MONASH University

Bayesian methods for identifying non-protein coding genomic regions contributing to diseases

Manjula D Algama Appuhamilage Dona

B.Sc.(Hons), University of Colombo, Sri Lanka

A thesis submitted for the degree of Doctor of Philosophy

at the School of Mathematical Sciences

Monash University, Australia

October 2015

Contents

Copyright Notice

Abstract

Identifying and discerning the function of non-coding RNAs (ncRNAs) is an important goal of genetic research. Much evidence suggests that ncRNAs play an important role in the aetiology of many complex genetic diseases. Therefore the task of developing methods to identify these elements in genomes has become increasingly urgent.

In this research my focus was to use a Bayesian approach to identify putative functional non-coding genomic sequences contributing to various diseases. The analysis was mainly carried out using a Bayesian segmentation model, implemented in the software package *changept*, designed to segment discrete genomic data. In the first phase of the research, I developed methods to expand the capabilities of *changept*. One simple but powerful innovation was to develop several ways of encoding an alignment of sequences using a D-character representation (D is a positive integer). This enables sequence alignments to be segmented based on multiple data types: specifically conservation, GC content and transition/transversion ratio and significantly generalizes the capacity of *changept*, which previously could only segment on the basis of one of these characteristics at a time. Incorporating multiple data types greatly helped to clearly identify complex segmentation patterns and functional signatures among species, especially between closely related species. A second methodological innovation was a new model selection procedure to decide the optimal model for the data. A third, and most important, methodological innovation was to build a process for systematically discovering genome-wide putative ncRNAs, including data selection, cleaning, encoding, analysis and

post-processing. To validate these findings, both experimental methods and currently available bioinfomatics resources were used.

In the second phase of the research, my focus turned to application of *changept*, and the new methods developed, to identify genome-wide putative non-coding elements that may be associated with diseases. I was able to discover more than a thousand highly conserved non-coding sequences in human, mouse and zebrafish genomes. A complementary analysis focused on a set of genes involved in muscle development. Some of these elements identified may contribute to muscle diseases. Discovery of putative small ncRNAs in the bacterium *Wolbachia pipientis* is another successful application of the new methods; this work was undertaken as part of the eradicate dengue project. Application to malaria genomes revealed genetic mechanisms important in infecting multiple hosts. I also identified putative regulatory sequences in $3'$ UTRs in 3 closely related *Drosophila* species. Although this work focussed on *Drosophila* rather than human diseases, mutations in $3'$ UTRs have been shown to play a crucial role in human health and diseases.

Acknowledgement

First and foremost I would like to thank my main supervisor, Assoc Prof Jonathan Keith, the most amazing person I have ever worked with. Jon's wealth of knowledge, endless patience in teaching, taking active interest in my research and high availability had a significant impact on the completion of this work. From the very beginning he encouraged me to write as I go which helped to own many publications during my PhD. I am very grateful for his guidance, motivation, having faith in my work and specially giving me freedom to work from home.

I would also like to thank my associate supervisor, Dr Robert Bryson-Richardson at the School of Biological Sciences. Rob was very approachable and has been taking care of biological aspects of my research. I am very grateful for his abundance of knowledge and many insightful discussions, comments and suggestions. Also a big thanks to Edward Tasker who worked as a research assistant, for being supportive. I would also like to express my appreciation to Dr Meg Woolfit, Dr Anders Goncalves da Silva, Dr Sarah Boyd and Caitlin Williams for their collaborations on Chapters 4-6. I gratefully acknowledge Philip Chan at Monash IT for helping to solve all my IT related problems.

Most important from all, my beloved husband Gayan, words cannot express how grateful I am for your love and support. Thank you for all the sacrifices you made on my behalf. To my dearest daughter, Mineli, thank you for being such a good girl. Finally with greatest respect a special thanks to my mom, aunt and all my friends for their unconditional love and care especially during this period.

The List of Publications

Chapter 2 : Algama M, Keith JM. (2014). Investigation of genomic structure using *changept*: A Bayesian segmentation model. *Computational and Structural Biotechnology Journal* 10, 107-115.

Chapter 3 : Algama M, Oldmeadow C, Tasker E, Mengersen K, Keith JM. (2014). Drosophila 3′ UTRs are more complex than protein-coding sequences. *PLoS ONE* 9(5): e97336. doi:10.1371/journal.pone.0097336.

Chapter 4 : Algama M, Tasker E, Williams C, Parslow AC, Bryson-Richardson RJ, Keith JM. (2015). Genome-wide identification of ncRNAs using a Bayesian segmentation approach. In preparation. Target journal: *Genome Research.*

Chapter 5 : Woolfit M, Algama M, Keith JM, McGraw EA, Popovici J. (2015). Discovery of putative small non-coding RNAs from the obligate intracellular bacterium Wolbachia pipientis. *PLoS ONE* 10(3): e0118595. doi:10.1371/journal.pone.0118595.

Chapter 6 : Goncalves da Silva A, Algama M, Tasker E, Sunnucks P, Clarke RH, Keith JM. (2015). Host-generalism in blood parasites: a case for reversible host-specialization. In preparation. Target journal: *Nature Genetics.*

PART A: General Declaration

Monash University

Declaration for thesis based or partially based on conjointly published or unpublished work

General Declaration

In accordance with Monash University Doctorate Regulation 17.2 Doctor of Philosophy and Research Master's regulations the following declarations are made:

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes 3 original papers published in peer reviewed journals and 2 unpublished publications. The core theme of the thesis is identifying non-protein-coding genomic regions using Bayesian methods. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the candidate, working within the School of Mathematical Sciences at Monash University under the supervision of A/Prof. Jonathan Keith and Dr. Robert-Bryson Richardson.

The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.

In the case of Chapters, 2 (part 2) to 6 my contribution to the work involved the following:

Thesis chapter	Publication title	Publication status*	Nature and extent of candidate's contribution
2-part 2	Investigation of genomic structure using *changept*: A Bayesian segmentation model	Published	Located and reviewed articles, wrote the paper.
3	Drosophila 3' UTRs are more complex than protein-coding sequences	Published	Method developments, performed experiments, analysed data, wrote the paper.
4	Genome-wide identification of ncRNAs using a Bayesian segmentation approach	To be submitted	Conceived methods, scripting, designed and performed computational experiments, analysed data, wrote the paper.
5	Discovery of putative small non-coding RNAs from the obligate intracellular bacterium Wolbachia pipientis	Published	Scripting, performed experiments, analysed data, wrote parts of the paper.
6	Host-generalism in blood parasites: a case for reversible host-specialization	To be submitted	Method developments, scripting, performed experiments, analysed data, wrote parts of the paper.

I have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Signed: [redacted] .....

Date: 9/10/15 .....

Chapter 1

Introduction

Genetic information is stored in Deoxyribonucleic acid (DNA), it is transferred into Ribonucleic acid (RNA) through a process called transcription and then possibly transferred to a protein molecule via translation. RNAs are divided into two classes: messenger RNAs (mRNAs), which are translated into proteins, and non-coding RNAs (ncRNAs), that do not encode a protein. It has been estimated that around 98% of the human genomic output is ncRNAs [1]; however, what proportion of ncRNAs are functional remains debatable [2, 3].

There are 2 groups of ncRNAs, short ncRNAs (ones shorter than $\approx$ 200nt: micro RNAs -miRNAs; small nucleolar RNAs -snoRNAs; small interfering RNAs -siRNAs; small nuclear RNAs -snRNAs; PIWI interacting RNAs -piRNAs; transfer RNAs -tRNAs) and long ncRNAs, those longer than $\approx$ 200nt (lncRNAs). Based on the regions in which they are expressed, lncRNAs fall into 4 categories: (1) Antisense lncRNAs - transcribed from the opposite strand of protein- coding genes; (2) Intronic lncRNAs - transcribed completely from within a single intron of a protein- coding gene; (3) Bidirectional lncRNAs which share promoters with protein-coding genes, but transcribed in the opposite direction; and (4) Intervening lncRNAs - transcribed from regions that are at least 5kb or more from protein-coding genes [4]. ncRNAs have been found to

carry out a variety of diverse functions, including transcription, splicing, translation, regulation of gene expression, chromatin modification, regulation of differentiation and development, regulation of epigenetic processes, and RNA modification [4–25]. Much evidence suggests that ncRNAs contribute to diseases. Alzheimer's disease (AD) is a chronic neurodegenerative disease associated with expression in *BACE1-AS*, an antisense RNA [26]. It's also found that the lncRNA *BC200* is expressed at higher levels in AD patients in comparison to controls [27, 28]. Many ncRNAs, including miRNAs, snoRNAs and lncRNAs show abnormal expression patterns in cancerous tissues [4, 29–42]. A recent study found that the expression of PIWI interacting RNA, *piR-651*, in gastric, colon, lung, and breast cancer tissues was higher than that in paired non-cancerous tissues [43]. More recently, Cheng *et al.* [44] discovered that the expression level of another piRNA, *piR-823*, in gastric cancer tissues was significantly lower than that in non-cancerous tissues. Several lines of evidence suggest key roles of miRNAs in diseases, for example cardiovascular disorders [45, 46] and spinal motor neuron disease [47]. Among other evidence, the role of miRNAs in type 2 diabetes was first established in 2004 [48] where it is shown that *miR-375* is directly involved in the regulation of insulin secretion. The recent work published by Fernandez-Valverde *et al.* [49] reveals the roles and effects of miRNAs in type 1 and type 2 diabetes. Many of the diseases associated with ncRNAs are reviewed in [50–59].

Given the emerging picture of the importance of ncRNAs, the task of developing methods to identify these elements has become increasingly urgent. Various experimental and computational methods used to identify ncRNAs are presented in [60–64]. One statistical technique for identifying putative functional elements in genomes, including ncRNAs, is known as *sequence segmentation*. This technique involves partitioning of genomic sequences into compositionally homogeneous blocks. Genomes can be segmented based on atypical sequence characteristics, such as conservation levels relative to other genomes, GC content, SNP frequency, transition/transversion ratio and potentially many others [65].

Keeping all these in mind, the main objective of this thesis is to:

'develop methods to identify putative functional non-protein coding genomic regions contributing to diseases'

In achieving this objective, I used a Bayesian DNA segmentation algorithm, *changept*, [66, 67] throughout this research. The main method developments implemented were centred around this program. In brief, *changept* can be described as a segmentation-classification model. It is capable of simultaneously segmenting a genomic alignment and classifying segments into one of a predefined number of segment classes. In Chapter 2 - Literature Review: Part 2, I have discussed many of the currently used genome segmentation methods including change-point analysis. I have also explained the mathematics of *changept* and have summarised few applications of it. This work has been published in Computational and Structural Biotechnology journal [68].

The focus of Chapter 3 is to introduce new methods into program *changept*, so that it can be effectively applied on different genomes to find putative functional elements. Introducing an 8-character representation to encode a pair-wise alignment and a 32-character representation to encode a 3-way alignment capturing information about conservation, GC content and transition/transversion ratios was a simple but powerful methodological development described in this chapter. The second methodological innovation was an alternative model selection procedure, less conservative than an earlier method based on investigating DICV values (type V of Deviance Information Criterion, [69]) to identify the most likely model for the data. I wrote several scripts in perl and R for various kinds of processing related to these two main methodological developments (for example a perl code to transform alignments and R code to produce different types of trace plots). These codes are currently being integrated into a complete package with GUI (Graphical User Interface). I applied these new methods to a data set of three closely related *Drosophila* species. An unexpected discovery was made: that segmentation patterns in $3'$ UTRs of the species *D. melanogaster*, *D. simulans* and *D. yakuba* are more complex than in the protein-coding regions. In

this analysis, a number of known and putative miRNA targets in 3′ UTR regions in *D. melanogaster* were also identified. This work was published in the journal PLoS ONE [70] and also presented as a poster at two international conferences: the 13^{th} International Conference of Bioinformatics in Sydney (awarded as the best poster -Gold) and the BioInfo Summer 2014 conference held in Melbourne (where it won the 2^{nd} place poster prize).

I continued to explore methods and bioinformatics resources currently available to discover functional ncRNAs and other regulatory sequences in various genomes. This motivates the work of Chapter 4, where I developed a process to systematically identify genome-wide intronic putative functional elements (PFEs) in human, mouse and zebrafish using *changept* and methods developed in the previous chapter (Chapter 3). The majority of the PFEs identified were compared with regions predicted by other computational methods (EvoFold [71] and RNAz [72]) and bioinformatics resources (DNase I footprints data [73] and fRNAdb entries -functional RNA database [74]) in addition to experimental validation. In the same chapter, I also carried out a pathway-specific analysis discovering 27 PFEs in a set of genes involved in muscle development. Although the specific functions are unknown, these elements may contribute to muscle disease and variation of severity of several diseases. It is known that mutations in these genes can cause many conditions. For example, *eya1* can cause a syndrome including deafness; *eya4*- deafness and cardiomyopathy; *pax3*- waardenburg syndrome and rhabdomyosarcoma; *pax7*- rhabodmyosarcoma; *six1*- branchiotic syndrome and deafness; and *wnt1*- osteogenesis imperfect (`http://omim.org/`). Further research is required to assess the specific roles of these PFEs. This work is currently being revised for submission to the high profile journal *Genome Research*.

In the second phase of the research, the main objective was to discover non-coding genomic regions associated with diseases by using *changept* and methods developed in Chapters 3 and 4. In Chapter 5, this was achieved using a dataset from *Wolbachia*

pipientis, a bacterium that induces a wide range of effects in its insect hosts, including manipulation of reproduction and protection against pathogens. In particular, *Wolbachia* infected mosquitoes are currently being tested in trials with the aim of reducing dengue virus transmission [75, 76]. This study was aimed at understanding how *Wolbachia* interacts with its host species. In particular, two novel putative sRNAs that may play significant roles in the biology of *Wolbachia* were identified. This work has been published with the journal PLoS ONE [77].

By using a different disease related dataset, in Chapter 6, I successfully used *changept* to segment three malaria genomes, *Plasmodium falciparum* - the human malaria; *P. reichnowi* - which infects chimpanzees, and *P. gallinaceum* - which infects jungle fowls. The main analysis performed in this chapter was examining the relationship between different functional groups of genes (eg: general transcription factors, chromatin related factors, specific transcription factors [78]) and the *changept* segment classes which represent different functionally constrained genomic regions. The goal of this particular application was to identify genetic mechanisms that make *host generalists* or species that are able to thrive on a variety of resources. These results will help to warn us about the next possible host of these malaria species, hence will contribute to build *early -warning* systems for disease emergence. This work is in preparation to be submitted as a letter to *Nature*.

As this thesis is written in fulfilment of the requirement for 'thesis by publication', Chapters 2 (part 2) to 6 are comprised of journal articles. Each of these chapters contain a section on *changept* modelling for the purpose of publication. Therefore, the method sections partially overlap. Furthermore, some of the *changept* applications summarised in Chapter 2 (part 2) pre-empt the more detailed discussions in succeeding chapters. The bibliography of each chapter can be found at the end of the chapter.

In summary, the main objective of this thesis is to develop methods to identify putative functional non-protein coding genomic regions associated with disease development. To achieve this, I have explored the following aspects:

1. Develop methods to enhance the applicability of program *changept* (Chapter 3).

2. Build a process to systematically identify putative ncRNAs and other regulatory elements using *changept* and current bioinformatics resources (Chapter 4).

3. Identify putative functional non-coding elements that may contribute to various diseases.

   (a) Muscle disease (Chapter 4)

   (b) Dengue (Chapter 5)

   (c) Malaria (Chapter 6)

Bibliography

[1] J S Mattick. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep*, 2:986–991, 2001.

[2] W F Doolittle. Is junk DNA bunk? a critique of ENCODE. *Proceedings of the National Academy of Sciences of the USA*, 110:5294–5300, 2013.

[3] G Graur, Y Zheng, N Price, R B R Azevedo, R A Zufall, and E Elhaik. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution*, 5:578–590, 2013.

[4] G G Carmichael, editor. *Regulatory NonCoding RNAs*, chapter Diverse Functions and Mechanisms of Mammalian Long Noncoding RNAs, pages 1–14. Springer, New York, 2015.

[5] N Brockdorff, A Ashworth, G F Kay, V M McCabe, D P Norris, P J Cooper, S Swift, and S Rastan. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell*, 71:515–526, 1992.

[6] C J Brown, B D Hendrich, J L Rupert, R G Lafreniere, Y Xing, J Lawrence, and H F Willard. The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, 71:527–542, 1992.

[7] M A Ripoche, C Kres, F Poirier, and L Dandolo. Deletion of the H19 transcription unit reveals the existence of a putative imprinting control element. *Genes Dev*, 11:1596–1604, 1997.

[8] V H Meller, K H Wu, G Roman, M I Kuroda, and R L Davis. roX1 RNA paints the X chromosome of male drosophila and is regulated by the dosage compensation system. *Cell*, 88:445–457, 1997.

[9] J T Lee, L S Davidow, and D Warshawsky. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nature Genet*, 21:400–404, 1999.

[10] T Sado, Z Wang, H Sasaki, and E Li. Regulation of imprinted X-chromosome inactivation in mice by Tsix. *Development*, 128:1275–1286, 2001.

[11] F Sleutels, R Zwart, and D P Barlow. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, 415:810–813, 2002.

[12] N Thakur, V K Tiwari, H Thomassin, R R Pandey, M Kanduri, A Gondor, T Grange, R Ohlsson, and C Kanduri. An antisense RNA regulates the bidirectional silencing property of the Kcnq1 imprinting control region. *Mol. Cell. Biol*, 24:7855–7862, 2004.

[13] T L Young, T Matsuda, and C L Cepko. The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. *Curr. Biol*, 15:501–512, 2005.

[14] M R Ginger, A N Shore, A Contreras, M Rijnkels, J Miller, M F Gonzalez-Rimbau, and J M Rosen. A noncoding RNA is a potential marker of cell fate

during mammary gland development. *Proc. Natl Acad. Sci. USA*, 103:5781–5786, 2006.

[15] J S Mattick and I V Makunin. Non-coding RNA. *Human Molecular Genetics*, 15: R17–R29, 2006.

[16] S Swiezewski, F Liu, A Magusin, and C Dean. Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature*, 462:799–802, 2009.

[17] A M Khalil, M Guttman, M Huarte, M Garber, A Raj, D Rivea Morales, and J L Rinn. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci*, 106:11667–11672, 2009.

[18] R W Carthew and E J Sontheimer. Origins and Mechanisms of miRNAs and siRNAs. *Cell*, 136:642–655, 2009.

[19] J S Mattick. RNA as the substrate for epigenome-environment interactions: RNA guidance of epigenetic processes and the expansion of RNA editing in animals underpins development, phenotypic plasticity, learning, and cognition. *Bioessays*, 32:548–552, 2010.

[20] M J Koziol and J L Rinn. RNA traffic control of chromatin complexes. *Curr Opin Genet Dev*, 20:142–148, 2010.

[21] J S Mattick. The central role of RNA in human development and cognition. *FEBS Lett*, 585:1600–1616, 2011.

[22] M E Askarian-Amiri, J Crawford, J D French, C E Smart, M A Smith, M B Clark, K Ru, T R Mercer, E R Thompson, S R Lakhani, and et al. SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer. *RNA*, 17:878–891, 2011.

[23] M Guttman, J Donaghey, B W Carey, M Garber, J K Grenier, G Munson, and E S Lander. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, 477:295–300, 2011.

[24] M Kretz, Z Siprashvili, C Chu, D E Webster, A Zehnder, K Qu, C S Lee, R J Flockhart, A F Groff, J Chow, and et al. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature*, 493:231–235, 2013.

[25] Mattick J S Mercer T R. Structure and function of long noncoding RNAs in epigenetic regulation. *Nature Structural and Molecular Biology*, 20:300–307, 2013.

[26] M A Faghihi, F Modarresi, A M Khalil, D E Wood, B G Sahagan, T E Morgan, C E Finch, G Laurent, P J Kenny, and C Wahlestedt. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat Med*, 14:723–730, 2008.

[27] E Mus, P R Hof, and H Tiedge. Dendritic BC200 RNA in aging and in Alzheimers disease. *Proc Natl Acad Sci U S A*, 104:10679–10684, 2007.

[28] H Tiedge, W Chen, and J Brosius. Primary structure, neural-specifi c expression, and dendritic location of human BC200 RNA. *J Neurosci*, 13:2382–2390, 1993.

[29] L Pibouin, J Villaudy, D Ferbus, M Muleris, M T Prosperi, Y Remvikos, and G Goubin. Cloning of the mRNA of overexpression in colon carcinoma-1: a sequence overexpressed in a subset of colon carcinomas. *Cancer Genet Cytogenet*, 133:55–60, 2002.

[30] X Fu, L Ravindranath, N Tran, G Petrovics, and S Srivastava. Regulation of apoptosis by a prostate-specific and prostate cancer-associated noncoding gene, PCGEM1. *DNA Cell Biol*, 25:135–141, 2006.

[31] M Mourtada-Maarabouni, M R Pickard, V L Hedge, F Farzaneh, and G T Williams. GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. *Oncogene*, 28:195–208, 2009.

[32] X Y Dong, P Guo, J Boyd, X Sun, Q Li, W Zhou, and J T Dong. Implication of snoRNA U50 in human breast cancer. *J Genet Genomics*, 36:447–454, 2009.

[33] C P Christov, E Trivier, and T Krude. Noncoding human Y RNAs are overexpressed in tumours and required for cell proliferation. *Br J Cancer*, 98:981–989, 2008.

[34] R A Gupta, N Shah, K C Wang, J Kim, H M Horlings, D J Wong, M C Tsai, T Hung, P Argani, J L Rinn, and et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 464:1071–1076, 2010.

[35] R Kogo, T Shimamura, K Mimori, K Kawahara, S Imoto, T Sudo, F Tanaka, K Shibata, A Suzuki, S Komune, S Miyano, and M Mori. Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modifi cation and is associated with poor prognosis in colorectal cancers. *Cancer Res*, 71:6320–6326, 2011.

[36] D Khaitan, M E Dinger, J Mazar, J Crawford, M A Smith, J S Mattick, and R J Perera. The melanoma-upregulated long noncoding RNA SPRY4-IT1 modulates apoptosis and invasion. *Cancer Res*, 71:3852–3862, 2011.

[37] Z Yang, L Zhou, L M Wu, M C Lai, H Y Xie, F Zhang, and S S Zheng. Overexpression of long non-coding RNA HOTAIR predicts tumor recurrence in hepatocellular carcinoma patients following liver transplantation. *Ann Surg Oncol*, 18:1243–1250, 2011.

[38] T Gutschner, M Hammerle, M Eissmann, J Hsu, Y Kim, G Hung, A Revenko, G Arun, M Stentrup, M Gross, and et al. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res*, 73:1180–1189, 2013.

[39] T Chiyomaru, S Fukuhara, S Saini, S Majid, G R Deng, V Shahryari, I Chang, Y Tanaka, H Enokida, M Nakagawa, R Dahiya, and S Yamamura. Long non-coding RNA HOTAIR Is targeted and regulated by miR-141 in human cancer cells. *J Biol Chem*, 289:12550–12565, 2014.

[40] R Bottcher, A M Hoogland, N Dits, E Verhoef, C Kweldam, P Waranecki, C H Bangma, G J van Leenders, and G Jenster. Novel long non-coding RNAs are specific diagnostic and prognostic markers for prostate cancer. *Oncotarget*, 6: 4036–4050, 2015.

[41] M Morlando, M Ballarino, and A Fatica. Long Non-Coding RNAs: New Players in Hematopoiesis and Leukemia. *Frontiers in Medicine*, 2:23, 2015.

[42] M Vitiello, A Tuccoli, and L Poliseno. Long non-coding RNAs in cancer: implications for personalized therapy. *Cellular Oncology*, 38:17–28, 2015.

[43] J Cheng, J M Guo, B X Xiao, Y Miao, Z Jiang, H Zhou, and Q N Li. piRNA, the new non-coding RNA, is aberrantly expressed in human cancer cells. *Clin Chim Acta*, 412:1621–1625, 2011.

[44] J Cheng, H Deng, B Xiao, H Zhou, F Zhou, Z Shen, and J Guo. piR-823, a novel non-coding small RNA, demonstrates in vitro and in vivo tumor suppressive activity in human gastric cancer cells. *Cancer Lett*, 315:12–17, 2012.

[45] C K Sen, G M Gordillo, S Khanna, and S Roy. Micromanaging vascular biology: tiny microRNAs play big band. *J Vasc Res*, 46:527–540, 2009.

[46] K G Barringhaus and P D Zamore. microRNAs: Regulating A Change of Heart. *Circulation*, 119:2217–2224, 2009.

[47] S Haramati, E Chapnik, Y Sztainberg, R Eilam, R Zwang, N Gershoni, E McGlinn, P W Heiser, A M Wills, I Wirguin, and et al. miRNA malfunction causes spinal motor neuron disease. *Proc Natl Acad Sci U S A*, 107:13111–13116, 2010.

[48] M N Poy, L Eliasson, J Krutzfeldt, S Kuwajima, X Ma, P E Macdonald, S Pfeffer, T Tuschl, N Rajewsky, P Rorsman, and M Stoffel. A pancreatic islet-specific microRNA regulates insulin secretion. *Nature*, 432:226–230, 2004.

[49] S L Fernandez-Valverde, R J Taft, and J S Mattick. MicroRNAs in -cell biology, insulin resistance, diabetes and its complications. *Diabetes*, 60:1825–1831, 2011.

[50] R J Taft, K C Pang, T R Mercer, M Dinger, and J S Mattick. Non-coding RNAs: regulators of disease. *The journal of pathology*, 220:126–139, 2010.

[51] M Esteller. Non-coding RNAs in human disease. *Nature Reviews Genetics*, 12: 861, 2011.

[52] O Wapinski and H Y Chang. Long noncoding RNAs and human disease. *Trends in Cell Biology*, 21:354–362, 2011.

[53] J Li, Z Xuan, and C Liu. Long Non-Coding RNAs and Complex Human Diseases. *Int. J. Mol. Sci*, 14:18790–18808, 2013.

[54] P Rao, E Benito, and A Fischer. MicroRNAs as biomarkers for CNS disease. *Front Mol Neurosci*, 6:39, 2013.

[55] Y Huang, J Wang, X Yu, Z Wang, T Xu, and X Cheng. Non-coding RNAs and diseases. *Molecular Biology*, 47:465–475, 2013.

[56] D Vucicevic, H Schrewe, and A Ulf. Molecular mechanisms of long ncRNAs in neurological disorders. *Frontiers in Genetics*, 5:48, 2014.

[57] H Soreq. Novel roles of non-coding brain RNAs in health and disease. *Front Mol Neurosci*, 7:55, 2014.

[58] T R Mercer, M E Dinger, and J S Mattick. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics*, 10:155–159, 2009.

[59] K V Morris and J S Mattick. The rise of regulatory RNA. *Nature Reviews Genetics*, 15:423–437, 2014.

[60] A Huttenhofer and J Vogel. Experimental approaches to identify non-coding RNAs. *Nucleic Acids Research*, 34:635–646, 2006.

[61] G Solda, I V Makunin, O U Sezerman, A Corradin, G Corti, and A Guffanti. An Ariadne's thread to the identification and annotation of noncoding RNAs in eukaryotes. *Brief Bioinform*, 10:475–489, 2009.

[62] N E Ilott and C P Ponting. Predicting long non-coding RNAs using RNA sequencing. *Methods*, 63:50–59, 2013.

[63] A Pauli, E Valen, M F Lin, M Garber, N L Vastenhouw, J Z Levin, L Fan, A Sandelin, J L Rinn, and et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res*, 22:577–591, 2012.

[64] J S Mattick and J L Rinn. Discovery and annotation of long noncoding RNAs. *Nature Structural and Molecular Biology*, 22:5–7, 2015.

[65] J M Keith. Sequence segmentation. *Methods Mol Biol*, 452:207–229, 2008.

[66] J M Keith. Segmenting eukaryotic genomes with the Generalized Gibbs Sampler. *Journal of Computational Biology*, 13:1369–1383, 2006.

[67] J M Keith, P Adams, S Stephen, and J S Mattick. Delineating slowly and rapidly evolving fractions of the Drosophila genome. *Journal of Computational Biology*, 15:407–430, 2008.

[68] M Algama and J M Keith. Investigating genomic structure using *changept*: A Bayesian segmentation model. *Computational and Structural Biotechnology Journal*, 10:107–115, 2014.

[69] C Oldmeadow and J M Keith. Model Selection in Bayesian Segmentation of multiple DNA alignments. *Bioinformatics*, 27:604–610, 2011.

[70] M Algama, C Oldmeadow, E Tasker, K Mengersen, and J M Keith. Drosophila 3' UTRs Are More Complex than Protein-Coding Sequences. *PLoS ONE*, 9:e97336, 2014.

[71] J S Pedersen, G Bejerano, A Siepel, K Rosenbloom, K Lindblad-Toh, E S Lander, J Kent, W Miller, and D Haussler. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol*, 2:e33. DOI: 10.1371/journal.pcbi.0020033, 2006.

[72] A R Gruber, S Findei, S Washietl, I L Hofacker, and P F Stadler. RNAz 2.0: Improved noncoding RNA detection. *Pac Symp Biocomput*, 15:69–79, 2010.

[73] S Neph, J Vierstra, A B Stergachis, A P Reynolds, E Haugen, B Vernot, R E Thurman, S John, R Sandstrom, A K Johnson, and et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489:83–90, 2012.

[74] T Mituyama, K Yamada, E Hattori, H Okida, Y Ono, G Terai, A Yoshizawa, T Komori, and K Asai. The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Res*, 37:D89–D92, 2009.

[75] A Hoffmann, B Montgomery, J Popovici, I Iturbe-Ormaetxe, P Johnson, F Muzzi, M Greenfield, M Durkan, Y S Leong, Y Dong, and et al. Successful establishment of Wolbachia in Aedes populations to suppress dengue transmission. *Nature*, 476: 454–457, 2011.

[76] T Walker, P H Johnson, L A Moreira, I Iturbe-Ormaetxe, F D Frentiu, C J McMeniman, Y S Leong, Y Dong, J Axford, P Kriesner, and et al. The wMel Wolbachia strain blocks dengue and invades caged Aedes aegypti populations. *Nature*, 476:450–453, 2011.

[77] M Woolfit, M Algama, J M Keith, E A McGraw, and J Popovici. Discovery of Putative Small Non-Coding RNAs from the Obligate Intracellular Bacterium Wolbachia pipientis. *PLoS ONE*, 10:doi:10.1371/journal.pone.0118595, 2015.

[78] E Bischoff and C Vaquero. In silico and biological survey of transcription-associated proteins implicated in the transcriptional machinery during the erythrocytic development of Plasmodium falciparum. *BMC Genomics*, 11:34, 2010.

[79] S R Cook, A Gelman, and D B Rubin. Validation of Software for Bayesian Models Using Posterior Quantiles. *Journal of Computational and Graphical Statistics*, 15: 675–692, 2006.

Chapter 2

Literature Review: Part 1

2.1 Introduction

The literature review is organised as follows. It consists of two parts: (1) Part 1 includes; some background to genetics, a review of few statistical methods, an introduction to Bayesian inference and Markov chain Monte Carlo (MCMC) method, and the mathematics of *changept* model; and (2) Part 2 consists of a review of segmentation methods including *changept* modelling, the new encoding method and a few applications of *changept*. Part 2 of the literature review has been published in the Computational and Structural Biotechnology journal [1].

2.2 Some Background to Genetics

2.2.1 Gene

A gene is a molecular unit of heredity of a living organism. Genes hold the information to build and maintain an organisms cells and pass genetic traits to offspring. All organisms have many genes corresponding to various biological traits, some of which are immediately visible, such as eye colour or height, and some of which are not,

such as blood type or increased risk for specific diseases, or the thousands of basic biochemical processes that comprise life. Genes are made from a long molecule called DNA, which is copied and inherited across generations.

2.2.2 DNA and Chromosomes

Deoxyribonucleic acid (DNA) is the carrier of genetic information. The information in DNA is stored using a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The order or sequence of these bases determines the information available for building and maintaining an organism. DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. Together, a base, sugar, and phosphate are called a nucleotide. Nucleotides are arranged in two long strands that form a spiral called a double helix. These two strands run in opposite directions to each other and are therefore said to be anti-parallel. An important property of DNA is that it can replicate, or make copies of itself.

In the nucleus of each cell, the DNA molecule is packaged into thread-like structures called chromosomes. Each chromosome is made up of DNA tightly coiled many times around proteins called histones that support its structure.

2.2.3 RNA

Like DNA, Ribonucleic acid (RNA) is made up of a long chain of components called nucleotides. Each nucleotide consists of a nucleobase, a ribose sugar, and a phosphate group. The sequence of nucleotides allows RNA to encode genetic information. The chemical structure of RNA is very similar to that of DNA, with two differences: (a) RNA contains the sugar ribose, while DNA contains the slightly different sugar deoxyribose, and (b) RNA contains the nucleobase uracil while DNA contains thymine.

Unlike DNA, most RNA molecules are single-stranded and can adopt very complex three-dimensional structures.

2.2.4 UTRs, Introns and Exons

UTRs are the regions of the mRNA molecule that are not translated into a protein (Figure 2.1). There is one on each side of a coding sequence (5′ and 3′). An intron is any nucleotide sequence within a gene that is removed by RNA splicing to generate the final mature RNA product of a gene. The term intron refers to both the DNA sequence within a gene, and the corresponding sequence in RNA transcripts. Sequences that are joined together in the final mature RNA after RNA splicing are called exons.

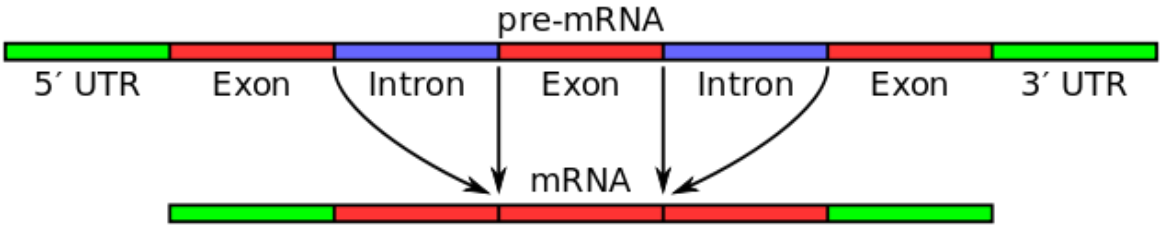


Figure 2.1: *An unspliced mRNA precursor, with UTR's, two introns and three exons (top). The mature mRNA sequence is made after the introns have been removed via splicing (bottom). Source: wikipedia*

2.2.5 The Central Dogma of Molecular Biology

The central dogma of molecular biology was first postulated by Francis Crick about 60 years ago [2]. In its simplest form, it states that genetic information is transferred from DNA to messenger RNA (mRNA, by a process known as transcription) and then to protein (translation) (Figure 2.2).

Figure 2.2: *The Central Dogma of Molecular Biology: DNA makes RNA makes proteins.*

2.2.6 Non-Coding RNA

Non-coding RNAs (ncRNAs) are functional RNA molecules that are transcribed from DNA, but are not translated into proteins (Figure 2.3). ncRNAs have been found to carry out diverse functions including causing a variety of diseases (see Introduction). There are two main groups of ncRNAs: (1) short ncRNAs (ones $\leq 200nt$, such as ribosomal RNAs, transfer RNAs, small nucleolar RNAs, microRNAs, PIWI-interacting RNAs), and (2) long ncRNAs ($>200nt$).

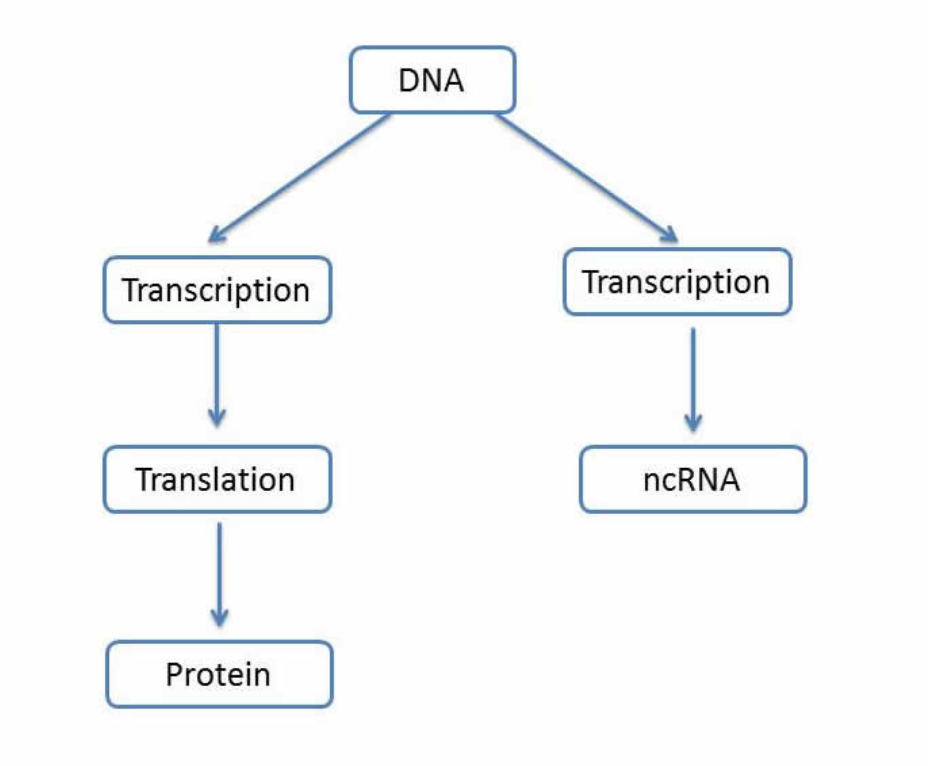


Figure 2.3: *Non-coding RNAs*

2.2.7 Host, Parasite and Pathogen

A *host* is a living organism (eg: human, animal, plant) that nourishes and supports a parasite. The term *parasite* refers to an organism that grows, feeds and is sheltered on or in a different organism while contributing nothing to the survival of its host. For example, in Chapter 6 we segmented three malaria parasites, *Plasmodium falciparum* - the human malaria; *P. reichnowi* - which infects chimpanzees, and *P. gallinaceum* -

which infects jungle fowls to identify genetic mechanisms that make species that are able to thrive on a variety of resources, thus to better understand the malaria disease.

A *pathogen* is defined as a microbe that is able to cause a disease or capable of causing host damage [3, 4]. All parasites that cause a disease in a specific organism are also considered as pathogens for this specific organism. Host-pathogen interactions provide information that can help scientists and researchers understand disease pathogenesis - the biological mechanisms that lead to the diseased state, the biology of one or many pathogens, as well as the biology of the host. For example in Chapter 5, we studied about the pathogenic *Wolbachia* strain *wMelPop*, which was originally identified during a survey of lab lines of *Drosophila melanogaster* for genetic mutations causing brain degeneration [5]. This strain over replicates in host cells, causing cellular damage and reducing lifespan by approximately one-half in flies [6, 7] and causes similar host effects when transinfected into the mosquito *Aedes aegypti* [8]. The life-shortening effect of *wMelPop* is being utilized as part of a novel biocontrol strategy to reduce dengue virus transmission by *A. aegypti* [9–11]. In Chapter 5, our goal was to identify candidate sRNAs that may play significant roles in its interactions with its host and this work was carried out as a part of eradicate dengue project.

2.3 Statistical Methods

Here I present some statistical methods used in up coming chapters.

2.3.1 Z-test to compare proportions

To compare two population proportions p_1 and p_2, the z-test is appropriate given the following conditions are met (used in Chapter 4 to compare the proportion of transcription factors in PFE genes to proportion of transcription factors in the alignment).

- The samples are independent.

- Each sample is large enough to justify using normal approximations to the distributions of the sample proportions. It is usually sufficient if each sample has a minimum of 10 successes and 10 failures.

- All individuals in the population are equally likely to be sampled.

Let $\hat{p_1}$ and $\hat{p_2}$ be the observed proportions of two populations, where $\hat{p_1} = \frac{x_1}{n_1}$ and $\hat{p_2} = \frac{x_2}{n_2}$. Here n_1 and n_2 are the sizes of each sample, and x_1 and x_2 are the number of successes in each sample respectively.

Steps in hypothesis test

1. State null (H_0) and the alternative (H_a) hypotheses. Three sets of statistical hypotheses can be formulated;
   a. $H_0 : p_1 - p_2 = 0$ versus $H_a : p_1 - p_2 \neq 0$; a two-tailed test.
   b. $H_0 : p_1 - p_2 \leq 0$ versus $H_a : p_1 - p_2 > 0$; an upper-tailed test.
   c. $H_0 : p_1 - p_2 \geq 0$ versus $H_a : p_1 - p_2 < 0$; a lower-tailed test.

2. Summarize data into a suitable test statistic. Here the test statistic is, $\frac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}} \sim N(0, 1)$ where p is the observed proportion from the combined samples, calculated by $p = \frac{x_1 + x_2}{n_1 + n_2}$.

3. Assuming the null to be true, find the $p - value$. P-value is "the probability that the observed statistic value (or more extreme value) could occur if the null model was correct".

4. Decide if the result is statistically significant based on the $p - value$ and report conclusion. If p-value $\geq$ level of significance (often 0.05), do not reject H_0; otherwise reject H_0.

2.3.2 Mann-Whitney U test

In hypothesis testing, parametric tests are used when confident the assumptions of the test are satisfied, otherwise non-parametric test are used, especially for small sample sizes. The Mann-Whitney U test is a non-parametric test which allows non-normally distributed and ordinal data sets to be compared (used in Chapter 5 to test significant differences in candidate sRNA expression). It tests the null hypothesis that two samples come from the same population against an alternative hypothesis, given below conditions are met.

- All observations from both groups are independent of each other.
- The responses are ordinal.

Steps in calculation

1. Assign numeric ranks to all the observations, beginning with 1 for the smallest value. Where there are groups of tied values, assign a rank equal to the midpoint of unadjusted rankings, ie the ranks of (3, 5, 5, 9) are (1, 2.5, 2.5, 4).
2. Add up the ranks for the observations which came from sample 1. The sum of ranks in sample 2 is determinate, since the sum of all the ranks equals $N(N+1)/2$, where N is the total number of observations.
3. U is then given by: $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$, where n_1 is the sample size for sample 1, and R_1 is the sum of the ranks in sample 1. Note that there is no specification as to which sample is considered sample 1. An equally valid formula for U is $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$ using sample 2.
4. Calculate $U = min(U_1, U_2)$.
5. Statistical tables can be used for the Mann-Whitney U test to find the probability of observing a value of U or lower. This is the $p - value$.
6. Decide if the result is statistically significant based on the $p - value$ and report conclusion.

NOTE: If the number of observations is such that n_1, n_2 are large enough (> 20), a normal approximation can be used with mean, $\mu_U = \frac{n_1 n_2}{2}$ and standard deviation, $\sigma_U = \sqrt{\frac{n_1 n_2 (N+1))}{12}}$, where $N = n_1 + n_2$.

2.3.3 LOESS Model - LOcally WEighted Scatter-plot Smoother

There are two general strategies for fitting a smooth curve: parametric and nonparametric fitting [12]. Parametric fitting requires the analyst to specify the functional form of the relationship in advance (eg: least squares regression). Often the correct functional form is unknown. Currently, the most popular nonparametric smoother is loess [13]. Thus the biggest advantage of loess is, it does not require the specification of a function to fit a model to all of the data in the sample. Loess provides a graphical summary of the relationship between a dependent variable and one or more independent variables.

The following parameters must be supplied prior to the loess model fitting procedure in order to guarantee that the loess curve really does pass through the center of the empirical data points [14].

The smoothing parameter, α

Parameter α (a value between 0 and 1) gives the proportion of observations that is to be used in each local regression (explained below). The fitted curve becomes smoother with larger values of α. However a decision about a proper value of α must be made on a case-by-case basis to avoid "over-fitting" or "lack of fitting" the model.

The degree of the loess polynomial, λ.

The λ parameter specifies the degree of the polynomial that the loess procedure fits to the data. If $\lambda = 1$, then linear equations are fit and when $\lambda = 2$, quadratic equations are used.

Fitting a loess smooth curve

1. Assume that the data consist of n observations on two variables, X and Y. These data are displayed in a bivariate scatter plot. The plotted points are the ordered pairs (x_i, y_i), where i ranges from 1 to n.

2. Select a series of m locations or evaluation points, v_j, with j running from 1 to m. These evaluation points are equally-spaced across the range of X.

3. Loess performs a series of m weighted regression analyses using a subset of observations, one at each of the v_j. These regressions are "local" in the sense that each one only uses the subset of observations that fall closest to that evaluation point along the horizontal axis of the scatter plot.

4. Specify the proportion of the total data that is included within each subset using a loess parameter, α.

5. Specify the functional form (either linear or quadratic) using the loess λ parameter. The observations included in each local regression are inversely weighted according to their distance from the evaluation point along the X axis.

6. The coefficients from each local regression are used to estimate a predicted or fitted value, designated $\hat{g}(v_j)$ for that evaluation point.

7. After all of the local regressions are completed, plot m different ordered pairs, $(v_j, \hat{g}(v_j))$ in the scatterplot, superimposed over the n data points that are already shown in the plot.

8. Finally, connect adjacent fitted points that is, the $(v_j, \hat{g}(v_j))$ for successive v_js by line segments (loess line).

2.4 Introduction to Bayesian Inference

The most frequently used statistical methods assume that unknown parameters are fixed constants, and define probability in terms of limiting relative frequencies. It follows from these assumptions that probabilities are objective and that we cannot make probabilistic statements about parameters because they are fixed. Bayesian methods offer an alternative approach; they treat parameters as random variables and define probability as 'degrees of belief' (that is, the probability of an event is the degree to which we believe the event is true). Bayesian methods provide a natural and principled way of combining prior information with data, and can incorporate past information about a parameter to form a prior distribution for use in future analysis.

Suppose we are interested in estimating a parameter θ from data $y = (y_1, ..., y_n)$ by using a statistical model described by a density $p(y|\theta)$. The Bayesian approach assumes that θ cannot be determined exactly and uncertainty about the parameter is expressed through probability statements and distributions. The following steps describe the essential elements of Bayesian inference:

1 A probability distribution for θ is formulated as $\pi(\theta)$, which is known as the prior distribution. The prior distribution expresses beliefs (for example, on the mean, the spread, the skewness) about the parameter before we examine the data.

2 Given the observed data y, choose a statistical model $p(y|\theta)$ (likelihood model) to describe the distribution of y given θ.

3 Update beliefs about θ by combining information from the prior distribution and the data through the calculation of the posterior distribution $p(\theta|y)$.

The third step is carried out by using Bayes′ Theorem, which enables to combine the prior distribution and the model in the following way:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)\pi(\theta)}{p(y)} = \frac{p(y|\theta)\pi(\theta)}{\int p(y|\theta)\pi(\theta)d\theta} \tag{2.4.1}$$

The quantity $\int p(y|\theta)\pi(\theta)d\theta$ is the normalizing constant of the posterior distribution. $p(y)$ is the marginal distribution of y (or marginal distribution of the data). The likelihood function of θ is any function proportional to $p(y|\theta)$.

Simply, Bayes′ Theorem tells how to update existing knowledge with new information. We begin with a prior belief $\pi(\theta)$, and after learning information from data y, we change or update our belief about θ and obtain $p(\theta|y)$.

In recent literature [15], the process of Bayesian data analysis is explained by dividing it into following three steps.

1. Set up a full probability model - a joint probability distribution for all observable and unobservable quantities in a problem $p(\theta, y)$. The model should be consistent with knowledge about the underlying scientific problem and the data collection process.

2. Condition on observed data - calculate and interpret the appropriate posterior distribution $p(\theta|y)$.

3. Evaluate the fit of the model and the implications of the resulting posterior distribution - how well does the model fit the data, are the substantive conclusions reasonable, and how sensitive are the results to the modelling assumptions in step 1? In response, one can alter or expand the model and repeat the three steps.

2.5 A Markov Chain

A Markov chain is a random process with the property that, conditional on its present value, the future is independent of the past. A typical random process Θ is a family

$\{\Theta_t : t \in T\}$ of random variables indexed by some set T. If $T = \{0, 1, 2, ...\}$, we call the process a *discrete-time* process; if $T = \mathbb{R}$ or $T = [0, \infty)$, we call it a *continuous-time* process [16].

State space: S: A state space (S) is the set of values which a process can take or the range of possible values for the random variables X. S is said to be countable if the elements of S can be put into a one-to-one correspondence with some subset of the integers. In other words, one can index all possible states using the integers or some subset thereof.

Discrete-time Markov chain: Let $\Theta_0, \Theta_1, \Theta_2, ...$ be a sequence of random variables which takes values in some countable space S, called the state space. Each Θ_n is a discrete random variable that takes one of N possible values, where $N = |S|$; it may be the case that $N = \infty$. The process Θ is a discrete-time first-order Markov chain if it satisfies the following condition.

$$p(\Theta_n = s|\Theta_0 = \theta_0, \Theta_1 = \theta_1, ..., \Theta_{n-1} = \theta_{n-1}) = p(\Theta_n = s|\Theta_{n-1} = \theta_{n-1}) \qquad (2.5.1)$$

for all $n \geq 1$ and all $s, \theta_0, \theta_1, ..., \theta_{n-1} \in S$.

A Markov chain is **irreducible** if it is possible to go from any state to any other state (not necessarily in one step). That is, all states communicate with each other. The chain is said to be **aperiodic** when the number of steps required to move between two states is not required to be a multiple of some integer, > 1. That is, the chain is not forced into some cycle of fixed length between certain states. Finally, a Markov chain is **recurrent**, if for any given state i, if the chain starts at i, it will eventually return to i with probability 1. It is said to be **positive recurrent** if the expected return time to state i is finite; otherwise it is *null recurrent*.

2.5.1 Stationary Distribution

A distribution π on target space S is stationary with respect to a transition matrix P, if $\pi\mathbf{P} = \pi$.

The *Transition matrix* $\mathbf{P} = (p_{ij})$ is the $|S| \times |S|$ matrix of *transition probabilities*:

$$p_{ij} = p(\Theta_{n+1} = j | \Theta_n = i) \tag{2.5.2}$$

If we can devise a Markov chain whose stationary distribution π is the desired posterior distribution $p(\theta/y)$, then we can run this chain to get draws that are approximately from $p(\theta/y)$ once the chain has converged. A sufficient condition for a unique stationary distribution is that the detailed balance equation holds (for all i and j),

$$P(i,j)\pi_i = P(j,i)\pi_j$$

2.5.2 Ergodic Theorem

Let $\theta_0, \theta_1, ..., \theta_M$ be M values from a Markov chain that is aperiodic, irreducible, and positive recurrent, and $E[g(\theta)] < \infty$, where $g(\theta)$ is some function of θ.

Then with $p\Big(\frac{1}{M}\sum_{i=1}^{M} g(\theta_i), M \to \infty\Big) = 1$.

$$\frac{1}{M}\sum_{i=1}^{M} g(\theta_i) \simeq \int_{\Theta} g(\theta)\pi(\theta)d\theta \tag{2.5.3}$$

as $M \to \infty$, where π is the stationary distribution.

2.5.3 Monte Carlo Method

Monte Carlo is a method developed by physicists to use random number generation to compute integrals. Suppose we wish to compute a complex integral:

$$\int_a^b h(\theta)d\theta$$

If we can decompose $h(\theta)$ into the product of a function $f(\theta)$ and a probability density function $p(\theta)$ defined over the interval (a, b), then the integral $\int_a^b h(\theta)d\theta$ can be expressed as an expectation of $f(\theta)$ over the density $p(\theta)$ as below.

$$\int_a^b h(\theta)d\theta = \int_a^b f(\theta)p(\theta)d\theta = E_{p(\theta)}[f(\theta)] \tag{2.5.4}$$

Thus if we draw a large number of random variables, $\theta_1, ..., \theta_n$ from the density $p(\theta)$, then we can write;

$$\int_a^b h(\theta)d\theta = E_{p(\theta)}[f(\theta)] \simeq \frac{1}{n}\sum_{i=1}^{n} f(\theta_i) \tag{2.5.5}$$

This is referred to as **Monte Carlo integration.**

2.5.4 Markov Chain Monte Carlo Method

A major limitation of Bayesian approaches is that obtaining the posterior distribution often requires the integration of high-dimensional functions. This can be computationally very difficult. If we have a Markov chain that has converged to the stationary distribution, the draws in our chain appear to be like draws from the posterior distribution, $f(\theta|y)$. The ergodic theorem allows us to perform Monte Carlo Integration to find quantities of interests ignoring the dependence between draws.

The Markov chain Monte Carlo (MCMC) approach is to construct a Markov chain having the following two properties [16]

(a) The chain has π as the unique stationary distribution.
(b) The transition probabilities of the chain have a simple form.

Property (a) ensures the distribution approaches the required distribution (the posterior distribution in Bayesian statistics) and property (b) ensures the easy simulation of the chain. Therefore MCMC sampling is frequently used in Bayesian inference to simulate sampling of a posterior distribution and compute posterior quantities of interest.

Consider a first-order Markov chain with a sequence of random variables $\theta^1, \theta^2, ...$ in a target space S, for which the random variable θ^t depends on all previous θs only through its immediate predecessor θ^{t-1},

$$p(\theta^t|\theta^1, \theta^2, .., \theta^{t-1}) = p(\theta^t|\theta^{t-1}), \tag{2.5.6}$$

Monte Carlo integration can be used to approximate posterior (or marginal posterior) distribution required for a Bayesian analysis [17]. Thus the integral $I(y) = \int_a^b f(y|\theta)p(\theta)d\theta$ can be approximated by:

$$I'(y) = \frac{1}{n}\sum_{i=1}^{n} f(y|\theta_i) \tag{2.5.7}$$

where the θ_i are drawn from density $p(\theta)$.

In Bayesian statistics, there are two general MCMC algorithms that are commonly used: (1) the Metropolis-Hastings algorithm; and (2) the Gibbs sampler.

Metropolis-Hastings Algorithm

The Metropolis algorithm is named after its inventor, the American physicist and computer scientist Nicholas C. Metropolis. The algorithm is simple but practical,

and it can be used to obtain random samples from an arbitrarily complicated target distribution of any dimension, where the normalizing constant may not be known.

Suppose our goal is to draw samples from some distribution $p(\theta|y)$. The Metropolis algorithm [18, 19] generates a sequence of draws from this distribution as follows:

(1). Start with any initial value θ_0 satisfying $p(\theta_0|y) > 0$.

(2). Using current θ_0 value, sample a candidate point θ^* from a jumping distribution $q(\theta_1, \theta_2)$, which is the probability of returning a value of θ_2 given a previous value of θ_1. This distribution is also referred to as the proposal distribution. The only restriction on the jump density in the Metropolis algorithm is that it is symmetric, i.e., $q(\theta_1, \theta_2) = q(\theta_2, \theta_1)$.

(3). Given the candidate point θ^*, calculate an acceptance ratio of the density (α) at the candidate (θ^*) and current (θ_{t-1}) points,

$$\alpha = min\left\{\frac{p(\theta^*|y)}{p(\theta_{t-1}|y)}, 1\right\} \tag{2.5.8}$$

(4). Accept candidate point θ^* as θ_t with probability α. If θ^* is not accepted, $\theta_t = \theta_{t-1}$.

(5). Repeat steps 2-4 (M times).

This generates a Markov chain $(\theta_0, \theta_1, ..., \theta_k, ..., \theta_M)$, as the transition probabilities from θ_t to θ_{t+1} depend only on θ_t and not $(\theta_0, ..., \theta_{t-1})$. Following a sufficient *burn-in* period (of say, k steps), the chain approaches its stationary distribution and samples $\theta_{k+1}, ..., \theta_M$ are samples from $p(\theta|y)$.

Hastings (1970) generalized the Metropolis algorithm using an asymmetric proposal distribution, $q(\theta^*, \theta_t) \neq q(\theta_t, \theta^*)$. The difference in its implementation comes in calculating the ratio of densities:

$$\alpha = min\left\{\frac{p(\theta^*|y)q(\theta^*, \theta_{t-1})}{p(\theta_{t-1}|y)q(\theta_{t-1}, \theta^*)}, 1\right\} \tag{2.5.9}$$

Other steps remain the same.

The Gibbs Sampler

The Gibbs sampler, named by Geman and Geman (1984) after the American physicist Josiah W. Gibbs, is a special case of the Metropolis and Metropolis-Hastings Algorithms in which the proposal distributions exactly match the posterior conditional distributions and proposals are accepted 100% of the time. Gibbs sampling requires us to decompose the joint posterior distribution into full conditional distributions for each parameter in the model and then sample from them. The sampler can be efficient when the parameters are not highly dependent on each other and the full conditional distributions are easy to sample from.

Suppose $\theta = (\theta_1, \theta_2, ..., \theta_k)$ is the parameter vector, $p(y|\theta)$ is the likelihood, and $\pi(\theta)$ is the prior distribution. The full posterior conditional distribution of $\pi(\theta_i|\theta_j; i \neq j, y)$ is proportional to the joint posterior density. That is,

$$\pi(\theta_i|\theta_j; i \neq j, y) \propto p(y|\theta)\pi(\theta)$$

For instance, the one-dimensional conditional distribution of θ_1 given $\theta_j = \theta_{jnew}, 2 < j < k$, is computed as the following:

$$\pi(\theta_1|\theta_j = \theta_{jnew}, 2 \leq j \leq k, y) \propto p(y|\theta = (\theta_1, \theta_{2new}, .., \theta_{knew}))\pi(\theta = (\theta_1, \theta_{2new}, ..., \theta_{knew}))$$

The Gibbs sampler works as follows:

(1) Set, $t = 0$ and choose an arbitrary initial value of $\theta^0 = (\theta_1^0, ..., \theta_k^0)$

(2) Generate each component as follows:

• draw θ_1^{t+1} from $\pi(\theta_1|\theta_2^t, \theta_3^t, ..., \theta_k^t, y)$

• draw θ_2^{t+1} from $\pi(\theta_2|\theta_1^{t+1}, \theta_3^t, ..., \theta_k^t, y)$

•

•

• draw θ_k^{t+1} from $\pi(\theta_k|\theta_1^{t+1}, \theta_2^{t+1}, ..., \theta_{k-1}^{t+1}, y)$

(3) Set $t = t+1$. If $t < M$, the number of desired samples, return to step 2. Otherwise, stop.

The result is a Markov chain with M draws of θ that are approximately from the posterior, $p(\theta|y)$. We can perform Monte Carlo integration on those draws to obtain quantities of interest.

Generalized Gibbs Sampler

Here I summarise the Generalized Gibbs Sampler (GGS) presented in [20].

The conventional Gibbs sampler is used to sample from a distribution $p(x)$ over a space X in which points have fixed dimension d. Each iteration of the Gibbs sampler involves d coordinate updates in which new values for each of the d coordinates are drawn from the one-dimensional conditional distributions of p with the other coordinates fixed. On the other-hand, GGS can be used when points in X do not have fixed dimension, and may not even have a representation in terms of coordinates. It is formalized by augmenting a space I of move-types to X and defining a set $\mathcal{Q}(x) \subset I \times \{x\}$ to be the set of move-types available at x for each $x \in X$. The move types are analogous to the coordinate updates of the conventional Gibbs sampler.

Let $U \equiv \cup_{x \in X} \mathcal{Q}(x)$. The GGS generates a Markov chain in U such that the projection of the chain onto X has the limiting distribution f. The GGS makes use of two transition matrices Q and R. The first of these, Q is used to select a move type from $\mathcal{Q}(x)$, where x is the most recently sampled element of X. The matrix Q assigns probability zero to a transition between any element of $\mathcal{Q}(x)$ and any element of $\mathcal{Q}(y)$, for $y \neq x$. This ensures that the selected move type is one that is available at x. The matrix Q is otherwise arbitrary. Selecting a move type using Q is analogous to selecting a coordinate to update in the conventional Gibbs sampler.

The second transition matrix, R, selects an element of the set $\mathcal{R}(u)$, where $u = (i, x)$ is the element of $\mathcal{Q}(x)$ selected using Q, and $\mathcal{R}(u)$ is a subset of U containing u. The sets $\{\mathcal{R}(u) : u \in U\}$ must form a partition of U, but are otherwise arbitrary. The transition matrix R is given by:

$$R(u, v) = \begin{cases} \dfrac{f(y)q_y(v)}{\displaystyle\sum_{w \in \mathcal{R}(u)} f(z)q_z(w)} & \text{for } v \in \mathcal{R}(u) \\ \\ 0 & \text{otherwise} \end{cases}$$

where $u = (i, x)$, $v = (j, y)$, $w = (k, z)$, and q_x is a distribution on $\mathcal{Q}(x)$ that is stationary with respect to the transition matrix Q.

GGS algorithm

Starting with an arbitrary U_0, perform the following steps iteratively:

(1) [**Q-step**]: Given $U_n = (i, x)$, generate $V \in \mathcal{Q}(x)$ by drawing from the distribution with density $Q((i, x), .)$.

(2) [**R-step**]: Given $V = (j, y)$, generate $W \in \mathcal{R}(j, y)$ by drawing from the distribution with density $R((j, y), .)$.

(3) Let $U_{n+1} = W$.

This algorithm generates a Markov chain $\{U_0, U_1, ...\} = \{(I_0, X_0), (I_1, X_1), ...\}$ such that the limiting distribution of X_n as $n \to \infty$ is f, provided that P is irreducible and aperiodic.

Consider a Markov chain $\{U_1, U_2, ...\}$ on U with transition matrix P defined by:

$$P = QR$$

For $u = (i, x)$ and $v = (j, y)$,

$$P(u, v) = \sum_{w \in U} Q(u, w)R(w, v) = \sum_{w \in \mathcal{Q}(x) \cap \mathcal{R}(v)} Q(u, w)R(w, v)$$

Let μ be the distribution on U defined by $\mu(i, x) = f(x)q_x(i, x)$.

Note that μ is stationary with respect to R

Consider

$$\mu(i, x)R((i, x), (j, y)) = \begin{cases} \dfrac{f(x)q_x(i, x)f(y)q_y(j, y)}{\displaystyle\sum_{(k,z) \in \mathcal{R}(i,x)} f(z)q_z(k, z)} & \text{for } (j, y) \in \mathcal{R}(i, x) \\ 0 & \text{otherwise} \end{cases}$$

Now consider

$$\mu(j, y)R((j, y), (i, x)) = \begin{cases} \dfrac{f(y)q_y(j, y)f(x)q_x(i, x)}{\displaystyle\sum_{(k,z) \in \mathcal{R}(j,y)} f(z)q_z(k, z)} & \text{for } (j, y) \in \mathcal{R}(i, x) \\ 0 & \text{otherwise} \end{cases}$$

We know that

$$(j, y) \in \mathcal{R}(i, x), (k, z) \in \mathcal{R}(j, y) \Longrightarrow (k, z) \in \mathcal{R}(i, x)$$

Therefore we can write

$$\mu(i, x)R((i, x), (j, y)) = \mu(j, y)R((j, y), (i, x))$$

μ is stationary with respect to R and consequently

$$\sum_{(i,x)\in U} \mu(i,x)R((i,x),(j,y)) = \mu(j,y)$$

Note that μ is stationary with Q

$\sum_{(i,x)\in U} \mu(i,x)Q((i,x),(j,y))$ is the probability at (j,y) after applying transition matrix Q to the distribution $\mu(i,x)$.

$Q((i,x),(j,y)) = 0$ unless $(i,x) \in \mathcal{Q}(y)$ and if $(i,x) \in \mathcal{Q}(y)$ then $x = y$.

Therefore we have

$$\begin{aligned}\sum_{(i,x)\in U} \mu(i,x)Q((i,x),(j,y)) &= \sum_{(i,y)\in \mathcal{Q}(y)} \mu(i,y)Q((i,y),(j,y)) \\ &= \sum_{(i,y)\in \mathcal{Q}(y)} f(y)q_y(i,y)Q((i,y),(j,y)) \\ &= f(y) \sum_{(i,y)\in \mathcal{Q}(y)} q_y(i,y)Q((i,y),(j,y))\end{aligned}$$

As q_y is stationary with respect to $\mathcal{Q}(y)$

$$\sum_{(i,x)\in U} \mu(i,x)Q((i,x),(j,y)) = f(y)q_y(j,y) = \mu(j,y)$$

Therefore μ is stationary with respect to Q.

Note that the distribution μ is stationary with respect to P

μ is stationary with respect to Q and R. Then in matrix notation

$$\mu P = \mu QR = \mu R = \mu$$

Therefore μ is stationary with respect to P.

If P is irreducible and aperiodic, μ is the limiting distribution of the process P .

The GGS is summarized in the following algorithm.

2.5.5 Assessing MCMC Convergence

There are several methods for assessing the convergence of MCMC chains. The most simple way to see if the chain has converged is by visual inspection using a traceplot.

Visual analysis via trace plots

A traceplot is a plot of the iteration number against the value of a sampled parameter at each iteration (Figure 2.4). The trace indicates if the chain has not yet converged to its stationary distribution. A trace can also indicate whether the chain is mixing well. The aspects of stationarity that are most recognizable from a trace plot are a relatively constant mean and variance.

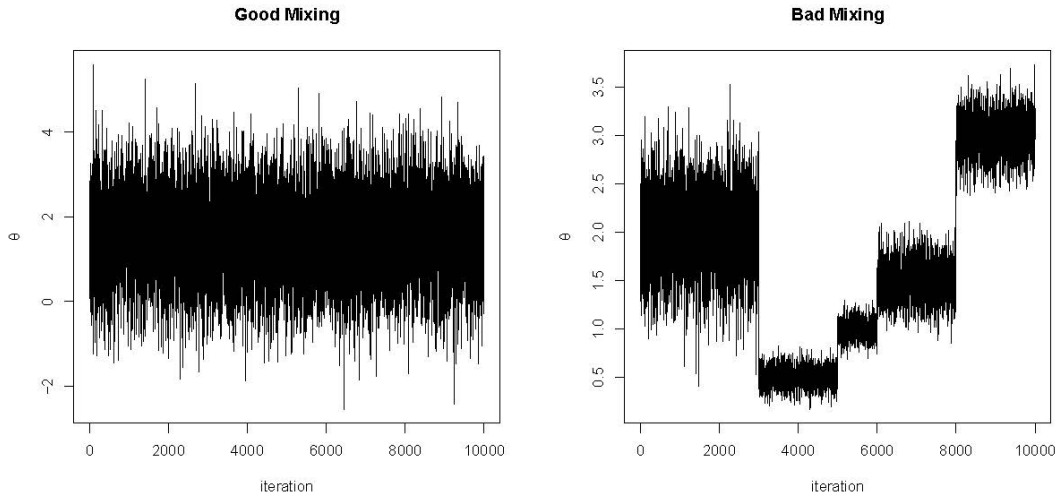


Figure 2.4: *Examples of traceplots for parameter θ: The trace in the right seems to have failed to converge. One can consider reparameterizing the model or run the Markov chain for a long time. The chain in the left can be considered as converged with mean around value of 2 and small fluctuations.*

A number of more formal methods are prevalent in the literature (reviewed in [21, 22]). Most of these diagnostics are implemented in the package CODA, which is a popular

program for convergence diagnostics written for R [23, 24]. In this thesis, we used the method developed by Heidelberger-Welch to assess the convergence [25].

Heidelberger and Welch Diagnostics

This test consists of two parts: a stationary portion test and a half-width test [25, 26]. The convergence test uses the Cramer-von-Mises statistic to test the null hypothesis that the sampled values come from a stationary distribution. The test is successively applied, firstly to the whole chain, then after discarding the first 10%, 20%, ... of the chain until either the null hypothesis is accepted, or 50% of the chain has been discarded. The latter outcome constitutes 'failure' of the stationarity test and indicates that a longer MCMC run is needed. If the stationarity test is passed, the number of iterations to keep and the number to discard are reported.

The part of the chain that is deemed stationary is put through a half-width test, which checks whether the Markov chain sample size is adequate to estimate the mean values accurately. The half-width test calculates a 95% confidence interval for the mean, using the portion of the chain which passed the stationarity test. If the half-width is less than ϵ times the sample mean (where ϵ is a small fraction), the half-width test is passed and the retained sample is deemed to estimate the posterior mean with acceptable precision. If the half-width test is failed, this implies that a longer run is needed to increase the accuracy of the posterior estimates for the given variable. The CODA default for ϵ is 0.1.

2.6 The changept model

Changept model can mainly be used to segment either a pairwise or a multiple alignment using two types of encoding methods: (1) binary encoding - enables segmentation based on a single property of interest (eg: degree of conservation); (2) letter encoding - enables segmentation based on multiple properties of interest (eg: conservation levels, GC content, transition/transversion ratio). Here I present the mathematics of each

type of *changept* modelling in detail [27, 28]. A review of segmentation methods including *changept* analysis is presented in Chapter 2- part 2 [1].

2.6.1 Changept modelling for segmenting a binary sequence

Generating the binary sequence

The *changept* model currently does not consider indels (insertions and deletions) of an alignment. Thus the binary sequence is generated by first stripping the columns containing an indel. Suppose we are interested in segmenting a pairwise alignment based on the degree of conservation between two species. Then the alignment is converted into a binary sequence by replacing alignment coloumns in which genomes match with a '1′ and mismatch with a '0′ (Table 2.1). The boundaries between alignment blocks are marked using a '#′ character and these are considered as fixed change-points.

Table 2.1: *Generating binary sequence of a pairwise alignment based on conservation levels*

Species 1	A	C	G	G	A	C	G	T
Species 2	A	A	C	G	G	G	T	T
Symbol	1	0	0	1	0	0	0	1

Binary sequence: 10010001

Modelling

Suppose that the length L of a binary sequence and the positions of fixed change-points are given. Then for each position in the sequence except the first and those immediately following the fixed change-points, a decision is made as to whether to start a new segment at that position (at 1^{st} position and the positions after fixed change-points, a new segment has to start, therefore no decision to make). The probability of starting a new segment is denoted by ϕ.

Thus the probability of generating a new segmentation with total k change-points including k' fixed change-points at positions $c = (c_1, c_2, ..., c_k)$ is given by the joint probability of k and c conditional on ϕ:

$$p(k, c|\phi) = \phi^{k-k'}(1-\phi)^{(L-1-k)} \tag{2.6.1}$$

Here $k+1$ segments are numbered from 0 to k and for convenience $c_0 = 1$ and $c_k = L + 1$.

Each segment is assigned to one of v conservation classes. The probability of assigning any given segment to a class v is denoted by π_v and $\pi = (\pi_0, \pi_1, ..., \pi_{v-1})$. The class to which segment i is assigned is denoted by $g_i \in (0, 1, .., v-1)$; let $g = (g_0, g_1, ..., g_k)$.

The probability of a specific assignment of the $k+1$ segments such that x_0 segments are assigned to class $0, ..., x_{(v-1)}$ segments are assigned to class $v-1$ is given by:

$$p(g|k, \pi) = \pi_0^{x_0}\pi_1^{x_1}...\pi_{v-1}^{x_{v-1}} = \prod_{i=0}^{k} \pi_{gi} \tag{2.6.2}$$

Each segment is then assigned a Bernoulli parameter representing the probability of generating a '1′ at each position in that segment. For each segment i in class t, the Bernoulli parameter θ_i is drawn from a beta distribution with as yet unspecified parameters $\alpha_0^{(t)}$ and $\alpha_1^{(t)}$. (θ_i is the probability of generating a '1′ at each position of segment i in class t).

$$B(\theta_i|\alpha_0^{(t)}, \alpha_1^{(t)}) = \frac{\Gamma(\alpha_0^{(t)} + \alpha_1^{(t)})}{\Gamma(\alpha_0^{(t)})\Gamma(\alpha_1^{(t)})}\theta_i^{\alpha_1^{(t)}-1}(1-\theta_i)^{\alpha_0^{(t)}-1} \tag{2.6.3}$$

Here $\theta = (\theta_0, \theta_1, ..., \theta_k), \alpha^{(t)} = (\alpha_0^{(t)}, \alpha_1^{(t)})$ and $\alpha = (\alpha^{(0)}, .., \alpha^{(v-1)})$

Finally, the binary sequence within each segment is generated by independent Bernoulli trials at each position in that segment. Thus the probability that segment i contains a

specific sequence S_i including m_i zeros and n_i ones is:

$$p(S_i|L_i, \theta_i) = \theta_i^{n_i}(1-\theta_i)^{m_i} \tag{2.6.4}$$

Let $L_i = c_{i+1} - c_i$ be the length of segment i. Let S be the final binary sequence obtained by concatenating $S_0.....S_k$. Thus the joint distribution of k, c, g, θ, S is given by:

$$p(k, c, g, \theta, S|\phi, \pi, \alpha) = p(k, c|\phi)p(g|k, \pi) \times \prod_{i=0}^{k} B(\theta_i|\alpha^{(g_i)})p(S_i|L_i, \theta_i) \tag{2.6.5}$$

To complete the Bayesian model, the prior probabilities of unspecified parameters $\phi, \pi, \alpha_0^{(t)}$ and $\alpha_1^{(t)}$ are assigned as follows.

For ϕ and π uniform prior densities $p(\phi) = 1$ and $p(\pi) = 1$ are used on the interval $[0, 1]$. For $\alpha^{(t)}$, the uniform priors on mean μ and standard deviation of the beta distribution, given by $\mu_j^{(t)} = \alpha_j^{(t)}/(\alpha_0^{(t)} + \alpha_1^{(t)})$ and $\sigma^t = \sqrt{\mu_0^{(t)}\mu_1^{(t)}/(\alpha_0^{(t)} + \alpha_1^{(t)} + 1)}$ for $j = 0, 1$ are used.

Now using Bayes rule, integrating over ϕ and θ and summing over g the following posterior distribution is obtained:

$$p(k, c, \pi, \mu, \sigma|S) \propto \Gamma(k - k' + 1)\Gamma(L - k) \times \prod_{i=0}^{k} f(m_i, n_i|\pi, \alpha) \tag{2.6.6}$$

where $\mu = (\mu^{(0)}, \mu^{(1)}, ..., \mu^{(v-1)})$ and $\sigma = (\sigma^{(0)}, \sigma^{(1)}, .., \sigma^{(v-1)})$, α is a function of μ and σ and:

$$f(m, n|\pi, \alpha) = \sum_t \left[\pi_t \frac{\Gamma(\alpha_0^{(t)} + \alpha_1^{(t)})}{\Gamma(\alpha_0^{(t)})\Gamma(\alpha_1^{(t)})} \times \frac{\Gamma(m + \alpha_0^{(t)})\Gamma(n + \alpha_1^{(t)})}{\Gamma(m + \alpha_0^{(t)} + n + \alpha_1^{(t)})} \right] \tag{2.6.7}$$

Figure 2.5 shows the parameters of the model and their conditional dependencies. A parameter at the head of the arrow is conditionally dependent on the parameter at the tail.

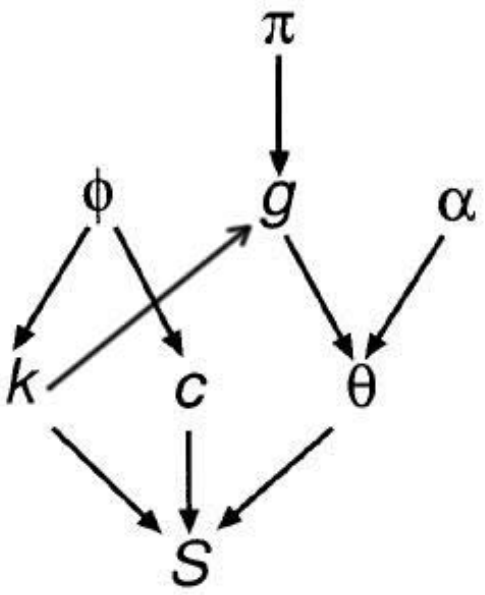


Figure 2.5: *The parameters and their conditional dependencies of the model: source [28].*

Sampling

In order to estimate the parameters k, c, π and α, a sample from the posterior distribution in equation 2.6.6 is drawn using the Generalised Gibbs sampler (GGS) [20]. The sampler involves updating parts of the current element of a Markov chain while holding other parts fixed, in a manner resembling the conventional Gibbs sampler. Unlike the conventional Gibbs sampler, the GGS can sample from spaces in which the dimension varies from point to point.

The GGS algorithm is separated in different steps using the move-types defined below.

- (I, i): decide whether to insert a new change-point in segment i, and at what position.
- (D, i): decide whether to delete a new change-point i, if it is not a fixed and permanent change-point.

- (S, i): slide change-point i, to a new position between c_{i-1} and c_{i+1} if it is not a fixed and permanent change-point.
- (π_{t1}, π_{t2}): simultaneously update π_{t1}, π_{t2} for $(t_1, t_2) \in \{0, ..., v-1\}^2$ keeping their sum constant.
- (t_1, t_2): simultaneously update π_{t1}, π_{t2} and α^{t1}, α^{t2} for $(t_1, t_2) \in \{0, ..., v-1\}^2$.
- π_t: update π_t, scaling all other π values by a constant factor.
- σ^t: update $\sqrt{1/(z^t + 1)}$ while holding μ^t constant.
- μ^t: update μ^t while holding σ^t constant.

The number of moves available from any given segmentation depends on k and it is given by:

$$N(k) = 3k + 1 + v(v-1) + 3v$$

where v is the number of conservation classes. Here the value of $N(k)$ was obtained by adding up the following moves.

- (I, i): $k + 1$ moves
- (D, i): k moves
- (S, i): k moves
- π_t, σ^t, μ^t : $3v$ moves - 1 move for each parameter in each group
- (π_{t1}, π_{t2}) : $v(v-1)/2$ moves
- (t_1, t_2) : $v(v-1)/2$ moves

In each of the first 3 moves, there is a possibility no change is made, in which case the current segmentation is repeated. The sampler cycles through the available moves in a systematic manner, illustrated in Figure 2.6 below.

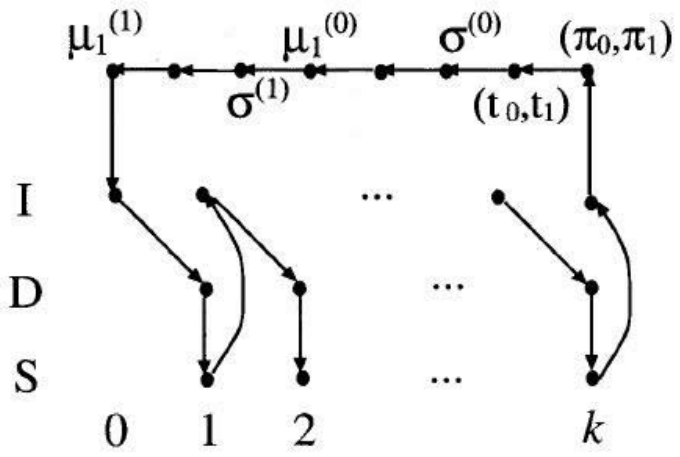


Figure 2.6: *The order in which the updates are carried out. Note that I updates run from* 0 *to k whereas D and S updates run from* 1 *to k. The updates shown here are for* $t = 0, 1$ *only. source [28].*

Insertion Step:(I, i)

For each segment $i = (0, ..., k)$ an insertion move is performed as follows.

- Determine the conditional distribution over the set of segmentations that can be obtained by inserting a new change-point between c_i and c_{i+1}, while holding π, α and the positions of the existing change-points constant.

- Using equation 2.6.6, the conditional probability of the current segmentation is proportional to $(L - k - 1)f(m_i, n_i|\pi, \alpha)/(k - k' + 1)$.

- Using equation 2.6.6, the conditional probability of the segmentation with a new change-point at z is proportional to $f(m_i^1, n_i^1|\pi, \alpha)f(m_{i+1}^1, n_{i+1}^1|\pi, \alpha)$ where m_i^1 and n_i^1 are, respectively, the number of '0's and number of '1's in the new segment between c_i and $z - 1$, and m_{i+1}^1 and n_{i+1}^1 are, respectively, the number of '0's and number of '1's in the new segment between z and $c_{i+1} - 1$.

- Select a new segmentation with probability proportional to $w_- = (1/N(k))(L - k - 1)f(m_i, n_i|\pi, \alpha)/(k - k' + 1)$ or select a new segmentation with a new change-point at z with probability proportional to $w_z = f(m_i^1, n_i^1|\pi, \alpha)f(m_{i+1}^1, n_{i+1}^1|\pi, \alpha)/N(k + 1)$ for each $z \in \{c_i + 1, ..., c_{i+1} - 1\}$.

Improving the efficiency of insertion step (I, i)

The final step of the procedure outlined in the previous section can be improved using the following procedure. Decide whether to insert a new change-point with probability:

$$min\left[1, \frac{\sum_{z=c_i+1}^{c_{i+1}-1} w_z}{w_-}\right]$$

If the decision is made to insert a new change-point, its position $z \in \{c_i+1, ..., c_{i+1}-1\}$ is selected with probability proportional to w_z. This modification increases the probability of accepting an insertion, and thus slightly improves the efficiency of the algorithm. Note that if a change-point is inserted, the move type is updated to $(D, i+1)$, otherwise it remains (I, i). In either case, the move-type is then updated as in Figure 2.6.

Deletion Step: (D, i)

For each non-fixed change-point $i = (1, .., k)$, a deletion move is performed as follows.

- Determine the conditional distribution over the set of segmentations consisting of the segmentations obtained by deleting change-point i and the segmentation that can be obtained by sliding c_i to a (possibly) a new change-point between c_{i-1} and c_{i+1}, while holding π, α and the positions of the other change-points constant.

- Using equation 2.6.6, the conditional probability of the segmentation obtained by deleting change-point i is proportional to $(L-k)f(m_i, n_i|\pi, \alpha)/(k-k')$.

- Using equation 2.6.6, the conditional probability of the segmentation obtained by sliding change-point i to the (possibly) new z is proportional to $f(m_i^1, n_i^1|\pi, \alpha)f(m_{i+1}^1, n_{i+1}^1|\pi, \alpha)$, where m_i^1 and n_i^1 are, respectively, the number of '0's and number of '1's in the segment with end-points c_{i-1} and $z-1$, and

m^1_{i+1} and n^1_{i+1} are, respectively, the number of '0's and number of '1's in the segment with end-points z and $c_{i+1} - 1$.

- A straight forward GGS update would be to delete the change-point with probability proportional to $w_- = (1/N(k-1))(L-k)f(m_i, n_i|\pi, \alpha)/(k-k')$ or slide the change-point to position z with probability proportional to $w_z = f(m^1_i, n^1_i|\pi, \alpha)f(m^1_{i+1}, n^1_{i+1}|\pi, \alpha)/N(k)$ for each $z \in \{c_{i-1}+1, ..., c_{i+1}-1\}$.

Improving the efficiency of deletion step (D, i)

Decide whether to delete the change-point i with probability:

$$min\left[1, \frac{w_-}{\sum_{z=c_{i-1}+1}^{c_{i+1}-1} w_z}\right]$$

If the decision is made not to delete the change-point, its position remains unchanged. Note that if a change-point is deleted, the move-type is updated to $(I, i-1)$ otherwise it remains (D, i). In either case, the move-type is then updated as in Figure 2.6. Also note that for a fixed change-point i, the (D, i) move is replaced by the trivial move of repeating the current segmentation.

Slide Step: (S, i)

For each non-fixed change-point $i = (1, ..., k)$ a slide move is performed as follows.

- Determine the conditional distribution over the set of segmentations obtained by sliding c_i to a (possibly) new change-point between c_{i-1} and c_{i+1}, while holding π, α and the positions of the other change-points constant.

- Using equation 2.6.6, the conditional probability of the segmentation obtained by sliding change-point i to z is proportional to $f(m^1_i, n^1_i|\pi, \alpha)f(m^1_{i+1}, n^1_{i+1}|\pi, \alpha)$, where m^1_i and n^1_i are, respectively, the number of '0's and number of '1's in the

segment with endpoints c_{i-1} and $z-1$, and m^1_{i+1} and n^1_{i+1} are, respectively, the number of '0's and number of '1's in the segment with end-points z and $c_{i+1}-1$.

- Using equation 2.6.6, the conditional probability of the current segmentation is proportional to $f(m_i, n_i|\pi, \alpha)f(m_{i+1}, n_{i+1}|\pi, \alpha)$, where m_i and n_i are, respectively, the number of '0's and number of '1's in between c_{i-1} and $c_i - 1$, and m_{i+1} and n_{i+1} are, respectively, the number of '0's and number of '1's in between c_i and $c_{i+1} - 1$.

- A straight forward GGS update would be to re-select the current segmentation with probability proportional to $w_- = f(m_i, n_i|\pi, \alpha)f(m_{i+1}, n_{i+1}|\pi, \alpha)/N(k)$ or slide the change-point to position z with probability proportional to $w_z = f(m^1_i, n^1_i|\pi, \alpha)f(m^1_{i+1}, n^1_{i+1}|\pi, \alpha)/N(k)$ for each $z \in \{c_{i-1}+1, ..., c_{i+1}-1\}$.

Improving the efficiency of sliding step (S, i)

Here the probability of sliding the change-point c_i to x is:

$$\left[\frac{w_x}{w_- + \sum_{z=c_{i-1}+1}^{c_{i+1}-1} w_z} \right]$$

Note that for a fixed change-point i, the (S, i) move is replaced by the trivial move of repeating the current segmentation.

Steps π**,** α**,** μ**:**

Updates for the parameters π_t, σ^t and μ^t are conventional Gibbs updates, and involve sampling the conditional posterior distributions over various one-dimensional subspaces of the target space, in particular holding k and c constant. The conditional distributions are straight forward to obtain, though care must be taken to multiply by the appropriate Jacobian when a change of variables is performed. In all cases a change of variables are performed in such a way that the one-dimensional subspace that we wish to sample requires varying only one parameter, while holding others constant [28].

Monte Carlo Integration

For each character position in the binary sequence, thus for each column of the pair wise alignment that does not contain an indel, the posterior probability that position is contained within a given conservation class is estimated by Monte Carlo integration. The posterior probability that each character position belongs to the class in question, given an element of the sample, is then calculated and averaged over the sample [28].

2.6.2 Changept modelling for segmenting a letter encoded sequence

In computational biology, because the data to be analysed are usually categorical (DNA sequence with a four letter alphabet or protein sequences with a 20 letter alphabet), the binomial and multinomial distributions are most commonly used. The unknown parameters often correspond to the frequencies of each letter in the alphabet. The conjugate priors for the multinomial families are the Dirichlet distributions, among which Beta distribution is a special case for the binomial family. In analysing DNA sequences, we often let $\theta = (\theta_a, \theta_t, \theta_g, \theta_c)$ represent the unknown probabilities of the four nucleotides. With the simple model that each residue in the observed sequence is independent and identically distributed with frequency θ, the likelihood of an observed DNA sequence can be written as:

$$p(n_a, n_t, n_g, n_c|\theta) \propto \theta_a^{n_a}\theta_t^{n_t}\theta_g^{n_g}\theta_c^{n_c}$$

where n_a, n_t, n_g, n_c are the count of the four types of nucleotides in the sequence [29]. Thus the conjugate prior for θ is of the form:

$$\Pi(\theta) \propto \theta_a^{\alpha_a-1}\theta_t^{\alpha_t-1}\theta_g^{\alpha_g-1}\theta_c^{\alpha_c-1}$$

which is a Dirichlet distribution with parameters $\alpha = (\alpha_a, \alpha_t, \alpha_g, \alpha_c)$.

Generating the letter encoded sequence

There are various possibilities of encoding multiple-sequence alignments into a D-character alphabet, where the value of D is chosen based on two main criteria: (1) the number of species in the alignment, and (2) types biological information interested in. Below I present few different options of D-character representations.

16-character representation

Suppose we are interested in segmenting a pairwise alignment based on multiple properties of interest, eg: the degree of conservation between two species, GC level of the species and transition/transversion ratio. This can be done using a 16-character encoded sequence as the input to the program *changept* (Table 2.2).

Table 2.2: *Generating 16-character representation to encode a pairwise alignment*

Species 1	A	A	A	A	C	C	C	C	G	G	G	G	T	T	T	T
Species 2	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
Symbol	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p

eg: characters 'a', 'f', 'k' and 'p' represent the conserved bases, characters from 'e' to 'l' represent the GC content in species 1

32-character representation

Suppose we are interested in segmenting a 3 way-alignment based on the same properties mentioned above and we are not interested in strand-specific information. This can be done using a 32-character alphabet shown below by encoding alignment columns with complementary bases using the same letters. If one is also interested in extracting strand specific information, we can use a different set of alphabet to encode the complementary bases.

```
Species 1:  AAAAAAAAAAAAAAAACCCCCCCCCCCCCCCC
Species 2:  AAAACCCCGGGGTTTTAAAACCCCGGGGTTTT
Species 3:  ACGTACGTACGTACGTACGTACGTACGTACGT
Symbol:     abcdefghijklmnopqrstuvwxyzUVWXYZ
```

Modelling

The input sequence S is assumed to be formed from a finite alphabet $1, ..., D$. A segmentation of S is composed of the number of change-points k and a vector of change-point positions $A = (A_1, ..., A_k)$, where A_i is the position of the left most character in segment $i + 1$. Within each segment, the sequence is supposed to have

been generated by independent trials with D possible outcomes. The probabilities of these outcomes for segment $i = 1, .., k+1$ are $\Theta_i = (\theta_{i1}, ..., \theta_{iD})$. Thus the probability of the observed sequence is a product of binomial distributions.

$$p(S|k, A, \Theta) = \prod_{i=1}^{k+1} \prod_{j=1}^{D} \theta_{ij}^{m_{ij}} \tag{2.6.8}$$

where θ_{ij} is the probability of generating character j in segment i and m_{ij} is the number of times character j appears in segment i.

Let ϕ be the probability that any particular sequence position (except the first position) start a new segment. Thus the probability of generating a new segmentation with total k change-points at positions $A = (A_1, A_2, ..., A_k)$ is given by the joint probability of k and A conditional on ϕ:

$$p(k, A|\phi) = \phi^k (1 - \phi)^{(L-1-k)} \tag{2.6.9}$$

where L is the length of S.

We can write,

$$p(k, A, \Theta) = p(k, A) \prod_{i=1}^{k+1} p(\Theta_i) \tag{2.6.10}$$

with $p(\Theta_i)$ a Dirichlet distribution with parameter vector $\alpha = (\alpha_1, ..., \alpha_D)$ as $p(k, A)$ assumed to be independent of $p(\Theta)$.

Further, a beta prior $B(a, b)$ is adopted for ϕ and the non standard prior density:

$p(\alpha_1, ..., \alpha_D) \propto \left[\frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)}\right]^{c-1} e^{-d \sum_j \alpha_j}$ is adopted to α.

Most of the inferences are performed using $a = b = 0.001$ and $c = d = 0$.

Using the Bayes' rule the posterior distribution is given by:

$$p(k, A, \Theta, \phi, \alpha|S) \propto p(\phi) p(k, A|\phi) p(\alpha) \prod_{i=1}^{k+1} p(\Theta_i|\alpha) p(S|k, A, \Theta) \tag{2.6.11}$$

Sampling

The posterior distribution $p(k, A, \Theta, \alpha|S)$ can be sampled using the GGS. The first step is to define a set of move types, analogous to the coordinate updates of the conventional Gibbs sampler.

- Insertion move (i, I): Decides whether to reselect the current segmentation Θ_i or to insert a new change point in segment i and new values are selected for Θ_i and Θ_{i+1}.

- Deletion move (i, D): Decides whether to delete the change point i and new value is selected for Θ_i or to slide the change point i to the new position x.

- Conventional Gibbs updates are done for each of the parameters $\alpha_1, ..., \alpha_D$.

- An additional Gibbs update for the sum $\beta = \sum_j \alpha_j$ while holding the proportions $\gamma_j = \alpha_j / \sum_j \alpha_j$ constant, to improve the convergence and mixing times of the Markov chain.

The total number of move types available, $N(k) = 2k + D + 2$ can be obtained by adding up, $k + 1$ - I moves, k - D moves and $D + 1$ hyper parameter updates. The algorithm cycles through these moves in the order shown in Figure 2.7 below with $D+1$ hyper parameter updates performed after the last insertion move of each iteration and before the first insertion move of the next iteration.

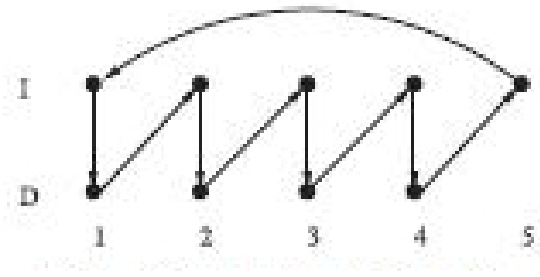


Figure 2.7: *Order of move types for the sampler. source [27].*

Insertion Step:(i, I)

If a decision is made not to insert a change-point, Θ_i updated from the Dirichlet distribution with parameter vector $(m_{i1} + \alpha_1, ..., m_{iD} + \alpha_D)$. Therefore the probability of not inserting a change-point is proportional to:

$$w_- = \frac{\Gamma(k+a)\Gamma(L-1-k+b)\prod_j \Gamma(m_{ij}+\alpha_j)}{\Gamma(L-1+a+b)\Gamma\left(\sum_j (m_{ij}+\alpha_j)\right)} \frac{1}{N(k)}$$

If a decision is made to insert a change-point at some position in segment i, Θ_i and Θ_{i+1} are sampled from Dirichlet distributions with parameter vectors $(m^1_{i1}+\alpha_1, ..., m^1_{iD}+\alpha_D)$ and $(m^1_{(i+1)1} + \alpha_1, ..., m^1_{(i+1)D} + \alpha_D)$ respectively. Therefore the probability of inserting a change-point at any given position x in segment i is proportional to:

$$w_x = \frac{\Gamma(k+1+a)\Gamma(L-2-k+b)\prod_j \Gamma(m^1_{ij}+\alpha_j)\Gamma(m^1_{(i+1)j}+\alpha_j)}{\Gamma(L-1+a+b)\Gamma\left(\sum_j (m^1_{ij}+\alpha_j)\right)\Gamma\left(\sum_j (m^1_{(i+1)j}+\alpha_j)\right)} \frac{1}{N(k+1)}$$

where $(k+1, A^1)$ is the new segmentation and m^1_{ij} and $m^1_{(i+1)j}$ are the number of times character j appears in the segments left and right to the x.

Deletion Step:(i, D)

If a decision is made to delete the change-point i, the probability is proportional to:

$$w_- = \frac{\Gamma(k-1+a)\Gamma(L-k+b)\prod_j \Gamma(m_{ij}+\alpha_j)}{\Gamma(L-1+a+b)\Gamma\left(\sum_j (m_{ij}+\alpha_j)\right)} \frac{1}{N(k-1)}$$

where m_{ij} is the number of times character j appears in segment i after deleting the change-point.

If a decision is made to move the change-point i to the new position x, the probability of sliding the change-point is proportional to:

$$w_x = \frac{\Gamma(k+a)\Gamma(L-1-k+b)\prod_j \Gamma(m^1_{ij}+\alpha_j)\Gamma(m^1_{(i+1)j}+\alpha_j)}{\Gamma(L-1+a+b)\Gamma\left(\sum_j(m^1_{ij}+\alpha_j)\right)\Gamma\left(\sum_j(m^1_{(i+1)j}+\alpha_j)\right)}\frac{1}{N(k)}$$

where m^1_{ij} and $m^1_{(i+1)j}$ are number of times character j appears in the segments to the left and right of the change-point if it is moved to the new position x.

Hyper-parameter Updates

The new value of α_j is selected from the non-standard distribution proportional to:

$$\left[\frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)}\right]^{k+c} (e^{-d}\prod_j \theta_{ij})^{\alpha_j}$$

α_j is sampled using the adaptive rejection sampling technique.

Similarly the new value of $\beta = \sum_j \alpha_j$ is sampled from

$$\left[\frac{\Gamma(\beta)}{\prod_j \Gamma(\gamma_j\beta)}\right]^{k+c} \left(e^{-d}\prod_j\left(\prod_i \theta_{ij}\right)^{\gamma_j}\right)^{\beta}$$

using adaptive rejection sampling. These update formulae are obtained using $s(u,v)=1$ as the conventional Gibbs sampler.

Additional Move Types

Two additional move types were added to further enhance convergence and mixing properties of the chain.

Updating Θ_i

Updating Θ_i for each segment was done using conventional Gibbs updates. This was performed after each successful deletion or failed insertion move involving segment i.

Sliding Move

This was involved sliding change-point $i+1$ to a new position between change-point i and $i+2$, with probability proportional to w_x of choosing position x, then updating Θ_i and Θ_{i+1}. Such moves are also conventional Gibbs updates, and were performed after each failed deletion or successful insertion move involving change-point $i+1$.

These 2 additional move types add an additional $2k+1$ moves for a segmentation with k change-points, and now $N(k) = 2(2k+1) + D + 1$.

Bibliography

[1] M Algama and J M Keith. Investigating genomic structure using *changept*: A Bayesian segmentation model. *Computational and Structural Biotechnology Journal*, 10:107–115, 2014.

[2] F H C Crick. On protein synthesis. *Symp. Soc. Exp. Biol.*, 12:139–163, 1956.

[3] A Casadevall and L Pirofski. Host-Pathogen Interactions: Redefining the Basic Concepts of Virulence and Pathogenicity. *Infection and Immunity*, 67:37033713, 1999.

[4] A Casadevall and L Pirofski. Ditch the term pathogen. *Nature*, 516:165–166, 2014.

[5] K T Min and S Benzer. Wolbachia, normally a symbiont of Drosophila, can be virulent, causing degeneration and early death. *Proc Natl Acad Sci U S A*, 94: 10792–10796, 1994.

[6] K T Min and S Benzer. Wolbachia, normally a symbiont of drosophila, can be virulent, causing degeneration and early death. *Proceedings of the National Academy of Sciences of the United States of America*, 94:10792–10796, 1997.

[7] C J McMeniman, A M Lane, A W C Fong, D A Voronin, I Iturbe-Ormaetxe, R Yamada, and S L ONeill. Host adaptation of a Wolbachia strain after long-term serial passage in mosquito cell lines. *Appl Environ Microbiol*, 74:6963–6969, 2008.

[8] C J McMeniman, R V Lane, B N Cass, A W Fong, M Sidhu, Y F Wang, and S L O'Neill. Stable introduction of a life-shortening Wolbachia infection into the mosquito Aedes aegypti. *Science*, 323:141–144, 2009.

[9] A Hoffmann, B Montgomery, J Popovici, I Iturbe-Ormaetxe, P Johnson, F Muzzi, M Greenfield, M Durkan, Y S Leong, Y Dong, and et al. Successful establishment of Wolbachia in Aedes populations to suppress dengue transmission. *Nature*, 476: 454–457, 2011.

[10] T Walker, P H Johnson, L A Moreira, I Iturbe-Ormaetxe, F D Frentiu, C J McMeniman, Y S Leong, Y Dong, J Axford, P Kriesner, and et al. The wMel Wolbachia strain blocks dengue and invades caged Aedes aegypti populations. *Nature*, 476:450–453, 2011.

[11] E A McGraw and S L ONeill. Beyond insecticides: new thinking on an ancient problem. *Nat Rev Microbiol*, 11:181–193, 2013.

[12] W S Cleveland. *Visualizing data*. Hobart Press, Summit, NJ, USA, 1st edition, 1993.

[13] W S Cleveland and S J Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.*, 83:596–610, 1988.

[14] W G Jacoby. Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, 19:577613, 2000.

[15] A Gelman, J B Carlin, H S Stern, D B Dunson, A Vehtari, and D B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, third edition edition, 2013.

[16] G R Grimmett and D R Stirzaker. *Probability and Random Processes*, chapter Markov chains. Oxford University Press, New York, 3rd edition, 2001.

[17] B Walsh. Markov Chain Monte Carlo and Gibbs Sampling: Lecture notes for EEB 581, 2004.

[18] N Metropolis and S Ulam. The Monte Carlo method. *J.Amer.Statist.Assoc.*, 44: 335–341, 1949.

[19] N Metropolis, A W Rosenbluth, M N Rosenbluth, A H Teller, and E Teller. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21:1087–1091, 1953.

[20] J M Keith, D P Kroese, and D Bryant. A Generalized Markov Sampler. *Methodology and Computing in Applied Probability*, 6:29–53, 2004.

[21] M K Cowles and B P Carlin. Markov chain monte carlo convergence diagnostics: A comparative review. *J. Amer. Statist. Assoc.*, 91:883–904, 1996.

[22] A C Favre, B Bernard, and E A Salaheddine. Comparison of methodologies to assess the convergence of markov chain monte carlo methods. *Computational Statistics and Data Analysis*, 50:2685–2701, 2006.

[23] M Plummer, N Best, K Cowles, and K Vines. CODA:convergence diagnosis and output analysis for MCMC. *R News*, 6:7–11, 2006.

[24] R Development Core Team. *R: A Language and Environment for Statistical Computing*. The R Foundation for Statistical Computing, Vienna, Austria, 2011.

[25] P D Welch and P Heidelberger. Simulation run length control in presence of an initial transient. *Operations Research*, 31:1109–1144, 1983.

[26] P D Welch and P Heidelberger. A spectral method for confidence interval generation and run length control in simulations. *Comm. ACM.*, 24:233–245, 1981.

[27] J M Keith. Segmenting eukaryotic genomes with the Generalized Gibbs Sampler. *Journal of Computational Biology*, 13:1369–1383, 2006.

[28] J M Keith, P Adams, S Stephen, and J S Mattick. Delineating slowly and rapidly evolving fractions of the Drosophila genome. *Journal of Computational Biology*, 15:407–430, 2008.

[29] J S Liu and C E Lawrence. Bayesian inference on biopolymer models. *Bioinformatics*, 15:38–52, 1999.

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 2- Part 2

Declaration by candidate

In the case of Chapter 2- Part 2, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Located and reviewed articles, wrote the paper, made modifications to the manuscript as suggested by co-author and the reviewers	90

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Jonathan Keith	Provided helpful guidance and editorial work	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature	[redacted]	**Date** 9/10/15
Main Supervisor's Signature	[redacted]	**Date** 9/10/15.

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Literature Review: Part 2

Investigating Genomic Structure Using Changept: A Bayesian Segmentation Model

Objectives

This 2^{nd} part of the literature review aims to introduce the technique *sequence segmentation*, discuss a number of segmentation methods including the *changept* algorithm and also to summarise a few applications of the *changept* model.

Authorship

Manjula Algama, Jonathan M Keith

School of Mathematical Sciences, Monash University, Clayton, VIC 3800, Australia

Reference

Algama M, Keith JM. (2014). Investigating genomic structure using *changept*: A Bayesian segmentation model. *Computational and Structural Biotechnology Journal* 10(17): 107-115.

Computational and Structural Biotechnology Journal 10 (2014) 107–115

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini Review

Investigating genomic structure using *changept*: A Bayesian segmentation model

Manjula Algama, Jonathan M. Keith *

School of Mathematical Sciences, Monash University, Clayton, VIC 3800, Australia

ARTICLE INFO

Available online 27 August 2014

Keywords:
Sequence segmentation
Bayesian modelling
Generalised Gibbs sampler
Conservation levels
GC content
Non-coding RNA

ABSTRACT

Genomes are composed of a wide variety of elements with distinct roles and characteristics. Some of these elements are well-characterised functional components such as protein-coding exons. Other elements play regulatory or structural roles, encode functional non-protein-coding RNAs, or perform some other function yet to be characterised. Still others may have no functional importance, though they may nevertheless be of interest to biologists. One technique for investigating the composition of genomes is to segment sequences into compositionally homogenous blocks. This technique, known as 'sequence segmentation' or 'change-point analysis', is used to identify patterns of variation across genomes such as GC-rich and GC-poor regions, coding and non-coding regions, slowly evolving and rapidly evolving regions and many other types of variation. In this mini-review we outline many of the genome segmentation methods currently available and then focus on a Bayesian DNA segmentation algorithm, with examples of its various applications.

Contents

1. Role of genome segmentation

Identifying the distinct components of the human and other genomes is a core task in current bioinformatics, and a necessary

* Corresponding author.

http://dx.doi.org/10.1016/j.csbj.2014.08.003
2001-0370/

pre-requisite to a full understanding of the connections between genomes and phenotypes. Yet the annotation of complex eukaryotic genomes is still far from complete. Even the proportion of the genome that performs biological functions is still hotly debated, with estimates varying from 5% [1] to 80% [2]. Whatever the true figure may be, it is clear that a vast amount of the biology underlying the structure of genomes remains to be discovered. Bioinformatics has an important role to play in this endeavour, and one of its tasks is to identify segments of the genome representing elements that require annotation.

2. Segmentation methods

Several techniques have been developed to analyse variation in properties of interest across a genome and to provide clues to the nature of its components. In this article we review some of the most widely used segmentation methods and discuss the main ideas behind each technique.

2.1. Sliding window analysis

Although not technically a segmentation method, 'sliding window analysis' is the most commonly used way to profile variation in a property of interest across a genome. This technique involves averaging the property of interest over a sliding window of a predetermined length along the sequence. For example if the window size is 10, the first point is obtained by averaging the property of interest over nucleotides 1–10, the second point is the average over nucleotides 2–11, and so on. Determining the window size can be crucial: a smaller window allows for a more precise localisation of changes, however this can increase the noise. Tajima in 1991 has proposed an algorithm to determine window size [3]. The main drawback of the sliding window analysis is that it does not identify boundaries where statistically significant changes to the property in question occur. To avoid some of the disadvantages of the sliding window approach, a windowless technique based on the Z curve was introduced to analyse GC content of genomic sequences [4]. This method enables calculation of GC content at any resolution, even at a base position. Some applications of the sliding window analysis can be found in papers [5–16].

2.2. Hidden Markov models

More precise segmentation methods have been developed to identify homogenous segments as well as the locations (change-points) at which sharp changes in a particular property of interest occurs. Hidden Markov models (HMMs) are one approach capable of inferring segment boundaries. The HMM methodology is well-established, dating from the 1950s [17]. In these models, the observed sequence is considered to be composed of segments, with the sequence of each segment generated by a Markov process. The transition probabilities for each segment are determined by a hidden state, and transitions between hidden states occur at segment boundaries. The sequence of hidden states is also modelled as a Markov process. A key parameter of an HMM is the *order* of the Markov chain, that is, the number of preceding sequence positions required to condition the transition probabilities of the observed sequence. This is unknown a priori, and usually needs to be specified, although some approaches are able to infer the order, or determine it adaptively.

HMMs were first used in biological sequence analysis by Churchill [18,19]. The parameters of the model, including segment boundaries, were estimated by using the maximum likelihood method based on the expectation–maximisation (EM) algorithm [20]. HMMs have since been widely used for sequence analysis problems in bioinformatics, and an extensive literature now exists. Two important developments were the 1998 GeneMark.hmm algorithm which used an HMM to find exact gene boundaries [21] and an HMM developed by Peshkin and Gelfand in 1999 to segment yeast DNA sequences [22]. Some other important examples are included in [23–29]. The Sarment package of Python modules built by Gueguen for easy building and manipulation of sequence segmentations uses both sliding window and HMM methods [30].

HMM models have also been implemented from a Bayesian perspective. One advantage of adopting a Bayesian approach is that it provides quantification of the uncertainties in parameter estimates in the form of probability distributions. In fact, one can dispense with point estimates of parameters altogether, instead reporting marginal distributions for key parameters, such as the locations of change-points. Boys et al. in 2000 presented a Bayesian method of segmentation using HMMs when the number of segments is known [31] and later generalised this method for an unknown number of segments [32]. In 2006, the segmentation method developed by Kedzierska and Husmeier was a combination of the sliding window analysis and the Bayesian HMM [33]. Nur and co-workers in 2009 performed sensitivity analysis on priors used in the Bayesian HMM to show the impact of prior choice on posterior inference [34]. One challenge for Bayesian HMM approaches is that they are computationally intensive and are typically infeasible for segmenting large-scale sequences, without simplifying heuristics.

2.3. Multiple change-point analysis

This approach arose independently of HMMs, and has an extensive literature dating back to the 1970s [35,36]. Change-point analysis differs from HMMs in that it typically assumes no Markov dependence in either the observed sequence or the underlying sequence of hidden states. In this sense change-point models are simpler than HMMs, and have fewer parameters. However, the two types of analysis are clearly related, and it may be useful to think of change-point models as zeroth order HMMs. A key advantage of change-point models, due to their simplicity, is their reduced computational burden, a point which is of particular relevance when implementing them within a Bayesian framework.

The use of multiple change-point models in bioinformatics was pioneered by Liu and Lawrence in 1999, using a Bayesian framework [37]. In 2000, Ramensky et al. developed a similar method which uses a Bayesian estimator to measure the degree of homogeneity in segmentation [38]. In this method, optimal segmentation is obtained by maximising the likelihood function using the dynamic programming technique presented in [39]. After completion, the partition function approach is used to obtain segmentation with longer segments by filtering the boundaries. In contrast to the approach of Liu and Lawrence, this method does not use probability distributions for segment boundaries and does not use sampling. A related method is presented in [40], which uses reversible jump Markov chain Monte Carlo (RJMCMC) sampling method to estimate posterior probabilities [41]. In contrast to Liu and Lawrence, they have used Poisson intensity models as the underlying model (as opposed to multinomial likelihood). The method has been tested by applying to modelling the occurrence of ORFs along the human genome. Another Bayesian model can be found in [42].

The method on which we focus in the main part of this article [43,44] is also of this type. The method can be described as a segmentation–classification model as it not only detects change-points but also groups segments based on their sequence characteristics. The group to which a segment belongs is essentially a hidden state, in the terminology of HMMs, and the classification is unsupervised, in the terminology of machine learning. There are two main innovations in this method. The first is that the character frequencies (emission probabilities) for a given segment are not constant for all segments in a group. Instead, the character frequencies are drawn from a Dirichlet distribution specific to the group to which that segment belongs, and it is the parameters of this distribution that characterise the group. There is thus an additional layer to this hierarchical model, and this layer is

another characteristic distinguishing the model from HMMs. Allowing variation in the character frequencies for segments in a group means that this model can be used to dissect multi-modal distributions of properties of interest, a central feature in recent applications [45,46]. The second innovation in this method is the use of the Generalised Gibbs Sampler (GGS) [47], a new technique in Markov chain Monte Carlo simulation. The GGS provides highly efficient sampling from a varying dimensional space (important here as the number of change points is variable).

2.4. Recursive segmentation method

The recursive segmentation method finds segment boundaries that maximise the difference in base compositions between adjacent segments with respect to some predefined compositional measure (Jensen–Shannon divergence — D_{JS}). The process is repeated until further segmentation of sequence segments produces no statistically significant improvements. The recursive segmentation method has been widely applied to segmentation problems such as isochore detection or detection of CpG islands [48–52]. More recent applications include locating borders between coding and non-coding regions of bacteria genomes [53] and in developing IsoPlotter: a tool for studying the compositional architecture of genomes [54].

The recursive segmentation method presented in [55] is significant in that it does not require specification of the number of segment classes (something most of the other methods require). This method has been successfully used to identify alien DNAs in bacterial genomes, detect structural variants in cancer cell lines and perform alignment-free genome comparisons.

2.5. Other segmentation methods

Methods based on least squares estimation [56] and wavelet analysis [57] have also been used. Sequential importance sampling (SIS) [58], the cross-entropy method [59] and the Bayesian adaptive independence sampler [60] have also been used to find segment boundaries and parameters of the process in each segment.

Olshen et al. developed the circular binary segmentation method (CBS) in 2004 for the analysis of array-based comparative genomic hybridisation (array-CGH) data [61]. CGH (comparative genomic hybridization) is a technique for measuring DNA copy numbers at thousands of locations on a genome. The modification of conventional CGH to obtain high resolution data is called array-CGH. The variation in DNA copy number is often used to identify cancer progression. The CBS algorithm divides the genome into regions of equal DNA copy number and identifies the genomic locations of copy number transitions (change-points). In 2007, changes were made to the original CBS algorithm to enhance the speed by introducing, (1) a hybrid approach for the computation of the p-value and (2) a stopping rule for early identification of change-points [62].

In 1996, Tibshirani proposed a new method called 'lasso' (least absolute shrinkage and selection operator) for estimation in regression models, which involves constraining the sum of the absolute values of the regression coefficients [63]. This produces some coefficients that are exactly zero and hence gives interpretable models. In 2006 'fused lasso' — a generalisation of 'lasso' — was introduced to handle problems with features that can be ordered in some meaningful way [64]. The fused lasso penalises the sum of the absolute values of the coefficients and their successive differences. The method was applied along with the CBS method to estimate the copy number alterations in breast tumour data (CGH data of breast cancer cell line MDA157) [65]. CBS had difficulties in detecting change points whose alteration signals are weak (chromosome 7 and 15 of the selected cell line), but the fused lasso successfully recognised various copy number alterations. Besides identifying gains and losses in CGH data, the fused lasso can also be generalised to other analysis; for example, understanding the interactions between copy number alternations and mRNA expression levels.

Determining the number of change-points is an important aspect of change-point analysis. In 2007, Zhang et al. proposed the modified Bayes Information Criterion (BIC) as a model selection procedure for array-CGH data analysis [66]. The first term of the modified BIC is similar to the classic BIC (consisting of the log likelihood), but it differs in the terms that penalise for model dimension. One of the advantages of using the modified BIC is that it does not require a specific prior or tuning parameters, but it can only be applied to normally distributed, uncorrelated and homoscedastic data. However the modified BIC is not limited to the analysis of array-CGH data. Some other methods that adaptively determine the number of change-points can be found in [41,46,67].

The multi-scale segmentation method developed by Futschik and co-workers also estimates the number of segments and their boundaries simultaneously [68]. One advantage of this method is that it does not require distributional assumptions regarding the lengths of segments. Another feature is that this method is able to choose an appropriate number of segments with user specified probability $1 - \alpha$.

Many early statistical segmentation methods were reviewed in [69]. Elhaik et al. reviewed the performance of seven recent algorithms by segmenting human chromosome 1 based on variability of GC content [70].

3. *Changept* analysis

In the remainder of this mini-review, we focus on the *changept* program developed by Keith et al. [43,44]. This is a Bayesian multiple change-point algorithm capable of simultaneously segmenting a genomic alignment and classifying segments into one of a predefined number of segment classes. Segments can be classified according to multiple properties including level of evolutionary conservation between species, GC content and transition/transversion ratio. Program *readcp* is a part of the *changept* package that takes the outputs produced by *changept* and estimates, for each genomic position, the probability that genomic position belongs to each segment class. The package uses a highly efficient sampling technique known as the Generalised Gibbs Sampler [47] resulting in a highly efficient algorithm that enables chromosome or even genome-wide analysis. The algorithm can be used to segment a genomic alignment based on a single property of interest or multiple properties. There is no limit on number of aligned species.

4. Method

4.1. Transforming alignment

4.1.1. Segmentation based on a single property of interest (e.g.: conservation level)

Suppose we want to segment a pairwise alignment of size L based on the degree of conservation between two species. The first step is to convert the alignment into a binary sequence by replacing the alignment columns in which two DNA sequences match with a '1' and replacing columns in which they mismatch with a '0'. The gaps between alignment blocks are marked by a '#' symbol and these are considered as fixed change-points by the model. The indels (alignment gaps) in the reference species are not encoded while indels in other species are encoded using letter 'I' which will be excluded from the final analysis of the sequence. The binary sequence generated in this way is used as the input for the program *changept*.

4.1.2. Segmentation based on multiple properties (e.g.: conservation level, GC content and transition/transversion ratio)

In segmenting a pairwise alignment based on more than one property of interest, one possibility is to use a 16-character representation

Table 1
16-character representation used to encode a pairwise alignment.

Species 1	A	A	A	A	C	C	C	C	G	G	G	G	T	T	T	T
Species 2	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
Symbol	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p

($A = (a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p)$ to encode the alignment (Table 1).

In the case of a 3-way alignment, a 32-character representation ($A = (a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, U, V, W, X, Y, Z)$ is used to transform the alignment into the *changept* input sequence. Table 2 depicts the possible encoding. Indel positions in Species 2 and Species 3 are encoded using letter 'I' which will be excluded from the final analysis.

In the 3-way alignment, alignment columns with complementary bases were encoded using the same characters.

For example:

Species 1 'A', Species 2 'A', Species 3 'A' = Species 1 'T', Species 2 'T', Species 3 'T' = 'a'

Species 1 'A', Species 2 'A', Species 3 'C' = Species 1 'T', Species 2 'T', Species 3 'G' = 'b'.

In the 16-character representation, the symbols 'a', 'f', 'k' and 'p' represent the conserved bases in the alignment. Similar information is represented by symbols 'a' and 'v' in the 32-character representation. Both input sequences also contain other biologically significant information such as GC content in species and transition/transversion ratio. For example in the 16-character representation, symbols from 'e' to 'l' correspond to 'C' or 'G' content in Species 1 and similar information is represented by symbols from 'q' to 'Z' in the 32-character representation.

In the case of more than 3 aligned species, we have proposed two methods that can be used to transform an alignment. The first method is known as 'maximum frequency transformation' in which a score is assigned for each alignment column equivalent to the maximum number of nucleotides that are identical. The second method uses Fitch's algorithm [71] to compute Parsimony score — the smallest number of mutations along the evolutionary tree. See [45] which uses both methods in transforming a 4-way alignment into the *changept* input sequence.

4.2. Modelling

The complete model is presented in [43,44]. Here we only present the main idea behind the model.

The process of Bayesian modelling consists of 3 main steps [72]: (1) set up a joint probability distribution for all the variables in a problem; (2) calculate posterior distribution — the conditional probability distribution of the unobserved parameters of interest, given the observed data; (3) evaluate the model. Step (1) starts with writing down the likelihood function of the model, i.e. probability of the observed quantities given unknown parameters. This describes the stochastic process by which sequences are generated, and consequently it quantifies the probability of generating the observed sequence for any given parameter values.

In writing down the likelihood function of our model, we denote the probability of starting a new segment by ϕ, the number of fixed change-points by k' and the total number of change-points (including fixed change-points) by k. The positions of change-points are denoted by $C = (c_1, c_2, \ldots, c_k)$. We set $c_0 = 1$. For each position in the sequence, except for the first position and those immediately following a fixed change-point (marked by '#'s), a decision has to be made whether to start a new segment. Thus the probability of generating a segmentation with k change-points at $C = (c_1, c_2, \ldots, c_k)$ positions is given by:

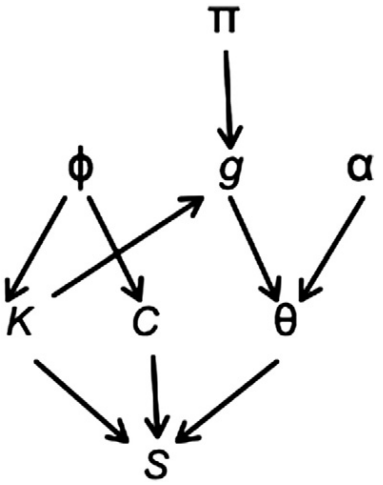


Fig. 1. Parameters of the *changept* model and their conditional dependencies. The parameter at the head of the arrow is conditionally dependent on the parameter at the tail.

$$p(k,\ C|\phi) = \phi^{k-k'}(1-\phi)^{L-1-k}$$

where L is the length of the sequence S.

Each segment is then assigned to one of ω conservation classes. Let π_t denotes the probability of assigning a segment to class t. We denote the class to which segment i is assigned by $g_i \in \{0, 1, \ldots, \omega - 1\}$ and let $g = (g_0, g_1, \ldots, g_k)$. The probability that x_0 segments are assigned to class 0, x_1 segments are assigned to class 1, …, $x_{\omega-1}$ segments are assigned to class $\omega - 1$ is:

$$p(g|k,\pi) = \pi_0^{x_0} \times \pi_1^{x_1} \times \ldots \ldots \times \pi_{\omega-1}^{x_{\omega-1}} = \prod_{i=0}^{k} \pi_{g_i}.$$

In the case of the binary representation of the sequence S, let θ_i represent the probability of generating a '1' in each position of segment i in class t. Each θ_i is independently drawn from the following beta distribution with unknown parameters $\alpha_0^{(t)}$ and $\alpha_1^{(t)}$.

$$p\left(\theta_i|\alpha_0^{(t)}, \alpha_1^{(t)}\right) = \frac{\Gamma\left(\alpha_0^{(t)} + \alpha_1^{(t)}\right)}{\Gamma\left(\alpha_0^{(t)}\right)\Gamma\left(\alpha_1^{(t)}\right)} \theta_i^{\alpha_1^{(t)}-1}(1-\theta_i)^{\alpha_0^{(t)}-1}.$$

Here $\theta = (\theta_0, \theta_1, \ldots, \theta_k)$, $\alpha^{(t)} = (\alpha_0^{(t)}, \alpha_1^{(t)})$ and $\alpha = (\alpha^{(0)}, \alpha^{(1)}, \ldots, \alpha^{(\omega-1)})$.

This can be generalised when S represents the alignment formed using a finite alphabet $\{1, \ldots, D\}$ (D-character representation). Let θ_{ij} represent the probability of generating character j in segment $i = 0, \ldots, k$. We denote $\Theta_i = (\theta_{i1}, \ldots, \theta_{iD})$. Then for each segment i in class

Table 2
32-character representation used to encode a 3-way alignment.

Species 1	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
Species 2	A	A	A	A	C	C	C	C	G	G	G	G	T	T	T	T	A	A	A	A	C	C	C	C	G	G	G	G	T	T	T	T
Species 3	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
Symbol	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	U	V	W	X	Y	Z

g_i, Θ_is are drawn from a Dirichlet distribution $p(\Theta_i|\,\alpha, g_i)$ with parameter vector $\alpha = (\alpha_1^{(t)}, \ldots, \alpha_D^{(t)})$ for each class.

The binary sequence within each segment i is generated by independent Bernoulli trials at each position in the segment. Thus the probability that segment i contains specific sequence S_i including m_i number of '0's and n_i number of '1's is given by:

$$p(S_i|L_i,\theta_i) = \theta_i^{n_i}(1-\theta_i)^{m_i}$$

where $L_i = c_{i+1} - c_i$ is the length of segment i.

In using the D-character representation, we assume that within each segment, the sequence is generated by independent trials with D possible outcomes. Let m_{ij} be the number of times character j appears in segment i. Thus the likelihood of an observed DNA sequence can be written as:

$$p(S|k,C,\Theta) = \prod_{i=0}^{k}\prod_{j=1}^{D}\Theta_{ij}^{m_{ij}}.$$

The final sequence is obtained by concatenating sequences $S_0,\ldots,S_k$. Therefore the joint distribution of parameters k, c, g, θ and S is given by:

$$(k,c,g,\theta,S|\phi,\pi,\alpha) = p(k,c|\phi)\,p(g|k,\pi)\prod_{i=o}^{k}B\left(\theta_i|\alpha^{(g_i)}\right)p(S_i|L_i,\theta_i).$$

The prior probabilities assigned to parameters ϕ, π and α are given in [44]. Using Bayes theorem, integrating over ϕ and θ, and summing over g, the following posterior distribution is obtained:

$$p(k,c,\pi,\alpha|S) = \Gamma(L-k)\Gamma(k-k'+1)\prod_{i=0}^{k}f(m_i,n_i|\pi,\alpha)$$

where

$$f(m,n|\pi,\alpha) = \sum_t \pi_t \frac{\Gamma\left(\alpha_0^{(t)}+\alpha_1^{(t)}\right)}{\Gamma\left(\alpha_0^{(t)}\right)\Gamma\left(\alpha_1^{(t)}\right)}\frac{\Gamma\left(m+\alpha_0^{(t)}\right)\Gamma\left(n+\alpha_1^{(t)}\right)}{\Gamma\left(m+\alpha_0^{(t)}+n+\alpha_1^{(t)}\right)}.$$

Table 3
8-character representation used to encode a pairwise alignment.

Species 1	A	T	A	T	A	T	A	T	C	G	C	G	C	G	C	G
Species 2	A	T	C	G	G	C	T	A	A	T	C	G	G	C	T	A
Symbol	a	a	b	b	c	c	d	d	e	e	f	f	g	g	h	h

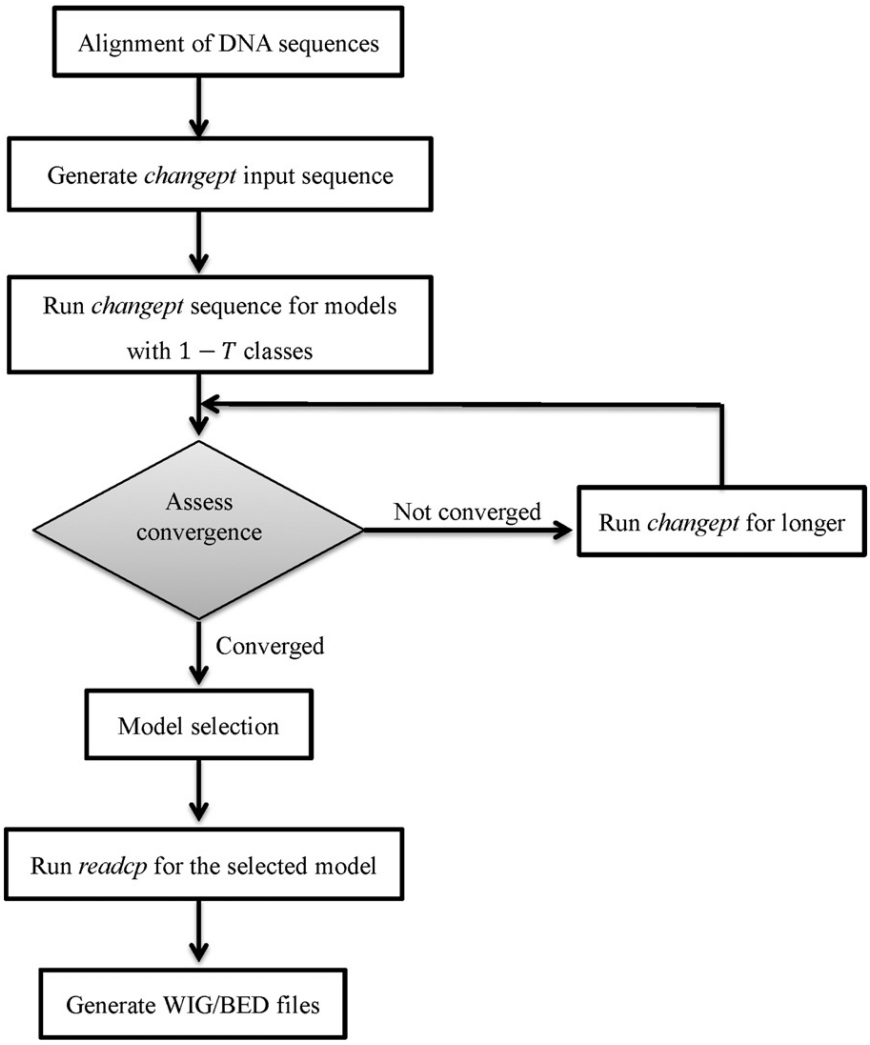


Fig. 2. The *changept* workflow. This figure illustrates the sequence of steps generally followed in analysing a set of DNA sequences by using the program *changept*. In step 3, T represents the number of segment classes specified by the user.

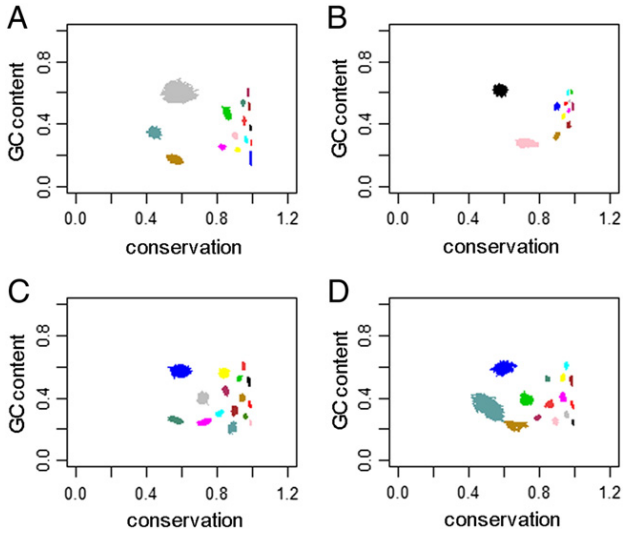


Fig. 3. GC content versus conservation level for selected models. GC content (in the first named species of each pair) versus the proportion of alignment matches, for each model is shown. The different colours represent different classes, and each class is plotted for the post burn-in samples; A) 15-class model for the *D. melanogaster* versus *D. simulans* 3′UTR alignment, B) 12-class model for the *D. melanogaster* versus *D. simulans* first coding sequence (coding 1) alignment, C) 16-class model for the *D. melanogaster* versus *D. yakuba* 3′UTR alignment and D) 15-class model for the *D. simulans* versus *D. yakuba* 3′UTR alignment.

In the case of the D-character representation, the posterior distribution is given by:

$$p(k, C, \Theta, \phi, \alpha, g, \pi|S) \propto p(\phi)p(k, C|\phi)p(\alpha)p(\pi)\ p(g|k, \pi) \prod_{i=1}^{k+1} p(\Theta_i|\alpha,\ g_i)p(S|k, C, \Theta).$$

Here $p(\phi)$, $p(\alpha)$ and $p(\pi)$ denote the prior probabilities assigned to parameters ϕ, α and π [43]. In simplifying further, it is possible to integrate the above equation over ϕ and Θ and to take sum over g to obtain the posterior distribution of $p(k, C, \alpha, \pi|S)$.

Fig. 1 shows the parameter dependencies of the model.

4.3. Sampling

The posterior distribution is sampled using the Generalised Gibbs Sampler (GGS), a Markov chain Monte Carlo technique [47]. Unlike the conventional Gibbs sampler, the GGS takes into account the fact that the number of change-points is varying and thus provides an alternative to the reversible-jump sampler [41]. It cycles through each segment and either inserts a change-point, deletes a change-point or updates the change-point positions. These different types of updates are referred as 'move-types' which are analogous to the coordinate updates of the conventional Gibbs sampler.

Once the alignment is transformed into the *changept* input sequence, it is then run through the program *changept* (source code is available upon request) to produce a user specified number of samples.

The next step of *changept* analysis is to check if convergence to the limiting distribution has occurred. This is most commonly assessed by inspecting a time-series plot of the log-likelihood against the sample number. The same plot is used to decide the length of the 'burn-in' period. *Changept* currently requires the user to specify the number of segment classes (T). Selecting the model with the most appropriate number of classes can be done by using either of the following methods: (1) investigating AIC, BIC and DICV plots [67]; and (2) investigating the stability of each segment class [46]. The final model is then run through the program *readcp* to calculate profile values. The profile shows the probability that each position in the input sequence belongs to one of the segment classes in the selected model. These posterior probabilities are estimated using Monte Carlo integration. These outputs (a profile file for each segment class in the final model) are used to generate WIG/BED files that can be uploaded to a genome browser (e.g. http://genome.ucsc.edu/) for viewing gene-related information.

This workflow is illustrated in Fig. 2 and a full description of how to use *changept* and *readcp* can be found in [73].

4.4. Applications of changept

In this section we discuss several applications of program *changept*. These can be categorised into sub-headings:

- Investigate segmentation patterns of genomic regions
- Identify alternatively spliced exons
- Identify putative transcription factor binding sites (TFBS)
- Identify putative non-coding RNAs
- Identify rapidly evolving genomic regions.

In each sub-heading we provide examples to illustrate the performance of the program *changept*.

4.4.1. Investigate segmentation patterns of genomic regions

This section summarises the results of [46]. The program *changept* was applied to three possible pairwise alignments of 3′UTR among three closely related *Drosophila* species: *Drosophila melanogaster*, *Drosophila simulans* and *Drosophila yakuba*. We also segmented three randomly selected portions of the alignment of *D. melanogaster* to *D. simulans* protein-coding sequences of the same length as the 3′UTR alignment of that pair. This was required as the number of segment classes detectable is sensitive to the length of the *changept* input sequence. These alignments were obtained from http://genomics.princeton.edu/AndolfattoLab/Andolfatto_Lab.html. Each pairwise alignment is encoded using an 8-character representation (Table 3) that

Table 4
Segmentation characteristics of two genomic regions.

Alignment	Component	Model	No. of alignment columns	No. of fixed change-points	Posterior average no. of change-points	Posterior average length of segments
Dme[a] vs Dsi[b]	3′UTR	15	2,678,635	9112	50,001	54
Dme vs Dya[c]	3′UTR	16	2,486,711	8622	53,051	47
Dsi vs Dya	3′UTR	15	2,481,568	8607	51,547	48
Dme vs Dsi	Coding 1[d]	12	2,680,987	6760	11,086	242
Dme vs Dsi	Coding 2[d]	12	2,681,121	6626	10,190	263
Dme vs Dsi	Coding 3[d]	14	2,681,284	6463	9982	268

[a] Dme: *D. melanogaster*.
[b] Dsi: *D. simulans*.
[c] Dya: *D. yakuba*.
[d] Coding 1, 2, 3: three different randomly selected protein-coding sequences.

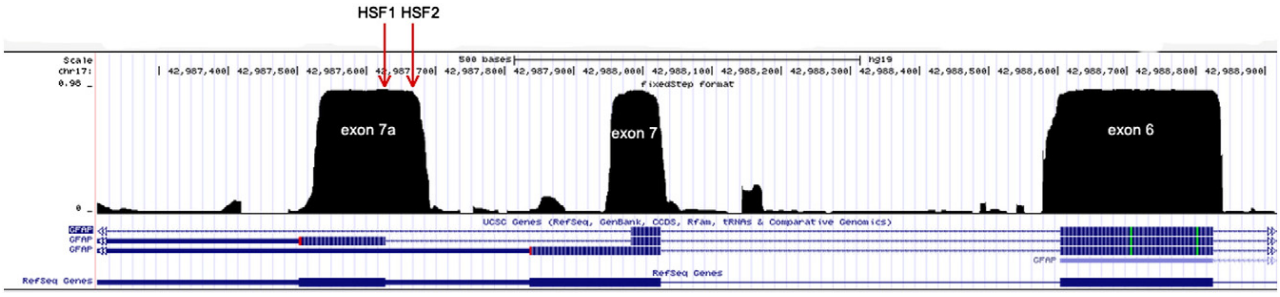


Fig. 4. Conserved features across exon 6/7/7a of GFAP. This profile corresponds to the most conserved segment class of the 4-class model. The profile value shows the probability that the base at each position of the GFAP gene belongs to the most conserved class. Exons (wide bars), UTRs (narrow bars) and introns (arrowed lines) are shown for three genes in the UCSC collection and one in RefSeq. HSF1 and HSF2 mark the actual and possible acceptor sites identified by Human Splice Finder (scores 93.19 and 76.63 respectively).

captures degree of conservation between two species, GC content and transition/transversion ratio.

In order to select the optimal number of segment classes for each alignment, we performed separate segmentation analysis using models with 1–20 segment classes ($T = 1,..., 20$). After assessing stability of segment classes in each model of 3'UTRs, we selected the 15-class model for the *D. melanogaster* versus *D. simulans* alignment, the 16-class model for the *D. melanogaster* versus *D. yakuba* alignment and the 15-class model for the *D. simulans* versus *D. yakuba* alignment. Further we selected the 12-class model for the *D. melanogaster* versus *D. simulans* two protein-coding sequences (coding 1 and coding 2) and the 14-class model for the third protein-coding sequence (coding 3).

The figure (Fig. 3) shows the segmentation patterns of each of the alignments based on the conservation levels between two species and the GC content of the first species in each pair. It can be seen that segment classes identified in *D. melanogaster* versus *D.yakuba* (Fig. 3C) and *D. simulans* versus *D. yakuba* (Fig. 3D) 3'UTR alignments have very similar characteristics. Although classes detected in the 3′UTR alignment of *D. melanogaster* versus *D. simulans* (Fig. 3A) show a similar pattern, corresponding classes appear to be compressed towards the right of the figure (i.e. higher conservation levels). This must be due to the shorter evolutionary distance between *D. melanogaster* and *D. simulans*. By contrast, the classes shown in Fig. 3B, representing the first coding sequence alignment of *D. melanogaster* versus *D. simulans*, exhibit a pattern distinct from the other three, making it difficult to identify class correspondences.

Table 4 summarises further evidence of distinct segmentation patterns of two genomic regions; 3′UTR and protein-coding.

According to these segmentation results (Table 4) it is clear that a greater number of segment classes is identified in *Drosophila* 3′UTR components compared to protein-coding regions. The number of change-points estimated in 3′ UTRs is nearly five times that estimated for coding sequence, and consequently the average segment length in 3′UTRs is about one fifth of that in the coding sequence. This evidence suggests that *Drosophila* 3′UTRs contain more numerous sub-units than protein-coding sequences.

4.4.2. Identify alternatively spliced exons

This example was extracted from work presented by Boyd SE and co-workers in segmenting a 3-way alignment (human, mouse and rat DNA sequences) of the GFAP gene [74].

Fig. 4 shows a section of the WIG file (uploaded to the UCSC genome browser) of the segment class that corresponds to regions of high conservation among human, mouse and rat of the GFAP gene. In general, the start and end points of the conserved features occur at or very close to the boundaries of the exons (e.g. exon 6 in right of the screen). In the case of exons 7 and 7a (as labelled), the conserved features do not terminate immediately after the end of the annotated exon boundaries. The conserved feature corresponding to exon 7 extends for 30 nucleotides into intron 7 and the feature corresponding to exon 7a begins 50 nucleotides upstream of the start of exon 7a.

To find the possible novel splicing sites associated with exon 7a, the human DNA sequence of the extended region has been submitted to the Human Splicing Finder server (http://www.umd.be/HSF/HSF.html). The HSF predicts a potential acceptor splice site located 40 nt upstream of the conserved region (marked by HSF2 in Fig. 4), supporting the hypothesis of a new splice variant of the GFAP gene.

4.4.3. Predict transcription factor binding sites (TFBS)

Identifying putative TFBS is yet another interesting application of the program *changept*. To test this, we selected the pairwise alignment (human versus mouse) of the SHH gene which contains experimentally identified regulatory elements within the upstream regulatory region [75]. We used LAGAN (http://lagan.stanford.edu/lagan_web/index.shtml) [76] to align the two DNA sequences. The alignment was encoded using the 16-character representation. Based on the

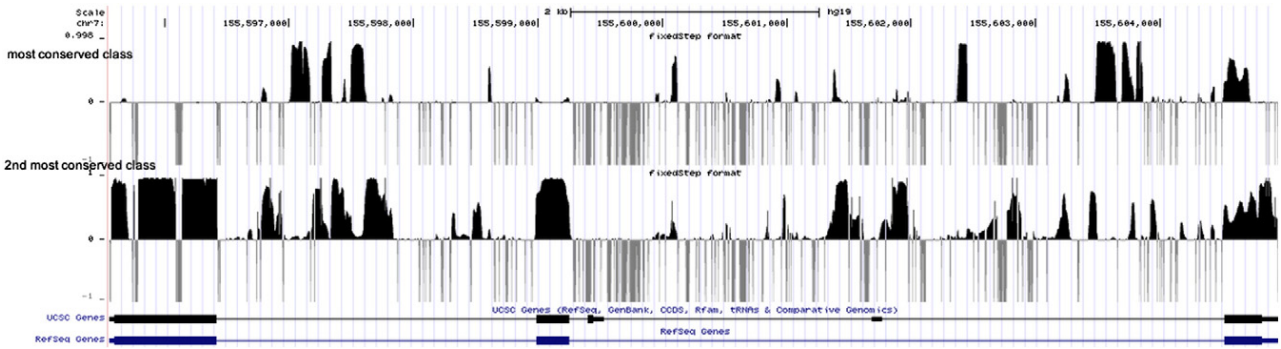


Fig. 5. WIG profiles of the two most conserved segment classes of the SHH gene. The figure shows the profiles (uploaded to UCSC genome browser) of the two most conserved classes (90% and 85% conservation levels), as identified by the program *changept* applied to the 2-way alignment of human and mouse DNA sequences. The two rows below the 2nd most conserved class profile display the exons (wide bars), the UTRs (narrow bars) and the introns (thin lines) of the SHH gene recorded in the UCSC and RefSeq collections respectively. The grey vertical lines with value −1 represent the gaps (insertions and deletions) in the original alignment as assigned by program *readcp*.

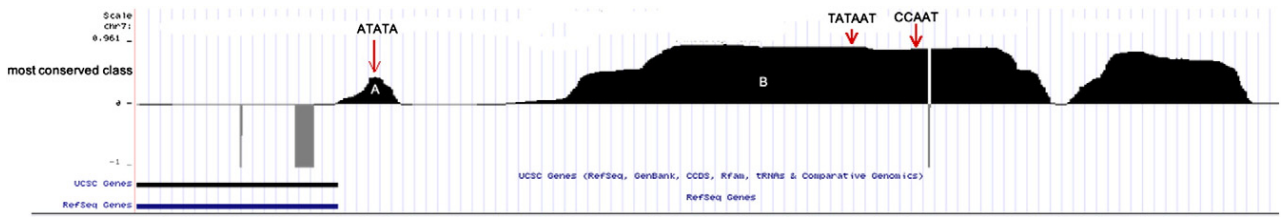


Fig. 6. Conserved regions correspond to TFBS in SHH identified by program *changept*. The profile shows the conserved features predicted by program *changept* in the upstream of SHH gene, genomic coordinates — chr7:155,604,884–155,605,370. The locations of TFBS are marked by red arrows.

investigation of DICV values, the 6-class model was selected for human and mouse 2-way alignment. Interestingly, for SHH, the positions of annotated exons were not identified as belonging to the most conserved segment class (90% conservation), rather they were identified to belong to the second most conserved class (85% conservation). Fig. 5 depicts the WIG profiles of these two most conserved segment classes.

Features A and B (Fig. 6) are regions identified as belonging to the most conserved class. These regions have been experimentally identified as regulatory elements [75].

This result confirms that regions predicted by *changept* (features A and B) are in appropriate locations for transcription factor binding. We are currently investigating the potential of *changept* for genome-wide detection of TFBS.

4.4.4. Identify putative non-coding RNAs

Non-coding RNA (ncRNA) is an RNA molecule that is not translated into a protein. It has been estimated that 98% of human genomic output is ncRNAs, however what proportion of ncRNAs are functional and the functions of many ncRNAs remain unknown [77]. The program *changept* can be used to identify highly conserved non-coding regions in genomes that are likely to be functional. To provide an example, we can use the WIG profiles of the two most conserved segment classes of SHH gene (Fig. 5). The top profile shows features that are even more conserved than the annotated protein-coding regions. Further, *changept* has predicted conserved features in the 2nd most conserved class that are equally conserved as exons. These highly conserved elements could contain either ncRNAs or regulatory sequences. In a recent project, we are working with biologists to investigate these and other putative ncRNAs identified using *changept* in a number of genomes.

4.4.5. Identify rapidly evolving genomic regions

The work presented in [44] provides an example for this *changept* application. To summarise the main findings, program *changept* has been applied on three whole-genome and three partial-genome pairwise alignments of eight *Drosophila* species. Three main classes of conservation level have been identified, comprising slowly evolving, rapidly evolving and intermediate segments. In a recent project, we are applying *changept* to three malaria species to identify genomic regions likely to be involved in the ability of the malaria parasite to infect their host species.

5. Summary

In this mini-review, we discussed various algorithms that can be used to segment genomic sequences. We also outlined the mathematics and methods of program *changept*, a Bayesian segmentation algorithm that is capable of segmenting an alignment while simultaneously classifying segments into different segment classes that share similar properties. We have demonstrated the effectiveness of this method through examples. The program *changept* can be used to identify putative functional elements in genomes such as non-coding RNAs, alternatively spliced exons and transcription factor binding sites. Other applications of program *changept* include identifying rapidly evolving genomic regions and inferring various segmentation patterns in genomic regions.

Acknowledgements

We would like to thank Dr. Robert Bryson-Richardson and Edward Tasker for their collaboration on *changept* applications. This work was supported by the Australian Research Council (grant DP1095849).

References

[1] Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. Initial sequencing and comparative analysis of the mouse genome. Nature 2002;420:520–62.
[2] Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
[3] Tajima F. Determination of window size for analyzing DNA sequences. J Mol Evol 1991;33:470–3.
[4] Zhang CT, Wang J, Zhang R. A novel method to calculate the G + C content of genomic DNA sequences. J Biomol Struct Dyn 2001;19:333–41.
[5] Bernardi G. Misunderstandings about isochores. Part 1. Gene 2001;276:3–13.
[6] Clay O, Carels N, Douady C, Macaya G, Bernardi G. Compositional heterogeneity within and among isochores in mammalian genomes. I. CsCl and sequence analyses. Gene 2001;276:15–24.
[7] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
[8] Costantini M, Auletta F, Bernardi G. Isochore patterns and gene distributions in fish genomes. Genomics 2007;90:364–71.
[9] Costantini M, Clay O, Auletta F, Bernardi G. An isochore map of human chromosomes. Genome Res 2006;16:536–41.
[10] Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW. Identification of conserved regulatory elements by comparative genome analysis. J Biol 2003;2:13.
[11] Turner TL, Hahn MW, Nuzhdin SV. Genomic islands of speciation in *Anopheles gambiae*. PLoS Biol 2005;3:e285.
[12] Spellman PT, Rubin GM. Evidence for large domains of similarly expressed genes in the Drosophila genome. J Biol 2002;1:5.
[13] Takami H, Nakasone K, Takaki Y, Maeno G, Sasaki R, Masui N. Complete genome sequence of the alkaliphilic bacterium Bacillus halodurans and genomic sequence comparison with Bacillus subtilis. Nucleic Acids Research 2000;28:4317–31.
[14] Karlin S. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends Microbiol 2001;9:335–43.
[15] Fares MA, Elena SF, Ortiz J, Moya A, Barrio E. A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. J Mol Evol 2002;55:509–21.
[16] Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, et al. Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res 2005;15:1553–65.
[17] Stratonovich R. Conditional Markov processes. Theory Probab Appl 1960;5:156–78.
[18] Churchill GA. Stochastic models for heterogeneous DNA sequences. Bull Math Biol 1989;51:79–94.
[19] Churchill GA. Hidden Markov chains and the analysis of genome structure. Comput Chem 1992;16:107–15.
[20] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B Methodol 1977;39:1–38.
[21] Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 1998;26:1107–15.
[22] Peshkin L, Gelfand MS. Segmentation of yeast DNA using hidden Markov models. Bioinformatics 1999;15:980–6.
[23] Nicolas P, Bize L, Muri F, Hoebeke M, Rodolphe F, Ehrlich SD, et al. Mining Bacillus subtilis chromosome heterogeneities using hidden Markov models. Nucleic Acids Res 2002;30:1418–26.
[24] Azad RK, Borodovsky M. Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory. Brief Bioinform 2004;5:118–30.

[25] Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology: applications to protein modeling. J Mol Biol 1994;235:1501–31.
[26] Stjernqvist S, Ryden T, Skold M, Staaf J. Continuous-index hidden Markov modelling of array CGH copy number data. Bioinformatics 2007;23:1006–14.
[27] Marioni JC, Thorne NP, Tavare S. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics 2006;22:1144–6.
[28] Willenbrock H, Fridlyand J. A comparison study: applying segmentation to array CGH data for downstream analyses. Bioinformatics 2005;21:4084–91.
[29] Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain AN. Hidden Markov models approach to the analysis of array CGH data. J Multivar Anal 2004;90:132–53.
[30] Gueguen L. Sarment: Python modules for HMM analysis and partitioning of sequences. Bioinformatics 2005;21:3427–8.
[31] Boys RJ, Henderson DA, Wilkinson DJ. Detecting homogeneous segments in DNA sequences by using hidden Markov models. J R Stat Soc: Ser C: Appl Stat 2000;49: 269–85.
[32] Boys RJ, Henderson DA. A Bayesian approach to DNA sequence segmentation. Biometrics 2004;60:573–88.
[33] Kedzierska A, Husmeier D. A heuristic Bayesian method for segmenting DNA sequence alignments and detecting evidence for recombination and gene conversion. Stat Appl Genet Mol Biol 2006;5 [Article27].
[34] Nur D, Allingham D, Rousseau J, Mengersen KL, McVinish R. Bayesian hidden Markov model for DNA sequence segmentation: a prior sensitivity analysis. Comput Stat Data Anal 2009;53:1873–82.
[35] Hawkins DM. Testing a sequence of observations for a shift in location. J Am Stat Assoc 1977;72:180–6.
[36] Worsley KJ. On the likelihood ratio test for a shift in location of normal populations. J Am Stat Assoc 1979;74:365–7.
[37] Liu JS, Lawrence CE. Bayesian inference on biopolymer models. Bioinformatics 1999; 15:38–52.
[38] Ramensky VE, Makeev V, Roytberg MA, Tumanyan VG. DNA segmentation through the Bayesian approach. J Comput Biol 2000;7:215–31.
[39] Finkelstein AV, Roytberg MA. Computation of biopolymers: a general approach to different problems. Biosystems 1993;30:1–19.
[40] Salmenkivi M, Kere J, Mannila H. Genome segmentation using piecewise constant intensity models and reversible jump MCMC. Bioinformatics 2002;18(Suppl. 2): S211–8.
[41] Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 1995;82:711–32.
[42] Husmeier D, Wright F. A Bayesian approach to discriminate between alternative DNA sequence segmentations. Bioinformatics 2002;18:226–34.
[43] Keith JM. Segmenting eukaryotic genomes with the Generalized Gibbs Sampler. J Comput Biol 2006;13:1369–83.
[44] Keith JM, Adams P, Stephen S, Mattick JS. Delineating slowly and rapidly evolving fractions of the *Drosophila* genome. J Comput Biol 2008;15:407–30.
[45] Oldmeadow C, Mengersen K, Mattick JS, Keith JM. Multiple evolutionary rate classes in animal genome evolution. Mol Biol Evol 2010;27:942–53.
[46] Algama M, Oldmeadow C, Tasker E, Mengersen K, Keith JM. *Drosophila* 3′ UTRS are more complex than protein-coding sequences. PLoS ONE 2014;9:e97336.
[47] Keith J, Kroese D, Bryant D. A Generalized Markov Sampler. Methodol Comput Appl Probab 2004;6:29–53.
[48] Bernaola-Galvan P, Roman-Roldan R, Oliver JL. Compositional segmentation and long-range fractal correlations in DNA sequences. Phys Rev 1996;53:5181–9.
[49] Oliver JL, Carpena P, Hackenberg M, Bernaola-Galvan P. IsoFinder: computational prediction of isochores in genome sequences. Nucleic Acids Res 2004;32:W287–92.
[50] Oliver JL, Roman-Roldan R, Perez J, Bernaola-Galvan P. SEGMENT: identifying compositional domains in DNA sequences. Bioinformatics 1999;15:974–9.
[51] Li W, Bernaola-Galvan P, Haghighi F, Grosse I. Applications of recursive segmentation to the analysis of DNA sequences. Comput Chem 2002;26:491–510.
[52] Cohen N, Dagan T, Stone L, Graur D. GC composition of the human genome: in search of isochores. Mol Biol Evol 2005;22:1260–72.
[53] Deng S, Shi Y, Yuan L, Li Y, Ding G. Detecting the borders between coding and non-coding DNA regions in prokaryotes based on recursive segmentation and nucleotide doublets statistics. BMC Genomics 2012;13(Suppl. 8):S19.
[54] Elhaik E, Graur D, Josic K, Landan G. Identifying compositionally homogeneous and nonhomogeneous domains within the human genome using a novel segmentation algorithm. Nucleic Acids Res 2010;38:e158.
[55] Azad RK, Li J. Interpreting genomic data via entropic dissection. Nucleic Acids Res 2013;41:e23.
[56] Haiminen N, Mannila H. Discovering isochores by least-squares optimal segmentation. Gene 2007;394:53–60.
[57] Wen SY, Zhang CT. Identification of isochore boundaries in the human genome using the technique of wavelet multiresolution analysis. Biochem Biophys Res Commun 2003;311:215–22.
[58] Sofronov G, Evans G, Keith J, Kroese D. Identifying change-points in biological sequences via sequential importance sampling. Environ Model Assess 2009;14: 577–84.
[59] Evans GE, Sofronov GY, Keith JM, Kroese DP. Estimating change-points in biological sequences via the cross-entropy method. Ann Oper Res 2011;189:155–65.
[60] Sofronov G. Change-point modelling in biological sequences via the Bayesian adaptive independent sampler, 5. , International Conference on Telecommunication Technology and Applications; 2011. p. 22–126.
[61] Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 2004;5:557–72.
[62] Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 2007;23:657–63.
[63] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol 1996;58:267–88.
[64] Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. J R Stat Soc Ser B (Stat Methodol) 2005;67:91–108.
[65] Tibshirani R, Wang P. Spatial smoothing and hot spot detection for CGH data using the fused lasso. Biostatistics 2008;9:18–29.
[66] Zhang NR, Siegmund DO. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics 2007;63: 22–32.
[67] Oldmeadow C, Keith JM. Model selection in Bayesian segmentation of multiple DNA alignments. Bioinformatics 2011;27:604–10.
[68] Futschik A, Hotz T, Munk A, Sieling H. Multiscale DNA partitioning: statistical evidence for segments. Bioinformatics 2014. http://dx.doi.org/10.1093/bioinformatics/btu1180.
[69] Braun JV, Muller H-G. Statistical methods for DNA sequence segmentation. Stat Sci 1998;13:142–62.
[70] Elhaik E, Graur D, Josic K. Comparative testing of DNA segmentation algorithms using benchmark simulations. Mol Biol Evol 2010;27:1015–24.
[71] Fitch WM, Margoliash E. Construction of phylogenetic trees. Science 1967;155: 279–84.
[72] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. 2nd ed. Taylor & Francis; 2003.
[73] Keith JM. Sequence segmentation. Methods Mol Biol 2008;452:207–29.
[74] Boyd SE, Nair B, Ng SW, Keith JM, Orian JM. Computational characterization of 3′ splice variants in the GFAP isoform family. PLoS ONE 2012;7:e33565.
[75] Kitazawa S, Kitazawa R, Tamada H, Maeda S. Promoter structure of human sonic hedgehog gene. Biochim Biophys Acta 1998;1443:358–63.
[76] Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 2003;13:721–31.
[77] Mattick JS. Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep 2001;2:986–91.

Reviewing an additional approach not mentioned in the published paper in Literature review – part 2

ChromHMM

ChromHMM developed based on Multivariate Hidden Markov Models (HMMs) provides a potential alternative approach to *changept* - that it is also designed to handle very long sequences. Multivariate HMMs are graphical probabilistic models that model multiple `observed' inputs as generated by unobserved `hidden' states, using transitions between hidden states to model spatial relationships [1]. High-dimensional multivariate datasets occur in a large number of problem domains. In many cases, these datasets have either a sequential or temporal structure [2]. To uncover which combinations of histone modifications are biologically meaningful, Ernst and Kellis [1] has developed an automated computational system - *ChromHMM*. *ChromHMM* is useful for learning chromatin sites, characterizing their biological functions and correlations with large scale functional datasets and visualizing the resulting genome-wide maps of chromatin-state annotations.

Bibliography

[1] J Ernst and M Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol, 28:817-825, 2010.

[2] K E Lee and H S Park. Review of Three Different Studies on Hidden Markov Models for Epigenetic Problems: A Computational Perspective. Genomics Inform., 12:145-150, 2014.

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 3

Declaration by candidate

In the case of Chapter 3, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Methodology developments, performed experiments, analysed data, wrote the paper, made modifications to the manuscript as suggested by co-authors and the reviewers	65

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Christopher Oldmeadow	Performed experiments, analysed data, provided editorial work	
Edward Tasker	Performed experiments, analysed data, provided editorial work	
Kerrie Mengersen	Provided editorial work	
Jonathan Keith	Conceived and designed experiments, contributed analysis tools and editorial work	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature	[redacted]	**Date** 9/10/15
Main Supervisor's Signature	[redacted]	**Date** 9/10/15

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 3

Drosophila $3'$ UTRs Are More Complex than Protein-Coding Sequences

Chapter Objectives

The overall objective of this thesis is to develop methods to identify non-protein coding genomic regions contributing to diseases. This chapter describes two main methodological developments of the *changept* model: (1) a new method to encode a pair-wise and a 3-way alignment by integrating multiple data types (on conservation, GC content, transition/transversion ratios) making it possible to clearly identify functional signatures even between very closely related species; and (2) a new model selection procedure making it possible to discover new motifs not identified by models selected using DICV values. These methods were tested on a dataset corresponding to $3'$ UTRs of three closely related *Drosophila* species to identify complex segmentation patterns. This study also discovered a number of known and predicted miRNA targets and other regulatory sequences in $3'$ UTRs of *Drosophila melanogaster*.

Authorship

Manjula Algama[1], Christopher Oldmeadow[2], Edward Tasker[1], Kerrie Mengersen[3], Jonathan M Keith[1]

1. School of Mathematical Sciences, Monash University, Clayton, VIC 3800, Australia

2. School of Medicine and Public Health, University of Newcastle, Newcastle, NSW 2300, Australia

3. School of Mathematical Sciences, Queensland University of Technology, Brisbane, QLD 4000, Australia

Reference

Algama M, Oldmeadow C, Tasker E, Mengersen K, Keith JM. (2014). Drosophila 3′ UTRs Are More Complex than Protein-Coding Sequences. *PLoS ONE* 9(5): e97336. doi:10.1371/journal.pone.0097336.

OPEN ACCESS Freely available online

PLOS ONE

Drosophila 3′ UTRs Are More Complex than Protein-Coding Sequences

Manjula Algama[1☯], Christopher Oldmeadow[2☯], Edward Tasker[1☯], Kerrie Mengersen[3], Jonathan M. Keith[1]*

1 School of Mathematical Sciences, Monash University, Clayton, Victoria, Australia, **2** School of Medicine and Public Health, University of Newcastle, Newcastle, New South Wales, Australia, **3** School of Mathematical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia

Abstract

The 3′ UTRs of eukaryotic genes participate in a variety of post-transcriptional (and some transcriptional) regulatory interactions. Some of these interactions are well characterised, but an undetermined number remain to be discovered. While some regulatory sequences in 3′ UTRs may be conserved over long evolutionary time scales, others may have only ephemeral functional significance as regulatory profiles respond to changing selective pressures. Here we propose a sensitive segmentation methodology for investigating patterns of composition and conservation in 3′ UTRs based on comparison of closely related species. We describe encodings of pairwise and three-way alignments integrating information about conservation, GC content and transition/transversion ratios and apply the method to three closely related Drosophila species: *D. melanogaster*, *D. simulans* and *D. yakuba*. Incorporating multiple data types greatly increased the number of segment classes identified compared to similar methods based on conservation or GC content alone. We propose that the number of segments and number of types of segment identified by the method can be used as proxies for functional complexity. Our main finding is that the number of segments and segment classes identified in 3′ UTRs is greater than in the same length of protein-coding sequence, suggesting greater functional complexity in 3′ UTRs. There is thus a need for sustained and extensive efforts by bioinformaticians to delineate functional elements in this important genomic fraction. C code, data and results are available upon request.

Citation: Algama M, Oldmeadow C, Tasker E, Mengersen K, Keith JM (2014) Drosophila 3′ UTRs Are More Complex than Protein-Coding Sequences. PLoS ONE 9(5): e97336. doi:10.1371/journal.pone.0097336

Editor: Sudhindra R. Gadagkar, Midwestern University, United States of America

Received December 11, 2013; **Accepted** April 18, 2014; **Published** May 13, 2014

Funding: This work was supported by the Australian Research Council (grants DP0556631, DP0879308 and DP1095849), the National Health and Medical Research Council (grant ID 389892) and by a Vice Chancellor's Research Fellowship funded by Queensland University of Technology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* [redacted]

☯ These authors contributed equally to this work.

Introduction

The fundamental role played by non-protein-coding functional DNA and RNA in cellular processes is no longer contentious. Various lines of evidence have contributed to recognition of its importance. Ever since it became possible to compare two mammalian genomes, it has been clear that far more is conserved than just the protein-coding component [1]. In mammals, unsurprisingly since the encoded proteome is relatively stable, it has been determined that non-coding elements are the predominant source of evolutionary innovation [2], much of which is due to variation in the regulatory architecture [3]. In the human genome, genetic association studies have identified numerous disease-associated genetic variants in non-protein-coding regions [4–6]. The ENCODE project, which aims to catalogue all components of the human genome, has found evidence that at least ~80% of the human genome is functional, where a functional element is defined as "a discrete genome segment that encodes a defined product (for example, protein or non-coding RNA) or displays a reproducible biochemical signature (for example, protein binding, or a specific chromatin structure)" [7]. Moreover, the ENCODE study identifies that ~60% of the genome is included in at least one long (>200 bases) RNA transcript. The ENCODE definition of function, and the 80% estimate, have been sharply criticised [8,9] but this debate does not obscure a broad consensus that the functional component of the genome far exceeds the ~1.2% that codes for proteins. It is also becoming increasingly clear that genome-wide transcription is regulated and profoundly complex [10].

The 3′ UTRs of protein-coding genes are a likely source of as yet uncharacterised functional non-protein-coding elements, because this genomic fraction is not only transcribed but also associated with known functional elements (the corresponding genes). There is growing awareness of the crucial importance of 3′ UTRs in post-transcriptional regulation of protein expression (for example [11]). Mutations in 3′ UTRs have been shown to play a crucial role in human health and disease, perhaps as much as that of coding sequences [12]. Our own interest in 3′ UTRs stems from previous work in which we found that a highly conserved component of Drosophila genomes was highly enriched in fragments of sequence from 3′ UTRs [13].

A recent review [14] catalogues a wide range of functional elements in 3′ UTRs. One motif found in 3′ UTRs is the polyadenylation signal with consensus sequence AAUAAA. This signal occurs approximately 10–30 nucleotides upstream of the site at which a pre-mRNA is cleaved prior to polyadenylation, and acts as a protein binding site around which a complex multi-protein assembly forms. A number of other motifs are also known to

participate in the process of polyadenylation. More than half of human genes contain alternative polyadenylation sites, resulting in isoforms that differ only in the length of the 3′ UTR. Individual isoforms are also differentially expressed in different cell types and developmental stages. This has important consequences for post-transcriptional regulation, as isoforms with shorter 3′ UTRs tend to be more stable, partly because the shorter isoforms may exclude binding sites for microRNAs. Such binding sites are another common functional element in 3′ UTRs, and in fact most miRNA binding sites are located in 3′ UTRs.

Other key regulatory sequences found in 3′ UTRs include: AU-rich elements and GU-rich elements, to which proteins involved in mRNA degradation bind; a CU-rich element known as the differentiation control element (DICE) to which proteins that inhibit translation initiation bind; other CU-rich elements bound by proteins including polypyrimidine-tract binding protein (PTB), which modulates a variety of mRNA processes including splicing and polyadenylation; CA-rich elements to which proteins that stabilise mRNAs bind; and motifs that form stem-loop structures recognised by specialised regulatory proteins. Repetitive motifs within 3′ UTRs have previously been demonstrated to direct the cellular localisation of mRNA transcripts [15]. Andken *et al.* [15] identify computationally a CAG repeated motif common to many mammalian genes which localise to the dendrites of neurons, and validate experimentally in two specific cases that the correct localisation is dependent on the presence of this motif. Numerous other functional binding sites in 3′ UTRs are known. The database UTRsite maintains a list of experimentally validated functional motifs in UTRs [16].

In this paper, we assess the complexity of 3′ UTRs relative to that of protein-coding sequences, by comparing the extent to which segmental substructures can be detected within these two genomic fractions based on sequence composition and conservation. We argue that the degree of segmental substructure is a useful proxy for functional complexity. We find that segmental substructures in 3′ UTRs are shorter on average, more numerous and more varied in type than in protein-coding sequence. Annotation of function in 3′ UTRs will therefore not be complete until it is rather more detailed than the annotation of protein domains in protein-coding sequences. We therefore echo [17] in calling for bioinformaticians to turn their attention to annotation of this important genomic fraction.

Our methodology involves comparing closely related species, which may seem unusual given that functional signatures are more clearly distinguishable from background patterns at greater evolutionary distances. However, we suspect that full elucidation of the functional component of 3′ UTRs may require comparison of closely related species, in addition to conventional comparisons of more distantly related species. Furthermore, it may require consideration of additional data not based on species comparisons, and perhaps unique to individual species. The reason for this is that some functional components of genomes may be ephemeral, that is, may persist in genomes only briefly relative to evolutionary time-scales, perhaps so briefly as to be unique to one extant species.

The existence of such ephemeral functional elements is an inevitable consequence of genetic drift. In finite populations, beneficial mutations are not guaranteed to become fixed, and those that do may subsequently be eliminated in the lottery of genetic drift, particularly if the advantage conferred is slight. Recently evolved functional elements whose integration into the system is not yet optimal are perhaps more vulnerable to random extinction, despite the selective pressures that favour their survival. Such functional turnover is certain to occur in evolving genomes, but the proportion of the human and other genomes currently under ephemeral constraints is not known.

Evidence possibly indicative of ephemeral constraints was uncovered by the ENCODE pilot project [18], which found that not all bases within experimentally defined functional genomic regions show evidence of constraint, and that many functional elements are seemingly unconstrained across mammalian evolution. The authors of that paper proposed that the genome contains a large pool of "neutral elements that are biochemically active but provide no specific benefit to the organism" [18]. We consider that explanation contradictory, since it is intended to address the observation that *functional* elements are seemingly unconstrained, and function implies a benefit to the organism. A more natural conclusion is that a significant proportion of the human genome is subject to ephemeral functional constraints, visible to comparative genomics studies only for closely related species, if at all. More recent ENCODE publications support this latter interpretation, for example finding that elements without detectable mammalian constraint do show evidence of negative selection in primates [7].

Evidence of large-scale turnover of transcription factor binding sites (TFBSs) has been found in Drosophila. Moses *et al.* [19] identified numerous known regulatory binding sites in *D. melanogaster* that were not present in closely related species, including *D. simulans*. There is also mounting evidence that binding of transcription factors (TFs) to seemingly non-functional 'decoy' TFBSs has subtle effects on the regulation of target gene expression [20,21]. Low information content decoy TFBSs are frequently created and destroyed by point mutations and are likely candidates for functional elements under ephemeral constraints. Similarly, post-transcriptional binding sites in 3′ UTRs are mostly low information content sequences that are potentially subject to rapid turnover.

In this paper, we present a sensitive methodology for investigating patterns of conservation and sequence composition in pairwise and three-way alignments of closely related species. Segmentation models are well suited to detecting subtle variations in sequences, and have a long history of use in bioinformatics [22]. In such models, it is assumed that the sequence (usually, but not limited to, DNA) can be partitioned into a series of segments, each with some degree of internal homogeneity. The challenge is to find the positions that delineate the segments (known as change-points). Bayesian models are attractive in these circumstances as they are apt for modelling complex hierarchies, and also provide a natural framework to model uncertainty. The seminal paper for such models is [23], and the approach has recently been developed and extended [13,24–26]. Our Bayesian model and associated Markov chain Monte Carlo (MCMC) sampler were developed for the segmentation of sequences derived from pairwise and multiple alignments.

In earlier work [13], three main classes of conservation level were identified in Drosophila, corresponding to slowly evolving, rapidly evolving and intermediate segments. A more recent analysis involved generalizing the Bayesian segmentation technique to identify patterns of conservation variation in multiple sequence alignments [26]. The method was able to distinguish multiple classes of evolutionary rate; 7 in an alignment of four mammals (including humans) and 9 for an alignment of four drosopholids. The classes were indicative of different degrees of selection acting in a segmented pattern over the genome, the scale of which was much finer than could be attributed to local variations in the neutral mutation rate. These findings indicated a significant problem with the conventionally assumed dichotomy of conservation level (conserved or not) used in many previous analyses based on evolutionary rates [1,18,27–30]. They also

highlighted the importance of sophisticated analyses capable of detecting sub-classes of evolutionary rates, for investigating the vastly complex landscape of evolution. A recent simulation study by the authors [31] demonstrated that this technique does not detect superfluous modes, confirming the above conclusions.

Despite the success of the segmentation approach, it is clear that conservation data alone will not provide sufficient power to detect unique functional signatures. This point is particularly relevant in the analysis of closely related species, where distinctions in conservation level are likely to be fine and difficult to detect. We therefore generalise the segmentation approach for sequences formed from characters of an arbitrary alphabet, making it well suited to incorporate other sequence characteristics that are also suggestive of function. We consider the problem of integrating multiple data types, with the aim of identifying classes on a finer scale than previously. This issue is explored briefly in [13], and raised as area which requires further study. Here we segment and classify the 3′ UTR sequence of *D. melanogaster* based on three data types: conservation relative to one or two other species (based on alignment matches and mismatches), GC content, and transition/transversion rates. We illustrate the methodology for the three pairwise, and one 3-way alignment of *D. melanogaster*, *D. simulans* and *D. yakuba* 3′ UTR sequences. The classes thus identified represent a resource for the future discovery of novel functional elements in Drosophila. We also examined several of our identified classes and investigated the extent to which they display properties consistent with function, and explore potential functional roles of motifs identified to be enriched within the different classes.

Results, Discussion and Conclusions

We applied our segmentation method to the 3-way alignment and three possible pairwise alignments of 3′ UTRs among the species *D. melanogaster*, *D. simulans* and *D. yakuba*. We also applied the method to four different types of control sequence. To compare the segmentation patterns detected in 3′ UTRs to those of known functional sequences, we segmented a randomly selected portion of the alignment of *D. melanogaster* to *D. simulans* protein-coding sequences, of the same length as the 3′ UTR alignment for that species pair. The requirement that this coding alignment be the same length is necessary because the number of segment classes identified is sensitive to the length of the input sequence. In general, more classes can be detected with a longer input sequence. This process was repeated three times with different coding sequences, to ensure that the results were reproducible. In order to demonstrate the advantage of incorporating multiple data types into an 8-character representation, we segmented a binary representation of conservation (matches/mismatches) in the *D. melanogaster* versus *D. simulans* 3′ UTR alignment. Similarly, we segmented a binary representation of GC content in *D. melanogaster* 3′ UTRs. Lastly, we segmented an artificially generated control sequence with only one class of segments. The artificial sequence was generated using the same overall character frequencies, and to be the same length as the *D. melanogaster* versus *D. simulans* 3′ UTR alignment.

Model Selection

At present our segmentation algorithm requires the user to specify the number of segment classes T. Separate segmentations were therefore performed for each value of T in the range 1–20. Two different procedures were then applied to select the number of classes for each alignment; investigating Deviance Information Criterion V (DICV) values (Procedure 1) and investigating the stability of the classes (Procedure 2). Figure 1 shows plots of the model selection criterion (DICV) versus T for the segmentations of four 8-character alignment representations. Based on these plots, using Procedure 1, we selected the 12-class model for the *D. melanogaster* and *D. simulans* 3′ UTR alignment (Figure 1A), the 10-class model for the *D. melanogaster* and *D. yakuba* 3′ UTR alignment (Figure 1C), the 12-class model for the *D. simulans* and *D. yakuba* 3′ UTR alignment (Figure 1D), and the 14-class model for the 3-way 3′ UTR alignment (Figure S1).

Using Procedure 2, we selected the 15-class model for the *D. melanogaster* versus *D. simulans* alignment, the 16-class model for the *D. melanogaster* versus *D. yakuba* alignment, the 15-class model for the *D. simulans* versus *D. yakuba* alignment, and the 15-class model for the 3-way alignment. The numbers of classes selected for each sequence by each procedure are summarised in Table 1. In general, Procedure 1 selects a model with fewer classes than Procedure 2.

Comparison to Control Sequences

Table 1 indicates that twelve to fourteen segment classes with distinct character frequencies can be distinguished in each of the three coding sequence alignments, using Procedure 1 or Procedure 2. The DICV values used for Procedure 1 and one of the three coding sequence alignments are shown in Figure 1B. It is not surprising that such a large number of classes can be detected in coding sequence, given that it consists of numerous sub-units (protein domains) subject to a variety of structural and functional constraints. What is perhaps surprising is that a similar number of classes can be detected in 3′ UTRs, and in fact Procedure 2 consistently identifies a greater number of classes in 3′ UTRs. The implication is that 3′ UTRs contain numerous sub-units subject to an even greater variety of structural and functional constraints than coding sequence. This is in line with the continuing focus in genomics on the significant regulatory and evolutionary role of non-coding sequences, particularly in regard to the regulation of gene expression. Further evidence that 3′ UTRs may have more complex sub-structures than coding sequences is shown in Table 2. The number of change-points estimated in 3′ UTRs is nearly **five times** that estimated for coding sequence, and consequently the average segment length in 3′ UTRs is about one fifth that in coding sequence. Many of these change-points may correspond to the boundaries of functional elements. The values shown in Table 2 were obtained using models selected by Procedure 2, but the same conclusions were reached using models selected by Procedure 1.

Both model selection procedures identified a significantly larger number of segment classes than our previous studies using binary sequence representations of pairwise alignments [13,25]. Figures 2 and 3 demonstrate why this is the case. The figures show, for the two model selection procedures and the four 8-character alignments, the estimated GC content versus conservation level (proportion of matches) for the classes identified. These are time series plots over the MCMC sample, so the size of the 'blobs' is an indication of uncertainty. It is clear from these plots that the use of multiple data types has enabled a greater number of classes to be distinguished, because projection onto either of the 'GC content' or 'conservation' axes would make many of these classes indistinguishable. The same information for the 3-way alignment of 3′ UTRs is shown in Figure S2.

To further clarify this point, we compared the number of classes found using the 8-character representation to the number obtained using the binary sequence representing the conservation of *D. melanogaster* relative to *D. simulans* 3′ UTRs (see Table 1). Similarly, we also determined the number of classes found using the binary sequence representing GC content of *D. melanogaster* 3′

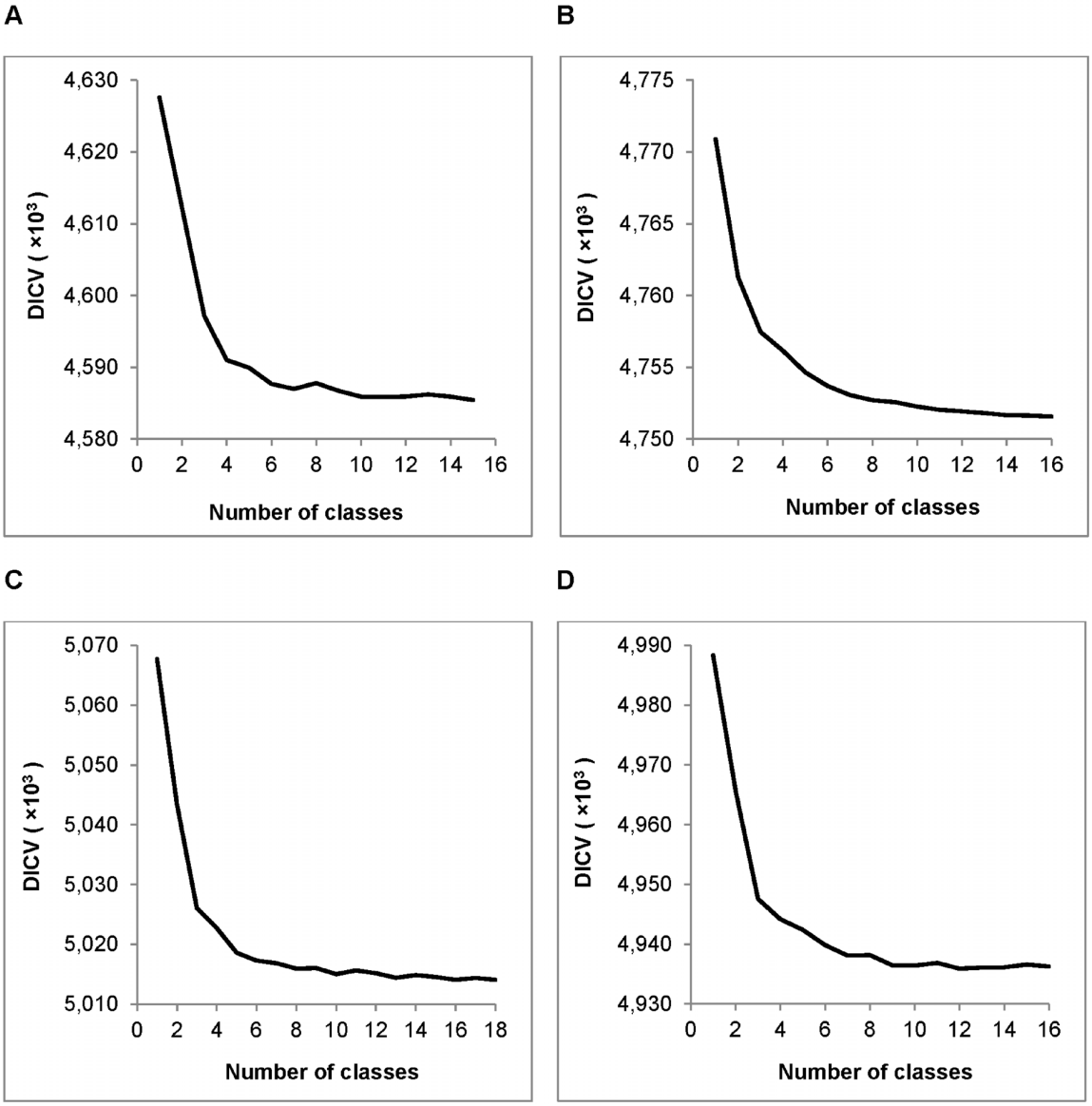


Figure 1. DICV values for segmentation of four alignments. DICV values obtained using a varying number of classes, for four input sequences derived from A) *D. melanogaster* versus *D. simulans* 3′ UTR alignment, B) *D. melanogaster* versus *D. simulans* first coding sequence (Coding 1) alignment, C) *D. melanogaster* versus *D. yakuba* 3′ UTR alignment and D) *D. simulans* versus *D. yakuba* 3′ UTR alignment.
doi:10.1371/journal.pone.0097336.g001

UTRs. Figure 4 shows the DICV values with $T = 1$–10 for the segmentation of each of the binary representations. Based on these plots, using Procedure 1, the 4-class model was selected for GC content (Figure 4A), and the 2-class model was selected for conservation (Figure 4B). Using Procedure 2, the 2-class model was selected for GC content, and the 3-class model was selected for conservation. It is clear that the numerous classes identified using the 8-character representation cannot be resolved using either GC content or conservation in isolation.

The final control sequence was artificially generated and was designed to have only one class of segments. Figure S3 shows DICV values for segmentation of this control sequence with $T =$ 1–5. Note that Procedure 1 correctly selects the 1-class model, thus supporting the use of DICV values for model selection. Figure S4 shows the time-series plot of conservation level versus sample number for segmentations of the artificially generated control sequence with $T = 1$ and $T = 2$. Figure S4A shows the 1-class model is stable, whereas Figure S4B shows that one of the two classes has a widely varying conservation level. This unstable class also had a very low mixture proportion and thus the 1-class model was again selected for the control sequence using Procedure 2. This confirms results of our previous study [31] demonstrating that models selected using DICV do not typically contain superfluous modes, and are generally conservative in the number of components identified.

Consistency of Segment Classes

In this study, we have used two different model selection procedures to decide how many distinct segment classes can be identified, with Procedure 1 being generally more conservative

Table 1. Models selected using two procedures.

Alignment	Component	Encoding	Procedure 1	Procedure 2
Dme vs Dsi	UTR	8-char	12	15
Dme vs Dya	UTR	8-char	10	16
Dsi vs Dya	UTR	8-char	12	15
Dme, Dsi, Dya	UTR	32-char	14	15
Dme vs Dsi	Coding 1	8-char	12	12
Dme vs Dsi	Coding 2	8-char	12	12
Dme vs Dsi	Coding 3	8-char	14	14
Dme vs Dsi	UTR	GC alone (binary)	4	2
Dme vs Dsi	UTR	Conservation alone (binary)	2	3

Dme: *D. melanogaster*; Dsi: *D. simulans*; Dya: *D. yakuba*; Procedure 1: Models selected based on DICV values; Procedure 2: Models selected by investigating stability of classes; Coding 1, 2, 3: three different coding sequences.
doi:10.1371/journal.pone.0097336.t001

than Procedure 2, in that it favours fewer classes. The question naturally arises whether the selected number of classes T radically alters the classification, or whether the segment classes are consistent in the sense that increasing T merely results in some classes resolving into two or more subclasses. A similar question arises concerning the consistency of classes identified in the three pairwise alignments and the 3-way alignment. Given that each Drosophila species is involved in two pairwise alignments, one wonders whether comparable classifications result in all three cases.

First, we compared the models chosen by the two model selection procedures, investigating specifically the *D. melanogaster* versus *D. simulans* 3′ UTR alignment. Nine of the classes identified in the 12-class model map directly to individual classes in the 15-class model. The remaining 3 classes from the 12-class model mapped to weighted averages of two classes each from the 15-class model, indicating that the primary difference between the 12-class and 15-class models was the splitting of three classes into two subclasses each. The results of the mapping are summarised in Table S1: characteristics considered include mixture proportions, conservation levels, GC content and transition/transversion ratio.

Many of the segment classes contain, in the corresponding *D. melanogaster* regions, characteristic tandem repeat sequences detected as highly significant motifs using MEME (see Methods section 'Class Profiling'), the significance of which are discussed further in the following section. To further investigate the consistency of the 12- and 15-class models, we investigated whether the same characteristic tandem repeats were identified in corresponding classes. In the 12-class model, ten motifs were identified within six classes; within the ten motifs there were six distinct types of motif. In the 15-class model, eleven motifs were identified within eight classes; within the eleven motifs there were six distinct types of motif. Similar motif types to each of the six distinct motif types from the 12-class model were identified in the 15-class model, and in general the motif types found to be common to both models were found in the corresponding classes as identified by the previously mentioned mapping (Table S1). The 15-class model identified two additional motif types not identified in the 12-class model. For this reason, and given that difference between the 12 and 15-class models is only the splitting of three classes, our further analysis of detected motifs focuses on models identified by Procedure 2. A more detailed summary of these results is provided in Tables S2 and S3.

Secondly, we compared the classes identified in the different alignments. Figures 2 and 3 provide an initial indication that the classes detected in the three 2-way alignments of 3′ UTRs are fairly consistent. Figures 3C and 3D in particular, corresponding respectively to alignments of *D. melanogaster* versus *D. yakuba* and *D. simulans* versus *D. yakuba*, are strikingly similar, and many of the classes detected in one alignment can immediately be placed in

Table 2. Segmentation characteristics of models selected by Procedure 2.

Alignment	Component	Length	Nfixed	*k*	*L*
Dme vs. Dsi	UTR	2678635	9112	50001	54
Dme vs. Dya	UTR	2486711	8622	53051	47
Dsi vs. Dya	UTR	2481568	8607	51547	48
Dme, Dsi, Dya	UTR	2247759	8260	54523	41
Dme vs. Dsi	Coding 1	2680987	6760	11086	242
Dme vs. Dsi	Coding 2	2681121	6626	10190	263
Dme vs. Dsi	Coding 3	2681284	6463	9982	268

Length: number of alignment columns in the component; Nfixed: number of fixed change-points, corresponding to the boundaries of alignment blocks; *k*: posterior average number of change-points; *L*: posterior average length of segments. Note the length of the coding sequence is equal to that of the 3′ UTRs for the same species pair, once the number of fixed change-points (corresponding to the ends of alignment blocks) is added to the length.
doi:10.1371/journal.pone.0097336.t002

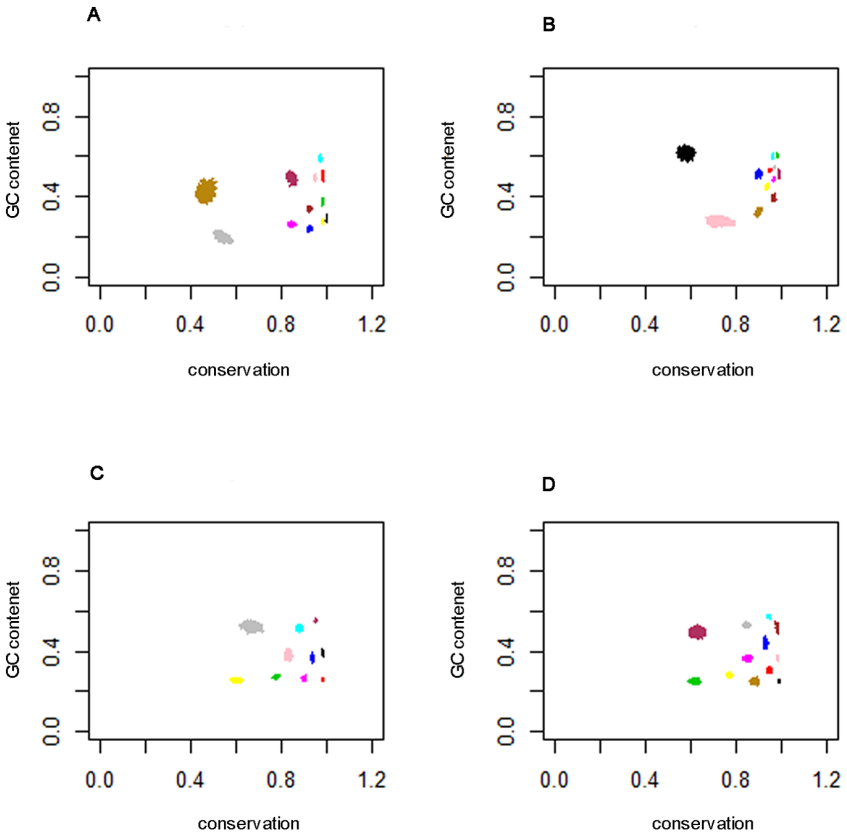


Figure 2. GC content versus conservation level for models selected by Procedure 1. GC content (in the first named species of each pair) versus the proportion of alignment matches, for each model selected by Procedure 1. The different colours represent different classes, and each class is plotted for the post burn-in samples; A) 12-class model for the *D. melanogaster* versus *D. simulans* 3′ UTR alignment, B) 12-class model for the *D. melanogaster* versus *D. simulans* first coding sequence (Coding 1) alignment, C) 10-class model for the *D. melanogaster* versus *D. yakuba* 3′ UTR alignment and D) 12-class model for the *D. simulans* versus *D. yakuba* 3′ UTR alignment. This is a crude diagnostic used to determine if the model has converged in distribution and also indicates how well separated the classes are.
doi:10.1371/journal.pone.0097336.g002

correspondence with classes detected in the other. Figure 3A, corresponding to the alignment of *D. melanogaster* versus *D. simulans* also shows the same pattern, but corresponding classes appear compressed towards the right of the figure relative to their counterparts in Figures 3C and 3D. This is no doubt due to the shorter evolutionary distance between *D. melanogaster* and *D. simulans*, leading to generally higher conservation levels in most classes. By contrast, the classes shown in Figure 3B, representing the coding sequences alignment, exhibit a pattern distinct from the other three, and it does not appear possible to identify class correspondences.

Further evidence of consistency among the three 2-way 3′ UTR alignments is shown in Table 3. Based on mixture proportions, conservation levels, GC content and transition/transversion ratios, twelve classes were directly comparable among the three 2-way alignments (although the correspondence is more convincing in some cases than in others). There were four cases in which classes were comparable in only two of three alignments, and there were only two cases in which a class was unable to be matched with a class from another alignment. The correspondence between classes identified for different alignments is even more clear when individual character frequencies are compared (Table S5). We also compared the significant motifs detected in the *D. melanogaster* versus *D. simulans* classes (Table S3) to those detected in the *D. melanogaster* versus *D. yakuba* alignment (Table S4). In most cases, classes that correspond in Table 3 were found to contain the same or similar characteristic tandem repeat sequences (Table S5).

The pattern shown in the plot of GC content versus conservation for the 3-way alignment (Figure S1), upon visual inspection, does not display an obvious similarity to the 2-way alignment plots. However, all but two of the classes can be mapped to classes from the 2-way alignments by considering the frequency of the individual characters within the segment classes (Table S6). While the encodings used for 2-way and 3-way alignments are different, a conserved A or T is represented by the character 'a' in both encodings, and a conserved G or C is represented respectively by the characters 'f' and 'v' in the 2-way and 3-way alignments; thus these characters were used in the comparison of the classes between 2-way and 3-way alignments.

Exploring Class Content

That such a large number of clearly distinguishable segments and segment classes can be identified in the 3′ UTRs of Drosophila genes is indicative of a surprisingly intricate compositional and mutational complexity. We hypothesize that this complexity results from a wide variety of structural and functional constraints, and we speculate about some of these constraints in this section. We focus on classes from the 15-class model of the *D. melanogaster* versus *D. simulans* 3′ UTR alignment that contain characteristic tandem repeat sequences identified by MEME as

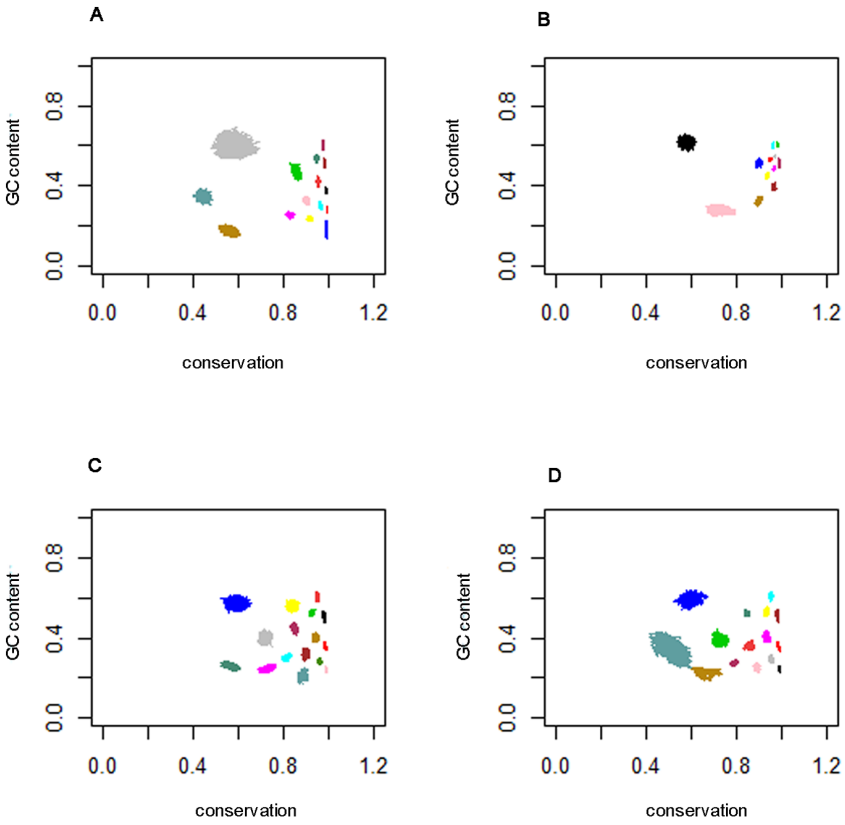


Figure 3. GC content versus conservation level for models selected by Procedure 2. GC content (in the first named species of each pair) versus the proportion of alignment matches, for each model selected by Procedure 2. The different colours represent different classes, and each class is plotted for the post burn-in samples; A) 15-class model for the *D. melanogaster* versus *D. simulans* 3′ UTR alignment, B) 12-class model for the *D. melanogaster* versus *D. simulans* first coding sequence (Coding 1) alignment, C) 16-class model for the *D. melanogaster* versus *D. yakuba* 3′ UTR alignment and D) 15-class model for the *D. simulans* versus *D. yakuba* 3′ UTR alignment.
doi:10.1371/journal.pone.0097336.g003

highly significant, and which are enriched in elements from the UTRdb, and PicTar annotation databases (see Methods section 'Class profiling').

One important concern regarding repetitive motifs is to ensure that they are not in some way artifacts of sequence composition. To test this, we artificially generated 100 control classes for each class from the 15-class segmentation of the *D. melanogaster* versus *D.*

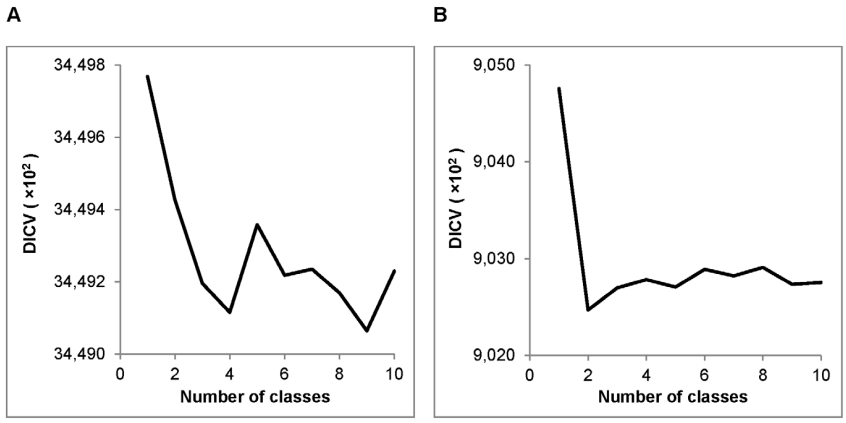


Figure 4. DICV values for segmentation of binary sequences. DICV values versus the number of classes (1–10) for segmentation of: A) the binary representation of GC content in *D. melanogaster* 3′ UTRs, and B) the binary representation of conservation in the *D. melanogaster* versus *D. simulans* 3′ UTR alignment.
doi:10.1371/journal.pone.0097336.g004

Table 3. Model comparisons.

Alignment	Class	MP	Conservation	GC content	T/T
Dme vs Dsi	0	15.9%	99%	38%	1.18
Dme vs Dya	1	11.8%	98%	36%	0.96
Dsi vs Dya	1	13.5%	98%	37%	0.97
Dme vs Dsi	1	14.3%	99%	28%	0.80
Dme vs Dya	15	13.8%	96%	28%	0.82
Dsi vs Dya	7	13.6%	95%	29%	0.88
Dme vs Dsi	2	2.0%	86%	47%	0.94
Dme vs Dyaa	7	2.0%	72%	40%	1.03
Dsi vs Dya	2	2.3%	72%	39%	1.00
Dme vs Dsi	3	2.3%	99%	18%	0.50
Dme vs Dya	8	8.5%	99%	24%	0.95
Dsi vs Dya	0	11.0%	99%	25%	0.81
Dme vs Dsi	4	17.1%	96%	30%	0.91
Dme vs Dya	4	7.5%	81%	30%	0.88
Dme vs Dsi	5	2.9%	83%	25%	0.73
Dme vs Dya	13	1.6%	58%	26%	0.71
Dsi vs Dya	11	1.6%	65%	24%	0.71
Dme vs Dsi	6	7.7%	92%	24%	0.67
Dme vs Dya	14	3.9%	89%	22%	0.67
Dsi vs Dya	8	6.9%	89%	25%	0.73
Dme vs Dsi	7	0.3%	58%	60%	0.91
Dme vs Dya	3	0.8%	60%	57%	0.78
Dsi vs Dya	3	0.7%	60%	59%	0.87
Dme vs Dsi	8	8.0%	90%	33%	0.98
Dme vs Dya	10	11.1%	90%	32%	0.92
Dsi vs Dya	12	9.5%	86%	36%	1.03
Dme vs Dsi	9	3.0%	97%	60%	1.48
Dme vs Dya	12	2.3%	95%	60%	1.30
Dsi vs Dya	4	2.2%	95%	61%	1.24
Dme vs Dsi	10	8.2%	98%	51%	1.45
Dme vs Dya	0	4.1%	98%	51%	1.34
Dsi vs Dya	10	4.3%	98%	52%	1.24
Dme vs Dsi	12	11.0%	95%	42%	1.07
Dme vs Dya	11	11.7%	94%	40%	1.11
Dsi vs Dya	5	12.8%	93%	41%	1.08
Dme vs Dsi	13	5.9%	95%	54%	1.32
Dme vs Dya	2	7.9%	93%	53%	1.33
Dsi vs Dya	6	7.8%	93%	53%	1.35
Dme vs Dsi	14	0.7%	44%	34%	0.70
Dsi vs Dya	14	0.5%	52%	34%	0.83
Dme vs Dya	5	3.2%	74%	25%	0.75
Dsi vs Dya	9	6.4%	78%	27%	0.81
Dme vs Dya	6	2.5%	84%	56%	0.95
Dsi vs Dya	13	6.8%	85%	52%	1.06

Comparison of the three models selected by Procedure 2, for each pairwise alignment of 3′ UTRs. MP: mixture proportions; T/T: Transition/Transversion ratio. Class 11 of Dme vs Dsi (MP: 0.7%, Conservation: 56%, GC content: 17% and T/T: 0.5) and the class 9 of Dme vs Dya (MP: 7.5%, Conservation: 85%, GC content: 45% and T/T: 1.1) alignments did not match with other models.

doi:10.1371/journal.pone.0097336.t003

simulans alignment which had significant motifs detected (Classes 0, 1, 3, 7, 9, 10, 12, 13; 800 in total). Each control class was generated independently such that the number and lengths of the segments corresponded exactly with one of the observed classes, and such that the frequency of bases was the same as observed in that corresponding class. Each of the control classes was run through MEME. No significant motifs were detected in any of these 800 control classes.

Class 1 had the equal highest proportion of conserved bases (~99%) and a relatively low GC content (~28%). MEME identified two motifs within Class 1 segments: an AT repeat motif common to 171 of 1491 Class 1 segments (E-value: 4.00E-36), and a polyA motif common to 136 segments (E-value: 3.70E-43, Figure 5A). The polyA motif consensus sequence matched the Polyadenylation Signal (PAS, UTRsite motif: U0043), according to the software UTRscan: a program for identifying known UTR regulatory motifs within a given sequence [16]. Class 1 segments were also found to be enriched in the PAS annotation in the UTRdb database (observed: 866, expected: 360, associated p-value: negligible). Given that poladenylation of the 3′ end of mRNAs is near ubiquitous in eukaryotes, it is perhaps unsurprising that our segmentation of 3′UTRs, based on sequence composition and conservation, identified a class of segments enriched in PASs. Cytoplasmic polyadenylation can occur for mRNAs which have been tranlastionally repressed, for example maternally inherited mRNAs which are activated on fertilization [14]. Class 6 segments were found to be enriched in the Cytoplasmic Polyadenylation Element (CPE, UTRsite motif: U0006; observed: 9, expected: 4, associated p-value: 3.26E-9). The median length of 3′ UTRs which contained Class 6 segments was 262 bases (IQR = 480), the shortest of all 15 segment classes, this is perhaps indicative of the inverse relationship between 3′ UTR length and mRNA stability, given that mRNAs requiring cytoplasmic polyadenylation are also required to be stable [14].

Along with Class 1, Class 0 also had the equal highest proportion of conserved bases (~99%), differing on GC content (~38%). A CAA tri-nucletide repeat motif was identified in Class 0 segments (E-value: 3.0E-34). Both Class 0 and 1 were found to be enriched in multiple miRNA targets, as predicted by PicTar [32]. miRNA targets represent a class of elements found in 3′ UTRs which are important in gene regulation, miRNAs (in cooperation with a protein complex) bind 6–8 mer sites in mRNAs promoting the degradation of the bound mRNA [33] PicTar predictions are partly based on sequence conservation so it is somewhat unsurprising that there is significant overlap between our highly conserved segments classes and PicTar predictions.

Class 9 had the equal highest GC content of the classes (~60%), a relatively high proportion of conserved bases (~97%), the longest segments (median = 142 bases, IQR = 138), the highest transition/transversion ratio (1.48) and a bias towards the coding end of 3′ UTRs, with a median distance to the coding sequence of 21.5 bases (IQR = 240). Class 10 was notable for a relatively high GC content (~51%), relatively high conservation (~98%), and a relatively high transition/transversion ratio (1.45). Relatively high GC content, high conservation and positional bias are all independently indicative of enrichment in functional elements. MEME identified a CAG tri-nucleotide repeat motif (Figure 5B) in both segment classes, common to 124 of the 298 Class 9 segments and 114 of the 1023 Class 10 segments (E-values, respectively: 5.30E-138, 1.60E-21). TOMTOM identified matches in both the "All vertebrates" and the "All Drosophila" database for both motifs. In the "All Drosophila" database, both CAG repeat motifs matched the binding site of odd, a Drosophila zinc-finger protein. The CAG-repeat motif resembles a repeated E-box: a basic helix-loop-helix (bHLH) binding site with consensus sequence (CANNTG). The matches in the "All Vertebrates" database were both to proteins with bHLH DNA-bonding domains; the Class 9 motif matched the mouse Ascl2 primary binding site (E-value: 2.17E-5), and the Class 10 motif matched the mouse Tcf12 binding site (E-value: 2.47E-5). bHLH protein structures are common to DNA binding proteins involved in transcriptional regulation in all eukaryotes [34]. In Drosophila, twist, acheate-

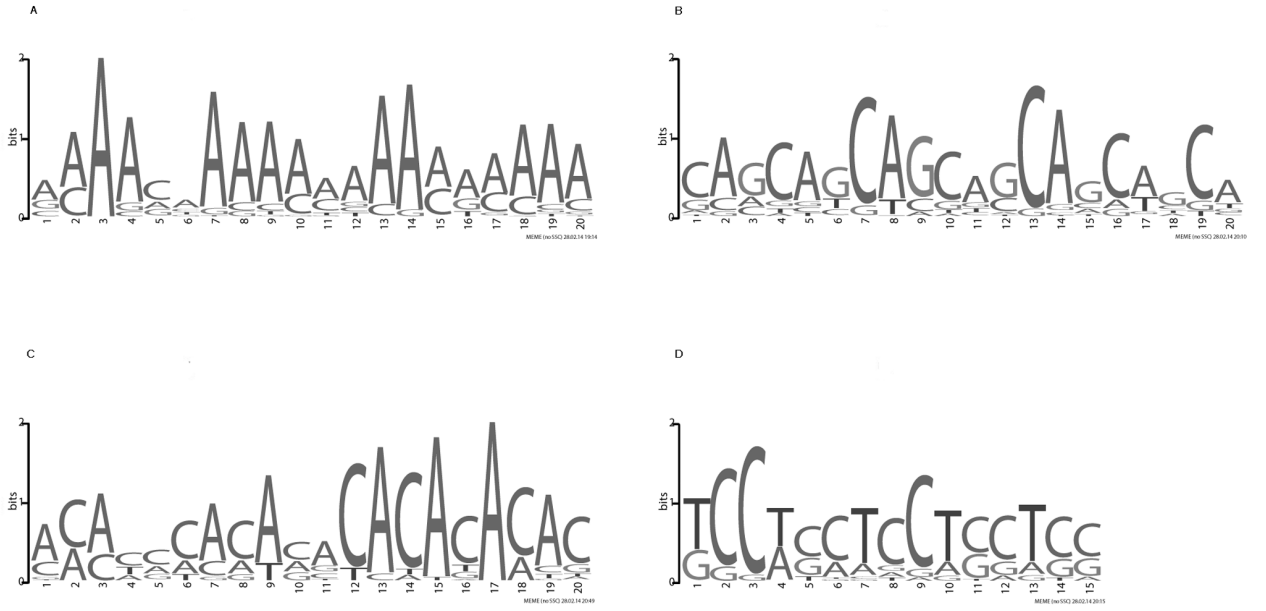


Figure 5. Motifs identified by MEME. Sequence LOGOs for four of the motifs identified by MEME in the 15-class model for the *D. melanogaster* versus *D. simulans* 3′ UTR alignment: A) a polyA motif identified in Class 1, B) a CAG repeat motif identified in Class 9, C) a CA repeat motif identified in Class 12, D) a TCC repeat motif identified in Class 9.
doi:10.1371/journal.pone.0097336.g005

scute, D-mef2 and daughterless are examples of bHLH proteins with well documented regulatory roles that bind E-Box like regulatory elements in order to regulate target gene expression [35,36]. Furthermore, there are at least 56 known genes in Drosophila coding for proteins with the bHLH DNA binding domain [37].

A CA di-nucletide repeat motif was identified in Class 12, common to 35 of 849 segments (E-value: 3.80E-12, Figure 5C). A possible function for such sites is the documented CA-rich elements (CAREs) which are known to interact with heterogenous nuclear ribonucleoprotein L in order to stabilise mRNAs [14]. In addition, TOMTOM identified matches to three Drosophila zinc-finger protein binding sites in the "All Drosophila" database: klumpfuss, stripe and fruitless (E-values, respectively: 9.43E-3, 2.70E-2, 3.02E-2). TOMTOM also identified matches in the "All Vertebrates" database to the human zinc-finger protein RREB1 and the mouse zinc-finger protein EGR2 binding sites (E-values, respectively: 1.31E-2, 2.55E-2). While many of motifs identified by MEME have similarities with TFBSs, we note that regulatory elements in 3′ UTRs are primarily thought to operate post-transcriptionally and hence to interact with proteins (and miRNAs) that bind RNA, not DNA. The CA-dinucleotide repeat motif was one of two motifs from the 15 class segmentation of the *D. melanogaster* versus *D. simulans* 3′UTR alignment in which TOMTOM identified matches in the "RNA-binding motifs" database. (Recognising a deficiency in knowledge of RNA-binding motifs, the "RNA-binding motif" database was generated by a large-scale experiment for determining binding motifs of known RNA-binding proteins [38]. Synthetic RNA molecules were generated for all possible sequences of length 7, 8 and 9 nucleotides, binding affinity to each motif was measured for 193 unique RNA-binding proteins - 141 with no previously known motif - including 61 from Drosophila.) RNA-binding proteins are known to play a crucial role in gene expression, including roles in splicing, polyadenylation and controlling mRNA stability. One of the most well characterised RNA-binding proteins is the Drosophila Sxl, well known for its role in the complex Drosophila sex determination mechanism [39]. Classes 4, 5, and 6 were enriched in the Sxl binding motif (Table S8). The CA repeat motif matched eleven different RNA-binding motifs in the database, five of which were for Drosophila proteins. Thus it has been shown there are Drosophila proteins which will bind the sequences generating the CA repeat motif. The second motif with a match in the "RNA-binding motif" database is a TCC tri-nucleotide repeat motif, common to 96 of 298 Class 9 segments (E-value: 3.70E-5, Figure 5D). The TCC repeat motif matched the binding site of the human RNA-binding protein SRSF1, a splicing factor.

The positions of segments from each segment class for the segmentation models chosen by Procedure 2 are available in BED format as part of supplementary materials (File S1, S2, S3). A full summary of the motifs identified can be found in Tables S2, S3 and S4, and a full summary of the enrichment of PicTar and UTRdb annotations can be found in Tables S7 and S8. As discussed, several of these repetitive motifs resemble binding sites of common regulatory proteins. While it is possible that TFBSs located within 3′ UTRs could act as enhancer elements [40], in general 3′ UTRs are not considered to play a significant role in transcription activation. It is more likely that these motifs participate in post-transcriptional regulatory interactions with RNA-binding proteins and miRNAs. However, we note in passing that many zinc-finger proteins are capable of binding RNA in addition to DNA, and transcription factors that bind both DNA and mRNAs are known (for example [41]).

Conclusions

A pairwise alignment can be encoded as an 8-character sequence containing information about sequence conservation, GC content and transition/transversion ratios. A similar approach can be used to encode a three-way alignment as a 32-character sequence. Such sequences can then be segmented and the segments classified according to character frequencies. Here and elsewhere [31] we have shown that DICV provides a method for selecting the number of classes that is conservative in the sense that it does not generally favour models with superfluous classes. We have also proposed a second, less conservative, model selection procedure. Using these encodings, it is possible to distinguish segment classes that could not be resolved on the basis of sequence similarity or GC content considered in isolation. We have therefore proposed the method as suitable for analysing pairwise alignments of closely related species.

An unexpectedly large number of clearly distinguishable segment classes were identified in pairwise and three-way alignments of 3′ UTRs for the species *D. melanogaster*, *D. simulans*, and *D. yakuba*. The number of classes found is comparable to and possibly exceeds the number identified in equal length alignments of protein-coding sequences. The estimated number of change-points in 3′ UTRs exceeds the corresponding estimate for protein-coding sequences by a factor of five. This is suggestive of intricate functional complexity in Drosophila 3′ UTRs, far exceeding that of protein-coding sequences. Similar classes were identified in all three pairwise alignments, suggesting similar constraints are maintained in all three species.

Several of the segment classes we identified were highly enriched in low information content sequences. Although care must be taken to ensure that such motifs are not artifactual, we have used rigorous controls to demonstrate that is not the case here. Moreover, many of the known regulatory sequences in 3′ UTRs have precisely this low information character. We speculate that such regulatory sequences may be frequently created and destroyed in 3′ UTRs, resulting in rapid turnover of functional elements, individual variation in regulatory profiles, and ephemeral conservation. We further speculate that some extended low information content regions of 3′ UTRs may be functional only in the sense that they regularly produce and lose binding sites, thus facilitating changes in regulatory profiles in response to changing selective pressures. A full elucidation of functional elements in 3′ UTRs may therefore require comparisons of closely related species, as well as examination of non-comparative indicators of function.

Materials and Methods

Data Transformation

A three-way multiple sequence alignment (MSA) of *D. melanogaster*, *D. simulans* and *D. yakuba* genes was obtained from http://genomics.princeton.edu/AndolfattoLab/w501_genome.html (see also (Hu *et al.* 2013)). The data is made available by the Andolfatto Lab, and incorporates a second generation assembly of the *D. simulans* genome performed in 2012. Annotations of the *D. melanogaster* genome are also provided, and were used to separate the alignments into genic sections, in particular coding regions and 3′ UTRs. The three-way MSA was analysed as three pairwise sequence alignments of *D. melanogaster* to *D. simulans*, *D. melanogaster* to *D. yakuba*, and *D. simulans* to *D. yakuba*.

We used an 8-character sequence representation ($A = \{a,b,c,d,e,f,g,h\}$) of the pairwise alignments, in which each character in the sequence corresponds to a non-directional mono-nucleotide alignment combination:

Species 1: ATATATATCGCGCGCG
Species 2: ATCGGCTAATCGGCTA
Symbol: a a b b c c d d e e f f g g h h.

Insertions and deletions relative to *D. melanogaster* are excluded from the representation of the alignment.

For each of the three pairwise alignments, the 8-character sequences for the 3′ UTRs of each gene on chromosome arms 2R, 2L, 3R, 3L were concatenated into a single sequence. Each 3′ UTR segment was separated from the next by a # symbol. The *D. melanogaster* versus *D. simulans* alignment of protein-coding sequences was constructed in a similar manner, with each exon separated by a # symbol. Three randomly selected subsequences were then selected, each the same length as the *D. melanogaster* versus *D. simulans* 3′ UTR sequence. This was done by choosing a uniform random starting position and then an end position such that that the lengths were the same.

The 3-way alignment of *D. melanogaster*, *D. simulans* and *D. yakuba* was converted to a 32-character sequence representation ($B = \{a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u,v,w,x,y,z,U,V,W,X,Y,Z\}$).

Species 1: AAAAAAAAAAAAAAAACCCCCCCCCCCCCCCC
Species 2: AAAACCCCGGGGTTTTAAAACCCCGGGGTTTT
Species 3: ACGTACGTACGTACGTACGTACGTACGTACGT
Symbol: a b c d e f g h i j k l m n o p q r s t u v w x y z U V W X Y Z

The alignment columns with complementary bases were encoded using the same characters. For example:

Species 1 'A', Species 2 'A', Species 3 'A' = Species 1 'T', Species 2 'T', Species 3 'T' = 'a'

Species 1 'A', Species 2 'A', Species 3 'C' = Species 1 'T', Species 2 'T', Species 3 'G' = 'b'.

Two binary sequence representations were also constructed: a binary representation of the GC content in *D. melanogaster* 3′ UTRs (1 for 'G' or 'C', and 0 for 'A' or 'T') and a binary representation of conservation in the *D. melanogaster* versus *D. simulans* 3′ UTR alignment (1 for a match, 0 for a mismatch). Both binary sequences involved concatenation in a similar manner as for the 8-character sequences. Note that the binary representations can be recovered from the 8-char representation of the *D. melanogaster* versus *D. simulans* 3′ UTR alignment (as discussed under the heading 'Assessing Convergence' below).

Change-point Modeling

We constructed a Bayesian multiple change point model for the sequences described above. The model is described in detail for binary sequences in previous papers [13,24,25] and for larger alphabets in [26,31]. In summary, this approach estimates positions in the sequence that delineate homogenous segments (known as change-points), the number of which is unknown. The # symbol is considered as a fixed change-point. Each segment is drawn from a multinomial distribution with parameters drawn from one of T Dirichlet distributions with uniformly sampled probabilities. As the number of classes T is not known *a priori*, independent runs with values of T from 1 to 20 were performed. We used an efficient varying-dimensional MCMC technique for simulating from the posterior distribution for the number of change-points, k, and segment parameters for different numbers of classes. Each model was run for 20,000 iterations and then tested for convergence.

To test our model selection procedures, we also constructed an 8-character control sequence. The sequence was generated such that it was the same length as the *D. melanogaster* versus *D. simulans* 3′ UTR alignment, with fixed change-points in the same positions. Each segment had parameter $\theta = (\theta_a,\theta_b,\theta_c,\theta_d,\theta_e,\theta_e,\theta_g,\theta_h)$ drawn from the same Dirichlet distribution ($T = 1$), based on the character frequencies of the *D. melanogaster* versus *D. simulans* 3′ UTR alignment.

Assessing Convergence

To assess convergence of the MCMC sampler in 8-character sequence representation, the mean proportion of no mutations (alignment matches: represented by input symbols 'a' and 'f') was calculated for each iteration of the sampler:

$$E[\theta_{cons}] = \frac{\theta_a + \theta_f}{\sum_{j \in A} \theta_j}$$

This was plotted against the GC proportions (represented by characters 'e', 'f', 'g' and 'h'), again calculated for each iteration of the sampler:

$$E[\theta_{GC}] = \frac{\theta_e + \theta_f + \theta_g + \theta_h}{\sum_{j \in A} \theta_j}$$

Such plots show a striking trend during the 'burn-in' phase of MCMC, at the end of which is a dense 'blob' indicating that convergence has occurred. Figures 2 and 3 are examples of such plots, but show only the post-burn-in phase.

For 32-character representation, similar information is given by symbols 'a' and 'v' for alignment matches and by symbols 'q', 'r', 's', 't', 'u', 'v', 'w', 'x', 'y', 'z', 'U', 'V', 'W', 'X', 'Y' and 'Z' for GC proportion in species 1 (Figure S2).

Model Selection

Our current segmentation model assumes that the number of classes (T) is known; in reality this is not the case. We used two procedures to select the number of classes, after fitting the model for a range of T. In both procedures, a model containing classes considered to be empty (low mixture proportions) was considered to be an over-fitted model and thus a model with a fewer number of classes would be selected in which the main criterion was still fulfilled (see [42] for a discussion of this approach to model selection).

Procedure 1: Investigating DICV. Deviance Information Criterion (DIC) is a criterion for model selection related to the better known Akaike Information Criterion (AIC) and Bayesian (or Schwarz) Information Criterion (BIC). Here we use type V DIC, which we investigate as a model selection criterion for sequence segmentation in [31]. DICV is defined:

$$DICV = Pv + \overline{D(\theta)}$$

where $\overline{D(\theta)} = -2\times$ average of log-likelihood over the set of segmentations sampled by MCMC and $Pv = 1/2\times$ variance of log-likelihood over the set of samples.

Models with smaller DICV are preferred; however, it often happens that there is no clear minimum. In general we select the value of T which corresponds to the first local minimum of DICV. However, a subjective judgement is used when it appears obvious that the DICV values continue to decrease significantly with larger values of T. For a detailed discussion of using information criterion to select the number of classes, see [31].

Procedure 2: Investigating the stability of classes. In this procedure the model selected was the model with the largest

number of classes in which each class was considered stable. Stability of classes was assessed based on time-series plots of conservation levels and GC content versus sample number. Classes which were highly variable in either GC content or conservation level were deemed unstable (again this involved a subjective judgement). As previously mentioned, the mixture proportions of the classes was used as a second criteria to assess the selected model, and a model with a smaller number of classes was selected if any of the classes were deemed empty.

Class Profiling

The positions of segments in each of the segment classes in each of models chosen by Procedure 2 were recorded in BED format (BED files submitted with supplementary material), with genomic coordinates relative to the *D. melanogaster* genome (release R5.33). The *D. melanogaster* sequence corresponding to each segment for each of the segment classes was also extracted in fasta format. We defined segments as belonging to a particular class as contiguous runs of at least eight sequence positions at which the posterior probability of belonging to the given class is >0.5. The use of the threshold (>0.5) is discussed in [13], and is demonstrated to be an effective compromise between false negative and false positive allocation of positions to segment classes.

MEME motif identification. MEME [43] was used to search for motifs shared by segments from the profiled classes. We allowed the option of zero or one motif per sequence in all queries, with a maximum motif size of 20 base pairs and for the reverse complement of each sequence to be considered. For each motif identified by MEME with an E-value <0.05, TOMTOM [44] (web interface: http://meme.nbcr.net/meme/cgi-bin/tomtom.cgi) was then used to search for similar motifs in each of four motif databases: "All Drosophila"; "JASPAR-insects"; "All Vertebrates"; "RNA-binding motifs" (descriptions of the motif databases are found at the web interface). Motifs reported by TOMTOM with an E-value <0.05 were considered significantly similar.

Annotation enrichment. Drosophila 3′ UTR annotations were obtained from UTRdb [16] and PicTar output [32], then segment classes were tested for enrichment in each of the annotation types. The Drosophila subset of the UTRdb dataset of annotations (UTRef) was obtained from http://ebi.edu.au/ftp/databases/UTR/data/. All Drosophila annotation in UTRef are based on pattern similarity identified using the tool UTRscan. PicTar is a program for predicting miRNA binding sites from multiple species alignments, sites predicted in Drosophila were obtained from http://dorina.mdc-berlin.de/rbp_browser/dm3.html.

The positions of annotations in *D. melanogaster* were compared with the positions of each of the segment classes of the 15-class model of the *D. melanogaster* versus *D. simulans* 3′ UTR alignment. For each annotation type we test whether there is evidence for enrichment of that annotation type in our segment classes. For the null hypothesis of no enrichment, the expected number of annotations in each segment class is based on the proportion of the *D. melanogaster* sequence covered by each segment class. The bagFFT algorithm [45] (web interface: http://www.cs.cornell.edu/w8/~niranjan/llr.html) was used to calculate p-values for an exact multinomial goodness-of-fit test. Annotation types with p-value <0.05, after Bonferroni correction for multiple testing, are considered significant. Only annotation types with more than one match in the segment classes are considered for testing. For annotation types with significant p-values, classes containing more occurrences of that type than expected are considered enriched in that element.

Supporting Information

Figure S1 DICV values for segmentation of 3-way alignment. DICV values obtained using 1–20 segment classes for *D. melanogaster*, *D. simulans* and *D. yakuba* 3′ UTR alignment. The 14-class model was selected as minimum DICV has occurred at class 14.
(TIFF)

Figure S2 GC content versus conservation level for models selected for 3-way alignment. GC content of *D. melanogaster* versus the proportion of alignment matches, for each model selected for the 3-way 3′ UTR alignment. A) 14-class model selected by Procedure 1 and B) 15-class model selected by Procedure 2. The different colours represent different classes, and each class is plotted for the post burn-in samples. This plot was used to access the convergence of the selected models.
(TIF)

Figure S3 DICV values for the control sequence. DICV values were obtained for an artificially generated sequence having only one class of segments. The minimum DICV has occurred at 1-class; therefore justifies models selected by Procedure 1.
(TIFF)

Figure S4 Conservation level vs sample number for control sequences. Figure shows time-series plots of conservation level versus sample number for segmentations of artificially generated control sequence with A) 1 segment class and B) 2 segment classes.
(TIF)

Table S1 Model comparisons - Procedure 1 versus Procedure 2. Comparing characteristics of the two models selected by Procedure 1 and Procedure 2 (12-class model and 15-class model respectively) for 3′ UTR alignment of *D. melanogaster* versus *D. simulans*.
(XLSX)

Table S2 Types of motif identified in 12-class model of D. melanogaster vs D. simulans alignment. Types of motif identified in *D. melanogaster* versus *D. simulans* 12-class model selected by Procedure 1.
(XLSX)

Table S3 Types of motif identified in 15-class model of D. melanogaster versus D. simulans alignment. Types of motif identified in *D. melanogaster* versus *D. simulans* 15-class model selected by Procedure 2.
(XLSX)

Table S4 Types of motif identified in 16-class model of D. melanogaster versus D. yakuba alignment. Types of motif identified in *D. melanogaster* versus *D. yakuba* 16-class model selected by Procedure 2.
(XLSX)

Table S5 Class comparisons of 3′ UTR pairwise alignments. Comparison of change-point character frequencies in each of the classes identified by Procedure 2 for each pairwise alignment of *D. melanogaster* (D. mel), *D. simulans* (D. sim), and *D. yakuba* (D. yak) 3′ UTRs. Classes from different models with similar character frequencies are grouped together.
(XLSX)

Table S6 Class comparisons of 3′ UTR pairwise and 3-way alignments.
(XLSX)

Table S7 Enrichment of PicTar miRNA targets in segment classes.
(XLSX)

Table S8 Enrichment of UTRdb motifs in segment classes.
(XLSX)

File S1 Positions of segments for the 15-class model of D. melanogaster versus D. simulans alignment.
(BED)

File S2 Positions of segments for the 16-class model of D. melanogaster versus D. yakuba alignment.
(BED)

File S3 Positions of segments for the 15-class model of 3-way D. melanogaster, D. simulans, D. yakuba alignment.
(BED)

Acknowledgments

The authors would like to thank Professor Peter Adams for numerous helpful discussions.

Author Contributions

Conceived and designed the experiments: JMK. Performed the experiments: MA CO ET. Analyzed the data: MA CO ET. Contributed reagents/materials/analysis tools: JMK. Wrote the paper: MA CO ET KM JMK.

References

1. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril J, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.
2. Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, et al. (2007) Genome of the marsupial monodelphis domestica reveals innovation in non-coding sequences. Nature 447: 167–177.
3. Mattick JS (2005) The functional genomics of noncoding RNA. Science 309: 1527–1528.
4. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakkerand PIW, et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316: 1331–1336.
5. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, et al. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316: 1341–1345.
6. Zeggini E,Weedon MN, Lindgren CM, Frayling TM, Elliott KS, et al. (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316: 1336–1341.
7. Dunham I (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.
8. Graur G, Zheng Y, Price N, Azevedo RBR, Zufall RA, et al. (2013) On the immortality of television sets: ''function'' in the human genome according to the evolution-free gospel of ENCODE. Genome Biology and Evolution 5: 578–590.
9. Doolittle WF (2013) Is junk DNA bunk? a critique of ENCODE. Proceedings of the National Academy of Sciences of the USA 110: 5294–5300.
10. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, et al. (2012) Targetted RNA sequencing reveals the deep complexity of the human transcriptome. Nature Biotechnology 30: 99–104.
11. Kuersten S, Goodwin EB (2003) The power of the 3′ UTR: translational control and development. Nature Reviews Genetics 4: 626–637.
12. Chatterjee S, Pal JK (2009) Role of 5′- and 3′-untranslated regions of mRNAs in human diseases. Biology of the Cell 101: 251–262.
13. Keith JM, Adams P, Stephen S, Mattick JS (2008) Delineating slowly and rapidly evolving fractions of the drosophila genome. Journal of Computational Biology 15: 407–430.
14. Matoulkova E, Michalova E, Vojtesek B, Hrstka R (2012) The role of the 3′ untranslated region in post-transcriptional regulation of protein expression in mammalian cells. RNA Biology 9: 563–576.
15. Andken BB, Lim I, Benson G, Vincent JJ, Ferencand MT, et al. (2007) 3′-UTR SIRF: a database for identifying clusters of whort interspersed repeats in 30 untranslated regions. BMC Bioinformatics 8: 274.
16. Grillo G, Turi A, Licciulli F, Mignone F, Liuni S, et al. (2010) UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Research 38: D75–80.
17. Ahmed F, Benedito VA, Zhao PX (2011) Mining functional elements in messenger RNAs: overview, challenges, and perspectives. Frontiers in Plant Science 2: 84.
18. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras T, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816.
19. Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, et al. (2006) Large-scale turnover of functional transcription factor binding sites in drosophila. PLoS Computational Biology 2: 1219–1231.
20. Burger A, Walczak AM, Wolynes PG (2010) Abduction and asylum in the lives of transcription factors. Proceedings of the National Academy of Sciences of the United States of America 107: 4016–4021.
21. Lee TH, Maheshri N (2012) A regulatory role for repeated decoy transcription factor binding sites in target gene expression. Molecular Systems Biology 8: 1–11.
22. Braun JV, Muller HG (1998) Statistical methods for DNA sequence segmentation. Statistical Science 13: 142–162.
23. Liu JS, Lawrence CE (1999) Bayesian inference on biopolymer models. Bioinformatics 15: 38–52.
24. Keith JM, Kroese DP, Bryant D (2004) A generalized markov sampler. Methodology and Com-puting in Applied Probability 6: 29–53.
25. Keith JM (2006) Segmenting eukaryotic genomes with the generalized gibbs sampler. Journal of Computational Biology 13: 1369–1383.
26. Oldmeadow C, Mengersen K, Mattick JS, Keith JM (2010) Multiple evolutionary rate classes in animal genome evolution. Molecular Biology and Evolution 27: 942–953.
27. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803–819.
28. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15: 1034–1050.
29. Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, et al. (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Research 17: 760–774.
30. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research 20: 110–121.
31. Oldmeadow C, Keith JM (2011) Model selection in bayesian segmentation of multiple dna alignments. Bioinformatics 27: 604–610.
32. Grun D, Wang Y, Langenberger D, Gunsalus KC, Rajewsky N (2005) microRNA target predictions across seven drosophila species and comparison to mammalian targets. PLOS Computational Biology 1: e13.
33. Bartel DP (2009) MicroRNAs: Target recognition and regulatory functions. Cell 136: 215–233.
34. Massari ME, Murre C (2000) Helix-loop-helix proteins: Regulators of transcription in eucaryotic organisms. Molecular and Cellular Biology 20: 429–440.
35. Molkentin JD, Olson EN (1996) Combinatorial control of muscle development by basic helix-loop-helix and mads-box transcription factors. Proceedings of the National Academy of Sciences of the United States of America 93: 9366–9373.
36. Murre C, McCaw PS, Baltimore D (1989) A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. Cell 56: 777–783.
37. Ledent V, Vervoot M (2001) The basic helix-loop-helix protein family: Comparative genomics and phylogenetic analysis. Genome Research 11: 754–770.
38. Ray D, Kazan H, Cook KB, et al. (2013) A compendium of RNA-binding motifs for decoding gene regulation. Nature 499: 172–177.
39. Penalva LOF, Sanchez L (2003) RNA binding protein sex-lethal (sxl) and control of drosophila sex determination and dosage compensation. Microbiology and Molecular Biology Reviews 67: 343–359.
40. Splinter E, de Laat W (2011) The complex transcription regulatory landscape of our genome: control in three dimensions. The EMBO Journal 30: 4345–4355.
41. Morrison AA, Viney RL, Ladomery MR (2008) The post-transcriptional roles of wt1, a multifunctional zinc-finger protein. Biochimica et Biophysica Acta 1785: 55–62.
42. Rousseau J, Mengersen K (2011) Asymptotic behaviour of the posterior distribution in overfitted mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73: 689–710.
43. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Research 34: W369–W373.
44. Gupta S, Stamatoyannopolous JA, Bailey T, Noble WS (2007) Quantifying similarity between motifs. Genome Biology 8: R24.
45. Keich U, Nagarajan N (2006) A fast and numerically robust method for exact multinomial goodness-of-fit test. Journal of Computational and Graphical Statistics 15: 779–802.

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 4

Declaration by candidate

In the case of Chapter 4, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Conceived methods, scripting, designed and performed computational experiments, analysed data, wrote the paper, made modifications to the manuscript as suggested by co-authors	90

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Edward Tasker	Assisted in pathway-focussed analysis	
Caitlin Williams	Designed and performed laboratory experiments, wrote the section for experimental validation	
Adam Parslow	Designed and performed laboratory experiments	
Robert Bryson-Richardson	Conceived the idea, designed laboratory experiments, provided guidance in interpretation and writing	
Jonathan Keith	Conceived the idea, contributed to analysis tools, provided guidance in interpretation and writing	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature	[redacted]	**Date** 9/10/15
Main Supervisor's Signature	[redacted]	**Date** 9/10/15

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 4

Genome-wide Identification of ncRNAs using a Bayesian Segmentation Approach

Chapter Objectives

In this chapter, I continued to develop methods to identify ncRNAs. In particular, my aim was to build a systematic process to discover genome-wide non-coding putative functional elements (PFEs) using a Bayesian approach. To achieve this, I used the *changept* model and the methods developed in the previous chapter. This analysis revealed over 1000 intronic and intergenic PFEs, conserved in human, mouse and zebrafish. I compared these results with the findings of other methods which are readily available to use for this purpose (EvoFold and RNAz predictions, DNAse I footprints data, fRNAdb entries, and RNA-seq data). These results out-performed the results of EvoFold. We experimentally validated 26 PFEs identified in a set of genes involved in muscle development. These PFEs may play a role in muscle diseases.

Authorship

Manjula Algama[1], Edward Tasker[1], Caitlin Williams[2], Adam C. Parslow[2], Robert J. Bryson-Richardson[2], Jonathan M Keith[1]

1 School of Mathematical Sciences, Monash University, Clayton, VIC 3800, Australia

2 School of Biological Sciences, Monash University, Clayton, VIC 3800, Australia

Reference

Algama M, Tasker E, Williams C, Parslow AC, Bryson-Richardson RJ, Keith JM. (2015). Genome-wide Identification of ncRNAs using a Bayesian Segmentation Approach. In preparation. Target journal: *Genome Research*.

4.1 Abstract

Non-coding RNAs (ncRNAs) play crucial roles in a variety of biological processes. They are sometimes encoded within the introns of protein-coding genes, and may regulate and interact in networks involving the containing gene. In this study, we carried out a genome-wide analysis using a Bayesian segmentation-classification model to identify intronic elements highly conserved between three evolutionarily distant vertebrate species: human, mouse and zebrafish. These elements may include ncRNAs or domains of ncRNAs with crucial functions. They may also include other functional sequences including regulatory sequences. We identified 655 such intronic sequences, which we refer to as putative functional elements (PFEs). This included 97 regions not identified by EvoFold. As indicated by our analysis, there was a significant over-representation of transcription factors in the genes containing PFEs (p-value: 1.2e-56). We also performed a pathway-focussed analysis using a set of genes involved in muscle development. We detected 27 intronic PFEs in 7 transcription factors. The expression of 26 PFEs was experimentally validated using RT-PCR. This provides further evidence that PFEs are predominantly representatives of a class of ncRNA modulating expression or otherwise interacting with transcription factors. Our method extended the length of the predicted functional regions of EvoFold. This study demonstrates the success of our Bayesian approach in identifying putative ncRNAs and other regulatory elements using improved alignments.

4.2 Introduction

It has become evident that non-coding RNAs play a significant role in gene regulation. These elements have been implicated in a variety of biological functions including transcription [1], RNA splicing [2, 3], editing [4], translation [5] and chromatin modification [6–8]. Disruption of ncRNAs is associated with many diseases including cancer [9], leukaemia [10], diabetes [11, 12] and neurological disorders [13–16]. Further roles

of ncRNAs include regulation of differentiation and development (retinal and erythroid development, breast development, epidermal differentiation - [17–21]), regulation of epigenetic processes (X chromosome dosage compensation, parental imprinting in mammals, vernalization in plants - [22–31]), and RNA modification and evolution [32, 33].

Identification of conserved intronic elements is of interest due to their potential to contain regulatory sequences and non-coding RNAs. Many regulatory elements in introns are enriched in transcription factor binding sites [34–36] and evolutionary conservation has been identified as a property of functional transcription factor binding sites [37–39]. Non-coding RNAs are found in intergenic regions as well as in introns, but their transcription as part of an intron is a potential mechanism for regulatory and other interactions with the gene in which they occur [40, 41]. The ncRNAs can be mainly classified into two groups: (1) short ncRNAs (<200nt, such as ribosomal RNAs, transfer RNAs, small nucleolar RNAs, microRNAs, endogenous short interfering RNAs, PIWI-interacting RNAs); and (2) long ncRNAs (>200nt).

The lack of identifying features, such as those used to predict protein-coding genes, makes the identification of ncRNAs from sequence data alone very challenging. Current computational methods to identify ncRNAs frequently rely on formation of secondary structure of a potential ncRNA sequence (such as Mfold- [42], RNAfold- [43], RNAz- [44]), or combine this approach with comparative sequence analysis (such as EvoFold - [45]).

Commonly used basic analysis of DNA sequence conservation has two main disadvantages, both of which limit the results that can be obtained. When using an alignment of orthologous sequences to inspect conservation, *sliding window analysis* is often used. This technique involves counting the number of matches/mismatches in overlapping windows of a predetermined length, to obtain a profile of conservation level across the sequence. A smaller window allows for more precise localisation of changes in the property of interest; however a smaller window also allows for noise within the sequence

to more significantly affect the output. Thus sliding window analysis is inherently a compromise between these two factors [46]. This compromise fails to fully recognise the known discrete, modular nature of DNA functionality, for example the boundaries between exons and introns, the ends of transcription factor binding sites (TFBSs), and the transcription start sites of expressed RNAs. Thus this technique will not be able to accurately identify such positions in DNA sequences, and more sophisticated segmentation methods are required ([47], reviewed in [48]). The second disadvantage is the common consideration of conservation as a dichotomy (conserved or not-conserved). It is reasonable to expect that the constraint on any given region will have changed over the course of evolution, also that for different regions the constraints will have varied differently over time. A demonstration of this is found in [49], 7 evolutionary rate classes were identified within mammals and 9 evolutionary rate classes within drosophilids. Further discussion of the merits of rejecting the idea of a dichotomy of conservation levels can also be found in [49].

To overcome the above-mentioned disadvantages of conventional analysis of sequence alignments, we performed an analysis using a Bayesian segmentation model *changept* [50, 51]. Adopting a Bayesian approach is beneficial as it provides quantification of the uncertainties in parameter estimates in the form of probability distributions. The *changept* model can be described as a segmentation-classification model, which is capable of simultaneously segmenting a genomic alignment and classifying segments into one of a predefined number of segment classes. Segments were classified according to multiple sequence characteristics including level of evolutionary conservation between species, GC content and transition/transversion ratio.

Using *changept*, we carried out a genome-wide analysis using an automated alignment corresponding to all zebrafish chromosomes. We identified 655 intronic putative functional elements (PFEs) distributed among 193 zebrafish genes. To determine if PFEs correspond to ncRNAs or other regulatory elements, the locations of PFEs were compared with the findings of other methods (EvoFold, RNAz, DNase I footprints

regions and fRNAdb entries). PFEs were highly enriched in transcription factors. To examine if there were conserved elements between different members of the same pathway, we performed a pathway-focussed analysis on 24 genes involved in muscle development (myogenesis). Understanding the genetic and biochemical processes that contribute to myogenesis is an important step in developing treatments for muscle diseases. The interaction of protein-coding genes in the regulation of myogenesis has been well characterised (reviewed in [52]). Many microRNAs are known to down-regulate target gene expression by repressing the translation of protein-coding mRNAs involved in myogenesis [53–55]; in addition lncRNAs have been demonstrated to up-regulate gene expression by providing a secondary target for microRNAs to bind [56]. Using our segmentation method, we identified 27 PFEs with clearly defined boundaries that belong to the class of most highly conserved segments. All PFEs were detected in transcription factors, consistent with the results of the genome-wide analysis. We validated our findings experimentally, confirming the expression of PFEs in zebrafish embryos.

4.3 Results

To identify putative functional non-coding elements conserved between human, mouse and zebrafish, we performed a genome-wide analysis using the readily available multiz 8-way alignment. For each zebrafish chromosome, a zebrafish-referenced 3-way alignment was extracted, giving 25 alignments in total. Approximately 4%-5% of each chromosome was aligned, however this captured 50% of the Ensembl genes.

4.3.1 Identification of conserved non-coding elements

To search for the most conserved elements in each gene, *changept* was applied to each alignment independently. Alignments were segmented into T classes, with the value of T determined using either of the following methods: (1) investigating approximations

to 3 information criteria; Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Type V of Deviance Information Criterion (DICV) [57] (Fig. 4.1); (2) investigating the stability of segment classes based on the conservation levels of each class, calculated using the mean proportion of alignment matches between 3 species [48] (Fig. 4.2). Next, for each class we calculated profile values, using the post burn-in iterations of the sampler [58]. These values are posterior probabilities that each sequence location belongs to the class being profiled. The profiles were visualised in context using BED files uploaded to the UCSC genome browser.

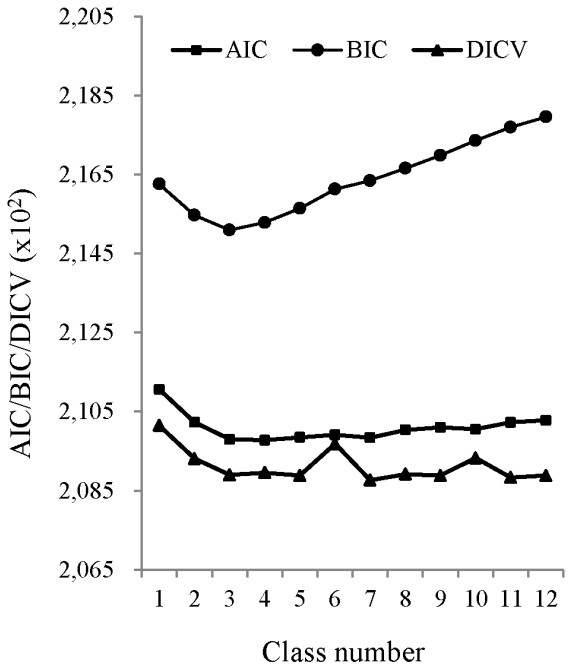


Figure 4.1: *Model selection for eya1. Approximations to well-known information criteria AIC, BIC and DICV for 1-12 classes. Generally, a lower value of the information criteria indicates a better model. BIC clearly suggests a 3-class model. The first local minimum of AIC and DICV has also occurred at the 3-class model. Therefore we selected a 3-class model for this data.*

Not all the genes on the same chromosome have similar levels of conservation; we therefore used a gene-specific approach to identify the most conserved regions in each gene. For each gene, intronic regions that have similar or higher conservation levels compared to exons were the focus of this analysis. Sometimes there was more than one class of segment overlapping with exons of the gene (Fig. 4.3). In that case

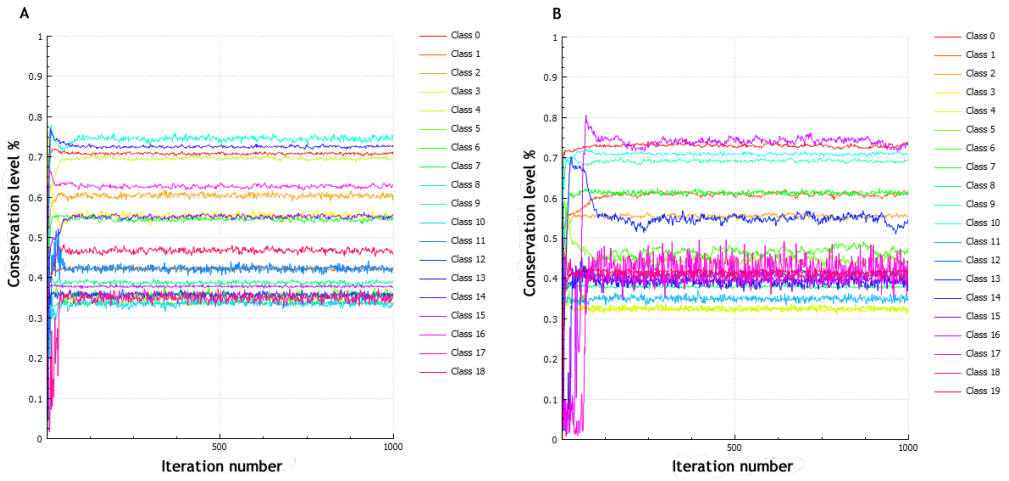


Figure 4.2: *Model selection of chromosome 1 alignment. Figure shows the time series plots of conservation level versus iteration number for each class of (A) 19-class model; and (B) 20-class model. In (A), all classes have stable conservation levels and in (B), one of the classes has a widely varying conservation level. Thus the 19-class model was selected for chromosome 1 alignment. Figure (A) also shows that the model has converged rapidly.*

we considered the segment class or classes that overlap with annotated exons (using annotations in Ensembl) as highly conserved. Notably, there are regions within the introns which show distinct boundaries and probabilities of belonging to the highly conserved classes, but which have no annotated function (Fig. 4.3) and some intronic regions that are more conserved than coding regions (Fig. 4.4).

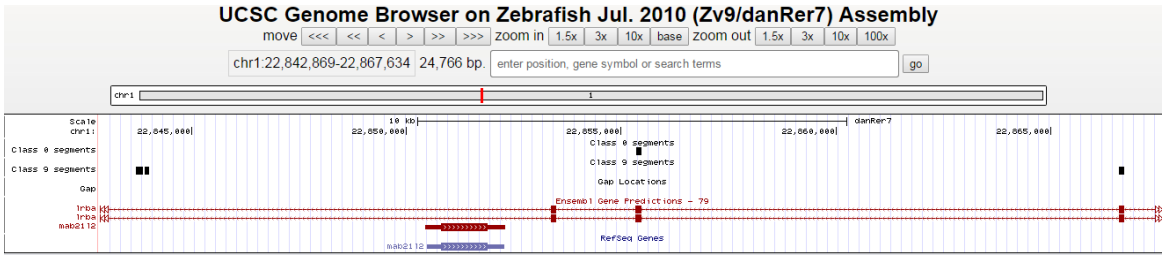


Figure 4.3: *Most conserved segment classes of lrba gene. Two BED files uploaded to UCSC genome browser correspond to Class 0 (conservation - 71%) and Class 1 (conservation - 75%) segments of zebrafish chromosome 1. Segments in each of Class 0 and Class 1 coincide with annotated exons (wide bars) of lrba (ENSDARG00000031108). Segments in the left end of Class 9 do not correspond to any annotated functional elements.*

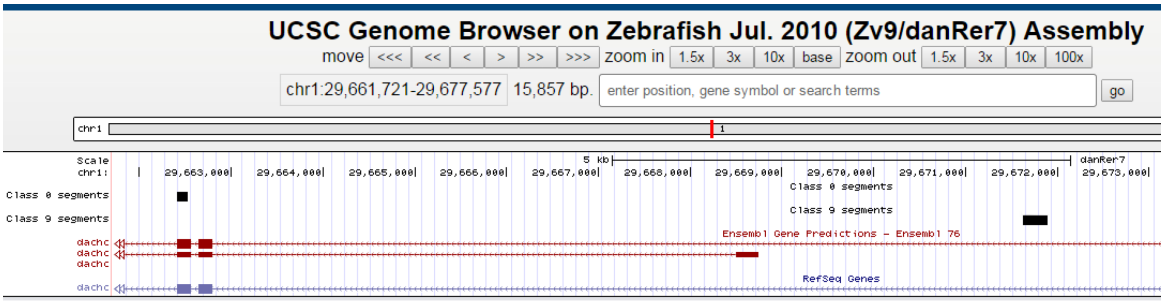


Figure 4.4: *An intronic region more conserved than exons. Figure shows a section of BED files corresponding to Class 0 and Class 9 of zebrafish chromosome 1 uploaded to UCSC genome browser. The conservation levels of Class 0 and Class 9 are 71% and 75% respectively. The annotated exon (wide bars) of dachc (ENSDARG00000003142) coincides with the segment in Class 0. The 261nt long segment in the right end also belongs to Class 9, hence is more conserved than the marked exon.*

Identification of putative functional elements (PFEs)

To identify putative functional non-coding elements in each gene we filtered intronic segments of at least 100nt in length, such that each position in the region had ≥ 0.9 probability of belonging to the highly conserved class/classes of each gene in question. Regions which passed this filtering were referred to as PFEs. These conservative threshold values were chosen to ensure there would be very few false positives. Consequently, many conserved functional elements that do not pass the criteria will have been missed. For example, a PFE length threshold of $\geq 100nt$ will exclude short matches that could be individual transcription factor binding sites, but it will also exclude many spurious alignments. The probability that such long regions could be so well preserved over vast evolutionary time-scales without some form of constraint is vanishingly small, and thus PFEs are nearly certain to be biologically meaningful. The later analysis of comparing PFEs with other supporting evidence and lab results confirmed that there were very few false positives.

In the genome-wide analysis, we identified 665 PFEs distributed among 200 zebrafish genes (including paralogues). Of these, we discarded 10 PFEs as they were located in 7 alignments of non-homologous genes (different genes in human and zebrafish; Table B.2). The medium length of the remaining 655 PFEs was 168nt (based on zebrafish

sequences) and 33% of the PFEs were longer than 200nt. Forty-seven PFEs were located in zebrafish paralogues corresponding to 23 PFEs in human (highlighted in yellow in Supplemental Table 1; http://dx.doi.org/10.6084/m9.figshare.1510941). All other PFEs were in one-to-one correspondence between zebrafish and human. The number of PFEs identified in each zebrafish chromosome is summarised in Fig. 4.5. There were 17 genes containing 10 or more PFEs and another 20 genes containing 5 to 9 PFEs. Thirty-four PFEs were identified in *foxp2* (ENSDARG00000005453); this was the highest number of PFEs located in a single gene. The UCSC genomic coordinates of identified PFEs and Ensembl gene IDs are recorded in Supplemental Table 1 (http://dx.doi.org/10.6084/m9.figshare.1510941).

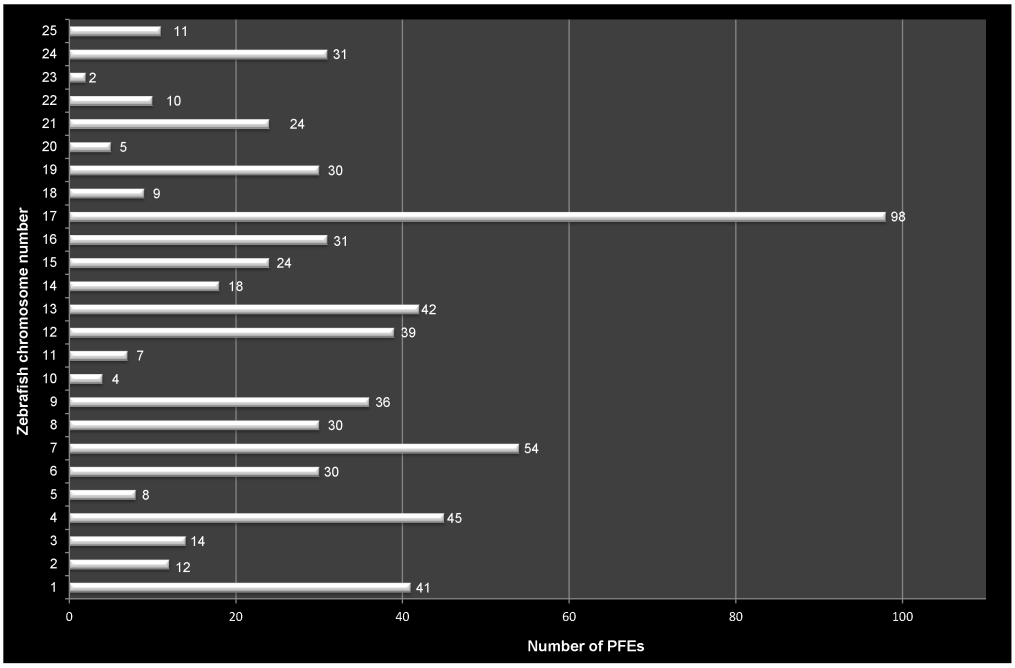


Figure 4.5: *Number of intronic PFEs identified in each zebrafish chromosome. Six hundred and fifty five intronic PFEs were identified in 25 zebrafish chromosomes in total. The highest number of PFEs (98) was detected in zebrafish chromosome 17. Thirty-four PFEs were identified in foxp2 (ENSDARG00000005453) in chromosome 4 and this is the highest number of PFEs found in a single gene followed by 28 PFEs in npas3 (ENSDARG00000079182 chromosome 17).*

Comparing PFEs with other supporting evidence

Two possible reasons that a region of non-coding DNA may be conserved are: (1) that the region is expressed as a ncRNA; and (2) the region is a regulatory element for gene expression. To determine if PFEs represent functional elements, we compared the locations of PFEs with functional regions identified by 4 other methods, (1) EvoFold; (2) RNAz; (3) DNase I footprints; and (4) fRNAdb entries.

Ninety-four percent of PFEs (616) identified in the genome-wide analysis were also identified by other methods (Fig. 4.6). Of those 616 PFEs, 534 (87%) overlapped with EvoFold predicted regions and another 24 PFEs were located near EvoFold regions (within 30nt either side). One hundred and seventy-four PFEs overlapped with multiple EvoFold regions. Of 58 PFEs that do not contain or are not located near EvoFold predictions, 5 were in experimentally validated mouse ncRNAs. EvoFold has identified $21,859$ putative functional intronic elements, genome-wide. The median length of an EvoFold predicted region is 31.5nt (Q1: 21nt; and Q3: 43nt). Of these, 1445 elements were $\geq 100nt$ as opposed to 632 identified by *changept* (relative to the human sequences). However, only 260 EvoFold regions overlap with PFEs. A large number (496) of the EvoFold regions were absent from the alignment we used, and an additional 130 were only partially aligned. The remaining 559 EvoFold predictions were present in the alignment, but were not detected by our method.

According to DNase I footprints data, 342 PFEs (56%) were in protein binding regions (expected overlap -15.2%, this was calculated using simulated segments with the same average length as PFEs using BEDtool-random - see page 120 for details). Ninety-two PFEs (15%) were predicted as structured ncRNAs by RNAz (Supplemental Table 2; http://dx.doi.org/10.6084/m9.figshare.1510936). Comparing with fRNAdb results, 47 PFEs matched with experimentally identified ncRNA transcripts in the database (Fig.4.6 and Supplemental Table 2; http://dx.doi.org/10.6084/m9.figshare.1510936). Of these, 45 mapped to ncRNAs identified in an analysis of the mouse transcriptome [59, 60]. The remaining 2 PFEs

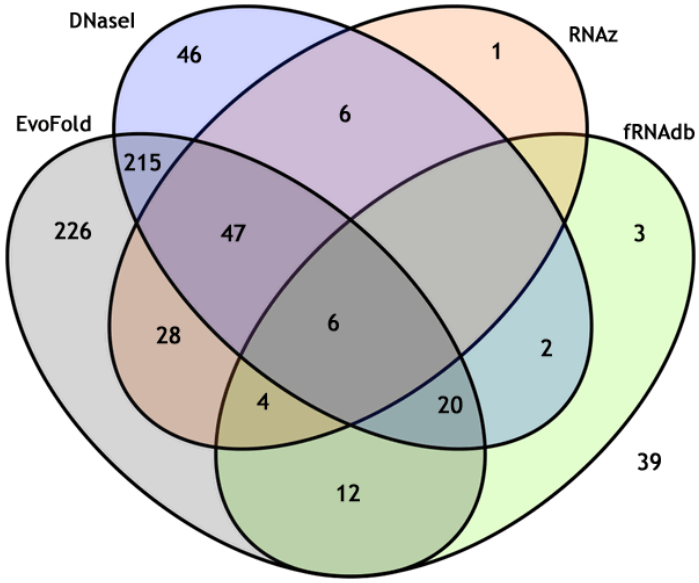


Figure 4.6: *Venn diagram showing the number of genome-wide intronic PFEs supported by other methods. 94% of the PFEs found in the genome-wide analysis were overlapped with the functional elements (predicted or experimentally validated) identified in 4 other databases, EvoFold, fRNAdb, RNAz and DNase I footprints. Most of the PFEs overlapped with entries in EvoFold and there were 47 matches with experimentally identified ncRNA transcripts in fRNAdb.*

were contained in human ncRNA transcripts [61]. Except for one of the human ncRNA transcripts (fRNAdb reference FR407542/FR407474), all other transcripts were substantially longer than the PFEs which they matched. This suggests that regions identified as PFEs represent functional domains within longer RNA transcripts. As an added check to determine if PFEs correspond to ncRNAs, we compared the locations of PFEs with long non-coding RNAs (lncRNAs) identified in zebrafish by [62–64]. There were 8 PFEs overlapping with lncRNA regions (Supplemental Table 2; http://dx.doi.org/10.6084/m9.figshare.1510936). Of 655 PFEs, 39 were not identified by other methods we used for comparisons, and thus can be classified as new predictions.

Investigation of expression using RNA-seq data

Next we used zebrafish RNA-seq data to determine if the PFEs predicted by *changept* were expressed. A large number of zebrafish RNA-seq reads ($\approx$ 327 million from Sanger Institute) were mapped to the zebrafish genome (assembly Zv9). We recorded the

properly paired RNA-seq read counts overlapping with each of the 655 PFEs (Supplemental Table 2; http://dx.doi.org/10.6084/m9.figshare.1510936). Ninety-three percent of PFEs contained at least a single hit showing evidence of expression (expected overlap - 85%; later analysis validated expression even with a single hit; eg: PFE # 2 of *shha* and # 1 of *pax3b*). However no hits may also mean that these are expressed at very low levels, or in tissue types or developmental stages that are not in the database. Supporting this, the remaining 49 PFEs that had 0 reads included 46 PFEs for which other evidence is available (EvoFold, RNAz, DNase I footprints or fRNAdb entries).

Characterizing genes containing PFEs

Intronic ncRNAs are known to be enriched in transcription factors [40]. To find if transcription factors were similarly overrepresented in the 193 genes containing PFEs as identified in the genome-wide analysis, we compared the proportion of transcription factors in 193 genes to proportion of transcription factors in the alignment, using AnimalTFDB database [65]. Results indicated that 40.9% of genes with PFEs (79/193) are transcription factors and 4.7% (9/193) are transcription co-factors. Of the genes represented in the alignment, 10.6% (1733/16296) were identified as transcription factors, 1.5% (240/16296) as transcription co-factors and another 0.5% (85/16296) as chromatin remodelling factors. Therefore PFEs are highly enriched in transcription factors (p-value: 1.2e-56, Z-test for comparing proportions). As an additional analysis, we examined the distribution of Gene Ontology (GO) terms (http://geneontology.org; [66]) in 193 genes with PFEs. According to the results, GO terms associated with transcription factors (eg: sequence-specific DNA binding transcription factor activity, sequence-specific DNA binding RNA polymerase II transcription factor activity, regulation of transcription DNA-templated, transcription from RNA polymerase II promoter, nucleic acid-templated transcription) were significantly overrepresented in genes containing PFEs (Table B.3). This analysis identified 8 more transcription factors not included in AnimalTFDB database (Supplemental Table 3; http://dx.doi.org/10.6084/m9.figshare.1510937).

Identification of intergenic PFEs

In the genome-wide study, we also identified 352 regions that satisfy PFE criteria, that is, are allocated to the most conserved class of the model selected for each chromosome, but are intergenic according to genes annotated by Ensembl and RefSeq. We referred to these regions as *intergenic PFEs*. Of these, 340 intergenic PFEs (97%) were found to overlap with regions identified by other methods (EvoFold, RNAz, DNase I footprints or fRNAdb entries, Supplemental Table 4; http://dx.doi.org/10.6084/m9.figshare.1510938). This also included 12 intergenic PFEs that were in ncRNA transcripts according to fRNAdb entries and 11 intergenic PFEs that overlapped with intergenic lncRNAs identified in [62]. There were 12 highly conserved intergenic regions only identified by program *changept*.

4.3.2 Examination of Genome-wide results in a specific pathway

The second part of our study was a pathway-focussed analysis. This was performed to: (1) repeat the PFE analysis with an improved alignment; (2) identify ncRNAs involved in a well characterised pathway, since that may facilitate future work to determine their function; and (3) identify ncRNAs that may play a role in myogenesis. Pathway-focussed analysis was performed on 11 genes encoding transcription factors known to play important roles in myogenesis, and 13 genes encoding other muscle proteins. For each gene, human-referenced 3-way alignments were generated independently using LAGAN alignment tool [67].

Identification of putative functional elements (PFEs)

To search for the most conserved elements in each gene, we applied *changept* to the 3-way alignments corresponding to each of the 24 myogenesis genes (as opposed to each chromosome in the genome-wide analysis). The profiles were visualised in context using WIG files uploaded to the UCSC genome browser. Fig. 4.7 demonstrates the remarkable effectiveness with which the distinct boundaries of functional elements can

be identified. Class 1 is the most conserved class, and sharp changes (from low to high probabilities) in the WIG profile for Class 1 coincide closely with the annotated positions of exons. Regions within the intron of the gene have not been reported as functional, but are confidently predicted as belonging to the same conservation class that includes all the other exons. These regions were considered for PFE analysis using the same criteria used in the genome-wide analysis (segment length $\geq 100nt$; profile ≥ 0.9).

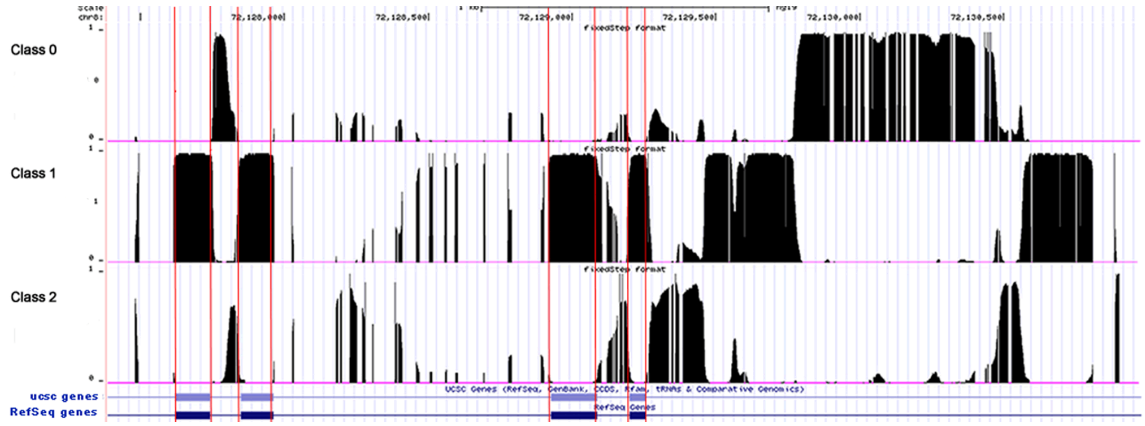


Figure 4.7: *WIG profile of the eya1. The top three profiles show, for each sequence position in the human eya1 DNA sequence (UCSC genomic coordinates chr8:72,109,668-72,268,979), the probability that any base at that position belongs to Class 0 (50% conservation), Class 1 (65% conservation), Class 2 (45% conservation) respectively. At any position, the sum of the three profiles is 1. The two rows below the Class 2 profile display the exons (wide bars) and the introns (thin lines) of eya1 recorded in the UCSC and RefSeq collections respectively. Exon boundaries are indicated with red vertical lines. Class 1 corresponds mainly to the mapped exons of eya1, and covers regions of high conservation between human, mouse and zebrafish.*

We identified 27 PFEs in total and all were found in introns of 7 transcription factors (Table B.4). Of 27 PFEs, only 5 PFEs (3 of *pax3a* and 2 of *eya1* PFEs) were identified in our genome-wide analysis. The majority of PFEs were distributed among *eya1*, *pax3a* and *pax7*. The median length of PFEs was 222nt (based on zebrafish sequences) and there were 15 PFEs longer than 200nt. In contrast, no PFEs were identified in the other muscle genes examined, in either the pathway focussed or genome-wide analyses.

Comparing PFEs with other supporting evidence

We analysed the pathway-focussed PFEs using the same methods used in the genome-wide analysis (EvoFold, RNAz, DNase I footprints, and fRNAdb entries). An example WIG profile of a 169nt long PFE identified in the 3-way alignment of *eya1* is shown in Fig. 4.8. Three possible translation phases (top) indicate a lack of open reading frame within the region. The overlap of the PFE with a sequence protected in DNA footprinting assays indicates protein binding in this region. Furthermore, the PFE is also predicted to be a functional ncRNA by EvoFold.

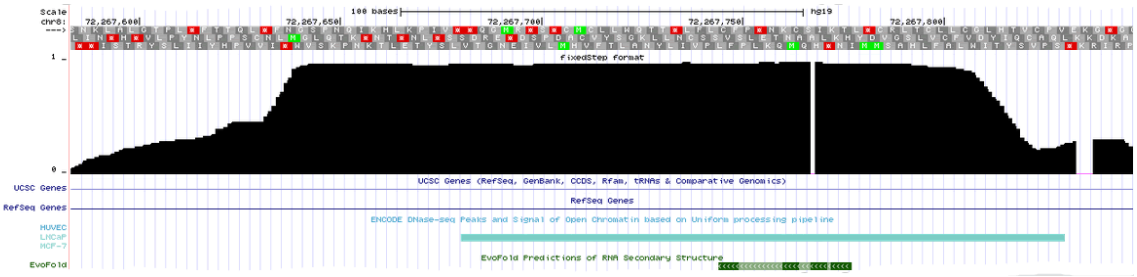


Figure 4.8: *WIG profile of eya1 PFE 4. This PFE is located within intron 2 of human eya1 (UCSC genomic coordinates chr8:72,267,639-72,267,809). The third bar from the top contains single letter amino acid codes corresponding to the actual protein translation phase. At the bottom, the light blue bar indicates a DNase-seq peak track and the green bar shows that there is an EvoFold prediction within the PFE which also suggest that this region is functional.*

The Venn diagram in Fig. 4.9 depicts the number of PFEs supported by other evidence and summarised in Table 4.1 (full details in Supplemental Table 5; http://dx.doi.org/10.6084/m9.figshare.1510939). Of 27 PFEs, 24 were also identified by other methods, providing strong additional evidence of a functional role. Out of those 24, the majority of PFEs were identified by either EvoFold (67%) or DNase I footprint regions (75%). Three PFEs overlapped with multiple EvoFold regions (PFE # 1 of *pax7b*, # 3 and # 4 of *pax3a*). In all cases where PFEs overlap with EvoFold regions, the PFEs are longer; this suggests that our analysis has identified extended functional regions of EvoFold predictions.

EvoFold has predicted 44 intronic functional regions in the same human genes containing PFEs. The median length of an EvoFold region is 31nt (Q1: 20nt and Q3:

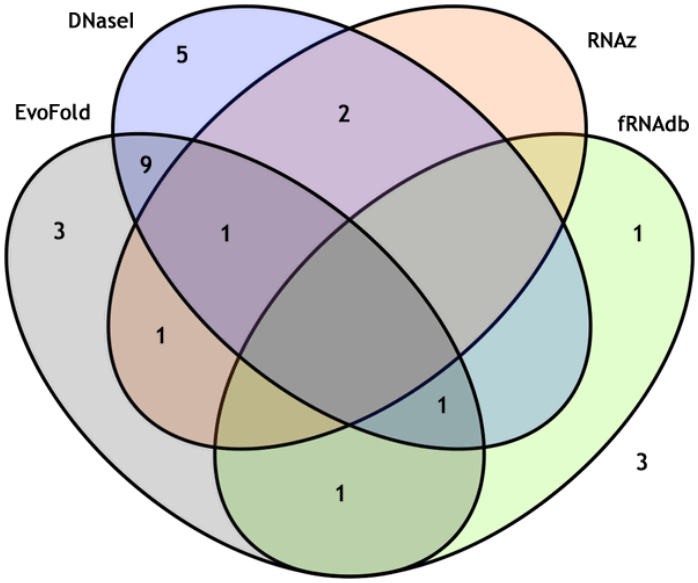


Figure 4.9: *Venn diagram showing the number of pathway-focussed PFEs supported by other methods. 88% of the PFEs found in the pathway-focussed analysis overlapped with the functional elements (predicted or experimentally validated) identified in 4 other databases, EvoFold, fRNAdb, RNAz and DNase I footprints. Most of the PFEs overlapped with entries in either EvoFold or DNase I footprints and there were 3 matches with experimentally identified ncRNA transcripts in fRNAdb.*

Table 4.1: Pathway-focussed results: Number of PFEs supported by other methods suggestive of function

		No. of PFEs contained				
Gene	No. of PFEs identified	EvoFold	DNase I footprints	RNAz	ncRNA transcripts (fRNAdb)	RNA-seq reads
eya1	6	5	6	0	1	6
eya4	2	1	1	0	0	2
pax3 (ZFa)[a]	7	5	4	0	1	7
pax3(ZFb)	2	1	1	0	1	2
pax7(ZFb)	6	4	3	3	0	6
shh(ZFa)	2	0	1	0	0	2
myf5	1	0	1	1	0	1
six4(ZFsix4.3)	1	0	1	0	0	1
Total	27	16	18	4	3	27

$(ZFa)^a$ human and mouse DNA sequences of pax3 is aligned with zebrafish paralog a. Similarly corresponding zebrafish paralog is mentioned within brackets for other genes if any.

52nt). Only 4 of the EvoFold regions were longer than 100nt. These were contained in PFE # 5 of *eya1*, # 2 and # 4 of *pax3a*, and PFE # 4 of *pax7b*. Of 44 EvoFold predictions, only 50% overlapped with PFEs and all of these regions were shorter (Q1: 27nt; median: 47nt; Q3: 70nt) than PFEs identified in the pathway-focussed analysis (which are at least 100nt).

Three PFEs matched with two experimentally identified ncRNA transcripts in mouse (Table 4.2). Both transcripts that mapped to the corresponding region in the mouse genome were substantially longer than the PFEs that they matched. This is consistent with our earlier observation that regions identified as PFEs in the genome-wide analysis, where they overlap with known ncRNAs, are typically shorter than those ncRNAs, and thus may represent functional domains within longer RNA transcripts. The remaining 3 PFEs (PFE # 2 of *shha*, PFE # 1 of *pax3a* and PFE # 6 of *pax7b*) were not identified by any of the 4 methods we used here.

Table 4.2: *Pathway-focussed results: PFEs matching with experimentally identified ncRNAs in fRNAdb*

Gene	Human UCSC coordinates	PFE length (nt)	fRNAdb reference	Length mouse transcript (nt)
eya1	chr8:72267639-72267809	169	FR127136	3697
pax3(ZFa)	chr2:223153695-223153821	126	FR205645	1521
pax3(ZFb)	chr2:223153529-223153656	113	FR205645	1521

Investigation of expression using RNA-seq data

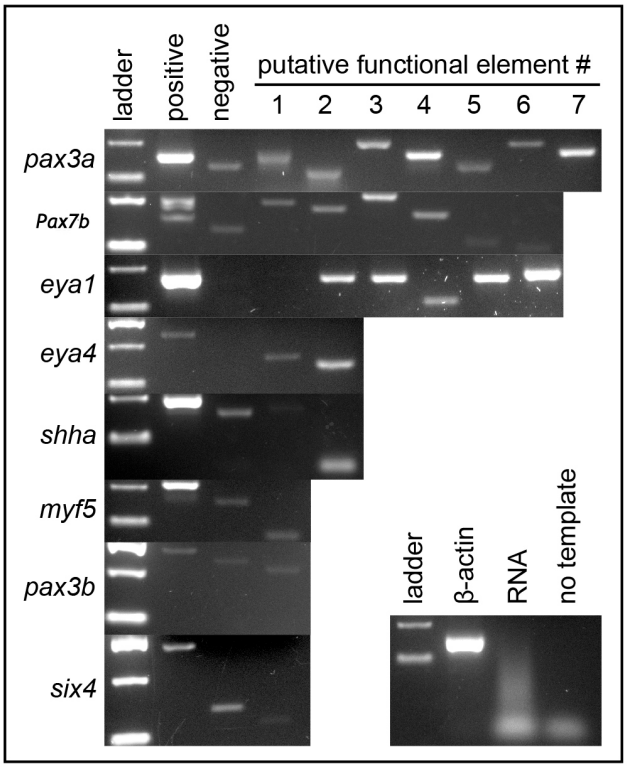
To investigate if PFEs identified in our pathway-focussed analysis show evidence of expression, the number of properly paired reads overlapping with each of the PFEs is recorded in Supplemental Table 6 (http://dx.doi.org/10.6084/m9.figshare.1510940). To ensure the robustness of our analysis, we also examined a set of negative

controls (25) selected from each of the genes containing PFEs. These were intronic regions randomly selected from aligned regions $\geq 100nt$ which do not belong to the most conserved segment class (the class corresponding to exons) of the selected model. Results indicates that all PFEs have at least a single hit from the zebrafish RNA-seq database, suggesting that these regions were all expressed. However, 76% of the control regions also contained RNA-seq reads (Supplemental Table 6; http://dx.doi.org/10.6084/m9.figshare.1510940). This suggests that RNA products encoded within introns are common in these transcription factors, whether the sequence is a PFE or not.

Experimental validation of PFEs

PFEs are transcribed

To investigate whether the intronic PFEs identified were transcribed, RT-PCR analysis was performed using cDNA extracted from 24 hours post-fertilisation (hpf) zebrafish embryos (Fig. 4.10). Primers were designed to amplify short products within the PFE sequences (except for PFE # 2 of *pax3b* as it was too short to design a primer). All of the genes investigated had at least one positive PCR result for a PFE. In total 92% (24/26) of the tested PFEs showed a positive PCR result indicating transcription of the PFE region. The positive control in each case confirmed that the gene of interest, from which the intronic PFE is derived, is also expressed at 24hpf. The negative controls were initially chosen from intronic regions within the gene of interest that were not identified as PFEs. The expected result was that there would be no PCR product as is seen for *eya1* and *eya4*. Contrary to expectations the other 6 intronic regions of interest showed a positive PCR result indicating that this intronic region was also being transcribed. This supports the suggestion that PFEs may be regions within larger transcripts.



	pos	neg	PFE1	PFE2	PFE3	PFE4	PFE5	PFE6	PFE7
pax3a	+	+	+	+	+	+	+	+	+
Pax7b	+	+	+	+	+	+	+	+	
eya1	+	-	-	+	+	+	+	+	
shha	+	+	-	+					
eya4	+	-	+	+					
myf5	+	+	+						present +
pax3b	+	+	+						present (faint) +
six4	+	+	+						absent -

Figure 4.10: *RT-PCR of intronic putative functional elements (PFEs) showing their presence or absence in 24hpf zebrafish cDNA pools. Each gene has between 1 and 7 PFEs. Positive lane represents an exonic region, spanning an intron, of the gene of interest. Negative lane represents a randomly selected intronic region that was not identified as a PFE. Primers were designed to amplify products with sizes ranging 57-274bp. 3 bands of the ladder showing are the 100, 200 and 300bp bands. The gels with 2 bands of the ladder showing are the 100 and 200bp bands. The panel insert is a cDNA control. β-actin (exonic spanning an intron) and RNA (RNA used as a template) lanes demonstrate there is no genomic contamination. No template lane rules out contamination of other PCR reagents.*

Intronic transcripts are associated with PFEs

Given the detection of intronic transcripts for 6 out of 8 of the PFE containing genes we wanted to determine if intronic transcripts were found more frequently in PFE containing genes. We examined 20 additional muscle expressed genes via RT-PCR (Fig. 4.11). Five of the genes (*actn2*, *flnca*, *myod1*, *tpma* and *wnt7ab*) were not detected, of the remaining 15 genes only 1, *wnt7aa*, showed a band in the intronic region.

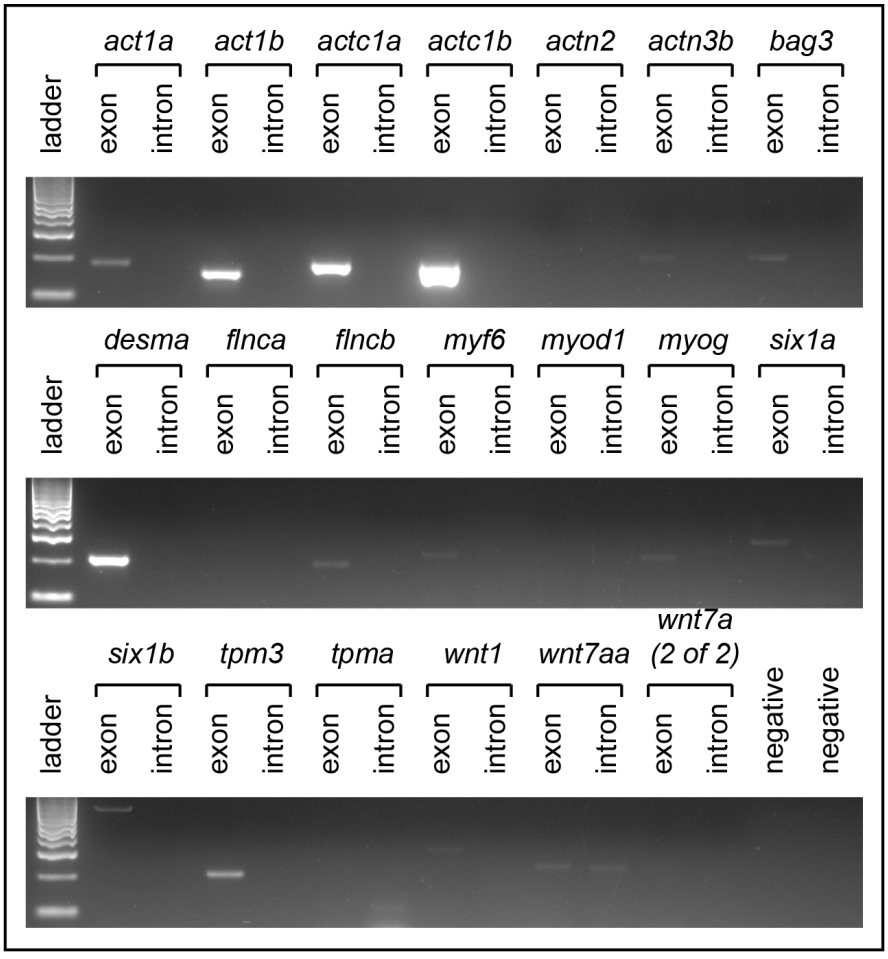


Figure 4.11: *RT-PCR of muscle expressed genes not containing PFEs. The positive controls are marked exon. Exonic, intron spanning, controls are represented as exon. Intronic regions are represented with intron. Primers were designed to amplify products with sizes ranging 100-638bp. Lane 1 for each gel contains a 100bp ladder. The negative lanes are no template controls to rule out contamination.*

4.4 Discussion

We carried out two independent analyses: (1) a genome-wide analysis; and (2) a pathway-focussed analysis of 24 genes involved in myogenesis. The main advantage of the pathway-focussed analysis was that it was based on manually curated alignments performed with the aid of LAGAN, whereas the genome-wide analysis was performed on pre-computed publicly available alignments. Both analyses identified intronic sequences that are highly conserved in the genes of three vertebrate species: human, mouse and zebrafish. We have termed these elements *Putative Functional Elements* (PFEs).

As the name suggests, there is as yet little indication of what function these PFEs might have, or how diverse these functions might be. One clue to the possible functions of PFEs is their prevalence in the introns of transcription factors. This was strikingly demonstrated by the pathway-focussed analysis: all PFEs were found in introns of transcription factors, and none in introns of muscle proteins without transcription factor activity. In the genome-wide analysis, 49.6% of the genes containing PFEs were identified in transcription factors (p-value: 1.2e-56, Z-test for comparing proportions), supporting the finding that PFEs are significantly enriched in transcription factors. PFEs were also found in genes that were not transcription factors, but given that the defining criteria for PFEs are based only on conservation level and length, a mixture of functional types is expected.

PFEs found in the introns of transcription factors could contribute to regulatory interactions in various ways, including:

- Containing binding sites for other transcription factors,
- Containing auto-regulatory binding sites,
- Folding into ncRNAs that interact or form complexes with the containing gene, and

- Folding into ncRNAs that interact or form complexes with other genes in a manner that coordinates their expression levels and activity with that of the containing gene.

It is also possible that PFEs in the introns of transcription factors encode ncRNAs whose function is unrelated to that of the containing gene, but this would not explain why so many PFEs are so located.

Our RT-PCR results showed that PFEs from the introns of muscle-related genes are expressed and suggest that they may play a functional role at the RNA level. In fact, these experimental results indicate something stronger: expression of intronic sequences (not just PFEs) is much more common in transcription factors, at least for genes expressed in muscle tissue at 24 hpf. We found many non-PFE sequences from transcription factor introns were also expressed, but very few sequences from the introns of other muscle-related genes were expressed. A plausible explanation of these results is that the PFEs identified in our pathway-focused analysis are conserved functional domains within longer ncRNAs encoded within the introns of transcription factors. This conclusion is supported by the 47 PFEs that matched experimentally verified ncRNAs in human and mouse: all but one of these were from ncRNAs substantially longer than the PFE. The fact that PFEs represent deeply conserved portions of larger functional elements that are not as strongly conserved, supports the finding that functional sequences need not be conserved (reviewed in [68]).

One surprising finding is that only 5 of the 27 PFEs identified in the pathway-focussed analysis were found in the genome-wide analysis. We attribute this to the superior quality of the alignments used in the pathway-focussed analysis, due not only to the use of LAGAN, but also to manual interventions to improve alignment quality. This was not feasible genome-wide. It does suggest, however, that the genome-wide analysis may be finding only a fraction of the intronic elements conserved between human and zebrafish.

To determine if PFEs correspond to ncRNAs or other regulatory sequences, we compared them to other bioinformatics resources (EvoFold, RNAz, DNase-seq footprints and fRNAdb entries). The majority (85%) of our PFEs identified in the genome-wide study contain EvoFold predicted regions. EvoFold has identified 1445 intronic regions longer than 100nt in the human genome with the potential to form RNA structures. However a larger number of these regions were absent from the alignment we used. This could be due in part to using different alignments with different assemblies and even different species. Our analysis was performed using a more recent alignment including the human 2009 assembly, whereas EvoFold findings are based on an earlier 8-way alignment including the human 2004 assembly. The alignments contain only 4 species in common: human, mouse, zebrafish and fugu. On the other hand, we failed to detect 559 EvoFold predictions that were present in our alignment. This could be due to: (1) failing to satisfy the PFE gap criteria (we rejected segments with a gap of ≥ 20 alignment columns or if the total length of gaps within the segment was $\geq 10\%$ the length of the segment); or (2) the segments may not be as highly conserved as exons.

This situation was reversed in the pathway-focussed analysis, where we identified 27 PFEs and EvoFold only found 4 regions $\geq 100nt$ in the same human genes. This could be attributed to the success of our Bayesian method applied to an improved alignment used in the pathway-focussed analysis.

Ninety-seven (15%) of the PFEs identified in the genome-wide analysis do not contain EvoFold regions and are not within 30nt of an EvoFold region. Of these, 61% (59) overlap either RNAz, DNase I footprints or fRNAdb entries and 35/38 of the remaining PFEs not identified by these methods/resources contained at least 1 RNAseq hit. Moreover, 11 PFEs identified in the pathway-focussed analysis do not contain EvoFold predictions but were all found to be expressed in our RT-PCR results. In addition to identifying putative ncRNAs not identified by EvoFold, our method typically extends the length of the predicted functional regions, so much so that many of our PFEs

contain two or more EvoFold predictions. In particular, in the pathway-focussed results, PFEs that contain an EvoFold prediction are substantially longer than that Evofold prediction.

One of the limitations in validating PFEs using other resources is the fact that some of the ncRNAs and regulatory sequences will only have a functional role at a particular developmental stage in a specific tissue. For example, if one looks for expression or DNAse sensitivity of such a PFE at a different developmental stage or in a different tissue to that in which it has a functional role, validation will not be successful. For this reason, even those PFEs that we were not able to validate via other methods may yet be functional.

In summary, our study provides a systematic process centred around a Bayesian segmentation method to identify putative intronic functional elements in genomes that may contain ncRNAs and other regulatory sequences. That these elements are enriched in transcription factors provides further evidence that they are functional domains of ncRNAs, given that ncRNAs are also known to be enriched in transcription factors [40].

4.5 Methods

4.5.1 The list of genes used in pathway-focussed analysis

Transcription factors of myogenesis pathway: *eya1, eya4, pax3, pax7, six4, myf5, shh, six1, myod1, myog, myf6*

Other muscle proteins: *wnt1, wnt7a, acta1, actc1, actn2, actn3, bag3, des, flnc, tpm3, myh7, tnnt1, nebulin*

4.5.2 Sequence and alignment of data

Genome-wide analysis

Multiz 8-way alignment was downloaded from UCSC genome browser (http://hgdownload-test.cse.ucsc.edu/goldenPath/danRer7/multiz8way/). The assemblies used in the alignments were: zebrafish: Zv9/danRer7; human: hg19/GRCh37 and mouse: GRCM38/mm9. For each zebrafish chromosome, the 3-way alignment (zebrafish-referenced) was extracted using program mafExtractor (https://github.com/dentearl/mafTools/tree/master/mafExtractor) giving 25 alignments in total, one for each zebrafish chromosome.

Pathway- focussed analysis

Human, mouse and zebrafish DNA sequences for each of 24 genes were downloaded from Ensembl genome browser (http://www.ensembl.org/index.html; zebrafish: Zv9; human: GRCh37 and mouse: NCBIM37). For 10 of these 24 genes (*pax3, shh, six1, wnt7a, acta, actc, actn3, desm, flnc, tpm3*), there are 2 paralogues in zebrafish and for *myh7* there are 3 paralogues. Thus a separate 3-way alignment was generated for each of these, giving a total of 36 alignments (for *pax7*, only *pax7b* was used as we could not find the complete sequence of *pax7a*). We used LAGAN [67] to perform the 3-way alignments (human-referenced) using default parameters. For the few cases where we noticed mis-alignments of exons (eg: *myf6, wnt7aa*), those sequences were aligned separately using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) effectively forcing exons to align. We then combined the ClustalW2 results (partial alignments) with the original LAGAN alignments.

4.5.3 Transform alignments

Each of the 3-way alignments was transformed into a single 32-character sequence *(A=a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u,v,w,x,y,z,U,V,W,X,Y,Z)* using the following

encoding using a perl script. This sequence was used as the input for program *changept*. The alignment columns with complementary bases were also encoded using the same characters.

```
Human:      AAAAAAAAAAAAAAAACCCCCCCCCCCCCCCC
Mouse:      AAAACCCCGGGGTTTTAAAACCCCGGGGTTTT
Zebrafish:  ACGTACGTACGTACGTACGTACGTACGTACGT
Symbol:     abcdefghijklmnopqrstuvwxyzUVWXYZ
```

The insertions and deletions (indels) in the alignment were excluded from analysis. In the genome-wide analysis, discontinuous alignment blocks with respect to each species were also separated by using a # symbol. The # symbol is considered as a fixed change-point in the model. Occasionally *changept* identified only 1-class of segments in segmenting the 3-way alignments of relatively short genes (for example *shh, myog, six1, six4* in pathway-focussed analysis). This problem was overcome by concatenating the 32-character sequences of such genes, thus providing *changept* a larger sample to segment.

4.5.4 Change-point analysis

A full description of the *changept* model can be found in previous papers [50, 51, 69]. In summary, the sequences generated for 3-way alignments for each of the genes/chromosomes were separately run through *changept* to find positions (change-points) in the sequences that delineate homogeneous segments. Character frequencies within each segment are modelled as a multinomial distribution with parameter $\Theta = (\theta_a, \theta_b, ..., \theta_Y, \theta_Z)$, where θ is drawn from one of T Dirichlet distributions. As the number of classes (T) is unknown a priori, independent runs with different numbers of classes were performed. The generalized Gibbs sampler [69] was used to sample from the varying dimensional space: it allows the number of change-points to vary. Each

model was run with varying values of T for 1,000 iterations. Information criteria was then used to select the value of T.

4.5.5 Assessing convergence

The convergence of the model was assessed by plotting the log-likelihood of each of the 1000 iterations using an R script. The burn-in phase is characterised by an upward trend in the log-likelihood. In addition, we used Heidelberger and Welch convergence diagnostic test [70, 71] of the CODA package [72, 73] to validate the convergence results (Table B.1).

4.5.6 Model selection

To determine the optimal number of classes for each alignment, we calculated approximations to three information criterion values - AIC, BIC and DIC - using post burn-in samples. These approximations are discussed in [57]. The model with the smallest information criterion value is generally considered optimal. However, model selection was not purely based on this method. A subjective judgement was made on which model to choose by investigating the mixture proportions; a model containing classes with very low mixture proportions was considered to be an over-fitted model and thus a model with a smaller number of classes was selected. In combination with this method, we also used an alternative model selection method, by investigating the stability of segment classes [48]. Stability of classes was assessed based on time-series plots of conservation levels versus sample number. Classes which were highly variable in conservation levels were deemed unstable (again this involved a subjective judgement).

4.5.7 Quantifying the conservation level of segment classes

Changept employs Markov Chain Monte Carlo (MCMC) sampling. The individual character frequencies within each class were calculated at each iteration. To determine

the conservation level of each class for the selected model, the mean proportion of alignment matches, $E(\theta)$ was calculated for each iteration of the sampler.

$$E[\theta] = \frac{\theta_a + \theta_v}{\sum_{j \in A} \theta_j}$$

Here characters a and v represent conserved bases. These values were plotted against each iteration number (Fig. 4.2). These conservation plots were also used to assess the convergence as a second method (eg: Fig. 4.2(A) shows that convergence to the limiting distribution has occurred).

4.5.8 The readcp program

We used *readcp* program (part of the *changept* package) to calculate profile values showing the probability that each sequence position belongs to a given class of the chosen model. These posterior probabilities are estimated by Monte Carlo integration. A complete description of how to use programs *changept* and *readcp* can be found in [58].

4.5.9 Identifying putative functional elements

PFEs were identified for the 3-way alignments of each gene using the following criteria: an intronic segment of at least 100nt in length, such that each position had ≥ 0.9 probability of belonging to the most conserved segment class/classes. As *changept* skips gaps in the alignment, gaps were considered in the following manner: a segment was not considered continuous if there was a gap of ≥ 20 alignment columns or if the total length of gaps within the segment was $\geq 10\%$ the length of the segment. In the genome-wide analysis, regions that satisfy PFE criteria belonging to the most conserved class of the selected model corresponding to each zebrafish chromosome, but not located in genic regions were referred as *intergenic* PFEs.

4.5.10 Creating wiggle tracks and BED files

The *readcp* output was used to generate BED files or wiggle tracks (one for each class in the final model) so that results could be plotted as a profile alongside gene tracks and other information in the UCSC browser. In the genome-wide analysis, we used the more compact BED file format to handle the large amount of data. The positions of segments matching PFE criterion (minimum segment length of 100nt with profile ≥ 0.9 and same gap criterion as above) in each class and in each model were recorded in BED format with genomic coordinates relative to zebrafish. We used *intersect* BEDtool (http://bedtools.readthedocs.org/en/latest/content/tools/intersect.html) to find the segment class (or classes) that overlap with annotated exons (3′ UTR exons, 5′ UTR exons and the coding exons downloaded from UCSC table browser) of the gene in question. Sometimes there was more than one class corresponding to annotated exons of the gene (Fig. 4.3) and occasionally segments satisfying PFE criteria were found to be located in a class more highly conserved than a class corresponding to marked exons (for example, there is a PFE in Class 9 in Fig. 4.4). Thus in each gene, segments that were conserved at a level comparable or higher than exons were considered for PFE analysis. In our analysis we only reported PFEs with conservation level >50%. Wiggle tracks were used in the pathway-focussed analysis. The WIG profile for a selected class shows the probability that the base at a particular position in the sequence belongs to the class in question, thus every position has an associated value between 0 and 1 (Fig. 4.7). In this analysis, we examined the wiggle track of the most conserved segment class (for example, Class 1 of Fig. 4.7).

4.5.11 Mapping with zebrafish RNA-seq data

The RNA-seq reads were downloaded from the European Nucleotide Archive (ENA) web application accessible at http://www.ebi.ac.uk/ena/data/view/ERP000016. These paired-end reads were 36, 37, 54 and 76 base pairs long and had been extracted from

zebrafish embryonic and adult tissues [74]. We performed the initial quality control (QC) checks using FASTQC program (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were filtered for quality using Trimmomatic [75] by removing all leading and trailing bases with quality less than 30, and then discarding: (1) reads shorter than 20nt after trimming; and (2) any unpaired reads. We used Bowtie2 [76] with default options to align a total of 327,019,912 QC passed paired-end reads to the zebrafish genome. 86% of the reads were mapped. We then used BEDtools - multicov to count the number of properly paired reads overlapping with each of the PFEs to get an indication whether the regions identified by the *changept* were expressed.

4.5.12 Other supporting evidence

EvoFold

Human genomic coordinates of EvoFold regions were downloaded in BED format using UCSC table browser. To check the overlap between PFEs and EvoFold regions, we used BEDtool -intersect.

DNase I footprints

We used the database of DNase-seq footprints identified by the ENCODE project [77] in their large-scale analysis of 41 different human cell types. The data (combined.fps.gz) was downloaded from link ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/footprints/jan2011/. Once again BEDtool -intersect was used to check the overlap between PFEs and DNase-seq footprints.

fRNAdb

The BLAST function of fRNAdb database (accessible at http://www.ncrna.org/frnadb/blast) was used to search for fRNAdb entries (ncRNA transcripts and RNAz

regions) with high sequence similarity to human sequences of each PFE identified in our analysis.

Calculating expected "overlaps"

To find what proportion of PFEs would overlap DNAse I hypersensitive regions if PFEs were randomly distributed in the aligned portions of the human genome, first we generated a BED file of random segments of the aligned portion of the genome using BEDtool- random. We set the segments' length to average PFE length. Next, we used BEDtool- intersect to find the overlap between randomly distributed PFEs and DNAse I hypersensitive regions. The same method was used to find the expected overlaps between RNA-seq properly paired reads and randomly distributed PFEs in zebrafish genome.

4.5.13 Experimental validation

Zebrafish maintenance and cDNA synthesis

Zebrafish were maintained as previously described in [78]. RNA was collected from 24hpf wild-type embryos using TRI-Reagent (Sigma-Aldrich) and treated with DNAse (Promega) to remove genomic DNA. cDNA was synthesised using the ProtoScript II First Strand cDNA Synthesis Kit (NEB) according to the manufacturer's instructions.

Designing primers

Positive control sequences were obtained using Ensembl Genome Browser (`http://www.ensembl.org/index.html`) and regions spanning introns of the genes of interest were selected. PFE and negative control sequences were obtained after analysis with *changept* and primers were designed using the online software Primer3 (`http://bioinfo.ut.ee/primer3`). Polymerase chain reaction and Gel electrophoresis Reverse transcriptase PCR was performed using GoTaq Green Master Mix (Promega). Samples were amplified for 30 cycles with an annealing temperature of 57°C. 15 μl of each

sample was run on a 3% TBE gel, supplemented with GelRed (Biotium), at 60V for 3 hours.

PCR and Gel

Reverse transcriptase PCR was performed using GoTaq Green Master Mix (Promega). Samples were run for 30 cycles with an annealing temperature of 57°C. 15 μl of each sample was run on a 3% TBE gel, supplemented with GelRed (Biotium), at 60V for 3 hours.

4.5.14 Identification of proportion of transcription factors in genes with PFEs

AnimalTFDB (http://bioinfo.life.hust.edu.cn/AnimalTFDB/index.shtml; [79]) is a comprehensive database including classification and annotation of genome-wide transcription factors, transcription co-factors and chromatin remodelling factors in 65 animal genomes including zebrafish. To examine the proportion of genes containing PFEs that belong to each of these 3 categories, we first downloaded the Ensembl gene list associated with each category. In total, there were 2,345 transcription factors, 315 transcription co-factors and 100 chromatin remodelling factors in the database. Next we used BEDtool-intersect to check how many genes were represented in genome-wide 3 way alignments. 16,296 genes (from total 32,475 Ensembl genes) overlapped with the segments recorded in our BED files. The final step was to examine the proportion of transcription factors, transcription co-factors and chromatin remodelling factors in aligned 16,296 genes using the 3 lists downloaded from AnimalTFDB. To perform GO enrichment analysis, we used *AmiGO* web interface accessible at http://amigo.geneontology.org/amigo [80]. We obtained significant GO terms (with p-value <0.05) in each of three sub-ontologies: Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) using 193 zebrafish genes containing PFEs. Further, we manually filtered GO terms associated

with *DNA binding*, *regulation of gene expression*, *sequence-specific DNA binding* and *nucleic acid binding* to check if any of the genes in the sample were classified as transcription factors using existing evidence.

4.6 Data access

The zebrafish positions of the intronic and intergenic PFEs identified in the genome-wide analysis were recorded in BED format and are available as part of supplemental materials (Supplemental File 1, http://dx.doi.org/10.6084/m9.figshare.1517694 and Supplemental File 2, http://dx.doi.org/10.6084/m9.figshare.1517695, respectively).

4.7 Acknowledgements

This work was supported by the Australian Research Council (grant DP1095849). We thank Dr. Sarah Boyd for initial helpful discussions and Dr. Nathan S. Watson-Haigh for insightful discussions in analysing RNA-seq data.

Bibliography

[1] D R Corey. Regulating mammalian transcription with RNA. *Trends Biochem. Sci*, 30:655–658, 2005.

[2] J S Mattick and I V Makunin. Small regulatory RNAs in mammals. *Hum. Mol. Genet.*, 14:R121–R132, 2005.

[3] S Kishore and S Stamm. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science*, 311:230–232, 2006.

[4] J S Mattick and I V Makunin. Non-coding RNA. *Human Molecular Genetics*, 15: R17–R29, 2006.

[5] G Storz, J A Opdyke, and A Zhang. Controlling mRNA stability and translation with small, non-coding RNAs. *Curr. Opin. Microbiol.*, 7:140–144, 2004.

[6] A M Khalil, M Guttman, M Huarte, M Garber, A Raj, D Rivea Morales, and J L Rinn. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci*, 106:11667–11672, 2009.

[7] M J Koziol and J L Rinn. RNA traffic control of chromatin complexes. *Curr Opin Genet Dev*, 20:142–148, 2010.

[8] J L Rinn, M Kertesz, J K Wang, S L Squazzo, X Xu, S A Brugmann, L H Goodnough, J A Helms, P J Farnham, E Segal, and H Y Chang. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129:1311–1323, 2007.

[9] M Fabbri. *Non-coding RNAs and cancer*. Springer, New York, NY, 2014. ISBN 1461484448.

[10] G A Calin, C G Liu, M Ferracin, T Hyslop, R Spizzo, C Sevignani, M Fabbri, A Cimmino, E J Lee, S E Wojcik, and et al. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell*, 12:215–229, 2007.

[11] S L Fernandez-Valverde, R J Taft, and J S Mattick. MicroRNAs in -cell biology, insulin resistance, diabetes and its complications. *Diabetes*, 60:1825–1831, 2011.

[12] M N Poy, L Eliasson, J Krutzfeldt, S Kuwajima, X Ma, P E Macdonald, S Pfeffer, T Tuschl, N Rajewsky, P Rorsman, and M Stoffel. A pancreatic islet-specific microRNA regulates insulin secretion. *Nature*, 432:226–230, 2004.

[13] S S Hebert, A S Papadopoulou, P Smith, M C Galas, E Planel, A N Silahtaroglu, N Sergeant, L Buee, and De Strooper B. Genetic ablation of dicer in adult forebrain

neurons results in abnormal tau hyperphosphorylation and neurodegeneration. *Hum. Mol. Genet.*, 19:3959–3969, 2010.

[14] J Kim, K Inoue, J Ishii, W B Vanti, S V Voronov, E Murchison, G Hannon, and A Abeliovich. A microRNA feedback circuit in midbrain dopamine neurons. *Science*, 317:1220–1224, 2007.

[15] A Schaefer, D O'Carroll, C L Tan, D Hillman, M Sugimori, R Llinas, and P Greengard. Cerebellar neurodegeneration in the absence of microRNAs. *J. Exp. Med.*, 204:1553–1558, 2007.

[16] G Wang, J M van der Walt, G Mayhew, Y J Li, and S Zuchner. Variation in the miRNA-433 binding site of FGF20 confers risk for Parkinson disease by overexpression of alpha-synuclein. *AAm. J. Hum. Genet.*, 82:283–289, 2008.

[17] T L Young, T Matsuda, and C L Cepko. The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. *Curr. Biol*, 15:501–512, 2005.

[18] M R Ginger, A N Shore, A Contreras, M Rijnkels, J Miller, M F Gonzalez-Rimbau, and J M Rosen. A noncoding RNA is a potential marker of cell fate during mammary gland development. *Proc. Natl Acad. Sci. USA*, 103:5781–5786, 2006.

[19] M E Askarian-Amiri, J Crawford, J D French, C E Smart, M A Smith, M B Clark, K Ru, T R Mercer, E R Thompson, S R Lakhani, and et al. SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer. *RNA*, 17:878–891, 2011.

[20] W Hu, B Yuan, J Flygare, and H F Lodish. Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation. *Genes Dev.*, 25:2573–2578, 2011.

[21] M Kretz, Z Siprashvili, C Chu, D E Webster, A Zehnder, K Qu, C S Lee, R J Flockhart, A F Groff, J Chow, and et al. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature*, 493:231–235, 2013.

[22] N Brockdorff, A Ashworth, G F Kay, V M McCabe, D P Norris, P J Cooper, S Swift, and S Rastan. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell*, 71:515–526, 1992.

[23] C J Brown, B D Hendrich, J L Rupert, R G Lafreniere, Y Xing, J Lawrence, and H F Willard. The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, 71:527–542, 1992.

[24] V H Meller, K H Wu, G Roman, M I Kuroda, and R L Davis. roX1 RNA paints the X chromosome of male drosophila and is regulated by the dosage compensation system. *Cell*, 88:445–457, 1997.

[25] Mattick J S Mercer T R. Structure and function of long noncoding RNAs in epigenetic regulation. *Nature Structural and Molecular Biology*, 20:300–307, 2013.

[26] M A Ripoche, C Kres, F Poirier, and L Dandolo. Deletion of the H19 transcription unit reveals the existence of a putative imprinting control element. *Genes Dev*, 11:1596–1604, 1997.

[27] J T Lee, L S Davidow, and D Warshawsky. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nature Genet*, 21:400–404, 1999.

[28] T Sado, Z Wang, H Sasaki, and E Li. Regulation of imprinted X-chromosome inactivation in mice by Tsix. *Development*, 128:1275–1286, 2001.

[29] F Sleutels, R Zwart, and D P Barlow. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, 415:810–813, 2002.

[30] N Thakur, V K Tiwari, H Thomassin, R R Pandey, M Kanduri, A Gondor, T Grange, R Ohlsson, and C Kanduri. An antisense RNA regulates the bidirectional silencing property of the Kcnq1 imprinting control region. *Mol. Cell. Biol*, 24:7855–7862, 2004.

[31] S Swiezewski, F Liu, A Magusin, and C Dean. Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature*, 462:799–802, 2009.

[32] J S Mattick. RNA as the substrate for epigenome-environment interactions: RNA guidance of epigenetic processes and the expansion of RNA editing in animals underpins development, phenotypic plasticity, learning, and cognition. *Bioessays*, 32:548–552, 2010.

[33] J S Mattick. The central role of RNA in human development and cognition. *FEBS Lett*, 585:1600–1616, 2011.

[34] J A Hural, M Kwan, G Henkel, M B Hock, and M A Brown. An Intron Transcriptional Enhancer Element Regulates IL-4 Gene Locus Accessibility in Mast Cells. *The Journal of Immunology*, 165:3239–3249, 2000.

[35] J Majewski and J Ott. Distribution and characterization of regulatory elements in the human genome. *Genome Res*, 12:1827–1836, 2002.

[36] Z Li and C K S Carlow. Characterization of Transcription Factors That Regulate the Type IV Secretion System and Riboflavin Biosynthesis in *Wolbachia* of *Brugia malayi*. *PLoS ONE*, 7:e51597, 2012.

[37] F Vallania, D Schiavone, S Dewilde, E Pupo, S Garbay, R Calogero, M Pontoglio, P Provero, and V Poli. Genome-wide discovery of functional transcription factor binding sites by comparative genomics: the case of Stat3. *Proc Natl Acad Sci USA*, 106:5117–5122, 2009.

[38] J Wang, J Zhuang, S Iyer, X Lin, T W Whitfield, M C Greven, B G Pierce, X Dong, A Kundaje, Y Cheng, and et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res*, 22:1798–1812, 2012.

[39] T W Whitfield, J Wang, P J Collins, E C Partridge, S F Aldred, N D Trinklein, R M Myers, and Z Weng. Functional analysis of transcription factor binding sites in human promoters. *Genome Biol*, 13:R50, 2012.

[40] H I Nakaya, P P Amaral, R Louro, A Lopes, A A Fachel, Y B Moreira, T A El-Jundi, A M da Silva, E M Reis, and S Verjovski-Almeida. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome Biology*, 8:R43, 2007.

[41] R Louro, A S Smirnova, and S Verjovski-Almeida. Long intronic noncoding RNA transcription: expression noise or expression choice? *Genomics*, 93:291–298, 2009.

[42] M Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31:3406–3415, 2003.

[43] I L Hofacker and P F Stadler. Memory efficient folding algorithms for circular RNA secondary structures. *Bioinformatics*, 22:1172–1176, 2006.

[44] A R Gruber, S Findei, S Washietl, I L Hofacker, and P F Stadler. RNAz 2.0: Improved noncoding RNA detection. *Pac Symp Biocomput*, 15:69–79, 2010.

[45] J S Pedersen, G Bejerano, A Siepel, K Rosenbloom, K Lindblad-Toh, E S Lander, J Kent, W Miller, and D Haussler. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol*, 2:e33. DOI: 10.1371/journal.pcbi.0020033, 2006.

[46] F Tajima. Determination of window size for analysing DNA sequences. *Journal of Molecular Evolution*, 33:470–473, 1991.

[47] J V Braun and H-G Muller. Statistical methods for DNA sequence segmentation. *Statistical Science*, 13:142–162, 1998.

[48] M Algama, C Oldmeadow, E Tasker, K Mengersen, and J M Keith. Drosophila 3' UTRs Are More Complex than Protein-Coding Sequences. *PLoS ONE*, 9:e97336, 2014.

[49] C Oldmeadow, K Mengersen, J S Mattick, and J M Keith. Multiple evolutionary rate classes in animal genome evolution. *Molecular Biology and Evolution*, 27: 942–953, 2010.

[50] J M Keith. Segmenting eukaryotic genomes with the Generalized Gibbs Sampler. *Journal of Computational Biology*, 13:1369–1383, 2006.

[51] J M Keith, P Adams, S Stephen, and J S Mattick. Delineating slowly and rapidly evolving fractions of the Drosophila genome. *Journal of Computational Biology*, 15:407–430, 2008.

[52] R J Bryson-Richardson and P D Currie. The genetics of vertebrate myogenesis. *Nature Reviews Genetics*, 9:632–646, 2008.

[53] C G Crist, D Montarras, G Pallafacchina, D Rocancourt, A Cumano, S J Conway, and M Buckingham. Muscle stem cell behavior is modified by microRNA-27 regulation of Pax3 expression. *Proc Natl Acad Sci U S A*, 106:13383–13387, 2009.

[54] J F Chen, Y Tao, J Li, Z Deng, Z Yan, X Xiao, and D Z Wang. microRNA-1 and microRNA-206 regulate skeletal muscle satellite cell proliferation and differentiation by repressing Pax7. *Journal of Cell Biology*, 190:867–879, 2010.

[55] B K Dey, J Gagan, and A Dutta. miR-2006 and -486 induce myoblast differentiation by downregulating Pax7. *Mol Cell Biol*, 31:203–214, 2011.

[56] M Cesana, D Cacchiarelli, I Legnini, T Santini, O Sthandier, M Chinappi, A Tramontano, and I Bozzoni. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell*, 147:358–369, 2011.

[57] C Oldmeadow and J M Keith. Model Selection in Bayesian Segmentation of multiple DNA alignments. *Bioinformatics*, 27:604–610, 2011.

[58] J M Keith. Sequence segmentation. *Methods Mol Biol*, 452:207–229, 2008.

[59] Y Okazaki, M Furuno, T Kasukawa, J Adachi, H Bono, S Kondo, I Nikaido, N Osato, R Saito, H Suzuki, and et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, 420:563–573, 2002.

[60] P Carninci, T Kasukawa, S Katayama, J Gough, M C Frith, N Maeda, R Oyama, T Ravasi, B Lenhard, C Wells, and et al. The transcriptional landscape of the mammalian genome. *Science*, 309:1559–1563, 2005.

[61] T Imanishi, T Itho, Y Suzuki, C ODonovan, S Fukuchi, K O Koyanagi, R A Barrero, T Tamura, Y Yamaguchi-Kabata, M Tanino, and et al. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol*, 2:856–875, 2004.

[62] A Pauli, E Valen, M F Lin, M Garber, N L Vastenhouw, J Z Levin, L Fan, A Sandelin, J L Rinn, and et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res*, 22:577–591, 2012.

[63] K Kaushik, V E Leonard, K V Shamsudheen, M K Lalwani, S Jalali, A Patowary, A Joshi, V Scaria, and S Sivasubbu. Dynamic expression of long non-coding RNAs (lncRNAs) in adult zebrafish. *PLoS ONE*, 8:e83616, 2013.

[64] I Ulitsky, A Shkumatava, C H Jan, H Sive, and D P Bartel. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, 147:1537–1550, 2011.

[65] H Zhang, H Chen, W Liu, H Liu, J Gong, H Wang, and A Guo. AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Research*, 2011. doi: 10.1093/nar/gkr965.

[66] Gene Ontology Consortium. Gene Ontology annotations and resources. *Nucleic Acids Res*, 41:D530–D535, 2013.

[67] M Brudno, C B Do, G M Cooper, M F Kim, E Davydov, NISC Comparative Sequencing Program, E D Green, A Sidow, and S Batzoglou. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*, 13:721–731, 2003.

[68] K C Pang, M C Frith, and J S Mattick. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet*, 22:1–5, 2006.

[69] J M Keith, D P Kroese, and D Bryant. A Generalized Markov Sampler. *Methodology and Computing in Applied Probability*, 6:29–53, 2004.

[70] P D Welch and P Heidelberger. A spectral method for confidence interval generation and run length control in simulations. *Comm. ACM.*, 24:233–245, 1981.

[71] P D Welch and P Heidelberger. Simulation run length control in presence of an initial transient. *Operations Research*, 31:1109–1144, 1983.

[72] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[73] R Development Core Team. *R: A Language and Environment for Statistical Computing*. The R Foundation for Statistical Computing, Vienna, Austria, 2011.

[74] J E Collins, S White, S M Searle, and D L Stemple. Incorporating RNA-seq data into the zebrafish Ensembl genebuild. *Genome Res*, 22:2067–2078, 2012.

[75] A M Bolger, M Lohse, and B Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30:2114–2120, 2014.

[76] B Langmead and S Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9:357–359, 2012.

[77] S Neph, J Vierstra, A B Stergachis, A P Reynolds, E Haugen, B Vernot, R E Thurman, S John, R Sandstrom, A K Johnson, and et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489:83–90, 2012.

[78] M Westerfield. *The zebrafish book. A guide for the laboratory use of zebrafish (Danio rerio), Publisher=University of Oregon Press* . Eugene, 4 edition, 2000.

[79] G Zhang, M Hussain, S L ONeill, and S Asgari. Wolbachia uses a host microRNA to regulate transcripts of a methyltransferase, contributing to dengue virus inhibition in Aedes aegypti. *Proceedings of the National Academy of Sciences*, 110:10276–10281, 2013.

[80] S Carbon, I Ireland, C J Mungall, S Q Shu, B Marshall, S Lewis, the AmiGO Hub, and the Web Presence Working Group. AmiGO: online access to ontology and annotation data. *Bioinformatics*, 25:288–289, 2008.

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 5

Declaration by candidate

In the case of Chapter 5, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Scripting, performed experiments, analysed data, wrote parts of the paper, made modifications to the manuscript as suggested by co-authors and the reviewers	45

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Megan Woolfit	Conceived and designed the experiments, analysed data, contributed reagents/materials/analysis tools, wrote parts of the paper	
Jonathan Keith	Conceived and designed the experiments, analysed data, contributed reagents/materials/analysis tools	
Elizabeth A McGraw	Analysed the data, contributed reagents/materials/ analysis tools, wrote parts of the paper	
Jean Popovici	Conceived and designed experiments, performed experiments, analysed data, wrote parts of the paper	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature	[redacted]	**Date** 9/10/15
Main Supervisor's Signature	[redacted]	**Date** 9/10/15

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 5

Discovery of Putative Small Non-Coding RNAs from the Obligate Intracellular Bacterium Wolbachia Pipientis

Chapter Objectives

This chapter addresses the second objective of this thesis, which is to identify putative functional non-coding genomic regions contributing to diseases. This work was carried out as part of the eradicate dengue project. Dengue fever is transmitted by the mosquito, *Aedes aegypti*. It has been revealed that the presence of *Wolbachia* in mosquitoes blocks the ability of the dengue virus to grow in mosquitoes and we hypothesise that small ncRNAs play a significant role in the biology of *Wolbachia*. To identify these elements and to understand how *Wolbachia* interacts with their hosts, two independent methods were used: (1) comparative genomics (by applying *changept*); and (2) using RNA-seq data. Use of a 16-character representation to encode the pair-wise alignment between two *Wolbachia* strains - *w*Mel and *w*Pip helped to clearly distinguish a large number of segment classes. This analysis revealed a number

of putative small ncRNAs that may play a significant role in reducing dengue virus transmission.

Authorship

Megan Woolfit[1], Manjula Algama[2], Jonathan M Keith[2], Elizabeth A McGraw[1], Jean Popovici[1,#]

1 School of Biological Sciences, Monash University, Clayton, VIC 3800, Australia

2 School of Mathematical Sciences, Monash University, Clayton, VIC 3800, Australia

Current address: Jean Popovici, Malaria Molecular Epidemiology Unit, Institut Pasteur in Cambodia, Phnom Penh, Cambodia

Reference

Woolfit M, Algama M, Keith JM, McGraw EA, Popovici J. (2015). Discovery of Putative Small Non-Coding RNAs from the Obligate Intracellular Bacterium Wolbachia pipientis. *PLoS ONE* 10(3): e0118595. doi:10.1371/journal.pone.0118595.

PLOS | ONE

RESEARCH ARTICLE

Discovery of Putative Small Non-Coding RNAs from the Obligate Intracellular Bacterium *Wolbachia pipientis*

Megan Woolfit[1], Manjula Algama[2], Jonathan M. Keith[2], Elizabeth A. McGraw[1], Jean Popovici[1¤]*

1 School of Biological Sciences, Monash University, Clayton, Victoria, Australia, **2** School of Mathematical Sciences, Monash University, Victoria, Australia

¤ Current address: Jean Popovici, Malaria Molecular Epidemiology Unit, Institut Pasteur in Cambodia, Phnom Penh, Cambodia
* [redacted]

Citation: Woolfit M, Algama M, Keith JM, McGraw EA, Popovici J (2015) Discovery of Putative Small Non-Coding RNAs from the Obligate Intracellular Bacterium *Wolbachia pipientis*. PLoS ONE 10(3): e0118595. doi:10.1371/journal.pone.0118595

Academic Editor: Kostas Bourtzis, International Atomic Energy Agency, AUSTRIA

Received: September 12, 2014

Accepted: January 21, 2015

Published: March 4, 2015

Data Availability Statement: The RNA-Seq sequence data have been deposited at NCBI under Bioproject PRJNA266744, sample numbers SAMN03174110, SAMN03174111, SAMN03174113, SAMN03174115 and SAMN03174116.

Funding: This work was supported by a grant from the National Health and Medical Research Council of Australia and by a grant from the Australian Research Council (DP1095849). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Wolbachia pipientis is an endosymbiotic bacterium that induces a wide range of effects in its insect hosts, including manipulation of reproduction and protection against pathogens. Little is known of the molecular mechanisms underlying the insect-*Wolbachia* interaction, though it is likely to be mediated via the secretion of proteins or other factors. There is an increasing amount of evidence that bacteria regulate many cellular processes, including secretion of virulence factors, using small non-coding RNAs (sRNAs), but sRNAs have not previously been described from *Wolbachia*. We have used two independent approaches, one based on comparative genomics and the other using RNA-Seq data generated for gene expression studies, to identify candidate sRNAs in *Wolbachia*. We experimentally characterized the expression of one of these candidates in four *Wolbachia* strains, and showed that it is differentially regulated in different host tissues and sexes. Given the roles played by sRNAs in other host-associated bacteria, the conservation of the candidate sRNAs between different *Wolbachia* strains, and the sex- and tissue-specific differential regulation we have identified, we hypothesise that sRNAs may play a significant role in the biology of *Wolbachia*, and in particular in its interactions with its host.

Introduction

Wolbachia pipientis is a vertically transmitted endosymbiotic Alphaproteobacteria that is thought to infect up to 40% of arthropod species [1]. Different *Wolbachia* strains induce a diverse range of effects in their hosts, including multiple forms of reproductive manipulation that enhance transmission of the endosymbiont to the next host generation [2,3]. More recently it has also been discovered that a number of *Wolbachia* strains inhibit the replication of viral and other pathogens in both their natural hosts, such as *Drosophila melanogaster*, and heterologous hosts such as *Aedes aegypti* [4–6]. These effects make *Wolbachia* an attractive biocontrol

Competing Interests: The authors have declared that no competing interests exist.

agent for vector-borne diseases, and field releases of *Wolbachia*-infected *A. aegypti* are currently being tested in trials with the aim of reducing dengue virus transmission [7].

The molecular mechanisms by which *Wolbachia* causes these different host phenotypes remain largely unknown. Recent work has demonstrated that *Wolbachia* infection modulates expression of mosquito host miRNAs that regulate diverse genetic targets, including host metalloprotease and methylase genes [8–10]. *Wolbachia* infection in other taxa has also been shown to affect transcription of host genes involved in iron metabolism and the oxidative stress response [11–13]. At least some host responses to *Wolbachia* infection are likely to be induced by effectors secreted by the endosymbiont. *Wolbachia* has a conserved and functional type IV secretion system (T4SS) [14], and these systems are known to play a role in infection, survival and proliferation in many other symbiotic and pathogenic intracellular prokaryotes [15]. *Wolbachia* genomes also contain an unusual number of genes encoding ankyrin domains. Host-interacting ankyrin proteins are secreted via the T4SS in other intracellular Alphaproteobacteria such as *Anaplasma phagocytophilum* and *Ehrlichia chaffeensis*, and these proteins are considered the most likely candidates to underlie the molecular dialogue between *Wolbachia* and its host [16–20].

Numerous *Wolbachia* genes, including those encoding ankyrin domains, show host sex- and tissue-specific expression patterns [21,22], further suggesting that they may be involved in host interaction. The mechanisms by which *Wolbachia* regulates the expression of these genes are currently unknown. Few transcription factors have been identified in *Wolbachia* genomes, and these factors have so far been shown to regulate only a small number of genes [23]. Recently, however, numerous other species of facultative or obligate intracellular bacteria have been shown to use small non-coding RNAs (sRNAs) to regulate the expression of genes associated with diverse aspects of host interaction, including iron homeostasis [24], the cell cycle [25], quorum-sensing [26], secretion systems [27] and secreted virulence factors [28–30]. These small RNAs are highly variable in sequence and function, and vary in number from a few tens to a few hundreds in many bacterial genomes [31].

There are at least five main classes of sRNAs, which regulate gene expression in several ways [31,32]. Antisense sRNAs are typically 50–500 nt in length, are transcribed from the opposite strand of the genes that they regulate, and act via extensive complementarity with their target mRNAs. Trans-encoded sRNAs, in contrast, are often shorter (around 100 nt), are usually encoded intergenically or with partial overlap of one or more CDSs, may regulate many different mRNAs, and have much more limited complementarity with their targets. Both antisense and trans-encoded sRNAs may interact with mRNA targets to enhance or inhibit translation. A third kind of sRNA, also encoded outside CDSs, are 5' riboswitches, which do not operate as independent transcripts but are part of the mRNA they regulate. Fourth, there are a small number of sRNAs, such as 6S sRNA, that interact with proteins rather than mRNA. Finally, bacteria also encode a number of 'housekeeping' sRNAs that do not pair with mRNAs or regulate proteins; these include the ribozyme RNase P, the 4.5S RNA component of the signal recognition peptide, and tmRNA. Genes encoding tmRNA, 4.5S sRNA, RNase P and 6S sRNA are present in *Wolbachia* genomes, and the latter two show host tissue-specific expression in filarial nematodes [21]. To our knowledge, however, no antisense or trans-encoded sRNAs have previously been identified in *Wolbachia* genomes.

The majority of trans-encoded sRNAs described to date are expressed under specific growth conditions [31], and this class of sRNA may therefore be of particular interest in elucidating host sex- and tissue-specific gene regulation in *Wolbachia*. In this study, two independent methods have allowed the identification of candidate trans-encoded sRNA in several *Wolbachia* strains. The first method is based on examination of RNA-Seq data from the *Wolbachia* strains *w*MelPop, *w*MelPop-CLA and *w*MelCS. To investigate the potential utility of RNA-Seq

for *Wolbachia* gene expression studies, we had previously performed a number of trial runs of this sequencing technology. These data were not ideal for detection of sRNAs, as we did not perform strand-specific sequencing and had chosen to sequence DNA fragments of ~300 nt, which is longer than many known sRNAs. Despite these limitations, however, we serendipitously identified a number of sRNA candidates while analysing the sequencing reads for other purposes. The second method we used to identify candidate sRNAs is bioinformatic, and based on comparative genomics of the strains *w*Mel and *w*Pip.

To increase the probability that the candidates identified using the methods above are true sRNAs, we have conservatively focused on transcripts that are encoded entirely within intergenic regions rather than overlapping a CDS, and that are transcribed specifically rather than as an intergenic component of a polycistronic mRNA. We identified several candidate sRNAs, and have experimentally confirmed the differential expression of one putative sRNA in four strains of *Wolbachia*, and in different host sexes and tissues.

Materials and Methods

Fly rearing and cell culture

Drosophila melanogaster (w^{1118}) stock lines stably infected with the *w*Mel, *w*MelPop, *w*MelCS and *w*Au strains of *Wolbachia* were maintained on standard molasses and cornmeal medium at a constant temperature of 25°C with a 12h light/dark cycle [33,34].

C6/36 cells infected with *w*MelPop-CLA were routinely passaged in RPMI 1640 medium supplemented with 10% FBS [35].

Sample preparation for RNA-Seq experiment

We performed RNA-Seq sequencing on five trial libraries. Three libraries were created using material from C6/36 cells infected with *w*MelPop-CLA, and two libraries were created from the heads of flies infected with either *w*MelPop or *w*MelCS. In an attempt to minimize the number of experimental manipulations that could affect the transcriptomic profile, we created the two fly libraries without performing either purification of *Wolbachia* from the host material or depletion of host or *Wolbachia* rRNA. For each of these libraries, we dissected the heads from 10 flies. *w*MelCS was obtained from *D. melanogaster* Canton S virgin female flies at 3 days of age, and *w*MelPop was obtained from *D. melanogaster* w^{1118} virgin female flies at 3 days of age. In each case, total RNA was isolated after homogenization of dissected heads in 100ul of Trizol (Invitrogen). RNA was then purified according to the manufacturer's instructions and DNase-treated (DNase I recombinant, Roche) before being sent for Illumina sequencing.

The three samples derived from *w*MelPop-CLA-infected cell culture were each subject to different treatment. For the first, total RNA was isolated from a 175cm^2 flask of *w*MelPop-CLA-infected C6/36 cells at ~80% confluence using Trizol according to the manufacturer's instructions, and the RNA was DNase treated and sent for Illumina sequencing. For the second sample, we extracted total RNA from a single flask of cell culture as above, while for the third sample, we purified *Wolbachia* from the cell culture using the method of Iturbe-Ormaetxe et al [36], then performed total RNA extraction. For both samples, RNA was DNase-treated and then depleted for host and bacterial rRNA using successively the RiboMinus Eukaryote kit (Ambion) and the MicrobExpress bacterial mRNA Enrichment kit (Ambion) according to the manufacturer's instructions. After depletion, first and second-strand cDNA synthesis was done using SuperScript III Reverse Transcriptase (Life Technologies) and DNA Polymerase I, Klenow Fragment (NEB) according to the manufacturer's instructions. cDNAs were purified using the MinElute Reaction Cleanup Kit (Qiagen) before being sent for Illumina sequencing.

The second and third cell culture samples (those with rRNA depletion) were indexed on a single lane of Illumina GAII, and sequenced at Micromon (Monash University), with 300 bp size selection and 75 bp paired-end reads. The remaining three samples were sent to Macrogen (South Korea) for library preparation and sequencing indexed on a single lane of HiSeq, with 300 bp size selection and 70 bp paired-end reads.

The RNA-Seq sequence data have been deposited at NCBI under Bioproject PRJNA266744, sample numbers SAMN03174110, SAMN03174111, SAMN03174113, SAMN03174115 and SAMN03174116.

Data analysis and mapping of RNA-Seq reads

We filtered reads for quality using Trimmomatic [37] by removing all trailing bases with quality less than 30, and then discarding (1) reads shorter than 40 nt after trimming and (2) any unpaired reads. We then performed read mapping and downstream analyses using the Nesoni toolset (http://www.vicbioinformatics.com/software.nesoni.shtml). Filtered paired reads were mapped to the reference *w*Mel genome [18] using BWA [38], and then mappings were filtered so that read pairs with multiple equally good alignments were randomly assigned to one of those alignments. We then created a modified *w*Mel gff file that listed intergenic regions as well as the more typical annotation features (CDSs, rRNAs, tRNAs, etc), and used a custom Perl script to count the alignments to each feature.

Because some intergenic regions are smaller than the mean fragment size sequenced, and because there appears to be a substantial amount of polycistronic transcription occuring in *w*Mel, many "intergenic" mapping counts actually reflect transcription of flanking genes. Nonetheless, these counts provided us with a preliminary list of candidate intergenic regions with high transcription levels. We then inspected the read mapping data for these candidate regions in the Artemis genome browser [39]. We identified regions that appeared to have intergenic-specific transcription, based on mapping of read pairs, for further investigation.

Change-point analysis: prediction of conserved candidate sRNAs

We used the program *changept* to identify a class of segments characterized by a high degree of conservation. The process followed in this analysis is described below.

Sequence and alignment of data. We used the published complete genome sequences of *w*Mel and *w*Pip (NCBI accession numbers NC_002978.6 and NC_010981.1 respectively) to identify intergenic regions that were highly conserved between these strains. Fragments of the genome may show low levels of divergence between strains for at least two reasons. The first possibility is that they are evolving under selective constraint, and these are the regions we wish to identify. Alternatively, however, genomic regions that were horizontally transferred between *w*Mel and *w*Pip after the divergence of these strains will also be more similar to one another than expected, not due to selection but because they have a more recent common ancestor than the rest of the genome. We took two approaches to exclude these regions. First, we masked prophage regions [40] and a known region of horizontal gene transfer (WD0507-WD0517 [41]) in the genomes before analysis. Secondly, we also performed a post hoc check for horizontal transfer after candidates were identified by extracting their nucleotide sequences from the *w*Mel genome and using them as megablast queries against the NCBI NT database. To attempt to exclude regions that have artifactually low levels of divergence due to recent horizontal transfer between supergroups, we accepted only those candidate regions that had better hits to all available A group genomes than to all available B group genomes. The *changept* procedure we used to identify these conserved non-CDSs is described below. We aligned the masked genomes using progressive Mauve [42], and used the accessory script stripSubsetLCBs (available

Table 1. *changept* 16-character code used for conversion of pairwise alignment of *w*Mel and *w*Pip.

*w*Mel	A	A	A	A	C	C	C	C	G	G	G	G	T	T	T	T
*w*Pip	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
Symbol	a	b	C	d	e	f	g	h	i	j	k	l	m	n	o	p

doi:10.1371/journal.pone.0118595.t001

from http://gel.ahabs.wisc.edu/mauve/snapshots/) to extract local colinearity blocks (aligned core genome blocks) at least 500 nt in length. We then used a custom script to convert this XMFA output file into Fasta format.

Data transformation. The pairwise alignment of *w*Mel and *w*Pip was then converted into a *changept* input sequence using a 16-character code (A = (a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p)), (Table 1). Insertions and deletions were excluded. The alignment blocks were separated using the '#' symbol; these are considered as fixed change-points by the model. In 16-character representation, characters 'a', 'f', 'k' and 'p' represent conserved bases. Although this sequence also contains other biologically significant information, such as the GC content of each species and transition/transversion ratio, we were mainly focused on the different levels of conservation between *w*Mel and *w*Pip.

Model selection. *Changept* currently requires the user to specify the number of segment classes. Separate segmentation analyses were performed for models with 1–12 segment classes. Each model was run for 1,000 iterations. Selecting the model with the most appropriate number of classes involved model selection criteria discussed in [43]. In summary, we calculated approximations to three information criteria: AIC, BIC and DIC (Figure A in S1 File). The model with the smallest information criterion is generally considered optimal. However a subjective judgment was made on which model to choose; models containing classes with very low mixture proportions were considered to be over-fitted and thus a model with fewer classes was selected. For this data, we selected the 7-class model. BIC increases with the number of classes, displaying an unusual behaviour (Panel B, Figure A in S1 File); therefore we based our decision on AIC and DICV (Panel A, Figure A in S1 File). The first local minimum of both these criteria occurred at seven classes.

The most conserved class. To determine the most conserved class of the selected model, the mean proportion of alignment matches was calculated for each iteration of the sampler:

$$E(\theta) = \frac{\theta_a + \theta_f + \theta_k + \theta_p}{\sum_{j \in A} \theta_j}$$

where $\theta_j; j \in A$ is the frequency of character j in the most conserved segment class.

These values were plotted against iteration number (Figure B in S1 File). This plot was also used as an added check to determine if the model had converged along with the loglikelihood plot (obtained by plotting the log likelihood at each iteration for the 1000 iterations).

Calculating profile values and generating wiggle track. We used the program *readcp* (part of the *changept* package) to calculate profile values for the most conserved segment class. The profile shows the probability that each sequence position belongs to the specified class. These posterior probabilities are estimated using the samples from the post burn-in phase (that is, the first 150 samples were discarded) by Monte Carlo integration. A complete description of the *changept* and *readcp* programs can be found in [44,45]. The *readcp* output is then used to generate a wiggle track (https://cgwb.nci.nih.gov/goldenPath/help/wiggle.html). This file was uploaded to the UCSC browser (http://microbes.ucsc.edu/) for viewing in order to identify highly conserved non-coding segments in *w*Mel and *w*Pip. A section of the wiggle track corresponding to one of the candidate sRNAs (*w*Mel coordinates 1,039,579–1,039,870) predicted by

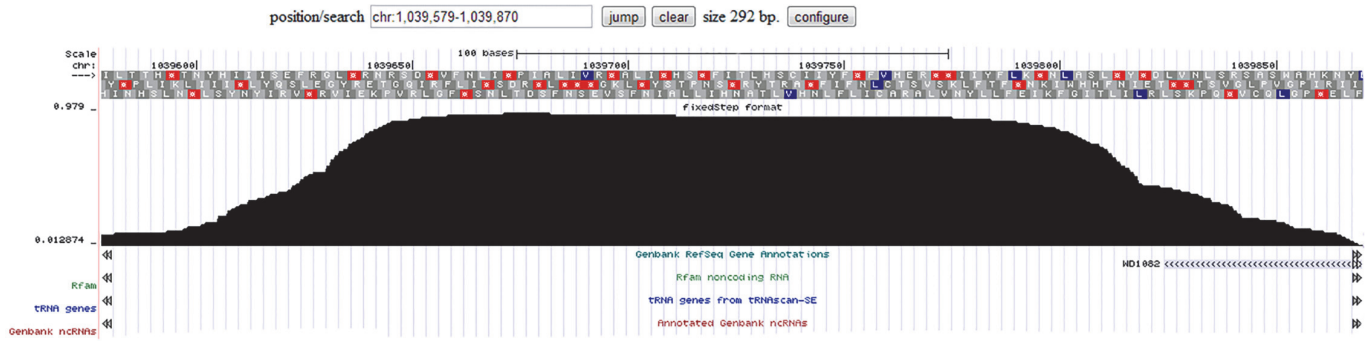


Fig 1. The WIG profile of a highly conserved feature. This profile shows the probability that each position in the region belongs to the most conserved class. The conserved non-coding region is positioned in wMel coordinates 1,039,579–1,039,870. The top 3 bars containing single letter amino acid codes show 3 possible protein translation phases. At the bottom, coding region WD1082 is shown (in light blue with arrows). There are no previously annotated non-coding regions corresponding to this region.

doi:10.1371/journal.pone.0118595.g001

the *changept* program is shown in Fig. 1. The non-CDs regions longer than 50 nt with profile value ≥ 0.5 are considered as most likely candidates.

Validation of 5′ ends of candidate sRNAs by Rapid Amplification of cDNA Ends

A Rapid Amplification of cDNA Ends (RACE) procedure was used to determine the 5' ends of a selection of candidate sRNAs [46]. The protocol used is based on the procedure described by [47]. Briefly, total RNA was isolated from whole bodies of ~100 *w*Mel-infected 1-day-old adult male w^{1118} flies using Trizol, DNAse-treated, and treated with Tobacco Acid Pyrophosphatase (Epicentre) for 30 min at 37°C. T4 RNA ligase (NEB) was used to add a short RNA adaptor on the 5' ends of RNA. Reverse transcription of RNA was performed using primers complementary for candidate sRNAs and Superscript III (Invitrogen). Primers and adaptor sequences are listed in Table A in S1 File. The resulting cDNA was used as template for PCR using candidate sRNA specific primers and RNA adaptor specific primers. PCR products were run on a 1% agarose gel, gel-extracted using the QIAEX II kit (QIAGEN), and cloned in pGEMTeasy vector (Promega). *E. coli* DH5α were transformed with the constructs and plasmids were purified using the QIAprep spin miniprep kit (QIAGEN) prior to sequencing using SP6 primer (Micromon, Monash University, Australia).

Verification of transcription of candidate sRNA within intergenic regions

In order to verify that expression in each intergenic region was due to the transcription of the candidate sRNA itself and not the result of transcription together with a downstream gene in a single RNA molecule, a set of RT-PCRs were done using primers overlapping genes downstream of the intergenic regions. These "overlapping RT-PCRs" were done to verify that the 3' end of a 5'RACE-confirmed candidate sRNA was indeed within the intergenic region. PCR was performed on cDNA from *w*Mel-infected w^{1118} 1 day-old female abdomens and ovaries. Briefly, total RNA was isolated from pools of 5 abdomens or pools of 10 dissected ovaries using Trizol. RNA was further purified according to Trizol instructions and DNase treated. cDNAs were synthesized from 1 μg of total RNA using random primers and SuperScript III, in accordance with the manufacturer's instructions. PCR was performed on cDNA and for every primer pair on DNA controls. PCR products were then visualised on a 2% agarose gel. All primers used are listed in Table A in S1 File.

RNA extraction, cDNA synthesis and quantitative PCR analysis of candidate sRNA

All RNA extractions were performed on 1-day-old flies. Three types of biological material were used for RNA extraction and further qRT-PCR analysis: whole bodies of male flies, abdomens of male and female flies, and dissected tissues (head, carcass, gonads) of male and female flies. Total RNA was isolated from individual whole bodies of male w^{1118} flies infected with *w*Mel, *w*MelCS, *w*MelPop or *w*Au strains (n = 15 individuals per line) using Trizol. Total RNA was isolated from pools of 5 abdomens of either male or female w^{1118} flies infected with *w*Mel, *w*MelCS, *w*MelPop or *w*Au strains (n = 15 pools per line) using Trizol. Total RNA was isolated from pools of dissected tissues of 10 male or female w^{1118} flies infected with *w*Mel (n = 12 pools per tissue per sex). Flies were ice-anaesthetized, then head, gonad and carcass were dissected in ice-cold PBS and immediately transferred into Trizol. All RNA samples were further purified according to Trizol instructions and DNase treated. cDNAs were synthesized from 1 μg of total RNA using random primers and SuperScript III, in accordance with the manufacturer's instructions. Expression of candidate sRNA in all samples was measured by qPCR using the LightCycler480 SYBR Green I Master (Roche) on a LightCycler480 II instrument (Roche) in duplicate on a 2–5 times dilution of the cDNAs. Primers are listed in Table A in S1 File. *Wolbachia* surface protein *wsp* expression was used as reference to normalize candidate sRNA expression and account for *Wolbachia* density [48]. On a subset of samples, to confirm differential expression, we compared the use of *wsp* as reference gene to *Wolbachia* 16S and *Drosophila melanogaster rps17*. Relative quantification of expression was calculated using the LightCycler480 II software. Significant differences in candidate sRNA expression were tested by Mann-Whitney *U* test using GraphPad Prism 5 software (GraphPad Software, San Diego, California USA).

Results

We used two approaches to identify candidate novel sRNAs in *Wolbachia*. For the first approach, we extracted RNA from *D. melanogaster* flies or *Aedes albopictus* C6/36 cell lines infected with *Wolbachia*, performed RNA-Seq, and mapped the resulting reads to the *w*Mel and host genomes (Table 2). As these sequencing runs were exploratory trials and our treatments (rRNA depletion and *Wolbachia* purification) were not replicated, we cannot draw any firm conclusions about the effects of each treatment on the quality or content of the resulting sequence data. It is clear, however, that some kind of purification, rRNA and/or host RNA depletion is necessary to obtain reasonable coverage of the *Wolbachia* transcriptome without wasted sequencing effort as shown in other studies [49,50]. Even though our RNA-Seq experiments were not designed for this purpose, we observed reads mapping to intergenic regions of *Wolbachia* (S1 Data). Given the variable coverage of the *Wolbachia* genome we obtained from the different RNA-Seq experiments, we have focused our work using the dataset with the highest number of reads mapping to *Wolbachia*: RNA extracted from *Wolbachia* in cell culture.

RNA-Seq reveals transcription from multiple *Wolbachia* intergenic regions

A large number of RNA-Seq reads from each sequencing library mapped to intergenic regions of the *w*Mel reference genome, as listed in S1 Data. However, many of these reads may derive not from independent intergenic transcripts, but rather from 5' or 3' UTRs, or the intergenic regions of polycistronic transcripts. We identified candidate sRNAs by selecting intergenic regions with high transcription levels, and then inspecting the mapping of paired reads in these

Table 2. Summary of the RNA-Seq data obtained.

Source material Treatment *Wolbachia* strain Sequencing instrument	Cells Purification and rRNA depletion wMelPop-CLA GAII	Cells rRNA depletion wMelPop-CLA GAII	Cells Untreated wMelPop-CLA HiSeq	Fly head Untreated wMelPop HiSeq	Fly head Untreated wMelCS HiSeq
Number of filtered reads	31,348,758	18,079,262	135,928,880	130,923,592	132,780,000
Number of mapped reads	27,244,266	12,094,331	119,609,644	129,883,646	131,671,602
Number mapping to *Wolbachia*	22,734,365	932,604	21,641,258	4,176,245	1,061,022
(% of mapped reads)	(83%)	(8%)	(18%)	(3%)	(1%)
Number mapping to host	4509901	11161727	97968386	125707401	130610580
(% of mapped reads)	(17%)	(92%)	(82%)	(97%)	(99%)
Number mapping to host rRNA	4276896	10822638	97535071	122477846	126476269
(% of mapped reads)	(16%)	(89%)	(82%)	(94%)	(96%)
(% of host reads)	(95%)	(97%)	(99.5%)	(97%)	(97%)
Among reads mapped to *Wolbachia*:					
16S	13%	9%	44%	44%	43%
23S-5S	79%	54%	50%	51%	53%
CDS	8%	37%	6%	5%	4%
% of CDS with $>$ 10 reads	95%	83%	94%	74%	49%

doi:10.1371/journal.pone.0118595.t002

regions. We excluded from further analysis transcribed intergenic regions in which one read of any pair mapped to the intergenic region and the other mapped to a flanking CDS, and focused only on those regions in which both ends of all read pairs mapped within the intergenic region. Candidate sRNAs presented here were identified using the data obtained from *w*MelPop-CLA in C6/36 cells.

Bioinformatic prediction of candidate conserved intergenic sRNAs

Our second, independent approach to identifying candidate sRNAs was based on comparative genomics. Previous work has demonstrated that, while some sRNAs are specific to a single bacterial strain or species, others show conservation of sequence across broader taxonomic distances [51,52]. If conserved intergenic sRNAs are undergoing purifying selection to maintain functionality, we would expect them to evolve more slowly than other intergenic regions that are not functionally constrained in this way. To search for such regions, we aligned the published genome sequences of the moderately divergent *Wolbachia* strains *w*Mel and *w*Pip (from the A and B supergroups of *Wolbachia*, respectively [18,53]), and used the program *changept* to identify highly conserved non-CDS regions. A full description of the *changept* model can be found in previous papers [54–56]. In summary, the algorithm takes as input a sequence of characters (which may represent pairwise or multiple alignments) and estimates positions (called change-points) that delineate homogeneous segments. Using the *changept* program, we identified a class of segments (Class 4 of the 7-Class model, Figure B in S1 File) characterized by the highest degree of conservation ($\sim$95% between *w*Mel and *w*Pip). Non-coding segments longer than 50 nt and with $\geq$ 0.5 probability of belonging to this class form the focus of our analysis. There are 42 non-CDS regions in the most conserved class (Table B in S1 File). These included the 16S, 23S and 5S rRNA genes, 17 tRNA genes, a recent pseudogene (WD0002), and the housekeeping sRNAs RNase P and tmRNA. We also identified 19 highly conserved intergenic regions (Table 3) with no previous annotation, which represented a preliminary list of candidate conserved sRNAs.

Table 3. Intergenic regions predicted by *changept* to be highly conserved.

Coordinates in *w*Mel genome	Upstream/downstream CDS	Length (nt)
44,380–44,468	dnaJ/tRNA-Arg-1	89
85,867–85,929	dprA/WD0093	63
279,526–279,619	WD0299/coxB	94
547,479–547,732	nuoD/WD0562	254
611,202–611,370	WD0625/WD0626	169
612,281–612,391	WD0626/WD0627	111
622,779–622,923	WD0632/WD0633	145
623,094–623,293	WD0632/WD0633	200
639,293–639,403	fabG/WD0651	111
719,048–719,171	WD0744/WD0745	99
723,861–724,026	WD0749/WD0750	166
764,459–764,871	WD0790/WD0791	413
768,936–768,988	rho/WD0796	53
850,067–850,142	WD0878/trx	76
932,596–932,693	WD0973/WD0974	98
940,039–940,142	nuoI/trmE	104
941,823–941,975	tRNA-Ser-2/WD0982	153
1,039,579–1,039,870	WD1081/WD1082	292
1,105,661–1,105,744	tRNA-Thr-2/mutM	84

doi:10.1371/journal.pone.0118595.t003

We then checked the read pairs mapping to these regions in the *w*MelPop-CLA RNA-Seq data, as described above. Reads were mapped to all but one of these intergenic regions, indicating that they were transcribed in this strain. However, only one of these intergenic regions showed evidence of specific transcription, rather than transcription as part of a unit with one or both flanking genes. This region (*w*Mel coordinates 1,039,579–1,039,870) was therefore selected for experimental validation together with the other candidate sRNAs identified using RNA-Seq data above.

This approach is limited to the comparison of strains with an intermediate level of divergence. We repeated the *changept* analysis comparing the genomes of two more closely related *Wolbachia* strains, *w*Mel and *w*Ri. These genomes have undergone only limited divergence, and we found that over 70% of the genome was assigned to the most highly conserved class of segments, greatly reducing the predictive power of the method. All highly conserved non-CDS regions identified in the *w*Mel-*w*Pip comparison were, however, also identified in the *w*Mel-*w*Ri comparison. At the other end of the taxonomic scale, we also attempted to repeat this analysis comparing the genomes of *w*Mel or *w*Pip with the more distantly related D group strain *w*Bm. Unfortunately, the genomes of these strains have undergone extensive rearrangement since their divergence, and too few genomic regions with conserved synteny and of sufficient length could be identified to perform the analysis.

Experimental validation of candidate sRNAs

For all subsequent experiments to investigate expression of our candidate sRNAs, we used *Wolbachia*-infected insects, rather than cell culture, to ensure that our results reflect the natural biology of the symbiont. As an initial step, we tested whether the candidate sRNAs identified by our two methods were transcribed specifically, rather than as part of a single transcript with a flanking gene, in *w*Mel in *D. melanogaster*. We first used a 5' RACE procedure to identify the 5'

Table 4. Intergenic regions (IGR) selected for 5'RACE experimentsa and name and position of the two putative*Wolbachia* small non-coding RNAs we identified.

Coordinates in *w*Mel genome	IGR size (bp)	Upstream/downstream CDS (IGR ID[b])	5' end coordinate[c]	3' end within IGR?[d]	Name of putative sRNA
67,695–68,395	700	WD0072/WD0073 (IG-60)	67,860	Yes	*ncrwmel01*
170,838–171,549	711	WD0187/mutS (IG-151)	NA	NA	
461,742–462,763	1021	WD0478/WD0480 (IG-292)	NA	NA	
527,661–528,615	954	hemC/sucB (IG-446)	NA	NA	
587,733–588,439	706	WD0609/WD0610 (IG-498)	587,739	No	
896,038–896,357	319	WD0931/WD0932 (IG-760)	NA	NA	
915,181–915,500	319	WD0955/WD0956 (IG-781)	NA	NA	
978,741–979,093	352	WD1015/WD1016 (IG-834)	NA	NA	
1,039,579–1,039,870	291	WD1081/WD1082 (IG-884)	1,039,620	Yes	*ncrwmel02*[e]
1,080,340–1,081,546	1206	tRNA-Arg-4/WD1131 (IG-921)	NA	NA	
1,189,867–1,190,417	550	WD1243/WD1244 (IG-1021)	1,190,012	No	
1,216,366–1,216,864	498	ispH/WD1275 (IG-1047)	1,216,539	NA	
1,217,297–1,217,892	595	WD1276/htpG (IG-1049)	NA	NA	

[a]Note that we did not demonstrate sRNA-like intergenic-specific transcription for most of these regions

[b]IGR ID as designed in the S1 Data

[c]Determined by 5'RACE

[d]Determined by RT-PCR with downstream CDS

[e]Also predicted by the bioinformatic approach.

doi:10.1371/journal.pone.0118595.t004

end of those candidate sRNA transcripts for which it was possible to design a combination of RACE primers specific to the intergenic region. In many cases this was not possible, due to the high levels of repetitive sequence in the *w*Mel genome. These transcripts were discarded as candidates for further analysis.

We performed the 5'RACE procedure on 13 candidate sRNAs. Of these, five amplified successfully, and sequencing of the resulting plasmid showed that the 5' end of each RNA was indeed within the intergenic region (Table 4). The sequences of the plasmids are provided in Table C in S1 File. The other regions did not amplify, which could be due to multiple factors: no expression in the given biological conditions (age or tissue for example), expression below our detection limit, or expression from the opposite DNA strand as we designed all the primers on the positive strand only for this preliminary analysis. These non-amplifying regions were not considered further as candidate sRNAs.

In addition to performing 5'RACE, 'overlapping RT-PCR' was done to verify that the 3' end of each of the candidate sRNAs was also within the intergenic region. Of the five regions with a 5' end confirmed by 5'RACE to be within an intergenic region, it was possible to design specific RT primers for four. Two regions could be amplified using a forward intergenic primer and reverse downstream gene primer, indicating that the transcription of these regions occurred as a single RNA molecule with a flanking gene. These two intergenic regions might contain sRNAs transcribed as part of an operon, but could alternatively be 5' UTR regions, and so were excluded from further analysis. However, two regions showed no amplification when RT-PCR was performed using reverse primers in the downstream flanking gene, while amplification occurred using forward and reverse primers that both bound to the intergenic region. These results demonstrate that the transcription of RNA from these regions begins and ends in the intergenic region, and these two RNA molecules can consequently be considered to be

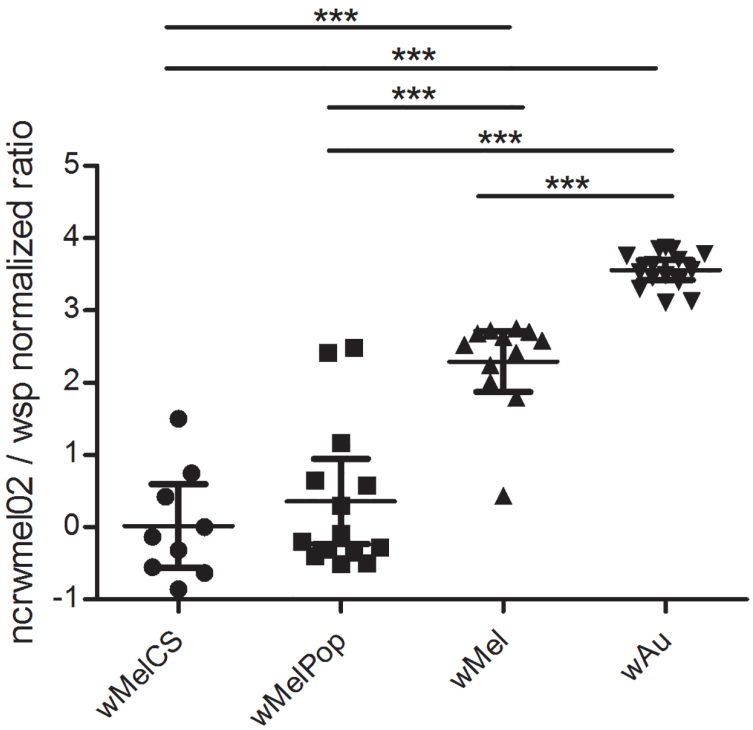


Fig 2. Expression of *ncrwmel02* in four *Wolbachia* strains from the whole body of 1-day-old male *D. melanogaster*. Expression (mean ± 95% CI) normalized to *wsp* expression (Mann-Whitney *U* test ** $p < 0.01$, **** $p < 0.0001$)

doi:10.1371/journal.pone.0118595.g002

intergenic sRNAs. We labeled them non-coding RNA *Wolbachia* wMel 01 and 02 (*ncrwmel01* and *ncrwmel02*). The putative sRNA *ncrwmel02* is a highly conserved intergenic region that was predicted by both RNA-Seq and comparative genomics approaches (Table 4).

Putative sRNA shows sequence conservation but differential transcript levels across *Wolbachia* strains

We then used qPCR to test for differences in the expression of *ncrwmel02* in four different *Wolbachia* strains (*w*Mel, *w*MelPop, *w*MelCS and *w*Au) in the whole body of 1 day-old male *D. melanogaster* flies. We selected this putative sRNA because its sequence is conserved (Figure C in S1 File), and we were able to design specific qPCR primers for the intended template and successfully amplify cDNA from all strains, while we could not for *ncrwmel01*.

The expression of *ncrwmel02* was normalized against the expression of the *Wolbachia* Surface Protein encoding gene, *wsp*, to account for differences in *Wolbachia* density between the strains. *ncrwmel02* was expressed at a relatively low level in *w*MelCS and *w*MelPop, but approximately twice as highly in *w*Mel, and seven-fold more highly in *w*Au (Fig. 2).

Wolbachia putative sRNA expression is differentially regulated in host tissues and sexes

In order to assess whether the expression of this *Wolbachia* putative sRNA is constitutive or regulated, we analysed its expression in different tissues of male and female flies. First, *ncrwmel02* expression was compared in the abdomens of male and female 1-day-old flies for

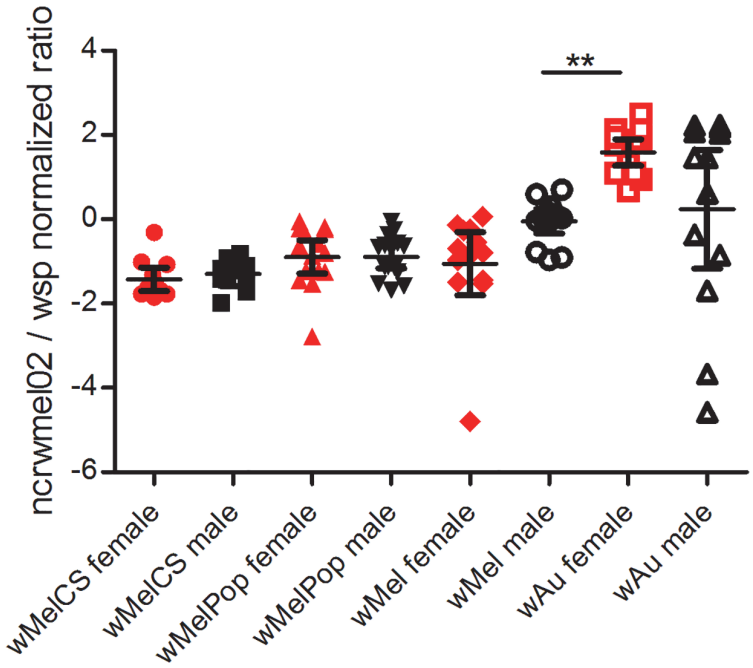


Fig 3. Expression of *ncrwmel02* in four *Wolbachia* strains from abdomens of 1-day old male (black) or female (red) *D. melanogaster*. Expression (mean ± 95% CI) normalized to *wsp* expression (Mann-Whitney *U* test, ** $p < 0.01$)

doi:10.1371/journal.pone.0118595.g003

the four *Wolbachia* strains (Fig. 3). We observed only one significant difference in *ncrwmel02* expression in these tissues: *ncrwmel02* in *w*Mel was more highly expressed in male than in female abdomens, demonstrating that in some conditions its expression is differentially regulated.

Because *ncrwmel02* expression in whole abdomens might not reflect the level of regulation occurring in specific *Wolbachia*-infected tissues, we performed a second experiment in which its expression was analyzed in dissected tissues (gonads, head and carcasses) of male and female flies infected with the *w*Mel strain (Fig. 4). In contrast to our observation of generally stable *ncrwmel02* expression in male and female whole abdomens, its expression in dissected tissues showed clear evidence of differential expression.

The greatest differences were observed between gonads (ovaries and testes), which have very different physiological activity and regulation. Expression was significantly upregulated in testes compared to ovaries with more than ten-fold difference. For body parts that are expected to be more similar in terms of regulation and activity between males and females, such as heads and carcasses, no differential regulation of *ncrwmel02* in different host sexes was observed (Fig. 4). In addition, we also observed differential expression of *ncrwmel02* when comparing different tissues in the same sex. For example, expression was upregulated in female carcasses compared to female gonads, and upregulated in male gonads compared to male heads. To validate the differential expression we observed on those dissected tissues we compared the use of *wsp*, *Wolbachia* 16S and *Drosophila melanogaster rps17* as reference genes. Whatever the gene used as reference, all the patterns of expression remain the same and all differential expression remains significant (Figure D in S1 File).

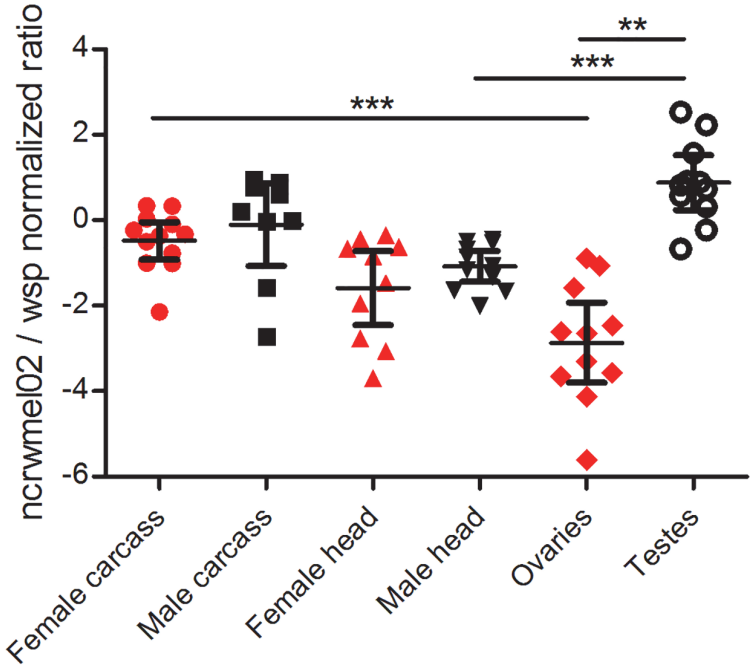


Fig 4. Expression of *ncrwmel02* in the *w*Mel strain from gonads, heads and carcasses of 1-day-old male (black) or female (red) *D. melanogaster*. Expression (mean ± 95% CI) normalized to *wsp* expression (Mann-Whitney *U* test, ** $p < 0.01$ *** $p < 0.001$)

doi:10.1371/journal.pone.0118595.g004

In total, these expression experiments show that the *Wolbachia* putative sRNA *ncrwme02* is expressed in the four *Wolbachia* strains in a regulated pattern that differs according to the sex of the host and the tissue in which the bacterium is localized.

Discussion

Recent research has begun to uncover the critical roles played by small RNAs in the regulation of cellular processes ranging from highly conserved housekeeping functions to rapid responses to environmental or host cues. Most research on sRNAs to date has focused on free-living or facultatively intracellular bacteria. Yet we might expect that obligately host-associated bacteria would rely on sRNAs at least as much as, if not more than, free-living bacteria, for two reasons. First, sRNAs offer bacteria a flexible and rapidly adaptable mode of gene regulation that may be ideally suited to the constantly changing co-adaptive interplay between host and symbiont. Secondly, many endosymbiotic bacteria, including *Wolbachia*, have undergone at least some degree of genome reduction, often resulting in the loss of genes encoding canonical transcriptional regulatory proteins [18]. Intergenic regulatory regions associated with sRNAs show evidence of retention and conservation even in some of the most reduced endosymbiotic genomes [57], indicating that sRNA-based regulation may remain necessary and be under sufficient selection to resist loss via genome reduction.

We have identified two novel putative sRNAs in *Wolbachia* genomes, using two independent methods that are likely to detect different subsets of sRNAs. The comparative genomics approach we used could detect sRNA candidates that are conserved at the nucleotide level across strains from A and B supergroups of *Wolbachia*, regardless of the conditions under which they are expressed. It would not, however, identify candidates that are not present in

both strains, have originated since the divergence of the supergroups, or are conserved at the level of secondary structure rather than nucleotide sequence. In contrast, RNA-Seq data could be used to detect these latter classes of sRNAs, but would not be able to identify sRNAs if they were not expressed under the experimental conditions used to generate the data. The two putative sRNAs characterized here were identified by our analysis of *w*MelPop-CLA RNA-Seq data, while only *ncrwmel02* was predicted using comparative genomics. This probably reflects the level of sequence conservation of these putative sRNAs: when used as a blastN query against the NCBI NT and WGS databases, *ncrwmel02* has longer hits, with higher percentage sequence identity, to a broader range of other *Wolbachia* strains (from the A, B, C and D supergroups), than *ncrwmel01*.

We showed *ncrwmel02* was present and transcribed in the four A group strains we used to experimentally characterize the expression of this putative sRNA. The strains *w*Mel, *w*MelCS and *w*MelPop are closely related [58], naturally infect *D. melanogaster*, and all induce host cytoplasmic incompatibility (CI), the most frequently observed type of reproductive manipulation caused by *Wolbachia*. *w*Mel and *w*MelCS are otherwise benign, but *w*MelPop is pathogenic, causing its adult hosts to die prematurely. In contrast, *w*Au is somewhat more distantly related [59], infects *D. simulans*, and does not cause CI. All four strains were placed into the same genetic background (*D. melanogaster* w^{1118}) for these analyses, to limit the effects of different host species on sRNA expression patterns.

Almost all pairwise strain comparisons are significantly different at the whole body level for *ncrwmel02* expression; most strikingly, it is substantially more highly expressed in *w*Au than in the three other strains. Host sex-specific differences in expression in *w*Mel become apparent at the tissue level. *ncrwmel02* is more highly expressed in testes than in ovaries, but does not show evidence of male-specific upregulation in other tissues. The significant upregulation of *ncrwmel02* in testes compared to ovaries suggests that sRNAs might possibly be involved in some aspect of host reproductive manipulation, although the data we provide here only suggest this hypothesis and future experimental demonstration would be required on a range of CI and non-CI inducing strains. CI involves *Wolbachia*-induced modification of sperm in infected hosts [2], and increased transcription of sRNAs may play a role in that process. More generally, given the regulatory roles of sRNAs in other bacterial species, differential expression of sRNA could contribute to a range of phenotypic differences between strains.

Developing a full understanding of the roles of sRNA in *Wolbachia* will require not only searching for additional sRNAs and characterizing their expression in different strains and host tissues, but identifying the targets of these molecules. Many of these targets, whether genes, mRNA or proteins, are expected to be of *Wolbachia* origin, but it is also possible that *Wolbachia* sRNAs could directly target host gene expression. Although secretion of functional bacterial sRNAs into eukaryotic host cells has not been observed, it is a possibility worth considering [60]. Viral sRNAs are known to target host genes [61], and bacterial sRNAs may have the same ability. In addition, it has already been shown that interplay occurs between *Wolbachia* and its insect host via eukaryote miRNA [8,9], and that the virus-blocking phenotype induced by *Wolbachia* involves changes in the expression of host miRNA [10]. It is possible that *Wolbachia*-induced phenotypes such as dengue inhibition may occur as part of a molecular dialogue between the bacterial endosymbiont and its eukaryotic host involving uni- or bi-directional gene regulation by small non-coding RNAs.

Conclusions

Considering (a) the fundamental roles played by sRNA in other bacteria, especially in quorum-sensing, pathogenesis and virulence, (b) the conservation of *ncrwmel02* between different

Wolbachia strains, and (c) the strain-, sex- and tissue-specific differential regulation of *ncrwmel02* expression, we hypothesize that sRNAs may play significant roles in the biology of *Wolbachia*. The analyses described here are preliminary and had limited power, and the two putative sRNAs we have identified are likely to represent only the largest, most highly expressed and/or conserved sRNAs in *Wolbachia* genomes. Additional RNA-Seq experiments with different size selection of RNA, bacterial purification, host and rRNA depletion [49,50], and strand-specific library preparation might allow the identification of many more of these molecules, and further research will be required to assess the roles of sRNAs in the insect-*Wolbachia* interaction. Nonetheless, the descriptive work presented here opens a new path in understanding the molecular mechanisms underlying the complex and diverse range of phenotypes induced by *Wolbachia* within its host.

Supporting Information

S1 Data. RNA-Seq reads mapping to *w*Mel features. Reads mapping to CDS and intergenic regions are indicated for all 5 RNA-Seq experiments.
(XLS)

S1 File. Figure A, Selection of optimal number of classes. We used approximations to the well-known information criteria AIC, BIC and DIC to identify the number of distinct classes of conservation levels. Generally, a lower value of the information criteria indicates a better model. BIC favoured a 1-class model, which is inappropriate. We therefore based our judgement on AIC and DICV and selected the 7-class model as the first local minimum of AIC and DICV has occurred at seven classes. **Figure B, Identifying the most conserved class**. The mean proportion of alignment matches was plotted against each iteration of the sampler to identify the class that contains the most conserved segments in *w*Mel and *w*Pip (Class 4). The different colours represent different classes in the 7-class model. **Figure C, Sequence alignment of the *ncrwmel02* amplicon from the published genome data of *w*Mel [18], *w*MelCS, *w*MelPop [58] and *w*Au [62]. Figure D, Validation of *ncrwmel02* differential expression observed using *wsp* as reference gene in dissected tissues of *w*Mel-infected male (black) or female (red) *D. melanogaster***. *ncrwmel02* expression calculated using *wsp*, 16S or *rps17* is represented for the three significant differential expression observed using *wsp*. Expression (mean ± 95% CI) normalized to *wsp*, 16S or *rps17* expression (Mann-Whitney U test, * $p < 0.1$, ** $p < 0.01$ *** $p < 0.001$). Panel A: *ncrwmel02* expression in male and female gonads. Panel B: *ncrwmel02* expression in female dissected tissues. Panel C: *ncrwmel02* expression in male dissected tissues. **Table A, Oligonucleotides used in this study. Table B, Highly conserved non-coding region predicted by changept**. Thresholds used: 1. Conservation = 0.95 (conservation level of the most conserved class-Class 4); 2. Profile value $\geq$0.5 (probability that each position in the conserved feature belongs to Class 4); 3. Length >50 nt (length of the conserved feature). **Table C, 5' RACE of intergenic regions (IGR) plasmid sequences**. Insert in pGEMTeasy in bold.
(DOC)

Acknowledgments

We thank Thomas Walker and Iñaki Iturbe-Ormaetxe for helpful discussion about the manuscript, Jyotika Taneja de Bruyne for technical help with fly rearing, and Yi Dong and Alison Carrasco for laboratory assistance.

Author Contributions

Conceived and designed the experiments: MW JK JP. Performed the experiments: MA JP. Analyzed the data: MW MA JK EAM JP. Contributed reagents/materials/analysis tools: MW JK EAM. Wrote the paper: MW MA JK EAM JP.

References

1. Zug R, Hammerstein P (2012) Still a host of hosts for Wolbachia: analysis of recent data suggests that 40% of terrestrial arthropod species are infected. PLoS One 7: e38544. doi: 10.1371/journal.pone.0038544 PMID: 22685581
2. Werren JH, Baldo L, Clark ME (2008) Wolbachia: master manipulators of invertebrate biology. Nature Reviews Microbiology 6: 741–751. doi: 10.1038/nrmicro1969 PMID: 18794912
3. Hoffmann A, Turelli M (1997) Cytoplasmic incompatibility in insects. In: O'Neill SL, Hoffmann A, Werren JH, editors. Influential passengers Inherited microorganisms and arthropod reproduction. Oxford: Oxford University Press. pp. 42–80.
4. Hedges LM, Brownlie JC, O'Neill SL, Johnson KN (2008) Wolbachia and Virus Protection in Insects. Science 322: 702–702. doi: 10.1126/science.1162418 PMID: 18974344
5. Teixeira L, Ferreira A, Ashburner M (2008) The Bacterial Symbiont Wolbachia Induces Resistance to RNA Viral Infections in Drosophila melanogaster. Plos Biology 6: 2753–2763.
6. Moreira LA, Iturbe-Ormaetxe I, Jeffery JAL, Lu G, Pyke AT, et al. (2009) A Wolbachia symbiont in Aedes aegypti limits infection with dengue, Chikungunya and Plasmodium Cell 139: 1268–1278. doi: 10.1016/j.cell.2009.11.042 PMID: 20064373
7. Hoffmann A, Montgomery B, Popovici J, Iturbe-Ormaetxe I, Johnson P, et al. (2011) Successful establishment of Wolbachia in Aedes populations to suppress dengue transmission. Nature 476: 454–457. doi: 10.1038/nature10356 PMID: 21866160
8. Osei-Amo S, Hussain M, O'Neill SL, Asgari S (2012) Wolbachia-Induced aae-miR-12 miRNA Negatively Regulates the Expression of MCT1 and MCM6 Genes in Wolbachia-Infected Mosquito Cell Line. PLoS One 7: e50049. doi: 10.1371/journal.pone.0050049 PMID: 23166816
9. Hussain M, Frentiu FD, Moreira LA, O'Neill SL, Asgari S (2011) Wolbachia uses host microRNAs to manipulate host gene expression and facilitate colonization of the dengue vector Aedes aegypti. Proceedings of the National Academy of Sciences 108: 9250–9255. doi: 10.1073/pnas.1105469108 PMID: 21576469
10. Zhang G, Hussain M, O'Neill SL, Asgari S (2013) Wolbachia uses a host microRNA to regulate transcripts of a methyltransferase, contributing to dengue virus inhibition in Aedes aegypti. Proceedings of the National Academy of Sciences 110: 10276–10281. doi: 10.1073/pnas.1303603110 PMID: 23733960
11. Brennan LJ, Keddie BA, Braig HR, Harris HL (2008) The Endosymbiont <italic>Wolbachia pipientis</italic> Induces the Expression of Host Antioxidant Proteins in an <italic>Aedes albopictus</italic> Cell Line. PLoS ONE 3: e2083. doi: 10.1371/journal.pone.0002083 PMID: 18461124
12. Kremer N, Voronin D, Charif D, Mavingui P, Mollereau B, et al. (2009) *Wolbachia* Interferes with Ferritin Expression and Iron Metabolism in Insects. PLoS Pathog 5: e1000630. doi: 10.1371/journal.ppat.1000630 PMID: 19851452
13. Kremer N, Charif D, Henri H, Gavory F, Wincker P, et al. (2012) Influence of Wolbachia on host gene expression in an obligatory symbiosis. BMC Microbiology 12: S7. doi: 10.1186/1471-2180-12-S1-S7 PMID: 22376153
14. Rances E, Voronin D, Tran-Van V, Mavingui P (2008) Genetic and functional characterization of the type IV secretion system in Wolbachia. Journal of Bacteriology 190: 5020–5030. doi: 10.1128/JB.00377-08 PMID: 18502862
15. Voth DE, Broederdorf LJ, Graham JG (2012) Bacterial Type IV secretion systems: versatile virulence machines. Future microbiology 7: 241–257. doi: 10.2217/fmb.11.150 PMID: 22324993
16. Siozios S, Ioannidis P, Klasson L, Andersson SG, Braig HR, et al. (2013) The Diversity and Evolution of Wolbachia Ankyrin Repeat Domain Genes. PLoS One 8: e55390. doi: 10.1371/journal.pone.0055390 PMID: 23390535
17. Walker T, Klasson L, Sebaihia M, Sanders MJ, Thomson NR, et al. (2007) Ankyrin repeat domain-encoding genes in the wPip strain of Wolbachia from the Culex pipiens group. Bmc Biology 5.
18. Wu M, Sun LV, Vamathevan J, Riegler M, Deboy R, et al. (2004) Phylogenomics of the reproductive parasite Wolbachia pipientis wMel: A streamlined genome overrun by mobile genetic elements. Plos Biology 2: 327–341.

19. Iturbe-Ormaetxe I, Burke GR, Riegler M, O'Neill SL (2005) Distribution, expression, and motif variability of ankyrin domain genes in Wolbachia pipientis. Journal of Bacteriology 187: 5136–5145. PMID: 16030207
20. Rikihisa Y, Lin M, Niu H (2010) Microreview: Type IV secretion in the obligatory intracellular bacterium Anaplasma phagocytophilum. Cellular Microbiology 12: 1213–1221. doi: 10.1111/j.1462-5822.2010.01500.x PMID: 20670295
21. Darby AC, Armstrong SD, Bah GS, Kaur G, Hughes MA, et al. (2012) Analysis of gene expression from the Wolbachia genome of a filarial nematode supports both metabolic and defensive roles within the symbiosis. Genome Research 22: 2467–2477. doi: 10.1101/gr.138420.112 PMID: 22919073
22. Papafotiou G, Oehler S, Savakis C, Bourtzis K (2011) Regulation of Wolbachia ankyrin domain encoding genes in Drosophila gonads. Research in Microbiology 162: 764–772. doi: 10.1016/j.resmic.2011.06.012 PMID: 21726632
23. Li Z, Carlow CKS (2012) Characterization of Transcription Factors That Regulate the Type IV Secretion System and Riboflavin Biosynthesis in <italic>Wolbachia</italic> of <italic>Brugia malayi</italic>. PLoS ONE 7: e51597. doi: 10.1371/journal.pone.0051597 PMID: 23251587
24. Boughammoura A, Matzanke BF, Böttger L, Reverchon S, Lesuisse E, et al. (2008) Differential role of ferritins in iron metabolism and virulence of the plant-pathogenic bacterium Erwinia chrysanthemi 3937. Journal of Bacteriology 190: 1518–1530. doi: 10.1128/JB.01640-07 PMID: 18165304
25. Grieshaber NA, Grieshaber SS, Fischer ER, Hackstadt T (2006) A small RNA inhibits translation of the histone-like protein Hc1 in Chlamydia trachomatis. Molecular Microbiology 59: 541–550. PMID: 16390448
26. Bejerano-Sagie M, Xavier KB (2007) The role of small RNAs in quorum sensing. Current Opinion in Microbiology 10: 189–198. PMID: 17387037
27. Murphy ER, Payne SM (2007) RyhB, an Iron-Responsive Small RNA Molecule, Regulates Shigella dysenteriae Virulence. Infection and Immunity 75: 3470–3477. PMID: 17438026
28. Podkaminski D, Vogel J (2010) Small RNAs promote mRNA stability to activate the synthesis of virulence factors. Molecular Microbiology 78: 1327–1331. doi: 10.1111/j.1365-2958.2010.07428.x PMID: 21143308
29. Bradley ES, Bodi K, Ismail AM, Camilli A (2011) A genome-wide approach to discovery of small RNAs involved in regulation of virulence in Vibrio cholerae. Plos Pathogens 7: e1002126. doi: 10.1371/journal.ppat.1002126 PMID: 21779167
30. Koo JT, Alleyne TM, Schiano CA, Jafari N, Lathem WW (2011) Global discovery of small RNAs in Yersinia pseudotuberculosis identifies Yersinia-specific small, noncoding RNAs required for virulence. Proceedings of the National Academy of Sciences 108: E709–E717. doi: 10.1073/pnas.1101655108 PMID: 21876162
31. Gottesman S, Storz G (2011) Bacterial Small RNA Regulators: Versatile Roles and Rapidly Evolving Variations. Cold Spring Harbor Perspectives in Biology 3.
32. Storz G, Vogel J, Wassarman Karen M Regulation by Small RNAs in Bacteria: Expanding Frontiers. Molecular Cell 43: 880–891. doi: 10.1016/j.molcel.2011.08.022 PMID: 21925377
33. Yamada R, Iturbe-Ormaetxe I, Brownlie JC, O'Neill SL (2011) Functional test of the influence of Wolbachia genes on cytoplasmic incompatibility expression in Drosophila melanogaster. Insect Molecular Biology 20: 75–85. doi: 10.1111/j.1365-2583.2010.01042.x PMID: 20854481
34. Min KT, Benzer S (1997) Wolbachia, normally a symbiont of Drosophila, can be virulent, causing degeneration and early death. Proceedings of the National Academy of Sciences of the United States of America 94: 10792–10796. PMID: 9380712
35. Frentiu FD, Robinson J, Young PR, McGraw EA, O'Neill SL (2010) Wolbachia-Mediated Resistance to Dengue Virus Infection and Death at the Cellular Level. PLoS One 5: e13398. doi: 10.1371/journal.pone.0013398 PMID: 20976219
36. Iturbe-Ormaetxe I, Woolfit M, Rances E, Duplouy A, O'Neill SL (2011) A simple protocol to obtain highly pure Wolbachia endosymbiont DNA for genome sequencing. J Microbiol Methods 84: 134–136. doi: 10.1016/j.mimet.2010.10.019 PMID: 21047535
37. Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, et al. (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Research 40: W622–W627. doi: 10.1093/nar/gks540 PMID: 22684630
38. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754–1760. doi: 10.1093/bioinformatics/btp324 PMID: 19451168
39. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA (2012) Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics 28: 464–469. doi: 10.1093/bioinformatics/btr703 PMID: 22199388

40. Shelby KS, Popham HJR (2007) Increased plasma selenium levels correlate with elevated resistance of Heliothis virescens larvae against baculovirus infection. Journal of Invertebrate Pathology 95: 77–83. PMID: 17316679

41. Woolfit M, Iturbe-Ormaetxe I, McGraw EA, O'Neill SL (2009) An Ancient Horizontal Gene Transfer between Mosquito and the Endosymbiotic Bacterium Wolbachia pipientis. Molecular Biology and Evolution 26: 367–374. doi: 10.1093/molbev/msn253 PMID: 18988686

42. Darling ACE, Mau B, Blattner FR, Perna NT (2004) Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Research 14: 1394–1403. PMID: 15231754

43. Oldmeadow C, Keith JM (2011) Model selection in Bayesian segmentation of multiple DNA alignments. Bioinformatics 27: 604–610. doi: 10.1093/bioinformatics/btq716 PMID: 21208984

44. Keith J (2008) Sequence Segmentation. In: Keith J, editor. Bioinformatics: Humana Press. pp. 207–229.

45. Algama M, Keith JM (2014) Investigating genomic structure using changept: A Bayesian segmentation model. Computational and Structural Biotechnology Journal 10: 107–115. doi: 10.1016/j.csbj.2014.08.003 PMID: 25349679

46. Frohman MA, Dush MK, Martin GR (1988) Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proceedings of the National Academy of Sciences 85: 8998–9002. PMID: 2461560

47. Gerhart E, Wagner H, Vogel J (2008) Approaches to Identify Novel Non-messenger RNAs in Bacteria and to Investigate their Biological Functions: Functional Analysis of Identified Non-mRNAs. Handbook of RNA biochemistry: 614–642.

48. Walker T, Klasson L, Sebaihia M, Sanders MJ, Thomson NR, et al. (2007) Ankyrin repeat domain-encoding genes in the wPip strain of Wolbachia from the Culex pipiens group. BMC Biol 5: 39. PMID: 17883830

49. Darby AC, Christina Gill A, Armstrong SD, Hartley CS, Xia D, et al. (2014) Integrated transcriptomic and proteomic analysis of the global response of Wolbachia to doxycycline-induced stress. ISME J 8: 925–937. doi: 10.1038/ismej.2013.192 PMID: 24152719

50. Kumar N, Creasy T, Sun Y, Flowers M, Tallon LJ, et al. (2012) Efficient subtraction of insect rRNA prior to transcriptome analysis of Wolbachia-Drosophila lateral gene transfer. BMC research notes 5: 230. doi: 10.1186/1756-0500-5-230 PMID: 22583543

51. Albrecht M, Sharma CM, Dittrich MT, Muller T, Reinhardt R, et al. (2011) The transcriptional landscape of Chlamydia pneumoniae. Genome Biol 12: R98. doi: 10.1186/gb-2011-12-10-r98 PMID: 21989159

52. Skippington E, Ragan MA (2012) Evolutionary Dynamics of Small RNAs in 27 Escherichia coli and Shigella Genomes. Genome Biology and Evolution 4: 330–345. doi: 10.1093/gbe/evs001 PMID: 22223756

53. Klasson L, Walker T, Sebaihia M, Sanders MJ, Quail MA, et al. (2008) Genome evolution of Wolbachia strain wPip from the Culex pipiens group. Molecular Biology and Evolution 25: 1877–1887. doi: 10.1093/molbev/msn133 PMID: 18550617

54. Keith J, Kroese D, Bryant D (2004) A Generalized Markov Sampler. Methodology And Computing In Applied Probability 6: 29–53.

55. Keith JM, Adams P, Stephen S, Mattick JS (2008) Delineating slowly and rapidly evolving fractions of the Drosophila genome. Journal of Computational Biology 15: 407–430. doi: 10.1089/cmb.2007.0173 PMID: 18435570

56. Keith JM (2006) Segmenting eukaryotic genomes with the generalized Gibbs sampler. Journal of Computational Biology 13: 1369–1383. PMID: 17037964

57. Degnan PH, Ochman H, Moran NA (2011) Sequence conservation and functional constraint on intergenic spacers in reduced genomes of the obligate symbiont Buchnera. PLoS genetics 7: e1002252. doi: 10.1371/journal.pgen.1002252 PMID: 21912528

58. Woolfit M, Iturbe-Ormaetxe I, Brownlie JC, Walker T, Riegler M, et al. (2013) Genomic evolution of the pathogenic Wolbachia strain, wMelPop. Genome Biol Evol 5: 2189–2204. doi: 10.1093/gbe/evt169 PMID: 24190075

59. Paraskevopoulos C, Bordenstein SR, Wernegreen JJ, Werren JH, Bourtzis K (2006) Toward a Wolbachia multilocus sequence typing system: Discrimination of Wolbachia strains present in Drosophila species. Current Microbiology 53: 388–395. PMID: 17036209

60. Storz G, Vogel J, Wassarman KM (2011) Regulation by small RNAs in bacteria: expanding frontiers. Molecular Cell 43: 880–891. doi: 10.1016/j.molcel.2011.08.022 PMID: 21925377

61. Smith NA, Eamens AL, Wang M-B (2011) Viral small interfering RNAs target host genes to mediate disease symptoms in plants. PLoS pathogens 7: e1002022. doi: 10.1371/journal.ppat.1002022 PMID: 21573142

62. Sutton E, Harris S, Parkhill J, Sinkins S (2014) Comparative genome analysis of Wolbachia strain wAu. BMC Genomics 15: 928. doi: 10.1186/1471-2164-15-928 PMID: 25341639

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 6

Declaration by candidate

In the case of Chapter 6, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Method developments, scripting, performed experiments, analysed data, wrote parts of the paper	50

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Anders Gonçalves da Silva	Conceived idea and designed the experiments, analysed data, wrote parts of the paper	
Edward Tasker	Analysed data	
Jonathan M Keith	Conceived and designed the experiments, contributed reagents/materials/analysis tools, provided guidance in interpretation	
Paul Sunnucks	Conceived idea, provided guidance	
Rohan H. Clarke	Provided guidance	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature	[redacted]	**Date** 9/10/15
Main Supervisor's Signature	[redacted]	**Date** 9/10/15

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 6

Host-Generalism in Blood Parasites: a Case for Reversible Host-Specialization

Chapter Objectives

It is well known that malaria causes enormous human loss every year. There are many species of malaria parasites, and hence this group is ideally suited for comparative analyses. Many malaria parasite species occur in a single vertebrate species. From the parasites' perspective, the shared evolutionary history (monophyly of Haemoproteidae) that has honed its ability to use haemoglobin as a fundamental resource, coupled with complex and divergent vertebrate immune systems that work to preclude access to the resource, suggests that the genomes of these creatures are likely to be a mosaic of highly conserved and divergent regions that reflect this tug-of-war. Thus a simple, two-category, classification of genome segments (conserved, divergent) seems inadequate, naive and too simplistic in such a system. In this chapter, using the *changept* model, I performed a comparative analysis of three *Plasmodium* genomes: (1) *P. falciparum*,

which infects humans; (2) *P. reichenowi*, which infects chimpanzees; and (3) *P. gallinaceum*, which infects jungle fowls (it is the only *Plasmodium* genome available for a species that infects something other than mammals). The *changept* application allows a rich categorisation of the genome, transforming the linear three-way alignment of nucleotides, into a high-dimensional detailed description of the mosaic that is the genome. The goal of this work was to use new *changept* methods to better understand the malaria disease by focusing on the distribution of conservation levels in DNA and look for signatures of a shared ancestral mechanism that might explain how malaria parasites cope with host uncertainty.

Authorship

Anders Gonalves da Silva[1+], Manjula Algama[2], Edward Tasker[2], Paul Sunnucks[1], Jonathan M Keith[2], Rohan H. Clarke[1]

1 School of Biological Sciences, Monash University, Clayton, VIC 3800, Australia

2 School of Mathematical Sciences, Monash University, Clayton, VIC 3800, Australia

+ Corresponding author

Reference

Gonalves da Silva A, Algama M, Tasker E, Sunnucks P, Keith JM, Clarke RH. (2015). Host-generalism in blood parasites: a case for reversible host-specialization. In preparation as a letter to *Nature* .

6.1 Introduction

Eukaryotic parasites that require multiple hosts in order to complete their life-cycle have limited control over which host species they will infect next [1]. This is especially true of Haemosporidians; Apicomplexan Protozoans (single-cell Eukaryotes) that include *Plasmodium* species causing malaria in humans, other mammals, birds and reptiles [2]. In the *Plasmodium* life-cycle, uncertainty about the host environment arises twice: once when the parasite moves from mosquito to vertebrate host, and again when it moves from vertebrate host to mosquito [2]. This puts pressure on the parasite to reduce uncertainty by: (1) manipulating its current host to increase the chance of a particular transmission pathway, and/or; (2) increasing the range of its tolerances through the evolution of *host-generalism* [3, 4]. In *Plasmodium*, there is abundant evidence for frequent hosts [5–10], and thus for host-generalism. However, it has its limitations, as switching between bird/reptile and mammal hosts is rare [11, 12]. We do not know what is the mechanism that drives this characteristic, but it is believed that such a mechanism would generate substantial genetic diversity in order to evade the myriad host immune systems [13]. The variable surface antigen (VSA) gene families of mammalian *Plasmodium* [14] are likely candidates. In *P. falciparum*, the *vir* gene family is at the core of the parasite′s ability to rapidly generate large amounts of genetic variation, making *falciparum* malaria one of the most formidable challenges to public health officials. However, while the *P. falciparum vir* genes share common characteristics with other mammalian *Plasmodium* VSA gene families, they do not share a common evolutionary origin [15], and equivalent gene families have not yet been found in *Plasmodium* that infect vertebrates other than mammals. Thus, it remains unclear if the mammalian *Plasmodium* VSA gene families are a remarkable case of convergent evolution, or if there is a single ancestral root that spans all *Plasmodium* that might indicate a universal mechanism underlying host-generalism in this group of parasites. We posit that while the VSA gene families provide the raw genetic material necessary for host-generalism, the crucial unifying component rests on the parasites'

ability to control the expression of the genetic variation stored in the VSA genes [14]. Here, we present comparative genomic analysis that supports our hypothesis that the mechanism controlling the expression of variable surface antigen is ancestral, and occurs in at least one *Plasmodium* species that does not infect mammals. We thus propose that host-generalism in *Plasmodium* is achieved through preferentially expressing copies of VSA genes that are best suited for the individual host, a strategy that *P. falciparum* has perfected to continually evade individual human immune systems.

6.2 Results

The ability to generate high levels of genetic variation is thought to be crucial to host-generalism [3]. In *P. falciparum*, the variable surface antigens are one component of an immune evasion strategy that not only generates large amounts of genetic variation, but also carefully controls how this variation is expressed. The strategy has three components: (1) multiple copies of the VSA genes spread throughout the genome [16]; (2) continuous generation of new variation through recombination among gene copies whilst still in the vertebrate host [17]; and, (3) fine epigenetic control of gene expression that ensures a single copy is expressed in any given cell [14, 18, 19]. The last step, in particular, ensures that at least some individual *P. falciparum* in a population of otherwise genetically nearly identical *P. falciparum* are able to elude the host's immune system [20, 21]. The variants that successfully elude the immune system are maintained, and allow the parasite to complete its life-cycle [22, 23]. We propose that *P. falciparum*′s strategy is a candidate mechanism for how *Plasmodium* in general, and other Haemosporidae, are able to frequently host-switch. We call this strategy *reversible instant host-specialisation*, where the parasite generates sufficient genetic variation to almost instantaneously adapt to a particular host's immune system.

At the frontline of this immune evasion strategy are the variable surface antigens, coded by large gene families, often with >100 copies across the *Plasmodium* genome. Considerable effort has been expended in identifying these genes in mammalian *Plasmodium* [15], and comparative work suggests that at least some VSA gene families may have separate evolutionary origins [24]. Searches for genes that match mammalian *Plasmodium* VSA gene families in the currently available non-mammalian *Plasmodium* genome have not yielded any hits. However, this is not entirely surprising. Even within mammalian *Plasmodium*, high sequence variation among closely related gene families makes it hard to identify homology [15], and there is no evidence to suggest that currently identified mammalian *Plasmodium* VSA gene families have a common ancestor that predates the switch to mammalian hosts [24, 25]. The apparent lack of VSA gene families in non-mammalian *Plasmodium*, coupled with the inferred difficulty *Plasmodium* experienced when first switching to mammalian hosts [11, 12] suggest that perhaps the VSA gene family are part of a strategy that evolved after the switch to mammalian hosts. However, the strength of *P. falciparum*′s strategy lies not only in the amount of variation it can produce, but also in how the variation is selectively expressed [18, 19, 23].

We performed an *in silico* gene capture experiment in order to compare *P. falciparum* (human), *P. reichenowi* (chimpanzee), and *P. gallinaceaum* (jungle fowl). Our comparative analysis sought to identify genome-scale patterns of functional evolutionary conservation and divergence that might explain why host-switching between mammals and non-mammals is rare. Most importantly, however, we focused our efforts on measuring the conservation levels at genes that make up *P. falciparum*′s gene expression machinery described in [26] of which many underlie the carefully controlled expression of VSAs. Our premise is that a high degree of conservation at these genes constitutes indirect evidence suggesting that *P. gallinaceum* and more generally *Plasmodium* that infect birds and reptiles has a similar immune evasion strategy, however the VSA gene families might be completely novel to those in mammalian *Plasmodium*.

Our *in silico* experiment proceeded in two steps with the aim of identifying shared genomic segments across the three species, and classifying the segments along a gradient of functional evolutionary divergence. In our first step, we obtained a three-way, whole-genome alignment of *P. falciparum* [16], *P. reichenowi* [27], and *P. gallinaceum* (Sanger Institute). *P. falciparum* has the best-curated *Plasmodium* genome; *P. reichenowi* is the closest known relative to *P. falciparum*; and, *P. gallinaceum* is the only non-mammal-infecting *Plasmodium* for which there is appreciable genomic data. For our analyses, we kept alignment blocks that included all three species and were at least 100 bases in length, totaling over 4 million bases (approximately $1/5^{th}$ of the *P. falciparum* genome). These alignment blocks represent shared genomic regions across the three species. In the second step, we aimed to partition the shared genomic regions into segments grouped into classes that shared similar levels of evolutionary divergence. A segment is a portion of the three-way alignment assigned to a particular segment class. Segments that fall within annotated genes of the *P. falciparum* genome were further characterized by the Gene Ontology (GO) terms [28] annotated to the gene. This was done by concatenating the alignment blocks, and applying a Bayesian segmentation model [29]. In total, the model identified 18 segment classes. Segments classes were then ranked in order of functional evolutionary conservation. The ranking and validation of our ranking are described in the methods.

Homology across the three genomes was almost exclusively confined to coding regions (96.1% of segments overlapped with an annotated coding region of the *P. falciparum* genome, $N = 13,818$). As expected, very few of the segments overlapped with VSA gene families. In particular, there were no segments that overlapped with genes belonging to the VSA gene families (*var, rfin, stevor*, and *Pfmc-2TM*) largely responsible for *P. falciparum* pathogenesis [30–33]. While *var* and *Pfmc-2TM* are thought to be exclusive to the *P. falciparum*/*P. reichenowi* clade [27], *rfin* and *stevor* are found in all mammalian *Plasmodium* investigated thus far [15, 27]. This corroborates previous results suggesting that mammalian-like VSA genes do not likely

exist in *P. gallinaceum* [15]. On the other hand, we see substantial overlap of segments with genes associated with key biological processes, implying these processes have been conserved following the switch to mammalian hosts. The alignment blocks included 48.8% of the *P. falciparum* annotated genes (2,688 of the 5,507 annotated nuclear genes). These spanned 85% ($N = 1,672$) of GO terms currently annotated to the *P. falciparum* genome. Across the Biological Processes Gene Ontology terms, segments were found in 79% of genes annotated to *P. falciparum*′s metabolic process, 60% in protein phosphorylation, 60% in translation, 59% in transport, and 59% in regulation of transcription (Table D.1). This is in contrast to genes annotated to biological processes directly associated with *P. falciparum*′s ability to evade host immune systems and invade host cells. Here, segments were found in only 6% of the genes annotated to pathogenesis, >1% to antigenic variation, and in none of the genes annotated to mediation of erythrocyte aggregation (Table D.1). The great majority of genes annotated to these GO terms belong to the mammalian VSA gene families.

In spite of the relatively high proportion of segments found across many of *Plasmodium*′s core biological processes, there is substantial variation in proportion of segments from the different classes that overlap with genes of these core processes (Figure 6.1, Table 6.1). Our results show that segments overlapping genes associated with transcription and translation were generally from evolutionarily divergent classes, while those associated with metabolism, DNA replication, and ubiquitin-dependent protein catabolism are mostly associated with evolutionarily conserved classes (Table 6.1). Given the importance of transcription to the carefully regulated expression of the VSA genes this result did not support our hypothesis. However, a closer look at the 128 genes of the 202 [26] thought to be involved in controlling gene expression in *P. falciparum* for which there were overlapping segments reveals an important detail: genes involved in the fundamental processes associated with the control of expression of *vir* gene family in *P. falciparum* were predominantly associated with segments from classes ranked as evolutionarily conserved; while genes involved in controlling the

parasites developmental cycle and general timing of expression were predominantly associated with segments from classes ranked as most evolutionarily divergent.

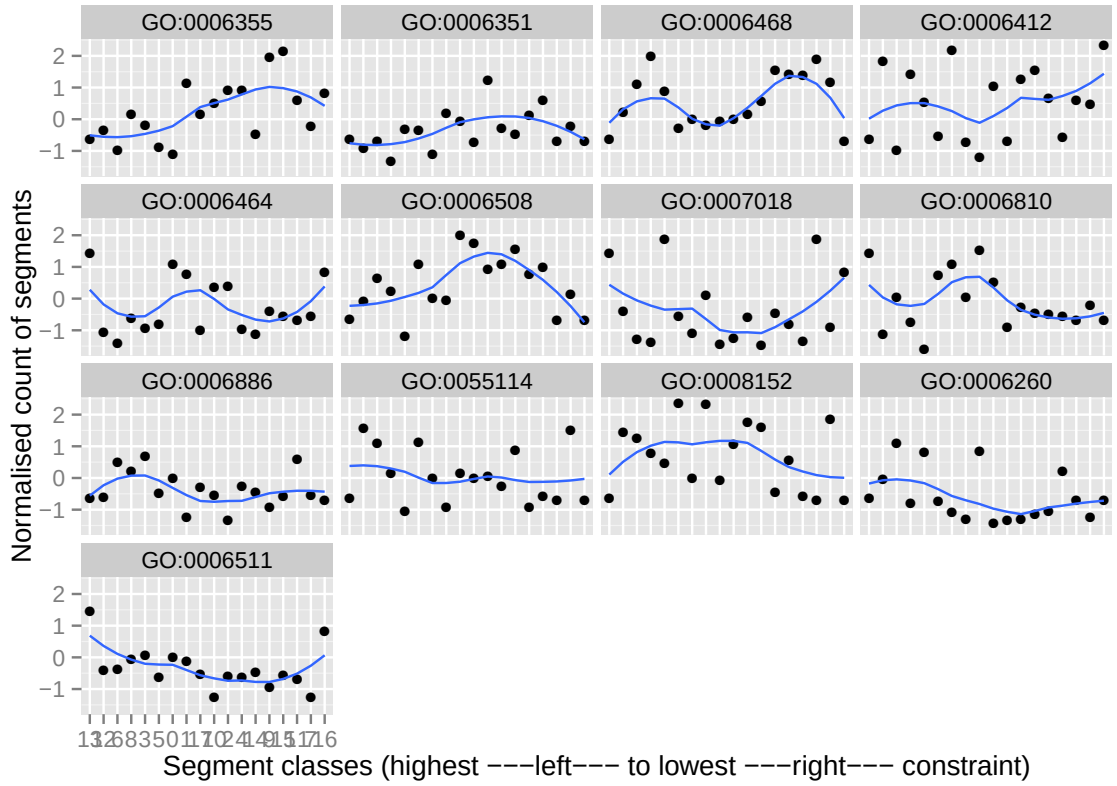


Figure 6.1: *Scatter plot for all Biological Process GO terms with >200 overlapping segments of segment class rank along a functional evolutionary divergence gradient (left to right from least to most divergent along the x-axis) and z-scores for counts of segments within segment classes (y-axis). A loess line was fitted to explore for monotonic increase/decrease of segment class rank and z-score.*

In particular, we found that the fundamental process of chromatin remodelling and modification that drives the peculiar mode of *vir* gene expression in *P. falciparum* (reviewed in [34, 35]) is well represented (39 of 55 histone modifying genes), and are slightly enriched for highly functionally constrained segments (Spearman rank r = -0.36; single tail $p = 0.071$; Figure 6.2). The histones themselves are poorly represented (1 of 8), but this is likely an artefact of the low coverage of *P. gallinaceum* data [36]. This is in contrast to the shared genes in the *Plasmodium*-specific transcription factors (48 out 73 genes shared), where there was an enrichment for segment classes ranked as most evolutionarily divergent (Spearman rank $r = 0.56$; single tail $p = 0.008$; Figure

Table 6.1: *Spearman rank correlation results for all Biological Process GO terms with >200 overlapping segments*

GO	Description	r	p-value
GO:0006355	Regulation Of Transcription DNA-templated	0.564	0.016
GO:0006351	Transcription DNA-templated	0.348	0.158
GO:0006468	Protein Phosphorylation	0.226	0.366
GO:0006412	Translation	0.191	0.446
GO:0006464	Cellular Protein Modification Process	0.158	0.530
GO:0006508	Proteolysis	0.129	0.609
GO:0007018	Microtubule-based Movement	-0.015	0.948
GO:0006810	Transport	-0.172	0.487
GO:0006886	Intracellular Protein Transport	-0.207	0.403
GO:0055114	Oxidation-reduction Process	-0.234	0.344
GO:0008152	Metabolic Process	-0.240	0.331
GO:0006260	DNA Replication	-0.325	0.185
GO:0006511	Ubiquitin-dependent Protein Catabolic Process	-0.490	0.040

r:Spearman Rank correlation between segment class functional evolutionary divergence rank and segment class z-scores *(normalized and standardized counts of segments in each segment class) for each GO term. P-values correspond to two-sided hypothesis test under the null hypothesis of no association between two variables.*

6.2). The *Plasmodium*-specific transcription factors include the ApiAP2 family of transcription factors thought to control the parasite′s developmental cycle [37], and speculated to be responsible for differences between *P. falciparum* and *P. vivax* [38]. Furthermore, disruption of gene expression can lead to severe cases of the disease [39], highlighting its importance to successful immune evasion and infection.

6.3 Conclusion

Our comparative analysis indicates that non-mammalian *Plasmodium* have the machinery to control the expression of VSA gene families. This suggests that *P. gallinaceum* has the capacity to employ a similar immune evasion strategy to that described in *P. falciparum*, and is therefore likely to be an ancestral strategy shared by mammalian and non-mammalian *Plasmodium* alike. We call this strategy *reversible host-specialisation*. The lack of VSA family-like genes in *P. gallinaceum* could be a function of the low

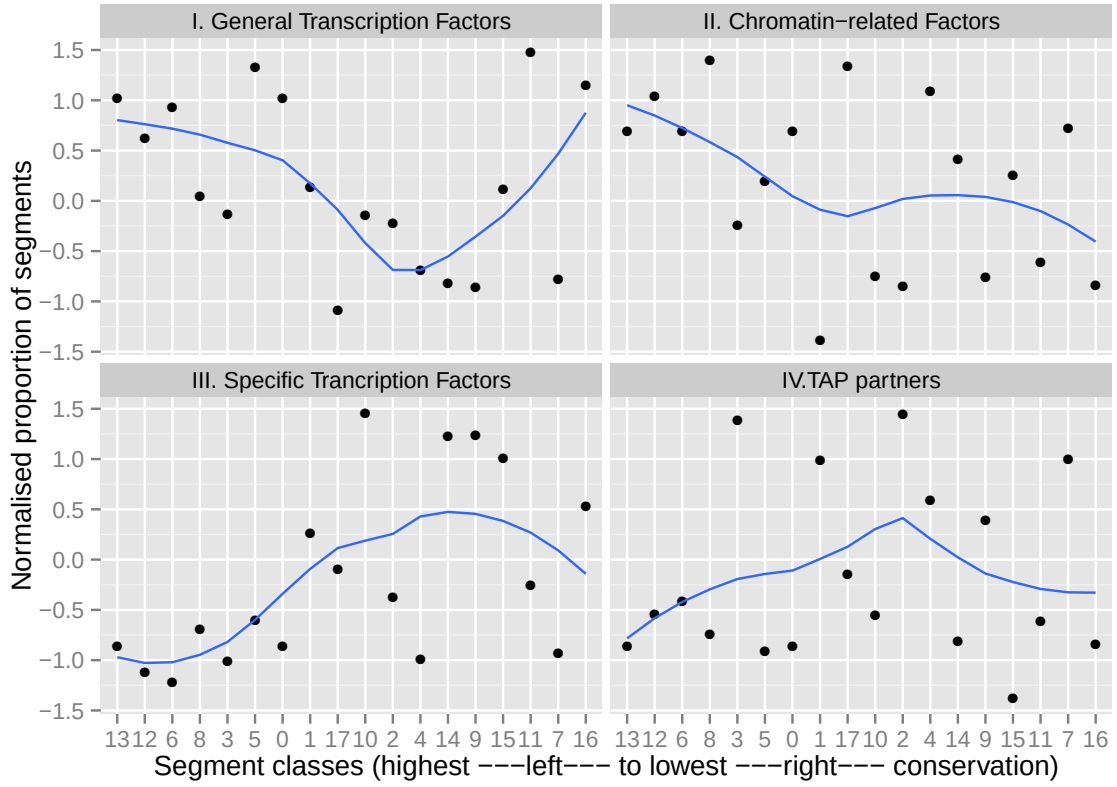


Figure 6.2: *Scatter plot for the four groups of transcriptionally related genes defined in [26] of segment class rank along a functional evolutionary divergence gradient (left to right from least to most divergent along the x-axis) and z-scores for counts of segments within segment classes (y-axis). A loess line was fitted to explore for monotonic increase/decrease of segment class rank and z-score.*

coverage of the available genomic data, but could also be that it neither has homology nor does it share a common ancestor with any of the currently described mammalian VSA gene families. In *P. falciparum*, the strategy is suited to evading individually variable human immune systems, and the long-history of humans and *P. falciparum* [40] does not suggest it facilitates switching to other vertebrate host species. However, the mechanism carries the hallmark characteristics that would be required to support a host-generalist life strategy, and we propose *P. falciparum*′s mechanism evolved from a host-generalist ancestor. On the other hand, our analysis also indicates why a switch to mammalian hosts from reptile/avian hosts has been rare. In particular, it suggests the timing of gene expression throughout the parasite′s life cycle might be an important constraint to host-switching. Sequencing the genome of *P. relictum*, an avian P*lasmodium* that infects over 100 host species, and exploring its transcriptome across

multiple hosts, would give us further insight into the evolution of the host-generalism, and the constraints to host-switching in Haemosporida.

6.4 Methods

6.4.1 Whole genome alignment

Whole genome alignment was performed using three *Plasmodium* species: *P. falciparum*, *P. reichenowi* and *P. gallinaceum* (abbreviated *Pf*, *Pr*, and *Pg* in the remainder of the document). Summary information on the data is provided in Table D.2. Genomic data in chromosome scaffolds were available for *Pf* (Assembly Version 3) and *Pr* (version available online on September 2013). The *Pf* genome was first published by [16], and the *Pr* genome has been recently published [27]. Three different *Pg* assemblies, build from 3X coverage Sanger sequencing data, were available in September 2013. Without additional data to choose among the different assemblies, we chose the one with the most nucleotides (assembly labelled: *P_gallinaceum.phusion_supercontigs.180705*).

The whole genome alignment suggests that *Pf* and *Pr* are 23% divergent, while *Pg* is >62% divergent to both *Pf* and *Pr* (Table D.3). The alignment between *Pf* and *Pr* suggests high synteny between these two species, as is expected from previous work [27]. The alignment between *Pg* and the mammal *Plasmodium* suggests that most of the alignment blocks seen between *Pf* and *Pr* are represented in *Pg* (Figure 6.3). However, a large portion of the *Pg* sequence data did **not** align simultaneously to *Pf* and *Pr* genomes. As we see, the majority of three-way aligned sequences map to coding regions.

6.4.2 Locally collinear blocks (LCBs)

We used the program *stripSubsetLCBs* (`http://darlinglab.org/mauve/snapshots/2015/2015-01-09/linux-x64/`) to output locally collinear blocks (LCBs) of a specified

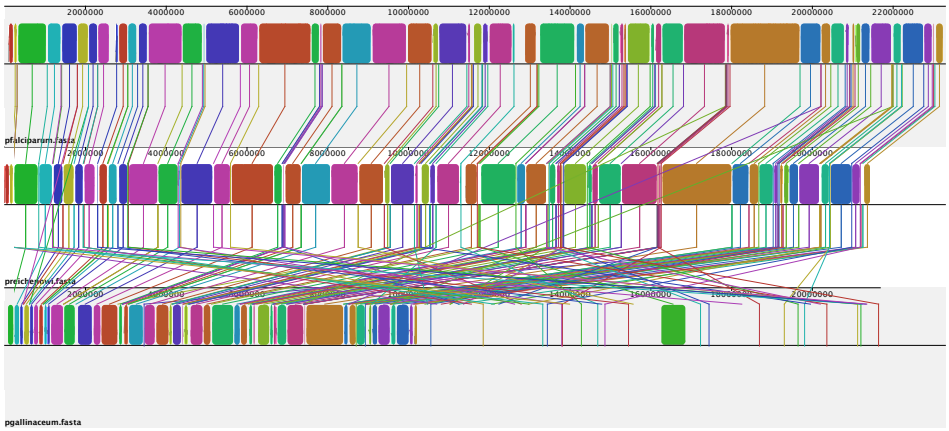


Figure 6.3: *Three way whole genome alignment between Pf, Pr, and Pg produced using progressiveMauve. Species are ordered from top to bottom as: Pf, Pr, and Pg.*

minimum length. We examined LCBs of minimum length 100, 200, 300, 400, and 500 nucleotides. The total number of nucleotides covered over the three species decreased with the increasing minimum LCB length (Table D.4 and Figure 6.4). At a minimum length of 100 nucleotides, approximately 18% of the *Pf* genome is covered by the three-way alignment. This reduces to approximately 11% by increasing the minimum length to 500 nucleotides. Because protein domains and other functional elements can be small, we settled on LCBs with a minimum 100 nucleotides. This value maximizes genome coverage in our analysis, but still guarantees a low likelihood of spurious alignments.

6.4.3 Bayesian genome segmentation

We applied the *changept* model [29, 41] to classify our three-way alignment of *Pf*, *Pr*, and *Pg* into classes of segments with distinct levels of functional evolutionary divergence. Here, the two dimensional three-way alignment was transformed into a single string using a 32 character redundant alphabet that encodes each possible permutation of four nucleotides taken in a group of three (Figure 6.5).

The model is implemented as a Bayesian hierarchical model that estimates the posterior probability of each position in the alignment belonging to each possible class. Posterior

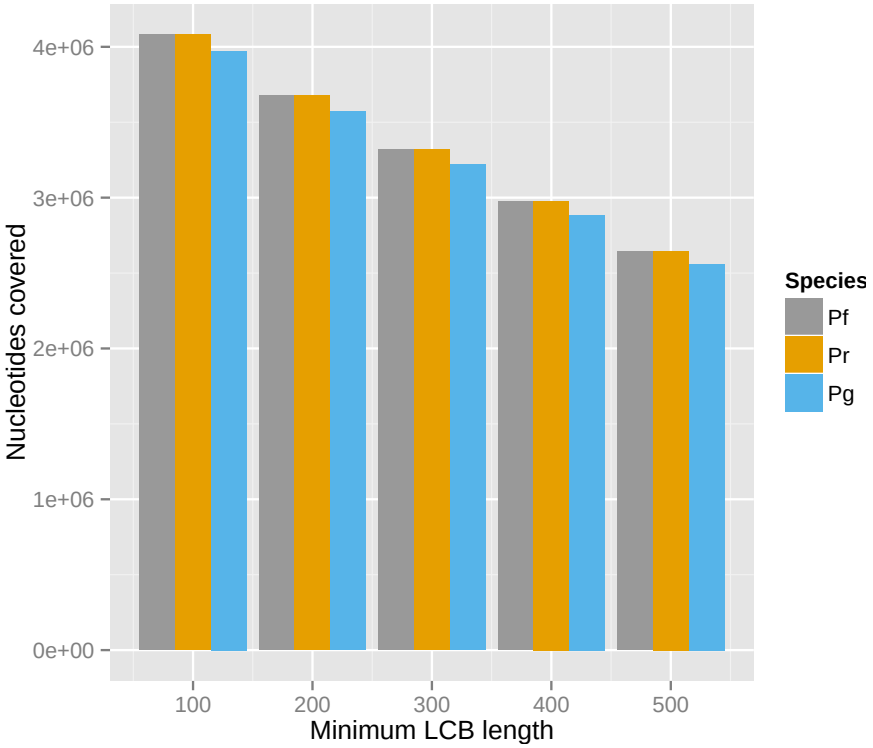


Figure 6.4: *Change in nucleotides covered by three way alignment across LCBs of different minimum lengths.*

distributions were estimated using an MCMC approach with Gibbs sampling [41]. A total of 650 samples were taken from the posterior, with an additional 350 samples being discarded as burn-in. We accepted that a particular position belonged to a certain class k if the $\mathrm{P}(K = k|\text{data}) > 0.5$. The total number of classes K is fixed, thus the model was run separately for K ranging from 1 to 25. The model with *K=18* had the lowest type V Deviance Information Criterion (DICV) [42], and also had segments across all classes and displayed class stability [43] (Figure 6.6). Thus, we accepted *K=18* as the model with the best fit to the data.

6.4.4 Characterizing segments

Summary characteristics of each segment class inferred by the *changept* model with K = 18 classes is shown in Table D.5. Using a MySQL database, *P. falciparum* genome annotations were cross referenced with segment class identity. As shown elsewhere [29, 44], segment classes inferred with this method can exhibit distinct functional

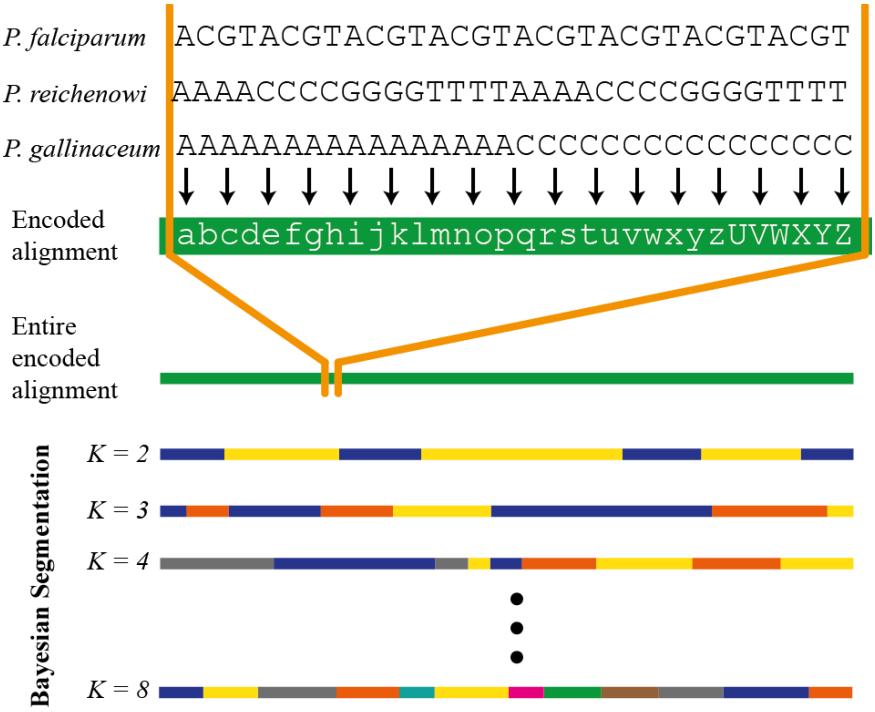


Figure 6.5: *Illustration depicting the process of Bayesian segmentation. It starts by obtaining a three-way alignment (top). The two-dimensional alignment is compressed into a one- dimensional string using a 32 character alphabet (demonstrated in* encoded alignment*). The alphabet is redundant, in that a position in the alignment with 'AAA' has the same code as a position with 'TTT'. It also uniquely encodes each type of mutation, as a position with 'AAT' gets a different code from a position that is 'ATA'. The one-dimensional string representing the whole-alignment is then partitioned into* K *separate segment classes (*Bayesian segmentation*) using a Bayesian hierarchical model, in which the number of classes,* K*, is a fixed hyperprior. The optimal number of classes* K *is found by selecting the* K *that minimizes the Deviance Information Criterion, has segments in all classes, and has stable segment classes.*

constraints and evolutionary rates. Similarly, we demonstrate that the identified segment classes can be ranked along a gradient of functional evolutionary divergence [45, 46] based on the probability of observing a mutation in the second codon position [47] (Table D.6, Figure 6.7). In particular, segments from classes $0, 1, 3, 5, 6, 8, 12, 13$ and 17 seem to show high levels of functional constraint. While the remaining segments seem particularly divergent. We chose this approach because we found evidence for mutation-saturation making it unlikely that commonly-used models for estimating natural selection would return robust results [48].

Mutation-saturation is detected when genetic divergence is less than expected given the amount of time since the most recent common ancestor between two lineages [49, 50].

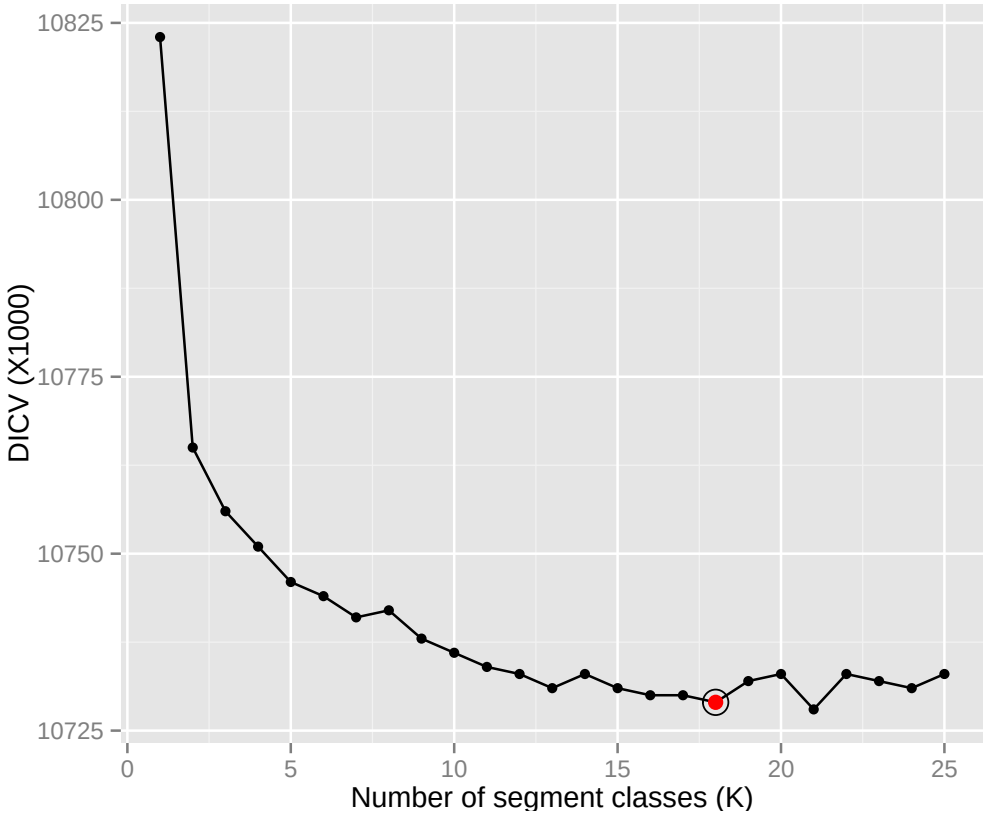


Figure 6.6: *DICV values across different values of* K. K= *18 is marked in red and circled.* K = *21 had smaller DICV, but did not meet the other criteria of segment stability and mixture proportions.*

To test for mutation-saturation we performed pairwise comparisons of the proportion of species-specific changes across segment classes. Across segment classes we found a high correlation between observed *P. falciparum* and *P. reichenowi*-specific changes (r = 0.99, R^2 *adjusted* = 0.99, p <0.05; Figure 6.8, bottom left). However, when comparing *P. gallinaceum*-specific changes to *P. falciparum*, top left- or *P. reichenowi*-specific changes, the correlation disappears (*Pg* to *Pf* comparison: r = -0.48, R^2 *adjusted* = 0.035, p >0.4; similar results were observed between *Pr* and *Pg*, Figure 6.8, top left and top right). Thus, while there is no evidence for mutation saturation between *P. falciparum* and *P. reichenowi*, the same cannot be said for *P. gallinaceum* and the other two species.

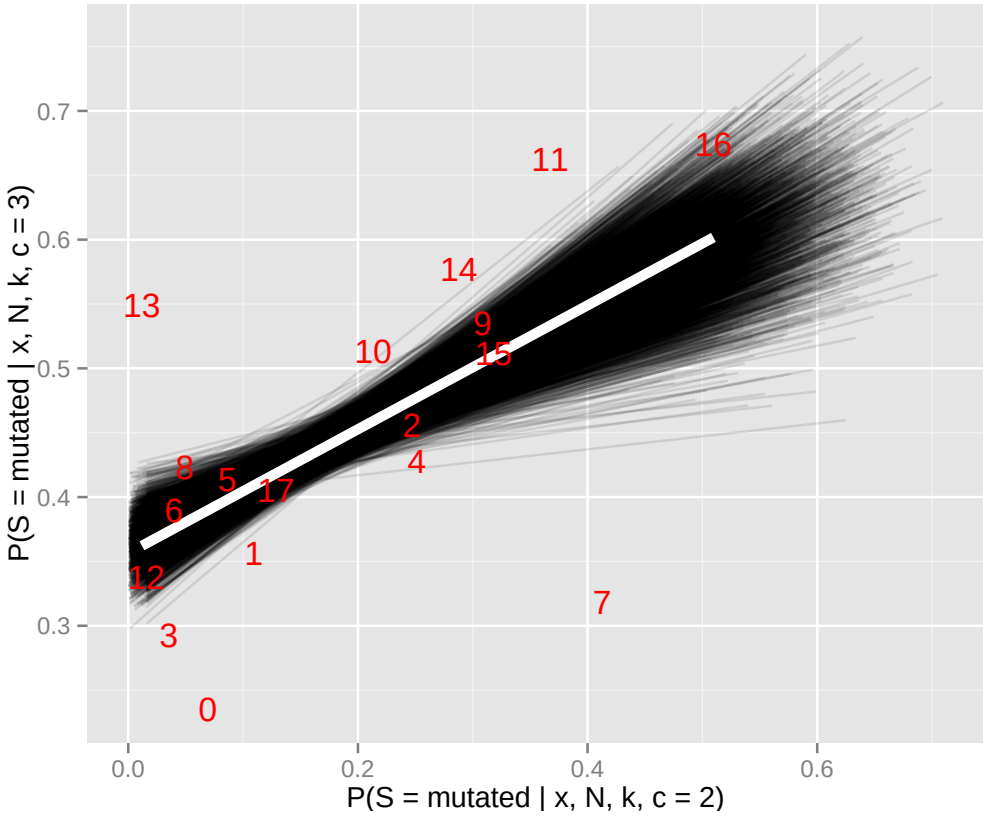


Figure 6.7: *Posterior P(S = mutated — x, N, k, c = 2) against the posterior P(S = mutated — x, N, k, c =3). Median values for each segment class are indicated by red numbers. Black lines are linear regression lines fitted to 5000 samples from the posterior. The bold white line is a linear regression line fitted to the median values.*

6.4.5 Validating assumption of functional constraint in Bayesian segmentation output

We validated our ranking of the segment classes by examining the distribution of segment classes across two genes: *actin1*, known to be highly evolutionarily constrained [51], and *msp1*, known to be highly divergent [52]. *actin1* forms an integral part of the parasite′s molecular motor [53]. The gene′s entire open reading frame (ORF) was classified as a single segment from a class with a low probability of observing mutations at the second codon position ($p = 0.016$, 95%HPD (highest probability density) 0.015-0.017; Figure 6.9).

msp1 plays a significant role in red-blood cell invasion [54, 55], and is known to be shared across the three species in this study [56]. In this case, we observed portions of

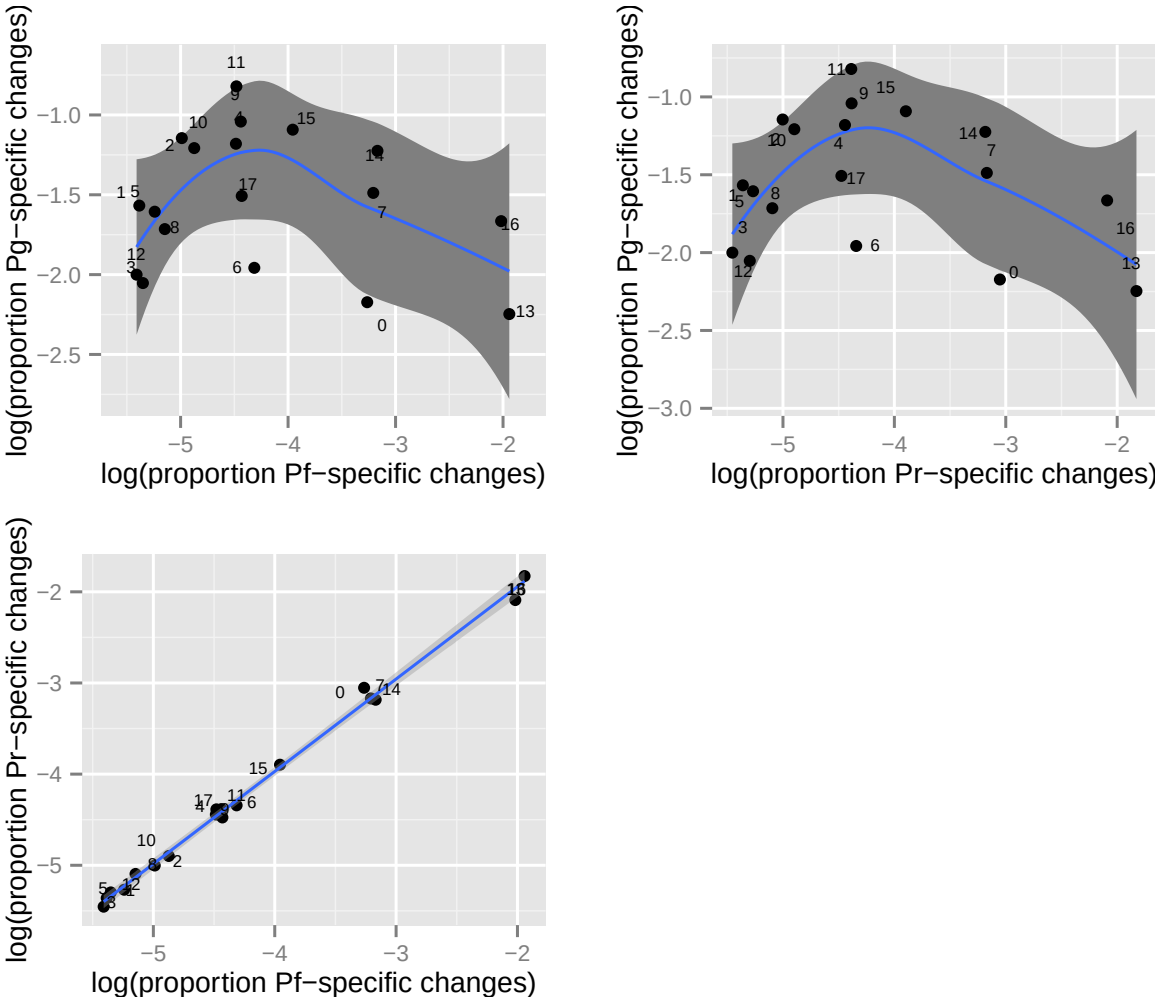


Figure 6.8: *Pairwise comparison between individual species log proportion specific changes. Top left: Pf vs Pg; Top right: Pr vs Pg; and Bottom left: Pf vs Pr. For comparisons including Pg, a loess model was fitted, using span of 0.99 and polynomial of degree 2. For the remaining comparison, a linear model was fitted. In all cases, 0.99 confidence interval envelopes are plotted. Points are plotted within jitter, while some jitter was applied to text labels to improve readability. Text labels refer to segment class ID, as noted in Table D.5.*

the ORF being classified into three different classes, all with relatively high posterior probabilities of observing a mutation at the second codon position (the segment class with lowest $p = 0.213$, 95%HPD 0.203-0.223; Figure 6.10). Together these results gave us confidence that our method was able to correctly classify and rank the recovered shared genomic segments along a gradient of functional evolutionary conservation. Studies in other organisms demonstrate further the strength of Bayesian segmentation in uncovering biologically relevant features in genomic data [57].

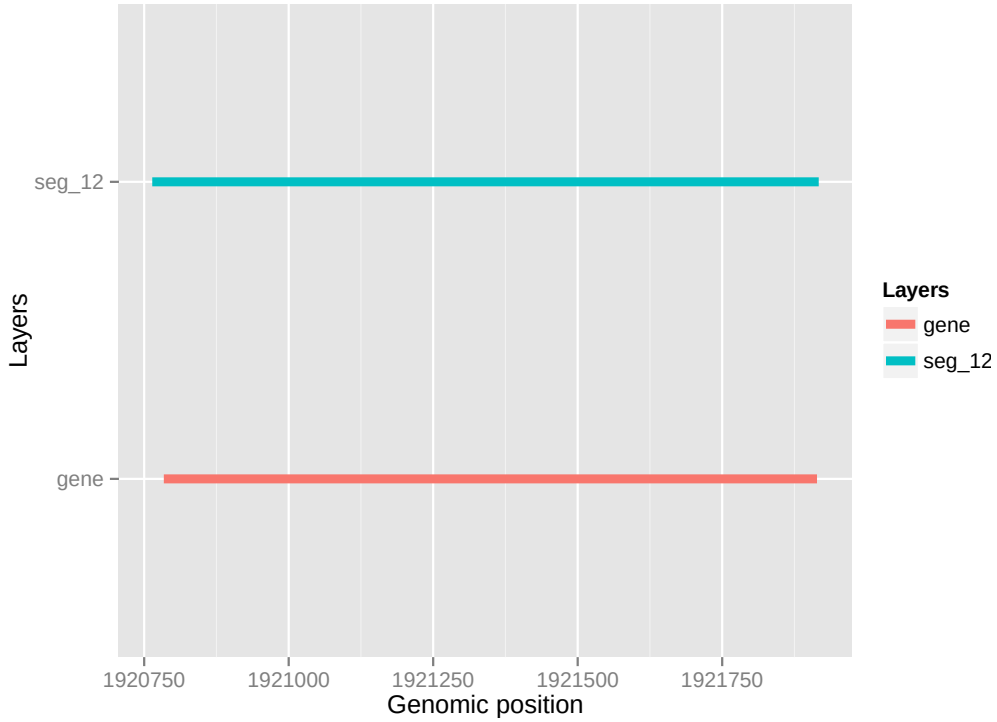


Figure 6.9: *Segment mapping to actin 1 gene. A single segment from class 12 mapped to the whole of the annotated coding region of actin 1. The segment includes a few bases both up and downstream from the annotation.*

6.4.6 Biological Processes Gene Ontology (GO) term analysis

The goal of this analysis was to examine the relationship between Biological Process GO terms and segment classes.

Cross-referencing GO terms to segments and segment classes

In order to cross-reference GO terms to segment classes, we used the `Bioconductor` package `org.Pf.plasmo.db` (version 3.1.2). The package provides access within `R` to the *Pf* annotation for version 3 of the *Pf* draft genome. Our first step was to determine how many genes were in the `org.Pf.plasmo.db`, ensure that a correct mapping of the genes could be made to our MySQL database, and then subset the nuclear genes. In total, we mapped **5507** nuclear genes between the `org.Pf.plasmo.db` database and our MySQL database that either had a peptide and/or mRNA annotation. Two nuclear genes present in the `org.Pf.plasmo.db` were not present in our database (PF3D7_1039300, PF3D7_1102000). It is unclear why these two genes are missing.

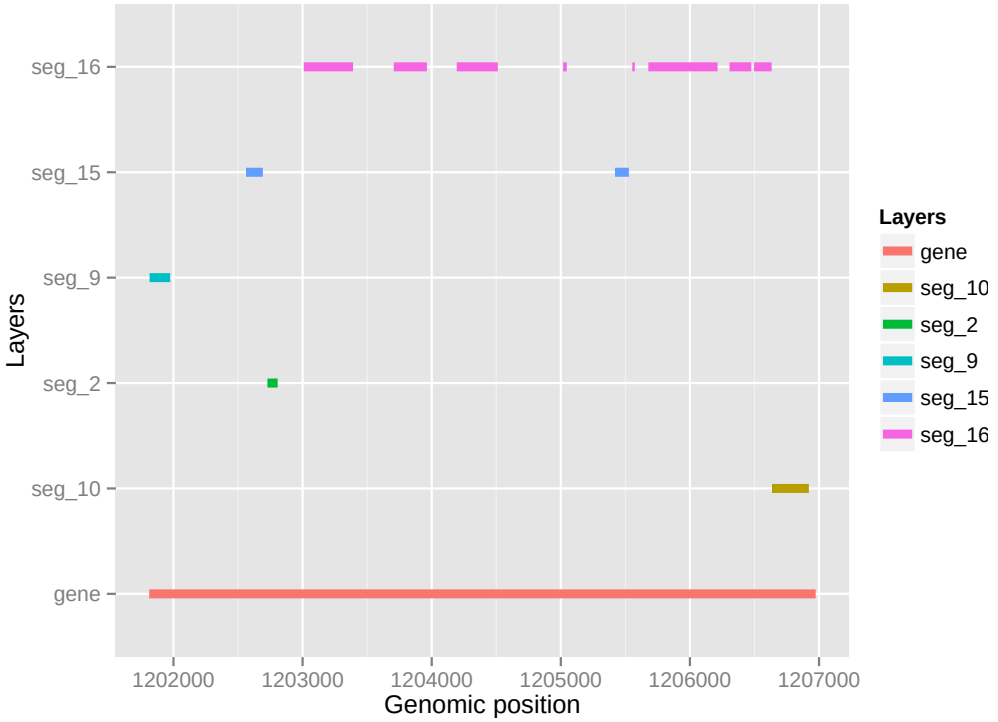


Figure 6.10: *Segments mapping to msp 1 gene. Segments from five classes mapped to the coding region of the msp 1 gene. Segment classes are plotted from top to bottom in decreasing probability of observing a mutation at the second codon position. The segment of class 9 is mapped to a portion annotated as a signal peptide. The segment of class 10 is mapped to the EGF domain 1. Many of the class 16 segments, and one of the class 15 segments mapped to the MSP1 C-terminus domain. The segment of class 2 did not map to any known protein domain, but maps closely to an N-glycosylation site.*

In `plamodb.org` these two genes are annotated as producing "unspecified products". Given the scope of our analysis, we did not pursue the matter further. An additional 97 genes, assigned to the apicoplast and mitochondrial genomes in the `org.Pf.plasmo.db` were ignored.

Overall, we found 13,295 segments mapped to the 2,688 genes of the 5,507 identified above. This leaves 2,819 genes without a mapping segment. To examine the possibility that the three-way alignment contributed to the missing genes, we attempted to map the Sanger reads available for *Pg* onto the *Pf* reference genome using `BWA` [58, 59], `Bowtie2` [60], and `LastZ` [61]. These failed to improve the number of genes. Thus, the missing genes are likely a combination of low coverage of the *Pg* genome, and true biological differences between the compared species.

A total of 1,672 GO terms were associated with 4,543 genes of the 5,507 nuclear genes of interest in our analysis (958 had no GO term annotation in the database). Of these, 1,411 are associated with 2,370 genes that have a mapped segment, and 867 were mapped to 2,173 genes without a mapping segment. Therefore, almost 85% of the GO terms associated with the *Pf* genome are represented in the 2,370 genes for which we observed at least one mapping segment. Count of GO terms by ontology are tallied in Table D.7.

Spearman rank correlation between segment class evolutionary constraint rank and segment class *z*-score for individual GO terms

In order to identify associations between segment class evolutionary constraint ranks and proportion of segments for each class overlapping genes of a single GO term, we performed Spearman rank correlations using *pvrank* function in R package: https://cran.r-project.org/web/packages/pvrank/pvrank.pdf. Spearman rank correlations test for monotonic changes in two ranks. A positive correlation coefficient (r) indicates a positive relationship between the two variables, while a negative correlation coefficient express a negative relationship. The closer r is to $+1$ or -1, the stronger the monotonic relationship. The closer r is to 0, the weaker the association between the ranks. To determine, based upon sample data, whether there is any or no evidence to suggest that a correlation is present in the population, one needs to perform hypothesis testing. If the information on the direction of the relationship is available (ie positive or negative correlation, eg: plot (ii) and (iii) of Figure 6.2 show a negative and a positive relationship between 2 ranks respectively), then a one-tailed test is appropriate. Otherwise a two-tailed test should be performed using the null hypothesis of 'there is no association between variables in the underlying population' against the alternative hypothesis, 'there is an association between variables in the underlying population'. A significant p - value (< 0.05) will favour the alternative hypothesis. The aim of this analysis was to identify GO terms with particularly high proportion of conserved segments relative to diverged segments, and vice-versa. To

accomplish this, we first standardised and normalized the count of segments of a particular class across GO terms. For the purposes of this analysis, we only examined Biological Processes GO terms with > 100 associated genes. Thus, for each GO term, we had a z-score associated with all segment classes, which measured how much more or less likely was a segment of that class mapping to genes associated with the GO term relative to all other GO terms. We also had the rank of segment classes based on our evolutionary constraint analysis. The results are presented in Table D.8.

The chromatin and gene transcription regulation machinery

The transcription machinery is essential for the *Pf*'s ability to evade the humans immune system. The machinery is responsible for the fine-tuned expression of the VAR genes, which ensure there is sufficient diversity across *Pf* clones that ensures at least some clones will go unnoticed by the human immune system. A total of 202 genes across four categories have been identified as important in regulating transcription in *Pf* [26], of which 128 had overlapping segments (Table D.9). Similarly to the GO term analysis, we also performed a Spearman rank correlation analysis. Here, z-scores were taken based on the mean count of segments per gene. Results are displayed in Table D.10.

Bibliography

[1] M C I Medeiros, G L Hamer, and R E Ricklefs. Host compatibility rather than vector-host-encounter rate determines the host range of avian Plasmodium parasites. *Proceedings of the Royal Society B-Biological Sciences*, 280:2012–2947, 2013.

[2] P C C Garnham. *Malaria Parasites and other Haemosporidia*. Blackwell Scientific, Oxford, 1966.

[3] M E Woolhouse, L H Taylor, and D T Haydon. Population biology of multihost pathogens. *Science*, 292:1109–1112, 2001.

[4] S Remold. Understanding specialism when the Jack of all trades can be the master of all. *Proc Biol Sci*, 279:4861–4869, 2012.

[5] C III van Riper, S G van Riper, M L Goff, and M Laird. The epizootiology and ecological significance of malaria in Hawaiian land birds. *Ecol Monogr*, 56: 327–344, 1986.

[6] R E Ricklefs, S M Fallon, and E Bermingham. Evolutionary relationships, cospeciation, and host switching in avian malaria parasites. *Syst Biol*, 53:111–119, 2004.

[7] J Mu, D A Joy, J Duan, Y Huang, J Carlton, J Walker, J Barnwell, P Beerli, M A Charleston, O G Pybus, and X Su. Host Switch Leads to Emergence of Plasmodium vivax Malaria in Humans. *Molecular Biology and Evolution*, 22: 1686–1693, 2005.

[8] C S Lim, L Tazi, and F J Ayala. Plasmodium vivax: Recent world expansion and genetic identity to Plasmodium simium, journal = P Natl Acad Sci USA. 102: 15523–15528, 2005.

[9] O Hellgren, J Perez-Tris, and S Bensch. A jack-of-all-trades and still a master of some: prevalence and host range in avian malaria and related blood parasites. *Ecology*, 90:2840–2849, 2009.

[10] J G Ewen, S Bensch, T M Blackburn, C Bonneaud, R Brown, P Cassey, R H Clarke, and J Perez-Tris. Establishment of exotic parasites: the origins and characteristics of an avian malaria community in an isolated island avifauna. *Ecol Lett*, 15:1112–1119, 2012.

[11] K S C Yotoko and C Elisei. Malaria parasites (Apicomplexa, Haematozoea) and their relationships with their hosts: is there an evolutionary cost for the

specialization? . *Journal of Zoological Systematics and Evolutionary Research*, 44:265–273, 2006.

[12] D C Outlaw and R E Ricklefs. Comparative gene evolution in haemosporidian (apicomplexa) parasites of birds and mammals. *Mol Biol Evol*, 27:537–542, 2010.

[13] M E Woolhouse and S Gowtage-Sequeria. Host range and emerging and reemerging pathogens. *Emerging Infectious Diseases*, 279:1842–1847, 2005.

[14] M T Duraisingh, T S Voss, A J Marty, M F Duffy, R T Good, J K Thompson, L H Freitas-Junior, A Scherf, B S Crabb, and A F Cowman. Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in Plasmodium falciparum. *Cell*, 121:13–24, 2005.

[15] C S Janssen, R S Phillips, C M R Turner, and M P Barrett. Plasmodium interspersed repeats: the major multigene superfamily of malaria parasites. *Nucleic Acids Res*, 32:5712–5720, 2004.

[16] M J Gardner, N Hall, E Fung, O White, M Berriman, R W Hyman, J M Carlton, A Pain, K E Nelson, S Bowman, and et al. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419:498–511, 2002.

[17] A Claessens, W L Hamilton, M Kekre, T D Otto, A Faizullabhoy, J C Rayner, and et al. Generation of antigenic diversity in Plasmodium falciparum by structured rearrangement of var genes during mitosis, journal = PLoS Genetics. 10:e1004812, 2014.

[18] A Scherf, R Hernandez-Rivas, P Buffet, E Bottius, C Benatar, B Pouvelle, J Gysin, and M Lanzer. Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in Plasmodium falciparum. *EMBO J*, 17:5418–5426, 1998.

[19] Q Chen, V Fernandez, A Sundstrom, M Schlichtherle, S Datta, P Hagblom, and M Wahlgren. Developmental selection of var gene expression in Plasmodium falciparum. *Nature*, 394:392–395, 1998.

[20] D J Roberts, A G Craig, A R Berendt, R Pinches, G Nash, K Marsh, and C I Newbold. Rapid switching to multiple antigenic and adhesive phenotypes in malaria. *Nature*, 357:689–692, 1992.

[21] P C Bull, A Pain, F M Ndungu, S M Kinyanjui, D J Roberts, C I Newbold, and K Marsh. Plasmodium falciparum antigenic variation: relationships between in vivo selection, acquired antibody response, and disease severity. *J. Infect. Dis.*, 192:1119–1126, 2005.

[22] G M Warimwe, T M Keane, G Fegan, J N Musyoki, C R Newton, A Pain, M Berriman, K Marsh, and P C Bull. Plasmodium falciparum var gene expression is modified by host immunity. *P Natl Acad Sci USA*, 106:21801–21806, 2005.

[23] N Rovira-Graells, A P Gupta, E Planet, V M Crowley, S Mok, L Ribas de Pouplana, P R Preiser, Z Bozdech, and A Cortes. Transcriptional variation in the malaria parasite Plasmodium falciparum. *Genome Res*, 22:925–938, 2012.

[24] C Frech and N Chen. Variant surface antigens of malaria parasites: functional and evolutionary insights from comparative gene family classification and analysis. *BMC Genomics*, 14:427, 2013.

[25] E J Lauron, K S Oakgrove1, L A Tell, K Biskar, S W Roy, and R N M Sehgal. Transcriptome sequencing and analysis of Plasmodium gallinaceum reveals polymorphisms and selection on the apical membrane antigen-1. *Malar J*, 13:382, 2014.

[26] E Bischoff and C Vaquero. In silico and biological survey of transcription-associated proteins implicated in the transcriptional machinery during the erythrocytic development of Plasmodium falciparum. *BMC Genomics*, 11:34, 2010.

[27] T D Otto, J C Rayner, U Bohme, A Pain, N Spottiswoode, M Sanders, M Quail, B Ollomo, F Renaud, A W Thomas, and et al. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nat Comms*, 5:4754, 2014.

[28] 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, and et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.*, 25:25–29, 2000.

[29] J M Keith, P Adams, S Stephen, and J S Mattick. Delineating slowly and rapidly evolving fractions of the Drosophila genome. *Journal of Computational Biology*, 15:407–430, 2008.

[30] J D Smith, C E Chitnis, A G Craig, D J Roberts, D E Hudson-Taylor, D S Peterson, R Pinches, C I Newbold, and L H Miller. Switches in expression of Plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell*, 82:101–110, 1995.

[31] X Z Su, V M Heatwole, S P Wertheimer, F Guinet, J A Herrfeldt, D S Peterson, J A Ravetch, and T E Wellems. The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum -infected erythrocytes. *Cell*, 82:89–100, 1995.

[32] Q Cheng, N Cloonan, K Fischer, J Thompson, G Waine, M Lanzer, and A Saul. stevor and rif are Plasmodium falciparum multicopy gene families which potentially encode variant antigens. Mol. Biochem. Parasitol. *Mol. Biochem. Parasitol.*, 97: 161–176, 1998.

[33] T Y Sam-Yellowe, L Florens, J R Johnson, T Wang, J A Drazba, K G Le Roch, and et al. A Plasmodium gene family encoding Maurer's cleft membrane proteins: structural properties and expression profiling. *Genome Res*, 14:1052–1059, 2004.

[34] L Cui and J Miao. Chromatin-mediated epigenetic regulation in the malaria parasite Plasmodium falciparum. *Eukaryotic Cell*, 9:1138–1149, 2010.

[35] M F Duffy, S A Selvarajah, G A Josling, and M Petter. The role of chromatin in Plasmodium gene expression. *Cellular Microbiology*, 14:819–828, 2012.

[36] S C Nardelli, F-Y Che, N C Silmon de Monerri, H Xiao, E Nieves, C Madrid-Aliste, S O Angel, Jr Sullivan, J William, R H Angeletti, K Kim, and L M Weiss. The Histone Code of Toxoplasma gondii Comprises Conserved and Unique Posttranslational Modifications. *MBio*, 4:e0092213, 2013.

[37] S Balaji, M M Babu, L M Iyer, and L Aravind. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Res*, 33:3994–4006, 2005.

[38] Z Bozdech, S Mok, G Hu, M Imwong, A Jaidee, B Russell, H Ginsburg, F Nosten, N P J Day, N J White, and et al. The transcriptome of Plasmodium vivax reveals divergence and diversity of transcriptional regulation in malaria parasites. *P Natl Acad Sci USA*, 105:16290–16295, 2008.

[39] C J Merrick, C Huttenhower, C Buckee, A Amambua-Ngwa, N Gomez-Escobar, M Walther, D J Conway, and M T Duraisingh. Epigenetic dysregulation of virulence gene expression in severe Plasmodium falciparum malaria. *J. Infect. Dis.*, 205:1593–1600, 2012.

[40] J C Silva, A Egan, C Arze, J L Spouge, and D G Harris. A new method for estimating species age supports the coexistence of malaria parasites and their Mammalian hosts. *Mol Biol Evol*, 32:1354–1364, 2015.

[41] J M Keith. Segmenting eukaryotic genomes with the Generalized Gibbs Sampler. *Journal of Computational Biology*, 13:1369–1383, 2006.

[42] C Oldmeadow and J M Keith. Model Selection in Bayesian Segmentation of multiple DNA alignments. *Bioinformatics*, 27:604–610, 2011.

[43] M Algama, C Oldmeadow, E Tasker, K Mengersen, and J M Keith. Drosophila 3' UTRs Are More Complex than Protein-Coding Sequences. *PLoS ONE*, 9:e97336, 2014.

[44] C Oldmeadow, K Mengersen, J S Mattick, and J M Keith. Multiple evolutionary rate classes in animal genome evolution. *Molecular Biology and Evolution*, 27: 942–953, 2010.

[45] B S Gaut and J F Doebley. DNA sequence evidence for the segmental allotetraploid origin of maize. *P Natl Acad Sci USA*, 94:6809–6814, 1997.

[46] M K Hughes and A L Hughes. Evolution of duplicate genes in a tetraploid animal, Xenopus laevis. *Mol Biol Evol*, 10:1360–1369, 1993.

[47] L Bofkin and N Goldman. Variation in evolutionary processes at different codon positions. *Mol Biol Evol*, 24:513–521, 2007.

[48] M Nei and T Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, 3:418–426, 1986.

[49] W M Brown, E M Prager, A Wang, and A C Wilson. Mitochondrial DNA sequences of primates: Tempo and mode of evolution. *J Mol Evol*, 18:225–239, 1982.

[50] W M Brown, M George, and A C Wilson. Rapid evolution of animal mitochondrial DNA. *P Natl Acad Sci USA*, 76:1967–1971, 1979.

[51] H V Goodson and W F Hawse. Molecular evolution of the actin family. *J Cell Sci*, 115:2619–2622, 2002.

[52] L H Miller, T Roberts, M Shahabuddin, and T F McCutchan. Analysis of sequence diversity in the Plasmodium falciparum merozoite surface protein-1 (MSP-1). *Mol. Biochem. Parasitol*, 59:1–14, 1993.

[53] J Baum, D Richard, J Healer, M Rug, Z Krnajski, T W Gilberger, J L Green, A A Holder, and A F Cowman. A conserved molecular motor drives cell invasion and gliding motility across malaria life cycle stages and other Apicomplexan parasites. *J. Biol. Chem.*, 281:5197–5208, 2006.

[54] M J Blackman, H G Heidrich, S Donachie, J S McBride, and A A Holder. A single fragment of a malaria merozoite surface protein remains on the parasite during red cell invasion and is the target of invasion-inhibiting antibodies. *J. Exp. Med.*, 172:379–382, 1990.

[55] M J Blackman, T J Scott-Finnigan, S Shai, and A A Holder. Antibodies inhibit the protease-mediated processing of a malaria merozoite surface protein. *The Journal of Experimental Medicine*, 180:389–393, 1994.

[56] S D Polley, G D Weedall, A W Thomas, L M Golightly, and D J Conway. Orthologous gene sequences of merozoite surface protein 1 (MSP1) from Plasmodium reichenowi and P. gallinaceum confirm an ancient divergence of P. falciparum alleles. *Mol. Biochem. Parasitol.*, 142:25–31, 2005.

[57] S E Boyd, B Nair, S W Ng, J M Keith, and J M Orian. Computational characterization of 3' splice variants in the GFAP isoform family. *PLoS ONE*, 7:e33565, 2012.

[58] H Li and R Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26:589–595, 2010.

[59] H Li. Towards better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30:btu356–2851, 2014.

[60] B Langmead and S Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9:357–359, 2012.

[61] R S Harris. Improved pairwise alignment of genomic DNA. *Pennsylvania State University*, 2007.

Chapter 7

Summary, Conclusions and Future Work

The main objective of this thesis is to apply a Bayesian approach to identify genome-wide putative ncRNAs and other regulatory sequences contributing to diseases. These elements were identified by applying the statistical technique of *sequence segmentation* to partition an alignment of multiple species. In the past, other authors have attempted to divide a sequence alignment into conserved and divergent segments [1–5]. This dichotomous approach fails to capture the complexity of genome-wide conservation landscapes, which have evolved under a diverse range of structural and functional constraints. To overcome this limitation, I used a Bayesian segmentation model (*changept* model) that classifies genomic segments into more than two segment classes based on degree of conservation and other sequence characteristics (eg: GC level, transition/transversion ratio) [6, 7].

In the first part of the thesis, I introduced the concept of integrating multiple data types into segmentation. To encode multiple information in an alignment (conservation, GC content and transition/transversion ratio), I used a D-character representation (D is a positive integer). A 3-way alignment was encoded using a 32-character representation,

where in each alignment coloumn, the complementary bases were also encoded using the same letters. This helped to reduce the number of parameters to be estimated, and hence the computational cost opposed to using a 64-character representation which also captures the strand-specific information. *Changept* was then applied to segment this sequence, and segments were classified according to character frequencies. Using this encoding greatly increased the number of segment classes identified (see Chapter 3). The method is especially beneficial in analysing segmentation patterns in closely related species, as conservation alone cannot identify fine differences between these species. The main drawback of D-character encoding is that it can be computationally expensive when the number of species increases (ie >3). In this case, two other methods can be used to transform an alignment: (1) Maximum frequency transformation; (2) Parsimony scores (see Literature Review Part 2).

Investigating stability of segment classes to facilitate selection of the number of segment classes was the second main methodological development I introduced in this thesis. In general, this method identifies a larger number of segment classes than an earlier method based on investigating DICV values. Typically, the additional classes were resulted from splitting some of the classes identified by the DICV method. This new method led to the discovery of additional features (eg: new motifs, see Chapter 3) not identified using the models selected by the DICV method.

These methods were first applied to segment $3'$ UTR regions of three closely related *Drosophila* species: *D. melanogaster*, *D. simulans*, and *D. yakuba* (Chapter 3). I demonstrated that $3'$ UTRs have more complex sequence structure than coding sequence, and argue that this is indicative of greater functional complexity. Several segment classes were highly enriched in low information content sequences and we propose that certain low information content regions are functional only in the sense that they frequently gain or lose 'decoy' TF binding sites, thus facilitating rapid, coordinated, adaptive responses of expression levels over many genes. We refer to this idea as the 'poised genome', and in future work we are interested in pursuing this idea.

In chapters 4 to 6, I used the new methods to identify putative functional non-coding elements (PFEs) that may be associated with or relevant to several diseases: (1) muscle disease; (2) dengue; and (3) malaria. In Chapter 4, I also built a systematic process to identify genome-wide PFEs based on identifying deeply conserved regions between human and zebrafish using *changept* classifications. In addition to genome-wide study, a pathway-focussed analysis was carried out using 24 genes involved in myogenesis. Results revealed the advantages of applying *changept* to a high quality alignment. A larger number of PFEs was identified in the pathway-focussed analysis using the alignment software LAGAN (and manual interventions) than in the genome-wide analysis restricted to the same human genes using the readily available multiz-8way alignment. All PFEs tested (26 PFEs) in this pathway-focussed analysis using RT-PCR were found to be expressed. Furthermore, PFEs identified in this pathway-focussed analysis were substantially longer than EvoFold predicted regions, and this suggests that the method I applied has identified extended functional regions surrounding EvoFold predictions. Both genome-wide and pathway-focussed analyses provided further evidence that ncRNAs are enriched in transcription factors [8]. Given that the quality of the alignment increases the number of PFEs identified, future work should focus on repeating the genome-wide analysis using an improved alignment. In filtering PFEs, a set of thresholds were used to ensure that laboratory validation of the approach would be relatively easy to achieve without the hindrance of false positives (eg: length of the PFE $\geq 100nt$, number of gaps in the PFE segment < 20 alignment columns or if the total length of gaps within the segment was $< 10\%$ the length of the segment, profile value > 0.9, etc.). However, ideally the thresholds should have chosen to strike an optimal balance between sensitivity and specificity, and this is an issue for future research. Identifying the specific functions of PFEs detected in the pathway-focussed analysis will help to determine if these regions play a role in muscle diseases.

Certain strains of *Wolbachia* inhibit replication of mosquito-borne pathogens, such as dengue viruses, the malaria parasite, and filarial nematodes [9]. In Chapter 5, two experimentally validated small ncRNAs in two *Wolbachia* strains: *w*Mel and *w*Pip were identified. In addition to these, I identified 18 highly conserved intergenic regions using *changept*. The main limitation of the *changept* method is that it would not identify candidates that are not present in both strains, have originated since the divergence of the supergroups, or are conserved at the level of secondary structure rather than nucleotide sequence. To overcome these limitations, we used a second approach based on mapping RNAseq data. However, this method would not be able to identify putative sRNAs if they were not present in the RNAseq library. Further research is required to identify the specific functions of these candidate sRNAs.

A simple, two-category, classification of genome segments (conserved, divergent) would miss nuances of the data, such as GC-content, transition-transversion ratios, dN/dS ratios, and the fact that conserved and divergent are opposite ends of a continuous scale. This limitation was overcome by the new *changept* method - integrating multiple data types into segmentation. Using this method, in Chapter 6, I carried out a comparative genomic analysis on three malaria species - *P. falciparum*, *P. reichenowi* and *P. gallinaceaum* - to identify genetic mechanisms that facilitate host jumping. Analysis suggests that the mechanism controlling the expression of variable surface antigen is ancestral, and occurs in at least one *Plasmodium* species that does not infect mammals. The variable surface antigen (VSA) gene families of mammalian *Plasmodium* are likely candidate genes contributing to *host-generalism* [10]. In turn, these results will help to warn us about the next possible host of these malaria species, hence will contribute to build early -warning systems for disease emergence. Findings will also help to better understand the deadly malaria disease and might suggest novel regions for drug-discovery.

The next step of this work would be to sequence a *Plasmodium* that is truly a host-generalist, like *Plasmodium relictum*, that has been found in over 100 bird species.

That would allow us to examine their gene transcription apparatus, and see how it differs from *P. falciparum*. It would also allow us to search for VSA gene families, and determine how similar they are to those in *P. falciparum*, and other mammalian *Plasmodium*.

A wealth of genomic data is presented as profiles of continuous measurements across a genome (eg: RNA expression levels, copy number variation) and the *changept* model is not applicable to this type of data. Developing a model to overcome this limitation and integrating it with *changept* will enable simultaneous segmentation of both types of data, discrete and continuous.

One advantage of the Bayesian frame work is the use of priors, when prior knowledge of parameters is available. However, *changept* model has been implemented using uninformative priors. Using the boundaries of known functional elements, and forcing them to be in the same segment class are some ways to incorporate prior information to the model.

Bibliography

[1] R H Waterston, K Lindblad-Toh, E Birney, J Rogers, JF Abril, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.

[2] E Birney, J A Stamatoyannopoulos, A Dutta, R Guigo, TR Gingeras, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816, 2007.

[3] K Lindblad-Toh, C M Wade, T S Mikkelsen, E K Karlsson, D B Jaffe, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438:803–819, 2005.

[4] A Siepel, G Bejerano, J S Pedersen, A S Hinrichs, M Hou, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15:1034–1050, 2005.

[5] K S Pollard, M J Hubisz, K R Rosenbloom, and A Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20:110–121, 2010.

[6] J M Keith. Segmenting eukaryotic genomes with the Generalized Gibbs Sampler. *Journal of Computational Biology*, 13:1369–1383, 2006.

[7] J M Keith, P Adams, S Stephen, and J S Mattick. Delineating slowly and rapidly evolving fractions of the Drosophila genome. *Journal of Computational Biology*, 15:407–430, 2008.

[8] H I Nakaya, P P Amaral, R Louro, A Lopes, A A Fachel, Y B Moreira, T A El-Jundi, A M da Silva, E M Reis, and S Verjovski-Almeida. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome Biology*, 8:R43, 2007.

[9] J G Mayoral, M Hussain, D A Joubert, I Iturbe-Ormaetxe, S L ONeill, and S Asgar. Wolbachia small noncoding RNAs and their role in cross-kingdom communications. *PNAS*, 111:52:18721–18726, 2014.

[10] M T Duraisingh, T S Voss, A J Marty, M F Duffy, R T Good, J K Thompson, L H Freitas-Junior, A Scherf, B S Crabb, and A F Cowman. Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in Plasmodium falciparum. *Cell*, 121:13–24, 2005.

Appendix A

Appendix Chapter 3

File S1: Positions of segments for the 15-class model of *D. melanogaster* versus *D. simulans* alignment (BED file).

http://dx.doi.org/10.6084/m9.figshare.1517635

File S2: Positions of segments for the 16-class model of *D. melanogaster* versus *D. yakuba* alignment (BED file).

http://dx.doi.org/10.6084/m9.figshare.1517636

File S3: Positions of segments for the 15-class model of 3-way *D. melanogaster*, *D. simulans*, *D. yakuba* alignment (BED file).

http://dx.doi.org/10.6084/m9.figshare.1517637

Supplementary Table 1 - Characteristics of 12-class and 15-class models for *D. mel* versus *D. sim* 3'UTR alignment

The classes of the 12-class model are labelled 12 -0 to 12 -11, and the classes of the 15-class model are labelled 15 -0 to 15 -14.

The averages 15 - 3&4, 15 - 12&13 and 15 - 7&14 are weighted according to the mixture proportions of the classes averaged.

T/T: Transition/Transversion ratio.

Model-Class	Mixture proportion	Conservation	GC content	T/T	Type of motif
12 - 0	9.10%	100%	29%	0.46	poly A
15 - 1	14.30%	99%	28%	0.80	AT; poly A
12 - 1	7.20%	98%	50%	1.49	CAG; CA
15 - 10	8.20%	98%	51%	1.45	CAG
12 - 2	20.30%	98%	38%	1.13	A[CA]
15 - 0	15.90%	99%	38%	1.18	CAA
12 - 3	8.80%	92%	24%	0.70	
15 - 6	7.70%	92%	24%	0.67	
12 - 4	4.60%	97%	59%	1.50	CAG; C[AT]C; CA[CG]
15 - 9	3.00%	97%	60%	1.48	CAG; CA[AG]; TCC
12 - 5	4.40%	84%	26%	0.75	
15 - 5	2.90%	83%	25%	0.73	
12 - 6	19.90%	98%	28%	0.84	
15 - 3&4	19.40%	97%	29%	0.89	AT (15:3)
12 - 7	0.90%	54%	20%	0.53	
15 - 1	0.70%	56%	17%	0.50	
12 - 8	9.80%	95%	49%	1.18	CA; CAG
15 - 12&13	16.90%	95%	46%	1.16	CA (15:12) / CAG (15:13)
12 - 9	1.80%	85%	49%	0.96	poly G
15 - 2	2.00%	86%	47%	0.94	
12 - 10	12.50%	92%	34%	1.00	
15 - 8	8.00%	90%	33%	0.98	
12 - 11	0.60%	46%	42%	0.78	
15 - 7&14	0.90%	48%	42%	0.75	poly G (15:7)

Supplementary Table 2 - Types of motif identified in D. melanogaster-D. simulans 12-class model

Class	Type of motif	E-value	Motif width	# sites
0	poly A	1.10E-11	20	82
1	CAG repeat	5.10E-13	19	62
	CA repeat	1.60E-07	20	32
2	A[CA] repeat	4.50E-46	20	223
4	CAG repeat	8.60E-88	14	120
	C[AT]C repeat	1.00E-48	20	130
	CA[CG] repeat	0.049	20	38
8	CAG repeat	6.40E-04	15	63
	CAG repeat	1.00E-19	20	47
9	poly G	0.0028	15	36

Supplementary Table 3 - Types of motif identified in D. melanogaster-D. simulans 15-class model

Class	Type of motif	E-value	Motif width	# sites	TOMTOM matches (Number - Database)
0	CAA repeat	3.00E-34	20	240	8 - All Drodophila
1	AT repeat	4.00E-36	19	171	1 - All Drosophila
	ployA	3.70E-43	20	136	4 - All Drosophila
3	AT repeat	1.30E-13	20	25	no matches
7	polyG	4.60E-02	15	16	8 - All Drodophila; 11 - All Vertebrates; 2 - JASPAR-insects
9	CAG repeat	5.30E-138	20	124	1 - All Vertebrates; 1 - JASPAR-insects
	CA[AG] repeat	5.00E-08	20	32	no matches
	TCC repeat	3.70E-05	15	96	1 - All Vertebrates; 1 - RNA-binding motifs
10	CAG repeat	1.60E-21	20	114	1 - All Vertebrates; 1 - JASPAR-insects
12	CA repeat	3.80E-12	20	35	3 - All Drosophila; 2 - All Vertebrates; 11- RNA-binding motifs
13	CAG repeat	3.20E-33	19	64	3 - All Vertebrates

Supplementary Table 4 - Types of motif identified in D. melanogaster-D. yakuba 16-class model

Classs	Type of motif	E-value	Motif width	# sites	TOMTOM matches (Number - Database)
0	CAG repeat	3.60E-23	20	64	2 - All Vertebrates
1	CAA repeat	3.90E-11	20	151	6 - All Drosophila
2	CAG repeat	1.70E-15	19	107	2 - All Vertebrates
3	polyG	6.20E-03	15	28	7 - All Drosophila; 10 - All Vertebrates; 1- JASPAR-insects; 1 - RNA-binding motifs
6	CAG repeat	4.10E-09	20	34	1 - All Vertebrates; 1 - JASPAR-insects
8	polyA	2.10E-21	20	65	4 - All Drosophila; 6 - RNA-binding motifs
	AT repeat	2.30E-21	18	108	1 - All Drosophila
11	CA repeat	3.30E-16	20	46	1 - All Drosophila; 11 - RNA-binding motifs
12	CAG repeat	9.60E-78	20	78	no matches
	CCN repeat	2.70E-18	20	79	1 - All Vertebrates; 1 - JASPAR-insects
15	polyA	5.40E-06	20	95	6 - All Drosophila; 7 - RNA-binding motifs
	AT repeat	7.90E-30	20	80	no matches

Supplementary Table 5 - Class comparisons

Comparison of change-point character frequencies in each of the classes indentified by Procedure 2 for each paiwise alignment of D. melanogaster (D. mel), D. simulans (D. sim), and D. yakuba (D. yak) 3' UTRs

Classes from different models with similar character frequencies are grouped together.

Alignment	Class	a	b	c	d	e	f	g	h	Type of motif
D. mel - D. sim	**0**	61.6%	0.2%	0.4%	0.2%	0.1%	37.2%	0.1%	0.2%	CAA
D. mel - D. yak	**1**	62.9%	0.2%	0.4%	0.4%	0.2%	35.4%	0.1%	0.4%	CAA
D. sim - D. yak	**1**	62.4%	0.2%	0.4%	0.4%	0.2%	35.9%	0.1%	0.5%	
D. mel - D. sim	**1**	72.0%	0.1%	0.2%	0.2%	0.1%	27.3%	0.0%	0.1%	AT; poly A
D. mel - D. yak	**15**	69.3%	0.6%	0.9%	0.9%	0.5%	26.7%	0.2%	0.9%	AT; poly A
D. sim - D. yak	**7**	68.0%	0.5%	1.0%	1.1%	0.6%	27.3%	0.2%	1.2%	
D. mel - D. sim	**2**	45.2%	2.2%	3.9%	2.0%	1.9%	40.6%	1.3%	3.0%	
D. mel - D. yak	**7**	44.3%	3.5%	7.3%	4.9%	3.2%	27.8%	2.1%	6.9%	
D. sim - D. yak	**2**	46.3%	3.4%	6.7%	4.8%	3.3%	26.2%	2.2%	7.0%	
D. mel - D. sim	**3**	80.8%	0.2%	0.2%	0.5%	0.2%	17.8%	0.1%	0.2%	AT
D. mel - D. yak	**8**	75.0%	0.2%	0.3%	0.3%	0.1%	23.9%	0.0%	0.3%	AT; ploy A
D. sim - D. yak	**0**	74.5%	0.1%	0.2%	0.3%	0.1%	24.5%	0.1%	0.3%	
D. mel - D. sim	**4**	67.9%	0.5%	1.0%	0.7%	0.4%	28.6%	0.2%	0.7%	
D. mel - D. yak	**4**	58.5%	2.4%	4.5%	4.2%	2.2%	22.7%	1.2%	4.3%	
D. mel - D. sim	**5**	64.3%	2.3%	3.6%	4.5%	2.2%	18.6%	0.9%	3.6%	
D. mel - D. yak	**13**	47.9%	5.7%	9.3%	11.5%	5.4%	9.8%	2.2%	8.3%	
D. sim - D. yak	**11**	54.5%	4.6%	7.4%	10.0%	4.6%	10.2%	1.5%	7.2%	
D. mel - D. sim	**6**	71.0%	1.2%	1.7%	2.3%	1.2%	20.5%	0.4%	1.7%	
D. mel - D. yak	**14**	71.0%	1.5%	2.1%	3.4%	1.5%	17.7%	0.5%	2.4%	
D. sim - D. yak	**8**	68.1%	1.4%	2.1%	2.9%	1.3%	21.1%	0.6%	2.4%	
D. mel - D. sim	**7**	19.1%	6.3%	10.0%	4.1%	6.5%	38.9%	5.0%	9.9%	poly G
D. mel - D. yak	**3**	22.6%	6.5%	8.8%	4.7%	6.9%	37.0%	4.7%	8.9%	poly G
D. sim - D. yak	**3**	21.6%	6.2%	8.6%	4.2%	6.4%	38.5%	4.5%	10.0%	
D. mel - D. sim	**8**	61.7%	1.2%	2.6%	1.9%	1.1%	28.5%	0.7%	2.2%	
D. mel - D. yak	**10**	62.0%	1.3%	2.3%	2.3%	1.1%	27.8%	0.6%	2.6%	
D. sim - D. yak	**12**	56.5%	1.7%	3.3%	2.7%	1.6%	29.3%	1.0%	3.9%	

Alignment	Class	a	b	c	d	e	f	g	h	Type of motif
D. mel - D. sim	**9**	38.7%	0.3%	1.0%	0.3%	0.2%	58.8%	0.3%	0.5%	CAG; CA[AG]; TCC
D. mel - D. yak	**12**	37.0%	0.7%	1.6%	0.6%	0.4%	57.7%	0.6%	1.4%	CAG; CCN
D. sim - D. yak	**4**	36.8%	0.5%	1.2%	0.6%	0.5%	58.4%	0.6%	1.4%	
D. mel - D. sim	**10**	47.9%	0.2%	0.7%	0.2%	0.2%	50.3%	0.1%	0.3%	CAG
D. mel - D. yak	**0**	47.1%	0.2%	0.8%	0.3%	0.2%	50.6%	0.2%	0.5%	CAG
D. sim - D. yak	**10**	47.2%	0.2%	0.6%	0.3%	0.2%	50.7%	0.3%	0.6%	
D. mel - D. sim	**12**	55.0%	0.7%	1.5%	0.8%	0.5%	40.3%	0.4%	0.9%	CA
D. mel - D. yak	**11**	56.3%	0.6%	1.5%	1.0%	0.6%	37.9%	0.5%	1.5%	CA
D. sim - D. yak	**5**	55.9%	0.7%	1.5%	1.3%	0.7%	37.3%	0.5%	2.0%	
D. mel - D. sim	**13**	43.2%	0.8%	1.8%	0.6%	0.5%	51.4%	0.5%	1.3%	CAG
D. mel - D. yak	**2**	43.4%	0.9%	2.1%	0.8%	0.7%	49.3%	0.7%	2.1%	CAG
D. sim - D. yak	**6**	43.3%	0.7%	1.8%	0.7%	0.7%	49.9%	0.7%	2.1%	
D. mel - D. sim	**14**	33.3%	8.0%	12.1%	12.3%	8.3%	11.1%	4.0%	10.9%	
D. sim - D. yak	**14**	38.7%	6.7%	11.0%	9.8%	6.5%	13.4%	3.3%	10.7%	
D. mel - D. yak	**5**	58.7%	3.6%	5.7%	6.9%	3.2%	14.9%	1.4%	5.6%	
D. sim - D. yak	**9**	59.8%	2.8%	4.8%	5.3%	2.6%	18.7%	1.2%	4.8%	
D. mel - D. yak	**6**	35.9%	2.2%	3.9%	2.1%	2.1%	48.4%	1.7%	3.8%	CAG
D. sim - D. yak	**13**	40.3%	1.7%	3.5%	2.0%	2.1%	44.3%	1.6%	4.4%	

Classes with no comparable class in other models

	Class	a	b	c	d	e	f	g	h	Type of motif
D. mel - D. sim	**11**	52.1%	5.6%	7.4%	17.5%	5.2%	3.9%	1.1%	7.2%	
D. mel - D. yak	**9**	47.1%	1.7%	4.0%	2.2%	1.7%	37.7%	1.4%	4.2%	

Supplementary Table 6 - Class comparisons pairwise and 3-way alignments

Comparison of change-point models indetified by Procedure 2 for each paiwise alignments and the 3-way alignment of 3' UTRs
Classes from different models with similar character frequencies are grouped together.

Alignment	Class	mixture proportion	conservation	GC	a	f	Type of motif
D. mel - D. sim	**0**	15.9%	99.0%	38.0%	61.6%	37.2%	CAA
D. mel - D. yak	**1**	11.8%	98.0%	36.0%	62.9%	35.4%	CAA
D. sim - D. yak	**1**	13.5%	98.0%	37.0%	62.4%	35.9%	
3-way	**3**	14.4%	97.4%	40.5%	57.9%	39.4%	CAA; CA
D. mel - D. sim	**1**	14.3%	99.0%	28.0%	72.0%	27.3%	AT; poly A
D. mel - D. yak	**15**	13.8%	96.0%	28.0%	69.3%	26.7%	AT; poly A
D. sim - D. yak	**7**	13.6%	95.0%	29.0%	68.0%	27.3%	
D. mel - D. sim	**2**	2.0%	86.0%	47.0%	45.2%	40.6%	
D. mel - D. yak	**7**	2.0%	72.0%	40.0%	44.3%	27.8%	
D. sim - D. yak	**2**	2.3%	72.0%	39.0%	46.3%	26.2%	
D. mel - D. sim	**3**	2.3%	99.0%	18.0%	80.8%	17.8%	AT
D. mel - D. yak	**8**	8.5%	99.0%	24.0%	75.0%	23.9%	AT; ploy A
D. sim - D. yak	**0**	11.0%	99.0%	25.0%	74.5%	24.5%	
3-way	**0**	16.8%	98.1%	26.1%	72.7%	25.4%	poly A
D. mel - D. sim	**4**	17.1%	96.0%	30.0%	67.9%	28.6%	
D. mel - D. yak	**4**	7.5%	81.0%	30.0%	58.5%	22.7%	
3-way	**14**	7.2%	80.1%	32.7%	55.9%	24.2%	
D. mel - D. sim	**5**	2.9%	83.0%	25.0%	64.3%	18.6%	
D. mel - D. yak	**13**	1.6%	58.0%	26.0%	47.9%	9.8%	
D. sim - D. yak	**11**	1.6%	65.0%	24.0%	54.5%	10.2%	
3-way	**7**	1.5%	57.2%	26.1%	47.4%	9.8%	
D. mel - D. sim	**6**	7.7%	92.0%	24.0%	71.0%	20.5%	
D. mel - D. yak	**14**	3.9%	89.0%	22.0%	71.0%	17.7%	
D. sim - D. yak	**8**	6.9%	89.0%	25.0%	68.1%	21.1%	
3-way	**6**	6.5%	85.9%	23.9%	67.3%	18.6%	
D. mel - D. sim	**7**	0.3%	58.0%	60.0%	19.1%	38.9%	poly G
D. mel - D. yak	**3**	0.8%	60.0%	57.0%	22.6%	37.0%	poly G
D. sim - D. yak	**3**	0.7%	60.0%	59.0%	21.6%	38.5%	
D. mel - D. sim	**8**	8.0%	90.0%	33.0%	61.7%	28.5%	
D. mel - D. yak	**10**	11.1%	90.0%	32.0%	62.0%	27.8%	
D. sim - D. yak	**12**	9.5%	86.0%	36.0%	56.5%	29.3%	
3-way	**1**	18.9%	92.7%	31.9%	63.8%	28.9%	
D. mel - D. sim	**9**	3.0%	97.0%	60.0%	38.7%	58.8%	CAG; CA[AG]; TCC
D. mel - D. yak	**12**	2.3%	95.0%	60.0%	37.0%	57.7%	CAG; CCN
D. sim - D. yak	**4**	2.2%	95.0%	61.0%	36.8%	58.4%	
D. mel - D. sim	**10**	8.2%	98.0%	51.0%	47.9%	50.3%	CAG
D. mel - D. yak	**0**	4.1%	98.0%	51.0%	47.1%	50.6%	CAG
D. sim - D. yak	**10**	4.3%	98.0%	52.0%	47.2%	50.7%	
3-way	**9**	8.6%	94.0%	56.4%	40.5%	53.5%	CAG; C[AC][AC]
D. mel - D. sim	**12**	11.0%	95.0%	42.0%	55.0%	40.3%	CA
D. mel - D. yak	**11**	11.7%	94.0%	40.0%	56.3%	37.9%	CA
D. sim - D. yak	**5**	12.8%	93.0%	41.0%	55.9%	37.3%	
D. mel - D. sim	**13**	5.9%	95.0%	54.0%	43.2%	51.4%	CAG
D. mel - D. yak	**2**	7.9%	93.0%	53.0%	43.4%	49.3%	CAG
D. sim - D. yak	**6**	7.8%	93.0%	53.0%	43.3%	49.9%	
3-way	**12**	11.2%	88.0%	47.2%	46.5%	41.5%	CA
D. mel - D. sim	**14**	0.7%	44.0%	34.0%	33.3%	11.1%	
D. sim - D. yak	**14**	0.5%	52.0%	34.0%	38.7%	13.4%	
3-way	**4**	0.3%	39.7%	34.5%	28.3%	11.4%	
D. mel - D. yak	**5**	3.2%	74.0%	25.0%	58.7%	14.9%	
D. sim - D. yak	**9**	6.4%	78.0%	27.0%	59.8%	18.7%	
3-way	**2**	4.5%	70.6%	26.9%	54.9%	15.7%	
D. mel - D. yak	**6**	2.5%	84.0%	56.0%	35.9%	48.4%	CAG
D. sim - D. yak	**13**	6.8%	85.0%	52.0%	40.3%	44.3%	
3-way	**8**	3.6%	79.6%	55.0%	34.7%	44.9%	
D. mel - D. sim	**11**	0.7%	56.0%	17.0%	52.1%	3.9%	
3-way	**11**	0.4%	47.1%	22.4%	43.5%	3.6%	
D. mel - D. yak	**9**	7.5%	85.0%	45.0%	47.1%	37.7%	
3-way	**5**	4.3%	81.4%	37.3%	52.1%	29.3%	
Classes with no comparable class in other models							
3-way	**10**	0.2%	47.9%	47.1%	25.6%	22.4%	
3-way	**13**	1.8%	63.2%	46.8%	33.4%	29.8%	

Supplementary Table 7 - Enrichment of PicTar miRNA targets in segment classes

Expected number of observations calculated by the total number of annotations covered by the *D. melanogaster* sequence in the D. melanogaster vs. D. simulans alignment multiplied by the proportion of bases which is covered by the given segment class

O: Observed; E: Expected

Only showing the elements considered as significant at the 0.05 level, after Bonferroni correction (actual p-vaue cutt-off: 0.00034722)

Class	**0**		**1**		**2**		**3**		**4**		**5**		**6**		**7**		**8**		**9**		**10**		**11**		**12**		**13**		**14**		**No Class**		**TOTAL**	**p-value**
proportion of alignment	0.08036		0.050558		0.005		0.0021		0.04623		0.011		0.018		6E-04		0.0274		0.02		0.0426		0.002		0.0301		0.025		0.002		0.6373281		1	
miRNA	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E		
dme-miR-972-5p	30	32.2	115	20.3	0	2.1	19	0.8	30	18.5	0	4.5	7	7.2	0	0.2	5	11.0	0	8.1	0	17.1	0	0.6	3	12.1	1	9.9	0	0.6	191	255.6	401	3.54E-71
dme-miR-289-5p	4	7.6	18	4.8	0	0.5	5	0.2	6	4.4	0	1.1	1	1.7	0	0.1	0	2.6	0	1.9	0	4.0	0	0.2	1	2.9	0	2.3	0	0.2	60	60.5	95	1.40E-11
dme-miR-1000-5p	11	6.7	16	4.2	0	0.4	2	0.2	3	3.8	0	0.9	0	1.5	0	0.1	0	2.3	0	1.7	0	3.5	0	0.1	1	2.5	0	2.0	0	0.1	50	52.9	83	1.35E-07
dme-miR-6-3p	16	8.2	15	5.2	0	0.5	0	0.2	10	4.7	0	1.2	0	1.8	0	0.1	0	2.8	0	2.1	3	4.3	0	0.2	0	3.1	0	2.5	0	0.2	58	65.0	102	9.96E-07
dme-miR-2c-3p	15	7.7	14	4.9	0	0.5	0	0.2	10	4.4	0	1.1	0	1.7	0	0.1	0	2.6	0	1.9	3	4.1	0	0.2	0	2.9	0	2.4	0	0.2	54	61.2	96	3.03E-06
dme-miR-2a-3p	14	7.7	15	4.9	0	0.5	0	0.2	8	4.4	0	1.1	0	1.7	0	0.1	0	2.6	0	1.9	3	4.1	0	0.2	0	2.9	0	2.4	0	0.2	56	61.2	96	4.79E-06
dme-miR-13a-3p	13	7.3	14	4.6	0	0.5	0	0.2	8	4.2	0	1.0	0	1.6	0	0.1	0	2.5	0	1.8	3	3.9	0	0.1	0	2.7	0	2.2	0	0.1	53	58.0	91	1.49E-05
dme-miR-5-3p	13	7.3	14	4.6	0	0.5	0	0.2	8	4.2	0	1.0	0	1.6	0	0.1	0	2.5	0	1.8	3	3.9	0	0.1	0	2.7	0	2.2	0	0.1	53	58.0	91	1.49E-05
dme-miR-308-3p	13	7.3	14	4.6	0	0.5	0	0.2	8	4.2	0	1.0	0	1.6	0	0.1	0	2.5	0	1.8	3	3.9	0	0.1	0	2.7	0	2.2	0	0.1	53	58.0	91	1.49E-05
dme-miR-2b-3p	13	7.3	14	4.6	0	0.5	0	0.2	8	4.2	0	1.0	0	1.6	0	0.1	0	2.5	0	1.8	3	3.9	0	0.1	0	2.7	0	2.2	0	0.1	53	58.0	91	1.49E-05
dme-miR-11-3p	13	7.3	14	4.6	0	0.5	0	0.2	8	4.2	0	1.0	0	1.6	0	0.1	0	2.5	0	1.8	3	3.9	0	0.1	0	2.7	0	2.2	0	0.1	53	58.0	91	1.49E-05
dme-miR-13b-3p	13	7.3	14	4.6	0	0.5	0	0.2	8	4.2	0	1.0	0	1.6	0	0.1	0	2.5	0	1.8	3	3.9	0	0.1	0	2.7	0	2.2	0	0.1	53	58.0	91	1.49E-05
dme-miR-277-3p	22	10.0	12	6.3	0	0.7	1	0.3	8	5.8	0	1.4	0	2.3	0	0.1	0	3.4	0	2.5	4	5.3	0	0.2	3	3.8	0	3.1	0	0.2	75	79.7	125	1.98E-05
dme-miR-1014-3p	13	5.6	11	3.5	0	0.4	1	0.1	3	3.2	0	0.8	1	1.3	0	0.0	0	1.9	0	1.4	0	3.0	0	0.1	1	2.1	0	1.7	0	0.1	40	44.6	70	0.00012977
dme-miR-263b-5p	13	3.9	3	2.5	0	0.3	0	0.1	3	2.3	0	0.6	0	0.9	0	0.0	0	1.3	0	1.0	0	2.1	0	0.1	7	1.5	0	1.2	0	0.1	23	31.2	49	0.000259158

Supplementary Table 8 - Enrichment of UTRdb motifs in segment classes

Expected number of observations calculated by the total number of annotations covered by the *D. melanogaster* sequence in the *D. melanogaster* vs. *D. simulans* alignment multiplied by the proportion of bases which is covered by the given segment class

O:Observed; E:Expected

Only showing elements considered significant at the 0.05 level, after Bonferroni correction (actual p-vaue cutt-off: 0.005)

Class	0		1		2		3		4		5		6		7		8		9		10		11		12		13		14		No Class		TOTAL	p-value
proportion of	0.08036		0.05056		0.00524		0.00210		0.04623		0.01134		0.01805		0.00062		0.02738		0.02022		0.04260		0.00161		0.03014		0.02461		0.00161		0.63733		1	
alignment	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E		
PAS	432	572.6	866	360.3	23	37.4	59	15.0	771	329.4	185	80.8	428	128.6	1	4.4	278	195.1	5	144.1	47	303.6	84	11.4	123	214.8	25	175.4	26	11.5	3773	4541.6	7126	0
CPE	2	16.7	11	10.5	0	1.1	3	0.4	6	9.6	1	2.4	9	3.8	0	0.1	2	5.7	0	4.2	0	8.9	0	0.3	0	6.3	0	5.1	0	0.3	174	132.6	208	3.26E-09
BRD-BOX	40	31.3	21	19.7	0	2.0	0	0.8	23	18.0	7	4.4	6	7.0	0	0.2	9	10.7	0	7.9	10	16.6	0	0.6	13	11.7	1	9.6	2	0.6	257	247.9	389	2.18E-05
K-BOX	60	34.5	30	21.7	2	2.2	1	0.9	18	19.8	5	4.9	2	7.7	1	0.3	7	11.7	3	8.7	11	18.3	0	0.7	16	12.9	4	10.6	0	0.7	269	273.4	429	3.01E-05
SXL_BS	3	7.3	3	4.6	1	0.5	0	0.2	8	4.2	5	1.0	5	1.6	0	0.1	5	2.5	0	1.8	1	3.9	1	0.1	1	2.7	0	2.2	0	0.1	58	58.0	91	0.0002481

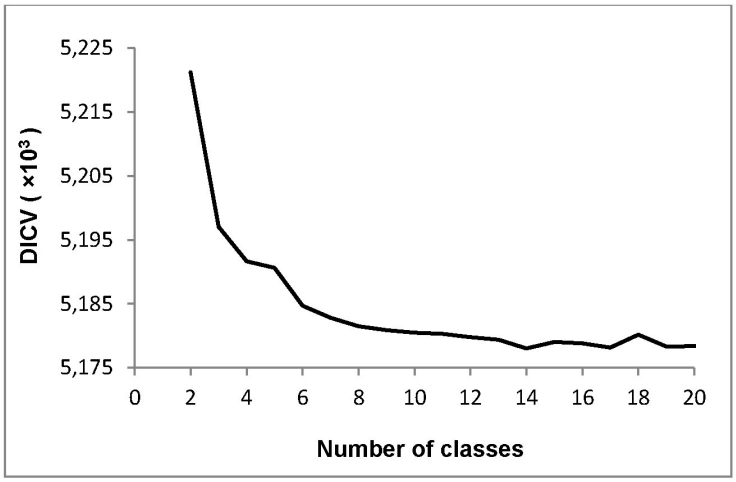


***Figure S1**: DICV values for segmentation of 3-way alignment. DICV values obtained using 1-20 segment classes for D. melanogaster, D. simulans and D. yakuba 3′ UTR alignment. The 14-class model was selected as minimum DICV has occurred at class 14.*

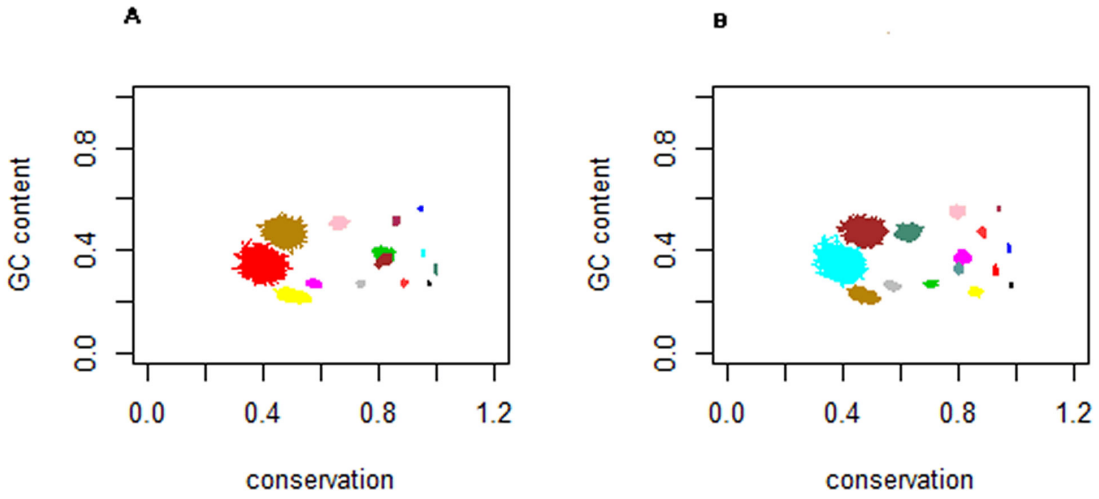


***Figure S2**: GC content versus conservation level for models selected for 3-way alignment. GC content of D. melanogaster versus the proportion of alignment matches, for each model selected for the 3-way 3′ UTR alignment. A) 14-class model selected by Procedure 1 and B) 15-class model selected by Procedure 2. The different colours represent different classes, and each class is plotted for the post burn-in samples. This plot was used to access the convergence of the selected models.*

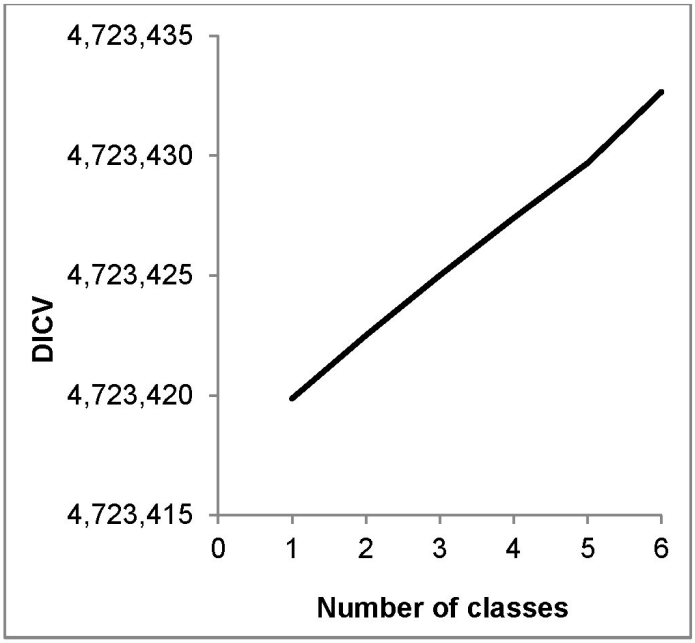


Figure S3: *DICV values for the control sequence. DICV values were obtained for an artificially generated sequence having only one class of segments. The minimum DICV has occurred at 1-class; therefore justifies models selected by Procedure 1.*

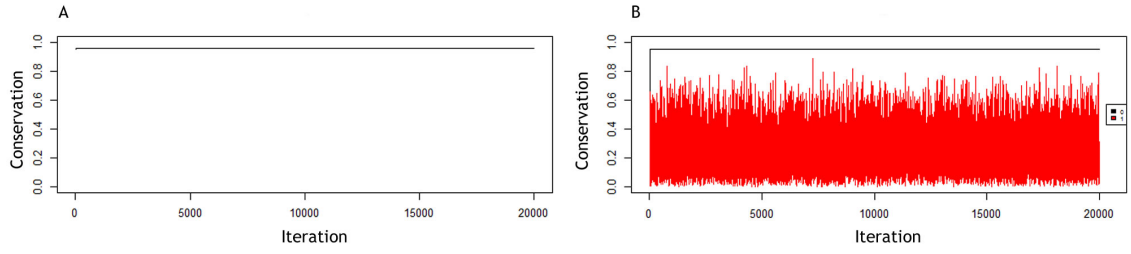


Figure S4: *Conservation level vs sample number for control sequences. Figure shows time-series plots of conservation level versus sample number for segmentations of artificially generated control sequence with A) 1 segment class and B) 2 segment classes.*

Appendix B

Appendix Chapter 4

Table B.1: *Assessing the convergence of chromosome 1: 19-class model using Heidelberger and Welch test*

Class	Stationarity test result	Start iteration	p-value	Half-width test	Mean	Half-width
class0	passed	201	0.27	passed	0.698	0.000505
class1	passed	1	0.473	passed	0.515	0.0598
class2	passed	101	0.254	passed	0.352	0.000726
class3	passed	301	0.55	passed	0.543	0.000619
class4	passed	201	0.649	passed	0.34	0.000458
class5	passed	401	0.387	passed	0.39	0.000359
class6	passed	301	0.734	passed	0.445	0.00101
class7	passed	101	0.459	passed	0.709	0.000411
class8	passed	1	0.359	passed	0.614	0.0159
class9	passed	301	0.411	passed	0.599	0.00058
class10	passed	101	0.589	passed	0.35	0.000583
class11	passed	101	0.0509	passed	0.379	0.000184
class12	passed	201	0.471	passed	0.42	0.000563
class13	passed	1	0.192	passed	0.338	0.0278
class14	passed	201	0.154	passed	0.56	0.00111
class15	passed	1	0.381	passed	0.73	0.0541
class16	passed	1	0.148	passed	0.724	0.00358
class17	passed	1	0.0583	passed	0.427	0.00162
class18	passed	301	0.058	passed	0.369	0.00166

The stationarity test results confirm that the 19-class model selected for chromosome 1 has achieved the convergence. The half-width tests indicate that posterior samples provide precise estimates of conservation level (mean values) of each class using the CODA accuracy criterion of $\epsilon = 0.12$.

Table B.2: *PFEs discarded from the genome-wide analysis*

Zebrafish genomic position	Zebrafish gene ID	Zebrafish gene name	Human genomic position	Human gene name
chr22:19331908-19332130	ENSDARG00000061658	*polrmt*	chr5:44392191-44392395	*fgf10-as1*
chr15:47316561-47316687	ENSDARG00000060524	*zgc:153039*	chr5:178740426-178740551	*adamts2*
chr8:3492911-3493335	ENSDARG00000086603	*zgc:136963*	chr9:127147857-127148219	*psmb7*
chr7:446331-446531	ENSDARG00000090143	*cabZ01074659.1*	chr10:135151556-135151746	*znf511*
chr11:9441898-9442024	ENSDARG00000086489	*cu179643.1*	chr3:174811682-174811801	*naaladl2*
chr19:510861-510972	ENSDARG00000077668	*cabz01010103.1*	chr2:178215677-178215783	*nfe2l2*
chr21:220613-220766	ENSDARG00000007915	*jak2a*	chr7:158128055-158128194	*ptprn2*
chr21:221576-221720	ENSDARG00000007915	*jak2a*	chr7:158129217-158129341	*ptprn2*
chr21:222010-222165	ENSDARG00000007915	*jak2a*	chr7:158129790-158129927	textitptprn2
chr21:222861-222961	ENSDARG00000007915	*jak2a*	chr7:158130852-158130952	textitptprn2

Table B.3: *GO terms related to Transcription Factors*

GO term	GO category	Backgroud frequency	Sample frequency	Expected	p-value
nucleic acid binding transcription factor activity	MF	653	26	4.12	1.79E-10
sequence-specific DNA binding transcription factor activity	MF	653	26	4.12	1.79E-10
sequence-specific DNA binding RNA polymerase II transcription factor activity	MF	239	9	1.51	0.0449
transcription from RNA polymerase II promoter	BP	218	9	1.38	0.0317
regulation of transcription, DNA-templated	BP	1468	55	9.26	1.71E-24
regulation of nucleic acid-templated transcription	BP	1470	55	9.28	1.83E-24
nucleic acid-templated transcription	BP	893	33	5.63	6.37E-13
transcription, DNA-templated	BP	893	33	5.63	6.37E-13
regulation of transcription from RNA polymerase II promoter	BP	452	13	2.85	0.0185

Background frequency- frequency of the GO term in question in all zebrafish genes; Sample frequency- frequency of the GO term in 193 genes containing PFEs identified in genome-wide analysis; MF: Molecular Function; BP: Biological Process

Table B.4: *UCSC genomic coordinates of PFEs identified in pathway-focussed analysis*

Gene	PFE number	Human genomic position	Zebrafish genomic position
eya1	1	chr8:72129606-72129770	chr24:13909370-13909591
eya1	2	chr8:72130576-72130808	chr24:13907414-13907687
eya1	3	chr8:72155916-72156061	chr24:13891420-13891667
eya1	4	chr8:72267639-72267809	chr24:13836991-13837100
eya1	5	chr8:72270664-72270891	chr24:13832262-13832548
eya1	6	chr8:72271490-72271729	chr24:13831700-13831937
eya4	1	chr6:133652590-133652771	chr23:31734041-31734235
eya4	2	chr6:133778149-133778268	chr23:31712298-31712447
shha	1	chr7:155601378-155601490	chr7:43615955-43616135
shha	2	chr7:155603492-155603599	chr7:43614779-43614886
pax3a	1	chr2:223078919-223079145	chr2:47629182-47629489
pax3a	2	chr2:223083770-223083947	chr2:47622371-47622515
pax3a	3	chr2:223133797-223134194	chr2:47593793-47594123
pax3a	4	chr2:223137113-223137779	chr2:47590024-47590492
pax3a	5	chr2:223153695-223153821	chr2:47587337-47587505
pax3a	6	chr2:223156764-223157217	chr2:47584072-47584367
pax3a	7	chr2:223135284-223135480	chr2:47592582-47592824
pax3b	1	chr2:223105881-223106300	chr15:40712700-40712983
pax3b	2	chr2:223153529-223153656	chr15:40722774-40722842
pax7b	1	chr1:18965463-18965874	chr23:21155541-21155912
pax7b	2	chr1:18973120-18973324	chr23:21159261-21159464
pax7b	3	chr1:18984276-18984573	chr23:21160493-21160791
pax7b	4	chr1:19037049-19037309	chr23:21193990-21194229
pax7b	5	chr1:19050405-19050528	chr23:21205291-21205404
pax7b	6	chr1:19057495-19057596	chr23:21214580-21214683
myf5	1	chr12:81111797-81111914	chr4:20679147-20679234
six4.3	1	chr14:61189618-61189728	chr18:35590841-35590970

Appendix C

Appendix Chapter 5

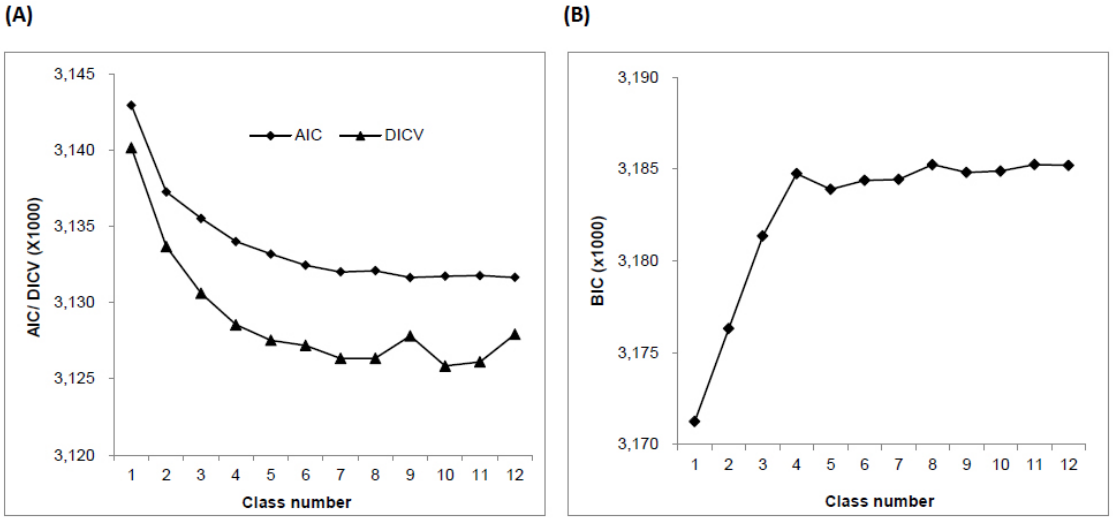


Figure A*: Selection of optimal number of classes. We used approximations to the well-known information criteria AIC, BIC and DIC to identify the number of distinct classes of conservation levels. Generally, a lower value of the information criteria indicates a better model. BIC favoured a 1-class model, which is inappropriate. We therefore based our judgement on AIC and DICV and selected the 7-class model as the first local minimum of AIC and DICV has occurred at seven classes.*

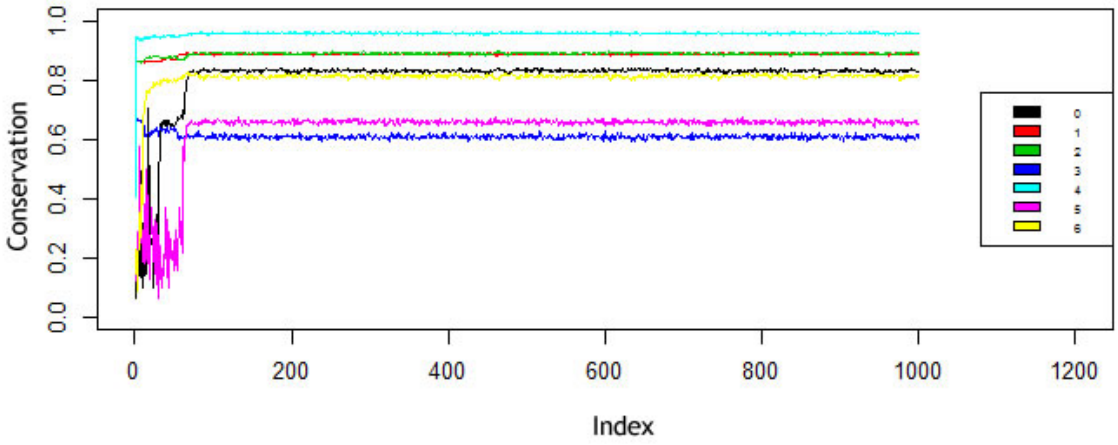


***Figure B**: Identifying the most conserved class. The mean proportion of alignment matches was plotted against each iteration of the sampler to identify the class that contains the most conserved segments in* w*Mel and* w*Pip (Class 4). The different colours represent different classes in the 7-class model.*

```
wMel     TGTAGCGTTATGAATTAGGAGTGCTATATTAAAGCTTACCTCACTATTAAAGCTATCGGTCAGATTAGATTAAAAACCTAATCTGACCGGTTTC
wMelCS   TGTAGCGTTATGAATTAGGAGTGCTATATTAAAGCTTACCTCACTATTAAAGCTATCGGTCAGATTAGATTAAAAACCTAATCTGACCGGTTTC
wMelPop TGTAGCGTTATGAATTAGGAGTGCTATATTAAAGCTTACCTCACTATTAAAGCTATCGGTCAGATTAGATTAAAAACCTAATCTGACCGGTTTC
wAu      TGTAGCGTTATGAATTAGGAGTGCTATATTAAAGCTTACCTCACTATTAAAGCTGTCGGTCAGATTAGATTAAAAACCTAATCTGACCGGTTTC
```

Fig. C. Sequence alignment of the *ncrwmel02* amplicon from the published genome data of *w*Mel (Wu et al 2004), *w*MelCS, *w*MelPop (Woolfit et al 2013) and *w*Au (Sutton et al 2014).

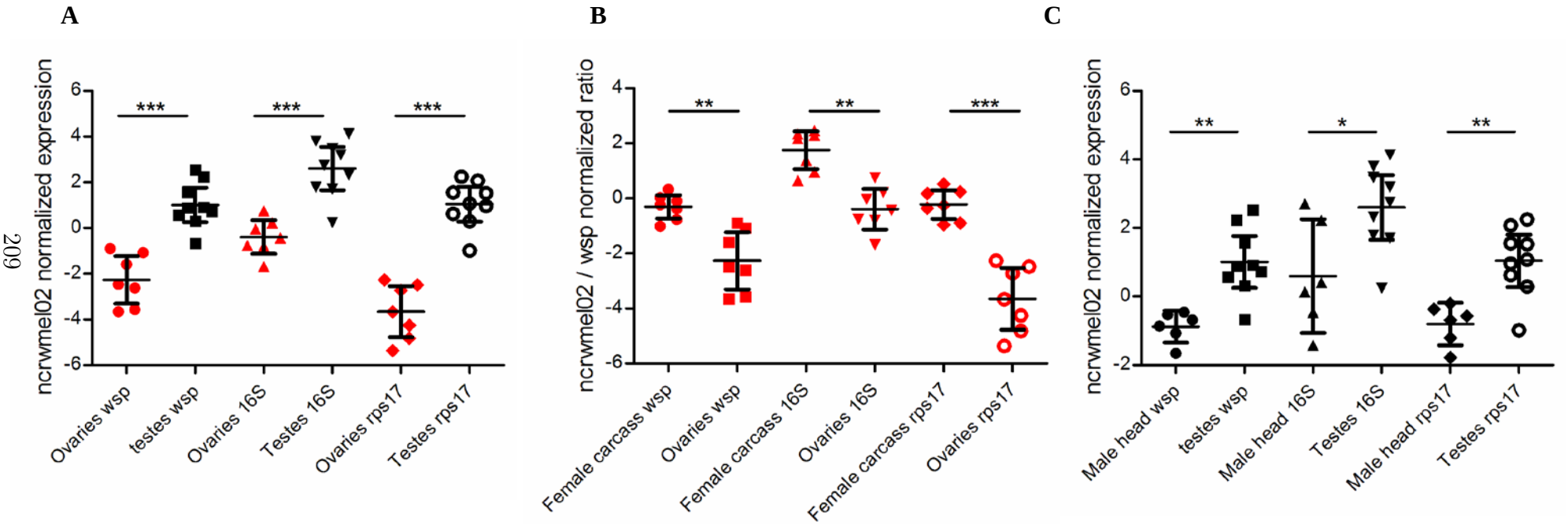


Fig. D. Validation of *ncrwmel02* differential expression observed using *wsp* as reference gene in dissected tissues of *w*Mel-infected male (black) or female (red) *D. melanogaster*. *ncrwmel02* expression (mean ± 95% CI) normalized to *wsp*, 16S or *rps17* expression (Mann-Whitney U test, * $p < 0.1$, ** $p < 0.01$ *** $p < 0.001$). Panel A: *ncrwmel02* expression in male and female gonads. Panel B: *ncrwmel02* expression in female dissected tissues. Panel C: *ncrwmel02* expression in male dissected tissues

Table A. Oligonucleotides used in this study.

Sequence (5'-3')	Description	Reference
5'RACE primers		
AUAUGCGCGAAUUCCUGUAGAACGAACACUAGAAGAAA	RNA adaptor	[47]
GCGCGAATTCCTGTAGA	Adaptor specific PCR primer	[47]
GGATCTATGTTAAGAGATACCGTGAA	IGR-60-specific RT primer	This study
ATGACGGTTCGTGACGGTAT	IGR-60-specific PCR primer	This study
GCAGCTTAATCTTGCTTGTCA	IGR-151-specific RT primer	This study
ACGCCAATATTTTAAAGCGGATA	IGR-151-specific PCR primer	This study
GCAAAAGAAGCCCTGAGGTT	IGR-392-specific RT primer	This study
AGAAGCCCTGAGGTTATTATCCGCT	IGR-392-specific PCR primer	This study
TCGCACTACGTGCATCGCAT	IGR-446-specific RT primer	This study
CTACGTGCATCGCATGTCTT	IGR-446-specific PCR primer	This study
TTTCAAGCTTTGCCAAAAGAA	IGR-498-specific RT primer	This study
CCCCAATCAAAACAGCCTTA	IGR-498-specific PCR primer	This study
CACTTGAGCGATGCAACAAAGCCA	IGR-760-specific RT primer	This study
AACAAAGCCATCCCAGTGTC	IGR-760-specific PCR primer	This study
ATGGGAGGGAAGCAAAATCT	IGR-781-specific RT primer	This study
GGGAAGCAAAATCTGGCTTAATGGC	IGR-781-specific PCR primer	This study
TTGCATGACACCCTGACAAC	IGR-834-specific RT primer	This study
GCTACGTGTTAGCGGGATCT	IGR-834-specific PCR primer	This study
CGCTCGTGCACAAATTAAAA	IGR-884-specific RT primer	This study
TGTAGCGTTATGAATTAGGAGTGC	IGR-884-specific PCR primer	This study
CATAGATCCCGCTAACACGTAG	IGR-921-specific RT primer	This study
AGCCCCGTGGTTATTATCTG	IGR-921-specific PCR primer	This study
CCCGTGTTAGCTAGTTGTCACTCCC	IGR-1021-specific RT primer	This study
ATCCTGCAAATTGGCGTACT	IGR-1021-specific PCR primer	This study
AGCAGTGGGATGACGAGACT	IGR-1035-specific RT primer	This study
AAAGAAGCCCCGTGGTTGGC	IGR-1035-specific PCR primer	This study
CGAGATTCAGCCGCTTTTA	IGR-1047-specific RT primer	This study
GCAACTAACCTACGCTGCAA	IGR-1047-specific PCR primer	This study
TGTATTTGGCGTAAATCATGC	IGR-1049-specific RT primer	This study
GCACTATGTGCACCTCATGTCT	IGR-1049-specific PCR primer	This study
RT-PCR primers		
TGGATCCCAGTGTCAAGCAC	IGR-60 Fwd in IGR	This study
ACGACAATCGTCATCCCAGC	IGR-60 Rev in downstream CDS	This study
CGACGGCATGACGATAAGGT	IGR-60 Rev in IGR	This study
GCTGTTTTGATTGGGGTCTT	IGR-498 Fwd in IGR	This study
TCGTATCGGGCAAGAACGTA	IGR-498 Rev in downstream CDS	This study
TTTCAAGCTTTGCCAAAAGAA	IGR-498 Rev in IGR	This study
GAAACCGGTCAGATTAGGTTTTT	IGR-884 Fwd in IGR	This study
CCGTAACCGGCACTGAAGTA	IGR-884 Rev in downstream CDS	This study
TGTAGCGTTATGAATTAGGAGTGC	IGR-884 Rev in IGR	This study
AACAACGTAGTTGGCGTCTT	IGR-1021 Fwd in IGR	This study
AGCACTGGGATGACACCATT	IGR-1021 Rev in downstream CDS	This study
AACAACGTAGTTGGCGTCTT	IGR-1021 Rev in IGR	This study
qPCR primers		
Fwd GAAACCGGTCAGATTAGGTTTTT Rev TGTAGCGTTATGAATTAGGAGTGC	*ncrwmel02*qPCR primers	This study
Fwd ATCTTTTATAGCTGGTGGTGGT Rev GGAGTGATAGGCATATCTTCAAT	*wsp* qPCR primers	[6]
Fwd CGGTGAATACGTTCTCGGGTC Rev CACCCCAGTCACCGATCCC	*16S* qPCR primers	This study
Fwd CACTCCCAGGTGCGTGGTAT Rev GGAGACGGCCGGGACGTAGT	*rps17* qPCR primers	[63]

Table B. Highly conserved non CDS predicted by *changept*.*

wMel coordinates	Type of conserved feature	Profile value	Length (nt)
1,739-2,162	pseudo WD0002	1.0	424
2,274-2,503	pseudo WD0002	0.5	230
3,024-3,118	tRNA	0.5	95
44,380-44,468	intergenic	1.0	89
83,877-83,957	tRNA	0.7	81
85,867-85,929	intergenic	1.0	63
117,042-117,328	ncRNA tmRNA	0.6	287
124,753-124,836	tRNA	0.9	84
182,216-185,396	rRNA 23S+5S	1.0	3181
279,526-279,619	intergenic	0.9	94
372,011-372,117	tRNA	0.8	107
513,727-513,814	tRNA	0.9	88
547,479-547,732	intergenic	1.0	254
611,202-611,370	intergenic	1.0	169
612,281-612,391	intergenic	0.7	111
622,779-622,923	intergenic	0.5	145
623,094-623,293	intergenic	1.0	200
639,293-639,403	intergenic	1.0	111
706,682-707,007	ncRNA rnpB1	0.9	326
719,048-719,171	intergenic	0.5	99
722,484-722,594	tRNA	0.8	111
723,861-724,026	intergenic	1.0	166
764,459-764,871	intergenic	1.0	413
768,936-768,988	intergenic	0.7	53
793,553-793,636	tRNA	0.8	84
840,429-840,513	tRNA	1.0	85
850,067-850,142	intergenic	0.9	76
932,596-932,693	intergenic	0.5	98
934,908-935,042	tRNA	1.0	135
935,321-935,403	tRNA	0.8	83
940,039-940,142	intergenic	1.0	104
941,714-941,808	tRNA	0.6	95
941,823-941,975	intergenic	1.0	153
970,671-970,776	tRNA	1.0	106
1,039,579-1,039,870	intergenic	1.0	292
1,105,661-1,105,744	intergenic	0.9	84
1,107,303-1,107,403	tRNA	1.0	101
1,152,142-1,152,229	tRNA	0.7	88
1,158,600-1,158,694	tRNA	1.0	95
1,167,332-1,169,526	rRNA 16S	1.0	2195
1,186,283-1,186,373	tRNA	1.0	91
1,208,797-1,208,903	tRNA	0.5	107

* Thresholds used: 1. Conservation = 0.95 (conservation level of the most conserved class-Class 4); 2. Profile value ≥0.5 (probability that each position in the conserved feature belongs to Class 4); 3. Length >50 nt (length of the conserved feature)

Table C. 5' RACE of intergenic regions (IGR) plasmid sequences*.

IGR coordinates in wMel genome	Sequence
IGR-60	CACTAGTGATT**GCGCGAATTCCTGTAGAACGAACATTAGAAGAAAAAAAAACATTG TATTTTAACGTAAAACAGCTATTTTTATGCTCACCAACTTAATAAAATTCCTGGATC CCAGTGTCAAGCACTGGGATGACAAGATATAAACCTTATCGTCATACCGTCACGAA CCGTCAT**AATCGAATTCCCGCGGCCGCC
IGR-498	CACTAGTGATT**GCGCGAATTCCTGTAGACTAGAAGAAAAAAAACTTCTTTTAAGATT AAGCGTTTGAAAGGTTTTTGAGTAAGGCTGTTTTGATTGGGG**AATCGAATTCCCGCG GCCGCC
IGR-884	CACTAGTGATT**GCGCGAATTCCTGTAGAACACGAAGAAAAGAGTTTAGAGGGTTAT AGAGAAACCGGTCAGATTAGGTTTTTAATCTAATCTGACCGATAGCTTTAATAGTG AGGTAAGCTTTAATATAGCACTCCTAATTCATAACGCTACA**AATCGAATTCCCGCGG CCGCC
IGR-1021	CACTAGTGATT**ATCCTGCAAATTGGCGTACTATACTGTCTTAAACGACTTATAAGCG CGTTTCAGCTTGTGCAGGTAAAAACCTAGAATATTGTGAAGACATAAGGTGCACAT AGTGCAAAAAATTAAAAATAAGACGCCAACTACGTTGTTTTCTTGCTGTTTAATCT GCACAGATGAAGATAACTGAATGCCTTCTTTCTTCTAGTTTCTACAGG**AATTCGCGC AATCGAATTC CCGCGGCCGCC
IGR-1047	GGCCGCGGGAATTCGATT**GCGCGAATTCCTGTAGAATAGAAGAAAGAAGCTATTGT ATTTGCTTTCGCCAATCTGCAGATTAAAAGGTAAGGATTACTTAATGTATCGGCGT CTTATGTTCAATTTTTTGCAGTATATAGATACTGTATGTCTTTACAAAACTTCATCT ACATCTAGATTTTTATCTAAATAAGCTGAACGCGCTTATAAAGCGTTACAAGACGT CAAAAAATGCCAATACTCGACAGAGATAGTAAAAGACTAGCTAACTCGGGGATTCT TTGTCTTTTTTTCTGCTTAGTAAATTTCTTAAACATTTGCAGCGTAGGTTAGTTGCA**ATCACTAGTGAATTCGCG

* insert in pGEMTeasy in bold

Appendix D

Appendix Chapter 6

Table D.1: *Proportion of genes annotated to Biological Process Gene Ontology (with over 100 genes) terms with overlapping segments.*

GO term	Description	Proportion
GO:0020033	Antigenic Variation	0.01
GO:0006468	Protein Phosphorylation	0.60
GO:0006412	Translation	0.60
GO:0006810	Transport	0.59
GO:0008152	Metabolic process	0.79
GO:0006355	Regulation of Transcription DNA-templated	0.59

Table D.2: *Genome data summary.*

Species	Preferred Hosts	Assembly Version	Description	GC content	Reference
Plasmodium falciparum	*Homo sapiens*	V3	22.9Mb across 14 chromosomes	0.190	Gardner et al. 2002. *Nature*
Plasmodium reichenowi	*Pan troglodytes*	September 2013	21.4Mb across 14 chromosomes	0.184	*Unpublished*
Plasmodium gallinaceum[1]	*Gallus sp.*	September 2013	21.6Mb across 4996 contigs		*Unpublished*

Table D.3: *Genome content distance matrix produced by* progressiveMauve.

	Pf	**Pr**	**Pg**
Pf	0.000		
Pr	0.230	0.000	
Pg	0.636	0.621	0.000

Table D.4: *Number of nucleotides covered by three way alignments across LCBs of different minimum lengths.*

Minimum Size	*Pf*	*Pr*	*Pg*
100	4,082,818	4,082,692	3,972,363
200	3,677,293	3,677,681	3,572,734
300	3,320,095	3,320,956	3,221,897
400	2,976,512	2,977,592	2,884,842
500	2,643,857	2,645,655	2,560,179

Table D.5: *Individual segment class summary characteristics as estimated from the Bayesian segmentation model*

Class ID	Mixture Proportion	Conservation Level	Proportion Pg-specific changes	Proportion Pf-specific changes	Proportion Pr-specific changes	Proportion all three species change	Pf/Pr percent identity	Pf/Pg percent identity	Pr/Pg percent identity	Pf- GC level
0	0.004	0.784	0.114	0.038	0.047	0.016	0.898	0.831	0.822	0.111
1	0.158	0.780	0.209	0.005	0.005	0.003	0.988	0.784	0.784	0.209
2	0.113	0.663	0.318	0.007	0.007	0.005	0.981	0.670	0.670	0.197
3	0.126	0.854	0.135	0.004	0.004	0.002	0.989	0.858	0.858	0.220
4	0.050	0.662	0.307	0.011	0.012	0.008	0.969	0.674	0.673	0.122
5	0.075	0.785	0.201	0.005	0.005	0.004	0.985	0.790	0.790	0.283
6	0.042	0.824	0.141	0.013	0.013	0.008	0.965	0.837	0.837	0.267
7	0.015	0.672	0.226	0.040	0.042	0.019	0.898	0.714	0.713	0.032
8	0.104	0.804	0.180	0.006	0.006	0.004	0.984	0.810	0.810	0.305
9	0.073	0.612	0.353	0.012	0.012	0.011	0.965	0.624	0.624	0.257
10	0.063	0.679	0.299	0.008	0.007	0.007	0.978	0.687	0.687	0.282
11	0.005	0.524	0.440	0.011	0.012	0.012	0.964	0.537	0.536	0.197
12	0.077	0.859	0.128	0.005	0.005	0.003	0.987	0.864	0.863	0.303
13	0.001	0.506	0.106	0.143	0.161	0.084	0.611	0.666	0.649	0.272
14	0.007	0.590	0.294	0.042	0.041	0.033	0.884	0.631	0.632	0.309
15	0.033	0.606	0.335	0.019	0.020	0.019	0.942	0.627	0.625	0.195
16	0.002	0.475	0.189	0.133	0.124	0.080	0.664	0.598	0.608	0.219
17	0.051	0.747	0.221	0.012	0.011	0.009	0.968	0.758	0.759	0.208

Mixture Proportions: proportion of segments belong to each segment class; Conservation levels: alignment columns where all species are same; Pg-specific changes: alignment coloumns where human (Pf) and chimp (Pr) are same and jungle fowl (Pg) is different; Pf/Pr percent identity: Conservation levels + Pg specific changes;

Table D.6: *Posterior probability distribution of mutation for each segment class and codon position*

Class ID	Codon position	Q1	Median	Q3
0	First	0.009	0.064	0.196
0	Second	0.009	0.064	0.196
0	Third	0.090	0.215	0.394
1	First	0.169	0.173	0.177
1	Second	0.106	0.109	0.113
1	Third	0.352	0.357	0.362
2	First	0.338	0.347	0.355
2	Second	0.239	0.247	0.255
2	Third	0.447	0.456	0.465
3	First	0.075	0.077	0.079
3	Second	0.034	0.035	0.037
3	Third	0.289	0.293	0.297
4	First	0.337	0.379	0.422
4	Second	0.214	0.250	0.289
4	Third	0.383	0.426	0.469
5	First	0.155	0.160	0.165
5	Second	0.083	0.086	0.090
5	Third	0.408	0.414	0.421
6	First	0.094	0.098	0.102
6	Second	0.037	0.040	0.043
6	Third	0.383	0.390	0.396
7	First	0.326	0.454	0.586
7	Second	0.276	0.399	0.531
7	Third	0.195	0.307	0.436
8	First	0.110	0.114	0.119

Continued on next page

Table D.6 – *Continued from previous page*

Class ID	Codon position	Q1	Median	Q3
8	Second	0.046	0.049	0.052
8	Third	0.417	0.423	0.430
9	First	0.395	0.410	0.424
9	Second	0.295	0.308	0.322
9	Third	0.521	0.535	0.550
10	First	0.298	0.308	0.319
10	Second	0.204	0.213	0.223
10	Third	0.501	0.513	0.525
11	First	0.298	0.411	0.530
11	Second	0.249	0.357	0.476
11	Third	0.524	0.643	0.751
12	First	0.049	0.051	0.053
12	Second	0.015	0.016	0.017
12	Third	0.334	0.338	0.342
13	First	0.037	0.092	0.181
13	Second	0.000	0.011	0.059
13	Third	0.408	0.533	0.654
14	First	0.239	0.290	0.344
14	Second	0.236	0.286	0.340
14	Third	0.516	0.574	0.630
15	First	0.402	0.431	0.460
15	Second	0.291	0.318	0.345
15	Third	0.482	0.511	0.540
16	First	0.485	0.615	0.735
16	Second	0.363	0.491	0.620
16	Third	0.522	0.651	0.766

Continued on next page

Table D.6 – *Continued from previous page*

Class ID	Codon position	Q1	Median	Q3
17	First	0.213	0.220	0.227
17	Second	0.123	0.129	0.135
17	Third	0.397	0.405	0.414

Table D.7: *Count of unique GO terms for genes with mapping segments and genes without mapping segments. Total count, GO terms that are associated with genes that have mapping segments and genes that do not have mapping segments (shared), GO terms that are associated exclusively with genes that have no mapping segments (missing) are also displayed.*

					Number of genes
	BP	CC	MF	Total	
With Segments	534	237	640	1411	2370
Without Segments	346	174	347	867	2173
Total GO terms	634	289	749	1672	4543
Shared GO Terms	246	122	238	606	–
Missing GO Terms	100	52	109	261	–

Table D.8: *Spearman rank correlation ρ values and associated* p*-values for each of the Biological Process GO terms with > 100 associated genes.*

GO	Description	rho	pvalue
0006355	Regulation Of Transcription, DNA-templated	0.6677	0.003165
0006468	Protein Phosphorylation	0.2838	0.2529
0006464	Cellular Protein Modification Process	0.1662	0.5085
0006412	Translation	0.1187	0.6384
0006508	Proteolysis	0.1187	0.6384
0007018	Microtubule-based Movement	0.01548	0.9541
0006457	Protein Folding	-0.07327	0.767
0006886	Intracellular Protein Transport	-0.2281	0.3567
0006810	Transport	-0.2549	0.302
0055114	Oxidation-reduction Process	-0.2652	0.2825
0008152	Metabolic Process	-0.2693	0.2749
0006511	Ubiquitin-dependent Protein Catabolic Process	-0.3395	0.1655
0006260	DNA Replication	-0.3725	0.1263

Table D.9: *Tally of genes with overlapping segments across the four categories of transcription genes.*

classes	found	total
I. General Transcription Factors	33	56
II. Chromatin-related Factors	40	63
III. Specific Trancription Factors	48	73
IV.TAP partners	7	10

Table D.10: *Spearman rank correlation ρ values and associated* p*-values for each of four categories of genes that control transcription in* Pf.

class	rho	pvalue
I. General Transcription Factors	-0.2343	0.348
II. Chromatin-related Factors	-0.3602	0.1401
III. Specific Trancription Factors	0.5562	0.01767
IV.TAP partners	0.06295	0.7986