Advances in Statistical Methodology and Analysis in a Study of ARC Syndrome

Anne-Marie Lyne

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Statistical Science
University College London

April 19, 2016

I, Anne-Marie Lyne, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

This thesis presents statistical analysis and methodology development for a systems analysis of ARC syndrome. ARC is a genetic disease caused by mutations in one of two proteins, *VPS33B* and *VIPAS39*, of whose function little is known. Transcriptomic and metabolomic data are analysed to identify differentially expressed genes and pathways, and to highlight processes which are perturbed. Results consistently point to processes involved in cell polarisation and cell-cell adhesion, which is corroborated by experimental work. Beneficial suggestions for future experimental work are included and have already yielded interesting results.

Motivated by the desire to incorporate knowledge of genetic dependencies into this analysis, methodology is developed to enable Bayesian inference for 'doubly-intractable distributions'. These models have a likelihood normalising term which is a function of unknown model parameters and which cannot be computed. This means that standard methods for sampling from the posterior, such as Markov chain Monte Carlo (MCMC), cannot be used. In the developed method, the likelihood is expressed as an infinite series which is then stochastically truncated. These unbiased, but possibly negative, estimates can then be used in a Pseudo-marginal MCMC scheme to compute expectations with respect to the posterior.

Finally, methodology is developed to enable unbiased estimation for models in which data can be generated but no tractable likelihood is available. The main motivation for this is stochastic kinetic models used to describe complex and heterogeneous biological systems, but models of this type can be found across the sciences. Approximate Bayesian Computation is used to define a sequence of consistent Monte Carlo estimates, and these are then combined to produce an estimator which is unbiased with respect to the true posterior. Both approaches are demonstrated on a range of examples followed by a critical assessment of their strengths and weaknesses.

Thesis overview

The advent of experimental techniques such as genomic sequencing, proteomics and metabolomics, has meant that the molecular building blocks of biological organisms have more or less been catalogued. Current work in Biology therefore focuses on studying the interactions between these molecules. Statistical and computational modelling of molecular interactions is what sets the Systems Biology approach apart from more traditional methods in Biology. The aim is to describe the salient features of a biological system in order to simulate its behaviour and make testable predictions. Statistical inference is a very important part of this process as it enables the fitting of models to noisy and incomplete data, and the ranking of competing models. The Bayesian paradigm is particularly useful in this respect as uncertainty is consistently propagated through the modelling process.

Whilst the implementation of Bayesian methodology may be desirable, it is often difficult to apply for computational reasons. Analytic solutions are rarely available for the complex models which describe biological systems. In this work, statistical methodology is developed to enable Bayesian inference for a set of models which are used in Systems Biology and beyond. The motivation for the methodology development comes from analysis of data relating to a rare genetic disorder called ARC syndrome. ARC syndrome is caused by mutations in one of two genes, *VPS33B* or *VIPAS39*, and little is known of their function. Loss of these proteins, however, results in a severe multisystem disorder, and generally in loss of life by the age of one year. These proteins must therefore play an important role in cellular function across a range of tissues.

Chapter 1 provides an introduction to the thesis. It begins by describing ARC syndrome, both its clinical presentation and what is currently known of its pathogenesis from cellular biology. It then goes on to introduce the type of data to be analysed, as well as the role of stochastic modelling and Bayesian inference in a systems

Biology approach. A review of the computational statistics required for the later methodology chapters is also included.

In Chapter 2 transcriptomic and metabolomic data from cell lines knocked-down for *VPS33B* or *VIPAS39* is analysed. The aims of the analysis are to assess the similarity of the different cell lines and to identify differentially expressed genes, pathways and networks of interactors giving insight into the development of ARC syndrome. The work corroborates conclusions drawn from experimental work as well as suggesting avenues for future lab work.

During the course of the transcriptomic data analysis, it became clear that it was desirable to incorporate knowledge about the topology of protein interactions when identifying differentially expressed genes. One way to do this is using an undirected graphical model to describe the dependencies between the differentially expressed genes. However, Bayesian posterior inference for these types of models is extremely difficult due to the presence of an intractable normalising term in the likelihood. This situation, and its associated difficulties, are commonly found in many areas of Statistics in which dependent data is modelled, such as spatial statistics and image analysis. These types of distributions have been termed 'doubly-intractable'. In Chapter 3, current inference methods are reviewed and then a novel Markov chain Monte Carlo (MCMC) approach for Bayesian inference is developed and presented.

A common ingredient in a systems analysis of a biological process is to define a stochastic model describing the process and then fit the model to available noisy observations. For many models, it is not possible to write down or compute a likelihood, but it is possible to generate data according to the model. This makes it very difficult to apply standard procedures from either frequentist or Bayesian statistics. Methods such as Approximate Bayesian Computation (ABC), which require only simulation at the expense of introducing some bias, have therefore had a surge in popularity as modelling complex biological systems has become common place. However, the bias introduced is not well characterised. In Chapter 4, literature on ABC methods and unbiased estimation is reviewed and then methodology is developed to allow unbiased estimation of functions or parameters with respect to the posterior distribution.

Finally, Chapter 5 discusses the contributions of the thesis and suggests areas for future work.

Contributions of this thesis

The contributions of this thesis are threefold:

1. Transcriptomic and metabolomic data from three knock-down cell lines is analysed to gain insight into the pathogenesis of ARC syndrome. A mixture of univariate, multivariate and network based statistical methods are used. The similarity of the three cell lines is compared and differentially expressed genes and pathways are identified. The findings are then discussed in detail with reference to recent literature and independent experimental observations from the lab. Suggestions for future experimental work are also made. This work forms part of a submitted paper written in collaboration with other members of the Gissen lab: 'Regulation of post-Golgi PLOD3 trafficking is essential for collagen maturation' (Banushi et al., 2015).

2. Motivated by the idea of using dependencies between proteins in the identification of differentially expressed genes, statistical methodology is developed to enable Bayesian inference for probabilistic models with an intractable normalising term. It is shown that these 'doubly-intractable' distributions are commonly encountered across a range of disciplines. The methodology is based on Pseudo-marginal MCMC, in which only an unbiased estimate of the target posterior is required at each iteration. A method for generating such unbiased estimates is developed and an approach allowing negative estimates to be used is described. The methodology is tested on several examples including undirected graphical models. This work has been accepted for publication in Statistical Science as 'On Russian Roulette Estimates for Bayesian inference with Doubly-Intractable Likelihoods' (Lyne et al., 2015).

3. Methodology is developed to enable unbiased posterior estimation for models with fully intractable likelihoods. This situation is extremely common in

Systems Biology where complex molecular interactions are often described by stochastic kinetic equations. These models can be simulated from, but likelihoods cannot be computed. Currently, Approximate Bayesian Computation (ABC) methods are often used for inference in this case, but the introduced bias is not well characterised. In the newly developed methodology, a simple series is constructed which enables a sequence of Monte Carlo ABC estimates to be combined to form an unbiased estimator. This can then either be used in a Pseudo-marginal MCMC scheme or directly to produce unbiased estimates of posterior expectations. This is the first time unbiased estimation has been developed in the likelihood-free context without the need to introduce further assumptions.

Acknowledgements

I would first like to thank my supervisors, Professors Mark Girolami and Paul Gissen. They have provided support and guidance throughout the course of this PhD, and it could not have been completed without them. I would also like to express my gratitude to UCL Systems Biology for funding this research.

I am sincerely grateful to everyone at the UCL Department of Statistical Science, for creating a stimulating environment in which to work and for being such good company over the last three years. I'll miss our lunchtimes and Friday nights at ULU. I am also grateful to the members of the Gissen Lab for providing me with experimental results, for always being available to answer my questions and for generally being a great group to collaborate with.

I am indebted to my parents, brothers and extended family for always encouraging me in whatever I do.

And finally, I am eternally grateful to Christophe for his support, good humour and composure in the face of a crisis, as well as for proofreading this thesis.

Contents

1 Introduction **25**

1.1 ARC syndrome . . . 25
1.2 Experimental models of ARC syndrome . . . 27
1.3 Transcriptomics: Affymetrix microarrays . . . 28
1.4 Metabolomics . . . 31
1.5 Stochastic modelling in Systems Biology . . . 32
1.6 Bayesian inference . . . 34
1.7 Monte Carlo methods . . . 35
1.7.1 Monte Carlo integration . . . 35
1.7.2 Importance sampling . . . 36
1.7.3 Markov chain Monte Carlo . . . 37
1.7.4 The Metropolis-Hastings algorithm . . . 38
1.7.5 The Gibbs sampler . . . 39
1.7.6 Annealed Importance sampling and Sequential Monte Carlo 40
1.7.7 Perfect sampling . . . 42
1.8 Approximate Bayesian Computation . . . 44
1.9 Conclusion . . . 45

2 Multivariate analysis of transcriptomic and metabolomic data **47**

2.1 Aims of analysis . . . 47
2.2 Initial data processing . . . 49
2.2.1 Microarrays: initial processing . . . 49
2.2.2 Metabolomics: initial processing . . . 51
2.3 Extent of similarity between knock-downs . . . 51
2.4 Differentially expressed genes/metabolites . . . 55
2.5 Principal component analysis . . . 60
2.5.1 PCA results . . . 61
2.6 Partial Least Squares . . . 62

2.6.1 PLS results . . . 63
2.7 Functional analysis . . . 67
2.8 Network analysis . . . 69
2.8.1 Network: results . . . 72
2.9 Discussion . . . 74
2.10 Motivation for the next two chapters . . . 79
2.10.1 Methodology for Doubly-intractable distributions . . . 79
2.10.2 Methodology for unbiased ABC . . . 80

3 Roulette for Doubly-intractable distributions 83
3.1 Introduction . . . 83
3.2 Inference methods for doubly-intractable distributions . . . 85
3.2.1 Approximate Bayesian inference . . . 85
3.2.2 Exact MCMC methods . . . 88
3.2.3 Valid Metropolis-Hastings-type transition kernels . . . 89
3.3 An alternative approach using Pseudo-marginal MCMC . . . 91
3.3.1 Proposed methodology . . . 93
3.3.2 The Sign Problem . . . 94
3.4 Pseudo-marginal MCMC for doubly-intractable distributions . . . 96
3.4.1 Geometric Series Estimator . . . 98
3.4.2 Unbiased estimators using an exponential auxiliary variable 100
3.4.3 Possible choices for $\hat{\mathcal{Z}}_i(\theta)$ and $\tilde{\mathcal{Z}}(\theta)$ . . . 101
3.5 Unbiased Truncation of Infinite Sums: Russian Roulette . . . 103
3.5.1 Single Term Weighted Truncation . . . 103
3.5.2 Russian Roulette . . . 104
3.5.3 Comparison with current algorithms . . . 107
3.5.4 Computational Complexity . . . 107
3.6 Experimental Evaluation . . . 109
3.6.1 Simple example . . . 109
3.6.2 Ising Lattice Spin Models . . . 112
3.6.3 The Fisher-Bingham Distribution on a Sphere . . . 115
3.7 Discussion and Conclusion . . . 120

4 Unbiased posterior estimation using ABC 123
4.1 Introduction . . . 123
4.2 Stochastic models in Systems Biology . . . 124
4.3 Approximate Bayesian Computation . . . 126

4.4 Unbiased estimation using biased estimates 129
4.5 Unbiased estimation using ABC estimates 131
4.5.1 Unbiased rejection ABC 131
4.5.2 Pseudo-marginal MCMC ABC 134
4.5.3 Designing truncation distributions 136
4.5.4 Debiasing with low dimensional sufficient statistics 137
4.6 Experimental validation 137
4.6.1 Toy example 137
4.6.2 Simple molecular system 140
4.7 Discussion and Conclusions 143

5 General Conclusions **147**

Appendices **150**

A Top 100 PCA loadings for transcriptomic data for Principal Components 1 and 2 **151**

B Enriched DAVID annotations for transcriptomic data **155**

C Russian Roulette **157**

Bibliography **161**

List of Figures

1.1 The trafficking of a molecule in yeast which is eventually degraded. CORVET functions via Rab5 to fuse two early endosomes. HOPS then promotes fusion of the late endosome with the vacuole via Rab7. Reprinted with permission of The Company of Biologists, Journal of Cell Science, kleine Balderhaar and Ungermann (2013). . 26

1.2 Mouse IMCD cells grown in 3D culture to form spheres with a central lumen. The image on the left shows control cells forming a cohesive epithelial layer, in the three knock-down cell lines junction and polarisation are disrupted. Picture provided by A. Straatman-Iwanowska (Gissen Lab). 28

1.3 Processes involved in measuring mRNA abundance using an Affymetrix microarray. RNA from the original sample is reverse-transcribed to give cDNA and then re-transcribed with biotinylated nucleotides. This mixture is then hybridised with the array and scanned giving an image file which requires processing to produce values indicating relative mRNA abundance. Reprinted by permission from Macmillan Publishers Ltd, Nature Reviews Genetics, The Tumor Analysis Best Practices Working Group (2004). 30

1.4 Processes involved in measuring metabolite abundance using GC-MS. The sample is first vapourised and then transferred to a chromatographic column. Compounds are separated in the column via their interaction with the column walls. Samples leaving the column are ionised and passed through a magnetic field which separates the the ionised compounds based on their charge-to-mass ratio. GC-MS schematic was created by K. Murray and shared under CC BY-SA 3.0. 32

1.5 The cycle of Systems Biology research. From Kitano (2002b), reprinted with permission from AAAS. 33

2.1 Flow chart showing initial analysis carried out on transcriptomic data from knock-down IMCD cell lines. 48
2.2 Top row shows log-intensity histograms and bottom row shows box plots of log-intensity. Figures on the left are before RMA processing while figures on the right are after. The data is from knock-down IMCD cell lines. 51
2.3 The three boxplots show the metabolomic data for each sample (a) log transformed, unnormalised (b) log-transformed, normalised to an internal standard (c) log transformed, normalised to median of each sample. 52
2.4 Hierarchical clustering of (a) transcriptomic and (b) metabolomic samples based on Euclidean distance. 52
2.5 Heatmaps showing Spearman's rank correlation between (a) transcriptomic and (b) metabolomic samples. 53
2.6 PCA score plots for (a) transcriptomic data and (b) metabolomic data. The original high-dimensional data is projected onto two axes which explain variation in the data. 54
2.7 Venn diagrams showing the overlap in differentially expressed genes ((a) and (b)) and metabolites (c). Differentially expressed genes were identified using the limma package in R and differentially abundant metabolites using a standard t-test. In both cases p-values were Benjamini-Hochberg corrected for multiple testing and a cut-off of FDR<0.05 applied. In (a) and FDR cut-off of <0.05 was used. In (b) the top 100 genes ranked by p-value were used. . . 56
2.8 Heat maps showing RNA transcripts which are significantly differentially expressed based on comparing all controls against VIPAR and VPS33B knock-downs. Expression levels in PLOD3 are shown although this data was not used in the analysis. 58
2.9 Metabolites with significantly different levels in either the VIPAR or VPS33B knock-down cell lines. Green indicates up compared to control, red down compared to control. The third column shows metabolites up or down in mock-transfected compared to wild type cells. 59
2.10 Heat maps showing metabolites which are significantly differentially expressed based on comparing (a) all controls against Vipar and Vps33b knock-downs (b) Mock transfected controls against wild type cells 60

2.11 Metabolomic PCA loading plot 62

2.12 Root mean square error in class prediction using PLS with varying numbers of components for (a) transcriptomic data (b) metabolomic data. 64

2.13 Scores from the first two components of a PLS model for (a) transcriptomic data (b) metabolomic data. 65

2.14 Loadings from the first two components of a PLS model for metabolomic data. 66

2.15 (a) Root mean square error of prediction in the full PLS model of metabolomic data predicted by transcriptome data. The median has been taken across the prediction of all metabolite abundances. (b) Scores for the first two components in the full PLS model of metabolomic data predicted by transcriptome data. 67

2.16 Overview of the method developed by Banerji et al. (2015) to detect genes whose interaction profiles have changed significantly between a healthy and diseased state. Figure created by C. Banerji and shared under CC BY 4.0. 71

2.17 Schematic of tight junctions, adherent junctions and desmosomes which are the three main junction complexes connecting adjacent epithelial cells. Reprinted by permission from Macmillan Publishers Ltd, Nature Reviews Gastroenterology and Hepatology, Neunlist et al. (2013). 75

3.1 MCMC traces for 2000 samples using three different methods to sample from the posterior of the example from Murray et al. (2006). 110

3.2 Histogram showing number of unbiased estimates required at each MCMC iteration when using Roulette method on example in Murray et al. (2006). 111

3.3 Autocorrelation function up to lag 20 for three different methods for samples from example in Murray et al. (2006). 111

3.4 Running mean of 10,000 MCMC samples from posterior from example in Murray et al. (2006) using three different methods. 111

3.5 Traces of samples using the debiasing infinite series with (a) Russian Roulette, (b) Poisson truncation, and (c) the approximate Exchange algorithm (d) the Exchange algorithm using perfect samples and (e) an MCMC chain with the partition function calculated using the matrix transfer method. Note in (a) and (b) the samples are not drawn from the posterior distribution, $p(\beta|\mathbf{y})$, but from the (normalised) absolute value of the estimated density. 114

3.6 Autocorrelation plots for samples drawn from the posterior distribution $p(\beta|\mathbf{y})$ of a 10×10 Ising model using five methods: (a) Debiasing series with roulette truncation, (b) Debiasing series with Poisson truncation (c) approximate Exchange (d) the Exchange algorithm using perfect samples and (e) an MCMC chain with the partition function calculated using the matrix transfer method. . . . 116

3.7 Plots of the running mean for the posterior distribution $p(\beta|\mathbf{y})$ of a 10×10 Ising model using three methods: (a) Debiasing series with roulette truncation, (b) Debiasing series with Poisson truncation (c) approximate Exchange (d) the Exchange algorithm using perfect samples and (e) an MCMC chain with the partition function calculated using the matrix transfer method. 117

3.8 Plots of the running standard deviation for the posterior distribution $p(\beta|\mathbf{y})$ of a 10×10 Ising model using three methods: (a) Debiasing series with roulette truncation, (b) Debiasing series with Poisson truncation (c) approximate Exchange (d) the Exchange algorithm using perfect samples and (e) an MCMC chain with the partition function calculated using the matrix transfer method. 118

3.9 Sample traces and autocorrelation plots for the Fisher-Bingham distribution for the geometric tilting with Russian Roulette truncation ((a) and (b)) and Walker's auxiliary variable method ((c) and (d)).) . 119

4.1 Realisations of (a) continuous deterministic and (b) discrete stochastic processes for a simple model in which a single protein is produced and degraded with rates α and μ respectively. Reproduced by permission from Macmillan Publishers Ltd, Nature Reviews Genetics, Wilkinson (2009). 125

4.2 Running estimates of (a) the mean and (c) standard deviation of the toy example posterior with $N = 1$ using unbiased rejection ABC. Error estimates are based on the distribution of sample means and standard deviations. (b) and (d) give histograms of the 139
4.3 Running estimates of (a) the mean and (b) the standard deviation of the toy example posterior with $N = 1$ using pseudo-marginal MCMC ABC. Error estimates are based on the distribution of sample means and standard deviations using an estimate of the effective sample size as the number of samples. (c) shows the trace and (d) shows the autocorrelation between MCMC samples. 140
4.4 Running estimates of the mean of the toy example posterior with $N = 10$ using independent debiased estimates (top row) and pseudo-marginal MCMC ABC (bottom row). Error estimates are based on the distribution of sample mean using either the number of samples for (a) or an estimate of the effective sample size as the number of samples for (c). (b) shows a histogram of the debiased estimates and (d) shows the MCMC trace. 141
4.5 Data simulated for the simple dimerisation model using the Gillespie algorithm. Eight data points were used for species A and AB, shown as the large circles. 142
4.6 (a) and (b) Running mean for parameters k_1 and k_2 for stochastic dimerisation model. Note that the red lines denote the values used to simulate the data, not the true mean of the posterior. (c) MCMC trace for the Pseudo-marginal samples. (d) Autocorrelation for MCMC samples. 143

B.1 Full list of enriched Gene Ontology annotations for 150 genes with largest loadings in the full PLS model using online software DAVID. 155

List of Tables

2.1 Table showing the top 20 most down-regulated transcripts, as detected by modified t-test, and some detail on what is known of their function. Wild type and mock-transfected cells were treated as controls and VPS33B and VIPAR knock-down cell lines as one class. . 57

2.2 Table showing enriched GO annotations for lists of significantly differentially expressed genes for knock-down cell lines using the 150 genes with the largest loadings in the first component of a full PLS model. 68

2.3 Table showing enriched pathways for knock-down cell lines (Vps33b and Vipar knock-downs analysed together). Gene sets were taken from the Kegg, Biocarta and Reactome online databases. 70

2.4 Top 20 genes and their functions as identified by Kullback-Leibler divergence between their interaction distribution in controls and knock-downs. 82

3.1 Table showing the exact posterior mean and variance of the parameter θ from the simple example in Murray et al. (2006), as well as the estimates from the three sampling methods: Roulette, Exchange and standard Metropolis. 110

3.2 Monte Carlo estimates of the mean and standard deviation of the posterior distribution $p(\beta|\mathbf{y})$ using the five algorithms described. The debiasing series estimates have been corrected for negative estimates. The exact chain was run for 100,000 iterations and then the second half of samples used to achieve a 'gold standard' estimate. An estimate of the effective sample size (ESS) is also shown based on 10,000 MCMC samples. 115

3.3 Estimates of the posterior mean and standard deviation of the posterior distribution using roulette and Walker's method for the Fisher-Bingham distribution. An estimate of the effective sample size (ESS) is also shown based on 10,000 MCMC samples. 119

Chapter 1

Introduction

1.1 ARC syndrome

Arthrogryposis, renal dysfunction and cholestasis (ARC) syndrome is an autosomal recessive genetic disease. Its name describes three of the main diagnostic features: joint contractures, kidney tubular dysfunction and disruption of bile flow from the liver to the duodenum. ARC is a multisystem disorder with defects of the musculoskeletal system, kidneys, liver, skin and platelets. The syndrome is rare but severe; children affected by the syndrome suffer from a severe failure to thrive and most die in the first year of life.

Approximately 75% of cases are caused by mutations in *VPS33B* which encodes the protein Vacuolar Protein Sorting 33 Homolog B (VPS33B). The other $\sim$25% of cases are caused by mutations in *VIPAS39* which encodes VPS33B Interacting Protein, Apical Basolateral Polarity Regulator (VIPAR) (Gissen et al., 2004; Smith et al., 2012; Gissen et al., 2006; Cullinane et al., 2010). From the severity and multisystem nature of ARC syndrome, it is clear that the proteins are involved in a vital cellular process across many tissues. Further, it is highly likely that they function together as loss of either leads to the same disease symptoms. The precise functions of VPS33B and VIPAR are unknown, but information gleaned via homology as well as experimental results from the Gissen lab and elsewhere, has elucidated several interactors and roles for the two proteins.

VPS33B and VIPAR are homologous to yeast Vps33p and Vps16p which are constituents of the HOmotypic fusion and vacuole Protein Sorting (HOPS) and class C CORe Vacuole/Endosome Tethering (CORVET) protein complexes crucial for vac-

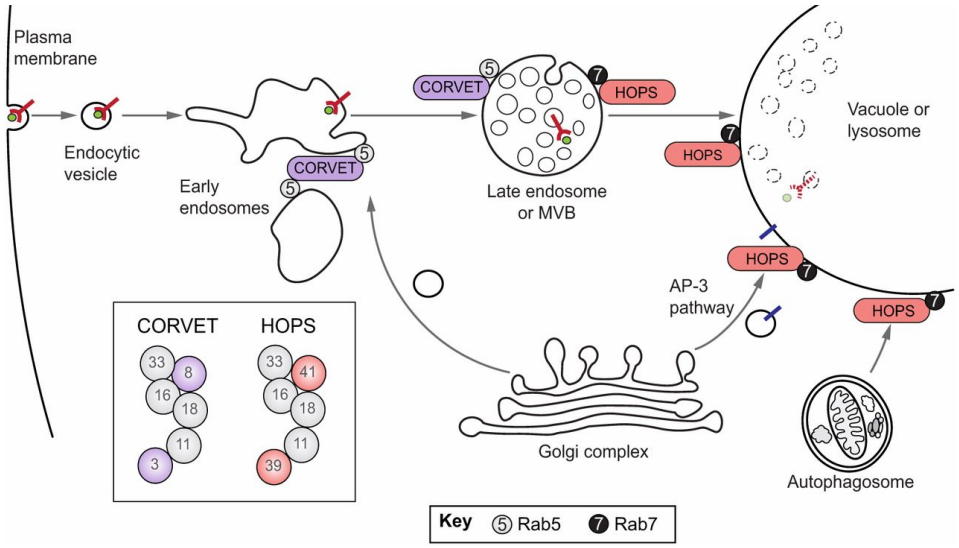


Figure 1.1: The trafficking of a molecule in yeast which is eventually degraded. CORVET functions via Rab5 to fuse two early endosomes. HOPS then promotes fusion of the late endosome with the vacuole via Rab7. Reprinted with permission of The Company of Biologists, Journal of Cell Science, kleine Balderhaar and Ungermann (2013).

uolar biogenesis (kleine Balderhaar and Ungermann, 2013). This is a process which involves mechanisms such as the sorting of vacuolar proteins away from secretory pathways, endocytosis of material from the plasma membrane and transport pathways which deliver proteins to the vacuole (Bryant and Stevens, 1998). The HOPS and CORVET complexes carry out multiple roles including tethering membranes and interacting with RAB GTPases which regulate trafficking (Solinger and Spang, 2013).

The HOPS and CORVET complexes have mammalian equivalents which are involved in vesicular trafficking in endocytosis and autophagy (Pols et al., 2013; Kim et al., 2010; Huizing et al., 2001; McEwan et al., 2015). Metazoans have two homologues of Vps33p (VPS33A and VPS33B) and Vps16p (VPS16 and VIPAR). Some studies have proposed that both sets of homologues participate in HOPS and CORVET function in multicellular organisms (Tornieri et al., 2013; Zhu et al., 2009). However, experimental evidence from studies in mammalian cells has identified VPS33A and VPS16, and not VPS33B and VIPAR, as members of the conventional mammalian HOPS and CORVET complexes (Baker et al., 2013; Wartosch et al., 2015; Graham et al., 2013).

Previous work from the Gissen lab has suggested that VPS33B and VIPAR form a complex that functions as a Rab11a effector due to its specific interaction with the

active form of Rab11a (Cullinane et al., 2010). Rab11a is a small GTPase known to regulate the recycling of internalised cargo back to the cell membrane, and to participate in epithelial cell polarity. RNAi knockdown (kd) of Vps33b and Vipas39 in mouse Inner Medullary Collecting Duct (mIMCD-3) cells leads to downregulated expression of apical junction complex proteins and loss of polarity (Cullinane et al., 2010). This links the VPS33B-VIPAR complex to diseases such as cancer, in which the disruption of cell-cell junctions and polarity leads to epithelial-to-mesenchymal transition (EMT) and metastasis.

Recent work from the lab has indicated a role for VIPAR/VPS33B in a trafficking pathway which transports Procollagen-Lysine, 2-Oxoglutarate 5-Dioxygenase type 3 enzyme (PLOD3) from the TGN to its functional sites in procollagen carriers. Collagens are the most abundant family of proteins in the human body and the major fibrous protein in the extracellular matrix. Lysine hydroxylation of procollagen is a post-translational modification carried out by three PLOD enzymes (PLOD1-3). The specific collagen modifications catalyzed by PLOD3 are critical for the process of fibre crosslinking, which stabilises the procollagen fibril into the overall collagen structure (Knott and Bailey, 1998). Clinical features of the single patient described with PLOD3 deficiency overlap with those of ARC syndrome, supporting the idea that VPS33B-VIPAR interacts with PLOD3, and include severe growth retardation, hypotonia (abnormally increased muscle tone), arthrogryposis, low bone mineral density and bone fractures (Salo et al., 2008).

1.2 Experimental models of ARC syndrome

Experimental tools have been developed in the Gissen Lab to allow investigation into the functions of VPS33B and VIPAR, and how mutating them might cause the ARC syndrome phenotype. These experimental models can be used to investigate, for example, the localisation of proteins, possible protein interactors and the morphological impacts of low levels of VPS33B/VIPAR. They can also be analysed via transcriptomic, proteomic or metabolomic experiments to assess how the global expression of various molecules is affected by the mutations.

Mouse Inner-Medullary Collecting Duct (mIMCD-3) cells have been stably knocked down using silencing short hairpin (sh-) RNA for *VPS33B*, *VIPAS39* and *PLOD3*. mIMCD-3 cells are a polarised epithelial cell line which form tubules and tight junctions and are therefore a good context in which to study the impact

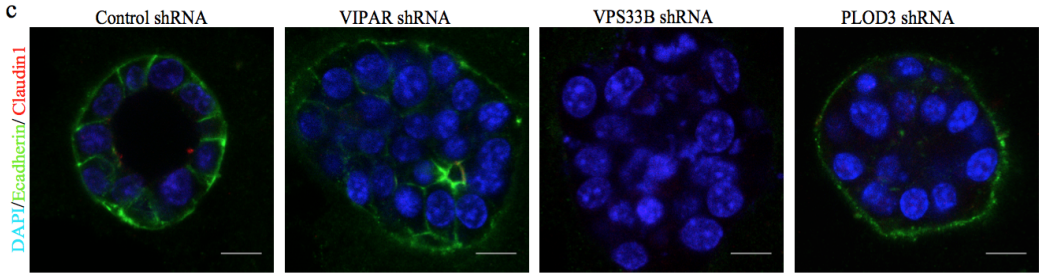


Figure 1.2: Mouse IMCD cells grown in 3D culture to form spheres with a central lumen. The image on the left shows control cells forming a cohesive epithelial layer, in the three knock-down cell lines junction and polarisation are disrupted. Picture provided by A. Straatman-Iwanowska (Gissen Lab).

of VPS33B and VIPAR mutations on polarisation and epithelial integrity. The phenotype observed in knock-down cells can be clearly seen in Figure 1.2. Control cells form a central lumen when grown in 3D culture with correctly polarised cells, but the three knock-downs suffer from disrupted cell junctions and do not form a cohesive cell layer. This corroborates previous work suggesting that VPS33B, VIPAR and PLOD3 are involved in cell polarisation and adhesion.

In the first chapter of this thesis, transcriptomic and metabolomic data from mIMCD cells knocked-down for *VPS33B*, *VIPAS39* and *PLOD3* are analysed. The first aim is to analyse how similar the transcriptional and metabolite response is in the three different cell lines. Phenotypic observations suggest that the three proteins have some overlapping function and this analysis could further validate or dispel this hypothesis. The second aim is to identify genes and metabolites which have significantly different expression when VPS33B, VIPAR and PLOD3 have reduced levels. By identifying these genes and their functions, it is hoped that processes relevant to the pathogenesis of ARC syndrome will be discovered. The final aim is to determine pathways and networks of interactors whose functions are substantially perturbed in the knock-down cell lines and discuss in reference to recent literature how these might be involved in the pathogenesis of ARC syndrome.

1.3 Transcriptomics: Affymetrix microarrays

Messenger RNA (mRNA) is a polymeric molecule whose main role is to carry genetic information from chromosomal DNA to cellular sites where proteins are synthesised. Within a given organism, all cells contain the same sequence information in their DNA, and yet distinct cell types exist with different morphologies and func-

tions. One of the major differences between cell types is differential expression of mRNA; certain genes will be more highly expressed in one cell type compared to another. We can think of each cell type as having a characteristic gene expression profile which allows that cell type to carry out its specific function. The ability to measure this expression profile is incredibly useful to gain insight into the genetic interactions within cells as well as for diagnostic/investigative medicine. It has revolutionised the way molecular biology proceeds; genes which are differentially expressed between two experimental conditions can immediately be identified with no bias introduced by the state of current knowledge.

Microarrays are a high-throughput assay which enable researchers to measure expression levels of a large number of RNA molecules within one sample simultaneously. The following description is based on Affymetrix arrays as these were the ones used in our experiments and the specific type of array is relevant to how the proceeding analysis is carried out. The concept is simple (see Figure 1): a chip is designed such that it contains 'probes' (short oligonucleotide sequences) complimentary to regions on the ('target') mRNA molecules in a given organism. Before the sample can be applied to the chip, double-stranded cDNA is synthesised from the RNA in the sample. This is then transcribed using biotin-labelled ribonucleotides to produce antisense mRNA molecules. The chip and the labelled mRNA are hybridised and then stained with a fluorescent molecule which binds to biotin. The chip is then scanned producing an image in which the fluorescence of each spot is related to the amount of biotin and hence to the amount of that particular mRNA molecule. Each chip contains multiple probes for each target mRNA which are combined to obtain a more precise measurement of the expression level.

Microarray chips have been extremely useful in identifying functionally relevant genes across a range of processes and diseases, as they provide a more complete picture of gene transcription than more traditional low-throughput methods. However there are several problems with the technique which limit its accuracy, for example:

- Different probes for the same gene will have different (unknown) binding affinities.
- Probes frequently 'cross-hybridise' to the wrong target.
- There is a lack of consensus on how to convert fluorescence measurements

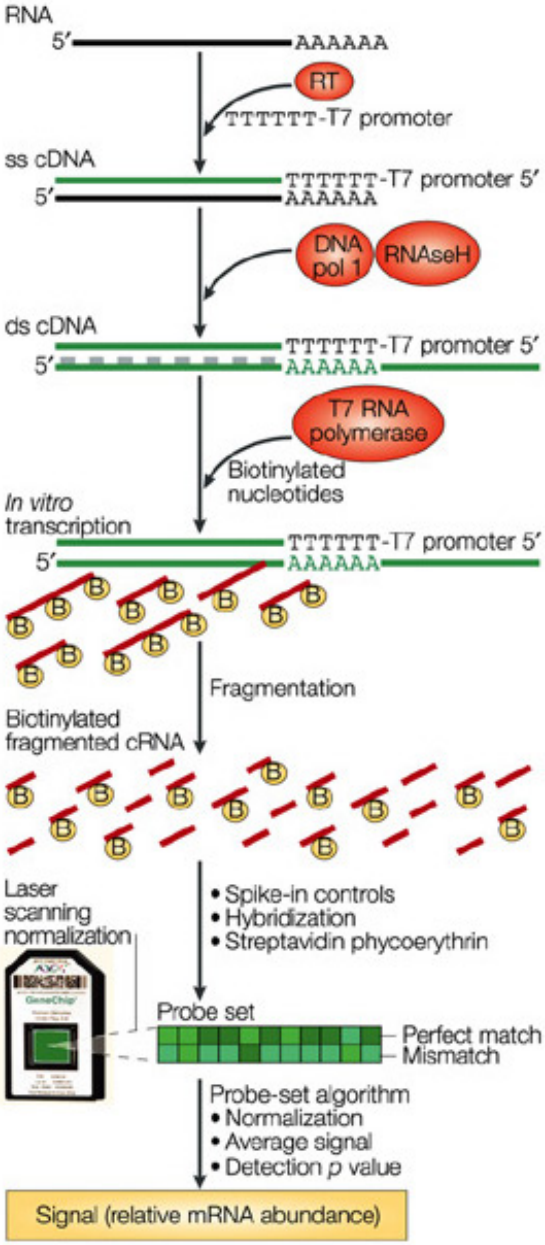


Figure 1.3: Processes involved in measuring mRNA abundance using an Affymetrix microarray. RNA from the original sample is reverse-transcribed to give cDNA and then re-transcribed with biotinylated nucleotides. This mixture is then hybridised with the array and scanned giving an image file which requires processing to produce values indicating relative mRNA abundance. Reprinted by permission from Macmillan Publishers Ltd, Nature Reviews Genetics, The Tumor Analysis Best Practices Working Group (2004).

into mRNA expression values.

A more holistic point worth noting, is that microarrays provide information about mRNA levels in the sample, but this does not necessarily translate directly to information about protein levels, as transcriptional regulation is not the only way of controlling protein concentrations in a cell. There is debate over the level of correlation between the proteome and the transcriptome, and hence how informative transcript levels are about protein expression, for example Ghazalpour et al. (2011) and Nagaraj and Wisniewski (2011) find $r = 0.27$ and $r = 0.6$ respectively. However, results in the first paper suggest that transcript levels may correlate more strongly with clinical traits than protein levels (at least with currently available experimental techniques) and that therefore transcriptomic analysis does yield biologically relevant information.

1.4 Metabolomics

Metabolites are small molecules which are intermediates or products of the chemical reactions in living organisms. Metabolomics is an experimental and analytical field in which the concentrations of many metabolites are measured simultaneously in a given sample, based on the idea that these concentrations can reflect changes in disease state or cellular function and provide insight into disease pathogenesis. Small molecules are extracted from the sample, and then separated and quantified. The techniques involved include liquid and gas chromatography (LC and GC) for separation and nuclear magnetic resonance (NMR) spectroscopy and mass spectroscopy (MS) for quantification. Combinations of these techniques are often used, meaning that extensive pre-processing of the data is required before expression levels are obtained.

The metabolomic data obtained from the mIMCD cell lines was produced using GC-MS and this technique is therefore described in more detail. In the gas chromatograph, the sample is first vapourised before travelling up a capillary column. Differences in chemical properties of the various molecules lead to them separating as they travel, and they can then be eluted separately. The mass spectrometer is then used to identify the amount and type of molecules present in the sample. The molecules are converted into charged fragments which are then detected based on their mass-to-charge ratio.

MS platforms can be used to characterise, identify and quantify a large num-

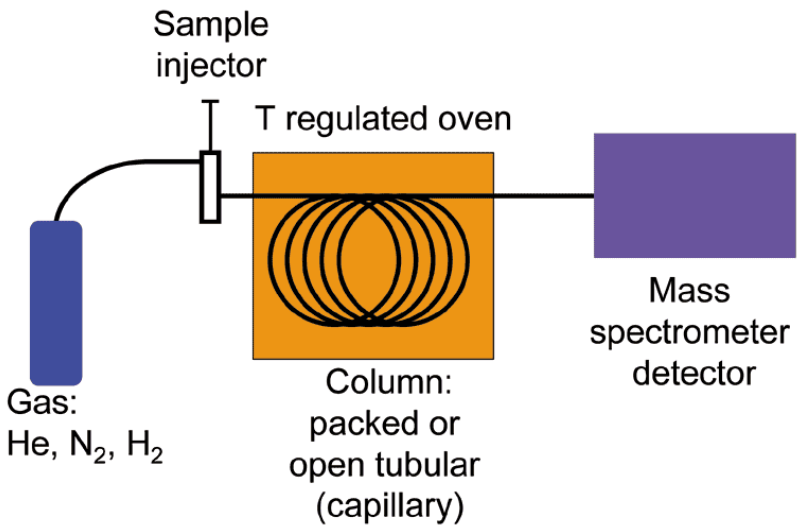


Figure 1.4: Processes involved in measuring metabolite abundance using GC-MS. The sample is first vapourised and then transferred to a chromatographic column. Compounds are separated in the column via their interaction with the column walls. Samples leaving the column are ionised and passed through a magnetic field which separates the the ionised compounds based on their charge-to-mass ratio. GC-MS schematic was created by K. Murray and shared under CC BY-SA 3.0.

ber of metabolites where the concentrations might cover a large range of values (Kaddurah-Daouk et al., 2008). In particular they can be used to measure low concentration molecules such as signalling molecules and to measure the abundance of molecules which are not yet identified. GC-MS provides structural information, reasonable quantitative precision and high-throughput data, however it cannot be used to study molecules which cannot be readily vapourised.

1.5 Stochastic modelling in Systems Biology

Mathematical modelling has a long history in Biology. Topics have included the modelling of predator-prey populations (Lotka, 1925; Volterra, 1927), quantifying the effect of cow-pox inoculation on the spread of smallpox (Bernoulli, 1760) and excitation and conduction in neurons (Hodgkin and Huxley, 1952). However, the advent of high-throughput technologies enabling the simultaneous quantification of many molecules, has a led to a massive increase in the number of researchers working in the field of mathematical and computational biology. It is no longer possible to follow up experimentally on all findings of interest, and a computational model can be a cheaper and more efficient way to perturb the system of interest. Two of the key aims are to simulate the process so as to make testable predictions and to inves-

tigate the interactions between multiple components of one organism. In addition to these practical considerations, has come the awareness that Biology cannot be understood simply by drawing diagrams of interactions between molecules (Kitano, 2002a). An understanding of its dynamics is key to understanding its complexity.

Kitano (2002a) describes a cycle of Systems Biology research (see Figure 1.5) in which models are created, simulated to make predictions and then tested experimentally. Some models will be eliminated and those that are consistent with evidence can be further developed and analysed to improve understanding.

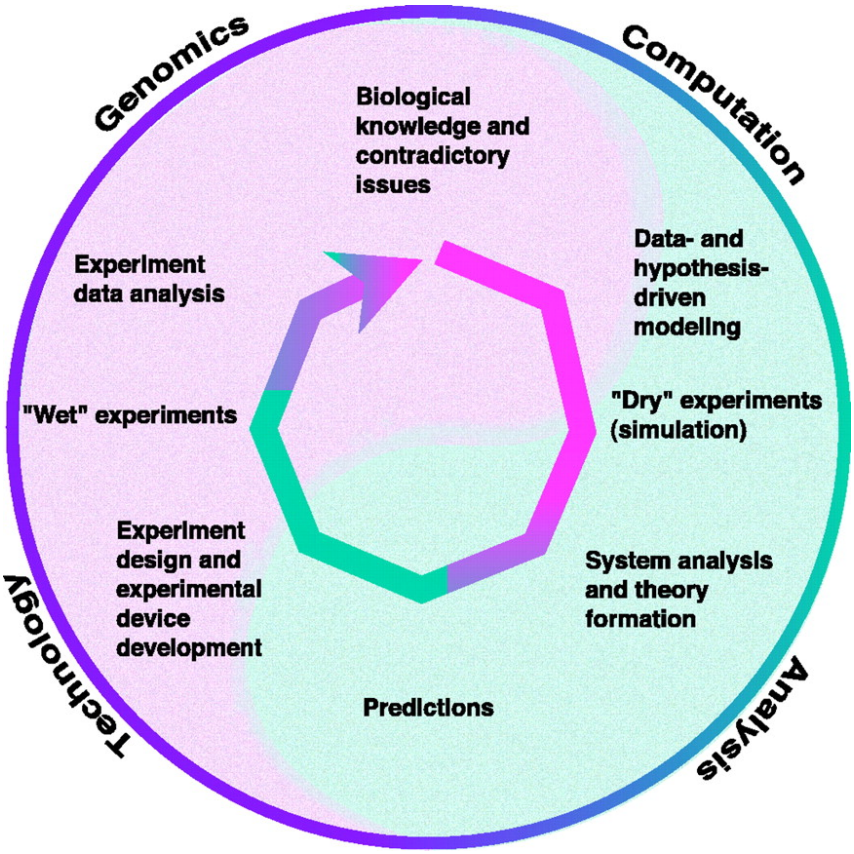


Figure 1.5: The cycle of Systems Biology research. From Kitano (2002b), reprinted with permission from AAAS.

The models in use have become more complex to incorporate the vast wealth of knowledge collated on biological systems, but it is not possible to build models of molecular interactions of a scale to be useful. Hence all models of biological systems are approximations which leave out many details and include only the salient features of interest. Further, experimental conditions can never be completely controlled in biological systems, for example hormonal oscillations, respiration and

individual genetic variations. In this modelling regime, the kinetics of biological processes are inherently stochastic and therefore require stochastic models (Wilkinson, 2011).

Accordingly, a key part of the Systems Biology cycle is statistical inference: fitting and comparing stochastic models to observed data. Bayesian methods provide a powerful framework for analysis due to the fully probabilistic approach for incorporating prior beliefs, statistical models and data. Probability distributions describing full beliefs about model parameters given the observed data can be obtained. For models of real world situations, analytic solutions are rarely available for Bayesian inference and hence computational methods are important. Increases in computational power have combined with methodological developments such as those in Markov chain Monte Carlo theory to allow Bayesian methods to be applied to a wide range of situations, however there are still many models and big data scenarios in which full Bayesian solutions remain elusive. In the next section the details of Bayesian inference are described.

1.6 Bayesian inference

Across the sciences, a key aim is to understand the underlying processes which produce observed data. Statistical modelling and inference are useful tools to aid in this understanding and to enable scientists to make testable predictions. There are various stages involved in inference after collecting some data:

1. The formulation of a statistical model, often several competing models
2. Fitting the models, finding which parameter values best fit the observed data
3. Model selection, choosing which model best describes the underlying process or is most useful for the purpose
4. Testing how well the model fits the data, possibly by collecting more data.

The methodology developed in the later part of this thesis is associated with the Bayesian paradigm in which model parameters are viewed as random variables and probability is used to quantify uncertainty or beliefs about the parameter. Before any data is collected, information about the parameters is included in the *prior* distribution. These beliefs are then updated using information in the likelihood to produce the *posterior* distribution, from which all inferences proceed.

Throughout this thesis, unless stated otherwise, the observed data is denoted by $\mathbf{y}$, and the unknown model parameters are denoted by θ. The posterior distribution, $\pi(\theta|\mathbf{y})$, is proportional to the data likelihood, $p(\mathbf{y}|\theta)$ and the prior $\pi(\theta)$

$$\pi(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)\pi(\theta)}{p(\mathbf{y})}, \tag{1.1}$$

where $p(\mathbf{y}) = \int_\Theta p(\mathbf{y}|\theta)\pi(\theta)\mathrm{d}\theta$. Often we are interested in computing the expectation of some function with respect to the posterior distribution i.e.

$$\mathbb{E}_\pi[\varphi(\theta)] = \int_\Theta \varphi(\theta)\pi(\theta|\mathbf{y})\mathrm{d}\theta$$

where the simplest function, and often the one of interest, is $\varphi(\theta) = \theta$, the parameter values themselves. The denominator of (1.1), $p(\mathbf{y})$, is known as the *marginal likelihood* or the *model evidence* and is generally not easy to compute. Whether or not this matters depends on the problem at hand. For parameter inference, it is generally not required as it is not a function of θ and inference can therefore be based on $\pi(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)\pi(\theta)$. If, on the other hand, it is model comparison that is of interest, then the model evidence is crucial as the posterior odds of two models is computed as

$$\frac{\pi(M_1|\mathbf{y})}{\pi(M_2|\mathbf{y})} = \frac{p(M_1)}{p(M_2)} \times \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}.$$

1.7 Monte Carlo methods

1.7.1 Monte Carlo integration

Monte Carlo methods are a set of simulation tools utilising random numbers to estimate integrals or expectations with respect to a given distribution. They are therefore extremely useful in Bayesian statistics, both for parameter inference and model choice. Assume we wish to estimate an integral of the form

$$I = \int_{\Theta} \varphi(\theta) p(\theta) \mathrm{d}\theta, \tag{1.2}$$

where $p(\theta)$ could equal $p(\theta|\mathbf{y})$. In the simplest application of Monte Carlo, N points are sampled from $p(\cdot)$ and the integral is approximated by

$$\hat{I} = \frac{1}{N} \sum_{n=1}^{N} \varphi(\theta_n) \qquad \theta_n \sim p(\cdot).$$

The estimator is unbiased and its variance converges at the canonical rate of $1/N$ as long as it is bounded appropriately. This rate is independent of dimension, meaning it can outperform quadrature in high dimensions.

1.7.2 Importance sampling

It is not always possible to sample from the density p above and indeed it is not always optimal in terms of the variance of the estimator. Importance sampling is a technique which allows the integral I to be approximated by sampling from a different distribution, referred to as the importance density, and weighting each sample to compensate for the discrepancy between the importance density and p. We can rewrite (1.2) as

$$I = \int_{\Theta} \frac{\varphi(\theta) p(\theta)}{g(\theta)} g(\theta) \mathrm{d}\theta,$$

by introducing the importance density, $g(\cdot)$, which has the same support as $p(\cdot)$.The integral can now be approximated via Monte Carlo with

$$\hat{I}_{\text{imp}} = \frac{1}{N} \sum_{n=1}^{N} \frac{\varphi(\theta_n) p(\theta_n)}{g(\theta_n)} \qquad \theta_n \sim g(\cdot),$$

where $p(\theta_n)/g(\theta_n)$ is the importance weight for the n-th sample. Clearly impor-

tance sampling also gives unbiased estimates of the integral, however, the importance sampling estimate can often have a lower variance if the importance density is appropriately designed, see for example Ripley (2009, Chap. 5) for details of optimal implementation.

1.7.3 Markov chain Monte Carlo

A prerequisite for the use of Monte Carlo integration is the ability to sample from the probability distribution of interest, $p(\cdot)$. In Bayesian statistics the distribution of interest is generally the posterior distribution, but for many combinations of likelihoods and priors it has proven difficult to design methods to sample efficiently from this distribution. Indeed this meant that for a long time Bayesian inference was only implemented in situations with a simple likelihood and a conjugate prior. The development of Markov chain Monte Carlo (MCMC) methods revolutionised the field of Bayesian statistics by enabling samples to be drawn from a much greater range of posterior distributions.

There are of course many other (sometimes preferable) ways to sample from distributions, such as rejection sampling and inverse transform sampling, however as the methodology development in this thesis largely focuses on MCMC methods, we assume for now that these methods are not available.

To describe MCMC, we first define a Markov chain. A stochastic process, $\{X_n\}$, is Markovian if the conditional distribution of X_{n+1} given $X_1, \dots X_n$ depends only on the previous sample, X_n. This conditional distribution, K, is called the transition distribution or kernel and we will assume it is stationary (i.e. does not depend on n). The joint distribution of the Markov chain is completely defined by the initial distribution of X_1 and the iteratively applied transition kernel. The aim in MCMC is to design and simulate a Markov chain which has the target density as its unique stationary density. The invariant or stationary density, π, for transition kernel, $K(x, \mathrm{d}y)$ is defined as

$$\pi(\mathrm{d}y) = \int_{x \in \mathcal{X}} \pi(\mathrm{d}x) K(x, \mathrm{d}y), \tag{1.3}$$

which in words means that if we have a sample from π, and the transition kernel is applied, the marginal distribution of the next state of the chain is also π. Surpris-

ingly, MCMC algorithms which have π as their invariant density can be designed relatively simply (for example using the algorithms described in the next two subsections), however two other conditions must also be satisfied to guarantee that the chain has a unique stationary distribution and is guaranteed to converge to it:

1. *irreducibility*, the ability to reach any x with $\pi(x) > 0$ in a finite number of steps.

2. *aperiodicity*, the chain must not get trapped in cycles.

With these three conditions met (π as the stationary distribution, irreducibility and aperiodicity) then for π-a.e. $x \in \mathcal{X}$

$$\lim_{n\to\infty} ||K^n(x,\cdot) - \pi(\cdot)|| = 0,$$

(see Roberts and Rosenthal (2004) for proof).

The next two subsections outline details of common MCMC algorithms which can be used to construct Markov chains which converge to π.

1.7.4 The Metropolis-Hastings algorithm

It is often easy to construct Markov chains which are *reversible* (also referred to as satisfying *detailed balance*), defined as follows

$$\pi(\mathrm{d}y)K(y,\mathrm{d}x) = \pi(\mathrm{d}x)K(x,\mathrm{d}y) \qquad \text{for all } x,y \in \mathcal{X}. \tag{1.4}$$

This is important because integrating both sides with respect to x returns Equation (1.3), and hence it is a useful condition to derive MCMC algorithms. The Metropolis-Hastings algorithm is the simplest way to make use of this sufficient (but not necessary) condition, and the steps are outlined in Algorithm 1. To implement the algorithm, all that is required is the design of a proposal distribution, $q(x,\cdot)$.

To show that the algorithm produces a Markov chain which satisfies detailed bal-

Algorithm 1 The Metropolis-Hastings algorithm

Initialise with a sample x_0
for $n = 1$ to N **do**
Sample $x' \sim q(x_{n-1}, \cdot)$
Set $x_n = x'$ with probability, α

$$\alpha(x, x') = \min\left[1, \frac{\pi(x')q(x', x)}{\pi(x)q(x, x')}\right]$$

Or with probability $1 - \alpha$ set $x_n = x_{n-1}$
end for

ance consider two situations: (a) $x = x'$, for which detailed balance is satisfied trivially and (b) $x \neq x'$. It suffices to show that the left-hand side (or right-hand side) of Equation (1.4) is symmetric in x and y for the transition kernel of the Metropolis-Hastings algorithm.

$$\begin{aligned}\pi(\mathrm{d}y)K(y, \mathrm{d}x) &= \pi(y)\mathrm{d}y\, q(y,x)\alpha(y,x)\mathrm{d}x = \pi(y)q(y,x)\min\left[1, \frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}\right]\mathrm{d}x\mathrm{d}y \\ &= \min\left[\pi(y)q(y,x), \pi(x)q(x,y)\right]\mathrm{d}x\mathrm{d}y\end{aligned}$$

Note that the normalising constant of the target distribution, π, is not required in the acceptance ratio, as it cancels.

1.7.5 The Gibbs sampler

The Gibbs sampler is a special case of the Metropolis-Hastings algorithm used to sample from a d-dimensional target distribution. At each iteration, the i-th component is drawn from its conditional distribution dependent on all the other components which are fixed. This can either be implemented with the components updated in sequential order (which generally does not satisfy detailed balance) or with the component to be updated decided randomly (which does). Either way, the Gibbs sampler generates a Markov chain with π as its invariant distribution. The Gibbs sampler can be applied fairly automatically, as it does not require the design or tuning of a proposal distribution as in the Metropolis-Hastings algorithm.

1.7.6 Annealed Importance sampling and Sequential Monte Carlo

When implementing Bayesian statistics, the aim often is to calculate the expectation of some function, $\varphi(\cdot)$, with respect to the posterior distribution. If the posterior distribution is very complex and/or multimodal it may be difficult to generate samples according to it in a reasonable amount of time as standard MCMC methods tend to get stuck in one mode. Annealed importance sampling (AIS) (Neal, 2001) combines importance sampling with the use of Markov chains to overcome these problems.

In AIS a high-dimensional target distribution which has the posterior as a marginal distribution is constructed. Importance sampling is then used to estimate the expectation of functions with respect to this distribution.

Denote by $p_0(x)$ the distribution of interest, and design a sequence of densities, $p_1(x)$ to $p_n(x)$, each of which must be known up to a constant of proportionality, $f_1(x)$ to $f_n(x)$ (indexing is consistent with Neal (2001)). There must be available a Markov chain transition kernel, T_j, that leaves each p_j invariant. One suggestion is to construct the densities as follows

$$f_j(x) = f_0^{\beta_j} f_n(x)^{1-\beta_j},$$

where $1 = \beta_0 > \beta_1 > ... > \beta_n = 0$. For Bayesian statistics f_n would be the prior (which can generally be sampled) and f_0 would be the unnormalised posterior (which generally can not).

Next an expanded target density is constructed

$$p(x_0 \dots x_{n\text{-}1}) \propto f_0(x_0)\tilde{T}_1(x_0,x_1)\tilde{T}_2(x_1,x_2)\dots\tilde{T}_{n\text{-}1}(x_{n\text{-}1},x_{n\text{-}1}), \tag{1.5}$$

which has the distribution of interest, $p_0(x_0)$, as its marginal. $\tilde{T}_1(x_0,x_1)$ is the reverse of transition $T_1(x_1,x_0)$ i.e. $T_1(x_1,x_0)p(x_1) = \tilde{T}_1(x_0,x_1)p(x_0)$. We cannot sample from (1.5), but we can sample from

$$g(x_0 \dots x_{n\text{-}1}) = f_n(x_{n\text{-}1}) T_{n\text{-}1}(x_{n\text{-}1}, x_{n\text{-}2}) \dots T_2(x_2, x_1) T_1(x_1, x_0), \tag{1.6}$$

by sampling x_n from the simple distribution p_n and then successively sampling each x_j using the invariant kernel T_j. Expectations can then be computed with respect to p_0 by drawing N samples from the importance density (1.6) and defining an importance weight

$$w^{(i)} = \frac{f_{n-1}(x_{n\text{-}1}) f_{n\text{-}2}(x_{n\text{-}2}) \dots f_0(x_0)}{f_n(x_{n\text{-}1}) f_{n\text{-}1}(x_{n\text{-}2}) \dots f_1(x_0)},$$

using

$$\mathbb{E}_{p_0}[\varphi(x)] \approx \sum_{i=1}^{N} w^{(i)} \varphi(x^{(i)}) \Big/ \sum_{i=1}^{N} w^{(i)}.$$

Integrating with respect to the importance density, g, shows that the expectation is correct. An estimate of the normalising term of p_0 can also be obtained using

$$\hat{\mathcal{Z}} = \sum_{i=1}^{N} w^{(i)}.$$

By noting that the definition of the extended target distribution in (1.5) is arbitrary, a whole range of algorithms can be defined. This is the idea that the more general algorithm Sequential Monte Carlo (SMC) is based on. In addition to the generalisation of the target, an extra step, known as resampling takes place at each stage. The importance samples are resampled according to their normalised weights after they are sampled from their invariant kernel T_j. This reduces the variance of the estimate, by removing particles with low weight, whilst still retaining the unbiasedness property (Pitt et al., 2012). There is a large literature on the various ways to implement SMC and in particular on ways to design the high dimensional target distribution, for example Moral et al. (2006); Cappé et al. (2007).

1.7.7 Perfect sampling

In MCMC, one attempts to sample from a distribution π by designing a Markov chain with π as its invariant distribution. However, whilst it is possible in some cases to bound the mixing time and therefore ensure that the distribution sampled is close to the target distribution, in most cases this isn't possible and practitioners must rely on subjective judgements and heuristics to decide whether or not a chain has converged.

Perfect sampling or 'coupling from the past' (Propp and Wilson, 1996; Fill, 1997) is a more ambitious idea utilising Markov chains, which aims to draw samples from the exact target distribution. The idea is to simulate from a Markov chain which has been running for infinitely long without having to simulate the entire chain.

Assume we wish to sample from a discrete distribution π with a state space $\mathcal{S}$ of M states (work has been done to generalise the method to continuous spaces (Murdoch and Green, 1998)). In the perfect sampling literature, two Markov chain are 'coupled' if they use the same sequence of random numbers. If two chains are coupled and their trajectories meet, they will follow the same trajectory for all subsequent time steps, they will 'coalesce'.

The transitions in the Markov chain can be thought of as a deterministic function of the current state and the random numbers used in the update. Consider a Markov chain with a unique stationary distribution, π, and deterministic update function, ϕ, and imagine running the chain from the infinite past up to the present. If this were possible, we would be drawing samples from π. In reality, we cannot run the chain from $-\infty$ to $t = 0$; instead Algorithm 2 is implemented to draw one sample from π.

Algorithm 2 The Coupling from the Past (CFTP) algorithm

for $t = 1$ to R (where R is a random variable) **do**
  Start chains from each state in $\mathcal{X}$ at time $-t$ and run to time $t = 0$.
  Use the same uniform random variables for each chain.
  At time $t = 0$
  if chains have coalesced **then**
    Set $t = R$, return X_0
  else
    Set $t = t + 1$.
  end if
end for

To see why the algorithm works, we first set up some notation based on the description given in Dimakos (2001). The m-th state of $\mathcal{S}$ is denoted x_m and $X^{t_2}(t_1, x_m)$ is the state at time t_2 of a chain started in state x_m at time $t_1 < t_2$. Transitions in the Markov chain are viewed as deterministic functions of the current state and the random numbers involved

$$X^0(-t, x_m) = \phi(\phi(\cdots\phi(x_m, U^{-t+1}), \cdots U^{-1}), U^0).$$

To shorten notation, define the transition function

$$X^{t_2}(t_1, x_m) = \Phi_{t_1}^{t_2}(x_m, U^{t_1+1}, \ldots, U^{t_2}),$$

and the event

$$A_{t_1}^{t_2} = \{\Phi_{t_1}^{t_2}(x_m, U^{t_1+1}, \ldots, U^{t_2}) \text{ equal the same value for all } x_m \in \mathcal{S}\}.$$

For an ergodic Markov chain with $P(A_0^L) > 0$, events $A_{-kL}^{-(k-1)L}$, $k = 1, 2\ldots$ are independent and have same positive probability of occurring. As there are an infinite number of these events, the probability that one of them will occur is 1. Let T^* be the smallest t for which A_{-t}^0 occurs. As the same random numbers are used for each run, A_{-t}^0 also occurs for all $t > T^*$ including for $A_{-\infty}^0$. Therefore $X^0(-T^*, x_m)$ is the state visited by an infinitely long Markov chain with π as its stationary distribution, $X^0(-T^*, x_m) \sim \pi$.

Useful results on monotonicity of the state space mean that for certain models (including Ising models) it is only required to start chains in the 'maximum' and 'minimum' states (Propp and Wilson, 1996), which considerably simplifies the algorithm and reduces the computation.

1.8 Approximate Bayesian Computation

The use of standard Markov chain Monte Carlo methods is limited by the requirement to compute the likelihood. For many models, computing the likelihood is infeasible, perhaps for computational reasons or because the model is purely generative. Approximate Bayesian Computation (ABC) has been developed over the past 20 years to allow inference in situations where data can be simulated according to a model, but the value of the likelihood cannot be computed (Marjoram et al., 2003; Fearnhead and Prangle, 2012).

The ABC method stems from Algorithm 3 which draws samples directly from the true posterior.

Algorithm 3 Likelihood-free rejection algorithm

for $n = 1$ to N **do**
  Simulate θ' from the prior $\pi(\cdot)$.
  Simulate pseudo-data $\mathbf{x}$ from the likelihood $p(\cdot|\theta')$
  Accept θ' if $\mathbf{x} = \mathbf{y}$.
end for

Whilst Algorithm 3 does not require computation of the likelihood, the acceptance probability is proportional to $p(\mathbf{y})$ and hence it is often not computationally feasible for discrete data, and cannot be used for continuous data. The algorithm is therefore 'made approximate' by removing the strict requirement that the pseudo-data equal the observed data, and instead accepting θ' if some distance, d, between the data and pseudo-data $d(\mathbf{x},\mathbf{y}) < \varepsilon$, or if the distance between some low dimensional statistics, η, of the data $d(\eta(\mathbf{x}),\eta(\mathbf{y})) < \varepsilon$. This algorithm draws samples from $p(\theta|d(\eta(\mathbf{x}),\eta(\mathbf{y})) < \varepsilon)$, and the hope is that if the statistics are informative enough about the parameter values and ε small enough, then the distribution will approximate the true posterior.

There have been many developments towards making this methodology more efficient and widely applicable, such as the development of likelihood-free MCMC algorithms (Marjoram et al., 2003), the development of SMC methods for ABC (Toni et al., 2009; Moral et al., 2012) as well as efforts to automate the selection of the low dimensional statistics (Fearnhead and Prangle, 2012). These will be discussed in more detail in Chapter 4.

1.9 Conclusion

This Chapter has summarised and introduced the necessary background for the main contributions of the thesis. ARC syndrome and what is currently known of its pathogenesis has been described, as well as the data to be analysed in the next chapter. Mathematical/computational modelling within Systems Biology has been introduced and its importance in modern biological research stressed. Finally, the use of the Bayesian framework to propagate uncertainty through to parameter estimates has been justified and the computational techniques used in its application outlined.

Chapter 2

Multivariate analysis of transcriptomic and metabolomic data

2.1 Aims of analysis

The aims of this chapter are to analyse the transcriptomic and metabolomic data to identify changes in genes/metabolites, pathways and networks of interactors when the expression of VPS33B, VIPAR and PLOD3 are significantly reduced. Experimental work suggests that VPS33B and VIPAR form a complex together and knocking-down/mutating either protein leads to similar cell and disease phenotypes, so the first question is how similar are the transcriptomes/metabolomes of the two knock-downs. Similarly, PLOD3 has been identified as an interactor of the VPS33B-VIPAR complex and PLOD3 knock-down cell lines have a similar phenotype to VPS33B/VIPAR knock-downs. Experiments in mouse IMCD cells suggest that PLOD3 is trafficked to its functional site in collagen secreting carriers via a VPS33B-VIPAR dependent pathway. Hence we might expect a subset of the changes observed in VPS33B/VIPAR knock-downs to be seen in the PLOD3 knock-down cells. An exploratory data analysis is performed to identify genes, pathways and networks of interactors which may be relevant to the pathogenesis of ARC syndrome and which can suggest future avenues for experimental work.

The pipeline followed is shown in Figure 2.1. The first stage is processing the data so that useful information about the differences between experimental conditions can be inferred. This is very important as in both cases the measurement process

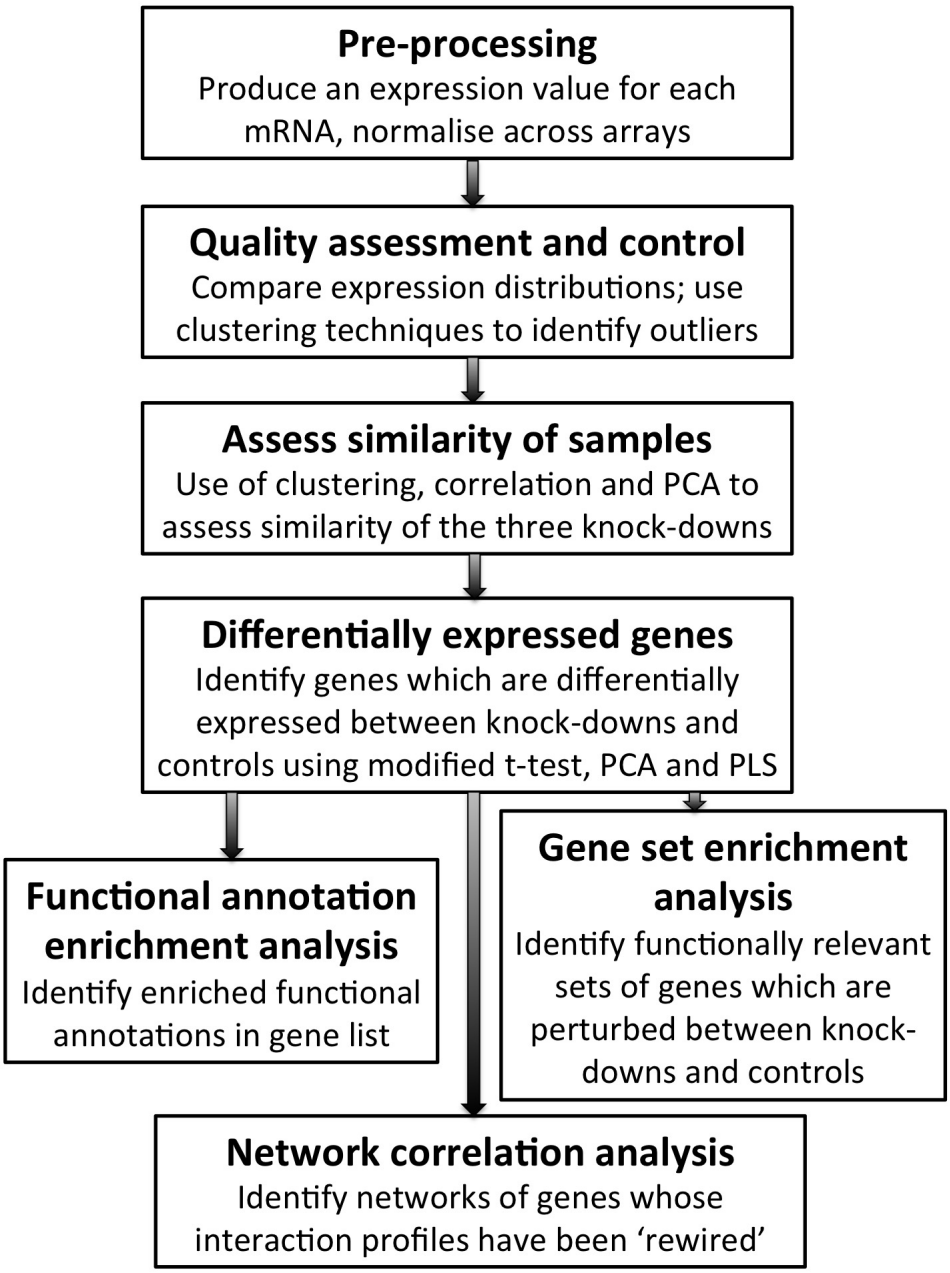


Figure 2.1: Flow chart showing initial analysis carried out on transcriptomic data from knock-down IMCD cell lines.

is complicated and involves multiple stages. The raw transcriptomic data, for example, consists of fluorescence measurements for multiple probes targeting each mRNA molecule which need to be transformed into an expression level for each mRNA. The data also needs to be normalised so that measurements across samples can be compared. Then the quality of the data needs to be assessed to check if any of the samples are outliers and if samples within the same experimental groups are similar.

Once pre-processing is complete, hierarchical clustering, correlation measures and Principal Component Analysis (PCA) projections are used (described in Section 2.3) to assess the similarity of the transcriptomes and metabolomes across samples.

After this, differentially expressed genes/metabolites are identified. This can be done using a t-test (modified to account for small sample size and multiple testing) or using multivariate techniques such as PCA or Partial Least Squares (PLS) (described in Sections 2.4, 2.5 and 2.6). A comprehensive review of the literature relating to the functions of these genes gives insight into the differences between the control and knock-down cells, however, in order to better understand which underlying biological functions are perturbed, techniques such as Gene Set Enrichment analysis (GSEA) and functional annotation enrichment can be used. These look at groups of genes and assess which particular functions or processes are perturbed by the knock-down (described in more detail in Section 2.7).

Finally network approaches, in which pairwise correlations between gene expressions are used as a proxy for gene interaction, are used to identify nodes or sets of nodes whose interaction profiles differ between the controls and knock-downs (more detail in Section 2.8).

Descriptions of the techniques used, results and a discussion are presented in the remainder of this chapter.

2.2 Initial data processing

2.2.1 Microarrays: initial processing

The raw transcriptomic data consists of fluorescence measurements for multiple probes (called a ‘probe set’) targeting each mRNA transcript. The initial processing

aims to provide an expression level for each mRNA molecule in the sample. In order to do this, fluorescence measurements from each probe set must be combined in such a way that the result is strongly related to the actual amount of mRNA in the sample. A technique called Robust Multiarray Average (RMA) (Bolstad et al., 2003; Bolstad, 2004; Irizarry et al., 2003) is a set of processes which aim to 'clean up' and summarise the data through

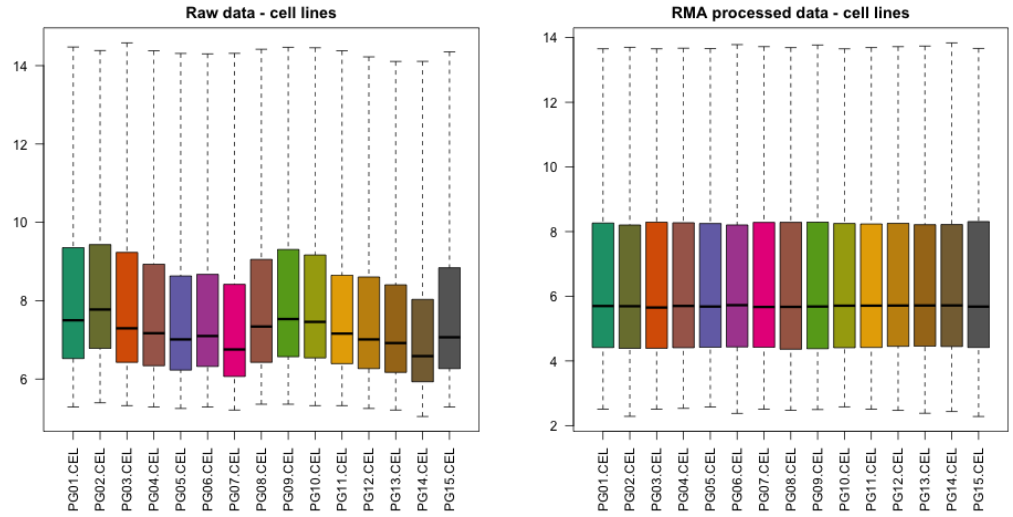
1. Background correction
2. Normalisation across arrays
3. Probe summarisation.

Background correction primarily aims to deal with background noise and processing effects, however it can also adjust for cross-hybridisation. The RMA method assumes that the observed signal (O) is the sum of an exponential true signal (S) and a truncated normal background (so that the expected value will always be positive). The expected value of the true signal given the observed signal, $\mathbb{E}[S|O]$, can then be calculated analytically (see Bolstad (2004) for detailed description).

Normalisation removes unwanted variation resulting from non-biological factors such as machine settings, different amounts of sample etc. To normalise across all arrays, the assumption is made that only a few genes change their expression significantly, and therefore that the differences in the expression of most other genes are due to other factors. Hence, even for different experimental conditions the distribution of intensities should be the same. In the quantile normalisation used in RMA, each array (considered as a vector) is sorted so that the entries are in intensity order, some summary statistic of the n-th entry across arrays (e.g. the mean or the median) replaces the n-th entry in each vector and then the vectors are sorted again back to their original order.

The final stage deals with the multiple probes each targeting a different region on the same mRNA molecule. Probe summarisation requires combining the measurements from each probe within the set to give one overall value for the expression level of the mRNA. A robust multi-chip linear model on the log intensities is used to fit the data.

A multitude of techniques for initial data processing have been developed and there is no consensus as to which is best. RMA processing was selected as it is the most



(a) **(b)**

Figure 2.2: Top row shows log-intensity histograms and bottom row shows box plots of log-intensity. Figures on the left are before RMA processing while figures on the right are after. The data is from knock-down IMCD cell lines.

commonly used procedure in the literature and has been shown to be effective in reducing between-microarray variation (Bolstad, 2004).

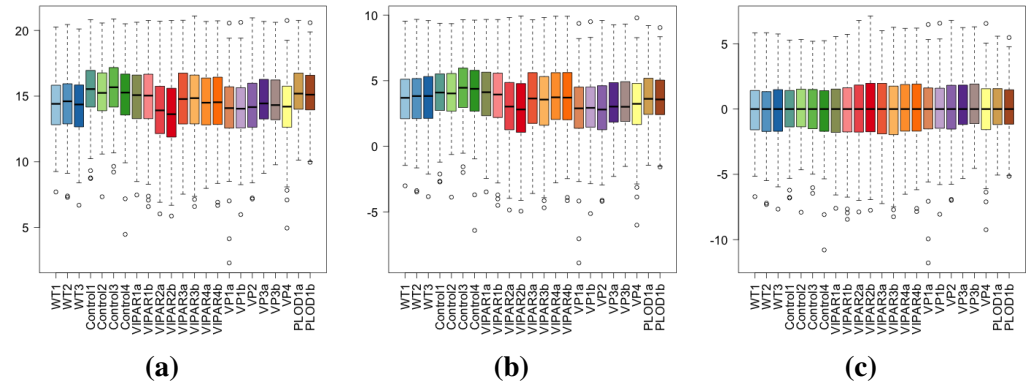
The normalisation was found to reduce significantly between-array variation in terms of both variance and bias (see Figure 2.2), as well as being significantly faster than competing methods.

2.2.2 Metabolomics: initial processing

The metabolomics data had already been processed such that it consisted of an abundance measure for each metabolite. This data was log-transformed as it was right-skewed and abundances of individual metabolites varied considerably over several orders of magnitude. The data was then normalised to enable comparison between samples by scaling with the median of each sample. As can be seen in Figure 2.3, samples are more comparable after subtraction of the median (log abundance) for each sample.

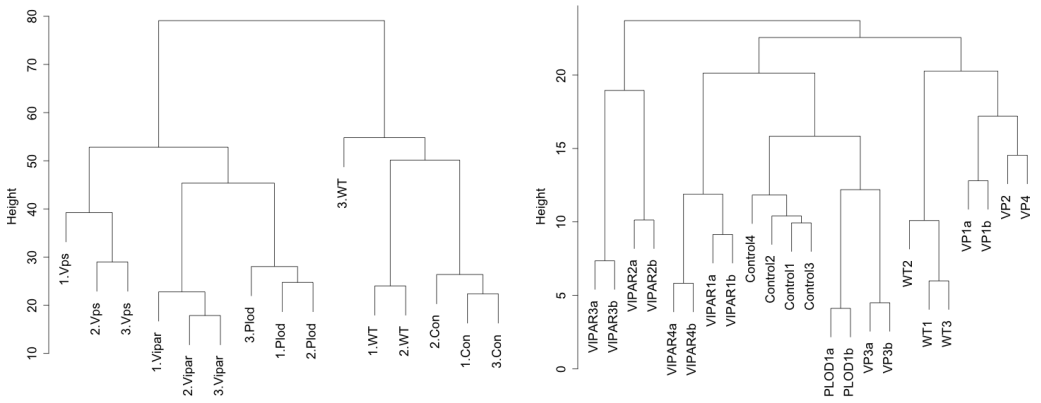
2.3 Extent of similarity between knock-downs

There are several ways to assess the similarity of samples. Here we focus on three techniques: hierarchical clustering, correlation analysis and PCA.



(a) **(b)** **(c)**

Figure 2.3: The three boxplots show the metabolomic data for each sample (a) log transformed, unnormalised (b) log-transformed, normalised to an internal standard (c) log transformed, normalised to median of each sample.

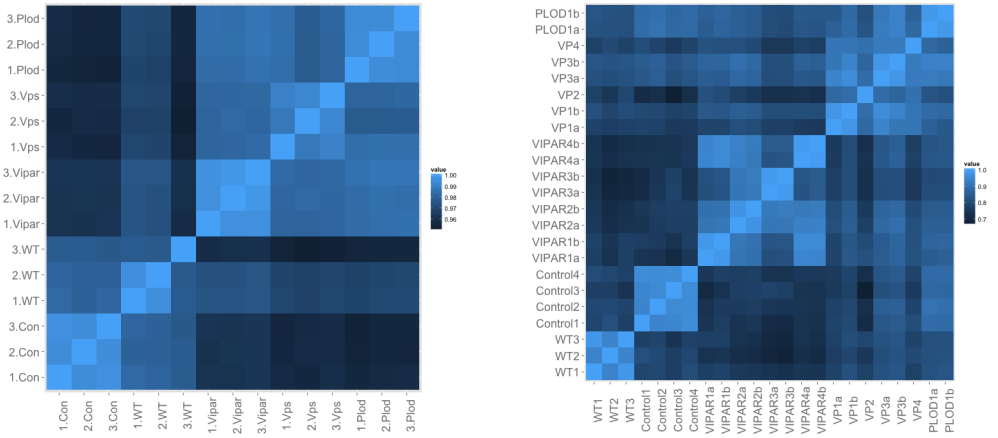


(a) **(b)**

Figure 2.4: Hierarchical clustering of (a) transcriptomic and (b) metabolomic samples based on Euclidean distance.

Hierarchical clustering algorithms can either be agglomerative, in which each sample is treated as a singleton and pairs are successively merged based on how 'similar' they are, or divisive, in which all samples start in one cluster which is split based on dissimilarity between two subgroups. A variety of different distance measures can be used to define similarity e.g. Euclidean or the maximum across dimensions and there are a number of ways to compute the distance between clusters e.g. the maximum, minimum or average distance between the samples in two clusters.

Agglomerative clustering based on Euclidean distance and average linkage were used to produce the dendrograms in Figure 2.4 (although other choices produced



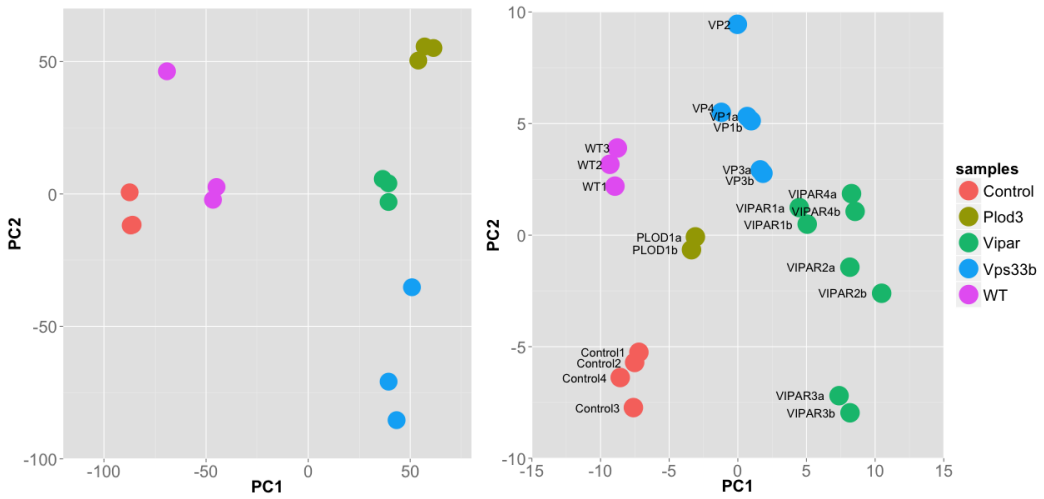
(a) **(b)**

Figure 2.5: Heatmaps showing Spearman's rank correlation between (a) transcriptomic and (b) metabolomic samples.

the same clusters). The transcriptomic data, shown on the left, clusters both into experimental groups, as well as into larger clusters of knock-downs, which *a priori* would be expected to be similar, and controls. One wild type sample appears to have considerably different expression levels to the other five control samples, however it is still clustered correctly. The metabolomic data, on the right, also broadly clusters into experimental groups, although the wild type and control data do not cluster together and the Vps33b knock-down cells do not cluster as one big group. From the heat maps in Figures 2.8 and 2.10 (discussed in more detail below) it seems clear that the transcriptomic samples are more consistent within experimental groups than the metabolomic samples and that the RNA transcript measurements are less affected by the transfection process.

The Spearman's correlation between each pair of samples is shown in Figure 2.5. This measure of the correlation is based on ranks rather than absolute values to reduce the impact of outlier measurements. A similar picture is seen here as for the hierarchical clustering. For the transcriptomic data shown in (a), stronger correlations are seen within experimental groups and between knock-downs and controls. Again, one wild type sample appears different to the others. There are also strong correlations within experimental groups for the metabolomic data, although the wild type and control samples do not appear to be very well correlated, again implying that the silencing sh-RNA transfection process affects metabolite measurements strongly.

Finally, PCA is a technique which can be used to reduce the dimensionality of



(a) Transcriptomics score plot **(b)** Metabolomic score plot

Figure 2.6: PCA score plots for (a) transcriptomic data and (b) metabolomic data. The original high-dimensional data is projected onto two axes which explain variation in the data.

and visualise high-dimensional data. This will be explained in more detail in Section 2.5, however for now it can be viewed as a technique in which the data is projected onto two orthogonal axes, known as the Principal Components, which explain large proportions of the variation in the data. From Figure 2.6, similar patterns to those seen above can be seen for both types of data, with experimental groups largely clustering together and knock-downs separated from controls by the first principal component.

Overall, from the hierarchical clustering, correlation analysis and PCA it appears that, certainly for the transcriptomic data, the VPS33B and VIPAR knock-down cell lines have very similar expression profiles, distinct from those of the controls. This certainly does not prove that VPS33B and VIPAR are involved in similar functions but does mean that this hypothesis is feasible. The PLOD3 knock-down also has some similarity to the VPS33B/VIPAR knock-downs. The metabolomic data is considerably more variable and doesn't show the same level of consistency between wild type/control cells and VPS33B/VIPAR knock-downs. However, there is still consistency within groups and so the data can be further investigated for insights into ARC syndrome.

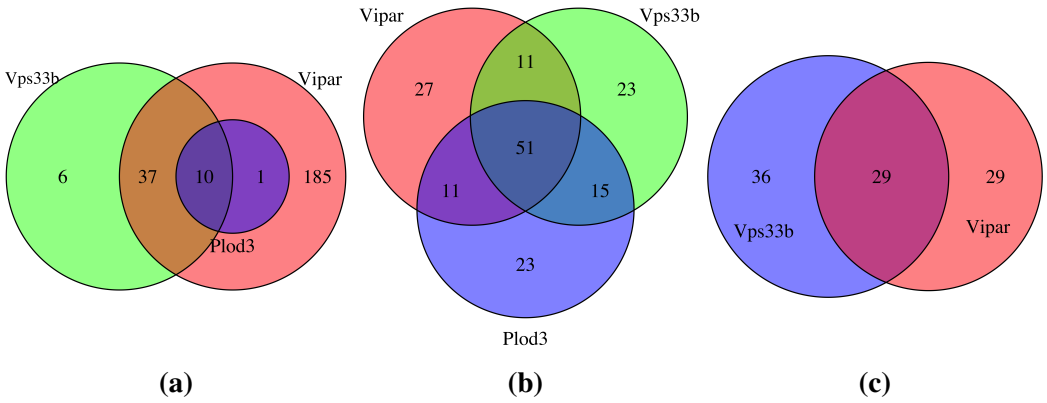
2.4 Differentially expressed genes/metabolites

In order to identify genes which are differentially expressed between the knock-downs and controls, the simplest initial analysis involves using a t-test. A modified version of the t-test from the limma R package (Smyth, 2005) was used for the transcriptomic data as the empirical Bayes approach taken stabilises estimates when the number of samples is small. As many tests are performed at the same time, the p-values need to be adjusted to avoid a large number of false positives, and here the Benjamini-Hochberg correction controls the false discovery rate (expected proportion of false positives in all discoveries).

Lists of differentially expressed (DE) genes for each of the knock-downs were compiled, both to see what processes these genes are involved in and to compare how similar the lists are across the knock-downs. Using a cut-off of FDR< 0.05 there are 53, 233 and 11 DE genes respectively in the VPS33B, VIPAR and PLOD3 knock-downs. Figure 2.7 (a) shows a Venn diagram of the list overlaps. All of the PLOD3 genes are differentially expressed in at least one of the other two knock-down cell lines which is compatible with the hypothesis that PLOD3 functions downstream of VPS33B-VIPAR. A large percentage of the VPS33B genes are also DE in the VIPAR knock-down cell line with all overlaps highly significant as assessed by hypergeometric test.

It is not clear whether the large number of DE genes in the Vipar knock-down cells is because the protein levels were reduced more than the other cell lines leading to more detectably differentially expressed genes, or whether VIPAR has extra functions compared to the other two proteins. Taking the top 100 DE genes in each knock-down, (shown in Figure 2.7 (b)) the strong overlap between all three cell lines is maintained implying it may be the former.

Given that VPS33B and VIPAR form a complex, we can combine the data from these two knock-downs and combine all the controls. This will increase the statistical power and help to identify genes specifically related to the joint function of VPS33B/VIPAR. Table 2.1 and Figure 2.8 show the functions of the top 20 down-regulated genes and a heat map of the most differentially expressed genes. The majority of the top 20 genes were down-regulated, only one was up-regulated: Galm (Galactose mutarotase) which catalyzes the interconversion of the α- and β-anomers of either galactose or glucose (Frey, 1996). Of the down-regulated genes, several have (at least some part of their) function known. Many of these genes are



(a) **(b)** **(c)**

Figure 2.7: Venn diagrams showing the overlap in differentially expressed genes ((a) and (b)) and metabolites (c). Differentially expressed genes were identified using the limma package in R and differentially abundant metabolites using a standard t-test. In both cases p-values were Benjamini-Hochberg corrected for multiple testing and a cut-off of FDR<0.05 applied. In (a) and FDR cut-off of <0.05 was used. In (b) the top 100 genes ranked by p-value were used.

expressed in epithelial cells and are involved in polarity (Ap1m2, Mal2), trafficking (Cc2d2a, Rab25) or in cell-cell adhesion (Cldn7, Macc1, Sema5a), corroborating published work stating that these functions are perturbed in patients and cellular models of ARC syndrome (Gissen et al., 2004, 2006; Cullinane et al., 2010).

We can perform a similar analysis with the metabolomic data. In this case we use a standard t-test, with p-values adjusted for multiple testing so that FDR<0.05. Figure 2.9 shows all metabolites identified as up (green) or down (red) relative to mock-injected controls, as well as whether the same metabolites were identified as up or down in controls compared to wild type cells. Heat maps showing expression levels of these metabolites are shown in Figure 2.10 and it is clear that some of the differences detected are very small. The most likely candidates for a true difference (i.e. those with a small p-value when comparing knock-downs with controls and with large p-values when comparing controls to wild type) are Serine, Beta Alanine, Monomethylphosphate, Adenosine-5-monophosphate and Threonine. A Venn diagram showing the overlap in metabolites identified as differentially abundant in VPS33B and VIPAR knock-downs is shown in Figure 2.7 (c) (overlap not significant as assessed with a hypergeometric test).

Symbols	Function
Esrp1	Along with Esrp2, regulator of an epithelial splicing regulatory network, in particular promote splicing of the epithelial variant of the FGFR2, ENAH, CD44, and CTNND1 transcripts (Warzecha et al., 2009)
Mal2	Membrane protein, part of basolateral-to-apical transcytosis machinery, essential for formation of central lumen (Madrid et al., 2010)
Atp8b1	P-type ATPase, implicated in the inward translocation of phospholipids in biological membranes. Deficiency causes Progressive Familial Intrahepatic Chloestasis Type 1 and Benign Recurrent Intrahepatic Cholestasis Type 1 (Paulusma et al., 2008)
Tmprss2	Androgen regulated serine protease, may be involved in virus activation via cleavage (Glowacka et al., 2011)
Cldn7	Tight junction component, loss reduces integrin expression and promotes mesenchymal traits via regulation of Rab25 (Bhat et al., 2014)
Macc1	Regulator of the HGF/Met signaling pathway which plays an important role in cell motility, metastatis and invasiveness (Stein et al., 2009)
Nipal2	Mutations in homologue Nipal4 cause autosomal recessive congenital ichthyosis (RodriguezPazos et al., 2011)
Tmem30b	
Tmem184a	
Cc2d2a	Mutations cause ciliopathy diseases, may facilitate protein transport through a role in Rab8-dependent vesicle trafficking (Bachmann-Gagescu et al., 2011)
Fermt1	Involved in organisation and anchorage of the actin cytoskeleton to integrin-associated platforms (Lai-Cheong et al., 2009; Mas-Vidal et al., 2010)
Sema5a	Membrane protein, involved in axon guidance in neuronal cells and regulates cell adhesion and motility in epithelial cells (Capparuccia and Tamagnone, 2009)
Rab25	From family of small GTPases involved in membrane trafficking, specifically involved in sorting of integrin to lysosomes (Dozynkiewicz et al., 2012)
Epha1	Mainly expressed in epithelial cells where it regulates cell morphology and motility (Yamazaki et al., 2009)
Ano9	
Ehf	ETS transcription factor subfamily, expressed in epithelial cells, represses the expression of key EMT genes such as TWIST1, ZEB2, BMI1, and POU5F1 (Albino et al., 2012)
Fam132a	
Ap1m2	This gene encodes a subunit of the heterotetrameric adaptor-related protein complex 1 (AP-1), mediates protein sorting to regulate epithelial cell polarity and proliferation (Hase et al., 2013)
Slit2	May be involved in cell motility and apoptosis in epithelial cells (Alajez et al., 2011)

Table 2.1: Table showing the top 20 most down-regulated transcripts, as detected by modified t-test, and some detail on what is known of their function. Wild type and mock-transfected cells were treated as controls and VPS33B and VIPAR knock-down cell lines as one class.

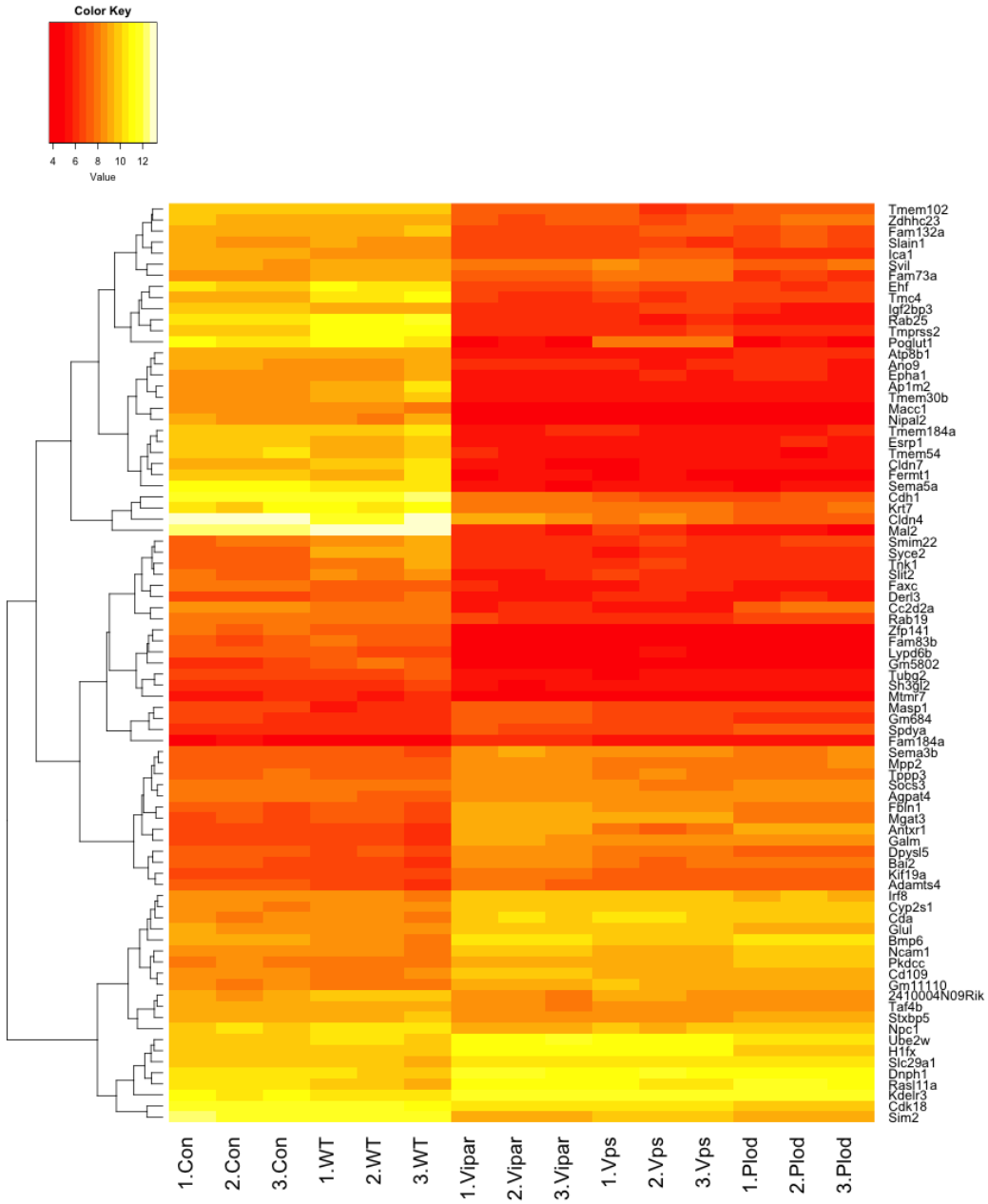


Figure 2.8: Heat maps showing RNA transcripts which are significantly differentially expressed based on comparing all controls against VIPAR and VPS33B knock-downs. Expression levels in PLOD3 are shown although this data was not used in the analysis.

Metabolite	Vipar	Vps33b	Control
Cysthionine			
Gluconic.acid			
Proline			
o-Phosphoetanolamine			
Glycerol 3 phosphate			
Myo.inositol			
Serine			
Lysine			
Monomethylphosphate			
Adenosine 5 Monophosphate			
Beta-alanine			
Tyrosine			
Pyruvic.acid			
Oxalic.acid			
Threonine			
Phenylalanine			
Glyceric acid 3 phosphate			
Glycine			
Phosphoric acid			
Lactic.acid			
L.Cysteine			
Ornithine			
Hexadecanoic.acid			
4-Aminobutyric acid (GABA)			
Malic.acid			
Citric.acid			
Stearic.acid			
Aspartic.acid			
Cholesterol			
Glutamine			
Asparagine			
Spermine			
Tryptophan			
L Glutamic acid			
Spermidine			

Figure 2.9: Metabolites with significantly different levels in either the VIPAR or VPS33B knock-down cell lines. Green indicates up compared to control, red down compared to control. The third column shows metabolites up or down in mock-transfected compared to wild type cells.

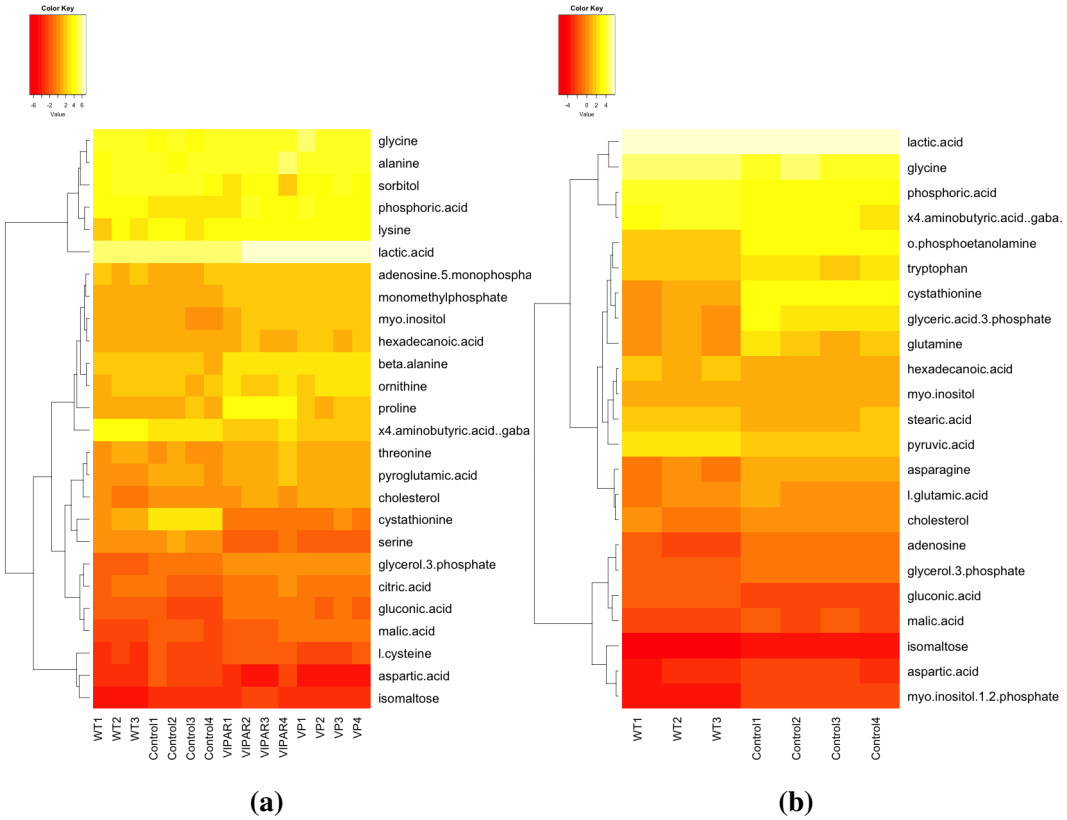


Figure 2.10: Heat maps showing metabolites which are significantly differentially expressed based on comparing (a) all controls against Vipar and Vps33b knock-downs (b) Mock transfected controls against wild type cells

2.5 Principal component analysis

Principal component analysis (PCA) is an unsupervised exploratory data analysis technique in which a set of observations are transformed such that the new variables are linearly uncorrelated and identify the principal directions in which the data varies. The first principal component is the linear combination of the original variables with maximal variance. The second principal component is the linear combination with second greatest variance, constrained to be orthogonal to the first, and so on. As the principal components are a linear combination of the original axes and are mutually orthogonal, the problem involves only a change of basis making PCA soluble with linear algebra decomposition techniques. The key assumption is that directions with large variance are believed to be directions of importance, i.e. representing signal as opposed to noise.

Start with a data matrix, $\mathbf{X}$, with n rows containing samples and p columns containing mean-centred variables, and then take a linear combination of the original

variables, $\mathbf{T} = \mathbf{XP}$. We would like the covariance matrix of $\mathbf{T}$ to be diagonal, i.e. for each of the new variables to be uncorrelated.

$$\mathrm{Cov}(\mathbf{T}) = \frac{1}{n-1}\mathbf{T}'\mathbf{T} = \frac{1}{n-1}(\mathbf{XP})'(\mathbf{XP}) = \frac{1}{n-1}\mathbf{P}'(\mathbf{X}'\mathbf{X})\mathbf{P}$$

From linear algebra, $\mathrm{Cov}(\mathbf{T})$ will be diagonal if the matrix $\mathbf{P}$ has columns which are the normalised eigenvectors of $1/(n-1)\mathbf{X}'\mathbf{X}$, i.e. the eigenvectors of the covariance matrix of $\mathbf{X}$. The columns of $\mathbf{P}$ are called the *principal components* of $\mathbf{X}$. Further, with this choice of $\mathbf{P}$, the diagonal elements of $\mathrm{Cov}(\mathbf{T})$, i.e. the variance of each new variable, are given by the eigenvalues of $\mathrm{Cov}(\mathbf{X})$, $\{\lambda_i\}_{i=1}^{p}$. Due to the orthogonality of the new basis, the variance of each component contributes independently to the overall variance. It is hence possible to compute the explained variance per component by dividing each individual variance (eigenvalue) by the trace of $\mathrm{Cov}(\mathbf{T})$,

$$\text{variance accounted for by } i\text{-th component} = \frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j}.$$

This is a major benefit of PCA as one of the hopes is that the original high-dimensional data can be reasonably represented using only a few of the principal components and hence the dimensionality can be significantly reduced.

2.5.1 PCA results

Figure 2.6 (already discussed) shows the transcriptomic and metabolic scores (transformed variables) projected onto the first two principal components. In both cases the various experimental groups are well separated and the first principal axis separates the knockdowns and controls. The second axis separates the three types of knockdown (not quite so cleanly for the metabolomic data). The first principal component should therefore be most informative as to the difference between knock-downs and controls as it explains 32% of the variance in the metabolic data and 29% in the transcriptomic data.

Figure 2.11 shows the PCA loadings for the metabolic data i.e. the extent to which

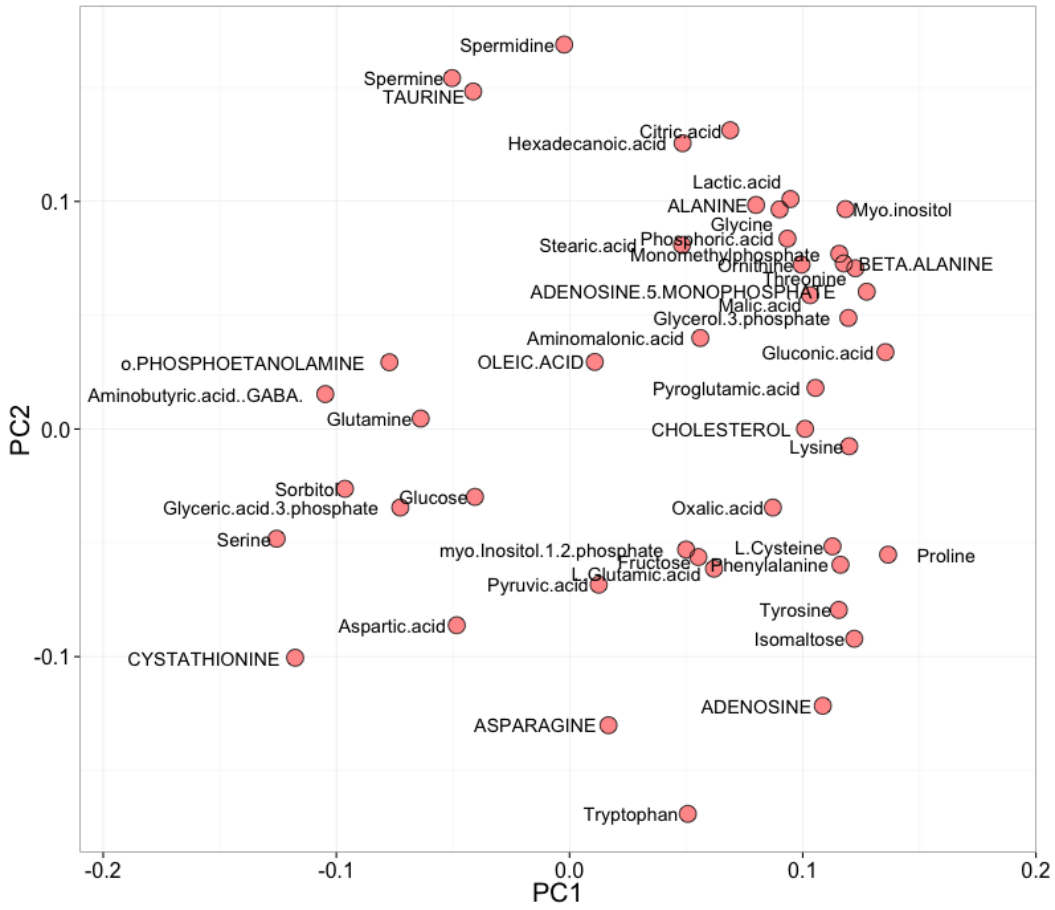


Figure 2.11: Metabolomic PCA loading plot

each original variable contributes to each principal component. As expected, many of the metabolites with large loadings coincide with metabolites already identified as having differential abundance, however, now Serine, Cystathionine, Proline and Gluconic acid contribute most to the first principal component. Given the large number of variables in the transcriptomic dataset, the loadings are provided in a table in Appendix A rather than in plot form, but many of the genes with large loadings overlap with the genes detected by moderated t-test.

2.6 Partial Least Squares

PCA is a useful technique to distill information in high-dimensional data, but it does not take the response vector into account when computing the principal components, and therefore predictive power may be reduced. It could be the case that a variable which explains very little variance in $\mathbf{X}$ is very strongly correlated with the response, $\mathbf{Y}$. Partial Least Squares (PLS) regression is an alternative approach

which takes the response variables into account when finding latent variables, so that they both model the variation and predict the response well (Wold et al., 1984, 2001). Both the dependent and independent variables are projected to a new space. To understand how PLS works, recall that in a general linear model we have

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E},$$

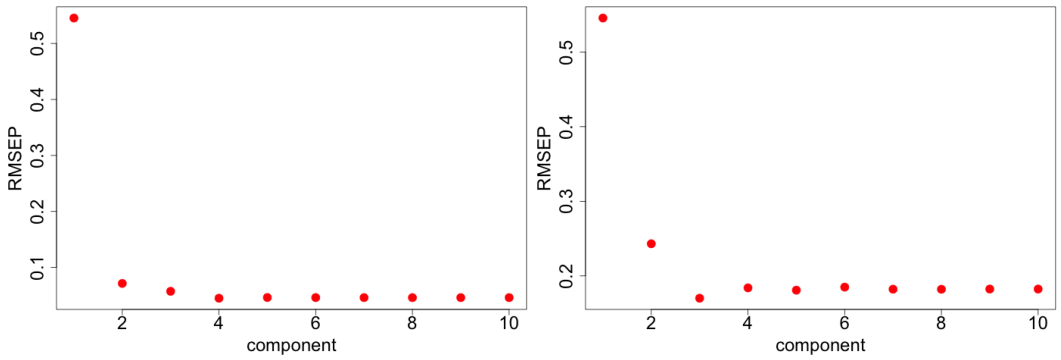
where $\mathbf{Y}$ is a $n \times m$ matrix of response variables, $\mathbf{X}$ is an $n \times p$ matrix of factors or predictor variables, $\mathbf{B}$ is an $m \times p$ matrix of parameters and $\mathbf{E}$ is an $n \times m$ matrix of errors. The aim is to explain or predict the responses using the measured predictor variables. If the predictors are few and not collinear, multiple linear regression can be used. However, with omic data there are often many, correlated predictors. In this case we may wish to replace the original variables of $\mathbf{X}$ with a smaller number of variables that have better properties. Instead of modelling exclusively the $\mathbf{X}$ variables, both the $\mathbf{X}$ and $\mathbf{Y}$ variables are modelled as follows:

$\mathbf{T} = \mathbf{XW}^*$	scores, $\mathbf{T}$, are a linear combination of $\mathbf{X}$
$\mathbf{X} = \mathbf{TP}' + \mathbf{E}$	$\mathbf{X}$ is approximated by the scores multiplied by a loading matrix, $\mathbf{P}$
$\mathbf{Y} = \mathbf{TC}' + \mathbf{F}$	$\mathbf{Y}$ is modelled by a multivariate linear regression on the scores, $\mathbf{T}$

PLS estimates the latent variable (LV) model parameters $\mathbf{W}^*$, $\mathbf{P}$ and $\mathbf{C}$ so as to obtain the linear combination of the x-variables which have maximum covariance with a certain linear combination of the y-variables (this is slightly different depending on the algorithm used). Both $\mathbf{X}$ and $\mathbf{Y}$ are assumed to be functions of a small number of common LVs $\mathbf{T}$. There are many different algorithms for computing weights, loadings and scores but one of of the most commonly used is the Non-Linear Iterative Partial Least Squares (NIPALS) algorithm in which the above matrices are computed iteratively.

2.6.1 PLS results

Separate PLS models were first fitted for the transcriptomic and metabolomic data using a univariate response vector, $\mathbf{y}$, with 0s for controls and 1s for knock-downs



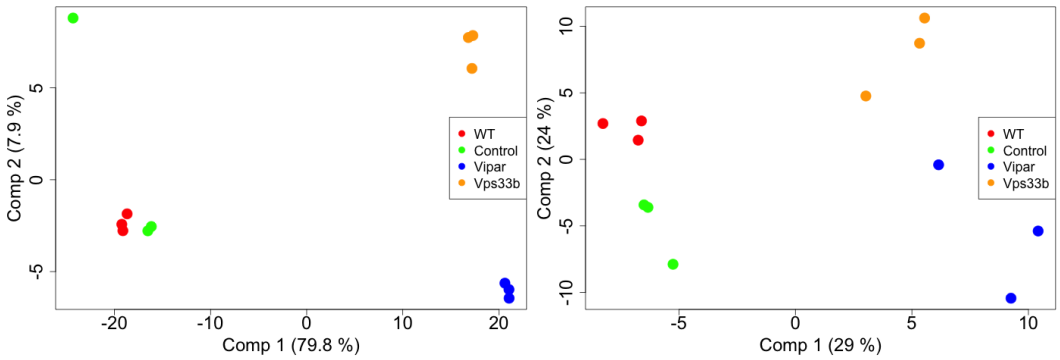
(a) transcriptomic data **(b)** Metabolomic data

Figure 2.12: Root mean square error in class prediction using PLS with varying numbers of components for (a) transcriptomic data (b) metabolomic data.

(VPS33B/VIPAR and Control/WT data combined). Figure 2.12 shows the root mean square prediction error for both models with various numbers of components, and it was decided based on this data to use two components in the transcriptomic model and three in the metabolomic model. Fewer predictive components are required in the PLS model compared to a Principal Component Regression model as the PLS algorithm takes the variation in **y** into account (data not shown).

Figure 2.13 shows the scores (**T**) projected onto the first two PLS components and clearly the first component separates knock-down cell lines from controls and the second generally separates the two different knock-downs. Note that both data types are now well separated by the first component, and so the variables which contribute to this component should be informative as to the difference between the knock-downs and controls. The first component explains a large proportion of the variance in the transcriptomic data, and the two components combined explain almost 90% of the variation. For the metabolomic data, the variation explained by the first two components is less, approximately 55%, however, adding more components does not increase the prediction error and an inspection of the scores projected onto higher components does not reveal further components (in addition to the first one) which separate the controls from the knock-downs.

To interpret the PLS model, the loadings (**P**) are inspected as these show the degree to which the latent variables contribute to **X**. The first and second component loadings for the metabolomic data are plotted in Figure 2.14, with Cystathionine, Proline and Serine again standing out as the major contributors to the first component. Cystathionine is generated from Serine and is an intermediate in the produc-



(a) transcriptomic data **(b)** Metabolomic data

Figure 2.13: Scores from the first two components of a PLS model for (a) transcriptomic data (b) metabolomic data.

tion of Cysteine (both of which have large loadings in Figure 2.14) (Berg et al., 2002, Chap. 23). Cysteine is an important amino acid structurally due to its ability to form disulfide bonds which form cross-links between polypeptide chains (Berg et al., 2002, Chap. 3). This is particularly important in the extracellular matrix (disulfide bonds are generally not stable in the cytosol). Serine is a non-essential amino acid which is concentrated in cell membranes due to its involvement in the production of membrane phospholipids (Berg et al., 2002, Chap. 26). Proline is a non-essential amino acid which is an essential component of collagen and which stabilises the collagen triple helix, and hence is important for cell structure and the proper functioning of joints and tendons (Berg et al., 2002, Chap. 8).

The genes with the top 20 loadings (Mal2, Sema5a, Peg3, Rab25, Fermt1, Cdh1, Poglut1, Tmprss2, Cldn7, Nipal2, Esrp1, Tmem184a, Tmem54, Cldn8, Cldn4, Epcam, Macc1, Atp8b1, Ap1m2) are heavily biased towards membrane proteins including three Claudins (which form the tight junction seal in epithelial tissue) and Cadherin 1 (one of the most important molecules in epithelial cell adhesion, located in the adherens junction (Pećina-Šlaus, 2003)). There are also a number of regulatory genes including Mal2 which regulates basolateral-to-apical transcytosis (Madrid et al., 2010), Macc1 which regulates the HGF/Met signalling pathway (Stein et al., 2009), Esrp1 which regulates a splicing network in epithelial cells (Warzecha et al., 2009) and Ap1m2 which regulates epithelial cell polarity via a role in protein sorting (Hase et al., 2013). Interestingly, several of the highlighted transcripts come much further down the list using PCA or univariate approaches, supporting the existence of variables which contribute little to the variation in $\mathbf{X}$ but which correlate strongly with the response variable.

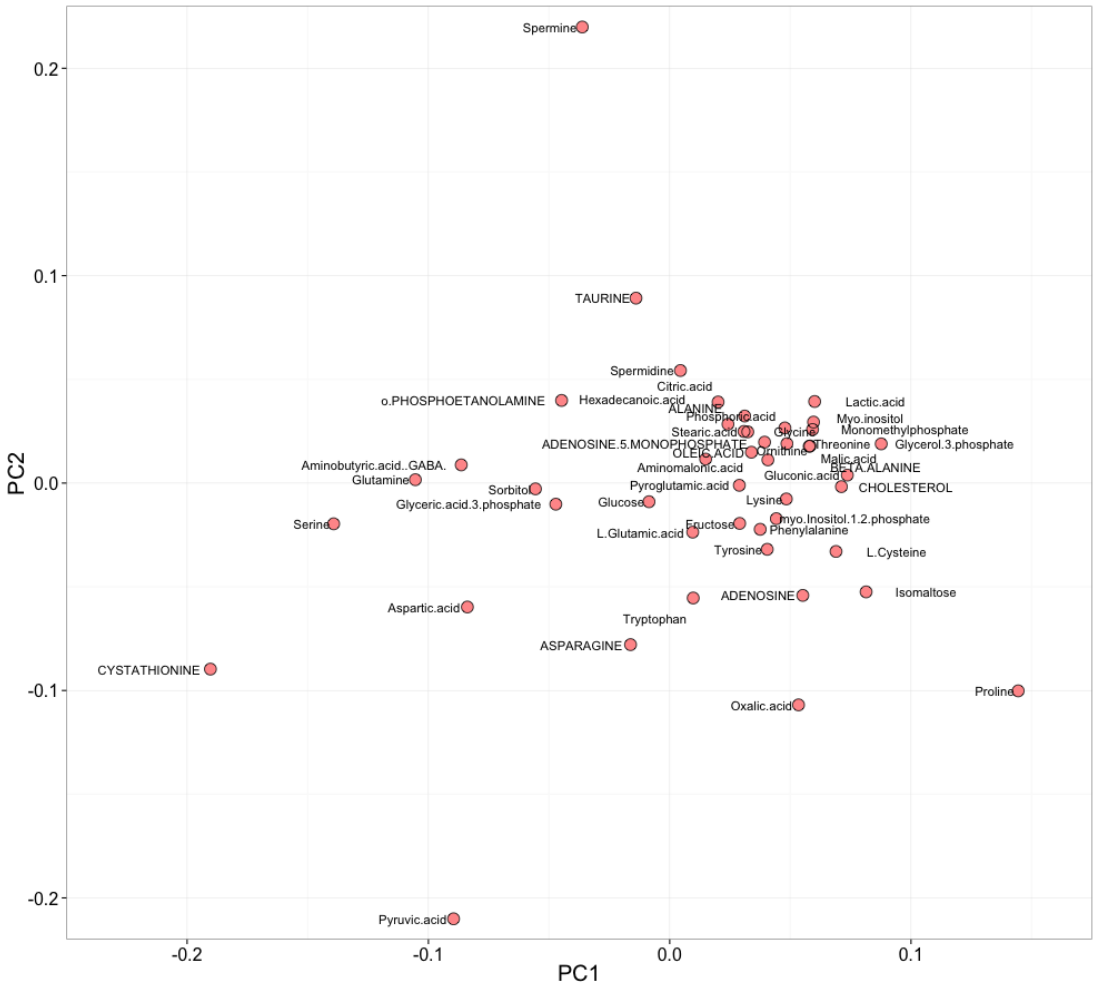
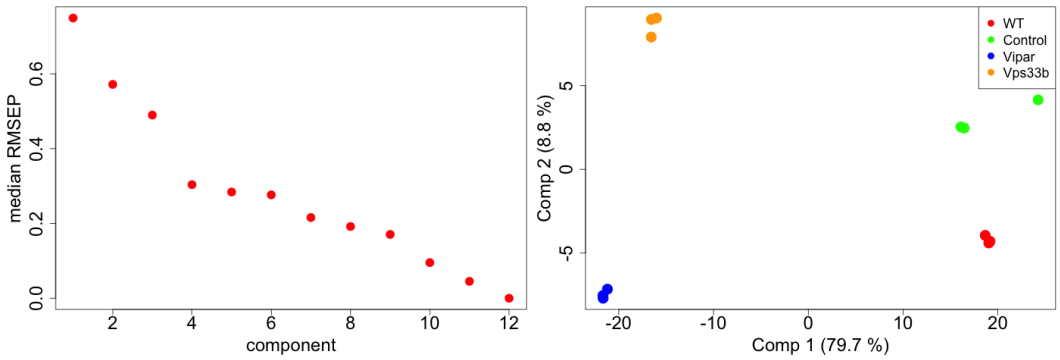


Figure 2.14: Loadings from the first two components of a PLS model for metabolomic data.

Given the nature of the data, we can also define a PLS model in which the transcript levels are the explanatory variables $\mathbf{X}$, and the metabolomic data is the response, $\mathbf{Y}$. This reflects the idea that metabolite abundances can be explained by changing RNA transcript levels. A PLS model was fitted and the root mean square error in prediction is shown in Figure 2.15 (a) using the median prediction error across the $\mathbf{Y}$ variables. A model was fitted using all the components and Figure 2.15 (b) shows the scores projected onto the first two components, with the first clearly separating knock-downs and controls. The top twenty genes are more or less the same as those identified in the PLS model of transcript levels predicting class membership, albeit in a slightly different order.

The second component may also be of interest as this separates the VPS33B and VIPAR samples. The genes with the largest loadings are enriched for glycopro-



(a) transcriptomic data **(b)** Metabolomic data

Figure 2.15: (a) Root mean square error of prediction in the full PLS model of metabolomic data predicted by transcriptome data. The median has been taken across the prediction of all metabolite abundances. (b) Scores for the first two components in the full PLS model of metabolomic data predicted by transcriptome data.

teins which are often integral membrane proteins and can play a role in cell-cell interactions. It could be that one of VPS33B or VIPAR has some unique function relating to this, or it could be that these genes are variable due to being perturbed by the transfection method (as the second component also separates the knock-downs from the control samples).

2.7 Functional analysis

Once lists of differentially expressed genes/metabolites have been compiled, the objective is to analyse the results further to increase understanding of the biological themes present. In this part of the analysis we make use of the vast amount of information saved in online databases about the cellular function of proteins and groups of proteins functioning in pathways. One question of interest is: are the differentially expressed genes enriched in any particular functional annotation compared to the full set of genes. There are multiple tools available to analyse gene lists in this way, the most commonly used being DAVID (Huang et al., 2008), which annotates the genes in a list with Gene Ontology (GO) terms and then uses Fisher's exact test to test the enrichment compared to a background list. The genes with the top 150 loadings in the full PLS model were input to DAVID, with the Mus Musculus genome used as the background. The top ten enriched Gene Ontology (GO) annotation terms are shown in Table 2.2 (full list given in Appendix B). The annotation terms highlighted are functionally relevant to ARC syndrome with the top annotations relating to cell-cell adhesion and junctions. This confirms the assessment

of the top 20 genes as well as adding weight to the experimental observation that IMCD cells with lower levels of VPS33B/VIPAR no longer form cohesive epithelial layers.

Cell line GO annotations	**P-value**	**Benjamini-Hochberg corrected**
Cell adhesion	0.0006	0.3
Biological adhesion	0.0006	0.2
Cell-cell adhesion	0.0015	0.3
Apical junction complex	0.0068	0.7
Apicolateral plasma membrane	0.0072	0.4
Plasma membrane part	0.0078	0.3
Cell-cell junction	0.0093	0.3
Calcium-independent cell-cell adhesion	0.012	0.8
Basement membrane	0.017	0.4
Tight junction	0.019	0.4

Table 2.2: Table showing enriched GO annotations for lists of significantly differentially expressed genes for knock-down cell lines using the 150 genes with the largest loadings in the first component of a full PLS model.

An alternative approach is to look at pre-defined *sets* of genes (pathways) and ask: do any of these sets contain more of the perturbed genes than would be expected by chance? Statistically, analysing sets of genes rather than individual genes increases power and reduces dimensionality, and biologically, it introduces a more direct link to disease pathology by analysing whole pathways of known function. The approach can be summarised as follows:

- Compute a statistic for each gene e.g. fold change or t-test statistic
- Transformation of gene level statistic - e.g. square or rank
- Compute a statistic for each gene set e.g. mean or median of individual gene statistics
- Assess statistical significance

Ackermann and Strimmer (2009) provide a general overview of the multi-step approach describing the various options at each step and applying many combinations to experimental data. Their study found that the use of simple univariate statistics to summarise genes/sets and permutation to test the significance reliably detected

gene sets with diverse expression signatures and correlation structures. Based on their recommendation, the squared loading for each gene in the first component of the full PLS model was used as the individual gene statistic, the mean was used to obtain a gene set statistic and permutation was used to assess significance (permuting both the set of genes and the samples).

There are many databases available with collections of gene sets representing current knowledge on molecular interactions. A combination of the pathways stored in Kegg (Kanehisa and Goto, 2000; Kanehisa et al., 2014), Biocarta (Nishimura, 2001) and Reactome (Joshi-Tope et al., 2005) databases was used, as these are kept up to date and together cover the majority of known pathways.

The 15 most significantly enriched pathways (p-values calculated based on sampling 10000 gene sets of the same size as the set being assessed) are shown in Table 2.3. As with the most over-represented GO terms, the most perturbed pathways relate to cell-cell adhesion and communication. The top gene set is those involved in cell-cell communication which has as subsets the second and third sets: cell-cell junction organisation and tight junction interactions, as well as adherens junctions interactions (lower down the table). These gene sets are most perturbed on average due to the presence of several genes which have large loadings in the first component of the PLS model, for example, several Claudins (4,7 and 8) and E-cadherin. There are also pathways relating to axon guidance, in particular Semaphorin interactions. At first glance this may seem spurious, as Semaphorins were discovered and are best known for their role in axon guidance in neuronal growth. However, recent research has shown that Semaphorins are widely expressed in many cell types and that through their receptors, Plexins, they can regulate Integrins which are crucial for transmitting signals from the extracellular matrix to the interior of the cell (Yazdani and Terman, 2006; Capparuccia and Tamagnone, 2009; Tamagnone, 2012). Finally, several cancer pathways are implicated, likely because cancerous cells often undergo an epithelial-to-mesenchyme (EMT) transition in which cells become depolarised and junctions are disrupted, hence exhibiting a similar phenotype to the VPS33B/VIPAR/PLOD3 knock-down cells.

2.8 Network analysis

The previous approaches all fundamentally rely on identifying genes or metabolites depending on changes in their levels. If instead, we now view the data as a network

Pathway	p-value	FDR
Cell-cell communication (Reactome)	0.0001	0.02
Cell-cell junction organisation (Reactome)	0.0001	0.02
Tight junction interactions (Reactome)	0.0001	0.02
Cell junction organisation (Reactome)	0.0001	0.02
Adhesion molecules (CAMs) (Kegg)	0.0001	0.02
Leukocyte transendothelial migration (Kegg)	0.0001	0.02
Tight junction (Kegg)	0.0004	0.025
Other semaphorin interactions (Reactome)	0.0006	0.04
Apoptotic cleavage of cell adhesion proteins (Reactome)	0.0008	0.06
Axon guidance (Kegg)	0.0008	0.06
Pathways in cancer (Kegg)	0.0015	0.09
Apoptotic cleavage of cellular proteins (Reactome)	0.0017	0.09
Bladder cancer (Kegg)	0.0025	0.09
Adherens junction interactions (Reactome)	0.003	0.09
Downregulated of MTA-3 in ER-negative Breast Tumours (Biocarta)	0.0037	0.1

Table 2.3: Table showing enriched pathways for knock-down cell lines (Vps33b and Vipar knock-downs analysed together). Gene sets were taken from the Kegg, Biocarta and Reactome online databases.

of protein interactions, we may aim to identify regions of the network whose connectivity or interactions have changed significantly between the knock-downs and the controls. Many network approaches have been applied to omic data (e.g. Kim et al., 2011; Komurov et al., 2010) but no definitive way of identifying networks of differential interaction has been found.

A simple method was developed recently in the study of facioscapulohumeral muscular dystrophy (Banerji et al., 2015), which uses pairwise correlations between transcript expressions as a proxy for interaction strength and then builds an 'interaction distribution' for each gene in the control and knock-down states. Figure 2.16 gives an overview of how the method works.

Two things are required to implement the method: data measuring RNA transcript levels in knock-down/disease samples and control samples, and a database of known protein-protein interactions. In step 1, correlations between each protein pair are computed and used to weight the edge between those two proteins in the interaction network. In step 2, two interaction distributions are defined for each gene, one in the knock-down samples and one in the control samples. The 'interaction probability' is proportional to the absolute value of the correlation, which is then normalised across all the interactions for each gene. Genes which have very different interaction profiles in controls and knock-downs can then be identified by using the Kullback-Leibler divergence. This yields the original full set of nodes (genes)

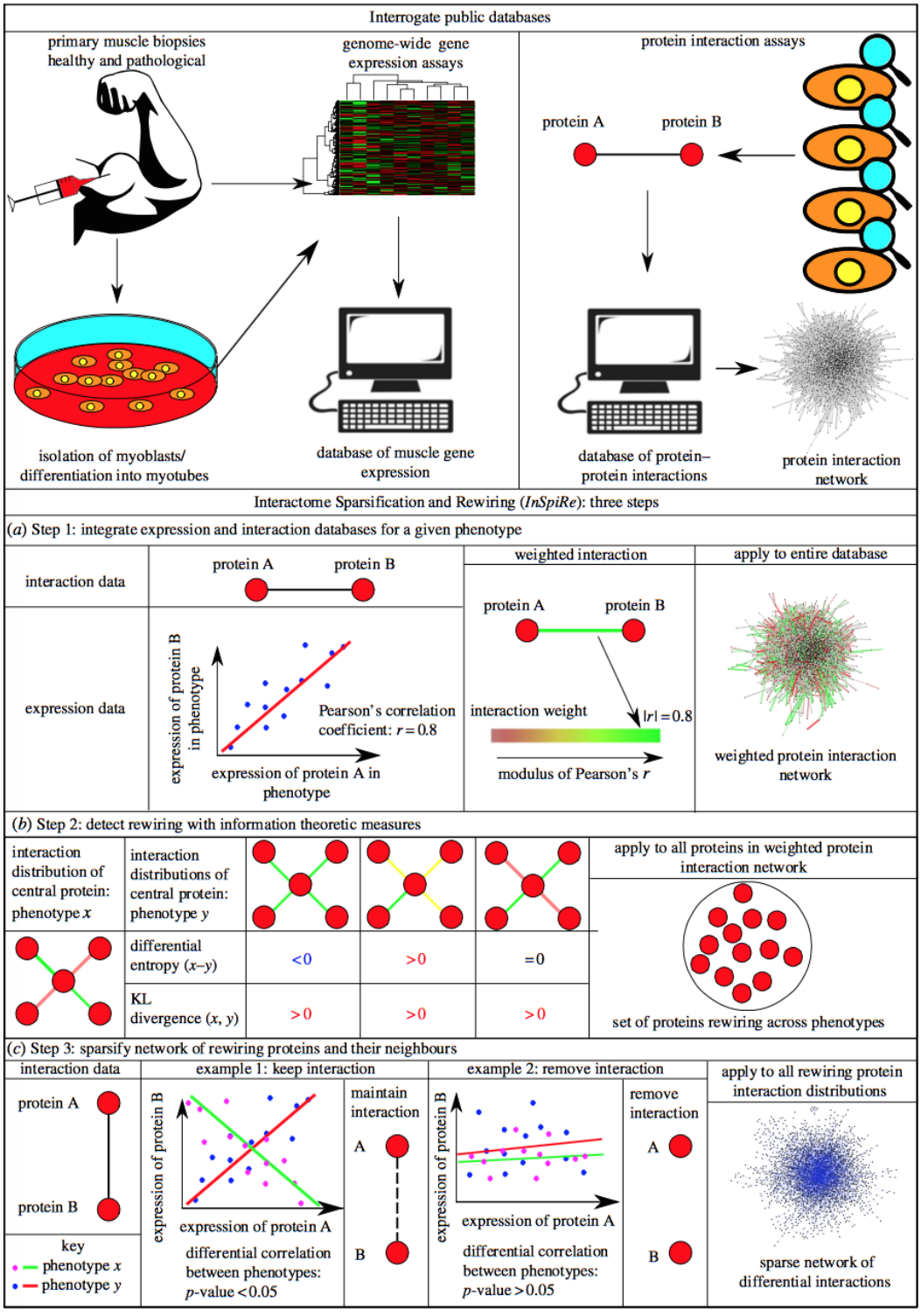


Figure 2.16: Overview of the method developed by Banerji et al. (2015) to detect genes whose interaction profiles have changed significantly between a healthy and diseased state. Figure created by C. Banerji and shared under CC BY 4.0.

with a measure of how much each one's interactions have changed. In step 3, the network is sparsified to retain only interactions in which the pairwise correlation has changed significantly.

A few changes need to be made in order to make the method suitable for the ARC cell line data. For example, as relatively little is known about the functions and interactions of the mutated proteins in ARC syndrome, it seems sensible to not limit investigated interactions to those already published. Instead we can compute all pairwise correlations and use the values as a proxy for interaction strength on a complete graph. Of course, this will identify both direct and indirect interactions.

Another small adaptation to the original version of the method, is to rescale the correlation values by adding one so that they are on a scale $[0,2]$. This is to avoid taking the absolute value of the correlation and hence losing considerable amounts of information. Before computing correlations, the genes were filtered with a non-stringent cut-off of $p < 0.5$ in order to remove genes which have very low probability of being affected by VPS33B/VIPAR knock-down. Any correlations with these genes are likely to be noise rather than true signal.

2.8.1 Network: results

The top 20 genes identified as having very different interaction distributions in knock-downs and controls are listed in Table 2.4, along with a brief description of what is currently known of their function. As this method identifies genes whose *interactions* have changed rather than just those with a different expression level, the genes could be more functionally relevant to the pathogenesis of ARC syndrome in the kidney.

Of immediate interest is the fact that two of the genes are Collagen IVα1 and α4, members of the Collagen IV subfamily which only occur in basement membranes. Experimental work in the Gissen lab has established that one of the functions of VPS33B and VIPAR is to traffic PLOD3 to procollagen carrying vesicles so that it can hydroxylate lysine residues. This is vital for the stability of intermolecular crosslinks as the resultant hydroxylysyl groups are attachment sites for carbohydrates in collagen (Khoshnoodi, 2008). It is therefore striking that these two component of Collagen IV have been identified as interacting differently.

Many of the genes have a role in cellular signalling, for example, Smad2 is a member of the Smad family of signalling effectors activated by Transforming growth factor-β (TGF-β) or Activin Type I receptors. Smad-2 is a receptor Smad, meaning that it is released from the receptor complex to translocate to the nucleus where it acts as a transcriptional repressor (Derynck and Zhang, 2003). Bone morphogenic protein 7 (BMP7) is a member of the TGF-β superfamily of secreted signalling molecules, and like other bone morphogenic proteins it can induce ectopic bone growth. BMP7 induces mesenchyme-to-epithelial transition (MET) in the kidney which is crucial for the formation of the kidney, and it also functions as an endogenous inhibitor of TGF-β-induced EMT and hence is important for kidney homeostasis (Kalluri and Weinberg, 2009).

Pde8a and Akap12 both regulate the second messenger cyclic adenosine monophosphate (cAMP), which mediates several intracellular signals. Pde8a is a phosphodiesterase which hydrolyses cAMP whereas Akap12 associate with Protein kinases A and C (PKA and PKC) to compartmentalise cAMP signals. In the collecting duct of the kidney, cAMP is involved in the regulation of water reabsorption by activating PKA. This results in apical plasma membrane accumulation of Aquaporin 2 (AQP2) which allows water to be reabsorbed from urine (Rieg et al., 2010).

There are genes involved in immune response such as the interferon-γ receptor 1 which binds interferon-γ to activate the Jak-Stat pathway which causes the transcription of various target genes. Traf3 is a Tumour necrosis factor receptor (TNFR)-associated factor (TRAF) protein which are essential components of signalling pathways activated by TNFR or Toll-like receptor (TLR) family members. TRAF3 is a versatile regulator that positively controls type I interferon production, but negatively regulates mitogen-activated protein kinase activation and alternative nuclear factor-κB signalling (Häcker et al., 2011).

Finally, Gdpd5 is a glycerophosphocholine phosphodiesterase (GPC-PDE) which catalyses the degradation of glycerophosphocholine (GPC), an abundant osmoprotective renal medullary osmolyte. GPC levels are altered in most cancers although the mechanism is poorly understood (Moestue et al., 2012; Stewart et al., 2012). It is of note that Gdpd5 hydrolyses GPC to (Choline and) Glycerol-3-phosphate which is one of the metabolites with a larger loading in Figure 2.14 (Gallazzini et al., 2008; Zablocki, 1991).

2.9 Discussion

The first aim of this analysis was to assess how similar the three knock-downs are in terms of their transcriptomes and metabolomes. It has already been shown experimentally that reducing the expression of either VPS33B, VIPAR or PLOD3 produces a similar phenotype; all three cell lines become depolarised and cell adhesion is disrupted (Cullinane et al., 2010).

Unsupervised hierarchical clustering of the microarray data showed that the knock-down samples and the control samples clustered into two distinct groups, and that within these two groups, samples from the same experimental group clustered together. Correlation analysis and PCA score plots showed similar results, indicating that not only are the knock-down samples different from the controls, but that there are very strong similarities between the VPS33B, VIPAR and PLOD3 samples. The metabolomic data is not quite so clear: whilst samples from the same experimental group clustered together, there is a lot more variability in the data and correlations between VPS33B and VIPAR samples were only slightly stronger on average than those between VPS33B and controls cells. However, the PCA and PLS score plots (Figures 2.6 and 2.13) show that the first Principal Component does separate controls from knock-downs demonstrating that these multivariate approaches may be useful to identify relevant metabolites. Overall, the data suggests that knocking down VPS33B, VIPAR and PLOD3 produces a similar and consistent response, fitting with the phenotypic observations and other experimental evidence. The similarity in the transcriptional response adds weight to the hypothesis that the three gene products are involved in the same process.

The second aim was to identify which specific genes and pathways are functionally relevant to the pathology of ARC syndrome to inform future experimental work. Differentially expressed (DE) genes were first identified using a modified t-test for each of the knock-downs separately. Corroborating the findings in the first part, there was a strongly significant overlap in the top 100 DE genes for each knock-down, with $\sim$75% of genes in each list also differentially expressed in one or both other knock-down cells.

To increase statistical power, and to identify genes relevant to the function of VPS33B/VIPAR in ARC syndrome, the data from the VPS33B and VIPAR knock-down cell lines was combined, and the data from the controls also combined giving six biological repeats in each experimental group. The DE genes identified in this

context were highly relevant to both the pathogenesis of ARC syndrome and to the observed cellular phenotype. Many are membrane proteins (e.g. Mal2, Sema5a) and/or involved in processes which maintain cell polarity or junctional integrity (e.g. Cldn7, Rab25).

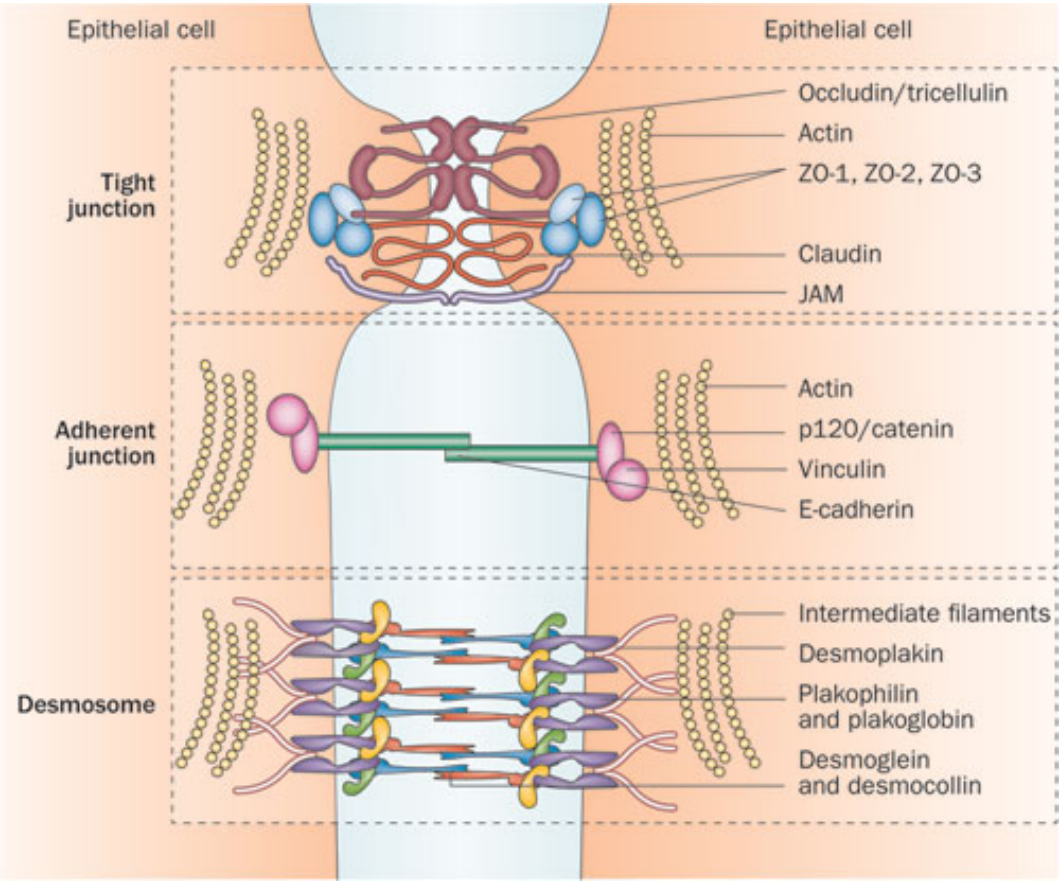


Figure 2.17: Schematic of tight junctions, adherent junctions and desmosomes which are the three main junction complexes connecting adjacent epithelial cells. Reprinted by permission from Macmillan Publishers Ltd, Nature Reviews Gastroenterology and Hepatology, Neunlist et al. (2013).

A closer look at the functions of these genes provides insight into which biological processes are disrupted by loss of VPS33B/VIPAR. Claudin 7 is a member of the Claudin family of proteins which form the major components of tight junctions (see Figure 2.17 and Krause et al. (2008)). Claudin 7 is expressed in the collecting duct of the kidneys and is unusual for a Claudin in that it is located in the basolateral membrane. The function of basolateral claudins is not well known, but recent work has shown that Cldn7 interacts with Integrinβ1 to maintain epithelial cell attachment in lung cancer cells (Lu et al., 2015).

Semaphorin 5a (Sema5a), is a member of the Semaphorin family, first identified

as axon guidance factors but now known to have roles outside the nervous system, including an emerging role in cancer progression (Tamagnone, 2012; Neufeld and Kessler, 2008). Plexins are the main receptors for Semaphorins and the typical outcome of Plexin activation is the inhibition of Integrin-mediated cell-substrate adhesion and cytoskeleton remodelling. The action of Sema5a appears to be context dependent as it has been reported that increased expression of Sema5A and its receptor PlexinB3 significantly correlates with invasion and metastasis in human gastric and pancreatic tumours (Pan et al., 2009) whereas others have found that Sema5A-PlexinB3 signalling inhibits the migration of human glioma cells (Li et al., 2012).

Mal, T-cell differentiation protein 2 (Mal2) is a transmembrane protein which is an essential component of the basolateral-to-apical transcytosis machinery in polarised HepG2 cells (Madrid et al., 2010; de Marco et al., 2002). After basolateral endocytosis of apical cargo, a fraction of MAL2 redistributes from the apical region into peripheral endosomes which concentrates internalised cargo. These endosomes then progressively fuse and move toward the apical surface for cargo delivery (de Marco et al., 2002). Madrid et al. (2010) show that Cell Division Control Protein 42 homolog (Cdc42) controls apical transcytosis and correct lumen formation by regulating Mal2 dynamics.

Rab proteins such as Rab25 are members of the RAS superfamily of small GTPases that are involved in membrane trafficking. Rab25 is a member of the Rab11 family, which regulate the recycling of internalised membrane molecules. Rab25 is only expressed in epithelial cells and studies in polarised MDCK cells have indicated that Rab25 regulates transcytosis of cargo in both directions and therefore is an important regulator of cell surface composition (Tzaban et al., 2009; Casanova et al., 1999; Wang et al., 2000). Rab25 has also been shown to regulate the recycling of Integrin, a large family of transmembrane receptors which enable cells to bind and respond to other cells and the extracellular matrix (ECM) (Dozynkiewicz et al., 2012; Agarwal et al., 2009).

Fermitin Family Member 1 (Fermt1) is also involved in Integrin signalling and hence in the linkage of the cytoskeleton to the ECM. Loss-of-function mutations cause Kindler Syndrome which is an autosomal recessive disorder characterised by skin atrophy and blistering. Studies on Kindler Syndrome skin show an altered distribution of several basement membrane proteins, including types IV, VII, and XVII collagens, and reduced levels of active Integrin $\beta 1$ (Mas-Vidal et al., 2010;

Lai-Cheong et al., 2009).

Results from the multivariate methods such as PLS overlapped considerably with those found using the univariate methods, however, a number of extra genes were identified. E-cadherin was identified as a gene with a large loading, a member of the Cadherin family of transmembrane proteins which form adherens junctions (see Figure 2.17). Cadherin 1 or E-Cadherin is a classical Cadherin, and its loss or mutation is associated with multiple types of cancer. This is believed to be because loss of its function reduces the integrity of cell-cell junctions and increases proliferation (Paredes et al., 2012).

Metabolites highlighted having large loadings in the PLS model were also of interest. Proline is an essential component of collagen and hence required for the proper functioning of bones and tendons (Grant and Prockop, 1972). Proline has increased abundance in the knock-down cell lines, fitting with experimental observations from the lab of increased Collagen deposits in ARC syndrome models. Cystathionine is an intermediate in the formation of cysteine, and itself is produced from serine (both of which also have large loadings in the PLS model, all three with decreased expression). The production of cysteine is important given its ability to form disulfide bonds which cross-link polypeptide chains, particularly in the ECM (Berg et al., 2002, Chap. 3). The presence of all three molecules indicates that this pathway may be disrupted in cells lacking in VPS33B/VIPAR. Serine, is concentrated in cell membranes and has a central role in cell proliferation (Berg et al., 2002, Chap. 26).

The prevalence of membrane and junction proteins is maintained when the top 150 DE genes are interrogated for over represented Gene Ontology (GO) annotation. Table 2.2 lists the annotations which are most over-represented compared to those expected for the mouse transcriptome. All terms relate to either cell-cell adhesion or cell membranes, confirming that the transcriptional signature of cell lines with reduced VPS33B/VIPAR fits with experimental observations.

The pathways highlighted as perturbed by the GSE analysis are also strongly related to cell adhesion with ‘Cell-Cell Communication’ and its subsets connected to cell junctions the most significant results. However, other pathways relating to Axon Guidance and Semaphorin interactions are also identified. As already discussed, Semaphorins have recently been described as a family of widely expressed proteins, which activate Plexin and Neuropilin receptors to transduce signals. These

molecules carry out diverse roles including cardiovascular development and growth (Neufeld et al., 2012), tumour progression (Capparuccia and Tamagnone, 2009) and immune cell regulation (Kumanogoh et al., 2002). Pertinently to ARC syndrome, plexin-mediated signalling has been implicated in the inhibition of Integrin-mediated cellular adhesion and cytoskeletal remodelling (Serini et al., 2003; Takamatsu and Kumanogoh, 2012).

In the network approach, relevant genes were identified by computing the difference between their interaction distribution in control samples and knock-down samples. The genes identified in this way were also relevant to ARC syndrome but provided different information on the changes in the transcriptome. For example, the most obvious finding is two Collagen IV α chains, a major component of the basement ECM. This directly fits with current experimental work from the Gissen lab which has found that PLOD3, which hydroxylates lysyl residues in Collagen, requires VPS33B and VIPAR for its trafficking to vesicles containing procollagen.

A variety of signalling molecules were also pinpointed, such as Smad2 and Bmp7, both involved in the TGF-β pathway, which transcriptionally regulates the expression of a wide range of genes involved in many cellular processes. Two proteins involved in the regulation of cAMP, Pde8a and Akap12, were also highlighted. cAMP is a second messenger which mediates many different cell responses.

The overall picture from the analysis is a strong signature of cell-cell adhesion and junctions being disrupted in VPS33B and VIPAR knock-downs. The proteins identified are sometimes directly involved in adhesion, such as Claudins which form tight junctions, or E-cadherin which form part of adherens junctions. Other proteins are involved in adhesion in a regulatory sense, such as Semaphorin 5a which has been shown to regulate cell motility and invasiveness in a context dependent manner (Pan et al., 2009; Li et al., 2012), or Mal2 which regulates basolateral-to-apical transcytosis in polarised cells (Madrid et al., 2010). Functions and genes relating to Integrin activation and regulation are repeatedly identified, so further investigation of its expression, cellular location and function may be of interest. Integrin has already been identified as an interactor of VIPAR in a yeast two-hybrid screen (unpublished data from lab) adding more weight to the hypothesis that it may be functionally relevant to the pathogenesis of ARC syndrome.

2.10 Motivation for the next two chapters

2.10.1 Methodology for Doubly-intractable distributions

In this chapter, transcriptomic and metabolomic data has been analysed to shed light on the pathogenesis of ARC syndrome and to suggest possible routes for future experimental work. Standard statistical techniques have been used, largely because these are the accepted and trusted way of extracting information from Biological data. However, clearly the identification of the differentially expressed genes could be carried out in a more sophisticated way. For example, differentially expressed genes have been identified using only measurements of their abundances. It would be preferable to incorporate the vast wealth of pathway information in this process, as these dependencies will enable the detection of smaller changes. This type of analysis has already been suggested, for example the work of Wei and Li (2007) uses a Markov Random Field (MRF) to model the dependencies between interacting genes.

MRF models are probabilistic networks in which each node is a random variable. Each node has a defined 'neighbourhood' or set of nodes on whose values its conditional probability depends. Dependencies can propagate a long way through the network via these short range connections. In the work of Wei and Li (2007), each node is a protein, and its neighbourhood is the proteins with which it interacts. Their work introduces a latent unobserved variable, $\mathbf{x}$, whose dimension equals the total number of genes, and which takes values $x_i = 1$ if gene i is differentially expressed and $x_i = 0$ otherwise. y_i is the observed mRNA level for gene i. A Markov Random Field is used to model $p(\mathbf{x};\theta)$ such that if a gene is differentially expressed, other genes in its neighbourhood are also more likely to be differentially expressed. The conditional likelihood, $p(\mathbf{y}|\mathbf{x},\theta)$, is then defined as a product of Gamma distributions, one for each gene, with parameterisation dependent on being differentially or equally expressed genes.

This approach incorporates more of the available information into the detection of differentially expressed genes. However, it comes with the drawback that inference for Markov Random Fields, frequentist or Bayesian, is extremely difficult due to the presence of an intractable normalising term which cannot be computed. Wei and Li (2007) use a pseudo-likelihood which is a simple approximation of the MRF; this removes the computational difficulty but also models the long-range dependencies poorly.

In Chapter 3, methodology is developed to enable Bayesian inference for MRFs and other models with the same drawback of an intractable normalising term, called 'doubly-intractable' distributions. It emerges that these types of models occur across a wide range of disciplines. The methodology is general and applies to all models of this form, which means it is applicable in Systems Biology and beyond.

2.10.2 Methodology for unbiased ABC

At this stage in the molecular study of ARC syndrome, not enough is known of the interactions and dynamics of the proteins involved to draw up a model of the processes involved. However, once further experimental work and analysis has been carried out, it would be desirable to define a model describing the molecular interactions involved. This would allow simulation and investigation of the system without having to carry out costly experiments. It is crucial to use stochastic models so that all the inherent sources of variability in biological systems and experiments are modelled fully. Use of the more standard deterministic models results in 'apparent unpredictability' (Wilkinson, 2009) and unexplained heterogeneity across biological replicates.

However, the price of using more realistic stochastic models is the increased computational complexity involved in fitting them to data. One bonus is that algorithms, such as the Gillespie algorithm, are available to simulate from these models (Gillespie, 1977). However, as the likelihood does not have a tractable form, standard statistical methods cannot easily be applied. When implementing a Bayesian analysis, Approximate Bayesian computation (ABC) is often used as a method to draw samples from a distribution which is 'close' to the true Bayesian posterior. No likelihood need be computed, but data must be simulated from the model, and hence this method can be applied to stochastic models of interacting molecules.

The ABC method has permitted inference for a host of complex models for which statistical analysis was previously difficult. However, as samples are not drawn from the true posterior distribution, a bias is introduced into Bayesian inference and this bias is not well characterised. In Chapter 4 methodology is developed to allow unbiased estimation for models for which no likelihood can be computed but from which data can be simulated, at the expense of increased computation. The method utilises Monte Carlo ABC estimates and combines them in such a way that the overall estimator is unbiased. This allows unbiased estimation for stochastic

models and hence makes a strong contribution to both Systems Biology and other areas where such models are used.

Gene	K-L divergence	P-value estimate	Function
Epb4.1l3	0.71	0.12	May be involved in tethering the F-actin skeleton to membrane, tumour suppressor activity demonstrated in variety of cancers (Cavanna et al., 2007; Dafou et al., 2010)
Smad2	0.71	0.009	Receptor-regulated SMAD (R-SMAD), an intracellular signal transducer and transcriptional modulator activated by Activin type 1 receptor kinases (Derynck and Zhang, 2003)
Tbc1d9	0.70	0.004	GTPase-activating protein for Rab family protein(s), possibly not expressed in kidney (Nakamura et al., 2015)
Cdcp1	0.69	0.004	Transmembrane protein, involved in cell adhesion and ECM association via activation of Src-family kinases (Liu et al., 2011)
Ifngr1	0.69	0.004	Receptor for interferon gamma, a cytokine that is critical for innate and adaptive immunity against infections (Farrar and Schreiber, 1993)
Pde8a	0.68	0.021	Hydrolyzes the second messenger cAMP, a key regulator of many important physiological processes (Fisher et al., 1998; Patrucco et al., 2010)
Col4a1	0.68	0.017	On component of Type IV collagen, major structural component of glomerular basement membranes
Akap12	0.68	0.024	A-kinase anchoring protein, binds to cAMP-dependent protein kinase (PKA) to direct the kinase to discrete intracellular locations (Colledge and Scott, 1999)
Gdpd5	0.68	0.004	A GPC-PDE which contributes to osmoregulation of GPC in the renal medulla (Gallazzini et al., 2008)
Syt14	0.68	0.006	May be involved in the trafficking and exocytosis of secretory vesicles in non-neuronal tissues (Pang and Südhof, 2010)
Col4a4	0.68	0.006	On component of Type IV collagen, major structural component of glomerular basement membranes
Zfp296	0.67	0.003	
Anks1	0.67	0.009	Regulator of different signaling pathways via interactions with Epha2 and Arap3 (Mercurio et al., 2013)
Bmp7	0.67	0.004	Bone morphogenic protein, which induces cartilage and bone formation. Also important in kidney homeostasis by inhibiting EMT (Zeisberg et al., 2003)
Tiparp	0.67	0.006	Regulator of Ahr, which may have important roles in functions such as growth, differentiation and immunity (McMillan and Bradfield, 2007; MacPherson et al., 2014)
Traf3	0.67	0.09	TNF receptor associated factor (TRAF) protein, mediates Cd40 signals important for immune response (Häcker et al., 2011)
Dock5	0.67	0.03	Guanine nucleotide exchange factor (GEF) for small GTPases Rho and Rac, exchanges bound GDP for free GTP (Vives et al., 2011)
Ppbp	0.67	0.005	Involved in neuro-protection, possibly linked to CREB activation (Yang and Alkayed, 2009)
Ahr	0.66	0.004	Ligand activated transcription factor which regulates xenobiotic-metabolising enzymes such as cytochrome P450 (MacPherson et al., 2014)
B4galt6	0.66	0.005	Type II membrane-bound glycoproteins which have specificity for the donor substrate UDP-galactose (Tokuda et al., 2013)

Table 2.4: Top 20 genes and their functions as identified by Kullback-Leibler divergence between their interaction distribution in controls and knock-downs.

Chapter 3

Roulette for Doubly-intractable distributions

3.1 Introduction

At the end of the previous Chapter, an approach to identifying differentially expressed genes was introduced which incorporated knowledge of protein interactions using a probabilistic undirected graphical model. This would allow more subtle changes to be detected as dependencies in gene expression are fully accounted for. It was also stated that standard inference techniques cannot be used for these types of models due to the presence of an intractable normalising term. In this chapter methodology is developed to enable Bayesian inference for these types of models which are used in Systems Biology and many other research areas.

The term *doubly-intractable* has been used to described posterior distributions associated with these likelihoods, and this was first coined by Murray et al. (2006). To illustrate what constitutes a doubly-intractable posterior, take some data $\mathbf{y} \in \mathcal{Y}$ used to make posterior inferences about the variables $\theta \in \Theta$ that define a statistical model. A prior distribution defined by a density $\pi(\theta)$ with respect to Lebesgue measure $\mathrm{d}\theta$ is adopted and the data density is given by $p(\mathbf{y}|\theta) = f(\mathbf{y};\theta)/\mathcal{Z}(\theta)$, where $f(\mathbf{y};\theta)$ is an unnormalised function of the data and parameters, and $\mathcal{Z}(\theta) = \int f(\mathbf{x};\theta)\mathrm{d}\mathbf{x}$ is the likelihood normalising term which *cannot be computed*. The posterior density follows in the usual form as

$$\pi(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta) \times \pi(\theta)}{p(\mathbf{y})} = \frac{f(\mathbf{y};\theta)}{\mathcal{Z}(\theta)} \times \pi(\theta) \times \frac{1}{p(\mathbf{y})}, \tag{3.1}$$

where $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)\pi(\theta)d\theta$. 'Doubly-intractable' refers to the fact that not only is $p(\mathbf{y})$ intractable (this is common in Bayesian inference and does not generally present a problem for inference), but $\mathcal{Z}(\theta)$ is also intractable.

To construct a Markov chain with invariant distribution $\pi(\theta|\mathbf{y})$, the standard recourse would be to the Metropolis-Hastings algorithm; a transition kernel is constructed by designing a proposal distribution $q(\theta'|\theta)$ and accepting the proposed parameter value with probability

$$\alpha(\theta', \theta) = \min\left\{1, \frac{\pi(\theta'|\mathbf{y})q(\theta|\theta')}{\pi(\theta|\mathbf{y})q(\theta'|\theta)}\right\} = \min\left\{1, \frac{f(\mathbf{y};\theta')\pi(\theta')q(\theta'|\theta)}{f(\mathbf{y};\theta)\pi(\theta)q(\theta'|\theta)} \times \frac{\mathcal{Z}(\theta)}{\mathcal{Z}(\theta')}\right\}. \tag{3.2}$$

Clearly a problem arises when the value of the normalising term for the data density, $\mathcal{Z}(\theta)$, cannot be obtained either due to it being non-analytic or uncomputable with finite computational resource. This situation forms a major challenge to methodology for computational statistics currently (e.g. Møller et al., 2006; Besag and Moran, 1975; Besag, 1974; Green and Richardson, 2002; Møller and Waagepetersen, 2003).

This Chapter is organised as follows. In Section 3.2 examples of doubly-intractable distributions are described along with current inference approaches. These encompass both approximate and exact methods which have been developed in the Statistics, Epidemiology and Image analysis literature. In Section 3.3 a novel approach based on Pseudo-marginal MCMC (Beaumont, 2003; Andrieu and Roberts, 2009) is suggested in which an unbiased estimate of the intractable target distribution is used in an MCMC scheme to sample from the exact posterior distribution. In Sections 3.4 and 3.5 we describe how to realise such unbiased estimates of a likelihood with an intractable normalising term. This is achieved by writing the likelihood as an infinite series in which each term can be estimated unbiasedly. Then Russian Roulette techniques are used to truncate the series such that only a finite number of terms need be estimated whilst maintaining the unbiasedness of the overall esti-

mate. Section 3.6 contains experimental results for posterior inference over doubly-intractable distributions: Ising models and the Fisher-Bingham distribution. Section 3.7 contains a discussion of the method and suggests areas for further work.

3.2 Inference methods for doubly-intractable distributions

3.2.1 Approximate Bayesian inference

Many models describing data with complex dependency structures are doubly-intractable. Examples which have received attention in the Statistics literature include:

1. undirected or directed graphical models to incorporate prior knowledge of genetic dependencies into the detection of differentially expressed genes (Wei and Li, 2007; Li and Li, 2008; Wei and Pan, 2008)

2. the Ising model (Ising, 1925). Originally formulated in the Physics literature as a simple model for interacting magnetic spins on a lattice. Spins are binary random variables which interact with neighbouring spins.

3. the Potts model and autologistic models. Generalisations to the Ising model in which spins can take more than two values and more complex dependencies are introduced. Used, *inter alia*, in image analysis (Besag, 1986; Hughes et al., 2011), disease mapping (e.g. Green and Richardson, 2002) and to model cellular adhesion (Turner and Sherratt, 2002).

4. Spatial point processes. Used to model point pattern data for example ecological data (e.g. Silvertown, 2001; Møller and Waagepetersen, 2003) or epidemiological data (e.g. Diggle, 1990).

5. Exponential Random Graph (ERG) models. Used to model social network structure in terms of local graph statistics such as the number of triangles (e.g. Goodreau et al., 2009).

6. Massive Gaussian Markov random field (GMRF) models. Used in image analysis and spatial statistics, amongst others (e.g. Rue and Held, 2005).

Standard Bayesian inference techniques such as drawing samples from the posterior using MCMC cannot be used due to the intractability of the likelihood normalising term, and hence a number of approximate inference methods have been developed. A common approach when the full likelihood cannot be computed is to use a pseudo-likelihood (Besag, 1974; Besag and Moran, 1975), in which an approximation to the true likelihood is formed using the product of the conditional probabilities for each variable. This can normally be computed efficiently and can therefore replace the full likelihood in an otherwise standard inference strategy to sample from the posterior (e.g. Heikkinen and Hogmander, 1994; Zhou and Schmidler, 2009). This approach scales well with the size of the data and can give a reasonable approximation to the true posterior, but inferences may be significantly biased as long range interactions are not taken into account (this has been shown to be the case for ERG models (Duijn et al., 2009), hidden Markov random fields (Friel et al., 2009) and autologistic models (Friel and Pettitt, 2004)). Methods based on composite likelihoods have also been used for inference in massive scale GMRF models, in which an approximation to the likelihood is based on the joint density of spatially adjacent blocks (Eidsvik and Shaby, 2014). This has the advantage that the separate parts of the likelihood can be computed more efficiently and in parallel.

Another pragmatic approach is that of Green and Richardson (2002), in which the interaction parameter in the Potts model is discretised to a grid of closely spaced points and then a prior is set over these values. Estimates of the normalising term are then pre-computed using thermodynamic integration (as described by Gelman and Meng (1998)) so that no expensive computation is required during the MCMC run. This allowed inference to be carried out over a model for which it would not otherwise have been possible. However it is not clear what impact this discretisation and use of approximate normalising terms has on parameter inference and it seems preferable, if possible, to retain the continuous nature of the variable and to not use approximations unless justified.

Approximate Bayesian Computation (ABC) (Marin et al., 2012; Tavaré et al., 1997; Beaumont et al., 2002) has already been briefly described in the Introduction and at the end of the previous chapter, and as it is a technique which does not require the computation of the likelihood, it can also be used for doubly intractable models. The types of models for which ABC was originally developed are implicit; meaning data can be simulated from the likelihood but the likelihood cannot be written down.

In its simplest form it proceeds by proposing an approximate sample from the joint distribution, $p(\mathbf{y}, \theta)$, by first proposing θ' from the prior and then generating a dataset from the model likelihood conditional on θ'. This data set is compared to the observed data and the proposed parameter value accepted if the generated data is 'similar' enough to the observed data. An obvious drawback to the method is that it does not sample from the exact posterior, although it has been shown to produce comparable results to other approximate methods and recent advances mean that it can be scaled up to very large datasets (Moores et al., 2014; Grelaud et al., 2009; Everitt, 2012).

Several approximate but consistent algorithms have been developed based on Monte Carlo approximations within MCMC methods. For example, an approach was developed by Atchadé et al. (2013) in which a sequence of transition kernels are constructed using a consistent estimate of $\mathcal{Z}(\theta)$ from the Wang-Landau algorithm (Wang and Landau, 2001). The estimates of the normalising term converge to the true value as the number of iterations increases and the overall algorithm gives a consistent approximation to the posterior. Bayesian Stochastic Approximation Monte Carlo (Jin and Liang, 2014) works in a similar fashion, sampling from a series of approximations to the posterior using the stochastic approximation Monte Carlo algorithm (Liang et al., 2007) which is based on the Wang-Landau algorithm. These algorithms avoid the need to sample from the model likelihood but in practice suffer from the curse of dimensionality as the quality of the importance sampling estimate depends on the number and location of the grid points. These points need to grow exponentially with the dimension of the space limiting the applicability of this methodology. They also require a significant amount of tuning to attain good approximations to the normalising term, and hence ensure convergence is achieved.

Alternative methodologies have avoided sampling altogether and instead used deterministic approximations to the posterior distribution. This is particularly the case for GMRF models which often have complex parameter dependencies and are very large in scale, rendering MCMC difficult to apply. INLA (integrated nested Laplace approximations) (Rue et al., 2009) was designed to analyse latent Gaussian models and has been applied to massive GMRFs in diverse areas such as spatio-temporal disease mapping (Schrödle and Held, 2011) and point processes describing the locations of muskoxen (Illian et al., 2012). By using Laplace approximations to the posterior and an efficient programming implementation, fast Bayesian inference can

be carried out for large models. However, this benefit also constitutes a drawback in that users must rely on standard software, and therefore model extensions which could be tested simply when using an MCMC approach are not easy to handle. Further, it is of course necessary to ensure that the assumptions inherent in the method apply so that the approximations used are accurate. It should also be noted that the work of Taylor and Diggle (2014) found that in the case of spatial prediction for log-Gaussian Cox processes, an MCMC method using the Metropolis-adjusted Langevin Algorithm (MALA) algorithm gave comparable results in terms of predictive accuracy and was actually slightly more efficient than the INLA method.

3.2.2 Exact MCMC methods

As well as approximate inference methods, a small number of exact algorithms have been developed to sample from doubly-intractable posteriors. These are described below as well as advice as to when these algorithms can be used. 'Exact' in this context means that the Markov chain has the true posterior distribution as its invariant distribution.

3.2.2.1 Introducing auxiliary variables

An exact sampling methodology for doubly-intractable distributions is proposed in Walker (2011), which uses a similar approach to those described in Adams et al. (2009) and Section 9 of Beskos et al. (2006). A Reversible-Jump MCMC (RJMCMC) sampling scheme is developed that cleverly gets around the intractable nature of the normalising term. Consider the univariate distribution $p(y|\theta) = f(y;\theta)/\mathcal{Z}(\theta)$ where N independent and identically distributed (i.i.d.) observations, y_i are available. In its most general form, it is required that y belongs to some bounded interval $[a,b]$, and that there exists a constant $M < +\infty$ such that $f(y;\theta) < M$ for all θ and y (it is assumed that $[a,b] = [0,1]$, and $M = 1$ in the following exposition). The method introduces auxiliary variables $\nu \in (0,\infty)$, $k \in \{0,1,\ldots\}$, $\{s\}^{(k)} = (s_1,\ldots,s_k)$, to form the joint density

$$f(\nu,k,\{s\}^{(k)},\mathbf{y}|\theta) \propto \frac{\exp(-\nu)\nu^{k+N-1}}{k!}\prod_{j=1}^{k}\left(1-f(s_j;\theta)\right)\mathbb{1}(0<s_j<1)\prod_{i=1}^{N}f(y_i;\theta).$$

Integrating out ν and $s^{(k)}$ and summing over all k returns the data distribution $\prod_{i=1}^{N} p(y_i|\theta)$. An RJMCMC scheme is proposed to sample from the joint density $f(\nu,k,\{s\}^{(k)},\mathbf{y}|\theta)$ and this successfully gets around the intractable nature of

the normalising term. The scheme has been used to sample from the posterior of a Bingham distribution (Walker, 2014).

However the methodology has some limitations to its generality. Firstly, the unnormalised density function must be strictly bounded from above to ensure the positivity of the terms in the first product. This obviously limits the generality of the methodology to the class of strictly bounded functions, however this is not overly restrictive as many functional forms for $f(y_i;\theta)$ are bounded e.g. when there is finite support, or when $f(y_i;\theta)$ takes an exponential form with strictly negative argument. Even if the function to be sampled is bounded, finding bounds that are tight is extremely difficult and the choice of the bound directly impacts the efficiency of the sampling scheme constructed, see e.g. El Ghaoui and Gueye (2008) for bounds on binary lattice models. Ideally we would wish to relax the requirement for the data, $\mathbf{y}$, to belong to a bounded interval, but if we integrate with respect to each s_j over an unbounded interval then we can no longer return $1-\mathcal{Z}(\theta)$ and the sum over k will therefore no longer define a convergent geometric series equaling $\mathcal{Z}(\theta)$. This last requirement particularly restricts the generality and further use of this specific sampling method for intractable distributions.

3.2.3 Valid Metropolis-Hastings-type transition kernels

An ingenious MCMC solution to the doubly-intractable problem was proposed by Møller et al. (2006) in which the posterior state space is extended as follows

$$\pi(\theta,\mathbf{x}|\mathbf{y}) \propto p(\mathbf{x}|\theta,\mathbf{y})\pi(\theta)\frac{f(\mathbf{y};\theta)}{\mathcal{Z}(\theta)}. \tag{3.3}$$

This extended distribution retains the posterior as a marginal. The method proceeds by taking the proposal for $\mathbf{x},\theta$ to be $q(\mathbf{x}',\theta'|\mathbf{x},\theta) = \frac{f(\mathbf{x};\theta')}{\mathcal{Z}(\theta')}q(\theta'|\theta)$ so that at each iteration the intractable normalising terms cancel in the Metropolis-Hastings acceptance ratio. This algorithm has been named the Single Auxiliary Variable Method (SAVM). A drawback of the algorithm is the need to choose the marginal for $\mathbf{x}$, $p(\mathbf{x}|\theta,\mathbf{y})$, particularly as the authors suggest that ideally this distribution would approximate the likelihood, thereby reintroducing the intractable normalising term.

Murray et al. (2006), simplified and extended the algorithm to the Exchange algo-

rithm, and in the process removed this difficulty, by defining a joint distribution as follows

$$p(\mathbf{x}, \mathbf{y}, \theta, \theta') \propto \frac{f(\mathbf{y}; \theta)}{\mathcal{Z}(\theta)} \pi(\theta) q(\theta' | \theta) \frac{f(\mathbf{x}; \theta')}{\mathcal{Z}(\theta')},$$

which also has the posterior as a marginal. At each iteration, MCMC proceeds by first Gibbs sampling θ' and $\mathbf{x}$, and then proposing to swap the values of θ and θ' using Metropolis-Hastings. Again, the intractable normalising terms cancel in the acceptance ratio.

Both of these algorithms use only valid MCMC moves and therefore target the exact posterior. However they both require the capability to sample from the likelihood using a method such as perfect sampling (Propp and Wilson, 1996; Kendall, 2005). This can be considered a restriction to the widespread applicability of this class of methods as for many models this is not possible e.g. ERG models in social networks. Even when perfect sampling is possible, e.g. for the Ising and Potts models, it becomes prohibitively slow as the size of the model increases. Attempts have been made to relax the requirement to perfectly sample by instead using an auxiliary Markov chain to sample approximately from the model at each iteration (Caimo and Friel, 2011; Liang, 2010; Everitt, 2012; Alquier et al., 2014). Theoretical justification for this approach is given in Everitt (2012), where it is shown that as the number of auxiliary MCMC iterations is increased, the invariant distribution of the approximate chain becomes 'closer' to the true posterior distribution, under certain conditions. The paper by Alquier et al. (2014) analyses and suggests multiple approximate MCMC algorithms for doubly-intractable distributions and then applies results from Markov chain theory to bound the total variation distance between the approximate chains and a hypothetical exact chain. These types of approximate algorithms were in use due to their computational feasibility and so it is pleasing to see some theoretical justification for their use emerging in the Statistics literature.

3.3 An alternative approach using Pseudo-marginal MCMC

As has been seen, there are many approximate methods for sampling from doubly-intractable posteriors. There are also exact methods available, but these can only be applied when it is possible to perfectly sample from the data model and hence can only be applied to small datasets and certain models. Now we would like to approach the question of whether it is possible to relax this requirement and develop methodology for 'exact' MCMC sampling of the posterior when perfect sampling is not possible. To do so an approach is developed based on the Pseudo-marginal methodology (Beaumont, 2003; Andrieu and Roberts, 2009; Doucet et al., 2012), and hence the algorithm is now described.

The Pseudo-marginal class of methods is particularly appealing in that they have the least number of restrictions placed upon them and provide one of the most general MCMC methods for intractable distributions. They are sometimes referred to as Exact-approximate methods, based on the property that the invariant distribution of the Markov chain produced is the exact target distribution despite the use of an approximation in the Metropolis-Hastings acceptance probability. To use the scheme, an unbiased and positive estimate of the target density is substituted for the true density giving an acceptance probability of the form

$$\alpha(\theta',\theta) = \min\left\{1,\ \frac{\hat{\pi}(\theta'|\mathbf{y})}{\hat{\pi}(\theta|\mathbf{y})} \times \frac{q(\theta|\theta')}{q(\theta'|\theta)}\right\} = \min\left\{1,\ \frac{\hat{p}(\mathbf{y}|\theta')\pi(\theta')}{\hat{p}(\mathbf{y}|\theta)\pi(\theta)} \times \frac{q(\theta|\theta')}{q(\theta'|\theta)}\right\}, \tag{3.4}$$

where the estimate at each proposal is propagated forward as described in Beaumont (2003); Andrieu and Roberts (2009). For the case of doubly-intractable distributions, assuming the prior is tractable, this equates to a requirement for an unbiased estimate of the likelihood as seen on the right in (3.4) above. The remarkable feature of this scheme is that the corresponding transition kernel has an invariant distribution with θ-marginal given precisely by the desired posterior distribution, $\pi(\theta|\mathbf{y})$. To see this, denote all the random variables generated in the construction of the likelihood estimator by the vector $\mathbf{u}$, and its density $p(\mathbf{u})$. These random variables are, for example, those used when generating an importance sampling estimate of the target. The estimator of the likelihood is denoted $\hat{p}_N(\mathbf{y}|\theta,u)$ with N symbol-

ising, for example, the number of Monte Carlo samples used in the estimate. The estimator of the likelihood must be unbiased i.e.

$$\int \hat{p}_N(\mathbf{y}|\theta,\mathbf{u})p(\mathbf{u})\mathrm{d}\mathbf{u} = p(\mathbf{y}|\theta). \tag{3.5}$$

A joint density for θ and $\mathbf{u}$ is now defined which returns the posterior distribution after integrating over $\mathbf{u}$

$$\begin{aligned}\pi_N(\theta,\mathbf{u}|\mathbf{y}) &\propto \hat{p}_N(\mathbf{y}|\theta,\mathbf{u})\pi(\theta)p(\mathbf{u}) \\ &= \frac{\hat{p}_N(\mathbf{y}|\theta,\mathbf{u})\pi(\theta)p(\mathbf{u})}{p(\mathbf{y})}.\end{aligned}$$

It is simple to show using Equation (3.5) that $\pi_N(\theta,\mathbf{u}|\mathbf{y})$ integrates to 1 and has the desired marginal distribution for $\theta|\mathbf{y}$. Now consider sampling from $\pi_N(\theta,\mathbf{u}|\mathbf{y})$ using the Metropolis-Hastings algorithm, with the proposal distribution for $\mathbf{u}'$ being $p(\mathbf{u}')$. In this case the densities for $\mathbf{u}$ and $\mathbf{u}'$ cancel and we are using the acceptance probability in (3.4). Hence, this algorithm samples from $\pi_N(\theta,\mathbf{u}|\mathbf{y})$ and the samples of θ obtained are distributed according to the posterior.

This is a result that was highlighted in the statistical genetics literature (Beaumont, 2003) then popularised and formally analysed in Andrieu and Roberts (2009), with important developments such as Particle MCMC (Doucet et al., 2012) proving to be extremely powerful and useful in a large class of statistical models. Due to its wide applicability, the Pseudo-marginal algorithm has been the subject of several recent papers in the statistical literature, increasing understanding of the methodology. These have covered how to select the number of samples in the unbiased estimate to minimise the computational time (Doucet et al., 2012), optimal variance and acceptance rates to maximise efficiency of the chain (Sherlock and Thiery, 2014) and results to order two different pseudo-marginal implementations in terms of the acceptance probability and asymptotic variance (Andrieu and Vihola, 2014). It is interesting to note that the problem of Exact-Approximate inference was first considered in the Quantum Chromodynamics literature almost thirty years ago, see for example Kennedy and Kuti (1985); Bhanot and Kennedy (1985); Bakeyev and Forcrand (2001); Lin et al. (2000); Joo et al. (2003).

Note that the approach of Møller et al. (2006) is a Pseudo-marginal type algorithm. In order to sample from a doubly-intractable posterior using the Pseudo-marginal algorithm, an unbiased estimate of $1/\mathcal{Z}(\theta)$ is required. This can be achieved by introducing an auxiliary density as in Equation (3.3), $p(x|y,\theta)$, and rewriting as follows

$$\begin{aligned}\frac{1}{\mathcal{Z}(\theta)} &= \frac{1}{\mathcal{Z}(\theta)}\int p(x|y,\theta)\,\mathrm{d}x = \int \frac{p(x|y,\theta)}{f(x;\theta)}\frac{f(x;\theta)}{\mathcal{Z}(\theta)}\,\mathrm{d}x \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\frac{p(x_i|y,\theta)}{f(x_i;\theta)} \qquad\qquad x_i \sim p(\cdot|\theta).\end{aligned}$$

Setting $N = 1$, and drawing one sample from the likelihood at each iteration, this is then equivalent to the Møller et al. (2006) algorithm. The algorithm remains valid if $N > 1$, and this should improve the mixing of the resulting Markov chain (although not necessarily the efficiency).

Note further, that in the Exchange algorithm, the intractable ratio $\mathcal{Z}(\theta)/\mathcal{Z}(\theta')$, is replaced by $f(x;\theta)/f(x;\theta')$ with x simulated from the likelihood, which can also be viewed as a one-sample importance estimate. However, as the estimate is a function of both θ and θ', the Exchange algorithm is not an implementation of the Pseudo-marginal algorithm and more than one sample cannot be used to improve mixing.

3.3.1 Proposed methodology

The aim of this Chapter is to develop a pseudo-marginal MCMC algorithm to carry out exact Bayesian inference for doubly-intractable models of a range of sizes. Hence we require unbiased estimates of the likelihood. For each θ and y, it is shown that one can construct random variables $\{V_\theta^{(j)},\ j \geq 0\}$ (where dependence on $\mathbf{y}$ is omitted) such that the series defined as

$$\pi(\theta, \{V_\theta^{(j)}\}|\mathbf{y}) := \sum_{j=0}^{\infty} V_\theta^{(j)}$$

is finite almost surely, has finite expectation, and $\mathbb{E}\left(\pi(\theta,\{V_\theta^{(j)}\}|\mathbf{y})\right) = \pi(\theta|\mathbf{y})$. We propose a number of ways to construct such series. Although unbiased, these estimators are not practical as they involve infinite series. A computationally feasible truncation of the infinite sum is therefore employed which, crucially, remains unbiased. This is achieved using Russian Roulette procedures well-known in the Physics literature (Hendricks and Booth, 1985; Carter and Cashwell, 1975). More precisely, a random time τ_θ is introduced, such that with $\mathbf{u} := (\tau_\theta, \{V_\theta^{(j)}, 0 \leq j \leq \tau_\theta\})$ the estimate

$$\pi(\theta,\mathbf{u}|\mathbf{y}) : \sum_{j=0}^{\tau_\theta} V_\theta^{(j)} \quad \text{satisfies} \quad \mathrm{E}\left[\pi(\theta,\mathbf{u}|\mathbf{y})|\{V_\theta^{(j)}, j \geq 0\}\right] = \sum_{j=0}^{\infty} V_\theta^{(j)}.$$

As in the notation used previously, $\mathbf{u}$ is a vector of all the random variables used in the unbiased estimate, i.e. those used to estimate terms in the series, as well as those used in the roulette methods to truncate the series. As the posterior is only required up to a normalising constant in $\mathbf{y}$ and the prior is assumed tractable, in reality we require an unbiased estimate of the likelihood.

3.3.2 The Sign Problem

If the known function $f(\mathbf{y};\theta)$ forming the estimate of the target is bounded then the whole procedure can proceed without difficulty, assuming the bound provides efficiency of sampling. However in the more general situation where the function is not bounded there is a complication here in that the unbiased estimate $\pi(\theta,\mathbf{u}|\mathbf{y})$ is not guaranteed to be positive (although its expectation is non-negative). This issue prevents us from plugging-in directly the estimator $\pi(\theta,\mathbf{u}|\mathbf{y})$ in the Pseudo-marginal framework for the case of unbounded functions. The problem of such unbiased estimators returning negative valued estimates turns out to be a well-studied issue in the Quantum Monte Carlo literature, see e.g. (Lin et al., 2000). The problem is known as the Sign Problem which in its most general form is NP-hard (non-deterministic polynomial time hard) (Troyer and Wiese, 2005) and at present no general and practical solution is available. Indeed, recent work by Jacob and Thiery (2013) showed that given unbiased estimators of $\lambda \in \mathbb{R}$, no algorithm exists to yield an unbiased estimate of $f(\lambda) \in \mathbb{R}^+$, where f is a non-constant real-valued function. Therefore, we will need to apply a different approach to this problem.

We apply methodology developed in Lin et al. (2000) and show that with a weighting of expectations it is still possible to compute any integral of the form $\int h(\theta)\pi(\theta|\mathbf{y})d\theta$ by Markov Chain Monte Carlo.

Suppose that we have an unbiased, but not necessarily positive, estimate of the likelihood $\hat{p}(\mathbf{y}|\theta,\mathbf{u})$ and we wish to sample from $\pi(\theta,\mathbf{u}|\mathbf{y}) = \hat{p}(\mathbf{y}|\theta,\mathbf{u})\pi(\theta)p(\mathbf{u})/p(\mathbf{y})$ where $p(\mathbf{y}) = \int\int p(\mathbf{y}|\theta,\mathbf{u})\pi(\theta)p(\mathbf{u})\mathrm{d}\theta\mathrm{d}\mathbf{u}$ is an intractable normaliser. Although $\pi(\theta,\mathbf{u}|\mathbf{y})$ integrates to one, it is not a probability as it is not necessarily positive. Define $\sigma(\mathbf{y}|\theta,\mathbf{u}) := \mathsf{sign}(\hat{p}(\mathbf{y}|\theta,\mathbf{u}))$, where $\mathsf{sign}(x) = 1$ when $x > 0$, $\mathsf{sign}(x) = -1$ if $x < 0$ and $\mathsf{sign}(x) = 0$ if $x = 0$. Furthermore denote $|\hat{p}(\mathbf{y}|\theta,\mathbf{u})|$ as the absolute value of the estimate, then we have $\hat{p}(\mathbf{y}|\theta,\mathbf{u}) = \sigma(\mathbf{y}|\theta,\mathbf{u})\,|\hat{p}(\mathbf{y}|\theta,\mathbf{u})|$.

Suppose that we wish to compute the expectation

$$\int h(\theta)\pi(\theta|\mathbf{y})d\theta = \int\int h(\theta)\;\pi(\theta,\mathbf{u}|\mathbf{y})\mathrm{d}\mathbf{u}\;\mathrm{d}\theta. \tag{3.6}$$

The above integral can be written

$$\begin{aligned}
\int h(\theta)\pi(\theta|\mathbf{y})d\theta &= \int\int h(\theta)\;\pi(\theta,\mathbf{u}|\mathbf{y})\mathrm{d}u\;\mathrm{d}\theta \\
&= \frac{1}{p(\mathbf{y})}\int\int h(\theta)\;\hat{p}(\mathbf{y}|\theta,\mathbf{u})\pi(\theta)p(\mathbf{u})\;\mathrm{d}\mathbf{u}\;d\theta \\
&= \frac{\int\int h(\theta)\sigma(\mathbf{y}|\theta,\mathbf{u})\;|\hat{p}(\mathbf{y}|\theta,\mathbf{u})|\;\pi(\theta)p(\mathbf{u})\;\mathrm{d}\mathbf{u}\;\mathrm{d}\theta}{\int\int \sigma(\mathbf{y}|\theta,\mathbf{u})\;|\hat{p}(\mathbf{y}|\theta,\mathbf{u})|\;\pi(\theta)p(\mathbf{u})\;\mathrm{d}\mathbf{u}\;\mathrm{d}\theta} \\
&= \frac{\int\int h(\theta)\sigma(\mathbf{y}|\theta,\mathbf{u})\;\breve{\pi}(\theta,\mathbf{u}|\mathbf{y})\;\mathrm{d}\mathbf{u}\;\mathrm{d}\theta}{\int\int \sigma(\mathbf{y}|\theta,\mathbf{u})\;\breve{\pi}(\theta,\mathbf{u}|\mathbf{y})\;\mathrm{d}\mathbf{u}\;\mathrm{d}\theta},
\end{aligned} \tag{3.7}$$

where $\breve{\pi}(\theta,u|\mathbf{y})$ is the distribution

$$\breve{\pi}(\theta,\mathbf{u}|\mathbf{y}) := \frac{|\hat{p}(\mathbf{y}|\theta,\mathbf{u})|\pi(\theta)p(\mathbf{u})}{\int\int|\hat{p}(\mathbf{y}|\theta,\mathbf{u})|\pi(\theta)p(\mathbf{u})\;\mathrm{d}\mathbf{u}\;\mathrm{d}\theta}.$$

We can sample from $\breve{\pi}(\theta,\mathbf{u}|\mathbf{y})$ using a Pseudo-marginal scheme. At each iteration we propose a new value θ', generate an unbiased estimate of the likelihood

$p(\mathbf{y}|\theta',\mathbf{u}')$, and accept it with probability

$$\min\left\{1,\ \frac{|\hat{p}(\mathbf{y}|\theta',u')|\pi(\theta')}{|\hat{p}(\mathbf{y}|\theta,u)|\pi(\theta)}\times\frac{q(\theta|\theta')}{q(\theta'|\theta)}\right\},$$

remembering to save the value and sign of the accepted estimate. We can then use Monte Carlo to estimate the expectation in (3.6) using (3.7) with

$$\int h(\theta)\pi(\theta|\mathbf{y})d\theta=\frac{\sum_{i=1}^{N}h(\theta_i)\sigma(\mathbf{y}|\theta_i,u_i)}{\sum_{i=1}^{N}\sigma(\mathbf{y}|\theta_i,u_i)}. \tag{3.8}$$

The output of this MCMC procedure gives an importance-sampling-type estimate for the desired expectation $\int h(\theta)\pi(\theta|\mathbf{y})d\theta$, which is consistent but biased (as with estimates from all MCMC methods). Importantly, this methodology gives us freedom to use unbiased estimators which may occasionally return negative estimates.

It is important to note that without a *strictly positive* unbiased estimator to plug into the Pseudo-marginal acceptance ratio, the samples are drawn from $\breve{\pi}(\theta,\mathbf{u}|\mathbf{y})$ and *not* the true posterior. However, consistent expectations with respect to the true posterior can still be computed using the importance sampling-style estimator in (3.8).

The following section addresses the issue of constructing the unbiased estimator to be used in the overall MCMC scheme.

3.4 Pseudo-marginal MCMC for doubly-intractable distributions

The foundational component of Pseudo-marginal MCMC is the unbiased and positive estimator of the target density. In the methodology developed here, it is not essential for the estimate of the intractable distribution to be strictly positive and we exploit this characteristic. Note, that whilst there are many methods for unbiasedly estimating $\mathcal{Z}(\theta)$ such as importance sampling, Sequential Monte Carlo (SMC)

(Moral et al., 2006) and Annealed Importance Sampling (AIS) (Neal, 2001), if we then take some non-linear function of the estimate, for example the reciprocal, the overall estimate of the likelihood is no longer unbiased.

It is possible to directly construct an estimator of $1/\mathcal{Z}(\theta)$ using an instrumental density $q(\mathbf{y})$ as follows

$$\frac{1}{\mathcal{Z}(\theta)} = \frac{1}{\mathcal{Z}(\theta)} \int q(\mathbf{y}) \mathrm{d}\mathbf{y} = \int \frac{q(\mathbf{y})}{f(\mathbf{y};\theta)} p(\mathbf{y}|\theta) \mathrm{d}\mathbf{y} \approx \frac{1}{N} \sum_{i=1}^{N} \frac{q(\mathbf{y}_i)}{f(\mathbf{y}_i;\theta)} \qquad \mathbf{y}_i \sim p(\cdot|\theta),$$

however this requires the ability to sample from the likelihood, and if we can do this then we can implement the Exchange algorithm. Further, the variance of the estimate depends strongly on the choice of the instrumental density. A biased estimator can be constructed by sampling the likelihood using MCMC (e.g. Zhang et al., 2012), but a Pseudo-marginal scheme based on this estimate will not target the correct posterior distribution. Very few methods to estimate $1/\mathcal{Z}(\theta)$ can be found in the Statistics or Physics literature, presumably because in most situations a consistent estimate will suffice. Therefore, we have to look for other ways to generate an unbiased estimate of the likelihood.

In outline, the intractable distribution is first written in terms of a nonlinear function of the intractable normalising term. For example, in Equation (3.1), the non-linear function is the reciprocal $1/\mathcal{Z}(\theta)$, and an equivalent representation would be $\exp(-\log \mathcal{Z}(\theta))$. This function is then represented by a convergent Maclaurin expansion which has the property that each term can be estimated unbiasedly using the available unbiased estimates of $\hat{\mathcal{Z}}(\theta)$. The infinite series expansion is then stochastically truncated without introducing bias so that only a finite number of terms need be computed. These two components—(1) unbiased independent estimates of the normalising constant, and (2) unbiased stochastic truncation of the infinite series representation—then produce an unbiased, though not strictly positive, estimate of the intractable distribution. The final two components of the overall methodology consist of (3) constructing an MCMC scheme which targets a distribution proportional to the absolute value of the unbiased estimator, and then (4) computing Monte Carlo estimates with respect to the desired posterior distribution as detailed in the previous section.

This method draws together ideas from several places in the Statistics and Physics literature. In the Physics literature, researchers used a similar method to obtain unbiased estimates of $\exp(-U(x))$ when only unbiased estimates of $U(x)$ were available (Kennedy and Kuti, 1985; Bhanot and Kennedy, 1985). They further showed that even when using such unbiased estimates in place of the true value, detailed balance still held. The method for realising the unbiased estimates at each iteration is also similar to that suggested by Booth (2007), in which he described a method for unbiasedly estimating the reciprocal of an integral, which is of obvious relevance to our case. In the Statistics literature, Douc and Robert (2011) used a geometric series to estimate an inverse probability, and Beskos et al. (2006); Fearnhead et al. (2008) also used techniques to truncate a series unbiasedly in their work on likelihood estimation for stochastic diffusions. Finally both Rhee and Glynn (2012) and McLeish (2011) use roulette methods to realise an unbiased estimate when only biased but consistent estimates are available. This is achieved by writing the quantity to be unbiasedly estimated as an infinite series in which each term is a function of the consistent estimates which can be generated, and then truncating the series using roulette methods. However, they do not utilise these estimates in a Pseudo-marginal MCMC scheme.

In the following sections, two series expansions of a doubly-intractable likelihood are presented, in which each term can be estimated unbiasedly using unbiased estimates of $\mathcal{Z}(\theta)$. Following this comes a description of unbiased truncation methods.

3.4.1 Geometric Series Estimator

In this subsection we show how the intractable likelihood can be written as a geometric series in which each term can be estimated unbiasedly. Take a biased estimate of the likelihood $\tilde{p}(\mathbf{y}|\theta) = f(\mathbf{y};\theta)/\tilde{\mathcal{Z}}(\theta)$, where $\tilde{\mathcal{Z}}(\theta) > 0$ is ideally an upper bound on $\mathcal{Z}(\theta)$, or alternatively an unbiased importance sampling estimate or a deterministic approximation. Then, using a multiplicative correction

$$p(\mathbf{y}|\theta) = \tilde{p}(\mathbf{y}|\theta) \times c(\theta) \left[1 + \sum_{n=1}^{\infty} \kappa(\theta)^n \right], \tag{3.9}$$

where $\kappa(\theta) = 1 - c(\theta)\mathcal{Z}(\theta)/\widetilde{\mathcal{Z}}(\theta)$ and $c(\theta)$ ensures $|\kappa(\theta)| < 1$, the convergence of a geometric series gives

$$\tilde{p}(\mathbf{y}|\theta) \times c(\theta) \left[1 + \sum_{n=1}^{\infty} \kappa(\theta)^n \right] = \tilde{p}(\mathbf{y}|\theta) \times \frac{c(\theta)}{1-\kappa(\theta)} = \tilde{p}(\mathbf{y}|\theta) \times \frac{\tilde{\mathcal{Z}}(\theta)}{\mathcal{Z}(\theta)} = p(\mathbf{y}|\theta).$$

Based on this equality, and with an infinite number of independent unbiased estimates of $\mathcal{Z}(\theta)$ each denoted $\hat{\mathcal{Z}}_i(\theta)$, an unbiased estimate of the target density is

$$\hat{\pi}(\theta|\mathbf{y}) = \frac{\pi(\theta)\tilde{p}(\mathbf{y}|\theta)}{p(\mathbf{y})} \times c(\theta) \left[1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \left(1 - c(\theta) \frac{\hat{\mathcal{Z}}_i(\theta)}{\tilde{\mathcal{Z}}(\theta)} \right) \right]. \tag{3.10}$$

Notice that the series in (3.10) is finite a.s. and we can interchange summation and expectation if

$$E\left(\left| 1 - c(\theta) \frac{\hat{\mathcal{Z}}_i(\theta)}{\tilde{\mathcal{Z}}(\theta)} \right| \right) < 1.$$

Since $E(|X|) \leq E^{1/2}(|X|^2)$, a sufficient condition for this is $0 < c(\theta) < 2\tilde{\mathcal{Z}}(\theta)\mathcal{Z}(\theta)/E\left(\hat{\mathcal{Z}}_1^2(\theta)\right)$, which is slightly more stringent than $|\kappa(\theta)| < 1$. Under this assumption, the expectation of $\hat{\pi}(\theta|\mathbf{y})$ is

$$\begin{aligned} E\left\{ \hat{\pi}(\theta|\mathbf{y}) | \tilde{\mathcal{Z}}(\theta) \right\} &= \frac{\pi(\theta)\tilde{p}(\mathbf{y}|\theta)}{p(\mathbf{y})} \times c(\theta) \left[1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \left(1 - c(\theta) \frac{E\left\{ \hat{\mathcal{Z}}_i(\theta) \right\}}{\tilde{\mathcal{Z}}(\theta)} \right) \right] \\ &= \frac{\pi(\theta)\tilde{p}(\mathbf{y}|\theta)}{p(\mathbf{y})} \times c(\theta) \left[1 + \sum_{n=1}^{\infty} \kappa(\theta)^n \right] \\ &= \pi(\theta|\mathbf{y}). \end{aligned}$$

Therefore, the essential property $E\{\hat{\pi}(\theta|\mathbf{y})\} = \pi(\theta|\mathbf{y})$ required for Exact-Approximate MCMC is satisfied by this geometric correction. However, there are difficulties with this estimator. It will be difficult in practice to find $c(\theta)$ that ensures the series in (3.10) is convergent in the absence of knowledge of the actual value of $\mathcal{Z}(\theta)$. By ensuring that $\tilde{\mathcal{Z}}(\theta)/c(\theta)$ is a strict upper-bound on $\mathcal{Z}(\theta)$, denoted

by $\mathcal{Z}_U$, guaranteed convergence of the geometric series is established. Even if an upper bound is available, it may not be computationally practical as upper bounds on normalising constants are typically loose (see, for example El Ghaoui and Gueye, 2008), making the ratio $\mathcal{Z}(\theta)/\mathcal{Z}_U$ extremely small, and, therefore, $\kappa(\theta) \approx 1$; in this case, the convergence of the geometric series will be slow. A more pragmatic approach is to use a pilot run at the start of each iteration to characterise the location and variance of the $\mathcal{Z}(\theta)$ estimates, and use this to conservatively select $\tilde{\mathcal{Z}}(\theta)/c(\theta)$ such that the series converges. Of course, if the distribution of the estimates is not well enough characterised then we may not be able to guarantee with probability 1 that $|\kappa(\theta)| < 1$ and hence approximation will be introduced into the chain.

In the next section we describe an alternative to the geometric series estimator which does not have the practical issue of ensuring the region of convergence is maintained.

3.4.2 Unbiased estimators using an exponential auxiliary variable

In this section we show how the introduction of an auxiliary variable can enable the posterior density to be written in terms of a Taylor series expansion of the exponential function. The introduction of $v \sim \mathsf{Expon}(\mathcal{Z}(\theta))$ defines a joint distribution of the form of

$$\begin{aligned}
\pi(\theta, v|\mathbf{y}) &= [\mathcal{Z}(\theta)\exp(-v\mathcal{Z}(\theta))] \times \frac{f(\mathbf{y};\theta)}{\mathcal{Z}(\theta)} \times \pi(\theta) \times \frac{1}{p(\mathbf{y})} \\
&= \exp(-v\mathcal{Z}(\theta)) \times f(\mathbf{y};\theta) \times \pi(\theta) \times \frac{1}{p(\mathbf{y})} \\
&= \left[1 + \sum_{n=1}^{\infty} \frac{(-v\mathcal{Z}(\theta))^n}{n!}\right] \times f(\mathbf{y};\theta) \times \pi(\theta) \times \frac{1}{p(\mathbf{y})}.
\end{aligned}$$

Integrating over v returns the posterior distribution and therefore if we sample from this joint distribution, our θ samples will be distributed according to the posterior. As hinted at in the previous section, the methods used to truncate the series are more computationally feasible if the series converges quickly. Therefore, we introduce $\tilde{\mathcal{Z}}(\theta)$ which is preferably an upper bound on $\mathcal{Z}(\theta)$ or if unavailable, some other approximation. The exponential can then be expanded as follows

$$\begin{aligned}\exp(-\nu\mathcal{Z}(\theta)) &= \exp(-\nu\tilde{\mathcal{Z}}(\theta)) \times \exp\left(\nu(\tilde{\mathcal{Z}}(\theta)-\mathcal{Z}(\theta))\right) \\ &= \exp(-\nu\tilde{\mathcal{Z}}(\theta)) \times \left(1+\sum_{n=1}^{\infty}\frac{\nu^n}{n!}\left(\tilde{\mathcal{Z}}(\theta)-\mathcal{Z}(\theta)\right)^n\right).\end{aligned}$$

If $\tilde{\mathcal{Z}}(\theta)$ is an upper bound on $\mathcal{Z}(\theta)$ then its introduction prevents the terms in the Taylor series from alternating in sign, by ensuring the exponent is positive; this helps to reduce the impact of returning negative estimates. Even if $\tilde{\mathcal{Z}}(\theta)$ is not a strict upper bound, its presence reduces the absolute value of the exponent which improves the convergence properties of the series, and therefore makes the truncation methods described in the next section more efficient.

An unbiased estimator of the series is

$$\widehat{\exp(-\nu\mathcal{Z}}(\theta)) = \exp(-\nu\tilde{\mathcal{Z}}(\theta))\left[1+\sum_{n=1}^{\infty}\frac{\nu^n}{n!}\prod_{i=1}^{n}(\tilde{\mathcal{Z}}(\theta)-\hat{\mathcal{Z}}_i(\theta))\right], \tag{3.11}$$

where $\{\hat{\mathcal{Z}}_i(\theta), i \geq 1\}$ are i.i.d. random variables with expectation equal to $\mathcal{Z}(\theta)$. The magnitude of the exponent can present computational barriers to the implementation of this scheme. If $\mathcal{Z}(\theta)$ is very large it is easier to carry out the division $\hat{\mathcal{Z}}(\theta)/\mathcal{Z}(\theta)$ in (3.10) (which can be computed in log space), than the subtraction $\mathcal{Z}(\theta)-\hat{\mathcal{Z}}(\theta)$ in (3.11). On the other hand, since $n!$ grows faster than the exponential, this series is always well defined (finite almost surely).

In Fearnhead et al. (2008), the *Generalised Poisson Estimator*, originally proposed in Beskos et al. (2006), is employed to estimate transition functions that are similar to (3.11). Here again, this series is finite almost surely with finite expectation. The choice of which estimator to employ will be problem dependent and, in situations where it is difficult to guarantee convergence of the geometric series, this form of estimator may be more suitable.

3.4.3 Possible choices for $\hat{\mathcal{Z}}_i(\theta)$ and $\tilde{\mathcal{Z}}(\theta)$

At this point, it is useful to describe in a bit more detail the possible choices for $\hat{\mathcal{Z}}_i(\theta)$ and $\tilde{\mathcal{Z}}(\theta)$. For the unbiased estimates of the normalising term, some form of Monte Carlo estimator is envisaged, either simple importance sampling or Sequen-

tial Monte Carlo (SMC). These methods produce unbiased estimates, of which the variance can easily be estimated and controlled. Whilst in theory the variance, and hence efficiency, of Monte Carlo methods is independent of dimension, in practice the variance is highly influenced by the dimension as it becomes more difficult to design good importance distributions as the dimension increases. The methodology therefore suffers to some extent from the same problem as the Wang-Landau type methods. The advantage of our method, however, is that the equilibrium distribution of the Markov chain is constant throughout the process and so all the samples can be used, whereas for the Wang-Landau methods, the equilibrium distribution converges to the desired posterior as the chain is run, and it is this convergence that is slowed as the dimension increases.

If an upper bound is available for all all the estimates of $\mathcal{Z}(\theta)$, then it is possible to produce a strictly positive estimate of $1/\mathcal{Z}(\theta)$ and hence the true posterior can be sampled from. The data for many doubly-intractable models e.g. Ising and Potts models are discrete, and in these cases it is often possible to find an upper bound. However, it will often be the case that an upper bound of this nature is not available, and so the estimates returned will sometimes be negative and the distribution sampled will not be the true posterior. This is obviously a drawback to the methodology as one of the major benefits of Bayesian methodology is obtaining probability distributions for parameters as opposed to point estimates. However, it is still possible to obtain consistent estimates of functions with respect to the true posterior, including functions such as the second moment, and often this together with a mean is all that is required.

For $\tilde{\mathcal{Z}}(\theta)$, a range of options is available. In the best case scenario, an upper bound is available on $\mathcal{Z}(\theta)$ and the geometric series construction can then be used with a guarantee of convergence. Upper bounds are most likely to be available in the case where the state space is finite or the data is confined to a bounded interval. For example, a naive bound for the Ising model can be found by setting all spins to +1 and multiplying by the total number of graphs in the state space. Tighter bounds, which require more computation, have been found by Wainwright et al. (2005); Liu and Ihler (2011); Ghaoui and Gueye (2008) amongst others.

In other cases, such as the Fisher-Bingham distribution introduced later in the Chapter, the unnormalised function, $f(\mathbf{y};\theta)$, is an exponential function with a convex sum i.e. $\sum_{i=1}^{M} a_i x_i$ where the x_i are real vectors and $a_i > 0$ for all i and $\sum_{i=1}^{M} a_i = 1$. An upper bound can therefore be determined by giving a weight of 1 to the largest

x_i.

It is also possible prior to the MCMC run to estimate values of $\mathcal{Z}(\theta)$ for a grid of θ points and use sample estimates of the mean and variance to determine empirical upper bounds with very high probability. This is obviously a more feasible approach when the dimension of the parameter space is low, however, this is often the case with Markov Random Field models as inference is even more challenging when the parameter space becomes large.

In the following section, the final element of the proposed methodology is discussed: unbiased truncation of the infinite series estimators.

3.5 Unbiased Truncation of Infinite Sums: Russian Roulette

Two unbiased estimators of nonlinear functions of a normalising constant have been considered. Both of them rely on the availability of an unbiased estimator for $\mathcal{Z}(\theta)$ and a series representation of the nonlinear function. We now require a computationally feasible means of realising the desired estimator without explicitly computing the infinite sum and without introducing any bias into the final estimate. It transpires that there are a number of ways to randomly truncate the convergent infinite sum $\mathcal{S}(\theta) = \sum_{i=0}^{\infty} \phi_i(\theta)$ in an unbiased manner; these stem from work by von Neumann and Ulam in the 1940s, see Papaspiliopoulos (2009) for a good review of such methods.

3.5.1 Single Term Weighted Truncation

The simplest unbiased truncation method is to define a set of probabilities and draw an integer index k with probability q_k then return $\phi_k(\theta)/q_k$ as the estimator. It is easy to see that the estimator is unbiased as $E\{\hat{\mathcal{S}}(\theta)\} = \sum_k q_k \phi_k(\theta)/q_k = \mathcal{S}(\theta)$. The definition of the probabilities should be chosen to minimise the variance of the estimator, see e.g. Fearnhead et al. (2008). An example could be that each index is drawn from a Poisson distribution $k \sim \mathsf{Poiss}(\lambda)$ with $q_k = \lambda^k \exp(-\lambda)/k!$. However in the case of a geometric series where $\phi_k(\theta) = \phi^k(\theta)$, the variance of the estimator will be infinite with this choice since the factorial function $k!$ grows faster than the exponential. Using the geometric distribution as our importance distribution, the variance is finite subject to some conditions on the choice of p, the

parameter of the geometric distribution. To see this, note that, as k is chosen with probability $q_k = p^k(1-p)$, the second moment,

$$\mathbb{E}[\hat{S}^2] = \sum_{k=0}^{\infty} \mathbb{E}[\hat{S}_k^2]\, q_k = \sum_{k=0}^{\infty} \frac{\mathbb{E}[\phi_k^2]}{p^k(1-p)}$$

is finite if

$$\lim_{k\to\infty} \left| \frac{\mathbb{E}[\phi_{k+1}^2]}{p\mathbb{E}[\phi_k^2]} \right| < 1. \tag{3.12}$$

As the values of ϕ_k are unknown, the best way to design the probabilities q_k is to precompute some estimates of the first few ϕ_k and then choose the tail probabilities to be geometric such that the overall estimator has finite variance.

3.5.2 Russian Roulette

An alternative unbiased truncation that exhibits superior performance in practice is based on a classic Monte Carlo scheme, known as Russian Roulette in the Physics literature (Lux and Koblinger, 1991; Carter and Cashwell, 1975). The procedure is based on the simulation of a finite random variable (stopping time) τ according to some probabilities $p_n = \mathbb{P}(\tau \geq n) > 0$ for all $n \geq 0$ with $p_0 = 1$. Define the weighted partial sums as $S_0 = \phi_0$ and for $k \geq 1$

$$S_k = \phi_0 + \sum_{j=1}^{k} \frac{\phi_j}{p_j}.$$

The Russian Roulette estimate of S is $\hat{S} = S_\tau$. Russian Roulette implementations in the Physics literature commonly choose a stopping time of the form

$$\tau = \inf\{k \geq 1 : \ U_k \geq q_k\},$$

where $\{U_j,\ j \geq 1\}$ are i.i.d. $\mathcal{U}(0,1)$, $q_j \in (0,1]$ and $\hat{S} = S_{\tau-1}$. In this case $p_n = \prod_{j=1}^{n-1} q_j$.

It can be shown that the expectation of the estimate is as required

$$\sum_{k=0}^{n} S_k P(\tau = k) = \sum_{k=0}^{n} S_k (p_k - p_{k+1}) = \phi_0 + \sum_{k=0}^{n-1} S_{k+1} p_{k+1} - \sum_{k=0}^{n} S_k p_{k+1} = \sum_{k=0}^{n} \phi_k - S_n p_{n+1}.$$

By Kronecker's lemma $\lim_{n\to\infty} p_n S_n = 0$, and $|p_{n+1} S_n| = (p_{n+1}/p_n) p_n |S_n| \leq p_n |S_n| \to 0$, as $n \to \infty$. We conclude that $\mathrm{E}[\hat{S}(\theta)] = \sum_{k=0}^{\infty} S_k P(\tau = k) = \sum_{k=0}^{\infty} \phi_k = S(\theta)$. Refer to Appendix C for a more detailed discussion of the variance of such an estimator and how to design the sequence of probabilities (p_n).

Based on results presented in the Appendix, for a geometric series where $\phi_k(\theta) = \phi^k(\theta)$, if one chooses $q_j = q$, then the variance will be finite provided $q > \phi(\theta)^2$. In general there is a trade-off between the computing time of the scheme and the variance of the returned estimate. If the selected q_j's are close to unity, the variance is small, but the computing time is high. But if q_j's are close to zero, the computing time is fast but the variance can be very high, possibly infinite. In the case of the geometric series, $\phi_k(\theta) = \phi^k(\theta)$, choosing $q_j = q = \phi(\theta)$ works reasonably well in practice.

Results from Rhee and Glynn (2013), where a similar construction is analysed taking full account of the random nature of each ϕ_k, show that the variance of the estimator will be finite if

$$\sum_{i=0}^{\infty} \frac{\mathbb{E}[(\phi_i - S)^2]}{p_i} < \infty. \tag{3.13}$$

For some series, properties of the numerator of (3.13) may be known, in which case the probabilities p_i can be designed to ensure the second moment of the estimator is finite. If this is not the case, then pre-computation may be used to estimate the magnitude of the expectations in (3.13) and then design the probabilities such that they decay more slowly than the expectations.

As an illustrative example, consider the joint density

$$p(\theta,\nu,\mathbf{u}|\mathbf{y}) = \exp(-\nu\tilde{\mathcal{Z}}(\theta)) \times \left(1 + \sum_{n=1}^{\tau_\theta} \frac{\nu^n}{q^n n!} \prod_{i=1}^{n} \left(\tilde{\mathcal{Z}}(\theta) - \hat{\mathcal{Z}}_i(\theta)\right)\right) \times \frac{f(\mathbf{y};\theta)\pi(\theta)}{p(\mathbf{y})}, \tag{3.14}$$

where the random variable $\mathbf{u}$ represents the random variables in the estimates $\hat{\mathcal{Z}}_i(\theta)$ and the random variable used in Russian roulette truncation, and $q^n = \prod_{l=1}^{n} q_l$ denotes the probabilities in the Russian Roulette truncation. If we define a proposal for ν' as $q(\nu'|\theta') = \tilde{\mathcal{Z}}(\theta')\exp(-\nu'\tilde{\mathcal{Z}}(\theta'))$ and a proposal for θ' as $q(\theta'|\theta)$ then the Hastings ratio for a transition kernel with invariant density $\pi(\theta,\nu,\mathbf{u}|\mathbf{y})$ follows as

$$\frac{f(\mathbf{y};\theta')}{f(\mathbf{y};\theta)} \times \frac{\tilde{\mathcal{Z}}(\theta)}{\tilde{\mathcal{Z}}(\theta')} \times \frac{\pi(\theta')}{\pi(\theta)} \times \frac{q(\theta|\theta')}{q(\theta'|\theta)} \times \phi(\nu,\nu',\theta,\theta') \tag{3.15}$$

where

$$\phi(\nu,\nu',\theta,\theta') = \frac{1 + \sum_{m=1}^{\tau_{\theta'}} \frac{(\nu')^m}{q^m m!} \prod_{j=1}^{m} \left(\tilde{\mathcal{Z}}(\theta') - \hat{\mathcal{Z}}_j(\theta')\right)}{1 + \sum_{n=1}^{\tau_\theta} \frac{\nu^n}{q^n n!} \prod_{i=1}^{n} \left(\tilde{\mathcal{Z}}(\theta) - \hat{\mathcal{Z}}_i(\theta)\right)}. \tag{3.16}$$

It is interesting to note that $\phi(\nu,\nu',\theta,\theta')$ acts as a multiplicative correction for the Hastings ratio that uses the approximate normalising term $\tilde{\mathcal{Z}}(\theta)$ rather than the actual $\mathcal{Z}(\theta)$. The required marginal $\pi(\theta|\mathbf{y})$ follows due to the unbiased nature of the estimator.

The Russian roulette methodology has been used in various places in the literature. McLeish (2011) and Rhee and Glynn (2012); Glynn and Rhee (2014) use the Russian roulette estimator to 'debias' a biased but consistent estimator. In their construction the aim is to unbiasedly estimate X, for which only a sequence of approximations is available, X_i, with $\mathrm{E}[X_i] \to \mathrm{E}[X]$ as $i \to \infty$. Define an infinite series, $S = \mathrm{X}_0 + \sum_{n=1}^{\infty}(X_n - X_{n-1})$; an unbiased estimate of S is an unbiased estimate of X, assuming that the estimates are good enough to interchange expectation and summation. To achieve a computationally feasible and unbiased estimator of X, the Roulette or Poisson truncation schemes can be applied. In the context of our work this provides an alternative to the geometric or exponential series described above, in which only a consistent estimator is required. One drawback to this 'debiasing'

scheme for use in Pseudo-marginal MCMC is that there is no obvious way to reduce the probability of the final estimate being negative.

3.5.3 Comparison with current algorithms

Both the Exchange and SAVM can also be used to sample from doubly-intractable posteriors, however the main drawback of these methods is that they can only be implemented when it is possible to sample from the likelihood. On the other hand, the methodology described above, can be implemented whenever an unbiased or consistent estimate of $\mathcal{Z}(\theta)$ is available. This means that this new methodology can be applied to a wider class of problems.

The cost of this increased applicability is computational; for problems where it is possible to sample from the likelihood, the Exchange algorithm will generally be least computationally intensive of the exact methods, as only one sample from the likelihood is required per iteration. However, this computational advantage can also be a disadvantage as the Markov chain can mix badly as a result.

The SAVM can also be applied with only one sample, however, the mixing can be improved by averaging more than one estimate. This algorithm has the additional drawback of the need to design the auxiliary distribution $p(x|y,\theta)$ which has a large impact on the efficiency. If this auxiliary density is not designed well enough, more likelihood samples will be required in order to reduce the variance of the importance sampling estimate and to improve the mixing.

3.5.4 Computational Complexity

The number of terms required for the unbiased estimate at each iteration is random, and therefore the running time for a given number of samples is also random. However, it is possible to look at the expected computation for each iteration. The computation required to produce each unbiased estimate depends on:

1. the expected value of the random variable used to truncate the series,

2. the amount of computation required per term in the infinite series.

However, both of these depend in some way on the intrinsic 'difficulty' of the problem, and it is therefore difficult to give general results on the computational complexity of the methodology. If it is possible to produce low variance estimates

(either unbiased or consistent) of the normalising term, $\mathcal{Z}(\theta)$, then the amount of computation required for each series term will be low and the random variable used to truncate can have a low expected value. If, on the other hand, it is difficult to produce low variance estimates of $\mathcal{Z}(\theta)$, then each series term will require a large amount of computation and the optimum truncation distribution will have a high expectation.

One specific example is now investigated to illustrate these points: the geometric series estimator with Single Term Weighted Truncation using a geometric stochastic truncation variable, N, parameterised by p. It was shown in Section 3.5.1 that the variance will be finite if

$$\mathrm{E}[(1-\frac{\hat{\mathcal{Z}}}{\tilde{\mathcal{Z}}})^2] = \mathrm{Var}\,(1-\frac{\hat{\mathcal{Z}}}{\tilde{\mathcal{Z}}}) + \mathrm{E}[(1-\frac{\hat{\mathcal{Z}}}{\tilde{\mathcal{Z}}})]^2 < p. \tag{3.17}$$

Recall that in the geometric construction, the n-th series term is the product of n independent estimates of $1-\mathcal{Z}/\tilde{\mathcal{Z}}$, and that when using Single Term Weighted Truncation only one series term is computed per likelihood estimate. Therefore, the expected computation per likelihood estimate is simply the expectation of N multiplied by the work required to produce one estimate $\hat{\mathcal{Z}}$, denoted by w. The expectation of the geometric random variable N is $p/(1-p)$, and therefore the expected compute time for each unbiased estimate is $pw/(1-p)$. Clearly smaller values of p result in lower computational costs, and from 3.17 it can be seen that both the variability of the estimates, $\hat{\mathcal{Z}}$, and the 'tightness' of the upper bound, $\tilde{\mathcal{Z}}$, impact how low p can be set and hence the computation required per likelihood estimate.

The size of w, the computation required to compute each Monte Carlo estimate, $1-\hat{\mathcal{Z}}/\tilde{\mathcal{Z}}$, varies depending on the specific model. Agapiou and Papaspiliopoulos (2015) suggest that the number of samples required for accurate importance sampling scales exponentially with the Kullback-Leibler divergence between the proposal and target distributions, and therefore the computational complexity of this methodology depends strongly on whether a proposal distribution close to the likelihood can be found. They further show that the number of samples required for accurate importance sampling increases exponentially with the dimension of the state space, and therefore the number of data points also strongly impacts the amount of

computation required for one estimate of $1 - \hat{\mathcal{Z}}/\tilde{\mathcal{Z}}$.

Now that the complete Exact-Approximate MCMC scheme has been detailed, the following section illustrates the methodology on some models that are doubly-intractable, considering the strengths and weaknesses.

3.6 Experimental Evaluation

3.6.1 Simple example

We start by using a simple example from Murray et al. (2006) so as to compare the various algorithms available. Consider sampling from the posterior distribution of the precision parameter of a univariate normal distribution which has a conjugate gamma prior distribution. With N i.i.d. data points, the model is specified as follows:

$$p(\mathbf{y}|\theta) = \prod_{i=1}^{N} \mathcal{N}(y_i; 0, 1/\theta) \qquad p(\theta|\alpha,\beta) = \text{Gamma}(\alpha,\beta).$$

The corresponding posterior distribution is

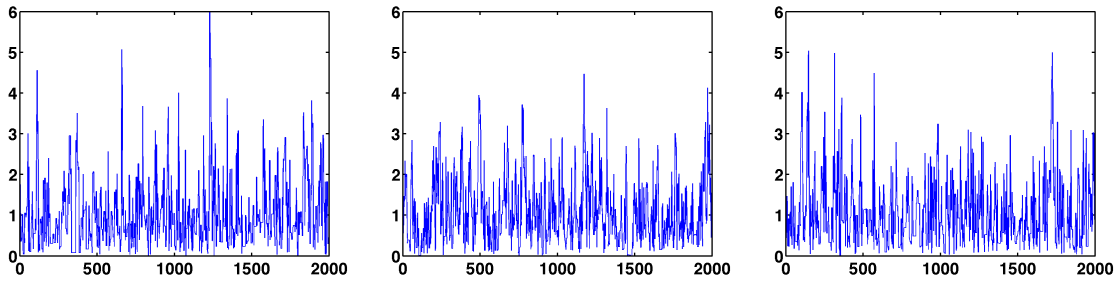
$$p(\theta|\mathbf{y}) = \text{Gamma}(N/2+\alpha, \sum_n y^2/2+\beta),$$

which can be sampled easily using standard methods, however we pretend in this example that the normalising term in the likelihood is unknown.

As the posterior distribution has an analytic form, the mean and the variance of the parameter with respect to the posterior are known, enabling comparison with our approximate methods. The Exchange algorithm can easily be implemented as all that is required at each MCMC iteration is a draw from a normal distribution. A standard MCMC chain assuming the normalising constant is known was also run for comparison. Our methodology was implemented with importance sampling used to estimate the normalising term and Russian roulette used to terminate the infinite series. For all three methods, the chain was run for $100,000$ iterations and the second

	Roulette	Exchange	Metropolis	Exact
Mean	0.9988	1.0021	0.9995	1
Variance	0.6642	0.6602	0.6614	2/3
Relative CPU time	1.23	1	1.02	NA

Table 3.1: Table showing the exact posterior mean and variance of the parameter θ from the simple example in Murray et al. (2006), as well as the estimates from the three sampling methods: Roulette, Exchange and standard Metropolis.



(a) Roulette **(b)** Exchange **(c)** Metropolis

Figure 3.1: MCMC traces for 2000 samples using three different methods to sample from the posterior of the example from Murray et al. (2006).

half of the samples used to estimate the means and variances. A normal distribution was used as the proposal and acceptance rates tuned to around 40%.

Standard importance sampling estimates were used for the $\hat{Z}_i(\theta)$ with a Gaussian importance distribution and 10 samples per estimate. At each iteration five preliminary estimates were computed, and the maximum of these was used as $\tilde{Z}(\theta)$.

The distribution is univariate and unimodal and hence all three methods easily sample from the posterior and take much the same time. From Table 3.1 it is clear that all three methods estimate the mean and variance well. The computational time is also shown; for this example the Exchange algorithm takes about the same amount of time as the standard Metropolis-Hastings algorithm as drawing a sample from a normal distribution does not require much additional computation. The Roulette algorithm takes the longest amount of time as at each iteration at least one estimate of the normalising term is required. From the histogram in Figure 3.2 however, it is clear that at most iterations, only one estimate is required. Figures 3.1 and 3.4 show traces and running means from the three methods respectively. The only difference between the three methods is that the Exchange samples have stronger autocorrelations (Figure 3.3) than the other two methods meaning that for the same number of samples there are fewer independent samples.

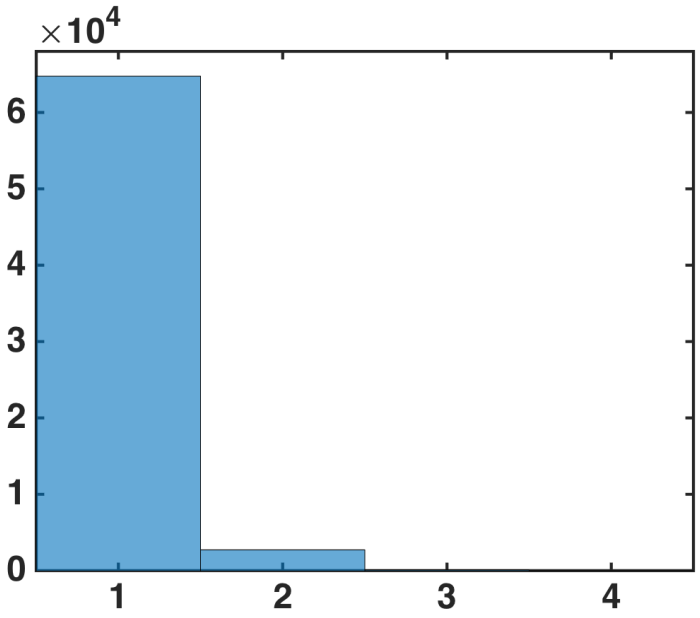
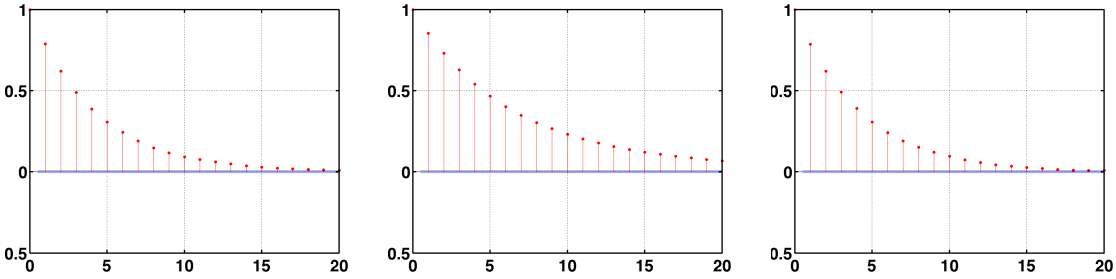
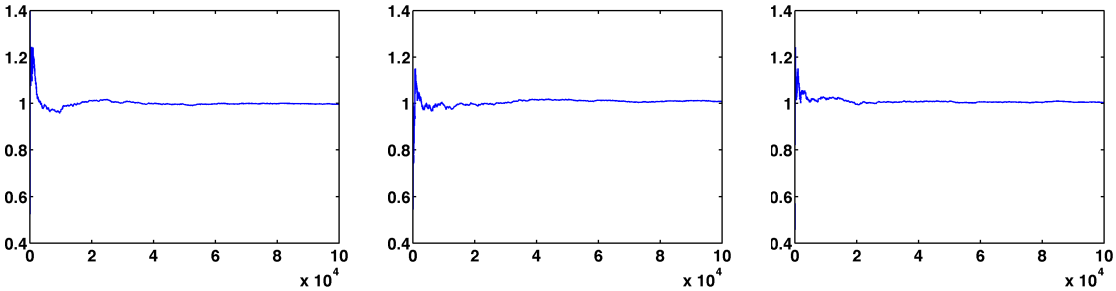


Figure 3.2: Histogram showing number of unbiased estimates required at each MCMC iteration when using Roulette method on example in Murray et al. (2006).



(a) Roulette **(b)** Exchange **(c)** Metropolis

Figure 3.3: Autocorrelation function up to lag 20 for three different methods for samples from example in Murray et al. (2006).



(a) Roulette **(b)** Exchange **(c)** Metropolis

Figure 3.4: Running mean of 10,000 MCMC samples from posterior from example in Murray et al. (2006) using three different methods.

3.6.2 Ising Lattice Spin Models

Ising models are examples of doubly-intractable distributions over which it is challenging to perform inference. They form a prototype for priors for image segmentation and autologistic models e.g. (Hughes et al., 2011; Gu and Zhu, 2001; Møller et al., 2006). Current exact methods such as the Exchange algorithm (Murray et al., 2006) require access to a perfect sampler (Propp and Wilson, 1996) which, while feasible for small grids, cannot be scaled up. A practical alternative is employed in (Caimo and Friel, 2011), where an auxiliary MCMC run is used to approximately simulate from the model. This is inexact and introduces bias, but it is hoped that the bias has little practical impact. In this section, the Exchange algorithm and its approximate version are compared with our Pseudo-marginal methodology.

For an $N \times N$ grid of spins, $\mathbf{y} = (y_1, \ldots y_{N^2})$, $y \in \{+1, -1\}$, the Ising model has likelihood

$$p(\mathbf{y}; \alpha, \beta) = \frac{1}{\mathcal{Z}(\alpha, \beta)} \exp\left(\alpha \sum_{i}^{N^2} y_i + \beta \sum_{i \sim j} y_i y_j \right), \tag{3.18}$$

where i and j index the rows and column of the lattice and the notation $i \sim j$ denotes summation over nearest neighbours. Periodic boundary conditions are used in all subsequent computation. The parameters α and β indicate the strength of the external field and the interactions between neighbours respectively. The normalising constant,

$$\mathcal{Z}(\alpha, \beta) = \sum_{\mathcal{Y}} \exp\left(\alpha \sum_{i}^{N^2} y_i + \beta \sum_{i \sim j} y_i y_j \right), \tag{3.19}$$

requires summation over all 2^{N^2} possible configurations of the model, which is computationally infeasible even for moderately sized lattices. This is, in fact, a naive bound as the transfer matrix method (see for example MacKay (2003)) which has complexity $N2^N$ can also be used to compute the partition function.

Experiments were carried out on a small 10×10 lattice to enable a detailed comparison of the various algorithms. A configuration was simulated using a perfect sampler with parameters set at $\alpha = 0$ and $\beta = 0.2$. Inference was carried out over the posterior distribution $p(\beta|\mathbf{y})$ ($\alpha = 0$ was fixed). A standard Metropolis-Hastings

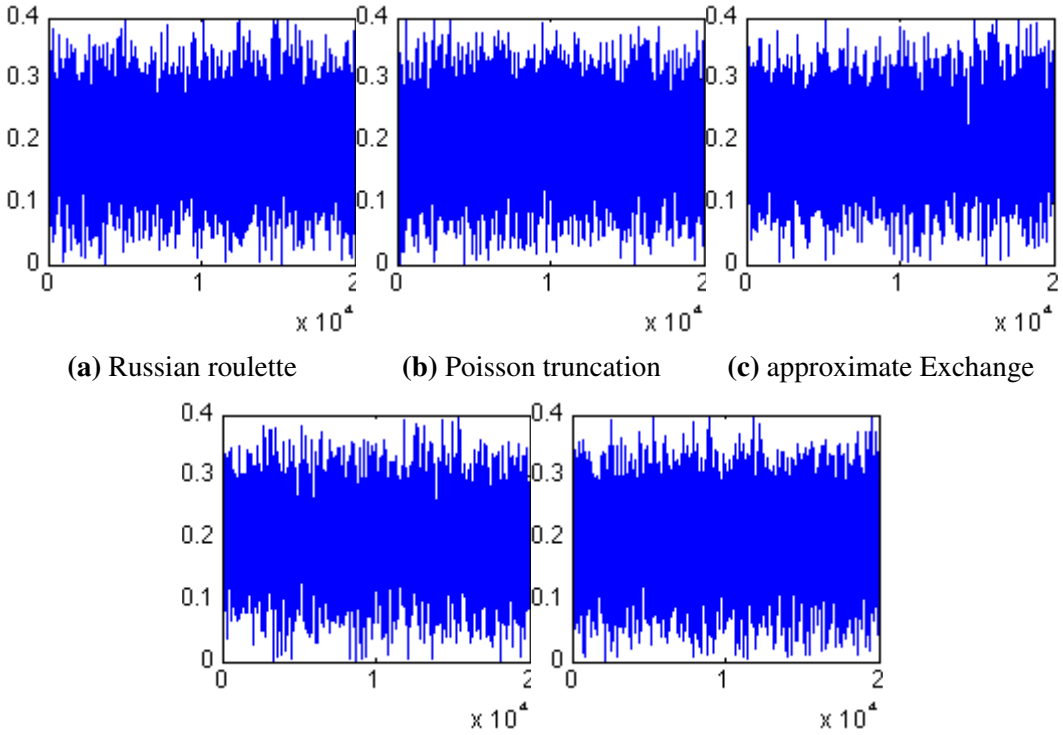
sampling scheme was used to sample the posterior, with a normal proposal distribution centred at the current value and acceptance rates tuned to around 40%. A uniform prior on $[0,1]$ was set over β. As no tight upper bound is available on the normalising term $\mathcal{Z}(\theta)$, the debiasing series construction of McLeish (2011) and Glynn and Rhee (2014) described at the end of Section 3.5.2, was used to construct an unbiased estimate of the likelihood. The sequence of biased but consistent estimates of $1/\mathcal{Z}(\theta)$ was produced by taking the reciprocal of unbiased SMC estimates of $\mathcal{Z}(\theta)$ with an increasing number of importance samples and temperatures (see Moral et al. (2006) for a good introduction to SMC). SMC proceeds by defining a high-dimensional importance density which is sampled sequentially, and in this case we used a geometric schedule (Gelman and Meng, 1998; Neal, 2001) to define the sequence of distributions

$$p(\mathbf{y}|\theta)_n \propto p(\mathbf{y}|\theta)^{\phi_n} U(\mathbf{y})^{1-\phi_n},$$

with $0 \leq \phi_1 < \ldots \phi_p = 1$ and $U(\cdot)$ a uniform distribution over all the grids in $\mathcal{Y}$. A Gibbs transition kernel, in which one spin was randomly selected and updated according to its conditional distribution, was used to sequentially sample the high-dimensional space. The initial estimate, $1/\mathcal{Z}(\theta)_0$, used 100 temperatures and 100 importance samples; the i-th estimate used 100×2^i temperatures and importance samples.

The infinite series was truncated unbiasedly using both Poisson truncation and Russian Roulette. A geometric distribution was used as the stopping distribution in both cases with $p = 0.7$, chosen such that the variance of the log estimator was approximately 1 as suggested by Doucet et al. (2012). For comparison, the posterior distribution was also sampled using the Exchange algorithm, the approximate form of the Exchange algorithm (Caimo and Friel, 2011) with an auxiliary Gibbs sampler run for 50,000 steps at each iteration, and an 'exact' MCMC chain using the matrix transfer method to calculate the partition function at each iteration. All chains were run for 20,000 iterations and the second half of the samples used for Monte Carlo estimates.

The exact posterior mean and standard deviation are not available for comparison but the estimates from the five methods agree well (Table 3.2). The traces in Fig-



(a) Russian roulette **(b)** Poisson truncation **(c)** approximate Exchange

(d) Exchange algorithm **(e)** Exact MCMC chain

Figure 3.5: Traces of samples using the debiasing infinite series with (a) Russian Roulette, (b) Poisson truncation, and (c) the approximate Exchange algorithm (d) the Exchange algorithm using perfect samples and (e) an MCMC chain with the partition function calculated using the matrix transfer method. Note in (a) and (b) the samples are not drawn from the posterior distribution, $p(\beta|\mathbf{y})$, but from the (normalised) absolute value of the estimated density.

ure 3.5 show that the algorithms mix well and Figures 3.7 and 3.8 show that the estimates of the mean and standard deviation agree well. Estimates of the Effective sample size (ESS) are also included in Table 3.2, which give an idea of how many independent samples are obtained from each method per 10,000 samples.

Approximately 5% of estimates were negative when using roulette truncation and 10% when using Poisson truncation, however using the correction in Equation (3.8), expectations with respect to the posterior still converge to the correct values. If we had opted to implement the geometric series construction of Section 3.4.1 in order to reduce the number of negative estimates, we have available only a naive upper bound for the partition function corresponding to setting all spins to $+1$. This bound is very loose and therefore impractical, as the series converges very slowly. Hence the availability of a method to deal with negative estimates frees us from atrocious upper bounds that would explode the asymptotic variance of the chains.

	Roulette	Poisson	Exchange (approx)	Exchange (exact)	Exact
Mean	0.2004	0.2005	0.2013	0.2010	0.2008
Standard deviation	0.0625	0.0626	0.0626	0.0626	0.0625
ESS	2538	2660	1727	1732	3058
Relative CPU time	2.31	3.85	1	1.12	45.8

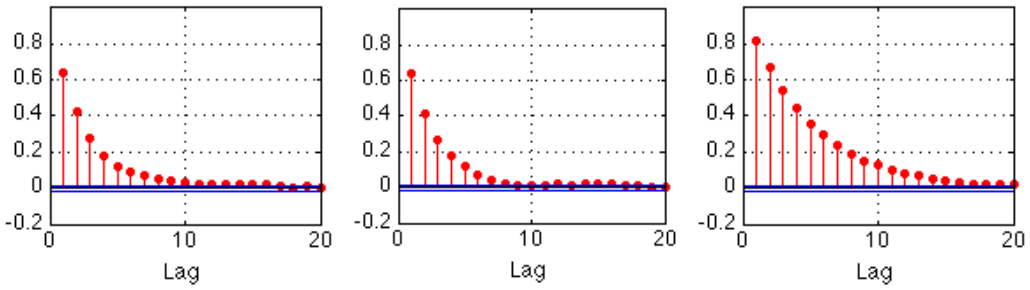
Table 3.2: Monte Carlo estimates of the mean and standard deviation of the posterior distribution $p(\beta|\mathbf{y})$ using the five algorithms described. The debiasing series estimates have been corrected for negative estimates. The exact chain was run for 100,000 iterations and then the second half of samples used to achieve a 'gold standard' estimate. An estimate of the effective sample size (ESS) is also shown based on 10,000 MCMC samples.

The autocorrelation functions (Figure 3.6) and the effective sample size (Table 3.2) of both Russian Roulette and Poisson truncation outperform the approximate and exact Exchange algorithm in this example and are comparable to the exact implementation; of course it is possible to improve the performance of our algorithm by using more computation, whereas this is not possible with the Exchange algorithm. It should be noted that the Exchange algorithm in this guise is less computationally intensive. However, it becomes impossible to perfectly sample as the size of the lattice increases, whereas our algorithm can still be implemented, albeit with considerable computational expense. Note that even at this small lattice size, the approximate version of Exchange looks noticeably less stable.

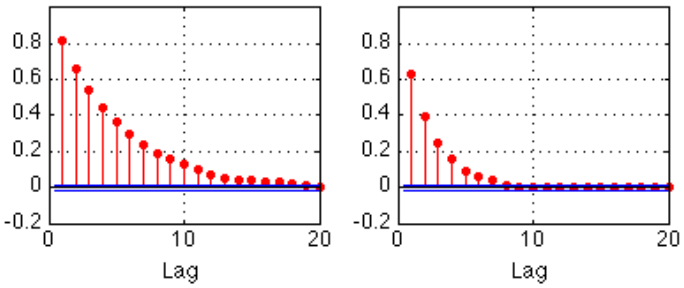
We have further experimented on larger lattices, for example both the Exchange algorithm and our methodology have been used to carry out inference over a 40×40 grid. At this size it is not possible to use the matrix transfer method to run an 'exact' chain. Sequential Monte Carlo (SMC) was used to estimate $\mathcal{Z}_i(\theta)$ at each iteration in the Roulette implementation. Again, the estimates of the means and the standard deviations from both methods agreed well. We have also carried out inference over a 60×60 grid, however it is no longer possible to perfectly sample at this size, particularly for parameter values near the critical value.

3.6.3 The Fisher-Bingham Distribution on a Sphere

The Fisher-Bingham distribution (Kent, 1982) is constructed by constraining a multivariate Gaussian vector to lie on the surface of a d-dimensional unit radius sphere, S_d. Its form is



(a) Russian roulette **(b)** Poisson truncation **(c)** approximate Exchange



(d) Exchange algorithm **(e)** Exact MCMC

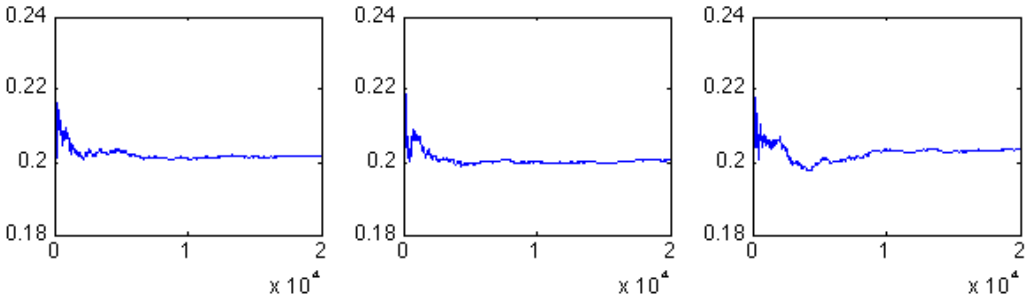
Figure 3.6: Autocorrelation plots for samples drawn from the posterior distribution $p(\beta|\mathbf{y})$ of a 10×10 Ising model using five methods: (a) Debiasing series with roulette truncation, (b) Debiasing series with Poisson truncation (c) approximate Exchange (d) the Exchange algorithm using perfect samples and (e) an MCMC chain with the partition function calculated using the matrix transfer method.

$$p(\mathbf{y}|\mathbf{A}) \propto \exp\{\mathbf{y}'\mathbf{A}\mathbf{y}\},$$

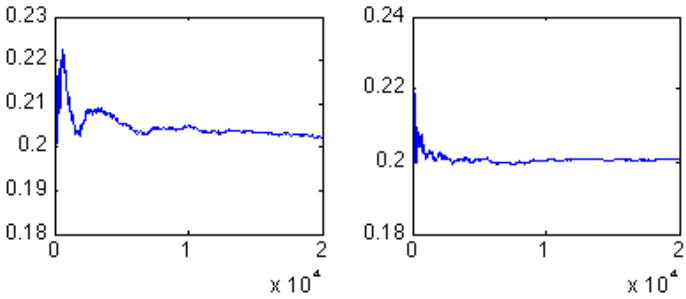
where $\mathbf{A}$ is a $d \times d$ symmetric matrix and, from here on, we take $d = 3$. After rotation to principal axes, $\mathbf{A}$ is diagonal and so the probability density can be written as

$$p(\mathbf{y}|\lambda) \propto \exp\left\{\sum_{i=1}^{d} \lambda_i y_i^2\right\}.$$

This is invariant under addition of a constant factor to each λ_i so for identifiability we take $0 = \lambda_1 \geq \lambda_2 \geq \lambda_3$. The normalising constant, $\mathcal{Z}(\lambda)$ is given by



(a) Russian roulette **(b)** Poisson truncation **(c)** approximate Exchange



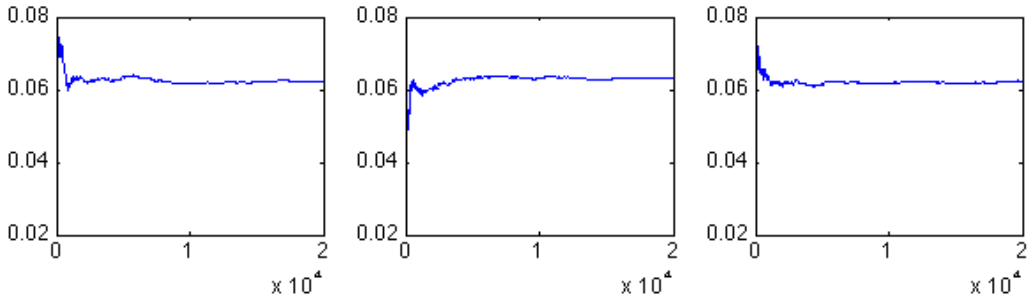
(d) Exchange algorithm **(e)** Exact MCMC

Figure 3.7: Plots of the running mean for the posterior distribution $p(\beta|\mathbf{y})$ of a 10×10 Ising model using three methods: (a) Debiasing series with roulette truncation, (b) Debiasing series with Poisson truncation (c) approximate Exchange (d) the Exchange algorithm using perfect samples and (e) an MCMC chain with the partition function calculated using the matrix transfer method.

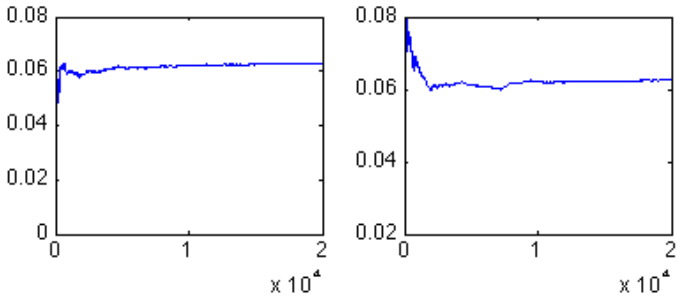
$$\mathcal{Z}(\lambda) = \int_S \exp\left\{\sum_{i=1}^{d} \lambda_i y_i^2\}\right\} \mu(\mathrm{d}\mathbf{y})$$

where $\mu(\mathrm{d}y)$ represents Hausdorff measure on the surface of a sphere. Very few papers have presented Bayesian posterior inference over the distribution due to the intractable nature of $\mathcal{Z}(\lambda)$. However in a recent paper, Walker uses an auxiliary variable method (Walker, 2011) outlined in Section 3.2 to sample from $p(\lambda|\mathbf{y})$. We can apply our version of the Exact-Approximate methodology as we can use importance sampling to get unbiased estimates of the normalising constant.

Twenty data points were simulated using an MCMC sampler with $\lambda = [0, 0, -2]$ and posterior inference was carried out by drawing samples from $p(\lambda_3|\mathbf{y})$ i.e. it was assumed $\lambda_1 = \lambda_2 = 0$. Our Exact-Approximate methodology was applied using the geometric construction with Russian Roulette truncation. A geometric distribution was used as the stopping distribution with $p = 0.7$, chosen such that the variance



(a) Russian roulette **(b)** Poisson truncation **(c)** approximate Exchange



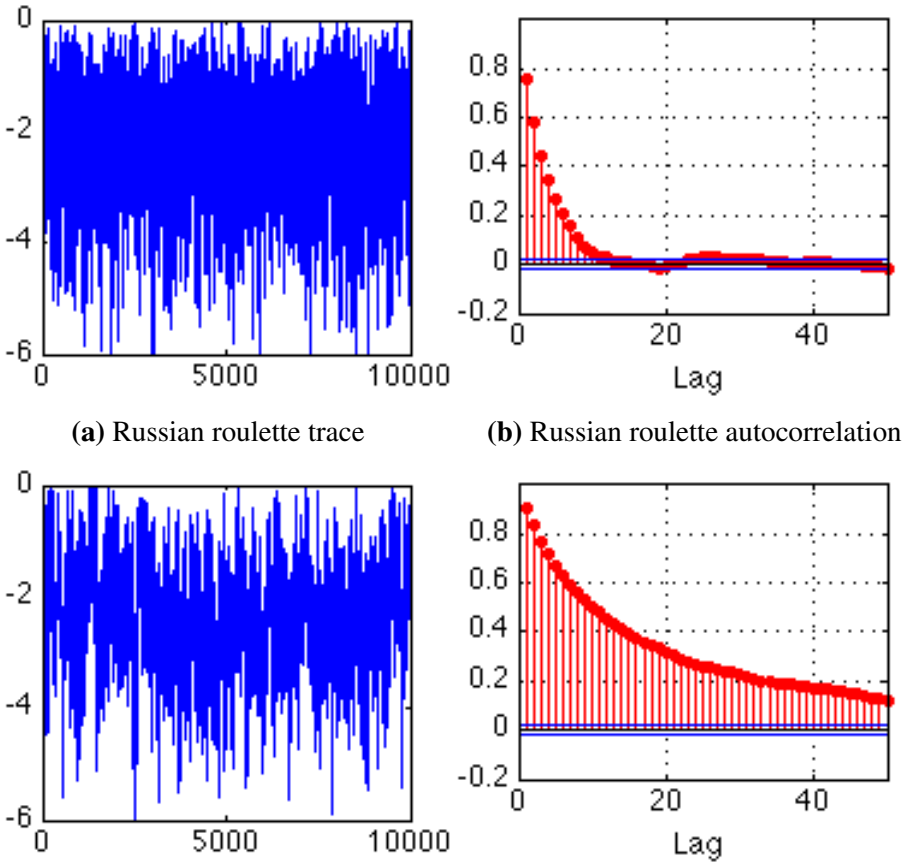
(d) Exchange algorithm **(e)** Exact MCMC

Figure 3.8: Plots of the running standard deviation for the posterior distribution $p(\beta|\mathbf{y})$ of a 10×10 Ising model using three methods: (a) Debiasing series with roulette truncation, (b) Debiasing series with Poisson truncation (c) approximate Exchange (d) the Exchange algorithm using perfect samples and (e) an MCMC chain with the partition function calculated using the matrix transfer method.

of the log estimator was approximately 1 as suggested by Doucet et al. (2012). A uniform distribution on the surface of a sphere was used to draw importance samples for the estimates of $\mathcal{Z}(\lambda)$. Prior to the MCMC run, estimates of $\mathcal{Z}(\lambda)$ were computed for a grid of points and an empirical upper bound set based on the mean and standard deviation of the estimates.

The proposal distribution for the parameters was Gaussian with mean given by the current value, a uniform prior on $[-5,0]$ was set over λ_3, and the chain was run for 20,000 iterations. Walker's auxiliary variable technique was also implemented for comparison using the same prior but with the chain run for 200,000 samples and then the chain thinned by taking every 10th sample to reduce strong autocorrelations between samples. In each case the final 10,000 samples were then used for Monte Carlo estimates.

In the Russian Roulette method, six negative estimates were observed in 10,000 estimates. The estimates of the mean and standard deviation of the posterior agree well (Table 3.3), however the effective sample size and autocorrelation of the Russian



(a) Russian roulette trace **(b)** Russian roulette autocorrelation

(c) Auxiliary variable trace **(d)** Auxiliary variable autocorrelation

Figure 3.9: Sample traces and autocorrelation plots for the Fisher-Bingham distribution for the geometric tilting with Russian Roulette truncation ((a) and (b)) and Walker's auxiliary variable method ((c) and (d)).)

	Roulette	Walker
Estimate of mean	-2.377	-2.334
Estimate of standard deviation	1.0622	1.024
ESS	1356	212
Relative CPU time	2.54	1

Table 3.3: Estimates of the posterior mean and standard deviation of the posterior distribution using roulette and Walker's method for the Fisher-Bingham distribution. An estimate of the effective sample size (ESS) is also shown based on 10,000 MCMC samples.

Roulette method are superior as seen in Figure 3.9. Note that it is also possible to get an upper bound on the importance sampling estimates for the Fisher-Bingham distribution. If we change our identifiability constraint to be $0 = \lambda_1 \leq \lambda_2 \leq \lambda_3$, we now have a convex sum in the exponent which can be maximised by giving unity weight to the largest λ i.e. $\sum_{i=1}^{d} \lambda_i y_i^2 < \lambda_{max}$. We can compute $\tilde{\mathcal{Z}}(\theta)$ as $1/N \sum_n \exp(\lambda_{max})/g(y_n)$, where $g(y)$ is the importance distribution.

3.7 Discussion and Conclusion

The capability to perform Pseudo-marginal MCMC on a class of doubly-intractable distributions has been established in this chapter. The methods described are not reliant on the ability to simulate exactly from the underlying model, only on the availability of unbiased or consistent estimates of the normalising term, which makes them applicable to a wider range of problems than has been the case to date.

The methodology is based on the stochastic truncation of a series expansion of the desired density. If the intractable likelihood is composed of a bounded function and non-analytic normalising term, then the proposed methodology can proceed to full MCMC with no further restriction. However, in the more general case, where an unbounded function forms the likelihood, then the almost sure guarantee of *positive* unbiased estimates is lost. The potential bias induced due to this lack of strict positivity is dealt with by adopting a scheme employed in the QCD literature where an absolute measure target distribution is used in the MCMC and the final Monte Carlo estimate is 'sign corrected' to ensure that expectations with respect to the posterior are preserved. What has been observed in the experimental evaluation is that, for the examples considered, the sign problem is not such a practical issue when the variance of the estimates of the normalising terms is well controlled and this has been achieved by employing Sequential Monte Carlo estimates in some of the examples. Hence one of the areas for future work is efficient estimators of the normalising term, which can be either unbiased or merely consistent. The inherent computational parallelism of the methodology, due to it only requiring a number of independent estimates of normalising constants, indicates that it should be possible to implement this form of inference on larger models than currently possible, however it is also clear that there is some limit to how much the method can be scaled up.

It has been shown (Jacob and Thiery, 2015) that it is not possible to realise strictly

positive estimates of the target distribution using the series expansions described in this paper, unless the estimates of the normalising term lie in a bounded interval. In its most general representation it is recognised that the sign problem is NP-hard implying that a practical and elegant solution may remain elusive for some time to come. However, other ideas from the literature, such as the absolute measure approach (Lin et al., 2000) can be used to tackle the sign problem.

The motivation for this methodology development was to use a Markov Random Field to model known dependencies in gene expression for the ARC syndrome data. In order for models on the scale of the number of genes to be analysed, the method needs to be optimised and in particular, techniques for generating low variance estimates of partition functions need to be developed. Further, the inherent parallel nature of the scheme must be more fully utilised to compute multiple estimates simultaneously. Methods such as the delayed acceptance scheme of Christen and Fox (2005) can also be used to improve efficiency. The methodology described in this chapter provides a general scheme with which Exact-Approximate MCMC for Bayesian inference can be deployed on a large class of statistical models, including those used to model dependencies in Systems Biology and other areas.

Chapter 4

Unbiased posterior estimation using ABC

4.1 Introduction

The previous chapter dealt with Bayesian inference for doubly-intractable distributions, models for which the likelihood normalising term was a function of the parameters and could not be computed. In this chapter we deal with a further level of intractability: models which can be simulated but for which the likelihood cannot be computed at all. This situation is common for many complex models in the Biological sciences, particularly in genetics (Siegmund et al., 2008), epidemiology (Blum and Tran, 2010) and population biology (Ratmann et al., 2007). As already mentioned in the Introduction, it is also the case for stochastic models of molecular interactions, which must be used in order to model all sources of variability. For these types of models, neither standard frequentist nor standard Bayesian techniques are available, and hence methods such as Approximate Bayesian Computation (ABC) have been developed. ABC requires only the ability to simulate from the data model, at the expense of introducing a bias into parameter estimation. Despite a growing amount of literature analysing and utilising ABC methods (see Beaumont (2010); Csilléry and Blum (2010); Marin et al. (2012) for recent reviews), this bias is not well characterised.

In this chapter, methodology is developed to enable unbiased likelihood-free Bayesian parameter estimation. The methodology utilises Monte Carlo estimates using samples from an ABC posterior. It also builds on work by Rhee and Glynn (2012) and McLeish (2011), analysed in detail by Agapiou et al. (2014). In these

papers, an unbiased infinite series estimator is constructed from consistent estimates and then unbiasedly truncated. This is a very similar approach to that of the previous chapter, but with the considerable advantage that only consistent estimates are required. The overall unbiased estimate can either be used directly to estimate expectations with respect to the posterior or in a Pseudo-marginal MCMC scheme. This novel approach allows unbiased parameter estimation with respect to the true posterior distribution, something previously not possible without making additional assumptions.

In the next section, stochastic models which are commonly used as a modelling tool in Systems Biology are introduced. These models are relevant to the ARC syndrome study as they will be used in a later stage of the analysis. They are also an example of models for which inference is difficult due to an intractable likelihood. In Section 4.3 Approximate Bayesian Computation is described in detail, including a review of some of the extensions to the method which have been published in recent years. In Section 4.4, a simple 'debiasing' method is described for obtaining an unbiased estimate when a sequence of consistent estimates are available, based on the work of Rhee and Glynn (2012) and McLeish (2011). In Section 4.5 these ideas are combined and developed to enable unbiased parameter and function estimates with respect to the true posterior distribution. The method is then illustrated on a simple example in Section 4.6 before being discussed and concluded in the final section.

4.2 Stochastic models in Systems Biology

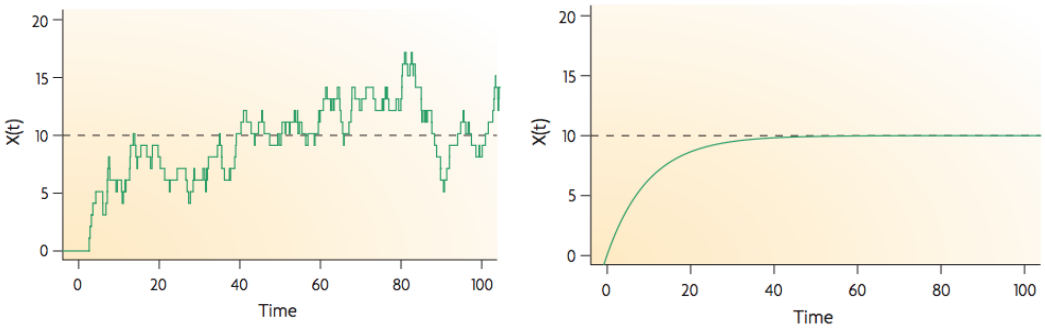
Thanks to modern experimental methods, most of the molecular building blocks of life have been recorded, so the challenge now is to explain how they interact to produce the versatile yet robust range of behaviours exhibited by living organisms. Interacting systems can exhibit a range of behaviours and so mathematical models are required to investigate the dynamics fully. Models of biological systems can operate at a range of scales, from the cellular level right up to the large-scale movement of animals; in this exposition, we focus on biochemical reaction kinetics at the single cell level.

The two main approaches to modelling biochemical kinetics are

1. Deterministic, continuous models using ordinary differential equations

2. Stochastic, discrete models in which state changes occur probabilistically.

In the first, it is assumed that all the reactants are abundant and that their concentration can therefore be measured on a continuous scale. If this assumption is met, and interest is only in the broad behaviour of the system, then this course of action can provide accurate solutions. If, on the other hand, molecular numbers are low and the behaviour is inherently stochastic, then the model will not reproduce the observed variability of the biological system, and the second approach is required. Figure 4.1 shows realisations of a simple reaction network in which one species, X, is produced and degraded at rates α and μ respectively. The deterministic simulation is clearly unrealistically smooth and simplistic.



(a) Stochastic **(b)** Deterministic

Figure 4.1: Realisations of (a) continuous deterministic and (b) discrete stochastic processes for a simple model in which a single protein is produced and degraded with rates α and μ respectively. Reproduced by permission from Macmillan Publishers Ltd, Nature Reviews Genetics, Wilkinson (2009).

In the second, probabilistic approach, numbers of molecules are discrete and changes occur randomly with probability determined by the current state. This is known as a *Markov jump process*. Consider a reaction in which two chemical species react to form a third

$$X + Y \to Z.$$

In a container of fixed volume and in a small time interval, dt, the probability of a reaction occurring is proportional to cxydt, for some constant c and where x and y are the number of molecules of X and Y respectively. This probability depends only on the current state of the system and hence satisfies the Markov property.

The constant, c, is a rate constant specific to the particular reaction. Changes to the system occur at random times, with probability also dependent on the current state.

An area of key importance and interest is statistical inference for reaction networks so that parameters such as rate constants and initial conditions can be inferred from experimental data. As already discussed, the Bayesian paradigm, in which prior beliefs, observed data and other sources of uncertainty can all be jointly modelled and propagated through to inferences made, is a particularly beneficial approach. However, as the likelihood has no simple tractable form, most attempts at inference have followed an *ad hoc* procedure in which parameter values are tuned to match experimental data (e.g. Arkin et al., 1998). New developments in Bayesian inference have allowed inference to be carried out when simulation methods are available (e.g. Boys et al., 2008; Golightly and Wilkinson, 2006, 2008). However, the algorithms are computationally intensive and generally require approximations to be introduced (Wilkinson, 2009). Approximate Bayesian Computation is one way to carry out Bayesian inference when no likelihood is available, and this is reviewed in the next section.

4.3 Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) methods have been developed over the last 15 years to deal with Bayesian inference problems in which the likelihood, $p(\mathbf{y}|\theta)$, is unavailable. The method was originally developed around fifteen years ago starting with the work of Marjoram et al. (2003); Tavaré et al. (1997); Pritchard et al. (1999).

Take a Bayesian inference problem in which we would like to find the expectation of the model parameters or some function of the model parameters with respect to the posterior distribution. The data $\mathbf{y} \in \mathcal{Y}$ is used to make posterior inferences about the variables $\theta \in \Theta$ which define the data density given by $p(\mathbf{y}|\theta)$. A prior distribution defined by $\pi(\theta)$ is adopted and the posterior is defined as

$$\pi(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)\pi(\theta).$$

The likelihood could be unavailable for one of several reasons. A process may be described by a generative model involving unknown parameters, for which it is not possible to write down or compute a likelihood e.g. stochastic kinetic models of chemical reactions (Wilkinson, 2009, 2007) or models of the domestication of plants and animal (Gerbault et al., 2014). It may not be available in closed form, as is sometimes the case when the likelihood is a marginal distribution of a higher-dimensional distribution

$$p(\mathbf{y}|\theta) = \int p(\mathbf{y}, \mathbf{u}|\theta) \mathrm{d}\mathbf{u},$$

where $\mathbf{u}$ is a latent vector. Often the dimensionality of $\mathbf{u}$ is so high that including these variables in the likelihood prohibits the use of standard Markov chain Monte Carlo approaches, but the integral above cannot be computed analytically, see for example inference for coalescent models in population genetics (Tavaré et al., 1997). Alternatively, the likelihood may have an intractable normalising term, $Z(\theta)$

$$p(\mathbf{y}|\theta) = \frac{f(\mathbf{y}|\theta)}{Z(\theta)},$$

which cannot be computed because the sum or integral of the unnormalised function $f(\mathbf{y}|\theta)$, over $\mathcal{Y}$ cannot be computed. This case was seen in the previous chapter.

Approximate Bayesian Computation has been developed as a method to carry out parameter inference when it is not possible to compute a likelihood but when a generative model is available. A simple implementation of the method is outlined in Algorithm 4, often referred to as rejection ABC. Values of θ' are simulated from the prior and then pseudo-data, $\mathbf{x}$, is simulated from the likelihood as a function of θ'. In the final step, the pseudo-data is compared to the real data via some summary statistics, and the proposed values of θ' are accepted or rejected based on a user defined distance, d, and threshold, ε.

If the summary statistics used to compare the true data to the simulated pseudo-data

Algorithm 4 Rejection ABC

for $n = 1$ to N **do**
  repeat
    Sample θ' from the prior $\pi(\cdot)$
    Generate $\mathbf{x}$ from the likelihood $p(\cdot|\theta')$
  until $d(\eta(\mathbf{x}), \eta(\mathbf{y})) < \varepsilon$
  Set $\theta_n = \theta'$.
end for

are sufficient for the parameters to be estimated or the identity (i.e. the distance is computed directly between the data and the pseudo-data), then as the threshold $\varepsilon \to 0$, ABC estimates tend to the true value (Barber et al., 2015). For some models, in which the data is discrete, ε can be set to zero and the method can be used to sample from the exact posterior. However, in most cases, this is not possible as the acceptance rate becomes prohibitively low, and this cannot be done when the data is continuous. A further complication is that sufficient statistics are generally not available so less informative summary statistics, $\eta(\mathbf{y})$, are often used instead, and much research has focused on how to choose these summary statistics (e.g. Fearnhead and Prangle, 2012; Beaumont et al., 2002; Nunes and Balding, 2010). In the majority of implementations ε is set to some non-zero value and samples are drawn from a distribution, $\pi(\theta|d(\eta(\mathbf{x}), \eta(\mathbf{y})) < \varepsilon)$, which it is hoped approximates the true posterior well.

The rejection ABC algorithm outlined above works well if areas of high density in the prior and posterior coincide, but this is often not the case. If, for example, non-informative priors are used, then most of the pseudo-data sets will be very different to the observed data, and the acceptance rate will be very low. For this reason, MCMC algorithms have been developed (Marjoram et al., 2003) which have considerably higher acceptance rates because the Markov chain spends the majority of its time in regions of high posterior probability. An implementation is outlined in Algorithm 5; the Markov chain has $\pi(\theta|d(\eta(\mathbf{x}), \eta(\mathbf{y})) < \varepsilon)$ as its stationary distribution and, as with all MCMC algorithms, produces correlated samples.

When $\varepsilon \neq 0$, ABC estimates are biased. The aim of this work is to combine these biased ABC estimates such that the bias is removed. The next section therefore, introduces a simple 'debiasing' method which does just this, albeit at the expense of added randomness.

Algorithm 5 MCMC ABC

Initialise θ_0
for $n = 1$ to N **do**
  Propose θ' from the proposal distribution $q(\cdot|\theta_{n-1})$
  Generate $\mathbf{x}$ from the likelihood $p(\cdot|\theta')$
  if $d(\eta(\mathbf{x}), \eta(\mathbf{y})) < \varepsilon$ **then**
    Set $\theta_n = \theta'$ with probability $\alpha = \min\left[1, \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}\right]$
  else
    Set $\theta_n = \theta_{n-1}$
  end if
end for

4.4 Unbiased estimation using biased estimates

In this section, a simple method is described for producing an unbiased estimate when only a sequence of biased but consistent estimators is available. The method has recently received attention in the literature (e.g. Agapiou et al., 2014; Glynn, 1984; Glynn and Rhee, 2014; Rhee and Glynn, 2013; McLeish, 2011; Strathmann et al., 2015) and stems from a method for unbiasedly estimating infinite sums used by von Neumann and Ulam in the 1950s.

The following exposition is based on results from Rhee and Glynn (2013). We would like to unbiasedly estimate the expectation of some random variable Y, $\mathrm{E}[Y]$, and we have available a sequence of approximations, Y_i, such that $\mathrm{E}[Y_i] \to \mathrm{E}[Y]$ as $i \to \infty$. Define the random variable

$$X = \sum_{i=0}^{\infty} Y_i - Y_{i-1}, \tag{4.1}$$

($Y_{-1} \triangleq 0$), and if the approximations are good enough that Fubini's theorem applies, then X is an unbiased estimate of $\mathrm{E}[Y]$. In this form, the estimate requires an infinite amount of computation, and therefore is not useful. To circumvent this problem, a non-negative integer-valued random variable, N, is introduced such that the random variable,

$$Z = \sum_{i=0}^{N} \frac{Y_i - Y_{i-1}}{P(N \geq i)}, \tag{4.2}$$

is a computationally realisable, unbiased estimate of $\mathrm{E}[Y]$. If condition (4.3) is met:

$$\sum_{i=1}^{\infty} \frac{\mathrm{E}[(Y_{i-1} - Y)^2]}{P(N \geq i)} < \infty, \tag{4.3}$$

then the estimator, Z, has finite variance and the expected value of Z^2 is

$$\mathrm{E}[Z^2] = \sum_{i=0}^{\infty} \frac{\mathrm{E}[(Y_{n-1} - Y)^2] - \mathrm{E}[(Y_n - Y)^2]}{P(N \geq i)}, \tag{4.4}$$

(see Rhee and Glynn (2013) for proof). If in addition, the estimator also has finite expected compute time, τ

$$\mathrm{E}[\tau] = \sum_{j=0}^{\infty} (\sum_{i=0}^{j} t_i) P(N = j) = \sum_{j=0}^{\infty} \bar{t}_j P(N \geq j),$$

where $\bar{t}_i$ is the incremental work required to compute Y_i, then a central limit theorem result holds as follows

$$c^{1/2}(\bar{\alpha}(c) - \mathrm{E}[Y]) \sim (\mathrm{E}[\tau]\mathrm{Var}(Z))^{1/2} \mathcal{N}(0,1),$$

where $\bar{\alpha}(c)$ is the estimator available after c units of computer time have been expended (Glynn and Whitt, 1992). This is desirable so that the estimator, Z, has the canonical square root convergence rate.

From the forms of Equations 4.1 and 4.1, it is clear that there is a trade-off when designing $P(N \geq n)$ to ensure that it decays fast enough to give finite and reasonable computing time, but slow enough to ensure the variance is finite. Note that only the finiteness of $\mathrm{Var}(Z)$ is required to build confidence intervals for the estimator Z.

Further results from Rhee and Glynn (2013) provide insight into designing the optimal probability distribution when both the variance and expected compute time are finite, by minimising the product of the variance and the expected work.

4.5 Unbiased estimation using ABC estimates

It is reasonable to assume that the scheme described above can be applied to estimates obtained using an ABC method. For any non-zero value of the tolerance, ε, estimates based on samples from the ABC posterior are biased, but the estimates converge in expectation to the true value as $\varepsilon \to 0$. These consistent estimates can therefore be combined as described in the previous section, to produce an unbiased estimator. Note that this is only the case when the summary statistics are either the identity (i.e. the distance d is directly between the data and the simulated pseudo-data) or sufficient for the parameters to be estimated. In the case where non-sufficient statistics are used, the estimation will be unbiased with respect to $\pi(\theta|d(\eta(\mathbf{x}),\eta(\mathbf{y})) < \varepsilon)$.

In the next following subsections, two methods for unbiased estimation with respect to the posterior distribution are described: unbiased rejection ABC and unbiased pseudo-marginal MCMC ABC.

4.5.1 Unbiased rejection ABC

This method combines the simple debiasing scheme of the previous section with the rejection ABC algorithm described in Algorithm 4. To implement the debiasing scheme, we need to design a sequence of estimators, $\{Y_i\}_{i=0}^{\infty}$, satisfying $\mathrm{E}[Y_i] \to \mathrm{E}[Y]$. The random variable Y is either the parameter of interest or some function of it, and expectations are with respect to the posterior distribution.

The sequence can be designed by setting each random variable, Y_i to be the mean of $f(i)$ samples from the ABC posterior $\pi(\theta|d(\eta(\mathbf{x}),\eta(\mathbf{y})) < \varepsilon_i)$, where the summary statistics are either the identity or sufficient. $f(i)$ is an increasing schedule for the

number of samples to be drawn from the i-th ABC posterior, defined by a tolerance ε_i which tends to zero as $i \to \infty$. Therefore, in the limit $i \to \infty$, an infinite number of samples are drawn from the true posterior, and Y_∞ is a constant equal to the expectation we wish to estimate.

To produce one unbiased estimate, the random variable, N, is drawn, estimates $\{Y_i\}_{i=0}^{N}$ are produced, and finally combined as in (4.2). An implementation to produce one unbiased estimate of $\mathrm{E}[\theta]$ is outlined in Algorithm 6, and this procedure can then be repeated M times and the estimates averaged to produce an unbiased estimator of lower variance.

Algorithm 6 Unbiased estimation with ABC

Design sequences $\{f(i)\}_{i=0}^{\infty}$ and $\{\varepsilon_i\}_{i=0}^{\infty}$ for the number of samples and value of ε at each level respectively.
Sample $k \sim P(N = k)$
for $j = 0$ to k **do**
  for $m = 1$ to $f(j)$ **do**
    repeat
      Sample θ' from the prior $\pi(\cdot)$
      Generate $\mathbf{x}$ from the likelihood $p(\cdot|\theta')$
    until $d(\eta(\mathbf{x}), \eta(\mathbf{y})) < \varepsilon_j$
    Set $\theta_m^{(j)} = \theta'$.
  end for
end for
Compute $Z = \sum_{j=0}^{k} \frac{Y_j - Y_{j-1}}{P(N \geq j)}$, where Y_j is the mean of $f(j)$ samples from the ABC posterior $\pi(\theta|d(\eta(\mathbf{y}), \eta(\mathbf{x})) < \varepsilon_j)$, with η either the identity or sufficient for θ.

Ideally this estimator would have both finite variance and finite expected compute time. In order for the variance to be finite, condition (4.3) must be met. To verify that this can be achieved, results from work by Barber et al. (2015) can be applied. They study the mean square error of Monte Carlo estimates based on samples from an ABC posterior as $\varepsilon \to 0$ and $n \to \infty$ where n is the number of Monte Carlo samples. Their results are therefore directly applicable as this is the function in the numerator of each term in (4.3).

They prove that for the optimal schedule of $\varepsilon_n \propto n^{-1/4}$, in the limit $\varepsilon \to 0$, the computational cost is $\propto n\varepsilon^{-q} = n^{(q+4)/4}$ and the mean square error is $\propto \mathrm{cost}^{-2/(q+4)} = n^{-1/2}$ where q is the dimension of the observation (or sufficient statistic, if used). Therefore, setting a sequence in which the number of samples doubles with each successive term, $f(i) = 2^i$, and setting the sequence ε_i according to the optimal

schedule, it is possible to design $P(N \geq i)$ such that the overall estimator, Z, has finite variance. For example, $P(N \geq i) = 2^{-pi}$ with $p < 1$ leads to the series in (4.3) being convergent.

Unfortunately, by comparing the conditions for finite variance and finite expected compute time, it is clear that both cannot be true. For the estimator to have finite variance, it must be true that

$$\sum_{i=1}^{\infty} \frac{\mathrm{E}[(Y_{i-1} - Y)^2]}{P(N \geq i)} = \sum_{i=1}^{\infty} \left(\frac{\varepsilon_{i-1}^q}{f(i\text{-}1)} \right)^{\frac{2}{q+4}} \frac{1}{P(N \geq i)} < \infty,$$

and if this is a convergent series then necessarily

$$\lim_{i \to \infty} \left(\frac{\varepsilon_{i-1}^q}{f(i\text{-}1)} \right)^{\frac{2}{q+4}} \frac{1}{P(N \geq i)} = 0.$$

Note that as $\lim_{i\to\infty} \frac{\varepsilon_{i-1}^q}{f(i\text{-}1)} < 1$ and $2/(q+4) < 1$, then also

$$\lim_{i \to \infty} \left(\frac{\varepsilon_{i-1}^q}{f(i\text{-}1)} \right) \frac{1}{P(N \geq i)} = 0. \tag{4.5}$$

Now looking at the form of the expected compute time, and using the expressions from Barber et al. (2015) with $\bar{t}_j$ denoting the incremental compute time for term j, we have

$$\begin{aligned} \mathrm{E}[\tau] = \sum_{j=0}^{\infty} \bar{t}_j P(N \geq j) &= \frac{f(0)}{\varepsilon_0^q} P(N \geq 0) + \sum_{j=1}^{\infty} \left(\frac{f(j)}{\varepsilon_j^q} - \frac{f(j\text{-}1)}{\varepsilon_{j-1}^q} \right) P(N \geq j) \\ &> \frac{f(0)}{\varepsilon_0^q} P(N \geq 0) + \sum_{j=1}^{\infty} \left(\frac{f(j)}{\varepsilon_j^q} P(N \geq j+1) - \frac{f(j\text{-}1)}{\varepsilon_{j-1}^q} P(N \geq j) \right). \end{aligned}$$

If the final series on the right is to be convergent, the sequence of its partial sums

$$\frac{f(0)}{\varepsilon_0^q}P(N \geq 0) + \sum_{j=1}^{m}\left(\frac{f(j)}{\varepsilon_j^q}P(N \geq j+1) - \frac{f(j\text{-}1)}{\varepsilon_{j-1}^q}P(N \geq j)\right) = \frac{f(m)}{\varepsilon_m^q}P(N \geq m+1)$$

should converge. Clearly comparing with (4.5) shows that, in fact, the partial sums diverge. When using ABC estimates in the debiasing series, if the variance is finite then the expected compute time is not. A similar result was found by Rhee and Glynn (2013) in the context of unbiased estimation for stochastic differential equations for which only discretisation schemes with a strong order greater than $1/2$ can have finite expected compute time. This unfortunately means that the central limit theorem results do not apply.

Of course, once a value for N has been sampled, the time taken to compute that estimate is finite, and the mean of k estimates will be unbiased even if the rate of convergence is not the canonical Monte Carlo rate of $k^{-1/2}$. Further, asymptotically valid confidence intervals can still be built along standard lines, however the estimator is likely to be computationally costly.

The scheme is a valid way to obtain unbiased estimates with respect to the true posterior distribution when the likelihood cannot be computed. However, it will inherit the disadvantages associated with the original ABC rejection scheme, namely that if the priors are uninformative or have mass in a different location to the mass of the posterior, then the rejection rate will be high. The simulation load will therefore be high whenever a high value of N is drawn. A simple way to reduce this impact is to sample from an importance distribution instead of from the prior and weight the samples accordingly as described in Fearnhead and Prangle (2012). This significantly increases the efficiency by proposing parameter values in regions of high posterior probability.

4.5.2 Pseudo-marginal MCMC ABC

As already described, MCMC ABC schemes can improve efficiency, as the Markov chain spends the majority of its time in regions of high posterior probability. A pseudo-marginal MCMC scheme (Andrieu and Roberts, 2009), in which an unbi-

ased estimate of the likelihood is substituted for the true value, can also be implemented here. The substituted estimates need only be unbiased with finite variance and do not require a central limit theorem result.

As in the previous subsection, a sequence of consistent estimates is required, but this time the estimates should be of the likelihood itself. To see how this may be achieved, note that the approximation to the likelihood used in ABC is as follows:

$$p^{\varepsilon}(\mathbf{y}|\theta) = \frac{1}{V(\varepsilon)} \int \mathbb{1}(y \in B_{\varepsilon,\mathbf{x}}) p(\mathbf{x}|\theta) \mathrm{d}\mathbf{x},$$

where $B_{r,\mathbf{z}}$ denotes a ball of radius r around $\mathbf{z}$, and $V(r) = \int \mathbb{1}(x \in B_{r,0}) \mathrm{d}\mathbf{x}$ is the volume of the ball centred at 0. An unbiased estimate of this artificial likelihood, $\widehat{p^{\varepsilon}(\mathbf{y}|\theta)}$, can be computed via Monte Carlo by generating pseudo-data according to the likelihood and assigning the sample a weight of 1 if the pseudo-data is within a ball of radius ε of the data, and 0 if not.

$$\widehat{p^{\varepsilon}(\mathbf{y}|\theta)} = \frac{1}{V(\varepsilon)} \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}(y \in B_{\varepsilon,\mathbf{x}}) \qquad \mathbf{x} \sim p(\cdot|\theta).$$

The value of $V(\varepsilon)$ is also required, however if Euclidean distance is used, then this is simply the volume of a hypersphere in N-space.

A sequence of converging likelihood estimates can therefore be constructed by designing a decreasing schedule for ε_i such that $\varepsilon_i \to 0$ as $i \to \infty$, and an increasing schedule for the number of samples. These can then be combined as previously described to produce an unbiased estimate of the likelihood, which can then be substituted in the Metropolis-Hasting acceptance ratio in order to sample from the posterior distribution. As was the case for the unbiased series estimator used in the doubly-intractable case, these unbiased estimates can be negative, and hence the absolute-measure MCMC methodology described in the previous chapter is again utilised. Algorithm 7 outlines the steps in the method.

A similar result is found here with regards to the variance and expected compute

Algorithm 7 Pseudo-marginal MCMC using ABC

Initialise θ_0
for $n = 1$ to N **do**
  Propose a value of $\theta' \sim q(\cdot|\theta_{n-1})$
  Sample $k \sim P(N = k)$
  for $i = 0$ to k **do**
    Generate $f(i)$ pseudo-data samples from the likelihood $p(\cdot|\theta')$
  end for
  Compute $\hat{f}(\mathbf{y}|\theta') = \sum_{i=0}^{k} \frac{Y_i - Y_{i-1}}{P(N \geq i)}$, where Y_i is the mean of $f(i)$ indicator variables with tolerance ε_i
    Set $\theta_n = \theta'$ with probability $\alpha = \min\left[1, \frac{|\hat{p}(\mathbf{y}|\theta')|\pi(\theta')q(\theta|\theta')}{|\hat{p}(\mathbf{y}|\theta)|\pi(\theta)q(\theta'|\theta)}\right]$
  Otherwise set $\theta_n = \theta_{n-1}$. Save θ_n, $|\hat{p}(\mathbf{y}|\theta_n)|$ and $\text{sign}(\hat{p}(\mathbf{y}|\theta_n))$
end for

time of the overall estimator. Schedules for ε_i, $f(i)$ and $P(N \geq i)$ can be found such that the variance of the estimator is finite. But if it is, then the expected compute time is not finite. Whilst this means that the method will be computationally intensive, it is still valid in that consistent expectations with respect to the posterior distribution can be computed. It is also expected that the method will be more efficient than the unbiased rejection ABC method, as simulated pseudo-data is more likely to be similar to the true data when the parameter value is in a region of high posterior probability.

4.5.3 Designing truncation distributions

It has been shown that stopping distributions can be designed such that the overall variance is finite, but that this will result in infinite expected compute time. Therefore, the key concern when designing stopping distributions is to reduce the computational time as much as possible. The results of Barber et al. (2015) can be used to inform the schedules for ε_i, $f(i)$ and $P(N \geq i)$. If, for simplicity, we choose a schedule in which the number of samples doubles for each estimate, then the optimal way to decrease ε so as to minimise the mean square error is $\sim 2^{-i/4}$. Then, as described in Section 4.5.1, the variance of the overall unbiased estimator is finite for $P(N \geq i) = 2^{-ip}$ with $p < 1$. So, in this case, it would be sensible to set p just below 1 in order to keep the computation as low as possible.

It should be noted, that as there is no requirement for the approximations to be independent, subsets of the pseudo-data simulated for the final series term can be re-used for the estimates of earlier terms in Russian roulette-type estimates, which

reduces computation considerably.

Due to the strong dependence of the computational cost of estimate $Y_n^{\varepsilon_n}$ on the dimension of the data, $\mathrm{E}[\text{cost } Y_n^{\varepsilon_n}] \sim n\varepsilon_n^q$, where n is the number of samples, the method is likely to work best in cases where the size of the dataset is small. However, the methodology can be easily parallelised, either by computing unbiased estimates in parallel and averaging over them afterwards, or by simulating the many independent pseudo-datasets required at a specific value of θ' in parallel.

4.5.4 Debiasing with low dimensional sufficient statistics

It is very rare in practice to have sufficient statistics available for complex models, so this subsection is included more for interest than practicality. If sufficient statistics, $T(\mathbf{y})$, are available, then due to the factorisation theorem the posterior depends on the data only through the sufficient statistics (e.g. Rice, 2006, Chapter 8). The posterior may be computed or sampled directly in terms of $p(T(\mathbf{y})|\theta)$ so that $p(\theta|\mathbf{y}) = p(\theta|T(\mathbf{y})) \propto p(T(\mathbf{y})|\theta)\pi(\theta)$. Therefore, both of the debiasing algorithms described above can be implemented with a distance between $T(\mathbf{y}_{obs})$ and $T(\mathbf{x})$ substituted for one directly between $\mathbf{y}_{obs}$ and $\mathbf{x}$, and this will allow unbiased (for Algorithm 6) or consistent (Algorithm 7) estimation with respect to the posterior distribution. In the case of the unbiased rejection algorithm, Barber et al. (2015) showed that the computation required to draw one sample from the ABC posterior as $\varepsilon \to 0$ is ε^{-q} where q is the dimension of the data or sufficient statistic if used. Therefore, the use of lower dimensional statistics to summarise the data will lead to significant computational savings.

4.6 Experimental validation

4.6.1 Toy example

The methodology is demonstrated on a toy example from Murray et al. (2006) in which the posterior distribution and relevant expectations of the parameter are known analytically. The data consists of M i.i.d. data points from a zero-mean Gaussian $p(y_i|\theta) = \mathcal{N}(0, 1/\theta)$ with unknown precision, θ. A conjugate prior $\pi(\theta|\alpha,\beta) = \text{Gamma}(\alpha,\beta)$ is set over θ. The posterior is given by $p(\theta|\mathbf{y}) = \text{Gamma}(\alpha + M/2, \beta + \sum_i y_i^2/2)$. We pretend that we cannot compute the likelihood but that we can simulate from it. Two datasets are used, $N = 1$, $y = 1$ as in

Murray et al. (2006) and a dataset with $N = 10$ in which independent data points are simulated $\sim \mathcal{N}(0, 1/1.5)$.

We aim to compute the expectation and standard deviation of the posterior distribution, $\pi(\theta|\mathbf{y})$, which can be computed analytically. Both schemes are implemented: unbiased ABC rejection and pseudo-marginal MCMC. The schedule for the number of Monte Carlo samples in each estimate was set as $f(i) \propto 2^i$. The schedule for ε_i was set as $\varepsilon_i = 2^{-i/4}$. The probability distribution was designed such that the estimator had finite variance, $P(N \geq i) = 2^{-im}$, with $m < 1$. For the Markov chain, a normal proposal distribution was used with acceptance rate tuned to 30%.

Figure 4.2 shows the running mean (top row) and standard deviation (bottom row) estimates for the unbiased ABC rejection algorithm. Each individual estimate is independent, and the standard error of the mean is used to indicate the uncertainty. Clearly the estimates converge to the correct values, although as expected the estimate of the standard deviation takes longer to converge than the estimate of the mean. The running time was 40 seconds for 10,000 unbiased estimates.

Figure 4.3 shows running estimates of the mean (left) and standard deviation (right) for the Pseudo-marginal unbiased ABC algorithm. As the unbiased estimates can be negative ($\sim$20% were negative), and the samples were not drawn from the true posterior, the estimates of the running mean and running standard deviation were computed using the sign corrected formulae from the previous chapter. A simple estimate of the uncertainty was based on the standard error of the mean using effective sample size rather than the iteration number. The trace of the MCMC samples and the autocorrelation are shown on the bottom row. The estimates clearly converge to the correct values, the trace shows little sticking and the autocorrelation between samples is low. It took 50 minutes to complete 500,000 samples, so in comparison to the unbiased ABC rejection, each sample can be produced much more quickly, but as the overall Markov chain converges slowly, many more samples are required.

For the $N = 10$ analysis, both sets of samples were produced using parallel computation to reduce the running time. For the unbiased ABC rejection algorithm, simulation can easily be spread across multiple compute nodes as each node can independently simulate from the prior, simulate corresponding pseudo-data and compare the pseudo-data to the true data. These results can then be sent back to a central node and combined to produce unbiased estimates. For the unbiased MCMC ver-

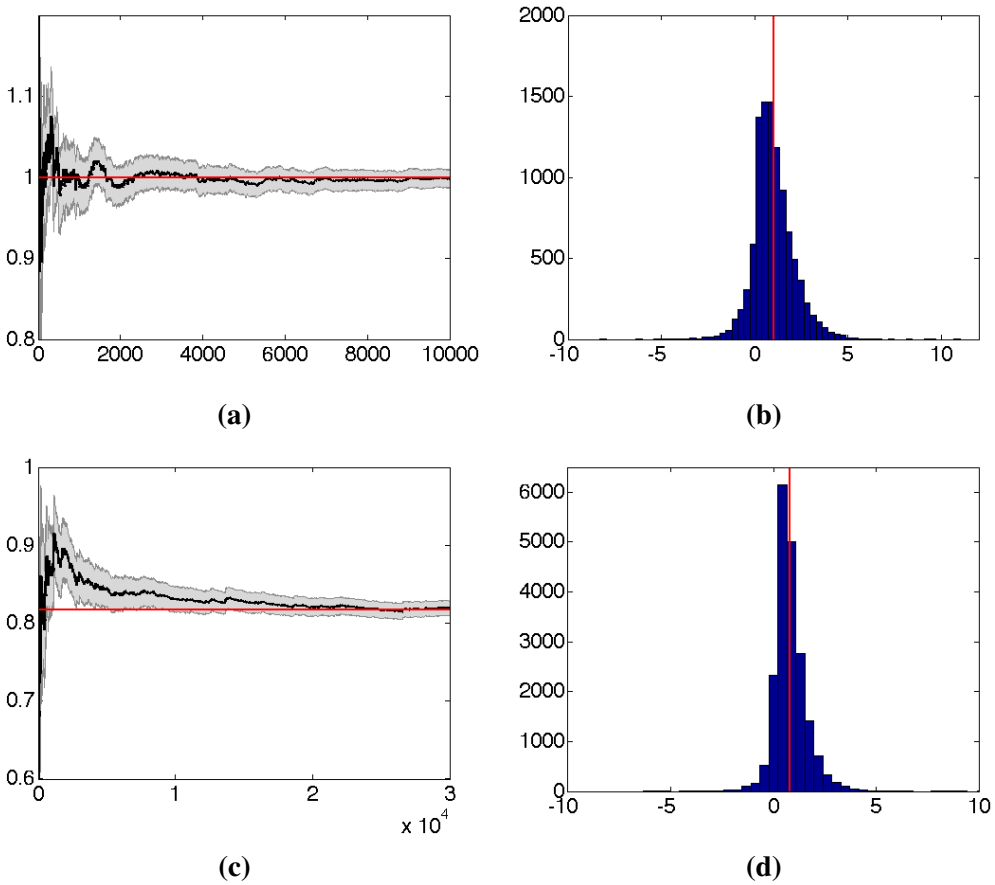


Figure 4.2: Running estimates of (a) the mean and (c) standard deviation of the toy example posterior with $N = 1$ using unbiased rejection ABC. Error estimates are based on the distribution of sample means and standard deviations. (b) and (d) give histograms of the

sion, simulation of pseudo-data can also be shared across many nodes once a value has been proposed for θ'. In each case 20 nodes were used to parallelise computation and the algorithms ran for 3hrs (unbiased ABC rejection) and 1hr 40mins (Pseudo-marginal).

Results for the $N = 10$ example are shown in Figure 4.4. The MCMC chain mixes well, although a small amount of sticking is seen, as is to be expected from a Pseudo-marginal chain. On the higher-dimensional example, the Pseudo-marginal algorithm is considerably more efficient. This is because the chain spends the majority of its time in regions of high posterior probability, and therefore the generated pseudo-data has a higher probability of being accepted, so overall fewer pseudo-data simulations are required. In the unbiased rejection algorithm, many of the prior samples are rejected because the posterior mass becomes more concentrated when additional data is available. This discrepancy in running time depends strongly on

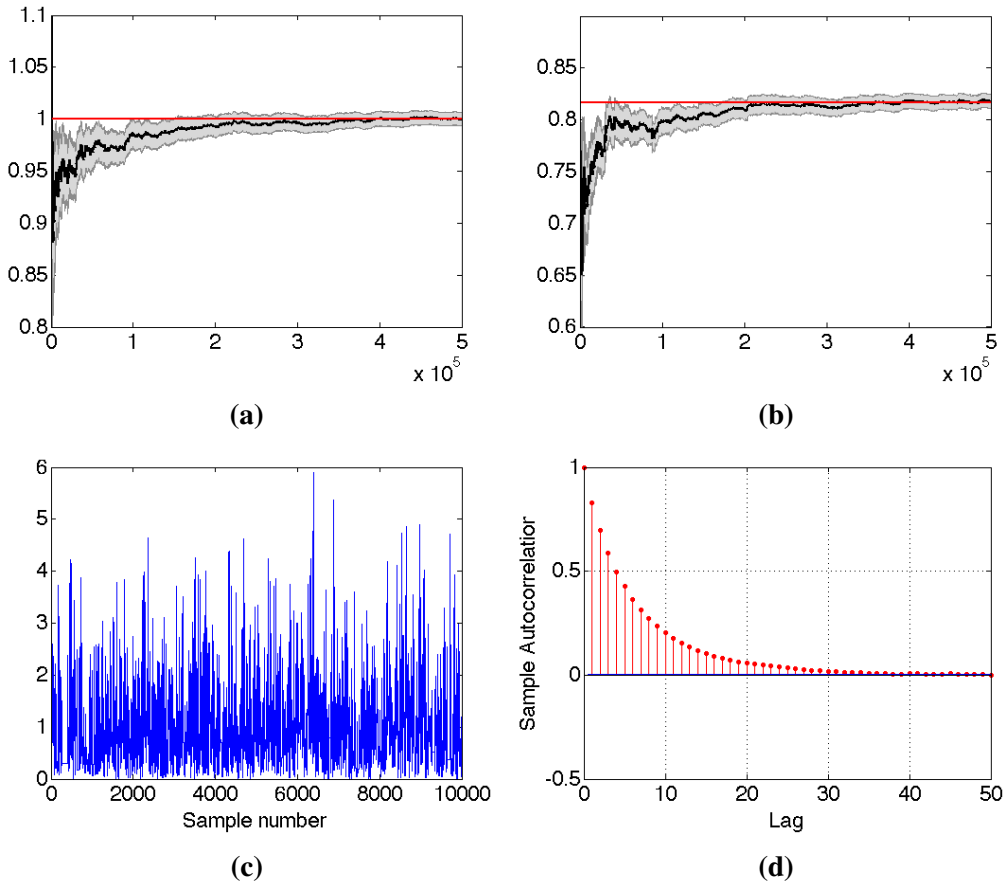


Figure 4.3: Running estimates of (a) the mean and (b) the standard deviation of the toy example posterior with $N = 1$ using pseudo-marginal MCMC ABC. Error estimates are based on the distribution of sample means and standard deviations using an estimate of the effective sample size as the number of samples. (c) shows the trace and (d) shows the autocorrelation between MCMC samples.

the shape of the prior compared to the posterior, as the pseudo-data has a higher chance of being accepted if the prior and posterior distributions have a similar scale and location.

4.6.2 Simple molecular system

Consider a reversible dimerisation reaction in which two molecules, A and B, react to form a dimer AB. k_1 and k_2 are the rate constant associated with dimerisation and disassociation respectively. This is a simple example of a model for which data can be simulated, but no likelihood can be computed.

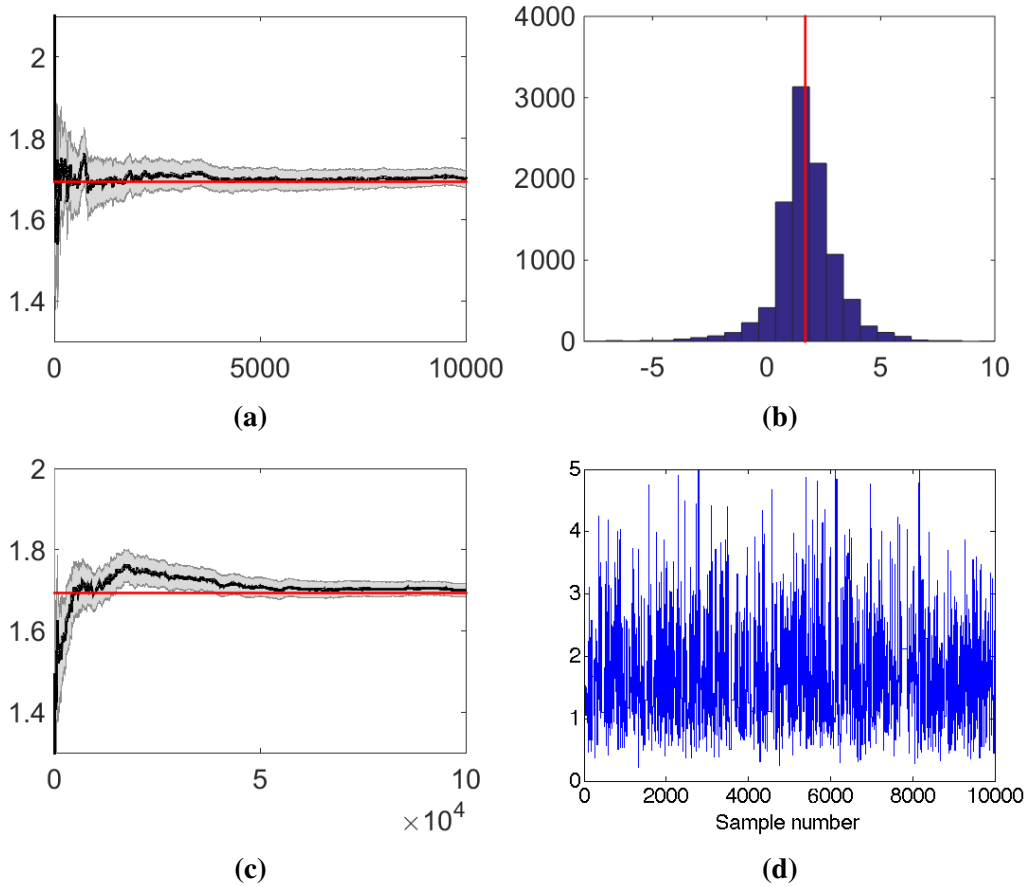


Figure 4.4: Running estimates of the mean of the toy example posterior with $N = 10$ using independent debiased estimates (top row) and pseudo-marginal MCMC ABC (bottom row). Error estimates are based on the distribution of sample mean using either the number of samples for (a) or an estimate of the effective sample size as the number of samples for (c). (b) shows a histogram of the debiased estimates and (d) shows the MCMC trace.

$$\mathrm{A} + \mathrm{B} \underset{k_2}{\overset{k_1}{\rightleftharpoons}} \mathrm{AB}$$

Data was simulated for measurements of A and AB with initial values, $\mathrm{A}_0 = \mathrm{B}_0 = 20$, $\mathrm{AB} = 0$ and $k_1 = 1.5$ and $k_2 = 1$. As the molecule numbers are low, the fluctuations are relatively large as can be seen in Figure 4.5, where the data points are denoted by large circles. We now proceed to carry out Bayesian posterior inference for the rate parameters k_1 and k_2.

Gamma priors were set over the rate constants with shape parameter $\alpha = 2$ and inverse scale parameter $\beta = 2$. A Pseudo-marginal MCMC algorithm was run as

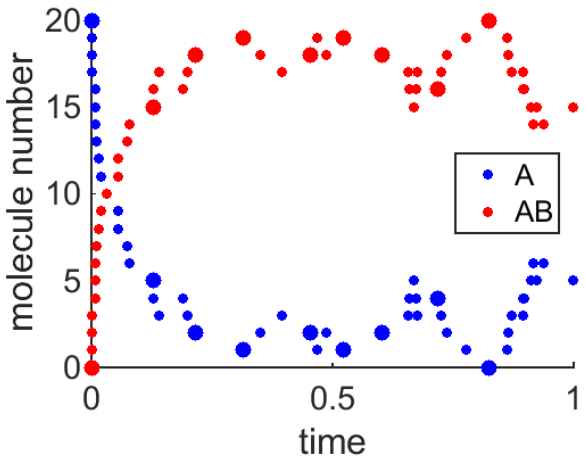


Figure 4.5: Data simulated for the simple dimerisation model using the Gillespie algorithm. Eight data points were used for species A and AB, shown as the large circles.

described in Section 4.5.2. The schedule for ε_i was set as $\varepsilon_i = e/(1+i\xi)$ with $e = 3$ and $\xi = 0.02$. The number of samples was doubled for each successive Monte Carlo ABC estimate. A pre-run was used to estimate the numerator of each term in the finite variance condition (4.3), and the probability distribution $P(N \geq i)$ was then set so as to ensure the series was convergent. The MCMC proposal distribution was a normal distribution with mean as the current value and standard deviation set to achieve an acceptance rate of 25%. At each iteration, simulations of pseudo-data were performed in parallel spread over 20 cores and the overall running time to produce 30,000 samples was 2hrs.

Results for the Pseudo-marginal MCMC run are shown in Figure 4.6. No obvious sticking is seen in the chain, but the samples had quite high autocorrelations so the chain was run for 30,000 iterations and then thinned by taking every third sample. All plots were produced with the thinned samples. The running posterior mean of the rate constants k_1 and k_2 are shown on the top row, as expected these are close to the simulation values, but not exactly the same as this depends on the specific data simulated. Error estimates in grey are based on the standard error of the mean using the iteration number corrected for effective sample size. Standard rejection ABC was also used to analyse the data, with a large number of samples and a small value of $\varepsilon = 2$, and the posterior means were very similar.

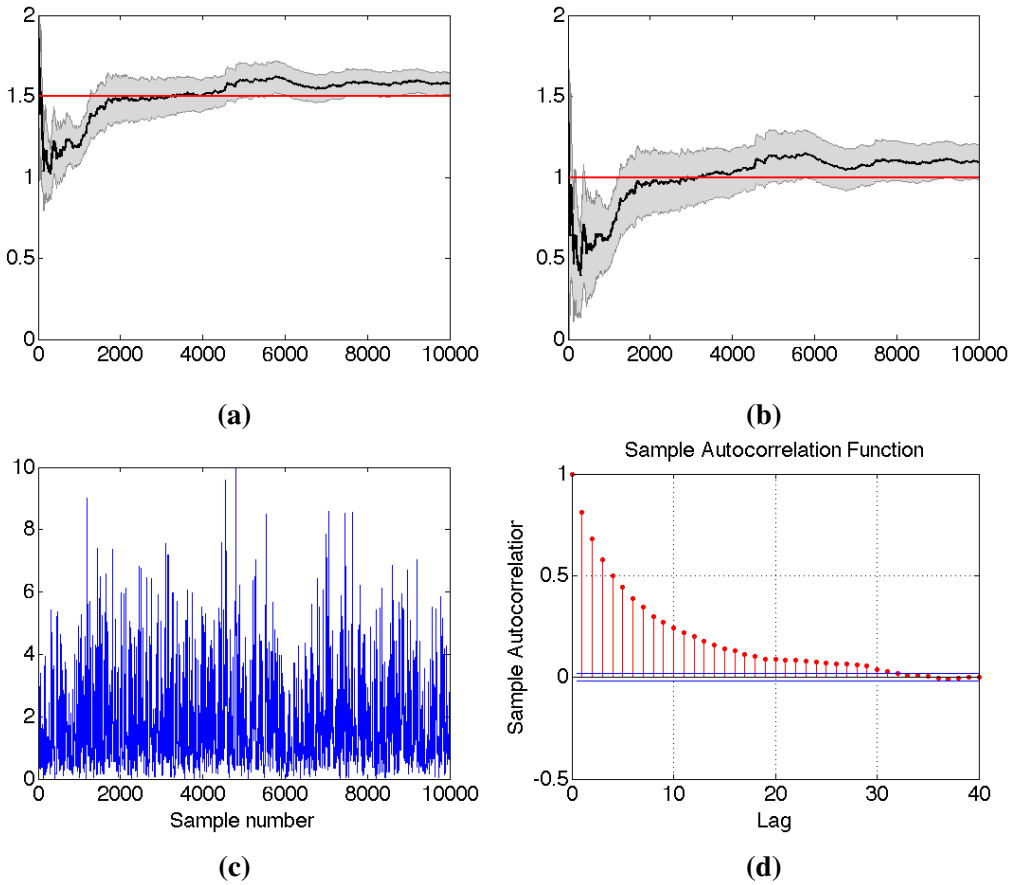


Figure 4.6: (a) and (b) Running mean for parameters k_1 and k_2 for stochastic dimerisation model. Note that the red lines denote the values used to simulate the data, not the true mean of the posterior. (c) MCMC trace for the Pseudo-marginal samples. (d) Autocorrelation for MCMC samples.

4.7 Discussion and Conclusions

In this chapter, a method has been developed to enable the unbiased estimation of parameters or functions with respect to the Bayesian posterior distribution when a likelihood is unavailable. The approach only requires the ability to simulate from the likelihood, and therefore can be implemented in situations which previously only had ABC methods available. In all work published to date, Bayesian parameter estimation in the likelihood-free context has always introduced some bias, unless further assumptions are made (for example Wilkinson (2013)'s work showed that ABC gives unbiased results under the assumption of a uniform additive model error). The work in this chapter shows that a theoretically valid method is available to unbiasedly estimate parameter values with respect to the posterior distribution without making this assumption.

The methodology is based on the stochastic truncation of an infinite series, constructed from a sequence of biased but consistent estimators. This sequence of biased estimates can be formulated using a sequence of Monte Carlo ABC estimates with an increasing schedule for the number of Monte Carlo samples and a decreasing schedule for ε. The infinite series is then constructed from the sum of the differences between successive estimates such that the expectation of the series is the required expectation which cannot be simulated (Rhee and Glynn, 2013). These unbiased estimates can either be used directly to estimate posterior expectations, or used in a Pseudo-marginal MCMC scheme. The MCMC scheme is preferable in terms of efficiency as the main computational cost comes from generating pseudo-data which matches the observed data, and this is more likely for parameter values of high posterior probability. However, as the unbiased estimates cannot be guaranteed to be positive, the absolute measure approach described in the previous chapter needs to be implemented.

The main drawback to the scheme, as shown in this Chapter, is that if the variance of the unbiased estimator is finite, then the expected compute time is not. This means that the central limit results derived in Glynn and Whitt (1992) do not apply, and the method will have high compute time. However, as long as the variance is finite, theoretically valid confidence intervals can still be computed based on standard asymptotic arguments to give an indication of its variability. The Pseudo-marginal scheme does not require anything other than the unbiasedness property for its theoretical validity (Andrieu and Roberts, 2009).

Models for which no likelihood is available are common in Systems Biology, with examples including stochastic models of chemical reactions (Wilkinson, 2009), genetic evolution (Ratmann et al., 2007), population genetics (Tavaré et al., 1997) and parasitology (Drovandi and Pettitt, 2011). This method therefore represents an important step forward in allowing unbiased estimation for many models where previously an uncharacterised bias was unavoidable.

The method has been demonstrated on a two toy examples, although as expected the computational load was high. The most important part of further work is to optimise the procedure to reduce computation. The choice of schedules for ε and number of Monte Carlo samples are all left to the practitioner and therefore can be optimised. It has also been shown that ABC SMC (Toni et al., 2009; Moral et al., 2012) can be used to draw samples from the ABC posterior more efficiently than naively simulating from it. In this technique a sequence of target distributions

is designed and sampled successively. This could therefore form the basis of the consistent estimate in each series term.

Chapter 5

General Conclusions

The research presented in this PhD thesis focuses on statistical analysis and methodology for Systems Biology. High-throughput data is analysed to extract information about the pathogenesis of ARC syndrome and methodology is developed to enable Bayesian inference for intractable models used to describe complex, dynamic systems. Both of these are key parts in a systems analysis of a genetic disease.

In the second chapter, transcriptomic and metabolomic data were analysed to gain insight into ARC syndrome, a rare autosomal genetic disorder. Three different knock-down cell lines were analysed and compared to controls. It was confirmed that the transcriptomes of the three knock-downs were very similar, a finding which fits with experimental evidence that the three genes, VPS33B, VIPAR and PLOD3 are involved in the same process. Genes/metabolites with differential expression and interaction profiles were identified and found to be largely involved in membrane and trafficking processes. Over-represented Gene Ontology analysis found that annotations pertaining to cell-cell adhesion and junction complexes were most over-represented compared to the background gene set. The analysis is validated by microscopy experiments which have shown that cell-cell adhesion is disrupted and that the IMCD-3 cells no longer form a cohesive epithelial layer. The suggestions for future experimental work, such as investigating the role of Mal2 in epithelial polarisation and examining changes in Integrin interactions, are therefore extremely valuable.

The detection of differentially expressed genes was implemented using only mRNA expression levels, and the idea of using a Markov Random Field (MRF) to model known genetic dependencies motivated the methodology development of the next

chapter. MRFs are doubly-intractable, meaning the likelihood normalising term is a function of the unknown parameters and cannot be computed. This makes Bayesian inference difficult, a problem which is encountered in a multitude of modelling scenarios from social networks to modelling of disease outbreaks.

The approach developed used Pseudo-marginal MCMC and hence required an unbiased estimate of the likelihood. This could not be done simply using Monte Carlo as an estimate of the *reciprocal* of the normalising term was required. A series construction was implemented in which multiple unbiased estimates of the normalising term were combined to produce an unbiased estimate of the reciprocal of the normalising term. As the estimates cannot be guaranteed to be positive without the availability of a bound on the normalising term, a weighted sum of estimates was used to ensure expectations with respect to the posterior could still be realised. The method was demonstrated on several examples, including Ising models and the Fisher-Bingham distribution. This is a general contribution to Bayesian methodology and allows posterior inference to be carried out even when perfect simulation from the model is not available.

Future work involves scaling up the methodology and implementing it on more realistic examples. Central to achieving this goal is developing methods for estimating normalising terms, as well as fully utilising the parallel nature of the algorithm in computation. The largest model on which the methodology was demonstrated was an Ising model of 3,600 spins, and this was implemented by computing unbiased estimates in parallel at each MCMC iteration. Whilst this dataset is close to the order of magnitude required for analysis of mRNA expression levels, applying the methodology to a more complex MRF modelling genetic dependencies will require a combination of parallel computation and state-of-the-art Sequential Monte Carlo estimation. It will also be prudent to make use of other developments in the MCMC literature, such as adaptive proposals (Roberts and Rosenthal, 2007) and delayed acceptance (Christen and Fox, 2005), to reduce the computation as much as possible.

It is crucial to use stochastic models to characterise all the sources of variability in biological systems. Indeed, once further experimental results including time-course data are available for the molecular species involved in ARC syndrome, the aim would be to build a stochastic model to investigate and simulate its behaviour. However, the price to pay for the use of realistic, stochastic models is increased difficulty in fitting the models to data. The final contribution of this thesis, therefore,

concentrated on unbiased inference for models with no tractable likelihood. This situation is common across the sciences where complex generative models can often be defined but the likelihood cannot be easily computed. In particular, the biological sciences have produced many models where likelihoods cannot be computed, for example population genetic models, epidemiological models, or agent-based models in ecology.

The aim of this chapter was to develop methodology for unbiased Bayesian parameter estimation when the likelihood cannot be computed. The approach taken was similar to that in the doubly-intractable case, except this time the infinite series used was based on the work of Rhee and Glynn (2013) as only consistent (not unbiased) estimates were available. It was shown that these consistent ABC estimates, which use a decreasing schedule of ε and an increasing number of Monte Carlo samples, can be combined to produce an unbiased estimate with finite variance. However, it was also shown that if the variance is finite, then the expected compute time is not. Unbiased estimates with respect to the posterior can therefore be produced for stochastic models where previously some bias had to be accepted, albeit with the acceptance of a high computational load. Reducing this computational cost by optimising the user-specified parts of the method, and generating the Monte Carlo estimates as efficiently as possible are the key areas for future work.

In conclusion, this thesis has made contributions to the area of Systems Biology and Statistical methodology through the analysis and interpretation of data relating to ARC syndrome and the development of methodology for Bayesian inference of commonly used models with intractable likelihoods. The methods have been investigated empirically and theoretically on multiple examples. Ultimately, the key area for future work is the application of the methods to large-scale real-world problems.

Appendix A

Top 100 PCA loadings for transcriptomic data for Principal Components 1 and 2

Gene	PC 1	PC 2
Tmem54	-0.9931513	0.0182024
Esrp2	-0.9927839	0.01846715
Sema5a	-0.987471	0.00755073
9930014A18Rik	-0.9861582	-0.0313508
Fhdc1	-0.9859477	-0.0388543
Fam84b	-0.9858694	-0.0205286
Btd	0.98449204	0.06327265
Fermt1	-0.9843311	0.08087599
Snhg18	-0.9832074	0.06438172
Cldn4	-0.9829758	-0.030578
Il18r1	-0.9826365	0.06004387
Hspbap1	-0.9821981	0.07215234
Dapk1	-0.9816951	0.12889227
Tom1l1	-0.9811454	-0.1818591
Fam198b	-0.9808098	-0.0124572
Exoc2	0.98055516	-0.0269386
Rad51c	0.980277	0.06377451
Pxk	0.98011253	-0.0068555

Continued on next page

Gene	PC 1	PC 2
Cldn8	-0.9801095	0.05194399
Prim1	0.98006869	-0.0227844
Esrp1	-0.980058	0.10675977
Mctp2	-0.9798622	-0.1289667
Peg3	-0.9795938	-0.0087553
Igf2bp3	-0.9790267	-0.1057278
Tfrc	-0.9786331	0.06099299
Atp8b1	-0.9782377	0.11606252
Limch1	-0.9779123	-0.0638448
Tmsb4x	-0.9778612	0.04164798
Aldh3b1	0.97680755	-0.1782945
Slc25a17	0.97660151	-0.0237803
Ano9	-0.9762801	0.09946054
Nol7	0.9762221	0.01135309
Adssl1	0.97616967	0.00757651
Fam132a	-0.9752967	0.02541575
Marveld3	-0.9751795	0.08197014
Noa1	-0.9747917	0.026807
Pole2	0.97469755	0.06337233
Sim2	-0.9745076	-0.0119788
Cpn1	0.97412192	-0.0835262
Epha1	-0.9739234	0.11725178
Mfsd12	0.97372093	0.17203841
Lypd6b	-0.973235	-0.0136111
Faxc	-0.9731959	-0.1148237
Tmem184a	-0.9730842	0.11936074
Etl4	-0.9727828	-0.0314996
Nipal2	-0.9725023	0.12816219
Cyp2s1	0.9724902	-0.0282878
Phf19	0.97245229	-0.0525579
Nolc1	-0.9722985	0.00374698
Obfc1	0.97212305	-0.0902529
Elfn1	0.97197699	-0.0609808
Ripk4	-0.9719298	0.04252755

Continued on next page

Gene	PC 1	PC 2
Ica1	-0.9715738	-0.1625626
Zfp72	-0.9707526	0.03359738
Zak	-0.9702061	-0.0082185
Snta1	0.96990495	-0.0590663
Spdya	0.96962669	0.1296889
Evc2	-0.9695774	0.17541329
Igsf5	-0.9692433	-0.1427075
Cdh1	-0.9687867	0.185722
Fut11	0.96850057	0.14306967
Tmem39a	-0.9684142	-0.0091815
Mettl7a1	0.96828986	0.03673417
Egln3	0.96747923	0.03681717
Lonrf3	0.96744486	0.14691301
Mxd3	0.96718427	0.06619411
Tspan17	0.96712937	-0.1411708
Palm	0.96702226	0.06184463
Cerk	0.96698977	-0.0391274
Kremen1	-0.9669527	0.00110735
Grhl2	-0.9667478	0.18045015
Lgals8	0.96668103	-0.0030784
Mkl2	-0.9664827	0.04349228
Zfp141	-0.9664672	0.07133049
Lman2	0.9661122	0.06430423
Ngly1	0.96592995	0.08012644
Aldh7a1	-0.965575	-0.1331071
Efna2	0.96534501	0.06586912
Tgm2	0.96500224	-0.0047286
Lbp	0.96462453	0.19427303
Polr1b	-0.9644367	-0.0759986
Mum1l1	0.96440604	0.06391846
Adamts5	-0.9643612	-0.0565225
Tcta	0.96414569	-0.0560248
Igsf9	-0.9640769	-0.1396267
Marveld1	0.96398097	0.19587427

Continued on next page

Gene	PC 1	PC 2
Abcd4	0.96372524	-0.0800626
Idnk	0.96366777	-0.0001867
Arfgap3	0.96362895	0.12339597
Zfp605	-0.9635515	0.08820809
Inadl	-0.9634451	0.00081499
Dnajc9	0.96298761	-0.101629
Abt1	0.96293616	-0.0714704
Zfp459	-0.962836	-0.0457722
Amot	-0.9627988	0.1433027
Abcd1	0.96276412	-0.1284864
Sh3kbp1	0.96268241	-0.0355288
Rala	0.96223109	-0.0218123
Myzap	-0.9620316	0.00055472
Usp16	-0.9618319	-0.0712475

Appendix B

Enriched DAVID annotations for transcriptomic data

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
GOTERM_BP_FAT	cell adhesion	RT		13	9.4	5.8E-4	3.0E-1
GOTERM_BP_FAT	biological adhesion	RT		13	9.4	5.8E-4	1.7E-1
GOTERM_BP_FAT	cell-cell adhesion	RT		8	5.8	1.4E-3	2.5E-1
GOTERM_CC_FAT	apical junction complex	RT		5	3.6	6.5E-3	6.5E-1
GOTERM_CC_FAT	plasma membrane part	RT		22	15.8	6.8E-3	4.2E-1
GOTERM_CC_FAT	apicolateral plasma membrane	RT		5	3.6	7.0E-3	3.1E-1
GOTERM_CC_FAT	cell-cell junction	RT		6	4.3	8.9E-3	3.0E-1
GOTERM_BP_FAT	calcium-independent cell-cell adhesion	RT		3	2.2	1.1E-2	8.3E-1
GOTERM_CC_FAT	basement membrane	RT		4	2.9	1.7E-2	4.1E-1
GOTERM_CC_FAT	tight junction	RT		4	2.9	1.8E-2	3.9E-1
GOTERM_CC_FAT	occluding junction	RT		4	2.9	1.8E-2	3.9E-1
GOTERM_CC_FAT	cell junction	RT		9	6.5	2.2E-2	4.0E-1
GOTERM_CC_FAT	intrinsic to membrane	RT		54	38.8	3.0E-2	4.5E-1
GOTERM_CC_FAT	extracellular matrix part	RT		4	2.9	3.0E-2	4.2E-1
GOTERM_CC_FAT	integral to membrane	RT		52	37.4	3.7E-2	4.5E-1
GOTERM_MF_FAT	identical protein binding	RT		6	4.3	4.0E-2	1.0E0
GOTERM_CC_FAT	plasma membrane	RT		30	21.6	4.2E-2	4.6E-1
GOTERM_MF_FAT	calcium ion binding	RT		11	7.9	4.4E-2	9.9E-1
GOTERM_MF_FAT	racemase and epimerase activity, acting on carbohydrates and derivatives	RT		2	1.4	4.4E-2	9.6E-1
GOTERM_CC_FAT	external side of plasma membrane	RT		5	3.6	6.5E-2	5.9E-1
GOTERM_BP_FAT	regulation of RNA splicing	RT		2	1.4	6.8E-2	1.0E0
GOTERM_MF_FAT	racemase and epimerase activity	RT		2	1.4	6.9E-2	9.8E-1
GOTERM_MF_FAT	structural molecule activity	RT		7	5.0	7.1E-2	9.6E-1
GOTERM_BP_FAT	protein oligomerization	RT		3	2.2	9.2E-2	1.0E0

Figure B.1: Full list of enriched Gene Ontology annotations for 150 genes with largest loadings in the full PLS model using online software DAVID.

Appendix C

Russian Roulette

Consider approximating the finite sum $S = \sum_{k\geq 0} \alpha_k$. Let τ denote a finite random time taking positive integer values such that $p_n := P(\tau \geq n) > 0$ for all $n \geq 0$. The fact that τ is finite almost surely means that

$$P(\tau = \infty) = \lim_{n\to\infty} p_n = 0. \tag{C.1}$$

We consider the weighted partial sums with $S_0 = \alpha_0$, and for $k \geq 1$,

$$S_k = \alpha_0 + \sum_{j=1}^{k} \frac{\alpha_j}{p_j}.$$

For completeness, we set $S_\infty = \infty$. The Russian roulette random truncation approximation of S is $\hat{S} = S_\tau$.

If τ can be easily simulated and the probabilities p_n are available then $\hat{S}$ can be computed. The next result states that $\hat{S}$ is an unbiased estimator of S.

Proposition C.0.1. *The random variable $\hat{S}$ has finite expectation, and $\mathrm{E}(\hat{S}) = S$.*

Proof. Set $\bar{S}_0 = |\alpha_0|$, and $\bar{S}_k = |\alpha_0| + \sum_{j=1}^{k} |\alpha_j|/p_j$. Then for all $n \geq 1$

$$\begin{aligned}\sum_{k=0}^{n}|S_k|P(\tau=k) &\le \sum_{k=0}^{n}\bar{S}_k P(\tau=k) = \sum_{k=0}^{n}\bar{S}_k\left(p_k-p_{k+1}\right)\\ &= \bar{S}_0 p_0 + \sum_{k=1}^{n}\left(\bar{S}_k-\bar{S}_{k-1}\right)p_k + \sum_{k=1}^{n}\bar{S}_{k-1}p_k - \sum_{k=0}^{n}\bar{S}_k p_{k+1}\\ &= \sum_{k=0}^{n}|\alpha_k| - \bar{S}_n p_{n+1} \le \sum_{k=0}^{n}|\alpha_k|.\end{aligned}$$

Since $\sum_n |\alpha_n| < \infty$, we conclude that $\sum_n |S_n| P(\tau = n) < \infty$, hence $\mathrm{E}(|\hat{S}|) < \infty$. A similar calculation as above gives for all $n \ge 1$,

$$\sum_{k=0}^{n} S_k P(\tau = k) = \sum_{k=0}^{n}\alpha_k - S_n p_{n+1}.$$

By Kronecker's lemma $\lim_{n\to\infty} p_n S_n = 0$, and $|p_{n+1}S_n| = (p_{n+1}/p_n)p_n|S_n| \le p_n|S_n| \to 0$, as $n \to \infty$. We conclude that $\mathrm{E}(\hat{S}) = \sum_{k=0}^{\infty} S_k P(\tau = k) = \sum_{k=0}^{\infty}\alpha_k$. □

This random truncation approximation of the series $\sum_n \alpha_n$ is known in the Physics literature as Russian roulette. It has been independently re-derived by McLeish (2011). In the Physics literature it is common to choose τ as a stopping time of the form

$$\tau = \inf\{k \ge 1 : \ U_k \ge q_k\},$$

where $\{U_j,\ j \ge 1\}$ are i.i.d. $\mathcal{U}(0,1)$, $q_j \in (0,1]$ and $\hat{S} = S_{\tau-1}$. In this case $p_n = \prod_{j=1}^{n-1} q_j$. The random time τ can be thought of as the running time of the algorithm. It is tempting to choose τ such that the Russian roulette terminates very quickly. The next result shows that the resulting variance will be high, possibly infinite.

Proposition C.0.2. *If*

$$\sum_{n\ge 1}\frac{|\alpha_n|}{p_n}\sup_{j\ge n}\left|\sum_{\ell=n}^{j}\alpha_\ell\right| < \infty,$$

then $\mathsf{Var}(\hat{S}) < \infty$ *and*

$$\mathsf{Var}(\hat{S}) = \alpha_0^2 + \sum_{n\geq 1} \frac{\alpha_n^2}{p_n} + 2\sum_{n\geq 1} \alpha_n S_{n-1} - S^2.$$

If $\{\alpha_n\}$ *is a sequence of nonnegative numbers and* $\sum_{n\geq 1} \alpha_n S_{n-1} = \infty$*, then* $\mathsf{Var}(\hat{S}) = \infty$.

Proof. $\mathsf{Var}(\hat{S}) = \mathrm{E}(\hat{S}^2) - S^2$. So it suffices to work with $\mathrm{E}(\hat{S}^2)$. $\mathrm{E}(\hat{S}^2) = \sum_{k=0}^{\infty} S_k^2 P(\tau = k) = \lim_{n\to\infty} \sum_{k=0}^{n} S_k^2 P(\tau = k)$. For any $n \geq 1$, we use the same telescoping trick used in Proposition C.0.1 to get

$$\begin{aligned}\sum_{k=0}^{n} S_k^2 P(\tau = k) &= \sum_{k=0}^{n} S_{k-1}^2 (p_k - p_{k+1}) \\ &= \alpha_0^2 + \sum_{k=1}^{n} \frac{\alpha_k^2}{p_k} + 2\sum_{k=1}^{n} \alpha_k S_{k-1} - S_n^2 p_{n+1}. \end{aligned} \tag{C.2}$$

By Jensen's inequality $S_n^2 \leq \left(\sum_{k=1}^{n} p_k^{-1}\right)\left(\sum_{k=1}^{n} p_k^{-1}\alpha_k^2\right)$. Hence, using Kronecker's lemma, we see that

$$p_{n+1}S_n^2 \leq p_n S_n^2 \leq \left(p_n \sum_{k=1}^{n} \frac{1}{p_k}\right)\left(\sum_{k=1}^{n} \frac{\alpha_k^2}{p_k}\right) = o\left(\sum_{k=1}^{n} \frac{\alpha_k^2}{p_k}\right), \text{ as } n \to \infty. \tag{C.3}$$

so it suffices to show that the sequence $\sum_{k=1}^{n} \frac{\alpha_k^2}{p_k} + \sum_{k=1}^{n} \alpha_k S_{k-1}$ is bounded. But

$$\begin{aligned}\left|\sum_{j=1}^{n} \frac{\alpha_k^2}{p_k} + \sum_{k=1}^{n} \alpha_k S_{k-1}\right| &= \left|\alpha_0 \sum_{j=0}^{n} \alpha_j + \sum_{j=1}^{n} \frac{\alpha_j}{p_j}\left(\sum_{k=j}^{n} \alpha_k\right)\right| \\ &\leq |\alpha_0| \sum_{j\geq 0} |\alpha_j| + \sup_n \sum_{j=1}^{n} \frac{|\alpha_j|}{p_j}\left|\sum_{k=j}^{n} \alpha_k\right|,\end{aligned}$$

and the two terms on the right-hand-side are bounded under the stated assumptions. Therefore the series $\sum_n S_n^2 P(\tau = n)$ is summable and the variance formula follows by taking the limit as $n \to \infty$ in (C.2).

To establish the rest of the proposition, we deduce from (C.3) that for n large enough

$$\sum_{k=1}^{n} S_k^2 P(\tau = k) \geq \alpha_0^2 + 2 \sum_{k=1}^{n} \alpha_k S_{k-1},$$

which implies the statement. □

Remark C.0.1. *As an example, for a geometric sequence $\alpha_i = \alpha^i$ for $\alpha \in (0,1)$, and we choose $q_i = q$ for some $q \in (0,1)$, then for $\alpha^2/q < 1$, the condition of Proposition C.0.2 is satisfied and $var(\hat{S}) < \infty$. If $q > \alpha^2$ the variance is infinite. The average computing time of the algorithm is $\mathrm{E}(\hat{\tau}) = \frac{1}{1-q}$. Although this variance/computing speed trade-off can be investigate analytically, a rule of thumb that works well in simulations is to choose $q = \alpha$.*

Bibliography

Ackermann, M. and Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):47.

Adams, R., Murray, I., and MacKay, D. (2009). Nonparametric Bayesian density modeling with Gaussian processes. *arXiv preprint arXiv:0912.4896*.

Agapiou, S. and Papaspiliopoulos, O. (2015). Importance Sampling: Computational Complexity and Intrinsic Dimension. *arXiv preprint arXiv:1511.06196*.

Agapiou, S., Roberts, G., and Vollmer, S. (2014). Unbiased Monte Carlo: posterior estimation for intractable/infinite-dimensional models. *arXiv preprint arXiv:1411.7713*.

Agarwal, R., Jurisica, I., Mills, G., and Cheng, K. (2009). The emerging role of the RAB25 small GTPase in cancer. *Traffic*, 10(11):1561–1568.

Alajez, N., Lenarduzzi, M., Ito, E., Hui, A., Shi, W., Bruce, J., Yue, S., Huang, S., Xu, W., and Waldron, J. (2011). MiR-218 suppresses nasopharyngeal cancer progression through downregulation of survivin and the SLIT2-ROBO1 pathway. *Cancer Research*, 71(6):2381–2391.

Albino, D., Longoni, N., Curti, L., Mello-Grand, M., Pinton, S., Civenni, G., Thalmann, G., D'Ambrosio, G., Sarti, M., and Sessa, F. (2012). ESE3/EHF controls epithelial cell differentiation and its loss leads to prostate tumors with mesenchymal and stem-like features. *Cancer Research*, 72(11):2889–2900.

Alquier, P., Friel, N., Everitt, R., and Boland, A. (2014). Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, pages 1–19.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725.

Andrieu, C. and Vihola, M. (2014). Establishing some order amongst exact approximations of MCMCs. *arXiv preprint arXiv:1404.6909*.

Arkin, A., Ross, J., and McAdams, H. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected Escherichia coli cells. *Genetics*, 149(4):1633–1648.

Atchadé, Y., Lartillot, N., and Robert, C. (2013). Bayesian computation for statistical models with intractable normalizing constants. *Brazilian Journal of Probability and Statistics*, 27(4):416–436.

Bachmann-Gagescu, R., Phelps, I., Stearns, G., Link, B., Brockerhoff, S., Moens, C., and Doherty, D. (2011). The ciliopathy gene cc2d2a controls zebrafish photoreceptor outer segment development through a role in Rab8-dependent vesicle trafficking. *Human Molecular Genetics*.

Baker, R., Jeffrey, P., and Hughson, F. (2013). Crystal structures of the Sec1/Munc18 (SM) protein Vps33, alone and bound to the homotypic fusion and vacuolar protein sorting (HOPS) subunit Vps16. *PLoS One*, 8(6):e67409.

Bakeyev, T. and Forcrand, P. D. (2001). Noisy Monte Carlo algorithm reexamined. *Physical Review D*, 63(5).

Banerji, C., Knopp, P., Moyle, L., Severini, S., Orrell, R., Teschendorff, A., and Zammit, P. (2015). β-catenin is central to DUX4-driven network rewiring in facioscapulohumeral muscular dystrophy. *Journal of The Royal Society Interface*, 12(102).

Banushi, B., Forneris, F., Straatman-Iwanowska, A., Strange, A., Lyne, A.-M., Rogerson, C., Burden, J., Heywood, W., Hanley, J., Straatman, K., Smith, H., Bem, D., Kriston-Vizi, J., Ariceta, G., Risteli, M., Wang, C., Waddington, S., Howe, S., Ferraro, F., Gjinovci, A., Lawrence, S., Marsh, M., Girolami, M., Bozec, L., Mills, K., and Gissen, P. (2015). Regulation of post-Golgi PLOD3 trafficking is essential for collagen maturation. *Submitted*.

Barber, S., Voss, J., and Webster, M. (2015). The rate of convergence for approximate Bayesian computation. *Electronic Journal of Statistics*, 9(80-105).

Beaumont, M. (2010). Approximate Bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics*, 41:379–406.

Beaumont, M., Zhang, W., and Balding, D. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160.

Berg, J., Tymoczko, J., and Stryer, L. (2002). *Biochemistry. 5th*. New York: WH Freeman.

Bernoulli, D. (1760). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir. *Histoire de l'Acad. Roy. Sci.(Paris) avec Mém. des Math et Phys*, pages 1–45.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 259—-302.

Besag, J. and Moran, P. (1975). On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika*, 62(3):555–562.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O., and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):333–382.

Bhanot, G. and Kennedy, A. D. (1985). Bosonic lattice gauge theory with noise. *Physics Letters B*, 157(1):70–76.

Bhat, A., Pope, J., Smith, J., Ahmad, R., Chen, X., Washington, M., Beauchamp, R., Singh, A., and Dhawan, P. (2014). Claudin-7 expression induces mesenchymal to epithelial transformation (MET) to inhibit colon tumorigenesis. *Oncogene*.

Blum, M. and Tran, V. (2010). HIV with contact tracing: a case study in approximate Bayesian computation. *Biostatistics*, 11(4):644–660.

Bolstad, B. (2004). *Low-level analysis of high-density oligonucleotide array data: background, normalization and summarization*. PhD thesis, University of California, Berkeley.

Bolstad, B., Irizarry, R., Åstrand, M., and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

Booth, T. E. (2007). Unbiased Monte Carlo estimation of the reciprocal of an integral. *Nuclear Science and Engineering*, 156(3):403–407.

Boys, R., Wilkinson, D., and Kirkwood, T. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, 18(2):125–135.

Bryant, N. and Stevens, T. (1998). Vacuole biogenesis in Saccharomyces cerevisiae: protein transport pathways to the yeast vacuole. *Microbiology and Molecular Biology Reviews*, 62(1):230–247.

Caimo, A. and Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, 33(1):41–55.

Capparuccia, L. and Tamagnone, L. (2009). Semaphorin signaling in cancer cells and in cells of the tumor microenvironment-two sides of a coin. *Journal of Cell Science*, 8(8):632—-645.

Cappé, O., Godsill, S., and Moulines, E. (2007). An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, 95(5):899 – 924.

Carter, L. and Cashwell, E. (1975). Particle Transport Simulation with the Monte Carlo method.

Casanova, J., Wang, X., Kumar, R., Bhartur, S., Navarre, J., Woodrum, J., Altschuler, Y., Ray, G., and Goldenring, J. (1999). Association of Rab25 and Rab11a with the apical recycling system of polarized Madin-Darby canine kidney cells. *Molecular biology of the cell*, 10(1):47–61.

Cavanna, T., Pokorná, E., Vesely, P., Gray, C., and Zicha, D. (2007). Evidence

for protein 4.1 B acting as a metastasis suppressor. *Journal of cell science*, 120(4):606–616.

Christen, J. and Fox, C. (2005). MCMC using an approximation. *Journal of Computational and Graphical Statistics*, 14(4):795–810.

Colledge, M. and Scott, J. (1999). AKAPs: from structure to function. *Trends in cell biology*, 9(6):216–221.

Csilléry, K. and Blum, M. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in ecology & evolution*, 25(7):410–418.

Cullinane, A., Straatman-Iwanowska, A., Zaucker, A., Wakabayashi, Y., Bruce, C., Luo, G., Rahman, F., Gurakan, F., Utine, E., Ozkan, T., Denecke, J., Vukovic, J., Di Rocco, M., Mandel, H., Cangul, H., Matthews, R., Thomas, S., Rappoport, J., Arias, I., Wolburg, H., Knisely, A., Kelly, D., Muller, F., Maher, E., and Gissen, P. (2010). Mutations in VIPAR cause an arthrpogryposis, renal dysfunction and cholestasis syndrome phenotype with defects in epithelial polarization. *Nature Genetics*, 42(4):303–312.

Dafou, D., Grun, B., Sinclair, J., Lawrenson, K., Benjamin, E. C., Hogdall, E., Kruger-Kjaer, S., Christensen, L., Sowter, H. M., Al-Attar, A., Edmondson, R., Darby, S., Berchuck, A., Laird, P., Pearce, C., Ramus, S., Jacobs, I., and Gayther, S. (2010). Microcell-mediated chromosome transfer identifies EPB41L3 as a functional suppressor of epithelial ovarian cancers. *Neoplasia*, 12(7).

de Marco, M., Martín-Belmonte, F., Kremer, L., Albar, J., Correas, I., Vaerman, J., Marazuela, M., Byrne, J., and Alonso, M. (2002). MAL2, a novel raft protein of the MAL family, is an essential component of the machinery for transcytosis in hepatoma HepG2 cells. *The Journal of cell biology*, 159(1):37–44.

Derynck, R. and Zhang, Y. (2003). Smad-dependent and Smad-independent pathways in TGF-β family signalling. *Nature*, 425(6958):577–584.

Diggle, P. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 349–362.

Dimakos, X. (2001). A guide to exact simulation. *International statistical review*, 69(1).

Douc, R. and Robert, C. (2011). A vanilla Rao-Blackwellization of Metropolis-Hastings algorithms. *The Annals of Statistics*, 39(1):261–277.

Doucet, A., Pitt, M., and Kohn, R. (2012). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *arXiv preprint arXiv:1210.1871*.

Dozynkiewicz, M., Jamieson, N., MacPherson, I., Grindlay, J., van den Berghe, P. V., von Thun, A., Morton, J. P., Gourley, C., Timpson, P., Nixon, C., McKay, C., Carter, R., Strachan, D., Anderson, K., Sansom, O., Caswell, P., and Norman, J. (2012). Rab25 and CLIC3 collaborate to promote integrin recycling from late endosomes/lysosomes and drive cancer progression. *Developmental cell*, 22(1):131–145.

Drovandi, C. and Pettitt, A. (2011). Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, 67(1):225–233.

Duijn, M. V., Gile, K., and Handcock, M. (2009). A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks*, 31(1):52–62.

Eidsvik, J. and Shaby, B. (2014). Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics*, 23(2):295–315.

El Ghaoui, L. and Gueye, A. (2008). A convex upper bound on the log-partition function for binary graphical models. In *Proc. NIPS*.

Everitt, R. G. (2012). Bayesian parameter estimation for latent Markov random fields and social networks. *Journal of Computational and Graphical Statistics*, 21(4):940–960.

Farrar, M. and Schreiber, R. (1993). The molecular cell biology of interferon-gamma and its receptor. *Annual review of immunology*, 11(1):571–611.

Fearnhead, P., Papaspiliopoulos, O., and Roberts, G. O. (2008). Particle filters for partially observed diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):755–777.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semiautomatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474.

Fill, J. (1997). An interruptible algorithm for perfect sampling via Markov chains. *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 688–695.

Fisher, D., Smith, J., Pillar, J., Denis, S. S., and Cheng, J. (1998). Isolation and characterization of PDE8A, a novel human cAMP-specific phosphodiesterase. *Biochemical and biophysical research communications*, 246(3):570–577.

Frey, P. (1996). The Leloir pathway: a mechanistic imperative for three enzymes to change the stereochemical configuration of a single carbon in galactose. *The FASEB Journal*, 10(4):461–470.

Friel, N. and Pettitt, A. (2004). Likelihood estimation and inference for the autologistic model. *Journal of Computational and Graphical Statistics*, 13(1).

Friel, N., Pettitt, A. N., Reeves, R., and Wit, E. (2009). Bayesian Inference in Hidden Markov Random Fields for Binary Data Defined on Large Lattices. *Journal of Computational and Graphical Statistics*, 18(2):243–261.

Gallazzini, M., Ferraris, J., and Burg, M. (2008). GDPD5 is a glycerophosphocholine phosphodiesterase that osmotically regulates the osmoprotective organic osmolyte GPC. *Proceedings of the National Academy of Sciences*, 105(31):11026–11031.

Gelman, A. and Meng, X. L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185.

Gerbault, P., Allaby, R., Boivin, N., Rudzinski, A., Grimaldi, I. M., Pires, J. C., Vigueira, C., Dobney, K., Gremillion, K., and Barton, L. (2014). Storytelling and story testing in domestication. *Proceedings of the National Academy of Sciences*, 111(17):6159–6164.

Ghaoui, L. and Gueye, A. (2008). A Convex Upper Bound on the Log-Partition

Function for Binary Distributions. *Advances in Neural Information Processing Systems (NIPS)*, pages 409–416.

Ghazalpour, A., Bennett, B., and Petyuk, V. (2011). Comparative analysis of proteome and transcriptome variation in mouse. *PLoS genetics*, 7(6).

Gillespie, D. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(2):2340–2361.

Gissen, P., Johnson, C., Morgan, N., Stapelbroek, J., Forshew, T., Cooper, W., McKiernan, P., Klomp, L., Morris, A., Wraith, J., McClean, P., Lynch, S., Thompson, R., Lo, B., Quarrell, O., Di Rocco, M., Trembath, R., Mandel, H., Wali, S., Karet, F., Knisely, A., Houwen, R., Kelly, D., and Maher, E. (2004). Mutations in VPS33B, encoding a regulator of SNARE-dependent membrane fusion, cause arthrogryposis, renal dysfunction, cholestasis (ARC) syndrome. *Nature Genetics*, 36(4):400–404.

Gissen, P., Tee, L., Johnson, C. A., Genin, E., Caliebe, A., Chitayat, D., Clericuzio, C., Denecke, J., Di Rocco, M., Fischler, B., and Others (2006). Clinical and molecular genetic features of ARC syndrome. *Human Genetics*, 120(3):396–409.

Glowacka, I., Bertram, S., Müller, M., Allen, P., Soilleux, E., Pfefferle, S., Steffen, I., Tsegaye, T., He, Y., Gnirss, K., and Pohlmann, S. (2011). Evidence that TMPRSS2 activates the severe acute respiratory syndrome coronavirus spike protein for membrane fusion and reduces viral control by the humoral. *Journal of virology*, 85(9):4122–4134.

Glynn, P. (1984). Some asymptotic formulas for markov chains with applications to simulation. *Journal of Statistical Computation and Simulation*, 19(2):97–112.

Glynn, P. and Rhee, C. (2014). Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51:377–389.

Glynn, P. and Whitt, W. (1992). The asymptotic efficiency of simulation estimators. *Operations Research*, 40(3):505–520.

Golightly, A. and Wilkinson, D. (2006). Bayesian sequential inference for stochastic kinetic biochemical network models. *Journal of Computational Biology*, 13(3):838–851.

Golightly, A. and Wilkinson, D. (2008). Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis*.

Goodreau, S., Kitts, J., and Morris, M. (2009). Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks. *Demography*, 46(1):103—-125.

Graham, S., Wartosch, L., Gray, S., Scourfield, E., Deane, J., Luzio, J., and Owen, D. (2013). Structural basis of Vps33A recruitment to the human HOPS complex by Vps16. *Proceedings of the National Academy of Sciences*, 110(33):13345–13350.

Grant, M. and Prockop, D. (1972). The biosynthesis of collagen. *New England Journal of Medicine*, 286(6):291–300.

Green, P. and Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, 97(460):1055–1070.

Grelaud, A., Robert, C., and Marin, J. (2009). ABC methods for model choice in Gibbs random fields. *Comptes Rendus Mathematique*, 347(3):205—-210.

Gu, M. G. and Zhu, H. T. (2001). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):339–355.

Häcker, H., Tseng, P., and Karin, M. (2011). Expanding TRAF function: TRAF3 as a tri-faced immune regulator. *Nature Reviews Immunology*, 11(7):457–468.

Hase, K., Nakatsu, F., Ohmae, M., Sugihara, K., Shioda, N., Takahashi, D., Obata, Y., Furusawa, Y., Fujimura, Y., and Yamashita, T. (2013). AP-1B - Mediated Protein Sorting Regulates Polarity and Proliferation of Intestinal Epithelial Cells in Mice. *Gastroenterology*, 145(3):625–635.

Heikkinen, J. and Hogmander, H. (1994). Fully Bayesian approach to image restoration with an application in biogeography. *Applied Statistics*, pages 569–582.

Hendricks, J. and Booth, T. (1985). MCNP variance reduction overview. *Monte-Carlo Methods and Applications in Neutronics, Photonics and Statistical Physics*, pages 83–92.

Hodgkin, A. and Huxley, A. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500–544.

Huang, D., Sherman, B., and Lempicki, R. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1):44–57.

Hughes, J., Haran, M., and Caragea, P. C. (2011). Autologistic models for binary data on a lattice. *Environmetrics*, 22(7):857–871.

Huizing, M., Didier, A., Walenta, J., Anikster, Y., Gahl, W., and Kramer, H. (2001). Molecular cloning and characterization of human VPS18, VPS 11, VPS16, and VPS33. *Gene*, 264(2):241–247.

Illian, J., Soerbye, S., Rue, H., and Hendrichsen, D. (2012). Using INLA to fit a complex point process model with temporally varying effects-a case study. *Journal of Environmental Statistics*.

Irizarry, R., Bolstad, B., and Collin, F. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*, 31(4).

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258.

Jacob, P. and Thiery, A. (2013). On non-negative unbiased estimators. *arXiv preprint arXiv:1309.6473*.

Jacob, P. and Thiery, A. (2015). On nonnegative unbiased estimators. *The Annals of Statistics*, 43(2):769–784.

Jin, I. and Liang, F. (2014). Use of SAMC for Bayesian analysis of statistical models with intractable normalizing constants. *Computational Statistics & Data Analysis*, 71:402–416.

Joo, B., Horvath, I., and Liu, K. (2003). The Kentucky noisy Monte Carlo algorithm for Wilson dynamical fermions. *Physical Review D*, 67(7).

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., and Matthews, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33.

Kaddurah-Daouk, R., Bruce, K., and Weinshilboum, R. (2008). Metabolomics: a global biochemical approach to drug response and disease. *Annu. Rev. Pharmacol. Toxicol.*, 48:653–683.

Kalluri, R. and Weinberg, R. (2009). The basics of epithelial-mesenchymal transition. *The Journal of clinical investigation*, 119(6):1420.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research*, 42(D1).

Kendall, W. (2005). Notes on perfect simulation. *Dept. of statistics, University of Warwick*.

Kennedy, A. D. and Kuti, J. (1985). Noise without noise: a new Monte Carlo method. *Physical review letters*, 54(23):2473–2476.

Kent, J. (1982). The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society. Series B*, 44(1):71–80.

Khoshnoodi, J. (2008). Mammalian collagen IV. *Microscopy research and technique*, 71(5):357–370.

Kim, S., Wairkar, Y., Daniels, R., and DiAntonio, A. (2010). The novel endosomal membrane protein Ema interacts with the class C Vps-HOPS complex to promote endosomal maturation. *The Journal of cell biology*, 188(5):717–734.

Kim, Y.-A., Wuchty, S., and Przytycka, T. M. (2011). Identifying Causal Genes and Dysregulated Pathways in Complex Diseases. *PLoS Computational Biology*, 7(3):e1001095.

Kitano, H. (2002a). Computational systems biology. *Nature*, 420(6912):206–210.

Kitano, H. (2002b). Systems biology: a brief overview. *Science*, 295:1662–1664.

kleine Balderhaar, H. and Ungermann, C. (2013). CORVET and HOPS tethering complexes-coordinators of endosome and lysosome fusion. *Journal of Cell Science*, 126.

Knott, L. and Bailey, A. (1998). Collagen cross-links in mineralizing tissues: a review of their chemistry, function, and clinical relevance. *Bone*, 22(3):181–187.

Komurov, K., White, M., and Ram, P. (2010). Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data. *PLoS Computional Biology*, 6(8).

Krause, G., Winkler, L., Mueller, S., Haseloff, R., Piontek, J., and Blasig, I. (2008). Structure and function of claudins. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1778(3):631–645.

Kumanogoh, A., Marukawa, S., Suzuki, K., Takegahara, N., Watanabe, C., Chng, E., Ishida, I., Fujimura, H., Sakoda, S., and Yoshida, K. (2002). Class IV semaphorin Sema4A enhances T-cell activation and interacts with Tim-2. *Nature*, 419(6907):629–633.

Lai-Cheong, J., Parsons, M., Tanaka, A., Ussar, S., South, A., Gomathy, S., Mee, J. B., Barbaroux, J., Techanukul, T., Almaani, N., Clements, S., Hart, I., and McGrath, J. (2009). Loss-of-function FERMT1 mutations in kindler syndrome implicate a role for fermitin family homolog-1 in integrin activation. *The American journal of pathology*, 175(4):1431–1441.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182.

Li, X., Law, J., and Lee, A. (2012). Semaphorin 5A and plexin-B3 regulate human glioma cell motility and morphology through Rac1 and the actin cytoskeleton. *Oncogene*, 31(5):595–610.

Liang, F. (2010). A double Metropolis-Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*.

Liang, F., Liu, C., and Carroll, R. J. (2007). Stochastic approximation in Monte Carlo computation. *Journal of the American Statistical Association*, 102(477):305–320.

Lin, L., Liu, K. F., and Sloan, J. (2000). A noisy Monte Carlo algorithm. *Physical Review D*, 61(7):74505.

Liu, H., Ong, S., Badu-Nkansah, K., Schindler, J., White, F., M, and Hynes, R. (2011). CUB-domain-containing protein 1 (CDCP1) activates Src to promote melanoma metastasis. *Proceedings of the National Academy of Sciences*, 108(4):1379–1384.

Liu, Q. and Ihler, A. (2011). Bounding the Partition Function using Holder's Inequality. *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 849–856.

Lotka, A. (1925). *Elements of physical biology*. Williams & Wilkins company.

Lu, Z., Kim, D., Fan, J., and Lu, Q. (2015). A non-tight junction function of claudin-7 - Interaction with integrin signaling in suppressing lung cancer cell proliferation and detachment. *Molecular cancer*, 14(1).

Lux, I. and Koblinger, L. (1991). *Monte Carlo particle transport methods: neutron and photon calculations*. CRC press Boca Raton.

Lyne, A., Girolami, M., Atchade, Y., Strathmann, H., and Simpson, D. (2015). On Russian Roulette Estimates for Bayesian inference with Doubly-Intractable Likelihoods. *Statistical Science*, 30(4):443–467.

MacKay, D. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

MacPherson, L., Ahmed, S., Tamblyn, L., Krutmann, J., Forster, I., Weighardt, H., and Matthews, J. (2014). Aryl hydrocarbon receptor repressor and TiPARP (ARTD14) use similar, but also distinct mechanisms to repress aryl hydrocarbon receptor signaling. *International journal of molecular science*, 15(5):7939–7957.

Madrid, R., Aranda, J., Rodriguez-Fraticelli, A., Ventimiglia, L., Andres-Delgado, L., Shehata, M., Fanayan, S., Shahheydari, H., Gomez, S., Jimenez, A., Byrne, J., and Alonso, M. (2010). The formin INF2 regulates basolateral-to-apical transcytosis and lumen formation in association with Cdc42 and MAL2. *Developmental cell*, 18(5):814–827.

Marin, J., Pudlo, P., Robert, C., and Ryder, R. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte

Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

Mas-Vidal, A., Minones-Suarez, L., Toral, J., Mallo, S., and Perez-Oliva, N. (2010). A novel mutation in the FERMT1 gene in a Spanish family with Kindler's syndrome. *Journal of the European Academy of Dermatology and Venereology*, 24(8):978–979.

McEwan, D., Popovic, D., and Gubas, A. (2015). PLEKHM1 regulates autophagosome-lysosome fusion through HOPS complex and LC3/GABARAP proteins. *Molecular cell*, 57(1):39–54.

McLeish, D. (2011). A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods and Applications*, 17(4).

McMillan, B. and Bradfield, C. (2007). The aryl hydrocarbon receptor sans xenobiotics: endogenous function in genetic model systems. *Molecular pharmacology*, 72(3):487–498.

Mercurio, F., Marasco, D., Pirone, L., Scognamiglio, P., Pedone, E., Pellecchia, M., and Leone, M. (2013). Heterotypic Sam-Sam Association between Odin-Sam1 and Arap3-Sam: Binding Affinity and Structural Insights. *Chembiochem*, 14(1):100–106.

Moestue, S., Giskeodegard, G., Cao, M., Bathen, T., and Gribbestad, I. (2012). Glycerophosphocholine (GPC) is a poorly understood biomarker in breast cancer. *Proceedings of the National Academy of Sciences*, 109(38):E2506—-E2506.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Møller, J. and Waagepetersen, R. (2003). *Statistical inference and simulation for spatial point processes*. CRC Press.

Moores, M., Drovandi, C., Mengersen, K., and Robert, C. (2014). Pre-processing for approximate Bayesian computation in image analysis. *Statistics and Computing*, 25(1):23–33.

Moral, P. D., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society. Series B*, 68(3):411–436.

Moral, P. D., Doucet, A., and Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22(5):1009–1020.

Murdoch, D. and Green, P. (1998). Exact sampling from a continuous state space. *Scandinavian Journal of Statistics*, 25(3):483–502.

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 359–366. AUAI Press.

Nagaraj, N. and Wisniewski, J. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology*, 7(1).

Nakamura, Y., Asano, A., Hosaka, Y., Tskeuchi, T., Iwanaga, T., and Yamano, Y. (2015). Expression and intracellular localization of TBC1D9, a Rab GTPase-accelerating protein, in mice testes. *Experimental Animals*.

Neal, R. (2001). Annealed importance sampling. *Statistics and Computing*, 11(2):125–139.

Neufeld, G. and Kessler, O. (2008). The semaphorins: versatile regulators of tumour progression and tumour angiogenesis. *Nature Reviews Cancer*, 122(11):1723–1736.

Neufeld, G., Sabag, A., Rabinovicz, N., and Kessler, O. (2012). Semaphorins in angiogenesis and tumor progression. *Cold Spring Harbor perspectives in medicine*, 2(1).

Neunlist, M., Landeghem, L. V., Mahe, M., Derkinderen, P., des Varannes, S., and Rolli-Derkinderen, M. (2013). The digestive neuronal-glial-epithelial unit: a new actor in gut health and disease. *Nature Reviews Gastroenterology and Hepatology*, 10(2):90–100.

Nishimura, D. (2001). BioCarta. *Biotech Software & Internet Report: The Computer Software Journal for Science*, 2(3):117–120.

Nunes, M. and Balding, D. (2010). On optimal selection of summary statistics

for approximate Bayesian computation. *Statistical applications in genetics and molecular biology*, 9(1).

Pan, G., Ren, H., Zhang, S., Wang, X., and Wen, J. (2009). Expression of semaphorin 5A and its receptor plexin B3 contributes to invasion and metastasis of gastric carcinoma. *World journal of gastroenterology: WJG*, 15(22).

Pang, Z. and Südhof, T. (2010). Cell biology of Ca 2+-triggered exocytosis. *Current opinion in cell biology*, 22(4):496–505.

Papaspiliopoulos, O. (2009). Publisher statement : None A methodological framework for Monte Carlo probabilistic inference for diffusion processes. Technical report.

Paredes, J., Figueiredo, J., Albergaria, A., Oliveira, P., Carvalho, J., Ribeiro, A. S., Caldeira, J., Costa, {\^A}. M., Sim{\~o}es-Correia, J., Oliveira, M., A, H. P., Pinho, S., Mateus, R., Reis, C., Leite, M., Fernandes, M., Schmitt, F., Carneiro, F., Figueiredo, C., Oliveira, C., and Seruca, R. (2012). Epithelial E-and P-cadherins: role and clinical significance in cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1826(2):297–311.

Patrucco, E., Albergine, M., Santana, L., and Beavo, J. (2010). Phosphodiesterase 8A (PDE8A) regulates excitation-contraction coupling in ventricular myocytes. *Journal of molecular and cellular cardiology*, 49(2):330–333.

Paulusma, C., Folmer, D., Ho-Mok, K. S., de Waart, D. R., Hilarius, P. M., Verhoeven, A. J., and Oude Elferink, R. P. (2008). ATP8B1 requires an accessory protein for endoplasmic reticulum exit and plasma membrane lipid flippase activity. *Hepatology*, 47(1):268–278.

Pećina-Šlaus, N. (2003). Tumor suppressor gene E-cadherin and its role in normal and malignant cells. *Cancer Cell International*, 3(1).

Pitt, M., Silva, R., Giordani, P., and Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151.

Pols, M., Brink, C., Gosavi, P., Oorschot, V., and Klumperman, J. (2013). The HOPS proteins hVps41 and hVps39 are required for homotypic and heterotypic late endosome fusion. *Traffic*, 14(2):219–232.

Pritchard, J., Seielstad, M., Perez-Lezaun, A., and Feldman, M. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random structures and Algorithms*, 9(1-2):223–252.

Ratmann, O., Jørgensen, O., Hinkley, T., Stumpf, M., Richardson, S., and Wiuf, C. (2007). Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum. *PLoS Computional Biology*, 3(11).

Rhee, C. and Glynn, P. (2012). A new approach to unbiased estimation for SDE's. *Proceedings of the Winter Simulation Conference*.

Rhee, C. and Glynn, P. (2013). Unbiased estimation with square root convergence for SDE models. *Submitted for publication*.

Rice, J. (2006). *Mathematical statistics and data analysis*. Cengage Learning.

Rieg, T., Tang, T., and Murray, F. (2010). Adenylate cyclase 6 determines cAMP formation and aquaporin-2 phosphorylation and trafficking in inner medulla. *Journal of the Journal of the American Society of Nephrology*, 21(12):2059–2068.

Ripley, B. (2009). *Stochastic simulation*. Wiley-Interscience Paperback Series.

Roberts, G. and Rosenthal, J. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71.

Roberts, G. and Rosenthal, J. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of applied probability*, pages 458–475.

RodriguezPazos, L., Ginarte, M., Fachal, L., Toribio, J., Carracedo, A., and Vega, A. (2011). Analysis of TGM1, ALOX12B, ALOXE3, NIPAL4 and CYP4F22 in autosomal recessive congenital ichthyosis from Galicia (NW Spain): evidence of founder effects. *British Journal of Dermatology*, 165(4):906–911.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B*, 71(2):319—-392.

Salo, A., Cox, H., Farndon, P., Moss, C., Grindulis, H., Risteli, M., Robins, S., and Myllyla, R. (2008). A connective tissue disorder caused by mutations of the lysyl hydroxylase 3 gene. *The American Journal of Human Genetics*, 83(4):495–503.

Schrödle, B. and Held, L. (2011). Spatio-temporal disease mapping using INLA. *Environmetrics*, 22(6):725–734.

Serini, G., Valdembri, D., Zanivan, S., Morterra, G., Burkhardt, C., Caccavari, F., Bussolino, F., Zammataro, L., Primo, L., Tamagnone, L., and Logan, M. (2003). Class 3 semaphorins control vascular morphogenesis by inhibiting integrin function. *Nature*, 424(6947):391–397.

Sherlock, C. and Thiery, A. (2014). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *The Annals of Statistics*, 43(1):238–275.

Siegmund, K., Marjoram, P., and Shibata, D. (2008). Modeling DNA methylation in a population of cancer cells. *Statistical applications in genetics and molecular biology*, 7(1).

Silvertown, J. (2001). *Integrating Ecology and Evolution in a Spatial Context: 14th Special Symposium of the British Ecological Society*. Cambridge University Press.

Smith, H., Galmes, R., Gogolina, E., Straatman-Iwanowska, A., Reay, A., Banushi, B., Bruce, C., Cullinane, A., Romero, R., Chang, R., Ackermann, O., Baumann, C., Cangul, H., Cakmak Celik, F., Aygun, C., Coward, R., Dionisi-Vici, C., Sibbles, B., Inward, C., Ae Kim, C., Klumperman, J., Knisely, A., Watson, S., and Gissen, P. (2012). Associations among genotype, clinical phenotype, and intracellular localization of trafficking proteins in ARC syndrome. *Human mutation*, 33(12):1656–1664.

Smyth, G. (2005). Limma: linear models for microarray data. *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420.

Solinger, J. and Spang, A. (2013). Tethering complexes in the endocytic pathway: CORVET and HOPS. *FEBS Journal*, 280(12):2743–2757.

Stein, U., Smith, J., Walther, W., and Arlt, F. (2009). MACC1 controls Met: what a difference an Sp1 site makes. *Cell Cycle*, 8(15):2467–2469.

Stewart, J., Marchan, R., Lesjak, M., Lambert, J., Hergenroeder, R., Ellis, J., Lau, C.-H., Keun, H., Schmitz, G., Schiller, J., and Hengstler, J. (2012). Choline-releasing glycerophosphodiesterase EDI3 drives tumor cell migration and metastasis. *Proceedings of the National Academy of Sciences*, 109(29):8155–8160.

Strathmann, H., Sejdinovic, D., and Girolami, M. (2015). Unbiased Bayes for Big Data: Paths of Partial Posteriors. *arXiv preprint arXiv:1501.03326*.

Takamatsu, H. and Kumanogoh, A. (2012). Diverse roles for semaphorin-plexin signaling in the immune system. *Trends in immunology*, 33(3):127–135.

Tamagnone, L. (2012). Emerging role of semaphorins as major regulatory signals and potential therapeutic targets in cancer. *Cancer cell*, 22(2):145–152.

Tavaré, S., Balding, D., Griffiths, R., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518.

Taylor, B. and Diggle, P. (2014). INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. *Journal of Statistical Computation and Simulation*, 84(10):2266–2284.

The Tumor Analysis Best Practices Working Group (2004). Expression profiling best practices for data generation and interpretation in clinical trials - Google Search. *Nature Reviews Genetics*, 5:229–237.

Tokuda, N., Numata, S., Li, X., Nomura, T., Takizawa, M., Kondo, Y., Yamashita, Y., Hashimoto, N., Kiyono, T., and Urano, T. (2013). β4GalT6 is involved in the synthesis of lactosylceramide with less intensity than β4GalT5. *Glycobiology*.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202.

Tornieri, K., Zlatic, S., Mullin, A., Werner, E., Harrison, R., L'Hernault, S., and Faundez, V. (2013). Vps33b pathogenic mutations preferentially affect VIPAS39/SPE-39-positive endosomes. *Human molecular genetics*, 22(25):5215–5228.

Troyer, M. and Wiese, U. (2005). Computational complexity and fundamental limitations to fermionic quantum Monte Carlo simulations. *Physical review letters*, 94(17).

Turner, S. and Sherratt, J. (2002). Intercellular adhesion and cancer invasion: a discrete simulation using the extended Potts model. *Journal of Theoretical Biology*, 216(1):85–100.

Tzaban, S., Massol, R., Yen, E., Hamman, W., Frank, S., Lapierre, L., Hansen, S., Goldenring, J., Blumberg, R., and Lencer, W. (2009). The recycling and transcytotic pathways for IgG transport by FcRn are distinct and display an inherent polarity. *The Journal of Cell Biology*, 185(4):673–684.

Vives, V., Laurin, M., and Cres, G. (2011). The Rac1 exchange factor Dock5 is essential for bone resorption by osteoclasts. *Journal of Bone and Mineral Research*, 26(5):1099–1110.

Volterra, V. (1927). *Variazioni e fluttuazioni del numero d'individui in specie animali conviventi*. C. Ferrari.

Wainwright, M., Jaakkola, T., and Willsky, A. (2005). A new class of upper bounds on the log partition function. *Information Theory, IEEE Transactions on*, 51(7):2313–2335.

Walker, S. (2011). Posterior sampling when the normalizing constant is unknown. *Communications in Statistics - Simulation and Computation*, 40(5).

Walker, S. (2014). A Bayesian analysis of the Bingham distribution. *Brazilian Journal of Probability and Statistics*, 28(1):61–72.

Wang, F. and Landau, D. (2001). Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical review letters*, 86(10):2050.

Wang, X., Kumar, R., and Navarre, J. (2000). Regulation of vesicle trafficking in madin-darby canine kidney cells by Rab11a and Rab25. *Journal of Biological Chemistry*, 275(37):29138—-29146.

Wartosch, L., Günesdogan, U., Graham, S., and Luzio, J. (2015). Recruitment of VPS33A to HOPS by VPS16 Is Required for Lysosome Fusion with Endosomes and Autophagosomes. *Traffic*, 16(7):727–742.

Warzecha, C., Sato, T., Nabet, B., Hogenesch, J., and Carstens, R. (2009). ESRP1 and ESRP2 are epithelial cell-type-specific regulators of FGFR2 splicing. *Molecular cell*, 33(5):591–601.

Wei, P. and Pan, W. (2008). Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics*, 24(3):404–411.

Wei, Z. and Li, H. (2007). A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, 23(12):1537–1544.

Wilkinson, D. (2007). Bayesian methods in bioinformatics and computational systems biology. *Briefings in bioinformatics*, 8(2):109–116.

Wilkinson, D. (2009). Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, 10(2):122–133.

Wilkinson, D. (2011). *Stochastic modelling for systems biology*. CRC Press.

Wilkinson, R. (2013). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical applications in genetics and molecular biology*, 12(2):129–141.

Wold, S., Ruhe, A., Wold, H., Dunn, W., and III (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130.

Yamazaki, T., Masuda, J., Omori, T., Usui, R., Akiyama, H., and Maru, Y. (2009). EphA1 interacts with integrin-linked kinase and regulates cell morphology and motility. *Journal of cell science*, 122(2):243—-255.

Yang, S. and Alkayed, N. (2009). Cyclic adenosine monophosphate response element-binding protein phosphorylation and neuroprotection by 4-phenyl-1-(4-phenylbutyl) piperidine (PPBP). *Anesthesia and analgesia*, 108(3).

Yazdani, U. and Terman, J. (2006). The semaphorins. *Genome Biology*, 7(3):211.

Zablocki, K. (1991). Accumulation of glycerophosphocholine (GPC) by renal cells:

osmotic regulation of GPC: choline phosphodiesterase. *Proceedings of the National Academy of Sciences*, 88(17):7820–7824.

Zeisberg, M., Hanai, J., Sugimoto, H., Mammoto, T., Charytan, D., Strutz, F., and Kalluri, R. (2003). BMP-7 counteracts TGF-β1-induced epithelial-to-mesenchymal transition and reverses chronic renal injury. *Nature medicine*, 9(7):964–968.

Zhang, Y., Sutton, C., Storkey, A., and Ghahramani, Z. (2012). Continuous Relaxations for Discrete Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems 25*, pages 3203–3211.

Zhou, X. and Schmidler, S. C. (2009). Bayesian Parameter Estimation in Ising and Potts Models: A Comparative Study with Applications to Protein Modeling. Technical report, Technical report, Duke University.(a).

Zhu, G., Salazar, G., Zlatic, S., Fiza, B., Doucette, M., Heilman, C., Levey, A., Faundez, V., and L'Hernault, S. (2009). SPE-39 family proteins interact with the HOPS complex and function in lysosomal delivery. *Molecular biology of the cell*, 20(4):1223–1240.