Bioinformatics approaches for functional predictions in diverse informatics environments.

Paula Maria Moolhuijzen, BSc

This thesis is presented for the degree of Doctor of Philosophy of Murdoch University

2011

Declaration

I declare that this thesis is my own account of my research and contains as its main content work, which has not previously been submitted for a degree at any tertiary education institution.

Signature:

Name: Paula Moolhuijzen

Date: 4th October 2011

Abstract

Bioinformatics is the scientific discipline that collates, integrates and analyses data and information sets for the life sciences. Critically important in agricultural and biomedical fields, there is a pressing need to integrate large and diverse data sets into biologically significant information. This places major challenges on research strategies and resources (data repositories, computer infrastructure and software) required to integrate relevant data and analysis workflows. These challenges include:

- The construction of processes to integrate data from disparate and diverse resources and legacy systems that have variable data formats, qualities, availability and accessibility constraints.
- Substantially contributing to hypothesis driven research for biologically significant information.

The hypothesis proposed in this thesis is that in organisms from divergent origins, with differing data availability and analysis resources, *in silico* approaches can identify genomic targets in a range of disease systems. The particular aims were to:

1. Overcome data constraints that impact analysis of different organisms.
2. Make functional genomic predictions in diverse biological systems.
3. Identify specific genomic targets for diagnostics and therapeutics in diverse disease mechanisms.

In order to test the hypothesis three case studies in human cancer, pathogenic bacteria, and parasitic arthropod were selected, the results are as follows.

In case study 1 sequence information was integrated to make novel predictions, and generate novel findings for the role of the Alu repeat element in cancer. An under representation of Alu was found in cancerous transcript and most noncancerous Alu transcript found were of an unknown function. These findings led to an Alu-mediated siRNA model for the down regulation of Alu containing mRNA in cancer.

Case study 2, comparative genomic analyses identified venereal diagnostic targets that discriminated *Campylobacter fetus* subspecies *venerealis* from other *Campylobacter* species and subspecies. Plasmid borne virulence Type IV secretory pathway genes specificity however varied for biovars, compromising their use for diagnostics. These findings resulted in the targeted sequencing of *Campylobacter fetus* subspecies *venerealis* biovar genomes.

Case study 3, in cattle tick ectoparasite (*Rhipicephalus microplus*), a large highly complex and under researched genome, transcript sequence was analysed and tick vaccination targets identified. These vaccine candidates successfully imparted immunity in the bovine host. The developed high throughput vaccine target identification system is now being applied to other disease systems.

Through the shared bioinformatics approaches, novel functional targets and models in disease were determined. This thesis has developed and demonstrated *in silico* approaches for:

1. The collation, annotation and integration of data from divergent organisms with variable data constraints.
2. Novel functional predictions in diverse biological systems.
3. Novel vaccine and diagnostic candidate identification, in diverse disease mechanisms, substantially contributing to hypothesis driven research.

Table of Contents

Declaration ii

Abstract iii

Acknowledgements ix

Abbreviations x

Publications associated with this PhD thesis xiv

1 Chapter One – Introduction 1
- **1.1 Thesis structure** 1
- **1.2 Background and significance of study** 1
- **1.3 The aims of this thesis** 5

2 Chapter Two - Literature Review 9
- **2.1 General introduction** 9
- **2.2 Data mining and integration** 10
  - 2.2.1 Semantics and ontology 12
  - 2.2.2 Major data resources 15
  - 2.2.3 Bioinformatics workflow systems 20
- **2.3 Genomic analysis and functional predictions** 20
  - 2.3.1 Identification of disease and pathogenic targets 21
- **2.4 Informatics genomic case studies** 24
  - 2.4.1 Human TE disease 24
  - 2.4.2 Pathogenic Bacteria 26
  - 2.4.3 Ectoparasites and vector borne disease control 28

3 Chapter three - The transcript repeat element: the Human Alu sequence as a component of gene networks influencing cancer 32
- **3.1 Introduction** 32
  - 3.1.1 The functional roles of human TE elements 32
  - 3.1.2 Disease targets in human TE elements 36
- **3.2 Material and Methods** 40
  - 3.2.1 Databases and resources 40
  - 3.2.2 Sequence ontology 40
  - 3.2.3 Bioworkflows 41
  - 3.2.4 Alu Analysis Within the cDNA 41
  - 3.2.5 Gene Ontology 41
  - 3.2.6 PostgreSQL DB 42
  - 3.2.7 Mapping H-Inv Loci 42
- **3.3 Results and discussion** 43
  - 3.3.1 Bioinformatics workflow 43
  - 3.3.2 Differentiating between the cancerous and normal tissue information within the integrated H-Inv database 44
  - 3.3.3 Human Alu repeat sequence as a component of gene networks 46

3.3.4 Alu-transcript functions 50
3.3.5 Alu families and subfamilies 51
3.3.6 Alu locations within the transcript 52
3.3.7 Alu sequence quantification within transcripts 54
3.3.8 Incorporation of Alu elements as part of transcribed genes 55
3.3.9 The impact of Alu elements within the transcript UTR 56
3.3.10 The impact of Alu elements within transcribed exons 58
3.3.11 Alu-siRNA mediated feedback model in disease 60
3.4 Conclusion 64

4 Chapter Four - Genomic analysis of *Campylobacter fetus* subspecies: identification of candidate virulence determinants and diagnostic assay targets 66
4.1 Introduction 66
4.1.1 Bacterial virulence factors 67
4.1.2 Virulence targets for diagnostics 68
4.2 Materials and Methods 72
4.2.1 Bacterial strains, culture conditions and DNA preparation 72
4.2.2 Library construction, DNA sequencing and assembly 73
4.2.3 Genomic data 74
4.2.4 Alignment of genomic *Cfv* contigs based on *Cff* 74
4.2.5 *Cfv* Open reading frame identification & annotation 75
4.2.6 *Campylobacter* protein similarity to *Cfv* ORF 75
4.2.7 Putative virulence genes 76
4.2.8 Primer design 76
4.3 Results 77
4.3.1 Bioinformatics workflow 77
4.3.2 Assembly of *Cfv* for identifying targets for diagnostics 78
4.3.3 *Cfv* open reading frame analysis 83
4.3.4 *Cfv* Open reading frame analysis of the *Cfv* specific suite of genomic regions 83
4.3.5 *Cfv* IS*Cfe*1 insertion elements 84
4.3.6 Genomic plasmid analysis 85
4.3.7 COG Analysis -Virulence Genes 90
4.3.8 PCR diagnostics based on sequence identified in *Cfv* 95
4.4 Discussion 98
4.5 Conclusion 104

5 Chapter Five - Predicting gene targets in complex genomes: *Rhipicephalus microplus* target gene predictions for parasite control 106
5.1 Introduction 106
5.1.1 Functional roles of genes in ectoparasite required for feeding 106
5.1.2 Genomic targets for tick control 107
5.2 Materials and Methods 110
5.2.1 BAC end sequences 110
5.2.2 BAC genomic DNA extraction, library construction, and BAC screening and sequencing 111
5.2.3 BAC sequencing 111
5.2.4 BAC assembly 111
5.2.5 BES analyses 113
5.2.6 Gene prediction 113

5.2.7 Sequence alignment and phylogeny 114
5.2.8 Repeat identification 114
5.2.9 cDNA preparation 115
5.2.10 *Papilin* PCR amplification and sequencing 115
5.2.11 *Papilin* cloned products 117
5.2.12 *Helicase* PCR amplification and sequencing 117
5.2.13 BM-012-E08 PCR 117
5.2.14 BM-012-E08 Long range PCR 118
5.2.15 qRT-PCR analysis 119
5.2.16 Cot selected genomic DNA 119
5.3 Results 120
5.3.1 Bioinformatics workflow 120
5.3.2 Genome sequence via BES and Cot DNA 121
5.3.3 BAC Analysis 122
5.3.4 Selection of BAC clones for gene content: *Serpin* and *rRNA* 128
5.3.5 BAC BM-005-G14 assembly and analysis 128
5.3.6 BAC BM-012-E08 assembly and analysis 142
5.3.7 Bioinformatics workflow for vaccine candidate identification 152
5.3.8 Vaccine candidate tests 155
5.4 Discussion 161
5.4.1 Tick genomic structure: assembly and predictive models 161
5.4.2 Tick gene structure: predictive models 162
5.4.3 Tick DNA comparative studies: Identifying tick-specific sequence differences 166
5.4.4 Tick gene expression analysis 166
5.4.5 The analysis of genome sequence via BAC end sequencing and Cot DNA 167
5.4.6 Vaccine candidate identification 169
5.5 Conclusion 169

6 Chapter Six - Conclusion 172
6.1 Thesis contribution to the field of bioinformatics 172
6.2 Case study chapter results 174
6.3 Discussion and future work 180
6.4 Summary Conclusion 181

Appendix 183

References 234

Acknowledgements

This thesis would not have been possible without the support of many good people. Thank you to my supervisors Professor Rudi Appels and Professor Matthew Bellgard for the opportunity to under take this thesis, and for their expert guidance and support. To the team at the Centre for Comparative Genomics, especially Mr David Schibeci, Mr Mark O'Shea, Dr Roberto Barrero and Mr Adam Hunter, thank you for your expert help and support. I would also like to acknowledge key collaborators at the DEEPI in Queensland and the BeefCRC, especially Dr Ala Lew-Tabor, Dr Manuel Rodreguez-Valle, Dr Felix Guerrero (ARS-USDA) and Dr Jessica Morgan. Last but not least, my deepest gratitude to my family and partner for their support and encouragement.

Abbreviations

Ab - Antibodies

Ag - Antigens

API - Application programming interface

BES – BAC End Sequence

BLAST - Basic Local Alignment Search Tool

BmiGI – *Boophilus microplus* Gene Index

CCG - Center for Comparative Genomics

COG - Clusters of Orthologous Groups

cDNA - cloned DNA

DDBJ - DNA Data Bank of Japan

DE - Differentially expressed

DFCI - Dana Faber Cancer Institute

DNA - Deoxyribonucleic acid

DPI - Department of Primary Industries, QLD

dsRNA - double-stranded RNA

EMBL - European Molecular Biology Laboratory

EBI - European Bioinformatics Institute

EST - Expressed sequence tag

FB - Fold Back

GI - Gene Index / Genomic Islands

GIRI – Genetic Information Research Institute

GO - Gene Ontology

GWAS - Genome-Wide Association Studies

HDI - Histone Deacetylase Inhibition

HMM - Hidden Markov Model

HPLC - High-Performance Liquid Chromatography

HR – Highly Repetitive

HTP – High ThroughPut

IGP – *Ixodes scapularis* Genome Project

IS - Insertion Sequence

JBIRC - Japan biological Information Research Center

KOG - Eukaryote Clusters of Orthologous Groups

LSU – Long SubUnit

LTR - Long Terminal Repeat

MGE - Mobile Genetic Elements

MR – Moderately Repetitive

miRNA - micro RNA

mRNA - messenger RNA

MSA - Multiple sequence alignment

NCBI - National Center for Biotechnology Information

ncRNA – non-protein coding RNA

NIH - National Institute of Health

NLM - National Library of Medicine

NMD - Nonsense Mediated Decay

OIE – Office International des Epizooties

ORF - Open Reading Frame

PAI – Pathogenic Islands

PANTHER - Protein ANalsis THrough Evolutionary Relationships

PCR - Polymerase Chain Reaction

pFAM - protein FAMily

PFGE – Pulse Field Gel Electrophoresis

PID - Percent IDentity

qRT-PCR - quantitative Real-Time PCR

QTL - Quantitative Trait Loci

RNA - RiboNucleic Acid

RNAi - RNA interference

SINE – Short Interspersed repetitive Element

siRNA - small interfering RNA

SNP - Single Nucleotide Polymorphism

SSH – Suppressive Subtractive Hybridization

SSU – Small SubUnit

SQL - Simple Query Language

TE - Transposable Element

TC - Tentative Consensus sequences

TGI -TIGR Gene Index

UNSAM - La Universidad Nacional de San Martín

UPM - Universal Primer Mix

UTR - Un-Translated Region

VF – Virulence Factors

VI – Vaccine Identification

Publications associated with this PhD thesis

1. Moolhuijzen P, Kulski JK, Dunn DS, Schibeci D, Barrero R, Gojobori T, Bellgard M: The transcript repeat element: the human Alu sequence as a component of gene networks influencing cancer. Funct Integr Genomics 2010, 10(3):307-319.
2. Moolhuijzen PM, Lew-Tabor AE, Wlodek BM, Aguero FG, Comerci DJ, Ugalde RA, Sanchez DO, Appels R, Bellgard M: Genomic analysis of *Campylobacter fetus* subspecies: identification of candidate virulence determinants and diagnostic assay targets. BMC Microbiol 2009, 9:86.
3. Lew AE, Guo S-Y, Venus B, Moolhuijzen P, Sanchez D, Trott D, Burrell P, Wlodek B, Bellgard M: Comparative genome analysis applied to develop novel PCR assays to characterise and identify *Campylobacter fetus* subsp. *venerealis* isolates. Zoonoses and Public Health 2007 54(Supplement 1):154.
4. Moolhuijzen P, Lew-Tabor A, Morgan ATJ, Rodriguez Valle M, Peterson GD, Dowd S. E, Guerrero F, Bellgard M, Appels R: The complexity of *Rhipicephalus* (*Boophilus*) *microplus* genome characterised through detailed analysis of two BAC clones. BMC Research Notes 2011, 22;4:254.
5. Bellgard MI, Moolhuijzen PM, Guerrero F.D., Appels R, Schibeci D, Rodriguez-Valle M, Barrero R, Hunter A, Lew-Tabor AE: CattleTickBase: Internet-based analysis tools and bioinformatics repository of available

genomics resources for *Rhipicephalus* (*Boophilus*) *microplus*. International Journal for Parasitology 2011, In review.

6. Guerrero FD, Moolhuijzen PM, Peterson DG, Bidwell S, Caler E, Appels R, Bellgard M, Nene VM, Djikeng A: Reassociation kinetics-based approach for partial genome sequencing of the cattle tick, *Rhipicephalus* (*Boophilus*) *microplus*. BMC Genomics 2010, 11:374.
7. Lew-Tabor AE, Moolhuijzen PM, Vance ME, Kurscheid S, Valle MR, Jarrett S, Minchin CM, Jackson LA, Jonsson NN, Bellgard MI et al: Suppressive subtractive hybridization analysis of *Rhipicephalus* (*Boophilus*) *microplus* larval and adult transcript expression during attachment and feeding. Vet Parasitol 2009, 167(2-4):304-320.