Using somatic evolution to unravel cell fate dynamics to gain tissue level insights



Ryan Schenck
University College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy in Genomic Medicine and Statistics

Trinity 2022

Statement of Originality

This thesis is submitted to the University of Oxford in fulfillment of the requirements of the degree of Doctor of Philosophy. Except for portions of Chapter 3 and 4, all work was carried out solely by myself. Patient data in Chapter 3 was collected by Doctor Darryl Shibata at the University of Southern California. The master equation and inference framework presented in Chapter 3 (Section 4.5 and B.1.6) was developed by Mr Callum Gabbutt in our collaboration and included with permission. Patient samples in Chapter 4 were collected by and diagnosed by Doctor Niels Asger Jakobsen and Professor Paresh Vyas and included here with permission. All other contributions are noted directly in text or cited where appropriate.

Acknowledgements

First and foremost, I would like to thank my supervisors Professor Simon Leedham and Professor Alexander Anderson. Professor Leedham, I am grateful for your contagious passion for new ideas and your tutelage. Professor Anderson, you provided me with a chance to pursue something incredible. You have been unwavering in your support and confidence in my ideas, I will forever be grateful. You have both ensured that I have grown as a scientist and as an individual. I also must acknowledge the pervasive guidance of Doctor Darryl Shibata. I am greatly indebted to you for your persistent guidance in my research; I must acknowledge our numerous conversations as the genesis to most of the work that follows.

I also thank Professor Mark Coles and Professor Skirmantas Kriaucionis for ensuring my ideas, research, and progress were sound by serving as my transfer of status and confirmation of status examiners. I thank my co-authors Doctor Chandler Gatenbee, Mr Rafael Bravo, Doctor Mark Robertson Tessi, Doctor Jeffrey West, Mr Callum Gabbutt, and Professor Trevor Graham for stimulating conversations that helped to shape the direction of my research.

Many individuals aided me along the way by providing their support in my efforts or at least by providing a welcomed distraction. Doctor Maximillian Strobl, you have not only been my student colleague throughout my thesis, providing me with invaluable discussions, but you have been a steadfast friend providing unforgettable memories during my time in Oxford and Tampa. Thank you Maxi. Anka, Ester, Nadia, Gabriel, and OUYC thank you for the memories during my time in Oxford. To Jill, Rebecca B, and the rest of IMO thank you for your lunchtime conversations and many adventures. To my DIYC family, especially *Breezin' Bayou* and her crew, thank you for the friendship, refuge, and sailing. I also thank H. Lee Moffitt Cancer Center, Wellcome Centre For Human Genetics, University College, Wellcome Trust Foundation, and Clarendon for their support.

To my mother, Brenda Oldham, thank you for your love, unconditional guidance, and critical support that made this all a possibility. To my son, Talon, I will never forget the sacrifices you have made for me to pursue this doctorate. To my partner Rebecca, thank you for your unwavering strength, thankless support, and unending encouragement. Rebecca and Talon, I love you dearly.

COVID-19 Statement

It is difficult to put into words or quantify, at this time, the impact that COVID-19 has had on myself and all others involved in supporting, supervising, or working alongside me during the research conducted herein. My research was conducted across two research groups and institutions. Upon the commencement of the second half of my research, spent at Moffitt Cancer Center in Tampa, Florida, COVID-19 spread globally, reaching Europe and the United States in the winter of 2020. This pandemic stretched the capabilities of everyone, all institutions, and all governments. As it pertains to the work presented here, COVID-19 had the largest impact in three specific ways. During the beginning of the global pandemic until a year and a half later I unexpectedly served as instructor and administrator of my son's education at home. Second, the research that had been originally planned for travel back to Oxford from Tampa, Florida did not come to fruition. Lastly, due to the extent of isolation from colleagues several collaborations fell to the wayside, that may have made substantial impacts to the work here, to focus only on essential research activity. Every effort has been made to ensure the highest quality of work despite these significant impacts.

Abstract

Cancer initiation and progression is an evolutionary process where somatic alterations convey fitness advantages to cells that exist in a complex environment acting to shape cell populations. Homeostasis, being the normal state of a tissue prior to disease, prioritizes and protects long lived stem cells that are capable of incurring these somatic alterations and contributes to the early evolutionary dynamics from which cancer initiation occurs. Research over the past two decades has made substantial inroads into our understanding of somatic mutations through characterization of healthy tissue. In parallel, agent based mechanistic models have emerged as an important tool for understanding somatic mutation and stem cell dynamics in both cancer and homeostatic tissues. In this thesis we combine first principles mathematical modeling of homeostasis, complete with high-resolution genotypic information, with healthy patient sequence data in various tissue types to answer critical biological questions about somatic evolution, stem cell dynamics, and cancer initiation.

We first develop a three dimensional mechanistic model of the homeostatic epidermis and methodology necessary to incorporate base pair resolution mutation data. We use this model to examine neutrally expanding subclonal populations and compare simulations to healthy patient data. We then introduce functionally heterogenous oncogenic drivers in our *in silico* model to examine non-neutral dynamics in the epidermis. Next we develop a novel lineage tracing method using fluctuating methylation sites through analysis of patient data combined with a spatial model of a homeostatic intestinal crypt to determine stem cell dynamics in glandular tissue. Our spatial model reveals the necessary dynamics of fluctuating methylation sites to be used as fluctuating methylation clocks that allows us to determine stem cell behavior over short timescales *in vivo*.

In the last chapters of this thesis we focus on using our novel fluctuating methylation clock and somatic evolution as an important biomarker to understand clonal hematopoiesis. As the hematopoietic stem cells become homogenized through disease, a distinctive distribution emerges within patient data that can be exploited to diagnose clonal hematopoiesis without the need to understand somatic drivers. Combined, the work in this thesis reveals the power of somatic evolution on various timescales as a marker that can provide tissue level insights. We conclude by outlining a future direction where combining fluctuating methylation markers with DNA alteration could be a powerful tool in diagnostics and prognosis within the clinic.

Contents

List of Figures vii

List of Tables viii

Abbreviations and Acronyms x

1 Introduction 1

1.1 Role of mechanistic modeling in understanding tissue dynamics . . . 3

1.2 Somatic evolution is commonplace in normal tissue 4

1.3 Biological fidelity and lineage tracing 6

1.4 Modeling of genomic alterations 7

1.4.1 An important consideration for modeling genotypic information 8

1.5 Modeling homeostasis 10

1.5.1 The epidermis, intestinal mucosa, and hematopoietic stem cells 11

1.5.1.1 Epidermis 11

1.5.1.2 Intestinal crypt 11

1.5.1.3 Hematopoietic stem cells 13

1.6 Clinical relevance of understanding somatic evolution and cell fates . 14

1.7 What has changed? 14

1.8 Outline of project aims 15

2 Materials and Methods 17

2.1 Human samples 18

2.1.1 Epidermis Patient Data 18

2.1.2 Tissue Collections from human intestines and endometrium . . 18

2.1.2.1 Data availability for methylation data from intestines and endometrium 18

2.1.3 Publicly available methylation data 20

2.1.4 Study Participants of clonal hematopoiesis of indeterminate potential study . . . 20
2.2 DNA . . . 20
2.2.1 Methylation Arrays . . . 20
2.2.2 DNA extractions from peripheral blood and bone marrow . . . 21
2.2.3 Targeted DNA Sequencing of clonal hematopoiesis of indeterminate potential samples . . . 21
2.2.3.1 Variant calling in targeted sequencing of mononuclear cells . . . 22
2.2.3.2 Annotation of clonal hematopoiesis driver mutations 23
2.2.3.3 Comparison of bone marrow and peripheral blood allele frequencies . . . 23
2.3 Methylation . . . 25
2.3.1 Methylation arrays and analysis . . . 25
2.3.2 Copy number analysis on publicly available methylation data . 25
2.4 Equations and Statistics . . . 26
2.4.1 First incomplete moment . . . 26
2.4.2 Epidermis model parameterization . . . 27
2.4.3 Departure from homeostasis in the epidermis . . . 28
2.5 Introducing model error . . . 28
2.5.1 Calculating variant allele frequencies from simulations . . . 28
2.5.2 Fluctuating CpG distribution error model . . . 29
2.6 RNA data from the intestinal crypts . . . 30

3 Homeostasis and Carcinogenesis in a First Principles Model of the Epidermis 31
3.1 Background . . . 31
3.1.1 Chapter aims . . . 32
3.2 First principles model of the Epidermis . . . 33
3.2.1 Growth Factor . . . 33
3.2.2 Keratinocyte dynamics . . . 34
3.3 Epidermis Model Results . . . 35
3.3.1 Gattaca: Mutation induction tracking . . . 38
3.3.1.1 Gattaca: Implementation and Case Studies . . . 39
3.3.1.2 Gattaca in the epidermis . . . 41
3.3.2 Visualizing Evolutionary Dynamics in the Epidermis . . . 42

3.4 Neutral evolution . . . 43
3.5 Modeling Fitness Advantage in NOTCH1 Mutant Keratinocytes. . . . 46
3.6 Modeling TP53 mediated advantage in the UV exposed epidermis . . 47
3.7 Discussion . . . 52
3.7.1 Conclusions . . . 52

4 Exploiting biological fidelity to understand homeostasis in the intestinal crypt 54
4.1 Background . . . 54
4.1.1 Chapter aims . . . 55
4.2 The exclamatory bowel model . . . 55
4.2.1 Crypt dynamics of the exclamatory bowel . . . 57
4.2.1.1 Evaluating an alternative spatial crypt model . . . 59
4.3 Fluctuating CpG theory explained by the exclamatory bowel . . . 60
4.4 Patient data and fluctuating CpG sites (fCpG) . . . 61
4.4.1 Early attempts to define fCpG filtering criteria . . . 63
4.4.2 fCpG loci are enriched in minimally expressed genes . . . 63
4.4.3 Methylation status of fCpGs is conserved along the crypt . . . 65
4.5 Mathematical model of the methylation distribution of fCpGs. . . . 65
4.6 Stem cell dynamics are inferred with high accuracy *in silico* . . . 66
4.7 Measurement of stem cell dynamics in human intestines and tissues . 69
4.7.1 Stem cell dynamics in human endometrial glands . . . 69
4.7.2 Departure from homeostasis: familial adenomatous polyposis . 73
4.8 Discussion . . . 73

5 Exploiting fluctuating methylation clocks identified in peripheral blood 75
5.1 Background . . . 75
5.1.1 Chapter aims . . . 76
5.2 Hematopoiesis model . . . 77
5.3 Fluctuating methylation clock dynamics in hematopoiesis simulations 78
5.4 Fluctuating CpGs in publicly available blood samples . . . 79
5.5 Clonal hematopoiesis . . . 81
5.5.1 Clonal hematopoiesis patient cohort . . . 82
5.5.2 Clonal hematopoiesis fluctuating methylation clocks . . . 82
5.6 Mutation agnostic tool for diagnosing clonal hematopoiesis . . . 84
5.6.1 Classifier performance on clonal hematopoiesis patient cohort 86

5.6.2 *In-silico* evaluation of classification model . . . 87
5.6.3 Retrospective examination of normal whole blood cohorts . . . 88
5.6.3.1 Validation of newly diagnosed clonal hematopoiesis patients . . . 88
5.7 Clonally heterogeneous landscapes in the hematopoietic stem cell pool 92
5.7.1 Confounding effects of multiple drivers *in silico* . . . 94
5.7.2 Multiple subclonal drivers in the data . . . 95
5.7.3 Paired bone marrow from patients with clonal hematopoiesis . 97
5.7.4 Single cell sequencing of clonal hematopoiesis patients confirms subclonal structures . . . 98
5.7.4.1 Modeling nested subclonal expansions . . . 99
5.8 Discussion . . . 100

6 Conclusions **103**
6.1 Summary of results . . . 103
6.2 Discussion . . . 104
6.2.1 Somatic evolution in homeostatic tissue . . . 104
6.2.1.1 Somatic mutations and mechanistic models . . . 105
6.2.2 Clinical insights and considerations . . . 106
6.3 Future directions . . . 107
6.3.1 Future direction for fluctuating methylation clocks . . . 108
6.3.1.1 Inference of clonal hematopoiesis expansion rates and induction times . . . 108
6.3.1.2 Progression of clonal hematopoiesis to a malignancy 111
6.3.1.3 Early detection of clonal hematopoiesis and oncogenic progression using fluctuating methylation clocks . . . 112
6.4 Closing remarks . . . 114

A Appendix: Chapter 2 **115**
A.1 Gattaca Case Studies . . . 115
A.1.1 Gattaca Case Study 1: Dimensionality . . . 115
A.1.2 Gattaca Case Study 2: Wounding . . . 117
A.2 EvoFreq R Package . . . 119
A.2.1 EvoFreq Results and Example . . . 120
A.3 Additional Figures . . . 122

B Appendix: Chapter 3 **127**
B.1 Additional Information . . . 127
B.1.1 Derivation of fluctuating CpG loci . . . 127
B.1.2 Evidence for fluctuating human CpG sites . . . 129
B.1.3 Mathematical model of fluctuating CpG distributions in the stem cell niche . . . 131
B.1.4 Error model . . . 133
B.1.5 Likelihood and prior functions . . . 134
B.1.6 Bayesian inference . . . 135

C Appendix: Chapter 4 **143**
C.1 Additional Information . . . 143
C.1.1 Further restrictions to derivation of fluctuating CpG loci . . . 143
C.1.2 Hematopoietic stem cell model additional considerations . . . 144
C.1.3 Hematopoietic stem cell model beta value corrections . . . 144

Bibliography **148**

List of Figures

1.1 Genomic alterations, evolutionary processes, and timescales 5
1.2 The epidermis, intestinal crypt, and hematopoietic stem cells 12

2.1 Overview of primary datasets 17

3.1 The homeostatic epidermis model and Gattaca method 32
3.2 Three dimensional hybrid cellular automaton keratinocyte lifecycle . . 36
3.2 Three dimensional hybrid cellular automaton keratinocyte lifecycle . . 37
3.3 Gattaca method 38
3.4 Neutral evolution of the homeostatic epidermis 44
3.5 Homeostasis, *a priori*, constrains clones in a functionally heterogeneous tissue, dictating an age-dependent clonal expansion. 45
3.6 Combined effects reveal homeostatic recovery and recapitulate patient clone distributions 48
3.7 Logistic growth model is able to capture spatial dynamics to aid in evaluating TP53 sun days and dynamics 49
3.8 TP53 parameter evaluation using a space limited ordinary differential equation 51
3.9 Combined effects adhere to age-dependent expansions 53

4.1 Fluctuating methylation status as a lineage tracing marker 56
4.2 Exclamatory Bowel Model 58
4.3 Fluctuating CpG distributions in the exclamatory bowel/ring 62
4.4 Identification of selectively neutral fluctuating CpG loci 64
4.5 W-shaped methylation distributions are indicative of clonal dynamcis 67
4.6 Tissue-specific differences in stem cell dynamics 70
4.7 The fluctuating methylation clock method is generalizable to other glandular tissue 72

5.1 Fluctuating methylation clock idealization in whole blood 76

5.2 Fluctuating CpG dynamics observed in chronic and acute leukemia . 80
5.3 Characterization of CH patient samples 83
5.4 FMC-CH classifier 85
5.5 Retrospective analysis of 1,388 normal whole blood samples 89
5.6 Concurrent driver mutations in the model 93
5.7 Concurrent driver mutations in the peripheral blood and bone marrow 96
5.8 Single cell sequencing confirms subclonal composition 101

6.1 Clonal hematopoiesis expansion rates and induction times 109
6.2 Current progression paradigm for clonal hematopoiesis 111
6.3 Early detection of clonal hematopoiesis and oncogenic progression using fluctuating methylation clocks 112

A.1 Gattaca Case Study 1: Dimensionality 116
A.2 Gattaca Case Study 2: Wounding 118
A.3 Evofreq functionality 120
A.4 Evofreq with subclonal reconstruction software 121
A.5 Years for a single progenitor to sweep through basal layer 122
A.6 PD20399 pairwise comparisons with simulations. 123
A.7 PD21910 pairwise comparisons with simulations. 124
A.8 PD18003 pairwise comparisons with simulations. 125
A.9 PD13634 pairwise comparisons with simulations. 126

B.1 Additional analysis of fluctuating CpG identification 137
B.2 Additional evidence supporting the existence of fluctuating CpG loci 139
B.3 Variances of fluctuating CpG distributions of the Exclamatory Bowel 141
B.4 Fluctuating CpG distribution at various (de)methylation rates for the Exclamatory Bowel 142

C.1 Model results from multiple founders 146
C.2 Model steady state of fluctuating CpG beta variance 147

List of Tables

1.1 HAL provides the necessary components to incorporate high resolution genomic information. 9

2.1 Fluctuating Methylation Clock Patient Samples 19
2.2 Paired peripheral blood and bone marrow allele frequencies 24

3.1 Epidermis model simulations 33

A.1 Comparison of the features within the currently available packages for plotting evolutionary dynamics. EvoFreq is capable of visualizing the evolutionary dynamics from ClonEvol, PhyloWGS, and CALDER. . . 119

B.1 Exclamatory bowel parameters 128

C.1 Whole Blood Simulation Parameters. 144

Abbreviations and Acronyms[1]

0D zero-dimensional (non-spatial) . . . . . 39, 113, 115 f.

2D two-dimensional . . . . . 25, 39, 41, 106, 113, 115 f.

3D three-dimensional . . . . . 14, 25, 30, 32, 39, 113, 115 f.

3D-HCA three-dimensional hybrid cellular automaton 26, 31, 45, 47, 49

ABM agent based model . . . . . 4, 7 f., 36 f., 39

AMG aorta/gonad/mesonephros 12

AML acute myeloid leukemia 17, 77

BM bone marrow . . . . . 19, 22, 73, 95 ff.

BMP bone morphogenetic protein . . . . . 11 f.

CH clonal hematopoiesis13, 22, 24, 73–78, 80, 82, 84–90, 92 f., 95–98, 100, 102, 105, 108 f., 111, 141

CHIP clonal hematopoiesis of indeterminate potential 13, 15, 102, 106, 108 f., 111

CML chronic myeloid leukemia 17, 80

CNA copy number alteration . . . . . 22, 24, 38, 78, 89 f., 103, 109, 111, 118

EB exclamatory bowel 55, 58, 64, 66, 131, 142

ET essential thrombocythaemia 17, 141

FAP familial adenomatous polyposis . . . . . 69, 72, 101 f., 104

fCpG fluctuating CpG 15, 28, 53, 55, 58 f., 61, 63 f., 66 f., 69, 71–74, 77, 80, 82, 84, 86, 88 ff., 92 f., 95–98, 100 f., 106, 108, 111, 125, 127 f., 131, 141 f.

[1] f. denotes page listed and the following page. ff. denotes page listed and the following pages.

FMC fluctuating methylation clock 15, 53, 58, 64, 67, 69, 71–74, 78, 84, 90, 93, 98, 100 ff., 104 ff., 109, 111, 125, 128, 133

GLM generalized linear model 67

HAL hybrid automata Library 31, 35, 37, 50, 53

HSC hematopoietic stem cell . 10, 12 f., 15, 22, 73 ff., 77, 84 ff., 88 ff., 93, 95 f., 98, 100, 102, 106, 109, 111, 142

LOH loss of heterozygosity 6, 118

MDS myelodysplastic syndromes 17, 80, 109

MNC mononuclear cell 19

NGS next generation sequencing . . . . . 3, 8 f., 27, 35

ODE ordinary differential equation 45, 47

OR odds ratio 61, 87

PB peripheral blood . . . . . 19, 22, 73, 84, 95 ff.

PMF primary myelofibrosis 17, 141

PV polycythaemia vera 17, 141

SNP single nucleotide polymorphism . . . . . 5 f., 9, 109, 111, 118

SNV single nucleotide variant . . . . . 5–9, 14, 35, 78, 85, 88 f., 101, 109, 118

TCGA The Cancer Genome Atlas 28, 61, 115

tMN therapy-related myeloid neoplasm 111

UV ultraviolet 45, 47, 49

VAF variant allele frequency . . . . . 21 f., 27, 38, 40 f., 80, 84 f., 92 f., 96 ff., 115

Chapter 1

Introduction

There are two common ways that a biomedical cancer researcher can contribute their efforts to the cure and treatment of cancer. The first is to approach cancer as a focused, singular entity; something that is distinct from human tissue and homeostasis[1]. In contrast, the second views cancer as an emergent part of our tissue. The former paradigm is the most commonplace. It is the idea that treatment for cancer begins only when someone walks into the clinic, receives a diagnosis, and treatment options are considered. These often include debulking[2]; alongside chemotherapeutics and/or targeted therapies in some combination that is based on current standards of care fit to all patients [2–5]. This is a worthwhile approach that has provided tremendous extended survival time to patients that would have otherwise succumbed to their cancers much sooner. Since the advent of our first pharmacological cancer treatment, chemotherapy, aminopterin—thanks to the horrifically serendipitous observation that nitrogen mustard (colloquially, mustard gas) significantly suppresses hematopoiesis [6]—we have made enormous gains in understanding cancer. At the same time, we have become dishearteningly aware that treatment doesn't begin with diagnosis.

At the time chemotherapy was discovered, and since the few decades that have passed between then and now, there has been an exponential gain in our knowledge of what cancer is and how to treat it. As part of this knowledge, we learned there doesn't appear to be a singular treatment that is likely to treat every cancer due to the vertigo-inducing number of cancer subtypes within each tissue that can exist [7]. We learned that treatments fail rather quickly—sometimes with staggering speeds—

[1]Homeostasis here meaning the balance of cellular and extracellular factors that contribute to a healthy cell and tissue; as opposed to formal definition where homeostasis is the tendency towards a stable equilibrium [1], which can paradoxically encompass some states of diseases.

[2]Debulking is the surgical reduction of as much tumor as possible, either to aid in relieving symptoms, as a single treatment option, or as a (neo)adjuvant therapy to make a combination treatment more successful.

due to our inability to remove all malignant cells or via the emergence and growth of treatment resistant cancer cells [8]. Large statistical studies of cancer incidence continue to yield insights into correlations between certain environmental exposures (chemicals, foods, materials, etc.) and increases in the incidence of various cancers [9]. One example is the recent interest in understanding how inflammation can prime the body for cancer development [10]. We also discovered just how similar cancer cells are to normal cells, allowing cancer cells to evade–normally highly effective—immunosurveillance. Most recently, we have a new understanding that, not only is there a vast landscape of different types of malignant neoplasia, but there exists a highly heterogeneous landscape of molecularly distinct cells for evolutionary processes to act on, even within a single tumor [11].

This is where the second type of contribution to the cure and treatment of cancer begins, with the understanding that cancer is a part of ourselves. A part that has broken free of the shackles and cellular checks and balances of homeostatic tissue; those cells within us, relatively (to their ancestral lineages) free to grow, unfettered beyond their cooperative role in support of a tissue and organism. We have learned that treatment of cancer truly begins with the ambitious goals of prevention and early detection. Both require knowledge not only of cancer as a separate, distinct entity, but also of how cancer cells retain vestigial components of their normal developmental selves and the preceding, homeostatic tissue from which cancer emerges. Tumorigenesis is the all-encompassing term describing the gain of malignant properties, resulting in the gradual loss of lineage identity (as would have been encoded for during embryonic development). Tumorigenesis is the result of evolutionary processes driven by selection for one, or several, hallmarks of cancer (providing several fitness advantages over homeostatic cells) [12] that facilitate cancer development. Tumorigenesis itself, is the consequence of somatic evolution acting on the functional differences present in normal populations of stem cells, some of which possess altered regulatory networks that convey net proliferation gains and disruptions in homeostasis due to genomic alterations. However, it is not a single genomic alteration that directly conveys the unquestionable ability to transform a cell into a malignancy, despite mutation being a requirement, it is a complex dance whereby the right timing and environment selects for mutated cells that go on to become a malignancy.

It is this complex dance between a changing environment and the ever-changing cells within it that this body of work attempts to exploit. To gain a deeper understanding of normal tissue somatic evolution and the corruption of this process in neoplasia initiation, through the integration of mechanistic modeling with genetic

markers of cell fate. The act of somatic evolution within various tissues allows us to gain insight into collective tissue homeostasis, as well as individual cell fates, and to exploit the early departure from homeostasis for diagnostic purposes, as well as diagnose and delineate patients at risk.

1.1 Role of mechanistic modeling in understanding tissue dynamics

The term 'model' is used often and can cause confusion for many. Its definition is quite straightforward "a system or thing used as an example to follow or imitate." This has broad applicability in everything from judicial law (as precedence) to the equations describing diffusion (partial differential equations). Mechanistic modeling is an approach that seeks to establish a mechanistic explanation in an observed phenomena. Baker *et al.* summarized mechanistic modeling as 'focused on the causality of input-output relationships' [13]. To understand what mechanistic modeling can provide it is necessary to frame it in the context of the broader research ecosystem. We begin with an observed phenomena where we collect data from patients on various scales (e.g., transcriptome, genome, and proteome from imaging and *next generation sequencing* (NGS)). Our goal is to understand an observed phenomenon to be able to provide insights into prevention, diagnosis, prognosis, and treatment. There are many ways to accomplish this, but all approaches can be compressed to the five 'W's: who, what, when, where, and why (elaborated further in my perspective piece [14]). Statistical and machine learning models answer what is likely to happen and when might it occur with little regard for underlying mechanism. This is classical correlation versus causation. Experimental models are more like mechanistic models in that they strive to understand who and why. Who are the important players and why are they important? That's not to say that mechanistic models are not predictive or able to answer similar questions to statistical or machine learning models, they simply do so by considering additional information to uncover causation. This highlights two very important steps for any research question. What are the models that are required and what validation is necessary. Integrating this understanding of the research ecosystem with the perspective piece by Baker *et al.* [13], we deduce that mechanistic modeling has only just begun to make contributions to biomedical research and its full potential integration with high throughput data, predictive powers of machine learning, and the confidence provided from statistical analysis is only just being realized.

Mechanistic models have long been used to understand aging and cancer incidence in tissue, it is through this observed relationship that the theory of multistage carcinogenesis was conceived [15–17]. However, until recently, mechanistic models have been focused on cancer initiation and progression without considering the normal homeostatic dynamics that precede cancer. Further still, these models focus largely on phenotypic information rather than incorporating any genetic information. In the most recent review that covers mechanistic modelling's role in understanding mutations in normal tissue, Altrock *et al.* [17] highlight several models that have been used to understand mutations in normal tissue. However, rarely do these approaches examine mutation dynamics in homeostatic tissue or examine how spatially defined differences can emerge based on homeostatic tissue constraints. The two most popular of these types of models include branching process models and Moran process models [17]. These models prevent us from being able to incorporate any high-resolution genetic information due to their precisely solvable approach. When considering time, space, individual cell fates, and high-resolution genetic information a more apt approach, given today's computational power, is to use an *agent based model* (ABM). These simulation-based models utilize a discretized agent (e.g., cell) that operates either stochastically or deterministically; meaning, behaviors are governed by a probability for a stochastic process rather than the same outcome always occurring given the same parameters and initial conditions. Importantly, these models have already demonstrated their power to understand timing of somatic mutations in normal tissue and cancer [18, 19].

1.2 Somatic evolution is commonplace in normal tissue

Some of the earliest studies that identified mutations in normal somatic cells (1996-2001) were confounded by a desire to find oncogenic drivers in normal tissues. Their presence was explained by stating it was due to an environmental stimulus and/or likely the earliest stages of transformation, their presence, able to serve as an early detection marker [20–22]. We now know that oncogenic driver mutations within histologically normal tissue is commonplace [23–37]. These early driver mutations are the earliest constituents of multi-stage carcinogenesis, a theory first proposed by Armitage and Doll in 1954 [15]. Somatic mutations are commonly subject to neutral drift, where mutations have little to no effect on a cells function, but due to random birth and death these mutations frequencies change over time and eventually

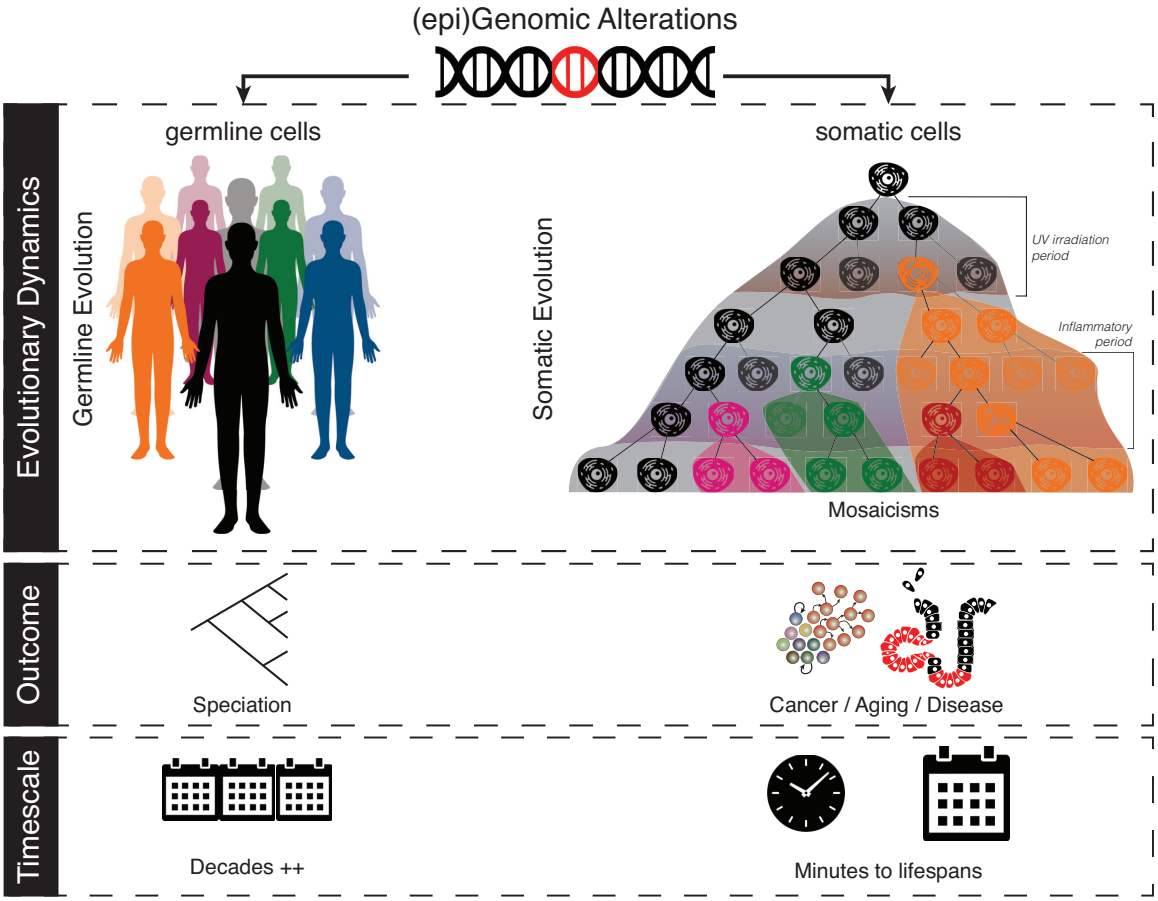


Figure 1.1: Genomic alterations, evolutionary processes, and timescales. Evolution occurs over different timescales and leaves its record of change hidden within the genome from individual, germline evolution, and cell to cell, somatic evolution. Various mutational processes with associated biological fidelities such as (de)methylation markers and DNA mutations can be leveraged to understand homeostasis, aging, and disease.

fixation can occur within a population [38–40]. In contrast, functionally advantaged populations of cells are selected for, determined by their environment, such that their population grows larger than other, less fit, populations of cells.

Somatic alterations in the gametes (*single nucleotide polymorphism* (SNP)[3] and large structural variations as the result of recombination) can drive evolutionary change, offering fitness advantages that can ensure adaptations and continuity of species. In contrast, these same processes leading to genomic alterations in somatic cells are unable to offer an improved functionality for evolution to select upon and to improve the species, unless the selfish cell is the primary focus, in which case evolution

[3]SNPs are the common term used to describe point mutations in the germline. The appropriate terminology for point mutations in somatic tissue is *single nucleotide variants* (SNVs)

is working as expected by introducing functional heterogeneity, increasing the odds that a single cell lineage will survive, driving aging and in some cases, cancer. The collection of somatic mutations within seemingly normal tissue in different population sizes is often termed 'mosaicisms' [41]. Due to the continued increase in detection of mutations and the ability to assay normal tissue, as well as disease, it has become more apparent that our somatic tissue's genomes are mosaic as the rule, rather than the exception [42].

1.3 Biological fidelity and lineage tracing

In aggregate, any measure of somatic evolution assesses the outcomes of several challenge events—iterative genomic changes followed by selection or drift—over some timescale. These dynamics leave a record within tissues over time. Changes to the genome occur and are passed from one cell to the next with an associated fidelity at various scales. With regards to hereditary material, in somatic cells these scales are changes to the base pair composition of DNA (SNVs, copy neutral *loss of heterozygosity* (LOH), and aneuploidy) and epigenetic markers, primarily methylation of cystine at palindromic CpG sites, although histone methylation and acetylation can be considered. This fidelity can be exploited to understand evolutionary dynamics in tissues and provide insights into cell fates but the information we can gain is limited by how rapidly changes occur in the tissue and in DNA alterations.

Within tissues, cell fate events have been understood through cell fate markers combined with microscopy or histology. This has been extraordinarily useful to understand tissue development, homeostasis, and disease through experimental manipulation of cells in animal models [43]. In the simplest examples DNA synthesis proteins can be used as markers to assess proliferation status such as Ki-67 (MKI67[4]), PCNA, and MCM-2. These approaches provide a fixed snapshot to characterize the suspected cell fates (describing the proportion of expected progeny at that time) of a tissue. No long-term cell fate determination can be gained from this approach. For *in vitro* or *in vivo* cell fate tracking, one can use bromodeoxyuridine (BrdU) or 5-Ethynyl-2′-deoxyuridine (EdU), both of which are thymidine analogues that are incorporated into cells and then detected through BrdU and EdU antibodies over time. Cell fates and tissue dynamics in normal human tissue are challenging to untangle as it is not possible to perform exogenous cell labeling to easily trace cell lineages,

[4]Throughout this dissertation the official gene names and symbols for *Homo sapiens* are used without expansion of their full names. These official names and symbols are defined by the *HUGO Gene Nomenclature Committee* (HGNC)[44].

as is possible with animal models that these markers are used in [45–54]. Instead, we rely on the DNA that is inherited by daughter cells from its ancestors with a particular fidelity or mutation rate, estimated to be between 10^{-7} and 10^{-8} per base pair per generation [55]. We can contrast these rates with epigenetic fidelity, where methylation sites change with those that occur at a much faster scale, three to four orders of magnitude faster, but are loci dependent.

To determine cell fates and uncover evolutionary dynamics, the important component of genetic change is that a collection of alterations is shared across a collection of cells within a lineage, representing a clonal population[5]. Various types of somatic DNA alterations that alter base pair composition have been exploited for cell lineage reconstruction in human tissues, including *mitochondrial DNA* (mtDNA) mutations [56–69], DNA methylation at selectively neutral loci [70–77], allelic loss at heterozygous loci [78, 79] and SNVs detected by genome sequencing [25, 31, 35, 36, 80–88]. In normal tissue however, these mutations fail to expand to a frequency large enough (where a clone can be detected) until several years have passed due to limited growth and expansion. Conversely, methylation status changes rapidly and only short-term dynamics can be examined before the signal changes. This introduces a problem where only short term or long-term cell fates can be uncovered in normal human tissue.

1.4 Modeling of genomic alterations

Studies of somatic evolution in normal tissues have largely focused on mutation characterization, with limited analysis of the dynamics driving observed evolutionary trajectories, with only a few notable exceptions using murine models [30]. Fewer still have begun incorporating ABMs as a tool to explore somatic evolution in spatially constrained tissue [30, 89, 90], something that has been of great use for understanding cancer progression [19, 91–93]. Historically, genomes within agent-based models have been represented as simple counters or as binary arrays. These binary arrays serve only as a proxy to then relate back to the genome of patients or animal models through heterogeneity measures and simple counts. Most notably, passenger/driver model genomes are simple counters that increase, and decrease based on a set muta-

[5]The definition of a clone or subclone is different depending on the context or what field is using these terms. In many fields, a clone is described as a population of cells that carry the same phenotype, here we are concerned with genetic clones and subclones. A clone or subclone is a population of cells that shares a particular mutation or daughter population of this clone with a shared set of mutations (subclone). In normal tissue, the widely adopted term for multiple existing subclones from the 'background' normal is mosaicism.

tion rate and probability of driver acquisition. These are then tied to some function defined in the model, sometimes in very arbitrary ways. These studies have lacked the ability to compare, at base pair resolution the mutation spectra or utilize common tools designed for genotypical data (such as dN/dS and subclonal reconstruction tools designed for noisy biological data).

To date there is not a mechanistic, ABM framework that allows for somatic mutation induction and tracking at a base pair resolution for mechanistic ABMs. Previously, our understanding of mutational processes and sequencing technologies has not had the resolution or technical understanding necessary to create such a tool, not to mention the computational power needed for such a task. Due to substantial advances this has quickly become possible. First was our improved understanding of mutational processes [94]. Then we began to grasp the importance of heterogeneity in tumors through multi-region sequencing studies [95][6]. Now, we have added to the wealth of research to understand mutation in normal, homeostatic tissues (see section 1.2 Somatic evolution is commonplace in homeostatic tissue) [23–37]. Currently, there is an ongoing debate over neutral and non-neutral dynamics in both tumors and normal tissue [82, 99–107]. Lastly, only recently have mechanistic models begun to really intersect with bioinformatics [19, 30, 32, 92, 108]. Together, these have facilitated the need for new methodologies where realistic genetic information can be incorporated into mechanistic models for future researchers. There is great value in developing a methodology that provides a means of inducing and tracking base pair resolution SNVs within an ABM framework with temporal, spatial, and genomic positional information. Prior to embarking on the research contained herein I worked to help develop a modeling framework that had the necessary components to integrate this genomic information into an ABM[109], a summary of the features of this framework (HAL), can be found in Table 1.1. This will form part of our work in Chapter 3.

1.4.1 An important consideration for modeling genotypic information

Cheap computational processing power, storage space, and increasingly high-throughput sequencing technologies created a boom of genetic data that has been placed in pub-

[6]Numerous studies now exist that employ multi-region sequencing, some of which I have previously contributed heavily to (see NeoPredPipe [96], the first and only neoantigen prediction pipeline designed specifically for multi-region NGS samples, and subsequent studies I have contributed to as part of NeoPredPipe [97, 98]), but the 2012 Gerlinger *et al.* study is widely considered the seminal work in this area.

Feature	Feature used in	HAL	Chaste	Physicell	Mason
On-Lattice ABMs	Chapter 3	X	X		X
Off-Lattice Point ABMs	Chapter 4	X	X		X
Off-Lattice Spherical ABMs	Chapter 4	X	X	X	
Multinomial Population ABMs	Chapters 3-4	X			
Diffusion PDEs	Chapter 3	X	X		
Real-Time Visualization	Chapters 3-4	X		X	X
Mac Compatible	Chapters 3-4	X	X	X	X
Linux Compatible	Chapters 3-4	X	X	X	X
User-Facing Programming Language	*N/A*	Java	C/C++	C/C++	Java

Table 1.1: HAL provides the necessary components to incorporate high resolution genomic information. Comparison of select modeling frameworks and the chapter that each listed feature was used in.

licly available repositories for future inquiry [110–115]. One thing that is common across all these repositories and various sequencing approaches and technologies is the error associated with each approach. Vast resources in funding, time, and expertise have been and are still being dedicated to correcting error associated with array based and NGS methodologies [116–118].

Microarrays are the primary approach used as a diagnostic in the clinic since they are reliable and provide consistent results for clinical decision making. In essence, a DNA molecule is labeled with a fluorescent dye, where the difference in state dictates which label it receives (e.g., SNP A versus B or methylated versus unmethylated CpG), then this product is hybridized in parallel to denatured DNA probes spotted on a glass slide [118, 119]. Microarrays can be used to assess a multitude of information such as gene expression, SNV, SNP, methylation, and aneuploidy. The difference in all of these is the preparation of the source DNA/RNA. The similarity across all of these is that systematic error is introduced through spot intensity-dependent, feature-specific, and spot position-dependent contributions [118], necessitating a normalization procedure to deal with variances of saturation levels [120].

NGS sequencing introduces errors in a slightly different way and is not yet adopted in the clinic as widely as microarray technologies due to its cost, storage requirements, and extraneous information (i.e., not useful for clinical decision making) it can provide. NGS sequencing provides short and long reads (fragments of DNA) through a variety of sequencing approaches, after library preparation, where whole genome or exome amplification prepared via enzymatic or mechanical fragmentation, and adapter ligation can be mapped to a reference genome and used to call somatic variants. These sequencing approaches include measuring voltage differences for each base as DNA crosses a membrane or measuring fluorescence from terminating dinucleotides

as they bind to amplified DNA in real time [121, 122]. There are inherent sources of error that these methodologies present tied to the sample preparation, preferential amplification of source templates, sequencing error rates (approximately 0.1-1% per base), and in the bioinformatics alignment methods required for interpretation [117].

In modeling genetic information using mechanistic models, this error presents a special challenge not often seen in the broader mathematical oncology community. Simulations provide us with data inherently free of technical error. This is an important feature of mathematical modeling and one of its strengths; however, this can cause problems in attempts to quantitatively fit a model to observed clinical and biological data. This necessitates the presence of an error model that should be applied to synthetic genetic data, from mechanistic models, prior to comparisons. In each chapter to follow, we utilize and mechanistically model sequence and (de)methylation information to directly compare with patient derived samples (Figure 2.1). Each chapter utilizes at least one approach to introduce a small amount of error that is an important step in comparing NGS and microarray data with the outputs of mechanistic models.

1.5 Modeling homeostasis

Homeostasis has a few fundamental truths. These truths can be the guiding principles for the development of any model of homeostatic tissue dynamics.

- A steady number of cells must exist within a given tissue, yielding balanced birth and death due to progenitor loss and replacement.
- A population of cells must divide regularly and maintain the ability for future divisions.
- Genetic mutations must accrue and have the ability to be passed on to daughter cells during somatic divisions providing inherited states.

Of course, these three fundamental truths of homeostasis, neglect a vast number of important considerations for the continued integrity and maintenance of this homeostasis. Some important considerations here are the morphogens that arise from various other cell types that govern cell behavior or the immune cells that can help perform immune surveillance to maintain homeostasis. However, depending on the question, these can be incorporated into the assumptions leading to these three behaviors. Importantly, these fundamental truths allow a model to be biologically realistic,

with the smallest number of assumptions, since the only important factor is the emergence of homeostatic behavior. These fundamental truths can be applied to several different tissues, and we apply them here in two very different epithelium tissues, the epidermis, and the intestinal crypt and one connective tissue, the hematopoietic system. A brief outline of these hierarchical tissues provides us with the biological background necessary for the work herein.

1.5.1 The epidermis, intestinal mucosa, and hematopoietic stem cells

Modeling homeostasis and embryogenesis are very different undertakings but some understanding of the origins of these tissues is helpful. The epidermis, intestinal crypt, and blood are distinct in their origins during embryogenesis and distinct in their organization, maintenance, and healing, but share some important features (Figure 1.2). The two epithelial tissues, the epidermis and intestinal crypt, form from the ectoderm and endoderm, respectively [123], while *hematopoietic stem cells* (HSCs) have a more complicated etiology.

1.5.1.1 Epidermis

The epidermis forms from a single sheet of epithelial cells called the basal layer after cells are committed to the surface ectoderm based on WNT and *bone morphogenetic protein* (BMP) signaling activity. This sheet of epithelial cells, the basal layer with progenitor keratinocytes, then follows a stepwise process with rounds of keratin expression at each stage before forming a stratified squamous epithelium. This is the tissue type of the epidermis, esophagus, and vagina, but differing in their thickness, microbiota, and support cells. In homeostasis, the epidermis is comprised of keratinocytes, progenitor cells within the basal layer maintaining tissue structure with balanced births and deaths [124] (Figure 1.2, Epidermis). Contributions to this understanding of homeostasis through somatic evolution can be found in Chapter 3.

1.5.1.2 Intestinal crypt

In contrast to the stratified squamous epithelium, is the simple epithelium found in the intestines, simple in that it's a single layer of epithelial cells. The gastrointestinal tract forms from the endoderm where a sheet of uniform progenitor epithelial cells are bent into crypt structures where the concentration of morphogens (interplay of WNT, *sonic hedgehog* (SHH), and BMP) differentiates the uniform intestinal

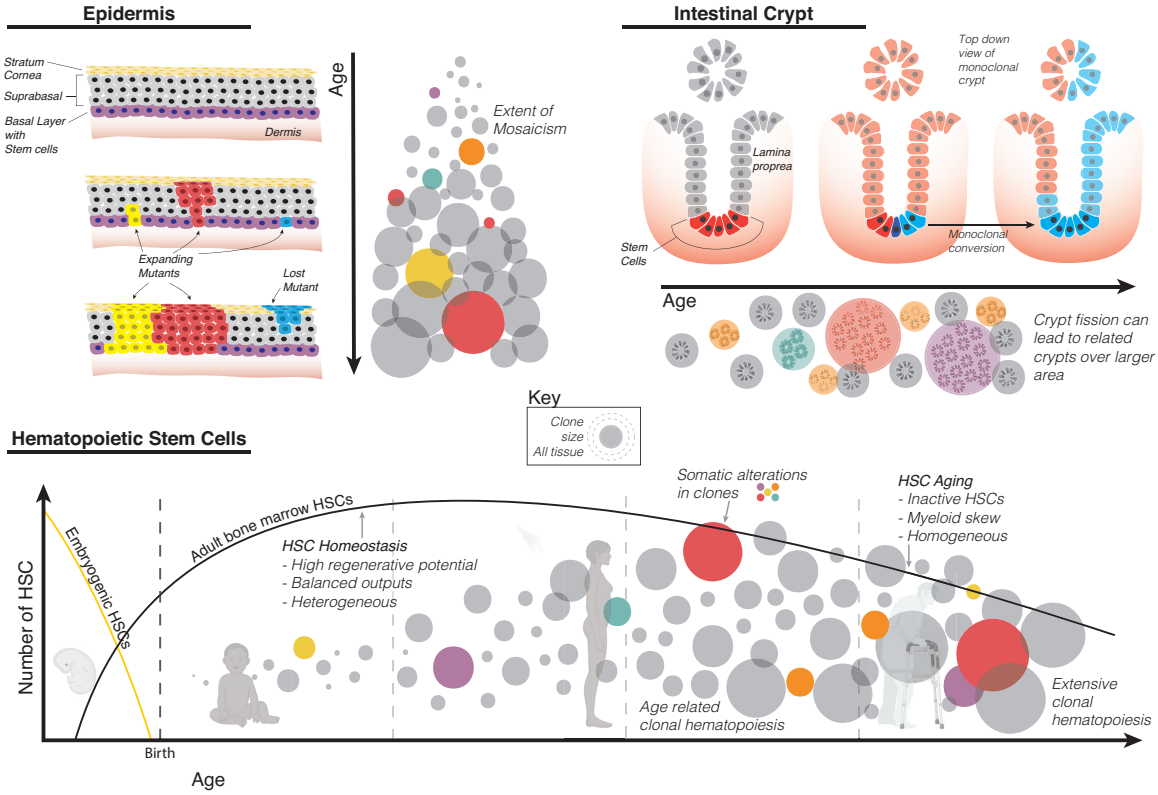


Figure 1.2: The epidermis, intestinal crypt, and hematopoietic stem cells. Somatic mutations accumulate with age, resulting in mosaicisms throughout various tissues in the human body. Within the epidermis this results in a patchwork of clonal populations, some with oncogenic driver mutations. The size of these clones increases with age. In the intestinal crypt, monoclonal conversions occur where a single mutation can sweep through the crypt giving rise to independent lineages. Through fission, small clusters of related crypts exist in the intestines with age. Different from epithelia, the hematopoietic system is loosely organized in the bone marrow, highly polyclonal, and stem cells are migratory. Hematopoietic stem cell numbers differ by age as well as their clonality.

stem cells to aggregate a handful of progenitors at the crypt base [125]. Once established, the intestinal mucosa is a complex, interdependent ecosystem, whereby the epithelial layer is supported by the underlying microenvironmental constituents. Inter-compartmental crosstalk within this system regulates epithelium cell fate determination and overall crypt structures [126–129]. Within the crypt the stem cell niche is maintained at the base of the crypt where the highest concentration of WNT is present [127]. As cells divide away from this WNT source, they differentiate and move up the side of the crypt, the crypt wall, and undergo anoikis where they break away from the epithelial sheet and shed into the lumen or are incorporated into the villi in the small intestines. We exploit this stereotypical structure and turnover to develop a novel cell fate marker in Chapter 4 (Figure 1.2, Intestinal Crypt).

1.5.1.3 Hematopoietic stem cells

HSCs form the postnatal hematopoietic system and can give rise to all blood cell lineages. They reside in a perivascular niche within trabecular bone concentrated in the pelvis, sternum, portions of the long bones, and ribs, after they emerge from the *aorta/gonad/mesonephros* (AMG) region inside the embryo[7] [130, 132]. HSCs are responsible for giving rise to an array of cell types derived from common myeloid or lymphoid progenitors, the two cell types that an HSC can commit itself to. The mature cells of myeloid progenitors dominate the blood, while mature lymphoid cells predominantly reside throughout tissues in the body [133]. The HSC pool of cells is highly heterogeneous and while hematopoiesis appears regional (i.e., which cells come from which HSC compartment), HSC themselves are highly polyclonal and migratory [134, 135].

In epithelial tissues, homeostasis exists throughout ones life where the long-term homeostatic constitution of this tissue is maintained, with only slight changes, such as diminished height of the epidermis with age. The hematopoietic system differs in that its constitution changes over time. At the time of birth, you have relatively few HSCs, these expand and acquire distinct mutations over time making them highly heterogeneous and polyclonal by the age of twenty [136]. HSCs are also present across many other parts of the bone marrow, until twenty to thirty where they become

[7]Determining the site of HSCs during embryogenesis was highly challenging to researchers. The embryo is hemogenic before the AMG region develops. First attempts to uncover the source of postnatal HSCs indicated that HSCs emerged from the yolk sac. This was proven not to be the case. Instead, distinct stages of hematopoietic development occur from the yolk sac before moving to the fetal liver. The HSCs, postnatally, are independent of these early embryogenic hematopoietic cells [130, 131].

concentrated in the pelvis and sternum. As we age the number of HSCs changes and the pool of stem cells undergoes *clonal hematopoiesis* (CH) (Figure 1.2, Hematopoietic Stem Cells). CH defines the condition of homogenization of the HSCs and serves as a marker for disease, being diagnosed based on the frequency of somatic alterations. In the presence of specific drivers, the term for this CH is *clonal hematopoiesis of indeterminate potential* (CHIP), otherwise it is called *age-related clonal hematopoiesis* (ARCH) (if cell counts are within normal ranges). In Chapter 5 we show that our cell fate marker developed in Chapter 4 can be used to assess this homogenization in HSCs.

1.6 Clinical relevance of understanding somatic evolution and cell fates

Clinical relevance of basic research is often overstated. The work that follows does not directly offer any direct clinical impacts by way of treatment. Rather, the work presented here strives to understand somatic evolution within homeostatic tissue, to focus on prevention and early detection as applied to epithelial and hematological diseases. We do not yet have a full understanding of the temporal clonal dynamics within the homeostatic tissues discussed above, or the extent to which we can use them to understand underlying tissue dynamics, or use them to develop early diagnostic and prognostic markers. We will revisit this discussion in the final chapter, highlighting how the findings here can be understood in the scope of their clinical relevance and the direction that these projects may drive future inquiry.

1.7 What has changed?

As highlighted in each section throughout this introduction, there has been a coalescence of both knowledge and ideas that make it possible to carry out the research within this thesis that would not have been possible five to ten years prior:

- We now know that oncogenic driver mutations accrue within normal cells throughout the human body.
- High resolution spatial and temporal datasets are now routinely generated from both disease and normal tissues.
- Mathematical oncology has brought to light the power of mechanistic models and is being integrated with experimental approaches as a critical tool for un-

derstanding tumorigenesis as it allows full access to dynamics over time and space.

To date there has been minimal research focused on the integrated understanding of homeostasis and somatic evolution of stem cells and tissue through mechanistic modeling. Motivated by the fact that homeostasis is the critical, often missing component, to understanding disease development, treatment response, and disease progression I have set out to gain insight into homeostasis using somatic evolution.

1.8 Outline of project aims

In this thesis I aim to integrate clinical omics data with mechanistic models across different tissue types to better understand somatic evolution, progenitor cell dynamics, and early departure from homeostasis. The first project utilizes publicly available data and contributes to the paradox of oncogenic SNVs expanding in normal stratified squamous epithelium (the epidermis) driven by underlying stem cell turnover. In the next two chapters I turn my attention to measures of stem cell dynamics and homeostasis using methylation, first in another epithelial tissue (intestinal crypts) and then on connective tissue (blood). For each chapter, the aims can be summarized as follows:

- **Chapter 3:** We aim to develop a first principles *three-dimensional* (3D) mechanistic model of the homeostatic epidermis containing cells with base pair resolution mutation data. Using this model and novel method for mutation tracking *in silico* we aim to examine neutrally expanding subclonal populations, and compare simulations to patient data. By including mechanisms for how NOTCH1 and TP53, oncogenic drivers, are limited by tissue homeostasis we aim to explore clonal dynamics given functionally heterogeneous clonal populations.

- **Chapter 4:** We aim to uncover *fluctuating methylation clocks* (FMCs) by developing a mechanistic spatial model of the homeostatic crypt complete with cells containing epigenomes. We then aim to use this model to aid in identifying candidate *fluctuating CpG* (fCpG) sites within the colon and small intestinal crypts suitable for use as markers of stem cell fate within single crypts. Using this mechanistic model we aim to validate an inference approach to resolve stem cell dynamics in glandular tissues for numerous patients.

- **Chapter 5:** Expanding on our knowledge of FMCs, we aim to explore whether they could be a marker for loss of polyclonality in the HSC compartment. We

must first develop a simple model of HSC and explore the theory of expanding subclonal populations. We then aim to examine thousands of patients, through publicly available methylation data, to determine whether HSCs can serve as a marker of HSC clonal expansions from normal turnover to acute leukemias. Lastly, we hope to provide a novel dataset of CHIP patients to develop a, mutation agnostic, diagnostic model based on FMCs.

We begin with an overview of datasets used in each chapter in Figure 2.1 and outline the specific methods used for data collection, processing, and analysis in Chapter 2. To conclude, in Chapter 6 we summarize and critically review the insights gained from the work in this thesis. We further discuss how the findings herein may be extended in the future.

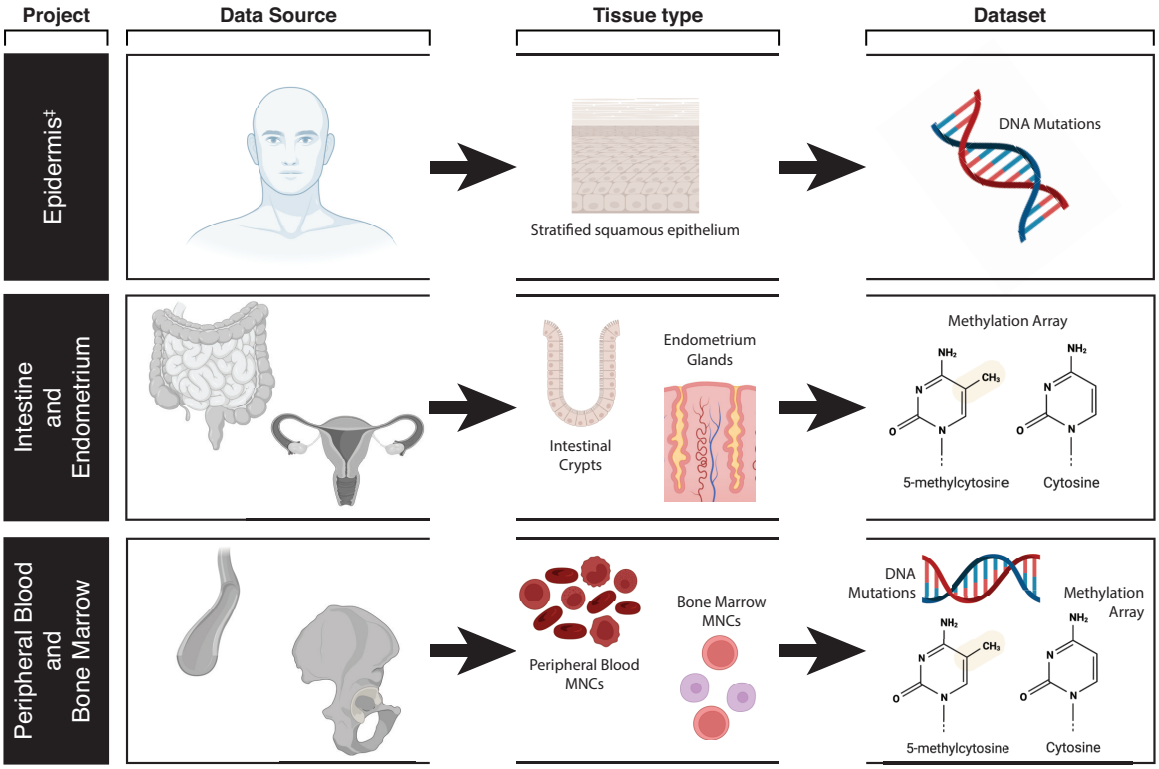


Figure 2.1: Overview of primary datasets. This figure highlights the three primary projects throughout this body of work and their corresponding datasets: Epidermis mutational data obtained from the eyelid, methylation data from single intestinal crypts and endometrium glands, and methylation with paired DNA sequencing data from *peripheral blood* (PB) and *bone marrow* (BM). ‡ denotes that only publicly available data is used.

Chapter 2

Materials and Methods

2.1 Human samples

2.1.1 Epidermis Patient Data

Patient data was previously procured and exhaustively characterized by Martincorena *et al.* [25]. However, the data is filtered to remove patient biopsies present in multiple biopsies as this may prove problematic in its interpretation, as previously noted [31]. The clonal area is taken as the product of the biopsy size (mm^2) and twice the variant allele frequency. This curated dataset provides the necessary information to perform direct comparisons with simulation data.

2.1.2 Tissue Collections from human intestines and endometrium

Tissues were collected at the University of Southern California Keck School of Medicine from excess surgical samples taken in the course of routine clinical care, with Institutional Review Board approval. Additional normal colon specimens were obtained from the University College London Hospital (UCLH) Cancer Biobank (REC approval: 15:YH/0311).

2.1.2.1 Data availability for methylation data from intestines and endometrium

Illumina EPIC array data (colon, small intestine and endometrium) collected in the process of this study are currently available at the European Genome–Phenome Archive (EGA) (accession EGAS00001005514).

ID	Age	Location	Disease	Cancer	Samples
CE	77	Colon	-	CRC	CE19,CE20,CE21,CE18
EAN	59	Colon	-	CRC	EAN10,EAN11,EAN12,EAN9
CB	35	Colon	AFAP	-	CB1,CB2,CB4,CB5
PA	24	Colon	-	-	PA2,PA3,PA4,PA5
KN	50	Colon	-	adenoma	KN4,KN5,KN6,KN2
O	17	Colon	-	CRC	O5,O6,O7,O8
FA	19	Colon	FAP	-	FA1,FA2,FA4,FA5
CD	82	Colon	-	CRC	CD4,CD5,CD16,CD17
CG	22	Colon	FAP	-	CG1,CG2,CG4,CG5
UQ	42	Endometrium	-	-	UQ25,UQ22,UQ23,UQ24
UAH	55	Endometrium	-	-	UAHB16,UAHB17,UAHB18,UAHB20
UF	38	Endometrium	-	-	UF11,UF13,UF14,UF15
UT	85	Endometrium	-	-	UT19,UT21,UTL22,UTL23
UR	32	Endometrium	-	-	UR1,UR3,UR4,UR28
UAG	56	Endometrium	-	-	UAGB11,UAGM13,UAGW14,UAGW15
US	33	Endometrium	-	-	US16,US17,US20,US29
UH	17	Endometrium	-	-	UHL13,UHL14,UH15,UH16
N	75	Small Intestine	-	CRC	N1,N2,N3,N4
KS	17	Small Intestine	-	-	KS22,KS24,KS25,KS26,KC27,KC28,KC29,KC30
M	27	Small Intestine	IBD	-	M5,M7,M8,M9
BS	67	Small Intestine	-	pancreaticCA	BS10,BS8,BSs6,BSs7
E	58	Small Intestine	-	CRC	E1,E2,E3,E5
F	46	Small Intestine	-	ovarianCA	F6,F7,F9,F10
A	76	Small Intestine	-	CRC	A14,A15,A6,A6-2

Table 2.1: Fluctuating Methylation Clock Patient Samples. Sample information, including individual ID, age, sex, tissue the sample was derived from, any underlying diseases, and whether the individual was diagnosed with cancer (*colorectal cancer* (CRC), *attenuated familial adenomatous polyposis* (AFAP), *familial adenomatous polyposis* (FAP), *ovarian cancer* (ovarianCA), *pancreatic cancer* (pancreaticCA), and *irritable bowel disorder* (IBD)). In total, 43 colon samples from 9 patients, 28 small intestine samples from 7 patients, and 32 endometrium samples from 8 patients.

2.1.3 Publicly available methylation data

Blood methylation data were obtained from the *Gene Expression Omnibus* (GEO) [137, 138] using beta values as provided. The data sets are GSE40279 [139] (normal blood, Figure 5.2A), GSE73115 [140] (ten-year serial samples, Figure 5.2B), GSE51759 [141] (*myelodysplastic syndromes* (MDS)), GSE42042 [142] (*essential thrombocythaemia* (ET), *polycythaemia vera* (PV), *primary myelofibrosis* (PMF)), GSE106600 [143] (*chronic myeloid leukemia* (CML)), GSE105420 [144] (*chronic myelomonocytic leukaemia* (CMML)), GSE62298 [145] (*acute myeloid leukemia* (AML)) and GSE69229 [146] (ALL) (Figure 5.2C). Additional methylation data for normal samples GSE87571 [147] and pediatric AML cases GSE133986 [145].

2.1.4 Study Participants of clonal hematopoiesis of indeterminate potential study

Patient samples were collected with informed consent under the Mechanisms of Age-Related Clonal Hematopoiesis (MARCH) Study. Written informed consent was obtained from all participants in accordance with the Declaration of Helsinki. This study was approved by the Yorkshire & The Humber - Bradford Leeds Research Ethics Committee (REC Ref: 17/YH/0382). Participants were recruited from individuals due to undergo elective total hip replacement surgery at the Nuffield Orthopedic Centre, Oxford. Exclusion criteria were: History of rheumatoid arthritis or other inflammatory arthritis, history of septic arthritis in the limb undergoing surgery, history of hematological cancer, bisphosphonate use, and oral steroid use. Patient characteristics are summarized in (Figure 5.3).

2.2 DNA

2.2.1 Methylation Arrays

Crypt and endometrial tissues were collected at the University of Southern California Keck School of Medicine from excess surgical samples taken in the course of routine clinical care, with Institutional Review Board approval. Crypts or endometrial glands were isolated using an EDTA-washout method as previously described [71, 148].

2.2.2 DNA extractions from peripheral blood and bone marrow

At the time of surgery, trabecular bone fragments and *bone marrow* (BM) aspirates were obtained from the femoral canal and collected in anticoagulated buffer containing acid-citrate-dextrose, heparin sodium and DNase. Samples of *peripheral blood* (PB) were collected in EDTA vacutainers prior to surgery. PB and BM *mononuclear cell*s (MNCs) were isolated by Ficoll density gradient centrifugation and viably frozen in liquid nitrogen. PB granulocytic cell pellets were frozen at -20 for later DNA extraction. Genomic DNA was extracted from BM MNCs, PB granulocytes using a DNeasy Blood & Tissue Kit (Qiagen).

2.2.3 Targeted DNA Sequencing of clonal hematopoiesis of indeterminate potential samples

Targeted DNA sequencing was performed on BM MNC and PB granulocyte DNA samples. Pre-capture DNA libraries were prepared using the KAPA HyperPlus protocol (KK8514, Roche) either by hand or using a Biomek FXp Automated Workstation (Beckman Coulter). $100ng$ of genomic DNA was fragmented by enzymatic fragmentation. Following end repair and A-tailing, adapter ligation was performed using KAPA dual-indexed adapters (KK8722, Roche). Library cleanup and double-sided size selection was performed using Agencourt AMPure XP beads (Beckman-Coulter) to obtain fragments of $\approx 320bp$. Libraries were amplified by ligation-mediated PCR for 6 cycles using a KAPA HiFi HotStart high-fidelity DNA polymerase (Roche) and purified using AMPure XP beads.

Targeted capture was performed using a custom pool of biotinylated capture probes (SeqCap EZ Prime Choice, Roche) targeting 97 genes recurrently mutated in clonal hematopoiesis and myeloid malignancies (see Supplementary Table x)[149–155]. Amplified DNA libraries were hybridized to the capture probes in pools of $10-12$ samples according to the manufacturer's instructions. The captured DNA libraries were amplified by 14 cycles of PCR using a KAPA HiFi HotStart DNA polymerase and purified using AMPure XP Beads.

Post-capture amplified DNA libraries were quantified by Qubit (Life Technologies) and size distribution and quality analyzed using a Bioanalyser chip (Agilent Technologies). Libraries were pooled in equimolar concentrations and were sequenced on an Illumina NextSeq 500 using paired end reads.

2.2.3.1 Variant calling in targeted sequencing of mononuclear cells

Sequencing data were analyzed in a custom pipeline based on the GATK Best Practices (GATK v4.1.2.0 and Picard v2.21.3). Raw sequencing reads were converted to an unmapped BAM file and adapter sequences soft-clipped using Picard MarkIlluminaAdapters. Following conversion back to a FASTQ file, reads were mapped to the hg38 human reference genome assembly using the Burrows–Wheeler aligner (BWA-mem v0.7.17)[156] with alternate contig-aware alignment. Mapped and unmapped BAM files were merged using MergeBamAlignment and reads from different sequencing lanes were combined. Duplicate reads were marked using Picard MarkDuplicates and base quality scores recalibrated with GATK BaseRecalibrator and ApplyBQSR. Somatic variant calling was performed on the pre-processed BAM files using VarDict (v1.7.0) [157] and Mutect2 [158] in tumor-only mode. For VarDict, variants were called with a minimum variant allele frequency of 0.01, minimum base quality score of 25 and minimum supporting reads of 2, with indel realignment and removal of adapter sequences. For Mutect2, a minimum tumor LOD of 2 was used and variants were filtered for sequence context-dependent artifacts using FilterMutectCalls and FilterByOrientationBias. Indels were left-aligned and normalized using bcftools norm (v1.9). Variants were annotated using ANNOVAR [159]. Target enrichment metrics and coverage was calculated using Picard CollectHsMetrics and custom scripts.

VarDict and Mutect2 variant calls were analyzed separately to identify a consensus list of high-confidence variants. The following post-processing filters were applied to VarDict calls to exclude likely sequencing artefacts: (1) Minimum of 5 variant reads for SNVs (with at least 2 reads in forward and reverse directions), or minimum of 10 variant reads for indels (with at least 4 reads in forward and reverse directions). (2) Minimum base quality score 30. (3) Minimum mapping quality score 40. (3) Maximum strand bias Fisher p-value of 0.0001. (4) No position bias towards beginning or end of reads. The following post-processing filters were applied to Mutect2 calls: (1) Passed all default Mutect2 filters or only failing the clustered_events filter. 2) Minimum of 5 variant reads for SNVs (with at least 2 reads in forward and reverse directions), or minimum of 10 variant reads for indels (with at least 4 reads in forward and reverse directions). Variants were flagged as likely germline or sequencing artefacts if any of the following applied: (1) Variant allele with a population allele frequency > 1 in 1,000 according to any of three large polymorphism databases (Gnomad, 1000 Genomes Project, ESP6500) that is not a hotspot driver mutation with COSMIC recurrence > 100 or is present in a list of clonal hematopoiesis-associated mutations compiled from five large studies [151, 152, 155, 160, 161]. (2) *variant allele*

frequency (VAF) between 0.4-0.6 or > 0.9 unless recurrent in COSMIC > 5 times, or previously reported in clonal hematopoiesis. (3) Present in a panel of normal cord blood samples. (4) Recurrent in the cohort unless present at least 5 times in COSMIC or at least 2 times in the clonal hematopoiesis studies.

2.2.3.2 Annotation of clonal hematopoiesis driver mutations

Samples were annotated as having CH based on the presence of driver mutations in bone marrow sequencing. Variants were annotated as driver mutations using the following criteria: (1) Truncating mutations (nonsense, splice site or frameshift indel) in the following genes: DNMT3A, TET2, ASXL1, NF1, IKZF1, RAD21, WT1, KMT2D, SH2B3, TP53, CEBPA, RUNX1, BCOR, KDM6A, STAG2, PHF6, KMT2C, PPM1D and ZRSR2. (2) Truncating variants in CALR exon 9. (3) Non-synonymous variants at the following hotspot residues: CBL E366, L380, C384, C404, R420 and C396; DNMT3A R882; FLT3 D835; IDH1 R132; IDH2 R172 and R140; JAK2 V617F; KIT W557, V559 and D816; KRAS A146, Q61, G13 and G12; MPL W515; NRAS Q61, G12 and G13; SF3B1 K700 and K666; SRSF2 P95; U2AF1 Q157, R156 and S34. (4) FLT3 internal tandem duplication. (5) Non-synonymous variants occurring within the following residues of DNMT3A: p.292-350, p.482-614 and p.634-912; or TET2: p.1104-1481 and p.1843-2002. (6) Non-synonymous variants reported at least 10 times in COSMIC with VAF $< 40\%$. (7) Non-synonymous variants falling within an annotated InterPro domain with VAF $< 40\%$. (8) Non-synonymous variants reported in COSMIC > 100 times.

2.2.3.3 Comparison of bone marrow and peripheral blood allele frequencies

Variants were called in peripheral blood DNA sequencing data as described above, using a minimum variant allele frequency of 0.005. Unfiltered variant calls were intersected with the list of curated BM variants to compare the VAF between BM and PB. For those mutations known to be present in BM which were not called in PB, raw allele counts were performed using bcftools mpileup and the allele frequency calculated.

Sample	Putative Driver (PB/BM)	PB Driver(VAF_{PB}/VAF_{BM})	BM Driver(VAF_{PB}/VAF_{BM})	Δ_{VAF} ($\mid VAF_{PB} - VAF_{BM} \mid$)
NOC007_PB	ETV6	0.0202 / -	- / 0.0184	0.0018
NOC008_PB	ASXL1	0.1932 / -	- / 0.1665	0.0267
NOC009_PB	DNMT3A	0.0196 / -	- / 0.0188	0.0008
NOC013_PB	GNB1	0.2678 / -	- / 0.1814	0.0864
NOC015_PB	TET2	0.0113 / -	- / 0.0188	0.0075
NOC020_PB	DNMT3A	0.0295 / -	- / 0.0535	0.024
NOC030_PB	NF1	0.0272 / -	- / 0.0206	0.0066
NOC031_PB	TET2	0.0907 / -	- / 0.0802	0.0105
NOC037_PB	DNMT3A	0.0932 / -	- / 0.0793	0.0139
NOC056_PB	DNMT3A	0.1531 / -	- / 0.1268	0.0263
NOC062_PB	DNMT3A	0.4275 / -	- / 0.3657	0.0618
NOC075_PB	TP53 / ATRX	0.0103 / 0.0342	- / 0.1055	-
NOC078_PB	TET2	0.1002 / -	- / 0.0557	0.0445
NOC079_PB	SF3B1 / TET2	0.0147 / 0.0114	- / 0.018	-
NOC082_PB	ASXL1	0.0777 / -	- / 0.0804	0.0027
NOC083_PB	DNMT3A	0.0251 / -	- / 0.0325	0.0074
NOC084_PB	DNMT3A	0.0971 / -	- / 0.0626	0.0345
NOC090_PB	DNMT3A / SF3B1	0.0178 / 0.0129	0.0168 / 0.0247	-
NOC092_PB	ASXL1	0.027 / -	- / 0.0337	0.0067
NOC097_PB	TET2	0.0521 / -	- / 0.0427	0.0094
NOC098_PB	TET2	0.0672 / -	- / 0.0796	0.0124
NOC103_PB	KDM6A	0.0659 / -	- / 0.0358	0.0301
NOC108_PB	DNMT3A	0.0213 / -	- / 0.0715	0.0502
NOC113_PB	DNMT3A / SMC1A	0.0246 / 0.0192	- / 0.023	-
NOC114_PB	SF1	0.014 / -	- / 0.018	0.004
NOC115_PB	TET2	0.175 / -	- / 0.1007	0.0743
NOC125_PB	DNMT3A / KMT2D	0.069 / 0.0399	- / 0.0542	-
NOC131_PB	DNMT3A	0.0452 / -	- / 0.0607	0.0155
NOC132_PB	DNMT3A	0.2494 / -	- / 0.1275	0.1219
NOC137_PB	ASXL1	0.1634 / -	- / 0.1413	0.0221
NOC138_PB	DNMT3A	0.0796 / -	- / 0.0461	0.0335
NOC146_PB	TET2	0.0523 / -	- / 0.0318	0.0205
NOC150_PB	EP300 / TET2	0.0361 / 0.0245	0.028 / 0.025	-
NOC170_PB	TET2	0.0503 / -	- / 0.0262	0.0241
NOC171_PB	DNMT3A	0.233 / -	- / 0.2297	0.0033
NOC176_PB	PPM1D	0.1828 / -	- / 0.1828	0.0
NOC177_PB	DNMT3A	0.0294 / -	- / 0.0163	0.0131
NOC190_PB	DNMT3A	0.0195 / -	- / 0.0107	0.0088

Table 2.2: Paired PB and BM for the largest driver mutations in each of the paired samples for the CH cohort.

2.3 Methylation

2.3.1 Methylation arrays and analysis

DNA methylation from the intestine and endometrium samples was measured with EPIC bead arrays (Illumina) using the Restore protocol and the manufacturers protocols [162].

DNA methylation analysis was performed on samples of peripheral blood granulocyte DNA, selecting samples with CH variants in a range of VAFs and controls without CH. DNA methylation was bisulfite converted and measured with Infinium MethylationEPIC bead arrays (Illumina) using the manufacturers protocols at King's College London IoPPN Genomics & BioMarker Core Facility.

IDAT files were processed with the noob normalization function using the R package minfi and then further filtered using the Chip Analysis Methylation Pipeline (ChAMP) default filters [163–166].

2.3.2 Copy number analysis on publicly available methylation data

Copy number alterations (CNAs) were called using intensity normalized methylation data on the publicly available normal blood samples using ChAMP in R [163, 167, 168]. The publicly available datasets and CH cohort presented here use different probe sets, 450k and 850k EPIC arrays respectively. Due to this, each cohort CNAs were called based on the average probe intensities within each cohort individually across all samples. Filtering of the segmentation means was performed in Python using custom scripts where cutoffs were applied manually such that a CNA had to be supported by ≥ 3 markers as used previously [167]. A minimum segment length cutoff is used of 5kb. A segmentation mean had to be above or below 0.4 the median ratio for determining a gain or loss. For analysis of age-related frequency of CH classified individuals under the age of 20 were excluded from analysis as the HSC because the HSC pool undergoes drastic changes leading to a heterogeneous stem cell pool until the age of ≈ 20 [169]; patients younger than this are excluded from additional analysis involving CNAs.

To analyze genes and recurrent regions with aneuploidy we performed a new set of copy number calls using the classified normals as the baseline to assess differences in the classified CH samples for each cohort. Recurrent CNAs were identified using a custom function with a defined overlap of losses/gains of 5kb in at least 3 losses/gains. These recurrent genomic positions were labelled with their respective cytobands for

the regions where 100% overlap exists. Genes found within each CNA were annotated using bedtools [170] and the NCBI Reference Genome Database (hg19) [171]. Gene set enrichment analysis was performed on genes that fall within CNA segments using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) [172, 173]. Gene set enrichment analysis was conducted for genes associated with various diseases to check whether there was aneuploidy in genes associated with hematological malignancies or other hematological disorders. We used a minimum number of genes required in a particular annotation term to be two with a maximum P-value of 0.05 and a false discovery rate of 0.01; we also excluded all HLA genes from the analysis as these are highly polymorphic and can increase false discoveries. Finally, Using a custom list of 569 genes implicated in driving CH (based on literature and driver genes defined in Section 2.2.3.2) and the list of unique driver genes from intOGen [174]. Genes within CNA segments are cross referenced with this list of CH and oncogenic drivers.

2.4 Equations and Statistics

2.4.1 First incomplete moment

Given a neutrally expanding keratinocyte population driven by progenitor proliferation with a basal progenitor loss/replacement rate given by $r\lambda$, it follows that the clone, n, distribution can be calculated for any given patient of age t:

$$P_{n\,0}(t) = \frac{1}{ln(r\lambda t)} \frac{e^{-\frac{n}{r\lambda t}}}{n} \tag{2.1}$$

The cumulative probabilities of clone size distributions can then be used to assess neutrality, under an expectation of exponential size dependency given by the first incomplete moment [31]:

$$\mu_1(n,t) = \frac{1}{\langle n(t) \rangle} \sum_{m=n}^{N} m P_m(t) \tag{2.2}$$

Here, N represents the largest clone observed and $\langle n(t) \rangle$ is the average mutant clone size. This assessment of neutrality allows for a straightforward method when applied to simulation data and patient biopsies. Clone size can be calculated by $2 * VAF * B$, where B is either patient biopsy size or simulated biopsy area in mm^2.

2.4.2 Epidermis model parameterization

Parameter values for the spatial model were selected based on iterative *redundancy analysis* (RDA) using R [168] (vegan package [175]), where constraining variables were the emergent properties measured as a necessary component of homeostasis:
Primary Constraining Variables:

1. Mean Tissue height
2. Mean Loss/replacement rate within the basal layer

Secondary Constraining Variables:

1. Mean Overall tissue turnover time
2. Mean cell age
3. Mean Basal cell density

During the first iteration of parameterization variables are set randomly choosing values between zero and one for 20,000 two-year simulations. Diffusible parameters (ψ, γ, λ), where values were examined between zero and 1/6, were not subject to this iterative process. For the diffusion coefficient (ψ), values greater than 1/6 (for 3D, 0.25 in *two-dimensional* (2D)) results in numerically unstable solutions for a given lattice point because it fails to satisfy Courant-Friedrichs-Lewy condition that guarantees convergence of a forward-time central-space finite difference method ($\psi + \frac{\delta t}{h^2} \leq 1/6$ (3D)). After each model run we collect the constraining variable data and in conjunction with the model parameters, we then perform RDA. Variance explained by each parameter are used to determine the influence of each parameter on the constraining variables. Refinement of the model parameters is then performed based on the variance explained and the linearity of that parameter on the constraining variables. In this way we are able to constrain the variable range iteratively by reducing the number of simulations to $5,000$ (two-year simulations) while converging on parameter values. Iterations are repeated until target values are reached for each of the constraining variables of homeostasis where the combination of parameter values yields homeostasis. The secondary constraining variables served as a sanity check rather than a strict parameterization. The process can be described by by Algorithm 1.

Algorithm 1 Execute Simulations given parameter ranges for RDA Analysis

Require: *Initial prior range*

1. *Randomly select parameter values from prior range*
2. *Run 2-year simulation of model for parameter set*
3. *Collect and calculate constraining variable values*

while *Homeostasis not reached* **do**

1. *Execute steps 1-3 above (20,000 initially, 5,000 each round after)*
2. *Perform RDA and analyze parameter influence on simulations*
3. *Run 2-year simulation of model for parameter set*
4. *Constrain prior range*

end while

2.4.3 Departure from homeostasis in the epidermis

In order to assess how the *three-dimensional hybrid cellular automaton* (3D-HCA) behaves in response to functional heterogeneity being introduced through NOTCH1 and TP53 mutants we use a distance metric d_H. This distance metric summarizes the departure from homeostasis given a normal morphology such that the overall population size is constant throughout the simulation (e.g. all neutral simulations). It is defined as

$$d_H = log\left(\frac{1}{n}\sqrt{\sum_{i=0}^{n}(N_k(t_i) - N_s(t_i))^2}\right) \tag{2.3}$$

where N_k and N_s are the population size at the given time, t_i, for the homeostatic and simulation in question, respectively. This provides a single, relative value to assess how far from homeostasis any simulation departs.

2.5 Introducing model error

2.5.1 Calculating variant allele frequencies from simulations

As expounded in the introduction (Section 1.4.1) we see that there is an error associated with measurements from NGS approaches that lead to the downstream calculated VAFs. Simulations from Chapter 3 and Chapter 5 require some amount of correction for these perfect outputs that the simulations provide. The primary error associated with NGS approaches is one of sampling, where base calls for homozygous sites are identical and base calls for heterozygous sites follow a 1:1 binomial distribution [117]. For this reason variant calling tools have successfully used a binomial distribution or multinomial distribution for approximating heterozygous base counts

[176–181]. More nuanced approximations are now taken into account that understand better how sources of error can skew heterozygous allele ratios, but these differ by which individuals perform library preparations, which bioinformatics tools are used, and the presence of aneuploidy. Hence, we build upon a simple binomial distribution approach to adding a small amount of noise to the VAF calculated model data [82].

The true VAF (assuming heterozygosity), VAF_t, is given from $VAF_t = \frac{N_i}{2N_e}$, where N_i is the number of cells with a given mutation and N_e is the population size (i.e., number of cells in a simulation). To perform *in silico* sequencing that would yield a corrected VAF we must translate this ratio into counts. We first define D_i as the number of reads for at a variant's loci. In cases where a single value is described as the target depth, d, the number of reads is drawn from a Poisson distribution, which yields $D_i \sim Poisson(d)$. In cases where a collection of variants or depths from NGS is known a maximum likelihood estimation [182] is used to determine the shape (k_c) and scale (p_c) parameters of a Gamma distribution to be used so that $D_i = Gamma(k = k_c, p = p_c)$. The number of reads for a given variant (f_i) is determined by $f_i = B_0(n = D_i, p = VAF_t)$. Using these counts, the number of variant reads (f_i) and the total number of reads for that locus (d_i), provides us with a 'sequenced' VAF where $VAF_s = \frac{f_i}{D_i}$.

2.5.2 Fluctuating CpG distribution error model

Models of CpG distributions allow for values at 0, 50, or 100%. However, these values aren't ever observed from the data obtained from methylation microarrays due to background noise, normalization, and technical sources of error (see Section 1.4.1). We address the saturation observed in the observed β values by performing a linear transformation on the agent based model's output methylation status representing the methylated and unmethylated fractions (*in silico* β values, m_β). The linear transformation yields corrected beta values, $\vec{m_\beta}$, similar to that of the observed data where $\hat{\vec{m}_\beta} = (\epsilon - \Delta) * \vec{m_\beta} + \Delta$. The parameters Δ and ϵ are determined by the saturation observed in the data. A small amount of noise is then added to each of the transformed β values by modeling the $\hat{m}_\beta$ values as β-distributed random variables which yields a vector of "corrected" beta values ($\hat{\vec{\beta}} = [\hat{\beta}_0, \hat{\beta}_1, \dots]$), where $\hat{\beta}_i \sim B(\kappa_{\beta_i}, \kappa(1 - \beta_i))$ where $\kappa \geq 1000$.

2.6 RNA data from the intestinal crypts

RNA expression data for normal tissue was derived from 40 individuals that were retrieved from *The Cancer Genome Atlas* (TCGA) [114] and used to examine fCpG association with gene expression.

Chapter 3

Homeostasis and Carcinogenesis in a First Principles Model of the Epidermis

3.1 Background

Recent studies have documented that substantial numbers of mutations accumulate in normal human tissues, with evidence for selection of canonical, oncogenic driver mutations [25, 26, 28, 35, 183, 184]. Typically, selection for these driver mutations is modeled by increased cell proliferation; however, requirements of skin homeostasis imposes spatial constraints that inherently limit selective sweeps. Because these mutations do not appear to disrupt homeostasis, these mutations can also provide important new information on normal human tissue dynamics, which are typically studied with experimental fate markers in animal systems. Although substantial numbers of mutations accumulate with age, and a quarter of all cells carry driver mutations [26], skin thickness and mitotic activity do not significantly change. A controversy currently exists between our understanding of mutation selection and neutral evolution within tissues [25, 31], in part because our understanding of how a mutant clone is able to expand within normal tissue is completely lacking. Furthermore, most mutations are passengers and skin subclone size/frequency distributions are consistent with neutral evolution or the absence of detectable selection. There is strong evidence of selection by driver mutations (NOTCH and TP53) manifested by dN/dS ratios and their generally larger subclone sizes in normal skin [25]. Varying exposures to sunlight between individuals also complicates our understanding of mutation accumulation. The challenge is to develop a model that integrates selection, neutral evolution, and sunlight. The reward would be a realistic model of lifelong

human skin cell dynamics.

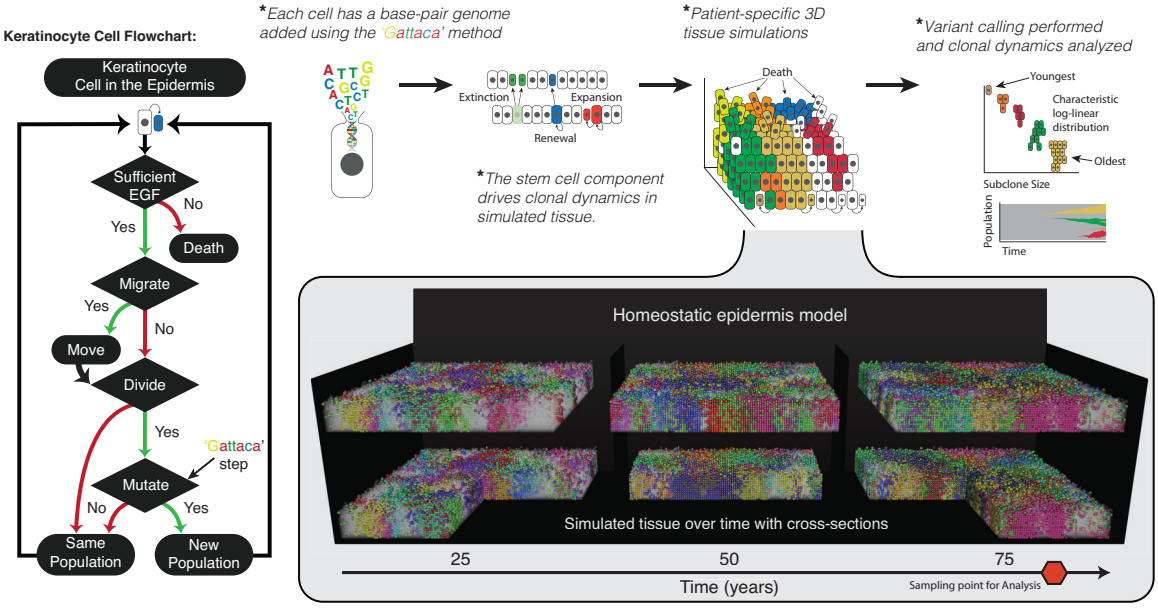


Figure 3.1: The homeostatic epidermis model and Gattaca method. The homeostatic epidermis is comprised of individual keratinocyte cells that are governed by a set of decisions (Flowchart). Spatial structure of the *three-dimensional* (3D) model and a *growth factor* (GF) gradient governs the loss/replacement in the stem cell niche at the basal layer. The model is entangled with DNA sequence data by the 'Gattaca' step, where base-pair mutations can be introduced with every *in silico* cell division. These mutations act as cell fate markers and different subclones are represented by different cell colors. Therefore, cells in the simulated skin produce a patchwork of different sized subclones, where each colored population differs by at least one mutation.

3.1.1 Chapter aims

1. Construct a 3D first principles model of a homeostatic epidermis.

2. Develop a method for interrogating and tracking base pair resolution somatic mutations *in silico* to study evolutionary dynamics.

3. Examine neutral clonal dynamics of the homeostatic epidermis to evaluate evidence of neutrality in published patient data.

4. Introduce *in silico* functional heterogeneity (defined by 'driver' mutations with a fitness advantage) through integrating TP53 and NOTCH1 functional alterations.

3.2 First principles model of the Epidermis

3D-HCA Model Parameters

Parameter	Role
α	Level at which apoptosis can occur by chance.
θ	Random probability of death.
ξ	Scaling factor to control roliferation rate.
ϱ	Probability of moving into an empty position.
ω	Probability governing direction of division.

Growth Factor Diffusible Parameters

Parameter	Role
ψ	Rate of GF diffusion from basal layer.
λ	Rate of GF decay.
γ	Rate of GF consumption by keratinocytes.

Table 3.1: Model Parameters. Parameters are separated into the cellular automata parameters and those belonging to the diffusible *growth factor* (GF).

Here we have constructed a homeostatic, 3D-HCA to mechanistically model the human epidermis using the *hybrid automata Library* (HAL) framework [109]. We based our model on a simplified version of the vSkin model previously developed [185] to investigate the clonal dynamics of keratinocyte populations (Figure 3.1). In this simpler model, while we incorporate keratinocyte dynamics, we do not explicitly consider melanocytes and stromal cells (not in the epidermis) but rather, a single diffusible growth factor acting to define the stem niche on the basal layer. We define our area to approximate different sized biopsies assuming a median cell size of approximately $15\mu m$.

3.2.1 Growth Factor

The single growth factor governing keratinocyte apoptosis and proliferation is defined by a partial differential equation that dictates how our diffusible behaves throughout the domain over time. Similar to Kim et al. [185] & Basanta et al. [186], due to the interactions between discrete, on lattice cells, and the continuous growth factor variable we describe the discretized version of the equation. Over time our growth factor diffuses from an imposed, unchanging source (E_C) along the basal layer (i.e. $E(x, y, 0) = E_C$) with a diffusion rate (ψ, Table 3.1), representative of production from the fibroblast population in the dermis. The keratinocytes (K) consume the growth factor at a rate, γ (Table 3.1), and it decays naturally (λ, Table 3.1). The spatial and temporal dynamics of this microenvironmental factor in its discretized

form, using a first order Euler method, on a three-dimensional lattice is defined as follows, from time t to time $t + \delta t$ of $E([E(\eta, t)])$:

$$\frac{[E(\eta, t + \delta t)] - [E(\eta, t)]}{\delta t} = \psi \Delta_3 [E(\eta, t)] - \gamma [E(\eta, t)] K_\eta - \lambda [E(\eta, t)]. \tag{3.1}$$

Where $\eta \equiv (\eta_x, \eta_y, \eta_z)$, δt denotes the time step, K_η is set to 1 if the lattice point η is occupied by a keratinocyte and 0 otherwise. The Δ_3 denotes the central difference approximation of the Laplacian operator in three-dimensional space:

$$\begin{aligned} \Delta_3 f(\eta_x, \eta_y, \eta_z) \equiv & \frac{f(\eta_x + h, \eta_y, \eta_z) + f(\eta_x - h, \eta_y, \eta_z)}{h^2} \\ & + \frac{f(\eta_x, \eta_y + h, \eta_z) + f(\eta_x, \eta_y - h, \eta_z)}{h^2} \\ & + \frac{f(\eta_x, \eta_y, \eta_z + h) + f(\eta_x, \eta_y, \eta_z - h)}{h^2} \\ & - \frac{6 f(\eta_x, \eta_y, \eta_z)}{h^2}, \end{aligned} \tag{3.2}$$

where h is the grid size. Boundary conditions on the top ($\eta_z = Z$,Z: height) are no-flux ($\frac{\delta E}{\delta Z}|_{z=Z} = 0$) and on the bottom ($\eta_Z = 0$) are Dirichlet ($E(\eta_x, \eta_y, 0) = E_C$), while periodic boundary conditions are present on the left and right, front and back ($E(0, \eta_y, \eta_z) = E(X, \eta_y, \eta_z)$, and $E(\eta_x, 0, \eta_z) = E(\eta_x, Y, \eta_z)$)). Each time step we solve equation 3.1 until steady state reached (< 100 iterations). A steady state value of the growth factor is used to determine keratinocyte birth and death in the model.

3.2.2 Keratinocyte dynamics

Cell death can occur in one of two ways at each time step for each cell (a cell time step, δt, is equal to 1 day), either through apoptosis or random turnover. First, the cell is dependent upon growth factor concentration to prevent a cell from undergoing apoptosis. The level at which apoptosis can occur by chance (α, Table 3.1)), is a model specific parameter and translates to a cell undergoing apoptosis when a randomly selected number, c_α ($c_\alpha \in U(0, 1)$), is less than $P(K_\eta \; Death) = (1 - \frac{E(\eta,t)}{\alpha})^5$. If a cell survives in low growth factor concentrations or is not in an area with sufficiently low growth factor it is then subject to random death representative of an intrinsic probability of death if a randomly drawn number ($c_\theta \in (0, 1)$) is less than the random death parameter θ. Provided a cell does not die within a given time step it is capable of moving to an empty lattice position in its 6 immediate nearest neighbors i.e. the 3D von Neumann neighborhood ($N^v_{x,y,z} = \{(x, y, z + h), (x, y, z - h), (x + h, y, z), (x -$

$h, yz), (x, y-h, z), (x, y+h, z)\}$). This location is determined by assessing if any empty lattice positions exist within the cell's neighborhood. If multiple spaces are empty, one is chosen randomly. The model specific parameter dictating the probability of movement ($P_{Move}(K_\eta) = \rho$, Table 3.1) governs the cells ability to move into that empty space so that when a randomly selected number c_ρ from (0,1) is greater than ρ, then the cell moves into that empty position (i.e., if $c_\rho \in U(0,1) \geq \rho$, the cell will move into a randomly selected empty neighborhood, $N^v_{x,y,z}$).

Cells divide based on the underlying growth factor concentrations ($P_{Birth}(K_\eta) = \xi[E]_\eta$), where ξ is the proliferation scale factor (Table 3.1). This results in a stem–like population of cells dependent on the underlying growth factor concentration. A cell (K_η at η) may divide into two identical daughter cells ($D_{1,2}$) if a randomly selected number c_d ($c_d \in U(0,1)$) is less than $\xi[E]_\eta$ regardless of there being an empty position. One daughter cell occupies the position of the parent cell (location of $D_1 = \eta_(x, y, z)$) while the other daughter cell (D_2) occupies one of five locations, being above or on one of the four orthogonal neighbors of the parent cell. This location is governed by ω, the division location probability, whereby the second daughter cell is placed in a position whose location ($D_2(c_\omega, \omega)$) is determined according to

$$D_2(c_\omega, \omega) = \begin{cases} \eta_{x,y,z+h}, & \text{if } 0 \leq c_\omega \leq \omega, \\ \eta_{x+h,y,z}, & \text{if } \omega \leq c_\omega \leq \omega + \frac{1-\omega}{4}, \\ \eta_{x-h,y,z}, & \text{if } \omega \leq c_\omega + \frac{1-\omega}{4} \leq \omega + \frac{2(1-\omega)}{4}, \\ \eta_{x,y+h,z}, & \text{if } \omega \leq c_\omega + \frac{2(1-\omega)}{4} \leq \omega + \frac{3(1-\omega)}{4}, \\ \eta_{x,y-h,z}, & \text{if } \omega \leq c_\omega + \frac{3(1-\omega)}{4} \leq \omega + \frac{4(1-\omega)}{4}, \end{cases} \tag{3.3}$$

where η_{xyz} is the coordinates of the parent cell's lattice position, h is the grid size, and c_ω is a randomly selected number from $(0, 1)$ (i.e., $c_\omega \in U(0, 1)$. Boundary conditions for cells are the same as those for our diffusible growth factor, being periodic on the left and right, and front and back to enable spatial competition, no-flux on the bottom ($\eta_z = 0$), and Dirichlet (Cell $(\eta_x, \eta_y, Z) = 0$) on the top. The model parameters are determined using redundancy analysis (Methods 2.4.2).

3.3 Epidermis Model Results

Niche driven behavior, while simplified, is an appropriate reduction in a homeostatic epidermis model. Several studies have provided insights into the heterogeneity of stem cell populations that influence clonal dynamics of the epidermis ([187–190]). The primary types of stem cells sustaining the epidermis being holoclones, undergoing extensive self-renewal. The remaining stem cells, transiently dividing before being

lost, meroclones and paraclones, make up the bulk of the cells of the epidermis. We can see that the observed frequency of holoclones within the population is a rare event (those that persist within the basal layer the longest periods of time, Figure 3.2). Most recently, in a commendable study by Tobias Hirsch et al. [187] we see that the proportion of holoclones is between 1.92% and 11.26%; we observe that 95% of stem cells within the basal layer are lost within 4 months, consistent with the findings

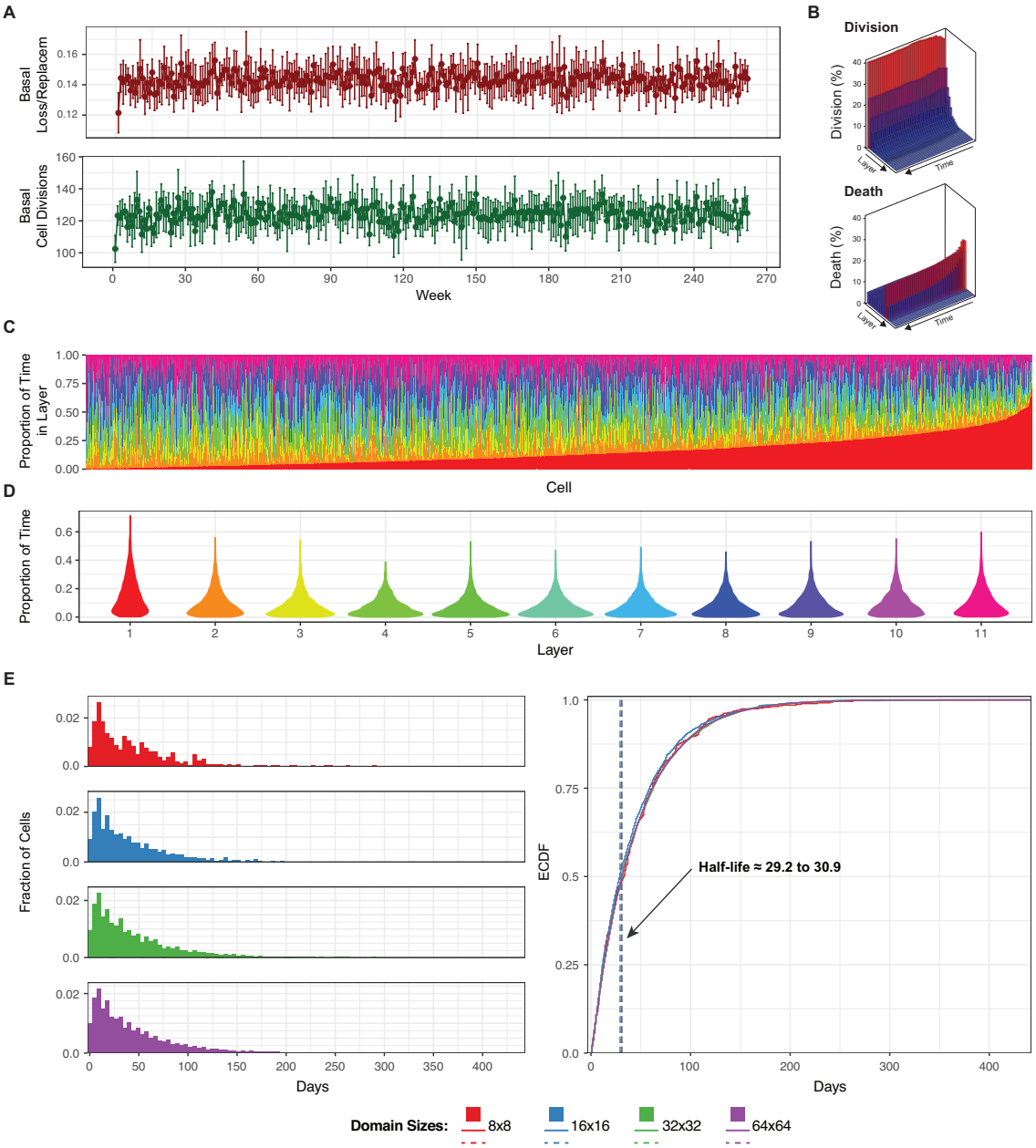


Figure 3.2: Three dimensional hybrid cellular automaton keratinocyte life-cycle (Caption next page.)

Figure 3.2: For a simulation of $1,024$ cells (X and Y dimensions of $32x32$) we see the weekly loss/replacement of the basal/progenitor cells (red) and the corresponding number of births (green)(**A**). For these same dimensions, the contributions of birth (top) and death (bottom) towards the overall population size of the simulated tissue are given for each layer of the epidermis over the course of a 2 year simulation (**B**). For **C** through **E** a three year period is simulated while positional and time information is only recorded for the second year. **C** and **D** show the proportion of time spent in each of the epidermis layers for each cell (**C**) or overall (**D**). **C** is sorted from least to greatest time spent in the basal layer for the time spent in the basal/progenitor layer and the colors for the layers are the same in **C** and **D**. **E** provides the distribution of time spent in the basal layer over four different domain sizes (**E**, left). The *empirical cumulative density function* (ECDF) (**E**, right) distributions provide additional information for the half-life ($t1/2 = \tau ln(2)$, τ is the mean time in the basal/progenitor layer) of cells within the basal/progenitor layer.

that meroclones and paraclone half-lives are $3-4$ months [187]. The distribution is a very heavy tailed distribution whereby most clones are lost regularly (non-holoclones) with few persisting in the basal layer where the overall half-life of all stem cells is ~ 4 weeks, an observation invariant to differing domain sizes (Figure 3.2C).

Within our homeostatic epidermis approach, instead of a simple stochastic birth/death model in two dimensions or non-spatial approach, we can see that wounded tissue regenerates and heals due to the models capability of regenerating tissue in response to a surge of our microenvironmental diffusible, a generic *growth factor* (GF). This diffusible increases in concentration in the wounded space where consumption of the diffusible is no longer taking place by the lost keratinocytes. We use experimentally derived measurements showing a progenitor loss/replacement rate of 0.51 per week [31] and calculate a basal layer loss/replacement rate that we then use to parameterize the model (basal/progenitor layer half-life of ~ 30 days) (Figure 3.2E). While this niche driven behavior, incorrectly allows for some transient amplifier cell division above the basal layer these cells are inconsequential as they are lost in the stratum cornea (Figure 3.2). Overall population size (mean cell number of $\approx 120,000$ within a simulated $1mm^2$ biopsy) is maintained by balanced birth with death, where the top layers of the epidermis transition into the stratum cornea due to low diffusible concentrations. Hence, random stem cell survival and homeostasis results in a mosaic of subclones of different sizes and ages.

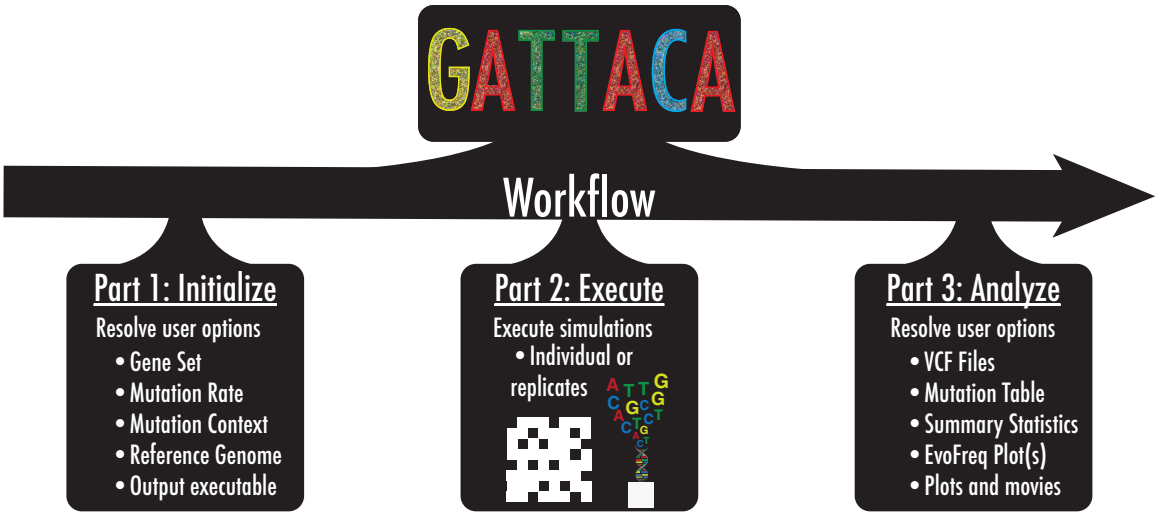


Figure 3.3: Gattaca is a three part workflow for simulating base pair resolution mutations within the human genome for somatic evolution *in silico* studies.

3.3.1 Gattaca: Mutation induction tracking

To incorporate patient mutational data, we developed the 'Gattaca method,' a standalone tool to be used alongside HAL [109] simulations to track base pair resolution mutation data that a user is interested in. Gattaca has been published in Molecular Biology and Evolution [191].Several studies have began to examine histologically normal human and murine tissue, showing an admixture of somatic mutations that can exist and expand to significant clonal areas [25, 26, 30, 31, 36]. These studies have largely focused on mutation characterization and have had limited tools to offer explanations for the dynamics driving observed evolutionary trajectories, with only few notable exceptions that have begun incorporating agent based models as a tool to explore somatic evolution in spatially constrained tissue [30, 90]. Due to this increasing body of work we set out to develop a method capable of incorporating base pair resolution information to compare to these NGS studies. The above mentioned studies and others that sought to incorporate representative genomes in their mechanistic models have been represented as simple counters or as binary arrays. These studies have lacked the ability to compare, at base pair resolution, the mutation spectra or utilize common tools designed for genotypic NGS data. Gattaca is the first to provide a means of tracking base pair resolution *single nucleotide variants* (SNVs) within an agent based modeling framework. This crucially provides an ability to accurately capture mutation data on a level comparable to sequencing experiments from the clinic or research settings.

3.3.1.1 Gattaca: Implementation and Case Studies

Gattaca is provided as an easily executed python script consisting of three parts: initialization, execution, and analysis (Figure 3.3). After the initial setup Gattaca produces a java file that allows for integration into an ABM. The only pre-requisite for Gattaca is an installation of snpEff [192], which provides the necessary base pair resolution reference genome and the tools to access it before Gattaca digests this information for downstream uses.

Part 1: Initialization

The setup resolves user inputs that includes mutation rates, mutation context probabilities, a gene set, and reference genome choice. Once resolved Gattaca extracts the gene locations from within the users reference genome, a *browser extinsible format* (BED) file is created that snpEff uses to extract bases for each gene. Gattaca then reads a provided mutational context file, in the event of none being provided a uniform probability is used. This file represents the probability of observing a given mutation from the 96 possible mutations within their trinucleotide contexts. Lastly, the mutation rates are scaled to the desired mean mutation rate. The mutation rates are adjusted from the gene specific mutation rates derived from a pan-cancer study [94]. This information is then prepared to generate a Gattaca java class tailored for execution within an ABM framework such as HAL [109].

The heart of Gattaca is its ability to track mutations within simulations at a base pair resolution. This requires a series of steps during each cell division where a user checks for mutation. The expected number of mutations per division is given for each gene (g_i) by the product of its individual mutation rate μ_{g_i} and its length L_{g_i}. Within each mutation check during division a Poisson distribution is used to determine the number of mutations accrued for each gene (X_g), so that $X_g \sim Poisson(\mu_g * L_g)$.

Determining the specific base that acquires a mutation is based on a multinomial of the 32 possible mutation positions based on trinucleotide contexts. This is drawn from a multinomial distribution based on the 32 possible positions. Once the trinucleotide is determined the base mutation is determined using the mutation context probabilities to determine the mutation type.

Part 2: Execution

Simulations utilizing Gattaca require the two files that are output by the Gattaca initialization step. These files, a java Gattaca class and a csv file with loci information,

will be placed within the scope of your executable HAL model [109]. Details on using HAL can be found at (http://halloworld.org). Once these are added to HAL, the Gattaca class will require initialization for a founding clone/population. Gattaca ties conveniently into the HAL phylogeny tracker requiring minimal additional computational overhead. Once Gattaca is initialized a function call to *_RunPossibleMutation* will be required during each division that will trigger the possibility of mutation upon division as outlined above. A detailed tutorial on integrating Gattaca and HAL can be found at https://github.com/MathOnco/Gattaca.

Out of the box, Gattaca provides its users with the necessary genetic underpinnings for a purely neutral simulation, e.g. mutations convey no functional advantage to a cell with any particular mutation. Should the user be interested in introducing functional heterogeneity based on mutations, the user would have to predetermine which mutation(s) will result in a functional difference. If this is tied to a specific mutation the user must simply check the Gattaca genome during simulations where the functional advantage would be conveyed for that cell. However, due to the stochastic nature of when and where in the genome the mutation will occur, it may be valuable to induce the mutation manually at a specific time to evaluate simulations with non-neutral dynamics. If a user wants to evaluate mutation number, say for a Muller's ratchet study [193, 194], they would simply check the number of acquired mutations for a cell and define its fitness advantage. However, mutation annotation is not completed until the Analysis section of Gattaca below, so any knowledge of mutation type or effect would have to be known a-priori. Further, a user could integrate a distribution of fitness effects using Gattaca as the underlying genome in the same way as the above two non-neutral examples. All of these genome queries/changes can be completed for any cell and/or genome during the simulation by accessing the genome class, details on mid-simulation genome access are provided in the documentation [191].

Part 3: Analysis

Once simulations are complete Gattaca introduces the appropriate noise for each mutation type, using either the Poisson or Gamma distribution approaches for deteremining total read counts (as described in Materials and Methods Section 2.5.1). Gattaca assumes that mutations are heterozygous and does not explicitly track *copy number alterations* (CNAs); however, the outputs of the Gattaca analysis pipeline provides users with the necessary numbers to examine the assumption of heterozygosity by *post-hoc* analysis of mutation data.

Once the variants are called based on the corrected VAF, variants are annotated with snpEFF and mutational position information is obtained. The user can output this information as a mutational table for every desired time point and every replicate simulation. As an additional output option users can also export variants from their simulations as a *variant call format* (VCF) file. This option allows for easy use in several bioinformatics downstream tools. Lastly, the execution of the analysis component of Gattaca provides several summary statistics for evolutionary dynamics, such as $1/f$ [82], first incomplete moment [31, 195], a Muller plot (using the EvoFreq [196] package), and a crude dN/dS measurement. We note that a true dN/dS would be expected to be the same across all simulations unless the user implements functional heterogeneity within their simulations based on a single, or collection of, point mutations.

3.3.1.2 Gattaca in the epidermis

Gattaca is the first base pair resolution mutation tracking tool for *in silico* studies. Due to this, it has been created as a standalone tool to use beyond the scope of the epidermis study presented within this chapter. To show Gattaca's broadly applicable use we constructed two case studies to demonstrate Gattaca's functionality. For the first case study we constructed two simple *agent based models* (ABMs) of cell turnover in three different dimensions, *zero-dimensional (non-spatial)* (0D), 2D, and 3D, to showcase and compare the mutational profiles and clonal dynamics that Gattaca allows its users to evaluate. This first case study reveals important considerations for model choice in a fully seeded model (such as the epidermis) versus a single cell seeded model. The second case study expands on this work to illustrate how wounding can have an effect on the measures of evolutionary dynamics. Both of these case studies have been incorporated into Appendix A case study 1 (A.1.1) and case study 2 (A.1.2).

Within the scope of the epidermis we incorporate 72 specific genes with gene specific mutation rates and profiles. These are embedded within each keratinocyte, accurately capturing the mutation distribution consistent with UV exposure. Each cell contains a base pair resolution genome of the 72 genes containing genomic positional information constructed from the reference $hg19$ genome. Each gene has a gene specific mutation rate such that the mean mutation rate (μ_g) is normalized to $\approx 3.2 * 10^{-9}\ bp^{-1}\ division^{-1}$ (likely a conservative estimate, but normal somatic mutation rates can vary wildly [24, 197–199]. This method allows for a more realistic mutation accrual, since a genome wide mutation rate would be biologically unrealistic due to a large number of co-variate factors (e.g. level of expression, location

on chromosome, and GC content, etc.) leading to differences in mutation rates [94]. Specifically, a mutation rate applied uniformly across the genome leads to the longest genes, such as ERBB4, PTPRT and BAI3 accumulating the most mutations (in our model) while NOTCH1-4 would acquire almost no mutations. This is not observed in normal epidermis, where NOTCH1 and a number of various other shorter genes are frequently mutated.

Upon initialization of the model we construct Poisson distributions for each gene given the mutation rate, μ_g, and the length of the gene, L_g, providing the expected number of mutations upon division. Where the number of mutations for each gene, X_g, is $X_g \sim Poisson(\mu_g L_g)$, and the specific base affected is determined by a multinomial distribution constructed for each A, C, T, and G allowing us to define and capture the specific mutational signature observed in the data. VAFs are calculated as described in the Materials and Methods Section 2.5.1 using a Gamma distribution for each individual patient's read depth distributions to define D_i for each variant and a 0.005 VAF cutoff.

3.3.2 Visualizing Evolutionary Dynamics in the Epidermis

Changes in genotype frequencies are often visualized using Muller plots, wherein each polygon represents a genotype (clone), and the thickness of the polygon indicates either the number of individuals with that genotype, or the frequency of the genotype in the total population at each time point. Nesting of genotypes represents evolutionary relationships, i.e. one genotype emerging from within another genotype's polygon indicates that the former was created by an individual with the latter genotype. Muller plots thus provide an excellent way to visualize how the genetic composition of a population changes over time. As these changes are governed by mutation, selection, and drift, the visualization of these dynamics facilitates an understanding of which forces are dominating the population.

One example where this form of data visualization has proven insightful is in understanding tumor evolution and subclonal compositions from both mechanistic models and bulk or multi-region sequencing. The hereditary nature of mutations in somatic tissue and cancers allows us to recapitulate, the temporal dynamics associated with mutational events through subclonal reconstruction. To date, two visualization packages exist in R to visualize this temporal data fishplot [200] and ggmuller [201]. For the purposes of understanding mutations within the homeostatic epidermis we constructed an alternative library that includes the features of these two libraries while expanding functionalities (see Table A.1). Notably, tying EvoFreq into com-

monly used subclonal reconstruction next generation sequencing tools. EvoFreq offers users ease of implementation and a way to generate many visualizations for assessing the temporal dynamics of clonal Evolutionary Frequencies. Evofreq is published in BMC Bioinformatics [196] and its implementation and functionalities are described in Appendix A.2.

3.4 Neutral evolution

Simulations of variable biopsy sizes (0.75,1.0, and $3.14mm^2$), age matched to comparable patients, reconstructs a log-linear subclone size-age distribution with neutral mutations that easily captures the extent of clone size variability within patients (Figures 3.4A and 3.4B). We see that small clones dominate the highest frequencies, while a heavy right tailed distribution is indicative of the clonal exponential size dependency (Figures 3.4A and 3.4B) consistent with patient data [31, 202, 203]. Most recently introduced mutations dominate the lowest frequencies and thus are difficult to observe [31]. By using a VAF cutoff of 0.005, the lowest VAF observed in the patient data [25], a large fraction of mutations in the simulated data are filtered out and analysis is conducted on what is likely to be observable within high resolution patient sequence data. For each patient and each biopsy we construct a *first incomplete moment* (FIM) distribution (Methods 2.4.1, Eq. 2.2). We then performed pairwise Kolmogrov Smirnov tests for each pair of these distributions (Figures A.6, A.7, A.8, A.9). We fail to reject the null hypothesis, that the distribution of clone sizes are from different distributions, for the vast majority of our simulation versus patient biopsy comparisons, but are able to reject the null for a select few comparisons (Figure 3.2B). However, for all patient biopsies at least one of the patient specific simulations fails to reject the null hypothesis. In a homeostatic epidermis, most subclones are rife with non-functional (i.e. neutral) mutations that are small, induced in the most recent past, and destined for extinction due to random stem cell turnover (Figures 3.2A, 3.2B, 3.5E, and A.5). Persistent, older subclones are rarer and larger (Figure 3D). Hence, the observed log-linear subclone mutation size distributions (Figure 3.2B) are an emergent property of the homeostatic constraints of normal human skin. These results reveal that much of the clinically observed mutation data largely reflects stochastic stem cell survival, where large subclones (e.g. greater mutation frequencies) represent the survival and expansion of older lineages.

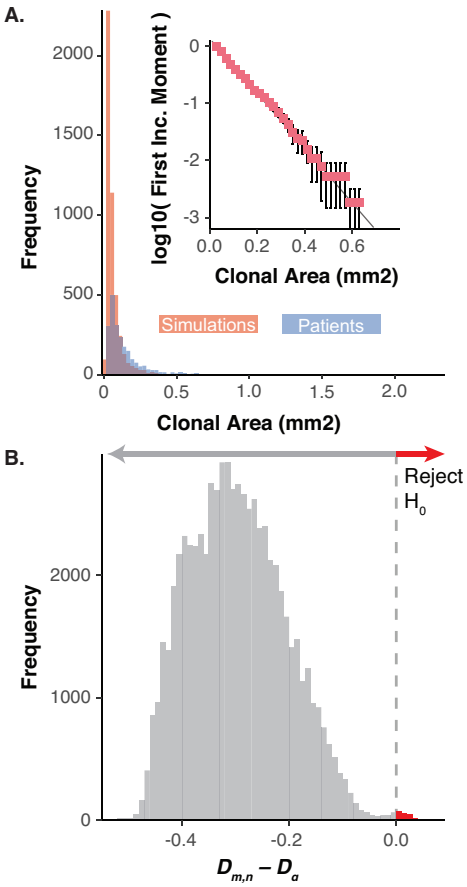


Figure 3.4: Neutral model dynamics from *three-dimensional* (3D) simulations of various sizes where (**A**) is the cumulative clonal area frequencies of all patient biopsies (blue) and randomly chosen simulations for each of the patient specific biopsies such that the number of simulations chosen is equal across the three simulated sizes and the sum of the number of biopsies equals that of the total biopsies for the patient (red). The inset figure shows the log-10 transformed *first incomplete moment* (FIM) for the same random sampling of patient comparable simulations for PD20399 (error bars denote standard deviation for 100 repeated samplings). (**B**) Shows the difference from the Komlogrov-Smirnov test statistic ($D_{m,n}$) and critical value (D_{α}) for all patient biopsy to patient specific model simulation's *first incomplete moment* (FIM) distributions for all simulation sizes. The red arrow denotes comparisons where the null hypothesis can be rejected. Individual patient test statistics are available in Appendix Figures A.6,A.7, A.8, A.9.

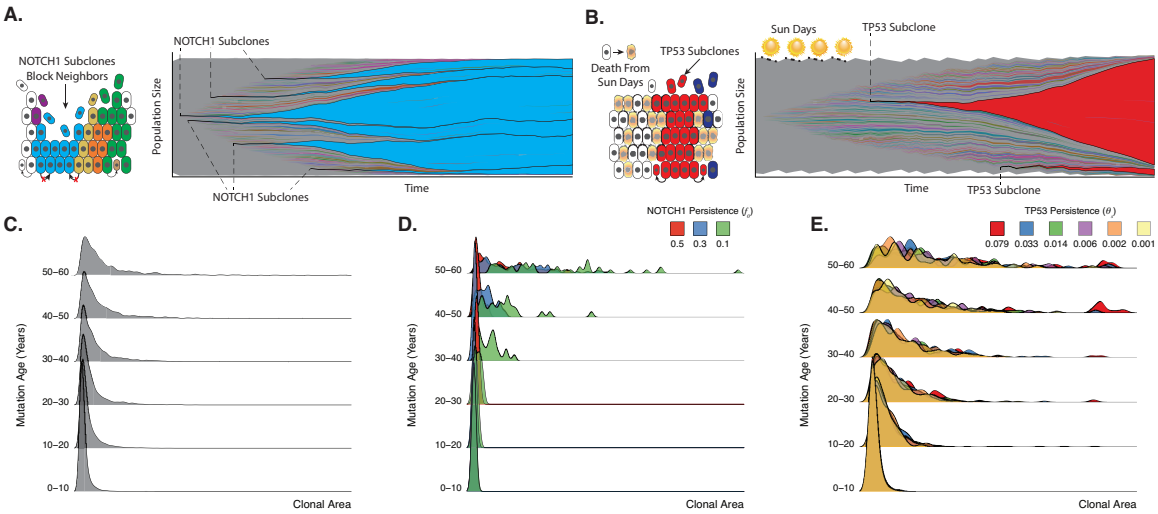


Figure 3.5: Clonal dynamics in homeostasis are a function of complex interactions between the microenvironment and the external environment. During homeostasis, every cell is equal relative to its neighbors, i.e. neutral. NOTCH1 disrupts neutral dynamics by offering a slight advantage via a blocking probability (f_0) which prevents neighboring cells from dividing into their positions (**A** schematic). Whereas, TP53 mutations aren't subject to *ultraviolet* (UV) damage, when a proportion (θ_s) of non-TP53 mutant cells may be killed by *ultraviolet* (UV) damage (**B** schematic). (**A**) and (**B**) next to each schematic, show clonal expansions for a single simulation for NOTCH1 advantaged clones, blue, and $TP53$ advantaged clones, red, in a $1mm_2$ simulation to 58 years, respectively. The age dependence is shown in the density plots broken by mutation age for 100 replicate simulations up to 58 years for neutral (**C**), NOTCH1, with varying strengths (f_0) (**D**), and $TP53$, with varying strength (θ_s) (E).

3.5 Modeling Fitness Advantage in NOTCH1 Mutant Keratinocytes.

Previously, NOTCH1 and TP53 mutations were observed to have larger subclone sizes relative to neutral mutations [25], indicating positive selection. Interestingly, age-related clonal size dependencies also exist even for advantageous driver mutations in the homeostatic simulations (Figures 3.5D-E). Unlike Lynch et al. [203] which lacks the homeostatic control in a 2D model of the epidermis, our findings implicate the constraints of the homeostatic tissue architecture as a primary factor in defining the age-related exponential size dependency of selective mutations and their subclones. To better define this selection within the context of homeostasis and without changing division rates, we simulated a decreased ability of NOTCH1 mutant subclones to be displaced by neighboring cells, as observed experimentally (Lowell, Jones et al., 2000). Mechanistically, we introduce functional change to the model when a NOTCH1 mutation occurs through a parameter (blocking probability, f_o) leading to increased persistence of NOTCH1 mutant cells over dividing non-mutant cells.

Specifically, upon induction of a mutation within NOTCH1, a fitness advantage is given to that cell and its progeny, if any. The fitness advantage gained through a NOTCH1 mutation is non-proliferative and allows the cell to remain within the basal layer longer than it would have otherwise. f_o is independent for each cell regardless of clonal lineage (e.g. cells with this advantage are non cooperative and can block one another). This indirectly impacts cell division at the basal layer, since successful division requires displacement of neighboring cells towards the corneal layer. We assessed a range of f_o values ($f_0 = [0.0, 1.0]$, 0.0 being no advantage and 1.0 never being displaced); at values close to one the tissue height is a quarter of its parameterized height (Figure 3.5).

The increased persistence of NOTCH1 mutated subclones increases both subclone sizes and frequencies, consistent with positive selection (Figure 3.5E). However, the requirement to maintain homeostasis indicates this mechanism comes with a cost, which inherently limits the strength of selection (Figures 3.5 and 3.6). The stronger the advantage, the greater the loss of homeostasis, creating a catch-22 where a NOTCH1 mutant clone with too much of a selective advantage destroys the epidermis, which is not observed in mice or human epidermis. Therefore, the simulations indicate a "homeostatic" limit to the strength of NOTCH1 selection such that young, rapidly expanding large NOTCH1 subclones are seldom observed (Figures 3.5 and 3.9). This is consistent with the observation that patient NOTCH1 mutations are smaller but

more frequent than TP53 mutations [25, 203]. Instead, similar to the neutral mutations, the largest NOTCH1 subclones are those with the oldest NOTCH1 mutations. Although driver mutations can only confer limited selective advantages without disrupting homeostasis, even small persistent advantages expressed at earlier patient ages pays dividends and inherently leads to larger and more frequent NOTCH1 subclones in adult skin (relative to neutral mutations).

3.6 Modeling TP53 mediated advantage in the UV exposed epidermis

TP53 mutations protect against cell death and sunlight enhances selection because TP53 subclones are smaller in non-sun exposed skin [90]. Normal sun-exposed skin does not exist because the extent of sun exposure varies massively between patients. Here we incorporate this important environmental variability through "sun days", where our homeostatic tissue takes a sun-filled vacation and a proportion of non-TP53 mutant cells are killed by *ultraviolet* (UV) irradiation (Figure 3.6B). Introducing functional heterogeneity by a TP53 advantaged clone does not involve a cell intrinsic parameter. Rather, UV damage kills cells on a given sun damage day, S, where the proportion of non-TP53 mutant cells killed is given by θ_s (Figure 3.6B). The TP53 mutant cells are subject to the same dynamics aside from this reprieve from UV induced death. We determined the parameters to use for the sun damage day set, defined by the number of days and the spacing of those days throughout the year, and θ_s, by approximating our 3D-HCA using an *ordinary differential equation* (ODE) that models keratinocyte population dynamics in a space limited condition. The total population size of a typical simulation over time for our homeostatic, spatial model can be defined by an ordinary differential equation model that describes population growth dynamics over time in a space limited condition. We use the classic

$$\frac{dN}{dt} = rN(1 - \frac{N}{K}) \tag{3.4}$$

logistic growth equation where N is the number of keratinocytes, r indicates a growth rate, K stands for a carrying capacity representing space limitation. Equation 3.4 allows us to approximate the same initial conditions as the spatial model (3D-HCA) by the same total number of cells in the ODE as the simulated domain. After initialization the population reduces to the carrying capacity over time. In order to incorporate death from UV-damage we introduce a new death component:

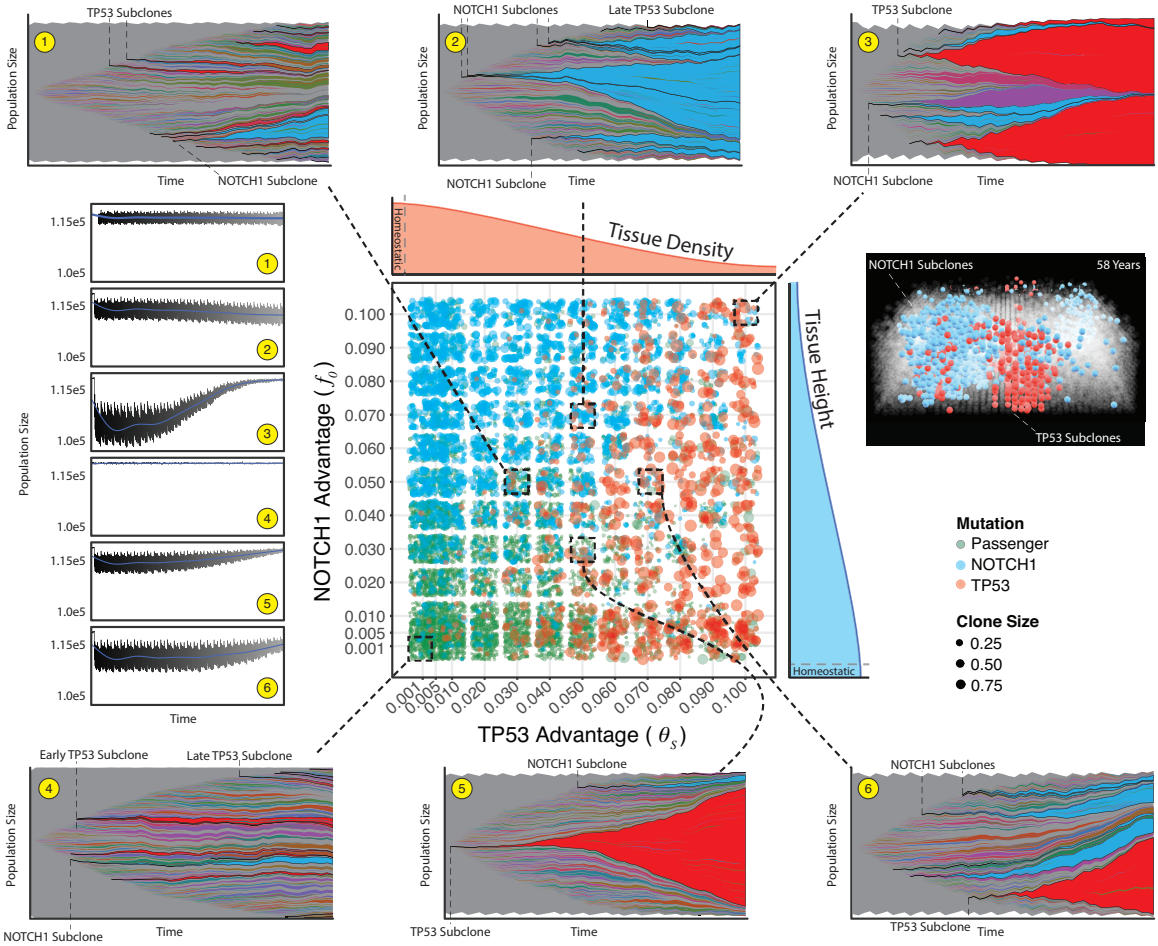


Figure 3.6: The primary, central, plot reveals clone sizes for $1mm^2$ simulations for four different ages with five replicates across a manually defined scale of parameter pairs for NOTCH1 (f_0) and TP53 (θ_s) advantages ($n = 2,880$ simulations). Density subplots show an idealized example of how each mutation type effects homeostasis (red and blue for TP53 and NOTCH1, respectively). The dashed boxes and lines show the evolutionary dynamics for a single simulation for a 58 year simulation (population frequencies greater than 0.001) for six parameter pairs and the corresponding overall population sizes over time ($1-6$ yellow circles, blue line is a locally weighted smoothing line). For a single parameter pair ($\theta_s = 0.03$, $f_0 = 0.05$) a $0.4mm^2$ tissue is shown where only NOTCH1 (blue) and TP53 (red) clones are colored.

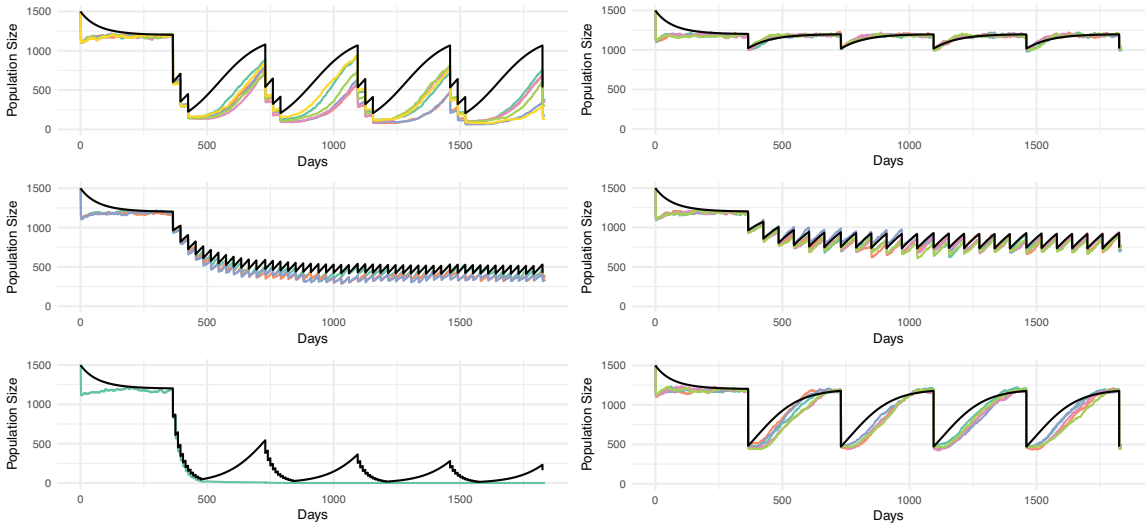


Figure 3.7: Here we define various sun day spacings, S, and the proportion of cells killed during sun days, θ_s. Black line in each plot shows the ODE logistic growth model and the colored lines are for the spatial model (Methods). The first three plots on the left show severe and infrequent or, simply frequent perturbations. Whereas, for the right three plots, we see less severe perturbations through parameter combinations.

$$\frac{dN}{dt} = rN(1 - \frac{N}{K}) - \theta_S \mathbb{1}_S(t)N, \tag{3.5}$$

where the proportion of cells, θ_S, at time, t, is killed by sun exposure. The indicator function defined as,

$$\mathbb{1}_S(t) = \begin{cases} 1, & \text{if } t \in S, \\ 0, & \text{otherwise,} \end{cases} \tag{3.6}$$

The set of sun days, $S = \{t, t \in \mathbb{Z}, t \leq 365\}$ because $S \subset Y$ (the total days in a year), where UV damage occurs depends on the number of days chosen to repeat each year and the spacing of those days. We see that at realistic values for θ_s the ODE model is able to approximate our homeostatic spatial simulations well, while a larger set of sun days and higher θ_s result in complete, or nearly complete, loss of tissue (Figure 3.7).

In addition, the spatial model (3D-HCA) follows a logistic growth after a delayed period. This delayed period is exacerbated by higher values of θ_s and/or by less spacing in the number of sun days. This is the result of cells settling upon large death, creating a sparse tissue (in the absence of a TP53 mutant), the cells settle, partially, towards the basal layer prior to re-establishing homeostasis (Figure 3.7 and

3.8).

Here we assess a range of values for θ_s and sun day sets, S. The number of sun days varies for each individual as does the number of spacings between sun exposure days. We attempt to address both of these variables with regard to homeostasis. We assessed a number of sun days throughout the year (7, 20, and 100) and the spacing of those sun days (1 day to evenly spaced throughout the year). We can see that the fewer the number of sun days the least amount of tissue is lost (Figures 3.8 and 3.8). Likewise, the more spaced out these sun days are the closer to homeostasis the tissue remains (Figure 3.8). This is true for both the ODE and spatial models.

We use this understanding of the growth behavior of TP53 mutant clones in the epidermis in our 3D-HCA parameters. Due to sun exposures varying within the lifetime of each individual we assume that a baseline UV exposure is taken into account for the parameterization of the model, encompassing the empirically derived turnover (birth/death) of cells within the basal layer. Thus, it is necessary to have a method of UV damage and subsequent TP53 mutant advantaged clones without a direct increase in proliferation rates. The number of sun damage days within a given year defines the sun day set (S) and can take on a number of values depending on how those sun damage days are spaced throughout the year. There are 365 possible sun damage days a year. It is unlikely that an individual would be subjected to all of these sun days, but there is a high degree of variability between individuals. This ODE approximation takes a relatively complex model, reduces its complexity, and spatial dynamics necessary to evaluate clonal dynamics in order to assess UV damage on the spatial model (Figures 3.7 and 3.8). This mechanism, unlike NOTCH1 does not intrinsically increase persistence but protects TP53 subclones from death after sun exposure, which allows rapid expansion as normal cells are eliminated when skin is exposed to more sun days (Figure 3E), consistent with mice studies that show short term expansions of TP53 mutant clones upon UV damage [90]. However, similar to NOTCH1, simulated TP53 subclonal expansions are also constrained by homeostasis, because too much simulated UV exposure causes a loss in tissue density and homeostasis.

Given the patient data (and normal skin in general) is most likely from patients with significant sun exposure, we consider the situation where both NOTCH1 and TP53 are found within the same simulated biopsy (Figures 3.6 and 3.9). In the absence of sunlight, TP53 mutations are subject to drift, along with all other mutations. Consistent with patient data we see evidence of selection with more frequent occurrences of NOTCH1 mutations at subclone sizes larger than neutral mutations.

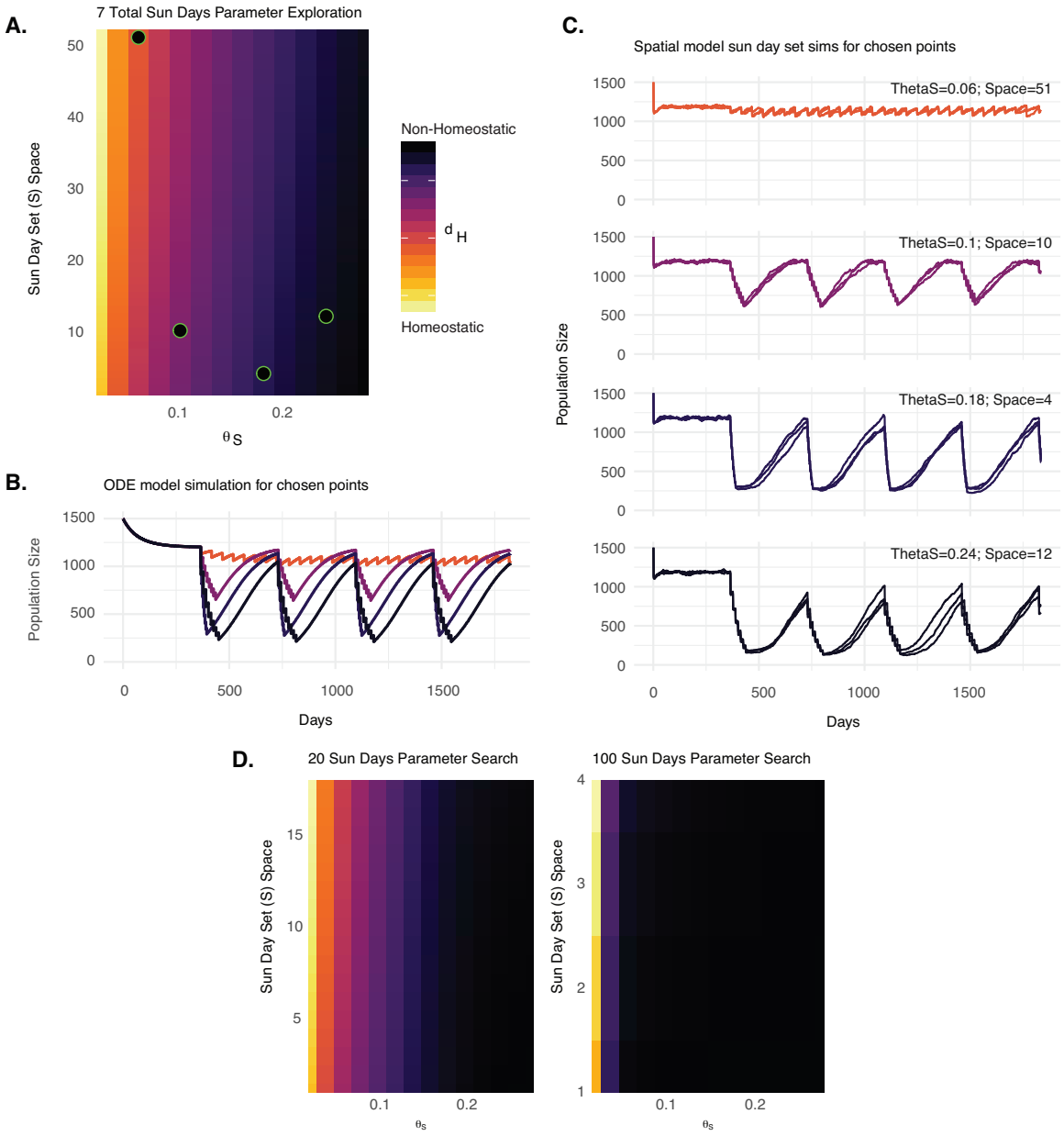


Figure 3.8: Evaluation of the number of sun days and θ_s values using the ODE model. In (**A**) we see the homeostatic measure d_H (Equation 2.3) as a response to sun day spacing and θ_s repeating yearly. For each black point with a green border the population sizes over time is given for the ODE (B) and the spatial 3D-HCA model (**C**). For (**B**) and (**C**) the colored lines are colored for where they fall within (**A**). (**D**) shows the same information, but for different numbers of sun days. The available spacing of these sun days for (**A**) and (**D**) differ due to how many days are available to space the sun days throughout the year.

Upon sun exposure, TP53 mutations are able to expand, even when they arise later in life, thereby leading to larger TP53 mutant subclones and shifting the age-dependent distribution to a more recent point in time. Passenger mutations are occasionally carried to larger clonal areas, but they are mainly relegated to the lowest frequencies (Figure 3.6). Interestingly, sunlight can reduce the selective effects of NOTCH1 mutations (i.e. reduce subclone sizes) because TP53 mutant subclones expand with UV irradiation (Figure 3.6). Even with the presence of sunlight and both mutations, we reveal an age-dependent log-linear relationship consistent with patient data where the earliest mutations expand to appreciable sizes within the tissue (Figure 3.9).

3.7 Discussion

3.7.1 Conclusions

Here we have presented the first data-driven *in silico* model able to capture the homeostatic cell dynamics of sun exposed human epidermis with high-resolution base pair tracking using the 'Gattaca' method [191] embedded within the HAL [109] framework. Our *in silico* epidermal model reveals that mutational persistence is a likely key to observed subclone frequency-size distributions. Strikingly, both neutral and driver mutations exhibit similar frequency-size distributions broadly consistent with neutral evolution, reflecting that random stem cell turnover is a salient feature of normal human epithelial homeostasis. Although mitotic stem cells accumulate mutations, random stem cell death with replacement typically leads to mutation flushing such that most *de novo* mutations are lost or are limited to small subclones. This work broadens our current understanding of selection and fitness acting in a homeostatic, normal tissue, where selection reflects persistence rather than selective sweeps, leading to larger but limited subclone sizes in human adult skin. Although the underlying skin model does not include all of the biological variables facilitating skin integrity, it illustrates that by explicitly considering the prime directive of normal tissue to maintain homeostasis as well as the effects of environmental damage, we can readily account for the major features of patient normal skin mutation data. The increasing availability of mutation data from normal human tissues, combined with new advances in lineage clone tracing *in vivo*, will further aid to parameterize our agent-based tissue simulations for future experimental validation and hypothesis testing.

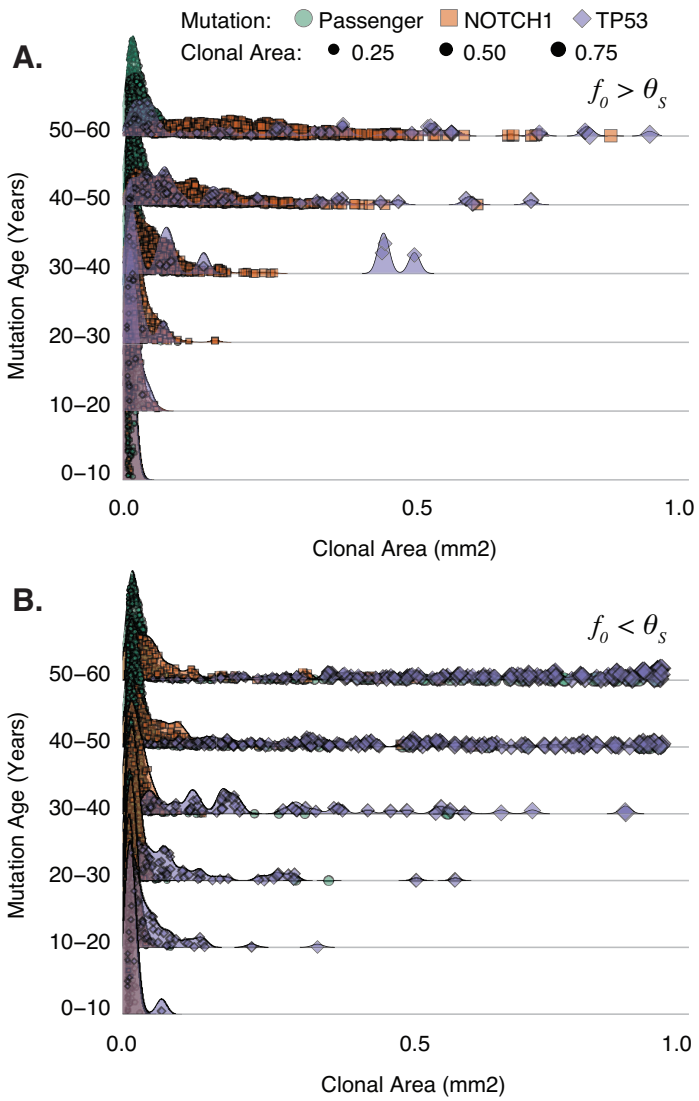


Figure 3.9: All 58 year $1mm^2$ paired parameter simulations from Figure 3.6 are displayed where TP53 persistence (θ_s)is greater than NOTCH1 persistence (f_0) (**A**) and $f_0 > \theta_s$ (**B**). Individual clones are broken into three groups (passenger, NOTCH1, and TP53) with corresponding sizes of each clone.

Chapter 4

Exploiting biological fidelity to understand homeostasis in the intestinal crypt

4.1 Background

Most molecular clocks use 'unidirectional' measurements that count the accumulation of changes since birth to infer the relatedness of lineages. The resolution at which a molecular clock can track clonal ancestry is a function of the rate at which genomic alterations accrue. A slow rate of alteration accrual can only reveal clonal dynamics occurring over long timescales. For example, genome sequencing studies of normal skin [25], blood [36], intestinal crypts [81], and endometrial glands [35] identified multiple subclones in each tissue, but in most cases the reconstructed lineages diverged many years in the past and recent cell turnover was not evident in the data. In comparison, a faster rate of alteration accrual has the potential to reveal rapid and/or recent clonal dynamics, but in practice, these approaches are compromised by 'saturation' wherein the same pattern of alterations evolve convergently in distinct clonal populations [88], and effectively recording stops in childhood.

In particular, somatic cell turnover is pervasive in mammalian tissues, but the dynamics of birth, death and replacement are difficult to measure. For example, in the intestine, small numbers of mitotic epithelial stem cells maintain intestinal crypts and undergo random turnover such that only one lineage persists. It takes several months in mice for one stem cell lineage to repopulate the entire crypt [45, 46]. The expansion and fixation of stem cell clones presumably recur throughout life, but most fate-marker methods can only record a single clonal dominance cycle.

Previous work has shown that DNA methylation at specific CpG loci can oscillate

back and forth under specific conditions [204, 205], including stem-cell exit from pluripotency [206, 207]. In these cases, CpG oscillations occur with a period of hours to days and are therefore less useful in timing replacement dynamics that occur over months and years.

Here, we introduce the concept of *fluctuating methylation clocks* (FMCs) whereby epigenomic alterations reversibly change state. We test the hypothesis that certain CpG sites stochastically and measurably fluctuate in their DNA methylation levels (specifically the fraction methylated, typically referred to as the beta value) between 0% (homozygously unmethylated CpG), 50% (heterozygous methylation), and 100% (homozygous methylation) in individual diploid cells (Figure 4.1A). When this fluctuation occurs at a timescale on the order of decades, we show that these *fluctuating CpG* (fCpG) sites can be used to infer recurrent dynamics of contemporary cell populations. Unlike traditional lineage tracing methods, which typically use a single molecular marker and thus rely on aggregating information over multiple patients to infer the average population dynamics (exemplified in Nicholson, A. M. et al. [208]), the presence of thousands of fluctuating methylation clocks enables individual clone-specific measurements to be made. This allows inter- and intra-patient heterogeneity of the stem cell dynamics to be directly probed.

4.1.1 Chapter aims

1. Develop a mechanistic spatial model of the homeostatic crypt complete with epigenomes that can incorporate fCpGs.
2. Resolve method for identification of candidate fCpG sites within patient Colon and Small Intestine crypts.
3. Develop a mathematical inference methodology to extract ancestral information encoded within fCpG loci.
4. Validate inference methodology using spatial model of the homeostatic crypt and apply this methodology to patient data.
5. Orthogonally validate FMC methodology using Endometrial glands.

4.2 The exclamatory bowel model

A crypt ignoring villi in the small intestine, forms a cylindrical geometry with stem cells at the base and a crypt wall moving up the crypt. Here we have developed an

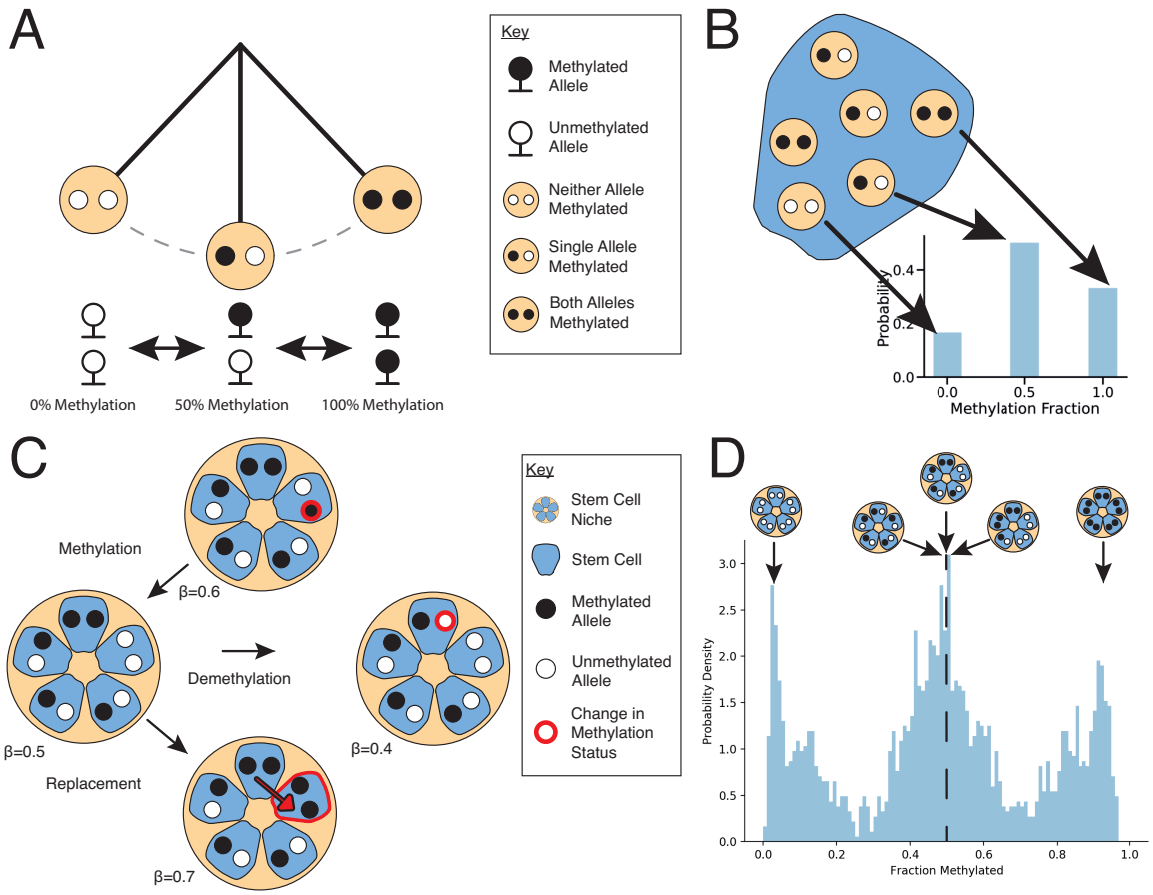


Figure 4.1: Fluctuating methylation status as a lineage tracing marker. (**A**) Illustration of the three possible methylation states at a specific CpG locus within a particular cell. A cell can either be homozygously (de)methylated or heterozygously methylated at that CpG locus. It is the spontaneous transitions between these states that allow methylation to act as a lineage tracing marker. (**B**), Illustration of the link between the methylation status of a given CpG locus within a particular cell and the β value (the fraction of methylated DNA at that locus) associated with that cell. (**C**), Graphical representation of how the methylation status in a small population of five stem cells at a particular CpG locus can change over time due to (1) methylation, (2) demethylation or (3) cell replacement. (**D**), Methylation (β) distributions from an individual crypt; the peaks near 0% and 100% correspond to a clonal methylated or unmethylated CpG locus, respectively, whereas the peak at 50% corresponds both to clonal heterozygous CpG loci and subclonal populations caught mid-sweep.

off-lattice mechanistic agent-based model of the human crypt using the HAL modeling framework [109] capable of representing a crypt of the small intestine or colon. The cylindrical unit is separated into two compartments, the stem cell compartment represented as a pool at the base of the crypt and then the wall of the crypt where transient amplifying cells are pushed upwards until they are removed from the top of the crypt (Graphical summary of the model can be found in Figure 4.2).

4.2.1 Crypt dynamics of the exclamatory bowel

The spatial model of the crypt, termed *exclamatory bowel* (EB) because of its resemblance to an exclamation mark, is dynamic in the sense that the x and y dimensions are calculated using the total populations size (N_T) and the stem cell pool radius (ψ). The x dimension is defined as $x = 2\pi\psi$. The center of the stem cell pool is placed such that the origin of the center of this circular stem cell pool whose size, and thus number of stem cells allowed within this pool, is placed at (h, k) where $h = x/2$ and $k = \psi + 5$. Division for each stem cell, is defined by ρ_c, which is randomly assigned as the hourly cell cycle defined by $\rho_c \sim U(\rho_{min}, \rho_{max})$ where ρ_{min} and ρ_{max} are $\rho \pm 4$ hours (Table B.1).

As a cell approaches ρ_c the cells diameter doubles for the five hours/timesteps preceding the cells division. Upon division, both daughter cells occupy this space. When a stem cell, defined by $d(x_c, y_c) \leq \psi$ where $d(x_c, y_c) = \sqrt{(x_c - h)^2 + (y_c - k)^2}$, divides the daughter cells can be placed in any arrangement around the parent cells x_c and y_c position, differentiated cells can only be placed vertically (i.e. the x_c values are equal). The base of the crypt wall is set just above the origin of the stem cell pool plus ψ and a small offset to provide space so that no cell forces interact between the stem cell pool and the base of a stem cell wall. If $d(x_c, y_c) < \psi$ then the cell is moved to the base of the stem cell wall where the cells new position (x_2, y_2) is given as y_2 and x_2 is given by the cells exit radians, rad_s, given by $atan2(y_c, x_c)$ so that the cells position along the x dimension is $x_2 = (rad_s + \pi)(\frac{x}{2\pi})$. Boundary conditions for the cells within the crypt wall are periodic (i.e. allowed to wrap around) and no-flux at the top and bottom of the crypt (i.e. no cell can breach these boundaries). A run step in the model is hourly and updates to cell positions occur for the whole crypt and applied at each run step.

In order to incorporate a comparable measure to patient data we must integrate an *in silico* CpG loci. These fCpG are incorporated into the exclamatory bowel initially as $1,000$ fCpG loci (later refined to exactly 1794 fCpGs) on a paternal and maternal copy of the chromosome providing a binary representation of methylation. By using

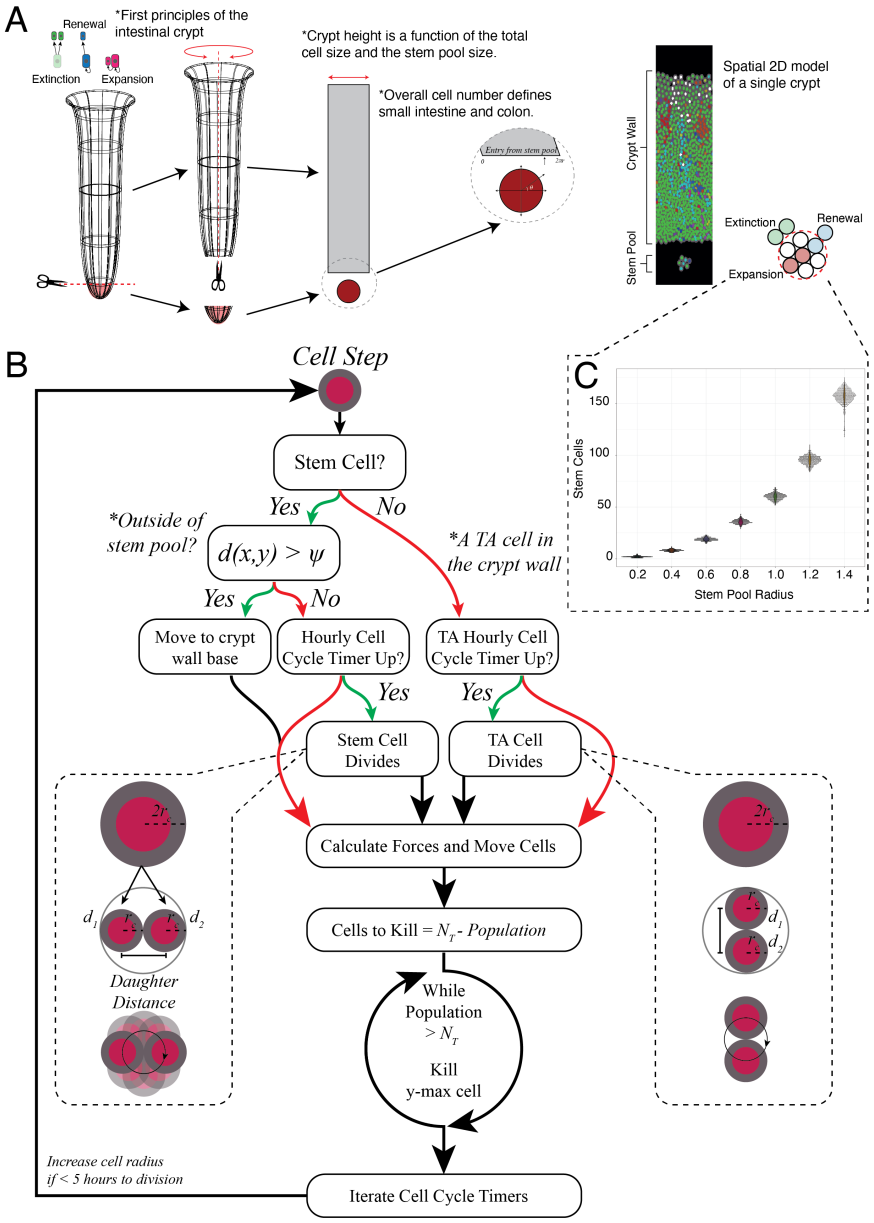


Figure 4.2: Exclamatory Bowel Model. **A** provides a conceptualization of the *exclamatory bowel* (EB) model (i.e. a single crypt) within the simulated domain. Stem cells are governed by first principles of renewal, extinction, and expansion; while the height of the crypt is governed by the total population size. **B** Model step flowchart for the EB. First a check is completed to determine if a cell is still a stem cell, moving it to the base of the crypt wall if it falls outside of the defined stem pool size. Divisions are allowed to occur in any direction for a stem cell and only above or below for a differentiated cell. The total population size is maintained by removing cells at the uppermost crypt wall position until the population is back to the homeostatic population size. Inset in the upper right corner shows an example of the exclamatory bowel model with the emergent stem cell dynamics driving extinction, renewal, and expansions. **C** illustrates how dynamic the stem pool is based on the defined stem pool radius. The increased size of the stem cell pool causes a sharp increase in the number of stem cells.

what we term a mitotic array we allow for accurate *in silico* chromosomal segregation to ensure that each cell gets a copy of a paternal and maternal chromosome during meta- and anaphase. Methylation changes occur within the array as the strand is copied into the daughter cell. These loci can switch methylation status at a rate defined by ω upon division. Here we ignore varying forms of methylation as the method of detection is unable to differentiate between (or detect) 5-*methylctyosine* (5-mC), 5-*hydroxymethylctyosine* (5-hmC), 5-*formylcytosine* (5-fC), or 5-*carboxycytosine* (5-caC) differences in methylation types or determine hemi-hemimethylation status.

At each hourly time step we assume that the forces acting on each individual cell to be at equilibrium, $F_{c_i} = 0$, where F_{c_i} is equal to the contact force between cell i and its neighbors. For two cells whose radius is R_i and R_j, respectively, their contact force between them is based on a linear spring constant model (Hooke's law) and is calculated as

$$F_{c_{ij}} = \begin{cases} k_i \frac{\Delta R_{ij}}{R_i + R_j} & \text{if } \frac{\Delta R_{ij}}{R_i + R_j} > 0 \\ 0 & \text{if } \frac{\Delta R_{ij}}{R_i + R_j} < 0 \end{cases} \tag{4.1}$$

Assuming that each cell has the same spring constant k, the overlap of cells ($\frac{\Delta R_{ij}}{R_i + R_j}$), and the overall number of cells in contact with any given cell (n_i) gives the velocity for an individual cell is $v_i = k \sum_{j=1}^{n_j} \frac{\Delta R_{ij}}{R_i + R_j}$. The modeling framework can be obtained from `https://github.com/MathOnco/flipflopspatialmodel.git`[209].

4.2.1.1 Evaluating an alternative spatial crypt model

A typical representations of the intestinal stem cell pool is a ring where cells are not capable of moving freely except during divisions where expansions, renewals, and extinctions can take place. This configuration neglects the contributions of the crypt wall, but is useful to understand stem cell dynamics. Notably, stem cell replacement rates and time to monoclonal conversion (fixation). In order to evaluate whether the exclamatory bowel differs in dynamics from this typical model, especially as it relates to our representation of methylation, we implemented a version of the exclamatory bowel where stem cells were restricted to a ring structure and immobile. This additional crypt model is termed the *exclamatory ring* (ER), as it's simply swapping out the stem pool for a ring with two big parameter differences. The stem cell pool radius, ψ, is replaced by defining a specific number of stem cells, S. These stem cell positions are arranged equally around the center point of the stem pool in the spatial model and restricted to that position until a division event occurs. Upon division a cell has equal probability of expansion and renewal (Figure 4.3A). If a cell expands

both daughter cells become stem cells on the ring and a neighboring stem cell is displaced (equal chance for cells on the left and right of the dividing cell) and moved to the base of the crypt wall. If a stem cell renews one daughter is moved to the stem cell wall. In this model, extinction events happen by cells displacing their neighbors through division explicitly. The exclamatory bowel model has this same process, but the off lattice movements can cause more distant stem cells to go extinct as they fall out of the stem pool (Figure 4.3A).

4.3 Fluctuating CpG theory explained by the exclamatory bowel

We set out to understand the dynamics of CpG sites fluctuating within a crypt, hypothesizing that stem cell dynamics can be uncovered by exploiting (de)methylation associated with CpG sites fluctuating with a 'just-right' rate. Using the EB model we analyzed differences in (de)methylation rates for total population size (i.e. small intestine and colon), stem cell numbers, and division times. For the EB we did a parameter sweep by discretizing a grid space of parameters such that $\theta_i = (N_T, \rho_c, \psi, \omega)$ (Table B.1). The total grid space of θ_i is 288 parameter combinations with 20 replicates for each. First, the EB methylation outputs were examined for the entire crypt, as the individual stem cell compartment can not be directly pulled out of patient data. The (de)methylation rate (i.e. stem cell sweeps through the stem pool population) proved to be the most important parameter for identifying the FMC signal (Figures 4.3B, B.3A and B.3C). Only subtle differences exist in fCpG beta distributions within a single (de)methylation rate across the additional parameters. There is a relatively large range over which the largest variances of fCpG beta distributions can be found that results in three persistent peaks (Figures 4.3 and B.4). This analysis allows us to maximize the information that can be gained from the fCpG beta distribution with little *a posteriori* knowledge of filtering criteria to define the 'just-right' rate.

For the crypt FMC, this 'just-right' rate is indicated by a 'W' shaped distribution observed in the individual crypt simulation outputs and patient data (Figures 4.3B 'Just right', B.3, and B.4). Beta values from the exclamatory bowel reveal major peaks at 0%, 50%, and 100%, along with additional minor peaks near ∼10-40% and ∼60-90% which are due to recent (de)methylation events that had expanded to some, but not all, crypt cells (subclonal (de)methylation events). This can be deduced by simulating different stem cell numbers using the exclamatory bowel model, examining the fCpG distributions for the whole crypt, the crypt wall, and the stem cell

compartment alone for a given 'just-right' de(methylation) rate (Figure 4.3C). Indeed, when we tease apart these different compartments we observe that the underlying distribution of beta peaks correlates with the number of stem cells, S, such that the number of beta peaks is approximately equal to $2S + 1$ (Figure 4.3D). These peaks follow a separation of approximately $\frac{1}{2S}$. Hence, the positions of these subclonal peaks hold information on the number of stem cells within the niche. Similarly, the number of fCpG loci in each peak ought to contain information regarding the (de)methylation and replacement rates.

At low (de)methylation rates (where the clock fluctuated too slowly), the methylation distribution is essentially concentrated near 0% and 100% methylated, with a small minority of fCpG loci in the intermediate 50% methylation state, mainly due to clonal heterogeneous methylation. This is because at such a low (de)methylation rate, very few of the fCpG sites had changed their methylation status even once, and further, (de)methylation events that could distinguish a subclone are unlikely to occur. If this too slow system is left until it has converged to the steady state, the distribution exhibits three sharp peaks near 0%, 50% and 100%, with the 50% peak containing approximately twice as many fCpG loci as the 0% or 100% peaks (due to the multiplicity of the clonal heterogeneous state) (Figures 4.3B 'Too slow', B.3, and B.4).

Conversely, at high (de)methylation rates (where the FMCs fluctuate too fast) the methylation distribution approaches a binomial-like distribution centered at 50% (Figure 4.3B 'Too fast', B.3, and B.4). The intuition behind this behavior is that when the (de)methylation rate is very fast, the record of clonal dynamics caused by the stem cell replacement process changing methylation allele frequency is immediately lost; hence, the system is effectively equivalent to $2S$ independent binary oscillators, with a probability of a given fCpG being in the methylated state equal to 0.5.

4.4 Patient data and fluctuating CpG sites (fCpG)

We isolated DNA from individual single colon or small intestinal crypts (31 colon samples originating from eight individuals and 28 small intestinal samples originating from seven individuals from colonoscopies or flexible sigmoidoscopies; Supplementary Table 2.1) and measured DNA methylation in each crypt using Illumina EPIC arrays (Methods 2.2.1). Samples from each tissue were treated separately to account for tissue-specific differences in the DNA (de)methylation processes.

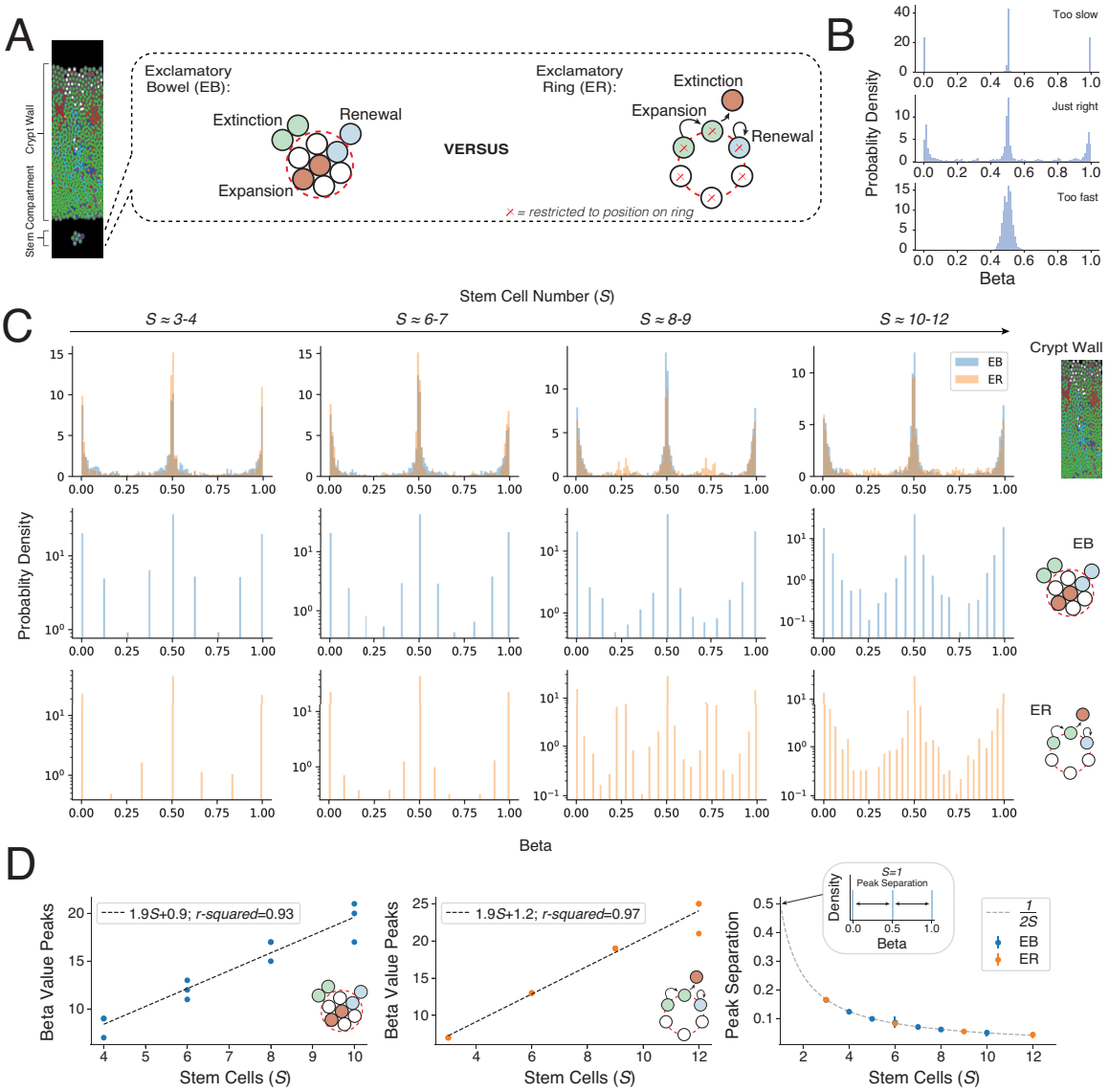


Figure 4.3: Fluctuating CpG distributions in the exclamatory bowel/ring. **A** provides a graphical comparison of the *exclamatory bowel* (EB) and *exclamatory ring* (ER) models. In both, the stem cell dynamics of extinction, renewal, and expansions are driven by division events, but their configurations alter these dynamics by explicitly restricting them in the ER model. Histograms in **B** show the too slow, just right, and too fast *fluctuating CpG* (fCpG) distributions for EB 20 year simulations with 4 stem cells at 0.0005, 0.05, and 0.5 (de)methylation rates, respectively. **C** shows beta distributions of the simulated fCpG sites are consistent with the characteristic 'W' shape in both the crypt wall (top row) and stem cell compartment across all spatial models (middle row = EB, bottom row = ER). Here, we illustrate the distinctive peaks observed in the stem cell pool separate from the crypt wall. Each row is a different number of average stem cells. **D** shows the correlation ($\approx 2S + 1$ peaks) of stem cells and beta value peaks where major peaks exist at 0%, 50%, and 100%, minor peaks are present between these due to recent (de)methylation events with a separtion of $\approx \frac{1}{2S}$. This observation persists regardless of model type (EB on left, ER on right, and Eq. B.4).

4.4.1 Early attempts to define fCpG filtering criteria

To select fCpG sites, we identified CpG loci unlikely to be actively regulated and that exhibited a high degree of intraindividual heterogeneity (Figure 4.4A and Methods B.1.1). This process identified 7,073 putative fCpGs within the colon sample cohort and 8,828 fCpGs within the small intestine cohort, of which 1,794 fCpGs were shared between tissue types (Figure B.2A). There was a good correlation ($R^2 = 0.62$) in the heterogeneity scores between colon and small intestine samples (Figure B.1B), and fCpGs loci that were exclusive to the colon had a substantially higher average variability score in the small intestine than all CpG loci (Figure B.1C), suggesting that the relatively large number of non-overlapping loci was due to the arbitrary strictness of our identification procedure (Methods B.1.1). Further analysis was performed on these shared 1,794 fCpG loci[210] to aid the generalizability of our approach. Methylation of the 1,794 fCpG loci when averaged across the 31 colon crypts was $\sim$ 50%, consistent with uncorrelated stochastic (de)methylation at fCpG sites occurring independently in each crypt, but in individual crypts, we observed a characteristic trimodal 'W-shaped' distribution of methylation values that likely was similar to the methylation pattern of the most recent common ancestor of the crypt population (Figure 4.1D).

4.4.2 fCpG loci are enriched in minimally expressed genes

For CpG loci to act as a molecular clock, the loci should not be subject to strong evolutionary selection or cell-specific regulation. We compared the proportion of the 1,794 fCpG sites that were associated with a specific gene to the 428,511 CpG sites that were not identified as fluctuating (Methods). fCpG loci were strongly enriched for non-genic CpG sites (Figure 4.4B; *odds ratio* (OR) = 1.8, chi-squared test, $P < 0.001$). We tested RNA expression using 40 normal colon samples from TCGA [114] and found that the mean expression of genes associated with fCpG loci was lower than that of genes associated with the non-fCpG loci (Figure 4.4B; -0.24 Cohen's d calculated for log-transformed expression, Welch's unequal variance t-test, $P < 0.001$). Furthermore, none of the genes that had fCpG loci in their promoter region had intermediate (10 *transcripts per million* (TPM)) or greater expression in single-cell RNA sequencing data of normal colon (Figure B.2A). Together, these analyses were suggestive that methylation at the fCpG sites was unlikely to be under strict regulation or evolutionary selection in the colon.

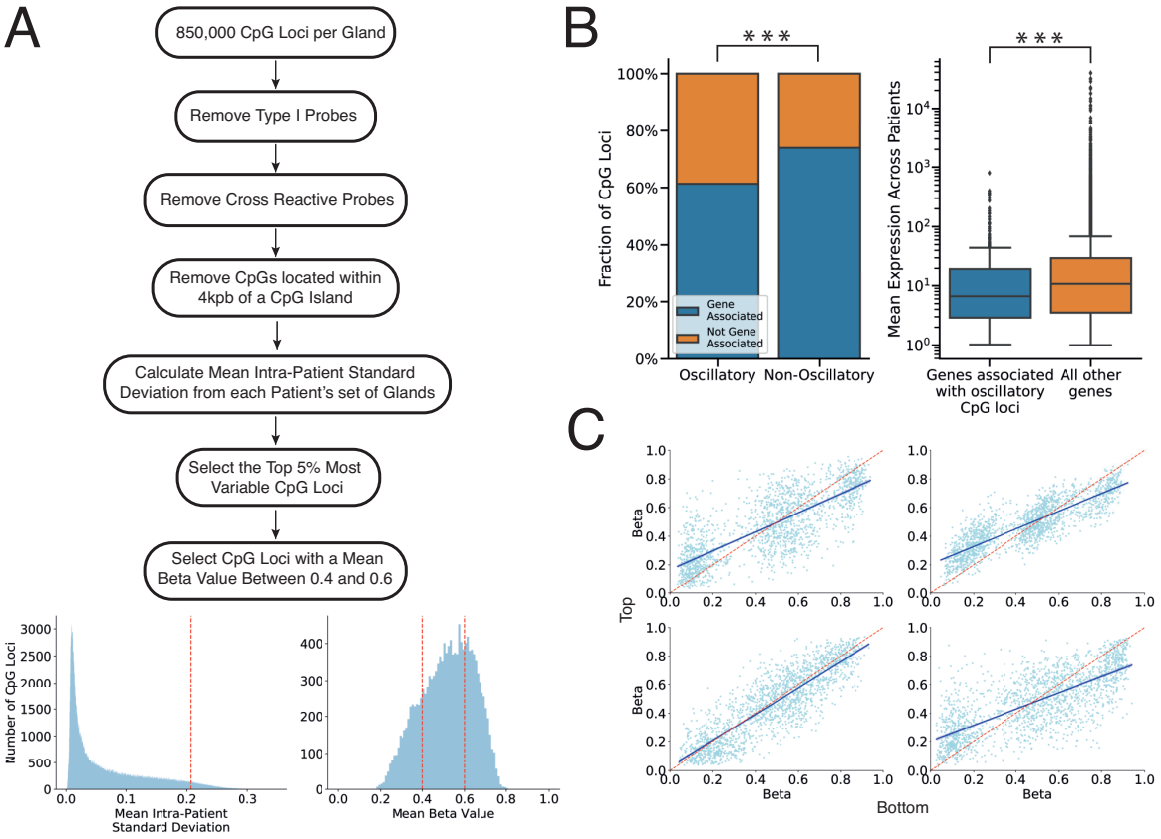


Figure 4.4: Identification of selectively neutral fluctuating CpG loci. (**A**) Workflow used to identify *fluctuating CpG* (fCpG) loci that exhibit high intraindividual heterogeneity. Input data were the 850,000 CpG loci assayed by an Illumina EPIC array. We removed type I probes and probes that cross-hybridize highly homologous DNA regions. For each CpG locus, we calculated the standard deviation for each set of approximately four crypts per individual and then calculated the mean standard deviation across the cohort as a metric for the intraindividual heterogeneity. We selected the top 5% most highly variable CpG loci and then removed CpG loci that have a mean β value (across the entire cohort) less than 0.4 or greater than 0.6; kb, kilobases. (**B**), Left: fCpGs are enriched for CpG loci not associated with any genes ($P = 6.5 * 10^{-34}$, chi-squared test). Right: the set of genes associated with fCpG loci exhibit lower average RNA expression ($P = 6.6 * 10^{-6}$, two-sided Welch's t-test performed following the log-transformed data) in normal colon than those genes associated with non-fCpG loci (center line, median; box limits, upper and lower quartiles; whiskers, 1.5 interquartile range). $***P < 0.001$. (**C**), β values of fCpG loci are correlated between the bottom and top halves of a crypt.

4.4.3 Methylation status of fCpGs is conserved along the crypt

Methylation status of fCpGs is conserved along the crypt. Previous research has found that the methylation profile of the whole crypt is typically representative of the stem cell population at the base of the crypt [211]. To ensure that this was the case specifically for the fCpG loci identified above, we split seven crypts into their respective tops and bottoms (Figure B.2C) and ran Illumina EPIC arrays on both halves using the same protocol described previously. Due to the low input DNA amounts, three of the samples failed the quality control step. The remaining four crypts exhibited a good correlation between the β values of the fCpG loci in the tops and bottoms of the crypts (Fig. 4.4C; $R^2 > 0.6$, $P < 0.001$ in all cases), and binarizing the CpG calls ($\beta < 0.2$ encoded as 0 versus $\beta > 0.8$ encoded as 1) showed only 2/988 (0.2%) of fCpGs changed methylation status between crypt base and top (Figure B.2D).

4.5 Mathematical model of the methylation distribution of fCpGs.

We hypothesized that the precise shape of the methylation β value distribution for fCpGs was determined by the underlying dynamics of cell evolution. To test this hypothesis in the context of intestinal crypts, we developed a mathematical model and associated Bayesian inference framework to relate the competitive dynamics of stem cells within their crypt to the measured distribution of fCpG methylation.

The mathematical model consisted of a hidden Markov model that described the time-dependent probability distribution of the number of methylated and unmethylated copies of a single CpG locus within a stem cell niche of fixed size S. We considered three possible processes that changed the methylation status at a given CpG locus: (1) spontaneous methylation (at constant rate μ per allele per stem cell per year), (2) spontaneous demethylation (constant rate γ per allele per stem cell per year) and (3) one stem cell replacing another stem cell (constant rate λ per stem cell per year) (Figure 4.1C). We further assumed that the stem cells could be treated as a well-mixed population such that each stem cell could replace any other stem cell within the niche with equal probability. The system could be fully characterized with just two state variables: k, which represents the number of stem cells in the crypt with one allele methylated, and m, which represents the number of stem cells with

both alleles methylated. By considering the possible $(k, m) \rightarrow (k', m')$ transitions, we derived a system of ordinary differential equations describing how the probability $(P(k, m|\lambda, \mu, \gamma; t))$ of the system being in state (k, m) changes over time (Appendix B.1.3). For a pool of S stem cells, there are $2S + 1$ discrete states the niche methylation level could take, with a β value of $\frac{z}{2S}$ (for $z \in [0, 2S]$). To link the probability that a particular CpG locus has a population methylation status z to the output of our stem cell dynamics model, we marginalized over the various combinations of k and m that correspond to a particular z value.

We developed a Bayesian inference framework (Appendix B.1.6) [212] that allowed for simultaneous inference of the number of stem cells (S), the replacement rate per stem cell (λ), and the methylation (μ) and demethylation (γ) rates per stem cell per allele per year for an individual gland. Thus, we could fit our model of stem cell dynamics to the data from individual crypts, allowing us to probe tissue-specific stem cell dynamics while accounting for intra- and interindividual heterogeneity.

4.6 Stem cell dynamics are inferred with high accuracy *in silico*

To verify that our Bayesian inference framework was able to accurately infer the stem cell dynamics of a crypt from FMC patterns, we first used the non-spatial stem cell model (Section 4.5) to generate three 'synthetic' crypts each containing five stem cells, a mean replacement rate of 1 per stem cell per year and a de novo (de)methylation event rate of 0.0005, 0.05 or 0.5 per allele per stem cell per year (Figure 4.5A) and used our inference framework to attempt to recover the (known) underlying parameter values from the simulated methylation distributions. With the knowledge gained surrounding the theory of FMCs from the EB spatial model we expected that the inference framework would be unable determine the posterior for the number of stem cells unless fluctuating 'just right,' (i.e. ~ 0.05).

As expected, the Bayesian inference could not satisfactorily determine the posterior for the number of stem cells for the too slow crypt, as there were too few fCpG sites with intermediate values that held information on stem cell number. By contrast, the inference framework accurately recovered the number of stem cells for the too fast crypt, as subclonal methylation events were abundant, but the replacement rate could not be inferred accurately. This was because the clonal information that is propagated by stem cell replacement (increase/decrease in β values from the expanding clone) was almost immediately lost due to the high (de)methylation rate.

When the simulated (de)methylation rate was just right, the model was accurately able to recover all known parameter values with good confidence (Figure 4.5 B and C). We note that this *in silico* analysis shows that we were able to confidently confirm that the (de)methylation rate for a given set of CpG loci is within the just-right range by the presence of the characteristic 'W' shape. Note that the range of the methylation error rates that give rise to the W shape and which are suitable for timing using our analysis is relatively broad, covering over two orders of magnitude. Despite the apparent correlations between the rate parameters, the parameters were

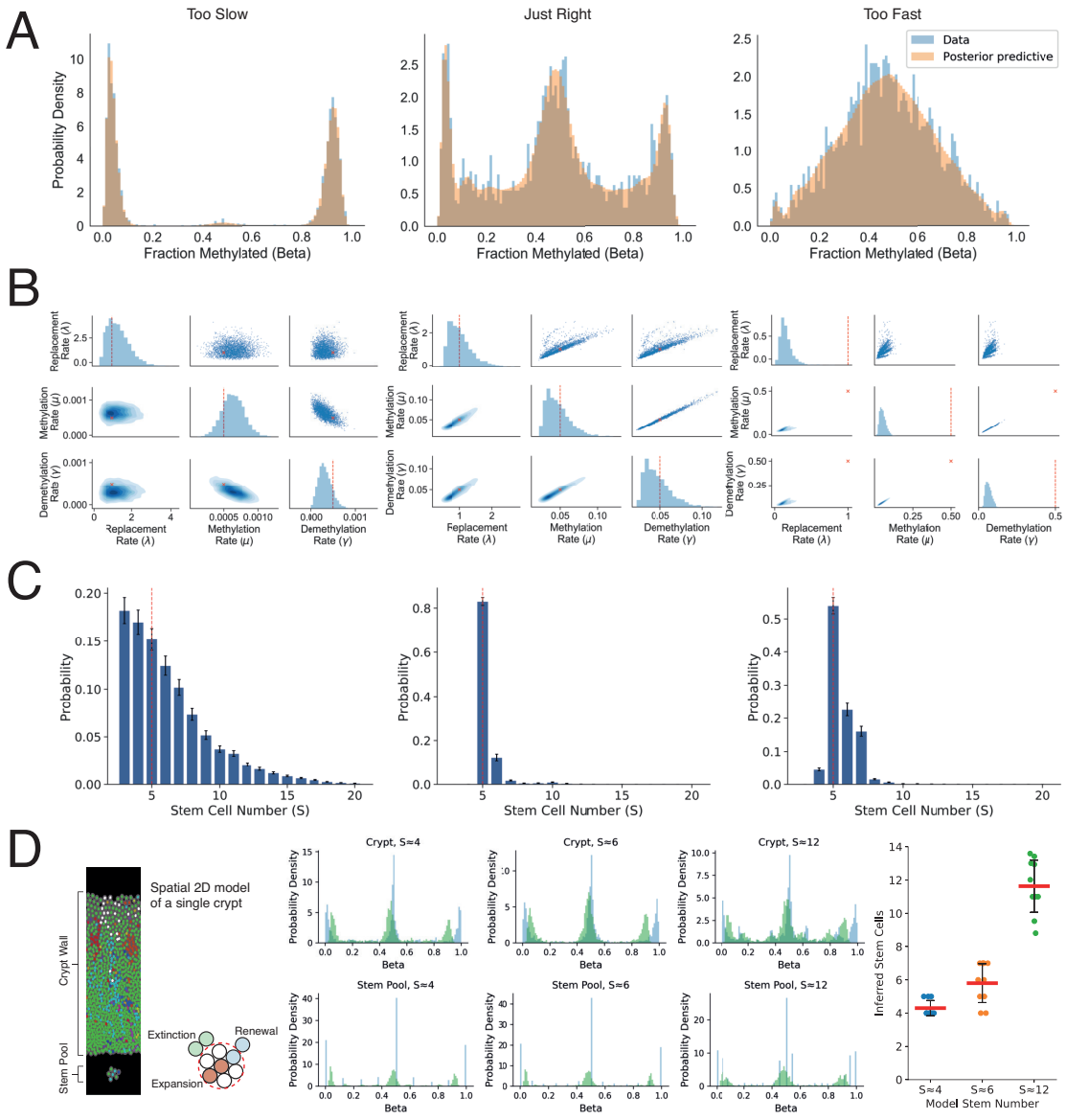


Figure 4.5: W-shaped methylation distributions are indicative of clonal dynamcis.(Caption next page.)

Figure 4.5: W-shaped methylation distributions are indicative of clonal dynamics. (Figure previous page.) *In silico* evaluation of the accuracy of Bayesian inference on stem cell number (S), replacement rate (λ) and (de)methylation rates (μ,γ) as a function of input (de)methylation rates. Three regimens were evaluated: $\mu = \gamma = 0.0005$ methylation events per allele per stem cell per year ('too slow'), very high methylation rates ($\mu = \gamma = 0.5$ per allele per stem cell per year ('too fast')) and intermediate methylation rates ($\mu = \gamma = 0.05$ per allele per stem cell per year ('just right')). (**A**) Simulated *fluctuating CpG* (fCpG) methylation distributions from individual crypts at each of three input (de)methylation rates. The characteristic W distribution is only evident for the just-right (de)methylation rate. (**B**) Posterior distributions of inferred replacement and (de)methylation rates for each input (de)methylation rate. (**C**) Posterior distributions of inferred stem cell number. The stem cell number posterior mean was calculated by taking the softmax of the log evidence, while the error bars were calculated from the estimated error (1 s.d.) on the log evidence. In (**B**) and (**C**) red dashed lines indicate the true (inputted) value of the parameter. The simulated datasets each contained $S = 5$ stem cells, had a replacement rate of $\lambda = 1.0$ per stem cell per year, and the noise due to sampling was simulated with offsets due to background noise $\Delta = 0.04$, $\epsilon = 0.92$ and peak specific noise with sample size $kz = 100$. (**D**) Independent validation of the inference method on a spatial representation of the single crypt with varying stem cells. Methylation distributions are noise adjusted (Methods) for the inferences on the stem pool only. Mean inferred stem cell numbers are shown for ten replicate simulations, and the red bar represents the mean of the ten replicates while the error bars denote 1 s.d.

separately identifiable within the region of the parameter space our model explores (Supplementary Information and Supplementary Figure 7 [210]).

The mathematical model and inference framework relied on a number of assumptions, such as the stem cell niche being well mixed and that individual fCpG loci act independently, the impact of which was explored via generating synthetic crypts with these assumptions loosened. Our analysis, combined with the inferences on the exclamatory bowel, is generally robust to altering these assumptions.

To further validate our Bayesian inference framework, we implemented our EB spatial model of crypt cell evolution (Section 4.2). We used the EB crypt model to generate fCpG patterns from a range of stem cell pool sizes. In order to provide insights into the fluctuating fCpG signal from the exclamatory bowel model of the homeostatic crypt we have to add noise to the output data of the spatial model. This is because the inference framework is designed to fit the noisy experimental data and the fact that fCpG sites with values of zero or one are not captured in the data. In order to add a small amount of noise of the perfect methylation distributions output by the spatial model, noise is added as described in Materials and Methods

(Section 2.5.2) where $\kappa = 1000$ and an offset from zero ($\Delta = 0.04$) and an offset from one ($\epsilon = 0.92$) is used (Figure 4.5D). Once the beta values with noise are added the inference framework is executed for each model simulation's β distributions across stem cell number ranges from $2-9$, $3-10$ and $8-15$ respectively using 400 live points for the dynesty sampler [213]. Applying our inference framework on the resulting CpG patterns, we were able to accurately recover the stem cell numbers (Figure 4.5D) for each of the three different pool sizes (4.3 ± 0.46, 5.8 ± 1.17, and 11.63 ± 1.56 stem cells (mean $\pm$ s.d.)).

4.7 Measurement of stem cell dynamics in human intestines and tissues

We measured human colon and small intestinal crypt stem cell dynamics using our FMC methodology. Methylation data were generated for each crypt individually, followed by crypt-by-crypt inference of stem cell dynamics, producing crypt-specific posterior estimates of effective stem cell number and replacement rate (Figure 4.6A).

The mean number of stem cells was found to be similar across tissues, with 5.8 ± 1.7 stem cells in normal colon samples and 6.5 ± 1.0 stem cells within small intestinal glands (mean$\pm$1 s.d.; Figure 4.6B). The replacement rate in normal colon was 1.1 ± 0.8 replacements per stem cell per year and was reduced to 0.79 ± 0.5 replacements per stem cell per year in small intestine (Figure 4.6C).

We used a hierarchical Bayesian *generalized linear model* (GLM) to account for the hierarchical structure in our data and compared stem cell numbers and replacement rates between tissues. We found that glands from the small intestine had a greater number of stem cells (Figure 4.6D; $P < 0.05$, GLM) but a lower replacement rate per stem cell than normal colon (Figure 4.6E), such that the total number of replacements per crypt was not significantly different between colon and small intestine (Figure 4.6F; $P < 0.05$, GLM).

4.7.1 Stem cell dynamics in human endometrial glands

We analyzed fCpG methylation in 32 endometrial glands derived from eight individuals using the same methodology as for intestinal crypts (Figure 4.7; Table 2.1). We derived a set of 7,721 fCpG sites, of which 807 were shared with the set of loci identified in the colon. The resulting methylation distributions exhibited the same characteristic W shape as in the intestine (Figure 4.7). We then applied our Bayesian inference pipeline to each endometrial gland to infer the effective stem cell dynamics

[148]. The inferred stem cell replacement rate was broadly similar compared to colon at 1.2 ± 0.3 (mean±s.d.) replacements per stem cell per year (Figure 4.7B), whereas the number of stem cells per gland was significantly higher in endometrium than in colon ($P < 0.05$, GLM), with each endometrial gland containing 8.6 ± 2.9 stem cells (Figure 4.7C-E). Intriguingly, the endometrium exhibited a significantly greater degree of intraindividual variability regarding the number of stem cells ($P < 0.05$, GLM), perhaps due to the dynamic nature of the endometrium through menstrual cycles and age-related changes. We acknowledge that the stem cell structure of en-

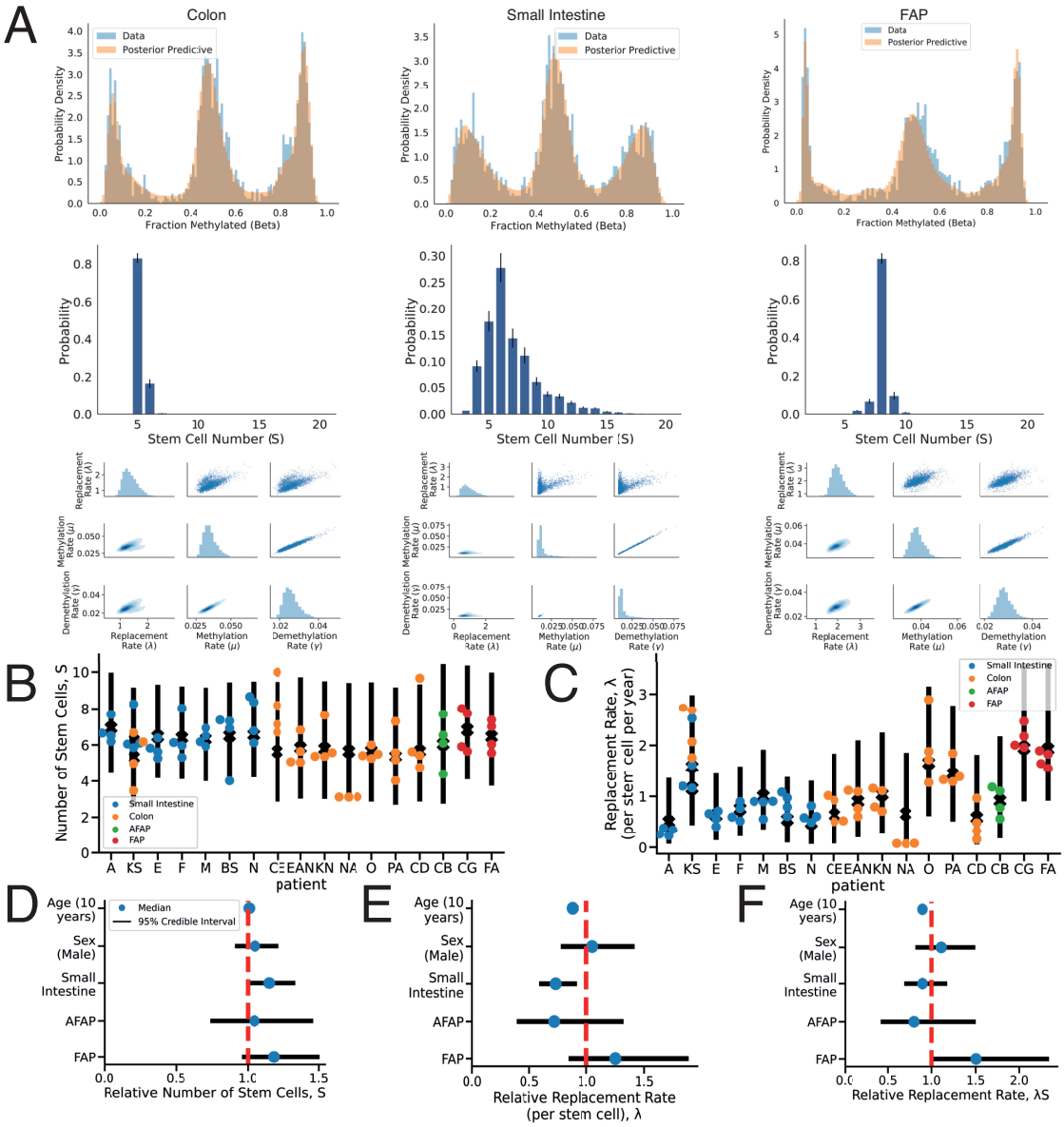


Figure 4.6: Tissue-specific differences in stem cell dynamics.(Caption next page.)

Figure 4.6: Tissue-specific differences in stem cell dynamics. (Figure previous page.) The stem cell dynamics Bayesian inference framework was applied to 71 individual intestinal crypts originating from 17 individuals. (**A**) Examples of the posterior predictive distributions, the discrete stem cell number posterior and the posterior for the replacement rate, methylation rate and demethylation rate in crypts derived from normal colon, small intestine and the colon of individuals with *familial adenomatous polyposis* (FAP) (left to right). Error bars were calculated from the estimated error (1 s.d.) on the log evidence. (**B**) Individual crypt and posterior mean per individual for the stem cell number; *attenuated familial adenomatous polyposis* (AFAP). (**C**), Replacement rate per stem cell with the 95% credible range of the GLM expectation, accounting for age, sex, tissue, disease state and intra- and interindividual heterogeneity. (**D-F**) Posterior distributions for the effect of age (per decade), sex (with female encoded as reference), tissue type and disease state on the relative number of stem cells ((**D**), replacement rate per stem cell (**E**) and total number of replacements (**F**) compared to normal colon. A Bayesian parameter estimation hypothesis testing approach was taken, such that a difference was called significant if the 95% credible region did not overlap.

dometrial glands is likely more complex than that of colon crypts [214], limiting the degree to which our simple model reflects the underlying biology. Nevertheless, the fact that we still observe large clonal peaks near 0% and 100% methylation suggests that monoclonal conversion does still occur, and our model is still applicable as a simplified caricature of the complicated dynamics present in endometrial glands.

The above analysis of human intestinal crypts and endometrial glands indicates that these small populations are maintained by small numbers of stem cells that stochastically and recurrently turnover throughout life. Experimental lineage markers can record a single clonal replacement cycle in murine crypts [45, 46, 215], but FMCs continuously record ongoing stem cell dynamics that otherwise lack definitive starts or ends in adults. Although stem cell pool sizes remain constant, replacement bottlenecks and succession to a single lineage recur with a mean fixation time of 8.3 ± 5.5 years (mean±s.d.) in the small intestine, 7.0 ± 6.4 years in the colon and 6.8 ± 4.2 years in the endometrium. Furthermore, we found that the inferred mean fixation time decreases with age, suggesting that the dynamics of stem cell replacement slow over the course of one's lifetime.

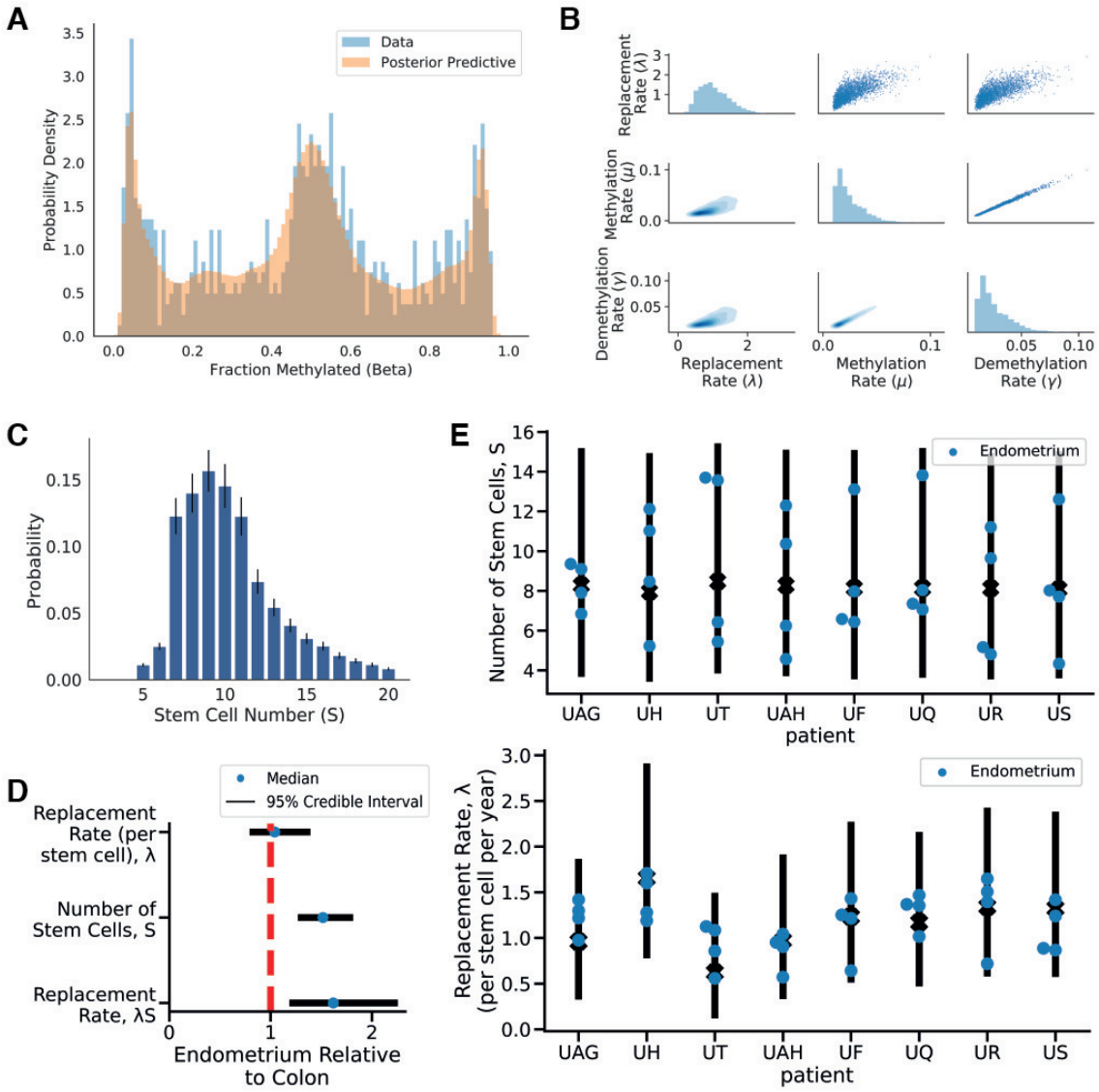


Figure 4.7: The fluctuating methylation clock method is generalizable to other glandular tissue. The stem cell dynamics Bayesian inference framework was applied to 32 individual endometrial glands originating from eight individuals. (**A**) Measured methylation β values (blue bars) and posterior predictive distribution (salmon overlay) are shown for an example endometrial gland. The methylation patterns exhibit a similar 'W' shape to that observed in intestinal crypts. (**B**) Posterior distributions for the replacement rate per stem cell and (de)methylation rates for the gland shown in (**A**). (**C**) Posterior distribution for stem cell number for the gland shown in a. Error bars were calculated from the estimated error (1 s.d.) on the log evidence. (**D**) Inferred relative replacement rate per stem cell, number of stem cells and total replacement rate in endometrium versus colon, indicating that there are significantly more stem cells per gland in endometrium than in colon. Bars show 95% credible intervals derived from a *generalized linear model* (GLM). (**E**) Inferred number of stem cells and replacement rate per stem cell for each individual gland from each individual (dots) with the 95% credible range of the GLM expectation (bars).

4.7.2 Departure from homeostasis: familial adenomatous polyposis

Individuals with *familial adenomatous polyposis* (FAP) carry a heterozygous germline mutation in the APC gene and are at increased risk of developing colorectal cancer [216–218]. APC is a key regulator of Wnt signaling, and, therefore, pathogenic APC mutations cause alterations to Wnt signaling [219–221]. Wnt signaling is essential for the maintenance of intestinal stem cells [222–224]. Consequently, we hypothesized that individuals with FAP would have altered stem cell dynamics. Inference on fCpG sites in individual FAP crypts from morphologically normal colon showed that stem cell numbers were similar in FAP crypts and non-FAP colon (6.7 ± 0.3 stem cells per crypt), but the stem cell replacement rate was almost doubled at 1.9 ± 0.3 replacements per stem cell per year (Figure 4.7A-C), resulting in a significantly higher total number of replacements per crypt per year in FAP (Figure 4.7F).

4.8 Discussion

In this chapter we have demonstrated the theory behind FMCs and how they can be used to reconstruct human cell population dynamics that start or recur at different times during life from standard bulk methylation assays. Large numbers of fCpG sites that are tissue specific reversibly flip–flop their methylation status between 0%, 50% and 100% (representing homozygous and heterozygous (de)methylation). When the appropriate fCpG sites are chosen that align with the replacement rates of the underlying cells, a characteristic W-shaped distribution emerges. At any given point in time this distribution of FMCs provides information about the clonal dynamics occurring from the last common ancestor to the population of cells sampled.

Within glandular tissue, such as an intestinal crypt, multiple stem cells are subject to neutral drift as a succession occurs with a previously inferred periodicity using the limited number of point mutations [45, 46]. FMCs provide far more loci to assess the underlying clonal dynamics occurring within the intestinal crypt and is broadly applicable to other tissues that harbor their own FMCs. The fCpG loci have different (de)methylation rates and identifying those that are not too fast, or slow is critical to observing the underlying clonal dynamics. In cases where fluctuations occur too fast even closely related cells will become asynchronous and provide little information. Conversely, if sites fluctuate too slowly there will never be any changes in CpG methylation status and cells will maintain synchrony. Here we show how various dynamics within the crypt leads to consistent patterns of fCpG distributions,

how to isolate and determine candidate CpG sites to serve as FMCs using several modeling approaches and in-depth analysis of fCpG sites.

Using an inference framework and model of fCpG distributions directly we analyze colon and small intestine crypts to determine their stem cell numbers, replacement rates, and (de)methylation rates after validating this methodology using a spatial model of the intestinal crypt. We then apply this same inference methodology to pre-malignant FAP and *attenuated familial adenomatous polyposis* (AFAP) crypts. To demonstrate the universality of FMC in glandular tissue we then apply this technique to endometrium glands that undergo regular self-renewal. This inference approach relies on several assumptions and is only suitable for determining dynamics in a relatively small number of stem cells.

FMCs have potential for use in numerous tissues and possess the capability to answer many contemporary questions about human cell populations. The erratic flip–flop behavior of fCpG sites in polyclonal populations is suspected to become detectable only after expansions towards a clonal cell population. FMCs can measure alterations that start or recur throughout life and are capable of being used to track changes over a few years. FMCs have potential use, not only in uncovering dynamics of homeostatic tissue, but also to provide a means of early detection for pre-malignant tissues. This may be especially pertinent when considering the homogenization of stem cells in hematological diseases.

Chapter 5

Exploiting fluctuating methylation clocks identified in peripheral blood

5.1 Background

In Chapter 4 we discovered FMCs and exploited this methodology to determine the stem cell numbers and replacement rates in various healthy and precancerous glandular tissues. Within the colon crypts and endometrium glands regular monoclonal conversions occur via neutral drift over a predictable period, in healthy and pre-malignant tissue, due to their small numbers of stem cells and spatially confined organization [45, 46, 62]. This differs from the hematopoietic system where polyclonality of the HSC population is extensive and clonal expansions are only likely in the most severe malignancies due to the multiple orders of magnitude more stem cells in a spatially unstratified tissue. The fCpG behavior seen in intestinal crypts and endometrial glands (both epithelial tissues) are likely to be present across other tissue types.

We hypothesize that the loss of polyclonality, or homogenization of the HSC pool, should be reflected in fCpG distributions of normal, pre-malignant, and malignant samples (Figure 5.1). This idea is easily explored through abundant publicly available datasets. Unlike intestinal crypts which recurrently drift to clonality, blood is a large, well-mixed tissue with diverse cell types and is normally polyclonal because it is produced by thousands of bone marrow stem cells [36, 80]. Normal HSC turnover is not synchronized as in the intestines, CpG loci that randomly fluctuate through 0, 50 and 100% methylation in individual cells will have average methylation around 50% in normal polyclonal blood samples.

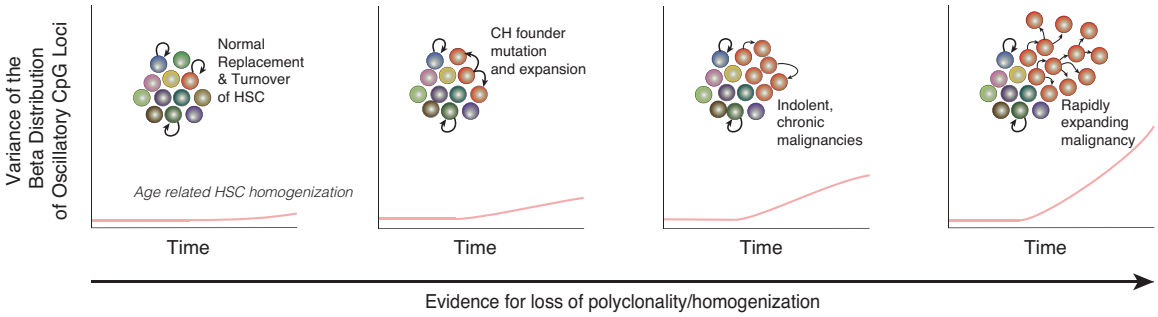


Figure 5.1: Fluctuating methylation clock idealization in whole blood. Illustration of a diverse HSC compartment with normal turnover and *fluctuating CpG* (fCpG) variances. Through several possible routes the variances move from the normal, steady state at 50% to the "W" like distribution seen in the crypt in the most extreme, rapidly expanding malignancies. Various stages of *hematopoietic stem cell* (HSC) homogenization is reflected in the fCpG variances.

Here, we build a mechanistic HSC model to show, through symmetric divisions, how fCpG variance increases as polyclonality is lost. We turn to the abundant public methylation array datasets for normal, pre-malignant, and malignant whole blood samples to derive fCpG sites. Unlike the crypt, a separate, less refined, identification process is necessary as heterogeneity of CpG β values is inaccessible within patients due to a lack of measurements within the same individual. Unfortunately, no methylation array datasets currently exist for CH. We collect and analyze 38 patients whole blood samples who have paired methylation array and mutation data, from paired PB and BM, with confirmed CH along with ten confirmed normal control samples. Using this cohort, we develop a novel approach to diagnose CH agnostic to mutation status, evaluating evidence of CH in 1,388 patients.

5.1.1 Chapter aims

1. Evaluate fluctuating methylation clocks as a marker for a loss of polyclonality via development of a mechanistic hematopoietic stem cell model.

2. Identify and analyze fluctuating CpG sites in publicly available methylation datasets of normal and malignant whole blood samples.

3. Perform analysis on methylation array data of new cohort of clonal hematopoiesis patients and identify fluctuating CpG sites.

4. Develop a machine learning classification tool to diagnose clonal hematopoiesis and perform retrospective analysis on 'normal' whole blood samples.

5. Evaluate clonal hematopoiesis classifier and hematopoietic stem cell dynamics using the hematopoietic stem cell model.

5.2 Hematopoiesis model

Whole blood was simulated in Java using the HAL framework [109] as a non-spatial agent-based model using 27,634 fCpG sites as measured in the experimental data. Parameters (Table C.1) for normal hematopoiesis are numbers of HSCs (N), number of possible division events (T), (de)methylation rates (S) for the fCpG sites, and HSC replacement dynamics (λ). To model clonal expansion, a single, random, cell is selected to grow upon induction, with the added parameters of expansion rate (E) and final blood frequency of its clonal expansion (ω). These clonal expansions result in the overall population size growing until the appropriate final blood frequency was reached. The output of these simulations provided β values at the fCpG sites as well as the overall distribution variance over time.

The number of HSCs is set at a lower value of 1000 initiating cells. This was much lower than the 30,000 based on the large number of HSC inferred by DNA sequencing studies [36, 80]; however, the results shown here are invariant to more than 1000 initiating cells as the output of methylation status is similar across various ranges of HSCs. (De)methylation rates varied between CpG sites and were assigned based on the distribution averages of the 656 normal, healthy, individuals from GSE40279 [139]. We found that some of the whole blood fCpGs did not appear to have equal (de)methylation rates because their averages tended to always be above or below 50% in multiple individuals. Hence, to better model and match the data, we used a look-up distribution table in the simulations to initialize a cell's fCpG parameters, with lower and unequal (de)methylation rates at CpG sites with average methylation typically found near 0.4 (demethylation > methylation) or near 0.6 (methylation > demethylation) to maintain the variance of the 27,634 fCpG sites around 0.1 during cell divisions. The (de)methylation rates varied between 0.0001 to 0.001 changes per division, with the highest (de)methylation rates being more equal at CpG sites near 50% methylation.

Cell survival was set at exact replacement (one cell produces one living offspring), and results did not vary much if random replacement was simulated, where a random stem cell is removed and a newly initiated cell replaces it. A proportion of cells underwent replacement at each timestep (Table C.1). For the neoplastic simulations in Figure 5.2D the expansion rate (E) was varied to model either rapid expansion

(visible or more than 5% leukemic cells within 1 year or 200 divisions) akin to acute leukemia, modest expansion (visible within 4 years or 12,000 divisions), or very slow expansion (visible within 6 years or 18,000 divisions). The extent of blood involvement was varied between 20% (black lines), 50% (blue lines) and 90% (red lines). These simulations indicated that clonal expansions change whole blood fCpG variances depends both on how fast the expansion grows and to what extent it involves the blood. Rapid growth to high levels like acute leukemias results in high fCpG variances and characteristic W-shaped fCpG distributions. Slower growth to lower levels like chronic leukemias results in low fCpG variances and broader distributions that lack the W-shape. Interestingly, very indolent clonal expansions which may occur with CH [225] can result in small increases in fCpG variances, which may account for the age-related increase in fCpG variances seen in Figure 5.2A.

More sophisticated modelling with a better selection of whole blood fCpG sites could improve the extraction of ancestral information. For example, a selection of slower fCpG sites may improve the detection and analysis of indolent clonal expansions, where many of the faster fluctuations return to average $\sim 50\%$ methylation by the time the expansion reaches detectable blood levels.

The simulation framework can be obtained, along with sample simulation results, on GitHub through `https://github.com/MathOnco/flipflopblood.git`[226] A GUI compatible with most operating systems is accompanied to allow for rapid evaluation of different parameters.

5.3 Fluctuating methylation clock dynamics in hematopoiesis simulations

We simulated hematopoiesis to better understand how fluctuating sites detect clonality in whole blood (Figure 5.2). Methylation fluctuates between 0%, 50% and 100% in single cells and the simulations indicate polyclonal whole blood variance is low and stable through time because human hematopoiesis is maintained by large numbers of stem cells. Clonal expansion by a single cell synchronizes fluctuations and results in higher whole blood variances that depend on growth rates. As in the crypts, there is a balance between clonal expansion rates, which increase population variances, and the rates at which fluctuating sites drift back to 50% average methylation, which decreases variance. A rapid expansion (less than 2 years) to high blood levels as in acute leukemias produces high variances and W-shaped distributions. The "W" methylation pattern resembles the methylation at 0, 50 and 100% methylation of the

initiating cell. Expansions that grow more slowly have variances greater than normal blood but lack the W-shape as methylation fluctuations become increasing desynchronized with time. These more indolent expansions are more consistent with the experimental data for chronic myeloproliferative neoplasms, which may be asymptomatic and persist for years. Clones that grow even slower and arise later, as may occur with age related CH [225], lead to slightly higher variances, as seen with aging in the normal whole blood cohort. Hence, a simple model with $27,634$ fCpG sites and different rates of clonal expansion is broadly consistent with the experimental data from hundreds of clinical samples.

5.4 Fluctuating CpGs in publicly available blood samples

We identified suitable fCpG loci by averaging normal whole blood DNA methylation at $\sim 450,000$ autosomal CpG loci from a commonly used aging database of 656 healthy individuals [139]. We selected all loci ($N = 27,634$) with average values between 40 and 60% methylation in these 656 specimens. fCpG appear tissue-specific because only $\sim 5\%$ of the intestinal loci were in the blood set. Fluctuating methylation for each individual sample revealed tight distributions around 50% methylation, which can be described by its variance (Figure 5.2C). Consistent with random fluctuations, average fluctuating methylation remained 50% with aging. Serial samples ten years apart [228] revealed variance to be relatively stable for an individual, with a slight significant trend for increases with age (Figure 5.2D), which was also observed throughout aging (Figure 5.2C). HSCs start out homogenous and expand, introducing mutations and individual lineages until patients are approximately 20 years old [169, 229]. We observe this in the fCpG variance in two AML cohorts where a predominantly older cohort (70.3% over the age of 50; GSE62298) exhibits significantly lower fCpG variance compared to the pediatric AML cases, all under the age of 18 (n=64; GSE133986) (Figure 5.2E). HSC homogenization is more exaggerated within these pediatric samples in early HSC homeostatic development.

CH in the blood is an early step in the evolution of neoplasia and will increase variances because clonal cells will initially share the 0, 50, 100% methylation pattern of the progenitor. For rapid clonal expansions (i.e., acute leukemias), W-shaped blood distributions like those observed in the crypts are expected. Consistent with these expectations, whole blood samples from different types of major hematopoietic neoplasms had higher than normal variances (Figure 5.2A). Acute lymphoblastic

leukemias (ALL) and acute myeloid leukemias (AML) had the highest variances and characteristic W-shaped distributions. More indolent chronic myeloproliferative or myelodysplastic whole blood specimens showed more modest variance increases and generally lacked the "W" shape of the acute leukemias, crypts and glands.

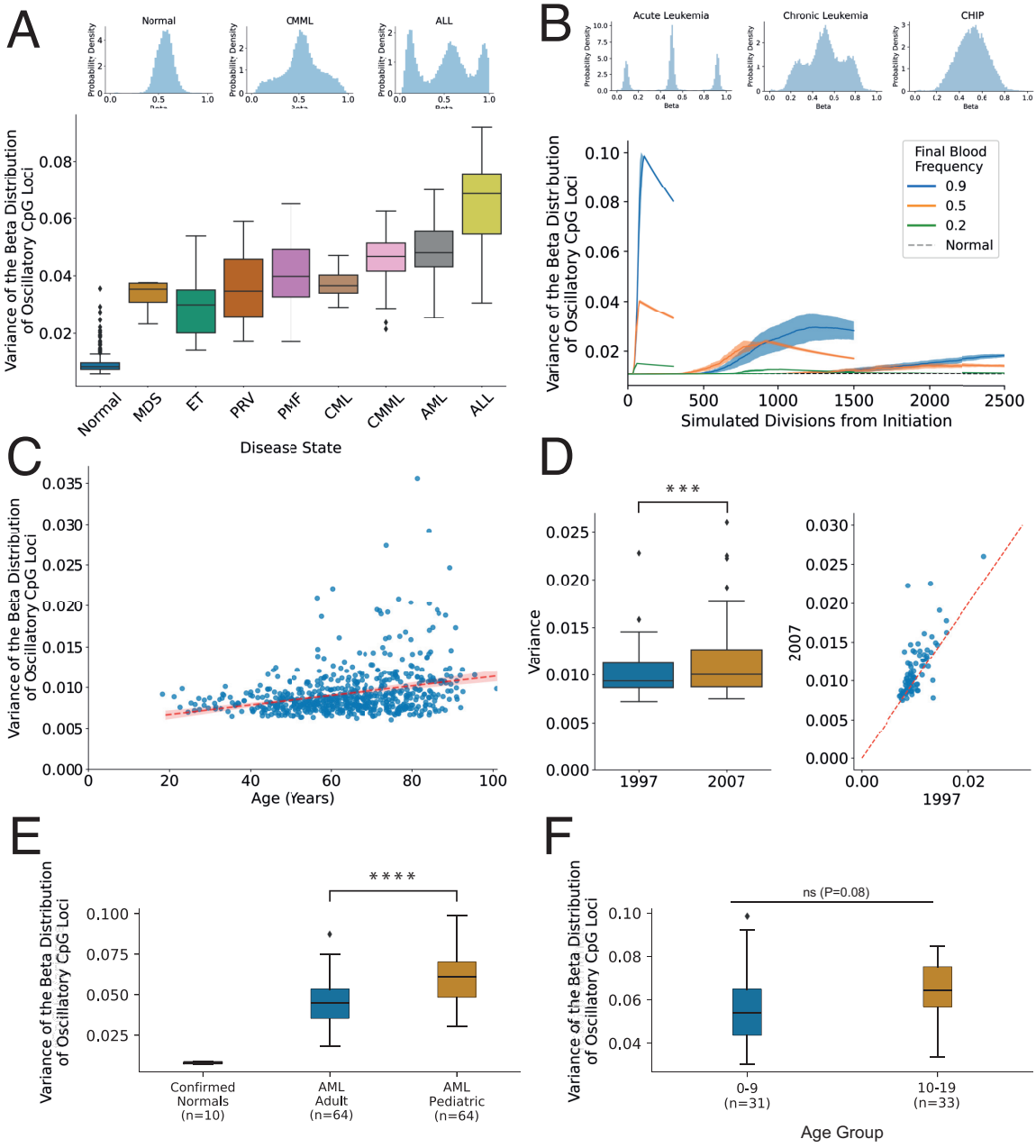


Figure 5.2: Fluctuating CpG dynamics can be observed in chronic and acute leukemia.(Caption next page.)

Figure 5.2: Fluctuating CpG dynamics can be observed in chronic and acute leukemia. (Figure previous page.) **A**: The variance of the fCpG methylation distribution is a proxy for the rapidity of the clonal expansion within the blood. In normal samples the large stem cell population size leads to the methylation distribution being concentrated near 50% (as one would expect for uncorrelated oscillators). However, as a clonal cancerous population expands, clonal peaks begin to separate from the 50% peak. In the case of ALL, the large, well-separated peaks near 0% and 100% are indicative of a single clonal population making up the majority of the remaining stem cells following rapid growth. **B**: Simulations confirm that a simple HSC model can recapitulate the observed methylation distribution observed in patient data. Data presented as mean±SEM. **C**: The variance of the fCpG methylation distribution experiences a gradual increase with age in normal patients. Confidence band was calculated via bootstrapping and represents 95% confidence intervals. **D**: The variance of paired blood samples taken 10 years apart (1997 and 2007) also exhibits a small but marked increase (0.37 Cohen's d, $p = 2.8 * 10^{-5}$ two-sided paired t-test). **E:** Significant differences are observed in AML patients that are predominantly older than 50 years old compared to pediatric patients younger than 18 years old. Comparisons of fCpG β distribution variance shows that pediatric samples with ages less than 18 (GSE133986; [227]) exhibit higher variance than an older patient cohort (19 patients in this cohort are < 50 years old; GSE62298 [145]) (two-sided Welch's t-test, $P = 7.1 * 10^{-8}$). Confirmed normals are those presented within this study. **F:** Stratification of the pediatric cohort fails to reveal any significant differences in fCpG β distribution variances (two-sided Welch's t-test, $P = 0.08$). For **A**, **D**, **E**, and **F** center line, median; box limits, upper and lower quartiles; whiskers, $1.5IQR$.

5.5 Clonal hematopoiesis

Clonal hematopoiesis (CH) is diagnosed based on somatic alterations whose frequencies are greater than 2% (generally SNVs and small insertions and deletions assessed from peripheral blood samples) in the absence of hematologic malignancy [160, 230, 231]. The prevalence of CH increases as an individual ages and conveys a non-negligible risk for progression to various hematopoietic malignancy [151–153, 161, 232, 233]. While these studies focus on specific somatic alterations, there are others that have more generally found an increased risk of hematologic malignancies by defining CH as samples with high numbers of somatic mutations [234]. Further still, chromosomal anomalies such as large structural variants or CNAs are also associated with increased risk of hematopoietic malignancies [161, 232, 233, 235–238], but are not generally defined as a diagnostic method for CH, despite their more common occurrence in individuals who subsequently develop myeloid or lymphoid leukemia, as observed through longitudinal studies [235]. While these risks for all

somatic alterations carry an increased risk of developing malignancies, the absolute risk is low; however, incorporating a more broadly universal method for identifying CH is of great value, especially if information could be gained for teasing apart risk for hematopoietic malignancies. Here we develop a method agnostic to the underlying somatic alteration type by using FMC behavior.

5.5.1 Clonal hematopoiesis patient cohort

Patients undergoing elective total hip replacement surgery were diagnosed as either normal or CH based on the VAF of putative driver mutations. Here, we present ten patients with no evidence of CH and 38 patients subdivided into different VAF groups ([1,2)% n=8, [2,5)% n=10, [5,10)% n=10, ≥10% n=10) whose VAF of putative drivers is greater than 1%. Most patients present with a DNMT3A driver mutation (≈55% (21/38) CH patients), where ≈81% (17/21) of those DNMT3A drivers are the highest frequency driver mutation for that patient. TET2 is the second most frequently observed driver (≈26% (10/38) CH patients), where ≈80% (8/10) TET2 drivers are the highest frequency driver mutation for that patient (Figure 5.3B). A smattering of additional driver mutations are observed, such as ASXL1 (n=5), TP53 (n=2), and SF3B1 (n=3). We don't observe any significant differences between the predominant driver mutations (DNMT3A and TET2) for their corresponding variant frequencies or patient ages ($P = 0.76$ and $P = 0.17$, respectively, two-sided Welch's t-test, Figure 5.3B). In addition, there is no significant increase in the observed VAF of driver mutations with age (Figure 5.3B).

5.5.2 Clonal hematopoiesis fluctuating methylation clocks

fCpG in the blood form a predictable distribution around 50% methylation in normal samples (Figures 5.2A and 5.3C). When founder mutations leading to CH occur, this increases the variance of these fCpG methylation distributions as evidence for loss of polyclonality or homogenization occurs. This homogenization is when we begin to see fCpG sites shift to being concentrated at 0, 50, and 100% methylation, similar to those observed in glandular tissues. Consistent with these findings we see that the mean variance for the control samples: confirmed normal is $8.1 * 10^{-3} \pm 5.7 * 10^{-4}$ (mean±SD, n=10); while the mean variance for the CH samples is $0.012 \pm 3.8 * 10^{-3}$ (mean±SD, n=38). Normal samples taken from GEO[139] exhibit a similar mean, but the distribution of fCpG variance has a left skew towards loss of polyclonality as reflected in a higher standard deviation ($8.9 * 10^{-3} \pm 3.7 * 10^{-3}$, mean±SD; n=656);

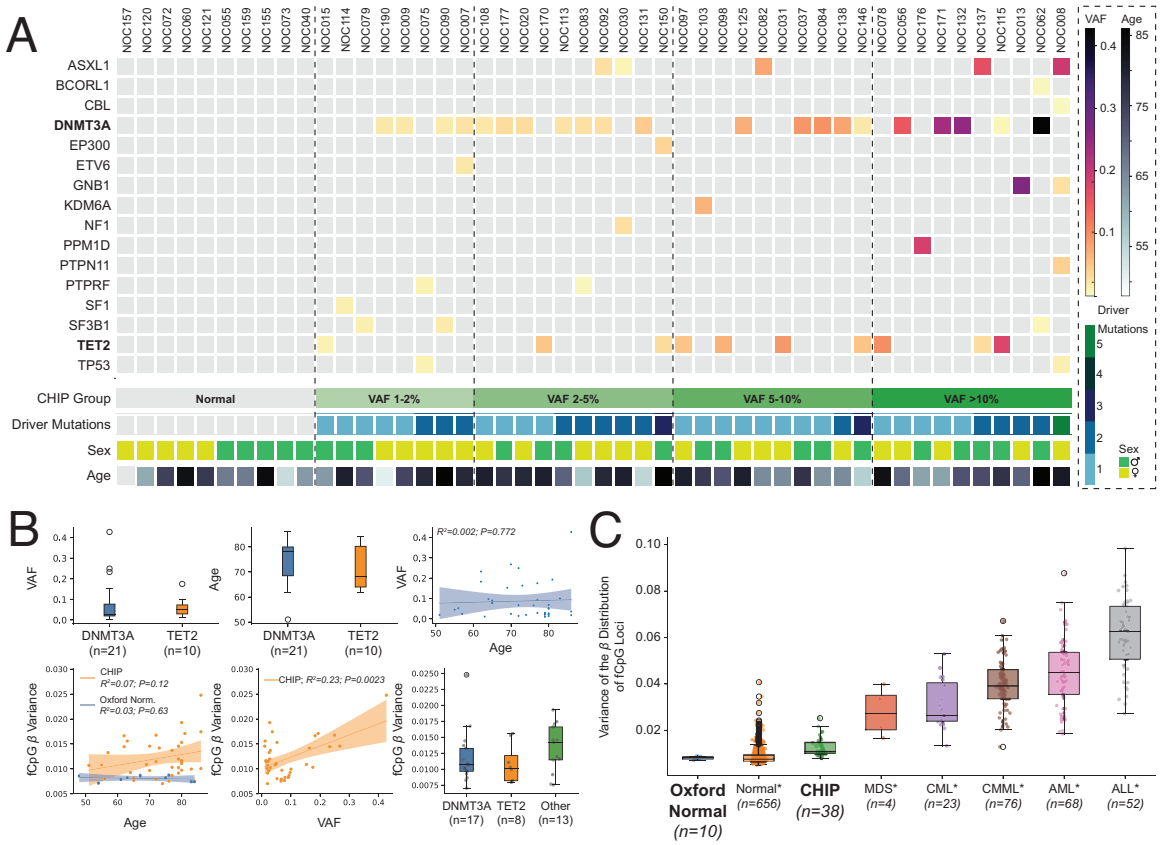


Figure 5.3: Characterization of CH patient samples. A: Cohort characteristics for CH samples presented here, including driver mutation VAF and burden, sex, age, and cohort group. **B:** There is no difference between patients containing DNMT3A and TET2 mutations in both their VAF or age (center line, median; box limits, upper and lower quartiles; whiskers, 1.5 IQR; two-sided Welch's t-test, $P > 0.05$). Overall, there is no increase in mutation's VAF across the age of the patients. The bottom row shows minimal increase in fCpG variance correlated with age and a slight correlation with driver VAF. No significant differences in fCpG variance is observed between DNMT3A, TET2, and other drivers. **C:** Variances of fCpG methylation distributions now including the confirmed normal controls ('Oxford Normal') and CH samples. For **B** and **C** plots show center line, median; box limits, upper and lower quartiles; whiskers, 1.5 IQR. All ribbons show the calculated 95% confidence interval for the regression

however, low frequency CH driver mutations cannot be ruled out for this sample group as paired mutational data was not collected (refined fCpG site filtering can be found in Materials and Methods C.1.1). The next highest fCpG variance, at an order of magnitude higher, are in malignant samples for MDS ($0.028 \pm 9.0 * 10^{-3}$ mean±SD; n=4) [141] and CML (0.031 ± 0.01 mean±SD; n=23) [143]. Recent work has shown that 92.4% of clones expand steadily with slow exponential expansions after driver acquisition, primarily for TET2 and DNMT3A mutations [239]. In support of this, we expect and observe no differences between fCpG variances for TET2 and DNMT3A ($0.012 \pm 4.0 * 10^{-3}$ and $0.011 \pm 2.9 * 10^{-3}$ mean±SD, respectively) driver mutations (two-sided Welch's t-test, $P > 0.05$; Figure 5.3C). We do see a slight increase in fCpG variances for other driver mutations who are expected to expand rapidly, but there are insufficient numbers of these drivers to draw definitive conclusions from this dataset.

5.6 Mutation agnostic tool for diagnosing clonal hematopoiesis

To evaluate whether CH can be diagnosed using fCpG sites we used our cohort of confirmed CH and normal samples. Publicly available normal methylation data does not have any paired mutational data to rule out the presence of CH in the study's cohort of patients, especially patients with early CH where clones would be observed only at very low frequencies. Here we sub-sampled our 12,000 fCpG sites, without replacement, to bolster our numbers of samples to roughly 1500 samples of both Normal and CH (1500 and 1482, respectively). Each patient is sub-sampled equally so that representation from each sample is equal across our normal and CH VAF groups. This leaves us with 2982 samples of roughly equal representation from normal and CH samples (Figure 5.4A). In addition to the 2000 fCpG sites sampled to create each sub-sample, three summary statistics for each sub-sample (mean, variance, and standard deviation) is used to bolster our predictive power (Figure 5.4A). We perform a train/test split of 80%/20% for these samples and train a Gaussian Naive Bayes supervised learning algorithm to classify our samples as either CH or normal. The algorithm is trained using a five-fold cross validation scheme with prediction probabilities calibrated using non-parametric isotonic regression. This FMC-CH classifier is capable of differentiating CH and normal samples with an accuracy of 90.5% and an area under the receiver operating characteristic curve (AUC) of 0.903 (Figure 5.4B). Additionally, the balanced F-score (F_1, balance in precision and recall where

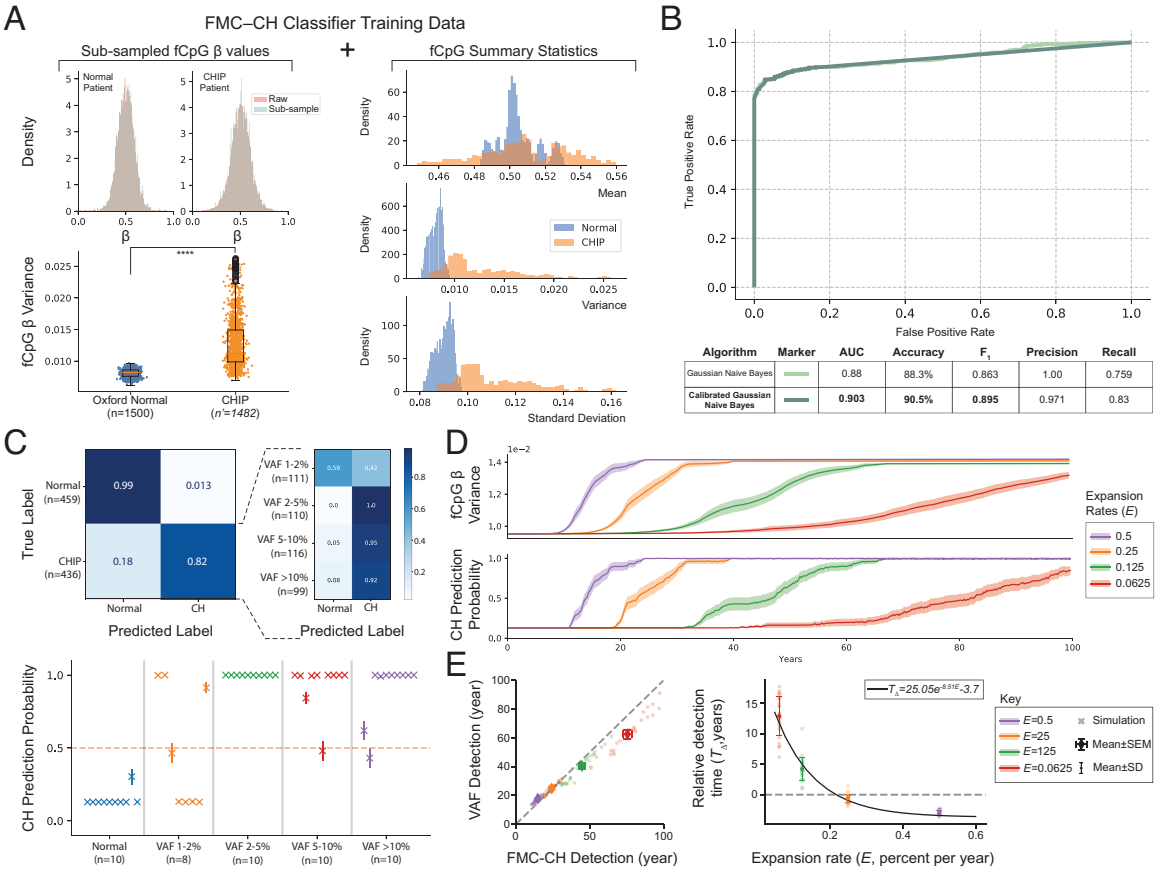


Figure 5.4: FMC-CH classifier. A: Illustration of an exemplary fCpG β distribution for a normal and CH sample where bootstrapping provides a larger number of samples for machine model train/test datasets used for the FMC-CH classifier (center line, median; box limits, upper and lower quartiles; whiskers, $1.5IQR$). **B:** Gaussian Naive Bayes machine learning model provides the greatest performance where data pre-processing balances the precision and recall. **C:** The FMC-CH classifier is most useful for clones greater than 2%. Error bars (bottom) denote 95% confidence intervals for the CH prediction probability from 200 classifications. **D:** Simulations with varying expansion rates, E, determine time when the classifier will diagnose CH; however, prediction probabilities increase as the subclonal population expands (plots show mean of 25 simulations with mean and ribbons showing SEM). **E, left:** Comparison of the earliest time when diagnosis would be made using the gold standard VAF with the FMC-CH classifier shows faster expanding clones are detected earlier, while slower clones are detected later. This relationship is quantified to show that the relative detection time decreases exponentially as expansion rates increase (right) (Data shown is from 26 replicates for each parameter combination shown in **D**).

$F_1 = 2 * \frac{precision*recall}{precision+recall}$) for this classifier is 0.895 (Figure 5.4B). In the absence of the calibration and five-fold cross validation (i.e., no isotonic regression) we see that the accuracy is very close at 88.3%, but the F_1 score is lower at 0.863 indicating a larger imbalance in precision and recall.

5.6.1 Classifier performance on clonal hematopoiesis patient cohort

The FMC-CH classifier proves to be a robust tool for diagnosing CH based only on fluctuating methylation clocks. The confusion matrix between the normal and CH samples yields a false positive rate of only 1.8% and false negative rate of 1.3% (Figure 5.4C). Our FMC-CH classification tool is a binary classifier where samples are either CH or normal. However, our CH cohort has been carefully curated for samples with known degrees of CH subclonal expansions, even those that are below clinical detection (< 2% VAF). fCpG variance correlates with the VAF of CH drivers (Figure 5.3B), thus we would expect that the classifier would struggle with sub-clinical CH samples. We evaluated how well the model performs across the different VAF groups by examining the true CH samples. When we pull these out of the confusion matrix, we can see that the lowest accuracy is for the CH group at 1-2% VAF where the false negative rate is 58% (Figure 5.4C). Across the remaining groups the false negative rate does not exceed 8%.

A diagnostic or research tool is only useful with an appropriate interface and an easily integrated backend functionality. For the FMC-CH classifier a function is provided that takes an array of fCpG β values. The function pre-processes these samples, by performing the appropriate sub-sampling (allowing backwards compatibility with 450k, 850k, and earlier CpG probe sets), extracting the relevant summary statistics, and then performing the classification. This process is performed 100 times (by default, but the user can specify) providing summary statistics, such as a 95% confidence interval of the prediction probabilities. The threshold for CH classification from the replicate predictions is based on the 95% confidence interval, where CH is diagnosed if the upper limit of the 95% confidence interval is > 50%. We illustrate this process by performing predictions on the entire CH cohort (Figure 5.4C), revealing that samples with the smallest frequency CH driver clones are the most difficult to diagnose.

5.6.2 *In-silico* evaluation of classification model

The gold standard for identifying CH is through identification of deleterious somatic alterations from PB with a VAF cutoff of 2%. We know that FMC dynamics are a function of several underlying processes. At their core it is the turnover, cell number, expansion rates, (de)methylation rates, and final blood frequency of subclones/malignant populations 5.2B. Using our *in silico* model of HSC turnover and subclonal expansions we evaluated under what parameters the FMC-CH classifier may struggle to resolve CH status. The primary consideration for whether a growing subclone will be detected by the classifier is the rate of expansion. The rate of expansion, combined with the frequency of subclones, are the largest contributors to changes in the fCpG variance. Clones that expand with high rapidity result in the earliest detection using the FMC-CH classifier (5.4D; 50 percent per year). However, this steep increase in fCpG variance is similar to those seen in malignancies (Figure 5.2B and 5.3C) where clonal frequencies are generally >> 20%. Clones expanding slowly take much longer to reach a point where they are classified as CH, since the number of fCpG sites changing with the stem cell expansion are too few and remain 'hidden' in the fCpG β distribution, maintaining a distribution similar to the normal polyclonal HSCs fCpG β distributions (Figure 5.2A and B). However, once these subclones have reached a high enough frequency, comparable to those that are seen in CH, the FMC-CH classifier accurately determines the presence of CH within the simulations (Figure 5.4D and 5.6A).

While the FMC-CH classifier can accurately determine that CH is present with increasing accuracy as a clone expands (Figure 5.6C), there are differences in how soon the classifier reliably determines when CH is present based on the expansion rate of a clone. The gold standard for diagnosis of CH is based on a collection of different putative driver genomic alterations (SNVs as well as small insertions/deletions) that result in fitness advantages for that HSC population. Here we calculate a corresponding VAF likeness for a growing subclone within our HSC model to compare a gold standard VAF diagnosis with the FMC-CH classifier. This is accomplished using the method described in the Materials and Methods 2.5.1 defing the depth, d, as 200 and defining VAF_t for the expanding subclone as half its final blood frequency (assuming a heterozygous single clone). When a clone arises rapidly (50% per year on average), the classifier works better than the gold standard with a mean detection time 2.97 ± 0.4 years (mean$\pm$ standard deviation) earlier than if the VAF was used (Figure 5.4D). However, there is a relationship between how quickly a clone grows and how soon it will be detected with the gold standard VAF diagnosis and the FMC-CH

classifier. When a clone grows slowly ($\approx 6.25\%$ per year) the difference in time to diagnosis using the classifier increases to 12.9 ± 3.2 years (mean±standard deviation) after a VAF diagnosis. We see that the time between diagnosis in Figure 5.4D increases faster for every halving of the expansion rate. Overall, this relationship can be quantified where the relative time to diagnosis from the VAF gold standard, T_Δ, given the expansion rate, E, is $T_\Delta \approx 25.05e^{-8.51E} - 3.7$ (Figure 5.4E (right)).

Subclones within the simulations are all induced at 5 years. This early induction allows us to visualize the dynamics over the course of a 100-year simulation. While induction times of a driver mutation is possible this early in an individuals life, these drivers would likely expand very slowly, either through persistence (i.e., not lost during homogenization of HSCs during aging) or through slow continuous expansions, such as in our model. These clones will be detected at a point when their frequencies are of an appreciable size, where a second oncogenic driver mutation has a higher probability due to population size. If the second hit occurs earlier leading to a rapid increase then this population would be detected early using our method. This distinction in detecting slow clones when they are larger or fast clones earlier could be a salient feature of this diagnostic approach.

5.6.3 Retrospective examination of normal whole blood cohorts

Based on the fCpG variances of the normal samples from the publicly available dataset we wanted to evaluate whether there was any evidence of CH present within these samples even if it may be below the clinical threshold of 2% VAF, given this has never been queried for these samples. For this we used the normal dataset of 656 patients used to derive fCpGs (GSE40279 [139]; Figures 5.2A and 5.3C) and another large data set of normal patients with 732 patients (GSE87571 [147]). Using our CH classifier on each cohort we find support for classifying 29.1% (191 of 656) and 19.8% (145 of 732) as having CH within GSE40279 and GSE87571, respectively (Figure 5.5A and B). Across both normal cohorts (n=1,388) our FMC-CH classifier finds 24.2% of individuals have CH.

5.6.3.1 Validation of newly diagnosed clonal hematopoiesis patients

The current gold standard for validating the 24.2% of newly characterized patient samples would be to perform DNA sequencing on a panel of CH drivers to examine evidence of mutations with $VAF \geq 2\%$. Neither of these cohorts have paired mutational data for validation purposes. However, we can examine whether expected

characteristics of a CH cohort is present, and we can look for evidence of copy number alterations that could be significantly enriched within the newly identified CH patients.

The likelihood of a clonal hematopoiesis diagnosis increases with age

CH, as outlined above, is typically a disease largely limited to the elderly. By the age of 70, 10-15% of individuals will present with CH [240] and by 85 years, more than 30% will have CH [241]. As a sanity check we would expect that our FMC-CH classifier would classify CH in predominantly older patients. Across the two studies evaluated we see that the median age of normal samples is 54 years, significantly different from those with CH, whose median age is 70 years (-0.67 Cohen's d, $P = 2.09 * 10^{-25}$ two-sided paired t-test). Individually for each cohort, we see that the median age for samples classified as CH from GSE40279 and GSE87571 is 73 and 62 years compared to the normal samples of 62 and 45 years, respectively (GSE40279, -0.77 Cohen's d,

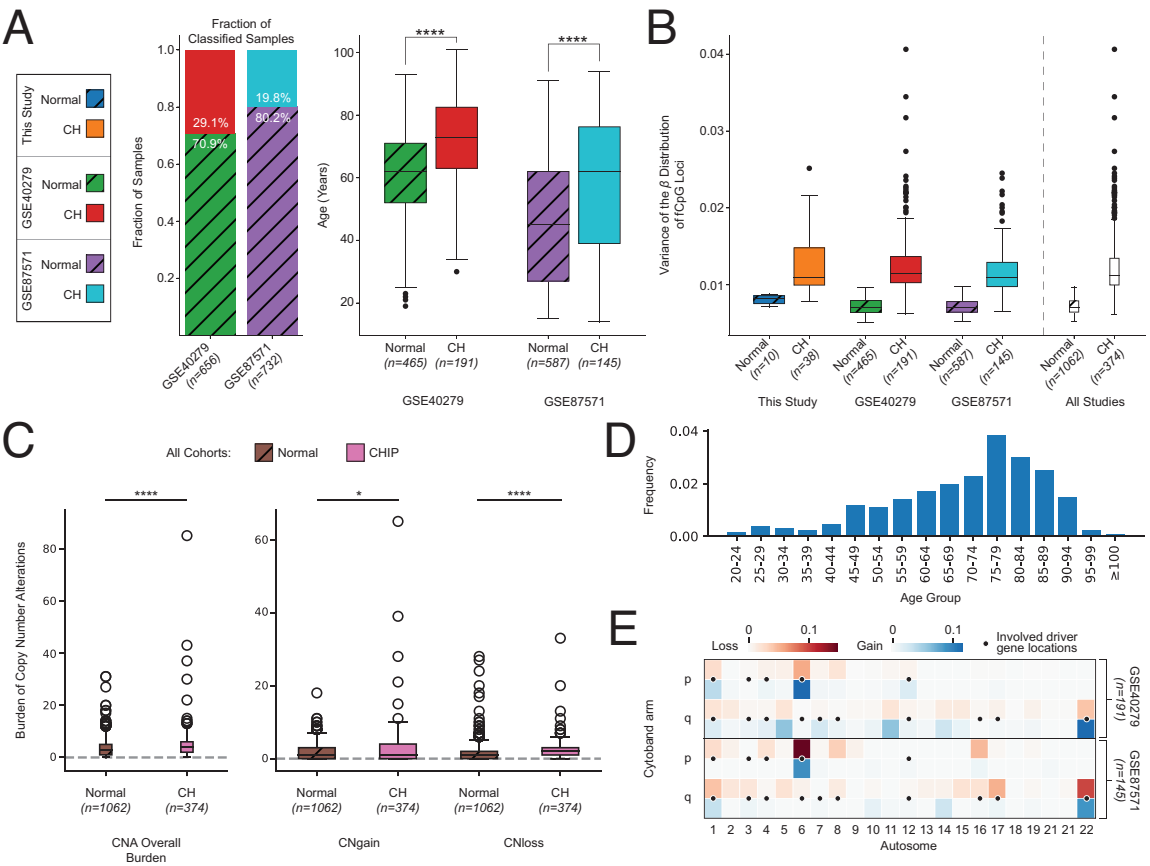


Figure 5.5: Retrospective analysis of 1,388 normal whole blood samples. (Caption next page.)

Figure 5.5: Retrospective analysis of 1,388 normal whole blood samples. (Figure previous page.) **A (left):** Retrospective classification using the FMC-CH classifier on two large publicly available normal cohorts' peripheral blood fCpG methylation sites (GSE40279 (n=656) and GSE87571 (n=732)). Across the cohorts the CH classifier finds evidence of 29.1% to 19.8% in each cohort respectively. **A (right):** A significant difference in age for the individuals in each cohort is apparent, where those classified as having CH are significantly older (GSE40279: -0.77 Cohen's *d*, $P = 3.13 * 10^{-18}$; GSE87571: -0.54 Cohen's *d*, $P = 7.90 * 10^{-09}$, both P-values from two-sided Welch's t-tests). **B:** fCpG *beta* distribution variances from the classified CH are not significantly different from those used to train and evaluate the FMC-CH classifier. Combined (to the right of the black, dashed line), we see the expected increase in fCpG *beta* distribution variances for samples exhibiting CH. **C:** CNA analysis on newly classified CH and normal samples reveals a significant increased overall CNA burden, a slightly significant increase in CNA gains (CNgains), and a more significant increase in CNA losses (CNloss) for the classified CH samples (CNA burdens $P = 8.26 * 10^{-5}$; CNgains $P = 0.022$; CNloss $P = 0.00025$). **D:** The frequency of individuals with a presence of CNAs in the classified CH samples across both cohorts increases with age from 0.27% for those younger than 50 to 11.14% for those > 70 years old ($P = 0.0013$, chi-squared test). **E:** CNA calls using the samples not classified as CH as the reference we see several recurrent genomic positions across the autosomes where losses and gains are observed. Shown is the frequency of gains/losses in the shown genomic regions for each cohort's CH samples. Several of these losses/gains involve regions where oncogenic/CH driver genes are present (black dots). All boxplots show center line, median; box limits, upper and lower quartiles; whiskers, 1.5IQR.

$P = 3.13 * 10^{-18}$; GSE87571, -0.54 Cohen's *d*, $P = 7.90 * 10^{-09}$). When comparing the age distributions for each of these two cohorts we see that GSE40279 is a significantly older cohort (64.0 ± 13.7 years; mean±SD) compared to GSE87571, which also has a broader sampling of patient ages (47.4 ± 20.9 years; mean±SD). This is reflected in the differences between the proportion of samples that are classified as CH between the two cohorts, where the older cohort, GSE40279, had 9.3% more CH classifications. For the entire cohort of patients (n=1,388) the median age is 58 years. Patients older than 58 years are more likely to be diagnosed with CH compared to those 58 or younger across these two cohorts (*odds ratio* (OR)=3.02, two-sided fisher exact test, $P = 1.28 * 10^{-17}$).

Copy number alterations are greater in clonal hematopoiesis samples

Newly diagnosed CH samples from GSE40279 and GSE87571 exhibit fCpG variances consistent with the verified CH samples presented here (Figure 5.5). As illustrated

through the analysis of publicly available malignancy methylation data with our samples (Figures 5.2A and 5.3C) and alongside the HSC simulations (Figures 5.2B and 5.4D) we expect that fCpG variances would be driven higher by homogenization of the HSC pool, which we show correlates with age above. Within the newly classified CH samples no mutation data is present to be evaluated. Thus, we can't confirm that the newly classified CH do or do not have CH driver SNVs. However, because CH is also driven by copy number alterations there are several insights that can be gained by understanding the copy number burden differences between the classified normal versus CH samples. Due to these two normal cohorts not having DNA sequencing we are unable to rule out that either sample group has driver SNVs, but we hypothesize that several samples would exhibit a higher degree of aneuploidy.

We performed copy number calls using the methylation array data across the CH and normal cohorts to assess evidence of differences between the two groups of patients (CNA calling method in Materials and Methods 2.3.2). We see that there are significant CNA burden differences overall, as well as for copy number losses and gains (Figure 5.5C; two-sided Welch's t-test, $P = 8.26 * 10^{-5}$,$P = 0.022$, and $P = 0.00025$, respectively). The average burden per patient across the CH cohorts is 4.7 ± 6.1 CNAs (mean±standard deviation) while normal samples have significantly fewer CNAs at 3.7 ± 3.3 (mean±standard deviation) (Figure 5.5C). However, while the overall burden is significantly different, the greatest difference driving that overall burden is the differences in the number of gains where the average copy number gains per patient is 2.6 ± 4.8 (mean±standard deviation) for CH samples compared to 1.9 ± 2.13 (mean±standard deviation) for the normal samples. While some retrospective, longitudinal studies diagnosed CH based simply on burden of SNVs ($>$ 20 SNVs) [234], this is not the gold standard for diagnosis or validation of CH. Justification for this diagnostic approach was the finding that this high burden of somatic alterations carried an 11.1 hazard ratio for the development of hematologic malignancy in CH patients. To evaluate differences in the CNA in the classified CH samples several analyses were conducted to evaluate additional evidence supporting the classification.

While burden alone may not support evidence of CH, we posit that the distributions of CNA burdens is similar to those seen within clonal mosaicisms [235]. In our data we lack information about subclonal proportions to perform the same analysis; however, given our understanding of what drives fCpG variance (loss of polyclonality in HSCs) we can deduce that subclonal expansions within samples with higher fCpG variances are likely. Studies have shown that chromosomal abnormalities are present in expanded clones at frequencies of 7-95% representing clonal mosaicism, defined

as CNA events with corresponding subclonalities above a threshold [235, 242]. The SNP array methodologies can resolve subclonal admixtures of normal to expanded subclones, something that is not possible using array-based methylation data. However, Jacobs *et al.* [235] analyzed the frequency of detectable clonal mosaic events by age in both cancer patients and cancer free patients. Similar to the age distributions seen for CH, whereby 10-15% of patients present with CH by the age of 70 and 30% by the age of 80 [240], the frequency of individuals with detectable clonal mosaic events increases with age from 0.23% to 1.91% for those under 50 and between 75-79 [235]. The mosaic proportions within this study needed to be greater than 7%. Within our cohort of classified CH samples we see a similar reflection to the age distributions of the patients classified as being CH samples. The frequency of individuals with a presence of CNAs in the classified CH samples increase with age from 0.27% for those younger than 50 to 11.14% for those $>$ 70 years old ($P = 0.0013$, chi-squared test; Figure 5.5D). These results support our FMC-CH classified samples as having CH.

We next performed copy number calls, using our classified normals as the controls, to examine specific differences in genes and recurrent CNAs across different genomic regions that may be implicated in driving CH within the classified CH samples. Using our filtered, high confidence segmentation calls we annotated cytobands, determined genes within aneuploidy segments, and analyzed the genes to determine enrichment in a particular disease area or if overlap exists with known CH or cancer driver genes. On this set of CNAs we see enrichment in several disease classes associated with hematological diseases and malignancies. Of interest, we see significant gene enrichment for genes in regions exhibiting aneuploidy for disease classes related to acute and chronic lymphoblastic and myeloid leukemias as well as various other malignancies (see Materials and Methods 2.3.2 for filtering criteria). In addition, we find 25 known oncogenic drivers associated with recurrent CNA regions exhibiting gains/losses in the CH cohorts (Figure 5.5E). Taken together, with the CNA burden differences and age of individuals classified with CH we add to our evidence that our FMC-CH classifier can accurately identify CH using fCpG sites.

5.7 Clonally heterogeneous landscapes in the hematopoietic stem cell pool

Our model of HSC dynamics, analysis of the CH cohort presented here, and analysis of publicly available data thus far has revealed that fCpG variances reflect underlying turnover and clonal expansions within the HSC pool using peripheral blood. FMCs in

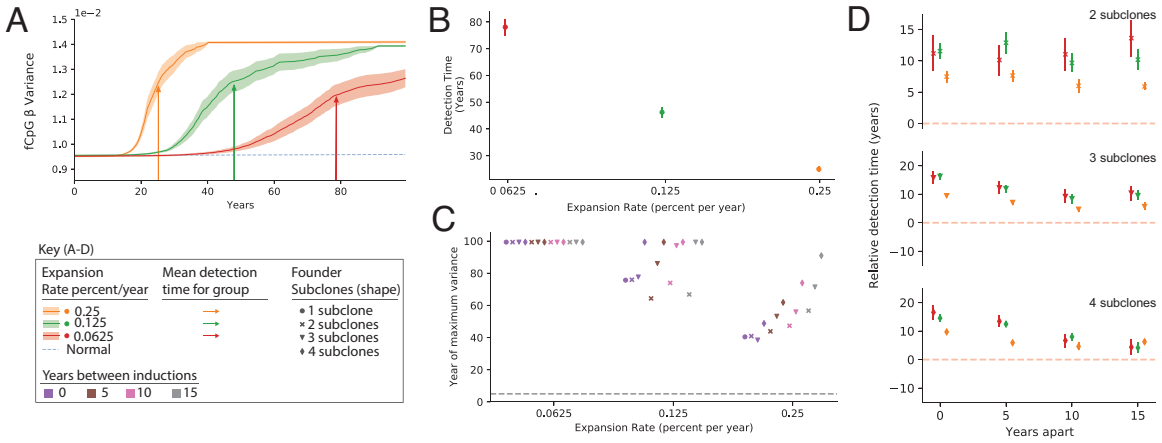


Figure 5.6: Concurrent driver mutations in the model. A: Baseline simulation of a single clone with three different expansion rates (0.0625, 0.125, and 0.25) in an initial population of 500 HSCs revealing the delayed detection time for slower expanding clones. Lines are mean of 20 replicates with ribbons representing the *standard error of the mean* (SEM). Arrows of the same color denote the mean detection time based on classification using the FMC-CH classifier. **B:** Mean detection time for the single subclone simulations in **A** along with the SEM. Plots of fCpG variance over time for simulations with concurrent founding subclones and varied induction times can be found in Figure C.1. **C:** The time to reach the mean steady state fCpG variance increases with slower expanding clones, by spreading out induction times this time is increased further (maximum allowed is 100 years) (steady state determination found in Figure C.2). **D:** Presence of multiple subclones allows for earlier or the same detection time as a single clone, regardless of how far apart the clones are when they emerge. Points show the mean while error bars denote *standard error of the mean* (SEM).

the peripheral blood can be used to diagnose CH; however, it is necessary to evaluate how the make-up of multiple predominant subclones could confound the diagnostic capabilities of the FMC-CH classifier and alter fCpG β distributions. To this end we will leverage our HSC model to further explore the CH data.

5.7.1 Confounding effects of multiple drivers *in silico*

To assess the presence of multiple subclones we first must establish our baseline fCpG variance for a single clone and the corresponding detection times. From the *in silico* validation of the FMC-CH classifier, we showed that given a single clone, the clone's expansion rate is the most important variable impacting CH detection time. A rapidly expanding clone can be detected quickly, but an indolent expansion will take more time to be detected (Figure 5.4D and Figure C.1A). However, in our data multiple subclonal drivers are present (Figure 5.3A) motivating an analysis of the model where multiple clones can emerge at the same time or are acquired at different time points within simulations. We decided to examine the presence of 2 to 4 independent subclonal CH drivers (1 driver population serves as our baseline). We then separated the induction of these concurrent founding populations out by 5, 10, and 15 years. For each of these concurrent drivers and their induction times we evaluated three different expansion rates that are all relatively slow (when compared to those that form a characteristic 'W' fCpG distribution). Expansions of symmetrically dividing populations were halted once they reached 20% of the population (roughly corresponding to 10% VAF for purposes of comparisons, Figures 5.6 and C.1.

For our control, a single subclone induced at year 5 with varied expansion rates, 0.25, 0.125, and 0.625, we see that the mean detection time across the three expansion rates is 20.0, 41.2, and 73.08 years, respectively (Figure 5.6A-B). This relationship is observed during the validation of the FMC-CH classifier as well, where the relative time to detection decreases exponentially as expansion rates increase (Figures 5.4D-E). Comparing the time to maximum fCpG variance, when we introduce multiple concurrent subclones (determined in Figure C.2) with different expansion rates and induction timings, we see that the slowest clones rarely reach their maximum fCpG variances (and thus subclonal frequencies) in the 100-year simulation times irrespective of the induction spacing of multiple clones. Not until expansion rates are doubled, do we see the clones that are inducted at the same time reaching their maximum fCpG variances, when the induction times are spaced out, they fail to reach the maximum allowable frequency (10-15 years induction spacing of 3-4 subclones). For those third and fourth subclonal populations inducted with 15 year spacing means

induction times are 35 and 50 years after the initial subclone, which at the slower expansion rates means they will not expand to an appreciable size by the end of the simulation (Figures 5.6C and C.1).

The detection time for CH, in the presence of multiple subclones, decreases the time to diagnosis (Figure 5.6D). We would expect that the expansion of multiple, concurrent subclones would increase the fCpG variance if expansions weren't proportional across all HSCs. When expansion rates are equal, and thus no functional heterogeneity and competition exists, the most important factors in driving detection time is the expansion rate, number of subclones, and how far apart their induction times are. The presence of two to four concurrent subclones, irrespective of their induction spacing, results in an earlier detection time by 5-15 years. For the largest number of concurrent clones, the expansion rates converge onto no relative difference in detection times at 15 years apart. This emphasizes that detection is driven by those first two clones that were inducted since later clones would have insufficient time to expand to an appreciable size to impact the fCpG distribution seen in normal samples.

We know that the HSC compartment is highly heterogeneous, and we observe multiple subclones within our CH patients. Our model results suggest that there is an additive or multiplicative increase on fCpG variance as the number of subclones increases with various expansion rates. This prompts us to explore the presence of multiple subclones and their relationship with the FMC β distributions that we observe.

5.7.2 Multiple subclonal drivers in the data

Within the CH samples presented, we observe a weak, positive correlation between the largest VAF driver (the gold standard in the clinic for CH diagnosis) with fCpG variance (Figure 5.3B; r-squared=0.23, $P = 2.3 * 10^{-}3$). Based on the results above from the HSC model, we expect that the sum of independent subclones could be an important consideration to re-evaluate the correlation of VAF with fCpG variance. We separated the CH cohort into samples that have multiple and single driver mutations, regardless of their frequency. When we examine the differences in fCpG variances we observe a nearly significant increase in variance for patients with more than one driver mutation ($0.013 \pm 4.0 * 10^{-3}$ mean±SD; n=16) compared to a single driver mutation ($0.011 \pm 2.9 * 10^{-3}$ mean±SD; n=22) (Cohen's d=-0.58; two-sided Welch's t-test, P= 0.089) (Figure 5.7A).

Sequencing data is difficult to perform subclonal deconvolution on without deeper

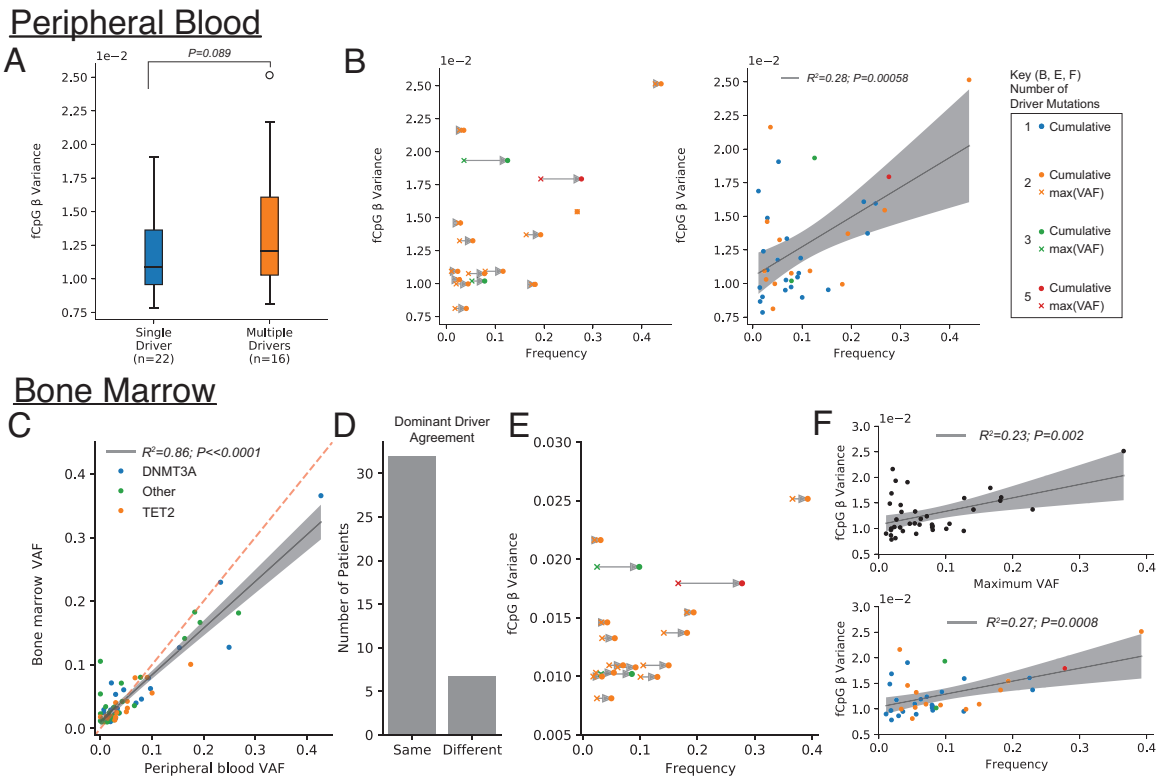


Figure 5.7: Concurrent driver mutations in the peripheral blood and bone marrow. **A** and **B** show mutation results from the PB samples in the CH patient cohort. **A:** CH patient samples with multiple drivers' trend towards a higher fCpG variance (Cohen's d=-0.58; two-sided Welch's t-test, $P = 0.089$; center line, median; box limits, upper and lower quartiles; whiskers, 1.5 IQR). **B**, left: Shows the shift in frequency from the driver subclone with the maximum VAF (e.g., clone used for diagnosis) to the sum of all driver mutation subclonal VAF. The arrow connects these adjusted frequencies. **B**, right: Cumulative frequencies have a stronger correlation compared to the subclonal driver used for diagnosis (Figure 5.3B). **C** through **F** show mutation results from paired BM patient samples in the CH patient cohort. **C:** There is a strong correlation between the frequencies of the same mutation in the PB and the BM (red dashed line shows the location for equal frequencies; linear regression shows $VAF_{BM} \approx VAF_{PB} * 0.73 + 0.01$; r-squared=0.86; P=$3.5 * 10^{-30}$). **D**: The dominant driver differs in a small number of CH paired PB/BM samples. **E:** Frequency shifts between the largest subclonal driver and the cumulative sum of subclonal drivers just as in the PB (**B**). **F:** Just as in the PB results we see a stronger correlation of fCpG variance with the cumulative sum of driver subclones (bottom) compared to only the largest subclonal driver mutation (top). All ribbons show the calculated 95% confidence interval for the linear regression.

whole exome/genome sequencing, and even then, the resolution of low frequency subclones (such as those detected in CH with driver VAFs of 2-10%) is difficult to unravel. Several statistical frameworks attempt to perform this deconvolution [243, 244], but cannot be applied here. Due to this, we can deduce two possibilities for this increased variance for samples with multiple driver mutations. The first being that the driver subclones arose independently and expand with an unknown slow expansion rate. This first possibility is far easier to assess within the data as we can simply sum the VAFs of all observed clones (ignoring the fact that other independent non-'driver' subclones are likely present). When we do this, we see that the cumulative frequency of the subclones for patients with more than one mutation have a significant shift, in what the HSC model would consider to be the final blood frequency corresponding, to higher fCpG variances (Figure 5.7B (left) and Figure C.1). The model results show that even at low expansion rates the presence of multiple subclones expanding leads to a larger fCpG variance. This means that multiple clones won't confound the classifiers' ability to detect CH. We re-evaluate the correlation between the cumulative frequency of the observed mutations and the fCpG variance and we observe a slightly stronger correlation than the single driver VAF (Figure 5.7B (right); r-squared=0.28, $P = 5.8 * 10^{-4}$). The second possibility is that the smaller frequency drivers are acquired mutations in daughter cells of the first clone resulting in a modulated expansion rate conveyed by this second driver. Importantly, this modulation of the expansion rate, as seen in the *in silico* HSC simulations wouldn't lead to a decrease in the fCpG variance.

5.7.3 Paired bone marrow from patients with clonal hematopoiesis

The PB is not the site of HSC, rather, HSCs are primarily located in the axial skeleton BM where hematopoiesis begins with cells eventually committing to their identities (one of several cell types, lymphoid or myeloid terminally differentiated cell) and move out to the peripheral blood (Section **??**). Most measurements within the PB are reflections of events that occur within the BM, but few studies confirm this. Importantly for fCpG distributions, we would like to accurately capture the underlying distribution of the HSCs from the terminally differentiated PB. For the fCpG β distribution to accurately reflect what is seen in the BM, individual subclones would need to give rise to equal proportions of various blood cell types. This is still a wide open research area, but one study found that the production of blood is highly polyclonal, deriving from a large number of HSCs [245]. In our study we have paired

methylation data derived from PB with corresponding DNA sequences to evaluate the presence of CH subclones. Here we also present the paired BM DNA sequencing results to examine whether PB or BM subclones are of equal sizes and whether their subclonal frequencies correlate with fCpG variance (Figure 5.7C-F).

The presence of paired BM provides several sanity checks for both the interpretation of the PB findings as well as the fCpG variances. We see that all variant allele frequencies are similar between the BM and PB (Figure 5.7C). We also see that as the frequency of the CH driver mutations increase in the BM (presumably the oldest subclones) they are slightly higher in the PB, but a strong correlation exists (r-squared=0.86; $P = 3.5 * 10^{-30}$). We evaluated whether DNMT3A or TET2 mutations saw a departure from this correlation and no significant difference was observed across mutation types (r-squared=0.91 and 0.77 with P-values $\ll 0.0001$, respectively). Beyond this correlation, pairwise comparisons across patients reveal there is no significant differences between the variant allele frequencies of the driver subclones between PB and BM (Wilcoxon signed-rank test; $P = 0.36$). For some of the dominant driver mutations (used for diagnosis) we see slight differences in their VAFs, but these differences are often within an expected margin of technical error from DNA sequencing (Figure 5.7D; Table 2.2).

We next wanted to evaluate how well the fCpG β distribution correlates with the VAF of the putative drivers from the BM to see if similar correlations are found as those seen in the PB (Figure 5.7B). When we perform the same correction as applied to the samples with multiple subclonal drivers to evaluate the correlations we see the same increased correlation when taking the cumulative sum of putative driver VAF (Figure 5.7F).

5.7.4 Single cell sequencing of clonal hematopoiesis patients confirms subclonal structures

Our understanding of CH, aging of HSCs populations, and accumulation of mutations within normal tissues resulting in mosaic tissue suggests that we are likely to observe a vast collection of distinct populations of HSC subclones. Some of these subclonal populations ought to have acquired the necessary drivers for CH with their likelihood increasing as an individual ages. This prompted our analysis of concurrent driver clones as it relates to our ability to detect CH with our FMC-CH classifier (Figure 5.6) and the cumulative frequency correlation with fCpG variance in the data (Figure 5.7A-E). However, based on our understanding of multi-stage tumorigenesis it is highly likely that some of the observed driver mutations present in our CH samples

with multiple drivers are nested and acquired at some point within the parent driver's lineage. With high resolution whole exome/genome sequencing we could potentially deconvolve the subclonal structure of these mutations [243, 244, 246], but these are low frequency mutations that would be difficult to disentangle from one another. Instead, single cell sequencing offers the best resolution currently to reliably uncover the subclonal structures of these mutations to determine if a frequency adjustment is necessary to examine correlations with fCpG variance.

We performed single cell sequencing on four of our CH samples who had drivers with VAFs $> 5\%$ and showed presence of a DNMT3A or TET2 mutation (NOC062, NOC137, NOC115, and NOC131). Fortuitously, two of these samples exhibited subclonal structures that were nested (NOC062 and NOC137), while the other two were concurrent CH subclones (NOC115 and NOC131) (Figure 5.8A). Interestingly, we see a relationship within these four samples where there appears to be an unknown threshold where the frequency of a subclone is larger in the bone marrow while the mutation is small enough, while it is larger in the PB if it is a larger clone. Unfortunately there are too few samples to draw any definitive conclusions here. However, this could indicate that the clones that have been around longer have given rise to cells with longer half-lives, persisting in the PB rather than cells with rapid turnover. This could occur for an indolently expanding clone, that arose early and has a low population frequency and thus a low fCpG variance, or for a clone with a moderate expansion rate that arose some time in the not too distant past; however, in turn this would result in a higher fCpG variance which occurs for larger cumulative subclones or faster expanding clones (in the model this is shown in Figures 5.2A and 5.6A-D; for the data we illustrate this with our correlations in Figures 5.7). PB cell half-lives once they have fully differentiated, exiting the BM, are all relatively short with lineage tracing studies revealing half-lives between 12 hours for monocytes and 4-7 weeks for B cells [247–249].

5.7.4.1 Modeling nested subclonal expansions

So far, we have presented model simulations with distinct subclonal populations with varied expansion rates, various years between mutation induction, and numbers of subclones. However, given that nested structures exist within the data it's necessary to also simulate the occurrence of multiple subclones with one originating as the daughter from the initial driver population as confirmed in two of our samples. Here we initialize the first clone at five years with the slowest expansion rate of 6.25 percent per year. Unlike the previous simulations we introduce a new founder clone

as a daughter of the initial subclone at 60 years (the timing of this second subclone was chosen arbitrarily, but it needed to occur early enough to be seen by the 100-year stopping point of the simulations). There are numerous approaches to determining fitness gains conveyed by the acquisition of passenger mutations and additional drivers acquired within a subclonal population, but here we assume that a modest fitness increase is conveyed as a faster expansion rate for that subclonal population. Where its expansion rate increases 200% (2x) to 12.5 percent per year (the moderate expansions seen in previous simulations). We permit these subclones to expand until the driver reaches a final population frequency of 20% (Figure 5.8C).

We see that the nested subclonal causes an increase in the fCpG variance over the control where a single subclone grows at a steady 6.25 percent per year (Figure 5.8E). We see that the frequency of the initial population diverges ten years after the introduction of the 2nd driver with an appreciable growth of the 2nd clone between induction and reaching its final frequency. Qualitatively, these frequencies are approximately similar to those observed for patient NOC137 whose frequency of the primary driver VAF is 0.16 for ASXL1 and its nested driver VAF, TET2, is 0.041 with a corresponding fCpG variance just under 1.5 (Figure 5.8C-E). This suggests that having a known frequency and corresponding fCpG β distribution could be valuable in providing insights into expansion times and clone inductions.

5.8 Discussion

Through the work presented within this chapter we have orthogonally validated FMCs to show that not only are fCpGs found in the crypts of the colon, small intestine, and endometrium, but they exist in HSCs as well. Large numbers of fCpG sites reversibly switch their methylation status like an erratically swinging pendulum between 0%, 50% and 100% (representing homozygous and heterozygous (de)methylation). In the polyclonal populations, exemplified by HSC, fluctuations are unsynchronized between individual cells and fCpG methylation is saturated at 50%. This is distinct from clonal populations that form the characteristic W-shaped distribution shown in Chapter 4 with modal peaks at 0%, 50% and 100% methylation.

Within the scope of hematopoietic cells we show that fCpG dynamics are present and useful to reconstruct clonal dynamics. The identity of the fCpG sites in hematopoietic cells differs from those in the epithelium, likely reflecting that fCpGs tend to be found within non-expressed genes and the fact that gene expression patterns vary between tissues. We illustrate the ability of fCpG sites to be used to de-

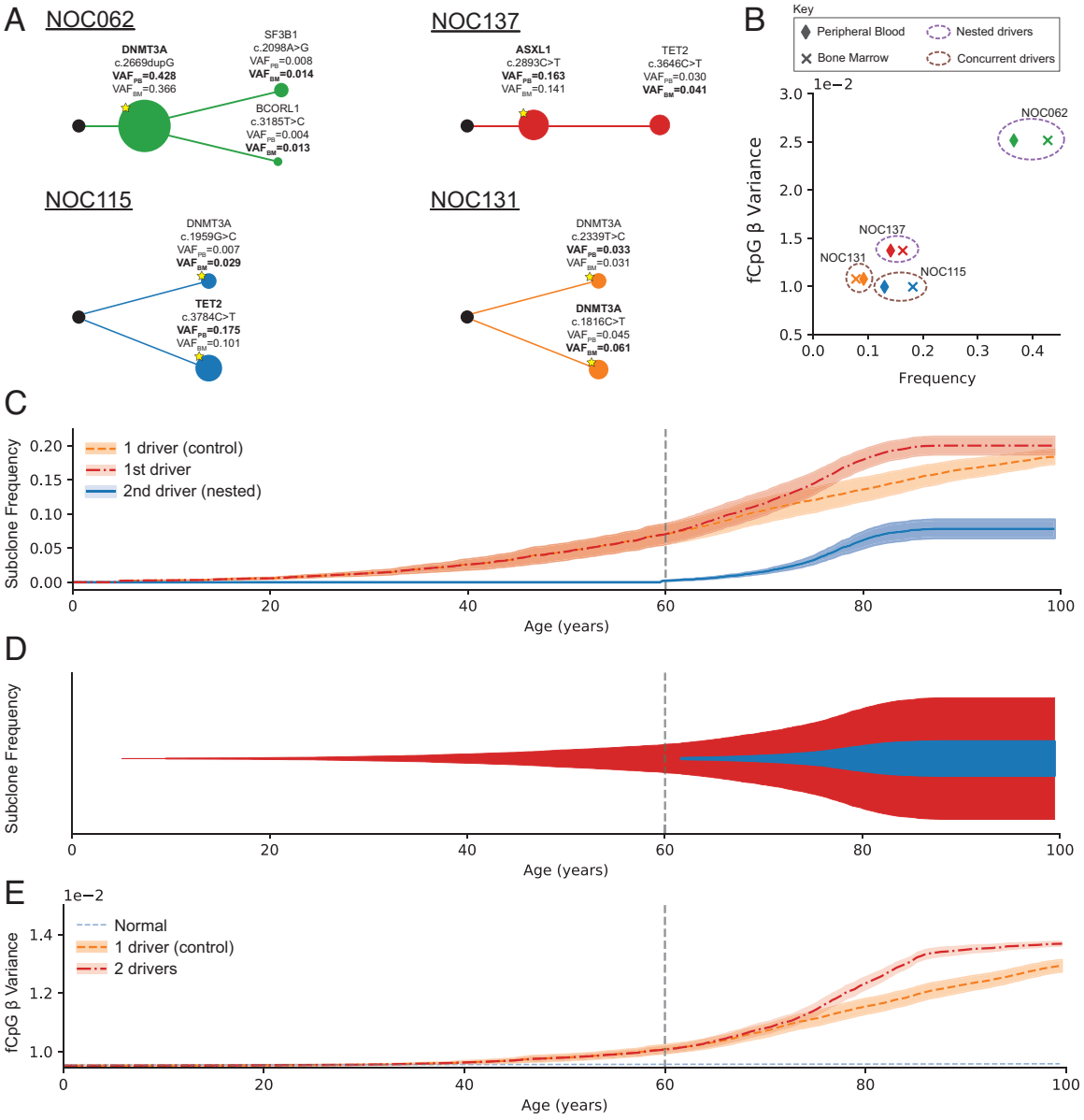


Figure 5.8: Single cell sequencing confirms subclonal composition. Single cell sequencing conducted on four samples to confirm subclonal composition and nested structure of mutations. **A:** Subclonal compositions confirmed for four patient samples from single cell sequencing for two patients with nested subclonal drivers (NOC062 and NOC137) and two patients with concurrent drivers (NOC115 and NOC131). **B:** Corrected frequencies with confirmed subclonal structures, where the largest driver VAF for nested drivers and the cumulative sum of frequencies for concurrent drivers show correlation with fCpG variance for both BM and PB (stars in **A** indicate frequencies for each sample used). **C-E:** Results when a nested subclone conveying an increased expansion rate is considered for a clone induced at 5 years and a nested subclone induced at 60 years doubling the expansion rate. **C:** Population size frequencies for each clone independently with the control present for a single clone. **D:** Population frequencies using a Muller plot [196] showing the lineage of the second subclone. **E:** fCpG variance for a single clone compared to the clone with a nested, subclonal driver.

tect CH through our HSC model of symmetrically expanding subclonal populations corresponding to an increase in average fCpG variances with clonality and characteristic W-shaped distributions present in acute leukemias (Figures 5.2A-B). Chronic leukemias had intermediate fCpG variance increases and generally lacked W-shaped distributions, likely reflecting their slower growth.

Development of HSC pools ends how they begin, more homogeneous in children and elderly. HSCs begin relatively homogenous during embryogenesis and expanding for the first two decades ($<$ 20 years) giving rise to a highly heterogeneous mixture of HSCs before losing some HSCs through aging [169]. In acute leukemia cohorts of older patients versus pediatric patients we see that the latter exhibit higher fCpG variance likely reflecting differences in initial HSC dynamics. At the other end of the life span, where HSCs diminish in numbers and exhibit an age related incidence of CH we see fCpG variances increased [225].

Due to the age-related incidence of CH it is important to consider how important CH may be in the development of malignancies. Given that across most tissues in the human body (including the ones presented in Chapters 3 and 4) somatic alterations are ubiquitous, and may not pose much risk to the further development of malignancies, it is paramount to be able to tease apart what underlying dynamics are clinically relevant to future malignancies and which are simply a part of human aging. To this end we show that we can leverage FMCs to determine the extent of CH. The approach developed here adds significant value in that it is more sensitive to the underlying dynamics than the current gold standard based on variant allele frequencies. FMC have the added value of being agnostic to the underlying genomic alterations for the purposes of diagnosing CH; further, they are more sensitive to faster expanding clones or the presence of multiple subclones with malignant potential.

Chapter 6

Conclusions

6.1 Summary of results

In this thesis I have presented three studies that integrate clinical omics data with mechanistic models across different tissue types to better understand somatic evolution, progenitor cell dynamics, and early departure from homeostasis. The first project utilized publicly available data and contributes to the paradox of oncogenic SNVs expanding in normal stratified squamous epithelium (the epidermis) driven by underlying stem cell turnover. In the next two chapters I turned my attention to measures of stem cell dynamics and homeostasis using methylation, first in another epithelial tissue (intestinal crypts) and then on connective tissue (blood). In more detail:

- **Chapter 2:** We highlighted the methods used throughout the results chapters excluding the mechanistic models that were developed as an integral component for each results chapter individually.

- **Chapter 3:** We developed a first principles model of the homeostatic epidermis, complete with cells containing base-pair resolution somatic mutations using our novel tool developed as part of this chapter, Gattaca [191]. With this model, we examined neutrally expanding subclonal populations and compare simulations to high resolution sequencing of spatially defined patient epidermis samples. We then explored mechanisms for how non-neutral dynamics in NOTCH1 and TP53–genes that are considered to be oncogenic drivers–are limited by tissue homeostasis.

- **Chapter 4:** We uncovered FMCs by developing a spatial agent-based homeostatic crypt model, complete with methylation markers in each *in silico* cell. We

showed how information from this methylation can be used as lineage tracing for stem cells in the crypt. We then determined criteria for identifying fCpG sites in the intestinal crypts from multiple single crypts in the colon and small intestine of 16 patients. Lastly, using a mathematical framework and inference method validated with our spatial model, we revealed stem cell dynamics of several patients, including colon crypts of FAP patients and endometrium glands.

- **Chapter 5:** Expanding on our knowledge of FMCs, we explored whether they could be a marker for loss of polyclonality in the HSC compartment. We first developed a simple model of HSCs and explored the theory of expanding subclonal populations. We examined thousands of patients publicly available methylation data revealing that HSCs can serve as a marker of HSC clonal expansions. This is the case in normal turnover to highly monoclonal hematological malignancies such as acute leukemias. We then analyzed a first of its kind paired mutation and methyaltion data of a cohort of CHIP patients with varying degrees of CH. Using this cohort of CHIP patients we developed a mutation agnostic diagnostic machine learning approach based on FMCs.

6.2 Discussion

6.2.1 Somatic evolution in homeostatic tissue

Somatic mutations spread in a population by arising in a long lived stem cell lineage. In each of the chapters, mutations arise by chance events and then spread based on rules that govern the cells turnover. In the epidermis, most subclonal 'patches' of clones are small and are lost with that cells lineage (Figure 3.2). In the intestinal crypts, the only cells that can give rise to mutations able to become fixed in the crypt are the limited number of stem cells. Monoclonal conversion, where a single population sweeping through the entirety of the crypt, takes an average of 7-8 years (Figure 4.6). Homeostasis and stem cell turnover inherently limits the spread of mutations in these long lived stem cells across both of these tissues in the absence of functional heterogeneity. The spatial architecture with which these cells are restrained aids in the ability for one single mutation to dominate. When a single mutation does manage to dominate, our results suggest that it can do so by one of two methods. The first method is persisting, allowing the cell to divide normally, but failing to be lost and replaced by a neighboring cell. This approach maintains a longer lived stem like

cell able to divide and become larger than its non-persisting neighbors. This has been shown in intestinal crypt where APC mutant cells secrete WNT antagonists and cause neighboring, non-mutant cells to differentiate [250, 251]. The second approach is for a cell to divide faster than its neighboring cells. This is one method that is expected for sporadic APC mutations within the crypt as indicated by the faster replacement rates observed in FAP crypts that already have one APC mutation. Through a comparative study of intestinal crypts across numerous mammals it was revealed that mutation burden is the same at the end of a mammals life regardless of body mass, life histories, and lifespan [252]. This indicates that somatic mutation rates are evolutionary constrained and likely a contributing factor in aging. In the case of FAP vs sporadic colorectal cancers this helps to explain the observed differences between acquisition of necessary drivers and timescales of when an FAP patient develop begins presenting with polyps in the clinic versus the non-FAP patients where sporadic APC mutated polyps are present at about twice the age of FAP patients (mean time for presentation of polyps is about 32 years old). In the blood, similar to the crypt, we see the accrual of mutations over an individuals lifetime, but the large pool of stem cells in this system prevents a single clone from ever dominating. In the homeostatic hematopoietic system, committed progenitors (myeloid and lymphoid progenitors) further aid in making sure that a single lineage doesn't dominate. In this work, we leveraged somatic evolution to understand stem cell dynamics and homeostasis and developed several new, impactful methods that can be extended for future studies.

6.2.1.1 Somatic mutations and mechanistic models

As part of the work presented here we introduced Gattaca, the first base pair resolution mutation induction and tracking *in silico* tool to model genomes within agent based models. Gattaca provides a powerful tool to induce and track mutations through time and space to compare with patient and murine samples. Gattaca provides the necessary genomic underpinnings to evaluate 'null' simulations that are capable of explicitly modeling neutral somatic mutation within spatial contexts that the user defines. We have demonstrated this by comparing the genomes and clonal dynamics that Gattaca provides across different modeling dimensions and model choices. We then show, through a second use case, that wounding can show evidence of selection, but only in the $2D$ wounding case. This sets an important precedent that modeling choices around dimensionality can significantly impact the measures of neutrality.

Gattaca at its heart is a simple tool built around user defined parameters that

assumes mutations incurred are heterozygous, impart no fitness advantages to the cell, and can be easily deployed within most HAL simulations. This simplicity also allows for flexibility, in that users can evaluate CNAs *post-hoc* with relative ease and introduce fitness advantages through careful consideration. Genotype-phenotype mapping is highly specific to an individual research question. If users decide to use Gattaca to introduce a genotype-phenotype mapping, great care should be taken to understand Gattacas' limitations and the complexities/assumptions involved with introducing functional heterogeneity. Although Gattaca was built for and tested using the HAL agent based modeling framework, with some effort, one could write a set of wrapper functions so that Gattaca could be used across different modeling frameworks.

Gattaca provides a highly customizable framework that is easily implemented into users' agent based simulations for evaluating somatic evolution in normal or disease tissue. Through the incorporation of common bioinformatics and genotypic outputs (variant call format files) used frequently in clinical and experimental approaches users can quickly analyze and compare mutation spectra, burden, heterogeneity, and selection between their samples and *in silico* models.

6.2.2 Clinical insights and considerations

In the opening paragraphs of this work I mentioned that the clinical relevance of basic research is often overstated (Section 1.6). While I believe this to be true, the contributions of basic research to providing future clinical insights is powerful and can often be explored by another research team with a different expertise or through future inquiry. There are important findings and future directions in each of the presented projects.

In the case of the epidermis we show that, in homeostatic tissue, oncogenic mutations (e.g., NOTCH1 and TP53) must arise early in order to expand to observed clonal frequencies in patients. We also know that various non-melanoma skin cancer incidences increase with age after lifetimes of sun exposure. Together, this indicates that mutations, while a necessary precursor, are insufficient to be the sole drivers in the transition to dysplasia. This also supports the idea that the most important time to guard tissue from ultraviolet damage from the sun (i.e., wear sunscreen) is generally in the first half of ones life. The homeostatic epidermis model reveals that it may be possible to infer sun exposure over a lifetime by examining clone sizes, mutation types, and spectra. The resolution of this approach would be determined by the resolution of the longitudinal data set that would have the information necessary.

When considering clinical contributions and insights in Chapter 4 we see that most of the findings contribute to a theoretical understanding of stem cell behavior in the crypt and glandular tissues. The most important part of this work is the development of the FMC theory to serve as *in vivo* barcodes that can be used on relatively short timescales. At the genesis of this project we had hoped to answer questions about why the small intestinal cancers are between 14 to 50 times less likely to occur when compared to the colon [253, 254]. The number of crypts and accumulated mutations with age are the same between the colon and small intestine [81]; therefore, the differences in cancer incidence must be explained by observable differences in stem cell turnover and their numbers. This was proposed as the "bad luck hypothesis" [255]. However, we find that there is only 16% more stem cells in the small intestine crypts compared to the colon and their turnover is the same, suggesting that stem cell dynamics alone are insufficient to explain the differences in cancer between these two tissues. We did provide confirmatory evidence that first hit somatic APC mutation, such as those in sporadic colorectal cancers, leads to a fitness advantage with increased division since FAP samples exhibit higher replacement rates.

In the case of clonal hematopoiesis (Chapter 5), there are much clearer clinical contributions worth highlighting. At present current clinical guidelines state that it is not advisable for the general population to diagnose CH and that regular screening is unnecessary. This is critical guidance for clinicians overseeing an aging population fraught with ailments of more pressing concern that do have clinical interventions. It would be unwise to diagnose a condition with a large amount of prognostic uncertainty and few options for clinical intervention. That being said, it may be only a few years before prognostic markers are found that can determine which CH cases are those likely to progress to some form of malignancy. When combined with mutation and copy number data yielding a clone size, FMC could prove to be a powerful prognostic marker. There is more work to be done to determine the full extent of how FMCs could be leveraged to understand the hematopoietic stem cells and disease.

6.3 Future directions

I have been told that many theses result in a clear direction forward for future inquiry that can make it challenging to determine where to end the work that becomes the body of the thesis. I am not an exception to this experience and the last few months of my thesis I found this to be the most challenging part. Through discussion I decided that the work that I would have liked to include in Chapter 5 would be best

outlined in this forward thinking section. Before discussing the future directions of that chapter I want to highlight work that I think is an important continuation and use of the homeostatic model of the epidermis.

Research examining stratified epithelium in the oesophagus, similar to tissue seen in the epidermis, examined clonal dynamics and competition using murine models and a 2D model of normal cell turnover [30]. They also highlight the importance of spatial competition in homeostatic tissue in shaping the observed clonal dynamics. The oesophagus is not routinely exposed to irradiation, but does suffer regular chemical and thermal insult (by way of carrying food and liquid). In the case of the epidermis, a murine model that examines irradiated epidermis and growth dynamics of TP53 in a longitudinal *in vivo* study would serve to validate the findings of those in Chapter 3 where we introduced non-neutral dynamics in the epidermis. In the context of carcinogenesis, we stopped just short of attempts to model cancer initiation as the data lacks any information on this critical step and was beyond the scope of the primary aims. However, this would be an interesting application of the homeostatic epidermis model.

6.3.1 Future direction for fluctuating methylation clocks

The development of FMC and subsequent application in Chapter 4 directly contributes to the future direction of how FMC were applied to orthogonally validate the 'W' shaped distribution found in clonal cell populations at the beginning of Chapter 5. While there are many tissues whose stem cell dynamics could be examined using FMCs I think the most clear direction to move forward is in the context of understanding hematopoietic stem cells and hematological disease diagnostics, prognostics, and monitoring.

6.3.1.1 Inference of clonal hematopoiesis expansion rates and induction times

In the endometrium and crypt, the defined algorithm for examining fCpG status is simplistic in nature, but computational power and time increases exponentially with the number of stem cells and fCpG sites under consideration. While the number of HSCs is unresolved, we know it is more than 10^3 with estimates ranging between $\approx 10,000$ [256, 257] to $\approx 200,000$ [36, 80]. Together, this makes the analytical model of directly simulating methylation distributions (Section 4.5) unsuitable for assessing FMCs in HSCs. A novel approach is necessary in order to determine dynamics of the HSCs that is computationally feasible and complex enough to capture dynamics of the

fCpGs and HSCs. Our approach in Chapter 5 uses a non-spatial agent-based model of HSC turnover with fCpG undergoing (de)methylation. In order to perform inferences we need a model quick enough to execute millions of simulations. To satisfy these requirements we developed a comparable model using a Markov Process to determine the state of an allele where a single loci can have no alleles, one allele, or both alleles being methylated (Figure 6.1A). We can initialize the number of stem cells with fCpGs in the same way as our original model, but separate the normal HSCs (N_H) and mutant population (N_C). Our normal population is dividing asymmetrically, therefore we only need to initialize and then determine the state of our fCpG loci for each timestep so that when the mutation is induced we can copy the state of the fCpGs in order to simulate our expanding population carrying CHIP driver mutations. Assuming exponential growth, we can couple the induction time, T, and expansion rate, E, in a straightforward manner. Assuming a clone (N_C) arises at time T and grows exponentially with an expansion rate E, then its population size follows:

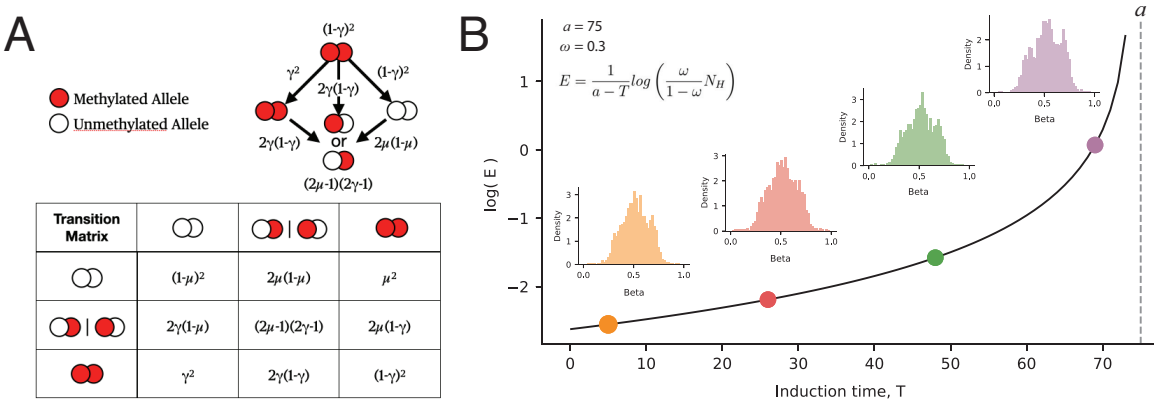


Figure 6.1: Clonal hematopoiesis expansion rates and induction times. A: describes the transition state matrix for the Markov Process governing the transition of methylation status for a loci with two alleles where μ is methylation and γ is demethylation. **B:** Given equation 6.7 we can visualize the differences in fCpG distributions as output from the new CH model where the oldest, most indolent clones expand slowly but arise earlier, compared to those with higher fCpG beta variance corresponding to a faster expansion rate and more recent induction time.

$$N_c(t) = N_c(t)e^{E(t-T)}, \tag{6.1}$$

so that at the time of induction $N_c(t) = N_c(t)$ and at age, a, of a given patient:

$$N_c(a) = N_c(T)e^{E(a-T)}. \tag{6.2}$$

We know from our simulations that a clone arises from one cell at the time of induction,

$$N_c(T) = 1 \tag{6.3}$$

and that the final frequency of a clone at the sample time (i.e., age of the patient), a, is given as ω, so that

$$\omega = \frac{N_c(a)}{N_H + N_c(a)} \tag{6.4}$$

With this knowledge we are able to constrain the possible expansion rates that satisfy the above conditions. Solving equation 6.1 for E, we get:

$$E = \frac{1}{a - T} log \left(\frac{N_c(a)}{N_c(T)} \right). \tag{6.5}$$

We know most variables in equation 6.5 except for $N_c(a)$; however, we can derive $N_c(a)$ from equation 6.4:

$$N_c(a) = \frac{\omega}{1 - \omega} N_H. \tag{6.6}$$

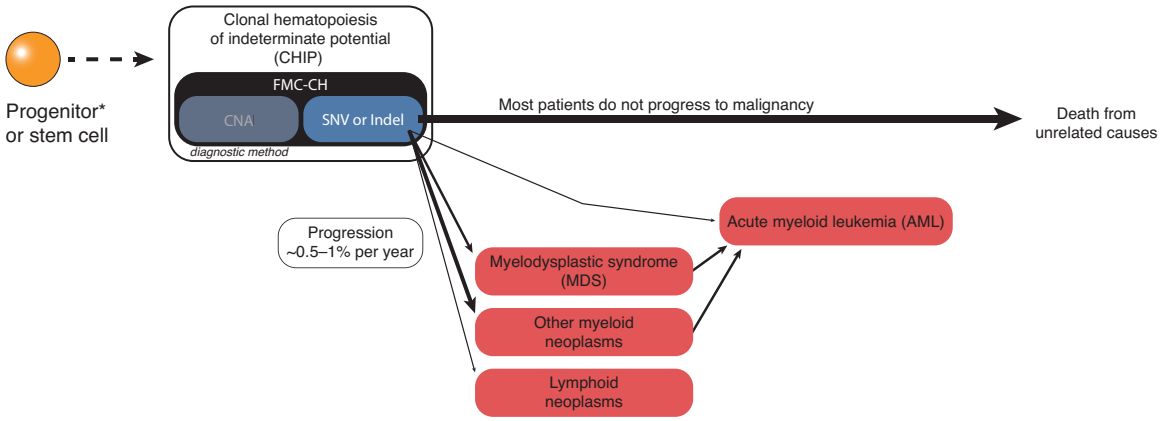
Substituting this into equation 6.5, we get:

$$E = \frac{1}{a - T} log \left(\frac{\omega}{1 - \omega} N_H \right). \tag{6.7}$$

Importantly, similar to the fCpG variance, as the number of stem cells considered, N_H, increases the values of E and T combinations converge and becomes insensitive to the total number of stem cells allowing us to focus and constrain our inference priors on the two primary variables we are interested in determining. For a given clone frequency and a patients age, we have a set of points that provide possible (E,T) pairs. A basic rejection sampling inference framework is sufficient to resolve these parameters for each of our CHIP patients presented in Chapter 5. When we examine the corresponding fCpG distributions for equation 6.7 we see that variance is higher for the fastest expanding clones (Figure 6.1B).

The mutation induction time in CH is still a widely open question, as are the clones corresponding expansion rates, with a few studies currently examining the timing of DNMT3A and TET2 mutations in CHIP [239, 258]. Using whole genome sequencing these CH clones have been inferred to expand at a median of 26% per year and arise decades prior to detection [258], consistent with results in Chapter 5 and initial inference attempts using the model described here (not shown).

6.3.1.2 Progression of clonal hematopoiesis to a malignancy



* Progenitor could be one of a myeloid progenitor, lymphoid progenitor, or a multipotential progenitor

Figure 6.2: Current progression paradigm for clonal hematopoiesis. If clonal hematopoiesis of indeterminate potential is observed for an expanding progenitor or stem cell it is currently diagnosed through targeted sequencing of select oncogenic drivers. Of those confirmed cases of CHIP the majority will never progress to a malignancy. The rate at which those with CHIP progress to MDS, AML, or other malignancies is $\approx$ 0.5–1% per year. The relative proportion of CHIP cases that progress to each of the different malignancies is shown by the arrows linewidth. Figure adapted from Figure 1B of Steensma *et al.* [230].

In the case of clonal hematopoiesis, across various patients, there is no clear predictor of which CH cases will progress to a malignancy. CH could, but rarely does, progress to one of many different types of malignancies involving myeloid or lymphoid cells diagnosed through molecular testing (i.e., DNA sequencing) and cytomorphogenetic analysis (Figure 6.2). The most broadly characterized hematological malignancy that exhibits incredibly similar mutational characteristics to CH is MDS. Low risk MDS is considered to be cases with possible cytopenia, $>$ 10% VAF of more than one oncogenic mutation, and little to no dysplastic morphologies [230]. The drivers of transitions between malignancy beyond mutation are poorly understood with proposed driving mechanisms spanning CNAs, addition of oncogenic SNVs, or methylation. While we have yet to uncover these dynamics, FMC can serve as a powerful tool to determine progression, as it captures the behavior of HSCs and progression to various hematological malignancies or diseases, resulting in clonal homogenization. This could be true for many hematological conditions and clinical applications.

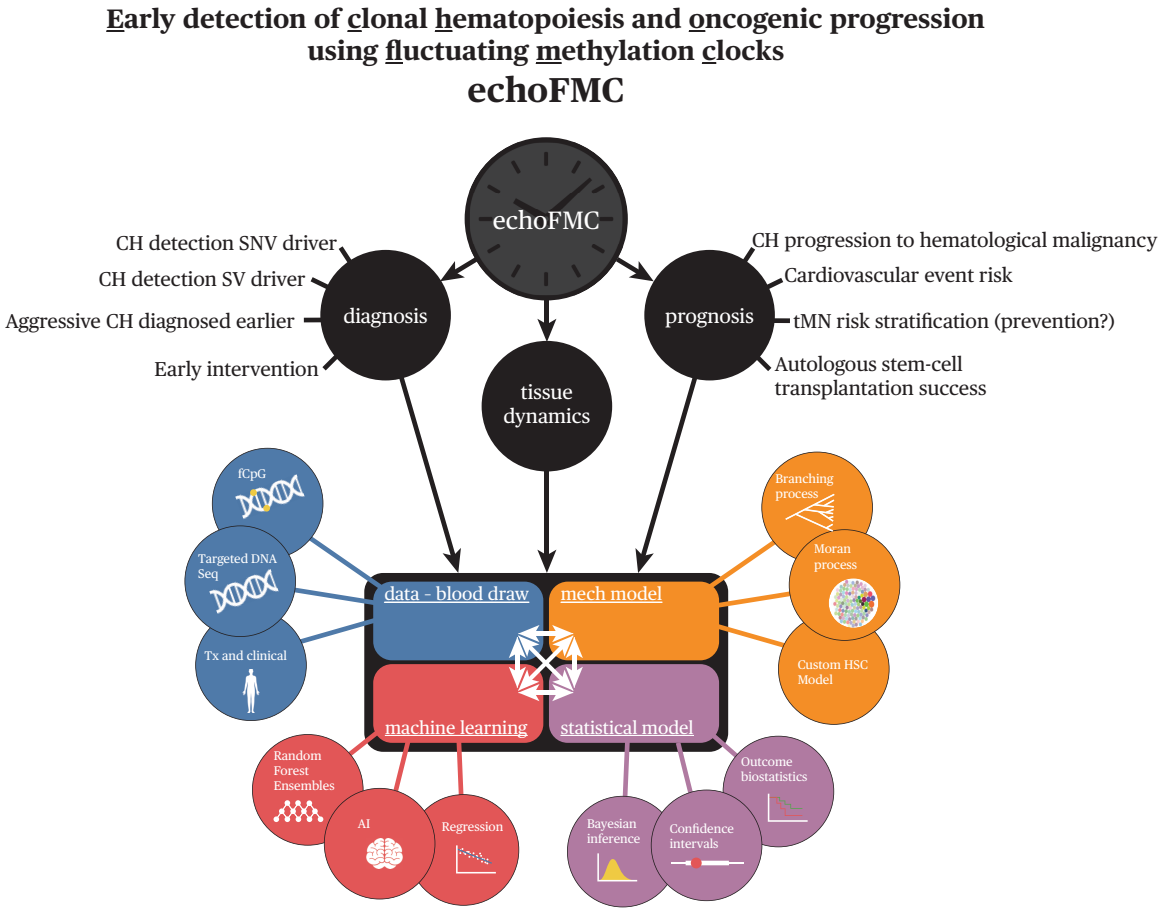


Figure 6.3: Early detection of clonal hematopoiesis and oncogenic progression using fluctuating methylation clocks. Infographic outlining a future direction to fully investigate early detection of clonal hematopoiesis and oncogenic progression using fluctuating methylation clocks (echoFMC for short). Similar to the work in this thesis, echoFMC would integrate high resolution data on clonal dynamics and leverages machine learning with mechanistic and statistical modeling to provide diagnosis, prognosis, and HSC dynamics.

6.3.1.3 Early detection of clonal hematopoiesis and oncogenic progression using fluctuating methylation clocks

Thinking beyond the work here, I propose a future area of inquiry where FMCs, combined with additional genetic information (i.e., mutation and CNAs), can serve as a critical evolutionary biomarker that captures HSC dynamics in numerous applications. A large portion of the necessary groundwork to pursue such a biomarker has been completed through Chapters 4 and 5. Here we have already shown how FMC can diagnose CH. Through pairing methylation data with SNP information (SNP data is already the gold standard for CHIP diagnosis) we have outlined a way in which we can infer the growth dynamics of expanding clones. If this information is

combined with longitudinal datasets for patient outcomes, FMCs could be useful in determining truly indolent, low risk clones, from those more aggressive and likely to progress.

However, CH is not only associated with hematological malignancies. CH has also been associated with an increased risk of all cause mortality, especially an increased risk of stroke and epithelial cancers [151, 259]. Understanding these relationships and being able to assess risk based on CH and the extent of HSC homogenization may also be a powerful tool for understanding ones own health as a consumer. Inflammation has been linked to CH as an important correlative for poor outcomes (reviewed by Cook *et al.* [260]). If consumers understood their risks they may be inclined to incorporate healthier diets, added stress relief, and avoid alcohol (i.e., strive for an overall healthier lifestyle that has been shown to aid in reduced systemic inflammation). Finally, we are beginning to understand how *therapy-related myeloid neoplasm* (tMN) occurs in patients that exhibit CH at the time of treatment of solid tumors through cytotoxic chemotherapy or radiation therapy. Risk of tMN may be assessed by the monitoring of HSC via FMCs.

It is not beyond the scope of reality that the outlined approaches could be pursued. Especially given long term studies that already have mutation, methylation, and health outcome data for hundreds of thousands of patients. Some of these consortia that possess the appropriate datasets include the European Prospective Investigation into Cancer and Nutrition (EPIC), InChianti, The Framingham Study, UK Biobank, Finnish Twin Study, and FINRISK. By curating these studies to examine only those with methylation and SNP data prior to CHIP, cardiovascular events, or the development of any cancers the needed datasets may already exist. These long term outcome studies paired with molecular data would provide the necessary longitudinal datasets needed to explore the prognostic capabilities of FMCs from blood. For future patients, the use of a customized approach to collect a patients fCpG methylation data, targeted sequencing, and CNA information could be used to replace the current gold standard providing higher resolution data at a cheaper price providing better clinical insights to clinicians and greater benefits to the patient.

The proposed research outlines a future direction to fully investigate early detection of clonal hematopoiesis and oncogenic progression using fluctuating methylation clocks (echoFMC for short; Figure 6.3). Similar to the work in this thesis, echoFMC integrates high resolution data on clonal dynamics and leverages machine learning with mechanistic and statistical modeling to provide diagnosis, prognosis, and HSC dynamics.

6.4 Closing remarks

There are two common ways that a biomedical cancer researcher can contribute their efforts to the cure and treatment of cancer. One of these approaches is heavily prominent throughout this body of work. All of that which is presented herein, alongside several other recent high profile publications cited throughout this thesis, provides critical insights into how integration of bioinformatics, mechanistic modeling, experimental biology, and clinical data collection is a necessary and critical way forward to fully understand evolutionary dynamics in somatic tissue. This integrative approach requires new, forward thinking, team science initiatives, capable of supporting this research and training of new interdisciplinary scientists and data driven clinicians. Through these initiatives, these individuals, trained in bioinformatics and mechanistic modeling, would integrate with computationally literate experimentalists and data driven clinicians, providing novel insights previously unreachable.

I sincerely believe that through understanding homeostasis and normal tissue we can gain better insights into tumorigenesis, diagnosis and early detection, and prognosis. Eventually these insights will yield therapeutic interventions that will target the earliest stages of tumorigenesis. Leveraging somatic evolution to understand stem cell fate dynamics is a critical step in unraveling the mysteries of homeostatic tissue and the evolutionary dynamics that stabilize and maintain healthy tissue.

Appendix A

Appendix: Chapter 2

A.1 Gattaca Case Studies

A.1.1 Gattaca Case Study 1: Dimensionality

Gattaca allows us to track base pair resolution genomes across any agent based modeling dimension. Recent interest in understanding how spatial architecture may affect clonal dynamics and measurements of neutrality motivates our case study [26, 31, 92, 108].

Here we have constructed two simple agent based models (ABM) of cell turnover in three different dimensions, 0D, 2D, and 3D, to showcase and compare the mutational profiles and clonal dynamics that Gattaca allows its users to evaluate. In addition, we perform the simulations for these three dimensions and two model types for three different total final population sizes to demonstrate the functionalities and outputs of Gattaca (Figure A.1). The two model types differ only in the number of cells that are present at initialization. The fully seeded model initializes by placing an agent with its unique genome at every lattice point or until the carrying capacity is reached in the 0D case. The second simple ABM is initialized only with a single cell at a random position within the simulated domain, or simply a population size of one for the 0D case (Figure A.1). These two simple model types can be conceptualized as a naive tissue type of model to compare with a stem cell growth model similar to the idea that cancer originates from a single transformed clone. Here we introduce no functional heterogeneity across the different genomes that emerge through mutation at each timepoint governed by the conditions set in Gattaca.

Within the two models we use the same parameters so as to be able to more accurately compare across the different dimensions. Each model across all dimensions uses the same birth/death function. The birth rate (λ, $\lambda = 0.4$) is scaled by the

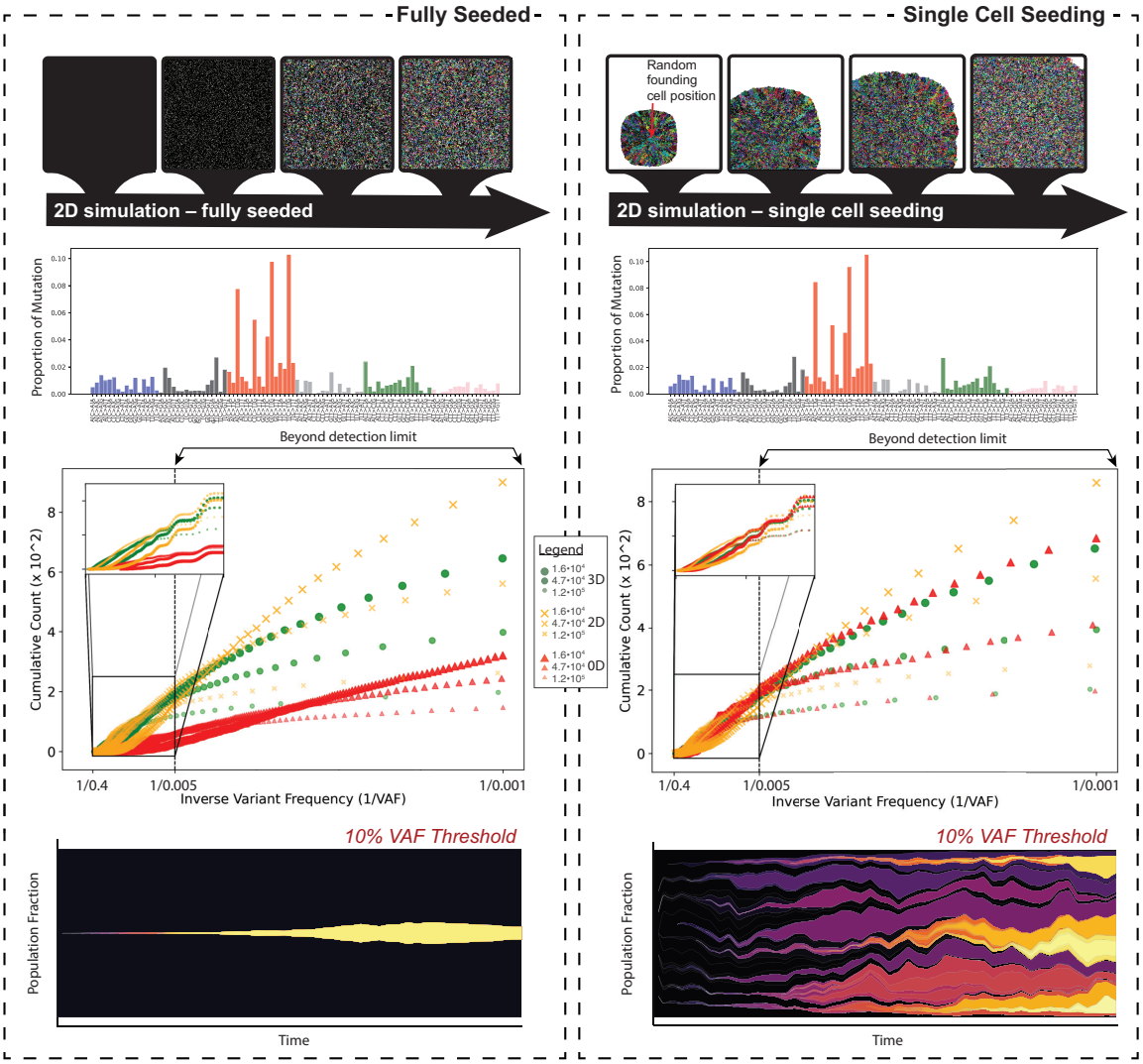


Figure A.1: Results comparing the fully seeded (*left*) and single cell seeding (*right*) model types and their corresponding mutation profiles and clonal dynamics. For each case, the mutation proportion across the 96 mutation trinucleotides is shown for one of the $2D$ simulation replicates. The $1/f$ values for each of the modeled dimensions is shown for three different population sizes, the inset shows the $1/f$ distribution for that which would be within the limits of detection (a generous 0.005 VAF at high depths). Beneath this, the same $2D$ replicate that is shown in the mutation spectrum plot is used to highlight the differences in clonal dynamics using an EvoFreq plot with a 10% VAF cutoff between fully seeded and single cell seeding.

carrying capacity (k) and population size (N_T) at every time point of either the domain (e.g. number of lattice points) or as a set parameter in the 0D case. The equation governing this scaled birth rate (λ_T) is given by $\lambda_T = \lambda\frac{k-N_T}{k}$. If a random number ($[0, 1]$) is less than the death parameter (ρ) plus λ_T a death or birth may happen for a given cell. The probability of a birth event given an empty lattice position (2D and 3D only) is given by $P(Birth) = \rho + \lambda_T$. If a random number ($[0, 1]$) is less than this birth event value a cell will die, if not the cell is able to divide.

When initializing Gattaca for these simulations an overall mutation rate of $3.2 * 10^{-}9$ was used and the mutation spectrum defined was given from a randomly sampled cohort of *diffuse large B-cell lymphoma* (DLBCL) *whole exome sequencing* (WES) from TCGA (this is available in the gattaca example code). When we analyze these mutation spectrums, post simulation we observe similar distributions of mutation types across all dimensions and model types consistent with mutation processes expected, based on the Gattaca initialization (Figure A.1 mutation spectrums). This cohort was chosen at random from the collection of TCGA sample types, but because of the way Gattaca uses this information to initialize distributions it is highly generalizable and any mutation spectra can be used. In the cases where the domain (or carrying capacity for 0D) is fully seeded we see that the $1/f$ distributions of variant allele frequencies is similar in the 3D and 2D cases (Figure A.1). Contrasting this with the single cell seeding case we see that the 0D and 3D cases are the most similar while 2D appears to reveal a different distribution (Figure A.1). These results suggest that the modeling dimension is an important consideration for the research question. As expected most of the clones that are obsered are below the detection limits of common methodologies, but can be captured here. The clonal dynamics, as demonstrated by the EvoFreq plots [196] illustrates that spatially constrained clones competing with one another are rarely able to expand beyond 10% VAF in the fully seeded cases while several clones reach this size during simulations with single cell seeding.

A.1.2 Gattaca Case Study 2: Wounding

Within the first case study we utilized Gattaca across two different types of models and three different dimensions. Next we wanted to evaluate if wounding within these models would alter the observed clonal dynamics as the spatial constraints for certain clones is relaxed when cells are removed in a wounding event (Figure A.2A). In all simulations, each ABM is seeded by a single cell. Wounding begins once the thousandth timestep is reached (Figure A.2B). After this, wounding occurs at time

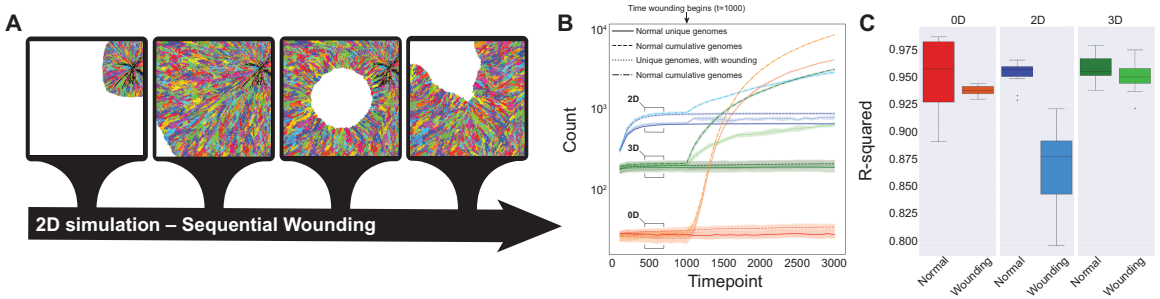


Figure A.2: Illustration of the repeated wounding of the single cell model in 2D where colors represent clones that differ by at least one mutation (A). For each of the dimensions the cumulative and unique genomes is given over the course of simulations (B). R-squared values for the linear regression on $1/f$ distributions for mutations is plotted for all dimensions with and without wounding (C) for all replicate simulations.

steps where the population is greater or equal to 85% of the total possible population (as dictated by the domain size). For the 0D and 3D simulations cells are killed by wounding within a circular and spherical manner, respectively. The number of cells killed through each wounding event is kept similar by adjusting the radius between 2D and 3D simulations, while in the 0D case, the number of cells killed is an equivalent number of cells. The same birth/death dynamics and equations used in case study one are used here, because the probability of birth is modulated by the number of cells (*i.e.* the probability of birth is reduced as the carrying capacity of the system is reached) a wounding event acts to increase cell divisions where empty sites are present and thus allows clones to expand into the wounded areas.

When we examine the differences between the wounding and non-wounding simulation's cumulative and unique genomes over time we see a clear signal at the time wounding occurs. At this point, space is open and rapid cell proliferation refills the areas where the wound occurred (in 0D this results in rapid proliferation back to carrying capacity). As cells divide and mutate a large number of unique genomes appear over time (Figure A.2B). We see that the number of unique genomes in the 0D case increases drastically faster than those in the 2D and 2D cases, this is due to the mechanism where clones in the 0D case are chosen at random to be killed while whole or near whole subclonal populations are removed in the 3D and 3D simulations. Interestingly, when we compare the $1/f$ distributions through their R-squared values, from linear regression analysis, we see that in the 2D wounding case the relaxation of spatial constraints appears to drive a signal of non-neutral dynamics in a system that is functionally homogeneous where slight fitness advantages are conferred through

	ggmuller	fishplot	EvoFreq
Frequency data		✓	✓
Size data	✓		✓
Long format	✓		✓
Wide format		✓	✓
Smoothing		✓	✓
Uses ggplot	✓		✓
Animations			✓
Create graphs			✓
Integration with subclonal reconstruction software		✓	✓

Table A.1: Comparison of the features within the currently available packages for plotting evolutionary dynamics. EvoFreq is capable of visualizing the evolutionary dynamics from ClonEvol, PhyloWGS, and CALDER.

room to expand (Figure A.2C).

A.2 EvoFreq R Package

EvoFreq is built on top of the leading visualization library in R, ggplot2 [261]. Input data can be either clone sizes or frequencies, and in long or wide formats. Given such data, EvoFreq can create Muller plots and/or dendrograms, revealing the clonal dynamics over time. Additional customizations include the ability to color clones based on a user defined attribute (such as fitness), provide custom colors, and label polygons. Due to EvoFreq's utilization of ggplot2 as the primary, underlying library, further customization to EvoFreq's plots is possible. Using the optional dependency [262], users may also create animations of evolving Muller plots and "growing" dendrograms. EvoFreq is capable of visualizing frequency dynamics at a single point in time, as a phylogenetic tree representation, a graph representation, or as a frequency plot over time similar to FishPlot [200] and ggMuller [201], but with extended capabilities (Table A.1).

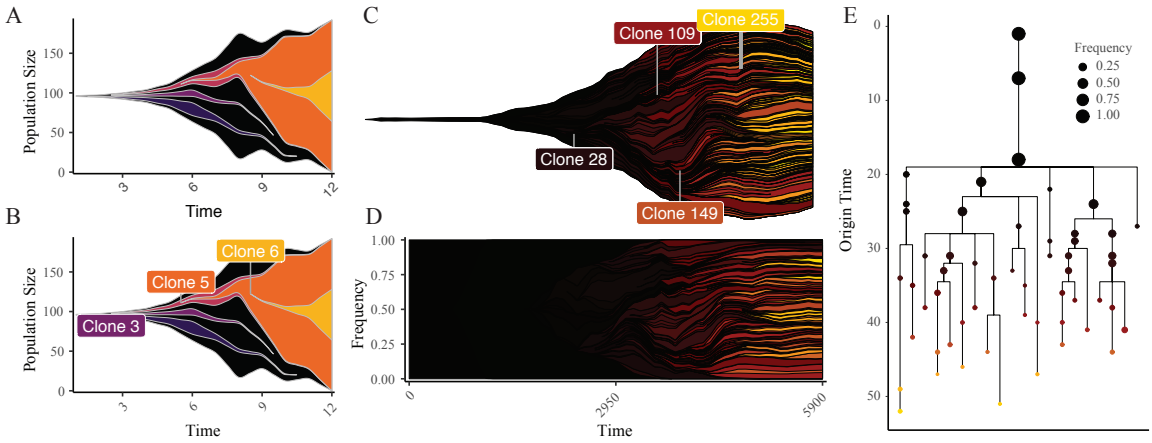


Figure A.3: EvoFreq is a comprehensive and flexible R package for the visualization of longitudinal data. **A** and **B** show an EvoFreq plot for one of the provided datasets with **B** and without **A** the function to add labels. For more complicated data EvoFreq provides a powerful means to quickly filter data (**C** and **D** thresholded at 0.2 frequency), color by an attribute (driver strength), and visualize dynamics as a frequencies rather than population size. Using a dendogram styled to ensure that origin time is easily conveyed (termed here as an EvoGram), a more quantitative view is provided **E**.

A.2.1 EvoFreq Results and Example

At the core of EvoFreq is the ability to visualize relational data structures over time, whether this is from simulations or inferred from data. Here we illustrate the usefulness of generating these results using EvoFreq. Our first example utilizes data generated by West et al. which quantifies how spatial constraints alter the evolutionary trajectory of a single tumor using a passenger-driver model [108]. EvoFreq was used extensively within this publication and a subset of this data is shown in Figure A.3. This figure highlights how the user can manually or automatically add labels to provide informative details of clones, a particularly useful feature of EvoFreq.

Second, we analyze sequence data using three different clonal reconstruction tools from 15 initial engraftments and serial propagation of primary and metastatic breast cancers [263] and visualize these results using EvoFreq. We applied ClonEvol [246], PhyloWGS [244], and CALDER [243] to infer the clonal dynamics from each of the longitudinal xenografts (select inferences are illustrated in Figure A.4). Initial processing and reformatting of the somatic SNV, CNA, and LOH data from *whole-genome shotgun sequences* (WGSS) and Affymetrix SNP Array 6.0 was carried out to extract read data using custom python scripts and prepare inputs for PhyloWGS and CALDER. Each of these tools requires different pre-processing and infers subclonal

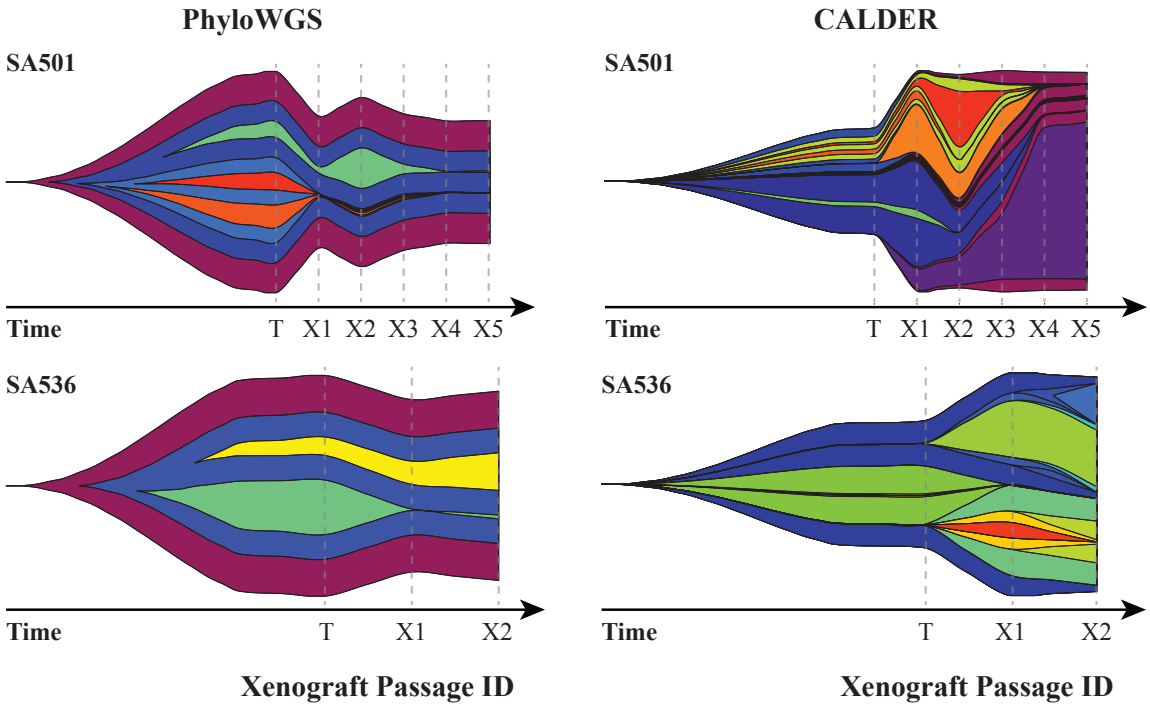


Figure A.4: EvoFreq can be easily used to visualize outputs from CloneEvol, PhyloWGS, and CALDER. Data parsing functions are included within EvoFreq to rapidly visualize subclonal reconstructions. Each column above illustrates a method of subclonal reconstruction for two separate human breast cancer xenograftments from Eirew et al. [263] for PhyloWGS (left) and CALDER (right). Originating tumors (T) and their subsequent xenograft passages (X1, X2, etc.) are shown for SA501 (top) and SA536 (bottom)

reconstructions in different ways. We have incorporated functions within EvoFreq to parse outputs of each of these tools to visualize inferred clonal dynamics using EvoFreq. When the output provides numerous solutions for subclonal reconstructions, the user has an option of returning one or all of these solutions. Examples of this process have also been provided within EvoFreq's documentation.

EvoFreq is a versatile library capable of, generating publication and presentation ready images as well as video for, visualizing clonal frequencies over time. EvoFreq's design allows for broad access to all users through robust support for input data and input validation functions. EvoFreq can be used for all relational data structures providing many different visualizations from one user-facing library, making it applicable in a number of fields. EvoFreq has currently been adopted by a number of research groups, focused on cancer genomics and mechanistic modelling. All source code, read me, and issue support is available at https://github.com/MathOnco/EvoFreq.

A.3 Additional Figures

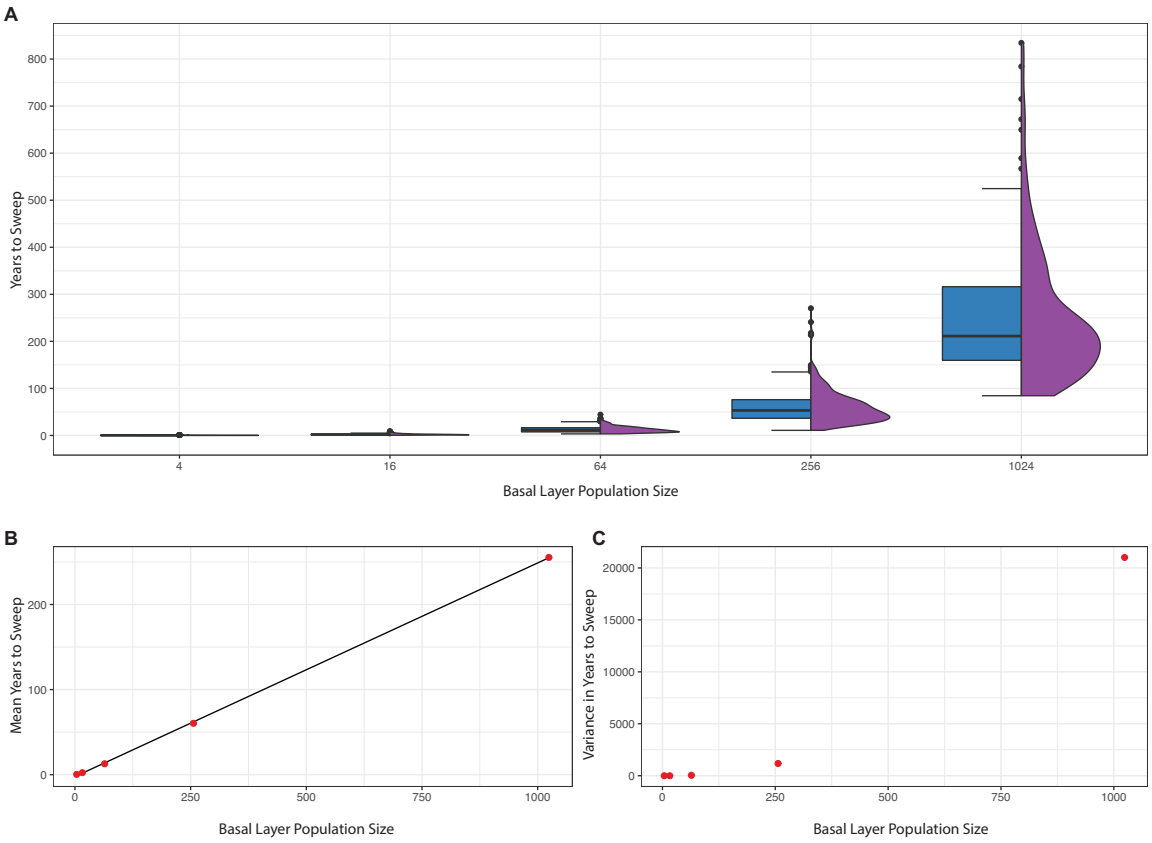


Figure A.5: For 500 replicate simulations ranging from as small as $2x2$ cells within the basal layer up to $32x32$ cells (population size of 1024) within the basal layers we show the time it takes a single basal layer cell to sweep through the basal layer alone (**A**). For these same simulations a linear regression model ($R - squared = 0.9998$, p-value < 0.001) is fit to the mean of the different simulations (**B**). While **B** illustrates the linearity of the mean time to sweep, the variance increases non-linearly as the population size increases (**C**).

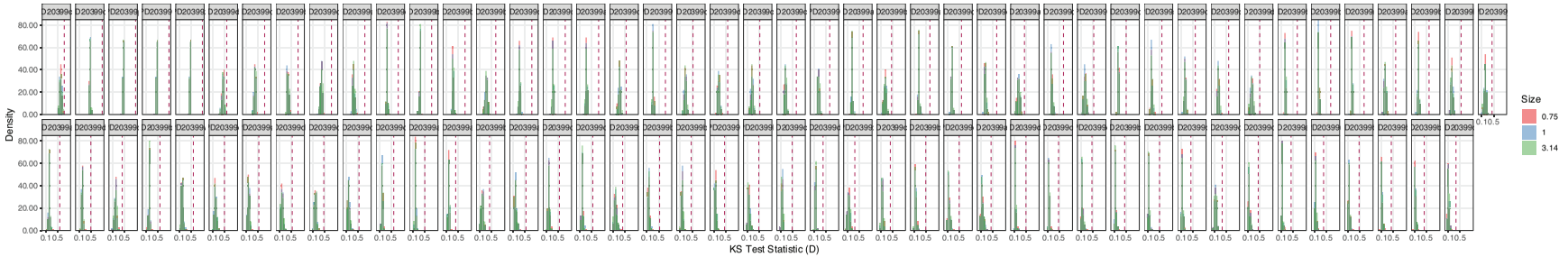


Figure A.6: Pairwise comparisons for 55 year old patient PD20399 illustrating the Komlogrov-Smirnov test statistic ($D_{m,n}$) distribution for the three biopsy simulation sizes. Minimum critical values (D_α) needed to reject the null hypothesis is illustrated by the red dashed line.

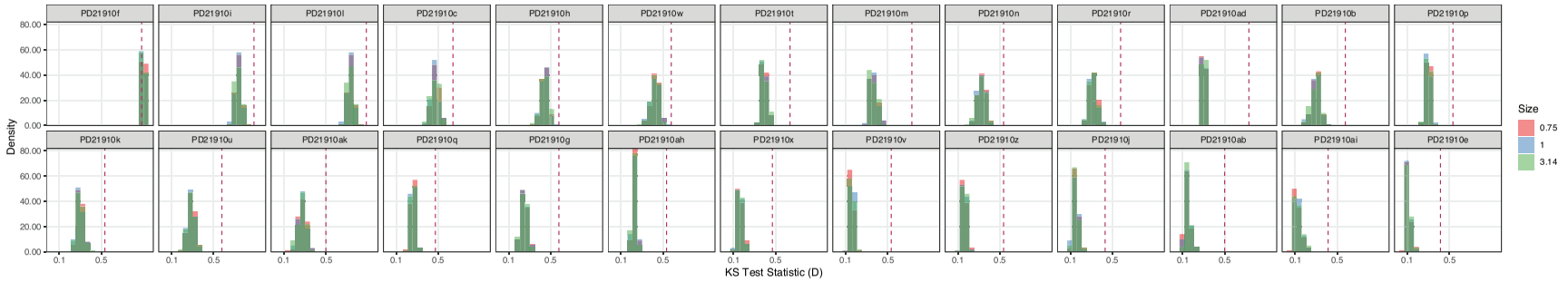


Figure A.7: Pairwise comparisons for 58 year old patient PD21910 illustrating the Komlogrov-Smirnov test statistic ($D_{m,n}$) distribution for the three biopsy simulation sizes. Minimum critical values (D_α) needed to reject the null hypothesis is illustrated by the red dashed line.

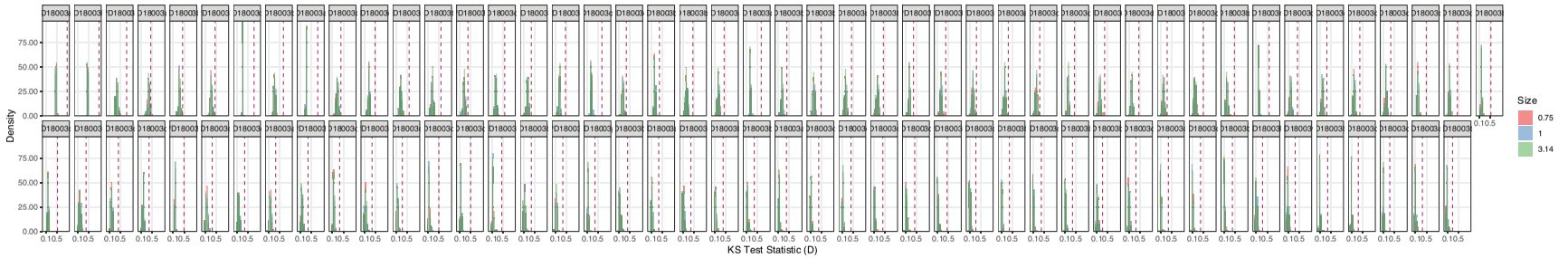


Figure A.8: Pairwise comparisons for 65 year old patient PD18003 illustrating the Komlogrov-Smirnov test statistic ($D_{m,n}$) distribution for the three biopsy simulation sizes. Minimum critical values (D_α) needed to reject the null hypothesis is illustrated by the red dashed line.

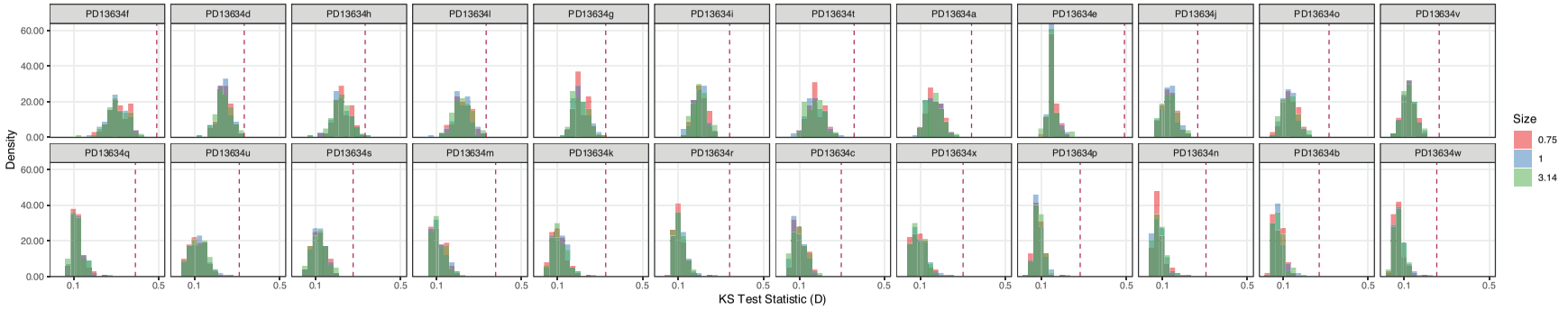


Figure A.9: Pairwise comparisons for 73 year old patient PD13634 illustrating the Komlogrov-Smirnov test statistic ($D_{m,n}$) distribution for the three biopsy simulation sizes. Minimum critical values (D_α) needed to reject the null hypothesis is illustrated by the red dashed line.

Appendix B

Appendix: Chapter 3

B.1 Additional Information

B.1.1 Derivation of fluctuating CpG loci

To isolate those CpG sites that behave as FMCs, it was first necessary to filter out those loci that are likely to have a regulatory function or change their methylation status over the length of the crypt. This was done by selecting only those CpG sites that lie in the 'open sea' (further than 4kb from a CpG island). Furthermore, probes of CpG loci that were identified as being cross-reactive [265, 266] were filtered out, along with CpG loci positioned on sex-determinant chromosomes. Given the relatively low amounts of DNA contained within a single crypt, we also filtered out probes that were likely to have experienced incomplete binding by restricting our analysis to probes that had a total intensity greater than 1,200 (arbitrary units).

The Illumina EPIC array features two different probe types, type I and type II [266]. Type I probes feature a higher dynamic range, leading to the two probe types having different underlying distributions of β values. Due to difficulties in simultaneously modeling the two different probe types, and given that type I probes are over-represented in CpG-dense regions of the genome, the analysis was restricted to type II probes.

CpG sites with fluctuating methylation were then detected by comparing between-individual to within-individual heterogeneity in methylation value. At fluctuating sites, we expect the average methylation in non-clonal bulk samples to follow a distribution centered on 0.5 (because methylation at the site is uncorrelated between the multiple lineages that make up the bulk sample), whereas in individual clonal samples, the methylation value can take any value between 0 and 1. Thus, to select for fCpG loci, we selected CpG sites that had the highest 5% of variance in β value

Definition	Variable	Values for θ_i
Population size	N_T	200, 1000
Stem cell division rate ranges (hours)	ρ_c	(12, 24), (24, 48), (48, 72), (72, 120)
Stem cell pool radius (shown as stem cells)	ψ	6-7, 9-10, 12-13, 15-16, 21-22
(De)methylation rate	ω	0.001, 0.005, 0.01, 0.05, 0.1, 0.5

Table B.1: Exclamatory bowel parameters. Exclamatory bowel parameters and values used for discretized grid search of parameters. These parameter values make up 288 different parameter combinations such that a combination is given as $\theta_i = (N_T, \rho, \psi, \omega)$. Parameters governing cell movement are not shown and are constant based on off lattice parameters defined by HAL [264].

between individual samples and then filtered these for sites with mean methylation across all samples and individuals of ~ 0.5 (mean β value between 0.4 and 0.6) (Fig. 2a TODO!!!).

To demonstrate technical accuracy in methylation measurement from the small amounts of DNA in single small intestinal crypts (~ 400 cells), colon crypts ($\sim 2,000$ cells) or endometrial glands ($\sim 5,000$ cells), we identified similar fCpG sites on the X chromosome and compared methylation between male and female individuals. In males, there is only a single copy of the X chromosome; hence, only two modal peaks near 0% and 100% methylation should be present in clonal populations, as opposed to the trimodal distribution observed on autosomes. Consistent with the ability to measure fluctuating methylation in small tissue samples, the X chromosome fluctuating sites exhibited trimodal W-shaped distributions in female colon crypts and bimodal 'U-shaped' distributions in male colon crypts (Supplementary Figure B.1D). This observation is supportive of the hypothesis that the methylation distribution of fCpG loci is reflective of that of the most recent recurrent clone rather than varying with cell type or differentiation status.

We compared methylation of fluctuating sites between crypts from the same individual. If fluctuating methylation occurs stochastically and without biological regulation, then each crypt should independently evolve a unique pattern of fCpG site methylation. Intercrypt comparison between crypts within the colon or small intestine, both across the set of crypts sampled from each individual and across crypts from different individuals, showed that fluctuating methylation patterns between crypts were uncorrelated (Figure B.2E). There was weak correlation of fluctuating methylation patterns between crypts for younger individuals (age < 30 years), but this was lost with advancing age (Figure B.2F).

B.1.2 Evidence for fluctuating human CpG sites

Oscillatory DNA methylation has been previously documented in cell line systems where changes in DNA methylation have been directly observed and correlated with changes in gene expression and developmental cell fates. Cyclical DNA methylation of a transcriptionally active promoter [204–207]. These active changes in DNA methylation are modulated by Tet and DNMT family enzymes, and occur over hours or a few days. By contrast, the erratic fluctuations at fCpG loci outlined here occur over years and in adult human tissues. For these fCpG loci, the proposed mechanism is more likely stochastic replication errors, where DNA methylation and demethylation are equally likely.

Such stochastic, neutral changes in DNA methylation are more likely to occur in non-genic or non-regulatory CpG sites rather than in expressed genes that could alter cell phenotypes. The identified fCpG loci were relatively enriched for non-genic CpG sites (Figure B.2B) and were about three times less likely to be in promoter regions that largely control gene expression. Moreover, genes with promoter fCpG sites had significantly lower expression than genes without promoter fCpG loci ($p < 0.001$ Welch's t-test performed upon the log-transformed data, Figure B.2A). We note that none of the genes with promoter fCpG sites had an expression greater than 10TPM, a typical cutoff for "intermediate" expression. Hence, fCpG methylation variation is unlikely to be associated with cell differentiation or actively confer significant changes in phenotype.

Although direct serial observations of the methylation and demethylation of fCpG loci in the same human colon crypt are impractical, the proposed stochastic mechanism has testable predictions. We illustrate some of these predictions by comparing methylation extremes, defined as "0" if the methylation is less than 0.2 and "1" if the methylation is greater than 0.8 (Figure B.2B). The clonal crypts have hundreds of fCpG sites with values of either "0" or "1", which should effectively serve as lineage barcodes over short time intervals (CpG sites with methylation between 0.2 and 0.8 are not compared). For example, two identical samples should have the same barcode. With increasing time, the fluctuations should randomize the fCpG methylation, and the probability of matching should be about 0.5, or the same as flipping a coin.

We compared FMCs between the top and bootoms of colon crypts (Figure B.2C). Crypt stem cells are located in the bottoms of crypts and cell differentiation occurs at crypt bottoms. Diffentiated cells migrate upwards and the tops of crypts are differentiated cells that are lost within a week. As expected for a relatively short interval of a week, there was minimal switching between 0 to 1 or 1 to 0 between the top and bottom of the same crypt (Figure B.2D). This data also indicates the fluctuations in methylation pattern are not assocated with cell differentation.

By contrast, essentially a random distribution was found when FMCs were compared between adult crypts from different indivduals or within an individual, with an average identity of barcodes of about 0.5 (Figure B.2E). The youngest individual was 17 years old, providing many years for methylation fluctuations to randomize between crypts that last share a common ancestor before birth. Consistent with a requirement for time for randomization, there was a trend for increasingly more random barcodes with age, with younger individuals tending to have more similar fCpG barcodes (Figure B.2F).

In summary, although we are unable to directly observe changes in the fCpG loci methylation, these observations are consistent with fluctuations in methylation at CpG sites with change occurring over many years. The fCpG sites do not switch during cell differentiation occurring over a week, and are rarely found in promoter regions or in highly expressed genes. Consistent with stochastic fluctuations that occur independently in different cells, FMC barcodes are essentially randomized between adult crypts from different individuals, and become increasingly different with aging within an individual. These findings are more consistent with dynamic methylation fluctuations that randomly change between 0, 0.5 and 1.0, rather than active regulatory oscillations or static random states. The FMCs distribution reflect the population structure of their cells. In stable polyclonal populations, the fCpG sites are asynchronous between cells and average methylation will be around 0.5, with narrow distributions as seen in normal whole blood. However, fCpG sites become synchronized in clonal populations and exhibit "W" shaped distributions, as seen in normal human crypts, endometrial glands, and clonal hematopoiesis.

B.1.3 Mathematical model of fluctuating CpG distributions in the stem cell niche

We developed a stochastic model to describe how the fraction of methylated alleles (β value) in the stem cell niche of a given CpG locus changes over time. This model draws on previous attempts [267, 268] to model the behavior of the stem cell niche in colonic crypts but with a number of modifications that account for the differences when considering methylation as a lineage tracing marker rather than DNA. Namely, while DNA mutations occur relatively infrequently, allowing for a model that only considers a single mutant population expanding or contracting with reference to a single wild-type population, the relatively high methylation switching rate requires us to consider the potential of multiple clones existing simultaneously. Further, while DNA mutations can be generally regarded as irreversible, the methylation status of a given cell (that is, whether a particular cell is homozygously (de) methylated or heterozygously methylated) can theoretically flip–flop, necessitating a careful consideration regarding the possible ways the overall β value can change.

For this reason, we made the simplifying assumption that the population was well mixed, such that any of the S stem cells can replace any of the other S–1 stemcells with equal probability and that these replacements occur at a constant rate λ per stem cell. This assumption greatly simplified our analysis, as the system can be fully characterized using just two state variables: k – the number of stem cells containing a

single methylated allele, and m – the number of stem cells containing two methylated alleles. The admitted states are constrained by the inequality $0 \leq k+m \leq S$ for a total of $\frac{1}{2}(S+1)(S+2)$ states.

Along with the replacement process, we assumed that a previously unmethylated CpG locus could spontaneously become methylated with a rate μ per year and, conversely, that a previously methylated CpG locus could spontaneously become demethylated with a rate γ per year.

To develop the series of ordinary differential equations that fully determine the system, we considered the ways in which a state (k,m) could transition to a state (k',m'). As an example, if we consider Figure 4.1C, we observe that of the $S=5$ stem cells, 3 of the stem cells are heterozygously methylated, and 1 of the cells is homozygously methylated; hence, the system is initially in state $(k=3,m=1)$. To transition to state $(k'=3,m'=2)$, the homozygously methylated stem cell must clonally expand, replacing the homozygously demethylated cell. The rate at which any one of the stem cells replaces another is $\lambda S = 5\lambda$, but of the $S(S–1) = 20$ possible transitions, only 1 would lead to the desired $(k'=3,m'=2)$ state; hence, the rate at which the system transitions $(3,1) \rightarrow (3,2)$ is $\frac{1}{20} * 5\lambda = \frac{1}{4}\lambda$. We continue this process considering the general transition $(k,m) \rightarrow (k',m')$, deriving the following master equation (full derivation can be found in the supplementary information of our publication [210]):

$$\begin{aligned}
\frac{dP(k,m|\lambda,\mu,\gamma;t)}{dt} = \\
(S-m-(k-1))((k-1)\frac{\lambda}{S-1}+2\mu)P(k-1,m|\lambda,\mu,\gamma;t) \\
+(m-1)(S-(m-1)-k)\frac{\lambda}{S-1}P(k,m-1|\lambda,\mu,\gamma;t) \\
+(k+1)((m-1)\frac{\lambda}{S-1}+\mu)P(k+1,m|\lambda,\mu,\gamma;t) \\
+(k+1)((S-m-(k+1))\frac{\lambda}{S-1}+\gamma)P(k+1,m|\lambda,\mu,\gamma;t) \\
+(m+1)(S-(m+1)-k)\frac{\lambda}{S-1}P(k,m+1|\lambda,\mu,\gamma;t) \\
+(m+1)((k-1)\frac{\lambda}{S-1}+2\gamma)P(k-1,m+1|\lambda,\mu,\gamma;t) \\
-(2(k(S-k)+m(S-k-m))\frac{\lambda}{S-1}+(2S-(k+2m))\mu \\
+(k+2m)\gamma)P(k,m|\lambda,\mu,\gamma;t)
\end{aligned} \tag{B.1}$$

This linear series of differential equations can be solved computationally by rewrit-

ing the equations into a matrix equation, $\frac{dP(t)}{dt} = T\vec{P}(t)$ and applying matrix eponentiation to the resulting transition matrix T,

$$\vec{P}(t) = e^{tT}\vec{P}(t=0). \tag{B.2}$$

During the very early stages of embryogenesis, the existing methylation patterns inherited from parental gametes are largely erased before a large wave of de novo methylation remodels the entire genome, resulting in a bimodal methylation distribution [269]. Given that all the stem cells within a niche are initially clonal, we thus assumed that it was equally likely to find a given fCpG locus as homozygously methylated or unmethylated across all the stem cells within the niche at time 0. Further study is necessary to ensure the validity of this assumption.

$$P(k-1, m+1|\lambda, \mu, \gamma; t=0) = \begin{cases} 0.5 & \text{if } k = 0 \wedge m = S \\ 0.5 & \text{if } k = 0 \wedge m = 0 \\ 0 & \text{otherwise} \end{cases} \tag{B.3}$$

However, the methylation status of individual cells is not available using methylation arrays; hence, the hidden states must be marginalized over to calculate the probability of there being z methylated copies within the stem cell niche (note that $0 \leq z \leq 2S$). This can be achieved by summing the various combinations of k and m states that satisfy the equation $z = k + 2m$.

$$P(z|\lambda, \mu, \gamma; t) = \sum_{m=0}^{S} \sum_{k=0} S - m(k, m|\lambda, \mu, \gamma; t)\delta_{k+2m,z} \tag{B.4}$$

The resulting distribution of $P(z|\lambda, \mu, \lambda; t)$ can qualitatively reproduce the characteristic W shape exhibited in the methylation fraction of individual crypts.

B.1.4 Error model

The probablity distribution calculated above (Appendix B.1.3 Equation B.4), $P(z|\lambda, \mu, \gamma; t)$, gives the probability that exactly z of the $2S$ alleles (across S stem cells) are methylated at a particular CpG locus; however, the Illumina EPIC array quantifies the methylation level at specific loci aggregated over the whole crypt similar to the output β values the spatial EB model provides without any noise or technical error. As such, we introduced an error model to link the measured β value with the 'true' z value at a specific site. We chose to model the measured β values as a mixture of z β-distributed random variables, each with a mean value determined by z and a scale parameter κ_z.

To account for the background noise of the array, the mean value of each β peak was set to be equal to a linear transform of $z : x = (\epsilon - \Delta)\frac{z}{2S} + \Delta$, with the parameters describing this transform (ϵ and Δ) to be inferred. The scale parameters (sometimes referred to as the sample size), $\vec{\kappa}$, of each β peak were modeled as hierarchical, with each κ_z being drawn from a lognormal distribution parameterized in terms of the population mean, θ, and its standard deviation, σ. These hyperparameters were also inferred during the Bayesian inference.

B.1.5 Likelihood and prior functions

As rate parameters are naturally positive quantities, λ, μ and γ were constrained to positive real values by defining the prior distributions in terms of positive half-normals with a scale informed by prior literature. Following the finding of Nicholson et al. [208] that the replacement rate is approximately 1.3 replacements per stem cell per year, we set the scale of the prior on the replacement rate equal to 1. Similarly, θ and σ were also constrained to positive values using broad half-normal prior distributions, with a scale of 500 and 50, respectively. Previous work has found that methylation fidelity can vary dramatically across the genome, from $\sim 10^{-4}$ to 10^{-2}, and we will take an estimate of 10^{-3} per division as a reasonable scale [270]. If we assume intestinal stem cells divide every $\sim$ 3 days and we consider that our definition of μ, γ is in units of (per allele per year), this corresponds to a (de)methylation rate of $\sim$ 0.05. We note that the inference is relatively insensitive to the exact choice of prior on the (de)methylation rate (Supplementary Fig. 7b,c TODO).

The lognormal hierarchical prior distribution naturally constrains $\vec{\kappa}$ to real values. The 'offsets' in the linear transform, Δ and ϵ, were constrained to lie between 0 and 1 by placing a β distribution on each parameter, such that the mean prior value was 0.05 and 0.95, respectively.

The behavior of individual CpG loci was assumed to be independent, such that the likelihood of all N = 1,794 CpG loci was simply the product of the per-CpG likelihood computed according to the mathematical model outlined above (Appendix B.1.3).

Likelihood:

$$x = (\epsilon - \Delta)\frac{z}{2S} + \Delta \tag{B.5}$$

$$P(\beta_i | z, \Delta, \epsilon, \kappa_z) = \frac{\beta_i^{\kappa_z x - 1}}{B(\kappa_z x, \kappa_z(1-x))} \tag{B.6}$$

$$\mathcal{L}(\lambda, \mu, \gamma, \Delta, \epsilon, \vec{\kappa}, S|\vec{\beta}) = \prod_{i=1}^{N} \sum_{z=0}^{2S} P(\beta_i|z, \Delta, \epsilon, \kappa_z) P(z|\lambda, \mu, \gamma; t) \tag{B.7}$$

Hyperpriors:

$$\theta \sim \text{halfnormal}(500)$$

$$\sigma \sim \text{halfnormal}(50)$$

Priors:

$$\lambda \sim \text{halfnormal}(1.0)$$

$$\mu \sim \text{halfnormal}(0.05)$$

$$\gamma \sim \text{halfnormal}(0.05)$$

$$\Delta \sim \beta(5, 95)$$

$$\epsilon \sim \beta(95, 5)$$

$$\kappa_z \sim \text{halfnormal}\left(ln\left(\frac{\theta^2}{\sqrt{\theta^2 + \sigma^2}}\right), \sqrt{ln\left(1 + \frac{\sigma^2}{\theta^2}\right)}\right)$$

B.1.6 Bayesian inference

A Bayesian inference methodology was developed to infer the biological model parameters (number of stem cells within the stem cell niche (S), replacement rate per stem cell per year (λ), and methylation (μ) and demethylation (γ) rate per CpG locus per stem cell per year) directly from the distribution of FMC β values for each crypt.

Investigation of simulated datasets revealed that the resulting posterior distributions were multimodal, with each S value associated with a local maxima (due to the correlation in the posterior between S and λ). This multimodality can make the posterior difficult to explore using traditional Markov chain Monte Carlo techniques, such as Hamiltonian Monte Carlo. To overcome this, a nested sampling method [271, 272] was developed to calculate the Bayesian evidence (marginal probability density, Z) of each S value considered ($S \in [3..20]$) while simultaneously generating samples from the posterior associated with each value of S. The probability of S for a given crypt can then be calculated as

$$P(S|\vec{\beta}) = \frac{Z(S|\beta)}{\sum_j Z(S_j|\vec{\beta})} \tag{B.8}$$

The full posterior can be approximated by drawing samples from each S mode with a weight equal to the inferred probability of S. The nested sampling was performed using dynesty [213], a Python implementation of the nested sampling algorithm, using the 'rwalk' sampling option, such that new live points are generated from existing live points under random walk behavior.

To ensure that the posterior samples had converged to the equilibrium distribution, four independent samples were run with random initializations for each sample, and the rank-normalized potential scale reduction statistic ($\hat{R}$) was calculated [273]. $\hat{R}$ was found to be less than 1.1 (a typical threshold used to determine convergence) in all cases. The inference code can be obtained from `https://github.com/CalumGabbutt/flipflop.git` (ref [212]).

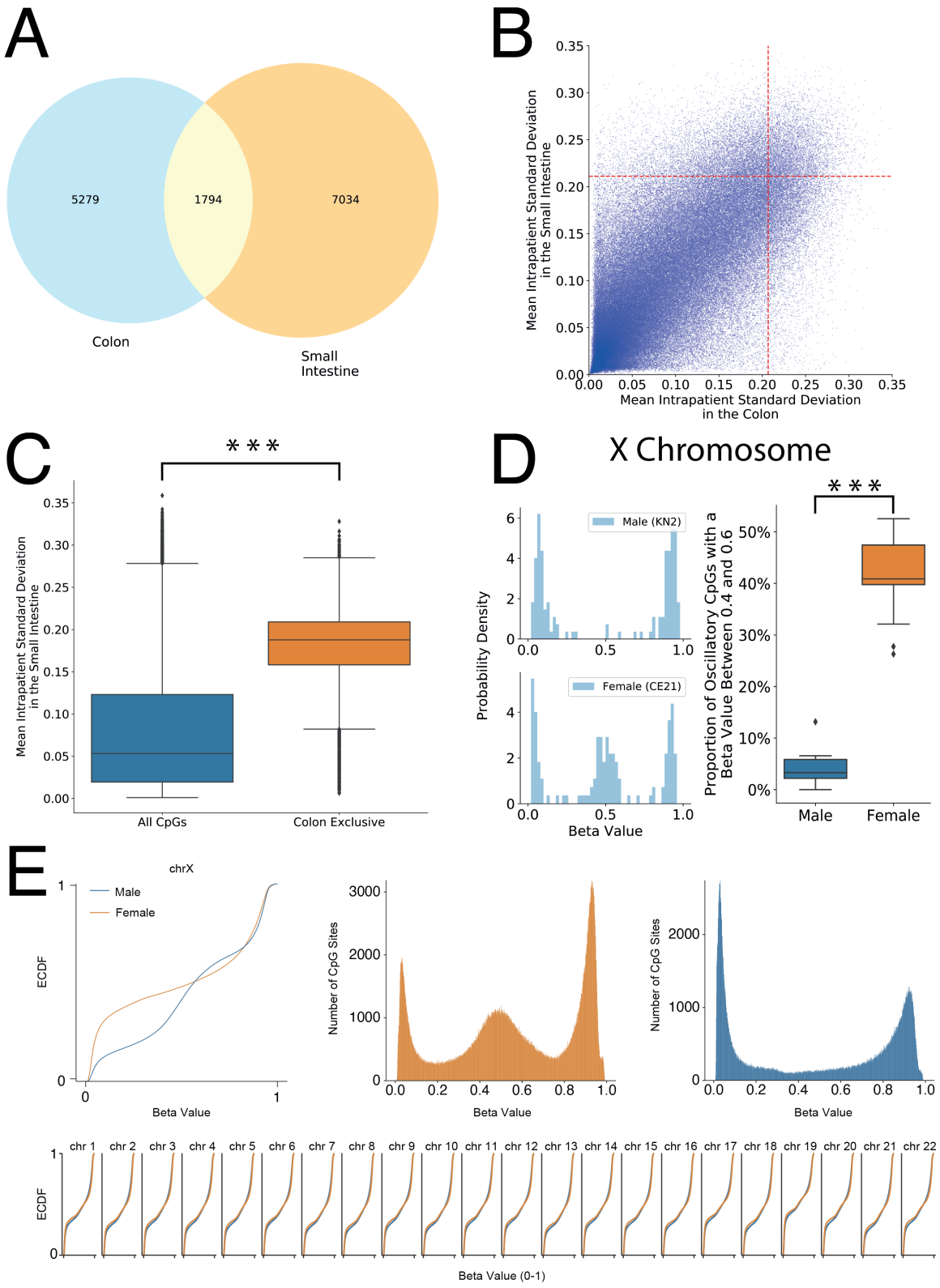


Figure B.1: Additional analysis of fluctuating CpG identification.(Caption next page.)

Figure B.1: Additional analysis of fluctuating CpG identification. (Figure previous page.) **A:** Venn diagram showing the overlap of CpG loci identified as *fluctuating CpGs* (fCpGs) in the colon and the small intestine. **B:** Scatter density plot (with the density plotted on the log-scale) of the heterogeneity metric (mean intra-patient standard deviation) of CpGs in the colon and small intestine, with the cutoff of the top 5% most heterogenous loci indicated in red. **C:** Comparison of the heterogeneity metric of the colon exclusive fCpG loci (i.e. those identified in the colon but not the small intestine) to all type II CpGs, within the small intestine samples. The colon exclusive CpG loci are significantly more variable ($p < 2 * 10^{-16}$, two-sided Mann Whitney U test). An extension of the fCpG identification process to CpG loci located on the X chromosome. **D:** (left) We present example methylation distributions for these X-chromosome fCpG loci for a male and a female crypt, confirming the predictions from theory that the male crypts lack the peak at 50% as they contain only a single copy. (right) To test whether this relationship holds in general, we compare the proportion of fCpG's on the X chromosome with an intermediate beta value ($0.4 \leq \beta \leq 0.6$) between all colon crypts from males and females, confirming that males have a significantly lower probability mass near 50% ($p = 3.9 * 10^{-6}$, two-sided Mann Whitney U test). **E:** This lack of peak at 50% can be shown across all patient fCpG sites for each chromosome. We show an *empirical cumulative density function* (ECDF) plot as well as histogram for chromosome X (chr X), and *empirical cumulative density function* (ECDF) plots for all autosomes. All boxplots show center line, median; box limits, upper and lower quartiles; whiskers, 1.5 IQR.

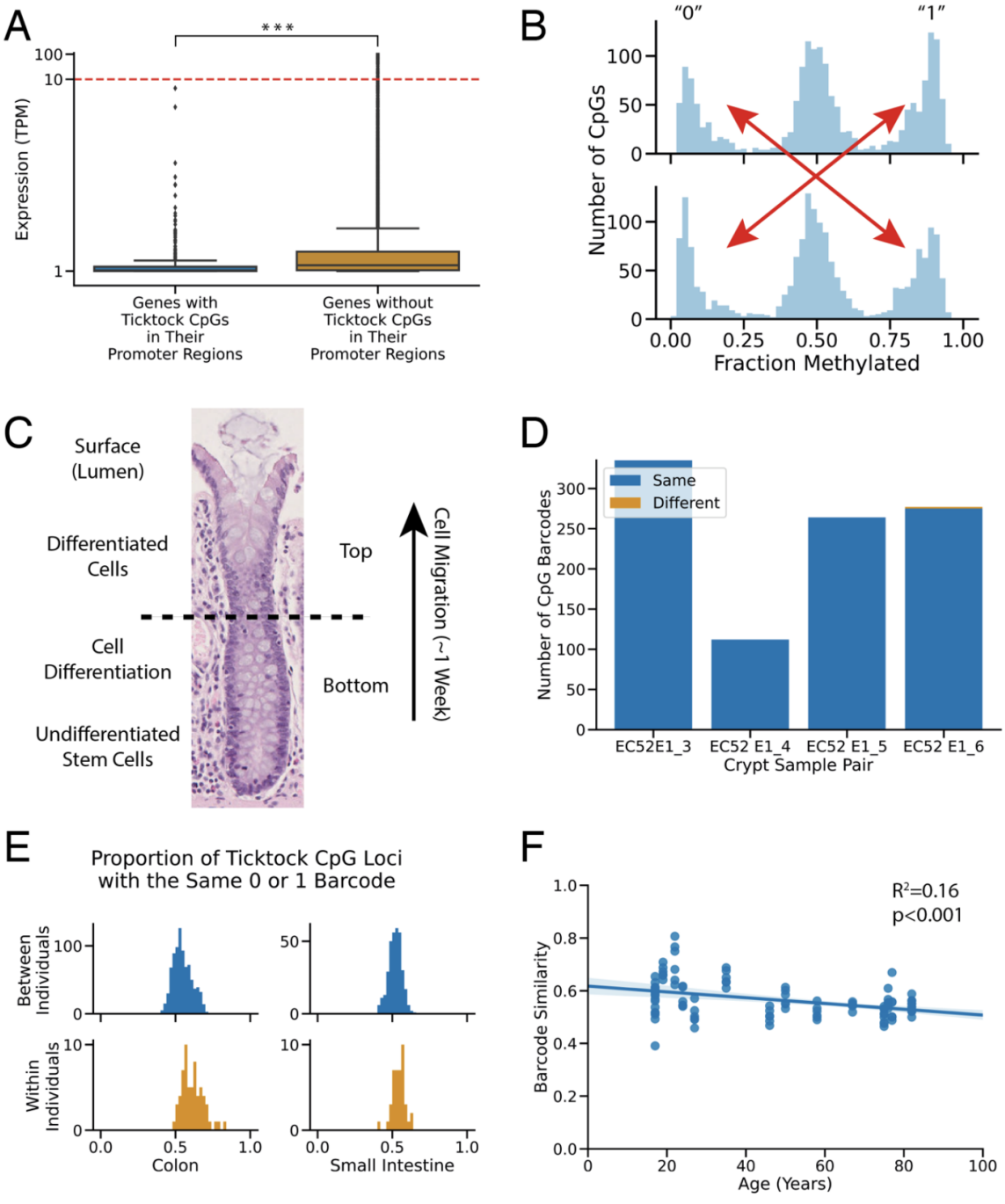


Figure B.2: Additional evidence supporting the existence of fluctuating CpG loci.(Caption next page.)

Figure B.2: Additional evidence supporting the existence of fluctuating CpG loci. (Figure previous page.) **A:** RNA expression is on average 3 times lower ($p = 3.0 * 10^{-28}$ two-sided Welch's t-test performed upon the log-transformed data) for genes with fCpGs sites in their promoter regions compared to genes without promotor fCpG loci in the colon (RNAseq single cell data from GSE132257. *fluctuating methylation clock* (FMC) methylation reversibly fluctuates between 0, 50 and 100% methylated. Two samples can be compared by counting the number of fCpG loci that switch from 0 to 1 or vice versa. **C:** Colon crypts have small numbers of undifferentiated stem cells at their bases/bottoms. Cell differentiation occurs near crypt bottoms and fully differentiated cells migrate upwards and are lost within a week or two. **D:** Comparisons between four top and bottom halves of the same crypts reveals FMC barcode methylation (0 or 1) is nearly always identical (986/988 barcode comparisons). Comparisons of colon and SI crypt FMC barcode methylation between and within individuals. **E:** The probability of another crypt having the same 0 or 1 barcode is approximately 0.5, or essentially random. **F:** Performing pairwise comparisons of the proportion of fCpG loci with the same 0 or 1 barcode, there was a slight trend towards increasingly random barcodes with age. Confidence band was calculated via bootstrapping and represents 95% confidence intervals. All boxplots show center line, median; box limits, upper and lower quartiles; whiskers, 1.5 IQR.

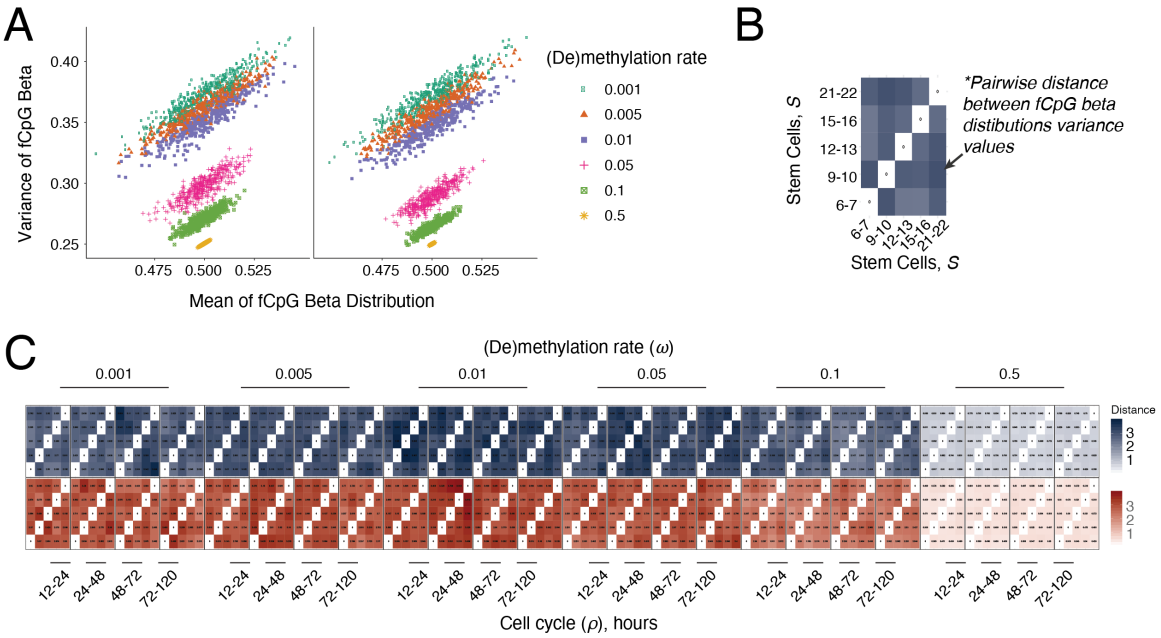


Figure B.3: Variances of fluctuating CpG distributions of the Exclamatory Bowel. These plots show the outputs for 288 unique parameter combinations, θ_i, with 20 replicates each. **A:** plots the mean and variance for a discretized grid space of the 5,760 simulations where the color denotes the (de)methylation rates. **B:** provides a legend for **C** to explain what each set of pairwise comparisons illustrates. On both axis we are comparing the number of stem cells across multiple ranges of cell cycles and (de)methylation rates are the average euclidean distance between *fluctuating CpG* (fCpG) distribution outputs. A further explanation of these plots, with fCpG distributions can be found in Figure B.4

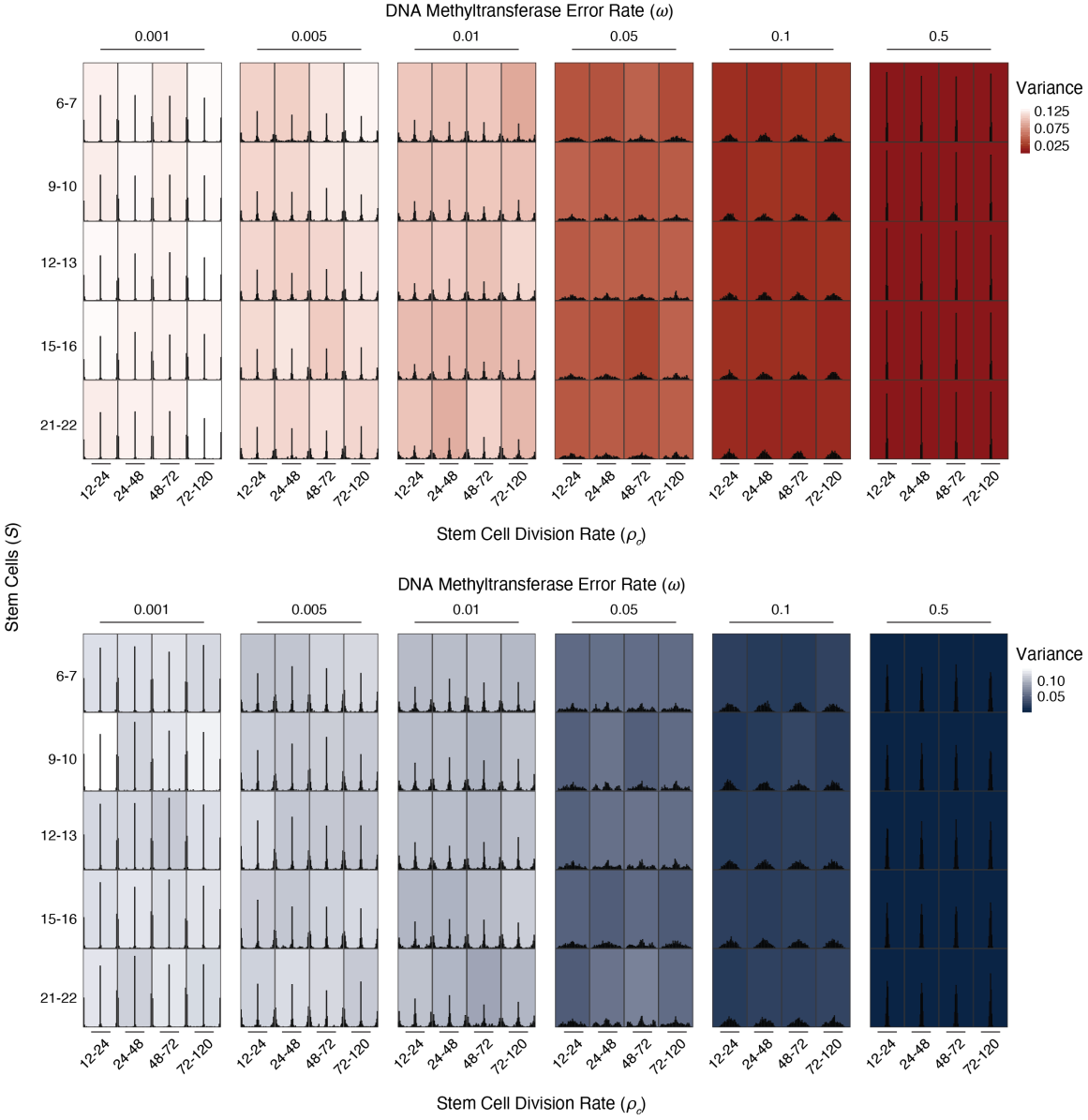


Figure B.4: Fluctuating CpG distribution at various (de)methylation rates for the Exclamatory Bowel. TODO Insert info here

Appendix C

Appendix: Chapter 4

C.1 Additional Information

C.1.1 Further restrictions to derivation of fluctuating CpG loci

Initial descriptions of how fCpG sites were derived can be found in 5.4 and published in our manuscript [210] are simple and naive to different methodologies used to assess methylation for the publicly available data. fCpG site isolation has also been previously described for sites within the colon and endometrium described in Chapter 4. Here we applied extra effort to restrict and normalize fCpG selection across the various publicly available datasets that utilize 450k, 850k, and HumanMethylation27 Illumina BeadChip technologies. Meaning some samples have a great deal more CpG sites to evaluate as candidate fCpGs. Using the raw, intensity normalized, methylation data, CpG sites are restricted to Type II probes and autosomes (i.e., no sex chromosomes). We then take only CpG sites that fall in the 'open sea', defined as 4kb from a CpG island. We then calculate the mean β values for each probe within each group, requiring the average beta value to fall between 0.4 and 0.6. From here we require that probes have beta values greater than and less than 0.5. Across the various array technologies this can result in a different total number of probes requiring that we take the same number of probes across all disease/normal groups. To accomplish this, we rank probes based on their variance from the most variable sites to the least variable. The top $12,000$ most variable sites are taken for further analysis for each group as the fCpGs of interest. Due to this filtering approach, Illumina HumanMethylation27 BeadChip arrays used for GSE42042 (i.e., PV, ET, and PMF samples) are eliminated due to lack of acceptable probes. Differences between these filtering approaches can be seen across the CH samples and publicly available data

Parameter	Description	Values
N	Number of HSCs, initial population	100
T	Simulation time	$2,500$
S	CpG error rates	$0.0001 - 0.001$ per division
λ	Cell survival, exact replacement	$0.0001 - 0.001$
E	Disease expansion rate	$(0.1, 0.005, 0.00225)$
ω	Final blood frequency	$(0.2, 0.5, 0.9)$

Table C.1: Parameters used in simulations describing how the methylation distribution of well-mixed HSCs changes in response to the expansion of a single clonal population used to generate Figure 5.2B.

in Figures 5.2A and 5.3C.

C.1.2 Hematopoietic stem cell model additional considerations

When performing blood model simulations where we consider concurrent founding clones and acquisition of new drivers (nested) when a subclone reaches the defined frequency symmetric division stops. This is an assumption that is not biologically likely. Based on the data and what is known about stem cell biology the growing subclones are likely to drift in and out without a selective advantage; however, here we pause this process by essentially freezing the signal at the appropriate frequencies to examine the fCpG β distribution variance. Likely these subclones would continue to expand at their respective expansion rates as the patient ages. However, normal HSCs cells are quiescent and divide infrequently [169] motivating the introduction of quiescence once a frequency is reached.

C.1.3 Hematopoietic stem cell model beta value corrections

Importantly, our data provides us with perfect information like that in the EB model and requires a transformation to reflect the expected technical error and uncertainty (this approach is first explained in Section B.1.4). We first perform a linear transform on the beta values ($\vec{m_\beta}$) from the model, where ($\hat{\vec{m}_\beta} = (\epsilon - \Delta) * \vec{m_\beta} + \Delta$), providing a beta value distribution more representative of the data. Noise is then added to each of the transformed beta values through an error model, which yields a list of "corrected" beta values ($\hat{\vec{m}_\beta} = [\hat{\beta}_0, \hat{\beta}_1, \dots]$), where $\hat{\beta}_i \sim B(\kappa\beta_i, \kappa(1-\beta_i))$ where $\kappa = 1000$. The variance of this transformed fCpG β distribution is denoted by ϕ, so that $\phi = Var(\hat{\vec{m}_\beta})$. For all transformations of the data output from the blood model

where the FMC-CH classifier and comparisons to patient data is conducted we use $\Delta = 0.04$ and $\epsilon = 0.92$.

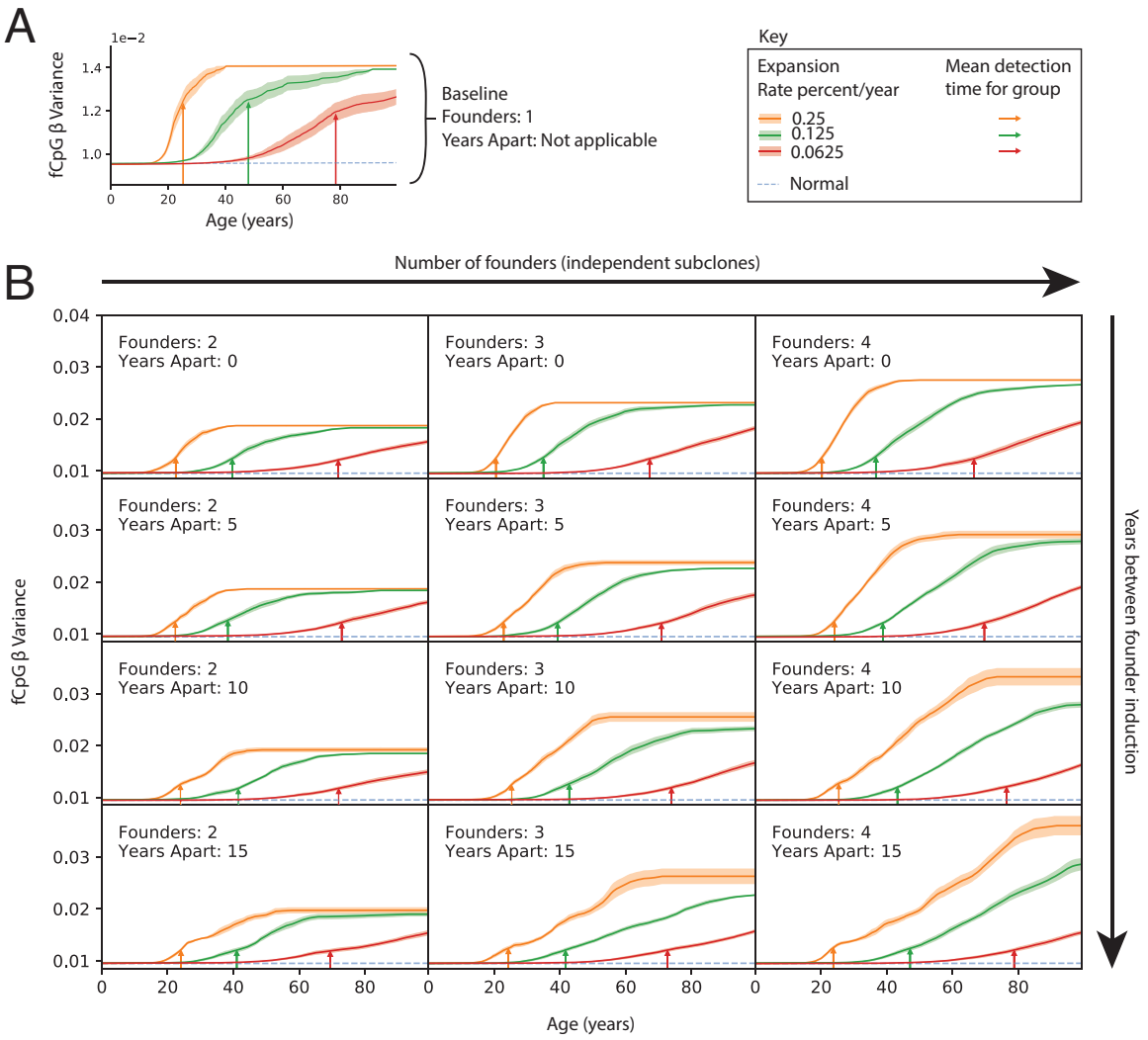


Figure C.1: Model results from multiple founders. Simulations with 500 HSCs to compare different expansion rates with multiple subclones (founders) and timings of founder induction. **A:** Baseline simulations with one founding subclone emerging at 5 years with three different expansion rates showing fCpG beta (β) distribution variance over time. **B:** Grid of fCpG β distribution variance over time for various founding subclones (2-4) on each column with induction times of each clone on each row (0, 5, 10, and 15 years apart). All expansion rates per year show the mean with *standard error of the mean* (SEM) as similar colored bands for 20 replicate simulations. Arrows denote the mean classification/detection time from the FMC-CH classifier for each corresponding expansion rate.

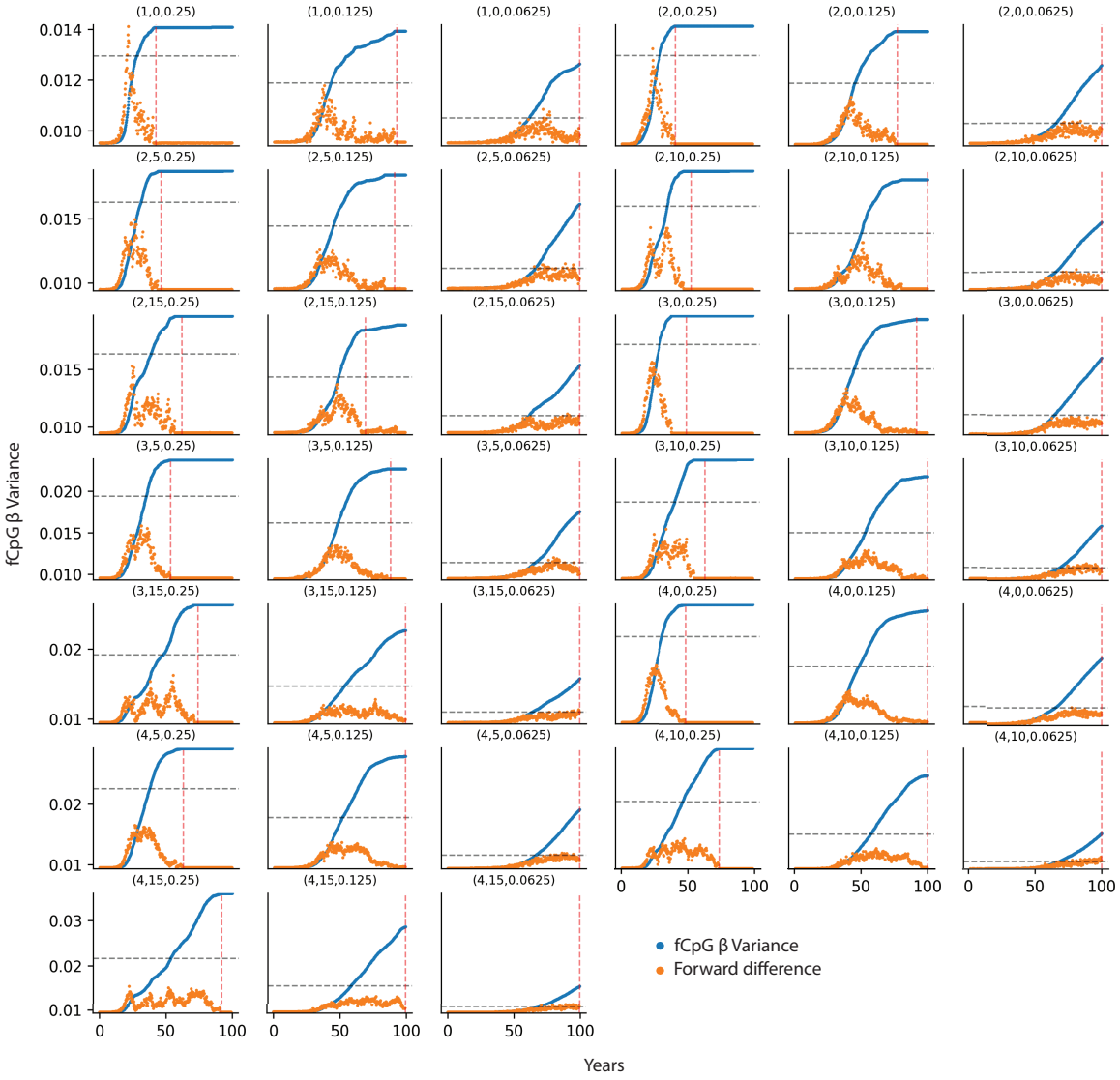


Figure C.2: Model steady state of fluctuating CpG beta variance. Illustration for determining when the steady state of the fCpG β distribution variance is reached for simulations from Figure 5.6C. The forward difference is taken from the fCpG β distribution variance and normalized, for purposes of plotting, to be on the same scale as the fCpG β distribution variance values. Only time points where the fCpG β distribution variance is above the mean fCpG β distribution variance (grey dashed line) are considered. The steady state is reached once the forward distance is less than $1.0 * 10^{-8}$ (red dashed line). Titles above each plot denote the number of subclones, the years between their induction, and the expansion rates.

References

[1] Merriam-Webster Dictionary. Merriam-webster. *On-line at http://www. mw. com/home. htm*, 14, 2002.

[2] MS Litwin and HJ Tan. The diagnosis and treatment of prostate cancer: a review. *Jama*, 317(24):2532–2542, 2017.

[3] AG Waks and EP Winer. Breast cancer treatment: a review. *Jama*, 321(3): 288–300, 2019.

[4] A McGuigan, P Kelly, RC Turkington, C Jones, HG Coleman, and RS McCain. Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes. *World journal of gastroenterology*, 24(43):4846, 2018.

[5] KJ Napier, M Scheerer, and S Misra. Esophageal cancer: A review of epidemiology, pathogenesis, staging workup and treatment modalities. *World journal of gastrointestinal oncology*, 6(5):112, 2014.

[6] EB Krumbhaar. Rôle of the blood and the bone marrow in certain forms of gas poisoning: I. peripheral blood changes and their significance. *Journal of the American Medical Association*, 72(1):39–41, 1919.

[7] C Curtis, SP Shah, SF Chin, G Turashvili, OM Rueda, MJ Dunning, D Speed, AG Lynch, S Samarajiwa, Y Yuan, and et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403): 346–352, 2012.

[8] N Vasan, J Baselga, and DM Hyman. A view on drug resistance in cancer. *Nature*, 575(7782):299–309, 2019.

[9] A Balmain. The critical roles of somatic mutations and environmental tumor-promoting agents in cancer risk. *Nature genetics*, 52(11):1139–1143, 2020.

[10] FR Greten and SI Grivennikov. Inflammation and cancer: triggers, mechanisms, and consequences. *Immunity*, 51(1):27–41, 2019.

[11] Mariam Jamal-Hanjani, GA Wilson, N McGranahan, NJ Birkbak, TBK Watkins, S Veeriah, S Shafi, DH Johnson, R Mitter, R Rosenthal, et al. Tracking the evolution of non–small-cell lung cancer. *New England Journal of Medicine*, 376(22):2109–2121, 2017.

[12] D Hanahan and RA Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[13] RE Baker, JM Peña, J Jayamohan, and A Jérusalem. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology letters*, 14(5):20170660, 2018.

[14] R Schenck. The research ecosystem: Where does mathematical oncology fit. https://ryanoschenck.com, 2019.

[15] P Armitage and R Doll. A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *British journal of cancer*, 11(2):161, 1957.

[16] MR Stratton, PJ Campbell, and PA Futreal. The cancer genome. *Nature*, 458 (7239):719–724, 2009.

[17] PM Altrock, LL Liu, and F Michor. The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer*, 15(12):730–745, 2015.

[18] K Sprouffske, JW Pepper, and CC Maley. Accurate reconstruction of the temporal order of mutations in neoplastic progression. *Cancer prevention research*, 4(7):1135–1144, 2011.

[19] A Sottoriva, H Kang, Z Ma, TA Graham, MP Salomon, J Zhao, P Marjoram, K Siegmund, MF Press, D Shibata, and C Curtis. A big bang model of human colorectal tumor growth. *Nature Genetics*, 47:209 EP –, 2015.

[20] EJ Mao, SM Schwartz, JR Daling, D Oda, L Tickman, and AM Beckmann. Human papilloma viruses and p53 mutations in normal, pre-malignant and malignant oral epithelia. *International journal of cancer*, 69(2):152–158, 1996.

[21] P Keohavong, HH Mady, WM Gao, JM Siegfried, JD Luketich, and MF Melhem. Topographic analysis of k-ras mutations in histologically normal lung

tissues and tumours of lung cancer patients. *British journal of cancer*, 85(2): 235–241, 2001.

[22] D Zhu, P Keohavong, SD Finkelstein, P Swalsky, A Bakker, J Weissfeld, S Srivastava, and TL Whiteside. K-ras gene mutations in normal colorectal tissues from k-ras mutation-positive colorectal cancer patients. *Cancer research*, 57 (12):2485–2492, 1997.

[23] R Li, L Di, J Li, W Fan, Y Liu, W Guo, W Liu, L Liu, Q Li, L Chen, Y Chen, C Miao, H Liu, Y Wang, Y Ma, D Xu, D Lin, Y Huang, J Wang, F Bai, and C Wu. A body map of somatic mutagenesis in morphologically normal human tissues. *Nature*, 597(7876):398–403, 2021.

[24] I Martincorena and PJ Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483, 2015.

[25] I Martincorena, A Roshan, M Gerstung, P Ellis, P Van Loo, S McLaren, DC Wedge, A Fullam, LB Alexandrov, JM Tubio, L Stebbings, A Menzies, S Widaa, MR Stratton, PH Jones, and PJ Campbell. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348 (6237):880–886, 2015.

[26] I Martincorena, JC Fowler, A Wabik, ARJ Lawson, F Abascal, MWJ Hall, A Cagan, K Murai, K Mahbubani, MR Stratton, RC Fitzgerald, PA Handford, PJ Campbell, K Saeb-Parsy, and PH Jones. Somatic mutant clones colonize the human esophagus with age. *Science*, 362(6417):911–917, 2018.

[27] X Cai, GD Evrony, HS Lehmann, PC Elhosary, BK Mehta, A Poduri, and CA Walsh. Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell reports*, 8(5):1280–1289, 2014.

[28] K Yizhak, F Aguet, J Kim, JM Hess, K Kübler, J Grimsby, R Frazer, H Zhang, NJ Haradhvala, D Rosebrock, D Livitz, X Li, E Arich-Landkof, N Shoresh, C Stewart, AV Segrè, PA Branton, P Polak, KG Ardlie, and G Getz. Rna sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science*, 364(6444), 2019.

[29] SR Kennedy, Y Zhang, and RA Risques. Cancer-associated mutations but no cancer: insights into the early steps of carcinogenesis and implications for early cancer detection. *Trends in cancer*, 5(9):531–540, 2019.

[30] B Colom, MP Alcolea, G Piedrafita, MWJ Hall, A Wabik, SC Dentro, JC Fowler, A Herms, C King, SH Ong, and et al. Spatial competition shapes the dynamic mutational landscape of normal esophageal epithelium. *Nature genetics*, 52(6):604–614, 2020.

[31] BD Simons. Deep sequencing as a probe of normal stem cell fate and preneoplasia in human epidermis. *PNAS; Proceedings of the National Academy of Sciences*, 113(1):128–133, 2016.

[32] MWJ Hall, PH Jones, and BA Hall. Relating evolutionary selection and mutant clonal dynamics in normal epithelia. *Journal of the Royal Society Interface*, 16 (156):20190230, 2019.

[33] JD Krimmel, MW Schmitt, MI Harrell, KJ Agnew, SR Kennedy, MJ Emond, LA Loeb, EM Swisher, and RA Risques. Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic tp53 mutations in noncancerous tissues. *Proceedings of the National Academy of Sciences*, 113(21): 6005–6010, 2016.

[34] J Matas, B Kohrn, J Fredrickson, KT Carter, M Yu, T Wang, X Gui, T Soussi, V Moreno, WM Grady, and et al. Colorectal cancer is associated with the presence of cancer driver mutations in normal colon. *medRxiv*, 2021.

[35] L Moore, D Leongamornlert, THH Coorens, MA Sanders, P Ellis, SC Dentro, KJ Dawson, T Butler, R Rahbari, TJ Mitchell, and et al. The mutational landscape of normal human endometrial epithelium. *Nature*, 580(7805):640–646, 2020.

[36] H Lee-Six, NF Øbro, MS Shepherd, S Grossmann, K Dawson, M Belmonte, RJ Osborne, BJP Huntly, I Martincorena, E Anderson, and et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature*, 561 (7724):473–478, 2018.

[37] MH Lee, B Siddoway, GE Kaeser, I Segota, R Rivera, WJ Romanow, CS Liu, C Park, G Kennedy, T Long, and et al. Somatic app gene recombination in alzheimer's disease and normal neurons. *Nature*, 563(7733):639–645, 2018.

[38] M Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

[39] RA Fisher. *The genetical theory of natural selection*. 1958.

[40] S Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931.

[41] J Vijg. Somatic mutations, genome mosaicism, cancer and aging. *Current opinion in genetics & development*, 26:141–149, 2014.

[42] JR Lupski. Genome mosaicism—one human, multiple genomes. *Science*, 341 (6144):358–359, 2013.

[43] K Kretzschmar and FM Watt. Lineage tracing. *Cell*, 148(1-2):33–45, 2012.

[44] S Tweedie, B Braschi, K Gray, TEM Jones, RL Seal, B Yates, and EA Bruford. Genenames. org: the hgnc and vgnc resources in 2021. *Nucleic Acids Research*, 49(D1):D939–D946, 2021.

[45] C Lopez-Garcia, AM Klein, BD Simons, and DJ Winton. Intestinal stem cell replacement follows a pattern of neutral drift. *Science*, 330(6005):822–825, 2010.

[46] HJ Snippert, LG Van Der Flier, T Sato, JH Van Es, M Van Den Born, C Kroon-Veenboer, N Barker, AM Klein, J Van Rheenen, BD Simons, and et al. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing lgr5 stem cells. *Cell*, 143(1):134–144, 2010.

[47] M Monné, KW Chan, DJ Slotboom, and ERS Kunji. Functional expression of eukaryotic membrane proteins in lactococcus lactis. *Protein Science*, 14(12): 3048–3056, 2005.

[48] S Han, J Fink, DJ Jörg, E Lee, MK Yum, L Chatzeli, SR Merker, M Josserand, T Trendafilova, A Andersson-Rolf, and et al. Defining the identity and dynamics of adult gastric isthmus stem cells. *Cell stem cell*, 25(3):342–356, 2019.

[49] HJ Snippert, AG Schepers, JH van Es, BD Simons, and H Clevers. Biased competition between lgr5 intestinal stem cells driven by oncogenic mutation induces clonal expansion. *EMBO reports*, 15(1):62–69, 2014.

[50] MS Andersen, E Hannezo, S Ulyanchenko, S Estrach, Y Antoku, S Pisano, KE Boonekamp, S Sendrup, M Maimets, MT Pedersen, and et al. Tracing the cellular dynamics of sebaceous gland development in normal and perturbed states. *Nature cell biology*, 21(8):924–932, 2019.

[51] X Lan, DJ Jörg, FMG Cavalli, LM Richards, LV Nguyen, RJ Vanner, P Guilhamon, L Lee, MM Kushida, D Pellacani, and et al. Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. *Nature*, 549(7671):227–232, 2017.

[52] A Sánchez-Danés, E Hannezo, JC Larsimont, M Liagre, KK Youssef, BD Simons, and C Blanpain. Defining the clonal dynamics leading to mouse skin tumour initiation. *Nature*, 536(7616):298–303, 2016.

[53] M Aragona, S Dekoninck, S Rulands, S Lenglez, G Mascré, BD Simons, and C Blanpain. Defining stem cell dynamics and migration during wound healing in mouse skin epidermis. *Nature communications*, 8(1):1–14, 2017.

[54] C Blanpain and BD Simons. Unravelling stem cell dynamics by lineage tracing. *Nature reviews Molecular cell biology*, 14(8):489–502, 2013.

[55] S Oota. Somatic mutations–evolution within the individual. *Methods*, 176: 91–98, 2020.

[56] RW Taylor, MJ Barron, GM Borthwick, A Gospel, PF Chinnery, DC Samuels, GA Taylor, SM Plusa, SJ Needham, LC Greaves, and et al. Mitochondrial dna mutations in human colonic crypt stem cells. *The Journal of clinical investigation*, 112(9):1351–1360, 2003.

[57] AM Baker, C Gabbutt, MJ Williams, B Cereser, N Jawad, M Rodriguez-Justo, M Jansen, CP Barnes, BD Simons, SAC McDonald, and et al. Crypt fusion as a homeostatic mechanism in the human colon. *Gut*, 68(11):1986–1993, 2019.

[58] DL Lavery, AM Nicholson, R Poulsom, R Jeffery, A Hussain, LJ Gay, JA Jankowski, SS Zeki, H Barr, R Harrison, and et al. The stem cell organisation, and the proliferative and gene expression profile of barrett's epithelium, replicates pyloric-type gastric glands. *Gut*, 63(12):1854–1863, 2014.

[59] NT Gaisa, TA Graham, SAC McDonald, S Cañadillas-Lopez, R Poulsom, A Heidenreich, G Jakse, PJ Tadrous, R Knuechel, and NA Wright. The human urothelium consists of multiple clonal units, each maintained by a stem cell. *The Journal of pathology*, 225(2):163–171, 2011.

[60] NT Gaisa, TA Graham, SAC McDonald, R Poulsom, A Heidenreich, G Jakse, R Knuechel, and NA Wright. Clonal architecture of human prostatic epithelium

in benign and malignant conditions. *The Journal of pathology*, 225(2):172–180, 2011.

[61] LS Ludwig, CA Lareau, JC Ulirsch, E Christian, C Muus, LH Li, K Pelka, W Ge, Y Oren, A Brack, and et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell*, 176(6):1325–1339, 2019.

[62] C Stamp, A Zupanic, A Sachdeva, EA Stoll, DP Shanley, JC Mathers, TBL Kirkwood, R Heer, BD Simons, DM Turnbull, and et al. Predominant asymmetrical stem cell fate outcome limits the rate of niche succession in human colonic crypts. *EBioMedicine*, 31:166–173, 2018.

[63] M Moad, E Hannezo, SJ Buczacki, L Wilson, A El-Sherif, D Sims, R Pickard, NA Wright, SC Williamson, DM Turnbull, and et al. Multipotent basal stem cells, maintained in localized proximal niches, support directed long-ranging epithelial flows in human prostates. *Cell reports*, 20(7):1609–1622, 2017.

[64] B Cereser, M Jansen, E Austin, G Elia, T McFarlane, CHM van Deurzen, AM Sieuwerts, MG Daidone, PJ Tadrous, NA Wright, and et al. Analysis of clonal expansions through the normal and premalignant human breast epithelium reveals the presence of luminal stem cells. *The Journal of pathology*, 244 (1):61–70, 2018.

[65] L Gutierrez-Gonzalez, M Deheragoda, G Elia, SJ Leedham, A Shankar, C Imber, JA Jankowski, DM Turnbull, M Novelli, NA Wright, and et al. Analysis of the clonal architecture of the human small intestinal epithelium establishes a common stem cell for all lineages and reveals a mechanism for the fixation and spread of mutations. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 217(4):489–496, 2009.

[66] LC Greaves, SL Preston, PJ Tadrous, RW Taylor, MJ Barron, D Oukrif, SJ Leedham, M Deheragoda, P Sasieni, MR Novelli, and et al. Mitochondrial dna mutations are established in human colonic stem cells, and mutated clones expand by crypt fission. *Proceedings of the National Academy of Sciences*, 103 (3):714–719, 2006.

[67] AM Baker, B Cereser, S Melton, AG Fletcher, M Rodriguez-Justo, PJ Tadrous, A Humphries, G Elia, SAC McDonald, NA Wright, and et al. Quantification of crypt and stem cell evolution in the normal and neoplastic human colon. *Cell reports*, 8(4):940–947, 2014.

[68] TG Fellous, S Islam, PJ Tadrous, G Elia, HM Kocher, S Bhattacharya, L Mears, DM Turnbull, RW Taylor, LC Greaves, and et al. Locating the stem cell niche and tracing hepatocyte lineages in human liver. *Hepatology*, 49(5):1655–1663, 2009.

[69] A Humphries, B Cereser, LJ Gay, DSJ Miller, B Das, A Gutteridge, G Elia, E Nye, R Jeffery, R Poulsom, and et al. Lineage tracing reveals multipotent stem cells maintain human adenomas and the pattern of clonal expansion in tumor evolution. *Proceedings of the National Academy of Sciences*, 110(27): E2490–E2499, 2013.

[70] TA Graham, A Humphries, T Sanders, M Rodriguez-Justo, PJ Tadrous, SL Preston, MR Novelli, SJ Leedham, SAC McDonald, and NA Wright. Use of methylation patterns to determine expansion of stem cell clones in human colon tissue. *Gastroenterology*, 140(4):1241–1250, 2011.

[71] Y Yatabe, S Tavaré, and D Shibata. Investigating stem cells in human colon by using methylation patterns. *Proceedings of the National Academy of Sciences*, 98(19):10839–10844, 2001.

[72] P Nicolas, KM Kim, D Shibata, and S Tavaré. The stem cell population of the human colon crypt: analysis via methylation patterns. *PLoS computational biology*, 3(3):e28, 2007.

[73] MW Chu, KD Siegmund, CL Eckstam, JY Kim, AS Yang, GC Kanel, S Tavaré, and D Shibata. Lack of increases in methylation at three cpg-rich genomic loci in non-mitotic adult tissues during aging. *BMC medical genetics*, 8(1):1–10, 2007.

[74] JY Kim, KD Siegmund, S Tavaré, and D Shibata. Age-related human small intestine methylation: evidence for stem cell niches. *BMC medicine*, 3(1):1–11, 2005.

[75] JY Kim, S Tavaré, and D Shibata. Human hair genealogies and stem cell latency. *BMC biology*, 4(1):1–10, 2006.

[76] KD Siegmund, P Marjoram, YJ Woo, S Tavaré, and D Shibata. Inferring clonal expansion and cancer stem cell dynamics from dna methylation patterns in colorectal cancers. *Proceedings of the National Academy of Sciences*, 106(12): 4828–4833, 2009.

[77] KM Kim and D Shibata. Tracing ancestry with methylation patterns: most crypts appear distantly related in normal adult human colon. *BMC gastroenterology*, 4(1):1–10, 2004.

[78] F Campbell, CE Fuller, GT Williams, and ED Williams. Human colonic stem cell mutation frequency with and without irradiation. *The Journal of pathology*, 174(3):175–182, 1994.

[79] M Novelli, A Cossu, D Oukrif, A Quaglia, S Lakhani, R Poulsom, P Sasieni, P Carta, M Contini, A Pasca, and et al. X-inactivation patch size in human female tissue confounds the assessment of tumor clonality. *Proceedings of the National Academy of Sciences*, 100(6):3311–3314, 2003.

[80] CJ Watson, AL Papula, GYP Poon, WH Wong, AL Young, TE Druley, DS Fisher, and JR Blundell. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science*, 367(6485):1449–1454, 2020.

[81] F Blokzijl, J De Ligt, M Jager, V Sasselli, S Roerink, N Sasaki, M Huch, S Boymans, E Kuijk, P Prins, and et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, 538(7624):260–264, 2016.

[82] MJ Williams, B Werner, CP Barnes, TA Graham, and A Sottoriva. Identification of neutral tumor evolution across cancer types. *Nature genetics*, 48(3): 238–244, 2016.

[83] MJ Williams, B Werner, T Heide, C Curtis, CP Barnes, A Sottoriva, and TA Graham. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics*, 50(6):895–903, 2018.

[84] S Galandiuk, M Rodriguez-Justo, R Jeffery, AM Nicholson, Y Cheng, D Oukrif, G Elia, SJ Leedham, SAC McDonald, NA Wright, and et al. Field cancerization in the intestinal epithelium of patients with crohn's ileocolitis. *Gastroenterology*, 142(4):855–864, 2012.

[85] CP Pipinikas, TS Kiropoulos, VH Teixeira, JM Brown, A Varanou, M Falzon, A Capitanio, SE Bottoms, B Carroll, N Navani, and et al. Cell migration leads to spatially distinct but clonally related airway cancer precursors. *Thorax*, 69 (6):548–557, 2014.

[86] SJ Leedham, SL Preston, SAC McDonald, G Elia, P Bhandari, D Poller, R Harrison, MR Novelli, JA Jankowski, and NA Wright. Individual crypt genetic

heterogeneity and the origin of metaplastic glandular epithelium in human barrett's oesophagus. *Gut*, 57(8):1041–1048, 2008.

[87] C Thirlwell, OCC Will, E Domingo, TA Graham, SAC McDonald, D Oukrif, R Jeffrey, M Gorman, M Rodriguez-Justo, J Chin-Aleong, and et al. Clonality assessment and clonal ordering of individual neoplastic crypts shows polyclonality of colorectal adenomas. *Gastroenterology*, 138(4):1441–1454, 2010.

[88] SJ Leedham, TA Graham, D Oukrif, SAC McDonald, M Rodriguez-Justo, RF Harrison, NA Shepherd, MR Novelli, JAZ Jankowski, and NA Wright. Clonality, founder mutations, and field cancerization in human ulcerative colitis–associated neoplasia. *Gastroenterology*, 136(2):542–550, 2009.

[89] RO Schenck, E Kim, RR Bravo, J West, S Leedham, D Shibata, and ARA Anderson. How homeostasis limits keratinocyte evolution. *bioRxiv*, 2019.

[90] K Murai, G Skrupskelyte, G Piedrafita, M Hall, V Kostiou, SH Ong, T Nagy, A Cagan, D Goulding, AM Klein, and et al. Epidermal tissue adapts to restrain progenitors carrying clonal p53 mutations. *Cell Stem Cell*, 23(5):687–699.e8, 2018.

[91] J West, R Schenck, C Gatenbee, M Robertson-Tessi, and ARA Anderson. Tissue structure accelerates evolution: premalignant sweeps precede neutral expansion. *bioRxiv*, page 542019, 2019.

[92] R Noble, D Burri, C Le Sueur, J Lemant, Y Viossat, JN Kather, and N Beerenwinkel. Spatial structure governs the mode of tumour evolution. *Nature ecology & evolution*, 6(2):207–217, 2022.

[93] K Chkhaidze, T Heide, B Werner, MJ Williams, W Huang, G Caravagna, TA Graham, and A Sottoriva. Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. *PLoS computational biology*, 15(7):e1007243, 2019.

[94] MS Lawrence, P Stojanov, P Polak, GV Kryukov, K Cibulskis, A Sivachenko, SL Carter, C Stewart, CH Mermel, SA Roberts, and et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499 (7457):214–218, 2013.

[95] M Gerlinger, AJ Rowan, S Horswell, J Larkin, D Endesfelder, E Gronroos, P Martinez, N Matthews, A Stewart, P Tarpey, and et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl j Med*, 366:883–892, 2012.

[96] RO Schenck, E Lakatos, C Gatenbee, TA Graham, and ARA Anderson. Neopredpipe: high-throughput neoantigen prediction and recognition potential pipeline. *BMC bioinformatics*, 20(1):1–6, 2019.

[97] E Lakatos, MJ Williams, RO Schenck, WCH Cross, J Househam, L Zapata, B Werner, C Gatenbee, M Robertson-Tessi, CP Barnes, and et al. Evolutionary dynamics of neoantigens in growing tumors. *Nature genetics*, 52(10):1057–1066, 2020.

[98] C Gatenbee, J West, AM Baker, N Guljar, L Jones, TA Graham, M Robertson-Tessi, and ARA Anderson. Macrophage-mediated immunoediting drives ductal carcinoma evolution: space is the game changer. *bioRxiv*, page 594598, 2019.

[99] AM Leroi, B Lambert, J Rosindell, X Zhang, and GD Kokkoris. Neutral syndrome. *Nature human behaviour*, 4(8):780–790, 2020.

[100] TO McDonald, S Chakrabarti, and F Michor. Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution. *Nature genetics*, 50(12):1620–1623, 2018.

[101] MJ Williams, B Werner, T Heide, CP Barnes, TA Graham, and A Sottoriva. Reply to 'revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data'. *Nature genetics*, 50(12):1628–1630, 2018.

[102] A Balaparya and S De. Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data. *Nature genetics*, 50(12):1626–1628, 2018.

[103] B Werner, MJ Williams, CP Barnes, TA Graham, and A Sottoriva. Reply to 'currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution'. *Nature genetics*, 50(12):1624–1626, 2018.

[104] M Tarabichi, I Martincorena, M Gerstung, AM Leroi, F Markowetz, PT Spellman, QD Morris, OC Lingjærde, DC Wedge, and P Van Loo. Neutral tumor evolution? *Nature genetics*, 50(12):1630–1633, 2018.

[105] T Heide, L Zapata, MJ Williams, B Werner, G Caravagna, CP Barnes, TA Graham, and A Sottoriva. Reply to 'neutral tumor evolution?'. *Nature genetics*, 50(12):1633–1637, 2018.

[106] BD Simons. Reply to martincorena et al.: Evidence for constrained positive selection of cancer mutations in normal skin is lacking. *Proceedings of the National Academy of Sciences*, 113(9):E1130–E1131, 2016.

[107] I Martincorena, PH Jones, and PJ Campbell. Constrained positive selection on cancer mutations in normal skin. *Proceedings of the National Academy of Sciences*, 113(9):E1128–E1129, 2016.

[108] J West, RO Schenck, C Gatenbee, M Robertson-Tessi, and ARA Anderson. Normal tissue architecture determines the evolutionary course of cancer. *Nature communications*, 12(1):1–9, 2021.

[109] RR Bravo, E Baratchart, J West, RO Schenck, AK Miller, J Gallaher, CD Gatenbee, D Basanta, M Robertson-Tessi, and ARA Anderson. Hybrid automata library: A flexible platform for hybrid modeling with real-time visualization. *PLOS Computational Biology*, 16(3):1–28, 2020.

[110] GEO NCBI. archive for functional genomics data sets—update. *Google Search*, 2012.

[111] D Karolchik, R Baertsch, M Diekhans, TS Furey, A Hinrichs, YT Lu, KM Roskin, M Schwartz, CW Sugnet, DJ Thomas, and et al. The ucsc genome browser database. *Nucleic acids research*, 31(1):51–54, 2003.

[112] C Kanz, P Aldebert, N Althorpe, W Baker, A Baldwin, K Bates, P Browne, A van den Broek, M Castro, G Cochrane, and et al. The embl nucleotide sequence database. *Nucleic acids research*, 33(suppl_1):D29–D33, 2005.

[113] N Siva. 1000 genomes project. *Nature biotechnology*, 26(3):256–257, 2008.

[114] Cancer Genome Atlas Network and et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330, 2012.

[115] KD Pruitt, T Tatusova, and DR Maglott. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(suppl_1):D61–D65, 2007.

[116] GA Logsdon, MR Vollger, and EE Eichler. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10):597–614, 2020.

[117] SH Wu, RS Schwartz, DJ Winter, DF Conrad, and RA Cartwright. Estimating error models for whole genome sequencing using mixtures of dirichlet-multinomial distributions. *Bioinformatics*, 33(15):2322–2329, 2017.

[118] Y Fang, A Brass, DC Hoyle, A Hayes, A Bashein, SG Oliver, D Waddington, and M Rattray. A model-based analysis of microarray experimental error and normalisation. *Nucleic acids research*, 31(16):e96–e96, 2003.

[119] A Schumacher, P Kapranov, Z Kaminsky, J Flanagan, A Assadzadeh, P Yau, C Virtanen, N Winegarden, J Cheng, T Gingeras, and et al. Microarray-based dna methylation profiling: technology and applications. *Nucleic acids research*, 34(2):528–542, 2006.

[120] D Skvortsov, D Abdueva, C Curtis, B Schaub, and S Tavare. Explaining differences in saturation levels for affymetrix genechip® arrays. *Nucleic acids research*, 35(12):4154–4163, 2007.

[121] M Meyer and M Kircher. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, 2010 (6):pdb–prot5448, 2010.

[122] L Liu, Y Li, S Li, N Hu, Y He, R Pong, D Lin, L Lu, and M Law. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 2012, 2012.

[123] TW Sadler. *Langman's medical embryology*. Lippincott Williams & Wilkins, 2018.

[124] FM Watt. Epidermal stem cells: markers, patterning and the control of stem cell fate. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1370):831–837, 1998.

[125] AE Shyer, TR Huycke, CH Lee, L Mahadevan, and CJ Tabin. Bending gradients: How the intestinal stem cell gets its home. *Cell*, 161(3):569–580, 2015.

[126] SJ Leedham, TA Graham, D Oukrif, SAC McDonald, M Rodriguez-Justo, RF Harrison, NA Shepherd, MR Novelli, JAZ Jankowski, and NA Wright. Clonality, founder mutations, and field cancerization in human ulcerative colitis–associated neoplasia. *Gastroenterology*, 136(2):542–550.e6, 2009.

[127] S Biswas, H Davis, S Irshad, T Sandberg, D Worthley, and S Leedham. Microenvironmental control of stem cell fate in intestinal homeostasis and disease. *The Journal of Pathology*, 237(2):135–145, 2015.

[128] S Brabletz, O Schmalhofer, and T Brabletz. Gastrointestinal stem cells in development and cancer. *The Journal of Pathology*, 217(2):307–317, 2009.

[129] SJ Leedham, AT Thliveris, RB Halberg, MA Newton, and NA Wright. Gastrointestinal stem cells and cancer. *Stem Cell Reviews*, 1(3):233–241, 2005.

[130] M Tavian, K Biasch, L Sinka, J Vallet, and B Péault. Embryonic origin of human hematopoiesis. *International journal of developmental biology*, 54(6-7): 1061–1065, 2010.

[131] K Bollerot, C Pouget, and T Jaffredo. The embryonic origins of hematopoietic stem cells: a tale of hemangioblast and hemogenic endothelium. *Apmis*, 113 (11-12):790–803, 2005.

[132] S Pinho and PS Frenette. Haematopoietic stem cell activity and interactions with the niche. *Nature Reviews Molecular Cell Biology*, 20(5):303–320, 2019.

[133] M Jagannathan-Bogdan and LI Zon. Hematopoiesis. *Development*, 140(12): 2463–2467, 2013.

[134] MJ Kiel, T Iwashita, OH Yilmaz, and SJ Morrison. Spatial differences in hematopoiesis but not in stem cells indicate a lack of regional patterning in definitive hematopoietic stem cells. *Developmental biology*, 283(1):29–39, 2005.

[135] MA Rieger and T Schroeder. Hematopoiesis. *Cold Spring Harbor perspectives in biology*, 4(12):a008250, 2012.

[136] T Köhnke and R Majeti. Clonal hematopoiesis: From mechanisms to clinical intervention. *Cancer Discovery*, 11(12):2987–2997, 2021.

[137] R Edgar, M Domrachev, and AE Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30 (1):207–210, 2002.

[138] T Barrett, SE Wilhite, P Ledoux, C Evangelista, IF Kim, M Tomashevsky, KA Marshall, KH Phillippy, PM Sherman, M Holko, and et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41 (D1):D991–D995, 2012.

[139] G Hannum, J Guinney, L Zhao, LI Zhang, G Hughes, SV Sadda, B Klotzle, M Bibikova, JB Fan, Y Gao, and et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell*, 49(2):359–367, 2013.

[140] M I Sierra, A F Fernández, and M F Fraga. Epigenetics of aging. *Current genomics*, 16(6):435–440, 2015.

[141] X Zhao, F Yang, S Li, M Liu, S Ying, X Jia, and X Wang. Cpg island methylator phenotype of myelodysplastic syndrome identified through genome-wide profiling of dna methylation and gene expression. *British journal of haematology*, 165(5):649–658, 2014.

[142] C Pérez, M Pascual, JI Martín-Subero, B Bellosillo, V Segura, E Delabesse, S Álvarez, MJ Larrayoz, J Rifón, JC Cigudosa, and et al. Aberrant dna methylation profile of chronic and transformed classic philadelphia-negative myeloproliferative neoplasms. *haematologica*, 98(9):1414, 2013.

[143] S Maupetit-Mehouas, F Court, C Bourgne, A Guerci-Bresler, P Cony-Makhoul, H Johnson, G Etienne, P Rousselot, D Guyotat, A Janel, and et al. Dna methylation profiling reveals a pathological signature that contributes to transcriptional defects of cd 34+ cd 15- cells in early chronic-phase chronic myeloid leukemia. *Molecular oncology*, 12(6):814–829, 2018.

[144] L Palomo, R Malinverni, M Cabezón, B Xicoy, M Arnan, R Coll, H Pomares, O García, F Fuster-Tormo, J Grau, and et al. Dna methylation profile in chronic myelomonocytic leukemia associates with distinct clinical, biological and genetic features. *Epigenetics*, 13(1):8–18, 2018.

[145] HJ Ferreira, H Heyn, M Vizoso, C Moutinho, E Vidal, A Gomez, A Martínez-Cardús, L Simó-Riudalbas, S Moran, E Jost, and et al. Dnmt3a mutations mediate the epigenetic reactivation of the leukemogenic factor meis1 in acute myeloid leukemia. *Oncogene*, 35(23):3079–3082, 2016.

[146] AS Gabriel, FM Lafta, EC Schwalbe, S Nakjang, SJ Cockell, A Iliasova, A Enshaei, C Schwab, V Rand, SC Clifford, and et al. Epigenetic landscape correlates with genetic subtype but does not predict outcome in childhood acute lymphoblastic leukemia. *Epigenetics*, 10(8):717–726, 2015.

[147] AA Johansson, S Enroth, and U Gyllensten. Continuous aging of the human dna methylome throughout the human lifespan. *PloS one*, 8(6):e67378, 2013.

[148] JY Kim, S Tavaré, and D Shibata. Counting human somatic cell replications: methylation mirrors endometrial stem cell divisions. *Proceedings of the National Academy of Sciences*, 102(49):17739–17744, 2005.

[149] Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine*, 368(22):2059–2074, 2013.

[150] E Papaemmanuil, M Gerstung, L Bullinger, VI Gaidzik, P Paschka, ND Roberts, NE Potter, M Heuser, F Thol, N Bolli, and et al. Genomic classification and prognosis in acute myeloid leukemia. *New England Journal of Medicine*, 374(23):2209–2221, 2016.

[151] S Jaiswal, P Fontanillas, J Flannick, A Manning, PV Grauman, BG Mar, RC Lindsley, CH Mermel, N Burtt, A Chavez, and et al. Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine*, 371(26):2488–2498, 2014.

[152] G Genovese, AK Kähler, RE Handsaker, J Lindberg, SA Rose, SF Bakhoum, K Chambert, E Mick, BM Neale, M Fromer, and et al. Clonal hematopoiesis and blood-cancer risk inferred from blood dna sequence. *New England Journal of Medicine*, 371(26):2477–2487, 2014.

[153] M Xie, C Lu, J Wang, MD McLellan, KJ Johnson, MC Wendl, JF McMichael, HK Schmidt, V Yellapantula, CA Miller, and et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature medicine*, 20(12):1472–1478, 2014.

[154] M Buscarlet, S Provost, YF Zada, A Barhdadi, V Bourgoin, G Lépine, L Mollica, N Szuber, MP Dubé, and L Busque. Dnmt3a and tet2 dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions. *Blood, The Journal of the American Society of Hematology*, 130(6): 753–762, 2017.

[155] R Acuna-Hidalgo, H Sengul, M Steehouwer, M van de Vorst, SH Vermeulen, LALM Kiemeney, JA Veltman, C Gilissen, and A Hoischen. Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life. *The American Journal of Human Genetics*, 101(1): 50–64, 2017.

[156] H Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.

[157] Z Lai, A Markovets, M Ahdesmaki, B Chapman, O Hofmann, R McEwen, J Johnson, B Dougherty, JC Barrett, and JR Dry. Vardict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic acids research*, 44(11):e108–e108, 2016.

[158] K Cibulskis, MS Lawrence, SL Carter, A Sivachenko, D Jaffe, C Sougnez, S Gabriel, M Meyerson, ES Lander, and G Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213–219, 2013.

[159] H Yang and K Wang. Genomic variant annotation and prioritization with annovar and wannovar. *Nature protocols*, 10(10):1556–1566, 2015.

[160] S Abelson, G Collord, SWK Ng, O Weissbrod, NM Cohen, E Niemeyer, N Barda, PC Zuzarte, L Heisler, Y Sundaravadanam, and et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature*, 559(7714):400–404, 2018.

[161] P Desai, N Mencia-Trinchant, O Savenkov, MS Simon, G Cheang, S Lee, M Samuel, EK Ritchie, ML Guzman, KV Ballman, and et al. Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nature medicine*, 24(7):1015–1023, 2018.

[162] *Infinium HD FFPE Restore Protocol*. Illumina, 2011.

[163] TJ Morris, LM Butcher, A Feber, AE Teschendorff, AR Chakravarthy, TK Wojdacz, and S Beck. Champ: 450k chip analysis methylation pipeline. *Bioinformatics*, 30(3):428–430, 2014.

[164] MJ Aryee, AE Jaffe, H Corrada-Bravo, C Ladd-Acosta, AP Feinberg, KD Hansen, and RA Irizarry. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014.

[165] W Zhou, PW Laird, and H Shen. Comprehensive characterization, annotation and innovative use of infinium dna methylation beadchip probes. *Nucleic acids research*, 45(4):e22–e22, 2017.

[166] J Nordlund, CL Bäcklin, P Wahlberg, S Busche, EC Berglund, ML Eloranta, T Flaegstad, E Forestier, BM Frost, A Harila-Saari, and et al. Genome-wide signatures of differential dna methylation in pediatric acute lymphoblastic leukemia. *Genome biology*, 14(9):1–15, 2013.

[167] A Feber, P Guilhamon, M Lechner, T Fenton, GA Wilson, C Thirlwell, AM Flanagan, AE Teschendorff, JD Kelly, and S Beck. Cna profiling using high density dna methylation arrays. *Genome Biol*, 2013.

[168] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[169] GA Challen and MA Goodell. Clonal hematopoiesis: mechanisms driving dominance of stem cell clones. *Blood*, 136(14):1590–1598, 2020.

[170] AR Quinlan and IM Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[171] KD Pruitt, T Tatusova, and DR Maglott. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 33(suppl_1):D501–D504, 2005.

[172] DW Huang, BT Sherman, and RA Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4 (1):44–57, 2009.

[173] DW Huang, BT Sherman, and RA Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2009.

[174] F Martínez-Jiménez, F Muiños, I Sentís, J Deu-Pons, I Reyes-Salazar, C Arnedo-Pac, L Mularoni, O Pich, J Bonet, H Kranas, and et al. A compendium of mutational cancer driver genes. *Nature Reviews Cancer*, 20(10): 555–572, 2020.

[175] J Oksanen, FG Blanchet, M Friendly, R Kindt, P Legendre, D McGlinn, PR Minchin, RB O'Hara, GL Simpson, P Solymos, MHH Stevens, E Szoecs, and H Wagner. *vegan: Community Ecology Package*, 2017. R package version 2.4-4.

[176] R Goya, MGF Sun, RD Morin, G Leung, G Ha, KC Wiegand, J Senz, A Crisan, MA Marra, M Hirst, and et al. Snvmix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, 26(6):730–736, 2010.

[177] RA Cartwright, J Hussin, JEM Keebler, EA Stone, and P Awadalla. A family-based probabilistic method for capturing de novo mutations from high-throughput short-read sequencing data. *Statistical applications in genetics and molecular biology*, 11(2), 2012.

[178] PA Hohenlohe, S Bassham, PD Etter, N Stiffler, EA Johnson, and WA Cresko. Population genomics of parallel adaptation in threespine stickleback using sequenced rad tags. *PLoS genetics*, 6(2):e1000862, 2010.

[179] H Li, J Ruan, and R Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851–1858, 2008.

[180] M Lynch. Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics*, 182(1):295–301, 2009.

[181] T Maruki and M Lynch. Genotype-frequency estimation from high-throughput sequencing data. *Genetics*, 201(2):473–486, 2015.

[182] DJ Venzon and SH Moolgavkar. A method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 37(1):87–94, 1988.

[183] MD Lynch, CNS Lynch, E Craythorne, K Liakath-Ali, R Mallipeddi, JN Barker, and FM Watt. Spatial constraints govern competition of mutant clones in human epidermis. *Nature Communications*, 8(1):1119, 2017.

[184] A Yokoyama, N Kakiuchi, T Yoshizato, Y Nannya, H Suzuki, Y Takeuchi, Y Shiozawa, Y Sato, K Aoki, SK Kim, and et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature*, 565(7739):312–317, 2019.

[185] E Kim, V Rebecca, IV Fedorenko, JL Messina, R Mathew, SS Maria-Engler, D Basanta, KSM Smalley, and ARA Anderson. Senescent fibroblasts in melanoma initiation and progression: an integrated theoretical, experimental, and clinical approach. *Cancer research*, 73(23):6874–6885, 2013.

[186] D Basanta, DW Strand, RB Lukner, OE Franco, DE Cliffel, GE Ayala, SW Hayward, and ARA Anderson. The role of transforming growth factor-β–mediated tumor-stroma interactions in prostate cancer progression: An integrative approach. *Cancer research*, 69(17):7111–7120, 2009.

[187] T Hirsch, T Rothoeft, N Teig, JW Bauer, G Pellegrini, L De Rosa, D Scaglione, J Reichelt, A Klausegger, D Kneisz, and et al. Regeneration of the entire human epidermis using transgenic stem cells. *Nature*, 551(7680):327–332, 2017.

[188] C Blanpain and E Fuchs. Epidermal stem cells of the skin. *Annu. Rev. Cell Dev. Biol.*, 22:339–373, 2006.

[189] E Enzo, AS Seconetti, M Forcato, E Tenedini, MP Polito, I Sala, S Carulli, R Contin, C Peano, E Tagliafico, and et al. Single-keratinocyte transcriptomic analyses identify different clonal types and proliferative potential mediated by foxm1 in human epidermal stem cells. *Nature communications*, 12(1):1–15, 2021.

[190] Y Barrandon and H Green. Three clonal types of keratinocyte with different capacities for multiplication. *Proceedings of the National Academy of Sciences*, 84(8):2302–2306, 1987.

[191] RO Schenck, G Brosula, J West, S Leedham, D Shibata, and ARA Anderson. Gattaca: Base-Pair Resolution Mutation Tracking for Somatic Evolution Studies using Agent-based Models. *Molecular Biology and Evolution*, 39(4), 2022. msac058.

[192] P Cingolani, A Platts, LL Wang, M Coon, T Nguyen, L Wang, SJ Land, X Lu, and DM Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly*, 6(2):80–92, 2012.

[193] HJ Muller. Some genetic aspects of sex. *The American Naturalist*, 66(703): 118–138, 1932.

[194] HJ Muller. The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 1(1):2–9, 1964.

[195] RJ Butler and JB McDonald. Using incomplete moments to measure inequality. *Journal of Econometrics*, 42(1):109–119, 1989.

[196] CD Gatenbee, RO Schenck, RR Bravo, and ARA Anderson. Evofreq: visualization of the evolutionary frequencies of sequence and model data. *BMC Bioinformatics*, 20(1):710, 2019.

[197] DJ Araten, DW Golde, RH Zhang, HT Thaler, L Gargiulo, R Notaro, and L Luzzatto. A quantitative measurement of the human somatic mutation rate. *Cancer Res; Cancer Research*, 65(18):8111–8117, 2005.

[198] M Lynch. Rate, molecular spectrum, and consequences of human mutation. *PNAS; Proceedings of the National Academy of Sciences*, 107(3):961–968, 2010.

[199] B Milholland, X Dong, L Zhang, X Hao, Y Suh, and J Vijg. Differences between germline and somatic mutation rates in humans and mice. *Nature Communications*, 8(1), 2017.

[200] CA Miller, J McMichael, HX Dang, CA Maher, L Ding, TJ Ley, ER Mardis, and RK Wilson. Visualizing tumor evolution with the fishplot package for r. *BMC genomics*, 17(1):1–3, 2016.

[201] R Noble. *ggmuller: Create Muller Plots of Evolutionary Dynamics*, 2018. R package version 0.5.1.

[202] MWJ Hall, PH Jones, and BA Hall. Relating evolutionary selection and mutant clonal dynamics in normal epithelia. *bioRxiv*, 2019.

[203] MD Lynch, CNS Lynch, E Craythorne, K Liakath-Ali, R Mallipeddi, JN Barker, and FM Watt. Spatial constraints govern competition of mutant clones in human epidermis. *Nature Communications*, 8(1), 2017.

[204] S Kangaspeska, B Stride, R Métivier, M Polycarpou-Schwarz, D Ibberson, RP Carmouche, V Benes, F Gannon, and G Reid. Transient cyclical methylation of promoter dna. *Nature*, 452(7183):112–115, 2008.

[205] R Métivier, R Gallais, C Tiffoche, C Le Péron, RZ Jurkowska, RP Carmouche, D Ibberson, P Barath, F Demay, G Reid, and et al. Cyclical dna methylation of a transcriptionally active promoter. *Nature*, 452(7183):45–50, 2008.

[206] S Rulands, HJ Lee, SJ Clark, C Angermueller, SA Smallwood, F Krueger, H Mohammed, W Dean, J Nichols, P Rugg-Gunn, and et al. Genome-scale oscillations in dna methylation during exit from pluripotency. *Cell systems*, 7 (1):63–76, 2018.

[207] A Parry, S Rulands, and W Reik. Active turnover of dna methylation during cell fate decisions. *Nature Reviews Genetics*, 22(1):59–66, 2021.

[208] AM Nicholson, C Olpe, A Hoyle, AS Thorsen, T Rus, M Colombe, R Brunton-Sim, R Kemp, K Marks, P Quirke, and et al. Fixation and spread of somatic mutations in adult human colonic epithelium. *Cell Stem Cell*, 22(6):909–918, 2018.

[209] Schenck RO, Gabbutt C, Weisenberger D, Kimberley C, Berner A, Household J, Lakatos E, Robertson-Tessi M, Martin I, Patel R, Clark S, Latchford A, Barnes C, Leedham SJ, Anderson ARA, Graham TA, and Shibata D. Cell lineage tracing with molecular clocks based on fluctuating DNA methylation – Flip flop spatial model, 2021.

[210] C Gabbutt, RO Schenck, DJ Weisenberger, C Kimberley, A Berner, J Household, E Lakatos, M Robertson-Tessi, I Martin, R Patel, and et al. Fluctuating methylation clocks for cell lineage tracing at high temporal resolution in human tissues. *Nature biotechnology*, pages 1–11, 2022.

[211] LTJ Kaaij, M van de Wetering, F Fang, B Decato, A Molaro, HJG van de Werken, JH van Es, J Schuijers, E de Wit, W de Laat, and et al. Dna methylation dynamics during intestinal stem cell differentiation reveals enhancers driving gene expression in the villus. *Genome Biology*, 14(5):R50, 2013.

[212] Gabbutt C, Schenck RO, Weisenberger D, Kimberley C, Berner A, Household J, Lakatos E, Robertson-Tessi M, Martin I, Patel R, Clark S, Latchford A, Barnes C, Leedham SJ, Anderson ARA, Graham TA, and Shibata D. Cell lineage tracing with molecular clocks based on fluctuating DNA methylation - flipflop, 2021.

[213] JS Speagle. dynesty: a dynamic nested sampling package for estimating bayesian posteriors and evidences. *Monthly Notices of the Royal Astronomical Society*, 493(3):3132–3158, 2020.

[214] N Tempest, M Jansen, AM Baker, CJ Hill, M Hale, D Magee, D Treanor, NA Wright, and DK Hapangama. Histological 3d reconstruction and in vivo lineage tracing of the human endometrium. *The Journal of Pathology*, 251(4): 440–451, 2020.

[215] L Ritsma, SIJ Ellenbroek, A Zomer, HJ Snippert, FJ de Sauvage, BD Simons, H Clevers, and J van Rheenen. Intestinal crypt homeostasis revealed at single-stem-cell level by in vivo live imaging. *Nature*, 507(7492):362–365, 2014.

[216] I Nishisho, Y Nakamura, Y Miyoshi, Y Miki, H Ando, A Horii, K Koyama, J Utsunomiya, S Baba, P Hedge, and et al. Mutations of chromosome 5q21 genes in fap and colorectal cancer patients. *Science*, 253(5020):665–669, 1991.

[217] KW Kinzler, MC Nilbert, LK Su, B Vogelstein, TM Bryan, DB Levy, KJ Smith, AC Preisinger, P Hedge, D McKechnie, and et al. Identification of fap locus genes from chromosome 5q21. *Science*, 253(5020):661–665, 1991.

[218] J Groden, A Thliveris, W Samowitz, M Carlson, L Gelbert, H Albertsen, G Joslyn, J Stevens, L Spirio, M Robertson, and et al. Identification and characterization of the familial adenomatous polyposis coli gene. *Cell*, 66(3): 589–600, 1991.

[219] V Korinek, N Barker, PJ Morin, D Van Wichen, R De Weger, KW Kinzler, B Vogelstein, and H Clevers. Constitutive transcriptional activation by a β-catenin-tcf complex in apc-/- colon carcinoma. *Science*, 275(5307):1784–1787, 1997.

[220] T Zhan, N Rindtorff, and M Boutros. Wnt signaling in cancer. *Oncogene*, 36 (11):1461–1473, 2017.

[221] OJ Sansom, KR Reed, AJ Hayes, H Ireland, H Brinkmann, IP Newton, E Batlle, P Simon-Assmann, H Clevers, IS Nathke, and et al. Loss of apc in vivo immediately perturbs wnt signaling, differentiation, and migration. *Genes & development*, 18(12):1385–1390, 2004.

[222] V Korinek, N Barker, P Moerer, E van Donselaar, G Huls, PJ Peters, and H Clevers. Depletion of epithelial stem-cell compartments in the small intestine of mice lacking tcf-4. *Nature genetics*, 19(4):379–383, 1998.

[223] D Pinto and H Clevers. Wnt control of stem cells and differentiation in the intestinal epithelium. *Experimental cell research*, 306(2):357–363, 2005.

[224] D Pinto, A Gregorieff, H Begthel, and H Clevers. Canonical wnt signals are essential for homeostasis of the intestinal epithelium. *Genes & development*, 17 (14):1709–1713, 2003.

[225] S Jaiswal and BL Ebert. Clonal hematopoiesis in human aging and disease. *Science*, 366(6465), 2019.

[226] Schenck RO, Gabbutt C, Weisenberger D, Kimberley C, Berner A, Household J, Lakatos E, Robertson-Tessi M, Martin I, Patel R, Clark S, Latchford A, Barnes C, Leedham SJ, Anderson ARA, Graham TA, and Shibata D. Cell lineage tracing with molecular clocks based on fluctuating DNA methylation – Flip flop blood model, 2021.

[227] G Yamato, T Kawai, N Shiba, J Ikeda, Y Hara, K Ohki, SI Tsujimoto, T Kaburagi, K Yoshida, Y Shiraishi, and et al. Genome-wide dna methylation analysis in pediatric acute myeloid leukemia. *Blood Advances*, 2022.

[228] Q Tan, BT Heijmans, JB Hjelmborg, M Soerensen, K Christensen, and L Christiansen. Epigenetic drift in the aging genome: a ten-year follow-up in an elderly twin cohort. *International journal of epidemiology*, 45(4):1146–1158, 2016.

[229] JS Welch, TJ Ley, DC Link, CA Miller, DE Larson, DC Koboldt, LD Wartman, TL Lamprecht, F Liu, J Xia, and et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell*, 150(2):264–278, 2012.

[230] DP Steensma, R Bejar, S Jaiswal, RC Lindsley, MA Sekeres, RP Hasserjian, and BL Ebert. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood, The Journal of the American Society of Hematology*, 126(1):9–16, 2015.

[231] SR Mitchell, J Gopakumar, and S Jaiswal. Insights into clonal hematopoiesis and its relation to cancer risk. *Current Opinion in Genetics & Development*, 66:63–69, 2021.

[232] PR Loh, G Genovese, and SA McCarroll. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature*, 584(7819):136–141, 2020.

[233] AG Bick, JS Weinstock, SK Nandakumar, CP Fulco, EL Bao, SM Zekavat, MD Szeto, X Liao, MJ Leventhal, J Nasser, and et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature*, 586(7831):763–768, 2020.

[234] F Zink, SN Stacey, GL Norddahl, ML Frigge, OT Magnusson, I Jonsdottir, TE Thorgeirsson, A Sigurdsson, SA Gudjonsson, J Gudmundsson, and et al.

Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood, The Journal of the American Society of Hematology*, 130 (6):742–752, 2017.

[235] KB Jacobs, M Yeager, W Zhou, S Wacholder, Z Wang, B Rodriguez-Santiago, A Hutchinson, X Deng, C Liu, MJ Horner, and et al. Detectable clonal mosaicism and its relationship to aging and cancer. *Nature genetics*, 44(6):651–658, 2012.

[236] CC Laurie, CA Laurie, K Rice, KF Doheny, LR Zelnick, CP McHugh, H Ling, KN Hetrick, EW Pugh, C Amos, and et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature genetics*, 44(6):642–650, 2012.

[237] A Niroula, A Sekar, MA Murakami, M Trinder, M Agrawal, WJ Wong, AG Bick, MM Uddin, CJ Gibson, GK Griffin, MC Honigberg, SM Zekavat, K Paruchuri, P Natarajan, and BL Ebert. Distinction of lymphoid and myeloid clonal hematopoiesis. *Nature Medicine*, 2021.

[238] PR Loh, G Genovese, RE Handsaker, HK Finucane, YA Reshef, PF Palamara, BM Birmann, ME Talkowski, SF Bakhoum, SA McCarroll, and et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature*, 559(7714):350–355, 2018.

[239] MA Fabre, JG de Almeida, E Fiorillo, E Mitchell, A Damaskou, J Rak, V Orru, M Marongiu, MS Vijayabaskar, JE Baxter, et al. The longitudinal dynamics and natural history of clonal hematopoiesis. *Blood*, 138:LBA–2, 2021.

[240] K Karner, TI George, and JL Patel. Current aspects of clonal hematopoiesis: implications for clinical diagnosis. *Annals of laboratory medicine*, 39(6):509–514, 2019.

[241] DP Steensma. Clinical implications of clonal hematopoiesis. In *Mayo Clinic Proceedings*, volume 93, pages 1122–1130. Elsevier, 2018.

[242] LA Forsberg, C Rasi, HR Razzaghian, G Pakalapati, L Waite, KS Thilbeault, A Ronowicz, NE Wineinger, HK Tiwari, D Boomsma, and et al. Age-related somatic structural changes in the nuclear genome of human blood cells. *The American Journal of Human Genetics*, 90(2):217–228, 2012.

[243] MA Myers, G Satas, and BJ Raphael. Inferring tumor evolution from longitudinal samples. *bioRxiv*, 2019.

[244] AG Deshwar, S Vembu, CK Yung, GH Jang, L Stein, and Q Morris. Phylowgs: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35, 2015.

[245] H Lee-Six and DG Kent. Tracking hematopoietic stem cells and their progeny using whole-genome sequencing. *Experimental hematology*, 83:12–24, 2020.

[246] HX Dang, BS White, SM Foltz, CA Miller, J Luo, RC Fields, and CA Maher. Clonevol: clonal ordering and visualization in cancer sequencing. *Annals of Oncology*, 28(12):3076–3082, 2017.

[247] DA Fulcher and A Basten. B cell life span: a review. *Immunology and cell biology*, 75(5):446–455, 1997.

[248] Z He, C Allers, C Sugimoto, N Ahmed, H Fujioka, WK Kim, ES Didier, and MJ Kuroda. Rapid turnover and high production rate of myeloid cells in adult rhesus macaques with compensations during aging. *The Journal of Immunology*, 200(12):4059–4067, 2018.

[249] L Khodadadi, Q Cheng, A Radbruch, and F Hiepe. The maintenance of memory plasma cells. *Frontiers in immunology*, page 721, 2019.

[250] SM van Neerven, NE de Groot, LE Nijman, BP Scicluna, MS van Driel, MC Lecca, DO Warmerdam, V Kakkar, LF Moreno, FA Vieira Braga, et al. Apc-mutant cells act as supercompetitors in intestinal tumour initiation. *Nature*, 594(7863):436–441, 2021.

[251] DJ Flanagan, N Pentinmikko, K Luopajärvi, NJ Willis, K Gilroy, AP Raven, L Mcgarry, JI Englund, AT Webb, S Scharaw, et al. Notum from apc-mutant cells biases clonal competition to initiate cancer. *Nature*, 594(7863):430–435, 2021.

[252] A Cagan, A Baez-Ortega, N Brzozowska, F Abascal, THH Coorens, MA Sanders, ARJ Lawson, LMR Harvey, S Bhosle, D Jones, et al. Somatic mutation rates scale with lifespan across mammals. *Nature*, pages 1–8, 2022.

[253] RL Siegel, KD Miller, HE Fuchs, and A Jemal. Cancer statistics, 2021. *CA: A Cancer Journal for Clinicians*, 71(1):7–33, 2021.

[254] K Raghav and MJ Overman. Small bowel adenocarcinomas—existing evidence and evolving paradigms. *Nature reviews Clinical oncology*, 10(9):534–544, 2013.

[255] C Tomasetti, L Li, and B Vogelstein. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, 355(6331):1330–1334, 2017.

[256] SN Catlin, P Guttorp, and JL Abkowitz. The kinetics of clonal dominance in myeloproliferative disorders. *Blood*, 106(8):2688–2692, 2005.

[257] J Xu, Y Wang, P Guttorp, and JL Abkowitz. Visualizing hematopoiesis as a stochastic process. *Blood advances*, 2(20):2637–2645, 2018.

[258] V Koerber, N Ansari-Pour, NA Jakobsen, R Moore, N Claudino, M Metzner, B Usukhbayar, MA Salazar, S Newman, BJL Kendrick, et al. A simple and direct method to define clonal selection in somatic mosaicism. *bioRxiv*, 2021.

[259] S Jaiswal, P Natarajan, AJ Silver, CJ Gibson, AG Bick, E Shvartz, M McConkey, N Gupta, S Gabriel, D Ardissino, et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *New England Journal of Medicine*, 377 (2):111–121, 2017.

[260] EK Cook, M Luo, and MJ Rauh. Clonal hematopoiesis and inflammation: Partners in leukemogenesis and comorbidity. *Experimental hematology*, 83:85–94, 2020.

[261] H Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4.

[262] TL Pedersen and D Robinson. *gganimate: A Grammar of Animated Graphics*, 2019. R package version 1.0.3.9000.

[263] P Eirew, A Steif, J Khattra, G Ha, D Yap, H Farahani, K Gelmon, S Chia, C Mar, A Wan, E Laks, J Biele, K Shumansky, J Rosner, A McPherson, C Nielsen, AJL Roth, C Lefebvre, A Bashashati, C de Souza, C Siu, R Aniba, J Brimhall, A Oloumi, T Osako, A Bruna, JL Sandoval, T Algara, W Greenwood, K Leung, H Cheng, H Xue, Y Wang, D Lin, AJ Mungall, R Moore, Y Zhao, J Lorette, L Nguyen, D Huntsman, CJ Eaves, C Hansen, MA Marra, C Caldas, SP Shah, and S Aparicio. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518:422 EP –, 2014.

[264] R Bravo, E Baratchart, J West, RO Schenck, AK Miller, J Gallaher, CD Gatenbee, D Basanta, M Robertson-Tessi, and ARA Anderson. Hybrid automata library: A modular platform for efficient hybrid modeling with real-time visualization. *bioRxiv*, page 411538, 2018.

[265] Y Chen, M Lemire, S Choufani, DT Butcher, D Grafodatskaya, BW Zanke, S Gallinger, TJ Hudson, and R Weksberg. Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray. *Epigenetics*, 8(2):203–209, 2013.

[266] R Pidsley, E Zotenko, TJ Peters, MG Lawrence, GP Risbridger, P Molloy, S Van Djik, B Muhlhausler, C Stirzaker, and SJ Clark. Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome dna methylation profiling. *Genome biology*, 17(1):1–17, 2016.

[267] C Lopez-Garcia, AM Klein, BD Simons, and DJ Winton. Intestinal stem cell replacement follows a pattern of neutral drift. *Science*, 330(6005):822, 2010.

[268] L Ricci-Vitiani, DG Lombardi, E Pilozzi, M Biffoni, M Todaro, C Peschle, and R De Maria. Identification and expansion of human colon-cancer-initiating cells. *Nature*, 445(7123):111–115, 2007.

[269] H Cedar and Y Bergman. Programming of dna methylation patterns. *Annual review of biochemistry*, 81:97–117, 2012.

[270] T Ushijima, N Watanabe, E Okochi, A Kaneda, T Sugimura, and K Miyamoto. Fidelity of the methylation pattern and its variation in the genome. *Genome research*, 13(5):868–874, 2003.

[271] J Skilling. Nested sampling for general bayesian computation. *Bayesian analysis*, 1(4):833–859, 2006.

[272] A Vehtari, A Gelman, D Simpson, B Carpenter, and PC Bürkner. Rank-normalization, folding, and localization: An improved (r) for assessing convergence of mcmc. *arXiv preprint arXiv:1903.08008*, 2019.

[273] A Gelman and DB Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.