PhD Dissertation

Tackling the development of hormone therapy resistance in breast cancer through mathematical modelling

Martín Parga Pazos

Advisors
Prof. Elena Akhmatskaya
Dra. María del Mar Vivanco

2024

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

PhD Dissertation

Tackling the development of hormone therapy resistance in breast cancer through mathematical modelling

Martín Parga Pazos

Advisors
Prof. Elena Akhmatskaya
Dra. María del Mar Vivanco

2024

basque center for applied mathematics

EXCELENCIA SEVERO OCHOA

CIC bioGUNE
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Abstract

Patients suffering from estrogen-driven breast cancer frequently develop hardly predictable resistance to hormone therapy, which significantly complicates treatment. Current approaches for tackling this problem include cell models and clinical studies, both supported by sequencing technologies like RNA-seq, and offering different strengths and limitations. This dissertation addresses the challenge of predicting resistance to hormone therapy in breast cancer by merging advances in bioinformatics and Bayesian statistics, and applying them to two types of data – RNA-seq data and clinical data. First, we explore the statistical analysis of clinical data through Bayesian inference combined with enhanced Markov Chain Monte Carlo techniques, and introduce a novel algorithm for adaptive integration in prospective Modified Hamiltonian Monte Carlo (MHMC) methods. We demonstrate its positive effect on performance of MHMC in biomedical applications using clinical data of breast cancer patients. Next, we propose and implement an RNA-seq pipeline within our interactive web-app for the analysis of resistant breast cancer cell lines sequenced at CIC bioGUNE. Finally, we propose an original approach based on a Bayesian logistics regression model coupled with a simulated annealing-like algorithm for a combined analysis of RNA-seq and clinical data, and apply it to ad hoc data to obtain and validate in-silico and in-vitro a novel 6-gene signature for stratifying patient response to hormone therapy.

This research was supported by the Spanish State Research Agency and the Ministry of Science and Innovation through BCAM Severo Ochoa accreditation CEX2021-001142-S/MICINN/AEI/10.13039/501100011033, PID2019-104927GB-C22, PID2022-136585NB-C22 and CEX2021-001136-S/MCIN/AEI/10.13039/501100011033 (M.V.), PID2020-118464RB I00 (M.V:); by the Basque Government through the BERC 2022–2025 program, the Elkartek program (KK-2020/00008, KK-2022/00006), and the BMTF project (Mathematical Modelling applied to health).

Summary

Patients suffering from estrogen-driven (ER+) breast cancer frequently develop resistance to hormone therapy, which represents a significant clinical challenge. Important contributions have been made in unveiling the mechanisms behind this resistance by utilizing cell models and clinical studies. However, both approaches present limitations when taken individually. On the one hand, experiments with cell models are highly reproducible but do not reflect the indubitable heterogenous landscape of breast cancer. On the other hand, clinical studies account for this complexity but introduce uncontrolled noise due to external factors. Moreover, the types of data used in these studies are very different in terms of how they were obtained, processed and analysed. This thesis addresses these challenges by studying cell and clinical data individually, as well as in combination, to gain new insights into breast cancer therapy resistance mechanisms. RNA-sequencing data serves as a common platform for the integration of these datasets.

The thesis has two major objectives. Objective 1 focuses on the development of new methods and tools for the study of several *ad hoc* datasets related to the resistance problem. Our contributions regarding this objective can be found in Chapters 2, where we focus on clinical data, and Chapter 3, centered in RNA-sequencing data. In Chapter 4, we address Objective 2, in which we utilize the methods and tools we have explored and developed for Objective 1 with the aim of obtaining a biologically meaningful result related to the problem of resistance to hormone therapy in breast cancer.

After an introductory Chapter 1 where we lay down the context and objectives of the thesis, in Chapter 2 we explore the statistical analysis of clinical data through a Bayesian lens. In this context, Bayesian statistics offer an excellent framework for biomedical data due to its unique features and characteristics, such as the ability to include information known *a priori* in the model, through the use of prior distributions, or the built-in uncertainty quantification, offered directly by the posterior distributions. At the same time, it also contains many practical advantages, like a more straightforward interpretation of significance tests or a direct regularization approach. Both of which can be exploited to obtain directly results that need further assumptions in frequentist analyses, as we demonstrate at the end of this chapter. In general, the Bayesian approach is used for three different tasks, namely, calculation of model parameters (via marginalization), prediction of incoming data and hypothesis testing. All of these are explored in this thesis using two *ad-hoc* datasets related to the problem of resistance to endocrine therapy.

Despite its advantages, one of the major drawbacks halting a wider use of Bayesian statistics is the intensive computational demand that is often required by Bayesian models. The Bayesian methodology involves the calculation of complex integrals for essentially any practical application. For example, the evaluation of an expected value over a given posterior distribution is central to many Bayesian applications. For most practical cases, these integrals have no analytical solutions and require the use of numerical methods. Moreover, the complexity greatly increases for high-

dimensional parameter spaces. Consequently, numerical integration techniques are often essential for approximating posterior distributions in Bayesian methods. Recent advancements in computing power, coupled with novel numerical algorithms, have facilitated the development of methods that address these challenges more effectively.

Sampling algorithms play this crucial role in Bayesian statistics by enabling the evaluation of complex posterior distributions. Markov Chain Monte Carlo (MCMC) algorithms offer flexibility and accuracy in this task and, as a result, have emerged as primary solutions for the computational issues frequently associated with Bayesian problems. MCMC methods create a Markov chain whose invariant distribution is a target distribution $\pi(\boldsymbol{\theta})$, which in this case is the posterior of the Bayesian problem. By iteratively generating samples from the target distribution, MCMC methods explore the parameter space efficiently and obtain representative samples that characterize the shape and uncertainty of the posterior distribution. Advances in sampling algorithms, including more sophisticated variants like Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 1994), have further improved the efficiency and convergence properties of Bayesian inference, making it more accessible and practical for a broader range of applications. The main result of Chapter 2 focuses precisely on enhancements to numerical integration in Hamiltonian Monte Carlo-based methods.

Hamiltonian Monte Carlo, as an MCMC method, is an algorithm that samples from the target canonical distribution $\pi(\boldsymbol{\theta}) \sim \exp(-\beta H(\boldsymbol{\theta}, \mathbf{p}))$, with a Hamiltonian function H, defined as

$$H(\boldsymbol{\theta}, \mathbf{p}) = K + U = \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} + U(\boldsymbol{\theta}) \equiv A + B,$$

where the potential energy term $U(\boldsymbol{\theta})$ is related to the target distribution $\pi(\boldsymbol{\theta})$ as $U(\boldsymbol{\theta}) = -\log \pi(\boldsymbol{\theta}) +$ const., and K is the kinetic energy.

The HMC proposals are obtained by numerically integrating the Hamiltonian equations of motion, which results in larger and thus less correlated moves across the sample space compared with traditional MCMC. Besides, using gradients of the potential energy leads to faster convergence to the target distribution. Clearly, integration is an essential part of the algorithm and its performance can be greatly improved by taking advantage of the separable form of the Hamiltonian and considering multi-stage splitting integration schemes, such as 2- and 3-stage families (Blanes et al., 2008, 2014; Radivojević et al., 2018) defined as

$$\begin{aligned}
\Psi_h^{2-s} &= \psi_{bh}^B \circ \psi_{h/2}^A \circ \psi_{(1-2b)h}^B \circ \psi_{h/2}^A \circ \psi_{bh}^B \\
\Psi_h^{3-s} &= \psi_{bh}^B \circ \psi_{ah}^A \circ \psi_{(\frac{1}{2}-b)h}^B \circ \psi_{(1-2a)h}^A \circ \psi_{(\frac{1}{2}-b)h}^B \circ \psi_{bh}^B \circ \psi_{ah}^A.
\end{aligned}$$

The schemes are constructed as palindromic compositions of solutions flows explicitly given by

$$\psi_h^A(\boldsymbol{\theta}, \mathbf{p}) = (\boldsymbol{\theta} + hM^{-1}\mathbf{p}, \mathbf{p}), \quad \psi_h^B(\boldsymbol{\theta}, \mathbf{p} - h\nabla_\theta U(\boldsymbol{\theta})).$$

and associated with the split system

$$\begin{aligned}
A: \quad & \dot{\boldsymbol{\theta}} = \nabla_p K(\mathbf{p}) = M^{-1}\mathbf{p} && \dot{\mathbf{p}} = -\nabla_\theta K(\mathbf{p}) = \mathbf{0} \\
B: \quad & \dot{\boldsymbol{\theta}} = \nabla_p U(\boldsymbol{\theta}) = \mathbf{0} && \dot{\mathbf{p}} = -\nabla_\theta U(\boldsymbol{\theta}),
\end{aligned}$$

In Chapter 2, we focus on the recent variation of Hamiltonian Monte Carlo — Mix & Match HMC, or MMHMC (Radivojević and Akhmatskaya, 2020). This is a generalized HMC importance sampler which, in contrast to HMC, is irreversible by construction, uses Modified Hamiltonians in

the Metropolis test and performs partial momentum refreshment. The method demonstrated superior sampling performance over conventional HMC in various Bayesian models (Radivojevic, 2016). Our objective is to further enhance accuracy and sampling efficiency of MMHMC through 2- and 3-stage adaptive splitting integration schemes specifically developed for modified Hamiltonian Monte Carlo methods.

Our methodology, which we call s-MAIA (i.e., statistical Modified Adaptive Integration Approach), provides the optimal choice of parameters — b or (a, b) — for these 2 and 3-stage splitting integrators, in terms of the best conservation of modified energy for Gaussian targets. To achieve this, we propose a method which combines analysis of expected acceptance rates and expected modified energy error of multivariate Gaussian models with the data generated at the burn-in stage of an MMHMC simulation to calculate the optimal parameters for 2- and 3-stage splitting integrators, as well as dimensional stability limits for any given model and integration stepsize.

After defining the appropriate metrics for the analysis, for which we introduce a novel approach to Effective Sample Size (ESS) estimation for irreversible MCMC importance sampling, we show numerical experiments confirming the superior sampling capabilities of MMHMC combined with the proposed s-MAIA methodology over MMHMC with the state of the art modified integrators and HMC with its best integration schemes. We conclude Chapter 2 by applying MMHMC coupled with s-MAIA to biomedical problems, with the emphasis on the breast cancer case study.

Chapter 3 focuses mainly on the analysis of RNA-sequencing data, using for that a cell line dataset provided by the Cancer Heterogeneity Lab at CIC bioGUNE (bG-RNA-seq dataset). We introduce the foundations of an analysis pipeline for RNA-seq data, starting from an overview of the sequencing process, explaining how data is processed and genes are quantified. The result of this process is a gene count matrix, a $G \times M$ matrix (G is the total amount of genes identified and M is the number of samples), where each value for a gene g in a sample m accounts for the number of times a gene has been identified in the sequencing process. In the rest of the chapter, we explore the mathematical methods needed to extract the biological information encapsulated in it.

Sequencing has multiples sources of variability attached to the process, from technical biases related to the equipment or sample preparation, to inherent biological variability. Because of this, the counts obtained for the gene counts matrix need to go through a process of normalization that makes them comparable across samples and allows for accurate comparisons of gene expression levels between samples. We explore 4 different methods for normalization and provide a comparison of their accuracy and efficiency using the bG-RNA-seq dataset.

Based on the comparison results, we select the Relative log expression (RLE) normalization method which is also, in turn, part of the broader methodology DESeq2 (Love et al., 2014), commonly used for differential expression analysis (DEA). DEA is the key component of RNA-seq data analysis employed to identify genes that are significantly up- or down-regulated between experimental conditions. This analysis involves statistical comparisons of gene expression levels between different groups and it typically applies a negative binomial distribution to model the count data. For each gene g in sample m, it is possible to fit a generalised linear model (GLM) and use the resulting coefficient $\beta_{g,m}$ to perform the comparison between genes and conditions and obtain a result for differential expression. This is followed by the identification of statistically significant changes in gene expression.

From DEA analysis results, plenty of functional analyses can be performed. These play a crucial role in extracting meaningful biological insights from high-throughput data. While differential expression analysis identifies genes that are significantly altered between conditions, functional analyses provide a deeper understanding of the underlying biological processes, pathways, and functions associated with these gene sets. Among them, we explore mainly Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) and pathway analysis (Khatri et al., 2012).

Chapter 3 concludes with the presentation of a bioinformatics web tool, called Vivanco LabSeq,

which represents our central result in this area. It condenses most of the analysis introduced in this chapter and facilitates efficient online access for researchers at the Cancer Heterogeneity Lab. The platform of choice for the development of this web-app is R-Shiny, which allows for the accessibility and easy-to-use features essential for our objective. At the same time, this package for web development is native to R where most of the packages for analysis presented in this thesis are hosted.

The Vivanco LabSeq tool offers access to the in-house cell line bG-RNA-seq data sourced from the Cancer Heterogeneity Lab and to various tools necessary for analyses of this data. The tool supplies interactive tables with DEA and GSEA results, as well as interactive plots for both of these analyses. Though the Vivanco LabSeq tool was initially aimed at providing support for the research of the Cancer Heterogeneity Lab, the core of the platform could easily be adapted to host external data.

In Chapter 4, we integrate two types of data — clinical and RNA-seq — in one study and propose a new methodology for biomarker discovery that is based on a combined analysis of sequencing data from controlled cell experiments and heterogeneous clinical samples, that include clinical and sequencing information from The Cancer Genome Atlas (TCGA), a widely used public repository for cancer patients data. First, the clinical data is classified between those patients responding well and those showing resistance to the therapy. Further, we perform DEA analysis between both groups, as well as between the resistant and non-resistant cells to obtain a set of relevant genes significant in both analyses. The initial set of genes is analysed and ultimately refined using a Bayesian logistic regression model coupled with an original simulated annealing-type algorithm, which incorporates Bayesian model selection techniques to improve predictive accuracy of the posterior distribution.

The central result of this Chapter is the discovery of a novel 6-gene signature able to stratify patients response to hormone therapy, the task which constitutes Objective 2. The signature, which we call EnCaRes (Endocrine Cancer Resistance) undergoes several computational validations, using two more independent cohorts of patients. The results are validated using survival analysis, which indicates a decrease of 14% in the 10-year relapse free survival probability for hormone therapy-treated tumours with high expression levels of the gene signature. Moreover, we use Cox proportional hazard models to confirm the competence of the signature for classifying patients over other clinical covariates, but also indicate the superior predictive performance of this gene set over previously known signatures of similar scope. The signature also withstands experimental validation in MCF7 resistant cells. In conclusion, our findings revealed a new gene signature for identification of patients with breast cancer with an increased risk of developing resistance to endocrine therapy, which outperforms the known resistance-related gene signatures.

In conclusion, this thesis explores and develops techniques for mathematical modelling and bioinformatic analysis, and utilizes them to offer insights into the mechanisms of drug resistance in breast cancer treatment, uncovering a signature of biomarkers that can be used to accurately predict the risk of patients developing this resistance.

Resumen

Los pacientes que padecen cáncer de mama sensible a estrógenos (ER+) desarrollan con frecuencia resistencia a la terapia hormonal, lo que representa un importante reto clínico. Se han realizado importantes contribuciones para desvelar los mecanismos que subyacen a esta resistencia utilizando modelos celulares y estudios clínicos. Sin embargo, ambos enfoques presentan sus respectivas limitaciones. Por un lado, los experimentos con modelos celulares son fácilmente reproducibles, pero no consiguen abarcar inherente heterogeneidad del cáncer de mama. Por otro lado, los estudios clínicos tienen en cuenta esta complejidad, pero añaden información no relevante para el problema debido multitud de factores externos. Además, los tipos de datos que se utilizan en estos estudios son muy diferentes en cuanto a la forma de obtenerlos, procesarlos y analizarlos. Esta tesis aborda los retos que se presentan en este contexto, estudiando datos celulares y clínicos individualmente, así como en conjunto, para obtener ahondar en los mecanismos de resistencia a la terapia del cáncer de mama. Para ambos casos, los datos de secuenciación de ARN (RNA-seq) sirven plataforma común para la integración de datos celulares y de pacientes.

Esta tesis tiene dos objetivos principales. El Objetivo 1 se centra en el desarrollo de nuevos métodos y herramientas para el estudio de varios conjuntos de datos *ad hoc* relacionados con el problema de resistencia a terapia hormonal. Nuestras contribuciones en este aspecto se abordan en el Capítulo 2, donde nos centramos en los datos clínicos, y en el Capítulo 3, centrado en los datos de secuenciación de ARN. En el Capítulo 4, abordamos el Objetivo 2, en el que recurrimos a los métodos y herramientas que hemos explorado y desarrollado para el Objetivo 1 con el fin de obtener un resultado significativo desde el punto de vista biológico relacionado con el problema de la resistencia a la terapia hormonal en el cáncer de mama.

Tras el Capítulo 1, que sirve de introducción para establecer el contexto y los objetivos de la tesis, en el Capítulo 2 exploramos el análisis estadístico de datos clínicos a través de una perspectiva Bayesiana. La estadística Bayesiana ofrece un marco metodológico excelente para los datos biomédicos debido a sus singulares propiedades y características, como son la capacidad de incluir información conocida de antemano en el modelo, mediante el uso de distribuciones *a priori*, o la cuantificación de incertidumbre que ofrecen directamente las distribuciones *a posteriori*. Al mismo tiempo, también posee muchas ventajas prácticas, como una interpretación más directa de las pruebas de significación estadística o una forma de aplicar la regularización a modelos directamente desde el mismo método. Ambas pueden aprovecharse para obtener directamente resultados que necesitan la asunción de ciertas hipótesis en los análisis estadísticos clásicos, como demostramos al final de este capítulo. En general, el enfoque Bayesiano se utiliza para tres tareas diferentes como son el cálculo de parámetros del modelo (mediante marginalización), la predicción de los nuevos datos y la validación de hipótesis. En esta tesis exploramos todas ellas utilizando dos conjuntos de datos *ad-hoc* relacionados con el problema de la resistencia a la terapia endocrina.

A pesar de sus ventajas, uno de los principales inconvenientes que frenaron durante mucho

tiempo el uso de los modelos Bayesianos es la intensa demanda computacional que suelen requerir. La metodología Bayesiana trae consigo el cálculo de integrales complejas para prácticamente cualquier aplicación práctica. Por ejemplo, la evaluación del valor esperado sobre una distribución *a posteriori* es algo fundamental para muchas aplicaciones Bayesianas. Sin embargo, para la mayoría de los casos prácticos, estas integrales no tienen soluciones analíticas y requieren el uso de métodos numéricos. Además, la complejidad aumenta considerablemente en espacios de alta dimensión. Por lo tanto, las técnicas de integración numérica suelen ser esenciales para aproximar las distribuciones *a posteriori* obtenidas con los estos métodos. En los últimos años, ha habido un gran salto en potencia de cálculo que, junto con el desarrollo de nuevos algoritmos para integración numérica, ha facilitado el desarrollo de técnicas para abordar problemas en estadística Bayesiana de una manera más fácil y accesible.

Los algoritmos diseñados para extraer muestras de una distribución desempeñan este papel crucial en la estadística Bayesiana al permitir la evaluación de las complejas distribuciones *a posteriori* que mencionábamos previamente. Los algoritmos de Markov Chain Monte Carlo (MCMC) ofrecen flexibilidad y precisión en esta tarea y, en consecuencia, se presentan como la principal solución para solucionar los problemas computacionales frecuentemente asociados a los problemas Bayesianos. Los métodos MCMC generan una cadena de Markov cuya distribución invariante es una distribución objetivo $\pi(\boldsymbol{\theta})$, que en este caso es la distribución *a posteriori* del problema planteado. Mediante la generación iterativa de muestras de la distribución objetivo, los métodos MCMC son capaces de explorar el espacio de forma eficiente y obtienen muestras representativas que caracterizan la forma y la incertidumbre de la distribución *a posteriori*. Los avances en estos algoritmos para obtener muestras, incluidas variantes más sofisticadas como Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 1994), han mejorado todavía más su eficiencia, haciendo que los métodos Bayesianos sean cada vez más accesibles y utilizados en una amplia gama de aplicaciones. El principal resultado del Capítulo 2 se centra, precisamente, en las mejoras de la integración numérica en los métodos basados en Hamiltonian Monte Carlo.

Hamiltonian Monte Carlo, en un algoritmo que, como el resto de MCMC, genera muestras de una distribución objetivo $\pi(\boldsymbol{\theta}) \sim \exp(-\beta H(\boldsymbol{\theta}, \mathbf{p}))$, donde la función Hamiltoniana H se define como

$$H(\boldsymbol{\theta}, \mathbf{p}) = K + U = \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} + U(\boldsymbol{\theta}) \equiv A + B,$$

siendo $U(\boldsymbol{\theta})$ el término de energía potencial que se relaciona con la distribución objetivo $\pi(\boldsymbol{\theta})$ de la forma $U(\boldsymbol{\theta}) = -\log \pi(\boldsymbol{\theta}) + \text{const.}$, y K el término de energía cinética.

HMC genera propuestas integrando numéricamente las ecuaciones de movimiento Hamiltonianas, lo que da lugar a movimientos más largos y, por tanto, menos correlacionados a través del espacio en comparación con los métodos clásicos de MCMC. Además, el uso de gradientes de la energía potencial produce una convergencia más rápida a la distribución objetivo. En consecuencia, la integración numérica es una parte esencial del algoritmo y su rendimiento puede mejorarse enormemente aprovechando la forma separable del Hamiltoniano y considerando esquemas de integración en varias etapas, como las familias de 2 y 3 etapas (Blanes et al., 2008, 2014; Radivojević et al., 2018) definidas de la forma

$$\begin{aligned}
\Psi_h^{2-s} &= \psi_{bh}^B \circ \psi_{h/2}^A \circ \psi_{(1-2b)h}^B \circ \psi_{h/2}^A \circ \psi_{bh}^B \\
\Psi_h^{3-s} &= \psi_{bh}^B \circ \psi_{ah}^A \circ \psi_{(\frac{1}{2}-b)h}^B \circ \psi_{(1-2a)h}^A \circ \psi_{(\frac{1}{2}-b)h}^B \circ \psi_{bh}^B \circ \psi_{ah}^A.
\end{aligned}$$

Estos esquemas de integración se construyen como composiciones palindrómicas de los flujos dados por

$$\psi_h^A(\boldsymbol{\theta}, \mathbf{p}) = (\boldsymbol{\theta} + hM^{-1}\mathbf{p}, \mathbf{p}), \quad \psi_h^B(\boldsymbol{\theta}, \mathbf{p} - h\nabla_\theta U(\boldsymbol{\theta})).$$

y asociados con el sistema separable

$$
\begin{aligned}
A: \quad & \dot{\boldsymbol{\theta}} = \nabla_p K(\mathbf{p}) = M^{-1}\mathbf{p} && \dot{\mathbf{p}} = -\nabla_\theta K(\mathbf{p}) = \mathbf{0} \\
B: \quad & \dot{\boldsymbol{\theta}} = \nabla_p U(\boldsymbol{\theta}) = \mathbf{0} && \dot{\mathbf{p}} = -\nabla_\theta U(\boldsymbol{\theta}),
\end{aligned}
$$

En el Capítulo 2, nos centramos en la reciente variante de Hamiltonian Monte Carlo - Mix & Match HMC, o MMHMC (Radivojević and Akhmatskaya, 2020). Se trata de un método HMC generalizado basado en *importance sampling* que, a diferencia de HMC, es irreversible por construcción, utiliza Hamiltonianos Modificados en el test de Metropolis y realiza una actualización parcial del momento en cada iteración. El método muestra un rendimiento superior a las variantes HMC convencional en varios modelos Bayesianos (Radivojevic, 2016). Nuestro objetivo es mejorar aún más la precisión y la eficiencia del método de MMHMC mediante esquemas de integración adaptativos de 2 y 3 etapas desarrollados específicamente para los métodos de Monte Carlo usando Hamiltonianos Modificados.

El nuevo método que proponemos, que denominamos s-MAIA, selecciona los paramétros óptimos - b o (a, b) - para estos integradores en 2 y 3 etapas, en términos de la mejor conservación de energía (modificada) en distribuciones Gaussianas. Para lograrlo, proponemos un método que combina el análisis de las tasas de aceptación del test de Metropolis y el error en la energía modificada de modelos gaussianos. Para ello, usamos los datos generados en la etapa de preproducción de una simulación MMHMC para calcular los parámetros óptimos de los integradores en 2 y 3 etapas, así como los límites de estabilidad dimensional para cualquier modelo y tamaño de paso de integración dados por el problema.

Tras definir las métricas apropiadas para el análisis, para lo cual introducimos un enfoque novedoso de estimación del *Effective Sample Size* (ESS) para métodos MCMC irreversibles que usan *importance sampling*. Mostramos experimentos numéricos que confirman la capacidad superior de muestreo de MMHMC combinado con la metodología s-MAIA propuesta sobre MMHMC con otros integradores avanzados y sobre HMC con sus mejores esquemas de integración. Concluimos el Capítulo 2 aplicando MMHMC acoplado con s-MAIA a problemas biomédicos, con énfasis en el caso de estudio del cáncer de mama.

El capítulo 3 se centra principalmente en el análisis de datos de secuenciación de ARN (RNA-seq), utilizando para ello un conjunto de datos de líneas celulares proporcionado por el *Cancer Heterogeneity Lab* del CIC bioGUNE (bG-RNA-seq). Introducimos los fundamentos para el análisis de datos de RNA-seq, comenzando por una visión general del proceso de secuenciación, explicando cómo se procesan los datos y se cuantifican los genes. El resultado de este proceso es una matriz de cuentas de genes, una matriz $G \times M$ (donde G es la cantidad total de genes identificados y M es el número de muestras biológicas), donde cada valor para un gen g en una muestra m representa el número de veces que se ha identificado un gen en el proceso de secuenciación. En el resto del capítulo, exploramos los métodos matemáticos necesarios para extraer la información biológica subyacente encapsulada en él.

Durante el proceso de secuenciación existen múltiples partes que pueden convertirse en fuentes de variabilidad para el resultado final, desde sesgos técnicos relacionados con el equipo o la preparación de la muestra, hasta la propia variabilidad inherente a las muestras biológicas. Debido a esto, los datos obtenidos para la matriz de cuentas de genes necesitan pasar por un proceso de normalización que los haga comparables entre distintas muestras y permita comparaciones precisas de los niveles de expresión génica entre ellas. En el capítulo 3, exploramos 4 métodos diferentes para esta normalización y proporcionamos una comparación de su precisión y eficacia utilizando el conjunto de datos bG-RNA-seq.

Basándonos en los resultados de la comparación, seleccionamos el método de normalización *Relative log expression* (RLE) que, a su vez, también forma parte de una metodología más amplia, DESeq2 (Love et al., 2014), utilizada habitualmente para el análisis de expresión diferencial (DEA).

El análisis DEA es el componente clave del análisis de datos de RNA-seq empleado para identificar genes que están significativamente regulados al alza o a la baja entre condiciones experimentales. Este análisis implica realizar comparaciones estadísticas entre los niveles de expresión génica entre diferentes grupos. Para ellos, suele aplicarse una distribución binomial negativa para modelar los datos de cuentas de genes. Para cada gen g de la muestra m, es posible ajustar un modelo lineal generalizado (GLM) y utilizar el coeficiente resultante $\beta_{g,m}$ para realizar la comparación entre genes y condiciones y obtener un resultado de expresión diferencial. A partir de ello se identifican los cambios en la expresión génica y se realizan test de hipótesis para discernir si las alteraciones son estadísticamente significativas.

A partir de los resultados del análisis DEA, pueden realizarse numerosos análisis funcionales que van más allá de los cambios de expresión de genes entre muestras. Estos análisis expanden la utilidad de los datos de RNA-seq y desempeñan un papel crucial en la extracción de conocimiento sobre el estado de la muestra y sus procesos biológicos internos. Mientras que el análisis de expresión diferencial identifica los genes que están significativamente alterados entre las condiciones, los análisis funcionales proporcionan una comprensión más profunda de los procesos biológicos subyacentes como las rutas que siguen ciertos genes y las funciones asociadas con estos conjuntos de genes. Entre ellos, exploramos principalmente el análisis de enriquecimiento de conjuntos de genes (GSEA) (Subramanian et al., 2005) y el análisis de *pathways* (Khatri et al., 2012).

El capítulo 3 concluye con la presentación de una herramienta web para bioinformática, denominada Vivanco LabSeq, que representa nuestro resultado central en este ámbito. Condensa la mayor parte de los análisis introducidos en este capítulo y facilita un acceso en línea a los investigadores del *Cancer Heterogeneity Lab* del CIC bioGUNE. La plataforma elegida para el desarrollo de esta web-app es R-Shiny, que nos ofrece las características de accesibilidad y facilidad de uso que creíamos esenciales para desarrollar este objetivo. Al mismo tiempo, este paquete para el desarrollo web es nativo de R, donde se alojan la mayoría de los paquetes de análisis bioinformático presentados a lo largo de esta tesis.

En el Capítulo 4, integramos dos tipos de datos - clínicos y de RNA-seq - en un estudio en el que proponemos una nueva metodología para el descubrimiento de biomarcadores. Nuestra propuesta se basa en un análisis combinado de datos de secuenciación de experimentos celulares homogéneos y muestras clínicas heterogéneas, que incluyen información clínica y de secuenciación de la base de datos The Cancer Genome Atlas (TCGA), un repositorio público ampliamente utilizado para datos de pacientes con cáncer. En primer lugar, los datos clínicos se clasifican entre los pacientes que responden bien y los que muestran resistencia a la terapia. Además, realizamos un análisis DEA entre ambos grupos, así como entre las células resistentes y no resistentes para obtener un conjunto de genes relevantes significativos en ambos análisis. El conjunto inicial de genes se analiza y finalmente se refina mediante un modelo Bayesiano en el que usamos una regresión logística acoplada a un algoritmo original basado en *simulated annealing*, que incorpora técnicas bayesianas de selección de modelos para mejorar la precisión en la predicción de la distribución *a posteriori* resultante de nuestro modelo.

El resultado central de este capítulo es el descubrimiento de una nueva firma de 6 genes capaz de estratificar la respuesta de los pacientes a la terapia hormonal y establecer grupos de riesgo. Esto constituye el resultado buscado de cara al Objetivo 2 previamente introducido. La firma, que denominamos EnCaRes (Endocrine Cancer Resistance) es sometida a varias validaciones computacionales, utilizando otros dos conjuntos de datos de pacientes independientes. Los resultados se validan mediante análisis de supervivencia, que indica una disminución del 14% en la probabilidad de supervivencia sin recaída a 10 años para los tumores tratados con hormonoterapia y que poseen altos niveles de expresión de la firma génica encontrada. Además, utilizamos modelos de riesgos de Cox para confirmar el buen rendimiento de la firma para identificar el riesgo de los pacientes, estando por encima de otras covariables clínicas, pero de otras firmas previamente conocidas para

este problema, mostrando una capacidad predictiva superior a otras firmas. Nuestra firma también resiste la validación experimental en células MCF7 resistentes al tamoxifeno. En conclusión, nuestros hallazgos revelan una nueva firma génica para la identificación de pacientes con cáncer de mama con un mayor riesgo de desarrollar resistencia a la terapia endocrina, que supera a las firmas génicas conocidas hasta ahora y relacionadas con la resistencia.

En conclusión, esta tesis explora y desarrolla técnicas de modelización matemática y análisis bioinformático, y las utiliza para ofrecer una visión de los mecanismos de resistencia a terapias hormonales en el tratamiento del cáncer de mama, descubriendo una firma de 6 biomarcadores que puede utilizarse para predecir con precisión el riesgo de que las pacientes desarrollen esta resistencia.

Contents

1 Introduction **1**

1.1 Cancer research and data analysis . . . 3
1.1.1 Clinical data . . . 3
1.1.2 High-throughput data . . . 7
1.2 Problem at hand: Resistance to hormone therapy in breast cancer . . . 9
1.2.1 Heterogeneity of breast cancer . . . 9
1.2.2 Drug resistance in breast cancer . . . 11
1.2.3 Gene signatures for breast cancer . . . 12
1.3 Objectives . . . 13

2 Statistical Analysis of *ad hoc* Clinical Data **16**

2.1 Collection and treatment of datasets . . . 17
2.1.1 SOX2 Clinical Dataset . . . 17
2.1.1.1 Data description and collection . . . 17
2.1.1.2 Data curation . . . 19
2.1.2 TCGA Clinical dataset . . . 21
2.1.2.1 Data description and collection . . . 21
2.1.2.2 Data curation . . . 22
2.2 Mathematical modelling of clinical data . . . 23
2.2.1 Bayesian inference . . . 23
2.2.2 Markov Chain Monte Carlo Methods . . . 24
2.2.2.1 Metropolis-Hastings algorithm . . . 26
2.2.3 Hamiltonian Monte Carlo (HMC) . . . 27
2.2.4 Enhancing performance of HMC: Irreversibility . . . 30
2.2.4.1 Generalized Hamiltonian Monte Carlo (GHMC) . . . 30
2.2.4.2 Mix & Match Hamiltonian Monte Carlo (MMHMC) . . . 31
2.2.5 Enhancing performance of HMC: Numerical integrators . . . 34
2.2.6 Performance evaluation of HMC methods . . . 37
2.3 Boosting sampling in biomedical applications with the reinforced MMHMC method . . . 42
2.3.1 Effective Sample Size for importance sampling HMC . . . 42
2.3.2 Modified Adaptive Integration Approach for MMHMC (s-MAIA) . . . 47
2.3.2.1 Objective . . . 47
2.3.2.2 Upper bound for expected modified energy error with respect to modified density . . . 47
2.3.2.3 Nondimensionalization of the stepsize . . . 49
2.3.2.4 Algorithm . . . 54
2.3.2.5 Implementation . . . 54

2.3.2.6 Performance evaluation: Comparison with state-of-the-art modified integrators . . . 57
2.3.2.7 Performance evaluation: Optimal MMHMC vs Optimal HMC . . . 59
2.3.3 Applications . . . 60
2.3.3.1 Wisconsin Breast Cancer dataset — A biomedical benchmark . . . 61
2.3.3.2 SOX2 Clinical dataset — Posterior analyses . . . 61
2.4 Conclusions . . . 65

3 Bioinformatic Tools for *ad hoc* RNA-Seq Data 68
3.1 Pre-processing . . . 69
3.1.1 Negative binomial model . . . 70
3.2 RNA-seq datasets . . . 71
3.2.1 Cell line bG-RNA-seq data . . . 72
3.2.1.1 Data description . . . 72
3.2.1.2 Data exploration . . . 73
3.2.2 TCGA patients RNA-seq data . . . 75
3.3 RNA-seq data analysis . . . 75
3.3.1 Normalization . . . 75
3.3.2 Comparison of normalization techniques . . . 78
3.3.3 Differential Expression Analysis . . . 79
3.4 Functional analysis . . . 82
3.4.1 Gene Set Enrichment Analysis . . . 82
3.4.1.1 GSEA — Classical analysis . . . 82
3.4.1.2 Enrichr — Bulk analysis . . . 83
3.4.2 Pathway analysis . . . 83
3.5 A practical application: Vivanco LabSeq tool . . . 84
3.6 Conclusions . . . 89

4 Integration of Clinical and Transcriptomic Data for Biomarker Identification 92
4.1 Current landscape . . . 92
4.2 Methodology . . . 94
4.3 Data collection and classification . . . 95
4.3.1 MCF7 cell lines . . . 95
4.3.2 TCGA patients dataset . . . 96
4.3.2.1 Selection of clinical cases . . . 96
4.3.2.2 Classification of patients responses . . . 97
4.4 Statistical analysis of sequencing data . . . 97
4.4.1 Identification of common biomarkers . . . 99
4.4.2 BLR-LOO model . . . 99
4.4.3 Simulated Annealing-type algorithm . . . 102
4.5 Validation . . . 105
4.5.1 Validation cohorts . . . 105
4.5.2 Survival analysis . . . 105
4.5.2.1 Kaplan–Meier curves . . . 105
4.5.2.2 Cox proportional hazard models . . . 107
4.5.3 Experimental validation . . . 109
4.6 Biological implications . . . 110
4.7 Conclusions . . . 112

5 Conclusions, Main Contributions and Future Work 114
5.1 Conclusions . . . 114

5.2 Main Contributions 115
5.3 Future Work 119

A SOX2 Clinical Dataset **122**
A.1 Original SOX2 Clinical dataset 122
A.2 SOX2 Clinical dataset after data imputation 124

B Supplementary information for EnCaRes **128**
B.1 Individual survival analysis of the genes in EnCaRes 128
B.2 Clinical information of selected patients 129

Bibliography **132**

List of Figures

1.1 Schematic representation of the major breakthroughs in biology and computer science. . 3
1.2 Collection of relevant genes for each of the major molecular subtypes from Nolan et al. (2023) (left) and illustration of the antagonist mechanism of tamoxifen (right). Adapted from "Tamoxifen Mechanism of Action in Breast Cancer", by BioRender.com (2023). . . . 12
1.3 Thesis outline. 14

2.1 General pipeline for the mathematical analysis of clinical data. 16
2.2 Distribution of patient's response to tamoxifen treatment. All patients with an Allred score of 3 or more show a resistant response to the treatment. 19
2.3 Pattern plot for missing data in the original SOX2 Clinical dataset. 20
2.4 Pattern plot for missing data in the SOX2 Clinical dataset after removing patients with more than 4 variables missing. 21
2.5 Density plots for the possible imputation values for missing data. Generated using *mice*. 21
2.6 Behaviour of several ESS methods in HMC, GHMC and MMHMC (unweighted) scenarios. 45
2.7 Comparison of several ESS metrics with the reference value obtained using i.i.d. samples from a 1000-D multivariate Gaussian. 46
2.8 Comparison of possible combinations of MCMC and IS estimators within the MCIS estimator. 46
2.9 Values of the coefficient b for 2-stage integrators (left) and 3-stage integrators (right). 2-, 3- s-MAIA parameters vs h are shown in green. Stepsize values for h are normalized with respect to the number of stages to fit in the VV stability interval. 55
2.10 Values of the $\rho_r(h, \Gamma)$ function (in $\log_{10}$ scale) for the 2-stage integrators (left) and 3-stage integrators (right). 2-, 3- s-MAIA parameters vs h are shown in green. Stepsize values for h are normalized with respect to the number of stages to fit in the VV stability interval. . 56
2.11 Values of the coefficient b after correction for 2-stage integrators (left) and 3-stage integrators (right). 2-, 3- s-MAIA parameters vs h are shown in green. Stepsize values for h are normalized with respect to the number of stages to fit in the VV stability interval. . . 56
2.12 Values of the $\rho_r(h, \Gamma)$ function (in $\log_{10}$ scale) using patched values of b for the 2-stage integrators (left) and 3-stage integrators (right). 2-, 3- s-MAIA parameters vs h are shown in green. Stepsize values for h are normalized with respect to the number of stages to fit in the VV stability interval. 57
2.13 Performance of s-MAIA2 (green) against other 2-stage modified integrators combined with MMHMC and tested on the 1000-dimensional multivariate Gaussian model. 58
2.14 Performance of s-MAIA3 (green) against other 3-stage modified integrators combined with MMHMC and tested on the 1000-dimensional multivariate Gaussian model. 58
2.15 Performance of s-MAIA2 (green) against other 2-stage modified integrators when combined with MMHMC and tested on the BLR German model. 59

2.16 Performance of s-MAIA3 (green) against other 3-stage modified integrators when combined with MMHMC and tested on the BLR German model. . . . 60
2.17 Performance comparison of HMC and MMHMC combined with 2-, 3-stage adaptive integrators s-AIA (in yellow and blue) and s-MAIA (in green and black), respectively, using the Multivariate Gaussian model . . . 61
2.18 Performance of s-MAIA2 (green) against other 2-stage modified integrators when combined with MMHMC and tested on the BLR WBC model. . . . 62
2.19 Performance of s-MAIA3 (green) against other 3-stage modified integrators when combined with MMHMC and tested on the BLR WBC model. . . . 62
2.20 Posterior distributions for three priors studied for the SOX2 clinical dataset. . . . 63
2.21 Posterior distributions (zoomed around 0) for the SOX2 clinical BLR model using the $\mathcal{N}(0,1)$ prior. The colour indicates the percentage of the curve under it. . . . 65

3.1 General pipeline for the mathematical analysis of *-omics* data. . . . 68
3.2 Illustration of the mapping process. The input consists of a set of reads and a reference genome. In the middle, it gives the results of mapping: the locations of the reads on the reference genome. The objective of the alignment is to provide appropriate matches, even accounting for possible mismatches. Image from Galaxy Training Network (Wolff et al., 2024) used under Creative Commons Copyright license. . . . 70
3.3 Histogram of count data in an RNA-seq experiment, with the (scaled) Negative-Binomial distribution superimposed in red. . . . 71
3.4 Map depicting all available cell lines sequenced for the bG-RNA-seq dataset and the relations between them. . . . 72
3.5 PCA analysis of the MCF7 cell lines from the bG-RNA-seq data, with the “sh_TAMR” and “TamR” conditions grouped with the pink label “TAMR”. The shSOX2 condition (in green) is scattered across the first principal component more visibly than other conditions 74
3.6 PCA of all cell lines from the bG-RNA-seq dataset. MDA-MB-231 (labelled as "MDA231") and T47D form their own clusters. Those without a cell line name before the condition are MCF7. These form 2 clusters, one for the original sequencing batch and another for the sh_SOX2 and PLKO conditions, which were sequenced again. . . . 74
3.7 Difference in library size between conditions. . . . 75
3.8 Fitness terrain obtained for all admissible combinations of lower and upper trimming percentages in the bG-RNA-seq data set. The asterisk indicates the combination corresponding to the highest AUCVC value. . . . 78
3.9 Number of genes left for normalization after removing genes below a threshold value across samples. . . . 79
3.10 Comparison of CoV curves for all normalization methods considered. . . . 80
3.11 Volcano plot showing statistically significant differentially expressed genes. . . . 81
3.12 Example of Enrichr view for the MCF7 CRTL vs TamR analysis. . . . 84
3.13 Alterations to the KEGG estrogen signaling pathway using the MCF7 CTRL vs TamR analysis. Blue is chosen for an average negative expression — meaning expression is down in the TamR condition —, while red represents genes upregulated in that node in the resistant condition. . . . 85
3.14 Home page. . . . 86
3.15 Table with DEA results for the MCF7 CTRL vs MCF7 TamR study. . . . 87
3.16 Volcano plot highlighting several genes significantly differentially expressed in the MCF7 CTRL vs MCF7 TamR study. . . . 88
3.17 Boxplot showing the expression changes between MCF7 CTRL and MCF7 TamR conditions of HERC1, a gene found to be relevant for tamoxifen resistance in our mathematical models (see Chapter 4). . . . 88
3.18 GSEA results using the KEGG platform. . . . 88

3.19 GSEA table. 89

3.20 Download page. 89

4.1 Schematic representation of the proposed work pipeline. The data collection and data treatment are followed by the joint analysis of cell and patients data, and the refinement of the results. The validation process relies on statistical analysis and experiment. 94

4.2 Heatmap of all the ER+ & tamoxifen-treated patients before cleaning. The patient's last follow-up status was used to separate groups in the DEA analysis, although no clear structure in the data can be detected. 96

4.3 Timeline visualization of TCGA patients treatment data. Each row in the y-axis shows the major events in the treatment of a patient. Patients with a star are responded well to the treatment and patients with a cross showed resistance to treatment. 98

4.4 Heatmap of the differentially expressed genes in TCGA-resistant patients (black dotted box) and well-responded patients (green dotted box) from the TCGA dataset. The right-most black box highlights the only 2 patients labelled as resistant by our classification scheme that do not cluster with the rest. 98

4.5 Gene cloud showing the distribution of genes in the joint analysis of the cell and patients data as well as the effect of the applied filters on the final shortlist. **A** No filter. **B** Filter for expression levels of $|\log_2 \text{FC}| > 0.5$ set for both analyses. **C** Only genes expressed in the same direction are kept. **D** Remaining 17 genes after setting a significance threshold of FDR < 0.1 for both DEA. 100

4.6 Schematic representation of the gene signature selection algorithm, BLR-SA. The BLR model gives a posterior distribution from which it is possible to make predictions on resistance. This value is compared to previous best and recorded before proposing a new signature for a new iteration. 103

4.7 Correlation between *ELPD* and Overall Prediction Accuracy for all the gene signatures tested with the BLR-SA algorithm. The trend in the mean of the boxplots suggests that more accurate models also show lower negative *ELPD*. Final selection is marked with a star. 104

4.8 Precision–Recall curve for the 6-gene signature obtained using the BLR-SA algorithm (blue) and compared with the Precision–Recall curve produced by using all 17 genes. . . 104

4.9 Relation between mean expression levels of the EnCaRes signature and recurrence free survival evaluated by Kaplan-Meier analysis using the KMplotter tool. On the left, tamoxifen treated patients. In the middle, patients treated with any type of hormone therapy and on the right, patients with ER- tumours non-susceptible to hormone therapy treatment. In the treated patients, the EnCaRes signature separates high and low risk groups using an upper tercile cut-off. No significant effect is seen in the ER- patients. . . 106

4.10 Relation between mean expression levels of the EnCaRes signature and recurrence free survival in the METABRIC cohort for patients treated with any type of hormone therapy (left) and for patients with ER- tumours non-susceptible to hormone therapy treatment (right). An upper tertile threshold for high/low expression based on the mean expression levels of the EnCaRes signature was used. In the treated patients, the EnCaRes signature separates high and low risk groups, but no significant effect is seen in the ER- patients. . 107

4.11 Univariate (top) and multivariate (bottom) Cox proportional-hazards regression models for testing the predicting ability of the EnCaRes signature against the known resistance signatures on the METABRIC cohort. The selected signatures abbreviations stand for (from top to bottom): the 18-gene SET ER/PR signature (Sinn et al., 2019), EnCaRes signature (this study), ODX — Oncotype DX (Paik et al., 2006), the Men 10GS — 10-gene signature by Men et al. (Men et al., 2018), the HOXB13/IL17BR ratio (Ma et al., 2004), CRISPR — the signature, obtained with CRISPR-Cas9 edited ESR1 (Harrod et al., 2022) and 5-Cand Path. — the 5-Candidate Pathway (Rahem et al., 2020). HR is the Hazard Ratio and CI is the confidence interval. A smaller p-value indicates a more statistically significant result 108
4.12 Multivariate Cox proportional-hazards regression model compares the predicting ability of our EnCaRes signature with the known clinical covariates on the METABRIC cohort. HR is the Hazard Ratio and CI is the confidence interval. 109
4.13 **Experimental validation by qPCR**. Relative transcript levels from qPCR analysis of the EnCaRes signature in MCF7 CTRL (black bars) and MCF7 TamR (red bars) cells show a significant increase in all but one of the genes in the signature. Error bars represent standard deviation (SD) for n=5 experiments. Asterisks indicate statistical significance from a one-sided t-test. (*) p-value < 0.05, (**) p-value< 0.01. 110
4.14 Gene Set Enrichment Analysis of the 17 common genes using the Hallmark collection from the Molecular Signature Database (left), and the KEGG database (right). 112
4.15 Expression levels of the genes in EnCaRes after 6, 18, 24, 48 and 72 hours of tamoxifen treatment compared to the expression in MCF7 cells. 112

B.1 Individual Kaplan-Meier survival curves for each genes in the EnCaRes signatures. The lack of an effect in any of them individually emphasizes their impact as a signature. . . . 128

List of Tables

2.1 Extract from the complete SOX2 Clinical dataset. In green, patients that responded to therapy; in red, those who developed a recurrence. . . . 18
2.2 Description of the features available in the SOX2 Clinical dataset. . . . 19
2.3 Brief fragment of the complete TCGA Clinical dataset. Death, Start and End of treatment are measured in days. . . . 23
2.4 Subset of the splitting integrator families for modified HMC used in this work and derived in Radivojević et al. (2018). . . . 37
2.5 Metrics used in the numerical tests for diagnostics and estimation of performance. . . . 41
2.6 Relevant for ESS estimation properties of the MCMC samplers presented in this study. . . . 43
2.7 Methods used in Figure 2.6. . . . 44
2.8 Methods used in Figure 2.7. . . . 45
2.9 Methods used in Figure 2.8. . . . 46

4.1 Genes forming the EnCaRes signature. . . . 104

A.1 Original dataset obtained from Hospital Universitario Galdakao-Usansolo for the study of the biomarker SOX2 on tamoxifen treated patients. In green, patients that responded to therapy; in red, those who developed a recurrence. . . . 122
A.2 Complete SOX2 clinical dataset after filling in the missing data with the MICE methodology. In ***emphatic bold***, inputted data points. In green, patients that responded to therapy; in red, those who developed a recurrence. . . . 124

B.1 Clinical characteristics of patients from TCGA (tamoxifen and all hormone therapy, ALL HT) datasets. In parentheses, distributions of resistant (R) and good responder (GR) patients in each category are presented. . . . 129

List of Algorithms

1 Pseudo-code for the Random Walk Metropolis-Hastings (RW-MH) algorithm. 26

2 Pseudo-code for the Hamiltonian Monte Carlo algorithm. Highlighted in ***emphatic bold*** are the new steps in the method compared to the previous algorithm (Metropolis-Hastings). 29

3 Pseudo-code for the Generalized Hamiltonian Monte Carlo algorithm. In ***emphatic bold*** we highlight the new steps in the method compared to the previous algorithm (HMC). 31

4 Pseudo-code for the Mix & Match Hamiltonian Monte Carlo algorithm. In ***emphatic bold***, new steps in the method compared to previous algorithm (GHMC) are highlighted. 35

5 Adaptive selection of optimal integration parameters for MMHMC. 54

*

Chapter 1

Introduction

In his book *The Structure of Scientific Revolutions* (Kuhn, 1962), philosopher of science Thomas Kuhn postulated that scientific progress historically develops in a cyclic system. Kuhn defended that scientific research eventually leads to a point where current theories can not give answers to the problems observation presents. This indicates the need for a different theory that can offer a new view on the present problems and push the field forward, allowing further research to continue. In his work, he claims that, whenever science leads to previously unknown and hard to explain matters, a *paradigm shift* occurs. These new paradigms aim to fill the gaps of knowledge present in previous theories and provide a baseline for science to advance and establish until it again fails to account for observations. Kuhn offered a few examples, mostly based in physics, such as the heliocentric view proposed by Copernicus, which broke past conventions about celestial mechanics and paved the way to Newton's theory of gravity. Science then evolved until new problems arose and Einstein introduced his theory of relativity and, with it, a new paradigm for gravitation.

In the context of biology, a relatively new science compared to astronomy or physics, these types of revolutions are still uncommon. Until recently, biology has been mostly a descriptive discipline with roots in medicine and observations of the living world. In this sense, it is possible to connect Darwin's theory of evolution with Mendel's law of inheritance and, ultimately, with Watson's and Crick's unveiling of the structure of DNA. All of them constituted the beginning and end of the search for the mechanism that carries information from one generation to the next. With the formulation of the central dogma of molecular biology (Crick, 1970), which outlines the flow of genetic information from DNA to RNA to proteins, a new era of research began for biology. The collective effort to untangle the secrets of DNA gave birth to several new branches of science devoted to its study.

At the forefront of these efforts was the Human Genome Project, an ambitious international endeavour aimed at deciphering the entire sequence of the human genome. The mapping of the key piece in the complex mechanism of life was an incredible feat. However, in the process of answering some of the most fundamental questions about human biology, it opened many more questions oriented towards the discovery of the mechanisms that operated such a complex biological machine. It represented a new *shift* in the paradigm. Not because it broke completely with the old theories — it didn't — but because of the new frontiers it defined for the study of human biology (Horwich, 1993). Simply put, it led to new horizons to explore and a completely new landscape where every piece of information was practically unknown. Instead of breaking with the previous theories it branched out and created novel ways of thinking into old unsolved problems, in a similar way as Einstein's relativity took a more limited understanding of gravity and amplified it to a completely new dimension. With the basic instructions of life identified, it was necessary to obtain deeper understanding of how genes function, interact and are regulated, as well as untangling their roles in health and disease (Keller, 2005).

In parallel, another revolution was occurring predominantly driven by the development of increasingly intricate and swifter computation. The digital revolution distinguished itself by its reliance on the tremendous technological advancements occurring in the second half of the 20th century. This revolution was not fuelled by a change in any fundamental theory, but it did create unprecedented abilities to explore previously uncharted territory and caused a tremendous shift in the way problems could be approach all together. Its origins can be traced to World War II where the computers, still in their infancy and taking up entire warehouses worth of space, were vital in cracking cryptographic codes and in aiding scientists in the Manhattan Project (also birthplace of the Monte Carlo methods) in the resolution of complex nuclear calculations. Both problems illustrated the potential of computers in a key area related to problems that were mathematically solvable, but involving inhumanly long calculations. The next half of the century was marked by incredible improvements both in hardware, with increasingly more powerful machines, and software, the development of more refined algorithms and a better understanding of the mathematics governing computations. These advancements caused a significant leap in processing power and computation capabilities compared to earlier computing devices, and consequently revolutionized the way information was collected, stored and analysed.

The culmination of the revolution for informatics came with the advent of internet connectivity and the appearance of personal computers. Computers had become an indispensable part of the scientific process, and an incredible aid in solving a huge variety of problems. The widespread access to personal devices with computing power surpassing that of their predecessors by several orders of magnitude marked a turning point and opened the access to these resources to nearly everyone. All while information can be shared instantly with anyone in the world, fundamentally transforming the world into a data-centric culture where a vast amount of data is gathered daily and used for optimizing nearly every aspect of life. As a consequence of this unprecedented availability of data, processing and analysis of large datasets have become important topics for research lately in computer science and mathematics.

The impact of these technological advancements in biomedical applications was two-fold. On the one hand, it made feasible the use of large sequencing techniques, capable of reading millions of base pairs at once (Kelly, 1989). On the other hand, all this data, alongside other clinical and biological features, could now be shared, analysed and compared all over the world thanks to the advent of internet connectivity. Effectively, this created a much wider sample size to study all kinds of diseases. It also opened the door for the study of clinical data to a wider range of scientists besides those in direct contact with the patient. This fact made medicine and biology a much more interdisciplinary field and opened the door for mathematical innovation at a greater scale.

Hence, it was at the junction between these revolutions where bioinformatics was born. From its origins it had an inherent interdisciplinary nature which brought together biology and computer science as building blocks of this new paradigm. The emergence of this new field not only enhanced a deeper understanding of biological processes but also facilitated the constant development of innovative methodologies and tools that were instrumental in advancing research in all these domains. As the field continues to grow, its interdisciplinary approach proves to be key for discovering the complex mechanisms of life and addressing critical challenges in biology and medicine.

Figure 1.1 presents a schematic visualization of the major breakthroughs in biology and computer science and provides context for this thesis.

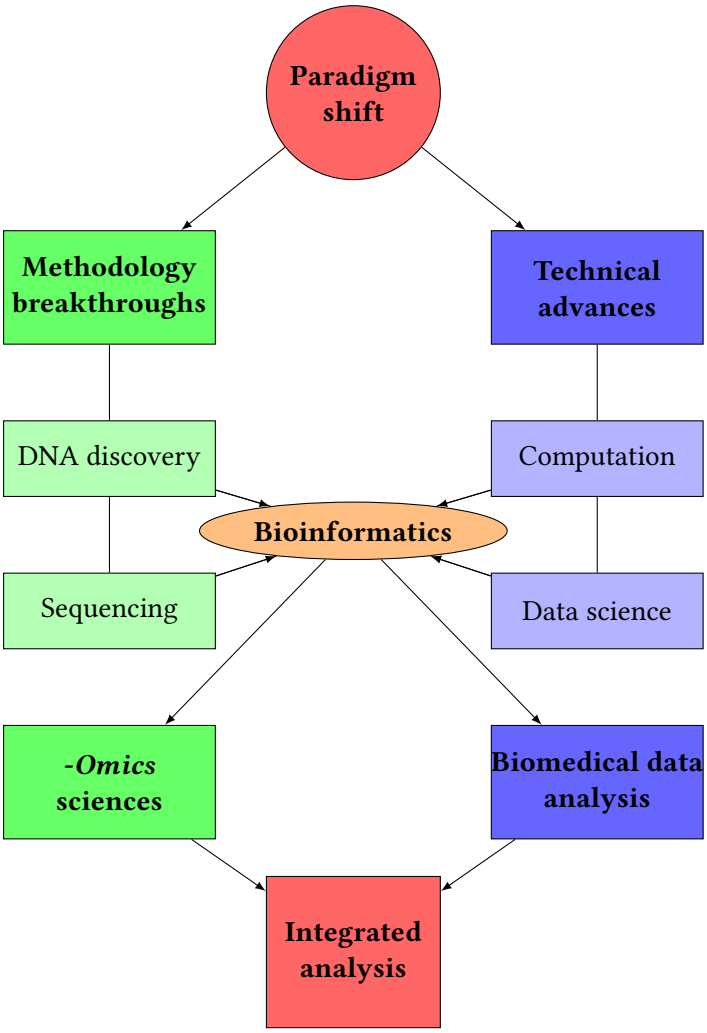

Figure 1.1: Schematic representation of the major breakthroughs in biology and computer science.

1.1 Cancer research and data analysis

1.1.1 Clinical data

In medicine, information is one of the most valuable assets of a clinician. It is thanks to careful testing, explorations and analysis that a doctor can diagnose a disease and recommend a treatment. Until very recently, a direct inspection of the patient has been the only way to study a disease. As a consequence, medicine has been a rather imprecise science for most of its history, being full of clinical treatments that, in the light of our current understanding of human biology, were less than adequate, wrong or, in several instances, harmful for the patients. This is completely understandable as the available information was rather limited. For example, the root cause of many infectious diseases was not understood until the discovery of viruses and bacteria in the late 19^{th} century. Despite this, the doctors of the medieval times realized the contagious nature of several diseases based on observations alone. This led to preventive measures, such as the isolation of patients with contagious disease, and gave birth to the term *quarantine*. Although a very primitive example, this illustrates the power of observations and inference in medicine to save lives.

An even more practical example of the unequivocal value of information in clinical practice is the development of our understanding of cancer. Despite being known since ancient times, cancer has always posed a difficult challenge for physicians. Before the advent of modern pharmacology, the only way a cancer could be treated was by surgically removing the tumours. However, surgery alone

is generally not an effective treatment against the majority of cancers and it was already known in ancient Roman times that tumours could come back after surgery (Ford and Finlayson, 2010). Although some progress was made in surgical removal of cancerous tumours, it was not until the end of the 19th century that clinicians began to uncover some of the key aspects of cancer that would propel the development of new therapies. This was possible due to the work by pioneers in oncology such as Sir George Beatson, who noticed that advanced breast cancer patients improved when the ovaries were removed. This planted the seeds for further research into hormone-driven tumours (Beatson, 1896). These new extracts of information, suggesting a link between cancer and other biological process, were cemented soon thanks to the advancements in microscopy technology and the better understanding of the role of cells in life. From that point onwards, molecular biology also became a key source of information in the clinical practice.

These new insights marked a turning point in the understanding of cancer as a complex disease with a high level of interplay with other biological process and external factors. Discoveries such as Beatson's prompted larger studies of the role of hormones in cancer and their link with other pathological data. Besides, during the first half of the 20th, an increasing number of clinicians started drawing associations between lung cancer incidence and certain environmental hazards, such as cigarette smoking or the inhalation of chemicals often found in factories and other industrial-heavy environments (such as arsenic, asbestos or radon (Ruegg, 2015)).

Besides *pathological* and *external* influences, a third major player causing the appearance of cancer is its *hereditary* component. One of the first reports on the matter was by the French surgeon Paul Broca. He investigated his wife's family ascendancy to discover that, among 24 woman across 5 generations, 10 died of cancer (Broca, 1866). There has always been a certain grasp about the hereditary nature of many diseases through history, based on common re-occurrences within family clusters. However, it was not until the advent of sequencing techniques that a conclusive answer on the cause of this observation was discovered. In 1994, BRCA1 was identified as the first gene associated with a cancer (Miki et al., 1994). A mutation in this gene, which can be passed from one generation to the next, was identified as a risk factor for breast cancer. Since then, sequencing has played a major part in cancer research, diagnostic and treatment.

Hence, having a complete picture of a patient's history can facilitate inferences between patient information and clinical outcome. In clinical practice, pathological information is usually recorded in a more or less standardized manner alongside other demographic factors. Among the two other information sources, we find that genetic information may provide better insight into the problem under study in comparison with the information on environmental factors. This is because environmental risk factors (such as smoking condition or alcohol consumption) are already known and accounted for (Katzke et al., 2015). Therefore, unless the main focus of the study is the research of a proposed external risk factor and a given condition, new relevant patient specific features are needed to establish connections between them and the disease. On the contrary, sequencing data offers thousands of new elements to explore in connection to the clinical feature of study.

Thus, in the current landscape, it is much more common to find genetic information of some kind accompanying demographic and pathological data. Similarly to the examples for external factors, the inclusion of already known disease-related genes in these studies is relevant for day-to-day clinical practice (Qin, 2019). But this fails to add value when the objective is the identification of novel features related to the disease. Nowadays, it is much easier and more common to attach complete genetic profiles to a clinical record thanks to the ever-increasing accessibility of sequencing techniques. Compared to clinical records, sequencing data offers the advantage of being standardized to some extent, which facilitates its bulk analysis and access. However, technical differences arising from different techniques and equipment for sequencing can pose a challenge for the analysis. In the next section, we will discuss more in-depth the nature and analysis of sequencing data.

Clinical studies cover an ample spectrum of data variables, and the analysis of these datasets

requires flexible methodologies that can be adapted to different problems and ranges of data. One of the most crucial aspects of a clinical study is the scientific questions that it intends to extract from a specific dataset. This uniquely defines the type and amount of data needed, as well as the analysis and validation of the results. Hence, we can identify four major steps in the span of a clinical study:

Step 1: Collection of clinical data and design of clinical trials.

Step 2: Data treatment and curation.

Step 3: Mathematical modelling and analysis.

Step 4: Testing and result validation.

In some cases it is difficult to have direct control over all aspects of a study, specially from the perspective of a data analyst whose work usually begins with an already defined dataset. Regardless, it is important that a robust mathematical approach is used rigorously from the start to end. In this sense, the main approach to statistics in medicine has mostly been a frequentist one. However, Jack Lee and Chu (Jack Lee and Chu, 2012) presented the case for a Bayesian approach to the analysis of clinical data in this manner:

> "Considering the treatment effect, θ, which is the parameter of interest, the frequentist framework assumes that θ is fixed yet unknown. Through clinical trials, we can collect data to inform θ. Hence, the inference on the treatment effect can be made by evaluating the probability: *P(Data | θ)*, where the data are considered to be random and the parameter θ is fixed. Conversely, the Bayesian framework assumes that the data is fixed and the unknown parameter θ is modelled as a random variable of a probability distribution. Bayesian inference is made by computing *P(θ | Data)* [...], in other words, all relevant information for making inference on θ is contained in the observed data and not in other unobserved quantities"

Hence, the Bayesian approach makes a relevant assumption that is critical and gains more value in the medical context. In the Bayesian approach, the data is fixed, and the unknown parameters of the model are the ones that need to be modelled as random variables. Besides, as new data becomes available, we can update our beliefs about a set of parameters based on the observed data using Bayes' theorem[1] (Bayes and Price, 1763)

$$\overbrace{P(\boldsymbol{\theta}|\mathbf{D})}^{\text{Posterior}} = \frac{\overbrace{P(\boldsymbol{\theta})}^{\text{Prior}}\overbrace{P(\mathbf{D}|\boldsymbol{\theta})}^{\text{Likelihood}}}{\underbrace{P(\mathbf{D})}_{\text{Marginal Likelihood}}}. \tag{1.1}$$

The *posterior distribution*, which encapsulates our updated knowledge, is proportional to the product of the *likelihood function* and the *prior distribution*. The likelihood is a function that represents the statistical model used to describe the data, D, while the prior is used to pass onto the model our beliefs about the parameters $\boldsymbol{\theta}$ and their uncertainty. Ultimately, the posterior distribution offers not only a description of the values of the model parameters, but also inherently accounts for the uncertainty associated with them. The *marginal likelihood* $P(\mathbf{D})$ is used to ensure that the posterior distribution is a proper probability distribution, i.e. its definite integral over the entire space is equal to 1.

[1] Following Jack Lee and Chu's notation

In essence, Bayesian statistics offers an appealing approach for modelling, both from a theoretical point of view (data informs the model) and a practical one (built-in uncertainty quantification). Besides these advantages, the ever-expanding access to more powerful computational methods made Bayesian models an appealing option. As such, they can be present in all four major steps for clinical studies as summarised above.

In the design of clinical trials (Step 1), Bayesian methods offer the advantage of their adaptivity to new incoming data, and thus are frequently used in the estimation of an optimal sample size. Estimating the number of patients in a clinical study is vital to obtain appropriate results. A low number may not be enough to detect the desired effect or response, while a high number may constitute a problem too if, for example, a new drug is being tested and the side effects can be potentially harmful. To tackle this, some properties of the Bayesian methods can be used. By imposing informative priors, based on previous clinical results, the size effect could be accounted for *a priori*, reducing the range of patients needed for achieving a certain probability of success (Kunzmann et al., 2021). It has been noted that the use of carefully built priors based on expert knowledge in Bayesian statistics creates a more defined view of the study and its expectations from the beginning (Sutton and Abrams, 2001). But perhaps, more importantly, Bayesian methods can be re-evaluated with gradual updates of information, which is precisely the type of methodology that fits best with the non-predictable influx of patients in most diseases (Berry, 2006).

Data treatment and curation (Step 2) are crucial parts in the actual landscape of data analysis. As mentioned before, it is common that with data coming from various sources such as different doctors or hospital, the final dataset has missing instances. For this, one of the most popular algorithms is the Multiple Imputation by Chained Equations (MICE) (Van Buuren, 2007), which creates an imputation model for a feature with missing values conditional on all other features. A value for missing data can be obtained by generating multiple possible datasets from this imputation model, sampling from the resulting posterior and combining the results. This methodology improves significantly over other methods popular in data science, such as those using random data or the median or mean imputation. Bayesian approaches to data imputation have also been proposed (Jang et al., 2021), but so far they are more limited and, for other than exceptional cases, the gains are not considerable when computational time is taken into account.

In terms of mathematical modelling (Step 3), recent years have seen a surge in the application of novel Bayesian methodologies to different stages of cancer research. The aforementioned availability of public repositories has opened the possibility to researchers to test their models in real-world scenarios without the need to obtain a new cohort of patients. This means that plenty of machine learning models can be developed, implemented and tested with clinical data such as prediction of diseases (Rabiei et al., 2022), diseases evolution and staging (Zaballa et al., 2023) or response to treatment (Mani et al., 2013). Within these methodologies, again Bayesian methods should be acknowledged as they have shown to offer incredible accuracy in classification of tumours (Karabatak, 2015) or patient survival (Teng et al., 2022) using only clinical information. A side note should be made here to mention the impact of mathematical modelling in the last major health crisis. Among global efforts made to combat the COVID-19 pandemic, several contributions were aimed at creating epidemiological models to predict the spread of the disease (Hauser et al., 2020) or to obtain efficient use of hospital resources (Inouzhe et al., 2023; Self et al., 2022) when these were dwindling under the heavy needs created by the pandemic. Plenty of more examples of uses of mathematical modelling to provide support during the pandemic can be drawn, for which we refer the reader to Cao and Liu (2022) and references therein.

Finally, the increase in the use of purely computational (or *in-silico*) experiments in biological contexts, has increased the need for multiple testing and independent validation (Step 4). Given the scarcity of data and the high susceptibility of biological conditions even to small changes, it is crucial to keep mathematical models general enough and test them against overfitting to a specific dataset.

Common techniques, such as cross-validation and regularization, can be used to minimize the risk of excessively adapting a model to the data (Chicco, 2017). However, even after validation, assessing the significance of a result can be tricky. Here, the Bayesian approach offers its own hypothesis-testing mechanism in the Bayes factor, which could be comparable to the frequentist p-value (Held and Ott, 2018). Moreover, it also provides metrics specific for cross-validation scenarios, such as the expected log-pointwise density (ELPD) (Vehtari et al., 2015) or the direction of the effect of a variable in an outcome, such as the p-direction (Makowski et al., 2019). Both of these are used later in this thesis to assess the quality of the results obtained within the Bayesian framework.

In summary, the study of clinical data could be seen as a problem with a lot of room for improvement, not only in terms of modelling, but also in the way these models are solved. As we have seen, the use of Bayesian models is on the rise, but it brings new challenges. Bayesian statistics is usually more computationally intensive than its frequentist counterpart, so progress needs to be made towards developing enhanced methodologies that optimize performance and reduce computational burden. As medicine progresses towards a world where personalized medicine and computer-aided diagnostics will play a major role, it is becoming increasingly important that new methods appear to ensure an optimal performance of the mathematical models at play. Chapter 2 explores in more depth Bayesian analysis and methods to enhance sampling performance in Bayesian scenarios, which are presented as practical applications towards the end of the chapter.

1.1.2 High-throughput data

The completion of the Human Genome Project marked a key moment in our understanding of the human genome and significantly boosted our confidence in untangling the secrets behind the mechanism of life. However, this brilliant feat primarily relied on a technique called Sanger sequencing, a method to identify the sequence of bases (nucleotides) in a DNA molecule that Frederick Sanger and his colleagues initially presented in 1977 (Sanger et al., 1977). The study of DNA was revolutionised by this technique where small fragments of DNA were replicated in the presence of modified bases called dideoxynucleotides which lacked the hydroxyl group necessary for the formation of a bond with the next nucleotide, leading to the termination of DNA synthesis. This resulted in small DNA fragments that could be separated physically (by size and electric charge using electrophoresis) and analysed individually. Although this method was revolutionary at the time and allowed for the sequencing of separate DNA fragments, it was a laborious and time-consuming process, which restricted its ability to analyse genomes on a large scale.

Because of these restrictions, Sanger sequencing is considered as a low-throughput sequencing technique, meaning that only a relatively small sequence of nucleotides could be read at a time. However, by the turn of the 21st century, a new collection of sequencing technologies emerged, usually referred to as Next Generation Sequencing (NGS). The leap from the first generation sequencing techniques to the next one was enormous, not only in the reading capability, but also in the parallelization that the new generation brought with it. Although this new generation had multiple procedures for DNA base reading, they were no longer based on terminating DNA replication to produce the reads. Instead, DNA was fragmented for sequencing but then aligned to a reference genome, allowing for the reconstruction of the complete sequence. As this could be done in parallel, these machines allowed for the simultaneous sequencing of thousands of DNA fragments (Schuster, 2008). This dramatic increase in sequencing throughput also carried with it a reduction in cost per base pair and boosted the emergence of bioinformatics and computational biology as essential disciplines for the analysis and interpretation of complex biological datasets.

These techniques could be applied for reading of not only DNA, but nearly all the biomolecule chains present in an ordinary cell. As a consequence, genomics, which stems from the analysis of the genome (DNA), branched out into several disciplines according to the biomolecule of study, the so-called *-omics* sciences (Manzoni et al., 2018). In particular, transcriptomics focuses on the study of

the transcriptome (RNA), which is responsible for transcribing the genetic information encoded in DNA to the cellular processes that will lead to the synthesis of proteins. Proteomics focuses precisely on the study of proteins, while the study of lower components created by a metabolism is referred to as metabolomics.

In this thesis, we will focus exclusively in transcriptomic studies, mostly using RNA-sequencing (RNA-seq) data (Wang et al., 2009). RNA-seq experiments quantify the amount of RNA in a sample and are used to gain insights into gene expression patterns, RNA processes and regulatory mechanisms within cells and tissues. By examining the transcriptome, or the complete set of RNA transcripts present in a biological sample, transcriptomics provides valuable information about the dynamic nature of gene expression and the functional roles of different RNA molecules in diverse biological processes.

The transition from the technical output of the NGS machines to the mathematical description of the genes involved is straightforward. In essence, the machine encodes the fragmented reads of DNA as sequences of nucleotides in text format (A, C, T and G for adenine, cytosine, thymine and guanine, respectively). These sequences are then compared to a reference genome to identify regions matching these reads. If a unique and positive match is found, a read count is added to the gene that is coded in the corresponding part of the genome. Hence, the array of nucleotides initially outputted by the sequencer can be converted into a matrix of counts. The resulting gene count matrix for an RNA-seq experiments usually has over 24.000 entries/rows, one for each known gene, showing the number of matches obtained for every gene or transcript.

One key aspect of gene counts is that there is no reference for the absolute value of a given set of counts to compare with. Essentially, this means that it is difficult to interpret, without further information, if a gene count in the thousands is a high or a low one. In practice, this limits the use of RNA-seq experiments mostly to situations where it is possible to compare two separate biological conditions. Besides, due to the inherit biological variability of most samples, the variance in the data tends to be far larger than the mean, which makes comparing RNA-seq experiments a challenging process.

Hence, as discrete count data with overdispersion, the gene counts K_{gm} for a gene g in sample m are usually modelled using a Negative-Binomial distribution[2], $K_{g,m} \sim NB(\mu_{g,m}, \alpha_g)$. Here, the mean $\mu_{g,m}$ for a given gene g and sample m, is obtained from observed data; while the dispersion parameter α_g is usually estimated gene-wise (Anders and Huber, 2010). The exact estimation of these parameters is addressed in Chapter 3. For such a distribution of counts K, the probability mass function can be modelled as Poisson distribution with a rate parameter given by Gamma distribution as:

$$P(x = K)_{NB} = \frac{\Gamma(\alpha_g + K)}{K!\Gamma(\alpha_g)} \left(\frac{\alpha_g}{\alpha_g + \mu_{g,m}}\right)^{\alpha_g} \left(\frac{\mu_{g,m}}{\alpha + \mu_{g,m}}\right)^{K}. \tag{1.2}$$

This approach is not only very accurate to model the count data from sequencing experiments, but also quite flexible and able to capture a wide range of expression levels, including low count transcripts, which are the most abundant. However, technical factors, such as the preparation of the samples, the number of total transcripts reads (called library size) or the characteristics of the equipment, will influence heavily the observed values for the counts (Conesa et al., 2016). To limit the influence of these mostly uncontrollable factors, various technical replicates are usually made, where the measurements are repeated to account for this variability. Similarly, it is standard practice to also use biological replicates (different samples of the same tissue or cell) to account for the inherent biological variability.

[2] Also known as Gamma-Poisson, as it is a generalization of the Poisson distribution that allows for overdispersion in the data by having the rate parameter follow a Gamma distribution.

As a consequence, raw data counts should be normalized to account for these sources of variability. The resulting normalized counts across replicates could then be used to extract information about dissimilarities in gene expression levels between two different biological conditions. At this point it is important to note that, despite "gene" being a term originally used for genomics, it is often exploited as a general term to refer to the resulting counts, regardless of the type of *-omics* methodology in use. This is carried forward in the pipeline, and the most usual analysis for RNA-seq data is referred to as *differential gene expression (DGE) analysis*.

Many methodologies exist for analysing RNA-seq data from the raw data counts to the DGE results. The most popular of them include DESeq2 (Love et al., 2014), edgeR (Robinson et al., 2010) and baySeq, an empirical Bayesian approach to DGE (Hardcastle and Kelly, 2010). A detailed description of each of these methods and following steps of the process in the pipeline of RNA-seq data analysis will be provided in chapter 3.

Beyond RNA-seq, high-throughput sequencing technologies continue to evolve and recent efforts have been directed towards improving sequencing accuracy, adding spatial information and reducing costs. In parallel, mathematical methods for their analysis, including Bayesian methodologies, keep up with these emerging technologies. In this regard, novel high-throughput methods, such as single-cell RNA-seq (scRNA-seq), offer a jump in sequencing precision by moving from sample sequencing to the single-cell level. This allowed for a much finer study of cellular heterogeneity and has aided classification of cellular subtypes within many problems. Bayesian methods for scRNA-seq offer advantages when passing prior information (often from bulk RNA-seq) into the new models, while also providing a better scenario for multi-group comparison (Nault et al., 2022). More recently, the development of spatial transcriptomics has opened a door for the study of the spatial distribution of gene expression within a tissue. This has brought sequencing closer to microscopy and created a new dimension to elucidate which components of a sample correlate with over or under expression of a given gene. Within these methodologies, Bayesian methods are again at the forefront of its analyses. As an example, BayesSpace (Zhao et al., 2021) uses neighbour information within a grid to enhance these methodologies and increase significantly spatial resolution compared to frequentist approaches.

Finally, RNA-seq data has proven to be a versatile tool with many uses besides quantifying DGE. Pathway analysis (García-Campos et al., 2015) employs RNA-seq data to discover enhanced or silenced pathways within a biological system. This creates a direct association between changes in gene expression and the underlying molecular mechanisms, which allows some insight into the functional relations between biological process. Gene Set Enrichment Analysis (GSEA) is another powerful technique that relies on RNA-seq data (Subramanian et al., 2005). GSEA evaluates predefined sets of genes and shows differences in expression of the complete set between experimental conditions. This enhances the more classical approach to DGE by helping in the identification of coordinated changes in gene expression that may arise from a shift at a higher order than the gene level. Overall, the nature of RNA-seq data enables a wide range of applications, facilitating deeper exploration of gene expression patterns and functional implications.

1.2 Problem at hand: Resistance to hormone therapy in breast cancer

1.2.1 Heterogeneity of breast cancer

A tumour is usually defined as an abnormal cluster of cells grown from unrestricted proliferation. Some tumours do not represent a major health issue (when this is the case we talk about benign tumours), however, tumours with the ability to spread and infiltrate nearby or distant tissues do pose a serious problem for those who suffer them. The collection of diseases originated by tumours of this kind, called malignant tumours, are known as cancers. The treatment and prognosis of each cancer

type greatly depends on the area affected and the characteristics of cells involved in the disease.

One of the most diagnosed cancer worldwide is breast cancer, with nearly 2.3 million new cases in the year 2020. It accounts for $\sim 25\%$ of all cancer cases in women (although it is not exclusive to woman) and for $\sim 15\%$ of female cancer-related deaths, making it the most prevalent cancer among women (Sung et al., 2021). Its incidence continues to rise, as global data from 1980 estimated ~ 640.000 cases worldwide, compared to the current value well over 2 million. However, incidence rates show that there exists a higher incidence in high-income regions than in developing regions. This contrasts with the mortality rate of the disease, which follows the opposite trend and has the highest indexes in low-income regions. This is attributed mainly to the availability of screening programs in higher-income countries and the significant advantage of an early detection, as a tumour at an early-stage and non-metastatic is associated with a better prognosis overall. In low-income regions, patients are diagnosed with later stages of the disease, often associated with a poorer prognosis. (Harbeck et al., 2019)

However, these numbers arise from global statistics and, although useful to comprehend the prevalence of breast cancer in our world, eventually fail to truly represent a disease that exhibits many forms. Breast cancer is not a unique disease but rather a spectrum of diseases that present distinct biological characteristics, treatment and prognosis. This heterogeneity appears in intra and inter-tumoural form, which can cause complications even for diagnosis and characterization of a single tumour (Koren and Bentires-Alj, 2015). Thus, heterogeneity is a crucial aspect of breast cancer and a correct identification of a tumour within the landscape of breast cancer is essential to provide an accurate assessment of the disease and, ultimately, an adequate treatment. A tumour is mainly characterized by the type of cell that experiences an abnormal growth. In the human breast, there are three major groups of cells present: luminal cells, which can be found in the lobules and whose function is to produce milk; basal cells, responsible for pushing the milk into the ductal tubes by muscular contraption and, finally, all the connective tissue holding everything in place, which is mainly comprised of fibrous and fatty tissue. Depending on where the tumour develops, breast cancers are classified in the first place into either lobular or ductal, with the latter representing over 80% of the diagnosed cases (Makki, 2015). This *histological classification* can be subsequently divided into lobular or ductal carcinoma *in situ* (LCIS or DCIS) and the invasive lobular or ductal carcinomas (ILC and IDC). The invasive type represents a much higher health risk due to its ability to infiltrate other tissues, both in the same breast and in distal organs.

Apart from the original location of the tumours, another classification of breast cancer exists that can more accurately be used to predict the behaviour of the tumours and prepare a response to it. It is based on the presence or absence of specific receptors on the cancer cell such as the estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2). An abundance of any of these indicators is usually pointed out by referring to it as positive (+), while its absence or low presence in the tumour is usually indicated by stating the receptor is negative (-). Besides these, other biomarkers can also be used for classification purposes, including mutations in key genes (such as the ESR1 gene, coding the estrogen receptor) or an abundance of other proliferation markers (such as Ki67). Hence, the most widely accepted classification of breast cancer subtypes follows this system and consists of four major molecular subtypes (Sørlie et al., 2001):

- Luminal A (ER+, PR+, HER2-) is the most common breast cancer subtype, representing over 70% of the cases. It is characterized by an abundance of expression of hormone receptors (ER and PR) and an absence of HER2. It has the most favourable prognosis due to a small proliferation rate and aggressiveness compared to other subtypes. It can be treated with hormone-targeted therapies, which will be discussed in the next section.

- Luminal B (either ER/PR+, HER2+) is a far less common luminal subtype (between 10-15% of cases) that also shows expression of ER and PR, but in lower quantities than Luminal A. This

is paired with a high expression of HER2 and usually also high Ki67 index meaning higher proliferation rates. These tumours usually have an intermediate prognosis and accompany the hormone targeted therapies with HER2-targeted therapy.

- HER2-enriched (ER-, PR-, HER2+) is the least common cancer type (around 5%) and is characterised for not having any hormone receptor expression. HER2-enriched tumors are treated with therapies like trastuzumab (Herceptin) and other HER2-targeted drugs such as Lapatinib.

- Basal (ER-, PR-, HER2-). Also known as triple-negative breast cancer (TNBC) because of their lack of any of the major receptors that define these subtypes. With an incidence ranging from 10 to 15% of cases, it has the poorest prognosis as none of the treatments available for the other subtypes can be used. Therefore, a combination of surgery and chemotherapy (often used in the rest of subtypes as well, alongside the specific therapy) is the most common treatment.

This classification can be traced with RNA-sequencing techniques and it provides a common profile for each molecular subtype. Certain biomarkers are common to a specific subtype and can be seen altered within the corresponding subtype, as seen in Figure 1.2 (Nolan et al., 2023). An alternative classification based on clinical immunohistochemistry analysis reduces the classification to three groups, mainly ER+, HER2+ and TNBC meaning that there is not a total concordance between molecular subtypes and clinical subtypes (López-Ruiz et al., 2022). Although still in use in clinical practice, this method oversimplifies the breast cancer landscape and can lead to an incomplete definition of the cells present within the tumour. Throughout this text, we will be referring to the molecular subtypes of breast cancer previously explained, unless explicitly mentioned at any point in the analysis of clinical data.

Among the molecular subtypes, luminal (ER+) subtypes are not only the most common subtype but, in the majority of cases, offer the best treatment path for a complete recovery. They are characterised by a high abundance of estrogen receptors present in their breast cancer cells, which fuel their growth. Therefore, estrogen targeted therapy is key to tackle them. Despite the relatively high degree of success of anti-estrogen therapies, the number of cases is so high that even small percentages of failed therapies still amounts to a significant number of deaths each year. Therefore, understanding the intricate relationship between estrogen and breast cancer is pivotal in tailoring effective treatment strategies and underscores the importance of an accurate breast cancer diagnosis and targeted treatment.

1.2.2 Drug resistance in breast cancer

Estrogen is a key female sex hormone and plays a complex and significant role in the development and progression of breast cancer. Estrogen promotes cell division and growth in breast tissue, hence, it can stimulate the proliferation of cancer cells by binding to estrogen receptors on these cells (Yaşar et al., 2017). Consequently, a straightforward approach to halting proliferation of breast cancer cells is stopping in some form the binding between estrogen and the correspondent receptor. The therapies developed with this intent are called hormone or endocrine therapies and they include aromatase inhibitors and estrogen receptor modulators (SERMs). The most widespread SERM treatment is tamoxifen, an estrogen antagonist which binds to the estrogen receptor without activating it, as shown in Figure 1.2. Clinically, this drug is the most extensively used in ER+ breast cancer at all stages with a standard 5-year treatment after surgery and it has shown to reduce recurrence by $\sim 50\%$ and mortality by $\sim 30\%$ (Group, 2011). On the other hand, aromatase inhibitors like anastrozole, letrozole or exemestane target the aromatase, an enzyme responsible for producing estrogen in females, hence reducing the estrogen levels in the organism. By inhibiting its production they can stop cancer cells from growing, specially in post-menopausal women, when ovarian estradiol (E2) production is

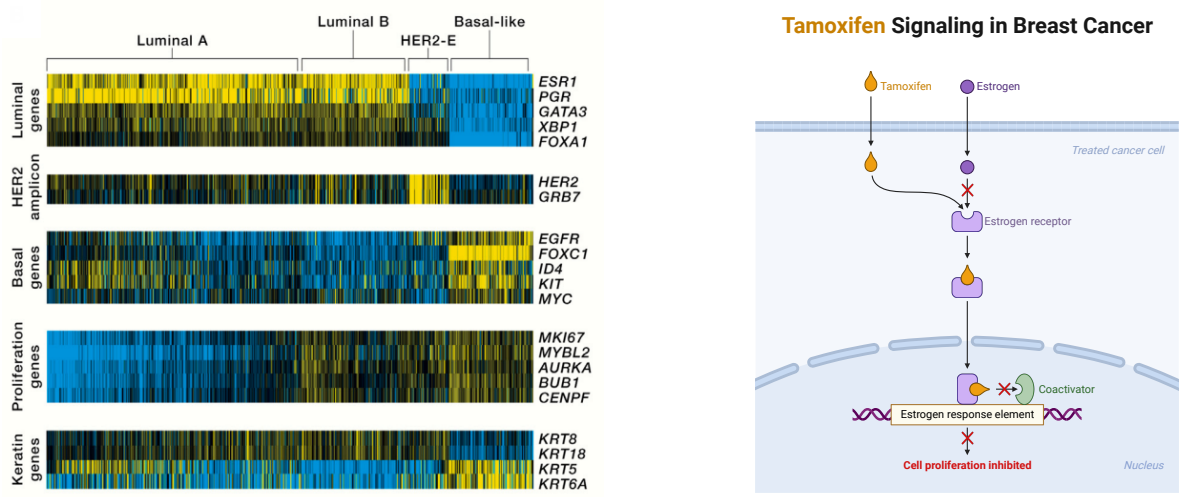

Figure 1.2: Collection of relevant genes for each of the major molecular subtypes from Nolan et al. (2023) (left) and illustration of the antagonist mechanism of tamoxifen (right). Adapted from "Tamoxifen Mechanism of Action in Breast Cancer", by BioRender.com (2023).

minimal and estrone (E1) is the major postmenopausal hormone. This means that this treatment is not recommended for pre-menopausal women, where the standard treatments are therefore SERM.

The development of drugs like tamoxifen and aromatase inhibitors, have significantly improved the prognosis for ER+ breast cancer patients (Pan et al., 2017). However, the success of these treatments is limited by an intrinsic or acquired tumour resistance, and 30-50% of patients (and almost all with metastases) treated with endocrine therapy suffer a relapse as the therapy fails (Szostakowska et al., 2019). The mechanisms behind this resistance are not completely understood yet. For example, crosstalk between estrogen and a number of pathways that modulate ER activity or mutations (Jeselsohn et al., 2015) in key genes modulating estrogen such as ESR1 (mostly in tumours treated with aromatase inhibitors), were found to be linked to tumor resistance and were proposed as possible sources for this resistance. Among them, it was found that the Wnt signalling pathway, activated by SOX2 (Piva et al., 2014), as well as pathways related to the epidermal growth factor family (Osborne et al., 2005) had impact on the development of this resistance.

What seems clear is that multiple phenotypes can be linked to the appearance of resistance. Therefore, there is an increasing interest in developing predictive models of such a resistance that could be tested before treatment. A standard tamoxifen treatment encompasses 5 years and if resistance is present or acquired during that timeframe, the patient loses precious time that could be used to try alternative treatments. In this context, mathematical modelling can offer a possible solution to the problem by providing the tools for the study and analysis of resistant patients as a whole.

1.2.3 Gene signatures for breast cancer

A gene signature can be defined as a set of genes and their expression values, that are characteristic of a particular biological condition. Genes signatures are extremely practical in the clinical environment. An analysis of the expression of the collection of genes that conforms the signature can help to identify patients with the related condition. Gene signatures are often obtained by analysing large datasets and finding which genes are consistently up or downregulated for a given medical condition. Integration of high-throughput sequencing results such as signatures into clinical practice and research settings emphasizes its critical role in the future of personalized medicine.

For breast cancer, the landmark results in terms of gene signatures and their application in

clinical practice are the 70-gene signature Mammaprint (Van't Veer, 2002) and the 21-gene signature OncoType (Paik et al., 2006), both widely successful methods where certain thresholds for expression of the signature genes can provide an accurate prediction of the risk of metastasis or recurrence respectively. It is important to remark that even these successful signatures work only for certain types of breast tumours, which should serve as an indication again of the heterogeneous landscape that exists in breast cancer.

In more recent years, prognostic signatures have been proposed for a wide range of breast cancer subtypes and clinical scenarios, and several studies have tackled specifically the problem of resistance to endocrine therapy. For instance, some authors explored both RNA and DNA mutations using patients sequencing data (Xia et al., 2022), while access to new gene editing methods have led to the proposal of a 6-gene signature extracted from the study of mutant cell lines where the ESR1 gene was modified using CRISPR-Cas9 (Harrod et al., 2022). The signatures introduced for general endocrine resistance nominated the MAPK pathway (Miller et al., 2015) or ER-PR related genes as prognostic predictors (Sinn et al., 2019). When accounting for tamoxifen resistance uniquely, lists of individual markers (Hermawan et al., 2020; Mihály et al., 2013; Wang and Wang, 2021) and ratios of gene expression (Ma et al., 2004) were identified as potential risk factors. The impact of some key pathways in the development of resistance to tamoxifen was also investigated, leading to the identification of a signature of 5 pathway-representative genes (Rahem et al., 2020).

However, some concerns have been raised recently about the significance and benefit of gene signatures in breast cancer. Enhanced proliferation of cells is one of the major drivers of cancer, and consequently, some studies (Goh and Wong, 2018; Venet et al., 2011) suggested that a signature including proliferation-related genes or genes showing correlation with proliferation markers could not really be considered significant. As many genes are in one way or another related to proliferation, avoiding confounding effects is extremely difficult. On the other hand, Venet et al. also pointed towards a lack of statistical significance (that can be enhanced by poor validation) as a major factor in poorly effective signatures. Finally, a high number of genes in the signature, with an upper bound of 25 genes suggested by Manjang et al. (2021), was found to be yet another reason for generating seemingly significant but not necessarily meaningful prognostic signatures. At the same time, all of these recommendations are to be taken carefully, as some signatures including proliferation genes and/or made of a considerable number of genes have proven to be reliable predictors. Therefore, there exists a relevant problem in gene signature development that could be solved with the applications of more refined statistical methods, while keeping the focus on the biological context.

1.3 Objectives

This dissertation has resulted from the collaboration between the Modelling and Simulation in Life & Material Sciences (MSLMS) group at the Basque Center for Applied Mathematics (BCAM), led by Prof. Elena Akhmatskaya, and the Cancer Heterogeneity Lab at CIC bioGUNE, led by Dr. María Vivanco. In the upcoming chapters, we intend to delve deeper into the interdisciplinary nature of bioinformatics, which was central to creating a bridge between both groups. On the one hand, it helped the MSLMS group in the development and testing of newly created sampling methodologies for Bayesian inference in high dimensional problems for biomedical applications. On the other hand, it was used to analyse the data generated at the Cancer Heterogeneity Lab after sequencing several types of resistant breast cancer cells and to provide solutions to the problem at hand. Finally, the expertise from both groups converged for a multi-level study of the problem of resistance to hormone therapy in breast cancer. A detailed layout of the thesis is discussed below and summarized in Figure 1.3.

This dissertation is structured in three major chapters, each featuring an introduction to the core problem being addressed, a detailed explanation of the methodologies employed, and a compilation

	Chapter 2	Chapter 3	Chapter 4
Problem	Statistical analysis of clinical data	Exploration & analysis of RNA-seq data	Integration of cell & patients data for biomarker discover
Datasets	SOX2 Clinical TCGA Clinical	bG-RNA-seq	TCGA Clinical & RNA-seq bG-RNA-seq
Strategy	Exploit advantageous properties of Bayesian methods	Propose an analysis pipeline for *ad hoc* RNA-seq data	Combine homogeneous cell models with patients heterogeneous data
Goal	Enhance performance of chosen samplers in Bayesian methods	Create a tool for quick bioinformatic analyses	Identify key genes in tamoxifen resistance
Contibution	Adaptive integration of Importance Sampling HMC	Vivanco LabSeq Tool	EnCaRes gene signature
	Objective 1		**Objective 2**

Figure 1.3: Thesis outline.

of the novel results obtained within the context of the chapter. The results of Chapters 2 and 3 have a similar goal in general, which we refer to as **Objective 1**. It centers in developing new methods and tools for the study of several *ad hoc* datasets, although their potential application ultimately goes beyond the specific case-studies shown here. Chapter 4 is built upon the foundations laid in previous chapters and has a different approach, which we refer to as **Objective 2**. Here, the focus is on achieving a biologically meaningful result using the tools we have explored and developed for **Objective 1**. The thesis is structured as follows:

Chapter 2 explores the statistical analysis of clinical data from a Bayesian viewpoint, with the emphasis on the *ad hoc* data provided by the Cancer Heterogeneity Lab and also extracted from the public repository specially for this study (we will refer to the data as SOX2 clinical data). We place the dataset into the context of the thesis and continue with its curation and pre-processing. The chapter then goes in-depth into Hamiltonian Monte Carlo (HMC) methodologies, which can be used to improve sampling in Bayesian problems. We explore several ways in which HMC can be enhanced and present a novel integration approach which improves the quality of sampling. After defining the appropriate metrics for the analysis, we show the numerical experiments backing our proposed methodology and exemplify how it can be applied to biomedical problems, with an emphasis on the breast cancer case study previously mentioned.

In chapter 3 we present the pipeline needed for the analysis of RNA-seq data from sequencing to functional analyses. We follow it to explore another *ad hoc* dataset, the cell line bG-RNA-seq data, by using, and adapting when required, the appropriate bioinformatics tools. The proposed pipeline of bioinformatics tools and the relevant analyses are then compiled into an all-in-one bioinformatics toolkit, VivancoLabSeq, for quick online access by researchers working at the Cancer Heterogeneity Lab.

Finally, in chapter 4 we tackle the most ambitious goal of the project and identify a gene signature for hormone therapy resistance in breast cancer. For that, we create a methodology that uses the knowledge acquired in the previous two chapters to develop a model that combines clinical and RNA-seq data within a Bayesian framework, with the aim of extracting relevant biomarkers for the problem. The signature is validated with computational and experimental methods to ensure its

credibility. We remark that the approach is general enough to be applicable to data beyond the one used within this study.

We conclude with a summary of this thesis, emphasizing major contributions and providing insight into possible future developments in chapter 5.

Chapter 2

Statistical Analysis of *ad hoc* Clinical Data

The rapid increase in the production and availability of clinical data constitutes a grand challenge and at the same time an excellent opportunity for statisticians. Its widespread availability offers a great canvas to demonstrate the power of cutting-edge data analysis solutions in impactful real-world applications, while the unique characteristics of these datasets usually require novel or specific analyses. Statistics can provide a new insight into many issues common in a clinical context: from fixing problems with the data itself, such as filling missing data, to offering solutions at the patient's end, e.g. by producing clinical predictive models or diagnostic tools.

In this thesis, we illustrate the use of different statistical techniques to tackle problems in a biomedical context. From missing data imputation to classification models, statistical methods are an integral part of every aspect of this dissertation. In particular, throughout this chapter, we follow three major steps for statistical analysis as depicted in Figure 2.1, using two datasets related to breast cancer and, more accurately, to the problem of the development of resistance to endocrine therapy in breast cancer. They are used as case studies for the major applications of this dissertation, and as such we introduce them here in order to present their nuances and the statistical techniques required for their pre-processing.

The majority of this chapter revolves around Bayesian statistics, which we combine with Hamiltonian Monte Carlo (HMC) methodologies to use in practical applications. After presenting the necessary background, we explore methodologies to enhance the performance of HMC and its extensions, mainly focusing on importance sampling and numerical integration. Within the latter, we introduce a novel Modified Adaptive Integration Approach for computational statistics (s-MAIA), for which we present the necessary metrics to assess its quality. Among them, we include our new estimator for the Effective Sample Size (ESS) for importance sampling Hamiltonian Monte Carlo. Finally, we perform numerical experiments to demonstrate the improvements obtained with s-MAIA on standard benchmarks models, as well as the potential of s-MAIA in breast cancer-related applications.

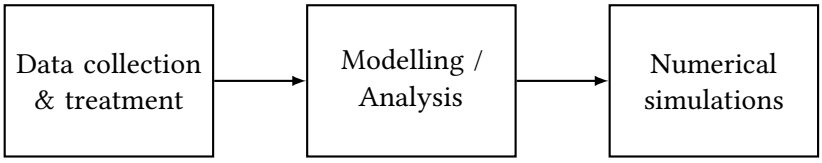

Figure 2.1: General pipeline for the mathematical analysis of clinical data.

2.1 Collection and treatment of datasets

The collection and treatment of biomedical data pose several challenges that can be solved with an accurate use of statistical methods. Even for a well formulated clinical trial, many problems can appear in the recollection process and the recording of patients information. These issues, which usually scale with the number of patients analysed, often take the form of missing data points and poor compatibility of instances within the dataset. Ultimately, the problem stems from the need for a significant number of patients within a dataset so that statistically meaningful conclusions can be extracted from a study.

Setting aside a handful of diseases, most serious illnesses do not have a high enough prevalence for a single clinician to have a sufficient amount of patients to develop a new study. This means that data from several hospitals is usually aggregated, which can lead to a non-alignment of the data. Often this is a consequence of disparities in the individual input between clinicians, or of differences in the availability of tests from one institution to another (Dhruva et al., 2020). Ultimately, solving these issues requires a great deal of human effort, as the curation of this type of information needs to be supervised, creating an unwanted overhead for any subsequent analysis done with the data (Peng et al., 2020). Nowadays, we face an extreme variant of this phenomenon. Information can be shared in easier and faster ways through the internet, but the handling of this information usually requires even greater attention. Thus, it is important that each instance is clearly defined so that there is no doubt about the type of information provided or its origin.

In this section, we present two datasets which will be later used for testing our proposed methodologies and for searching for biomarkers related to the problem of resistance to endocrine therapy in Chapter 4. Each dataset is representative of distinct approaches that can be taken for the data collection process, which, in turn, greatly affect their analysis. We first introduce the **SOX2 Clinical** dataset, created *ad hoc* to answer a specific scientific question and carried out in a single hospital. It is used at the end of this chapter in the context of **Objective 1** (Figure 1.3) for testing our proposed methodology s-MAIA. The second dataset we present was extracted from The Cancer Genome Atlas (TCGA) public online repository, and we refer to it as the **TCGA Clinical** dataset. It combines patients from several hospitals and regions within the USA (The Cancer Genome Atlas Network et al., 2013). Crucially, this dataset contains a combination of a wide variety of *clinical features* and *high-throughput data* for each patient, something critical to develop **Objective 2** (Figure 1.3) by associating biomarkers with patient condition.

2.1.1 SOX2 Clinical Dataset

2.1.1.1 Data description and collection

The data collection process for this dataset was tailor-made to suit the needs of the study it was conceived (Piva et al., 2014). The purpose of this dataset was to aid in the research on tamoxifen resistance mechanisms, which is the main research line at the Cancer Heterogeneity Lab in CIC-bioGUNE. For this, a group of ER+ breast cancer patients was tested for the presence of a specific gene marker, ***SOX2***, among other standard clinical covariates. The dataset was originally comprised of 55 breast cancer patients with samples extracted from main tumours, from which 22 patients suffered a recurrence at least once (4 of them, twice). In total, this resulted in 81 samples shown in full in Table A.1 in Appendix A. A representative selection of patients can be seen in Table 2.1.

This table contains the clinico-pathological features of 81 tumours, labelled in the first column with only a number if the tumour responded well to tamoxifen therapy. If the patient was resistant to treatment, a letter accompanies the number. **T** indicates a primary tumour, **R** is used for a recurrence, and **Rb** is used for subsequent recurrences. The rest of features and the values they take are detailed in Table 2.2.

Patient	Age	Type	Grade	Size	Node	ER	PR	HER2	p53	Ki67	Sox2
28	46	ILC	G2	T1		70%	90%	0	0	5%	1
29	61	IDC	G2	T1a	N0	90%	40%	0	0	5%	1
30	78	IDC	G2	T2	N1	90%	90%	3+	0	8%	0
31	37	IDC	G2	T1c	N0	26%	87%	3+	63%	39%	0
34T	43	IDC	G2	T2	N0	60%	65%	0	0	10%	3
34R	49	IDC	G2	T1		40%	0	1+	0	2%	8
42T	77	IDC	G3	T1c	N2a	40%	10%	3+	0	40%	3
42R	83	IDC	G3			90%	0	3+	40%	30%	3
42Rb	87	IDC	G3	rT2		23%	0	3+	5%	16%	3

Table 2.1: Extract from the complete SOX2 Clinical dataset. In green, patients that responded to therapy; in red, those who developed a recurrence.

Variable	Description
Age	Age of the patient in years.
Type	Origin of the tumoural cells.
IDC	Invasive ductal carcinoma ($\sim$ 80% cases). Starts in the milk ducts of the breast.
ILC	Invasive lobular carcinoma. Starts in the lobules, breast glands resposible for producing milk.
Grade	Measures cancer cell distinctiveness compared to healthy cells.
G1	Grade 1 or low. Cancer cells still resemble healthy ones. Maintain close to regular growth patterns and pace.
G2	Grade 2 or intermediate. Cancer cells start to look abnormal. They divide faster than healthy cells.
G3	Grade 3 or high grade. Cancer cells differ significantly. They show irregular growth and faster division rate.
Size	Characterises tumour size (T) and extent of spread.
T0	No evidence of a primary tumour.
T1 ***T2*** ***T3*** ***T4***	Represent the sizes of the tumour and the extent to which it has grown into neighbouring breast tissue. The higher the T value, the larger the tumour and/or the more it may have grown into the breast tissue.
Node	Indicates amount of spread to nearby lymph nodes (N).
NX	Cancer in nearby lymph nodes cannot be measured.
N0	Nearby lymph nodes do not contain cancer.
N1 ***N2*** ***N3***	Increase with number of lymph nodes involved and growing cancer presence. The higher the N number, the greater the extent of the lymph node involvement.
ER	Shows estrogen receptor staining intensity as a percentage.
PR	Shows progesterone receptor staining intensity as a percentage.
HER2	Rates HER2 expression on a 0 to 3 scale. Cancers with HER2 amplification are called HER2+. They are more aggressive but have specific targeted treatments.
p53	Quantifies the percentage of cells with mutant p53 protein. This mutation can induce higher proliferation in those cells.

Ki67	Quantifies, as a percentage, the abundance of the Ki67 proliferation marker. Indicates faster proliferation rate of cells.
SOX2	Factor under study to determine its power as a biomarker for resistance to tamoxifen therapy. Measured in an Allred score.

Table 2.2: Description of the features available in the SOX2 Clinical dataset.

The response variable we want to model is the event of relapse after 5 years, which happened in 42 of the cases studied, versus the 33 patients that responded well to the treatment. All the non-respondent patients had a value for the Allred score for the SOX2 gene over 2. The Allred score is a semi quantitative scoring system that takes into consideration the proportion of positive cells (on a scale of 0-5) and staining intensity (on a scale of 0-3). The proportion and intensity are then summed to produce total scores from 0 to 8. A score of 0–2 was regarded as negative, while 3–8 as positive (Allred et al., 1998).

In a practical sense, this implies that SOX2 *perfectly separates* the data into two categories where all the members have the same response to treatment. Hence, this variable can completely determine the outcome on its own in this specific dataset. This can be seen clearly in Figure 2.2 where patients responded well to the treatment if and only if they have a SOX2 Allred score lower than 3.

Although this seems like a great feature of our data, it can easily become a problem, depending on the intended application. In this case, we intended to use data to obtain information about the effects of clinical covariates beyond SOX2. However, the appearance of separation implies that maximum likelihood estimates do not exist for some binary classification models such as logistic regression (Rainey, 2016) in the classical frequentist interpretation. Circumventing this issue was initially one of the driving forces for using a Bayesian methodology for this analysis. In section 2.3.3, we explore the effects of this phenomenon in a Bayesian logistic regression (BLR) scenario and provide an analysis on how the use of different priors for the model affects the output.

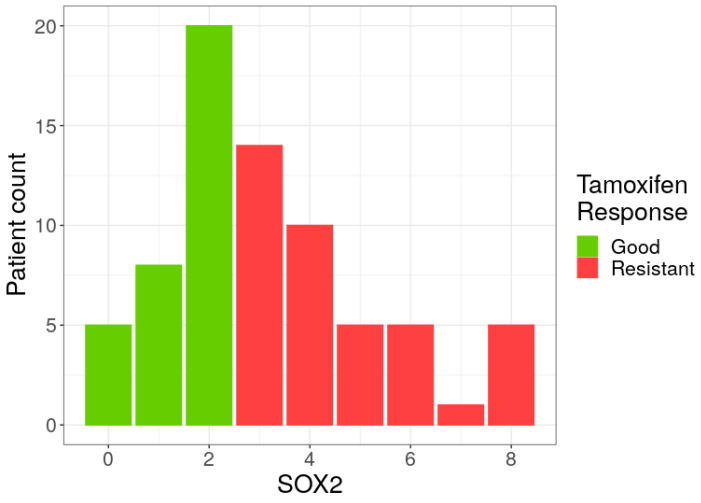

Figure 2.2: Distribution of patient's response to tamoxifen treatment. All patients with an Allred score of 3 or more show a resistant response to the treatment.

2.1.1.2 Data curation

Before the data can be used for any type of analysis, it is necessary to treat it and curate it in a way that a mathematical model can process it. For this, we need to transform the non-numerical data to mathematical form and remove or complete the missing instances in the dataset. This dataset, despite

being collected locally and being relatively small, still has a fair number of missing instances. Most of the variables in the SOX2 Clinical dataset are already numerical, or are easily convertible to this format, such as Grade or Size. The only exception is the type of tumour (IDC or ILC), but it can be easily converted into a binary variable without loss of information. With these changes, we can focus on completing the missing data points.

Missing data and imputed data

First, we explored the abundance and distribution of missing data points in the dataset with the help of the R package *mice* (Van Buuren, 2007) and the graphical tool *ggmice* (Oberman, 2023) which helped us visualize the distribution of the missing data.

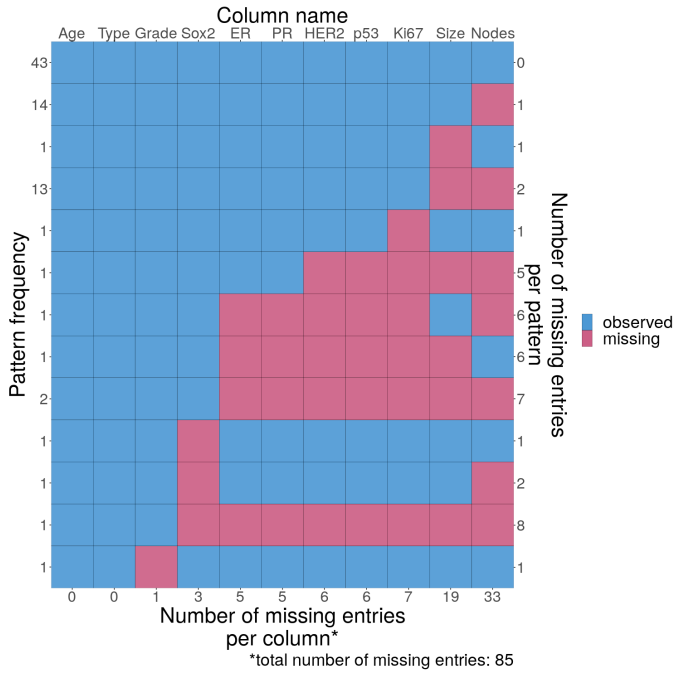

Figure 2.3: Pattern plot for missing data in the original SOX2 Clinical dataset.

In Figure 2.3 we see that 8, 7, 6 and 5 data points are missing for 1, 2, 2 and 1 patients, respectively. Those patients were eliminated from the dataset. The resulting dataset is then composed of 75 entries, 33 having a positive response to tamoxifen and 42 with a negative one.

After this, for two cases only one data point is missing, namely, *Grade* and *Ki67*. Since the *Ki67* column is numerical and comes as a percentage, we used the mean of the rest of the data points to fill the missing value, while for the categorical *Grade* we used the mode as the consensus value for the missing instance (Sessa and Syed, 2016). After this, only the SOX2, Size and Nodes categories had missing data. Figure 2.4 shows the distribution of missing data after removing those cases.

To complete the data treatment, we applied the *mice* function to the rest of data points in the dataset to create a regression model and suggest a value for imputation. This was done iteratively, and the proposed value improved on every iteration.

In Figure 2.5, we observe the distribution of the values generated by the imputation algorithm for 5 different chains. We can see that both Sizes and Nodes proposed values (in red) follow closely the real distribution of values (in blue). For the SOX2 case, the algorithm fails to accurately reproduce the behaviour. This is due to the fact that with only 3 values missing (compared with 14 for Size and 28 for Nodes) the shape of the real curve is almost impossible to reproduce. However, the proposed

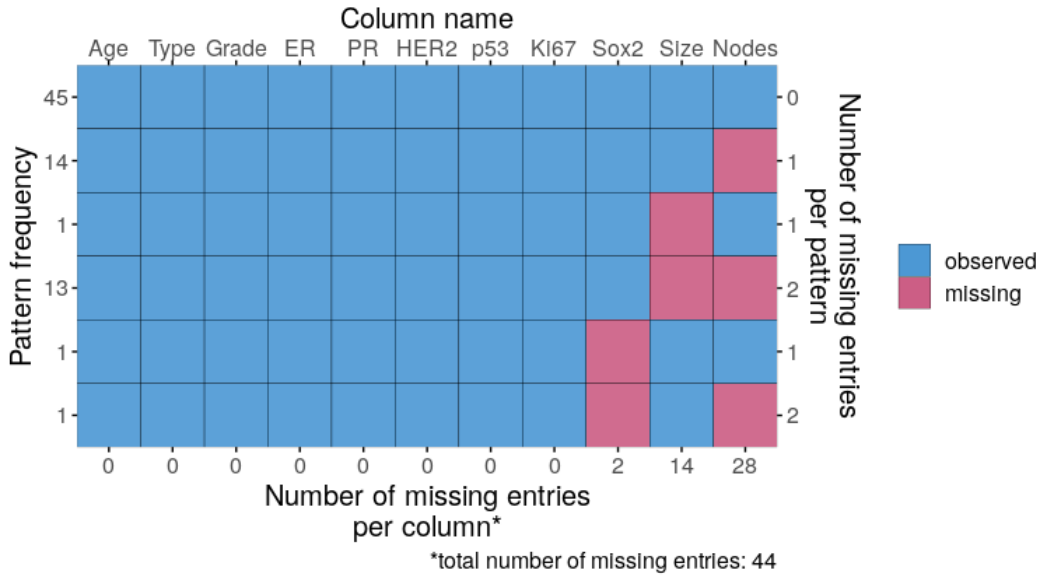

Figure 2.4: Pattern plot for missing data in the SOX2 Clinical dataset after removing patients with more than 4 variables missing.

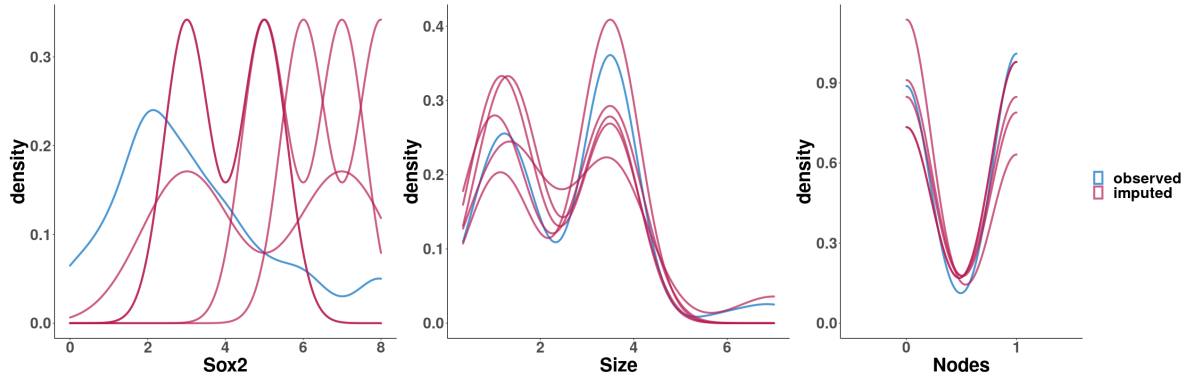

Figure 2.5: Density plots for the possible imputation values for missing data. Generated using *mice*.

values do not break the separation patterns followed by this variable in the rest of the dataset, which shows the ability of the imputation system to capture the nuances of this specific dataset.

The complete set of data resulting from this imputation methodology can be seen in Table A.2 in Appendix A. This dataset is featured in 2.3.3 as a test case within Objective 1 (Figure 1.3) for the improved Bayesian sampling methodology we derive in 2.3.2. In addition to its use for the exploration of relevant clinical factors others than SOX2, the dataset represents an interesting application for Bayesian analysis due to its separation property.

2.1.2 TCGA Clinical dataset

2.1.2.1 Data description and collection

The TCGA is a landmark research program centred on the study of cancer genomics. It compiles pan-cancerous data from 38 different types and subtypes of cancer. The cohorts in this program are

comprised of clinical patient information combined with several types of biotechnological analysis. Among them, we find sequencing analysis as well as methylation or mutation information. One of the biggest and most widely used cohorts within the TCGA is the Breast invasive carcinoma (BRCA) dataset (Network, 2012).

The complete BRCA dataset is composed of 1097 breast cancer patients with abundant clinical and sequencing information, from which 209 were ER+ & tamoxifen-treated patients. However, unlike the previously presented SOX2 Clinical dataset, the clinical information on this dataset is aggregated from different sources, meaning that some information is present for only a fraction of patients. General information, such as ER or PR positivity, is present in over 95% of patients. However, only 39% of patients show a numerical value of either ER or PR positivity, meaning that, for practical purposes, the variable should become a binary one just recording positive or negative status. However, the abundant sequencing data present in the TCGA database allows the extraction of information on key genes related to the resistance process.

2.1.2.2 Data curation

Due to the specific scientific questions that we plan to address regarding resistance to treatment, not all the 209 tamoxifen-treated patients included in the dataset are equally useful. In fact, in order to guarantee that the treatment had an effect on the patient, we should exclude patients who received tamoxifen during a brief window of time, or those whose clinical information is too inexact to accurately measure the treatment time or dosage. In many instances, the information on the duration or response to the treatment was incomplete and the methods for statistical imputation we introduced in section 2.1.1.2 were not reliable to use for such a sparse case and a time-sensible variable. Hence, we proposed and applied two filters for a patient to be selected in the first place:

1. Receiving a continuous treatment with tamoxifen for a minimum of 2 years.

2. Availability of complete information on follow-ups to monitor the appearance of resistance.

In a practical sense, the first filter even before these was the ER positivity, which was filled according to the status of the column "*breast_carcinoma_estrogen_receptor_status*" from the original clinical information category in the TCGA database. For the treatment condition, we took the column "*drug_name*" which indicated the drug of use, while "*days_to_drug_therapy_start*" and "*days_to_drug_therapy_end*" were used to estimate the duration of the treatment. Finally, the column "*days_to_last_followup*" was taken as temporal reference of subsequent follow-ups (up to 9), with the columns "*person_neoplasm_cancer_status*", "*new_tumor_event*" and "*vital_status*" needed to establish the persistence or disappearance of the tumour, the appearance of a recurrence or metastatic event, and the eventual death of a patient respectively.

A simplified version of the dataset used for this purpose is shown next in 2.3. Only a few selected columns are presented due to space restrictions from the more than 100 available clinical variables related to other clinical factors (Menopause, HER2 status), treatment (subsequent therapies) and in-depth patient follow-up. The source of this table can be extracted in full from the BRCA database at (http://firebrowse.org/) where the names of the columns appear as described above.

This resulted in a dataset with only 37 tamoxifen-treated ER+ breast cancer patients remaining. Using a similar approach, we could also select 127 patients treated with any hormone therapy. Both these cohorts play a central role in the final study of this dissertation, where the clinical and sequencing datasets are used in combination. Thus, more detailed information on the selection and classification of their response is presented in Chapter 4.

Age	ER%	PR%	Grade	Nodes	Size	Treatment	Response	Day Start	Day End	Follow-up	Death
42	75	55	G2a	N0	T2	tamoxifen	NA	322	follow-up	tumour free	NA
39	95	50	G2a	N0	T2	tamoxifen	NA	107	1258	tumour	1920
73	+	+	G3b	N1a	T4b	anastrozole (HT)	yes	335	follow-up	tumour	NA
82	40-49	NA	G4	N2	T2	tamoxifen	NA	479	1365	tumour	1365
72	50-59	95	G2a	N0	T2	aromasin (HT)	no	799	2535	tumour free	NA
54	+	+	G1	N0	T1c	tamoxifen	yes	219	2051	tumour free	NA
61	95	95	G1	N0	T1C	arimidex (HT)	NA	200	follow-up	tumour free	NA

Table 2.3: Brief fragment of the complete TCGA Clinical dataset. Death, Start and End of treatment are measured in days.

2.2 Mathematical modelling of clinical data

In section 1.1.1 we presented a motivation for the use of Bayesian statistics in biomedical applications and provided examples of its application at various phases of the data analysis pipeline. While the use of Bayesian statistics is beneficial at many stages, including the construction of clinical trials and data treatment, the centerpiece of these applications is the development of patient-oriented prognostic models. These models can be used as decision tools for diagnostic and treatment purposes in clinical practice. They are often related to evaluating risks for patients or predicting events in the evolution of a disease (Abu-Hanna and Lucas, 2001). Because of this, classification and prediction models are the ones most frequently used in practice.

In this section, we focus on introducing a Bayesian framework that we choose for developing such models. Our contribution to enhancing its performance as well as its applications will be explained in detail and further discussed in section 2.3.

2.2.1 Bayesian inference

We briefly explained Bayes' theorem in Chapter 1. However, here we plan to explore in depth the implications that it carries. We start by introducing the notations that will be used throughout the remainder of this dissertation.

Given a set of N observations constituting data points $\mathbf{y} = (y_1, ..., y_N)$, we assume that the model defined by a likelihood function has D parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_D)$. Then, the *posterior distribution* of the parameters $\boldsymbol{\theta}$ given the data $\mathbf{y}$, $p(\boldsymbol{\theta}|\mathbf{y})$, expresses the variability within model parameters and reads as

$$\overbrace{p(\boldsymbol{\theta}|\mathbf{y})}^{\text{Posterior}} = \frac{\overbrace{p(\boldsymbol{\theta})}^{\text{Prior}}\ \overbrace{p(\mathbf{y}|\boldsymbol{\theta})}^{\text{Likelihood}}}{\underbrace{p(\mathbf{y})}_{\text{Marginal Likelihood}}}. \tag{2.1}$$

Here, the *likelihood function* models the problem by linking the data $\mathbf{y}$ and the model parameters $\boldsymbol{\theta}$, while the *prior distribution* accounts for any prior information known about the distribution of the parameters $\boldsymbol{\theta}$. The *marginal likelihood* — also known as prior predictive distribution — is the sum (or integral for the continuous cases) of y over all possible values of $\boldsymbol{\theta}$ (Gelman et al., 1995)

$$p(y) = \int p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta},$$

and can be viewed as a normalizing constant. Usually this is reflected using a more general formulation of Bayes' theorem with an unnormalised posterior distribution

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

Even when an explicit evaluation of the marginal likelihood is not needed, the calculation of complex integrals is a key part of the Bayesian paradigm, essential for most applications. For example, the evaluation of an expected value over a given posterior is central to many Bayesian applications. For a function $f(\boldsymbol{\theta})$ this results in the following integral

$$\mathbb{E}_p[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \tag{2.2}$$

The Bayesian framework can be applied mainly for three tasks, which we explore throughout this thesis:

1. Calculation of model parameters, θ, via marginalization.
2. Prediction of new incoming data, $\tilde{y}$.
3. Model comparison and hypothesis testing.

The first task is explored in section 2.3.3 using the SOX2 Clinical dataset. The key fact in the Bayesian framework is that a model parameter is represented by a (posterior) probability distribution and not a single value. The other two tasks are major pieces of the algorithm developed in Chapter 4. There, we generate several models which are compared in a Bayesian framework. Each model makes predictions for incoming observations using the posterior distribution. Taking a new observation, $\tilde{y}$, one can obtain predictions by using its *posterior predictive distribution* conditional on the posterior distribution of the $\boldsymbol{\theta}$ as

$$p(\tilde{y}|\mathbf{y}) = \int p(\tilde{y}, \boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \int p(\tilde{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}, \tag{2.3}$$

assuming conditional independence of $\tilde{y}$ and $\mathbf{y}$ given $\boldsymbol{\theta}$.

Hence, applications of Bayesian inference to the majority of relevant problems lead, almost always, to computation of integrals over complex posterior distributions that have no analytical solution. Historically, this has been a major drawback, halting the use of Bayesian statistics outside a handful of simple cases. The complexity greatly increases for high-dimensional parameter spaces, as with the public datasets we showed before; or when dealing with intricate dependencies between variables, as in the case of the SOX2 clinical dataset.

As a result, numerical methods are often required to approximate the posterior distributions arising in Bayesian problems. The recent increase in computing power alongside the advances in the development of novel numerical algorithms have allowed the emergence of methods capable of handling these issues in a much easier manner. Markov Chain Monte Carlo algorithms stand-out among them due to their flexibility and accuracy.

2.2.2 Markov Chain Monte Carlo Methods

Markov Chain Monte Carlo (MCMC) methods are part of the broader family of Monte Carlo (MC) methods. The goal of an MC method is to draw samples from the target distribution $\pi(\boldsymbol{\theta})$ to estimate an integral of interest by a sample average estimator, for example, to calculate the expected value of a function $f(\boldsymbol{\theta})$,

$$I = \mathbb{E}_\pi[f(\boldsymbol{\theta})] = \int_{\mathbb{R}^d} f(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{2.4}$$

This can be done by collecting samples $\boldsymbol{\theta}_n$ from the distribution $\pi(\boldsymbol{\theta})$ and evaluating $f(\boldsymbol{\theta}_n)$ to yield an estimator

$$\hat{I} = \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{\theta}_n), \tag{2.5}$$

that converges almost surely to I as $n \to \infty$ by the Strong Law of Large Numbers (SLLN). This methodology is followed by all MC methods and the differences among them emerge mainly from the approach taken to draw the samples.

While the original formulation of Monte Carlo (Metropolis and Ulam, 1949) relied on taking independent and identically distributed (i.i.d.) random samples, MCMC iteratively constructs a Markov Chain whose invariant distribution is the target distribution $\pi(\boldsymbol{\theta})$. Generating a large number of steps in the chain leads to its eventual convergence to the target distribution. However, specific conditions are required to ensure that the distribution obtained at stationarity is the intended target.

At the core of Markov chains, we find two main components: *stochasticity* and the *Markov property*. The former is introduced in the form of randomness when estimating the probability of transition between states. The latter states that the probability of transitioning to a new state depends only on the previous state. This implies that new states of the chain $\boldsymbol{\theta}_n$ can be obtained from the previous one $\boldsymbol{\theta}_{n-1}$ using a *transition probability* $\rho_T(\boldsymbol{\theta}_n|\boldsymbol{\theta}_{n-1})$. For a distribution $\pi(\boldsymbol{\theta})$ to be a stationary distribution under these conditions, the transition probability must satisfy

$$\pi(\boldsymbol{\theta}') = \int \rho_T(\boldsymbol{\theta}'|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \quad ; \quad \forall \boldsymbol{\theta}'. \tag{2.6}$$

A Markov chain converges to its unique invariant distribution $\pi(\boldsymbol{\theta})$ if it is irreducible and aperiodic. *Irreducibility* means that there is a non-zero probability of transitioning between any pair of states within a finite number of steps, while *aperiodicity* rules out any fixed periodicity in the sequence of state transitions. As a consequence of the SLLN, these properties imply that the chain is also ergodic. *Ergodicity* guarantees that the Markov chain converges, in the limit, to a unique stationary distribution. These properties ensure that the Markov chain:

- Can explore the entire state space (irreducibility)
- Does not get stuck in loops during exploration (aperiodicity)
- Converges to the unique stationary target distribution (ergodicity).

Another important property a Markov chain may possess is reversibility. The chain is reversible if the detailed balance condition (DBC)

$$\pi(\boldsymbol{\theta})\rho_T(\boldsymbol{\theta}|\boldsymbol{\theta}') = \pi(\boldsymbol{\theta}')\rho_T(\boldsymbol{\theta}'|\boldsymbol{\theta}) \quad ; \quad \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \tag{2.7}$$

holds for any two states of the chain. It is easy to see that 2.6 can be obtained by integrating both sides of 2.7, meaning that satisfying the DBC implies that the distribution $\pi(\boldsymbol{\theta})$ is the unique invariant target distribution.

The DBC is a *sufficient but not necessary* condition to ensure samples are still being drawn from $\pi(\boldsymbol{\theta})$. As such, it is also possible to construct chains that do not satisfy this condition but still have $\pi(\boldsymbol{\theta})$ as its stationary distribution, as is the case with many irreversible MCMC methods.

MCMC stands at the basis of plenty of state-of-the-art research, where more powerful algorithms, sampling techniques or performance upgrades are being constantly proposed. Besides, due to their

flexibility and adaptability, MCMC-based methods are a clear go-to algorithm in complex problems in physics, chemistry and biology. A wide range of literature is available, exploring in greater depth the intricacies of MCMC, see e.g., (Brooks et al., 2011; Robert et al., 2005).

2.2.2.1 Metropolis-Hastings algorithm

The first and one of the most recognised MCMC methods is Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953). Initially developed for simulations in the realm of statistical mechanics, the method was later generalized for using in a computational statistical framework and its popularity and ever-lasting presence in many fields reliant on numerical simulations has made of it one of the most important algorithms of all time (Gelman et al., 1995).

Every iteration of the method consists of two major steps. First, a move in the parameter space is generated from a proposal density $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$. Then, an acceptance/rejection test is used to determine if the chain changes to the proposed state or stays in the previous one. For that purpose, an *acceptance probability* α, which depends on the current and previous states, is calculated and used to perform the acceptance or rejection by comparing it with a random number in the unit interval. We can summarise this steps in the following Algorithm 1.

Input: $\boldsymbol{\theta}_0$: Initial position,
N: Number of samples,
$q(\boldsymbol{\theta}'|\boldsymbol{\theta})$: Proposal distribution

for $n = 1$ *to* N **do**
  $\boldsymbol{\theta} = \boldsymbol{\theta}_{n-1}$
  Propose a new state from:

$$\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}'|\boldsymbol{\theta}).$$

  Calculate acceptance probability:

$$\alpha = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}.$$

  Perform a Metropolis test:
  Draw $r \sim \mathcal{U}(0,1)$
  if $r < \alpha$ **then**
    $\boldsymbol{\theta}_n = \boldsymbol{\theta}' \Longrightarrow$ *Acceptance*
  else
    $\boldsymbol{\theta}_n = \boldsymbol{\theta} \Longrightarrow$ *Rejection*
  end
end

Result: $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N)$: Samples from the target distribution

Algorithm 1: Pseudo-code for the Random Walk Metropolis-Hastings (RW-MH) algorithm.

In its original formulation by Metropolis et al., the proposal distribution was symmetric, i.e.

$$q(\boldsymbol{\theta}|\boldsymbol{\theta}') = q(\boldsymbol{\theta}'|\boldsymbol{\theta}),$$

which greatly simplified the evaluation of the acceptance probability. Another useful property of the algorithm is its independence of the computationally demanding normalizing constant of the target distribution. Indeed, at the acceptance step, it cancels out.

Thus, the Metropolis-Hastings algorithm is an incredibly versatile algorithm with a wide range of applications and extensions, making it one of the most widely used in computational statistics. Its application goes beyond extracting samples and, with small tweaks, it can be used for example, in optimization problems as in the Simulated Annealing algorithm (Černỳ, 1985; Kirkpatrick et al., 1983). However, the Metropolis-Hastings algorithm presents certain limitations that leave plenty of room for other methods to improve its efficiency. The main problem behind such limitations is the random walk behaviour of the chain. This results in a poor exploration of the parameter space and, as a consequence, in a slow convergence towards the desired target distribution. This issue only becomes worse with an increasing number of dimensions, creating a great handicap for its application to many real-world problems. In the upcoming sections, we explore some variations of the algorithm that greatly improve its performance.

2.2.3 Hamiltonian Monte Carlo (HMC)

The concept of the HMC methodology was first proposed by Duane et al. (1987) to perform lattice field theory simulations. The idea of the method, which they called Hybrid Monte Carlo, was to combine *deterministic proposals* (molecular dynamics) with *stochastic Monte Carlo* to benefit from the complementary properties of both sampling techniques. Later, Neal adapted the method for Bayesian statistics (Neal, 1994) and the name Hamiltonian Monte Carlo soon replaced Hybrid Monte Carlo in computational statistics, despite the fact that the dynamics considered in that case are fictitious, in contrast to molecular simulation applications.

As an extension of MCMC, a main goal of Hamiltonian Monte Carlo is to generate samples from a target distribution $\pi(\boldsymbol{\theta})$. For that, the methodology creates a Markov chain that has $\pi(\boldsymbol{\theta})$ as its stationary distribution. HMC achieves this by constructing an augmented target distribution related to the Hamiltonian function $H(\boldsymbol{\theta}, \mathbf{p})$

$$\pi(\boldsymbol{\theta}, \mathbf{p}) = \pi(\boldsymbol{\theta})p(\mathbf{p}) \propto \exp\left(-H(\boldsymbol{\theta}, \mathbf{p})\right), \tag{2.8}$$

which uses an auxiliary momentum variable $\mathbf{p}$ that is usually drawn from a Gaussian zero-centered distribution, $p(\mathbf{p}) = \mathcal{N}(0, M)$.

The Hamiltonian is a mathematical function stemming from physics, where it is used to describe the energy of a system of D particles and its time evolution in t. The system is represented by two (generalized) canonical coordinates, one for the position $\boldsymbol{\theta} = (\theta_1, ..., \theta_D)$ and one for the momentum $\mathbf{p} = (p_1, ..., p_D)$. The pair of D-dimensional vectors $(\boldsymbol{\theta}, \mathbf{p})$ defines a state in the phase space $\Omega \subseteq \mathbb{R}^{2D}$. The HMC method considers a separable Hamiltonian with two terms that are independent of each other

$$H(\boldsymbol{\theta}, \mathbf{p}) = K + U = \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} + U(\boldsymbol{\theta}), \tag{2.9}$$

where U and K denote the potential and kinetic energy functions, respectively. The kinetic energy is defined using the auxiliary momentum variable $\mathbf{p}$ and a mass matrix M, which is a symmetric positive definite one. Meanwhile, the potential energy term is related to the target distribution $\pi(\boldsymbol{\theta})$ as

$$U(\boldsymbol{\theta}) = -\log \pi(\boldsymbol{\theta}) + \text{const.} \tag{2.10}$$

The dynamics of the system are described by the Hamiltonian equations of motion, which, as Newton's equations, are ordinary differential equations for the generalized canonical coordinates

$$\frac{d\boldsymbol{\theta}}{dt} = M^{-1}\mathbf{p}, \qquad \frac{d\mathbf{p}}{dt} = -\nabla_\theta U(\boldsymbol{\theta}). \tag{2.11}$$

In simulations involving real physical systems, the kinetic and potential terms may be imposed by or can be estimated from the physical model. However, in computational statistics, these quantities do not possess such a direct interpretation. Nevertheless, the same strategies can be applied by cleverly selecting functions that resemble separable Hamiltonians. In computational statistics, **HMC is used to solve Bayesian inference problems**, where $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})$ is the posterior distribution of $\boldsymbol{\theta}$ given the data $\mathbf{y}$. Having determined a likelihood function $L(\mathbf{y}|\boldsymbol{\theta})$ and a suitable prior $p(\boldsymbol{\theta})$, one can set the potential energy function in 2.10 to

$$U(\boldsymbol{\theta}) = -\log \pi(\boldsymbol{\theta}) = -\log L(\mathbf{y}|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}). \tag{2.12}$$

The key addition to a convencional MCMC that HMC brings is the use of Hamiltonian trajectories in the generation of new proposed states. This is done by integrating the Hamiltonian equations of motion, which use information from the gradients of the system to generate better proposals. Since the actual form of the Hamiltonian can be very complex, integration schemes are applied in practice for this task. One can take advantage of the separable form of the Hamiltonian 2.9 to split the system 2.11 into two systems

$$\begin{aligned} A: \quad & \dot{\boldsymbol{\theta}} = \nabla_p K(\mathbf{p}) = M^{-1}\mathbf{p}\,, \qquad && \dot{\mathbf{p}} = -\nabla_\theta K(\mathbf{p}) = \mathbf{0} \\ B: \quad & \dot{\boldsymbol{\theta}} = \nabla_p U(\boldsymbol{\theta}) = \mathbf{0}\,, && \dot{\mathbf{p}} = -\nabla_\theta U(\boldsymbol{\theta}) \end{aligned} \tag{2.13}$$

which have two solution flows, one for momentum and another for position (h is an integration stepsize)

$$\psi_h^A(\boldsymbol{\theta}, \mathbf{p}) = (\boldsymbol{\theta} + hM^{-1}\mathbf{p}, \mathbf{p}), \quad \psi_h^B(\boldsymbol{\theta}, \mathbf{p} - h\nabla_\theta U(\boldsymbol{\theta})). \tag{2.14}$$

Integration schemes constructed from composition of these two flows, which are usually referred as drifts (position) and kicks (momentum), are called splitting schemes. The best known integrator in this family is the Velocity Verlet (VV) integrator (Verlet, 1967)

$$\Psi_h^{VV} = \psi_{h/2}^B \circ \psi_h^A \circ \psi_{h/2}^A. \tag{2.15}$$

The numerical schemes for HMC, including the ones resulting from the composition of these flows, need to be symplectic and reversible to ensure invariance of $\pi(\boldsymbol{\theta}, \mathbf{p})$. Symplectiness demands that Ψ_h satisfies

$$\Psi_h'(\mathbf{z})^T \mathbf{J}^{-1} \Psi_h'(\mathbf{z}) = \mathbf{J}^{-1} \quad ; \quad \forall \mathbf{z} \in \Omega,$$

where Ψ_h' is the Jacobian of Ψ_h, $\mathbf{z} = (\boldsymbol{\theta}, \mathbf{p})$ belongs to an open set in phase space Ω and $\mathbf{J}$ is

$$\mathbf{J} = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix},$$

I being the $D \times D$ identity matrix. Symplectic integrators have good numerical stability (Skeel et al., 1997) and, crucially for HMC, imply conservation of volume during integration. In addition, the numerical scheme should be reversible, i.e, $\Psi_h \circ \mathcal{F} = (\Psi_h \circ \mathcal{F})^{-1}$, where the map $\mathcal{F}(\boldsymbol{\theta}, \mathbf{p}) = (\boldsymbol{\theta}, -\mathbf{p})$ is called the *momentum flip*.

Given an initial state $(\boldsymbol{\theta}, \mathbf{p})$ at the beginning of each Monte Carlo iteration, a new proposal $(\boldsymbol{\theta}', \mathbf{p}')$ is obtained by iteratively applying a numerical scheme Ψ_h to the Hamiltonian equations 2.11 for $L \in \mathbb{N}$ steps

$$(\boldsymbol{\theta}', \mathbf{p}') = \Psi_{h,L}(\boldsymbol{\theta}, \mathbf{p}) = \underbrace{\psi_h \circ \cdots \circ \psi_h}_{L \text{ times}}(\boldsymbol{\theta}, \mathbf{p}). \tag{2.16}$$

The HMC algorithm is summarized in Algorithm 2.

Input: $\boldsymbol{\theta}_0$: Initial position,
N: Number of samples,
M: ***Mass matrix***,
h: ***Stepsize***,
L: ***Length of the trajectory***,
$\Psi_{h,L}$: ***Integration scheme***

for $n = 1$ *to* N **do**
  $\boldsymbol{\theta} = \boldsymbol{\theta}_{n-1}$
  Draw $\mathbf{p} \sim \mathcal{N}(0, M)$
  Propose a new state ***by integrating Hamiltonian dynamics with stepsize*** h ***for*** L ***steps using*** Ψ:

$$(\boldsymbol{\theta}', \mathbf{p}') = \Psi_{h,L}(\boldsymbol{\theta}, \mathbf{p}).$$

  Calculate acceptance probability ***based on*** $\Delta H = H(\boldsymbol{\theta}', \mathbf{p}') - H(\boldsymbol{\theta}, \mathbf{p})$:

$$\alpha = \min\{1, \exp(-\Delta H))\}.$$

  Perform a Metropolis test:
  Draw $r \sim \mathcal{U}(0, 1)$
  if $r < \alpha$ **then**
    $\boldsymbol{\theta}_n = \boldsymbol{\theta}' \Longrightarrow$ *Acceptance*
  else
    $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} \Longrightarrow$ *Rejection*
  end
  Discard momentum $\mathbf{p}'$
end
Result: $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N)$: Samples from the target distribution

Algorithm 2: Pseudo-code for the Hamiltonian Monte Carlo algorithm. Highlighted in ***emphatic bold*** are the new steps in the method compared to the previous algorithm (Metropolis-Hastings).

In practice, an HMC simulation is subdivided in two stages. The first stage is a *burn-in stage* where the algorithm evolves the chain until it has achieved convergence to the intended target distribution. The samples taken during this phase are discarded, however several computations can be done during this stage for calibration and diagnostics. After the burn-in stage is finished and convergence has been reached, samples are collected at the *production stage*.

In summary, HMC makes efficient use of gradient information to reduce random walk behaviour and reach faster convergence to the target distribution. Its proposals based on the numerical integration of Hamiltonian dynamics move across the sample space in larger steps compared with traditional Markov Chain Monte Carlo, and therefore the samples are less correlated. This makes HMC more efficient than MCMC for many applications. The acceptance probability in the HMC method is influenced by the differences in the Hamiltonian values at initial and final states within each iteration, as seen in Algorithm 2. Moving through Hamiltonian trajectories maintains the system's energy constant, so, in practice, only the error in numerical integration affects the acceptance probability.

This highlights a clear avenue for enhancing HMC methodologies through improvements in numerical integration. Additionally, the current algorithm discards the momentum at the end of each iteration, resulting in the loss of dynamical information about the system, which could otherwise contribute to better proposals. In the following sections, we explore some extensions of HMC designed to enhance the performance of a standard HMC algorithm.

2.2.4 Enhancing performance of HMC: Irreversibility

2.2.4.1 Generalized Hamiltonian Monte Carlo (GHMC)

In HMC, the momentum variable is discarded after each iteration. For physical applications, this means that potentially valuable dynamic information between two steps of the chain is completely ignored. In more general problems, one may lose track of a good search direction by resorting to this *full* momentum update at each iteration, which can slow down convergence. Hence, some extensions of HMC are focused on solving this issue to retain information about the dynamic part as well.

One solution is a *partial momentum update* (PMU) instead of the complete refreshment of momenta. The main idea behind the PMU is to construct a new momentum proposal from the previous one, instead of discarding it entirely after each Monte Carlo iteration. This approach helps maintain good search directions more successfully, while effectively rejecting bad ones. Initially proposed in Horowitz (1991) in the Generalised guided Monte Carlo method, the PMU was employed in conjunction with HMC in the Generalized Hybrid Monte Carlo (GHMC) method (Kennedy and Pendleton, 2001).

Let us denote the points obtained in the previous Monte Carlo iteration by $(\boldsymbol{\theta}, \mathbf{p})$. In the PMU, instead of drawing a new candidate $\mathbf{p}^*$ every iteration, we draw an additional noise vector $\mathbf{u} \sim \mathcal{N}(0, M)$ and compute

$$\begin{aligned} \mathbf{p}^* &= \sqrt{1-\varphi}\mathbf{p} + \sqrt{\varphi}\mathbf{u}, \\ \mathbf{u}^* &= -\sqrt{\varphi}\mathbf{p} + \sqrt{1-\varphi}\mathbf{u}, \end{aligned} \tag{2.17}$$

where the role of $\varphi \in (0, 1]$ is to control the extent to which the momentum can deviate from its current direction. If $\varphi \to 0$ we have $\mathbf{p}^* \to \mathbf{p}$, trying to maintain the direction of the previous search. On the contrary, if $\varphi = 1$ then $\mathbf{p}^* = \mathbf{u}$, i.e., the momentum is fully randomized, completely discarding the previous value and recovering classical HMC behaviour.

From a theoretical point of view, the PMU corresponds to an augmentation of the target distribution by a normal random variable. Note that the transformation 2.17 is orthogonal and, therefore, $(\mathbf{p}^*, \mathbf{u}^*)$ follow the same distribution as $(\mathbf{p}, \mathbf{u})$. Hence, no additional acceptance-rejection test is required for GHMC. However, as the momentum is not discarded completely from one iteration to the next, the Metropolis test in GHMC includes a momentum flip upon rejection

$$\mathcal{F}(\boldsymbol{\theta}, \mathbf{p}) = (\boldsymbol{\theta}, -\mathbf{p}), \tag{2.18}$$

to ensure that the detailed balance condition is satisfied. For GHMC, this is not done in a straightforward manner following 2.7. The condition 2.7 is satisfied for both steps of the algorithm: i) the proposal generation, using Hamiltonian dynamics followed by the Metropolis test, and ii) the partial refreshment of the momentum done, i.e., the PMU. Both of these steps leave the canonical distribution π invariant. But crucially, the composition of these two steps is non-symmetric and does not satisfy the DBC, hence the resulting Markov Chain is irreversible. However, the method satisfies the modified (or generalized) detailed balance condition, ensuring convergence to the invariant target distribution (Fang et al., 2014). Here it is important to note that the irreversible transition kernel for this modified detailed balance does not act on $H(\boldsymbol{\theta}, \mathbf{p})$ but on the augmented $H(\boldsymbol{\theta}, \mathbf{p}, \mathbf{u})$ (Song and Tan, 2022).

In terms of convergence, it is known that irreversible processes accelerate convergence to equilibrium and the reversible counterpart has the slowest rate of convergence to equilibrium among all diffusion processes that are ergodic with respect to π (Duncan et al., 2016; Ottobre, 2016). Besides, Duncan et al. also provide an in-depth analysis of how irreversibility of a Markov chain leads to a reduction in asymptotic variance. These two results indicate the advantages that irreversible samplers may offer over reversible ones, and motivate the pursuit of more advanced irreversible methods. We study the impact of irreversibility on performance of some HMC-based methods in section 2.3.1. A pseudo-code of a GHMC algorithm is presented in Algorithm 3.

Input: $(\boldsymbol{\theta}_0, \mathbf{p}_0)$: Initial state,
N: Number of samples,
M: Mass matrix,
h: Stepsize,
L: Length of the trajectory,
$\Psi_{h,L}$: Integration scheme,
$\varphi \in (0, 1]$: ***Momentum noise parameter***

for $n = 1$ *to* N **do**
  $(\boldsymbol{\theta}, \mathbf{p}) = (\boldsymbol{\theta}_{n-1}, \mathbf{p}_{n-1})$
  Draw $\mathbf{u} \sim \mathcal{N}(0, M)$ ***and perform partial momentum update***:

$$\mathbf{p}^* = \sqrt{1-\varphi}\mathbf{p} + \sqrt{\varphi}\mathbf{u}$$
$$\mathbf{u}^* = -\sqrt{\varphi}\mathbf{p} + \sqrt{1-\varphi}\mathbf{u}.$$

  Propose a new state by integrating Hamiltonian dynamics with stepsize h for L steps:

$$(\boldsymbol{\theta}', \mathbf{p}') = \Psi_{h,L}(\boldsymbol{\theta}, \mathbf{p}^*).$$

  Calculate acceptance probability based on $\Delta H = H(\boldsymbol{\theta}', \mathbf{p}') - H(\boldsymbol{\theta}, \mathbf{p}^*)$:

$$\alpha = \min\{1, \exp(-\Delta H)\}.$$

  Perform a Metropolis test:
  Draw $r \sim \mathcal{U}(0, 1)$
  if $r < \alpha$ **then**
    $(\boldsymbol{\theta}_n, \mathbf{p}_n) = (\boldsymbol{\theta}', \mathbf{p}') \Longrightarrow$ *Acceptance*
  else
    $(\boldsymbol{\theta}_n, \mathbf{p}_n) = \mathcal{F}(\boldsymbol{\theta}, \mathbf{p}^*) \Longrightarrow$ ***Rejection and momentum flip***
  end
end
Result: $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N)$: Samples from the target distribution

Algorithm 3: Pseudo-code for the Generalized Hamiltonian Monte Carlo algorithm. In ***emphatic bold*** we highlight the new steps in the method compared to the previous algorithm (HMC).

2.2.4.2 Mix & Match Hamiltonian Monte Carlo (MMHMC)

Importance sampling (IS) methodologies (Kahn and Marshall, 1953) encompass a family of Monte Carlo methods where instead of sampling directly from the desired distribution, samples are taken from a different one, often called the *importance distribution*. This change is usually motivated by the difficulty of obtaining samples from the desired distribution, or by some type of benefit arising from

a clever choice of importance distribution. To achieve this, one can easily express the expected value of a function f with respect to π as an evaluation on the importance density q instead

$$I = E_\pi[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = \int f(\boldsymbol{\theta})\frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}q(\boldsymbol{\theta})d\boldsymbol{\theta} = \int f(\boldsymbol{\theta})\omega q(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{2.19}$$

where $\omega = \dfrac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}$ is the importance weight function to be used for computing estimates for the integral 2.19 from samples $\boldsymbol{\theta}_i \sim q(\boldsymbol{\theta})$ as

$$\hat{I}_N = \frac{\sum_{i=1}^N w_i f(\boldsymbol{\theta}_i)}{\sum_{i=1}^N w_i}, \tag{2.20}$$

which converges almost surely to I as $n \to \infty$ by the SLLN.

An accurate choice of importance distribution is key for importance sampling methods and can be a complex task in practice. The method works best when the importance density $q(\boldsymbol{\theta})$ is fairly similar to $\pi(\boldsymbol{\theta})$, which also implies that there is a small variability between weights.

One of the main applications of the method are variance reduction techniques (Owen, 2013) and hence, the choice of importance distribution is often made with that objective. However, the methodology has been expanded to encompass multiple uses. As such, the importance distribution can be taken to be one where certain features of the original distribution are accentuated. In this line, it is commonly used in rare event estimation, where higher importance (weights) can be assigned to rare events, resulting in more accurate estimates than those provided by sampling in the original distribution (Blanchet and Lam, 2012).

The use of importance sampling in Hamiltonian Monte Carlo methods was motivated by the objective to maximize the acceptance rate of HMC. This gave rise to a class of HMC algorithms known as Modified Hamiltonian Monte Carlo.

Modified Hamiltonian Monte Carlo

Let us recall that the acceptance probability depends exclusively on the energy error, i.e, the differences between the evaluation of the Hamiltonian in the current and proposed states, $\Delta H = H(\boldsymbol{\theta}', \mathbf{p}') - H(\boldsymbol{\theta}, \mathbf{p})$.

Exact solutions preserve the Hamiltonian. However, in practice, one resorts to numerical methods to solve Hamiltonian equations. A simple example of the type of symplectic integrator adequate to use in this context was presented in section 2.2.3 in the form of the Velocity Verlet integrator (2.15). The use of numerical integrators results in a Hamiltonian that is not exactly preserved, having an error tied to the numerical algorithm used for the task. This, in turn, results in an error introduced between the initial and final state of a Hamiltonian that depends on the order p of the numerical scheme used, the integration stepsize h and the dimension of the problem, D, as

$$\mathbb{E}(\Delta H) = \mathcal{O}(Dh^{2p}). \tag{2.21}$$

For a given scheme of order p, one can define a modified Hamiltonian as an asymptotic expansion in powers of the integration stepsize as

$$\tilde{H} = H + h^p H_{p+1} + h^{p+1} H_{p+2} + \dots . \tag{2.22}$$

For reversible integrators, it is ensured that the modified Hamiltonian has an expansion in even powers (Radivojevic, 2016). Modified Hamiltonians are also Hamiltonians and, as such, solutions of

the modified differential equations of 2.11 if and only if the numerical scheme used is symplectic (Sanz-Serna and Calvo, 1994). For practical purposes, modified Hamiltonians are truncated to order k. A *k-order truncated modified Hamiltonian* (with $k > p$) reads as

$$\tilde{H}^{[k]} = H + h^p H_{p+1} + ... + h^{k-1} H_k, \tag{2.23}$$

and differences between modified Hamiltonians depends on k as

$$\mathbb{E}(\Delta \tilde{H}^{[k]}) = \mathcal{O}(Dh^{2k}). \tag{2.24}$$

This implies that using a k-order truncated Hamiltonian instead of a true Hamiltonian should lead to a reduction in the expected modified energy error in the integration. This is evident from comparing 2.21 and 2.24 with $k > p$. Hence, the application of modified Hamiltonians to HMC methods seems beneficial, as one may expect a higher acceptance rate by replacing H with $\tilde{H}^{[k]}$. The explicit construction of $H^{[k]}$ is discussed in length in Engle et al. (2005); Moan (2014); Skeel and Hardy (2001). For the particular case that attains us, i.e., the parametric family of symplectic splitting integrators, we refer to Radivojevic (2016). Here we can find the construction of $\tilde{H}^{[4]}$ and $\tilde{H}^{[6]}$ for 2-stage and 3-stage integrators of this family.

The first methods that introduced the use of modified Hamiltonians in Generalised guided Monte Carlo and GHMC were developed by Akhmatskaya and Reich as the Targeted Shadow Hamiltonian Monte Carlo, TSHMC (Akhmatskaya and Reich, 2006) and Generalized Shadow Hamiltonian Monte Carlo, GSHMC (Akhmatskaya and Reich, 2008), respectively. In these methods, sampling was done with respect to a modified distribution

$$\pi(\boldsymbol{\theta}, \mathbf{p}) \propto \exp\left(-\tilde{H}^{[k]}(\boldsymbol{\theta}, \mathbf{p})\right). \tag{2.25}$$

As samples are obtained from the modified distribution, this method can be viewed as an importance sampling Monte Carlo method, where the target density is that given by the true Hamiltonian, but the samples are drawn from the importance distribution generated using the modified Hamiltonian. The original distribution can be recovered by using the importance weights

$$w_i = \exp\left(-\left(H(\boldsymbol{\theta}_i, \mathbf{p}_i) - \tilde{H}^{[k]}(\boldsymbol{\theta}_i, \mathbf{p}_i)\right)\right). \tag{2.26}$$

The computational cost for evaluating $\tilde{H}^{[k]}$ typically increases tremendously with k, since higher order derivatives of the potential function $U(\boldsymbol{\theta})$ are involved. As an example, $\tilde{H}^{[k]}$ with $k = 4$ and $p = 2$ requires the evaluation of the Hessian, $\nabla_{\theta\theta} U(\boldsymbol{\theta})$):

$$\tilde{H}^{[4]}(\boldsymbol{\theta}, \mathbf{p}) = H(\boldsymbol{\theta}, \mathbf{p}) + h^2 \left(c_{21} \mathbf{p}^T M^{-1} \nabla_{\theta\theta} U(\theta) M^{-1} \mathbf{p} + c_{22} \nabla_\theta U(\boldsymbol{\theta})^T M^{-1} \nabla_\theta U(\boldsymbol{\theta})\right),$$

where both c_{21} and c_{22} depend on the specific integrator used. We explore more in-depth numerical integration for HMC-related methodologies and higher order modified Hamiltonians in the following section.

Using modified Hamiltonians in the importance distribution offers several advantages. First, the preservation of the modified Hamiltonians to a higher accuracy than that of the true Hamiltonians, should lead to higher acceptance rates, implying better sampling. Second, it naturally overcomes major challenges in importance sampling related to the construction of an adequate importance distribution and the variability of weights. The modified Hamiltonians are easy to construct, and the weights from 2.26 are small due to the proximity between the true and modified targets.

The state-of-the-art importance sampling Hamiltonian Monte Carlo method in computational statistics is the Mix & Match Hamiltonian Monte Carlo (MMHMC) methodology (Radivojević and Akhmatskaya, 2020). The method consists of three main steps.

First, the Partial Momentum Monte Carlo (PMMC) performs a partial update of the momentum based on a noise vector $\mathbf{u} \sim \mathcal{N}(0, M)$ which preserves the importance density. For that, the momentum needs to be accepted or rejected using a Modified Metropolis test with respect to an extended Hamiltonian

$$\hat{H}(\boldsymbol{\theta}, \mathbf{p}, \mathbf{u}) = \tilde{H}^{[k]}(\boldsymbol{\theta}, \mathbf{p}) + \frac{1}{2}\mathbf{u}^T M^{-1}\mathbf{u},$$

where the acceptance probability relies on $\Delta\hat{H}$ calculated for the momenta and noise vector before and after applying the noise as in equation 2.17.

Then, in the Hamiltonian Dynamics Monte Carlo (HDMC) step, a move to a proposed state is performed by integrating 2.11 using the appropriate symplectic integrator. The move is accepted with another Metropolis test, this time with respect to the truncated modified Hamiltonians $\tilde{H}^{[k]}$.

The final step is reweighting using 2.20 in order to estimate 2.19. Algorithm 4 shows the pseudo-code for the MMHMC method.

2.2.5 Enhancing performance of HMC: Numerical integrators

By this point, it has been stated in many forms the key role of numerical integration in HMC-related methodologies. In an ideal scenario with an exact flow for Hamiltonian trajectories, a 100% acceptance rate in the Metropolis tests within all HMC algorithms would be achieved, implying a reduced computational time needed for a simulation due to the lack of rejected proposals. However, in nearly all practical cases, the necessity of employing numerical integration underscores the significance of choosing an appropriate integrator in HMC simulations. Lowering the error in integration corresponds to a higher acceptance rate, thus increasing the efficiency of the (costly) generation of proposals via Hamiltonian dynamics.

Integration schemes can be constructed using the solution flows in 2.14. The most popular scheme of this type is the Velocity Verlet (2.15). A family of integrators can be derived from an alternating composition of momentum and position flows. As compositions of Hamiltonians flows they are symplectic and if they are constructed using a palindromic structure, reversibility is also ensured. This results in a family of *r-stage palindromic splitting integrators*, where *r* indicates the number of gradient evaluations per integration step done by the scheme. The family of 2-stage integrators of this kind is parameterized by $0 < b < 1/2$, and one integration step of length h of a generic member of this family is defined in Blanes et al. (2008) as

$$\Psi_h^{2-s} = \psi_{bh}^B \circ \psi_{h/2}^A \circ \psi_{(1-2b)h}^B \circ \psi_{h/2}^A \circ \psi_{bh}^B. \tag{2.27}$$

Notice that taking $b = 1/4$ corresponds to two sequential steps of the Velocity Verlet algorithm with stepsize $h/2$ and the degenerate case $b \to 1/2$ corresponds to one step of the VV integrator with stepsize h.

The 3-stage family can be similarly defined and parameterized by including $0 < a < 1/2$ as an additional parameter

$$\Psi_h^{3-s} = \psi_{bh}^B \circ \psi_{ah}^A \circ \psi_{(\frac{1}{2}-b)h}^B \circ \psi_{(1-2a)h}^A \circ \psi_{(\frac{1}{2}-b)h}^B \circ \psi_{bh}^B \circ \psi_{ah}^A. \tag{2.28}$$

Although r-stage splitting integrators can be defined generally for even and odd *r* following similar structures, here we restrict our attention to the 2- and 3-stage families as the most promising

Input: $(\boldsymbol{\theta}_0, \mathbf{p}_0)$: Initial state,
N: Number of samples,
M: Mass matrix,
h: Stepsize,
L: Length of the trajectory,
$\Psi_{h,L}$: Integration scheme,
$\varphi \in (0, 1]$: Momentum noise parameter

for $n = 1$ *to* N **do**

$(\boldsymbol{\theta}, \mathbf{p}) = (\boldsymbol{\theta}_{n-1}, \mathbf{p}_{n-1})$

PMMC step

Draw $\mathbf{u} \sim \mathcal{N}(0, M)$ and perform partial momentum update:

$$\begin{aligned} \mathbf{p}^* &= \sqrt{1-\varphi}\mathbf{p}_{n-1} + \sqrt{\varphi}\mathbf{u} \\ \mathbf{u}^* &= -\sqrt{\varphi}\mathbf{p}_{n-1} + \sqrt{1-\varphi}\mathbf{u}. \end{aligned}$$

Accept momentum $\hat{\boldsymbol{p}} = \boldsymbol{p}^*$ ***with probability*** P ***or reject and set*** $\hat{\boldsymbol{p}} = \boldsymbol{p}$**:**

$$P = \min\{1, \exp\Big(-(\hat{H}(\boldsymbol{\theta}, \mathbf{p}^*, \mathbf{u}^*) - \hat{H}(\boldsymbol{\theta}, \mathbf{p}, \mathbf{u}))\Big)\}.$$

MDMC step

Propose a new state by integrating Hamiltonian dynamics with stepsize h for L steps:

$$(\boldsymbol{\theta}', \mathbf{p}') = \Psi_{h,L}(\boldsymbol{\theta}, \hat{\mathbf{p}}).$$

Calculate acceptance probability with respect to $\tilde{H}^{[k]}$**:**

$$\alpha = \min\{1, \exp\Big(-(\tilde{H}^{[k]}(\boldsymbol{\theta}', \mathbf{p}') - \tilde{H}^{[k]}(\boldsymbol{\theta}, \hat{\mathbf{p}}))\Big)\}.$$

Perform a Metropolis test:
Draw $r \sim \mathcal{U}(0, 1)$

if $r < \alpha$ **then**

$(\boldsymbol{\theta}_n, \mathbf{p}_n) = (\boldsymbol{\theta}', \mathbf{p}') \Longrightarrow$ *Acceptance*

else

$(\boldsymbol{\theta}_n, \mathbf{p}_n) = \mathcal{F}(\boldsymbol{\theta}, \hat{\mathbf{p}}) \Longrightarrow$ *Rejection and momentum flip*

end

Compute weight $w_n = \exp\Big(-\big(H(\boldsymbol{\theta}_n, \mathbf{p}_n) - \tilde{H}^{[k]}(\boldsymbol{\theta}_n, \mathbf{p}_n)\big)\Big)$.

end

Estimate an integral of interest as

$$\hat{I}_N = \frac{\sum_{i=1}^N w_i f(\boldsymbol{\theta}_i)}{\sum_{i=1}^N w_i}.$$

Algorithm 4: Pseudo-code for the Mix & Match Hamiltonian Monte Carlo algorithm. In ***emphatic bold***, new steps in the method compared to previous algorithm (GHMC) are highlighted.

schemes for both HMC (Blanes et al., 2014; Nagar et al., 2024) and MHMC methodologies (Radivojević et al., 2018). For a detailed review on palindromic splitting integrators we refer to Bou-Rabee and Sanz-Serna (2018), while information on the derivation of r-stage integrators for $r > 3$ with modified Hamiltonians can be found in Radivojevic (2016).

Integrators 2.27 and 2.28 are part of a family of r-stage palindromic splitting integrators which are fully characterised by $r - 1$ parameters Γ, where

$$\Gamma = \begin{cases} b & \text{for } r = 2 \\ a, b & \text{for } r = 3. \end{cases} \tag{2.29}$$

By cleverly selecting values of Γ, one can potentiate some properties of interest of an integrator. Among the important properties of an integrator are the so-called *stability limit* and, related to it, the closed interval from 0 to the stability limit, referred to as *stability interval*.

For a given b (or $\{a, b\}$ pair in the 3-stage scenario), the stability limit of a given integrator represents the longest stepsize h at which the integrator is stable for the underlying model, whereas the numerical solution grows unboundedly for bigger values of the stepsize. It is known that, among all choices of integrator parameter, $b = 1/2r$ (corresponding to r-stage Verlet integrators) leads to the longest stability limit within the class. Since the stability interval is model-dependent, it is convenient and common to consider a dimensionless stepsize $\bar{h} > 0$ so that, for any system, the standard one-stage VV has a stability limit of 2 (that amounts to 4 for the 2-stage implementation of the VV and 6 for the 3-stage case). Even though the Verlet schemes possess the longest stability intervals, other methods can offer better performance elsewhere, at the cost of having a smaller stability interval, i.e, a narrower range of stable stepsizes to perform the integration.

Selecting optimal coefficients for integrators in HMC or MMHMC can easily become a delicate parameter tuning problem. The common strategy is to choose integrator parameters as minimizers of an energy error-related figure of merit.

The Minimum Error (ME) integrator, first suggested as a 2-stage integrator in McLachlan and Atela (1992) and expanded to the 3-stage family in Predescu et al. (2012), aims to minimize the *Hamiltonian truncation error* as $h \to 0$. The error introduced in one iteration is of order $\mathcal{O}(h^{p+1})$, p being the order of the integrator. The size of this error for a 2-stage integrator is then proportional to Ch^3, and the value of the constant C depends on the choice of an integrator coefficient. For 2-stage integrators (2.27), Mclachlan and Atela found that the value that results in the minimum Hamiltonian truncation error is $b \approx 0.1932$. Our interest in this thesis lies in the modified Minimum Error integrator (M-ME), derived using similar ideas in Radivojević et al. (2018) for its application in Modified Hamiltonian Monte Carlo.

An alternative way to minimize the energy error was suggested by Blanes et al. in Blanes et al. (2014), and was based on the analysis of the Hamiltonian of a standard harmonic oscillator — which in computational statistics is equivalent to zero-mean univariate Gaussian distribution with unit variance. For an r-order multi-stage splitting integrator, they obtained the associated r coefficients ($r = 2, 3, 4$) by minimizing the maximum of the tight upper bound for the expected value of the energy error, $\rho(h, \Gamma)$, within a half stability interval, i.e., $h \in (0, r)$.

The method is usually called BCSS after the scientists that proposed it (Blanes, Casas and Sanz-Serna). The derivation of even more efficient 3-stage BCSS integrators for HMC, as well as their stability analysis, can be found in Campos and Sanz-Serna (2017). In this thesis, we make use of the 2- and 3-stage versions of the BCSS integrator for modified Hamiltonians, for which we refer again to Radivojević et al. (2018). Here we remark that the modified 2- and 3-stage integrators were derived using a tight upper bound for the expected modified energy error.

A key consideration when choosing an integrator is the stepsize for the simulation. Both integra-

Integrator	N. of stages	Parameters	Stability limit
M-ME2s	2	b=0.230907	4.089
M-ME3s	3	b= 0.142757 a = (2b- 1)/4(3b- 1)	4.887
M-BCSS2s	2	b=0.238016	4.144
M-BCSS3s	3	b=0.144115 a = (2b- 1)/4(3b- 1)	4.902

Table 2.4: Subset of the splitting integrator families for modified HMC used in this work and derived in Radivojević et al. (2018).

tors discussed above, (M-)ME and (M-)BCSS, are based on assumptions that limit the effectiveness of their choices of parameters to specific stepsize regions. ME works best for small values of h while BCSS optimal performance region is around the half stability limit. As a result, several approaches have emerged recently to face the problem of selecting an integrator adaptively, i.e., obtaining the optimal coefficients for the splitting integrator of choice given a stepsize. A 2-stage Adaptive Integration Approach (AIA) was originally developed by Fernández-Pendás et al. for molecular simulations (Fernández-Pendás et al., 2016). For a given problem and stepsize, this method automatically chooses an optimal (in terms of the best conservation of energy for harmonic forces) integration scheme within the family 2.27.

The method offers greater flexibility in the selection of stepsizes and blends with the previously shown method by offering values close to ME and VV for the values of the stepsize in the regions closer to zero and the stability limit, respectively. Crucially, the method does not carry with it any major computational overhead, as the optimal coefficients are computed at the preprocessing stage.

The AIA methodology was expanded for the use with modified Hamiltonians (MAIA), also for molecular simulations, in Akhmatskaya et al. (2017). MAIA proved to be more efficient and robust in MMHMC than its HMC counterpart, while demonstrating its superiority in sampling efficiency over all fixed parameter 2-stage integrators.

However, molecular simulations present a favourable environment for the adaptive approaches AIA and MAIA that is difficult to replicate in their application to Bayesian inference problems. In molecular dynamics, physical conditions imply that harmonic forces are dominant, and for many problems, angular frequencies and resonance conditions are known. Moreover, the stepsize is usually fixed in contrast to computational statistics, where randomization of the stepsize has proven to increase efficiency of HMC algorithms. Hence, the translation of an adaptative methodology to computational problems in statistics is far from trivial. Very recently, a methodology that addresses those issues was proposed in Nagar et al. (2024) for HMC. Here, our aim is to introduce a novel adaptive approach for modified HMC when applied to Bayesian inference problems (s-MAIA). We derive and test such a methodology in section 2.3.2.

2.2.6 Performance evaluation of HMC methods

Up to this point, we have discussed methods constructed by adding complexity to previously existing MCMC methodologies to enhance simulation quality. However, we have not delved into how these enhancements can be quantified. In this section, we introduce the metrics for assessing performance of sampling methodologies.

The methods proposed so far aim to improve certain inherently valuable properties of MCMC methods, such as convergence, space exploration or sample quality. Some of these, such as convergence or error estimation, provide information on the reliability and the goodness of the simulations. Other, such as sampling quality or exploration, allow us to judge efficiency of the methodology, which

ultimately relates to computational time or computational resources used.

In practice, several metrics exist to assess the quality of a method or any proposed modifications and improvements. We have already discussed the *acceptance rate (AR)*. It measures the percentage of accepted proposals in the Metropolis test and we have already seen how different methodologies target improvements in this area. Another common metric is the *Potential Scale Reduction Factor* (PSRF), also known as $\hat{R}$, which is used to assess convergence to the stationary distribution of M-parallel Markov chains. Originally conceived in Gelman and Rubin (1992), the method was further refined by Brooks and Gelman (1998). It relies on within-chain (W) and between-chain (B) variances. For M Markov chains each drawing N samples from the posterior distribution, the within-chain variance uses the classical formula for the variance of a given chain, σ^2 and computes the average over all chains as

$$W = \frac{1}{M}\sum_{m=1}^{M}\sigma_m^2,$$

while the between-chain variance computes the variance of the individual chain means $\bar{\theta}_m$ with respect to the global mean across chains $\bar{\theta}_G$, i.e.

$$B = \frac{N}{M-1}\sum_{m=1}^{M}(\bar{\theta}_m - \bar{\theta}_G)^2.$$

The sample variance from all chains combined is obtained by taking a weighted average of both these values as

$$\hat{\sigma} = (1 - \frac{1}{N})W + \frac{1}{N}B,$$

which is used to define

$$\hat{V} = \hat{\sigma} + \frac{B}{MN}.$$

With this, the potential scale reduction factor is defined as:

$$\hat{R} = \sqrt{\frac{d+3}{d+1}\frac{\hat{V}}{W}},$$

where d is the number of degrees of freedom of a t-distribution with mean $\bar{\theta}_G$ and variance $\hat{V}$ estimated by the method of moments

$$d \approx \frac{2\hat{V}^2}{Var(\hat{V})}.$$

A value of $\hat{R} \approx 1$ indicates a good convergence, and the closer to 1 the values are, the better convergence is. In the original methodology, Gelman and Rubin recommended running longer simulations if $\hat{R} > 1.1$. Nowadays, the threshold is usually stricter and, to ensure independence with respect to the initial point, a value of at least $\hat{R} < 1.01$ is recommended for achieving convergence to the intended stationary distribution (Vehtari et al., 2021).

Beyond diagnostic measures, the evaluation of performance is measured by estimating the quality of the samples produced. The MCMC methods do not gather samples directly from the target

distribution and use various algorithms for generating samples — from producing correlated chains of samples, such as in Metropolis Hastings, to sampling from an importance distribution as, e.g., in MMHMC. Thus, the samples extracted using these algorithms are not equivalent to i.i.d. samples extracted directly from the target distribution. Hence, we need to define a way to quantify the efficiency of a sampler in producing independent samples from the target distribution.

The *Effective Sample Size* (ESS) is defined as the number of equivalent i.i.d samples from the target the algorithm in use can generate. The classical approach to its estimation in MCMC is based on the autocorrelation of samples

$$ESS_{MCMC} = N \frac{1}{1 + 2\sum_{k=1}^{\infty} \gamma_k}, \tag{2.30}$$

where γ_k is the autocorrelation function at lag k, and the dominator represents the autocorrelation within the sequence with different lags, let us call it, $\mathcal{F}_{acc}$. The more disconnected samples are in the chain, the quicker this function declines. For completely independent samples we naturally have $ESS = N$, but most cases have $ESS < N$, indicating the quality of the samples obtained is lower than the N samples actually taken. The function $\mathcal{F}_{acc}$ begins at 1 (correlation of a sample with itself) and decays until it eventually starts oscillating around 0. The oscillatory nature of the function means that it is typically truncated at some point to avoid the introduction of unwanted noise in the calculation. The most common truncation index is the *initial monotone sequence estimator* (Geyer, 1992) which truncates the sum in 2.30 at the minimum k such that the estimated sequence is monotone. We denote this value as m and obtain

$$ESS_{Geyer} = N \frac{1}{1 + 2\sum_{k=1}^{m} \gamma_k}. \tag{2.31}$$

However, this formulation works under the assumption of having **reversible chains** which, as we saw in previous sections, is a property that some advanced methods do not possess. Instead, a triangular Bartlett window of width K (Thiébaux and Zwiers, 1984) that overcomes this assumption can be used to define ESS(Ortigueira, 2010):

$$ESS_{Bartlett} = \frac{N}{1 + 2\sum_{k=1}^{K} (1 - \frac{k}{K})\gamma_k}. \tag{2.32}$$

Various methods have explored different routes to estimate ESS other than directly truncating this sum. These methods can also be used for irreversible chains. The R package CODA (Convergence, Diagnosis and Output Analysis) (Plummer et al., 2006) fits an autoregressive model to the chain and uses its spectral density at frequency zero to estimate ESS. A different approach based on batch-means estimation (Vats et al., 2019) divides the chain in a_n batches of size b_n with mean $\overline{Y}_k$. The method computes an estimate for the asymptotic (by the Central Limit Theorem.) covariance matrix in the Markov chain as

$$\Sigma_n = \frac{b_n}{a_n - 1} \sum_{k=1}^{a_n} (\overline{Y}_k - \theta_n)(\overline{Y}_k - \theta_n)^T.$$

Taking Λ to be the sample covariance matrix, Σ the estimate of the Monte Carlo covariance matrix for the Markov chain and N the current sample size, the method, called multivariate ESS

(multiESS), uses the ratio of covariance matrices to obtain an estimate of the ESS, which in addition takes into account the dimensionality of the problem, D

$$multiESS = N\frac{|\Sigma|^{1/D}}{|\Lambda|^{1/D}}, \tag{2.33}$$

while a similar univariate estimation of the ESS per dimension can be done by taking the diagonal values of both matrices, σ_i and λ_i respectively

$$uniESS_i = \frac{\sigma_i^2}{\lambda_i^2}. \tag{2.34}$$

When using **importance sampling**, correlation is not usually an issue but, as before, the samples are not taken directly from the stationary target distribution. In this case, the ESS is usually defined through the importance weights $\omega_n, n = 1, ..., N$, as first introduced by Kong (Kong et al., 1994)

$$ESS_{IS} = \frac{\left(\sum_{n=1}^{N}\omega_n\right)^2}{\sum_{n=1}^{N}\omega_n^2} = \frac{1}{\sum_{n=1}^{N}\bar{\omega}_n^2}, \tag{2.35}$$

where $\bar{w}_i$ are the normalized weights defined as

$$\bar{\omega}_i = \frac{\omega_i}{\sum_{j=1}^{N}\omega_j}. \tag{2.36}$$

Recently, some efforts have been made to improve on this original approach (Elvira et al., 2022; Martino et al., 2017). Elvira and his collaborators identified several undesirable properties of Kong's original formulation, namely:

1. Upper-bounded by N and lower-bounded by 1.
2. Independent of the function of interest f.
3. Independent of the samples taken $\{\theta_i\}_{n=1}^{N}$.

The first issue is a key one, as it presents a theoretical limitation of the formula by restricting $1 \leq ESS_{IS} \leq N$. However, in practice, reduction of variance techniques can force $ESS_{IS} \geq N$, as well as a very poor choice of importance distribution can make $ESS_{IS} = 0$. The other two points can be considered as a trade-off between the simplicity of only using weights and the accuracy of inserting more complex terms into the estimate. Therefore, Elvira and Martino presented several variations of the initial formula, taking care of the first point and issuing recommendations for the inclusion of the other two.

Among the proposed formulations, they selected one based on the Huggins-Roy's family of metrics $H_N^{(\beta)}(\bar{\omega})$ (Huggins and Roy, 2019):

$$ESS_{H_N} = H_N^{(\beta)}(\bar{\omega}) = \left(\frac{1}{\sum_{i=1}^{N}\bar{\omega}_i^{\beta}}\right)^{\frac{1}{\beta-1}} = \left(\sum_{i=1}^{N}\bar{\omega}_i^{\beta}\right)^{\frac{1}{1-\beta}}, \qquad \beta \geq 0.$$

The Huggins-Roy's family is related to the *Rényi entropy* $R_N^{(\beta)}(\bar{\omega})$ of the probability mass function defined by $\bar{\omega} = \{\bar{\omega}_i\}_{i=1}^N$ (Cover, 1999) as

$$R_N^{(\beta)}(\bar{\omega}) = \log(H_N^{(\beta)}(\bar{\omega})) = \frac{1}{1-\beta}\log\left(\sum_{i=1}^{N}\bar{\omega}_i^{\beta}\right), \qquad \beta \geq 0. \tag{2.37}$$

From this family they derived several estimators for different values of the parameter β

$$\begin{aligned} ESS_{sqrt}^{1/2} &= \left(\sum_{n=1}^{N}\sqrt{\bar{\omega}_n}\right)^2, \qquad \beta = 1/2, \\ ESS_{Renyi}^{1} &= exp\left(-\sum_{n=1}^{N}\bar{\omega}_n^{\log\bar{\omega}}\right), \qquad \beta = 1, \\ ESS_{max}^{\infty} &= \frac{1}{\max[\bar{\omega}_1, ..., \bar{\omega}_N]}. \qquad \beta = \infty. \end{aligned} \tag{2.38}$$

Another recent approach for ESS estimators in importance samplers uses the ideas of (Vats et al., 2019) and the ratio of variances formulation presented in 2.33 (Agarwal et al., 2022). In this IS adaptation, the unnormalized importance sampling estimator and the self-normalized importance sampling estimator take the roles of Σ and Λ in equation 2.33 respectively.

However, existing formulations fail to account for all the sources of variance appearing in importance sampling methods that rely on MCMC algorithms (i.e. GSHMC & MMHMC). For such cases, it is necessary to introduce novel methods that take into account the effects in ESS of both correlation and importance sampling simultaneously. In the next section, we provide an approach to calculate ESS for MMHMC based on the metrics we just presented.

Finally, an important performance metric derived from the effective sample size is the *Monte Carlo Standard Error* (MCSE), which evaluates the precision and reliability of Monte Carlo simulations. It measures the variability in estimated parameters, indicating the deviation of simulation results from N samples θ_n and true values taken to be the estimated mean value $\bar{\theta}$. A lower MCSE signifies higher precision and reliability, implying effective convergence to the target distribution. The MCSE expression further combines variance estimation with ESS, providing a value calculated over the number of effective samples rather than the total samples taken:

$$MCSE = \frac{\sigma^2}{ESS} = \frac{\frac{1}{N-1}\sum_{n=1}^{N}(\theta_n - \bar{\theta})^2}{ESS}. \tag{2.39}$$

A summary of the relevant metrics to be used in numerical experiments in section 2.3 can be found in Table 2.5.

Metric	Determines
PSRF	Convergence
AR	Space Exploration
MCSE	Variance
ESS	Sampling Efficiency

Table 2.5: Metrics used in the numerical tests for diagnostics and estimation of performance.

2.3 Boosting sampling in biomedical applications with the reinforced MMHMC method

Our objective now is to further improve performance of the promising HMC-based sampling method, MMHMC, and adapt it upmostly to biomedical applications. This section revolves around our novel approach for adaptive integration in Modified Hamiltonian Monte Calo for Bayesian inference, which we call s-MAIA. We start with refining the performance metrics for the methodologies relying on both importance sampling (IS) and MCMC and propose a new insight into the selection of appropriate ESS estimators for Modified Hamiltonian Monte Carlo methods. We then derive the s-MAIA method and test its performance in MMHMC on standard benchmarks against the other state-of-the-art integrators, as well as against HMC with its optimal settings.

The new methodologies and all the simulations in this section were implemented in and performed with the BCAM in-house software HaiCS (Hamiltonians in Computational Statistics), originally developed in Radivojevic (2016). The package was designed to perform Bayesian inference and parameter estimation in high dimensional and complex models using MCMC and HMC based methods. HaiCS offers all the state-of-the-art samplers and integrators we discussed in previous sections (and many more), providing a versatile platform not just for their application but also for the development of novel methodologies. While written in C for optimal performance, the package HaiCS is complemented by the analysis and diagnostics tools coded in R, which makes its output compatible with many popular packages for Bayesian analytics such as *mcmc* (Geyer and Johnson, 2020) or *CODA* (Plummer et al., 2006).

2.3.1 Effective Sample Size for importance sampling HMC

In the previous section, we pointed out that estimating the effective sample size for samplers involving both importance sampling and Markov Chain Monte Carlo remains an open problem. In this section, we analyse the behaviour of several approaches for calculation of ESS in MCMC under different sampling conditions, namely, irreversibility and importance sampling, in order to identify the most accurate approaches.

To the best of our knowledge, the first approach for computing ESS for irreversible importance sampling MCMC was proposed in Radivojević and Akhmatskaya (2020). The idea was to apply 2.35 to M uncorrelated samples out of N selected ones, identified with the help of the CODA package (Plummer et al., 2006) along with a thinning procedure at a distance of N/M. This resulted in

$$ESS_{thinned} = \frac{1}{\sum_{n=1}^{M} \bar{\omega}_n^2} = \frac{\left(\sum_{n=1}^{M} \omega_n\right)^2}{\sum_{n=1}^{M} \omega_n^2}. \tag{2.40}$$

However, there are several practical issues with this approach:

1. The thinning distance is taken as the closest integer to N/M rounded down $\rightarrow$ Resulting ESS estimation is rather rough, especially, for $N/M < 1.5$.
2. The approach requires $M \leq N \rightarrow$ It excludes the realistic scenario in which $ESS > N$.
3. The use of thinning implies a loss of already generated samples $\rightarrow$ Extra computational cost required to replicate a number of uncorrelated samples close to N.
4. It is constricted to the classical definition of ESS_{IS} reliant on weights.

Here, we propose a formulation of ESS that overcomes these challenges by not being bounded and allowing for the inclusion of any ESS_{IS} or ESS_{MCMC} metrics. For that, we recall the definition of ESS as a ratio of variances which is valid for both MCMC (Vats et al., 2019) and IS (Agarwal et al., 2022). Let us denote ESS of an MCMC method sampling from π as ESS_π and the variance observed in the estimation of 2.4 using this method as Var_π. On the other hand, we call ESS_{iid} and Var_{iid} the estimator and variance, respectively, of the method which samples from π by generating uncorrelated i.i.d. samples only, meaning that $ESS_{iid} = N$. Then ESS_π can be calculated as:

$$ESS_\pi = N\frac{Var_{iid}}{Var_\pi}. \tag{2.41}$$

In most cases, $Var_\pi > Var_{iid} \implies ESS < N$, either due to correlation in the construction of the Markov Chain $\pi(\boldsymbol{\theta})$ or due to sampling from the importance distribution $q(\boldsymbol{\theta})$ and reweighting samples instead of extracting them from the intended target.

By taking the variance introduced by both estimators simultaneously, we can obtain an ESS estimator for MCMC methods that use both, correlated samples and importance sampling:

$$ESS = \frac{1}{N}\overbrace{N\frac{Var_{iid}}{Var_\pi}}^{MCMC}\underbrace{N\frac{Var_\pi}{Var_q}}_{IS}, \tag{2.42}$$

which leads to the following expression for the ESS estimator for Monte Carlo & Importance Sampling

$$ESS_{MCIS} = \frac{ESS_{MCMC}}{N}ESS_{IS}. \tag{2.43}$$

This equation essentially operates as 2.40 if Kong's methodology is used for ESS_{IS}, where ESS_{MCMC} takes the role of the summation index, M. However, this new methodology is more general and flexible than 2.40 and can be adapted for use with different ESS formulations in order to obtain the most accurate estimation of the effective sample size.

The approach we choose to use, i.e. 2.43, should be able to handle the impact of several properties of the sampling methodologies such as correlation, irreversibility of the chain or importance sampling in the correct manner. This is something that has not been studied in detail before. In fact, one of the most popular approaches is still the initial monotone sequence estimator by Geyer (2.31), which is defined specifically for reversible chains. In table 2.6 we present the properties that could affect the ESS estimation for each of the MCMC samplers discussed in section 2.2. Correlation is the basic property that these samples share and the property that ESS targets in the first place. Next, we intend to study the impact of irreversibility and importance sampling on ESS using available metrics in order to identify the proper metric for each of the MCMC sampler of interest.

	RW-MH	HMC	GHMC	MMHMC
Correlation	✓	✓	✓	✓
Irreversibility	X	X	✓	✓
Importance Sampling	X	X	X	✓

Table 2.6: Relevant for ESS estimation properties of the MCMC samplers presented in this study.

Irreversibility

We start with the analysis of the effect of irreversibility on the ESS estimations in GHMC and MMHMC (without reweighting). For the numerical experiments, we selected a Banana-shaped distribution, as presented in Girolami and Calderhead (2011). We are interested in comparing the behaviour of different ESS metrics under optimal simulation conditions for HMC, GHMC and MMHMC. Optimal parameters for the Banana-shaped distribution in these scenarios were identified in Radivojević and Akhmatskaya (2020). However, in our experiments, we doubled the recommended trajectory length as it resulted in a better overall performance. The likelihood of this Banana distribution is given as

$$y_k \sim \mathcal{N}(\theta_1 + \theta_2^2, \sigma_y^2), \quad k = 1, ...K,$$

where the prior is $\theta_1, \theta_2 \sim \mathcal{N}(0, \sigma_\theta^2)$ with $\theta_1 + \theta_2 = 1$, $\sigma_y = 2$ and $\sigma_\theta = 1$.

We took K=100 and ran simulations with HaiCS for 5.000 burn-in and 5.000 production iterations using L=14, h=1/9 and $\phi = 0.5$ (for GHMC and MMHMC). We repeated this for 10 chains and the results were averaged over chains. Figure 2.6 shows the mean ESS for each variate calculated using the five ESS approaches presented in the previous section for MCMC (see Table 2.7).

Name	Method
Bartlett	Eq. 2.32
Coda	(Plummer et al., 2006)
MultiESS	Eq. 2.33
UniESS	Eq. 2.34
Geyer	Eq. 2.31

Table 2.7: Methods used in Figure 2.6.

In all cases, except for Geyer's, we see the expected behaviour with optimal parameters, i.e, $ESS_{HMC} < ESS_{GHMC} < ESS_{MMHMC}$. Moreover, since Geyer's approach does not support irreversible samplers, its results for GHMC and MMHMC cannot be considered as meaningful. The other results are well consistent for all estimators and all samplers. The estimators which are solely based on autocorrelation, like Bartlett and CODA, demonstrate very similar behaviour, whereas their results visibly differ from those obtained by Geyer's (HMC), and the Univariate and Multivariate methods, i.e. the methods relying on the ratio of variances approach.

The difference in estimation between both groups is around 10% across all samplers and it seems to arise from the methodological differences. In terms of behaviour, we can only rule out Geyer's equation for the irreversible cases. As we intend to study the effects of autocorrelation, we will move forward with both Bartlett and CODA for our next analysis, although we will also keep the uniESS model to investigate further the implications of these differences.

Importance sampling

To make the final selection, we tested several combinations of various ESS_{MCMC} and ESS_{IS} estimators to find the most appropriate ones for using in 2.43. We performed the test on a 1000-dimensional multivariate Gaussian $\mathcal{N}(0, \Sigma)$ where the precision matrix, Σ^{-1}, is generated from a Wishart distribution following (Hoffman et al., 2014). The motivation for this particular choice of Σ is that the resulting multivariate Gaussian has a strong correlation across its dimensions. The high dimensionality of the problem provides a great environment for benchmarking HMC-derived methodologies. More importantly, in this case, having the exact distribution allows us to take i.i.d. samples directly from it to obtain a "true" ESS estimator as a reference for further comparison of proposed estimators.

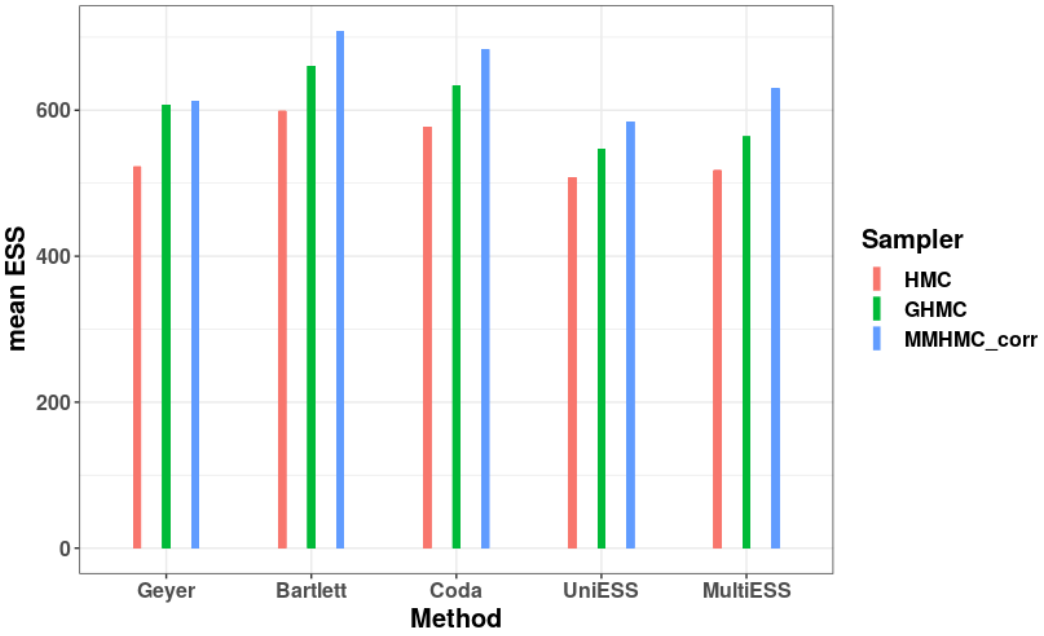

Figure 2.6: Behaviour of several ESS methods in HMC, GHMC and MMHMC (unweighted) scenarios.

The chosen model was run with MMHMC using HaiCS for 5000 burn-in and 20000 production iterations, and the simulations were repeated 10 times for averaging. The parameters were chosen as proposed in Radivojević and Akhmatskaya (2020). With the resulting trajectories, we obtained the ESS estimates for three different ESS methodologies in MCMC conditions and another four for IS, summarised in Table 2.8.

Name	Method	Feature
Coda	(Plummer et al., 2006)	MCMC
Bartlett	Eq. 2.32	MCMC
UniESS	Eq. 2.34	MCMC
IS Kong	Eq. 2.35	IS
IS sqrt	Eq. 2.38 (top)	IS
IS ωmax	Eq. 2.38 (bottom)	IS
IS Renyi	Eq. 2.38 (center)	IS

Table 2.8: Methods used in Figure 2.7.

Figure 2.7 shows that MCMC estimators (depicted in green) have normalized ESS values close to N (ESS is displayed normalized to the total number of samples, N), in contrast to the true average being below 0.5. As such, from this MCMC metrics group the most noteworthy conclusion is that the Bartlett metric is the only one offering the $ESS < N$. Among the IS estimators (shown in red), the only metric that shows a good comparison with the "true" value is the one using the maximum weight, i.e. $IS_{\omega max}$. All other metrics offer mean values of ESS that are similar to the results obtained with the MCMC estimators and significantly exceed the "true" value.

Figure 2.8 shows the average normalized ESS values for various combinations of the MCMC and IS estimators forming the ESS_{MCIS} in 2.43. Table 2.9 provides the details of such combinations.

We notice that all three combinations using $IS_{\omega max}$, bring the value closer to the “true” ESS, with ωmax-Barlett and ωmax-UniESS showing the best comparison. On the contrary, the Thinned metric which relies on the Kong’s approach 2.35 was not able to reproduce the "true" value. Based on these results, we choose to use the ωmax-Barlett metric for all the upcoming numerical experiments.

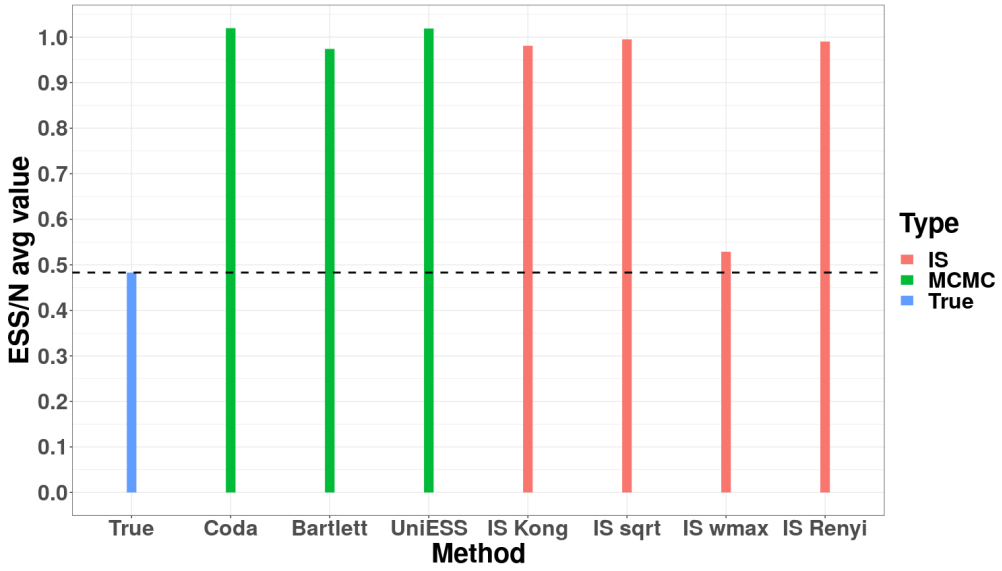

Figure 2.7: Comparison of several ESS metrics with the reference value obtained using i.i.d. samples from a 1000-D multivariate Gaussian.

Name	Method
Bartlett + IS ωmax	Eq. 2.32 + Eq. 2.38 (bottom)
Coda + IS ωmax	(Plummer et al., 2006) + Eq. 2.38 (bottom)
UniESS + IS ωmax	Eq. 2.34 + Eq. 2.38 (bottom)
Thinned	Eq. 2.40

Table 2.9: Methods used in Figure 2.8.

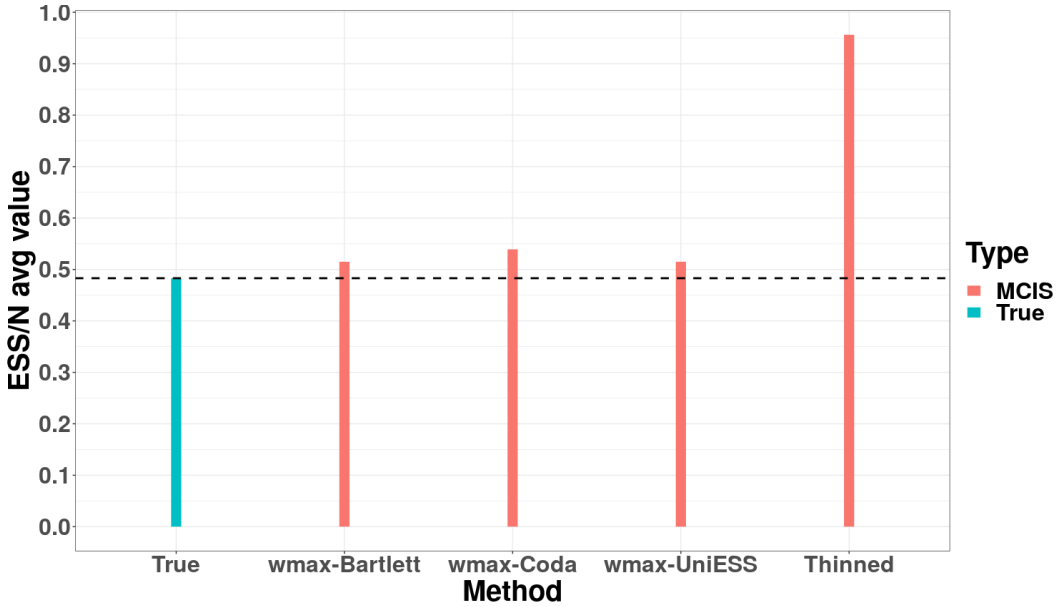

Figure 2.8: Comparison of possible combinations of MCMC and IS estimators within the MCIS estimator.

2.3.2 Modified Adaptive Integration Approach for MMHMC (s-MAIA)

2.3.2.1 Objective

We now take on the open problem left in section 2.2.5 related to adaptive integration in MMHMC. There, in the closing remarks, we identified the issues to be addressed while developing a statistical version of the MAIA methodology. In order to correctly handle adaptivity in a statistical framework, we need to develop a method that, for a given model and stepsize Δt, consistently provides the optimal choice of parameters - b or (a,b) - for r-stage integrators ($r = 2, 3$) in terms of the best conservation of modified energy for harmonic forces/Gaussian targets. To achieve it, we follow the procedure proposed in Akhmatskaya et al. (2017) for finding optimal values b^r_{opt} as

$$b^r_{opt} = \underset{0<b\leq 1/2r}{\arg\min} \max_{0<h<\bar{h}} \tilde{\rho}_r(h, b), \tag{2.44}$$

where $\tilde{\rho}_r(h, b)$ is the tight upper bound for expected modified energy error with respect to modified density $\tilde{\pi}$, and $\bar{h}$ is a nondimensional counterpart of the stepsize Δt. We use b instead of Γ from 2.29, as a is fully determined by b through

$$a = \frac{2b - 1}{4(3b - 1)}, \tag{2.45}$$

to guarantee good stability properties of a 3-stage integrator (Campos and Sanz-Serna, 2017). Crucially, since h in 2.44 is dimensionless, it is possible to tabulate once for each r the optimal integration coefficients b^r_{opt} at small increments of h to avoid the extra computational effort due to minmax operations in 2.44. Such tables then can be reused for any model and any simulation of interest. For this, the knowledge of the tight upper bounds of expected modified error with respect to modified density for each r (2,3) is required. We address this issue in the next section.

Besides, the methodology should include a procedure for nondimensionalising (or *scaling*) the stepsize Δt (which is problem-dependent and has the units given by the problem) into a dimensionless stepsize $\bar{h}$.

The AIA (Fernández-Pendás et al., 2016) and MAIA (Akhmatskaya et al., 2017) methodologies, developed for molecular simulations where the frequencies do have a physical interpretation, resolve these issue by using a safety factor, common for the whole 2-stage family, to avoid both linear and nonlinear resonances (Schlick et al., 1998). The overall idea is to write

$$h = \Delta t \cdot \tilde{S} \cdot \tilde{\omega}, \qquad \tilde{\omega} = \max_{1\leq j\leq D} \omega_j, \tag{2.46}$$

where $\tilde{S}$ is a safety factor closely related to the stability limits on $\omega\Delta t$ discussed in Schlick et al. (1998) and $\tilde{\omega}$ is the highest angular frequency. Such a scaling approach is generally possible in molecular simulations but becomes a hard problem to overcome when applied to computational statistics where frequencies have no immediate interpretations and might be costly to obtain. To find a reasonable adimensionalization procedure in the general setting of statistics, $\tilde{S}$ needs to be adapted to the underlying system, not fixed. We explore in depth the challenges of this problem and offer solutions to it in 2.3.2.3.

2.3.2.2 Upper bound for expected modified energy error with respect to modified density

We begin by taking an assumption that gives the angular frequencies a reasonable interpretation for statistical problems. Let us assume that the true distribution of the underlying system can be approximated by a multivariate Gaussian model. This is analogous to a system of D coupled harmonic oscillators with angular frequencies $\omega_1, \omega_2, ..., \omega_D$. Under this assumption, the angular frequencies

of the system are linked to the Hessian $U_{\boldsymbol{\theta},\boldsymbol{\theta}}(\boldsymbol{\theta})$ of the potential function and can be eventually estimated during a simulation.

Under this Gaussian assumption, we can describe the expected error of the modified Hamiltonian with respect to modified density $\tilde{\pi}$, $\mathbb{E}_{\tilde{\pi}}[\Delta\tilde{H}^{[k]}]$, as a sum of modified energy errors of D one-dimensional systems (Akhmatskaya et al., 2017)

$$0 \leq \mathbb{E}_{\tilde{\pi}}[\Delta\tilde{H}^{[k]}] \leq \sum_{j=1}^{D} \tilde{\rho}_r(h_j, b_j), \tag{2.47}$$

where $\tilde{\rho}_r(h_j, b_j)$ is the tight upper bound for one-dimensional expected modified energy error with respect to the modified density for a given integrator at the dimensionless stepsize $h_j = \Delta t \cdot S \cdot \omega_j$ associated to the frequency ω_j and a safety factor $S > 0$, independent of j, for all possible stepsizes h. For many models, the computation of the Hessian of $U(\boldsymbol{\theta})$ and its eigenvalues with sufficient accuracy can be very expensive or even not feasible. So, we avoid determining all the frequencies by making the additional assumption that we can approximate each term on the right-hand side of 2.47 by inserting one and the same value for h into each individual summand

$$\sum_{j=1}^{D} \tilde{\rho}_r(h_j, b_j) = D \cdot \tilde{\rho}_r(h, b), \qquad h = \Delta t \cdot \tilde{S} \cdot \max_{1 \leq j \leq D} \omega_j = \Delta t \cdot \tilde{S} \cdot \tilde{\omega}. \tag{2.48}$$

Note that in fact, we only need to determine the product $\tilde{S}\tilde{\omega}$ to render the stepsize dimensionless. This effectively allows us to develop a scaling formulation that avoids the calculation of frequencies, as we show in section 2.3.2.3.

Continuing with the Gaussian assumption, we need a formula linking the product $\tilde{S}\tilde{\omega}$ for the tight upper bound for the expected modified energy error with respect to modified density for a one-dimensional harmonic oscillator, i.e., $\tilde{\rho}_r(h, b)$. To simplify the analysis, but without any loss of generality, we set the mass matrix M to the identity and consider the 4^{th}-order truncated modified Hamiltonian of the form

$$\tilde{H}^{[4]}(\theta, p) = \frac{1}{2}\left(1 + 2h^2 c_{22}\right)\theta^2 + \frac{1}{2}\left(1 + 2h^2 c_{21}\right)p^2 \equiv \frac{1}{2}D_{h,1}\theta^2 + \frac{1}{2}D_{h,2}p^2. \tag{2.49}$$

The coefficients $c_{i,j}$ are listed in Radivojević et al. (2018) alongside the expression for the expected modified energy error with respect to modified density $\tilde{\pi}$ after L steps, which was derived in Akhmatskaya et al. (2017) and can be written as

$$\mathbb{E}_{\tilde{\pi}}[\Delta\tilde{H}^{[4]}] = s^2 \frac{(D_h B_h + C_h)^2}{2D_h(1 - A_h^2)}, \tag{2.50}$$

where $s = \sin(L \arccos a_h)$, meaning that $s^2 \leq 1$, which sets the upper bound

$$\tilde{\rho}_r(h, b) = \frac{(D_h B_h + C_h)^2}{2D_h(1 - A_h^2)}, \tag{2.51}$$

with $D_h = D_{h,1}/D_{h,2}$ extracted from the corresponding terms in 2.49. The rest of the terms needed to construct $\tilde{\rho}_r(h, b)$ for modified Hamiltonians take the following forms for the 2-stage splitting integrators

$$
\begin{aligned}
A_h &= \frac{h^4}{4} b(1-2b) - \frac{h^2}{2} + 1, \\
B_h &= -\frac{h^3}{4}(1-2b) + h, \\
C_h &= -\frac{h^5}{4} b^2(1-2b) + h^3 b(1-b) - h.
\end{aligned}
\tag{2.52}
$$

Combining 2.51 and 2.52 one gets an expression for 2-stages schemes

$$
\tilde{\rho}_2(h,b) = \frac{h^8(b(12+4b(6b-5)+b(1+4b(3b-2))h^2)-2)^2}{4(2-bh^2)(4+(2b-1)h^2(2+b(2b+1)h^2)(12+(6b-1)h^2)(6+(1+6(b-1)b)h^2)}
$$

as shown in Akhmatskaya et al. (2017).

Similarly, it is also possible to derive A_h, B_h and C_h for the 3-stage splitting integrators

$$
\begin{aligned}
A_h &= \frac{h^6}{4} a^2(2a-1)b(1-2b)^2 + \frac{h^4}{4} a(1-4b^2-2a(1-2b)) - \frac{h^2}{2} + 1, \\
B_h &= \frac{h^5}{4} a^2(1-2a)(1-2b)^2 - h^3 a(1-a)(1-2b) + h, \\
C_h &= \frac{h^7}{4} a^2(1-2a)b^2(1-2b)^2 + \frac{h^5}{2} a(2a(1-b)-1)b(1-2b) + \frac{h^3}{4}(1-2a(1-2b)^2) - h.
\end{aligned}
\tag{2.53}
$$

And $\tilde{\rho}_3$ can be obtained by combining 2.51 and 2.53.

The families of integrators derived using 2.44 are stable for sets of Γ coefficients resulting in $|A_h| < 1$ and $D_h > 0$ (Akhmatskaya et al., 2017). We analyse the behaviour of $\tilde{\rho}_r$ for $r = 2, 3$ in section 2.3.2.5.

2.3.2.3 Nondimensionalization of the stepsize

The quantities $\tilde{\omega}$ and $\tilde{S}$ in 2.46 are fully determined by the underlying system and are independent of the selected Δt, L, and the employed integrator. In order to find the product $\tilde{S}\tilde{\omega}$, we resort to the Verlet integrator with $L = 1$, for which a simple closed form expression for the expected modified energy with respect to modified density can be obtained. We recall that the VV can be recovered from 2.27 by letting $b \to 1/2$. We denote the expected modified energy error under these parameter values by $\tilde{\rho}_{VV}$ and using 2.50 obtain

$$
\tilde{\rho}_{VV}(h) = \frac{h^{10}}{16(72+6h^2-h^4)} = h^{10}\varepsilon(h) \qquad 0 \le h \le 2, \tag{2.54}
$$

with

$$
\varepsilon(h) = 1/\left(16(72+6h^2-h^4)\right). \tag{2.55}
$$

For a fixed stepsize Δt_{VV} let us denote

$$
h_j = \Delta t_{VV} S \omega_j \qquad \text{and} \tag{2.56}
$$

$$
h_{VV} = \Delta t_{VV} \tilde{S}\tilde{\omega}. \tag{2.57}
$$

For one integration step of size Δt_{VV} with the VV integrator, 2.48 is equivalent to

$$\sum_{j=1}^{D} \tilde{\rho}_{VV}(h_j) = D \cdot \tilde{\rho}_{VV}(h_{VV}) \overset{(2.54)}{\Leftrightarrow} \sum_{j=1}^{D} h_j^{10} \varepsilon(h_j) = D \cdot h_{VV}^{10} \varepsilon(h_{VV}).$$

Rearranging the above equation and inserting 2.56 yields

$$h_{VV} = \sqrt[10]{\frac{\sum_{j=1}^{D} h_j^{10} \varepsilon(h_j)}{\varepsilon(h_{VV}) D}} = \Delta t_{VV} S \sqrt[10]{\frac{\sum_{j=1}^{D} \omega_j^{10} \varepsilon(h_j)}{\varepsilon(h_{VV}) D}}.$$

And taking 2.57 we obtain an expression for the $\tilde{S}\tilde{\omega}$ product

$$\tilde{S}\tilde{\omega} = S \sqrt[10]{\frac{\sum_{j=1}^{D} \omega_j^{10} \varepsilon(h_j)}{\varepsilon(h_{VV}) D}}. \tag{2.58}$$

To estimate S, we continue our analysis of the expected modified energy error for one step of VV with the stepsize Δt_{VV}. We indicate this choice of parameters by the subscript VV in the modified energy error. We then have

$$\mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}_{VV}^{[4]}] = \sum_{j=1}^{D} \tilde{\rho}_{VV}(h_j) = \sum_{j=1}^{D} h_j^{10} \varepsilon(h_j) \overset{(2.56)}{=} \Delta t_{VV}^{10} S^{10} \sum_{j=1}^{D} \omega_j^{10} \varepsilon(h_j).$$

Or rearranging

$$S = \frac{1}{\Delta t_{VV}} \sqrt[10]{\frac{\mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}_{VV}^{[k]}]}{\sum_{j=1}^{D} \omega_j^{10} \varepsilon(h_j)}}.$$

By inserting this into the expression 2.58 for $\tilde{S}$ we finally obtain

$$h = \Delta t \tilde{S}\tilde{\omega} = \frac{\Delta t}{\Delta t_{VV}} \sqrt[10]{\frac{\mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}_{VV}^{[4]}]}{D}} \cdot \tilde{\varepsilon}_{VV}, \qquad \tilde{\varepsilon}_{VV} = 1/\sqrt[10]{\varepsilon(h_{VV})}. \tag{2.59}$$

The quantities on the right-hand side of 2.59 can usually be chosen, with the two exceptions $\mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}_{VV}^{[4]}]$ and $\tilde{\varepsilon}_{VV}$, which are still unknown. However, the former can be estimated from simulations and the latter can be approximated as well.

Recalling 2.55, a straightforward analysis shows that in the stability interval of the VV integrator – $h \in (0, 2)$ – the function on the right-hand side can be uniformly bounded by $1152 \leq \varepsilon^{-1} \leq 1296$. Hence, we can constrict the possible values of ε to

$$2.02369 < \tilde{\varepsilon}_{VV} = 1/\sqrt[10]{\varepsilon(h_{VV})} < 2.04768.$$

Given the small range of possible values for $\tilde{\varepsilon}_{VV}$, we propose to approximate $\tilde{\varepsilon}_{VV}$ with $\tilde{\varepsilon}_0$, such as

$$\tilde{\varepsilon}_0 = \mathbb{E}[\tilde{\varepsilon}_{VV}] = \frac{1}{2} \int_0^2 \sqrt[10]{16(72 + 6h^2 - h^4)} dh \approx 2.036629.$$

Then, from 2.59 one obtains

$$h = \frac{\Delta t}{\Delta t_{VV}} \sqrt[10]{\frac{\mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}_{VV}^{[4]}]}{D}} \cdot \tilde{\varepsilon}_0, \qquad \varepsilon_0 = 2.036629. \tag{2.60}$$

This leaves the value of $\mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}_{VV}^{[4]}]$ as the only unknown quantity in 2.60.

Computation of $\mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}_{VV}^{[4]}]$

An essential part of every HMC or MHMC algorithm is the burn-in stage, during which one simulates the chain for a fixed amount of HMC/MHMC iterations to make sure that it converges well enough to the sought distribution before sampling collection starts. The simulation data generated during burn-in is usually discarded. We propose to use the information from this stage for estimating $\mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}_{VV}^{[4]}]$.

We cannot take $\mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}_{VV}^{[4]}]$ directly from simulation data, as it is not dimensionless. However, we show that it is possible to link the expected acceptance rate to the expected energy error of the nondimensionalised modified Hamiltonian. Note that the acceptance rate is dimensionless by default, so that we can use its empirical values from burn-in.

An asymptotic relation for $D \to \infty$ between the expected acceptance rate and the dimensionless expected energy error of the true Hamiltonian was obtained in Theorem 2 in Calvo et al. (2021). It states that, under certain conditions, if the expected energy error of the true Hamiltonian converges to some $\mu \geq 0$ for $D \to \infty$, then

$$\lim_{D\to\infty} \mathbb{E}[a_D] = 2\Phi(-\sqrt{\mu}/2), \tag{2.61}$$

where a_D denotes the acceptance rate for the D-dimensional problem and Φ is the cumulative distribution function of the standard normal density. This relation continues to hold when we consider modified Hamiltonians and the expectation is taken with respect to the corresponding modified density. This fact is not obvious and we establish it for our modified setting in Theorem 2.1 further below. Since this statement does not depend on a specific splitting integrator in use we drop the index VV from now on in this section.

Let $\mu = \mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}^{[4]}]$. Assuming 2.61 holds for the modified density we take the corresponding $\tilde{\mu}$ in place of μ and we can use the asymptotic relation

$$2\Phi(-\sqrt{\tilde{\mu}/2}) = 1 - \frac{1}{2\sqrt{\pi}}\sqrt{\tilde{\mu}} + \mathcal{O}(\tilde{\mu}^{3/2}), \qquad \tilde{\mu} \to 0,$$

to obtain the useful approximation for expected modified energy error linking it to the simulation-based value a_D, which can be extracted during the burn-in phase, i.e.

$$\mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}^{[4]}] = \tilde{\mu} \approx 4\pi(1 - \mathbb{E}_{\tilde{\pi}}[a_D])^2, \qquad \tilde{\mu} \to 0. \tag{2.62}$$

A crucial ingredient for the proof of 2.61 is Lemma 2 in (Calvo et al., 2021), which here we adapt to our setting using modified Hamiltonians.

Lemma 2.1. *Consider the one-dimensional problem of a single harmonic oscillator, $H(\theta, p) = \theta^2/2 + p^2/2$. With $\tilde{\mu} = \mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}^{[4]}]$ and using a 2-stage splitting integrator of the form 2.27 we have*

$$\mathbb{E}_{\tilde{\pi}}[(\Delta\tilde{H}^{[4]}])^2] = 2\tilde{\mu} + 3\tilde{\mu}^2,$$
$$\mathbb{E}_{\tilde{\pi}}[(\Delta\tilde{H}^{[4]}])^3] = 18\tilde{\mu}^2 + 15\tilde{\mu}^3,$$
$$\mathbb{E}_{\tilde{\pi}}[(\Delta\tilde{H}^{[4]}])^4] = 36\tilde{\mu}^2 + 180\tilde{\mu}^3 + 105\tilde{\mu}^4.$$

Proof. We follow very closely the strategy of the proof of Lemma 2 in Calvo et al. (2021). We consider the modified energy error after L integration steps and express it as

$$2\Delta\tilde{H}^{[4]} = \theta^2 A + p^2 C + 2\theta p B, \tag{2.63}$$

for some coefficients A, B, C. Subsequently, we use an algebraic relation between these coefficients to simplify 2.63. Considering that θ and p are normally distributed, we can compute their moments and, thus, verify the claim.

Formulas for A, B, C can be found in Akhmatskaya et al. (2017); Radivojević et al. (2018) and are expressed as

$$A = s^2\left(\chi_h^{-2} D_{h,2} - D_{h,1}\right), \qquad B = s^2\left(\chi_h^2 D_{h,1} - D_{h,2}\right),$$
$$C = sc\left(\chi_h D_{h,1} - \chi_h^{-1} D_{h,2}\right),$$

for $s = \sin(L\arccos A_h)$, $c = \cos(L\arccos A_h)$, $\chi_h = B_h/s$. $D_{h,1}, D_{h,2}$ were defined in 2.49 and A_h and B_h are determined in 2.52, 2.53 for 2- and 3- stage integrators, respectively. The algebraic relation between the coefficients is given by

$$B^2 - AC = D_{h,2}A + D_{h,1}C.$$

Finally, since $\theta \sim \mathcal{N}(0, D_{h,1}^{-1})$ and $p \sim \mathcal{N}(0, D_{h,2}^{-1})$, we have for every positive integer k that $\mathbb{E}_{\tilde{\pi}}[\theta^{2k-1}] = \mathbb{E}_{\tilde{\pi}}[p^{2k-1}] = 0$, as well as

$$\mathbb{E}_{\tilde{\pi}}[\theta^{2k}] = \frac{m_k}{D_{h,1}^k} \quad \text{and} \quad \mathbb{E}_{\tilde{\pi}}[p^{2k}] = \frac{m_k}{D_{h,2}^k}, \quad \text{where} \quad m_k = \prod_{\ell=1}^{k}(2\ell - 1),$$

and, we get the expression for $\tilde{\mu}$ as

$$2\tilde{\mu} = 2\mathbb{E}_{\tilde{\pi}}[\Delta\tilde{H}^{[4]}] = \frac{A}{D_{h,1}} + \frac{C}{D_{h,2}},$$

which we can use to verify the claim, following the steps of Lemma 2 in Calvo et al. (2021). □

The remainder of the proof of 2.61 for our modified setting follows the proof of Theorem 2 in Calvo et al. (2021) verbatim. Nevertheless, we summarize the result in Theorem 2.1 below and present the main steps of the proof.

We consider a non-degenerate system of D harmonic oscillators with Hamiltonians $H_{D,j}$, $1 \leq j \leq D$, similarly as in multivariate Gaussian assumption leading to 2.47. In this case, the Hamiltonian of the entire system is given by $H_D = H_{D,1} + \cdots + H_{D,D}$. Since the numerical integration step commutes with a diagonalisation of the covariance matrix of these systems, it is safe to assume that the individual components are mutually independent (Bou-Rabee and Sanz-Serna, 2018).

Theorem 2.1. *Under the assumptions*

$$M_D = \max_{1\le j\le D} \mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}^{[4]}_{D,j}] \to 0 \qquad \text{for } D \to \infty$$

and

$$\mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}^{[4]}_{D}] = \sum_{j=1}^{D} \mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}^{[4]}_{D,j}] \to \tilde{\mu}, \qquad \text{for } 0 \le \tilde{\mu} < \infty,\ D \to \infty,$$

we have

(i) $(\Delta \tilde{H}^{[4]}_{D} - \tilde{\mu})/\sqrt{2\tilde{\mu}}$ *converges in distribution towards the standard normal distribution for* $D \to \infty$.

(ii) $\lim_{D\to\infty} \mathbb{E}_{\tilde{\pi}}[a_D] = 2\Phi(-\sqrt{\tilde{\mu}/2})$.

Proof. We denote $\tilde{\mu}_{D,j} := \mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}^{[4]}_{D,j}]$ and begin by showing

$$\lim_{D\to\infty} \sum_{j=1}^{D} \tilde{\mu}^k_{D,j} = 0, \qquad \text{for all integers } k \ge 2. \tag{2.64}$$

And indeed, we have

$$0 \le \sum_{j=1}^{D} \tilde{\mu}^k_{D,j} \le M_D^{k-1} \sum_{j=1}^{D} \tilde{\mu}_{D,j} = M_D^{k-1}\tilde{\mu} \to 0.$$

Let us now focus on (i). Since we assumed a diagonalised covariance matrix, we have by Lemma 2.1

$$\sigma_D^2 = \mathrm{Var}_{\tilde{\pi}}[\Delta \tilde{H}^{[4]}_{D}] = \sum_{j=1}^{D} \mathrm{Var}_{\tilde{\pi}}[\Delta \tilde{H}^{[4]}_{D,j}] = 2\sum_{j=1}^{D} \left(\tilde{\mu}_{D,j} + \tilde{\mu}^2_{D,j}\right) \to 2\tilde{\mu},$$

where we used 2.64 to compute the limit. With a view to applying the Central Limit Theorem with the Lyapunov condition for the 4th moment, we consider the expression

$$\frac{1}{\left(\sigma_D^2\right)^2} \sum_{j=1}^{D} \mathbb{E}_{\tilde{\pi}} \left[(\Delta \tilde{H}^{[4]}_{D,j} - \tilde{\mu}_{D,j})^4 \right].$$

The denominator σ_D^4 tends to $4\tilde{\mu}^2$, while the numerator converges to 0 as a result of Lemma 2.1 and 2.64. This concludes the proof of (i).

For the second claim, we use (i) to find that

$$\lim_{D\to\infty} \mathbb{E}_{\tilde{\pi}}[a_D] = \mathbb{E}[\min\{1, \exp(-(\sqrt{2\tilde{\mu}}Z + \tilde{\mu})\}], \qquad Z \sim \mathcal{N}(0,1),$$

and the result follows directly from Lemma 3 in Calvo et al. (2021).

□

We conclude with the summary of the nondimensionalization step proposed in this section. Given Δt in simulation units the dimensionless counterpart $\bar{h}$ can be found as

$$\bar{h} = \frac{\Delta t}{\Delta t_{VV}} \sqrt[10]{\frac{4\pi(1-a_D)^2}{D}}\,\tilde{\varepsilon}_0 \qquad \tilde{\varepsilon}_0 = 2.036629, \tag{2.65}$$

where a_D is the expected acceptance rate with respect to modified density, extracted from the burn-in stage of an MMHMC simulation performed with the Velocity Verlet at Δt_{VV} and $L = 1$.

2.3.2.4 Algorithm

In view of the methodology we just presented, an algorithm for s-MAIA can be implemented in a two-stage process. First, we resort to 2.65 to perform the nondimensionalization of the problem stepsize. In practice, this requires a_D and Δt_{VV}, which can be easily obtained during the burn-in stage of the simulation. On the other hand, it is necessary to compute the b^r_{opt} values that define the integrator for any possible stepsize in the stability interval. In the next section, we address the calculation of b^r_{opt}, which can be done once and for all and stored in a table. This results in a more efficient use of computational resources, as the costly computation of 2.44 is reduced to a search in a file with pre-tabulated values.

We present an algorithmic representation of the methodology for finding optimal integrator coefficients in Algorithm 5.

Input: Δt: Stepsize in simulation units (defined by the user),
Δt_{VV}: Estimated stepsize in simulation units for VV,
a_D: Estimated acceptance rate from burn-in stage,
r: Stage of the selected integrator.

Nondimensionalize the stepsize

$\Delta t\ \bar{h} \leftarrow \frac{\Delta t}{\Delta t_{VV}} \sqrt[10]{\frac{4\pi(1-a_D)^2}{D}}\,\tilde{\varepsilon}_0$ where $\tilde{\varepsilon}_0 = 2.036629$

Check stability of the r-stage integrator at Δt
if $\bar{h} > 2r$ **then**
`// The r-stage integrator is unstable for resulting` $\bar{h}$
Abort integration. Choose another Δt
else
Search in table for the integrator parameters corresponding to $\bar{h}$
Find $b^r_{opt} = \underset{0<b\leq 1/2r}{\arg\min} \max_{0<h<\bar{h}} \rho_r(h, b)$.
end

Result: Integrator coefficient of the optimal r-stage splitting integrator to be used in the production stage of an MMHMC simulation.

Algorithm 5: Adaptive selection of optimal integration parameters for MMHMC.

2.3.2.5 Implementation

The algorithm we propose has been implemented in our in-house software, HaiCS, using the following procedure.

The objective is to run an MMHMC simulation of a given D-dimensional problem using an optimal r-stage integrator with a stepsize Δt and a trajectory length L for N_{prod} production iterations. For reducing a statistical error of averages, such simulations are run for n_{ch} chains in parallel. For a burn-in stage, which must precede any production stage, the settings are: a number of iterations N_{burn}, the Velocity Verlet integrator with $\Delta t_{VV} \approx 1/D$ and $L = 1$. The initial stepsize Δt_{VV} is optimised in the course of a burn-in simulation using the bisection algorithm. The aim is to achieve a

high enough target acceptance rate a_{burn} that can guarantee the good quality of averages obtained during a burn-in stage. We recommend choosing a_{burn} within the interval (98.5%, 99.5%).

Once the optimal Δt_{VV} is found, we proceed with the rest of the burn-in phase. For each chain, we simulate N_{burn} MMHMC iterations using one step of the one-stage VV integrator with the stepsize Δt_{VV}. The empirical mean of the obtained n_{ch} acceptance rates, is then used as an estimator for $\mathbb{E}[a_D]$ associated to Δt_{VV} to realize the scaling. With the resulting nondimensionalized value of h, 2.65, we can simply find the corresponding optimal parameter for the integrator in the pre-calculated tables. Next, we describe the tabulation procedure, which though is not part of the s-MAIA algorithm, is a necessary condition for its realization.

For tabulating the values of b^r_{opt} we begin by constructing a grid with a stepsize of $\Delta\bar{h} = 10^{-4}$ within the stability interval (0, $2r$) and $\Delta b = 10^{-6}$, with b in the interval (0, $1/2r$). For each $\bar{h}_i$ within the grid we find $b^r_{opt,i}$ as

$$b^r_{opt,i} = \underset{0<b\leq 1/2r}{\arg\min} \max_{0<h<\bar{h}_i} \tilde{\rho}_r(h, b), \tag{2.66}$$

to produce the tables for both the 2- and 3- stage cases($r = 2$ and $r = 3$, respectively). In the 3-stage case, a is fully determined by b as the pair must satisfy the relation 2.45 for the resulting integrator to have good stability properties.

The values obtained for b using such an approach can be seen in Figure 2.9 along with the values for b that define the M-BCSS, M-ME and VV integrators.

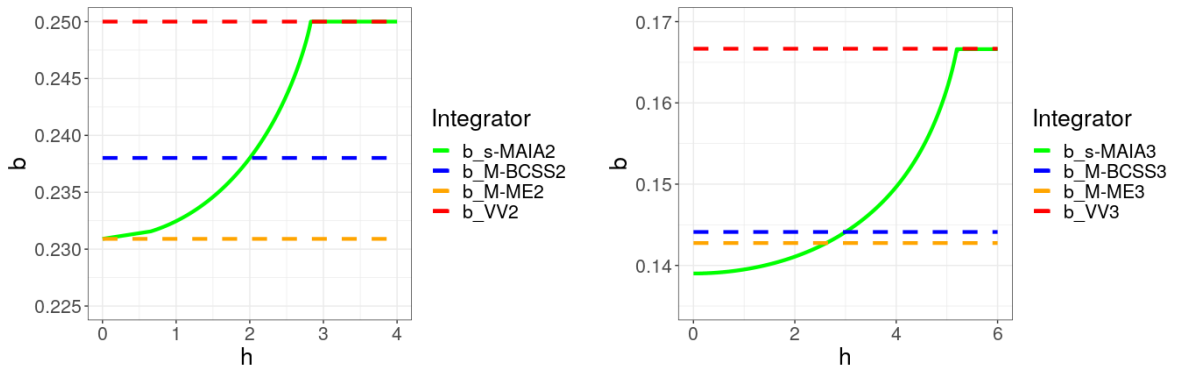

Figure 2.9: Values of the coefficient b for 2-stage integrators (left) and 3-stage integrators (right). 2-, 3- s-MAIA parameters vs h are shown in green. Stepsize values for h are normalized with respect to the number of stages to fit in the VV stability interval.

The corresponding behaviour of the function $\rho_r(h, b)$ in the VV stability interval can be seen in Figure 2.10.

At first glance, Figure 2.9 presents a reasonable set of values for b throughout the stability interval. Indeed, b parameters for small h obtained with s-MAIA and M-ME (optimised for small h) are close, whereas they are identical for s-MAIA and M-BCSS (optimised at h=r) in the centre of the stability interval as well as for s-MAIA and VV (possesses the longest stability limit) for long stepsizes. However, a closer look at Figure 2.10 reveals that $\rho_r^{\text{s-MAIA}}$ offers the minimum value for most of the interval except for the short spikes down obtained using b_{M-ME} and b_{M-BCSS} right before the half stability limit. The lack of such picks in the $\rho_r^{\text{s-MAIA}}$ function is likely the result of highly oscillating behaviour of the $\rho_r^{\text{s-MAIA}}$ function, which cannot be caught with the chosen grid size. In principle, it can be overcome by decreasing a grid size for the tabulation. However, this implies a significant increase of computational burden. We choose the alternative solution for tabulation and restrict the lower bound for the value of b to be b_{M-ME} (McLachlan and Atela, 1992; Predescu et al., 2012).

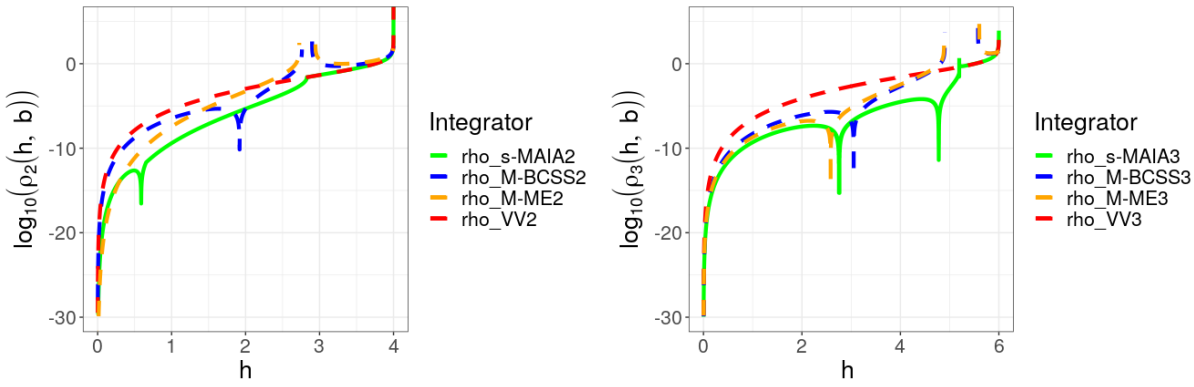

Figure 2.10: Values of the $\rho_r(h, \Gamma)$ function (in $\log_{10}$ scale) for the 2-stage integrators (left) and 3-stage integrators (right). 2-, 3- s-MAIA parameters vs h are shown in green. Stepsize values for h are normalized with respect to the number of stages to fit in the VV stability interval.

With this in mind, we proceed with *patching* the initially obtained values of b to account for these factors. The search is repeated for values of $b \in (b_{ME}, b_{VV})$, to add patches for those points where $\rho_r^{\text{s-MAIA}}(h, \{b_{ME}, b_{BCSS}\}) < \rho_r^{\text{s-MAIA}}(h, b_{unpatched})$. The resulting values of b are displayed in Figure 2.11.

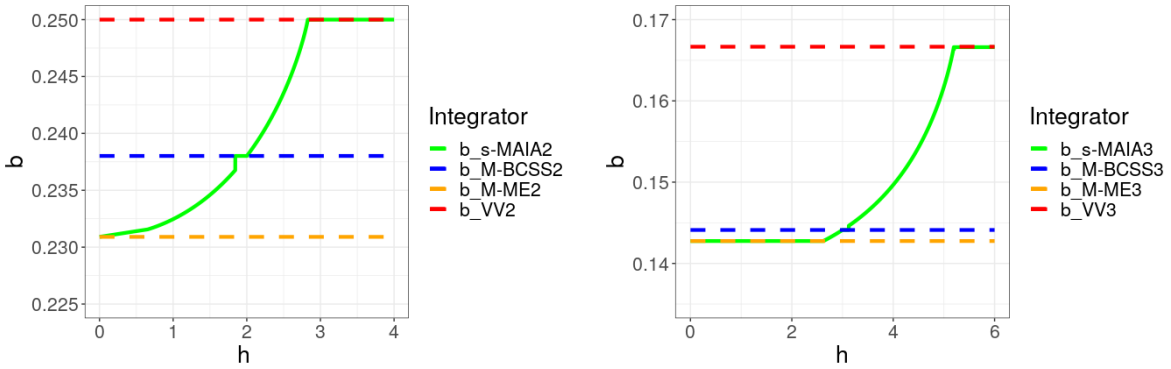

Figure 2.11: Values of the coefficient b after correction for 2-stage integrators (left) and 3-stage integrators (right). 2-, 3- s-MAIA parameters vs h are shown in green. Stepsize values for h are normalized with respect to the number of stages to fit in the VV stability interval.

And the function $\rho_r(h, \Gamma)$ for various 2-, 3-stage integrators in the VV stability interval can be seen in Figure 2.12. Clearly, as integration parameters b as functions ρ_r^{s-MAIA} presented in Figures 2.11 and 2.12, respectively, demonstrate the expected behaviour.

The tabulation of b_{opt}^r values offers a nice solution, not only due to the savings in computational cost of the method, but also because of the flexibility it provides. Randomization of the stepsize at each MC iteration is a common feature in many MCMC methods leading to enhancing performance (Neal et al., 2011). The pre-tabulation of values in s-MAIA, creates the opportunity to, first, generate in advance a vector of randomized stepsizes h_i — or Δt_i, if running in simulation units, — and then find the corresponding $b_{opt,i}$ straight away from the corresponding table. These pairs of randomized stepsizes and corresponding b_{opt} can be passed directly to the production stage of a simulation. This feature is present, for example, in s-AIA (Nagar et al., 2024), where adaptivity to randomization schemes has proven to be a good addition for HMC simulations. However, we do not show such a case in this dissertation. All tests presented next for s-MAIA were run with MMHMC, where randomization of the stepsize does not appear to offer any improvements. On the contrary, it potentially can lead to increasing the variability of weights, which implies the degradation of the MMHMC performance.

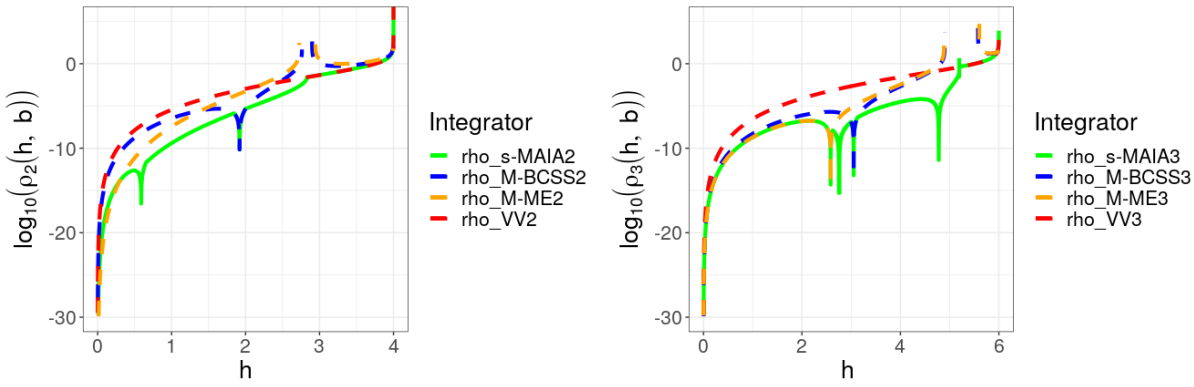

Figure 2.12: Values of the $\rho_r(h, \Gamma)$ function (in $\log_{10}$ scale) using patched values of b for the 2-stage integrators (left) and 3-stage integrators (right). 2-, 3- s-MAIA parameters vs h are shown in green. Stepsize values for h are normalized with respect to the number of stages to fit in the VV stability interval.

2.3.2.6 Performance evaluation: Comparison with state-of-the-art modified integrators

In this section, we evaluate the performance of the s-MAIA methodology by comparing it to other known integrators previously presented. The list includes the standard Velocity Verlet (VV), the modified Minimum Error (M-ME) integrator and the modified Blanes, Casas & Sanz-Serna (M-BCSS) (Radivojević et al., 2018), as well as the modified adaptive integration approach MAIA (Akhmatskaya et al., 2017). For comparison, we use 2 standard benchmarks and the evaluation metrics summarized in Table 2.5, where ESS uses the formulation 2.43. Crucially, ESS and MCSE are normalized to computational effort, for fair comparison. The most expensive calculations at each iteration are the evaluations of the gradient in the integration of Hamiltonian equations. Hence, we normalize these metrics using the number of gradient evaluations, rL, where r is the number of stages of the integrator and L is the average length of the trajectory (number of steps taken to perform the integration).

First, we show our methodology in action compared to other modified integrators using the 1000-dimensional multivariate Gaussian model, previously employed in the ESS metric selection experiment in section 2.3.1. The evolution of MMHMC performance in terms of the chosen metrics was monitored for 20 stepsizes within the stability interval, expressed in simulation units. Such an interval was obtained through the dimensionalization of the theoretical stability interval for the r-stage Velocity Verlet, i.e. $(0, 2r)$. The simulations were run with the randomized trajectory length $L \sim U(1, 2D-1)$ and the momentum noise parameter $\varphi \sim U(0, 0.2)$ which resulted in the mean values of L and φ to be 1000 and 0.1, respectively. The former is equal to the dimension of the problem, the reasonable choice for an HMC-based method.

The green line in Figures 2.13 and 2.14 shows the results obtained using MMHMC combined with s-MAIA. Across all metrics measured, MMHMC with s-MAIA offers the best performance overall, topping all metrics for the majority of values throughout the whole stability interval. The results are displayed separately for the 2- and 3-stage, as the 3-stage methodologies consistently result in better performance compared with the 2-stage counterparts for all integrators.

The second benchmark tested here is a Bayesian logistic regression model. The choice was motivated by the availability of relevant toy datasets and the potential applicability of binary classification, such as with the BLR models, in a biomedical context. Next, we present the formulation of the BLR model used in our tests.

For a set of given data $\{x_k, y_k\}_{k=1}^K$ where x_k is a vector of $D-1$ covariates and $y_k \in \{0, 1\}$ are the binary responses, the probability p_k of a particular binary outcome is given by the logit function:

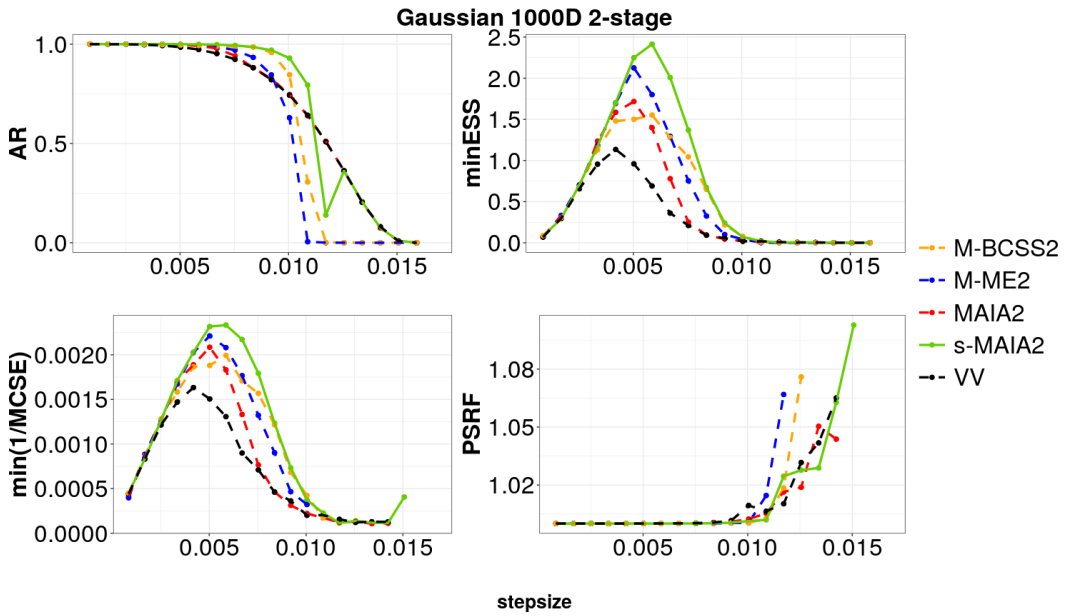

Figure 2.13: Performance of s-MAIA2 (green) against other 2-stage modified integrators combined with MMHMC and tested on the 1000-dimensional multivariate Gaussian model.

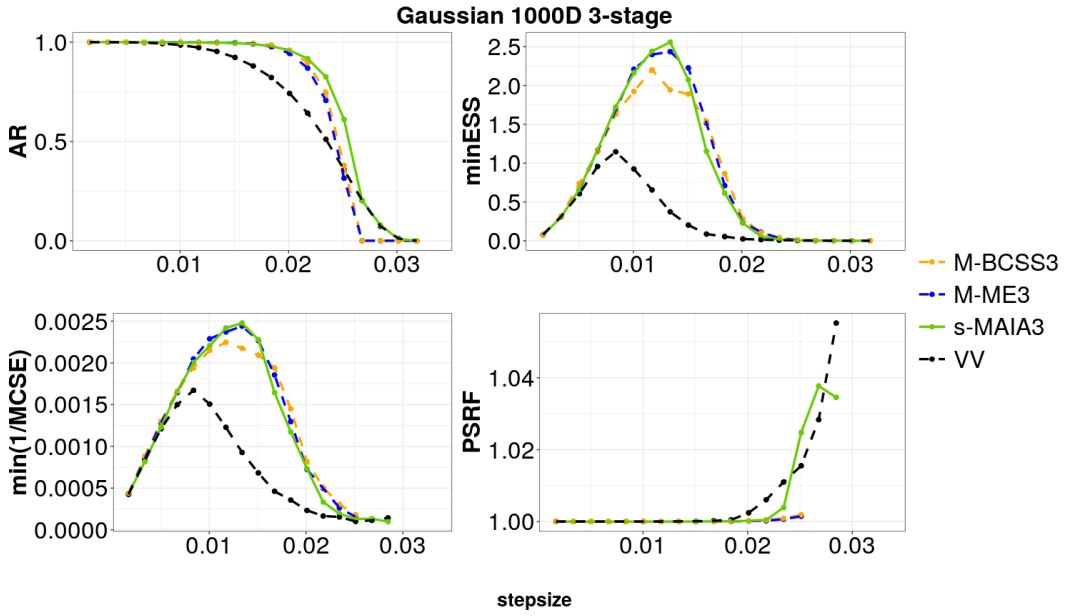

Figure 2.14: Performance of s-MAIA3 (green) against other 3-stage modified integrators combined with MMHMC and tested on the 1000-dimensional multivariate Gaussian model.

$$logit(p_k) = \theta_0 + \theta_1 x_{1,k} + \cdots + \theta_{D-1} x_{D-1,k},$$

which results in a likelihood function

$$L(y_i|\theta, x_i) = \prod_{i=1}^{N} \left[\left(\frac{e^{\theta x_i}}{1+e^{\theta x_i}} \right)^{y_i} \left(1 - \frac{e^{\theta x_i}}{1+e^{\theta x_i}} \right)^{1-y_i} \right],$$

where $\theta \in \mathbb{R}^D$ is the regression coefficient vector, for which we choose a Normal prior $\theta \sim \mathcal{N}(0, \sigma\mathbb{I})$. We performed the BLR benchmark tests for logistic regression, using the German credit dataset

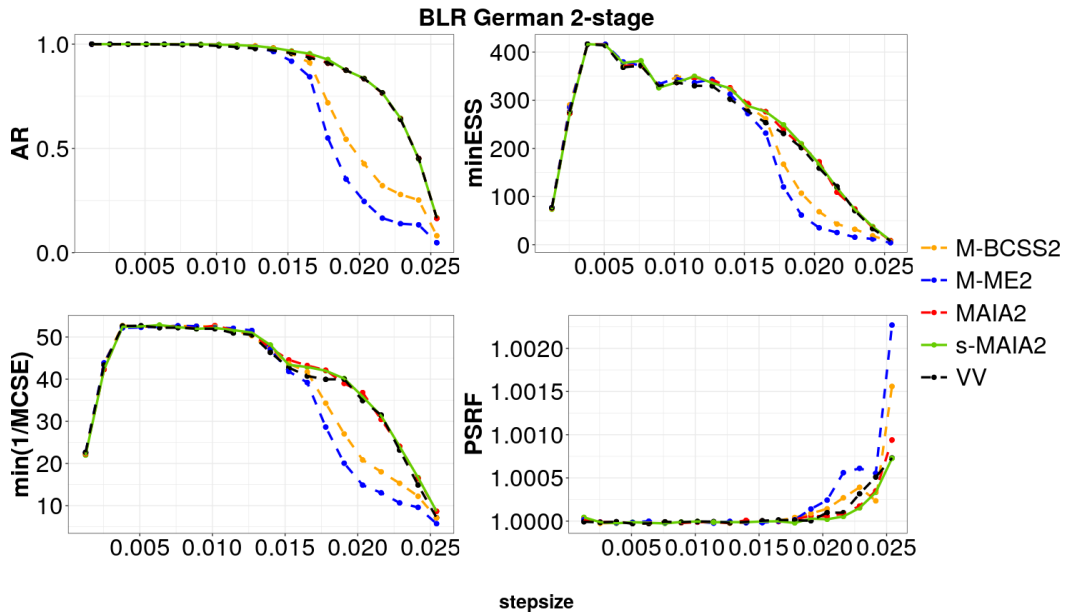

Figure 2.15: Performance of s-MAIA2 (green) against other 2-stage modified integrators when combined with MMHMC and tested on the BLR German model.

(Hofmann, 1994). The dataset contains 1000 observations and 25 features and is common for testing HMC methodologies in landmark papers, such as Hoffman et al. (2014).

As in the Gaussian case, we measured the performance of MMHMC combined with various 2- and 3-stage integrators using the selected performance metrics for 20 stepsizes within the theoretical stability limit for the r-stage Velocity Verlet. The mean value of the randomized trajectory length was selected again to be equal to the dimension of the problem, with the same type of uniform randomization as before. For the randomized momentum noise parameter, we selected a value of $\varphi = 0.9$.

In the case of BLR, we can see yet another advantage of the s-MAIA methodology over fixed-parameters integrators. As before, the method is performing at the top of each metric, although this time, other integrators can get to similar results. The key fact, however, is that though some integrators demonstrate high performance with some metrics, they are uncapable of maintaining a top performance across all studied metrics. This can be seen clearly with the VV example in Figure 2.16. Although it offers a great response in terms of AR and PSRF, its results for normalized minESS and MCSE^{-1} are visibly poorer than the ones observed with other 3-stage integrators, included s-MAIA. On the contrary, s-MAIA adapts well through the interval to maintain its top performance with all tested metrics.

2.3.2.7 Performance evaluation: Optimal MMHMC vs Optimal HMC

The advantages of MHMC methods over HMC had already been documented and studied in Akhmatskaya et al. (2017); Akhmatskaya and Reich (2008); Radivojević and Akhmatskaya (2020) using previously known integrators. Here, we show the advantages of MMHMC combined with our novel adaptive integrators s-MAIA over standard HMC with the top adaptive integrators proposed in Nagar et al. (2024). We again resort to the same multivariate Gaussian benchmark, for which we observed more noticeable positive impact of s-MAIA on the overall performance of MMHMC, than in the BLR model. We used the same simulation settings for both MMHMC and HMC as before. The only difference is that all HMC simulations were run with the s-AIA integrators (Nagar et al., 2024).

Figure 2.17 displays the results for HMC and MMHMC combined with the 2- and 3-stage versions

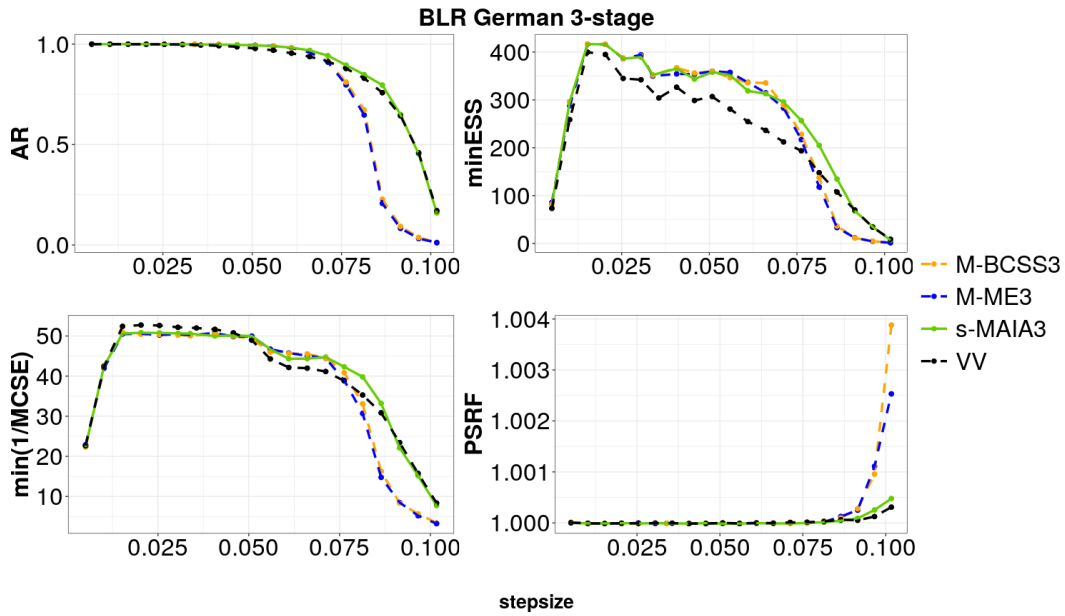

Figure 2.16: Performance of s-MAIA3 (green) against other 3-stage modified integrators when combined with MMHMC and tested on the BLR German model.

of appropriate integrators. The stepsize was normalized to be in the 1-stage Velocity Verlet interval. The figure clearly demonstrates the advantages of MMHMC with s-MAIA over the standard HMC, even with its best integration option. Moreover, the superiority of 3-stage adaptive integrators over 2-stage counterparts for both MMHMC and HMC is also confirmed. The major difference in behaviour of MMHMC and HMC, apart from the absolute values of the performance metrics, is the stepsize at which the top performance was achieved. Compared with s-MAIA, the s-AIA methodology seems to be able to guarantee the best results at longer stepsizes within the stability interval. However, this could be caused by its capability for more accurate adimensionalization and, consequently, a better approach to selecting an optimal stepsize. The HMC version, s-AIA, can easily work with the real frequencies of the system, which helps to obtain a more accurate dimensionless stepsize (Nagar et al., 2024). In the modified case, like MMHMC, the use of an importance distribution alters these values, making the approximation less accurate for this case. However, Figure 2.17 clearly indicates that the gain in overall performance compensates for this loss in the exact location of the stability interval.

2.3.3 Applications

In the upcoming section, we present three studies where we apply MMHMC with our s-MAIA methodology to different clinical datasets. These demonstrate three major areas of application of Bayesian analysis where s-MAIA can be used to enhance the result.

We first investigate the effect of s-MAIA on the performance of MMHMC when applied to a real clinical dataset. Then we move to the case where, using s-MAIA helps to effectively test the influence of priors in the separation problem in the SOX2 dataset.

Finally, we perform Bayesian inference with MMHMC (s-MAIA) of the BLR model using the SOX2 dataset and provide the posterior analysis with the focus on significance and probability of direction.

For the review of many other possible applications of Bayesian statistics to biomedical models, we refer the reader to Lopes et al. (2007).

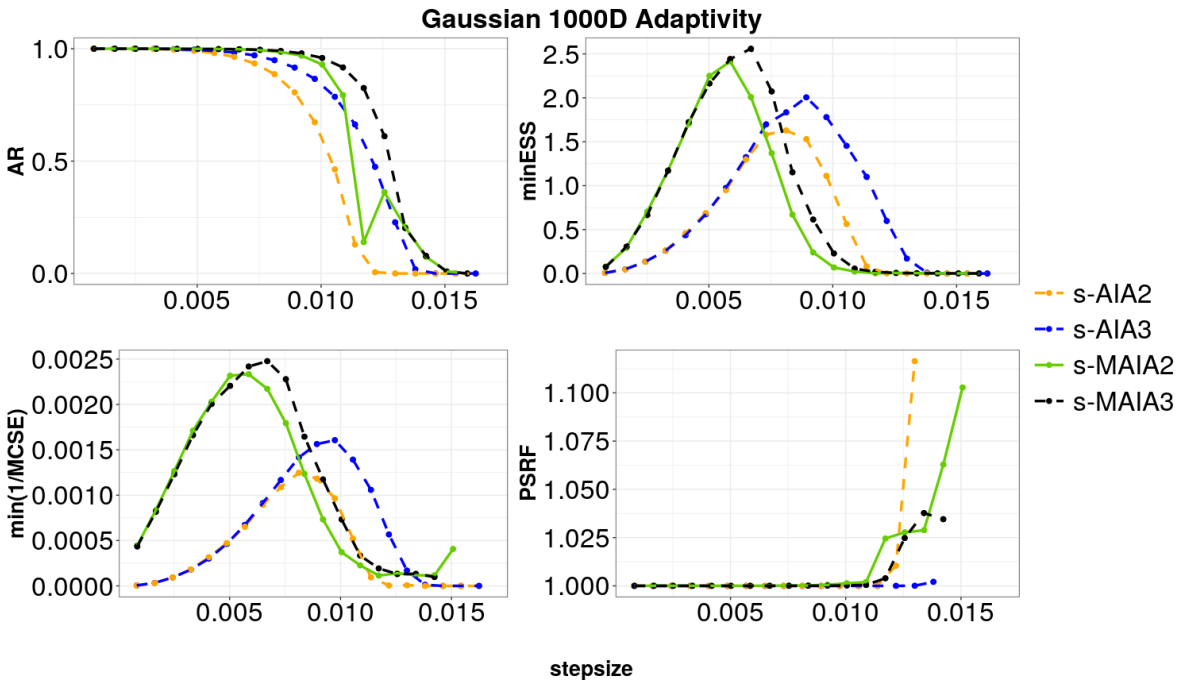

Figure 2.17: Performance comparison of HMC and MMHMC combined with 2-, 3-stage adaptive integrators s-AIA (in yellow and blue) and s-MAIA (in green and black), respectively, using the Multivariate Gaussian model

2.3.3.1 Wisconsin Breast Cancer dataset — A biomedical benchmark

The Wisconsin Breast Cancer (WBC) diagnostic dataset (Wolberg and Street, 1995) is a widely used dataset primarily focused on breast cancer diagnosis. It encompasses various measurements concerning morphological characteristics of breast tumours such as radius, area or smoothness, which can be found in the UCI Machine Learning Repository. The dataset is commonly employed in machine learning for the evaluation of predictive models for breast cancer classification.

Each tumour in the dataset is labelled as either malignant or benign, making it an ideal dataset for the BLR model presented in the previous section and tested on the German dataset. This time, our objective is to put the model in the context of breast cancer by combining it with the WBC dataset and use it for testing the sampling efficiency as well as applicability in biomedical research of the MMHMC method coupled with s-MAIA in Bayesian inference. The tests are performed in the similar manner and with the same settings as in section 2.3.2.6.

Figures 2.18 and 2.19 show similar impacts of the tested integrators on the MMHMC performance to those presented in Figures 2.15 and 2.16 for the BLR German dataset. The standard VV does perform better than M-BCSS and M-ME throughout the range of stepsizes studied. Meanwhile, s-MAIA proves to be most reliable around the half stability limit, while replicating the good behaviour of VV whenever this is the best option. Thus, we can confirm that the s-MAIA integrators remain the best integrator's choice for achieving reliable sampling efficiency of MMHMC in Bayesian inference of biomedical models.

2.3.3.2 SOX2 Clinical dataset — Posterior analyses

Separation and the influence of priors

Having asserted that s-MAIA offers the best performance overall in the benchmarks used, we resort to the dataset presented in section 2.1.1 to use MMHMC and the s-MAIA approach in practical scenarios involving different priors. The objective is to use the extreme case this dataset represents to learn

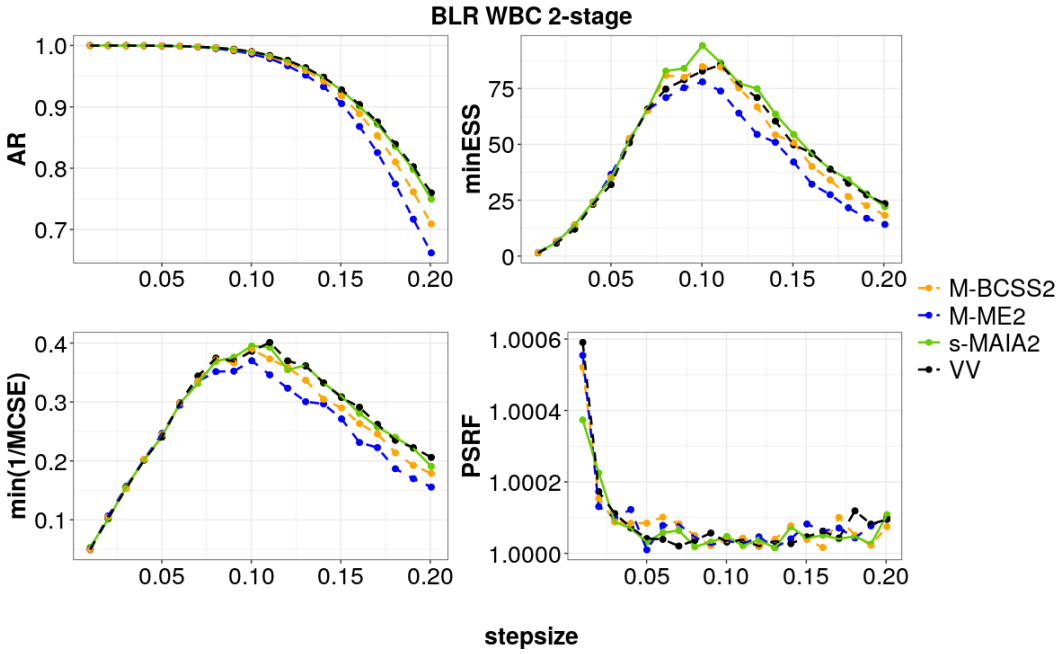

Figure 2.18: Performance of s-MAIA2 (green) against other 2-stage modified integrators when combined with MMHMC and tested on the BLR WBC model.

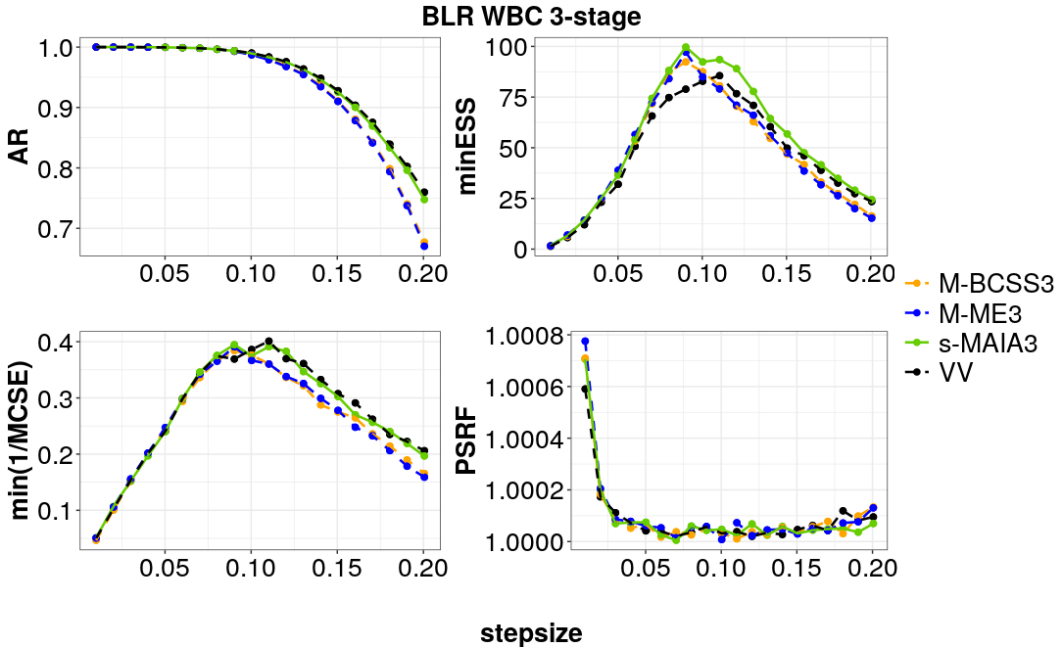

Figure 2.19: Performance of s-MAIA3 (green) against other 3-stage modified integrators when combined with MMHMC and tested on the BLR WBC model.

key information about the implications of a given prior for the shape of the posterior distribution.

We again plan to use a BLR model, but our focus now is on studying the parameters obtained from the MMHMC trajectories, i.e., the posterior distributions for each dimension in the model. We recall that the SOX2 variable is a *perfect predictor* for the response. In such a scenario, classical frequentist methods normally struggle due to their reliance on maximum likelihood estimators (MLE) to assign values to model parameters. For cases with separation, the likelihood function becomes extremely steep, making it challenging or even impossible to find a unique maximum. This results in classical parameter estimates with large standard errors or even infinite values, rendering the model

unreliable.

In particular, for the SOX2 Clinical dataset, the frequentist logistic regression model does not converge if the SOX2 variable is included. The frequentist framework does include some regularization techniques like Ridge (Hoerl and Kennard, 1970) or Lasso (Tibshirani, 1996) regressions. However, use of the penalty term in both Lasso and Ridge makes the calculation of standard errors for coefficient estimates a challenging task.

The Bayesian framework offers a natural solution to the separation problem in logistic regression, while at the same time naturally providing error estimates that can be directly extracted from the posterior distribution. When facing separation, carefully chosen informative priors can help regularize the estimation process. Incorporating prior information can make the model less sensitive to extreme observations and more robust in the face of separation. Bayesian methods also provide straightforward confidence intervals, which give a range of plausible values for the parameters, acknowledging the uncertainty inherent in statistical modelling while conveying information about the most likely values for each parameter.

In this context, we test the effect of three priors, an informative one $\mathcal{N}(0, 1)$, a weakly informative $\mathcal{N}(0, 2.5)$ - according to Gelman et al. (2008) - and a low informative $\mathcal{N}(0, 10)$. For all three cases we analyse the behaviour of all the gene-related features from Table 2.2, i.e., SOX2, Ki67, p53, HER2, PR and ER. Our goal is to assess the effect of using each type of prior on the location and shape of the posterior distribution of each parameter. The results can be seen next in Figure 2.20.

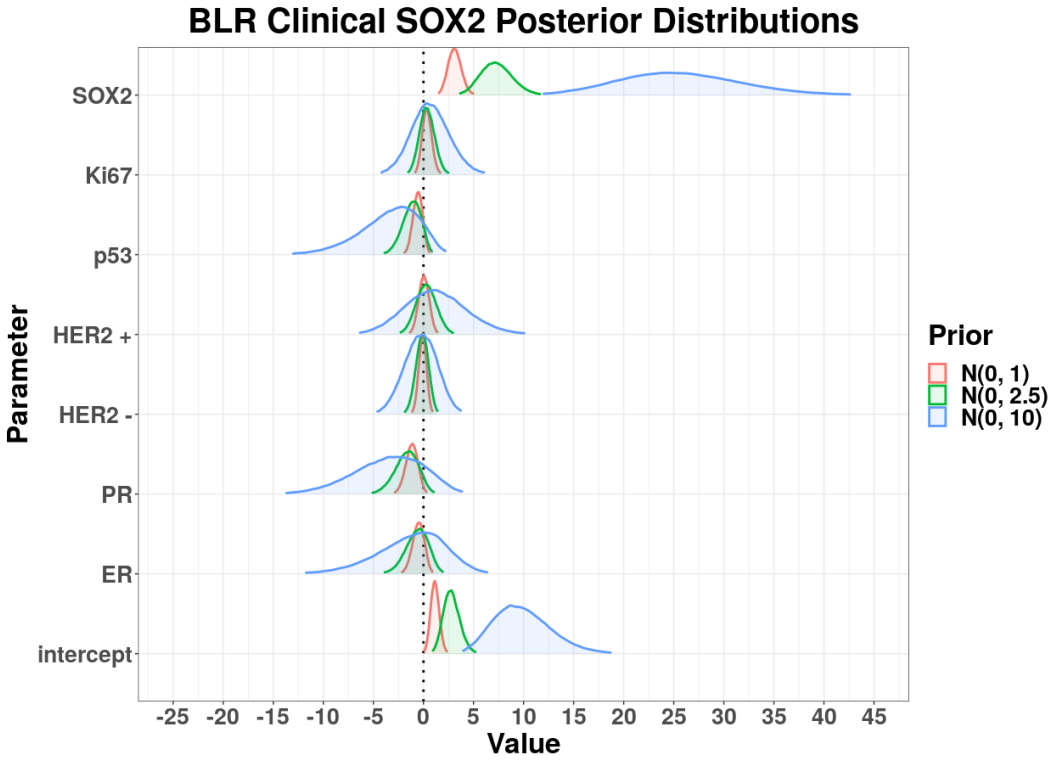

Figure 2.20: Posterior distributions for three priors studied for the SOX2 clinical dataset.

All three priors are able to deal with the separation effect introduced by the presence of the SOX2 variable, although the resulting posterior distributions differ greatly, up to the point where they barely overlap. However, the priors in this case were not based on *a priori* knowledge of the system, but rather represent three very different conditions. Since all priors tested address the separation problem, incorporating additional information for prior formulation should lead to notably more accurate estimates, especially when compared to the available frequentist alternatives. The width

of the priors used affects greatly all the variables including SOX2. We see a huge dispersion in the posteriors arising from the low informative prior, but much more accurate posterior distribution whenever some precise information is introduced in the model. In the next section, we will observe more carefully the behaviour of these posteriors around the origin and discuss the benefits that an informative prior can provide in analysis of the resulting posterior.

Significance and probability of direction

One of the consequences of the widespread use of frequentist statistics is the standard rule of measuring statistical significance through the use of p-values. However, it has been reported that, in spite of successful model predictions having significant p-values, in many cases strong statistical significance does not carry any predictive value in biomedicine (Bzdok et al., 2020). And in a more broad sense, the over-reliance on p-values is one of the reason behind the reproducibility crisis in science, accentuated by the use of small sample sizes (which lead to large variability in the results) and a lack of multiple independent validations (Gelman and Loken, 2014). The Bayesian framework can easily provide an answer to this problem without an over-reliance on statistical tests.

For many applications in computational biology, the primary goal of the computational method is to serve as the first approach to a problem which later should be extensively validated and studied in laboratory conditions. Hence, in many cases, the aim is to discover the direction in which a given variable affects an outcome over the exact value of the effect. Extremely precise models are harder to accurately produce with the type of data available, and usually the objective of a computational analysis is to provide a binary answer to whether a given feature has a positive or negative effect in the problem.

In this context, the Bayesian probability of direction, or pd (Makowski et al., 2019), can be defined as the certainty with which an effect goes in a particular direction (positive or negative). In other words, it corresponds to the percentage of the posterior probability with the same sign as its median. It can be computed by estimating the density function and then evaluating the area under the curve (AUC) of the density curve on corresponding (positive or negative) semi-planes.

Once the direction of the effect is confirmed, the research focus should be shifted toward its significance, including a precise estimation of its magnitude, relevance and importance. For example, Figure 2.21 shows a zoomed in version of Figure 2.20 for selected variables with a posterior density mostly lying in one semi-plane. The full density and various percentages (obtained by trimming half the missing percentage from each side for the plot) are shown to illustrate how the probability of direction metric works.

Figure 2.20 clearly shows that all priors can confidently assign a direction of effect (strictly positive) to the SOX2 variable. However, for other variables such as PR, only the informative prior $\mathcal{N}(0, 1)$ leads to a posterior where an important part of the density is confined in the negative semi-plane. In fact, Figure 2.21 confirms that, for PR, parts of 99.9% (in red) and 99% (in yellow) of the posterior distribution lay in the positive part of the plane. This means that there is a low but non-zero probability that the parameter associated to PR is positive (and hence, for the effect of this variable to be in either direction). However, if we restrict the posterior to 97.5%, we see that the density takes negative values only. The direct interpretation of this is that, with a probability of 97.5%, the effect of the PR gene in the resistance is opposite to that of SOX2.

In summary, the probability of direction can not be viewed as an indicator of significance (or a direct substitute of the classical p-value), but it does provide a robust, model-independent and easy to interpret approach to assess the direction of the effect in which a variable have an impact on a response. This represents a compelling example of how Bayesian methodologies could very easily be implemented in research without additional training in statistics due to the inherent probabilistic nature of the method.

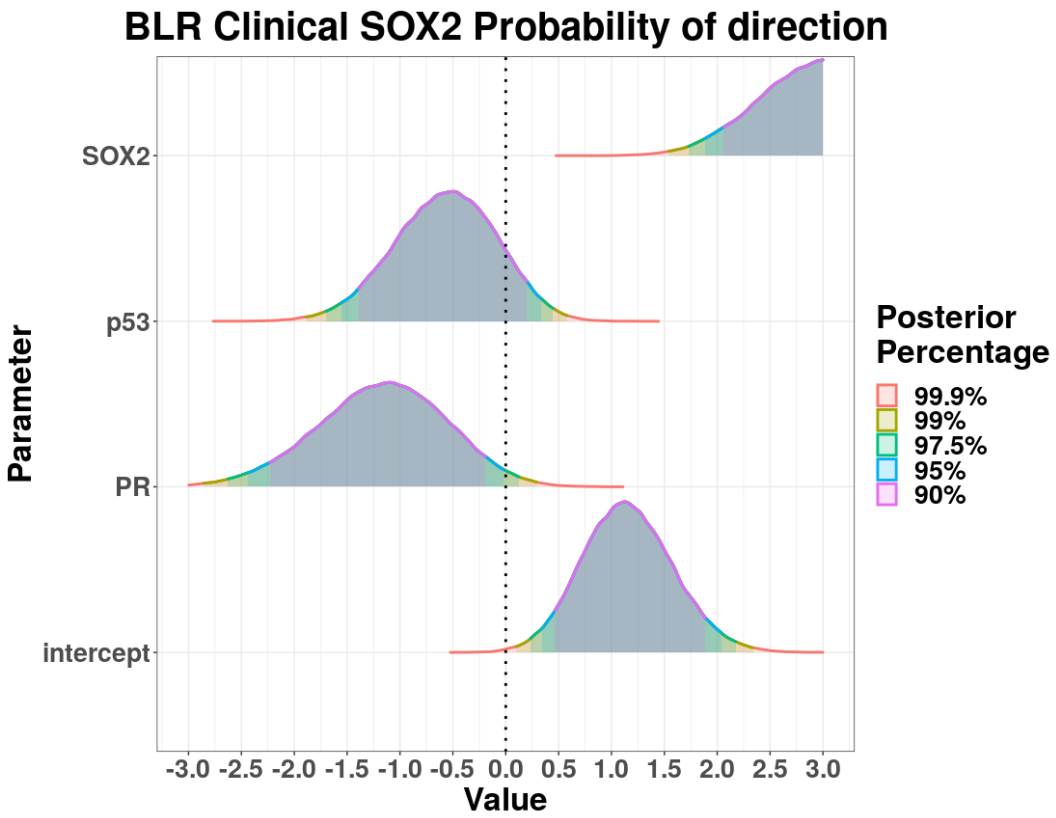

Figure 2.21: Posterior distributions (zoomed around 0) for the SOX2 clinical BLR model using the $\mathcal{N}(0,1)$ prior. The colour indicates the percentage of the curve under it.

2.4 Conclusions

In this chapter, we discussed each step of a general pipeline for the mathematical analysis of clinical data in the context of the available clinical datasets.

We introduced two datasets, the SOX2 clinical and the TCGA clinical. The former was thoroughly curated and pre-processed, while the latter was introduced and explained for further use in Chapter 4. We then reviewed the major components of the Bayesian inference and discussed in detail the most promising sampling techniques for the use in biomedical applications.

We also refined the performance evaluation tools and proposed an effective sample size (ESS) estimator that accounts for importance sampling, irreversibility and correlated samples. We then presented the main result of this chapter, our novel adaptive multi-stage integration approach, s-MAIA, for the importance sampling Generalized Hamiltonian Monte Carlo method — MMHMC. The s-MAIA algorithm can compute, for any chosen problem and stepsize (within the stability interval), the dimensional stability limit and the integrator coefficients for the family of 2- and 3-stage splitting schemes that guarantee the best conservation of modified energy for harmonic forces.

The improvements in the conservation of the modified energy imply an increase in acceptance rates for MMHMC and, as a consequence, an enhancement in sampling quality. The s-MAIA algorithm has been implemented in our in-house software HaiCS and does not introduce computational overhead in an MMHMC simulation. This is achieved by the efficient utilization of the simulation data generated during the burn-in stage and the use of the pre-tabulated values for optimal integration coefficients. The comparative analysis of performance of s-MAIA and the state-of-the-art fixed parameters 2- and 3- stage integrators using standard performance metrics revealed that s-MAIA outperforms the fixed-parameters integrators in accuracy, efficiency and stability.

We also applied MMHMC reinforced with s-MAIA to the SOX2 clinical dataset to demonstrate a potential of the improved method for solving biomedical problems.

Chapter 3

Bioinformatic Tools for *ad hoc* RNA-Seq Data

In previous chapters, we mentioned the role the advent of sequencing plays in offering new insights to biological and biomedical research. Sequencing techniques provide a window into the inner workings of cell structures and open paths to study intra- and inter-cellular mechanisms through the analysis of gene expression patterns. However, studying this vast landscape and unlocking the hidden information stored in the genetic material are not feasible without mathematical analyses. An accurate mathematical representation of these processes provides the lens through which one can learn the valuable information from RNA-seq data. RNA-seq offers a detailed and comprehensive view of the transcriptome by gathering simultaneously the expression of thousands of genes in a single experiment. Its scalability allows for the exploration of gene expression patterns in various biological contexts, and its sensitivity and accuracy permits their identification even at low gene expression levels. This unique place at the intersection of mathematics and molecular biology allows for a quantitative evaluation of the biological differences between distinctive conditions, facilitating the discovery of biomarkers, potential drug targets, and regulatory networks.

In this chapter, we lay out the fundamental properties of RNA-sequencing data and explore the principles that guide the mathematical analyses behind it. For this purpose, we discuss the data pipeline for RNA-seq experiments, depicted in simplified form in Figure 3.1, and explore the mathematical tools at play at various stages of the analytical process. In the first place, we contextualise the technique by looking at the technology that supports it and the subsequent mathematical treatment of the data produced. Then, we explore an RNA-seq dataset related to the problem of resistance to endocrine therapy in breast cancer and extracted from relevant cell lines developed in the Cancer Heterogenity Lab at CIC bioGUNE. We introduce the methods needed for a correct treatment and analysis of such data and select the most appropriate ones for our purposes. We end the chapter by presenting an all-in-one tool for analysis of RNA-seq data, supplemented with practical examples of its use. The tool offers various visualization techniques and is addressed to non-expert users with the aim of facilitating the RNA-seq data analysis.

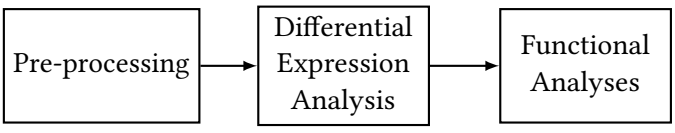

Figure 3.1: General pipeline for the mathematical analysis of *-omics* data.

3.1 Pre-processing

The term RNA-sequencing encompasses all the next generation sequencing techniques that result in some form of quantification of the transcriptome characteristics, the most common being the identification of a gene expression profile as gene counts[1]. There is no unique protocol or methodology to perform RNA-sequencing, nor an optimal one, and the choice may be influenced by the specific research goal or the characteristics of the samples to be sequenced (Conesa et al., 2016). However, in every RNA-seq experiment, there are two very distinguishable steps. First, the experimental part, which includes the preparation of the samples and the sequencing itself. The second one is the analysis of its output and the advanced analytical techniques connecting the result to various biological processes. The experimental side falls outside the scope of this thesis, however, we refer the reader to Hrdlickova et al. (2017) for a very informative review of the available methods with the practical recommendations for their use.

We start with presenting a quick summary of the main steps usually taken to go from raw data to gene counts. We remark that here many small details have been omitted for simplicity but, for more detailed information on the issue, we recommend the Galaxy platform (Afgan et al., 2016). Galaxy is a useful open-source web application where the complete RNA-seq analysis pipeline can be performed using the tools that will be introduced next, and many more.

The raw output of the sequencing experiment is stored as a text file containing the sequences of nucleotides read by the machine, as well as method-dependent codes for quality control purposes. This file is called a *FASTQC* file (Simon, 2015). Sequencing reads of this file need to be mapped to a reference genome, in order to determine their origin and location. Alignment algorithms are usually employed for this task, such as *Bowtie* (Langmead and Salzberg, 2012) or *STAR* (Dobin et al., 2013), which use heuristic methods to efficiently search for the best match between each read and the reference genome, as illustrated in Figure 3.2. After alignment, *SAM/BAM* (Sequence/Binary Alignment Map) (Li et al., 2009) files are generated, containing information on alignment positions and mapping quality. Next, a *GTF* (General Transfer Format) file is needed to link the aligned reads and the genomic coordinates of known genes. Finally, software such as *featureCounts* (Liao et al., 2014) or *HTSeq* (Anders et al., 2015) takes these two files as input to count the number of reads that align with each annotated gene. This results in a *gene count matrix*, which contains the raw counts for all genes and samples. In practice, such matrices have heavily unbalanced structures, with just a few samples but well over 20.000 gene features.

Many choices and options during these steps, such as library type (single or pair-ended) or strand-specificity (5' or 3'), may influence greatly downstream analysis and the final output data, but an in-depth explanation of each characteristic possible is outside the scope of this thesis. Here, we want to highlight three aspects that affect the shape or distribution of the final count matrix and appear as relevant players in different parts of the pipeline we present later in this section.

The *read length* corresponds to the number of base pairs that can be read at one time. Longer read lengths give more accurate information on the relative positions of the bases in a genome, however, the cost of sequencing also increases with read length. Shorter reads may result in more ambiguous mappings, potentially leading to lower mapping rates and a reduction in the number of available counts for gene expression estimation. Usually, a read length of 50 bases is sufficient for an accurate mapping of reads to a reference genome in counting experiments.

The *library size* (or *sequencing depth*) accounts for the total number of reads that are generated from an experiment. As before, the higher the number, the more precise the quantification. Differences in library size between samples can introduce biases in analyses downstream. Normalization methods are needed to account for library size differences (and other sources of variability) and ensure accurate

[1]We recall here the wide use of the word gene in some -*Omics* fields introduced in section 1.1.2

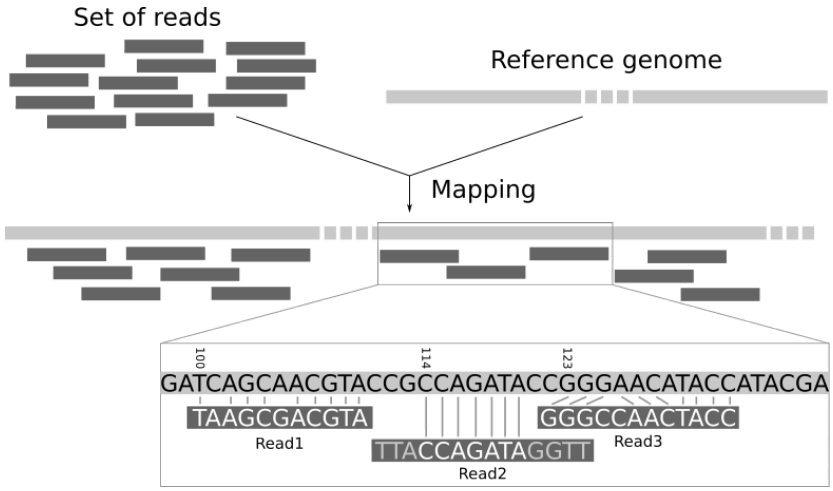

Figure 3.2: Illustration of the mapping process. The input consists of a set of reads and a reference genome. In the middle, it gives the results of mapping: the locations of the reads on the reference genome. The objective of the alignment is to provide appropriate matches, even accounting for possible mismatches. Image from Galaxy Training Network (Wolff et al., 2024) used under Creative Commons Copyright license.

comparisons of gene expression levels.

Finally, a crucial design factor is the *number of replicates*. The number of biological replicates influences the statistical power and reliability of differential expression analysis. A higher number of replicates allows for an easier identification of true biological differences and for a diminishing influence of experimental noise. It is standard practice that at least three replicates are required for an accurate identification of a sample.

We explore in more detail the impact of these factors as we discuss the structure of the available data and the mathematical treatment needed. For more information, we refer the reader to a complete review of best practices regarding RNA-seq analysis and many related processes in Conesa et al. (2016).

3.1.1 Negative binomial model

The experimental and pre-processing procedure that we just presented has a *gene count matrix* as its main output. In general, count data represents the number of occurrences of a particular event in a fixed unit of time. For RNA-seq, this corresponds to the number of sequencing reads aligning to a given gene for one experiment.

Mathematically, general count data follows discrete probability distributions, such as the Poisson distribution. However, the Poisson distribution assumes that the mean and variance of the distribution are equal but, as we saw so far, a multitude of factors add inherent variability to the experiments. Thus, the Poisson distribution is not a suitable candidate for modelling sequencing experiments.

In this context, the **Negative-Binomial distribution** appears as a useful candidate. It is also a discrete probability distribution but, crucially, it has the flexibility to handle overdispersion in the data. In fact, it has become a preferred choice for modelling the count data $K_{g,m}$ obtained in sequencing experiments:

$$K_{g,m} \sim \text{NB}(\mu_{g,m}, \alpha_g). \tag{3.1}$$

The mean parameter, $\mu_{g,m}$ in 3.1, is gene $(g = 1, ..., G)$ and sample $(m = 1, ..., M)$ specific, while the dispersion parameter, α_g, is gene-specific only. The dispersion parameter accounts for the inherent biological variability and technical noise in RNA-Seq experiments, providing a realistic

representation of the complex nature of gene expression. Figure 3.3 illustrates how this distribution can serve as a valid model for the type of data available

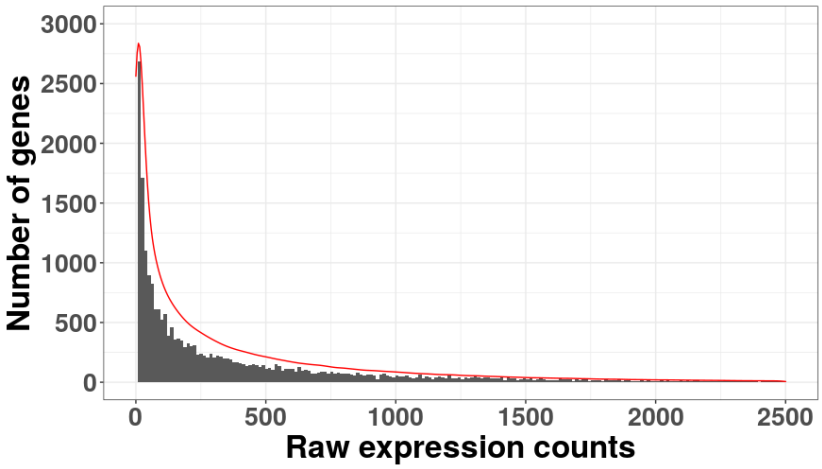

Figure 3.3: Histogram of count data in an RNA-seq experiment, with the (scaled) Negative-Binomial distribution superimposed in red.

This modelling choice is essential for robust statistical inference in differential gene expression analyses, where accurate representation of the data distribution is critical for identifying truly differentially expressed genes among the noise and variability in biological systems. We explore this issue further in section 3.3.3.

3.2 RNA-seq datasets

In this section, we introduce the RNA-seq data at our disposal, presenting its main features and its relevance for the study of the problem of resistance to endocrine therapy in breast cancer. We focus our attention on two very distinctive datasets.

First, we discuss the data extracted from sequencing several biological states of breast cancer cell lines used for research in the context of resistance to tamoxifen. From now on, we will refer to this dataset as bG-RNA-seq. Within the landscape of this thesis, this data will be first employed in the context of **Objective 1**, as our goal is to use it to develop tools and offer bioinformatic solutions. In this case, we have access to the raw data right after sequencing, so we use this dataset to show several techniques for normalization, treatment and analysis of the data, and illustrate our main contributions in this section. On the other hand, the bG-RNA-seq dataset will be featured as well in the next chapter as part of the combined analysis of cell and patients data, which is essential for **Objective 2**.

The other major cohort at play is the high-throughput data extracted from the TCGA database and introduced in section 2.1.2. Since the TCGA dataset, was accessed via a public repository, it is less flexible and was already processed to some degree when obtained. Additionally, patients data exhibits a significantly higher degree of heterogeneity compared to that derived from cell lines, offering a distinct setting to extract relevant information. In this chapter, we restrict ourselves to the presentation of the main characteristics of this TCGA dataset and commenting on the most relevant differences between both the TCGA and bG-RNA-seq datasets. The analysis of the TCGA dataset, central for the completion of **Objective 2** and used only in that context, is undertaken in the following chapter.

3.2.1 Cell line bG-RNA-seq data

3.2.1.1 Data description

Cell lines are immortalised cell cultures, derived from a specific tissue, which provide researchers with a reproducible and controlled environment to study cellular mechanisms, disease progression and response among other scientific questions. They play a crucial role in advancing our understanding by providing an easily replicable environment where new treatments can be tested and biological conditions can be manipulated.

ER+ breast cancer is the most common subtype, representing approximately 75% of all breast cancer cases. Cell lines commonly used to study ER+ breast cancer include MCF7 (Soule et al., 1973) and T47D (Keydar et al., 1979) cells. Triple-negative breast cancer (TNBC) is less common, approximately 10% of all cases and usually with worse prognosis. MDA-MB-231 cells (Cailleau et al., 1974) are one of the cell lines often used to investigate this breast cancer subtype. Despite great advances in breast cancer research, development of resistance to current forms of therapy is still widespread. For example, after five years of tamoxifen therapy for ER+ breast cancer, the risk of recurrence can range from 10% to 41%, depending on size, tumour grade and lymph node involvement (Pan et al., 2017).

One of the main research interest of the Cancer Heterogeneity Lab at CIC bioGUNE is understanding the clinical problem of resistance to hormone therapy. To this end, they have developed models of resistance to tamoxifen by exposing cells to this treatment over a long period of time to mimic the development of resistance to hormone therapy observed in patients with ER+ breast cancer (Domenici et al., 2019; Piva et al., 2014). As a result of these and other studies, they have identified a series of factors that are potentially implicated in resistance to hormone therapy and, therefore, are of interest for further research. Consequently, RNA-sequencing analyses were performed on various key factors in different breast cancer cell lines, as summarized in Figure 3.4. The data from these RNA-seq experiments is highly relevant to various studies performed at the Cancer Heterogeneity Lab, and most of it is still unpublished. Therefore, there exists a need for a platform that can be easily accessible and capable of quick and intuitive analysis and exploration of the experimental data. This is ultimately the main goal of this chapter and it is discussed in section 3.5.

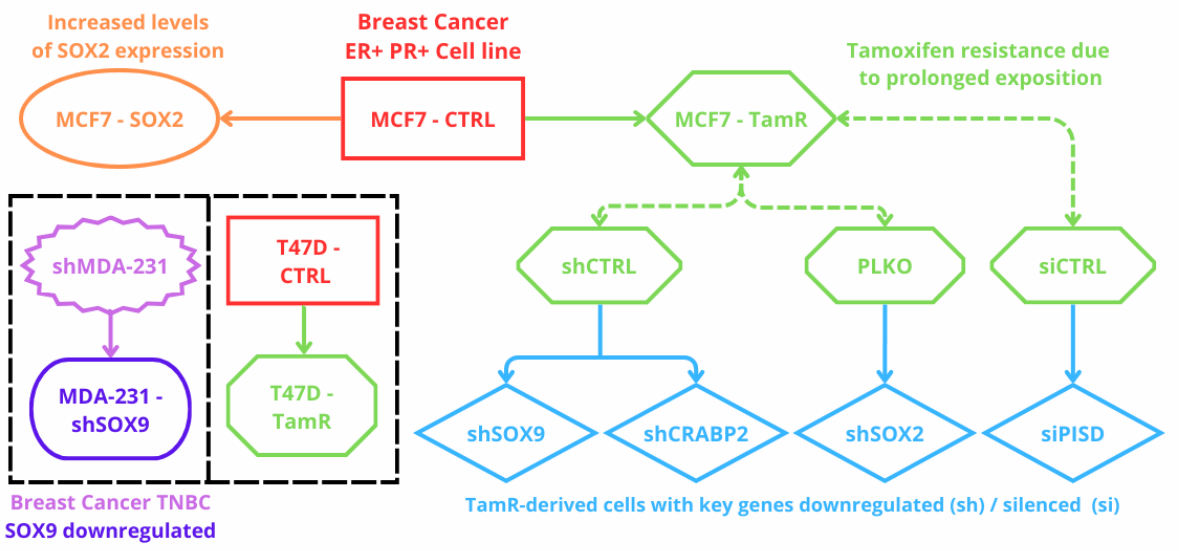

Figure 3.4: Map depicting all available cell lines sequenced for the bG-RNA-seq dataset and the relations between them.

Several of the conditions shown in Figure 3.4 (TAMR, sh_CRTL, PLKO and si_CTRL) represent

virtually the same state of resistance, and serve as different controls for the experiments derived from them. The sh_* prefix indicates that the target gene has been downregulated using shRNA (Short Hairpin RNA), a type of RNA molecule designed to interfere with the expression of specific genes by inducing the degradation of the messenger RNA (mRNA). For the case of PISD, the downregulation was not carried out using shRNA, but rather silenced using CRISPR/Cas9-mediated gene knockout. This means that the gene is nearly completely silenced, as opposed to shRNA, where the expression decreases but is not completely nullified.

The sequencing was done entirely at CIC bioGUNE, where the Cancer Heterogeneity Lab handled the library preparation, and the Genomics Platform performed the sequencing. In total, the bG-RNA-seq dataset consists of 14 different biological conditions, all sequenced in triplicate, amounting to 42 individual sequencing experiments. Sequencing libraries were prepared using the *TruSeq Stranded Total RNA with Ribo-Zero Human/Mouse/Rat kit* (Illumina Inc., Cat.# RS-122-2201), following the guidelines outlined in the *TruSeq Stranded Total RNA Sample Preparation Guide* (Part # 15031048 Rev. E). Libraries were sequenced on a HiScan-SQ platform (Illumina Inc.) with single-end 50-nucleotide reads.

3.2.1.2 Data exploration

A key part of all RNA-seq experiments is quality control. Several quality control mechanisms are in place at different points in the sequencing, alignment, and counting that can already raise the alarms. However, even with all these checks, it is possible to obtain the final count matrix that still possesses some unwanted alterations. In this case, a common quality control option is to perform a Principal Component Analysis (PCA) (Pearson, 1901). PCA is a dimensionality reduction technique widely used in statistics to transform high-dimensional data into a lower-dimensional space. The principal components capture the directions with the most significant variability in the data. In our case, PCA should be able to capture similarities between replicated conditions, and those should cluster together if no problem has occurred. Next, we demonstrate how the PCA analysis can help in cleaning the dataset from conditions non-consistent with the rest of the data for some technical reasons.

To test this, we performed PCA on the cell line bG-RNA-seq dataset trying to identify if this, or other condition, behaves unexpectedly. Since PCA assumes normality of the data, the counts are log-transformed first. Then all samples are plotted in a 2D plot where the axes are the first and second principal components, which describe the primary directions of variability in the dataset. PC1 captures the most significant variance, and PC2 captures the next most significant variance, orthogonal to PC1. They represent the directions in the feature space along which the data varies the most.

The PCA plot in Figure 3.5 shows that all experiments from each MCF7 cells cluster together, except for the sh_SOX2, which are more scattered than any other cell line. MDA-MB-231 and T47D are not shown in this figure, as they present no problem.

This behaviour of the sh_SOX2 experiments was the first indication of a problem, which was later confirmed by checking the levels of expression of SOX2. The gene was not downregulated as desired, which served as a final confirmation to discard the experiments and repeat the sequencing. Only this condition was sequenced again, alongside a proper control, named PLKO.

No clustering of sh_SOX2 and its control, PLKO, with the rest of MCF7 is observed after the new sequencing, as shown in Figure 3.6. However, the PC1 value for both MCF7 clusters is similar, with the differences being more accentuated in PC2, as both clusters still contain MCF7 cells and the differences are mostly related to technical issues. We remark that the cluster in the middle corresponding to the majority of MCF7 cell lines behaves similarly to Figure 3.5 without the sh_SOX2 condition.

The differences arising in PC2 values for various conditions in Figure 3.6 can be understood more clearly when looking at the distribution of library sizes. Figure 3.7 shows that the sequencing of the

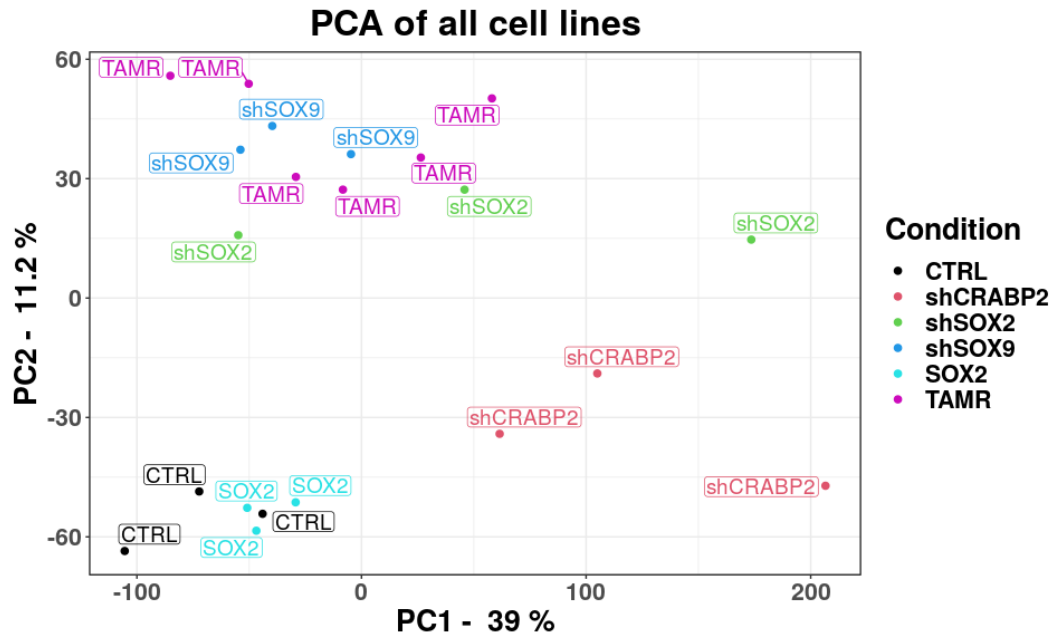

Figure 3.5: PCA analysis of the MCF7 cell lines from the bG-RNA-seq data, with the "sh_TAMR" and "TamR" conditions grouped with the pink label "TAMR". The shSOX2 condition (in green) is scattered across the first principal component more visibly than other conditions

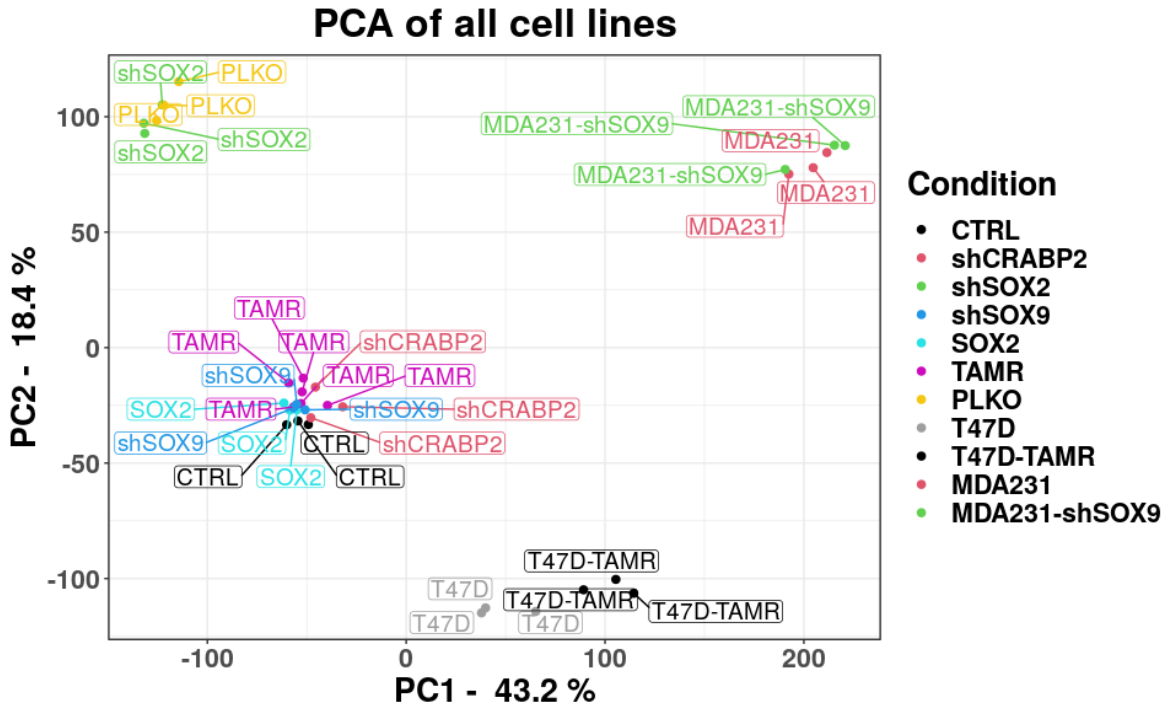

Figure 3.6: PCA of all cell lines from the bG-RNA-seq dataset. MDA-MB-231 (labelled as "MDA231") and T47D form their own clusters. Those without a cell line name before the condition are MCF7. These form 2 clusters, one for the original sequencing batch and another for the sh_SOX2 and PLKO conditions, which were sequenced again.

MDA-MB-231 conditions and sh_SOX2 share a similar library size and, as such, have more common features between them than with the rest of MCF7 and T47D conditions.

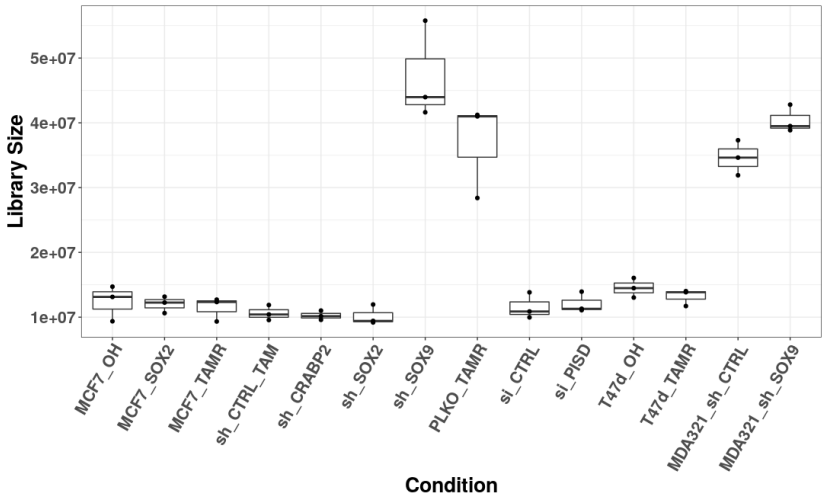

Figure 3.7: Difference in library size between conditions.

3.2.2 TCGA patients RNA-seq data

The cell line bG-RNA-seq dataset was created in-house using known biological conditions, with full control over nearly all aspects of the experiment. This means that we can identify any issues in the pre-processing and tackle them. Besides, one can expect a homogeneous behaviour for similar cell lines, making quality control and any subsequent analysis, much easier to achieve.

With the aim to include in our analysis more sequencing information on regular and resistant tumours, we next refer to the RNA-seq data from the TCGA patients presented in section 2.1.2. All experiments included in the dataset were done on tumour samples taken at the moment of diagnosis following the protocols for data generation and treatment, established by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), which play key roles in coordinating the TCGA database. For the purpose of our analysis in the next section, we consider the already aligned and counted gene count matrix.

In contrast to the cell line bG-RNA-seq data, heterogeneity is the dominant factor in the TCGA dataset. As we mentioned, breast cancer is a heterogeneous disease, and even tumours with similar phenotypes from different patients can present transcriptomic profiles differing greatly. We come back to this problem in section 4.3.2.2, when we classify the selected tumours in the TCGA dataset.

3.3 RNA-seq data analysis

In this section, we illustrate the analysis pipeline of RNA-seq experiments using the cell line bG-RNA-seq data from section 3.2.1.1. Our objective is to present the most relevant steps usually taken for the analysis from a mathematical perspective and show how relevant biological information from the data can be extracted. In addition, we wish to provide the necessary context and intuition for understanding the steps and options we included in our multipurpose tool that we present at the end of this chapter.

3.3.1 Normalization

One key factor when studying RNA-seq data is that the absolute number of reads obtained from the raw data, i.e., the gene count matrix, is not reflective of true biological expression levels. Adjusting for differences in read length, sequencing depth among samples or technical factors is essential to make comparisons between different biological conditions possible and to correct bias within a sample. In this context, normalization becomes a fundamental step to transform raw counts into values that

posses biologically sound information. Without proper normalization, the interpretation of gene expression data becomes impossible, as differences in counts emerging from technical variability might be erroneously attributed to biological differences.

The most common approach for normalization is to take the raw count matrix $K_{G\times M}$ and use sample-wide scaling factors s_m, where $m = 1, \ldots, M$, and M is the total number of samples. For each sample $m = 1, \ldots, M$, the counts in the original matrix are divided by the same factor for all genes $g = 1, \ldots, G$ in order to obtain the matrix of normalized counts N

$$N_{g,m} = \frac{K_{g,m}}{s_m}, \tag{3.2}$$

where the scaling s_m is sample-specific, but can change for different samples of the same biological condition.

A wide variety of normalization techniques exist nowadays, presenting different approaches for the calculation of the scaling factors. Some are based only on sequencing depth or gene length, and are usually better for within-sample analysis of gene abundance. Others take some reference value from the whole count matrix, or a subset of relevant genes from all samples, and create a scaling factor for between-sample comparisons. Next, we explore some methods, mostly of the between-sample group, and compare them to select a methodology for our analyses.

Counts per million (CPM)

The CPM approach is a basic gene expression unit that normalizes only for sequencing depth (number of mapped read counts), so the method is sometimes referred also as RPM (Reads per million). CPM is calculated by dividing the mapped read count by a per-million scaling factor taken from the total number of mapped reads:

$$s_m = \frac{1}{10^6} \sum_{g=1}^{G} K_{g,m}. \tag{3.3}$$

CPM is a rather simple approach that allows for a fair comparison of counts within the same sample. Through scaling the counts by the total number of reads, one removes much of the influence coming from the library size. However, the method fails to account for other factors, or for this particular factor relative to other samples, which can make it a poor choice for differential expression analysis — which requires between-sample comparisons.

Trimmed Mean of M-values (TMM)

The idea behind TMM (Robinson and Oshlack, 2010) is that certain genes have higher read counts due to technical reasons, and it may be desirable to avoid those when calculating the scaling factors. Therefore, the method first trims genes with high expression and then normalizes using the remaining set.

The method starts by scaling each sample by its total number of reads (library size, L_m) and selecting the sample with the upper quartile closest to the average as the *reference sample*, K' with library size L'. Then it calculates log-expression ratios and absolute expression levels (i.e., M- and A-values, respectively) relative to that sample.

$$\mathrm{M}_{g,m} = \log_2 \left(\frac{K_{g,m}/L_m}{K'_g/L'} \right), \tag{3.4}$$

$$\mathrm{A}_{g,m} = \frac{1}{2} \log_2 \left((K_{g,m}/L_m)(K'_g/L') \right). \tag{3.5}$$

Genes are then ordered by M and trimmed twice. Typically, M-values of 30% and A-values of 5% are trimmed at the upper and lower end of the data. The mean of the remaining M-values (the trimmed mean) is then calculated for each sample, and raw counts are finally scaled by the trimmed mean and the library size of their sample.

The construction of the method makes it very robust against outliers, as these are trimmed away. However, its dependence on a reference sample can cause problems if the changes in gene expression between all available samples are big and no sample can really be taken as a good reference point.

Relative log expression normalization (RLE)

The RLE method was originally presented for DESeq (Anders and Huber, 2010) methodology and is also included in the popular R package DESeq2 (Love et al., 2014). It is based on the construction of a pseudo-reference value for each gene, by taking the geometric mean of a gene counts across all samples. We define this pseudo-reference value as GM_g

$$\mathrm{GM}_g = \sqrt[M]{K_{g,1} \cdot K_{g,2} \cdot \ldots \cdot K_{g,M}}. \tag{3.6}$$

The scaling factor s_m for each sample is then obtained as the median of the values of the ratios of sample counts with respect to the pseudo-reference value

$$s_m = med \left(\frac{K_{g,m}}{\mathrm{GM}_g} \right) \quad \text{for} \quad g = 1, \ldots, G. \tag{3.7}$$

The method is based on the key assumption that not all genes G are differentially expressed between condition, i.e, that a sufficient number of genes are stably expressed across samples. This is usually a very accurate assumption, as comparisons often involve targeted changes rather than general differences. However, if a large proportion of genes are highly differentially expressed, the effectiveness of the normalization process may be compromised.

Total Ubiquitous genes (TU)

Originally, in the TU method (Glusman et al., 2013), a trimmed set of genes is obtained for each sample by excluding genes with zero values, sorting the non-zero genes by expression level in that sample, and by removing the upper and lower ends of the sample-specific expression distribution. Ubiquitous genes are then obtained as the intersection (common genes) of the trimmed sets of all samples considered. Scaling factors are defined such that the scaled counts for the ubiquitous genes in each set have the same value, i.e., the average across all samples.

Trimming percentages (and hence the set of ubiquitous genes) are selected to maximise the number of uniform genes. This means the trimming values are dataset dependent. In the original publication, it was defined manually to maximise differences between various normalization strategies. For the purposes of this thesis, we select trimming percentages such that they provide the best Area Under the Cross-Validation Curve (AUCVC) for post-normalization counts, similarly to the approach adopted in Wu et al. (2019). In the next section, we will explore in depth the ideas behind the AUCVC.

To test this approach, we first build a fitness terrain for the AUCVC function corresponding to all admissible combinations of upper/lower trimming percentages and then identify the optimal cuts for our dataset as 40% from the top and 15% from the bottom as seen in Figure 3.8. The results of testing this and other approaches will be discussed in the following section.

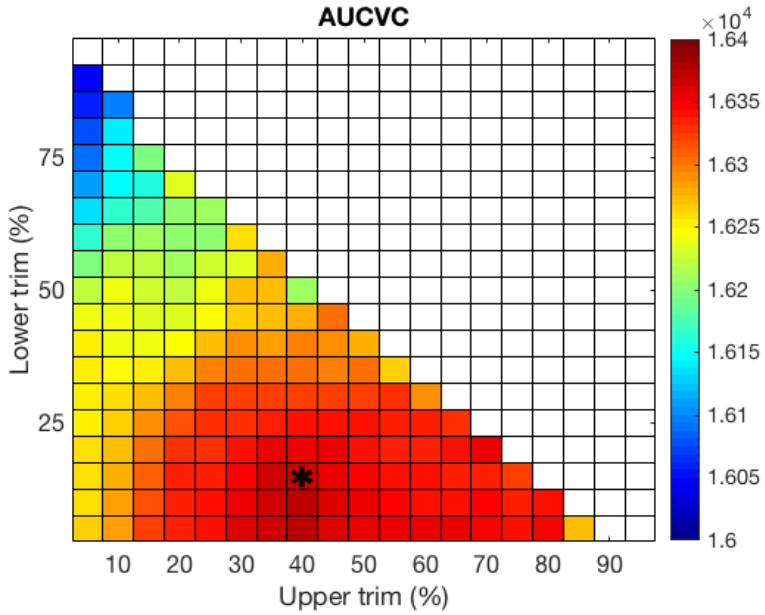

Figure 3.8: Fitness terrain obtained for all admissible combinations of lower and upper trimming percentages in the bG-RNA-seq data set. The asterisk indicates the combination corresponding to the highest AUCVC value.

3.3.2 Comparison of normalization techniques

We wish to decide which of these options for normalization is more suitable for our problem. For that, we take as our $K_{G\times M}$ the count data matrix that corresponds to the MCF7 cell lines clustering together in the bG-RNA-seq dataset (see Figure 3.6). It is common that the raw count matrix before normalization still contains plenty of genes with null or extremely low counts. Therefore, it is natural to do a first filter on total expression levels by removing those genes with a low count in total. In our case, the threshold was selected to be 30 counts across samples as

$$\sum_{m=1}^{M} K_{g,m} < 30\,, \quad g = 1, ..., G. \tag{3.8}$$

Here, the threshold value was selected *ad-hoc*, as there is no consensus on an optimal value. However, Figure 3.9 shows that using just a small filter can remove the majority of zero-counts genes and reduce dramatically the number of available genes. The value of 30 was taken so that if a gene has a mean count of 10 in a given condition (across 3 replicates), it passes the threshold.

Thus, after applying such a filter, the total number of rows (genes) went from $G_{raw} = 23459$ to $G_{>30} = 15814$. The number of columns (samples) is $M = 24$ (3 replicates per cell line).

To assess and compare performance of the reviewed normalization methods in our gene matrix, we need a reliable metric. Several options are available in the literature, but most of them rely on synthetic data to test the methods in extreme scenarios. As we are interested in a normalization method able to reduce between-sample variations of expression levels for a particular dataset, studying performance on general synthetic data is not helpful. On the other hand, many available metrics

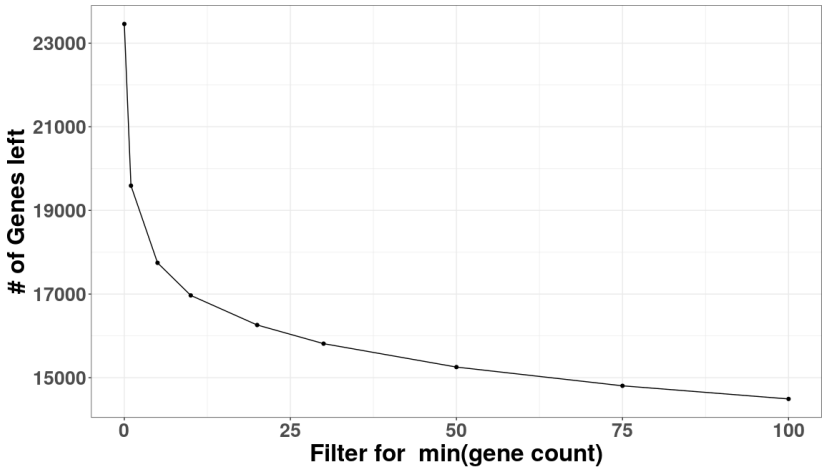

Figure 3.9: Number of genes left for normalization after removing genes below a threshold value across samples.

for real-data normalization count on the presence of experimental controls, which we do not have for our dataset (no spike-in controls or knowledge of "true" differential expression levels through housekeeping genes (Jain et al., 2020)). Hence, we need a metric that can allow us to select between normalization strategies by using just statistical properties of the resulting normalized matrix.

With this in mind, we set to use one of the metrics in Wu et al. (2019) for comparing the performance of the considered methods, namely, the coefficient of variation (CoV) defined for each gene as

$$\mathrm{CoV}_g = \frac{\sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(N_{g,j} - \bar{N}_g)^2}}{\bar{N}_g}, \tag{3.9}$$

where $\bar{N}_g = \frac{1}{M}\sum_{m=1}^{M} N_{g,m}$ is the mean count for gene g across the M samples in the already normalized matrix $N_{g,m}$. We then rescaled the CoV to the interval $[0, 1]$ and computed the number of genes with scaled CoV below a cut-off that varied between zero and one in 1% increments. That is, we computed the number of genes with scaled CoV value on the grid $x_i = (i-1)h$, with $h = 0.01$ and $i = 1, \ldots, 101$. The area under the CoV curve (AUCVC) was then approximated with the trapezoidal rule.

Figure 3.10 show that the RLE method gives the largest area under the CoV curve (AUCVC).

RLE has a slight advantage over both TMM and CPM, whereas the TU method falls far behind, despite being an AUCVC related metric. This may be attributed to the fact that there is no substantial amount of genes skewing the distribution, and the genes removed are actually needed for an accurate normalization. Similarly, the good performance of RLE could be explained by the assumption that most genes have more or less stable expression across samples.

Hence, we select RLE as the methodology for normalization. The RLE method is part of the widely used DESeq2 package (Love et al., 2014), which provides the straightforward implementation of differential gene expression analysis. In the upcoming section, we discuss this approach and present the DESeq2 pipeline for differential expression analysis that starts with RLE.

3.3.3 Differential Expression Analysis

Differential gene expression analysis (DGE or DEA) is one of the most relevant techniques in the analysis of sequencing data. It allows the quantification of changes in gene expression between differ-

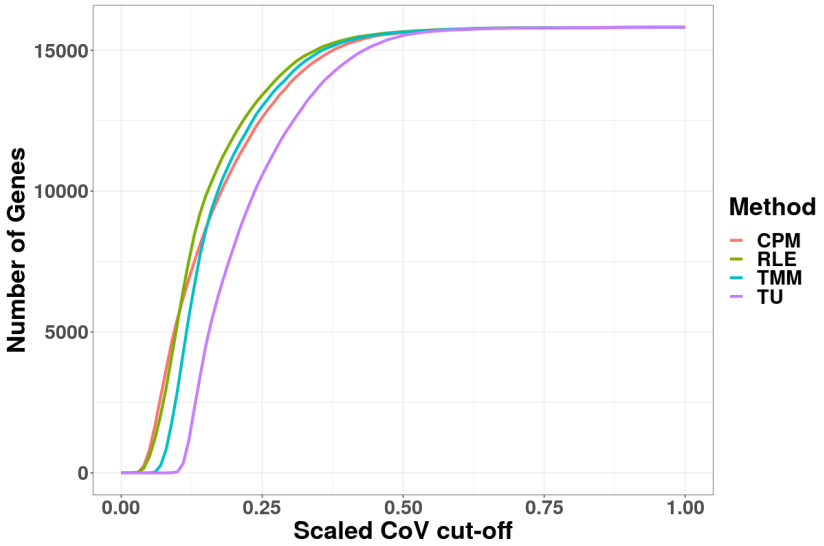

Figure 3.10: Comparison of CoV curves for all normalization methods considered.

ent biological conditions and provides valuable insights into the underlying mechanisms governing cellular processes. Besides, it helps to understand how genes can contribute or fuel changes in certain conditions or how they respond to them.

The fundamental goal of differential expression analysis is to discern genes that exhibit statistically significant changes in expression levels between experimental conditions. The purpose is not only to identify differentially expressed genes, but also to quantify the magnitude of these changes. This information is crucial for deciphering which genes drive certain biological phenomena, for example, those varying from normal cellular conditions to a state of disease. Highlighting genes associated with specific conditions, can uncover potential biomarkers, therapeutic targets and pathways implicated in some biological processes.

From a mathematical perspective, the count data is assumed to follow a Negative-Binomial distribution, as introduced in equation 3.1. Hence, the key elements for differential expression are an accurate estimation of the mean μ and dispersion α parameters for each gene in different conditions, and then a statistical test to identify if the changes in expression are statistically significant.

Attending to the results of the previous section, we follow the DESeq2 pipeline to obtain these parameters, which allow the comparisons between expression levels in different samples that are needed for differential expression. DESeq2 uses the sample-specific scaling factors s_m (eq. 3.7) explained in the previous section to estimate the mean expression level $\mu_{g,m}$ for a gene g in a sample m as

$$\mu_{g,m} = s_m q_{g,m}, \tag{3.10}$$

where $q_{g,m}$ is a quantity proportional to the concentration of DNA fragments from the gene g in the sample m, i.e, the sample reads. The dispersion parameter α_g describes the variance of the data as

$$\text{Var } K_{g,m} = \mu_{g,m} + \alpha_g \mu_{g,m}^2. \tag{3.11}$$

DESeq2 uses maximum-likelihood estimates (MLE) to obtain gene-wise dispersion values, which rely only on the data of each individual gene. Then, an empirical Bayes approach is used to shrink the values of the estimate. The strength of shrinkage depends on how close true dispersion values are to the fitted value and on the degrees of freedom (the more the sample size increases, the less the shrinkage).

To test for differential expression between conditions, DESeq2 employs a generalised linear model (GLM) with a logarithmic link based on the normalized counts

$$\log_2(q_{g,m}) = \sum_k x_{m,k}\beta_{g,k}, \tag{3.12}$$

where $x_{m,k}$ is the design matrix which hosts the information from the groups to compare. In the simplest case of a two-group comparison, as in altered vs control samples, the design matrix elements indicate whether a sample m belongs to the altered or control group. The GLM fit returns coefficients $\beta_{g,k}$ indicating the overall expression strength of the gene. The GLM fits are performed in two steps. First, the MLEs are obtained for the *log Fold Change (FC)* which are used to fit a zero-centered normal distribution to the observed distribution of MLEs over all genes. This distribution is used as a prior on a second round of GLM fits, and the maximum a posteriori estimates are kept as final estimates of the log FC between groups.

Finally, the FC estimates need to be analysed for significance. For this, first, the Wald test is employed as in Cule et al. (2011). The method uses the test statistic $W_{g,m-m'}$, taking the $\beta_{g,m}$ estimates for gene g in conditions $m = 1$ and $m' = 2$, resulting in:

$$W_{g,1-2} = \frac{\beta_{g1} - \beta_{g2}}{\sqrt{\mathrm{Var}(\beta_{g1} - \beta_{g2})}}. \tag{3.13}$$

The Wald test asymptotically follows a standard normal distribution $\mathcal{N}(0, 1)$, so one can get the associated p-values as

$$p(W_{g,12}) = 1 - \Phi(\sqrt{W_{g,12}}), \tag{3.14}$$

where Φ is a cumulative distribution function of the standard normal distribution.

These p-values are then adjusted for multiple testing using the procedure by Benjamini and Hochberg (Benjamini and Hochberg, 1995) to obtain a *False Discovery Rate (FDR)* value, which serves as a more restrictive metric than the p-value for significance.

These two values, $\log_2$ FC and FDR, are usually taken as the main output of DEA methodologies. In fact, it is a combination of these metrics that produces the most recognizable plot for these analyses, the *Volcano plot*.

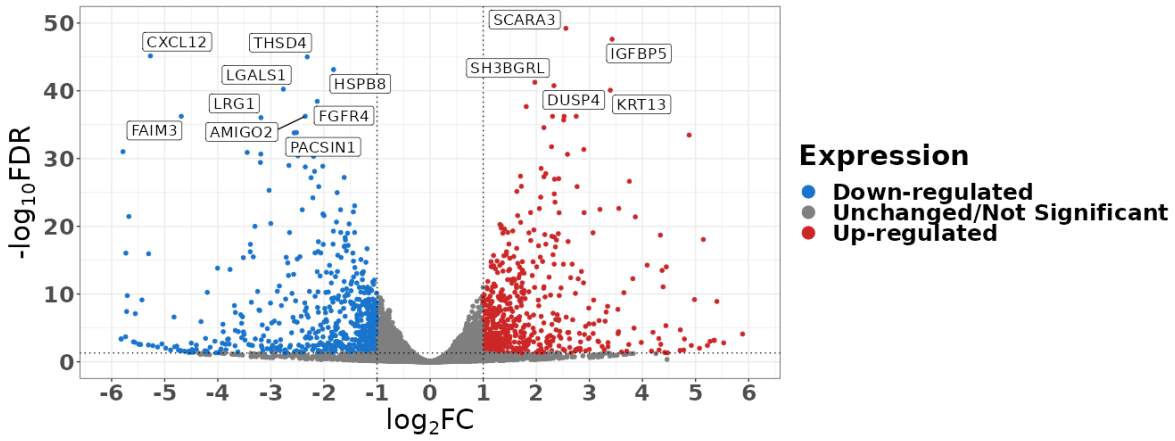

Figure 3.11: Volcano plot showing statistically significant differentially expressed genes.

The Volcano plot serves for easy identification of genes that are highly differentially expressed in a statistically significant manner. Figure 3.11 displays a two-group comparison between the control

MCF7 CTRL and MCF7_SOX2 conditions, as presented in the bG-RNA-Seq data. The x-axis displays the $\log_2$ FC while the y-axis usually displays significance (in this case $-\log_{10}(\text{FDR})$). Filters are used for both FC and FDR to remove genes that are not DE, or not deemed significant. These appear in dark grey in Figure 3.11 corresponding to genes with values of $|log_2\text{FC}| < 1$ or FDR < 0.1. The remaining genes, i.e., those that cleared the thresholds, are divided in blue (down-regulated) and red (up-regulated) groups.

Although this section just illustrates the most fundamental approach to DEA, the DESeq2 method includes plenty of corrections that increase the accuracy of the estimation of the dispersion and log FC. It also accounts for extreme cases where the number of samples is small, the number of genes with low counts is too high or the most highly expressed genes skew the distribution excessively. The original paper by Love et al. offers a more detailed view of these procedures (Love et al., 2014).

3.4 Functional analysis

The ever-expanding world of high-throughput sequencing offers a great variety of tools designed to extract the most information from the data available. Functional analyses of differential gene expression play a crucial role in extracting meaningful biological insights from high-throughput genomic data. While differential expression analysis identifies genes that are significantly altered between conditions, functional analyses provide a deeper understanding of the underlying biological processes, pathways, and functions associated with these gene sets.

In this section, we want to expand the scope of the analyses discussed so far, introducing some of the already available tools and platforms for functional analysis. This will help us establish a clearer picture regarding what information may be relevant from a biological point of view, which can guide us in this study. Besides, our focus is on interactive and easy-to-interpret analytical tools, since they play an important role in clearly communicating the importance of some key features or results.

3.4.1 Gene Set Enrichment Analysis

Many of the methods discussed in this section rely in some way or another on an enrichment score as defined by the Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005). This method takes as an input a set of genes that are over or under expressed in a certain biological condition compared to a contrast one, and determines if there is statistical significance based on the differences between two biological states (e.g. phenotypes). We try to explore ways in which these scores can be used to extract relevant information about the data in our hands.

3.4.1.1 GSEA — Classical analysis

The original Gene Set Enrichment Analysis software, GSEA, from the University of California San Diego, searches within pre-grouped sets of genes, for statistically significant differences between experimental conditions. For that, genes are first ranked based on a relevant metric (such as mean or differential expression), which captures the trend in gene expression changes across conditions. Then, a Kolmogorov–Smirnov enrichment score (ES) is calculated to represent the amount to which the genes in a set are over-represented at one of the extremes of a fold change ranking. The software uses its own gene set database, the Molecular Signatures DataBase (MSigDB).

After the calculation of the score, the software estimates its statistical significance. This calculation is done by shuffling random groups of genes and performing permutation tests in order to produce a null distribution for the ES. The p-value is determined by comparison to the null distribution. Calculating significance this way tests for the dependence of the gene set on the experimental or phenotypic labels in the analysis. The enrichment scores for each set are normalized (NES), and FDR is calculated using a Benjamini-Hochberg correction (Benjamini and Hochberg, 1995).

Despite this software being a reference in the field, the methodology for handling the data is not flexible and far from intuitive. Besides, the way it represents its results is in clear opposition to the objectives of clarity and accessibility that we intend.

To address these issues, we choose to perform GSEA using the R package fGSEA (Korotkevich et al., 2016), which implements in a fast an efficient manner a GSEA algorithm in an R environment, which is compatible with many DEA packages including DESeq2. As before, fGSEA takes a ranked list of genes as input and provides a Normalized Enrichment Score for the gene sets in the selected database. In this case, the user can select the database and one can choose again the MSigDB from the original GSEA, or resort to other known gene set or pathway databases such as the KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa and Goto, 2000) or Reactome (Croft et al., 2010) databases.

3.4.1.2 Enrichr — Bulk analysis

The last few years have seen a rise in the advent of all-in-one tools able to perform a multitude of analysis in one go. Similarly, online platforms have seen a major increase in numbers and functionality.

Enrichr is an online tool able to perform several enrichment analyses using transcription, pathways or gene ontologies data sets (Chen et al., 2013). The tool takes as an input a set of pre-identified genes and searches in its vast database for associations or presence of the genes in the set in pathway, cell line, transcription factor or disease sets. The most highly enriched sets for the inputted gene list provide knowledge about the possible relations between the inputted set and the biological features they can be associated to. For example, Figure 3.12 shows a compact view of several of the pathway databases available and the top altered pathways for the top 100 differentially expressed genes using as an example the MCF7 CRTL vs MCF7 TamR analysis from the cell line bG-RNA-seq dataset. When one of the databases is selected, an expanded view of that pathway analysis is shown.

Since there is no filter for the databases and barely no options for pre-computation ranking of the given set of genes, the program searches for relationship regardless of biological context. Counter-intuitively, the bulk analysis is much faster than classical GSEA, since the calculations involve only a p-value, z-score and simple combined score to assess relevance within the vast number of databases it has. The lack of information about the biological context of the sample leads to many results that should be treated with care to avoid creating spurious associations. However, it provides a quick and simple tool to search for associations of any kind which, after careful inspection, can highlight connections that may seem impossible to find otherwise.

3.4.2 Pathway analysis

The term Pathway Analysis serves as an umbrella under which several types of modern gene-data analysis can be found. We focus on the so-called *knowledge based-driven* pathway analysis (Khatri et al., 2012), which links the already known pathways catalogued in several databases with the statistically relevant genes found in a concrete experiment.

For its simplicity and visual impact, the most common way of representing a signalling pathway is a graph. Nodes represent genes and edges represent interactions between them. The representation is fairly similar for many of the biological processes involving pathways, functions or metabolic set, so analyses can be carried out using these graphs, with the same genes usually implicated in more than one mechanism.

In that context, EGSEA (Ensemble of Gene Set Enrichment Analyses) (Alhamdoosh et al., 2017) is a Bioconductor package in R that combines the analysis results of several standard GSEA algorithms to calculate collective significance scores for gene sets. The functionalities of the Pathview package,

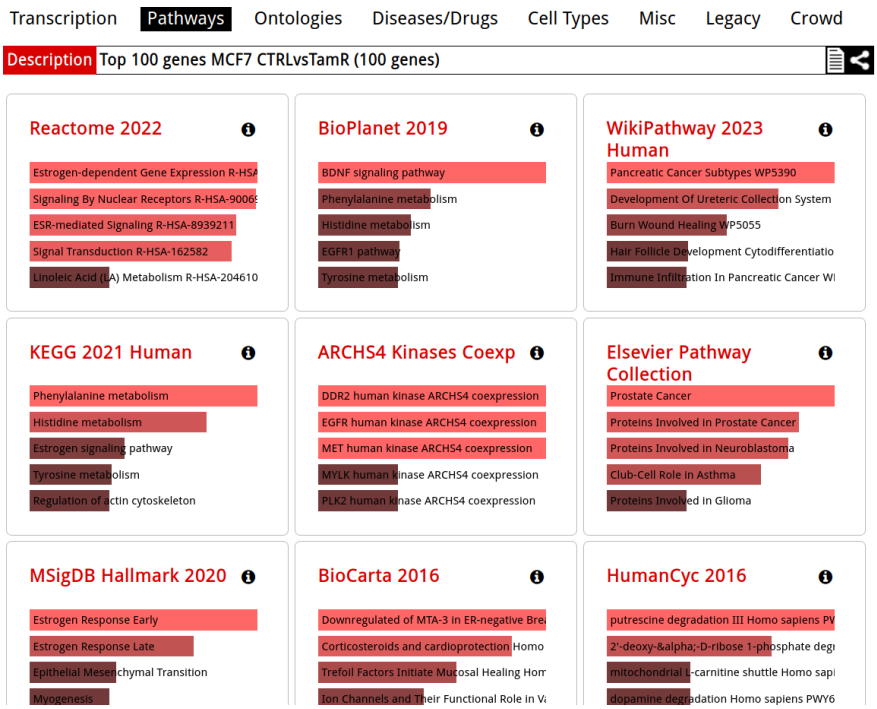

Figure 3.12: Example of Enrichr view for the MCF7 CRTL vs TamR analysis.

also part of Bioconductor, are integrated inside EGSEA so that the results obtained can be displayed on top of the pathways available in the most common databases (KEGG, MSigDB, Reactome...).

The integration of the packages results in a graph plot of highly altered pathways, where the colour of each node highlights the enrichment and significance score of each gene. However, the computational cost of this type of analysis increases heavily with the number of pathways analysed. Since EGSEA has access to all the major databases for functional annotation, the user has to specify a maximum number of pathways to be explored to maintain a good balance between computational effort and relevance of the result.

For the number of pathways selected, the package indicates in a table how many genes are altered in each one, alongside the average $\log_2$FC of these genes, as well as the enrichment score and the corresponding p-value. By selecting a pathway from this table, the data can be seen as the detailed pathway's graph, as shown in Figure 3.13. This allows for a quick visual exploration of the altered sections of the pathway, as well as the direction of this alteration.

One of the key advantages of EGSEA is the possibility of extracting an HTML analysis report, allowing for an interactive and more in-depth exploration of the results. It also makes the result easier to share and program-independent, as the HTML report is stored locally and can be opened in any browser. A downside of this approach is that the generation of the report is computationally expensive and time increases with the number of analysed pathways. But the report is only generated once and allows performing quick checks for parameter tuning without the computational burden of the full analysis.

3.5 A practical application: Vivanco LabSeq tool

The use and exploration of the methodologies and techniques shown so far resulted in the development of a web-app that condensed the analyses carried out on the available data into a single platform. The main motivation for the creation of this web-app was to enhance the accessibility and interest for exploring bioinformatic results. Some of the techniques or analyses that we showed throughout

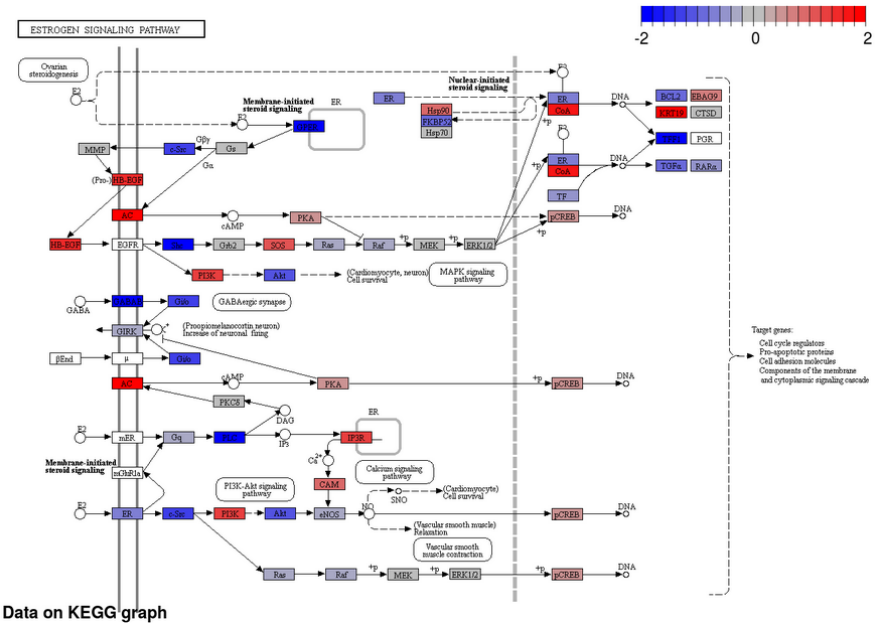

Figure 3.13: Alterations to the KEGG estrogen signaling pathway using the MCF7 CTRL vs TamR analysis. Blue is chosen for an average negative expression — meaning expression is down in the TamR condition —, while red represents genes upregulated in that node in the resistant condition.

this section may seem obscure or too dense for people that are not used to work with these tools or methodologies, so we constructed the platform over three key pillars:

1. **Accessibility:** The platform should be easy-to-use. Access and navigation should be straightforward.
2. **Clarity:** The platform should contain exactly the elements needed to perform the intended analyses and relevant information should accompany the tools.
3. **Visualization:** The results should be presented in an easy-to-interpret format, while data for more in depth analysis should be made available.

Hence, the platform of choice for the development of this web-app was R-Shiny (Chang et al., 2024), which allows for the accessibility and easy-to-use features that we were looking for. At the same time, this package for web development is native to R where most of the packages for analysis presented before are hosted.

The web offers access to the data and several analyses of the in-house cell line bG-RNA-seq data from section 3.2.1.1. This tool was designed for internal use of members of the Cancer Heterogeneity Lab. Its development was centred around extracting data from the cell lines employed in their research, aiming to provide support to their ongoing studies. However, the core of the platform could easily be adapted to host external data.

We included nearly all the analyses presented in this chapter. The most noticeable absentee in the web-app is the pathway analysis. As mentioned before, it is too cumbersome and computationally expensive to introduce in the platform. To account for this, a separate HTML analysis report using Pathview was created to provide this information to all group members.

Now, we illustrate the different parts of the platform using the MCF7 CRTL and MCF7 TamR cell lines from the bG-RNA-seq dataset. The MCF7 CRTL vs MCF7 TamR experiment is relevant and important for this thesis, and it forms the cell dataset for the cell-patients integrated data to be considered in chapter 4.

Home page

The Vivanco LabSeq web-app can be accessed through https://vivancoslabseq.shinyapps.io/RNASeqSOX/. The Home page provides a ***control menu*** on the left of the screen and a panel with several tabs where results are displayed. The ***Instructions*** tab is pre-selected and shows a guide for the various options the app provides.

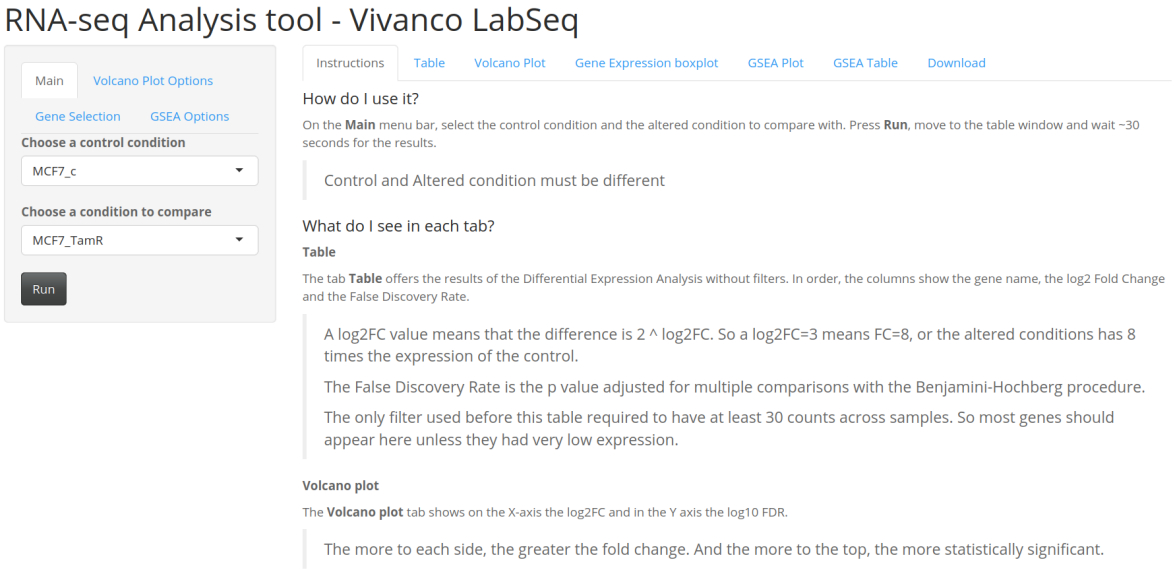

Figure 3.14: Home page.

The ***Main*** tab of the ***control menu*** has two dropdowns boxes to select the biological conditions to be analysed: "*Choose a control condition*" and "*Choose a condition to compare*". The options one can select are the cell lines in the bG-RNA-Seq dataset introduced in section 3.2.1.1. To perform an analysis, one should only select two compatible conditions (all altered conditions need to be compared using the appropriate control) and press the ***Run*** button to start the calculations. In a few seconds, the results are ready to display, and the user only needs to select the tab above the middle panel corresponding to the desired output (i.e., ***Table***, ***Volcano plot***, ***Gene expression boxplot***...). The options in the ***control menu*** (*Volcano plot options*, ***Gene selection*** and ***GSEA Options***) can be used after the computation has been done and allow real-time changes in the visualization of the results.

In the ***instructions*** tab, the middle panel provides the guidelines for each of the tabs, i.e., ***Table***, ***Volcano Plot***, ***Gene Expression Boxplot***, ***GSEA Plot***, ***GSEA Table*** and ***Download***; as well as the options within each of them.

DEA analysis

The ***Table*** tab displays the results of the Differential Expression Analysis performed using DESeq2, following the methodology explained previously in section 3.3.3.

The columns show, from left to right, the Gene Name, $\log_2$ Fold Change and the False Discovery Rate of the selected analysis. The table is automatically ordered by smallest FDR (most significant). The only filter to obtain this table was the requirement to have at least 30 counts across samples. This means the majority of genes should appear here unless their expression was extremely low in the selected conditions. One can search for a given gene of interest in the search bar in the upper right corner of the table.

The DEA results from the table can be visualized using a Volcano plot, which displays the $\log_2$ FC in the X-axis and the $\log_{10}$ FDR in the Y-axis. It can be accessed through the ***Volcano plot*** tab. The plot is interactive, the axis can be adjusted to fit any result, and the number of genes displayed can be

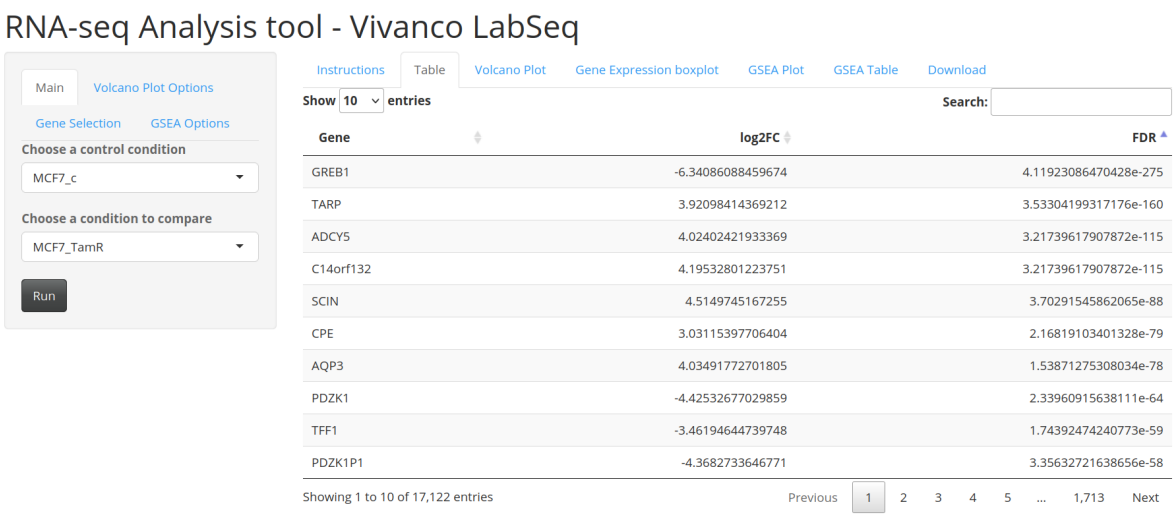

Figure 3.15: Table with DEA results for the MCF7 CTRL vs MCF7 TamR study.

filtered according to user preference with the help of the ***control menu*** on the left. The ***Volcano Plot Options*** tab displays the following plot options, from top to bottom:

1. **Number of highlighted genes in the Volcano plot:** Adjusts the number of labelled genes in the Volcano plot. It may not show all the requested genes if the labels don't fit in the plot.
2. **Statistical significance threshold (for FDR):** Controls how many genes are considered significant. Moves the line in the Y-axis that represents this threshold.
3. **Log2FC significance threshold (absolute log2FC):** Defines the minimum $\log_2$ FC to be considered significant. If 0, all genes are coloured (significant). If bigger, the vertical dotted lines show the thresholds.
4. **Regulate X axis display:** Adjusts the visible range of the X-axis in the plot. Can be used to zoom in/out. If zoomed in, request for more labels (1) to see the names of the new genes.
5. **Regulate Y axis display:** Adjusts the visible range of the Y-axis in the plot. Can be used to zoom in/out.

Finally, there is an option to retrieve the expression values of a given gene, to check for total amounts of expressions levels (normalized). For that, the ***Gene Expression boxplot*** tab in the middle panel should be selected, which displays a boxplot comparing the normalized counts in the two selected conditions for the gene inputted in the option menu ***Gene Selection*** in the control panel, as shown in Figure 3.17.

GSEA analysis

The ***GSEA Options*** tab in the ***control menu*** panel gives the user an option to select with a dropdown menu a database for GSEA analysis among the databases KEGG (Kanehisa and Goto, 2000), MSigDB Hallmark gene set (Liberzon et al., 2015), Reactome (Croft et al., 2010), Gene Ontology (Ashburner et al., 2000) and one with Common Oncogenic pathways (Sanchez-Vega et al., 2018). The number of sets to be displayed in the plot can also be selected from this panel.

The ***GSEA plot*** in the middle panel tab shows the Normalized Enrichment Score (NES) per pathway as seen in Figure 3.18. It reflects the degree to which a gene set is over-represented in the list of genes concerning a pathway. To be comparable, they are normalized to the size of the gene set.

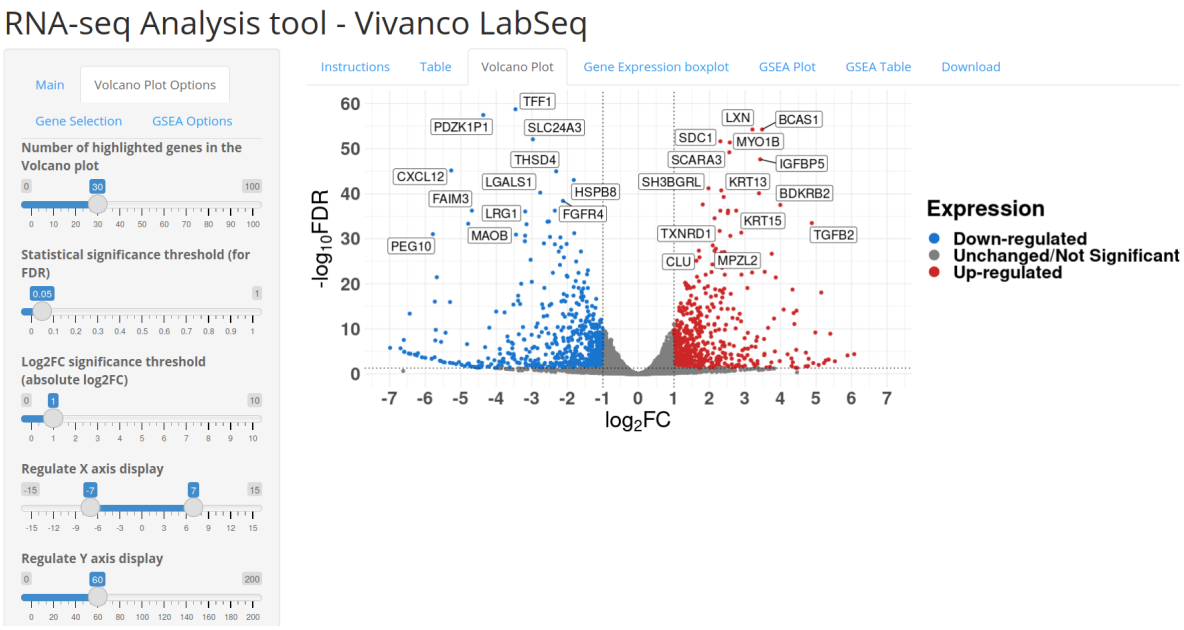

Figure 3.16: Volcano plot highlighting several genes significantly differentially expressed in the MCF7 CTRL vs MCF7 TamR study.

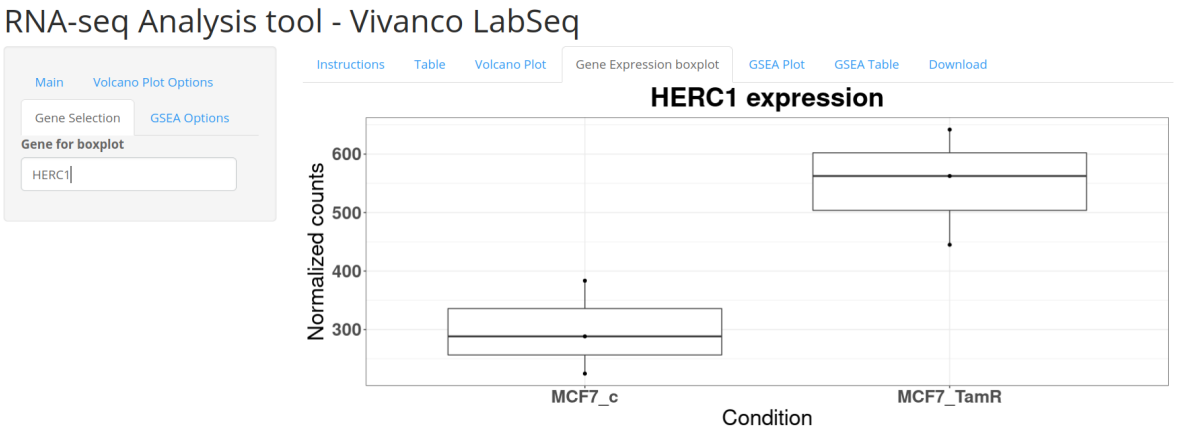

Figure 3.17: Boxplot showing the expression changes between MCF7 CTRL and MCF7 TamR conditions of HERC1, a gene found to be relevant for tamoxifen resistance in our mathematical models (see Chapter 4).

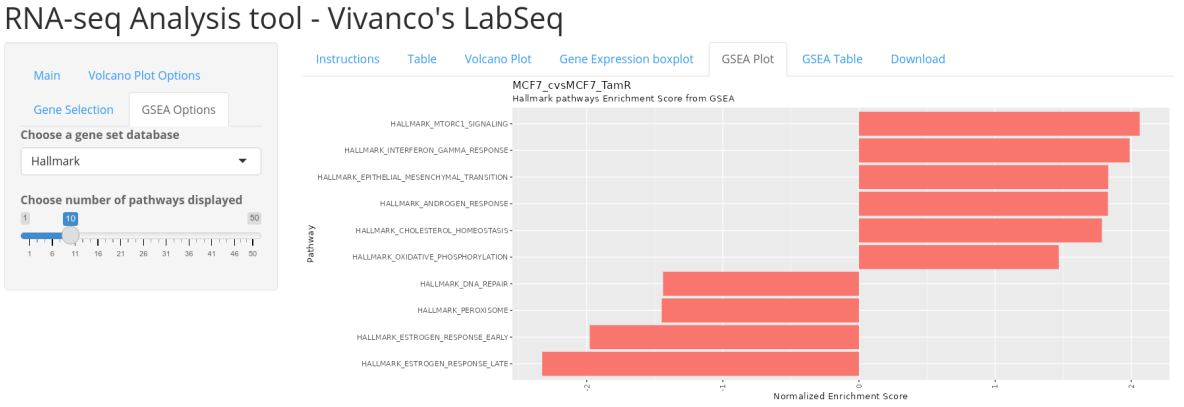

Figure 3.18: GSEA results using the KEGG platform.

The ***GSEA table*** tab in the middle panel (Figure 3.19) displays the results of the Gene Set Enrichment Analysis in table format. In order from left to right, it displays the pathway name, the NES and the FDR.

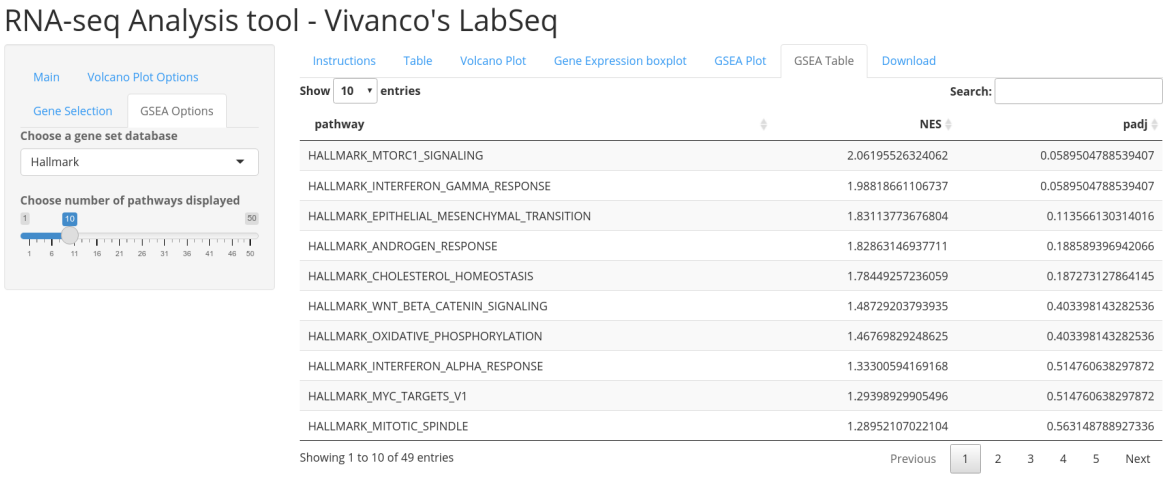

Figure 3.19: GSEA table.

Download page

Finally, the ***Download*** tab in the middle panel allows the user to download the tables for the DEA and GSEA analyses for personal use or application in other projects (Figure 3.20).

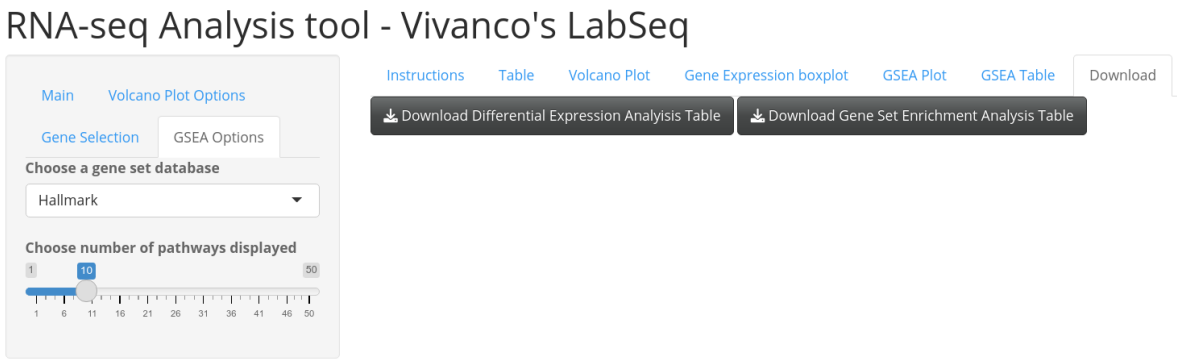

Figure 3.20: Download page.

3.6 Conclusions

In this chapter, we have introduced the most useful techniques for normalization, differential expression and functional analyses of RNA-seq data. As in the previous chapter, we present here two *ad hoc* datasets to be used in this study. The cell line bG-RNA-seq dataset served as the basis for all analyses in this chapter, while the TCGA patients RNA-seq, was presented here but will be mainly used in the next chapter.

The central part of this chapter was the development of Vivanco LabSeq web-app, which can be framed within Objective 1 in the context of the thesis. The web platform was created to allow the users from the Cancer Heterogeneity Lab to perform several types of DEA (Differential Expression Analysis) and GSEA (Gene Set Enrichment Analysis) of the pre-processed data related to the problem of resistance to tamoxifen in breast cancer therapy. The web-app was built with the aim of bridging the gap between experiments and computational biology, by offering to the researchers at the Cancer

Heterogeneity Lab an easy-to-access platform, with clear information and options for the visualization of the results.

Today, the platform is frequently used within the group to inspect gene behaviour (over/under expression) in the collection of cell lines employed daily for experiments at the lab. It provides a tool for discovering altered pathways that may point towards interesting experiments directions or novel approaches for researching the problem of resistance to tamoxifen. Finally, the plotting tools and the options within them, offer quick solutions for creating visual representations of gene expression and pathway enrichment to add a layer of computational validation to results from other laboratory experiments.

Chapter 4

Integration of Clinical and Transcriptomic Data for Biomarker Identification

In this chapter, we intend to implement the previously presented knowledge in a challenging and realistic scenario, where two different types of datasets are combined with the aim to identify a set of genes with significant impact on the problem of patient resistance to endocrine therapy. We start with a brief discussion of the context in which we frame this study, and proceed with an overview of the state of the art as well as an outlook on avenues for progress in the discovery of prognostic signatures in cancer research. We then present the data available and our proposed methodology to approach the task. The result is the identification of a new gene signature related to the problem, and its independent validation *in-silico* and *in-vitro*, to demonstrate its impact. We close this chapter with a brief discussion on the significance of our results.

In section 1.2, we outlined the context of breast cancer, focusing on the prevalence of the ER+ subtype and introducing the challenge of endocrine therapy resistance. In Chapters 2 and 3 we discussed various methodologies and biomedical datasets which can contribute in tackling the endocrine therapy resistance problem. Here, we take on the challenge of combining several of these techniques and sources of data in a way that the information from one source can add value to the rest. Connecting RNA-sequencing data with clinical information and outcomes has been lately a subject of study (Polewko-Klim et al., 2021; van Vliet et al., 2012), even from a Bayesian perspective (Gevaert et al., 2016). But there is still a lot of unfilled potential, giving the increasing amount of data available, and one of the ultimate goals should the development of robust methodologies allowing the efficient integration and analysis of the multi-platform data for identification of potential cancer biomarkers (Brueffer et al., 2018; De Sanctis et al., 2018).

We aim to integrate the data extracted from laboratory experiments and clinical analysis in order to find a genetic signature that could be linked to the acquisition of resistance to tamoxifen treatment in breast cancer patients. The sequenced cell lines and the controlled environment of sequencing will provide insights into the problem and a help to develop and study resistant conditions, while the clinical data will serve as a confirmation tool to translate laboratory findings into the clinical stage.

4.1 Current landscape

Acquired genomic changes, including losses of and mutations in the ESR1 gene (Jeselsohn et al., 2015; Priedigkeit et al., 2021), have been identified in long-term estrogen-deprived resistant tumours, although such mutations were found in only 15-20% of patients (Szostakowska et al., 2019), the

majority of whom received treatment with aromatase inhibitors. This reflects that, for most tumours, the acquisition of resistance occurs despite the expression of ER. A lot of research in hormone therapy (HT) resistance has been conducted either through cell line studies or Next Generation Sequence (NGS) analysis of data extracted from patients in clinical trials or from public repositories. On one hand, widely studied cell lines (such as MCF7) can be repeatedly assayed and modified in a variety of ways and hence allow for controlled experiments and reproducible behaviour. However, they may fail as models of an undoubtedly heterogeneous disease (Koren and Bentires-Alj, 2015). On the other hand, clinical data accounts for this heterogeneity, but can be susceptible to external influences or subjective descriptions that may affect its interpretation and can be difficult to identify or isolate.

Cell line studies in the field of resistance to anticancer therapies have already identified several relevant families of genes and signalling pathways. Among them, the SOX (Domenici et al., 2019; Piva et al., 2014) and Interleukin (Sarmiento-Castro et al., 2020) families, Notch (Magnani et al., 2013; Simoes et al., 2015) or Wnt (Piva et al., 2014) pathways and cell proliferation (Gao et al., 2014; Huang et al., 2011; Palafox et al., 2022). In addition, clinical studies have been successfully employed for the discovery of prognostic signatures in breast cancer. Among them are the widely used 70-gene signature Mammaprint (Van't Veer, 2002), or the 21-gene signature OncotypeDX (Paik et al., 2006), which used real-time reverse-transcriptase–polymerase-chain-reaction (RT-PCR) to analyse tumours from patients. These prognostic signatures serve for predicting metastasis or recurrence for some breast cancer subtypes.

In more recent years, prognostic signatures have been proposed for a wide range of breast cancer subtypes and clinical scenarios, and several studies have tackled specifically the problem of resistance to endocrine therapy. For instance, some authors explored both RNA and DNA mutations using patients sequencing data (Xia et al., 2022). Moreover, access to new gene editing methods have led to the proposal of a 6-gene signature extracted from the study of mutant cell lines where the ESR1 gene was modified using CRISPR-Cas9 (Harrod et al., 2022). The signatures introduced for general endocrine resistance nominated the MAPK pathway (Miller et al., 2015) or ER-PR related genes as prognostic predictors (Sinn et al., 2019). For tamoxifen resistance, lists of individual markers (Hermawan et al., 2020; Mihály et al., 2013; Wang and Wang, 2021) and ratios of gene expression (Ma et al., 2004) were identified as potential risk factors. The impact of some key pathways in the development of resistance to tamoxifen was also investigated, leading to the identification of a signature of 5 pathway-representative genes (Rahem et al., 2020).

However, some concerns have been raised recently about the *significance* and *utility* of gene signatures in breast cancer (Goh and Wong, 2018; Manjang et al., 2021; Venet et al., 2011). These studies suggest that there are three key factors that generate seemingly significant but not necessarily meaningful prognostic signatures:

1. Correlation with proliferation markers - such as meta-PCNA from Venet et al. (2011).
2. A lack of statistical significance, that can be enhanced by a poor validation of the signatures.
3. A high number of genes in the signature — over 25 genes according to Goh and Wong (2018).

Therefore, short predictive signatures that do not rely on proliferation biomarkers, which are generally indicators of bad prognosis for any cancer type, might be more relevant as they identify problem-specific genes. On top of that, independent and systematic validation studies are vital for generation of biologically significant gene signatures, as seemingly random sets of genes may achieve significant results in clinical outcome prediction without any biological context by overfitting to a training dataset.

Some efforts have also been made to integrate both cell line data and clinical sources as a strategy for discovering potential biomarkers related to the problem of resistance at the level of individual

genes (Men et al., 2018) or complete genomic profiles (Asghari et al., 2022). These studies improved understanding of the impact of individual biomarkers in the development of resistance but failed short of providing any advances in terms of resistance prediction.

4.2 Methodology

Here, we present a methodology to discover gene signatures and relevant biomarkers, which we apply to the problem of resistance to endocrine therapy in breast cancer. It takes advantage of experimental data generated from cell models (using the well-characterised MCF7 CTRL and MCF7 TamR cells as controls for ER+ breast cancer and resistance to tamoxifen, respectively) and from patients (publicly available transcriptomic and clinical patients information, from the TCGA). Our approach includes classification of these data, statistical analyses of common genes and validation. Thus, unlike previous attempts of using combinations of cell lines and patients data, this study represents, to the best of our knowledge, the first attempt to go beyond the mere identification of common traits between cell and patients data and offers a thorough computational and statistical analysis of the identified genes and their behaviour as a gene signature. All this is backed up with independent validations and further confirmed by experimental evidence. Furthermore, the identified 6-gene signature is shown to be useful when extended beyond tamoxifen to any kind of hormone therapy.

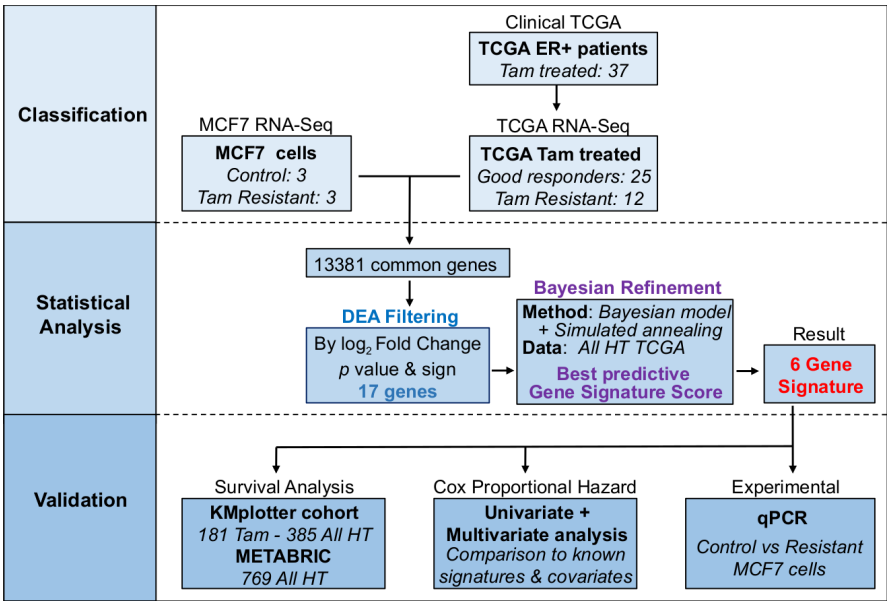

Figure 4.1: Schematic representation of the proposed work pipeline. The data collection and data treatment are followed by the joint analysis of cell and patients data, and the refinement of the results. The validation process relies on statistical analysis and experiment.

Armed with the tools we have seen in previous chapters, here we will explore the integration of cell and patients data. Our methodology can be divided in three steps, illustrated in Figure 4.1.

STEP 1: Data collection and classification

In this first step, we collect and treat data from cell and patients sources, and classify it into resistant and respondent to treatment. For this study, the cell line data is naturally classified in this sense, as we use the MCF7 CTRL and MCF7 TamR cell lines from the bG-RNA-seq dataset. The MCF7 CTRL responds well to tamoxifen treatment, while the MCF7 TamR is resistant to it. For the patient data, we use the TCGA cohort of tamoxifen treated patients, which does not offer this information explicitly and, as such, patients need to be classified into resistant and respondent to treatment groups in order

to perform differential expression analysis between them. We explain this in detail in section 4.3. The next step is to perform the DEA analysis on both cell and patient data to identify the common genes differentially expressed in both analyses. Those genes (13381 in this study) are then considered as likely related to the condition shared by both groups, i.e. resistance.

STEP 2: Statistical analysis and refinement of the common relevant genes

Next, the identified set of common 13381 genes is refined by setting up thresholds for False Discovery Rate (FDR) and $\log_2$ Fold Change and selecting the genes considered significantly altered in both, cell and patient DEA. This results in a smaller set of genes (17 in this work) which we use to perform an advanced search for the signature within this subset with the best predictive power. The selection of such a signature is performed using an original simulated annealing type algorithm BLR-SA (see for details section) combined with the Leave-One-Out (LOO) cross validation (Vehtari et al., 2017) incorporating an HMC-based sampling. In this particular study, we rely on a standard HMC algorithm (as presented in 2.2.3) as implemented in Stan (Stan Development Team, 2024) rather than on a more efficient MMHMC with s-MAIA (developed in 2.3.2) available in the in-house software package HaiCS only. The reason for that is the ready-to-use implementation of the LOO algorithm in Stan, whereas such a method is yet to be implemented in HaiCS.

STEP 3: Validation of the results

The last step includes multiple validations of the identified gene signature. We propose to use survival analysis and Cox proportional hazard models with various independent validation cohorts of hormone therapy resistance patients. In this study, we consider two independent cohorts extracted from Gene Expression Omnibus and METABRIC databases (see for details 4.5) that gave us over 1000 new ER+ patients and another 1000 ER- patients to use as negative control. Besides the *in-silico* validations, we also suggest including in this step *in-vitro* validation such as, for example, qPCR experiments presented in this work.

4.3 Data collection and classification

To find a genetic signature that could be linked to the acquisition of resistance to tamoxifen treatment in breast cancer patients, we use two datasets already introduced, the bG-RNA-seq data, from which we consider only MCF7 CTRL and MCF7 TamR cell lines, and the TCGA clinical datasets. From the TCGA dataset, we rely on both the clinical and RNA-seq patients data.

4.3.1 MCF7 cell lines

For the purposes of this thesis, we focus our analyses on those lines that are most directly related to the critical question regarding the mechanisms involved in resistance to hormone therapy in ER+ breast cancer. These candidates are the MCF7 CTRL and MC7 TamR, which represent the control breast cancer cell line and the cell line with naturally developed tamoxifen resistance.

The cells used for this experiment were obtained from the American Type Culture Collection. MCF7 tamoxifen-resistant (MCF7 TamR) cells were previously developed by exposing cells to tamoxifen for over 6 months to acquire resistance to treatment and, thus, were used as a model of resistance (Piva et al., 2014). Both cell lines were cultured in Dulbecco's Modified Eagle medium supplemented with 8% fetal bovine serum and 1% penicillin/streptomycin at 37 °C in 5% CO2.

Gene expression profiles of both MCF7 CTRL and MCF7 TamR cells were obtained by RNA-seq. Total RNA was extracted with the RNeasy Mini Kit (Qiagen). The quantity and quality of the RNAs were evaluated using the Qubit dsDNA Assay Kit (Life Technologies, Cat.# 32855) and Agilent RNA Nano Chips (Agilent Technologies, Cat.# 5067-1511), respectively. Sequencing libraries were

prepared using the "TruSeq Stranded Total RNA with Ribo-Zero Human/Mouse/Rat" kit (Illumina Inc., Cat.# RS-122-2201), following the *TruSeq Stranded Total RNA Sample Preparation Guide* (Part # 15031048 Rev. E). RNA extraction was performed by the Cancer Heterogeneity Lab while library preparation and sequencing was carried out by the Genomics platform, all within CIC bioGUNE. Libraries were sequenced on HiScan-SQ platform (Illumina Inc.) by single-end 50 nucleotides reads. Gene expression profiles were compared between control MCF7 cells and resistant MCF7TamR cells using a DESeq2 pipeline as presented in section 3.3.3 in the previous chapter. Three biological replicates were considered per cell line and processed as 3 independent experiments.

4.3.2 TCGA patients dataset

4.3.2.1 Selection of clinical cases

Breast cancer patients with hormone-dependent tumours that received endocrine therapy were selected from the TCGA database. The database from the TCGA contains detailed clinical and RNA-seq data collected from over 1.000 breast cancer patients of all types. Hence, RNA-seq and clinical records of patients were retrieved from the clinical biospecimen core resource XML files of the TCGA-BRCA (The Cancer Genome Atlas-Breast Invasive Carcinoma) using the Firebrowse platform (http://firebrowse.org/), where the raw count data was accessible. Initially, a total of 209 ER+ & tamoxifen-treated patients were identified in the TCGA breast cancer dataset as potential candidates to be included in the study, as discussed in section 2.1.2.

The heatmap in Figure 4.2 shows the top differentially expressed genes for the TCGA cohort taking all tamoxifen-treated patients before any cleaning. Here, cancer status in the last medical follow-up available ("with tumour" or "tumour free") was used as the sole indicator for resistance to treatment and helped to divide the cohort into two groups. No clear clustering or major differences in terms of transcriptomic profiles can be seen in Figure 4.2 using this analysis.

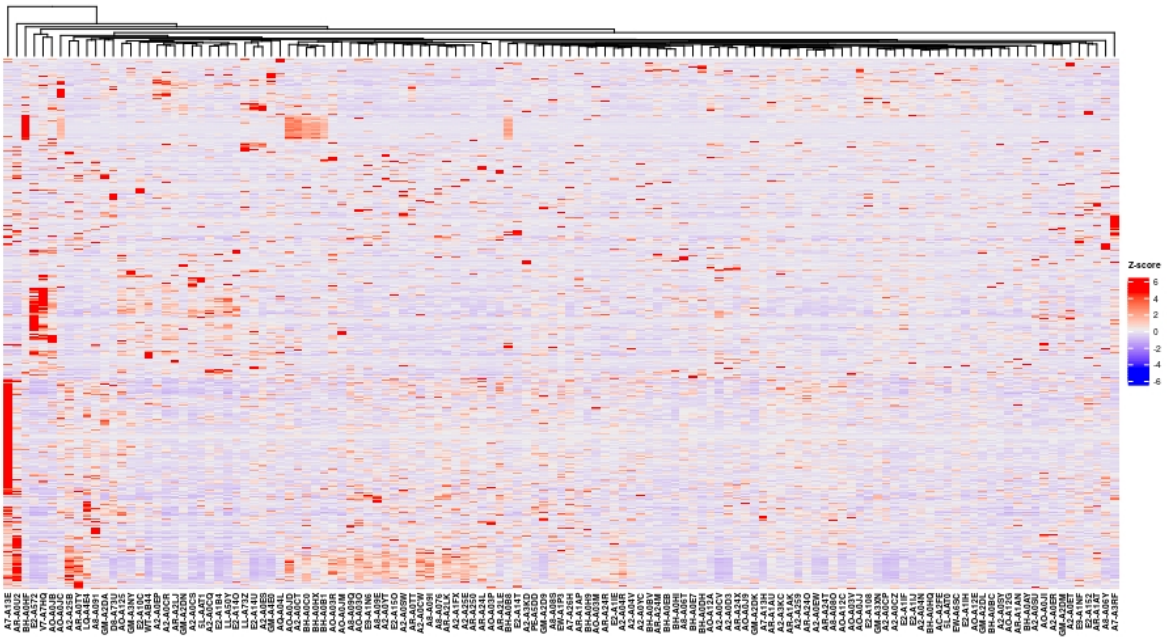

Figure 4.2: Heatmap of all the ER+ & tamoxifen-treated patients before cleaning. The patient's last follow-up status was used to separate groups in the DEA analysis, although no clear structure in the data can be detected.

Following the methodology presented in section 2.1.2.2, the initial selection was curated to choose, among all ER+ patients, those most suitable for the analysis. This resulted in a list of 37

tamoxifen-treated patients, increasing to 127 cases when all types of hormone therapy were included. A summary of the clinical information for these patients can be found in Appendix B.

4.3.2.2 Classification of patients responses

The outcome variable is the most crucial feature for classification models, as it dictates the group to which a given instance (in this case, a patient) belongs. For this case, resistance to treatment is not explicitly defined in most clinical records, with only some records specifically mentioning the patient to be "tumour free" or "with tumour" at the end of a given treatment. However, resistance to a treatment can be extrapolated from a poor progression of the tumour and a worsening of the patient's condition even after a long treatment time.

To account for this, a classification tool for endocrine resistance based on clinical information was created, allowing us to expand the number of available patients compared to other studies using TCGA (Men et al., 2018). The core of the developed classification tool is a Matlab script that analyses TCGA patients information and translates it to a timeline format. Several examples of this are presented in Figure 4.3. These timeline plots allow the detection of patients and treatments that, despite clearing the selected criteria, showed problematic behaviour in the context of the study, i.e, interrupted or incoherent treatments deviating from a regular tamoxifen schedule. The plots served as a visual representation of the evolution of each treatment and aided in the cleaning and classification process.

For a patient to be labelled as resistant, we established the following rule. The clinical history of a resistant patient after sustained treatment had to meet one of the conditions below:

- Disease progression.
- New tumoural event.
- Death during treatment.

Here, *disease progression* was assigned to a patient with persistent tumours after the last follow-up following sustained treatment. *New tumoural event* included the appearance of both metastatic events and tumour recurrences. Finally, *death* included all patients that died during the time window of the study. TCGA only states the death of a patient and not the cause and therefore it is not possible to accurately determine if the cancer itself was the main cause of the decease. However, such an extreme outcome could not be easily weighted down. Nevertheless, the classification tool shows that all deaths occurred either during treatment or shortly after relapse or a dissemination of the disease appeared. Therefore, we have no indicators suggesting that any of those cases should be excluded.

This methodology yielded a curated list of patients, including 25 patients that responded well to tamoxifen treatment and 12 deemed to be resistant.

As a reaffirmation of the validity of this classification, Figure 4.4 shows a heatmap for the reduced cohort of 37 patients selected and classified as resistant and good responders. Compared to the heatmap in Figure 4.2 we now see that the groups are much better defined. The heatmap clearly presents two distinct profiles, and the dendrogram over the plot suggest a similar division to our own. Most of the patients we identified as resistant cluster on the left side of the heatmap, with only 2 of their counterparts (highlighted within a box) not adhering to the same group but presenting similar transcriptomic profiles.

4.4 Statistical analysis of sequencing data

With the patients classified, we have all the ingredients to apply our methodology. For that, we will first create a shortlist of relevant genes common to both datasets. We will then exploit our previous

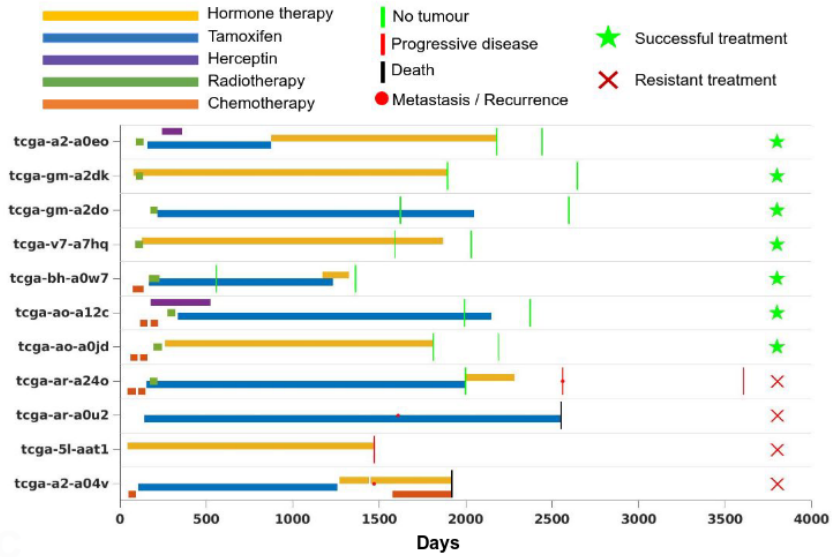

Figure 4.3: Timeline visualization of TCGA patients treatment data. Each row in the y-axis shows the major events in the treatment of a patient. Patients with a star are responded well to the treatment and patients with a cross showed resistance to treatment.

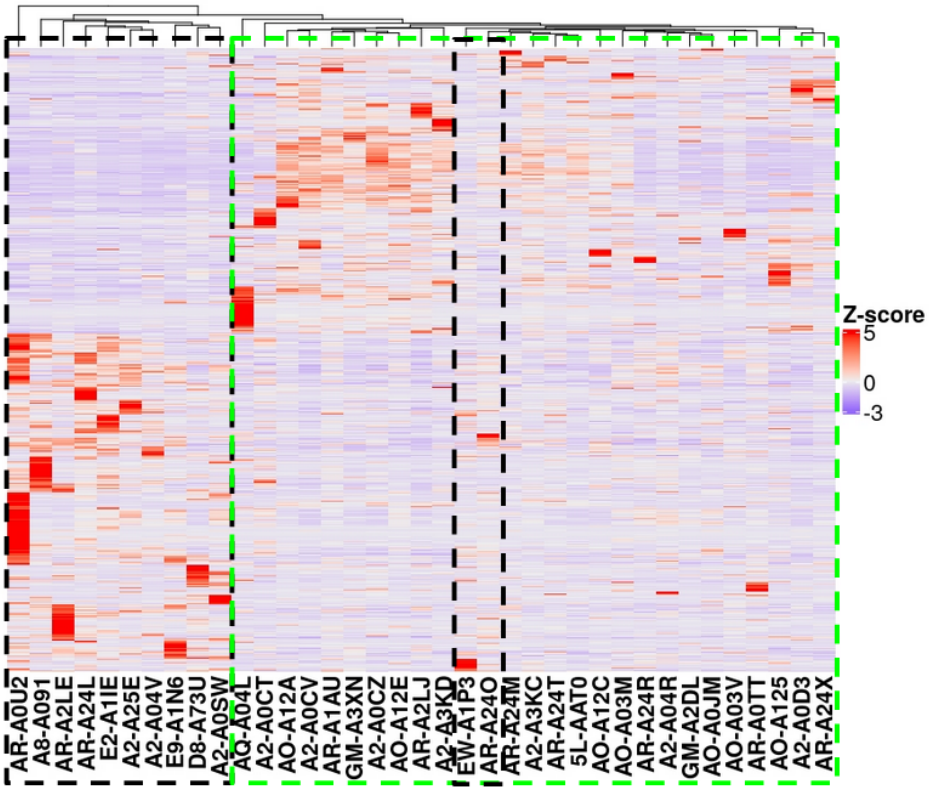

Figure 4.4: Heatmap of the differentially expressed genes in TCGA-resistant patients (black dotted box) and well-responded patients (green dotted box) from the TCGA dataset. The rightmost black box highlights the only 2 patients labelled as resistant by our classification scheme that do not cluster with the rest.

experience with BLR models to predict resistance using subsets of this gene list. With the original simulated annealing-type algorithm combined with the predictive model, we will be able to obtain the gene signature of the best prediction power.

4.4.1 Identification of common biomarkers

Following the DESeq2 pipeline presented in the previous chapter, i.e., RLE normalization and DEA as in 3.3.3, we performed differential expression analysis of the cell lines and patients RNA-seq data. The results of the DEA analysis were used to identify potential candidate genes related to resistance using both patients and cell lines data. Hence, two separate two-group DEA were performed — one, using the cell lines from the **MCF7 CTRL vs MCF7 TamR** conditions, and another one with the TCGA tamoxifen-treated patient dataset taking **Good Responders vs Resistant Patients** (following the classification presented in the previous section). To perform the DEA, genes with less than 30 counts across patients or cells were excluded from the analysis beforehand and counts were then normalized using the RLE method implemented in DESeq2 as discussed in 3.3.1.

For each two-group comparison, results were filtered according to False Discovery Rate (FDR), and $\log_2$ Fold Change so that only genes with FDR < 0.1 and $|\log_2 \text{FC}| > 0.5$ in both analyses were selected. The significance filter for FDR was slightly raised from common practice values due to the heterogeneity present in the TCGA samples, as all the selected genes cleared the threshold for the DEA of the cell lines data. For the next filter, we selected only those genes that were expressed in the same direction in both cases. This stems from our assumption that patients with a positive clinical response were considered comparable to MCF7 control cells, meaning that they are sensitive to hormone therapy. Similarly, MCF7TamR cells can be considered as models of resistance to tamoxifen, as observed previously in breast cancer patients (Piva et al., 2014). As MCF7TamR cells were already resistant when sequenced, such a grouping arises from our hypothesis that alterations leading to development of resistance might already be present in tumours before treatment, when they were sequenced.

The effect of these filters can be seen in Figure 4.5. Clearly, the number of relevant genes is reduced by sequentially applying the criteria for Fold Change, sign and FDR. The final selection of genes, those with similar behaviour in both sets, led to the set of 17 genes highlighted in panel D of Figure 4.5.

Before any further research, we explored the 17 genes extracted from the joint analysis for possible correlations with proliferation markers, as such correlations could potentially make a signature to be statistically significant without conveying any new biological information (Venet et al., 2011). In fact, prognostic signatures are often driven by proliferation markers (for example, in Oncotype DX the proliferation term is the most relevant for the calculation of the recurrence score). Thus, we checked for links with proliferation in the identified gene set using the methodology recommended by Venet et al. No link to proliferation was found in the identified set of 17 genes.

In addition, we used survival analysis to test whether **any individual gene**, **all up and down-regulated genes separately**, or **the 17 as whole**, could be used as predictors of resistance. For that we employed the KMplotter tool and database[1]. No statistically significant prediction of clinical outcomes (in terms of Relapse Free Survival) could be extracted from the performed tests.

4.4.2 BLR-LOO model

To refine the list of potential biomarkers obtained from this combined DEA-based analysis of patients and cell lines, we used the Bayesian logistic regression model presented in section 2.3.2.6. The model received as an input a Gene Signature Score (GSS) value 4.1, and classified patients according to their proneness to resistance. The GSS for each TGCA patient was estimated using the mean value of the Z-scores across all genes in the signature, separately for up- and down-regulated genes, similarly to other methodologies previously used in Hsiao et al. (2013). To obtain the Z-scores for each patient i and gene n we used the $\log_2$ transformed gene counts $(g_{i,n})$, normalized by extracting the mean

[1]The tool can be found in https://kmplot.com/ and a more detailed explanation of the platform and its data is provided in section 4.5.

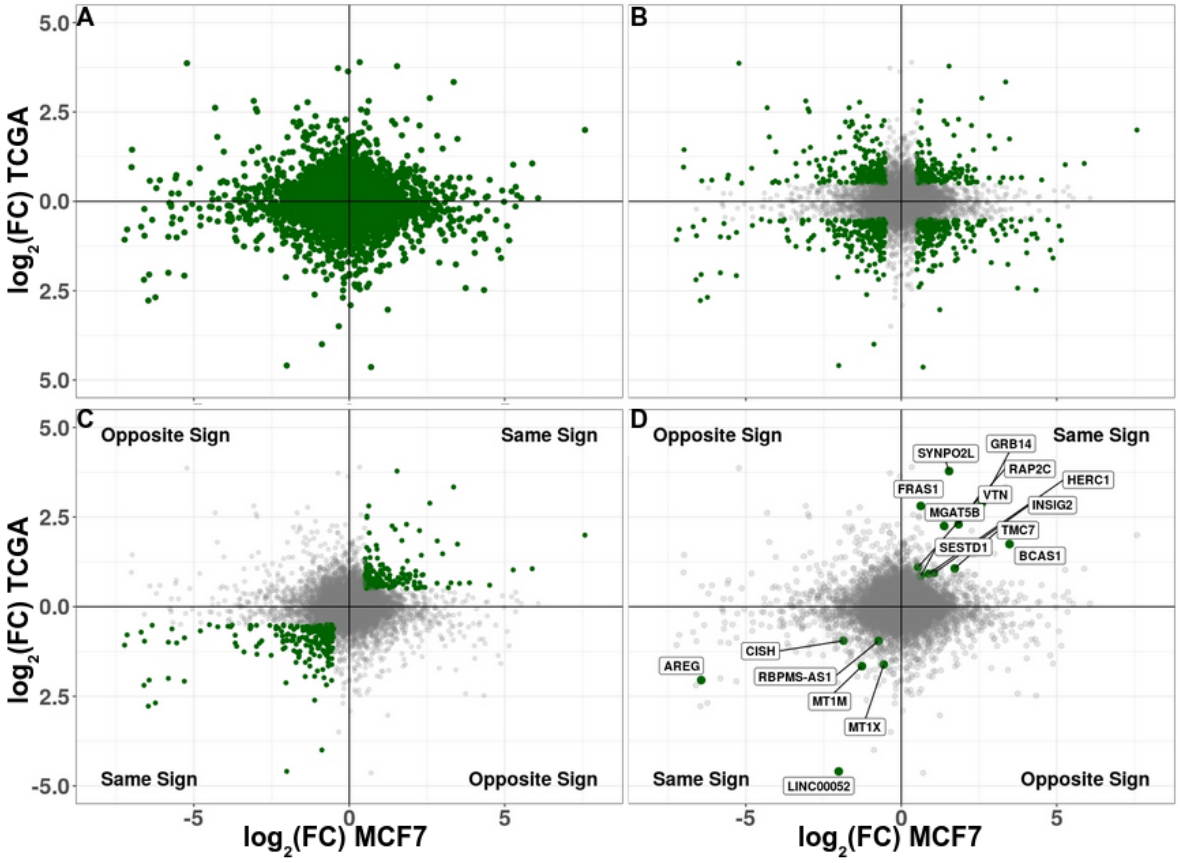

Figure 4.5: Gene cloud showing the distribution of genes in the joint analysis of the cell and patients data as well as the effect of the applied filters on the final shortlist. **A** No filter. **B** Filter for expression levels of $|\log_2$ FC$|$ > 0.5 set for both analyses. **C** Only genes expressed in the same direction are kept. **D** Remaining 17 genes after setting a significance threshold of FDR < 0.1 for both DEA.

across patients μ_n and dividing by the standard deviation σ_n. The GSS was then obtained by summing Z-scores across all the genes in the signature:

$$GSS = \frac{1}{N}\sum_{n=1}^{N}\frac{g_{i,n} - \mu_n}{\sigma_n}. \tag{4.1}$$

The use of a Bayesian framework allowed carrying information from the cell line data to this model. We do this by using a normal prior centered in the mean $\log_2^G$ FC value of each gene G in the signature. This value is taken from the MCF7 CTRL vs MCF7 TamR differential expression analysis in section 3.5,

$$\mathcal{P}(\theta) \sim \mathcal{N}(\log_2^G \text{FC}, 2.5), \tag{4.2}$$

where the standard deviation was chosen following the recommendations for the use of weakly informative priors in logistic regression by Gelman et al. (2008) as seen in section 2.3.3.2.

For the likelihood, we used the logistic function assuming a binary response of patients either developing resistance, $y_i = 1$, or responding well to the given treatment, $y_i = 0$

$$L(\theta|y, GSS_i) = \prod_{i=1}^{N}\left[\left(\frac{e^{\theta GSS_i}}{1+e^{\theta GSS_i}}\right)^{y_i}\left(1-\frac{e^{\theta GSS_i}}{1+e^{\theta GSS_i}}\right)^{1-y_i}\right] \; ; \; y_i \in \{0,1\}. \tag{4.3}$$

The resulting posterior distribution of the model coefficient θ together with a cross validation algorithm helped us to predict how well each GSS could classify patients into the correct response group. To evaluate the predictive power of a signature to be tested, the GSS value had to be obtained from patients data for all the genes in such a signature. Due to the low number of patients available to make predictions, we selected for cross validation the Leave-one-out (LOO) cross validation algorithm. Aiming to improve computational performance, we used the Bayesian LOO formulation developed by Vehtari and colleagues implemented in the R package *loo* (Vehtari et al., 2017, 2015) which is compatible with the popular software for Bayesian analysis Stan (Stan Development Team, 2024). Therefore, we implemented our BLR model in RStan (R version of Stan) and chose the Hamiltonian Monte Carlo method for sampling.

The metrics upon we based our decision regarding the predictive efficiency of a tested signature were the classical sensitivity, specificity, precision and accuracy. Those metrics rely on the ratios of True Positive (TP), True Negative (TN), False Positives (FP) and False Negatives (FN). In our study, these values account for the number of times a Good-Responder/Resistant patient has been correctly classified as such, or mislabelled as a member of the other group:

$$Sensitivity = \frac{TP}{TP + FN}, \tag{4.4}$$

$$Precision = \frac{TP}{TP + FP}, \tag{4.5}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{4.6}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.7}$$

Besides, due to the Bayesian nature of the problem, we included in the list of metrics a purely Bayesian metric, the expected log pointwise predictive density, or *ELPD* — a metric for model comparison specific to the Bayesian LOO framework as suggested in Vehtari et al. (2017)

$$ELPD = \sum_{i=1}^{N} \log(p(y_i|y_{-i})). \tag{4.8}$$

To evaluate it, one needs to look at the leave-one-out predictive density given the data y without the i^{th} data point removed for testing, i.e.

$$p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i})d\theta, \tag{4.9}$$

which can be tricky to compute. However, assuming that the data points in the model are conditionally independent, one can use importance sampling to obtain importance ratios that allow for estimation of equation 4.9 (Gelfand et al., 1992). For that, S draws θ^s from the full posterior are taken and with the use of the importance weight ratio

$$r_i^s = \frac{1}{p(y_i|\theta^s)} \propto \frac{p(\theta^s|y_{-i})}{p(\theta^s|y)}, \tag{4.10}$$

one can get the (IS) LOO predictive distribution,

$$p(\tilde{y}_i|y_{-i}) \approx \frac{\sum_{s=1}^{S} r_i^s p(\tilde{y}_i|\theta^s)}{\sum_{s=1}^{S} r_i^s}. \tag{4.11}$$

The method implemented in the *loo* R package further improves the LOO estimate by taking a Pareto smoothed importance sampling (Vehtari et al., 2015), which applies a smoothing procedure to the importance weights.

By using this approach, we wanted to seek further confirmation of the correlation between the found group of common 17 genes, or a subset of it, and the occurrence of resistance. The objective was to rule out the triviality of having genes arising purely due to a spurious correlation of both DEA. To this end, we set to study the predictive capabilities of the possible candidate signatures that could be generated using the genes within this list of 17 genes. We do this by analysing their ability to predict the response of the patients classified as resistant. However, the tamoxifen-treated patient dataset that we used to initially identify these relevant genes included only 37 patient, which is a low number for extrapolating predictions from, even by LOO standards.

Thus, we expanded our hypothesis to explore the possibility that some of the tested genes could also be associated with resistance to other forms of hormone therapy. We have previously built our hypothesis on the fact that tamoxifen-treated MCF7 and tamoxifen-treated patients in general share common traits that could implicate resistance development. The mechanisms for this resistance are not yet fully understood, but multiple signalling pathways have been named as potential factors. Moreover, Mills et al. (Mills et al., 2018) suggest that these pathways are not restricted to tamoxifen, opening an interesting path to explore if both forms of resistance (to tamoxifen and to aromatase inhibitors) share mechanisms and, thus, the identified genes might also be markers of resistance to aromatase inhibitors.

Therefore, we set to study the behaviour of the 17 candidate genes in a bigger cohort of 127 patients, which included all TCGA patients treated with any type of hormone therapy. These patients were chosen following the same selection criteria as the tamoxifen treated patients. By incorporating this new set of patients, we brought over additional benefits to our study. Among them, the generalisation of the methodology to a broader range of treatments, the analysis of the data on a dataset different from the one used for gene discovery — which should mitigate any overfitting to the original dataset — and, finally, an enhancement of the statistical significance of the findings, as the number of testing subjects was increased. The next step is to put the just introduced BLR-LOO approach into the framework of the simulated annealing-type refinement algorithm developed within this study.

4.4.3 Simulated Annealing-type algorithm

To efficiently explore the space of possible gene signatures and to identify the gene signature with the best predictive power, a simulated annealing (SA) (Černỳ, 1985; Kirkpatrick et al., 1983) – type algorithm, which we called BLR-SA, was proposed and implemented in-house. Using the data from 127 hormone therapy treated patients from the TCGA cohort, the algorithm probed 50.000 combinations of candidate signatures formed from the pool of genes previously identified.

The tested signatures were obtained by extracting genes from the 17-genes pool resulted from the cell-patient analysis. With the first signature selected randomly, every new gene signature was formed by either removing one randomly selected gene from the current signature or by adding a random gene selected from the remaining genes in the pool. In addition, the possibility to substitute one random gene in the current signature with the random one from the pool was also included with the same probability as the two aforementioned events.

At every iteration, the BLR model coupled with LOO was run for 5000 warm-up and 5000 production steps, sufficient for achieving convergence, and was used to estimate accuracy and *ELPD* as explained in the previous section. These metrics were used for model selection, alongside the Metropolis test within the SA-type algorithm. The overall prediction accuracy was considered as the default metric for the acceptance probability in the Metropolis test. The *ELPD* was applied as an auxiliary decision metric to resolve a tie in the competition between the top gene combinations

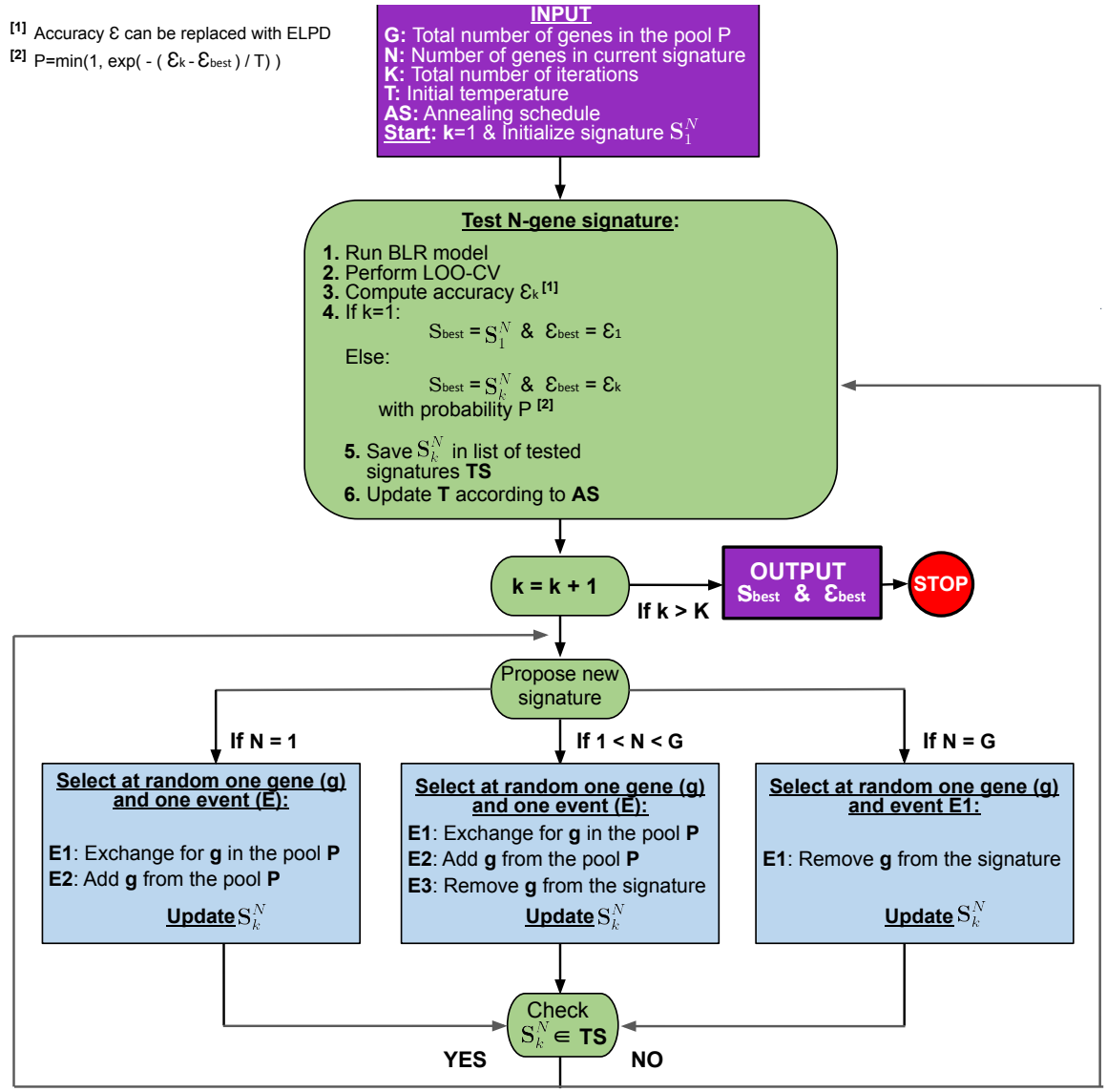

Figure 4.6: Schematic representation of the gene signature selection algorithm, BLR-SA. The BLR model gives a posterior distribution from which it is possible to make predictions on resistance. This value is compared to previous best and recorded before proposing a new signature for a new iteration.

that showed similar values in terms of predictive performance. The algorithm included a simulated annealing cooling scheme which halved the temperature parameter every 2.500^{th} iteration, gradually reducing it from 1.000 to 1. The BLR-SA algorithm is detailed in Figure 4.6.

Usual measurements for accuracy in prediction, i.e., sensitivity, specificity and overall accuracy were computed for all the tested signatures alongside *ELPD*. Among the tested signatures, several ones scored the best accuracy, reaching an overall accuracy in prediction of 76.4% (coloured points in Figure 4.7). To resolve the tie, the *ELPD* metric was used as a more advanced model comparison metric. The associated standard error for *ELPD* is often underestimated in scenarios involving a low number of instances and not independent models (Vehtari et al., 2017) - such as those generated using the LOO algorithm. The relationship between *ELPD* and prediction accuracy was carefully examined in our tests. We found a clear correlation between a higher and better (less negative) *ELPD* values and a more accurate prediction, which can be observed in Figure 4.7. This makes *ELPD* a good auxiliary metric for decision-making in cases where prediction metrics such as sensitivity, specificity

or accuracy display comparable results.

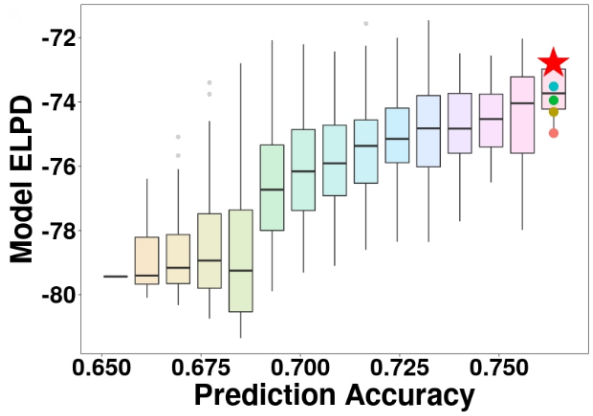

Figure 4.7: Correlation between *ELPD* and Overall Prediction Accuracy for all the gene signatures tested with the BLR-SA algorithm. The trend in the mean of the boxplots suggests that more accurate models also show lower negative *ELPD*. Final selection is marked with a star.

Therefore, the final selection recommended by the BLR-SA algorithm is that corresponding to the top right point in Figure 4.7, indicated with a star. This is a 6-gene signature which offered the overall optimal prediction. The signature, which we named EnCaRes (Endocrine Cancer Resistance) had a Sensitivity=0.813, Specificity=0.757, Overall Accuracy= 0.764 and *ELPD*=-72.82. The genes composing this signature are shown next in Table 4.1:

EnCaRes 6-gene signature					
VTN	**GRB14**	**TMC7**	**INSIG2**	**HERC1**	**FRAS1**

Table 4.1: Genes forming the EnCaRes signature.

In the Precision-Recall curve in Figure 4.8 we can see that the EnCaRes signature (red) improves the original 17-gene signature (blue) obtained using only the DEA results. The dotted line represents a dummy classifier, which gets a score below 0.5 due to the imbalanced nature of the dataset. The area under the Precision-Recall curve (AUC-PR) confirms the superiority in prediction performance of the new 6-gene signature generated using the BLR-SA algorithm over the initially identified 17 genes common to cell and patients.

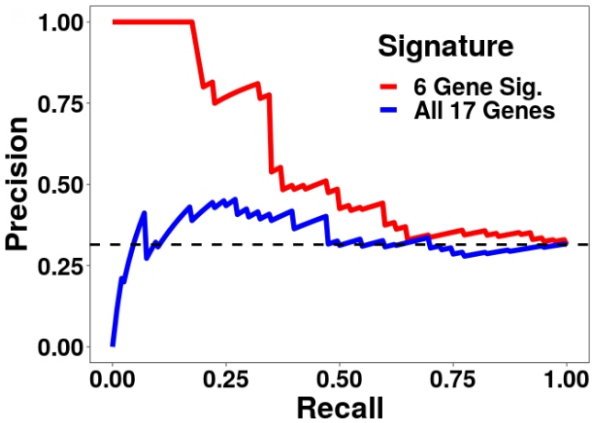

Figure 4.8: Precision–Recall curve for the 6-gene signature obtained using the BLR-SA algorithm (blue) and compared with the Precision–Recall curve produced by using all 17 genes.

4.5 Validation

To ensure the validity of the identified signature outside of the dataset used for its discovery, it is necessary to test it in other cohorts and check its ability for the identification of resistant patients. The amount of treatment detail offered by the TCGA database is generally not available in other databases, which can make the classification of resistant patients more difficult in other cohorts. However, relapse-free survival (RFS) information is widely accessible and could be used to perform extra validation on different datasets. RFS measures the time from the completion of treatment until the occurrence of a disease recurrence, progression, or the development of new tumours. The similarities between RFS and the conditions established in section 4.3.2.2 for the classification of patients in the TCGA cohort make RFS an ideal candidate metric for validation in different cohorts. For this purpose, two collections of patients were selected.

4.5.1 Validation cohorts

KMplotter

The KMplotter tool (Győrffy, 2021) is an online tool, frequently used to perform survival analysis. It offers access to a multitude of patients and microarray data (not RNA-seq) extracted from different cohorts in the Gene Expression Omnibus (GSE9195, GSE65194, GSE19615, GSE16391, GSE17907 and GSE21653). Microarray and RNA-seq are very different technologies for sequencing, so in most cases it is not advisable to mix them in the same experiments. However, in this case the datasets are not mixed or integrated directly, and the microarray data is only used to measure expression levels of genes in the signature for the cases of interest, meaning that it could be used to perform this independent validation. The chosen cohorts contained a total of 181 tamoxifen-treated and 385 endocrine therapy-treated patients. In addition, 470 other patients with ER- tumours were used as a negative control. The KMplotter tool was applied to compare the RFS curves of patients with high and medium/low signature expression levels.

METABRIC

The other dataset used for validation was taken from the cBioPortal online database (Gao et al., 2013) and consisted of RNA-seq and clinical data originally from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Curtis et al., 2012) database. This data was less detailed in terms of treatment than that of the TCGA, only indicating if endocrine therapy was administered or not. However, it did not include information on the specifics of the given treatment and, therefore, tests were conducted without distinguishing between hormone therapies. ER+ patients treated with hormone therapy were selected for validation (n=769), after excluding patients marked as deceased from other causes. As before, ER- patients were included as a negative control (n=429).

Since this cohort contains complete RNA-seq data (which was treated and analysed using DESeq2, following the same procedure as in 3.3.3), it was possible to test our gene signature and compare its performance with other known signatures for resistance/relapse and known covariates, as well as performing the same type of survival analysis as with the KMplotter cohort.

4.5.2 Survival analysis

4.5.2.1 Kaplan–Meier curves

Survival analysis is a statistical analysis widely used in medical research, to study the rate at which patients are affected by a certain event being investigated, such as death, disease recurrence, or relapse. It provides an understanding of the time-to-event data for a given study which may or may not account for censored observations, namely those where the event has not occurred by the end of

the study. In this context, the Kaplan-Meier estimator (Kaplan and Meier, 1958) generates survival curves that depict the probability of an event, i.e., a decreasing probability which over time goes down whenever an event occurs. These curves allow for the comparison of different groups and help identify the impact of one or various factors on the likelihood of an event happening. Since the curves depict probabilities, they are a powerful visualization tool for the influence of a variable in the probability of an event. Additionally, statistical tests such as the log-rank test are employed to calculate p-values and determine if there are significant differences between survival curves.

For this study, the survival analysis R packages *survival* (Terry M. Therneau and Patricia M. Grambsch, 2000; Therneau, 2023) and *survminer* (Kassambara et al., 2021) were used to conduct statistical comparisons among groups and create the Kaplan-Meier survival plots, in combination with the automatic tool from KMplotter. The patients in the validation cohorts were subject to right-censoring according to the respective clinical data, while the p-value accounting for statistical significance was derived from log-rank tests. The impact of the EnCaRes signature was tested by dividing the cohorts into two groups using the **upper tercile threshold**, i.e., taking the 66.6% threshold to separate the patients with high levels of gene expression from those with medium/low expression. These two groups represented the high and low risk groups for resistance according to our model.

For the KMplotter cohort, we studied the risk of patients from three groups: (1) treated with tamoxifen only, (2) treated using any form of hormone therapy and (3) patients with ER- tumours as control, as those are not susceptible to hormone treatments. Kaplan-Meier analysis showed a decrease of 12% in the 10-year RFS probability for tamoxifen-treated tumours with high expression levels of the gene signature (Figure 4.9, left). This effect was greater in patients treated with any type of hormone therapy (Figure 4.9, middle), with the reduction of RFS after 10 years from 78% to 64%. In concordance with our assumptions, there was no significant effect using EnCaRes for assessing risk of ER- cancers (Figure 4.9, right) as these are not susceptible to these types of therapies.

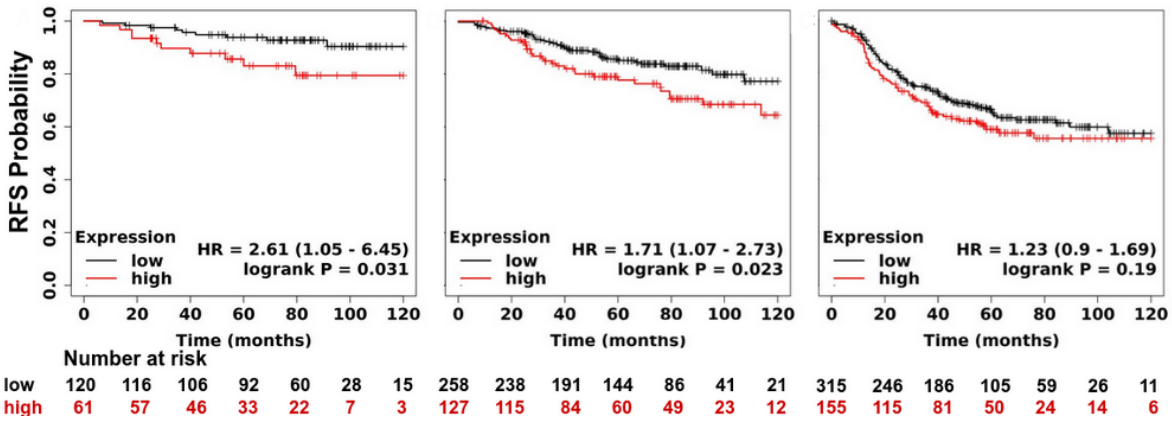

Figure 4.9: Relation between mean expression levels of the EnCaRes signature and recurrence free survival evaluated by Kaplan-Meier analysis using the KMplotter tool. On the left, tamoxifen treated patients. In the middle, patients treated with any type of hormone therapy and on the right, patients with ER- tumours non-susceptible to hormone therapy treatment. In the treated patients, the EnCaRes signature separates high and low risk groups using an upper tercile cut-off. No significant effect is seen in the ER- patients.

Next, we investigated our second validation cohort, containing METABRIC patients data and only included information about hormone therapy treated patients. However, the number of patients in this cohort was significantly larger than in the KMplotter cohort, leading to an improved statistical accuracy of the results. As before, the high-risk group was defined by expression levels higher than the upper tercile of all patients in the study, and the rest of the patients were classified as medium/low risk. Figure 4.10 presents the time evolution over 10 years of the RFS probability of the EnCaRes gene signature observed in two METABRIC cohorts. For patients receiving hormone therapy, high

expression levels of the gene signature revealed a 16.4% decrease in the 10-year RFS probability as seen in the left panel from Figure 4.10. In contrast, expression levels of this gene signature did not have any significant effect in terms of risk of recurrence in patients with ER- tumours, as seen in the right panel in Figure 4.10. Furthermore, none of the 6 genes of the signature showed potential as prognostic biomarker when tested individually, as seen in Appendix B.1. However, the combined enhanced expression of 6 genes demonstrated its association with an increased risk of recurrence in ER+ breast tumours.

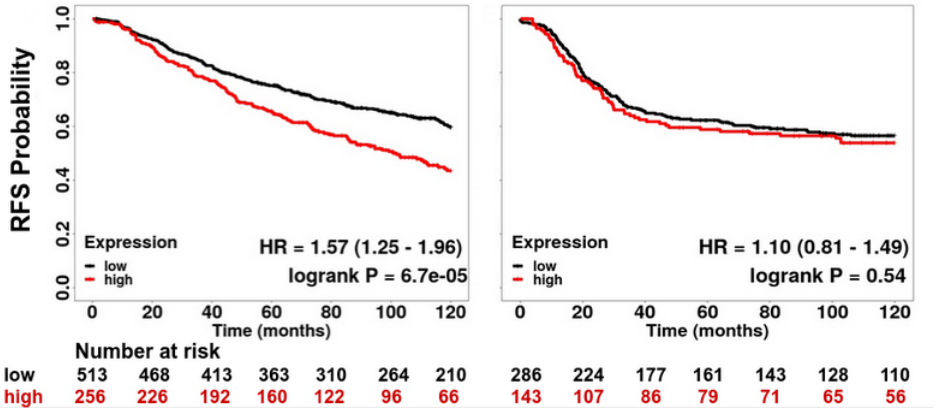

Figure 4.10: Relation between mean expression levels of the EnCaRes signature and recurrence free survival in the METABRIC cohort for patients treated with any type of hormone therapy (left) and for patients with ER- tumours non-susceptible to hormone therapy treatment (right). An upper tertile threshold for high/low expression based on the mean expression levels of the EnCaRes signature was used. In the treated patients, the EnCaRes signature separates high and low risk groups, but no significant effect is seen in the ER- patients.

4.5.2.2 Cox proportional hazard models

The Cox Proportional Hazards model is a specific type of survival model proposed by David Cox (Cox, 1972) to assess the effect of one or more covariates on a hazard occurring over time. Due to its prominent use within statistical medicine, the hazard usually refers to the rate of occurrence of a disease-related event. The key assumption of these models is that the hazard for any individual i is a constant multiple of the baseline hazard function, $h_0(t)$ and the impact of each additional covariate p is measured by a global parameter β_p.

Let $X_i = (X_{i,1}, ..., X_{i,p})$ represent the effect of all p covariates for a patient i. Then the Cox proportional hazard model for patient i takes the form

$$h_i(t|X_i) = h_0(t)\exp(\beta_1 X_{i,1} + ... + \beta_p X_{i,p}). \tag{4.12}$$

Since the baseline function $h_0(t)$ is common to all covariates, usually only the exponential part of 4.12 is used for the analysis. Hence, the term $\exp(\beta_p)$ for a given covariate p is called Hazard Ratio (HR) and a value of HR > 1 (or alternatively a value of $\beta_p > 0$) indicates that the covariate-related hazard increases and, as a consequence, the time of survival decreases due to the effect of that covariate.

The Hazard Ratio was already included in Figures 4.9 and 4.10, where this behaviour is exemplified in the clearly positive effect appearing in the ER+ resistant patients, meaning that an overexpression of the EnCaRes signature genes implies an increased hazard over time. For the ER- patients the HR value is also slightly above 1 but, crucially, the confidence interval in parentheses covers a range of both positive and negative effects, meaning that no significant conclusion can be taken regarding the effect on these patients.

In this section, we examine the impact of our signature EnCaRes as a predictor of risk in comparison with other relevant gene signatures using the METABRIC cohort. To do so, we utilized both the univariate model, where we analysed the hazard caused only by a single covariate; and a multivariate model, where we incorporated other major signatures as covariates. The univariate model allows assessing the relevance of the individual effect of each predictor tested, while the multivariate model offers information on the relative importance of all predictors, to compare between all covariates and elucidate which is more relevant.

The detailed survival information offered by the METABRIC database facilitated a comparison with other previously identified prognostic signatures. Three of the tested signatures were initially developed for tamoxifen resistance, namely the 5-Candidate Pathway signature (Rahem et al., 2020), the HOXB13/IL17BR ratio (Ma et al., 2004) and the 10-gene signature arising also from a MCF7-TCGA joint analysis (Men et al., 2018). Another three were developed for relapse under all forms of hormone therapy and included Oncotype DX (Paik et al., 2006), the 18-gene ER/PR related signature (Sinn et al., 2019) and a recent signature obtained from the study of ESR1 CRISPR-Cas9 edited breast cancer cell lines (Harrod et al., 2022).

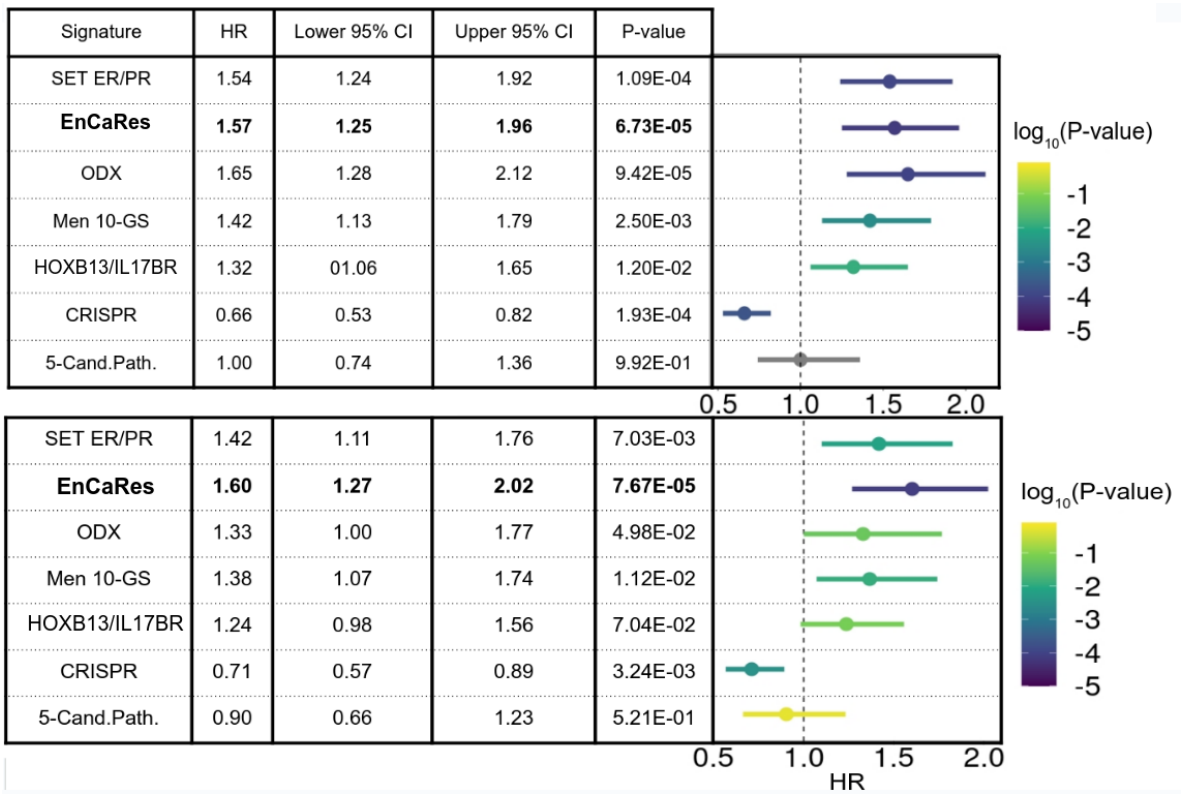

Figure 4.11: Univariate (top) and multivariate (bottom) Cox proportional-hazards regression models for testing the predicting ability of the EnCaRes signature against the known resistance signatures on the METABRIC cohort. The selected signatures abbreviations stand for (from top to bottom): the 18-gene SET ER/PR signature (Sinn et al., 2019), EnCaRes signature (this study), ODX — Oncotype DX (Paik et al., 2006), the Men 10GS — 10-gene signature by Men et al. (Men et al., 2018), the HOXB13/IL17BR ratio (Ma et al., 2004), CRISPR — the signature, obtained with CRISPR-Cas9 edited ESR1 (Harrod et al., 2022) and 5-Cand Path. — the 5-Candidate Pathway (Rahem et al., 2020). HR is the Hazard Ratio and CI is the confidence interval. A smaller p-value indicates a more statistically significant result

Figure 4.11 illustrates the performance comparison of EnCaRes with 6 signatures described above, where EnCaRes came out as the best predictor (in terms of p-value) both in the univariate (top) and multivariate (bottom) Cox proportional-hazard regressions. For the univariate case, which computes the predictive power of each signature individually, three signatures show fairly similar performance,

namely the in-house EnCaRes signature; ODX, widely used in clinical settings; and SET ER/PR, which uses genes correlated with ESR1 and PGR. Our signature differs from these by not relying on proliferation or breast cancer biomarkers, while also using fewer genes. For the multivariate case, our EnCaRes signature proved to be the best predictor for RFS among the known signatures.

In addition to the comparison with the previously known signatures, we also tested our signature against several clinical covariates available in the METABRIC dataset, such as menopausal status (Menopause), tumour grade (Grade), HER2 status and the administration of radio (Radiotherapy) or chemotherapy (Chemotherapy). All of these can be associated to a worse prognosis in some degree, with studies even pointing to a direct link between a positive HER2 status and a higher chance of developing resistance to tamoxifen (Menendez et al., 2021). Apart from these common clinical variables, we included in the analysis the Proliferative Index (Prol.Index) (Ramaker et al., 2017) defined through the meta-PCNA (Venet et al., 2011) to compare our result with proliferation markers, which are also frequently used in clinical practice. The results of this analysis can be seen in Figure 4.12, where, as before, our signature demonstrates the lowest p-value among the tested covariates and sets itself apart from the proliferation markers. This yet again suggests an implication of the EnCaRes signature in a specific treatment related mechanism, such as a resistance to the administered therapy and its potential for use as a predictive indicator compared to clinical variables currently in use.

Covariate	HR	Lower 95% CI	Upper 95% CI	P-value
Prol.Index	1.40	1.29	1.83	1.55E-04
EnCaRes	**1.62**	**1.29**	**2.04**	**3.22E-05**
Radiotherapy	0.78	0.62	0.99	4.42E-02
Menopause	0.51	0.34	0.75	8.88E-04
HER2 Status	1.74	1.25	2.42	9.20E-04
Chemotherapy	1.23	0.90	1.71	1.94E-01
Grade	1.45	1.19	1.76	1.65E-04

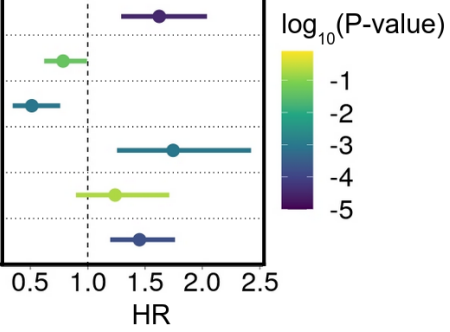

Figure 4.12: Multivariate Cox proportional-hazards regression model compares the predicting ability of our EnCaRes signature with the known clinical covariates on the METABRIC cohort. HR is the Hazard Ratio and CI is the confidence interval.

4.5.3 Experimental validation

The identified 6-gene signature, EnCaRes, suggests that the combined expression of 6 genes forming the signature is higher in breast cancer cells with increased risk of developing resistance to hormone therapy. Therefore, the final experimental validation step included the analysis of the mRNA expression levels in control and resistant MCF7 cells using quantitative PCR (qPCR). RNA was isolated using the Macherey-Nagel NucleoSpin® RNA, according to instructions of the manufacturer. Real-time PCR was performed on a ViiA 7 or a Quant-Studio 6 Flex Real-Time PCR Systems (Applied Biosystems) as previously described in Piva et al. (2014). The data shown here was extracted from experiments conducted by Dr. Miriam Rábano at the Cancer Heterogeneity Lab.

The expression of all genes present in the signature was validated by qPCR analysis. Expression of VTN, TMC7, INSIG2, GRB14 and FRAS1 using qPCR confirmed the upregulation of these genes in tamoxifen resistant cells as seen in Figure 4.13. While RNA-seq data initially suggested this, the results from RNA-seq and qPCR do not always fully coincide due to biological and technical variability, which can influence outcomes in both methods. Thus, the consistent results obtained with both RNA-seq and qPCR for 5 genes of the EnCaRes provide a positive confirmation of the initial findings.

This was not the case for HERC1, which was not found to be signficantly upregulated by qPCR. Incidentally, HERC1 is the only gene in the EnCaRes signature, with FDR in the TCGA DESeq2 analysis being slightly above the usual 0.05 threshold. Nevertheless, the signature EnCaRes works better with HERC1 included in every tested scenario and, therefore, the data supports including all 6 genes in the identified signature.

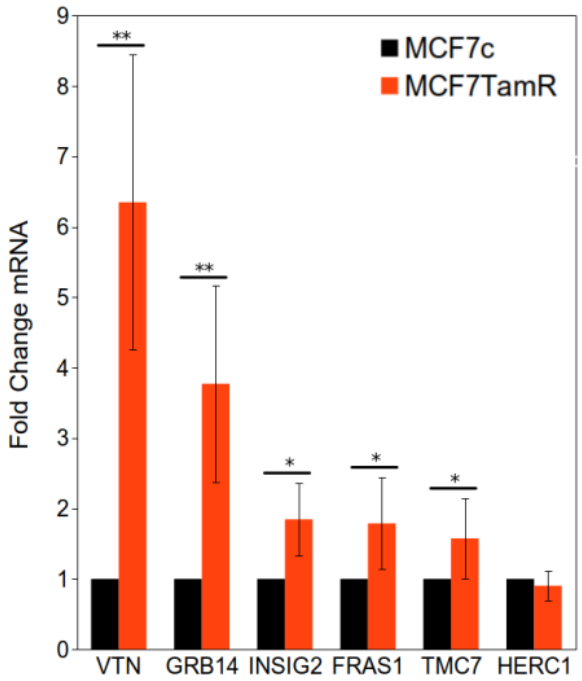

Figure 4.13: Experimental validation by qPCR. Relative transcript levels from qPCR analysis of the EnCaRes signature in MCF7 CTRL (black bars) and MCF7 TamR (red bars) cells show a significant increase in all but one of the genes in the signature. Error bars represent standard deviation (SD) for n=5 experiments. Asterisks indicate statistical significance from a one-sided t-test. (*) p-value < 0.05, (**) p-value< 0.01.

4.6 Biological implications

Our results suggest that overexpression of the genes in the EnCaRes signature is linked with a higher risk of resistance to endocrine therapies in ER+ breast cancer. After providing an exhaustive computational statistical analysis and validation, it is important that we complete the study with an appropriate biological context for the genes involved in the signature. We begin with a one-by-one description of the genes composing the signature, hyper-referenced to their GeneCard page, accompanied by the relevant information known to date on their function and origin.

- **VTN — Vitronectin**: A cell adhesion and spreading factor that interacts with glycosaminoglycans and proteoglycans, is recognized by certain members of the integrin family and serves as a cell-to-substrate adhesion molecule. The encoded protein functions, in part, as an adhesive glycoprotein. Differential expression of this protein can promote either cell adhesion or migration, as it links cells to the extracellular matrix. The protein can also promote extracellular matrix degradation and thus plays a role in tumorigenesis.
- **GRB14 — Growth factor receptor-bound protein 14**: Adapter protein which modulates coupling of cell surface receptor kinases. Binds to, and suppresses signals from, the activated insulin receptor (INSR). Potent inhibitor of insulin-stimulated MAPK3 phosphorylation. Plays a critical role in regulating PDPK1 membrane translocation in response to insulin stimulation

and serves as an adapter protein to recruit PDPK1 to activated insulin receptor, promoting PKB/AKT1 phosphorylation and transduction of the insulin signal.

- **INSIG2 — Insulin-induced gene 2 protein**: Mediates feedback control of cholesterol synthesis by controlling SCAP and HMGCR. Capable of retaining the SCAP-SREBF2 complex in the ER thus preventing it from escorting SREBPs to the Golgi. Seems to regulate the ubiquitin-mediated proteasomal degradation of HMGCR.

- **FRAS1 — Fraser extracellular matrix complex subunit 1**: This gene encodes an extracellular matrix protein that appears to function in the regulation of epidermal-basement membrane adhesion and organogenesis during development. Mutations in this gene cause Fraser syndrome.

- **TMC7 — Transmembrane channel-like protein 7**: It has been predicted to be involved in ion channel activity and transmembrane ion transport.

- **HERC1 — Probable E3 ubiquitin-protein ligase**: Member of the HERC protein family. The protein may be involved in membrane transport process via some guanine nucleotide exchange factor (GEF) activity.

In addition, nearly all the genes in the signature have been previously linked to breast cancer except for FRAS1. Recent studies associate to some degree expression of VTN (Bera et al., 2020), TMC7 (Song and Ran, 2021) and INSIG2 (Lu et al., 2021) to poorer prognosis. Meanwhile, HERC1 has been linked both to complications in breast cancer (Rossi et al., 2021) and, in particular, in estrogen-targeted treatments (Goto et al., 2011).

Among the genes forming our newly discovered signature, only GRB14 appears in another known gene signature for resistance (Men et al., 2018), though being expressed in the opposite direction, something found as well in Huang et al. (2013). However, every validation we could access, from our own sequencing, public repositories, and qPCR, reflects its overexpression in the context of resistance to hormone therapy. High GRB14 levels may indicate that ER transcriptional activity is compromised in resistant cells (Koren and Bentires-Alj, 2015; Piva et al., 2014), and GRB14 expression has been shown to be downregulated by estrogen (Kairouz et al., 2005).

Moreover, it is possible that the pathways where these genes are involved could have a downstream effect that relates to some resistance mechanism. In this regard, we observe that there are several links between the genes in this signature and the extra-cellular matrix (ECM). Gene set enrichment analysis of the 17 common genes performed using the R/Bioconductor package fast Gene Set Enrichment Analysis (fGSEA) (Korotkevich et al., 2019) showed that the estrogen response is reduced while angiogenesis and the ECM receptor interactions pathways are enhanced as seen in Figure 4.14.

Finally, one major thing to take into account is that the signature has been efficient for patients who had tumours sequenced at the time of diagnosis, but no data on the evolution of the gene expression patterns across time could be obtained. Therefore, the main assumption is that the genes in our gene signature are present in higher abundance in those patients who are prone to become resistant to hormone therapy.

We performed qPCR experiments that indicate that there is no immediate significant effect in the amount and direction of expression of these genes when MCF7 cancer cells are treated with tamoxifen. These experiments have been conducted by Dr. Miriam Rábano at CIC-bioGUNE and are displayed in Figure 4.15, where one can see that the response of these genes when treated with tamoxifen after 6, 18, 24, 48 and 72 hours does not follow any discernible pattern. These results suggest that there is no direct regulation of these genes by ER in the short term — hours or days. However, resistance can be developed over a longer period of time — months or years.

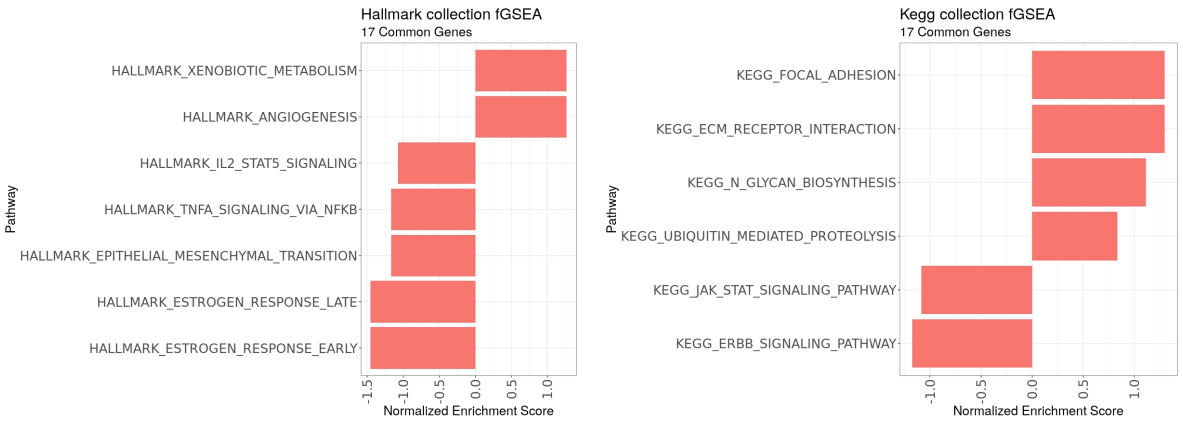

Figure 4.14: Gene Set Enrichment Analysis of the 17 common genes using the Hallmark collection from the Molecular Signature Database (left), and the KEGG database (right).

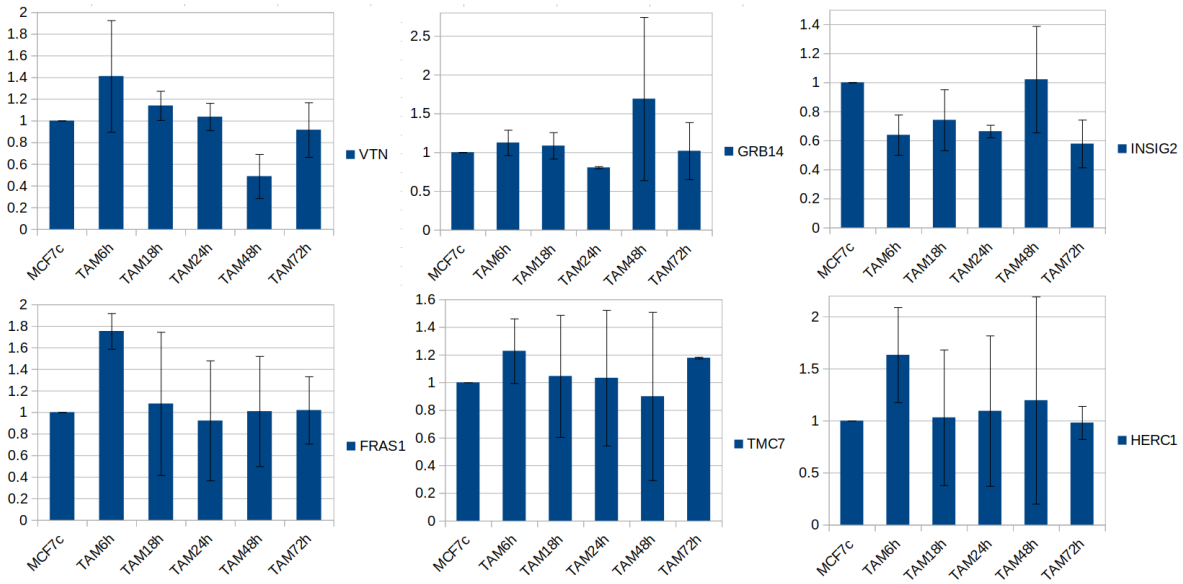

Figure 4.15: Expression levels of the genes in EnCaRes after 6, 18, 24, 48 and 72 hours of tamoxifen treatment compared to the expression in MCF7 cells.

New laboratory experiments are still needed to confirm the extent of the implication of this signature in the development of resistance, as well as the biological mechanisms behind it. However, the extensive computational validation and the preliminary qPCR results provide a solid ground upon which future experiments can be constructed.

4.7 Conclusions

In this chapter, we demonstrated the potential of the combined analysis of cell and patient data for the identification of relevant biomarkers for drug resistance in cancer. This was essentially the culmination of Objective 2, presented at the end of Chapter 1. With the methodology developed in this chapter, we were able to uncover a novel gene signature which we called EnCaRes signature for stratifying patient response to hormone therapy in ER+ breast cancer. This was achieved by detecting and exploring common alterations in resistance related sets of data, obtained and analysed in the laboratory using RNA-sequencing and in the clinic through the collection of patient statistics. This combined approach can assist in addressing several of the key issues usually associated with biomarker identification, as well as with applications of mathematical methods for studies of such

datasets. More precisely, it shows how a joint analysis of cell and patient data can contribute to solving two key problems in bioinformatics, one biological and another mathematical in nature.

From a mathematical perspective, the proposed methodology can serve as a feature selection mechanism, a helpful tool in modern big data statistics and machine learning. Despite recent advances in computational statistics, the efficient analysis of big amounts of data, such as outputs of transcriptomic studies, remains a great challenge. Therefore, a methodology like the one presented here should provide the aid necessary to filter data down to a manageable size, compatible with multiple mathematical algorithms, including the combined BLR-SA methodology developed and presented in this work. Moreover, the generality of the premise we proposed should facilitate implementations of this methodology in other types of cancer and disease studies, as well as applications of alternative types of refinement algorithms.

From a biological point of view, the new approach offers a framework where the usually heterogeneous clinical datasets can be guided and bounded by the laboratory experiments. A parallel analysis of carefully curated patients data and cell models should lead to the identification of common changes related to the biological question under study. By having the same underlying problem of resistance to therapy, the homogeneous nature of cell culture replicates can be used to focus the analysis of large heterogeneous transcriptomic profiles from real human patients towards the identification of promising biomarkers related to the problem in common. This methodology should help to manage the unavoidable heterogeneity observed in breast cancer and other types of tumours.

In addition, the presented here methodology BLR-SA for the refinement of the gene signature under study is applied before any validation. Thus, validation in our case is used for justification of the methodology rather than for further refinement of the signature, as commonly happens within the biomarker discovery process. In those cases, the final selection comes after hazard or survival analysis, resulting in a selection dependent on the validation cohort, something that we avoid by construction. We also remark that, in contrast to our proposed signature, other tamoxifen specific signatures were able to just marginally identify risk patients when tested in a scenario involving alternative hormone therapies as seen in Figure 4.11. Significantly, the stratification of patients based on risk of recurrence free survival was obtained in our case with just a 6-gene signature (compared to 21 genes in ODX and 18 in SET ER/PR) and without relying on proliferation genes. Extension of the signature from the tamoxifen specific scenario to all hormone therapies is particularly interesting in clinical research given that tamoxifen, which used to be the gold standard hormone therapy, is now often substituted by aromatase inhibitors for post-menopausal patients.

To conclude, this study shows that the mathematically enhanced combined analysis of well-characterised cell models and carefully curated patients data can help in the identification of biomarkers with a significant impact on the prediction of clinical responses. The approach proposed in this chapter is particularly useful for tackling the challenge posed by cancer heterogeneity, providing a tool for reducing model complexity and for efficient probabilistic data filtering, which should facilitate mathematical modelling of high-throughput sequencing data. Most of the results showed in this chapter were published in Parga-Pazos et al. (2024).

Chapter 5

Conclusions, Main Contributions and Future Work

5.1 Conclusions

Throughout this thesis, we have presented a series of novel studies and methods with two overall objectives in mind. Objective 1 revolved around the development of new methods and tools for the study of *ad hoc* datasets. We established two guiding principles for that, i.e., to work within a Bayesian framework and to use data related to the problem of resistance to hormone therapy in breast cancer for testing our proposed methodologies. Meanwhile, Objective 2 was tackling the development of hormone therapy resistance in breast cancer by means of the tools explored and developed to achieve Objective 1.

Within the scope of Objective 1, the main methodological advance presented in this thesis is our novel approach for adaptive integration of Modified Hamiltonian Monte Carlo methods for Bayesian inference, which we called s-MAIA. The algorithm offers an adaptive integration framework for Modified Hamiltonian Monte Carlo methods, such as Mix & Match Hamiltonian Monte Carlo (MMHMC), with the utmost aim to enhance sampling performance of MMHMC. For a given stepsize, the algorithm provides, first, a dimensional stability limit of 2- and 3-stage families of splitting integrators and then, the selection of optimal parameters for 2- and 3-stage splitting integration schemes that guarantee a minimal expected modified energy error. This translates into a better acceptance rates for the methodology MMHMC and, ultimately, to improved sampling.

In order to measure the impact of these enhancements, we proposed a new approach for calculation of Effective Sample Size (ESS) for irreversible Markov Chain Monte Carlo (MCMC) importance samplers, like MMHMC, which combines the most successful MCMC and importance sampling ESS estimators. We studied the effectiveness of several promising ESS estimators in irreversible and importance sampling conditions and selected a combination of those the best performance, namely: the Bartlett window ESS estimator to account for the correlated MCMC part of the chain, and $\omega \max$ ESS estimator for the importance sampling part. Using this and other popular performance metrics and two standard benchmark models, we compared performance of MMHMC coupled with s-MAIA against MMHMC with the state of the art modified 2-, 3-stage integrators as well as against HMC with its best performing integrators. Our tests revealed the superiority of MMHMC with s-MAIA over all tested methodologies. In addition, we applied MMHMC combined with s-MAIA to selected biomedical problems with the emphasis on the breast cancer study, and demonstrated that the efficient implementation of Bayesian inference can offer unique solutions to problems like perfect separation or the direction of influence of a given variable.

We can also frame within Objective 1 the development of a bioinformatics web-tool for the

analysis of cell line sequencing data extracted from tamoxifen resistant breast cancer cells. The tool, presented in Chapter 3 of this thesis, was specifically crafted to address the research requirements of the Cancer Heterogeneity Lab members at CIC bioGUNE, who contributed into this thesis by providing us with the dataset bG-RNA-seq for further analysis and with performing experimental validation of the gene signature developed within the thesis. This outcome is a great example of the positive synergies that can emerge from a close collaboration between statisticians and biologists. By establishing a common goal, this partnership facilitated the development of direct solutions to specific research challenges, showcasing the power of interdisciplinary cooperation.

Objective 2 was fulfilled with the work presented in Chapter 4, where we discovered a 6-gene signature related to resistance to endocrine therapy in ER+ breast cancer. This proved the potential of integrating cell and patients data analysis in identification of pertinent biomarkers for drug resistance in cancer. In this methodology, we selected a Bayesian approach in order to add more biological information in the proposed model. Our use of Bayesian priors reliant on differential expression data from the cell model made possible a flow of information from the cell to the patients' dataset.

Chapter 4 also includes the BLR-SA methodology, which is general enough to be considered as a major result on its own. The refinement of gene signatures, i.e., selecting a subset of highly relevant genes from a given set of genes, is not an easy task. It often requires additional biological information to perform the selection, and uses methods such the Cox proportional hazard model to remove genes that do not appear as significant contributors to the response one tries to model. The BLR-SA methodology utilizes a Leave-One-Out cross validation successfully to perform the selection within the original cohort while avoiding overfitting to the data. Since no extra biological information was assumed, the methodology can easily take any other disease with a binary response and use the same model to select the subset of genes most relevant for predicting the possible clinical outcomes.

In this study, we have maintained a coherent biological context through the mathematical analyses thanks to the close collaboration between BCAM and CIC bioGUNE. In addition, the EnCaRes signatures complies with several recommendations from the literature to avoid spurious correlations between signatures and clinical outcome (Goh and Wong, 2018; Manjang et al., 2021; Venet et al., 2011). We achieved this by proposing a short signature, not dependent on proliferation markers, that is able to create a high/low risk classification and that withstands multiple validations. We do not deny the relevance of proliferation biomarkers in this context, since gene signatures based on them can perform well, such as the widely popular and proved to be efficient Oncotype DX. However, identifying the markers that are not associated with proliferation helps to disentangle the role of proliferation from the role of other process-specific factors. On this regard, we believe that gene signatures like ours, that do not rely on proliferation markers and can successfully pass multiple independent validations, add value to the existing landscape of methods trying to predict response to breast cancer therapies.

Finally, we remark that prognostic signatures for endocrine therapy have also been proposed previously (Miller et al., 2015; Xia et al., 2022), but as far as we are aware, we are the first to use this combined cell-patient approach to obtain a signature with prognostic power in the broader context of endocrine therapy for ER+ breast cancer.

5.2 Main Contributions

The methodologies and results generated in this thesis have been implemented in multiple software packages and have been shared both in a peer-reviewed publication and in presentations at international and national conferences.

Publications

- **Parga-Pazos, M.**, Cusimano, N., Rábano, M., Akhmatskaya, E. and Vivanco dM. M (2024). A Novel Mathematical Approach for Analysis of Integrated Cell–Patient Data Uncovers a 6-Gene Signature Linked to Endocrine Therapy Resistance. *Laboratory Investigation*, 104(1), 100286. (DOI: https://doi.org/10.1016/j.labinv.2023.100286)

Abstract: A significant number of breast cancers develop resistance to hormone therapy. This progression, while posing a major clinical challenge, is difficult to predict. Despite important contributions made by cell models and clinical studies to tackle this problem, both present limitations when taken individually. Experiments with cell models are highly reproducible but do not reflect the indubitable heterogenous landscape of breast cancer. On the other hand, clinical studies account for this complexity but introduce uncontrolled noise due to external factors. Here, we propose a new approach for biomarker discovery that is based on a combined analysis of sequencing data from controlled MCF7 cell experiments and heterogenous clinical samples that include clinical and sequencing information from The Cancer Genome Atlas. Using data from differential gene expression analysis and a BLR model coupled with an original simulated annealing-type algorithm (BLR-SA), we discovered a novel 6-gene signature for stratifying patient response to hormone therapy. The experimental observations and computational analysis built on independent cohorts indicated the superior predictive performance of this gene set over previously known signatures of similar scope. Together, these findings revealed a new gene signature to identify patients with breast cancer with an increased risk of developing resistance to endocrine therapy.

Dissemination and Conferences

- E. Akhmatskaya, L. Nagar, M. Fernández-Pendás, **M. Parga-Pazos**, F. Puchhammer, T. Radivojević, J. M. Sanz-Serna. Hamiltonian Monte Carlo: Standard Practices Revisited, at *13^{th} International Conference on Monte Carlo Methods and Applications (MCM 2021)* in Mannheim, Germany, 18^{th} August 2021.

Abstract: With the recently increased interest in probabilistic models, such as Bayesian epidemic models or probabilistic deep learning, the efficiency of an underlying sampler becomes a crucial consideration. A Hamiltonian Monte Carlo (HMC) sampler is one popular choice for models of this kind. We revisit the standard practices of the HMC, and propose several promising alternatives. The topics of our discussion include a formulation of the HMC, numerical integration methods for Hamiltonian dynamics, and the choice of simulation parameters and settings.

- **M. Parga-Pazos**, E. Akhmatskaya, M. Vivanco. A Bayesian approach to tamoxifen resistance in breast cancer, in the talk series *Junior Research Seminar* at CIC bioGUNE in Bilbao, June 11^{th} 2022.

Abstract: Breast cancer is the most common type of cancer in women worldwide and most cases express the estrogen receptor, ER (ER+). These tumours are usually treated with hormone therapy, such as tamoxifen, but development of resistance to therapy is still a major clinical challenge. We aimed to tackle this problem by analysing RNA-Seq data extracted from breast cancer cell models and publicly available patient data, to identify common traits underlying the emergence of tumour relapse. In parallel, we developed a Bayesian logistic regression model to allow us to carry on information from cell models to the relevant context of clinical treatments. This model will be useful for validation of previous findings and identification of relevant genes implicated in mechanisms of resistance. In

addition, we anticipate prediction of the probability of a breast cancer patient to respond or not to hormone therapy.

- **M. Parga-Pazos**, E. Akhmatskaya, M. Vivanco. HaiCS in Action: A case study on Bayesian inference for genetics, at the meeting *Métodos de integración geométrica para problemas cuánticos, mecánica celeste y simulaciones Montecarlo* at Universidad Jaume I in Castellón, June 22^{nd} 2022.

Abstract: Breast cancer is the most common cancer type among women globally, with the majority of cases expressing the estrogen receptor (ER+), which makes them eligible for hormone treatment. However, a significant amount of tumours develop resistance, which ultimately results in a failed treatment and a loss of crucial time for treating the disease. As a result, prediction of resistance is a highly relevant problem with great implications in clinical practice. Here, we present a study where Bayesian inference is used to gain insight into this problem. We use advanced Hamiltonian Monte Carlo (HMC) techniques, such as Generalized HMC, with RNA-seq datasets implemented in BCAM's in-house software HaiCS (Hamiltonians in Computational Statistics) to analyse the behaviour of the posterior distributions of several key genes in the therapy resistance process.

- **M. Parga-Pazos**, E. Akhmatskaya, M. Vivanco. Integration of cell-patient data to tackle tamoxifen resistance in breast cancer, at the *UoE — Heriot-Watt Mathematical Biology Workshop* at the Bayes Centre in Edinburgh, March 31^{st} 2023.

Abstract: Breast cancer is the most prevalent cancer type among women globally, with the majority of cases expressing the estrogen receptor (ER+). Hormone therapies like tamoxifen are commonly used for treatment, yet resistance remains a significant clinical challenge. We aim to address this issue by integrating RNA-Seq data from breast cancer cell models with publicly available patient data to uncover common features associated with tumour resistance and relapse. Here, we present our BLR model developed to transfer insights from cell models to clinical treatment contexts, and the simulated annealing-type algorithm for refinement of gene signatures. This methodology results in a 6-gene signature capable of identifying breast cancer patients at higher risk of developing resistance to hormone therapy and which is validated independently across multiple independent cohorts.

- **M. Parga-Pazos**, L.Nagar, M. Fernández-Pendás, E. Akhmatskaya, M. Vivanco, V. Elvira, J.M. Sanz-Serna. Adaptive Integration Approach for Sampling with Hamiltonian Monte Carlo Based Methods, jointly with Lorenzo Nagar at the *14^{th} International Conference on Monte Carlo Methods and Applications (MCM 2023)* at the Sorbonne Université in Paris, June 29^{th} 2023.

Abstract: Hamiltonian Monte Carlo (HMC)-based methods have been widely recognized as a powerful sampling tool for Bayesian inference. Key performance factors of HMC are accurate numerical integration of underlying Hamiltonian equations and an appropriate choice of simulation parameters and settings. When applied to physical problems, natural constraints may offer a hint on such choices, which hardly can be translated into Bayesian inference applications. We present a novel adaptive integration approach (we call it s-AIA) that detects a system-specific multistage splitting integrator with a complete set of well-founded integrator-specific parameters to achieve a competitive sampling efficiency in HMC or a Generalized HMC-based Bayesian inference applications. The method relies on analysis of expected energy errors of multivariate Gaussian models combined with the data generated at the burn-in stage of a (G)HMC simulation. It automatically eliminates those values of simulation parameters, which may cause undesired extreme scenarios, such as resonance artefacts, low accuracy or poor sampling. We use the ideas of s-AIA to develop its advanced variant, s-MAIA, for importance sampling HMC. Testing s-(M)AIA on representative models demonstrates its superiority over popular symplectic integrators, e.g. Verlet, BCSS, ME. An illustrative example of s-(M)AIA application to

the study of uncovering novel biomarkers linked to endocrine therapy resistance in breast cancer is provided.

- **M. Parga-Pazos**, E. Akhmatskaya, M. Vivanco, V. Elvira. Impact of correlation, irreversibility and importance sampling on Effective Sample Size (ESS), at the *Hamiltonian Monte Carlo and Splitting Methods Workshop* at Universitat Jaume I in Castellón, November 23^{rd} 2023.

Abstract: Sampling methods for Bayesian inference do not gather samples directly from the target distribution and use various algorithms for generating samples — some produce correlated chains of samples, such as Markov Chain Monte Carlo (MCMC), or sampling from an importance distribution as in Importance Sampling (IS). The samples extracted using these algorithms are not equivalent to i.i.d. samples extracted directly from the target distribution. The Effective Sample Size (ESS) is a metric used to quantify the efficiency of a sampler in producing independent samples from the target distribution. Currently, several definitions of this metric co-exist, with no clear consensus reached for their use or range of applicability. Here, we explore how several key sampler properties affect ESS formulation and performance, and present a novel metric for cases where both MCMC and IS are in play.

Software Contributions

- **HaiCS (Hamiltonians in Computational Statistics)**

Authors: T. Radivojevic, E. Akhmatskaya, M. Fernandez-Pendas, F. Muller, J. Pérez-Heredia, L- Nagar, **M. Parga-Pazos**, F. Puchhammer, W. Isaac & H. Inouzhe.
Programming Language: C++ (calculations) & R (analysis).
Repository: https://gitlab.bcamath.org/mslms/haics
Description: MSLMS in-house software package for sampling in computational statistics using Hamiltonian Dynamics. The package is developed for statistical sampling of high dimensional and complex distributions and parameter estimation in different models through Bayesian inference using Hamiltonian Monte Carlo based methods. Different existing and recently developed numerical integrators, strategies for momenta update and flips can be employed within the Hamiltonian Monte Carlo (HMC), Generalized Hamiltonian Monte Carlo (GHMC), the in-house Mix & Match Hamiltonian Monte Carlo (MMHMC) methods and its variants. The package is suited for output analysis in CODA – a widely used R toolkit for Markov Chain Monte Carlo (MCMC) diagnostics. Currently, available numerical integrators include state of the art 2- and 3-stage splitting integrators for HMC and MMHMC (VV, (M-) ME, (M-) BCSS, as well as novel 2-, 3-stage integration schemes developed in MSLMS for HMC methods: s-AIA and for MMHMC method: (e)-s-MAIA.
Contributions:

- Implementation of novel adaptive integrators for modified Hamiltonians.
- Addition of new models based on Bayesian logistic regression and Gaussian mixtures.
- Implementation of a Prior Selection tool and its graphical interface.
- Contribution to the automation of the pre- and post-processing analysis.
- Addition of new visualization options for posterior distributions.
- Implementation of novel metrics for the calculation of Effective Sample Size.

- **TreaDS**

Authors: **M. Parga-Pazos**, S. Rusconi, J. Lemos
Programming Language: R
Repository: https://gitlab.bcamath.org/mslms/bayes-cancer
Description: Repository dedicated to tracking the development of resistance to anti-cancer therapy through mathematical modelling techniques. We include software to addresses challenges in clinical data analysis, specifically tackling the data separation issue and modelling gene counts across multiple biological conditions using RNA-sequencing data. Besides, it contains various approaches for RNA-seq data analysis such as, the treatment of raw unnormalized count data, pathway and gene set enrichment analyses, as well as the implementation of a Bayesian logistic regression model for assessing gene significance in a Bayesian context by means of probabilities of direction.

- **Vivanco LabSeq**

Authors: **M. Parga-Pazos**
Programming Language: R-Shiny
Repository: https://vivancoslabseq.shinyapps.io/RNASeqSOX
Description: Online application condensing several bioinformatic tools for DGE and GSEA analysis developed for use at Cancer Heterogeneity Lab at CIC-bioGUNE. The detailed description of the web is provided in section 3.5

- **BLR-SA-TamResistance**

Authors: **M. Parga-Pazos**
Programming Language: R & Matlab
Repository: https://github.com/chanfanetas/BLR-SA-TamResistance
Description: Source code for the use and replication of the methodology described in the paper *A novel mathematical approach for analysis of integrated cell-patient data uncovers a 6-gene signature linked to endocrine therapy resistance.* Includes the codes for the methodology presented in Chapter 4 of this thesis, i.e., the Matlab classification scripts, the differential expression analysis of cell and patients, the BLR-SA model and the validation checks. It also provides access to all the data (in the repository or referenced within it) for replication and validation of the results.

5.3 Future Work

Several of the ideas we explored in this thesis have great potential for expansion, paving the way for new research directions within the thesis topics and presenting opportunities for broader application across diverse fields.

In terms of enhancement of MHMC methodologies, we focused on one of the areas affecting the performance of these methods, i.e., numerical integration, and presented the s-MAIA algorithm. However, MHMC methodologies possess several simulation parameters, such as trajectory length or momentum noise, for which no clear optimization strategy exists so far. Further research is needed on optimization strategies for parameter tuning to automatize the choice of the best set of parameters for a given problem and to use it in tandem with our adaptive integration approach, s-MAIA.

Besides, the package HaiCS, where s-MAIA is currently implemented, needs modifications in order to extend its compatibility with several of the most widely used packages for Bayesian inference. In its current state, HaiCS provides highly efficient computations as it runs in C, but lacks the advanced interface for integrating various analysis packages written in R, that other state-of-the-art software

packages, such as Stan (Stan Development Team, 2024), provide thanks to its R implementation (RStan). An expansion of HaiCS in this direction is the next natural step to offer the opportunity of working with the highly efficient combination of MMHMC and s-MAIA to a wider public. The package itself will also benefit greatly from integration with packages such as *loo*, which should increase the range of use of the platform.

From the bioinformatics side, the possibilities for expansion involve the generalization of the ideas presented in this thesis. First of all, the code for web-app could easily be adapted to take as an input any form of RNA-seq data and perform these analyses for any type of biological condition. For such cases, the web-app should include an extra option to decide on key parameters regarding the differential expression analysis, but besides this, the rest of the platform can be used without any more significant changes. Moreover, it can be expanded to include a more direct approach to functional analysis, such as the Pathview graphs, currently outside of the pipeline due to the computational overheads introduced.

Finally, the methods used in Chapter 4 can also be generalized to any problem where a cell line model replicating patient response is available. The Bayesian framework, where cell line models inform the patients data, could very easily be reproduced for other types of cancer or even different diseases. However, extensive experimental work is still needed to confirm *in-vitro* and *in-vivo* the effectiveness of the EnCaRes signatures.

Appendix A

SOX2 Clinical Dataset

A.1 Original SOX2 Clinical dataset

Table A.1: Original dataset obtained from Hospital Universitario Galdakao-Usansolo for the study of the biomarker SOX2 on tamoxifen treated patients. In green, patients that responded to therapy; in red, those who developed a recurrence.

Patient	Age	Type	Grade	Size	Nodes	ER	PR	HER2	p53	Ki67	SOX2
1	50	IDC	G2	T2	N0	90%	70%	0	5%	10%	2
2	46	IDC	G2	T1c	N1a	90%	90%	0	0	20%	2
3	54	IDC	G2	T1c	N1a	90%	90%	0	0	0	2
4	53	IDC	G2	T2	N1a	90%	90%	0	0	25%	2
5	59	IDC	G2	T1	N1a	90%	80%	0	0	5%	2
6	42	IDC	G2	T1	N0	70%	60%	0	60%	0	2
7	43	IDC	G2	T1	N0	80%	30%	0	0	30%	2
8	43	IDC	G2	T2	N2a	80%	80%	0	0	10%	2
9	55	IDC	G2	T1	N0	90%	90%	0	0	10%	2
10	41	IDC	G2	T1	N0	80%	80%	0	0	15%	2
11	45	IDC	G1	T1c	N1a	90%	90%	0	0	5%	2
12	82	IDC		T2	N0	90%	50%	0	0	40%	2
13	65	IDC	G2	T1c	N1a	85%	30%	0	2%	10%	2
14	43	IDC	G3	T2	N1	70%	70%	2+	30%	25%	2
15	60	IDC	G1	T2	N0	80%	60%	0	0	5%	2
16	35	ILC	G2	T3	N1	70%	80%	0	5%	8%	1
17	53	IDC	G2	T2	N0	90%	90%	0	0	8%	0
18	81	IDC	G1	T2	N1	80%	70%	2+	0	6%	2
19	31	IDC	G3	T2	N1b	70%	90%	2+	0		2
20	78	IDC	G2	T2		90%	90%	2+	0	2%	1
21	55	ILC	G2	T1c	N1	80%	70%	0	10%	7%	2
22	73	IDC	G2	T4d	N1a	90%	90%	0	0	10%	1
23	52	IDC	G2	T2	N0	55%	27%	0	18%	28%	1
24	78	ILC	G2	T2	N3	90%	90%	2+	1%	5%	1
25	39	IDC	G2	T1c	N1a	70%	80%	3+	5%	15%	0
26	72	IDC	G3	T2		90%	10%	3+	30%	30%	2
27	63	ILC	G2	T1c		80%	50%	0	10%	10%	2

Continues on the next page

Patient	Age	Type	Grade	Size	Nodes	ER	PR	HER2	p53	Ki67	SOX2
28	46	ILC	G2	T1		70%	90%	0	0	5%	1
29	61	IDC	G2	T1a	N0	90%	40%	0	0	5%	1
30	78	IDC	G2	T2	N1	90%	90%	3+	0	8%	0
31	37	IDC	G2	T1c	N0	26%	87%	3+	63%	39%	0
32	63	IDC	G1	T1c	N0	90%	60%	0	5%	8%	1
33	35	IDC	G2	T2	N1c	30%	40%	3+	0	3%	0
34T	43	IDC	G2	T2	N0	60%	65%	0	0	10%	3
34R	49	IDC	G2	T1		40%	0	1+	0	2%	8
35T	56	IDC	G1	T1	N0	90%	60%	0	<5%	8%	4
35R	67	IDC	G1			60%	20%	0	0	10%	7
36T	44	IDC	G1	T1	N0	90%	90%	0	0	2%	3
36R	52	IDC	G2			95%	80%	1+	<5%	15%	8
37T	61	IDC	G1	T2	N0	90%	60%	1+	10%	35%	3
37R	65	IDC	G2	T1c		40%	0	1+	5%	30%	6
38T	57	IDC	G2	T2	N1a	80%	70%	0	<5%	10%	3
38R	69	IDC	G2	T1		98%	30%	1+	0	12%	6
39T	39	IDC	G2	T2	N0	90%	15%	3+	0	80%	3
39R	48	IDC	G3			90%	0	3+	0	60%	5
40T	59	IDC	G2	T2	N2a	80%	20%	2+	2%	25%	3
40R	62	IDC	G3			90%	0	1+	3%	15%	5
41T	86	IDC	G2	T2	Nx	85%	20%	0	6%	7%	4
41R	90	IDC	G2			90%	60%	0	60%	20%	8
42T	77	IDC	G3	T1c	N2a	40%	10%	3+	0	40%	3
42R	83	IDC	G3			90%	0	3+	40%	30%	3
42Rb	87	IDC	G3	rT2		23%	0	3+	5%	16%	3
43T	86	IDC	G2	T2	Nx	85%	0	3+	5%	40%	3
43R	89	IDC	G2								8
44T	70	IDC	G2	T2	N1a	95%	0	0	0	10%	3
44R	78	IDC	G2	T4b	N1a	95%	0	0	0	50%	6
45T	39	ILC	G2	T2	N1a	40%	40%	0	30%	15%	5
45R	44	ILC	G2	T2		10%	70%	2+	10%	20%	6
46T	55	IDC	G3	T2	N0	80%	65%	0	0	15%	3
46R	60	IDC	G3	T2		67%	7%	1+	0	17%	6
47T	52	IDC	G3								5
47R	62	IDC	G3	T2	N0	0	0	3+	0	20%	5
48T	45	IDC	G1	T2	N1a	90%	20%	0	0	30%	4
48R	51	IDC	G3	T1b		50%	0	0	0	40%	4
48Rb	54	IDC	G3			50%	0	0	0	40%	5
49T	49	ILC	G2	T1c	N1	60%	65%	2+	0	20%	4
49R	52	ILC	G3	T1b		0	0	3+	2%	20%	
50T	59	IDC	G2			80%	20%	0	2%	25%	4
50R	62	IDC	G2			90%	0	0	3%	15%	8
50Rb	63	IDC	G2			0	0	0	0	20%	8
51T	86	IDC	G3	T2		90%	0	2+	3%	40%	4
51R	89	IDC	G3								

Continues on the next page

Patient	Age	Type	Grade	Size	Nodes	ER	PR	HER2	p53	Ki67	SOX2
52T	33	IDC	G3	T1c	N0	30%	25%	3+	0	60%	
52R	44	IDC	G3			95%	10%	0	0	75%	4
53T	46	IDC, ILC	G2			50%	50%	0	0	5%	4
53R	57	IDC	G2			90%	20%	0	0	20%	4
54T	47	IDC	G2								3
54R	55	IDC	G3	T2		90%	20%	0	0	8%	3
55T	81	IDC	G2			40%	30%	0	0	20%	3
55R	82	IDC	G2	T1							3
55Rb	84	IDC	G2			70%	0				3

A.2 SOX2 Clinical dataset after data imputation

Table A.2: Complete SOX2 clinical dataset after filling in the missing data with the MICE methodology. In ***emphatic bold***, inputted data points. In green, patients that responded to therapy; in red, those who developed a recurrence.

Patient	Age	Type	Grade	Size	Nodes	ER	PR	HER2	p53	Ki67	SOX2
1	50	IDC	G2	T2	N0	90%	70%	0	5%	10%	2
2	46	IDC	G2	T1c	N1a	90%	90%	0	0	20%	2
3	54	IDC	G2	T1c	N1a	90%	90%	0	0	0	2
4	53	IDC	G2	T2	N1a	90%	90%	0	0	25%	2
5	59	IDC	G2	T1	N1a	90%	80%	0	0	5%	2
6	42	IDC	G2	T1	N0	70%	60%	0	60%	0	2
7	43	IDC	G2	T1	N0	80%	30%	0	0	30%	2
8	43	IDC	G2	T2	N2a	80%	80%	0	0	10%	2
9	55	IDC	G2	T1	N0	90%	90%	0	0	10%	2
10	41	IDC	G2	T1	N0	80%	80%	0	0	15%	2
11	45	IDC	G1	T1c	N1a	90%	90%	0	0	5%	2
12	82	IDC	***G2***	T2	N0	90%	50%	0	0	40%	2
13	65	IDC	G2	T1c	N1a	85%	30%	0	2%	10%	2
14	43	IDC	G3	T2	N1	70%	70%	2+	30%	25%	2
15	60	IDC	G1	T2	N0	80%	60%	0	0	5%	2
16	35	ILC	G2	T3	N1	70%	80%	0	5%	8%	1
17	53	IDC	G2	T2	N0	90%	90%	0	0	8%	0
18	81	IDC	G1	T2	N1	80%	70%	2+	0	6%	2
19	31	IDC	G3	T2	N1b	70%	90%	2+	0	***19.5***	2
20	78	IDC	G2	T2	***N1***	90%	90%	2+	0	2%	1
21	55	ILC	G2	T1c	N1	80%	70%	0	10%	7%	2
22	73	IDC	G2	T4d	N1a	90%	90%	0	0	10%	1
23	52	IDC	G2	T2	N0	55%	27%	0	18%	28%	1
24	78	ILC	G2	T2	N3	90%	90%	2+	1%	5%	1
25	39	IDC	G2	T1c	N1a	70%	80%	3+	5%	15%	0
26	72	IDC	G3	T2	***N1***	90%	10%	3+	30%	30%	2
27	63	ILC	G2	T1c	***N0***	80%	50%	0	10%	10%	2
28	46	ILC	G2	T1	***N1***	70%	90%	0	0	5%	1

Continues on the next page

Patient	Age	Type	Grade	Size	Nodes	ER	PR	HER2	p53	Ki67	SOX2
29	61	IDC	G2	T1a	N0	90%	40%	0	0	5%	1
30	78	IDC	G2	T2	N1	90%	90%	3+	0	8%	0
31	37	IDC	G2	T1c	N0	26%	87%	3+	63%	39%	0
32	63	IDC	G1	T1c	N0	90%	60%	0	5%	8%	1
33	35	IDC	G2	T2	N1c	30%	40%	3+	0	3%	0
34T	43	IDC	G2	T2	N0	60%	65%	0	0	10%	3
34R	49	IDC	G2	T1	***N1***	40%	0	1+	0	2%	8
35T	56	IDC	G1	T1	N0	90%	60%	0	<5%	8%	4
35R	67	IDC	G1	***T2***	**N1**	60%	20%	0	0	10%	7
36T	44	IDC	G1	T1	N0	90%	90%	0	0	2%	3
36R	52	IDC	G2	***T2***	***N1***	95%	80%	1+	<5%	15%	8
37T	61	IDC	G1	T2	N0	90%	60%	1+	10%	35%	3
37R	65	IDC	G2	T1c	***N0***	40%	0	1+	5%	30%	6
38T	57	IDC	G2	T2	N1a	80%	70%	0	<5%	10%	3
38R	69	IDC	G2	T1	***N1***	98%	30%	1+	0	12%	6
39T	39	IDC	G2	T2	N0	90%	15%	3+	0	80%	3
39R	48	IDC	G3	***T1c***	***N1***	90%	0	3+	0	60%	5
40T	59	IDC	G2	T2	N2a	80%	20%	2+	2%	25%	3
40R	62	IDC	G3	***T2***	***N1***	90%	0	1+	3%	15%	5
41T	86	IDC	G2	T2	Nx	85%	20%	0	6%	7%	4
41R	90	IDC	G2	***T2***	***N0***	90%	60%	0	60%	20%	8
42T	77	IDC	G3	T1c	N2a	40%	10%	3+	0	40%	3
42R	83	IDC	G3	***T1c***	***N2***	90%	0	3+	40%	30%	3
42Rb	87	IDC	G3	rT2	***N2***	23%	0	3+	5%	16%	3
43T	86	IDC	G2	T2	Nx	85%	0	3+	5%	40%	3
44T	70	IDC	G2	T2	N1a	95%	0	0	0	10%	3
44R	78	IDC	G2	T4b	N1a	95%	0	0	0	50%	6
45T	39	ILC	G2	T2	N1a	40%	40%	0	30%	15%	5
45R	44	ILC	G2	T2	***N1***	10%	70%	2+	10%	20%	6
46T	55	IDC	G3	T2	N0	80%	65%	0	0	15%	3
46R	60	IDC	G3	T2	***N1***	67%	7%	1+	0	17%	6
47R	62	IDC	G3	T2	N0	0	0	3+	0	20%	5
48T	45	IDC	G1	T2	N1a	90%	20%	0	0	30%	4
48R	51	IDC	G3	T1b	***N0***	50%	0	0	0	40%	4
48Rb	54	IDC	G3	***T1c***	***N0***	50%	0	0	0	40%	5
49T	49	ILC	G2	T1c	N1	60%	65%	2+	0	20%	4
49R	52	ILC	G3	T1b	***N0***	0	0	3+	2%	20%	***5***
50T	59	IDC	G2	***T1b***	***N1***	80%	20%	0	2%	25%	4
50R	62	IDC	G2	***T1c***	***N1***	90%	0	0	3%	15%	8
50Rb	63	IDC	G2	***T2***	***N1***	0	0	0	0	20%	8
51T	86	IDC	G3	T2	***N1***	90%	0	2+	3%	40%	4
52T	33	IDC	G3	T1c	N0	30%	25%	3+	0	60%	***4***
52R	44	IDC	G3	***T2***	***N0***	95%	10%	0	0	75%	4
53T	46	IDC, ILC	G2	***T1***	***N1***	50%	50%	0	0	5%	4
53R	57	IDC	G2	***T1***	***N1***	90%	20%	0	0	20%	4

Continues on the next page

Patient	Age	Type	Grade	Size	Nodes	ER	PR	HER2	p53	Ki67	SOX2
54R	55	IDC	G3	T2	***N1***	90%	20%	0	0	8%	3
55T	81	IDC	G2	***T1c***	***N0***	40%	30%	0	0	20%	3

Appendix B

Supplementary information for EnCaRes

B.1 Individual survival analysis of the genes in EnCaRes

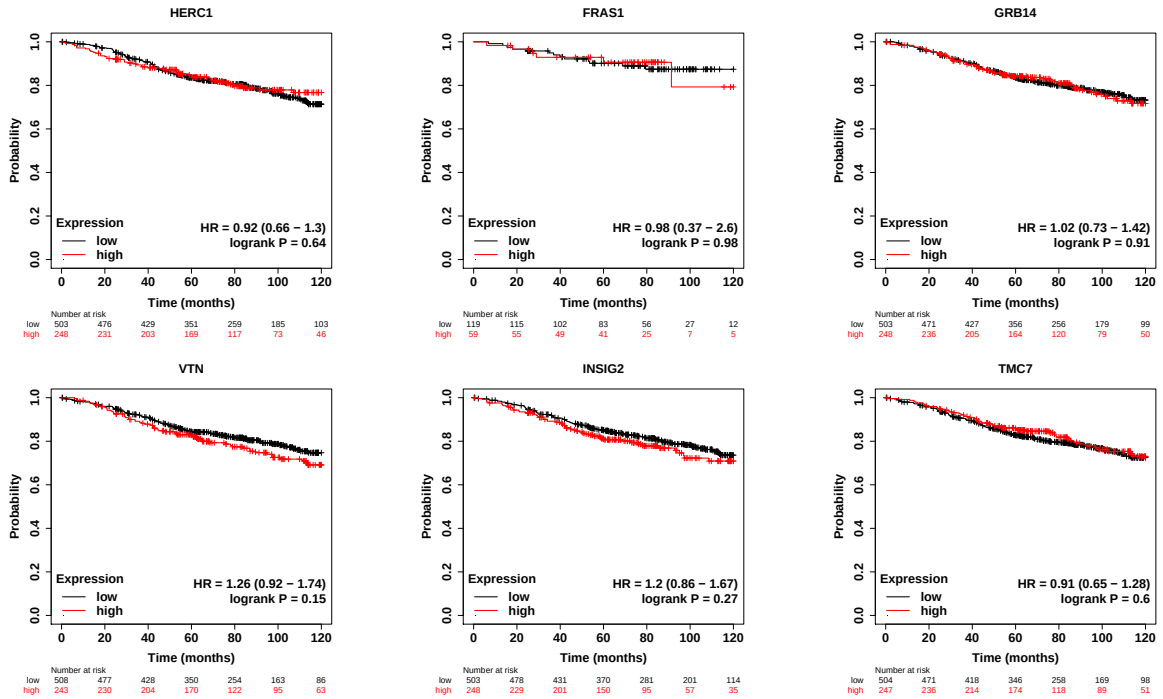

Figure B.1: Individual Kaplan-Meier survival curves for each genes in the EnCaRes signatures. The lack of an effect in any of them individually emphasizes their impact as a signature.

B.2 Clinical information of selected patients

Clinical information	Tamoxifen treated (R/GR)	All HT treated (R/GR)
ER IHC Status		
Positive	37 (12/25)	127 (40/87)
PR IHC Status		
Positive	31 (9/22)	111 (32/79)
Negative	6 (3/3)	16 (8/8)
HER2 Status		
Positive	5 (2/3)	18 (7/11)
Negative	32 (10/22)	109 (33/76)
Age		
>=50	25 (6/19)	30 (9/21)
<50	12 (6/6)	97 (31/66)
Menopause Status		
Pre	21 (6/15)	25 (8/17)
Post	12 (6/6)	97 (32/65)
Peri/NA	4 (0/4)	5 (0/5)
Tumour Grade		
T1	8 (1/7)	37 (7/30)
T2	24 (10/14)	70 (26/44)
T3	5 (1/4)	16 (4/12)
T4	0 (0/0)	4 (3/1)
Node affectation		
NX	0 (0/0)	2 (2/0)
N0	18 (5/13)	62 (14/48)
N1	9 (3/6)	37 (12/25)
N2	9 (4/5)	18 (8/10)
N3>	1 (0/1)	8 (4/4)
Metastatic Status		
MX	3 (1/2)	9 (5/4)
M0	33 (10/23)	113 (31/82)
M1	1 (1/0)	5 (4/1)
Cancer Stage		
I	6 (1/5)	27 (4/23)
II	19 (7/12)	63 (20/43)
III	11 (3/8)	32 (12/20)
IV	1 (1/0)	5 (4/1)

Table B.1: Clinical characteristics of patients from TCGA (tamoxifen and all hormone therapy, ALL HT) datasets. In parentheses, distributions of resistant (R) and good responder (GR) patients in each category are presented.

Bibliography

Abu-Hanna, A. and Lucas, P. J. (2001). Prognostic models in medicine. *Methods of information in medicine*, 40(01):1–5. See page 23.

Afgan, E., Baker, D., Van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., et al. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic acids research*, 44(W1):W3–W10. See page 69.

Agarwal, M., Vats, D., and Elvira, V. (2022). A principled stopping rule for importance sampling. *Electronic Journal of Statistics*, 16(2):5570–5590. See pages 41, 43.

Akhmatskaya, E., Fernández-Pendás, M., Radivojevic, T., and Sanz-Serna, J. (2017). Adaptive splitting integrators for enhancing sampling efficiency of modified Hamiltonian Monte Carlo methods in molecular simulation. *Langmuir*, 33(42):11530–11542. See pages 37, 47, 48, 49, 52, 57, and 59.

Akhmatskaya, E. and Reich, S. (2006). The targeted shadowing hybrid Monte Carlo (TSHMC) method. In *New Algorithms for Macromolecular Simulation*, pages 141–153. Springer. See page 33.

Akhmatskaya, E. and Reich, S. (2008). GSHMC: An efficient method for molecular simulation. *Journal of Computational Physics*, 227(10):4934–4954. See pages 33, 59.

Alhamdoosh, M., Law, C. W., Tian, L., Sheridan, J. M., Ng, M., and Ritchie, M. E. (2017). Easy and efficient ensemble gene set testing with EGSEA. *F1000Research*, 6. See page 83.

Allred, D., Harvey, J. M., Berardo, M., and Clark, G. M. (1998). Prognostic and predictive factors in breast cancer by immunohistochemical analysis. *Modern pathology: an official journal of the United States and Canadian Academy of Pathology, Inc*, 11(2):155–168. See page 19.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1. See pages 8, 77.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *bioinformatics*, 31(2):166–169. See page 69.

Asghari, A., Wall, K., Gill, M., Del Vecchio, N., Allahbakhsh, F., Wu, J., Deng, N., Zheng, W. J., Wu, H., Umetani, M., et al. (2022). A novel group of genes that cause endocrine resistance in breast cancer identified by dynamic gene expression analysis. *Oncotarget*, 13:600. See page 94.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29. See page 87.

Bayes, T. and Price, R. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London*, (53):370–418. See page 5.

Beatson, G. T. (1896). On the treatment of inoperable cases of carcinoma of the mamma: suggestions for a new method of treatment, with illustrative cases. *Transactions. Medico-Chirurgical Society of Edinburgh*, 15:153–179. See page 4.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300. See pages 81, 82.

Bera, A., Subramanian, M., Karaian, J., Eklund, M., Radhakrishnan, S., Gana, N., Rothwell, S., Pollard, H., Hu, H., Shriver, C. D., et al. (2020). Functional role of vitronectin in breast cancer. *PLoS One*, 15(11):e0242141. See page 111.

Berry, D. A. (2006). Bayesian clinical trials. *Nature reviews Drug discovery*, 5(1):27–36. See page 6.

Blanchet, J. and Lam, H. (2012). State-dependent importance sampling for rare-event simulation: An overview and recent advances. *Surveys in Operations Research and Management Science*, 17(1):38–59. See page 32.

Blanes, S., Casas, F., and Murua, A. (2008). Splitting and composition methods in the numerical integration of differential equations. *Bol. Soc. Esp. Mat. Apl.*, 45:89–145. See pages , 34.

Blanes, S., Casas, F., and Sanz-Serna, J. M. (2014). Numerical Integrators for the hybrid Monte Carlo Method. *SIAM Journal on Scientific Computing*, 36(4):A1556–A1580. See pages , 36.

Bou-Rabee, N. and Sanz-Serna, J. M. (2018). Geometric integrators and the Hamiltonian Monte Carlo method. *Acta Numerica*, 27:113–206. See pages 36, 52.

Broca, P. (1866). *Traité des tumeurs: I*, volume 1. P. Asselin. See page 4.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. CRC press. See page 26.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455. See page 38.

Brueffer, C., Vallon-Christersson, J., Grabau, D., Ehinger, A., Häkkinen, J., Hegardt, C., Malina, J., Chen, Y., Bendahl, P.-O., Manjer, J., et al. (2018). Clinical value of RNA sequencing–based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter Sweden Cancerome Analysis Network—Breast Initiative. *JCO Precision Oncology*, 2:1–18. See page 92.

Bzdok, D., Engemann, D., and Thirion, B. (2020). Inference and prediction diverge in biomedicine. *Patterns*, 1(8). See page 64.

Cailleau, R., Young, R., Olive, M., and Reeves Jr, W. (1974). Breast tumor cell lines from pleural effusions. *Journal of the National Cancer Institute*, 53(3):661–674. See page 72.

Calvo, M. P., Sanz-Alonso, D., and Sanz-Serna, J. M. (2021). HMC: Reducing the number of rejections by not using leapfrog and some results on the acceptance rate. *Journal of Computational Physics*, 437:110333. See pages 51, 52, and 53.

Campos, C. M. and Sanz-Serna, J. M. (2017). Palindromic 3-stage splitting integrators, a roadmap. *Journal of Computational Physics*, 346:340–355. See pages 36, 47.

Cao, L. and Liu, Q. (2022). COVID-19 modeling: a review. *medRxiv*, pages 2022–08. See page 6.

Černỳ, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of optimization theory and applications*, 45:41–51. See pages 27, 102.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2024). shiny: Web Application Framework for R. R package version 1.8.0.9000, https://github.com/rstudio/shiny. See page 85.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma'ayan, A. (2013). EnrichR: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, 14(1):1–14. See page 83.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData mining*, 10(1):35. See page 7.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology*, 17(1):1–19. See pages 8, 69, and 70.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons. See page 41.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202. See page 107.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563. See page 1.

Croft, D., O'kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., et al. (2010). Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(suppl_1):D691–D697. See pages 83, 87.

Cule, E., Vineis, P., and De Iorio, M. (2011). Significance testing in ridge regression for genetic data. *BMC bioinformatics*, 12:1–15. See page 81.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352. See page 105.

De Sanctis, G., Colombo, R., Damiani, C., Sacco, E., and Vanoni, M. (2018). -omics and clinical data integration. *Integration of Omics Approaches and Systems Biology for Clinical Applications*, pages 248–273. See page 92.

Dhruva, S. S., Ross, J. S., Akar, J. G., Caldwell, B., Childers, K., Chow, W., Ciaccio, L., Coplan, P., Dong, J., Dykhoff, H. J., et al. (2020). Aggregating multiple real-world data sources using a patient-centered health-data-sharing platform. *NPJ digital medicine*, 3(1):60. See page 17.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21. See page 69.

Domenici, G., Aurrekoetxea-Rodríguez, I., Simões, B. M., Rábano, M., Lee, S. Y., Millán, J. S., Comaills, V., Oliemuller, E., López-Ruiz, J. A., Zabalza, I., et al. (2019). A Sox2–Sox9 signalling axis maintains human breast luminal progenitor and breast cancer stem cells. *Oncogene*, 38(17):3151–3169. See pages 72, 93.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222. See pages , 27.

Duncan, A. B., Lelievre, T., and Pavliotis, G. A. (2016). Variance reduction using nonreversible Langevin samplers. *Journal of statistical physics*, 163:457–491. See page 31.

Elvira, V., Martino, L., and Robert, C. P. (2022). Rethinking the effective sample size. *International Statistical Review*, 90(3):525–550. See page 40.

Engle, R. D., Skeel, R. D., and Drees, M. (2005). Monitoring energy drift with shadow hamiltonians. *Journal of Computational Physics*, 206(2):432 – 452. See page 33.

Fang, Y., Sanz-Serna, J.-M., and Skeel, R. D. (2014). Compressible generalized Hybrid Monte Carlo. *The Journal of chemical physics*, 140(17). See page 30.

Fernández-Pendás, M., Akhmatskaya, E., and Sanz-Serna, J. M. (2016). Adaptive multi-stage integrators for optimal energy conservation in molecular simulations. *Journal of Computational Physics*, 327:434–449. See pages 37, 47.

Ford, M. and Finlayson, C. A. (2010). Surgical therapy of early breast cancer. In *Early Diagnosis and Treatment of Cancer Series: Breast Cancer*, chapter 12, pages 185–190. Elsevier Health Sciences. See page 4.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling*, 6(269):pl1–pl1. See page 105.

Gao, T., Han, Y., Yu, L., Ao, S., Li, Z., and Ji, J. (2014). CCNA2 is a prognostic biomarker for ER+ breast cancer and tamoxifen resistance. *PloS one*, 9(3):e91771. See page 93.

García-Campos, M. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2015). Pathway analysis: state of the art. *Frontiers in physiology*, 6:383. See page 9.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. *Bayesian statistics*, 4:147–167. See page 101.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC. See pages 23, 26.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. See pages 63, 100.

Gelman, A. and Loken, E. (2014). The statistical crisis in science. *American scientist*, 102(6):460–465. See page 64.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472. See page 38.

Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y., and De Moor, B. (2016). Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, 22(14):e184–90. See page 92.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical science*, pages 473–483. See page 39.

Geyer, C. J. and Johnson, L. T. (2020). mcmc: Markov Chain Monte Carlo. R package version 0.9-7. See page 42.

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2):123–214. See page 44.

Glusman, G., Caballero, J., Robinson, M., Kutlu, B., and Hood, L. (2013). Optimal scaling of digital transcriptomes. *PloS one*, 8(11):e77885. See page 77.

Goh, W. W. B. and Wong, L. (2018). Why breast cancer signatures are no better than random signatures explained. *Drug Discovery Today*, 23(11):1818–1823. See pages 13, 93, and 115.

Goto, N., Hiyoshi, H., Ito, I., Tsuchiya, M., Nakajima, Y., and Yanagisawa, J. (2011). Estrogen and antiestrogens alter breast cancer invasiveness by modulating the transforming growth factor-β signaling pathway. *Cancer science*, 102(8):1501–1508. See page 111.

Group, E. B. C. T. C. (2011). Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *The lancet*, 378(9793):771–784. See page 11.

Győrffy, B. (2021). Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. *Computational and structural biotechnology journal*, 19:4101–4109. See page 105.

Harbeck, N., Penault-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., Ruddy, K., Tsang, J., and Cardoso, F. (2019). Breast cancer. *Nature Reviews Disease Primers*, 5(1). See page 10.

Hardcastle, T. J. and Kelly, K. A. (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11:1–14. See page 9.

Harrod, A., Lai, C.-F., Goldsbrough, I., Simmons, G. M., Oppermans, N., Santos, D. B., Győrffy, B., Allsopp, R. C., Toghill, B. J., Balachandran, K., et al. (2022). Genome engineering for estrogen receptor mutations reveals differential responses to anti-estrogens and new prognostic gene signatures for breast cancer. *Oncogene*, 41(44):4905–4915. See pages , 13, 93, and 108.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. See page 26.

Hauser, A., Counotte, M. J., Margossian, C. C., Konstantinoudis, G., Low, N., Althaus, C. L., and Riou, J. (2020). Estimation of SARS-CoV-2 mortality during the early stages of an epidemic: A modeling study in Hubei, China, and six regions in Europe. *PLoS medicine*, 17(7):e1003189. See page 6.

Held, L. and Ott, M. (2018). On p-values and Bayes factors. *Annual Review of Statistics and Its Application*, 5:393–419. See page 7.

Hermawan, A., Putri, H., and Ikawati, M. (2020). Bioinformatic analysis reveals the molecular targets of tangeretin in overcoming the resistance of breast cancer to tamoxifen. *Gene Reports*, 21:100884. See pages 13, 93.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67. See page 63.

Hoffman, M. D., Gelman, A., et al. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623. See pages 44, 59.

Hofmann, H. (1994). Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77. See page 59.

Horowitz, A. M. (1991). A generalized guided monte carlo algorithm. *Physics Letters B*, 268(2):247 – 252. See page 30.

Horwich, P. (1993). World changes: Thomas Kuhn and the nature of science. See page 1.

Hrdlickova, R., Toloue, M., and Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*, 8(1):e1364. See page 69.

Hsiao, T.-H., Chen, H.-I. H., Lu, J.-Y., Lin, P.-Y., Keller, C., Comerford, S., Tomlinson, G. E., and Chen, Y. (2013). Utilizing signature-score to identify oncogenic pathways of cholangiocarcinoma. *Translational cancer research*, 2(1):6. See page 99.

Huang, L., Zhao, S., Frasor, J. M., and Dai, Y. (2011). An integrated bioinformatics approach identifies elevated cyclin E2 expression and E2F activity as distinct features of tamoxifen resistant breast tumors. *PloS one*, 6(7):e22274. See page 93.

Huang, O., Jiang, M., Zhang, X., Xie, Z., Chen, X., Wu, J., Liu, H., and Shen, K. (2013). GRB14 as an independent good prognosis factor for breast cancer patients treated with neoadjuvant chemotherapy. *Japanese journal of clinical oncology*, 43(11):1064–1072. See page 111.

Huggins, J. H. and Roy, D. M. (2019). Sequential Monte Carlo as approximate sampling: bounds, adaptive resampling via ∞-ESS, and an application to particle Gibbs. See page 40.

Inouzhe, H., Rodríguez-Álvarez, M. X., Nagar, L., and Akhmatskaya, E. (2023). Dynamic SIR/SEIR-like models comprising a time-dependent transmission rate: Hamiltonian Monte Carlo approach with applications to COVID-19. *arXiv preprint arXiv:2301.06385*. See page 6.

Jack Lee, J. and Chu, C. T. (2012). Bayesian clinical trials in action. *Statistics in medicine*, 31(25):2955–2972. See page 5.

Jain, N., Nitisa, D., Pirsko, V., and Cakstina, I. (2020). Selecting suitable reference genes for qPCR normalization: a comprehensive analysis in MCF-7 breast cancer cell line. *BMC Molecular and Cell Biology*, 21:1–19. See page 79.

Jang, J. H., Manatunga, A. K., Chang, C., and Long, Q. (2021). A Bayesian multiple imputation approach to bivariate functional data with missing components. *Statistics in medicine*, 40(22):4772–4793. See page 6.

Jeselsohn, R., Buchwalter, G., De Angelis, C., Brown, M., and Schiff, R. (2015). ESR1 mutations—a mechanism for acquired endocrine resistance in breast cancer. *Nature reviews Clinical oncology*, 12(10):573–583. See pages 12, 92.

Kahn, H. and Marshall, A. W. (1953). Methods of reducing sample size in Monte Carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278. See page 31.

Kairouz, R., Parmar, J., Lyons, R. J., Swarbrick, A., Musgrove, E. A., and Daly, R. J. (2005). Hormonal regulation of the GRB14 signal modulator and its role in cell cycle progression of MCF-7 human breast cancer cells. *Journal of cellular physiology*, 203(1):85–93. See page 111.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30. See pages 83, 87.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481. See page 106.

Karabatak, M. (2015). A new classifier for breast cancer detection based on Naïve Bayesian. *Measurement*, 72:32–36. See page 6.

Kassambara, A., Kosinski, M., and Biecek, P. (2021). survminer: Drawing survival curves using 'ggplot2'. R package version 0.4.9. See page 106.

Katzke, V. A., Kaaks, R., and Kühn, T. (2015). Lifestyle and cancer risk. *The Cancer Journal*, 21(2):104–110. See page 4.

Keller, E. F. (2005). The century beyond the gene. *Journal of Biosciences*, 30:3–10. See page 1.

Kelly, M. J. (1989). Computers: the best friends a human genome ever had. *Genome*, 31(2):1027–1033. See page 2.

Kennedy, A. and Pendleton, B. (2001). Cost of the generalised hybrid monte carlo algorithm for free field theory. *Nuclear Physics B*, 607(3):456 – 510. See page 30.

Keydar, I., Chen, L., Karby, S., Weiss, F., Delarea, J., Radu, M., Chaitcik, S., and Brenner, H. (1979). Establishment and characterization of a cell line of human breast carcinoma origin. *European Journal of Cancer (1965)*, 15(5):659–670. See page 72.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2):e1002375. See pages , 83.

Kirkpatrick, S., Gelatt Jr, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *science*, 220(4598):671–680. See pages 27, 102.

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American statistical association*, 89(425):278–288. See page 40.

Koren, S. and Bentires-Alj, M. (2015). Breast tumor heterogeneity: source of fitness, hurdle for therapy. *Molecular cell*, 60(4):537–546. See pages 10, 93, and 111.

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., and Sergushichev, A. (2016). Fast gene set enrichment analysis. *BioRxiv*, page 060012. See page 83.

Korotkevich, G., Sukhov, V., and Sergushichev, A. (2019). Fast gene set enrichment analysis. *bioRxiv*. See page 111.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago press. See page 1.

Kunzmann, K., Grayling, M. J., Lee, K. M., Robertson, D. S., Rufibach, K., and Wason, J. M. (2021). A review of Bayesian perspectives on sample size derivation for confirmatory trials. *The American Statistician*, 75(4):424–432. See page 6.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–359. See page 69.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The sequence alignment/map format and SAMtools. *bioinformatics*, 25(16):2078–2079. See page 69.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930. See page 69.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425. See page 87.

Lopes, H. F., Müller, P., and Ravishanker, N. (2007). Bayesian computational methods in biomedical research. *Computational Methods in Biomedical Research. Boca Raton: Chapman & Hall/CRC*, pages 211–259. See page 60.

López-Ruiz, J. A., Mieza, J. A., Zabalza, I., and Vivanco, M. d. M. (2022). Comparison of genomic profiling data with clinical parameters: Implications for breast cancer prognosis. *Cancers*, 14(17):4197. See page 11.

Love, M. ., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):1–21. See pages , 9, 77, 79, and 82.

Lu, N., Zhang, M., Lu, L., Liu, Y.-z., Liu, X.-d., and Zhang, H.-h. (2021). Insulin-Induced Gene 2 Expression Is Associated with Breast Cancer Metastasis. *The American Journal of Pathology*, 191(2):385–395. See page 111.

Ma, X.-J., Wang, Z., Ryan, P. D., Isakoff, S. J., Barmettler, A., Fuller, A., Muir, B., Mohapatra, G., Salunga, R., Tuggle, J. T., et al. (2004). A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer cell*, 5(6):607–616. See pages , 13, 93, and 108.

Magnani, L., Stoeck, A., Zhang, X., Lánczky, A., Mirabella, A. C., Wang, T.-L., Gyorffy, B., and Lupien, M. (2013). Genome-wide reprogramming of the chromatin landscape underlies endocrine therapy resistance in breast cancer. *Proceedings of the National Academy of Sciences*, 110(16):E1490–E1499. See page 93.

Makki, J. (2015). Diversity of breast carcinoma: histological subtypes and clinical relevance. *Clinical medicine insights: Pathology*, 8:CPath–S31563. See page 10.

Makowski, D., Ben-Shachar, M. S., Chen, S. A., and Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in psychology*, 10:2767. See pages 7, 64.

Mani, S., Chen, Y., Li, X., Arlinghaus, L., Chakravarthy, A. B., Abramson, V., Bhave, S. R., Levy, M. A., Xu, H., and Yankeelov, T. E. (2013). Machine learning for predicting the response of breast cancer to neoadjuvant chemotherapy. *Journal of the American Medical Informatics Association*, 20(4):688–695. See page 6.

Manjang, K., Tripathi, S., Yli-Harja, O., Dehmer, M., Glazko, G., and Emmert-Streib, F. (2021). Prognostic gene expression signatures of breast cancer are lacking a sensible biological meaning. *Scientific Reports*, 11(1):156. See pages 13, 93, and 115.

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., and Ferrari, R. (2018). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in bioinformatics*, 19(2):286–302. See page 7.

Martino, L., Elvira, V., and Louzada, F. (2017). Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, 131:386–401. See page 40.

McLachlan, R. I. and Atela, P. (1992). The accuracy of symplectic integrators. *Nonlinearity*, 5(2):541. See pages 36, 55.

Men, X., Ma, J., Wu, T., Pu, J., Wen, S., Shen, J., Wang, X., Wang, Y., Chen, C., and Dai, P. (2018). Transcriptome profiling identified differentially expressed genes and pathways associated with tamoxifen resistance in human breast cancer. *Oncotarget*, 9(3):4074. See pages , 94, 97, 108, and 111.

Menendez, J. A., Papadimitropoulou, A., Vander Steen, T., Cuyàs, E., Oza-Gajera, B. P., Verdura, S., Espinoza, I., Vellon, L., Mehmi, I., and Lupu, R. (2021). Fatty acid synthase confers tamoxifen resistance to ER+/HER2+ breast cancer. *Cancers*, 13(5):1132. See page 109.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092. See page 26.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American statistical association*, 44(247):335–341. See page 25.

Mihály, Z., Kormos, M., Lánczky, A., Dank, M., Budczies, J., Szász, M. A., and Győrffy, B. (2013). A meta-analysis of gene expression-based biomarkers predicting outcome after tamoxifen treatment in breast cancer. *Breast cancer research and treatment*, 140:219–232. See pages 13, 93.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L. M., Ding, W., et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, 266(5182):66–71. See page 4.

Miller, P. C., Clarke, J., Koru-Sengul, T., Brinkman, J., and El-Ashry, D. (2015). A novel MAPK–microRNA signature is predictive of hormone-therapy resistance and poor outcome in er-positive breast cancer. *Clinical cancer research*, 21(2):373–385. See pages 13, 93, and 115.

Mills, J. N., Rutkovsky, A. C., and Giordano, A. (2018). Mechanisms of resistance in estrogen receptor positive breast cancer: overcoming resistance to tamoxifen/aromatase inhibitors. *Current opinion in pharmacology*, 41:59–65. See page 102.

Moan, P. C. (2014). On an asymptotic method for computing the modified energy for symplectic methods. *Discrete & Continuous Dynamical Systems - A*, 34(1078-0947_2014_3_1105):1105. See page 33.

Nagar, L., Fernández-Pendás, M., Sanz-Serna, J. M., and Akhmatskaya, E. (2024). Adaptive multi-stage integration schemes for Hamiltonian Monte Carlo. *Journal of Computational Physics*, 502:112800. See pages 36, 37, 56, 59, and 60.

Nault, R., Saha, S., Bhattacharya, S., Dodson, J., Sinha, S., Maiti, T., and Zacharewski, T. (2022). Benchmarking of a Bayesian single cell RNAseq differential gene expression test for dose–response study designs. *Nucleic acids research*, 50(8):e48–e48. See page 9.

Neal, R. M. (1994). An improved acceptance procedure for the Hybrid Monte Carlo algorithm. *Journal of Computational Physics*, 111(1):194–203. See pages , 27.

Neal, R. M. et al. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov chain Monte Carlo*, 2(11):2. See page 56.

Network, T. C. G. A. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70. See page 22.

Nolan, E., Lindeman, G. J., and Visvader, J. E. (2023). Deciphering breast cancer: from biology to the clinic. *Cell*. See pages , 11, and 12.

Oberman, H. (2023). ggmice: Visualizations for 'mice' with 'ggplot2'. R package version 0.1.0. See page 20.

Ortigueira, M. (2010). On the estimation of the autocorrelation function. *Discussiones Mathematicae Probability and Statistics*, 30(1):103–115. See page 39.

Osborne, C. K., Shou, J., Massarweh, S., and Schiff, R. (2005). Crosstalk between estrogen receptor and growth factor receptor pathways as a cause for endocrine therapy resistance in breast cancer. *Clinical cancer research*, 11(2):865s–870s. See page 12.

Ottobre, M. (2016). Markov chain Monte Carlo and irreversibility. *Reports on Mathematical Physics*, 77(3):267–292. See page 31.

Owen, A. B. (2013). *Monte Carlo theory, methods and examples.* https://artowen.su.domains/mc/. See page 32.

Paik, S., Tang, G., Shak, S., Kim, C., Baker, J., Kim, W., Cronin, M., Baehner, F. L., Watson, D., Bryant, J., et al. (2006). Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor–positive breast cancer. *Journal of clinical oncology*, 24(23):3726–3734. See pages , 13, 93, and 108.

Palafox, M., Monserrat, L., Bellet, M., Villacampa, G., Gonzalez-Perez, A., Oliveira, M., Brasó-Maristany, F., Ibrahimi, N., Kannan, S., Mina, L., et al. (2022). High p16 expression and heterozygous RB1 loss are biomarkers for CDK4/6 inhibitor resistance in ER+ breast cancer. *Nature communications*, 13(1):5258. See page 93.

Pan, H., Gray, R., Braybrooke, J., Davies, C., Taylor, C., McGale, P., Peto, R., Pritchard, K. I., Bergh, J., Dowsett, M., et al. (2017). 20-year risks of breast-cancer recurrence after stopping endocrine therapy at 5 years. *New England Journal of Medicine*, 377(19):1836–1846. See pages 12, 72.

Parga-Pazos, M., Cusimano, N., Rábano, M., Akhmatskaya, E., et al. (2024). A Novel Mathematical Approach for Analysis of Integrated Cell–Patient Data Uncovers a 6-Gene Signature Linked to Endocrine Therapy Resistance. *Laboratory Investigation*, 104(1):100286. See page 113.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572. See page 73.

Peng, C., Goswami, P., and Bai, G. (2020). A literature review of current technologies on health data integration for patient-centered health management. *Health informatics journal*, 26(3):1926–1951. See page 17.

Piva, M., Domenici, G., Iriondo, O., Rábano, M., Simoes, B. M., Comaills, V., Barredo, I., López-Ruiz, J. A., Zabalza, I., Kypta, R., et al. (2014). Sox2 promotes tamoxifen resistance in breast cancer cells. *EMBO molecular medicine*, 6(1):66–79. See pages 12, 17, 72, 93, 95, 99, 109, and 111.

Plummer, M., Best, N., Cowles, K., Vines, K., et al. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R news*, 6(1):7–11. See pages 39, 42, 44, 45, and 46.

Polewko-Klim, A., Mnich, K., and Rudnicki, W. R. (2021). Robust data integration method for classification of biomedical data. *J. Med. Syst.*, 45(4):45. See page 92.

Predescu, C., Lippert, R. A., Eastwood, M. P., Ierardi, D., Xu, H., Jensen, M. Ø., Bowers, K. J., Gullingsrud, J., Rendleman, C. A., Dror, R. O., et al. (2012). Computationally efficient molecular dynamics integrators with improved sampling accuracy. *Molecular Physics*, 110(9-10):967–983. See pages 36, 55.

Priedigkeit, N., Ding, K., Horne, W., Kolls, J. K., Du, T., Lucas, P. C., Blohmer, J.-U., Denkert, C., Machleidt, A., Ingold-Heppner, B., et al. (2021). Acquired mutations and transcriptional remodeling in long-term estrogen-deprived locoregional breast cancer recurrences. *Breast Cancer Research*, 23(1):1–14. See page 92.

Qin, D. (2019). Next-generation sequencing and its clinical application. *Cancer biology & medicine*, 16(1):4. See page 4.

Rabiei, R., Ayyoubzadeh, S. M., Sohrabei, S., Esmaeili, M., and Atashi, A. (2022). Prediction of breast cancer using machine learning approaches. *Journal of Biomedical Physics & Engineering*, 12(3):297. See page 6.

Radivojevic, T. (2016). *Enhancing sampling in computational statistics using modified Hamiltonians.* PhD thesis, UPV/EHU. See pages , 32, 33, 36, and 42.

Radivojević, T. and Akhmatskaya, E. (2020). Modified Hamiltonian Monte Carlo for bayesian inference. *Statistics and Computing*, 30(2):377–404. See pages , 34, 42, 44, 45, and 59.

Radivojević, T., Fernández-Pendás, M., Sanz-Serna, J. M., and Akhmatskaya, E. (2018). Multi-stage splitting integrators for sampling with modified Hamiltonian Monte Carlo methods. *Journal of Computational Physics*, 373:900–916. See pages , 36, 37, 48, 52, and 57.

Rahem, S. M., Epsi, N. J., Coffman, F. D., and Mitrofanova, A. (2020). Genome-wide analysis of therapeutic response uncovers molecular pathways governing tamoxifen resistance in ER+ breast cancer. *EBioMedicine*, 61. See pages , 13, 93, and 108.

Rainey, C. (2016). Dealing with separation in logistic regression models. *Political Analysis*, 24(3):339–355. See page 19.

Ramaker, R. C., Lasseigne, B. N., Hardigan, A. A., Palacio, L., Gunther, D. S., Myers, R. M., and Cooper, S. J. (2017). Rna sequencing-based cell proliferation analysis across 19 cancers identifies a subset of proliferation-informative cancers with a common survival signature. *Oncotarget*, 8(24):38668. See page 109.

Robert, C. P., Casella, G., and Casella, G. (2005). *Monte Carlo statistical methods*, volume 2. Springer. See page 26.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140. See page 9.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, 11(3):1–9. See page 76.

Rossi, F. A., Calvo Roitberg, E. H., Enriqué Steinberg, J. H., Joshi, M. U., Espinosa, J. M., and Rossi, M. (2021). HERC1 regulates breast cancer cells migration and invasion. *Cancers*, 13(6):1309. See page 111.

Ruegg, T. A. (2015). Historical perspectives of the causation of lung cancer: nursing as a bystander. *Global qualitative nursing research*, 2:2333393615585972. See page 4.

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., Dimitriadoy, S., Liu, D. L., Kantheti, H. S., Saghafinia, S., et al. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell*, 173(2):321–337. See page 87.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467. See page 7.

Sanz-Serna, J.-M. and Calvo, M.-P. (1994). *Numerical hamiltonian problems*. London: Chapman and Hall. See page 33.

Sarmiento-Castro, A., Caamaño-Gutiérrez, E., Sims, A. H., Hull, N. J., James, M. I., Santiago-Gómez, A., Eyre, R., Clark, C., Brown, M. E., Brooks, M. D., et al. (2020). Increased expression of interleukin-1 receptor characterizes anti-estrogen-resistant ALDH+ breast cancer stem cells. *Stem Cell Reports*, 15(2):307–316. See page 93.

Schlick, T., Mandziuk, M., Skeel, R. D., and Srinivas, K. (1998). Nonlinear resonance artifacts in molecular dynamics simulations. *Journal of Computational Physics*, 140(1):1–29. See page 47.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature methods*, 5(1):16–18. See page 7.

Self, S. C. W., Huang, R., Amin, S., Ewing, J., Rudisill, C., and McLain, A. C. (2022). A Bayesian susceptible-infectious-hospitalized-ventilated-recovered model to predict demand for COVID-19 inpatient care in a large healthcare system. *Plos one*, 17(12):e0260595. See page 6.

Sessa, J. and Syed, D. (2016). Techniques to deal with missing data. In *2016 5th international conference on electronic devices, systems and applications (ICEDSA)*, pages 1–4. IEEE. See page 20.

Simoes, B. M., O'Brien, C. S., Eyre, R., Silva, A., Yu, L., Sarmiento-Castro, A., Alférez, D. G., Spence, K., Santiago-Gomez, A., Chemi, F., et al. (2015). Anti-estrogen resistance in human breast tumors is driven by JAG1-NOTCH4-dependent cancer stem cell activity. *Cell reports*, 12(12):1968–1977. See page 93.

Simon, A. (2015). FastQC. See page 69.

Sinn, B. V., Fu, C., Lau, R., Litton, J., Tsai, T.-H., Murthy, R., Tam, A., Andreopoulou, E., Gong, Y., Murthy, R., et al. (2019). SET ER/PR: a robust 18-gene predictor for sensitivity to endocrine therapy for metastatic breast cancer. *NPJ breast cancer*, 5(1):16. See pages , 13, 93, and 108.

Skeel, R. D. and Hardy, D. J. (2001). Practical Construction of Modified Hamiltonians. *SIAM Journal on Scientific Computing*, 23(4):1172–1188. See page 33.

Skeel, R. D., Zhang, G., and Schlick, T. (1997). A family of symplectic integrators: stability, accuracy, and molecular dynamics applications. *SIAM Journal on Scientific Computing*, 18(1):203–222. See page 28.

Song, J. and Ran, L. (2021). Pan-cancer analysis reveals the signature of TMC family of genes as a promising biomarker for prognosis and immunotherapeutic response. *Frontiers in Immunology*, 12:715508. See page 111.

Song, Z. and Tan, Z. (2022). On irreversible metropolis sampling related to Langevin dynamics. *SIAM Journal on Scientific Computing*, 44(4):A2089–A2120. See page 30.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874. See page 10.

Soule, H. D., Vazquez, J., Long, A., Albert, S., and Brennan, M. (1973). A human cell line from a pleural effusion derived from a breast carcinoma. *Journal of the national cancer institute*, 51(5):1409–1416. See page 72.

Stan Development Team (2024). RStan: the R interface to Stan. R package version 2.32.5. See pages 95, 101, and 120.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550. See pages , 9, and 82.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249. See page 10.

Sutton, A. J. and Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical methods in medical research*, 10(4):277–303. See page 6.

Szostakowska, M., Trębińska-Stryjewska, A., Grzybowska, E. A., and Fabisiewicz, A. (2019). Resistance to endocrine therapy in breast cancer: molecular mechanisms and future goals. *Breast Cancer Research and Treatment*, 173:489–497. See pages 12, 92.

Teng, J., Zhang, H., Liu, W., Shu, X.-O., and Ye, F. (2022). A Dynamic Bayesian Model for Breast Cancer Survival Prediction. *IEEE Journal of Biomedical and Health Informatics*, 26(11):5716–5727. See page 6.

Terry M. Therneau and Patricia M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York. See page 106.

The Cancer Genome Atlas Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120. See page 17.

Therneau, T. M. (2023). A package for survival analysis in r. R package version 3.5-7. See page 106.

Thiébaux, H. J. and Zwiers, F. W. (1984). The interpretation and estimation of effective sample size. *Journal of Applied Meteorology and Climatology*, 23(5):800–811. See page 39.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288. See page 63.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3):219–242. See pages 6, 20.

van Vliet, M. H., Horlings, H. M., van de Vijver, M. J., Reinders, M. J. T., and Wessels, L. F. A. (2012). Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. *PLoS One*, 7(7):e40358. See page 92.

Van't Veer, L. J. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536. See pages 13, 93.

Vats, D., Flegal, J. M., and Jones, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106(2):321–337. See pages 39, 41, and 43.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, 27:1413–1432. See pages 95, 101, and 103.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R ˆ for assessing convergence of MCMC (with discussion). *Bayesian analysis*, 16(2):667–718. See page 38.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*. See pages 7, 101, and 102.

Venet, D., Dumont, J. E., and Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS computational biology*, 7(10):e1002240. See pages 13, 93, 99, 109, and 115.

Verlet, L. (1967). Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Physical review*, 159(1):98. See page 28.

Wang, X. and Wang, S. (2021). Identification of key genes involved in tamoxifen-resistant breast cancer using bioinformatics analysis. *Translational Cancer Research*, 10(12):5246. See pages 13, 93.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63. See page 8.

Wolberg, William, M. O. S. N. and Street, W. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DW2B. See page 61.

Wolff, J., Batut, B., and Rasche, H. (2024). Mapping (galaxy training materials). See pages , 70.

Wu, Z., Liu, W., Jin, X., Ji, H., Wang, H., Glusman, G., Robinson, M., Liu, L., Ruan, J., and Gao, S. (2019). NormExpression: an R package to normalize gene expression data using evaluated methods. *Frontiers in genetics*, 10:400. See pages 77, 79.

Xia, Y., He, X., Renshaw, L., Martinez-Perez, C., Kay, C., Gray, M., Meehan, J., Parker, J. S., Perou, C. M., Carey, L. A., et al. (2022). Integrated DNA and RNA Sequencing Reveals Drivers of Endocrine Resistance in Estrogen Receptor–Positive Breast Cancer. *Clinical Cancer Research*, 28(16):3618–3629. See pages 13, 93, and 115.

Yaşar, P., Ayaz, G., User, S. D., Güpür, G., and Muyan, M. (2017). Molecular mechanism of estrogen–estrogen receptor signaling. *Reproductive medicine and biology*, 16(1):4–20. See page 11.

Zaballa, O., Pérez, A., Inhiesto, E. G., Ayesta, T. A., and Lozano, J. A. (2023). Learning the progression patterns of treatments using a probabilistic generative model. *Journal of Biomedical Informatics*, 137:104271. See page 6.

Zhao, E., Stone, M. R., Ren, X., Guenthoer, J., Smythe, K. S., Pulliam, T., Williams, S. R., Uytingco, C. R., Taylor, S. E., Nghiem, P., et al. (2021). Spatial transcriptomics at subspot resolution with BayesSpace. *Nature biotechnology*, 39(11):1375–1384. See page 9.