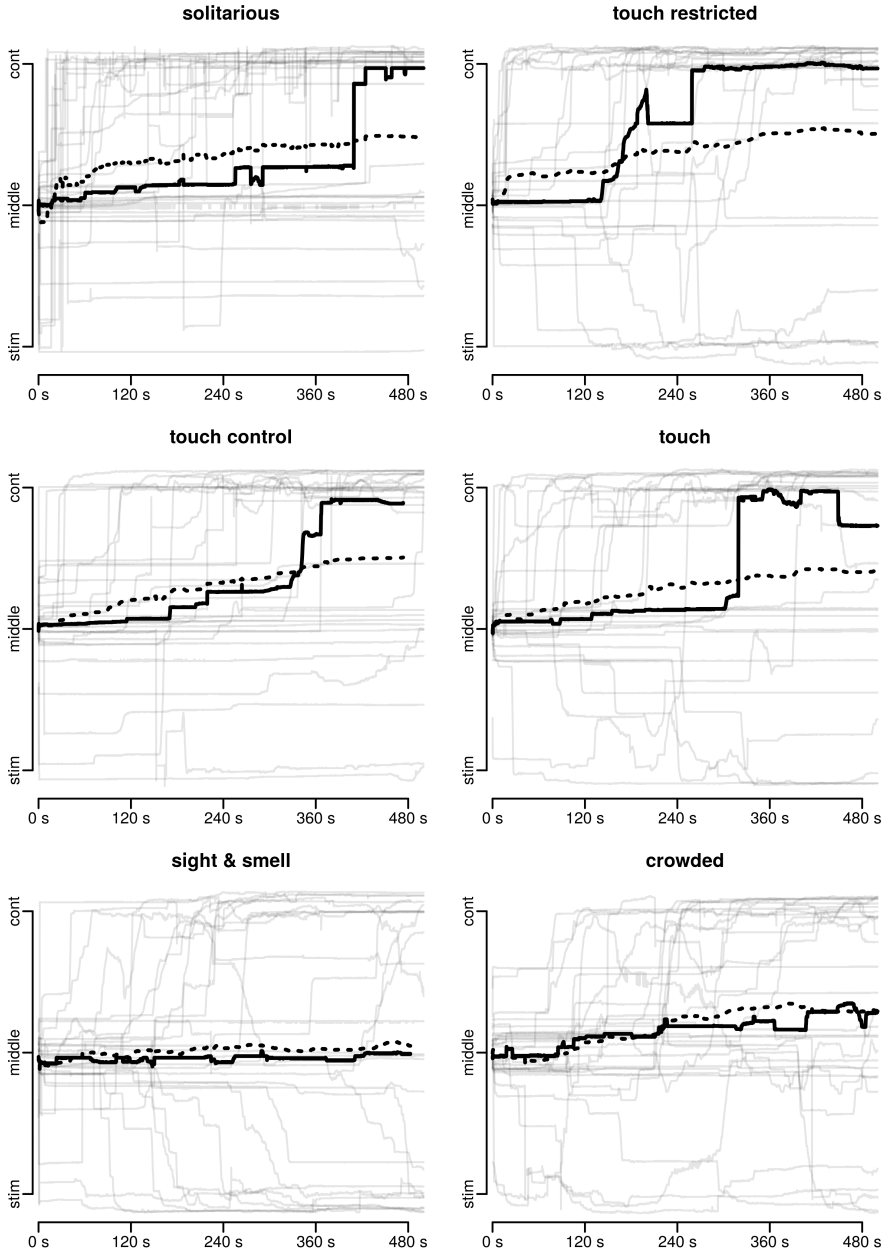



**Figure 5.8:** Distance from the stimulus group throughout the arena assay. Compared with solitarious animals in the training set in chapter 4 (fig. 4.30), the shift of the median position towards the control side happened later. Also, for the two-hours crowded animals, the median position barely shifted to the control side, while in the original crowding series (chapter 4, figure 4.30) it did after about 3 min.. None of the touch conditions shifted a lot, while the sight & smell animals were clearly visiting both sides of the arena.

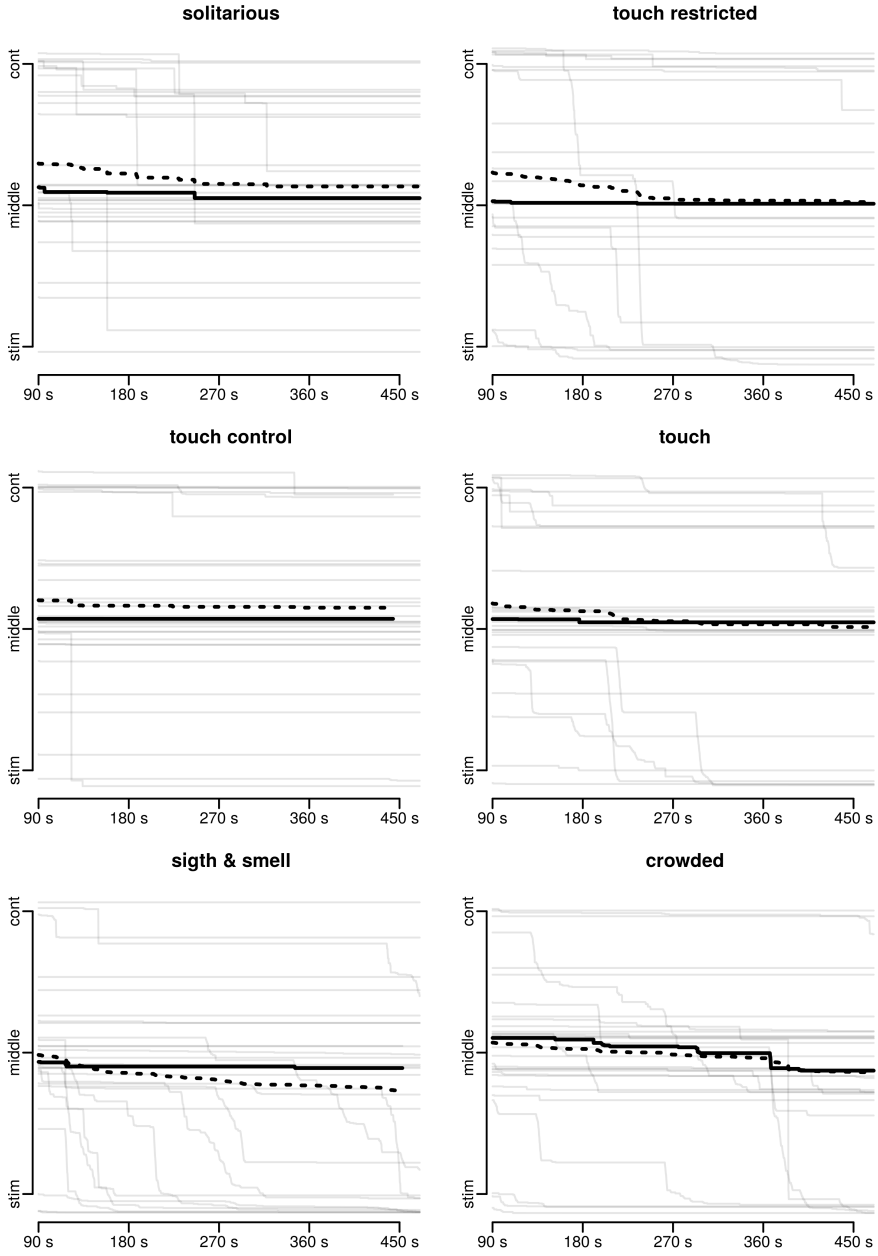



**Figure 5.9:** Cumulative closest distance to the stimulus group, starting after 90 seconds. For the solitarious animals, the median closest distance reached after 8 min. is similar to the picture in chapter 4 (fig. 4.31). Only mild effects were seen after 2 hours of sight & smell or crowding.

model 2. The solitary condition and "tc" were the only conditions that gregarized significantly less than the crowded group.

For the factor scores, signal (or statistical power) was low (see supporting materials). Hence, only trends could be observed. Factor scores for "tr", revealed high walking activity, and extremely high climbing activity, but the attraction - repulsion factor was down to solitarious levels. The "tc" animals however, were remarkably passive, but showed a slightly higher attraction (or lower repulsion).

Both for "t" and "ss", climbing and walking activity were higher than in solitarious animals, but perhaps slightly lower than the upsurge of activity seen after two hours of real crowding. Most strikingly, a drop in repulsion was seen, especially in the "ss" condition, with values similar to what is usually observed in gregarious animals. None of these differences were significant (the difference between "sa" and "ss" was significant before correction), but given my expectation that attraction/repulsion changes in the first hours of phase transition, the trends are worth mentioning.

**Table 5.2:** Adjusted p-values for pairwise differences between conditions, for the logit of $P_{greg}$ (Tukey HSD) and for factor scores (BH corrected Wilcoxon tests). * p-adj<0.05,** p-adj<0.01,*** p-adj<0.001.

| contrast | model 2 | model 5 | model RF | walking | attraction/ repulsion | climbing |
|---|---|---|---|---|---|---|
| tr-sol | *** | *** | *** | *** | | *** |
| tc-sol | | ** | * | | | |
| t-sol | *** | *** | *** | * | | |
| ss-sol | *** | *** | *** | ** | | |
| cr-sol | *** | *** | *** | *** | | |
| tc-tr | | | | ** | | * |
| t-tr | | | | | | |
| ss-tr | | | | | | |
| cr-tr | | | | | | |
| t-tc | | | | | | |
| ss-tc | *** | | * | * | | |
| cr-tc | *** | ** | ** | ** | | |
| ss-t | | | | | | |
| cr-t | | | | | | |
| cr-ss | | | | | | |

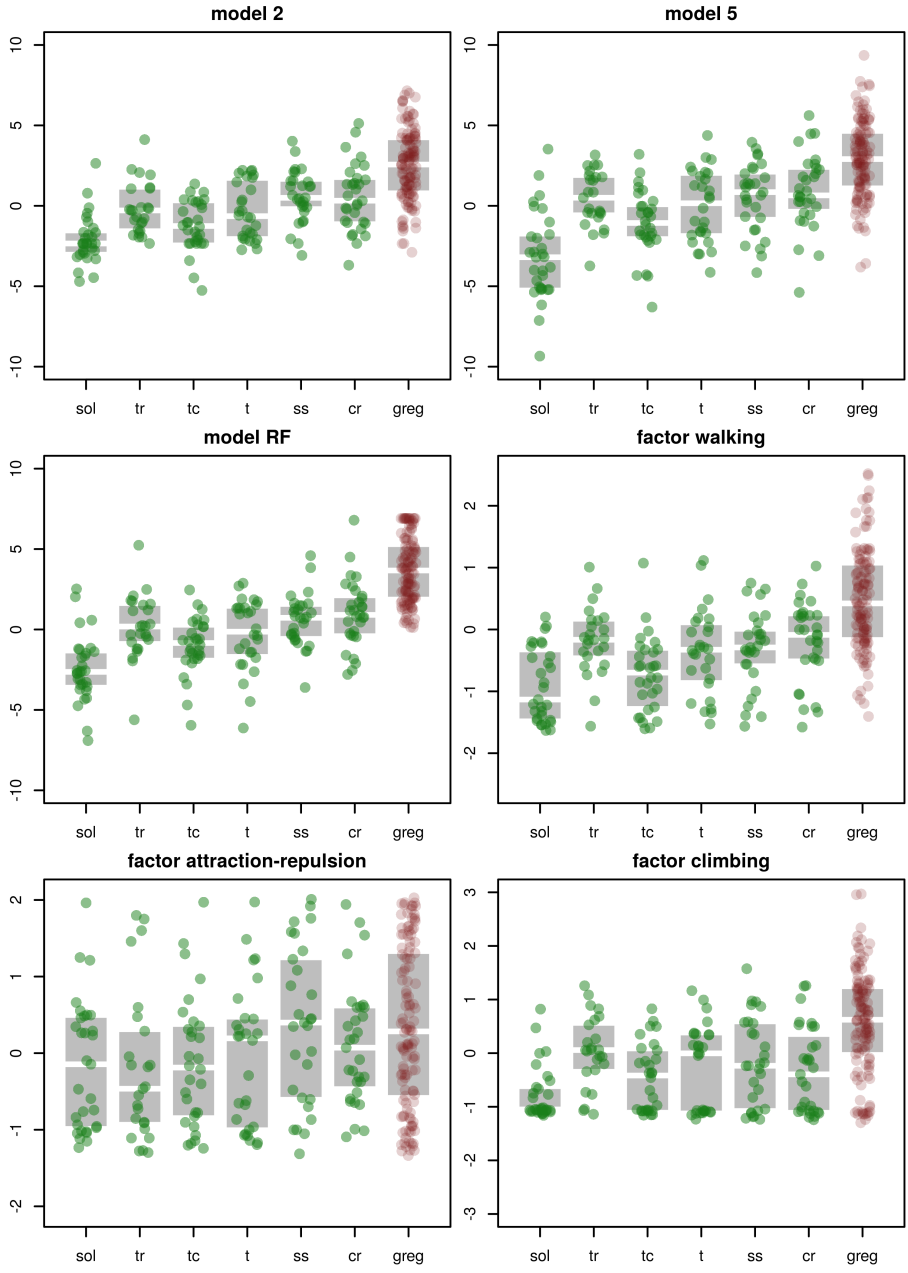



**Figure 5.10:** Box plots for 3 models ($P_{greg}$ values in logit scale), and 3 EFA factor scores for all experimental conditions. Gregarious values from the model population are given as a reference.

## 5.4.2 Transcriptomics

### 5.4.2.1 RNA quality

Concentration and quality measures depended heavily on which metric was used (fig. 5.11), but the 30 samples that were sent to BGI all met their criteria for sequencing. Detailed quality measures, both by our lab and BGI, can be found in the supporting materials.

### 5.4.2.2 Sequence quality and contamination

All samples contained between 13% and 40% (mean 18.6%) 16S ribosomal sequences. Figure 5.5 shows the mapping of a library from batch 1 onto the *Schistocerca gregaria* mitochondrial genome. It is clear that the majority maps exactly on the 16S rRNA genes, which mitigates the suggestion that the contamination would be mito-genomic. Similar figures were found for all libraries, and there was little difference between batch 1 and batch 2.

Apart from rRNA contaminations, the quality of the reads was excellent, and after filtering out reads that map to a collection of orthopteran mitochondrial and rRNA

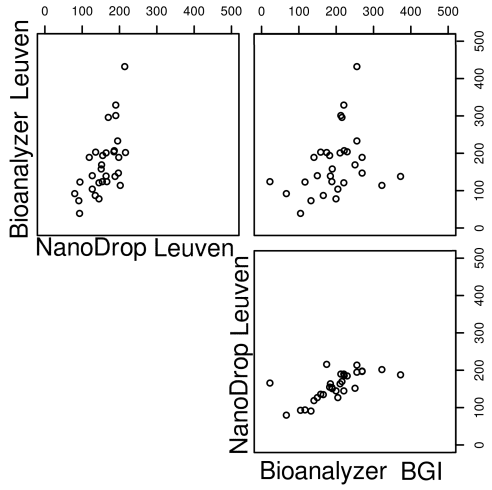



**Figure 5.11:** Concentrations of the same 30 samples as measured with Agilent 2100 Bioanalyzer and NanoDrop™ 1000 (Thermo Scientific™) in our lab, and with Agilent 2100 Bioanalyzer by BGI.

sequences, the concatenations of both batches all had more than 19.5 million clean reads.

#### 5.4.2.3 Mapping and read counting

For the first 28 samples, Between 85.4% and 87.7% of clean reads mapped to the reference at least once. The last two samples (SG32 and SG33) had slightly lower mapping rates (83.5%). Out of 21,595 reference contigs, 21,098 had at least one mapping from each sample, 16,571 at least 10, 10,275 at least 100 and 2,475 at least 1,000. Hexamerin 1 got the highest number of mappings (median 585,976) as well as the highest number of mappings per kilobase (median 29,279 fragments per kilobase of transcript per million mapped reads). Other hexamerin sequences also received many mappings, totalling about 6.5% of all counts. These numbers were also deviant in samples SG32 and SG33, with lower mapping counts for five different hexamerins. Pacifastin 1, a putative orthologue of *Periplaneta* Allergen Cr-PI (a storage protein) and a small unannotated fragment were also among the most abundant transcripts. Neither of these transcripts showed any condition-related expression pattern.

A nearest neighbour joining hierarchical clustering based on the correlation matrix of log transformed normalized counts for sets of both batches clearly showed near perfect coupling of the two batches within sample, indicating that batch 1 and batch 2 were genuine "technical replicates" (fig. 5.12). The 2 samples that contained RNA from the second round of extractions (SG32, SG33) clustered together. A similar pattern could be observed in the correlation matrices. The distribution of read counts was similar for all concatenated sets, except again for SG32 and SG33, and SG2731 to a lesser extent.

#### 5.4.2.4 Differentially expressed genes

Normalizations based on median and quantile, or on the trimmed mean of M-values (TMM) method of Robinson and Oshlack [211], all yielded quite similar scalar normalizers, and similar results after inferences. Normalization based on total counts, or on housekeeping genes only yielded different results. The latter also had the weakest match between observed and predicted q-values. The combination of the two batches provided most DEGs. In what follows, DEG is defined as found to be DE with FDR controlled at 0.1 (posterior probability > 0.9), in the combined batches, normalized with quantile, median or TMM.

EBSeq found few differentially expressed genes, and similar figures were seen with other methods. The low signal in the data was already suggested by the estimated prior probabilities: the prior for NDE after 10 iterations was > 99.8% for the two condition

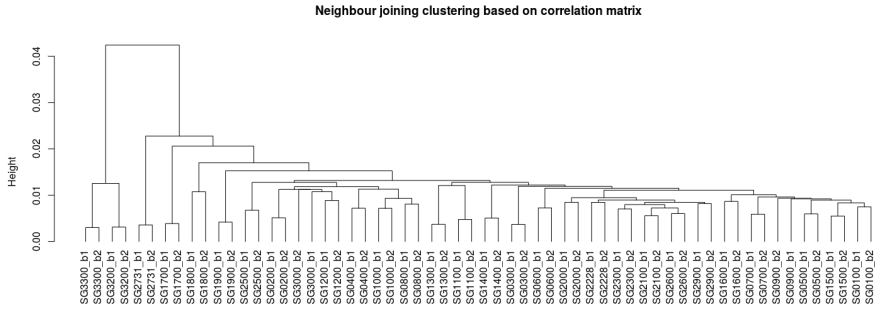



**Figure 5.12:** All libraries, both those originating from batch 1 and batch 2, were hierarchically clustered (based on the correlation matrix of log transformed normalized counts (mappings per total number of mappings), nearest neighbour joining method). Note how "technical" repeats cluster, and how the 2 deviant samples are strongly correlated to each other, but less to the other samples. Clustering was executed on sets that had already been filtered for all known *Schistocerca* mitochondrial and ribosomal sequences, and for sequences that got less than 10 mappings in each dataset. Note that the distances presented on the *Height* scale are to be interpreted as "at least", because of the neighbour joining. As an illustration, the lowest correlation in the set was 0.787, while the correlation between SG32 and SG33 was 0.987.

comparison, and 99.6% for the 5 condition comparison, regardless of normalization method used.

In a first round, only the two solitarious conditions were tested. 12 genes showed an increased expression in response to a mere visit to the arena (table 5.3, figure 5.13). The expression of these genes along all conditions showed general up-regulation compared to the "sb" condition without exception, suggesting the change was genuine. Inferences were more powerful when both batches were combined, but normalizations did not make a large difference. For the combined batches, 10 genes were DE for all 3 considered normalizations. Out of the 12 DE contigs, 8 had high throughput annotations, and some more information could be found with BLASTn. One contig was homologous to *Drosophila melanogaster ninaG* (Neither inactivation nor afterpotential protein G), which is an oxidoreductase involved in the biosynthesis of 3-hydroxyretinal, important for photo-reception [223]. 7 contigs had *cuticle-protein*-related hits, some very strong, others with only weak support. Upon manual inspection, they seemed to be 7 genuinely different transcripts, but some were hard to align with known transcripts from other insects over their entire length. 3 contigs were homologous to the molecular clock-regulated *Drosophila* gene *takeout*. These were 3 different sequences of about 1000 nucleotides length that shared some well conserved regions, suggesting they were paralogues. 1 contig had only weak resemblances to predicted or uncharacterized proteins in other insects.

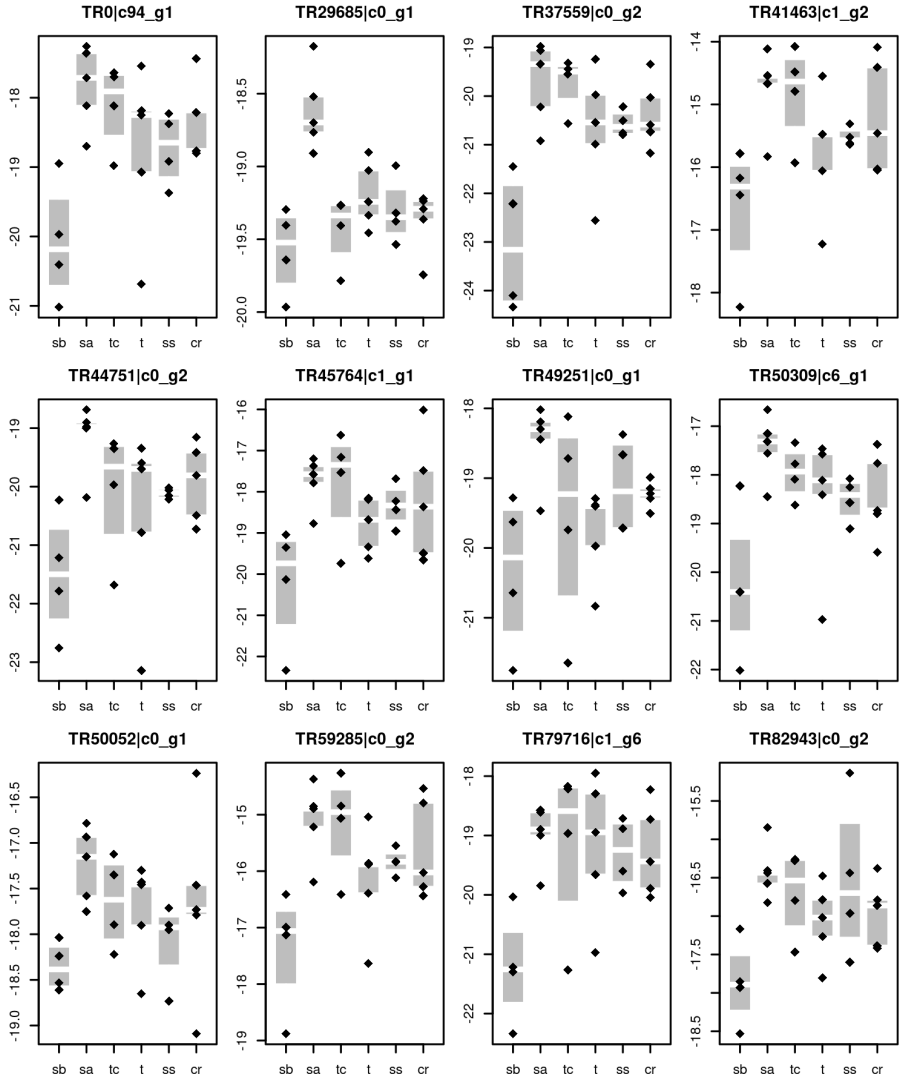



**Figure 5.13:** Boxplots of log transformed TMM-normalized mapping counts for all contigs where a significant arena effect was seen in the two solitarious conditions. Notice how in general, the other experimental conditions have higher counts than "sb", which suggests the differences are genuine.

**Table 5.3:** Posterior probabilities (p.p.) for differential expression between "sol blank" and "sol arena" for three different normalizations: quantile, median and TMM.

| Contig name | p.p. quant | p.p. median | p.p. TMM | Median counts | Annotation |
|---|---|---|---|---|---|
| $TR0\|c94_g1$ | 0.9973 | 0.9990 | 0.9991 | 48 | Protein takeout family |
| $TR29685\|c0_g1$ | 0.9352 | 0.8749 | 0.8814 | 27 | ? |
| $TR37559\|c0_g2$ | 0.9900 | 0.9937 | 0.9943 | 13 | Cuticle protein 19.8 |
| $TR41463\|c1_g2$ | 0.9453 | 0.9646 | 0.9640 | 388 | Weak hits in cuticle protein |
| $TR44751\|c0_g2$ | 0.9998 | 0.9999 | 0.9999 | 15 | Weak hits in cuticle proteins |
| $TR45764\|c1_g1$ | 0.9993 | 0.9996 | 0.9996 | 49 | Weak hits in cuticle proteins |
| $TR49251\|c0_g1$ | 0.9041 | 0.9384 | 0.9418 | 27 | Neither inactivation nor afterpotential protein G |
| $TR50309\|c6_g1$ | 0.9854 | 0.9851 | 0.9858 | 56 | Cuticle protein 18.7 |
| $TR50052\|c0_g1$ | 0.7234 | 0.8961 | 0.9125 | 81 | Protein takeout family |
| $TR59285\|c0_g2$ | 0.9835 | 0.9913 | 0.9916 | 281 | Cuticle protein 6 |
| $TR79716\|c1_g6$ | 1.0000 | 1.0000 | 1.0000 | 32 | Cuticle protein 21 |
| $TR82943\|c0_g2$ | 0.9987 | 0.9976 | 0.9975 | 155 | Protein takeout family |

**Table 5.4:** Posterior probabilities (p.p.) for differential expression upon experimental treatments for three different normalizations.

| Contig name | p.p. quant | p.p. median | p.p. TMM | Median counts | Annotation |
|---|---|---|---|---|---|
| **"Treatment stress"** | | | | | |
| $TR8281\|c0_g1$ | 1.0000 | 1.0000 | 1.0000 | 21 | Probable nuclear hormone receptor HR38 |
| $TR41522\|c0_g1$ | 0.9532 | 0.9274 | 0.9255 | 29 | ? |
| $TR47353\|c0_g1$ | 1.0000 | 1.0000 | 1.0000 | 420 | ? |
| $TR67676\|c0_g1$ | 0.9965 | 0.9986 | 0.9975 | 43 | ? |
| $TR80126\|c0_g1$ | 1.0000 | 1.0000 | 1.0000 | 296 | ? |
| $TR85424\|c0_g2$ | 0.9999 | 0.9995 | 0.9995 | 195 | ? |
| $TR86948\|c1_g1$ | 0.9996 | 0.9991 | 0.9985 | 5340 | ? |
| **"Plasticine stress"** | | | | | |
| $TR49151\|c0_g1$ | 0.9488 | 0.9368 | 0.9611 | 183 | ? |
| $TR70112\|c0_g1$ | 0.9801 | 0.9558 | 0.9723 | 90 | Cytochrome P450 18a1 |
| $TR82407\|c0_g1$ | 1.0000 | 1.0000 | 1.0000 | 165 | ? |
| $TR89121\|c0_g1$ | 0.9995 | 0.9981 | 0.9990 | 341 | Cytokine-inducible SH2-containing protein |
| **"Thoracic pathway"** | | | | | |
| $TR46101\|c0_g1$ | 0.9792 | 0.9905 | 0.9865 | 800 | ? |
| $TR65420\|c0_g1$ | 0.9827 | 0.9680 | 0.9575 | 10 | ? |

For the 5-condition tests, a set of 13 sequences was found with a posterior probability of being NDE lower than 0.1 for the combination of batch 1 and 2 with quantile, median or TMM normalization (table 5.4, figures 5.14, 5.15 and 5.16). Among these, 7 had a posterior probability higher than 0.9 to belong to the "general treatment stress" hypothesis, 4 would belong to the "plasticine stress" hypothesis and a mere two contigs had a posterior higher than 0.9 for best fitting the thoracic pathway hypothesis. Out of 13 sequences, only 3 annotations were found: *nuclear hormone receptor HR38*, *cytochrome P450 subunit 18a1*, and one putative orthologue of the vertebrate *cytokine inducible SH2 containing protein*.

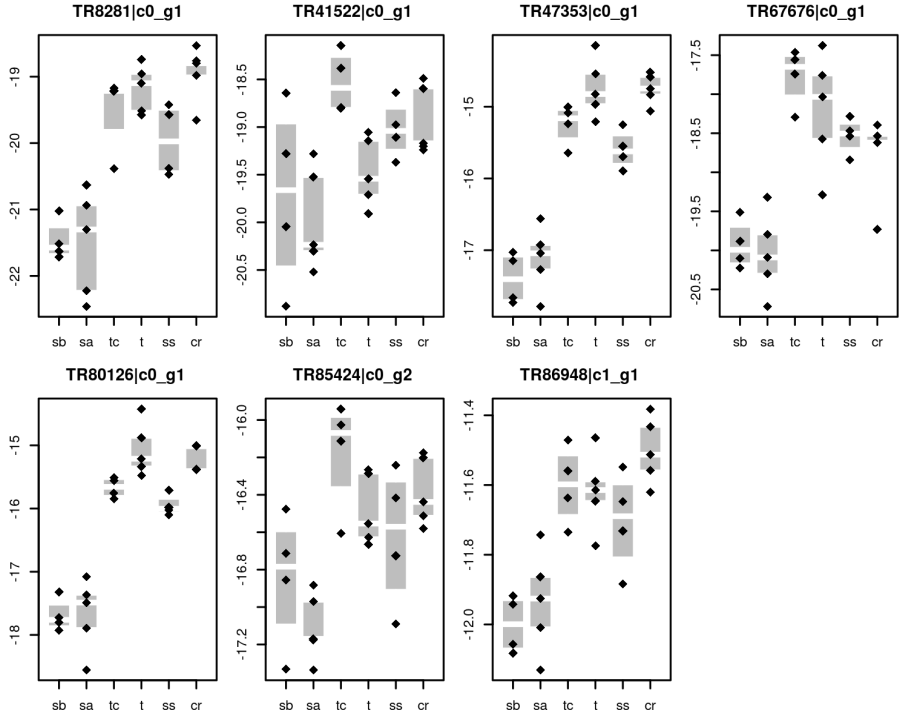



**Figure 5.14:** Boxplots of log transformed TMM-normalized mapping counts for all contigs where a significant "treatment stress" effect was found.

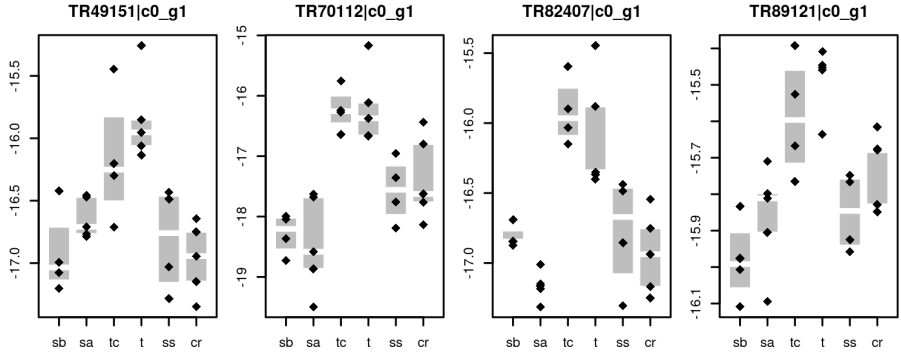



**Figure 5.15:** Boxplots of log transformed TMM-normalized mapping counts for all contigs where a significant "plasticine stress" effect was found.

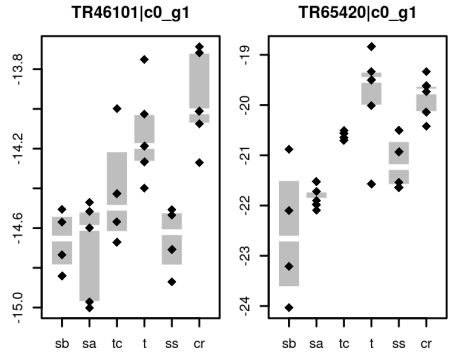



**Figure 5.16:** Boxplots of log transformed TMM-normalized mapping counts for all contigs where a significant "thoracic pathway" effect was found.

## 5.5 Discussion

### 5.5.1 Behaviour

Based on figures 5.7 - 5.9, there were minor differences between this experiment and the data presented in chapter 4; both for the solitarious and the crowded condition. In this chapter, groups were smaller, which makes the figures somewhat more erratic. Nothing changed between making recordings for chapter 4 and this experiment, and hence, although impossible to rule out, I see no good reason why there would be real differences.

The initial idea for a control for the "touch" condition was the "touch restricted" treatment, but halfway in the experiment, preliminary analysis of the behaviour showed that the "touch restricted" animals were remarkably active in the arena. Hence, the "touch control" condition was added.

According to Rogers *et al.* (2003) [218], the animals in the "touch restricted" condition should not have shown the observed shift in $P_{greg}$. A first explanation may be that the animals did partly gregarize because the coxal joint was insufficiently immobilized, allowing for movement and stretch while the animal attempted to extricate its hind legs. However, looking at the factor scores, a better explanation may be that these animals did not undergo phase transition, but rather, their $P_{greg}$ increased due to an increased propensity to walk and climb after having been immobilized.

It would be interesting to know if unusually high activity after hind leg immobilization was also observed in the original study. In an earlier study in *Chortoicetes terminifera*, it was observed that upon tickling of different body parts in solitarious animals, the

number of jumps increased together with $P_{greg}$ (with a model that did not include jumping), except for tickling of the antennae, where $P_{greg}$ increased dramatically but the animals barely jumped [63]. Both my observation of increased activity, and the observation of a disconnection between jumping and $P_{greg}$ when tactile stimuli were carefully isolated, demonstrate that under certain artificial conditions, the variables that constitute "gregariousness" can shift in multivariate space to visit positions that would not naturally occur during "canonical" phase transition. These are exactly the kind of situations where classifiers like logistic regression and random forest are of little use, and may even give spurious results without a warning.

A logistic regression would probably have allowed to discriminate between activity-only and authentic phase transition in case the treatments fully gregarized the animals, while here it seems that the observed differences were too subtle. As a result, changes in a single latent dimension were sufficient to push $P_{greg}$ into the range of short-term crowded animals. Unlike the original study [218], my treatments only took two hours, and few animals showed the big shifts in $P_{greg}$ that were observed earlier. Perhaps after longer treatments, the differences between conditions would become more pronounced. However, differences in the pace of behavioural change between colonies make it difficult to predict whether 4 hours would be sufficient for our animals and our behavioural assay. In any case, the animals in the "touch control" condition behaved as expected, with lower values for $P_{greg}$ and no obvious increase in activity. Hence, I think it was the right decision to use these animals in the RNA-Seq study.

## 5.5.2 Transcriptomics

### 5.5.2.1 Technical considerations

Throughout this multi-step procedure, decisions had to be made at numerous stages, and to permute over all alternatives would have been impossible. At some point a mapper and a count algorithm were chosen. I compared some possibilities, and chose for software tested and trusted by the community. The same is true for the packages that were used for statistical inferences, although here I had a clear preference for the Bayesian packages because of the way hypotheses were formulated, and *EBSeq* really stood out in ease of use. In what follows, I will briefly discuss some of the decisions that I made in my *EBSeq* workflow, including working on gene-level counts, choosing between normalizations, omitting 3 samples, and keeping 2 batches apart.

Although both RSEM and *EBSeq* are specifically designed for isoforms, I decided to work on gene level. The main reason was that the reference transcriptome was set-up to contain a single representative per gene. This was not really a choice, because the assembler produced very high numbers of redundant contigs, and I had little faith that all "isoforms" were *bona fide* splice variants. For some "chosen" genes, I

did include some manually curated splice variants, so in theory I could as well have worked on isoforms. However, in my pilot tests, *EBSeq* was never able to make use of the full range of isoform-specific possibilities, because the number of genes with multiple splice variants was too small, making it impossible to properly estimate the beta parameters exactly for those genes that I found most interesting.

Some contigs were clearly DE regardless of normalization while in other cases normalization was critical to pick up the signal (be it genuine or not). The most commonly used normalizations all yielded very similar normalizers, and hence similar end results. This phenomenon is described in literature [232]. The normalization based on housekeeping genes did not perform well, with weaker correspondence between fitted and observed q values, and with the least overlap in detected DEGs compared to other normalizations. It was well worth a try, but I think the results obtained with this normalization can be ignored. All the more because the housekeeping genes described in Van Hiel *et al.* (2009) [268] are involved in processes that have been described to be different between long term phases, or even to change during phase transition, such as metabolism [136, 49], cytoskeleton [136] and the ubiquitin cycle [14].

It is hard to say what makes the 2 deviant samples so heavily correlated with each other while clearly different from all others. Samples SG32 and SG33 showed distributions of read counts different from all others, and they also had the highest levels of ribosomal sequences. Even after filtering out ribosomal and mitochondrial sequences, they had lower mapping rates. Another remarkable observation was their much lower relative expression of hexamerins. Sample SG2731 had intermediate levels of hexamerins, and in the neighbour joining tree it came out as a sister group to all other samples except for SG32 and SG33.

The origin of the peculiarities of SG2731, SG32 and SG33 may have been merely biological, because these samples were taken at a later stage, with animals that had been reared in the old facility. This may even explain the intermediate nature of sample SG2731, which contained both old and fresh material. However, RNA was also extracted separately, so differences of technical origin cannot be excluded.

I suspect that there was an undetermined contaminant in all samples, that was correlated with the amount of ribosomal RNA. This would explain the lower mapping rates of the "cleaned" SG32 and SG33 sets, and it may also explain how the first principal component of all genes correlated with the amount of 16S rRNA. The identity of this contaminant remains an enigma. I looked into contigs with large contributions to the first principal component, tRNA and potential undetected rRNA, but could not find it. It may be any of those, or something else.

Whether the observed differences were reason enough to omit the samples from the analyses all together is open for debate. I have run pilot analyses many times,

excluding no samples, only SG32 and SG33, and excluding all three. Generally, these samples were harder to normalize, yielding many more outliers, or causing other outliers to remain unnoticed because variance was over-estimated. Looking at the mapping coverage graphs, I am quite sure that some of the DEGs that we found would not have been DE if those 3 samples had been kept on board, but at the end of the day one has to make decisions and settle with the results. The final argument to omit them, was the fact that their deviation from the rest of the samples could be explained by their history, and I already had a suspicion that the environment in which the animals were reared had an influence on their readiness to undergo behavioural change.

The sequencing of two batches was never my intention, and my decision to do all analysis in parallel on batch 1 only, and on the sum of counts in both batches, was based on a pilot analysis where both batches were concatenated before mapping. In this analysis, there seemed to be something peculiar about batch 2: the relative proportion of reads originating from this batch correlated strongly with the first principal component, even after all ribosomal sequences were filtered out. In hindsight I think this was an indirect correlation, through the relationship between 16S rRNA and an enigmatic contaminant. I did not find any other batch effect, despite thorough screening, so I think it is safe to just add counts of both batches. In an ideal world, I would repeat my analysis starting from mapping on the concatenations of both batches, because in some (hopefully rare) cases, RSEM would be slightly better informed about the most probable origin of multi-mappers, especially for those samples where batch 2 only contained a couple of millions of reads.

For 3 contigs, detection of differential expression was quite clearly caused by outliers. In these cases, there were always two samples that showed remarkably higher expression than all others (orders of magnitude difference, see figure in supporting materials). The reason why these contigs escaped outlier filtering, is because two samples showed this behaviour, blowing up the standard deviation. The fact that these genes were considered to be differentially expressed by *EBSeq* is worrisome, and the reason why this happened is a riddle to me. Furthermore, the reason why some contigs show homogeneous mapping densities up to hundreds of mappings per position over the length of thousand and more nucleotides by only a small subset of samples is puzzling, especially because each sample is a pool of six animals. I think it is unlikely that these patterns are merely caused by genetic variation, because in such case one would expect local drops in mapping density rather than dramatic differences over the entire length of the contig (although I cannot tell for sure whether RSEM may cause artefacts in case two allelic variants are present in the reference as two separate contigs). Another possibility is that they are contaminants, but in that case they should independently have been present in the reads that made up the reference as well. Unfortunately the 3 examples that I found had little or no annotation information, if they were contaminants, it would be difficult to guess their identity. They may come from non-brain locust tissue, or they may come from feed

or microorganisms.

#### 5.5.2.2 Biological signal

My behavioural data and earlier studies have shown that not just any stimulus can induce behavioural gregarization [213, 218]. It was even recorded that intense non-specific tactile stimulation may induce a partial reversal of recently acquired gregarious behaviour [99]. Hence the question remains why such a high proportion of the DEGs that were detected in this work, were positive in the control condition as well.

Since earlier studies did not contain such controls, it is impossible to tell whether this observation is unique to the experimental setting. However, the entanglement of early phase transition with a range of biological processes was reflected in earlier transcriptomic experiments in *Locusta* [280, 109] and *Schistocerca* [12]: genes that were observed to respond to gregarizing stimuli may be involved in processes that are related to altered sensory input in general rather than to phase transition. These include genes coding for products involved in chemosensory processing, photoreception, and signal transduction.

Another question is why overall, only few genes showed statistically significant differential expression. A first explanation may be a lack of statistical power. The variation in gene expression far exceeded mere "shot noise" (in other words, there was high overdispersion), but this is expected for RNA-Seq data, and four or five replicates are really common for transcriptomics studies. Perhaps the comparison of four conditions that were only subtly different from each other may have been an ambitious question, but even if just the two most extreme conditions (sol blank versus crowded) were compared without looking at any of the other samples, only a dozen DEGs were detected. This suggests that if the study was underpowered, it was not because of the complex design, but rather because of the multitude of tests and the large variance.

I do not believe that this study was underpowered. Even if differences were generally subtle and hence hard to individually confirm, patterns of differential gene expression would surface after cancelling out other sources of variation. However, samples did not cluster according to treatment in the space defined by any two among the first dozen principal components.

A better explanation may be that I overestimated the transcriptomic response to the treatments within the time frame of two hours. Perhaps dramatic changes should not have been expected: in earlier microarray-based studies only a limited number of genes responded to short periods of crowding. In *Locusta*, changes in the expression of 93 genes (uncorrected $p < 0.001$) were reported after 1 hour of crowding [109]. In

*Schistocerca* [12], 48 genes were found to be DE after 2 hours, despite the fact that samples came from two colonies. In these studies, the limited number of DEGs is not just a power issue, because similar studies contrasting long term phases report hundreds of DEGs [165, 14]. Direct comparison between this current study and the earlier microarray studies is difficult, because they were technically very different, and their samples contained more diverse tissues (entire CNS or heads).

In a way it seems inconceivable that so little would happen upon such impacting treatments. However, perhaps a large part of the physiological changes in the brain during early gregarization are independent of transcription. Fast responses do require proteins such as protein kinases, that are already present and only need to be switched on [187]. In 2010, Geva *et al.* demonstrated that a 30-minute crowding could induce behavioural gregarization that persisted even after 24 hours of re-isolation. This memory formation was sensitive to the protein synthesis blocker anisomycin [99]. Later, work by Stettin and colleagues suggested that acquired gregarious behaviour (2 hours crowding) was resistant to cold-shock anaesthesia [243], suggesting that at least some stable form of memory was formed during the first hours of crowding. However, they also found that the earliest hours of behavioural change are relatively robust against treatment with blockers of transcription (actinomycin D or alpha amanitin) or translation (anisomycin) [243].

In this work, I certainly did not prove that early behavioural change is not conditional on transcription. First, transcriptional responses may just be happening later on in the process of behavioural change, and the DEGs that were found may just be part of a smaller, initial response. Second, only one tissue was sampled on a single time point. Though the brain is the obvious candidate for memory formation, upstream processes may be happening elsewhere in the nervous system, or even at the level of sensory organs, as was seen in *Locusta migratoria* [109]. It may even be that some regions of the brain show strong transcriptional responses that are diluted by stable expression in other regions. As pointed out earlier, the insect brain is a complex structure with regions of different ontological origin and function. For example: microarray [231], RNA-Seq [178] and *in situ* hybridization [178] studies in honeybees have shown that different brain regions exhibit drastically different transcriptional responses to dance communication or the presentation of different kinds of food.

#### 5.5.2.3 Noteworthy findings

The most noteworthy findings, I would argue, are the effects of the arena assay on the expression of multiple *takeout* homologs and *cuticle proteins*.

The family of insect *cuticle proteins* (*CP* genes, InterPro family IPR000618) is conserved in different arthropod lineages [46, 212] and share stretches of hydrophobic amino acids and domains that are known to bind chitin [205]. The only gene ontology

term assigned to this protein family is the molecular function "structural constituent of cuticle". *CP* gene expression is generally activated by a pulse of ecdysteroids, and is most pronounced in epidermis and imaginal disks [46]. We found 7 different paralogues to be DE in the brain. It is hard to exclude contamination with cuticular tissues, but based on their mapping depth, some detected *CP* genes actually showed rather high expression. Cuticle proteins were also found to react to crowding and isolation in *Locusta migratoria* [109] but in this microarray study entire heads were used, so it is not known whether these genes were also expressed in the brain. The only report I found about expression of a *CP* gene in the CNS, was a study in the bee *Apis cerana*, where a single *CP* gene was found to be strongly expressed in the midgut, and weakly in the brain [52].

*Takeout* belongs to the *tubular lipid-binding* (*TULIP*) gene superfamily, homologous to a range of soluble ligand binding proteins including juvenile hormone binding proteins [141, 2]. In the honeybee genome, there are 8 putative genes in the *takeout/juvenile hormone binding protein* family, but their phylogeny is poorly resolved [113].

The *takeout* gene was originally described in *Drosophila melanogaster* as a circadian clock-regulated output gene involved in the integration of circadian rhythm and food status information [224]. In a variety of insects, *takeout* is involved in development and longevity on the one hand, and in the modulation of social or migratory behaviour on the other hand. In the fruit fly, it is thought to be one of the downstream effectors of lifespan extension through dietary restriction [19, 111], associated with altered JH signalling [44]. In the honey bee, the expression of at least one gene in the *takeout/juvenile hormone binding protein* family (GB19811) also has an age-dependent expression and responds to administered JH. In the burying beetle *Nicrophorus vespilloides*, a species with varying degrees of parental care, *takeout* is down-regulated during parenting and nursing of the larvae [189]. In the monarch butterfly *Danaus plexippus*, *takeout* is up-regulated in the summer morph compared to the migratory morph. In the termite *Reticulitermes flavipes*, knockdown of the *deviate* gene, a member of the *takeout/juvenile hormone binding protein* family, resulted in disrupted trailing behaviour without interfering with feeding, indicating a specific role in the social aspect of foraging [229].

In locusts, genes from the *takeout* family also play a role in the modulation of behaviour. In *Locusta migratoria*, 4 *takeout* genes were down-regulated upon crowding, especially in peripheral tissues such as antennae [109]. Moreover, upon knockdown of *LmigTO1*, solitarious animals were relatively less repelled by a stimulus group [109]. In their review on molecular mechanisms of phase transition, Wang and Kang propose that *takeout* is involved in the initiation of short-term phase change, probably through the modulation of olfactory perception [281].

We now find three putative *takeout* paralogues that are DE upon a 10 min. confrontation with hoppers in the arena. At first sight this would seem consistent

with the role in the early modulation of behaviour seen in *Locusta*, if it were not for the fact that in the case of *Schistocerca* an up-regulation was observed.

I speculate that events upstream of the integration of information from different sensory origins are evolutionary labile, and the co-optation of similar molecular components into these early processes may have independent (parallel) evolutionary histories. Different locust species use different proxies to monitor local population density, and the intensity of stimuli needed for gregarization or solitarization may vary as well. Perhaps *takeout* has properties that are most desirable in the fastest of both transitions, which is solitarization in *Locusta*, and gregarization in *Schistocerca*.

## 5.6 Acknowledgements

The idea of transcriptionally characterizing the different sensory pathways was originally coined by Swidbert Ott. Special thanks go to Michiel Holtof for his help in the practical management of the experiment, and for dutifully tickling the locusts together with his stand-ins Darron Cullen and Leen Van Lishout. All heavy duty computing was done on a server at the university of Leipzig that I was kindly granted access to by Katja Nowick and colleagues.

# Chapter 6

# Conclusions and perspectives

> We can only see a short distance ahead, but we can see plenty there that needs to be done.
>
> Alan M. Turing

## 6.1 The new desert locust reference transcriptome is a valuable resource

The new reference transcriptome assembly has proven to be a good resource of sequence information. The sequences of most expected neuropeptide precursors were recovered, and the selection of 21,595 contigs that served as a reference for my RNA-Seq experiment provided good coverage of the observed brain transcriptome, with 85 - 87 % of the reads mapping at least once. The number of annotated contigs, and the amount of annotation information per contig were also a lot higher than in earlier work, and are now comparable to other recent transcriptome studies in orthopterans [16, 289, 4, 275, 23, 137, 291, 201, 200, 132].

At the time of writing, several studies that relied on sequences found in one of the transcriptome assemblies discussed in this work, have already been published [3, 287, 73, 278, 150] or are in press. The newly discovered neuropeptide precursors are also a welcome addition to what was already known from earlier sequencing projects, and we are currently validating the predicted peptides with MS.

Since contig filtering (from 175,066 to 21,595 contigs) was based on the expression in $5^{th}$ instar brains, the assembly of 175,066 contigs undoubtedly contains additional contigs corresponding to genes that were not expressed in the brain, or to splice variants that were not selected as a reference. In future experiments focusing on different tissues, the 175,066 contigs can serve as a basis for other tissue specific transcriptomes, applying variations of the filtering strategy that I described in chapter 3.

In the selected set, 10,276 out of 21,595 contigs were annotated, leaving 11,319 contigs without annotation. Many of these are clearly protein coding, but show no obvious similarity to annotated (endopterygote) insect proteins. This suggests they may be specific to exopterygotes or to orthopterans. It will never be achievable, nor desirable, to experimentally determine the function of every transcript in every organism, but I do think there is a continued need for thorough functional characterization of gene products in a variety of organisms. Especially in poorly covered taxonomic branches, such as many exopterygote insect orders, new high quality annotations can serve as beacons for annotations in related organisms. In this light, some of the 11 unannotated DEGs described in chapter 5 may provide candidates for new lines of research.

Grasshoppers are known for their large genomes, and one of the driving mechanisms behind their genome expansion is thought to be high transposon activity [285, 281]. A large number of transposons were found in my assemblies as well (see supporting materials), but their identification and characterization were beyond the scope of this work. However, it would be interesting to know more about their diversity, and their potential role in phase transition. Large scale differences in the expression of

RNA-transposons between long-term phases in *Locusta migratoria* were suggested by Wei *et al.* (2009) [285], but their claims were based on a single replicate per phase.

## 6.2 Contributions to a quickly developing discipline

With the increased availability of high-throughput sequencing, transcriptomics experiments are now within reach for projects in non-model organisms. However, many of the methods in the field are still being developed, and there is room for improvements on many levels. In the methods presented in chapter 3, I used a pipeline that combined my own scripts with readily available tools. I do not claim that the trimming-merging-trimming approach is a revolutionary contribution, but it has turned out to work for this particular case, where small insert sizes caused an unusually high proportion of reads to overlap.

A recurring challenge when using bioinformatic applications is to choose optimal parameter settings. The estimation of overlap proportions prior to read merging has helped setting the parameters for the program FLASH, optimized for each individual input set of reads. Though I can think of better (fully probabilistic) methods that combine quality filtering and read merging, this is currently the best way to optimize parameters for FLASH.

Filtering contigs from large assemblies has been described as *more of an art than a science* by the developers of Trinity, which is problematic. Hence, I think that my filtering pipeline may be a valuable contribution to the field of transcriptomics, since it is not only based on mapping coverage, but also on sequence content.

## 6.3 The transcriptomic response to two hours of crowding is weak and non-specific

Nothing in my entire PhD project has taken more discussion and contemplation than the design of the experiment that I described in chapter 5. It is therefore especially painful that so little biological signal was picked up. The expectation was to find genes whose expression would swiftly respond to the different crowding regimes, but in the end, not a single annotated transcript was found to respond to treatments but not to controls. We already knew from earlier studies that responses to crowding [12, 109] are less pronounced than differences between long-term phases [165, 14]. I hypothesized that by looking at a more specific tissue, using a more precise technique, more DEGs would be recovered, but the opposite was true. In the discussion of chapter 5 I argue that low statistical power does not sufficiently explain the low

numbers of DEGs, and that my findings suggest genuinely low levels of differential gene expression.

## 6.4 Concerns about earlier studies

Although in this study, few targets for future research were found, my observations do put the literature on transcriptomics in locusts into perspective.

A first concern is that the only two real quantitative transcriptomics studies focusing on the process of phase transition [12, 109] did not include an experimental condition that controlled for processes that are unrelated, or only indirectly related to behavioural gregarization. In this present study, I show that the majority of transcripts that respond to crowding, or to crowd-mimicking regimes, also respond to such a control condition. This observation raises the question whether the DEGs found in earlier studies [12, 109] were part of a core molecular pathway leading to behavioural phase transition, or whether they were involved in peripheral processes such as the processing of altered sensory input, or stress reactions caused by manipulating the animals. As noted earlier, the molecular mechanisms of stress response and early phase transition are intertwined, and perhaps the kind of stress presented to the animals in this experiment is not very relevant under natural conditions. However, this does not justify the absence of control conditions in earlier experiments, and I argue that in future experiments, controls should again be included, and confirmed with high quality behavioural assays.

A second concern is that in the past, studies have been published that make claims about differential gene expression based on very limited evidence. Having seen and experienced the technical difficulties and statistical challenges typical of this kind of analysis, I find it worrisome that not a single publication on differential gene expression in locusts contains supporting materials (data and methods) that allow to repeat the analysis. In this present work, I used a false discovery rate cut-off of 0.1 (like Badisco *et al.* (2011) [14]), but the transcriptomic studies in *Locusta* do not mention if and how inferences were corrected for multiple testing. This may mean that if these datasets were re-analyzed by my statistical standards, fewer DEGs would be recovered. Furthermore, having seen the large variation in the expression of a single gene, even within a condition, I am especially sceptic towards those studies that did not apply biological replication [136, 285, 49, 281]. I do not claim that none of the reported DEGs are genuine, but I expect the proportion of false positives to be high.

## 6.5 An R-based behavioural assay

Building on a tradition of high quality studies [215, 242, 243, 217], I successfully implemented an arena-based behavioural assay with automated video tracking and classification of potentially phase-related behaviours. I modified the existing pipeline to the needs of our lab, and transferred it (almost) entirely to the R environment, which should make it more accessible for future users.

The collection of video-tracks can be a valuable resource for future behavioural studies, both as a study object or as baseline observations for novel experiments. It would be interesting to combine our videos with the video collections from the university of Cambridge and the university of Leicester, that were recorded in a comparable setting. In this way we could get a better idea about which observations generalize best, and which observations may depend on the specific arena or colony.

## 6.6 $P_{greg}$ does not say it all

For long, phase has been considered a "syndrome" [237, 240, 195, 64], which is reflected in the common use of a single aggregate measure to quantify the phase state of a locust [215]. Though the use of $P_{greg}$ has been of great utility to phase transition research, the case of the "touch restricted" condition shows that situations may occur where $P_{greg}$ does not provide an adequate measure of behavioural phase state. In this particular case, the culprit was the decoupling of two (latent) variables that are usually assumed to change together, leading to an uncommon combination of high activity without loss of repulsion.

In order to justify the use of $P_{greg}$ to quantify the behavioural phase state of intermediates between two polarities (to which the model is calibrated), behavioural phase change should really entail a series of concerted changes in activity and attraction/repulsion along a gradient between the extreme phase states. Recent studies have demonstrated that the transient gregarious-like state is physiologically and molecularly distinct from the long-term gregarious phase state [235, 220, 172], which supports the idea that during gregarization, behaviour does not necessarily follow a gradient of intermediates between long term phases.

Differences in the rate of change of activity and attraction/repulsion during phase transition have been reported in *Locusta migratoria* [109]. We know that in *Schistocerca*, traits such as colouration, melanization and gregarious behaviour are under independent physiological control and each change at their own pace, but also after my experiments it remains unclear whether this is equally true for different aspects of gregarious behaviour. Rogers *et al.* (2014) [217] demonstrated that in earlier

studies [234, 218, 5] changes in $P_{greg}$ as a response to gregarizing treatments could not be explained by a mere increase in locomotion, and that behavioural variables such as grooming frequency and attraction towards other locusts changed in concert with increasing activity. Also in my experiments it seems that both activity and attraction/repulsion change during the early hours of crowding, albeit without being correlated. Though these observations are reassuring for the credibility of earlier studies, they do not guarantee the reliability of $P_{greg}$ in future experiments.

From a pragmatic point of view, the distillation of the behaviour into a single measure like $P_{greg}$ may remain part of future analyses (if only for the sake of continuity with earlier literature), as long as the different aspects of behaviour are (seemingly) coupled and their rates of change are comparable, as long as no experimental condition influences (disrupts) their coupling, and as long as the advanced *transiens* behavioural phase state is close enough to the real gregarious phase state as far as the observed behaviours are concerned. This would, however, require a series of thorough assumption checks.

From a conceptual point of view however, I think we need to move away from the "syndrome" paradigm and recognise the potential multidimensionality of phase-related behaviours, especially during the transition. Accordingly, we should use multivariate methods that do not only allow to compare groups of individuals according to their main behavioural dimensions, but also allow to quantify the extent of coupling (correlation) between the different phase-related behaviours during phase change under various experimental conditions. Such a framework would also be of much more use in the exploration of phase-related behaviour in other (potential) locust species.

# References


[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, Oct. 1990.

[2] V. Alva and A. N. Lupas. The TULIP superfamily of eukaryotic lipid-binding proteins as a mediator of lipid sensing and transport. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, 1861(8):913–923, Aug. 2016.

[3] O. Andersson, L. Badisco, A. H. Hansen, S. H. Hansen, K. Hellman, P. A. Nielsen, L. R. Olsen, R. Verdonck, N. J. Abbott, J. Vanden Broeck, and G. Andersson. Characterization of a novel brain barrier ex vivo insect-based P-glycoprotein screening model. *Pharmacology Research & Perspectives*, 2(4):e00050, Aug. 2014.

[4] J. A. Andrés, E. L. Larson, S. M. Bogdanowicz, and R. G. Harrison. Patterns of Transcriptome Divergence in the Male Accessory Gland of Two Closely Related Species of Field Crickets. *Genetics*, 193(2):501–513, Feb. 2013.

[5] M. L. Anstey, S. M. Rogers, S. R. Ott, M. Burrows, and S. J. Simpson. Serotonin mediates behavioral gregarization underlying swarm formation in desert locusts. *Science (New York, N.Y.)*, 323(5914):627–30, Jan. 2009.

[6] S. W. Applebaum and Y. Heifetz. Density-dependent physiological phase in insects. *Annual review of entomology*, 44(159):317–341, 1999.

[7] G. Ariel and A. Ayali. Locust Collective Motion and Its Modeling. *PLOS Computational Biology*, 11(12):1–25, 2015.

[8] G. Ariel, Y. Ophir, S. Levi, E. Ben-Jacob, and A. Ayali. Individual pause-and-go motion is instrumental to the formation and maintenance of swarms of marching locust nymphs. *PLoS ONE*, 9(7), 2014.

[9] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M.

Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, May 2000.

[10] A. J. Ayme-Southgate, R. J. Southgate, R. A. Philipp, E. E. Sotka, and C. Kramp. The Myofibrillar Protein, Projectin, is Highly Conserved Across Insect Evolution Except for Its PEVK Domain. *Journal of Molecular Evolution*, 67(6):653–669, Dec. 2008.

[11] M. A. O. Babah and G. A. Sword. Linking Locust Gregarization to Local Resource Distribution Patterns Across a Large Spatial Scale. *Environmental Entomology*, 33(6):1577–1583, Dec. 2004.

[12] L. Badisco. *Study of the neurohormonal and molecular regulation of reproduction and phase transition in the desert locust, Schistocerca gregaria.* May 2010.

[13] L. Badisco, J. Huybrechts, G. Simonet, H. Verlinden, E. Marchal, R. Huybrechts, L. Schoofs, A. De Loof, and J. Vanden Broeck. Transcriptome analysis of the desert locust central nervous system: production and annotation of a Schistocerca gregaria EST database. *PloS one*, 6(3):e17274–e17274, Jan. 2011.

[14] L. Badisco, S. R. Ott, S. M. Rogers, T. Matheson, D. Knapen, L. Vergauwen, H. Verlinden, E. Marchal, M. R. J. Sheehy, M. Burrows, and J. Vanden Broeck. Microarray-based transcriptomic analysis of differences between long-term gregarious and solitarious desert locusts. *PloS one*, 6(11):e28110–e28110, Jan. 2011.

[15] K. A. Baggerly, L. Deng, J. S. Morris, and C. M. Aldaz. Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics*, 19(12):1477–1483, Dec. 2003.

[16] N. W. Bailey, P. Veltsos, Y.-F. Tan, A. H. Millar, M. G. Ritchie, and L. W. Simmons. Tissue-Specific Transcriptomics in the Field Cricket Teleogryllus oceanicus. *G3: Genes|Genomes|Genetics*, 3(2):225, Feb. 2013.

[17] W. Bao, K. K. Kojima, and O. Kohany. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6:11, 2015.

[18] P. Bateson and K. N. Laland. Tinbergen's four questions: an appreciation and an update. *Trends in Ecology & Evolution*, 28(12):712–718, Dec. 2013.

[19] J. Bauer, M. Antosh, C. Chang, C. Schorl, S. Kolli, N. Neretti, and S. L. Helfand. Comparative transcriptional profiling identifies takeout as a gene that regulates life span. *Aging*, 2(5):298–310, May 2010.

[20] S. Bazazi, J. Buhl, J. J. Hale, M. L. Anstey, G. A. Sword, S. J. Simpson, and I. D. Couzin. Collective Motion and Cannibalism in Locust Migratory Bands. *Current Biology*, 18(10):735–739, May 2008.

[21] Y. T. Belayneh. Acridid pest management in the developing world: a challenge to the rural population, a dilemma to the international community. *Journal of Orthoptera Research*, 14(2):187–195, Dec. 2005.

[22] S. Bennett. Solexa Ltd. *Pharmacogenomics*, 5(4):433–438, June 2004.

[23] E. L. Berdan, T. Blankers, I. Waurick, C. J. Mazzoni, and F. Mayer. A genes eye view of ontogeny: de novo assembly and profiling of the Gryllus rubens transcriptome. *Molecular Ecology Resources*, 16(6):1478–1490, Nov. 2016.

[24] E. L. Berdan, J. Finck, P. R. Johnston, I. Waurick, C. J. Mazzoni, and F. Mayer. Transcriptome profiling of ontogeny in the acridid grasshopper Chorthippus biguttulus. *PLOS ONE*, 12(5):e0177367, May 2017.

[25] K. Berthier, M.-P. Chapuis, S. J. Simpson, H.-J. Ferenz, H. Kane, M. Chérif, L. Kang, A. Lange, S. R. Ott, k. W. Rodenburg, S. M. Rogers, B. Torto, J. Vanden Broeck, J. J. van Loon, and G. A. Sword. Laboratory populations as a resource for understanding the relationship between genotypes and phenotypes: A global case study in locusts. *Advances in Insect physiology*, 39(10), 2010.

[26] Betts, Monica. The behaviour of a pair of Great Tits at the Nest. *British Birds*, 1955.

[27] E. Blanco, G. Parra, and R. Guigó. Using geneid to Identify Genes. In A. D. Baxevanis, D. B. Davison, R. D. Page, G. A. Petsko, L. D. Stein, and G. D. Stormo, editors, *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., Hoboken, NJ, USA, June 2007.

[28] B. Boerjan, F. Sas, U. R. Ernst, J. Tobback, F. Lemière, M. B. Vandegehuchte, C. R. Janssen, L. Badisco, E. Marchal, H. Verlinden, L. Schoofs, and A. De Loof. Locust phase polyphenism: Does epigenetic precede endocrine regulation? *General and comparative endocrinology*, 173(1):120–8, Aug. 2011.

[29] M. Boersma, P. Spaak, and L. De Meester. Predator-Mediated Plasticity in Morphology, Life History, and Behavior of Daphnia : The Uncoupling of Responses. *The American Naturalist*, 152(2):237–248, Aug. 1998.

[30] A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, Aug. 2014.

[31] T. Booth, M. Bicak, H. S. Gweon, D. Field, and E. Afgan. Bio-Linux as a tool for bioinformatics training. pages 578–582. IEEE, Nov. 2012.

[32] A. BouaïChi, P. Roessingh, and S. J. Simpson. An analysis of the behavioural effects of crowding and re-isolation on solitary-reared adult desert locusts (Schistocerca gregaria) and their offspring. *Physiological Entomology*, 20(3):199–208, Sept. 1995.

[33] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, Apr. 2016.

[34] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[35] M. R. Brent. How does eukaryotic gene prediction work? *Nature Biotechnology*, 25(8):883–885, Aug. 2007.

[36] J. Buhl, D. J. T. Sumpter, I. D. Couzin, J. J. Hale, E. Despland, E. R. Miller, and S. J. Simpson. From disorder to order in marching locusts. *Science (New York, N.Y.)*, 312(5778):1402–1406, June 2006.

[37] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268(1):78–94, Apr. 1997.

[38] M. Burrows, S. M. Rogers, and S. R. Ott. Epigenetic remodelling of brain, body and behaviour during phase change in locusts. *Neural Systems & Circuits*, 1(1):11, 2011.

[39] J. P. M. Camacho, F. J. Ruiz-Ruano, R. Martín-Blázquez, M. D. López-León, J. Cabrero, P. Lorite, D. C. Cabral-de Mello, and M. Bakkali. A step to the gigantic genome of the desert locust: chromosome sizes and repeated DNAs. *Chromosoma*, 124(2):263–275, June 2015.

[40] B. Canard and R. S. Sarfati. DNA polymerase fluorescent substrates with reversible 3 prime tags. *Gene*, 148(1):1–6, Oct. 1994.

[41] E. U. Canning. A New Eugregarine of Locusts, Gregarina garnhami n.sp., parasitic in Schistocerca gregaria Forsk. *The Journal of Protozoology*, 3(2):50–62, May 1956.

[42] A. J. Cease, S. Hao, L. Kang, J. J. Elser, and J. F. Harrison. Are color or high rearing density related to migratory polyphenism in the band-winged grasshopper, Oedaleus asiaticus? *Journal of Insect Physiology*, 56(8):926–936, Aug. 2010.

[43] M. Chaisson, P. Pevzner, and H. Tang. Fragment assembly with short reads. *Bioinformatics*, 20(13):2067–2074, Jan. 2004.

[44] K. H. Chamseddin, S. Q. Khan, M. L. Nguyen, M. Antosh, S. N. S. Morris, S. Kolli, N. Neretti, S. L. Helfand, and J. H. Bauer. takeout-dependent longevity is associated with altered Juvenile Hormone signaling. *Mechanisms of Ageing and Development*, 133(11-12):637–646, Nov. 2012.

[45] M.-P. Chapuis, C. S. Bazelet, L. Blondin, A. Foucart, R. Vitalis, and M. J. Samways. Subspecific taxonomy of the desert locust, *Schistocerca gregaria* (Orthoptera: Acrididae), based on molecular and morphological characters: Subspecific taxonomy of the desert locust. *Systematic Entomology*, 41(3):516–530, July 2016.

[46] J. Charles. The regulation of expression of insect cuticle protein genes. *Insect Biochemistry and Molecular Biology*, 40(3):205–213, Mar. 2010.

[47] R. A. Cheke. Theoretical rates of increase of gregarious and solitarious populations of the Desert Locust. *Oecologia*, 35(2):161–171, 1978.

[48] R. Chen, L. Hui, S. S. Cape, J. Wang, and L. Li. Comparative Neuropeptidomic Analysis of Food Intake via a Multi-faceted Mass Spectrometric Approach. *ACS chemical neuroscience*, 1(3):204–214, Mar. 2010.

[49] S. Chen, P. Yang, F. Jiang, Y. Wei, Z. Ma, and L. Kang. De Novo Analysis of Transcriptome Dynamics in the Migratory Locust during the Development of Phase Traits. *PLOS ONE*, 5(12):e15633, Dec. 2010.

[50] B. Chevreux, T. Pfisterer, B. Drescher, A. J. Driesel, W. E. G. Müller, T. Wetter, and S. Suhai. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, 14(6):1147–1159, June 2004.

[51] R. Chopra, G. Burow, A. Farmer, J. Mudge, C. E. Simpson, and M. D. Burow. Comparisons of De Novo Transcriptome Assemblers in Diploid and Polyploid Species Using Peanut ( Arachis spp.) RNA-Seq Data. *PLOS ONE*, 9(12):e115055, Dec. 2014.

[52] X. Chu, W. Lu, Y. Zhang, X. Guo, R. Sun, and B. Xu. Cloning, expression patterns, and preliminary characterization of AccCPR24, a novel RR-1 type cuticle protein gene from Apis cerana cerana. *Archives of Insect Biochemistry and Physiology*, 84(3):130–144, Nov. 2013.

[53] I. Claeys, B. Breugelmans, G. Simonet, S. Van Soest, F. Sas, A. De Loof, and J. Vanden Broeck. Neuroparsin transcripts as molecular markers in the process of desert locust (Schistocerca gregaria) phase transition. *Biochemical and biophysical research communications*, 341(2):599–606, Mar. 2006.

[54] E. Clynen, S. J. Husson, and L. Schoofs. Identification of new members of the (short) neuropeptide F family in locusts and Caenorhabditis elegans. *Annals of the New York Academy of Sciences*, 1163:60–74, Apr. 2009.

[55] E. Clynen and L. Schoofs. Peptidomic survey of the locust neuroendocrine system. *Insect biochemistry and molecular biology*, 39(8):491–507, Aug. 2009.

[56] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, Apr. 2010.

[57] M. Collett, E. Despland, S. J. Simpson, and D. C. Krakauer. Spatial scales of desert locust gregarization. *Proceedings of the National Academy of Sciences of the United States of America*, 95(22):13052–13055, Oct. 1998.

[58] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles. Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676, Sept. 2005.

[59] T. E. P. Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636–640, Oct. 2004.

[60] D. R. Cox. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242, 1958.

[61] K. Cressman. Monitoring Desert Locusts in the Middle East: An Overview. *Yale F&ES Bulletin*, 103:125–140, 1998.

[62] M. Croze, A. Wollstein, V. Božičević, D. Živković, W. Stephan, and S. Hutter. A genome-wide scan for genes under balancing selection in Drosophila melanogaster. *BMC Evolutionary Biology*, 17:15, 2017.

[63] D. A. Cullen, G. A. Sword, T. Dodgson, and S. J. Simpson. Behavioural phase change in the Australian plague locust, Chortoicetes terminifera, is triggered by tactile stimulation of the antennae. *Journal of Insect Physiology*, 56(8):937–942, Aug. 2010.

[64] D. a. Cullen, G. a. Sword, and S. J. Simpson. Optimizing multivariate behavioural syndrome models in locusts using automated video tracking. *Animal Behaviour*, 84(4):771–784, Oct. 2012.

[65] P. D. Danley, S. P. Mullen, F. Liu, V. Nene, J. Quackenbush, and K. L. Shaw. A cricket Gene Index: a genomic resource for studying neurobiology, speciation, and molecular evolution. *BMC Genomics*, 8(1):109, 2007.

[66] A. De Loof, I. Claeys, G. Simonet, P. Verleyen, T. Vandersmissen, F. Sas, and J. Huybrechts. Molecular markers of phase transition in locusts. *Insect Science*, 13(1):3–12, Feb. 2006.

[67] E. Despland, M. Collett, and S. J. Simpson. Small-scale processes in desert locust swarm formation: how vegetation patterns influence gregarization. *Oikos*, 88(3):652–662, Mar. 2000.

[68] E. Despland, J. Rosenberg, and S. J. Simpson. Landscape structure and locust swarming: a satellite's eye view. *Ecography*, 27(3):381–391, June 2004.

[69] E. Despland and S. J. Simpson. The role of food distribution and nutritional quality in behavioural phase change in the desert locust. *Animal Behaviour*, 59:643–652, 2000.

[70] E. Despland and S. J. Simpson. Small-scale vegetation patterns in the parental environment influence the phase state of hatchlings of the desert locust. *Physiological Entomology*, 25(1):74–81, Mar. 2000.

[71] E. Despland and S. J. Simpson. Food choices of solitarious and gregarious locusts reflect cryptic and aposematic antipredator strategies. *Animal Behaviour*, 69(2):471–479, Feb. 2005.

[72] E. Despland and S. J. Simpson. Surviving the change to warning colouration: density-dependent polyphenism suggests a route for the evolution of aposematism. *Chemoecology*, 15(2):69–75, June 2005.

[73] S. Dillen, R. Verdonck, S. Zels, P. Van Wielendaele, and J. Vanden Broeck. Identification of the short neuropeptide F precursor in the desert locust: evidence for an inhibitory role of sNPF in the control of feeding. *Peptides*, 53:134–139, Mar. 2014.

[74] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, Jan. 2013.

[75] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, Oct. 1998.

[76] S. R. Eddy. A new generation of homology search tools based on probabilistic inference. In *Proceedings of the 20th international conference*, pages 205–211, Pacifico Yokohama, Japan, Oct. 2009. Imperial college press.

[77] P. E. Ellis. Marching in Locust Hoppers of the Solitary Phase. *Nature*, 166(4212):151–151, July 1950.

[78] P. E. Ellis. *Marching Behaviour of Hoppers of African Migratory Locust in the Laboratory*. Centre for Overseas Pest Research, Dec. 1951.

[79] P. E. Ellis. The Gregarious Behaviour of Marching Locusta Migratoria Migratorioides (R. & F.) Hoppers. *Journal of Experimental Biology*, 30(2):214–234, June 1953.

[80] P. E. Ellis. Social Aggregation and Gregarious Behaviour in Hoppers of Locusta Migratoria Migratorioides (R. & F.) 1). *Behaviour*, 5(1):225–259, Jan. 1953.

[81] P. E. Ellis. Differences in Social Aggregation in Two Species of Locust. *Nature*, 178(4540):1007–1007, Nov. 1956.

[82] P. E. Ellis. Learning and social aggregation in locust hoppers. *Animal Behaviour*, VII(1-2):91–107, 1959.

[83] P. E. Ellis. Some factors influencing phase characters in the nymphs of the locust,Locusta migratoria migratorioides (R. and F.). *Insectes Sociaux*, 6(1):21–39, Mar. 1959.

[84] P. E. Ellis. Changes in the social aggregation of locust hoppers with changes in rearing conditions. *Animal Behaviour*, 11(1):152–160, Jan. 1963.

[85] P. E. Ellis and A. Pearce. Innate and learned behaviour patterns that lead to group formation in locust hoppers. *Animal Behaviour*, 10(3–4):305–318, July 1962.

[86] U. R. Ernst, M. B. Van Hiel, G. Depuydt, B. Boerjan, A. De Loof, and L. Schoofs. Epigenetics and locust life phase transitions. *The Journal of experimental biology*, 218(Pt 1):88–99, 2015.

[87] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8(3):186–194, Mar. 1998.

[88] B. Ewing, L. Hillier, M. C. Wendl, and P. Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, 8(3):175–185, Mar. 1998.

[89] C. Falckenhayn, B. Boerjan, G. Raddatz, M. Frohme, L. Schoofs, and F. Lyko. Characterization of genome methylation patterns in the desert locust Schistocerca gregaria. *Journal of Experimental Biology*, 216(8):1423–1429, Apr. 2013.

[90] J. C. Faure. *The Phases of Locusts in South Africa*. Imperial institute of entomology, 1932.

[91] K. Fawagreh, M. M. Gaber, and E. Elyan. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering*, 2(1):602–609, Dec. 2014.

[92] A. A. Fedotova and A. V. Kouprianov. Archival research reveals the true date of birth of the father of locust phase theory, Sir Boris Uvarov, FRS. *EUROASIAN ENTOMOLOGICAL JOURNAL*,, 15(4):321–327, 2016.

[93] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta. Pfam: the protein families database. *Nucleic Acids Research*, 42(D1):D222–D230, Jan. 2014.

[94] R. D. Finn, J. Clements, and S. R. Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(Web Server issue):W29–W37, July 2011.

[95] N. A. Fonseca, J. Rung, A. Brazma, and J. C. Marioni. Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–3177, Dec. 2012.

[96] H. Fricke and S. Fricke. Monogamy and sex change by aggressive dominance in coral reef fish. *Nature*, 266(5605):830–832, Apr. 1977.

[97] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, Dec. 2012.

[98] G. Fusco and A. Minelli. Phenotypic plasticity in development and evolution: facts and concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1540):547–556, Feb. 2010.

[99] N. Geva, M. Guershon, M. Orlova, and a. Ayali. Memoirs of a locust: density-dependent behavioral change as a model for learning and memory. *Neurobiology of learning and memory*, 93(2):175–82, Feb. 2010.

[100] S. Gillett. Airborne Factor affecting the Grouping Behaviour of Locusts. *Nature*, 218(5143):782–783, May 1968.

[101] S. D. Gillett. Social aggregation of adult Schistocerca gregaria and Locusta migratoria migratorioides in relation to the final moult and ageing. *Animal Behaviour*, 20(3):526–533, Aug. 1972.

[102] S. D. Gillett. Social determinants of aggregation behaviour in adults of the desert locust. *Animal Behaviour*, 21(3):599–606, Aug. 1973.

[103] S. D. Gillett. Changes in the social behaviour of the desert locust, Schistocerca gregaria, in response to the gregarizing pheromone. *Animal Behaviour*, 23:494–503, Aug. 1975.

[104] S. D. Gillett. Solitarization in the desert locust, Schistocerca gregaria (Forskål) (Orthoptera: Acrididae). *Bulletin of Entomological Research*, 78(04):623, Dec. 1988.

[105] S. Gotham and H. Song. Non-swarming grasshoppers exhibit density-dependent phenotypic plasticity reminiscent of swarming locusts. *Journal of Insect Physiology*, 59(11):1151–1159, Nov. 2013.

[106] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, May 2011.

[107] L. J. Gray, G. A. Sword, M. L. Anstey, F. J. Clissold, and S. J. Simpson. Behavioural phase polyphenism in the Australian plague locust (Chortoicetes terminifera). *Biology Letters*, 5(3):306–309, June 2009.

[108] R. Guigó, S. Knudsen, N. Drake, and T. Smith. Prediction of gene structure. *Journal of Molecular Biology*, 226(1):141–157, July 1992.

[109] W. Guo, X. Wang, Z. Ma, L. Xue, J. Han, D. Yu, and L. Kang. CSP and Takeout Genes Modulate the Switch between Attraction and Repulsion during Behavioral Phase Change in the Migratory Locust. *PLoS Genetics*, 7(2), Feb. 2011.

[110] W. Guo, X. Wang, D. Zhao, P. Yang, and L. Kang. Molecular cloning and temporal–spatial expression of I element in gregarious and solitary locusts. *Journal of Insect Physiology*, 56(8):943–948, Aug. 2010.

[111] M. Gáliková and T. Flatt. Dietary restriction and other lifespan extending pathways converge at the activation of the downstream effector takeout. *Aging*, 2(7):387–389, July 2010.

[112] B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. LeDuc, N. Friedman, and A. Regev. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature protocols*, 8(8), Aug. 2013.

[113] T. Hagai, M. Cohen, and G. Bloch. Genes encoding putative Takeout/juvenile hormone binding proteins in the honeybee (Apis mellifera) and modulation by age and juvenile hormone of the takeout-like gene GB19811. *Insect Biochemistry and Molecular Biology*, 37(7):689–701, July 2007.

[114] M. J. Hansen, J. Buhl, S. Bazazi, S. J. Simpson, and G. A. Sword. Cannibalism in the lifeboat — collective movement in Australian plague locusts. *Behavioral Ecology and Sociobiology*, 65(9):1715–1720, Sept. 2011.

[115] T. J. Hardcastle and K. A. Kelly. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1):422, 2010.

[116] D. G. C. Harper. Competitive foraging in mallards: 'Ideal free' ducks. *Animal Behaviour*, 30(2):575–584, 1982.

[117] A. W. Harvey. Schistocerca piceifrons (Walker) (Orthoptera: Acrididae), the swarming locust of tropical America: a review. *Bulletin of Entomological Research*, 73(02):171, June 1983.

[118] Y. Heifetz, S. Applebaum, and G. Popov. Phase characteristics of the Israeli population of the migratory locust, Locusta migratoria (L.)(Orthoptera: Acrididae). *Journal of Orthoptera Research*, 2(2):15–20, 1994.

[119] Y. Heifetz, H. Voet, and S. W. Applebaum. Factors affecting behavioural phase transition in the desert locust , Schistocerca gregaria (Forskal) (Orthoptera: Acrididae). *Journal of Chemical Ecology*, 22(9):1717–1734, 1996.

[120] I. Hen, A. Sakov, N. Kafkafi, I. Golani, and Y. Benjamini. The dynamics of spatial behavior: how can robust smoothing techniques help? *Journal of Neuroscience Methods*, 133(1-2):161–172, Feb. 2004.

[121] D. W. Hosmer, T. Hosmer, S. Le Cessie, and S. Lemeshow. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16(9):965–980, May 1997.

[122] B. Hoste, L. Luyten, I. Claeys, E. Clynen, M. M. Rahman, A. Loof, and M. Breuer. An improved breeding method for solitarious locusts. *Entomologia Experimentalis et Applicata*, 104(2-3):281–288, Sept. 2002.

[123] B. Hoste, S. J. Simpson, S. Tanaka, D. H. Zhu, a. De Loof, and M. Breuer. Effects of [His(7)]-corazonin on the phase state of isolated-reared (solitarious) desert locusts, Schistocerca gregaria. *Journal of insect physiology*, 48(10):981–990, Oct. 2002.

[124] H. Hotelling. Analysis of a complex of statistical variables into principal components. *ournal of Educational Psychology*, 24:417–441, 498–520., 1933.

[125] L. Hou, F. Jiang, P. Yang, X. Wang, and L. Kang. Molecular characterization and expression profiles of neuropeptide precursors in the migratory locust. *Insect Biochemistry and Molecular Biology*, 63:63–71, Aug. 2015.

[126] S. Imbeaud. Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces. *Nucleic Acids Research*, 33(6):e56–e56, Mar. 2005.

[127] M. Islam, P. Roessingh, S. J. Simpson, and A. R. McCaffery. Parental effects on the behaviour and colouration of nymphs of the desert locust Schistocerca gregaria. *Journal of Insect Physiology*, 40(2):173–181, Feb. 1994.

[128] M. S. Islam, P. Roessingh, S. J. Simpson, and A. R. Mccaffery. Effects of Population Density Experienced by Parents during Mating and Oviposition on the Phase of Hatchling Desert Locusts, Schistocerca gregaria. *Proceedings of the Royal Society B: Biological Sciences*, 257(1348):93–98, July 1994.

[129] K. J. Janssen, Y. Vergouwe, A. R. T. Donders, F. E. Harrell, Q. Chen, D. E. Grobbee, and K. G. Moons. Dealing with Missing Predictor Values When Applying Clinical Prediction Models. *Clinical Chemistry*, 55(5):994–1001, May 2009.

[130] T. Janssen, I. Claeys, G. Simonet, A. De Loof, J. Girardie, and J. V. Broeck. cDNA cloning and transcript distribution of two different neuroparsin precursors in the desert locust, Schistocerca gregaria: Desert locust neuroparsin transcripts. *Insect Molecular Biology*, 10(2):183–189, Apr. 2001.

[131] F. Jiang, M. Yang, W. Guo, X. Wang, and L. Kang. Large-Scale Transcriptome Analysis of Retroelements in the Migratory Locust, Locusta migratoria. *PLoS ONE*, 7(7):e40532, July 2012.

[132] Y. Jin, B. Cong, L. Wang, Y. Gao, H. Zhang, H. Dong, and Z. Lin. Differential Gene Expression Analysis of the Epacromius coerulipes (Orthoptera: Acrididae) Transcriptome. *Journal of Insect Science*, 16(1):42, Jan. 2016.

[133] I. Joliffe and B. Morgan. Principal component analysis and exploratory factor analysis. *Statistical Methods in Medical Research*, 1(1):69–95, Mar. 1992.

[134] M. Kabra, A. A. Robie, M. Rivera-Alba, S. Branson, and K. Branson. JAABA: interactive machine learning for automatic annotation of animal behavior. *Nature Methods*, 10(1):64–67, Dec. 2012.

[135] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(Database issue):D109–114, Jan. 2012.

[136] L. Kang, X. Chen, Y. Zhou, B. Liu, W. Zheng, R. Li, J. Wang, and J. Yu. The analysis of large-scale gene expression correlated to the phase changes of the migratory locust. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51):17611–17615, Dec. 2004.

[137] M. M. Kasumovic, Z. Chen, and M. R. Wilkins. Australian black field crickets show changes in neural gene expression associated with socially-induced morphological, life-history, and behavioral plasticity. *BMC Genomics*, 17(1):827, Oct. 2016.

[138] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, July 2002.

[139] S. A. Kelly, T. M. Panhuis, and A. M. Stoehr. Phenotypic Plasticity: Molecular Mechanisms and Adaptive Significance. In R. Terjung, editor, *Comprehensive Physiology*. John Wiley & Sons, Inc., Hoboken, NJ, USA, Apr. 2012. DOI: 10.1002/cphy.c110008.

[140] J. S. Kennedy. The behaviour of the desert locust (Schistocerca gregaria (Forsk.)) in an outbreak centre. *Transactions of the Royal Entomological Society of London*, 89(10):385–542, 1939.

[141] K. Kopec, V. Alva, and A. Lupas. Bioinformatics of the TULIP domain superfamily. *Biochemical Society Transactions*, 39(4):1033–1038, Aug. 2011.

[142] P. G. Kraft, C. E. Franklin, and M. W. Blows. Predator-induced phenotypic plasticity in tadpoles: extension or innovation? *Journal of Evolutionary Biology*, 19(2):450–458, Mar. 2006.

[143] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, 305(3):567–580, Jan. 2001.

[144] J. Kruppa, Y. Liu, G. Biau, M. Kohler, I. R. König, J. D. Malley, and A. Ziegler. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory: Probability estimation with machine learning methods. *Biometrical Journal*, 56(4):534–563, July 2014.

[145] J. Kruppa, Y. Liu, H.-C. Diener, T. Holste, C. Weimar, I. R. König, and A. Ziegler. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications: Probability estimation with machine learning methods: Applications. *Biometrical Journal*, 56(4):564–583, July 2014.

[146] E. M. LaFiandra and K. J. Babbitt. Predator induced phenotypic plasticity in the pinewoods tree frog, Hyla femoralis : necessary cues and the cost of development. *Oecologia*, 138(3):350–359, Feb. 2004.

[147] K. Lagesen, P. Hallin, E. A. Rødland, H.-H. Stærfeldt, T. Rognes, and D. W. Ussery. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35(9):3100–3108, May 2007.

[148] B. Langmead. Aligning Short Sequencing Reads with Bowtie. In A. D. Baxevanis, D. B. Davison, R. D. Page, G. A. Petsko, L. D. Stein, and G. D. Stormo, editors, *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., Hoboken, NJ, USA, Dec. 2010. DOI: 10.1002/0471250953.bi1107s32.

[149] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, Mar. 2012.

[150] C. Lenaerts, D. Cools, R. Verdonck, L. Verbakel, J. Vanden Broeck, and E. Marchal. The ecdysis triggering hormone system is essential for successful moulting of a major hemimetabolous pest insect, Schistocerca gregaria. *Scientific Reports*, 7:46502, Apr. 2017.

[151] N. Leng, J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. G. Smits, J. D. Haag, M. N. Gould, R. M. Stewart, and C. Kendziorski. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 29(8):1035–1043, Apr. 2013.

[152] R. Leo Lester, C. Grach, M. Paul Pener, and S. J. Simpson. Stimuli inducing gregarious colouration and behaviour in nymphs of Schistocerca gregaria. *Journal of insect physiology*, 51(7):737–47, July 2005.

[153] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

[154] H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, Mar. 2010.

[155] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug. 2009.

[156] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, July 2006.

[157] Z. Li, Y. Chen, D. Mu, J. Yuan, Y. Shi, H. Zhang, J. Gan, N. Li, X. Hu, B. Liu, B. Yang, and W. Fan. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Briefings in Functional Genomics*, 11(1):25–37, Jan. 2012.

[158] H. Li-Byarlay, Y. Li, H. Stroud, S. Feng, T. C. Newman, M. Kaneda, K. K. Hou, K. C. Worley, C. G. Elsik, S. a. Wickline, S. E. Jacobsen, J. Ma, and G. E. Robinson. RNA interference knockdown of DNA methyl-transferase 3 affects gene alternative splicing in the honey bee. *Proceedings of the National Academy of Sciences of the United States of America*, 110(31):12750–5, July 2013.

[159] A. Liaw and M. Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002.

[160] T. Lochmatter, P. Roduit, C. Cianci, N. Correll, J. Jacot, and A. Martinoli. SwisTrack - a flexible open source tracking software for multi-agent systems. pages 4004–4010. IEEE, Sept. 2008.

[161] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), Dec. 2014.

[162] N. Lovejoy, S. Mullen, G. Sword, R. Chapman, and R. Harrison. Ancient trans-Atlantic flight explains locust biogeography: molecular phylogenetics of Schistocerca. *Proceedings of the Royal Society B: Biological Sciences*, 273(1588):767–774, Apr. 2006.

[163] J. Lu, J. K. Tomfohr, and T. B. Kepler. Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, 6:165, 2005.

[164] J. Lüning. Phenotypic plasticity of Daphnia pulex in the presence of invertebrate predators: morphological and life history responses. *Oecologia*, 92(3):383–390, Dec. 1992.

[165] Z. Ma, W. Guo, X. Guo, X. Wang, and L. Kang. Modulation of behavioral phase changes of the migratory locust by the catecholamine metabolic pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 108(10):3882–7, Mar. 2011.

[166] Z. Ma, J. Yu, and L. Kang. LocustDB: a relational database for the transcriptome and biology of the migratory locust (Locusta migratoria). *BMC genomics*, 7:11, Jan. 2006.

[167] K. Maeno, S. Tanaka, and K.-I. Harano. Tactile stimuli perceived by the antennae cause the isolated females to produce gregarious offspring in the desert locust, Schistocerca gregaria. *Journal of insect physiology*, 57(1):74–82, Jan. 2011.

[168] K. O. Maeno, C. Piou, M. A. Ould Babah, and S. Nakamura. Eggs and hatchlings variations in desert locusts: phase related characteristics and starvation tolerance. *Frontiers in Physiology*, 4, 2013.

[169] T. Magoc and S. L. Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963, Nov. 2011.

[170] J. Magor, M. Lecoq, and D. Hunter. Preventive control and Desert Locust plagues. *Crop Protection*, 27(12):1527–1533, Dec. 2008.

[171] J. D. Malley, J. Kruppa, A. Dasgupta, K. G. Malley, and A. Ziegler. Probability Machines: Consistent Probability Estimation Using Nonparametric Learning Machines. *Methods of Information in Medicine*, 51(1):74–81, Sept. 2011.

[172] E. B. Mallon, H. E. Amarasinghe, and S. R. Ott. Acute and chronic gregarisation are associated with distinct DNA methylation fingerprints in desert locusts. *Scientific Reports*, 6:35608, Oct. 2016.

[173] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro,

X. V. Gomes, B. C. Goodwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. *Nature*, 437(7057):376–380, Sept. 2005.

[174] C. Mathé, M.-F. Sagot, T. Schiex, and P. Rouzé. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research*, 30(19):4103–4117, Jan. 2002.

[175] E. Mayr. *Animal Species and Evolution*. Jan. 1963.

[176] E. Maza, P. Frasse, P. Senin, M. Bouzayen, and M. Zouine. Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: A matter of relative size of studied transcriptomes. *Communicative & Integrative Biology*, 6(6), Nov. 2013.

[177] A. R. McCaffery, S. J. Simpson, M. S. Islam, and P. Roessingh. A gregarizing factor present in the egg pod foam of the desert locust Schistocerca gregaria. *Journal of Experimental Biology*, 201(3):347–363, Feb. 1998.

[178] M. S. McNeill, K. M. Kapheim, A. Brockmann, T. A. W. McGill, and G. E. Robinson. Brain regions and molecular pathways responding to food reward type and value in honey bees. *Genes, Brain and Behavior*, 15(3):305–317, Mar. 2016.

[179] A. E. Minoche, J. C. Dohm, and H. Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, 12:R112, 2011.

[180] A. Muñoz-Mérida, E. Viguera, M. G. Claros, O. Trelles, and A. J. Pérez-Pulido. Sma3s: A Three-Step Modular Annotator for Large Sequence Datasets. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, 21(4):341–353, Aug. 2014.

[181] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. J. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H.-H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter. A Whole-Genome Assembly of Drosophila. *Science*, 287(5461):2196–2204, Mar. 2000.

[182] K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. C. Linak, A. Hirai, H. Takahashi, M. Altaf-Ul-Amin, N. Ogasawara, and S. Kanaya. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, 39(13):e90–e90, July 2011.

[183] K. Nakasugi, R. Crowhurst, J. Bally, and P. Waterhouse. Combining Transcriptome Assemblies from Multiple De Novo Assemblers in the Allo-Tetraploid Plant Nicotiana benthamiana. *PLoS ONE*, 9(3), Mar. 2014.

[184] A. Nanjappa, L. Cheng, W. Gao, C. Xu, A. Claridge-Chang, and Z. Bichler. Mouse Pose Estimation From Depth Images. *arXiv:1511.07611 [cs]*, Nov. 2015. arXiv: 1511.07611.

[185] D. R. Nässel, M. G. Persson, and J. E. Muren. Baratin, a nonamidated neurostimulating neuropeptide, isolated from cockroach brain: distribution and actions in the cockroach and locust nervous systems. *The Journal of Comparative Neurology*, 422(2):267–286, June 2000.

[186] S. R. Ott and S. M. Rogers. Gregarious desert locusts have substantially larger brains with altered proportions compared with the solitarious phase. *Proceedings. Biological sciences / The Royal Society*, 277(1697):3087–3096, Oct. 2010.

[187] S. R. Ott, H. Verlinden, S. M. Rogers, C. H. Brighton, P. S. Quah, R. K. Vleugels, R. Verdonck, and J. Vanden Broeck. Critical role for protein kinase A in the acquisition of gregarious behavior in the desert locust. *Proceedings of the National Academy of Sciences of the United States of America*, 109(7):E381–7, Feb. 2012.

[188] F. S.-M. Pais, P. d. Ruy, G. Oliveira, and R. Coimbra. Assessing the efficiency of multiple sequence alignment programs. *Algorithms for Molecular Biology*, 9(1):4, 2014.

[189] D. J. Parker, C. B. Cunningham, C. A. Walling, C. E. Stamper, M. L. Head, E. M. Roy-Zokan, E. C. McKinney, M. G. Ritchie, and A. J. Moore. Transcriptomes of parents identify parenting strategies and sexual conflict in a subsocial beetle. *Nature Communications*, 6:8449, Sept. 2015.

[190] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. Technical Report biorxiv;021592v3, June 2015.

[191] R. Patro, S. M. Mount, and C. Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, 32(5):462–464, Apr. 2014.

[192] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical magazine*, 2(11):559–572., 1901.

[193] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8):2444–2448, Apr. 1988.

[194] M. P. Pener. Locust Phase Polymorphism and its Endocrine Relations. In P. D. Evans, editor, *Advances in Insect Physiology*, volume 23, pages 1–79. Academic Press, 1991. DOI: 10.1016/S0065-2806(08)60091-0.

[195] M. P. Pener and S. J. Simpson. *Locust Phase Polyphenism: An update*, volume 36. Advances in in insect physiology, vol 36, 2009.

[196] T. N. Petersen, S. Brunak, G. von Heijne, and H. Nielsen. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, 8(10):785–786, Sept. 2011.

[197] P. A. Pevzner, H. Tang, and M. S. Waterman. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753, Aug. 2001.

[198] S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T. Rattei, I. Letunic, T. Doerks, L. J. Jensen, C. von Mering, and P. Bork. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Research*, 40(Database issue):D284–D289, Jan. 2012.

[199] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn. The Pfam protein families database. *Nucleic Acids Research*, 40(Database issue):D290–301, Jan. 2012.

[200] Z. Qiu, F. Liu, H. Lu, and Y. Huang. Characterization and analysis of a de novo transcriptome from the pygmy grasshopper Tetrix japonica. *Molecular Ecology Resources*, July 2016.

[201] Z. Qiu, F. Liu, H. Lu, H. Yuan, Q. Zhang, and Y. Huang. De Novo Assembly and Characterization of the Transcriptome of Grasshopper Shirakiacris shirakii. *International Journal of Molecular Sciences*, 17(8):1110, July 2016.

[202] A. R. Quinlan. BEDTools: The Swiss-Army Tool for Genome Feature Analysis: BEDTools: the Swiss-Army Tool for Genome Feature Analysis. In A. Bateman, W. R. Pearson, L. D. Stein, G. D. Stormo, and J. R. Yates, editors, *Current Protocols in Bioinformatics*, pages 11.12.1–11.12.34. John Wiley & Sons, Inc., Hoboken, NJ, USA, Sept. 2014. DOI: 10.1002/0471250953.bi1112s47.

[203] M. M. Rahman, L. V. Bosch, G. Baggerman, E. Clynen, K. Hens, B. Hoste, K. Meylaers, T. Vercammen, L. Schoofs, a. De Loof, and M. Breuer. Search for peptidic molecular markers in hemolymph of crowd-(gregarious) and isolated-reared (solitary) desert locusts, Schistocerca gregaria. *Peptides*, 23(11):1907–14, Nov. 2002.

[204] P. Ramdya, J. Schneider, and J. D. Levine. The neurogenetics of group behavior in *Drosophila melanogaster*. *The Journal of Experimental Biology*, 220(1):35–41, Jan. 2017.

[205] J. E. Rebers and J. H. Willis. A conserved domain in arthropod cuticular proteins binds chitin1. *Insect Biochemistry and Molecular Biology*, 31(11):1083–1093, Oct. 2001.

[206] W. Revelle. psych: Procedures for psychological, Psychometric, and Personality Research, 2017.

[207] A. M. Reynolds, G. a. Sword, S. J. Simpson, and D. R. Reynolds. Predator percolation, insect outbreaks, and phase polyphenism. *Current biology : CB*, 19(1):20–4, Jan. 2009.

[208] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43(7), 2015.

[209] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, Jan. 2010.

[210] M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010.

[211] M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, Nov. 2007.

[212] R. Roer, S. Abehsera, and A. Sagi. Exoskeletons across the Pancrustacea: Comparative Morphology, Physiology, Biochemistry and Genetics. *Integrative and Comparative Biology*, 55(5):771–791, Nov. 2015.

[213] P. Roessingh, A. Bouaïchi, and S. J. Simpson. Effects of sensory stimuli on the behavioural phase state of the desert locust, Schistocerca gregaria. *Journal of Insect Physiology*, 44(10):883–893, Oct. 1998.

[214] P. Roessingh and S. J. Simpson. The time-course of behavioural phase change in nymphs of the desert locust, Schistocerca gregaria. *Physiological Entomology*, 19(3):191–197, Sept. 1994.

[215] P. Roessingh, S. J. Simpson, and S. James. Analysis of Phase-Related Changes in Behaviour of Desert Locust Nymphs. *Proceedings of the Royal Society B: Biological Sciences*, 252(1333):43–49, Apr. 1993.

[216] J. Roffey and G. Popov. Environmental and Behavioural Processes in a Desert Locust Outbreak. *Nature*, 219(5153):446–450, Aug. 1968.

[217] S. M. Rogers, D. A. Cullen, M. L. Anstey, M. Burrows, E. Despland, T. Dodgson, T. Matheson, S. R. Ott, K. Stettin, G. A. Sword, and S. J. Simpson. Rapid behavioural gregarization in the desert locust , Schistocerca gregaria entails synchronous changes in both activity and attraction to conspecifics. *Journal of Insect Physiology*, 65:9–26, 2014.

[218] S. M. Rogers, T. Matheson, E. Despland, T. Dodgson, M. Burrows, and S. J. Simpson. Mechanosensory-induced behavioural gregarization in the desert locust Schistocerca gregaria. *The Journal of experimental biology*, 206(Pt 22):3991–4002, 2003.

[219] S. M. Rogers, T. Matheson, K. Sasaki, K. Kendrick, S. J. Simpson, and M. Burrows. Substantial changes in central nervous system neurotransmitters and neuromodulators accompany phase change in the locust. *The Journal of experimental biology*, 207(Pt 20):3603–3617, 2004.

[220] S. M. Rogers and S. R. Ott. Differential activation of serotonergic neurons during short- and long-term gregarization of desert locusts. *Proceedings of the Royal Society B: Biological Sciences*, 282(1800):20142062–20142062, Dec. 2014.

[221] S. M. Rogers, J. Riley, C. Brighton, G. P. Sutton, D. A. Cullen, and M. Burrows. Increased muscular volume and cuticular specialisations enhance jump velocity in solitarious compared with gregarious desert locusts, Schistocerca gregaria. *Journal of Experimental Biology*, 219(5):635–648, Mar. 2016.

[222] M. Ronaghi. A Sequencing Method Based on Real-Time Pyrophosphate. *Science*, 281(5375):363–365, July 1998.

[223] S. Sarfare, S. T. Ahmad, M. V. Joyce, B. Boggess, and J. E. O'Tousa. The Drosophila ninaG Oxidoreductase Acts in Visual Pigment Chromophore Production. *Journal of Biological Chemistry*, 280(12):11895–11901, Mar. 2005.

[224] L. Sarov-Blat, W. V. So, L. Liu, and M. Rosbash. The Drosophila takeout gene is a novel molecular link between circadian rhythms and feeding behavior. *Cell*, 101(6):647–656, June 2000.

[225] C. M. Schalk. Predator-induced phenotypic plasticity in an arid-adapted tropical tadpole: Tropical Tadpole Plasticity. *Austral Ecology*, 41(4):409–416, June 2016.

[226] M. Schirmer, U. Z. Ijaz, R. D'Amore, N. Hall, W. T. Sloan, and C. Quince. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, 43(6):e37–e37, Mar. 2015.

[227] A. Schroeder, O. Mueller, S. Stocker, R. Salowsky, M. Leiber, M. Gassmann, S. Lightfoot, W. Menzel, M. Granzow, and T. Ragg. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*, 7:3, 2006.

[228] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)*, 28(8):1086–92, Apr. 2012.

[229] M. A. Schwinghammer, X. Zhou, S. Kambhampati, G. W. Bennett, and M. E. Scharf. A novel gene from the takeout family involved in termite trail-following behavior. *Gene*, 474(1-2):12–21, Mar. 2011.

[230] C. L. Schürmann and J. A. R. A. M. van Hooff. Reproductive strategies of the orang-utan: New data and a reconsideration of existing sociosexual models. *International Journal of Primatology*, 7(3):265–287, June 1986.

[231] M. Sen Sarma, S. L. Rodriguez-Zas, F. Hong, S. Zhong, and G. E. Robinson. Transcriptomic Profiling of Central Nervous System Regions in Three Species of Honey Bee during Dance Communication Behavior. *PLoS ONE*, 4(7):e6408, July 2009.

[232] F. Seyednasrollah, A. Laiho, and L. L. Elo. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, 16(1):59–70, Jan. 2015.

[233] S. J. Simpson, E. Despland, B. F. Hägele, and T. Dodgson. Gregarious behavior in desert locusts is evoked by touching their back legs. *Proceedings of the National Academy of Sciences of the United States of America*, 98(7):3895–7, Mar. 2001.

[234] S. J. Simpson, A. R. McCAFFERY, and B. F. HäGele. A behavioural analysis of phase change in the desert locust. *Biological Reviews*, 74(4):461–480, 1999.

[235] P. Simões, J. Niven, and S. Ott. Phenotypic Transformation Affects Associative Learning in the Desert Locust. *Current Biology*, 23(23):2407–2412, Dec. 2013.

[236] A. Smit, R. Hubley, and P. Green. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>.

[237] H. Song. Phylogenetic perspectives on the evolution of locust phase polyphenism. *Journal of Orthoptera Research*, 14(2):235–245, 2005.

[238] H. Song. Density-Dependent Phase Polyphenism in Nonmodel Locusts: A Minireview. *Psyche*, 2011:1–16, 2011.

[239] H. Song, M. J. Moulton, K. D. Hiatt, and M. F. Whiting. Uncovering historical signature of mitochondrial DNA hidden in the nuclear genome: the biogeography of *Schistocerca* revisited. *Cladistics*, 29(6):643–662, Dec. 2013.

[240] H. Song and J. W. Wenzel. Phylogeny of bird-grasshopper subfamily Cyrtacanthacridinae (Orthoptera: Acrididae) and the evolution of locust phase polyphenism. *Cladistics*, 24(4):515–542, Aug. 2008.

[241] D. J. Stekel, Y. Git, and F. Falciani. The Comparison of Gene Expression from Multiple cDNA Libraries. *Genome Research*, 10(12):2055–2061, Jan. 2000.

[242] K. Stettin and S. Ott. Phase change in the desert locust: A novel computerised behavioural assay to study the effects of anaesthesia on gregarisation. *Frontiers in Behavioral Neuroscience*, 6, 2012.

[243] Stettin, Katja. *Mechanisms driving the acquisition and maintenance of gregarious behaviour in the desert locust (Schistocerca gregaria).* PhD thesis, University of Cambridge, Cambridge, 2014.

[244] H. Stibor and J. Lüning. Predator-Induced Phenotypic Variation in the Pattern of Growth and Reproduction in Daphnia hyalina (Crustacea: Cladocera). *Functional Ecology*, 8(1):97–101, 1994.

[245] G. D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics (Oxford, England)*, 16(1):16–23, Jan. 2000.

[246] R. Sugahara, S. Saeki, A. Jouraku, T. Shiotsuki, and S. Tanaka. Knockdown of the corazonin gene reveals its critical role in the control of gregarious characteristics in the desert locust. *Journal of Insect Physiology*, 79:80–87, 2015.

[247] D. Sumpter, J. Buhl, D. Biro, and I. Couzin. Information transfer in moving animal groups. *Theory in Biosciences*, 127(2):177–186, May 2008.

[248] G. A. Sword. Density-dependent warning coloration. *Nature*, 397(6716):217–217, Jan. 1999.

[249] G. A. Sword. Tasty on the outside, but toxic in the middle: grasshopper regurgitation and host plant-mediated toxicity to a vertebrate predator. *Oecologia*, 128(3):416–421, Aug. 2001.

[250] G. A. Sword. A role for phenotypic plasticity in the evolution of aposematism. *Proceedings of the Royal Society B: Biological Sciences*, 269(1501):1639–1644, Aug. 2002.

[251] G. a. Sword. To be or not to be a locust? A comparative analysis of behavioral phase change in nymphs of Schistocerca americana and S. gregaria. *Journal of Insect Physiology*, 49(7):709–717, July 2003.

[252] G. a. Sword, M. Lecoq, and S. J. Simpson. Phase polyphenism and preventative locust management. *Journal of insect physiology*, 56(8):949–57, Aug. 2010.

[253] G. A. Sword, S. J. Simpson, O. T. M. El Hadi, and H. Wilps. Density-dependent aposematism in the desert locust. *Proceedings of the Royal Society B: Biological Sciences*, 267(1438):63–68, Jan. 2000.

[254] P. Symmons and K. Cressman. *Desert Locust Guidelines*. Food and Agriculture Organization of the united Nations (FAO), 2001.

[255] S. Tanaka. Corazonin and locust phase polyphenism. *Appl. Entomol.Zool.*, 41(2):179–193, 2006.

[256] S. Tanaka, K. Maeno, S. Ould Mohamed, S. Ould Ely, and M. A. Babah Ebbe. Upsurges of desert locust populations in Mauritania: Body coloration, behavior and morphological characteristics. *Applied Entomology and Zoology*, 45(4):641–652, 2010.

[257] S. Tanaka and Y. Nishide. Behavioral phase shift in nymphs of the desert locust, Schistocerca gregaria: Special attention to attraction/avoidance behaviors and the role of serotonin. *Journal of Insect Physiology*, 59(1):101–112, Jan. 2013.

[258] S. Tanaka, D. H. Zhu, B. Hoste, and M. Breuer. The dark-color inducing neuropeptide, [His(7)]-corazonin, causes a shift in morphometic characteristics towards the gregarious phase in isolated-reared (solitarious) Locusta migratoria. *Journal of insect physiology*, 48(11):1065–1074, Nov. 2002.

[259] J. D. Thompson, B. Linard, O. Lecompte, and O. Poch. A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. *PLOS ONE*, 6(3):e18093, Mar. 2011.

[260] N. Tinbergen. On aims and methods of Ethology. *Zeitschrift für Tierpsychologie*, 20(4):410–433, 1963.

[261] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, 31(1):46–53, Dec. 2012.

[262] S. S. Utami. Male bimaturism and reproductive success in Sumatran orang-utans. *Behavioral Ecology*, 13(5):643–652, Sept. 2002.

[263] B. Uvarov. *Locusts and grasshoppers. A handbook for their study and control,*. Imperial Bureau of Entomology, London, 1928.

[264] B. P. Uvarov. A revision of the genus Locusta, L. (= Pachytylus, Fieb.), with a new theory as to the periodicity and migrations of locusts. *Bulletin of Entomological Research*, 12:135–163, 1921.

[265] B. P. Uvarov. *Grasshoppers and Locusts: A handbook of General Acridology.*, volume 1. Cambridge University Press for the Anti-Locust Research Centre, 1966.

[266] B. P. Uvarov. *Grasshoppers and Locusts: A handbook of General Acridology.*, volume 2. Cambridge University Press for the Anti-Locust Research Centre, London, 1977.

[267] B. P. Uvarov and B. N. Zolotarevsky. Phases of Locusts and their Interrelations. *Bulletin of Entomological Research*, 20(03):261, Oct. 1929.

[268] M. B. Van Hiel, P. Van Wielendaele, L. Temmerman, S. Van Soest, K. Vuerinckx, R. Huybrechts, J. V. Broeck, and G. Simonet. Identification and validation of housekeeping genes in brains of the desert locust Schistocerca gregaria under different developmental conditions. *BMC molecular biology*, 10:56–56, Jan. 2009.

[269] P. Van Wielendaele, S. Dillen, S. Zels, L. Badisco, and J. Vanden Broeck. Regulation of feeding by Neuropeptide F in the desert locust, Schistocerca gregaria. *Insect Biochemistry and Molecular Biology*, 43(1):102–114, Jan. 2013.

[270] J. Vandermeulen, C. Bahr, E. Tullo, I. Fontana, S. Ott, M. Kashiha, M. Guarino, C. P. H. Moons, F. A. M. Tuyttens, T. A. Niewold, and D. Berckmans. Discerning Pig Screams in Production Environments. *PLOS ONE*, 10(4):e0123111, Apr. 2015.

[271] J. a. Veenstra. Mono- and dibasic proteolytic cleavage sites in insect neuroendocrine peptide precursors. *Archives of insect biochemistry and physiology*, 43(2):49–63, 2000.

[272] J. a. Veenstra. The contribution of the genomes of a termite and a locust to our understanding of insect neuropeptides and neurohormones. *Frontiers in Physiology*, 5(November):1–22, 2014.

[273] J. A. Veenstra. Allatostatins C, double C and triple C, the result of a local gene triplication in an ancestral arthropod. *General and Comparative Endocrinology*, 230-231:153–157, May 2016.

[274] R. Velasco, A. Zharkikh, J. Affourtit, A. Dhingra, A. Cestaro, A. Kalyanaraman, P. Fontana, S. K. Bhatnagar, M. Troggio, D. Pruss, S. Salvi, M. Pindo, P. Baldi, S. Castelletti, M. Cavaiuolo, G. Coppola, F. Costa, V. Cova, A. Dal Ri, V. Goremykin, M. Komjanc, S. Longhi, P. Magnago, G. Malacarne, M. Malnoy, D. Micheletti, M. Moretto, M. Perazzolli, A. Si-Ammour, S. Vezzulli, E. Zini, G. Eldredge, L. M. Fitzgerald, N. Gutin, J. Lanchbury, T. Macalma, J. T. Mitchell,

J. Reid, B. Wardell, C. Kodira, Z. Chen, B. Desany, F. Niazi, M. Palmer, T. Koepke, D. Jiwan, S. Schaeffer, V. Krishnan, C. Wu, V. T. Chu, S. T. King, J. Vick, Q. Tao, A. Mraz, A. Stormo, K. Stormo, R. Bogden, D. Ederle, A. Stella, A. Vecchietti, M. M. Kater, S. Masiero, P. Lasserre, Y. Lespinasse, A. C. Allan, V. Bus, D. Chagné, R. N. Crowhurst, A. P. Gleave, E. Lavezzo, J. a. Fawcett, S. Proost, P. Rouzé, L. Sterck, S. Toppo, B. Lazzari, R. P. Hellens, C.-E. Durel, A. Gutin, R. E. Bumgarner, S. E. Gardiner, M. Skolnick, M. Egholm, Y. Van de Peer, F. Salamini, and R. Viola. The genome of the domesticated apple (Malus × domestica Borkh.). *Nat Genet*, 42(10):833–839, 2010.

[275] N. N. Vellichirammal, A. J. Zera, R. J. Schilder, C. Wehrkamp, J.-J. M. Riethoven, and J. A. Brisson. De Novo Transcriptome Assembly from Fat Body and Flight Muscles Transcripts to Identify Morph-Specific Gene Expression Profiles in Gryllus firmus. *PLOS ONE*, 9(1):e82129, Jan. 2014.

[276] R. Verdonck, W. De Haes, D. Cardoen, G. Menschaert, T. Huhn, B. Landuyt, G. Baggerman, K. Boonen, T. Wenseleers, and L. Schoofs. Fast and Reliable Quantitative Peptidomics with labelpepmatch. *Journal of Proteome Research*, 15(3):1080–1089, Mar. 2016.

[277] H. Verlinden, L. Badisco, E. Marchal, P. Van Wielendaele, and J. Vanden Broeck. Endocrinology of reproduction and phase transition in locusts. *General and comparative endocrinology*, 162(1):79–92, May 2009.

[278] H. Verlinden, R. Vleugels, R. Verdonck, E. Urlacher, J. Vanden Broeck, and A. Mercer. Pharmacological and signalling properties of a D2-like dopamine receptor (Dop3) in Tribolium castaneum. *Insect Biochemistry and Molecular Biology*, 56:9–20, Jan. 2015.

[279] M. Volkonsky. Le comportement du criquet pelerin (Schistocerca gregaria Forsk.) dans le sahara Algéro-Nigérien. *Archives de L'Institut Pasteur d'Algérie*, 20:236–248, 1942.

[280] H.-S. Wang, X.-H. Wang, C.-S. Zhou, L.-H. Huang, S.-F. Zhang, W. Guo, and L. Kang. cDNA cloning of heat shock proteins and their expression in the two phases of the migratory locust. *Insect Molecular Biology*, 16(2):207–219, Apr. 2007.

[281] X. Wang, X. Fang, P. Yang, X. Jiang, F. Jiang, D. Zhao, B. Li, F. Cui, J. Wei, C. Ma, Y. Wang, J. He, Y. Luo, Z. Wang, X. Guo, W. Guo, X. Wang, Y. Zhang, M. Yang, S. Hao, B. Chen, Z. Ma, D. Yu, Z. Xiong, Y. Zhu, D. Fan, L. Han, B. Wang, Y. Chen, J. Wang, L. Yang, W. Zhao, Y. Feng, G. Chen, J. Lian, Q. Li, Z. Huang, X. Yao, N. Lv, G. Zhang, Y. Li, J. Wang, J. Wang, B. Zhu, and L. Kang. The locust genome provides insight into swarm formation and long-distance flight. *Nature Communications*, 5, Jan. 2014.

[282] X. Wang and L. Kang. Molecular Mechanisms of Phase Change in Locusts. *Annual Review of Entomology*, 59(1):225–244, Jan. 2014.

[283] X. V. Wang, N. Blades, J. Ding, R. Sultana, and G. Parmigiani. Estimation of sequencing error rates in short reads. *BMC Bioinformatics*, 13:185, 2012.

[284] J. H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[285] Y. Wei, S. Chen, P. Yang, Z. Ma, and L. Kang. Characterization and comparative profiling of the small RNA transcriptomes in two phases of locust. *Genome Biology*, 10(1):R6, 2009.

[286] E. C. Winnebeck, C. D. Millar, and G. R. Warman. Why Does Insect RNA Look Degraded? *Journal of Insect Science*, 10(159):1–7, Sept. 2010.

[287] N. Wynant, D. Santos, R. Verdonck, J. Spit, P. Van Wielendaele, and J. Vanden Broeck. Identification, functional characterization and phylogenetic analysis of double stranded RNA degrading enzymes present in the gut of the desert locust, Schistocerca gregaria. *Insect Biochemistry and Molecular Biology*, 46:1–8, Mar. 2014.

[288] S. Zels, S. Dillen, K. Crabbé, J. Spit, R. J. Nachman, and J. Vanden Broeck. Sulfakinin is an important regulator of digestive processes in the migratory locust, Locusta migratoria. *Insect Biochemistry and Molecular Biology*, 61:8–16, June 2015.

[289] V. Zeng, B. Ewen-Campen, H. W. Horch, S. Roth, T. Mito, and C. G. Extavour. Developmental Gene Discovery in a Hemimetabolous Insect: De Novo Assembly and Annotation of a Transcriptome for the Cricket Gryllus bimaculatus. *PLoS ONE*, 8(5):e61479, May 2013.

[290] D. R. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, May 2008.

[291] Z.-J. Zhou, X.-Y. Kou, L.-Y. Qian, and J. Liu. Transcriptome profile of Chinese bush cricket, Gampsocleis gratiosa: A resource for microsatellite marker development: Transcriptome and microsatellite markers. *Entomological Research*, 46(3):197–205, May 2016.

[292] M. Ziemann, Y. Eren, and A. El-Osta. Gene name errors are widespread in the scientific literature. *Genome Biology*, 17(1):1–3, 2016.

FACULTY OF SCIENCE
DEPARTMENT OF BIOLOGY
MOLECULAR DEVELOPMENTAL PHYSIOLOGY AND SIGNAL TRANSDUCTION
Naamsestraat 59
B-3000 Leuven